# CSE 343 : ML Project Monsoon 2023

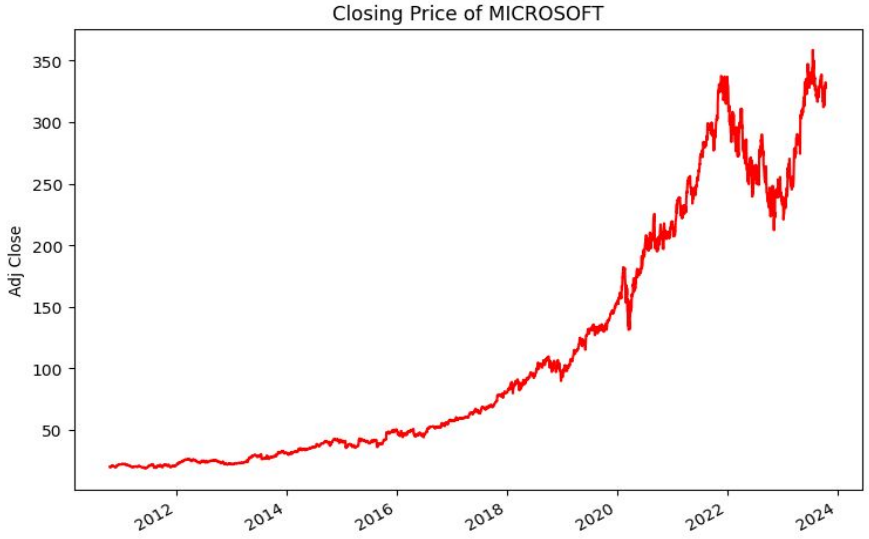
Aryan Dhull- 2021520
Deepanshu- 2021524
Pranav Aggarwal- 2021551
Prerak Gupta- 2021552

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**

# Finance Forecaster:

## Using Machine Learning to predict stock prices






# Motivation

- 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.

# Motivation

- 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.

# Literature Review

## 1. Stock Closing Price Prediction using Machine Learning Techniques

**Authors-**Mehar Vijh , Deeksha Chandola , Vinay Anand Tikkiwal , Arun Kumar

- This paper emphasis the use of ANN to improve the accuracy of price prediction model by introducing new variables created from the existing variables only to provide a more comprehensive view of market dynamics.

- 6 new variables were added which are :- 1. Stock High minus Low price (H-L) 2. Stock Close minus Open price (O-C) 3. 7 DAYS MA (moving average) 4. 14 DAYS MA 5. 21 DAYS MA 6. 7 DAYS STD DEV

- ANN is capable for finding hidden features through a self learning process and hence able to find the input and output relationship of a very large complex dataset.

# Literature Review

## 2. Evaluating multiple classifiers for stock price direction prediction

**Authors-**Michel Ballings , Dirk Van den Poel , Nathalie Hespeels , Ruben Gryp

- A great point of discussion in literature is whether stock price behavior is predictable or not. For a long time investors accepted the Efficient Market Hypothesis (EMH).

- This paper set out to benchmark the performance of ensemble methods (Random Forest, Adaboost and Kernel Factory) against single classifier models (Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbors) in predicting stock price direction.

- The direction of stock prices was predicted instead of absolute stock prices and it was established that  Random Forest is the top predictor followed at a distance by SVM.

# Dataset

We've harnessed yfinance, a library that taps into the Yahoo Finance API, for real-time stock price data. It's a crucial asset for making informed financial decisions.

Our dataset spans from 2010 to the present day, and it revolves around four technology giants: Google, Apple, Microsoft, and Amazon. These companies form the focal point of our analysis.

# Dataset

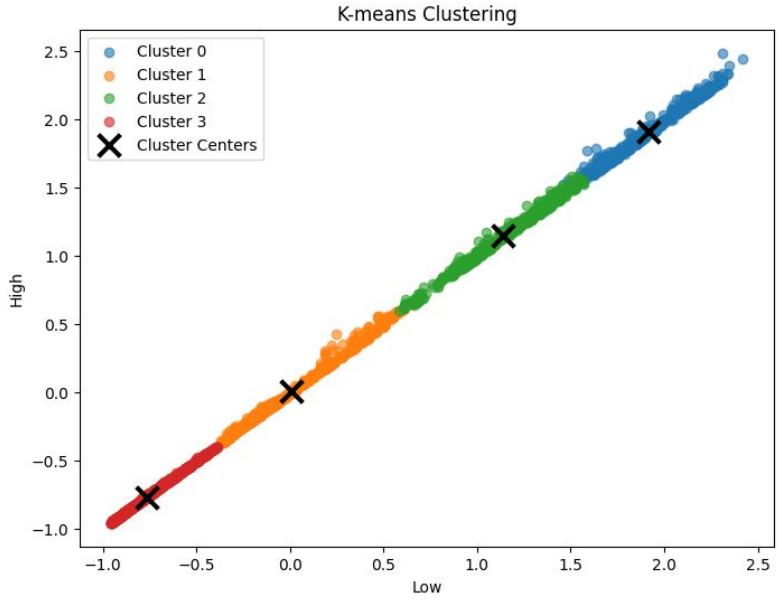
After augmenting the dataset with necessary features and conducting a comprehensive analysis of the return and risk metrics for each company, we have successfully compiled a refined dataset specifically focused on Microsoft. This dataset comprises 14 key features and encompasses over 3000 entries, offering a detailed perspective on Microsoft's stock prices.

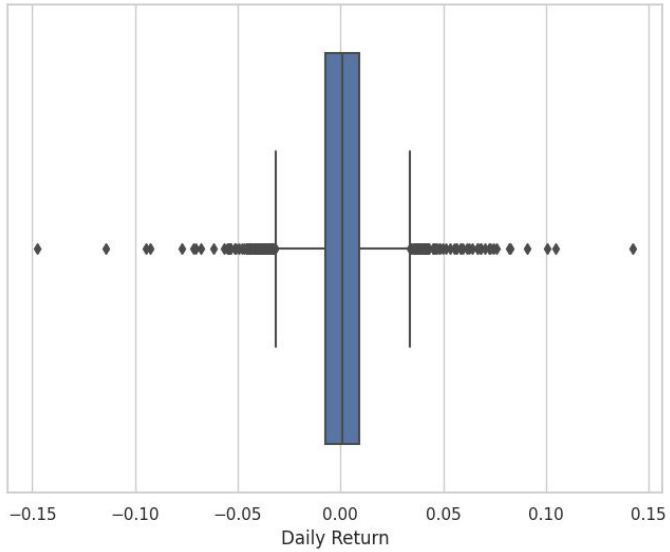
| Date | Open | High | Low | Close | Adj Close | Volume | MA for 10 days | MA for 20 days | MA for 50 days | Daily Return | Close1 | Close_2 | Close_3 | Next_Day_Closing_Price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2010-12-28** | 27.969999 | 28.170000 | 27.959999 | 28.010000 | 21.731487 | 23042200 | 21.708984 | 21.243864 | 20.579771 | -0.002137 | 28.010000 | 28.070000 | 28.299999 | 27.969999 |
| **2010-12-29** | 27.940001 | 28.120001 | 27.879999 | 27.969999 | 21.700445 | 19502500 | 21.736138 | 21.348991 | 20.615578 | -0.001428 | 27.969999 | 28.010000 | 28.070000 | 27.850000 |
| **2010-12-30** | 27.920000 | 28.000000 | 27.780001 | 27.850000 | 21.607349 | 20786100 | 21.736138 | 21.419206 | 20.660629 | -0.004290 | 27.850000 | 27.969999 | 28.010000 | 27.910000 |
| **2010-12-31** | 27.799999 | 27.920000 | 27.629999 | 27.910000 | 21.653900 | 24752000 | 21.729932 | 21.458774 | 20.703371 | 0.002154 | 27.910000 | 27.850000 | 27.969999 | 27.980000 |
| **2011-01-03** | 28.049999 | 28.180000 | 27.920000 | 27.980000 | 21.708214 | 53443800 | 21.736139 | 21.496015 | 20.745503 | 0.002508 | 27.980000 | 27.910000 | 27.850000 | 28.090000 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2023-10-06** | 316.549988 | 329.190002 | 316.299988 | 327.260010 | 327.260010 | 25645500 | 317.263004 | 323.117001 | 325.002360 | 0.024737 | 327.260010 | 319.359985 | 318.959991 | 329.820007 |
| **2023-10-09** | 324.750000 | 330.299988 | 323.179993 | 329.820007 | 329.820007 | 19891200 | 318.491003 | 322.711002 | 324.845659 | 0.007823 | 329.820007 | 327.260010 | 319.359985 | 328.390015 |
| **2023-10-10** | 330.959991 | 331.100006 | 327.670013 | 328.390015 | 328.390015 | 20557100 | 320.116003 | 322.542003 | 324.709253 | -0.004336 | 328.390015 | 329.820007 | 327.260010 | 332.420013 |
| **2023-10-11** | 331.209991 | 332.820007 | 329.140015 | 332.420013 | 332.420013 | 20063200 | 322.079004 | 322.360004 | 324.645065 | 0.012272 | 332.420013 | 328.390015 | 329.820007 | 331.160004 |
| **2023-10-12** | 330.570007 | 333.630005 | 328.720001 | 331.160004 | 331.160004 | 19313100 | 323.831003 | 321.983003 | 324.732104 | -0.003790 | 331.160004 | 332.420013 | 328.390015 | 327.730011 |

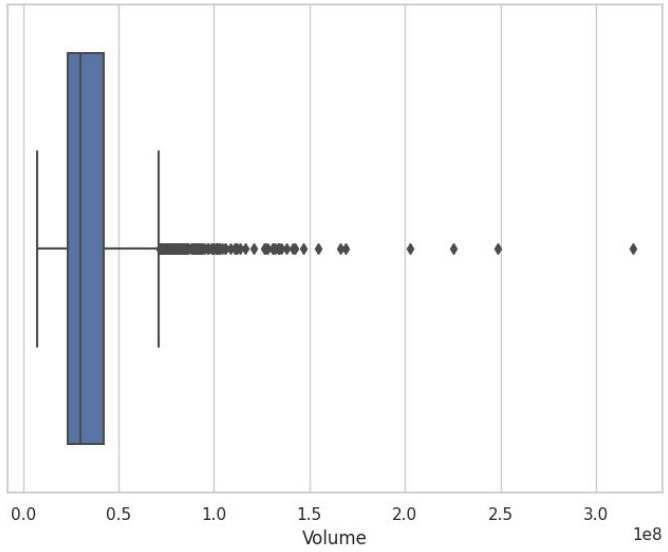
3220 rows × 14 columns

# Dataset

We standardized the dataset to ensure uniformity and then applied K-means clustering and boxplot analysis to identify outliers. Among the 14 features, outliers were detected specifically in the 'Daily Return' and 'Volume' variables, indicating deviations from the general patterns observed in these two aspects of the data.



# Methodology

- To ensure a realistic evaluation of our models on time series data, we've implemented a chronological train-test split with a **70:30** ratio.

- Our Next objective is to use a baseline **linear regression** model to predict Microsoft's (MSFT) next day's closing stock price based on historical data. We compute performance metrics, such as **Mean Squared Error (MSE)** and **R-squared (R2)**. All input features were modified so as to have the same scale.

- We then applied **Support Vector Regression (SVR)** to capture both linear and non-linear relationships and to enhance robustness against outliers in financial forecasting.

# Methodology

- Our experiments on historical stock price data involved using SVR to model the relationship between various features and stock prices. Results showed SVR to be effective, particularly with a linear kernel.

- We also delved into **Lasso** and **Ridge** Regularization to prevent overfitting and enhance model generalization. Lasso introduced sparsity, emphasizing key features, while Ridge mitigated multicollinearity risks.

- Fine-tuning the hyperparameter **alpha** was achieved through Cross Validation techniques.

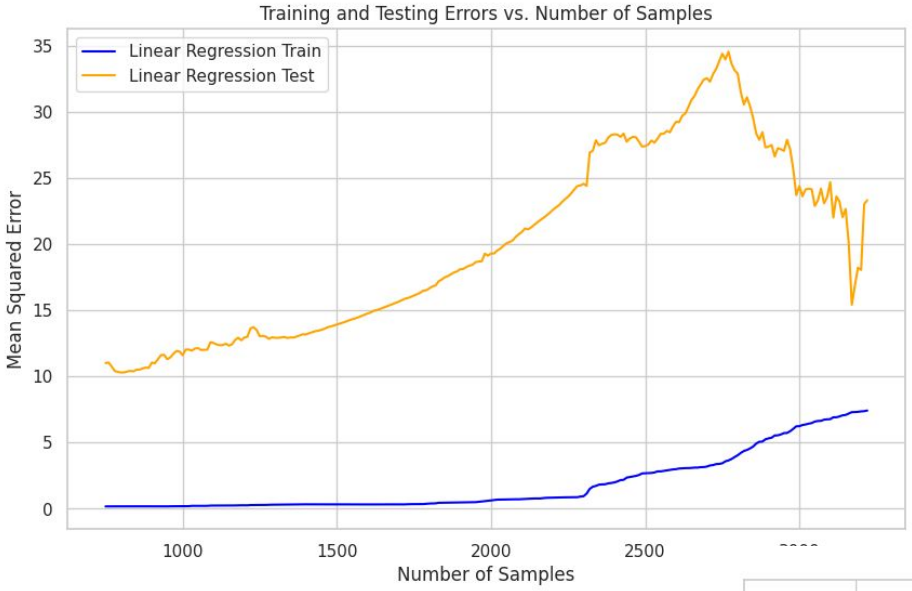
# Result & Analysis

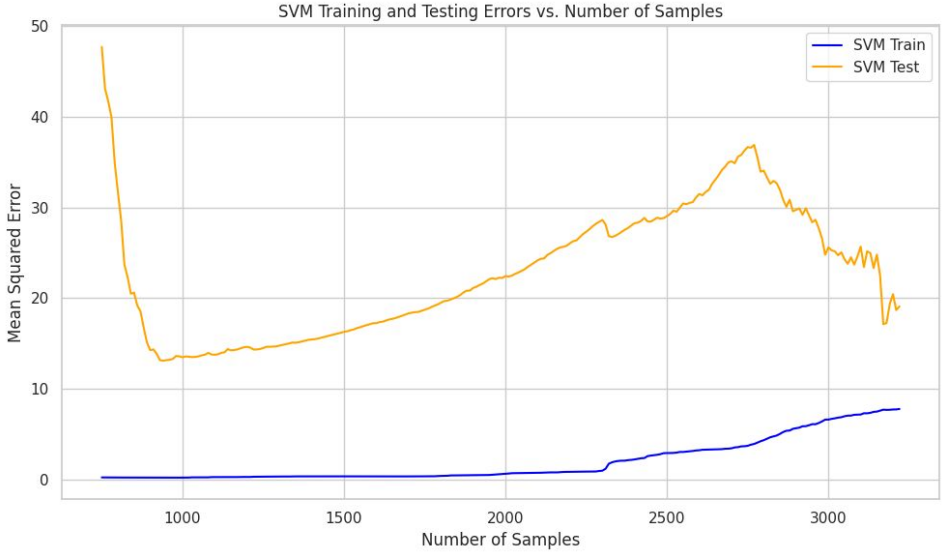
- **Linear Regression** performed quite well because of the data being heavily linearly related with an **MSE** of **23.63** and **R2 score** of **0.99**.

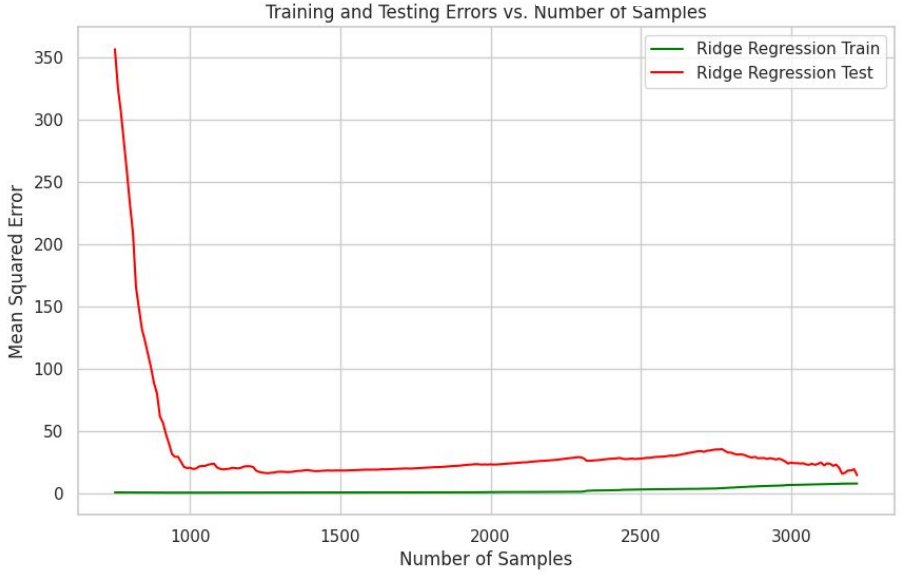
- **Support Vector Regression** also gave similar results with an **MSE** of **27.35** and **R2** score of **0.99**. Both the models showed extremely accurate predictions on the training set signifying overfitting and higher variance.

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

# Result & Analysis



# Timeline

Up until Week 7, our progress has aligned nearly with the planned timeline. Moving forward, we are committed to integrating newly acquired skills and techniques from our ongoing coursework. Our goal is to leverage these additional capabilities to achieve the most optimal outcomes in our project.

We plan to explore time series models like **ARIMA** and **SARIMAX** in our future analysis. These advanced techniques are expected to yield better results, enhancing the accuracy of our predictions as we move forward.

**Week 8:** Random Forest, Decision Tree and Neural Networks
**Week 9:** ARIMA and SARIMAX
**Week 10:** Hyperparameter Tuning and Outcome Analysis
**Week 11-12:** Final Report

# Contributions

| | |
|---|---|
| Aryan Dhull | Data Collection, Exploratory Data Analysis, Data Preprocessing, K-Means, Report + Presentation |
| Deepanshu | Data Collection, Data Preprocessing, Linear Regression, SVR, Report + Presentation |
| Pranav Aggarwal | Exploratory Data Analysis, K-Means, Naive-Bayes, Report + Presentation |
| Prerak Gupta | Data Collection, Linear Regression, SVR, Report + Presentation |

# QnA!