Topic project 9.1 Using time series analysis for sales and demand forecasting

Prepared By: Mohamed Nuri

Date - 04/06/2025

Client - Nielson BookScan Service

Table of Contents

Introduction

Neilson BookScan is a leading book sales tracking service that collects highly accurate point-of-sale data from major UK book retailers, covering nearly 90% of the market. To support small to medium sized publishers, Nielsen aims to develop a forecasting tool that predicts a book's sales profile after publication. This would enable publishers manage upfront investment risks by identifying titles with strong long-term and seasonal demand.

Datasets:
Nielson provided us with two .xlsx files of real-world data from Nielsen. The first file is an ISBN list that contains industry-standard metadata about each book, and the other is a UK weekly trended timeline that contains weekly sales data.

1. ISBN List File:

2. UK weekly trended timeline Dataset:

Data Cleaning:

The main dataset used was the UK weekly trended timeline, which provided book sales volume data by ISBN across four product categories. These were combined into a single Data Frame and resampled the data to weekly frequency using the end date of each reporting period for consistency. Missing weekly data was filled using linear imputation. Negative sales volumes, which are considered invalid, were replaced with NaN values and imputing them using the same approach.

After investigating all the ISBN's timeseries plots, it was decided to conduct further analysis on the Volume of Sales of two books, 'The Alchemist and 'The Very Hungry Caterpillar'.

The dataset for both books were filtered to only contain data from 01-01-2012 up to the latest week sales were made.

Results & Conclusion

Decomposition:
STL was selected as the ideal model to utilise to understand the structure of the time series as it is robust against outliers and can help identify sales spikes. We utilised period of 52 weeks to decompose our data as our timeseries plots for the two books showed seasonality. Reviewing the time series plots for the volume of sales of the two books, it can be assumed that both are multiplicative, showing the amplitude of the peaks increasing or decreasing in the dataset at times. However, when transformation was performed. The transformation flattened the trend, dampened the seasonal variation to near zero and so it was assumed to be additive. The Alchemists Volume of sales data showed a strong seasonality strength of 0.79, and a has a trend of 0.51. The Hungry Caterpillars Volume of sales data similarly showed moderately high seasonality of 0.67 and a strong trend strength of 0.78.

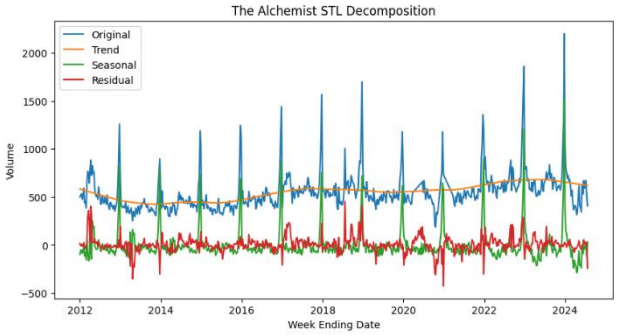

Figure 1

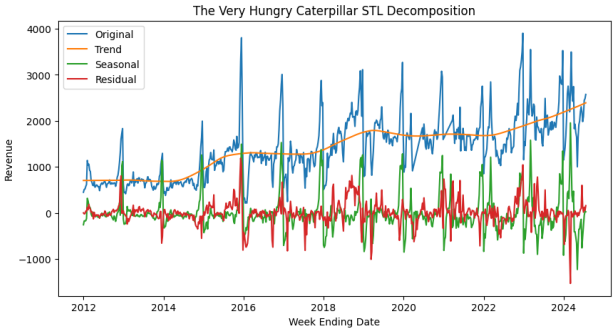

Figure 2

Auto-Correlation (ACF) of the Datasets:

The Alchemist's Volume of Sales, shows positive autocorrelation with a sharp decay in autocorrelation after the first lag. Only the first lag is significant and influences the current volume of sales. The Very Hungry Caterpillar volume sales show's decay after the first lag although decays slower than The Alchemist. Only the first lag is statistically significant to the current revenue. Both books show a high spike at week 52 indicating an annual seasonality for both books. Only the week before volume of sales influences current volume of sales. See figures 3 and 4.

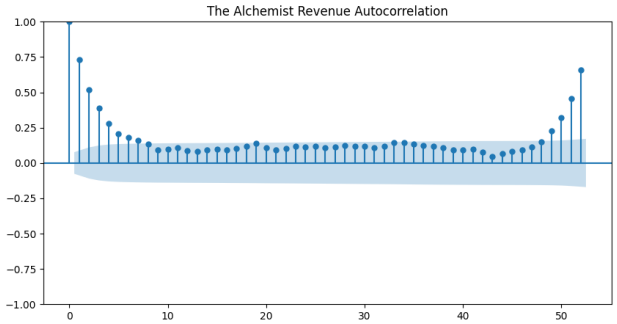

Figure 3

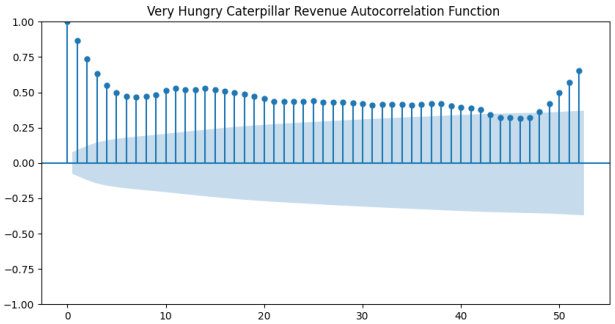

Figure 4

Partial-Autocorrelation (PACF) of the Datasets:

The PACF Plots for both books are similar, they show a strong positive correlation for the first two lags, however after that the PACF shows no Correlation. There is a spike after week 52 which indicates annual seasonality. See Figures 5 and 6.

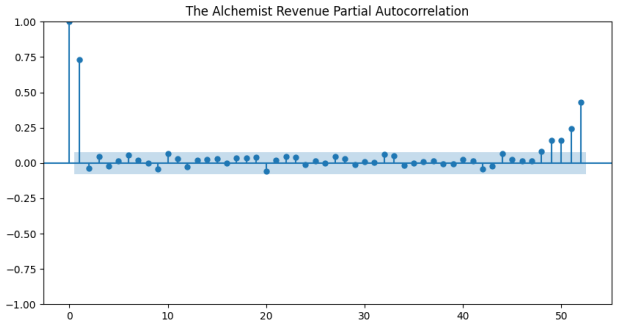

Figure 5

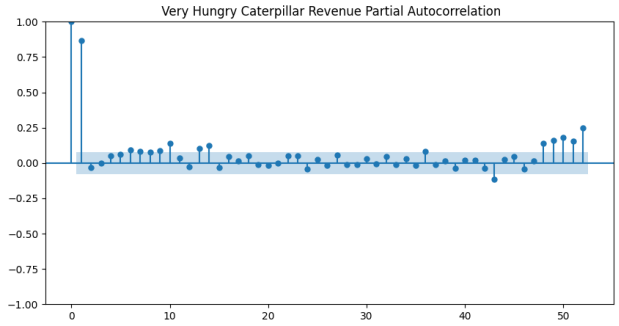

Figure 6

Stationarity ADF test for the Two Book Datasets:

The Alchemist shows a p value (1.36x10-17) which is less than 0.05. As a result, we reject the null hypothesis and assume the volume of sales feature is stationary. However, the p value (0.18) for The Very Hungry Caterpillar is greater than 0.05 thus, we fail to reject the null hypothesis and assume the volume sales feature is non-stationary based on the ad fuller stationarity test.

Auto-Arima modelling for each Book's Weekly Volume of Sales

The Alchemist Auto Arima Model Performance:

- The SARIMAX results for the Alchemist Volume of sales auto modelling showed a significant annual seasonal component (lag 52)
- The ljung box p-value < 0.05 and the null hypothesis is rejected meaning some of the residuals are autocorrelated.
- The AIC, BIC and HQIC are close to each other which suggests the model has reasonable complexity

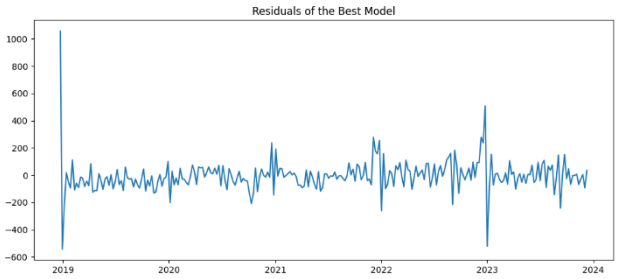

Figure 7

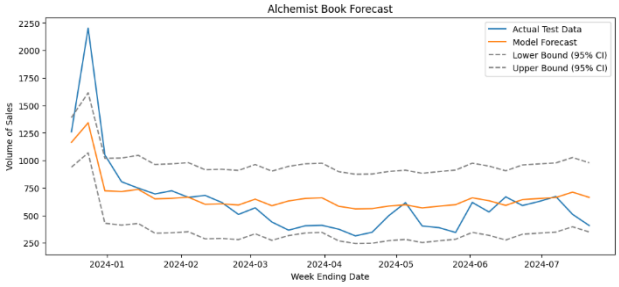

Figure 8

- In Figure 7, the residual plot for the model shows good performance as the residuals do not show clear structure or trend and behave like white noise around zero.
- Figure 8 shows the model forecast for 32 weeks compared to the actual, the model performs well on the alchemist dataset and the actual values fall within the 95% confidence interval.

Auto Arima Model Performance on The Very Hungry Caterpillar Dataset:

- The SARIMAX results showed a significant annual seasonal component (lag 52)
- The Ljung-Box p-value = 0.57, is not significant and suggests that the residuals are not autocorrelated, indicating the model has captured most of the patterns in the data.
- The AIC (3712.95), BIC (3730.73), and HQIC (3720.10) are close, suggesting the model has a balanced and reasonable complexity.

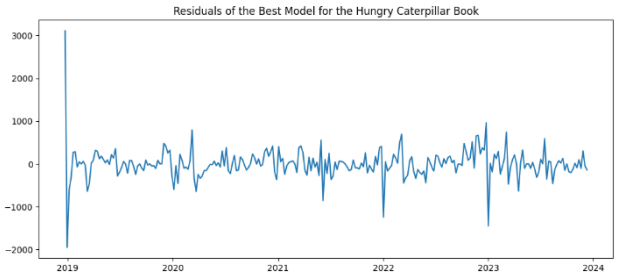

Figure 9

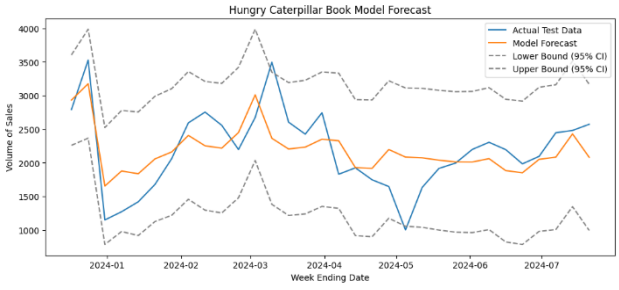

Figure 10

- In Figure 9, the residuals largely fluctuate around zero confirming they don't have a trend, seasonality or patterns.
- Figure 10 shows the Sarima model performance does well with most of the actual data inside the 95% confidence interval, the model captures the trend but underestimates sharp peaks and falls.

Machine Learning Techniques:

XGBOOST:

All data from 2012 were used as training data, excluding the last 32 weeks which was used as test data to evaluate model performance. To capture the seasonality and trend in the datasets, deseasonalising and detrending techniques were applied. The model was tuned on several hyperparameters. The best forecasting model was then used to predict the final 32 weeks.

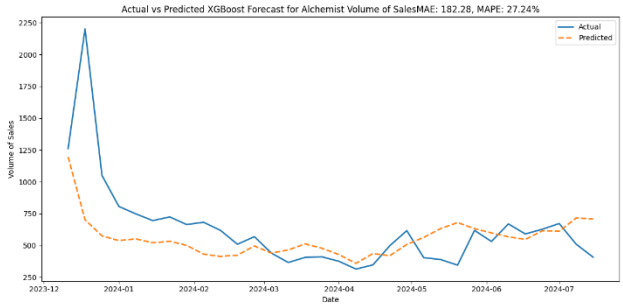

Figure 11

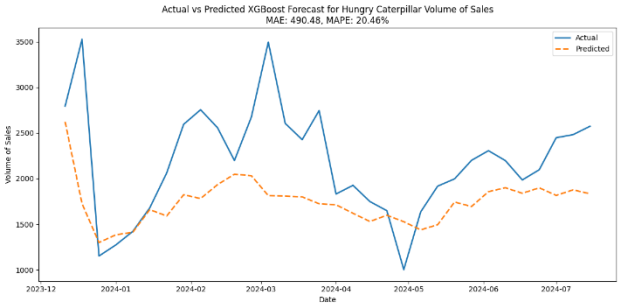

Figure 12

The XGBoost model performs well on the stationary Alchemist dataset, effectively capturing the overall pattern of sales volume. It achieves a Mean Absolute Error (MAE) of 182.28, (figure 11) which is relatively low compared to the dataset's mean sales volume of 550. The model successfully tracks the sharp rises and falls in the forecast period,

Similar to the Alchemist dataset, the XGBoost model performs well on the volume of sales data for 'The very Hungry Caterpillar'. The MAE of 490.48 (figure 12) is lower than the mean of 1405.37, with a MAPE of 20.6%. The model captures some of the seasonality towards the end of the year reflected in the original dataset but misses the sharper fluctuations in the rises and falls. Overall, the forecast is stable and tends to smooth out the peaks and troughs, producing a central estimate.

LSTM Modelling Process:

- the training data is weekly data over 10 years that excludes the last 32 weeks which it will be forecasting
- due to a strain of time, a window length of 52 was utilised to try to capture seasonality
- Recursive prediction strategy employed predicts one step forward, this was selected as the ACF and PACF showed that only one lag was significant in determining current volume, the ACF and PACF spikes at week 52 were included, as 52 was the window length employed

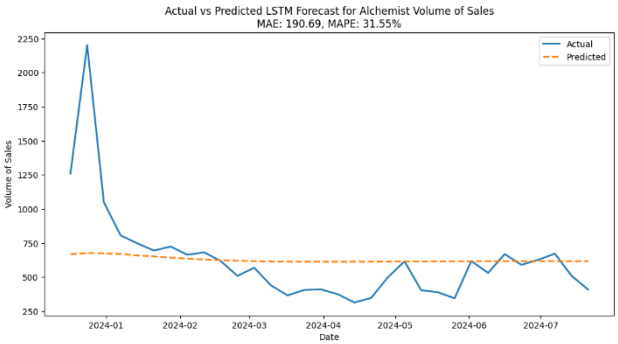

Figure 13

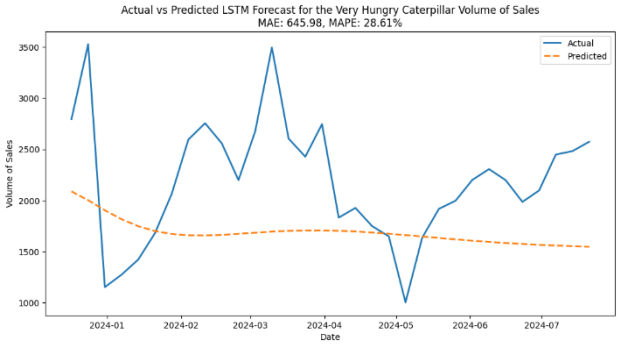

Figure 14

The LSTM Model has not performed too well on the Alchemist dataset, The MAE of 190 means that, on average, the model's predictions are off by 190 units. Given a typical sales volume of 550. The primary issue is the peak seasonality at the new year which is where the greatest deviation in forecasting the volume in sales occurs. The sharp rises, characterise an increase in volume of sales of the book at the New year, the model failed to capture this seasonality, though 10 years of historical data was passed through. A reason for this is is likely due to the recursive prediction strategy employed: the model predicts one step forward, then uses that prediction as part of the input for the next. This approach accumulates errors. Please figure 13.

Similar to the LSTM model applied to the Alchemist dataset, the model failed to capture the seasonal spike in sales that typically occurs in late December and early January for The Hungry Caterpillar. The model appears to approximate a central trend — generating forecasts that fall between the highs and lows. While it doesn't reflect short-term fluctuations or seasonal peaks, it identifies a relatively stable midpoint around which the actual sales vary. The model achieved a mean absolute error (MAE) of 645.98, compared to an actual weekly average of 1,405. Given the magnitude of the rises and falls, the model has predicative use as it captures the broader level around which sales fluctuate. Please see figure 14.

The Hybrid Model:

We performed further analysis on the modelling of the Alchemist Dataset, this time utilising a hybrid model. The hybrid model was formed using a SARIMA and a LSTM Model. The same lookback period of 10 years was utilised for the training data and used to predict 32 weeks.

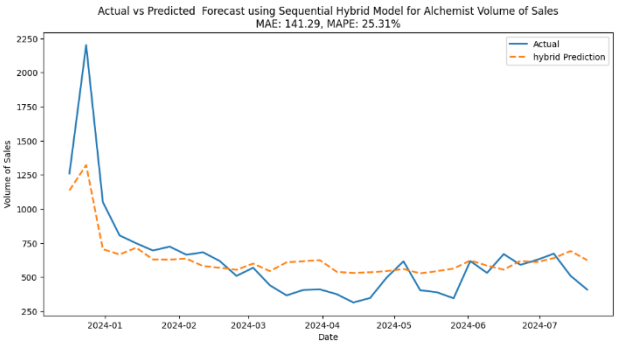

Figure 15

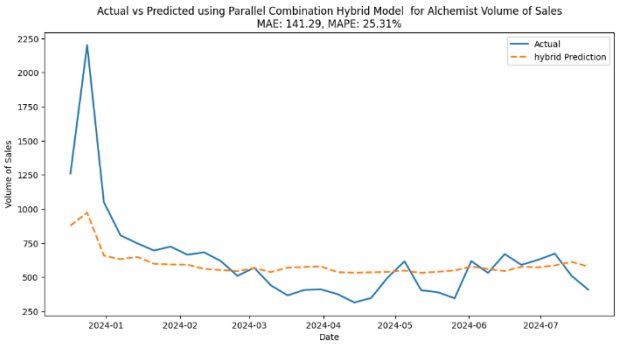

Figure 16

Please see figure 15, that shows the performance of the sequential Hybrid model, whereas figure 16 shows the performance of the parallel combination hybrid model.

The sequential hybrid model performs better at capturing the seasonal component at the start of the year, even though in the parallel combination sarimax was given a much higher weighting to improve model performance.

Monthly Prediction:

We aggregated weekly data into monthly for each dataset and retrained the XGBoost and SARIMAX model for comparison with models trained on the weekly datasets.

SARIMAX Model comparison between the monthly and weekly Datasets:

Figures 16 and 17 show SARIMAX modelled on monthly sales for the two datasets.

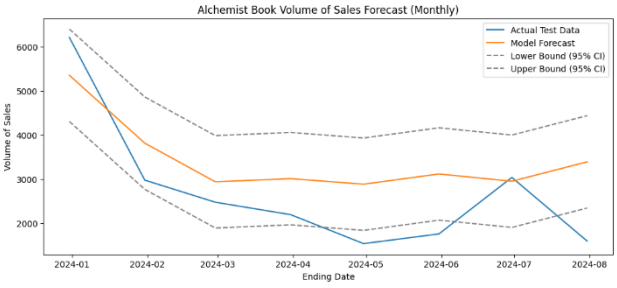

Figure 17

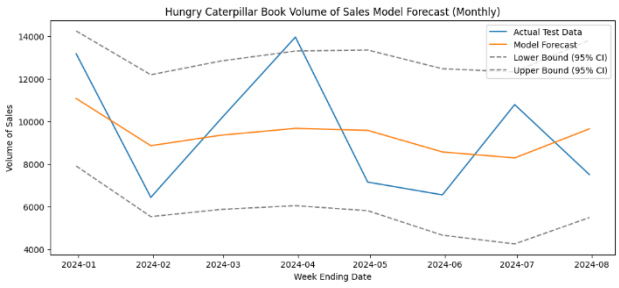

Figure 18

The Alchemist:
The weekly SARIMAX model identified a highly significant seasonal component at lag 52 ($p < 0.001$), confirming strong yearly seasonality in the weekly sales data. In comparison the SARIMAX best model for the monthly dataset, the seasonal component at lag 12 is only moderately significant ($p = 0.071$).

The weekly SARIMAX model demonstrates strong forecast performance, with all actual values falling within the 95% confidence interval, whereas for the SARIMAX trained on monthly some of the forecasted values fall outside the 95% confidence interval, particularly at peaks

The Very Hungry Caterpillar:
Both the weekly and monthly SARIMAX models detect strong seasonality. The weekly model more accurately captures rises and falls in sales volume. This is because it is more responsive to short-term shifts due to finer granularity.

XGBOOST Model comparison between the monthly and weekly Datasets:

Figures 19 and 20 show XGBoost modelled on monthly sales of the two books.

The Alchemist:
The weekly model performs significantly better, with a lower MAE (182.28 vs 781.52) and MAPE (27.24% vs 36.37%) than the monthly model. It captures short-term fluctuations and sharp peaks more accurately.

The Very Hungry Caterpillar
The weekly model also outperforms here, with a lower MAE (490.48 vs 1951.74) and slightly better MAPE (20.6% vs 21.15%). It is a better trend and seasonal fit, although still underestimates sharp shifts. The monthly model struggles more with variability and produces higher forecast errors.

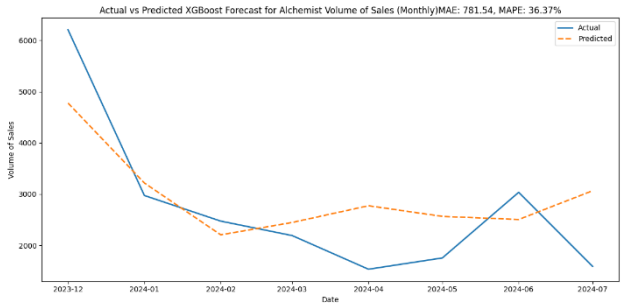

Figure 19

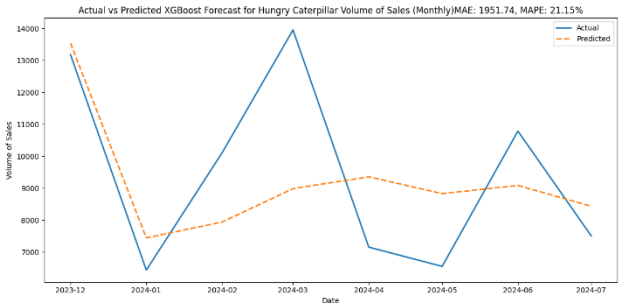

Figure 20

Evaluation:

- For the LSTM model, Volume of Sales was the only input into the model, End Date should have been included to improve model performance. Also a recursive prediction strategy was employed in the modelling.This approach accumulates errors

- Univariate Timeseries forecasting was conducted, exogenous variables were not explored,