Technical Report: Evaluation and Recommendation of Stock Price Prediction Models

Project Overview

This project aims to develop machine learning models to predict future Apple stock prices based on historical data from 2014 to 2023.

Project Stages

• Data Collection: Scripts and resources used to gather the data.

• Data Cleaning and Preprocessing: Methods applied to prepare the data for modeling.

• Modeling: Development and tuning of the predictive model.

• Evaluation: Metrics and techniques used to evaluate the model's performance.

Data Collection

Source: Historical Apple stock price data was collected yahoo finance via API.

Tools: Python scripts were used to automate the data collection process.

Data Processing

The data processing activities undertaken for this project included several key steps to ensure the accuracy and readiness of the data for time series analysis and model development:

1. **Selection of Price Column:** The dataset contained two types of stock price data: "Close" and "Adjusted Close" (Adj Close). The "Adj Close" column was chosen for the analysis as it accounts for corporate actions like dividends and stock splits, offering a more accurate reflection of the stock's value over time.

2. **Data Transformation:** The "Adj Close" column was renamed to "Price" to simplify references in the analysis. Additionally, the dataframe was filtered to include only the "Date" and "Price" columns, as these are the essential variables required for the time series analysis and model development in this project.

3. **Check for Missing Data:** An examination was conducted to identify any missing values within the dataset. It was confirmed that there were no missing data points, ensuring the completeness of the dataset for accurate analysis.

4. **Data Type Checks:** The data types of the columns were verified to ensure compatibility with the analytical methods used. It was determined that all columns contained the appropriate data types necessary for conducting time series analysis and model development.

5. **Data Manipulation:** Given that stock markets are closed on weekends and statutory holidays, adjustments were made to account for these non-business days. The stock price data from the most recent market day was carried forward to fill the dates corresponding to weekends and statutory holidays, maintaining continuity in the data for analysis purposes.

Predictive Models

This report details the evaluation of three predictive models developed to forecast the future prices of Apple stock, using historical stock price data from 2014 to 2023. The models evaluated include:

1. Holt Winters Model
2. SARIMAX Model (Order: (0,1,2))
3. SARIMAX Model (Order: (0,1,2) and Seasonal Order: (1,1,1,252))

Split Dataset into training and test sets

The dataset was divided into a training set (75%) and a test set (25%) to train and evaluate the models.

Figure 1: Plot of training and test datasets and splitting the historical dataset

The test dataset summary statistics are as follows:

```
test_data.describe()
```

	price
count	652.000000
mean	159.914230
std	17.486788
min	123.998459
25%	145.822430
50%	157.942818
75%	173.118259
max	197.589523

Figure 2: Test dataset summary statistics

Evaluation Metrics

The models were evaluated using the following statistical metrics:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- Directional Accuracy
- Mean of Predictions vs. Mean of Test Data

Model Performance Summary

The model performance is summarised in **Table 1** below.

Table 1: Models Performance Summary based on quantitative statistical Measures

Model	RMSE	MAE	MAPE	Directional Accuracy	Mean of Prediction	Mean Of Test data
Model 1	31.5025	28.3038	17.77%	48.00%	172.7274	159.9142
Model 2	15.6347	13.041	8.12%	50.46%	154.2093	159.9142
Model 3	24.6923	20.9097	13.41%	50.46%	174.2503	159.9142

Result Analysis:

Model 1 - Holt Winter Model

Figure 3: Plot showing Model 1, training and test dataset prices

The plot analysis of the Model 1 against the training and test datasets as shown in **Figure 3** reveals initial underperformance relative to the test data. As the model progresses, however, it begins to mimic the pattern of the test data more closely, albeit with a tendency to overestimate prices. Despite this overestimation, the directional movement of the model aligns with the test data, consistently indicating an uptrend. This suggests that, over time, the model accurately reflects the general increase in stock prices.

Model 2- SARIMAX of Order (0,1,2)

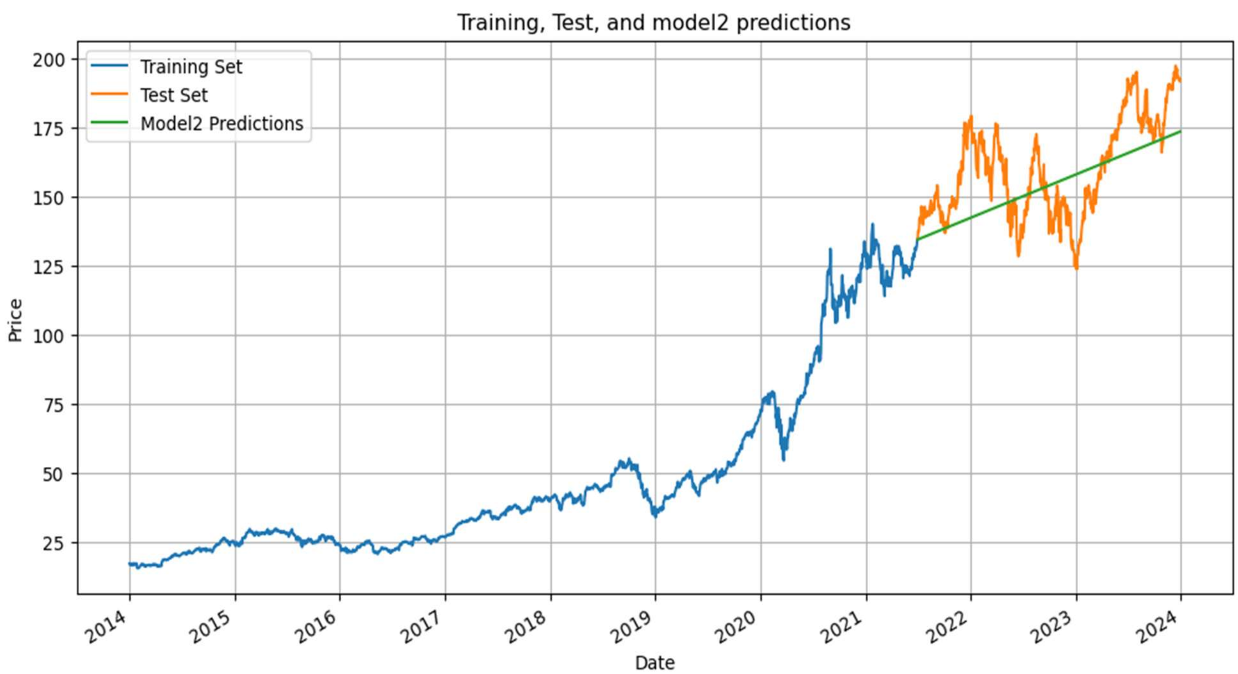

Figure 4: Plot showing Model 2, training and test dataset prices

The graphical plot of Model 2, juxtaposed against both training and test datasets as shown in **Figure 4**, reveals a consistently linear performance. The trendline of the model exhibits an upward trajectory that approximately bisects the mean of the test dataset. Furthermore, the peaks and troughs of the test data appear equidistant from the trendline of the model, suggesting that Model 2 effectively captures the underlying trend in stock price movements with considerable accuracy. In similarity to Model 1, the general trend is an uptrend indicating positive change in stock price over time.

Model 3- SARIMAX of Order (0,1,2) & Seasonal order (1,1,1,252)

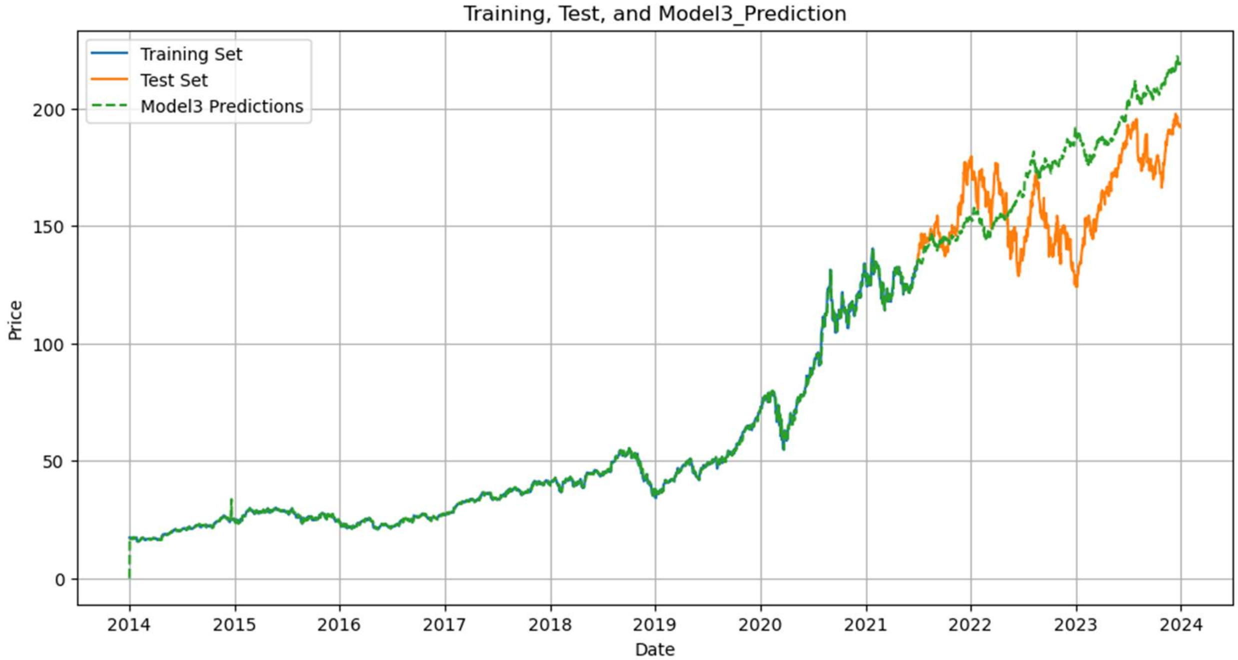

Figure 5: Plot showing Model 3, training and test dataset prices

The plot of Model 3, when compared against the training and test datasets, exhibits a general directional movement that aligns closely with that observed in Models 1 and 2, all indicating an upward trend in stock prices over time. The pricing patterns of Model 3 show a closer resemblance to the test dataset, particularly noticeable at the mid-year points of 2022 and 2023. This proximity in peak values suggests that Model 3 achieves better accuracy in predicting stock prices around these mid-year times. However, despite these periods of heightened accuracy, Model 3 generally tends to overestimate the predicted stock prices. This overestimation points to potential areas for recalibration to enhance the model's overall predictive precision.

Model Comparison

1. **Root Mean Squared Error (RMSE)**:
   - Model 2 has the lowest RMSE (15.6347), indicating the smallest average prediction error.
   - Model 3 has a moderate RMSE (24.6923).
   - Model 1 has the highest RMSE (31.5025).
2. **Mean Absolute Error (MAE)**:
   - Model 2 has the lowest MAE (13.0410).
   - Model 3 has a moderate MAE (20.9097).
   - Model 1 has the highest MAE (28.3038).
3. **Mean Absolute Percentage Error (MAPE)**:

- Model 2 has the lowest MAPE (8.12%), indicating smaller percentage errors relative to actual values.
- Model 3 has a moderate MAPE (13.41%).
- Model 1 has the highest MAPE (17.77%).

4. **Directional Accuracy**:
   - Model 2 and Model 3 both have a directional accuracy of 50.46%.
   - Model 1 has the lowest directional accuracy (48.00%).
5. **Mean of Predictions vs. Mean of Test Data**:
   - Model 2's mean of predictions (154.2093) is closest to the mean of the test data (159.9142).
   - Model 1 (172.7274) and Model 3 (174.2503) have higher means, suggesting potential overestimation.

Conclusion and Recommendation

Based on the evaluation metrics, **Model 2 (SARIMAX of order (0,1,2))** demonstrates superior performance. It has the lowest RMSE, MAE, and MAPE, indicating smaller errors in predictions. Additionally, its directional accuracy is on par with Model 3 and higher than Model 1. The mean of its predictions is closest to the actual mean of the test data, suggesting more accurate overall predictions. This is in line with the recommendation of the auto_rima grid search for best performing Arima models.

Recommendation:

- Model 2 (SARIMAX of order (0,1,2)) is recommended for predicting the future price of Apple stock due to its superior performance across multiple evaluation metrics. This model provides a good balance of accuracy and reliability, making it the best choice for forecasting Apple stock prices in this context.
- Cross Validation will be considered to improve the performance of Model 2
- The Long Short-Term Memory (LSTM) model, a type of recurrent neural network (RNN) specifically designed to address the challenge of learning long-term dependencies in sequence data may be developed and compared to Model 2 to determine optimal model for stock price prediction is chosen for future prediction of Apple stock prices.

https://github.com/nelsonaguele/Apple-Stock-Price-Prediction.git