Ta-feng Dataset Analysis

Contents

Introduction

- Ta-Feng is a grocery shopping dataset released by ACM RecSys, it covers products from food, office supplies to furniture.
- The dataset collected users` transaction data of 4 months, from November 2000 to February 2001.
- The total count of transactions in this dataset is 817741, which belong to 32266 users and 23812 products.
- Store located in Taipei.

Overview: Data Quality

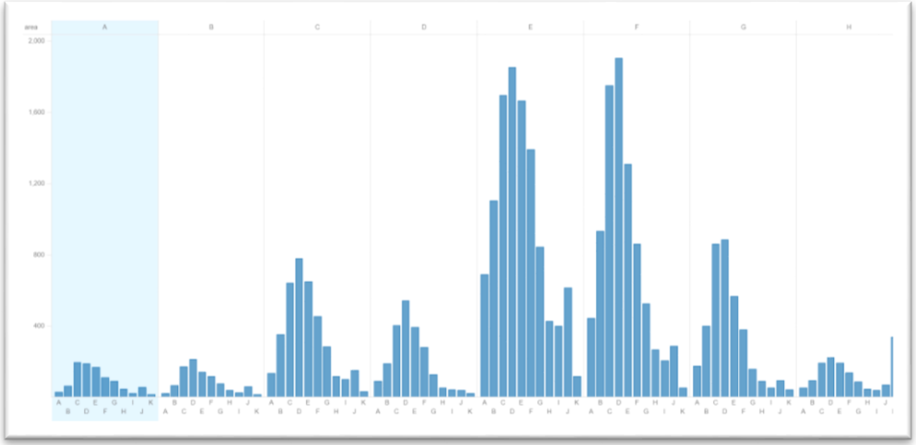
Evaluation of Data Quality

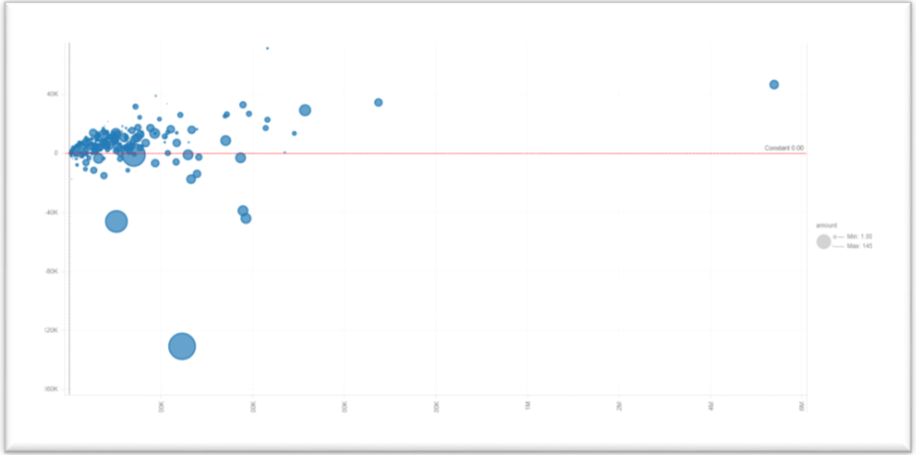
- No NULL in the table
- Since the grocery shop knows the product information, the unreliable data usually appears in customer side.

  For example:

  - One customer ID who purchased a lot may be an ID which is used to identify unknown customer.
  - Customers with other age information may be assigned “J” (>65) as their ages.
  - Too many “H” (Unknown Distance).
- If one customer has multiple transactions on one day, this table cannot indicate this situation.
- One product may belong to multiple subclass.

Overview: Descriptive Analysis





Overview: Descriptive Analysis

Overview: Descriptive Analysis

Top 10 Recorder Products
4714981010038
4711080010112
4711271000014
4710094097768
4710105015118
4710011401128
4710114128038
4710018004605
4719090900065
4710088433312

- Products with Top 5 Frequency in different areas and age groups

Top 5 Product	A	B	C	D	E	F	G	H
1	4714981010038	4714981010038	4711271000014	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038
2	4711271000014	4711271000014	4714981010038	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014
3	4719090900065	4710114128038	4710114128038	4710114128038	4719090900065	4719090900065	4711080010112	4711080010112
4	4710054380619	4713985863121	4719090900065	4710291112172	4711080010112	4711080010112	4710114128038	4713985863121
5	4710291112172	4719090900065	4710421090059	4719090900065	4710265849066	4710114128038	4710421090059	4710011401128

Top 5 Product	A	B	C	D	E	F	G	H	I	J	K
1	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038	4714981010038
2	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014	4711271000014
3	4711080010112	4711080010112	4710088410610	4711080010112	4710114128038	4719090900065	4719090900065	4719090900065	4719090900065	4710265849066	4713985863121
4	4710088433312	4710032501791	4710054380619	4710088410610	4719090900065	4710114128038	4710114128038	4710583996008	4710265849066	4719090900065	4710011401128
5	4710421090059	4710088410610	4710088410139	4710114128038	4713985863121	4713985863121	4710583996008	4711080010112	4710583996008	4710583996008	4711080010112

Promising Utilisations

- Retailer:

  (1) Association rule, if two things appear together, we can make them far from each other or closer to each other;

  (2) Recommendation system for weekly product advertisement;

  (3) New product: automated basket filling for online grocery shopping.

  (4) Customer segmentations

  (5) Price elasticity and discount evaluation

- Consumer:

  When will the price decrease on certain items? (However, consumers usually don’t have historical price information.)

- Wholesaler:

  Supply and demand balance. When should the wholesale price increase? (not my personal interests, and can be solved by universal approaches, time series prediction.)

Two Draft Models for Retailers

- I believe there are two groups of customers:

  (1) customers who dislike wasting time on grocery shopping, (like me, usually do it online)

  (2) customers who enjoy searching goods in supermarkets.

- Each group needs a product to meet their lifestyles.

- “Automated basket filler” helps the grocery shopping haters quickly fill the online shopping cart with the items they frequently purchased.

- “Personalised newsletter” helps the grocery shopping lovers explore other items they may like.

State-of-the-art works

- Retailer:

  Recommendation system:

  Sato, M., Izumo, H. and Sonoda, T., 2016. Model of Personal Discount Sensitivity in Recommender Systems. *IxD&A*, *28*, pp.110-123.

  Next basket prediction:

  Rendle, S., Freudenthaler, C. and Schmidt-Thieme, L., 2010, April. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web* (pp. 811-820). ACM.

  Yu, F., Liu, Q., Wu, S., Wang, L. and Tan, T., 2016, July. A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 729-732). ACM.

  Customer segmentations

  https://www.slideshare.net/jonsedar/customer-clustering-for-marketing

Next Basket Prediction

Simplify the problem and find a quick workable solution:

- I try to predict which items that a consumer bought in the previous two transactions will appear in the current transaction.
- The last transaction of each consumer form the testing set, the other transactions form the training set.
- Remove consumers who made less than 4 transactions and do not make any reorder in the training set.
- 5235 out of 32266 consumers are left.
- Very imbalanced dataset. Only 5.4% products in the previous two transactions have been reordered in the current transactions.

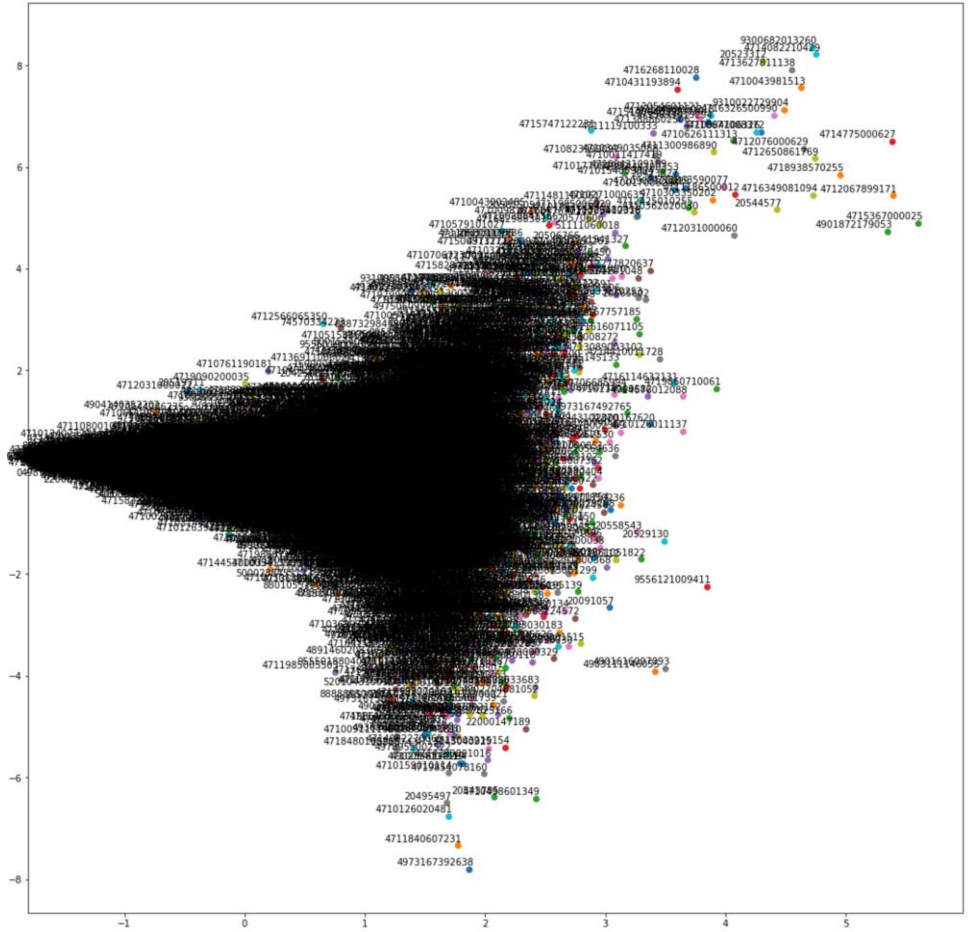
Next Basket Prediction

Feature preparation:

- For each product bought in the previous transaction:
  - Bought amount in the previous transaction
  - Bought amount in the transaction before the previous transaction
  - Time difference between current transaction and the last one in days
  - Time difference between current transaction and the transaction before the last one
  - Price difference…
  - Customers' location and age information
  - ….
- One Hot encoding for categorical feature
- For categorical features that have too many categories, such as the product ID and subclass, I use word2Vec to represent them as 20*1 vectors (since I plan to use XGBoost, the feature dimension cannot be too large).

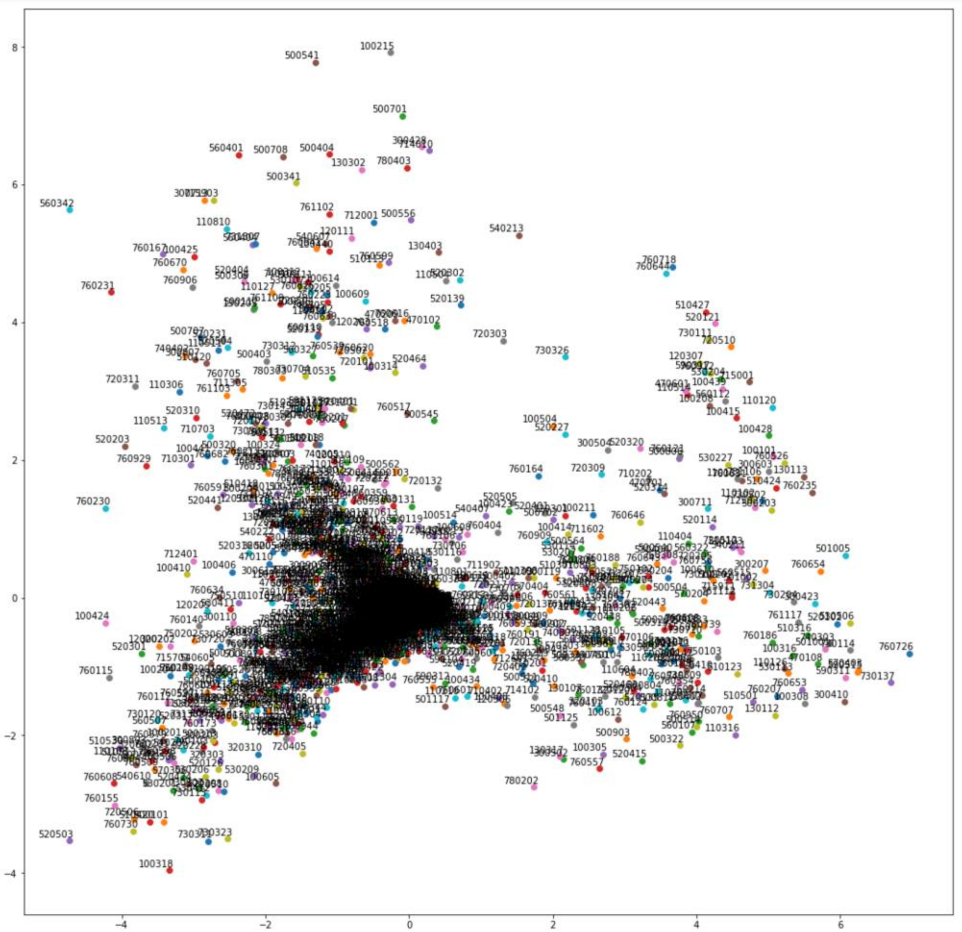
Next Basket Prediction

Visualise Word2Vec representation of Product ID



Next Basket Prediction

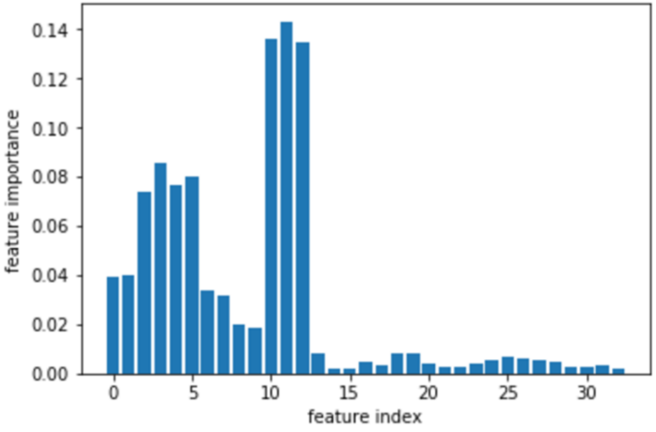
Visualise Word2Vec representation of Subclass



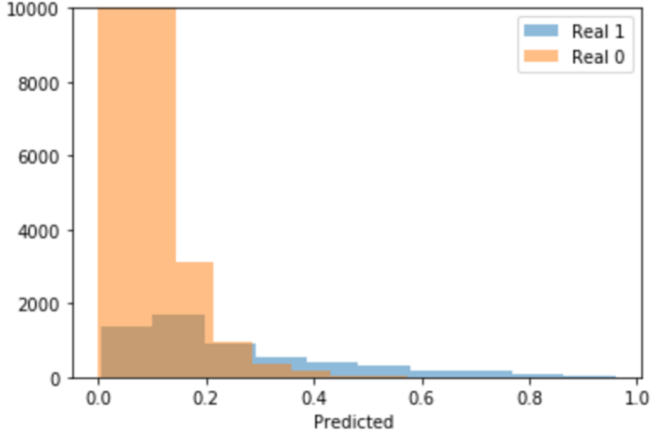
Next Basket Prediction

index	name
0	last_weekday',
1	'last_weekday_2',
2	'diff_last_day',
3	'diff_last_day_2',
4	'last_price',
5	'last_price_2',
6	'last_amount',
7	'last_amount_2',
8	'last_month',
9	'last_month_2',
10	'subcls',
11	'prod_id',
12	cust_id'

Feature Importance



Prediction Results



Confusion matrix

	Predict 0	Predict 1
Real 0	75045	2580
Real 1	2807	2877

Recall: 0.506,
Precision: 0.527,
Accuracy: 0.935,
F1 score: 0.516

Recommendation System

More sophisticated recommendation algorithms have been developed and evaluated by using Ta-feng dataset.

I just use the basic collaborative filtering method.

I compute the similarity between products and recommend to consumers.

Definition of rate:

$$\text{Rate} = \frac{\text{the number of time a consumer bought this product}}{\text{number of transactions of this customer}}$$

Evaluation:

Removed 1000 rates from the matrix. Use the recommendation system to predict the rates.

Absolute error is 0.0887. (average rate is 0.333)

Association Rule

Association Rule on Products

Support level 0.005

{'4710011401128', '4710011401135'},

{'4710011401128', '4710011405133'},

{'4711271000014', '4714981010038'}.

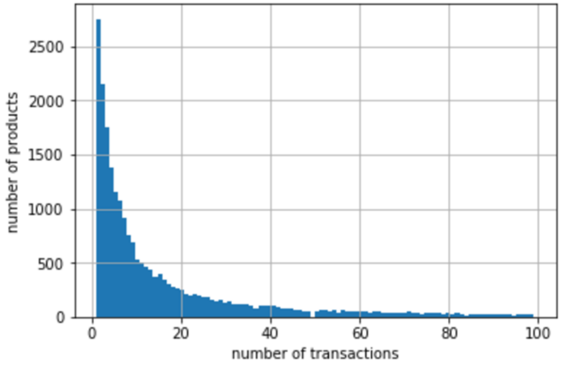
Confidence level 0.3

{'4710011401128'} --> {'4710011401135'} , conf: 0.42761148442272445

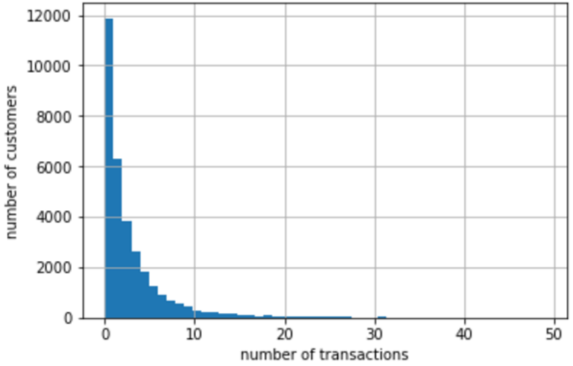
{'4710011401135'} --> {'4710011401128'} , conf: 0.7526881720430108

{'4710011405133'} --> {'4710011401128'} , conf: 0.6595744680851064

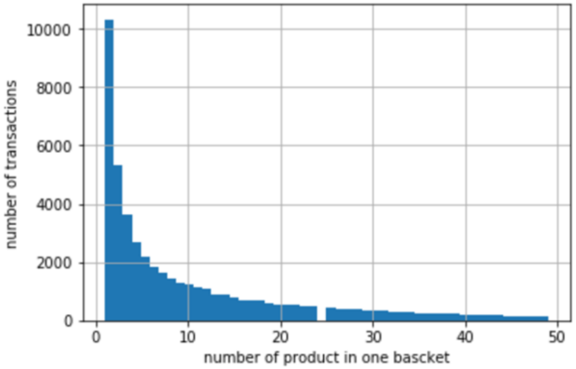
{'4710011401128'} --> {'4710011405133'} , conf: 0.37874160048869887



Distribution of product
vs. number of transactions



Distribution of customer
vs. number of transactions



Distribution of transactions
vs. different basket sizes

Association Rule

Association Rule on Subclass

Support level > 0.01

{100102, 100205},

{100205, 100505},

{100205, 110411},

{130204, 130315},

…

32 pairs

Confidence level > 0.3

{100312} --> {100205} , conf: 0.44585236481508767

{530103} --> {530101} , conf: 0.4327076041998091

{500203} --> {500201} , conf: 0.3620643069440692

{100201} --> {100205} , conf: 0.3576039633688635

{560402} --> {560201} , conf: 0.3194888178913738

{100323} --> {100205} , conf: 0.3698943159097401

Conclusion and Discussion

- Very interesting dataset capturing the transactions in the special time period
- Need more data to split this table to small ones
- More time to make something really useful