**Gene Chen**

**Assignment 2 - Association Rules**

**Executive Summary**

Dillard’s is a major department store chain with 292 stores across the United State. They provide a wide range of consumer goods but specialize primarily in different clothing brands. They have provided us with a point-of-sales data over a time frame.

Using this data, I have applied machine learning methods to find ‘association rules’. These rules are essentially if-then statements that can help us determine relationships between items within the set. This is especially useful in the market segment since it allows us to determine whether an item(s) (called the consequent) are more likely to be bought alongside the purchase of another (called the antecedent). This provides valuable information and subsequently allows the store to optimize its layout.

To find the association rules for Dillard’s, I used the Apriori method. Which takes item baskets of each transaction and calculates performance metrics for each combination of items in all item baskets. Each of these combinations consequently results in one of our association rules. Executing this method, I was able to rank the output of the algorithm and determine the best 100 rules and, consequently, the items that have the ‘strongest’ association with another item(s).

**Problem Statement**

Dillard’s has a large dataset with information on all transactions. My job was to use this data to determine ‘association rules’. This allows Dillard’s to pursue an optimized layout of their stores to maximize profit. Due to technology constraints, I decided to focus on the state of Idaho and determine rules for the 3 stores of store ID 3409, 3509 and 3609.

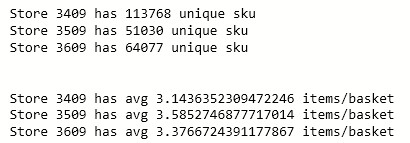
**Assumptions**

1. The data provided by Dillard’s is valid and correct
2. Parameters of Apriori algorithm are not such that rare items are omitted and/or not considered
3. Returns are not pertinent to determining the true association of items as these were filtered out

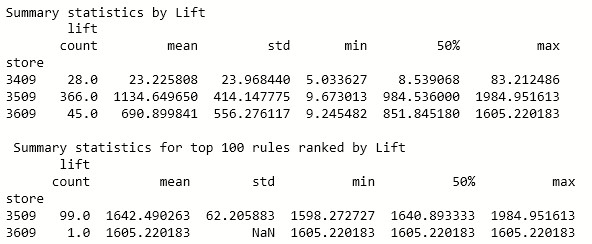
**Methodology**

I first began by pulling the point-of-sales data and initiating data cleanup. This was done in an external csv editor, EmEditor. Using this powerful tool, I removed all non-primary/foreign key columns if they were also not pertinent features for the analysis. The new data was now more easily read by python. I also took this time to subset the data by filtering out transactions labeled as returns. Finally, I took only the transactions that occurred in the state of Idaho and grouped them by the three different store ID’s.

The next step in my methodology was data exploration. I was able to determine, for each store, the number of unique SKU’s as well as the average number of items purchased per transaction. Note that Store 3409 has ~2x as much unique items compared to store 3509 and 3609. However, the average amount of items per transaction are all relatively equal.



Next, we began the Apriori algorithm, running it separately for each individual store. Since the function in python takes a list of item baskets, I first formatted the data as such. Then, I ran the function with specific parameters. The minimum support value I initially set was .05%. This means the algorithm will only include items that show up in more than .05% of transactions. This number is usually used as a means of saving computational power and/or memory. However, since the algorithm was still able to execute in an acceptable amount of time while handling a relatively small support threshold, I decided to keep the value as is. Note however, in cases where the function will not run properly in the time one needs, the minimum support value should be increased for the reasons mentioned before.

After running Apriori successfully, I began the analysis on the results. Since the algorithm outputs results in a congested, almost uninterpretable way, I decided to reformat for easier accessibility. The resultant table was then exported as a CSV file named allRules.csv, which I have included.

Shown above are the summary statistics for the final table of rules grouped by store. As one can see, the average value of lift for the rules of store 3509 are considerably higher than for store 3409 and 3609. This aligns with top 100 rules ranked by lift. As rules from store 3509 take up 99 of the 100 rules. These numbers are powerful as lift is defined as having the consequent be *(lift)*x’s as likely to be purchased alongside the purchase of the antecedent. However, not all 100 of these rules can be actionable since some of these might have interrelationships with one another. For example, rules a🡪b and b🡪a are redundant for our purposes. Because of this, I also provided a separate excel file compiling all unique rules.

**Conclusion**

Through this assignment, I was able to successfully determine association rules based on the point-of-sales data provided by Dillard’s using the Apriori algorithm. I subsequently ranked the list of these rules using the value of lift. From this, I determined that **store 3509** **has much potential to be optimized due to significantly high lift values associated with over 350+ rules.** However, optimization for the other stores should not be discouraged as rules for the other stores are low relative, but still significant. Some next steps include affirmation analysis. Through research I came across another metric called conviction. This metric explains how likely a rule would be *incorrect* if the association between the antecedent and consequent was by random chance. Executing this analysis could help allowing Dillard’s to omit certain rules and focus on the ones more concrete statistically.