Report: Analysis of S&P 500 Stock Data

Main Objective:

The primary aim of this analysis is to train linear regression models on S&P 500 stock data to predict the closing prices. The focus encompasses both prediction and interpretation of the outcome variable.

Data Description:

The dataset employed for this analysis contains historical stock data of various companies in the S&P 500 index. It encompasses attributes such as open, high, low, close prices, and volume traded. The dataset spans over a period of five years.

Data Exploration and Cleaning:

Upon loading the dataset, an exploration was conducted to understand its structure and contents. The 'date' column was identified as the datetime index, and the closing prices were visualized over time. The dataset was examined for missing values, and NaN values were found primarily in the 'Name' column, which was subsequently dropped. Additionally, the remaining NaN values were filled with the mean of each column.

Training Linear Regression Models:

Three variations of linear regression models were trained: Simple Linear Regression, Ridge Regression, and Lasso Regression. The independent variable used was the opening price ('open'), while the dependent variable was the closing price ('close'). These models were evaluated using metrics such as R^2 score and Mean Squared Error (MSE).

Model Performance:

Simple Linear Regression:

R^2 Score: 0.9998667604395625

MSE: 0.3564094413292893

Ridge Regression:

R^2 Score: 0.9995666328480284

MSE: 4.061682137320486

Lasso Regression:

R^2 Score: 0.9995667139486017

MSE: 4.06092203183255

Key Findings and Insights:

All three models achieved high R^2 scores, indicating robust predictive performance.

Outliers were detected and removed, leading to improved model performance.

The residual plot exhibited no clear patterns, suggesting that the models were well-fitted.

Suggestions for the next steps include exploring additional features and revisiting the analysis with different modelling techniques.

Conclusion:

Based on the evaluation metrics and insights gleaned from the analysis, the Simple Linear Regression model is recommended as the final model due to its high accuracy and explainability.

Next Steps:

Further analysing the impact of additional features on model performance.

Experiment with different modelling techniques, such as time series forecasting methods.

Exploring the use of ensemble methods to enhance predictive accuracy.