Hyung C. Park

Dillard Product Arrangement Project

Executive Summary

Dillard is a major retail chain with multiple stores. As major retail business that sells good directly to its client, the arranging of floors of the store could potentially have a large impact on the point of sales. To optimize their sales, it should take advantage of the data collected over point of sales at Dillard from mid 2004 to mid 2005. Analyzing the data could give an insight to customer buying pattern and also given an insight to how to arrange the floors in order to promote more sales.

Association is a rule-based analysis method more discovering interesting relation between variables in large data set. The success of this analysis relies on finding rules, and these rules are finding regularities between the products at Dillard. For example, if the customer buys a loaf of bread and peanut butter, then there is a high likelihood that the customer might also buy strawberry jam to make peanut butter and jelly sandwich or PB&J.

After cleaning and organizing a large set of data set provided by Dillard, the association rule based machine learning was used to identify a buying pattern for each transaction. Applying a few parameters based on the total revenue for selling each object, the items or candidates for re-arranging in the retail stores or in retail parlance to change the planograms.

Problem Statement

Dillard's data set contains information on each store, the product information, each individual transaction recorded at the stores for each item, and the department information. My job was to clean the data and extract the most useful data to find the candidates for changing the planogram at the stores.

Assumptions

- The subset of the dataset from the first five weeks from August 1st 2014 is used in the analysis for computation purpose, but it is assumed to be reprentative of all the times
- The data is representative of total of 332 stores, but I assumed that this data set will be will representative of all the stores
- The returned item plays a small part in the basket analysis, thus they could be disregarded for this analysis
- Confidence level of 20% and support of 500 is sufficient for the analysis

Methodology

Analysis of the data set involved three main components: understanding the data set stored in the database, cleaning up the data set, and finally running the association algorithms.

Before running the analysis, the important variables needed to be defined and understood. The most important feature of this analysis was to understand SKU, which is store keeping unit. Selecting the 100 SKUs as best candidates from over million SKUs in the system to modify the planogram means to chose the 100 best products that the Dillard store should move to maximize sales. The important features to note are the primary keys within the transaction table. The transaction table with over 120 million transactions outlines all the transactions of products for Dillard stores, and the primary keys that determine the transactions are SKU, Store, Register, Transaction Number, and Sale Date. Close exploration of the data set shows that transaction of each customer could be hypothesized very accurately by grouping by all the other variables except SKU. The SKUs could explain what the customers decided to purchase. So, the major task involved working with a very large data set that is stored in a remote setting.

While working with the data set, there were several problems that I encountered. First, the data set was too big to load in Python for association analysis, and also standard queries took extremely long time to execute, even failing due to time out connection from servers. In order to address this problem, a copy of the tables was made on my schema in order for modification. Second, the data sets in transaction table were all in character(255) types, which is very inefficient for the data stored in those tables. I was able to reduce the size of the table significantly by dropping features that were less important, and also changing the data types to small integers and dates. I further reduced the data set by dropping the return item transactions, since I was only interested in purchases and basket analysis. Finally, only the first five weeks of data were used for computational reasons, however considering that the data set consisted of one-year transaction data, around 10 percent seemed reasonable for analysis.

Using the data set, a new key was used to group SKUs into one transaction by the same person. A basket was formed using the following features: Store, Register, Transaction Number, and Sale Date. Using the orange3-associate python package, top rules were selected based on confidence level of 20% and support of 500. This was then used for choosing the 100 SKU candidates for modifying the planograms. Orange3-associate uses FP-growth algorithm for analysis.

Analysis

The rules from the association were outputted and examined. Among all the rules, the ones with the highest value for lift, confidence, and support were selected. This is very important because high lift confidence, and support corresponds to higher chance of customers buying more goods since they are more likely to buy large bundles of goods together. Based on the confidence level of 20% and support of 500 as minsup, there seems to be too many rules. Thus, there were some adjustment with higher level and support. Based on those rules, I normalized the support and confidence level, and found the rules with the highest sum of both. This corresponded with items and baskets that we are looking for, which are listed in the Appendix.

There were also several things to note about the Dillard transaction baskets. The rules demonstrated that there are so many goods that the minimum support of the entire dataset was small, mainly due to a lot of SKUs and the number of transactions. Also the selected SKUs had very high support, confidence, and lift values meaning that they good correlation. There is also a high chance that customers buying one item will trigger to buy another.

Conclusion

The association analysis is a very effective machine learning tool for analyzing large datasets. It allows analysts to group items into baskets and find attribute of the baskets that help identify items that generate most sales. Using this analysis, Dillard can potentially modify the planograms for the items listed in the Appendix to increase sales of their products by simply reorganizing.

Next Steps

Next steps of the analysis are doing a deeper analysis. Due to the time constraint and limited computational power, the association analysis made some big assumptions, and have not taken into account some factors that could be very influential.

First, the detailed analysis could encompass cost of items and retail sale value of the item in order to adjust the minsup. For example, electronics that create more revenue should have some priority in changes of the planograms compared to items like water and chips that might not generate as much revenue per sales.

In addition, there all the data could be used for analysis instead a small subset. Also, even if the stores are limited to modification of 100 SKU planograms, if the analysis could be run on individual stores, this could help maximize the revenues. For example, a store in the south could prioritize cold drinks and ice cream whereas the stores in the north could prioritize alcoholic beverages to fight the cold.

Finally, apart there could be a whole new set of analysis that could be performed in planogram modification. Given a specific planogram design, there could be analysis on where to move SKUs. Should SKU be moved towards the entrance or the corner of the store? Using the baskets and rules to estimate which items should be placed next to each other could also be very beneficial in increasing sales.

Appendix

1) Orange3-Associate is a open source package for association rule analysis The following link is the documentation for the package http://orange3-associate.readthedocs.io/en/latest/#

2) The dataset will all the results is attached in the csv files. Please read the headers for what each of the columns represent. Only the top 100 SKUs are displayed below:

```
[3582465,
 3844099,
 3013129,
 3968011,
 264715,
 9594893,
 4928011,
 9297426,
 6328344,
 8718362,
 173088,
 4296738,
 3864099,
 3988011,
 1310252,
 3346996,
 5509179,
 3053129,
 4008011,
 803921,
 4498011,
 2072671,
 7808101,
 9526376,
 7351914,
 2698353,
 4072567,
 7382655,
 3559555,
 3161221,
 1658506,
 8618636,
 3898011,
```

4218011,
3690654,
776350,
4008113,
2726578,
3868338,
5778109,
726718,
4976322,
39633,
9357022,
6420710,
9073382,
3611367,
4108011,
2288366,
6618353,
4112626,
6126322,
2783996,
5079809,
3631365,
6318344,
7313673,
3854099,
7967000,
3978011,
5528349,
7064350,
5036322,
3874099,
3998011,
8791356,
9402188,
348498,
3894099,
4440924,
1832285,
2688353,
9667426,
3968356,
4062567,
7261032,
3589483,
4138348,
2708353,

```
4208011,
5108107,
2716578,
3908011,
8798636,
9288109,
2366897,
9987000,
3524026,
2258366,
5957568,
29633,
2386897,
2571221,
9708505,
4992993,
3949538,
5978084,
9277426,
2784759,
6458364]
```