MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Mini Project Report of

Introduction to Data Analytics (CSE 2126)

STOCK PRICE PREDICTION

SUBMITTED
BY

Reeva Nanda, 220962310 ,Roll no. 48, AIML 'A'
Dev Vasudevan,220962,Roll no 60,AIML 'A'

Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal.
Oct 2023

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Manipal
00/00/2023

CERTIFICATE

This is to certify that the project titled **MiniProject Title** is a record of the bonafide work done by **Student(s) (Reg. No. 000000)** submitted in partial fulfilment of the requirements of **Introduction to Data Analytics (CSE 2126)** course of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Institute of Manipal Academy of Higher Education), during the academic year 2023-2024.

Name and Signature of Examiner:

Dr. Roopalakshmi R,
AssociateProfessor,
CSE Dept.

TABLE OF CONTENTS

INTRODUCTION

This report encapsulates the development and evaluation of a predictive model for stock closing prices, leveraging Linear Regression on historical market data. The objective is to forecast closing prices based on pertinent financial indicators, namely 'Open', 'High', 'Low', and 'Volume'. The dataset, obtained from Kaggle, undergoes meticulous preprocessing to ensure data reliability.

Our approach involves Evaluation metrics, including the model score and mean squared error, serve as benchmarks for reliability and accuracy. Complementing these metrics, Matplotlib facilitates visual analysis, presenting a graphical juxtaposition of actual and predicted close prices.

At the heart of our analysis lies Linear Regression, a fundamental statistical tool modeling the relationship between selected indicators and closing prices. The report succinctly introduces the core components of Linear Regression, emphasizing its role in predicting stock prices through minimizing the difference between predicted and actual values.

In essence, this report navigates through a streamlined workflow—from data preprocessing to model evaluation—aiming to deliver a concise yet comprehensive understanding of our predictive model for stock closing prices.

PROBLEM STATEMENT AND OBJECTIVES

The aim is to develop a predictive model that accurately forecasts the closing prices of stocks based on historical data.
The main objectives behind this model are :

- To use the Pandas library to structure and preprocess the data
- To implement Linear Regression , a machine learning algorithm , in order to build a predictive model
- To utilize the Open, Volume, High and Low values to forecast a closing price
- To assess the reliability and accuracy of the model
- To plot the predicted values against the actual values

METHODOLOGY

1.Dataset

We have used a dataset from Kaggle that has 7 columns , for Date , Open , High, Low , Close, Volume and OpenInt.
We uploaded the csv file onto an excel sheet and using the Pandas library , we have read the excel file in our program .
We need to use only Open , High,Low and Volume columns as a basis to predict our

2.Data Cleaning

Since our dataset is of stock prices , it is already free of garbage or NaN values , but to ensure that no noisy data interrupts our prediction making , we have used the dropna() function to remove any rows which may have null values.

The OpenInt values are not consistent hence we have dropped those columns
We remove the outliers in Volume by calculating the interquartile range from 25%ile to 75%ile and then creating upper and lower bounds. Volumes that lie outside those bounds are considered outliers and help us streamline our data

3. Predictive Analysis

We trained our model on the first stock file in our dataset. After testing it on the same file we have then used this model to test 2 other files in our dataset .
The model.score() method indicates the percentage of variability in the test data that is captured or explained by the linear regression model. For instance, if the resulting score is 80%, it implies that 80% of the variability in the actual closing prices can be explained by the features included in the model (i.e., 'Open', 'High', 'Low', 'Volume'). This score provides an understanding of how well the model fits the test data, with higher scores indicating better predictive performance.

After predicting close price values , we have calculated the mean squared error to assess the accuracy of the prediction .This function computes the mean of the squared differences between the actual values (y) and the predicted values (y_pred). It measures the average squared difference between the predicted

values and the actual values. Mathematically, it is calculated as the sum of squared differences divided by the number of samples.

4.Data Visualisation using Matplotlib

We have plotted a graph of actual close price vs predicted closing price to provide a visual analysis of the performance of the model . This has been carried out using the Matplotlib library.

5.Linear Regression

Linear regression is a fundamental statistical method used to understand and predict the relationship between two or more variables. In the context of our analysis for predicting stock prices, linear regression serves as a valuable tool to model the connection between specific stock market indicators, like 'Open', 'High', 'Low', and 'Volume', and the closing price of a stock.

Key Components:

- Dependent Variable (Target): In our case, the closing price of a stock is the variable we aim to predict or explain based on other factors.
- Independent Variables (Predictors): These are the factors, such as the opening price, highest and lowest prices reached, and trading volume, that we believe have an influence on the closing price.

The "linear" aspect of linear regression refers to its assumption that the relationship between the dependent and independent variables can be approximated by a straight line. The model attempts to find the best-fitting line through the data points that minimizes the difference between predicted and actual values.

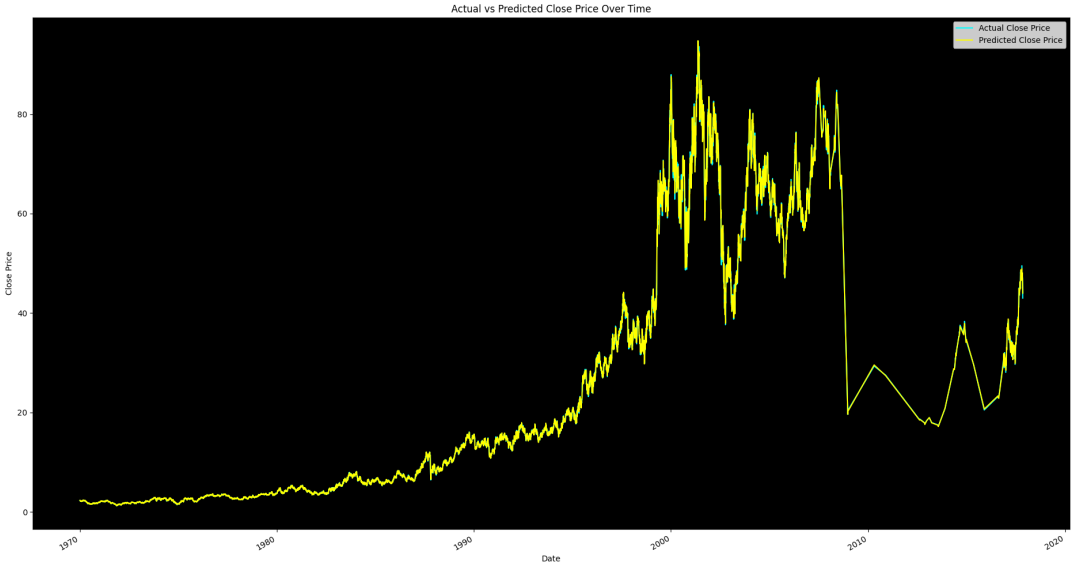
RESULTS AND SNAPSHOTS

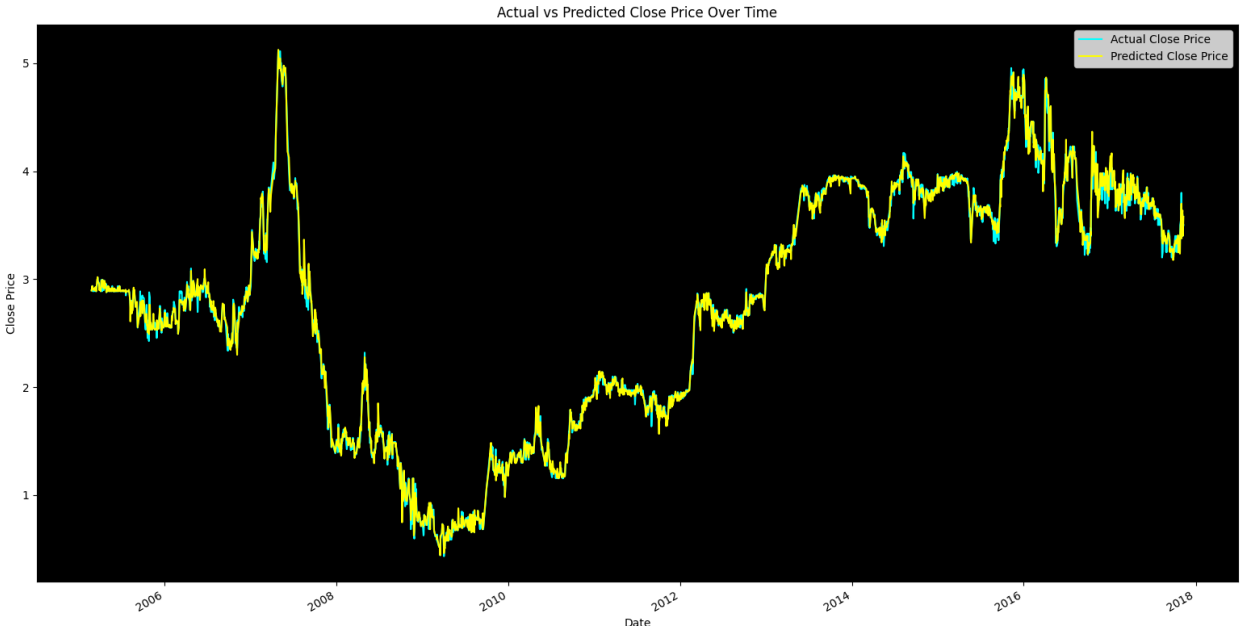
We have trained our model on data of 12,000 days and then tested it against 3 different scripts.

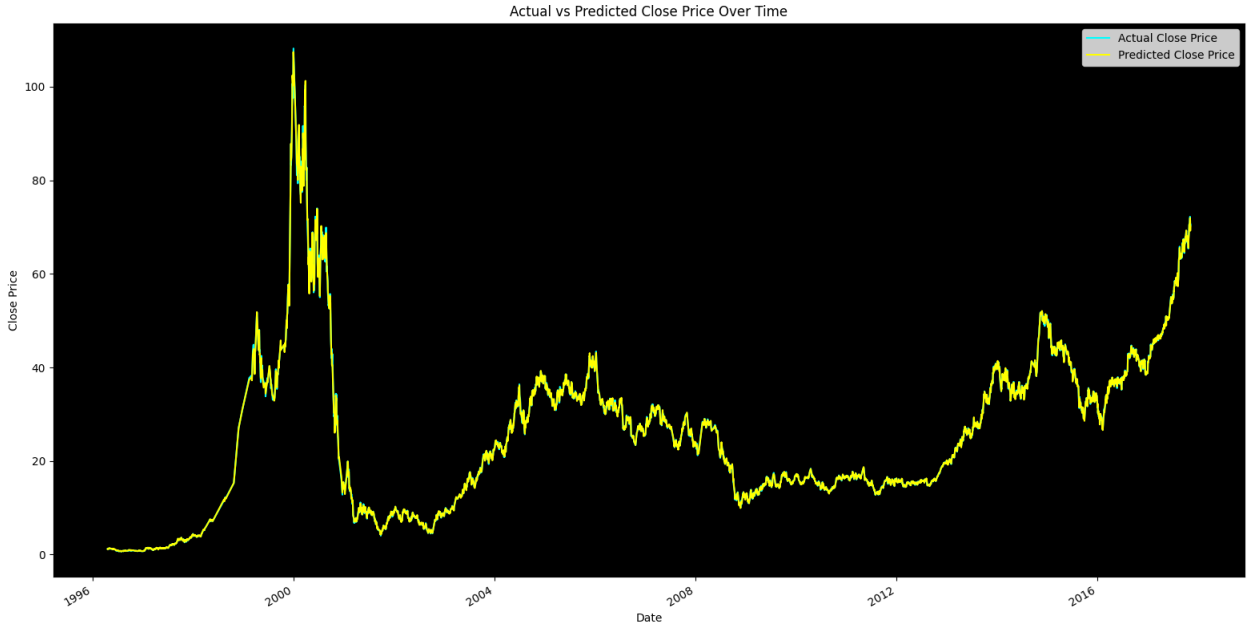
The code prints out the mean volume and we use the volume data to remove outliers. The models score and the Mean Squared Error is also printed out

```
Mean volume is 3337756.800894484
Model Score is 99.9881858101789
Mean squared error:0.07022240042668601

Mean volume is 6715.7614490772385
Model Score is 99.79752486399471
Mean squared error:0.0022907420322791066

Mean volume is 23217957.056496136
Model Score is 99.9655254779818
Mean squared error:0.10309301349785062
```

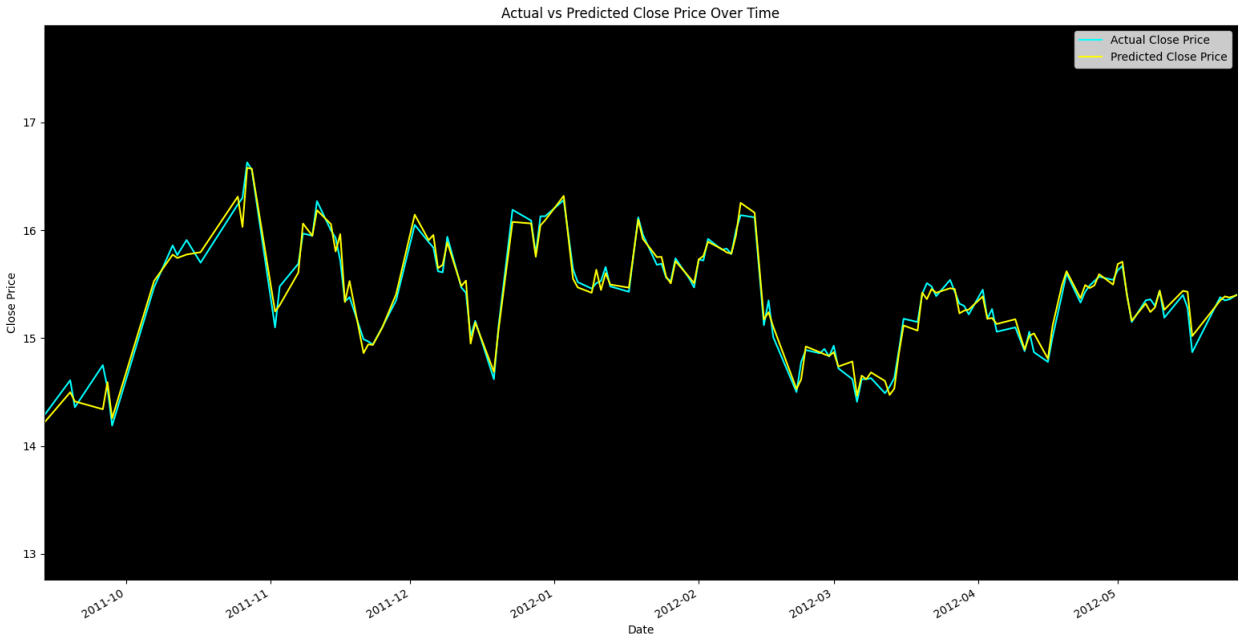
The estimated close price and the actual close price is plotted against the date.







On zooming in to a small time frame we see the 2 lines distinctly



CONCLUSION

In conclusion, this report has delved into the development and evaluation of a predictive model for stock closing prices, employing Linear Regression on historical market data. The journey encompassed key financial indicators, dataset preprocessing, and a meticulous train-test split to ensure a robust model. Evaluation metrics and visual analysis provided nuanced insights into the model's reliability and accuracy.

The significance of Linear Regression in modeling the relationship between selected indicators and closing prices has been highlighted, emphasizing its role in minimizing the disparity between predicted and actual values. The streamlined workflow, from data preprocessing to model evaluation, underscores our commitment to delivering a succinct and effective predictive model.

As the report concludes, the focus remains on the model's practicality and potential for real-world applications in forecasting stock closing prices. By combining quantitative assessments with visual representations, we aim to provide stakeholders with a comprehensive and insightful perspective on the model's performance and its implications in the realm of financial forecasting.

LIMITATIONS AND FUTURE WORK

Linear regression assumes a linear relationship between variables, which might not always hold true in complex real-world scenarios. It's also important to note that external factors not captured in the data can influence stock prices, impacting the model's accuracy.
Also,our model currently just predicts the closing price for a stock given its open ,high low and volume. In the future we would like to look into Recurrent Neural Networks to enhance the performance and usability of our model.

REFERENCES

https://www.geeksforgeeks.org/

https://www.youtube.com/

https://www.kaggle.com/