

---



# Learning From Data

---

Vedant Adawadkar - 730042744 - [va296@exeter.ac.uk](mailto:va296@exeter.ac.uk)

---

---

# Contents

- 01. The Why?
- 02. Introduction To The Dataset
- 03. Using Linear Regression To Predict Stock Prices
- 04. Which Machine Learning Models Were Used
- 05. An Overview On Data Analysis And The Results
- 06. Reflections
- 07. Limitations

---

# The Why?

In financial markets, predicting stock prices is a complex and very challenging task.

The constantly changing and volatile nature of stock markets makes it extremely hard to predict, mainly due to various unpredictable factors.

To understand the complexity of predicting current or future stock prices, in this coursework, we will analyse if a supervised machine learning technique, specifically regression, can help predict stock prices.

---



# Introduction To The Dataset

To study and perform our analysis, I have chosen a dataset from Kaggle that consists of data from the New York Stock Exchange.

The data consists of about 500 companies, although we will only look at three companies for this coursework while making our predictions.

The data spans from 2010 to 2016 or sometimes from 2012 to 2016.

The dataset is divided into three parts - prices, split-adjusted prices, and fundamentals.

---



# Data Pre-processing And Data Cleaning

Given that we need to use all three datasets provided, we need to perform some pre-processing and cleaning on the dataset before doing any operations.

Some of the operations included while preparing the data include:

- Dataset Merging
- Column Data Type Conversion.
- Handling NaN values.

---




# Using Linear Regression To Predict Stock Prices

For this coursework, I have chosen to use regression with linear regression specifically to try to predict the closing value of the stock for that day.

Using Linear Regression to predict the stock prices for the current day or future days is a relatively common approach in quantitative finance.

The reasoning behind Linear Regression is that regression provides a 'line of best fit' that minimizes the distance between actual and predicted scores.

---



# Feature Selection

One of the most essential things before employing machine learning models to train or test our dataset is to select different features we need to make our predictions.

The range of accuracy largely depends on these features.

In the case of linear regression, we need to set our target variable or the dependent variable and the independent variables or the predictors correctly.

For this research, we use the 'close' variable as the target variable, which will give the stock's closing price.

---



# Which Machine Learning Models Were Used

A supervised machine learning technique, specifically the Linear Regression Model, was used to predict the stock prices.

Linear Regression can help model the relationship between variables using linear equations.

Along with this model, we use a feature selection technique called Recursive Feature Elimination (RFE) to update the feature list by recursively removing the minor essential features.

Using RFE with linear Regression helps improve model performance and reduce model overfitting by only working with the most essential features.



---

# An Overview Of Data Analysis And Results

Three companies from the same sector were chosen: Microsoft, Electronic Arts, and Activision Blizzard. These companies are represented as securities, abbreviated as follows:

- MSFT - Microsoft
- EA - Electronic Arts
- Activision Blizzard - ATVI

Ten predictors that are known to influence the market are selected: open, low, high, volume, after-tax ROE, gross profit, gross margin, profit margin, total revenue, and earnings per share.

---

Three companies from the same sector were chosen: Microsoft, Electronic Arts, and Activision Blizzard. These companies are represented as securities, abbreviated as follows:

- MSFT - Microsoft
- EA - Electronic Arts
- Activision Blizzard - ATVI

Ten predictors that are known to influence the market are selected: open, low, high, volume, after-tax ROE, gross profit, gross margin, profit margin, total revenue, and earnings per share.

Because the number of predictors is too high, we use the RFE technique to ensure that only the most important features are selected.

In the case of all three companies, the following predictors are selected:

- Open
- Low
- High

---

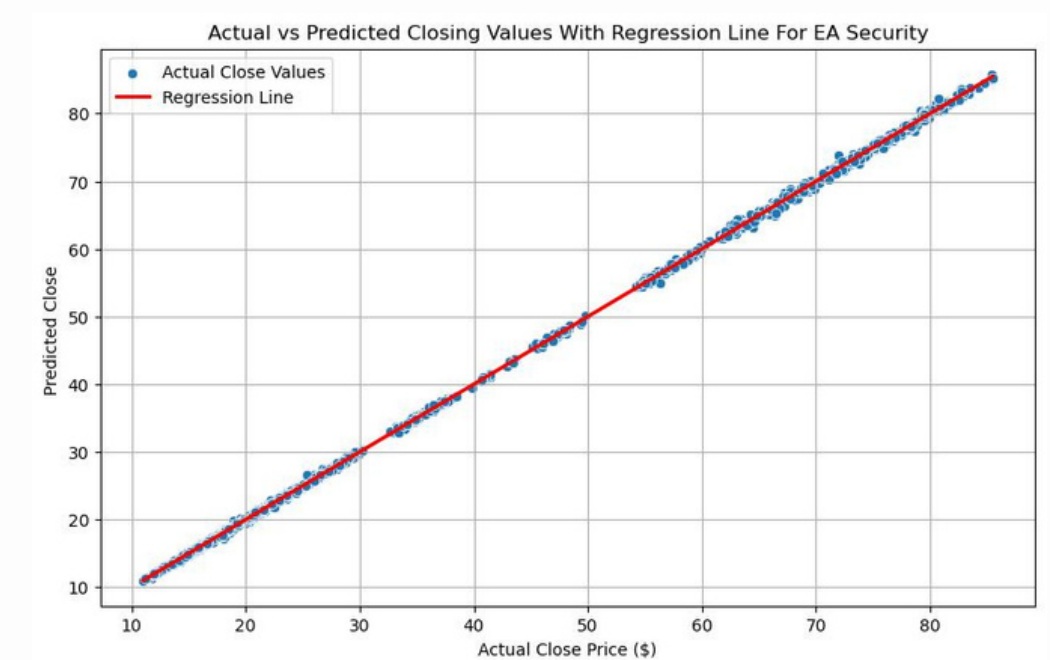
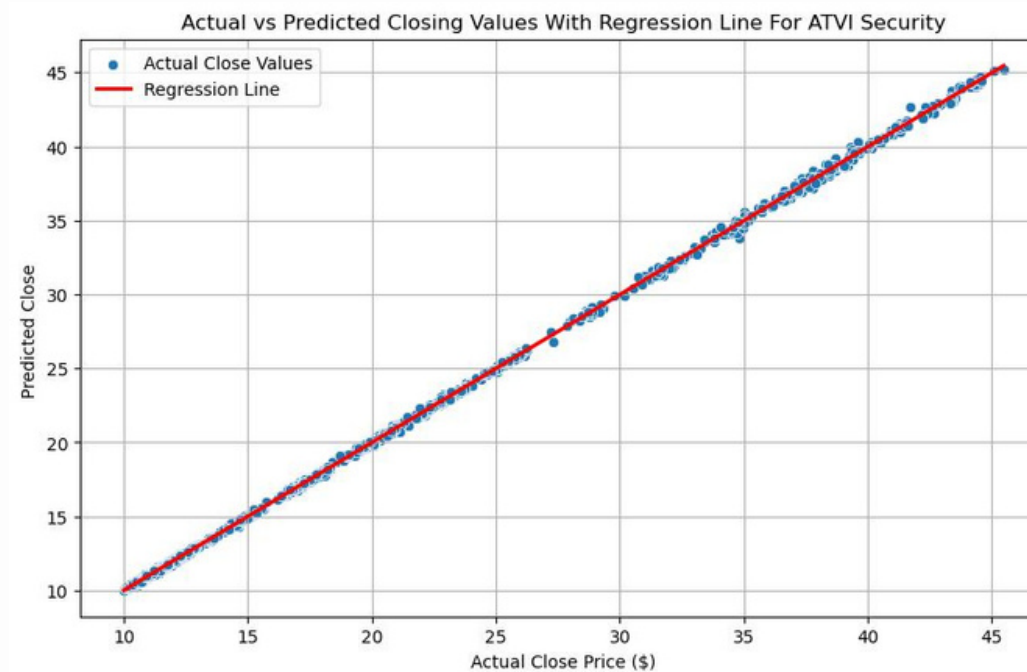
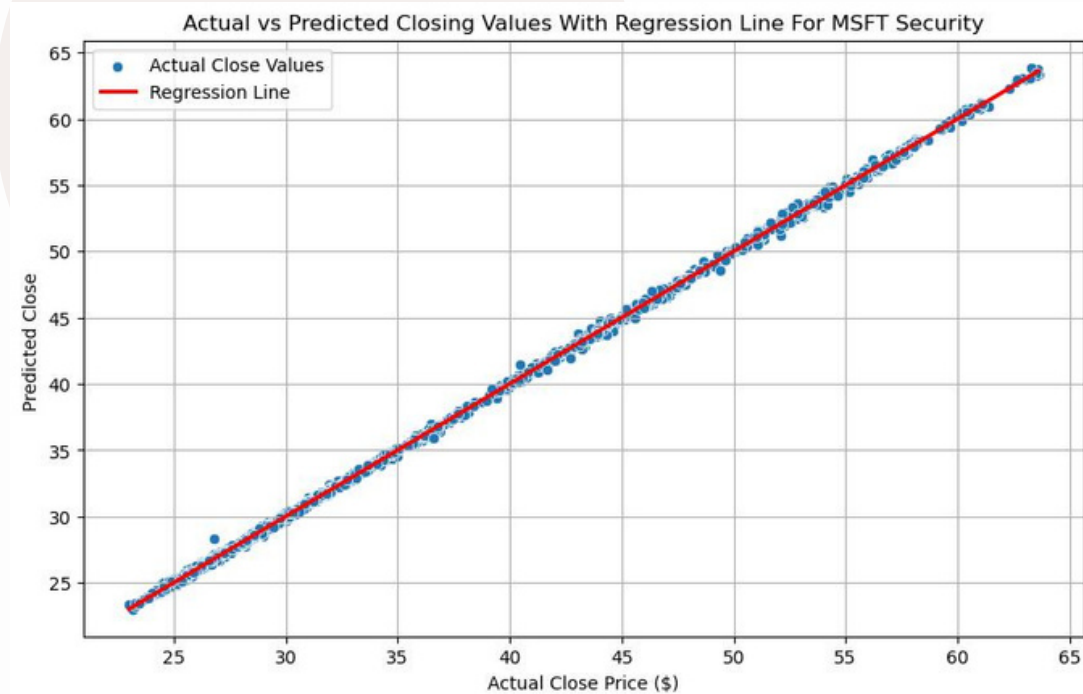
We then run the training model using the training data and then predict the 'close' values for the stock by running the test data on the trained model.

To evaluate the model's performance, we use Mean Absolute Error (MAE) and compare the values with the training and testing set. We also use the Root Mean Squared Error (RMSE) to compare the results with the training and the testing set.

Further, we also draw our results based on the R-squared score to understand the 'goodness-of-fit' measure among the independent and dependent variables.

Further visualisations are also done to make the regression analysis more interpretable, as seen on the subsequent slides.

# Actual vs Predicted Close



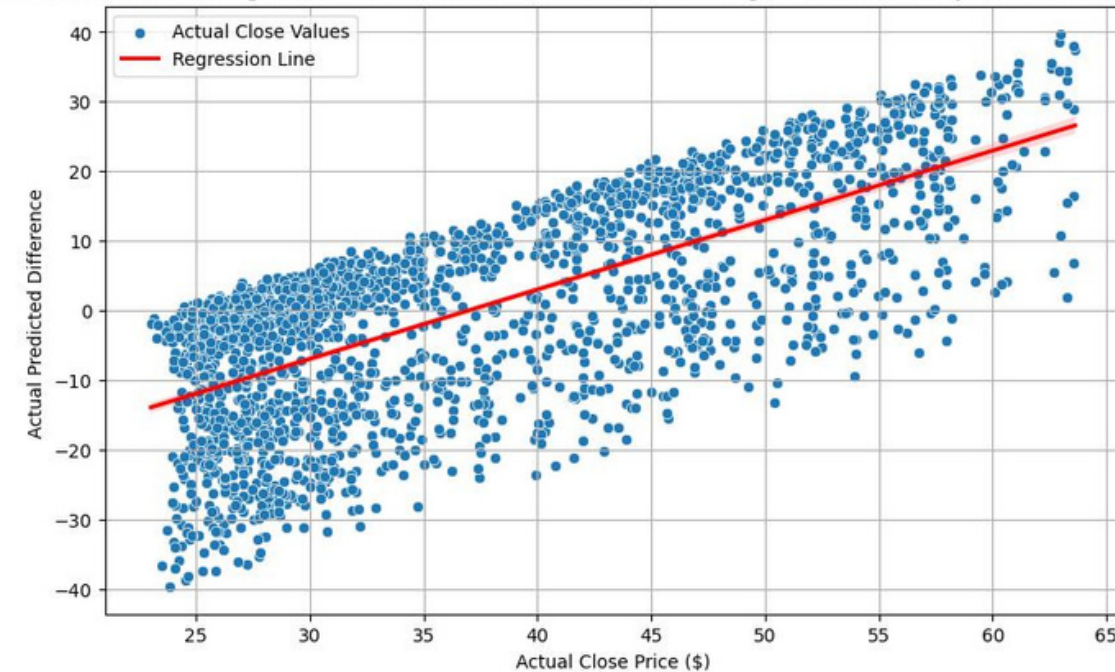
The figures above indicate a strong correlation between the variables, given how tightly different data points are closely grouped near the regression line.

The regression model can capture the trends in the data, making it a good fit.

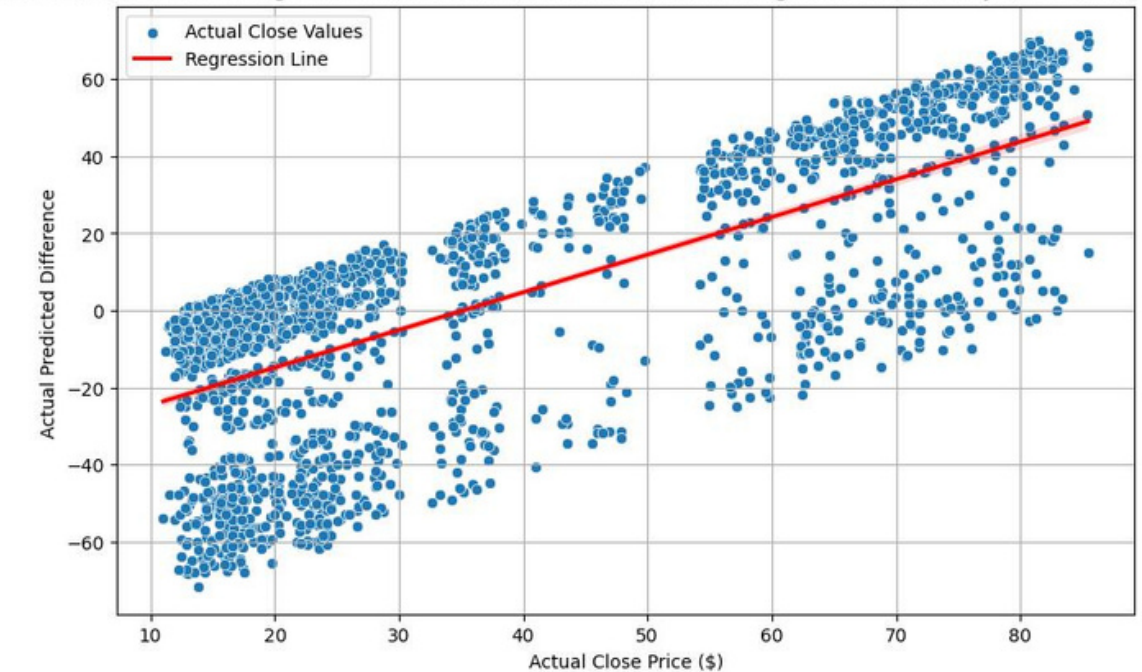
---

# Difference between Predicted Close and Actual Close of Next Day

Linear Regression Scatter Plot Showing Actual Close And Difference Between Closing Value Of Next Day and Predicted Close For MSFT Security



Linear Regression Scatter Plot Showing Actual Close And Difference Between Closing Value Of Next Day and Predicted Close For EA Security



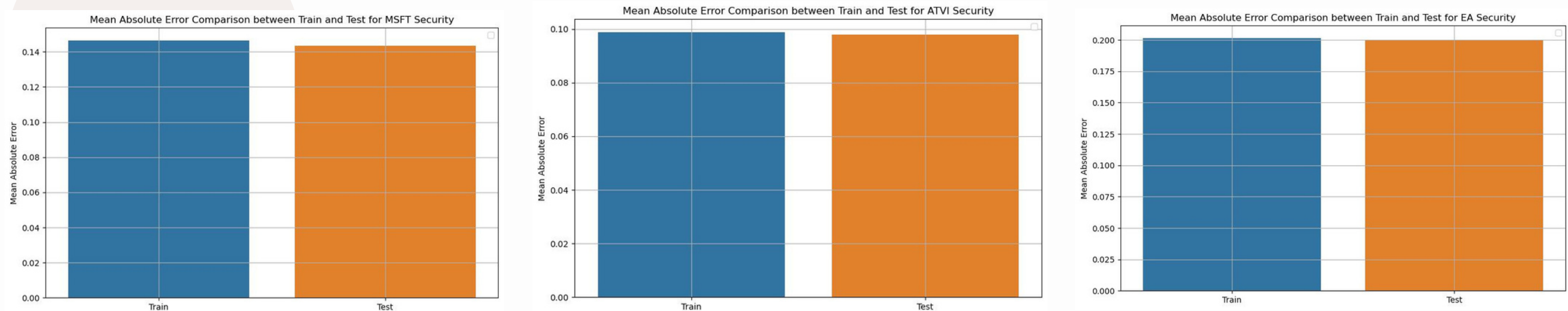
The figures above indicate that the model falters with points all over the plot when we try to see the difference between the predicted close for the day and the next day's actual close.

To capture the trends for the next day, we need to modify our approach so that the next day's predictions are more accurate in their results.



---

# Mean Absolute Error (MAE)

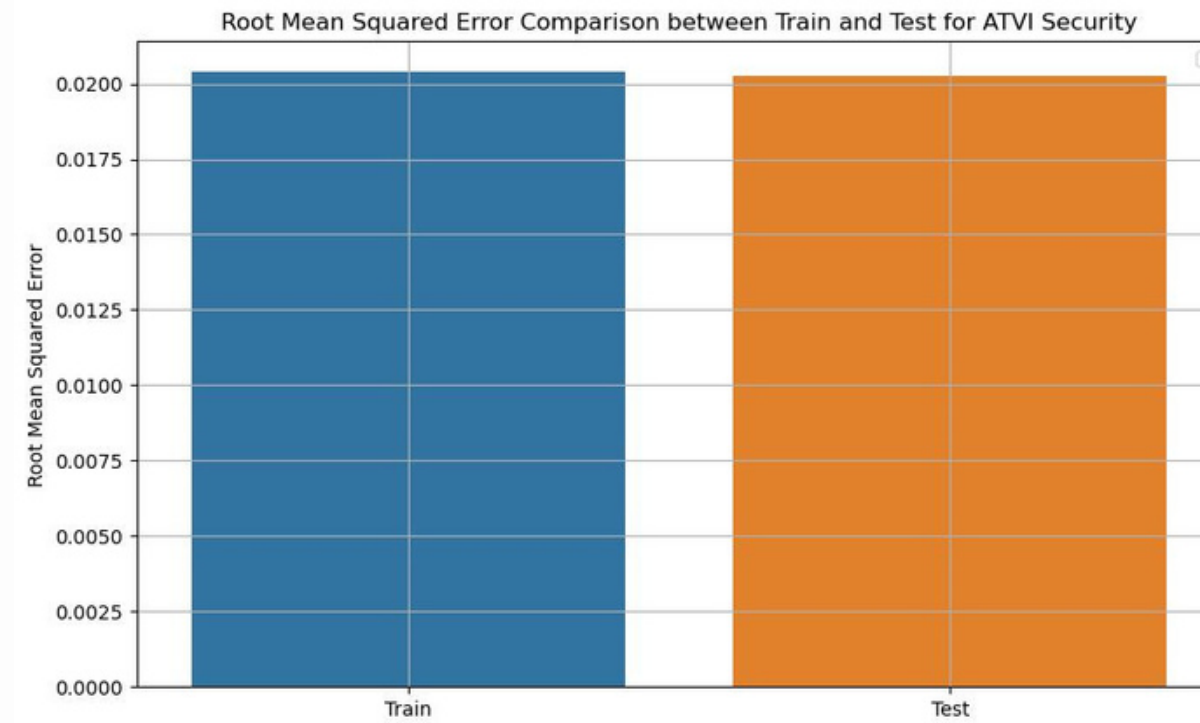
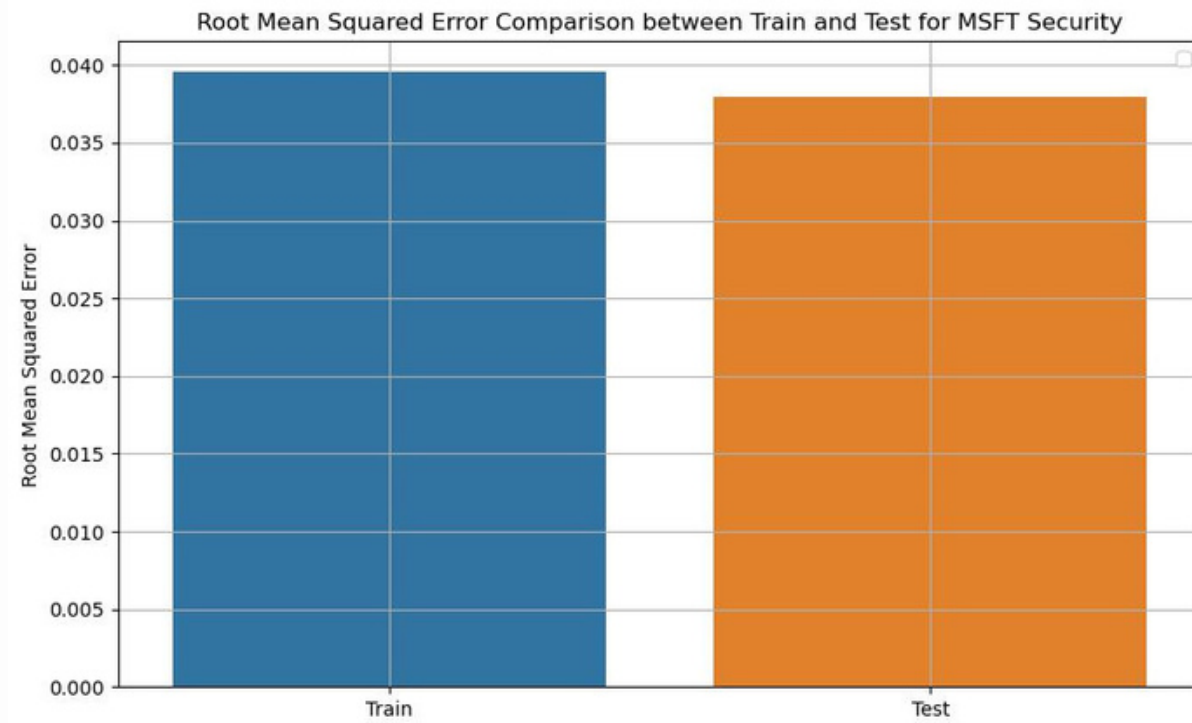


The Mean Absolute Error for all three chosen securities is well below one, which is positive, suggesting relatively low errors. Still, given how the training set is higher than the test set in all three securities, it indicates that the model is overfitting.

These values might result in a model that needs to be generalised and may lead to poor performance overall.

---

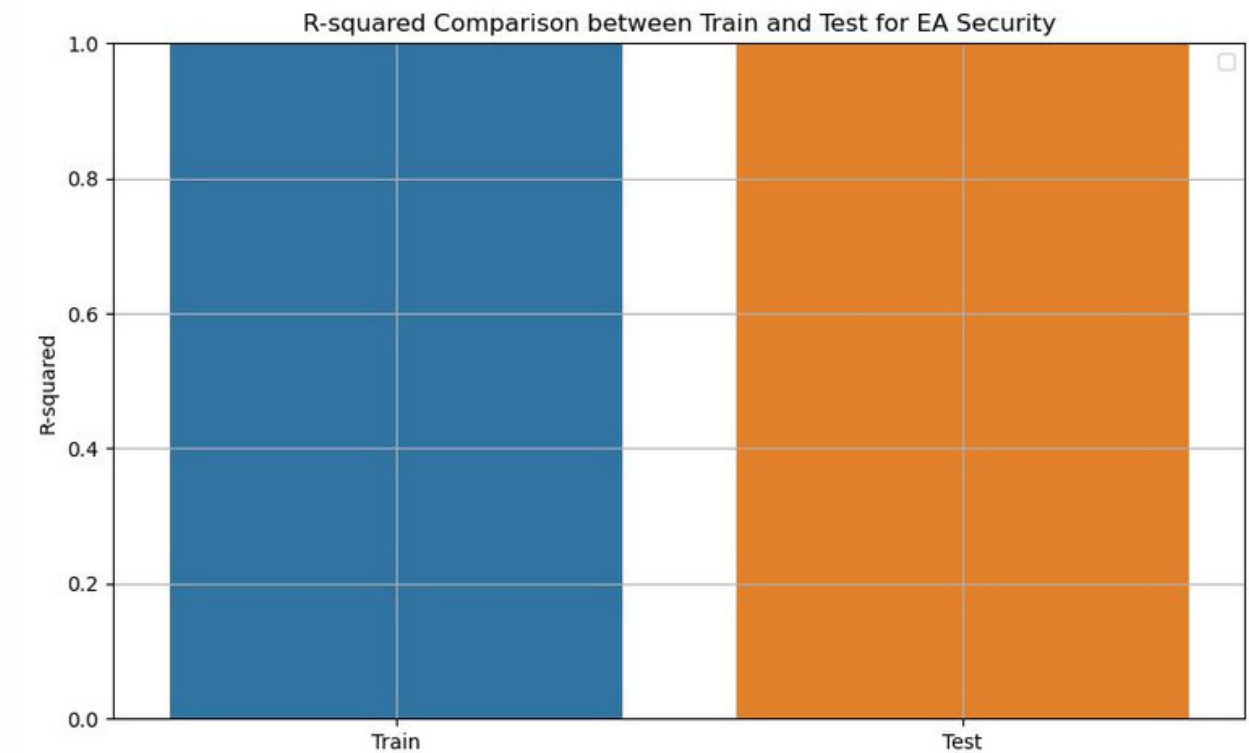
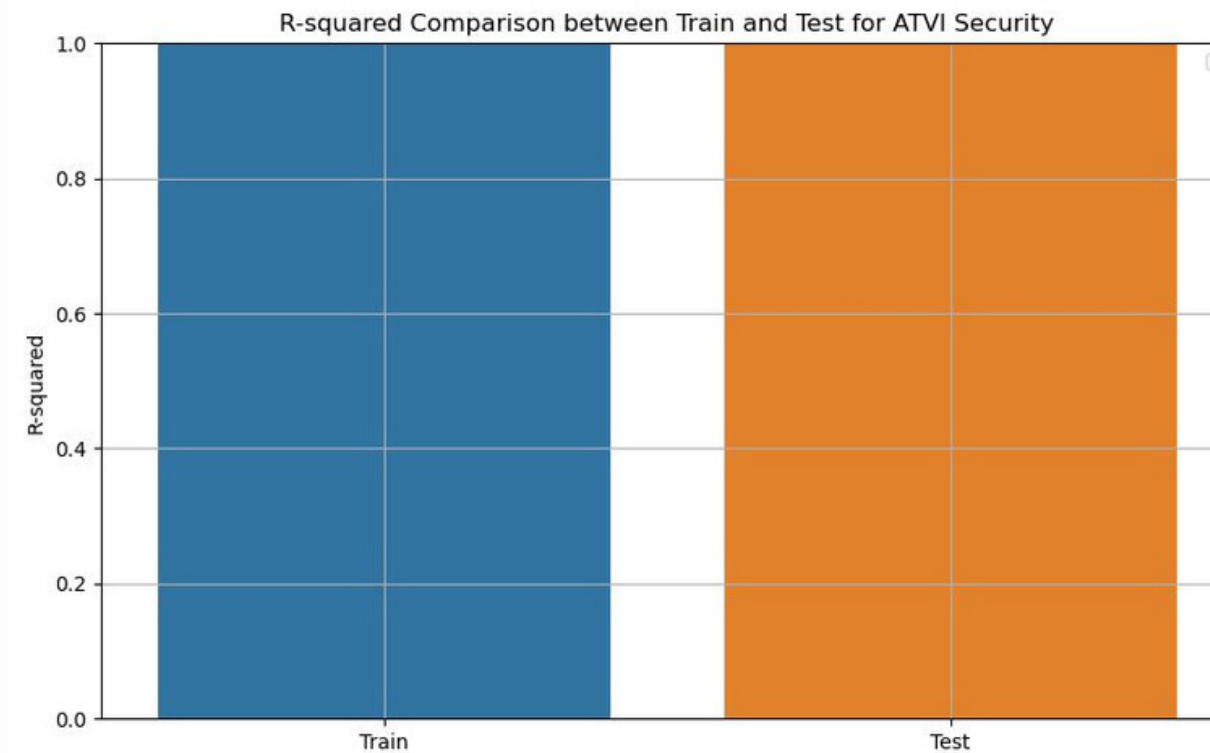
# Root Mean Squared Error (RMSE)



The root mean squared error also provides similar results to MAE in the case of MSFT Security and EA security. Interestingly, ATVI security has much better results where the model only fits the data a little.

---

# R-Squared Score



Given that the R-squared score is relatively high, reaching closer to values of 1. However, it indicates that the model is performing well; given the scores obtained in MAE and RMSE, we conclude that the model is overfitting.

The R-squared score means the model captures noise or outliers too well, so it might need to generalise real-world data better.



---

# Reflections

To correctly analyse a stock market dataset, it is essential first to understand the different features of the dataset and how one feature can affect the other.

Even though the model produces favourable results at first glance, the quantitative analysis suggests that the model is overfitting, i.e., there needs to be more optimisation to make the predictions more accurate.

It is essential to keep training and testing the model with different feature sets to improve interpretability.

We have to analyse the results using various quantitative means available, such as Mean Absolute Error, Root Mean Squared Error, and R-squared scores, to justify that we can rely on our results.

---

# Limitations

One of the most critical limitations of using the chosen dataset is that the dataset needs to consist of up-to-date data.

Further, the dataset consists of different features where the value of one feature relies on the other, making it challenging to perform pre-processing and cleaning of the dataset.

Although effective, linear regression is still not a great way to predict and analyze a complex topic like the stock market, which is already notoriously hard to predict.

As seen in the previous slides, we can conclude that without techniques such as cross-validation, regularisation, or hyperparameter tuning, simple ML models are prone to underfitting or, in this case, overfitting.

We should employ other machine learning models like Random Forests, Recurrent Neural Networks, or LSTMs.



Thank You!

---