CSE343: Machine Learning Final 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

Lili Yin et al in *Research on stock trend prediction method based on optimized random forest*[2] proposes a method for stock trend prediction based on optimized random forest. It highlights the challenges and limitations of these techniques, such as dealing with time series data, selecting technical indicators, and optimizing the model parameters. The paper highlights the advantages of random forest over other models, such as robustness to outliers, strong generalization ability, and better performance in ensemble learning. This study applies the exponential smoothing method to process the initial data and calculates the relevant technical indicators.

Tsai et al in *Stock Price Forecasting by Hybrid Machine Learning Techniques*[3] describes decision trees (DT) as another data mining technique that can provide reasonable classification and forecasting performances. DT has a tree structure that depends on different situations to create nodes and branches. Each node represents an output class and each branch represents a process for classification. The paper uses the CART technique to create and prune the decision tree model. The paper also mentions that DT can generate decision rules for further analyses.

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. It 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:

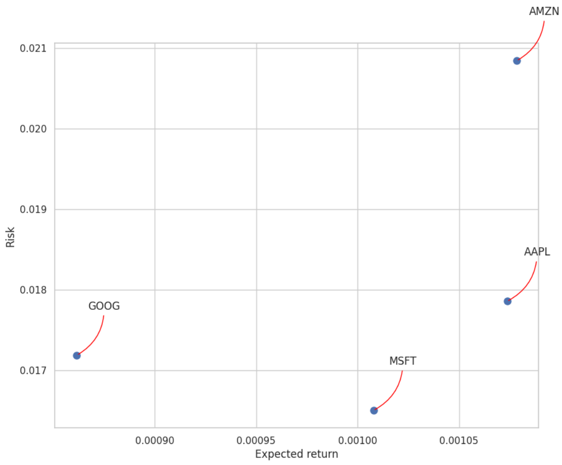

Figure 1. Risk vs Return for each Company

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.

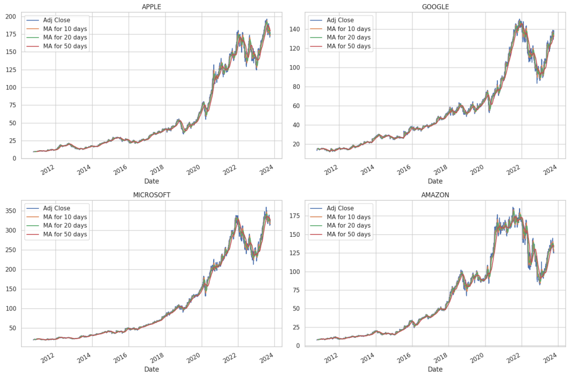

Figure 2. Moving Average for each Company

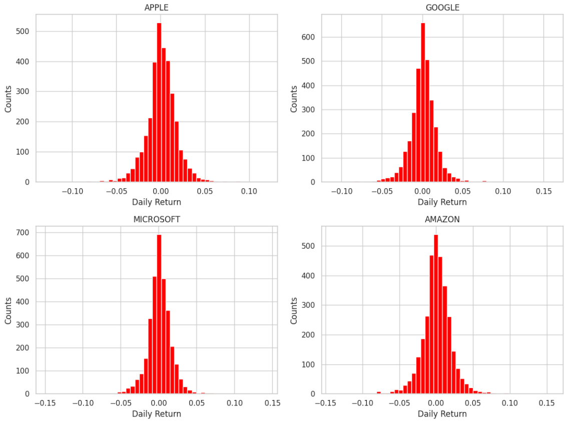

Figure 3. Daily Return for each Company

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.

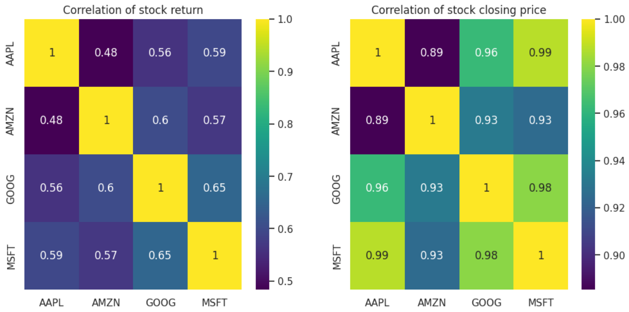

Figure 4. Correlation between each Company

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.

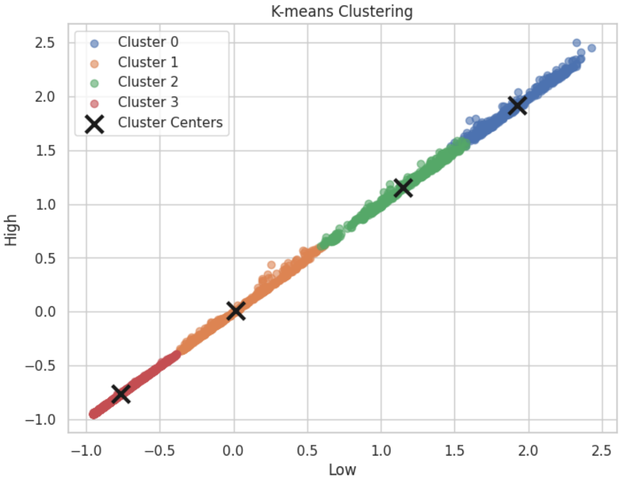

Figure 5. Clusters formed

4.5. Exponential Smoothing

Exponential smoothing is a widely employed technique in forecasting models. It builds upon the moving average method while addressing the limitation of not incorporating long-term data. Simultaneously, it accommodates certain aspects of overall period averaging in data processing. This involves assigning greater weight to data with shorter time distances and progressively decreasing weight to data with longer time distances. The exponential smoothing of a series Y can be represented recursively throughout the process.

$$ES_0 = Y_0, \qquad t = 0$$
$$ES_t = \alpha * Y_t + (1 - \alpha) * ES_{t-1}, \qquad t > 0$$

Among them, ES_t is the smoothing value at time t, Y_t is the actual value of time t, and α (the value range is (0, 1]) is the smoothing constant.

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.

5.4. Decision Trees

Decision trees in Machine Learning are used for building classification and regression models to be used in data mining and trading. A decision tree algorithm performs a set of recursive actions before it arrives at the end result.
The methodology involves Decision Tree Regressor for predictive modeling. It conducts a hyperparameter search using GridSearchCV, exploring various combinations of 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'max_features'. The best-performing model is then selected and used to make predictions on the test set.

5.5. Random Forests

The random forest algorithm is designed to construct an ensemble of decision trees. Each individual decision tree produces a distinct feature subset, facilitating the algorithm's selection of the most discriminative attribute for dataset segmentation. These trees independently contribute to a collective decision through a majority voting process, ultimately resulting in the creation of the random forest model.
The methodology involves a Random Forest Regressor for stock price prediction. It splits data into training and testing sets, trains the model, and evaluates its accuracy, enhancing predictions through ensemble learning and parameter tuning.

5.6. Artificial Neural Networks(ANNs)

Artificial Neural Networks (ANNs) are characterized by attributes such as robustness, fault tolerance, learning capability, generalizability, adaptability, universal function approximation, and parallel data processing. These inherent features empower ANNs to effectively address complex problems characterized by non-linear input-output relationships.
A sequential model of 6 hidden layers was build on the dataset using all the 14 features. The first hidden layer of the model consists of 64 neurons with the 'relu' activation function and dropout regularisation of 0.2. the second, fourth, fifth and sixth hidden layer consists of 32 neurons with other specifications same as previous layer. The seventh hidden layer (output layer) of the model consists of 1 neuron with the 'linear' activation function.
The model was then trained through 50 epochs with a batch size of 32, split into 80:20 ratio for training and validation. The trained model was thus used for predictions, and was evaluated using mse and r2 score.

5.7. ARIMA

The autoregressive integrated moving average (ARIMA) model is a technique applied to time series data which is defined by three parameters denoted as p, d, and q, where p signifies the order of the autoregressive (AR) term, q represents the order of the moving average (MA) term, and d indicates the order of differencing required to transform a non-stationary time series into a stationary one. The procedural steps begin with model identification based on the analysis of autocorrelation (ACF) and partial autocorrelation (PACF) plots.
To analyze and model the time series behavior of the closing prices of the financial instrument . The ARIMA model, with the specified order parameters (p=0, d=1, q=1), is then applied to capture any autoregressive, differencing, and moving average components in the 'Close' prices for predictive purposes.

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.

Decision Tree and Random Forest did not perform to the expectations, having overfit extremely. Pruning and hyperparameter tuning using GridSearchCV was tried but it did not give promising results with the mean squared error being 12000 compared to a very low MSE for training data.

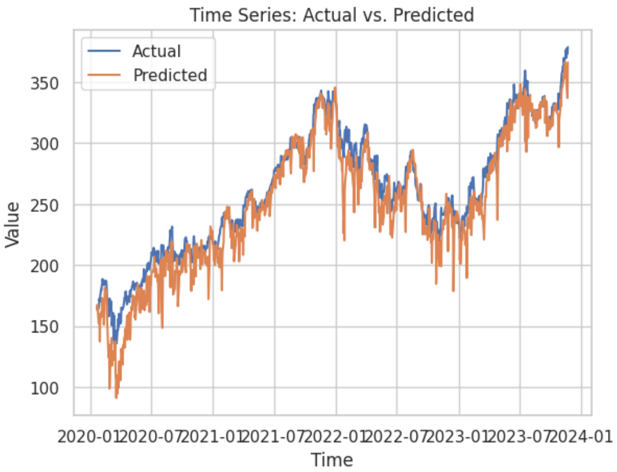

Figure 6. Predictions by ANN

ANNs, on the other hand, performed extremely well with a training MSE 34.78, testing MSE 308.5 and R2 Score 0.88. It was also observed that the model captured the trends of the stock fluctuations.

ARIMA Model was able to completely capture the patterns of the dataset it was fit on, but it is not able to predict a good forecast on that testing data as we can see in the graph, the step forecast keeps following the last forecast value (a horizontal line).

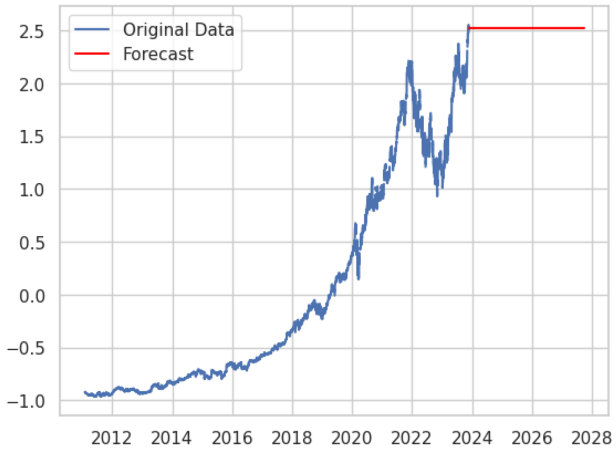

Figure 7. Predictions by ARIMA

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.

Decision Tree and Random Forest exhibited significant overfitting despite efforts at pruning and hyperparameter tuning. Complexity might be an issue, and further feature engineering is needed. Artifical Neural Networks outperformed the other models with a low MSE and high R2 Score, highlighting that it's well suited for modeling the underlying patterns. ARIMA Model effectively captured patterns in the data but struggled with accurate forecasting.

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.

Ensemble Learning: The use of Random Forests, an ensemble of decision trees, demonstrated effectiveness in predicting stock prices. By constructing individual decision trees and combining their predictions through a majority voting process, the random forest model provided a robust approach to stock price prediction. However, the model overfitted the training dataset and gave severe reults on testing.

Model Complexity and Overfitting: The observed overfitting in Decision Tree and Random Forest models emphasizes the importance of striking a balance between model complexity and simplicity; overly complex models may struggle to generalize to unseen data.

Neural Networks: The use of ANNs, characterized by attributes like robustness and adaptability, proved effective in addressing complex problems with non-linear input-output relationships. The sequential model with multiple hidden layers demonstrated the capability to learn and make predictions based on the provided features. The choice of activation functions, dropout regularization, and model architecture played key roles in achieving accurate predictions.

Time Series Modelling: The ARIMA model, applied to time series data, demonstrated its utility in predicting stock prices. The procedural steps involving model identification based on autocorrelation and partial autocorrelation plots provided a systematic approach to analyzing and modeling time-dependent patterns in the data.

7.2. Future Work

Following the work already done, future work will focus on the implementation of advanced machine learning techniques and deep learning models. While traditional machine learning models, excluding Artificial Neural Networks (ANN), have shown limited success in achieving satisfactory accuracy, we aim to explore methodologies to push the boundaries of predictive capabilities. Notably, the research will delve into the application of sophisticated ensemble learning methods such as Gradient Boosting Machines (GBM) and XGBoost, which have demonstrated success in various prediction tasks due to their ability to capture complex relationships within the data. Additionally, the integration of state-of-the-art deep learning architectures like Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs) will be a primary focus.

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, ANN, Report + Presentation

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

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

Prerak Gupta: Data Collection, Linear Regression, Support Vector Regression, ARIMA , Report + Presentation

References

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

[2] Lili Yin et al. *Research on stock trend prediction method based on optimized random forest* [2].

[3] Tsai et al. *Stock Price Forecasting by Hybrid Machine Learning Techniques*[3].