MGT 6203 Team 78 Final Report

Project Title: Predicting New Product Sales

Team Members: Joel Thiessen, Megha Maheshwari, Sharath Nataraj, Hannah Sailar

Executive Summary

Sales forecasting is a crucial activity for success in any consumer goods corporation. Most business decisions concerning production are made on top of analytical tools used to forecast product demand. Forecasting demand for existing products based on historical sales data is widely exercised, however, when past sales data is unavailable, more creative approaches are needed. Here we propose a methodology to predict sales data on a brand-new product in the marketplace. A series of regression models were trained on product data that was split between seasonal and non-seasonal products. XGBoost performed exceptionally well with non-seasonal data while an ensemble model consisting of XGBoost, Random Forest Regressor, and Extra Trees Regressor performed best on seasonal data. It was found that the XGBoost model improved upon the accuracy of current methods exercised in the industry for non-seasonal data and there is room for improvement in the future with seasonal data prediction.

Introduction

Target Corp.'s earnings and stock have suffered in recent months due to waves of excess inventory and low inventory in the distribution centers. Some of the inventory issues are due to logistical challenges, but for the most part, they are due to incorrect forecasting of sales demand. While it is possible to forecast the sales of standard existing items (like cereal, chocolates, etc.), it is a huge challenge to forecast the demand for new items and items with a faster turnaround (like apparel, soft home decor, bedding, etc.). The buying division at Target places orders for these items based on experience and best guesses what the demand will look like when the products hit the shelf. Demand planning for new products in the consumer products industry lacks analytical tools, causing a hit or a miss to the target.

This research carries a potential impact to not only improve business revenue, but to also cut down on waste from unsold products. Target's profits fell by 90% in Q2 2022 compared to the previous year, they were forced to resort to heavy discounts in order to sell off excess inventory. This is an example of poor demand prediction in a major retailer that can have huge impacts on revenue and stock prices. The ability to predict demand more accurately for a new product can balance over production, leaving too many units on the shelf, with under production where extra revenue could have been captured.

In this research project, we take a deeper look at predicting the demand for new products that do not have any sale history data for reference. The aim is to maintain an optimum inventory level at Target's distribution centers and stores, preventing excess or inventory deficits which in turn will maximize sales and cut losses for the company.

The research questions used to frame this project are as follows: how can we build a model to best predict demand for a brand-new product on the market? What product characteristics play the most important role in predicting demand? What outside data points of the social environment play a role in predicting demand for a new product?

A series of analytical models were built using Target Corp. data to aid in answering these questions. After the data was cleaned and new features were created, we were left with numerous variables used to describe a product. This high dimensionality makes modeling difficult. To address this problem, we utilized feature selection and dimensionality reduction techniques. After obtaining a more reasonable number of features to work with, regression models, such as multiple linear regression, decision tree regressor, random forest, or XGBoost regressor, were used to fit the data. This regression model used historical sales data from the product launch to predict sales of a new product.

We hypothesize this approach to be much more effective than the current industry standard of educated guesses based on experience.

Literature Survey

Our literature survey focuses on two studies, one that investigates the performance of various clustering algorithms and the other that explores common forecasting methods used in the fashion industry.

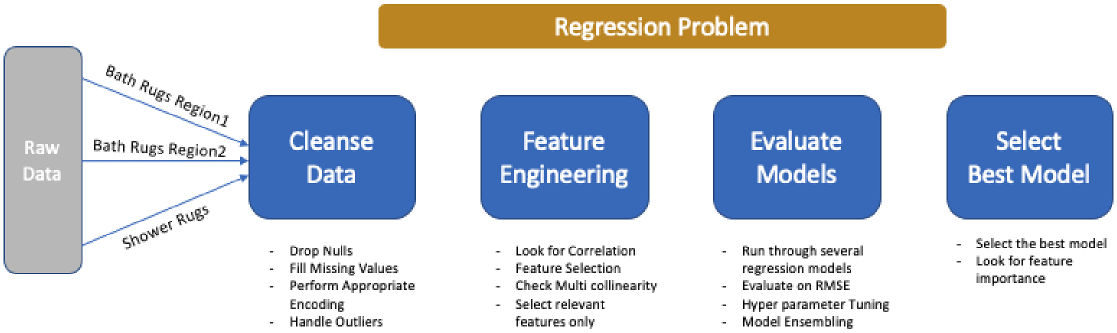
Hammouda (2000) looks at the comparative performances of several data clustering models including K-means, Fuzzy C-means, Mountain, and Subtractive clustering to diagnose heart disease. Of these four clustering models, K-means performed the best as well as the fastest while Subtractive performed the worst and Mountain was the slowest. The performance of K-means was a marginal improvement above the other algorithms but is worthwhile to consider over all others because of the reduced computation time. Hammouda (2000) states that Mountain clustering is only appropriate for low dimensional problems. In subtractive clustering, the programmer needs to choose the neighborhood radius value carefully because too small or too large of a radius will result in not including important data point or including too many data points to the point that clustering becomes irrelevant. They conclude that the reason none of the algorithms performed highly is because the data has very high dimensionality, too many variables tend to reduce clustering accuracy from complicated relationships amongst variables.

Beheshti-Kashi (2015) looks at state-of-the-art sales forecasting methods at the time through the lens of fashion and new consumer products. Forecasting methods were grouped into two camps: traditional methods such as exponential smoothing, Holt-Winters, and Box & Jenkins models, and new machine learning methods such as neural networks. It was found in this study that a hybrid approach to forecasting models produced the most accurate predictive results. As an analysis to this finding, Beheshti-Kashi (2014) states that although complex models may produce precise outcomes, real world fashion companies struggle to adopt these models in daily work.

Methodology

The below workflow illustrates the high-level data science workflow/methodology adopted for the project. We are considering this a multi-variate regression problem. The same workflow was used for Non-Seasonal and Seasonal datasets.

Figure 1: High-level workflow/methodology



Data Overview

Our data sets are provided by a supplier to Target and Walmart. The data focuses on products from the “soft home” category and includes features such as color, design, price, sales etc. over a four-year period.

There are three complex datasets used in this analysis. Each contains information of a particular type of product (bath rugs and shower rugs) that are sold in Target and Walmart stores in a designated region of the United States. The features contained in the datasets include department number, class, item description, item launch date, size, color, intended selling channel, brand, region, sales units, promotional sales units, sales in dollars, in stock percentage, and more.

These datasets are aggregated to create a diverse training dataset for modeling purposes. New, useful variables will be created for better prediction of the unit sales price. Class, color, seasonality, and sales price are expected to be significant factors.

Table 1: Description of raw datasets collected

Dataset	Description	Rows	Columns

Dataset64_2018_2022_Region1.csv	Bath rugs sales data for a particular region	122,289	25
Dataset64_2018_2022_Region2.csv	Bath rugs sales data for a particular region	119,766	26
Dataset64_2018_2022_Shower_Region.csv	Shower rugs sales data for a particular region	58,870	24

Data Cleaning & Feature Engineering

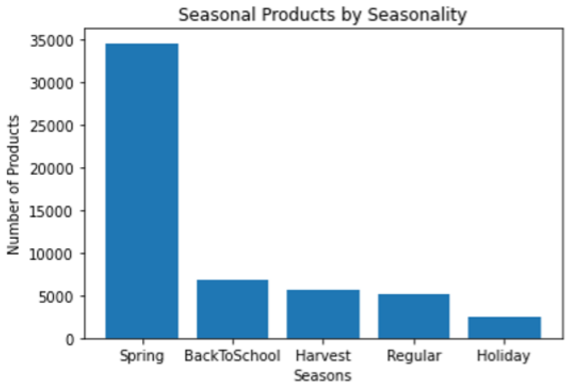
Data cleaning for four years of merged data began with data removal. A couple columns that had a very high rate of 'NaNs' were deleted because it is difficult to extrapolate that many missing values with any useful insight. Outliers were identified and removed, carefully identifying outliers will be a particularly important step in the original dataset because most of the year 2020 could be considered an outlier compared to other years, but we do not want to eliminate too much data. Clearance items which are to be discontinued from the product line were removed because we are only concerned with new products on the market

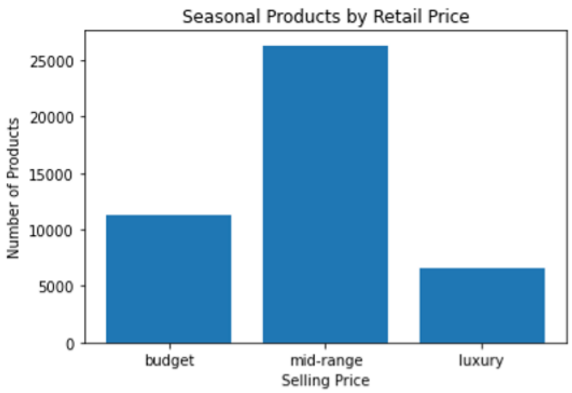
The next step in the data cleaning process involved adding data back into the data frames and adjusting column values and datatypes for ease of use. Columns with a few missing data points, like color, were filled with the column mean value. The numeric sales columns data types were converted to a more usable format and the categorical columns were label encoded for model use. The date columns were cleaned up to remove hourly and daily information because data was only collected on a weekly basis. Sales units were adjusted to account for out-of-stock items and clearance items.

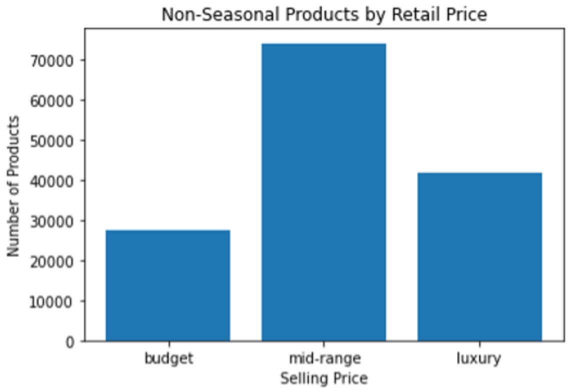
Finally, new columns were created from cleaned data. A new column was created for per unit sales price from the total sales number. Each product was assigned a pattern group based on the look of the material. A start date and end data column were created from the date column to indicate when a product hit the shelves and when it first went on clearance. This helped indicate when peak selling season for a product occurs, a seasonality factor was created based on this information. A product was marked as seasonal or non-seasonal, seasonal products given a categorical value as to when peak sales occurred (spring, back to school, harvest, holiday). An additional column was created based on the selling price of the product to indicate if it is a budget, mid-range, or luxury item.

The cleaned data was divided into two final data sets for modeling, one for seasonal products and another for non-seasonal products. Below, some visual information is shown on how the data is divided up amongst seasonal and non-seasonal data.

Figures 2-4: Seasonal Products by Seasonality, Seasonal Products by Retail Price, Non-Seasonal Product by Retail Price







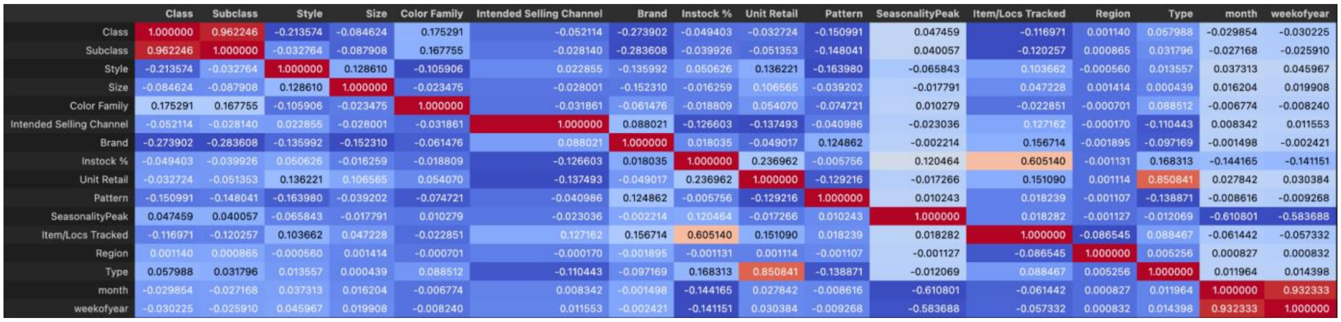
There is a high selling season each spring. We hypothesize this is due to tax refunds that are administered in the spring and as improving weather encourages people to leave their homes and go out to stores.

Feature Selection

After the data has been cleaned and new variables created, we are left with numerous features to describe each product. High dimensionality creates some statistical challenges when it comes to modeling such as overfitting. To address this, feature selection and dimensionality reduction techniques are introduced to the data.

A correlation matrix was used to visually inspect for highly correlated variables. Features with an absolute mutual correlation value greater than 0.4 were removed from the dataset to avoid multicollinearity. Only a single time related variable was selected because all others were highly correlated with each other. Item Description and DPCI were both eliminated because they are arbitrary and won’t result in good predictors. Class and subclass are highly correlated, subclass was selected.

Figure 5: Feature Correlation Matrix for Non-Seasonal Data

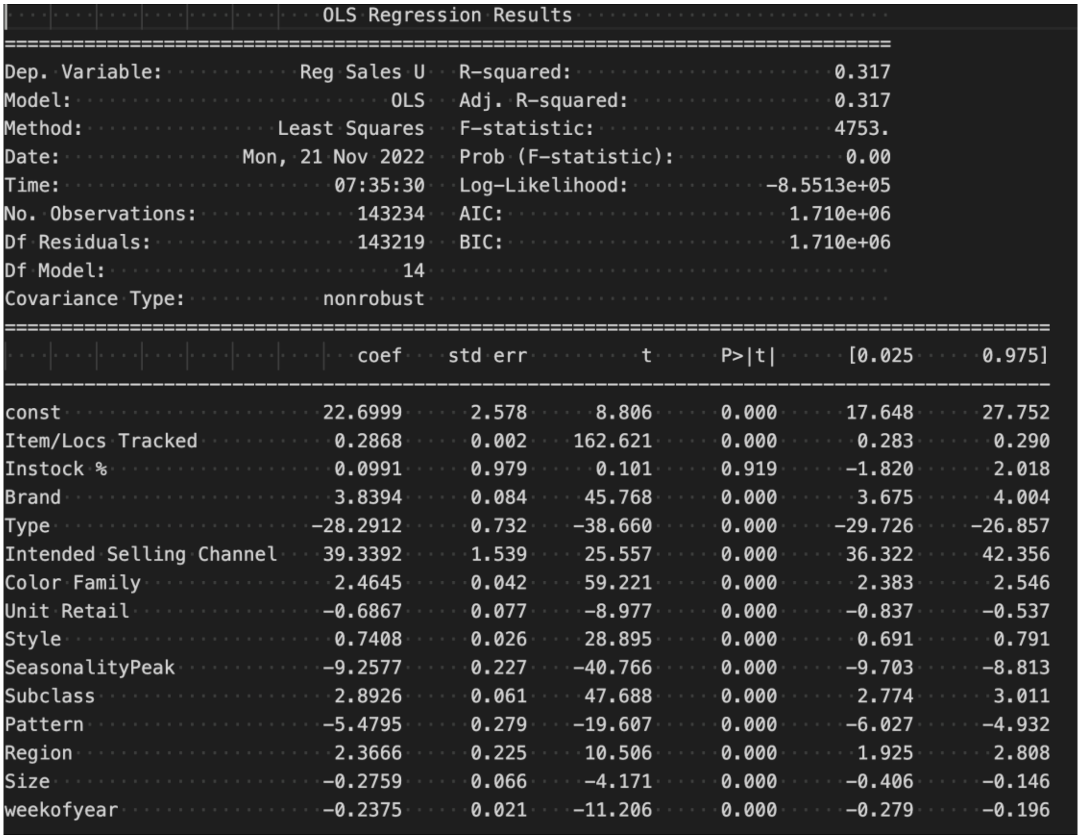


	Class	Subclass	Style	Size	Color Family	Intended Selling Channel	Brand	Instock %	Unit Retail	Pattern	SeasonalityPeak	Item/Locs Tracked	Region	Type	month	weekofyear
Class	1.000000	0.962246	-0.213574	-0.084624	0.175291	-0.052114	-0.273902	-0.049403	-0.032724	-0.150991	0.047459	-0.116971	0.001140	0.057988	-0.029854	-0.030225
Subclass	0.962246	1.000000	-0.032764	-0.087908	0.167755	-0.028140	-0.283608	-0.039926	-0.051353	-0.148041	0.040057	-0.120257	0.000865	0.031796	-0.027168	-0.025910
Style	-0.213574	-0.032764	1.000000	0.128610	-0.105906	0.022855	-0.135992	0.050626	0.136221	-0.163980	-0.065843	0.103662	-0.000560	0.013557	0.037313	0.045967
Size	-0.084624	-0.087908	0.128610	1.000000	-0.023475	-0.028001	-0.152310	-0.016259	0.106565	-0.039202	-0.017791	0.047228	0.001414	0.000439	0.016204	0.019908
Color Family	0.175291	0.167755	-0.105906	-0.023475	1.000000	-0.031861	-0.061476	-0.018809	0.054070	-0.074721	0.010279	-0.022851	-0.000701	0.088512	-0.006774	-0.008240
Intended Selling Channel	-0.052114	-0.028140	0.022855	-0.028001	-0.031861	1.000000	0.088021	-0.126603	-0.137493	-0.040986	-0.023036	0.127162	-0.000170	-0.110443	0.008342	0.011553
Brand	-0.273902	-0.283608	-0.135992	-0.152310	-0.061476	0.088021	1.000000	0.018035	-0.049017	0.124862	-0.002214	0.156714	-0.001895	-0.097169	-0.001498	-0.002421
Instock %	-0.049403	-0.039926	0.050626	-0.016259	-0.018809	-0.126603	0.018035	1.000000	0.236962	-0.005756	0.120464	0.605140	-0.001131	0.168313	-0.144165	-0.141151
Unit Retail	-0.032724	-0.051353	0.136221	0.106565	0.054070	-0.137493	-0.049017	0.236962	1.000000	-0.129216	-0.017266	0.151090	0.001114	0.850841	0.027842	0.030384
Pattern	-0.150991	-0.148041	-0.163980	-0.039202	-0.074721	-0.040986	0.124862	-0.005756	-0.129216	1.000000	0.010243	0.018239	-0.001107	-0.138871	-0.008616	-0.009268
SeasonalityPeak	0.047459	0.040057	-0.065843	-0.017791	0.010279	-0.023036	-0.002214	0.120464	-0.017266	0.010243	1.000000	0.018282	-0.001127	-0.012069	-0.610801	-0.583688
Item/Locs Tracked	-0.116971	-0.120257	0.103662	0.047228	-0.022851	0.127162	0.156714	0.605140	0.151090	0.018239	0.018282	1.000000	-0.086545	0.088467	-0.061442	-0.057332
Region	0.001140	0.000865	-0.000560	0.001414	-0.000701	-0.000170	-0.001895	-0.001131	0.001114	-0.001107	-0.001127	-0.086545	1.000000	0.005256	0.000827	0.000832
Type	0.057988	0.031796	0.013557	0.000439	0.088512	-0.110443	-0.097169	0.168313	0.850841	-0.138871	-0.012069	0.088467	0.005256	1.000000	0.011964	0.014398
month	-0.029854	-0.027168	0.037313	0.016204	-0.006774	0.008342	-0.001498	-0.144165	0.027842	-0.008616	-0.610801	-0.061442	0.000827	0.011964	1.000000	0.932333
weekofyear	-0.030225	-0.025910	0.045967	0.019908	-0.008240	0.011553	-0.002421	-0.141151	0.030384	-0.009268	-0.583688	-0.057332	0.000832	0.014398	0.932333	1.000000

To further reduce dimensionality, univariate feature selection was used to determine the top 13 variables. This method determines the relationship strength between each feature and the model target variable. We made use of Scikit-Learn's SelectKBest method to score the contribution of each feature. The highest 13 scores were selected for the final dataset, which include Style, Color Family, Unit Retail, Brand, Subclass, Intended Selling Channel, Month, Size, Pattern, Seasonality Peak, Cluster Type, Type.

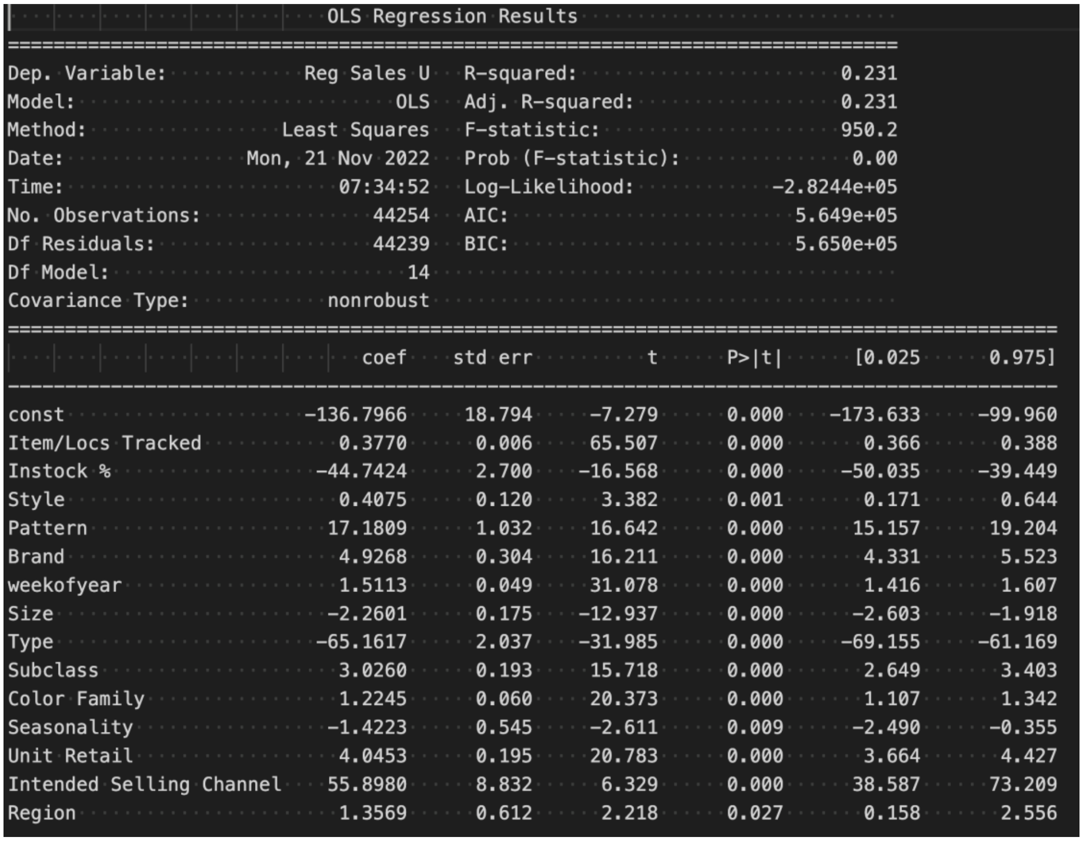
Below are the results of the multiple linear regression model run on Non-Seasonal and Seasonal Data.

Figure 6: MLR on Final Features Selected for Non-Seasonal



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Reg Sales U   R-squared:                       0.317
Model:                            OLS   Adj. R-squared:                  0.317
Method:                 Least Squares   F-statistic:                     4753.
Date:                Mon, 21 Nov 2022   Prob (F-statistic):               0.00
Time:                        07:35:30   Log-Likelihood:            -8.5513e+05
No. Observations:              143234   AIC:                         1.710e+06
Df Residuals:                  143219   BIC:                         1.710e+06
Df Model:                          14
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       22.6999      2.578      8.806      0.000      17.648      27.752
Item/Locs Tracked            0.2868      0.002    162.621      0.000       0.283       0.290
Instock %                    0.0991      0.979      0.101      0.919      -1.820       2.018
Brand                        3.8394      0.084     45.768      0.000       3.675       4.004
Type                       -28.2912      0.732    -38.660      0.000     -29.726     -26.857
Intended Selling Channel    39.3392      1.539     25.557      0.000      36.322      42.356
Color Family                 2.4645      0.042     59.221      0.000       2.383       2.546
Unit Retail                 -0.6867      0.077     -8.977      0.000      -0.837      -0.537
Style                        0.7408      0.026     28.895      0.000       0.691       0.791
SeasonalityPeak             -9.2577      0.227    -40.766      0.000      -9.703      -8.813
Subclass                     2.8926      0.061     47.688      0.000       2.774       3.011
Pattern                     -5.4795      0.279    -19.607      0.000      -6.027      -4.932
Region                       2.3666      0.225     10.506      0.000       1.925       2.808
Size                        -0.2759      0.066     -4.171      0.000      -0.406      -0.146
weekofyear                  -0.2375      0.021    -11.206      0.000      -0.279      -0.196
```

Figure 7: MLR on Final Features Selected for Seasonal



```
                        OLS Regression Results
==============================================================================
Dep. Variable:            Reg Sales U   R-squared:                       0.231
Model:                            OLS   Adj. R-squared:                  0.231
Method:                 Least Squares   F-statistic:                     950.2
Date:                Mon, 21 Nov 2022   Prob (F-statistic):               0.00
Time:                        07:34:52   Log-Likelihood:            -2.8244e+05
No. Observations:               44254   AIC:                         5.649e+05
Df Residuals:                   44239   BIC:                         5.650e+05
Df Model:                          14
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     -136.7966     18.794     -7.279      0.000    -173.633     -99.960
Item/Locs Tracked            0.3770      0.006     65.507      0.000       0.366       0.388
Instock %                  -44.7424      2.700    -16.568      0.000     -50.035     -39.449
Style                        0.4075      0.120      3.382      0.001       0.171       0.644
Pattern                     17.1809      1.032     16.642      0.000      15.157      19.204
Brand                        4.9268      0.304     16.211      0.000       4.331       5.523
weekofyear                   1.5113      0.049     31.078      0.000       1.416       1.607
Size                        -2.2601      0.175    -12.937      0.000      -2.603      -1.918
Type                       -65.1617      2.037    -31.985      0.000     -69.155     -61.169
Subclass                     3.0260      0.193     15.718      0.000       2.649       3.403
Color Family                 1.2245      0.060     20.373      0.000       1.107       1.342
Seasonality                 -1.4223      0.545     -2.611      0.009      -2.490      -0.355
Unit Retail                  4.0453      0.195     20.783      0.000       3.664       4.427
Intended Selling Channel    55.8980      8.832      6.329      0.000      38.587      73.209
Region                       1.3569      0.612      2.218      0.027       0.158       2.556
```

The Adjusted R-Squared for non-seasonal stood at .317 while that of Seasonal stood at .231.

Modeling Selection & Performance

In order to run models and test how well they perform, several models were tried on the training set. The models were evaluated using 10-Fold Cross-validation. A series of regression models were run against the seasonal and non-seasonal data to predict product sales volumes. The 10 models evaluated are below,

CART, K Nearest Neighbors, Random Forest, XGBoost, LightGBM, Bagging, Extra Trees Regressor, Linear Regression, LASSO, and Ridge Regression.

The models were cross validated for Root Mean Squared Error (RMSE), the error means and error standard deviations were compared to choose which yield the best results without overfitting to the data. The superior model will be used to predict a new launch in 2022 and will be verified against actual sales data using RMSE to evaluate model performance.

RMSE is described as $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\overline{y_i} - y_i)^2}{n}}$, this measures the difference between the predicted and actual values, squares to make all values positive, and averages across all predicted values.

Figure 8: Comparison of Various Model RMSE for Non-Seasonal Data

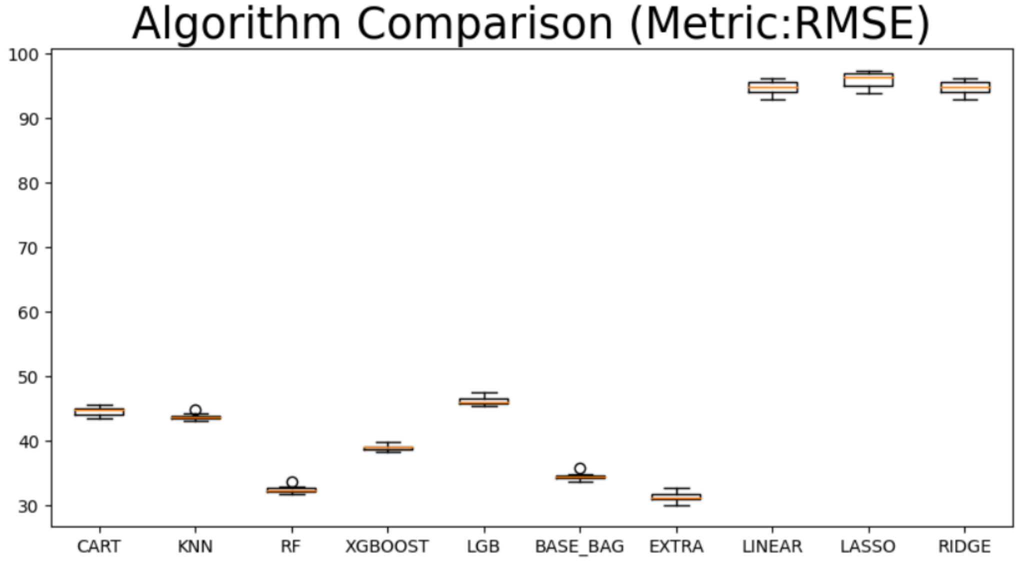
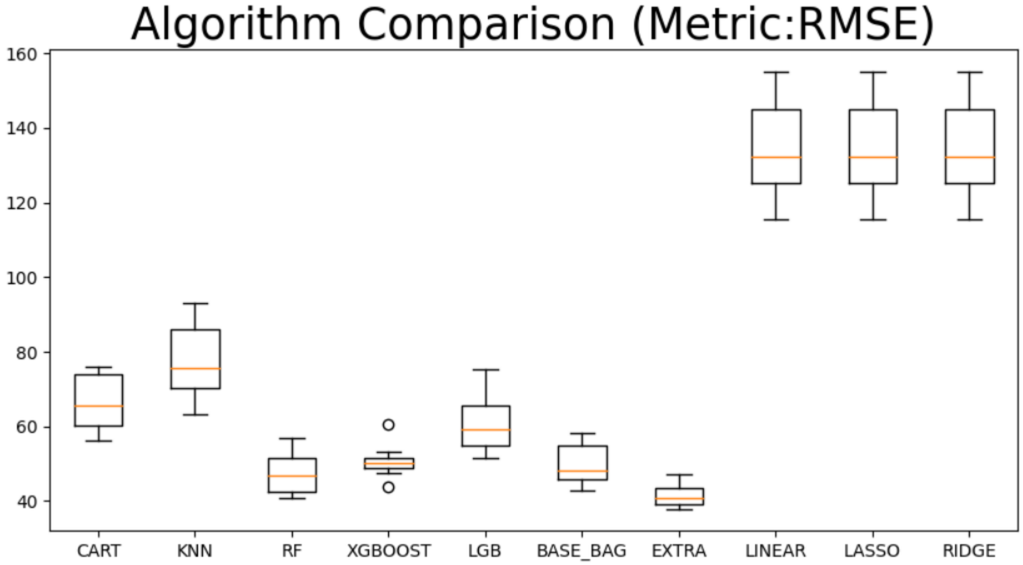


Figure 9: Comparison of Various Model RMSE for Seasonal Data



In the figure above we can see the RSME of each model when run on the training data. LASSO, Ridge, and Linear Regression performed the worst by far for both seasonal and non-seasonal data. The rest of the models performed around the same ballpark for both seasonal and non-seasonal data with the top five being Extra Trees, Bagging, Random Forest, XGBoost, and LightGBM. However, the model performance was much better all-around for non-seasonal data as compared to seasonal data based on the RSME scale on the y-axis.

The gridsearch function from the Sklearn library was used to tune the model's hyperparameters including tree depth, learning rate, number of estimators, and column sample by tree.

Final Model

The best model was evaluated as the one which gives lowest RMSE and at the same time doesn't overfit. As explained above, the validation set is a carefully curated dataset with brand new sales that is not present in the training set.

Our team chose to move forward with an ensemble of the top three performing models to see if it gives better performance. Results from XGBoost, Random Forest Regressor, and Extra Trees Regressor were averaged to conclude with the final sales volume prediction.

For Non-Seasonal data, the RMSE from Ensemble model turned out to be 120.8. However, XGBoost turned out to be the best model for non-seasonal data with an RMSE of 110.5.

For Seasonal data, the ensemble model of XGBoost and ExtraTrees regressor turned out to be best with RMSE of 654.5.

Extensive hyper parameter tuning using grid search was done to arrive at the final models. In addition to RMSE performance, due importance was given to overfitting while selecting the final model. Figures 8 and 9 show much lower RMSE values while cross-validating, however, such performances were not obtained when testing in the final validation set.

Results

The final model for non-seasonal was XGBoost with RMSE of 110.5 and final model for seasonal was ensemble of XGBoost and ExtraTrees Regressor with RMSE of 654.5. These results corroborate with the Multiple linear regression done on Non-Seasonal and Seasonal shown in Figure 6 and 7 respectively. Non-seasonal had better Adjusted R-Squared and was expected to perform better than seasonal with respect to predictions and explainability.

The performance of non-seasonal was quite good considering the fact these brands have products that are brand new and don't have any historical sales. Please see some stats below. By comparison, it was seen that about 72% of the predicted had % change between –50% and +50% of actual sales. About 16% had %change higher than 50% and about 11% had less than –50%.

Non-Seasonal - Results		
% Change	**Count**	**Proportion**
>50%	85	16%
<-50% and >50%	383	73%
<-50%	*57*	11%

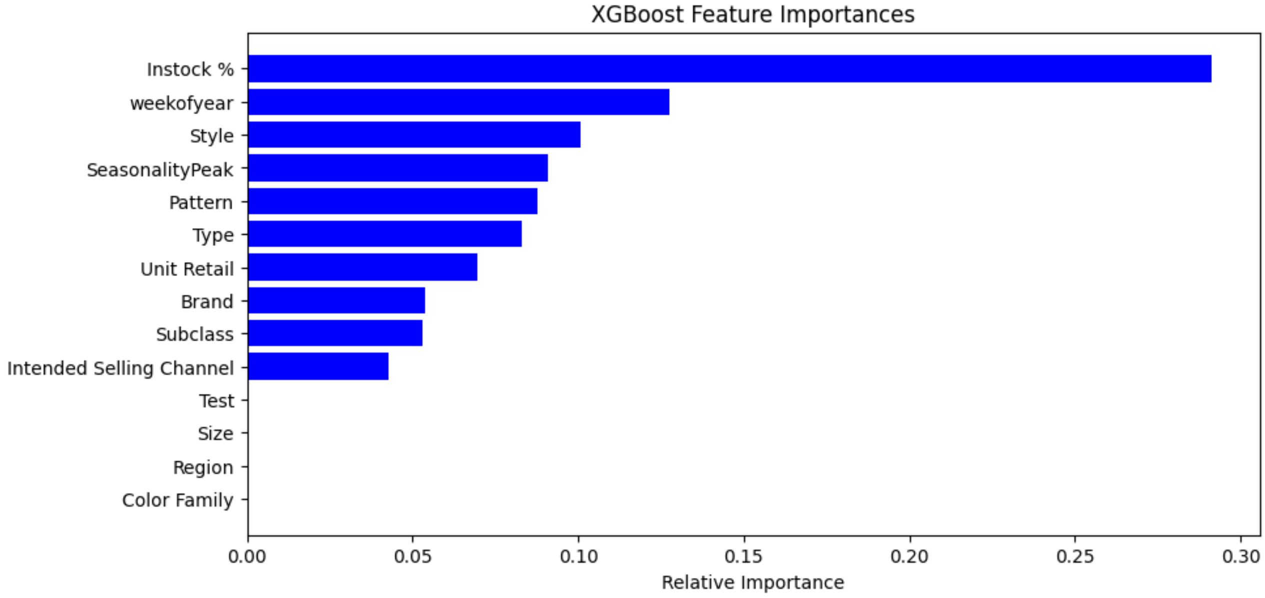
In stark contrast, the performance of Seasonal is not good. A large percentage ~ 55% of the predicted sales have variation of less than –50%. Only about 20% predicted sales fall between –50% and +50% of actual sales.

Seasonal - Results

% Change	**Count**	**Proportion**
>50%	70	25%
<-50% and >50%	55	20%
<-50%	*151*	55%

For non-seasonal, we were able to look at the feature importances. We can see that In Stock Percentage carries the most weight, this shows that maintaining inventory is very important to keep sales high. The next most important features include time of the year, style of product, seasonality, and so on.

Figure 10: Model Feature Importance for Non-Seasonal



The feature importance cannot be plotted for Seasonal because it is an Ensemble model.

Challenges

The non-seasonal data performed rather well for each model. However, the models did not perform well on the seasonal data in comparison. This could be due to the fact that the model needs more work to account for inconsistent sales throughout the year for these products.

Future Work

If given more time and resources, this analysis could be expanded in a few ways. First, we would like to address the deficit of model performance between seasonal and non-seasonal data. The seasonal dataset could benefit from additional analysis around the inconsistent sales throughout the year to contribute to better sales volume prediction.

In addition to splitting the data by seasonal and non-seasonal data, creating another dataset without data from the COVID-19 time period could provide a more generic model that would better fit new product launches outside of COVID-19 time periods.

Adding additional products and years of data to the training datasets will also strengthen the models and might pave way for more ensembling of models.

Interaction features and creation of new features is something that can be additionally looked at to strengthen the results.

Conclusions

Our approach improved on traditional methods currently used in the consumer product industry. The novelty of splitting the data by seasonal and non-seasonal products yielded a high-quality prediction for non-seasonal products and the potential for a very useful seasonal products model.

This analysis is worthwhile for any consumer product-based corporation to invest in because business areas across the entire corporation could be positively impacted. The buying team will change their order sizes to manufacturers, the marketing team will focus their advertisements on high selling products, and the financial department may prepare for better revenue statements as excess inventory and lost sales due to lack of inventory will be minimized. A successful model will also minimize the Bullwhip Effect, this is when small changes in demand are amplified in the downstream supply chain.

Works Cited

1. Hammouda, Khaled, and Fakhreddine Karray. "A comparative study of data clustering techniques." University of Waterloo, Ontario, Canada 1 (2000). https://www.researchgate.net/profile/Olga-Quintero/post/What_software_can_I_use_for_ANFIS_calibration/attachment/59d62d80c49f478072e9e8d1/AS%3A273562261229606%401442233734894/download/comparative+study+of+data+clustering+techniques.pdf
2. Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen & Michael Teucke (2015) A survey on retail sales forecasting and prediction in fashion markets, Systems Science & Control Engineering, 3:1, 154-161, DOI: 10.1080/21642583.2014.999389 https://www.tandfonline.com/doi/full/10.1080/21642583.2014.999389