Capstone Proposal

Dong Kim

February 17, 2018

There are many problems to solve in retail online stores to find the trends of the sales throughout the year. Each year, stores can conclude with what products were sold the most and what products were sold the least. Stores can gather information to find the margin and revenue yearly.

The key problem I want to solve is what products can be recommended the most and what products should be recommended the best based on customer’s sales data. Importantly, this information is very critical to customers who are doing online shopping. If we recommend the high selling products to customers, they tend to look more towards these products and assume they are verified by other customers. Based on my analysis, customers can measure what products are high sellers versus other similar products. Otherwise, people would not know what products are better than the others based on their interest of shopping. Also, when online retailers have list of products are bringing the least outcome, then they can advertise more or do bungle with other products to increase the outcome.

Description of Data: Online Product Sales from Kaggle.com The objective of the competition is to help us build as good a model as possible to predict monthly online sales of a product. Imagine the products are online self-help programs following an initial advertising campaign. Each row in this data set represents a different consumer product.The first 12 columns (Outcome\_M1 through Outcome\_M12) contains the monthly online sales for the first 12 months after the product launches.  
Date\_1 is the day number the major advertising campaign began and the product launched.  
Date\_2 is the day number the product was announced and a pre-release advertising campaign began. Other columns in the data set are features of the product and the advertising campaign. Quan\_x are quantitative variables and Cat\_x are categorical variables. Binary categorical variables are measured as (1) if the product had the feature and (0) if it did not.

Objective: 1) Clean out the dataset and summarize what products have high outcome throughout the year. 2) Make graphs month by month to see the trends of each products. 3) Find out how each variables (Date\_1,Date\_2,Quan and Cat) are co-related by using regression. 4) Find out what products should be bundled (Quantitative variables). This dataset has many different variables and it also includes advertising campaign dates. These variables can be very useful to see how advertising impact on sales. 5) Using regression to predict sales outcome of each products. 6) Summarize what products should be recommended based on the sales.

From now on, I would clean out the dataset to replace all the missing values to the mean of each columns and eliminate unnecessary variables for better visualization. By using ggplot method, I would create graphs of each month’s outcome to see the trends of sales values. Then, I would calculate the mean, median, minimum and maximum of each month to find out what products have sold the least and most. I will filter products by median or above to find what products are within these range. These can eliminate all other outliers and can focus on high sellers. I can use regression method to find how each variable are co-related and test my linear model to see if it’s effective.

Deliverables: 1) Recommend to customers what products are being sold the most throughout the year 2) Recommend to customers which products are being sold the least and why 3) Recommend to customers which products have best advertising impact