**Dillard’s Point of Sale Data Association Rules Analysis**

Charles Marshall

Prof. Klabjan

IEMS 308

13 February 2020

**Introduction:**

The premise of this business problem is to help Dillard’s, a struggling retail department store, rearrange the floors of the store in a way which will place items likely to be bought together next to one another. This movement will promote customer sales by making it easier for customers to access items which they will be statistically likely to buy together. The analysis of this report will present potential improvements to Dillard’s current store format by suggesting items which should be grouped together according to association rules.

**Data Exploration:**

One of the most important parts of this project was to understand and explore the data. I sought to initially understand the data by viewing the relational database schema and understanding how certain tables were connected to one another. The two most important tables in this project were the transaction data and the SKU information. The transaction data gave raw point of sale data and was used to do the market basket association rules analysis. The SKU information was used to find out more about items included in the rules, including the style, color, size and brand. This was useful for drawing conclusions from the rules.

Additionally, the most challenging part of data exploration for this project was dealing with the large datasets. The transaction data was so large that it could not be loaded into Jupyter Notebook in its entirety, instead it was read in using only the first 7 columns including all the primary keys and the SType, the type of the transaction. The other columns, such as AMT and MIC, were left out because they are not relevant parameters to creating market baskets and doing association rules analysis.

Lastly, because the transaction dataset was so large, approximately 121 million rows, it needed to be broken up in a creative way. I decided to subset the data using specific SKU values. Instead of using random SKUs, I decided to use SKUs which corresponded to specific departments, more specifically the Polo Men’s and Reebok Departments. I chose these departments because I was familiar with the brand and thought I would be able to more effectively analyze the rules and because they had a large number of unique SKU values, with Polo Men’s having 142k and Reebok having 27k unique values. A large number of unique SKUs shows that each department allows for a large number of choices.

Besides departments, I analyzed many other options to potentially use to subset the data. For instance, I decided against using specific store locations as a parameter because I wanted to create a rules which could be applied to all locations, not just one store. Additionally, I considered subsetting the data by only considering a specific time period, specifically the summer of 2005, because I know retail stores often change their configurations when new products come in in different seasons. However, I found that using the summer still left the data set with 23 million data points, and it would therefore need to be broken up more. Both of these approaches are valid and could have been used, but in the end I decided to go in a separate direction with the business question I answered.

Talk through why you decided to subset on the department including

**Solve the Problem:**

**Analysis:**