**MMA 860 Project**

**Technical Appendix**

**Team Chernoff:**

**Akshay Malhotra**

**Isaac Ng**

**Jenny Oh**

**Jiamin Mei**

**Leon Joulakian**

**Lilly Keramati**

**Shitangshu Maity**

**Wendy Fang**

**Feature Engineering**

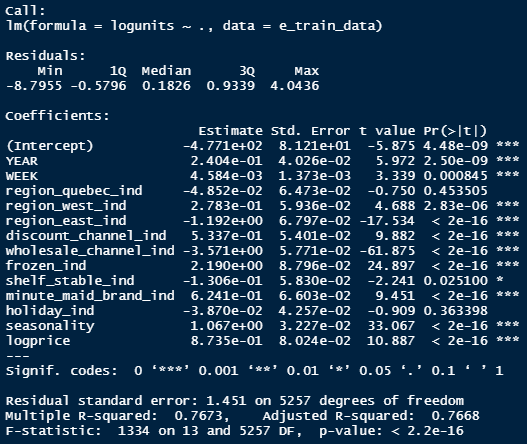
Feature engineering is the second technique we used ahead of the model predicting. Our data set has records from 2017 to 2019; most of them are strings that need to be recreated into features to calculate their impact on volume (unit) performance. The raw data were first cleaned and aggregated to brand-week-level, taking the revenue and units to ensure a single entry for each brand. Several indicators were build based on the client type and product hierarchy like Region, Channel, Category, Brand. For instance, there are two brand types: simply and minute maid; we take 'simply' as the base and create a binary column for 'minute maid.' Same for holiday indicators, they are flagged as 1 for the holiday week and 0 for the regular week. One seasonality feature was built by aggregating the data to a higher level and calculating the mean and standard deviation units. With it joined back to the base table, we can use the brand-week-level units minus the mean and divide by SD to capture the variation on the seasonal trend. Based on plot checks on the model output and our business experience, price and units usually have a skewed distribution instead of normal distribution, and the log transformation helped solve the issue. Several other features were created initially but eliminated or removed by the final model we had, for instance, flyer position indicator and list indicator, as they are not significant to the model.

**Predictive modeling**

After engineering our desired features, our model went through several iterations before landing on the final version (v8). In the screen capture, we can see that when looking at the summary of our training regression, we have an overall P-value of less than 0.05 and an adjusted R-squared of 0.7668. Overall, these two figures show us we are on the right path to creating our training regression model and are within the boundaries of what we were trying to achieve from our proposal. Looking at the engineered variables, we can see that price (which we logged along with the units), brand (minute maid), and channel (discount) had some of the most substantial impacts on our dependant variable. We used the 2017 & 2018 data to build our training regression and our 2019 data to test it on, resulting in an R-squared of 0.7530.

I want to touch upon why we see variables being included in the training model that are deemed not statistically significant by our regression summary. I’ll use the example of “region\_quebec\_ind,” the data we have divides the country into four regions (East, Quebec, Ontario, West). We have used Ontario as our Beta variable, and Quebec unit sales performance trends closely to Ontario. However, product pricing and promotions are varied between the two regions. It is deemed significant for the business to have Quebec as a variable because business strategy differs there vs. the rest of the country.

At first, we planned to build a detailed model that had numerous variables like clients & products. However, as we worked through it, we discovered that this could not be done in a single model but required multiple models for each client and product to be successful. Currently, our MAPE calculation shows us an accuracy of about 0.31 (meaning we are not correct 31% of the time). This model can show us the potential of how much the overall business can expect to sell and at what price point based on the 2022 calendar. However, we fall short of the company’s ask to be provided the ability to dissect the data and determine what can be fine-tuned. This model has shown us other inaccuracies in the detailed output for the 2022 predictions which deems that we cannot currently place this model into production for the business to use in day-to-day planning and requires more refining and testing which will take significantly more time.



**Managing problems with error terms**

We kept tracking the error terms and model explanatory capacity while we were building the model. When working with retail sales data, there can be lots of outliers and issues due to the nature of the Consumer-Packaged Goods industry. At first, we did not take the log transformation on price and units; although the model fits both train and test sets around the same percentage, the overall R-squared is around 0.28, which means the features only explain 28% of the unit sales. And the density plot also supports the points that the data is right-skewed; we need to take a log transformation to make it normal. At the same time, we have similar features in the model, which cause collinearity in the data set. By removing all the insignificant features and with log transformation on the skewed column, R-squared increased to 0.68, and the MSE did not rise. With the seasonality feature introduce and tune the model by removing and adding features, we finally reach an R-squared of 0.75 by the end. The model fits both test and train with different years of data. As the explanatory percentage increases, the error term also decreases. We calculated the MAPE (mean absolute percentage error) using the 2019 actual units and prediction units by the final model; we get the number around 31%, which is much smaller than it initially was.