Open Ended Lab Report

(*StockEzy*: Google Stock Prediction Web Application**)**

Group: G-2

Members:

Muhammad Ashar (CS-21062)
Habib Ullah (CS-21075)

Semester: 6th (Spring Semester)

Batch: 2021

Year: 3rd Year (T.E.)

Course Tite: Machine Learning
Course Code: CS-324

Submitted to: Miss Mahnoor Malik

Date: 07 July 2024

Predictive Modeling on Google Stock Data:

1. Introduction:

This project aims to explore and predict Google stock prices using historical data. The dataset contains daily stock prices and volume for Google from 2004 to 2023. The key steps involved in the project are data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model building, and evaluation.

2. Data Collection

The dataset is obtained from a reputable source, Kaggle, and contains the following columns:

- **Date**: The date of the stock price record.
- **Open**: Opening price of the stock.
- **High**: Highest price of the stock during the day.
- **Low**: Lowest price of the stock during the day.
- **Close**: Closing price of the stock.
- **Adj Close**: Adjusted closing price of the stock.
- **Volume**: Number of shares traded during the day.

The dataset includes 4858 entries, covering a substantial period that provides a comprehensive view of the stock's performance over time.

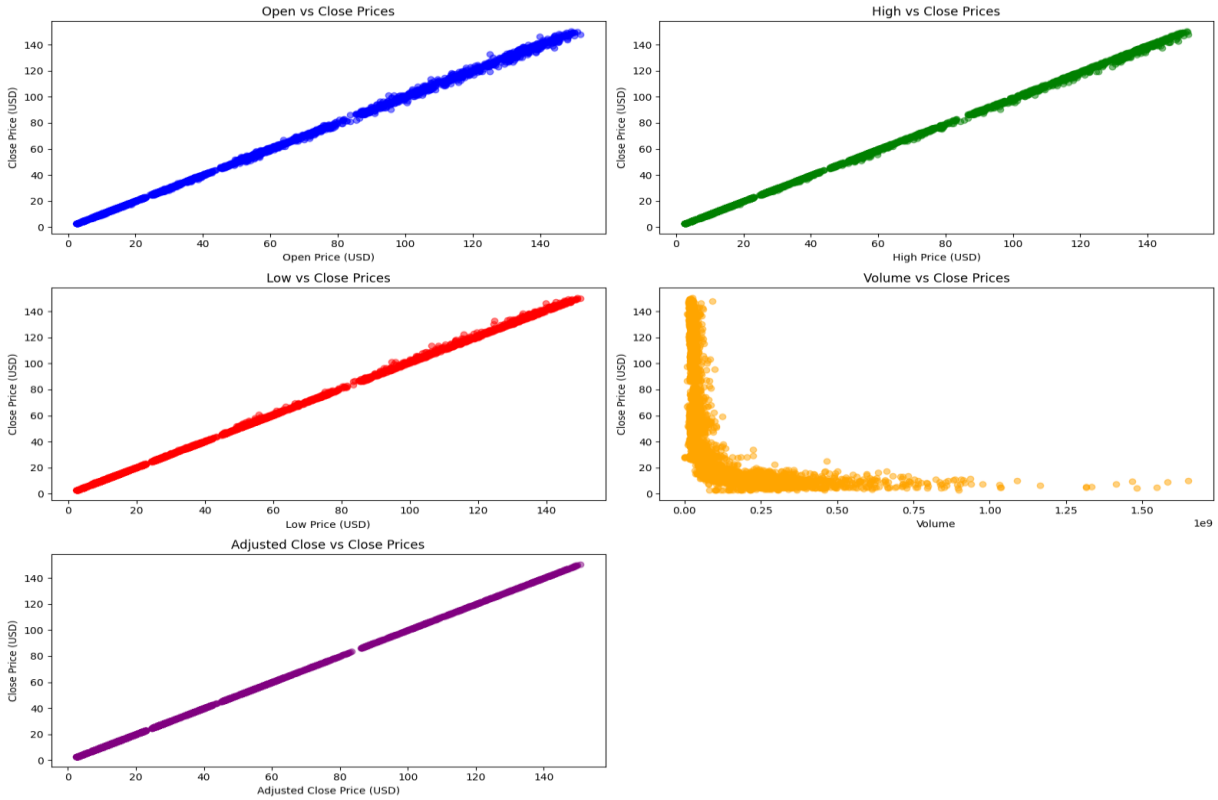
3. Data Preprocessing:

- **Missing Values**: No missing values were found in the dataset, ensuring data completeness.
- **Duplicates**: No duplicate entries were found, maintaining data integrity.
- **Date Conversion**: The Date column was converted to date-time format to facilitate time-series analysis.
- **Data Encoding**: No data encoding was needed as there are no categorical variables.

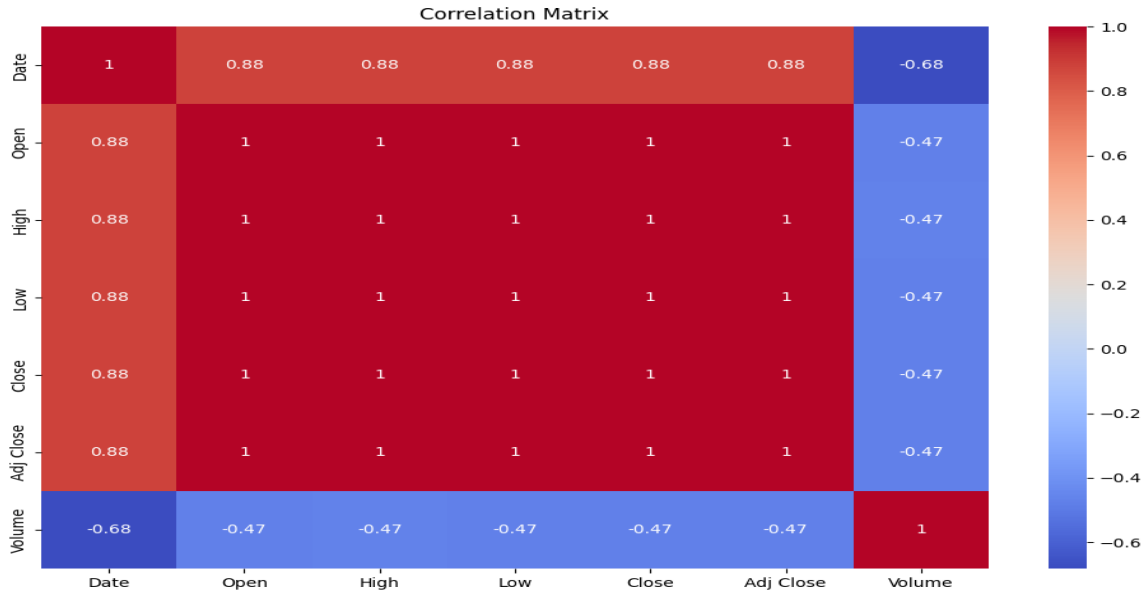
4. Exploratory Data Analysis (EDA):

EDA was conducted to understand the distribution of variables, identify correlations, and extract meaningful insights from the data. Key findings from the EDA include:

- **Distribution of Stock Prices**: The distribution of Open, High, Low, and Close prices showed a positive skew, indicating that most stock prices were lower, with a few very high values. This suggests that there were occasional significant price increases.

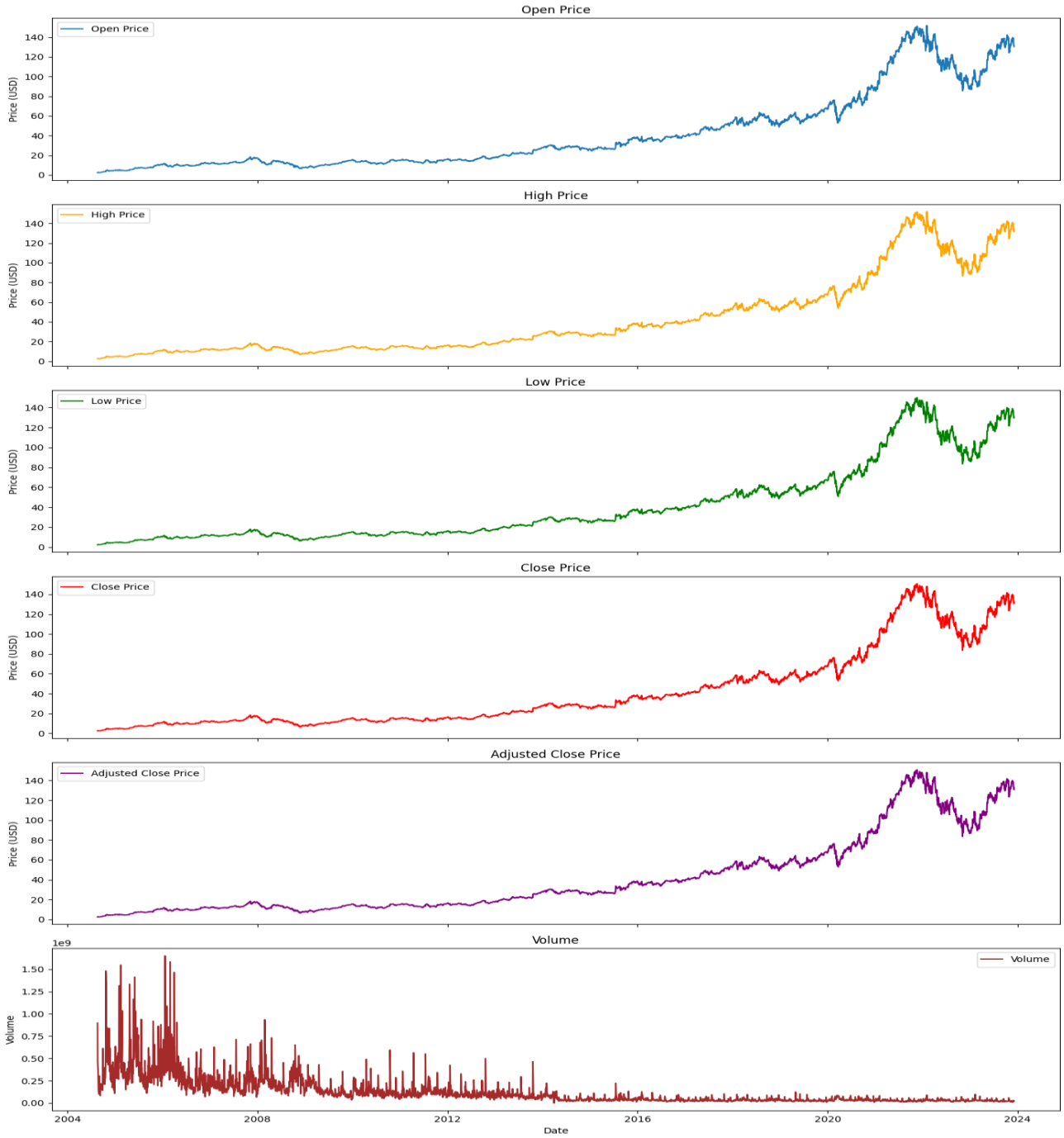


- **Volume Analysis**: The Volume of shares traded varied significantly, with occasional spikes corresponding to major market events or company announcements. High trading volumes often coincided with significant price movements, reflecting increased market activity.
- **Correlation Analysis**: A strong positive correlation was found between Open, High, Low, and Close prices, suggesting that these features move together. This is expected as these prices are interdependent throughout a trading day. The Volume had a weaker correlation with price

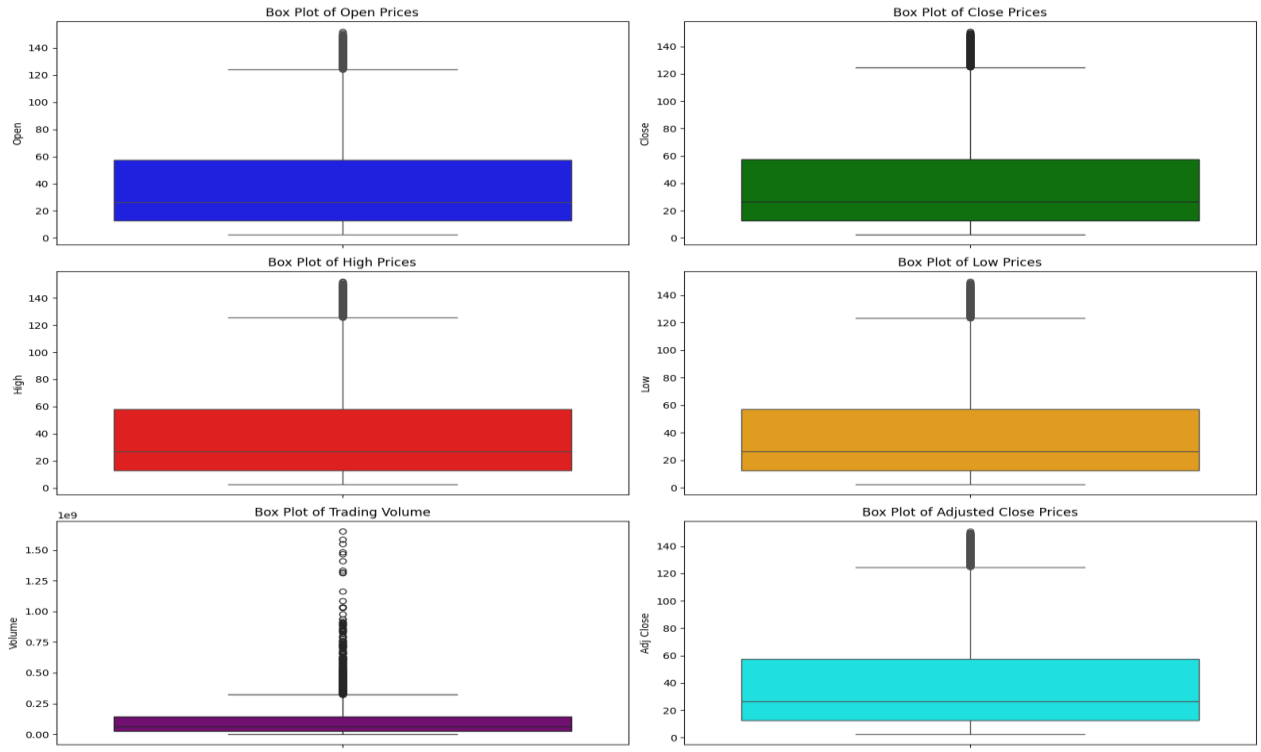


features, indicating that while trading volume affects prices, the relationship is not as strong as between the prices themselves.

- **Trend Analysis**: Over the years, Google stock prices showed an upward trend with occasional dips, reflecting overall market growth and specific downturns such as the 2008 financial crisis and the COVID-19 pandemic in 2020. This trend analysis is crucial for understanding the long-term growth and periodic volatility of the stock.



- **Volatility**: The stock exhibited periods of high volatility, especially during economic crises or significant company news. Volatility analysis helps in understanding the risk associated with the stock during different periods.



5. Feature Engineering:

- **Feature Creation**: No new features were created as the original dataset provided good predictions. The simplicity of the existing features was maintained to avoid overcomplicating the model.
- **Feature Scaling**: Feature scaling and normalization were attempted but did not improve predictions. Therefore, the original features were used, retaining their natural scale which seemed to work well with the models.
- **Feature Selection**: The Date and Adj Close columns were dropped as they were deemed irrelevant for prediction. The Date column, being a temporal identifier, was not necessary for the models. The Adj Close was redundant with the Close price already being used.

6. Model Building:

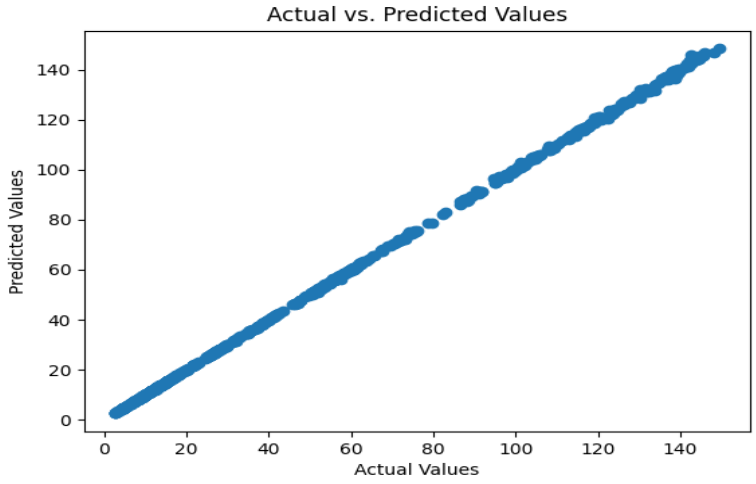
Three models were built using both Scikit-Learn and custom implementations:

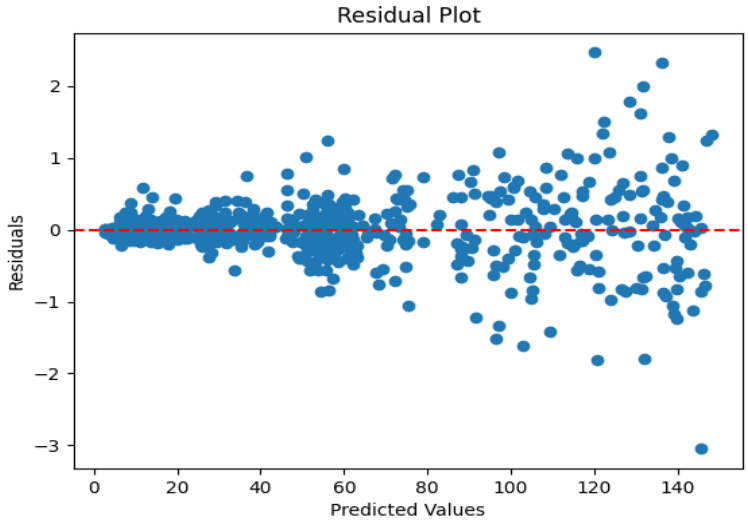
- **Linear Regression**: A simple yet effective model that fits a linear relationship between the features and the target variable. It performed well due to the strong linear relationships in the data.
- **Decision Tree Regression**: A non-linear model that splits the data into branches to make predictions. This model was useful for capturing non-linear patterns in the data.
- **Random Forest Regression**: An ensemble model that builds multiple decision trees and combines their predictions for better accuracy. This model generally provided robust predictions by averaging out the results of multiple trees.

7. Model Evaluation:

Models were evaluated using the following metrics:

- **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values. Lower values indicate better performance. It helped in understanding the overall prediction accuracy.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing error measurement in the same units as the target variable. RMSE made it easier to interpret the prediction errors.
- **Mean Absolute Error (MAE)**: Measures the average absolute difference between actual and predicted values. It provides a straightforward interpretation of the model's prediction error.
- **R-squared (R2)**: Indicates the proportion of variance in the target variable explained by the model. Higher values indicate better fit. It provided insights into the model's explanatory power.
- **K-Folds Cross-Validation**: A technique to assess model performance by dividing the data into k subsets and training/testing the model k times. This helped in evaluating the model's generalization capability and avoiding overfitting.



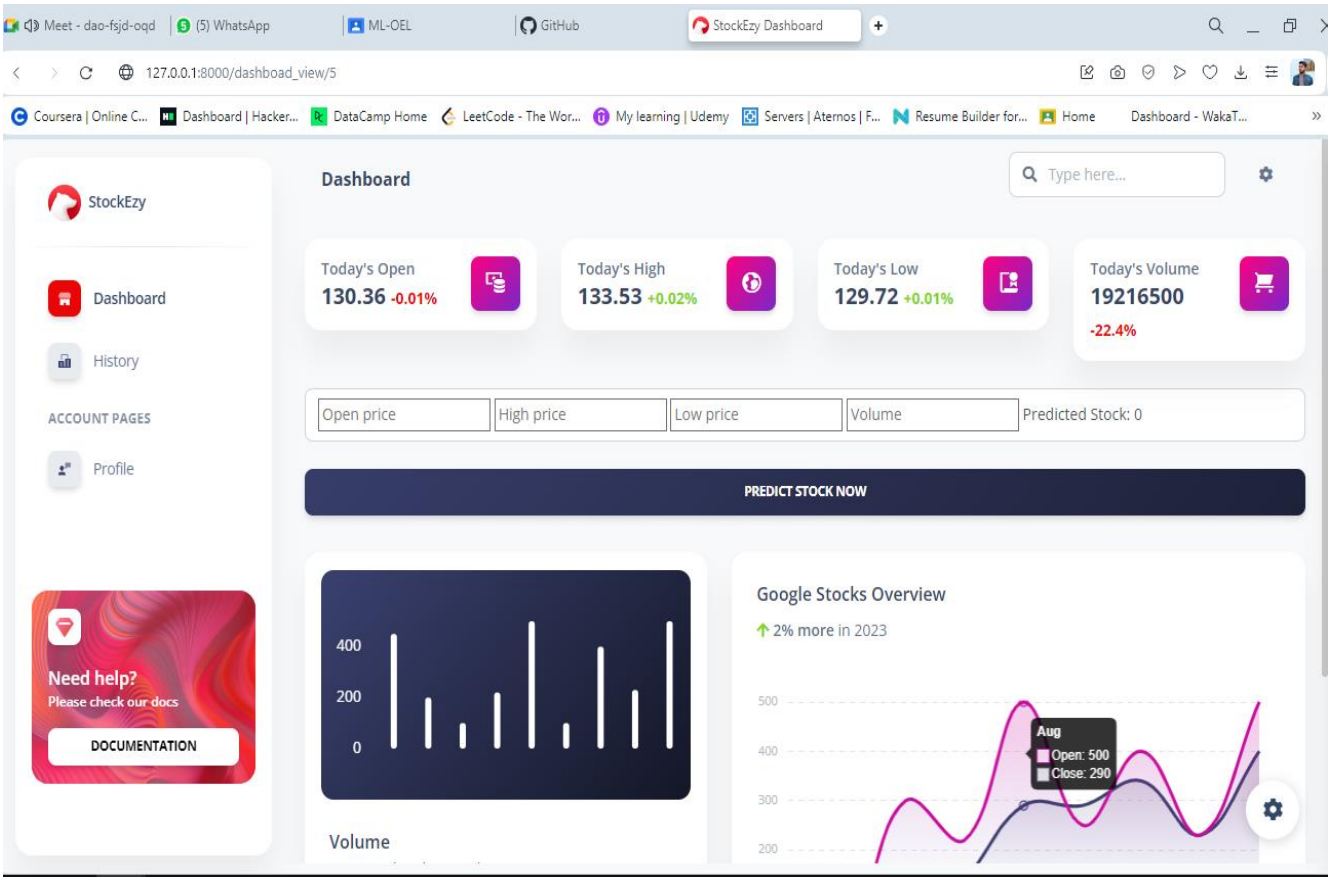


Linear Regression performed the best among the models, providing the most accurate predictions for Google stock prices. This was likely due to the strong linear relationships in the data.

8. Conclusion:

- **Findings**: Linear Regression provided the best predictions for Google stock prices, indicating a strong linear relationship between the features and the target variable. It effectively captured the overall trend and variations in the stock prices.
- **Insights**: The models can be further improved by incorporating additional features such as market trends, news sentiment, and other external factors. This could enhance the models' ability to capture complex market dynamics.
- **Limitations**: The models were limited to historical price data without considering external market factors. This limits their ability to predict future prices under changing market conditions.
- **Future Work**: Incorporating more features and trying advanced algorithms like LSTM (Long Short-Term Memory networks) for time series prediction could enhance model performance. LSTMs are particularly suited for time-series data and could better capture temporal dependencies and trends.

9. Application UI:



DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
BACHELORS IN COMPUTER SYSTEMS ENGINEERING
Course Code: CS-324
Course Title: Machine Learning
Open Ended Lab
TE Batch 2021, Spring Semester 2024
Grading Rubric
TERM PROJECT

Group Members:

Student No.	Name	Roll No.
S1		
S2		
S3		

CRITERIA AND SCALES				Marks Obtained		
				S1	S2	S3
Criterion1: Data Collection						
0	1	2	3			
The student has not chosen a suitable dataset for predictive modeling.	The student has chosen a dataset, but it may not be suitable for predictive modeling, or it lacks enough features.	The student has chosen a suitable dataset for predictive modeling, and it has enough features to work with.	The student has chosen an excellent dataset for predictive modeling, which has rich features and is well-suited for the task.			
Criterion 2: Data Preprocessing						
0	1	2	3			
The student has not performed data cleaning, handling missing values, or encoding categorical variables	The student has performed basic data cleaning and handled missing values, but has not encoded categorical variables.	The student has performed data cleaning, handled missing values, and encoded categorical variables.	The student has performed thorough data cleaning, handled missing values effectively, and encoded categorical variables efficiently.			
Criterion 3: Exploratory Data Analysis (EDA)						
0	1	2	3			
The student has not performed exploratory data analysis (EDA) or provided minimal analysis with no meaningful insights.	The student has performed basic exploratory data analysis, but the analysis lacks depth, and insights are limited	The student has performed thorough exploratory data analysis, identifying important variables, correlations, and providing meaningful insights.	The student has performed exceptional exploratory data analysis, providing comprehensive insights, and utilizing a variety of visualization techniques effectively.			
Criterion 4: Feature Engineering						
0	1	2	3			
The student has not performed feature engineering.	The student has performed basic feature engineering, but has not created new features or scaled/normalized existing features.	The student has performed feature engineering, creating new features and scaling/normalizing existing features if required.	The student has performed advanced feature engineering, creating meaningful new features and effectively scaling/normalizing existing features.			
Criterion 5: Model Building						
0	1	2	3			
The student has not built any predictive models.	The student has built models using machine learning algorithms, but the implementation lacks depth, and multiple algorithms were not used.	The student has built models using multiple machine learning algorithms, implementing them using Python packages, and evaluated their performance.	The student has built models using multiple machine learning algorithms, implemented them both using Python packages and without Python packages, and			

			thoroughly evaluated their performance.			
Criterion 6: Model Evaluation						
0	1	2	3			
The student has not evaluated model performance or has done so inadequately.	The student has evaluated model performance but has not used different techniques or compared the performance of different models.	The student has evaluated model performance using different techniques, compared the performance of different models, and selected the best-performing model.	The student has thoroughly evaluated model performance using various techniques, performed a detailed comparison of different models, and selected the best-performing model based on comprehensive evaluation metrics.			
Criterion 7: Conclusion						
0	1	2	3			
The student has not provided a conclusion or has provided a conclusion with minimal insights.	The student has provided a basic conclusion with some insights but has not discussed model limitations or suggested improvements.	The student has provided a detailed conclusion with meaningful insights, discussed model limitations, and suggested improvements.	The student has provided an exceptional conclusion with comprehensive insights, thorough discussion of model limitations, and insightful suggestions for improvements.			
Criterion 8: Report						
0	1	2	3			
The submitted report is unfit to be graded.	The report is partially acceptable.	The report is complete and concise.	The report is exceptionally written.			
			Total Marks:			