**THE BULLS AND BEARS**

**REPORT**

**Amazon Price Predictions**



Team Members:

* Sammy Sifuna
* Jael Akech
* Denis Kipper
* Hellen Mwangi
* Waruchu Kuria
* Alan Omondi
* Tobias Ng’ong’a

# **1. Business Understanding**

## **1.1 Overview**

Stock price prediction involves the use of several techniques to forecast the near or far future of stock prices. Stock prices play an important role in financial markets and accurate forecasts can have positive significant implications for investors and financial institutions if carefully analyzed.

Stock prices analysis plays an important role in helping investors and financial institutions to make decisions in terms of investments, deciding when to buy and sell shares in order to make profit and minimize losses.

## **1.2 Problem Statement**

The primary challenge in predicting stock prices, especially in the short term, is the inherent volatility and unpredictability of the stock market. Investors and fund managers often face the problem of making optimal investment decisions. They need accurate forecasts to decide when to buy, sell, or hold Amazon's stock, but the accuracy of predictions can vary widely.

Bulls and Bears seek to develop a reliable predictive model that helps investors make informed decisions about their Amazon stock holdings to maximize returns while managing risks. This problem encompasses the need for accurate predictions and the application of these predictions to real-world investment strategies.

## **1.3 Business Objective**

### **1.3.1 General Objective**

To develop a robust stock price prediction model for Amazon stock market.

### **1.3.2 Specific Objectives**

1. To build and implement different models for Amazon stock price prediction.
2. To evaluate the performance and accuracy of the models using R2 Score, and RMSE.
3. To use the best performing model to forecast Amazon stock prices.
4. To create a user-friendly dashboard/application for stakeholders to access predictions.

## **1.4 Success Criteria**

The success criteria for the models are as follows:

1. Achieve a low RMSE and a high R2-Score to accurately forecast stock prices.
2. Usability of the model by investors and financial analysts.
3. Developing a user-friendly application for stock prices prediction.
4. Have models that accurately predict stock prices.

# **2. Data Understanding & EDA**

Data understanding is the process of gaining a comprehensive understanding of patterns and relationships in the data that will be used during a project. Exploratory Data Analysis (EDA) is the process of analyzing datasets thus summarizing their characteristics and as well as showing visualizations in order to be able to see and analyze the trends in the datasets for better decision making. EDA analyzes both the categorical and numerical variables.

## **2.1 Data Features**

The data has the following features:

1. **Date:** date of the stock price observation.
2. **Open** **price**: opening price of the stock on the given date.
3. **High** **price**: highest price of the stock on the given date.
4. **Low** **price**: lowest price of the stock on the given date.
5. **Close** **price**: closing price of the stock on the given date.
6. **Adjusted** **Close** **price**: closing price after adjustments for all applicable splits and dividend distributions
7. **Volume**: number of shares of the stock traded on the given date.

**Shape**

* It has 3,960 rows and 5 columns.

**Data Types**

* All the data is numerical as expected.

**Missing Values**

* There are no missing values.

**Checking for Duplicates**

* There are no duplicates in the data.

**Time Series Conformity**

* The Date column is already in DateTime format and is the index.

# **2.2 Univariate Analysis**

### **2.2.1 Distribution of Data Within Columns using Histogram Plots**

## 

We analyzed the individual columns/variables to see how the values are distributed and get patterns for each of the columns.

**Observations**

* The **Open, High, Low, Close, Adj Close** plots have similar distributions throughout the period under review (2008 to 2023).
  + They are trimodal, having three different peaks.
  + 0-25 US dollars is the most frequent price.
* Volume seems to be heavily distributed around 100 million to 200 million.
* All variables showed a skewed distribution to the right.
* No outliers are visible from these graphs.

### **2.2.2 Time Series Plots for Open, High, Low, Close, and Adj Close Columns**

**Observations**

* The plots above visualize the historical price trends over the period under review
* Similar seasonality and trend characteristics were observed and further analysis is done to confirm the seasonality of these variables
* The time series exhibits seasonality, with a gradual price increase over the years, marked by a significant surge between 2018 and 2022. However, there is a decline as the year 2023 approaches, followed by a renewed increase towards the end of 2023.

### **2.2.3 Time Series Plot for Volume Column**

**Observation**

* The Volume column also looks seasonal and with trend.

### **2.2.4 Checking for Outliers or Anomalies Using Box Plots**

**Observations**

* The price is centered at approximately 10 and 90 dollars for all columns.
* In all features no outliers except Volume.

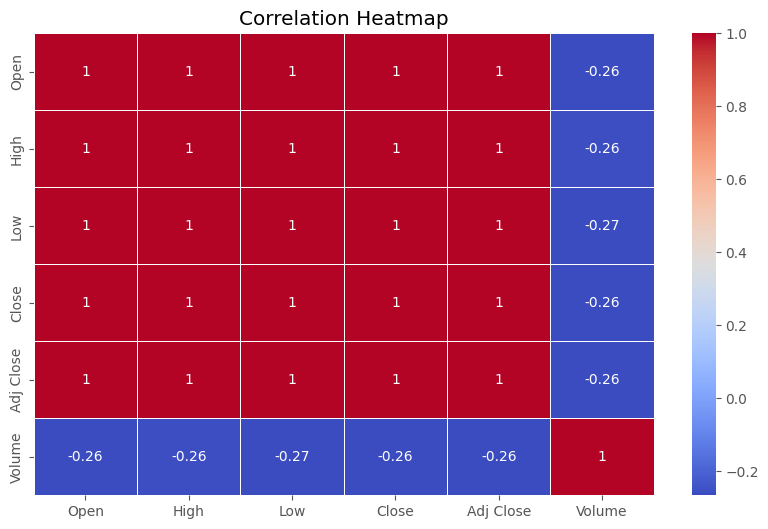
## **2.3 Bivariate and Multivariate Analysis**

### **2.3.1 Pairplots and Heatmap to Show Correlation**

### 

**Observations**

* The relationship among the variables is similar across the data.
* There is a linear relationship and very strong positive correlation among all features except with 'Volume'.
* Volume has non-linear relationship and weak negative correlation with all other variables.
* These observations are confirmed in the correlation heatmap below.

****

**External factors that may have caused the positive correlations can be explained as follows:**

1. Amazon stock price is driven by the overall performance of the company. When the company is performing well, its stock price tends to go up.
2. It is correlated with the overall performance of the stock market. When the stock market is doing well, Amazon's stock price tends to go up. This is because investors are more likely to buy risky assets, such as stocks, when the market is doing well.

**For the weak negative correlation with Volume:**

1. The volume of trading in Amazon stock is higher when the stock price is volatile. This is because investors are more likely to trade stocks when they are experiencing large price swings.
2. Volume of trading in Amazon stock is higher when there is a lot of news about the company. This is because investors may be more likely to buy or sell shares of the company based on news about its performance, products, or competitive landscape.

The state of the economy, interest rates, inflation and overall market sentiment are the general external factors that may have influenced the patterns visualized above.

# **3.Data Preparation**

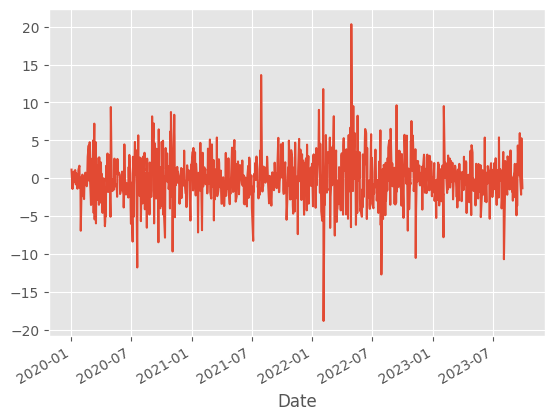
## **3.1 Feature Engineering**

Feature engineering is the process of extracting features from the data and transforming them into a format that is suitable for the machine learning model. In feature engineering we were identifying and selecting the most relevant and helpful data in the dataset we used.

The following features were added;

#### **3.1.1 Returns**

* A new column **Returns** is added to calculate the difference in price for two consecutive days.
* As seen in the plot, the column **Return** shows seasonality as the period progresses.



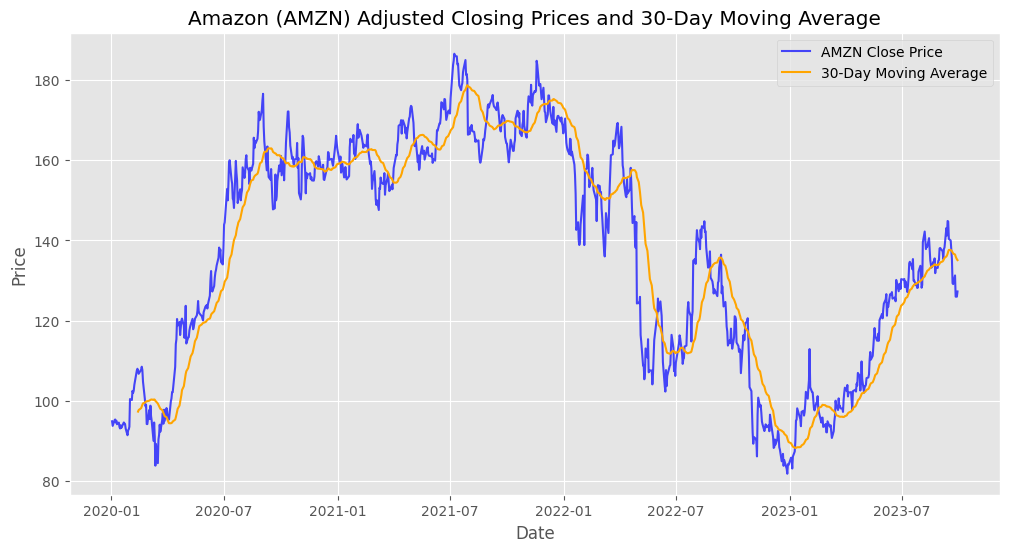
#### **3.1.2 Lag Features for Adj Close price**

* Creating lag features will capture the historical behavior of the stock prices.
* Lag features for the adjusted closing price have been calculated in the code below.

#### **3.1.3 Rolling Mean and Standard Deviation**

* The rolling mean and standard deviation columns serve to both smoothen the data and capture trends and fluctuations over a specific period, helping to identify and understand market dynamics.

#### **3.1.4 Moving Average**

* The 30-day moving average tracks the underlying trend in Amazon's closing stock prices over time, helping to smooth out short-term fluctuations and provide insights into longer-term price movements.

#### **3.1.5 Exponential Moving Average (EMA) Smoothing Factor**

* EMA can provide more weight to recent data points, which can be useful for capturing short-term trends.
* The small alpha value of 0.2 will result in a smoother EMA and higher sensitivity to recent price changes.

#### **3.1.6 Rate of Change (ROC) for Adj Close**

* This measures the percentage change in price over a specified period.
* Positive ROC values indicate upward momentum, while negative values indicate downward momentum.
* It helps traders and analysts identify potential trends or reversals in the price of an asset.

#### **3.1.7 Relative Strength Index (RSI)**

* It measures the speed and change of price movements.
* High RSI is normally placed as any value above 75% and a good indicator to sell.
* Low RSI is between 0 and 25%, a good indicator to buy.

#### **3.1.8 Average True Range (ATR)**

* A volatility indicator which measures the average range between the high and low prices over a specified period.
* It helps traders and analysts assess the level of volatility in the price movements of an asset.

## **3.2 Checking for Seasonality**

* First, we will check if there is seasonality and if it is statistically significant for each column in the time series.
* After checking for the seasonality, most of the columns were found to have seasonality
* We used the seasonal\_decompose function to remove seasonality for all the columns. This is important because if they are part of the time series, there will be effects in the forecast value

## **3.3 Checking for Stationarity**

* The **Augmented Dickey-Fuller (ADF)** test will be employed to assess the stationarity of the time series.
* The **ADF** test helps us assess whether a time series is stationary by comparing it to the null hypothesis that it has a unit root (meaning it's non-stationary).
* The test involves estimating the model's **co-efficients**, calculating a **test statistic**, and comparing it to **critical values** to determine whether the null hypothesis can be rejected.
* The **p-value** resulting from the test indicates the strength of evidence against the null hypothesis.
  + If the p-value is less than or equal to the significance level of 0.05, there's evidence to reject the null hypothesis, suggesting the series is stationary.
  + If the p-value is greater than the significance level of 0.05, there's weak evidence to reject the null hypothesis, indicating the series is likely non-stationary.
* After checking for stationarity, Most of the columns are not stationary apart from the following columns:
  + Volume
  + Returns
  + day\_of\_week
  + ROC
  + RSI
* We employed differencing to remove stationarity and make the model-building process accurate and reliable. The data's statistical properties will remain consistent, allowing for meaningful insights and accurate predictions.

## **3.4 Scaling**

This is the process of transforming and standardizing the data to make it more suitable for analysis and modeling. We used Min-Max Scaler to scale the data, making it easier to compare and analyse features. This leads to more accurate and reliable predictions.

## **3.5 Performing Train-Test Split**

The first step in building the model is to split the data into training and test sets. The training data is used by the machine learning algorithm to build a model while the test set is used to evaluate the model performance on unseen data. It is important to evaluate a model’s performance to guarantee that our metrics of success are met.

We used test size of 0.8 which splits 80% of the data to be used for training and the remaining 20% to be used for testing.

# **4. Modelling**

## **4.1 Baseline Model - SARIMAX-1**

For this project, we focus on the Adjusted Close prices which provides a more accurate representation of the stock values over time. The column of the Adjusted Close is therefore used in modeling as the target variable. The project uses the SARIMA model as the baseline model. This was done using the default values of p,d,q values which is 1,1,1 respectively.

The baseline model showcases a commendable level of accuracy as reflected in the Mean Absolute Error (MAE) of 0.0506, the Mean Squared Error (MSE) of 0.0047, and the Root Mean Squared Error (RMSE) of 0.0683. These relatively low error values indicate that the model's forecasts are close to the actual values, underscoring its capacity to provide reliable predictions.

The model however has a higher Mean Absolute Percentage Error (MAPE) value of 9.6815%, which shows that the model's predictions exhibit a 9.6815% relative error from the actual values. This may significantly affect the forecasting accuracy. The R-squared (r2) score is 0.0016, indicating a very low level of accuracy.

We did hyperparameters tuning using GridSearch-CV to get the optimal values of p,d,q and proceeded to do the SARIMA-2 model using these values.

## **Hyperparameters Tuning Using GridSearch CV**

Hyperparameter tuning using GridSearch-CV was done to find the optimal values of hyperparameters (p,d,q) to be used during modeling. The result were as follows:

* Best SARIMAX Order (p, d, q): (2, 0, 2)
* Best SARIMAX Order (p, d, q, s): (2, 0, 2, 5)

The optimal hyperparameters found(p,d,q,s) were used in SARIMA - 2 and the performance of the model was assessed.

## **4.2 SARIMAX-2**

We used the p,d,q,s generated from the GridSearch CV for SARIMAX -2 Modeling to evaluate the improvement of our model.

In our evaluation of the SARIMAX -2 model's performance, we found that it excels in multiple critical aspects. The model's ability to capture the nuances of the data is evidenced by its substantially lower AIC (-3523.004) and BIC (-3474.821) values, underscoring its effectiveness in modeling the underlying patterns and variations in the dataset. Additionally, the higher Log Likelihood (1770.502) signifies that the model is exceptionally well-suited to the data, providing strong support for its accuracy.

The model shows impressive predictive accuracy, as indicated by the relatively low error metrics. The Mean Absolute Error (MAE) of 0.0506, Mean Squared Error (MSE) of 0.0047, and Root Mean Squared Error (RMSE) of 0.0683 collectively highlight the model's capacity to make highly accurate forecasts, with predictions closely aligned with actual values. The Mean Absolute Percentage Error (MAPE) at 9.6815% further corroborates this, suggesting that the model's relative errors are limited, signifying a robust predictive performance.

The model's explanatory power is relatively weak, as indicated by the low R-squared (R2) score of 0.0025. While the model in forecasting accuracy, its ability to elucidate the underlying data dynamics is limited.

This SARIMAX -2 model indicates precise and reliable predictions, making it a valuable model for forecasting.

## **4.3 FBProphet**

We went ahead to do Facebook Prophet forecasting. In the evaluation of our FBProphet model's performance, we observed a combination of strengths and weaknesses. The Mean Absolute Error (MAE) indicates that, on average, the model's predictions exhibit a moderate error of approximately 0.0551 units, while the Mean Absolute Percentage Error (MAPE) stands at 10.5141%. These suggest that the model's forecasts are reasonably accurate, with predictions demonstrating a relatively minor deviation from the actual values.

The R-squared (R2) score for the test set forecasts is negative, registering at -0.0138. The negative value indicates that the model struggles to explain the variance in the data, suggesting a limited ability to capture the underlying patterns in the time series.

The FBProphet model results exhibits a reasonable level of predictive accuracy, making it a suitable choice for forecasting where a moderate level of error is tolerable. However, the presence of relatively large squared errors and a negative R2 score emphasize that the model may not be the ideal choice for applications that demand precise forecasts with minimal error.

We decided to explore RNN modeling and evaluate its performance

## **4.4 RNN (Recurrent Neural Network)**

A Simple Recurrent Neural Network (RNN) which is a foundational deep learning architecture designed for processing sequential data was used.

The evaluation of the Recurrent Neural Network (RNN) model's performance showed a moderate level of predictive accuracy, as indicated by a Mean Absolute Error (MAE) of 0.0779 and a Mean Absolute Percentage Error (MAPE) of 14.7780%. It has relatively large squared errors, reflected in a Root Mean Squared Error (RMSE) of approximately 0.1033. The negative R-squared (R2) score of -0.0471 indicates that the model struggles to explain the variance in the data, suggesting a limited ability to capture the underlying patterns in the time series.

The RNN model demonstrated a moderate level of accuracy, we explored LSTM modeling techniques.

## **4.5 LSTM (Long Short-Term Memory)**

### **4.5.1 LSTM (Using Original Features)**

We did the first LSTM modeling using the original features. The model had a Mean Absolute Percentage Error (MAPE) of 23.1175%, indicating that, on average, the model's predictions deviate by approximately 23.1175% from the actual values which is greater than the values for both SARIMAX, FBProphet, and RNN models. This suggests that the model’s predictions may not be consistently accurate. The Mean Absolute Error (MAE) is calculated at 0.0779, and the Mean Squared Error (MSE) at 0.0133. The Root Mean Squared Error (RMSE) is approximately 0.1153, reflecting that the magnitude of the model's errors can be significant. The R-squared (R2) score for the test set forecasts is negative, at -0.0025.

We concluded that the LSTM model using the original features demonstrated a degree of predictive accuracy, but may not capture the underlying patterns of the time series data.

We therefore decided to train another LSTM model using the 8 most important features and identified bu XGBoost

### **4.5.2 LSTM (Using 8 Most Important Features)**

The LSTM (Long Short-Term Memory) model using the 8 important features showed a better Mean Absolute Percentage Error (MAPE) of 14.9543%. This indicates that, on average, the model's predictions deviate by approximately 14.9543% from the actual values which is better as compared to the results of LSTM model using the original features. This metric suggests a moderate level of error in the forecasts, signifying that the model's predictions may exhibit some inconsistencies.

The Mean Absolute Error (MAE) is calculated at 0.0779, and the Mean Squared Error (MSE) at 0.0103. The Root Mean Squared Error (RMSE) is approximately 0.1013, reflecting the magnitude of the model's errors. Furthermore, the R-squared (R2) score for the test set forecasts is negative, at -0.0036. This shows that the model has a restricted ability to capture the underlying patterns within the time series data.

This LSTM model demonstrates a moderate level of predictive accuracy, but may not be the best alternative modeling technique to achieve more precise and reliable forecasts.

## **4.6 Evaluation of Models**

| **Model** | **MAE** | **MSE** | **RMSE** | **MAPE** | **R2-Score** |
| --- | --- | --- | --- | --- | --- |
| Baseline Model-SARIMA-1 | 0.0506 | 0.0047 | 0.0683 | 9.6815 | 0.0016 |
| SARIMA-2 | 0.0506 | 0.0047 | 0.0683 | 9.6815 | 0.0025 |
| FB Prophet | 0.0551 | 0.0058 | 0.0760 | 10.5142 | -0.0138 |
| Simple RNN | 0.0779 | 0.0107 | 0.1033 | 14.7780 | -0.0471 |
| LSTM-Original Features | 0.0779 | 0.0133 | 0.1153 | 23.1175 | -0.0025 |
| LSTM-Important Features | 0.0779 | 0.0103 | 0.1013 | 14.9543 | -0.0036 |

From the above model evaluation table we concluded as follows;

* **SARIMA-2** appears to be the best-performing model among the options listed. It has the lowest MAE, MSE, and RMSE, indicating that it provides the most accurate predictions with relatively low error rates. The MAPE is also below 10%, suggesting that, on average, its predictions are within 10% of the actual Amazon stock prices. It has the highest R2-Score of 0.0025.
* **FB Prophet**, while having a low MAPE, MAE, MSE, and RMSE but the values are higher as compared to SARIMA-2. It also has higher MAPE which is above 10%. This indicates that it might not be as accurate as SARIMA-2 in predicting Amazon stock prices.
* The **Simple RNN** model performs lower than FB Prophet in terms of MAE, MSE, and RMSE, it is not as accurate as SARIMA-2 and FB Prophet, as it has higher error rates.
* The **LSTM with original features** model with original features has lowest performance as compared to SARIMA-2, FB Prophet and Simple RNN, with slightly higher MAE, MAPE and RMSE. It has the highest MAPE of 23.1175%.
* The **LSTM model with important features** has a better performance than the LSTM with original features.. However, it still does not perform as well as SARIMA-2, FB Prophet, and Simple RNN in terms of accuracy.

In summary, the **SARIMA-2** model outperforms the other models in predicting Amazon stock prices, as it has the lowest MAE, MSE, and RMSE along with a reasonably low MAPE. It is also the only model which recorded the highest positive R-Squared Score value. The other models have higher errors and are less accurate in comparison. We proceeded to deploy the **SARIMA-2** Model

# **5. Deployment**

We created and deployed a web application on Streamlit. The app uses the chosen SARIMA model to train and predict stock prices using user-provided datasets.

# **6. Conclusion**

From our time series analysis SARIMA-2 model performed the best with MAE score of 0.0506, MSE score of 0.0047 RMSE score of 0.0683 and MAPE score of 9.6815 compared to other models we used which were:

1. FB Prophet
2. Simple RNN
3. LSTM-original features
4. LSTM-Important features

SARIMA-2 model performed well for short term predictions however long term predictions brought wide variations.

The top 8 features which highly influenced the price predictions in the amazon stock market are: Close, High, Returns, Rolling\_Std, Volume, Open, ROC, RSI.

# **7. Limitations**

1. Time series is an intensive machine learning models and hence it required more time for us to come up with optimum parameter and hyper-parameters for the model to perform much better which is a great constrain.
2. Stock prices are affected other external factors such as war and disease which are hard to predict and also influence stock prices.

# **8. Recommendations**

1. Investors and financial instittion can use the model for short term prediction of stock market prices to determine the general trend of the amazon prices. However other factor such as fundamental analysis need to be consider before making the final decision.
2. Our deployment model can also be improved and used for predicting other stock markets other than amazon stocks only.
3. For better performance of the LSTM model, more data is required for analysis. More data will enhance the model’s ability to recognize patterns and trends.
4. Carry out sentimental analysis alongside the model to factor in the impact of news and public sentiment on stock prices changes. This analysis will provide valuable contextual information.
5. Experiment with different train-test split ratios to evaluate how the model’s performance is affected by the division of data. This will help determine the optimal balance between the training and testing data that gives better performance of the model.
6. Coming up with a model that can predict other stock markets, not just Amazon. This will provide valuable insights from diverse stock markets to the investors.

# **9.Future Work**

1. Develop a more efficient techniques to optimize the parameters and hyperparameters of time series models like LSTM, reducing the time and resources needed for model fine-tuning.
2. Explore ways to better incorporate external factors such as geopolitical events (e.g., wars) and unforeseeable occurrences (e.g., pandemics) into predictive models, acknowledging their inherent unpredictability and understanding their influence on stock prices.
3. To boost LSTM model performance, there's room for improvement in collecting and integrating a wider array of data sources, including financial news sentiment data and economic indicators, to offer a more comprehensive view of market dynamics.
4. Expanding the model's coverage beyond Amazon to encompass various stock markets requires the development of multi-market models, enabling investors to make diversified, data-driven decisions across different asset classes.

### 