TITLE: Forecasting Book Sales Using Statistical, Machine Learning, and Hybrid Models

---

NAME: JOE ACHIRA

DATE: 13/05/2025

1. Introduction and Business Context

Accurate demand forecasting is crucial in the publishing industry for managing inventory, planning print runs, and supporting marketing decisions. Book sales can fluctuate due to factors such as seasonality, marketing efforts, holidays, and shifting consumer preferences. Traditional linear models often struggle to capture these dynamics, particularly for titles with non-linear trends or irregular peaks.

This project addresses this forecasting challenge by applying and evaluating various time series models on weekly sales data for two well-known books: *The Very Hungry Caterpillar* and *The Alchemist*. The goal is to identify models that can effectively forecast demand, support business decision-making, and adapt to different sales patterns.

---

2. Problem Statement and Task

The task is to use modelling to forecast the last 32 weeks of unit sales for both books using historical sales data. Specific objectives include:

- Developing and comparing the performance of multiple time series models.
- Identifying the most suitable modeling approach for each title based on performance.
- Recommending strategies for future forecasting and operational planning.

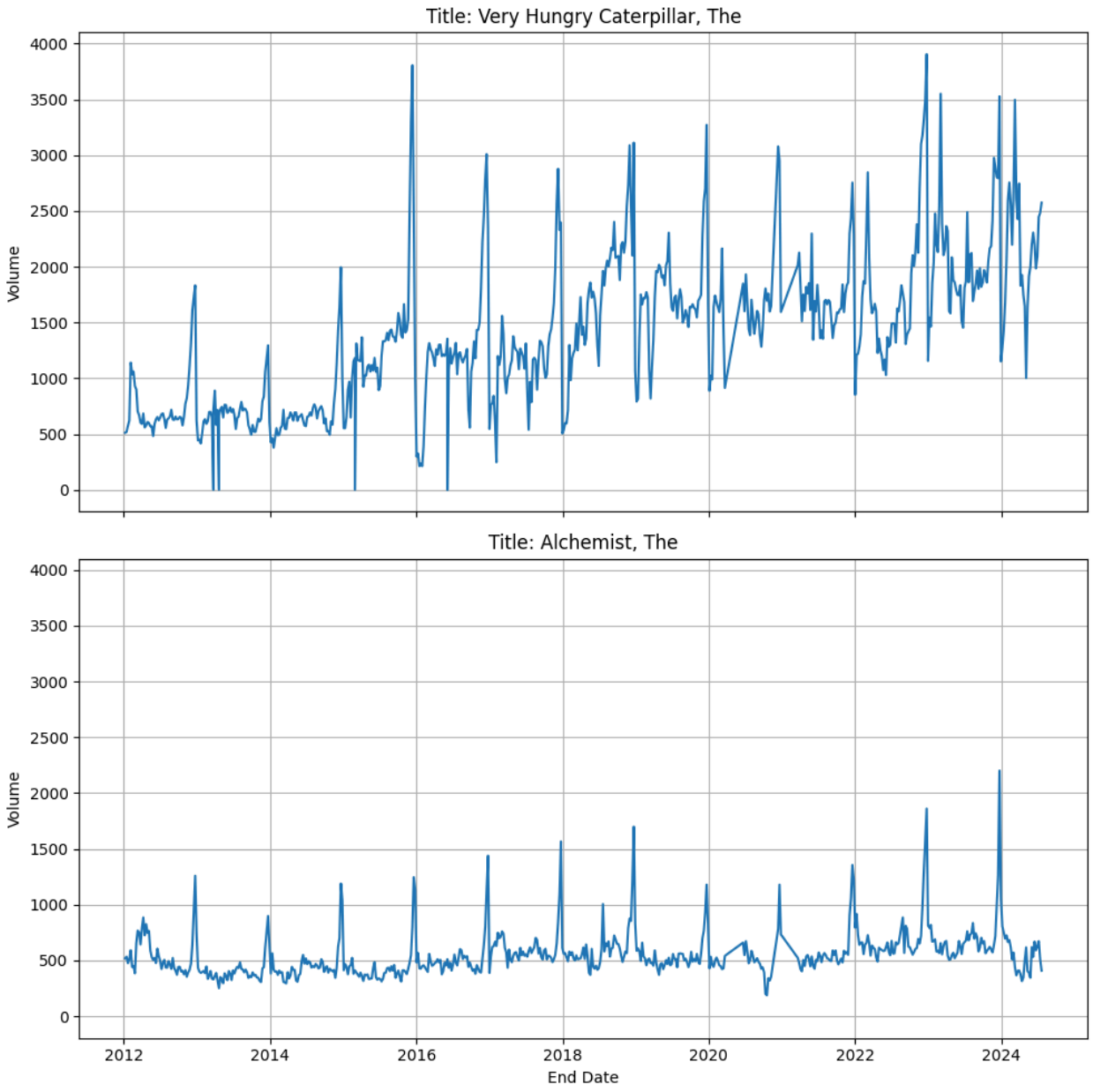
---

3. Dataset Overview

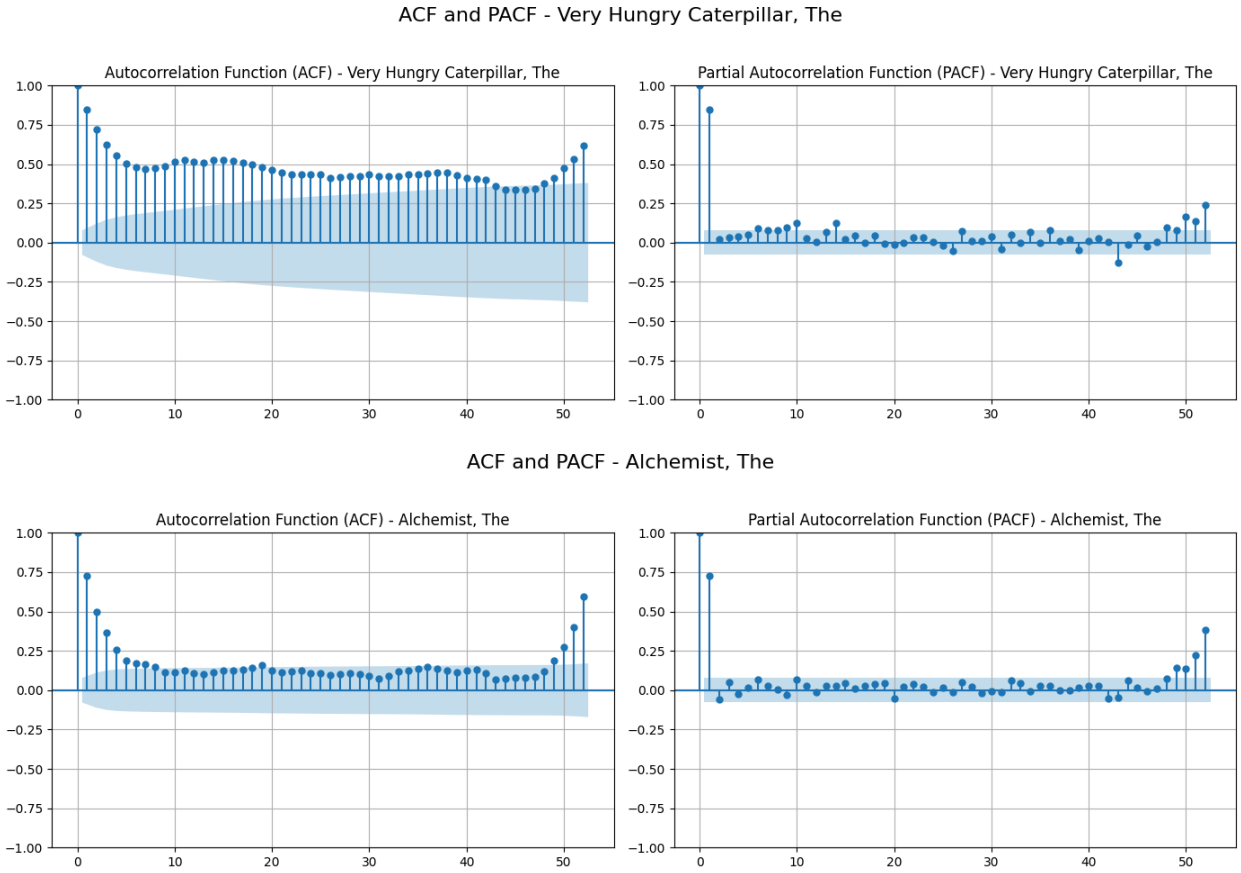
The dataset comprises weekly Nielsen BookScan sales for *The Very Hungry Caterpillar* and *The Alchemist* from **2012 to 2024 (**Fig 1 & 2**)**. Key attributes include:

- **Granularity**: Consistent Weekly sales counts.
- **Timeframe**: 12 years of data (2012–2024).
- **Titles**: Two distinct books with different sales patterns.
- **Preprocessing**: Handled missing weeks and smoothed out known anomalies.

We used data from 2012 to just before the 32-week forecast horizon for training, and held out the forecast horizon weeks as the test set for evaluation and performance comparison.


Title: Very Hungry Caterpillar, The
Volume
0
500
1000
1500
2000
2500
3000
3500
4000
Title: Alchemist, The
Volume
0
500
1000
1500
2000
2500
3000
3500
4000
2012
2014
2016
2018
2020
2022
2024
End Date

(Figs 1 & 2)



(Figs 3, 4, 5, 6)

The Autocorrelation Function (ACF) showed a slow decline (non-stationarity) compared to that of The Alchemist which showed a faster decline to near zero (stationarity).

Both Partial ACFs (PACF) showed a sharp cut off after the first lag which was indicative of AR model.

---

4. Modeling and Results

Each model was applied to both books independently. Below, we outline the modeling approach, performance, and observations for each.

4.1 ARIMA and SARIMA

Methodology

We applied the SARIMA model to capture both linear trends and seasonality. Grid search was used to optimise `(p, d, q)(P, D, Q, s)` parameters based on AIC values. Time series

decomposition confirmed strong weekly seasonality for *The Very Hungry Caterpillar*. For *The Alchemist*, seasonal effects were less pronounced.

Results and Performance

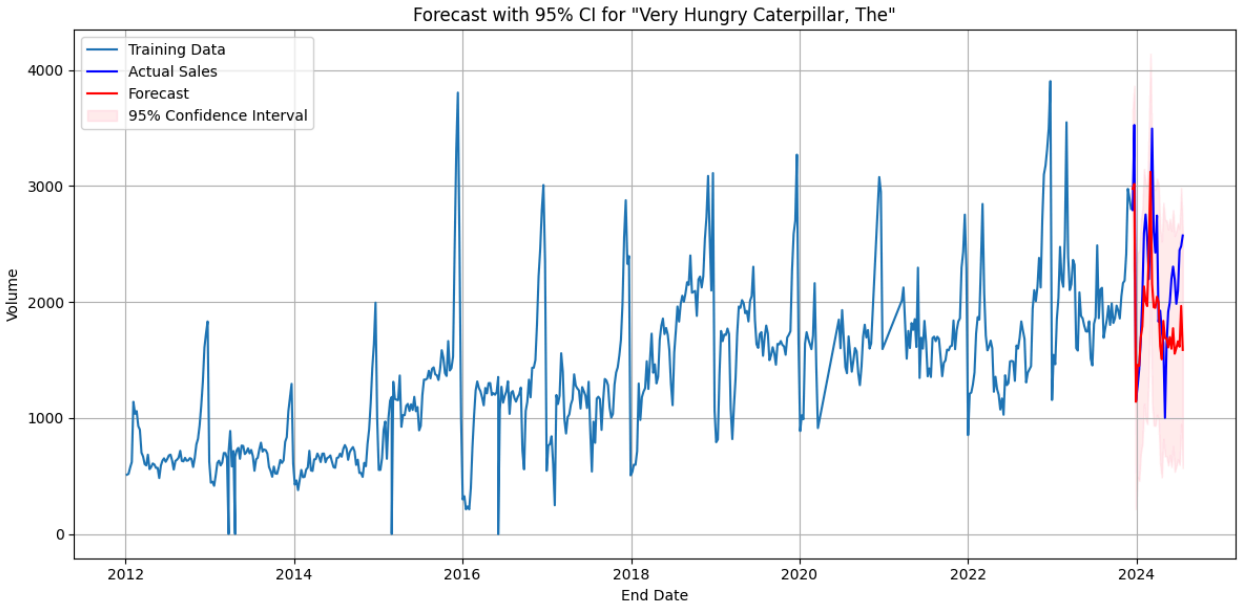
- For *The Very Hungry Caterpillar*, SARIMA captured regular seasonal peaks but underperformed during non-seasonal fluctuations (Fig 7).

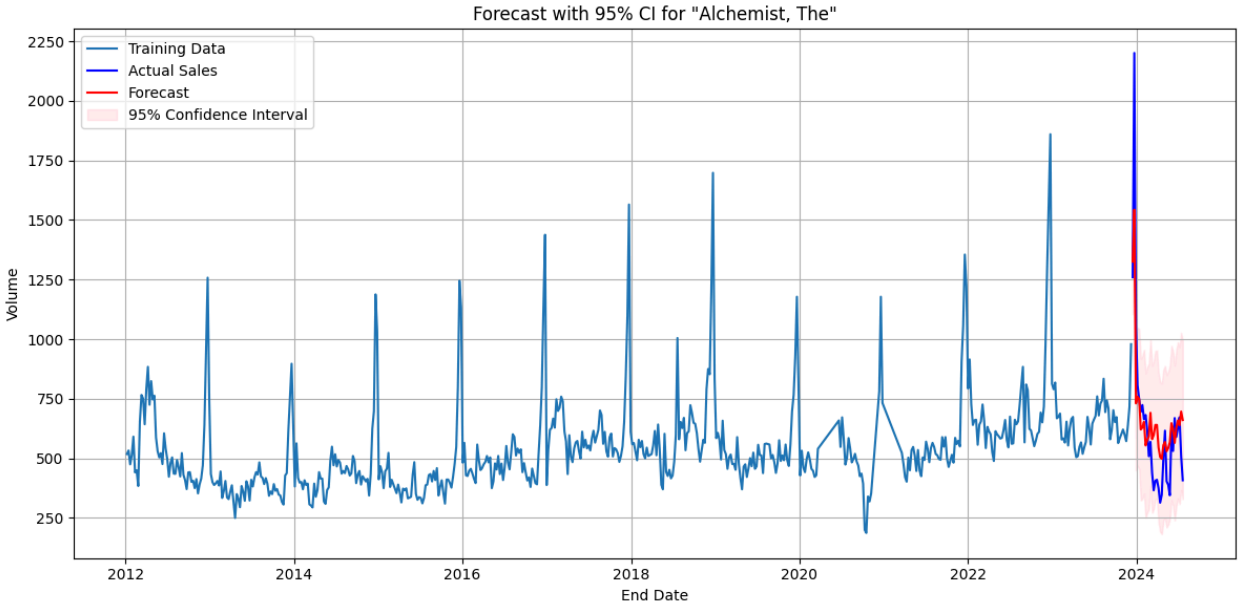
- For *The Alchemist*, SARIMA produced a reasonable trend forecast, though it struggled with minor fluctuations and shifts (Fig 8).

Error Metrics:

- Caterpillar: RMSE = 528

- Alchemist: RMSE = 183





Figs 7 & 8

4.2 XGBoost

Methodology

The time series was reframed as a supervised learning problem using lag features (e.g. previous 4-8 weeks of sales). Key hyperparameters like `n_estimators`, `max_depth`, and `learning_rate` were tuned using `TimeSeriesSplit` cross-validation. This approach allowed the model to learn patterns from past windows.

Results and Performance

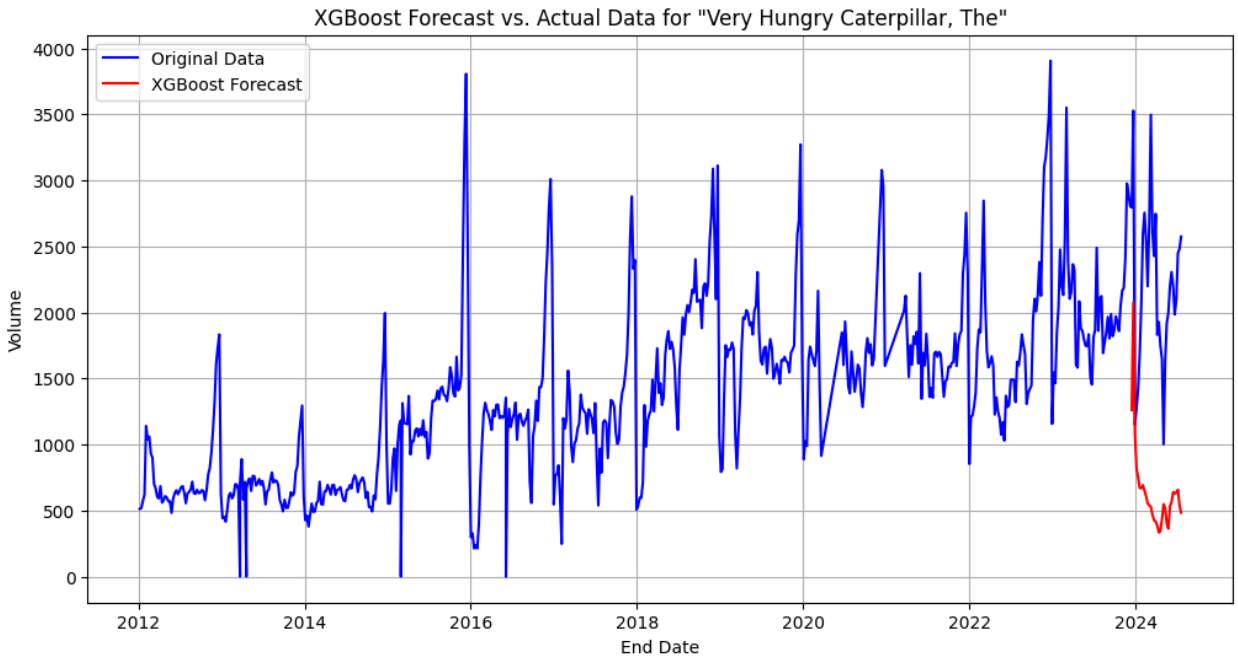
- For *The Very Hungry Caterpillar*, XGBoost struggled with sharp seasonal peaks unless those patterns were explicitly encoded (Fig 9).

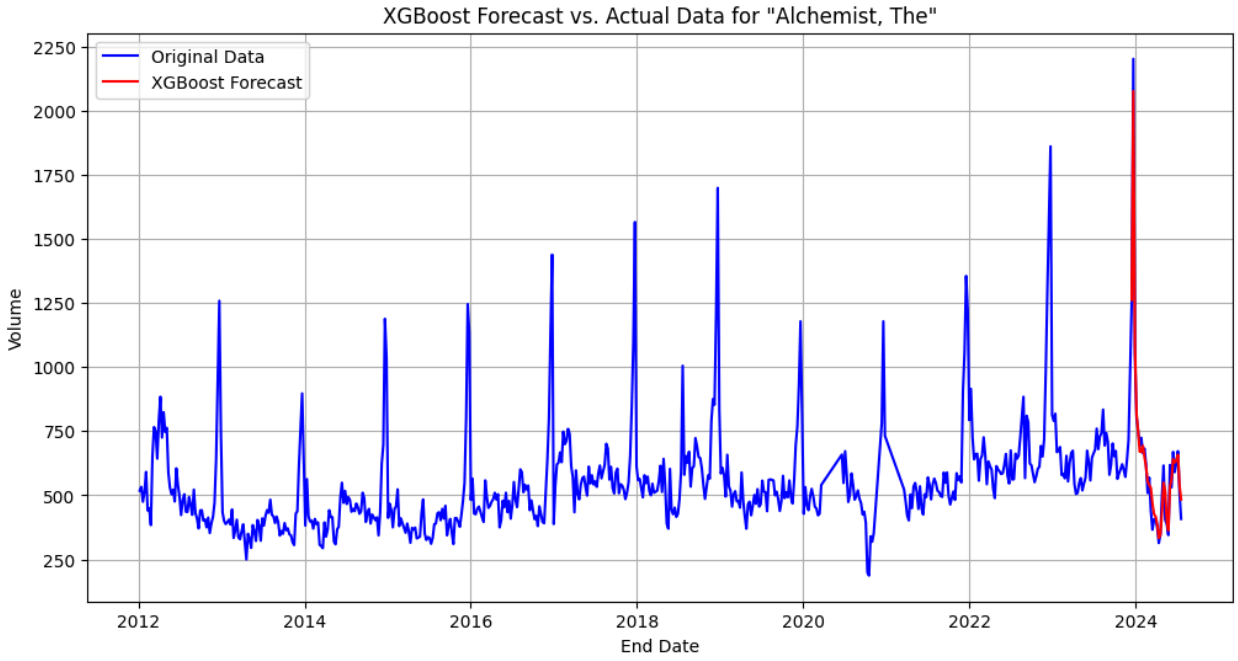
- For *The Alchemist*, XGBoost performed relatively well during stable periods but underpredicted sudden spikes in sales (Fig 10).

Error Metrics:

- Caterpillar: RMSE = 436, MAE = 353, MAPE = 17%

- Alchemist: RMSE = 144, MAE = 103, MAPE = 18%





Figs 9 & 10

4.3 LSTM Neural Networks

Methodology

We built LSTM models using TensorFlow/Keras, with normalised inputs and sequence lengths

of 8-12 weeks. Keras Tuner was used for hyperparameter optimisation (units, dropout, learning rate). Early stopping was applied to reduce overfitting.

Results and Performance

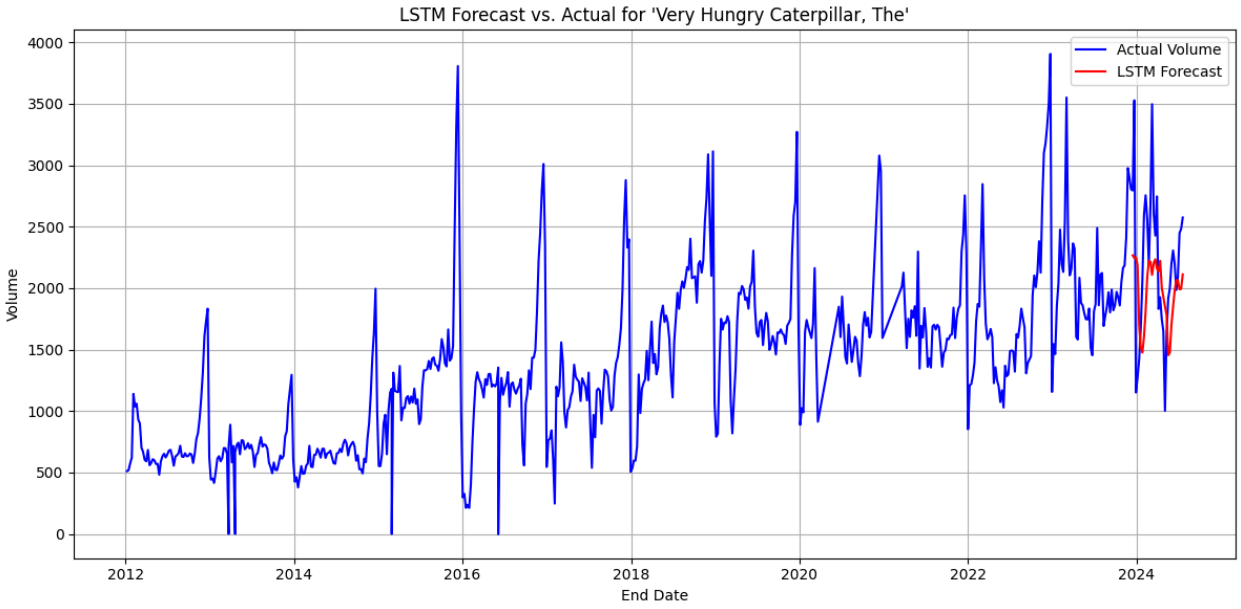
- For *The Very Hungry Caterpillar*, LSTM captured both short-term patterns and non-linear variations, handling fluctuations better than SARIMA (Fig. 11).

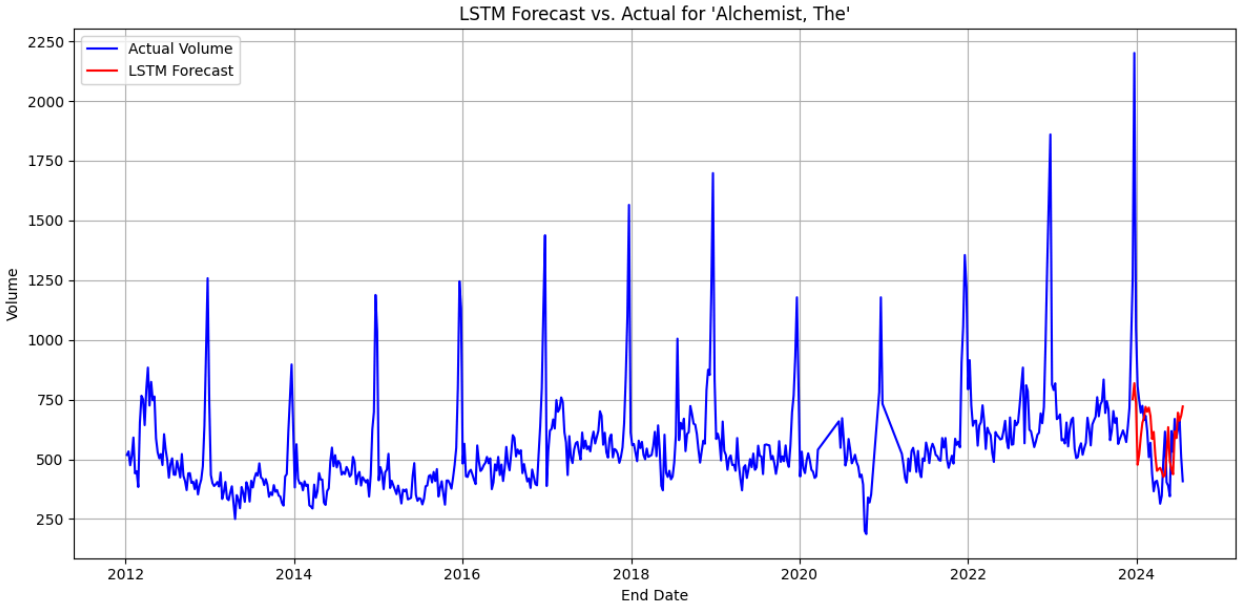
- For *The Alchemist*, LSTM slightly improved over traditional models, especially during local peaks (Fig. 12).

Error Metrics:

- Caterpillar: RMSE = 581, MAE = 480, MAPE = 25%

- Alchemist: RMSE = 316, MAE = 181, MAPE = 28%





Figs 11 & 12

4.4 Hybrid SARIMA-LSTM

Methodology

To combine the strengths of linear (SARIMA) and non-linear (LSTM) models, we implemented a **parallel hybrid approach**. SARIMA and LSTM were trained separately, and their outputs were combined via weighted averaging. The weightings were optimised on a validation window to minimise forecast error.

Results and Performance

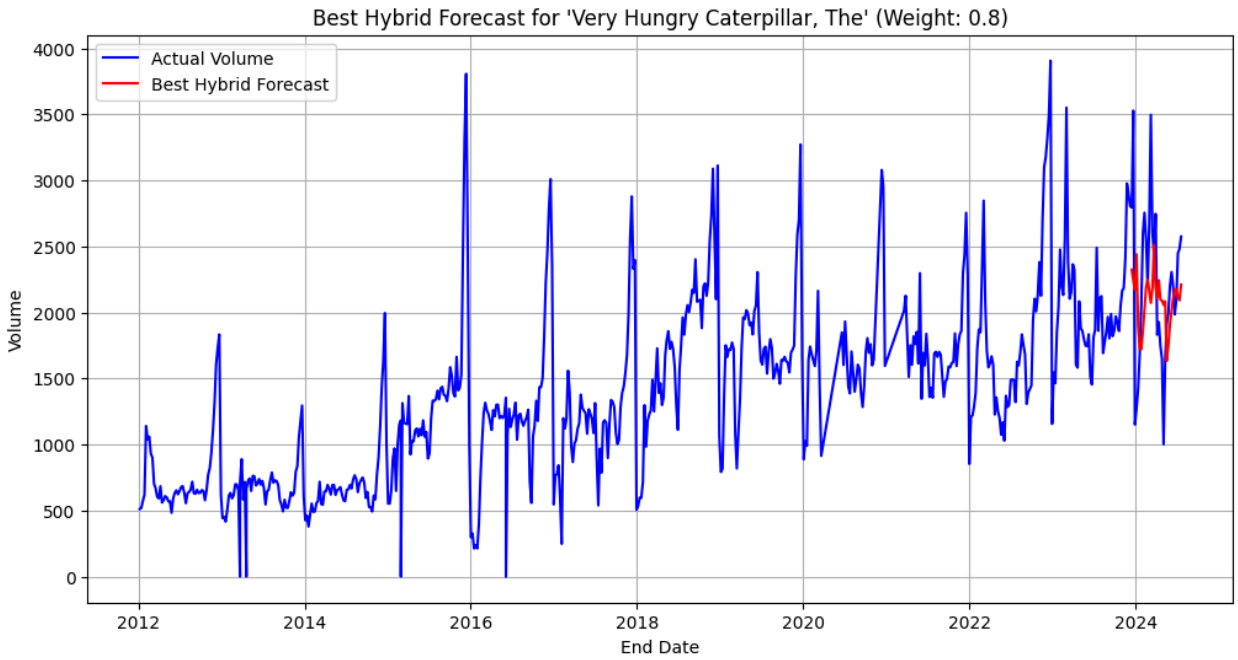
- For *The Very Hungry Caterpillar*, the hybrid model produced the most accurate forecasts, effectively blending seasonal accuracy and non-linear responsiveness (Fig. 13).

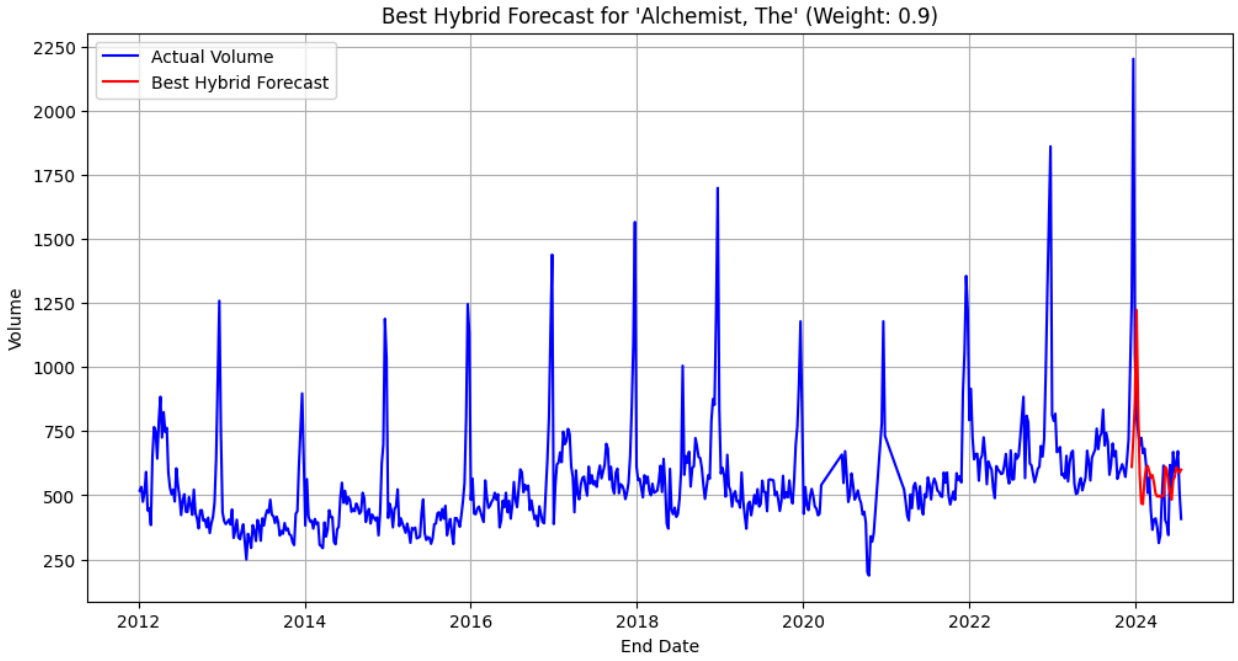
- For *The Alchemist*, improvements over standalone LSTM were marginal, due to less seasonal variation (Fig. 14).

Error Metrics:

- Caterpillar: RMSE = 592, MAE = 474, MAPE = 24%

- Alchemist: RMSE = 324, MAE = 190, MAPE = 29%





Figs 13 & 14

---

5. Final Thoughts and Recommendations

Summary of Findings

- **SARIMA** is effective for seasonality but limited in adapting to unpredictable patterns.
- **XGBoost** performs best in stable, trend-based sales patterns but requires extensive feature engineering.
- **LSTM** captures non-linearities and performs well across both titles, especially for variable demand.
- **Hybrid SARIMA-LSTM** delivers the best overall results, especially for titles with both strong seasonality and irregular demand spikes.

Recommendations

- Use hybrid models for children's or seasonal books with complex patterns.
- Consider LSTM-only models for books with smoother or more stable sales.
- Implement automated pipelines for regular retraining as new weekly sales data becomes available.

---

6. Future Work

To enhance forecast precision and business utility, future efforts should:

- Integrate external features such as promotions, holidays, or price changes (e.g. via SARIMAX or multivariate LSTM).
- Explore more advanced architectures.
- Extend the hybrid framework to multi-book or category-level forecasting.