CSE343: Machine Learning Interim Report
Finance Forecaster: Using Machine Learning to predict stock prices

Aryan Dhull 2021520 | Deepanshu 2021524 | Pranav Aggarwal 2021551 | Prerak Gupta 2021552

Abstract

Precise stock price prediction remains a challenging yet highly valuable venture in today's fast-paced financial markets. As traders and investors seek to make knowledgeable decisions, the ability to forecast the ups and downs of the market plays an important role. Eventually, successful price prediction could certify market participants with magnified insights, likely reducing risks and optimizing investments in an ever-evolving economic landscape.

This project aims to grasp algorithms and data analysis techniques to unlock patterns hidden within historical stock data. Machine learning techniques, such as time series forecasting and regression can be applied to historical stock market data to predict future price movements. Developing accurate predictive models can be precious for traders, investors, and financial institutions looking to make informed decisions.

1. Introduction

In today's fast-paced and ever-evolving financial landscape, making informed investment decisions is paramount. Investors, traders, and financial analysts constantly seek innovative tools and techniques to gain a competitive edge in the dynamic world of stock markets. Machine Learning (ML), has emerged as a powerful tool to analyze historical data and make predictions about future stock price movements. The project's primary objective is to harness the potential of advanced ML algorithms and techniques to forecast stock prices with greater accuracy and reliability.

Throughout this report, we will explore the datasets related to the project, objectives, methodology, model details, and the specific ML algorithms employed. Furthermore, we will present the results and performance metrics of the model, comparing its predictions against real-world stock price movements to assess its effectiveness and practicality.

The findings presented herein aim to contribute to the ongoing discourse surrounding the integration of ML into the world of finance and investment, offering a glimpse into the exciting possibilities and challenges that lie ahead in this field.

2. Literature Survey

Mehar et al in *Stock Closing Price Prediction using Machine Learning Techniques*[3] makes use of Artificial Neural Network (ANN) to improve the accuracy of price prediction model due to the perceived insufficiency of historical dataset. Existing ANN model relies primarily on the "OHCVL" (Open, High, Close, Volume, Low) variables and has shown limitations in consistently delivering accurate forecasts. To address this challenge, new variables are introduced from existing variables only to provide a more comprehensive view of market dynamics.

The use of prediction algorithms in finance has raised questions regarding the compatibility with the *Efficient Market Hypothesis (EMH)* [2] proposed by Malkiel and Fama in 1970. The Efficient Market Hypothesis (EMH) asserts analyzing past returns yields no advantage, as markets swiftly adjust. Debates persist, but Support Vector Machines (SVM) are acknowledged as potent tools in finance for predicting stock price movements, challenging the traditional EMH principles.

Adebiyi et al in *Stock Price Prediction Using the ARIMA Model*[1]explores the use of time series forecasting models, specifically the Autoregressive Integrated Moving Average (ARIMA) model, for predicting stock prices. The ARIMA model is used to convert non-stationary data into stationary data using the differencing technique. The literature reveals that the ARIMA model is widely used in stock price prediction due to its ability to handle time series data effectively. The model has been applied in various contexts, from predicting the stock prices of specific companies to forecasting the demand for different commodities.

3. Dataset

In our study of financial data analysis, we utilized the Python library 'yfinance' to fetch comprehensive stock data from Yahoo Finance. This data collection spanned from 2010 to the current date and focused on four prominent technology companies: Google, Apple, Microsoft, and Amazon. The objective was to conduct long-term predictions based on this dataset. During our analysis, we enhanced the raw dataset, consisting of 3270 entries for each company, with additional features including Open, High, Low, Close, Adjusted Close, and Volume.

3.1. Feature Enhancement:

Recognizing the need for a more comprehensive analysis, we augmented the dataset. To achieve this, we calculated Moving Averages for 10, 20, and 50 days for each entry. Furthermore, we computed the Daily Return percentage for every entry, facilitating a deeper understanding of the stocks' performance and risks. In our analysis, we expanded the dataset by incorporating three additional features. These features represent the closing prices of the financial instrument for the day immediately preceding, two days preceding, and three days preceding the focal trading day. This addition allows us to consider the influence of recent historical prices in our predictive modeling.

3.2. Risk Analysis and Company Selection:

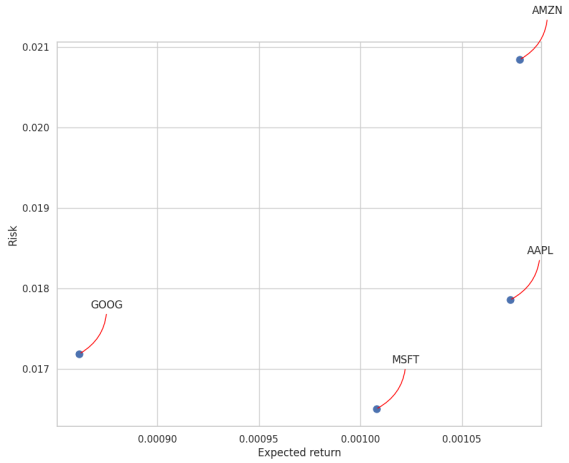

Figure 1. Risk vs Return for each Company

Upon completing the feature enhancement process, we employed the Daily Return percentage to assess the risks associated with each company's stock. Our meticulous analysis revealed that Microsoft exhibited the least risk, coupled with a high expected return. Consequently, we made an informed decision to proceed exclusively with Microsoft's data for our in-depth analysis.

3.3. Final Dataset:

After adding the necessary features and ensuring the removal of any NULL entries, our final dataset comprised 3221 entries, spanning the time frame from 2010 to the present day. This refined dataset featured 14 crucial attributes: Open, High, Low, Close, Adjusted Close, Volume, Moving Averages for 10, 20, and 50 days, Daily Return, Close1, Close2, Close3, and Next Day Closing Price.

4. Data Visualization, Preprocessing

4.1. Line Graphs and Histograms

The moving average (MA) stands as a fundamental tool in technical analysis, designed to provide a consistent, smoothed representation of price data. It achieves this by continually calculating an average price based on recent data points. In our analysis, we applied this method by plotting graphs illustrating moving averages over different periods, specifically 10, 20, and 50 days, for each company. This approach allowed us to effectively capture and comprehend the underlying trends within the stock prices, aiding in our understanding of the market dynamics.

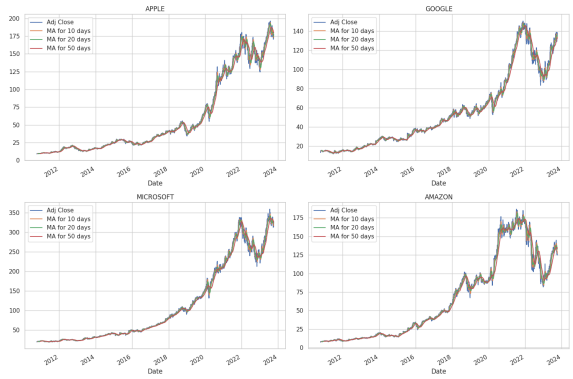

Figure 2. Moving Average for each Company

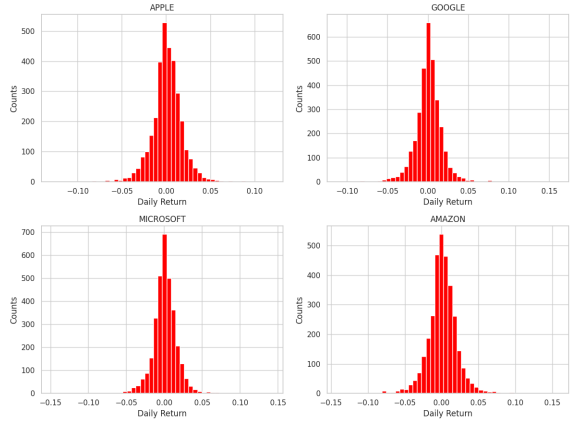

Figure 3. Daily Return for each Company

We employed histograms as a visual tool to gain insights into the daily returns of each company.

4.2. Correlation Heatmap

The difference in correlation between stock returns and stock closing prices among major tech companies offers valuable insights into market dynamics and investor behavior. When stock returns of major tech companies are not highly correlated, it suggests that the day-to-day fluctuations in the value of one company's stock are not strongly influenced by the fluctuations in another company's stock. On the other hand, the high correlation in stock closing prices indicates a strong link between the overall market trends and the closing values of these tech companies.

Figure 4. Correlation between each Company

4.3. Data Standardization

Standardization is a technique employed to transform data into a more manageable and comparable format. By standardizing data, the values are adjusted to center around the mean and have a unit standard deviation. The formula used for standardization is

$$z_i = \frac{x_i - x_{mean}}{\sigma}$$

Standardization ensures that the data follows a common scale, making it easier to analyze and interpret, especially when dealing with variables that may have different units or scales

4.4. Outlier detection Using BoxPlots and K-Means

Through the application of KMeans clustering and box plots, we successfully identified outliers within our dataset. Specifically, our analysis revealed that outliers were present only in the Volume and Daily Return variables. KMeans clustering allowed us to group similar data points together, enabling us to detect patterns and anomalies within the dataset. The box plots, on the other hand, provided a visual representation of the distribution of the data, making it easier to spot values that fell significantly outside the norm.

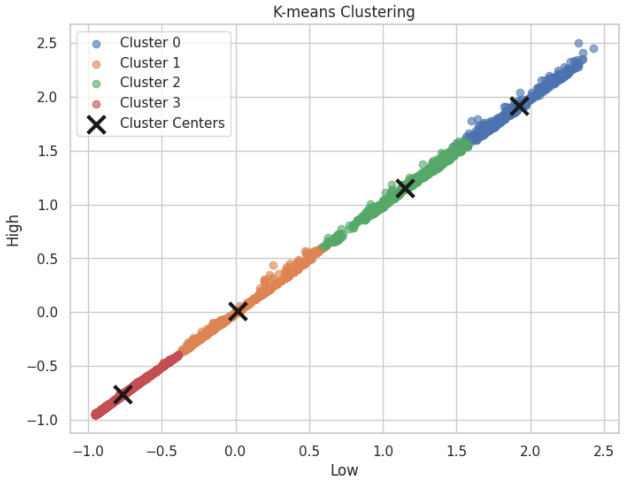

Figure 5. Clusters formed

5. Methodologies

The primary objective of this study is to identify the key features that significantly influence the next day's price of Microsoft (MSFT) stock.To accomplish this goal, we have employed various machine learning models, including classification with Naive Bayes and regression with Linear Regression and Support Vector Regression (SVR). We have undertaken a chronological train-test split with a ratio of 70:30 to ensure a realistic evaluation of our models on time series data.

5.1. Naive Bayes Classification

We started with the Naive Bayes classification algorithm on the clusters found during k-means Clustering, specially the Gaussian variant, as it is suitable for handling continuous numerical data. We evaluated the performance on the dataset using the precision-recall scores and overall accuracy.

5.2. Linear Regression

The objective of this analysis is to apply a baseline linear regression model to predict the next day's closing price of Microsoft (MSFT) stock using historical price data. Performance metrics including Mean Squared Error (MSE) and R-squared (R2) are computed.Feature scaling is applied to ensure that all input features have the same scale.

5.3. Support Vector Regression

SVR offers the advantage of capturing both linear and non-linear relationships and robustness to outliers, making it a valuable tool in the realm of financial forecasting. We conducted experiments on historical stock price data, employing SVR to model the relationship between various features and stock prices. Our findings indicate that SVR can

provide reasonably accurate predictions when linear kernel is used. We also explored the effect of Lasso and Ridge Regularization on SVR to prevent overfitting and improve the model's generalization. While Lasso can highlight the key features by introducing sparsity, Ridge can help mitigate the risk of multi-collinearity. The hyperparameter alpha was fine tuned using Cross Validation techniques.

6. Results and Analysis

Linear Regression performed quite well because of the data being heavily linearly related with an MSE of 23.63 and R2 score of 0.99. Applying k-Fold cross validation further showed that the model was not overfitting and gave as good results for different training sets.

Support Vector Regression also gave similar results with an MSE of 27.35 and R2 score of 0.99. Both the models showed extremely accurate predictions on the training set signifying overfitting and high variance.

Applying Lasso and Ridge Regularization with a hypertuned alpha led to improved observations of bias and variance but some amount of overfitting persisted.

7. Conclusion

The project aimed to predict the next day's closing stock price using machine learning techniques, specifically Linear Regression and Support Vector Regression (SVR). Regularization techniques Lasso and Ridge were applied to mitigate overfitting, given the linear nature of the data.
The models, particularly the regularized linear regression and SVR models, produced exceptionally accurate predictions for the next day's closing stock prices. The Mean Squared Error (MSE) and R-squared (R2) values indicated robust model performance.

7.1. Learnings

Model Performance: The regularized Linear Regression and SVR models demonstrated strong predictive capabilities for stock price movements. The use of regularization techniques helped prevent over fitting and improved model generalization.

Feature Importance: Feature extraction and engineering played a crucial role in model success. By selecting relevant OHLCV features and incorporating technical indicators like moving averages and daily returns, the models captured essential information to make accurate predictions.

Data Standardization: Standardizing the feature data through techniques like StandardScaler enhanced model performance and convergence, especially in the case of SVR.

Regularization: The incorporation of Lasso and Ridge regularization was pivotal in preventing over fitting, ensuring model stability, and improving generalization. It allowed the models to maintain accuracy while avoiding unnecessary complexity.

7.2. Work Left

The project has been well according to the timeline. We aim to try more models as per the proposed timeline. Eventually, in the coming weeks, we will apply several other models, such as using Decision Trees for regression, Random Forests, Neural Networks, ARIMA, and SARIMAX. In the end, we will finally account for issues of Bias and variance and try to combat them as well using Regularization and Hyperparameter tuning.

7.3. Contributions

We have all helped each other in various parts of the project, and the overall has been mainly a collective team effort.

Aryan Dhull: Data Collection, Exploratory Data Analysis, Data Preprocessing, K-Means, Report + Presentation

Deepanshu: Data Collection, Data Preprocessing, Linear Regression, SVR, Report + Presentation

Pranav Aggrawal: Exploratory Data Analysis, K-Means, Naive-Bayes, Report + Presentation

Prerak Gupta: Data Collection, Linear Regression, SVR, Report + Presentation

References

[1] Adebiyo et al. *Stock Price Prediction Using the ARIMA Model* [1].

[2] Michel Ballings et al. *Evaluating multiple classifiers for stock price direction prediction* [2].

[3] Mehar Vijh et al. *Stock Closing Price Prediction using Machine Learning Techniques* [3].