**Sales Forecasting Model for SkyMart: A Predictive Approach**

Justin Cassner

Department of Graduate Data Analytics, Franklin University

DATA 695: Capstone

Dr. Deanne Larson

May 10, 2025

**Background and Introduction**

SkyMart is a leader in the U.S. retail business sector with over 45 stores nationwide and is seeking to uncover valuable insights from newly gathered data relating to their business performance. With the progression of technology and its ability to generate loads of first-party data, retailers who utilize data analytic methods have a direct line to fully understanding their operations, products, and customers (8 Retail Analytics Benefits…, 2024). Beginning in 2010, SkyMart embarked on collecting this information and it is now time, in late 2012, to evaluate past performance and provide actionable insights for the future.

In general, data is collected in multiple ways in a retail environment. Data can be collected to the pinpointed granularity of each individual transaction via a cashier’s register or to a broader granularity of weekly store performance via end-of-day reports. Transactional data, also known as point-of-sale (POS) data, is generated for every purchase and can record both information on the customer and the item. Examples of customer data can include their gender, age, payment type, loyalty membership, and if coupons were used. Examples of the item data can include quantity, discount, product type/department, and venue of purchase (online versus in-person). Store performance data is usually generated on a wider range of time compared to POS data – day or week, usually. Examples of performance data can include total sales, if markdowns were applied, number of transactions, and external details that may influence performance. It is easy to see that the volume of data potentially generated in a retail business has the potential to become immense.

SkyMart’s performance evaluation is essentially asking, “Can we accurately predict SkyMart’s future sales using the data on hand?”. Investigating the data SkyMart has compiled to answer this main question also allows me to develop insights into secondary performance metrics such as 1) ranking store’s sales performance, 2) deciphering any seasonal sales fluctuations to facilitate improved inventory preparedness, and 3) determining if markdown promotions or other external factors (fuel price, unemployment rate, etc.) have any effect on overall sales.

**Data Description**

The data SkyMart has amassed consists of two separate datasets, both with unique features, and were sourced from Kaggle.com’s online repository. The first dataset spans a timeframe from 2010 to 2012 and is comprised of variables such as store number, department number, weekly date, weekly sales amounts, and if the week contained a holiday or not. The second dataset spans a timeframe from 2010 to 2013 and is comprised of variables such as store number, weekly date, and if the week contained a holiday or not – just like the first dataset. The second dataset also includes various external variables such as each store’s weekly markdown amount (if any), the weekly consumer price index (CPI), the average weekly temperature, the average weekly fuel price, and unemployment rate for the week.

An initial observation of the two datasets is that they possess different levels of granularity, department-level versus store-level, which needs to be recorded to guard against duplicative records for when they are combined. Another observation is the two datasets span two different lengths of timeframes. This needs to be recorded to guard against record deletion when joining datasets.

**Methodology**

The procedure to develop a future sales predictive model for SkyMart first started with an extensive data preprocessing stage. This in-depth step prepared the data for optimal use in the next stage of exploratory data analysis (EDA). It is here where the secondary insights, such as store performance, seasonality of sales, and effect of external factors were answered. Following that deep dive, the final stage of data modeling was performed with hyperparameter tuning and recursive feature elimination (RFE) to provide the best model for future use.

**Data Preprocessing**

Upon loading each dataset, I discovered some features were incorrectly formatted and I changed their datatype to more useable types. “Date” was originally an object datatype and was adjusted to a datetime datatype. “IsHoliday” was a feature depicting if a holiday was contained in the week or not and it was changed from a Boolean datatype to a binary 0/1 via encoding. I also discovered that there was a need to combine the five “Markdown” features, as there was no documentation as to why they were split, as well as impute zeros into their missing values to reflect no markdowns were given that week. Lastly, “Date” was parsed into year, month, and week number for best use in modeling.

The two datasets were then joined on the foreign keys of “Store”, “Date”, and “IsHoliday” to incorporate the target variable, “Weekly\_Sales”, with the other predicting features. This produced a dataset that was at the department level of granularity and was saved as “Department” for future use investigating the secondary insight requests. A dataset at the store level of granularity was then produced by aggregating “Weekly\_Sales” of all departments per store per date and saved as “Store” for future use in predicting future sales. With these cleansed datasets, it was possible to move forward into the next step of investigating SkyMart’s performance.

**Exploratory Data Analysis (EDA)**

The first task I dove into was the univariate analysis and the layouts of each feature. I utilized the SweetViz library to efficiently do this, which returned descriptive statistics on all features that were used to determine any concerning flags, such as skewness or lingering null values. The “Unemployment” and “Markdown” features were the two that produced the largest positively-skewed results with values of 1.18 and 3.89, respectively. Features with values above +1 warrant a transformation to provide a more normal distribution (Vadali, S., 2017). Furthermore, in anticipation of predicting the continuous variable of “Weekly\_Sales” with linear regressions, it is important to ensure features are well normalized in distribution beforehand. For these two reasons, both “Unemployment” and “Markdown” features were transformed and attained skewness values of 0.1 and 0.61, respectively. The original features were then eliminated from the dataset.

Now that the individual features were set up for best use, I dove into a bivariate analysis of how each related to each other and how they related to the target variable, “Weekly\_Sales”. A correlation matrix was developed to quantify each feature’s relationship to one another. It was evident that the “Weekly\_Sales” variable displayed the most correlation with “Store” and “Markdown”, which was noted for future modeling. It was also evident that some multicollinearity existed between “Date” and both “Fuel\_Price” and “Markdown”. This made sense as both increased over time and were noted for future modeling.

While in the EDA phase, I was also able to investigate answers to secondary requests from SkyMart. The first request was to rank each store’s sales performance from 2010 through 2012. I chose to display each store's weekly sales amounts over the 3 year period as a box and whisker plot (Figure 1). This helps display the overall distribution of their sales figures. Setting a success criteria of weekly sales at $600,000, it’s easy to see some stores perform better than others. This knowledge can be used to develop pinpointed plans of improving sales in stores 3, 5, 7, 9, 15-16, 29-30, 33, 36-38, and 42-44. If we knew the geographical location of each store, which is not contained in the datasets of this project, we could look into spatial relationships of over-performing stores versus under-performing stores. It is a good takeaway to note that 66% of stores are experiencing weekly sales above the $600,000 mark.

Figure 1. Store weekly sales 2010 - 2012

A graph of blue and white candlesticks

AI-generated content may be incorrect.

Another request from SkyMart was to determine any seasonality effect to store sales to assist in inventory preparedness for times of high sales. I performed a seasonal decomposition of the dataset to answer this question as it produces insights into general sales trends and seasonality. As we can see in Figure 2., the company’s sale’s trend is growing over time with a steady increase shown in the graph. We can also see that there are periods of repetitive sales spikes – both positive and negative. Sales jump up in the later months of the year, November and December, depicting both Black Friday and Christmas shopping. Inventory readiness efforts need to be in swing before mid-November to properly prepare for these times of higher sales figures. We also see a slight, repetitive slump in sales after the first of the year. This depicts the shopper’s lull in purchasing following the high purchasing time of Christmas. Coordinated inventory shifts during this period could be investigated to ensure the product on shelf is put where it’s best sold.

Figure 2. Sales trend and seasonality, 2010 - 2012

A graph of two people

AI-generated content may be incorrect.

**Data Modeling**

Now we have got to the main request of SkyMart – Can we develop a predictive model to forecast future sales? “Weekly\_Sales” is a continuous variable that we want to target in modeling, and, because of this, linear regression models are best to use. With each regression model investigated, I trained the model on 70% of the original dataset and tested on 30% of the original dataset and scaled the features since they were on vastly different scales. Each model was evaluated based on their coefficient of determination (R2 value), root mean squared error (RMSE), and residual plots. The R2 value tells us how well the model’s predictions match the actual data and a value closer to 1 is desirable. The RMSE tells us how far off the model’s predictions are from the real value, on average, in the scale of the target variable. A lower RMSE is desirable. The residual plots show you if the model is making predictive mistakes randomly (desirable) or systematically (undesirable).

I started investigating this request by implementing a multiple linear regression model targeting weekly sales and regressing on all cleaned features from the EDA. Due to subpar results combined with very large residual variability, the linearity of the features’ relationship to the target variable was called into question. Plotting the actual values versus the predicted values from the multiple linear regression displayed no linearity between the features and the target variable and made it clear that non-linear regressions needed to be investigated. The first non-linear model I implemented was a polynomial regression utilizing all features and I compared the same metrics for evaluation. As this model didn’t move the needle very much from the previous, it was clear that I needed to investigate more complex non-linear regression models.

The second model I implemented was a decision tree in which I performed a hyperparameter grid search to provide the best model specifics, performed recursive feature elimination (RFE) to provide the most predictive features, and then refit that best model on the reduced feature set. The hyperparameters I performed a grid search on were maximum tree depth (None, 10, 20), minimum samples split (2, 5, 10), and minimum samples leaf (1, 2, 4). The best parameters returned were None, 2, and 10, respectively. The RFE resulted in using 4 out of the 10 features (Store Number, CPI, Week, and the log transformation of Unemployment) and the model produced promising R2 and RMSE values. It is also wise note the computing time of running these more complex models as they can become quite cumbersome at times.

The third model I implemented was a random forest. I performed hyperparameter tuning in the same manner as a decision tree with the addition of number of estimators (50, 100, 200). The best parameters returned were no maximum tree depth, 1 minimum sample leaf, 5 minimum samples split, and 200 number of estimators. The same four features were selected as in the decision tree and its R2, RMSE values, and time to compute were captured. The final model I implemented was a gradient boosting model, specifically XGBoost. The hyperparameters tuned were number of estimators (50, 100, 200), maximum tree depth (3, 6, 10), learning rate (0.01, 0.1, 0.3), subsample (0.7, 1.0), and colsample\_bytree (0.7, 1.0). The best parameters returned were 200, 6, 0.1, 1.0, and 0.7, respectively. This model chose all ten features to include and its evaluation metrics were the same as the others.

**Results**

The multiple linear regression yielded a R2 value of 0.168 and an RMSE of $515,734, which are lackluster results and signified a far too simple model. Furthermore, this model’s residual plot displayed a very large degree of variability (±$2.5M) which points to it not identifying the relationships between features and the target well. I then pivoted to the polynomial regression due to the lack of linearity of the data and this model improved the R2 to 0.385 and reduced the RMSE to $442,143, which is still an unacceptable outcome when the average store’s weekly sales are around $1M. The residual plot exhibited a slightly reduced amount of variability but still displayed large variability (+$1M to -$2M) and a fanning out pattern when predicting higher values (Figure 3). This exercise told me that I would need to investigate more complex regression models as well as utilizing a grid search to find the best hyperparameters of each model and recursive feature elimination to eliminate unnecessary features.

Figure 3. Polynomial residual plot

A graph showing a number of values

AI-generated content may be incorrect.

The decision tree regression model produced far better metrics than the polynomial regression. The training versus testing R2 values were 0.983 versus 0.956 and the RMSE values were $74,147 versus $118,851. These values suggest a slight overfitting on the training set but the model still provides a solid generalization to new data as the testing R2 value is still high. Its residual plot (Figure 4) was much better and depicted more centralization of predicted values, but still with a fanning out with higher value predictions ±$1M. The computing time for this decision tree model only took 4.22 seconds to complete.

Figure 4. Decision tree regression residual plot

A graph showing a line of blue dots

AI-generated content may be incorrect.

The random forest regression model took more time to run than all three of the previous models combined – about 8.5 MINUTES. The model produced training versus testing R2 values of 0.99 versus 0.96 and the RMSE values were $56,576 versus $112,447. The R2 values were improved metrics to the decision tree, though very slight, but the RMSE values were worse with the training set roughly double the error in predictions. The residual plot showed much the same as the decision tree regression with the same degree of fanning out of the higher predicted values (Figure 5).

Figure 5. Random forest regression residual plot

A graph showing a line of blue dots

AI-generated content may be incorrect.

Lastly, the gradient boosting model with the XGBoost Regressor took more time to run than the decision tree but far less than the random forest – about 1.5 minutes. This model’s training versus testing R2 values were 0.995 vs 0.975 and the RMSE values were $41,019 versus $88,829. These were the lowest RMSE values seen during the modeling phase. The residual plot (Figure 6) is also improved with even more centralization of predictions and the fanning out is reduced to +$600K to -$400K.

Figure 6. XBGoost regression residual plot

A screen shot of a graph

AI-generated content may be incorrect.

**Conclusion**

Through the evaluation of SkyMart’s data, we have developed valuable insights that can make an impact on the business. We have pinpointed certain times of the year, particularly prior to Black Friday and Christmas holidays, where inventory needs to be prepared to provide the stores with the most ability to make sales. We have also looked within the company to highlight top-performing stores that could potentially get rewarded for their high performance and pinpoint low-performing stores that should be further investigated to best help them perform better.

Ultimately, we have investigated four predictive models that best suit SkyMart’s data to deliver the best model to predict future sales. Upon examining the performance of each model side by side (Table 1.) and their corresponding residual plots (Figure 7.), it’s recommended to implement the XGBoost regression model for future sales predictions. The XGBoost regression model provided the highest R2 value explaining the largest degree of variance in the target variable while also delivering the lowest RMSE explaining the lowest risk of prediction error. Furthermore, the XGBoost model provides the least amount of variability in its predictions while also doing so without a cumbersome delay in computation.

Table 1. Model comparison of evaluation metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Train R2 | Test R2 | Train RMSE | Test RMSE | Time to Run (sec) |
| Polynomial | 0.385 | 0.388 | $442,143 | $442,237 | 0.6 |
| Decision Tree | 0.983 | 0.956 | $74,147 | $118,851 | 4.2 |
| Random Forest | 0.990 | 0.960 | $56,576 | $112,447 | 598.4 |
| XGBoost | 0.995 | 0.975 | $41,019 | $88,829 | 119.2 |

Figure 7. Model comparison of residual plots

A collage of graphs showing different types of data

AI-generated content may be incorrect.

To operationalize the XGBoost regression model into SkyMart’s business, it would be integrated into the store’s analytics pipeline and updated with the most recent sales data on a weekly basis. This would ensure the model remains accurate by adapting to evolving consumer behavior and seasonal trends that exist. The output of the model would be visualized in store dashboards, allowing managers to easily interpret sales forecasts at the store level. These forecasts could support inventory planning, promotional timing, and even staffing decisions. Lastly, automated alerts could be developed to flag significant deviations from predicted trends and enable faster response to unexpected changes.

**Future Work**

One key assumption of the predictive model I developed is that historical sales patterns and promotional impacts will remain relatively consistent over time. However, this may not hold true during unusual events such as supply chain disruptions, altering foreign policy, economic downturns, and shifts in consumer preferences and need to be considered as they arise. Also, my model makes predictions on the whole dataset and, as time proceeds, this will impact its computational tax. A note for future work should entail developing lag sales features to predict on a more recent subset of the data – perhaps on a quarterly, bi-annual, or annual basis.

A limitation in my predictive model is that it predicts at a store level, which may obscure important patterns at a finer granularity such as product or department levels. A deeper dive into those granularities would greatly benefit SkyMart in a noticeably different way as it could pinpoint very specific actionable items. It is also important to note that XGBoost handles non-linearity well but it can be sensitive to data imbalance, such as the infrequent holidays we have in our dataset. This may be the cause of some of the high variability in the larger prediction errors of the model and should be investigated.

**References**

8 Retail Analytics Benefits You Should Know About. ThoughtSpot. (2024, February 2). <https://www.thoughtspot.com/solutions/retail-analytics/retail-analytics-benefits>

Vadali, S. (2017, December 29). *Day 8: Data transformations – Skewness, normalization and much more*. Medium. <https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55>