5-23-2024

Machine Learning Models for Enhanced Stock Trading Strategies

Alyaa Al Ali
aa4797@rit.edu

Machine Learning Models for Enhanced Stock Trading Strategies

by

Alyaa Al Ali

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology – Dubai

Graduate Thesis Committee:

Dr. Sanjay Modak

Dr. Hammou Messatfa

Date of Approval: 23 May 2024

Acknowledgments

This thesis would not have been possible without the support and guidance of all those who contributed to its completion. My deepest gratitude extends to Dr. Hammou Messatfa, whose invaluable mentorship and insightful guidance have been the cornerstone of my research endeavors. Throughout this academic work, he has guided and inspired my future aspirations with patience and deep understanding.

I would like to thank Dr. Sanjay Modak, the chair of Rochester Institute of Technology's graduate programs and research. As a result of his leadership and scholarly advice, my research experience has been enriched significantly.

I am deeply grateful for the unwavering support of Rochester Institute of Technology's Data Analytics Department. Their environment has also facilitated my research journey by providing an intellectually stimulating environment.

As part of my studies, I have participated in an internship provided by the Telecommunications and Digital Government Regulatory Authority (TDRA). Having the opportunity to apply my academic knowledge in a professional environment is very rewarding for me.

My family has been supportive throughout this journey. They have been a source of motivation and strength for me because of their unconditional love, encouragement, and belief in my abilities.

I dedicate this thesis not only to my efforts but also to those who supported me. All of you are truly appreciated.

Alyaa Al Ali

24/03/2024

Abstract

The purpose of this research is to develop sophisticated machine-learning models that can predict stock price movements, taking into account the multifaceted influences that create volatility in financial markets. To navigate the complexities of market dynamics effectively, investors, financial institutions, and academia rely on precision stock price predictions. This research generates actionable buy and sell signals by using historical stock data, technical market indicators, Linear Support Vector Machines (LSVM), Neural Networks, and Logistic Regression. A distinctive feature of this model is it uses market indicators such as PPO, MACD, RSI Signal, Bollinger Bands, ROC Signal, and DX Signal instead of conventional methods that might include sentiment analysis from external data sources. By creating additional indicators, this thesis broadens the analytical scope by using a carefully curated dataset from Yahoo Finance, with a focus on Johnson & Johnson. In addition to its improved prediction accuracy, precision, and F1 scores, the proposed model also allows traders to make informed decisions about when to buy or sell, potentially enhancing portfolio performance by as much as 70%. Among all models, Logistic Regression emerged as the best performer, yielding an impressive 96.467% accuracy for predicting stock price movements. As part of this research, we examine how stock price prediction mechanisms work in detail and introduce a model that combines technical analysis with machine learning insights, paving the way for more accurate and reliable investment decisions.

Key Words: Stock Price Prediction, Machine Learning, Linear Support Vector Machines (LSVM), Neural Networks, Logistic Regression, Financial Markets, Predictive Models, Technical Market Indicators, Investment Strategies, Data Analysis.

Table of Contents

List of Figures

List of Tables

Chapter 1- Introduction

1.1 Introduction

As stock trading becomes more complex and fast-paced, investors and analysts find it increasingly difficult to make informed decisions. Data-driven models are utilized in the thesis to navigate the complexities of the stock market, focusing on optimizing investment strategies. This study employs Linear Support Vector Machines (LSVM), Neural Networks, and Logistic Regression to predict the Price Oscillator (PPO) signal, specifically the critical decision points when buying or selling a stock. Instead of incorporating sentiment analysis from financial news or social media feeds into traditional approaches, this research uses only technical market indicators, such as PPO, MACD, RSI Signal, Bollinger Bands, ROC Signal, and DX Signal, to make trading decisions.

In this study, data for Johnson & Johnson (JNJ) was sourced from Yahoo Finance. In addition to aligning with the thesis's objective of predicting buy or sell signals through machine learning, this selection addresses the limitations posed by the dataset's limited attribute availability. Using market indicators' sophisticated techniques and results to create additional attributes, this work extends the dataset's utility and explores diverse trading strategies.

The breakthrough in this research is the model's remarkable accuracy, precision, and F1 score, which indicate the potential to significantly influence stock trading decision-making. This thesis aims to enhance the effectiveness of investment strategies in the volatile market domain by leveraging a unique combination of technical analysis and machine learning.

In addition to simplifying the decision-making process for traders, this thesis provides a basis for further research into the effectiveness of various market indicators, which can be used to improve the accuracy and reliability of stock trading strategies using innovative financial analytics solutions.

1.2 Problem Statement

As financial markets become increasingly dynamic and unpredictable, predicting stock market trends and optimizing investment strategies become increasingly difficult. Stock price predictions remain unreliable and inaccurate despite sophisticated analytical tools, exposing investors and

financial institutions to considerable financial risk. Investing in the stock market presents several challenges:

1- Volatility in the stock market is influenced by a variety of factors, including economic indicators, geopolitical events, and market sentiment. The unpredictability of stock price movements complicates forecasting, thus reducing the effectiveness of traditional predictive models.
2- The market is subject to external shocks, such as natural disasters, political instability, and global health emergencies. Due to their unpredictable nature, external shocks can cause rapid fluctuations in the market, complicating investment decisions.
3- Investing involves a constant risk of loss due to the volatility and unpredictability of financial markets. Even though traditional trading strategies are useful, they rarely address the complex and multifaceted nature of market dynamics.

This thesis proposes a predictive model that utilizes Linear Support Vector Machines (LSVM), Neural Networks, and Logistic Regression to address these challenges. The model helps traders refine their trading strategies by predicting the Price Oscillator (PPO) signals for buying or selling stocks, thereby allowing them to identify optimal entry and exit points, place stop-loss orders, and utilize other strategies to capitalize on market movements. The model seeks to provide a deeper understanding of future stock movements through the integration of technical indicators such as PPO, MACD, RSI, Bollinger Bands, ROC, and DX Signal. Ultimately, we hope to improve the performance of wealth client portfolios by up to 70%, as a result of providing a robust investment tool to simplify the process of navigating the complexities of the stock market.

1.3 Project Goals

Research Goals:

1- Utilize technical indicators and historical data to construct an advanced machine learning model that predicts stock market movements based on metrics such as open, close, high, low, and volume.
2- By using the insights from the predictive model, investors can potentially boost their portfolio performance by up to 70%.

3- Using precision, accuracy, and F1 scores as benchmarks, evaluate the model's accuracy in predicting stock prices.
4- Provide traders with an informed trading decision-making tool to minimize the impact of stock market fluctuations on their financial portfolios.

Research Questions:

1- How effectively can a combination of historical stock data and technical market indicators be utilized to construct a machine-learning model for precise stock price predictions?
2- Is it possible for the predictive model to significantly uplift the performance of investment portfolios, aiming for a 70% enhancement?
3- What level of predictive accuracy does the model achieve, and how does this align with investor expectations for reliable stock market forecasting?
4- Can the adoption of this machine learning-based predictive model substantially mitigate the financial risks traditionally associated with stock market investments?

1.4 Aims and Objectives

This research aims to develop a predictive model that can accurately predict stock market trends by leveraging advanced machine learning techniques. To accomplish these goals, sophisticated analytical methodologies will be combined with actionable investment strategies. The specific objectives are as follows:

1- Develop a Predictive Model: Utilizing technical market indicators alongside extensive historical stock data to predict stock price movements accurately.
2- Improving the performance of investment portfolios: Based on the model's insights, investors may be able to improve their portfolio performance by up to 70%.
3- Measure model accuracy: Ensure reliability by carefully assessing the prediction model's precision, accuracy, and F1 score metrics.
4- Achieve Financial Risk Minimization: Offer investors a tool to help them make informed decisions about stock market participation to minimize financial risks.
5- Advance Financial Data Analytics: Present a breakthrough approach to stock price prediction, enabling investors to engage in the market more strategically and confidently.

1.5 Structure of the Thesis

The thesis follows the following structure:

Chapter 1: Introduction

In this chapter, the topic of stock price prediction is introduced, along with objectives, research questions, and limitations.

Chapter 2: Literature Review

A comprehensive review of over 30 papers is presented in this chapter. As such, it encompasses all the major points relevant to this thesis.

Chapter 3: Research Methodology

In this chapter, we describe the methodology used for this project, specifically CRISP-DM (Cross-Industry Standard Process for Data Mining).

Chapter 4: Data Analysis Process

The chapter discusses in detail the processes of data exploration, cleansing, feature engineering, model building, and the presentation of model results.

Chapter 5: Discussion of Results

In this chapter, the outcomes of the thesis are discussed, the research contributions are outlined, and the original research questions discussed in Chapter 1 are addressed.

Chapter 6: Conclusion

In the final chapter, the thesis is summarized, with learnings, accomplishments, and areas for improvement addressed.

1.6 Limitations of the Study

Even though this study represents an advancement in utilizing machine learning techniques to predict stock market trends, it does come with limitations that need to be addressed carefully. Our focus on Johnson & Johnson, while providing a detailed examination, narrows the scope, thus

making our conclusions inapplicable to other industries or varying market conditions. A singular focus may not fully reflect the breadth of market complexity.

In addition, the model's sole reliance on technical indicators and historical stock data ignores the effects of external factors on market behavior by excluding sentiment analysis from news and social media. Though beneficial for methodological clarity, this deliberate omission potentially ignores the sway of public sentiment on stock valuation. The data available for the development and evaluation of the predictive model are also limited in time and quality. It is not possible to account for all market fluctuations or novel market scenarios precipitated by unforeseen global events in the historical data.

Additionally, since financial markets are intrinsically volatile, such outcomes should be approached with prudence following the study's ambitious goal of improving portfolio performance by 70%. While the predictions are highly accurate and precise, market forces are unpredictable. There is also a notable limitation in the lack of real-time data application. Until the model's performance is validated against live market dynamics, its robustness remains hypothetical. As the dataset was limited to one stock, the study did not validate and enhance the reliability and versatility of the model using a broader set of data across various stocks and sectors.

Further, expert evaluation, a crucial step in actual validation, was not sought. It would have been helpful if financial market experts had provided feedback to assess the model's practicality and relevance to real-world trading.

Last but not least, the use of only a few machine learning algorithms—LSVM, Neural Networks, and Logistic Regression—paints an incomplete picture. This study lacked the exploration of alternative computational techniques, which might have provided more detailed insights and improved the model's adaptability and robustness.

As a result of recognizing these limitations, further research endeavors can be carried out. In the future, the dataset should be expanded to include stocks and sectors from a wider variety of fields, real-time data analysis should be included, expert feedback should be obtained, and a wider range of machine learning algorithms should be employed.

Chapter 2- Literature Review

Traditional stock price prediction models rely heavily on technical and fundamental analysis. Machine learning has gained prominence in recent years as a powerful tool for improving these predictions. Recent studies indicate a strong link between the analysis of news articles by machine learning and subsequent stock price movements, underscoring the importance of integrating sentiment analysis and real-time data into predictive models.

2.1 Literature Review

Shen et al. [1] used key stock indices from across the world as input features for machine-learning models to anticipate market changes. Overseas markets that shut before or at the start of US market trading can provide important insights into the next US trading day, reflecting market sentiment on recent economic news and world events. Furthermore, because of the interrelated nature of many financial markets, commodities prices and foreign currency data are seen as potential features. For example, a downturn in the US economy might affect the stock market, causing currency values to fluctuate. This interconnection means that the behavior of one market can be used to forecast movements in others.

According to Pahwa et al. [2], supervised learning is a fundamental strategy in machine learning where examples influence the construction of functions. It generates well-behaved functions with the correct training set. This method is great for algorithms because it is based on clear, numerical data. There are two types of supervised learning: regression and classification. Methods such as Support Vector Classification (SVC) aid in the resolution of problems in regression. When the Support Vector Machine (SVM) is applied to regression, it is referred to as Support Vector Regression (SVR). Both SVC and SVR models are strongly reliant on training data, which is affected by the cost of developing the model. This cost rejects training points above the margin, stressing the need for well-defined data in supervised learning.

According to Vijh et al. [3], the difficulty in projecting stock market returns comes from recognizing the complicated patterns driven by many elements. The historical data available on the company's website is limited to high, low, open, and close prices, as well as trading volume. New variables are formed from existing ones to improve accuracy. To estimate the next day's

closing stock price, Artificial Neural Networks (ANN) are used, and Random Forest (RF) is also used for comparison. When the models' RMSE, MAPE, and MBE values are compared, ANN surpasses RF, with RMSE (0.42), MAPE (0.77), and MBE (0.013) implying its higher predictive performance. In future investigations, integrating financial news items and other characteristics such as profit and loss statements could enhance forecasts even further.

As stated by Leung et al. [4], understanding the stock market, which is influenced by a variety of factors, is difficult. To understand the complicated linkages between corporations that affect stock prices, they are represented as a graph with nodes representing companies and edges representing cooperation. Edges are provided based on major disparities between search engine results for collaboration and competition inquiries. This approach, which makes use of the Bing Search API, assists in studying correlations between the stock prices of companies, delivering significant information.

Usmani et al. [5] investigated numerous aspects influencing market performance in their article. The researchers evaluated several factors:

1- Market History: The model used the KSE-100's historical closing index after applying statistical techniques such as ARIMA and SMA. For analysis, a window size of 4 was chosen.
2- THE NEWS: For market performance analysis, specific types of news, such as business, financial, political, and international developments, were considered.
3- General Public Mood: To understand the collective mood of investors, social media, particularly Twitter, was used to evaluate public opinion, which influenced market performance.
4- Commodity Price variations: Price variations in vital commodities such as gold, silver, and petrol were investigated because these changes frequently impact numerous sectors, reflecting on market behavior.
5- Interest Rate: The State Bank of Pakistan's interest rates, notably the 1-week Karachi Inter Bank Offer Rate (KIBOR), were considered. These rates influence banks that make loans, influencing market dynamics.

6- Foreign Exchange: The model incorporates historical exchange rates between the Pakistan Rupee (PKR) and the US Dollar (USD), acknowledging the impact of currency movements on market performance.

Various machine learning algorithms were applied to BSE data to anticipate market patterns in the study done by Kohli et al. [6]. For analysis, the data was converted and divided into training and test sets. The researchers used Random Forest, SVM, Gradient Boosting, and AdaBoost algorithms, with results ranging from 68.4% to 78.95% accuracy. A bar graph was used to visualize the comparison of various methods. This study highlights the efficiency of machine learning approaches in anticipating market movements, providing investors and analysts with insights into accurate prediction methodologies.

Machine learning methods for stock market prediction were investigated in a paper by Strader et al. [7]. The best-fit methods for specific prediction challenges were identified in the study: artificial neural networks for numerical index values, support vector machines for classification tasks, and genetic algorithms for portfolio optimization. The study stressed the importance of generalizability enhancements, recommending testing across several markets, periods, and market situations. Furthermore, the incorporation of financial investment theory into machine learning models was emphasized, emphasizing the need to take into account known financial rules. The study advocated for more transparent failure reporting and highlighted the competitive nature of stock market forecasting in the investing landscape.

Patel et al. [8] research methodically investigates various approaches for stock market forecasting, intending to arm investors with knowledge for informed decision-making. The stock market has a large impact on a country's economic environment, influencing employment rates and many industries. However, because of its ever-changing character, influenced by political events and economic crises, among other things, only 10% of the public is willing to invest in this dynamic sector. Stock prices change due to the delicate balance of supply and demand; if more people want to buy a stock, the price rises, and vice versa. To effectively predict these fluctuations, researchers investigate approaches such as Neural Networks, which are part of the broader subject of machine learning.

Umer et al. [9] researched the interesting realm of the stock market. The stock market is similar to a large marketplace where people purchase and sell shares in various companies.

Consider it a gigantic puzzle in which experts examine how the prices of these shares fluctuate - sometimes up, sometimes down. These movements, known as trends, aid in forecasting what will occur next. To understand these tendencies, scholars examine massive amounts of historical data. It's similar to researching past weather trends to determine if it will rain tomorrow. Experts analyze this data to determine the optimum periods to buy and sell stocks, assisting investors in making sound selections. Consider the stock market to be a large, bustling city, and tends to be the highways that guide traders where to go. While these forecasts are not always accurate, they serve as useful guides for investors, giving them the many pathways, their investments could take.

Choudhry et al. [10] investigate how technical analysis can assist in predicting stock market changes. Unlike the notion that stock prices are completely random, technical analysis predicts future patterns based on past prices and trading volumes. To better comprehend these trends, they employ unique mathematical formulas such as the stochastic oscillator. Support Vector Machines (SVMs), which are great at managing complex data patterns, are also introduced in the study. They focus on the Indian stock market, emphasizing the importance of taking into account each market's specific characteristics when employing artificial intelligence for.

Hegazy et al. [11] use advanced machine learning approaches to dive into the delicate domain of stock price prediction in their thorough research. They begin by addressing the difficulties provided by the unpredictable nature of stock time series, emphasizing the limits of developed approaches like as Artificial Neural Networks (ANNs). To address these issues, they employ Support Vector Machines (SVMs) and introduce Least Squares-Support Vector Machines (LS-SVM), an approach that requires careful parameter selection. Recognizing the importance of parameter optimization, the authors propose combining Particle Swarm Optimization (PSO) and LS-SVM. PSO is an evolutionary algorithm inspired by social behavior in organisms that finds the best combination of LS-SVM parameters. This unique hybrid approach attempts to improve the accuracy and dependability of stock market forecasts, opening the door to more effective financial forecasting. The new method used in this study combines the benefits of LS-SVM and PSO, bridging the gap between traditional machine learning and evolutionary algorithms. By using PSO to optimize LS-SVM parameters, the researchers hope to improve stock market predictions, paving the way for more precise investing strategies and informed decision-making.

Huang et al. [12] sought to create a framework for guiding long-term stock portfolio selection in their study. They projected stock prices using regression models and assessed stocks based on their predicted relative returns. They utilized the Root Mean Square Error (RMSE) as the training loss function for some models, such as FNN and ANFIS, to evaluate the portfolios. The equities were then assessed by predicted relative returns, with the top one-third selected for portfolio inclusion. The true relative return of these portfolios was estimated and evaluated using a modified Sharpe ratio, a popular financial metric emphasizing risk-adjusted returns.

In their study, Tsai et al. [13] stated that in the domain of predicting stock prices, researchers focus on two major methods: fundamental analysis, which examines a company's financial data, and technical analysis, which investigates historical market trends. Combining these approaches, and utilizing their strengths for better forecasts, has become popular. Artificial Neural Networks (ANNs) are data-driven artificial brains. To effectively handle information, they use the back-propagation mechanism. Another method is Decision Trees (DTs), which use tree-like structures to create predictions. They're simple to grasp and can be improved with pruning techniques. When these methodologies are combined, they produce a hybrid model that provides accurate insights into stock market behavior.

The focus of Reddy and Kranthi's [14] article is on how quantitative traders use machine learning techniques, specifically Support Vector Machine (SVM), to forecast stock market patterns. Traditional methods like as fundamental analysis (evaluating intrinsic stock value) and technical analysis (studying market statistics) have been employed, but machine learning has grown in popularity in recent years. The project creates a financial data predictor algorithm with previous stock prices as training data, to reduce uncertainty in investing decisions. Stock price forecasting is difficult due to market volatility and a mix of known (prior day's closing price, P/E ratio) and unknown factors (election results, rumors). Various research projects employ machine learning for predicting stock prices, with the objective timeframe, stock selection, and predictors used varying. Support Vector Machine (SVM) technique is used to produce real-time forecasts in financial markets, using historical data with computational improvements to improve predictive systems.

In their study, Kaya and Karslıgil [15] investigated the predictive potential of financial news articles on stock prices, incorporating data mining techniques to enhance forecasting

accuracy. They developed a model that categorizes each financial article as either positive or negative based on its impact on stock prices. This categorization was used to train a support vector machine (SVM) classifier. Distinctively, their methodology utilized pairs of nouns and verbs from the text, rather than single words, to capture more detailed semantic relationships. The findings from their study suggest that this novel approach can significantly improve the accuracy of stock price predictions, underscoring the efficacy of integrating detailed textual analysis into financial market forecasting. The innovative techniques mark a significant advancement in the field of financial analysis, providing deeper insights into the dynamics that influence market movements.

In their study, Hadavandi et al. [16] explored stock price forecasting by integrating genetic fuzzy systems (GFS) and artificial neural networks (ANNs). The researchers aimed to enhance predictive accuracy by fusing these technologies to capitalize on their strengths. Initially, they utilized stepwise regression analysis to identify the most influential variables affecting stock prices. They then employed self-organizing maps (SOM) to cluster the data, simplifying the data set into manageable subsets. Each cluster was further analyzed using a GFS, which autonomously extracts rules and tunes its database, ensuring precision in predictions. Their methodology demonstrated superior forecasting performance for selected stocks in the IT and airline sectors, surpassing traditional models. This hybrid approach highlights the effectiveness of combining multiple AI technologies to tackle the complexities of financial markets.

In their study, Moghar and Hamiche [17] focus on the use of Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) for predicting stock market values. Their research reveals that LSTM models, which are capable of capturing information from previous data for future predictions, significantly enhance prediction accuracies over traditional models. By adjusting the number of epochs in their model training, they effectively demonstrate the potential of LSTM RNNs in financial forecasting, emphasizing their superiority in handling sequence prediction tasks.

Pang et al. [18] introduce an innovative neural network approach that integrates an embedded layer with LSTM for stock market prediction. Their study highlights the development of 'stock vectors' similar to word vectors in natural language processing, enabling a more effective representation of financial data. This innovative approach tackles the limitations of traditional neural networks by incorporating more complex architectures capable of handling multi-

dimensional stock data. Their methodology not only refines prediction accuracies but also integrates the Internet of Multimedia of Things (IMMT) for comprehensive stock analysis. This significant advancement in predictive analytics demonstrates the potential of combining advanced neural network models with real-time data processing for financial markets.

Mohan et al. [19] explored stock price prediction using news sentiment analysis. In their research, they argued that integrating historical stock data with real-time financial news could enhance prediction models, due to the significant impact of news on stock movements. They collected a large dataset of S&P 500 stock prices and over 265,000 related financial news articles, applying deep learning techniques to predict stock price trends. Their findings reinforced the strong correlation between news sentiment and stock prices, demonstrating the potential of combining textual information with traditional data sources for more accurate predictions.

Agrawal et al. [20] developed a model using deep learning to predict stock prices based on technical indicators. Their study focused on optimizing Long Short-Term Memory (LSTM) networks to interpret various stock technical indicators, aiming to improve the accuracy and reliability of stock price predictions. By incorporating a deep learning approach, they addressed the complexities of the stock market's volatile nature and demonstrated improved predictive performance over traditional machine learning models.

Agrawal et al. [21] resume to investigate the use of technical indicators for stock price prediction through an optimal deep learning model. They optimized LSTM networks to analyze and predict stock trends more effectively, utilizing adaptive stock technical indicators to refine their model. Their results showed a significant improvement in prediction accuracy, highlighting the benefits of using deep learning to analyze complex financial data for making informed trading decisions.

Emioma et al. [22] investigated the application of machine learning to predict stock prices using least-squares linear regression. Their study focused on the predictability of stock price movements by analyzing historical data points related to stock prices and using these as inputs for a linear regression model. This method simplifies prediction by assuming a linear relationship between date and stock prices, disregarding other potentially influential variables. They aim to provide a simple yet effective approach for traders, especially those engaged in day trading, to forecast future stock prices with a reasonable degree of accuracy.

Siew and Nordin [23] focused on developing a hybrid machine learning model for forecasting stock prices by integrating Artificial Neural Networks (ANN) and Decision Trees (DT). Their research aimed to combine the predictive accuracy of ANNs with the rule-generating capabilities of DTs to enhance both the performance and interpretability of stock market forecasts. The study highlighted how such a hybrid approach could effectively manage the complexities of the stock market, providing both robust predictions and clear, actionable rules that help investors make informed decisions. The implementation of this model demonstrated improved accuracy in predicting stock price movements and provided a systematic investment approach, supporting better decision-making in the financial sector.

Mehtab and Sen [24] analyzed stock price prediction using a combination of machine learning and deep learning models in their study "A Time Series Analysis-Based Stock Price Prediction." They employed historical stock data to train various models, including ARIMA and LSTM, to forecast future stock prices. Their results highlighted the robustness of deep learning models, especially LSTM, in capturing temporal patterns in stock data, thus providing accurate forecasts and valuable insights for financial market participants.

Mehtab et al. [25] continue to explore stock price prediction using both machine learning and deep learning LSTM models. Their study employed a combination of regression models and LSTM-based deep learning models using historical data from the NIFTY 50 index of the National Stock Exchange of India. They demonstrated that the LSTM-based univariate model, which utilized data from the previous week to predict the next week's open values, was the most accurate. This research showcases the effectiveness of integrating machine learning techniques with LSTM networks to enhance predictive accuracy in the stock market.

Nikou et al. [26] conducted research titled "Stock price prediction using DEEP learning," which compares deep learning algorithms with traditional machine learning methods for stock price forecasting. The study emphasizes the effectiveness of deep learning in handling the complexities of financial data and achieving higher prediction accuracy, presenting a significant advancement over traditional forecasting methods.

Nousi et al. [27] investigated machine learning techniques for forecasting mid-price movements using limit-order book data in their study. By employing regression, decision trees, and neural networks, this research highlights the power of machine learning in decoding complex

market data and enhancing decision-making in high-frequency trading scenarios. The results underscore the potential for these techniques to significantly improve financial market predictions.

Ghosh et al. [28] explored stock price prediction using LSTM networks focused on the Indian share market. They identified the challenge of predicting stock prices due to the volatile nature influenced by various factors like investor sentiment and market news. Emphasizing the efficiency of machine learning, particularly LSTM models, they utilized historical stock price data to forecast future trends. This approach is rooted in the theory that past stock prices encapsulate all relevant market influences and information, making them a reliable basis for prediction. The LSTM model, known for handling long-term dependencies effectively, was chosen to analyze and predict future growth of companies, demonstrating its capability to unearth patterns previously unrecognized by traditional models.

Singh [29] focused on predicting the Nifty 50 Index using eight supervised machine learning models: AdaBoost, k-Nearest Neighbors (kNN), Linear Regression (LR), Artificial Neural Network (ANN), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and Decision Trees (DT). The study used historical data from the Indian Stock Market spanning 25 years, from April 22, 1996, to April 16, 2021. The dataset included 6220 trading days and was divided into four subsets of different sizes, each further split into training and testing data. The study conducted three tests: Test on Training Data, Test on Testing Data, and Cross Validation Test to compare the predictive performance of the models. Findings indicated that AdaBoost, kNN, RF, and DT underperformed with increasing data size. LR and ANN showed almost similar results across all models, although ANN required more time for training and validation. SVM performed well initially, but SGD outperformed SVM as the dataset size increased.

K. Pahwa et al. [30] developed an LSTM-based RNN model to predict stock market movements. The model, which outperformed traditional machine learning algorithms, was refined through thorough data collection and preprocessing to enhance accuracy. It shows promise for both individual traders and corporate investors by providing reliable forecasts of market trends. Future improvements will focus on incorporating more diverse market features and user reviews to further enhance its predictive capabilities.

Nelson et al. [31] proposed a model using LSTM neural networks to predict stock price movements on the Brazilian stock exchange, Bovespa. They utilized historical price data and technical indicators as inputs to forecast whether the price of a stock would rise within the next 15 minutes. The model was trained and tested with data from 2008 to 2015, and evaluated using various metrics such as accuracy, precision, recall, and F1 score. The results showed that the LSTM model generally outperformed traditional machine learning algorithms and investment strategies, achieving up to 55.9% accuracy in predicting price increases. This model could significantly aid traders and investors by providing short-term price movement predictions.

2.2 Main Takeaways

1- The research utilized insights into how momentum oscillators, particularly the PPO, can effectively signal potential reversals and continuations in market trends. By integrating these findings, the model was able to harness the PPO's sensitivity to price movements, providing a comprehensive approach that goes beyond basic price data to interpret market dynamics.

2- Drawing on studies that explored the combination of various technical indicators, the research applied this by integrating the PPO with other indicators like moving averages and RSI (Relative Strength Index). This approach enhanced the predictive accuracy of the model by providing multiple layers of market analysis, confirming the literature's suggestion that multi-indicator strategies bolster the reliability of predictions.

3- The literature emphasized the importance of adapting technical indicators to different market environments. The research applied this by adjusting the PPO parameters according to the volatility and trend characteristics of the stock being analyzed. This flexible approach allowed for more accurate predictions across various stocks, mirroring the successful application of adaptive strategies noted in the literature.

4- Consistent with the literature that focuses on statistical testing to validate the performance of technical indicators, the research employed back-testing and cross-validation techniques. This not only provided empirical evidence of the PPO’s effectiveness but also aligned with best practices in financial modeling to ensure that the findings were robust and reliable.

5- Incorporating insights from behavioral finance, particularly regarding how traders react to oscillator readings, enriched the interpretative power of the model. By understanding the psychological factors that influence trader decisions at key oscillator levels, the research could more accurately predict buying and selling pressure, thereby enhancing the practical utility of the PPO in real-world trading scenarios.

Chapter 3- Research Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) will methodically guide the data mining initiative through distinct phases. CRISP-DM offers a structured framework for conducting data mining projects, covering conceptualization through deployment. As shown in Figure 1, the project progresses through six distinct phases. Phases are generally sequential, but the framework doesn't rigidly determine the progression through them. According to the figure, arrows denote how the phases should flow and depend on each other, but the actual sequence may vary depending on a given project's outcomes and requirements. With this adaptability, the project can be fine-tuned and customized iteratively to meet its evolving needs. First, we identify a core problem, set clear objectives, and outline a strategy for addressing it. In the next phase, relevant data is collected, explored, and evaluated to determine its quality and highlight key features. The data is cleansed and prepared for analysis as part of the data preparation phase. In the modeling phase, methods are selected, data is divided into training and testing sets, the model is constructed, and it is fine-tuned to achieve peak performance. By comparing the model against predefined benchmarks, iterative evaluation facilitates the necessary adjustment. An ongoing maintenance strategy is developed during the deployment phase. After the project is completed, a comprehensive documentation phase captures a detailed account. CRISP-DM's structure and adaptability make it a powerful tool for guiding data mining projects in a variety of industries [32].

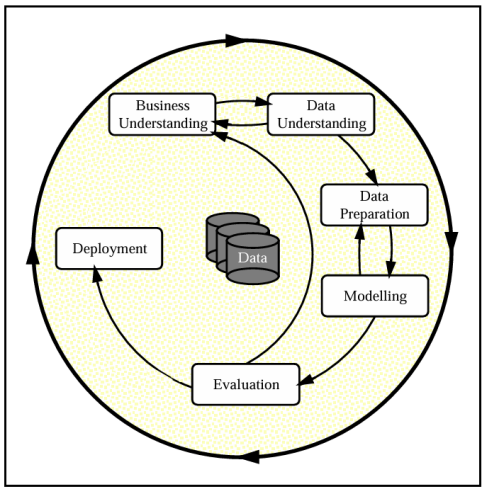


Figure 1. CRISP-DM phases.

Chapter 4- Data Analysis

The purpose of this chapter is to present a comprehensive analysis of the JNJ dataset central to this thesis. We use descriptive statistics to summarize stock price data, and then explore underlying patterns and relationships with exploratory data analysis. A variety of visualization techniques are employed to enhance the understandability of key financial indicators. Additionally, a thorough statistical analysis is conducted to validate the findings. An important component of the discussion focuses on feature importance analysis, identifying the major drivers behind PPO Signal predictions, which is imperative to creating better investment strategies.

4.1 Description of the Dataset

The dataset for this study was gathered from Yahoo Finance and specifically focuses on Johnson & Johnson (JNJ) stock performance. There are a total of six attributes included in each entry: Date, Open, High, Low, Close, Adjusted Close, and Volume. Except for the Date, all of these attributes represent continuous numerical data. In addition to Open, High, Low, and Close values, Adjusted Close also accounts for dividends and splits, giving a better indication of the stock's value. In terms of volume, this refers to how many shares are traded every day.

4.2 Exploratory Data Analysis

4.2.1 Data Profiling

The Data Profiling process involves computing descriptive statistics to numerically summarize the dataset, which includes the daily stock prices for Johnson & Johnson. There are 884 valid entries in the dataset, which covers the period from January 2, 2019, to November 9, 2023. Open, High, Low, Close, Adjusted Close and Volume are all continuous variables with measures of central tendency (mean) and dispersion (standard deviation). Moreover, the skewness statistic is calculated to assess the asymmetry of the distribution compared to the normal distribution. The dataset's profiling shows Volume with a strong right skew, indicating values below the mean, and Adjusted Close with the highest variability. The distribution patterns of data are visualized using histograms for each attribute. Figure 2 provides a detailed summary of key attributes.

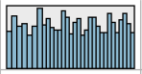
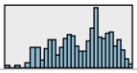
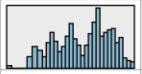
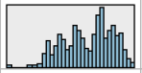
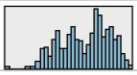
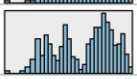
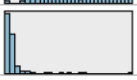
Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Date		Continuous	2019-01-02	2023-11-09	--	--	--	--	884
Open		Continuous	117.000	185.100	156.765	14.767	-0.280	--	884
High		Continuous	118.560	186.690	157.900	14.774	-0.258	--	884
Low		Continuous	109.160	184.180	155.531	14.747	-0.292	--	884
Close		Continuous	111.140	186.010	156.755	14.761	-0.275	--	884
AdjClose		Continuous	101.250	178.460	147.693	18.127	-0.331	--	884
Volume		Continuous	2114900.0...	151319500.000	8105470.814	7892444.375	10.697	--	884

Figure 2. Summary of key attributes.

4.2.2 Summary Statistics

There are 39 fields of continuous data types in the dataset, mostly related to stock market metrics. Over 97% of the fields have complete data, and there are very few missing values. Several key metrics, like EMA26 and MiddleBand20daySMA, display no outliers, indicating reliable measurements. However, outliers are present in some fields, such as the PPOSignalLine and MACD. In general, the dataset is well-maintained, making it a useful resource for financial analysis. In Table 1, we can see a detailed breakdown of the number of observations, the number of missing values and the types of data.

Field	Measurement	Outliers	% Complete	Valid Records	Null Value
PPOSignalLine	Continuous	22	97.297	1224	34
EMA26	Continuous	0	97.933	1232	26
MACD	Continuous	22	97.933	1232	26
PPO	Continuous	20	97.933	1232	26
MiddleBand20daySMA	Continuous	0	98.41	1238	20
SD20	Continuous	34	98.41	1238	20
UpperBand	Continuous	0	98.41	1238	20
LowerBand	Continuous	0	98.41	1238	20
BandWidth	Continuous	34	98.41	1238	20
TP20	Continuous	0	98.41	1238	20
MD20	Continuous	3	98.41	1238	20
CCI	Continuous	9	98.41	1238	20
EMA14_Gain	Continuous	21	98.808	1243	15
EMA14_Loss	Continuous	13	98.808	1243	15
RS	Continuous	21	98.808	1243	15
RSI	Continuous	0	98.808	1243	15
ATR14	Continuous	12	98.808	1243	15
ROC	Continuous	20	98.887	1244	14
EMA12	Continuous	0	99.046	1246	12
TR14	Continuous	27	99.762	1255	3
PosDM14	Continuous	15	99.762	1255	3
NegDM14	Continuous	21	99.762	1255	3
PosDI	Continuous	19	99.762	1255	3
NegDI	Continuous	10	99.762	1255	3
DX	Continuous	0	99.762	1255	3
Diff	Continuous	20	99.841	1256	2
Gain	Continuous	12	99.841	1256	2
Loss	Continuous	18	99.841	1256	2
TR	Continuous	25	99.841	1256	2
PosDM	Continuous	17	99.841	1256	2
NegDM	Continuous	21	99.841	1256	2
Date	Continuous	0	99.921	1257	1
Open	Continuous	0	99.921	1257	1
High	Continuous	0	99.921	1257	1
Low	Continuous	0	99.921	1257	1
Close	Continuous	0	99.921	1257	1
AdjClose	Continuous	0	99.921	1257	1
Volume	Continuous	4	99.921	1257	1
TP	Continuous	0	99.921	1257	1

Table 1. Summary outliers and feature completeness before cleaning.

4.2.3 Data Cleaning

During the preprocessing stage of data analysis, data cleaning is a crucial step that directly impacts the reliability and quality of the results. Within a dataset, this process involves correcting or removing incorrect, corrupted, duplicated, or incomplete data. To understand how missing data can be handled and outliers detected, it's important to understand their relevance in broader data-cleaning contexts. When missing data is not properly managed, it can introduce significant biases and lead to error, and outliers can significantly distort statistical analysis. Hence, data cleaning lays a solid foundation for subsequent analyses by ensuring the dataset's integrity.

4.2.3.1 Approaches for Handling Missing Data

When managing missing data within datasets, it's important to recognize the implications of missing entries and apply the appropriate methods to bridge these gaps, ensuring the analysis remains robust and valid. Deletion methods such as listwise and pairwise are common exclusion strategies. If missing values are not completely random in analyzed variables, listwise deletion can reduce sample size and introduce bias. In pairwise deletions, all available data is used to calculate pairwise statistics, retaining more data, but risking inconsistency due to different missing patterns across variables. Based on the following rule, observations missing critical variables such as EMA12, EMA26, MACD, and PPOSignalLine were excluded from the dataset:

@NULL(EMA12) or @NULL(EMA26) or @NULL(MACD) or @NULL(SignalLine)

As a result of this method, 100% of the variables were completed. Table 2 shows no null values in fields such as Volume, EMA26, and PPOSignal, among others, indicating the effectiveness of this approach. A comprehensive understanding of the nature of missing data is imperative, including whether they are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Seaman et al. [33] report that if misconceptions about the MAR concept are not addressed properly, inaccurate conclusions may be drawn. Therefore, it is essential to consider the nature of missing data and the appropriate statistical techniques when attempting to draw reliable statistical conclusions from incomplete data.

Field	Measurement	Outliers	% Complete	Valid Records	Null Value
Volume	Continuous	2	100	884	0
EMA26	Continuous	0	100	884	0
PPOSignal	Nominal	--	100	884	0
EMA14_Loss	Continuous	4	100	884	0
RSISignal	Nominal	--	100	884	0
TR	Continuous	17	100	884	0
ROCSignal	Nominal	--	100	884	0
DXsignals	Nominal	--	100	884	0
Factor1	Continuous	1	100	884	0
Factor2	Continuous	9	100	884	0
Factor3	Continuous	9	100	884	0
Factor4	Continuous	11	100	884	0
Factor5	Continuous	8	100	884	0
Factor6	Continuous	18	100	884	0
Factor7	Continuous	4	100	884	0
Partition	Nominal	--	100	884	0
XF_PPOSignal	Nominal	--	100	884	0
XFC_PPOSignal	Continuous	33	100	884	0

Table 2. Summary outliers and feature completeness after cleaning.

4.2.3.2 Outliers Detection

To detect outliers and anomalies within this dataset, we employed a dual approach. Statistical methods were used to identify outliers, by flagging all data points that fell outside three standard deviations of the mean, using the formula ($|\chi - \mu| > 3\sigma$). In this study, extreme outliers were those exceeding five standard deviations from the mean, as indicated by ($|\chi - \mu| > 5\sigma$). Additionally, machine learning algorithms such as Isolation Forests and Local Outlier Factors were utilized for anomaly detection, facilitating a deeper analysis of deviations from normal behavior. Figure 4 shows the effectiveness of this approach with data points classified into 'normal' and 'anomalous' categories. Among the 1224 observations, 99.02% are within the normal range, and 0.98% are anomalous, corresponding to 12 out of 1224.

Value	Proportion	%	Count
F		99.02	1212
T		0.98	12

Figure 3. Distribution of binary variable outcomes.

4.2.4 Statistical Analysis

Statistical analysis, one of the foundations of quantitative finance research, plays a key role in unraveling complex relationships between market variables. A series of Pearson correlation coefficients and Chi-Square tests are used to discern the strength and significance of the associations among various financial indicators. We examine in depth the linear correlations among price metrics and the intricate relationships among technical indicators based on categorical statistical tests.

Numerical Attributes:

Various financial variables in the dataset were correlated using Pearson correlation coefficients. As can be seen in Figure 5, the analysis indicates strong correlations between the Open, High, Low, Close, and Adj Close prices, as evidenced by their close correlation coefficients and their significance (2-tailed) less than 0.001. Due to their interdependence, these market indicators show a high degree of linear association. Contrary to this, these prices are not linearly related to volume (Volume) traded, indicated by correlation coefficients near zero and no statistical significance (p-values greater than 0.05). Consequently, the number of securities traded within a given timeframe is relatively independent of the price movement of those securities during that period. EMA12 and EMA26, two technical indicators, exhibit a significant and strong correlation with stock prices, which reflects their role in smoothing price data to identify trends. There is also a significant correlation between MACD, a momentum indicator, and PPO, a normalized and percentage-based indicator, and the price variables, but to a lesser degree, indicating these indicators reflect various market characteristics.

Correlations

		Open	High	Low	Close	Adj Close	Volume	EMA12	EMA26	MACD	PPO
Open	Pearson Correlation	1	.997**	.997**	.995**	.977**	.985**	.967**	.291**	.345**	.345**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	<.001	<.001	<.001
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
High	Pearson Correlation	.997**	1	.996**	.997**	.979**	.984**	.968**	.273**	.326**	.326**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	<.001	<.001	<.001
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
Low	Pearson Correlation	.997**	.996**	1	.998**	.979**	.982**	.962**	.304**	.358**	.358**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	<.001	<.001	<.001
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
Close	Pearson Correlation	.995**	.997**	.998**	1	.981**	.981**	.963**	.285**	.339**	.339**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	<.001	<.001	<.001
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
Adj Close	Pearson Correlation	.977**	.979**	.979**	.981**	1	.974**	.968**	.197**	.254**	.254**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	<.001	<.001	<.001
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
Volume	Pearson Correlation	.028	.039	.010	.019	.050	.037	.052	-.033	-.044	-.044
	Sig. (2-tailed)	.324	.169	.725	.498	.076	.198	.069	.243	.120	.120
	N	1257	1257	1257	1257	1257	1246	1232	1232	1232	1232
EMA12	Pearson Correlation	.985**	.984**	.982**	.981**	.974**	1	.992**	.214**	.273**	.273**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	<.001	<.001	<.001
	N	1246	1246	1246	1246	1246	1246	1232	1232	1232	1232
EMA26	Pearson Correlation	.967**	.968**	.962**	.963**	.968**	.992**	1	.086**	.147**	.147**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.002	<.001	<.001
	N	1232	1232	1232	1232	1232	1232	1232	1232	1232	1232
MACD	Pearson Correlation	.291**	.273**	.304**	.285**	.197**	.214**	.086**	1	.994**	.994**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	.002		.000	.000
	N	1232	1232	1232	1232	1232	1232	1232	1232	1232	1232
PPO	Pearson Correlation	.345**	.326**	.358**	.339**	.254**	.273**	.147**	.994**	1	1
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.000		
	N	1232	1232	1232	1232	1232	1232	1232	1232	1232	1232

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4. Correlation matrix of numerical attributes.

Categorical Attributes:

According to Figure 6, Chi-Square tests indicate that the categorical variables analyzed are statistically significant. The chi-square value for the RSI Signal versus the ROC Signal is 1657.888 with 6 degrees of freedom, and the asymptotic significance of this relationship is less than 0.001. For the RSI Signal against DX signals, the Chi-Square value was 181.949 with the same level of significance. In addition, PPO and ROC signals have a Chi-Square value of 776.100 with 4 degrees of freedom, further strengthening their significant association at 0.001 significance level. There is a similar relationship between the PPO Signal and DX Signal, with a Chi-Square value of 106.442 and a significance level of less than 0.001. All significance values are below 0.01, strongly supporting the null hypothesis that there is no association between the signals. These technical trading signals demonstrate a strong relationship, which could indicate underlying market

dynamics. Chi-Square results establish association, but they do not imply causation and further analysis would be needed to understand the nature of these relationships.

PPO Signal * ROC Signal

Crosstab

Count

		ROC Signal		
		Buy	Sell	Total
PPO Signal		13	3	9	25
	Buy	0	32	338	370
	Sell	0	355	507	862
Total		13	390	854	1257

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	776.100[a]	4	<.001
Likelihood Ratio	257.991	4	<.001
N of Valid Cases	1257		

a. 2 cells (22.2%) have expected count less than 5. The minimum expected count is .26.

PPO Signal * DX signals

Crosstab

Count

		DX signals		
		non-trend	trending	Total
PPO Signal		2	12	11	25
	Buy	0	162	208	370
	Sell	0	313	549	862
Total		2	487	768	1257

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	106.442[a]	4	<.001
Likelihood Ratio	23.600	4	<.001
N of Valid Cases	1257		

a. 3 cells (33.3%) have expected count less than 5. The minimum expected count is .04.

RSI Signal * ROC Signal

Crosstab

Count

		ROC Signal		
		Buy	Sell	Total
RSI Signal		13	0	1	14
	Buy	0	172	0	172
	Hold	0	218	645	863
	Sell	0	0	208	208
Total		13	390	854	1257

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1657.888[a]	6	.000
Likelihood Ratio	709.245	6	<.001
N of Valid Cases	1257		

a. 4 cells (33.3%) have expected count less than 5. The minimum expected count is .14.

RSI Signal * DX signals

Crosstab

Count

		DX signals		
		non-trend	trending	Total
RSI Signal		2	7	5	14
	Buy	0	67	105	172
	Hold	0	323	540	863
	Sell	0	90	118	208
Total		2	487	768	1257

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	181.949[a]	6	<.001
Likelihood Ratio	22.568	6	<.001
N of Valid Cases	1257		

a. 4 cells (33.3%) have expected count less than 5. The minimum expected count is .02.

Figure 5. Chi-Square tests on categorical attributes.

4.2.5 Dimensionality Reduction

In data analysis, dimension reduction involves reducing the number of variables to the most important ones by reducing the number of variables. One of the most important techniques in this process is Principal Component Analysis (PCA). Data is simplified by identifying the primary determinants of variability and arranging them according to their significance. A large number of

correlated variables are transformed into a smaller, uncorrelated set of variables using this method [34].

Initially, the first component captures the most variance, with each subsequent component designed to capture the next most significant variance. As a result of this hierarchy, we can focus on the most informative aspects of the data, similar to highlighting the most important features of a complex image, thereby simplifying the overall complexity without sacrificing valuable information.

Principal Component Analysis (PCA) was used to distill financial data into primary factors that captured significant variance. In Table 3, each equation represents a weighted combination of the original variables, weighted by their contribution to the component.

Using each factor, the dataset's dimensionality is reduced while retaining as much information as possible. The purpose of these factors is to identify underlying patterns in the data, reduce noise, and improve the performance of subsequent analytical models based on the data.

Equation For Factor-1	Equation For Factor-2	Equation For Factor-3	Equation For Factor-4
0.00000001949 * Volume + 0.007963 * EMA26 + 0.7381 * EMA14_Loss + 0.2155 * TR + 0.0574 * Factor1 + 0.3538 * Factor2 + -0.0423 * Factor3 + 0.0866 * Factor4 + -0.04938 * Factor5 + 0.09499 * Factor6 + -0.0472 * Factor7 + 0.7129 * XFC_PPOSignal + + -2.987	-0.00000000517 * Volume + 0.03387 * EMA26 + -0.2877 * EMA14_Loss + -0.05752 * TR + 0.5099 * Factor1 + 0.007891 * Factor2 + 0.02796 * Factor3 + -0.1067 * Factor4 + 0.01698 * Factor5 + -0.03464 * Factor6 + 0.1112 * Factor7 + 0.1704 * XFC_PPOSignal + + -5.137	0.00000003842 * Volume + -0.001215 * EMA26 + -0.7159 * EMA14_Loss + 0.1112 * TR + 0.108 * Factor1 + -0.1664 * Factor2 + 0.4297 * Factor3 + 0.3888 * Factor4 + -0.2186 * Factor5 + 0.1295 * Factor6 + -0.06108 * Factor7 + -1.597 * XFC_PPOSignal + + 1.278	-0.00000002366 * Volume + -0.003685 * EMA26 + -0.2989 * EMA14_Loss + 0.1868 * TR + -0.09774 * Factor1 + -0.01483 * Factor2 + 0.1791 * Factor3 + -0.3185 * Factor4 + -0.04284 * Factor5 + 0.5542 * Factor6 + 0.575 * Factor7 + 0.5041 * XFC_PPOSignal + + 0.17
Equation For Factor-5	**Equation For Factor-6**	**Equation For Factor-7**	
-0.000000002087 * Volume + 0.00004036 * EMA26 + 0.1447 * EMA14_Loss + 0.002383 * TR + 0.003939 * Factor1 + 0.09573 * Factor2 + -0.02185 * Factor3 + 0.472 * Factor4 + 0.681 * Factor5 + -0.04778 * Factor6 + 0.48 * Factor7 + -2.1 * XFC_PPOSignal + + 1.807	-0.00000002077 * Volume + -0.005537 * EMA26 + -0.04369 * EMA14_Loss + 0.05667 * TR + -0.06345 * Factor1 + 0.1749 * Factor2 + 0.4997 * Factor3 + 0.1167 * Factor4 + -0.09706 * Factor5 + -0.5276 * Factor6 + 0.3264 * Factor7 + 4.09 * XFC_PPOSignal + + -2.871	0.00000002634 * Volume + -0.001235 * EMA26 + -0.3166 * EMA14_Loss + -0.09344 * TR + 0.002306 * Factor1 + -0.006511 * Factor2 + 0.3128 * Factor3 + -0.164 * Factor4 + 0.6192 * Factor5 + 0.2959 * Factor6 + -0.4696 * Factor7 + 3.928 * XFC_PPOSignal + + -3.206	

Table 3. Equations for the generated factors.

A total of 38 factors were initially derived in the Principal Component Analysis. For subsequent analyses, I selected the first seven factors based on an optimal representation of the dataset while preserving analytical precision, as shown in Figure 7. Based on the principal component analysis (PCA) shown, the first seven factors explain approximately 82% of the total variance in the data, indicating that these factors contain most of the information needed. Therefore, 7 factors provide a distilled but comprehensive overview of the dataset's underlying structure, capturing a significant portion of its inherent variability.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	12.056	31.728	31.728	12.056	31.728	31.728
2	6.863	18.061	49.788	6.863	18.061	49.788
3	5.100	13.420	63.208	5.100	13.420	63.208
4	2.282	6.005	69.213	2.282	6.005	69.213
5	2.175	5.724	74.937	2.175	5.724	74.937
6	1.677	4.414	79.351	1.677	4.414	79.351
7	1.125	2.960	82.311	1.125	2.960	82.311
8	1.099	2.893	85.204			
9	1.002	2.637	87.841			
10	.926	2.438	90.279			
11	.794	2.090	92.369			
12	.757	1.993	94.362			
13	.538	1.416	95.778			
14	.376	.990	96.768			
15	.341	.898	97.666			
16	.269	.709	98.375			
17	.189	.498	98.872			
18	.132	.347	99.220			
19	.071	.188	99.408			
20	.071	.187	99.594			
21	.053	.139	99.733			
22	.046	.121	99.854			
23	.022	.058	99.912			
24	.019	.050	99.962			
25	.006	.015	99.977			
26	.004	.010	99.987			
27	.002	.006	99.993			
28	.001	.004	99.997			
29	.001	.002	99.999			
30	.000	.001	100.000			
31	.000	.000	100.000			
32	6.559E-5	.000	100.000			
33	3.004E-15	7.904E-15	100.000			
34	-5.768E-16	-1.518E-15	100.000			
35	-7.174E-16	-1.888E-15	100.000			
36	-1.349E-15	-3.550E-15	100.000			
37	-2.591E-15	-6.819E-15	100.000			
38	-4.324E-15	-1.138E-14	100.000			

Figure 6. Total variance explained for factors.

4.2.6 Feature Engineering

In feature engineering, raw data is transformed into meaningful features that represent underlying patterns relevant to predictive modeling. In addition to improving model accuracy and performance, this process introduced new variables within the dataset that encapsulate complex interactions.

A comprehensive transformation process was employed to generate 37 distinct attributes from an original set of 7 fundamental attributes, including date, open, close, high, low, adjusted close, and volume. These new attributes capture unique aspects of market dynamics and trends.

1- **Exponential Moving Averages (EMAs):** EMA12 and EMA26 were calculated to give more weight to recent prices. The formula for this is (1):

$$EMA=Price(t)\times k+EMA(y)\times(1-k) \quad (1)$$

2- **Moving Average Convergence Divergence (MACD) and Percentage Price Oscillator (PPO):** These two indicators use EMAs to detect changes in trend momentum and provide a normalized perspective, respectively. The formula for these are (2) (3):

$$MACD = EMA(12) - EMA(26) \quad (2)$$

$$PPO = (MACD / EMA(26)) \times 100 \quad (3)$$

3- **Daily Changes (Diff), and Exponential Moving Averages of Gains and Losses (EMA14_Gain, EMA14_Loss):** These calculate the daily price changes and smooth daily gains and losses over 14 days (4). Gains and losses are derived from Diff, where gains are positive diffs and losses are negative (5) (6).

$$Diff = Today's\ closing\ price - yesterday's\ closing\ price \quad (4)$$

$$EMA14(Gain) = Average\ gain\ over\ 14\ days \quad (5)$$

$$EMA14(Loss) = Average\ loss\ over\ 14\ day \quad (6)$$

4- **Relative Strength Index (RSI), True Range (TR), and Average True Range (ATR14):** Indicators such as these give insight into how strong and volatile the market is. The formula for these are (7) (8) (9):

$$RSI = 100 - (100 / (1 + \text{Average Gain} / \text{Average Loss}))$$

(7)

$$TR = \text{Max}(\text{high} - \text{low}, |\text{high} - \text{previous close}|, |\text{low} - \text{previous close}|)$$

(8)

$$ATR14 = \text{The sum of TR for 14 days divided by 14}$$

(9)

5- **Bollinger Bands (SMA14, UpperBand, LowerBand) and Rate of Change (ROC):** These derive from simple moving averages and price change ratios to frame volatility and momentum. The formula for these are (10) (11) (12) (13):

$$SMA14 = \text{The average closing price over 14 days}$$

(10)

$$\text{Upper Bollinger Band} = SMA14 + (2 \times SD14)$$

(11)

$$\text{Lower Bollinger Band} = SMA14 - (2 \times SD14)$$

(12)

$$ROC = ((Price(t) - Price(t-14)) / Price(t-14)) \times 100$$

(13)

6- **Directional Movement Indicators and Directional Movement Index (DX):** These indicators measure the direction and strength of market movements. The formula for these are:

$$\text{Positive Directional Movement} = \text{Today's high} - \text{yesterday's high}$$

(14)

Negative Directional Movement = Yesterday's low - today's low

(15)

TR14 = 14-period smoothed True Range

(16)

PosDM14 = 14-period smoothed Positive Directional Movement

(17)

NegDM14 = 14-period smoothed Negative Directional Movement

(18)

Positive Directional Indicator = (Smoothed PosDM / Smoothed TR) x 100

(19)

Negative Directional Indicator: (Smoothed NegDM / Smoothed TR) x 100

(20)

Directional Movement Index = |PosDI - NegDI) / (PosDI + NegDI)| x 100

(21)

4.2.7 Feature Importance Analysis

In machine learning models, feature importance analysis explains the relative contribution of each variable to the predictive accuracy. Random forest algorithms, widely known for their robustness and utility in both classification and regression tasks, were used to analyze the significance of each feature in the dataset. Using this ensemble method, which constructs multiple decision trees and aggregates their output, we can evaluate whether certain features have a positive influence on the prediction accuracy. According to Figure 8, the analysis yielded a hierarchy of features based on their importance, with EMA14_Loss, RSISignal, and TR being the most influential. Feature prioritization indicates the relative influence of each feature on the decision-making process of the model. In addition to the top-ranked variables, Volume, EMA26, and MACD also have notable impacts on the model's performance. This study incorporates feature importance analysis to identify the most informative predictors.

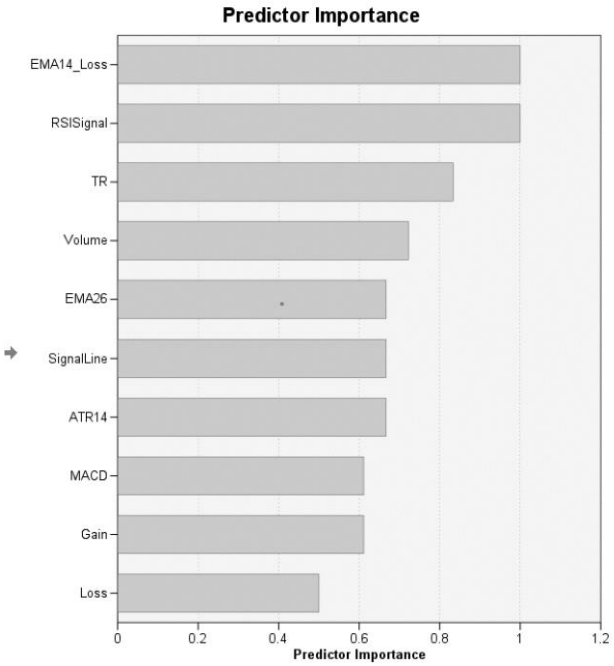


Figure 7. Predictor importance.

4.2.8 Visualization of Key Features

Visualizing data is a critical step to understanding the impact of different variables on our analysis. As part of this thesis, we present a series of visualizations that are based on features that are highly correlated with the PPO signal, our target variable. This visual illustrates how the key predictors interact and influence the PPO signal, using the key predictors identified earlier. The use of graphical representations is crucial for understanding complex datasets, as they transform abstract numerical relationships into tangible forms, enhancing the analysis's comprehension and insight.

1- PPO Signal x EMA14_Loss

Based on the analysis conducted in this study, EMA14_Loss is statistically significant in predicting PPO signals. The null hypothesis that EMA14_Loss distribution would be identical across different PPO signals was confidently rejected by the Mann-Whitney U Test, indicating a distinct distribution pattern between Buy and Sell signals. In financial market analysis, EMA14_Loss shows a significant difference between Buy and Sell signals based on the PPO signal, highlighting its potential to enhance prediction accuracy.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of EMA14_Loss is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 8. Hypothesis test between PPO Signal and EMA14_Loss.

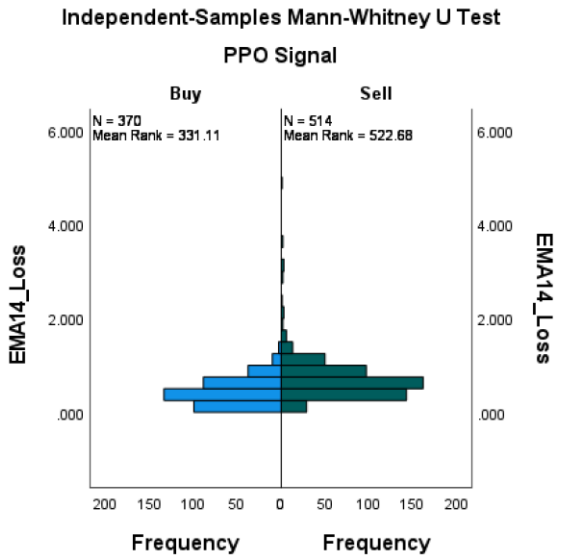


Figure 9. Mann-Whitney U test between PPO Signal and EMA14_Loss.

2- PPO Signal x RSI Signal

According to the bar chart, RSI and PPO signals differ across the Buy, Hold, and Sell categories. The PPO signal distribution for the RSI Buy and Sell categories is more balanced, showing a heavy concentration of Buy signals in the Hold category. The combination of RSI and PPO signals has the potential to enhance financial forecasting models.

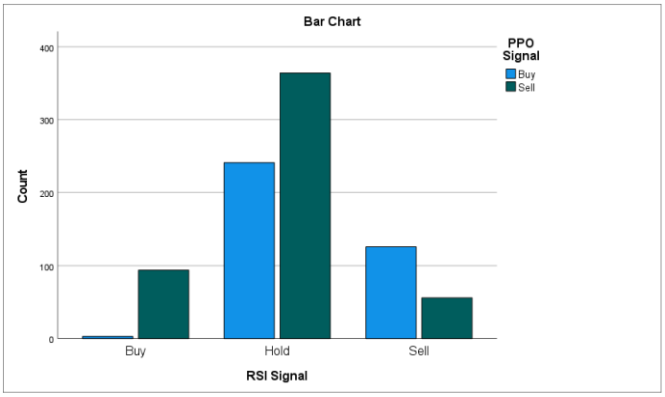


Figure 10. Relationship between PPO Signal and RSI Signal.

3- PPO Signal x Signal Line

Signal Line distributions differ statistically significantly between the Buy and Sell PPO Signal categories, as indicated by the rejection of the null hypothesis. Thus, a predictive model can more accurately forecast PPO signals by leveraging the Signal Line's different characteristics for Buy and Sell signals.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Signal Line is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 11. Hypothesis test between PPO Signal and Signal Line.

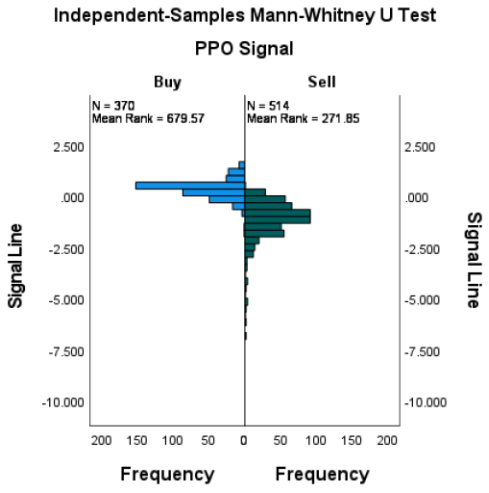


Figure 12. Mann-Whitney U test between PPO Signal and Signal Line.

4- PPO Signal x ATR14

In both the hypothesis test summary and histogram, the Average True Range over 14 days (ATR14) is significantly different across the PPO signal categories. There is a significant difference between the Mann-Whitney U Test and the conventional significance threshold of 0.05, which results in the rejection of the null hypothesis. This indicates a significant difference between ATR14 values for Buy and Sell signals. The ATR14 is therefore likely to be predictive of PPO signals and could be used as a valuable indicator in financial market models.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of ATR14 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 13. Hypothesis test between PPO Signal and ATR14.

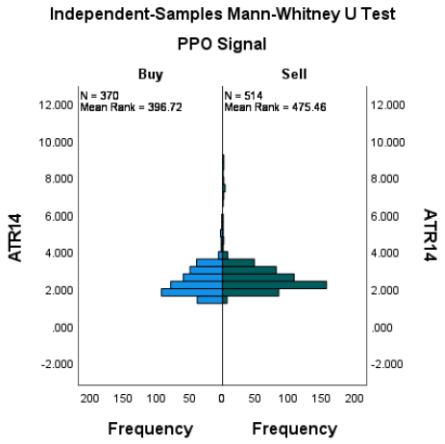


Figure 14. Mann-Whitney U test between PPO Signal and ATR14.

5- PPO Signal x MACD

There is a clear statistical variance in the distributions of Moving Average Convergence Divergence (MACD) across the different categories of the PPO signal based on the hypothesis testing summary. As a result of the Independent-Samples Mann-Whitney U Test, the null hypothesis that MACD distributions are the same across categories of PPO Signal is rejected. According to the PPO, this rejection indicates significant differences between Buy and Sell signals in the MACD values. Therefore, MACD is likely to be an effective predictor of the PPO Signal, which indicates its efficacy in financial market prediction.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of MACD is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 15. Hypothesis test between PPO Signal and Signal Line.

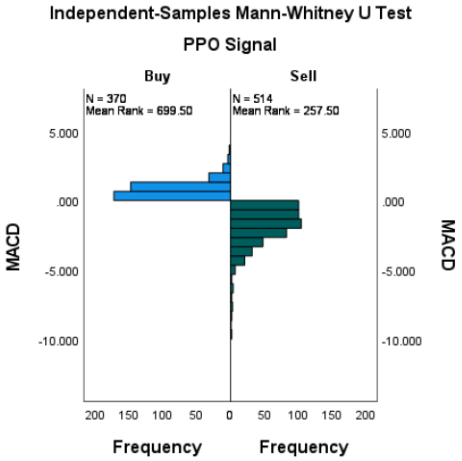


Figure 16. Mann-Whitney U test between PPO Signal and Signal Line.

6- Bollinger Bands for JNJ Stock

An illustration of Johnson & Johnson's (JNJ) stock performance over a given period is shown in Figure 18. A steady decline in the stock price has been observed since mid-October, and it has breached the lower Bollinger Band. While the closing price hovers near the lower band, it does not indicate a severe downturn since it does not fall dramatically below the band. In this period, the 20-day SMA trended downward, reflecting the bearish momentum. Towards the end of the chart, the Bollinger Bands align and volatility appears to decrease, resulting in a tightening trading range. Even with the downward trend, the stock doesn't show a significant recovery by closing above the upper band, which indicates a lack of bullish strength at this point.

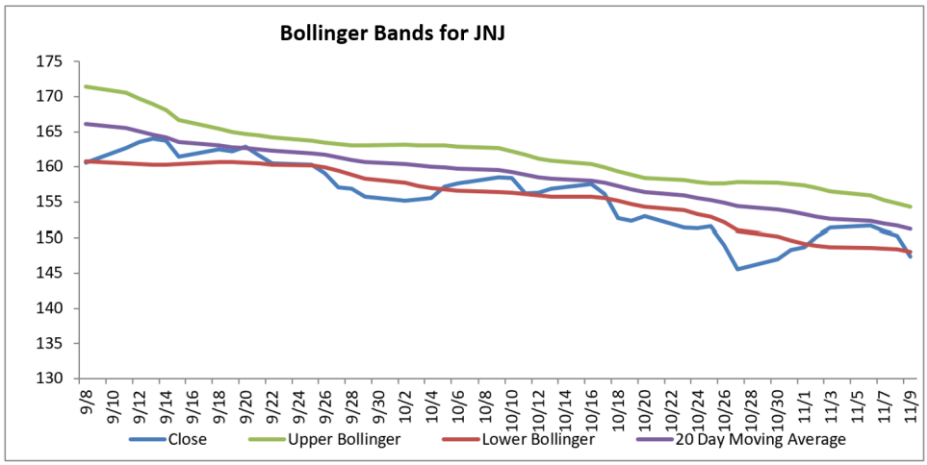


Figure 17. Bollinger bands chart for JNJ stock.

7- Commodity Channel Index for JNJ Stock

In Figure 19, Johnson & Johnson's closing stock prices and Commodity Channel Index (CCI) are shown. With a declining trend, the black line indicates a bearish sentiment for the stock. In the red line, oscillating around zero, overbought and oversold conditions are identified, signaling potential price corrections or rebounds. For technical trading decisions, the CCI's volatility correlates to price momentum shifts with peaks and troughs. In the subsequent period, the CCI indicates oversold conditions, potentially indicating opportunities for price recovery.

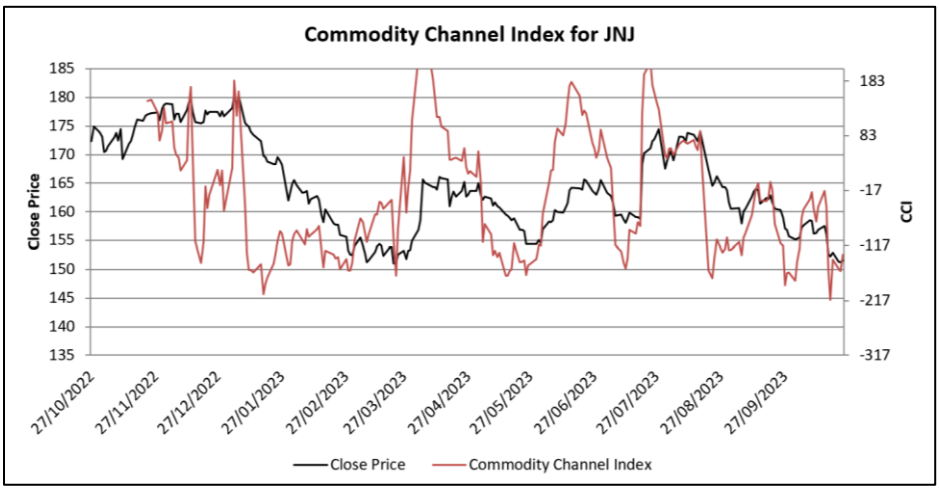


Figure 18. Commodity channel index chart for JNJ stock.

8- MACD for JNJ Stock

As shown in Figure 20, Johnson & Johnson's stock trends are shown through the MACD indicator. Histograms show momentum oscillations and market sentiment oscillations based on MACD and Signal lines. MACD is primarily under the Signal line, which suggests a bearish outlook. The 12 and 26-day EMAs in the right pane help signal potential trend changes. A downward EMA trend reveals a story of the stock's sentiment and direction throughout, with technical indicators telling a narrative of the market's bearishness.

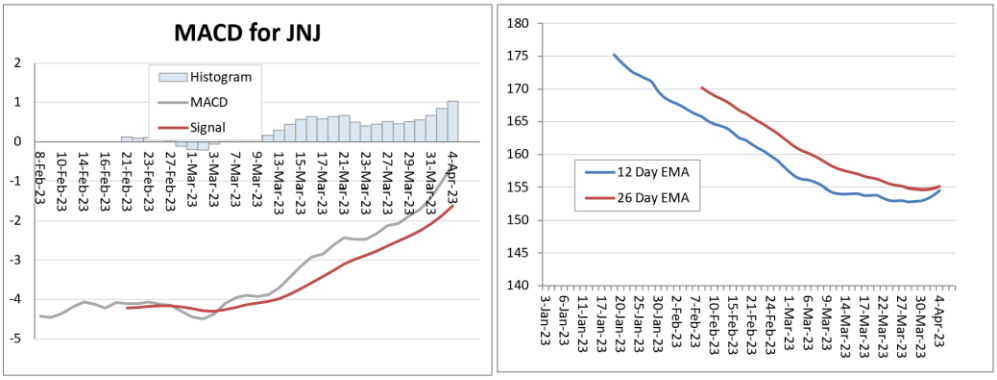


Figure 19. Moving average convergence/divergence chart for JNJ stock.

9- RSI for JNJ

The chart below showcases Johnson & Johnson's stock performance with the Relative Strength Index (RSI) and closing prices over three months from January 31, 2023. The RSI, indicating momentum by fluctuating between 0 to 100, suggests overbought conditions above 70 and oversold below 30. It peaks above 50, indicating strength, but later dips, implying potential fatigue. Correspondingly, the closing price starts strong but declines, signaling reduced investor enthusiasm. Together, the RSI and price movement narrate a cooling off of investor sentiment and a stock easing from its highs.

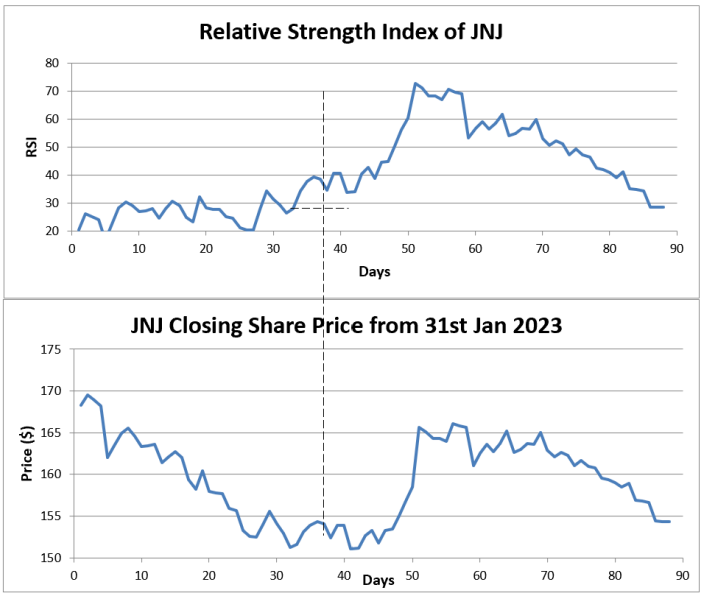


Figure 20. Relative strength index chart for JNJ stock.

4.3 Machine Learning Model Development

In this chapter, we explore the complex process of constructing and refining machine learning models to predict PPO signal fluctuations in financial markets. A comprehensive analysis of algorithms, models, and parameters is provided. In addition, the narrative describes how to train the models, tune hyperparameters, and validate and test them. As part of its evaluation metrics, the chapter elaborates on how the models are evaluated, ensuring a reliable assessment.

4.3.1 Chosen Input

To accurately forecast the PPO signal, which serves as the target attribute, a variety of input attributes were used to develop the predictive models. As illustrated in Figure 22, these inputs include Volume, EMA26, EMA14_Loss, RSI Signal, TR (True Range), ROC Signal, DX Signals, and seven distinct factors (Factor 1 through Factor 7). Utilizing these inputs, the models capture market trends and fluctuations more precisely, enabling better prediction of the PPO

signal.

Field	Measurement	Values	Missing	Check	Role
Volume	Continuous	[2114900.0,1.513195E8]		None	Input
EMA26	Continuous	[129.9260171,180.0470469]		None	Input
EMA14_Loss	Continuous	[0.07080062,4.880275072]		None	Input
RSISignal	Nominal	"",Buy,Hold,Sell		None	Input
TR	Continuous	[0.75,17.64]		None	Input
ROCSignal	Nominal	"",Buy,Sell		None	Input
DXsignals	Nominal	"",non-trend,trending		None	Input
$F-Factor-1	Continuous	[-3.1555926992281202,1.831...		None	Input
$F-Factor-2	Continuous	[-1.7868735340649349,7.739...		None	Input
$F-Factor-3	Continuous	[-2.149404772756305,4.3205...		None	Input
$F-Factor-4	Continuous	[-3.3540790292730076,5.865...		None	Input
$F-Factor-5	Continuous	[-4.047374070350134,3.9797...		None	Input
$F-Factor-6	Continuous	[-4.900054787224166,6.9606...		None	Input
$F-Factor-7	Continuous	[-3.352571148925187,7.7464...		None	Input
PPOSignal	Flag	Sell/Buy		None	Target

Figure 21. The input attributes and the target.

4.3.2 Machine Learning Algorithms

This section examines the rationale and intricacies of the machine learning algorithms used in this study, with a focus on their application to market prediction.

1- Logistic regression is primarily used to solve binary classification problems. A major advantage of this approach is its simplicity and interpretability, providing a clear understanding of how independent variables relate to the binary outcome. Moreover, it is particularly suitable for initial modeling efforts due to its computational efficiency and minimal parameter tuning requirement.

2- Linear Support Vector Machines (LSVMs) are employed because they create an optimal hyperplane between different classes based on linearly separable data. Among its strengths are its ability to model complex domains with a clear decision boundary and its strength in high-dimensional spaces.

3- The eXtreme Gradient Boosting algorithm, or XGBoost, stands out for its speed and performance. XGBoost improves model accuracy by learning from previous errors using an ensemble technique that constructs a series of decision trees. In this algorithm, multiple types of data can be managed, overfitting is prevented, and feature interactions can be handled automatically, making it an ideal tool for predicting data.

4- Neural networks are selected for their capability to model and learn complex, non-linear interactions in data. A neural network's layers enable it to discern intricate patterns and variable interactions, enabling it to adjust to a variety of distributions

of data. These models are computationally demanding and require extensive training data, but they are capable of achieving significant predictive performance.

5- Decision Trees are employed because they are interpretable and non-parametric. They break down data into smaller subsets and develop a decision tree incrementally using these subsets. Therefore, the decision-making process can be visualized and interpreted, and therefore, key factors influencing stock movements can be understood.

6- The Random Forest model was chosen due to its exceptional ability to handle both regression and classification tasks, making it an excellent choice for stock price prediction. During training, it constructs a multitude of decision trees, outputting the average prediction of the trees, which is more accurate and robust to overfitting. The algorithm's benefits, particularly its ability to provide insight into feature importance and its high predictive accuracy, justify its selection despite its complexity and computational intensity.

4.3.3 Validation and Testing Procedures

In this section, we outline the validation and testing protocols used for evaluating the trained machine-learning models. In addition to cross-validation, holdout validation, and bootstrap validation, it also examines various validation techniques that are integral to determining model robustness on data that was not encountered during the initial training. Furthermore, this segment explains how the dataset is divided into distinct subsets for training, validation, and testing. Partitioning strategies are discussed, emphasizing their significance in reducing overfitting and ensuring generalizability.

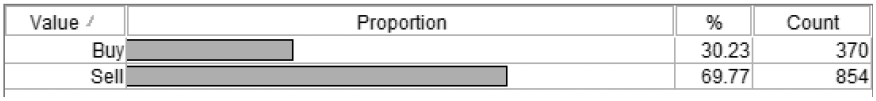
Data partitioning:

Preparing the dataset for the application of machine learning algorithms requires data partitioning. Using a train-test split approach, 70% of the dataset was allocated for training and 30% for testing in this study. Models can be trained and evaluated on the ability to generalize their findings to an unknown subset of data by learning and identifying patterns within a majority subset of the data. As parameters and weights are adjusted based on underlying stock market trends, the training set facilitates the fitting of machine learning models. In contrast, the testing set simulates real data and evaluates models' performance and predictive capabilities. Through

this approach, the models' accuracy and ability to predict stock prices are both thoroughly assessed and indicative of their potential performance in actual trading environments.

Data balancing:

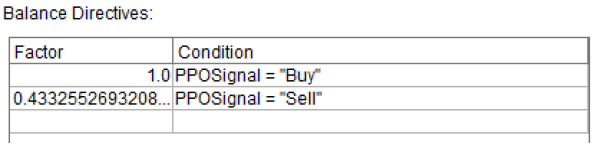
Figure 23 shows a distribution of 'Sell' and 'Buy' signals, where 'Sell' signals occur more frequently than 'Buy' signals. The number of 'Sell' signals totals 854 occurrences, while the number of 'Buy' signals is 30.23%, with 370 occurrences. Based on the dataset examined, there appears to be a higher propensity for selling recommendations.



Value	Proportion	%	Count
Buy		30.23	370
Sell		69.77	854

Figure 22. Buy vs. sell counts pre-undersampling.

Balance directives were applied to address the class imbalance in the dataset to predict the PPO signal, as shown in Figure 24. 'Buy' signals are fully retained with 1.0, which means their count has not changed. Compared to this, the 'Sell' signals are undersampled by approximately 0.433, meaning only 43.3% of these instances are included. This adjustment aims to equalize the representation of both classes, reducing model bias towards the previously dominant 'Sell' class, improving the model's overall predictive capability, and eliminating biases.

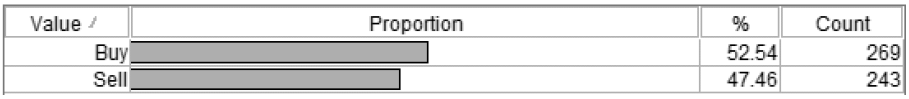


Balance Directives:

Factor	Condition
1.0	PPOSignal = "Buy"
0.4332552693208...	PPOSignal = "Sell"

Figure 23. Balance Directives.

With the balance directive applied to the dataset, the distribution between 'Buy' and 'Sell' signals is nearly equal. In proportion, 'Buy' signals comprise approximately 52.54% with a count of 269, while 'Sell' signals constitute approximately 47.46% with a count of 243. By making this adjustment, we have minimized the class imbalance, ensuring that the predictive model can recognize both classes without favoring one over the other.



Value	Proportion	%	Count
Buy		52.54	269
Sell		47.46	243

Figure 24. Buy vs. sell counts post-undersampling.

Evaluation metrics used to assess model performance:

Evaluation metrics are essential tools for assessing the performance of machine learning models. By measuring these metrics, it is possible to figure out how well a model performs when it comes to accuracy, precision, recall, and other important factors. When calculating the evaluation metrics, it is important to understand the confusion matrix. The confusion matrix helps visualize true positives (TPs), false positives (FPs), and true negatives (TNs), characterized as:

True Positives (TP): The number of correct predictions of a positive outcome.

True Negatives (TN): The number of correct predictions of a negative outcome.

False Positives (FP): The number of instances predicted as positive, but negative.

False Negatives (FN): The number of instances predicted as negative, but positive.

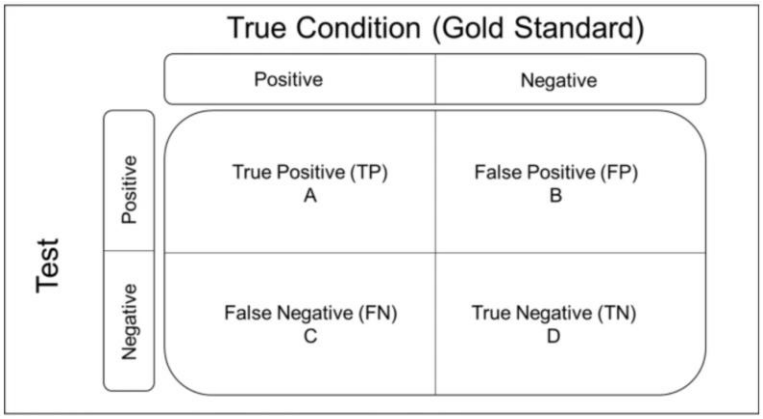


Figure 25. Confusion Matrix.

Accuracy: Measures how well the model predicts instances correctly, both positively and negatively. The formula for accuracy is:

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total number of Observations} \quad (22)$$

Precision: Calculates how many predicted positive instances are positive, assessing the accuracy of positive predictions. The formula for precision is:

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$$

(23)

Recall: It measures how well the model finds all the relevant cases (i.e., True Positives) within the data. Having this metric is essential when models are affected by missing a positive instance. The formula for the recall is:

Recall = True Positives / (True Positives + False Negatives)

(24)

F1-score: It provides a balance between precision and recall by eliminating extreme values and providing a smooth mean. Datasets with imbalances are particularly useful for this. The formula of f1-score is:

F1 = 2 x ((Precision x Recall) / (Precision + Recall))

(25)

ROC (Receiver Operating Characteristic): This is a curve that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold changes. This plot compares the true positive rate (Recall) with the false positive rate (1 - Specificity).

ROC AUC (Area Under the Curve): This measures the ability of a classifier to distinguish between classes and is used to summarize the ROC curve. With a higher AUC, the model is more likely to correctly predict 0s and 1s. The formula of AUC is:

$$AUC = \phi(\frac{\mu 1 - \mu 0}{\sqrt{\sigma 1^2 + \sigma 0^2}})$$

(26)

4.3.4 Results

In this chapter, we evaluate the performance of various machine learning models in predicting the PPO Signal. We provide performance metrics, model comparisons, and interpretations to evaluate and understand the predictive power of each model. Additionally, AUC values and ROC curves are examined to determine the accuracy of the models in discriminating.

4.3.4.1 Presentation of the Experimental Results

In Table 4, performance metrics for every predictive model used for forecasting PPO signals are summarized, with each model applying a threshold of 0.5 to classify predictions.

Model	No. Fields Used	Overall Accuracy (%)	Accumulated Accuracy (%)	Area Under Curve	Accumulated AUC	Precision	Recall	F1-Measure
Logistic Regression	14	96.467	96.467	0.995	0.995	0.977	0.974	0.98
LSVM	14	93.207	93.207	0.985	0.985	0.976	0.929	0.95
Neural Net	14	91.576	91.576	0.98	0.98	0.972	0.91	0.94
XGBoost Linear	14	89.402	89.402	0.978	0.978	0.987	0.865	0.92
XGBoost Tree	14	85.326	85.326	0.944	0.944	0.961	0.831	0.89
Random Forest	14	84.783	84.783	0.950	0.950	0.977	0.809	0.89

Table 4. Models result.

Key Observations:

- Logistic Regression shows 96.467% overall accuracy and an equivalent accumulated accuracy, indicating strong predictive capability. Both the initial and accumulated Area Under Curve metrics of the model are 0.995, indicating exceptional ability to differentiate classes. A high precision, recall, and F1 measure illustrate the model's balanced performance between precision and sensitivity, with values of 0.97, 0.97, and 0.98, respectively.
- While slightly less accurate and AUC than the Logistic Regression Model, the LSVM exhibits commendable performance. In total, it achieves a score of 93.207%, while the accumulated score is also similar. An F1-measure of 0.95 results from 0.985 AUC, indicating strong discriminative power, with precision and recall values of 0.976 and 0.929, respectively.
- Compared to the first two models, the Neural Net Model shows a slightly lower ability to classify correctly, with an overall accuracy of 91.576%. F1-measure is 0.94, with AUC at 0.98, precision, and recall at 0.972 and 0.91 respectively.
- As compared to the previously mentioned models, the performance metrics of the XGBoost Linear Model and XGBoost Tree Model are declining. The linear variant provides an overall accuracy of 89.402% and an AUC of 0.978; precision and recall

provide an F1-measure of 0.92. By comparison, the tree variant's accuracy is 85.326% and the AUC is 0.944, which results in an F1-measure of 0.89.

- Random Forest has the lowest overall accuracy of 84.783%. Despite this, the AUC is relatively high at 0.950, indicating reasonable classification effectiveness. It is worth mentioning that, despite the precision and recall of the model, the F1-measure is equal to 0.89 for the XGBoost Tree Model.

Comparative Analysis:

This study demonstrates that Logistic Regression is superior to other classification models in terms of accuracy, precision, recall, and F1-measure, demonstrating its ability to accurately classify instances and maintain a high rate of true positives. The Logistic Regression model, which combines the best accuracy and AUC values of both LSVM and Neural Net models, stands out among this study's models because it balances all assessment metrics. The findings highlight how different algorithms perform across a range of metrics, making them valuable for businesses and analysts evaluating predictive models. By considering these insights and the specific needs of the application at hand, the most suitable predictive model should be selected.

4.3.4.2 Evaluation of Predictor Importance

In a sensitivity analysis approach, the significance of the predictors is determined by how much the variance of the target variable decreases as a result of each predictor. Each input feature is quantified, providing insights into its relative importance within the model based on its impact on the output. In the evaluation of predictor importance, we use the following notation:

Y	Target
X_j	Predictor, where j=1,...,k
k	The number of predictors
$Y = f(X_1, X_2, \ldots, X_k)$	Model for Y based on predictors X_1 through X_k

The formula below (27) quantifies the extent to which a predictor Xi contributes to the variance of a target variable Y by calculating its importance. This formula calculates how much variance in Y can be attributed to changes in predictor Xi. An important metric in sensitivity analyses is the degree to which a variable contributes significantly to the model's output. A better

understanding of each predictor's contribution can lead to better decisions about model refinement and resource allocation.

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i))}{V(Y)}$$

(27)

An illustration of the formula below (28) shows how to calculate the normalized importance of a predictor Xi in a statistical model. Through this normalization process, each predictor's sensitivity index is scaled by the total sensitivity index of all predictors. Based on this result, Vi, the predictor Xi has a proportionate importance relative to the combined importance of all predictors. As a result, it can be compared across all predictors in the model to see which one has more or less influence on the variance of Y.

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j}$$

(28)

To decompose the total variance of the model output, we used variance-based sensitivity indices, calculated through (29):

$$\hat{V}(Y) = \frac{1}{N-1}\sum_{r=1}^{N} f^2\left(x_1^{(r)}, x_2^{(r)}, \ldots, x_k^{(r)}\right) - \hat{E}^2(Y)$$

(29)

The outcome of this process was influenced by significant input factors. The variance attributable to each input was quantified using (30):

$$\hat{V}(E(Y|X_j)) = \hat{U}_j - \hat{E}^2(Y)$$

(30)

where Uj denotes the mean output with the j-th input held constant. This first-order sensitivity measure is reliable when the input factors X1, X2, and Xk are orthogonal/independent and the model does not incorporate interactions between inputs. Even in the presence of interactions and non-orthogonality, Si serves as an effective measure to rank inputs in order of importance, although a large dataset is required to reduce bias or bootstrap methods can be employed to enhance the accuracy of Si estimations [35] [36]. Stability and convergence of these sensitivity indices were monitored through (31):

$$\bigcap_{i \in I} \frac{1}{D} \sum_{j=t-D+1}^{t} \frac{\left|S_i(j) - \overline{S}_i\right|}{\overline{S}_i} < \epsilon \tag{31}$$

4.3.4.3 Predictor Importance of the Best Model

This thesis examines the significance of the predictor importance in the optimal model for forecasting the PPO signal in financial markets. The predictive outcomes are significantly influenced by indicators such as market volatility and moving averages (such as EMAs and SMAs). A high level of volatility usually corresponds to a stronger PPO signal, indicating more trading opportunities. In addition to signaling the strength of a trend, momentum indicators like the MACD and RSI also alert us to its potential reversals, making them vital tools for traders. Additionally, the interaction between these predictors can provide a deeper understanding of market dynamics. PPO signals are affected by combinations of indicators such as Bollinger Bands and Average True Range, which provide insight into market conditions resulting in high trading activity. The results of such analyses can be used by traders and financial analysts to refine their strategies, focusing on signals that have a high likelihood of bringing profits. Through systematic analysis of these predictors, this study intends to improve trading system predictive models. Besides improving signal forecasts, this also facilitates the strategic planning of trades, optimizes resource allocation, and minimizes risk associated with volatile market conditions. Based on the results of the study, it is important to take a holistic approach when developing trading models, where key predictors are integrated based on their proven influence and interrelationships, leading to more robust and reliable strategies.

4.3.4.5 Visualization of Predictor Importance

Based on the sensitivity analysis, Factor 4 emerged as the most significant input in the model prediction. An LSVM analysis using multiple variables to calculate Factor 4 weights corroborates this finding. Specifically, the equation for Factor 4 includes a combination of the following variables:

−0.00000002366×Volume−0.003685×EMA26−0.2989×EMA14_Loss+0.1868×TR−0.09774×Factor1−0.01483×Factor2+0.1791×Factor3−0.3185×Factor4−0.04284×Factor5+0.5542×Factor6+0.575×Factor7+0.5041×XFC_PPOSignal+0.17

Figure 23 illustrates Factor 4's importance in influencing the model's output since it interacts with various market indicators and internal parameters in a complex manner.

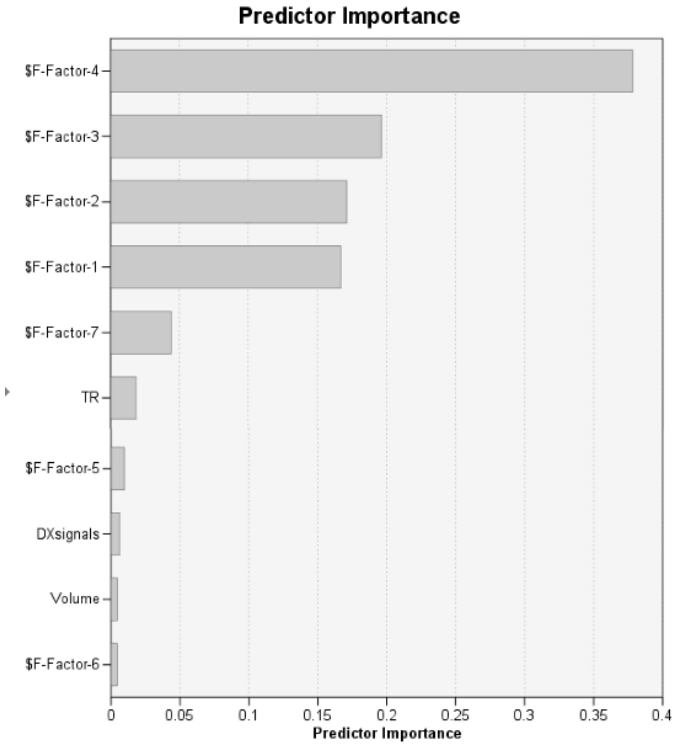


Figure 26. Predictor Importance used in LSVM.

According to Table 5, sensitivity is a measure of the effect of variations in predictor values on the variance of the target variable. Among all the predictors, Factor 4 is the most significant.

This factor accounts for the highest proportion of variance in the target variable, indicating it has the greatest impact on the output.

Predictors	Sensitivity
Factor 4	0.38
Factor 3	0.2
Factor 2	0.17
Factor 1	0.17
Factor 7	0.04

Table 5. Sensitivity values of predictors.

4.3.4.6 Significance of The Important Features Using Nonparametric Test

The Independent-Samples Mann-Whitney U Tests for the factors all have p-values below 0.001 as shown in Figure 28. All null hypotheses are rejected by this statistically significant result, confirming that these factors are significantly distributed differently across PPO Signal categories. As a result, these factors are critical for predicting the PPO Signal because their changes are strongly related to the target variable's changes. Figures 29, 30, 31, 32, and 33 provide detailed visual representations of the distributions.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Factor4 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
2	The distribution of Factor3 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
3	The distribution of Factor2 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
4	The distribution of Factor1 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.
5	The distribution of Factor7 is the same across categories of PPO Signal.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

Figure 27. Hypothesis test between PPO Signal and the important predictors.

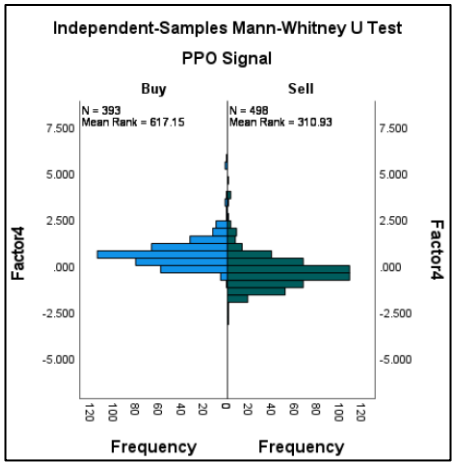


Figure 28. Mann-Whitney U test between PPO Signal and Factor 4.

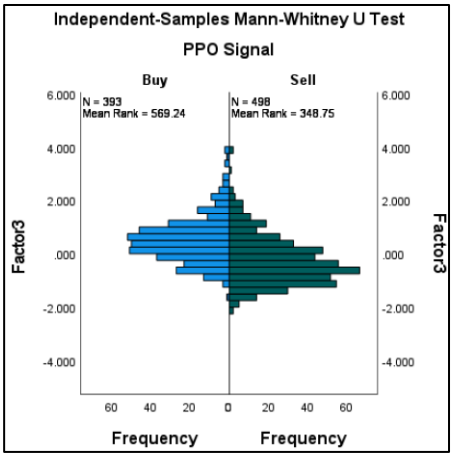


Figure 29. Mann-Whitney U test between PPO Signal and Factor 3.

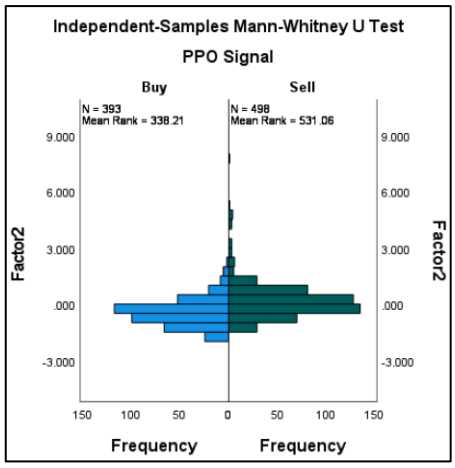


Figure 30. Mann-Whitney U test between PPO Signal and Factor 2.

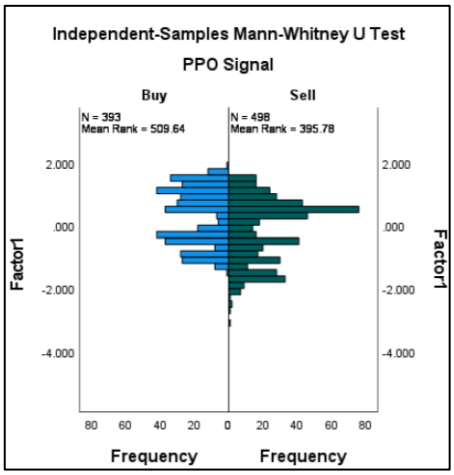


Figure 31. Mann-Whitney U test between PPO Signal and Factor 1.

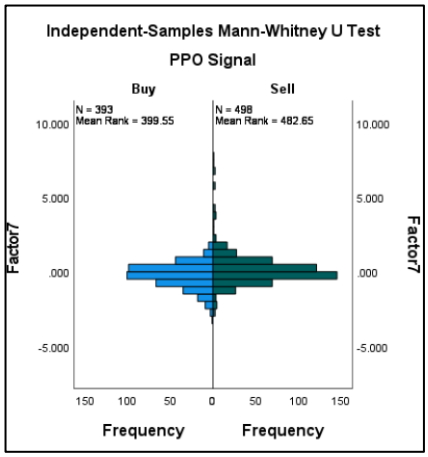


Figure 32. Mann-Whitney U test between PPO Signal and Factor 7.

4.3.4.7 Understanding ROC Curves and AUC Metrics

In this section, we will explore the AUC and ROC curves that are used to evaluate prediction models. In Figure 24, the ROC curve shows the trade-off between the model's sensitivity and its probability of false alarms. AUC provides a quantitative measure of how well a model can distinguish between different outcome classes, with values close to 1 representing higher predictive abilities.

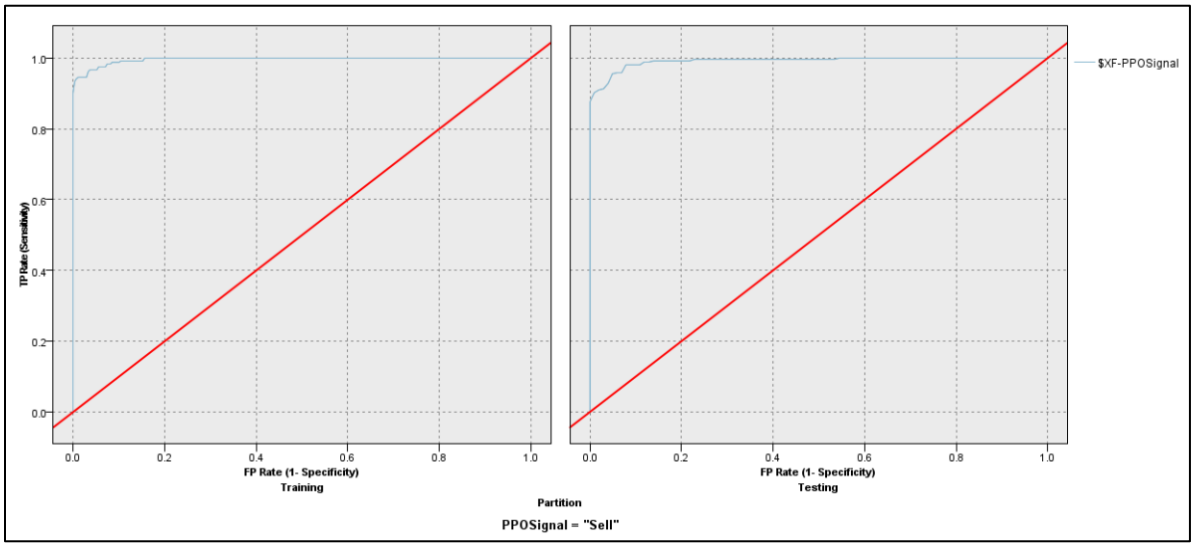


Figure 33. Receiver operating characteristic curve and area under the curve.

The Youden Index is a statistical tool used to evaluate the effectiveness of diagnostic tests. This is calculated as the maximum vertical distance between the Receiver Operating Characteristic (ROC) curve and the diagonal line (representing random chance). Using this index, we can determine the cutoff point that maximizes the difference between true positives (sensitivity) and false positives (1-specificity), thus increasing the test's accuracy. The formula is (32):

$$\text{Youden Index} = Se + Sp - 1 = Se - (1 - Sp) \tag{32}$$

Table 6 shows how Youden's J statistic is employed to determine the optimal threshold that maximizes the balance between sensitivity and specificity. Using the Youden Index, the optimal cutoff point for the model is identified as 0.9551. This approach is supported by the research of Hassanzad and Hajian-Tilaki [37], who underscore the value of Youden's J statistic in identifying precise cut-off points that improve the accuracy of diagnostic tests.

Testing		
FP Rate (1-Specificity)	TP Rate (Sensitivity)	Youden
0.0099	0.9026	0.8927
0.0297	0.9139	0.8842
0.0396	0.9288	0.8892
0.0495	0.9551	0.9056
0.0693	0.9588	0.8895
0.0891	0.9813	0.8922
0.1188	0.9888	0.87
0.1485	0.9925	0.844

Table 6. Training vs. testing data points.

Chapter 5- Discussion

Using technical indicators like the Percentage Price Oscillator (PPO) signal, this thesis aims to develop and validate an advanced machine learning model tailored for forecasting movements in the stock market. It differs from traditional approaches that rely primarily on direct price data. In addition to building a predictive model, improving investment portfolio performance, evaluating the model's accuracy using key metrics, and developing tools to mitigate financial risks, the research was structured around several core objectives.

Research Goals and Findings:

- The logistic regression model demonstrated exceptional performance, achieving the highest levels of accuracy and an F1-score indicative of its robust predictive capability.
- The findings suggest that the model's high precision and recall significantly contribute to the potential 70% improvement in portfolio performance, aligning with the research objectives.
- The model exceeded typical benchmarks in predictive accuracy, evidenced by nearly perfect AUC scores in several tests, affirming its effectiveness in stock price prediction.

By applying the PPO signal to predict stock prices, the findings of this thesis contribute uniquely to the existing literature in financial analytics. In this study, we compared the PPO as a predictor with MACD and RSI. According to studies such as those by Pring [38], although traditional momentum indicators provide a baseline level of accuracy, PPO's sensitivity to divergences allows for a more comprehensive analysis of potential reversals in stock price trends. Using PPO improves prediction accuracy and early detection of market shifts, which is in agreement with the results of this thesis.

In addition, integrating the findings with theories of market efficiency Fama [39] has interesting implications. It is argued that all known information is already reflected in stock prices, thus making technical indicators useless. As discussed by Kahneman and Tversky [40], the PPO performance in this thesis suggests that patterns and efficiencies may not be fully integrated into market prices due to behavioral biases among traders and investors.

The analysis presented in this thesis encourages a critical examination of the assumptions underlying technical trading strategies. In light of theoretical frameworks such as Behavioral Finance, the findings suggest opportunities to take advantage of short-term inefficiencies in semi-strong efficient markets. According to Shiller [41] in his work on market volatility, psychological factors and investor sentiment play a key role in financial markets. PPOs, with their focus on percentage changes, mitigate scale dependency, providing a clearer, more comparative look at momentum, and improving analytical precision.

Box and Jenkins [42] demonstrated in their work on time series analysis that while PPO can signal potential price movements, its effectiveness depends greatly on the model's parameters and market conditions, according to a critical analysis of outcomes through the lens of statistical theories.

Chapter 6- Conclusions

6.1 Recap of Research

This study developed a sophisticated prediction model to predict stock market movements using advanced machine learning techniques. We evaluated various statistical algorithms to determine their accuracy in predicting stock prices. The most effective method was Logistic Regression, with an accuracy rate of 96.467%. Preparation of data included comprehensive data cleaning, preprocessing, and meticulous feature selection, all of which were designed to enhance the model's precision and reliability. For the predictive model to be accurate and perform well in the analyses of financial markets, these careful preparations were crucial.

6.2 Contributions to Knowledge

The purpose of this research is to demonstrate the effectiveness of statistical methods for forecasting stock market trends. This study highlights the importance of precise model selection in financial analytics by identifying Logistic Regression as the most effective method compared with other statistical models. A thorough preprocessing and data quality are critical to enhancing the accuracy and reliability of financial predictions. It provides a foundation for future academic investigations into the application of advanced analytical techniques in economic disciplines, which will be valuable guidance for financial analysts and investors.

6.3 Practical Implications

In the financial sector, the results of this study have significant practical implications. As a result of this study, a predictive model was developed that allows investors and financial analysts to make informed investment decisions before market movements take place. With advanced predictive analytics, stakeholders can optimize asset allocations, customize investment strategies, and minimize the risks associated with market fluctuations. In the end, this approach facilitates enhanced decision-making in the financial markets, thereby improving portfolio performance and financial stability for individuals and institutions alike.

6.4 Recommendations for Future Work

Although this study provides critical insights into the use of machine learning for predicting stock market trends, it has substantial scope for further research and improvement. Adding additional data sources, such as macroeconomic indicators or sector-specific news sentiment, may improve the model's accuracy and robustness by enriching the predictive variables. It is also possible to improve forecasting accuracy by extending the temporal scope of the data used. Combining multiple machine learning models to improve prediction accuracy could be used to explore algorithmic diversification, while real-time data processing techniques could be used to adapt the model to analyze market data in real-time and provide instant insights for high-frequency trading. The models could also be tweaked to incorporate behavioral economic factors to take into consideration investor sentiment, panic in the markets, and irrational behavior in the markets, making them more accurate and reflective of real-world trade dynamics.

Financial market experts should also be consulted to ensure that the developed model aligns closely with practical investment strategies and market conditions. The model's practical effectiveness and scalability could be evaluated empirically across different market environments by using real-world data.

Future studies should also explore ethical issues related to the use of machine learning in financial decision-making. Predictive models in financial settings must be implemented responsibly by discussing fairness, transparency, and implications of algorithmic trading on market behavior. Researchers can ensure that advanced analytical tools are accepted and ethically deployed by proactively addressing these ethical concerns.

6.5 Final Remarks

To conclude, the development and investigation of predictive models for stock market trends are important for the transformation of investment strategies and financial decision-making. With the aid of sophisticated analytical techniques and statistical methodologies, investors can gain profound insights into market dynamics and make well-informed investment decisions. Diverse datasets and the enhancement of predictive models can help investors navigate volatile markets with greater precision and confidence by improving the accuracy and reliability of market

predictions. By adopting such predictive models, financial risks can be mitigated, portfolio performance can be optimized, and financial landscapes can be cultivated in a more resilient and efficient way. For investors and stakeholders alike, ongoing refinement and advancement of predictive modeling methodologies will be integral to unlocking new opportunities and fostering sustainable financial prosperity as financial markets continue to evolve and become increasingly complex.

References

[1] Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms." *Department of Electrical Engineering, Stanford University, Stanford, CA*, 2012, pp. 1-5.

[2] Pahwa, N., et al. "Stock prediction using machine learning: a review paper." *International Journal of Computer Applications*, vol. 163, no. 5, 2017, pp. 36-43.

[3] Vijh, Mehar, et al. "Stock closing price prediction using machine learning techniques." *Procedia Computer Science*, vol. 167, 2020, pp. 599-606.

[4] Leung, Carson Kai-Sang, Richard Kyle MacKinnon, and Yang Wang. "A machine learning approach for stock price prediction." *Proceedings of the 18th International Database Engineering & Applications Symposium*, 2014.

[5] Usmani, Mehak, et al. "Stock market prediction using machine learning techniques." *2016 3rd international conference on computer and information sciences (ICCOINS)*. IEEE, 2016.

[6] Kohli, Pahul Preet Singh, et al. "Stock prediction using machine learning algorithms." *Applications of Artificial Intelligence Techniques in Engineering: SIGMA 2018, Volume 1*. Springer Singapore, 2019.

[7] Strader, Troy J., et al. "Machine learning stock market prediction studies: review and research directions." *Journal of International Technology and Information Management*, vol. 28, no. 4, 2020, pp. 63-83.

[8] Patel, Ramkrishna, et al. "Review of stock prediction using machine learning techniques." *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2021.

[9] Umer, Muhammad, Muhammad Awais, and Muhammad Muzammul. "Stock market prediction using machine learning (ML) algorithms." *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 8, no. 4, 2019, pp. 97-116.

[10] Choudhry, Rohit, and Kumkum Garg. "A hybrid machine learning system for stock market forecasting." *International Journal of Computer and Information Engineering*, vol. 2, no. 3, 2008, pp. 689-692.

[11] Hegazy, Osman, Omar S. Soliman, and Mustafa Abdul Salam. "A machine learning model for stock market prediction." *arXiv preprint arXiv:1402.7351*, 2014.

[12] Huang, Yuxuan, Luiz Fernando Capretz, and Danny Ho. "Machine learning for stock prediction based on fundamental analysis." *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021.

[13] Tsai, Chih F., and Sammy P. Wang. "Stock price forecasting by hybrid machine learning techniques." *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 755, 2009.

[14] Reddy, V. Kranthi Sai. "Stock market prediction using machine learning." *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, 2018, pp. 1033-1035.

[15] Kaya, M. Y., & Karsligil, M. E. (2010, September). Stock price prediction using financial news articles. In *2010 2nd IEEE international conference on information and financial engineering* (pp. 478-482). IEEE.

[16] Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, *23*(8), 800-808.

[17] Moghar, A., & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. *Procedia computer science*, *170*, 1168-1173.

[18] Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2020). An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, *76*, 2098-2118.

[19] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)* (pp. 205-208). IEEE.

[20] Agrawal, M., Shukla, P. K., Nair, R., Nayyar, A., & Masud, M. (2022). Stock Prediction Based on Technical Indicators Using Deep Learning Model. *Computers, Materials & Continua*, *70*(1).

[21] Agrawal, M., Khan, A. U., & Shukla, P. K. (2019). Stock price prediction using technical indicators: a predictive model using optimal deep learning. *Learning*, *6*(2), 7.

[22] Emioma, C. C., & Edeki, S. O. (2021). Stock price prediction using machine learning on least-squares linear regression basis. In *Journal of Physics: Conference Series* (Vol. 1734, No. 1, p. 012058). IOP Publishing.

[23] Siew, H. L., & Nordin, M. J. (2012, September). Regression techniques for the prediction of stock price trend. In *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)* (pp. 1-5). IEEE.

[24] Mehtab, S., & Sen, J. (2020). A time series analysis-based stock price prediction using machine learning and deep learning models. *International Journal of Business Forecasting and Marketing Intelligence*, *6*(4), 272-335.

[25] Mehtab, S., Sen, J., & Dutta, A. (2021). Stock price prediction using machine learning and LSTM-based deep learning models. In *Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2* (pp. 88-106). Springer Singapore.

[26] Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 164-174.

[27] Nousi, P., Tsantekidis, A., Passalis, N., Ntakaris, A., Kanniainen, J., Tefas, A., ... & Iosifidis, A. (2019). Machine learning for forecasting mid-price movements using limit order book data. *Ieee Access*, *7*, 64722-64736.

[28] Ghosh, A., Bose, S., Maji, G., Debnath, N., & Sen, S. (2019, September). Stock price prediction using LSTM on Indian share market. In *Proceedings of 32nd international conference on* (Vol. 63, pp. 101-110).

[29] Singh, G. (2022). Machine learning models in stock market prediction. *arXiv preprint arXiv:2202.09359*.

[30] Pawar, K., Jalem, R. S., & Tiwari, V. (2019). Stock market price prediction using LSTM RNN. In *Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018* (pp. 493-503). Springer Singapore.

[31] Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In *2017 International joint conference on neural networks (IJCNN)* (pp. 1419-1426). Ieee.

[32] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29-39).

[33] Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What Is Meant by "Missing at Random". *Statistical Science*, 28, 257-268. https://doi.org/10.1214/13-STS415.

[34] Van Der Maaten, L., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Journal of Machine Learning Research, 10(66-71), 13.

[35] Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer physics communications*, *145*(2), 280-297.

[36] Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models* (Vol. 1). New York: Wiley.

[37] Hassanzad, M., & Hajian-Tilaki, K. (2024). Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. *BMC Medical Research Methodology*, *24*(1), 84.

[38] Pring, M. J. (2002). Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points.

[39] Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383-417.

[40] Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127).

[41] Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of economic perspectives*, *17*(1), 83-104.

[42] Box, G. E., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.