Time Series Analysis:
Book Sales and Demand Forecasting

1 Introduction

The objective of this project is to forecast future book sales using historical sales data to support independent publishers. Accurate sales forecasting can inform smarter stock control, reduce risk in reprinting decisions, and help identify titles with strong long-term potential.

The analysis focuses on two datasets provided by Nielsen BookScan, representing the characteristics of 500 books and their weekly sales figures. To model future sales, two forecasting approaches were employed: Auto ARIMA, a statistical method well-suited for capturing trend and seasonality, and XGBoost, a machine learning model capable of learning complex lag-based patterns from engineered time series features. These models were used to forecast sales volumes of two books on both a weekly and monthly basis.

2 Methodological Overview

The primary forecasting challenge in this project was to predict future book sales using historical data, capturing both short-term fluctuations and long-term seasonal patterns. Forecasts were generated for the final 32 weeks using weekly sales data and for the final 8 months using monthly aggregated data. These two timeframes enabled a comparison of model performance across different temporal resolutions.

To prepare the data, weekly sales records were sorted chronologically. Because the raw dataset excluded weeks with zero sales, the data was resampled to ensure a continuous timeline, with missing weeks filled using zeros. Figure 1 illustrates overall sales patterns for all books with data extending beyond July 2024. The plot highlights three key trends: (1) a steep decline in sales during the early 2000s, (2) a sharp drop to zero during the COVID-19 lockdowns, and (3) recurring spikes in sales toward the end of each year. The remainder of the analysis

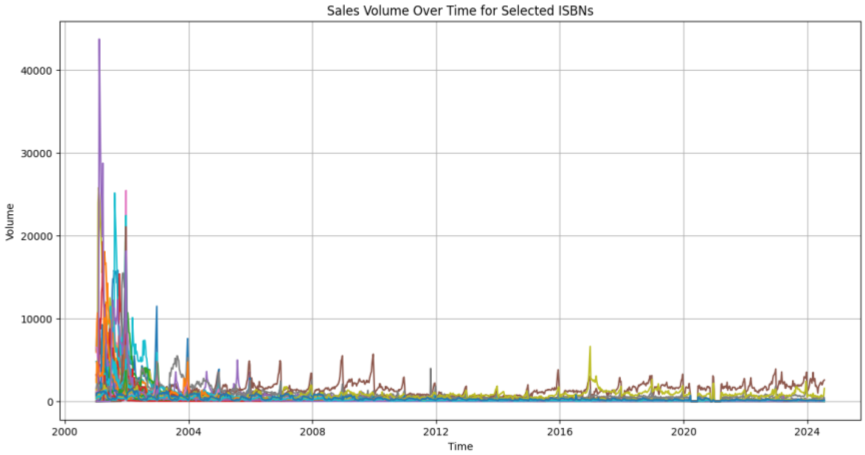


Figure 1

focused on two books with consistently available data from 2012 onward: *The Very Hungry Caterpillar* and *The Alchemist*, shown in Figure 2.

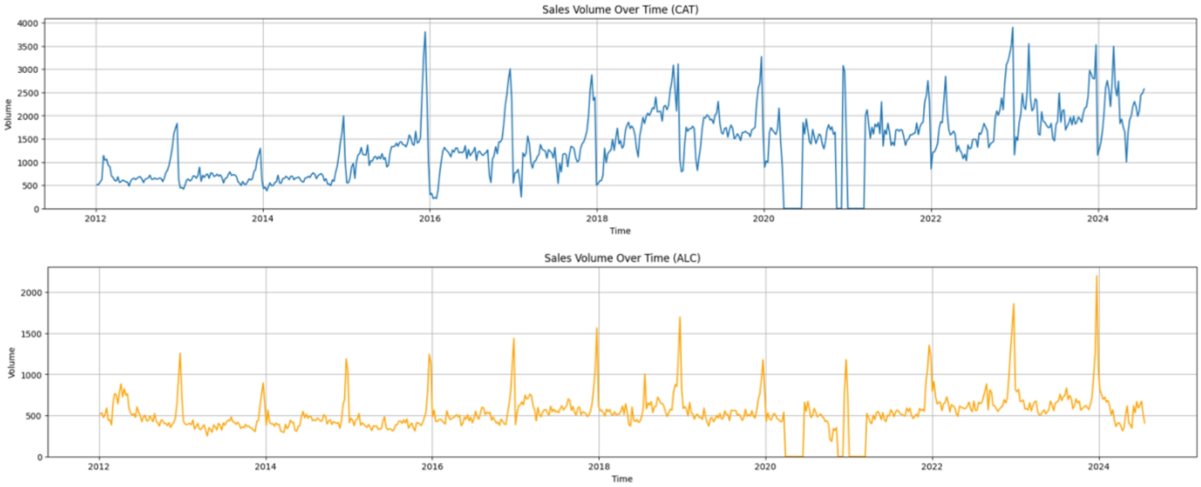


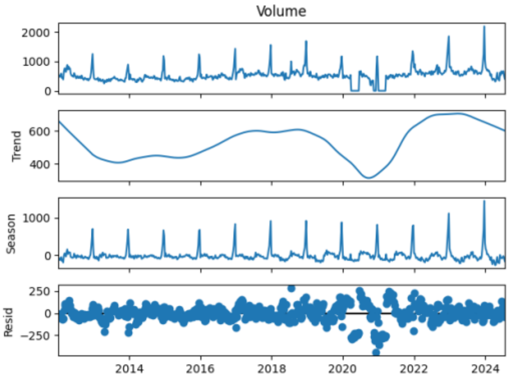
Figure 2

Two forecasting models were applied: Auto ARIMA and XGBoost. Auto ARIMA was chosen for its adaptive handling of seasonality and trend, while XGBoost used lagged sales values as features, with window length treated as a tuneable hyperparameter.

3 Exploratory Analysis

To understand the structure of the sales data, STL decomposition was applied to both books. Decomposition separates a time series into trend, seasonal, and residual components. For both *The Very Hungry Caterpillar* and *The Alchemist*, the seasonal component showed strong yearly patterns, with clear spikes in sales occurring at the end of each year. *The Caterpillar* also exhibited a rising long-term trend, while *The Alchemist* remained more stable in the long-term. Figure 3 shows the components of *The Alchemist.*[1]

Figure 3



Autocorrelation analysis was conducted using ACF and PACF plots (seen for *The Alchemist* in figure 4), which measure the relationship between observations at different time lags. These plots revealed residual seasonality and cyclic behaviour, supporting the presence of persistent seasonal sales patterns beyond just the end-of-year spikes.

To assess stationarity—the property of a time series having constant mean and variance over time—the Augmented Dickey-Fuller (ADF) test was used. The weekly datasets were found to be stationary, while the monthly data was non-stationary and required differencing for model fitting.

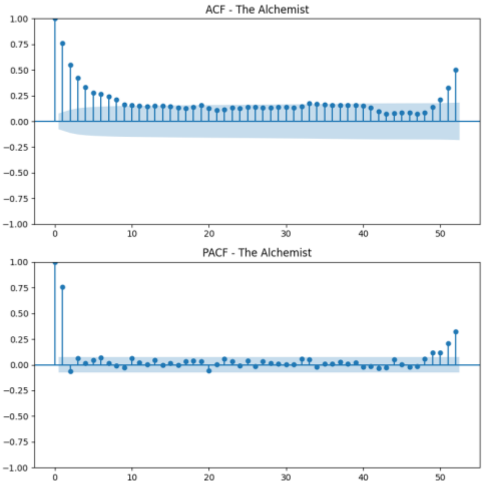


Figure 4

[1] Any additional figures can be found in the appendix of this document.

4 Model Design, Performance & Results

4.1 Auto ARIMA

Two models were used to forecast sales: Auto ARIMA and XGBoost. Auto ARIMA was a sensible choice, given its ability to automatically identify and adapt to patterns in the data by selecting optimal parameters for trend, seasonality, and differencing. Its built-in capacity to detect seasonal structures made it especially suitable for this task. As seen in the forecast plots for *The Very Hungry Caterpillar* (figure 5), Auto ARIMA consistently captured the annual sales spikes present in both the weekly and monthly data, aligning closely with known seasonal behaviour.

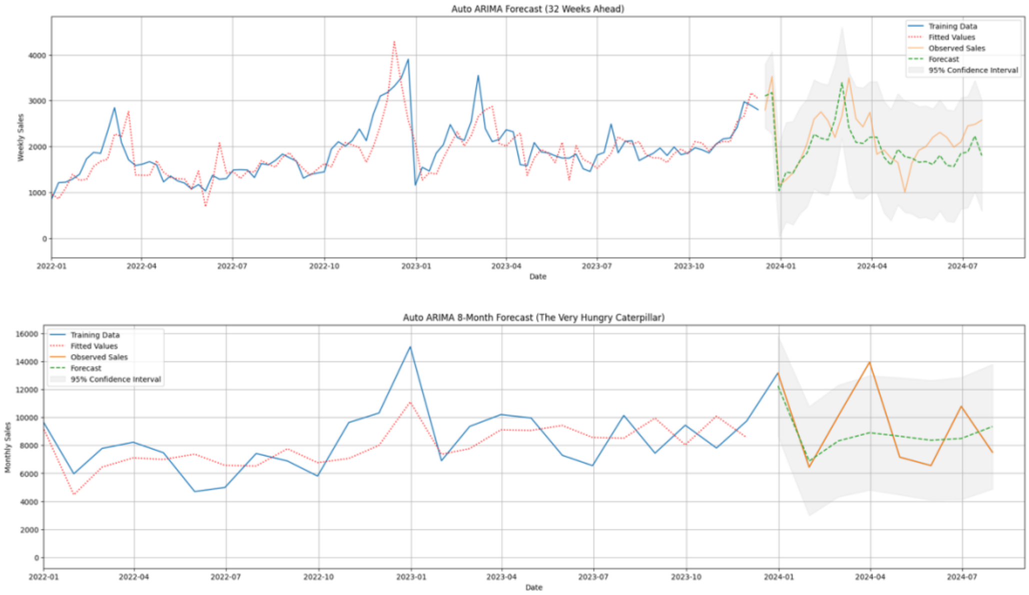


Figure 5

4.2 XGBoost

XGBoost was provided with lagged sales data as input features. Cross-validation was performed using an expanding window strategy to preserve temporal order. The hyperparameters *n_estimators*, *max_depth*, and *lag_window* were tuned using grid search. For the weekly data, while the *Mean Absolute Error* (MAE) and *Mean Absolute*

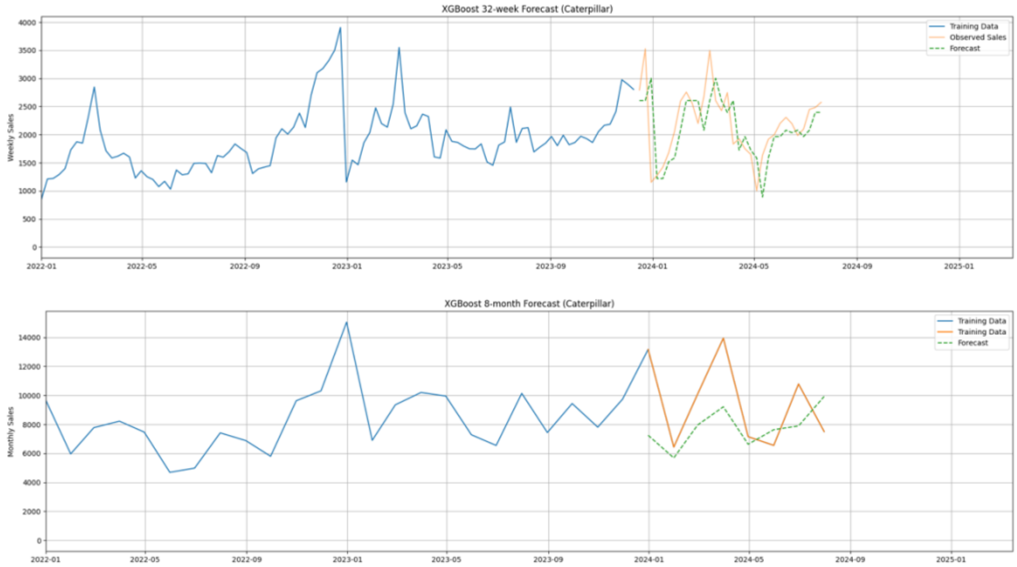


Figure 6

Percentage Error (MAPE) were higher than desired, the model's forecasts closely mirrored the shape of the true sales trajectory, indicating that it captured short-term dynamics well. However, when applied to monthly data, XGBoost struggled, failing to detect seasonal spikes and producing overly smooth forecasts. Figures 6 illustrates the predictions for *The Very Hungry Caterpillar* using weekly and monthly data. The contrast highlights each model's strengths: XGBoost seems to provide more reactive, fine-grained predictions, when the data is detailed as in the weekly case.

In summary, Auto ARIMA was more effective at forecasting major seasonal peaks, while XGBoost excelled at capturing subtler patterns in the more granular weekly data, though it lacked performance on the monthly series.

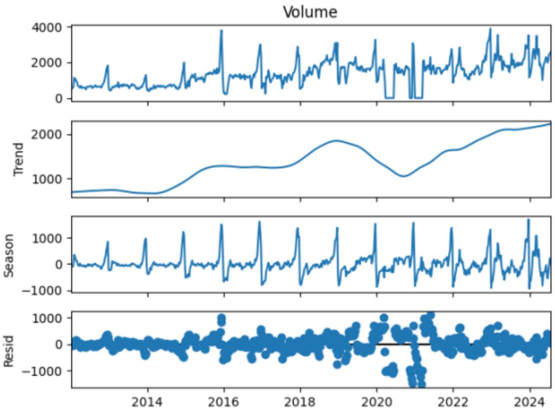
5 Final Insights & Conclusions

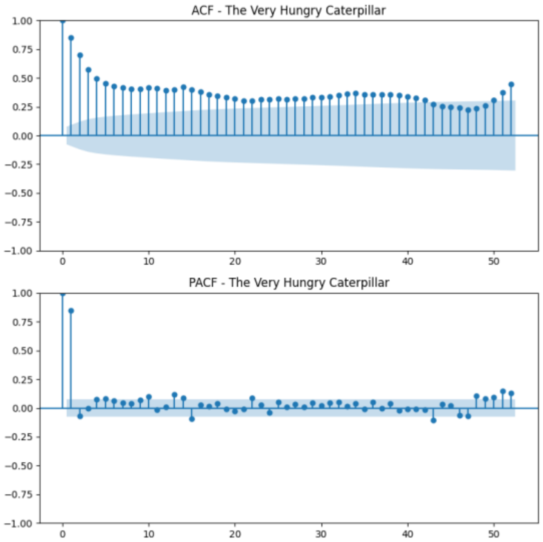
This project demonstrated that historical sales data can be effectively used to forecast future demand and guide better decision-making for independent publishers. Auto ARIMA proved highly capable of capturing seasonality and consistently predicted end-of-year sales spikes in both weekly and monthly formats. XGBoost was less successful at forecasting major peaks, particularly in the monthly data, but it performed well in modelling finer short-term fluctuations within the weekly data. These findings suggest that model choice should depend on the forecasting objective. Auto ARIMA is recommended when seasonality and long-term structure are key, while XGBoost may be better suited for short-term, high-frequency prediction.

In practical terms, publishers could rely on Auto ARIMA to estimate expected seasonal demand and use machine learning models to support more reactive, week-by-week inventory planning. As an alternative, providing a centralized index of books and their sales forecasts could greatly support independent vendors and represent a valuable business model in its own right.

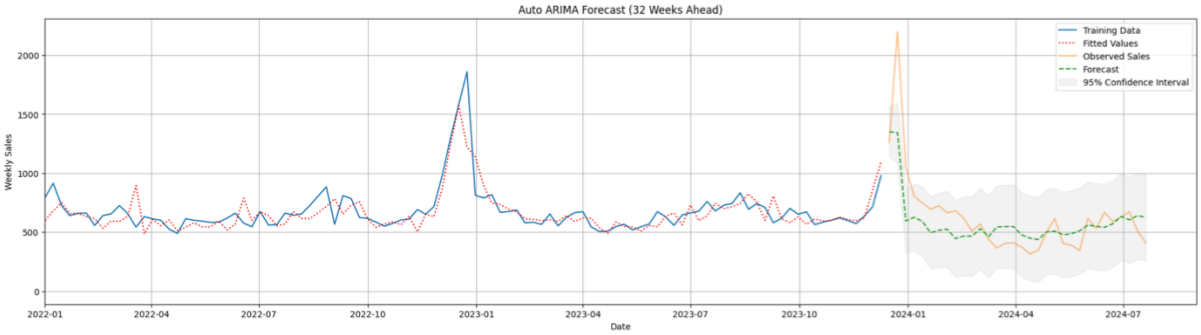
APPENDIX

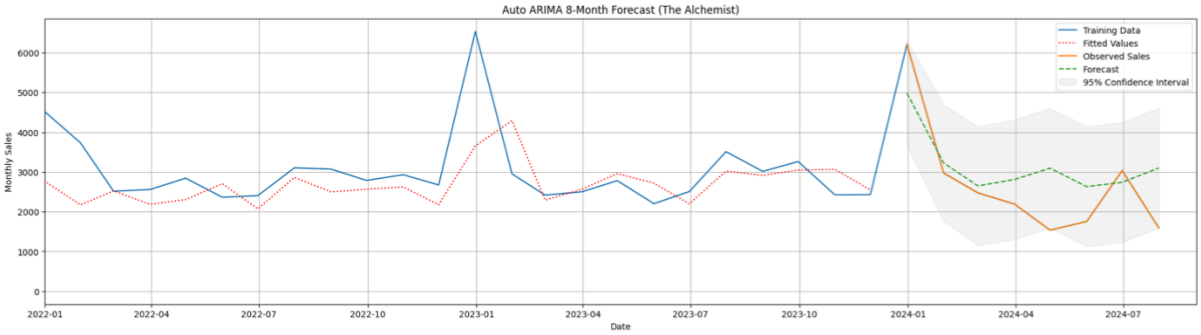
A – Decomposition and ACF/PACF for *The Very Hungry Caterpillar*



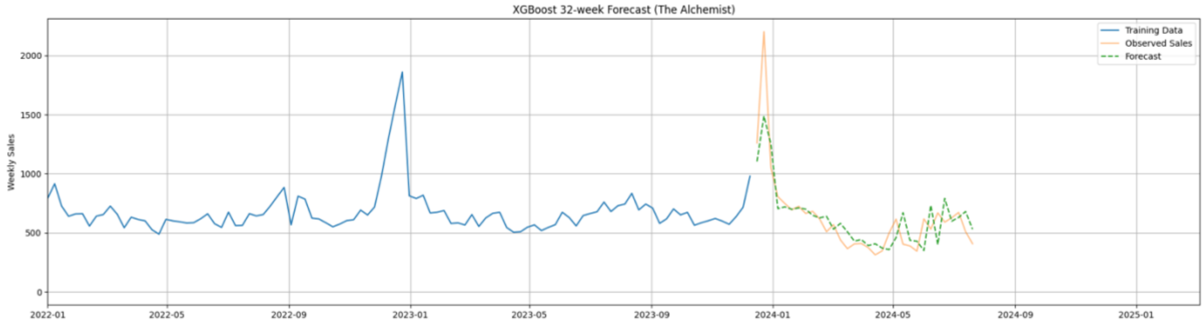


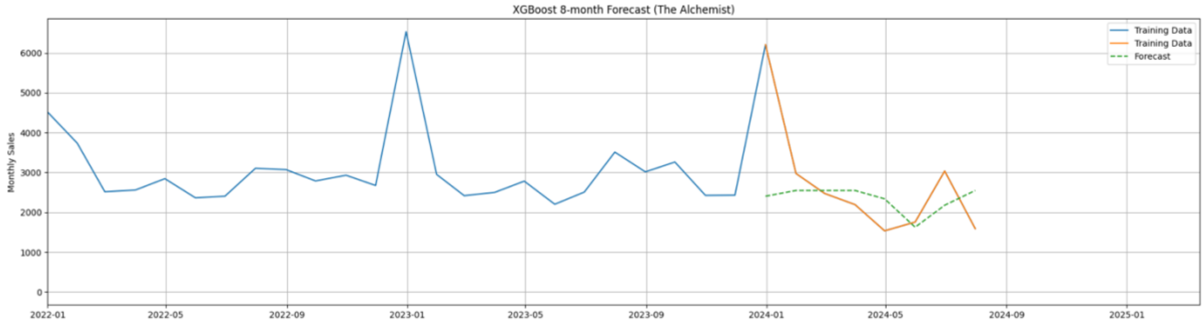
B – Auto ARIMA forecast *The Alchemist*





C – XGBoost forecast *The Alchemist*


XGBoost 32-week Forecast (The Alchemist)
Training Data
Observed Sales
Forecast
Weekly Sales


XGBoost 8-month Forecast (The Alchemist)
Training Data
Training Data
Forecast
Monthly Sales