Group Project

Data-Driven Analysis and Prediction of Stock Market Performance for Top 10 Companies (2017-2022)

BY

Manoj Mareedu (MXM220069)

Mounika Dantuluri (LXD210022)

Nihal Darak (NXD220015)

Kundana Rasi Tadikonda (KXT220013)

Under the Guidance of Prof. Thomas Lavastida

Table of Contents

Executive Summary

This project analyses data (stock prices) for top 10 companies from 2017-2022. The motivation behind this project is to identify trends and patterns in the stock market that could inform investment strategies or predictive models. The dataset includes information about the opening, highest, lowest, and closing prices, as well as the volume traded, for each day. The data is cleaned and preprocessed before being visualized and analyzed using a jupyter notebook. The findings of the project include several interesting insights, such as the overall upward trend in stock prices over the five-year period, as well as variations in stock performance across different companies and sectors. Overall, this analysis provides valuable information for investors and financial analysts who are interested in understanding the stock market and developing data-driven investment strategies.

Introduction

The stock market is a complex and dynamic system that can be influenced by a wide range of factors, including economic indicators, political events, and social trends. As such, investors and financial analysts are constantly seeking new ways to analyze and understand the stock market to make informed investment decisions.

The motivation behind this project is to explore stock price data for the top 10 companies from 2017-2022, with the goal of identifying trends and patterns that could inform investment strategies or predictive models. By analyzing this data, we hoped to gain insights into the factors that drive stock prices and to develop a deeper understanding of the stock market.

Additionally, the project aims to address the challenge of processing and analyzing large volumes of data. The stock price data for multiple companies over a five-year period is a large and complex dataset that requires specialized tools and techniques to make sense of. By demonstrating how to clean, preprocess, and analyze this type of data, we hoped to provide a valuable resource for other analysts and investors who are interested in exploring historical stock price data.

In summary, the project aims to address the challenges of analyzing and understanding the stock market by exploring historical stock price data for multiple companies from 2017-2022, with the goal of identifying trends and patterns that could inform investment strategies or predictive models.

Dataset and Preprocessing

The data set contains observations of the stocks of the top ten companies from 2017 to 2022. There is one categorical variable (Company) and seven quantitative variables (such as open value, closing value, volume, and so on) in this dataset. The data set consists of around 12,580 records. The dataset contains the stock information of APPLE, AMAZON, GOOGLE, JP-MORGAN, MICROSOFT, NETFLIX, NVIDIA, TESLA, VISA, and WALMART.

To understand the data, various statistical analysis and visualization techniques have been used. These techniques help to identify trends and patterns in the data, as well as potential outliers and influential observations.

Statistical analysis techniques used include:

• Descriptive statistics: used to summarize and describe the distribution of the data, such as the mean, median, and standard deviation.

• Panel Data Analysis: used to analyze data that contains observations over time and across multiple individuals or groups.

Visualization techniques used include:

• Line charts: used to visualize the trends in the data over time, such as the trend in the closing price or volume traded.

• Scatter plots: used to visualize the relationship between two variables, such as the relationship between the closing price and volume traded.

• Correlation analysis: used to explore the relationships between variables, such as the correlation between the closing price and other variables like volume or price change.

Preprocessing

To address the issue of duplicate records in the dataset, the '.drop_duplicates' method was used initially. Following that, the presence of null values in the dataset was identified, but it was found to be clean with no null values. To facilitate further analysis, the 'Date' column was converted to Datetime format. Additionally, the names of two companies, JP-MRGN and M-SOFT, were replaced with JPMRGN and MSOFT respectively to avoid errors while constructing OLS models. Duplicate variables were created for the 'Company' column to use in regression models. These steps helped in preparing a clean and accurate dataset for analysis.

Exploratory Data Analysis

Line Plot analysis

To understand the stock closing value trend over a five-year period, a line plot has been used to show the trend of close price. This visual analysis shows that for companies like Amazon, Netflix, Visa and Walmart, the stock prices have significantly increased compared to the rest of the companies.

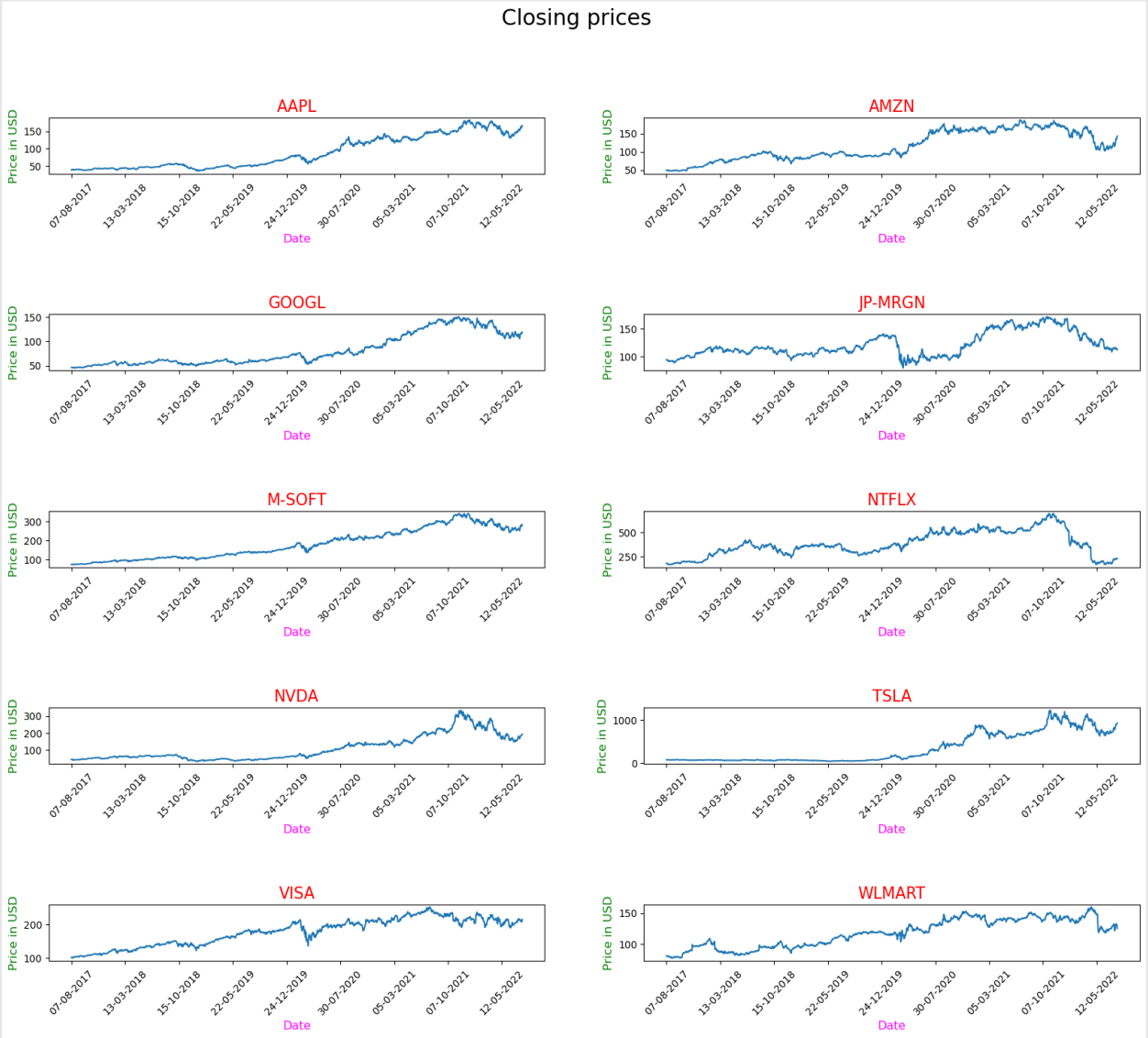

Fig 1. Close value trend over the five-year period

Pair plot Analysis

This plot depicts the trend of each column with respect to others. It only considers numerical values. We see that there's a negative correlation between volume and all other variables like open, close etc., indicating that as volume increases, the value of other variables is decreasing and so, volume has less significant impact on the stock prices.

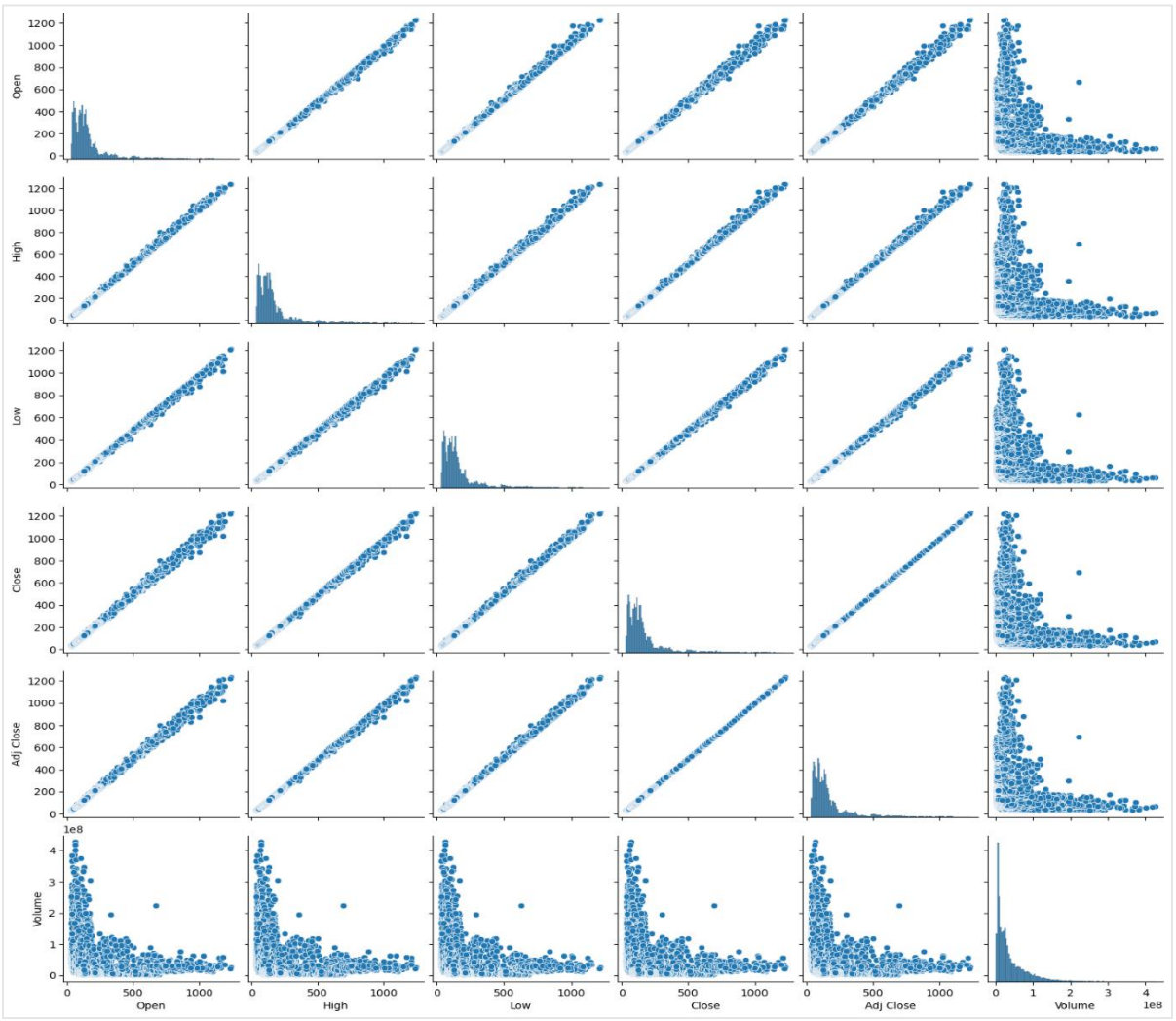

Fig2. Pairplot

Correlation Analysis

To avoid multicollinearity, we conducted a correlation analysis using *.corr()* method to examine the correlation between variables. The results indicated that the 'Volume' variable in the dataset showed no significant correlation with any of the independent variables or the dependent variable 'Close'. On the other hand, the variables 'Open', 'High', 'Low', and 'Adj Close' were highly correlated with each other as well as with the dependent variable 'Close'. Based on these findings, we decided to choose 'Open' and 'Volume' variables as the explanatory variables for the dependent variable 'Close'. These variables are less likely to suffer from multicollinearity issues and may offer more meaningful insights into predicting the 'Close' value. This analysis will be instrumental in building an effective regression model for stock price prediction.

Fig 3. Correlation matrix

From the correlation matrix the linear equation is as followed:

Close = (open) + (Volume) + (Company dummy variables)

In the model, we incorporated dummy variables for the categorical 'Company' variable, as it is essential in regression. We observed that the 'Open' variable is positively correlated with the 'Close' variable, but no significant correlation was found between 'Open' and 'Volume'. This suggests that multicollinearity is not an issue, and a linear regression model can be developed using the linear equation mentioned.

Models

There are several models that can be used for stock value predictions such as Linear Regression, Random Forest, Gradient Boosting, etc. Out of which we have used Linear regression because it is a simple and commonly used statistical method for modeling the relationship between a dependent variable and one or more independent variables. While random forest and gradient boosting can be powerful models for predicting complex relationships, linear regression can still produce accurate predictions in many cases, especially when the relationship between the independent and dependent variables is linear.

Linear Regression Model

Initially, a linear regression model was built to observe the impact of each independent variable and dummy variables on the dependent variable. One of the Company dummy variables is ignored as a

reference. ‘Train_test_split’ was imported and employed to split the data into training and test datasets. We trained the model on the training dataset and tested the model performance using the testing dataset.

The linear regression model that was developed revealed valuable insights regarding the relationship between the variables. The coefficient for 'Open' was found to be 0.997, indicating that an increase in the 'Open' variable will lead to a 0.997 unit increase in the 'Close' variable, provided that all other variables remain constant. On the other hand, the coefficient for 'Volume' was determined to have no significant impact on 'Close' when other variables are held constant. The remaining coefficients in the model illustrate the impact of each company on the target variable, with 'WLMART' as the reference dummy variable. The coefficient for 'NVDA' was found to be -0.079, implying that 'NVDA' has a lower expected value of 'Close' compared to 'WLMART', holding all other variables constant. The intercept, which is 0.394, reveals that the expected value of 'Close' is 0.394 units when all independent variables are zero, given that all other variables are held constant. These findings can offer valuable insights for further analysis and decision-making in the domain of stock market investments.

```
Co-efficents of linear model:
Open : 0.997467303553913
Volume : -2.0495972858322942e-10
Company_AAPL : -0.045033724335517486
Company_AMZN : -0.1306090379038141
Company_GOOGL : -0.13665539699037138
Company_JPMRGN : -0.10262564715478674
Company_MSOFT : 0.10055053831750548
Company_NTFLX : 0.368659923464483
Company_NVDA : -0.079711130641273
Company_TSLA : 0.3669402703276045
Company_VISA : 0.04221995808117321

 Intercept =  0.394523052381345
```

Fig 2. Model Co-efficient of Linear Model

We have predicted the close values for test dataset and calculated the Mean absolute error (MAE), Mean Squared Error (MSE), R-Squared to evaluate the model performance. The MAE train value is 2.549 and MSE train value is 36.75, MAE test value is 2.765, MSE test value is 46.93. The R-squared value is 0.9983, indicating that the linear regression model explains 99.8% of the variance in the target variable 'Close' using the selected independent variables. The plot showing the predicted Close values versus Actual Close values is as follows.

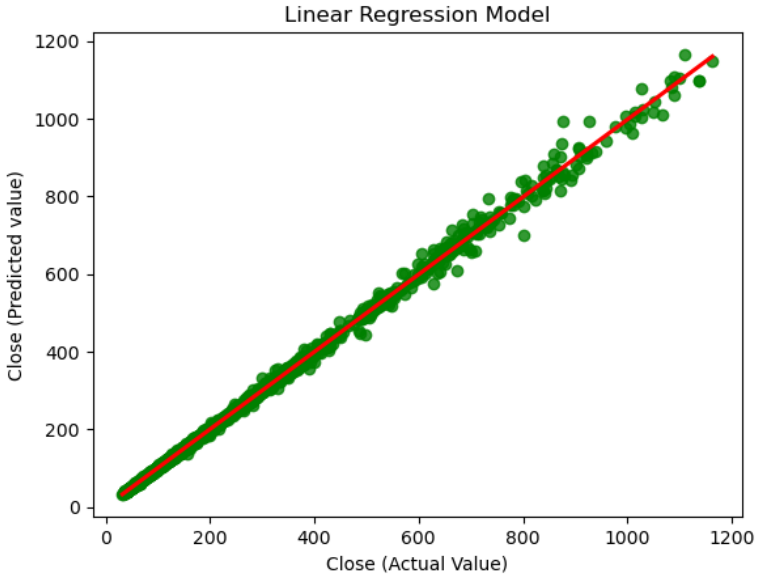

Fig 4. Predicted vs Actual Close Values plot using linear regression

In the linear Regression model, the 'Date' variable is ignored, which is also an important factor to be considered for stock predictions. So, to consider the time factor (Date) and identity factor (Company) we choose to perform Panel Data Analysis using fixed effects model.

Fixed Effects Model

Initially, a Pooled OLS model was constructed to investigate the requirements of the model. However, after analyzing the summary of the Pooled OLS model, it was observed that the model was suffering from issues such as autocorrelation between independent variables, multicollinearity, and insignificant variables, making it an unsuitable model. To address these issues and account for the fixed effects of individual variables being constant, the 'Categorical' class was utilized to represent different time periods in the 'time_period_col_name' column. The 'codes' function of the 'Categorical' class was then used to generate an array of codes for the categories in the 'Categorical' object. These codes are integers ranging from 0 to n-1, where n is the number of unique categories in the 'Categorical' object. By assigning these codes to the 'date_FE' column, a new column was created that represents the time periods using integer codes instead of the original values. In addition, dummy variables were used for the company variable. These steps were implemented to construct the Fixed-effects model.

```
Regression expression for fixed effects with dummies:

Close ~ Open + Volume + date_FE + AAPL + AMZN + GOOGL + JPMRGN + MOFT +
NTFLX + NVDA + TSLA + VISA
```

Dep. Variable:	Close	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	5.157e+05
Date:	Tue, 09 May 2023	Prob (F-statistic):	0.00
Time:	22:59:12	Log-Likelihood:	-28689.
No. Observations:	8806	AIC:	5.740e+04
Df Residuals:	8794	BIC:	5.749e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2123	0.220	0.964	0.335	-0.219	0.644
Open	0.9988	0.001	1824.685	0.000	0.998	1.000
Volume	-2.58e-09	2.56e-09	-1.008	0.314	-7.6e-09	2.44e-09
Company_AAPL	0.3097	0.412	0.751	0.453	-0.499	1.118
Company_AMZN	0.1195	0.359	0.333	0.739	-0.584	0.823
Company_GOOGL	-0.0116	0.304	-0.038	0.970	-0.608	0.585
Company_JPMRGN	0.0015	0.299	0.005	0.996	-0.584	0.587
Company_MSOFT	0.1002	0.307	0.327	0.744	-0.501	0.701
Company_NTFLX	0.2659	0.330	0.806	0.420	-0.380	0.912
Company_NVDA	-0.0124	0.315	-0.039	0.969	-0.630	0.605
Company_TSLA	0.5545	0.342	1.623	0.105	-0.115	1.224
Company_VISA	-0.0842	0.299	-0.281	0.778	-0.671	0.502

Omnibus:	6622.730	Durbin-Watson:	1.976
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2788367.620
Skew:	-2.464	Prob(JB):	0.00
Kurtosis:	90.035	Cond. No.	6.69e+08

Fig 5. Summary of Fixed-Effects Model

The R-squared value is 0.998, indicating that the linear regression model explains 99.8% of the variance in the target variable 'Close' using the selected independent variables. Performed VIF test to check the multicollinearity. After considering the tolerance calculated from this test, it is inferred that model is not suffering with multicollinearity. The model is a good model as it addresses the issues in OLS model and predicts the Close values accurately. But considering the other unobserved factors, the model is not the best model in the real time stock market scenario.

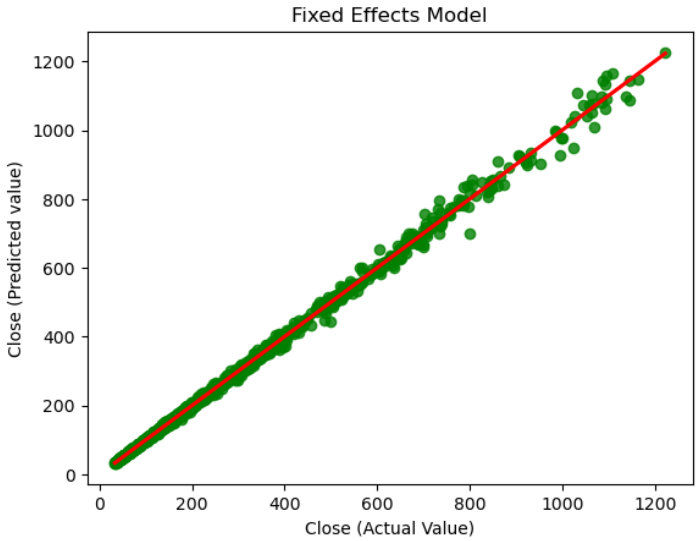

Fig 6. Predicted vs Actual Close Values plot using Fixed Effects Model

LSTM (Linear Short-Term Memory) Model

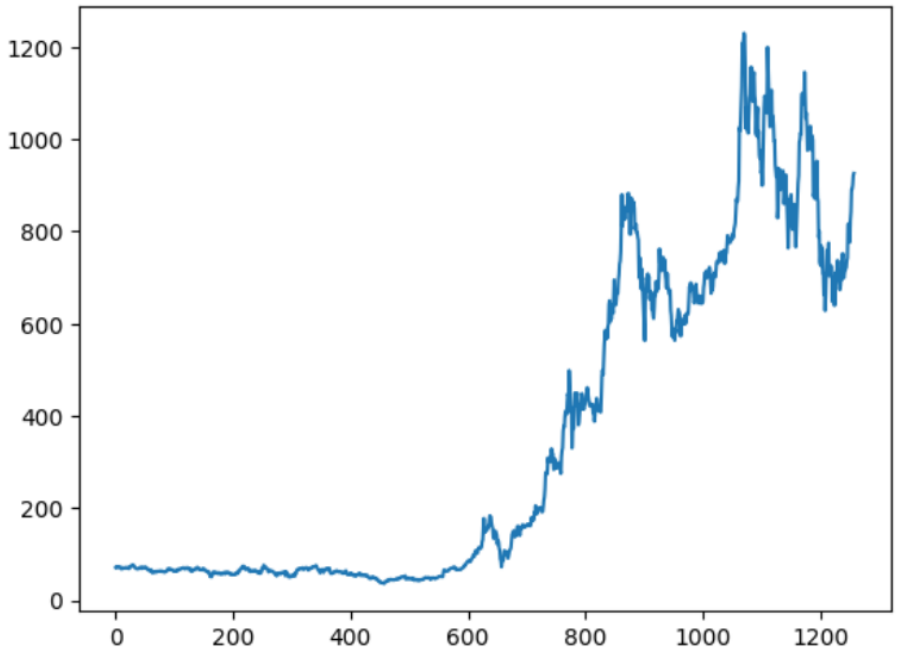

Fig 7. Tesla's stock price from 2017 to 2022

The chart shows that Tesla's stock price has been on a steady upward trend since 2017. The stock price reached a peak of $314.67 on August 16, 2022, but has since fallen back to around $170 per share. There are several factors that could be contributing to the recent decline in Tesla's stock price. These factors include:

- Rising interest rates, which make it more expensive for companies to borrow money.
- Supply chain disruptions, which have made it difficult for Tesla to produce and deliver cars.
- Increased competition from other electric car makers, such as Rivian and Lucid.

Despite the recent decline in its stock price, Tesla is still a valuable company. The company has a strong brand, a loyal customer base, and a leading position in the electric car market. Tesla is also well-positioned to benefit from the long-term growth of the electric car market.

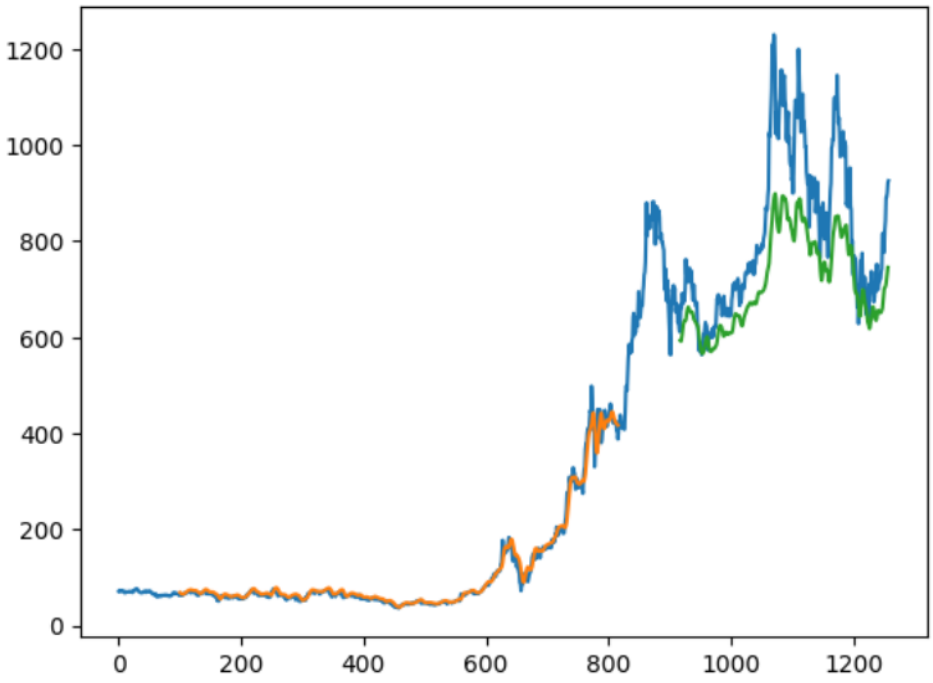

Fig 8. Train and Test MSE plot for Tesla stock

The above figure depicts the issue of overfitting. Overfitting is a problem that can occur when a model is trained on a dataset that is too small or too similar. When a model is overfit, it learns the specific details of the training data too well, and it is not able to generalize to new data.

In this case of Tesla's stock price, overfitting has occurred since the model is trained on a dataset that only includes data from the past few years. This is because Tesla's stock price has been volatile in recent months, and the model might learn the volatility too well. As a result, the model might not be able to accurately predict Tesla's stock price in the future.

To avoid overfitting, we have performed ridge regression, also it is important to train a model on a dataset that is large and diverse. This will help the model to learn the general trends of the data, and it will not be as likely to learn the specific details of the data.

Ridge Regression

Ridge regression is a type of linear regression that is used to overcome the problem of overfitting in a model. When there are many independent variables in a dataset, it is common to observe that some of the variables have a high correlation with each other. This can lead to a problem known as multicollinearity, where the model coefficients become unstable and difficult to interpret.

```
Co-efficents:
Open : 0.9992224700060124
Volume : -1.1879053628511405e-09
Company_AAPL : 0.05162525455325185
Company_AMZN : -0.019971726341392666
Company_GOOGL : -0.0391559117335694
Company_JPMRGN : -0.029125989918073275
Company_MSOFT : -0.0038468871811969865
Company_NTFLX : 0.051952907080722414
Company_NVDA : -0.05472974351161087
Company_TSLA : 0.16812659610180608
Company_VISA : -0.07586574816419199

 Intercept =  0.20924229024890906
```

Fig 9. Coefficients of Ride Regression Model

Ridge regression solves this problem by adding a regularization term to the cost function of the linear regression model. This regularization term adds a penalty to the model coefficients, which helps to reduce their values and thus reduce the impact of multicollinearity. We have observed that the linear regression model is overfitting for the above-mentioned tesla stock price analysis, in order to reduce the impact of multicollinearity and obtain a more stable and reliable model we have performed ridge regression.

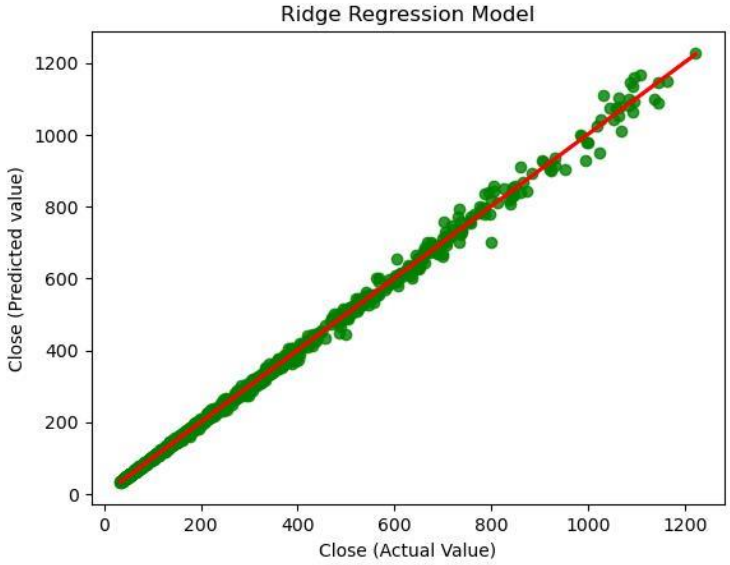

Fig 10. Predicted vs Actual Close Value plot.

Furthermore, while performing Ridge Regression we tend to identify correlations between independent and dependent variables, they may not be able to capture complex nonlinear relationships or account for unforeseen events. Therefore, it is important to use caution when interpreting the results of regression models and to supplement them with other forms of analysis, such as fundamental analysis and technical analysis.

Results

The predicted Close value over the period (2017-2022) is plotted as:

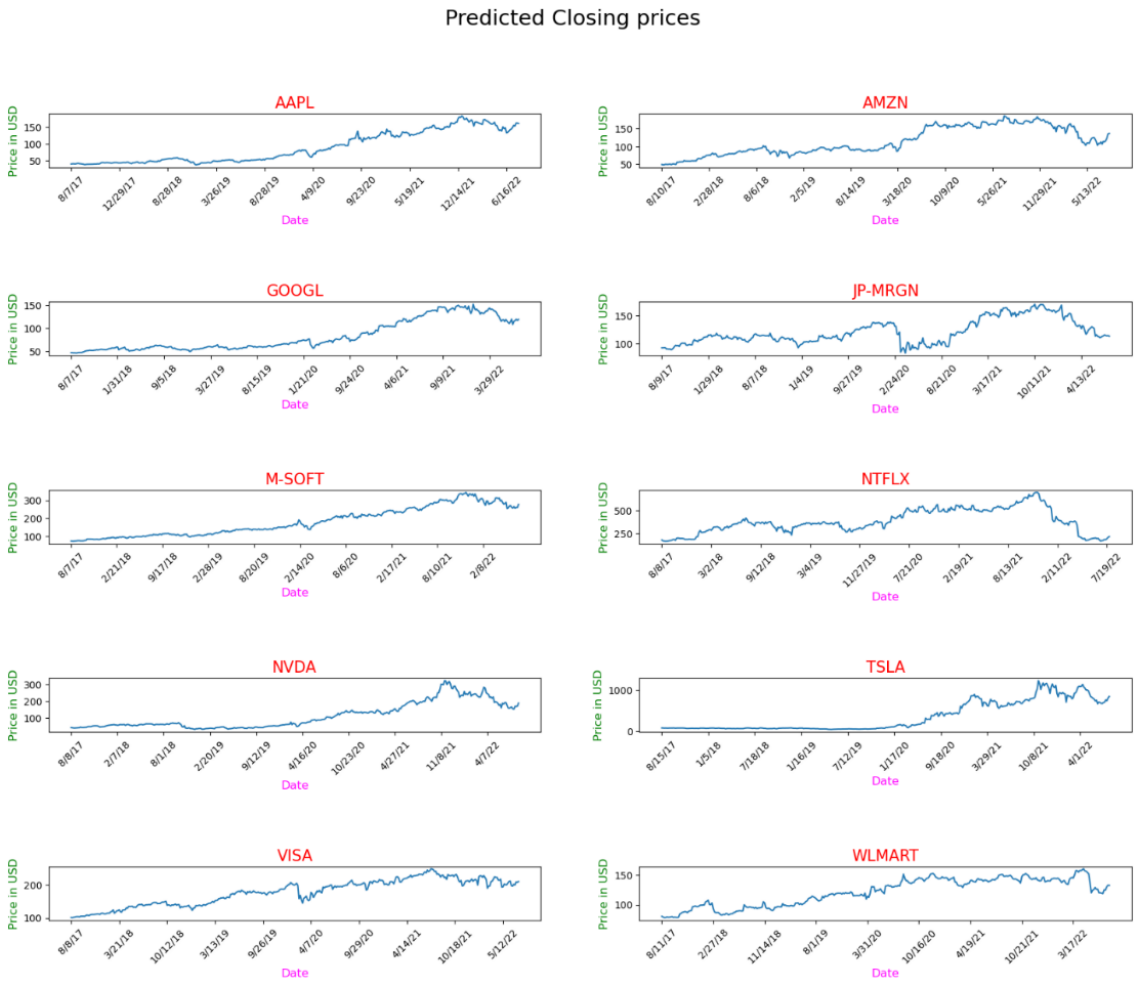

Fig 7. Predicted Close Values plot using Fixed Effects Model

From the above figure we conclude that the predicted stock values from 2017-2022 for top 10 companies are the same as the open values for the stocks. Though this model shows a high accuracy, model cannot be used for real time predictions because of many other influential factors such as economic indicators, global events, and company-specific news, which can make it difficult to accurately predict price movements. Additionally, stock market data is often noisy and unpredictable, making it a challenging problem for machine learning models. Nonetheless, regression models can be a useful tool in analyzing historical stock market data and identifying trends and patterns.

References

[1] https://www.kaggle.com/datasets/mdwaquarazam/stock-price-history-top-10-companies

[2] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplots.html

[3] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[4] The Fixed Effects Regression Model For Panel Data Sets – Time Series Analysis, Regression, and Forecasting (timeseriesreasoning.com)

[5] https://youtu.be/H6du_pfuznE