

UNIVERSITY OF BAHRAIN

**College of Business
Administration**



**Department of Management
& Marketing**

Predicting Firms' Performance in GCC Countries Using Data Mining Techniques

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy Degree in Business Analytics

Submitted by

Jawaher Khaled Aref Alanzi

University ID: 19983609

Supervised by

Prof. Minwir Al-Shammari

(Professor)

University of Bahrain

Kingdom of Bahrain

May 2025



Thesis Discussion Committee Decision

Form Code: UDNGS013

Student Name:	Jawaher Khaled Aref Alanzi
Student ID. No.	19983609
Department:	Management and Marketing
College:	Business Administration
Program:	PhD in Business Analytics
Student GPA:	3.73
Thesis Title:	Predicting Firms' Performance in GCC Countries Using Data Mining Techniques

The Discussion Committee met on 25 May 2025 in S1B-281 upon discussing the student whose details are stated above in public. The discussion committee recommends the following:

The student is awarded the grade PASS in the thesis course MGTBAA 798.

- ☐ Without making any changes to the thesis.
- ☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

1.	Supervisor Name:	Pr. Minwir Al-Shammari	Sign.:	<i>minwir</i>
2.	Internal Examiner:	Dr. Abdulsattar Al Azzawi	Sign.:	<i>abdulsattar</i>
3.	External Examiner:	Pr. Nasro Min Allah	Sign.:	<i>Nasro Min Allah</i>



Dissertation Deposit Form

Form Code: UDNGS025

Author Name:	Jawaher Khaled Aref Alanzi
Thesis Title:	Predicting Firms' Performance in GCC Countries Using Data Mining Techniques
Year of Publication:	2025
Department:	Management and Marketing
College:	Business Administration
E-mail:	19983609@stu.uob.edu.bh
Contact No.	39400390

Author's Declaration: I agree to the following Conditions:

The above thesis will be available (in the University of Bahrain Library and externally) and reproduced as necessary at the discretion of the University of Bahrain. It may also be digitized by the University of Bahrain and made available via the university's repository or via other parties on the Internet.

This condition shall apply to all copies including electronic copies.

The above thesis has been provided on the understanding that it is copyright protected material and that no quotation from this thesis may be published without paper acknowledgement.

Sign. of the Author:		Date:	29 May 2025
----------------------	---	-------	-------------

ABSTRACT

Purpose: This research aims to examine the accuracy of Data Mining (DM) techniques in predicting the financial performance of publicly listed companies in the Gulf Cooperation Council (GCC) countries (namely, Kingdom of Bahrain, Kuwait, Oman, Qatar, Kingdom of Saudi Arabia, and the United Arab Emirates (UAE)).

Design/Methodology/Approach: The study analyzed previous literature on predicting firms' performance using DM techniques, combining Systematic Literature Review (SLR) and bibliometric analysis to identify literature gaps. A conceptual model was developed to examine the relationship between independent variables: firm-specific (liquidity, leverage, solvency, asset utilization, and market value), macroeconomic (GDP and inflation), and control variables (firm size, age, industry type, and firm's country), and the dependent variable measured by Return On Investment (ROI). The dataset was sourced from the London Stock Exchange Group (LSEG) and the World Development Indicators (WDI) and covered 816 publicly listed companies across diverse industries of the GCC countries from 2003 to 2022. The study evaluated four prediction models using seven DM techniques implemented in Python. These included single DM techniques: Linear Regression (LR), Artificial Neural Networks (ANN), Regression Trees (RT), and k Nearest Neighbors (kNN), along with ensemble models such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost).

Findings: CatBoost outperforms other models, achieving the highest prediction accuracy, with R-squared (0.542) for the all-GCC dataset model. On the other hand, Bahrain's country-specific model demonstrated the best prediction accuracy (R-squared = 0.781, lowest error rates). In contrast, Saudi Arabia's model showed the weakest prediction accuracy (R-squared = 0.328, highest error rates). In terms of variable importance, asset utilization and solvency demonstrate a positive influence on profitability across most GCC countries. These results indicate that firm-specific factors have the most significant influence on profitability. Macroeconomic variables, such as GDP and inflation, have a

minimal influence on profitability, whereas control variables (age, industry type, country, and size) exert a moderate effect, with a negative impact of age on profitability.

Originality/Value: From a theoretical perspective, this study contributes novelty to the literature by exploring the capabilities of DM techniques in predicting ROI as a financial measure for profitability. This study examines the firm-specific factors that influence publicly listed companies' performance in the GCC region and assesses the macroeconomic variables that affect their performance across diverse industries. The study emphasizes the importance of implementing country-specific DM models rather than using an all-GCC dataset model to account for economic variations across countries. Furthermore, the findings of this research hold substantial empirical significance, offering valuable insights for informed decision-making among various stakeholders. Policymakers and regulators in the GCC, such as the Capital Market Authorities (CMA) and bankruptcy commissions, may utilize this information to take early interventions, preventing business failures and enabling proactive measures for firms identified as being at risk (e.g. low profitability prediction). Additionally, managers can leverage insights based on the importance of independent variables to optimize internal processes and resource allocation, ultimately enhancing financial performance. Furthermore, shareholders can make more informed strategic decisions regarding stock transactions based on the firm's predicted profitability outcomes.

DEDICATION

This thesis is lovingly dedicated to my beloved parents, Shaha Al-Azzam and Khaled Al-Anzi. May ALLAH protect them and continue to brighten my life with their endless love and warm smiles.

I also dedicate this work to my dearest sisters and brother Sami, whose unwavering moral support has been my strength throughout this journey. My success is a reflection of the love and encouragement of my family.

ACKNOWLEDGMENT

All praise and gratitude are due to Almighty ALLAH for granting me the strength, patience, and blessings to complete this work.

I would like to express my gratitude to my mentor and thesis supervisor, Professor Minwir Al-Shammari. His invaluable mentorship, insightful guidance, and unwavering support have been crucial in shaping the progress and success of my research. His dedication and encouragement have significantly contributed to my academic and personal growth.

I am grateful to everyone who supported me throughout my PhD journey. My heartfelt thanks go to my beloved sister and soulmate, *Fatima*, for her unwavering support and encouragement. I also profoundly thank my sister, *Meshael*, and her husband, Professor Khaled Al-Rubaie (King Faisal University), for their kindness and invaluable support.

Lastly, I am sincerely thankful to Ms. Thajba Aldosari and Ms. Hessa Aldosari from the Royal University for Women for their invaluable assistance in accessing essential data for this study.

PUBLICATIONS

JOURNAL PAPERS

1. Alanzi, J. K., & Al-Shammari, M. (2025). Data Mining for Predicting the Profitability of Publicly Listed Companies in the GCC Region. Manuscript submitted for publication to Engineering Applications of Artificial Intelligence.
2. Alanzi, J. K., & Al-Shammari, M. (2025). Performance Prediction Using Data Mining Techniques: A Systematic Literature Review with Bibliometric Analysis. Manuscript submitted for publication to the International Journal of Information Technology.

CONFERENCE PAPERS

1. Alanzi, J. K., & Al-Shammari, M. (2024, December). Machine Learning Techniques in Predicting Firms' Performance: A Bibliometric Analysis. In the 2024 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1-6). IEEE.
2. Alanzi, J. K., & Al-Shammari, M. (2025, April). Predicting Firm Profitability Using Data Mining in the GCC Region. In the 2025 International Conference on Computational Intelligence Approaches and Applications (ICCIAA). IEEE.
3. Alanzi, J. K., & Al-Shammari, M. (under review). Interpreting the Profitability Prediction Model for Bahraini Publicly Listed Firms Using SHAP. 6th International Conference on Data Analytics for Business and Industry (ICDABI 2025). IEEE.

TABLE OF CONTENTS

Thesis Discussion Committee Decision	A
Dissertation Deposit Form	B
Abstract.....	C
Dedication	E
Acknowledgment.....	F
Publications	G
Table of Contents	H
List of Tables	N
List of Figures.....	P
Abbreviations	S
CHAPTER 1: Introduction.....	1
1.1 Introduction.....	1
1.2 Research Problem	3
1.3 Research Significance.....	5
1.4 Settings of the Study	7
1.5 Research Objectives.....	8
1.6 Thesis Structure	9
1.7 Research Terminology	10
1.8 Summary	11
CHAPTER 2: Theoretical Background and Literature Review	13
2.1 Introduction.....	13
2.2 Firm's Performance	13
2.2.1 Concepts.....	13
2.2.2 Determinants of Firm's Performance.....	15
2.2.2.1 Dependent Variable	17
2.2.2.2 Independent Variables	20
2.2.2.3 Control Variables.....	26

2.2.3	Theoretical Perspectives	28
2.3	Data Mining	31
2.3.1	Concepts.....	31
2.3.2	Tasks and Methods	32
2.3.3	Prediction Techniques of a Firm's Performance	35
2.4	Previous Studies on Firms' Performance Prediction Using DM Techniques.....	38
2.4.1	Bibliometric Analysis	39
2.4.1.1	Evolution of Firm Performance Prediction Using DM	43
2.4.1.2	Most Cited Articles in Literature.....	45
2.4.1.3	Thematic Clusters of Literature	46
2.4.1.4	Literature Trends	50
2.4.2	Systematic Literature Review	52
2.4.2.1	Geographical Datasets in Literature	61
2.4.2.2	Associated Variables in Prediction Models.....	62
2.4.2.3	Utilization of DM Techniques and Analytic Tools	68
2.4.2.4	Preprocessing of the Dataset in Literature.....	68
2.4.2.5	Evaluation Metrics of Prediction Models.....	70
2.5	Research Gap	70
2.6	Conceptual Framework.....	75
2.7	Summary	77
CHAPTER 3: The GCC Market and Economic Context.....		78
3.1	Introduction.....	78
3.2	GCC Stock Markets	79
3.2.1	Kingdom of Bahrain	81
3.2.2	Kuwait.....	82
3.2.3	Oman.....	82
3.2.4	Qatar.....	83
3.2.5	Kingdom of Saudi Arabia	83
3.2.6	United Arab Emirates	84
3.3	GCC Economics	84

3.4	Regulatory and Legislative Developments	88
3.5	Summary	88
CHAPTER 4: Research Methodology		89
4.1	Introduction.....	89
4.2	Research Paradigm and Philosophies	89
4.3	Research Methods.....	90
4.4	Framework of Prediction Model.....	91
4.5	Data Collection and Analysis Tools	93
4.5.1	Dataset Sources	93
4.5.2	Study Variables	98
4.5.3	DM Software.....	100
4.6	Data Preprocessing	101
4.6.1	Handling Missing Values.....	102
4.6.2	Handling Outliers.....	103
4.6.3	Handling Text and Categorical Attributes	105
4.6.4	Variable Selection.....	105
4.6.5	Variable Scaling (Normalization)	107
4.7	Data Sampling and Splitting	108
4.7.1	Dataset Sampling	108
4.7.2	Dataset Splitting.....	109
4.7.2.1	Random Split	109
4.7.2.2	Chronological Split.....	109
4.8	Prediction Models	111
4.8.1	Hyperparameter Tuning	112
4.8.2	DM Techniques.....	113
4.8.2.1	Linear Regression	115
4.8.2.2	Artificial Neural Networks	116
4.8.2.3	Regression Trees.....	118
4.8.2.4	k-Nearest Neighbor.....	119
4.8.2.5	Random Forest.....	120
4.8.2.6	eXtreme Gradient Boosting.....	122

4.8.2.7	Categorical Boosting	123
4.9	DM Evaluation Metrics	124
4.9.1	R-squared	125
4.9.2	MAE.....	126
4.9.3	MSE	127
4.9.4	RMSE.....	127
4.10	Variable Importance and Sensitivity Analysis Using SHAP	128
4.11	Summary	131
CHAPTER 5: Data Analysis Results		132
5.1	Introduction.....	132
5.2	Descriptive Statistics.....	132
5.3	Data Preprocessing Results.....	140
5.3.1	I-RS-GCC Model	140
5.3.2	NI-RS-GCC Model	142
5.3.3	By Country (C) Model	144
5.3.4	CS-GCC Model.....	148
5.4	Summary	148
CHAPTER 6: Prediction Models Results and Discussion		149
6.1	Introduction.....	149
6.2	I-RS-GCC Model Experimental Results	150
6.2.1	Evaluation of Model Performance	150
6.2.2	Variable Importance.....	152
6.3	NI-RS-GCC Model Experimental Results	154
6.3.1	Evaluation of Model Performance	154
6.3.2	Variable Importance.....	156
6.4	C Model Experimental Results.....	158
6.4.1	Kingdom of Bahrain	158
6.4.1.1	Evaluation of Model Performance	158
6.4.1.2	Variable Importance	160
6.4.2	Kuwait.....	161
6.4.2.1	Evaluation of Model Performance	161

6.4.2.2	Variable Importance	163
6.4.3	Oman.....	164
6.4.3.1	Evaluation of Model Performance.....	164
6.4.3.2	Variable Importance	166
6.4.4	Qatar.....	168
6.4.4.1	Evaluation of Model Performance.....	168
6.4.4.2	Variable Importance	170
6.4.5	Kingdom of Saudi Arabia	171
6.4.5.1	Evaluation of Model Performance.....	171
6.4.5.2	Variable Importance	173
6.4.6	United Arab Emirates	174
6.4.6.1	Evaluation of Model Performance.....	174
6.4.6.2	Variable Importance	176
6.5	CS-GCC Model Experimental Results	177
6.5.1	I-CS-GCC Model Evaluation Results.....	177
6.5.2	NI-CS-GCC Model Evaluation Results	179
6.6	Discussion	180
6.6.1	DM Techniques.....	181
6.6.2	Handling Missing Values Methods.....	182
6.6.3	Sampling Methods	183
6.6.4	Data Splitting Methods	184
6.6.5	Variable Importance.....	187
6.7	Summary	190
CHAPTER 7: Conclusion		192
7.1	Introduction.....	192
7.2	Theoretical Implications	194
7.3	Managerial Implications	198
7.4	Limitations and Future Research	200
7.5	Summary	203
References.....		205
Appendices.....		220

Appendix 1. Thesis Dataset	221
Appendix 2. Optimal Hyperparameters.....	224
Appendix 3. Python Code.....	228
Appendix 4. Descriptive Charts of Data Distribution (C Models).....	230
Appendix 5. Pearson Correlation Matrices (C Models).....	234
Appendix 6. Distribution of Variables after Winsorization	240
Arabic Abstract.....	242

LIST OF TABLES

Table (2.1): Comparison of Deployed Review Methods (Donthu et al., 2021)	39
Table (2.2): Bibliometric Techniques Used to Map the Analysis	41
Table (2.3): Practical Criteria for Bibliometric Analysis	42
Table (2.4): Primary Information About the Collected Articles	42
Table (2.5): Top 5 Cited Publications in Performance Using DM Techniques	46
Table (2.6): Thematic Clusters of DM Research in Firm's Performance Prediction	49
Table (2.7): SLR Inclusion/ Exclusion Criteria	54
Table (2.8): SLR Quality Criteria	55
Table (2.9): Firm Performance Prediction Studies Utilizing DM Techniques	57
Table (2.10): Studies Based on the Dataset Region	62
Table (2.11): Studies Categorized by Dependent Variables and Data Mining Tasks	63
Table (2.12): Independent Variables Used in Previous Studies	65
Table (2.13): Data Split Types Used in Previous Studies	69
Table (2.14): Literature Summary and Research Gap	74
Table (3.1): Listed Firms in GCC Market & Allowable Share Ownership for GCC Citizens in 2023 (GCC-STAT, 2024)	79
Table (3.2): GCC Stock Markets Performance	81
Table (4.1): Number of Active Listed Firms Over 21-Year Period (LSEG, 2023)	94
Table (4.2): Variables Sources	95
Table (4.3): The Proportion of Industry Types Within the Dataset for FY 2022	97
Table (4.4): Definition of Study's Variables	99

Table (4.5): Nulls and Data Type for Research Variables	102
Table (4.6): Number of Observations (rows) in each GCC Country Dataset	108
Table (4.7): Chronological Approach to Split Training and Testing Sets..	110
Table (4.8): Libraries Deployed for DM Models	115
Table (5.1): Descriptive Statistics of Numerical Variables Before Data Preprocessing	134
Table (5.2): Correlated Variables ≥ 0.7 for all-GCC Dataset	139
Table (5.3): Remaining Nulls in Variables after Handling Missing Values in ROI.....	142
Table (6.1): Performance Results of I-RS-GCC Models.....	150
Table (6.2): Performance Results of NI-RS-GCC Models.....	154
Table (6.3): Evaluation Results of Bahrain Models	159
Table (6.4): Evaluation Results of Kuwait Models	161
Table (6.5): Evaluation Results of Oman Models	165
Table (6.6): Evaluation Results of Qatar Models	169
Table (6.7): Evaluation Results of Saudi Models.....	171
Table (6.8): Evaluation Results of UAE Models.....	174
Table (6.9): Evaluation Results for I-CS-GCC Models	178
Table (6.10): Evaluation Results for NI-CS-GCC Models	180
Table (6.11): Evaluation Results of CatBoost I-RS-GCC vs NI-RS-GCC Models.....	182
Table (6.12): Evaluation Results for the CatBoost in C and NI-CS-GCC Models.....	183
Table (6.13): Evaluation Results for I-RS-GCC, NI-RS-GCC, I-CS-GCC and NI-CS-GCC Models	185
Table (6.14): Variable Importance Rankings Using CatBoost for C and NI-CS-GCC Models	187

LIST OF FIGURES

Fig. 2.1: Relationship of factors affecting ROI (Flesher & Previts, 2013) ..	19
Fig. 2.2: Data Mining Taxonomy Tasks, Methods, and Algorithms adopted from (Sharda et al., 2021)	34
Fig. 2.3: Bibliometric analysis steps (Donthu et al., 2021)	39
Fig. 2.4: Annual scientific production of DM research in the field of firm's performance prediction plotted using R-Bibliometrix (Aria & Cuccurullo, 2017)	45
Fig. 2.5: Bibliographic coupling network.....	47
Fig. 2.6: Keyword evolution of DM research in a firm's performance prediction	51
Fig. 2.7: SLR phases, (Kitchenham & Charters, 2007).....	53
Fig. 2.8: The review process follows PRISMA guidelines (Page et al., 2021)	55
Fig. 2.9: Conceptual Framework of Firms' Performance Model and its Determinants	76
Fig. 3.1: Number of actively listed firms for the period 2003-2023 in GCC countries	80
Fig. 3.2: Real GDP (US\$ in Billions) for GCC countries, (World Bank, 2024)	85
Fig. 3.3: Real GDP Growth (Percent change, year over year; weighted averages), (IMF, 2024)	86
Fig. 3.4: Inflation rate for GCC countries (World Bank, 2024)	87
Fig. 3.5: CPI inflation in GCC, USA, and the world (year-over-year, average), (Fareed et al., 2023)	87
Fig. 4.1: Framework of DM Model	92
Fig. 4.2: The proportion of all GCC industry types within the dataset for FY 2022 (LSEG, 2023).....	96
Fig. 4.3: ANN Structure for the GCC Profitability Prediction Model	117

Fig. 5.1: Variables Distribution for GCC Dataset	136
Fig. 5.1 (cont.): Variables Distribution for GCC Dataset.....	137
Fig. 5.2: Pearson Correlation Matrix with P-Values for all GCC dataset variables	138
Fig. 5.3: Distribution of I-RS-GCC Model after Dataset Cleaning.....	141
Fig. 5.4: ROI Distribution for All GCC Before Winsorization.....	141
Fig. 5.5: ROI Distribution for All GCC After Winsorization	141
Fig. 5.6: Distribution of NI-RS-GCC Models after Dataset Cleaning	143
Fig. 5.7: Boxplot of ROI Distribution for NI-RS-GCC Model	143
Fig. 5.8: Variables Distribution for Bahrain After Dataset Cleaning.....	144
Fig. 5.9: Variables Distribution for Kuwait After Dataset Cleaning.....	145
Fig. 5.10: Variables Distribution for Oman After Dataset Cleaning.....	146
Fig. 5.11: Variables Distribution for Qatar After Dataset Cleaning.....	146
Fig. 5.12: Variables Distribution for Saudi Arabia After Dataset Cleaning	147
Fig. 5.13: Variables Distribution for UAE After Dataset Cleaning	147
Fig. 6.1: Comparison Results for the I-RS-GCC Models.....	151
Fig. 6.2: (a) Actual Versus Predicted ROI (b) Histogram of Residuals Distribution for I-RS-GCC CatBoost Model.....	152
Fig. 6.3: (a) SHAP Variable Importance (b) SHAP Summary Plots for I-RS-GCC Model using CatBoost	153
Fig. 6.4: Comparison Results for the NI-RS-GCC Model	155
Fig. 6.5: (a) Actual Versus Predicted ROI (b) Residuals Distribution for NI-RS-GCC CatBoost Model.....	156
Fig. 6.6: (a) SHAP Variable Importance (b) SHAP Summary Plots for NI-RS-GCC Model using CatBoost	157
Fig. 6.7: Comparison Results for Bahrain Models	158
Fig. 6.8: (a) Actual versus Predicted ROI (b) Residuals Distribution for Bahrain CatBoost Model	159

Fig. 6.9: (a) SHAP Variable Importance (b) SHAP Summary Plots for Bahrain Model using CatBoost	161
Fig. 6.10: Comparison Results for Kuwait Models	162
Fig. 6.11: (a) Actual versus Predicted ROI (b) Residuals Distribution for Kuwait CatBoost Model for Kuwait CatBoost Model	163
Fig. 6.12: (a) SHAP Variable Importance (b) SHAP Summary Plots for Kuwait Model using CatBoost.....	164
Fig. 6.13: Comparison Results for Oman Models	165
Fig. 6.14: (a) Actual versus Predicted ROI (b) Residuals Distribution for Oman XGBoost Model	166
Fig. 6.15: (a) SHAP Variable Importance (b) SHAP Summary Plots for Oman Model using XGBoost	167
Fig. 6.16: Comparison Results for Qatar Model	168
Fig. 6.17: (a) Actual versus Predicted ROI (b) Residuals Distribution for Qatar CatBoost Model	169
Fig. 6.18: (a) SHAP Variable Importance (b) SHAP Summary Plots for Qatar Model using CatBoost	170
Fig. 6.19: Comparison Results for Saudi Models.....	172
Fig. 6.20: (a) Actual versus Predicted ROI (b) Residuals Distribution for Saudi CatBoost Model	173
Fig. 6.21: (a) SHAP Variable Importance (b) SHAP Summary Plots for Saudi Model using CatBoost	173
Fig. 6.22: Comparison Results for UAE Model	175
Fig. 6.23: (a) Actual versus Predicted ROI (b) Residuals Distribution for UAE CatBoost Model	175
Fig. 6.24: (a) SHAP Variable Importance (b) SHAP Summary Plots for UAE Model from CatBoost	176
Fig. 7.1: Proposed framework of the early warning system for underperforming GCC-listed companies.....	200

ABBREVIATIONS

ADX	Abu Dhabi Securities Exchange
ADXG	Abu Dhabi Stock Exchange General Index
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AT	Asset Turnover
BAX	Bahrain Bourse All Share Index
BSE	Bahrain Stock Exchange
C Model	By Country Model
CaR	Cash Ratio
CAT	Current Asset Turnover
CatBoost	Categorical Boosting
CMA	Capital Market Authority
CPI	Consumer Price Index
CRISP-DM	Cross-Industry Standard Process for Data Mining
CuR	Current Ratio
CV	Cross-Validation
DFM	Dubai Financial Market
DGE	Dynamic General Equilibrium
DM	Data Mining
DR	Debt Ratio
DSM	Doha Securities Market
DSS	Decision Support System
DT	Decision Trees
EBDA	Earnings Before Depreciation and Amortization

EBIT	Earnings Before Interest and Taxes
EDA	Exploratory Data Analysis
ESCA	Emirates Securities and Commodities Authority
FAT	Fixed Asset Turnover
FY	Fiscal Year
GCC	Gulf Cooperation Council
GDP	Gross Domestic Product
ICR	Interest Coverage Ratio
I-CS-GCC	Imputation with Chronological Split- GCC Model
IPO	Initial Public Offering
IR	Inflation Rate
I-RS-GCC	Imputation with Random Split- GCC Model
IT	Industry Type
KDE	Kernel Density Estimate
KIA	Kuwait Investment Authority
kNN	k-Nearest Neighbors
KSE	Kuwait Stock Exchange
KWSE	Kuwait All Share Index
LN	Natural Logarithm
LR	Linear Regression
LSEG	London Stock Exchange Group
MAE	Mean Absolute Error
MDA	Multivariate Discriminant Analysis
MENA	Middle East and North Africa
ML	Machine Learning
MSE	Mean Squared Error

MSM	Muscat Securities MSM 30 Index
MSM	Muscat Securities Market
NI-CS-GCC	Non-Imputation with Chronological Split- GCC Model
NI-RS-GCC	Non-Imputation with Random Split- GCC Model
NOAT	Net Operating Asset Turnover
OPEC	Organization of the Petroleum Exporting Countries
P/B Ratio	Price-to-Earnings Ratio
P/E Ratio	Price-to-Earnings Ratio
P/S Ratio	Price-to-Sales Ratio
PBDITA	Profits Before Depreciation, Interest, Tax, and Amortization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QR	Quick Ratio
QSI	Qatar Exchange Index
RBV	Resource-Based View
RF	Random Forest
RMSE	Root Mean Squared Error
ROA	Return on Asset
ROE	Return on Equity
ROI	Return on Investment
ROS	Return on Sales
RS	Robust Scaling
RT	Random Trees
SAMA	Saudi Arabian Monetary Authority
SEMMA	Sample, Explore, Modify, Model, and Assess
SHAP	SHapley Additive exPlanations

SLR	Systematic Literature Review
SVM	Support Vector Machines
TACE	Total Assets to Common Equity
Tadawul	Saudi Stock Exchange
TASI	Tadawul All Share Index
TATE	Total Assets to Total Equity
TDPTE	Total Debt Percentage of Total Equity
TRBC	Thomson Reuters Business Classification
UAE	United Arab Emirates
WDI	World Development Indicators
WoS	Web of Science
WPI	Wholesale Price Index
XGBoost	eXtreme Gradient Boosting

CHAPTER 1: INTRODUCTION

1.1 Introduction

Firm performance constitutes a foundational pillar of economic resilience and diversification within the GCC region, where capital markets function as key drivers of growth and investor confidence (GCC-STAT, 2024). A firm's financial performance is a crucial factor that investors closely evaluate, as it serves as a key indicator of long-term economic development. It also holds significant importance for various stakeholders, including creditors, employees, managers, and governments (Esposito et al., 2025).

In this study, firm financial performance is measured using Return on Investment (ROI), a widely accepted financial metric that captures a firm's efficiency in utilizing capital to generate profit. ROI serves as a core indicator in both financial and management accounting, offering quantifiable insights into firms' efficiency and profitability (Lipták, 2022; Valentowitsch et al., 2024).

Within the GCC's capital markets, the interaction between firms, investors, and shareholders is fundamental to economic activity. Investors, in particular, prioritize firms with strong profitability prospects, as their primary goal is to achieve higher stock returns (Haghighinasab & Khobbakht, 2024). Therefore, understanding and predicting ROI is essential not only for evaluating individual firm performance but also for enhancing broader financial decision-making and market stability across the region.

The Gulf Cooperation Council (GCC) was established in 1981 by Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates (UAE) to foster regional cooperation and integration across political, economic, and social dimensions. It was initially driven by the shared cultural and historical ties among its member states and the strategic objective of leveraging their vast oil and gas resources to enhance global economic relevance. Over the years, the region has evolved into a key economic bloc, significantly contributing to global energy markets and geopolitical stability (The Supreme Council, 2024). In recent decades, the GCC has prioritized economic diversification to reduce its reliance on hydrocarbons (IMF, 2024). GCC countries are advancing broad reforms for socio-economic transformation, guided by initiatives such as Saudi Vision 2030, the UAE's "We the UAE 2031," and Oman Vision 2040. Post-pandemic implementation has accelerated, with a focus on social and business-friendly policies, fiscal sustainability, and strategic investments in digital and green infrastructure. Sustained reform efforts will boost productivity, economic diversification, and energy transition readiness (IMF, 2023). Bahrain's Economic Vision 2030 aims to transition from an oil-dependent economy to a productive, globally competitive one driven by a pioneering private sector. The vision aims to expand the middle class by creating high-wage jobs and enhancing productivity. Rooted in sustainability, competitiveness, and fairness, it ensures that all Bahrainis can achieve economic security and fulfill their potential (Government of Bahrain, 2008).

In recent years, there has been an increased focus on predicting firms' performance (Aljawazneh et al., 2021; Ainan et al., 2024; Arora & Saurabh, 2022; Assous, 2022; Balasubramanian et al., 2019; Barboza et al., 2017;

Chiadamrong & Wattanawarangkoon, 2023; du Jardin, 2021; Garcia, 2022; Halteh et al., 2018; Jabeur et al., 2021; J. P. Li et al., 2020; Khalil et al., 2022; Kim et al., 2018; Ledhem, 2022; M. Zhou et al., 2022; Nguyen et al., 2023; Park et al., 2021; P. V. H. Son & Duong, 2024; Saha et al., 2023; Smiti & Soui, 2020; H. Son et al., 2019; Wang & Wu, 2017), driven by the rise of Big Data (BD) (L. Zhou et al., 2017) and Data Mining (DM) (Sharda et al., 2021).

This chapter defines the research problem, significance, and study settings, outlines the study's objectives and structure, and clarifies key terminology. It lays the groundwork for understanding the research methodology and the study's contributions.

1.2 Research Problem

Recent regulatory reforms across the Capital Market Authorities (CMAs) of the GCC countries, including enhanced disclosure requirements, cross-border cooperation, and the implementation of comprehensive bankruptcy laws, have been introduced to strengthen transparency, corporate governance, and investor protection (GCC-STAT, 2024). The rollout of bankruptcy legislation began in the UAE in September 2016 (Ministry of Economy, 2016), followed by Bahrain in May 2018 (Legislation and Legal Opinion Commission, 2022), Saudi Arabia in September 2018 (Bankruptcy Commission, 2018), Oman in 2019 (Qaboos bin Said, 2019), and Kuwait in 2020 (KDIPA, 2020).

Despite these regulatory advancements, firm delisting's and financial instability persist as ongoing concerns. Between 2018 and 2021, Bahrain recorded 17 bankruptcy cases (Ministry of Justice and Islamic Affairs Bahrain, 2021). In 2023 alone, six firms were delisted in Kuwait (Boursa

Kuwait, 2024), two in Saudi Arabia (Saudi Exchange, 2024), and five in Oman (MSX, 2024). These developments underscore the critical importance of firm-level financial health, not only for shareholders but also for broader economic stability across the region.

While technological advancements and the increased availability of financial data have enabled the application of predictive modeling, a notable gap remains in research that applies DM techniques across diverse sectors in the GCC. Existing studies have predominantly focused on the banking sector (e.g., Assous, 2022; Li et al., 2020), often overlooking fundamental performance indicators such as ROI, a widely recognized measure of profitability and operational efficiency that remains underutilized in predictive research on GCC firms.

This thesis aims to address these gaps by developing predictive analytics models using DM techniques to forecast profitability for publicly listed companies across the GCC from 2003 to 2022. Emphasizing ROI as a core performance metric, the study extends the scope of analysis beyond the banking sector to provide a more comprehensive understanding of firm performance. In doing so, it evaluates the predictive accuracy of various DM methods and identifies the most influential financial drivers. The findings aim to offer stakeholders data-driven insights to enhance decision-making, improve early warning systems, and align with the region's evolving regulatory landscape.

1.3 Research Significance

This research contributes to the literature on firm performance prediction by assessing single and ensemble regression DM techniques in the GCC context. Using financial datasets from publicly listed companies in GCC stock markets, the dataset sourced from the London Stock Exchange Group (LSEG) spans annual records from 2003 to 2022 (LSEG, 2023).

This study is novel in its comprehensive scope. It analyzes all publicly listed companies in the GCC, in addition to each country, rather than focusing on a particular country or industry.

Furthermore, this study makes a novel contribution to the field of financial analytics by developing four predictive models that utilize seven distinct DM techniques. These models integrate a combination of sampling methods, data-splitting strategies (both random and chronological), and diverse approaches to handling missing data. Sensitivity analysis is also incorporated to evaluate the influence of different features on model predictions. By applying this comprehensive modeling framework to a dataset of publicly listed firms in the GCC region spanning the period from 2003 to 2022, the study advances existing literature on financial performance prediction. To conduct a sensitivity analysis, SHAP (Shapley Additive exPlanations) values are employed to quantify the impact of each input feature on the model's output. SHAP is chosen for its ability to unlock the black box of complex Machine Learning (ML) models by providing consistent, locally accurate, and model-specific explanations (Borgonovo et al., 2024). It captures both the magnitude and direction of feature influence across individual predictions, offering a deeper understanding of how specific inputs

drive model behavior. This approach emphasizes the pivotal role of model design and data preprocessing in determining predictive accuracy, offering actionable insights to inform data-driven decision-making in financial contexts.

Several research gaps were highlighted in existing literature. It is presented in a lack of DM techniques applied to predict the financial performance of all publicly listed GCC firms using fundamental factors. A review of research from 2017 to 2024 in leading journals (IEEE, Elsevier, Springer Nature, Emerald, Wiley, and Taylor & Francis) reveals that only two studies have focused on the performance of GCC banks (Assous, 2022; Li et al., 2020). Additionally, ROI has not been examined as a dependent variable despite its broad applicability in measuring profitability. While financial ratios have been widely used as independent variables, key metrics such as market value and specific macroeconomic indicators remain untested. Moreover, the CatBoost regressor, a DM technique, has not been examined for predicting firm performance; it has only been applied to forecasting solvency ratios (e.g., Khalil et al., 2022).

This study fills these gaps and contributes to the existing literature by evaluating the accuracy of DM techniques in predicting the financial performance of all publicly listed companies in the GCC. By focusing on predicting ROI as a measure of profitability, this research introduces a novel perspective to the field. Additionally, it explores the influence of previously untested financial metrics and macroeconomic indicators on profitability. Furthermore, it is the first study to apply the CatBoost regressor to ROI prediction.

More details regarding the literature gaps and the current study's contributions are provided in Section 2.5.

In addition, the research contributes to Bahrain's Economic Vision 2030 by providing data-driven insights into firms' performance (Government of Bahrain, 2008). Identifying and evaluating key profitability drivers supports transitioning from an oil-dependent economy to a competitive, knowledge-based market.

Overall, this research makes a significant contribution to both theory and practice in predicting firm financial performance, offering valuable insights for both academic and practical applications.

1.4 Settings of the Study

This research aims to examine the accuracy of DM techniques in predicting the financial performance of GCC-listed firms from 2003 to 2023. It analyzes key performance determinants from financial statements, including income statements and balance sheets of 816 publicly listed firms on GCC stock markets, as well as GCC macroeconomic factors. The analysis is limited to the fiscal years 2003 to 2022, as financial statements for 2023 were not yet available as of December 5, 2023. Datasets were sourced from LSEG Workspace, formerly Refinitiv (LSEG, 2023), a leading financial portal offering comprehensive information on GCC-listed firms. Macroeconomic datasets were sourced from World Development Indicators (WDI) (World Bank, 2024). The research used the Python programming language to conduct statistical analysis and inferential tests and to build and evaluate the prediction models.

1.5 Research Objectives

This thesis aims to examine the accuracy of DM techniques in predicting the financial performance of the GCC publicly listed companies, measured by profitability, from 2003 to 2022. To accomplish this primary objective, the thesis intends to fulfill the following specific research objectives:

4. To conduct a Systematic Literature Review (SLR) and bibliometric analysis to evaluate previous studies on predicting a firm's performance using DM and identify existing research gaps.
5. To utilize descriptive and inferential statistical tests to visualize the relationship among the study variables and compare the performance of publicly listed companies in the GCC.
6. To investigate the relationship between independent variables (e.g., liquidity, leverage, solvency, asset utilization, and market value) and firms' financial performance measured by profitability.
7. To employ single DM regression models including Linear Regression (LR), Artificial Neural Networks (ANN), Regression Trees (RT), and k-Nearest Neighbor (kNN), along with ensemble DM regression models including Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost); to predict firms' financial performance using dataset derived from financial market annual reports and firms' profiles.
8. To assess the predictive accuracy of different DM models for GCC publicly listed firms' profitability.

1.6 Thesis Structure

To achieve the research objectives, this thesis is organized into six chapters, providing a clear and cohesive roadmap from introduction to conclusion, as follows:

Chapter 1 presents the research problem, emphasizing the need to enhance the financial health of publicly listed firms in the GCC to support economic stability and mitigate insolvency risks. It then highlights the study's significance within the GCC context before outlining its objectives. Finally, the chapter overviews the thesis structure and defines key terminology.

Chapter 2 introduces the theoretical background on firm performance and DM and then addresses the first research objective by reviewing the empirical literature on DM techniques in predicting firm performance using bibliometric analysis and SLR. These reviews were deployed to identify research gaps, justify the study's focus on GCC firms, develop the study's conceptual framework, and establish a solid foundation for the subsequent chapters of the thesis.

Chapter 3 provides an overview of the research population, focusing on GCC publicly listed companies. It explores GCC stock markets and economic perspectives to establish the study's contextual foundation.

Chapter 4 outlines the research methodology for developing the firm performance prediction model. It begins with the research paradigm and methods and then proposes the DM prediction framework. Then, it explains data preprocessing, data sampling, and splitting. Later, it presents the

construction of four DM prediction models using seven DM techniques: single and ensemble, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost. Finally, it explains the deployed evaluation metrics and variable importance methods.

Chapter 5 presents the analysis results of the GCC dataset before and after data preprocessing for four predictive models. It addresses research objective two and describes the constructed models based on data preprocessing, sampling, and splitting methods to address research objective four.

Chapter 6 addresses research objectives three to five by presenting, analyzing, and comparing the prediction accuracy of four modeling approaches and defining the most influential variables on profitability.

Finally, Chapter 7 concludes the thesis by summarizing key findings from theoretical and managerial perspectives, outlining the research limitations, and offering recommendations for future research.

This structured approach ensures a systematic and coherent flow from the theoretical foundation to analysis, findings, and practical implications.

1.7 Research Terminology

This research defines key terms of firm performance and data-driven analysis to ensure clarity. The definitions are derived from the relevant literature, providing a foundation for understanding the concepts and methodologies used.

Firm Performance: refers to the economic outcomes that result from the interaction of a firm's internal operations, strategic decisions, and external market conditions. It serves as a key indicator of a firm's success and sustainability.

Data Mining: is the process of extracting meaningful patterns, trends, and insights from large datasets. In the context of the thesis, it plays a crucial role in extracting valuable knowledge for informed decision-making in business and finance.

Big Data: encompasses vast and complex datasets that exceed the capabilities of traditional data processing methods. It is characterized by high volume, velocity, and variety, requiring advanced analytical techniques for effective management and utilization.

Machine Learning: is a branch of Artificial Intelligence (AI) that enables machines to learn from data and past experiences, allowing them to recognize patterns and make predictions with little human involvement.

1.8 Summary

The research aims to examine the effectiveness of DM in predicting the financial performance of publicly listed firms in the GCC. This study will contribute to the literature by examining the accuracy of DM in predicting the profitability of GCC firms, using both firm-specific and macroeconomic factors as independent variables. It will identify patterns among factors, develop predictive models, and evaluate performance prediction across the GCC region. Additionally, it will identify the most influential factors driving

the profitability of GCC firms. The next chapter provides the theoretical background on firms' performance and DM, as well as a comprehensive literature review that employs bibliometric analysis and SLR on DM-based firm performance prediction. It then identifies the literature gap and introduces a conceptual framework to guide the thesis.

CHAPTER 2: THEORETICAL BACKGROUND AND LITERATURE REVIEW

2.1 Introduction

This chapter presents the theoretical background and literature review that underpin the study.

The theoretical background explores firm performance definitions, related theories, and determinants that shape performance outcomes. Additionally, it delves into DM concepts, tasks, methods, and techniques.

The literature review focuses on the firm performance prediction literature that utilizes DM techniques, detailing the planning, execution, and reporting of systematic reviews following bibliometric analysis.

Lastly, the chapter outlines the research gap. It draws a conceptual framework guiding this research, clearly defining the research variables used later to build DM prediction models for predicting a firm's performance.

2.2 Firm's Performance

2.2.1 Concepts

Firm performance is a crucial factor that investors evaluate closely, as it serves as a key indicator of long-term economic growth. It also holds significant importance for various stakeholders, including creditors, employees, managers, and governments (Vieira et al., 2019).

Firm performance is a key concept in strategic management research and is often used as a dependent variable. Despite its widespread use in academic studies, no universally agreed-upon definition or measurement exists. Without a standard operational definition, interpretations vary, influenced by individual perspectives. As a result, definitions of firm performance can range from broad and abstract to more specific and clearly outlined (Tudose et al., 2022).

Combs et al. (2005) explore the multidimensional nature of organizational performance and its implications for strategic management research. They argue that organizational performance is not a single construct but consists of distinct dimensions such as accounting returns, stock market performance, and growth. Their study differentiates between operational performance relating to specific internal processes, and the organizational performance, which reflects overall firm success. Their findings suggest that using multiple, uncorrelated measures without considering dimensionality can lead to inconsistent research outcomes. Their study highlights the need for strategic management research to clearly define and measure firm performance to enhance theoretical development and managerial insights.

In the 1960s and 1970s, organizations started adopting new approaches to assess their performance. During this period, performance was defined as an organization's ability to utilize its environment to access and manage limited resources effectively (Yuchtman & Seashore, 1967).

Adam (1994) emphasized that organizational performance depends heavily on the quality of employee performance. He argued that to maintain high performance, employees must continuously acquire updated knowledge

and skills, enabling them to adapt to market changes and enhance overall organizational effectiveness.

According to Lebas (1995), performance management is defined as a “philosophy that is supported by performance measurement”.

Combs et al. (2005) defined organizational performance as “economic outcomes resulting from the interaction of organizational characteristics, behavior, and environment”.

Harrison & Wicks (2013) defined firm performance consistent with Freeman’s (1984) fundamental idea that a firm should serve multiple stakeholders, as “the total value created by the firm through its activities, which is the sum of the utility created for each of a firm’s legitimate stakeholders”.

Defining firm performance is complicated because it involves finance, management, accounting, microeconomics, corporate governance, and other disciplines (Tudose et al., 2022).

2.2.2 Determinants of Firm’s Performance

"Measurement is tough, demanding, challenging, vital, exploited, and overused," yet, as Lord Kelvin famously stated, "If you cannot measure it, it does not exist" (Lebas, 1995).

Lebas (1995) argued that measurement allows people to simplify complex realities into numerical concepts, making communication and action

easier. This simplification helps identify key factors essential for successful management.

Measuring performance provides valuable information that enables management to monitor progress, report on achievements, enhance motivation and communication, and identify potential issues (Waggoner et al., 1999).

According to Gimbert et al. (2010), performance measurement systems are concise sets of (financial and/or non-financial) metrics that support the decision-making processes of an organization by gathering, processing, and analyzing quantified data of performance information.

The DuPont model is one of the original performance measurement systems developed by DuPont, the American chemicals giant. It is based on cost accounting theories and practices. DuPont created a framework that links accounting measures and financial ratios such as Return on Net Assets (RONA), ROI, and Return on Equity (ROE) to operational indicators and metrics. ROI is specifically used to assess the efficiency of an investment by comparing the profit generated to the investment cost. One of the main advantages of the DuPont model is its structured approach, which integrates financial measures with operational indicators. However, its primary drawback is that it focuses mainly on financial metrics, leading to criticism for being myopic and oriented toward short-term results (Bititci, 2015).

Profitability analysis and assessment of the fundamental drivers of profitability are critical components of evaluating financial performance (Melvin et al., 2004). This study considers profitability as the dependent

variable, while liquidity, leverage, solvency, asset utilization, and market value are independent variables.

2.2.2.1 *Dependent Variable*

Firm financial performance is the dependent variable in this thesis, measured by firm profitability.

Profitability is a crucial indicator of a firm's financial performance, reflecting a firm's ability to utilize resources and maintain a competitive edge efficiently. It is the primary goal for ensuring sustainability and stability. Enhancing profitability depends on effectively managing financial resources (Chiadamrong & Wattanawarangkoon, 2023).

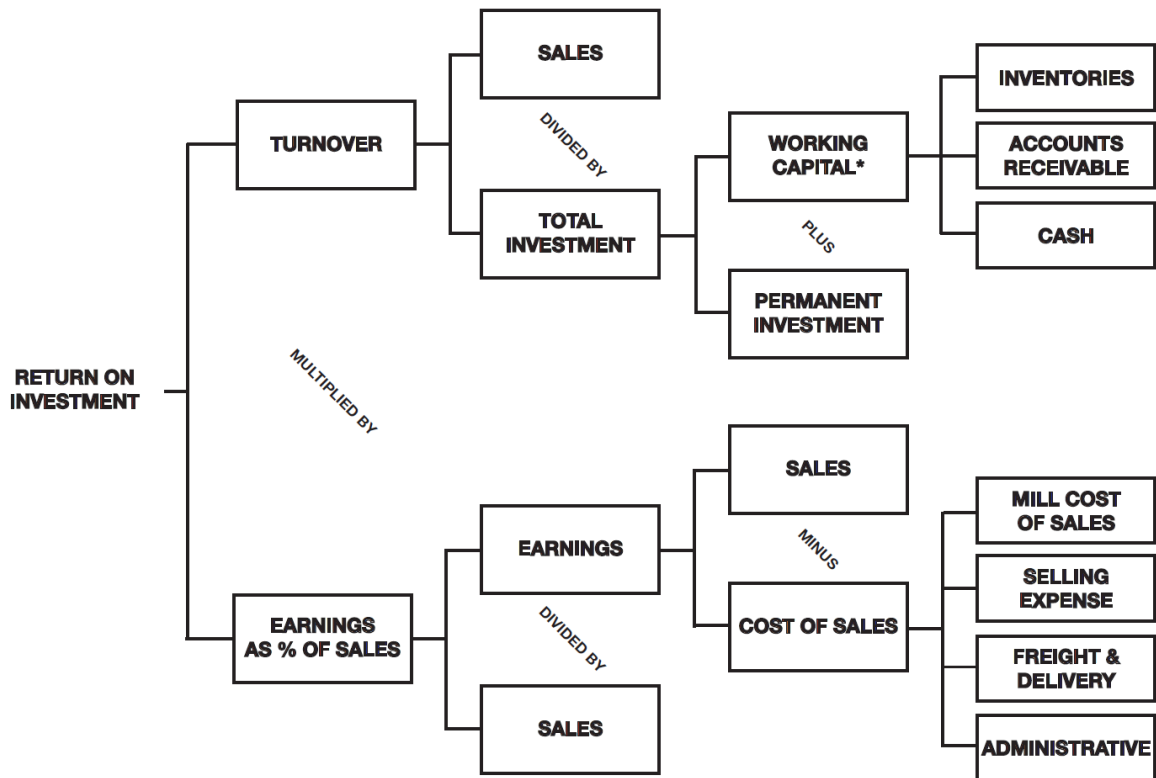
Pattitoni et al. (2014) highlight firm profitability as a key performance indicator, driven mainly by micro-level factors rather than macroeconomic conditions. While economic research examines industry-level influences, managerial and entrepreneurial perspectives in accounting and finance focus on firm-level analysis, emphasizing how strategic decisions directly affect profitability.

Standard metrics used to assess profitability include ROE, Return on Assets (ROA), ROI, Return on Sales (ROS), and Tobin's Q (Alareeni & Hamdan, 2020; Jain & Jain, 2020; Kasasbeh, 2021). Profitability measures organizational efficiency in generating earnings by utilizing various production factors (Delen et al., 2013). Failed firms exhibit negative profitability and efficiency ratios (Nguyen et al., 2023).

Ayoush et al. (2021) used statistical tests to examine the determinants of profitability for listed industrial firms on the Amman Stock Exchange. The determinants examined were liquidity, leverage, and solvency. They found that leverage has the highest relative impact among independent variables on profitability measured by ROA and ROE, followed by solvency and liquidity. They deployed firm size as the control variable.

Additionally, Garcia (2022) found that firms facing bankruptcy typically had lower profitability, particularly in terms of ROA. A higher ROI translates into greater profitability, improving a firm's financial performance. Firms that consistently generate an ROI below their cost of capital are at greater risk of financial distress, whereas firms with higher ROI tend to be in stronger financial health (Balasubramanian et al., 2019).

In 1914, Donaldson Brown introduced the classical ROI ratio, now widely recognized as the DuPont formula. That year, he was tasked with evaluating the performance of multiple operating departments at DuPont. To achieve this, he developed a procedure based on a simple mathematical equation: $R = T \times P$, where R represented the rate of return on invested capital, a key indicator of industrial efficiency and managerial effectiveness. T denoted the turnover rate of invested capital, while P signified the profit (earnings) margin on sales. Later, Brown extended his ROI framework to General Motors, emphasizing that a sustainable return on investment should align with long-term corporate growth. He viewed ROI as a fundamental measure of corporate productivity, reinforcing the idea that businesses exist to serve their owners' interests (Flesher & Previts, 2013). His complete formula was later formalized in a 1950 American Management Association publication, as illustrated in Fig. 2.1.



* Also includes small amounts of deferred charges which are not charted

Fig. 2.1: Relationship of factors affecting ROI (Flesher & Previts, 2013)

Since ROI is a management accounting metric, there are countless versions in the theoretical literature and practice. It covers several areas of applications. It is calculated by dividing the net profit of an investment by the cost of the investment. The outcome is expressed as a percentage or ratio (Balasubramanian et al., 2019; Valentowitsch et al., 2024):

$$\text{ROI} = \frac{\text{Net Profit}}{\text{Total Investments}} * 100$$

Net profit, also referred to as "net income" or "net earnings", represents the income after accounting for all operating and non-operating revenues and expenses, reserves, and income taxes. It is calculated before equity in earnings, minority interest, extraordinary items, after-tax adjustments,

discontinued operations, and preferred dividends. Total investments encompass the full value of trading assets, terms, property, loans, securities, private equity, and other investments, all evaluated using a non-differentiated approach (LSEG, 2023).

ROI is used to evaluate an investment's performance and compare the profitability of different investments. A positive ROI means an investment is profitable, while a negative ROI means an unprofitable investment. ROI is, therefore, a valuable tool for evaluating the cost-effectiveness of investments and drawing conclusions about firm performance (Valentowitsch et al., 2024).

ROI's simplicity and adaptability make it valuable for assessing both short-term performance and long-term strategic efficiency, which is crucial in competitive stock markets.

2.2.2.2 *Independent Variables*

This thesis considers the determinants of firm performance as independent variables. Five financial dimensions influencing firm profitability are examined: liquidity, leverage, solvency, asset utilization, and market value.

Liquidity reflects a firm's ability to generate cash flows from operations to meet short-term financial obligations (Hamann & Schiemann, 2021). It also represents the firm's capacity to cover short-term operating needs (Robinson et al., 2020). Similarly, Delen et al. (2013) defined liquidity as a firm's capability to meet short-term debts.

Empirical studies highlight the relationship between liquidity and firm performance. Tudose et al. (2022) found that the Current Ratio (CuR)

positively influences Return on Assets (ROA), suggesting that higher liquidity enhances profitability. Son et al. (2019) tested the predictive power of the Quick Ratio (QR) among other financial ratios in forecasting profitability, as measured by ROA. Additionally, Jabeur et al. (2021) examined liquidity ratios, such as CuR and Cash Ratio (CaR), in corporate failure prediction, concluding that non-distressed firms exhibit higher average liquidity while failed firms struggle to meet debt obligations.

Furthermore, Halteh et al. (2018) identified liquidity, measured by the "Working Capital/Total Assets" ratio, as one of the most predictive variables for financial distress in Islamic banks using the Altman Z-Score model. These findings emphasize the critical role of liquidity in maintaining financial stability and profitability.

Leverage represents the extent to which a firm relies on debt in its capital structure (H. Son et al., 2019). While leverage can amplify shareholder returns during profitable periods, it also increases financial risk when profits fall below borrowing costs (Balasubramanian et al., 2019).

Empirical studies provide mixed insights into the impact of leverage on firm performance. P. V. H. Son & Duong (2024) examined the Debt Ratio (DR), calculated as Total Debt/Total Assets, along with other financial ratios to predict profitability measured by Return on Assets (ROA). Their findings indicate that DR is the second most influential variable in profitability prediction. Similarly, Halteh (2018) tested Total Debt/Total Equity (TDPTE) based on Altman's (1968) Z-Score model and found it to be the second most important predictor of financial distress when using Stochastic Gradient Boosting.

Conversely, some studies highlight the risks associated with high leverage. Garcia (2022) concluded that highly leveraged firms are more prone to bankruptcy. Barboza et al. (2017) analyzed leverage using the ratio of $\text{Market Value of Shares} \times \text{Number of Shares} / \text{Total Debt}$, based on Altman's (1968) study, but found that leverage did not correlate significantly with other financial variables in bankruptcy prediction.

Furthermore, Chiadamrong & Wattanawarangkoon (2023) found that increasing leverage had no impact on firm performance, whereas reducing leverage levels negatively affected high, and medium-performing firms. These findings underscore the complex role of leverage in financial stability and performance, which varies based on firm-specific and industry factors.

Solvency reflects a firm's ability to meet long-term obligations and maintain financial stability. Solvency ratios fall into two main categories: debt ratios, which assess the proportion of debt relative to equity, and coverage ratios, which measure a firm's capacity to cover debt payments using earnings. These ratios are essential for evaluating the creditworthiness of a firm's bonds and other debt obligations (Robinson et al., 2020).

One critical coverage ratio is the Interest Coverage Ratio (ICR), which indicates how many times a firm's Earnings Before Interest and Tax (EBIT) can cover its interest payments. Often referred to as "times interest earned", an ICR above 2.5 suggests strong solvency, signifying lower default risk and better financial health (Ayoush et al., 2021; Balasubramanian et al., 2019).

Empirical research underscores the importance of solvency in financial performance and distress prediction. Zhou et al. (2022) tested eight solvency

ratios, including Interest Coverage (IC), to classify listed manufacturing firms into five financial performance levels. Arora & Saurabh (2022) used decision tree models incorporating Altman's variables ranking solvency as the fourth most significant predictor of bankruptcy in Indian firms. Additionally, Jabeur et al. (2021) identified solvency as the most critical factor in corporate failure prediction, reporting a positive correlation between solvency ratios and default probability. These findings highlight the essential role of solvency in assessing a firm's financial resilience and its ability to withstand economic downturns.

Asset utilization, also known as activity or operating efficiency, measures how efficiently the firm utilizes assets. Asset Utilization ratios reflect the efficient management of working capital and longer-term assets. Therefore, efficiency directly impacts liquidity (the ability of a company to meet its short-term obligations), therefore, some activity ratios are also proper in assessing liquidity (Robinson et al., 2020).

Chiadamrong & Wattanawarangkoon (2023) found that while an increase in asset utilization may not necessarily enhance firm performance, a decline in asset utilization can negatively impact medium- to high-performing firms. Barboza et al. (2017) analyzed asset utilization using the Asset Turnover (AT) ratio, defined as Sales/Total Assets, based on Altman's (1968) framework. Their study indicated that asset utilization had limited correlation with financial variables related to bankruptcy, unlike liquidity, profitability, and productivity ratios.

Several studies have employed activity ratios in financial failure prediction. Kim et al. (2018) used eight activity ratios to predict corporate

bankruptcy in Korean manufacturing firms. Wang & Wu (2017) focused on four ratios: inventory turnover, current assets turnover (CAT), total assets turnover, and accounts receivable turnover to assess business failure in Chinese listed manufacturing firms. Similarly, Aljawazneh et al. (2021), du Jardin (2021), and Park et al. (2021) incorporated various asset utilization ratios in financial failure models. Zhou et al. (2022) observed that firms with low CAT ratios often exhibited profitability concerns, with a significant proportion reporting negative net profits.

Nguyen et al. (2023) highlighted that firms with high asset turnover were more likely to face bankruptcy. The AT ratio is a significant component of the Dupont analysis, where higher turnover is generally linked to higher profitability (Robinson et al., 2020). Altman (1968) included this ratio among five financial indicators in his bankruptcy prediction model. However, industry-specific comparisons are crucial, as firms in sectors like wholesale trade tend to have high turnover but low margins, whereas luxury goods firms exhibit the opposite trend. Additionally, failed firms in Nguyen et al.'s (2023) sample had higher mean and median turnover ratios than non-failed firms.

Market value, or valuation, represents the quantity of an asset or flow (e.g., earnings) associated with ownership of a specified claim (e.g., a share or ownership of the enterprise) (Robinson et al., 2020). It measures the value of a firm's potential or existing investments (Fama & French, 1992). Garcia (2022) found that higher market capitalization reduces the likelihood of bankruptcy, indicating its role in enhancing financial stability. Given that capitalization is a market value metric, Garcia employed it as a proxy for corporate size. Barboza et al. (2017) examined the predictive power of market value on firm bankruptcy, utilizing the Change in Price-to-Book (CPB) ratio.

Firm performance is also shaped by broader national and regional factors, highlighting the interplay of macroeconomic conditions (Farooq et al., 2023). Macroeconomic drivers such as Gross Domestic Product (GDP) and inflation influence firm performance, impacting credit ratings and bankruptcy risk (Jiang & Jones, 2018).

GDP measures the monetary value of final goods and services produced within a country's borders over a specific period, reflecting the total output available for final consumption. It includes market-based production and nonmarket contributions, such as government-provided education and defense services. In contrast, Gross National Product (GNP) accounts for the total output generated by a country's residents, regardless of location. Since GDP is reported in nominal terms, direct comparisons across different periods require adjustments for inflation. Real GDP is obtained by adjusting nominal GDP for price changes using the GDP deflator, allowing for a more accurate assessment of economic growth (Callen, 2020).

GDP serves as an indicator of economic conditions and influences corporate investment decisions. Firms are more likely to expand investments in profitable ventures during periods of economic growth to maximize returns (Farooq et al., 2023). Additionally, Keynesian consumption theory suggests that rising GDP increases household income, driving consumer demand and stimulating further business investments. Jiang & Jones (2018) further argue that strong GDP growth reduces the likelihood of firm distress.

Inflation, another crucial macroeconomic variable, is commonly measured using the Consumer Price Index (CPI), which tracks the cost of a representative basket of essential goods and services compared to a base year.

Governments assess CPI through household surveys, with its percentage change representing consumer price inflation, which is the most widely used measure of inflation (Oner, 2012). J. P. Li et al. (2020) found that macroeconomic factors, including inflation, play a significant role in determining investment-grade ratings. Additionally, higher inflation often signals weaker economic conditions, increasing the risk of firm distress (Jiang & Jones, 2018).

These macroeconomic indicators provide essential insights into the economic environment, influencing firm performance and financial decision-making.

2.2.2.3 *Control Variables*

Firm size is commonly measured by total assets (Ainan et al., 2024; Aljawazneh et al., 2021; Chiadamrong & Wattanawarangkoon, 2023; Elamir, 2020; Halteh et al., 2018; Park et al., 2021; P. V. H. Son & Duong, 2024; Saha et al., 2023; Son et al., 2019) and has been shown to impact firm performance significantly. Alternatively, some studies classify firm size into categorical groups, small, medium, and large (e.g., Aljawazneh et al., 2021; H. Son et al., 2019). Barboza et al. (2017) enhanced the accuracy of prediction models by adding complementary financial indicators, including firm size (growth of assets), which proved crucial in short-term financial performance. Garcia (2022) further demonstrated that smaller firms, as measured by market capitalization, face a higher risk of bankruptcy, while larger market capitalization improves the Altman Z-score, reflecting a reduced bankruptcy risk. Similarly, Chiadamrong & Wattanawarangkoon (2023) found that larger

firms typically perform better, especially those seeking to recover from underperformance.

Firm listing age, which is measured from the Initial Public Offering (IPO), also influences firm performance. The longer a firm is listed, the more likely it is to have built a reputation, reducing the tendency toward risky behaviors associated with financial distress (Balasubramanian et al., 2019). Younger firms may face greater financial instability, whereas mature firms with established reputations experience greater financial stability.

Another important non-financial factor is industry type, which categorizes firms according to the economic sector in which they operate. Garcia (2022) found that a firm's industry, using the Bloomberg Industry Classification, can affect its bankruptcy risk; therefore, he included it as an additional predictor variable.

In this study, the Refinitiv Business Classification (TRBC) system is deployed since it is used by the LSEG Data & Analytics, which classifies firms into 13 economic sectors: Energy, Basic Materials, Industrials, Consumer Cyclical, Consumer Non-Cyclical, Healthcare, Financials, Technology, Telecommunications, Utilities, Real Estate, Government Activity, and Academic and Educational Services (Refinitiv, 2020). Each sector has distinct dynamics that influence firm performance, making industry classification a crucial determinant.

Furthermore, a firm's headquarters influence its performance, as economic conditions differ across GCC countries. Chapter three will explore GCC stock markets and economics.

The following section presents theories related to firm performance.

2.2.3 Theoretical Perspectives

Various theoretical frameworks, including stakeholder theory, Resource-Based View (RBV), and macroeconomic theory, shape firm performance in the capital markets.

Stakeholder theory provides a framework for evaluating firm performance by considering the interests of all stakeholders. Rather than focusing solely on maximizing shareholder wealth, this theory suggests that businesses should account for the interests of employees, customers, suppliers, and the broader community (Jensen, 2010). R. Edward Freeman introduced this approach in 1984 with the publication of *Strategic Management: A Stakeholder Approach* (Freeman & McVea, 2005). The motivation behind stakeholder management was to address the concerns of managers facing significant environmental instability and change. Harrison and Wicks (2013) argued that assessing stakeholders' well-being should consider multiple dimensions for each primary stakeholder, providing a valuable framework for firms to evaluate the value delivered to stakeholders.

The RBV theory, introduced by Barney (1991), posits that internal firm resources are the primary drivers of competitive advantage and profitability. RBV suggests that resources must be valuable, rare, inimitable, and non-substitutable to provide a sustainable competitive advantage. Unlike external-facing frameworks such as Porter's Five Forces, RBV focuses on internal capabilities and resources as the key to achieving superior performance. This inward perspective is fundamental to strategic management, emphasizing how

unique internal resources shape a firm's strategy and drive its success. Firms that effectively leverage their resources, such as knowledge, human capital, and technology, can achieve sustained competitive advantage in dynamic markets (Barney, 1991).

Macroeconomic theory, particularly the Dynamic General Equilibrium (DGE) approach by Wickens (2008), provides a framework for understanding how external economic variables impact firm performance. This theory posits that the economy functions as a general equilibrium system, where individuals and firms make decisions to optimize their welfare over time. Factors such as GDP growth, inflation, interest rates, and employment levels have a significant influence on a firm's profitability. For instance, GDP growth drives consumer demand, boosting revenue for firms, while inflation can raise input costs and squeeze profit margins unless firms pass those costs to consumers. Interest rates also affect borrowing costs; high rates typically reduce investment, whereas lower rates encourage changes in employment growth, and influence labor costs and consumer demand. Additionally, macroeconomic variables have a significant impact on the stock market. GDP growth and interest rates are key drivers of stock prices, affecting firms' profitability and future cash flow. Strong GDP growth can enhance investor confidence and increase stock prices, while high inflation or rising interest rates may depress them. Wickens (2008) notes that stock markets are sensitive to macroeconomic fluctuations, leading to volatility as investors adjust their portfolios in response to new information. It emphasizes the need for firms to align their strategies with the economy's equilibrium, ensuring investment decisions are integrated with broader patterns of economic growth.

In this study, only the most relevant theoretical frameworks were considered. Institutional Theory, while significant in understanding how regulatory, cultural, and normative environments shape organizational behavior, was excluded as it falls outside the scope of this research. This study focuses on quantitative financial data and predictive modeling, whereas Institutional Theory emphasizes how organizations adapt to evolving social environments, including changes in laws, norms, and professional standards (Brower & Dacin, 2020).

In empirical studies, Hamann and Schiemann (2021) emphasized the importance of focusing theoretical arguments and hypotheses on specific dimensions of organizational performance, such as growth, liquidity, stock market performance, and profitability. They argue that applying appropriate indicators for these dimensions in empirical tests can provide clearer insights. Without such a focus, empirical results may become unclear, limiting the advancement of knowledge in strategic management, particularly in understanding the factors contributing to organizational success variations.

The following section overviews DM concepts, tasks, methods, and techniques. Understanding and predicting financial conditions is vital for stakeholders like investors, creditors, and management to evaluate firm success and anticipate challenges (Hantono, 2019). Financial reports such as the balance sheet, income, and cash flow statement provide data on assets, earnings, and debts (Brigham & Ehrhardt, 2014). Traditional statistical methods struggle to handle the complexity of firms complicated BD's volume, velocity, and variety, prompting the rise of DM (H. Son et al., 2019), driving the adoption of DM as an effective tool for BD analysis, also known as data analytics and predictive analytics.

2.3 Data Mining

2.3.1 Concepts

DM is a crucial analytical approach for extracting meaningful patterns and insights from large and complex datasets. The term "big data" was first introduced by John Mashey (Mashey, 1999) to describe the handling and analysis of massive datasets. The concept gained prominence in the early 2000s when Doug Laney defined big data using the "3Vs": volume, velocity, and variety (Laney, 2001). According to this definition, "big data" refers to a rapidly increasing amount of data that is difficult, if not impossible, to manage using traditional methods. Today, BD is often defined by the "5Vs", which include veracity and value in addition to the original "3Vs" (L. Zhou et al., 2017). Below is a breakdown of the "5Vs":

1. Volume: Relates to the quantity and amount of data.
2. Velocity: Relates to the speed at which data is generated.
3. Variety: Relates to the heterogeneity of the data that describes the type, nature, and format of the data.
4. Veracity: Relates to the quality and trustworthiness of the captured data.
5. Value: Relates to the insights and impact derived from the data.

BD originates from distributed, large-volume, and heterogeneous sources, posing challenges for extracting meaningful insights (Wu et al., 2014). Effective BD analysis ensures timely and high-quality results, offering optimal value to firms (Erl et al., 2016).

In the era of BD, where organizations generate and collect massive amounts of information, the need for advanced analytical tools has become

more pressing (L. Zhou et al., 2017). DM represents a cornerstone in this evolution, offering effective methods to extract valuable insights from vast datasets (Ledhem, 2022). In the context of firm performance, DM methods effectively handle the large-scale data generated from financial reports and market data (Khalil et al., 2022; Saha et al., 2023; Son & Duong, 2024). This section delves into the concept of DM and explores its tasks, methods, and techniques.

According to (Sharda et al., 2021), most literature defines DM as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases”.

DM involves extracting meaningful patterns and insights from large datasets using techniques from statistics, ML, and data science (Sharda et al., 2021). It forms a key phase in the Knowledge Discovery in Databases (KDD) process, uncovering trends for informed decision-making (Sarkar et al., 2018). DM, often automated, is integral to predictive analytics and business intelligence, enabling organizations to identify trends and make data-driven decisions (Erl et al., 2016).

2.3.2 Tasks and Methods

DM utilizes existing and relevant data to build models that uncover patterns among a dataset's attributes. Generally, DM aims to identify three major types of patterns. DM tasks are generally categorized into three main types: prediction, association, and clustering (Sharda et al., 2021).

1. **Prediction:** Estimating the likelihood of future occurrences based on historical trends and data.

2. **Association:** Identifying frequently co-occurring groupings of items or events.
3. **Segmentation/ Clustering:** Detecting natural groupings of items or entities based on shared characteristics.

Learning techniques and algorithms drive the extraction of patterns from historical data. They are broadly classified into two learning types: supervised and unsupervised. Supervised learning relies on labeled data to train models for specific outcomes, while unsupervised learning identifies patterns or structures in unlabeled data.

Several methods for conducting DM prediction studies are available, including classification, regression, and time series. Most DM software tools utilize multiple techniques or algorithms for each method.

Depending on the nature of the dependent variable, prediction can be classified as either classification or regression. In classification, the outcome is a categorical label, such as "fail" or "non-fail," whereas regression predicts a continuous numerical value. Classification is one of the most commonly used DM methods for solving real-world problems (Sharda et al., 2021).

This study uses the supervised regression prediction method, highlighted by the bold line in Fig. 2.2, to address the research objectives. The model is a form of supervised learning where several independent variables are trained to predict the dependent variable, ROI. Regression is employed to predict the continuous ROI value, aiming to identify the most significant factors influencing firm performance. This method provides a robust basis to evaluate the relationship between the independent and dependent variables.

Supervised prediction is commonly used in firm performance prediction due to its interpretability and ability to capture complex data patterns.

Data Mining Tasks and Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naive Bayes, GA	Supervised
Regression	Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time series	Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA	Supervised
Association		
Market-Basket	Apriori, OneR, ZeroR, Eclat, GA	Unsupervised
Link Analysis	Expectation Maximization, Apriori Algorithm, Graph-Based Matching	Unsupervised
Sequence Analysis	Apriori Algorithm, FP-Growth, Graph-Based Matching	Unsupervised
Segmentation		
Clustering	K-means, Expectation Maximization (EM)	Unsupervised
Outlier Analysis	K-means, Expectation Maximization (EM)	Unsupervised

Fig. 2.2: Data Mining Taxonomy Tasks, Methods, and Algorithms adopted from (Sharda et al., 2021)

Two primary methods dominate the field of firm performance prediction based on dependent variable value: classification models and

regression models. Classification models are widely used for bankruptcy prediction, categorizing firms based on financial risk, while regression models focus on predicting continuous outcomes such as profitability and financial performance. Both approaches leverage financial ratios and other indicators, offering valuable insights into firm health and aiding industry decision-making (Kumbure et al., 2022). In Bankruptcy Prediction and Financial Distress Prediction, classification models are commonly used to categorize companies into predefined groups based on financial health. Most studies focus on binary classification, where firms are divided into categories such as "healthy" vs. "unhealthy," "non-bankrupt" vs. "bankrupt," or "financially sound" vs. "financially distressed." The specific definitions of these risk classes often vary across studies; some adopt legal definitions of bankruptcy, while others focus on broader financial distress criteria. In some cases, hybrid models combine bankruptcy and financial distress elements to provide a more comprehensive risk assessment. These classification models are critical in early detection, allowing stakeholders to proactively mitigate financial risk (Zhao et al., 2024).

2.3.3 Prediction Techniques of a Firm's Performance

This section provides a brief introduction to the DM techniques used to develop prediction models for firm performance. Furthermore, Section 4.8.2 provides a detailed explanation of the techniques employed in this study.

DM techniques are broadly categorized into two types: single and ensemble methods. Single or standalone techniques are individual algorithms that rely on their internal structures for predictions. While they are simpler

and faster to train, they may lack the robustness of ensemble methods (Sarkar et al., 2018; Sharda et al., 2021).

In contrast, ensemble techniques improve accuracy and reduce overfitting by combining predictions from multiple models. Ensemble learning methods, including bagging, boosting, and stacking, are widely used ML techniques that consistently deliver higher predictive accuracy compared to individual, standalone models. These methods involve training multiple ML models (i.e., weak or base learners, such as decision trees) to address the same predictive task, and then combining their outputs to produce a more accurate result (Khalil et al., 2022).

In both Bagging and Boosting algorithms, the same type of base learner is used; however, they differ in the way these learners are trained.

Bagging, short for Bootstrap Aggregation, involves creating multiple datasets through random sampling with replacement from the original dataset. Bagging was first introduced by Leo Breiman in 1996. The method involves creating multiple decision trees from resampled data and combining predictions through averaging or voting (Breiman, 2001).

Random Forest is a popular example of a bagging method (Sarkar et al., 2018). These sampled datasets are used to train several base learners in parallel, and their outputs are averaged to produce the final prediction. Bagging effectively reduces variance and mitigates overfitting, thereby enhancing the predictive capability of weaker models such as regressors or classifiers. Random forest classifiers and regressors outperform traditional

and other ensemble models in predicting firm performance (Arora & Saurabh, 2022; Barboza et al., 2017; Garcia, 2022; J. P. Li et al., 2020; Ledhem, 2022).

Boosting, on the other hand, is a sequential method in which base learners are trained one after another, with each learner focusing more on the instances misclassified by its predecessors. Boosting is widely regarded as the second most popular ensemble technique, following bagging. Yoav Freund and Robert E. Schapire were the pioneers behind the boosting algorithm, with initial separate publications in the early 1990s, followed by a collaborative paper in 1996. They are credited with developing the well-known Boost (Freund & Schapire, 1997).

Well-known boosting algorithms include AdaBoost, Gradient Boosting, XGBoost, and CatBoost, the latter being specifically optimized for handling categorical features efficiently. This adaptive approach reduces bias and contributes to building more robust predictive models (Sarkar et al., 2018).

Boosting models, both classifiers and regressors, have consistently outperformed traditional and other ensemble models in predicting firm performance, as demonstrated in several studies (Ainan et al., 2024; Nguyen et al., 2023; Khalil et al., 2022; Arora & Saurabh, 2022; Jabeur et al., 2021; du Jardin, 2021; Son et al., 2019).

In recent years, there has been a notable shift in focus towards enhancing the accuracy and reliability of financial performance predictions, driven by the complexities of the modern economic landscape. As businesses navigate unprecedented challenges and opportunities, understanding the key

indicators of financial health has become increasingly vital. Researchers and practitioners are dedicating significant attention to the methodologies employed to predict various aspects of financial performance.

2.4 Previous Studies on Firms' Performance Prediction Using DM Techniques

In the rapidly evolving domain of DM techniques for predicting firm performance, a thorough understanding of the existing literature is essential for identifying research gaps and shaping future studies. This section utilizes two methodological review approaches: bibliometric analysis and SLR. Together, these approaches offer a robust framework for mapping the current state of research, providing valuable insights to guide the objectives of this thesis.

The bibliometric analysis quantitatively evaluates research trends, focusing on publication frequency, influential authors, citation patterns, and collaboration networks. Complementing this, the SLR qualitatively identifies empirical studies aligned with the thesis objectives, enabling a detailed assessment of existing research.

This combined bibliometric and SLR approach lays the groundwork for a comprehensive review, effectively identifying gaps, guiding the research towards its objectives, and highlighting avenues for future research in firm performance prediction. Table 2.1 outlines and compares the characteristics of the review methods utilized in this thesis.

Table (2.1): Comparison of Deployed Review Methods (Donthu et al., 2021)

Review type	Goal	Scope	Dataset	Analysis
Bibliometric Analysis	Summarizes bibliometric large-volume data to show the intellectual structure and trends in a research field.	Broad	Large	• Quantitative (evaluation and interpretation) • Qualitative (interpretation only)
Systematic Literature Review	Summarizes and integrates the findings of existing literature on a research field.	Specific	Small	• Qualitative (evaluation and interpretation)

2.4.1 Bibliometric Analysis

The thesis adopted the bibliometric analysis procedure proposed by Donthu et al. (2021) that uses four steps for bibliometric reviews: (1) defining the review aims and scope; (2) choosing the analysis techniques; (3) collecting the data required for analysis; and (4) conducting the analysis and reporting the findings, as shown in Fig. 2.3.

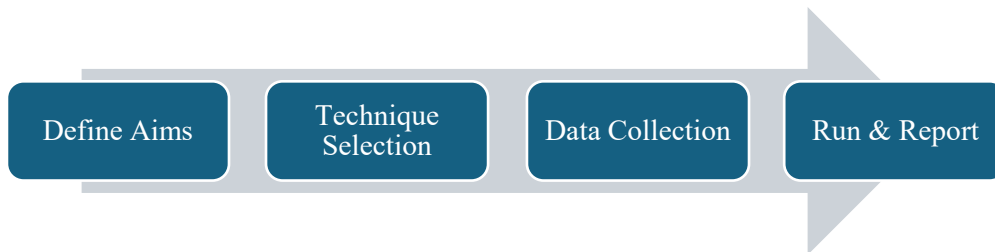


Fig. 2.3: Bibliometric analysis steps (Donthu et al., 2021)

Bibliometric analysis plays a crucial role in this process, allowing the researcher to pinpoint the most relevant and influential studies in the field. Bibliometric analysis can assess a vast amount of literature and provide a thorough summary of a particular field.

This section of the thesis aims to conclude the intellectual formation of DM techniques in firms' performance prediction research by examining the bibliometric structure, which encapsulates the publication trends of articles and countries. The scope of review is sufficiently large as DM transformative techniques are emerging widely in stock market research and applications to predict performance.

The bibliometric analysis maps the research landscape on DM techniques for predicting firm performance and identifies emerging trends and research gaps to guide the empirical objectives of this thesis. To achieve this, the following bibliometric goals were established:

1. Identify the publication (productivity) trend in firms' performance prediction research using DM techniques to establish the time frame of the SLR by identifying periods of significant research activity and understanding how interest in this area has evolved.
2. Identify the most impactful (most cited) articles in firms' performance prediction research using DM techniques to recognize foundational and highly influential research contributions. This serves as a benchmark for understanding the development and impact of DM techniques in the field.
3. Identifying thematic clusters in firms' performance prediction research using DM techniques aids in classifying and mapping the main research themes within the literature, providing insights into key areas of focus and enabling a clearer understanding of various aspects of firm performance prediction.
4. Identifying trends in firms' performance prediction research using DM techniques helps reveal shifts in research focus, emerging topics, and

technological advancements, which are critical for understanding the future direction of research and identifying areas that require further exploration.

Two main techniques of bibliometric analysis will be utilized to address the purpose of the analysis: performance analysis and science mapping. Specifically, performance analysis will address the first and second purposes, while science mapping will be used for the third and fourth purposes, as detailed in Table 2.2. The performance analysis approach summarizes the contributions of the most productive researchers, whereas the science mapping approach examines the bibliometric and intellectual structures.

Table (2.2): Bibliometric Techniques Used to Map the Analysis

Bibliometric Purposes	Technique
Evolution of firm performance prediction	Productivity per active year of publication (Publication-related metric)
Impactful articles	Total citation (Citation-related metric)
Thematic clusters	Bibliographic coupling
Literature trends	Co-word analysis

The literature search focused on articles from the Web of Science (WoS) core collection between 2017 and 2024, supplemented by additional methods, including website searches and citation searches. The year 2017 was chosen because research on DM techniques for predicting firm performance gained notable attention, with 28 relevant articles published, as shown in Fig. 2.7 and discussed in Section 2.4.1.1 later. The search strategy involved specific keyword combinations applied to the “Titles, Abstract, and Author Keywords” fields, as shown in Table 2.3. Using practical filters, 457 articles were identified on April 14, 2024, from an initial search of 4,025 documents.

Table (2.3): Practical Criteria for Bibliometric Analysis

Filter	Description
Search fields	Title, Abstract & Author Keywords
search string	(Firm OR Corporat* OR Compan* OR Enterprise OR Business) AND (Financial OR Performance OR Efficiency OR Profit* OR Growth) AND (Predict* OR Forecast*) AND (“AI” OR “Data Mining” OR “Machine Learning” OR “Deep Learning” OR “Ensemble”)
Document Type	Article
Publishers	IEEE, Elsevier, Springer Nature, Emerald Group Publishing, Wiley, and Taylor & Francis
Language	English
Period	Published between 2017- 2024
WoS Categories	"COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE" OR "COMPUTER SCIENCE INFORMATION SYSTEMS" OR "MANAGEMENT" OR "ECONOMICS" OR "BUSINESS" OR "BUSINESS FINANCE"
Citation Topics	"Artificial Intelligence & Machine Learning" OR "Economics" OR "Management"
Meso	

Table 2.4 presents key details regarding the 457 articles identified in the WoS database following the application of predefined search filters. These publications, dated between 2017 and 2024 and distributed across 176 distinct sources, consist of review and empirical research articles centered on applying DM techniques for firm performance prediction. The authorship analysis reveals contributions from 1,277 researchers, including 29 single-authored works, underscoring the predominantly collaborative nature of scholarly activity in this domain.

Table (2.4): Primary Information About the Collected Articles

Description	Results	Description	Results
Timespan	2017:2024	<u>Authors Collaboration</u>	
Sources (Journals, Books, etc)	176	Single-authored docs	32
Documents	457	Co-Authors per Doc	3.15
Annual Growth Rate %	1.46	International co-authorships %	31.07
Document Average Age	2.77	<u>Document Types</u>	
Average citations per doc	16.38	article	420
References	20915	article; early access	32
<u>Document Contents</u>		article; proceedings paper	4
Keywords Plus (ID)	958	article; retracted publication	1
Author's Keywords (DE)	1701		
<u>Authors</u>			
Authors	1277		
Authors of single-authored docs	29		

Table 2.4 supports the first research objective by presenting a comprehensive overview of all relevant studies, both quantitative and qualitative, published between 2017 and 2024. The timespan was defined based on the trend illustrated in Fig. 2.4, which shows that scholarly interest in this topic began to rise notably in 2017, with 28 articles published that year.

Two recognized bibliometric tools were used to run the analysis: RStudio 2023.12.1+402 with the Bibliometrix package and VOSViewer 1.6.20, which analyzed bibliometric data sourced from WoS.

The reporting stage involves presenting the results of the bibliometric analysis, which will be elaborated on in the following subsections.

2.4.1.1 Evolution of Firm Performance Prediction Using DM

Financial analysis began as a crucial tool for assessing firm performance in the late 19th and early 20th centuries, with researchers developing key financial ratios to evaluate companies' financial health (Wilcox & Bourne, 2003). In the 1930s, Winakor & Smith analyzed companies facing financial difficulties and found that the net working capital to total assets ratio was the most reliable predictor of potential failure (Horrigan, 1968).

In 1966, Beaver analyzed financial ratios' ability to predict firms' failure between 1954 and 1964; he found that some ratios predict failure up to five years in advance (Beaver, 1966). Later, in 1968, Altman developed a multivariate discriminant analysis (MDA) model called the Altman Z-Score model (Altman, 1968). This model uses five financial measures to classify firms as "bankrupt" or "non-bankrupt". In 1980, Ohlson used the logit for

probabilistic bankruptcy prediction (Ohlson, 1980), and in 1984, Zmijewski used the linear probit for financial distress prediction (Zmijewski, 1984).

Despite their effectiveness, these statistical models faced limitations, such as assumptions of linear separability, independence of variables, and multivariate normality, which were often unrealistic when dealing with multiple financial ratios (Smiti & Soui, 2020). To overcome these limitations, researchers have turned to DM techniques for bankruptcy prediction, which enable the development of more flexible and nonlinear models.

Fig. 2.4 presents the annual production evolution of articles published on the field firms' performance prediction using DM techniques. Early articles used the Artificial Intelligence (AI) knowledge base to predict financial firms' performance conducted by Abramson & Finizza in 1991 to predict oil prices using belief networks (Abramson & Finizza, 1991), followed by Martin-del-Brio & Serrano-Cinca in 1993, who applied Self-organizing Neural Networks to analyze the financial state of Spanish companies and make an assessment of bankruptcy risk by clustering the bankrupt and solvent regions together (Martin-del-Brio & Serrano-Cinca, 1993).

The research on DM techniques in predicting firms' performance started to gain more attention in 2017, with 28 published articles, and gradually increased to 103 and 112 in 2022 and 2023, respectively. In 2024, 31 articles were produced, increasing the production rate. Since this research area began to gain significant attention in 2017, the time frame for the SLR studies has been defined accordingly, covering the period from 2017 to 2024.

It is essential to recognize that the evolution of ratio analysis in assessing firm performance has followed distinct trajectories based on the purposes of creditors and managers. Credit analysis primarily evaluates an entity's ability to meet its financial obligations, with a focus on bankruptcy prediction models. In contrast, managerial analysis is focused on profitability, utilizing performance prediction models to estimate potential earnings. Although both approaches employ similar components of financial ratios, their applications and objectives differ significantly. Credit analysis is crucial in informing decision-making applications (Zhao et al., 2024).

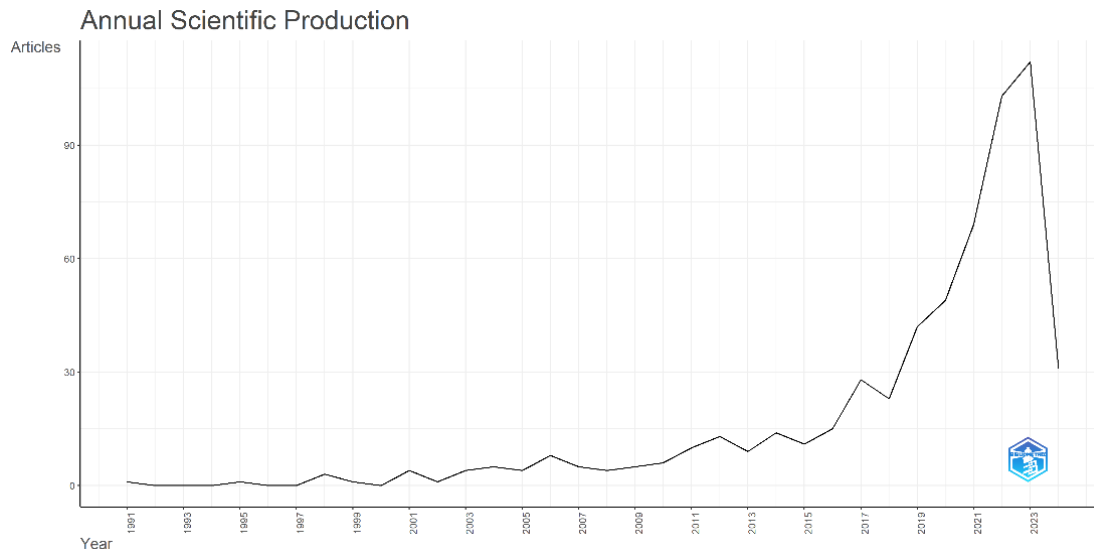


Fig. 2.4: Annual scientific production of DM research in the field of firm's performance prediction plotted using R-Bibliometrix (Aria & Cuccurullo, 2017)

2.4.1.2 Most Cited Articles in Literature

The top 5- globally cited publications of DM research in financial performance prediction are presented in Table 2.5. (Barboza et al., 2017) published the most impactful and influential article with the most citations in WoS (342), followed by (Lu et al., 2020) with 228 citations. (Barboza et al.,

2017) made a substantial improvement in bankruptcy prediction accuracy using single and ensemble methods of DM techniques for American and Canadian companies from 1985 to 2013, whereas (Lu et al., 2020) conducted a systematic review of the business literature to measure the impact of service robots on customers and employees.

Further assessment of the top-cited publications of DM research in financial performance prediction exposes those studies focused on using DM in fundamental and technical approaches to performance prediction (Barboza et al., 2017; Kraus & Feuerriegel, 2017) and customer-centric research (Lu et al., 2020; Steinhoff et al., 2019; Wamba-Taguimdje et al., 2020) are particularly impactful and influential.

Table (2.5): Top 5 Cited Publications in Performance Using DM Techniques

Article	PY	Title	TC
Barboza et al. (2017)	2017	Machine learning models and bankruptcy prediction	342
Lu et al. (2020)	2020	Service robots, customers, and service employees: what can we learn from the academic literature, and where are the gaps?	228
Kraus & Feuerriegel (2017)	2017	Decision support from financial disclosures with deep neural networks and transfer learning	175
Steinhoff et al. (2019)	2019	Online relationship marketing	173
Wamba-Taguimdje et al. (2020)	2020	Influence of AI on firm performance: The business value of AI-based transformation projects	164
Note(s): PY: Publication Year, TC: Total Citations			

2.4.1.3 Thematic Clusters of Literature

This review further examines the knowledge related to DM research in firms' performance prediction through bibliographic coupling. Bibliographic coupling relies on citing publications to explain the extant knowledge in the field (Kessler, 1963). It describes the connection between two articles that

refer to the same third work in their reference list. When two articles have more references in common, their bibliographic coupling becomes stronger. This process helps to group thematically similar documents into clusters.

Using VOSViewer software, 457 articles published between 2017 and 2024 were grouped under six clusters in the field of DM research related to firm performance prediction. Fig. 2.5 and Table 2.6 present the bibliographic coupling analysis of the documents with the full-counting method. Below is a summary of the clusters.

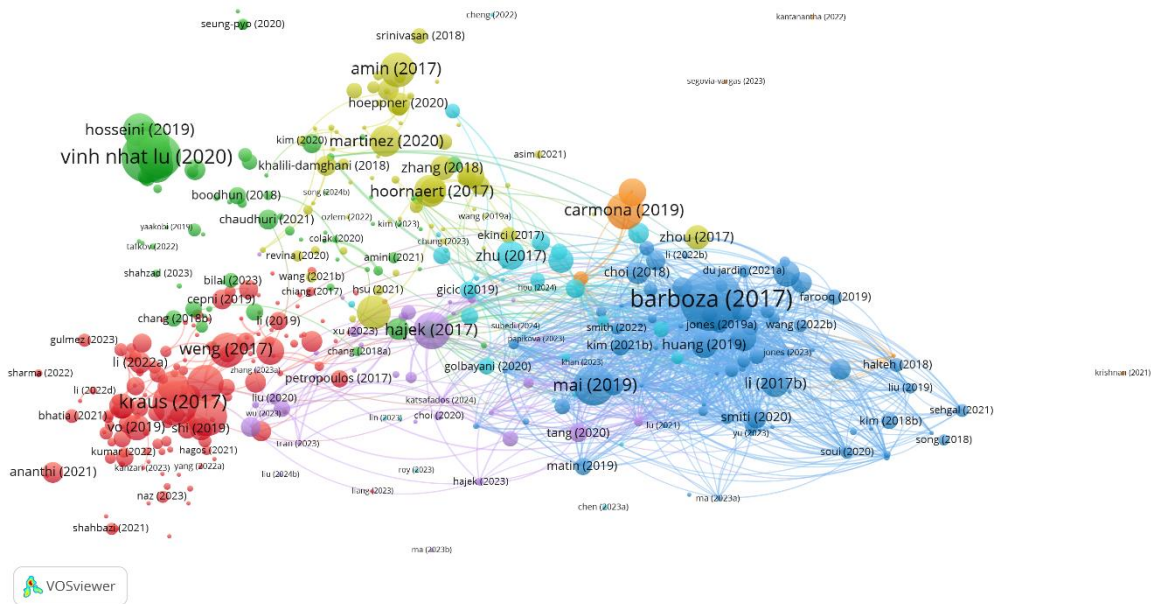


Fig. 2.5: Bibliographic coupling network

The first cluster, stock price prediction, highlights influential studies that leverage technical indicators within prediction models. Notable contributions include Kraus & Feuerriegel (2017) with 175 citations, Picasso et al. (2019) with 138, and Weng et al. (2018) with 114. This theme emphasizes the critical role of DM techniques in predicting stock prices,

illustrating the significance of market-driven predictive models. Specifically, Kraus & Feuerriegel (2017) explored the use of deep neural networks for financial decision support, aiming to predict stock price movements based on insights from financial disclosures.

The second cluster, customer-centric research, comprises 87 articles centered on the application of DM in customer service, collectively cited 1,516 times, according to WoS. The most highly cited articles in this cluster include Lu et al. (2020), with 228 citations, Steinhoff et al. (2019), with 173, and Wamba-Taguimdje et al. (2020), with 164. Lu et al. (2020) conducted a comprehensive literature review in 2017, examining the deployment of service robots in business contexts, encompassing publications from the Financial Times top 50 journals and a broad range of additional journals. Their analysis delved into the impact of service robots on customers and employees, providing critical insights into this evolving field.

The third cluster, bankruptcy and financial distress prediction, contains 77 articles cited 1,619 times based on WoS data. The leading contributions in this area are those of Barboza et al. (2017), with 342 citations, Mai et al. (2019), with 143, and Li et al. (2017), with 66. Barboza et al. (2017) introduced novel predictive variables, including the operating margin, changes in return-on-equity and price-to-book ratios, and growth indicators related to assets, sales, and employee numbers, to forecast bankruptcy one year prior to its occurrence. Their findings demonstrated that ensemble models, including bagging, boosting, and random forest, significantly outperformed other prediction techniques, providing a more reliable approach for forecasting financial distress.

Table (2.6): Thematic Clusters of DM Research in Firm's Performance Prediction

C	Theme	Article	TC
1	Stock price prediction	Kraus & Feuerriegel (2017)	175
		Picasso et al. (2019)	138
		Weng et al. (2018)	114
2	Customer-centric research	Lu et al. (2020)	228
		Steinhoff et al. (2019)	173
		Wamba-Taguimdje et al. (2020)	164
3	Bankruptcy/ financial distress prediction	Barboza et al. (2017)	342
		Mai et al. (2019)	143
		Z. Li et al. (2017)	66
4	DM in business-cycle analysis	Amin et al. (2017)	103
		Bussmann et al. (2021)	100
		Hoornaert et al. (2017)	89
5	Textual analysis	Hajek & Henriques (2017)	124
		Hájek (2018)	36
		Tang et al. (2020)	31
6	Ensemble DM approach in failure and credit risk	Carmona et al. (2019)	122
		Zhu et al. (2017)	72
		Climent et al. (2019)	66
Notes: C: Cluster, TC: Total Citations			

The fourth cluster, DM in business-cycle analysis, comprises 65 articles that were cited 1,094 times, as reported by WoS. The most cited articles in this category are Amin et al. (2017) with 103 citations, Bussmann et al. (2021) with 100, and Hoornaert et al. (2017) with 89. These studies applied ML techniques to address and analyze business challenges, including customer churn prediction, credit risk assessment, and the identification of innovative product concepts, showcasing the versatility of ML in optimizing business operations and informing strategic decision-making.

The fifth cluster, textual analysis, encompasses 38 articles focused on DM applications in textual analysis, with a total of 374 citations, according to WoS. The leading articles are Hajek & Henriques (2017), with 124 citations, Hájek (2018), with 36, and Tang et al. (2020), with 31. These studies employed textual analysis to address various issues, including fraud detection, stock return prediction, and the early identification of firm distress. This

theme highlights the significance of analyzing unstructured textual data to extract actionable financial and business insights.

The sixth cluster, the ensemble DM approach in credit risk and firms' default predictions, includes 37 articles cited a total of 682 times, based on WoS data. The top contributions in this field are Carmona et al. (2019), with 122 citations, Zhu et al. (2017), with 72, and Climent et al. (2019), with 66. These studies focus on employing ensemble ML techniques, such as bagging and boosting, to enhance the accuracy of credit risk evaluation and predict firm defaults, emphasizing the effectiveness of ensemble approaches in managing financial risk and improving predictive reliability.

2.4.1.4 *Literature Trends*

Fig. 2.6 shows the keyword evolution map generated using overlay visualization in VOSViewer software, which illustrates the progression of research themes in DM for firm performance prediction from 2017 to 2024. This was achieved by conducting a co-occurrence analysis of all terms using the "All Keywords" approach, combining author keywords with KeyWords Plus. The purpose is to uncover the thematic evolution of predicting financial performance topics, as shown in at least five articles in the review body. In VOSViewer, a complete counting method has been chosen, and the minimum number of occurrences of a term is set to 5 out of 2,325 terms; 162 meet this threshold. The closeness between the two terms directly relates to the strength of their relationship. Co-occurrence in documents establishes relatedness, with frequency of occurrence reflected in node size. In the field of DM techniques for predicting firms' performance, some of the most prominent terms are "machine learning", "model", "bankruptcy prediction", "prediction",

"performance", "classification", "deep learning", "models", "neural networks", and "impact".

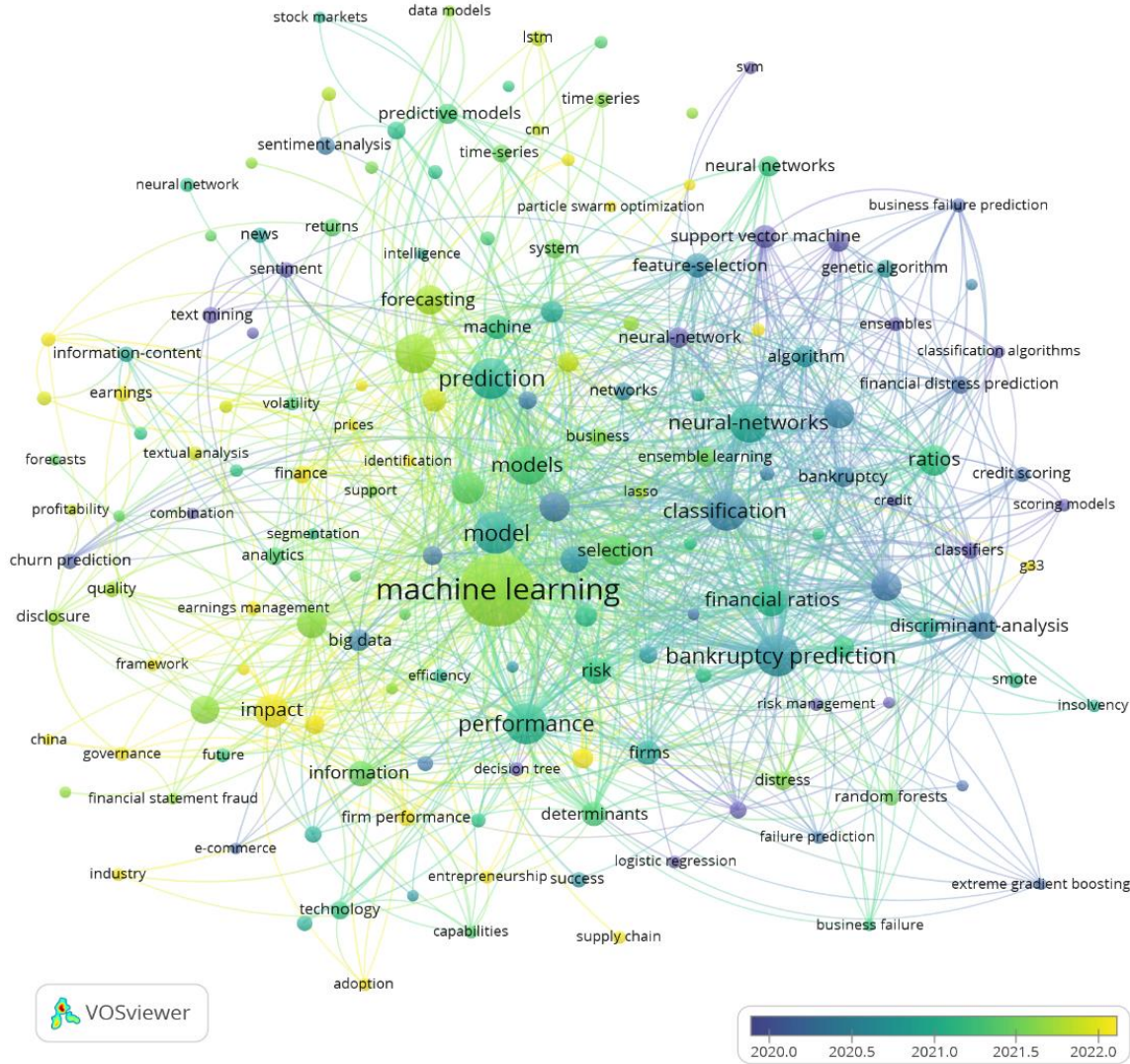


Fig. 2.6: Keyword evolution of DM research in a firm's performance prediction

The research trend of DM in firms' performance prediction before 2021 was focused on "bankruptcy prediction", "financial distress", "churn prediction", "business failure prediction", "credit risk" with "classification", "support vector machine", "neural networks", and "data mining". During 2021, the research started to shift towards "financial ratios", "firms",

"information", and "determinants", with "machine learning", "deep learning", "models", "forecasting", "regression", "neural networks", "ensemble learning" and "xgboost". In 2022, the research focused on "behavior", "finance", "impact", "corporate governance", and "firm performance".

As the keyword evolution reveals, research on predicting firms' performance using DM has progressively shifted from focusing on classification techniques and single models to emphasizing regression methods and ensemble approaches. This trend highlights the growing preference for more sophisticated and accurate predictive frameworks in financial performance prediction.

The next section will explore insight into the focused empirical studies of the research problem.

2.4.2 Systematic Literature Review

This research follows the SLR methodology outlined by Kitchenham & Charters (2007), which involves three chronological phases: (1) planning, (2) conducting, and (3) reporting the review, as illustrated in Fig. 2.7. This systematic mapping is conducted to summarize existing empirical studies on using DM to predict firm performance, identify research gaps for further investigation, and provide a framework to position new research activities. Furthermore, the SLR is documented using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Page et al. (2021), which detail the data collection and analysis processes at each stage of the study (Fig. 2.8).



Fig. 2.7: SLR phases, (Kitchenham & Charters, 2007)

The planning phase of the SLR establishes a comprehensive framework to define the scope and structure of the review. This phase is essential for ensuring methodological rigor and alignment with the study's objectives. It involves formulating precise, narrow SLR purposes and developing a detailed review protocol using the PRISMA guidelines. The main goals of SLR are:

1. Discovering the geographical distribution of datasets utilized in existing studies to identify gaps in the data sources used for firm performance prediction.
2. Identifying the variables commonly used in predicting firm performance.
3. Exploring the DM techniques and analytic tools used in existing research enables the evaluation of their effectiveness and popularity.
4. Exploring preprocessing methods in literature to enhance the accuracy of predictive models.
5. Exploring deployed evaluation metrics in the current literature to ensure that the DM models used in this study are assessed using reliable and widely accepted criteria.

For the SLR, the extraction of relevant journal articles was guided by predefined selection criteria. Table 2.7 outlines that inclusion and exclusion

criteria were systematically applied during the screening process to ensure alignment with the research objectives. For example, review articles (e.g. A. W. Li & Bastos, 2020; Lu et al., 2020), studies focusing on technical analysis such as stock prices or trading volume (e.g., Kraus & Feuerriegel, 2017; Picasso et al., 2019; Weng et al., 2018), and those employing purely statistical methods (e.g., Z. Li et al., 2017) were excluded. Conversely, only empirical studies that utilized data mining techniques and addressed firm-level financial performance through fundamental indicators were included in the analysis.

Table (2.7): SLR Inclusion/ Exclusion Criteria

Inclusion criteria	Exclusion criteria
Studies focused on fundamental analysis (Balance sheet and income statement variables).	Studies focused on technical analysis (close price, volume, etc.).
Studies used DM techniques.	Studies used statistical methods.
Studies used economic indicators.	Studies used sentiment analysis and text analysis.
Studies focused on financial performance.	Studies focused on daily/weekly stock returns.

In addition to the inclusion and exclusion criteria, a set of quality assessment questions was applied to ensure the methodological rigor of the selected studies, as shown in Table 2.8. Articles were retained only if they clearly stated their objectives, presented financial variables and DM techniques transparently, and provided well-defined evaluation metrics and results. This step ensured that only methodologically sound studies were included in the final analysis.

The second phase involves conducting the SLR. A total of 457 records from the bibliometric analysis were screened using inclusion and exclusion criteria, yielding 14 eligible articles. Additionally, nine relevant studies were

identified from other sources, resulting in a final set of 23 articles for data extraction.

Table (2.8): SLR Quality Criteria

Description
Are the objectives of the article clearly stated?
Is the methodology adequately presented?
Are financial variables clearly presented?
Are the data mining techniques clearly explained?
Are the evaluation metrics clearly defined, explained, and presented?
Are the results of the research study clearly explained?

The screening process consists of two stages: initially, titles and abstracts are reviewed, followed by a full-text review to ensure that studies meet the selection criteria for addressing the SLR goals. A descriptive (narrative) synthesis was employed to summarize the findings of the current study. The review process is thoroughly documented and presented in Fig. 2.8 using PRISMA, proposed by Page et al. (2021), to ensure the SLR reliability.

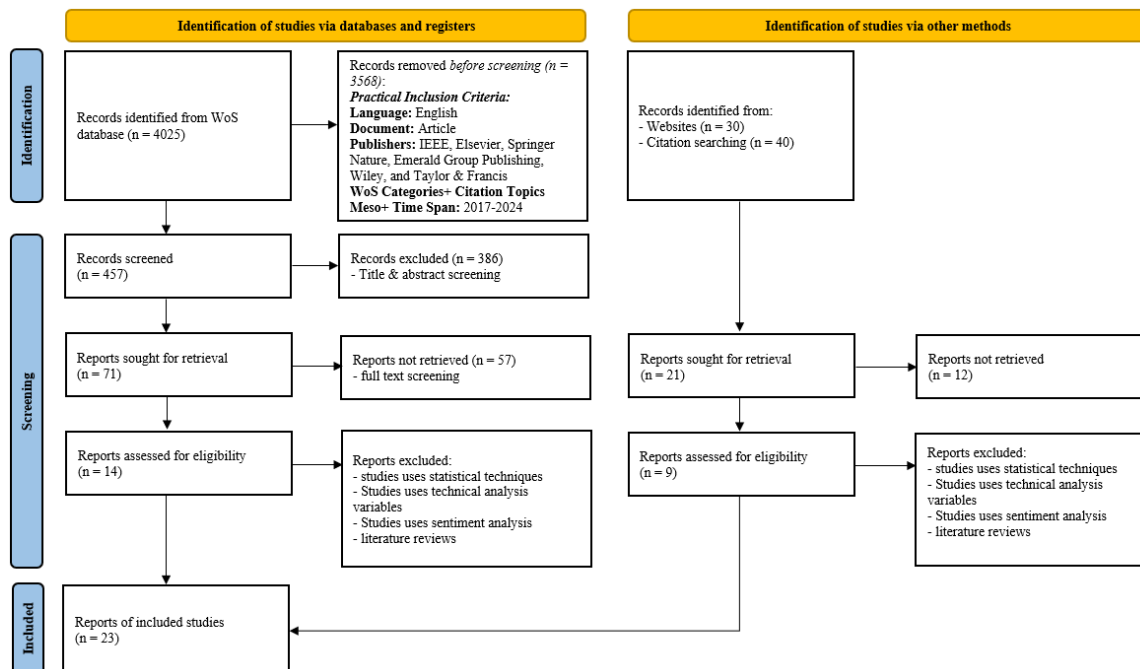


Fig. 2.8: The review process follows PRISMA guidelines (Page et al., 2021)

Table 2.9 analyzes 23 peer-reviewed articles published between 2017 and 2024, selected based on defined inclusion and exclusion criteria to align with the goals of this SLR on DM contributions to firm performance prediction.

The table summarizes essential insights related to study periods, datasets, variables, DM techniques, analytical tools, and evaluation metrics. It provides a structured synthesis of key themes in the literature, beginning with an examination of the geographical distribution of datasets to identify underrepresented regions and potential data gaps. It also highlights the most frequently used variables in predictive models, offering insight into the core indicators of firm performance. Furthermore, the table outlines the DM techniques and analytical tools applied across studies, enabling an evaluation of their effectiveness and popularity. Preprocessing methods are also reviewed, with an emphasis on their role in enhancing model accuracy and robustness. Finally, the table consolidates the evaluation metrics used in the literature to ensure that predictive models are assessed using reliable and widely accepted standards.

This comprehensive overview supports the development of more accurate and generalizable models, while also distinguishing between single and ensemble prediction approaches that have been explored in the field.

Table (2.9): Firm Performance Prediction Studies Utilizing DM Techniques

Article	Period	Dataset	Independent Variable(s)	Dependent Variable(s) /	DM Techniques	Data Split	Missing Values	Best Accuracy	Analytic Tool	Evaluation Metrics
Barboza et al. (2017)	1985- 2013	USA & Canada	11 financial indicators	Failed, non-failed	Bagging, Boosting, RF, SVM, ANN, LoR & MDA	RS	Keep missing values	Bagging, Boosting & RF	R- software	Type I & II Error & AUC
Wang & Wu (2017)	2012-2014	China	24 financial indicators	Failed, non-failed	FSOM, FSVM & ANN	CS 2y	Remove entire row	KFSOM	-	Accuracy
Halteh et al. (2018)	2014- 2015	Islamic banks	18 financial ratios & variables	Success, Failure	DT-CART, SGB & RF	CS 1y	-	DT-CART	Salford Systems Modeller 8.0	DT model
Kim et al. (2018)	2000- 2013	Korea	53 financial ratios	Failed, non-failed	Logit, MDA, ANN, data depth (DD-SVM)	RS 80%20% +CS 10y	-	DD-SVM	R- software	Accuracy, Type I & II error
Balasubramanian et al. (2019)	2011-2015	India	9 financial & 4 non-financial variables	sick, healthy	LoR	-	-	-	-	Confusion Matrix, Types I & II Error & AUC
H. Son et al. (2019)	2011-2016	Korea	108 financial ratios + 22 variables represent firm size	Normal, bankrupt	LoR, RF, Gradient Boosting (XGBoost, LightGBM), & ANN	RS	Replace with zero	XGBoost, LightGBM, & ANN	Python software	Recall, FallOut & AUC
J. P. Li et al. (2020)	2010- 2018	GCC Banks	51 Financial variables	Credit Rating (7 rating scales)	ANN, SVM, RF, CHAID & CART	CS	-	RF & CART	-	Accuracy, F1 Score & Specificity
Smiti & Soui (2020)	2002-2013	Poland	64 financial ratios	Bankrupt, nonbankrupt	A novel approach called BSM-SAES with kNN, DT, SVM, ANN & C5.0	RS 70%30% & 5y	-	BSM-SAES	Matlab 2017b platform	AUC & Sensitivity
Aljawazneh et al. (2021)	1998 -2013	Spain, Taiwan & Poland	33 Taiwan, 95 Polish & 64 Polish attributes	Bankrupt, Solvent	DBN, LSTM, & MLP, RF, SVM, kNN, AdaBoost & XGBoost	RS 90%10%	-	MLP with SMOTE-ENN balancing method	Python software	Accuracy, Recall, Specificity, Precision, & Type I & II Error

Article	Period	Dataset	Independent Variable(s)	Dependent Variable(s) /	DM Techniques	Data Split	Missing Values	Best Accuracy	Analytic Tool	Evaluation Metrics
du Jardin (2021)	2006- 2016	France	30 financial ratios	Failed, non-failed	LoR, DT-CART, k-NN, BBNNs, SVM, & ELM, bagging, AdaBoost, ensemble SONNs, RF, and XGBoost	CS	-	ensemble SONNs then XGBoost	Python package + Weka	AUC, type I & type II errors
Jabeur et al. (2021)	2014-2016	French firms	18 financial ratios	Bankrupt, nonbankrupt	CatBoost, XGBoost, GBM, SVM, RF, NN, DNN, DA & LoR	RS 70%30% + CS 3y	-	CatBoost	R-software (caret package)	Accuracy & AUC
Park et al. (2021)	2009- 2015	Korea	111 financial ratios	Bankrupt, nonbankrupt	XGB, LightGBM, & RF	RS	Replace with Median	LightGBM	-	AUC
Arora & Saurabh (2022)	2016- 2019	India	18 firm-level financial ratios	Financial distress, not- distress	LoR, lasso regression, DT, RF bagging, XGBoost & SVM	-	Replace with Median	Accuracy: RF bagging Precision: XGBoost	R- software	accuracy, recall, precision & AUC
Assous (2022)	2014- 2018	Saudi banks	17 financial ratios	Efficiency ratio	SVM, CHAID, LR, ANN	-	-	CHAID	-	R2, MSE, MAE, MBE, & RMSE
Garcia (2022)	2010-2018	USA	6 financial ratios & market variables	Bankrupt, nonbankrupt	LoR, Boosted LoR, LDA, PLS-DA, NB, kNN, NN, SVM, XGBoost, RF, & RF ensemble	RS 70%30%	-	RF ensemble	R-software (caret package)	Recall, Precision, Specificity, F measure, G mean & AUC
Khalil et al. (2022)	1999- 2019	Egyptian insurance companies	21 financial ratios	solvency ratio	SVR, RF & ensemble models (bagging & boosting)	RS 80%20%	Imputing missing values	CatBoost	Python software	MAE, RMSE & R-square
Ledhem (2022)	2011- 2020	Indonesian Islamic banks	3 financial ratios & variables	ROA	LASSO regression, RF, ANN, kNN	RS 75%25%	-	kNN & RF	Orange Libraries	MSE, RMSE, MAE, & R-squared
M. Zhou et al. (2022)	2011- 2020	China	36 financial indicators	Performance rating (category 1 to 5)	SOM-CNN, LSTM-CNN, Guassian, Linear & Quadratic SVM	-	Remove entire row	SOM-CNN	MATLAB 2018a	Accuracy

Article	Period	Dataset	Independent Variable(s)	Dependent Variable(s) /	DM Techniques	Data Split	Missing Values	Best Accuracy	Analytic Tool	Evaluation Metrics
Chiadamrong & Wattanawarangkoon (2023)	2003- 2014	Thailand	6 financial ratios	ROA+ROE+ROS / low, medium, high performance	MLR and LoR	CS 4y	-	LoR	Minitab 19	MAPE
Nguyen et al. (2023)	2013-2017	France	24 financial ratios	Failed, non-failed	LR, DA, SVM, ANN, RF, LightGBM, XGBoost, and NGBost	RS 80%20%	-	XGBoost	-	Accuracy, Type-I error, Type-II error, AUC-ROC curve
Saha et al. (2023)	2006-2019	Indian food industries	14 variables of firm size measures	WPI & PBDITA	RT, Bootstrap Forests, Boosted Trees & BART	-	-	BART	R & SAS software	R-squared & RMSEP
Ainan et al. (2024)	2000-2013	Poland	64 financial attributes	Bankrupt, successful	RF, XGBoost, SVM, RS 80%20% & hybrid XGBoost +ANN	RS 80%20%	Imputing missing values with Mean	hybrid XGBoost +ANN	-	Accuracy, Precision, AUC (Area Under the Curve), Sensitivity, Specificity
P. V. H. Son & Duong (2024)	2012 - 2022	Vietnam	14 financial variables	ROA	RF, GBR, XGBoost, KNN, SVR, Lasso, RR, & MLP	RS 70%30%	Remove entire row & impute	RF	-	R-square & RMSE

Article	Period	Dataset	Independent Variable(s)	Dependent Variable(s) /	DM Techniques	Data Split	Missing Values	Best Accuracy	Analytic Tool	Evaluation Metrics
---------	--------	---------	----------------------------	----------------------------	---------------	------------	-------------------	------------------	------------------	-----------------------

Notes:

RS: Random Split, CS: Chronological Split, y: year horizon to predict performance.

EPS: Earnings per Share, ROA: Return on Assets, WPI: Wholesale Price Index, PBDITA: Profits before depreciation, interest, tax, and amortization.

ANN: Neural Networks, SVM: Support Vector Machine, k-NN: k-Nearest Neighbor, MDA: Multivariate Discriminant Analysis, LoR: Logistic Regression, LR: Linear Regression, CLR: Cost-sensitive variation of Logistic Regression, AD: AdaBoost, AC: Ada Cost, SVM: Support Vector Machines, CSVM: Cost-sensitive Support Vector Machines, RF: Random Forest, XGBoost: Extreme Gradient Boosting, RT: Regression Tree, DBN: Deep Belief Network, BART: Bayesian Additive Regression Trees, CART: Classification & Regression Trees, CHAID: Chi-squared Automatic Interaction Detection, C4.5: classification tree, CatBoost: Categorical Boosting, XGBoost: Extreme Gradient Boosting, GBM: Gradient Boosting Machine, RF: Random Forest, DNN: Deep neural networks, DA: Discriminant Analysis, LDA: Linear discriminant analysis, BSM: Borderline Synthetic Minority oversampling technique, SAE: Stacked Auto Encoder, MLP: Multi-layer Perceptron, LSTM: Long Short-term Memory, CNN: Convolutional Neural Networks, CITs: Conditional Inference Trees, CRT: Cubist Regression Tree, MLR: Multiple Linear Regression, FSVM: Fuzzy Support Vector Machine, KFSOM: Kernel-based Fuzzy Self-Organizing Map, GBR: Gradient Boosting Regressor, SVR: Support Vector Regression, RR: Ridge Regression, SMOTE: Synthetic Minority Over-sampling Technique, ENN: Edited Nearest Neighbors, NGBoost: Natural Gradient Boosting.

RMSE: Root Mean Square Error, MSE: Mean Square Error, MAE: Mean Absolute Error, MAPE: Mean Absolute Percentage Error, MBE: Mean Bias Error, RAE: Relative Absolute Error, RSE: Relative Squared Error, AUC: Area under the ROC (Receiver Operating Characteristic) curve

The final stage of the SLR synthesizes the findings of 23 targeted empirical studies, as summarized in Table 2.9, which employ DM techniques to predict firm performance. These findings are elaborated upon in the subsequent sections.

2.4.2.1 Geographical Datasets in Literature

Table 2.10 highlights the geographical distribution of datasets used in firm performance prediction studies, as reported in 23 articles. It reveals that datasets from countries such as France, India, and Korea are most used, reflecting their strong presence in this field of DM research. These countries have contributed multiple studies that indicated a higher level of data availability or focused on firm performance prediction in these regions. For example, India is represented by three studies, and Korea and the USA each have several papers utilizing their respective datasets.

On the other hand, countries such as Canada, Egypt, Kuwait, and Poland have only one or two studies each, indicating less frequent use of datasets from these regions. In addition, some articles focus exclusively on the banking sector, such as Halteh et al. (2018), which examines Islamic banks; J. P. Li et al. (2020), which explores banks in the GCC region; and Assous (2022), which evaluates Saudi Banks. This highlights niche research areas within specific financial sectors.

It is important to note that the articles were selected based on inclusion criteria that focus on using fundamental analysis variables to measure firm performance. Fundamental analysis provides a more reliable prediction than technical analysis, as it evaluates a firm's financial health over time by

analyzing its financial statements, such as income, balance sheets, and cash flows, rather than relying on daily stock price movements. This approach offers more profound insights into a company's performance, making it a preferred method for firm performance prediction models (Barak et al., 2017).

The table also illustrates various countries, including emerging markets such as Vietnam, Thailand, and Indonesia, where firm performance prediction models are applied. These variations in dataset origin contribute to a broader understanding of how firm performance prediction models operate across diverse financial environments and regulatory frameworks.

Table (2.10): Studies Based on the Dataset Region

Articles	Country-related dataset
Ainan et al. (2024), Smiti & Soui (2020)	Poland
Aljawazneh et al. (2021)	Spain, Taiwan, Poland
Assous (2022)	Saudi Banks
Balasubramanian et al. (2019), Arora & Saurabh (2022), Saha et al. (2023)	India
Barboza et al. (2017)	Canada
Barboza et al. (2017), Garcia (2022)	USA
Chiadamrong & Wattanawarangkoon (2023)	Thailand
du Jardin (2021), Jabeur et al. (2021), Nguyen et al. (2023)	France
Halteh et al. (2018)	Islamic Banks
J. P. Li et al. (2020)	GCC Banks
Khalil et al. (2022)	Egypt
Kim et al. (2018), H. Son et al. (2019), Park et al. (2021)	Korea
Ledhem (2022)	Indonesia
Zhou et al. (2022), Wang & Wu (2017)	China
P. V. H. Son & Duong (2024)	Vietnam

2.4.2.2 *Associated Variables in Prediction Models*

Based on the reviewed articles, the dependent variables are categorized into two main types: classification and regression models, according to the

prediction model task. Furthermore, the independent variables are categorized into fundamental, macroeconomic, and non-financial factors.

As highlighted in Table 2.11, most reviewed articles have employed classification models, with a strong focus on predicting firm distress and financial stability through categorical outcomes. Classification models are used in both binary and multi-class formats to predict outcomes such as bankruptcy and credit ratings. Binary classification typically distinguishes between failed and non-failed firms (e.g., bankrupt vs. non-bankrupt), while multi-class classification assesses various credit or performance ratings. Furthermore, regression models, on the other hand, predict continuous variables such as ROA, ROE, profit, and efficiency metrics, providing a more detailed financial assessment.

Table (2.11): Studies Categorized by Dependent Variables and Data Mining Tasks

DM Task	Articles	Dependent Variable
Regression	Assous (2022), Khalil et al. (2022), Ledhem (2022), Saha et al. (2023), P. V. H. Son & Duong (2024)	Efficiency Ratio, Solvency Ratio, ROA, WPI & PBDITA
Binary Classification	Barboza et al. (2017), Wang & Wu (2017), Halteh et al. (2018), Kim et al. (2018), Balasubramanian et al. (2019), H. Son et al. (2019), Smiti & Soui (2020), Aljawazneh et al. (2021), du Jardin (2021), Jabeur et al. (2021), Park et al. (2021), Arora & Saurabh (2022), Garcia (2022), Nguyen et al. (2023), Ainan et al. (2024)	Failed vs. non-failed or Bankrupt vs. non-bankrupt
Multi-class Classification	J. P. Li et al. (2020)s, Zhou et al. (2022), Chiadamrong & Wattanawarangkoon (2023)	Credit Rating (7 classes), PR (5 classes), PR (3 classes)
Note: PR: Performance Rating		

Furthermore, as shown in Table 2.12, the independent variables in these studies are grouped into three broad types: fundamental indicators, macroeconomic factors, and non-financial variables.

- Fundamental indicators focus on variables commonly used in financial reporting, such as those derived from a firm's income statements, balance sheet, and cash flow statement (Wild & Shaw, 2019). Balance sheet ratios include liquidity ratios (measuring the firm's ability to meet its short-term obligations) and solvency ratios (measuring its ability to meet long-term and other obligations), where profitability ratios are mainly calculated from data derived from the income statement and the balance sheet.
- Macroeconomic factors encompass broader economic indicators that impact firm performance, including GDP growth rate, inflation, exchange rates, and commodity prices like oil and gold. These variables help capture the influence of external economic conditions on a firm's financial health (Kumbure et al., 2022). Only one study in the current literature tests the effect of macro and bank-specific variables on credit ratings (J. P. Li et al., 2020).
- Non-financial variables include factors such as industry type (Garcia, 2022), who used the Bloomberg industry classification, firm age (Aljawazneh et al., 2021; Balasubramanian et al., 2019), and size (Aljawazneh et al., 2021; Ainan et al., 2024; Chiadamrong & Wattanawarangkoon, 2023; Garcia, 2022; Halteh et al., 2018; H. Son et al., 2019; J. P. Li et al., 2020; Park et al., 2021; P. V. H. Son & Duong, 2024; Saha et al., 2023; Smiti & Soui, 2020). These variables play a crucial role in firm performance studies by accounting for sector-

specific dynamics and the firm's maturity, which can impact financial outcomes.

These independent variables provide a comprehensive view of firm performance, enabling models to accurately predict various financial states. The categorization of these variables into fundamental, macroeconomic, and non-financial dimensions reflects the multi-faceted approach that researchers take to predict firm performance using DM techniques.

Table (2.12): Independent Variables Used in Previous Studies

Dimensions and Variables	Articles
FUNDAMENTAL FACTORS:	
<u>Profitability</u> Retained earnings / Total assets, Return on Stockholders' Equity, ROA, Earnings per share (EPS), Net assets per share, Operating revenue per share, ROE, Return on Revenue, Cost of sales to net sales, Operating ratio, Non-operating income ratio, Financial expenses to sales, Ratio of expenditure to revenue, Financial expenses to total expenses, Accumulated depreciation ratio, Ratio of interest burden, Interest paid ratio, Financial expense to total borrowings, Accumulated earnings to stockholders' equity, Additional paid-in capital and retained earnings to stockholders' equity, Reserve ratio, R&D investment ratio, ROI, Earnings Power, Operating Income Margin, Return on Operating Assets, Return on Equity, Return on Assets, Cash flow / Shareholder funds, EBITDA / Permanent equity, EBITDA / Total assets, Profit before tax / Shareholder funds, Value added / Fixed assets, Return on shareholder funds, Return on capital employed, Return on total assets, Profit margin, Gross margin, EBITDA margin, EBIT margin, Cash ratio, Net income / Total assets, Net income / Sales, Net income before income tax expense to financial expenses, Net income before income tax expense to total assets, EBIT / Sales, Investment yield ratio, Loss ratio, Collective ratio, Operation ratio, Profit (Loss) of insurance activities to net investment income ratio, Earnings before taxes to total assets ratio, Earnings after taxes to shareholders' funds ratio, Main operating profit ratio, Operating profit ratio, Proportion of non-main business, Total return rate of assets, Major business profit ratio, Profit cost ratio, Ratio of sum of three expenses, Profit or Loss / Shareholder Funds, Profit or Loss / Capital Employed, EBIT, Gross margin, Return	Aljawazneh et al. (2021), Ainan et al. (2024), Arora & Saurabh (2022), Assous (2022), Balasubramanian et al. (2019), Barboza et al. (2017), du Jardin (2021), Garcia (2022), Halteh et al. (2018), Jabeur et al. (2021), J. P. Li et al. (2020), Khalil et al. (2022), Kim et al. (2018), M. Zhou et al. (2022), Nguyen et al. (2023), Park et al. (2021), P. V. H. Son & Duong (2024), Saha et al. (2023), Smiti & Soui (2020), H. Son et al. (2019), Wang & Wu (2017)
<u>Liquidity</u> Net working capital / Total assets, Current ratio, Loan to Deposit Ratio, Non-deposit Funding Ratio, Liquidity Coverage Ratio, Net Stability Funding Ratio, Financial Solvency, Acid Test, Cash / Current assets, Cash / Current liabilities, Cash / Total assets, Quick assets / Current liabilities, Current liabilities / Current assets, Cash / Sales, Current assets / Total assets, Working capital / Sales, Total loans / Total deposits, Current accounts + Saving accounts / Total deposits, Interest income deposits / Total deposits, Total deposits, Liquidity ratio, Total liabilities to liquid assets ratio, Liquidity	Aljawazneh et al. (2021), Ainan et al. (2024), Arora & Saurabh (2022), Assous (2022), Balasubramanian et al. (2019), Barboza et al. (2017), Chiadamrong & Wattanawarangkoon (2023), du Jardin (2021),

Dimensions and Variables	Articles
Management, Total current assets / Total current liabilities, Working Capital Requirement / Current Assets, Current Assets / Current Liabilities, Financial Debt / Cash Flow, Quick ratio.	Halteh et al. (2018), Jabeur et al. (2021), J. P. Li et al. (2020), Khalil et al. (2022), Ledhem (2022), Nguyen et al. (2023), Park et al. (2021), P. V. H. Son & Duong (2024), Smiti & Soui (2020), H. Son et al. (2019)
<u>Leverage</u> Market value of share * number of shares / Total debt, Total liabilities / Total assets, Total debt / Total capital, Debt Structure, Debt Cost, Debt Paying, Debt, Financial leverage.	Aljawazneh et al. (2021), Ainan et al. (2024), Barboza et al. (2017), Chiadamrong & Wattanawarangkoon (2023), Garcia (2022), Halteh et al. (2018), Smiti & Soui (2020), H. Son et al. (2019), P. V. H. Son & Duong (2024)
<u>Solvency</u> ICR, Financial expenses / Cash flow, Financial expenses / EBITDA, Financial expenses / Net income, Financial expenses / Total assets, Financial expenses / Value added, (Shareholder Funds / Total Assets) * 100, Gearing / Shareholder Funds, Financial expenses to sales, Equity to total assets, Debt to total assets, Financial expenses to liabilities, Total borrowings and bonds payable to total assets, Current liabilities to total assets, Debt to equity ratio, Receivables to payables, Market capitalization / Total debt, Cash flow from operations / Total debt, Total debt / Total assets, Solvency ratio, The shareholders' funds to total assets ratio, Technical provisions to shareholders' funds, Policyholders' funds and shareholders' funds to net premiums ratio, Interest cover, Debt equity ratio, Debt assets ratio, Capitalization ratio, Cash asset ratio, Capital immobilization ratio.	Arora & Saurabh (2022), Balasubramanian et al. (2019), du Jardin (2021), Garcia (2022), Jabeur et al. (2021), Khalil et al. (2022), M. Zhou et al. (2022), Park et al. (2021), H. Son et al. (2019), Wang & Wu (2017)
<u>Asset Utilization, Activity, Efficiency</u> Sales / Total assets, Inventory turnover, Current assets turnover, Total assets turnover, Account receivables turnover, Asset turnover, Net worth turnover ratio, Capital stock turnover, Turnover ratio of borrowing capital, Trade payable turnover period, Inventories to net sales, Goods in process turnover ratio, Account Receivable Collection Period, Net working capital turnover, Net Sales / Average Total Assets, Asset Rotation, Working Capital, Stock Turnover, Receivable Turnover, Accounts payable / Total sales, Current assets / Total sales, Current liabilities / Total sales, Inventories / Total sales, Receivables / Total sales, Total sales / Total assets, Total assets turnover ratio, Payables turnover ratio, Capital stock turnover ratio, Non-current assets turnover ratio, EBIT / Total assets, Sales / Total assets, Cash flow from operations / Total assets, Current assets turnover, Cash rate of sales, Ratio of net operating cash flow to liabilities, Cash flow ratio, Return on operating cash flow of an asset, Total sales / Total assets, (Cash flow / Total Sales) * 100, (Working Capital Requirement / Total Sales) * 360, (Net Current Assets / Total Sales) * 360, Total Sales / Tangible and Intangible Assets, (EBIT / Operating Revenue) * 100, (Value-Added / Total Sales) * 100, (Operating Cash flow / Value-Added) * 100, (Profit or Loss before Tax / Operating	Aljawazneh et al. (2021), Ainan et al. (2024), Arora & Saurabh (2022), Barboza et al. (2017), Chiadamrong & Wattanawarangkoon (2023), du Jardin (2021), Halteh et al. (2018), Kim et al. (2018), M. Zhou et al. (2022), Nguyen et al. (2023), Park et al. (2021), P. V. H. Son & Duong (2024), Smiti & Soui (2020), Wang & Wu (2017)

Dimensions and Variables	Articles
Revenue) * 100, Days sales outstanding, Days inventory outstanding, Days payable outstanding, Cash conversion cycle.	
<u>Productivity</u> Earnings before interest and taxes / Total assets, Growth of assets, Growth in sales, Growth in the number of employees, Operational margin, Change in return on equity, Change in price-to-book ratio, Value added per employee, Machinery & equipment per employee, Liabilities & stockholder's equity per employee, Ratio of value added to liabilities & stockholder's equity, Ratio of value added to tangible fixed assets, Ratio of value added to net sales.	Barboza et al. (2017), Kim et al. (2018)
<u>Risk Profile</u> Impaired Loans ratio, Net Charge-offs Ratio, Provisions to Income Before Provisions and Taxes, Non-performing loans coverage.	J. P. Li et al. (2020)
<u>Growth</u> Growth of assets, Growth in sales, Growth in the number of employees, Growth rate of shareholder's equity, Growth rate of total assets, Growth rate of sales, Growth rate of net income, Earnings per share growth rate, Operating revenue growth rate, Total profit growth rate, Net asset growth rate, Total assets growth rate, Net profit growth rate	Barboza et al. (2017), H. Son et al. (2019), Park et al. (2021), P. V. H. Son & Duong (2024), Wang & Wu (2017)
MACROECONOMIC FACTORS:	
GDP Growth Rate, Volatility in GDP Growth, GDP Per capita, Diversification, Volatility in Oil Prices, World Bank Government Effectiveness Index, World Bank Rule of Law Index, World Bank Control of Corruption Index, Inflation Level, Inflation Volatility.	J. P. Li et al. (2020)
NON-FINANCIAL FACTORS:	
Market Position by Business Lines, Diversity in Product Mix, Asset Size, Ownership Structure, Age of the firm, Small/medium/large, Sales, Bloomberg Industry Classification, ln(Total assets), Number of Employees (Employee count), Type of firm (Public Limited Liability Others), Linked to a Group (If the company is part of a group holding), Number of Partners (Total partners in the company), Province Code (Code of the firm's location), Number of Changes of Location (Frequency of relocations), Delay (If annual accounts were submitted on time), Historic Number of Judicial Incidences (Total judicial events since inception), Number of Judicial Incidences (Judicial events in the last year), Historical Amount of Money Spent on Judicial Incidences (Total spent since inception), Amount of Money Spent on Judicial Incidences (Amount spent last year), Historic Number of Serious Incidences (Total serious events since inception).	Aljawazneh et al. (2021), Ainan et al. (2024), Balasubramanian et al. (2019), Chiadamrong & Wattanawarangkoon (2023), Garcia (2022), Halteh et al. (2018), J. P. Li et al. (2020), Nguyen et al. (2023), Park et al. (2021), Saha et al. (2023), Smiti & Soui (2020), H. Son et al. (2019), P. V. H. Son & Duong (2024)

As shown in Table 2.12, the literature consistently utilizes a set of core independent variables commonly used in predicting firm performance. Profitability indicators such as ROA, ROE, and EBIT are the most frequently applied. The current ratio and quick ratio are the most widely used liquidity

measures, while the debt ratio is the most commonly used metric for leverage. For solvency, the ICR is the most cited. Asset turnover and current asset turnover are the primary metrics for asset utilization. Firm size, typically measured by total assets or its logarithmic form, also appears prominently across studies.

2.4.2.3 *Utilization of DM Techniques and Analytic Tools*

Table 2.7 presents the various DM techniques used in the 23 articles in the SLR, along with their best accuracy, analytical tools, and evaluation metrics. Ensemble methods, particularly bagging and boosting, were the most employed, followed by individual methods such as ANN, SVM, LR, and DT. NB was the technique that was used the least across the studies. The most frequently used analytic tools were R and Python packages, which facilitated the implementation of these models.

2.4.2.4 *Preprocessing of the Dataset in Literature*

Several studies have explored different dataset preprocessing techniques to enhance the accuracy of predictive models. Table 2.13 presents the different data split methods used in studies to train and test prediction models. Traditional random split was commonly applied, with studies using various train-test ratios such as 70%-30%, 75%-25%, 80%-20%, and 90%-10%. Some studies (Barboza et al., 2017; H. Son et al., 2019; Park et al., 2021) specified the exact number of firms to be considered in the validation and test sets.

The chronological split method, which considers time-based partitions, was also used. It applied different time horizons ranging from one year to 10

years to evaluate prediction accuracy over time. Additionally, other studies combined random and chronological splits, incorporating random with chronological splitting. This table highlights researchers' diverse strategies for splitting data for predictive modeling.

Table (2.13): Data Split Types Used in Previous Studies

Split Type	Articles	Ratio/Time Horizon
Random	Barboza et al. (2017), H. Son et al. (2019), Park et al. (2021)	Size of train/test sets by number of firms
	Aljawazneh et al. (2021)	90%10%
	Khalil et al. (2022), Nguyen et al. (2023), Ainan et al. (2024)	80%20%
	Ledhem (2022)	75%25%
	Garcia (2022), P. V. H. Son & Duong (2024)	70%30%
Chronological	Halteh et al. (2018)	1y
	J. P. Li et al. (2020)	train set (2010-2015), test set (2016-2018)
	du Jardin (2021)	3y
	Chiadamrong & Wattanawarangkoon (2023)	4y
Random & Chronological	Kim et al. (2018)	RS 80%20% & CS 10y
	Smiti & Soui (2020)	RS 70%30% & CS 5y
	Jabeur et al. (2021)	RS 70%30% & CS 3y
Notes:		
RS: Random Split, CS: Chronological Split, y: year horizon to predict performance		
Only 17 studies mention the data split method out of 23.		

The literature also addresses handling missing values. Table 2.7 highlights various methods researchers employ to address missing values in their data. Barboza et al. (2017) keep the missing values in the dataset. In contrast, M. Zhou et al. (2022) removed entire rows with missing values. Studies such as Arora & Saurabh (2022), Khalil et al. (2022), and Ainan et al. (2024) replace missing values with Median and Mean. While P. V. H. Son & Duong (2024) combined row removal with imputation, other studies replaced

missing values with specific measures, such as filling similar values of correlated variables and then Median for noncorrelated variables, as seen in the works of Park et al. (2021) or zero, as used by H. Son et al. (2019). These approaches reflect the diverse strategies employed to manage missing data, tailored to the study's context and objectives.

2.4.2.5 *Evaluation Metrics of Prediction Models*

In the 23 studies selected for the SLR, various metrics are utilized depending on the model type and task to evaluate the performance of prediction models in firm performance studies (Table 2.9). Commonly used metrics include Accuracy and AUC for classification tasks, which assess the model's overall prediction power. Additionally, Type I and Type II Errors are frequently reported to measure misclassification rates in binary classification tasks. For regression models, error-based metrics such as RMSE, MAE, and R-squared are often used to evaluate the models' ability to predict continuous variables, including profitability and solvency.

2.5 Research Gap

Despite the existing literature and progress within the field of firm performance prediction using DM techniques (Aljawazneh et al., 2021; Ainan et al., 2024; Arora & Saurabh, 2022; Assous, 2022; Balasubramanian et al., 2019; Barboza et al., 2017; Chiadamrong & Wattanawarangkoon, 2023; du Jardin, 2021; Garcia, 2022; Halteh et al., 2018; Jabeur et al., 2021; J. P. Li et al., 2020; Khalil et al., 2022; Kim et al., 2018; Ledhem, 2022; M. Zhou et al., 2022; Nguyen et al., 2023; Park et al., 2021; P. V. H. Son & Duong, 2024; Saha et al., 2023; Smiti & Soui, 2020; H. Son et al., 2019; Wang & Wu, 2017),

several research gaps remain. This section presents these ‘neglect-spotting’ to establish the foundation for this study. Four significant research gaps form the foundation of this PhD study:

1. There is a lack of DM techniques applied to predict the performance of all publicly listed companies in the GCC using fundamental factors, rather than technical factors. A review of recent research published between 2017 and 2024 in leading journals, including IEEE, Elsevier, Springer Nature, Emerald, Wiley, and Taylor & Francis, indicates that only two studies have examined the performance prediction of GCC banks (Assous, 2022; Li et al., 2020).
2. Previous studies have not examined ROI as a dependent variable to predict firms’ financial performance, despite its holistic applicability in measuring profitability across industries.
3. While previous studies have utilized various financial ratios as independent variables, certain key metrics, such as market value and specific macroeconomic indicators, remain untested in models predicting profitability.
4. Prior studies have not explored the CatBoost regressor for predicting firm profitability; it has only been tested as a regressor for predicting the solvency ratio (e.g., Khalil et al., 2022).

The first research gap highlights the lack of a comprehensive study applying DM techniques to predict firm performance across the GCC region. Existing studies predominantly focus on datasets from France, India, Korea, and Poland, USA, and China, likely due to greater data availability or research interest (Ainan et al., 2024; Arora & Saurabh, 2022; Balasubramanian et al., 2019; Barboza et al., 2017; du Jardin, 2021; Garcia, 2022; Halteh et al., 2018;

H. Son et al., 2019; Jabeur et al., 2021; Kim et al., 2018; Ledhem, 2022; Nguyen et al., 2023; Park et al., 2021; Saha et al., 2023; Smiti & Soui, 2020; Wang & Wu, 2017; Zhou et al., 2022). In contrast, only two studies have examined GCC datasets, both of which focused exclusively on the banking sector (e.g., Assous (2022) on Saudi banks and Li et al. (2020) on GCC banks) for a limited number of firms. This study examines the accuracy of DM prediction using the GCC's dataset of publicly listed firms to address the region's existing research gap. Sourced from LSEG Workspace, it comprises 20 years of annual financial and non-financial data. Seven DM techniques were applied to analyze and predict the financial performance of publicly listed GCC firms.

The second research gap concerns the nature of the dependent variables used in previous studies on firm performance prediction using DM techniques. These studies generally fall into two main categories: classification, which deals with binary or multi-class tasks, and regression, which predicts numerical outcomes. Most studies have focused on classification models to predict firm performance, such as failure/bankruptcy prediction or credit rating/performance scores. Conversely, regression models have been employed in fewer studies to predict firm performance using financial ratios, such as efficiency, profitability, and solvency. Examples include predicting cost efficiency via the efficiency ratio (Assous, 2022), profitability through ROA (Ledhem, 2022; Son & Duong, 2024), profit using real PBDITA (Saha et al., 2023), and insolvency via the solvency ratio measured by net earned premiums/total share funds (Khalil et al., 2022). However, ROI has not been utilized as a financial performance measure to compare firms' profitability, which will be addressed by examining ROI as a dependent variable and a

measure of the profitability of GCC firms. ROI is considered a widely used measure of investment efficiency and performance comparison (Phillips & Phillips, 2019). It provides a comprehensive assessment of financial performance and investment efficiency, aligning with the thesis objective.

The third research gap examines the independent variables employed in previous studies, which are categorized into financial, non-financial, and macroeconomic factors. Most literature focuses on financial ratios, such as profitability, liquidity, leverage, solvency, and asset utilization, often excluding market value metrics. The CPB ratio, used by Barboza et al. (2017), is an exception. Macroeconomic factors are only addressed by J.P. Li et al. (2020), using specific measures. Non-financial factors, such as firm size, are commonly used, while age (e.g., Balasubramanian et al., 2019; Aljawazneh et al., 2021) and industry type (e.g., Garcia, 2022) are less explored. To address this gap, new measures are introduced for the independent variables: Total Assets to Common Equity (TACE), Total Assets to Total Equity (TATE), for leverage, Net Operating Asset Turnover (NOAT) and Fixed Asset Turnover (FAT) for asset utilization, Price-to-Earnings Ratio (P/E), Price-to-Sales Ratio (P/S), and Price-to-Book Ratio (P/B) for market value, and real GDP and inflation rate for macroeconomic factors. Table 2.14 presents the utilization of variables in previous studies and the current research.

The fourth research gap highlights the limited use of the CatBoost technique, while other boosting ensemble DM methods dominate firm performance prediction due to their high predictive accuracy. CatBoost has been used as a regressor for predicting the solvency ratio (Khalil et al., 2022) and as a classifier for bankruptcy prediction (Jabeur et al., 2021). Jabeur et al. (2021) also recommended exploring CatBoost's application to datasets

beyond bankruptcy prediction. This study addresses this issue by implementing the CatBoost regressor to predict firm profitability, measured by ROI.

Table (2.14): Literature Summary and Research Gap

Study	Firm Performance Determinants (Independent Variables)												Firm Performance (Dependent Variable)	
	<u>Firm-Specific</u>										<u>Macro</u>		Regression	Classification
	Profitability	Liquidity	Leverage	Solvency	Asset Utilization	Market Value	Size	Age	Industry Type	Others	GDP	Inflation		
Barboza et al. (2017)	√	√	√		√	√				√				√
Wang & Wu (2017)	√			√	√					√				√
Halteh et al. (2018)	√	√	√		√		√			√				√
Kim et al. (2018)	√				√					√				√
Balasubramanian et al. (2019)	√	√		√				√		√				√
H. Son et al. (2019)	√	√	√	√			√			√				√
J. P. Li et al. (2020)	√	√					√			√	√	√	√*	
Smiti & Soui (2020)	√	√	√		√		√							√
Aljawazneh et al. (2021)	√	√	√		√		√	√		√				√
du Jardin (2021)	√	√		√	√					√				√
Jabeur et al. (2021)	√	√		√	√									√
Park et al. (2021)	√	√		√	√		√			√				√
Arora & Saurabh (2022)	√	√		√	√									√
Assous (2022)	√	√								√			√	
Garcia (2022)	√		√	√			√		√	√				√
Khalil et al. (2022)	√	√		√						√			√	
Ledhem (2022)		√								√			√	
M. Zhou et al. (2022)	√			√	√					√				√*
Chiadamrong & Wattanawarakoon (2023)		√	√		√		√			√				√*
Nguyen et al. (2023)	√	√			√		√			√				√
Saha et al. (2023)							√						√	
Ainan et al. (2024)	√	√	√		√		√			√				√
P. V. H. Son & Duong (2024)	√	√	√		√		√			√			√	
This Study		√	√	√	√	√	√	√	√		√	√	√	

Note: * Multi-class Classification

While previous studies have employed various DM techniques to predict firm performance, a significant gap remains in the literature regarding

the use of ROI as a dependent variable for predicting financial performance. Addressing this gap within the context of the GCC and exploring the factors influencing profitability for all publicly listed firms in the region will contribute to academic literature and provide valuable insights for policymakers aiming to promote sustainable economic development in the region.

2.6 Conceptual Framework

Fig. 2.8 presents the study's research model, which summarizes the relationships between the research variables. According to the proposed model, financial performance, measured by profitability, is the dependent variable.

Dependent variable: According to Phillips & Phillips (2019) and Robinson et al. (2020), ROI is a commonly used metric to evaluate an investment's efficiency or profitability and compare the performance of multiple investments. Consequently, this thesis deployed ROI as a dependent variable to predict the financial performance of listed firms in the GCC countries.

Independent variables: present the firm-specific and macroeconomic variables. Firm-specific variables are internal factors the firm manages (e.g., liquidity, leverage, solvency, asset utilization, and market value). Liquidity is measured by the CuR, QR, and CaR. Leverage is measured by DR, TACE, TATE, and TDPTE. Solvency is measured by ICR. Asset utilization is measured by AT, NOAT, CAT, and FAT. Market value is measured by P/E, P/S, and P/B ratios.

This study introduces several novel measures to assess financial dimensions. For leverage, the TACE and TATE are utilized. The NOAT and FAT ratios are employed to evaluate asset utilization. Finally, the P/E Ratio, P/S Ratio, and P/B Ratio are applied to measure market value. These measures are introduced to assess firm performance across different financial dimensions comprehensively. Furthermore, macroeconomic variables, which are external factors beyond the firm's control (e.g., real GDP, and the inflation rate), are tested as independent variables.

Control variables include firm size, age, industry type, and country.

Thus, predicting the profitability of GCC publicly listed companies measured by ROI is based on several independent and control variables, as presented in Fig. 2.9.

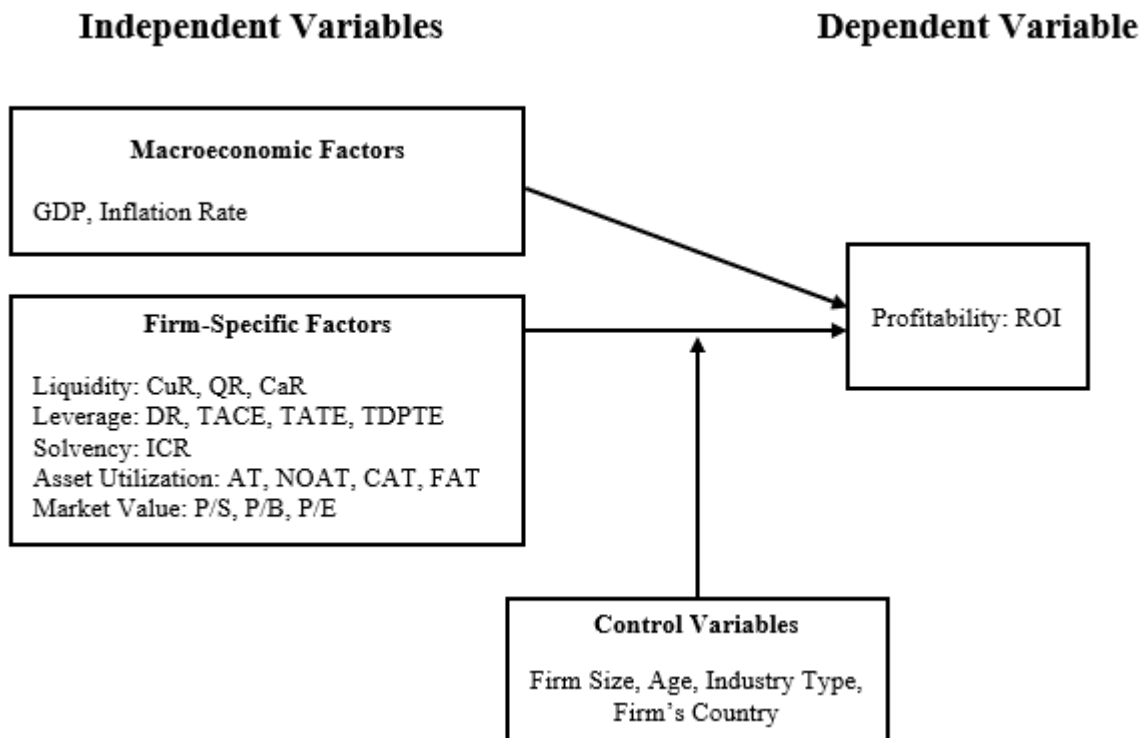


Fig. 2.9: Conceptual Framework of Firms' Performance Model and its Determinants

2.7 Summary

This chapter explored firm performance through relevant theoretical perspectives, emphasizing ROI as a primary profitability measure. It also reviewed the application of DM techniques in predictive modeling, encompassing classifiers and regressors.

Combining two review approaches, bibliometric analysis, and SLR, this research provides a pioneering comprehensive review framework for mapping the research landscape and identifying gaps in applying DM techniques for firm performance prediction.

The chapter concludes by introducing the conceptual framework, which delineates the relationships between independent and dependent variables within the context of GCC firms. Grounded in the reviewed literature, this framework will serve as the foundation for developing predictive models to assess the performance of GCC firms. Notably, testing ROI as a dependent variable will address a gap in literature by employing a comprehensive measure of profitability through DM techniques, furthermore, testing new measures (e.g., leverage: TACE, TATE; asset utilization: NOAT, FAT; market value: P/E, P/S, P/B Ratios) to assess ROI prediction.

CHAPTER 3: THE GCC MARKET AND ECONOMIC CONTEXT

3.1 Introduction

Stock markets serve as vital engines of economic growth, shaping the financial landscape of nations. This chapter delves into GCC countries' dynamic stock markets and macroeconomic conditions, highlighting their role in driving economic growth. An empirical study spanning 94 countries from 1960 to 1985 by Atje & Jovanovic (1993) demonstrated that well-developed stock markets positively correlate with economic growth. Their findings underscore the significance of financial markets in fostering prosperity, an insight that sets the stage for exploring the GCC's economic landscape.

Due to the volume and complexity of information, the financial data of publicly listed companies in GCC stock markets spanning over two decades can be classified as BD. Furthermore, conducting in-depth analyses for all listed firms' data stored in a Microsoft Excel file containing financial and nonfinancial information over 20 years (2003–2020) using various DM techniques presents a considerable computational challenge.

This chapter will explore the evolution of GCC stock markets and their current performance, providing an overview of key milestones and developments. Furthermore, it will highlight the impact of global economic factors and regulatory frameworks on these markets, emphasizing how external influences have shaped the performance of GCC stock exchanges.

3.2 GCC Stock Markets

In the 1970s and 1980s, the GCC countries began establishing financial markets to support their expanding economies. The Saudi Stock Exchange (Tadawul) was founded in 1984, making it the first stock market in the region. Other GCC countries followed suit, with the Kuwait Stock Exchange (KSE) established in 1977, the Bahrain Bourse (BSE) in 1987, the Qatar Stock Exchange (QSE) in 1995, the Dubai Financial Market (DFM) in 2000, the Abu Dhabi Securities Exchange (ADX) in 2000, the Muscat Securities Market (MSM) in 1988, and others (Bley & Chen, 2006).

Table (3.1): Listed Firms in GCC Market and Allowable Share Ownership for GCC Citizens in 2023 (GCC-STAT, 2024)

	Index	No. of Listed Firms in 2023	% of allowable share ownership for GCC citizens	Currently invested % of GCC citizens
Bahrain	Bahrain Bourse All Share Index (BAX)	41	100%	6%
Kuwait	Kuwait All Share Index (KWSE)	149	100%	20%
Oman	Muscat Securities MSM 30 Index (MSM)	109	100%	15%
Qatar	Qatar Exchange Index (QSI)	51	100%	7%
Saudi Arabia	Tadawul All Share Index (TASI)	231	97%	31%
Abu Dhabi	Abu Dhabi Stock Exchange General Index (ADXG)	96	92%	13%
Dubai	Dubai Financial Market General Index (DFM)	62	92%	8%
Total		739	97%	100%

Table 3.1 provides an overview of the listed firms in GCC markets, detailing the number of firms, the percentage of shares GCC citizens are

allowed to own, and their current investment levels as of 2023. GCC citizens have near or full ownership rights across the region, with the highest actual investment in Saudi Arabia (31%), while Bahrain records the lowest (6%). The data, sourced from (GCC-STAT, 2024), highlights significant variations in GCC citizens' market engagement.

Based on the dataset imported from LSEG Workspace, Fig.3.1 presents the number of actively listed firms in GCC countries from 2003 to 2023. Saudi Arabia demonstrates the most significant growth in listings, while countries such as Bahrain and Oman show relatively stable trends. Although 2024 data may appear relevant, it was excluded to maintain consistency, as the dataset was finalized in December 2023. Including partial-year data, which is available for the first half of 2024 from GCC-STAT (2025) could compromise the longitudinal integrity of the analysis.

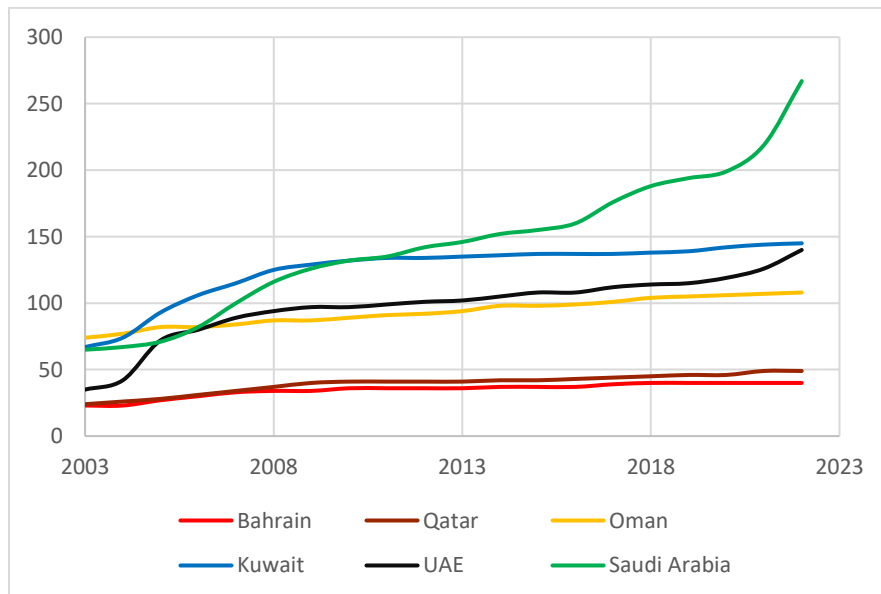


Fig. 3.1: Number of actively listed firms for the period 2003-2023 in GCC countries

In 2024, the aggregate market capitalization of GCC markets increased by \$126.53 billion, rising from \$2.25 trillion in 2023 to \$2.37 trillion. As detailed in Table 3.2, Saudi Arabia leads the GCC stock markets with a market capitalization of \$909.90 billion (excluding Saudi Aramco), followed closely by the UAE at \$1.06 trillion.

Saudi Aramco remained the most prominent company, with a market cap of \$1.81 trillion, far exceeding its closest peer, the UAE's International Holding Company, valued at \$242 billion (Argaam, 2025).

Table (3.2): GCC Stock Markets Performance

	Bahrain	Kuwait	Oman	Qatar	Saudi Arabia	UAE ^c	GCC
Stock trading began	1957	1952	1988	1997	1935	1989	
Electronic trading since	1987	1995	1988	2002	1988	2000	
Market cap. (US\$bn) ^a	2023	20.66	130.44	61.89	171.60	872.89 ^b	994.91
	2024	20.48	142.06	71.75	170.57	909.90 ^b	1064.17
		2379.12					

Notes:

^a Market capitalization (cap.) or market value is expressed in US\$ billion (bn), adopted from (Argaam, 2025)

^b Excluding Saudi Aramco: \$2.13 trillion in 2023 and \$1.81 trillion in 2024.

^c The statistics for UAE include the two stock markets: DFM and ADX.

The following outlines the evolution of each GCC financial market and the current economic landscape.

3.2.1 Kingdom of Bahrain

The Bahraini stock market dates back to 1957 with the National Bank of Bahrain's IPO but was formally established as the BSE in 1987, starting with 29 listed firms. Initially regulated by a government-appointed board, the

market opened to foreign investors in 1999, allowing full GCC ownership and 49% non-GCC ownership of local shares (Bley & Chen, 2006). Market capitalization grew from \$2.7 billion in 1989 to \$20.48 billion in 2024, though it remains the lowest in the GCC, with 41 listed firms (Argaam, 2025; GCC-STAT, 2024).

3.2.2 Kuwait

Share trading in Kuwait began in 1952 with the establishment of the Kuwait National Bank, followed by the KSE becoming an independent institution in 1983. The Iraqi invasion in 1990 halted trading until 1992. In 1995, an electronic trading system was introduced, and in 2000, GCC nationals gained access to the market, while non-GCC investors were restricted to overseas mutual funds. The Waad Future Financial Investment Service, the first equity futures market in the Arab world, was launched in 1996 for a limited number of stocks (Bley & Chen, 2006). Trading activities are mainly dominated by the Kuwait Investment Authority (KIA) and local commercial banks. At the end of 2023, 145 firms were listed at KSE (GCC-STAT, 2024), with a combined market value of about \$142.06 billion in 2024 (Argaam, 2025).

3.2.3 Oman

MSM was established in 1988 and later reorganized as an automated trading and settlement system in 1998. The stock exchange is a governmental entity regulated by the Capital Market Authority (CMA). The settlement procedure usually takes two days and is carried out through a settlement bank with a settlement guarantee fund. The Sultanate of Oman welcomes international investors and allows up to 100% foreign ownership (Bley & Chen, 2006). At

the end of 2023, 109 firms were listed at MSM, with a market capitalization reached \$71.75 billion in 2024 (Argaam, 2025).

3.2.4 Qatar

Trading on the Doha Securities Market started manually in 1997, with electronic central clearing and settlement, and became fully automated in 2002, noting that the market began operations with 17 companies (Bley & Chen, 2006). By the end of 2023, 53 listed firms were registered in the QSE, with a total market capitalization of \$170.57 in 2024 (Argaam, 2025).

3.2.5 Kingdom of Saudi Arabia

Saudi Arabia is the GCC country's most prominent stock market, with a market capitalization of \$909.90 billion in 2024, excluding Saudi Aramco, which topped the list of GCC firms with a total value of \$1.81 trillion. The total Saudi market capitalization reached \$2.72 trillion in total. The market is also the most active in the GCC region, with a trading value of \$496.6 billion, excluding Aramco's \$11.2 billion deals (Argaam, 2025).

Stock trading commenced with the Arabian Automobile Company in 1935, but it was irregular until the privatization of several electricity companies in the early 1960s. Most of the currently listed companies were made publicly available in the late 1970s and early 1980s, resulting in 231 listed firms by 2023. Market activities in Saudi Arabia were mainly unregulated until 1985, when the Saudi Arabian Monetary Authority (SAMA) granted exclusive authorization to 12 domestic and commercial banks to act as brokers. In 1988, electronic trading and settlement systems were

introduced. The Saudi Arabian CMA was established in 2003 to oversee and improve the country's capital markets (Bley & Chen, 2006).

3.2.6 United Arab Emirates

Since its establishment in 2000, the combined market capitalization of Dubai and Abu Dhabi has driven the UAE stock market to become the GCC's second largest, reaching \$1.06 trillion by December 2024. ADX surged to \$817.05 billion, up \$9.48 billion from 2023, while DFM grew by \$59.78 billion, bringing its total market value to \$247.12 billion (Argaam, 2025). The Emirates Securities and Commodities Authority (ESCA) oversees financial market activities and reports to the Ministry of Economy and Commerce (Bley & Chen, 2006).

3.3 GCC Economics

GCC countries continue to pursue economic diversification as a strategic response to reduce dependency on hydrocarbons and enhance long-term resilience. Recent efforts have focused on integrating financial markets, supporting non-oil sectors, and developing capital markets to attract foreign investment and stimulate private sector growth. The 2024 adoption of the cross-border investment fund registration is a critical milestone in fostering a unified financial environment that supports innovation, economic integration, and sustainable development across the region (GCC-STAT, 2025).

Fig. 3.2 illustrates the GDP growth of GCC countries from 2003 to 2023 in US dollars, based on WDI (World Bank, 2024).

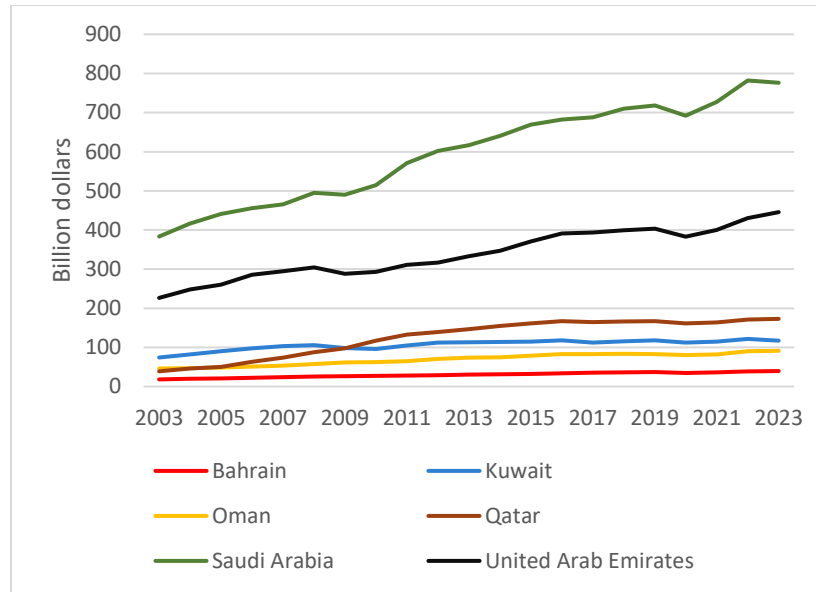


Fig. 3.2: Real GDP (US\$ in Billions) for GCC countries, (World Bank, 2024)

Saudi Arabia and the UAE consistently dominate in GDP size, with significant growth over time, particularly post-2020. Other countries, including Bahrain, Kuwait, Oman, and Qatar, exhibit steady but smaller GDP contributions, highlighting regional economic disparities. The marked increase in 2021 and 2022 reflects the recovery in global oil prices and the expansion of non-hydrocarbon sectors in these economies.

The non-hydrocarbon growth in the GCC region was supported by ongoing reforms aimed at diversifying the economy, high domestic demand, and substantial capital inflows (IMF, 2024).

Fig. 3.3 illustrates the GCC region's real GDP growth by quarter from 2021 to 2023. In 2023, voluntary oil production decisions had a significant impact on the growth of oil and gas producers in the Middle East and North Africa (MENA) region. Specifically, GCC countries saw a notable slowdown in hydrocarbon growth due to several rounds of voluntary production cuts

implemented by some Organization of the Petroleum Exporting Countries (OPEC)+ countries. As a result, real GDP growth in the GCC decreased sharply to 0.4 percent, despite non-hydrocarbon solid growth.

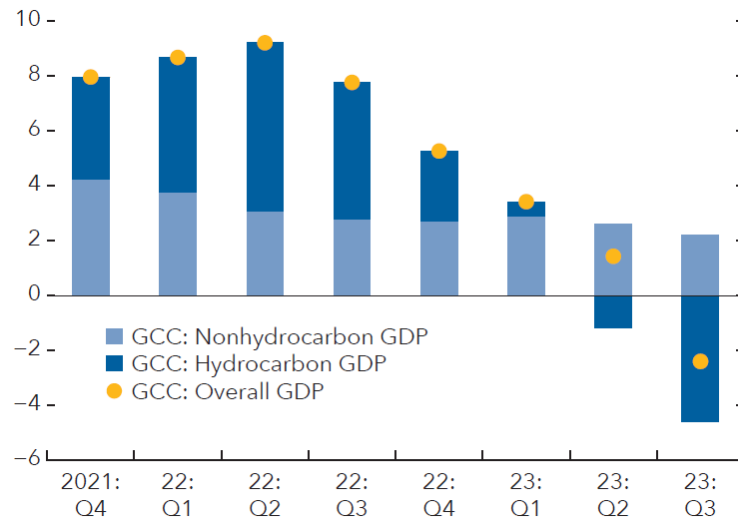


Fig. 3.3: Real GDP Growth (Percent change, year over year; weighted averages), (IMF, 2024)

Fig. 3.4 illustrates the inflation rate trends across GCC countries from 2003 to 2023 (World Bank, 2024). The chart highlights significant fluctuations in inflation rates, with particular volatility observed during the global financial crisis in 2008 and the COVID-19 pandemic in 2020. Bahrain, Qatar, and the UAE show the most pronounced fluctuations, while Saudi Arabia and Kuwait exhibit relatively stable inflation patterns throughout the COVID-19 period. This variation in inflation trends highlights the diverse economic factors that influenced each GCC nation during the observed period. In 2023, the inflation rate decreased across GCC countries.

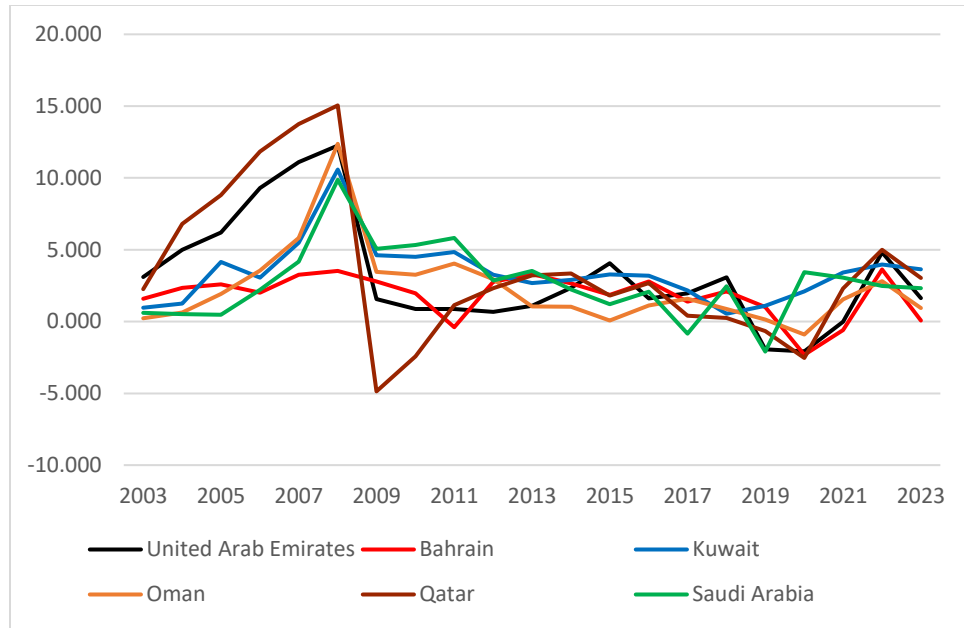


Fig. 3.4: Inflation rate for GCC countries (World Bank, 2024)

Post-COVID-19, inflation surged globally to multi-decade highs due to recovering demand, supply chain disruptions, and rising commodity prices, worsened by Russia's war in Ukraine, as shown in Fig. 3.5.

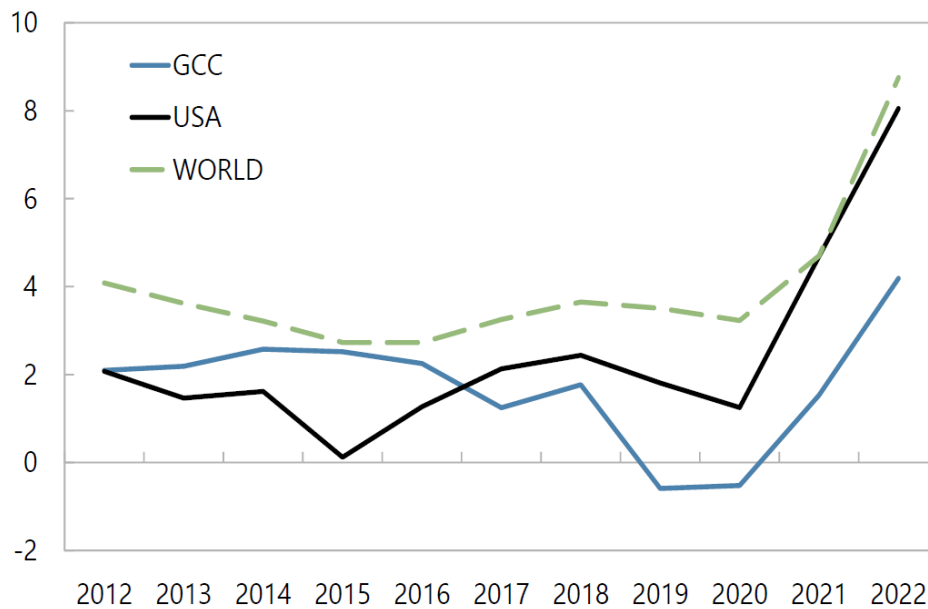


Fig. 3.5: CPI inflation in GCC, USA, and the world (year-over-year, average), (Fareed et al., 2023)

While GCC inflation remains lower than global levels, it has significantly increased. Identifying inflation drivers and crafting effective policies are vital to ensuring price stability, anchoring expectations, and sustaining economic activity (Fareed et al., 2023).

3.4 Regulatory and Legislative Developments

In 2023, GCC financial markets advanced regulations to improve transparency and governance. These reforms included stricter disclosure policies, the introduction of green bonds, securities registration frameworks, and reduced tax burdens for foreign investors. Regulatory bodies also aligned corporate governance with international best practices to safeguard investments and enhance market appeal. Cross-border collaboration increased, supported by memoranda of understanding with global organizations. These measures aim to strengthen investor confidence and position GCC markets as competitive global players (GCC-STAT, 2024).

3.5 Summary

Chapter 3 explored the evolution and performance of GCC financial markets, presenting key statistics on listed firms and overall market activity in recent financial years. It also examined the influence of global economic conditions and regulatory frameworks on the performance of GCC stock exchanges, highlighting the impact of both internal and external factors in shaping these markets over time. Additionally, the chapter emphasized the ongoing efforts of executive and supervisory authorities within GCC stock markets to enhance their supervisory, regulatory, and legislative frameworks.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Introduction

This chapter outlines the methodology of the experimental framework for performance prediction models specific to the GCC publicly listed companies. It discusses research paradigms, methods, and the DM framework. Additionally, it covers data collection, data preprocessing techniques, data sampling, and splitting. The framework is designed to systematically map the construction of four DM prediction models (explained in Section 4.8) based on seven DM techniques, single and ensemble, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost (detailed in Section 4.8.2). These DM models vary based on data preprocessing, sampling, and splitting methods.

Finally, the DM model evaluation and variable importance methods are explained in Sections 4.9 and 4.10.

4.2 Research Paradigm and Philosophies

The ontological assumptions in social research influence epistemological assumptions, which subsequently impact methodological concerns. This pertains to how reality or knowledge is acquired. It emphasizes identifying the correct methods for seeking reality or truth, mainly focusing on the relationship between the researcher and the knowledge or truth they aim to uncover. While ontology deals with the nature of truth, epistemology focuses on how to study and understand existing reality. In essence, ontology asks the

question "what," whereas epistemology asks the question "how" (Mulisa, 2022).

The current study follows the objectivism realism ontological paradigm in studying the performance of GCC firms, which is independent of the researcher's perceptions; it is, in fact, the discoverable existence of a measurable relationship between dependent and independent variables in the GCC firms.

Epistemology revolves around a theory of knowledge, concentrating on acquiring and validating knowledge. Positivists view social phenomena as a world of causal relationships where a single truth can be discovered and verified. This tradition uses case studies to test hypotheses, employing deductive approaches and relying on quantitative data collection and analysis techniques (Takahashi & Araujo, 2020).

This study adopts a positivist epistemological perspective to uncover the relationships among observable and measurable variables. Training DM models on GCC firms' data will enhance understanding firms' performance. DM generally aims to find patterns or make predictions based on data without subjectivity, free from subjective influences.

4.3 Research Methods

The study uses a predictive research approach, which focuses on making predictions based on analyzed phenomena (Collis & Hussey, 2021; Cooper & Schindler, 2014). It seeks to generalize findings by examining the accuracy of predicting the performance of publicly listed companies in the GCC. This is

done by exploring the relationships between specific dependent and independent variables that have been hypothesized.

Using a deductive approach, the study developed a conceptual framework in section 2.6, which was empirically tested to deduce relationships among variables. The study gathered quantitative data from LSEG Workspace (Section 4.5.1), which entails numerical data, and analyzed in detail in Section 5.2. Additionally, the researcher attempted to incorporate qualitative techniques by interviewing experts from the GCC CMAs; however, no response was received to these requests.

All methods discussed align with the objectivism ontological, scientific positivist epistemological paradigm. However, while traditional quantitative research emphasizes hypothesis testing, DM techniques adopt a predictive approach, enabling insights to guide future decisions rather than merely testing existing theories.

4.4 Framework of Prediction Model

The execution of DM projects typically adheres to a structured process designed to enhance the likelihood of success. DM researchers and practitioners have established a variety of workflows and step-by-step methodologies aimed at optimizing project outcomes. The development of these strategies has culminated in the establishment of several standardized processes (Sharda et al., 2021).

Sharda et al. (2021) noted that the 'My Own' category, which signifies user-defined methodologies, ranks as the second most frequently employed

approach, following the Cross Industry Standard Process for DM (CRISP-DM). This illustrates the adaptability and effectiveness of customized frameworks in meeting distinct research objectives, surpassing even the Sample, Explore, Modify, Model, and Assess (SEMMA) approach in utilization.

The current study presents a tailored framework for implementing DM models to predict the performance of GCC firms, as illustrated in Figure 4.1.

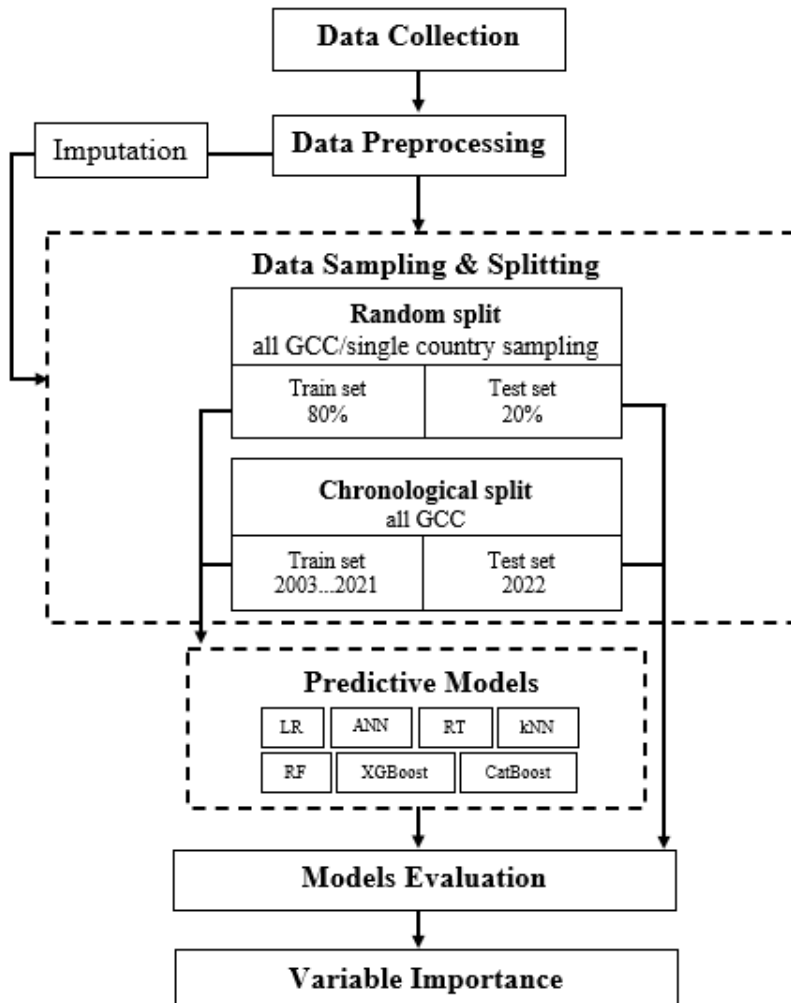


Fig. 4.1: Framework of DM Model

This DM model framework comprises a multi-stage process, which will be executed systematically. The stages include data collection (described in Section 4.5), data preprocessing (outlined in Section 4.6), data sampling and splitting (detailed in Section 4.7), construction of predictive models (detailed in Section 4.8), model evaluation using four performance metrics (elaborated upon in Section 4.9, Chapter 6), and, ultimately, the analysis of variable importance (presented in Section 4.10, Chapter 6).

4.5 Data Collection and Analysis Tools

4.5.1 Dataset Sources

The population under consideration includes all publicly listed companies in various industries of the GCC from 2003 to 2022 (20 years). Firm-specific data, including data from the firm's profile and financial statements presented in balance sheets and income statements, which are secondary data that were sourced from LSEG Workspace (formerly Refinitiv) on December 5, 2023, (LSEG, 2023). The macroeconomic variables were sourced from the WDI (World Bank, 2024).

The dataset, collected from LSEG Workspace, spans from Fiscal Year 2022 (FY0) and FY1 (2021) back to FY-19 (2003), resulting in an Excel Sheet of 816 observations (rows) and 23 variables. Additionally, the dataset obtained from the WDI presents the economic indicators of GCC countries over the same period, using real GDP (constant 2015 US\$) and the annual inflation rate of consumer prices. The sourced dataset was processed by converting the 20 columns (per year observation) for each row (firm) to rows (816 firms * 20 years) in an Excel Sheet. It resulted in 16,320 observations of all GCC-listed firms, with each entry representing firm data for a specific year.

Appendix 1 presents a sample of the GCC dataset used in this study, providing an overview of the financial and non-financial variables collected from publicly listed firms across the region.

Table 4.1 outlines the active publicly listed companies across the GCC countries during 2003-2023. As of 2023, there are 816 firms listed, with Saudi Arabia leading the count at 314 firms, followed by the UAE with 150 firms.

Table (4.1): Number of Active Listed Firms Over 21-Year Period (LSEG, 2023)

Year Became Public	Country of Headquarters						Total
	Bahrain	Kuwait	Oman	Qatar	Saudi Arabia	UAE	
2003 and before	23	67	74	24	65	35	288
2004	-	7	3	2	2	7	21
2005	4	19	5	2	4	30	64
2006	3	13	-	3	11	8	38
2007	3	9	2	3	18	9	44
2008	1	10	3	3	16	5	38
2009	-	4	-	3	10	3	20
2010	2	3	2	1	6	-	14
2011	-	2	2	-	3	2	9
2012	-	-	1	-	7	2	10
2013	-	1	2	-	4	1	8
2014	1	1	4	1	6	3	16
2015	-	1	-	-	3	3	7
2016	-	-	1	1	5	-	7
2017	2	-	2	1	16	4	25
2018	1	1	3	1	12	2	20
2019	-	1	1	1	6	1	10
2020	-	3	1	-	5	4	13
2021	-	2	1	3	20	7	33
2022	-	1	1	-	48	14	64
2023	-	-	2	4	41	6	53
NULL	-	-	4	-	6	4	14
Total	40	145	114	53	314	150	816

Most of these firms went public before 2003, accounting for 288 firms in the dataset. Notable years for public listings include 2005, 2007, and 2022, which experienced significant increases in firm listings, particularly in Saudi Arabia and the UAE. However, there has been a slowdown in new public

offerings from 2015 to 2020, as indicated by the smaller number of firms entering the public market during these years. Besides, the dataset includes 14 firms with missing values (NULL) regarding the year they became public.

As shown in Table 4.2, this study utilized 23 financial and non-financial variables sourced from LSEG Workspace and WDI. These variables were either directly extracted from the LSEG database or computed using imported LSEG data. ROI was calculated using the financial metrics (Net Income after Tax) and (Total Investments), serving as the dependent variable for the research.

Table (4.2): Variables Sources

	Research Variables	Source
1	Return On Investment	Researcher's computation using LSEG data
2	Current Ratio	LSEG
3	Quick Ratio	LSEG
4	Cash Ratio	LSEG
5	Debt ratio	Researcher's computation using LSEG data
6	Total Assets to Common Equity	LSEG
7	Total Assets to Total Equity	LSEG
8	Total Debt Percentage of Total Equity	LSEG
9	Interest Coverage Ratio	LSEG
10	Asset Turnover	LSEG
11	Net Operating Asset Turnover	LSEG
12	Current Asset Turnover	LSEG
13	Fixed Asset Turnover	LSEG
14	Price-to-Sales Ratio (P/S Ratio)	Researcher's computation using LSEG data
15	Price-to-Book Ratio (P/B Ratio)	Researcher's computation using LSEG data
16	Price-to-Earnings Ratio (P/E Ratio)	Researcher's computation using LSEG data
17	GDP (constant 2015 US\$)	WDI
18	Inflation, consumer prices (annual %)	WDI
19	Size of Firm	Researcher's computation using LSEG data
20	Age of Firm	Researcher's computation using LSEG data
21	Industry Type	LSEG
22	Firm's country	LSEG
23	year	LSEG

Fig. 4.2 presents the distribution of firms across 11 industry categories, as defined by the TRBC economic sectors, for publicly listed companies in GCC countries in 2022 (LSEG, 2023). The Financials sector emerges as the

most dominant, comprising 227 listed firms, followed by Basic Materials with 101 firms, and Real Estate with 88 firms.

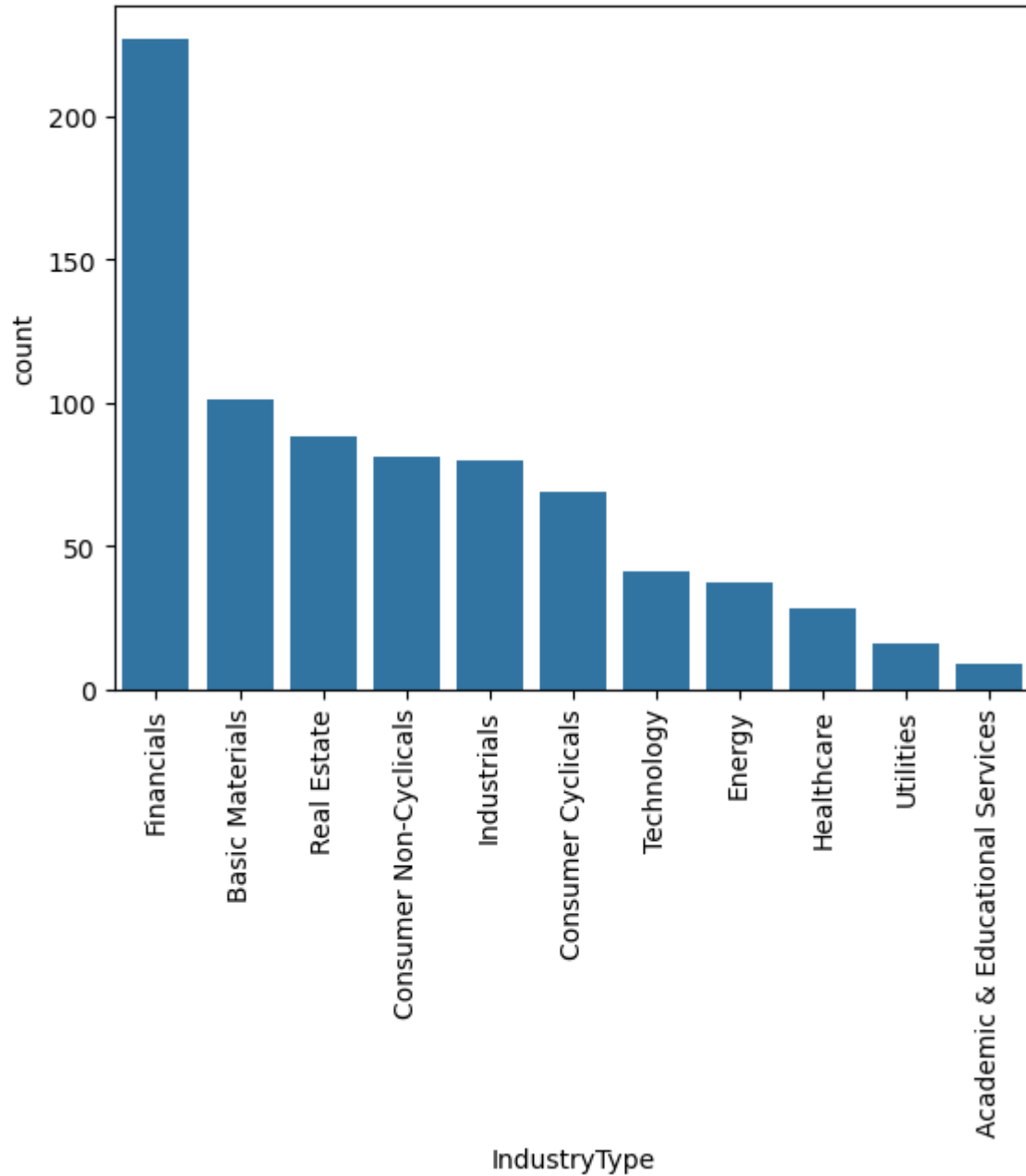


Fig. 4.2: The proportion of all GCC industry types within the dataset for FY 2022 (LSEG, 2023)

Table 4.3 presents the distribution of industry sectors among listed firms across GCC countries for the year 2022. The Financials sector demonstrates the highest level of representation across all member states, constituting 29.11% of the total sample. Within this sector, the UAE accounts for the largest number of firms (61), followed by Kuwait (51) and Saudi Arabia (46). In contrast, the Basic Materials sector is particularly concentrated in Saudi Arabia, which hosts 60 firms, representing 12.91% of the dataset. Additional sectors with notable representation include Real Estate, Consumer Cyclicals, and Consumer Non-Cyclical. Notably, the Government Activity and Institutions sectors are entirely unrepresented across all GCC stock exchanges, indicating a lack of publicly listed entities in these categories.

Table (4.3): The Proportion of Industry Types Within the Dataset for FY 2022

Sector	Bahrain	Kuwait	Oman	Qatar	Saudi Arabia	UAE	Proportion (%)
Academic & Educational Services	-	1	2	-	5	1	1.14
Basic Materials	1	10	16	4	60	11	12.91
Consumer Cyclicals	5	8	13	1	38	7	9.11
Consumer Non-Cyclicals	4	8	15	7	36	13	10.51
Energy		9	6	4	10	8	4.68
Financials	20	51	32	20	46	61	29.11
Government Activity	-	-	-	-	-	-	0.00
Healthcare	-	4	1	2	18	3	3.54
Industrials	5	13	10	3	36	15	10.38
Institutions, Associations & Organizations	-	-	-	-	-	-	0.00
Real Estate	3	34		6	34	13	11.39
Technology	2	6	2	4	18	9	5.19
Utilities	-	1	6	1	5	3	2.03
Total	40	145	103	52	306	144	

Note: The industry types are available for 790 (non-Nulls) out of 816 in the LSEG Workspace

4.5.2 Study Variables

In this study, several independent and control variables were deployed to predict the dependent variable, profitability measured by ROI, as presented in the study's conceptual framework in Chapter 2. Table 4.4. presents the definition of study's variables, some of which were adopted from prior research. In contrast, others were introduced as novel contributions to the literature, as discussed in the research gap section.

ROI is used to assess how efficiently net income is generated from investments. The independent variables include both firm-specific and macroeconomic factors. Firm-specific variables are categorized into five financial dimensions:

1. Liquidity, which is evaluated through CuR, QR, and CaR. These ratios measure a firm's ability to meet short-term liabilities with current assets, liquid assets, and cash equivalents, respectively.
2. Leverage is examined through the DR, TACE, TATE, and TDPTE ratios. These ratios reveal the degree of financial risk and the use of debt.
3. Solvency, assessed using the ICR, indicates the firm's ability to cover interest expenses and the return on invested equity.
4. Asset utilization, measured by AT, NOAT, CAT, and FAT. These ratios show how efficiently assets are used to generate sales.
5. Market value, assessed through the P/S, P/B, and P/E ratios, provides insights into the stock's market valuation relative to sales, book value, and earnings.

Table (4.4): Definition of Study's Variables

Financial Dimension	Variable	Acronym	Formula	Previous Studies
Dependent Variable				
V1	Profitability	Return On Investment	ROI	Net Income after Tax/ Total Investments * 100
Independent Variables				
<i>Firm-Specific Factors</i>				
V1	Liquidity	Current Ratio	CuR	Total Current Assets /Total Current Liabilities
V2		Quick Ratio	QR	(Total Current Assets-Total Inventories)/ Total Current Liabilities
V3		Cash Ratio	CaR	(Cash+Cash Equivalents)/Total Current Liabilities
V4	Leverage	Debt Ratio	DR	Total Debt/Total Assets
V5		Total Assets to Common Equity	TACE	Total Assets/Common Shareholders' Equity
V6		Total Assets to Total Equity	TATE	Total Assets/Total Shareholders' Equity
V7		Total Debt Percentage of Total Equity	TDPTE	(Total Debt/Total Shareholders' Equity) * 100
V8	Solvency	Interest Coverage Ratio	ICR	EBIT/Interest Expense
V9	Asset Utilization	Asset Turnover	AT	Total Revenue from Business Activities /Total Assets
V10		Net Operating Asset Turnover	NOAT	Total Revenue from Business Activities/Net Operating Assets
V11		Current Asset Turnover	CAT	Total Revenue from Business Activities/ Total Current Assets
V12		Fixed Asset Turnover	FAT	Total Revenue from Business Activities/ Net Fixed Assets
V13	Market Value	Price to Earnings Ratio	P/E	Stock Price / Earnings Per Share (EPS)
V14		Price to Sales Ratio	P/S	Market Capitalization/Revenue from Business Activities
V15		Price to Book Ratio	P/B	Stock Price/Book Value per Share
<i>Macroeconomics Factors</i>				
V16		GDP	GDP	Real GDP (constant 2015 billion US\$)
V17		Inflation Rate	IR	Inflation, consumer prices (annual %)
Control Variables				
V1		Size of Firm	Size	Log (Total Assets (US\$))
V2		Age of Firm	Age	2023 - Date Became Public
V3		Industry Type	IT	LSEG: TRBC Economic Sector
V4		Firm's Country	Country	Country of Firm Headquarters

The research employs macroeconomic variables, such as the real GDP (constant 2015 billion US\$) and the Inflation Rate (IR) in percentage, which provide insights into the overall macroeconomic conditions that affect GCC-listed firms' performance.

Control variables include firm Size (log of total assets), firm Age (years since going public), Industry Type (IT) according to Thomson Reuters Business Classification (TRBC) Economic Sectors by LSEG (Refinitiv, 2020), and the firm's country. These variables help adjust the effects of firm-specific variables that may influence the dependent variable outcomes.

Other variables, such as the observation year (2003 to 2022) and the firm's country of headquarters, help provide the chronological and geographical context for the observed variables.

4.5.3 DM Software

The DM models were implemented using Python Programming Language (version 3.11.7) within a Jupyter Notebook environment (version 7.0.8). The dataset was analyzed using Python's Pandas library to calculate descriptive statistics, while Matplotlib and Seaborn were employed to visualize trends and distributions. The scikit-learn (sklearn) library was utilized extensively for building, training, and evaluating DM models, including tasks such as train-test splitting, and hyperparameter tuning. The analysis and experiments were performed on a desktop computer with a 12th Gen Intel® Core™ i7-12700 CPU operating at 2.10 GHz and 16 GB of RAM. The system runs on a 64-bit Windows 10 operating system with an x64-based processor.

4.6 Data Preprocessing

Data preprocessing (cleaning) transforms raw data into a usable format (H. Son et al., 2019). Data preprocessing is often the most time-intensive phase in DM projects. This is due to the challenges of real-world data, which are frequently incomplete, noisy, or inconsistent, requiring significant effort to address issues like missing values, errors, outliers, and inconsistencies in codes or names (Sharda et al., 2021). Data preprocessing is crucial for accurately predicting firms' performance. It ensures model generalization by enhancing the quality of the model's variables and effectively managing outliers and missing values (H. Son et al., 2019).

One essential step before data preprocessing involves processing the imported Excel Sheet from LSEG by converting 20 Fiscal Years (FY) columns (e.g., FY0 (2022), FY-1 (2021), to FY-19 (2003)) into 'year' rows. This is achieved using Excel's INDEX function, which reshapes the 2D range (816 firms $\times$ 20 FYs) into a single column per variable, resulting in 16,320 rows for 23 variables.

After this essential step, data pre-processing techniques such as handling missing data, outliers, categorical attributes, and variable selection and normalization were applied to prepare the dataset. Once these steps are completed, the cleaned data will be ready to train the models. The following sections describe the specific preprocessing methods used in this study.

4.6.1 Handling Missing Values

The `info()` function provides an overview of the dataset, including variable names, the number of non-null values, and the data types (Table 4.5). The dataset comprises 16,320 observations (rows) and 22 variables (columns, features, or attributes). Notably, the CAT variable has only 7,820 non-null values, indicating that 8,500 observations are missing this variable.

Table (4.5): Nulls and Data Type for Research Variables

#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype
1	ROI	9370 non-null	float64	12	CAT	7820 non-null	float64
2	CuR	8289 non-null	float64	13	FAT	10496 non-null	float64
3	QR	7023 non-null	float64	14	P/E	10266 non-null	float64
4	CaR	8263 non-null	float64	15	P/S	10056 non-null	float64
5	DR	9482 non-null	float64	16	P/B	10371 non-null	float64
6	TACE	11566 non-null	float64	17	GDP	16320 non-null	float64
7	TATE	11566 non-null	float64	18	IR	16320 non-null	float64
8	TDPTE	9388 non-null	float64	19	Size	11673 non-null	float64
9	ICR	8650 non-null	float64	20	Age	16040 non-null	float64
10	AT	11088 non-null	float64	21	IT	15800 non-null	object
11	NOAT	10664 non-null	float64	22	Country	16320 non-null	object

Dtypes: indicates the data types; float64(20), object(2)

Missing values are typically handled through two methods: either deleting the entire row that contains the missing value for certain variables or imputing the missing values. Imputation involves filling in the missing values with the column's Mean, Median, or Mode. Imputation is generally applied when missing values are insignificant (P. V. H. Son & Duong, 2024). The choice of method depends on the nature of the data and the extent of the missing values. Proper imputation helps prevent biases that might arise from incomplete data and ensures that the statistical power of the dataset remains intact.

The study employed a two-step approach to address missing data:

1. First, Rows with missing values in significant variables, such as IT, Age, and Size, were removed using the `dropna()` function.
2. Then, the remaining missing values for other variables were imputed using the Median strategy `SimpleImputer(strategy='median')` function from the `scikit-learn` library. The Median strategy is deployed because the distribution of variables is skewed in the GCC dataset, with most values concentrated at the lower end and a few extreme values creating a long tail, as shown in the next chapter in the descriptive statistics results in Table 5.1 and the variable distribution chart in Fig. 5.1.

It is essential to note that the Mean strategy for imputation is employed when the data follows a normal distribution and the missing values are random. However, these conditions are not met in the dataset used in this thesis.

After addressing the missing values, the dataset is free of Null values. The next step in data preprocessing is to handle the outliers.

4.6.2 Handling Outliers

Outliers in financial datasets can distort analysis and lead to inaccurate conclusions, especially when extreme values do not reflect the overall dataset.

Effective outlier detection begins with examining the distribution of each variable, a process known as univariate analysis. This study adopts the Kurtosis/Skewness analysis and the Median and Interquartile Range (IQR)

Method (box-and-whisker plot), which was introduced by John Tukey in the 1970s. It employs the IQR to summarize data distribution (Ansari et al., 2024).

Various strategies have been proposed in the literature for handling outliers. Common approaches include trimming, where outliers are removed if their cause is known and justifiable; Winsorization, which caps extreme values to reduce their influence on statistical analyses; imputation, where outliers are replaced with estimated values such as the mean or predictions from regression models; and data transformation, which involves converting values into percentiles or using histogram-based methods to reshape the data (Ansari et al., 2024).

The Winsorization technique is applied at the 5th and 95th percentiles to address outliers in the current study. This method caps extreme values, especially in heavy-tailed distributions, at defined thresholds, minimizing their influence while preserving the overall dataset structure. Unlike the 1st/99th or 10th/90th percentiles, which may either under-correct or over-trim meaningful financial variations, the 5th/95th approach offers a balanced treatment of outliers. This threshold is consistent with Nguyen et al. (2023), who demonstrated its effectiveness in financial datasets. The `mstats.winsorize()` function is applied to all numerical columns in the dataset using the following algorithm:

```
from scipy.stats import mstats
df_winsorized = df.apply(lambda x: mstats.winsorize(x, limits=[0.05, 0.05]) if
x.dtype != 'object' else x)
```

After handling outliers, the next step is to address the text and categorical variables.

4.6.3 Handling Text and Categorical Attributes

Since most DM techniques require numerical values, encoding these categorical variables is essential (Sharda et al., 2021). Two categorical variables in the dataset, IT and Country, need to be handled as shown in Table 3.6. The function `pd.Categorical(x)` converts the text into a categorical type, and the `.codes` function assigns numerical codes to each category. This method converts categorical data into numerical values, assigning a unique integer to each category for DM. After applying the code, the output would be:

```
IndustryType Mapping (IndustryType Name -> Code):
{'Real Estate': 8, 'Basic Materials': 1, 'Consumer Non-Cyclicals': 3,
'Financials': 5, 'Consumer Cyclicals': 2, 'Industrials': 7, 'Energy': 4,
'Academic & Educational Services': 0, 'Utilities': 10, 'Healthcare': 6,
'Technology': 9}

Country Mapping (Country Name -> Code):
{'Kuwait': 1, 'Oman': 2, 'Bahrain': 0, 'Saudi Arabia': 4, 'United Arab
Emirates': 5, 'Qatar': 3}
```

4.6.4 Variable Selection

Variable (feature) selection is a crucial step in DM, as it reduces dimensionality, enhances model interpretability, and prevents overfitting (Sarkar et al., 2018). Including irrelevant or highly correlated variables can degrade model performance and lead to misleading conclusions.

This study employed Pearson Correlation Coefficients (PCC) to assess the strength and direction of linear relationships between variables. A correlation coefficient above 0.7 was considered indicative of a strong association (Chiadamrong & Wattanawarangkoon, 2023). PCC, with a

directional hypothesis (one-tail test), can examine the relationship between two variables, with a coefficient scale ranging from 1 (perfect positive correlation) to -1 (perfect negative correlation); the closer it is to 0, the weaker the relationship (Emma Bell et al., 2019).

High correlations among independent variables, known as multicollinearity, can increase variance, reduce interpretability, and lead to overfitting (Géron, 2017). Variance Inflation Factor (VIF) is one of the most commonly used diagnostic tools for detecting multicollinearity in regression models, as it quantifies the inflation in the variance of regression coefficients caused by multicollinearity (Gokmen et al., 2022). While this study did not apply VIF, it is discussed here for conceptual clarity.

Although VIF is widely used for diagnosing multicollinearity, PCC was selected in this study for its interpretability and direct focus on pairwise relationships. The `Pandas.corr()` function was used to compute PCC values, and Seaborn's `heatmap()` function visualized the correlation matrix.

It is important to distinguish between correlation and multicollinearity. Correlation measures the linear relationship between two variables, while multicollinearity refers to linear dependencies among multiple predictors. High pairwise correlation implies multicollinearity, but its absence does not rule it out. Therefore, correlation analysis serves as a preliminary diagnostic tool to identify potential issues of multicollinearity (Gokmen et al., 2022).

Barboza et al. (2017) maintained the correlated variables in the models since they tested only 11 variables related to American and Canadian firms. In contrast, Du Jardin (2021) discarded variables for which the p-values of the

test were higher than or equal to 0.05 and removed the most correlated variables when the correlation exceeded 0.6 among 30 financial ratios of 1.4 million French firms. In Arora and Saurabh's (2022) study, they dropped the highly correlated variables with a correlation coefficient exceeding 0.7 for 18 financial ratios of Indian firms.

Since the current study evaluates only 23 variables for 816 firms, it follows Arora and Saurabh's (2022) approach to handle only highly correlated variables. Variables with correlations of 0.7 or higher were removed from the dataset. Regardless of their statistical significance (p-values) in correlation analysis, all remaining variables were retained for modeling to allow the DM models to identify complex relationships and evaluate their predictive importance.

4.6.5 Variable Scaling (Normalization)

Variable (Feature) scaling is crucial for DM techniques, as they often perform poorly when input variables have different scales (Géron, 2017). Three common scaling methods are available in scikit-learn: *standardization*, *min-max scaling*, and *Robust Scaling (RS)*. The standard scaler adjusts values by subtracting the Mean and dividing by the SD. Min-max scaling transforms values to a 0-1 range but can be affected by outliers. In contrast, RS uses the IQR data scales without being impacted by outliers. The formula for RS is:

$$RS(X_i) = \frac{x_i - \text{median}(X)}{IQR_{(1,3)}(X)}$$

Each value of variable X is adjusted by subtracting X 's Median and then dividing the result by the IQR of X , which represents the difference between the first quartile (25th percentile) and the third quartile (75th percentile). In

this research, RobustScaler from `sklearn.preprocessing` is deployed. This ensures that the model focuses on data patterns, improving the reliability and generalizability of results in financial predictions.

4.7 Data Sampling and Splitting

4.7.1 Dataset Sampling

Sampling aims to create new subsets from a larger dataset (Barboza et al., 2017). This study involves constructing four models: three analyzing the entire GCC dataset using 16,320 observations: Imputation with Random Split-GCC (I-RS-GCC), Non-Imputation with Random Split-GCC (NI-RS-GCC), and the Chronological Split-GCC (CS-GCC), and one, by Country (C) model, focusing on individual GCC countries (Table 4.6). These models will be explained in detail in section 4.8.

The by-country (C) model was constructed by filtering the GCC dataset based on firms' headquarters. For example, the Bahrain dataset, containing 800 observations from 2003 to 2022, was extracted using the following code:

```
df_Bahrain = df[df['Country'] == 'Bahrain']
```

This process was repeated for the other GCC countries, resulting in datasets for each country, as shown in Table 4.6:

Table (4.6): Number of Observations (rows) in each GCC Country Dataset

Country	No. of Observations	Country	No. of Observations
Bahrain	800	Qatar	1,060
Kuwait	2,900	Saudi Arabia	6,280
Oman	2,280	UAE	3,000

4.7.2 Dataset Splitting

Data splitting (partitioning) is a fundamental step in developing the DM model, where a dataset is divided into two subsets: one for training the model and the other for testing its performance (Khalil et al., 2022). This splitting occurs after data preprocessing and is crucial for assessing the model's accuracy. There are two main ways to split the data:

1. Random Split (Classic Holdout): The data is shuffled randomly before being divided, which helps ensure that the subsets are unbiased and representative of the original dataset.
2. Chronological Split (Sequential Holdout): This method splits the data based on time or sequence, preserving the order of observations. It is significant for time-series data or datasets with temporal dependencies.

Both splitting methods are deployed in the research as described next.

4.7.2.1 *Random Split*

Data split into training and testing sets using an 80% to 20% ratio following Khalil et al. (2022), Nguyen et al. (2023), and Ainan et al. (2024), utilizing the `train_test_split()` function from the `sklearn` library was deployed in the research.

4.7.2.2 *Chronological Split*

Data split (partitioned) based on time sequences to evaluate the effect of prediction accuracy by adding more years to the training set as described in Table 4.7 for the all-GCC dataset. The data was split into training and testing sets using the prior year(s) and the post-year (2022). This approach tests the

effect of cumulative years to predict the firm's performance in the later year. The chronological data partitioning (split) is accomplished by employing the `loc()` function. The `loc()` function is a method provided by Pandas, a data manipulation library in Python. It is used to access a group of rows and columns by labels or a Boolean array. The split is based on the 'year' column within the dataset.

Table (4.7): Chronological Approach to Split Training and Testing Sets

Period	Total No. of Training Y	Training Y	Obs. No.	Testing Y	Obs. No.
1	1	2021	776	2022	777
2	2	2020-2021	1537	2022	777
3	3	2019-2021	2264	2022	777
4	4	2018-2021	2951	2022	777
5	19	2003-2021	10757	2022	777

Notes:

Total No. of Training Y: Total number of training years prior to prediction

Obs. No.: Number of observations (a specific observation in a year for a firm).

The data partitioning methodology for creating training and testing sets follows a chronological approach, with 2022 consistently used as the test set. The training sets are built cumulatively, progressively adding more years to evaluate historical data on predicting profitability (ROI) for 2022.

To assign the year 2022 as the test set, the following code is used:

```
test_set = df.loc[df['year'] == 2022]
```

The training set starts with only the previous year (2021):

```
train_set = df.loc[df['year'] == 2021]
```

This model tests the impact of using one prior year to predict the ROI for 2022.

For the second model, the training set includes two years (2020 and 2021):

```
train_set = df.loc[df['year'].between(2020, 2021)]
```

This setup assesses the effect of using two cumulative years of data to predict year 2022 profitability. The process is repeated for additional cumulative years, testing the impact of larger historical datasets on the model's accuracy as follows:

Three years of training data (2019-2021):

```
train_set = df.loc[df['year'].between(2019, 2021)]
```

Four years of training data (2018-2021):

```
train_set = df.loc[df['year'].between(2018, 2021)]
```

Nineteen years of training data (2003-2021):

```
train_set = df.loc[df['year'].between(2003, 2021)]
```

This approach allows the model to leverage increasing historical data, providing insights into how the training sets of prior years will affect the model's ability to predict future profitability.

4.8 Prediction Models

Seven DM techniques, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost, are utilized to build four predictive models to address the fourth thesis objective. These models incorporate diverse combinations of sampling methods, utilizing the entire GCC dataset as well as country-specific subsets for Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the UAE, along with data

preprocessing and data-splitting methods for training and testing the model. The four models are:

1. Imputation with Random Split- GCC (I-RS-GCC) Model
 - 7 models
2. Non-Imputation with Random Split- GCC (NI-RS-GCC) Model
 - 2 models
3. By Country (C) Model
 - 2 non-imputation models for each country
4. Chronological Split- GCC (CS-GCC) Model
 - 7 imputation models (I-CS-GCC)
 - 2 non-imputation models (NI-CS-GCC)

The goal is to evaluate each model's accuracy and effectiveness in predicting GCC firms' performance.

4.8.1 Hyperparameter Tuning

After splitting the dataset, the next step is hyperparameter tuning (or parameter optimization), a critical process in developing DM models. Usually, a DM technique uses learning parameters and hyperparameters. Learning parameters are learned during the training process, whereas a hyperparameter has a value set before the training begins. Different models require different hyperparameters, and tuning them is a crucial part of model selection (H. Son et al., 2019).

Hyperparameter tuning is essential for enhancing model performance, minimizing overfitting, and ensuring generalization to new data (Géron,

2017). Since hyperparameters are not learned during training, they must be optimized using randomized or grid search methods.

The current study deployed the randomized-search Cross-Validation (CV) method, which was implemented using the `RandomizedSearchCV` function from `scikit-learn`, a faster alternative than grid-search, as it evaluates random combinations of hyperparameters rather than all possible options. The research employs a 5-fold CV, splitting the dataset into five equal parts (folds). The model trains on four folds and validates on the fifth, rotating through all folds to ensure all data is used for training and validation. This process is repeated for each random hyperparameter combination, with performance averaged across folds. The best hyperparameters are selected based on cross-validated scores.

Appendix 2 provides detailed information on the optimal hyperparameters identified for each predictive model used in this study. Appendix 3(b) provides the Python code used for hyperparameter tuning and model evaluation. Using `RandomizedSearchCV` with 5-fold cross-validation, the code identifies the best parameters for each model and subsequently evaluates them using various performance metrics, as detailed in Section 4.9.

4.8.2 DM Techniques

As mentioned in Section 1.5, the fourth research objective is to implement single and ensemble DM regression techniques to develop performance prediction systems for GCC firms. In Chapter 2, Table 2.9 illustrates that many researchers have explored various individual and ensemble classification models. At the same time, fewer studies have examined

regression techniques specifically for building DM prediction models of firm performance. This section provides a summary of the DM techniques utilized in this study to address the fourth research objective of this thesis.

Seven supervised regression techniques, categorized into two types, single and ensemble regressor models, were trained and tested in the research.

Single Models, or standalone models, are individual algorithms that make predictions based on their internal structure. They are simpler and faster to train but may lack the robustness of ensemble methods. The single models are LR, ANN, RT, and kNN.

Ensemble methods combine the predictions of multiple models to improve accuracy and reduce overfitting. These methods can be categorized into two main techniques: bagging and boosting. The RF algorithm is a bagging or bootstrap aggregating technique that involves building multiple models in parallel using different subsets of the data. The predictions from these models are then aggregated to reduce variance. In contrast, XGBoost and CatBoost algorithms are boosting algorithms that build models sequentially. Each new model attempts to correct the errors made by the previous ones, which improves overall performance.

DM models, including LR, ANN, RT, kNN, and RF, were implemented using Scikit-Learn libraries. Advanced gradient boosting algorithms, such as XGBoost and CatBoost, were implemented using the XGBoost library (version 2.1.0) and the CatBoost library (version 1.2.7).

Table 4.8 summarizes the Python libraries and modeling classes used to construct supervised regression models specifically designed to predict the

profitability of GCC firms. Each model is presented alongside the corresponding library and function used in the implementation.

Table (4.8): Libraries Deployed for DM Models

DM Technique	Python Library	Modeling Classes
LR	sklearn.linear_model	LinearRegression
ANN	sklearn.neural_network	MLPRegressor
RT	sklearn.tree	DecisionTreeRegressor
kNN	sklearn.neighbors	KNeighborsRegressor
RF	sklearn.ensemble	RandomForestRegressor
XGBoost	xgboost	XGBRegressor
CatBoost	catboost	CatBoostRegressor

4.8.2.1 *Linear Regression*

LR is a basic statistical method used to analyze the relationship between a dependent variable Y and one or more independent variables (X). It assumes a linear form, represented by the equation:

$$Y = a + bX$$

Where a is the intercept and b is the slope.

The ordinary least squares method estimates the best-fitting line by minimizing the sum of squared errors between observed and predicted values (Sarkar et al., 2018). This study employed a multiple linear regression model to train and predict the dependent variable, ROI , of listed firms in GCC countries. The independent and control variables were deployed to predict profitability as follows:

$$\begin{aligned}
\text{Profitability}_{iy} = & \beta_0 + \beta_1 \text{Liquidity}_{iy} + \beta_2 \text{Leverage}_{iy} + \beta_3 \text{Solvency}_{iy} \\
& + \beta_4 \text{Asset Utilization}_{iy} + \beta_5 \text{Market Value}_{iy} + \beta_6 \text{GDP}_y + \beta_7 \text{IR}_y \\
& + \beta_8 \text{Size}_{iy} + \beta_9 \text{Age}_{iy} + \beta_{10} \text{IT}_{iy} + \beta_{11} \text{Country}_{iy} + \varepsilon_{iy}
\end{aligned}$$

Where:

β_0 is the intercept of the regression line,

$\beta_1, \beta_2, \dots, \beta_{11}$ are the coefficients that represent the degree of change in the ROI due to a change in the relevant independence and control variables for a specific firm i in a specific year y .

y represents the year of the corresponding observation between 2003 and 2022,

ε_{iy} represents the error term (residual).

The equation below shows how financial dimensions are expressed using financial ratios in the current study:

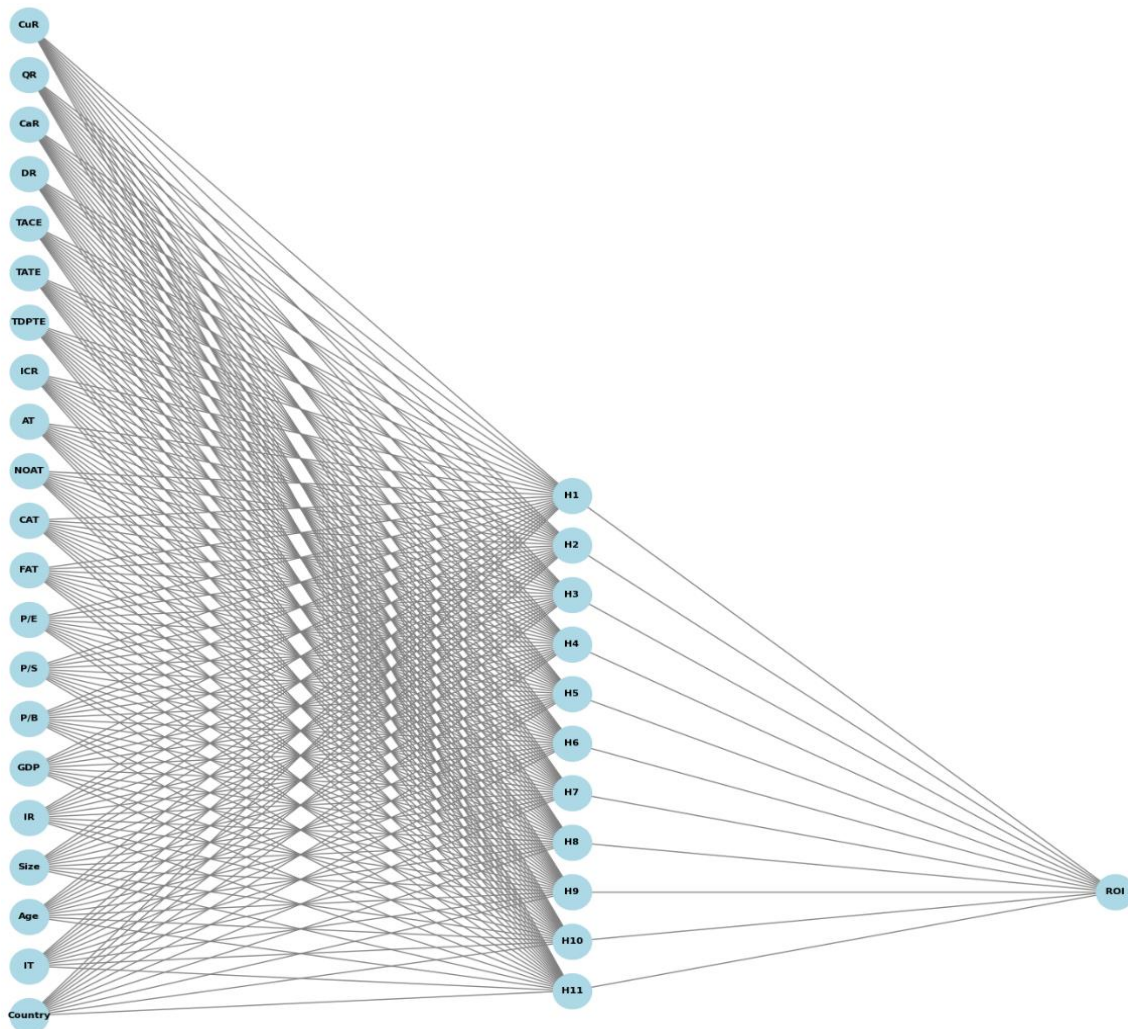
$$\begin{aligned} ROI_{iy} = & \beta_0 + (\beta_1 CuR_{iy} + \beta_2 QR_{iy} + \beta_3 CaR_{iy}) + (\beta_4 DR_{iy} + \beta_5 TACE_{iy} + \beta_6 TATE_{iy} \\ & + \beta_7 TDPTE_{iy}) + (\beta_8 ICR_{iy}) + (\beta_9 AT_{iy} + \beta_{10} NOAT_{iy} + \beta_{11} CAT_{iy} \\ & + \beta_{12} FAT_{iy}) + (\beta_{13} P/E_{iy} + \beta_{14} P/S_{iy} + \beta_{15} P/B_{iy}) + \beta_{16} GDP_y + \beta_{17} IR_y \\ & + \beta_{18} Size_{iy} + \beta_{19} Age_{iy} + \beta_{20} IT_{iy} + \beta_{21} Country_{iy} + \varepsilon_{iy} \end{aligned}$$

4.8.2.2 Artificial Neural Networks

Martín-del-Brío & Serrano-Cinca (1993) were the early ones who applied ANN in financial decision-making. They utilized self-organizing neural networks to examine the financial state of Spanish firms from 1990 to 1991 and to predict bankruptcy among Spanish banks during the 1977–1985 banking crisis (Martin-del-Brio & Serrano-Cinca, 1993).

ANNs are models inspired by the structure and function of the human brain, developed to perform various tasks. The foundational work on ANN started with McCulloch and Pitts in 1943, who introduced a simple binary artificial neuron model that mimicked some aspects of biological neurons.

Their model utilized a network of interconnected artificial neurons, laying the groundwork for modern neural networks. Over the decades, ANNs have evolved significantly and are widely used in DM applications. ANN research gained popularity from the late 1950s to the early 1960s and has been developed and utilized significantly ever since (Sharda et al., 2021). Fig. 4.3 shows the ANN model used in the current study to predict the financial performance of GCC-listed firms.



Note: The researcher visualizes the structure using Python, assuming a single hidden layer with 11 nodes for illustrative purposes. The actual number of hidden layers is determined during the hyperparameter tuning process. The code used to generate this figure is provided in Appendix 7.

Fig. 4.3: ANN Structure for the GCC Profitability Prediction Model

4.8.2.3 *Regression Trees*

RTs are part of the methodology of decision trees when the dependent variable is continuous and form the theoretical basis for the advancement of tree-based methods and ML. A regression tree is a piecewise constant or piecewise basically linear estimate of a regression function constructed by continuously partitioning the data and sample space (Saha et al., 2023).

In literature, most decision trees used the classification approach to predict firms' bankruptcy (e.g., du Jardin, 2021; Halteh et al., 2018; Smiti & Soui, 2020) rather than the regression approach (e.g., Saha et al., 2023).

A decision tree partitions the data space of all joint regressor values X into J - disjoint regions $(R_j)'$ (Saha et al., 2023). For a given set of joint regressor values X , the tree prediction $\hat{Y}_j = T_j(X)$ assigns as the response estimate, where T is defined as the original regression tree, the value assigned to each region containing X :

$$X \in R_j \Rightarrow T_j(X) = \hat{y}_j$$

A regression tree for predicting ROI builds a model by recursively dividing the dataset into smaller subsets based on the values of input variables. At each node, the tree chooses a split that minimizes the variance of ROI within the resulting subsets. The final output is the average ROI value for each leaf node.

To evaluate the quality of a split, the tree uses the Mean Squared Error (MSE) formula:

$$MSE = \frac{1}{N} \sum_{i=1}^N \left(ROI_i - \bar{ROI} \right)^2$$

Where:

ROI_i is the actual value for the i^{th} observation,

$\bar{ROI}$ is the mean value for the observations in the node,

N is the number of observations in the node.

4.8.2.4 *k-Nearest Neighbor*

DM algorithms often involve complex and computationally demanding processes. For instance, ANNs require extensive iterative calculations. In contrast, the kNN algorithm is much simpler but can still be effective for regression tasks. kNN is an instance-based or lazy learning method, meaning it does not build a model in advance but instead performs computations only at prediction time. For regression, kNN predicts a value by averaging the values of the k-nearest neighbors in the training set. This method is straightforward but highly dependent on the local data structure (Sharda et al., 2021).

According to Ledhem (2022), kNN can be used for regression when applied to a dataset consisting of pairs as $(x_i; y_i)$, presented as $(x_i; ROI_i)$ in the current study, where x_i represents numeric variables and ROI_i is the numeric value associated with each pair. This method allows for predicting a value $\hat{ROI}$ for a new instance vector x based on the information from kNN during the training process, which considers the simple cases of the dataset

defined by individual variables. Each data point in the dataset is characterized by the pair $(x_i; ROI_i)$, where x_i is a scalar value (Ledhem, 2022). Given a new case $(x = k; \hat{ROI} = ?)$, kNN will provide an output number $\hat{ROI}$ presenting in the following equation:

$$\hat{ROI} = \frac{1}{k} + \sum_{x_i \in N_k(x)} ROI_i$$

where: $N_k(x)$ is the set $\{x_i, \dots, x_k\}$ of neighbors from the training process of the new value x . Evidently, $N_k(x)$ hang on the selected value for k . kNN does not adopt any obvious form for the prediction role. Therefore, kNN regression is a nonparametric technique.

kNN has been deployed in many studies in the field of firm performance prediction (Aljawazneh et al., 2021; Garcia, 2022; Ledhem, 2022; Smiti & Soui, 2020).

4.8.2.5 *Random Forest*

RF, introduced by Leo Breiman in 2001, is an ensemble model for regression and classification that creates many DTs. In a regression context, the output is the average of all individual trees (Halteh et al., 2018). In regression, the method works by building multiple decision trees in the forest using randomly selected subsets during training. It outputs the average of the predictions of the individual trees to improve predictive accuracy and control over-fitting. A decision tree is trained on a bootstrapped sample of data, where for each node in the tree, only a subset of predictors (variables) is considered to split the node. This creates diverse trees that prevent overfitting (Breiman, 2001). *Bootstrapping* is a sampling technique where a fixed number of instances

from the original data is sampled (with replacement) for training, and the rest of the dataset is used for testing (Sharda et al., 2021)

RF classifiers and regressors have demonstrated superior performance compared to traditional models and other ensemble techniques in forecasting firm performance (Arora & Saurabh, 2022; Barboza et al., 2017; Garcia, 2022; J. P. Li et al., 2020; Ledhem, 2022). RF classifiers demonstrate their effectiveness in economics and finance (Nguyen et al., 2023).

Forests developed by Breiman rely on a collection of M randomized trees. The predicted value is assigned by (x, Θ_m, D_n) at a point x given by the m -th tree where $\Theta_1, \dots, \Theta_M$ are the independent random variables in the independent sample D_n . In the current study, $m_{M,n}(x, \Theta_1, \dots, \Theta_M, D_n)$ is the predicted ROI value from the n -th decision tree based on the randomly selected variables Θ of the study among these (liquidity, solvency, leverage, asset utilization, market value, GDP, IR, Size, Age, and Industry) where M is the total number of decision trees in the forest. The predictions of the M randomized trees are finally aggregated to produce a robust random forest prediction presented in the following equation:

$$m_{M,n}(x, \Theta_1, \dots, \Theta_M, D_n) = \frac{1}{M} + \sum_{m=1}^M m_n(x, \Theta_m, D_n)$$

RF is an effective technique for building decision trees that can handle noisy and complex data. Its ease of implementation and interpretability contribute to improved forecasting performance (Jabeur et al., 2021).

4.8.2.6 *eXtreme Gradient Boosting*

In 2016, Chen & Guestrin were the first to introduce XGBoost, an ensemble scalable tree model based on gradient boosting, which works by combining the predictions of multiple weak learners (decision trees) to form a stronger overall model. It focuses on reducing the residual errors of previous models by fitting new trees to correct the mistakes made by earlier ones (Chen & Guestrin, 2016). XGBoost has a built-in mechanism for handling missing values. It learns the best path for each missing value during training, allowing it to handle data with incomplete information more effectively than many other algorithms. Since the XGBoost model combines multiple decision trees, each tree corrects the residuals of the previous ones. The final prediction for i -th instance after M iterations (trees) is the following formula:

$$\hat{y}_i = F(x_i) = \sum_{m=1}^M f_m(x_i)$$

Where $f_m(x_i)$ is the prediction from the m -th tree, for instance, x_i , and each tree is learned to minimize the objective function.

The objective of the XGBoost *regularization* function Ω is to minimize errors while avoiding overfitting, which controls the complexity of the model. It does this by adding a *penalty* to large leaf weights (the output values of tree leaves), which helps prevent overfitting by encouraging simpler models, even if that means accepting a slightly less perfect fit.

XGBoost is more efficient and faster than traditional Gradient Boosting due to its regularized learning objective (Nguyen et al., 2023). The regularization term penalizes the complexity of each tree function in order to

avoid overfitting. The regularization term is included in the loss function as follows (Chen & Guestrin, 2016):

$$L(\theta) = \sum_i l(\hat{y}_i; y_i) + \sum_k \Omega(f_k) \quad \text{and}$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2$$

Where l is the loss function measuring the difference between the prediction and the real value, Ω is the regularization function calculated based on parameters T is the number of leaves in the tree, w is the weight of leaf, γ penalizes the number of leaves, and λ controls the second L regularization on leaf weights. The objective function becomes the original Gradient boosting if the regularization parameters are zero.

XGBoost has been widely adopted by researchers in the fields of finance and bankruptcy prediction (Arora et al., 2022; du Jardin, 2021; Garcia, 2022; Jabeur et al., 2021; Nguyen et al., 2023; P. V. H. Son & Duong, 2024).

4.8.2.7 Categorical Boosting

CatBoost, developed by Yandex in (Prokhorenkova et al., 2018), primarily focuses on gradient boosting using decision trees, with notable enhancements for managing categorical data. It provides high performance in both classification and regression tasks. CatBoost is distinguished from other gradient boosting techniques, such as XGBoost and LightGBM, in its effectiveness in handling categorical variables, making it especially valuable for datasets that contain many non-numeric attributes.

CatBoost stands out among gradient boosting algorithms due to several key features. It employs ordered boosting, a technique designed to prevent target leakage and enhance model reliability. Unlike many other algorithms, CatBoost performs well even on small datasets. Additionally, it natively supports categorical features, eliminating the need for extensive preprocessing by automatically converting categorical variables into numerical representations. Its versatility also makes it suitable for a wide range of data types and formats, contributing to its growing popularity in machine learning applications (Jabeur et al., 2021). This makes it suitable for application in the finance field (Khalil et al., 2022).

The estimated output described by Prokhorenkova et al. (2018) can be described as follows:

$$\hat{y} = \sum_{m=1}^M f_m(X)$$

Where $\hat{y}$ is the predicted value, M is the total number of trees in the model, and $f_m(X)$ is the output of the m -th tree for the input X .

4.9 DM Evaluation Metrics

This section introduces the evaluation metrics commonly used to assess the predictive performance of models on continuous variables, as highlighted by Khalil et al. (2022), Ledhem (2022), and Saha et al. (2023). Evaluating DM models on the GCC dataset is critical in developing robust prediction models. Beyond merely analyzing the predictive outcomes, the evaluation process determines the reliability and applicability of these models to real-world scenarios.

This research aims to measure the effectiveness of DM techniques in predicting the GCC-listed firms' performance represented in ROI. For this purpose, R-squared is employed with three 'Error-based' metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). R-squared is a performance metric used to evaluate the goodness-of-fit of the DM regression models, while error-based metrics measure the prediction errors of DM models. According to (Kumbure et al., 2022), RMSE was the most frequently used evaluation metric for measuring the forecasting performance of models among 138 articles of stock market prediction studies.

In the following subsections, an explanation of the evaluation metrics is provided.

4.9.1 R-squared

R-squared, or the coefficient of determination, indicates the proportion of the variance in the dependent variable that is predictable from the independent variables in the regression model (Sarkar et al., 2018). R-squared value ranged between 0 and 1. A value of 1 indicates that the independent variables can explain the variance in the dependent variable, while values closer to 0 suggest poor-fitting models (Sharda et al., 2021). The formula presented by (P. V. H. Son & Duong, 2024) is deployed to the thesis:

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(ROI_i - \hat{ROI}\right)^2}{\sum_{i=1}^n \left(ROI_i - \bar{ROI}\right)^2}$$

Where:

ROI_i is the actual ROI value for the i^{th} observation,

$\hat{ROI}_i$ is the predicted ROI value from the model for the i^{th} observation,

$\bar{ROI}$ is the mean of the actual ROI values,

n is the number of observations.

For the thesis, R-squared indicates the proportion of variance in ROI explained by the independent variables in the model. The high R-squared value signifies a better fit.

4.9.2 MAE

MAE measures the average absolute difference (errors) between predicted and actual values without considering their direction (Sarkar et al., 2018). It is the average absolute differences between actual values ROI_i and predicted values $\hat{ROI}_i$.

Since MAE uses absolute values, it gives an equal weight to all errors. Lower MAE values indicate better model performance. MAE is in the same units as the ROI variable. MAE is calculated as:

$$MAE = \frac{1}{n} \sum_{i=1}^n \left| ROI_i - \hat{ROI}_i \right|$$

4.9.3 MSE

MSE measures the average squared difference between the actual and predicted values of ROI . A lower MSE indicates that the model's predictions are closer to the actual values (Sarkar et al., 2018). MSE formula is given by:

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(ROI_i - \hat{ROI}_i \right)^2$$

Lower MSE values indicate better model performance. MSE is more sensitive to outliers because errors are squared. It is also in the square of the units of the ROI variable.

4.9.4 RMSE

The RMSE is calculated as the square root of the MSE. It is useful because it keeps the same units as the target variable, making it easier to interpret compared to MSE. RMSE is a good metric because it combines the advantages of both MAE and MSE. It penalizes larger errors like MSE, but it's also in the same units as the data like MAE. Lower RMSE values indicate better model performance. This metric is especially helpful when it is required to give more importance to significant errors. RMSE gives the magnitude of prediction errors in terms of the ROI, with smaller values indicating more accurate predictions (P. V. H. Son & Duong, 2024). RMSE is sensitive to large errors. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(ROI_i - \hat{ROI}_i \right)^2}$$

4.10 Variable Importance and Sensitivity Analysis Using SHAP

Model sensitivity measures how changes in input variables affect a model's output. It helps identify which inputs have the most significant influence, offering insights into the model's behavior, robustness to noise, and data quality (Borgonovo et al., 2024). In the current study, SHapley Additive exPlanations (SHAP) method is deployed to interpret and unlock the black box of complex DM models.

SHAP is a method developed by Lundberg & Lee (2017) to explain DM models. SHAP values assign an importance score to each variable, explaining its contribution to a specific model output. SHAP builds on Shapley values, initially developed in game theory, which determines the contribution of individual players (or variables in the context of DM) to the overall outcome (the model prediction).

In SHAP, each variable is considered a "player" in a game where the model's prediction is the value to be divided among the players. The SHAP value of a variable is the average contribution of that variable across all possible subsets of variables, ensuring a fair distribution of variable importance (Lundberg & Lee, 2017). The Shapley value for the variable i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (N - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Where:

N is the set of all variables,

S is a subset of variables not containing variable i ,

$f(S)$ is the model's prediction when only the variables in the subset S are included,

and $|S|$ is the size of subset S , and N is the total number of variables. The term $f(S \cup \{i\}) - f(S)$ represents the marginal contribution of variable i when it is added to the subset S . The weighting term $\frac{|S|!(N-|S|-1)!}{|N|!}$ ensures that the marginal contributions are fairly averaged over all possible subsets.

SHAP combines this approach with a linear explanation model where the prediction $f(x)$ for an instance x is approximated by the sum of SHAP values for all variables plus the model's baseline prediction:

$$f(x) = \phi_0 + \sum_i^M \phi_i$$

Here, ϕ_0 is the baseline value (the average prediction over all data points when no variables are included), and ϕ_i is the SHAP value for a variable i , showing how much a variable i contributed to deviating the prediction from the baseline. SHAP values have several desirable properties, such as local accuracy (the model's prediction is the sum of the SHAP values), consistency (if a model change increases a variable's marginal contribution, its SHAP value should improve), and additivity (allowing explanations to be added across multiple variables or instances).

SHAP provides two types of explanations: global and local. Global explanations provide insights into the overall model behavior, while local explanations focus on individual predictions.

Global explanations are deployed in this study; they give an overall view of how the model performs across the entire dataset. This involves understanding which variables are the most important in influencing predictions and how these variables interact with the model (Sarkar et al., 2018). SHAP accomplishes this by combining the SHAP values (variable contributions) across all observations in the dataset.

SHAP Variable Importance and SHAP Summary Plot are deployed for global model interpretation in the research. The SHAP Variable Importance plot visualizes the ranked significance of each variable in the DM model. Variables are ordered by their mean absolute SHAP values, indicating their contribution to the model's predictions. The SHAP Summary Plot provides an additional detailed visualization for interpreting the contributions of individual variables to a complex model's predictions. It shows the magnitude and direction of each variable's impact on individual predictions. It also highlights how different variable values (low vs. high) affect the model's output. Each dot on the plot represents a SHAP value for a single observation in the dataset, and each variable has its row. The x-axis shows the SHAP value (the impact on the model's prediction), while the color represents the variable's actual value, ranging from low (blue) to high (red). Variables are listed vertically based on their average impact on predictions, with the most influential variable appearing at the top.

Several studies within the firm performance DM literature have employed various explanatory methods to analyze complex firm performance prediction models. Notably, Nguyen et al. (2023) is the only study to use SHAP for interpreting the relationship between financial ratios and bankruptcy prediction. This research applies SHAP, implemented via the

SHAP library in Python, to the best-performing model trained using GCC firms' data, providing insights into how each input variable influences ROI.

Consequently, SHAP effectively addresses the third research objective by explaining the relationship between independent variables (e.g., liquidity, leverage, solvency, asset utilization, and market value) and firms' financial performance measured by profitability.

4.11 Summary

This chapter provided a comprehensive overview of the steps taken to establish the methodological foundations for the GCC firm's performance prediction model. It began by discussing the research paradigm, methodology, and DM framework. Additionally, it covers data collection, data preprocessing techniques, data sampling, and splitting. Later, it presented four DM prediction models built using seven DM techniques, single and ensemble, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost. These models varies based on different data preprocessing, sampling, and splitting methods.

In the next chapter, the analysis of datasets before and after data preprocessing is presented in detail showing the impact of cleaning the dataset on data distribution.

CHAPTER 5: DATA ANALYSIS RESULTS

5.1 Introduction

This chapter presents the analysis results of the dataset before and after data preprocessing. Section 5.2 introduces the GCC data before cleaning. In contrast, Section 5.3 provides the analysis after data cleaning for the four models: I-RS-GCC, NI-RS-GCC, C, and CS-GCC (addressing research objective two). Furthermore, Section 5.3 describes the construction of prediction models based on different data preprocessing, sampling, and splitting methods, addressing the fourth research objective.

5.2 Descriptive Statistics

Exploratory Data Analysis (EDA) is crucial for analyzing and understanding the patterns and characteristics of the data for the listed firms in the GCC countries, providing insights into the dataset and addressing the second research objective. This section addresses the second research objective by utilizing descriptive statistical tests to visualize and compare the financial performance of publicly listed companies in the GCC. The `describe()` function in Python provides a statistical overview of the dataset by displaying essential measures such as Count, Mean, Standard Deviation (SD), Minimum (Min), 25th percentile (25%), Median (50%), 75th percentile (75%), and Maximum (Max).

These statistics enable a foundational understanding of the GCC dataset's central tendency, dispersion, and range. Additionally, incorporating skewness and kurtosis through `skew()` and `kurtosis()` adds critical insights into the distribution's shape. Skewness measures asymmetry, indicating the presence of extreme values, while kurtosis captures the tailedness, showing the deviation from a normal distribution.

Table 5.1 summarizes the descriptive statistics of the study's variables before data preprocessing and modeling, highlighting the substantial variability and asymmetry in the dependent ROI variable. The Mean ROI is 400,924.13, with an exceptionally high SD of 18,782,548.38, reflecting significant variation among firms. A skewness of 48.80 indicates a highly right-skewed distribution, where most firms exhibit low ROI, while a few outliers inflate the Mean. The kurtosis of 2,399.91 highlights the presence of extreme outliers, which deviate significantly from a normal distribution. The descriptive statistics of other variables reveal significant variability and distributional asymmetries. For example, CuR and QR have high skewness values, indicating a right-skewed distribution driven by a few firms with exceptionally high values. DR shows moderate Mean and SD values but exhibits a heavy right tail with high kurtosis. Metrics like TATE and P/E Ratio display substantial dispersion and skewness, reflecting outliers with extreme financial characteristics among firms.

The macroeconomic variables, such as GDP and IR, display moderate skewness and kurtosis, reflecting relatively symmetric distributions. GDP has a high mean and standard deviation, indicating considerable economic differences across countries, while IR is relatively stable with lower variability.

Table (5.1): Descriptive Statistics of Numerical Variables Before Data Preprocessing

Variable	Count	Mean	SD	Min	25%	50%	75%	Max	Skew	Kurtosis
ROI	9370	400924.13	18782548.38	-15290139.19	2.80	12.49	58.01	1012222133.33	48.80	2399.91
CuR	8289	4.31	39.96	0	1.02	1.57	2.75	2641.45	45.91	2607.13
QR	7023	1.80	15.41	-1239.56	0.68	1.14	2.00	152.41	-74.12	6003.45
CaR	8263	1.29	17.08	-3.85	0.09	0.26	0.74	1176	56.90	3588.85
DR	9482	0.25	0.52	0.00	0.06	0.18	0.37	40.11	54.03	3902.53
TACE	11566	5.53	218.98	1.00	1.36	1.92	3.31	23483.33	106.61	11428.72
TATE	11566	5.13	218.55	1.00	1.33	1.86	3.14	23483.33	107.23	11520.27
TDPTE	9388	2.90	197.02	0	0.11	0.36	0.87	19086.17	96.83	9379.93
ICR	8650	63.36	3116.54	-132990.56	0.56	3.12	12.24	227871.35	37.99	3734.94
AT	11088	0.46	3.76	-0.32	0.07	0.25	0.57	392.04	101.82	10600.61
NOAT	10664	1.53	16.06	-9.84	0.17	0.48	1.11	1239.68	55.13	3718.02
CAT	7820	1.64	11.47	-1.44	0.67	1.22	1.88	1007.90	86.33	7571.68
FAT	10496	207.22	4660.20	-995.71	0.76	2.47	8.05	143932	27.32	763.71
P/E	10266	34.38	692.37	-9117.33	5.35	11.42	19.72	50172.30	45.84	2958.43
P/S	10056	2207164.03	95546374.87	-1495.25	1.25	3.18	7.59	6926533706	55.51	3442.88
P/B	10371	2.33	84.50	-3091.94	0.78	1.29	2.17	7486.34	60.92	6174.59
GDP (bn)	16320	342.42	280.98	11.07	105.97	257.92	528.21	1108.57	0.79	-0.47
IR	16320	2.91	3.02	-4.86	1.05	2.47	4.04	15.05	1.25	2.51
Size	11673	19.77	2.02	10.76	18.51	19.61	20.81	27.22	0.36	0.42
Age	16040	16.05	10.20	0.07	6.29	17.24	24.14	46.99	0.07	-0.65

Among the control variables, Size shows low skewness and kurtosis, indicating a relatively normal distribution. At the same time, Age has minimal skewness and negative kurtosis, suggesting a flatter, more uniform distribution, with firms varying in their years of establishment but without extreme outliers.

Fig. 5.1 presents a comprehensive visualization of the descriptive statistics for all GCC variables, generated using Seaborn's `histplot()` function to effectively display frequency distributions. Statistical measures, including Mean, SD, Skewness, and Kurtosis, are presented in the figure, revealing asymmetries and heavy tails across several variables, which are indicative of outliers and non-normal distributions.

Additional detailed visualizations for each GCC country's descriptive statistics are presented in Appendix 4, revealing asymmetries and heavy tails across several variables in each GCC dataset, which are indicative of outliers and non-normal distributions.

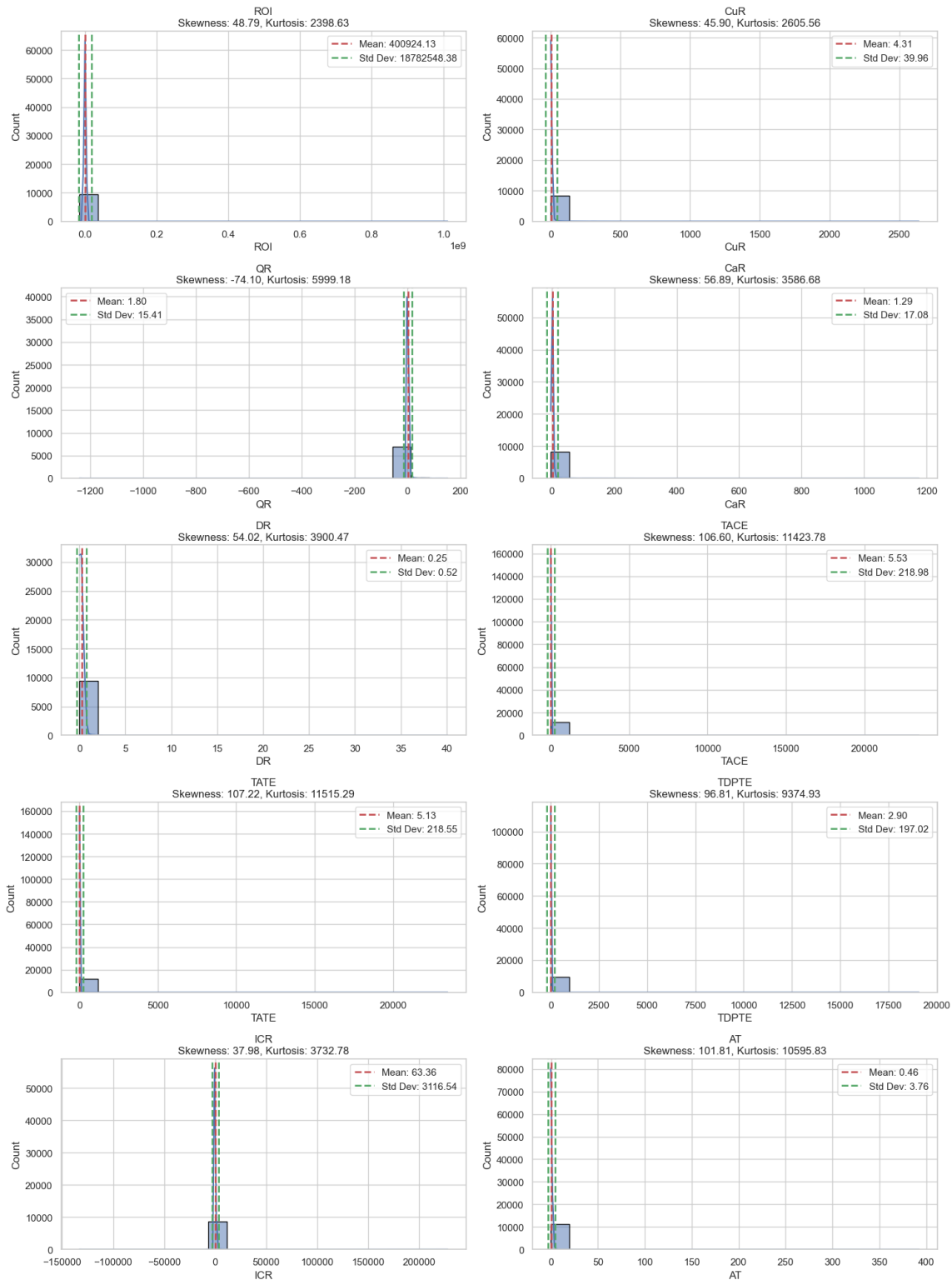


Fig. 5.1: Variables Distribution for GCC Dataset

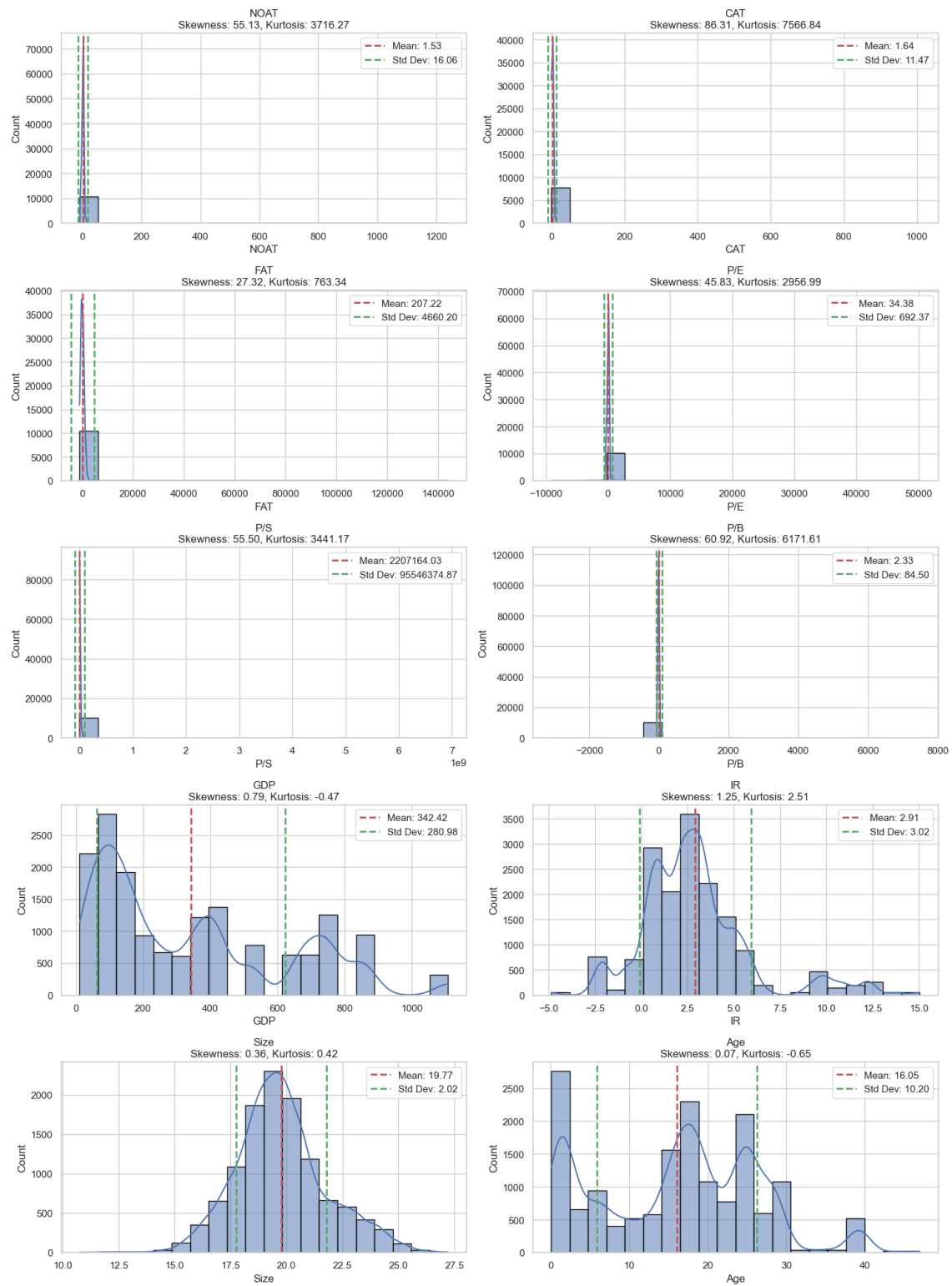


Fig. 5.1 (cont.): Variables Distribution for GCC Dataset

Fig. 5.2 presents the Pearson Correlation Coefficients and their corresponding p-values for the variables in the all-GCC dataset.

The correlation matrices for the country-specific (C) models are provided in Appendix 5.

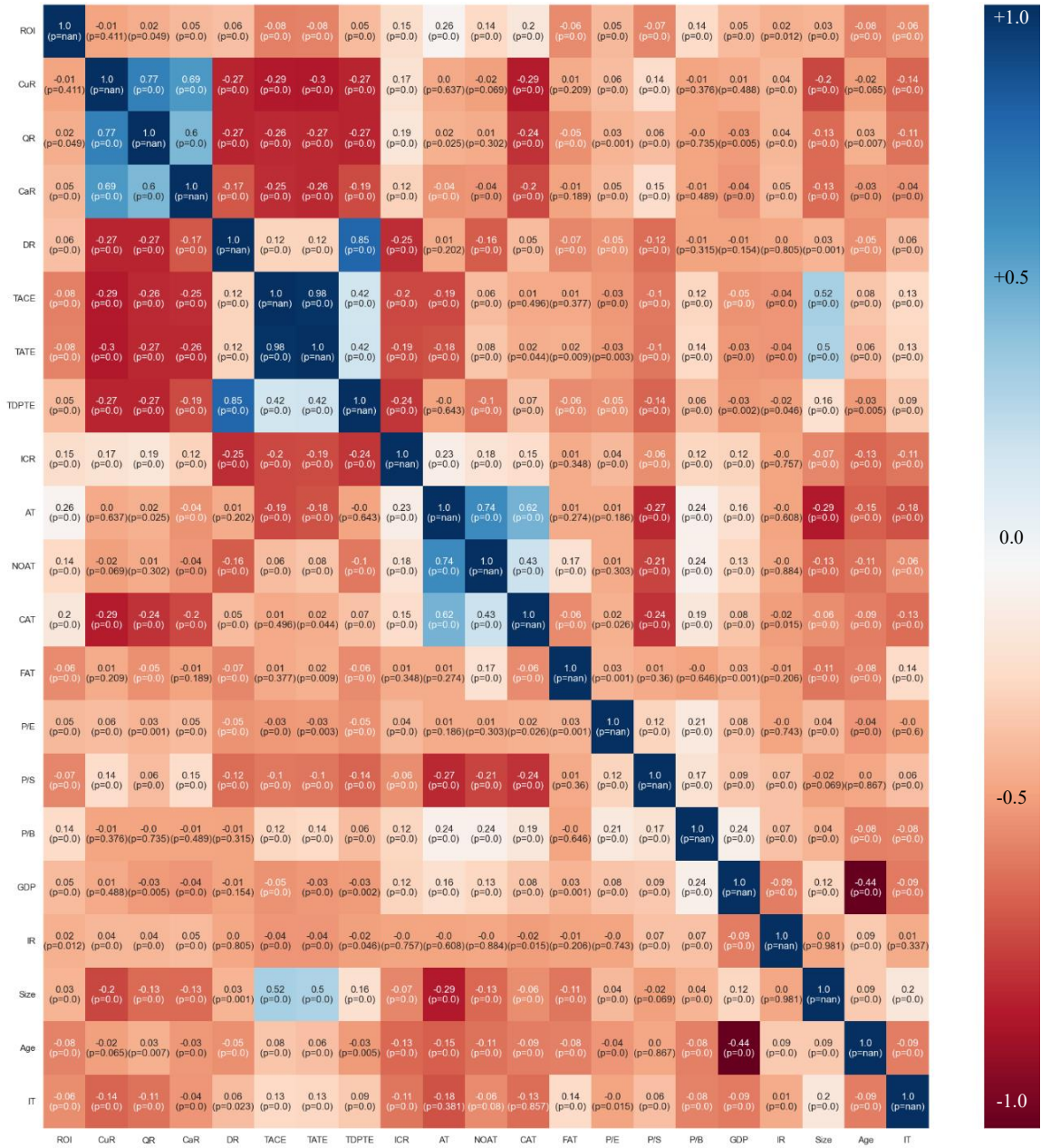


Fig. 5.2: Pearson Correlation Matrix with P-Values for all GCC dataset variables

As mentioned earlier in Section 4.6.4, the current study evaluates only 23 variables for 816 firms; therefore, it will follow Arora & Saurabh's (2022) approach to handle only highly correlated variables, noting that only 21 are analyzed in the Correlation Matrix, excluding Country and year. Variables with correlations equal to or exceeding 0.7 will be removed from the dataset, regardless of their statistical significance (p-values) in correlation analysis. Appendix 5 provides Pearson correlation matrices for each GCC country, noting that the correlation values vary depending on the dataset.

Table 5.2 illustrates the highly correlated variables ($r > 0.7$) and their corresponding dimensions, indicating that they are grouped within the same financial dimension.

Table (5.2): Correlated Variables ≥ 0.7 for all-GCC Dataset

Variable 1	Variable 2	Financial dimension	Correlation degree
1 QR	CuR (drop)	Liquidity	0.77
2 TATE (drop)	TACE	Leverage	0.98
3 TDPTE (drop)	DR	Leverage	0.85
4 NOAT (drop)	AT	Asset Utilization	0.74

The main finding from the above table is that ROI is not correlated with any independent variables. Second, QR and CuR are highly correlated (0.77). Since QR better reflects immediate liquidity, CuR will be dropped to reduce multicollinearity. Third, TACE and TATE are highly correlated (0.89) and represent the firm's leverage. TACE offers a more focused view of leverage risk for common shareholders; in contrast, TATE gives a more complete picture of the firm's use of equity in general; then, TACE will be dropped to reduce multicollinearity in the dataset. Fourth, TDPTE and DR are highly

correlated (0.85). The DR will be maintained as it is a more recognized measure of leverage. Fifth and finally, NOAT and AT correlate (0.74). NOAT will be dropped as AT provides a broader measure of asset utilization.

5.3 Data Preprocessing Results

This section analyzes the dataset after data cleaning, and it details the construction of prediction models using various preprocessing, sampling, and splitting methods.

5.3.1 I-RS-GCC Model

This model preprocessed all GCC datasets by deleting the entire observation (row) for missing values in control variables (IT, Age, Size, Country) and imputing Nulls for the other variables, resulting in 11,534 observations (2003–2022). Outliers were winsorized at the 5th and 95th percentiles, and categorical variables were encoded numerically.

Variables with correlations above 0.7 were removed. The dataset was randomly split into 80% for training and 20% for testing. Seven techniques were evaluated: LR, ANN, RT, kNN, RF, XGBoost, and CatBoost (Section 6.2).

Despite winsorization, variables like ROI and other financial ratios remained right-skewed as shown in Fig. 5.3 and Appendix 6.

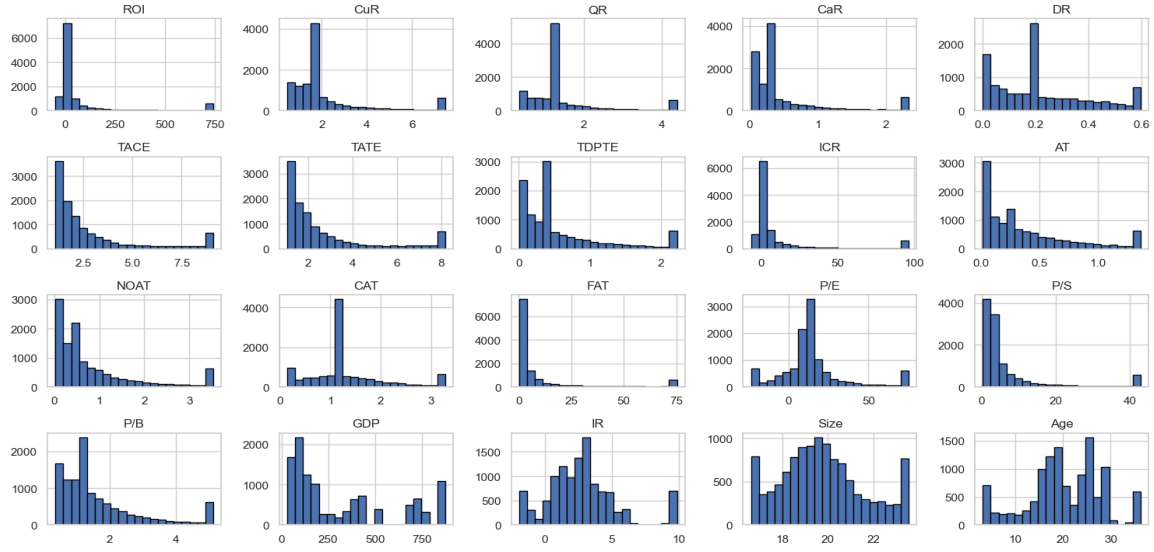


Fig. 5.3: Distribution of I-RS-GCC Model after Dataset Cleaning

Figures 5.4 and 5.5 show boxplots of ROI values before and after Winsorization. In Fig. 5.4, extreme outliers reach up to $1e9$, heavily skewing the data. After applying Winsorization, as shown in Fig. 5.5, extreme values are capped at the 5th and 95th percentiles, reducing the influence of outliers. The maximum ROI is capped at 750, thereby reducing the impact of outliers and improving data balance for more reliable modeling.

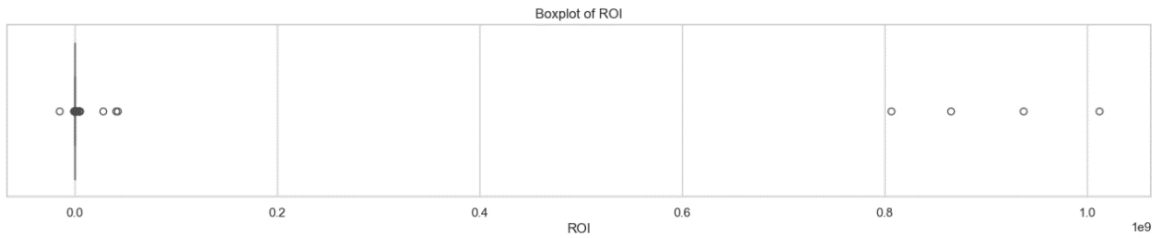


Fig. 5.4: ROI Distribution for All GCC Before Winsorization

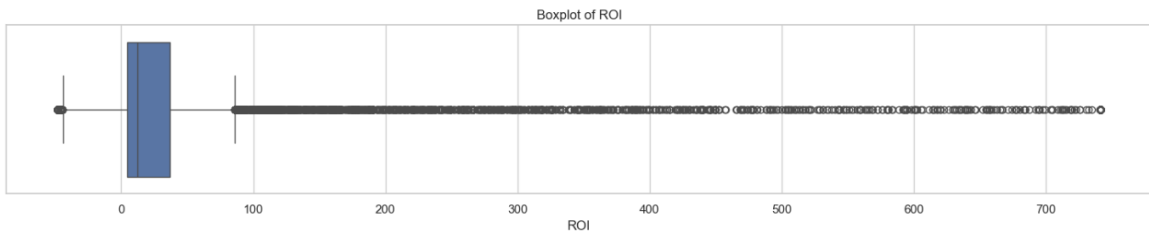


Fig. 5.5: ROI Distribution for All GCC After Winsorization

5.3.2 NI-RS-GCC Model

The non-imputation method in handling the missing data ensures that the dataset remains unaltered by preserving the integrity of the financial data. This model preprocesses the GCC firm datasets, handling missing values by deleting the entire row (observation) with Null in the dependent variable ROI and keeping other missing values for other independent variables.

The non-imputation model was built exclusively with XGBoost and CatBoost, utilizing the XGBRegressor and CatBoostRegressor classes, due to their inherent ability to handle missing values during both training and testing. Noting that XGBoost and CatBoost can only handle missing values in the input variables, but cannot do so in the output variable. As a result, the dataset was reduced to 9,370 observations with complete ROI values, while 25,427 missing values remained in other variables (columns), distributed as shown in Table 5.3.

Table (5.3): Remaining Nulls in Variables after Handling Missing Values in ROI

#	Variable	No. of Null Values	#	Variable	No. of Null Values
1	ROI	0	12	CAT	3500
2	CuR	3268	13	FAT	879
3	QR	4329	14	P/E	771
4	CaR	3296	15	P/S	951
5	DR	1923	16	P/B	776
6	TACE	41	17	GDP	0
7	TATE	41	18	IR	0
8	TDPTE	1959	19	Size	0
9	ICR	2545	20	Age	0
10	AT	346	21	IT	61
11	NOAT	741	22	Country	0

Outliers in the dataset were addressed using the Winsorization method at the 5th and 95th percentiles. It is important to note that the distribution of variables for this model, NI-RS-GCC, differs from the imputed model I-RS-GCC, because the missing values remain Null without changing the shape of the original GCC dataset (Fig. 5.6) and are detailed for ROI in Fig. 5.7. Note that the outliers range increased in the NI-RS-GCC Models compared to the I-RS-GCC Models after the dataset preprocessing stage.

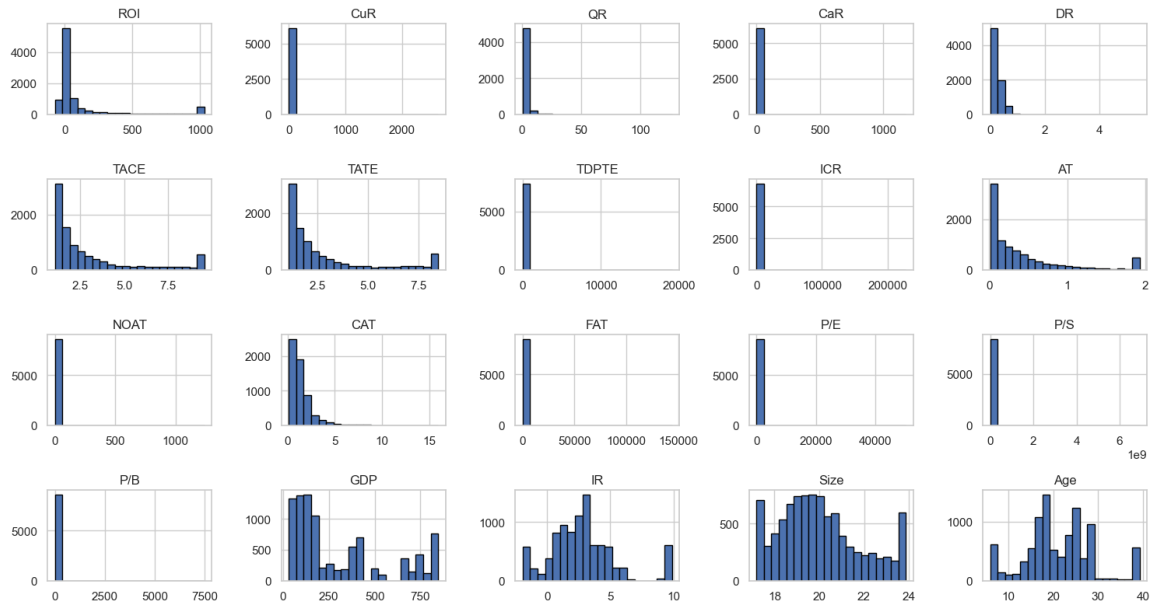


Fig. 5.6: Distribution of NI-RS-GCC Models after Dataset Cleaning

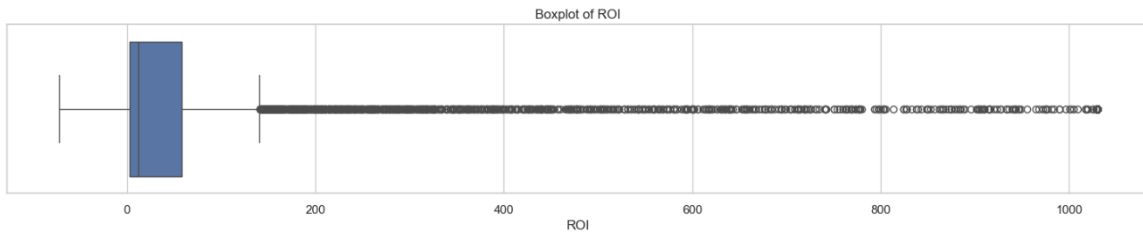


Fig. 5.7: Boxplot of ROI Distribution for NI-RS-GCC Model

The dataset for this model was split into training and testing sets using an 80/20 ratio, and the ensemble models, XGBoost and CatBoost, were evaluated in detail in Section 6.3.

5.3.3 By Country (C) Model

The by-country modeling approach utilizes the non-imputation method for data preprocessing. It employs only ensemble DM techniques, XGBoost and CatBoost. Missing values in the dependent variable ROI are addressed by deleting entire observations (rows) containing null ROI values. Outliers were winsorized at the 5th and 95th percentiles. Two ensemble models, XGBoost and CatBoost, were evaluated for each country (Section 6.4). Fig. 5.8 shows the histogram of variable distributions for Bahraini firms after data preprocessing. Most variables, including ROI, exhibit right-skewed distributions, with values concentrated at the lower range. GDP displays a relatively uniform distribution, while the IR includes negative values, reflecting deflation in specific years.

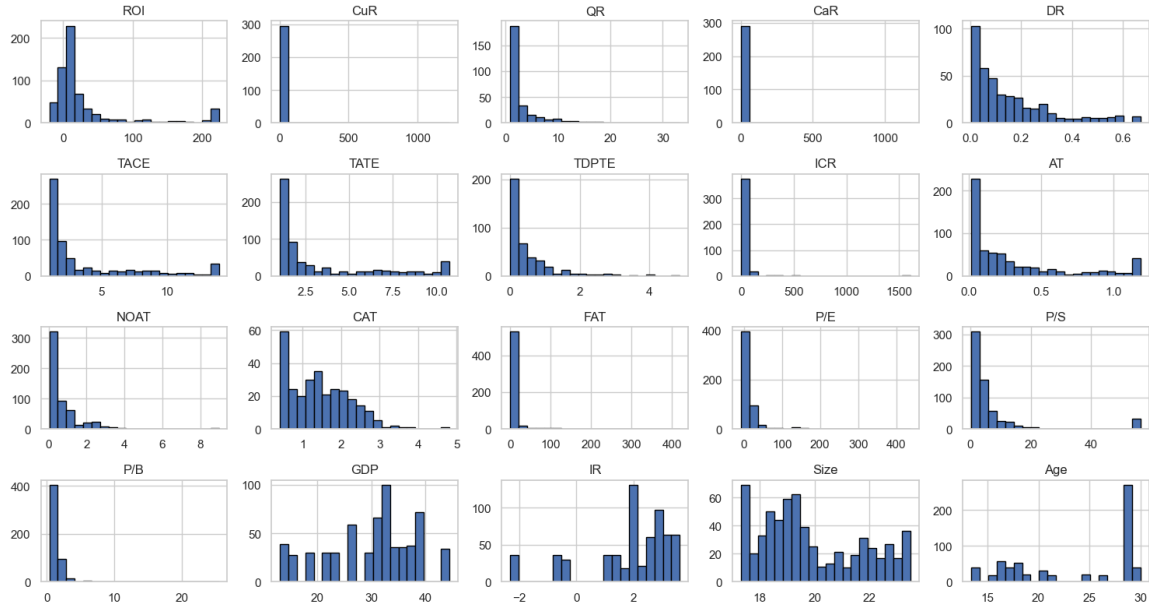


Fig. 5.8: Variables Distribution for Bahrain After Dataset Cleaning

Fig. 5.9 illustrates the distribution of Kuwaiti variables after preprocessing. Most variables exhibit right-skewed distributions, with a concentration of data points at lower values and a tail extending toward higher values. GDP and Size show relatively less skewness, indicating more symmetric distributions.

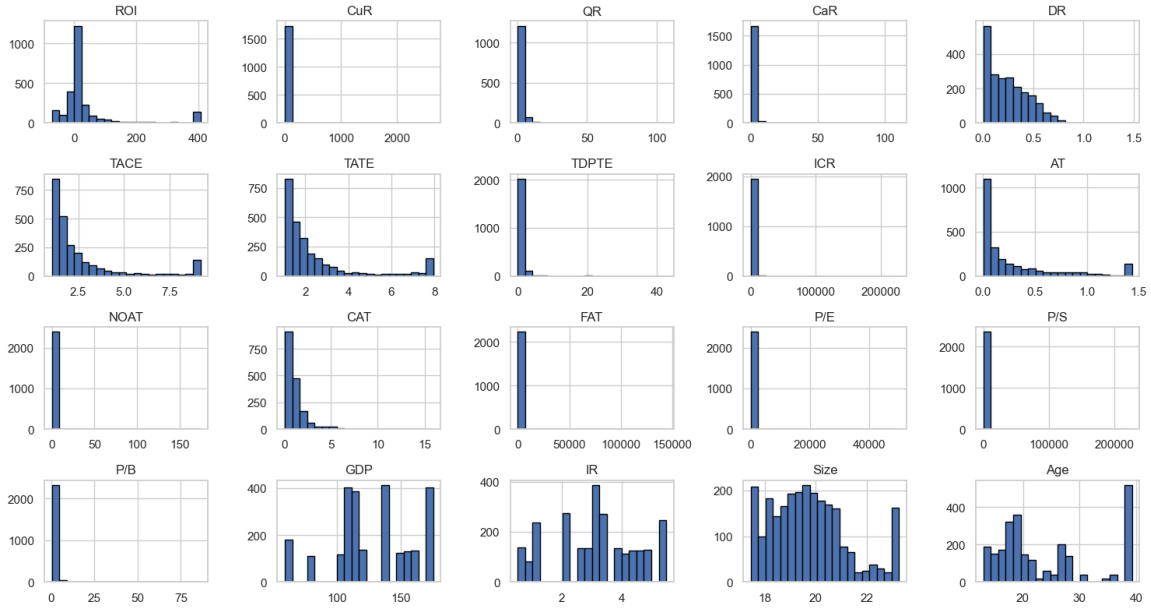


Fig. 5.9: Variables Distribution for Kuwait After Dataset Cleaning

Figures 5.10, 5.11, 5.12, and 5.13 present variable distributions for Oman, Qatar, Saudi Arabia, and the UAE.

Most variables in other GCC countries exhibit right-skewed distributions even after applying Winsorization, with a concentration of data points at lower values and a tail extending toward higher values.

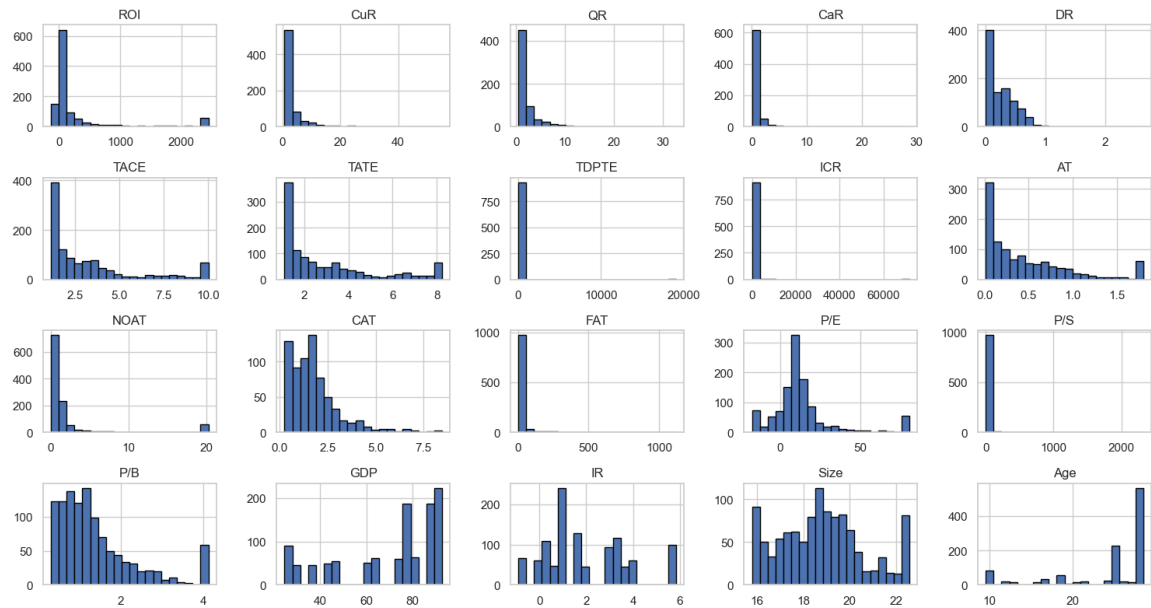


Fig. 5.10: Variables Distribution for Oman After Dataset Cleaning

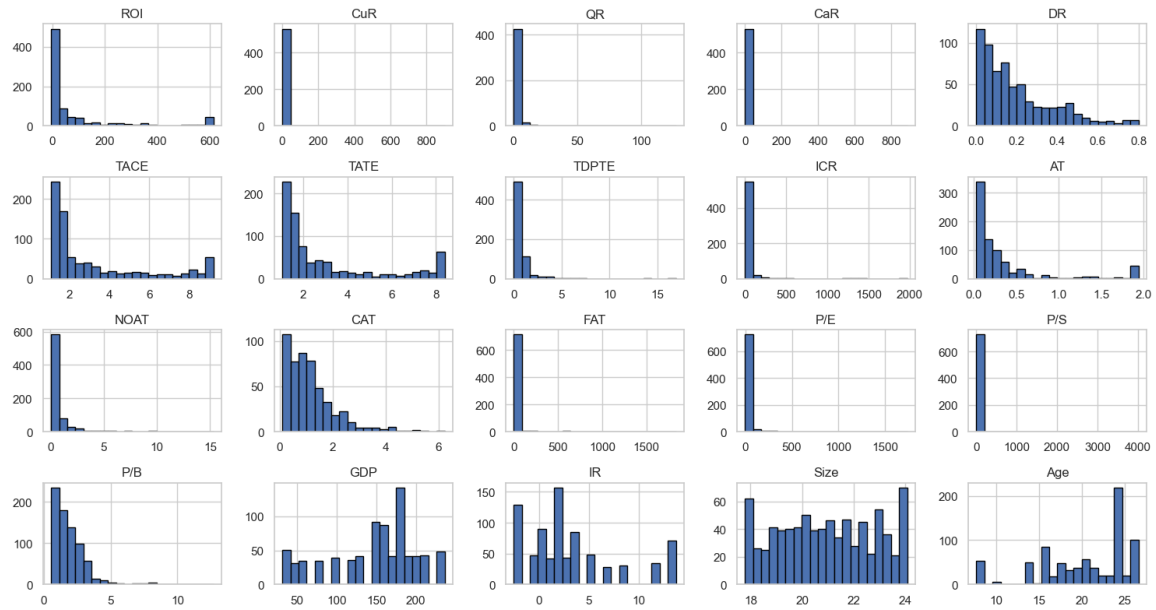


Fig. 5.11: Variables Distribution for Qatar After Dataset Cleaning

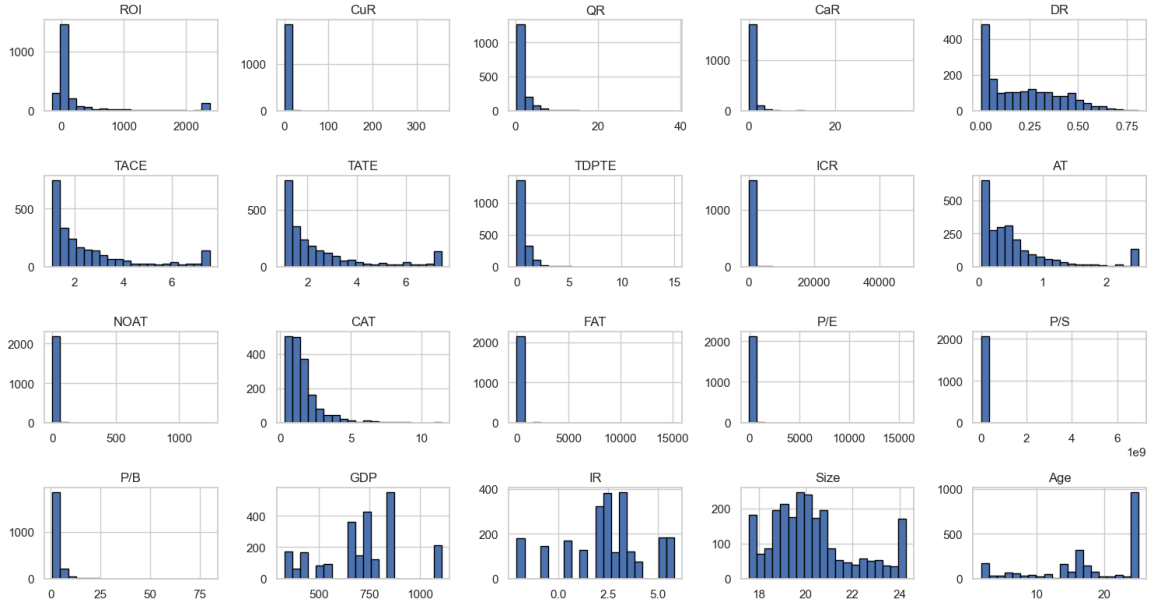


Fig. 5.12: Variables Distribution for Saudi Arabia After Dataset Cleaning

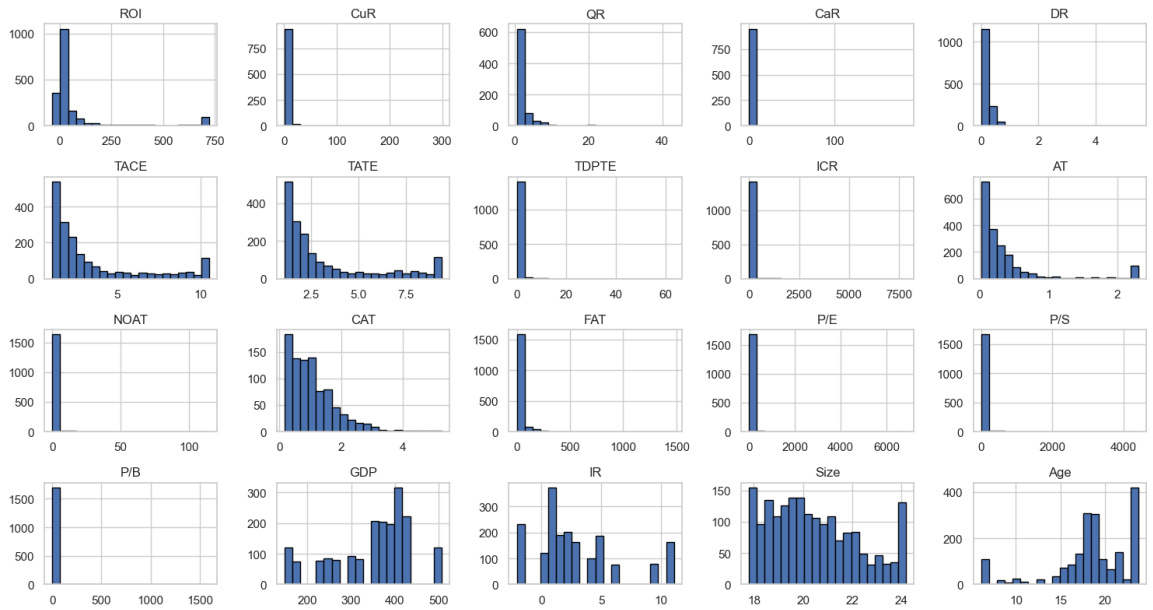


Fig. 5.13: Variables Distribution for UAE After Dataset Cleaning

5.3.4 CS-GCC Model

The chronological model is developed to evaluate the prediction accuracy across five periods (as described in Section 4.6.6), with imputation (I-CS-GCC) and without imputation (NI-CS-GCC). The I-CS-GCC data analysis results are similar to those of I-RS-GCC, while NI-CS-GCC is similar to NI-RS-GCC. Only the data split method for training and testing the models differs. Therefore, the results will not be presented in this section.

Outliers were winsorized at the 5th and 95th percentiles. Categorical variables were encoded as numerical values, and variables with correlations greater than 0.7 were removed. The data split method for creating training and testing sets follows a chronological approach, with 2022 consistently used as the test set. The training sets are built cumulatively, constantly adding more years to train historical data on predicting profitability ROI for 2022, as described in section 4.6.6. Seven techniques were evaluated: LR, ANN, RT, kNN, RF, XGBoost, and CatBoost (Section 6.5).

5.4 Summary

This chapter presented the analysis results of the GCC dataset before and after data preprocessing for four predictive models (I-RS-GCC, NI-RS-GCC, C, and CS-GCC). Additionally, four models were constructed based on data preprocessing, sampling, and splitting methods.

The subsequent chapter will present the results obtained from these models, providing insights into their performance across various evaluation metrics and identifying their most influential variables.

CHAPTER 6: PREDICTION MODELS

RESULTS AND DISCUSSION

6.1 Introduction

This chapter presents the empirical results of the study, addressing two key research objectives. First, it evaluates the predictive accuracy of the developed models, fulfilling the fifth research objective. Second, it explores the relative importance of independent variables using global model interpretability techniques, SHAP, in alignment with research objective three.

The chapter includes a comprehensive comparison and discussion of four predictive models, each employing distinct strategies for handling missing data, sampling techniques, and train-test splitting methods. This comparative analysis strengthens the conceptual link between Shapley values and sensitivity analysis, facilitating the extraction of both modeling insights and practical managerial implications. The four predictive models implemented in this study are as follows:

1. I-RS-GCC Model: Based on imputed missing data with random splitting applied to all GCC dataset.
2. NI-RS-GCC Model: Utilizes non-imputed data with random splitting applied to all GCC dataset.
3. C Model: Comprises country-specific models developed for each GCC member state.

4. CS-GCC Model: Applies chronological data split, assessed both with and without imputation, and applied to all GCC dataset.

6.2 I-RS-GCC Model Experimental Results

6.2.1 Evaluation of Model Performance

Table 6.1 presents the performance metrics for the models employed under the random sampling with the data imputation approach applied to all GCC firms. The optimal tuning hyperparameters for models are presented in Appendix 6.

Table (6.1): Performance Results of I-RS-GCC Models

Model	R-squared	MAE	MSE	RMSE
LR	0.101	98.650	29002.932	170.302
ANN	0.223	92.050	25084.283	158.380
RT	0.076	84.328	29829.823	172.713
kNN	0.395	69.426	19541.578	139.791
RF	0.419	75.453	18745.187	136.913
XGBoost	0.413	75.444	18932.636	137.596
CatBoost	0.482	69.886	16720.209	129.307

The table presents the key evaluation metrics for each model, including R-squared, MAE, MSE, and RMSE. The CatBoost model demonstrates the best performance, achieving the highest R-squared value of 0.482, indicating a more robust fit than other models; this implies that 48.2% of the variance in the dependent variable ROI is explained by the model, meaning the CatBoost does a better job than other models in capturing the relationship between the independent variables and ROI. CatBoost also has the lowest MSE (16720.209) and RMSE (129.307), suggesting greater accuracy in predicting firm performance. Similarly, RF and XGBoost are closely followed, with R-

squared values of 0.419 and 0.413, respectively, showing they also perform well in this context. In contrast, the LR model has the lowest R-squared (0.101) and the highest MSE (29002.932), indicating weaker predictive power. These results underscore the superior performance of ensemble models, particularly CatBoost, in predicting firm performance for the given dataset. Fig. 6.1 shows the comparison results for the I-RS-GCC model.

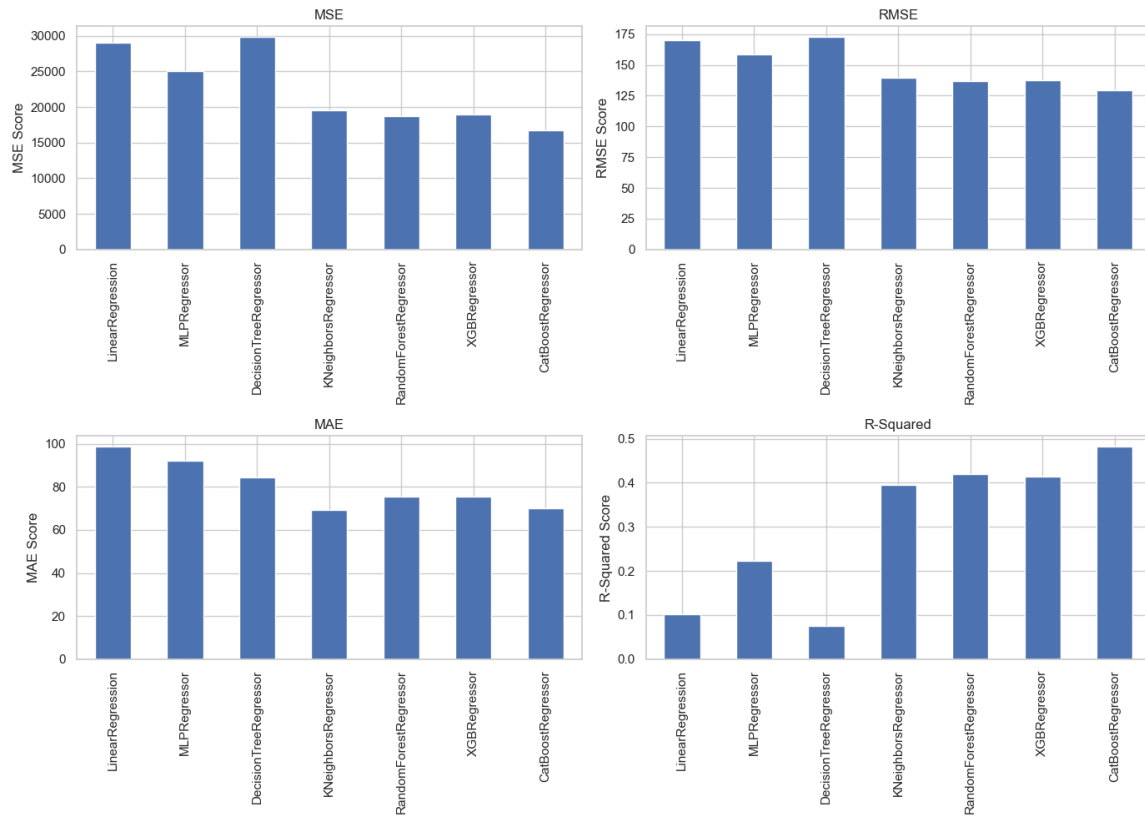


Fig. 6.1: Comparison Results for the I-RS-GCC Models

Fig. 6.2 shows the prediction performance and residual analysis of the I-RS-GCC CatBoost model. The left scatter plot (a) illustrates the actual versus predicted ROI for the I-RS-GCC CatBoost model, where most data points cluster below 200, but extreme values reach around 700. The right plot (b) shows Histograms of the residual distribution, with most errors centered

around zero, indicating correct predictions, but extending from approximately -600 to 800, indicating a right-skewed pattern.

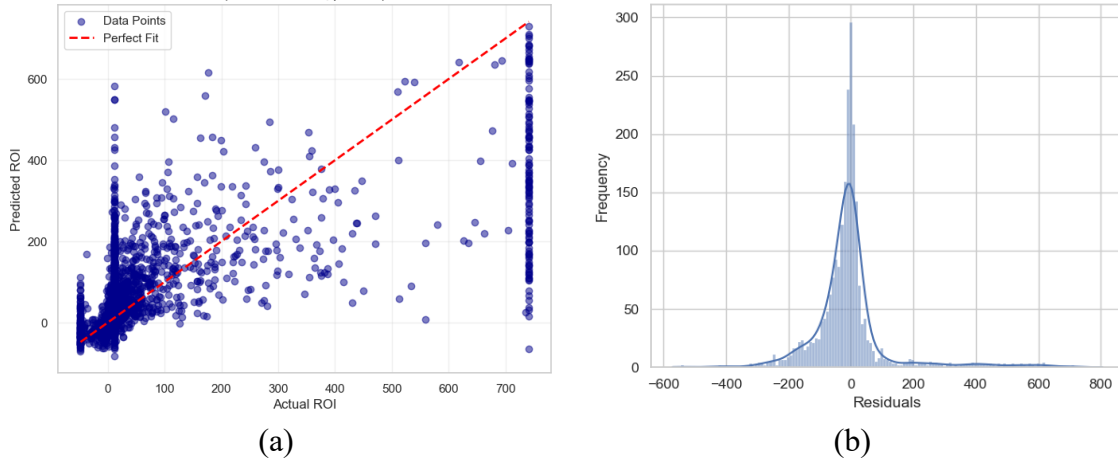


Fig. 6.2: (a) Actual Versus Predicted ROI (b) Histogram of Residuals Distribution for I-RS-GCC CatBoost Model

The prediction model tends to underestimate high ROI values by margins exceeding 700 while simultaneously overestimating several low ROI values by approximately 200, indicating a bias in the model's performance across different ROI ranges. These values suggest that while the model performs well for lower ROI values in general, it struggles with higher predictions, leading to larger deviations.

6.2.2 Variable Importance

As presented previously, CatBoost demonstrated the best performance among the predictive models. The SHAP TreeExplainer function was utilized to interpret the CatBoost model's predictions. The CatBoost model was trained with the best-fitting parameters (learning_rate=0.1, iterations=500, depth=10). Fig. 6.3 illustrates the significance of the SHAP variable, and summary plots offer a detailed understanding of the model's output and the contribution of individual variables in predicting ROI for the GCC model.

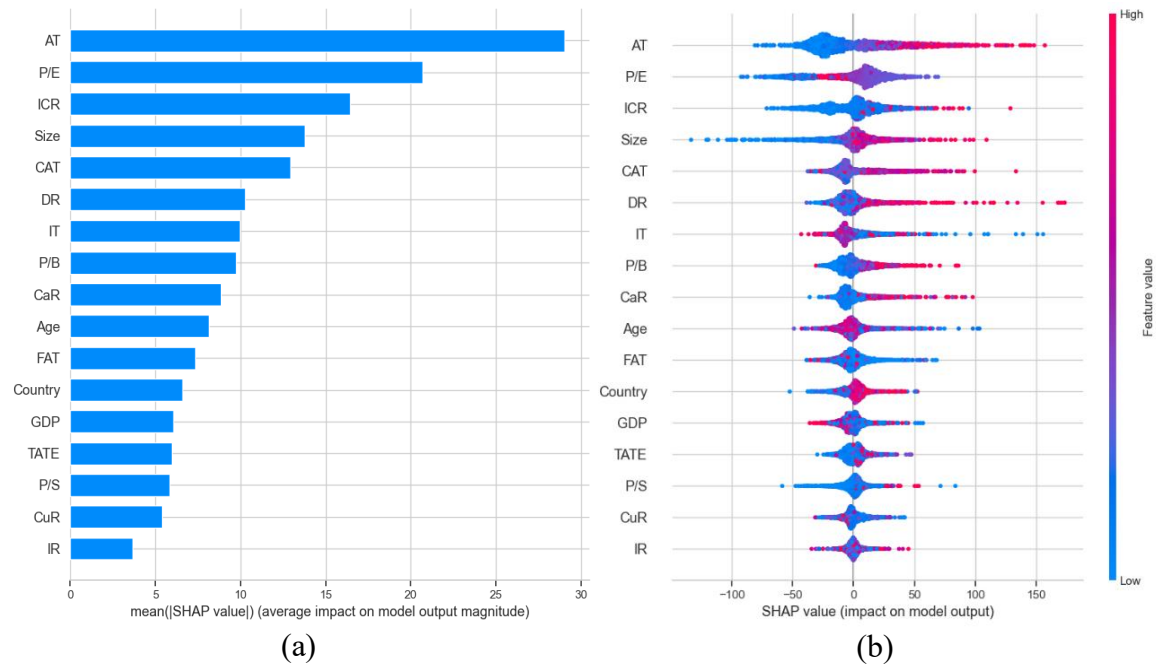


Fig. 6.3: (a) SHAP Variable Importance (b) SHAP Summary Plots for I-RS-GCC Model using CatBoost

The SHAP Variable Importance Plot in Subfigure 6.3 (a) ranks variables by their average impact on the model output, with higher SHAP values indicating greater significance. The Asset Turnover and Price-Earnings ratios are the most important for the All GCC model, followed by the Interest Coverage Ratio and firm Size, which are crucial for analyzing firm performance in the GCC. This finding aligns with Zhou et al. (2022), who observed that asset utilization correlates positively with firms' profitability and contradicts Barboza et al.'s (2017) finding that asset utilization has a limited correlation with financial performance.

The SHAP Summary Plot in Subfigure 6.3 (b) illustrates the magnitude and direction of variable effects. Each point represents a SHAP value for a single prediction, with a color gradient indicating the variable value (from low to high). As shown, Asset Turnover and firm Size positively correlate with

ROI predictions, while variations in Price-Earnings and Interest Coverage Ratios influence ROI predictions across the dataset.

6.3 NI-RS-GCC Model Experimental Results

6.3.1 Evaluation of Model Performance

Table 6.2 presents the performance metrics for the models employed under the random sampling with no data imputation approach applied to all GCC firms. The optimal tuning hyperparameters for models are presented in Appendix 6.

Table (6.2): Performance Results of NI-RS-GCC Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.533	88.470	31084.407	176.308
CatBoost	0.542	86.471	30496.782	174.633

XGBoost and CatBoost performed better than the imputation approach in the previous section. CatBoost outperforms XGBoost slightly, achieving the highest R-squared value of 0.542. This implies that 54.2% of the variance in the dependent variable, ROI, is explained by the model, indicating that CatBoost performs better than other models in capturing the relationship between independent variables and ROI. CatBoost also has the lowest MAE (86.471), MSE (30,496.782), and RMSE (174.633), indicating higher accuracy in predicting firm performance compared to XGBoost. Fig. 6.4 shows the comparison results for the NI-RS-GCC model.

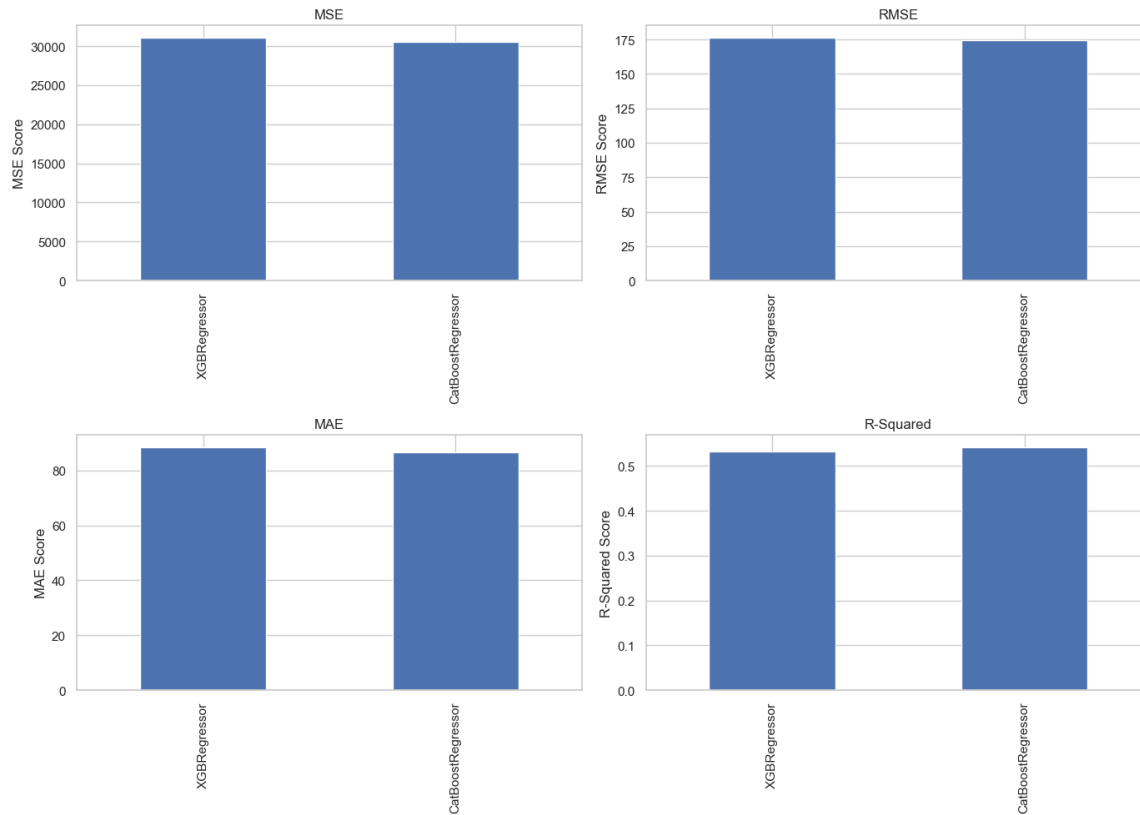


Fig. 6.4: Comparison Results for the NI-RS-GCC Model

The left plot (a) in Fig. 6.5 shows actual versus predicted ROI for the NI-RS-GCC CatBoost model. It reveals that the model tends to underestimate high ROI values by over 1,000 units and overestimate low ROI values (below zero) by up to 200 units and, in some cases. This indicates a prediction bias across different ROI ranges. While the model generally performs well for lower ROI values, it struggles with higher values, resulting in larger errors.

The right plot (b) displays residual distribution, with most errors clustering around zero but ranging from approximately -500 to over 1,000. This right-skewed pattern highlights the model's difficulty in handling extreme ROI values.

Overall, the residuals are notably non-normal, and the presence of significant outliers indicates that the model's accuracy could be improved, particularly for higher ROI predictions. Notably, the higher outliers increased by 200 in the NI-RS-GCC model compared to the I-RS-GCC model, suggesting a greater deviation in the non-imputed dataset and highlighting the impact of missing data on prediction stability.

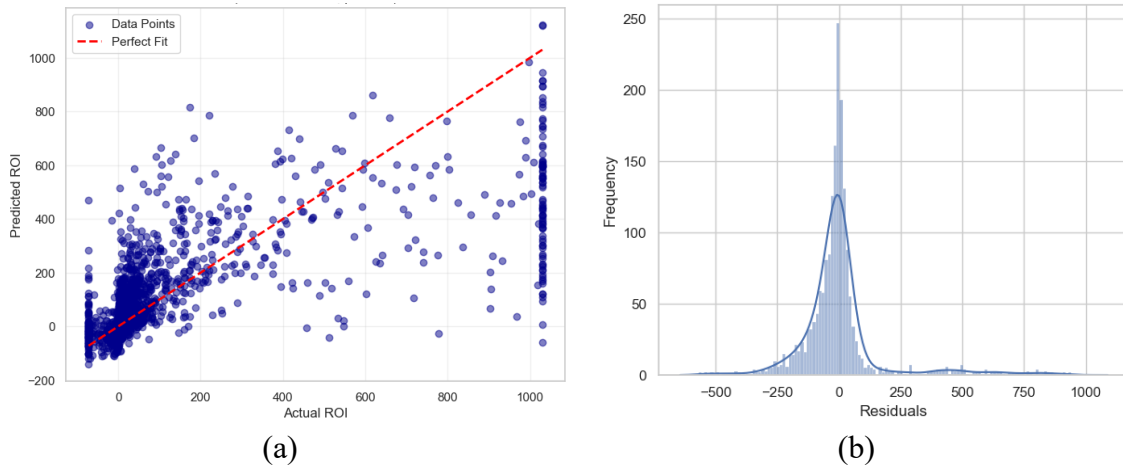


Fig. 6.5: (a) Actual Versus Predicted ROI (b) Residuals Distribution for NI-RS-GCC CatBoost Model

6.3.2 Variable Importance

Fig. 6.6 presents the variable importance analysis for the NI-RS-GCC CatBoost model. SHAP Variable Importance in Subfigure 6.6 (a) indicates that Asset Turnover, Price-to-Earnings Ratio, Current Asset Turnover, and Interest Coverage Ratio exhibit the highest mean SHAP values, indicating their strong influence on ROI predictions. These factors are followed by variables such as firm Age, Cash Ratio, Debt Ratio, and Industry Type, which contribute less significantly. The plot reflects the cumulative impact of each variable across all observations.

Fig. 6.6 (b) shows the SHAP Summary, which provides a more comprehensive view. Variables such as Asset Turnover, Current Asset Turnover, Debt Ratio, Cash Ratio, and Price-to-Book Ratio have a positive impact on ROI, as indicated by the concentration of red points in the positive SHAP region. This finding is consistent with Zhou et al. (2022), who reported a positive correlation between asset utilization and firm profitability. In contrast, Age affects the ROI negatively, meaning higher values of these variables decrease the firm's profitability. The Price-to-Earnings Ratio has a positive impact on ROI and becomes negative at significantly higher values. The interest coverage ratio does not generally impact ROI. Its positive effects are only observed at higher, limited values.

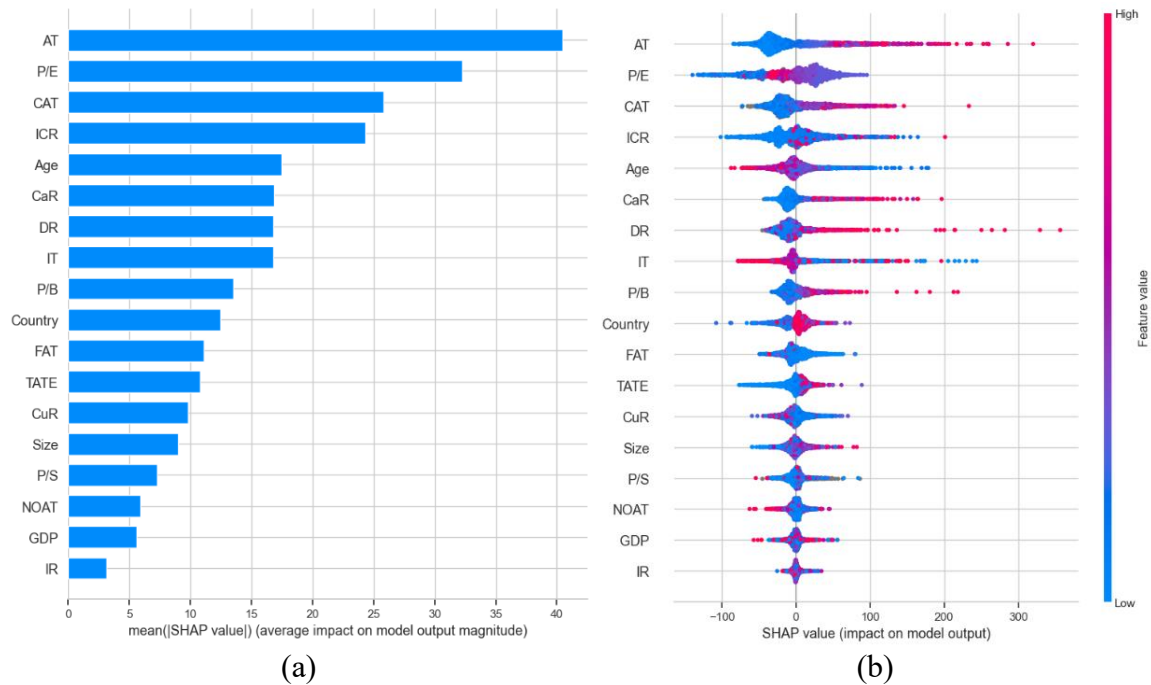


Fig. 6.6: (a) SHAP Variable Importance (b) SHAP Summary Plots for NI-RS-GCC Model using CatBoost

6.4 C Model Experimental Results

6.4.1 Kingdom of Bahrain

6.4.1.1 *Evaluation of Model Performance*

Fig. 6.7 and Table 6.3 present the performance metrics for Bahrain's model, which analyzed 40 Bahrain-listed firms from 2003 to 2022, resulting in 800 observations in the subset.

The results show that CatBoost outperforms XGBoost. CatBoost has a higher R-squared (0.781 versus 0.762), indicating better variance explanation and lower error rates.

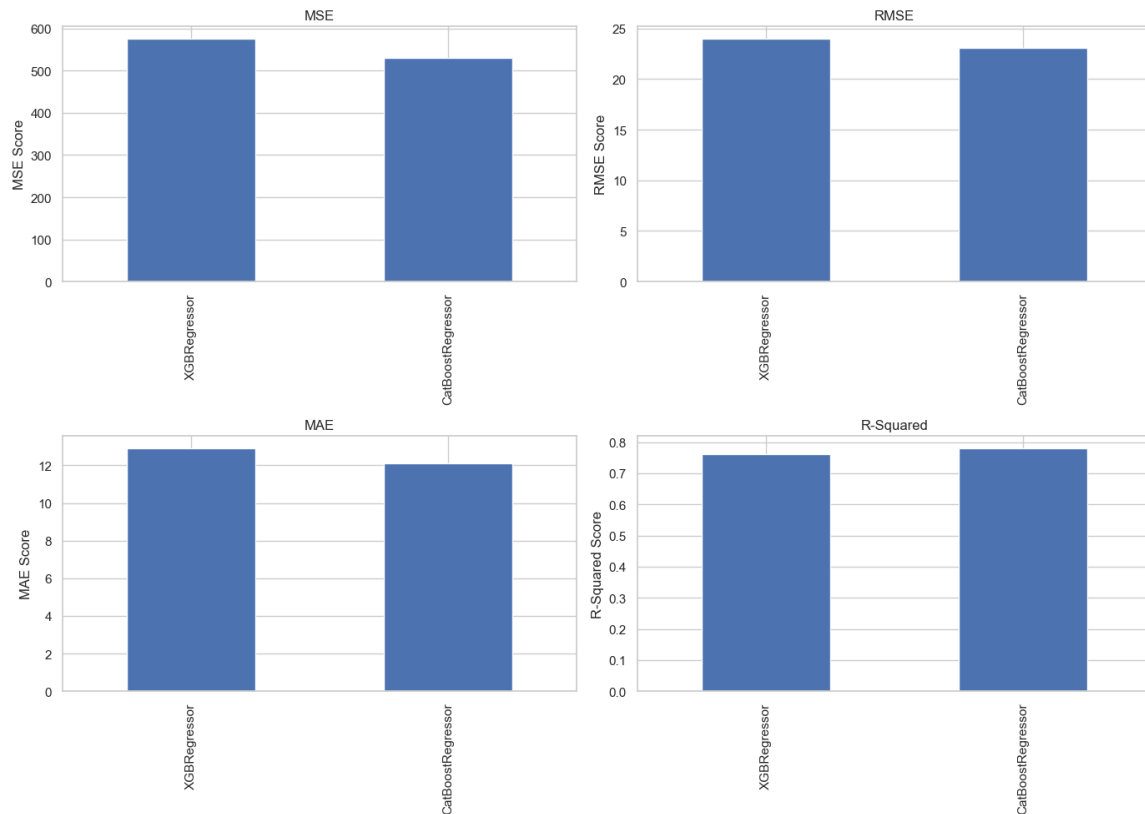


Fig. 6.7: Comparison Results for Bahrain Models

Table (6.3): Evaluation Results of Bahrain Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.762	12.910	575.924	23.998
CatBoost	0.781	12.110	531.042	23.044

Fig. 6.8 assesses the performance of the Bahrain CatBoost model. Subfigure (a) displays the relationship between actual and predicted ROI, revealing a strong correlation along the ideal fit line, particularly for lower ROI values. However, as ROI values increase, the model tends to underestimate, leading to greater prediction errors, especially for ROI values exceeding 200.

Subfigure (b) illustrates the distribution of the residuals, where the majority are clustered around zero, accompanied by a slight skew and the presence of occasional extreme values at 150.

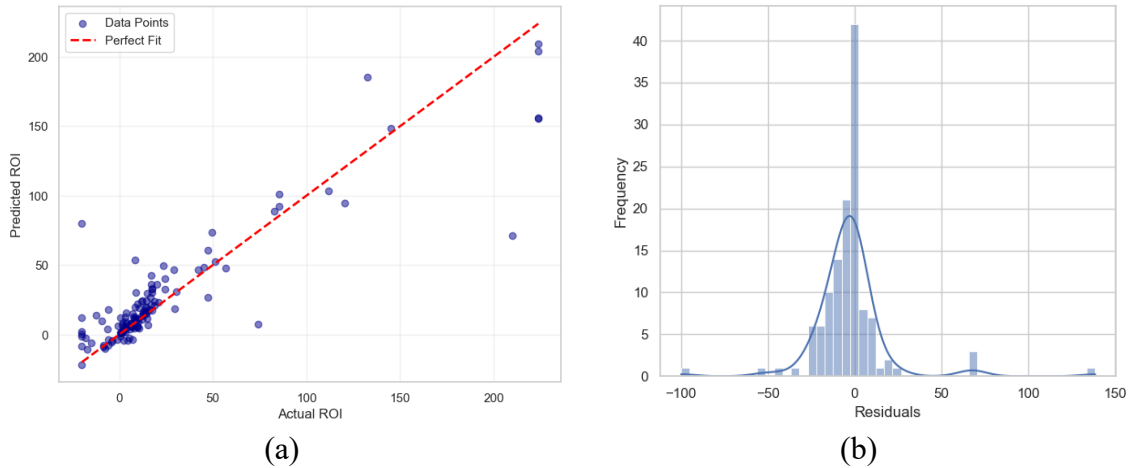


Fig. 6.8: (a) Actual versus Predicted ROI (b) Residuals Distribution for Bahrain CatBoost Model

6.4.1.2 *Variable Importance*

Given that CatBoost outperformed XGBoost in the Bahrain model, the SHAP explainer was trained using the optimal parameters (learning_rate: 0.05, iterations: 500, depth: 6).

Subfigure (a) in Fig. 6.9 illustrates that Asset Turnover is the most influential variable, followed by firm Size and Price-to-Earnings Ratio. These variables have the highest mean SHAP values, contributing the most to the variation in ROI predictions. This finding aligns with Zhou et al. (2022), who observed that asset utilization correlates positively with firms' profitability and contradicts Barboza et al.'s (2017) finding that asset utilization has a limited correlation with financial variables related to bankruptcy. Moreover, Subfigure (a) reveals that less influential variables, such as Current Asset Turnover and Industry Type, have much lower average SHAP values, indicating that they play a minor role in driving the model's predictions for ROI.

Subfigure (b) illustrates the relationships between various factors and the predicted ROI for Bahraini firms. The analysis indicates that Asset Turnover, firm size, Price-to-Book Ratio, Debt Ratio, Net Operating Asset Turnover, and Cash Ratio have a positive influence on ROI. Conversely, firm Age and GDP demonstrate a negative impact on ROI. Additionally, Price-to-Earnings Ratio, Current Asset Turnover, Fixed Asset Turnover, and other variables exhibit more varied impacts, sometimes pulling predictions in both positive and negative directions.

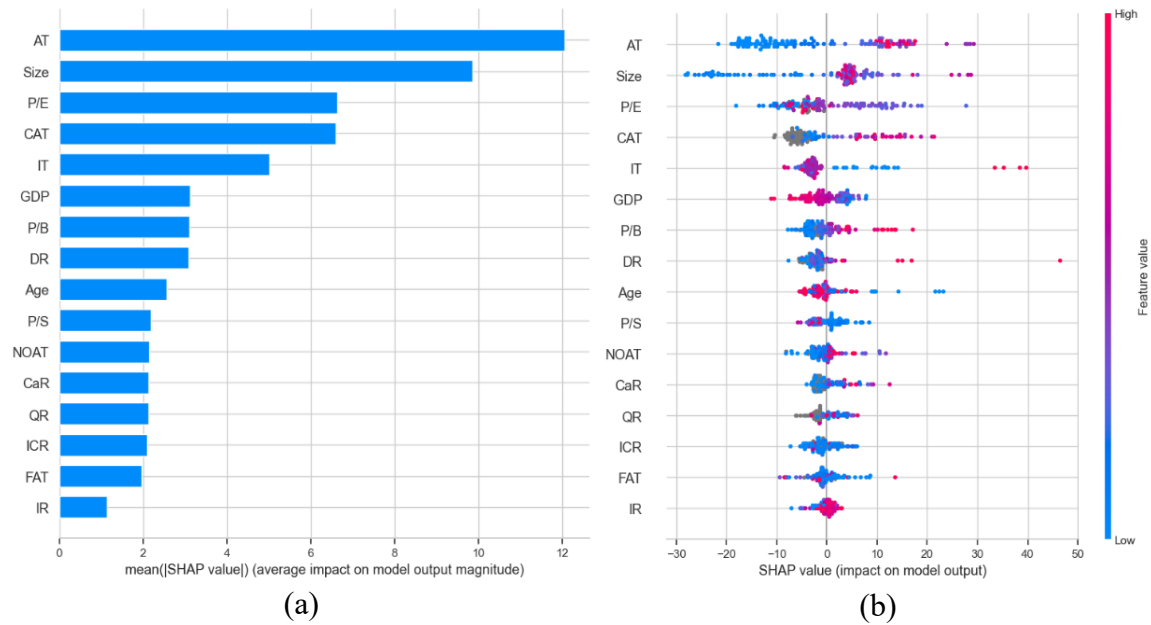


Fig. 6.9: (a) SHAP Variable Importance (b) SHAP Summary Plots for Bahrain Model using CatBoost

6.4.2 Kuwait

6.4.2.1 Evaluation of Model Performance

The XGBoost and CatBoost models, applied to 145 listed Kuwaiti firms with 2,900 observations, demonstrate similar performance with an R-squared of 0.577.

Both XGBoost and CatBoost models achieved a similar R-squared of 0.577, with the CatBoost model showing slightly better accuracy, achieving a lower MAE of approximately 33.155 and an RMSE of 69.866, as presented in Table 6.4 and Fig. 6.10.

Table (6.4): Evaluation Results of Kuwait Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.577	33.375	4885.020	69.893
CatBoost	0.577	33.155	4881.231	69.866

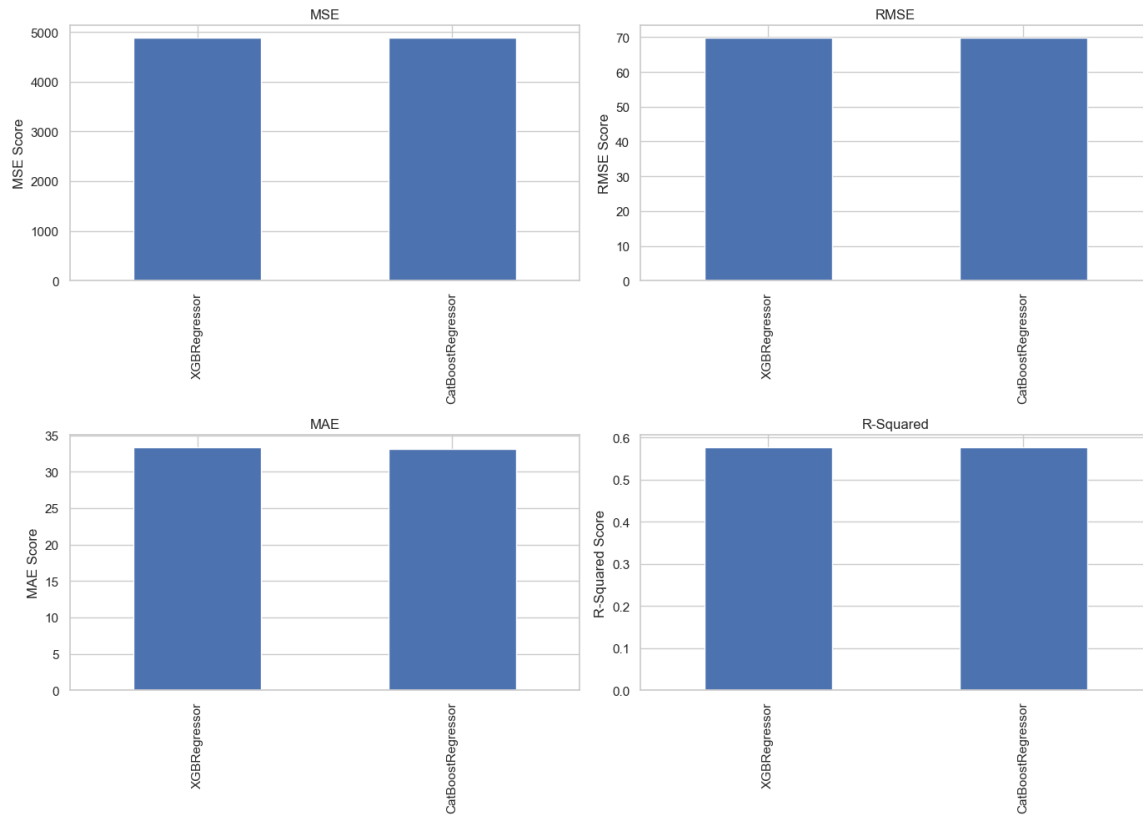


Fig. 6.10: Comparison Results for Kuwait Models

Fig. 6.11 evaluates the Kuwaiti CatBoost model performance using two key diagnostic plots. Subfigure (a) shows a concentration of data points below 50, with extreme values reaching around 400. The red dashed line represents the perfect fit, highlighting underestimation, particularly for ROI values exceeding 400.

Subfigure (b) presents the distribution of the residuals, where most errors are centered around zero but range from approximately -300 to 4,000, indicating a right-skewed pattern.

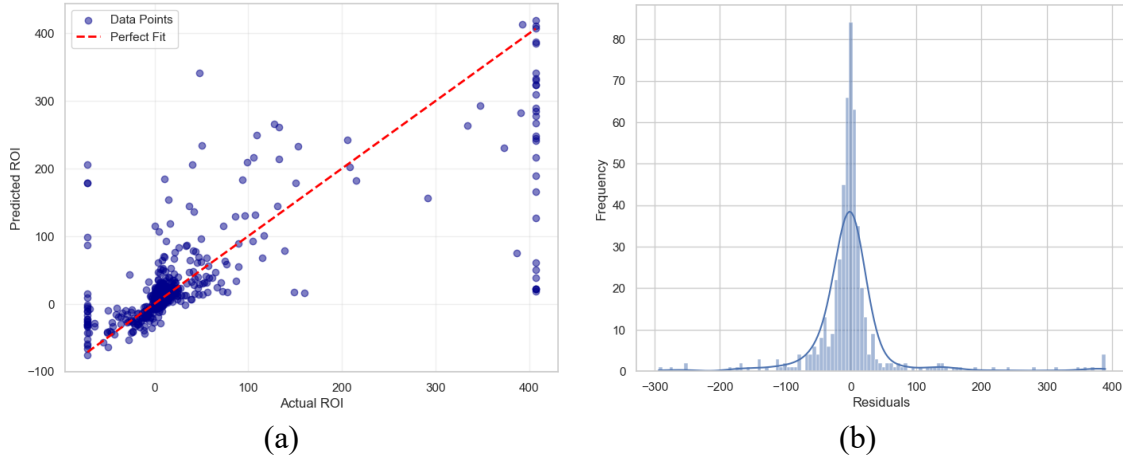


Fig. 6.11: (a) Actual versus Predicted ROI (b) Residuals Distribution for Kuwait CatBoost Model for Kuwait CatBoost Model

6.4.2.2 *Variable Importance*

This section presents the SHAP analysis of the Kuwait prediction model, exploring the variables that most influence the profitability of Kuwaiti firms from 2003 to 2022. Since CatBoost and XGBoost both explain 57.7% of the variance in ROI, the SHAP explainer is trained using CatBoost's optimal parameters (learning rate: 0.1, iterations: 200, depth: 6) to interpret the impact of these variables on ROI for Kuwaiti listed firms during this period.

Subfigures 6.12 (a) and (b) illustrate the SHAP values, identifying the most influential variables affecting the ROI prediction for Kuwaiti firms. The top variables impacting the model are Interest Coverage Ratio and Price-to-Earnings Ratio, followed by Asset Turnover and Current Asset Turnover, which contribute the most to the model's predictions. Jabeur et al. (2021) identified solvency as the most critical factor in predicting corporate failure, reporting a positive correlation between solvency ratios and default probability, which aligns with the current study's findings.

The Interest Coverage Ratio, Asset Turnover, Current Asset Turnover, and Price-to-Book Ratios consistently exhibit a substantial positive impact. Additionally, it is observed that some firms with high ROI face low Interest Coverage Ratio. This means that for some Kuwaiti firms, the Interest Coverage Ratio does not affect their ROI.

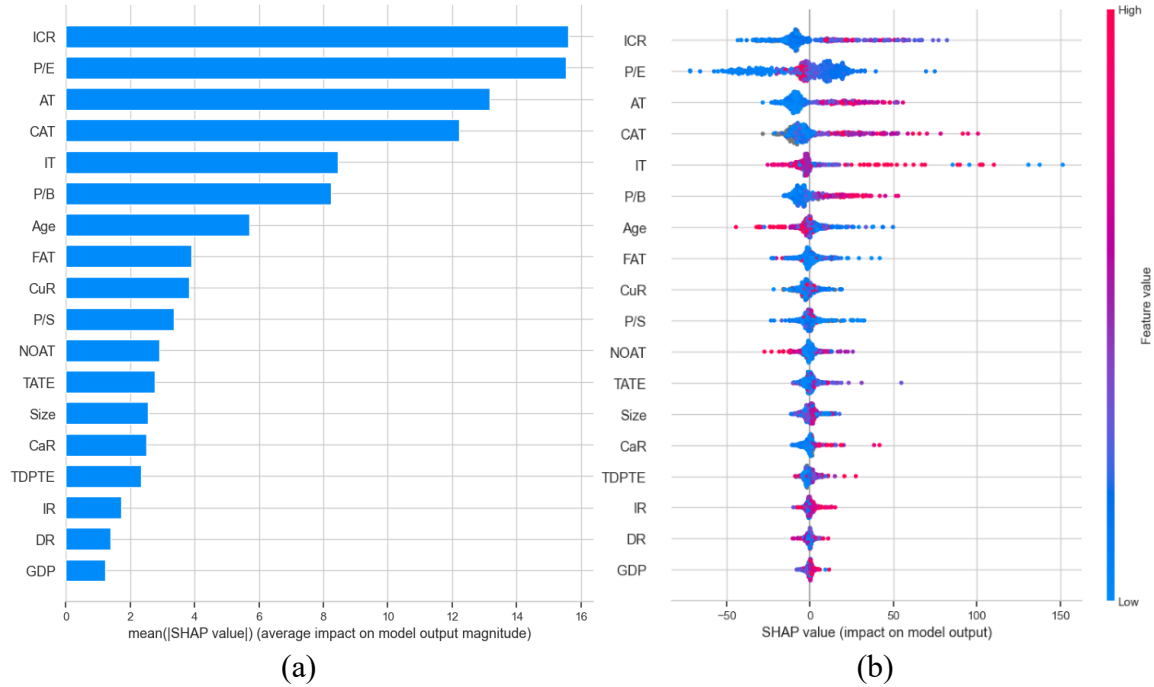


Fig. 6.12: (a) SHAP Variable Importance (b) SHAP Summary Plots for Kuwait Model using CatBoost

6.4.3 Oman

6.4.3.1 *Evaluation of Model Performance*

Table 6.5 presents the performance metrics for the Oman subset prediction models, comprising 2,280 observations from 114 actively listed Omani firms from 2003 to 2022, and Fig. 6.13 compares XGBoost and CatBoost. Unlike in previous models, XGBoost outperformed CatBoost in the Oman dataset,

achieving an R-squared value of 0.647 compared to 0.609. This indicates that XGBoost explains 64.7% of the variance in ROI. Additionally, XGBoost demonstrated superior error metrics, with a lower MAE of 175.742 and RMSE of 347.667.

Table (6.5): Evaluation Results of Oman Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.647	175.742	120872.316	347.667
CatBoost	0.609	177.930	133959.980	366.005

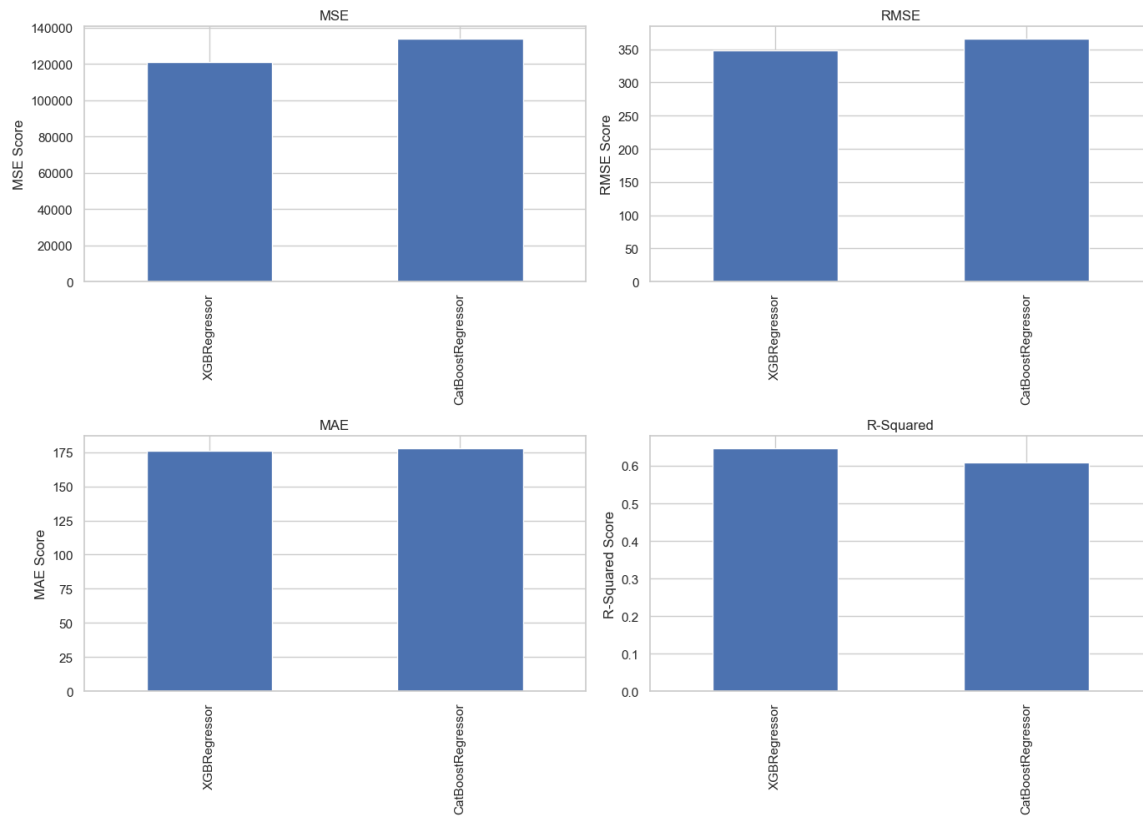


Fig. 6.13: Comparison Results for Oman Models

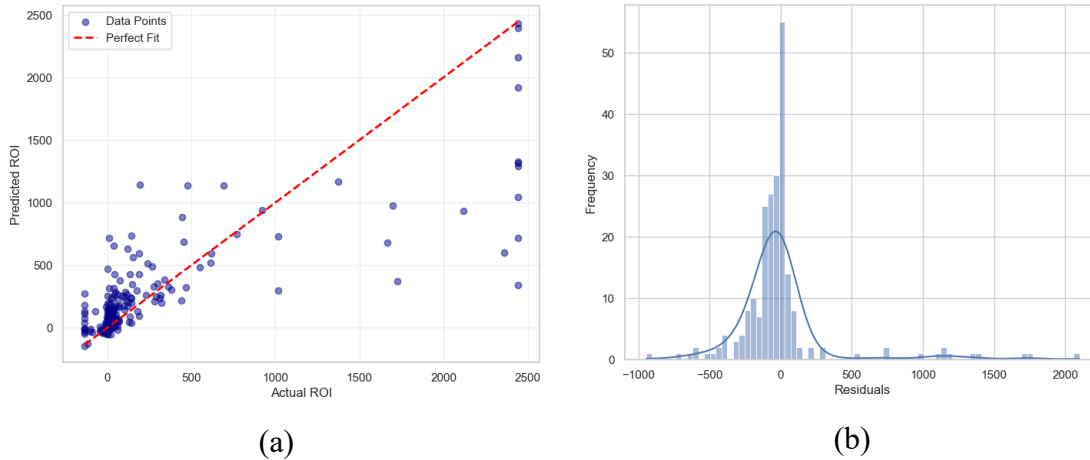


Fig. 6.14: (a) Actual versus Predicted ROI (b) Residuals Distribution for Oman XGBoost Model

Fig. 6.14 presents two diagnostic plots evaluating the Omani CatBoost model. Subfigure (a) shows that most data points are concentrated below an ROI of 250, while some extreme values reach up to 2,500. The model significantly underestimates these high ROI values, with prediction errors dropping below 500. Subfigure (b) displays residuals centered around zero but ranging from -1,000 to 2,000, indicating a right-skewed distribution. In comparison, residuals from the XGBoost model exhibit even more pronounced skewness, suggesting greater instability in extreme predictions.

6.4.3.2 Variable Importance

Given that XGBoost outperforms CatBoost in Omani models, explaining 64.7% of the variance in ROI, the SHAP explainer is trained using XGBoost's optimal parameters (learning rate: 0.1, iterations: 200, depth: 6) to interpret the variable impacts on ROI for Omani listed firms during this period.

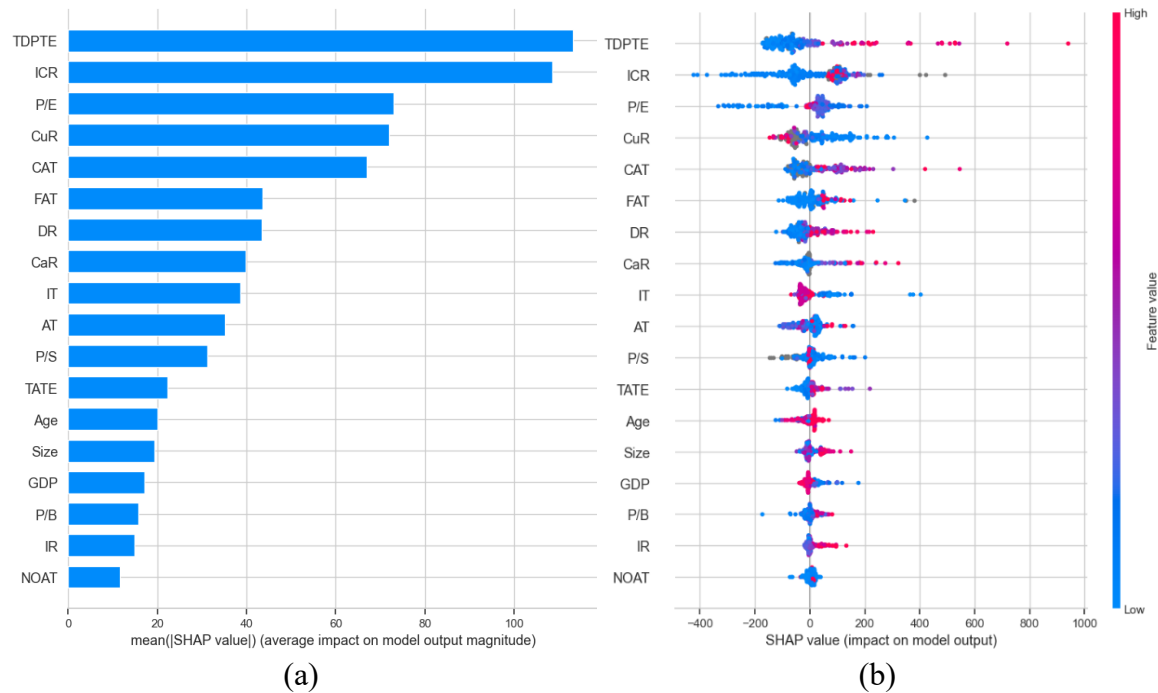


Fig. 6.15: (a) SHAP Variable Importance (b) SHAP Summary Plots for Oman Model using XGBoost

Subfigure 6.15(a) illustrates that the three most significant variables influencing profitability are Total Debt as a Percentage of Total Equity, Interest Coverage Ratio, and Price-to-Earnings Ratio. This finding aligns with Halteh (2018), who identified Total Debt as a Percentage of Total Equity as the second most important predictor of financial distress.

In Subfigure 6.15 (b), it is evident that the Total Debt Percentage of Total Equity has a positive effect on the profitability of Omani firms. Additionally, the Interest Coverage Ratio has a positive impact on profitability only at higher values. It has no effect on ROI for Omani firms that are already achieving high levels of profitability.

6.4.4 Qatar

6.4.4.1 *Evaluation of Model Performance*

Fig. 6.16 shows the ROI prediction results for the Qatar subset, comprising 1,060 observations from 53 actively listed Qatari firms from 2003 to 2022. CatBoost outperforms XGBoost, achieving an R-squared value of 0.773 compared to 0.734, indicating that CatBoost explains 77.3% of the variance in ROI.

Additionally, CatBoost demonstrates superior performance with a lower MAE of 43.733 and RMSE of 83.690, indicating a better fit for predicting ROI in Qatari firms (detailed in Table 6.6).

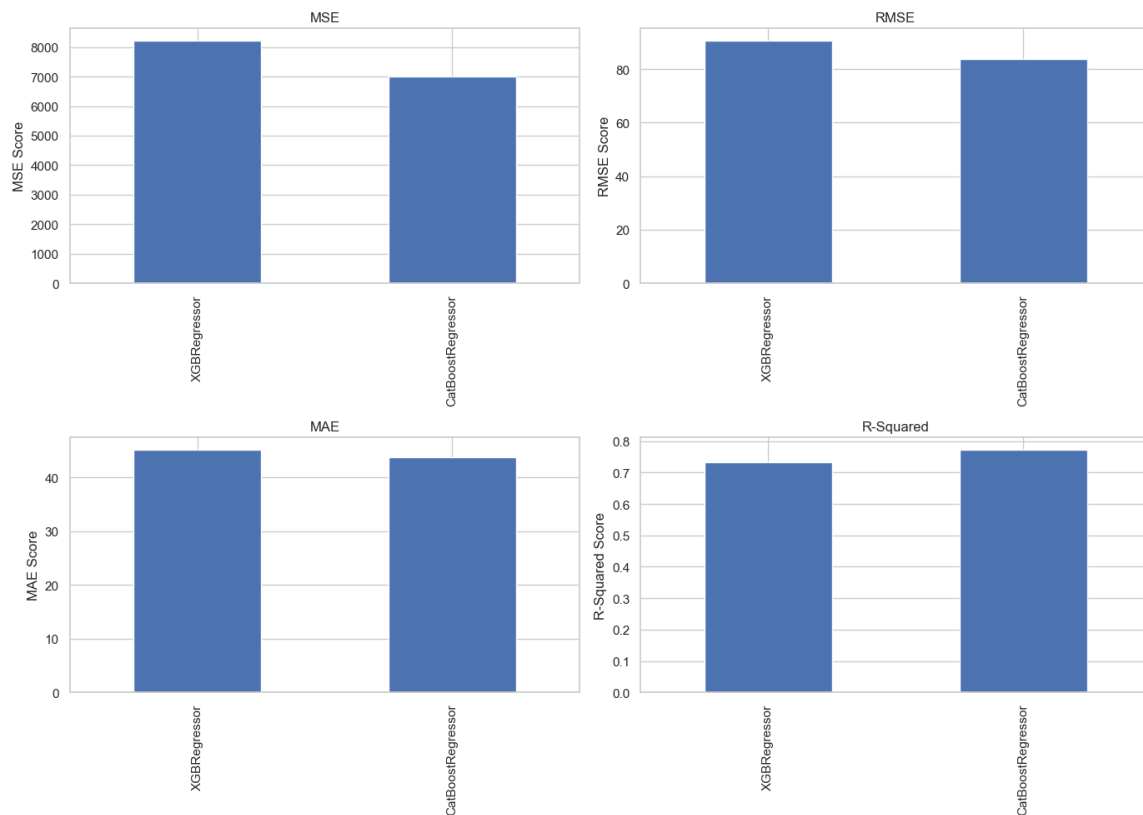


Fig. 6.16: Comparison Results for Qatar Model

Table (6.6): Evaluation Results of Qatar Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.734	45.148	8219.500	90.661
CatBoost	0.773	43.733	7003.996	83.690

Fig. 6.17 presents the performance of the Qatari CatBoost model using two diagnostic plots. Subfigure (a) shows that most data points are concentrated below the ROI value of 50, while extreme values reach up to 600. The model fails to capture these extremes, underestimating them by as much as 50. The red dashed line indicates the ideal fit, making deviations, particularly for higher ROI values, clearly visible. Subfigure (b) displays the residual distribution, with most errors centered around zero but ranging from -300 to 600, indicating a right-skewed pattern and challenges in predicting extreme ROI values.

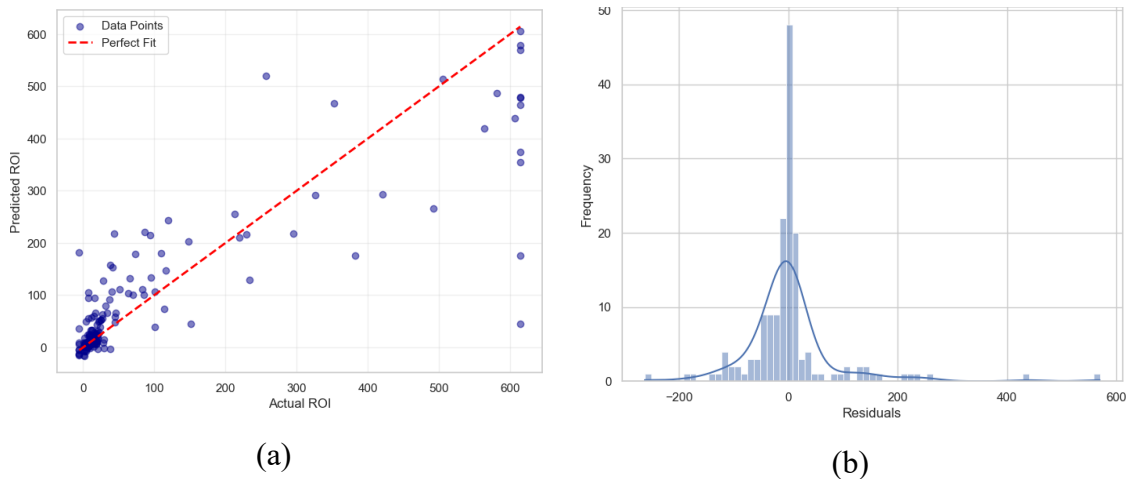


Fig. 6.17: (a) Actual versus Predicted ROI (b) Residuals Distribution for Qatar CatBoost Model

6.4.4.2 Variable Importance

Since CatBoost explains 77.3% of the variance in ROI better than XGBoost, the SHAP explainer is trained using CatBoost's optimal parameters (learning rate: 0.05, iterations: 500, depth: 6) to interpret the impact of these variables on ROI for Qatari-listed firms during this period. The two subfigures in Fig. 6.18 illustrate variables affecting the ROI prediction for Qatari firms.

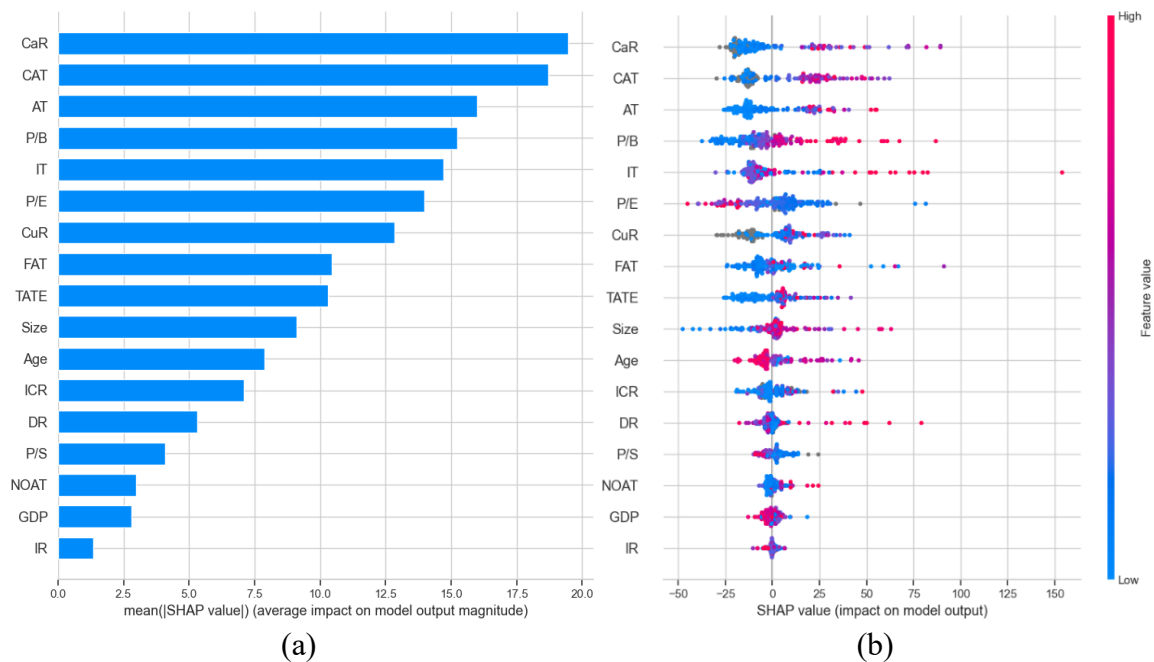


Fig. 6.18: (a) SHAP Variable Importance (b) SHAP Summary Plots for Qatar Model using CatBoost

The four most influential variables are the Cash Ratio, Current Asset Turnover, Asset Turnover, and Price-to-Book Ratio (Subfigure 6.18(a)). These variables also have a positive impact on Qatari profitability, indicating that higher liquidity, efficient asset utilization, and favorable market valuations contribute to increased returns for Qatari firms (Subfigure 6.18(b)). The strong positive influence of liquidity on Qatari firms aligns with the findings of Jabeur et al. (2021), who concluded that non-distressed firms

typically maintain higher average liquidity levels in corporate failure prediction, whereas failed firms often struggle to meet their debt obligations.

6.4.5 Kingdom of Saudi Arabia

6.4.5.1 *Evaluation of Model Performance*

Table 6.7 and Fig. 6.19 present the ROI prediction results for the Saudi subset, comprising 6,280 observations from 314 actively listed firms from 2003 to 2022. The analysis employs the XGBoost and CatBoost models, using the optimal hyperparameters in Appendix 6. Both models yield comparable R-squared values, 0.320 for XGBoost and 0.328 for CatBoost, indicating that approximately 32.8% of the variance in ROI is explained by the models.

However, CatBoost demonstrates a slight edge in performance, achieving a lower Mean Absolute Error (MAE) of 223.745 and a marginally lower Root Mean Squared Error (RMSE) of 428.819. These results suggest that CatBoost provides a slightly better fit for predicting ROI among Saudi firms.

Table (6.7): Evaluation Results of Saudi Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.320	228.210	186246.199	431.563
CatBoost	0.328	223.745	183885.537	428.819

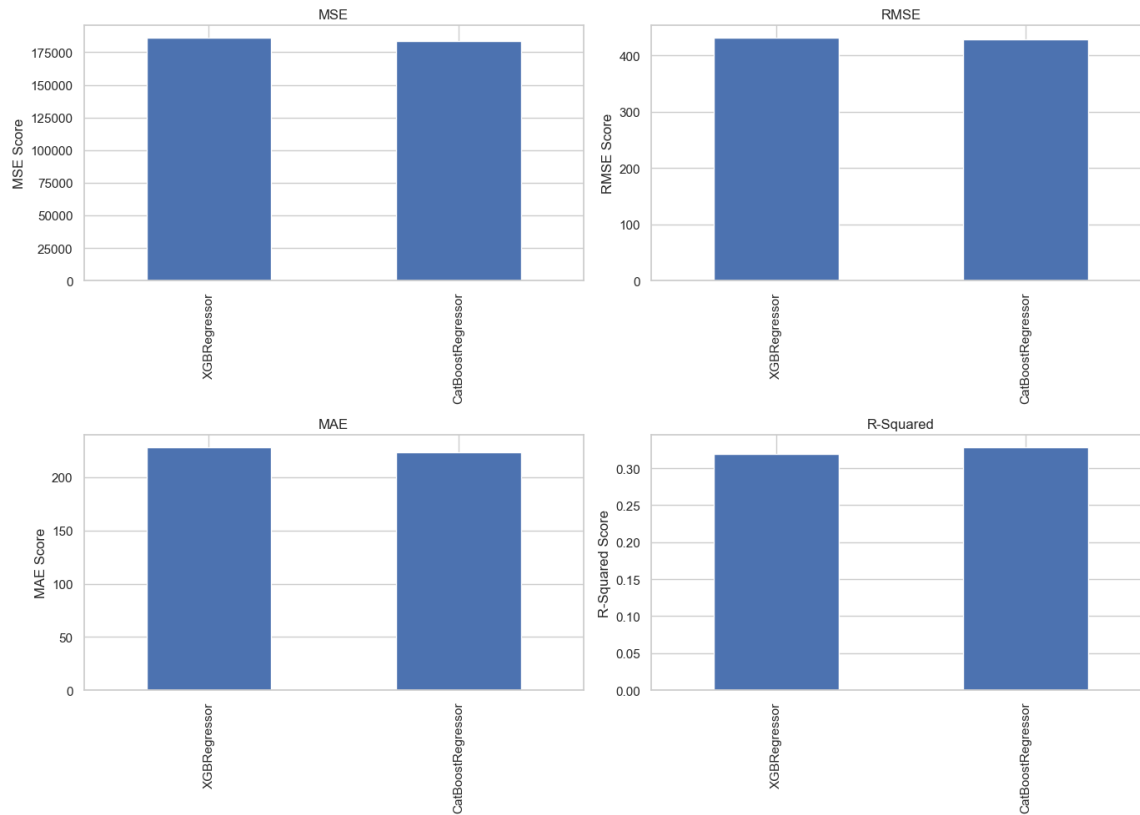


Fig. 6.19: Comparison Results for Saudi Models

Fig. 6.20 evaluates the performance of the Saudi CatBoost model using two diagnostic plots. Subfigure (a) shows that most data points are concentrated below an ROI of 400, while extreme values reach up to 2,500. The model severely underestimates these extremes, with predicted values dropping to zero. The red dashed line represents the ideal fit, highlighting significant deviations, particularly at higher ROI levels.

Subfigure (b) shows the residual distribution, where most errors are centered around zero but range from below -1,000 to approximately 2,400, indicating a highly right-skewed pattern and difficulty in capturing extreme ROI values.

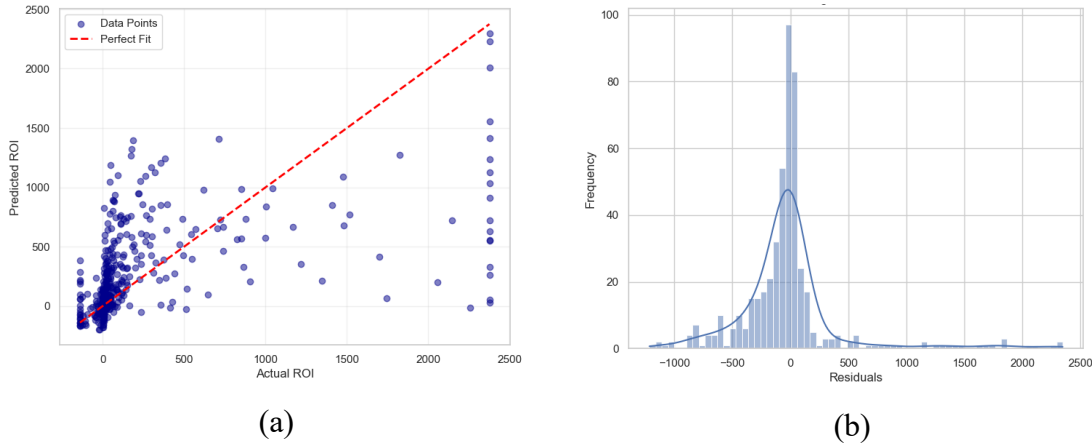


Fig. 6.20: (a) Actual versus Predicted ROI (b) Residuals Distribution for Saudi CatBoost Model

6.4.5.2 Variable Importance

Since CatBoost and XGBoost both explain 32% of the variance in ROI, the SHAP explainer is trained using CatBoost's optimal parameters (learning rate: 0.05, iterations: 500, depth: 6) to interpret the impact of these variables on ROI for Saudi listed firms (Fig. 6.21).

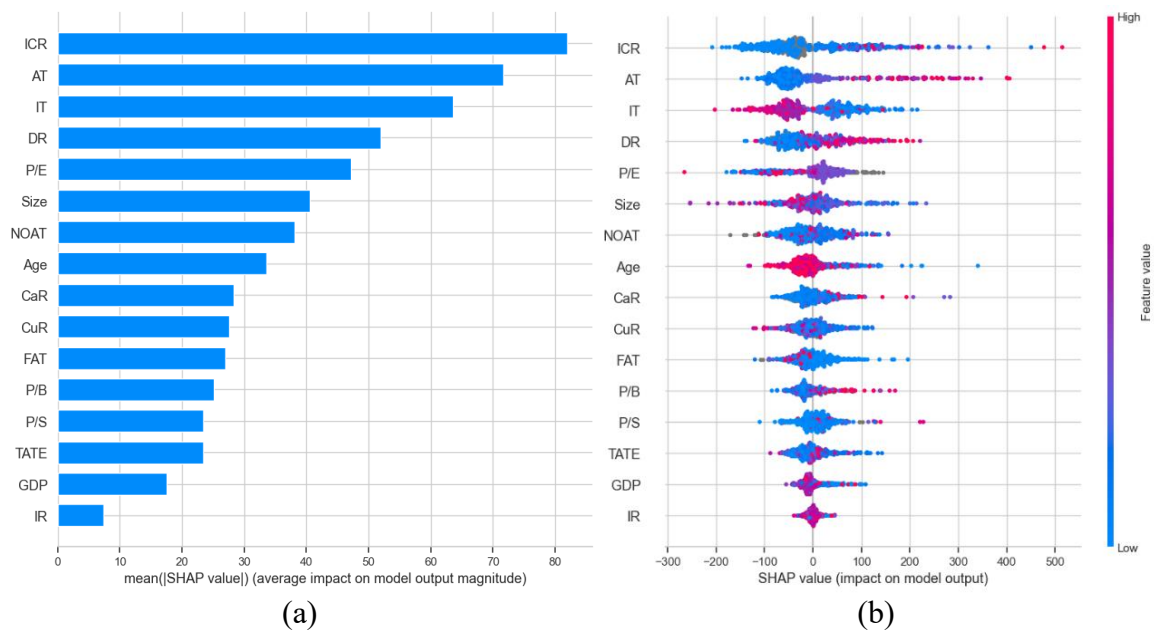


Fig. 6.21: (a) SHAP Variable Importance (b) SHAP Summary Plots for Saudi Model using CatBoost

The variable importance plot shows that the Interest Coverage Ratio, a representative of firm solvency, has the highest average impact on the model's output, followed by Asset Turnover, Industry Type, Debt Ratio, and Price-to-Earnings ratios (Fig. 6.21(a)). In Arora & Saurabh's (2022) study, they ranked solvency as the fourth most significant predictor of bankruptcy in Indian firms. The summary indicates that the Interest Coverage Ratio significantly influences the model's output only at higher values and only for a limited number of firms. Asset Turnover and Debt Ratio positively affect the profitability of Saudi firms, while Industry Type and Age have a negative impact (Fig. 6.21 (b)).

6.4.6 United Arab Emirates

6.4.6.1 *Evaluation of Model Performance*

Table 6.8 and Fig. 6.22 display the ROI prediction outcomes for the UAE model; the dataset includes 3,000 observations from 150 publicly listed UAE firms spanning the years 2003 to 2022. CatBoost outperforms XGBoost, with an R-squared of 0.586 compared to 0.551, explaining 58.6% of the variance in ROI. Additionally, CatBoost achieves lower error metrics, with an MAE of 58.285 and an RMSE of 121.173.

Table (6.8): Evaluation Results of UAE Models

Model	R-squared	MAE	MSE	RMSE
XGBoost	0.551	58.315	15941.869	126.261
CatBoost	0.586	58.285	14682.792	121.173

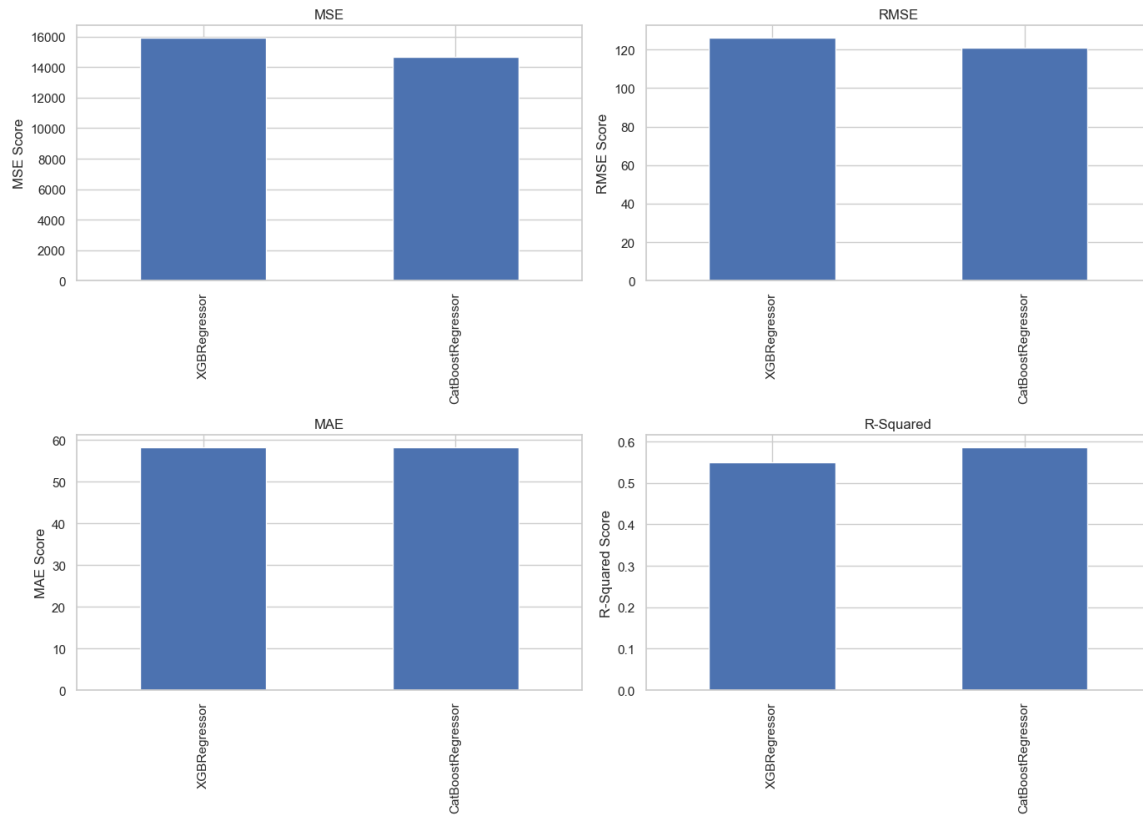


Fig. 6.22: Comparison Results for UAE Model

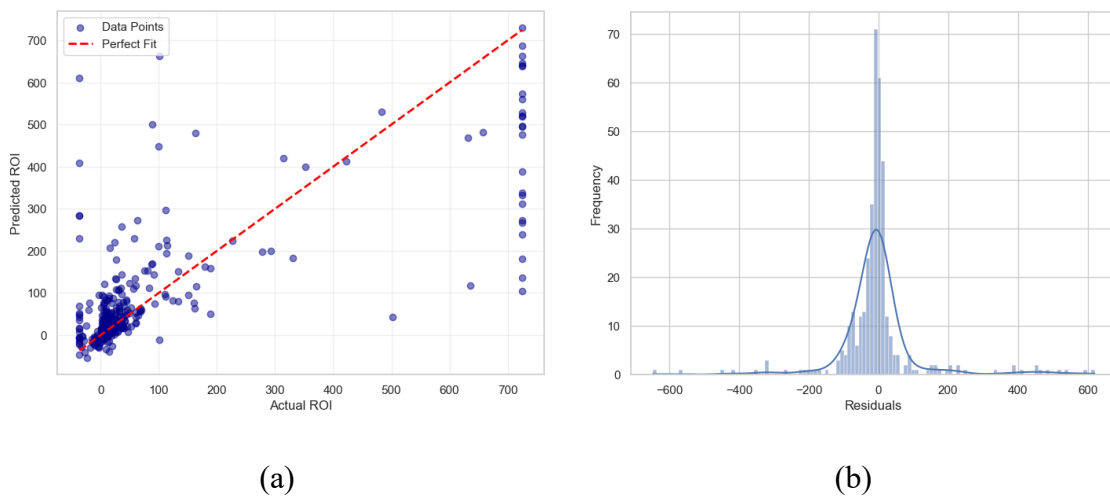


Fig. 6.23: (a) Actual versus Predicted ROI (b) Residuals Distribution for UAE CatBoost Model

Fig. 6.23 evaluates the performance of the UAE CatBoost model. Subfigure (a) shows most data points concentrated below an ROI of 80, while extreme values reach up to 700. The model significantly underestimates these high ROI values, with predictions dropping to around 100. The red dashed line represents the perfect fit, emphasizing the deviations, particularly at higher ROI levels. Subfigure (b) displays the residual distribution, with most errors centered around zero but ranging from below -600 to 600, indicating challenges in accurately predicting extreme values.

6.4.6.2 Variable Importance

The two figures illustrate the SHAP values, identifying the most influential variables affecting the ROI prediction for UAE firms.

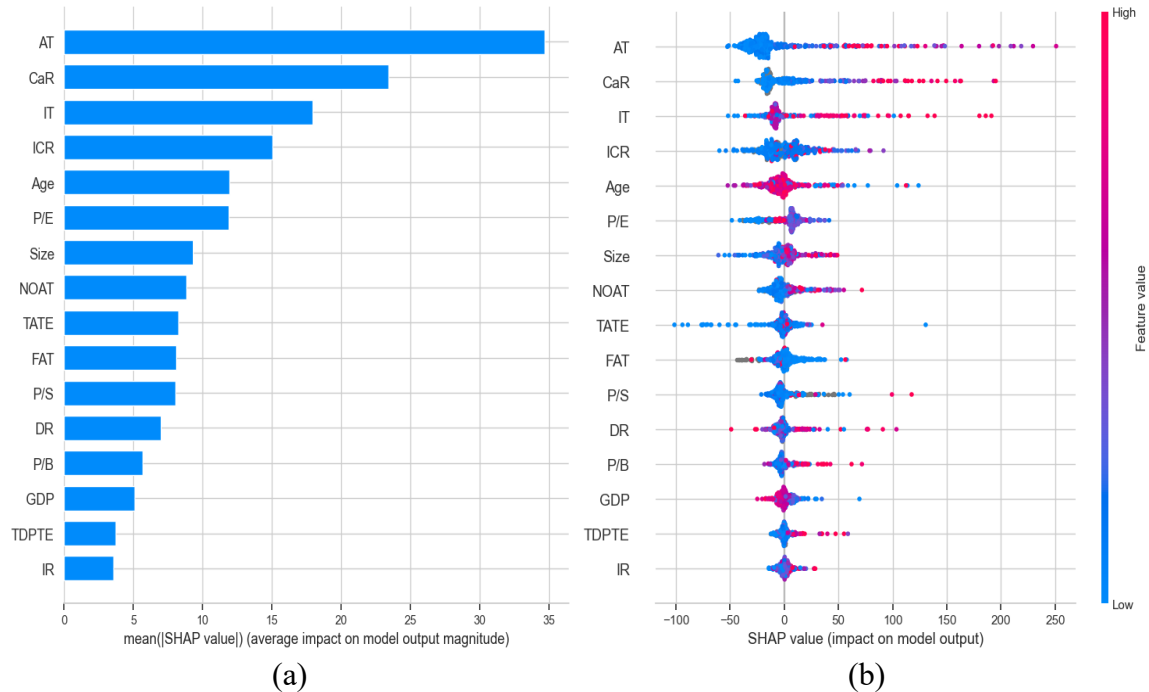


Fig. 6.24: (a) SHAP Variable Importance (b) SHAP Summary Plots for UAE Model from CatBoost

Fig. 6.24 indicates that Asset Turnover is the most influential variable, followed by the Cash Ratio and Industry Type. The SHAP summary reveals

a positive relationship between Asset Turnover and Cash Ratio with the profitability of UAE firms, while firm Age exhibits a negative relationship. This finding aligns with Zhou et al. (2022), who observed that asset utilization correlates positively with firms' profitability. The analysis also highlights a complex pattern in which specific industries have a positive influence on ROI predictions. In contrast, others have a diminishing effect, as reflected in the variability of SHAP values. Furthermore, variables such as the Total Assets to Total Equity ratio and Fixed Asset Turnover display non-linear effects.

6.5 CS-GCC Model Experimental Results

The following sections display the performance metrics for chronological partitioning with (I-CS-GCC) and without (NI-CS-GCC) imputation across all GCC countries. The optimal hyperparameters for each model are detailed in Appendix 6. Comparing the best parameters across models and periods shows that most models required adjustments in their complexity-related hyperparameters as the data size increased. While some models, like LR, remained static, others, like RF and CatBoost, adapted their depth and estimators to handle the additional data better.

6.5.1 I-CS-GCC Model Evaluation Results

Table 6.9 shows the performance metrics for chronological partitioning with imputation across different training periods, revealing that RF consistently outperforms other models regarding R-squared and error metrics. RF achieves the highest R-square of 0.281 in the two- and four-year training data models, with the lowest MAE (86.497) and RMSE (161.539) in the two-year and four-year training models, respectively. While kNN and CatBoost also show

competitive results, particularly in MAE and RMSE, they fall slightly behind RF in overall predictive power.

Table (6.9): Evaluation Results for I-CS-GCC Models

Period-1	R-squared	MAE	MSE	RMSE
LR	0.139	101.602	31260.673	176.807
ANN	0.176	91.903	29918.816	172.970
RT	-0.009	90.758	36678.018	191.515
kNN	0.242	93.557	27513.093	165.870
RF	0.255	89.685	27049.730	164.468
XGBoost	0.222	89.667	28327.114	168.307
CatBoost	0.235	90.227	27798.110	166.728
Period-2				
LR	0.141	101.614	31199.187	176.633
ANN	0.219	95.717	28360.173	168.405
RT	-0.132	92.207	41127.315	202.799
kNN	0.233	91.548	27879.483	166.971
RF	0.279	87.713	26201.354	161.868
XGBoost	0.201	90.896	29040.263	170.412
CatBoost	0.206	90.654	28847.461	169.854
Period-3				
LR	0.141	100.648	31212.530	176.670
ANN	0.219	96.671	28086.755	167.591
RT	-0.132	91.942	38027.089	195.005
kNN	0.233	87.833	26888.910	163.978
RF	0.279	86.533	26429.273	162.571
XGBoost	0.201	88.609	28360.949	168.407
CatBoost	0.206	90.843	28970.695	170.208
Period-4				
LR	0.142	101.368	31157.196	176.514
ANN	0.237	97.685	27726.481	166.513
RT	0.145	83.650	31057.043	176.230
kNN	0.25	89.495	27307.378	165.249
RF	0.281	86.497	26095.077	161.539
XGBoost	0.191	92.533	29370.538	171.378
CatBoost	0.170	94.134	30163.537	173.676
Period-5				
LR	0.130	105.177	31597.782	177.757
ANN	0.227	96.602	28096.198	167.619
RT	0.110	95.790	32340.547	179.835
kNN	0.155	95.412	30709.013	175.239
RF	0.257	91.498	27007.356	164.339
XGBoost	0.226	93.670	28105.656	167.656
CatBoost	0.216	96.104	28475.226	168.746

The LR and RT models generally underperform, with the lowest R-squared and higher error metrics, indicating poorer fitness and higher prediction errors.

Overall, the performance results from the chronological partitioning with the imputation method are weaker than those obtained from the random partitioning method.

The tables indicate that increasing the number of training years through chronological splitting does not significantly enhance prediction accuracy. R-squared values remain relatively stable, and improvements in error metrics are minimal across models. Random Forest consistently performs best, but the gains are marginal as more years are added.

Adding more historical data does not improve performance due to changing market conditions. Advanced time-series techniques may be needed to better capture temporal dependencies for more accurate predictions.

6.5.2 NI-CS-GCC Model Evaluation Results

Additional models were implemented without data imputation to assess whether prediction efficiency is impacted over cumulative years when missing values are left unfilled.

Only XGBoost and CatBoost were applied to these non-imputed models, with their optimal hyperparameters provided in Appendix 6 and the performance metrics in Table 6.10.

Table (6.10): Evaluation Results for NI-CS-GCC Models

Period-1	R-squared	MAE	MSE	RMSE
XGBoost	0.373	116.588	50446.430	224.603
CatBoost	0.377	118.936	50097.448	223.825
Period-2				
XGBoost	0.377	117.044	50109.529	223.852
CatBoost	0.413	114.417	47218.368	217.298
Period-3				
XGBoost	0.344	118.822	52793.863	229.769
CatBoost	0.399	114.519	48326.036	219.832
Period-4				
XGBoost	0.360	118.799	51461.028	226.850
CatBoost	0.405	114.765	47899.892	218.860
Period-5				
XGBoost	0.378	122.049	50027.218	223.667
CatBoost	0.350	128.469	52302.488	228.697

The results indicate improved prediction performance compared to I-CS-GCC imputation models; however, the models are still classified as weak-performing in terms of prediction accuracy.

The SHAP explainer will not be deployed to interpret the chronological splitting models due to their relatively poor performance compared to other models.

6.6 Discussion

This section discusses the performance of the four proposed modeling approaches, evaluating them based on DM techniques, methods for handling missing values, sampling strategies, train-test data splitting, and variable importance analysis. The discussion interprets the results and highlights each approach's findings.

6.6.1 DM Techniques

Using a financial dataset sourced from the LSEG Workspace, this study applied seven DM techniques, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost, to predict firm profitability, measured by the ROI, across the GCC region. Among these models, the ensemble-based CatBoost algorithm consistently demonstrated superior predictive performance, achieving the highest R-squared values and, in most cases, the lowest prediction errors, except for the Oman-specific model.

CatBoost's outperformance can be attributed to several of its distinctive features. Unlike traditional gradient boosting algorithms, CatBoost natively handles categorical variables without requiring extensive preprocessing. It employs ordered boosting, a technique designed to mitigate target leakage, thereby enhancing model generalizability. Additionally, CatBoost is well-suited for small datasets (Jabeur et al., 2021). These attributes likely contributed to its robust performance in this study.

Regarding similar studies conducted in the GCC region, Assous (2022) concluded that CHAID, a decision tree-based model, outperformed other techniques in predicting the efficiency of Saudi banks, achieving the highest R-squared values in both the training and testing phases. Additionally, J. P. Li et al. (2020) suggested that for stressed banks, both RF and CART models should be utilized to facilitate more accurate and informed risk assessments, particularly for analyzing credit risk and predicting credit ratings of banks in the GCC.

Furthermore, the findings are consistent with prior research. For instance, Khalil et al. (2022) reported CatBoost's superiority in regression tasks involving solvency ratio prediction, while Jabeur et al. (2021) highlighted its effectiveness in classification tasks such as bankruptcy prediction. Together, these results reinforce the algorithm's reliability in financial prediction contexts.

6.6.2 Handling Missing Values Methods

Table 6.11 compares performance results between the I-RS-GCC and NI-RS-GCC CatBoost models, utilizing random splitting for all GCC data. The findings suggest a trade-off between the amount of variance explained and the accuracy of predictions.

Table (6.11): Evaluation Results of CatBoost I-RS-GCC vs NI-RS-GCC Models

Method	R-squared	MAE	MSE	RMSE
Imputation	0.482	69.886	16720.209	129.307
Non-Imputation	0.542	86.471	30496.782	174.633

The non-imputation model has a higher R-squared (0.542), indicating a better ability to explain variability in ROI, likely due to capturing natural data patterns, including missing values. However, it also shows higher prediction errors (MAE 86.471, MSE 30496.782, RMSE 174.633), suggesting that missing data affects the model's precision. In contrast, the imputation model produces lower errors (MAE 69.886, MSE 16720.209, RMSE 129.307), indicating more accurate predictions. However, its R-squared value (0.482) is lower, indicating that it explains a smaller portion of the target variable's variance. In conclusion, while CatBoost performs best in both cases, the non-imputation model is preferable for explaining variance, whereas the

imputation model is better suited for achieving higher prediction accuracy. The choice between these approaches depends on the primary objective of stakeholder analysis.

6.6.3 Sampling Methods

The prediction performance of the CatBoost models for individual GCC countries shows notable variations in predictive accuracy and the ability to explain the variance in ROI, as shown in Table 6.12.

Table (6.12): Evaluation Results for the CatBoost in C and NI-CS-GCC Models

	R-squared	MAE	MSE	RMSE
Bahrain	0.781	12.110	531.042	23.044
Kuwait	0.577	33.155	4881.231	69.866
Oman *	0.647	175.742	120872.316	347.667
Qatar	0.773	43.733	7003.996	83.690
Saudi Arabia	0.328	223.745	183885.537	428.819
UAE	0.586	58.285	14682.792	121.173
GCC	0.542	86.471	30496.782	174.633
Note: * XGBoost outperforms CatBoost in the Oman model				

The Bahrain prediction model performs the best, with the highest R-squared (0.781), indicating a strong ability to explain variance in Bahraini firms' ROI. It also has the lowest error metrics, with MAE of 12.110, MSE of 531.042, and RMSE of 23.044, making it the most accurate prediction model among the GCC countries. The Qatar prediction model also performs well, with a relatively high R-squared (0.773) and moderate error values (MAE of 43.733, MSE of 7003.996, and RMSE of 83.690), suggesting a good balance between variance explanation and prediction accuracy. The prediction models for Oman, Kuwait, and the UAE demonstrate moderate performance, with R-

squared values of approximately 0.609 and 0.586. Oman exhibits the highest MAE (177.930), followed by the UAE (58.285) and Kuwait (33.155). The relatively high MAE and RMSE values indicate that these models produce less precise predictions than those for Bahrain and Qatar. The Saudi Arabia model demonstrates the weakest prediction performance, with a low R-squared (0.328) and the highest error metrics (MAE: 223.745, MSE: 183,885.537, RMSE: 428.819). This suggests that the model struggles to capture variance, resulting in the least accurate predictions. Overall, the GCC model achieves moderate prediction accuracy, with an R-squared of 0.542 and error metrics of MAE (86.471), MSE (30,496.782), and RMSE (174.633).

In conclusion, the Bahrain and Qatar prediction models demonstrate the strongest prediction performance, while the Saudi Arabia prediction model shows significant room for improvement. The differences in prediction performance across GCC countries may reflect varying levels of data quality, market conditions, or variable relevance in the models for each country. The Bahrain prediction model likely performs best due to its higher data quality, greater variable relevance, and a more stable, homogeneous market, which makes it easier for the model to capture underlying patterns and make accurate predictions. In contrast, the Saudi Arabia prediction model performs weaker due to the complexity and variability of its larger, more diverse economy, potential data quality issues, and less relevant variables, which make it harder for the model to generalize and accurately predict ROI.

6.6.4 Data Splitting Methods

Table 6.13 compares the performance metrics of XGBoost and CatBoost models, using both random and chronological splitting, with and

without imputation. Random split models outperform the chronological split models for the GCC dataset in both XGBoost and CatBoost, non-imputed with R-squared (0.533) and (0.542).

Table (6.13): Evaluation Results for I-RS-GCC, NI-RS-GCC, I-CS-GCC and NI-CS-GCC Models

	R-squared		MAE		MSE		RMSE	
	<u>Imputed</u>	<u>Non-Imp.</u>	<u>Imputed</u>	<u>Non-Imp.</u>	<u>Imputed</u>	<u>Non-Imp.</u>	<u>Imputed</u>	<u>Non-Imp.</u>
Random Split								
XGBoost	0.413	0.533	75.444	88.470	18932.636	31084.407	137.596	176.308
CatBoost	0.482	0.542	69.886	86.471	16720.209	30496.782	129.307	174.633
Chronological Split								
Period-1								
XGBoost	0.222	0.373	89.667	116.588	28327.114	50446.430	168.307	224.603
CatBoost	0.235	0.377	90.227	118.936	27798.110	50097.448	166.728	223.825
Period-2								
XGBoost	0.201	0.377	90.896	117.044	29040.263	50109.529	170.412	223.852
CatBoost	0.206	0.413	90.654	114.417	28847.461	47218.368	169.854	217.298
Period-3								
XGBoost	0.201	0.344	88.609	118.822	28360.949	52793.863	168.407	229.769
CatBoost	0.206	0.399	90.843	114.519	28970.695	48326.036	170.208	219.832
Period-4								
XGBoost	0.191	0.360	92.533	118.799	29370.538	51461.028	171.378	226.850
CatBoost	0.170	0.405	94.134	114.765	30163.537	47899.892	173.676	218.860
Period-5								
XGBoost	0.226	0.378	93.670	122.049	28105.656	50027.218	167.656	223.667
CatBoost	0.216	0.350	96.104	128.469	28475.226	52302.488	168.746	228.697
Note: Non-Imp.: Non-Imputed Model								

Furthermore, the comparison shows that increasing the number of training years through chronological splitting does not improve model performance in XGBoost and CatBoost. While CatBoost shows improvement in period 2, performance declines in periods 3, 4, and 5. The error metrics also exhibit no consistent improvement, indicating that adding more years beyond the second period does not enhance the R-squared value. Instead, performance

deteriorates in the period 5 model (19 training years), with worsening error metrics, highlighting the diminishing returns of extending the training period.

The performance metrics for chronological splitting, with and without imputation, indicate relatively weak predictive power across all GCC firms, as shown by low R-squared values and high error metrics. CatBoost generally outperforms XGBoost. This aligns with Jabeur et al. (2021), who found that CatBoost was the top performer in predicting firm failure up to three years in advance. They also observed a decline in accuracy with longer time horizons. However, while their CatBoost classifier excelled in predicting French firm failures, the current study's CatBoost regressor shows weaker performance in forecasting the profitability of GCC firms. The weak performance of these models is mainly due to issues with the chronological data split. This approach introduces chronological dependencies that models struggle to capture, exposing them to shifts in market or economic conditions over time. As a result, the models often overfit past trends, failing to generalize well to future data, which explains the low R-squared values. The high error metrics reflect their limited predictive power, indicating difficulty in handling the complexity and chronological shifts in the data.

In conclusion, using chronological data split for training and testing the GCC models in this context appears to be a limiting factor, resulting in low R-squared values and high prediction errors. The chronological nature of the data, combined with the evolving economic conditions in the GCC, likely contributes to the weak performance of these models. Therefore, alternative methods, such as incorporating time-based variables or more advanced time-series models, may be more suitable for enhancing predictive accuracy.

6.6.5 Variable Importance

Table 6.14 shows the variable importance rankings for the C and NI-RS-GCC models that utilize the non-imputation method. The comparison presents the top three variables influencing profitability for each GCC country, as well as the overall model for the GCC region.

The variable importance rankings for individual GCC models reveal country-specific differences in the drivers of ROI.

Table (6.14): Variable Importance Rankings Using CatBoost for C and NI-CS-GCC Models

	1 st important	2 nd important	3 rd important
Bahrain	AT (+)	Size (+)	P/E (+)
Kuwait	ICR (+)	P/E (-)	AT (+)
Oman*	TDPTE (+)	ICR (-)	P/E (-)
Qatar	CaR (+)	CAT (+)	AT (+)
Saudi Arabia	ICR (+)	AT (+)	IT
UAE	AT (+)	CaR (+)	IT
GCC	AT (+)	P/E (-)	CAT (+)

Notes:

* Oman variable importance plotted using the XGBoost Model because it outperforms other models

(+): variable contributes positively to ROI, (-): variable contributes negatively to ROI

The findings indicate that Asset Turnover is the most significant variable positively influencing ROI among publicly listed companies in Bahrain, the UAE, and the GCC models. It also ranked as the second most influential factor in Saudi Arabia and third in Kuwait and Qatar. This highlights the crucial role of asset utilization in enhancing firm profitability across the GCC region. Such insights underscore the need for companies to refine their asset management strategies to optimize overall financial performance. The current study's findings are consistent with the study of

Zhou et al. (2022), who observed that firms with low asset utilization ratios often exhibited profitability concerns, with a significant proportion reporting negative net profits. Moreover, it contradicts Nguyen et al. (2023), which highlights that firms with high Asset Turnover are more likely to face bankruptcy. Furthermore, it does not support Barboza et al. (2017), who found that asset utilization had a limited correlation with financial variables related to bankruptcy.

The Interest Coverage Ratio indicates a firm's ability to cover its interest expenses. It is notably the highest in positively influencing the profitability of publicly listed Kuwaiti and Saudi companies and ranked second in influencing Omani firms. This finding aligns with Jabeur et al. (2021), who identified solvency as the most critical factor in predicting corporate failure, reporting a positive correlation between solvency ratios and default probability. Likewise, Arora & Saurabh (2022) ranked solvency as the fourth most significant predictor of bankruptcy in Indian firms.

Regarding conventional and Islamic banks in Saudi Arabia, Assous (2022) concluded that their efficiency is highly influenced by profitability and that capital ratios are the best predictors of cost efficiency for Saudi banks.

In Qatar's publicly listed companies, liquidity, measured by the Cash Ratio, emerges as a critically influential factor positively impacting profitability. Liquidity reflects a firm's ability to generate cash flows from operations, thereby enabling it to meet short-term financial obligations. This finding aligns with the findings of Jabeur et al. (2021), who analyzed various liquidity ratios, including the Current Ratio and Cash Ratio, to predict corporate failures. They concluded that non-distressed firms exhibit higher

average liquidity, while failed firms struggle to meet their debt obligations. Additionally, it supports the study of Halteh et al. (2018), which concluded that liquidity is one of the most predictive variables for financial distress in Islamic banks.

In Omani publicly listed companies, leverage presented by Total Debt Percentage of Total Equity contributes the most positively to profitability. This finding is consistent with previous research that has identified leverage as a critical factor in predicting firm performance (Garcia, 2022; Halteh et al., 2018; P. V. H. Son & Duong, 2024).

This variation in variable importance regarding firm-specific factors among GCC countries highlights the distinct economic conditions and key drivers of firm performance across the GCC countries.

Furthermore, the examined macroeconomic factors, real GDP and inflation rate, have had a weak influence on the profitability of GCC firms. This contradicts Li et al. (2020), which found macroeconomic factors significant in investment score ratings.

Additionally, control variables, firm Size, Age, and Industry Type, had a moderate impact on the profitability of most prediction models in general, and were significant in limited models. At the same time, the Country has a moderate effect on the NI-RS-GCC model.

The effect of firm Size in Bahraini firms is consistent with previous research by Garcia (2022), which further demonstrates that smaller firms, as measured by market capitalization, face a higher risk of bankruptcy, while larger firms with higher market capitalization reduce bankruptcy risk.

Similarly, Chiadamrong & Wattanawarangkoon (2023) found that larger firms typically perform better, especially those seeking to recover from underperformance.

Firm Age had a negative impact on the profitability of publicly listed companies in the GCC. This finding contradicts Balasubramanian et al.'s (2019) finding that younger listed firms may face greater financial instability, whereas mature firms with established reputations experience greater financial stability.

Finally, regarding IT, Garcia (2022) found that industry does not impact prediction outcomes, contradicting the current study's findings that Industry Type has a moderate effect on profitability.

When comparing the current study to similar research in the region, J. P. Li et al. (2020) found that, for GCC banks with the highest credit quality, macroeconomic factors, such as economic, institutional, and fiscal strength, play a critical role. Additionally, a concentrated industry structure was shown to offer a competitive advantage to larger banks, thereby enhancing their credit quality. At the bank-specific level, franchise strength and earnings power were identified as key determinants of overall banking performance.

6.7 Summary

This chapter presented the study's results and discussed the findings from four proposed DM modeling approaches used to assess the accuracy of profitability prediction for GCC firms. The analysis considered various DM techniques, dataset sampling strategies, missing value handling methods, and train-test

splitting approaches. Additionally, the chapter compared the results of variable importance.

Ensemble DM techniques outperformed single DM techniques in terms of R-squared and most error metrics. CatBoost consistently outperforms all proposed models except for the Omani model; XGBoost outperforms CatBoost. Additionally, random split models, such as I-RS-GCC, NI-RS-GCC, and C, outperform the chronological split in predicting the profitability of GCC firms.

Non-imputed models outperformed imputed models, with Bahrain and Qatar achieving the highest prediction accuracy compared to the overall GCC models (I-RS-GCC and NI-RS-GCC) and other countries in the C models. Saudi Arabia's prediction model had the lowest prediction performance among GCC single models.

Control variables moderately impacted GCC firms' profitability, while macroeconomic variables had minimal impact, likely due to weak direct relationships with firm-level factors.

CHAPTER 7: CONCLUSION

7.1 Introduction

The primary objective of this thesis was to investigate the accuracy of DM techniques in predicting the financial performance of publicly listed GCC companies, as measured by profitability, with the aim of supporting firm sustainability for the benefit of stakeholders.

The study implemented four predictive modeling approaches utilizing seven DM regression techniques, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost, aiming to optimize prediction performance. The research began by collecting financial data from GCC publicly listed companies via LSEG Workspace and macroeconomic data from WDI, followed by data preprocessing and the implementation of imputation models, non-imputation models for the all-GCC dataset, then single-country models, and lastly chronological models. After all, a comparison was conducted to identify the best-performing regressor among the implemented models and to identify the most influential variables for each model.

Chapter 2 explored firm performance through theoretical perspectives, emphasizing Return On Investment as a key profitability measure. It reviewed DM techniques in predictive modeling and combined bibliometric analysis with SLR to map research gaps addressing the first research objective. The chapter introduced a conceptual framework outlining variable relationships in GCC firms, forming the basis for predictive modeling. This study addresses a

literature gap by testing Return On Investment as a dependent variable to measure profitability using DM techniques. It will also involve utilizing new measures as independent variables (e.g., leverage: Total Assets to Common Equity, Total Assets to Total Equity; asset utilization: Net Operating Asset Turnover, Fixed Asset Turnover; market value: Price-to-Earnings, Price-to-Sales, Price-to-Book Ratios) for profitability prediction.

Chapter 3 introduced the evolution and performance of GCC financial markets, highlighting statistics on listed firms and recent market activity. It also examined the impact of global economic conditions and regulatory frameworks on GCC stock markets, highlighting ongoing efforts by authorities to enhance supervisory, regulatory, and legislative frameworks.

Chapter 4 explained the research methodology, which follows an objectivist ontological, scientific positivist epistemological paradigm. Implementing the predictive DM models starting by collecting quantitative data from LSEG Workspace and identifying the research variables. Then it explained data preprocessing, data sampling and splitting. Later, it presented the construction of four DM prediction models built using seven DM techniques, single and ensemble, LR, ANN, RT, kNN, RF, XGBoost, and CatBoost.

Chapter 5 presented the analysis results of the GCC dataset before and after data preprocessing for four predictive models (I-RS-GCC, NI-RS-GCC, C, and CS-GCC). This addressed the research's objective two. Additionally, it described the constructed models based on data preprocessing, sampling, and splitting methods, which addressed objective four.

Chapter 6 presented the prediction models' results and discussed the accuracy of various models and their variable importance. It addressed the third and fifth research objectives.

The theoretical and managerial implications of the findings will be further explored in the following sections.

7.2 Theoretical Implications

This research contributes to the literature on firm performance prediction within the GCC context by demonstrating the effectiveness of both single and ensemble DM models. Additionally, it uniquely examines Return On Investment as a dependent variable in predicting firm performance, a first in the field. This section highlights the theoretical contributions of implemented DM models and the relevant research variables.

First, the study demonstrates that CatBoost is the best-performing technique for predicting Return On Investment across the GCC region and within individual countries, except for the Omani model, which aligns with the findings of Khalil et al. (2022). This reinforces the theoretical understanding that ensemble models, such as CatBoost and XGBoost, are superior to traditional single models for financial prediction tasks, as they handle complex, high-dimensional datasets more effectively, thereby improving prediction accuracy and reliability, aligning with (Ainan et al., 2024; Arora & Saurabh, 2022; Barboza et al., 2017; du Jardin, 2021; Jabeur et al., 2021; Khalil et al., 2022; Nguyen et al., 2023; H. Son et al., 2019).

Second, the study highlights the trade-offs between variance explanation and prediction accuracy, depending on the method used for handling missing values. The non-imputation CatBoost model showed higher R-squared values but with higher error rates, while the imputation approach reduced errors at the cost of lower variance explanation. This contributes to theoretical discussions on balancing these trade-offs and provides direction for future research to address this issue.

Third, the study shows notable differences in model performance across various GCC countries, highlighting the impact of market conditions, data quality, and the relevance of financial variables. For instance, the model for Bahrain achieved the highest R-squared value of 0.781. In contrast, the model for Saudi Arabia had the lowest performance, with an R-squared value of just 0.328. Meanwhile, the overall GCC model resulted in an R-squared value of 0.542. These findings underscore the importance of developing country-specific models that consider each region's unique economic dynamics and market characteristics.

Fourth, the research highlights the limitations of chronological data splitting into financial prediction tasks for GCC firms. Random split CatBoost models outperform their chronological counterparts for the non-imputed GCC dataset, achieving an R-squared of 0.542 compared to 0.413 (2 years earlier), with lower error values. CatBoost generally excels in the low-performing NI-CS-GCC models, aligning with Jabeur et al. (2021), who identified CatBoost as the top performer in predicting firm failure up to three years in advance. However, the CatBoost regressor in this study performs weaker than Jabeur et al.'s classifier. The poor performance of chronological splits suggests that stock market conditions introduce complexities that standard models fail to

capture, highlighting the need to incorporate more time-based variables or adopt advanced time-series modeling techniques to enhance predictive accuracy. Kim et al.'s (2018) study contradicts current research and found that Data Depth-SVM outperforms other DM techniques in terms of accuracy (0.9655), which declines slightly over the 10 years.

Fifth, regarding the variable importance of firm-specific factors, the research highlights the positive impact of key financial factors on the profitability of GCC publicly listed companies, measured by ROI. Specifically, asset utilization, represented by the Asset Turnover ratio, and solvency, measured by the Interest Coverage Ratio, contribute to higher profitability. This finding strengthens the theoretical understanding of firm profitability drivers in the GCC context. The positive relationship between asset utilization and profitability aligns with Zhou et al. (2022) but contradicts Nguyen et al. (2023). Similarly, the positive influence of solvency on profitability is supported by Jabeur et al. (2021). Moreover, this finding supports the RBV theory by Barney (1991), which emphasizes that a firm's internal resources are key drivers of competitive advantage and profitability, a perspective more relevant to explaining a firm's performance in the GCC stock markets.

Adding to the fifth theoretical contribution, studies from non-GCC emerging markets highlight varied priorities in predicting financial performance. In South Korea, Son et al. (2019) identified liquidity as the most critical indicator for firms listed on the Korea Exchange, aligning with the most influential financial dimension observed among listed firms in Qatar. In Vietnam, P. V. H. Son & Duong (2024) highlighted a strong internal performance link, where the Return On Equity significantly influenced the

Return On Assets . In India, Arora & Saurabh (2022) found that the Market Capitalization-to-Debt ratio is the strongest predictor of financial performance for firms listed on the Bombay Stock Exchange, reflecting a leverage perception emphasis similar to that observed in Omani firms, but with another leverage variable: the Total Debt Percentage of Total Equity. These regional variations reinforce the importance of contextualizing financial models. The prominence of asset utilization and the price-to-earnings ratio in GCC markets illustrates a distinct predictive dynamic that must be acknowledged when designing models or strategies in emerging economies.

Sixth, the research finding that macroeconomic variables, such as GDP and inflation, have a limited influence on Return On Investment contradicts the conclusions of Li et al. (2020) and the macroeconomic theory, particularly the DGE approach by Wickens (2008). The DGE framework suggests that stock markets are sensitive to macroeconomic fluctuations, causing volatility as investors adjust their portfolios.

Seventh, findings also support the stakeholder theory, which provides a framework for evaluating firm performance by considering the interests of all stakeholders rather than focusing solely on maximizing shareholder wealth. By improving Return On Investment prediction, the study enhances investor decision-making, promotes market transparency, and provides regulatory insights, reinforcing the theory's emphasis on balancing stakeholder interests for sustainable firm performance.

Finally, the study highlights the moderate impact of control variables, with a few showing significant effects. Firm size aligns with the findings of Chiadamrong & Wattanawarangkoon (2023) and Garcia (2022), while firm

Age contradicts Balasubramanian et al. (2019). Additionally, the impact of Industry Type contradicts Garcia (2022), who found no effect of Industry Type on prediction accuracy. These findings suggest the need for further research to understand the influence of these variables better.

Overall, the research enhances the theoretical understanding of DM techniques, related financial variables, and data-handling methods in predicting firm performance, particularly within the GCC context.

7.3 Managerial Implications

This research highlights several critical managerial implications for enhancing financial prediction and decision-making, particularly in deploying predictive models to predict ROI.

First, addressing missing data in financial datasets (Section 4.6.1) is crucial for improving prediction accuracy. Effective data management strategies and maintaining accurate financial records can significantly enhance the quality and reliability of predictive models, supporting more accurate financial analysis.

Second, the study reveals key financial metrics driving Return On Investment across GCC countries (Section 6.6.5), emphasizing the importance of developing country-specific models to account for regional economic variations and improve prediction accuracy.

Third, SHAP analysis results (Section 6.4) indicate that external economic factors have a minimal impact on ROI, suggesting that managers should optimize internal processes and resources to enhance firm financial

performance. Deploying predictive models also helps managers identify trends and insights that may not be apparent through traditional analysis, allowing for proactive decision-making and improved financial outcomes.

Finally, SHAP analysis provides transparency in interpreting model outcomes, enabling stakeholders, such as GCC CMAs and bankruptcy commissions, to make informed decisions on investments and risk assessments. These insights collectively guide managers in adopting data-driven strategies for effective financial decision-making and Return On Investment optimization.

CMAs' ongoing studies and research aim to refine and enhance the effectiveness of regulations addressing firm bankruptcy procedures (GCC-STAT, 2024). Figure 7.1 proposes a framework for an early warning system in the GCC for policy regulators in the CMA and bankruptcy commissions. In this system, financial reports of GCC-listed companies are analyzed using the GCC Performance Prediction (GCCPP) model and evaluated in line with CMA regulations. The system will trigger CMA interventions based on Return On Investment risk indicators, determining whether to adjust regulations or provide support to at-risk firms.

Additionally, firm managers can use the Return On Investment prediction model as a decision-support tool to proactively monitor financial health, identify early signs of declining performance, and implement corrective measures, such as restructuring debt, improving operational efficiency, or reallocating resources, before profitability deteriorates further.

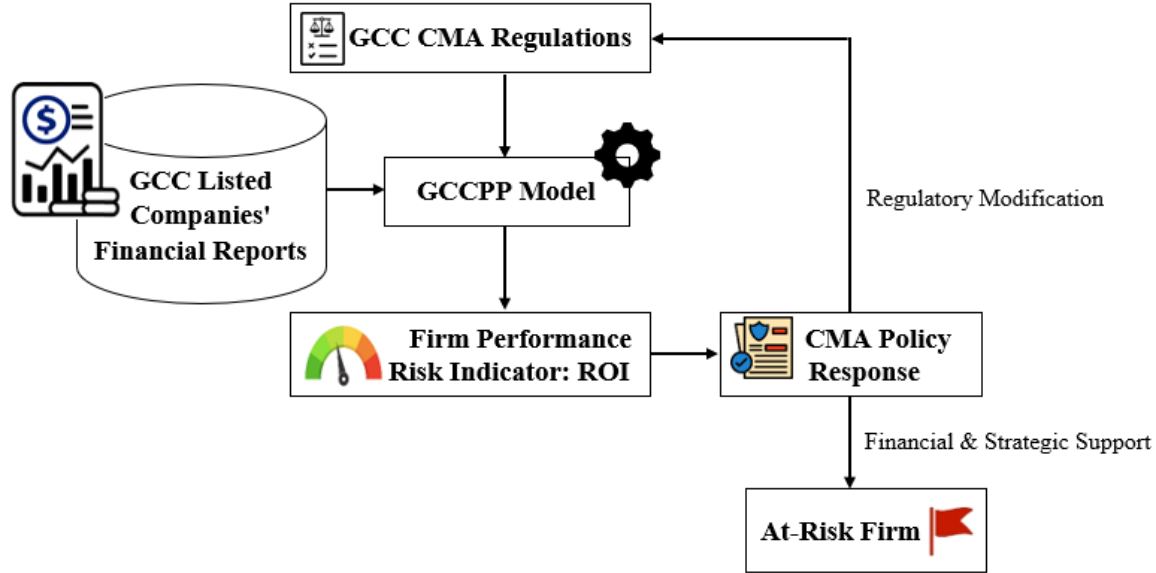


Fig. 7.1: Proposed framework of the early warning system for underperforming GCC-listed companies

7.4 Limitations and Future Research

This research acknowledges several limitations that offer potential directions for future studies aimed at improving the accuracy of performance predictions for listed companies in the GCC region.

The first significant limitation involves the presence of outliers and missing values in the dataset. While outliers and skewness are common characteristics of financial data (H. Son et al., 2019), the GCC dataset obtained from LSEG Workspace demonstrated particularly pronounced skewness and a high incidence of extreme values. To mitigate these issues, Winsorization was applied to handle outliers, and imputation techniques were used to address missing data. However, both methods can introduce bias and potentially distort the underlying structure of the data. This is a critical concern, as the

altered data distribution may affect the model's ability to generalize and accurately capture the relationships among variables. The impact of these preprocessing techniques is especially evident in the prediction results for ROI, where the model consistently underestimates high Return On Investment values and overestimates low ones.

This leads to the second key limitation, the model's performance at the distributional extremes. Such underestimation and overestimation patterns suggest poor performance at the skewed tails, which undermines the overall accuracy and reliability of the predictions. These errors highlight the model's difficulty in handling extreme financial cases, an essential challenge in financial prediction.

The third limitation is related to computational constraints. Support Vector Regression (SVR), a technique known for its accuracy in small to medium-sized datasets, was excluded from this study due to its high computational demands and the time-intensive process required for optimal parameter tuning. The unavailability of adequate computational resources limited the scope of model experimentation and may have restricted the identification of potentially better-performing algorithms.

The fourth limitation stems from the study's exclusive reliance on quantitative secondary data obtained from LSEG Workspace and WDI. Although these sources offer a wide coverage of firm-level and macroeconomic data, this dependence restricts the depth of analysis. Important qualitative factors, such as management quality, market sentiment, or regulatory impacts, were not incorporated, potentially omitting critical explanatory variables that affect firm performance.

The fifth and final limitation concerns the modeling of the temporal dimension. While the study examined the effect of time by applying chronological prediction models, the findings revealed a noticeable drop in accuracy over time. This temporal limitation suggests that the models may not fully capture dynamic patterns or structural changes across time, reducing their robustness for long-term predictions.

To address these limitations, several recommendations can be proposed for future research. First, researchers could consider using alternative financial data sources with fewer missing values, such as Bloomberg, to import and test the GCC dataset. Comparing prediction accuracy and variable importance across different data providers would provide a more comprehensive evaluation of model robustness.

Second, although SHAP was employed to interpret the model, future work could delve deeper into the behavior of SHAP values, specifically at the extremes of Return On Investment. This would help uncover how individual variables influence predictions in those critical regions and could support the engineering of more sensitive features tailored to high- and low-Return On Investment cases. Moreover, stratifying the dataset into ROI bands (e.g., low, medium, high) during the training and testing phases may help address data imbalance and enhance predictive performance across the full range of Return On Investment values.

Third, future studies could investigate the effectiveness of SVR and assess its effectiveness in this context by leveraging more powerful computing resources. A comparative analysis between SVR and other models may offer new insights into the prediction capabilities specific to the GCC region.

Additionally, integrating qualitative textual data extracted from financial reports of publicly listed GCC companies could enrich the dataset and provide a more comprehensive understanding of firm performance.

Lastly, applying advanced time series prediction models, such as Long Short-Term Memory (LSTM) Neural Networks, could address the limitations of static models in capturing temporal dependencies. These models may help improve accuracy over time and better reflect the evolving financial dynamics in the region.

Furthermore, future research could explore a classification-based approach by categorizing listed firms into distinct performance classes. For instance, firms could be grouped into high, moderate, and low performance. Applying classification models to all publicly listed GCC companies from 2003 to 2022 could provide complementary insights and enhance decision-making support when combined with the current regression-based analysis.

7.5 Summary

This chapter concludes the research by presenting theoretical and managerial contributions to the field of predicting firms' financial performance using DM techniques in the context of the GCC region. Furthermore, it discusses the research limitations and possible future research directions.

The main theoretical implication is that the study reveals CatBoost to be an effective DM technique in predicting GCC profitability, particularly in country-specific models such as those for Bahrain and Qatar. Additionally, firm-specific factors in publicly listed companies in the GCC were found to

significantly influence profitability, specifically asset utilization, while macroeconomic variables have a minimal impact.

Furthermore, the research's findings underscore the necessity of country-specific DM models to capture economic variations across GCC nations. Its findings hold practical significance for policymakers, regulators, managers, and investors, aiding in strategic decision-making, early risk detection, and financial optimization. Ultimately, this research advances the application of DM in financial analytics, bridging the gap between theory and real-world decision-making.

The limitations of this study open avenues for future research to enhance financial prediction models. Exploring alternative data sources, integrating qualitative data, and applying advanced techniques such as SVR and LSTM could enhance accuracy and provide deeper insights into the financial performance of GCC firms.

REFERENCES

- Abramson, B., & Finizza, A. (1991). Using belief networks to forecast oil prices. In *International Journal of Forecasting* (Vol. 7).
- Adam, E. E. (1994). Alternative quality improvement practices and organization performance. *Journal of Operations Management*, 12(1), 27–44.
- Ainan, U. H., Por, L. Y., Chen, Y. L., Yang, J., & Ku, C. S. (2024). Advancing Bankruptcy Forecasting With Hybrid Machine Learning Techniques: Insights From an Unbalanced Polish Dataset. *IEEE Access*, 12, 9369–9381. <https://doi.org/10.1109/ACCESS.2024.3354173>
- Alareeni, B. A., & Hamdan, A. (2020). ESG impact on performance of US S&P 500-listed firms. *Corporate Governance (Bingley)*, 20(7), 1409–1428. <https://doi.org/10.1108/CG-06-2020-0258>
- Aljawazneh, H., Mora, A. M., Garcia-Sanchez, P., & Castillo-Valdivieso, P. A. (2021). Comparing the Performance of Deep Learning Methods to Predict Companies' Financial Failure. *IEEE Access*, 9, 97010–97038. <https://doi.org/10.1109/ACCESS.2021.3093461>
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of FINANCE*, 23(4), 589–609.
- Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, 237, 242–254. <https://doi.org/10.1016/j.neucom.2016.12.009>
- Ansari, S., Nassif, A. B., Mahmoud, S., Majzoub, S., Almajali, E., Jarndal, A., Bonny, T., Alnajjar, K. A., & Hussain, A. (2024). Impact of Outliers on Regression and Classification Models: An Empirical Analysis. *Proceedings - International Conference on Developments in ESystems Engineering, DeSE*, 211–218. <https://doi.org/10.1109/DeSE63988.2024.10912020>
- Argaam. (2025, January 2). *How did GCC equity markets fare in 2024?* Argaam Special. <https://arg.am/3BC2F259>

- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Arora, P., & Saurabh, S. (2022). Predicting distress: a post Insolvency and Bankruptcy Code 2016 analysis. *Journal of Economics and Finance*, 46(3), 604–622. <https://doi.org/10.1007/s12197-022-09582-y>
- Assous, H. F. (2022). Prediction of Banks Efficiency Using Feature Selection Method: Comparison between Selected Machine Learning Models. *Complexity*, 2022. <https://doi.org/10.1155/2022/3374489>
- Atje, R., & Jovanovic, B. (1993). Stock markets and development. *European Economic Review*, 37(2–3), 632–640.
- Ayoush, M. D., Toumeh, A. A., & Shabaneh, K. I. (2021). Liquidity, leverage, and solvency: what affects profitability of industrial enterprises the most? *Investment Management and Financial Innovations*, 18(3), 249–259. [https://doi.org/10.21511/imfi.18\(3\).2021.22](https://doi.org/10.21511/imfi.18(3).2021.22)
- Balasubramanian, S. A., Radhakrishna, G. S., Sridevi, P., & Natarajan, T. (2019). Modeling corporate financial distress using financial and non-financial variables: The case of Indian listed companies. *International Journal of Law and Management*, 61(3–4), 457–484. <https://doi.org/10.1108/IJLMA-04-2018-0078>
- Bankruptcy Commission. (2018). *Saudi Bankruptcy law*. Bankruptcy Commission. <https://bankruptcy.gov.sa/>
- Barak, S., Arjmand, A., & Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market. *Information Fusion*, 36, 90–102. <https://doi.org/10.1016/j.inffus.2016.11.006>
- Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
- Barney, J. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, 17(1), 99–120.
- Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 71–111.

- Bititci, U. S. (2015). Overview of Popular Performance Measurement Models and Frameworks. In *Managing Business Performance* (pp. 251–262). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119166542.app1>
- Bley, J., & Chen, K. H. (2006). Gulf Cooperation Council (GCC) stock markets: The dawn of a new era. *Global Finance Journal*, 17(1), 75–91. <https://doi.org/10.1016/j.gfj.2006.06.009>
- Borgonovo, E., Plischke, E., & Rabitti, G. (2024). The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *European Journal of Operational Research*, 318(3), 911–926. <https://doi.org/10.1016/j.ejor.2024.06.023>
- Boursa Kuwait. (2024, December). *Delisted Companies*. Boursa Kuwait. <https://www.boursakuwait.com.kw/en/securities/company-information/delisted-companies>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.
- Brigham, E. F., & Ehrhardt, M. C. (2014). *Financial Management: Theory and Practice* (13th ed.). South-Western Cengage Learning.
- Brower, J., & Dacin, P. A. (2020). An Institutional Theory Approach to the Evolution of the Corporate Social Performance – Corporate Financial Performance Relationship. *Journal of Management Studies*, 57(4), 805–836. <https://doi.org/10.1111/joms.12550>
- Busmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, 57(1), 203–216. <https://doi.org/10.1007/s10614-020-10042-0>
- Callen, T. (2020). *Gross Domestic Product: An Economy's All*. <https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm>
- Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance*, 61, 304–323. <https://doi.org/10.1016/j.iref.2018.03.008>

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chiadamrong, N., & Wattanawarangkoon, T. (2023). Financial performance prediction model based on firms' internal capability determinants: evidence from listed firms in Thailand during the transition period of going public. *Cogent Engineering*, 10(1). <https://doi.org/10.1080/23311916.2023.2216860>
- Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. *Journal of Business Research*, 101, 885–896. <https://doi.org/10.1016/j.jbusres.2018.11.015>
- Collis, J., & Hussey, R. (2021). *BUSINESS RESEARCH* (fifth edition). Bloomsbury Publishing.
- Combs, J. G., Crook, T. R., & Shook, C. L. (2005). The Dimensionality of Organizational Performance and its Implications for Strategic Management Research. In *Research methodology in strategy and management* (pp. 259–286). [https://doi.org/10.1016/s1479-8387\(05\)02011-4](https://doi.org/10.1016/s1479-8387(05)02011-4)
- Cooper, D. R., & Schindler, P. (2014). *Business research methods* (Vol. 12). McGraw-hill.
- Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, 40(10), 3970–3983. <https://doi.org/10.1016/j.eswa.2013.01.012>
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- du Jardin, P. (2021). Forecasting corporate failure using ensemble of self-organizing neural networks. *European Journal of Operational Research*, 288(3), 869–885. <https://doi.org/10.1016/j.ejor.2020.06.020>

- Emma Bell, Alan Bryman, & Bill Harley. (2019). *Business Research Methods* (5th ed.). Oxford University Press.
- Erl, T., Khattak, W., & Buhler, P. (2016). *Big data fundamentals: concepts, drivers & techniques*. Prentice Hall Press.
- Esposito, P., Marrone, A., Pontrelli, V., & Rella, A. (2025). Exploring drivers and factors of benefit impact reports quality: Retracing a stakeholder theory approach. *Business Strategy and the Environment*, 34(1), 555–569. <https://doi.org/10.1002/bse.3974>
- Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2), 427–465. <https://doi.org/10.1111/j.1540-6261.1992.tb04398.x>
- Fareed, F., Rezghi, A., & Sandoz, C. (2023). *Inflation Dynamics in the Gulf Cooperation Council (GCC): What is the Role of External Factors?*
- Farooq, U., Tabash, M. I., Hamouri, B., Daniel, L. N., & Safi, S. K. (2023). Nexus between Macroeconomic Factors and Corporate Investment: Empirical Evidence from GCC Markets. *International Journal of Financial Studies*, 11(1). <https://doi.org/10.3390/ijfs11010035>
- Flesher, D. L., & Previts, G. J. (2013). Donaldson Brown (1885-1965): The power of an individual and his ideas over time. *Accounting Historians Journal*, 40(1), 79–101.
- Freeman, R. E. E., & McVea, J. (2005). A Stakeholder Approach to Strategic Management. In *The Blackwell handbook of strategic management* (pp. 183–201). The Blackwell handbook of strategic management. <https://doi.org/10.2139/ssrn.263511>
- Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Garcia, J. (2022). Bankruptcy prediction using synthetic sampling. *Machine Learning with Applications*, 9, 100343. <https://doi.org/10.1016/j.mlwa.2022.100343>

- GCC-STAT. (2024). *Performance of Stock Markets in the GCC Countries and the Regulatory and Legislative Developments during the year 2023- أداء أسواق الأوراق المالية في دول مجلس التعاون والتطورات التنظيمية والتشريعية خلال العام 2023*. <https://www.gccstat.org/ar/statistic/publications/statistical>
- GCC-STAT. (2025). *أداء أسواق الأوراق المالية في دول مجلس التعاون والتطورات التنظيمية والتشريعية خلال النصف الأول من العام 2024*. <https://www.gccstat.org/ar/statistic/publications/statistical-glance>
- Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow* (First Edition). O'Reilly Media, Inc. <http://oreilly.com/catalog/errata.csp?isbn=9781491962299>
- Gimbert, X., Bisbe, J., & Mendoza, X. (2010). The Role of Performance Measurement Systems in Strategy Formulation Processes. *Long Range Planning*, 43(4), 477–497. <https://doi.org/10.1016/j.lrp.2010.01.001>
- Gokmen, S., Dagalp, R., & Kilickaplan, S. (2022). Multicollinearity in measurement error models. *Communications in Statistics - Theory and Methods*, 51(2), 474–485. <https://doi.org/10.1080/03610926.2020.1750654>
- Government of Bahrain. (2008). *From regional Pioneer to global Contender: the economic Vision 2030 for Bahrain*.
- Haghighinasab, M., & Khobbakht, F. (2024). The Impact of Relationship Marketing Dimensions on Financial Indicators of Stock Exchange Market Companies. *Journal of Relationship Marketing*, 24(1), 1–33. <https://doi.org/10.1080/15332667.2024.2367334>
- Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications*, 29(1), 343–358. <https://doi.org/https://doi.org/10.1007/s00521-017-3194-2>
- Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, 128, 139–152. <https://doi.org/10.1016/j.knosys.2017.05.001>

- Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*, 44(6), 759–773. <https://doi.org/10.1108/MF-12-2016-0372>
- Hamann, P. M., & Schiemann, F. (2021). Organizational performance as a set of four dimensions: An empirical analysis. *Journal of Business Research*, 127, 45–65. <https://doi.org/10.1016/j.jbusres.2021.01.012>
- Hantono, H. (2019). Predicting financial distress using Altman score, Grover score, Springate score, Zmijewski score (case study on consumer goods company). *Accountability*, 8(1)(1), 1–16. <https://doi.org/10.32400/ja.23354.8.1.2019.1-16>
- Harrison, J. S., & Wicks, A. C. (2013). Stakeholder Theory, Value, and Firm Performance. *Business Ethics Quarterly*, 23(1), 97–124. <https://doi.org/10.5840/beq20132314>
- Hoornaert, S., Ballings, M., Malthouse, E. C., & Van den Poel, D. (2017). Identifying New Product Ideas: Waiting for the Wisdom of the Crowd or Screening Ideas in Real Time. *Journal of Product Innovation Management*, 34(5), 580–597. <https://doi.org/10.1111/jpim.12396>
- Horrigan, J. O. (1968). A Short History of Financial Ratio Analysis. In *Source: The Accounting Review* (Vol. 43, Issue 2).
- IMF. (2023). *Economic Prospects and Policy Challenges for the GCC Countries*. <https://doi.org/https://doi.org/10.5089/9798400263279.002>
- IMF. (2024). Regional Economic Outlook, Middle East and Central Asia. In *Regional Economic Outlook, Middle East and Central Asia, April 2024*. International Monetary Fund. <https://doi.org/10.5089/9798400272356.086>
- Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. Ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166. <https://doi.org/10.1016/j.techfore.2021.120658>
- Jain, J., & Jain, R. (2020). Impact of Solvency, Liquidity and efficiency on Profitability: A case study of coal India Ltd. *International Bilingual Peer Reviewed Refereed Research Journal Impact*, 7(27)(27), 32–38. <https://www.researchgate.net/publication/350873589>

- Jensen, M. C. (2010). Value Maximization, Stakeholder Theory, and the Corporate Objective Function. *Journal of Applied Corporate Finance*, 22(1), 32–42. <https://doi.org/10.1111/j.1745-6622.2010.00259.x>
- Jiang, Y., & Jones, S. (2018). Corporate distress prediction in China: a machine learning approach. *Accounting and Finance*, 58(4), 1063–1109. <https://doi.org/10.1111/acfi.12432>
- Kasasbeh, F. I. (2021). Impact of financing decisions ratios on firm accounting-based performance: evidence from Jordan listed companies. *Future Business Journal*, 7(1). <https://doi.org/10.1186/s43093-021-00061-0>
- KDIPA. (2020). *Kuwait Economic Laws*. <https://kdipa.gov.kw/invest-in-kuwait/economic-laws/>
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.
- Khalil, A. A., Liu, Z., Salah, A., Fathalla, A., & Ali, A. (2022). Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques. *IEEE Access*, 10, 117304–117314. <https://doi.org/10.1109/ACCESS.2022.3210032>
- Kim, S., Mun, B. M., & Bae, S. J. (2018). Data depth based support vector machines for predicting corporate bankruptcy. *Applied Intelligence*, 48(3), 791–804. <https://doi.org/10.1007/s10489-017-1011-3>
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. <https://www.researchgate.net/publication/302924724>
- Kraus, M., & Feuerriegel, S. (2017a). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38–48. <https://doi.org/10.1016/j.dss.2017.10.001>
- Kraus, M., & Feuerriegel, S. (2017b). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38–48. <https://doi.org/10.1016/j.dss.2017.10.001>

- Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. In *Expert Systems with Applications* (Vol. 197). Elsevier Ltd. <https://doi.org/10.1016/j.eswa.2022.116659>
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 1.
- Lebas, M. J. (1995). Performance measurement and performance management. *International Journal of Production Economics*, 41(1–3), 23–35.
- Ledhem, M. A. (2022). Data mining techniques for predicting the financial performance of Islamic banking in Indonesia. *Journal of Modelling in Management*, 17(3), 896–915. <https://doi.org/10.1108/JM2-10-2020-0286>
- Legislation and Legal Opinion Commission. (2022). *Reorganization and Bankruptcy Law*. <https://www.legalaffairs.gov.bh/Publications/Download/059.pdf>
- Li, A. W., & Bastos, G. S. (2020). Stock market forecasting using deep learning and technical analysis: A systematic review. In *IEEE Access* (Vol. 8, pp. 185232–185242). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ACCESS.2020.3030226>
- Li, J. P., Mirza, N., Rahat, B., & Xiong, D. (2020). Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, 161. <https://doi.org/10.1016/j.techfore.2020.120309>
- Li, Z., Crook, J., & Andreeva, G. (2017). Dynamic prediction of financial distress using Malmquist DEA. *Expert Systems with Applications*, 80, 94–106. <https://doi.org/10.1016/j.eswa.2017.03.017>
- Lipták, K. , H. N. , M. K. S. , & M. Z. (2022). Innovative Financial Indicators: Marketing ROI. *Eurasian Business and Economics Perspectives: Proceedings of the 35th Eurasia Business and Economics Society Conference*, 137–147.
- LSEG. (2023). *LSEG Screener [GCC Public Companies Dataset]*. <https://www.lseg.com>

- Lu, V. N., Wirtz, J., Kunz, W. H., Paluch, S., Gruber, T., Martins, A., & Patterson, P. G. (2020). Service robots, customers and service employees: what can we learn from the academic literature and where are the gaps? In *Journal of Service Theory and Practice* (Vol. 30, Issue 3, pp. 361–391). Emerald Group Holdings Ltd. <https://doi.org/10.1108/JSTP-04-2019-0088>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
- Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2), 743–758. <https://doi.org/10.1016/j.ejor.2018.10.024>
- Martin-del-Brio, B., & Serrano-Cinca, C. (1993). Self-organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases. In *Neural Comput & Applic* (Vol. 1).
- Mashey, J. R. (1999, June 6). Big data and the next wave of InfraStress problems, solutions, opportunities. *1999 USENIX Annual Technical Conference*.
- Melvin, J., Boehlje, M., Dobbins, C., & Gray, A. (2004). The DuPont profitability analysis model: An application and evaluation of an e-learning tool. *Agricultural Finance Review*, 64(1), 75–89. <https://doi.org/10.1108/00214660480001155>
- Ministry of Economy. (2016, September 29). *Federal Decree Law No (9) of 2016 on Bankruptcy*. UAE Ministry of Economy. <https://www.moec.gov.ae/en/federal-decree-law-no-9-of-2016-on-bankruptcy>
- Ministry of Justice and Islamic Affairs Bahrain. (2021, January). *Bankruptcy cases*. Ministry of Justice and Islamic Affairs Bahrain. <https://www.moj.gov.bh/ar/communication-guide-arb-6>
- MSX. (2024). *2023 Annual Bulletin*. www.msx.om

- Mulisa, F. (2022). When Does a Researcher Choose a Quantitative, Qualitative, or Mixed Research Approach? *Interchange*, 53(1), 113–131. <https://doi.org/10.1007/s10780-021-09447-z>
- Nguyen, H. H., Viviani, J. L., & Ben Jabeur, S. (2023). Bankruptcy prediction using machine learning and Shapley additive explanations. *Review of Quantitative Finance and Accounting*. <https://doi.org/10.1007/s11156-023-01192-x>
- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. In *Source: Journal of Accounting Research* (Vol. 18, Issue 1).
- Oner, C. (2012). *Inflation: Prices on the rise*. <https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Inflation>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. In *PLoS Medicine* (Vol. 18, Issue 3). Public Library of Science. <https://doi.org/10.1371/JOURNAL.PMED.1003583>
- Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of Machine Learning Models for Bankruptcy Prediction. *IEEE Access*, 9, 124887–124899. <https://doi.org/10.1109/ACCESS.2021.3110270>
- Pattitoni, P., Petracci, B., & Spisni, M. (2014). Determinants of profitability in the EU-15 area. *Applied Financial Economics*, 24(11), 763–775. <https://doi.org/10.1080/09603107.2014.904488>
- Phillips, P. P., & Phillips, J. J. (2019). *ROI Basics* (2nd ed.). Association for Talent Development.
- Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, 135, 60–70. <https://doi.org/10.1016/j.eswa.2019.06.014>

- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, 31*. <https://github.com/catboost/catboost>
- Qaboos bin Said. (2019, July). *Royal Decree 53/2019 Promulgating the Bankruptcy Law*. Decree- Omani Legislation Search Engine.
- Refinitiv. (2020). *The Refinitiv Business Classifications*. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/trbc-business-classification-methodology.pdf
- Robinson, T. R., Henry, E., & Broihahn, M. A. (2020). *International Financial Statement Analysis* (4th ed.). John Wiley & Sons.
- Saha, D., Young, T. M., & Thacker, J. (2023). Predicting firm performance and size using machine learning with a Bayesian perspective. *Machine Learning with Applications*, 11, 100453. <https://doi.org/10.1016/j.mlwa.2023.100453>
- Sarkar, D., Bali, R., & Sharma, T. (2018). Practical Machine Learning with Python. In *Practical Machine Learning with Python*. Apress. <https://doi.org/10.1007/978-1-4842-3207-1>
- Saudi Exchange. (2024). *2023 Annual Statistical Report*. <https://www.saudiexchange.sa>
- Sharda, R., Delen, D., & Turban, E. (2021). *Analytics, data science, & artificial intelligence: Systems for decision support* (11th ed.). Harlow: Pearson.
- Smiti, S., & Soui, M. (2020). Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE. *Information Systems Frontiers*, 22(5), 1067–1083. <https://doi.org/10.1007/s10796-020-10031-6>
- Son, H., Hyun, C., Phan, D., & Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 138. <https://doi.org/10.1016/j.eswa.2019.07.033>

- Son, P. V. H., & Duong, L. T. (2024). Research on applying machine learning models to predict and assess return on assets (ROA). *Asian Journal of Civil Engineering*, 25(5), 4269–4279. <https://doi.org/10.1007/s42107-024-01046-4>
- Steinhoff, L., Arli, D., Weaven, S., & Kozlenkova, I. V. (2019). Online relationship marketing. *Journal of the Academy of Marketing Science*, 47(3), 369–393. <https://doi.org/10.1007/s11747-018-0621-6>
- Takahashi, A. R. W., & Araujo, L. (2020). Case study research: opening up research opportunities. In *RAUSP Management Journal* (Vol. 55, Issue 1, pp. 100–111). Emerald Group Holdings Ltd. <https://doi.org/10.1108/RAUSP-05-2019-0109>
- Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*, 39(5), 769–787. <https://doi.org/10.1002/for.2661>
- The Supreme Council. (2024, December 1). *GCC foundation*. <https://www.gcc-sg.org/ar/AboutUs/Pages/FoundingDay.aspx>.
<https://www.gcc-sg.org/ar/AboutUs/Pages/FoundingDay.aspx>
- Tudose, M. B., Rusu, V. D., & Avasilcai, S. (2022). Financial performance – determinants and interdependencies between measurement indicators. *Business, Management and Economics Engineering*, 20(1), 119–138. <https://doi.org/10.3846/bmee.2022.16732>
- Valentowitsch, J., Kindig, M., & Burr, W. (2024). Toward an alternative measure of board diversity: an exploratory study on board polarization in German stock exchange-listed companies. *Baltic Journal of Management*, 19(6), 20–38. <https://doi.org/10.1108/BJM-09-2023-0367>
- Vieira, E. S., Neves, M. E., & Dias, A. G. (2019). Determinants of Portuguese firms' financial performance: panel data evidence. *International Journal of Productivity and Performance Management*, 68(7), 1323–1342. <https://doi.org/10.1108/IJPPM-06-2018-0210>

- Waggoner, D. B., Neely, A. D., & Kennerley, M. P. (1999). The forces that shape organisational performance measurement systems: An interdisciplinary review. *International Journal of Production Economics*, 60, 53–60.
- Wamba-Taguimdje, S. L., Fosso Wamba, S., Kala Kamdjoug, J. R., & Tchatchouang Wanko, C. E. (2020). Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. *Business Process Management Journal*, 26(7), 1893–1924. <https://doi.org/10.1108/BPMJ-10-2019-0411>
- Wang, L., & Wu, C. (2017). Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. *Knowledge-Based Systems*, 121, 99–110. <https://doi.org/10.1016/j.knosys.2017.01.016>
- Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, 112, 258–273. <https://doi.org/10.1016/j.eswa.2018.06.016>
- Wickens, M. (2008). *Macroeconomic Theory: a dynamic general equilibrium approach*. Princeton University Press. press.princeton.edu
- Wilcox, M., & Bourne, M. (2003). Predicting performance. *Management Decision*, 41(8), 806–816. <https://doi.org/10.1108/00251740310496314>
- Wild, J. J., & Shaw, K. W. (2019). *Fundamental Accounting Principles* (24th ed.). McGraw-Hill.
- World Bank. (2024, June 23). *World Development Indicators*. <https://databank.worldbank.org/source/world-development-indicators>
- Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97–107. <https://doi.org/10.1109/TKDE.2013.109>

- Yuchtman, E., & Seashore, S. E. (1967). A System Resource Approach to Organizational Effectiveness. *American Sociological Review*, 32(6), 891–903.
http://www.jstor.orgURL:http://www.jstor.org/stable/2092843http://www.jstor.org/stable/2092843?seq=1&cid=pdf-reference#references_tab_contents
- Zhao, J., Ouenniche, J., & De Smedt, J. (2024). Survey, Classification and Critical Analysis of the Literature on Corporate Bankruptcy and Financial Distress Prediction. *Machine Learning with Applications*, 100527. <https://doi.org/10.1016/j.mlwa.2024.100527>
- Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350–361. <https://doi.org/10.1016/j.neucom.2017.01.026>
- Zhou, M., Liu, H., & Hu, Y. (2022). Research on corporate financial performance prediction based on self-organizing and convolutional neural networks. *Expert Systems*, 39(9). <https://doi.org/10.1111/exsy.13042>
- Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, 28, 41–50. <https://doi.org/10.1007/s00521-016-2304-x>
- Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. In *Studies on Current Econometric Issues in Accounting Research* (Vol. 22).

APPENDICES

Appendix 1. Thesis Dataset

GCC Firms' Dataset Sample (LSEG, 2023)

S	ROI	CuR	QR	CaR	DR	TACE	TATE	TDPTE	ICR	AT	NOAT	CAT	FAT	P/E	P/S	P/B	GDP(bn)	IR	Size	Age	IT	FS	year	Country
1	13.94	1.91	1.10	0.28	0.20	1.86	1.59	0.32	3.72	0.85	1.12	1.87	3.48	Null	0.53	Null	35.92	3.63	18.76	30.00	Consumer Non-Cyclicals	GTFP.BH	2022	Bahrain
2	14.13	1.83	1.09	0.31	0.15	1.73	1.48	0.22	6.10	0.76	1.01	1.75	3.09	Null	0.62	Null	34.25	-0.61	18.65	30.00	Consumer Non-Cyclicals	GTFP.BH	2021	Bahrain
3	17.64	1.83	1.11	0.33	0.17	1.72	1.49	0.25	5.70	0.81	1.05	1.87	3.10	9.70	0.66	0.83	33.36	-2.32	18.62	30.00	Consumer Non-Cyclicals	GTFP.BH	2020	Bahrain
4	17.07	1.67	0.95	0.20	0.17	1.74	1.52	0.26	5.13	0.89	1.16	2.01	3.56	Null	0.66	Null	34.99	1.01	18.61	30.00	Consumer Non-Cyclicals	GTFP.BH	2019	Bahrain
5	23.69	1.96	1.12	0.21	0.08	1.55	1.38	0.11	13.28	0.99	1.33	2.11	4.41	Null	0.65	Null	34.24	2.09	18.47	30.00	Consumer Non-Cyclicals	GTFP.BH	2018	Bahrain
6	23.99	1.95	1.13	0.28	0.12	1.59	1.44	0.17	9.62	1.02	1.33	2.12	4.59	9.76	0.61	0.92	33.54	1.39	18.48	30.00	Consumer Non-Cyclicals	GTFP.BH	2017	Bahrain
7	25.13	1.75	0.91	0.09	0.14	1.60	1.48	0.20	8.82	1.08	1.34	2.31	4.79	7.63	0.46	0.76	32.16	2.79	18.44	30.00	Consumer Non-Cyclicals	GTFP.BH	2016	Bahrain
8	21.04	1.67	0.93	0.06	0.16	1.59	1.50	0.24	5.88	1.05	1.27	2.28	4.51	8.92	0.47	0.77	31.05	1.85	18.44	30.00	Consumer Non-Cyclicals	GTFP.BH	2015	Bahrain
9	17.35	1.43	0.73	0.05	0.24	1.72	1.63	0.39	2.68	1.01	1.23	2.11	4.42	10.25	0.44	0.73	30.30	2.65	18.49	30.00	Consumer Non-Cyclicals	GTFP.BH	2014	Bahrain
10	17.74	1.46	0.81	0.11	0.24	1.76	1.67	0.41	3.49	1.09	1.34	2.23	4.59	11.38	0.47	0.84	29.03	3.30	18.47	30.00	Consumer Non-Cyclicals	GTFP.BH	2013	Bahrain
11	16.75	1.39	0.82	0.08	0.28	1.87	1.77	0.49	1.48	1.02	1.24	2.10	4.08	13.82	0.47	0.90	27.54	2.76	18.42	30.00	Consumer Non-Cyclicals	GTFP.BH	2012	Bahrain
12	17.26	1.31	0.61	0.04	0.29	1.92	1.81	0.52	1.81	1.01	1.23	2.17	3.89	13.10	0.53	0.97	26.55	-0.40	18.49	30.00	Consumer Non-Cyclicals	GTFP.BH	2011	Bahrain
13	18.95	1.22	0.65	0.04	0.26	1.88	1.76	0.46	2.12	0.97	1.19	2.16	3.56	12.37	0.59	1.05	26.04	1.96	18.44	30.00	Consumer Non-Cyclicals	GTFP.BH	2010	Bahrain
14	19.87	1.17	0.65	0.02	0.28	1.94	1.84	0.51	2.90	0.90	1.13	1.96	3.95	11.76	0.64	1.10	24.95	2.80	18.45	30.00	Consumer Non-Cyclicals	GTFP.BH	2009	Bahrain
15	22.21	1.37	0.66	0.09	0.25	1.88	1.78	0.44	3.70	0.95	1.24	2.08	5.60	11.26	0.82	1.31	24.34	3.53	18.46	30.00	Consumer Non-Cyclicals	GTFP.BH	2008	Bahrain
16	20.32	1.50	0.82	0.10	0.18	1.76	1.66	0.30	4.37	0.98	1.28	2.25	6.76	10.29	0.84	1.28	22.91	3.26	18.30	30.00	Consumer Non-Cyclicals	GTFP.BH	2007	Bahrain
17	10.61	1.58	0.87	0.11	0.14	1.60	1.53	0.21	2.06	0.84	1.07	2.06	5.96	17.02	0.80	1.06	21.15	2.01	18.11	30.00	Consumer Non-Cyclicals	GTFP.BH	2006	Bahrain
18	14.32	1.71	0.95	0.18	0.16	1.54	1.46	0.24	3.64	0.89	1.10	2.30	6.12	11.88	0.98	1.02	19.87	2.59	18.13	30.00	Consumer Non-Cyclicals	GTFP.BH	2005	Bahrain
19	16.30	1.71	1.00	0.13	0.16	1.59	1.47	0.24	8.78	1.01	1.24	2.44	6.22	10.37	1.05	1.05	18.61	2.35	18.11	30.00	Consumer Non-Cyclicals	GTFP.BH	2004	Bahrain
20	31.73	1.58	0.92	0.06	0.18	1.73	1.58	0.29	12.72	1.15	1.46	2.45	6.49	7.71	0.86	1.21	17.39	1.59	17.85	30.00	Consumer Non-Cyclicals	GTFP.BH	2003	Bahrain
21	0.46	2.39	2.39	0.26	0.23	1.55	1.55	0.35	-0.14	0.14	0.20	0.48	2.62	115.70	2.95	0.66	114.89	3.98	19.14	24.47	Real Estate	ENMA.KW	2022	Kuwait
22	-3.97	1.29	1.28	0.06	0.19	1.59	1.59	0.29	-1.99	0.15	0.21	0.47	1.97	-17.99	3.21	0.82	105.54	3.42	19.19	24.47	Real Estate	ENMA.KW	2021	Kuwait
23	-0.94	1.43	1.42	0.12	0.22	1.70	1.70	0.38	0.66	0.14	0.19	0.40	2.94	-47.58	2.22	0.50	104.34	2.10	19.31	24.47	Real Estate	ENMA.KW	2020	Kuwait
24	3.41	1.65	1.65	0.06	0.18	1.59	1.59	0.28	4.01	0.21	0.29	0.68	24.67	16.39	1.84	0.63	114.48	1.09	19.28	24.47	Real Estate	ENMA.KW	2019	Kuwait
25	-1.74	1.00	1.00	0.11	0.21	1.69	1.69	0.35	2.53	0.24	0.32	0.85	34.14	-16.67	0.77	0.34	115.12	0.54	19.31	24.47	Real Estate	ENMA.KW	2018	Kuwait
26	-14.76	1.12	1.12	0.03	0.19	1.73	1.73	0.34	-5.42	0.18	0.24	0.62	22.36	-1.85	0.94	0.31	112.38	2.17	19.44	24.47	Real Estate	ENMA.KW	2017	Kuwait
27	-13.98	0.98	0.97	0.02	0.18	1.63	1.63	0.30	-7.76	0.25	0.32	0.77	11.41	-2.00	0.66	0.30	117.94	3.20	19.54	24.47	Real Estate	ENMA.KW	2016	Kuwait
28	7.58	1.05	0.96	0.05	0.19	1.60	1.60	0.31	2.14	0.43	0.54	1.33	12.11	5.66	0.60	0.41	114.59	3.27	19.67	24.47	Real Estate	ENMA.KW	2015	Kuwait
29	4.91	0.79	0.68	0.05	0.24	1.73	1.73	0.42	1.88	0.41	0.51	1.36	9.52	10.09	0.82	0.56	113.91	2.91	19.72	24.47	Real Estate	ENMA.KW	2014	Kuwait
30	5.08	0.85	0.81	0.02	0.23	1.68	1.68	0.39	2.65	0.36	0.44	1.25	6.78	13.23	1.12	0.66	113.34	2.68	19.68	24.47	Real Estate	ENMA.KW	2013	Kuwait
31	4.13	0.70	0.67	0.02	0.25	1.67	1.67	0.42	2.28	0.33	0.40	1.36	5.31	15.63	1.31	0.68	112.05	3.26	19.62	24.47	Real Estate	ENMA.KW	2012	Kuwait
32	2.11	1.02	0.86	0.05	0.25	1.53	1.53	0.39	2.03	0.22	0.26	0.97	4.07	38.73	2.36	0.80	105.09	4.84	19.51	24.47	Real Estate	ENMA.KW	2011	Kuwait
33	1.27	0.79	0.64	0.07	0.25	1.51	1.51	0.38	2.32	0.16	0.19	0.67	4.55	53.54	2.91	0.66	95.86	4.50	19.48	24.47	Real Estate	ENMA.KW	2010	Kuwait
34	-17.33	1.13	1.09	0.03	0.20	1.40	1.40	0.27	-3.68	0.19	0.24	0.72	6.97	-4.81	2.69	0.76	98.19	4.61	19.36	24.47	Real Estate	ENMA.KW	2009	Kuwait
35	12.05	1.44	1.31	0.18	0.13	1.39	1.39	0.19	9.80	0.34	0.45	0.99	10.97	12.40	3.26	1.34	105.67	10.58	19.58	24.47	Real Estate	ENMA.KW	2008	Kuwait
36	26.19	3.07	2.96	0.51	0.02	1.20	1.20	0.03	Null	0.43	0.55	1.04	11.97	7.36	2.66	1.22	103.11	5.48	19.39	24.47	Real Estate	ENMA.KW	2007	Kuwait
37	21.86	3.65	3.52	0.18	0.03	1.17	1.17	0.03	Null	0.49	0.62	1.21	11.06	8.87	2.55	1.32	97.28	3.06	19.23	24.47	Real Estate	ENMA.KW	2006	Kuwait
38	21.57	2.81	2.55	0.06	0.03	1.18	1.18	0.04	Null	0.42	0.56	1.10	9.31	11.33	4.03	1.78	90.48	4.14	19.11	24.47	Real Estate	ENMA.KW	2005	Kuwait
39	16.34	3.18	2.95	0.09	0.01	1.14	1.14	0.01	Null	0.31	0.48	0.77	11.09	9.96	4.97	1.42	81.80	1.25	18.98	24.47	Real Estate	ENMA.KW	2004	Kuwait
40	13.83	4.80	4.68	0.10	Null	1.12	1.12	Null	Null	0.29	0.51	0.63	12.79	13.67	6.02	1.71	74.20	0.96	18.89	24.47	Real Estate	ENMA.KW	2003	Kuwait
41	15.74	2.15	1.73	0.96	0.19	1.42	1.42	0.27	4.66	0.32	0.57	0.62	0.68	20.21	1.53	0.68	86.29	2.81	20.13	28.52	Basic Materials	OCOI.OM	2022	Oman
42	211.16	2.23	1.52	1.29	0.18	1.36	1.36	0.24	33.57	0.26	0.37	0.61	0.45	18.30	1.73	0.55	82.73	1.55	20.08	28.52	Basic Materials	OCOI.OM	2021	Oman
43	31.29	5.93	3.21	0.71	0.02	1.13	1.13	0.02	22.60	0.30	0.36	1.00	0.46	15.64	1.62	0.55	80.25	-0.90	19.88	28.52	Basic Materials	OCOI.OM	2020	Oman
44	18.37	3.17	1.31	0.22	0.05	1.17	1.17	0.06	5.53	0.27	0.31	1.13	0.42	19.38	1.62	0.52	83.05	0.13	19.93	28.52	Basic Materials	OCOI.OM	2019	Oman

S	ROI	CuR	QR	CaR	DR	TACE	TATE	TDPTE	ICR	AT	NOAT	CAT	FAT	P/E	P/S	P/B	GDP(bn)	IR	Size	Age	IT	FS	year	Country
45	21.79	2.97	1.30	0.28	0.07	1.19	1.19	0.08	12.66	0.26	0.32	1.16	0.43	13.57	1.95	0.64	84.00	0.88	20.00	28.52	Basic Materials	OCOI.OM	2018	Oman
46	21.62	2.75	1.51	0.22	0.10	1.24	1.24	0.13	16.41	0.29	0.34	1.33	0.47	14.14	2.36	0.84	82.93	1.60	20.06	28.52	Basic Materials	OCOI.OM	2017	Oman
47	24.43	2.34	1.10	0.20	0.11	1.25	1.25	0.14	31.63	0.28	0.33	1.49	0.47	12.18	2.77	0.96	82.68	1.11	20.09	28.52	Basic Materials	OCOI.OM	2016	Oman
48	22.95	2.73	1.22	0.31	0.11	1.23	1.23	0.14	42.78	0.27	0.33	1.28	0.46	13.01	2.92	0.95	78.71	0.07	20.06	28.52	Basic Materials	OCOI.OM	2015	Oman
49	24.99	3.66	2.46	0.53	0.08	1.20	1.20	0.10	62.73	0.27	0.34	1.13	0.47	13.03	3.33	1.07	74.95	1.02	20.02	28.52	Basic Materials	OCOI.OM	2014	Oman
50	27.44	3.23	2.05	0.36	0.07	1.18	1.18	0.08	81.65	0.28	0.34	1.30	0.49	18.07	5.19	1.69	73.99	1.05	20.02	28.52	Basic Materials	OCOI.OM	2013	Oman
51	34.73	3.08	1.99	0.22	0.07	1.17	1.17	0.08	79.07	0.32	0.37	1.45	0.53	12.19	3.76	1.36	70.32	2.95	19.99	28.52	Basic Materials	OCOI.OM	2012	Oman
52	44.65	2.90	2.03	0.21	0.05	1.17	1.17	0.05	39.47	0.28	0.33	1.16	0.28	11.15	2.98	0.96	64.59	4.04	19.93	28.52	Basic Materials	OCOI.OM	2011	Oman
53	67.59	3.22	2.47	0.25	0.04	1.16	1.16	0.05	Null	0.32	0.37	1.61	0.22	8.46	4.08	1.43	62.78	3.26	19.92	28.52	Basic Materials	OCOI.OM	2010	Oman
54	55.30	2.86	1.95	0.28	0.03	1.13	1.13	0.03	-132990.56	0.48	0.56	2.06	0.45	10.21	3.62	1.82	61.72	3.47	19.80	28.52	Basic Materials	OCOI.OM	2009	Oman
55	40.06	4.28	2.73	0.48	0.01	1.13	1.13	0.01	162.47	0.48	0.55	1.72	Null	7.94	1.57	0.84	58.16	12.38	19.66	28.52	Basic Materials	OCOI.OM	2008	Oman
56	41.30	4.12	3.56	0.30	0.02	1.12	1.12	0.02	784.01	0.39	0.44	1.90	Null	11.88	4.40	1.83	53.76	5.83	19.67	28.52	Basic Materials	OCOI.OM	2007	Oman
57	38.22	2.77	2.08	0.15	0.02	1.14	1.14	0.02	1718.78	Null	Null	Null	Null	9.41	3.87	1.79	51.46	3.54	19.58	28.52	Basic Materials	OCOI.OM	2006	Oman
58	Null	Null	Null	Null	Null	Null	Null	Null	991.37	Null	Null	Null	Null	Null	4.81	Null	48.84	1.91	Null	28.52	Basic Materials	OCOI.OM	2005	Oman
59	76.67	2.57	1.80	0.16	0.04	1.19	1.19	0.05	354.21	0.38	0.44	2.31	0.36	11.61	Null	1.96	47.65	0.62	19.30	28.52	Basic Materials	OCOI.OM	2004	Oman
60	80.82	2.41	1.49	0.17	0.05	1.20	1.20	0.06	109.31	0.32	0.37	2.34	0.26	11.24	Null	1.42	47.05	0.23	19.20	28.52	Basic Materials	OCOI.OM	2003	Oman
61	19.58	1.96	1.89	0.14	0.07	1.14	1.14	0.08	8.22	0.18	0.21	1.25	1.06	11.43	3.54	0.72	170.95	5.00	22.33	26.60	Industrials	QNNC.QA	2022	Qatar
62	13.99	1.54	1.47	0.22	0.10	1.19	1.18	0.12	3.48	0.16	0.18	1.44	0.84	11.94	3.14	0.59	164.04	2.30	22.29	26.60	Industrials	QNNC.QA	2021	Qatar
63	1.25	1.16	1.11	0.21	0.14	1.24	1.24	0.17	3.33	0.13	0.14	1.54	0.59	136.08	3.58	0.59	161.42	-2.54	22.27	26.60	Industrials	QNNC.QA	2020	Qatar
64	10.61	1.36	1.27	0.20	0.13	1.22	1.21	0.16	2.75	0.14	0.15	1.64	0.59	12.69	2.88	0.48	167.37	-0.67	22.30	26.60	Industrials	QNNC.QA	2019	Qatar
65	8.37	2.20	2.04	0.16	0.14	1.22	1.22	0.17	2.72	0.13	0.15	0.91	0.59	14.65	3.12	0.51	166.23	0.26	22.31	26.60	Industrials	QNNC.QA	2018	Qatar
66	7.01	1.87	1.81	0.19	0.23	1.39	1.38	0.32	2.30	0.12	0.16	0.52	0.55	13.14	2.57	0.47	164.20	0.39	22.37	26.60	Industrials	QNNC.QA	2017	Qatar
67	7.59	1.26	1.24	0.23	0.30	1.55	1.55	0.46	3.17	0.12	0.18	0.39	0.52	15.44	4.29	0.77	166.70	2.68	22.52	26.60	Industrials	QNNC.QA	2016	Qatar
68	21.30	1.30	1.26	1.03	0.22	1.62	1.61	0.36	8.04	0.15	0.22	0.51	0.67	9.73	3.65	0.79	161.74	1.81	22.53	26.60	Industrials	QNNC.QA	2015	Qatar
69	12.46	2.52	2.41	0.17	0.24	1.39	1.38	0.33	16.29	0.15	0.19	0.71	0.73	10.76	4.33	0.84	154.40	3.35	22.36	26.60	Industrials	QNNC.QA	2014	Qatar
70	15.52	2.10	1.98	0.18	0.14	1.24	1.24	0.18	15.68	0.16	0.19	0.83	0.72	9.95	4.12	0.75	146.58	3.22	22.18	26.60	Industrials	QNNC.QA	2013	Qatar
71	15.48	2.19	2.06	0.05	0.18	1.31	1.31	0.24	10.79	0.17	0.20	1.00	0.77	8.64	3.15	0.67	138.87	2.32	22.08	26.60	Industrials	QNNC.QA	2012	Qatar
72	17.10	0.94	0.88	0.47	0.15	1.26	1.25	0.19	10.43	0.16	0.19	1.21	0.73	12.29	4.00	0.83	132.59	1.14	22.02	26.60	Industrials	QNNC.QA	2011	Qatar
73	24.76	2.08	1.92	0.33	0.15	1.23	1.23	0.18	8.63	0.21	0.23	1.61	0.95	9.70	5.02	0.99	116.95	-2.43	22.02	26.60	Industrials	QNNC.QA	2010	Qatar
74	9.36	1.55	1.35	0.52	0.14	1.22	1.22	0.17	12.37	0.17	0.19	1.40	1.00	9.72	3.91	0.79	97.79	-4.86	21.41	26.60	Industrials	QNNC.QA	2009	Qatar
75	12.60	1.48	1.32	0.58	0.13	1.21	1.21	0.15	11.08	0.18	0.20	1.66	1.34	7.53	3.23	0.77	87.35	15.05	21.34	26.60	Industrials	QNNC.QA	2008	Qatar
76	7.15	0.55	0.48	0.14	0.18	1.28	1.28	0.23	3.81	0.16	0.18	1.29	1.42	13.53	5.77	0.99	74.24	13.76	21.49	26.60	Industrials	QNNC.QA	2007	Qatar
77	6.26	0.95	0.87	0.38	0.15	1.25	1.25	0.19	4.40	0.14	0.15	1.17	1.98	21.03	5.28	1.23	62.92	11.84	21.11	26.60	Industrials	QNNC.QA	2006	Qatar
78	3.68	1.43	1.38	0.74	0.07	1.09	1.09	0.07	9.12	0.13	0.14	1.13	2.09	34.43	9.21	1.21	49.87	8.81	21.63	26.60	Industrials	QNNC.QA	2005	Qatar
79	6.62	3.32	3.20	2.12	0.06	1.11	1.11	0.07	8.81	0.17	0.21	0.93	1.55	19.90	4.74	0.94	46.39	6.80	20.80	26.60	Industrials	QNNC.QA	2004	Qatar
80	11.91	3.44	3.29	2.08	0.10	1.16	1.16	0.11	8.82	0.16	0.20	0.62	1.02	8.26	5.95	0.75	38.91	2.26	20.36	26.60	Industrials	QNNC.QA	2003	Qatar
81	59.05	0.91	0.51	0.03	0.43	2.94	2.94	1.27	3.63	0.50	0.67	2.09	0.76	19.64	1.34	2.01	769.63	2.47	19.52	24.47	Basic Materials	2210.SE	2022	Saudi Arabia
82	45.56	0.84	0.50	0.03	0.46	3.47	3.47	1.59	1.78	0.46	0.59	2.40	0.68	24.09	1.37	2.28	708.16	3.06	19.54	24.47	Basic Materials	2210.SE	2021	Saudi Arabia
83	-20.60	0.22	0.14	0.01	0.59	4.44	4.44	2.63	-2.21	0.34	0.42	2.17	0.51	-18.93	1.78	2.71	678.79	3.45	19.64	24.47	Basic Materials	2210.SE	2020	Saudi Arabia
84	8.11	0.24	0.17	0.03	0.60	4.16	4.16	2.51	0.97	0.39	0.47	2.36	0.64	36.63	1.06	1.75	709.60	-2.09	19.67	24.47	Basic Materials	2210.SE	2019	Saudi Arabia
85	16.39	0.28	0.18	0.02	0.59	4.25	4.25	2.52	2.29	0.44	0.54	2.79	0.81	15.17	1.02	1.92	703.74	2.46	19.74	24.47	Basic Materials	2210.SE	2018	Saudi Arabia
86	1.14	0.24	0.19	0.03	0.66	5.20	5.20	3.42	0.61	0.39	0.49	2.39	0.65	145.73	0.69	1.50	684.83	-0.84	19.76	24.47	Basic Materials	2210.SE	2017	Saudi Arabia
87	-493.86	0.34	0.28	0.05	0.70	5.79	5.79	4.03	-4.10	0.23	0.27	1.33	0.32	-1.00	1.76	2.87	685.30	2.07	19.88	24.47	Basic Materials	2210.SE	2016	Saudi Arabia
88	-42.84	0.79	0.63	0.24	0.47	2.20	2.20	1.03	-3.09	0.18	0.22	0.96	0.25	-13.34	2.00	0.84	669.48	1.21	20.28	24.47	Basic Materials	2210.SE	2015	Saudi Arabia
89	-51.15	1.45	1.27	0.43	0.44	2.12	2.12	0.94	-2.89	0.25	0.31	1.14	0.37	-12.92	2.04	1.12	639.49	2.24	20.36	24.47	Basic Materials	2210.SE	2014	Saudi Arabia
90	-28.47	1.39	1.09	0.69	0.43	2.04	2.04	0.87	-1.19	0.28	0.33	1.49	0.40	-45.26	2.43	1.30	614.73	3.53	20.44	24.47	Basic Materials	2210.SE	2013	Saudi Arabia
91	-6.46	1.26	0.87	0.20	0.37	1.87	1.87	0.70	0.19	0.29	0.33	1.84	0.39	-147.51	2.30	1.21	597.70	2.87	20.34	24.47	Basic Materials	2210.SE	2012	Saudi Arabia
92	-169.98	1.71	1.22	0.47	0.35	1.73	1.73	0.60	-3.29	0.30	0.36	1.53	0.45	-5.38	1.76	0.93	566.93	5.83	20.27	24.47	Basic Materials	2210.SE	2011	Saudi Arabia
93	12.13	1.79	1.52	0.32	0.27	1.50	1.50	0.40	9.46	0.26	0.32	1.16	0.43	32.34	2.28	0.88	510.77	5.34	20.30	24.47	Basic Materials	2210.SE	2010	Saudi Arabia
94	-13.12	2.21	1.89	1.36	0.29	1.53	1.53	0.45	-8.27	0.16	0.21	0.62	0.30	-34.87	3.49	0.89	486.27	5.06	20.28	24.47	Basic Materials	2210.SE	2009	Saudi Arabia
95	-15.38	2.60	2.14	1.67	0.31	1.63	1.63	0.51	1.82	0.29	0.37	1.09	0.66	-16.74	1.77	0.71	496.49	9.87	20.34	24.47	Basic Materials	2210.SE	2008	Saudi Arabia
96	10.73	1.85	1.32	0.39	0.40	1.90	1.90	0.75	5.60	0.36	0.41	1.51	0.65	66.58	4.81	2.68	467.29	4.17	19.96	24.47	Basic Materials	2210.SE	2007	Saudi Arabia

S	ROI	CuR	QR	CaR	DR	TACE	TATE	TDPT	ICR	AT	NOAT	CAT	FAT	P/E	P/S	P/B	GDP(bn)	IR	Size	Age	IT	FS	year	Country
97	13.21	1.22	1.01	0.38	0.23	1.42	1.41	0.32	4.03	0.36	0.48	1.02	0.63	59.61	4.60	2.32	458.81	2.21	19.51	24.47	Basic Materials	2210.SE	2006	Saudi Arabia
98	10.87	2.72	2.43	0.98	0.18	1.58	1.33	0.24	-0.28	0.28	0.49	0.57	0.91	180.49	22.56	10.07	446.37	0.48	19.48	24.47	Basic Materials	2210.SE	2005	Saudi Arabia
99	-41.19	5.93	5.48	4.63	0.16	1.51	1.26	0.20	-4.20	0.31	0.48	0.72	0.85	-34.43	5.92	2.22	422.80	0.52	19.39	24.47	Basic Materials	2210.SE	2004	Saudi Arabia
100	-24.80	1.14	0.72	0.25	0.33	2.54	1.69	0.55	-8.55	0.29	0.32	1.50	0.60	-16.82	Null	2.55	391.63	0.61	18.78	24.47	Basic Materials	2210.SE	2003	Saudi Arabia
101	38.07	3.21	2.38	0.48	0.14	1.55	1.37	0.19	10.73	0.36	0.49	0.87	0.87	7.01	1.43	0.80	431.44	4.83	21.14	23.04	Industrials	ADAVIATION.AD	2022	UAE
102	30.90	2.60	2.01	0.47	0.15	1.66	1.48	0.22	13.39	0.32	0.44	0.77	0.72	7.10	1.30	0.65	400.04	-0.01	21.13	23.04	Industrials	ADAVIATION.AD	2021	UAE
103	19.21	3.63	2.64	0.57	0.18	1.59	1.43	0.26	5.60	0.31	0.43	0.76	0.64	10.41	1.03	0.51	383.34	-2.08	21.03	23.04	Industrials	ADAVIATION.AD	2020	UAE
104	39.57	4.62	3.56	0.65	0.18	1.64	1.46	0.26	6.94	0.43	0.58	1.10	0.82	4.43	0.64	0.44	403.34	-1.93	21.03	23.04	Industrials	ADAVIATION.AD	2019	UAE
105	36.05	2.58	1.85	0.28	0.20	1.67	1.50	0.30	5.99	0.38	0.51	1.10	0.69	5.36	0.84	0.53	398.91	3.07	20.99	23.04	Industrials	ADAVIATION.AD	2018	UAE
106	38.24	4.55	3.30	0.49	0.21	1.71	1.56	0.33	9.67	0.34	0.46	0.97	0.58	4.79	0.84	0.49	393.74	1.97	20.97	23.04	Industrials	ADAVIATION.AD	2017	UAE
107	69.73	3.14	2.34	0.33	0.24	1.87	1.72	0.40	11.20	0.38	0.53	1.11	0.63	4.76	0.74	0.52	390.87	1.62	20.99	23.04	Industrials	ADAVIATION.AD	2016	UAE
108	79.84	3.22	2.41	0.59	0.24	1.94	1.79	0.43	12.87	0.49	0.68	1.60	0.76	4.81	0.61	0.55	370.28	4.07	20.96	23.04	Industrials	ADAVIATION.AD	2015	UAE
109	90.00	2.77	1.93	0.42	0.22	1.96	1.78	0.39	11.37	0.42	0.56	1.38	0.67	5.45	0.83	0.61	346.74	2.35	20.88	23.04	Industrials	ADAVIATION.AD	2014	UAE
110	84.17	3.52	2.53	0.31	0.18	1.63	1.49	0.27	14.13	0.48	0.65	1.36	0.78	6.33	0.86	0.70	332.88	1.10	20.62	23.04	Industrials	ADAVIATION.AD	2013	UAE
111	75.32	3.00	2.24	0.37	0.22	1.91	1.75	0.39	8.08	0.50	0.73	1.50	0.81	4.41	0.65	0.62	316.86	0.66	20.70	23.04	Industrials	ADAVIATION.AD	2012	UAE
112	201.90	3.09	2.36	0.73	0.24	2.07	1.89	0.45	6.98	0.55	0.81	1.59	0.76	4.62	0.71	0.77	311.18	0.88	20.67	23.04	Industrials	ADAVIATION.AD	2011	UAE
113	167.09	4.02	3.16	1.18	0.23	2.05	1.89	0.43	6.89	0.52	0.74	1.74	0.68	5.10	0.69	0.76	292.97	0.88	20.58	23.04	Industrials	ADAVIATION.AD	2010	UAE
114	93.07	2.31	1.75	Null	0.28	2.36	2.12	0.60	4.21	0.46	0.65	1.76	0.60	7.31	0.70	0.75	288.35	1.56	20.57	23.04	Industrials	ADAVIATION.AD	2009	UAE
115	90.41	1.70	1.37	0.33	0.28	2.53	2.26	0.64	4.23	0.42	0.59	1.30	0.54	5.84	0.58	0.56	304.30	12.25	20.54	23.04	Industrials	ADAVIATION.AD	2008	UAE
116	25.85	1.71	1.37	0.21	0.32	1.83	1.83	0.58	1.88	0.29	0.38	0.71	Null	17.37	2.57	1.15	294.89	11.10	20.16	23.04	Industrials	ADAVIATION.AD	2007	UAE
117	20.42	1.73	1.43	0.38	0.19	1.45	1.45	0.28	5.49	0.34	0.45	0.88	Null	13.14	1.76	0.59	285.79	9.30	19.91	23.04	Industrials	ADAVIATION.AD	2006	UAE
118	37.29	1.86	1.56	0.05	0.20	1.41	1.41	0.29	Null	0.34	0.41	1.03	0.08	10.18	4.33	1.17	260.19	6.20	19.61	23.04	Industrials	ADAVIATION.AD	2005	UAE
119	47.62	2.60	2.06	0.28	0.15	1.29	1.29	0.19	Null	0.36	0.44	1.32	Null	9.72	5.71	1.06	248.14	5.00	19.42	23.04	Industrials	ADAVIATION.AD	2004	UAE
120	36.56	2.41	1.71	0.29	0.04	1.27	1.27	0.05	Null	Null	Null	Null	Null	7.85	4.03	0.59	226.48	3.10	19.29	23.04	Industrials	ADAVIATION.AD	2003	UAE

FS: Firm Symbol

Appendix 2. Optimal Hyperparameters

Optimal Hyperparameters Found by I-RS-GCC Models

Model	Best Params	Best CV Score
LR	{}	-29319.804
ANN	{'solver': 'adam', 'hidden_layer_sizes': (50, 50), 'alpha': 0.0001, 'activation': 'relu'}	-23832.677
RT	{'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10}	-29377.229
kNN	{'weights': 'distance', 'n_neighbors': 5, 'metric': 'manhattan'}	-21204.137
RF	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 30}	-19745.668
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}	-19368.933
CatBoost	{'learning_rate': 0.1, 'iterations': 500, 'depth': 10}	-18298.178

Optimal Hyperparameters Found by NI-RS-GCC Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}	-30618.271
CatBoost	{'learning_rate': 0.1, 'iterations': 500, 'depth': 10}	-29074.917

Optimal Hyperparameters Found by Bahrain Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.2}	-1024.780
CatBoost	{'learning_rate': 0.05, 'iterations': 500, 'depth': 6}	-971.002

Optimal Hyperparameters Found by Kuwait Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}	-3713.020833
CatBoost	{'learning_rate': 0.1, 'iterations': 500, 'depth': 6}	-3675.239184

Optimal Hyperparameters Found by Oman Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}	-166261.585
CatBoost	{'learning_rate': 0.05, 'iterations': 500, 'depth': 10}	-152969.026

Optimal Hyperparameters Found by Qatar Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 1.0, 'n_estimators': 200, 'max_depth': 5, 'learning_rate': 0.2}	-8956.610
CatBoost	{'learning_rate': 0.05, 'iterations': 500, 'depth': 6}	-8111.787

Optimal Hyperparameters Found by Saudi Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}	-211744.335
CatBoost	{'learning_rate': 0.05, 'iterations': 500, 'depth': 10}	-206413.110

Optimal Hyperparameters Found by UAE Models

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 5, 'learning_rate': 0.1}	-11642.643
CatBoost	{'learning_rate': 0.1, 'iterations': 500, 'depth': 6}	-11317.970

Optimal Hyperparameters Found by I-CS-GCC Models

(One Year of Training Data)

Model	Best Params	Best CV Score
LR	{}	-27694.470
ANN	{'solver': 'adam', 'hidden_layer_sizes': (100,), 'alpha': 0.0001, 'activation': 'tanh'}	-26895.440
RT	{'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': None}	-42681.280
kNN	{'weights': 'distance', 'n_neighbors': 9, 'metric': 'euclidean'}	-29876.370
RF	{'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': 20}	-29882.750
XGBoost	{'subsample': 0.8, 'n_estimators': 100, 'max_depth': 6, 'learning_rate': 0.01}	-27082.190
CatBoost	{'learning_rate': 0.01, 'iterations': 500, 'depth': 4}	-27056.010

Optimal Hyperparameters Found by I-CS-GCC Models

(Two Years of Cumulative Training Data)

Model	Best Params	Best CV Score
LR	{}	-26567.226
ANN	{'solver': 'adam', 'hidden_layer_sizes': (100,), 'alpha': 0.01, 'activation': 'tanh'}	-25467.447
RT	{'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 20}	-48405.471
kNN	{'weights': 'distance', 'n_neighbors': 9, 'metric': 'euclidean'}	-27843.475
RF	{'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': 20}	-27951.289
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.01}	-26276.391
CatBoost	{'learning_rate': 0.01, 'iterations': 200, 'depth': 6}	-25960.511

Optimal Hyperparameters Found by I-CS-GCC Models
(Three Years of Cumulative Training Data)

Model	Best Params	Best CV Score
LR	{}	-25804.139
ANN	{'solver': 'adam', 'hidden_layer_sizes': (100,), 'alpha': 0.01, 'activation': 'tanh'}	-24257.930
RT	{'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 10}	-39644.336
kNN	{'weights': 'distance', 'n_neighbors': 9, 'metric': 'euclidean'}	-27049.321
RF	{'n_estimators': 50, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': 30}	-25758.566
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.01}	-24994.998
CatBoost	{'learning_rate': 0.01, 'iterations': 200, 'depth': 6}	-24891.243

Optimal Hyperparameters Found by I-CS-GCC Models
(Four Years of Cumulative Training Data)

Model	Best Params	Best CV Score
LR	{}	-26332.006
ANN	{'solver': 'adam', 'hidden_layer_sizes': (100,), 'alpha': 0.0001, 'activation': 'tanh'}	-24838.049
RT	{'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10}	-44155.793
kNN	{'weights': 'uniform', 'n_neighbors': 9, 'metric': 'euclidean'}	-28530.489
RF	{'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_depth': 10}	-28079.962
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.01}	-25452.829
CatBoost	{'learning_rate': 0.01, 'iterations': 200, 'depth': 4}	-25742.104

Optimal Hyperparameters Found by I-CS-GCC Models
(Five Years of Cumulative Training Data)

Model	Best Params	Best CV Score
LR	{}	-29919.190
ANN	{'solver': 'adam', 'hidden_layer_sizes': (100,), 'alpha': 0.01, 'activation': 'tanh'}	-28401.100
RT	{'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10}	-39432.370
kNN	{'weights': 'uniform', 'n_neighbors': 9, 'metric': 'euclidean'}	-32728.610
RF	{'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': 10}	-29044.530
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.01}	-28475.930
CatBoost	{'learning_rate': 0.01, 'iterations': 500, 'depth': 6}	-28284.900

**Optimal Hyperparameters Found by NI-CS-GCC Models
(One Year of Training Data)**

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.01}	-45446.609
CatBoost	{'learning_rate': 0.05, 'iterations': 100, 'depth': 4}	-46145.921

**Optimal Hyperparameters Found by NI-CS-GCC Models
(Two Years of Cumulative Training Data)**

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.01}	-43118.545
CatBoost	{'learning_rate': 0.01, 'iterations': 500, 'depth': 6}	-43990.582

**Optimal Hyperparameters Found by NI-CS-GCC Models
(Three Years of Cumulative Training Data)**

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.01}	-44587.975
CatBoost	{'learning_rate': 0.01, 'iterations': 500, 'depth': 6}	-43931.114

**Optimal Hyperparameters Found by NI-CS-GCC Models
(Four Years of Cumulative Training Data)**

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.8, 'n_estimators': 200, 'max_depth': 4, 'learning_rate': 0.01}	-44644.903
CatBoost	{'learning_rate': 0.01, 'iterations': 500, 'depth': 6}	-46835.019

**Optimal Hyperparameters Found by NI-CS-GCC Models
(Five Years of Cumulative Training Data)**

Model	Best Params	Best CV Score
XGBoost	{'subsample': 0.6, 'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.01}	-48333.247
CatBoost	{'learning_rate': 0.05, 'iterations': 200, 'depth': 4}	-49441.482

Appendix 3. Python Code

(a) Plot ANN diagram:

```
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np
# Create a simple ANN diagram with 3 layers (input, hidden, output)

# Defining the layers
input_layer = ["Country", "IT", "Age", "Size", "IR", "GDP",
               "P/B", "P/S", "P/E", "FAT", "CAT", "NOAT", "AT", "ICR", "TDPT", "TATE", "TACE", "DR", "CaR", "QR", "CuR"]
hidden_layer = ["H11", "H10", "H9", "H8", "H7", "H6", "H5", "H4", "H3", "H2", "H1"] # 11 hidden
neurons for illustration
output_layer = ["ROI"]

# Creating a graph to visualize the ANN
G = nx.DiGraph()

# Add nodes for input, hidden, and output layers
for i in input_layer:
    G.add_node(i, layer='input')
for h in hidden_layer:
    G.add_node(h, layer='hidden')
for o in output_layer:
    G.add_node(o, layer='output')

# Connect input to hidden layer
for i in input_layer:
    for h in hidden_layer:
        G.add_edge(i, h)

# Connect hidden layer to output layer
for h in hidden_layer:
    for o in output_layer:
        G.add_edge(h, o)

# Position nodes in the graph for visualization
pos = {
    **{(i): (0, idx) for idx, i in enumerate(input_layer)}, # input layer on the left
    **{(h): (1, idx + 0.5) for idx, h in enumerate(hidden_layer)}, # hidden layer in the middle
    **{(o): (2, 2.5) for o in output_layer}, # output layer on the right
}

# Plot the ANN structure
plt.figure(figsize=(6, 6))
nx.draw(G, pos, with_labels=True, node_size=1000, node_color='lightblue', font_size=8,
        edge_color='gray', font_weight='bold', arrows=False)
plt.title("ANN Structure: Profitability Prediction")
plt.show()
```

(b) Hyperparameter Tuning with *RandomizedSearchCV*, then Model Evaluation:

```
import numpy as np
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

results = {}

for name, (model, param_grid) in models.items():
    print(f"Tuning hyperparameters for {name}")
    random_search = RandomizedSearchCV(
        estimator=model,
        param_distributions=param_grid,
        n_iter=100, # Number of parameter settings sampled
        scoring='neg_mean_squared_error',
        cv=5, # Cross-validation splitting strategy
        verbose=2,
        random_state=42,
        n_jobs=-1 # Use all available cores
    )
    random_search.fit(X_train_scaled, y_train)

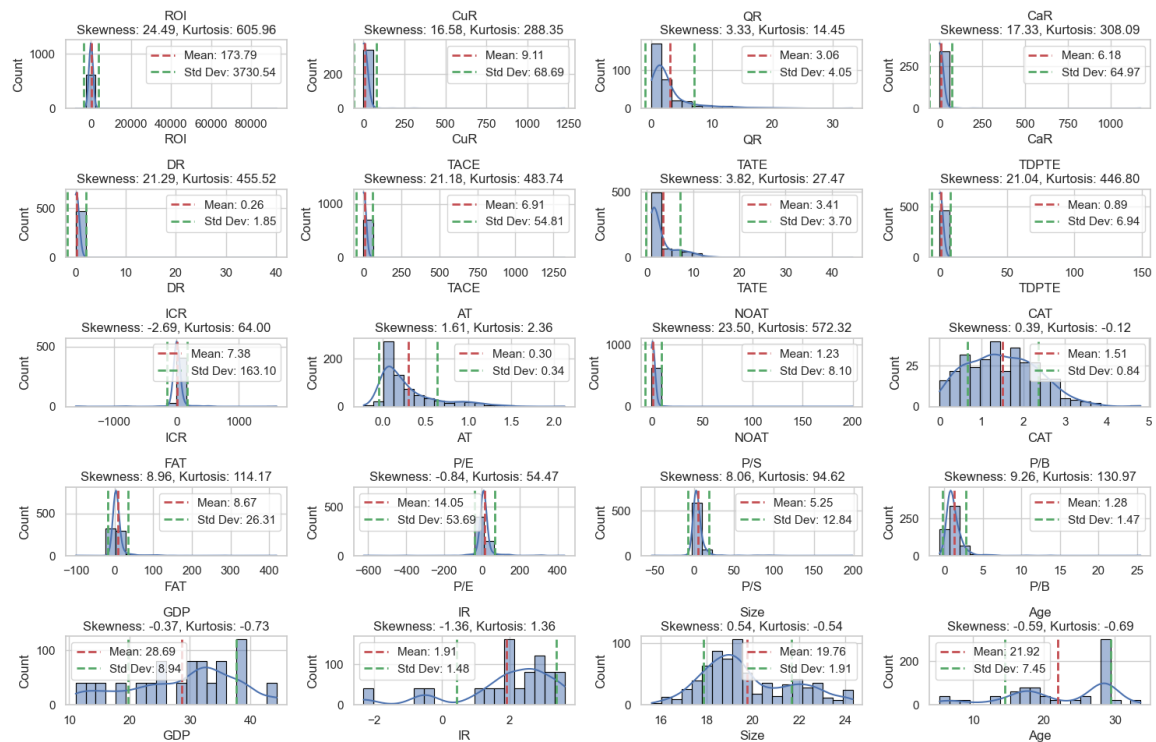
    # Get the best model parameters
    best_model = random_search.best_estimator_

    # Predict on the test set
    y_pred = best_model.predict(X_test_scaled)

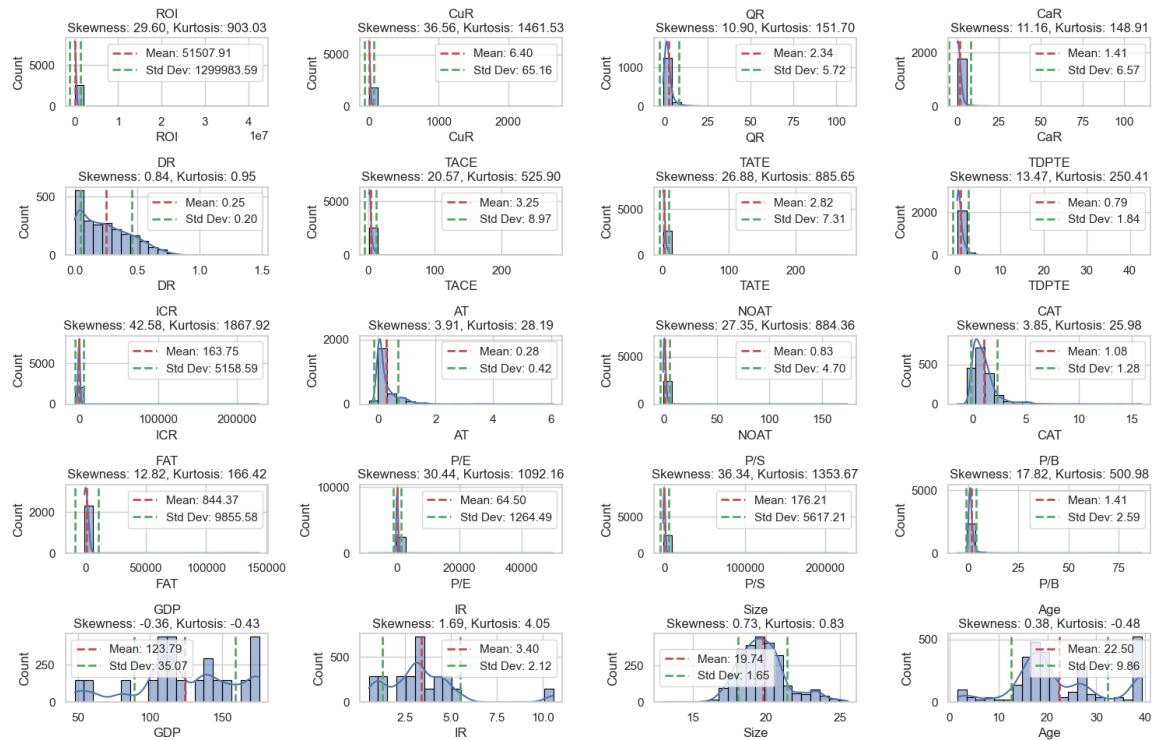
    # Evaluate the model
    mse = mean_squared_error(y_test, y_pred)
    rmse = np.sqrt(mse)
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    results[name] = {
        'best_params': random_search.best_params_,
        'best_score': random_search.best_score_,
        'mse': mse,
        'rmse': rmse,
        'mae': mae,
        'r2': r2
    }
```

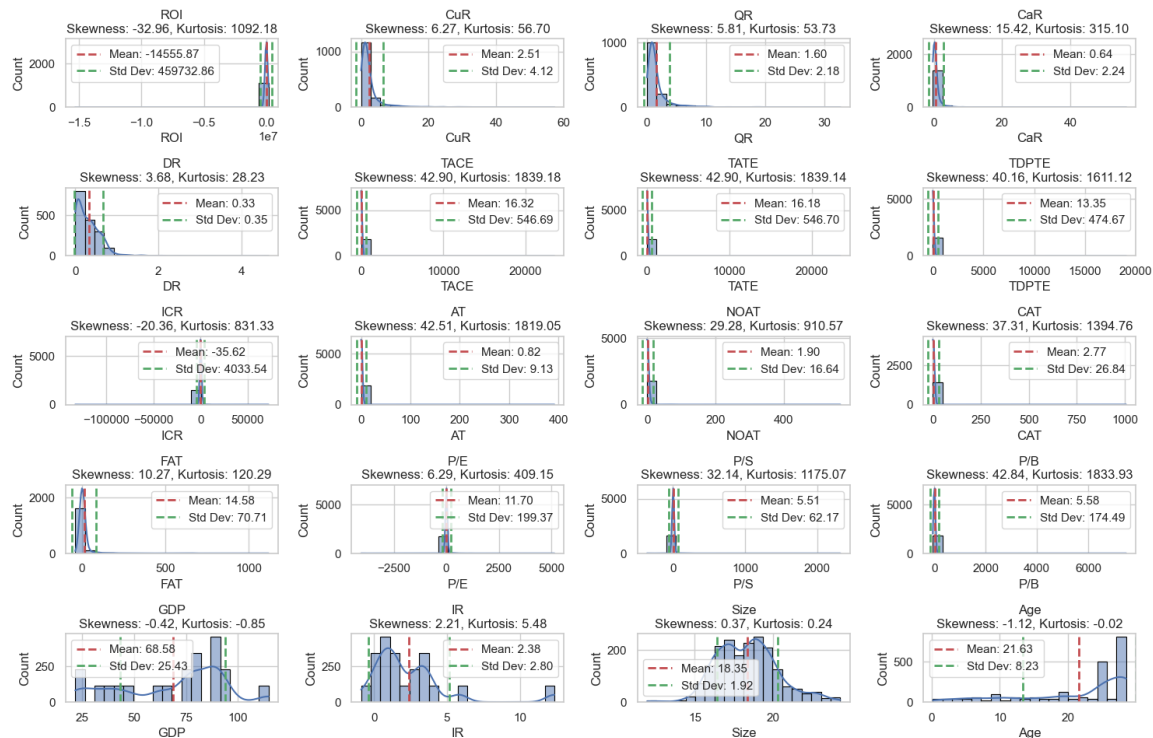
Appendix 4. Descriptive Charts of Data Distribution (C Models)



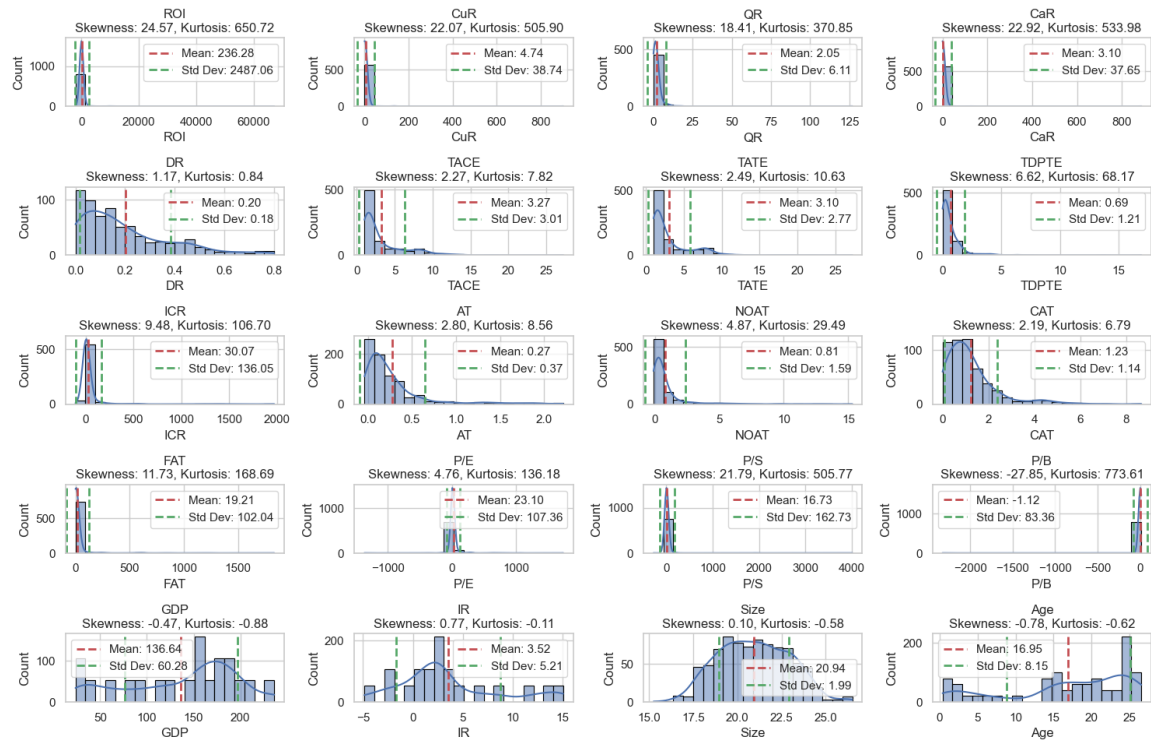
Bahrain's Variables Distribution



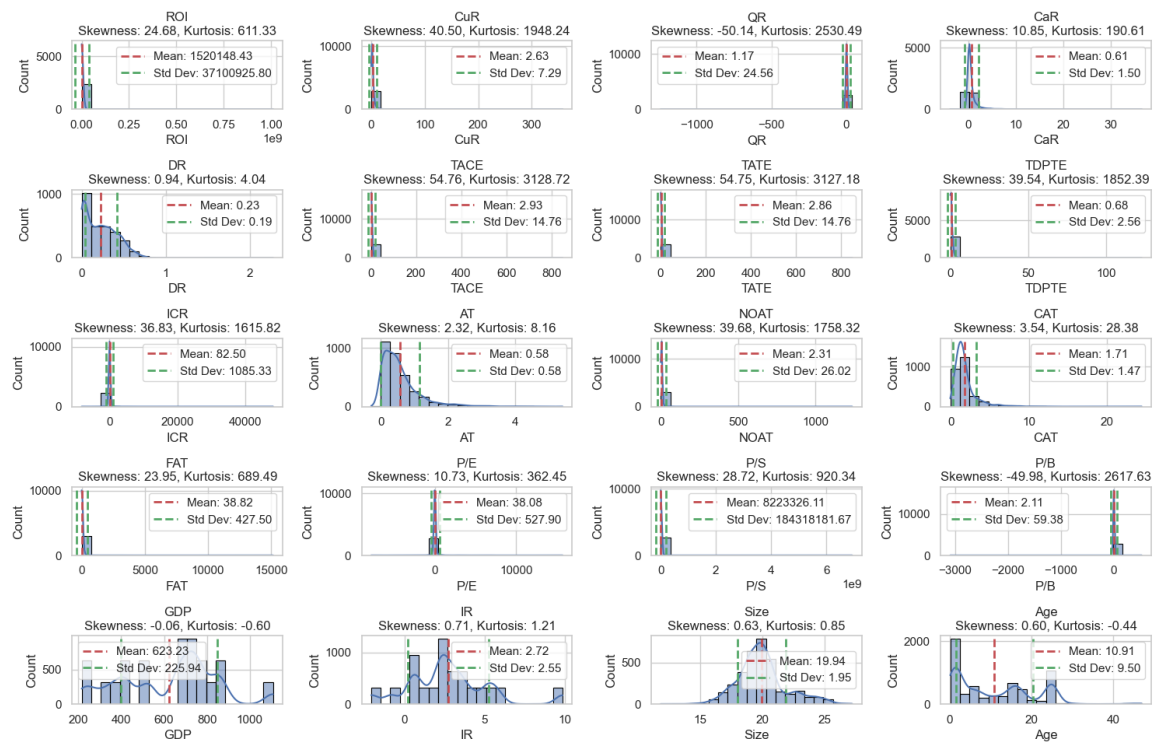
Kuwait's Variables Distribution



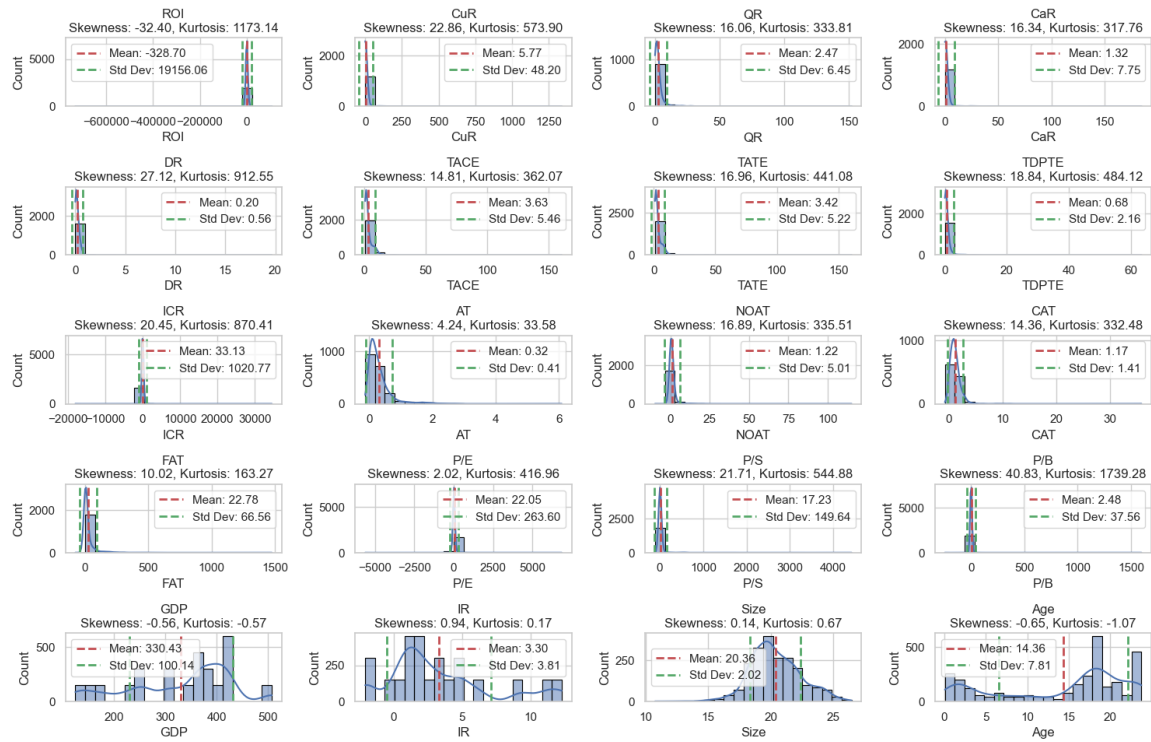
Oman's Variables Distribution



Qatar's Variables Distribution

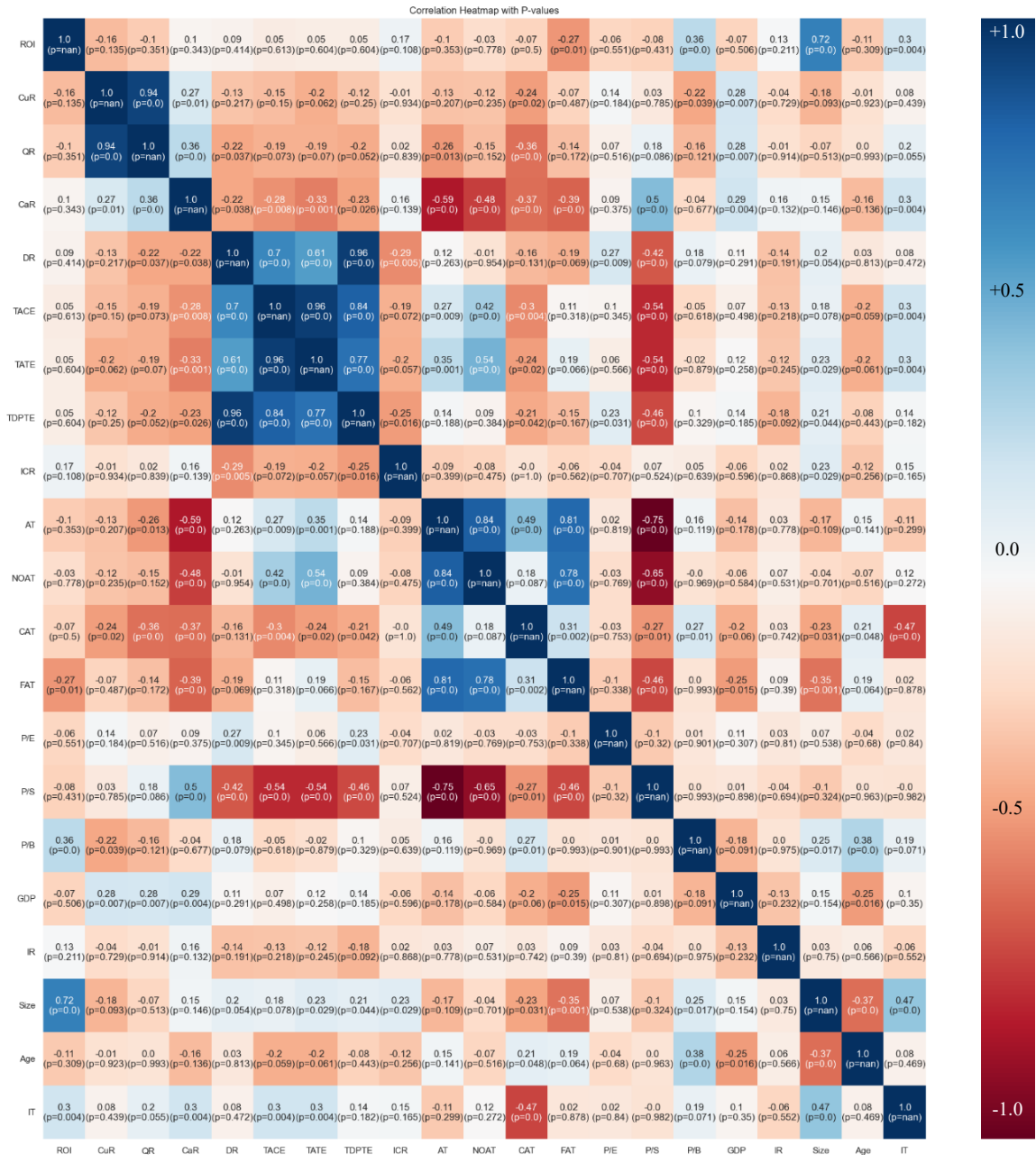


Saudi's Variables Distribution

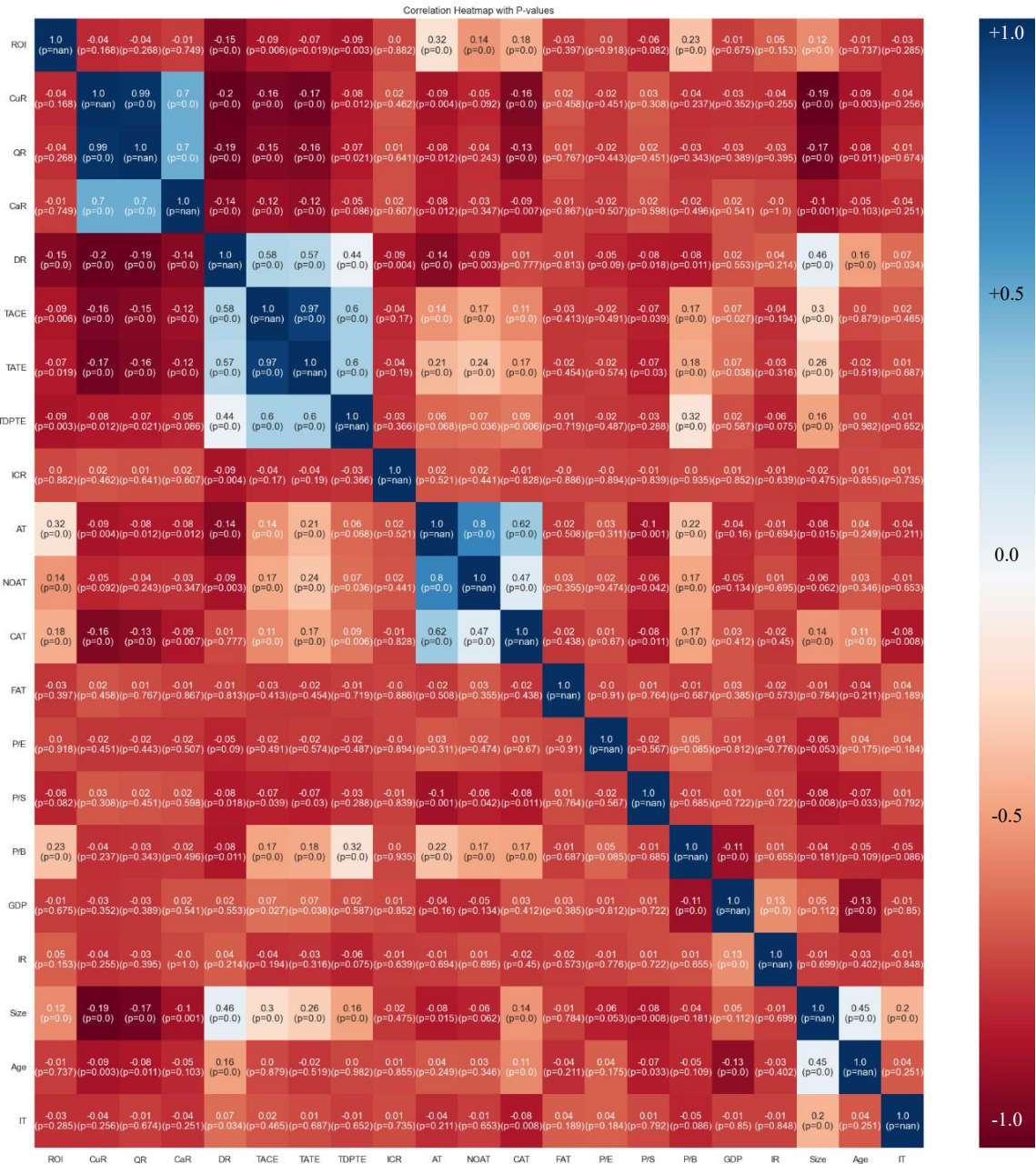


UAE's Variables Distribution

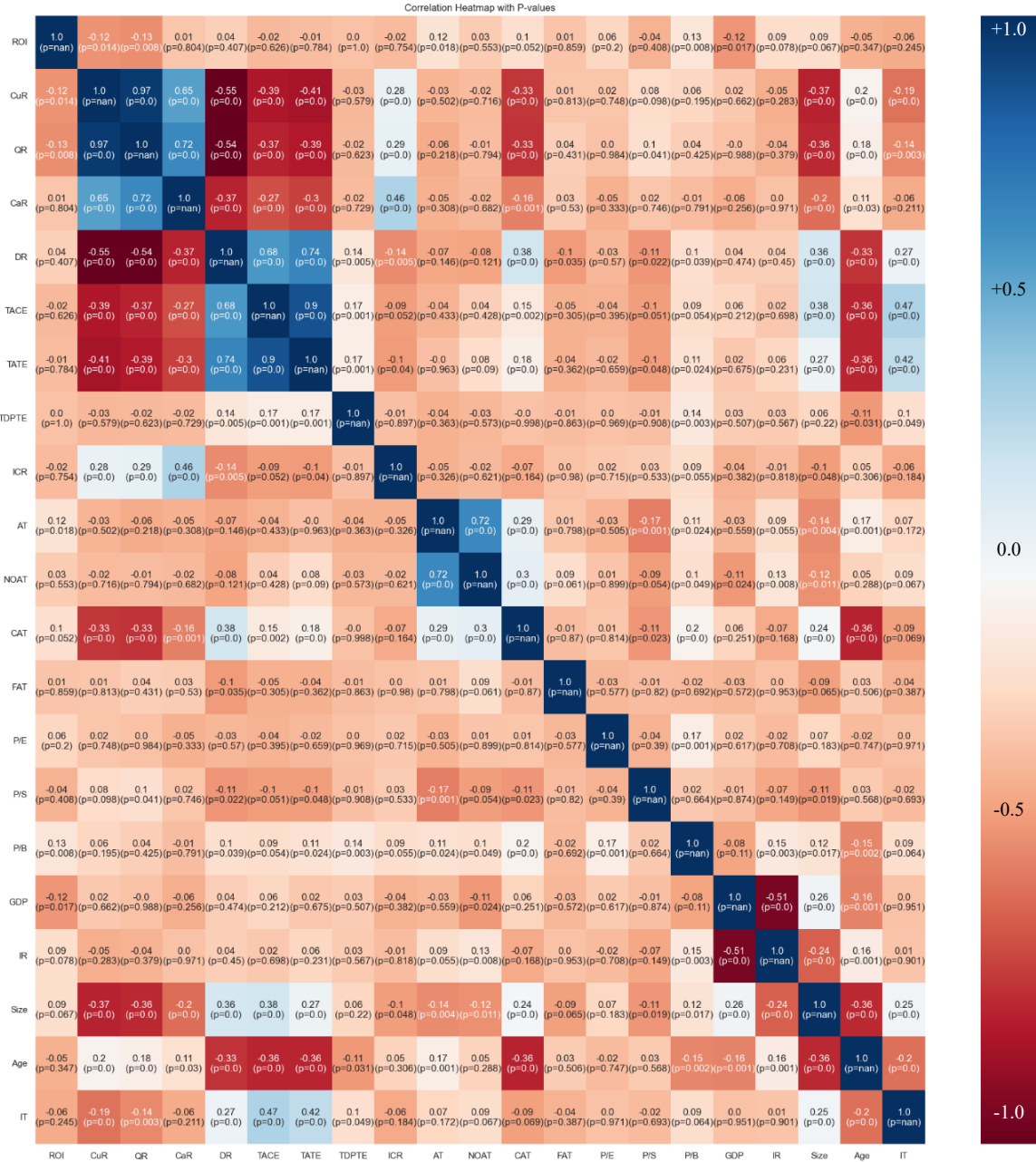
Appendix 5. Pearson Correlation Matrices (C Models)



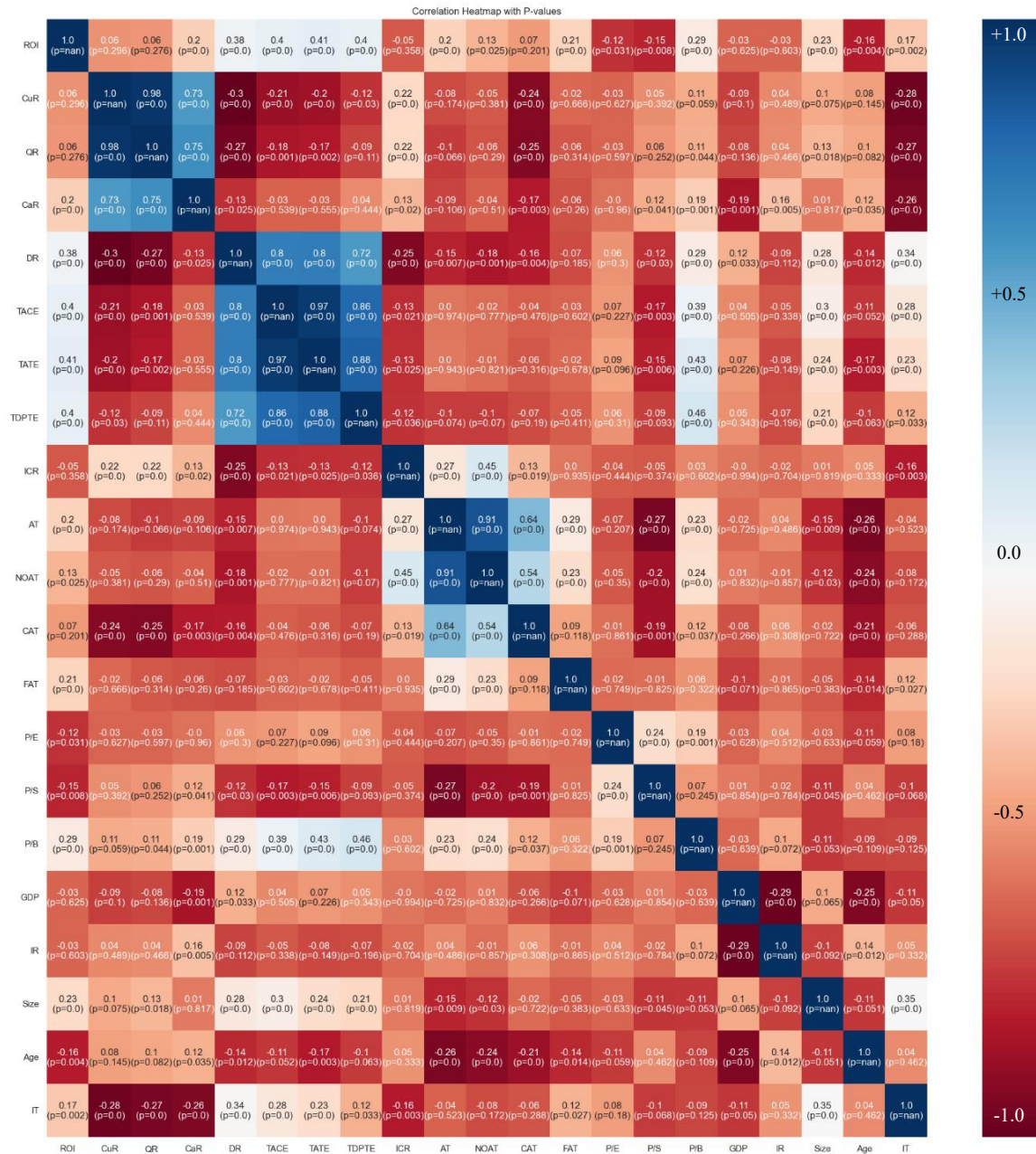
Bahrain's Variables Correlation Matrix



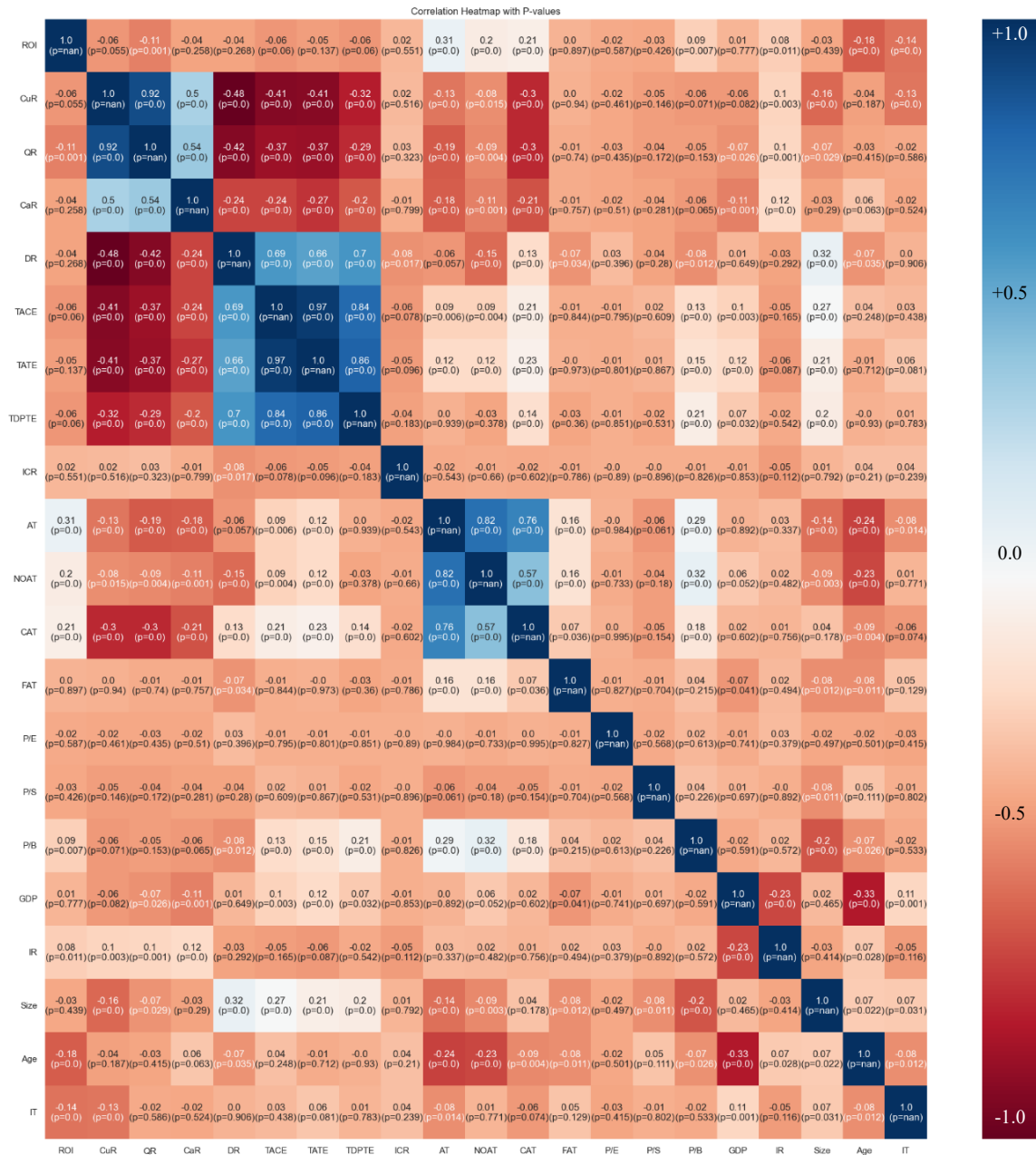
Kuwait's Variables Correlation Matrix



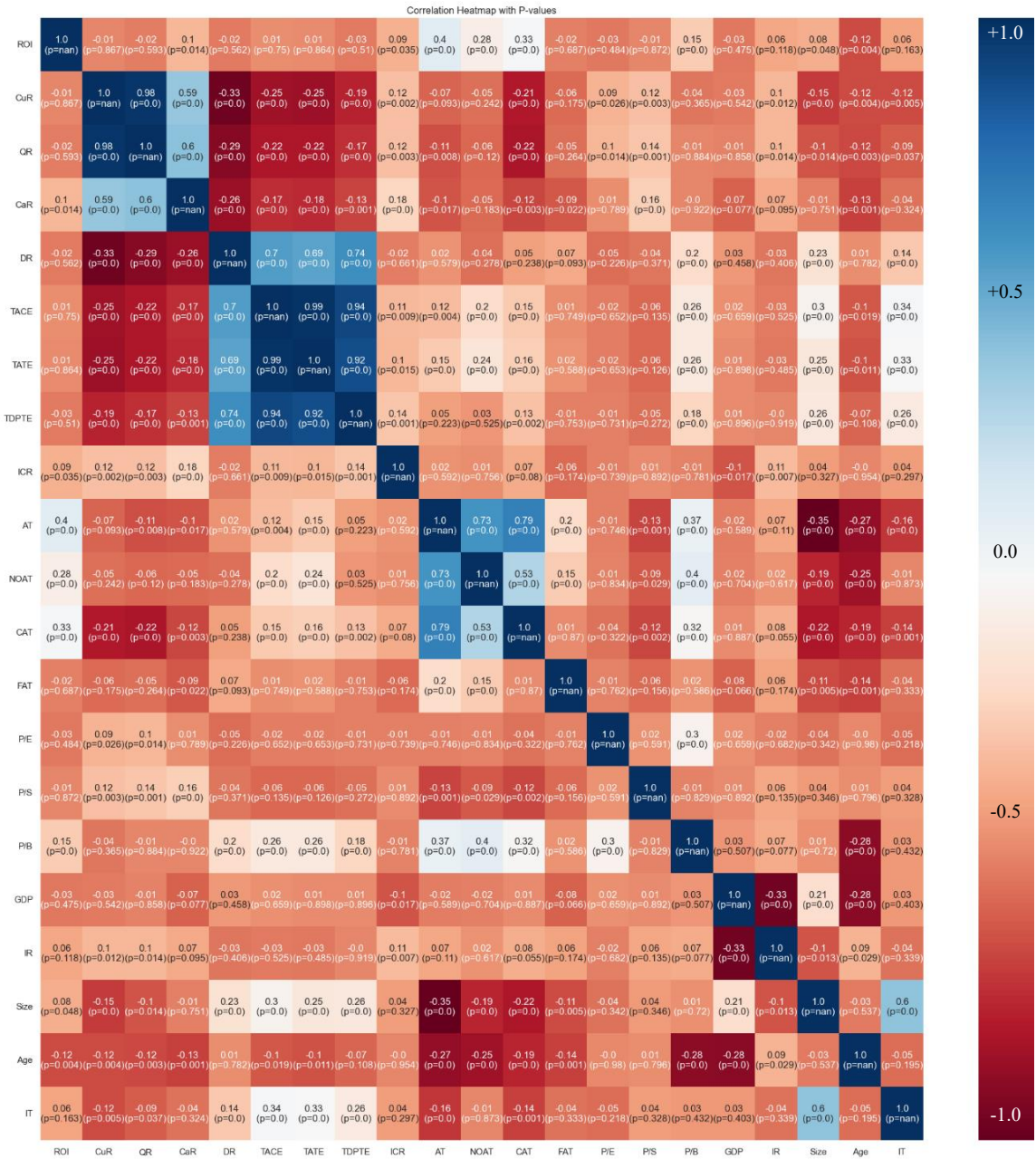
Oman's Variables Correlation Matrix



Qatar's Variables Correlation Matrix

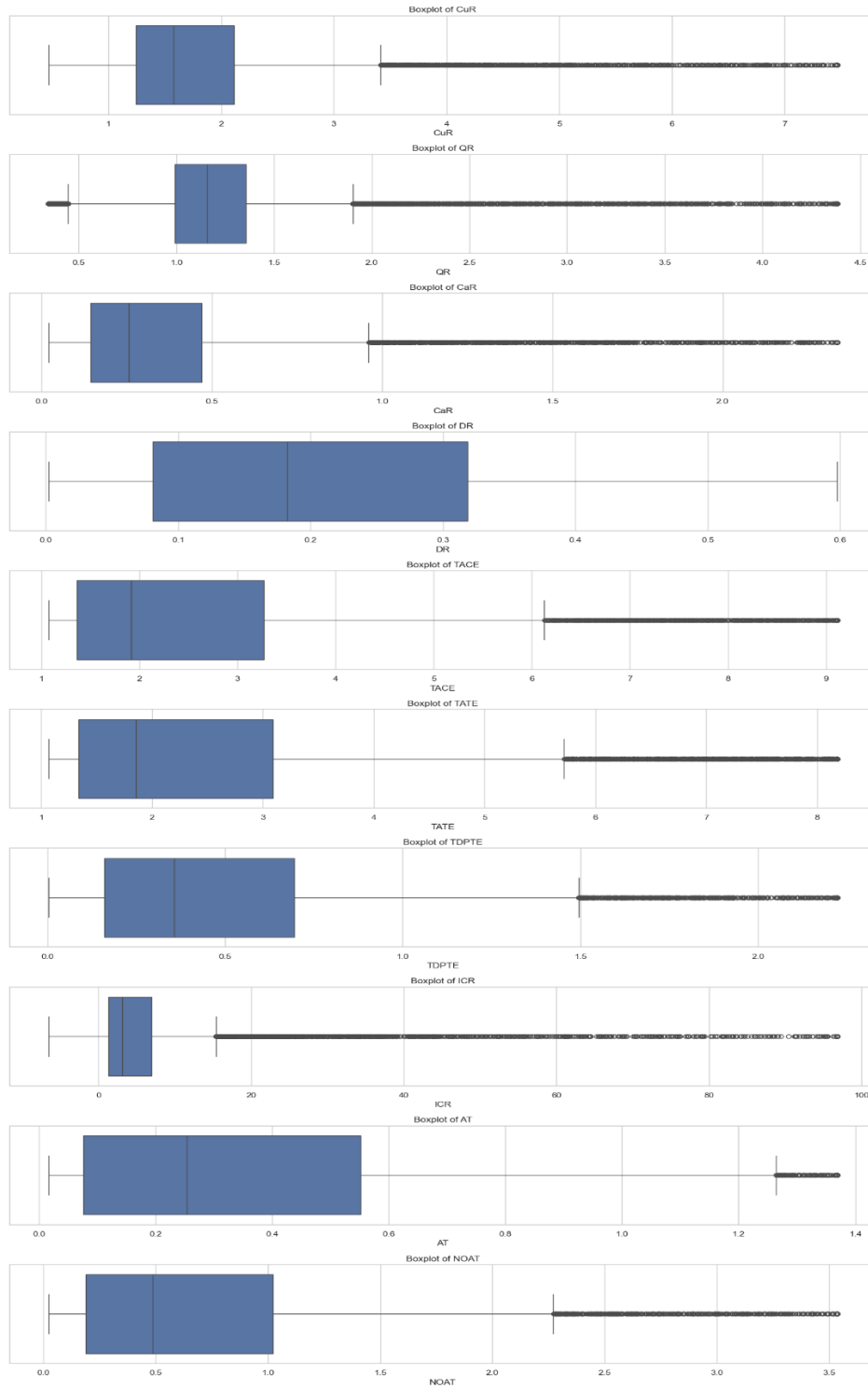


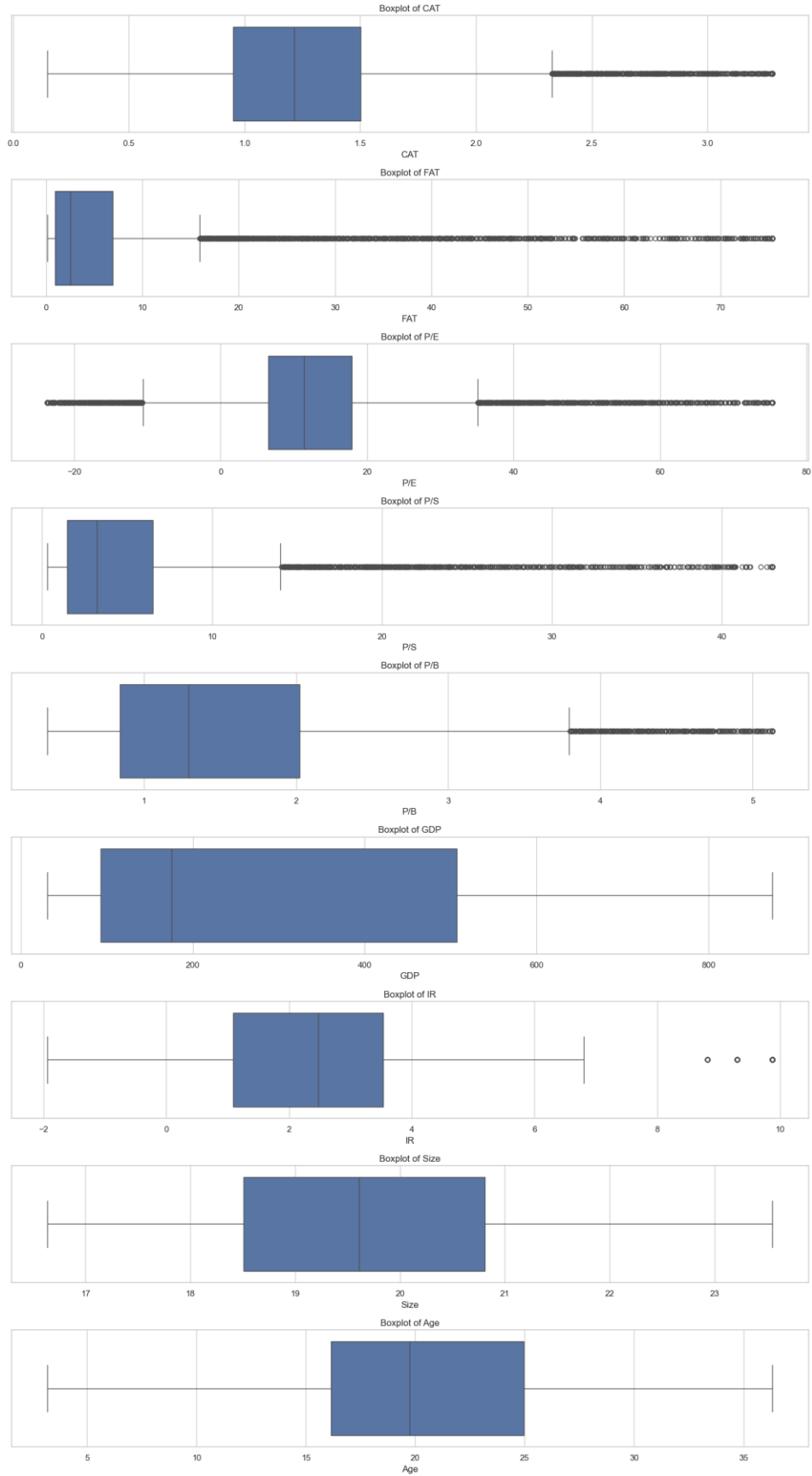
Saudi's Variables Correlation Matrix



UAE's Variables Correlation Matrix

Appendix 6. Distribution of Variables after Winsorization





ARABIC ABSTRACT

الهدف: يهدف هذا البحث إلى دراسة مدى دقة تقنيات التنقيب في البيانات (Data Mining) في توقع (تنبؤ) الأداء المالي للشركات المدرجة في بورصات دول مجلس التعاون الخليجي (مملكة البحرين، الكويت، عُمان، قطر، المملكة العربية السعودية، والإمارات العربية المتحدة).

التصميم/المنهجية/النهج: حللت الدراسة الأدبيات السابقة المتعلقة بتوقع أداء الشركات باستخدام تقنيات التنقيب في البيانات، ودمجت بين المراجعة المنهجية للأدبيات (SLR) والتحليل البليومتري لتحديد فجوات الأدبيات. بناء عليه، تم تطوير نموذج مفاهيمي لدراسة العلاقة بين المتغيرات المستقلة: المتغيرات الخاصة بالشركة (السيولة، والرافعة المالية، والملاءة المالية، واستخدام الأصول، والقيمة السوقية)، والمتغيرات الاقتصادية الكلية (الناتج المحلي الإجمالي والتضخم)، ومتغيرات الضبط (حجم الشركة، وعمرها، ونوع الصناعة، وبلد الشركة)، والمتغير التابع الذي يُقاس بعائد الاستثمار (ROI). تم الحصول على البيانات من مجموعة بورصة لندن (LSEG) ومؤشرات التنمية العالمية (WDI)، وغطت 816 شركة مدرجة في البورصة عبر قطاعات متنوعة في دول مجلس التعاون الخليجي من عام 2003 إلى عام 2022. طورت الدراسة أربعة نماذج توقع (تنبؤ) مختلفة وذلك باستخدام سبع تقنيات لتنقيب البيانات تم تطويرها باستخدام اللغة البرمجية بايثون. من ضمنها التقنيات الفردية: الانحدار الخطي (LR)، والشبكات العصبية الاصطناعية (ANN)، وأشجار الانحدار (RT)، وأقرب جيران (kNN)، إلى جانب تقنيات المجموعات مثل الغابة العشوائية (RF)، وتعزيز التدرج المتطرف (XGBoost)، والتعزيز الغشوي (CatBoost).

النتائج: تفوق نموذج CatBoost على النماذج الأخرى، محققاً أعلى دقة تنبؤ، $R\text{-squared} = 0.542$ لنموذج مجموعة بيانات دول مجلس التعاون الخليجي بالكامل. كما أظهر نموذج البحرين أفضل دقة تنبؤ على مستوى نماذج الدول الفردية ($R\text{-squared} = 0.781$)، مع أدنى معدلات خطأ. في المقابل، أظهر نموذج المملكة العربية السعودية أضعف دقة للتنبؤ ($R\text{-squared} = 0.328$)، مع أعلى معدلات الأخطاء. وفيما يتعلق بأهمية المتغيرات، تم ملاحظة بأن العوامل الداخلية الخاصة، كاستخدام الأصول والملاءة المالية هي الأكثر تأثيراً بشكل إيجابي على ربحية الشركات المدرجة في أغلب دول مجلس التعاون الخليجي. كما أظهرت الدراسة بأن المتغيرات الاقتصادية الكلية، مثل الناتج المحلي الإجمالي والتضخم، لها تأثيراً ضئيلاً على الربحية، بينما تمارس متغيرات التحكم مثل (العمر، نوع الصناعة، البلد، والحجم) تأثيراً متوسطاً، مع تأثير سلبي للعمر على الربحية.

الأصالة/القيمة: من الناحية النظرية، تُضيف هذه الدراسة حادثة إلى الأدبيات من خلال استكشاف قدرات تقنيات تنقيب البيانات في توقع قيمة العائد على الاستثمار كمقياس مالي للربحية. تدرس هذه الدراسة العوامل الخاصة بالشركة التي تؤثر على أداء الشركات المدرجة في بورصات دول مجلس التعاون الخليجي، وتُقيّم المتغيرات الاقتصادية الكلية التي تؤثر على أدائها عبر مختلف الصناعات. تؤكد الدراسة على أهمية تطبيق نماذج إدارة السوق الخاصة بكل بلد بدلاً من استخدام نموذج بيانات شامل لدول مجلس التعاون الخليجي لمراعاة الاختلافات الاقتصادية بين الدول. علاوة على ذلك، تحمل نتائج هذا البحث أهمية تجريبية جوهريّة، حيث تقدم رؤية قيّمة لاتخاذ قرارات مستنيرة بين مختلف الجهات المعنية. يمكن لصانعي السياسات والجهات التنظيمية في دول مجلس التعاون الخليجي، مثل هيئة سوق المال ولجان

الإفلاس، الاستفادة من هذه المعلومات للتدخل المبكر، ومنع فشل الشركات، وتمكين اتخاذ تدابير استباقية للشركات التي تم تحديدها على أنها معرضة للخطر (مثل توقع انخفاض الربحية). بالإضافة إلى ذلك، يمكن للمديرين الاستفادة من الرؤى القائمة على أهمية المتغيرات المستقلة لتحسين العمليات الداخلية وتخصيص الموارد، مما يؤدي في نهاية المطاف إلى تعزيز الأداء المالي. علاوة على ذلك، يمكن للمساهمين اتخاذ قرارات استراتيجية أكثر استنارة بشأن معاملات الأسهم بناءً على نتائج الربحية المتوقعة للشركة.

جامعة البحرين



قسم الإدارة والتسويق

كلية إدارة الأعمال

توقع أداء الشركات في دول مجلس التعاون الخليجي باستخدام تقنيات التنقيب في البيانات

أطروحة مقدمة استكمالاً لمتطلبات درجة الدكتوراه في تحليل الأعمال

مقدمة من

جواهر خالد عارف العنزي

الرقم الجامعي: 19983609

إشراف

أ.د. منور الشمري

(أستاذ)

جامعة البحرين

مملكة البحرين

مايو 2025