Stock Market Predictive Modeling

Introduction

The aim of this project is to forecast future stock prices using various regression models. We employ a dataset comprising stock market data with features such as open, high, low, close prices, volume, and technical indicators. The goal is to predict the target stock prices using machine learning techniques.

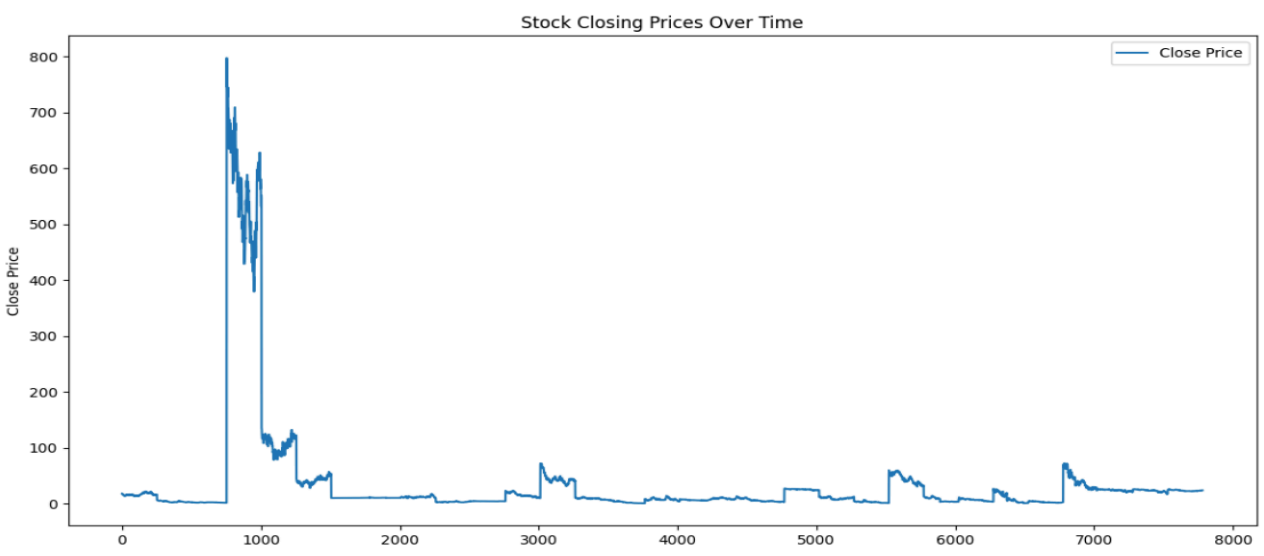
Data Preparation

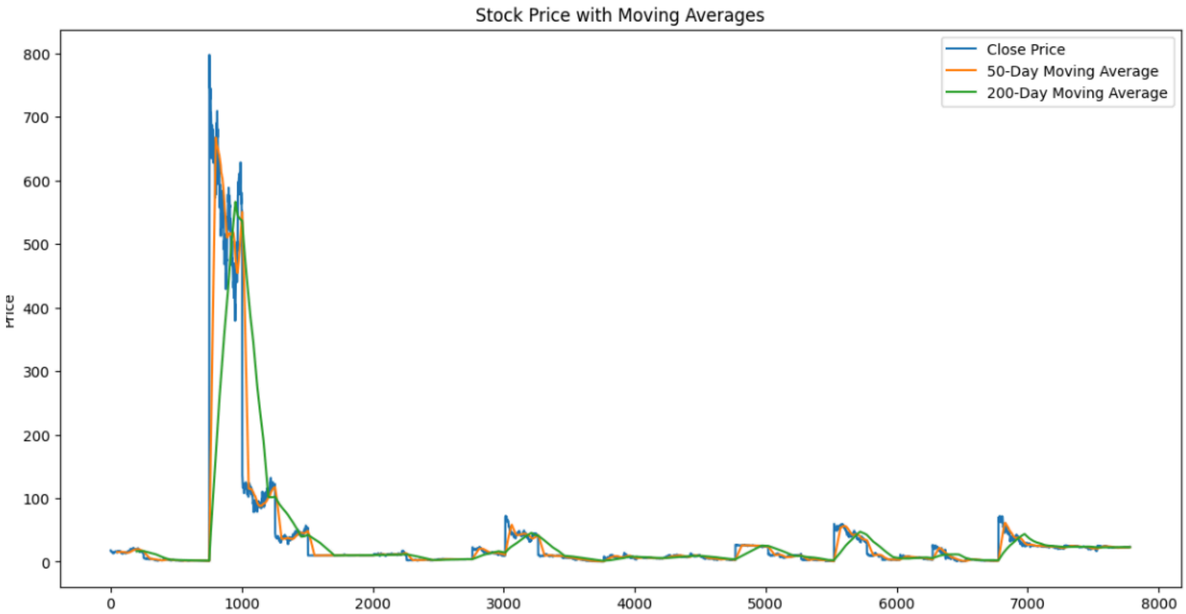
1. Load Dataset: The dataset is loaded from a CSV file and contains features such as \`open\`, \`high\`, \`low\`, \`close\`, \`volume\`, along with a \`TARGET\` column indicating the future stock price.
2. Date Parsing: The \`date\` column is converted to a datetime format and set as the index for time series analysis.
3. Data Cleaning: Ensured all columns are numeric, converting where necessary, and handled missing values by filling them with the mean of the respective columns.

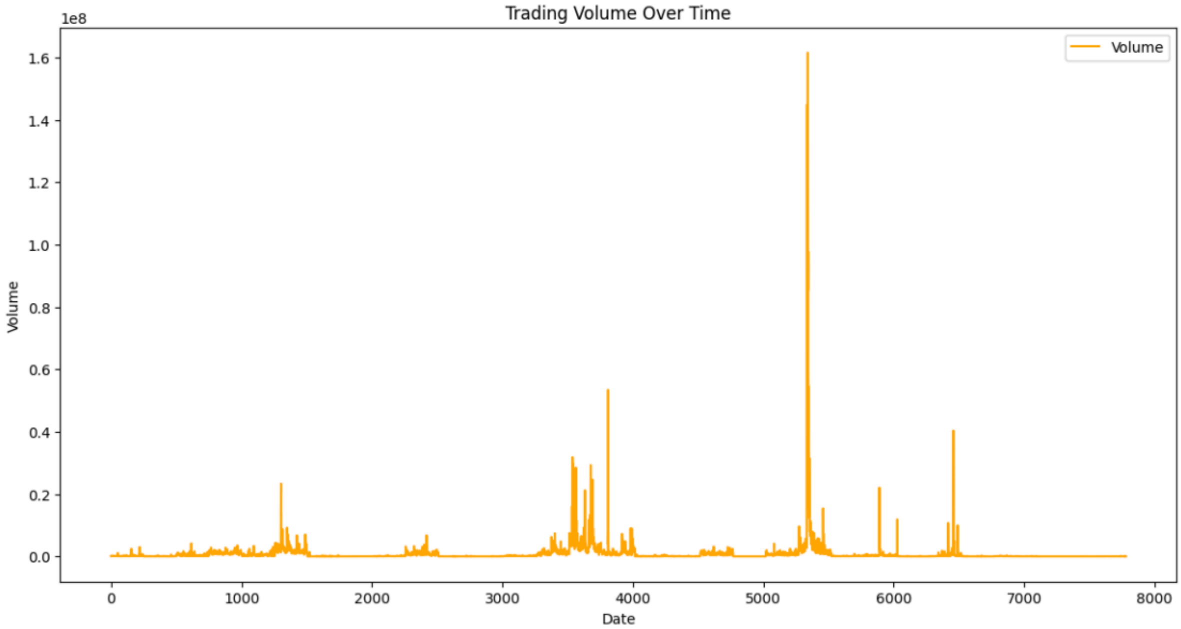
Exploratory Data Analysis (EDA)

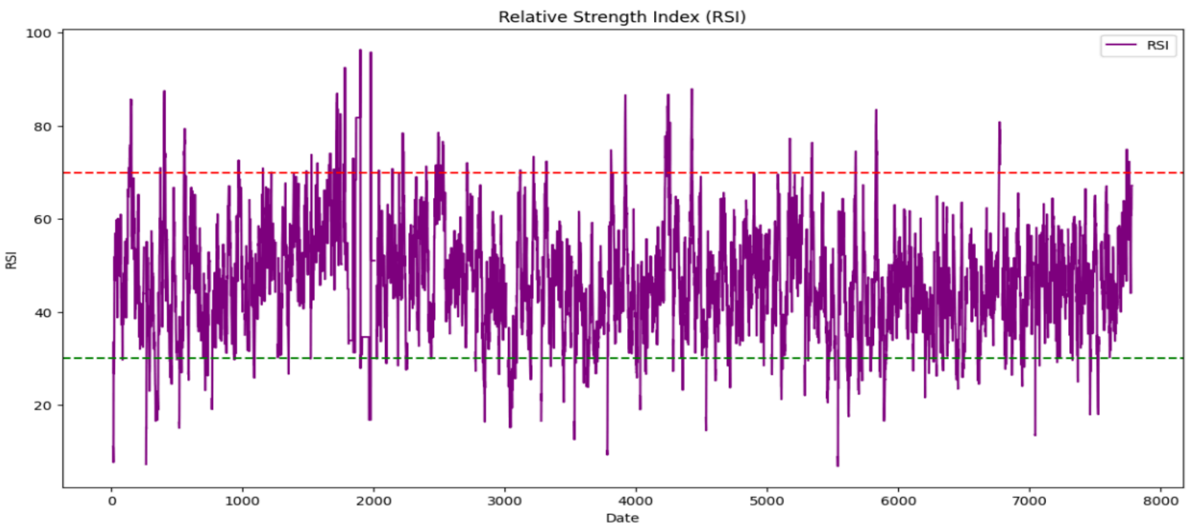
1. Provided an overview of the dataset using descriptive statistics.
2. Created plots for:
   - Time series of closing prices.
   - Distribution and box plots of the \`TARGET\` variable.
   - Pair plots for visualising relationships between key features and the \`TARGET\`.

Insights from EDA

- Observed general trends and patterns in stock prices over time.
- 


Stock Price with Moving Averages
Close Price
50-Day Moving Average
200-Day Moving Average
Price
800
700
600
500
400
300
200
100
0
0
1000
2000
3000
4000
5000
6000
7000
8000


Trading Volume Over Time
1e8
Volume
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
7000
8000
Date


Relative Strength Index (RSI)
RSI
100
80
60
40
20
0
1000
2000
3000
4000
5000
6000
7000
8000
Date

Model Training and Evaluation

1. Train-Test Split

The dataset was split into training and testing sets, ensuring that the feature columns (predictors) and the target column (`TARGET`) were correctly defined.

2. Chosen Predictive Models

I used four regression models to predict future stock prices:

1. Support Vector Regression (SVR)
2. Decision Tree Regressor
3. Random Forest Regressor
4. Linear Regression

1. Support Vector Regression (SVR)

SVR is chosen for its effectiveness in capturing non-linear relationships in the data. We used the Radial Basis Function (RBF) kernel, which is suitable for complex datasets with unknown distributions.

2. Decision Tree Regressor

Decision Trees are intuitive and provide easily interpretable models. They can capture non-linear relationships and handle both numerical and categorical data.

3. Random Forest Regressor

Random Forest is an ensemble method that builds multiple decision trees and averages their predictions. It improves accuracy and controls overfitting.

4. Linear Regression

Linear Regression serves as a baseline model. It assumes a linear relationship between the features and the target variable.

Results and Evaluation

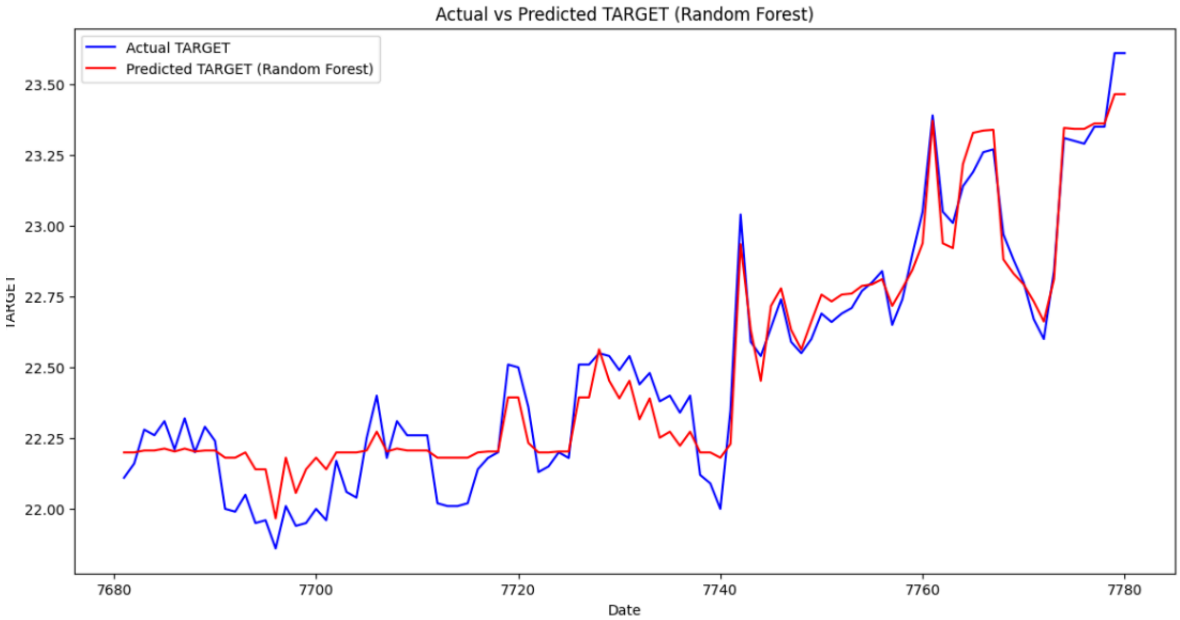
The models were evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

- **Support Vector Regression (SVR)**
  - RMSE: 95.31
  - MAE: 27.94

- **Decision Tree Regressor**
  - RMSE: 2.47
  - MAE: 0.53

- **Random Forest Regressor**
  - RMSE: 1.89
  - MAE: 0.44

- **Linear Regression**
  - RMSE: 1.02
  - MAE: 0.30



Insights

1. Model Performance:

- Random Forest and Linear Regression showed the best performance in terms of RMSE and MAE, indicating strong predictive power.
- SVR showed significantly higher errors, suggesting it may not be as suitable for this dataset.

2. Feature Importance:

- Random Forest provides insights into feature importance, which can be valuable for understanding the drivers of stock price movements.

3. Non-Linear Relationships:

- Models like Random Forest and Decision Tree can capture nonlinear relationships better than Linear Regression, leading to more accurate predictions.

Forecasting Future Stock Values

Using the trained models, particularly Random Forest due to its superior performance, we forecast future stock values and visualise the predictions.

Conclusion

In this project, we successfully applied and evaluated multiple regression models to predict stock prices. The Random Forest model emerged as the most effective, demonstrating the importance of considering non-linear relationships and ensemble methods in stock price prediction.

Future work could explore more advanced techniques like deep learning models and further feature engineering to improve predictive accuracy.