IEMS 308 Assignment 2 Report - JunHwa Lee

0. Executive Summary & Introduction

With the increase in online shopping services, such as Amazon and eBay, shoppers these days are so used to convenient shopping experiences. No matter where you are and no matter when it is, you can simply go to the website, look around different products, and order the product you want in one spot. On the other side, due to the increasing standard of shopping convenience, offline stores that offer less convenient shopping experience have been implementing different strategies to make the whole shopping journey seamless.

Modifying the planograms based on the technique called "association rules" would help Dillard's reach that goal. By going through transaction data, I identified 332 pairs of product groups that are correlated and that have a high likelihood of happening in many Dillard's stores. Since they are correlated, those pairs can be used to perform strategic promotions.

But more importantly, assuming that products with the same Stock Item Classification (CLASSID of POS) are displayed close to each other, I recommend 35 SKUs that could be moved to 36 different locations to decrease the inconvenience of buying one item here and finding the related item at the opposite side of the store. Luckily, other than those 35 cases, every product group pair was close to each other, and there was no need to change anything at the department level. This will reduce the estimated cost of this rearrangement operations.

In addition, all of those 35 SKUs that were recommended for rearrangements came from CLINIQUE and CELEBRT Department. Therefore, Dillard's does not need to rearrange all 60 departments but instead, start from two departments and expand if necessary. Also, considering that all those SKUs were by Clinique and LANCOME, the store managers can put extra focus on displaying products from those companies.

With all these efforts, it is expected that Dillard's would have a clear understanding of what products to move to improve the consumer's convenience and promote better experience at Dillard's.

1. Data preparing & preprocessing

1.1. Labeling Columns

All the labels of every dataset column were missing. Therefore, before any other steps, I started with opening each dataset and giving labels to its columns. By looking at data schema and value types of column description, most of the columns could be identified with no confusion. Also, the last column of all the data was deleted because it held non-meaningful binary values.

However, some special cases needed extra attention. For example, TRNSACT had one additional column that was not on data schema. I identified that column as *SPRICE* column. Also, for future convenience, new column *trnsact_identifier* was added to mark the unique transaction. When I defined one transaction as a combination of *STORE*, *REGISTER*, *TRANNUM*, *SALEDATE*, and *SEQ*, I ended up with only one *SKU* per transaction. However, excluding *SEQ* from the definition led to more reasonable results with more than one *SKU* per transaction. Therefore, I did not use *SEQ* for creating the *trnsact_identifier* column. In addition, the last two columns of SKUINFO were formatted irregularly, and those were ignored due to non-meaningful or unidentifiable data points.

After all labeling, all the tables were newly saved as "[original file name]+_labelled.csv" for the future convenience.

1.2. Understanding the Structure of Data

To form a better understanding of the data, I scrutinized the structure of the data. That scrutiny processes involved checking the shape, missing values, distribution of values in each column, and unique values of each column.

Fortunately, most of the tables had no missing value. There was one *SKU* on SKUINFO that had missing *COLOR* data point. Nothing had been done at this point because there was a possibility that that *SKU* would not be selected as we go through the filtering process. Also, when I checked the distribution of each column, there seemed to be no apparently wrong values (such as outliers), and all the values were in the reasonable range.

However, by checking the number of unique values, I found that I need to be extra careful about joining two tables. For example, the number of unique *SKU*s in the TRNSACT table was 714,499, which is smaller than 1,564,178 of SKUINFO. Also, the number of unique *STORE*s in SKSTINFO, which is 357, is larger than that of TRNSACT, which is 332. It is hard to identify why such discrepancies happened, but it told us that rechecking must be done whenever we merge or filter different tables.

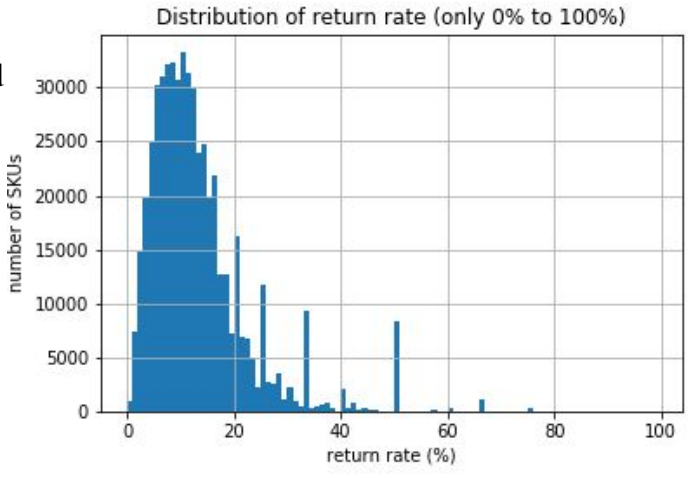
1.3. Filtering Data

For this analysis, I chose a subset of data. Specifically, I created a subset of *SKU*s and *STORE*s. There are two reasons for this decision. The first reason was that it was not realistic or possible for my machine to compute all the possible combinations of association rules that take a massive number of *SKU*s into account. However, the more important reason was that, because all the stores have different sizes and different portfolios of *SKU*s, association rules that are built from the whole dataset are likely to be affected by those differences. To control some of those differences and find the rules that can be applied to many, I decided to start with smaller but reasonable samples and validate my samples and conclusions throughout the analysis.

More in-depth exploratory analysis on the TRNSACT table was conducted to come up with appropriate standards to filter the data. TRNSACT table was appropriate because it was going to be the foundation for building association rules.

1.3.1. STYPE Selection

On the TRNSACT dataset, there is an *STYPE* column that represents the type of the transaction. Each *SKU* in each transaction can be either purchased or returned. We do not want to spend precious resources on rearranging products that have a high probability of return; Even if a particular product experiences large purchases, a large amount of return on that same product leaves our business almost no profit or benefit. Therefore, I tried to choose the products that have a low return rate where the return rate is defined as 100*(number of quantities returned)/(number of quantities purchased)%.



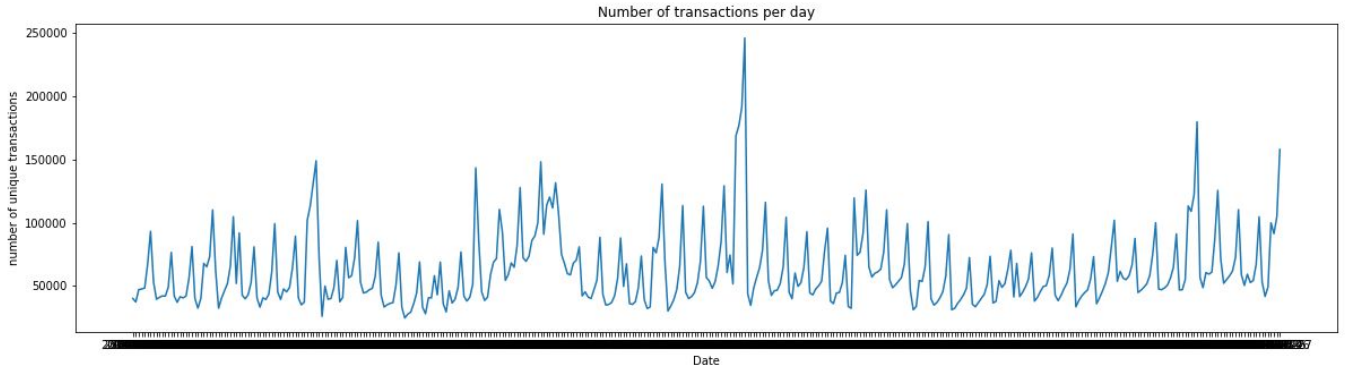
Based on my analysis, there were 24.51% of *SKU*s with no return, which is positive. On the other hand, the maximum return rate was 1,000% which means that that *SKU* had 10 times more return than the purchase. When I plotted the histogram of *SKU*s based on the return rate (as shown on the right), bulk of *SKU*s had lower than 20% of the return rate. Therefore, for further analysis, I chose the *SKU*s which have the return rate smaller

than 20%. Also, as we are interested in the association rules related to the purchasing decisions, return transactions were excluded from further analysis.

1.3.2. SALEDATE Selection

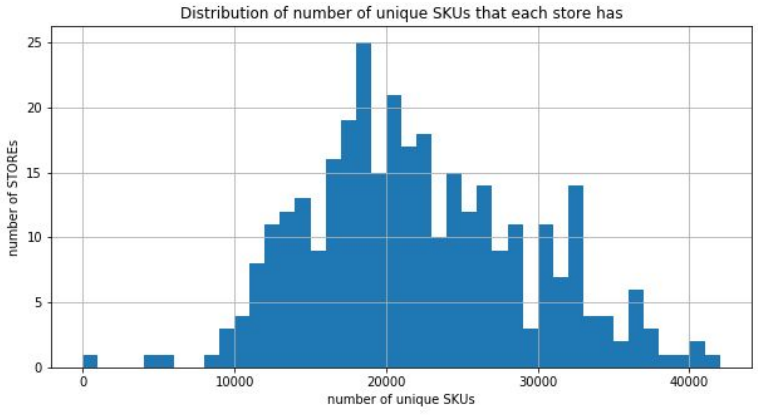
As the department chain and retailer, Dillard's unavoidably gets affected by the seasonal trend, and there might be changes in the number of transactions per day. If there is a specific period with a comparatively higher number of transactions, it might be better to tailor our product planograms for that period.

To check that assumption, I created a line graph (as shown below) that has *SALEDATE* on the x-axis and the number of transactions on the y-axis. Although there were on-going weekly fluctuations in the number of transactions, no specific period showed a significantly higher number of transactions. There was one peak on Feb 26th, 2005, caused by the increase in sales of lingerie and children's apparel, but we can see that it was a temporal success as we look at the general trend (source). With this graph, I could show that there was no specific time with a higher number of transactions, so I needed more qualitative reasons for choosing the subset of the sale date.



As explained in the assignment prompt, rearranging the floors of the stores is a costly process that it is not possible to have it done frequently. However, to capture consumer's demands and preferences that change throughout the year, some number of rearrangements are needed throughout one year. As a compromise plan for those two opposing reasons, I thought that conducting the planogram modification every three months is a reasonable frequency for both saving the cost of that process and capturing the market demand. Therefore, I chose data from 2004-08-28 to 2004-11-27, assuming that the data of that period will help us predict the upcoming transactions from late August to November 2005.

1.3.3. Store Selection



It would have been ideal if all the *STORE*s share many *SKU*s, but that was not the case. Based on my analysis, there was no single *SKU* that was shared by all the *STORE*s. Therefore, I needed to create a subset of *STORE*s and find *SKU*s that are common in that subset of *STORE*s.

Two opposing rationales got involved in the *STORE* selection. If we choose the store with a large number of *SKU*s, then the association rules that we find might not be applicable to other stores because other stores might not have *SKU*s in those rules. Contrarily,

if we choose the store with a small number of *SKU*s, the rules we find would be dependent on limited *SKU*s, opening the possibility of ignoring better rules that involve other *SKU*s.

Therefore, to handle that trade-off between applicability and effectiveness, I created a histogram of stores based on the number of *SKU*s that each *STORE* has. As it showed the relatively symmetrical shape, I chose the mode, which included 25 *STORE*s that had transactions of 18,000 ~ 19,000 *SKU*s.

1.3.4. SKU Selection

This was the last stage of the filtering that I conducted. Instead of taking a look at all the transactions of 25 *STORE*s that I chose, I looked for the common *SKU*s that were shared by those 25 *STORE*s. There were 343 *SKU*s that satisfy such a condition. By excluding *SKU*s that were not shared, I could prevent myself from getting the rules that were relevant to certain *STORE*s. Instead, I could get general association rules that were shared by all 25 *STORE*s.

1.4. Verifying whether our filtered data is representative enough

As we selected the specific set of *SKU*s from specific *STORE*s, I needed to demonstrate that this subset is representative enough and applicable to other *STORE*s too. 353 common *SKU*s that we identified in Step 1.3. appear in 70.78% of *STORE*s in the original unfiltered data set. Although 353 is far less than the unique number of *SKU*s in the dataset, we could identify that our selection is representative enough that finding effective association rules related to this filtered subset of *STORE*s would be applicable to many of Dillard's *STORE*s.

1.5. Filtering other tables

As we validated the representativeness of the filtered data set, I used that filtered TRNSACT table to filter all the other tables too. From here, although they were filtered, filtered datasets would still be called with their original names. Also, during this process, I combined SKUINFO and DEPTINFO and gave that cominbed table the name of SKU_DEPTINFO.

2. Exploratory Data Analysis

2.1. Rechecking the structure of data

Again, since I had gone through filtering processes, I rechecked our data and validated that there are no apparent issues with the data itself, and the distribution of each column seemed to be reasonable. I no longer needed to care about the missing *COLOR* value of one *SKU* in the SKUINFO table (the issue I identified in Step 1.2.) because I was no longer using that *SKU*.

2.2. Checking the relationship between data tables

All the tables for this assignment were connected to each other, and the data schema shows those relationships. In an ideal situation where not a single data point is missing, foreign keys of two connected tables should match.

While other pairs of tables were clear from problems, the relationship between SKSTINFO and TRNSACT table had an issue; while TRNSACT and SKSTINFO had the same list of *STORE*s, SKSTINFO were missing 99 *SKU*s out of 353 *SKU*s of TRNSACT table. However, I still included those 99 *SKU*s for creating the association rules, because, through Step 1.4., I became sure that those *SKU*s do exist in many *STORE*s and *COST* and *RETAIL* were not necessary information for my analysis.

3. Applying the Apriori Algorithm

With 71,229 transactions that involved 353 *SKU*s in 25 *STORE*s, I created the association rules by using Apriori Algorithm. Although this algorithm requires numerous database scans and computations, it was chosen in that it is easy to implement and that I could systematically get the association rules that satisfy the minimum support (minsup) and minimum confidence (minconf) conditions.

In terms of the data setup, I did not take quantities or profits into account, and no weights related to those were given. Instead, I only used the traditional structure that checks the likelihood of the person buying product B, given that the person has product A in his/her cart. The assumption behind that choice is that (1) the benefit of putting positively associated products together is more about giving comfort and convenience to users and that (2) the convenience cannot be represented with the quantity or product margins of the purchased product. Therefore, I decided to focus more on identifying (1) what are the products that are positively associated, and (2) where those products are located.

For each minsup and minconf, I used 0.0005 and 0.01, respectively. Although minsup seemed to be too low, considering that we were using the 71,229 transactions that spanned three months, we are selecting rules that are likely to happen more than 0.0005*71229 = 35.6 times per each quarter, which is not that low. Therefore, I stayed with 0.0005 for the minsup. For the minconf, 0.01 was used to find the rules that get applied at least 1% of the time when precedents were in the basket. Also, I set the minimum lift threshold as 1.5 because I want product sets that were not independent of each other but were rather positively correlated.

SKU	TO WHERE
108507	2
264715	1
348498	1
726718	1
803921	1
1206132	1
1400555	7
2716578	2
2726578	2
2783996	1
3013129	1
3524026	2
3559555	1
3582465	1
3611367	1
3631365	1
3690654	2
3898011	2
3968011	2
3978011	2
3998011	2
4108011	2
4112626	1
5079809	1
5528349	1
5528349	4
5928099	6
6318344	2
7064350	1
7261032	1
7808101	1
8618636	1
8718362	6
9402188	1
9526376	1
9594893	1

4. Insight & Discussion

4.1. Interpreting Basic Features of Created Rules

With the given parameters and the data, the Apriori Algorithm gave me 332 rules. Although I set the minsup to be 0.0005, the average support of the whole rule was 0.011, and the max was 0.006, which is the same as 427.3 times per quarter. Similarly, confidence had a mean of 0.1186, which was higher than the threshold (0.01). There was also a rule with a confidence value of 0.91, which is really high. Lift showed a significant standard deviation of 17.70, and it ranged from 1.51 to 123.39. These high values that are far from the minimum threshold demonstrate that I did not do oversampling and that I could find some meaningful rules that are effective in real life.

4.2. Identifying what SKUs to Move & where to move those SKUs

For this question, I assumed that *SKU*s with the same *DEPT* and *CLASSID* would be located on the same floor (or at least close to each other). Although there is no 100% guarantee, I thought that it is a reasonable assumption because Dillard's have not done any similar analysis and because many malls by default plan their floors based on the product types.

No rule had antecedents and consequents in different *DEPT*. It is interesting that it shows that many correlated purchases with a certain amount of support and

confidence actually happen in the same *DEPT*. Therefore, creating the floor planning that mixes different *DEPT* would cause inconvenience in shopper's shopping experience.

However, out of 332 rules, 102 rules (30.72%) had antecedents and consequents with different *CLASSID*. That is, reassigning locations of *SKU*s involved in those 102 rules would promote the convenience of the shoppers, which is the whole goal of this project. Overall, there were 36 possible rearrangements that include 35 unique *SKU*s, and I identified where each *SKU* should go based on its antecedent's or consequent's *CLASSID*. The result is shown as the table on the previous page.

4.3. Finding Patterns among SKUs that were recommended for Rearrangements

Then, I identified the shared characteristics among *SKU*s that were shown in the table above.

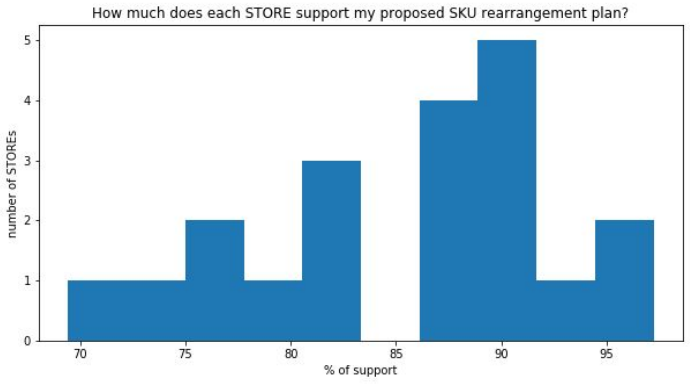
The first characteristic is that those products are mostly from DEPT 800 (CLINIQUE) and DEPT 2200 (CELEBRT). Especially, DEPT 800 take 29/35*100% = 82.8% of the products. This echoes with the previous finding that there is no need to mix up products from different *DEPT*s, but intra-DEPT rearrangements (especially DEPT 800 and 2200) would be enough to bring positive effects. It is great that there is no need for grand transformations on every single department, meaning Dillard's does not need huge investment.

DEPT	From CLASSID	To CLASSID	%
2200	1	6	2.78%
	4	6	2.78%
	6	1	11.11%
		4	2.78%
800	1	2	27.78%
	2	1	47.72%
		7	2.78%
	7	2	2.78%

Related to the first point, all the DEPT 800 has a brand of CLINIQUE, and all the DEPT 2200 has a brand of LANCOME. This finding will be good advice for store managers; the managers can pay additional attention when it gets to displaying and positioning products from those cosmetic conglomerates. Specifically, as shown on the table on the right, for DEPT 800, it is better to put the product with *CLASSID* of 2 with the product with *CLASSID* of 1.

*4.4. Validating identified SKU movement plan with other STORE*s

The last step involved validating our planogram change plan with other *STORE*s. Because we chose a subset of *STORE*s in Step 1.3.3., I needed to prove that our suggested *SKU* rearrangements are applicable to other *STORE*s too.



As mentioned in Step 1.4., almost 70% of *STORE*s have 343 *SKU*s that we used for the association rules. I randomly chose 20 *STORE*s and ran the Apriori Algorithm for each *STORE*. Then, I went through the same process to come up with a recommended *SKU* rearrangement plan for that specific plan. Lastly, I checked how many out of our 36 rearrangement plans are included in the plan of that store. If that proportion is high, it means that our 36 rearrangement plans are universal and applicable to other *STORE*s too.

As shown in the figure on the right, all *STORE*s have a proportion higher than 70%, the mode is near 90%. That means that the rules we identified are not limited to 25 *STORE*s that I used for the analysis. Rather, they are applicable to many more *STORE*s, which makes my recommendation stronger.

5. Limitations & Future Steps

The major limitation of my approach was that it could only come up with 35 *SKU*s for the modification of planograms. Although I identified more than 20 ways to move *SKU*s to different locations, 35 *SKU*s out of hundreds of *SKU*s in the store might offer too limited choices to shop managers. For the future step, I should find a way to come up with more *SKU* options without hurting rationales and assumptions that I developed for the filtering.

At first, I thought about using linear programming to create a floor plan that optimizes the profit. While there was no issue with making this into an optimization problem, the only limitation was that I could not quantify how much gain I would get from putting two correlated products together. With the current dataset, we can calculate the probability that the person buying product A would also buy product B. However, we do not know how much that probability would improve as we move product A and product B closer to each other. Also, if product A and product B have a high lift, it would be better to put them apart so that shoppers can purchase something else on their way from product A to product B. All these complexities made my original plan that takes profit into account not implementable. Therefore, for the next step, it would be great to research more on finding ways to quantify the benefits of putting associated products together.