SKU Association Analysis Project Report

Yuxiang (Alvin) Chen

Executive Summary

Dillard's is a major retail chain with several stores. Their point-of-sales (POS) data over a period of time between 2004 and 2005 is accessible to us.The retailer is interested in rearranging the floors of the stores (planogram). Because of budgetary reasons, Dillard's would like our team to find the 100 items (SKUs) that are the best candidates to modify the planogram.

The team accessed the provided POS data between two time points in 2004 and 2005 and conducted data exploration to learn more about the features and basic statistics of the data. Then the team focuses on purchasing transaction records. The team made several logical assumptions based on real life situations and observed characteristics of data. Since the dataset is large, the team decided to take random 1% sample transactions from all transactions to study the customer purchasing pattern. The team applied association rules analysis with application of Apriori algorithm to find the associations with the highest Lift metric and good support and confidence values.

The team identified the first 101 SKUs involved in the top Lift associations as the recommendations for Dillard's (101 instead of 100 because the last association rule involved adding multiple SKUs to the recommended list, the list is sorted so if 100 is a hard limit, the last one could be ignored). The team also records the corresponding association rules involved and ranks them based on Lift metrics for Dillard's to conduct future plans and further analysis. Among the association rules that are covered by these 101 SKUs, the minimum Lift is about , which is very high so we believe these recommendations will help Dillard's in modifying the planogram. We also provided another 100 SKU recommendations which focus on high individual appearing frequency in the basket for Dillard's future plan. These SKUs will affect more baskets and also involve association rules with significant lift.

Problem Statement

Dillard's, a major retail chain with several stores, would like to adjust it's planogram. Our team would like to find the 100 items (SKUs) that are the best candidates for Dillard's to modify its planograms of stores.

Assumptions

1. The dataset documentation does not specifically clearly show which combination of columns together indicate an individual transaction. Thus, the team conducted data exploration and logically assumed that the unique combination of three data columns ('STORE', 'SALEDATE', 'TRANNUM') indicate a single unique transaction. The logic of this assumption is based on the following three sub assumptions.

a. No single transaction will be separated and recorded as different sales dates
b. No single transaction will happen in two different stores
c. 'TRANNUM' is an ID that is unique for transactions that happen on the same day in the same store.

2. The team assumed that whether a good is returned or not does not impact the purchasing association because returns are primarily caused by product quality issues.

3. The team assumed that a valuable association rule should have the set of SKUs be purchased together for at least 0.01% of all baskets (support >= 0.0001) and the confidence value should be at least 10% to be considered good.

Methodology & Intermediate Analysis

The detailed step by step process taken is clearly recorded in the project jupyter notebook file with codes as well. Here I will present the steps we took.

1. **Feature Labelling -** Since the original dataset is not well labelled. The team carefully learned about each feature through reading the data schema & limited documentation provided. Together with the observations from the dataset, the team marked each dataset with feature columns based on our best judgement. An example of how the primarily used transaction information datatable is labelled is attached in the appendix.

2. **Data Exploration -** The team found the dataset contains Dillard's transaction data between 2004-08-01 and 2005-08-27. In this time period, there are 332 stores with transaction records. 714499 different SKUs involved in the transactions. The team also made assumptions based on definition of features and logical reasoning to group rows into transactions. There are 9991819 transactions. After grouping, on average, there are about 11.2 item purchase record rows for each transaction, which is reasonable.

3. **Data Selection** - Through learning about features and data explorations, the team filtered out records for item returning and only focused on records for item purchasing. Since the dataset only spans about 1 year, the team decided to use the whole time span but reduce the number of transactions analyzed through random sampling. The team used 1% of all transactions (99918 transactions). The team focused on transaction information dataset because it has all necessary information for the association analysis.

4. **Data Grouping** - Since whether an SKU is in a basket is the focused information and the quantity does not affect association analysis. The team grouped the data by transactions and focused on what SKUs are involved in selected sample transactions.

4. **Association Rule Analysis -** To find out the customer purchasing bundle patterns, the team decided to apply association analysis, specifically with the Apriori algorithm. The team, based on logical reasoning, decided that an association rule is valuable only if the

support metric surpasses 0.0001 and confidence surpasses 10%. Support threshold is 0.0001 because there are 714499 different SKUs and 9991819 transactions. Thus, if a bundle of multiple items appear in 0.01% of all transactions, it is a non negligible item. 10% confidence and minimum lift of 4 are also a common practice for retail items association in the industry. The association rules found through the criteria are recorded in association_rules_df.xlsx. There are 1665 rules in total.

5. **Top 100 SKU Selection -** Since Lift is commonly considered as the most important metric because it indicates the ratio between confidence & expected confidence. Thus, the top 100 SKUs are chosen by looking at the association rules in lift descending order and selecting the first 100 unique SKUs. The team in the end found 101 SKUs because the last association rule considered adding in multiple SKUs. These 101 SKUs are presented in 100SKUs_recommended.xlsx. We found that SKU csv dataset available has a formatting issue in it so we could not check what each SKU's specifics.

Result Analysis

Among the 1665 association rules found, many have high lift value and also many have low support that is not very far from the 0.0001 threshold we set. We found that high lift rules also most involve high confidence. However, some of them have support that's close to 0.0001 (just good). This means that these item bundles do not appear frequently. This means some of the 101 SKUs we chose would not affect large amounts of transactions although the association is ensured to be strong. Thus there is a tradeoff between association lift and support here.

So, to address this potential concern, we also generated another 101 SKUs list with large lift and also each SKU involved has at least 0.001 individual support as an individual SKU. The minimum rule support is still 0.0001 after trial and errors. Moving SKUs from this list would affect a much larger number of transactions. We would like to provide this second list as well. Thus, Dillard's could choose the SKUs from both based on their business needs. This will provide a better picture for their next step.

Conclusion & Next Step

There are many association rules with significant lift and confidence with good support among SKUs. They could potentially provide large business value for Dillard's. These rules are identified and presented clearly in the excel. In addition, we identified an important tradeoff between choosing SKUs purely based on rule lift and SKUs that are present in more transactions. We provided our SKUs recommendations based on lift because strong association should be the primary drive.

However, based on the observed tradeoff, we also came up with another 101 SKUs list. In the next step, we would like to communicate with Dillard's further to learn more about their business needs and further study the margin of these SKUs to give better recommendations.