University of Nottingham Ningbo China

Faculty of Business

Academic Year 2019/20 Semester

Analytics Specializations and Applications
(BUSI4392 UNNC)

Lecturer: Dr Jenny Xiaodie PU

Students: Mingyuan Zhu
20219292

Table of contents

Section 1: Executive Summary for business tasks 3

Section 2. Preprocessing and descriptions 4

Section 3: Data exploration and feature engineering.....6

Section 4 Clustering into segmentation and visualizations.....8

Section 5: Recovering segments and create profiles.....9

Section 6: Recommendations.....11

Appendix: Jupyter Notebook and Excel

1、 Executive Summary for Business Task

To be more profitable, sellers have attached great importance to understand buyer behaviour as a fundamental of marketing. The selling processes for purchasing products in the supermarket can be abstracted as the buying decision of groups of customers who have a similar need for products. Meanwhile, not only customer groups have different willing and buying abilities, but most products market to particular small segmentations. Targeted marketing research has made some progress, but what is effectively experiment and how to serve a specified group of people is still a severe issue for traditional business analysts. One mainstream machine learning testing for retailers for solving this is trading up or down the qualities of the commodity price from seasons to seasons, which considered to be an inefficient, original, time-consuming and expensive strategy.

This report aims at analyzing the content of a shopping database with more robust multiply methods of 3000 customers purchasing records over a half year (from March 01, 2017, to August 31, 2017). After data preprocessing, here we set customers as our minimal analysis target running the PCA, LDA and NMF to reduce dimensions and identify the most import features in protein segments. We then compared the customer segmentation results of both K-means and Gaussian Mixture model interpreting the character of each customer, and evaluate whether it is useful. In the end, total customers divided into five segmentations with pen portraits, like low-income groups and shopaholics. For elements of 3000 customers, we found that spend and purchasing quantity of each category have enormous variance. Meanwhile, statistical analysis for market segmentation provides some insights for profitability correlates with segmentation in selling, and we also offer some useful test hypothesis methods for further profound research like segmented A-B test

2、 Preprocessing and descriptions

The data structure and features we used for the following analysis shows in figure 1 below. After loading the four data frame from original CSV format files, we first apply data preprocessing steps including data cleaning, editing, reduction and selection. Data cleaning steps removing the " £ " sigmodal for all spending categories, while data editing step has edited formats into the correct data types. We also created new features, like "part_time" (counts for shopping behaviour of four divided periods, from morning to evening), "time_interval" (Recency of RFM model, shows data trunk of last purchase for each data

point) ， "total_category" and" average_category" from "baskets_sample". Data reduction steps only kept the new features we created in "basket_sample". Fortunately, there is no outliers and NULL values in the data, so it doesn't need to drop it. Finally, we merged all the data frames(3000* 32) including RFM featured and customer behaviour feature like product-based spends as well as temporal characteristics by primary key into final training set for next feature engineering steps.

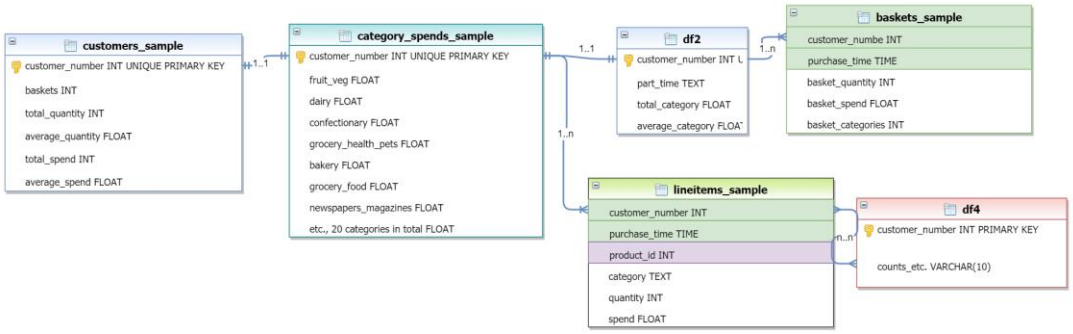
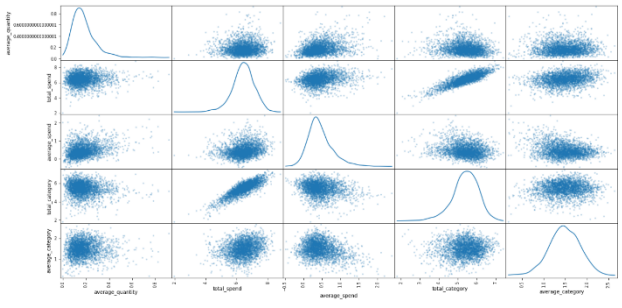


Figure 1: Data Relationships

3、 Data Exploration and Feature Engineering

As vital tools, correlation analysis is the quite common multivariate analysis and feature selection methods used in reducing dimensions. We first investigate and establish a relationship between 33 categories. By drawing the brief statistics summary in table and correlation plot in the heatmap, we found that "total_quality" and "basket" features are strongly correlated to the "total_spend" features. There exist many strongly correlated features, so next, we use the PCA, NMF, LDA methods to reduce the dimensions and correlations before statistical modelling and data analysis.

As hist graph shows, data not generally distribute with a massive variety of both mean and median. So here we applied natural logarithm first and replaced the 0 data to a specified tiny number. Then we apply the PCA to eliminate correlated feature and select the principal components we used for next segmentations. For the convenience of visualization, here we only keep the second two dimensions which can explain over 31% variance



of our data in table 1, and apply PCA transformation to assign the two PCs as our reduced data.

PC 1	PC 2
0.1893	0.3141
0.58 tobacco	0.29 tobacco
0.53 cashpoint	0.26 cashpoint
0.39 lottery	0.12 lottery
0.28 evening	0.03 customer_number
0.20 drinks	0.01 average_spend

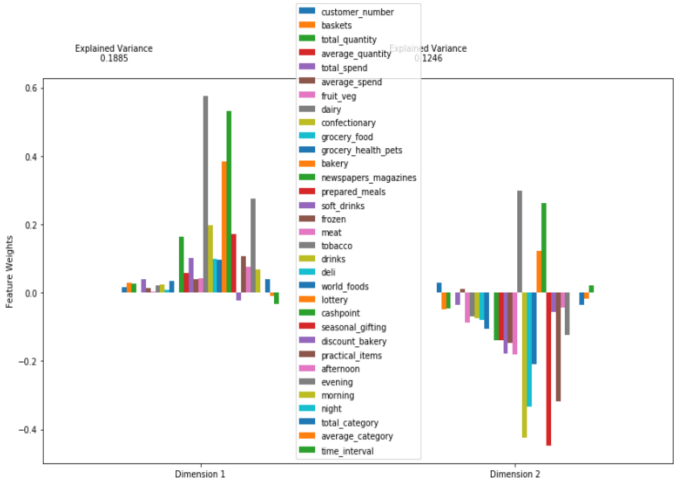


Table 1& figure 2： PCA for its interpreted dimensions(2 dimensions)

A definite increase in proportion corresponds with the positive-weighted rising and negative-weighted decreasing in features. For PCA, first dimensions show its interpretation with ' tobacco', 'drinks', 'lottery' and 'cashpoint' categories, consistent with previous scatterplot analysis. These categories are relaxing goods which shows the behaviour for preference. The next steps will prove the variable's contribution to group division and describe customer behaviour in all aspects.

We also compared the NMF and LDA results. We found that for the first clusters in these two methods, customer id contributes a lot for first dimensions, which is meaningless for interpreting. So here we conclude that PCA performed better and pickle it instead of LDA and NMF for the next analysis.

4、 Clustering into segmentation and visualization

Finally, we create our segments by K-Means algorithm and Gaussian Mixture models to identify the various customer segments hidden behind the information. We expect to have 5 to 7 number of clusters depending on the business problem, and the precious number would be made on features' structure hidden in the data. We test the group from five to seven on both GMM and K-means and calculating the silhouette coefficient for all data points. Silhouette coefficient evaluates each divided cluster performance for each position and calculates means intra-cluster distance as well as the nearest-cluster distance for each sample. Calculating the

mean silhouette coefficient is regarded as the simplest scoring method to determine the optimal number of segments. For the GMM and K-means silhouette score from figure 3 shows, they all have the highest ratings in a cluster of seven, and K-means performed better than GMM with higher optimal scores and more stable variance. Except for the calculation speed in two algorithms, it's recommended to use the GMM algorithm because GMM can output the probability that the data points belong to a particular category, so the output information is much more abundant than the K-means algorithm. As another criterion for optimal clusters, Bayesian information criterion (BIC) also has the lowest score to 7 groups which is consistent with our previous analysis using the silhouette score. All in all, seven groups was reduced to 2 dimensions with a mean silhouette score is 0.3410, and the BIC is 40343, which means we should be more focusing on it.

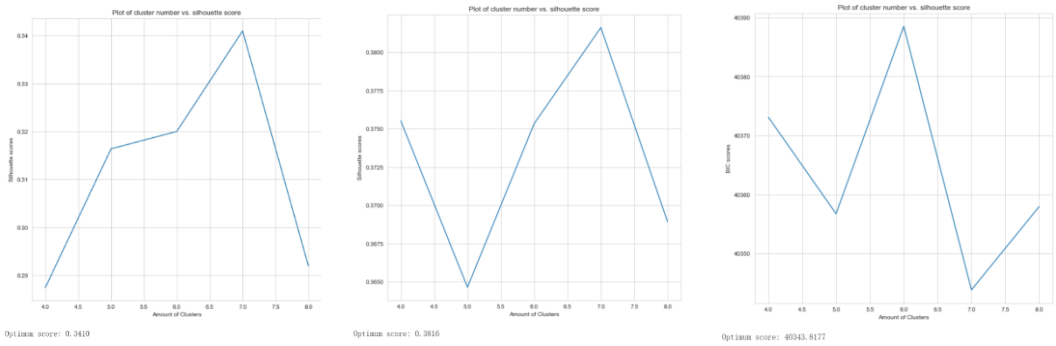


Figure 3: GMM and K-means silhouette score and BIC score

Here we pick the highest score of clusters equals to 7 in GMM models. As figure 4 shows, the bunch 1 to 7 have 516, 482, 385, 542, 600, 222,253 customers, and each of them can be divided into 2D dimensions well.

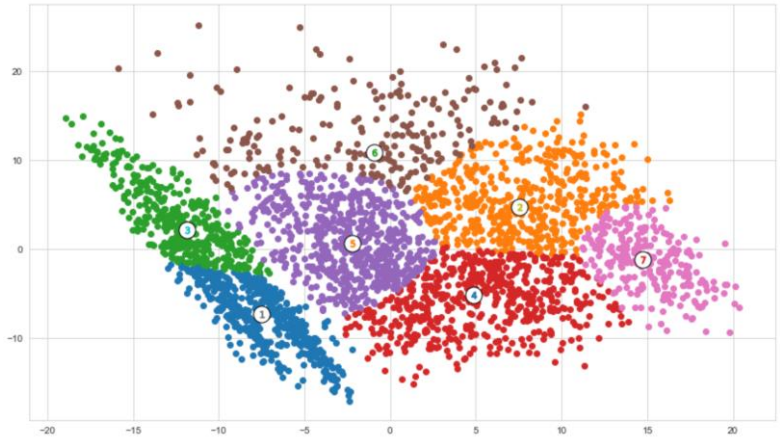


Figure4: Visualize the GMM results in a scatterplot of 2 dimensions

5、 Recovering Segment and Analysis for Profiles

First, we are recovering segment archetypes to original variables from two PCA dimensions. For all customers, we can see that some features in each group are equal to mean values like the average quantity is all the same of 1 item. They hardly buy the discount bakery, and they are seldom shopping at night. Therefore, a customer of those low-value goods would contribute small proportions for total spend. As one of purchasing preference, it has significant variance in the classification of clusters 2 and 5. We also notice that 50% or less customer spends 726 £ for these periods, while 75% of customers pay 957 Pounds or less. There is a massive gap between the most frequent buying and maximal buying decisions. For RMF analysis in 3000 customers, it performed not well, although they have many big spenders, the best presented and lost cheap customers still count for great propositions.

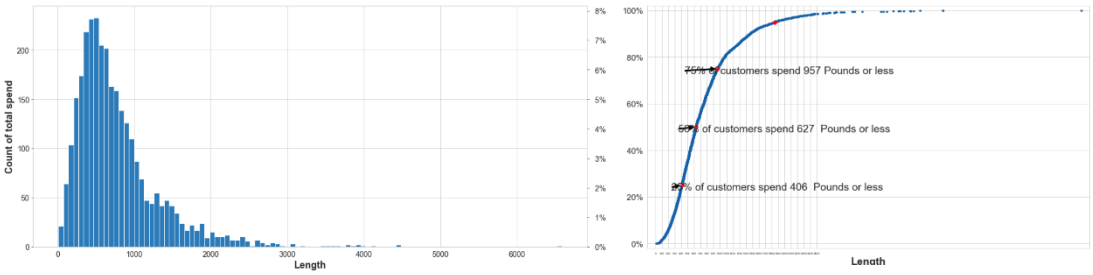


Figure 5: The hist gram for customer total spend

The centres represent the median values of each segmentations for different groups. Here produce profiles for each segment and individual statistical summaries as well as pen portrait of seven clusters. The colour depth corresponds to the size of the variable in the following heat maps.

Figure 6： Centres for each cluster

Next, we divided the customer segments with RFM Models and assigned the score from 1(lowest/worst) to 5(best/highest) to Recency, Frequency and Monetary. For segmentation classes and quartile of those tree features, high Recency means performed bad, while high frequency and monetary value are of great value. A final RFM score is calculated simply by combining individual RFM score numbers. So the total 3000 was also divided by the key RFM segments as following index{ RFMScore =="444":return "Best Customers", RFMScore== '244': "Almost Lost" RFMScore=='144': "Lost Customers", RFMScore=='111': 'Lost Cheap Customers', RFMScore=='X4X'"Loyal Customers", RFMScore=='XX4': "Big Spenders"}.

Best Customers:	**Loyal Customers:**	**Big Spenders:**	**Almost Lost:**	**Lost Customers:**	**Lost Cheap Customers:**
53	597	600	35	28	236

Table 2:RFM segmentations of 3000 customers

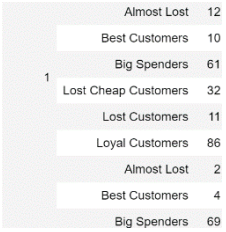
4	Big Spenders	75
	Lost Cheap Customers	7
	Lost Customers	9
	Loyal Customers	118
	Almost Lost	3
	Best Customers	12
	Big Spenders	55
	Almost Lost	6
	Best Customers	5
7	Big Spenders	28
	Lost Customers	1
	Loyal Customers	59

Both cluster 4 and 7 have higher values in spending and quantities with vast values of standard deviations and most of their customers with devotions. But for details, they are so different. Cluster 4 spend most on all kinds of meals, such as fruits, vegetables, drinks, prepared snacks as well as world foods. Their behaviours can be associated with full-time homemakers. On contrast, cluster 7 has a lower time interval from last shopping. This cluster prefers buying more relaxing goods like tobacco, newspapers and magazines. For shopping times behaviour compared with other groups, they have preference shopping in the afternoon and evening. This groups located at the shopaholics, expertly, who are busy working and want to shop after work to relieve stress.

3	Big Spenders	17
	Lost Cheap Customers	78
	Lost Customers	2
	Loyal Customers	36
	Almost Lost	10
	Best Customers	18
	Best Customers	3
6	Big Spenders	21
	Lost Cheap Customers	41
	Loyal Customers	21
	Almost Lost	6
	Best Customers	5

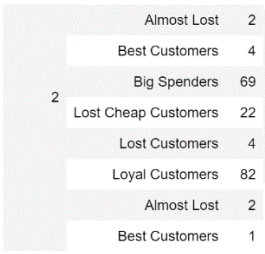
Cluster3 have the lowest basket quantity and spend among the seven groups. Except for their spends on food like fruits, vegetables and grocery foods, they hardly spend money on other categories. It contains most of the lost cheap customers. This groups with higher Wenger coefficient and belongs to the low-income group. Like cluster3, cluster 6 also have lower values in spending- most of the categories are always with little cost. Customer of segmentation six would fit with people who might shop little offline with balance categories, and perhaps they

spend very little money on this store (for example, frequent online shopping and shopping in different stores).

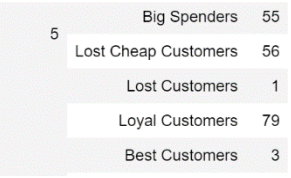


1	Almost Lost	12
	Best Customers	10
	Big Spenders	61
	Lost Cheap Customers	32
	Lost Customers	11
	Loyal Customers	86
	Almost Lost	2
	Best Customers	4
	Big Spenders	69

From statistic summary table, we can see that cluster 1 have a vast gap between min and max in spend and buying times. Group 1 is similar to cluster 4, but it has fewer categories every day. So the brand of bunch one can be described as the people prefer tradition foods. RFM analysis also shows this group consists of varieties of customer features; it's a varied group.



2	Almost Lost	2
	Best Customers	4
	Big Spenders	69
	Lost Cheap Customers	22
	Lost Customers	4
	Loyal Customers	82
	Almost Lost	2
	Best Customers	1



5	Big Spenders	55
	Lost Cheap Customers	56
	Lost Customers	1
	Loyal Customers	79
	Best Customers	3

Cluster 2 and cluster 5 have the median shopping counts and values among seven clusters. It's hard to distinguish those two groups from the previous 2D dimensions in PCA and centries, both with molecular size and steadies bunch preference. This two cluster is the typically traditional shoppers, without particular choice and shopping at the mean level. But group 5 performed severely on the RFM analysis, for its large quantities of lost cheap customers, which is of little value for retailers.

6 Recommendations

The recommendation is as follows. Based on previous analysis, here are some advice for marking:

1: A/B testing, also known as bucket testing, would be useful for companies to find the best marketing strategies of each segmentations. One of segmented A/B testing is to make small changes from currently categories and discount of goods to more attractive ones for each group, observing whether changing effect is negative or positive, and only make changes that react positively for customers. For the design of the test, we first can focus on the that five most significant segments, and create some hypothesis as to what would be work for each cluster. It's quick and more meaningful to start from large quantities of people, and then lets to granular tests for more specific segmentation.

For test intuition of our pen profiles, an A/B test would be performed as follows: If we want to improve the serving and delivery schedule of cluster4, which is of many big spenders, 20 per cent customers selected as samples for the service change. The remaining members of the

cluster would be the control group. Satisfaction surveys from each section pickled to determine the effect of the delivery schedule and serving changes on each of them. We can also set up the tracking on email market conversion and identify a goal for this on both RFM and GMM segments, to evaluate whether it is useful. The A/B testing would also be applied on an email campaign, where the control group receives a regular advertising version and the experiment group get the clustering tailored email for discounts, and analyze results of loyalty with spend and iterate on their performance.

2: Once the marketers have the precious pen portraits of those customers, they can relate differently to each persona, with the most correlated marketing interactions to each persona's product preferences. Each of these clusters will have distinct choices, and by targeting those groups, we have known from our previous analysis correctly while running experiments. Based on the segmentation for both seven clustering segmentation of GMM as well as precious RFM segmentation, we can make targeted recommendations as follows:

The test will get better results to focus on the clusters with extreme RFM segments like "Best Customers" and "Almost Lost" as well as significant features in selling. Section 4 represents that used to prepare foods for themselves or family members (i.e. homemakers, grandparents), and are most likely to keep a large stock of food on hand. They would be the most likely candidate to have shopping three days a week instead of more or less to keep the food fresh and save labour. The customers from segment 7 represent shopaholics, and according to their frequent shopping and less time interval, they are more likely to need delivery in an emergency. Other channels, such as online to offline shopping platform fit their need. Thus, from its long time interval, the customers from part 3 are more likely to need discount five days a week and should not be selected as long-term cultivation of concerns for its low loyalty.

All in all, we should try different kinds of tests instead of sticking in one specific test, like changing staffs, images, layout and user experience. Based on the validation of those business assumptions, a wish business decision will be made!