CMPT 353 Project

Deep Dive into BrightWhite Smile

Anmol Bajaj: 301265661

Earl Cooke: 301306562

1st December, 2019.

1. Introduction

The e-commerce industry in Canada is growing at 21.3% per year. With the fast growth comes unpredictable challenges in scaling manufacturing, fulfillment and logistics operations within the industry. E-commerce retailers need to be able to predict demand months in advance as the manufacturing and delivery of goods to the seller's warehouse often has a lead time of up to three (3) months. The competition within the industry ensures that only the retailers with the most efficient operations are successful. Ordering too little inventory can lead to being out-of-stock, resulting in a loss of revenue. Whereas, ordering too much inventory is an inefficient use of capital that could have otherwise been used for marketing and expanding the brand. It would be extremely useful if BrightWhite Smile management had tools to help predict product demand. The report also aims to provide the company with historical seasonal trends and statistical analysis of their sales.

1.1 Describing the Dataset

The data was exported from the retailer website located at BrightWhiteSmilePro.com. The data contains the data for user behaviour on the website. The dataset has data from January 1st, 2017 to December 31st, 2018. The business primarily sells one (1) SKU, the BrightWhite Smile Teeth Whitening Kit. Each row represents the figures for each hour, resulting in 17,521 rows of data. The hourly data is from the PST time zone since that is where the business operates. The three CSV files in the datasets contain hourly records for conversions, sales, and visits.

1. **Visits**: Data related to visits to the website.
2. **Sales**: Data containing total number of orders and order value in each hour.
3. **Conversions**: Data containing the percentage of users that resulted in purchases, total number of "Add to Cart" on the website, and the total number of orders.

1.2 Data Cleaning Challenges

The primary challenge of data cleaning involved grouping and merging data together. The input data had hourly timestamps along the associated data. The data had to be grouped in daily, monthly, and yearly increments in order to be analyzed. The three different CSV files also needed to be merged into one dataframe before using machine learning algorithms.

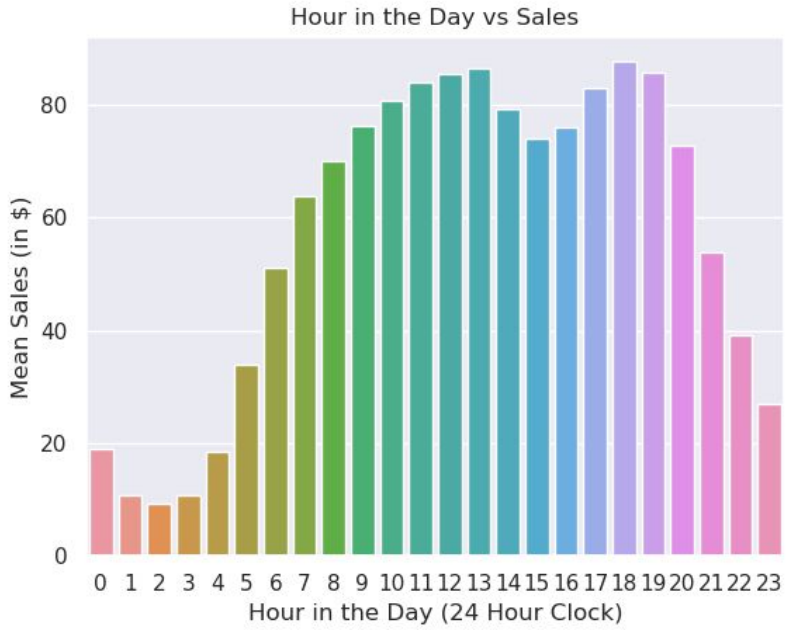
2. Preliminary Processing

This section of the report aims to understand the data and the context of the business. We aim to determine the possible trends and correlations. As the company is in the "Beauty/Cosmetic" industry, there may industry-related seasonal factors that drive the demand of products.

2.1. Hourly Trends

Hourly trends of the sales were analyzed. There is 2 years of hourly data in the dataset. Thus, there are 365*2 = 730 data points for each hour of the day. This satisfies the Central Limit Theorem. The normality test using the Python stats package results in the p-value 0.04. The p-value is less than 0.05. However, the value is close enough to 0.05 and the graph looks normal. The normality of the data can be assumed.

We can conclude that the time in the day with the most sales is 18:00, followed by 13:00, 19:00, and 12:00. The spike in sales 18:00 is likely due to after-work hours in both the PST and EST time zones. The second spike at 13:00 may be due to lunch-break for customers in PST and after-work for customers in EST.



2.2. Monthly Trends

The mean orders for each month will help the business to determine the seasonality trends of their products. The months of December, January, February, and March have the highest number of mean orders. This is likely due to various commerce-focused events such as holiday shopping, New Year's Resolutions, and Valentine's Day.

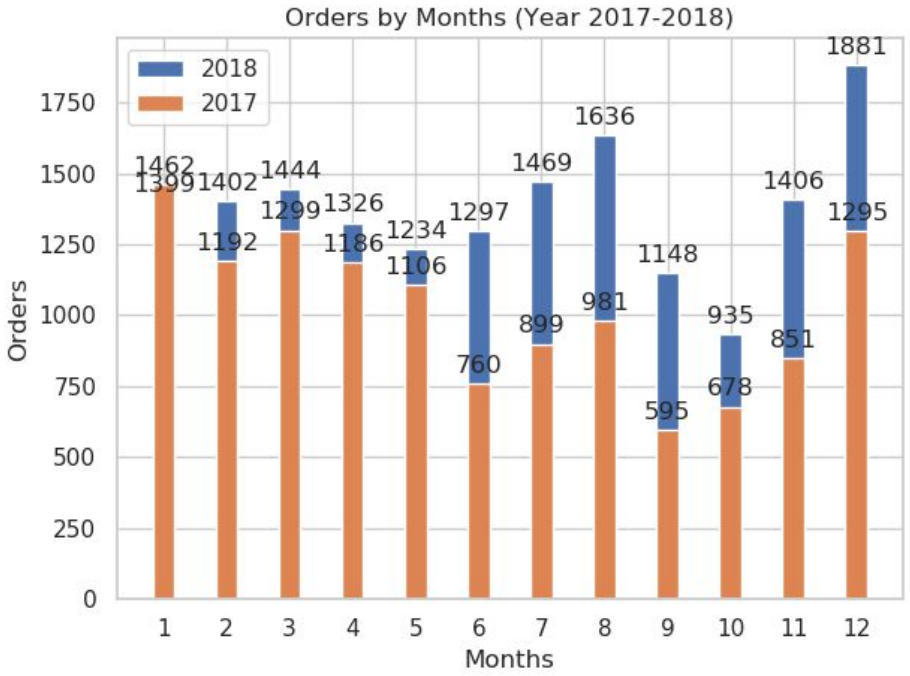
2.3. Yearly Trends

The data for yearly trends is limited as there is two years of data in the dataset. N = 24 as 2 years * 12 months. We will be performing a T-test to determine if the two years of 2017 and 2018 have different means. If the means are different, the business either sold more or less goods in the different years. T-test requires the normality and equal-variance in the data. Thus, we must perform those two tests prior to the T-test.

The normality test of data from the years 2017 and 2018 was 0.58 and 0.39 respectively. Both have the p-value > 0.05, so we can proceed to the equal-variance test before performing the T-test. The equal-variance test has the p-value of 0.248. The null-hypothesis of the equal-variance test states that the data has an equal-variance. P-value of 0.248 > 0.05, so the equal-variance between the orders in 2017 and 2018 can be assumed.

The normality and equal-variance conditions passed. Thus, T-test can be performed on the data. The T-test resulted in the p-value of 0.0024. This p-value < 0.05, so it can be concluded that the different years have different means of orders.

Plotting the orders of different years on the same barplot makes it evident that the orders in year 2018 grew when compared to the year 2017. Further statistical analysis was performed to conclude that the year 2018 grew in sales by 34.72% when compared to the year 2017. If the trends continue, the business can expect similar growth between the years 2018-2019.



3. Models and Results

This section of the report aims to determine the accuracy of various Machine Learning (ML) algorithms in predicting the number of orders. Successful implementation of this algorithm will result in higher efficiency for the business. It will allow the managers to accurately predict customer demand and ensure that there is enough stock in the warehouse to fulfill orders.

3.1. Feature Selection

The data contained many features that could be used to train the models and achieve a high accuracy score. However, we decided to simplify the model by only choosing a subset of features that the company could control. The models were trained using data for year, month, day, and total sessions to the website. The company can control total sessions to the website by increasing the online marketing spend. The models can be utilized before marketing campaigns by inputting the date and total visitors expected to the website. The models are expected to output the number of orders that would be purchased as a result.

3.2. Results

The accuracy scores of the different models are as follows:

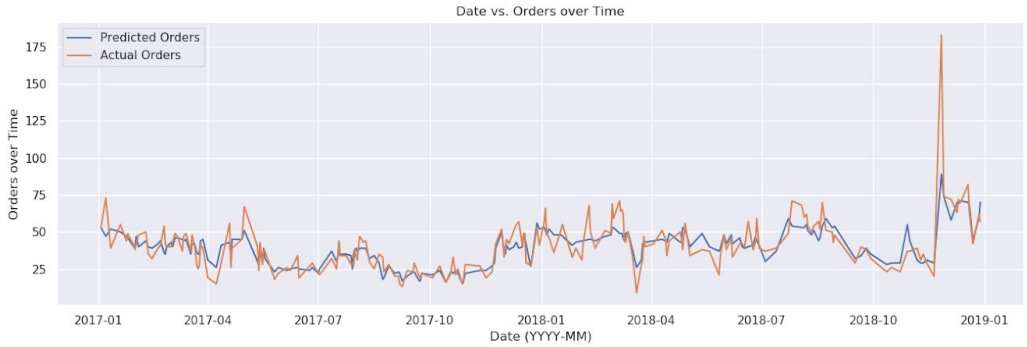
Models	Valid Score
SVC	0.0409
Gaussian NB	0.0136
K Neighbors Classifier	0.0545
Random Forest Classifier	0.0363
Random Forest Regressor (Default)	0.5847
Random Forest Regressor (Hypertuned)	0.6807

As evident in the table, predicting the number of orders is not a classification problem. For instance, it would be a classification problem if the models were expected to output a month with given number of orders, sessions, year and date as an input. Regression algorithms are more appropriate for our use-case of predicting the number of orders from given features. The Random Forest Regressor (RFR) algorithm with default parameters resulted in the accuracy of 58%. This means that RFR was able to predict the correct number of orders 58% of the time.

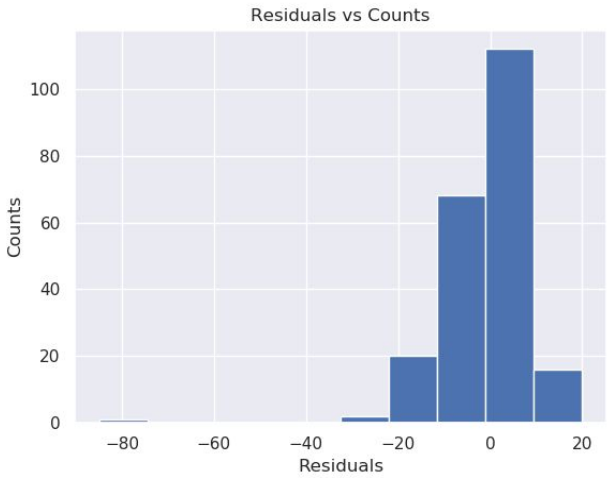
Technique called "Hypertuning Parameters" was used to attempt random variations of different parameter values. The resulting parameter values had the highest accuracy score of 68%. This hypertuned RFR model was then used to perform further analysis.

3.3. Analyzing Accuracy

The predicted number of orders generated by hypertuned RFR model, as well as the valid results were plotted for comparison. As evident in the plot, the predictions produced by RFR are extremely similar to actual number of orders for the given dates. The limitation of RFR model include the spike in orders on 2019 Black Friday. The RFR model was wrong by ~100 orders.



Predicted and actual orders can be further analyzed by plotting the values on a residual graph. The residual graph is used to visualize the difference between predicted and actual number of orders. The residual graph demonstrates that most predictions are within +- 10 of the actual number orders.



4. Conclusion and Future Work

This project gave valuable insight on the performance of different types of machine learning algorithms. We were able to achieve encouraging results using Random Forest Regression with hypertuned parameters. In the future, there is potential to incorporate more data to determine correlations. Dataset containing the national holidays could be used to determine the impact of holidays on sales. The accuracy of machine learning models can also be enhanced by incorporating other datasets, such as data from marketing campaigns and customer user profiles. Further exploration into the use of neural networks may lead to higher accuracy scores than classic machine learning algorithms.

Project Experience Summary

Anmol Bajaj:

- Prepared data for analysis by grouping various columns and merging different file which enabled the plotting of graphs.
- Utilized the stats package in Python to perform statistical tests that resulted in the acceptance or denial of different hypotheses about the data.
- Researched the implementation of parameter optimization of RFR that resulted in the increase in accuracy scores by 10%.

Earl Cooke:

- Plotted various graphs with mathplotlib and seaborn packages to perform visual analysis on the data.
- Developed the pipelines to train multiple machine learning models using given training and validation data to predict the number of orders from given features.
- Contributed to the written report by gathering the various plots and statistical tests which led to enhanced understanding of the order trends.