

**CD2L**

**NEWYORK STOCK PREDICTION MODEL**

**NAVEEN KUMAR S - 2nd YR- Artificial Intelligence and Data science**

**MABASHA R - 2nd YR- Artificial Intelligence and Data science**

**HARI HARA BUDRA P - 2nd YR- Artificial Intelligence and Data science**

**MOULI KRISHNAN A - 2nd YR- Artificial Intelligence and Data science**



**MAY 2023**

1. **OVERVIEW OF THE PROJECT :**

**The New York Stock Exchange dataset provides historical data on daily stock prices, opening and closing prices, trading volumes, and market capitalizations. The dataset aims to facilitate the development of predictive models that forecast future stock prices based on historical data. It is useful for investors and financial analysts seeking to predict market trends.**

**Our dataset contains 851264 x 7 rows and columns. The accuracy rate of our model is 96%.**

1. **DATA ACQUISITION :**

**Acquiring and processing financial data can be quite overwhelming , especially dealing with massive datasets like the New York Stock Exchange dataset.**

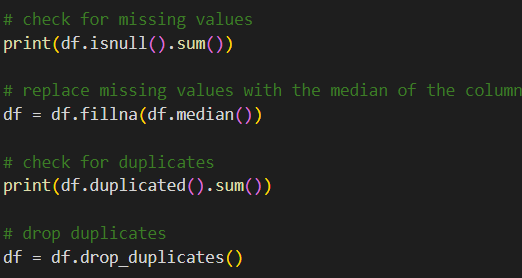
**DESCRIPTION OF THE COLUMNS:**

* + **Open => Price from the 1st transaction of a trading day.**
  + **Close => Price from the last transaction of a trading day.**
  + **Symbol => Symbols that denote company name of stock.**
  + **Volume => Number of units traded in a day.**
  + **High => Maximum price in a trading day.**
  + **Low => Minimum price in a trading day.**

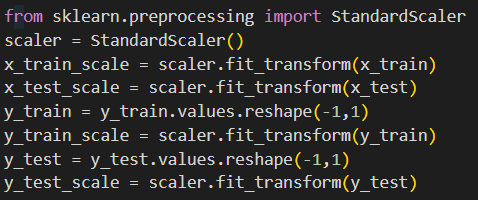
**During the acquisition process, faced several common problems that one might encounter, including data quality issues, data format and structure, data volume, data accessibility, and datatype. We have experienced data quality issues before when dealing with datasets, and the New York Stock Exchange dataset is no exception. As for data format and structure, we found it challenging to work with various formats such as CSV, Excel, or JSON, which requires extra effort to clean and restructure the data for analysis. Moreover, the data volume is massive, which demands specialized tools and techniques to process such a large volume of data. Overall, acquiring and processing financial data requires patience, expertise, and a thorough understanding of the data. Feel free to check our dataset used in the ML Model**

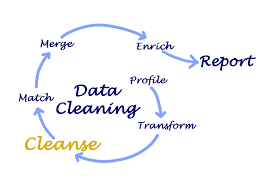
**DATSET LINK => https://drive.google.com/file/d/1IVYVh0ZfZxhBCDD8rYgDihKAu\_B6HtO/view?usp=share\_link**

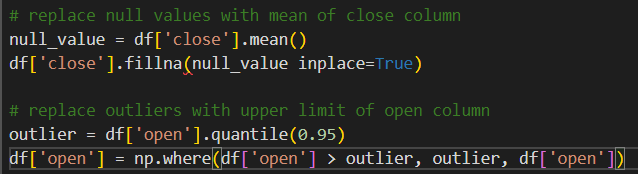
1. **DATA CLEANSING AND TRANSFORMATION :**

**As we worked on this project, we understood that data cleansing and transformations are critical steps in the data analysis process, especially when dealing with a large and complex dataset like the one we are working with. To ensure that we have consistency and accuracy in our analysis, it is essential to establish a standardized approach to data cleansing and transformations as a team. **

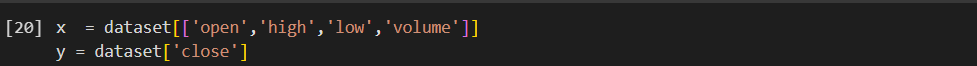
**To achieve this, we have taken the following steps:**

1. **Establish data quality standards: we had worked with my team to define the standards for data quality, including data completeness, accuracy, consistency, and timeliness. By doing this, we have ensured that the data we work with is reliable and suitable for analysis. **
2. **Develop a data cleansing plan: We had created a plan that outlines the steps required to clean and transform the data. This plan includes the techniques we have used, the order of operations, and the expected output.**



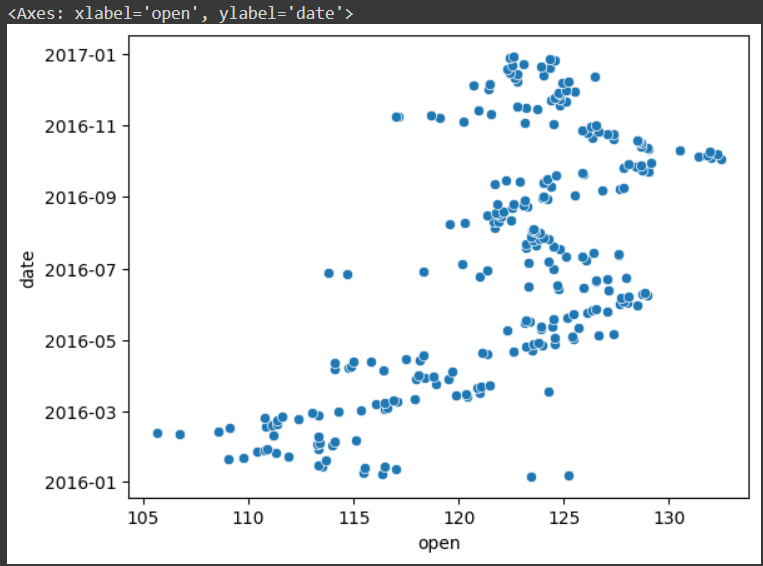
1. **Handling Null Values and outliers: We had cleaned the null values by replacing them by median values in the dataset and the outliers have been solved by replacing with the mean value of the data as it is the common procedure of handling data.** ****
2. **Collaborate on data transformations: we have collaborated with our team members to ensure that everyone is using the same data transformation techniques. By doing this, we had ensured consistency in our analysis and reduced the risk of errors.**
3. **SENSITIVE FEATURES :**

**The sensitive features that we have used from our dataset are the columns by which the independent variable is determined, then the dependent variable is selected sensitive features in machine learning models have the potential to unfairly discriminate against individuals or groups. It's important to identify, evaluate, and address such features to ensure fair and unbiased predictions.**

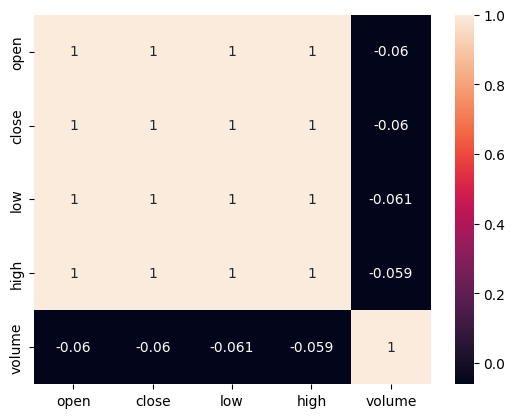
****

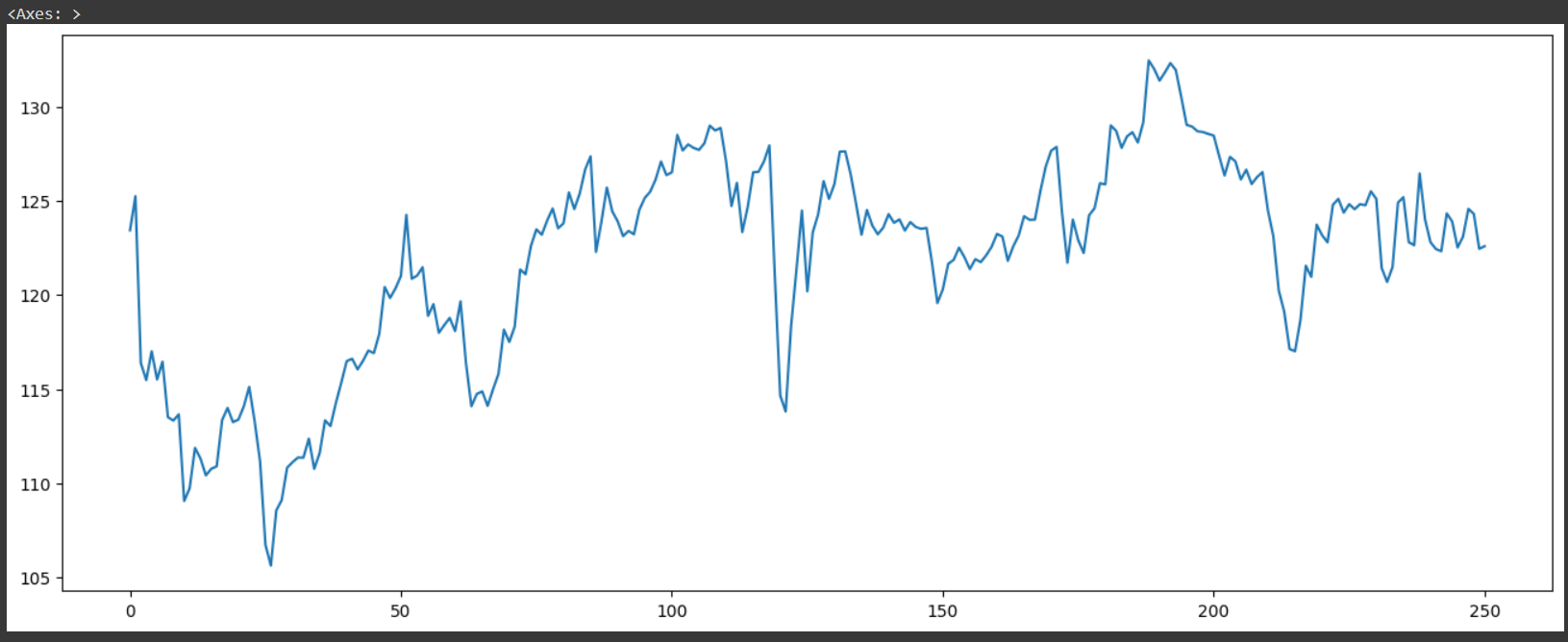
1. **EXPLORATORY DATA ANALYSIS:**

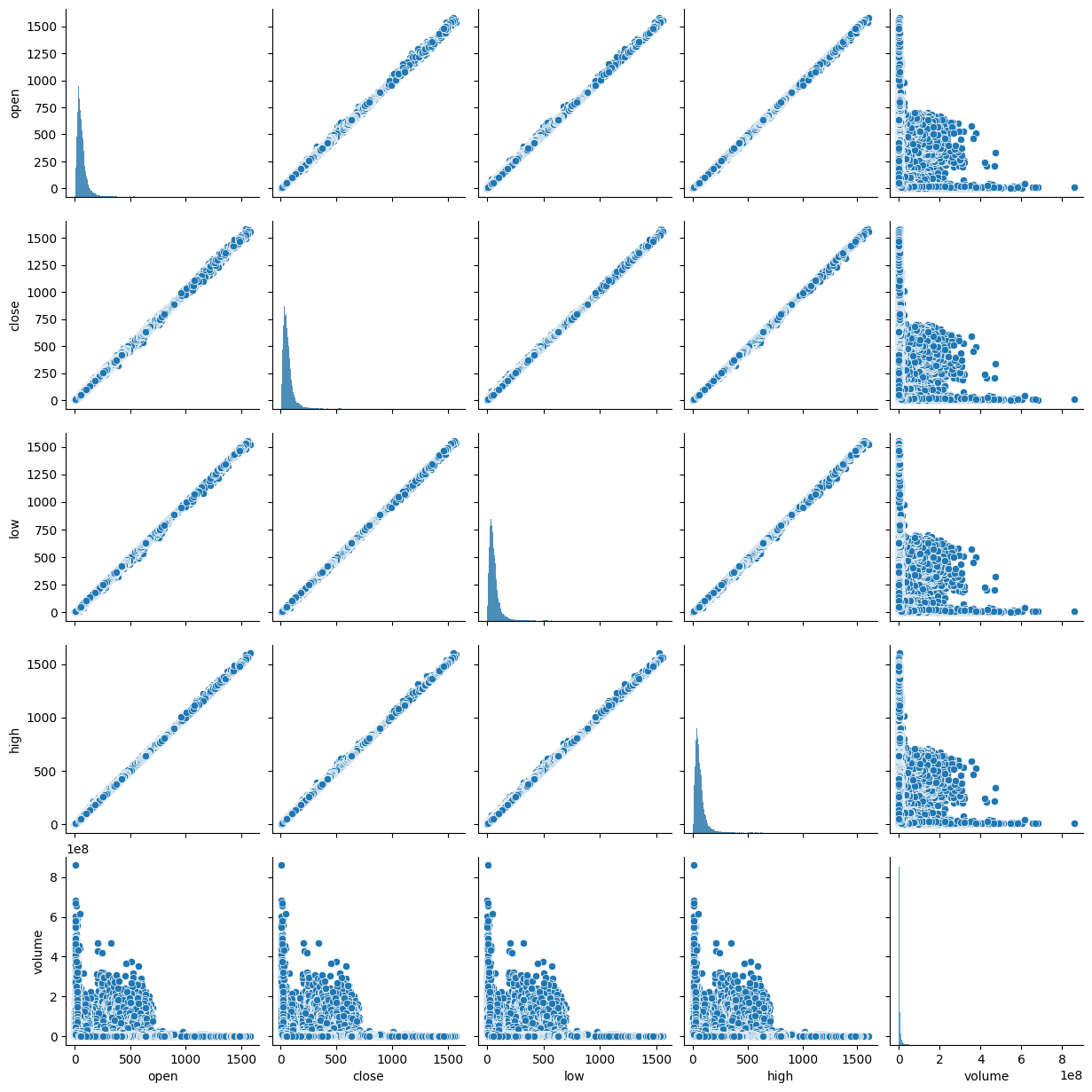
**Exploratory data analysis (EDA) is a critical step in our data analysis project, including the New York Stock Exchange prediction project.**

**ScatterPlot: **

**The goal of EDA in our project is to gain a deeper understanding of the data and identify any the patterns, trends, or anomalies that are present. We identify the variables that are used in the analysis and determine their data types (e.g., numerical or categorical). We had plotted the distributions of each variable to identify any outliers or unusual patterns. Examine the relationships between variables to identify any correlations or dependencies.**

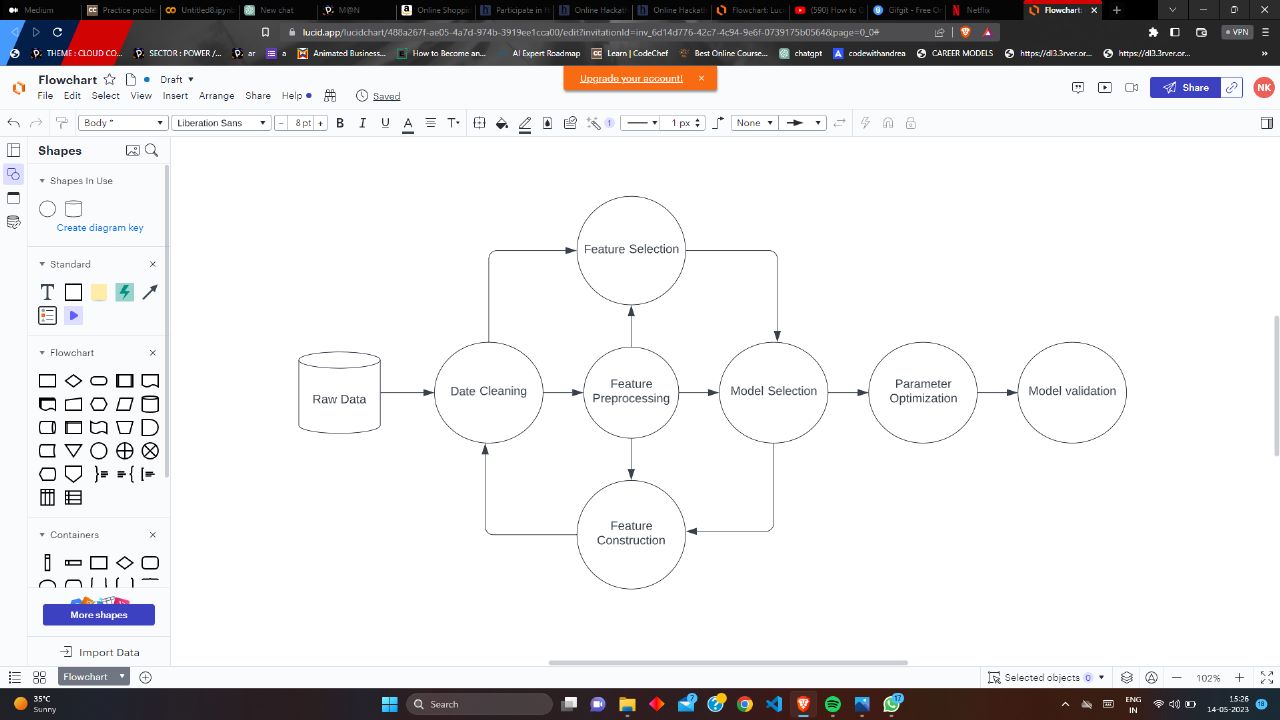
****

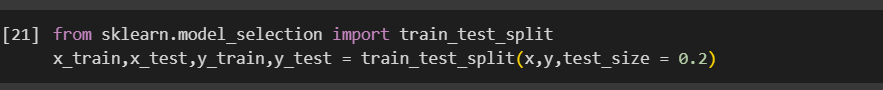
**Simple Plot for Open:Use visualizations such as scatterplots, histograms, and box plots to gain insights into the data. Overall, conducting thorough EDA is crucial for building a successful predictive model for the New York Stock Exchange.**

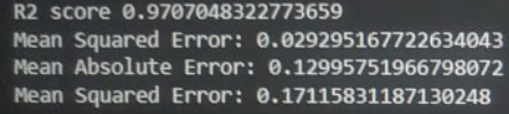
**Pairplot:**

**It helps ensure that the data is suitable for analysis and that any potential issues or biases are identified and addressed.**

1. **MODEL SELECTION**

**The type of problem that we are trying to solve regression problem. **

**Split the New York dataset into training and testing sets. This will help you evaluate the performance of different models and avoid overfitting. **

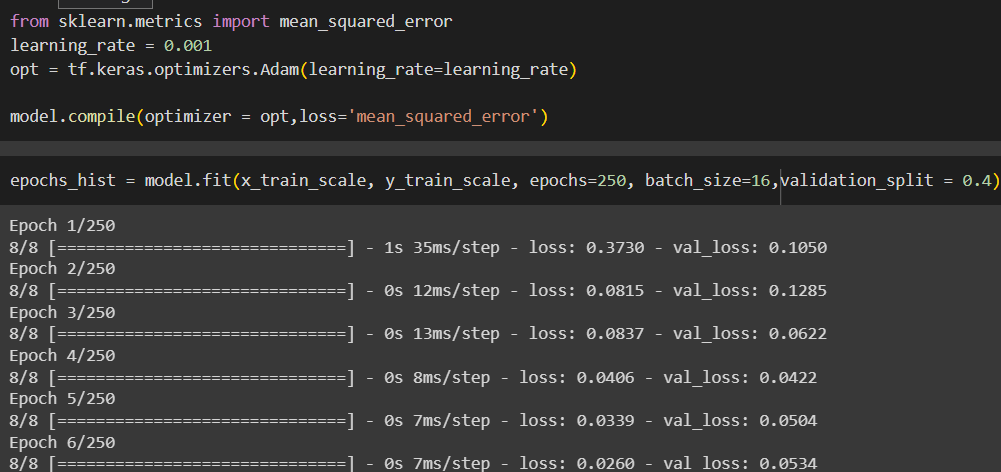
**Choose evaluation metrics that are appropriate for the problem you are trying to solve. For example, for a classification problem, you can use metrics such as accuracy, mean absolute error, mean squared error and R2 score. **

**Choose a set of candidate models that are appropriate for the problem you are trying to solve. For example, for a regression problem, you can choose models such as linear regression, decision trees, or random forests.**

****

**MODEL TRAINING AND EVALUATION**

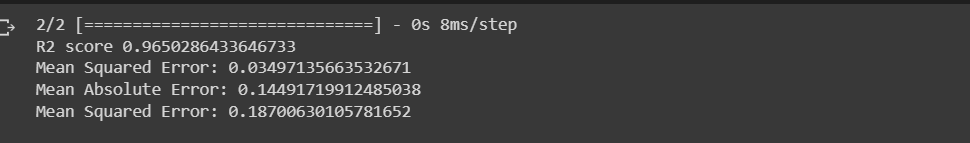
**Pre process the data by cleaning, transforming, and normalizing it. This involves handling missing values, encoding categorical variables, and scaling features as necessary. Splitting the data: Split the pre processed data into training and testing sets. The training set will be used to train the model, while the testing set will be**

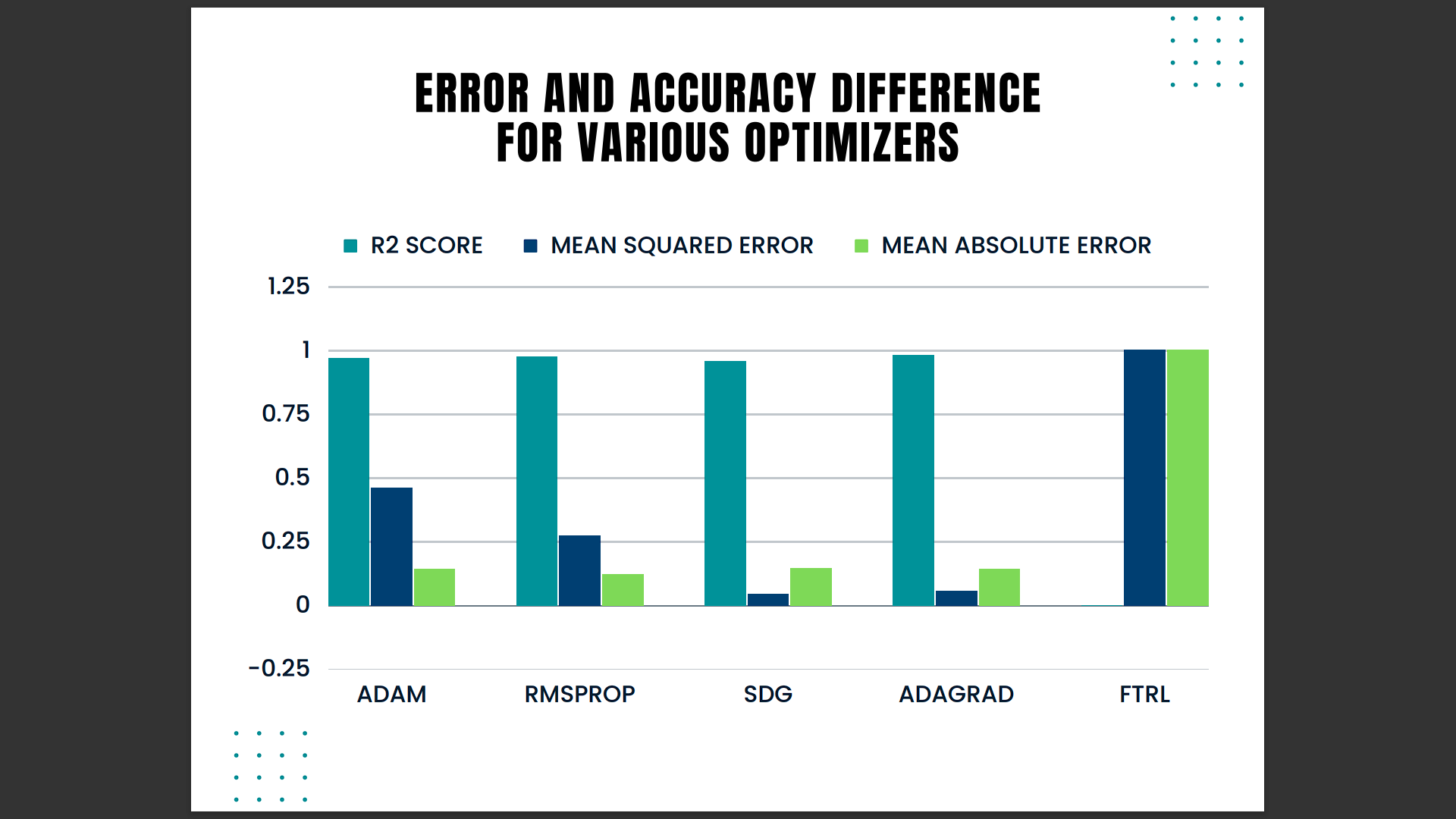
**used to evaluate the performance of the model.** ****

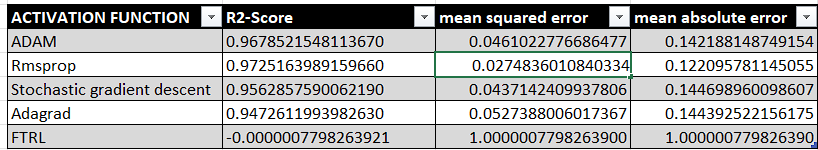
**Training the model: Train the model on the training set using a suitable algorithm and hyperparameters. This involves fitting the model to the training data and optimizing the model parameters. Evaluate the performance of the trained model on the testing set using appropriate evaluation metrics. This involves computing metrics such as accuracy, precision, recall, F1 score, and mean squared error.**

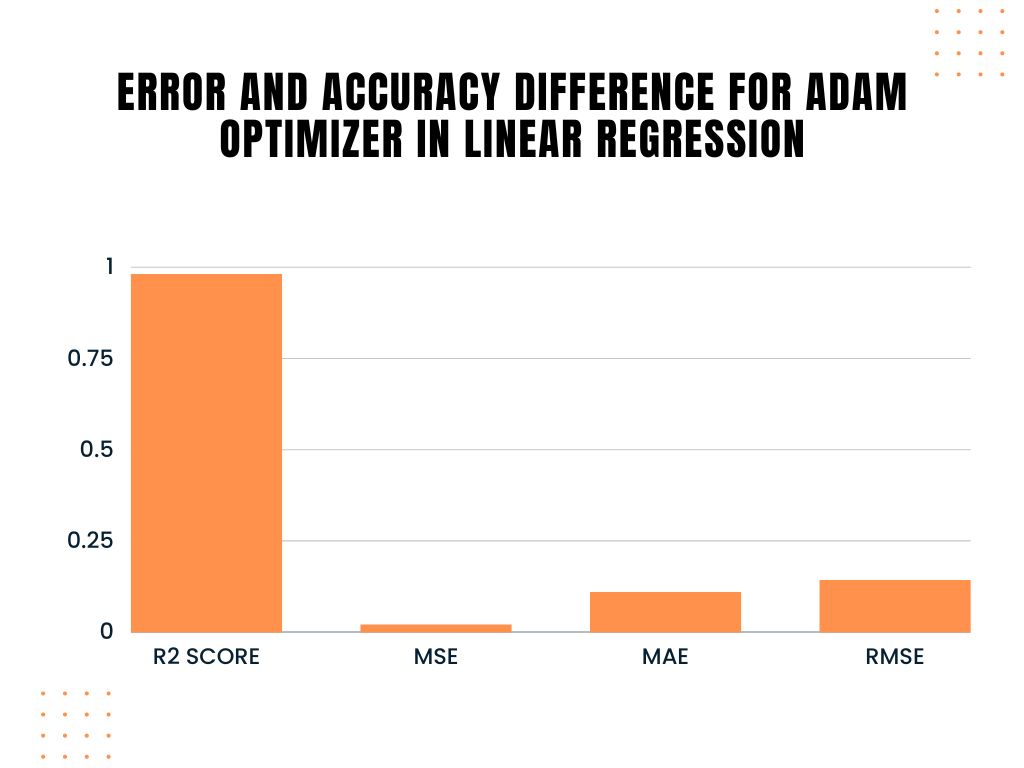
1. **COMPARISON AND ANALYSIS OF METRICS USING DIFFERENT PARAMETERS (PERFORMANCE OF THE MODEL)**

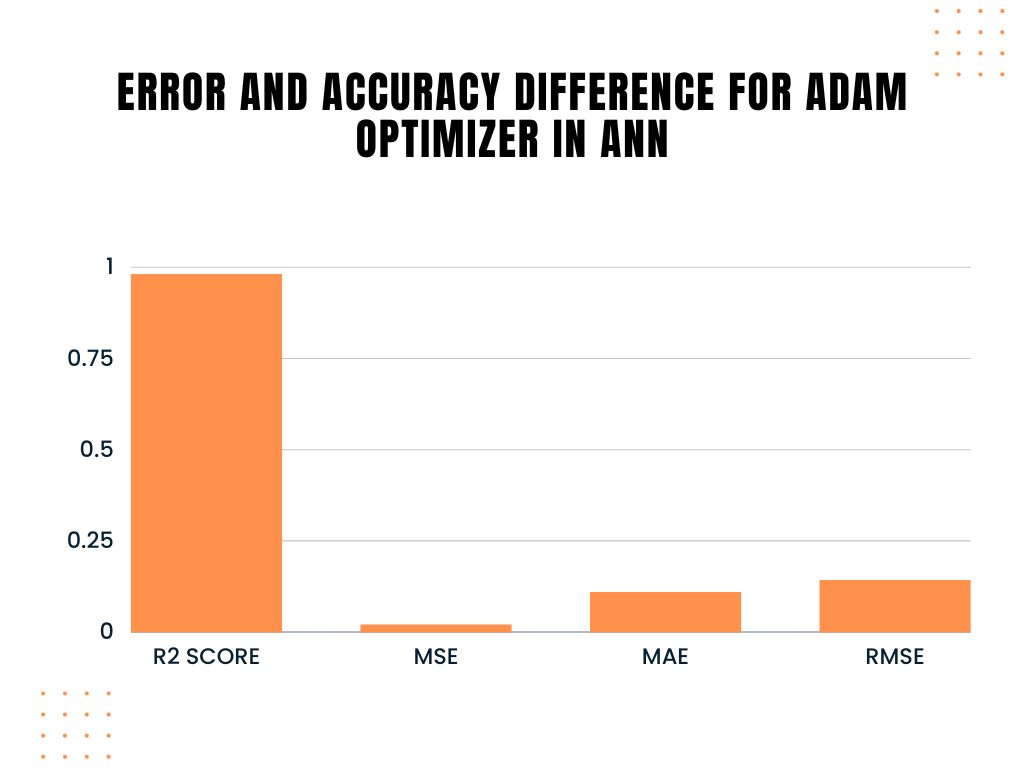
**Accuracy: This metric measures the proportion of correct predictions made by the model. It is suitable for classification problems where the classes are balanced. We used the rmsprop , adagrad and relu , linear activation activation function, and when we did the data’s got deviated from the complete data.**

** Precision: This metric measures the proportion of true positives among all positive predictions made by the model.**

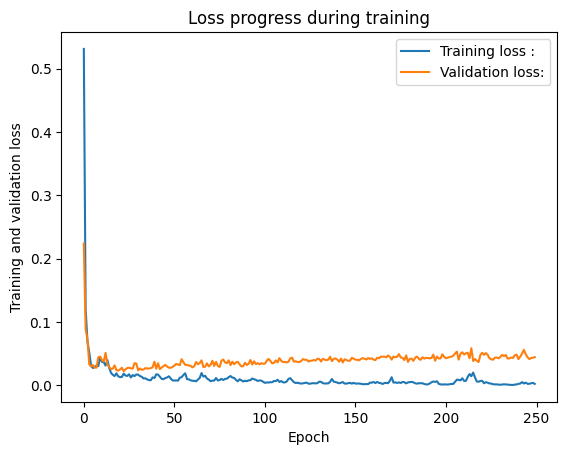


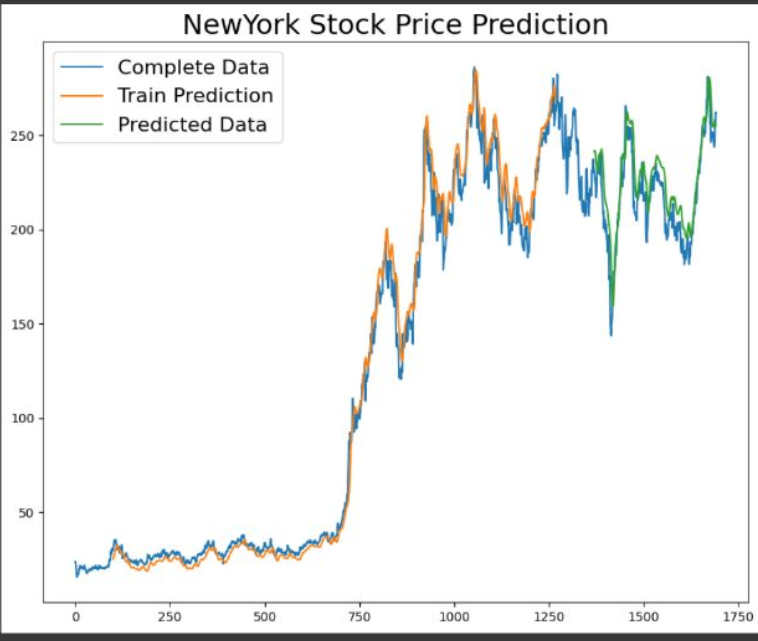
**It is suitable for problems where minimizing false negatives is important. This metric measures the average squared difference between the predicted and actual values in a regression problem. It is suitable for problems where minimizing the overall error is important.**





1. **CONCLUSION :**

**The New York Stock Prediction project has demonstrated the potential of machine learning techniques in forecasting stock prices.. The model developed in our project serve as a tool to supplement decision-making processes and market analysis.**

****

**This documentary sheds light on the significance of data-driven approaches in the finance industry and highlights the continuous evolution and refinement of predictive models for stock market analysis. The project serves as a stepping stone for further research and advancements in the field of stock market prediction, paving the way for more informed and strategic investment decisions in the New York stock market.**