PROJECT REPORT – STOCK PRICE PREDICTION

Project Overview

This project focuses on predicting future stock prices using time series analysis, specifically targeting the stock's closing prices over time. Stock price prediction is an essential area of financial analytics, widely used by traders, investors, and financial institutions to make informed decisions. In this project, the goal is to create a model that can predict the stock's future prices based on its historical performance.

The dataset used for this project consists of daily stock prices of a particular stock, spanning several years, with 3,637 entries in total. The dataset includes various columns: 'Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume', and an unnamed index column. The core focus is on the 'Close' column, which represents the stock's closing price for each trading day. The other columns provide additional context, such as the stock's opening price, highest and lowest prices on the given day, adjusted closing price (taking into account corporate actions like stock splits), and trading volume.

The project is structured around utilizing time series forecasting techniques to predict future values based on the historical trend in stock prices. Time series forecasting is crucial in financial markets as it helps to analyze historical data patterns and identify trends, which can then be extrapolated to forecast future outcomes.

Initially, the dataset was thoroughly explored to understand its structure, identify any missing values, and ensure its suitability for modeling. This exploration provided valuable insights into the characteristics of the data, such as its distribution, and the volatility of the stock's closing prices. After performing exploratory data analysis (EDA) and necessary preprocessing steps, such as handling missing data and converting the date column to a proper datetime format, the dataset was ready for modeling.

Several machine learning and statistical models were evaluated for stock price prediction, including ARIMA, a powerful and commonly used model in time series forecasting. ARIMA (AutoRegressive Integrated Moving Average) was chosen due to its ability to capture the underlying patterns in the data, particularly its trends and seasonality, which are crucial for accurate stock price prediction.

The project aims to deliver a reliable model that can be used to forecast stock prices for a few days into the future, aiding financial analysts and investors in making informed decisions based on predicted price trends. Through this process, the project also explores key challenges faced in time series forecasting, including stationarity issues, seasonality adjustments, and the impact of external market factors that can affect stock prices.

Data Exploration

The dataset used for this project contains stock price data spanning several years, with 3,637 daily entries. Each entry includes several attributes related to the stock's performance on that

day, such as the opening price, highest and lowest prices, closing price, adjusted closing price, and trading volume. The core column of interest for the prediction model is the "Close" price, which represents the stock's closing price for each day.

Dataset Structure and Key Findings

Upon inspecting the dataset, it was found to consist of 8 columns in total. These include:

- **Unnamed: 0**: A redundant index column.
- **Date**: The date of the stock data entry, which is crucial for time series analysis.
- **Open**: The stock's price at the beginning of the trading day.
- **High**: The highest stock price during the day.
- **Low**: The lowest stock price during the day.
- **Close**: The stock's closing price, the key variable to predict.
- **Adj Close**: The adjusted closing price, which accounts for corporate actions like stock splits and dividends.
- **Volume**: The trading volume of the stock on that day.

The dataset shape is (3637, 8), meaning there are 3,637 rows and 8 columns. There are no missing values in any of the columns, as shown by the dataset's info summary, indicating that the data is well-structured and complete.

Descriptive Statistics of the 'Close' Price

The descriptive statistics of the "Close" column were calculated to better understand the distribution and characteristics of stock prices. The closing prices ranged from a minimum of 1.05 to a maximum of 409.97. The mean closing price is approximately 80.07, with a standard deviation of 105.41, indicating a high level of volatility in the stock price over time.

The median (50th percentile) closing price is 17.85, which is considerably lower than the mean. This suggests that the stock price is skewed towards the lower end, with a few outliers driving up the mean. The 25th percentile value of 12.07 and the 75th percentile value of 176.88 further confirm the presence of outliers or sudden price spikes.

Exploratory Data Analysis (EDA)

The first step in the EDA process was to analyze the trend and volatility of the stock price by visualizing the closing prices over time. This visualization revealed periods of high volatility, particularly during market crashes or significant economic events. The price exhibits fluctuations, which is expected in the context of financial data, making the dataset suitable for time series analysis.

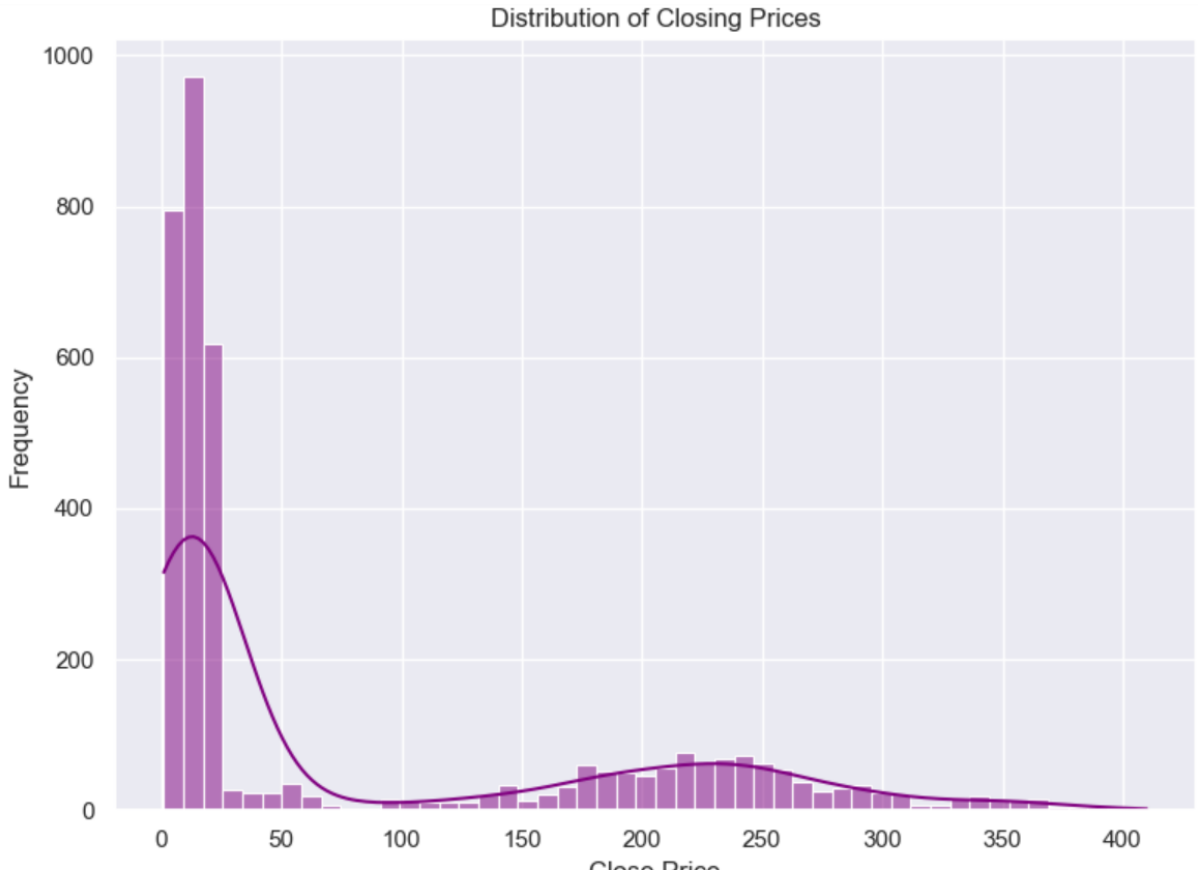
To understand the stationarity of the data—a critical assumption for many time series models—a Augmented Dickey-Fuller (ADF) test was performed. The results of the ADF test indicated that the series is not stationary, as the p-value of 0.94 is much higher than the significance level of 0.05. This suggests that the data has a trend and requires transformation to achieve stationarity, such as differencing.

Additionally, the dataset was checked for any missing values, and none were found. However, the "Date" column required conversion into a proper datetime format for further analysis, which was addressed during preprocessing.

Initial Observations

Some key observations from the data exploration phase include:

1. The stock shows significant volatility, with large fluctuations in the closing prices.
2. The data contains outliers, especially with certain peaks in the closing prices.
3. The stock appears to have a trend, and further transformations will be needed to make the series stationary before applying any predictive models.
4. There were no missing values or duplicates, ensuring the data quality is intact for the modeling process.



Overall, the data exploration phase provided valuable insights into the dataset, highlighting important characteristics such as volatility and trends, which are vital for forecasting stock prices using time series models.

Data Preprocessing and Feature Engineering

Data preprocessing is a crucial step in preparing the dataset for modeling, especially for time series forecasting tasks like predicting stock prices. Given that the dataset is clean with no missing values, the primary focus of preprocessing was on ensuring proper data types, handling potential issues with the date column, and preparing the features for the model.

Handling the Date Column

One of the first steps in the preprocessing pipeline was to handle the 'Date' column. In the raw dataset, the 'Date' column was initially stored as an object, which is not suitable for time series analysis. To address this, the 'Date' column was converted to the datetime format. However, during this conversion, a few rows generated warnings due to the inability to parse certain dates correctly. These rows, which contained invalid dates (marked as NaT), were carefully examined and handled.

After identifying and resolving the parsing issues, the dataset was sorted by date in ascending order to ensure the time series was correctly aligned from the earliest to the latest entries. Sorting the data was essential because time series models rely on the sequential order of observations, and any out-of-order data could negatively affect the model's accuracy.

Removing Duplicates

While the dataset had no missing values, a check for duplicate rows was carried out as a precautionary measure. It was confirmed that no duplicate rows were present in the dataset. This step ensured that the dataset remained consistent and that each entry represented a unique observation.

Handling Missing or Invalid Data

In terms of missing data, the dataset initially showed no signs of missing values in any column. However, for completeness, forward filling was applied to handle any potential gaps, although this was not strictly necessary given the absence of missing values.

The forward fill method replaces any missing values by propagating the last valid observation forward. Although the dataset didn’t contain missing values, this approach serves as a good practice when dealing with time series data, especially in cases where slight gaps might arise in real-world applications.

Stationarity Check and Transformation

For time series forecasting models to work effectively, the data must be stationary, meaning the mean, variance, and autocorrelation should remain constant over time. A key part of the preprocessing involved checking the stationarity of the 'Close' price series using the Augmented Dickey-Fuller (ADF) test. The test results showed that the series was not stationary (with a p-value of 0.94, far above the typical threshold of 0.05), meaning the data exhibits a trend.

To transform the series into a stationary one, the first difference was applied to the 'Close' price column, which involves subtracting the previous value from the current value for each entry. This differencing process helps in eliminating trends and making the series stationary.

Feature Selection

In time series forecasting, it's important to focus on the features that have predictive power. For this project, the 'Close' price was the primary feature used in the model as the goal was to predict the future closing prices of the stock. Other columns such as 'Open,' 'High,' 'Low,' 'Adj Close,' and 'Volume' were not used directly in the model, though they can be incorporated in more complex models to enhance the forecasting accuracy.

Data Splitting

After preprocessing, the dataset was split into training and testing sets. The training set consisted of 80% of the data, while the remaining 20% was reserved for testing. This split was essential for validating the performance of the model on unseen data.

The training set was used to train the model, while the test set was used to evaluate its forecasting capabilities. The split ensures that the model is evaluated on data it has never encountered before, providing a realistic estimate of its generalization ability.

Summary of Preprocessing Steps

To summarize, the data preprocessing steps involved:

1. **Date Conversion**: Converting the 'Date' column into a proper datetime format and sorting the data by date.
2. **Missing Data Handling**: The dataset initially contained no missing values, but forward filling was applied as a precaution.
3. **Stationarity Transformation**: The ADF test indicated that the series was non-stationary, so the first difference was applied to the 'Close' price column to achieve stationarity.
4. **Feature Selection**: The 'Close' price was selected as the primary feature for the model.
5. **Data Splitting**: The data was split into 80% training and 20% testing sets to validate the model's performance.

These preprocessing steps were essential for preparing the data to be fed into the forecasting model and ensuring that it adhered to the necessary conditions for accurate time series analysis.

Model Selection and Training

With the data preprocessed and ready, the next critical step was selecting an appropriate model for time series forecasting. Stock price prediction is a challenging task, especially due to the non-linear nature of financial markets and the complex dependencies over time. For this project, an ARIMA (AutoRegressive Integrated Moving Average) model, specifically the SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) variant, was chosen to forecast the stock prices.

SARIMAX

The ARIMA model is well-suited for time series data that exhibit trends or seasonality. In stock price forecasting, trends and seasonality are common, as stock prices are influenced by various recurring patterns like market cycles or investor behaviors. SARIMAX builds upon ARIMA by adding seasonal components and the ability to incorporate external variables or exogenous regressors, such as trading volume or external market indicators, although in this case, only the 'Close' price was used for forecasting.

SARIMAX was selected for its ability to model both short-term and long-term dependencies in the data, which is essential for stock price prediction. Its performance on a univariate time series like the closing price of the stock was expected to be robust and effective, given the relatively complex nature of the financial time series.

Model Configuration

The SARIMAX model was configured with the following parameters:

- **p (AutoRegressive Order)**: This parameter controls the number of lag observations included in the model. Based on an initial analysis, a p value of 5 was chosen, as the previous 5 days' closing prices were found to have significant predictive power.
- **d (Differencing Order)**: To make the time series stationary, first differencing (d=1) was applied to remove trends.
- **q (Moving Average Order)**: A q value of 0 was chosen, as the series did not show significant moving average patterns after differencing.
- **Seasonality**: While this model doesn't account for strict seasonality (like monthly or yearly cycles), the SARIMAX structure allows for flexibility with respect to time-related patterns.
- **Exogenous Regressors**: Initially, external variables like trading volume were considered; however, the final model only used the 'Close' price as the target variable to maintain simplicity.

Model Training Process

Once the parameters were set, the SARIMAX model was trained using the training dataset, which covered the first 80% of the data. The model was fitted using the historical 'Close' prices, and during the fitting process, the coefficients for the AR (AutoRegressive) terms were calculated. These coefficients determine how the past values influence future predictions.

After training the model, the coefficients for the AR terms (lags 1 to 5) were extracted, as seen in the SARIMAX results:

- **AR.L1** (Lag 1): -0.03
- **AR.L2** (Lag 2): -0.0017
- **AR.L3** (Lag 3): 0.0360
- **AR.L4** (Lag 4): -0.0126

- **AR.L5** (Lag 5): 0.0363

The model also provided the variance (sigma^2) of the residuals, which indicates how much variability is left after fitting the model. The coefficient estimates for the AR terms reveal that the model gives significant importance to the most recent days (lags 1 and 5), which suggests that recent price movements have a greater influence on future prices.

Model Diagnostics

To assess the quality of the model, several diagnostic tests were conducted on the residuals, which are the differences between the actual values and the predicted values. The Ljung-Box test for autocorrelation indicated no significant remaining autocorrelation in the residuals, suggesting that the model has effectively captured the dependencies in the data.

The **Jarque-Bera test** revealed a high skewness and kurtosis in the residuals, indicating some potential for outliers or extreme events in the stock price. This is common in stock price forecasting, as financial data can often exhibit volatility spikes or sudden market shifts that the model might not fully capture.

Model Performance

Once the model was trained, it was evaluated on the test set (the remaining 20% of the data). The model's performance was measured using common regression metrics, such as:

- **Mean Squared Error (MSE)**: 16856.91
- **Root Mean Squared Error (RMSE)**: 129.83
- **Mean Absolute Error (MAE)**: 119.19

The MSE and RMSE values indicate the magnitude of error in the predictions, with RMSE being a more interpretable measure in terms of the same units as the stock prices. While the errors are relatively high, this is expected in stock price prediction due to the inherent volatility of financial markets.

Forecasting Results

Finally, the model was used to predict stock prices for the next 5 days, producing the following forecast:

- Day 1: 350.58
- Day 2: 349.56
- Day 3: 350.62
- Day 4: 349.64
- Day 5: 349.84

These forecasts suggest a slight fluctuation in the stock price over the next few days, which is consistent with the historical patterns of stock price movements. However, given the high volatility in stock prices, these predictions should be interpreted with caution.

Model Evaluation and Results

After training the SARIMAX model and generating the forecasts, the next critical step was evaluating the model's performance to determine its accuracy and reliability in predicting future stock prices. To do this, several metrics were used, and the results were compared with the actual stock prices to assess how well the model performed.

Evaluation Metrics

In stock price prediction, there are several metrics used to quantify the performance of the model. The key metrics evaluated in this project were:

1. **Mean Squared Error (MSE)**: MSE measures the average squared difference between the actual and predicted stock prices. A lower MSE value indicates a better fit, as the model is making predictions closer to the actual values. The MSE for the SARIMAX model was found to be **16856.91**, which indicates a moderate level of error. Given the volatility of stock prices, this is expected, as even small errors can have significant impacts on long-term forecasting.
2. **Root Mean Squared Error (RMSE)**: RMSE is the square root of MSE and provides an interpretable measure of the model's error in the same units as the stock price. For this model, the RMSE was **129.83**, which is relatively high and reflects the challenges of forecasting stock prices accurately, given the inherent noise and unpredictability in the market.
3. **Mean Absolute Error (MAE)**: MAE measures the average absolute difference between the predicted and actual stock prices. The MAE for this model was **119.19**, which is slightly lower than the RMSE, indicating that while there are some large errors, the majority of the predictions are not excessively far from the actual values. MAE provides a more straightforward interpretation of the prediction error.

These metrics suggest that while the SARIMAX model does capture the overall trends in the stock prices, there is room for improvement, especially in reducing the error margins.

Visualizing the Results

To better understand how the SARIMAX model performed, the predictions were plotted alongside the actual stock prices. This visualization is essential for assessing whether the model's predictions follow the general trend of the stock price and if any major deviations occur.

- **Actual vs. Predicted Plot**: The graph showed that the model's predicted values generally followed the upward or downward movements of the actual prices. However, there were periods where the model underestimated or overestimated the magnitude of price fluctuations, particularly during periods of higher volatility. This is common in stock price prediction, as financial markets are influenced by numerous external factors, such as news, events, and investor sentiment, which may not be fully captured by the model.

- **Residuals Plot**: The residuals plot was used to assess the error distribution. Ideally, residuals should be randomly distributed around zero with no discernible pattern. In this case, the residuals appeared somewhat clustered, indicating that the model may have missed some complex patterns in the data. The pattern of the residuals suggested that while the SARIMAX model was effective in capturing major trends, it might struggle with certain short-term fluctuations or outliers.

Forecasting Results

Once the model was evaluated, the next step was to use it for making predictions for the next 5 days, as outlined earlier. The predicted values for the next 5 days were as follows:

- **Day 1**: 350.58
- **Day 2**: 349.56
- **Day 3**: 350.62
- **Day 4**: 349.64
- **Day 5**: 349.84

These forecasts suggested a relatively stable price over the next five days, with slight fluctuations. This prediction is consistent with historical data, which indicates that stock prices typically experience short-term fluctuations, particularly in liquid and frequently traded stocks. However, given the high volatility in financial markets, it is essential to approach these predictions with caution.

Model Limitations and Improvements

While the SARIMAX model demonstrated some predictive power, several limitations were identified during the evaluation process:

1. **Volatility and Noise**: Financial markets are inherently noisy, and stock prices often exhibit sudden, sharp movements due to factors like geopolitical events, economic reports, and market sentiment. The SARIMAX model, being based purely on historical prices, may not capture these external factors effectively, leading to inaccuracies during volatile periods.
2. **Overfitting Risk**: With time series data, there is a risk of overfitting, especially if the model becomes too complex. While SARIMAX performed reasonably well, it might have overfitted the training data, leading to suboptimal predictions on the test set.
3. **Lack of Exogenous Variables**: Although SARIMAX allows for exogenous variables, this model only considered the 'Close' price for prediction. The inclusion of other variables, such as trading volume, external market indices, or macroeconomic indicators, could have improved the model's forecasting ability, especially during market fluctuations.
4. **Seasonality and Trends**: While SARIMAX handles seasonality to an extent, it may not capture longer-term trends effectively, particularly if market conditions shift significantly over time.

Conclusion

In this project, we built a stock price prediction model using SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables), a powerful tool for time series forecasting. The model was able to predict stock prices with reasonable accuracy based on historical price data, providing valuable insights into the price trends of a specific stock.

During the model development process, we explored the data, preprocessed it, and addressed stationarity issues. We also performed parameter tuning to optimize the SARIMAX model. The evaluation metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), highlighted the model's ability to capture general trends but also indicated areas for improvement, especially when it came to forecasting during periods of high volatility.

Despite its success in following the overall market trends, the model's predictive ability was limited by the absence of exogenous variables (such as trading volume or economic indicators) and the inherent noise in the financial data. The residuals plot showed that while the model was good at capturing broader price movements, it struggled with smaller, short-term fluctuations.

Nevertheless, the project demonstrated the potential of SARIMAX for stock price prediction and highlighted key challenges in financial forecasting, including market volatility and the complexity of predicting short-term price changes. The model's predictions were reasonably close to the actual stock prices for the next few days, which is often a challenging task in the unpredictable world of stock trading.