Association Rules Project Report

This project entailed observing transactions at various Dillard’s locations over time and attempting to find associations between the data to move around stock keeping units in order to better facilitate customer activity. The first stage of this problem was coming to understand the data given to us and how to operate on it. This required reading and attempting to understand the data dictionary in order to properly label the datasets given but also getting and understanding of how Dillard’s functions as a retail outlet. After browsing the dictionary, I determined that the two required datasets to operate on were the transact.csv and skuinfo.csv as they contained the info of all of the individual transactions at the locations and further information on the sku’s I needed to make conclusions on. I also realized that based on the layout of a Dillard’s, moving sku’s between departments was impossible because they are grouped by brand. Thus, I decided only to look at sku associations on an interdepartmental level for the departments. Additionally, after trying to play around with the relatively large data set, I realized that the data would require quite a lot of pruning in order to run association rules practices on the data. As a result, I developed the methodology below in order to solve this problem.

First, I factored out returns because doing a proper analysis on sku’s and return rates would be very complicated and out of the scope of this class. Second, I factored out all single item purchases because when dealing with such a large amount of transactions I wouldn’t like to waste any time or computational speed working with data that could provide no associations. Then, I planned to sample different transactions from stores all around the country, as these sku changes would be applied on a nationwide level so not having data that is representative of all chains would be inaccurate. As a result, I ended up factoring out stores with a very small number of transactions because my next step was to randomly sample transactions from each store. The numbers I chose for store transaction numbers and number of transactions to sample came from evaluating the value counts of each store and the transaction numbers at those stores in order to determine a cutoff that would be loose enough to allow for diverse data but strict enough to allow for efficient computation. After this I factored out sku’s with less than 10 appearances in the remaining dataset, because the most computational expense in the association rules process comes from the creation of excess dummy variables, which are based on the number of sku’s present. Next, I began to separate the data into its separate departments via importing the skuinfo.csv file to get the department number of each sku. After all the sampling and separation, many of the departmental datasets were still too computationally intensive for the association rules functions to handle and as a result I had to limit the datasets some more. As the main issue is the high number of sku’s being observed along with getting enough datapoints for each sku to make valid conclusions, I had to try and target a limited number of sku’s with a high number of observations each. Thus, for each department I chose to target sku’s with more than 200 observations and limit the total number of observations to between 6000 and 25000 to avoid wasting time on departments with too little observations to draw valid conclusions but also keep out departments that would cause python to crash if association rule calculation was attempted. After this the mlxtend package was used on each valid department to obtain the association rules and related metrics. I then chose to observe rule with a lift grater than 10 and a confidence greater than .6, as I thought these were relatively good metric for determining strong association rules. Finally, after receiving the end data of this, which was around 250 rules, to choose the best 100 I decided to multiply lift confidence and support to get the best idea of which rules were most common and most correlated. My results on the top 100 rules according to my criteria are in the attached .csv file. These rules are quite strong in my opinion and should be heeded to move the associated items closer together in order to improve the customer experience and facilitate more sales.