



Predicting Stock Performance from Financial Ratios Using Machine Learning

Prepared By

Name: Marwa O. Essa

ID No: 243002281

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Ashraf Elsayed

College of Informatics, Midocean University, Comoros

Email:.....

Acknowledgements

I extend my deepest gratitude to my supervisor, Dr. Ashraf Elsayed, for his invaluable assistance, unwavering encouragement, and insightful critiques throughout my research journey. His steadfast commitment to academic excellence and meticulous attention to detail have significantly shaped the quality and depth of this dissertation.

I am also sincerely grateful to the Teaching Staff Members at the College of Informatics, Midocean University, whose guidance, resources, and encouragement have been invaluable to my academic development.

To my family, I express my deepest love and appreciation. Your unconditional love, patience, and sacrifices have been the foundation of my resilience and success. Without your constant encouragement, this achievement would not have been possible.

Abstract

Using machine learning for stock prediction has gained a lot of interest recently. Many studies show that machine learning algorithms can forecast stock movements effectively by using historical data. However, most current methods mainly focus on short-term predictions based on past stock prices and technical indicators, such as the RSI and MACD. In this study, we take a different approach. We use 41 quarters of financial statement data to explore the possibility of predicting long-term stock returns based on fundamental indicators. The model includes important financial ratios, such as profitability, leverage, efficiency, liquidity, and market value ratios, to create a predictive framework. We trained three machine learning algorithms: RF, LSTM, and GRU. We selected portfolios by selecting stocks with higher predicted returns using these machine learning models. Portfolios were constructed by selecting stocks with higher predicted returns according to these models. The constructed portfolios outperformed the TASI benchmark in the Saudi equity market, achieving higher returns and portfolio values. The findings emphasize the advantages of employing financial ratios in predicting the top 5 high-returning stocks in a mid-to-long-term investment situation, achieving over 33% higher excess returns for all models with lower risk profiles. Also, the selection of the top 3 high-return stocks achieved over 78% relative returns with a reasonable risk level, proving the effectiveness of applying ML techniques based on financial ratios for stock prediction.

List of Publications

Table of Contents

List of Abbreviations	vii
List of Figures.....	viii
List of Tables	viii
Chapter 1: Introducing	1
1.1 Problem Definition	2
1.2 Thesis Objectives	3
1.3 Thesis Contribution	4
1.4 Thesis Organization.....	5
Chapter 2: Overview.....	6
2.1 Introduction	6
2.2 Educational Data Mining	6
2.3 Applications of Educational Data Mining	7
Chapter 3: Related Work	31
3.1 Introduction	31
3.2 History	31
Chapter 4: Proposed Approach	78
4.1 Introduction	78
Chapter 5: Experimental Results and Discussion	86
Chapter 6: Conclusion and Future Work	127
7.1 Important Features	127
7.3 Future Work	129
References	130

List of Abbreviations

- **Acc:** Accuracy
- **GRU:** Gated Recurrent Unit
- **RF:** Random Forest
- **LSTM:** Long Short-Term Memory
- **MACD:** Moving Average Convergence Divergence
- **ML:** Machine Learning
- **OHLCV:** Open, High, Low, Close, Volume
- **ROA:** Return on Assets
- **ROC:** Return on Capital
- **ROCE:** Return on Common Equity
- **ROE:** Return on Equity
- **RF:** Random Forest
- **RSI:** Relative Strength Index
- **MSE:** Mean Squared Error
- **TASI:** Tadawul All Share Index
-

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1 Student Performance Data Set Description	21
2.2 Students' Academic Performance Dataset Description	23

Chapter 1 Introduction

1.1 Problem Definition

Despite the increasing use of artificial intelligence in financial markets, most stock prediction models focus mainly on technical analysis. They often overlook the valuable insights found in corporate financial statements. Traditional machine learning models also have difficulty capturing complex, non-linear relationships between fundamental financial indicators and stock price behavior. This research aims to fill this gap by creating a machine learning-based model that predicts stock trends using fundamental financial data. Specifically, the study looks at how key profitability ratios, such as ROA, ROC, ROE, and ROCE, affect stock price movements. The main goal is to improve the accuracy, clarity, and reliability of long-term stock price predictions by using AI-driven methods.

1.2 Thesis Objectives

This study proposes a new method for estimating the true value of stock returns. It uses information from financial statements and applies machine learning techniques. The goal is to improve the decision-making abilities of investors and financial analysts. This approach offers a data-driven way to evaluate stocks for the long term.

1.3 Thesis Contribution

Evaluating machine learning methods for long-term stock prediction based on fundamental analysis. We do so by comparing the prediction performance of three advanced machine learning methods based on fundamental analysis using financial statements.

1.4 Thesis Organization

Chapter 2 Overview

2.1 Introduction

The main reason for predicting stock price movements is the chance for profit. Since financial markets began, investors have looked for ways to forecast stock performance to find profitable opportunities. This involves buying stocks expected to increase in value and selling them for a profit. Over the years, many stock prediction methods have appeared, but their consistency and reliability are still debated.

Stock prediction methods can generally be grouped into three main categories: Fundamental Analysis. This approach predicts stock prices by looking at the financial health of companies using publicly available data, including financial statements and economic indicators. It aims to estimate a stock's true value.

Technical Analysis. This technique relies only on past market data such as prices and trading volumes. In contrast to fundamental analysis, technical analysis assumes that all public information is already included in a stock's price. It is based on three main ideas:

- Market action discounts everything.
- Prices are moving in trends.
- History repeats itself.

Sentiment Analysis. This method assesses investor sentiment by examining news articles, analyst reports, and online commentary related to specific stocks or the overall market. It seeks to identify the current mood or emotional tone influencing investor behavior.

Among these methods, technical analysis and sentiment analysis are mainly used for short-term predictions, for days or even hours, while fundamental analysis is usually applied for mid- to long-term forecasts, spanning quarters or years. Recently, ML and data mining techniques into financial modeling have gained

significant attention. Many existing ML-based stock prediction studies focus on short-term forecasting using technical indicators or sentiment data. However, these models often neglect real-world factors like transaction costs, which can affect the practical use and reliability of their results.

This research aims to fill this gap by examining machine learning methods for long-term stock prediction based on fundamental analysis. The study specifically investigates the predictive power of financial statement data in estimating future stock performance. To achieve this, we compare the prediction performance of three machine learning models trained on fundamental financial features. Our dataset includes quarterly financial data from 100 companies listed on the TASI (Tadawul All Share Index) between 2015 and 2025. We assess model performance by ranking 70 stocks based on their predicted relative returns and building portfolios accordingly. The actual returns of these portfolios serve as the main measure for evaluating prediction accuracy and model effectiveness.

Chapter 3 Related Work

3.1 Introduction

Stock price forecasting has always been a finance and data science problem since markets are dynamic, volatile, and multi-factorial. Previously, stock prediction was largely reliant on statistical and econometric models that involved histories of past prices. However, with swift advancements in ML and deep learning (DL) in recent years, efforts have been made to utilize more sophisticated methods involving both technical indicators and fundamentals.

3.2 History

In recent years, numerous studies have explored the application of machine learning techniques in stock prediction. Gupta, H., & Jaiswal, A. [1] in 2024 introduced a study on stock forecasting using deep learning and statistical models. The survey motive is to check various deep learning and statistical model techniques for stock price forecasting that are Moving Averages, ARIMA which are statistical techniques and LSTM, RNN, CNN, and FULL CNN which are deep learning models. The result of all the models is different and the best result comes from the LSTM model and we have also seen how Full CNN outweighs the pre-process CNN. This model helps the investors to know when to sell and buy the stock prices. As is also known how the stock prices are affected by the climatic or geographical changes, company profit or loss, change in geographical seasons, loss in the region, inflation, competition in the market etc. Additionally, Tsai, P.-F., Gao, C.-H., & Yuan, S.-M. [2] in 2023 delved into deep learning models for this purpose, they introduced a paper about, Stock selection using machine learning based on financial ratios, found that successful utilization of RF, FNN, GRU, and FinGAT models holds significant managerial implications for enhancing portfolio management strategies. On the other hand, Kunathananon, T. [3] in 2022 introduced Estimating stock price based on information from financial statement using machine learning approaches. This study introduces a tool for stock market investors and institutions constructed from Long Short-Term Memory (LSTM) for predicting stock prices. By

effectively analyzing financial statements and stock market data, the results indicate that LSTM can effectively capture complex relationships in the data and predict stock prices. This research highlights the potential of LSTM as a valuable and innovative tool for investors and institutions in the stock market. In 2022, Z.Y. Lu and S.M. Yuan [6] employed an FNN model with 18 financial ratios and sector dummy variables to predict potential stock returns for the next quarter. This approach improved prediction accuracy by incorporating sectors as dummy variables. In a historical context, the works of Heo, J., & Yang, J. Y. [7] in 2016 this study about stock price prediction based on financial statements using SVM. Han, S., & Chen, R.-C. [8] in 2007. The study used SVM with financial statement analysis for prediction of stocks.

Chapter 4 Proposed Approach

4.1 METHODOLOGY

4.1.1 Data Preprocessing

The sample in this study was collected from the firms included in Tadawul All Share Index (TASI) which has more than 200 companies listed on its main market. To ensure the study's relevance, attention was given to companies within the strict core industries of the index. For these companies, historical overtime data was available on a quarterly basis from the first quarter of 2015 to the first quarter of 2025, and this data was collected using web scraping techniques from publicly available data.

The original data set had a wide range of both financial and non-financial variables, but it was also a dataset with a lot of missing values. A large part of the missing values formed large clusters for some features and other missing values were randomly placed in the dataset. A combination of feature removal and data imputation techniques were found to be adequate to resolve these missing values.

We removed features that had over 50 percent missing data. Moreover, non-fundamental features such as stock high and low prices were removed to sharpen the focus on actual quantitative fundamentals. Even after the above steps were taken, the remaining features still contained some sparse missing values. These were resolved using linear interpolation with adjacent quarter values. If, for example, the ROE for Q3 2015 was absent, we would estimate it by taking the 2Q and 4Q values for 2015, averaging them, and assigning that as the ROE for Q3 2015.

After completing the data preparation procedures, the resulting dataset included 70 companies, each with 41 quarterly records over 10 years, yielding a total of 2,870 records. For further analysis, 13 financial features were extracted, which are detailed in Table 1.

Ratio	Calculation	Description
Current Ratio	Current Assets ÷ Current Liabilities	Measures a company's ability to pay short-term obligations with short-term assets.
Debt to Equity	Total Debt ÷ Shareholders' Equity	Shows the proportion of debt used relative to equity financing.
Long Term Debt/Capital	Long-Term Debt ÷ (Long-Term Debt + Shareholders' Equity)	Indicates the proportion of long-term debt in the total capital structure.
Asset Turnover	Revenue ÷ Average Total Assets	Measures how efficiently assets are used to generate revenue.
Inventory Turnover	Sales ÷ Inventories	Shows how often inventory is sold and replaced over a period.
Receivable Turnover	Net Credit Sales ÷ Receivables	Indicates how effectively receivables are collected.
Days Sales in Receivables	(Receivables ÷ Net Credit Sales) × 365	Measures the average collection period for receivables.
Book Value Per Share	(Shareholders' Equity – Preferred Equity) ÷ Total Outstanding Shares	Represents the equity value allocated to each outstanding share.
Gross Profit Margin	(Revenue – COGS) ÷ Revenue	Shows the percentage of revenue remaining after direct costs.
EBITA Margin	EBITA ÷ Revenue	Measures profitability before interest, taxes, and amortization.
EBIT Margin	EBIT ÷ Revenue	Measures profitability before interest and taxes.
Net Income Margin	Net Income ÷ Revenue	Shows the percentage of revenue remaining as net profit.
Return on Assets	Net Income ÷ Total Assets	Measures efficiency in using assets to generate profit.
Return on Capital	EBIT ÷ (Total Debt + Shareholders' Equity)	Evaluates returns generated from total capital employed.
Return on Equity	Net Income ÷ Shareholders' Equity	Measures return generated on total shareholders' equity.
Return on Common Equity	Net Income Available to Common Shareholders ÷ Common Equity	Focuses on returns to common shareholders, excluding preferred equity.

Table 1. 13 Financial Ratios

4.1.2. Moving Time Period

The RF models encompass a total of 32 quarters for each moving period. The initial data from the moving period span from 2015Q1 to 2023Q1, and so forth. We partitioned these data into three segments: the training set, the validation set, and the test set, comprising 24 quarters, 8 quarters, and 1 quarter, respectively as in Table(2). The training set facilitates model weight optimization, and the validation set aids in finetuning hyperparameters, ensuring model convergence without overfitting. Subsequently, we employ the test set to predict the relative return of the 51 stocks.

Moving time period	Training Set (24 Quarters)	Validation Set (8 Quarters)	Test Set (1 Quarters)	Moving Time Period During Training and Validation
Moving time period 1	2015Q1–2020Q4	2021Q1–2022Q4	2023Q1	Iterative training 1 Quarter as training input for each moving time period
Moving time period 2	2015Q2–2021Q1	2021Q2–2023Q1	2023Q1	
Moving time period 3				
.....				
Moving time period n				

The LSTM and GRU models span a total of 35 quarters for each moving period, with the initial data originating from 2015Q1 to 2023Q3 and so on. We segment this data into the training set, validation set, and test set, consisting of 20 quarters, 6 quarters, and 4 quarters, respectively as in Table (3). Like the previous models, the training set facilitates model weight optimization, and the validation set aids in hyperparameter fine-tuning, guaranteeing model convergence without overfitting.

Moving time period	Training Set (24 Quarters)	Validation Set (8 Quarters)	Test Set (4 Quarters)	Moving Time Period During Training and Validation
Moving time period 1	2015Q1–2020Q4	2021Q1–2022Q4	2023Q1–2023Q3	Iterative training 4 Quarters as training input for each moving time period
Moving time period 2	2015Q2–2021Q1	2021Q2–2023Q2	2023Q3–2024Q3	
Moving time period 3				
.....				
Moving time period n				

4.1.3. Training

4.1.4. Random Forest

We use Pytorch to develop the RF model. Hyperparameters are tuned with the validation set for preventing overfitting where the number of estimators is 1000 and the optimization criterion used SE.

The test result shows that the RF model for generating top 3 as a portfolio with a 109.2% relative return of 109.2% to TASI and top 5 as a portfolio with a 50.1% relative return of 50.1% to the TASI. The 51 composite index achieves only a 24.5% relative return of 24.5% to the TASI index.

In Figure 1, the importance of the random forest characteristics shows that all 13 financial indices are similar in importance.

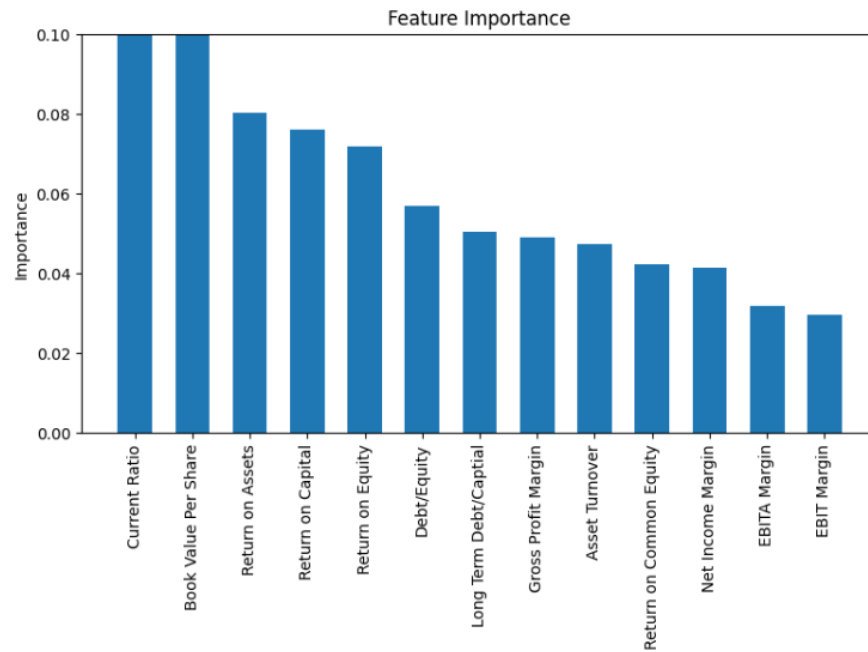


Figure 1. Importance of features of RF model

4.1.4 Prediction Model

LSTM

Name of Hyperparameters	Optimized Result
Hidden size	64
Epochs	100
Batch size	128
Dropout rate	0.4
Input size	13
Layers	2
Learning rate	0.0005
Optimizer	Adam

GRU

Name of Hyperparameters	Optimized Result
Hidden layers	1
Epochs	100
Batch size	128
Time step	12
Dropout rate	0.4
Input size	13
Layers	2
Learning rate	0.0005
Optimizer	Adam

4.1.3 Training

In this study, we tried out two different training techniques: constructing a single model to predict returns for all stocks and constructing a model for each stock individually. These two methods are often referred to as global learning and local learning. Global learning is when a single model is created to predict all stock

market data, offering a wider scope of data to understand relations between various stocks. In comparison, local learning builds individual models for each stock, allowing for bespoke learning tailored to that stock's profile. Global learning might have the advantage of a larger training dataset, while local learning usually outperforms global learning in capturing specific stock patterns richer in detail. In initial tests, the local learning method outperformed every other in the local setting. For this reason, we embraced local learning as a strategy, creating individual models for every stock using all three selected machine learning algorithms.

4.1.4 Evaluation Metrics

4.1.4.1. Excess Return

Excess returns are a useful measure to evaluate the performance of an investment portfolio. By comparing portfolio returns to those of a benchmark, investors can determine whether the portfolio is generating value above the benchmark.

4.1.4.2. Top-k Precision

The precision of k is used to measure the percentage of overlapping between a list of top k stocks with the highest predicted returns, which is 'predicted top k stocks', and a list of top k stocks with the highest actual return, which is 'actual top k stocks'. The higher the precision of the top-k, the more accurate the prediction of the model will be.

Chapter 5

Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

1. Gupta, H., & Jaiswal, A. (2024). A study on stock forecasting using deep learning and statistical models. arXiv.
2. Tsai, P.-F., Gao, C.-H., & Yuan, S.-M. (2023). Stock selection using machine learning based on financial ratios. *Mathematics*, MDPI, 4758.
3. Lu, C., & Xu, Y. (2023). *TRNN: An efficient time-series recurrent neural network for stock price prediction*. *Information Sciences*, 640, 119–134.
4. Kunathananon, T. (2022). *Estimating stock price based on information from financial statement using machine learning approach* (Master's thesis, Chulalongkorn University). Chula Digital Collections.
5. Lu, Z.-Y. A Deep Reinforcement Learning-Enabled Portfolio Management System with Quarterly Stock Re-Selection Based on Financial Statements. Master's Thesis, National Yang Ming Chiao Tung University, Taiwan, China, 2022.

6. Monahan, S. J. (2017). Financial statement analysis and earnings forecasting. *Foundations and Trends® in Accounting*, 12(2), 105–215.
7. Heo, J., & Yang, J. Y. (2016). Stock price prediction based on financial statements using SVM. *International Journal of Hybrid Information Technology*, 9(2), 57-66.
8. Han, S., & Chen, R.-C. (2007). Using SVM with financial statement analysis for prediction of stocks. *Communications of the IIMA*, 7(4), Article 8.