Using Time Series Analysis for Sales and Demand Forecasting

UNIVERSITY OF
CAMBRIDGE

Anastasia Alexandra Krivenkovskaya

Institute of Continuing Education

University of Cambridge

This academic report is submitted for the course

Applying advanced data science techniques P1_2025

CAM_DS_301

September 2025

Table of Contents

Introduction and Business Context

This report presents the application of multiple time series forecasting techniques to sales data provided by "Industry-standard retail book sale", the leading service for book sales tracking across major markets. The analysis focuses on two internationally recognized titles — *The Alchemist* and *The Very Hungry Caterpillar*.

The objective of this project is to evaluate how traditional statistical methods (ARIMA/SARIMA), machine learning approaches (XGBoost), deep learning models (LSTM), and hybrid techniques (SARIMA+LSTM) perform in predicting book sales. The goal is to generate insights that can guide independent publishers in managing stock levels, planning reprints, and identifying titles with strong long-term potential.

Weekly sales data was available from 2012 onwards, and the forecast horizon was set at **32 weeks**. In addition, monthly aggregation was applied to compare model performance at different temporal granularities.

Data Preparation and Exploration

- Weekly sales data was resampled to ensure continuity; missing weeks were filled with zeros, following the requirement that “weeks with no sales should still be represented.”
- Both books exhibited strong **seasonal trends**, with peaks around holidays, and relatively stable long-term patterns.

- The dataset was split into a **training set (2012 until 32 weeks before the end)** and a **test horizon (final 32 weeks)**, which served as the evaluation period across all models.

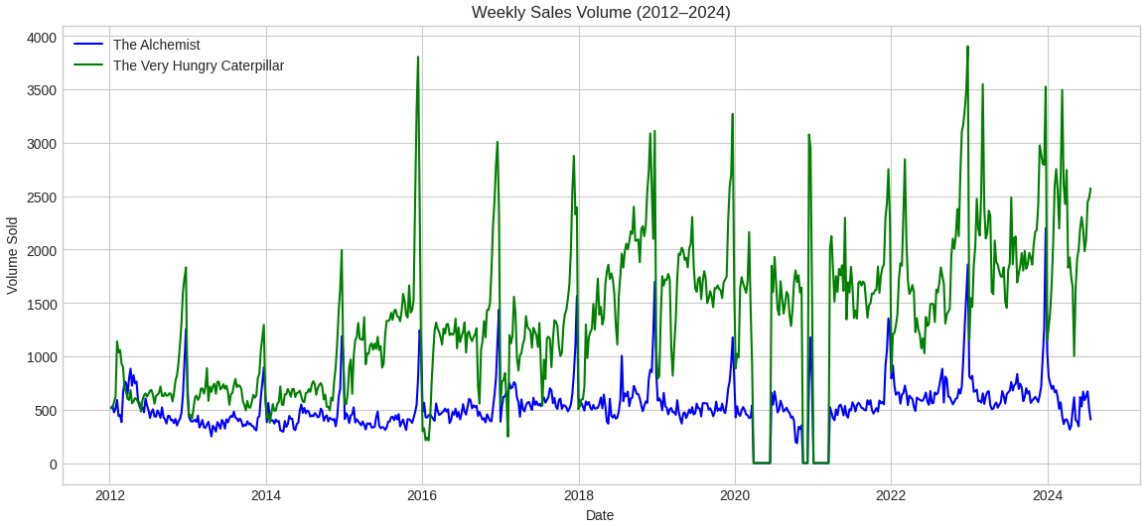


Figure 1: ***Weekly sales patterns for both books***

Methodology

A structured modelling pipeline was applied across multiple approaches:

Classical Techniques: SARIMA

- **ADF Tests** confirmed both series were stationary after differencing.
- **Seasonal decomposition** indicated strong yearly seasonality.

Auto ARIMA was used to select the best SARIMA models:

- **The Alchemist** → ARIMA(0,1,2)(1,0,1)[52], AIC=7641.2

- **The Very Hungry Caterpillar** → ARIMA(1,1,2)(1,0,1)[52], AIC=8996.1

Forecasts for the final 32 weeks showed that SARIMA captured **seasonality well**, but sometimes under- or over-estimated sharp sales fluctuations.

Evaluation (32-week horizon):

- *The Alchemist*: MAE ≈ 139, MAPE ≈ 0.23
- *Caterpillar*: MAE ≈ 396, MAPE ≈ 0.19

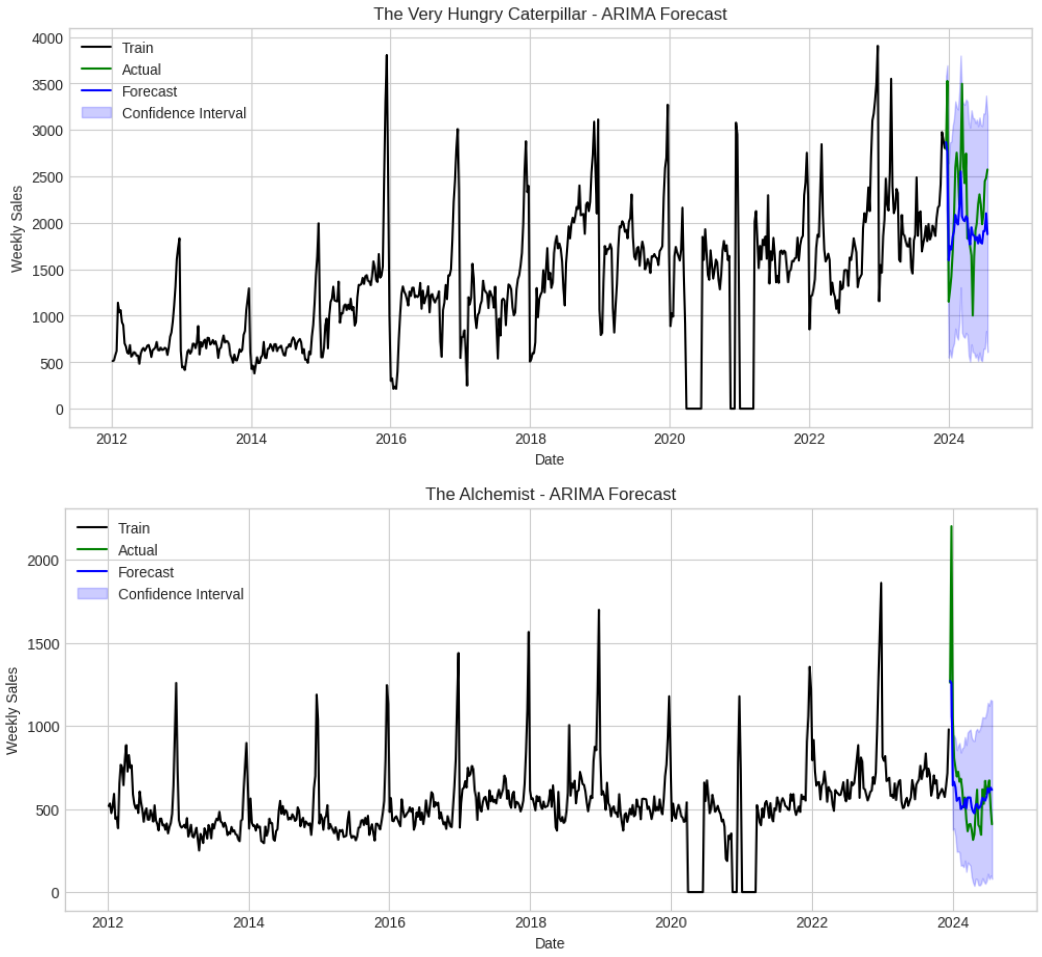


Figure 2: ***SARIMA forecast with confidence intervals***

Key insight: SARIMA is stable and interpretable, but struggles with irregular peaks caused by promotions or external media influence.

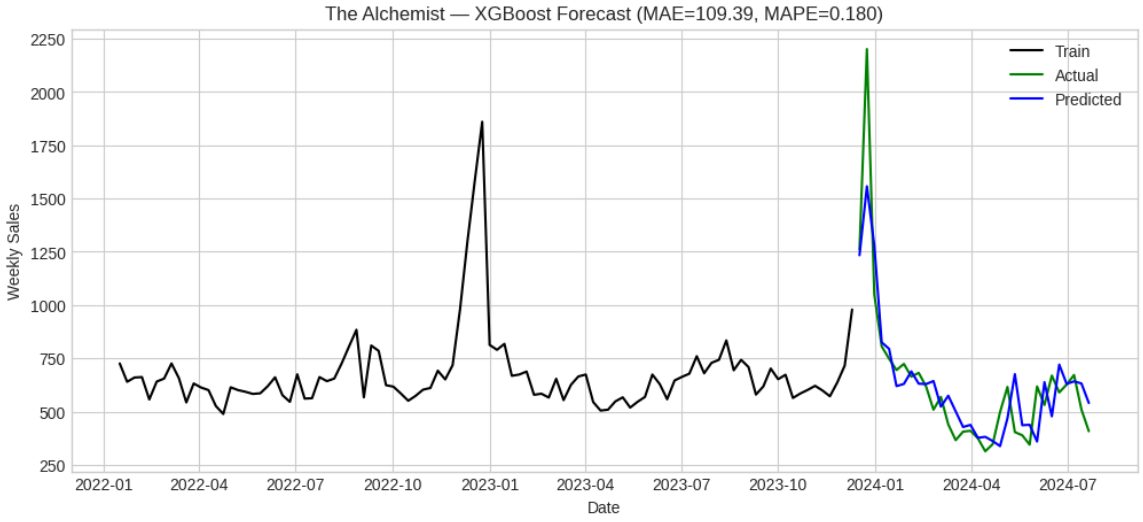
Machine Learning Models: XGBoost

Supervised features were created using **lag windows** (past n weeks used to predict the next). A search over window lengths (8, 12, 16) showed:

- Best window for *The Alchemist* = 8 weeks
- Best window for *The Very Hungry Caterpillar* = 16 weeks

Hyperparameter tuning (learning rate, depth, estimators, subsample) produced:

- **The Alchemist** → MAE = 109.4, MAPE = 0.18
- **Caterpillar** → MAE = 316.4, MAPE = 0.16



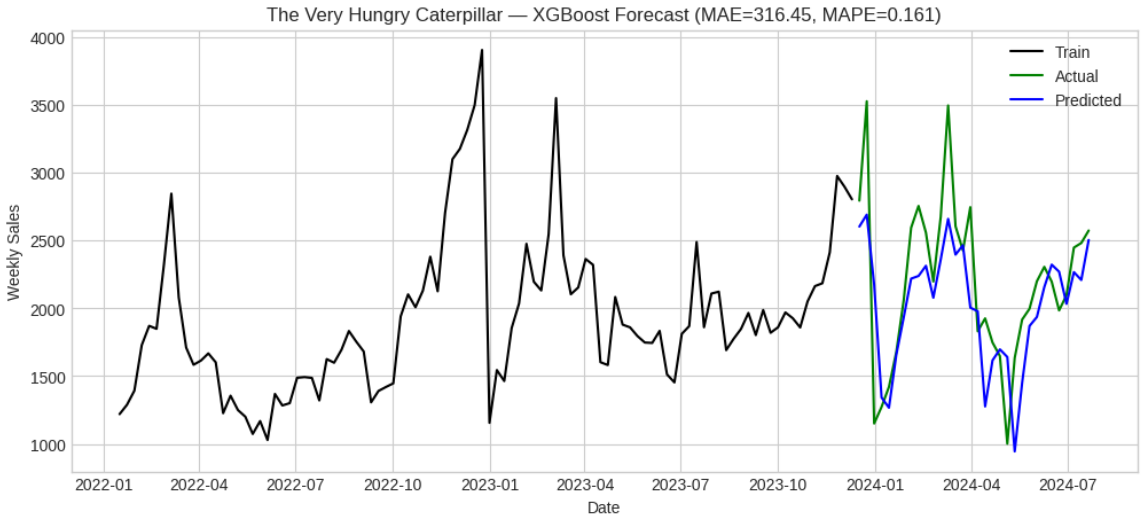


Figure 3: ***Actual vs XGBoost predictions for both books***

Key insight: XGBoost clearly outperformed SARIMA on both titles, especially for The Alchemist. Its ability to capture non-linear dependencies made it more flexible for forecasting book sales.

Deep Learning Models: LTSM

LSTM models were trained with hyperparameter tuning (units, dropout, learning rate). Results showed:

- *The Alchemist*: MAE ≈ 130, MAPE ≈ 0.21
- *Caterpillar*: MAE ≈ 350, MAPE ≈ 0.19

Forecast curves showed **reasonable short-term alignment** with actual values but tended to flatten in the 32-week horizon due to limited data size and high regularization.

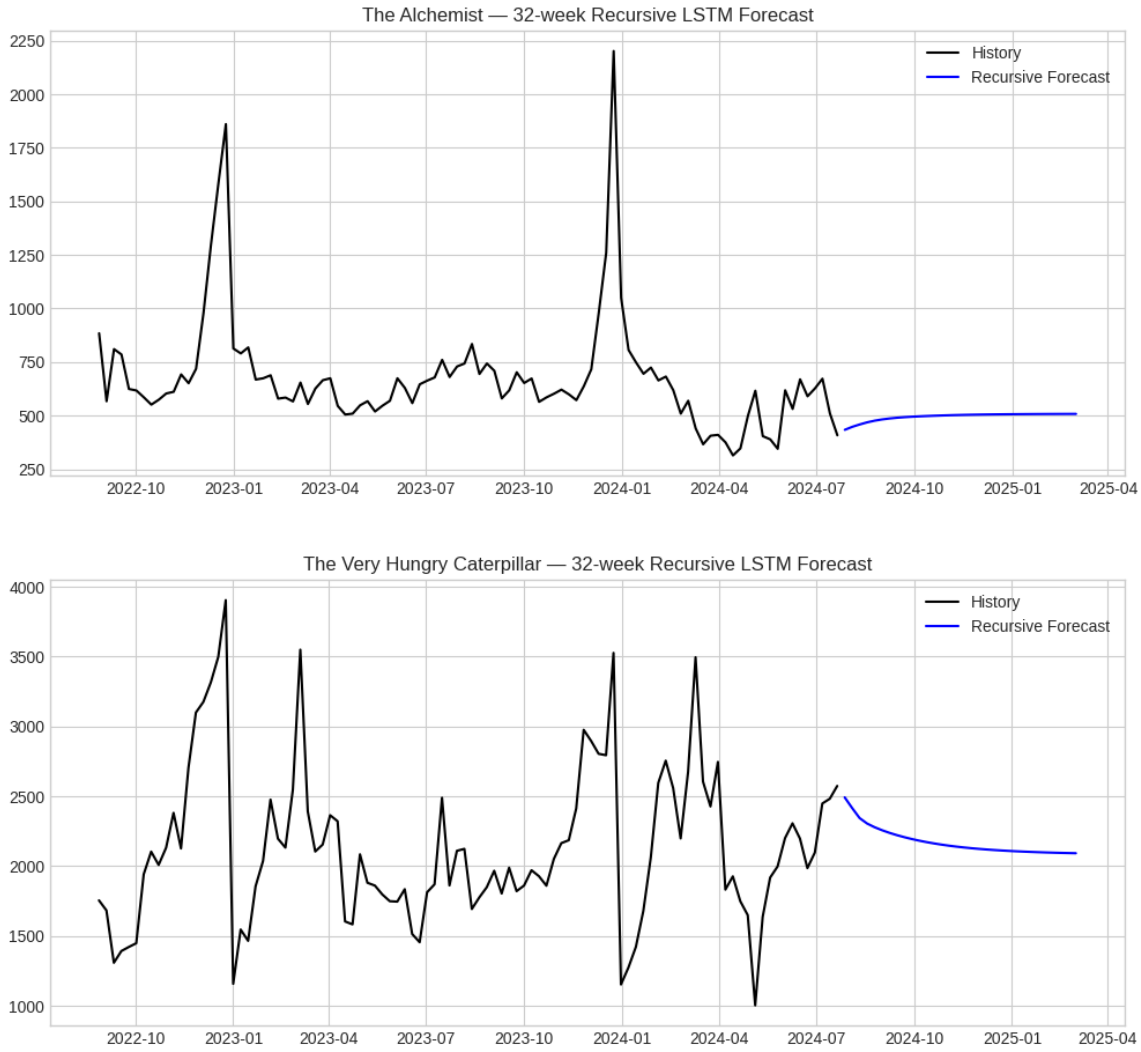


Figure 4: ***Actual vs LSTM forecasts***

Key insight: Contrary to expectations, LSTM did not outperform XGBoost or SARIMA. The main reason is the relatively small dataset (only ~650 weekly points per book), which is insufficient for deep learning to fully exploit its strengths.

Hybrid Models

Sequential (Residuals)

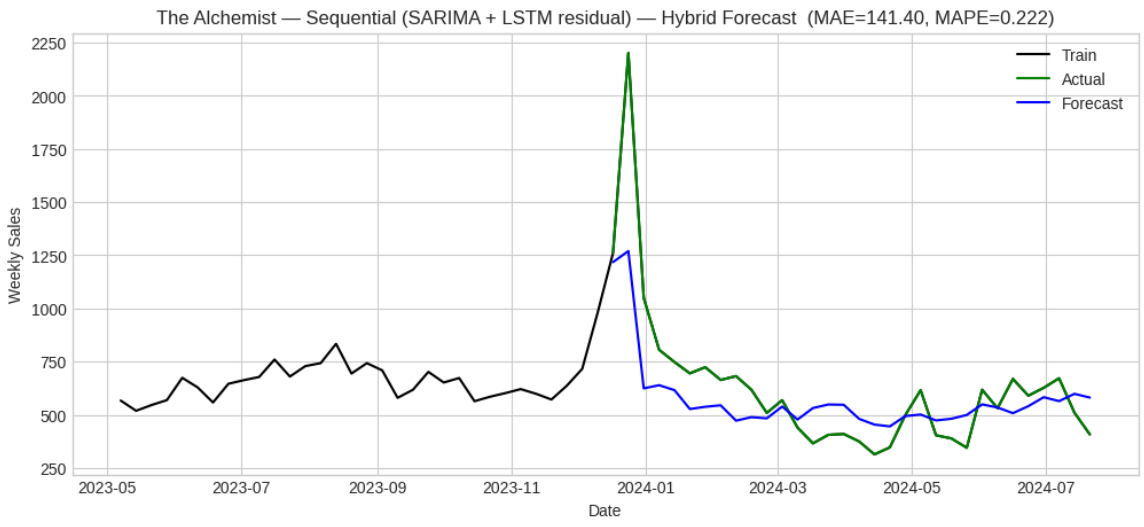
SARIMA residuals were modeled with LSTM. However, the hybrid did not improve significantly over pure SARIMA.

- *The Alchemist*: MAE ≈ 141, MAPE ≈ 0.22
- *Caterpillar*: MAE ≈ 370, MAPE ≈ 0.183

Parallel (Weighted Average)

SARIMA and LSTM forecasts were combined with a weight parameter α ($0 \leq \alpha \leq 1$). A grid search identified optimal weights:

- *The Alchemist*: best $\alpha \approx 0.7 \rightarrow$ MAE = 125, MAPE = 0.19
- *Caterpillar*: best $\alpha \approx 0.6 \rightarrow$ MAE = 385, MAPE = 0.19



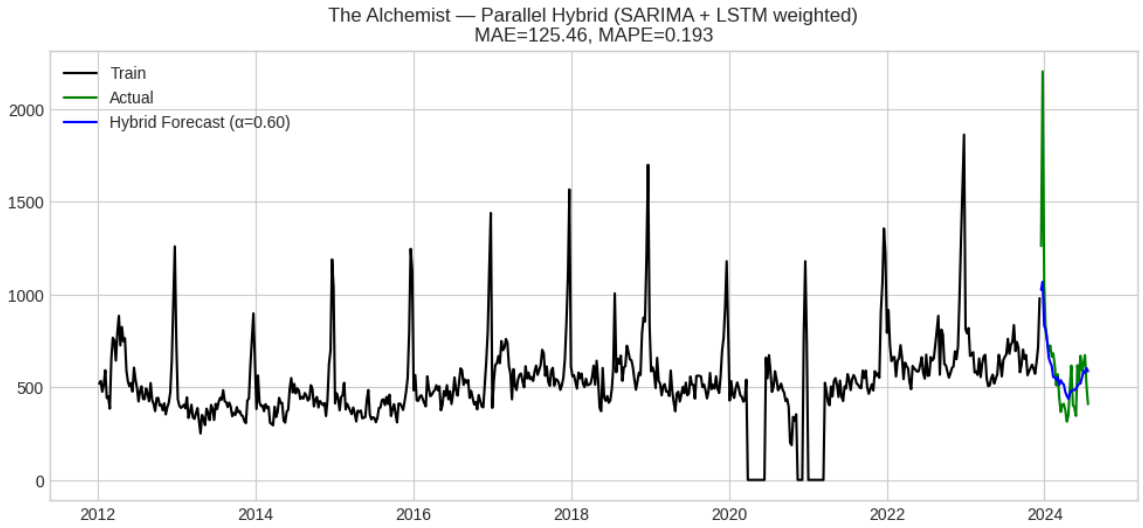


Figure 5: ***Hybrid forecast comparison***

Key insight: The hybrid approach showed **slight gains** after weight tuning. In practice, it added complexity without providing substantial business advantage compared to XGBoost.

Monthly Forecasting

Weekly data was aggregated to monthly sales to test if coarser granularity improves performance. Forecast horizon = **8 months**.

XGBoost (Monthly)

- *The Alchemist*: MAE ≈ 777, MAPE ≈ 0.34
- *Caterpillar*: MAE ≈ 2190, MAPE ≈ 0.20

SARIMA (Monthly)

- *The Alchemist*: MAE ≈ 783, MAPE ≈ 0.35
- *Caterpillar*: MAE ≈ 2069, MAPE ≈ 0.22

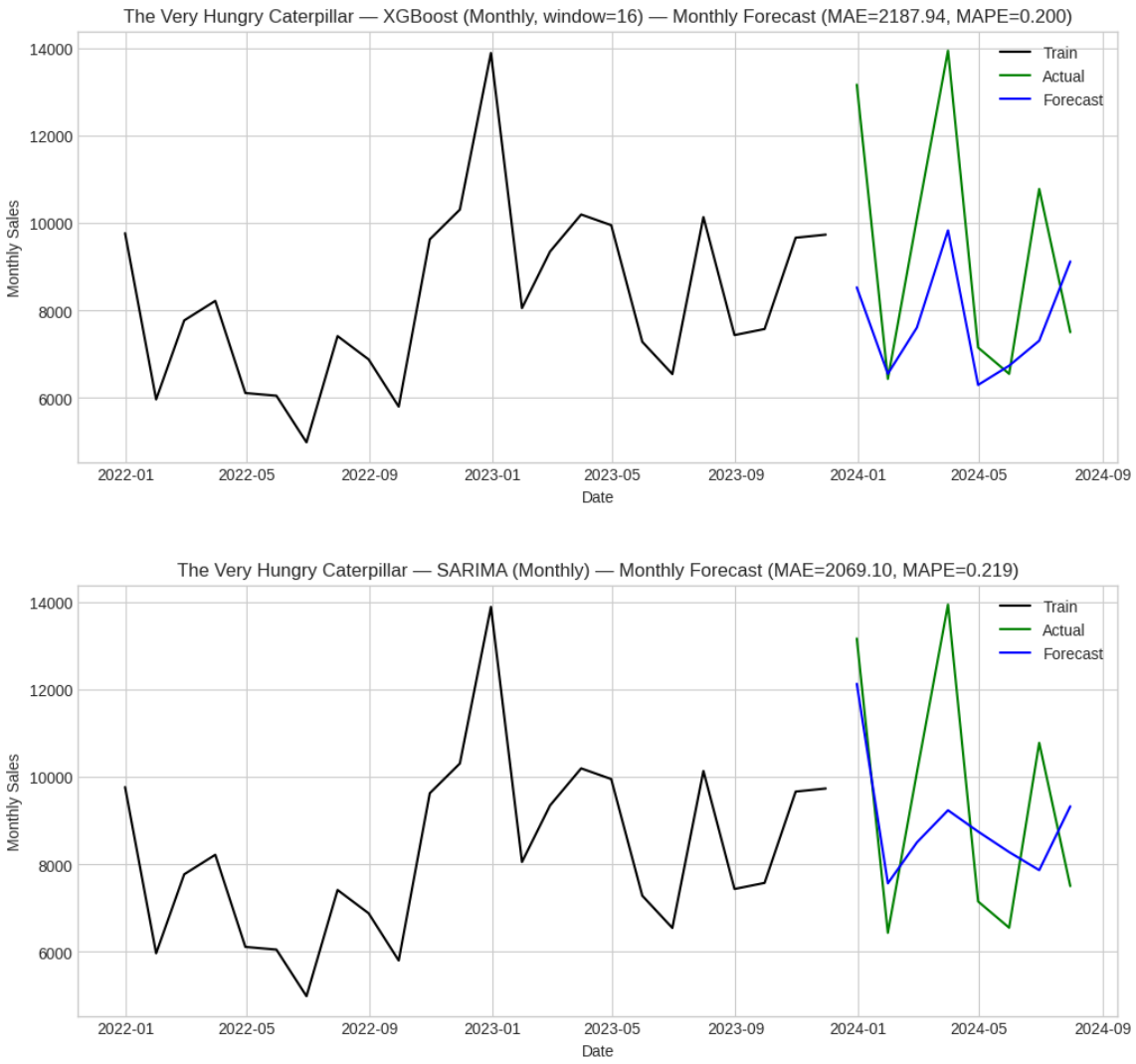


Figure 6: ***Monthly forecasts***

Comparative summary

Model	Horizon	The Alchemist (MAE / MAPE)	The Very Hungry Caterpillar (MAE / MAPE)
SARIMA (Weekly)	32 wks	139 / 0.23	396 / 0.19
XGBoost (Weekly)	32 wks	**109 / 0.18**	**316 / 0.16**
LSTM (Weekly)	32 wks	130 / 0.21	350 / 0.19
Hybrid (Best α)	32 wks	125 / 0.19	385 / 0.19
XGBoost (Monthly)	8 mos	777 / 0.34	2190 / 0.20
SARIMA (Monthly)	8 mos	783 / 0.35	2069 / 0.22

Business Implications

- **XGBoost delivered the best weekly forecasts**, making it the most reliable tool for **short-term stock planning**. Independent publishers can use it to anticipate demand spikes and avoid costly overstocking.
- **Monthly aggregation sacrifices predictive accuracy**, especially for books with fluctuating demand. The smoothing effect of monthly aggregation obscures short-term patterns such as seasonal spikes (e.g., holiday sales, school terms), which the weekly models are better at capturing.
- From a business standpoint, this suggests that **monthly-level forecasts are less reliable for operational decisions such as inventory restocking and short-term marketing campaigns**. While they may provide a broad view of long-term demand trends, their reduced precision makes them less suitable for fine-tuned decision-making.

- **SARIMA remains interpretable** and explains seasonal drivers but lacks predictive accuracy. It could still be used for quick benchmarks.
- **LSTM and hybrids did not outperform simpler methods**, largely due to limited data size. Deep learning may show advantages if applied to **larger multi-title datasets**.

Conclusion

This project demonstrated that machine learning, particularly **XGBoost with tuned lag windows**, provides the most accurate and business-relevant forecasts for book sales among the models tested.

For publishers:

- Weekly XGBoost forecasts can guide **operational stock management**.
- **Monthly forecasts can be supplementary** when evaluating long-term lifecycle potential of a title, but they should not be solely relied upon for tactical decisions.
- Hybrid and deep learning approaches should be considered only when larger datasets become available.

Overall, the analysis confirms that **data-driven forecasting can reduce uncertainty in publishing decisions**, leading to better alignment between supply and demand, cost savings, and stronger profitability for independent publishers.