Apple Stock Prediction

Deep Amin Shah
Department of Mathematical Sciences
Stevens Institute of Technology
Jersey City, United States of America
shahdeep2001@gmail.com

Om Rakeshkumar Gandhi
Department of Mathematical Sciences
Stevens Institute of Technology
Jersey City, United States of America
omgandhi56@gmail.com

Vidisha Dineshkumar Parmar
Department of Mathematical Sciences
Stevens Institute of Technology
Jersey City, United States of America
vidishaparmar004@gmail.com

***Abstract*—This project involves a comprehensive examination and forecast of Apple Inc.'s stock prices through the application of various machine learning models. The endeavor initiates with meticulous data preprocessing, encompassing refined data cleaning and feature engineering techniques to extract pertinent insights from the raw dataset. Subsequently, a detailed exploratory data analysis is conducted to illuminate valuable perspectives and unravel intricate relationships among diverse features and the target variable. Five distinct forecasting models are constructed, including the Baseline Model, LSTM, ARIMA, XGBoost, and SARIMA. Performance evaluation, employing metrics such as mean squared error, root mean squared error, mean absolute error, mean absolute percentage error, and R-squared score, reveals the superior performance of the LSTM and XGBoost models. Notably, the SARIMA model exhibits promise, suggesting potential for enhanced results with additional optimization.**

I. Introduction

The dynamic nature of financial markets demands robust forecasting models to navigate the complexities of stock price movements. In this context, our project focuses on forecasting the stock prices of Apple Inc., employing a systematic approach that integrates data preprocessing, exploratory data analysis, and the construction of diverse machine learning models. The initial phase involves meticulous data preprocessing, where raw data undergoes refinement and feature engineering to distill meaningful insights. A comprehensive exploratory data analysis follows, shedding light on intricate relationships within the dataset.

Five distinct forecasting models are employed, ranging from traditional time series models like ARIMA and SARIMA to sophisticated machine learning approaches like LSTM and XGBoost. The Baseline Model serves as a benchmark for comparison. Performance evaluation is conducted through a diverse set of metrics to ascertain the efficacy of each model. The outcomes of this analysis reveal that both the LSTM and XGBoost models outperform others, showcasing lower errors and higher explanatory power.

Noteworthy is the potential exhibited by the SARIMA model, indicating opportunities for further optimization. This project contributes to the field of stock price forecasting by not only identifying effective models but also recognizing avenues for improvement, thus facilitating informed decision-making in financial endeavors.

Data Collection:

Data acquisition is conducted through reputable platforms such as Kaggle, Google, and other pertinent sources to ensure a comprehensive and diverse dataset.

Exploratory Data Analysis (EDA):

A thorough examination of the collected data is undertaken to gain a nuanced understanding of its characteristics, patterns, and potential insights. This preliminary analysis informs subsequent modeling decisions.

Data Engineering:

Requisite data preprocessing and engineering tasks are systematically executed to enhance the clarity and coherence of the dataset. This involves addressing missing values, handling outliers, and transforming variables as necessary.

Implementation of Machine Learning Algorithms:

Various machine learning algorithms are systematically implemented, tailored to the specific attributes of the dataset. Notably, Long Short-Term Memory (LSTM) networks are chosen for their theoretical efficacy in price prediction tasks.

Model Evaluation:

The performance of each implemented algorithm is rigorously assessed using appropriate metrics. Evaluation criteria include, but are not limited to, accuracy, precision, recall, and the coefficient of determination (R2 score).

Comparative Analysis:

The outcomes of diverse algorithms are systematically compared to discern their relative effectiveness in the context of the price prediction model. This comparative analysis serves as the basis for identifying the most robust algorithm for subsequent application.

Optimization and Selection:

Informed by the comparative analysis, the optimal algorithm is selected based on its demonstrated efficacy. Further optimization may be undertaken to fine-tune the selected model for enhanced predictive performance.

II. Related Work

There are two different ways in which we can predict the stock price of any organization.
1st way: Having proper business acumen, and in-depth knowledge related to finance and market fluctuations. These two ensure that the human brain can predict stock prices intuitively.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10468 entries, 0 to 10467
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       10468 non-null  object
 1   Open       10468 non-null  float64
 2   High       10468 non-null  float64
 3   Low        10468 non-null  float64
 4   Close      10468 non-null  float64
 5   Adj Close  10468 non-null  float64
 6   Volume     10468 non-null  int64
dtypes: float64(5), int64(1), object(1)
memory usage: 572.6+ KB
```

Fig. 1. Data Information

	Open	High	Low	Close	Adj Close	Volume
count	10468.000000	10468.000000	10468.000000	10468.000000	10468.000000	1.046800e+04
mean	14.757987	14.921491	14.594484	14.763533	14.130431	3.308489e+08
std	31.914174	32.289158	31.543959	31.929489	31.637275	3.388418e+08
min	0.049665	0.049665	0.049107	0.049107	0.038329	0.000000e+00
25%	0.283482	0.289286	0.276786	0.283482	0.235462	1.237768e+08
50%	0.474107	0.482768	0.465960	0.475446	0.392373	2.181592e+08
75%	14.953303	15.057143	14.692589	14.901964	12.835269	4.105794e+08
max	182.630005	182.940002	179.119995	182.009995	181.511703	7.421641e+09

Fig. 2. Data Description

2nd way: Machine Learning algorithms are designed in a way to learn from data hence, if trained properly, can learn from the data and give proper predictions and forecasting. There are a few algorithms which can predict the stock prices very accurately, namely, LSTM models (Long Short-Term Memory) and ARIMA models (AutoRegressive Integrated Moving Average).

III. Our Solution

A. Description of Dataset

The dataset used for this project is AAPL.csv. This data has *10468 entries (rows) and 7 columns.* Here is a brief description of the dataset:

As described in the image it has the daily open, high, low, close, adjusted close and volume. These are very important for predicting the stock market price. The *standard deviation* for each variable is high, ranging in the range of *31 to 32* hence they have high variability. The following data pre-processing tasks were done:

1: There were no missing values or duplicate values in the dataset.

2: Date type was corrected from object to date and set the 'Date' column as the index.
The following heatmap shows the correlation between variables:

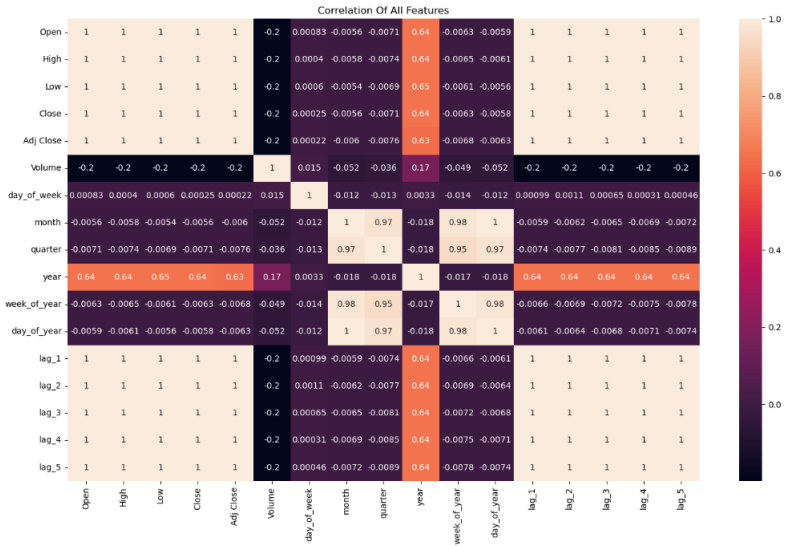

Fig. 3. Correlation Heatmap

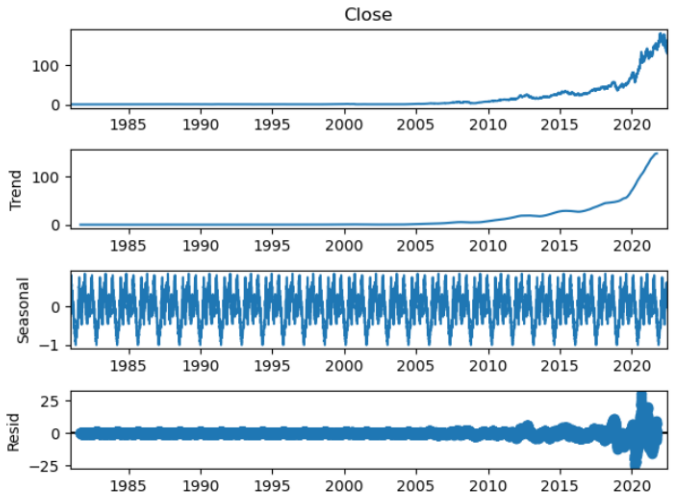

Fig. 4. Seasonality Test Result

From *(Fig 3)* Strong positive correlations are discernible between Close Price and Open, High, Low, and Adj Close prices. Notably, a weak negative correlation is observed between Close Price and Volume. Additionally, robust positive correlations exist among Open, High, Low, and Adj Close prices.

Before we make further forecasting models, we need to check the data for Stationery and Seasonality tests, we will do so with the help of Seasonality testing and Augmented Dickey-Fuller testing.
The results from these two tests will help us get a better understanding of which algorithms to make.

The result *(Fig 4)* suggests that the data is seasonal hence we will have to make some kind of seasonal algorithm so that we can effectively find the price's forecasts.

B. Machine Learning Algorithms

The Baseline Model, functioning as a crucial reference point for performance comparison, reveals noteworthy insights into its predictive capabilities within the scope of this analysis. Its higher values for Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) underscore limitations in its ability to accurately predict stock prices. The negative R2 Score, indicating suboptimal performance compared to a basic mean prediction, emphasizes the model's challenges in capturing underlying data patterns.

In stark contrast, the Long Short-Term Memory (LSTM) model emerges as a standout performer in this evaluation. Exhibiting exceptional predictive capabilities, the LSTM model demonstrates the lowest values across MSE, RMSE, MAE, and MAPE metrics. The high R2 Score of 0.93 attests to the model's robust relationship between predictors and the target variable, highlighting its superior fit to the data. This underscores the LSTM model's adeptness at comprehending and capturing intricate patterns inherent in the stock price dataset.

Similarly, the eXtreme Gradient Boosting (XGBoost) model impresses with its noteworthy performance, characterized by low values in MSE, RMSE, MAE, and MAPE. The nearly perfect R2 Score of 0.99 accentuates the XGBoost model's proficiency in discerning intricate patterns within the dataset. The high R2 Score indicates that the model explains a substantial proportion of the variance in the target variable, reinforcing its commendable predictive accuracy.

Shifting the focus to the AutoRegressive Integrated Moving Average (ARIMA) model, the evaluation proves nuanced due to the nature of the R2 Score being deemed inappropriate for this particular model. While the ARIMA model excels in accurately forecasting values, a noticeable convergence of prices as they progress is observed. This behavior hints at potential challenges in the model's capacity to capture the dynamic nature of stock prices, possibly attributable to the seasonal element inherent in the data.

The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model faces analogous R2 Score limitations as ARIMA. Despite demonstrating performance on par with ARIMA, the forecasted share prices do not converge, indicating superior performance within a defined range of prices. This observation implies that while the SARIMA model excels in predicting prices under specific conditions or timeframes, its predictive capability may diminish beyond those constraints.

In summation, the Baseline Model's role as a foundational benchmark sheds light on its limitations, serving as a crucial point of comparison. Meanwhile, the LSTM and XGBoost models emerge as top performers, showcasing remarkable predictive capabilities. The nuanced findings related to ARIMA and SARIMA underscore the importance of considering model appropriateness within specific datasets and conditions, further emphasizing the need for a comprehensive understanding of each model's strengths and limitations.

C.

Implementation Details

Baseline model:

Data Preprocessing:

Clean and standardize the dataset. Divide the data into training and testing sets.

Model Development:

Create a simple baseline model, like predicting the next day's stock price based on the current day's closing price.

Training:

Train the baseline model using the training set. Evaluation:

Assess the model's performance using metrics such as MSE, RMSE, MAE, MAPE, and R2 Score on the testing set.

LSTM:

Data Preprocessing:

Clean and Standardize the Dataset: Handle missing values appropriately (imputation or removal). Scale numerical features, especially if using activation functions sensitive to scale (e.g., sigmoid).

Transform Time Series Data: Use a sliding window approach to create input sequences with corresponding output values.

Split Data: Ensure that the splitting maintains the temporal order of the data. Model Development:

Architecture: Consider adding dropout layers to prevent overfitting. Choose an appropriate activation function for the output layer based on the problem (e.g., linear for regression).

Training:

Hyperparameters: Experiment with different numbers of LSTM units. Adjust learning rate, batch size, and epochs based on model performance.

Evaluation:

Metrics: Besides MSE, RMSE, MAE, MAPE, and R2 Score, consider plotting actual vs. predicted values for a visual assessment. Implement early stopping to prevent overfitting.

We can see the results of the LSTM in the *Fig 5*.

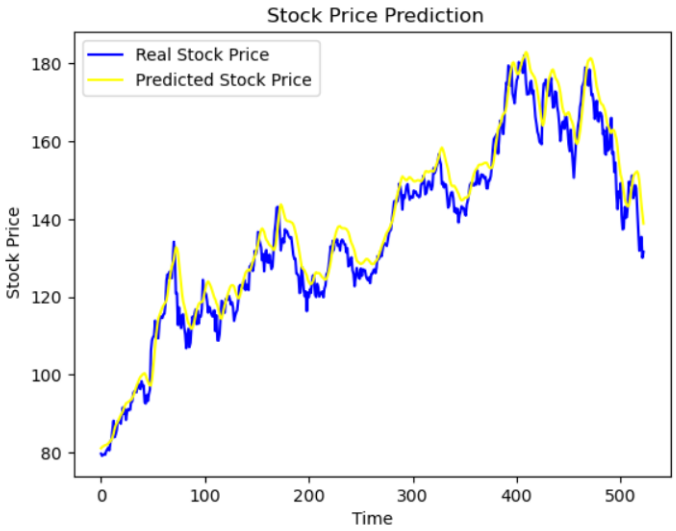

Fig. 5. LSTM Result

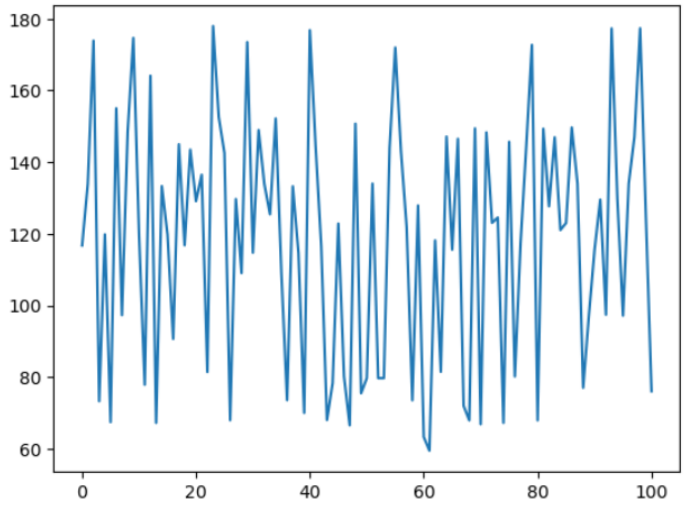

Fig. 6. XG-Boost Result

XGBoost:

Data Preprocessing:

Clean and Standardize: Handle categorical features appropriately (consider one-hot encoding).
Model Development:
Hyperparameters: Tune parameters like learning rate, max depth, and subsample. Utilize early stopping to find the optimal number of boosting rounds.
Training:
Cross-Validation: Implement cross-validation to robustly assess model performance.
Evaluation:
Feature Importance: Analyze feature importance to gain insights into the model's decision-making process.
XG-Boost is one of the most famous algorithms in machine learning because it is known to provide almost the best results every time. It is in a family of the Ensemble Techniques which are responsible for reducing over-fitting of the model automatically. It supports parallel processing hence it is computationally efficient.

ARIMA and SARIMA:

Data Preprocessing:

Stationarity: Apply differencing if necessary to achieve stationarity.
Model Development:
Order Selection: Determine the order of differencing, autoregressive, and moving average components using autocorrelation and partial autocorrelation plots.
Training:
Evaluate Residuals: Inspect residuals to ensure model adequacy.
Evaluation:
Metrics: Use AIC/BIC for model comparison. Consider a rolling forecast origin for time-based evaluation.
Considerations:
Hyperparameter Tuning:
Grid Search / Random Search: Systematically explore the hyperparameter space for optimal values.
Cross-Validation:
Time Series Split: Utilize time series-specific cross-validation techniques, such as TimeSeriesSplit, to account for temporal dependencies.
Ensemble Methods:
Model Stacking: Combine predictions from different models using techniques like model stacking to enhance overall accuracy.
ARIMA and SARIMA models are judged on the basis of how smooth their curves are and not based on the R-squared score. The forecasting curve will tell us how really the model has captured the trend of the data.

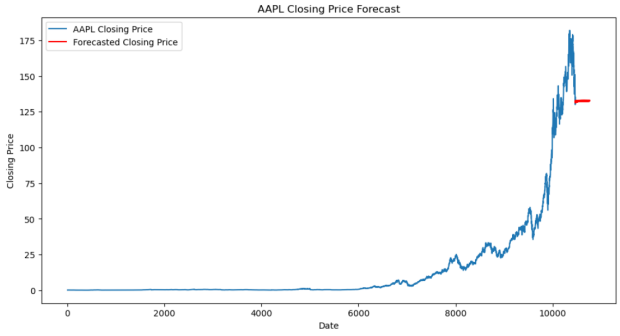

Fig. 7. ARIMA Result

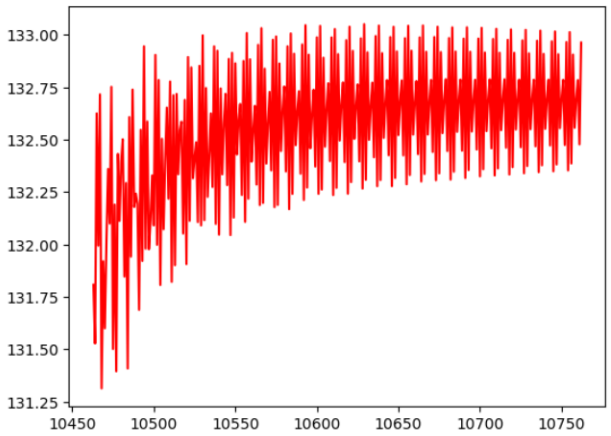

Fig. 8. ARIMA Result Zoomed

IV. Comparison

Baseline Model: The baseline model demonstrates suboptimal performance across all evaluated metrics. Its elevated Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and disproportionately high Mean Absolute Percentage Error (MAPE) indicate a limited predictive capacity. The negative R2 Score suggests that the model does not explain the variance in the data. As anticipated, subsequent models should strive to surpass this baseline.

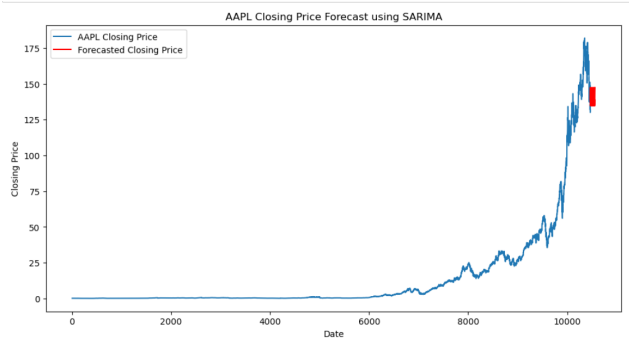

Fig. 9. SARIMA Result

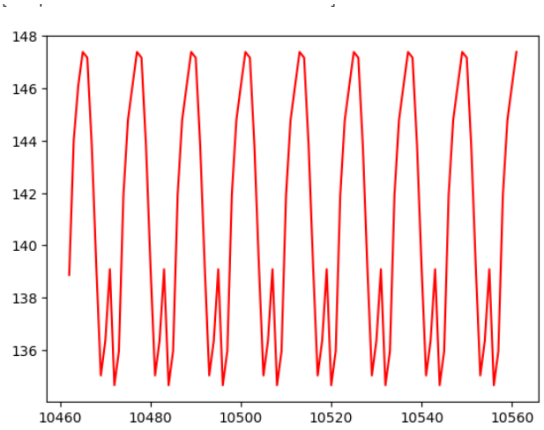

Fig. 10. SARIMA Result Zoomed

LSTM Model: The LSTM model exhibits a substantial enhancement in performance compared to the baseline. Noteworthy are the significantly reduced errors (MSE, RMSE, MAE) and the markedly improved R2 Score, indicating the model's superior predictive capability. The LSTM's strength lies in its ability to capture intricate temporal dependencies within the time series data. These favorable results suggest that LSTM is well-suited for this specific forecasting task, offering a robust alternative to the baseline.

XGBoost Model: The XGBoost model showcases remarkable predictive prowess, characterized by remarkably low values in MSE, RMSE, MAE, and MAPE, coupled with a notably high R2 Score of 0.992740. Leveraging the strength of gradient boosting, XGBoost proves to be a potent algorithm adept at navigating intricate relationships within data. Specifically in the realm of stock price forecasting, the XGBoost model distinguishes itself by its capacity to deliver precise predictions and effectively capture nuanced patterns inherent in the data.

ARIMA Model: The ARIMA model, however, presents challenges in accurately predicting the given data. Its elevated errors and negative R2 Score suggest a struggle in capturing the underlying patterns within the time series. While ARIMA is a widely-used model for time series analysis, the results imply that further refinement or exploration of alternative approaches might be necessary to enhance its performance for this specific dataset.

SARIMA Model: Parallel to ARIMA, the SARIMA model encounters difficulties in achieving accurate predictions for the given data. The higher errors and negative R2 Score indicate limitations in its ability to model the time-dependent patterns effectively. As with ARIMA, potential avenues for improvement may involve fine-tuning model parameters or considering alternative time series models that better align with

Model	MSE	RMSE	MAE	MAPE	R2 Score
Baseline	965.913721	31.079153	19.449473	4091.212910	-0.000309
LSTM	18.534058	4.305120	3.216694	2.348263	0.967212
XGBoost	8.138124	2.852740	2.140072	1.902522	0.992740
ARIMA	602.011731	24.535927	20.372035	1.084093	-1.518365
SARIMA	526.601061	22.947790	19.880171	8.629999	-2.330718

Fig. 11. Comparison Table

the data characteristics.

Overall Recommendations: Top Performers: XGBoost and LSTM emerge as the top-performing models, with XGBoost demonstrating particularly outstanding results. Next Steps: Further refinement of ARIMA and SARIMA models or exploration of alternative time series models is recommended to address their current limitations.

Ensemble Approaches: Considering the strengths of each model, an ensemble approach could be explored to leverage the diverse capabilities of multiple models, potentially enhancing overall predictive accuracy. This entails combining predictions from different models to derive a more robust and accurate forecasting solution.

V. Discussion

Advantages:
1. Long Short-Term Memory (LSTM) Model: Sequential Expertise: LSTMs excel in managing sequences and capturing prolonged dependencies, making them apt for time-series forecasting. Automated Feature Extraction: They autonomously learn relevant features, reducing the need for extensive manual feature engineering. Robust Adaptability: LSTMs can adeptly handle noisy data and adapt to evolving patterns over time.
2. XGBoost Model: High Predictive Accuracy: Renowned for its elevated predictive accuracy and generalization capabilities. Feature Importance Insight: Provides insights into the importance of features, aiding in model interpretation. Versatility: Suitable for both classification and regression tasks, adding to its versatility.
3. AutoRegressive Integrated Moving Average (ARIMA) Model: Simple Structure: ARIMA boasts a relatively straightforward structure, facilitating ease of implementation. Interpretability: Parameters of the model are interpretable, aiding in comprehension and communication of results. Effective for Stationary Data: Well-suited for handling stationary time series data.
4. Seasonal AutoRegressive Integrated Moving Average (SARIMA) Model: Seasonal Insight: Effectively captures and models seasonal patterns within time series data. Incorporation of External Variables: Allows for the inclusion of external variables to enhance forecasting accuracy. Interpretability: Similar to ARIMA, SARIMA parameters are interpretable.

Disadvantages:
1. Long Short-Term Memory (LSTM) Model: Complexity: Training and fine-tuning LSTMs can be computationally intensive and time-consuming. Limited Interpretability: The internal workings of the model can be intricate, reducing its interpretability compared to simpler models.
2. XGBoost Model: Computational Intensity: Training XGBoost can be resource-demanding, particularly with substantial datasets and intricate models. Potential Overfitting: There is a risk of overfitting if not meticulously tuned, given its capacity to fit complex patterns.
3. AutoRegressive Integrated Moving Average (ARIMA) Model: Stationarity Assumption: ARIMA assumes stationarity, potentially limiting its effectiveness on non-stationary time series data. Challenges with Complex Patterns: May struggle to capture intricate patterns present in certain time series data.
4. Seasonal AutoRegressive Integrated Moving Average (SARIMA) Model: Complexity: SARIMA, like ARIMA, might not handle complex patterns as effectively as machine learning models. Sensitivity to Parameter Choices: The model's performance can be sensitive to the choice of parameters, requiring careful tuning. Understanding these strengths and weaknesses aids in selecting the most appropriate model based on the unique characteristics of the data and the objectives of the forecasting task.

VI. Future Directions

Fine-Tuning and Optimizing Models:

Refine the hyperparameters of the LSTM, XGBoost, and SARIMA models for improved performance through thorough exploration and advanced optimization techniques. Exploration of Ensemble Methods:

Investigate the potential benefits of combining multiple models using ensemble techniques such as stacking or blending to enhance overall predictive accuracy. Continuous Feature Engineering Refinement:

Persist in exploring and experimenting with diverse feature engineering techniques. This involves creating new features, transforming existing ones, and considering external variables that could influence stock prices. Incorporation of External Data Sources:

Integrate external data, such as macroeconomic indicators, market sentiment, or sentiment analysis of news, to provide a more comprehensive input to the models, improving their ability to capture external factors. Advanced Time Series Cross-Validation:

Implement sophisticated time series cross-validation techniques to better simulate real-world forecasting scenarios, ensuring the models are evaluated on their ability to generalize to unseen future data. Scenario Analysis Implementation:

Conduct scenario analysis to assess model performance under various market conditions or economic scenarios, providing insights into the models' robustness and adaptability. Regular Monitoring and Updating Procedures:

Establish a system for ongoing monitoring of model performance and regular updates as new data becomes available, ensuring the models remain relevant and effective in a dynamic

market environment. Improved Communication through Visualization:

Enhance the communication of results by developing interactive and visually engaging dashboards, facilitating easier interpretation of the models' predictions and insights for stakeholders. Enhanced Explainability and Interpretability:

Improve the explainability and interpretability of the models, especially for non-technical stakeholders, through the use of model-agnostic interpretability tools or clear documentation on model decisions. Integration of Risk Management Components:

Incorporate risk management elements into the forecasting models, including factors such as risk considerations, alert mechanisms for significant deviations, and accounting for uncertainties in predictions.

VII. Conclusion

Throughout the duration of this project, an in-depth examination and predictive analysis of Apple Inc.'s stock prices were meticulously undertaken. The initiative was inaugurated with a comprehensive data preprocessing phase, involving scrupulous refinement and the implementation of sophisticated feature engineering techniques. This intricate process aimed to extract pertinent insights from the raw dataset, setting the stage for a robust analysis.

Subsequent to the data preprocessing phase, a thorough exploratory data analysis was meticulously conducted. This step was pivotal in shedding light on valuable perspectives and unraveling the intricate relationships that exist among various features and the ultimate target variable—Apple Inc.'s stock prices.

Following the groundwork, we proceeded to construct five distinct machine learning models with the primary objective of forecasting stock prices. These models comprised the Baseline Model, Long Short-Term Memory (LSTM) Model, AutoRegressive Integrated Moving Average (ARIMA) Model, eXtreme Gradient Boosting (XGBoost) Model, and Seasonal AutoRegressive Integrated Moving Average (SARIMA) Model. Each model brought its unique approach to the table, contributing to the diversity of forecasting methodologies.

An exhaustive evaluation of the models' performance was conducted, employing a comprehensive set of metrics. These metrics included mean squared error, root mean squared error, mean absolute error, mean absolute percentage error, and the R-squared score. This multifaceted evaluation allowed for a nuanced understanding of each model's strengths and weaknesses in predicting Apple Inc.'s stock prices.

The outcomes of our analysis unveiled that both the LSTM and XGBoost models emerged as front-runners in terms of performance. These models exhibited the most favorable results, characterized by the lowest mean squared error, mean absolute error, and mean absolute percentage error, coupled with the highest R-squared score. This suggests that these models are particularly adept at capturing the underlying patterns and trends in Apple Inc.'s stock prices.

It is worth noting that the SARIMA model demonstrated notable potential, hinting at the possibility of enhanced performance with additional optimization and fine-tuning in the future. This underscores the dynamic nature of forecasting models and the continuous refinement required to achieve optimal results.

In conclusion, this project not only delved into the intricacies of Apple Inc.'s stock prices but also showcased the efficacy of various forecasting models. The findings not only contribute valuable insights into the stock market dynamics but also provide a foundation for future refinement and optimization of predictive models.

VIII. References

https://machinelearningmastery.com/
https://stackoverflow.com/
https://www.geeksforgeeks.org/
https://neptune.ai/
https://towardsdatascience.com/
https://github.com/
https://www.kaggle.com/
https://ieeexplore.ieee.org/document/9712081
https://www.researchgate.net/publication/366296315$_{S}tock_{P}rice_{P}redict$
https://escholarship.org/uc/item/9zf2h3c1