DEMAND FORECASTING AND PRICE OPTIMIZATION

Table of Contents

[CHALLENGE: 3](#_Toc105424865)

[DATA USED: 3](#_Toc105424866)

[FEATURE ENGINEERING: 4](#_Toc105424867)

[MODELS: 4](#_Toc105424868)

[Linear Regression: 4](#_Toc105424869)

[ARIMA: 5](#_Toc105424870)

[XGBoost: 5](#_Toc105424871)

[Random Forest: 8](#_Toc105424872)

[RESULTS: 8](#_Toc105424873)

[CONCLUSIONS: 9](#_Toc105424874)

[FUTURE CONSIDERATIONS FOR DEMAND FORECASTING: 9](#_Toc105424875)

# CHALLENGE:

The goal is to build a price optimization engine. By forecasting the sales quantity of each item for the upcoming weeks and therefore sales of the next quarter we derive an understanding of how price impacts the sales quantity. Using this we need to optimize the price of each item in order to maximize sales.

To look at historical Q1 and Q2 (so far) sales data, implement a new demand forecasting model with this data and predict upcoming demand for the next two quarters. Use this forecasted demand as a basis for building a dynamic pricing model.

# DATA USED:

Ashcomm sales data was used to create a model that predicts future sales. All SKUs with sales in the last 180 days were taken. For each of those SKUs, all historic sales data starting from potentially 2019 (or whenever that SKU started selling) was gathered. Returns and cancellations were ignored. Sales information was gathered for each individual price zone. Also included was data from 2015 on when each tentpole sales event was, with details on pre-event, main event and post-event.

* Item/SKU list – All New, Current or Discontinued items with sales in the past 180 days except for Third Party Items, Items with names beginning with \*, items with adjustments and items in the list ADJPRO, FURNPRO, KADJPRO, OUTDRF, DS Small Parcel, Home Delivery, Threshold Delivery
* Previous Week sales -

**Data Dictionary:**

|  |  |
| --- | --- |
| **Data Field** | **Description** |
| ItemID | SKU or Item ID from Retail Item Master |
| ProductLine | Product line information from Retail Item Master. |
| Status | Item Status – New, Current and Discontinued items used |
| DateID | Date |
| PromoYear | Promo year value starting with 1 for 2015 upto 7 for 2022. The PromoYear starts on the Tuesday that has the first calendar week of the year. This can start either on the last week of the previous calendar year, or first week of current calendar year depending on when the Tuesday falls. |
| PromoQuarter | Promo quarter value ranging from 1 to 4 for each quarter in a year. The PromoQuarter starts on the Tuesday that has the first calendar week of the quarter. This can start either on the last week of the previous calendar quarter, or first week of current calendar quarter depending on when the Tuesday falls. |
| PromoMonth | 4-5 week period starting with the beginning of the year, with each PromoMonth period starting on a Tuesday. |
| PromoWeek | 52 or 53 weekly periods starting with the beginning of the PromoYear each on Tuesday. |
| FiscalYear | Fiscal year value starting with 1 for 2015 upto 7 for 2022. The FiscalYear starts on the Sunday that has the first calendar week of the year. This can start either on the last week of the previous calendar year, or first week of current calendar year depending on when the Sunday falls. |
| FiscalQuarter | Fiscal quarter value ranging from 1 to 4 for each quarter in a year. The FiscalQuarter starts on the Sunday that has the first calendar week of the quarter. This can start either on the last week of the previous calendar quarter, or first week of current calendar quarter depending on when the Sunday falls. |
| FiscalMonth | 4-5 week period starting with the beginning of the year, with each FiscalMonth period starting on a Sunday. |
| FiscalWeek | 52 or 53 weekly periods starting with the beginning of the FiscalYear. |
| Price Zone Assigned | The Price Zone against which each sale is logged. |
| Previous Week sales | Sales for the same SKU, same price zone in the previous fiscal week when compared to the week being processed. |
| Previous Week sales | Sales for the same SKU, same price zone in the previous fiscal week when compared to the week being processed. |
| Previous Week 2 sales | Sales for the same SKU, same price zone from two fiscal weeks when compared to the week being processed. |
| Previous Week 3 sales | Sales for the same SKU, same price zone from three fiscal weeks when compared to the week being processed. |

# FEATURE ENGINEERING:

* Holiday – If there is a particular day of the week is a holiday, mark as 1.
* Z scores – Z-Score of Difference from Average Price
* Z-Score of Logged Ratio of Price/AveragePrice
* Prev\_week\_sales, Prev\_week2\_sales, Prev\_week3\_sales
* Pre-holiday and post-holiday events as dummy variable
* Months as dummy variable
* Log price
* Holiday\_Prev\_Week, Holiday\_Prev\_Week2 – Interaction variable with holiday and previous week sales

# MODELS:

## Linear Regression:

The base model looked at 2022 E-Commerce data and built a model at the Item ID or SKU level. This was found to over forecast weekly sales.

The first iteration of revisions involved restricting the list of SKUs in the training data to just those from a single price zone (price zone Z16). Model building was also revised to build the model at a product line level so that each product line would have its own model. This brought down the number of models from the thousands to 10. Predicting was done at the item level so that demand (and sales) was forecast for each item within each of those models. Training and prediction were both for the promo week.

* **Conclusion:** This was also found to over forecast demand, but the model performed a little better.

The next iteration involved using data from all 5 price zones (Z16, Z23 and Z5) and building models at the product line level, while predicting demand and sales at the item level.

* **Conclusion:** This showed similar results to the previous iteration.

Further iterations of the linear regression model will look at item grouping or segmenting to focus more on items with more sales and to group lower sales items into categories based item price (high, medium and low) to see if that generates more useful data points. Another option is to group items either on RetailCategoryGroup or ItemCode to see if grouping like items yields better results.

Chart, scatter chart

Description automatically generated

## ARIMA:

ARIMA is a time-series oriented model that provides insights on seasonality in data.

The first iteration of building an ARIMA model used data from 2021 and 2022 (up to May 2022) and was trained on and predicted on the promo week level. The results were not conclusive as a lot of items were predicted to have zero sales.

* **Conclusion:** This was put on hold to pursue more viable options.

Chart, scatter chart

Description automatically generated

## XGBoost:

This tree-based model was the next approach to be tried.

1. The first iteration involved training the model for all price zones for 2022 (upto May) data. Training was done with sales at the promo week level and prediction was also done for the promo week. Model was built at the product line level and prediction of demand and sales were done for each item that had an associated price zone.

* **Conclusion:** The results were very promising with a favorable prediction of demand, but the predicted sales an under-forecast.

Chart, scatter chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Product line | R-Square | RMSE | MAE |
| Motion | 0.86 | 1.3 | 0.93 |
| Stationary | 0.88 | 2.07 | 1.39 |
| Accessories | 0.87 | 1.6 | 1.05 |
| Bedroom | 0.95 | 2.63 | 1.66 |
| Dining | 0.96 | 6.06 | 3.62 |
| Entertainment | 0.97 | 1.72 | 1.21 |
| Home Office | 0.95 | 1.29 | 0.93 |
| Bedding | 0.99 | 3.42 | 2.18 |
| Outdoor | 0.93 | 1.68 | 1.2 |
| Occasional | 0.91 | 1.84 | 1.28 |

1. The next iteration involved training the data with SKUs that have sales in the last 6 months and getting the sales data for these SKUs from 2015. Data was at the daily sales level including days that do not have sales. This provided information to the model about zero and positive sales. Model was trained with both zero sales and non-zero sales.

With Zero Sales

Chart, scatter chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Product line | R-Square | RMSE | MAE |
| Stationary | 0.92 | 1.16 | 0.24 |
| Accessories | 0.88 | 0.62 | 0.11 |
| Motion | 0.9 | 0.47 | 0.09 |
| Bedroom | 0.94 | 1.54 | 0.35 |
| Dining | 0.94 | 4.6 | 1.16 |
| Occasional | 0.93 | 1.23 | 0.37 |
| Bedding | 0.98 | 3.1 | 0.72 |
| Entertainment | 0.94 | 1.14 | 0.32 |
| Outdoor | 0.93 | 0.82 | 0.16 |
| Home Office | 0.92 | 1.25 | 0.38 |

Without Zero Sales

Chart, scatter chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Product line | R-Square | RMSE | MAE |
| Stationary | 0.91 | 2.7 | 1.35 |
| Accessories | 0.83 | 1.81 | 1.05 |
| Motion | 0.85 | 1.41 | 0.84 |
| Bedroom | 0.93 | 3.25 | 1.61 |
| Dining | 0.94 | 8.53 | 4.09 |
| Occasional | 0.91 | 2.38 | 1.41 |
| Bedding | 0.98 | 6.21 | 2.9 |
| Entertainment | 0.93 | 2.32 | 1.37 |
| Outdoor | 0.9 | 2.53 | 1.59 |
| Home Office | 0.91 | 2.46 | 1.52 |

* **Conclusion**: At the end of the second iteration of XGBoost, it was determined that we would proceed without zero sales data for the time being.

## Random Forest:

Another tree based algorithm, the first iteration for Random Forest was built in very similar lines to the first iteration of XGBoost. This involved training the model for all price zones for 2022 (upto May) data. Training was done with sales at the promo week level and prediction was also done for the promo week. Model was built at the product line level and prediction of demand and sales were done for each item that had an associated price zone.

* **Conclusion:** Random Forest also seemed to produce good results. Upon further analysis, it appeared that this model was overfitting the predictions. This was put on hold to pursue XGBoost.

Chart, scatter chart

Description automatically generated

# RESULTS:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Iteration | Iteration Description | Result | Conclusion |
| Linear Regression | 1 | Trained on one price zone, product line level modelling. | Actual Sales ~ $10m  Predicted Sales ~$20m | Over-forecasting. |
| Linear Regression | 2 | Use SKUs from 5 price zones (1,2,3,5,6) | Actual Sales ~ $10m  Predicted Sales ~$23m | Better results than Iteration 1. Over-forecasting. Can try item grouping or segmentation in another iteration. |
| ARIMA | 1 | Sales from 2021 and 2022 used for training. Product line level modelling. |  | Too many zero sales predictions. Put on hold. |
| XGBoost | 1 | Sales from 2022. All price zones. Product line level modelling. Promo week prediction. Only non-zero sales data. Prediction for Week 22. | Actual Qty: 32625  Predicted Qty: 37380.6  Predicted Sales: $12,850,689 | Good results. Under-forecasting sales. |
| XGBoost | 2 | Zero sales price data. Sales from 2019. Use price instead of average weekly price. Tried with and without previous week sales. Week 23. | With Zero Sales  Actual Qty: 23478  Actual Sales: $8,080,130  Predicted Qty: 42063.6  Predicted Sales: $19,366,351  Without Zero Sales  Actual Qty: 23462  Actual Sales: $8,070,431  Predicted Qty: 30129.1  Predicted Sales: $12,579,438 | For the time being, use sales without zero sales in order to refine this model. |
| Random Forest | 1 | Sales from 2022. All price zones. Product line level modelling. Promo week prediction. Only non-zero sales data. | Actual Qty: 32625  Predicted Qty: 37433.7  Predicted Sales: $12,659,106 | Overfitting results to training data. |

# CONCLUSIONS:

After considering several models, the model that performs the best so far and which seems the likeliest to provide valuable predictive data is the XGBoost model.

# FUTURE CONSIDERATIONS FOR DEMAND FORECASTING:

There are additional revisions that can be made to this model to improve the results. Some of those are listed here:

* Add ATP data. ATP data can be obtained for each SKU for each RDC as a weekly snapshot and has historical values from 2021.
* Add delivery message. This is the message that a user of the E-commerce website sees while placing
* Z score value – Find the average price value for each SKU and price zone combination and analyse the variation in the price value. This can be used as another data input to the model. This will provide more interpretability to the model.
* Post-prediction analysis – based on the results of the model, data analysis and post processing can measurably improve the results.
* Hyper-parametric tuning – Iteratively refine the parameters of
* Multi-model approach with a classifier – Split the model into two. The first model (classifier) uses sales data to predict whether a SKU in a price zone will have positive sales for the predicted week or not. Only for the SKUs that this model predicts as having sales, build and use a second model that predicts the amount of sales.

Other approaches that can be tried include:

* Time series based ARIMA models performed poorly in the initial work. Different time series based models including multiple ARIMA models that looks at weekly sales, tentpole event based sales on a year-over-year basis etc can be modelled differently to build a more accurate prediction engine.
* Vector Autoregression (VAR) is another time series based forecasting algorithm that can be used on seasonal data.