**Dillard’s: Association Rules**

Anand Raman

**Executive Summary**

At Dillard’s, many customers have baskets of just one item. By analyzing these top items, I found that many customers buy makeup from brands such as Clinique and Lancôme. One outstanding issue is that cost and retail price data is missing for several top selling products. This is an indication that Dillard’s should improve data quality so that better insights may be provided in future iterations. After running the association rules algorithm, with a minimum support of 0.001 (chosen for computational power reasons), I was able to produce a list of about 100 candidates for reorganizing the planograms. Upon closer inspection of a few of these associations, I noticed that all associations I looked into were homogenous across brand. Customers were purchasing two or three items made by the same company. Dillard’s can take advantage of this apparent brand loyalty and preference with promotional pricing to drive sales. Another possible course of action would be to try and stimulate customers to diversify their baskets. One interpretation of these homogenous baskets is that customers come in for a single purpose and then leave after that is fulfilled. One potential solution would be to spread out products of the same brand that are in the same department, which would have customers walk through more of the store and potentially see something they are willing to buy on impulse. In particular, Dillard’s could stock some of their top selling items (Clinique and Lancôme products) at the front of the store, near checkout so that customers are encouraged to purchase something on the way out. In particular, the less expensive make up/skincare items could be quick and easy purchases that require less financial consideration by customers.

**Methods**

To clean and process the transactions data, two files were used: the file containing store info and the transactions file. A random sample of five stores was drawn from the store info file. This number was chosen largely due to processing power limits of my machine. The transactions file was filtered for rows corresponding to the stores and transactions type “P”, which stands for purchase. It would not be useful to get association rules from a dataset that included purchases and returns.

The theory behind sampling stores, rather than just randomly sampling from transactions is a pseudo single stage cluster sampling technique. By capturing all the data from the sampled stores, it is certain that complete transactions are captured, as rows in the original transaction data do not correspond to unique transactions (SKU is part of the composite primary key).

After sampling from the transaction data, SKUs comprising less than 0.0001 of observations were filtered out. This could be considered a preliminary support filtration. I selected this threshold after trying several values and settled for 0.0001 as this would have no SKU observed fewer than 125 times. This threshold was chosen because it limits the possibility of combinatoric explosion and enabled my computer to output a reasonably sized matrix of dummy variables.

Following this filtering step, SKU was one-hot encoded and joined back to the transaction data. To identify unique transactions, I grouped by store, register, trannum, and seq and summed. Since apriori algorithm will not recognize numbers other than 1 or 0, any row-feature combination greater than 1 was recoded to 1. Essentially, if a customer purchased more than one of the same SKU in a transaction, this was recoded to 1.

The final dataframe was written to a csv and analyzed in a separate file. Simple count operations were performed to find the 10 most commonly purchased items in the sampled data. After analysis on the full dataset, rows with a sum equal to 1, or customers who only purchased one item were filtered out. Single purchase customers accounted for approximately 2/3 of the data. However, association rules will not work for customers who only buy one item. There is no benefit to running the analysis with these customers, so they were removed.

Finally, I used the mlxtend library’s apriori and association rules algorithm to find and analyze associations in the data. Once again for computation reasons, I selected a min\_support of 0.001. The metric selected for the association rules was lift. I chose this because I was interested in comparing how frequently items were bought together as compared to if they were independent. This metric seemed to provide the most efficient solution given the context of the problem: findings SKUs that should be moved on the sales floor.

**Results and Discussion**

*Exploratory Data Analysis*

The total number of transactions numbered over 136,000,000 rows. The data came from 453 unique stores in 31 distinct states. Overall, there were a total of 1,564,178 unique SKUs in the original dataset.

In the sampled dataset, there were 1,264,551 rows from the transaction file. After filtering for SKUs occurring less than 125 times then grouping by the relevant primary keys and summing, this yielded a total of 168,096 unique transactions and 550 columns.

In the sampled dataset, the average basket size was 1.52 items with a standard deviation of 0.943. Well over 50% of customers left the store after purchasing just one item. The basket size was right skewed. Outliers were not removed and there were relatively few. One individual bought 19 items from Dillard’s, which sounds like a rather expensive trip. Since association rules do not rely heavily on basket size or any particular distribution, the skewness of basket size should not have significant effect on the solution.

A picture containing clock

Description automatically generated

Figure 1: Basket Size Counts

In the final, sampled dataset, the average profit is approximately $8 per unit. Average cost is $11 and average retail price is $19. The vast majority of items have a profit margin of 40% (over 75% of items). The maximum profit margin is 80% On some products, the store posts a loss. Approximately 22 items recorded a loss. The maximum loss was $4.86 per unit. The product was made by a company with the name abbreviated as WESTPOIN. The item with the highest margin was made by a company with the name abbreviated MURANO N. The item cost $7.90 and was sold at retail for $39.50. The following plot describes the distributions of the costs, retail prices, and profits and margins of the sampled dataset. Profits are approximately normal, with a relatively small spread, particularly compared to cost and retail prices. Margin is also relatively normal, with a large number of observations occurring at about 40%.

A screenshot of a cell phone

Description automatically generated

Figure 2: SKU Cost, Retail, Profit, Margin Histograms

*Most Commonly Purchased Items*

Since many transactions were a single purchase, I analyzed some of the best-selling items at Dillard’s. The best-selling item was seen in 6% of transactions. It is an item made by Clinique. In fact, 8 of the 10 best-selling products at Dillard’s are made by Clinique. The other two are made by Lancôme. The maximum profit margin was 40%. An interesting feature of the data is that three of the best-selling products were not contained in the skstinfo file. This mean that data was unavailable for cost, retail and subsequent profit margin analysis. This is a considerable data quality issue as Dillard’s should be able to track the financial performance and viability of top selling products. The top selling product seemed to have erroneous data for cost and retail, as both were listed as 0. There are two possible explanations, the first is that this was a promotional item that was given for free. The second explanation is that the data in the skstinfo file has been misreported. Overall, my recommendation is that Dillard’s consistently stock the following items and potentially locate the makeup and cosmetics department in a central location in the store so that customers see it or walk through it no matter where they are going.

A screenshot of a cell phone

Description automatically generated

*Association Rules Results*

In total, 300 rules were found. Since the algorithm prints each unique rule twice, flipping the antecedent and consequent, this totals 150 unique rules. Average support was 0.2%. Lift averaged 73.79 and was heavily right skewed, with the median lift being 9.34. Confidence was on average 0.26 with a standard deviation of 0.26 as well. The maximum confidence value was 0.95, which means that no consequent-antecedent pair always occurred together.

The top ~100 candidates are printed in the table attached at the end of the document. They are taken directly from the association rules output, sorted by lift, and then deduplicated (mlxtend association rules output prints the “same” row twice, just flipping consequent and antecedent.

In output.txt, I specify about 100 candidates for items that should be arranged together. In this report, I will analyze a few candidates, all occurring in the list and provide more detailed analyses.

The pairing of SKUs 9600684, 9469364 are two undergarments both made by the company abbreviated CABERNET. SKU 9600684 has an above average margin at 57%. There is no pricing or cost data for the other SKU. This could be a good candidate for a promotional pricing program. If the margin on 9469364 is favorable, it could be reasonable to put 9600684 on sale to drive sales of 9469364. The two are bought together very often, with a support of 0.945. Once again, the lack of pricing data drives home the importance of data quality.

SKUs 6972521 and 7232521 are excellent options for promotional pricing. Both have above average margins, 41% and 62% respectively. The brand is abbreviated NOBLE EX, which supplies bedding (sheets, linens etc) at Dillard’s. If we had more information about these two items (i.e. product type) we could infer more about how and when to provide these promos.

SKUs 7039904 and 5329905 are two relatively high margin products at 60% each. However, the total profit is not very high at $2.40 each. I was unable to find what this company produces, but based on the profits, I think that there are better options than these two products to relocate. Especially consider that support is quite low ~0.1% for each.

I also thought it may be a good idea to look at a one of the three-way associations. SKUs 6752521, 6642521, and 6742521 are all manufactured by NOBLE EX, but only one has margins available, 6752521. This is the second time this SKU has appeared, so it could be a good idea to reorganize the floors so that these items are all in very close proximity in their respective aisle. The second most popular association with three SKUs were also three products manufactured by NOBLE EX.

**Conclusion**

Association rules were able to provide a list of at least 100 products worthy of consideration in rearranging of planograms. Many customers who purchase more than one item buy two or three items of the same brand. These homogenous baskets can be advantageous for Dillard’s as they can provide discounts by brand such as BOGOs or buy one get one half off. Dillard’s can address this homogeneity by spreading out brands within the same department and advertising many choices to the customer. In addition, Dillard’s should arrange it’s make up department, particularly products from Clinique and Lancôme in a central location that all customers either see or walk through. Furthermore, it could be advantageous for Dillard’s to put the some stock of the best-selling make up items described in the top 10 analysis in the front of the store, near checkout, so customers are encouraged to buy some of these items on their way out.



Figure 3: Potential Candidates