Integrating Random Forest and SARIMA Models for Future Export Sales Forecasting: A Machine Learning Ensemble Approach

by

Temidayo Olowoyeye

Data and Business Analyst

Terms of Use

This document was prepared by Temidayo Olowoyeye, herein referred to as the "Author", with the sole purpose of showcasing his analytical and modeling skills. The content pertains to a fictitious entity named ABC Ltd and is entirely hypothetical. The following terms and conditions apply to the use and interpretation of this document:

- **Confidentiality**: The information contained in this document is not based on any actual data from any real company. It demonstrates the author's analytical and modeling capabilities and does not disclose proprietary or confidential information.
- **Opinions and Interpretations**: All statements, findings, and views expressed in this document are solely those of the author and do not represent the views or opinions of any real company or individual. The author takes no responsibility for the accuracy or applicability of the content to any actual business entity.
- **Fictional Nature of Data**: Any data, figures, or statistics presented in this document are fictional and created to illustrate data analysis and modeling techniques. They do not reflect the operational performance of any real company.
- **Not to be used for Decision-Making**: This document is not intended for making business decisions or forming strategies for any company. It demonstrates the author's skills and should not be relied upon for practical purposes.
- **No Liability Assumed**: The author assumes no liability for any consequences arising from the use of this report. The document is provided without warranties or guarantees of accuracy, completeness, or reliability.
- **Unauthorized Use**: This document is the author's intellectual property and is intended for personal use and demonstration only. Unauthorized reproduction, distribution, or use of any part of this document is strictly prohibited.
- **Professional Consultation**: For any business decisions or actions, it is recommended to consult the author or qualified professionals and obtain relevant, accurate, and up-to-date information from authoritative sources.

You agree to abide by these terms and conditions by accessing and using this document. If you do not agree with any part of these terms, you are not authorized to use or rely on the information presented in this document.

Supplementary Materials

- Data Sources: The data utilized in this analysis was sourced from Kaggle.com, accessible via link.
- Python Script: The analysis was conducted using Python, and the corresponding code can be found at github.com/TemidayOlowoyeye

Conflicts of Interest

- The author declares no conflicts of interest.

1. Introduction

Machine learning has transformed numerous industries by facilitating predictive analytics to forecast future trends and outcomes accurately. Predictive modeling with machine learning involves utilizing algorithms and statistical models to scrutinize historical data, detect patterns, and anticipate forthcoming events or results. In the case of companies like ABC Ltd, machine learning techniques are harnessed to forecast export sales figures by considering factors such as past sales data and other pertinent variables.

This approach provides actionable insights and foresight into future export sales performance, crucial for effective planning and strategic decision-making. Traditional time series models like SARIMA (Seasonal Autoregressive Integrated Moving Average) have been widely employed for forecasting tasks. However, in recent years, ensemble approaches combining the strengths of multiple models have gained traction for their ability to improve predictive accuracy. In this study, the integration of Random Forest, a powerful machine learning algorithm, with SARIMA models is explored to enhance export sales forecasting for ABC Ltd. Leveraging the complementary strengths of both methodologies, robust predictions are aimed to be provided that account for both linear and nonlinear relationships within the data.

This analysis will predict ABC Ltd's projected performance of export sales relative to total sales over the next 38 months. By employing a machine learning ensemble approach, valuable insights are extracted, and forecast accuracy is improved, thereby assisting ABC Ltd in making informed decisions and optimizing its export sales strategies.

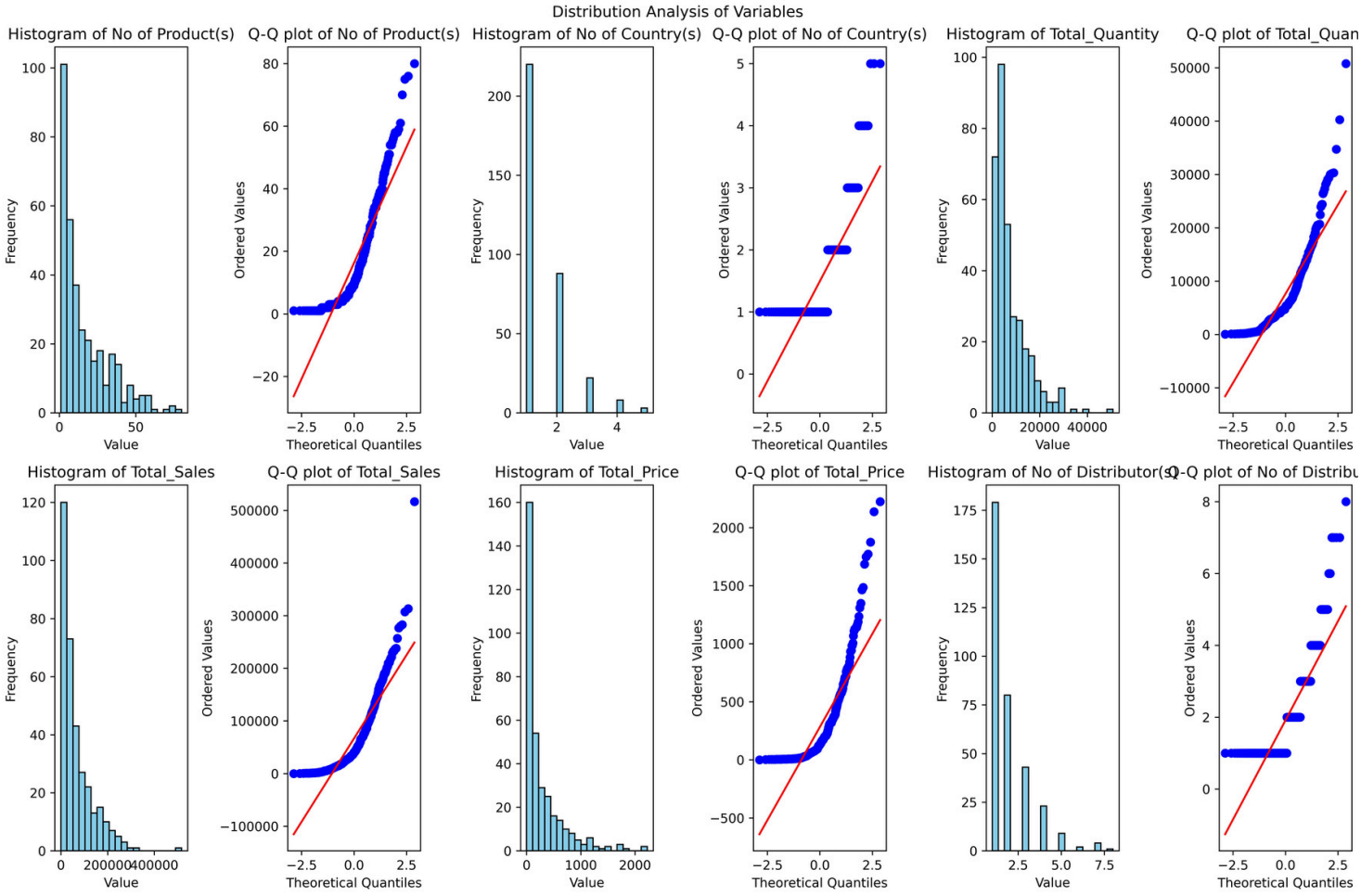
2. Data Processing

2.1. Data Cleaning and Manipulation

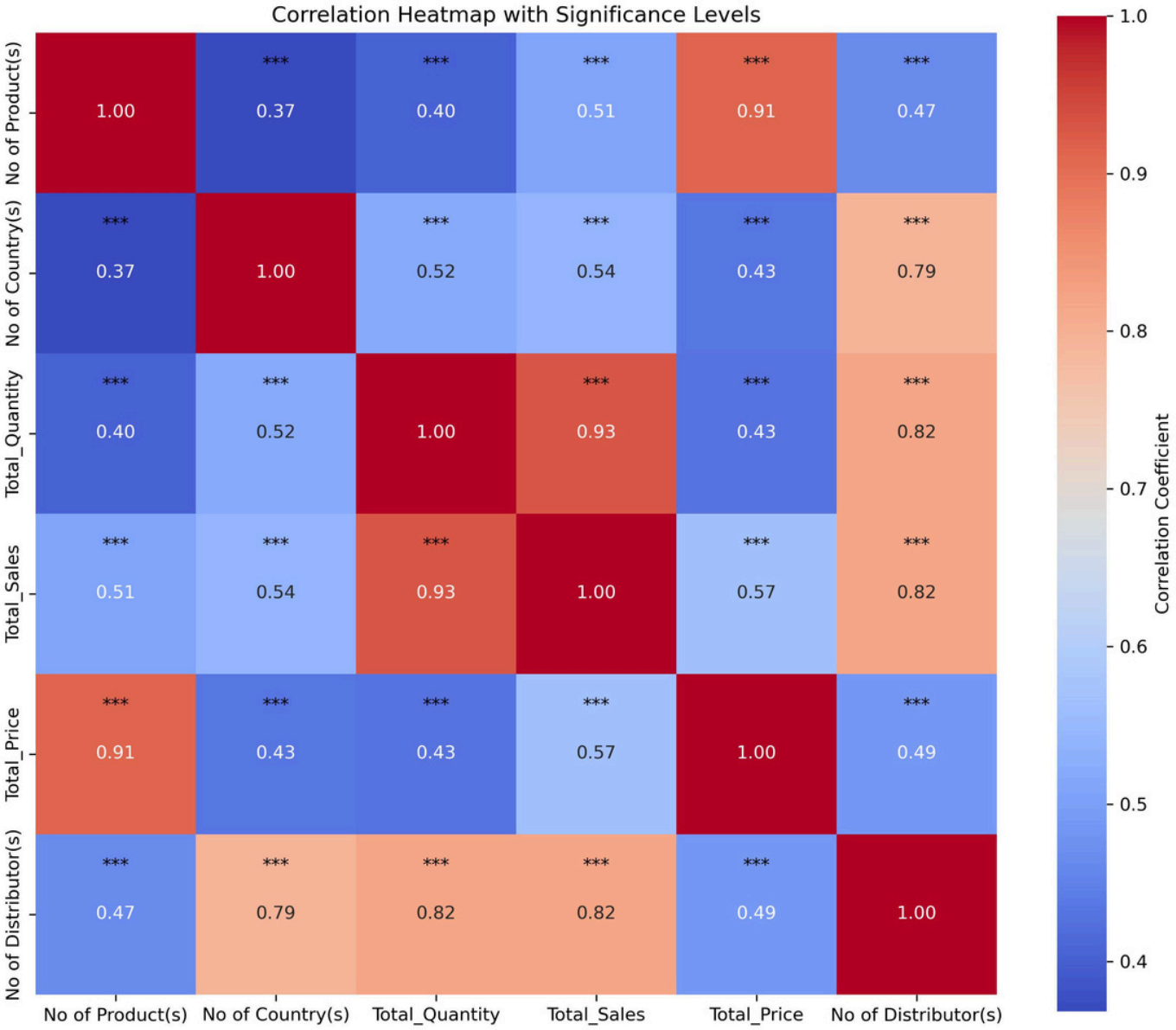
Cleaning and reshaping the dataset are vital steps in data analysis; once the data has been loaded into the data frame, practices such as checking for missing values, removing duplicates, adding additional column, and restructuring the data frame for efficient organization were performed to ensure data accuracy, consistency, and reliability, making the data ready for further analysis, modeling, or visualization activities.

2.2. Data Exploration

- **Normality Check:** This was conducted to assess the normality of each variable in the data frame, aiming to guide the selection of an appropriate model. This assessment involved utilizing Histograms, Q-Q plots, and the Kolmogorov-Smirnov test to evaluate data distribution and ascertain its adherence to normality assumptions. Based on the KS statistics, it was observed that none of the variables in the dataset follow a normal distribution with the p_value < 0.05.



- **Assess Variable Relationship:** It is essential to understand the relationships between variables to help select the most relevant features with a solid relationship with the target variable, which is more likely to contribute significantly to the model's predictive power, addresses multicollinearity, and provides insights into potential model performance. As shown below, all the variables are correlated and significant at <1% level, signifying a strong relationship among them and thus indicating that more stable and robust models can be built on them.



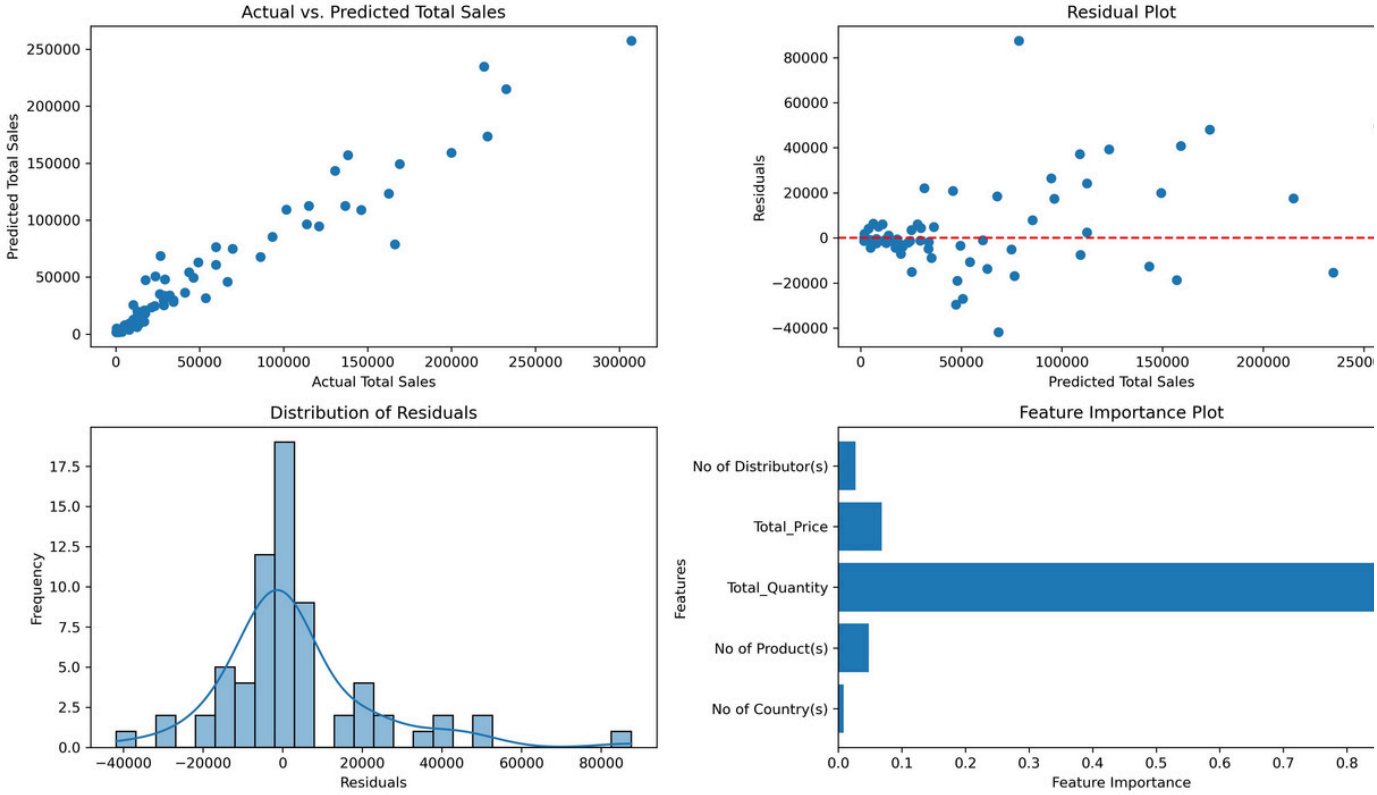
3. Machine Learning Algorithm

3.1. Random Forest Modeling (RF)

The normality check performed in 2.2 above influenced the choice of RF. RF is particularly well-suited for modeling datasets with variables that exhibit non-normal distributions due to its non-parametric nature, robustness to outliers, automatic feature selection capturing complex relationships, ensemble learning reducing overfitting, and scalability for handling large datasets with many variables. The application of this model involves three steps;

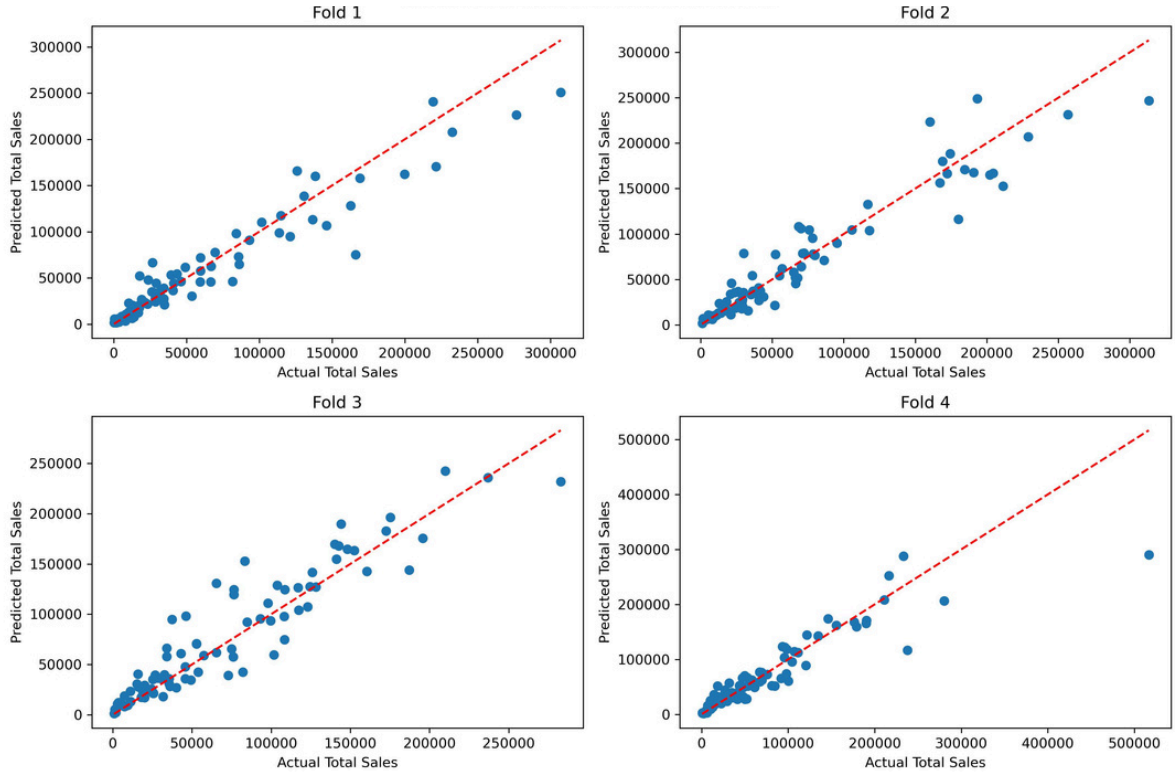
- Step I: The model was trained and fitted on the historical dataset using an 80/20 split, resulting in an R-squared (R2) value of 0.9177, indicating a strong correlation of 92% between the predictors and the target variable. The Mean Absolute Error (MAE) was calculated to be 12106, which is low relative to the scale of the target variable. This suggests a reasonable level of accuracy in predicting the target variable based on the selected features.

Diagnostic Tools for RF Model Evaluation

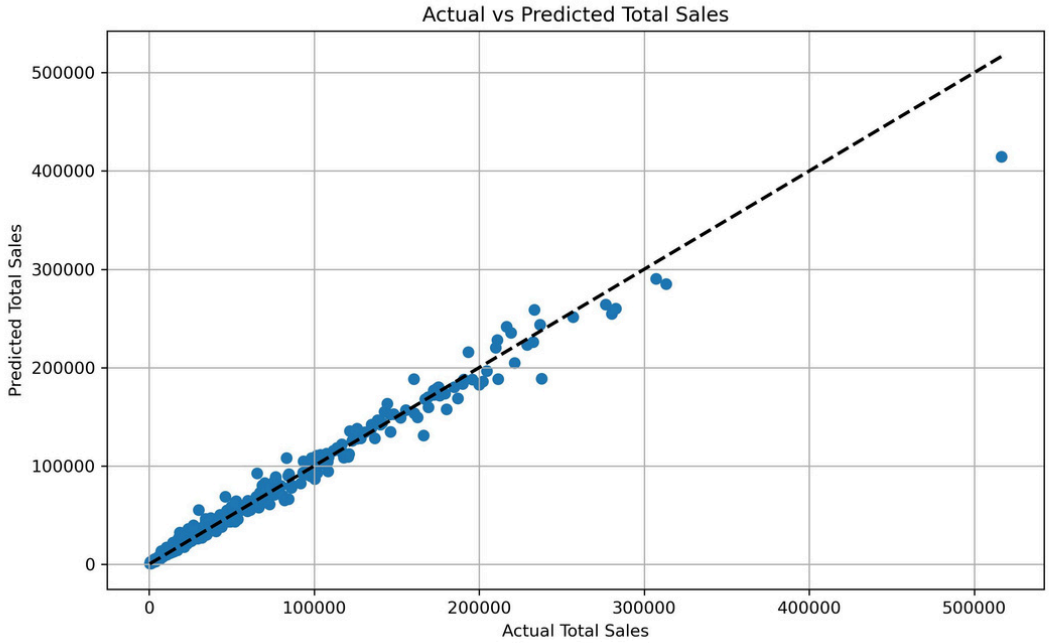


- Step II: The dataset was subjected to 4-fold cross-validation, splitting it into four subsets to run the model and ensure the reliability of model performance, reduce overfitting, handle data imbalance, and maximize data utilization. The average Mean Absolute Error (MAE) and R-squared (R2) values obtained were 14610 and 0.877, respectively. These results affirm the efficiency and effectiveness of the model in accurately predicting the target variable based on the selected features.

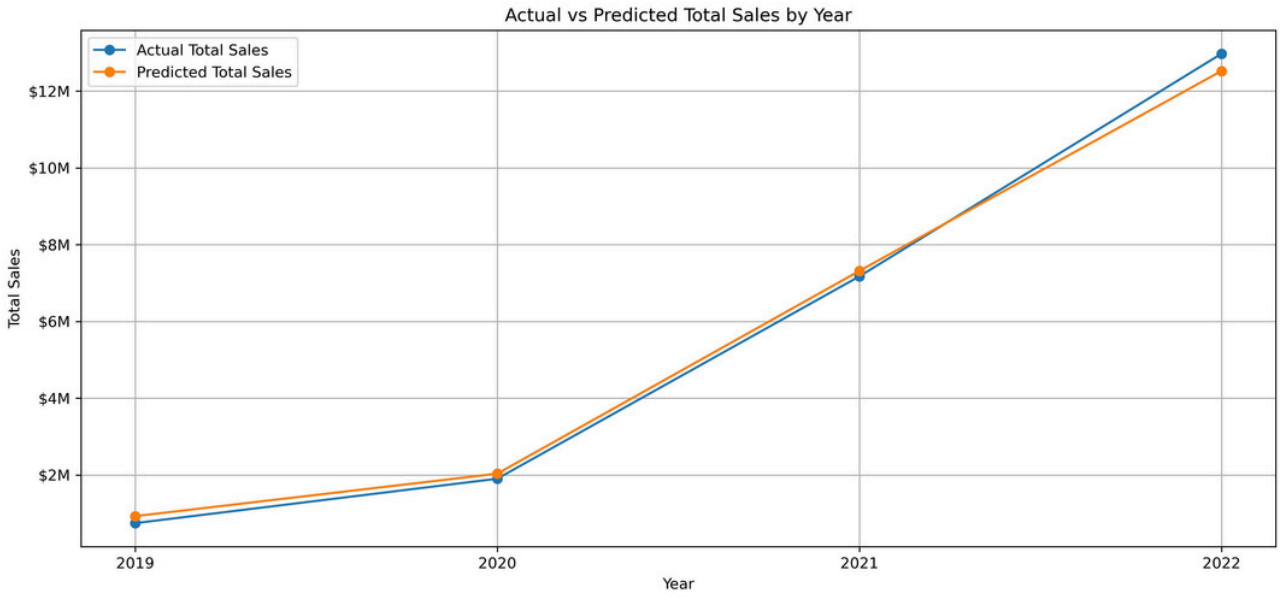
Actual vs predicted Total Sales for each Fold

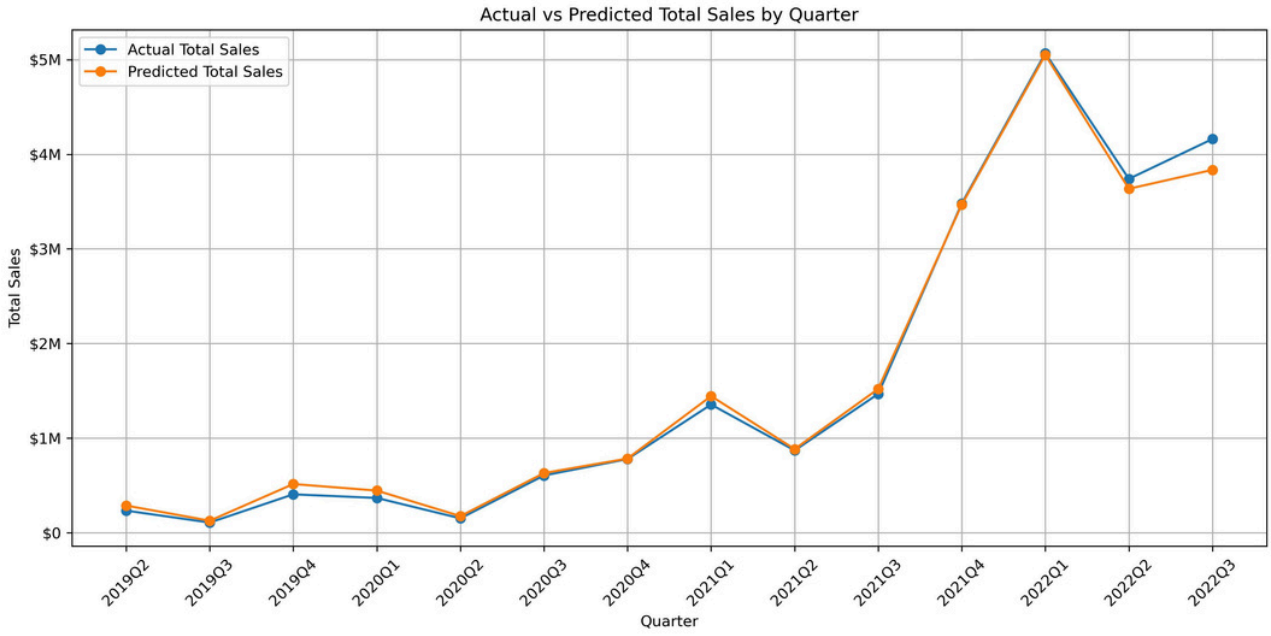


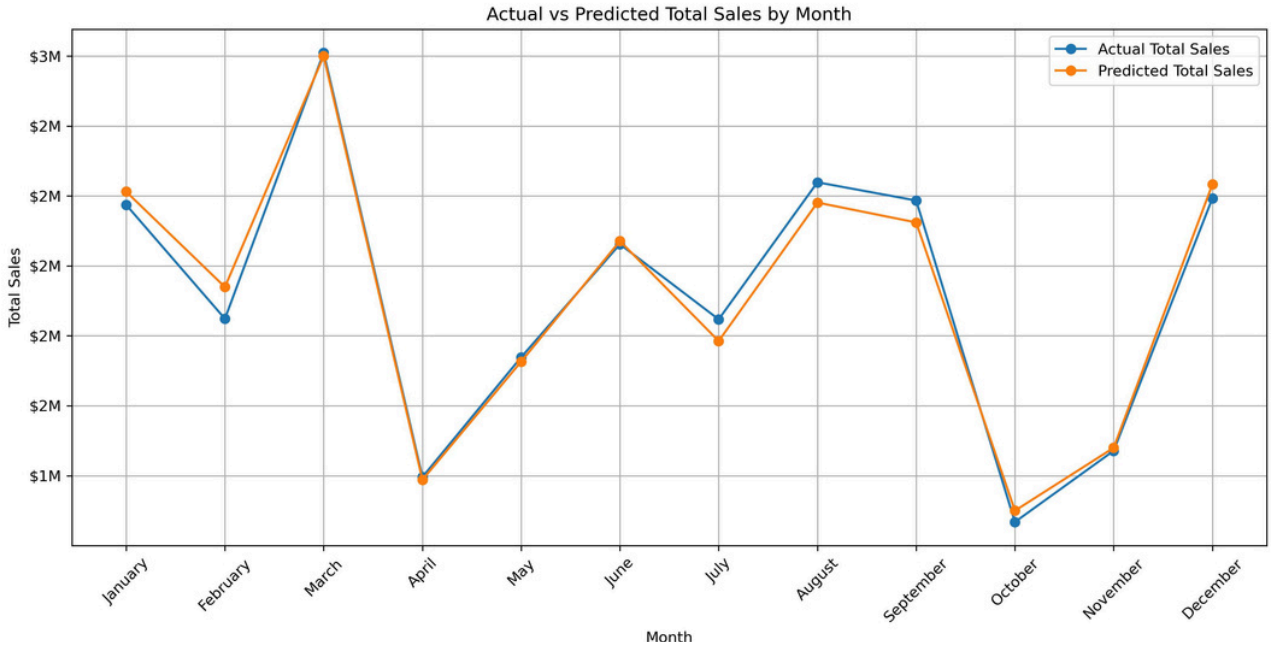
- Step III: After successfully passing the split test and undergoing 4-fold validation, which confirmed its reliability and usability based on performance metrics, the model was fit on the entire dataset. The resulting Mean Absolute Error (MAE) of 5514 and R-squared (R2) value of 0.98 indicate that the model performs even better when trained on the entire dataset, demonstrating its robustness and improved predictive power across the entire data range.



The model's performance on total export sales was viewed on a yearly, quarterly, and Monthly basis.





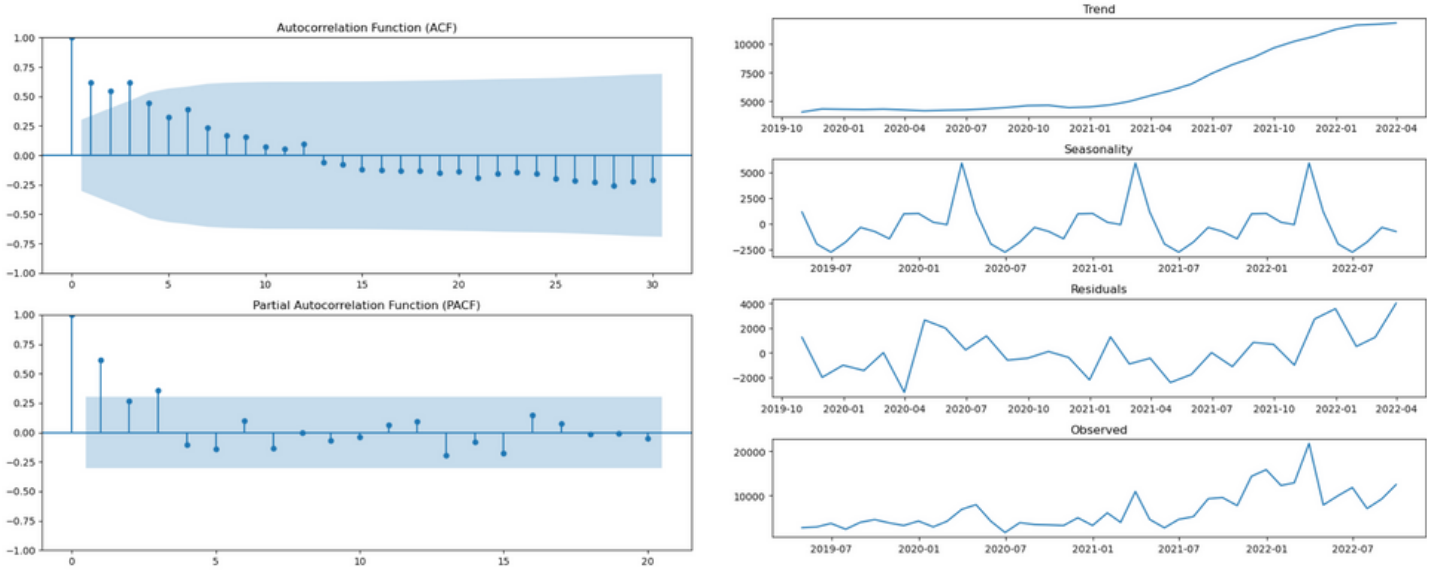


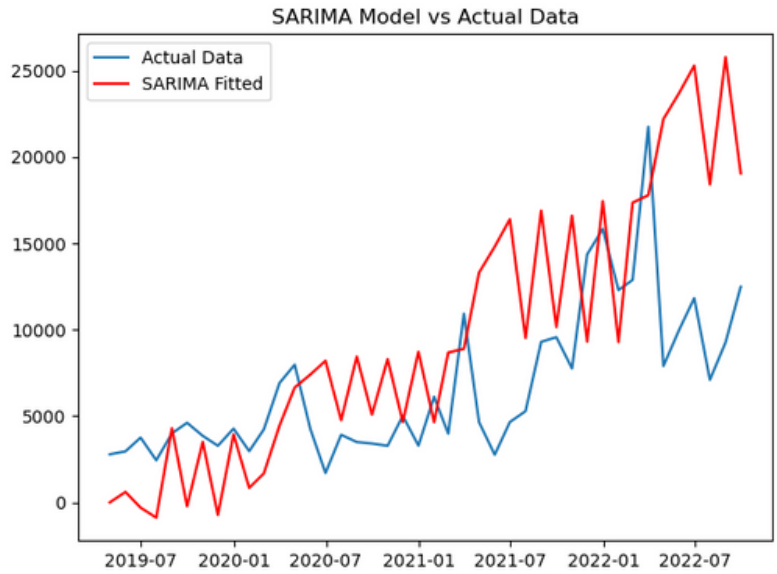
3.2. SARIMA Model

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model is chosen for time series forecasting due to its ability to capture complex patterns, dependencies, seasonality, trends, and autocorrelation within the data. This model was explicitly adopted because trends were detected from the initial model, indicating the need for a more sophisticated approach to capture and forecast these trends accurately. The application of the SARIMA model involves two main steps:

- Step I: The data was sorted and resampled to a Monthly frequency. This step takes into account the monthly trend observed in the data. Additionally, the most featured variable from the Random Forest (RF) model was incorporated into the SARIMA model, ensuring that important predictors are considered during the forecasting process. Autocorrelation, partial autocorrelation, and seasonal decomposition were performed to determine the range for grid-searching SARIMA orders. Multiple combinations were tested, and the combination resulting in the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values, indicating better model fit, was selected as the best SARIMA order.

Diagnostic Tools for SARIMA Model Evaluation





- Step II: The optimal SARIMA order obtained in Step 1 was utilized to predict the future values of the variables of interest, excluding the dependent variable, which was predicted using the Random Forest model.

SARIMA Forecast 09/2022 - 12/2025

```
In [203]: #check Forecasted data
          forecast_df.head()
```

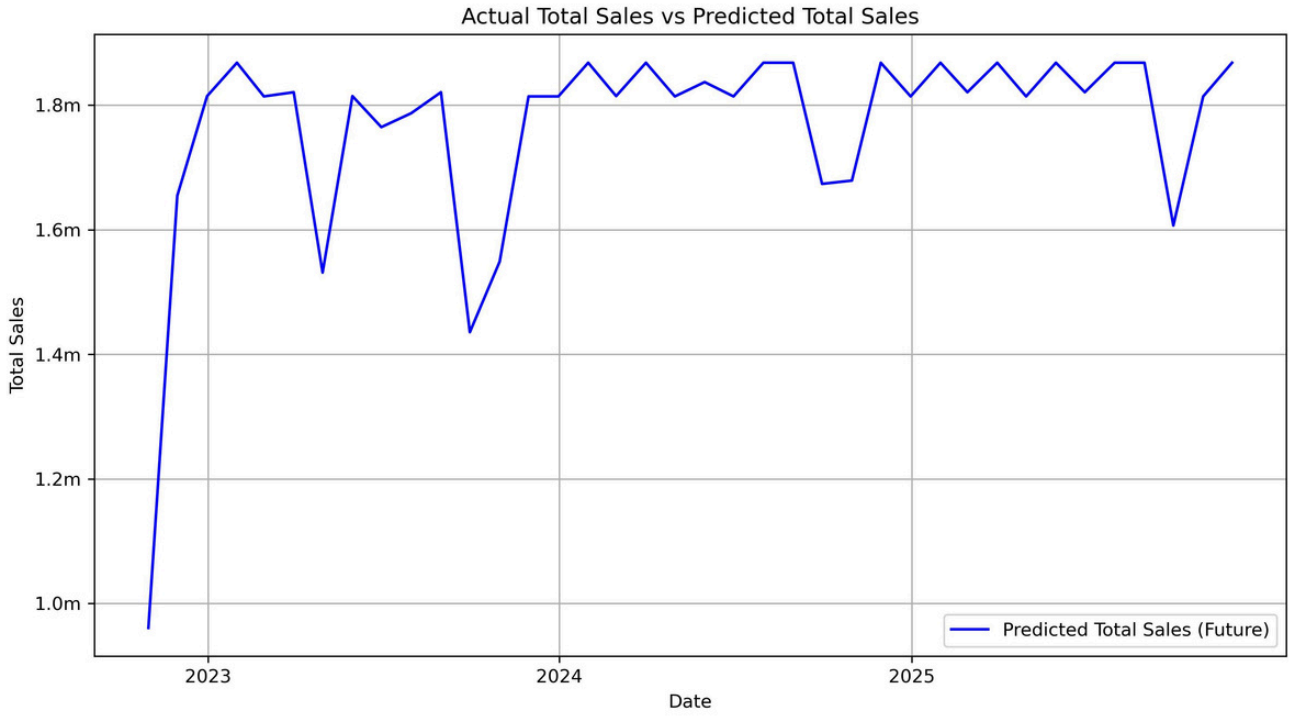
Out[203]:

Date_Time	No of Product(s)	No of Country(s)	Total_Quantity	Total_Price	No of Distributor(s)
2022-10-31	100	21	59951	5455.42	30
2022-11-30	263	32	183513	3319.04	61
2022-12-31	200	42	319391	6012.67	82
2023-01-31	304	51	326958	5482.38	91
2023-02-28	167	28	265087	5946.47	74

4. Predicting Future Export sales

The target variable (Total Sales) was extracted from the Historical data frame, and the feature variables forecasted with SARIMA were also extracted. RF was fit on this data to predict the future Total sales from October 2022 to December 2025.

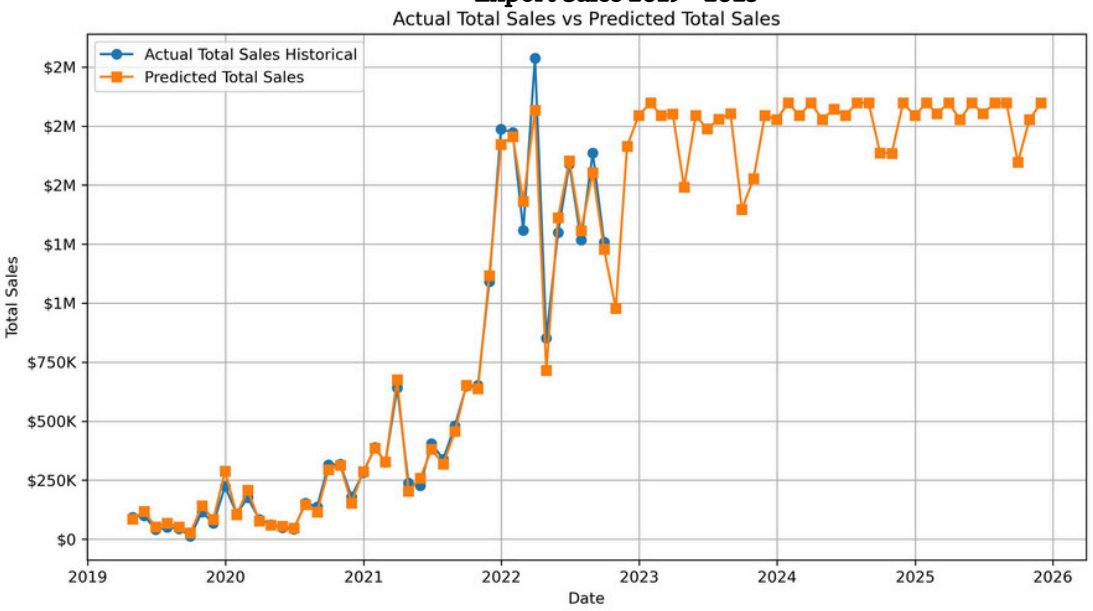
Predicted Export Sales 09/2022- 12/2025



4.1 Visualization and Insight on Future Export Sales

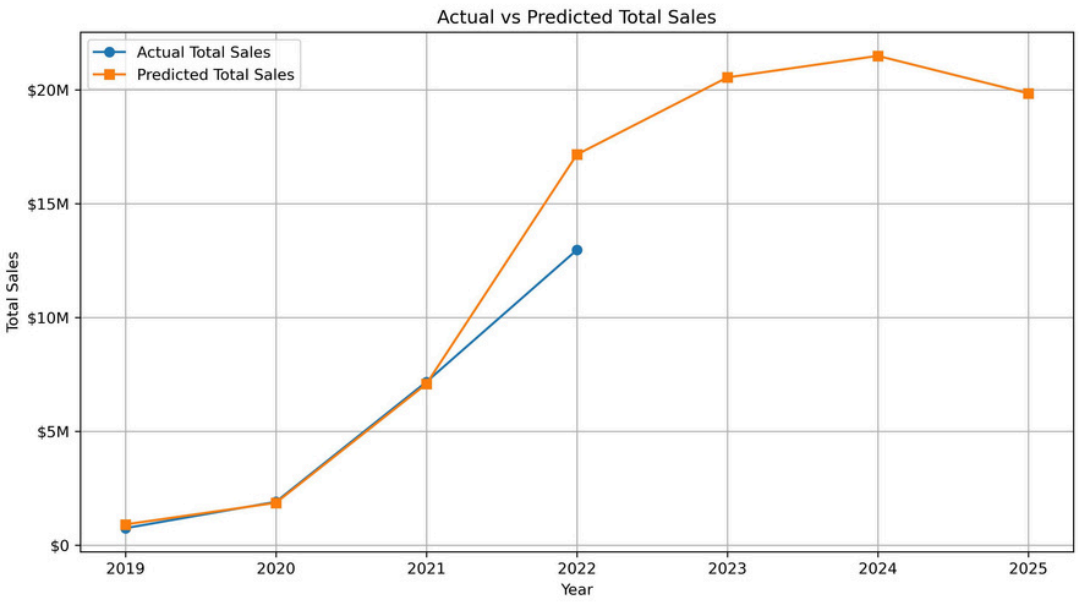
- **2023 - 2025 Overview**
  - There will be a fluctuating pattern in predicted total sales over the years, with periods of growth and decline.
  - Increasing trends will continue, indicating potential market acceptance and growth efforts.
  - A new pick level is expected In February 2023.
  - From November 2023 to November 2025, there's a relatively steady but moderate growth trend in predicted sales.

Export Sales 2019 - 2025



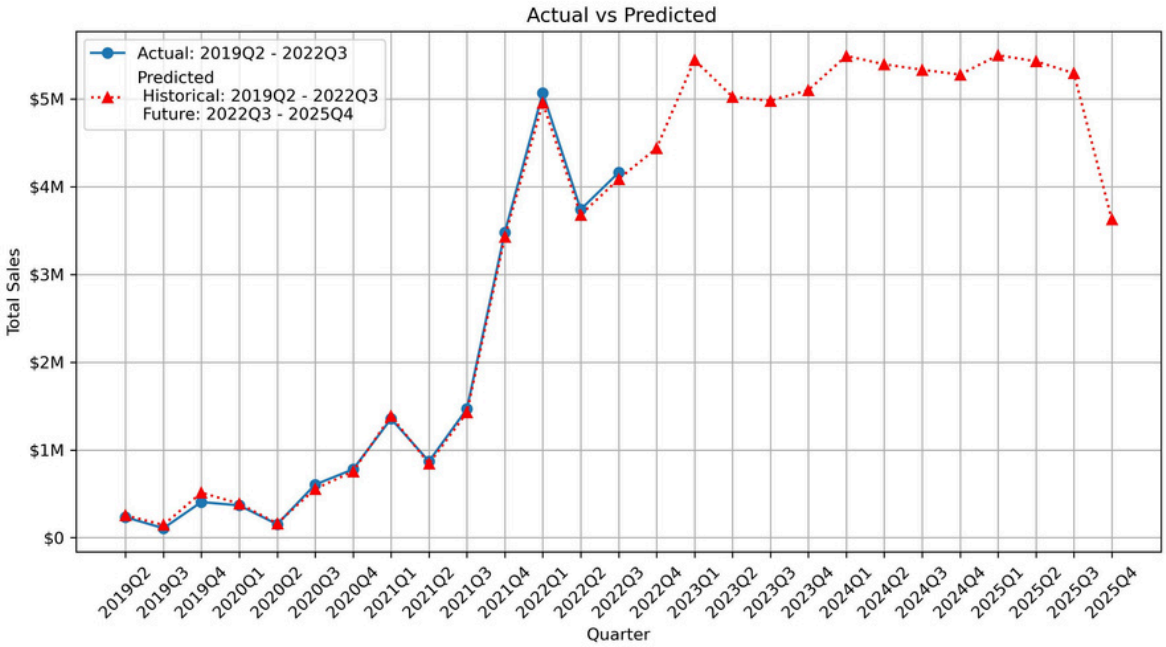
- **Yearly Prediction**
  - There is a continuous and clear upward trend in predicted total export sales from 2023 to 2025.
  - The growth trend will continue, albeit at a slightly slower pace, with an increase of approximately 19.10% from 2022 to 2023 and 4.97% from 2023 to 2024.
  - A slight decrease of approximately 7.70% in predicted total sales from 2024 to 2025 is expected, suggesting a potential growth or slowdown in market stabilization.
  - While the growth rates vary annually, the overall trend remains positive, indicating a healthy expansion in export activities.

Export Sales by Year



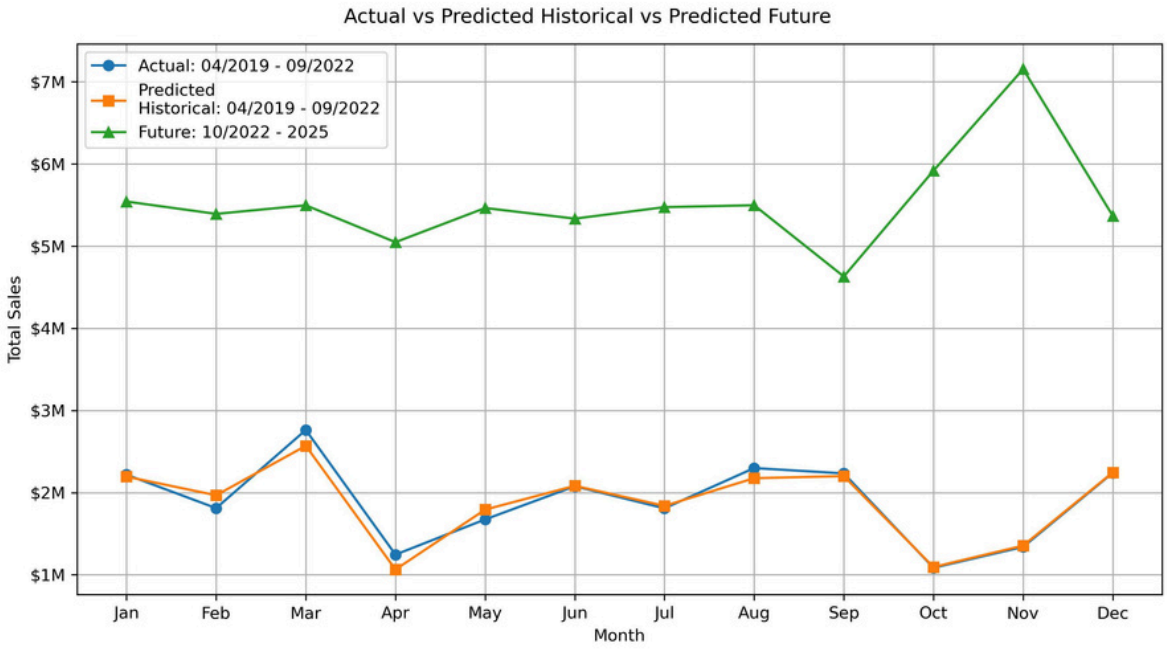
- **Quarterly Prediction**
  - Future total sales values suggest seasonal variations, with certain quarters showing higher sales than others.
  - From 2022Q2 to 2025Q3, there will be a stabilization in predicted total export sales.
  - A decline in total export sales is expected in the last quarter of 2025.

Export Sales by Quarter



- **Monthly Prediction**
  - There will be a continuous increase in total sales up to 2025.
  - The forecasted monthly total sales from January 2023 to December 2025 demonstrate a relatively stable trend compared to the actual total sales with variability and fluctuations.
  - There may be higher sales during certain months due to holidays or festivity, promotions, and seasonal demands.

Monthly Export Sales 2019 - 2025



5. Potential limitations

- The effectiveness of the models relies on the selection of relevant features. If significant predictors are overlooked or noisy variables are included, it can adversely impact the model's predictive power.
- While Random Forest can provide insights into feature importance, SARIMA models are often less interpretable, especially when dealing with complex time series patterns. This can limit the ability to explain the rationale behind specific predictions.
- SARIMA assumes a continuous time series, which may not always align with real-world business scenarios where abrupt changes or disruptions can occur.

6. Conclusion

Combining the two models adds robustness to the forecasting process, as one model can compensate for the shortcomings of the other in specific scenarios, resulting in more stable and reliable predictions overall. SARIMA is particularly adept at capturing time series patterns, seasonality, and autocorrelation, while Random Forest handles complex relationships, non-linearities, and feature interactions. This combined approach offers a comprehensive forecasting strategy that considers both time series-specific patterns and overall predictive modeling aspects, leading to better insights and informed decision-making based on the forecasts generated.

ABOUT THE AUTHOR

Temidayo Olowoyeye is a distinguished professional with a dual master's in Environmental Engineering and Agricultural Economics. With over 5+ years of experience, he is a proficient Data/Business Analyst dedicated to assisting individuals and organisations in translating data into actionable insights.

His expertise includes building compelling KPIs and target dashboards with PowerBI or Tableau, modifying and exploring data in relational databases using SQL, analysing big data and building models using Python or R, and preparing reports with key findings and recommendations capable of increasing business efficiency.

His unwavering commitment to excellence and passion for data-driven insights make him an asset to any organization seeking a seasoned Analyst with a diverse skill set.

Read more