DATA CAREER
ACCELERATOR

Course 3 Project: Using time series analysis for sales and demand forecasting

Juan Pablo Salazar

2025/04/14

FOURTH
REV

UNIVERSITY OF
CAMBRIDGE
INSTITUTE OF
CONTINUING EDUCATION

I. Introduction and problem statement:

This project analyzes and predicts weekly and monthly sales for two books of interest, *The Alchemist* and *The Very Hungry Caterpillar*, using real-world data from BookScan. We applied classical time series analytical methods, such as decomposition and ARIMA, alongside machine learning approaches, including XGBoost and Long Short-Term Memory (LSTM) neural networks. We employed hyperparameter tuning and explored hybrid forecasting strategies using sequential and parallel combinations of LSTM and SARIMA.

II. Data investigation

After resampling volume sales to weekly frequency, and filled in missing weeks with 0 sales. The dataset was filtered to identify books with sales beyond 2024-07-01, resulting in 61 unique ISBNs. These were visualized to identify patterns across time (Fig.1). The sales data shows only three books sold more than 500 copies per week after July 2024. Furthermore, comparing the 2001–2012 and 2012–2024 periods (Figs 2 and 3), a clear decline in sales volume is observed. While in the early 2000s weekly sales reached up to 200,000; the latter period peaked at just around 6,500 copies per week, even for the best-selling titles.

Next, we analyzed the sales of the two selected books after 2012 (Fig. 5) and applied seasonal decomposition to both time series (Figs. 6 and 7). The time series for *The Very Hungry Caterpillar* shows an increasing trend with seasonal fluctuations that grow in magnitude over time, indicating heteroscedasticity. This justified the use of multiplicative decomposition. However, several weeks during the COVID-19 lockdown recorded zero sales, which are incompatible with the multiplicative model due to division by zero. To address this, we applied spline interpolation to impute these values, enabling multiplicative decomposition. Additive decomposition was performed on *The Alchemist* data, both data sets show a seasonal component, stationary residuals, and an overall increasing trend that goes down during the COVID period.

For both books, the ACF plots (Fig. VII) show autocorrelation that gradually decays. In *The Very Hungry Caterpillar*, nearly all visible lags remain outside the insignificance region, suggesting that past values have a prolonged influence on future values. In contrast, *The Alchemist* autocorrelations fall within the insignificance bounds around lag 10, indicating that only the first few lags significantly influence future values.

In the PACF plots (Fig. VIII), both books show the first lag outside the insignificance region. For *The Alchemist*, all subsequent lags fall within the bounds, but for *The Very Hungry Caterpillar*, a few additional lags slightly exceed the bounds, possibly indicating weak seasonal effects.
Finally, applying the ADF test to both book time series, we obtained p-values below 0.05, indicating that both series are stationary. This means their mean and variance remain constant, making them suitable for SARIMA statistical modeling without differencing.

III. Classical and machine learning forecast

Auto ARIMA, XGBoost, LSTM, and hybrid methods combining LSTM and ARIMA were applied to both datasets to forecast the final 32 weeks of the database and compare the results with the actual values.

Auto ARIMA successfully identified appropriate seasonal ARIMA models for both datasets. Table I summarizes the parameter space and auto ARIMA model results. While both models pass residual checks, having p-values under 0.05 on the Ljung-Box test. *The Alchemist* exhibits more predictable sales, reflected in its lower MAPE (20.60%). This suggests it is more suited for ARIMA-style forecasting. In contrast, The *Very Hungry Caterpillar* shows a higher MAPE (33.16%), indicating that more advanced or nonlinear models may be more appropriate. Additionally, the residual plots reveal greater variance in *The Caterpillar* model, suggesting that its ARIMA specification may not fully capture the underlying dynamics of the series.

Moreover, we employed XGBoost, a tree-based decision model known for its robustness and efficiency in capturing complex nonlinear patterns. To optimize its performance, hyperparameter tuning with grid search was applied. The best models achieved a Mean Absolute Percentage Error (MAPE) of 8.42% for The Very Hungry Caterpillar and 16.01% for The Alchemist, as summarized in Table II.

We then applied a Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN) capable of learning long-term temporal dependencies thanks to its memory cell architecture. Through hyperparameter optimization of layers, units, dropouts, and learning rate with random search, we achieved a MAPE of 29.06% for The Very Hungry Caterpillar and 59.87% for The Alchemist with the best-performing LSTM models, summarized Table III.

Aiming to enhance predictive accuracy, we combined SARIMA (a classical time series model capturing seasonality and trends) and LSTM models through hybrid approaches. We explored both sequential and parallel combinations. The best hybrid model for The Very Hungry Caterpillar was a parallel model with SARIMA weight 0.25 and LSTM weight 0.75, yielding a MAPE of 21.00%. For The Alchemist, the best result came from a parallel combination with SARIMA weight 0.75 and LSTM weight 0.25, achieving a MAPE of 20.76%. These results are summarized in Table IV.

Finally, we employed monthly-level predictions using both SARIMA and XGBoost. For The Very Hungry Caterpillar, SARIMA produced a MAPE of 20.15%, while XGBoost achieved 11.57%. For The Alchemist, SARIMA resulted in a MAPE of 32.24%, and XGBoost achieved a notably better performance with a MAPE of 20.15%.

When comparing weekly and monthly predictions (Tables V and VI), XGBoost consistently performs well at both, although the model had higher accuracy at the weekly level. Meanwhile, SARIMA's performance degraded in the monthly prediction, particularly for *The Alchemist*, which exhibited approximately 10% more error. This suggests that while monthly predictions are useful for capturing broader sales trends, weekly forecasts provide finer accuracy and better adapt to short-term fluctuations, especially when using machine learning models like XGBoost.

Conclusion

While all forecasting methods achieved acceptable results with most MAPE values between 20% and 30%, XGBoost emerged as the most effective model, capturing complex, nonlinear sales patterns. In particular, weekly predictions with XGBoost consistently outperformed all other approaches, achieving the lowest error rates for both titles. (Best model performances are summarized in Tables V and VI).

Appendix:

Tables

Table I. ARIMA Model Comparison – *The Very Hungry Caterpillar* vs. *The Alchemist*

Feature	**The Very Hungry Caterpillar**	**The Alchemist**
Parameter Search Space	p: 0–5, d: 0, q: 0–5	p: 0–1, d: 0, q: 0–1
	P: 0–2, D: 0, Q: 0–2	P: 0–1, D: 0, Q: 0–1
Seasonal Frequency (m)	52 (weekly)	52 (weekly)
Best Model Chosen	(1, 0, 0) × (2, 0, [1], 52)	(1, 0, 1) × (1, 0, [], 52)
AIC	7571.403	7678.056
Ljung-Box Test , p-value	1.51 (p = 0.22)	0.06 (p = 0.80)
Heteroskedasticity (p-value)	3.45 (p = 0.00)	2.92 (p = 0.00)
Mean Absolute Error (MAE)	760.67	146.65
Mean Absolute Percentage Error	33.16%	20.60%

Table II. XGBoost Parameter Tuning and Best Results for Both Titles

Parameter	**Grid Search Values**	**Best Value – Caterpillar**	**Best Value – Alchemist**
n_estimators	[50, 100, 200]	200	200
learning_rate	[0.02, 0.03, 0.05, 0.1]	0.02	0.02
max_depth	[3, 4, 5, 6]	3	3
subsample	[0.5, 0.7, 1.0]	1	1
colsample_bytree	[0.5, 0.7, 1.0]	1	1
Window Lengths Tested	[2, 3, 4, 6, 12, 24]	24	2
Best MAE	–	**194.96**	**81.64**
Best MAPE	–	**8.42%**	**16.01%**

Table III. LSTM Parameter Tuning and Best Results for Both Titles

Parameter	Search Range	Best Value – Caterpillar	Best Value – Alchemist
num_lstm_layers	[1, 2, 3]	3	2
units_lstm_0	[4, 36, 68, 100, 128]	100	36
dropout_lstm_0	[0.1, 0.2, 0.3, 0.4, 0.5]	0.2	0.3
units_lstm_1	[4, 36, 68, 100, 128]	68	68
dropout_lstm_1	[0.1, 0.2, 0.3, 0.4, 0.5]	0.1	0.1
units_lstm_2	[4, 36, 68, 100, 128]	100	–
dropout_lstm_2	[0.1, 0.2, 0.3, 0.4, 0.5]	0.5	–
optimizer	['adam', 'rmsprop']	rmsprop	adam
learning_rate	[0.0001 – 0.01] (log-uniform)	0.00245	0.00206
Best MAE	–	**578.81**	**332.69**
Best MAPE	–	**29.06%**	**59.87%**

Table IV. Hybrid Model Results: SARIMA + LSTM Combinations

Dataset	Model Combination	SARIMA Weight	LSTM Weight	MAE	MAPE
Caterpillar	Sequential (ARIMA residuals → LSTM)	–	–	751.21	33.06%
	Parallel Combination	0.5	0.5	658.39	28.33%
	Parallel Combination	0.75	0.25	695.17	30.21%
	Parallel Combination	**0.25**	**0.75**	**489.86**	**21.00%**
Alchemist	Sequential (ARIMA residuals → LSTM)	–	–	139.46	20.06%
	Parallel Combination	0.5	0.5	127.64	20.82%
	Parallel Combination	**0.75**	**0.25**	**117.93**	**20.76%**
	Parallel Combination	0.25	0.75	150.87	20.86%

Table V. Best Model Performance – *The Very Hungry Caterpillar*

Model Type	**Parameters / Configuration**	**MAE**	**MAPE**
Best ARIMA	ARIMA(1,0,0)(2,0,1)[52]	760.67	33.16%
Best XGBoost	**n_estimators=200, learning_rate=0.02, max_depth=3, subsample=1.0, colsample_bytree=1.0, window=24**	**194.96**	**8.42%**
Best LSTM	num_lstm_layers=3, units=[100, 68, 100], dropouts=[0.2, 0.1, 0.5], optimizer=rmsprop, lr=0.0025	578.81	29.06%
Best Hybrid	Parallel combination SARIMA: 0.25, LSTM: 0.75	489.86	21.00%
Monthly XGBoost	n_estimators=50, learning_rate=0.05, max_depth=3, colsample_bytree=1.0, window=6	1000.31	11.57%
Monthly SARIMA	ARIMA(2,0,1)(0,0,2)[12]	2166	20.15%

Table VI. Best Model Performance – *The Alchemist*

Best ARIMA	ARIMA(1,0,1)(1,0,0)[52],	146.65	20.60%
Best XGBoost	**n_estimators=200, learning_rate=0.02, max_depth=3, subsample=1.0, colsample_bytree=1.0, window=2**	**81.64**	**16.01%**
Best LSTM	num_lstm_layers=2, units=[36, 68], dropouts=[0.3, 0.1], optimizer=adam, lr=0.0021	332.69	59.87%
Best Hybrid	Parallel combination SARIMA: 0.75, LSTM: 0.25	117.93	20.76%
Monthly XGBoost	n_estimators=100, learning_rate=0.02, max_depth=3, colsample_bytree=1.0, window=5	395.06	20.15%
Monthly SARIMA	SARIMA(1,0,1)(0,0,2)[12], with intercept	751.36	32.24%

Figures:

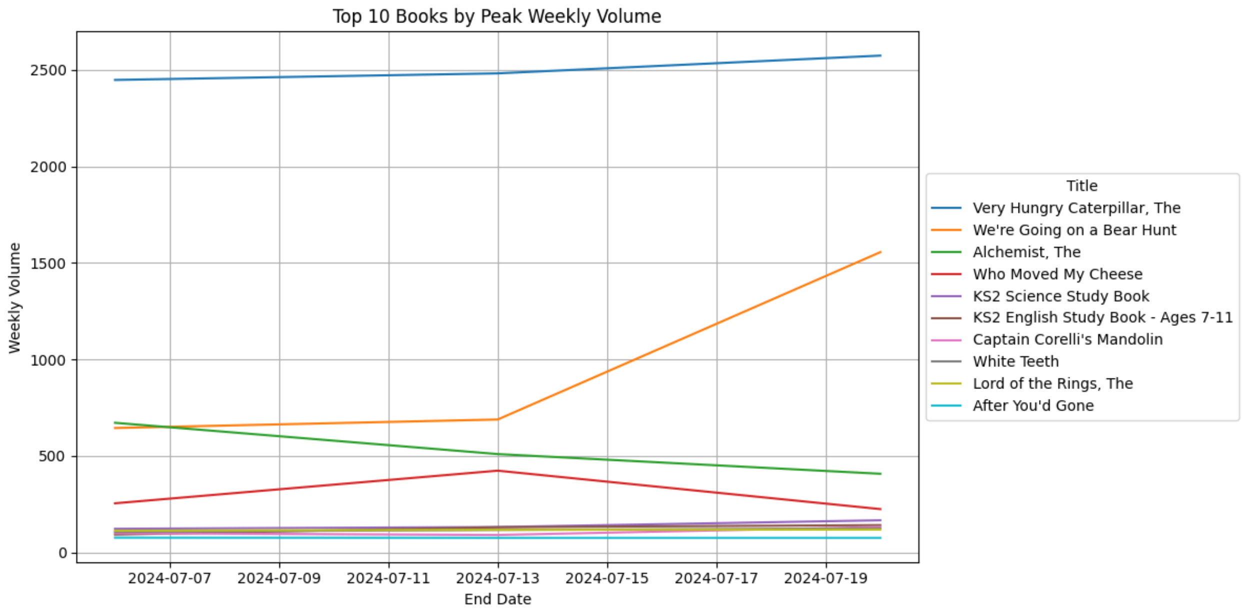

Figure I. Weekly sales volume for the top 10 best-selling books after January 1, 2024

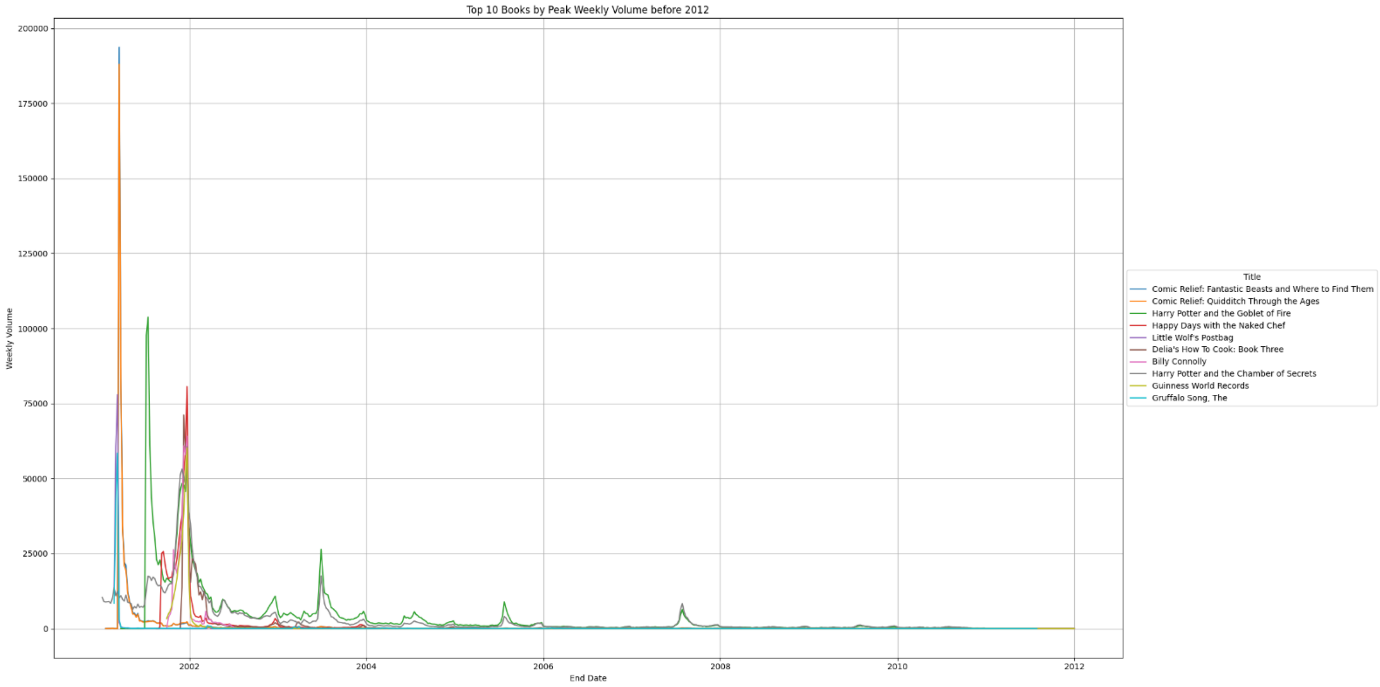

Figure II. Weekly sales volume for the top 10 best-selling books between 2001 and 2012

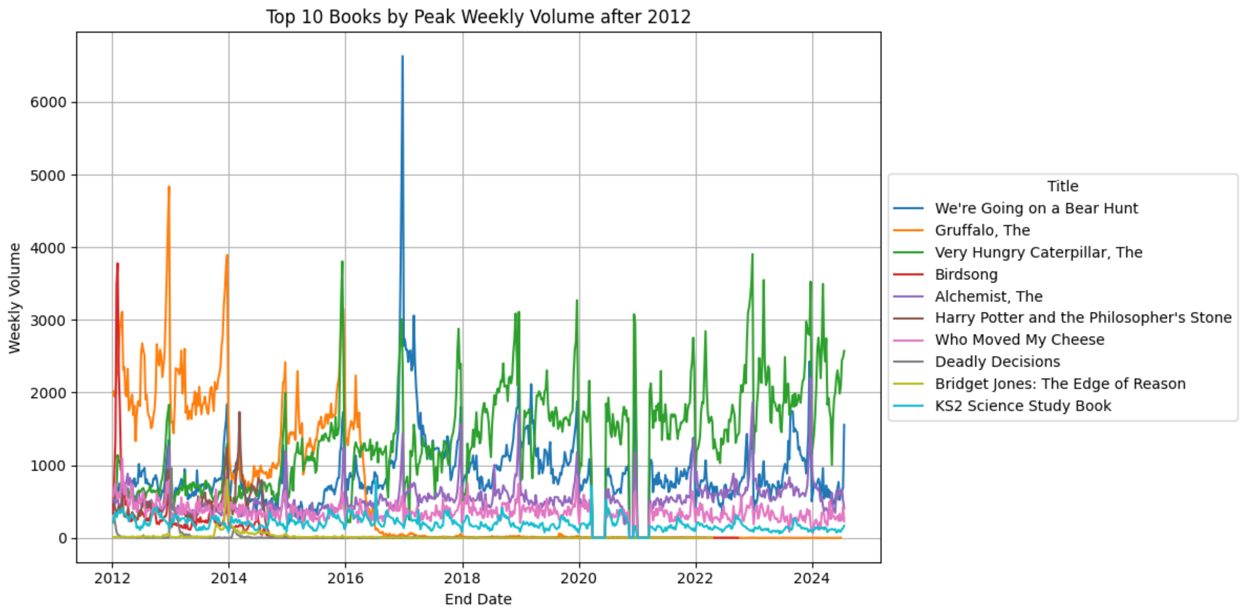

Figure III. Weekly sales volume for the top 10 best-selling books between 2012 and 2024

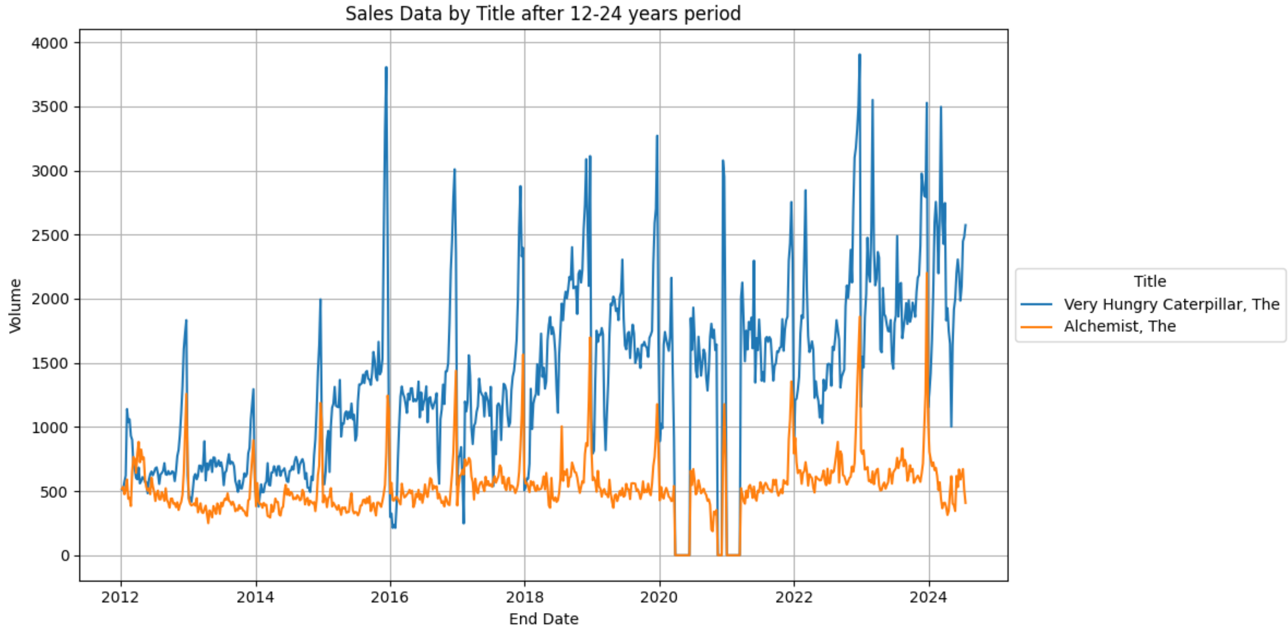

Figure IV. Weekly sales volumes for *The Very Hungry Caterpillar* and *The Alchemist* after 2012.

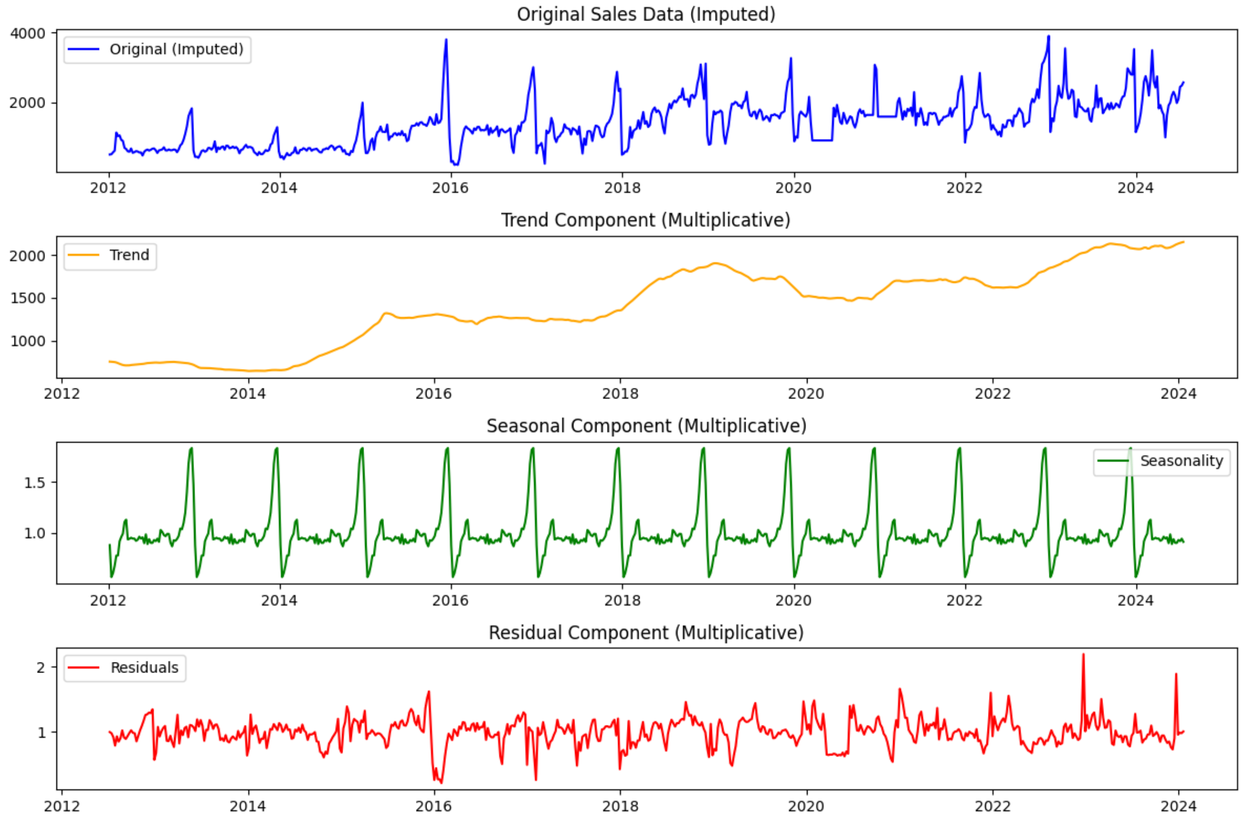

Figure V. Multiplicative decomposition of *The Very Hungry Caterpillar* book sales time series after 2012.

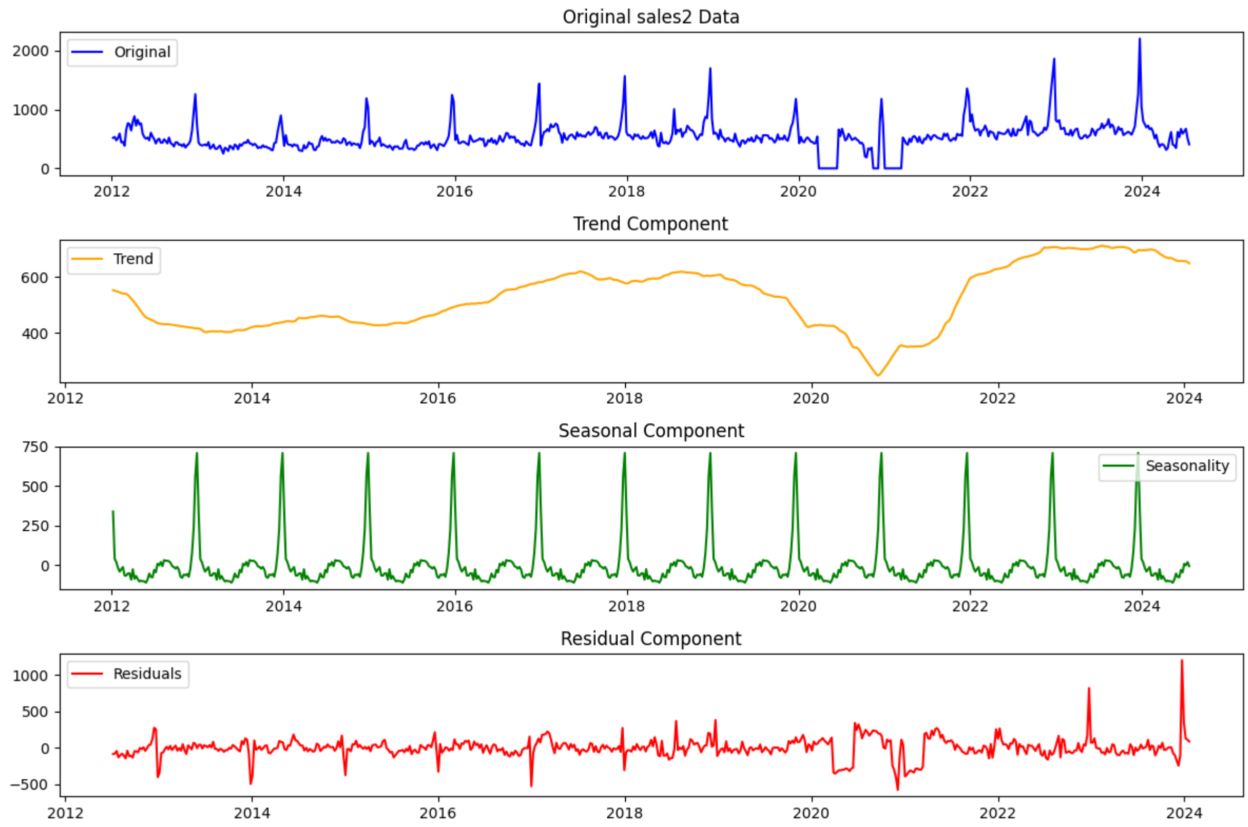

Figure VI. Additive decomposition of *The Alchemist* book sales time series after 2012.

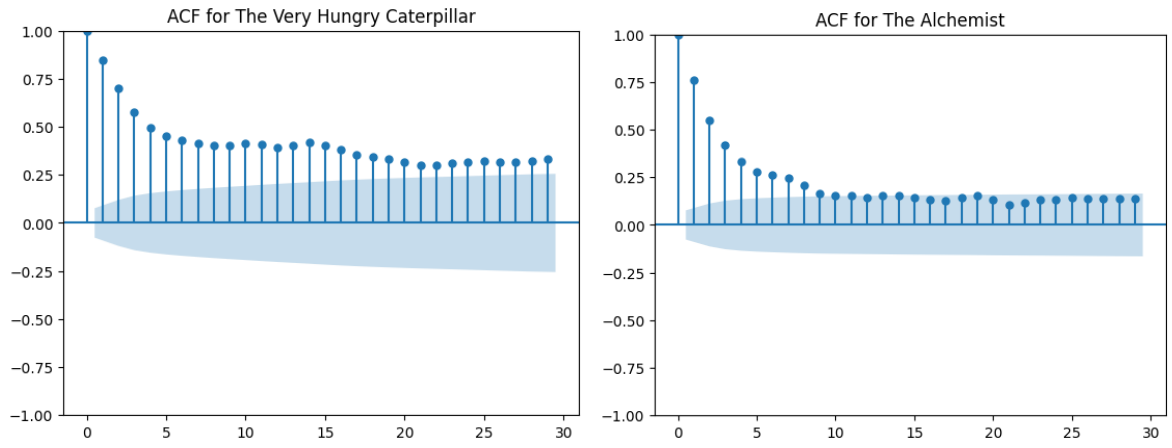

Figure VII. Autocorrelation function (ACF) plots for the time series of *The Very Hungry Caterpillar* and *The Alchemist*

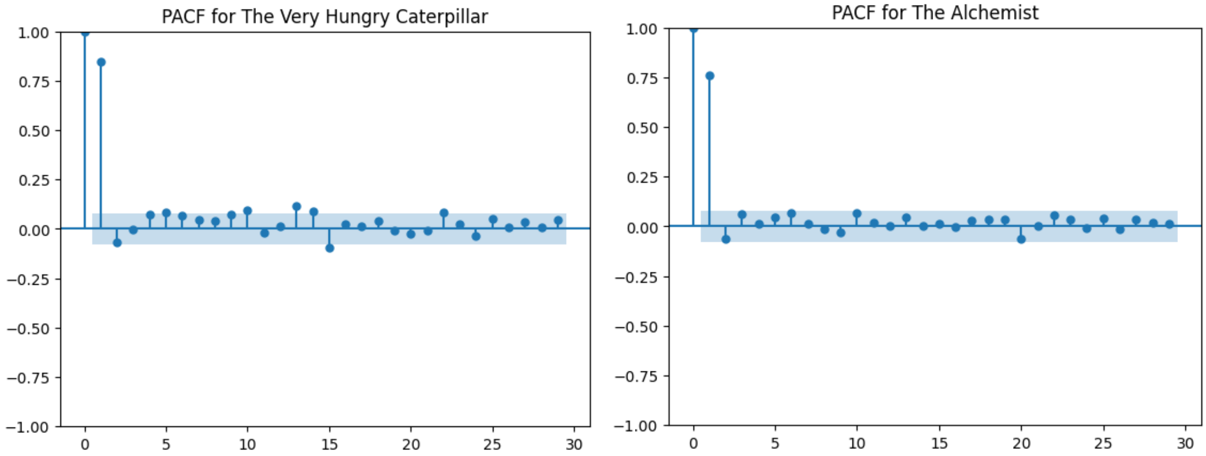

Figure VIII. Partial Autocorrelation function (PACF) plots for the time series of *The Very Hungry Caterpillar* and *The Alchemist*

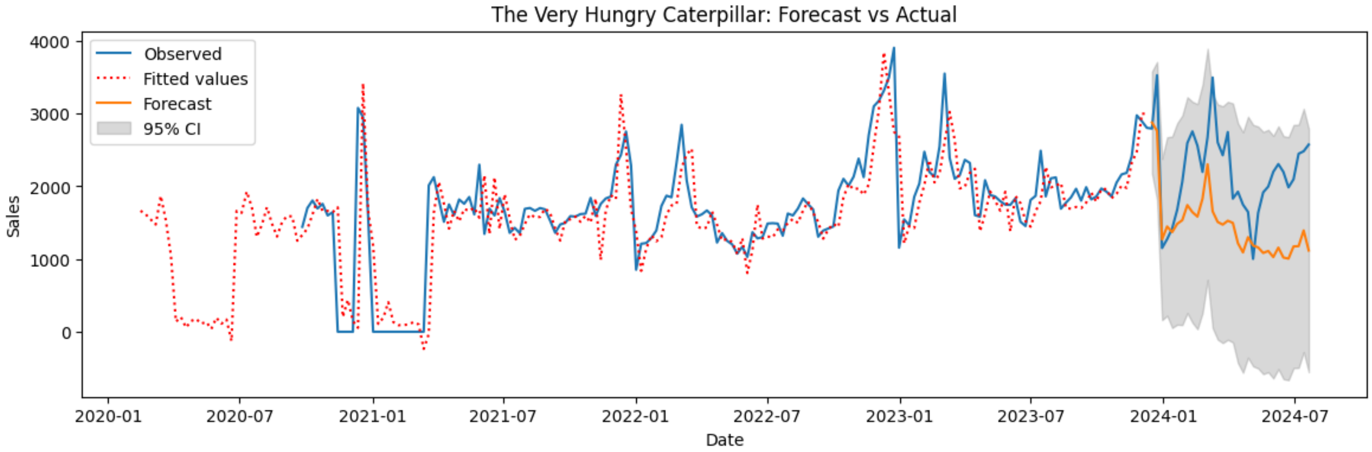

Figure IX. Auto ARIMA fitted values and forecast of *The Very Hungry Caterpillar* sales

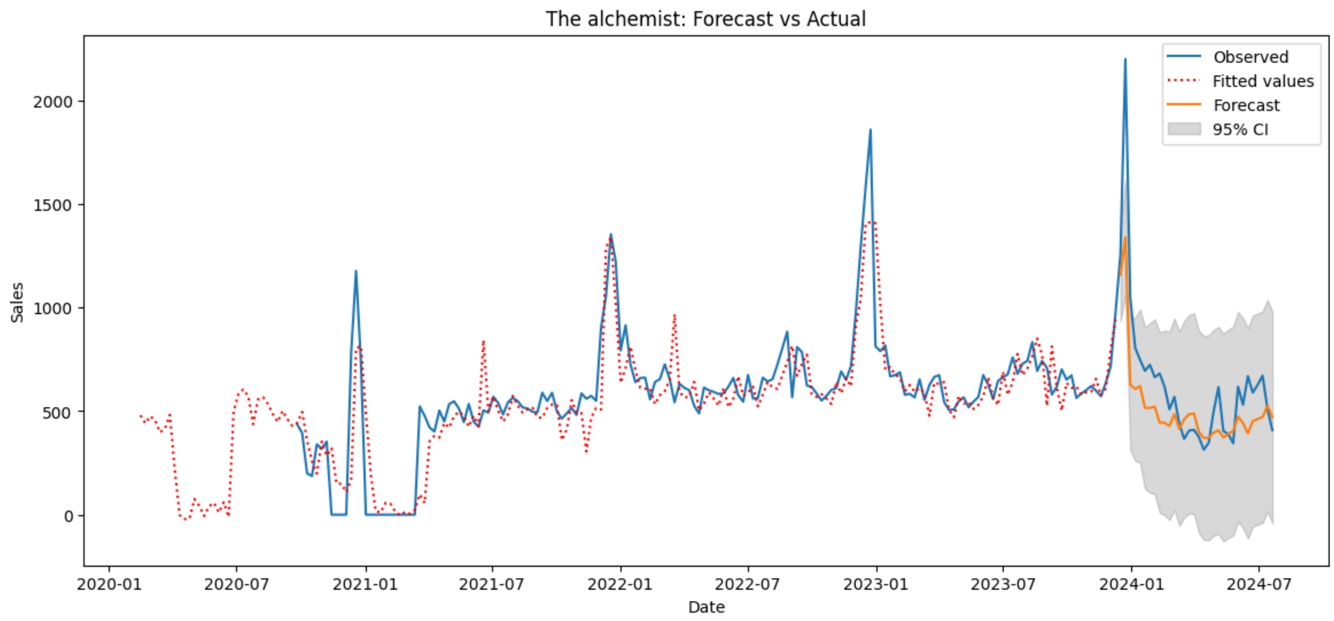

Figure X. Auto ARIMA fitted values and forecast of *The Alchemist* sales

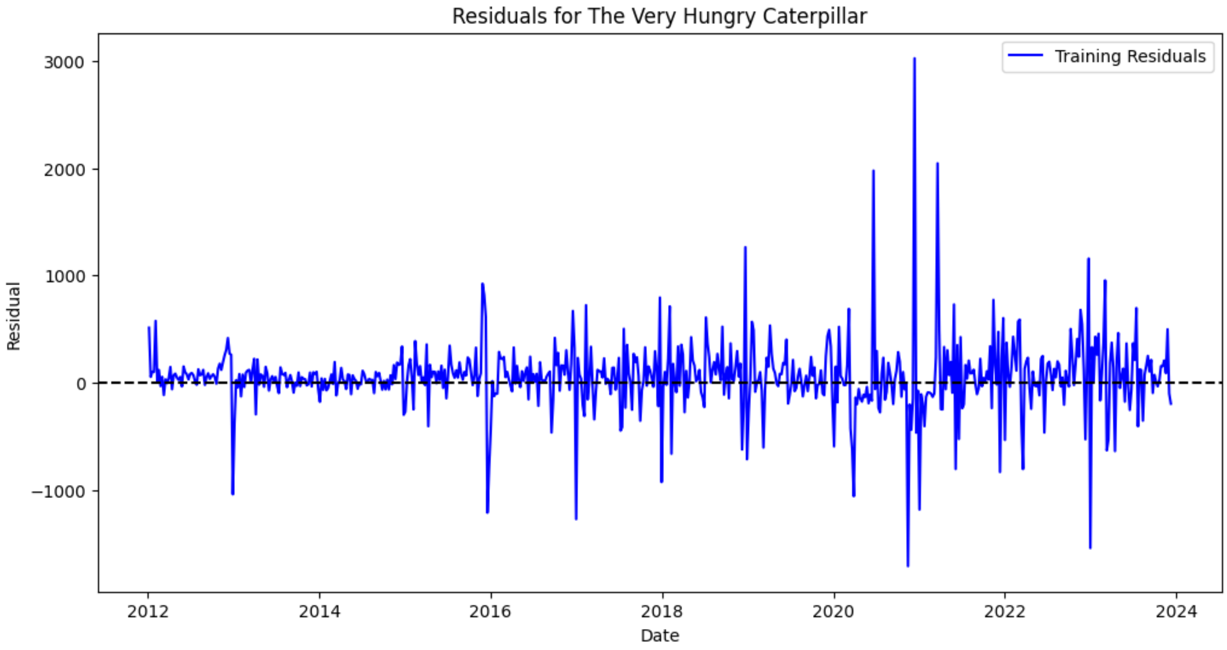

Figure XI. Residuals of the best Arima model for *The Very Hungry Caterpillar*

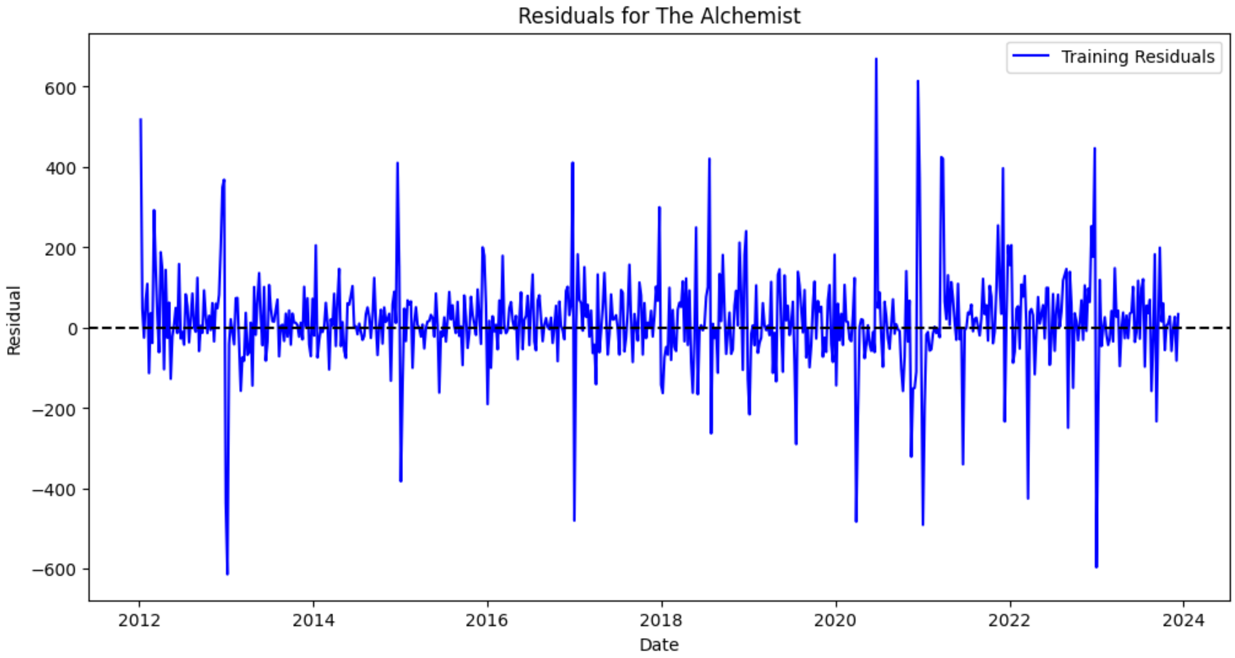

Figure XII. Residuals of the best Arima model for *The Alchemist*

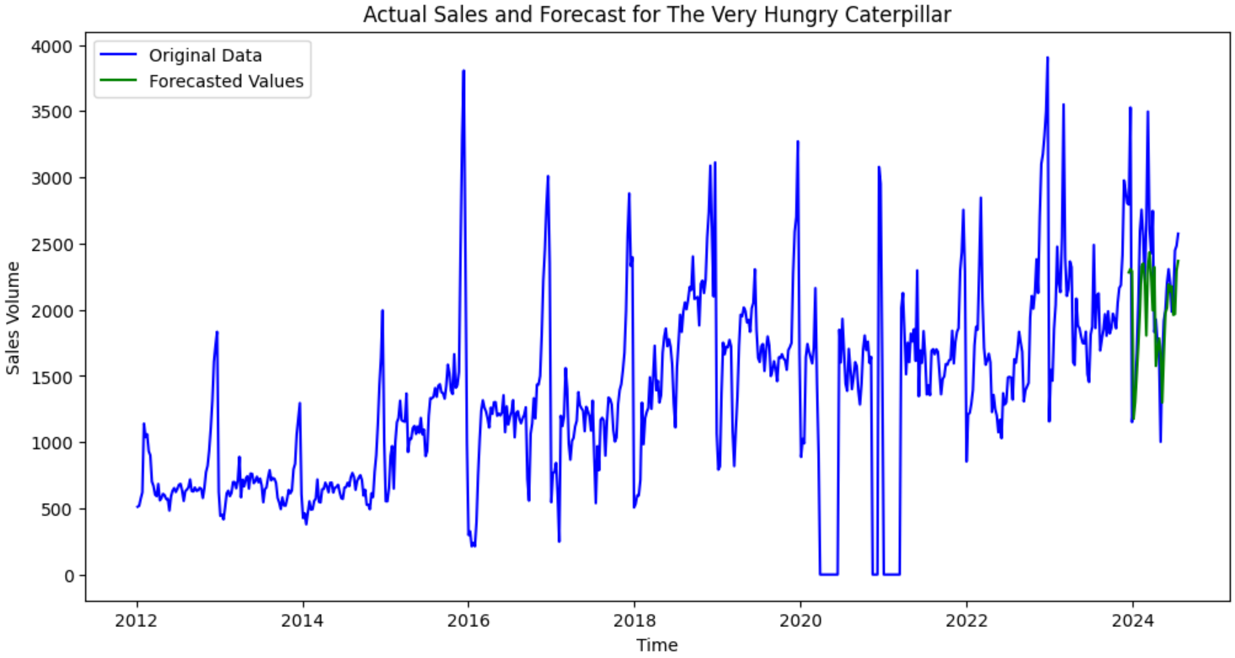

Figure XIII. XGBOOST forecast plot of *The Very Hungry Caterpillar* sales

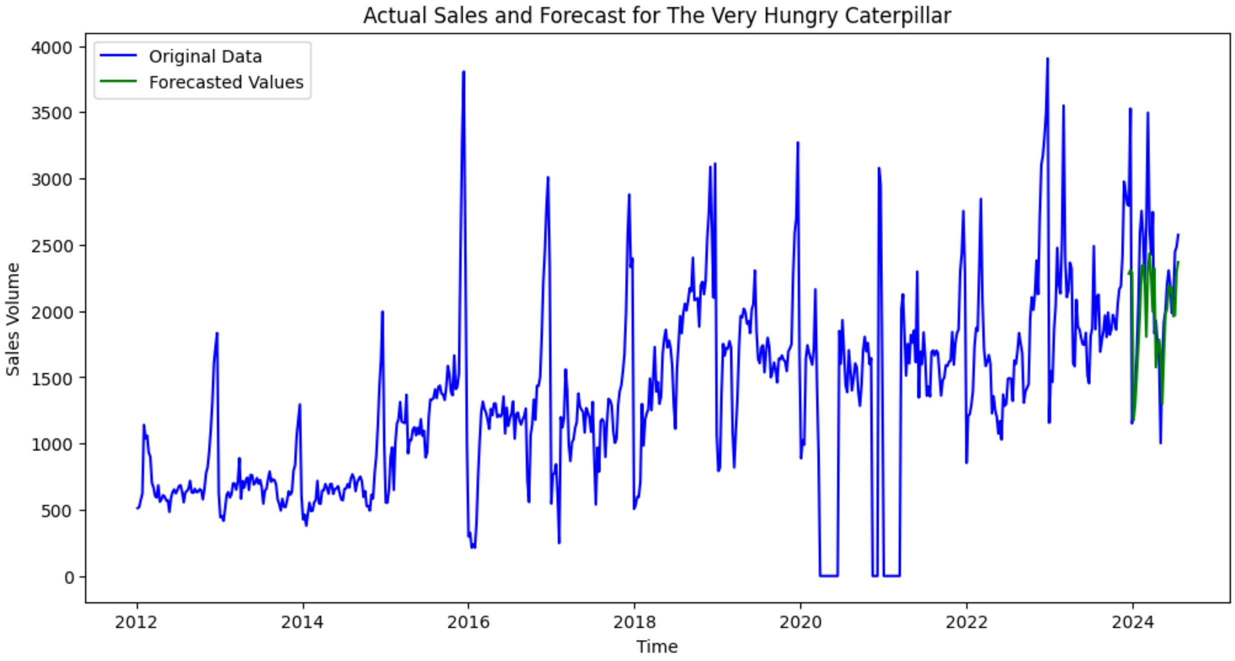

Figure XIV. XGBOOST forecast plot of *The Alchemist* sales

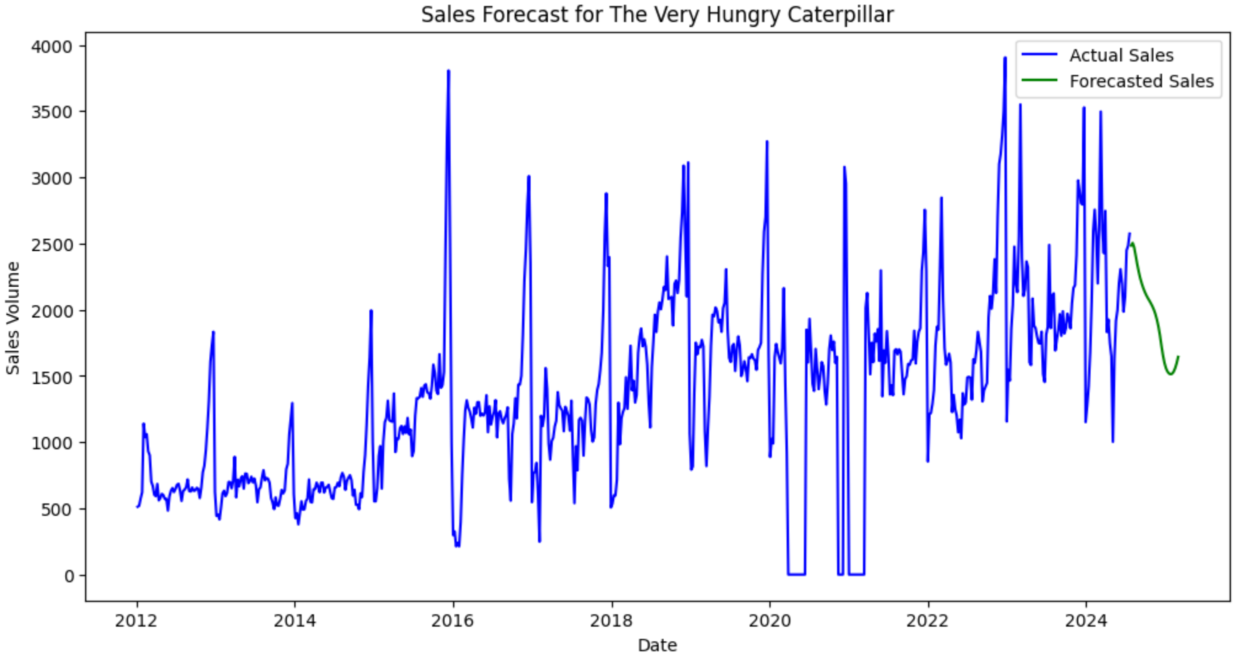

Figure XV. LSTM forecast plot of *The Very Hungry Caterpillar* sales

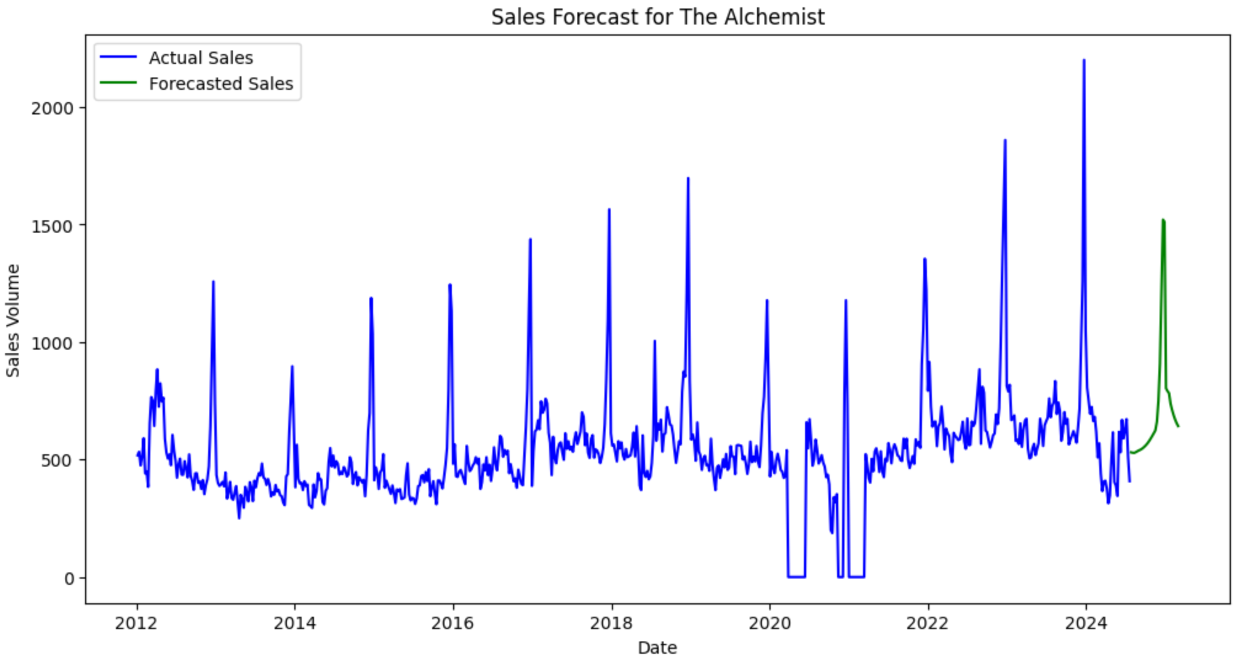

Figure XVI. LSTM forecast plot of *The Alchemist* sales

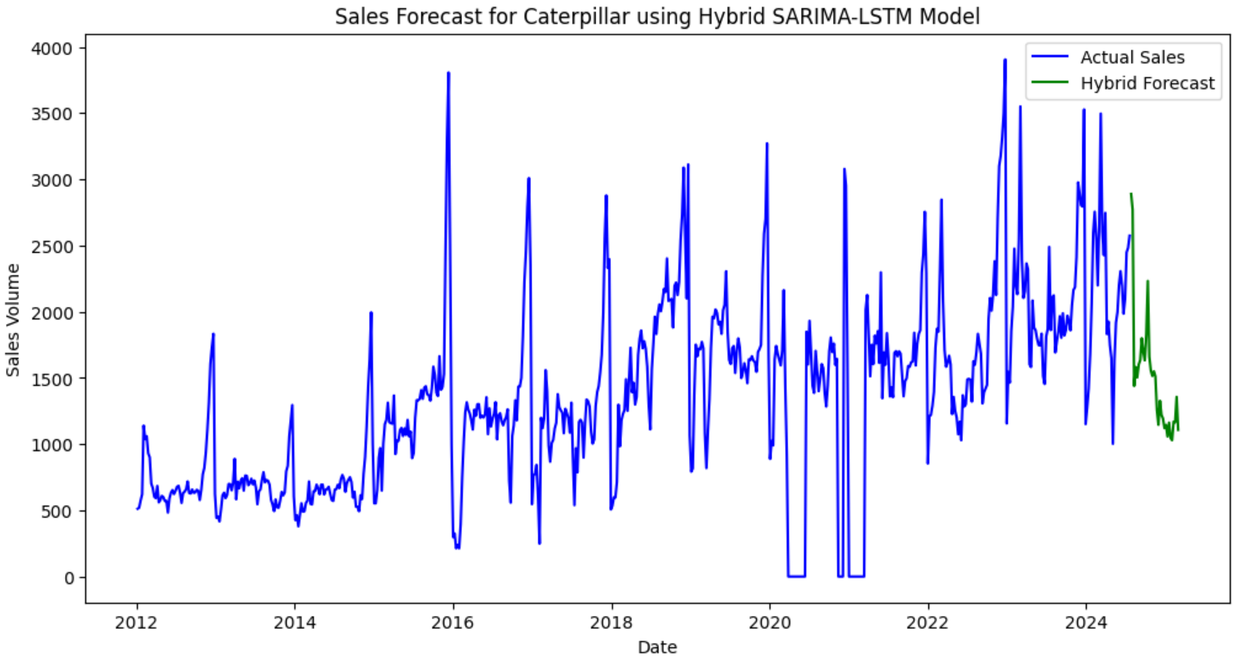

Figure XVII. LSTM and SARIMA hybrid forecast in a sequential combination of *The Very Hungry Caterpillar* sales.

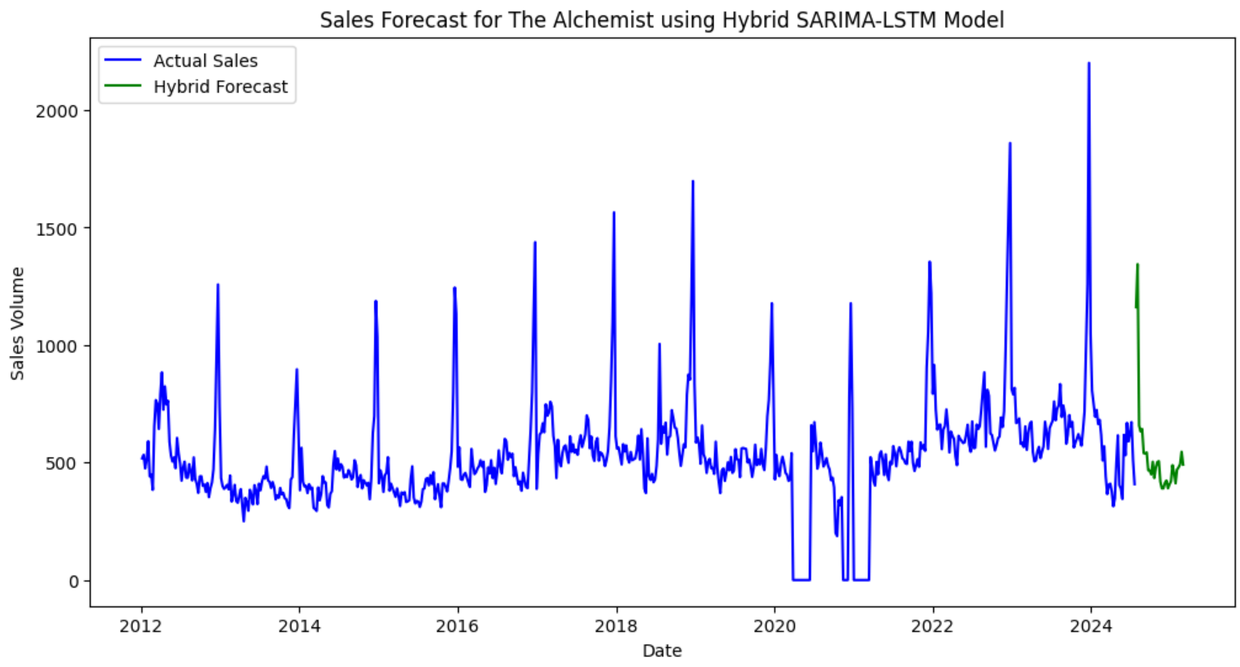

Figure XVIII. LSTM and SARIMA hybrid forecast in a sequential combination of *The Alchemist* sales.

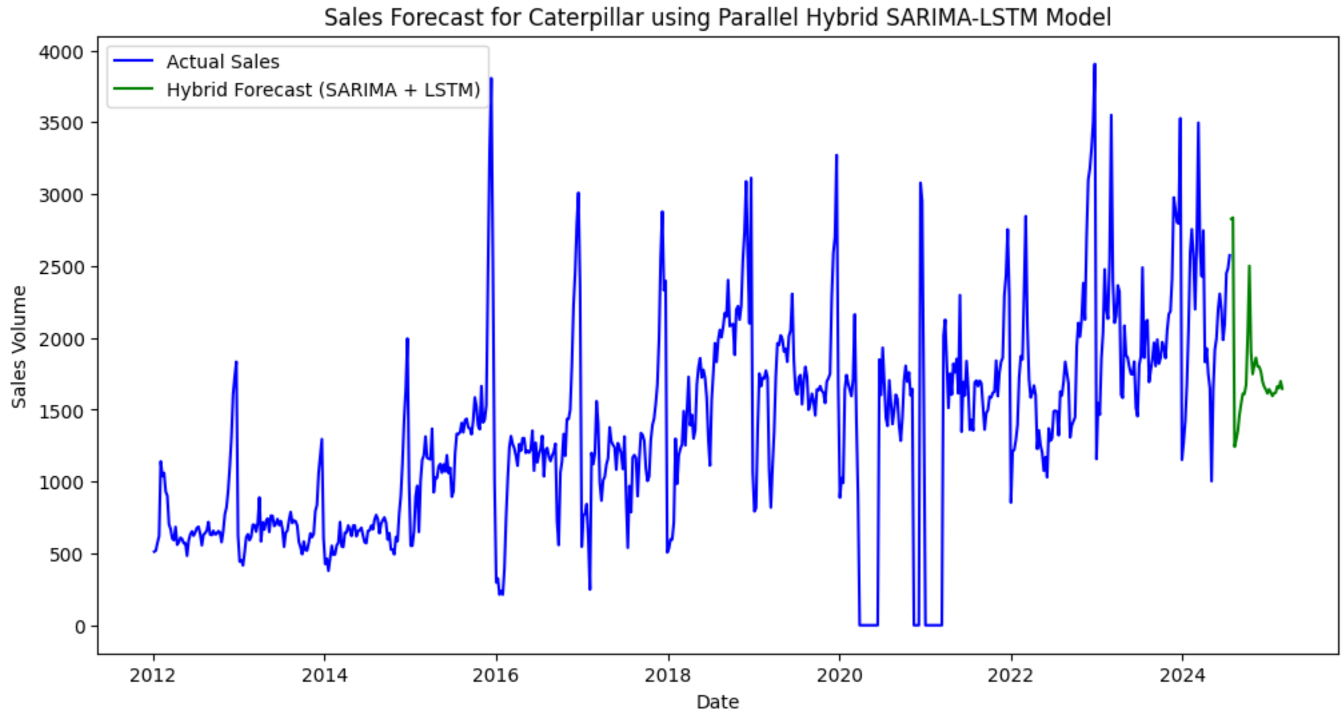

Figure XIX. Best LSTM and SARIMA hybrid forecast in parallel combination of *The Very Hungry Caterpillar* sales (sarima_weight=0.25, lstm_weight=0.75)

.

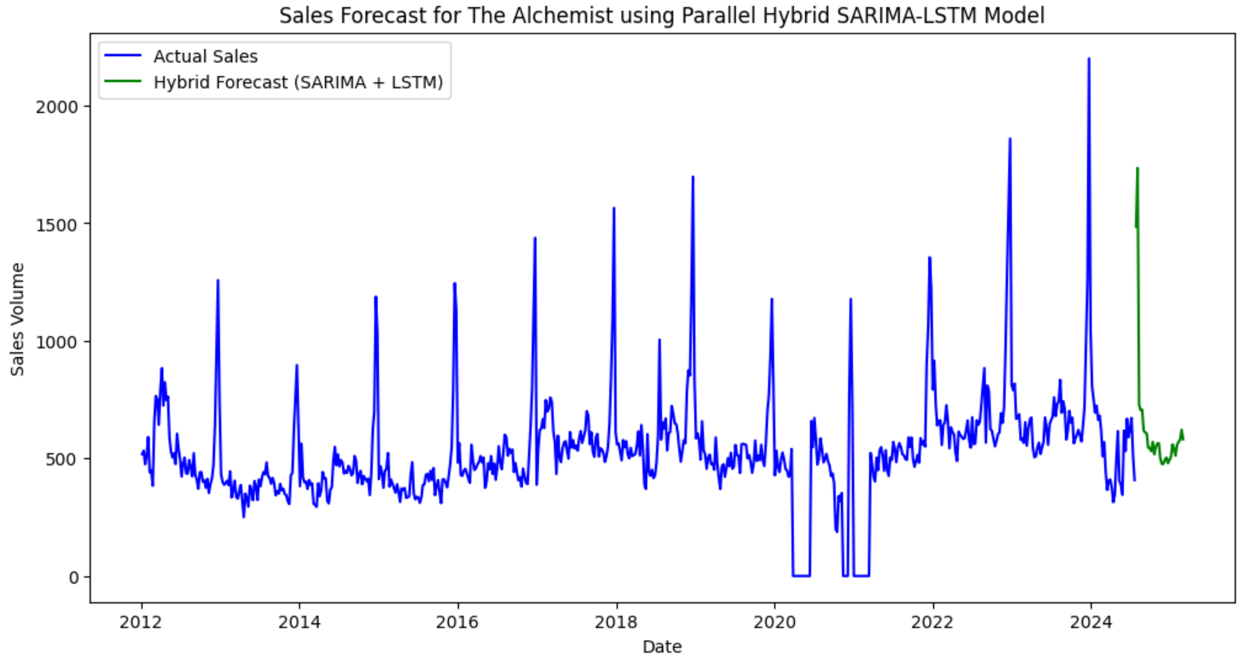

Figure XX. Best LSTM and SARIMA hybrid forecast in parallel combination of *The Alchemist* sales (sarima_weight=0.75, lstm_weight=0.25)

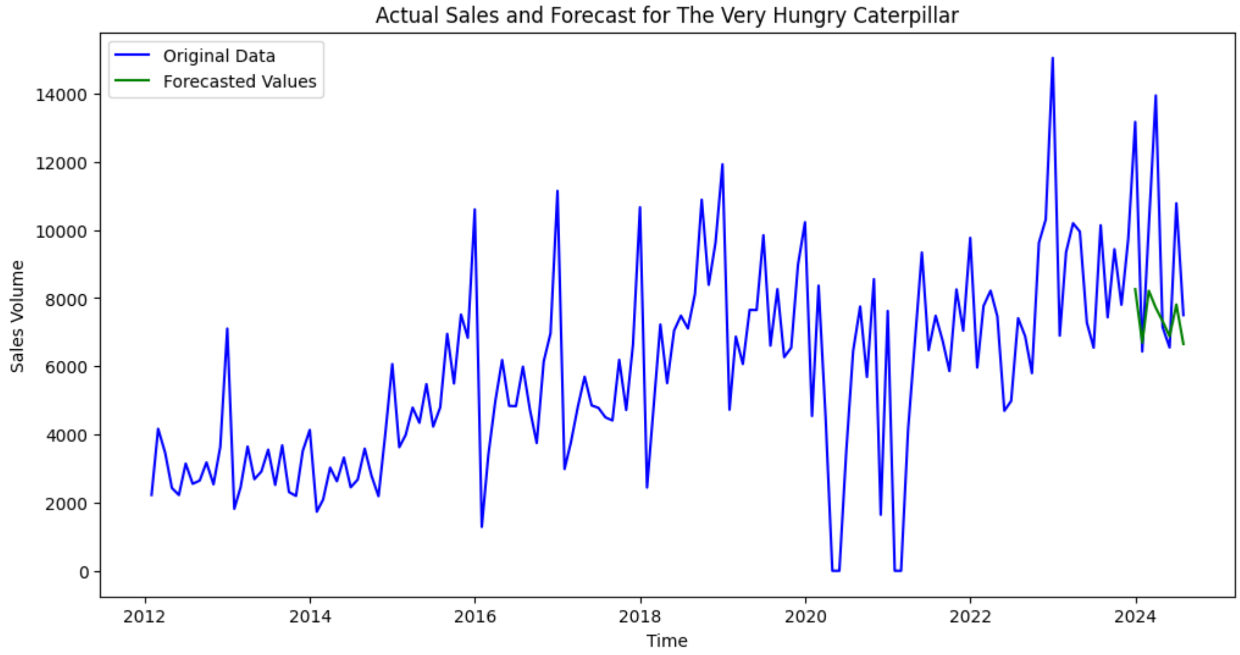

Figure XXI. XGBOOST monthly forecast of *The Very Hungry Caterpillar* sales

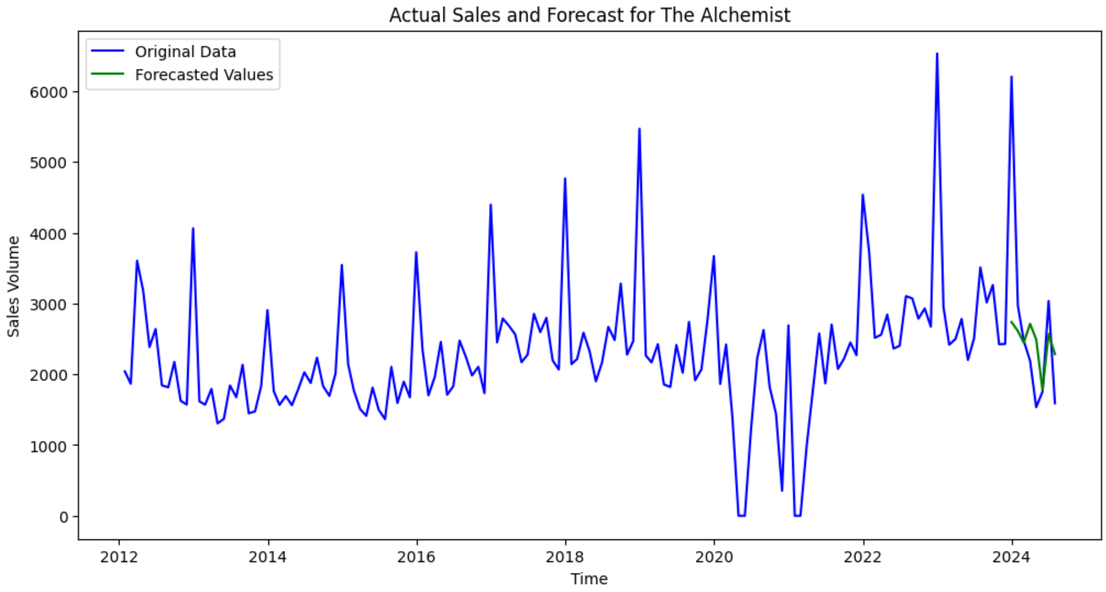

Figure XXII. XGBOOST monthly forecast of *The Alchemist* sales

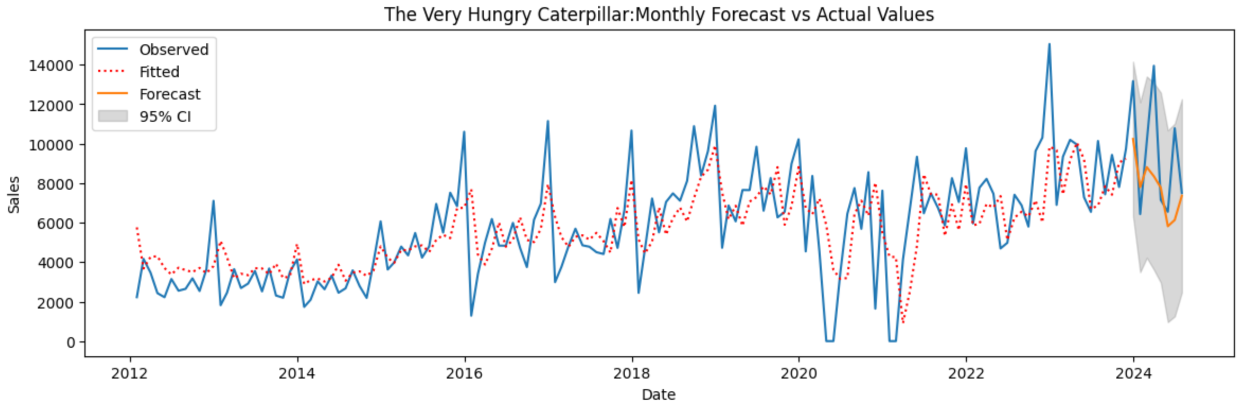

Figure XXIII. Auto ARIMA monthly forecast of *The Very Hungry Caterpillar* sales

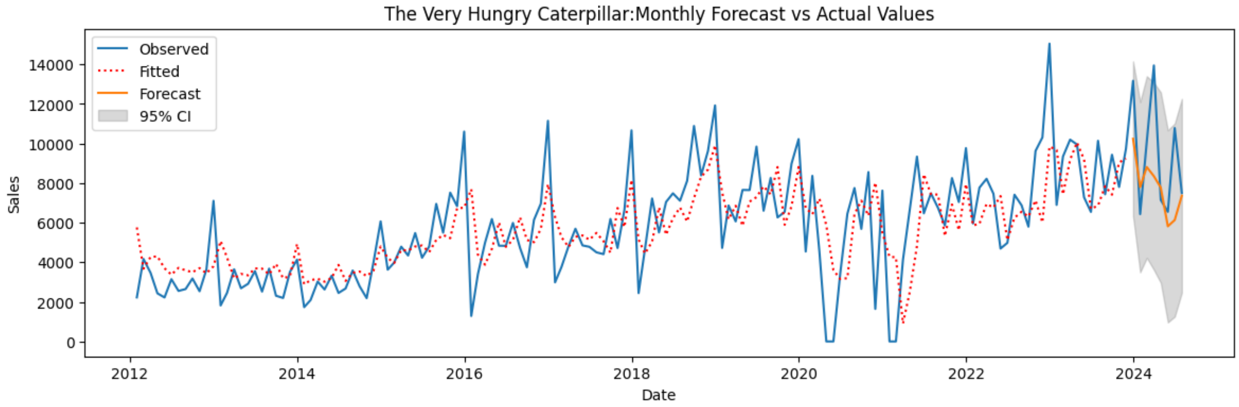

Figure XXIV. Auto ARIMA monthly forecast of *The Alchemist* sales