ASSOCIATION RULES LEARNING

Dillard's Point-of-Sales (POS) Data 2004~2005

James Jia, sjs8610, Northwestern IE308

EXECUTIVE SUMMARY

Dillard's, Inc., (NYSE: DDS) is an upscale American department store chain with approximately 300 stores in 30 states and headquartered in Arkansas. The 0.1 billion lines of transaction data is obtained from Dillard's point-of-sales (POS) data over the period between August of 2004 and August of 2005. POS data is usually (and true in this case) obtained directly through Dillard's department stores. It tracks every customer's every purchase, in order to render insights about the sales and customers.

Association rule learning is a rule-based machine learning method for discovering relations between variables. It discovers regularities (but not casualties) between products in large-scale datasets. For example, in the context of POS data, if there is a rule `{onions, potatoes} ⇒ {burger}`, it will indicate that if a customer buys onions and potatoes together, it is very likely for them to also purchase burger meat. In this analysis exercise, such a rule becomes very important as it can be used to identify best candidates to modify planograms (the store floor arrangement). Doing so may lead to promising business returns based on business intention. For example, arranging closely associated items together might provide customers with greater convenience, while arranging closely associated items far apart might provide more opportunities for customers to walk around and purchase other items on the way.

During the analysis process, we first identified the unique `basket ID` (the unique identifier of the list of items a particular customer purchased in 1 sitting). Afterwards, we conducted smart data reduction to reduce the data to-be-analyzed from 0.1 billion rows to around 700K rows. After further data processing and one-hot encoding, association rule algorithm is applied to the dataset to determine the best candidate SKUs to be recommended.

Taking a brief look at the results, it is determined that a combination of `store number, register number, transaction code, sale date` should be used together to serve as the unique identifier. Among the 500+ association rules generated, the mean support is around 0.19%, confidence is 34% and lift is 120. Interestingly, out of the top 100 SKU candidates selected (based on lift value), most of them belong to the *Clinique, Cabern, Mai, Celebrt, NOB, Frederi and LouisVL* departments. The retail price ranges from $70 to a few dollars.

PROBLEM STATEMENT

Dillard's is a major department store retail chain with dozens of stores across the country. The point-of-sales (POS) data is available from 2004.8 to 2005.8. The retailer is interested in rearranging the floors of the stores, a.k.a., changing the planograms. Due to limited manpower, they can only move around 20 SKU. This analysis aims to find the 100 SKUs that are the best candidates to modify the planograms.

METHODOLOGY

Unique Basket ID:

One of the first tasks to be undertaken is to understand the transaction data. In order to conduct association rule learning algorithms, it is important to identify the basket identification number. Because the transaction data is organized by the purchase of line items (such as apples and oranges), we need to first find a unique basket ID (or a combination of data columns) that can be used to identify the list of items that a particular customer purchased in one sitting. Mainly trial-and-error method was used, providing the information that items from the same basket must come from the same `store` and `saledate`. We combined those 2 info with either `SEQ` (purchase sequence number), and `trannum` (transaction number), and realize that neither can yield ideal results. For example, with reference to Table 1 below, if the basket ID is a combination of `store, saledate, trannum,` return items are also included together with purchase items, which cannot be the case.

	sku	store	register	trannum	intID	saledate	stype	quantity	am1	am2	am3	SEQ	mic
188340	66668	202	610	1300	862300112	2005-01-18	P	1	22.00	22.00	22.00	579400042	15
262153	187830	202	610	1300	862300112	2005-01-18	P	1	14.00	14.00	14.00	247500056	15
31449	236370	202	250	1300	0	2005-01-18	R	1	79.50	79.50	79.50	587900035	205
484476	1566080	202	120	1300	0	2005-01-18	P	1	78.00	39.99	39.99	500900030	601

Table 1

After several other attempts, it was discovered that a combination of 4 columns of information:`store, register, saledate, & trannum,` provides an accurate basket ID, as seen from Table 2. `(bkid)`

```
trn[(trn['store']==202) & (trn['saledate']=='2005-01-18') & (trn['register']==610) & (trn['trannum']==1300)]
```

	sku	store	register	trannum	intID	saledate	stype	quantity	am1	am2	am3	SEQ	mic	n
188340	66668	202	610	1300	862300112	2005-01-18	P	1	22.0	22.0	22.0	579400042	15	0
262153	187830	202	610	1300	862300112	2005-01-18	P	1	14.0	14.0	14.0	247500056	15	0
280220	3440673	202	610	1300	862300112	2005-01-18	P	1	22.0	22.0	22.0	235000005	15	0

Table 2

Data Exploration:

After obtaining the unique basket ID, data from all tables was briefly explored. Python package `DASK` was used in order to load the large datasets (over 10GB) without putting significant strain on the laptop memory. Summary of each dataset can be found below:

- `trnsact:` In total 120 million rows of data. People on average purchase 1 per item, and the average price of item purchase is around $2.50;
- `skuinfo:` In total 1.6 million SKUs are present in the system;
- `deptinfo:` In total 60 departments are present in the system;
- `strinfo:` In total 453 stores are present across the country, with Florida and Arizona having the most number of stores (20+);
- `skstinfo:` Different stores have different retail pricing schemes. The cost of any SKUs is on average $2.40, and the retail price is on average $4.30.

Data Reduction:

Much effort has been put into reducing the existing transaction data into a locally manageable size, in order to conduct further association rule learning. In particular, because it is known that the number of columns in the final dataset to be fed into machine learning will be over 100 (because at least 100 SKU candidates have to be recommended). Therefore, the goal of the number of rows for the final dataset is on the scale of 10^4.

The following steps were taken sequentially in order to carefully subset a logical and manageable dataset:

1. Assuming that people from different geography might have different purchase preferences, and each store will likely have different floor plans, the data was first subset to only include stores to a state-specific level. In this case, the state of Illinois (with 3 stores) was chosen. `[120M → 0.75M]`
2. As we are more concerned with the customers' purchase habits, only purchase data (indicated by "P" in the "stype" column) is included in the dataset. `[750K → 600K]`
3. As association rules mainly concern with items in a single basket, meaning, "what will a customer likely to purchase given the items already in their basket". Hence, baskets with only 1 item (single-item purchases) were removed from the dataset. `[600K → 450K]`
4. Based on the dataset, there are in total around 120K unique SKUs. The goal of this analysis is to find around 100 candidates. Hence, we ranked the number of occurrences for each SKU, and picked the top 500 SKUs that are most frequently appearing in customers' baskets. Based on analysis, in order for an SKU to be retained, it has to appear in at least 50 baskets. `[450K → 42K]`
5. Finally, the dataset is again cleaned for single-item baskets and remove any duplicates in the dataset. The dataset is now in the desired size. `[42K → 28K]`

One-hot Encoding, Basket Agglomeration, Association Rule:

With the reduced dataset of the desired size, we can formally proceed to apply one-hot encoding so that each column indicates an SKU, and then collapse the dataset so that each role represents a single basket with columns indicating the items purchased. The basket-level one-hot encoding result is shown in Table 3:

bkid	sku_39171	sku_39633	sku_47775	sku_68086	sku_98327	sku_106343	sku_108507	sku_112307	sku_136343	sku_152307	...
2512	0	0	0	0	0	0	0	0	0	0	...
5370	0	0	0	0	0	0	0	0	0	0	...
10443	0	0	0	0	0	0	0	0	0	0	...

Table 3

Finally, the dataset went through an association rule algorithm based on `mlxtend.frequent_patterns.apriori` (with `minsup` = 0.001) and `mlxtend.frequent_patterns.association_rules` to generate the results.

ANALYSIS OF RESULTS

Direct Result Summary:

For the data in final analysis, in total 11,695 baskets with in total 500 SKUs were analyzed. On average, each basket contains 2.4 items. However, baskets in this dataset have at most 16 items.

A summary of the association rule results can be seen below in Table 4.

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
count	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000
mean	0.013900	0.013900	0.001874	0.344452	119.808005	0.001590	inf
std	0.014458	0.014458	0.001222	0.289996	163.111273	0.001118	NaN
min	0.001112	0.001112	0.001026	0.018927	1.066997	0.000064	1.001937
25%	0.002565	0.002565	0.001197	0.087075	4.333996	0.001013	1.068763
50%	0.007995	0.007995	0.001454	0.237692	29.247762	0.001361	1.290668
75%	0.021035	0.021035	0.001967	0.584858	243.281569	0.001756	2.400837
max	0.054211	0.054211	0.009320	1.000000	830.414201	0.008649	inf

Table 4

As seen from the table above, `support` metrics for the association rule results are on average 0.19%. This means that, for the SKUs in one rule, on average their frequency of appearances among all samples is around 0.2%. The highest support is reaching 1%.

Additionally, the `confidence` metric is on average 34%, and goes as high as 100%. This means that for a rule such as {A} ⇒ {B}, the probability of occurrence of B in the same basket as A, given A is already is the basket -- the probability is on average 34%.

Finally, `lift` statistics show the strength of dependence for the items in the rule. As independence will yield lift = 1, a mean lift of 120 for our association rule results showcases a strength in association with our results. However, the degree of variance in lift statistics should also be considered. (s.d. ~= 1.5*mean)

In order to obtain the best 100 SKUs, the association rule results obtained from above are ranked by lift. Then, sequentially, each SKU is recorded in order of their lift with removal of duplicates. The resultant list will thus have the SKU with the highest lift in a rule it involved as the first item, and the SKU with the lowest lift in a rule it was involved as the last item. An example can be seen below in Table 5.

Using the method described above, SKU_4456421 and SKU_4766421 will be the first 2 items that get placed in priority to the recommendation list. It is possible that an item will appear in multiple rules, in that case, the order of its first sequence will be the one that gets recorded in the list.

	antecedents	consequents	support	confidence	lift
304	(sku_4456421)	(sku_4766421)	0.001026	0.923077	830.414201
305	(sku_4766421)	(sku_4456421)	0.001026	0.923077	830.414201
410	(sku_8616048)	(sku_8956048)	0.001454	0.944444	613.626543
411	(sku_8956048)	(sku_8616048)	0.001454	0.944444	613.626543
360	(sku_7612182)	(sku_6319962)	0.001197	0.700000	584.750000

Table 5

Recommendation List Result Analysis:

Based on the recommendation list, a brief analysis of the SKU candidates was conducted in order to see whether there is any perceivable trend or features.

Exploring the department at which those 100 SKUs belong to, it was surprising to find out that all 100 SKU candidates really just belong to 7 departments: Clinique, Cabern, Mai, Celebrt, Nob, Frederi and LouisVL. Considering there are over 60 departments, such distribution is quite interesting, and worth finding out in the next step what will be the common characteristics of these departments.

Exploring the pricing scheme for those candidates, the mean retail price of these candidates is $6.18, while the cost is $3.95. However, there are also outliers that are in the high range of $70 as retail price. Although more investigation is needed, it's suspected that lower price daily commodities are more often purchased, and sold as buy-along items.

LIMITATIONS

Over the course of the project, several limitations are also found. For example, the identification of basket ID can be more specific or scientifically thorough. Otherwise, more background research should be conducted in order to pick the correct basket ID.

Additionally, due to computational limitations, more data should have been included in the dataset to be analyzed through association rule algorithms. More scientific methods can be used in order to determine the sequence of subsetting via elimination.

Finally, sometimes different items have different occurrences in POS, but it doesn't mean that their significance to Dillard's is solely determined by their number of occurrences. Therefore, in the future, weighted association algorithms can be explored as a better alternative.