Nielsen
BookData

TIME SERIES TECHNIQUES FOR FORECASTING SALES AND DEMAND IN NIELSEN BOOKSCAN

PREPARED BY:
ONYEKA MUOKA

AUGUST, 2025

TABLE OF CONTENT

INTRODUCTION

Nielsen BookScan tracks 90% of UK print book purchases, giving them the data to benchmark performance, monitor pricing, and spot market trends. Yet they lack the tools to turn this data into demand forecasts, leading to poor stock decisions and risky investment choices.

This project tests whether time-series models can uncover repeatable sales patterns and improve demand prediction. Using Auto-ARIMA, XGBoost, LSTM, and hybrid models, we forecast weekly sales for two sample titles; The Alchemist and The Very Hungry Caterpillar. The ARIMA model and one of the Hybrid ARIMA-LSTM models performed best.

Limited by compute and scope, the analysis focuses on historical sales and excludes media shocks. Also there was limited tuning in the models used. Results show how forecasting can guide smarter stock and print-run decisions.

PROBLEM STATMENT

Nielsen BookScan lack the analytical tools and resources needed to extract clear demand patterns from its extensive historical sales data. This leads to poor stocking decisions, costly over or under ordering, and uncertain investments in new titles, ultimately lowering profitability and market competitiveness.

This project aims to develop a time-series forecasting solution that uncovers repeatable sales patterns to support smarter procurement, reordering, and print-run decisions for the company.

DATA DESCRIPTION

Two Nielsen datasets were used: the ISBN Metadata File, containing descriptive book details, and the UK Weekly Sales File, with weekly sales data per ISBN.

The datasets were merged to link book metadata with sales performance for demand analysis. Data preparation included importing libraries, consolidating Excel tabs, resampling weekly sales to fill missing weeks, standardizing ISBNs, and converting dates to datetime format. Titles with sales beyond July 2024 were flagged. Sales trends were visualized to detect seasonal and market patterns. Two books; The Alchemist and The Very Hungry Caterpillar, were analyzed post-2012. The project aims to develop time-series forecasts to improve stocking, ordering, and investment decisions.

METHODOLOGY

This project applied a structured approach to analyze and forecast book sales using Nielsen BookScan data. Weekly sales were resampled to ensure consistent intervals, missing weeks filled with zeros, ISBNs standardized, and dates converted to datetime format. Titles with sales beyond July 2024 were identified and visualized.

The Alchemist and The Very Hungry Caterpillar, were selected for detailed analysis from 2012 onward. Classical time series techniques, including decomposition, ACF/PACF, stationarity testing, and Auto ARIMA, were applied. Machine learning and deep learning models (XGBoost, LSTM) and hybrid SARIMA-LSTM approaches were developed. Monthly forecasts were generated to compare against weekly predictions, with accuracy evaluated using MAE and MAPE.

UNDERSTANDING GENERAL SALES PATTERN

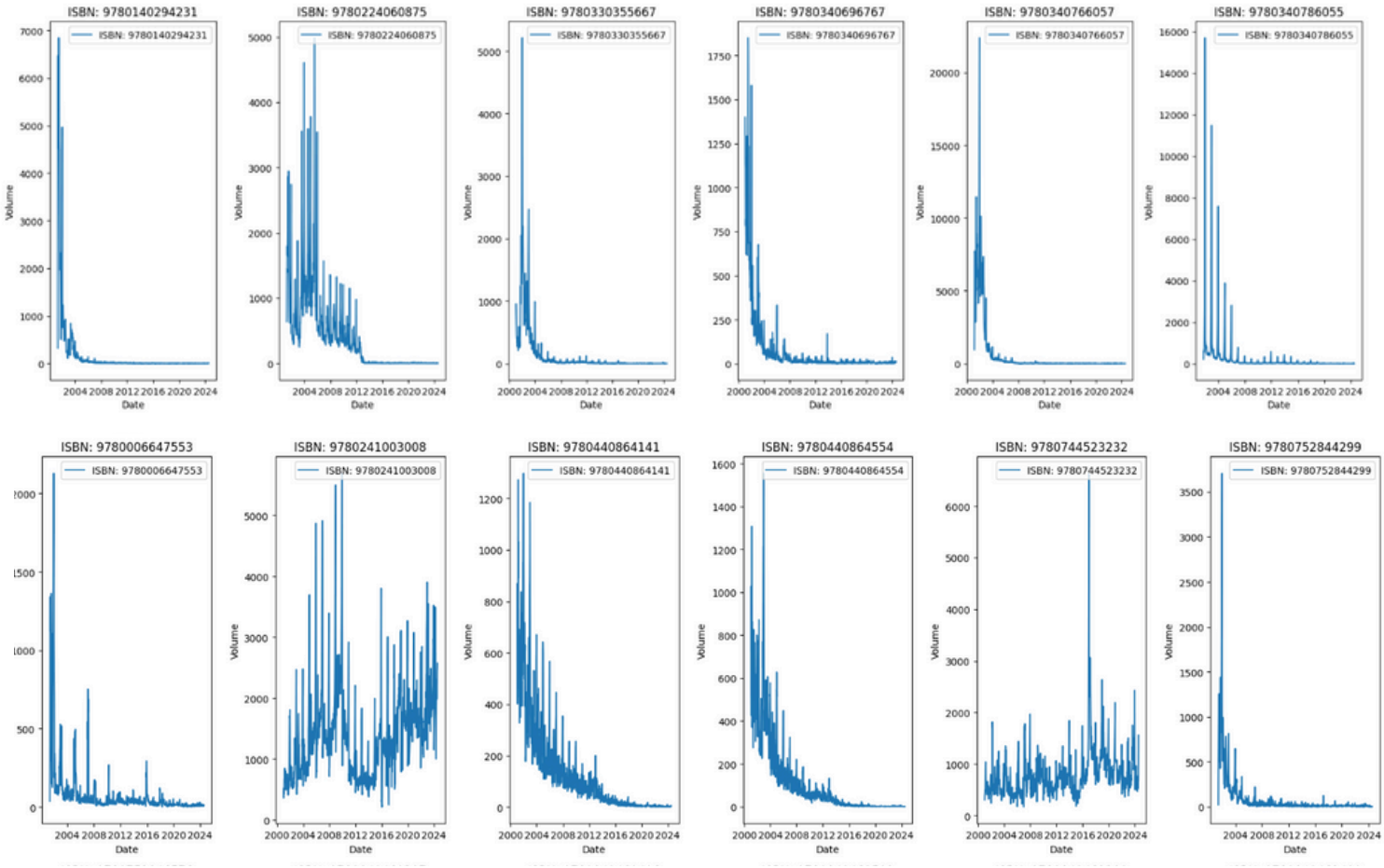


Figure 1. Sales pattern of some titles with sales beyond July 2024

The sales volume trends of books identified for sale after July 1, 2024, reveal a significant decline in the first 12 years, followed by a much slower, nearly flat decline from 2012 onward. This pattern aligns with broader industry trends, where print book sales have been decreasing due to factors like the rise of e-books and changing consumer preferences.

However, exceptions to this trend are notable. Children's and educational books, such as The Very Hungry Caterpillar (ISBN: 9780241003008) and We're Going on a Bear Hunt (ISBN: 9780744523232), have shown resilience in sales. These books are often used to establish early reading skills in children, suggesting that their enduring popularity may be linked to their educational value and continued demand in the market.

This phenomenon underscores the importance of considering genre-specific factors when analyzing sales trends. As not all books follow the same sales pattern. While most print books are selling less over time, some types of books, like educational or children’s books, still sell well. This shows that sales trends depend on the type of book, not just the overall market.

DECOMPOSITION, STAIONARITY, ACF AND PACF FOR BOTH BOOKS

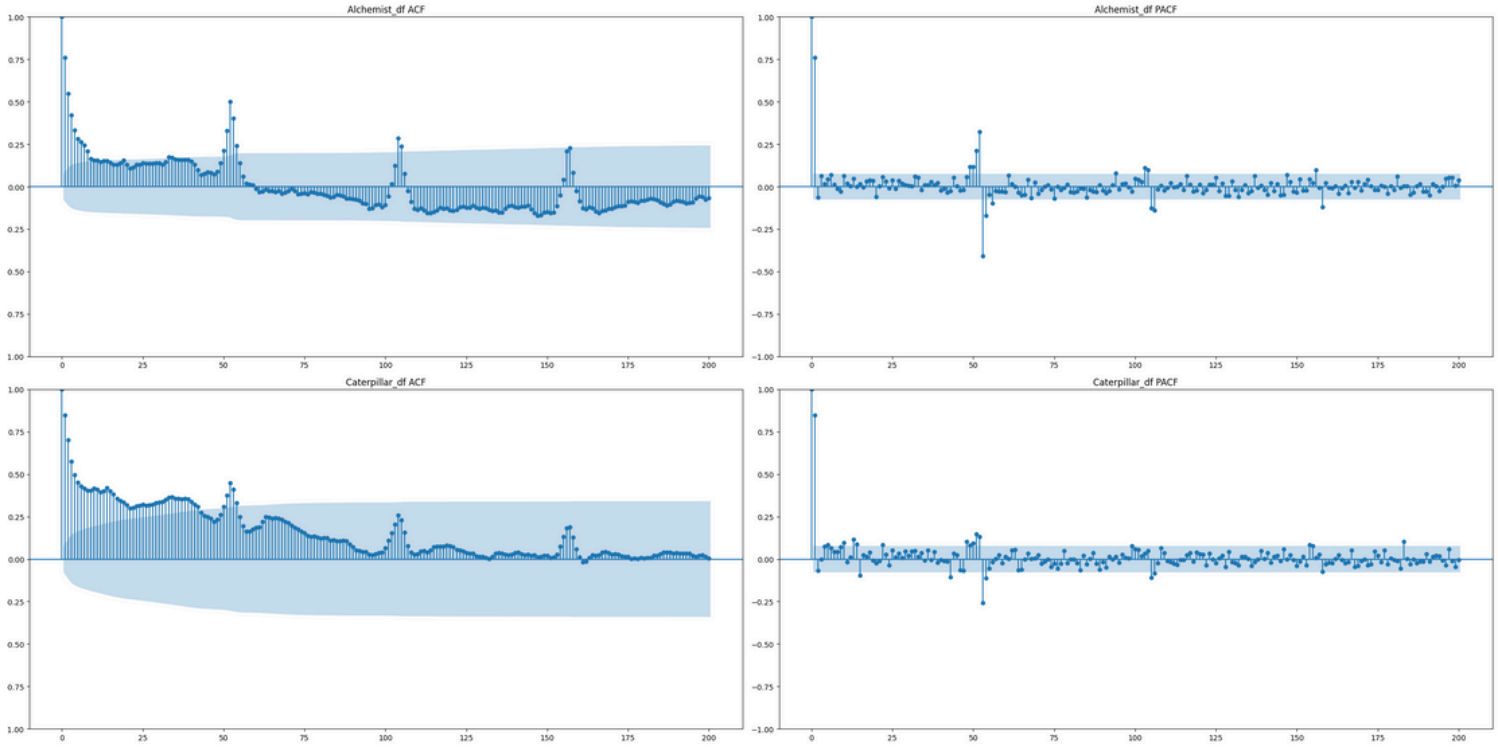


Figure 2. ACF and PACF plot for both books

Due to zero sales values during the 2020-2021 pandemic period, additive decomposition was employed for both datasets since multiplicative decomposition cannot handle zero values. ACF and PACF plots revealed strong autocorrelation with spikes beyond lag 50, indicating yearly seasonality in weekly data. ADF tests confirmed stationarity for both series: The Alchemist showed a test statistic far below critical values (p=0.0), while The Very Hungry Caterpillar had a statistic closer to zero but still significant (p=0.029), leading to rejection of non-stationarity in both cases.

CLASSICAL TIME SERIES ANALYSIS

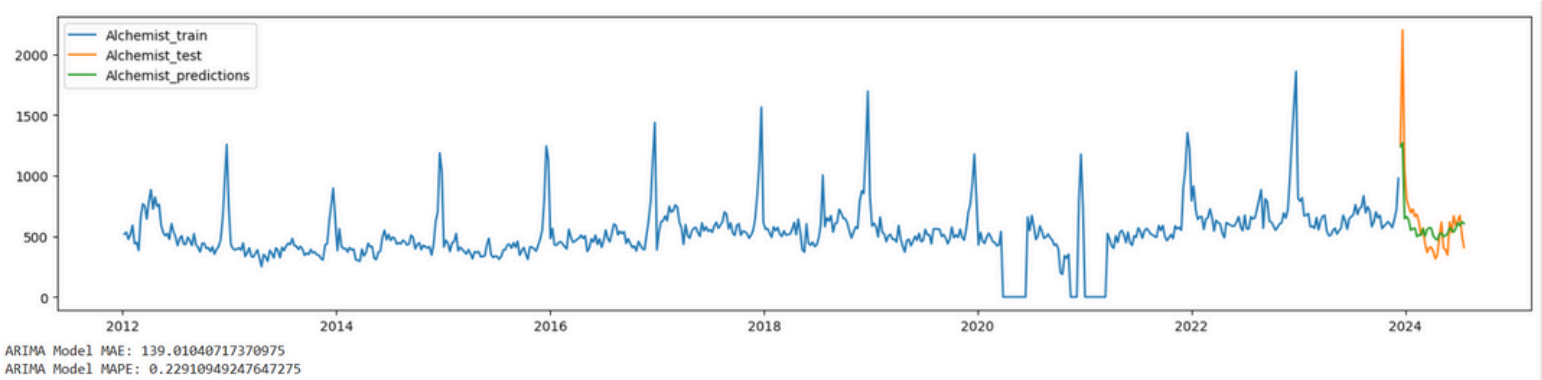


Figure 3. Seasonal Auto ARIMA Prediction for Alchemist dataset

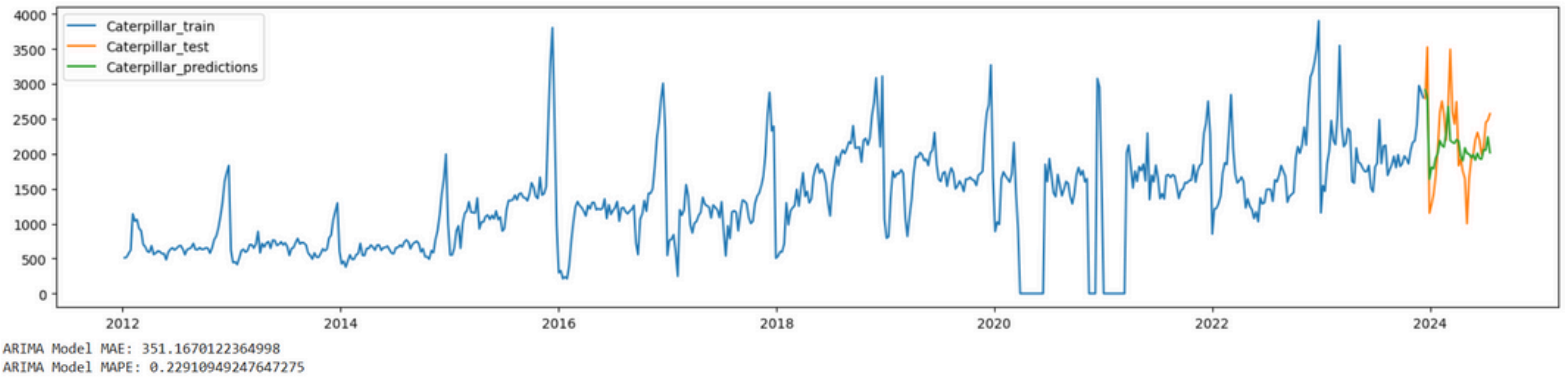


Figure 4. Seasonal Auto ARIMA Prediction for Caterpillar dataset

The best Auto ARIMA models capture key seasonal and trend patterns for both books. The Alchemist uses SARIMA(0,1,2)(1,0,1)[52], while The Very Hungry Caterpillar uses SARIMA(1,1,1)(1,0,1)[52]. Both show good fit via Ljung–Box tests, but elevated AIC, heteroskedasticity, and high kurtosis in residuals indicate potential for refinement. Residuals largely behave like white noise, yet variance and subtle seasonal patterns persist. Average MAPE is ~23%, showing predictions deviate moderately from actual sales.

MACHINE LEARNING (XGBOOST)

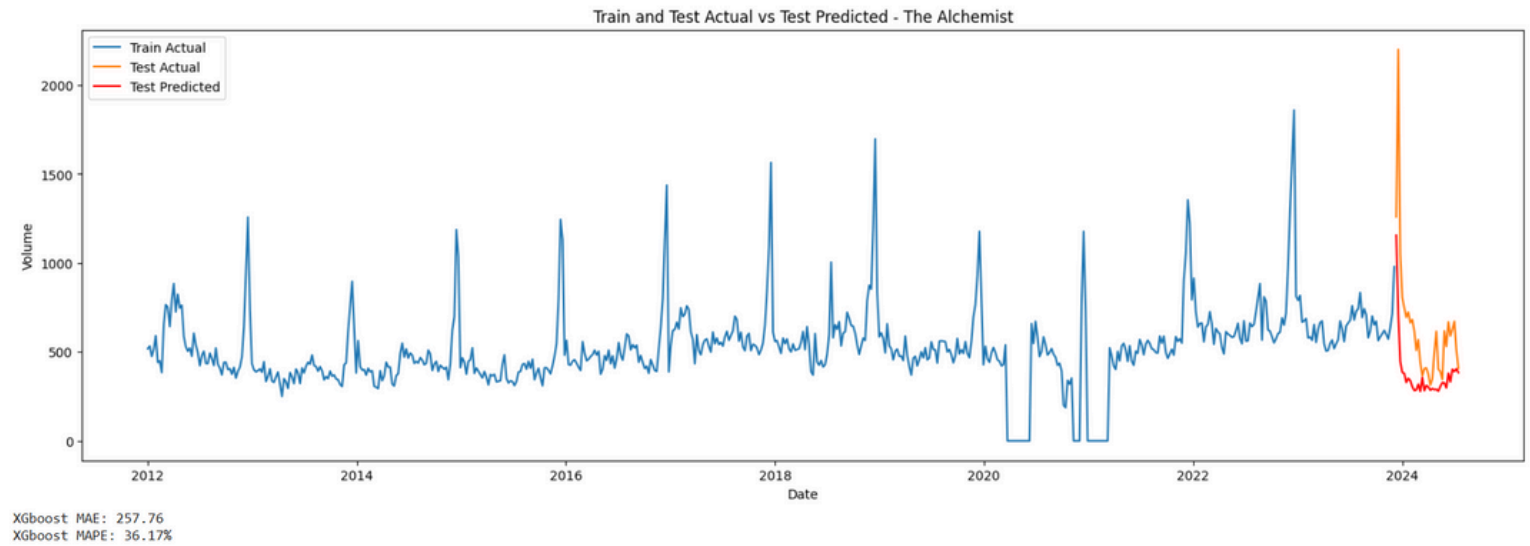


Figure 5. XGBoost Prediction for Alchemist dataset

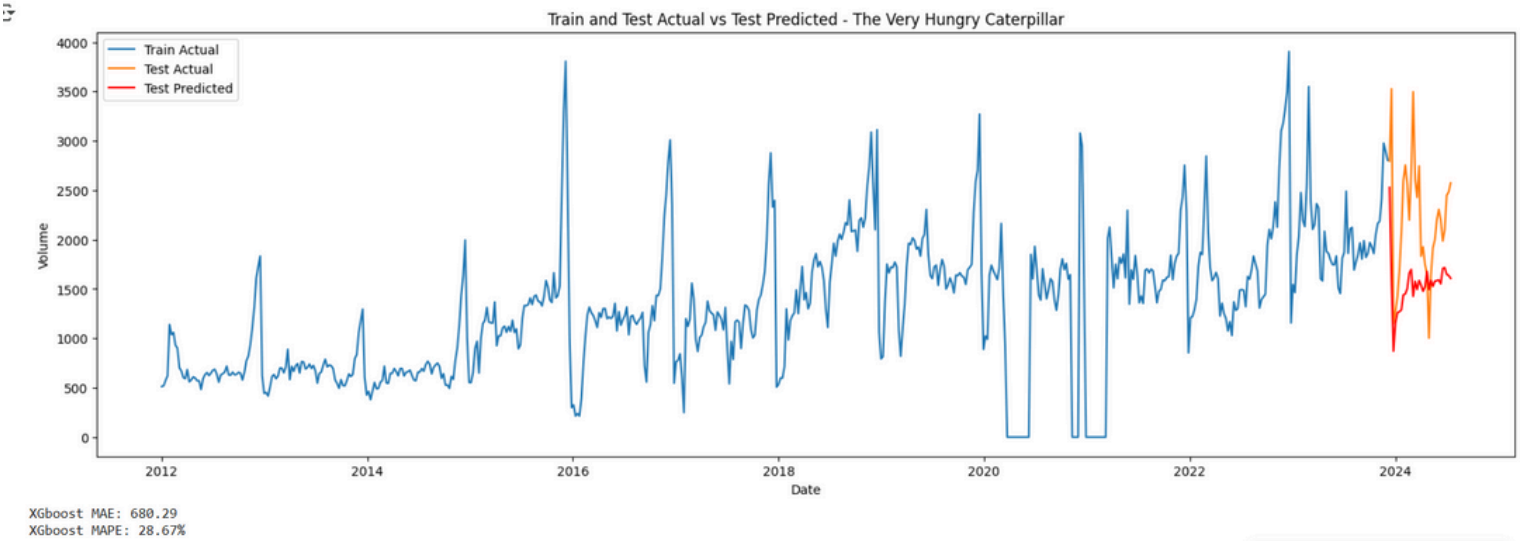
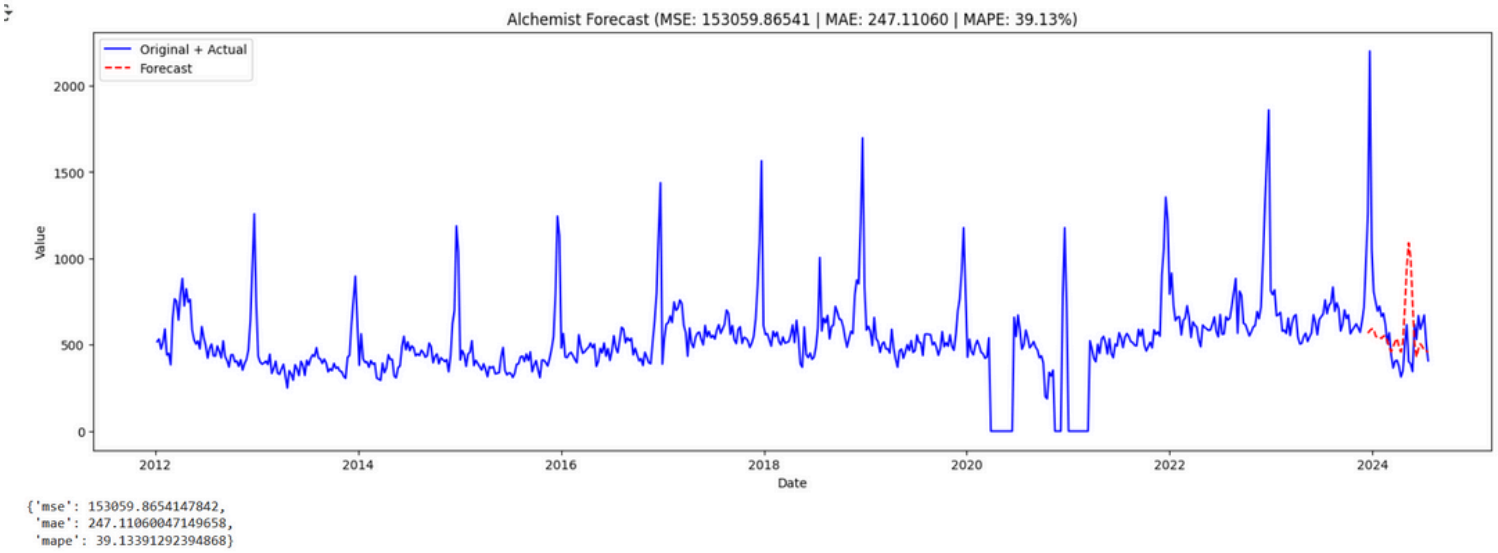


Figure 6. XGBoost Prediction for Caterpillar dataset

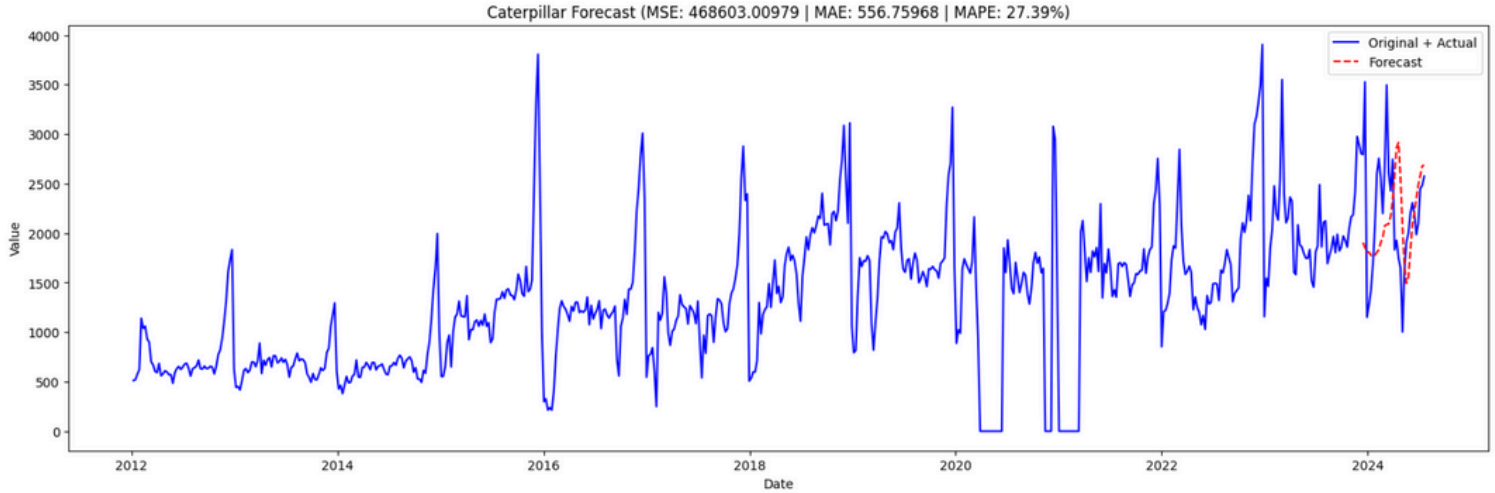
XGBoost underperformed compared to Auto ARIMA on both datasets. For The Alchemist, XGBoost achieved a MAPE of approximately 36%, versus 23% for Auto ARIMA. Similarly, for The Very Hungry Caterpillar, XGBoost reached 28% compared to Auto ARIMA's 23%. This gap likely arises because XGBoost is better suited for multi-feature regression than univariate time-series forecasting. The model used an additive deseasonalizer, which may limit accuracy with zero-sales weeks; a multiplicative approach could improve results if zero sales are handled first. Time constraints prevented testing this alternative.

DEEP LEARNING (LSTM)



{'mse': 153059.8654147842,
 'mae': 247.11060047149658,
 'mape': 39.13391292394868}

Figure 7. LSTM Prediction for Alchemist dataset



{'mse': 468603.0097884275, 'mae': 556.7596817016602, 'mape': 27.39067497531329}

Figure 8. LSTM Prediction for Caterpillar dataset

The LSTM model performed similarly to XGBoost on both datasets. For The Alchemist, it achieved a MAPE of approximately 40% and an MAE of 252, while for The Very Hungry Caterpillar, MAPE was 27% and MAE 556. This outcome is expected, as LSTMs generally require larger datasets and longer sequences for optimal performance. Only the number of LSTM units was tuned due to time constraints. Future work should explore additional hyperparameters, such as learning rate, layers, and dropout, to improve accuracy.

HYBRID MODELING

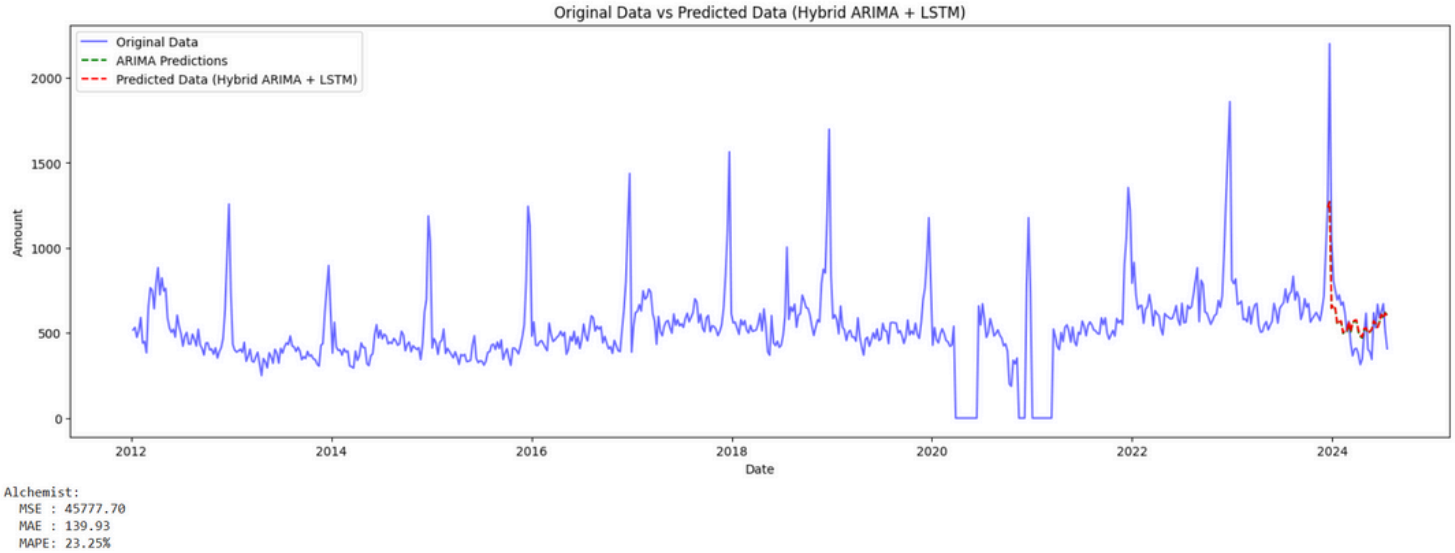


Figure 9. Hybrid ARIMA + LSTM(Residual) Prediction for Alchemist dataset

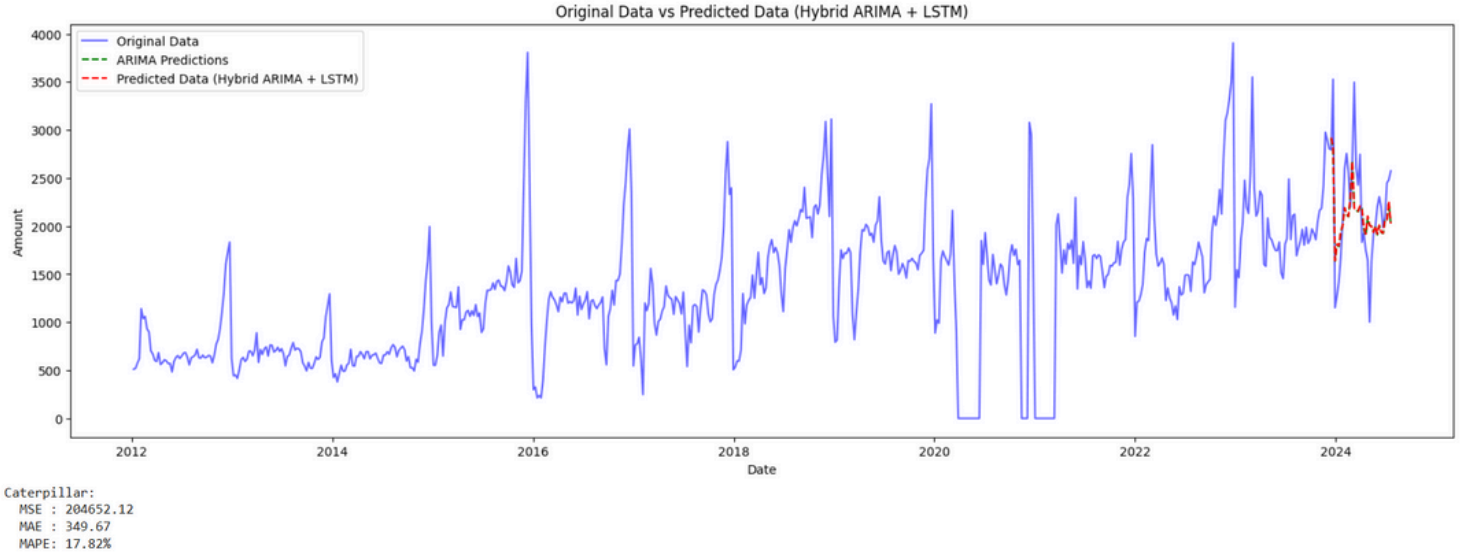


Figure 10. Hybrid ARIMA + LSTM(Residual) Prediction for Caterpillar dataset

The hybrid ARIMA + LSTM(Residual) model improved forecast accuracy for The Very Hungry Caterpillar, reducing MAPE from ~23% to ~18%, likely due to its variable sales patterns. In contrast, The Alchemist showed minimal gains, as ARIMA alone captured its stable trends. Parallel hybrid models with equal weighting underperformed, while weighted hybrids confirmed ARIMA's dominance. LSTM offered slight residual adjustments, but overall, hybrid benefits were dataset-dependent and modest.

MONTHLY PREDICTION

For further analysis, both datasets were resampled into monthly data and modeled using XGBoost and seasonal Auto ARIMA.

For The Alchemist, XGBoost produced a monthly MAE of approximately 826.6 and a MAPE of 40.4%, while ARIMA achieved slightly better performance with an MAE of 751.36 and a MAPE of 32.24%.

For The Very Hungry Caterpillar, both models performed better relative to The Alchemist in terms of the MAPE. XGBoost reached a MAE of 1814.3 and MAPE of 17.9%, and ARIMA had a MAE of 1954.26 and MAPE of 20.35%.

The improved performance on Caterpillar MAPE likely reflects its higher sales volume and more regular seasonal patterns, which provide clearer signals for both models to capture. In contrast, The Alchemist exhibits lower and more irregular sales, making predictions more challenging.

Note: Refer to the appendix for timeseries plots of monthly forecasts.

CONCLUSION AND RECOMMENDATION

Conclusion

This analysis demonstrates that time-series forecasting can significantly improve Nielsen BookScan's demand prediction capabilities. Seasonal Auto ARIMA and the hybrid ARIMA + LSTM (residual) models emerged as the top performers, with weekly predictions consistently outperforming monthly forecasts. The hybrid model particularly excelled for The Very Hungry Caterpillar, reducing MAPE from 23% to 18%, while Auto ARIMA proved optimal for The Alchemist's stable patterns.

Recommendations

Nielsen BookScan should implement a flexible modeling framework that selects different algorithms based on individual book sales behavior patterns. The company must establish clear criteria for model selection, deciding whether to prioritize MAE or MAPE based on business objectives and cost implications. Collaboration with industry experts would enhance model development and provide domain-specific insights into seasonal patterns and market dynamics. Additionally, investing in increased computing resources is essential to enable comprehensive hyperparameter tuning and testing of more sophisticated models, which was limited in this project. This strategic approach would optimize inventory management and reduce costly over or under-ordering decisions.

APPENDIX

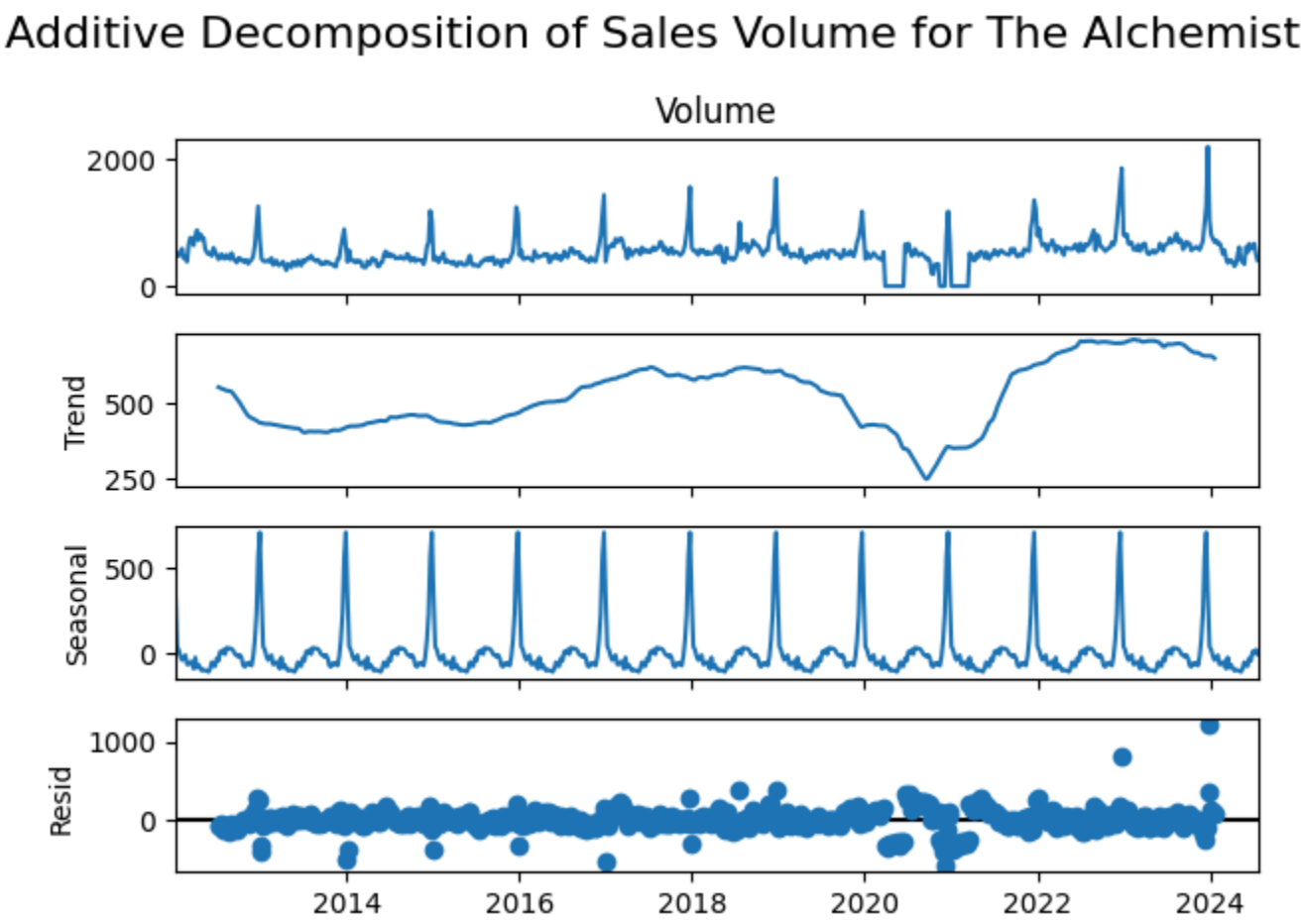


Figure 11. Additive decomposition for Alchemist dataset

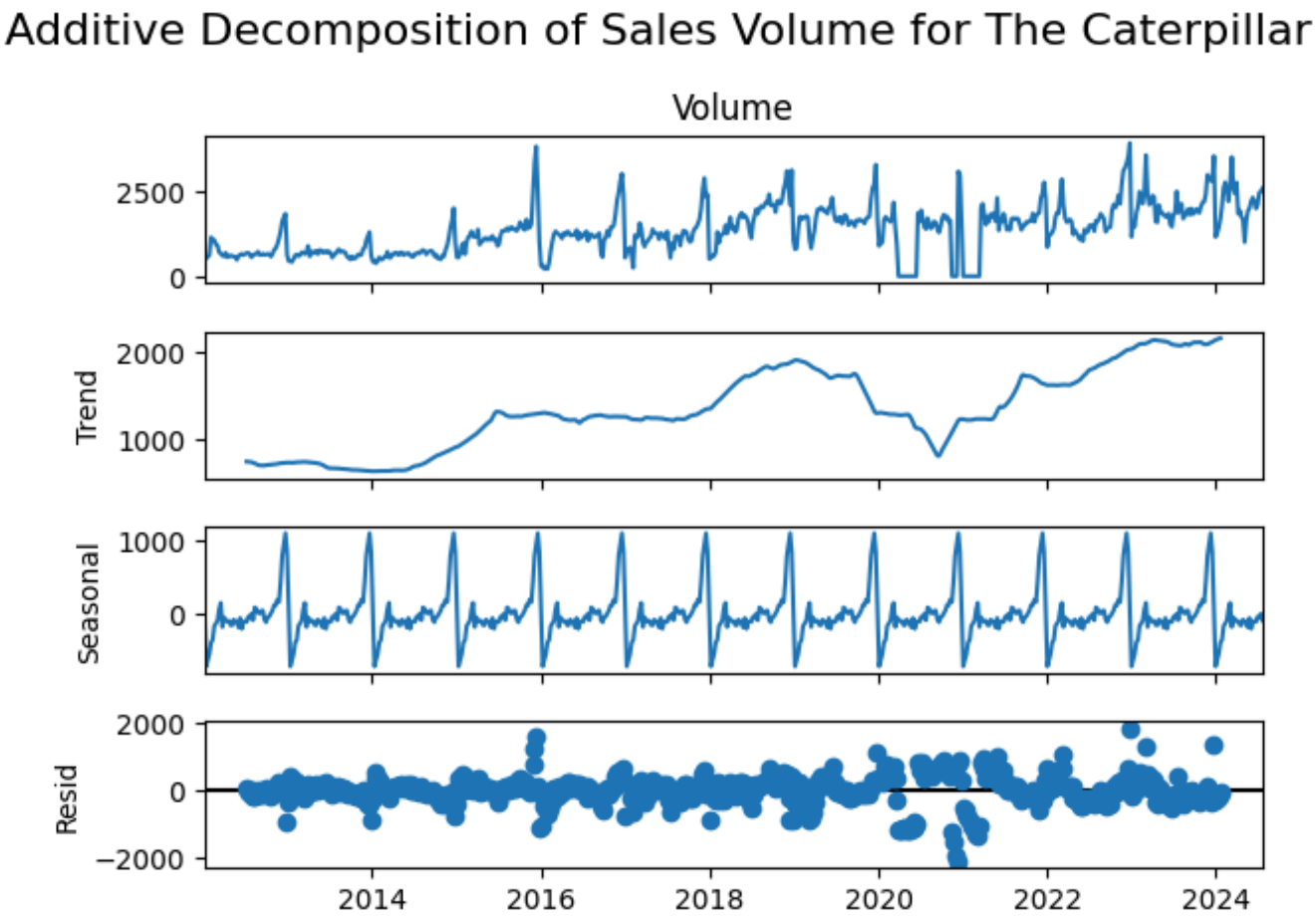


Figure 12. Additive decomposition for Caterpillar dataset

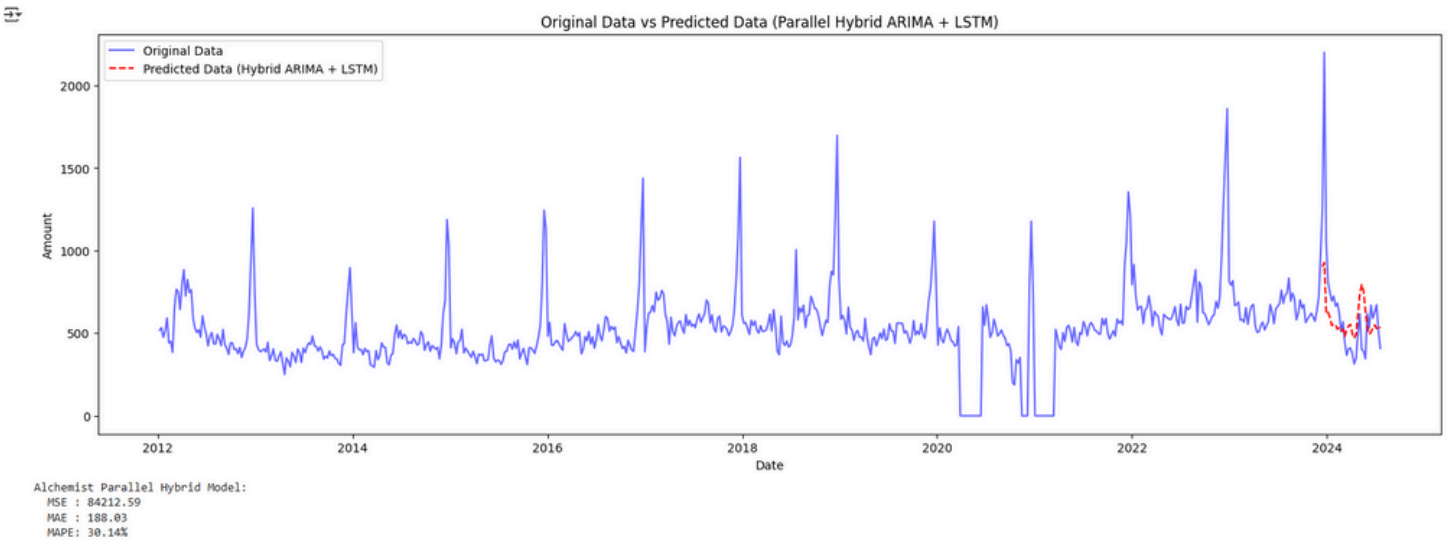


Figure 13. Parallel Hybrid Model Prediction for Alchemist dataset

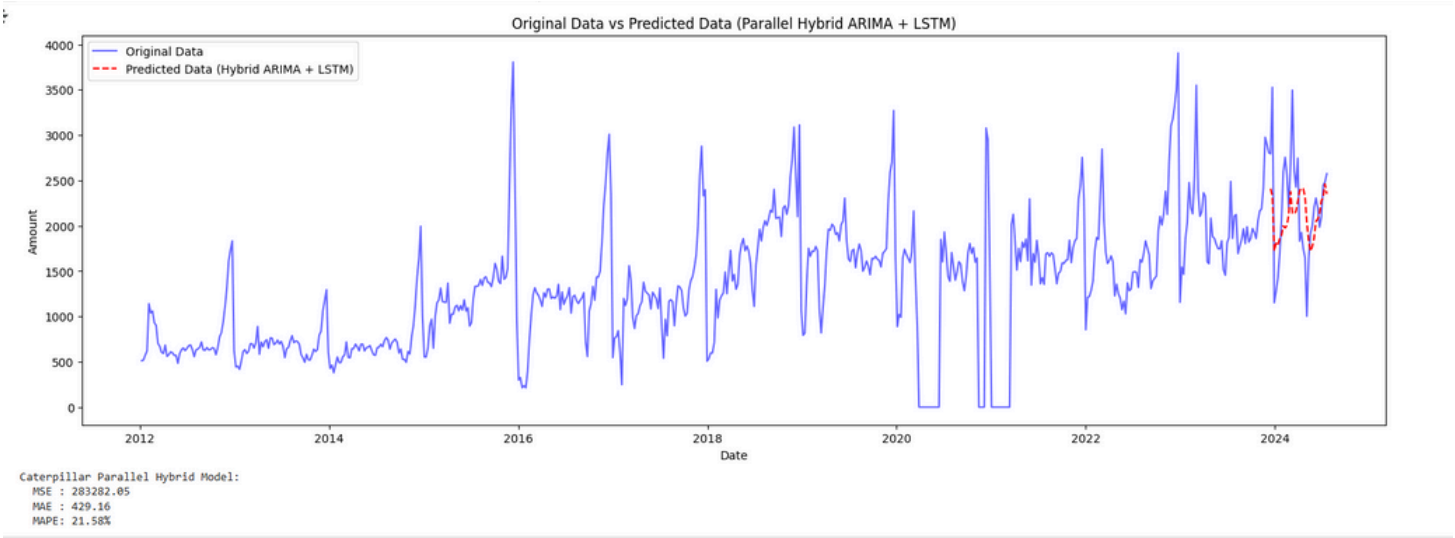


Figure 14. Parallel Hybrid Model Prediction for Caterpillar dataset

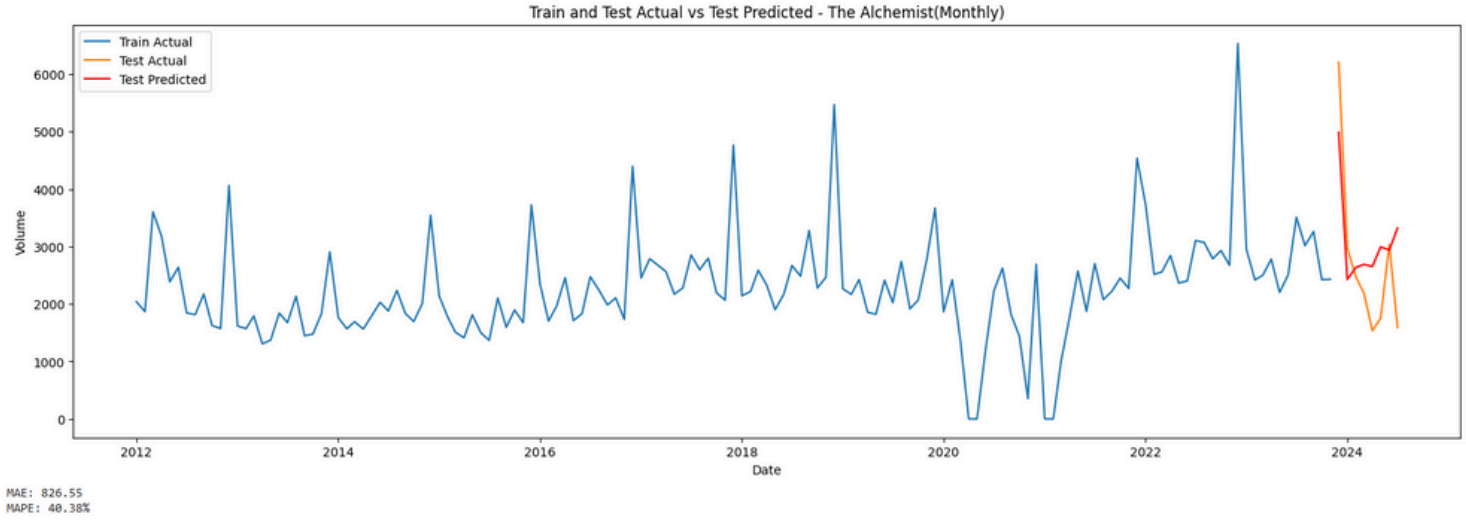


Figure 15. XGBoost Prediction for Monthly Alchemist dataset

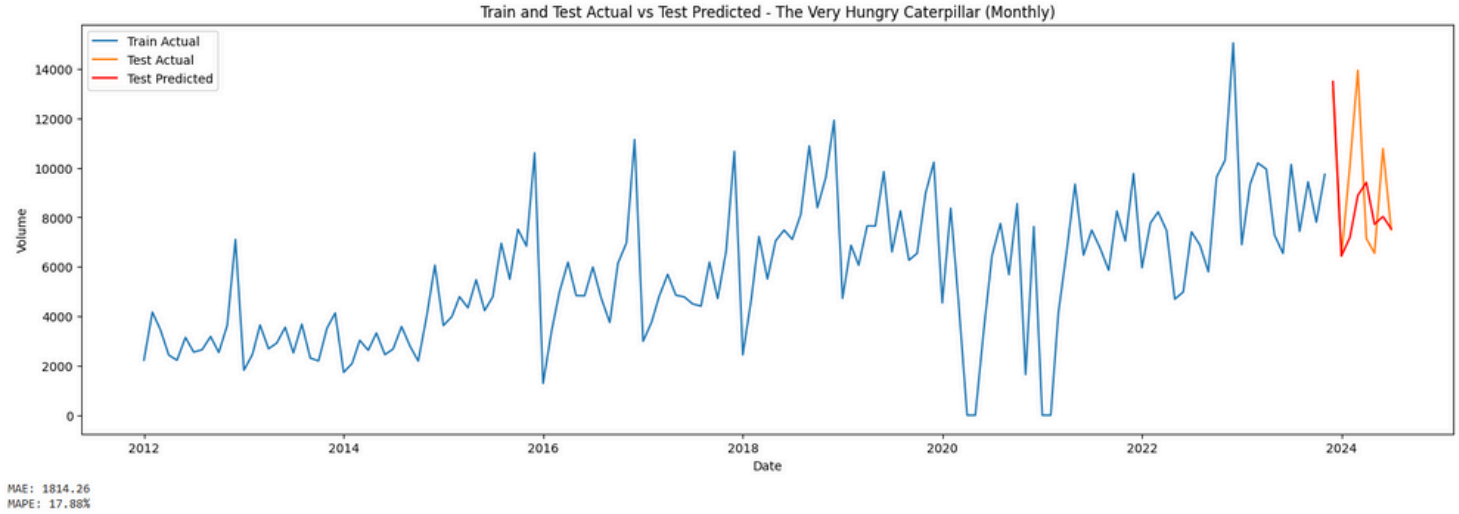


Figure 16. XGBoost Prediction for Monthly Caterpillar dataset

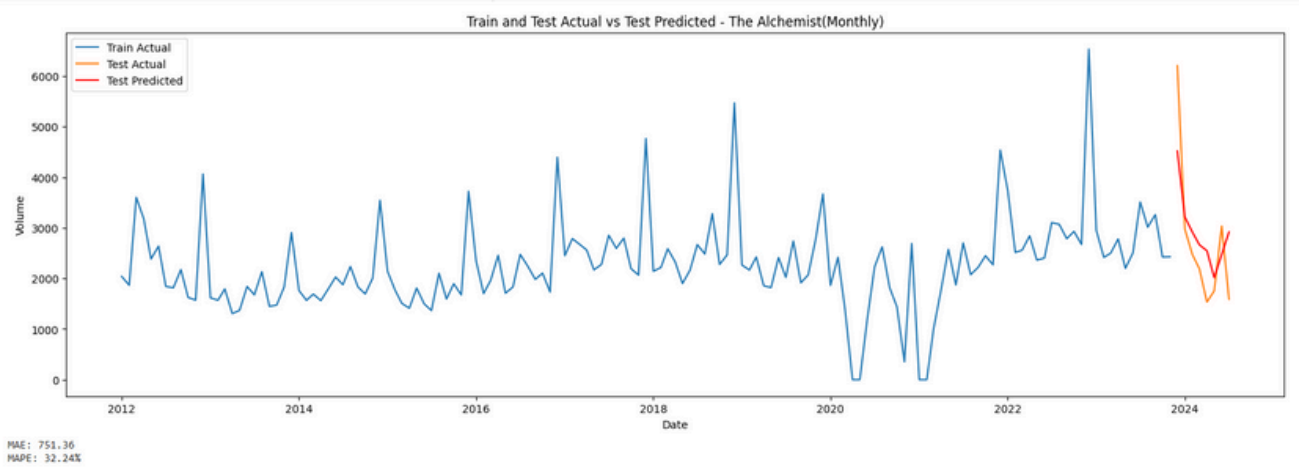


Figure 17. Auto ARIMA Prediction for Monthly Alchemist dataset

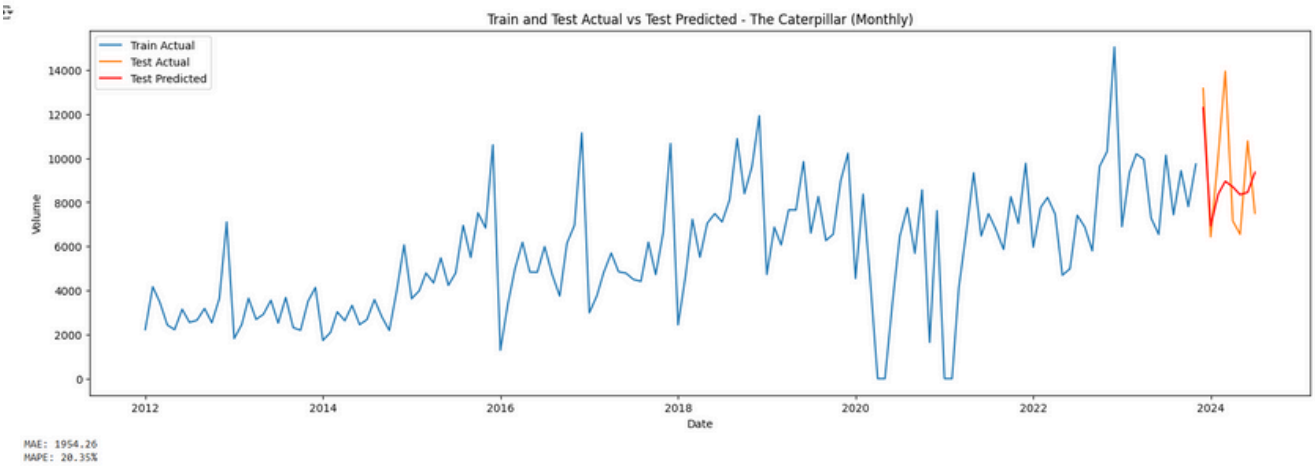


Figure 18. Auto ARIMA Prediction for Monthly Caterpillar dataset