Stock Market Prediction Documentation

1. Introduction

This documentation presents an analysis of stock market data using machine learning models. The analysis includes data collection, preprocessing, exploratory data analysis (EDA), and the implementation of various machine learning algorithms for stock price prediction. Dataset: Stock market prediction (kaggle.com)

2. Importing Libraries

Libraries such as pandas, NumPy, matplotlib, seaborn, and sklearn are imported for data manipulation, visualization, and machine learning tasks.

3. Data Collection and Preprocessing

- Data is loaded from a CSV file into a pandas DataFrame.
- Irrelevant columns are dropped, and missing values are handled.
- The data is converted to datetime format for time series analysis.
- Summary statistics are computed to understand the distribution of the data.
- Functions used:
  - → pd.read_csv()
  - → df.drop()
  - → df.isnull().sum()
  - → df.info()
  - → df.fillna()
  - → pd.to_datetime()
  - → df.head()
  - → df.describe()

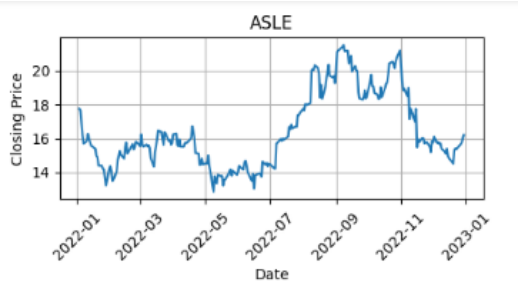
4. Exploratory Data Analysis (EDA)

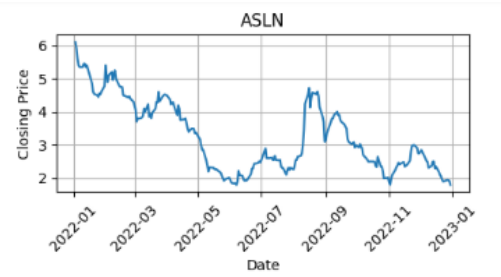
4.1 Time Series Analysis

Closing prices of different companies are plotted over time to observe trends and patterns.

Each company's closing price is visualized on individual subplots.

Example plots:

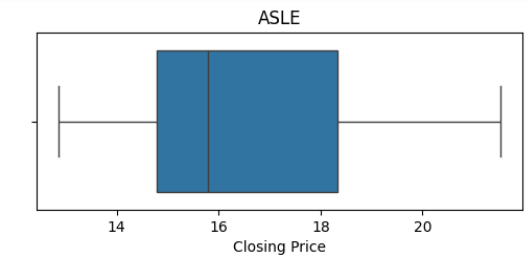




4.2 Box Plot for Closing Price

Box plots are created to visualize the distribution of closing prices for each company.

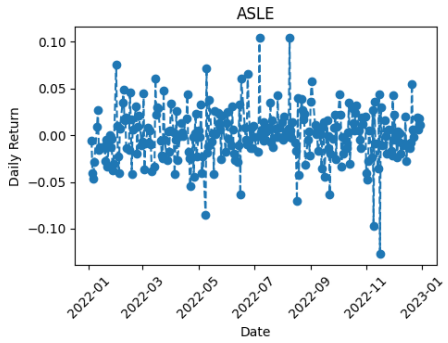
Example plots:

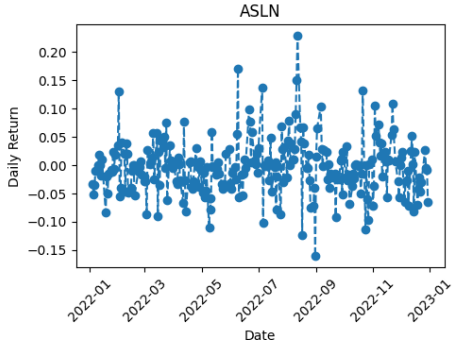


4.3 Analysis of Daily Returns

Daily returns for each company are computed and plotted over time to analyze volatility.

Example plots:





5. Machine Learning Models for Stock Price Prediction

5.1 Linear Regression

Description: Linear regression is a simple and commonly used statistical technique for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the target variable.

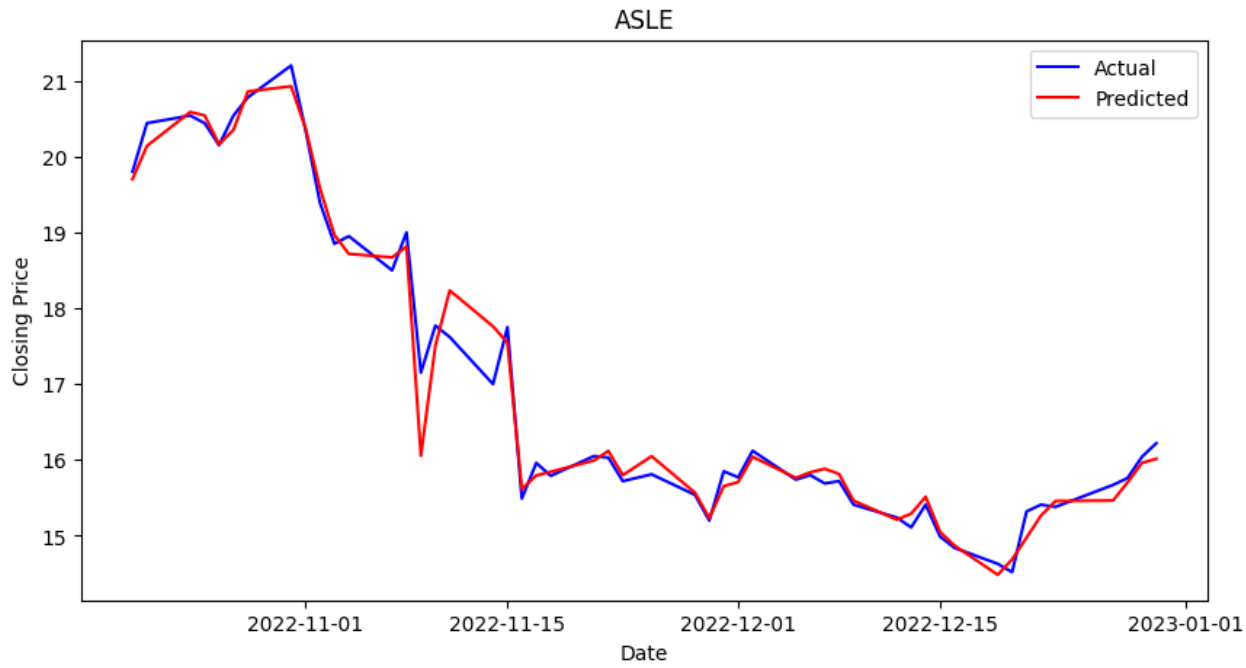
Usage: It is widely used for prediction and forecasting tasks when the relationship between variables can be approximated by a straight line.

For Stock market prediction:

- Linear regression models are trained and evaluated for each company.
- Mean Squared Error (MSE) and R-squared (R^2) scores are calculated.
- Actual vs. predicted closing prices are visualized for each company.

Example plot:

```
For ASLE:
Mean Squared Error: 0.06452565505159956
R^2 Score: 0.9839042125316576
```



5.2 XGBoost Regression

Description: XGBoost (Extreme Gradient Boosting) is a popular implementation of gradient boosting machines, which are ensemble learning methods used for regression and classification tasks. XGBoost builds a series of decision trees sequentially, where each tree corrects the errors of the previous one.

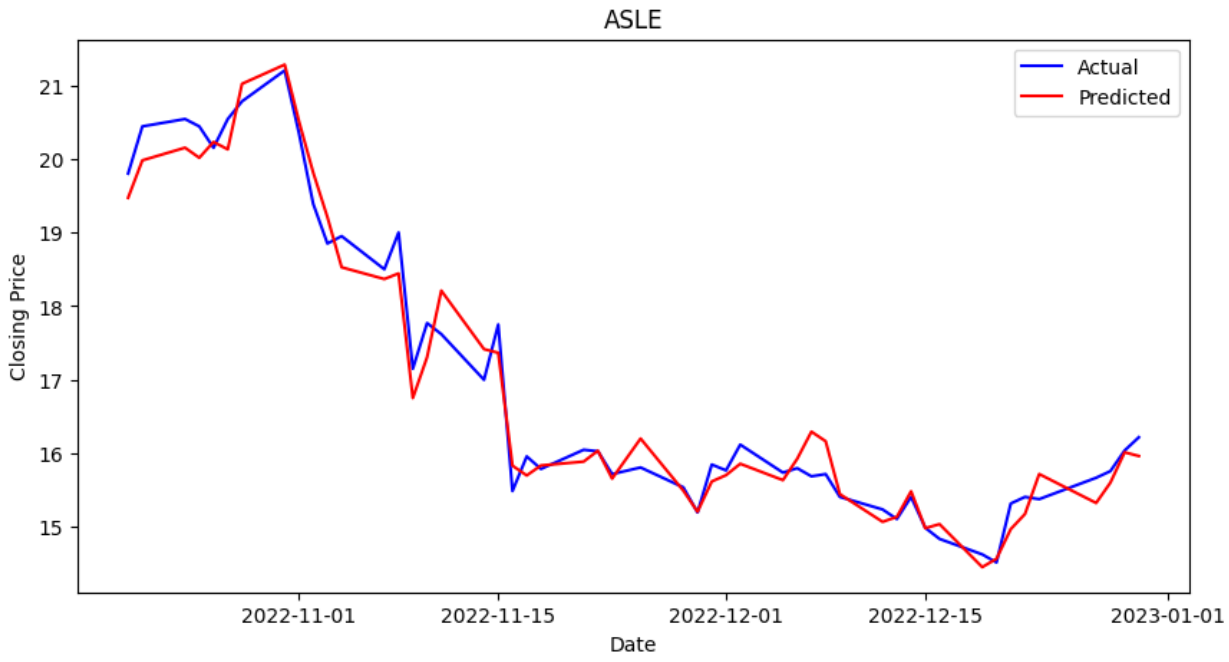
Usage: It is widely used in competitions and real-world applications for its high predictive accuracy and efficiency.

For Stock market prediction:

- XGBoost regression models are trained and evaluated for each company.
- MSE and R^2 scores are calculated.
- Actual vs. predicted closing prices are visualized for each company.

Example plot:

```
For ASLE:
Mean Squared Error: 0.08940232861490578
R^2 Score: 0.9776987792001539
```



5.3 Random Forest Classifier

Description: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. Each tree is trained on a random subset of the training data and a random subset of the features.

Usage: It is commonly used for classification and regression tasks where the input features have complex relationships with the target variable.

For Stock market prediction:

- Random forest classifier models are trained and evaluated for each company.
- Accuracy scores are calculated.
- Actual vs. predicted target values are displayed for each company.

Evaluation and Analysis:

ASLE:

	Actual	Predicted
201	True	True
202	True	True
203	False	True
204	False	False
205	True	True

Accuracy: 0.86

ASLN:

	Actual	Predicted
452	True	True
453	False	False
454	False	False
455	False	False
456	False	False

Accuracy: 0.84

ASMB:

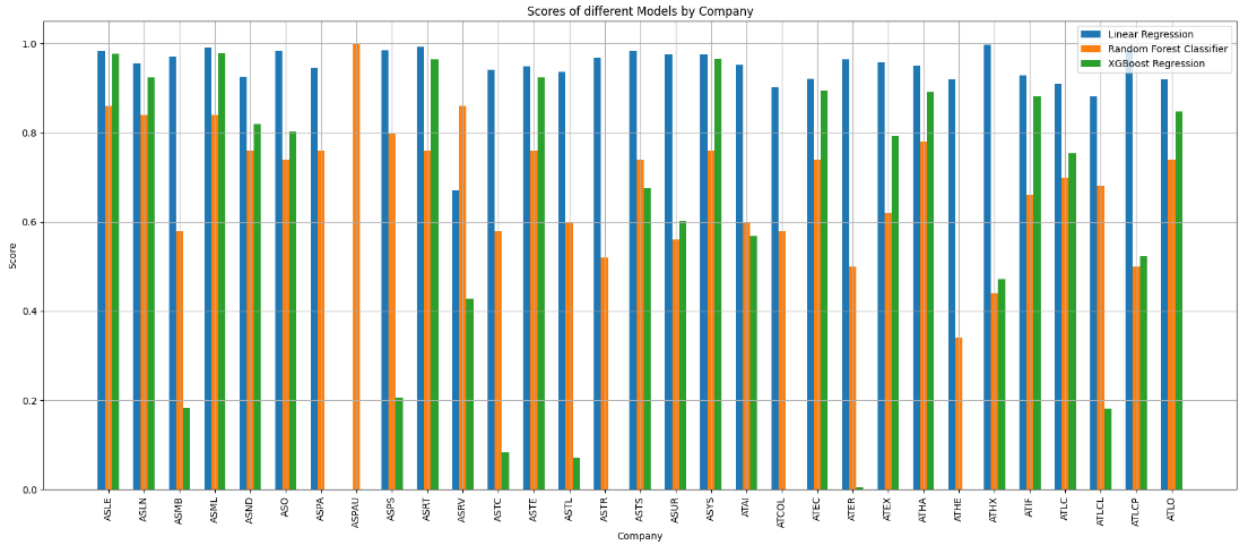
	Actual	Predicted
703	False	False
704	True	False
705	True	True
706	True	True
707	False	False

Accuracy: 0.58

6. Analysis of Model Performance

The performance of different machine learning models is compared using bar charts.

Mean R^2 score for linear regression, mean accuracy score for random forest classifier, and mean R^2 score for XGBoost regression are computed.




```
Mean r2 score of Linear Regression: 0.9015542182845016
Mean r2 score of XGBoost Regression: 0.28997662271996516
Mean accuracy score of Random Forest Regression: 0.6838709677419355
```

7. Conclusion

Based on the analysis, it is concluded that the stock market data fits best with the Linear Regression algorithm, achieving an R^2 score of 0.9.