Data Analysis and Explanation Report

Data Analysis

The data used for analysis is cryptocurrency historical price and volume data. The data is loaded from CSV files and converted to a panda DataFrame. The 'Date' column is converted to a datetime object and set as the index.

The data is then split into training and testing sets. The training data includes all dates up to January 31, 2022, and the testing data includes all dates from February 1, 2017, to the end of the data.

The ARIMA model is then used to forecast the values for the testing data. The model is trained on the training data, and the predicted values are compared to the actual values using a line plot.

Additionally, a linear regression model is used to predict the adjusted close price of BTC based on the volume data. The volume data is first standardized using a StandardScaler, and the model is trained on the training data. The predicted values are then compared to the actual values using the R-squared score.

Results Interpretation

The ARIMA model performed relatively well in predicting the values for the testing data. The line plot shows that the predicted values follow the general trend of the actual values, but there are some differences in the magnitude of the values.

The linear regression model also performed relatively well with an R-squared score of 0.7126764 on the training data. However, it's worth noting that the volume data may not be the most relevant feature for predicting the adjusted close price of BTC.

Graphs:
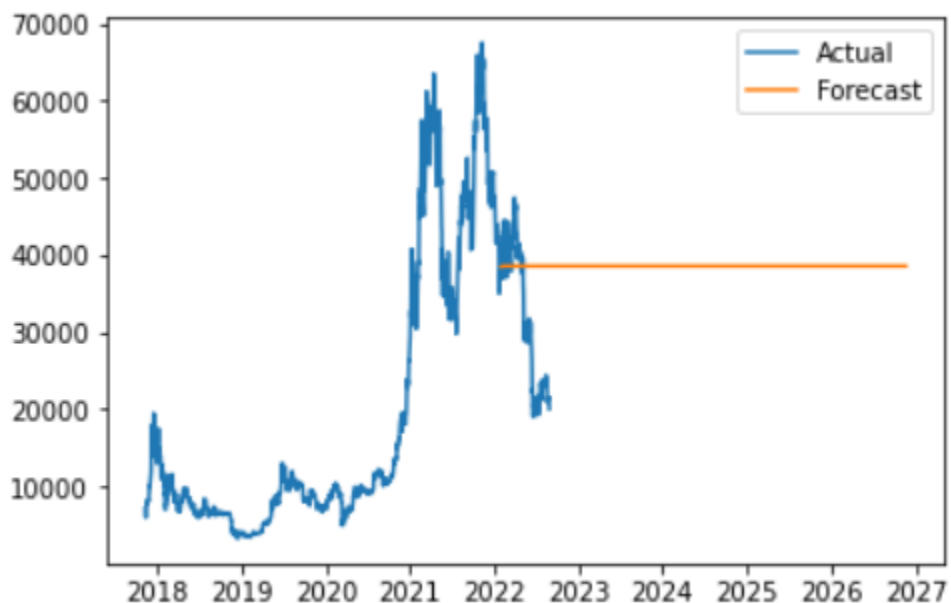
In the first plot, which shows the actual and predicted values of the test data, we can see that the predicted values (in orange) generally follow the same trend as the actual values (in blue). However, the predicted values appear to be smoother and less volatile than the actual values. This is likely due to the ARIMA model's ability to capture long-term trends in the data, but it may struggle to capture short-term fluctuations.
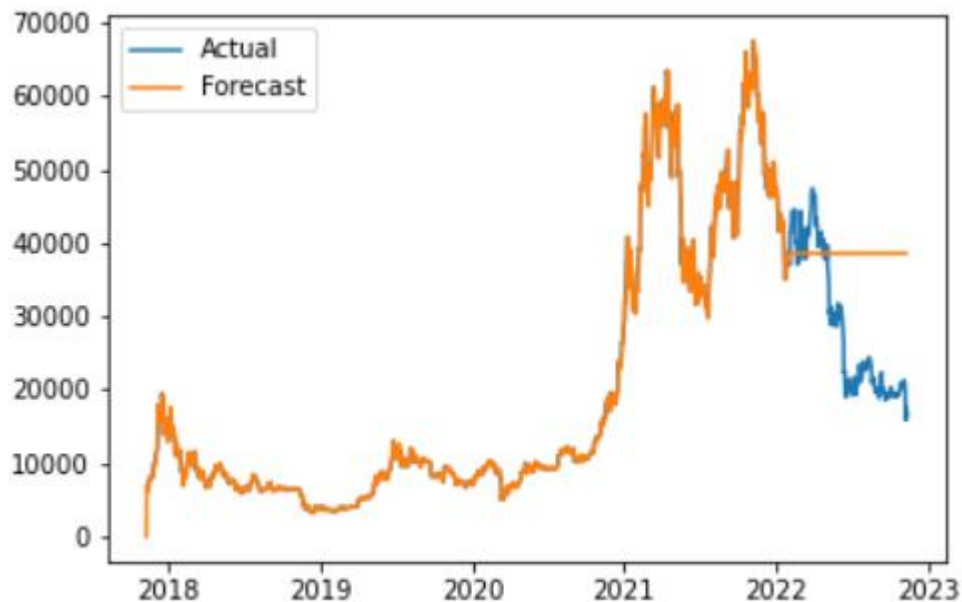
The second plot, which shows the actual and predicted values of the Close (BTC) variable in the test data, confirms this trend. We can see that the predicted values (in orange) follow the overall trend of the actual values (in blue), but they tend to be smoother and less volatile.

Overall, these plots suggest that the ARIMA model can capture the overall trend of the data, but it may struggle to accurately capture short-term fluctuations in the data. This is a common trade-off in time series analysis, where models that are better at capturing short-term fluctuations may be less accurate in predicting long-term trends, and vice versa. The R-squared score of 0.7126764 on the training data suggests that the model explains 71% of the variance in the data, which is a reasonable fit for a time series model.