

**MIS 620 – E-Business Infrastructure**

**Dr. Aaron Elkins**

**Ponpare Coupons - Data Analytics Project**

**By**

**Kishore Gauthaman Thambiram Arumugam**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Description** | **Page Number** |
| Executive Summary | 3 |
| Discovery | 4 |
| Data Preparation | 5 |
| Model Planning | 10 |
| Model Building | 14 |
| Results | 17 |
| Discussion and Recommendations | 19 |
| Appendix A – Data Dictionary | 22 |
| Appendix B – R Code for Data Preparation | 24 |
| Appendix C – R Code for Model Planning | 29 |
| Appendix D – R Code for Model Building | 30 |
| Appendix E – R Code for Data Visualizations | 32 |

**I. EXECUTIVE SUMMARY:**

Ponpare is Japan's leading joint coupon site, offering huge discounts on everything from hot yoga, to gourmet sushi, to a summer concert bonanza. Ponpare's coupons open doors for customers they've only dreamed of stepping through. They can learn difficult to acquire skills, go on unheard of adventures, and dine like (and with) the stars.

Investing in a new experience is not cheap. They fear wasting time and money on a product or service that customers may not enjoy or fully understand. Ponpare takes the high price out of this equation, making it easier for a customer to take the leap towards their first sky-dive or diamond engagement ring.

Using past purchase and browsing behavior, this competition required us to predict which coupons a customer will buy in a given period of time. The resulting models will be used to improve Ponpare's recommendation system, so they can make sure their customers don't miss out on their next favorite thing.

**Discovery Phase: Business Domain:** Ponpare is a global e-commerce marketplace connecting its customers with local merchants by offering discounts for different kinds of activities and services. We used k-means clustering to group the coupons into clusters based on price, catalog price, discount rate and discount period.

**Data Source:**Kaggle was our data source and we were provided with a year of transactional data for 22,873 users on the site ponpare.jp. The data set was divided into several tables such as: User Information (22,873), Coupon Information - Training set (19,413) and Test set (310), coupon views (2.8M), purchases (168,996).

**Data Preparation Phase:** To make predictions on 310 coupons, using browsing behavior of 22, 873 users, we started by performing Extract, Transform and Load (ETL) process on the training data. As the critical aspect of any data science project is to become familiar with the data, we spent most of our project time in learning about the data sets which are provided and conditioning the data and visualized the data using bar plots. Visualizations included the overall gender distributions of users, age groups, location wise data, number of coupons views vs purchases, etc.

**Model Planning Phase:** Based on our data structure and volume,we used k-means to cluster coupons into multi clusters. Our model was based mostly on price and discount rate of coupons. We then used logistic regression model to predict purchasing probability, by creating a master table with a combination of user and coupon data.

**Model Building Phase:** Ensuring that our model data is sufficiently robust,we created smaller test sets, for validating approach. We used k-means clustering and logistic regression to model our data.We chose R as the environment to build our model and workflows.

**Results:** We met our success criteria by successfully predicting the top 10 coupons that a particular user is most likely to purchase. To summarize our findings, we plotted a ROC curve showing the tradeoff between the true positive rate and the false negative rate. We found our model to be 73.4% accurate and will hence help improve Ponpare’s revenue and profit.

**Recommendations:** We created a recommendation system to predict the type of coupons a customer will buy, so that Ponpare can display the right kind of coupons to their customers to increase sales.

**II. DISCOVERY**

**Project Background:**

We had historical coupon purchasing data from Ponpare, along with customer’s data and coupon data. However, the data was separated into many tables and there are too many attributes that were not relevant to the analysis. The team had to choose which attributes to be used, such as, which customer characteristics and coupon characteristics will be used in the analysis. The next step was to change data formats into the appropriate formats, and combine all of the data into one table. We then developed prediction models after that and performed prediction model performance analysis to choose one model that would perform best on our testing data set.

**Framing the problem:**

The goal was to recommend a ranked list of coupons for each user in the dataset (found in user\_list.csv). Our predictions were then scored against the actual coupon purchases, made during the test set week, of the 310 possible test set coupons.

**Data source:**

Kaggle provided us with a year of transactional data for 22,873 users on the site ponpare.jp. The training set spans the dates 2011-07-01 to 2012-06-23. The test set spans the week after the end of the training set, 2012-06-24 to 2012-06-30. The dataset has a relational format, with hashed ID columns for each entity.

**Descriptive statistics:**

Our sample size is 22,873 (user’s browsing history on Ponpare), and predictions are to be made on 310 coupons which. The attributes we used included the following:

* Coupon area (Prefecture name, coupon id)
* Coupon list (catalog price, discount price, coupon id)
* Coupon details (purchase id, coupon id, user id)
* User list (User ID, Sex, Age)
* Coupon visit (purchase id, view coupon id)
* Prefecture locations

**Business case for selecting this data:**

We selected the attributes that were most important and relevant to make our predictions about which coupon a customer is most likely to purchase.

**Hypotheses:**

Our initial hypotheses were as listed below:

* Customers with different gender and age have different purchasing habits
* Coupons with higher discount rate tends to have higher purchase rate
* Coupons with longer validity period tends to have higher purchase rate

**III. DATA PREPARATION**

The competition wanted its participants to make use of 1 year purchase and browsing data of the 22,873 users of Ponpare and to predict which coupon will be purchased by the users over the next week. Predictions is to be made on the 310 coupons which is to be purchased on the following week.

The problem here is a cold-start recommendation problem. Cold-start problems are situations where the user`s ratings or clicks are unknown. So in order to calculate the predictions, we have to make use of user`s other attributes like gender, age, geographical location and the coupon`s attributes like genre, categories, popular coupons, discount rates, etc.

**Performing Extract, Transform and Load (ETL) process:**

Since, our data source was Kaggle, we were given with the test and training data for

* Coupon area
* Coupon list
* Coupon details
* User list
* Coupon visit and
* Prefecture locations.

Ponpare uses a relational database model for storing its records in a database. The coupon area, coupon list and coupon details were the intersection data of the coupon and user entity. These intersection tables were connected with the help of the primary keys in this case the coupon hash id and the user hash id. We extracted these data sets from the Kaggle website and used it for the analysis. Before performing the analysis, we had to do some transformation to the data sets. Since the data was about the Japanese coupon website, most of the data were in Japanese. We had to translate the Japanese script into English in order to better understand the data set. This was our initial challenge before moving to the actual analysis. This transformed data was loaded into the R Studio and we used it for further exploration and the original data was preserved.

**Data Conditioning:**

As the critical aspect of any data science project is to become familiar with the data, we spent most of our project time in learning about the data sets which are provided.

**i) User\_List.csv**

First, we started with the “user\_list.csv”. Since the goal was to predict the coupons for the users of Ponpare, we started analyzing the user list that has the master list of all the users in the data set. The attribute PREF\_NAME had the prefecture names which is the location of where the users are from. This was in Japanese in the original data set. We translated that into English using RStudio. We found that out of 22,873 users, only 48 locations were unique. This means these users visited the website from 48 different locations. From 22,873 records we were able to narrow it down to 48 which made our translation easier. A new data frame was created with the Japanese script and its equivalent English translation. This data frame was merged with the original data set with the PREF\_NAME as the common attribute. After the translation, the new variable age group was created with the help of the age variable given in the actual data set.

user\_list$AGE\_GROUPS <- cut (user\_list$AGE, breaks = c (14, 24, 34, 44, 54, 64, 74, 84),

labels = c ("14-23", "24-33", "34-43", "44-53", "54-63", "64-73", "74-83"))

**ii) Coupon\_list\_train.csv**

Similarly, in coupon list, there were five variables namely CAPSULE\_TEXT, GENRE\_NAME, large\_area\_name, ken\_name and small\_area\_name that were needed to be converted into English. Here, out of 19,413 observations only 136 values were unique from all 5 variables that needed the translation. A new data frame was created with these values and its corresponding English values. This data frame was merged with the original data set coupon\_list\_train with these five variables as the common attribute. The discount rates were grouped with the help of the price rate variable in the data set. Also, the validity of each coupon was also grouped to better understand the coupons nature. This was done to both the test and the training data sets.

coupon\_list\_train$PRICE\_RATE\_GROUPS <- cut (coupon\_list\_train$PRICE\_RATE, breaks = c (0, 20, 40, 60, 80, 100), labels = c ("0-19", "20-39", "40-59", "60-79", "80-100"))

coupon\_list\_train$VALIDPERIOD\_RANGE <- cut (coupon\_list\_train$VALIDPERIOD, breaks = c (0, 50, 100, 150, 200), labels = c ("0-49", "50-99", "100-149", "150-199"))

**iii) Coupon\_detail\_train.csv**

The translated data frame which was created using the coupon\_list\_train data set was used here to perform the translation. The two data frames were merged with the common variable small area name. Here only the variable Small area name was to be translated. Again to better understand the data set, the coupon purchase count was grouped and appended into a new variable item group.

coupon\_detail\_train$ITEM\_COUNT\_GROUP <- cut (coupon\_detail\_train$ITEM\_COUNT, breaks = c(0,10,20,30,40,50,60), labels = c("0-9", "10-19", "20-29", "30-39","40-49", "50-59"))

**Data Visualizations:**

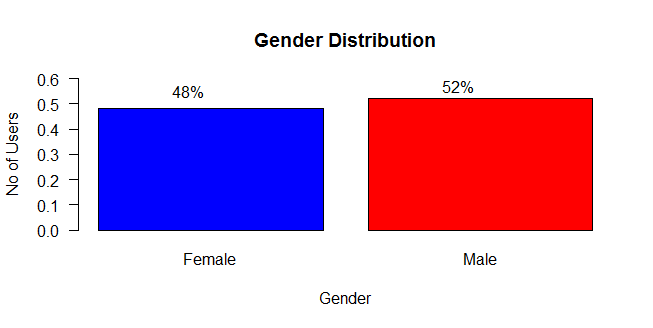


Figure 1: Gender Distribution

The above bar plot shows the overall gender distribution of the users of Ponpare. Out of the 22,873 users, we can see that 48% are Female and 52% are male users.

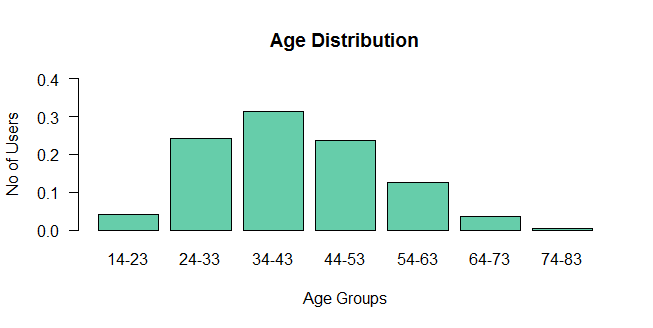


Figure 2: Age Distribution

On exploring further, we can see that the user age groups were the highest in the group 34-43, 24-33 and 44-53. The minimum age of the user is 14 and the maximum age is 83.

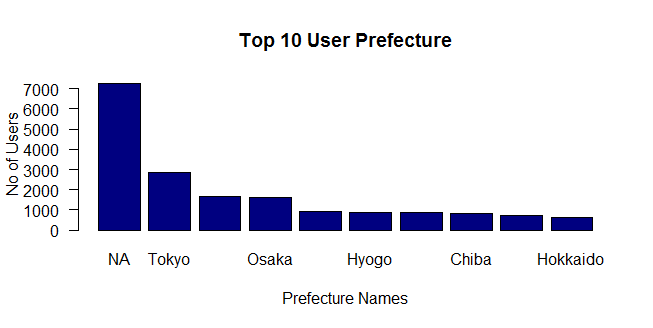


Figure 3: Top 10 User Prefecture

The above bar plot shows the top 10 locations where the users are residing. Clearly around 7000 user`s residence is not available. Apart from the NA, we can see that the most popular location where the users have visited the Ponpare website are from Tokyo. The second most popular location is Kanagawa and the third is Osaka.

While visualizing the coupon list data set, many insights were revealed on the coupon`s discount rate.

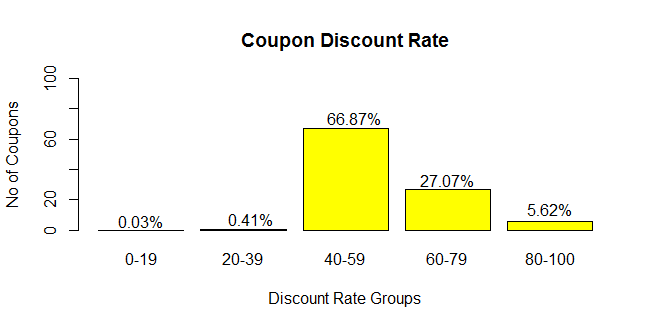


Figure 4: Coupon Discount Rate

The majority of the discount rate of the coupons from the Ponpare website falls under the 40-59 range. 66.87% of the coupon`s discount rate falls in this range. Only 5.62% of the discount count rates were in the range 80-100. 27.07% of discount rate falls under the 60-79 range.

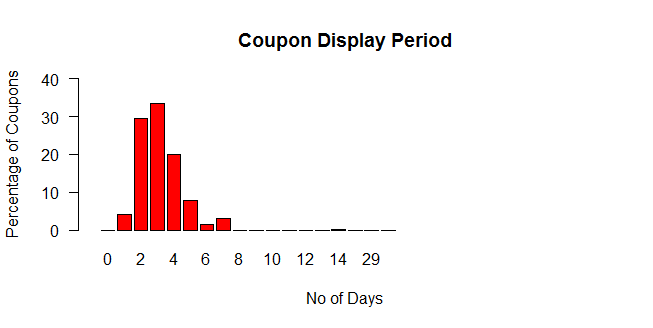


Figure 5: Coupon Display Period

If we look at the number of days the coupons are displayed on the website, 35% of the coupons were displayed for 3 days. 30% of the coupons were displayed for 2 days. The above bar plot shows us that most of the coupons were displayed from 1 to 7 days.

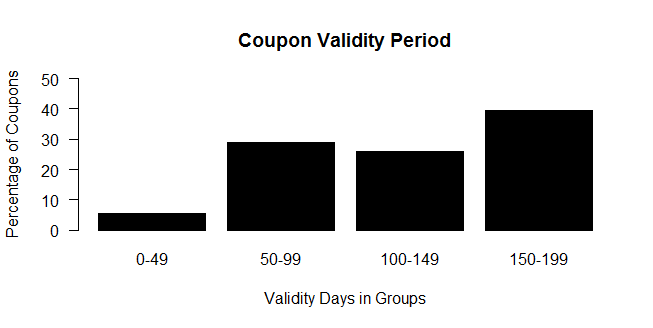


Figure 6: Coupon Validity Period

The above bar plot shows the coupon validity period. Only 8% of the coupons has a validity period of 0-50 days. Also majority of the coupons which is 40% of them has a validity period in the range 150-199 days. The rest of the coupons has the validity period from 50 to 150 days.

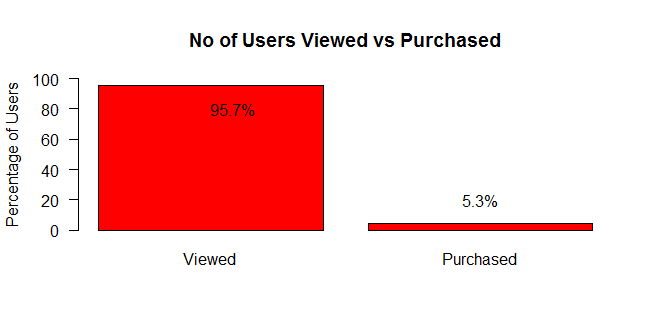


Figure 7: No. of users Viewed Vs Purchased

Out of the 2 million transaction records from 22,873 users, we can see that only 5.3% of the coupons are purchased. Some coupons were viewed first and then purchased. But 95.7% of the coupons were just viewed and didn’t got purchased. This shows that the user to coupon purchase rate is very low.

On exploring on the coupons that were purchased, we can see that the minimum value is 1 and the maximum value is 55. This means that some coupons were purchased just once and some coupons were purchased 55 times. This shows the coupons that are popular among the users and were purchased more frequently by different users.

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.000 1.000 1.431 2.000 55.000

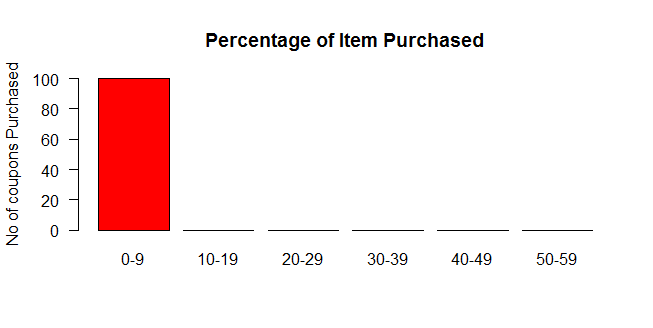


Figure 8: Percentage of Item Purchased

The majority of the coupons that were purchased were in the range 0-9. This means that almost all the coupons were only purchased 1 to 9 times by the users.

**IV. MODEL PLANNING**

After translating and visualizing the data sets from the previous phase we had a general idea about the datasets. The structure of the datasets is one factor that dictates the tools and analytical techniques used for the model building phase. We decided to dig further into the data sets to analyze the structure of the datasets.

**Data Exploration:**

Although some data exploration takes place in the data preparation phase, those activities focus mainly on data hygiene and on assessing the quality of data. In this phase, we explored the data to understand the relationships among the variables and to better understand the problem.

*Clustering:*

We used “coupon\_list\_train.csv” and “coupon\_list\_test.csv” to perform the cluster analysis. Since coupon\_list contains the master list of all the 310 coupons, we thought grouping them into different clusters will give a better idea of the coupons. This data set had information of the coupons like discount rate, coupon price rate, sales release date, sales end date and validity period.

For the analysis, we took the variables: “PRICE\_RATE”, “CATALOG\_PRICE”, “DISCOUNT\_PRICE” and “DISPPERIOD” from both the training and testing data. The optimal value for the number of clusters was got from within sum of squares value.

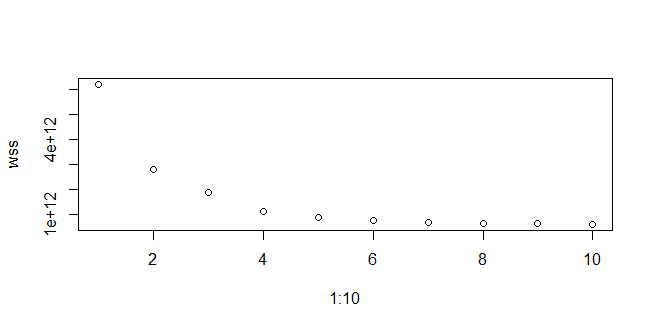


Figure 9: Clustering – WSS values

As the WSS value began to drop at 4, we were able to finalize our k value as 4. On performing the K means clustering, we got the following result.

PRICE\_RATE CATALOG\_PRICE DISCOUNT\_PRICE DISPPERIOD

1 76.61905 178641.27 40990.159 3.126984

2 66.31498 59481.52 19006.507 3.424009

3 56.89259 5665.76 2326.641 3.142028

4 62.28269 20799.44 7708.002 3.299245

We were able to see that the algorithm grouped them into clusters based on the discount rate and catalog price. Cluster 1 had the highest discount rate of 76.7% with the highest catalog price of 178,641.27. Cluster 3 had the least discount rate of 56.9% with the least catalog price of 5665.76.

Next, we wanted to find the no of visits and purchases each coupon had. For this, we had to merge two tables’ coupon\_list\_train and coupon\_visit\_train with the common variable coupon\_id\_hash. The relationship between the coupons visited and coupons purchased is shown in the scatter plot below. There is a positive correlation of 0.7789. By plotting it, we were able to identify an outlier. This could possibly be the most popular coupon in Ponpare website as it had more views and more purchases.

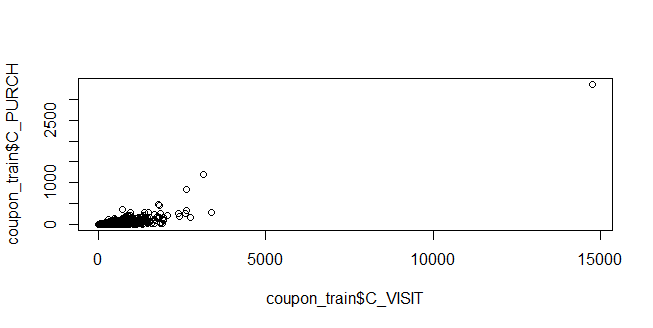


Figure 10: Coupon Visits Vs Purchases

The hypotheses which we framed in the discovery phase were researched further in this phase. The tool we used to visualize the data was Tableau. It is explained below.

Hypothesis 1: (Customers of different gender and age have different purchasing habits)

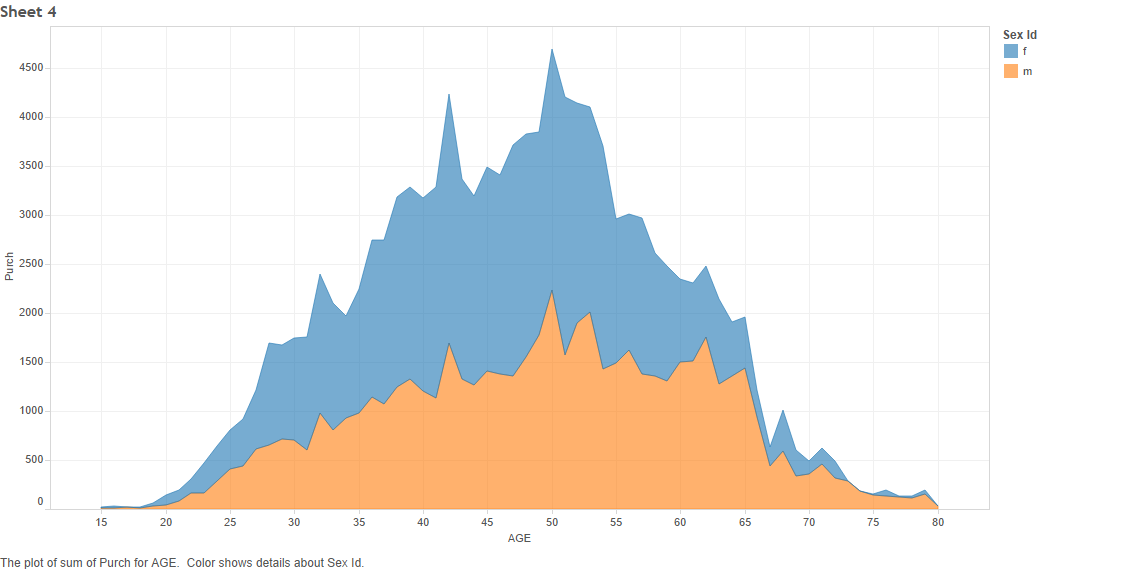


Figure 11: Plot of sum of Purchases for Age

From the plot above, we can see that females have purchased more coupons than the males. Also, users of age 35 – 65 has purchased more coupons than the users of other ages. With the help of this visualization, we were able to confirm our hypothesis that customers of different gender and different age groups have different purchasing habits.

Hypothesis 2: (Price of coupons and discount rate)

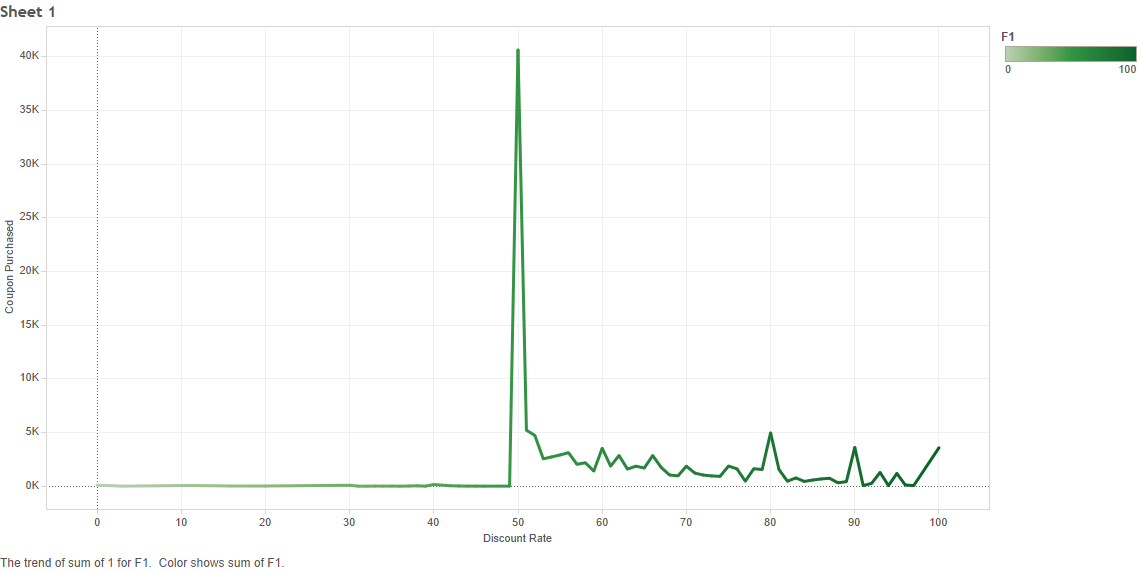


Figure 12: Discount rate for coupons purchased

From the above visualization, we can see that as the discount rate of the coupons increases, the coupon purchase rate is also increased. When the discount rate crosses 50%, there is huge spike in purchase rate. This shows that most of the coupons have a discount rate of 50% and the purchase rate is high.

Hypothesis 3: (Coupons with longer validity period tends to have higher purchase rate).

From the below visualization, we were able to prove our hypothesis about validity period and purchase rate. Initially there were few purchases if the validity period of each coupons lie in the range of 15-20 days. When the validity period lie in the range of 65 – 90 days, there were more purchases than the 15-20 days range. But when the validity period of the coupons lie in the 175 – 180 range, we can see that there is huge spike in purchases. This proves our hypothesis that more the validity days of the coupons, the more the chances of users to purchase those coupons.

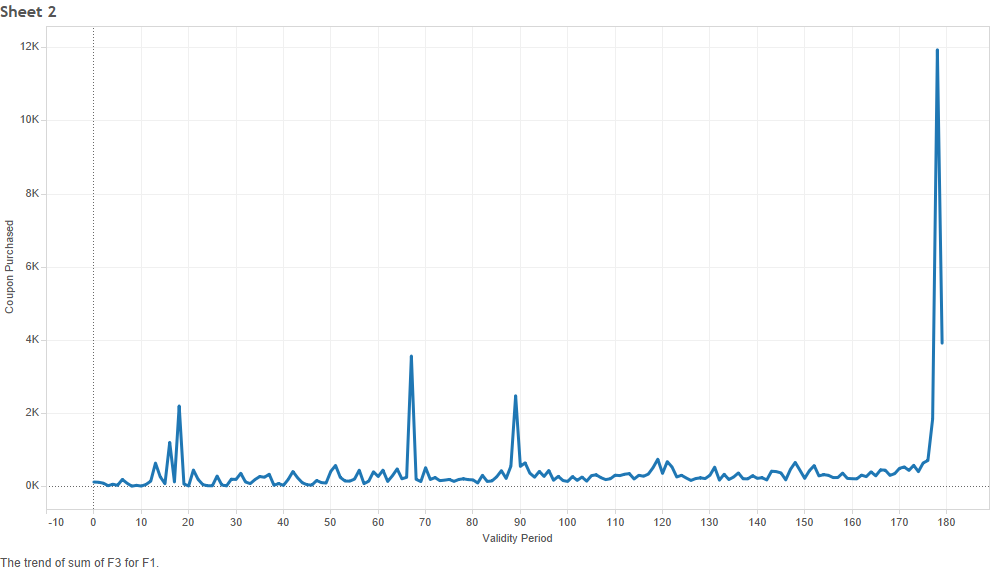


Figure 13: Validity period for Coupons purchased

**V. MODEL BUILDING**

**Coupon Clustering**

We used kmeans to cluster coupons into multi clusters. The clustering model will be based mostly on price and discount rate of coupons, so genre and availability of the coupons are not added in this model. Variables that have been added in the cluster are PRICE\_RATE, CATALOG\_PRICE, DISCOUNT\_PRICE, and DISPERIOD. Figure below shows within sum square error of each k value.

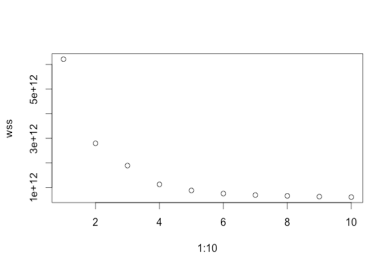


Figure 14: WSS error for each k value

We think that using k value at four will give the best performance. Table 1 below presents centroids of each cluster. Cluster 1 represents coupons that has high price rate, high catalog price, and high discount price. Cluster 3 represent coupons that has low price rate, low catalog price, and low discount price. We do not see much different in display period (DISPPERIOD) between the four clusters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | PRICE\_RATE | CATALOG\_PRICE | DISCOUNT\_PRICE | DISPPERIOD |
| 1 | 76.62 | 178,641.27 | 40,990.16 | 3.13 |
| 2 | 62.28 | 20,799.44 | 7,708.00 | 3.30 |
| 3 | 56.89 | 5,665.76 | 2,326.64 | 3.14 |
| 4 | 66.31 | 59,481.52 | 19,006.50 | 3.42 |

Table 1: Cluster Centroids

**Logistic regression:**

We use logistic regression model to predict purchasing probability. First of all, we created a master table that is a combination of users and coupons. This table included the flag that indicate whether a user has purchased a coupon, and we will use this flag as a class that we want to predict in the model. Figure 15 below shows all variables in this table.

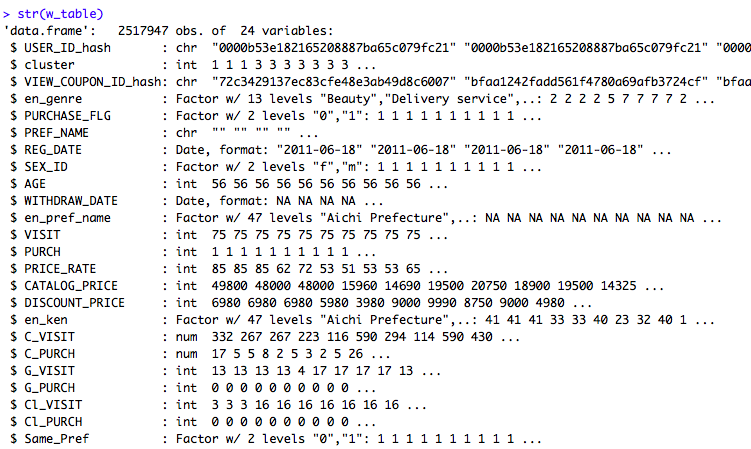


Figure 15: Master Table

Some variables in the table are highly correlated, so we only choose some variables that we think will be highly helpful in the model. Table below presents the names and description of the final candidate variables. G\_PURCH and Cl\_PURCH tells us a about the interest of a user. If the user has purchased any coupon in the genre or in the cluster before, he or she might be interested in this coupon too. C\_PURCH reflects a popularity of the coupon. PRICE\_RATE and DISCOUNT\_PRICE are coupon attributes that might affect user purchasing decision. Same\_Pref indicates whether the user and the coupon are close to each other geographically.

|  |  |
| --- | --- |
| G\_PURCH | A flag telling whether a user has been ever purchase a coupon in this genre. |
| Cl\_PURCH | A flag telling whether a user has been ever purchase a coupon in this cluster. |
| C\_PURCH | A number of time a coupon has been purchased by any users. |
| PRICE\_RATE | Discount rate of a coupon |
| DISCOUNT\_PRICE | Discount price of a coupon |
| Same\_Pref | A flag telling whether a user and a coupon have the same prefecture |

Table 2: Predictors used in logistic regression

The logistic regression has been trained using 70% data from the master table. We used glm function to create the regression model. Figure 16 below shows the model summary:

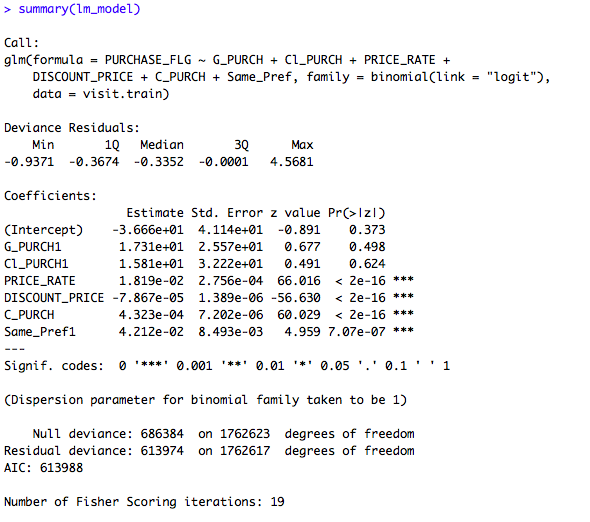


Figure 16: Logistic Regression Summary

**VI. RESULTS**

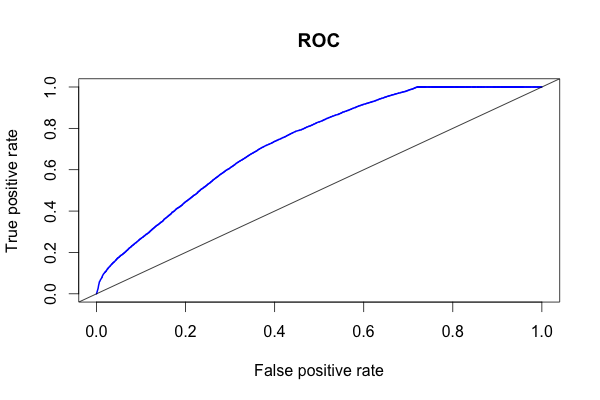
****

Figure 17: ROC curve of the logistic regression on the model

The above figure shows the ROC curve for our model. The ROC curve gives the tradeoff between the true positive rate and the false negative rate. The area under the curve indicates the accuracy rate of the model on the test data which is 73.4% in this case. The lift shows that our model performs better than the random classifier.

**Success Criteria’s:**

* **Use the past purchasing and browsing behavior to accurately predict and display coupons to customers, to increase the probability of them purchasing coupons from Ponpare**

****

Figure 18: Top ten coupons recommendation for a particular user

Our model recommends ten coupons to each user which improves the recommendation system of Ponpare. In the figure below, the coupons that were recommends to a particular user is highlighted.

* When right coupons are recommended to right customers, the customers tend to buy more coupons which increase revenue and profit.

There were four hypotheses that were build and were addressed. The hypothesis that were developed for the model gave us better understanding about the data. These hypotheses are explained in details in the model planning phase. List of the hypothesis are given below.

* Customers with different gender and age are different purchase habits.
* Coupons with discount rate of more than 50% are more likely to be purchased
* Coupons with higher discount rate tends to have higher purchase rate.
* Coupons with longer validity period tends to have higher purchase rate.

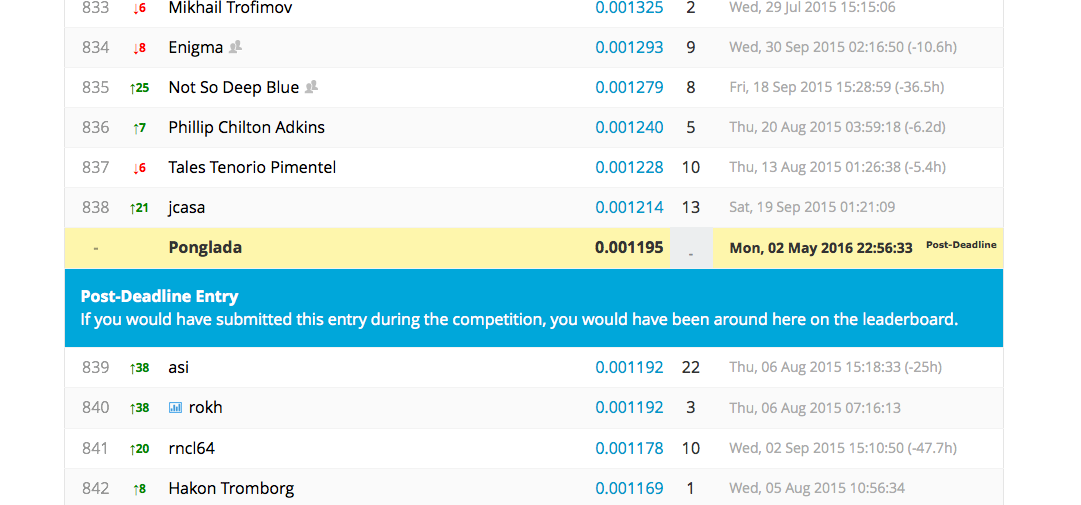
****

Figure 19: Rank for the submission of model on Kaggle website

One of method of evaluation of your model was done by submitted it on the kaggle website. Our model crossed the benchmark that was set by Kaggle and also we were ahead of about 300 participants.

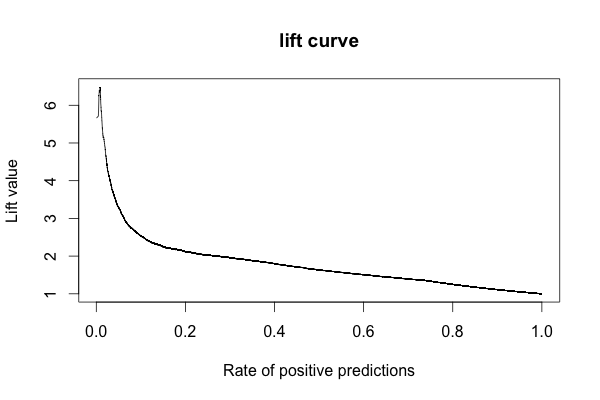
****

Figure 20: Lift Curve

**VII. DISCUSSION AND RECOMMENDATIONS**

Insights about data:

* Ponpare has 48% female and 52% male users.
* Majority of the users are between the age group of 24-53.
* Most of the users are from Tokyo, followed by Kanagawa and Osaka. So ponpare can sells coupons of events in these particular prefectures.
* Most of the coupons offered on the website have discount rate of 40-59%.
* Coupons are displayed on the website for about a weak period which have the validity of about 180 days.

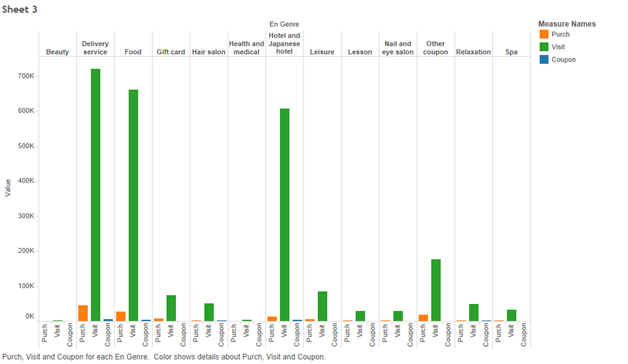
****

Figure 21: Purchase, Visit and Coupons for each category

The data visualizations helped us gain key insights about the data. Some of the facts that were learnt from the data that would help them improve the recommendation system were:

* Maximum number of views of coupons were from the genre Delivery services, Food, Hotel and japanese hotel, so with the help of this data we can suggest ponpare to have more coupons discounts offered in these area so that this will increase the rate of purchase.
* Majority of the coupons purchased were delivery services, so this will help ponpare gain insight to add more coupons for this particular genre and also help them learn on those genres that they are not doing good.
* Spa and Beauty have less views and purchases as well in comparison with the other genre coupons, so this will help Ponpare search for better and exciting deals to offer to the user.
* Health has least no of views and purchases, so this gives insight to the company to work more on this particular genre.

**Recommendations**

* Some of the data about the area were missing in the dataset, which we treated them as NA. If we had more information about them, it would help our model to be more accurate while recommending coupons considering the locations. We didn’t use this variable in our model because 37% of the data was missing.
* Ponpare had many coupons that had restricted usage days, like monday only, holiday only and so on. If the coupons were more generic there might be probability of increase in sales of the coupons as most of the customers would prefer using coupons in their spare time, which would usually be on weekends or holidays.
* The dataset could contain details about the frequency of views of a particular coupon. This would help Ponpare determine the popular coupon and try to fetch more deals related coupons or coupon genres.
* Purchase of coupons increase if the discount rate is more than 50%, so the company can recommend coupons that has deals which have higher discount rates.
* Purchases of coupons increase when they are valid for longer period, so the company can have the validity period extended on the coupons.

**Conclusion**

Female users in the age group of 40-59 is more likely to buy a coupon of genre Delivery of goods and service with the discount rate of about 50%.

**Appendix A - Data Dictionary**

Coupon\_list\_train.csv and coupon\_list\_test.csv

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Name | Description | Type | Length | Note |
| CAPSULE\_TEXT | Capsule text | VARCHAR2 | 20 | [JPN] |
| GENRE\_NAME | Category name | VARCHAR2 | 50 | [JPN] |
| PRICE\_RATE | Discount rate | NUMBER | 4 |  |
| CATALOG\_PRICE | List price | NUMBER | 10 |  |
| DISCOUNT\_PRICE | Discount price | NUMBER | 10 |  |
| DISPFROM | Sales release date | DATE |  |  |
| DISPEND | Sales end date | DATE |  |  |
| DISPPERIOD | Sales period (day) | NUMBER | 4 |  |
| VALIDFROM | The term of validity starts | DATE |  |  |
| VALIDEND | The term of validity ends | DATE |  |  |
| VALIDPERIOD | Validity period (day) | NUMBER | 4 |  |
| USABLE\_DATE\_MON | Is available on Monday | CHAR | 1 |  |
| USABLE\_DATE\_TUE | Is available on Tuesday | CHAR | 1 |  |
| USABLE\_DATE\_WED | Is available on Wednesday | CHAR | 1 |  |
| USABLE\_DATE\_THU | Is available on Thursday | CHAR | 1 |  |
| USABLE\_DATE\_FRI | Is available on Friday | CHAR | 1 |  |
| USABLE\_DATE\_SAT | Is available on Saturday | CHAR | 1 |  |
| USABLE\_DATE\_SUN | Is available on Sunday | CHAR | 1 |  |
| USABLE\_DATE\_HOLIDAY | Is available on holiday | CHAR | 1 |  |
| USABLE\_DATE\_BEFORE\_HOLIDAY | Is available on the day before holiday | CHAR | 1 |  |
| large\_area\_name | Large area name of shop location | VARCHAR2 | 30 | [JPN] |
| ken\_name | Prefecture name of shop | VARCHAR2 | 8 | [JPN] |
| small\_area\_name | Small area name of shop location | VARCHAR2 | 30 | [JPN] |
| COUPON\_ID\_hash | Coupon ID | VARCHAR2 | 32 |  |

user\_list.csv

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Description | Type | Note |
| USER\_ID\_hash | User ID | VARCHAR2 |  |
| REG\_DATE | Registered date | DATE | Sign up date |
| SEX\_ID | Gender | CHAR | f = female　m = male |
| AGE | Age | NUMBER |  |
| WITHDRAW\_DATE | Unregistered date | DATE |  |
| PREF\_NAME | Residential Prefecture | VARCHAR2 | [JPN] Not registered if empty |

coupon\_visit\_train.csv

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Description | Type | Note |
| PURCHASE\_FLG | Purchased flag | NUMBER | 0:Not purchased 1:Purchased |
| PURCHASEID\_hash | Purchase ID | VARCHAR2 |  |
| I\_DATE | View date | DATE | Purchase date if purchased |
| PAGE\_SERIAL |  | VARCHAR2 |  |
| REFERRER\_hash | Referer | VARCHAR2 |  |
| VIEW\_COUPON\_ID\_hash | Browsing Coupon ID | VARCHAR2 |  |
| USER\_ID\_hash | User ID | VARCHAR2 |  |
| SESSION\_ID\_hash | Session ID | VARCHAR2 |  |

**Appendix B - R Code in Data Preparation Phase**

**Translate2.R**

# Part of this script is retrieved from

# https://www.kaggle.com/anguyen/coupon-purchase-prediction/translate-everything-to-english-using-r

#################################################################################

# This script translates Japanese text to English in the data files

# and keeps the English translation in separate columns

#################################################################################

# Create master translation table from Japanese to English

coupon\_list\_train = read.csv("../input/coupon\_list\_train.csv", as.is=T) # Source file the English list is keyed by

trans = data.frame(

 jp=unique(c(coupon\_list\_train$GENRE\_NAME, coupon\_list\_train$CAPSULE\_TEXT,

             coupon\_list\_train$large\_area\_name, coupon\_list\_train$ken\_name,

             coupon\_list\_train$small\_area\_name)),

 en=c("Food","Hair salon","Spa","Relaxation","Beauty","Nail and eye salon","Delivery service","Lesson","Gift card","Other coupon","Leisure","Hotel and Japanese hotel","Health and medical","Other","Hotel","Japanese hotel","Vacation rental","Lodge","Resort inn","Guest house","Japanse guest house","Public hotel","Beauty","Event","Web service","Class","Correspondence course","Kanto","Kansai","East Sea","Hokkaido","Kyushu-Okinawa","Northeast","Shikoku","Chugoku","Hokushinetsu","Saitama Prefecture","Chiba Prefecture","Tokyo","Kyoto","Aichi Prefecture","Kanagawa Prefecture","Fukuoka Prefecture","Tochigi Prefecture","Osaka prefecture","Miyagi Prefecture","Fukushima Prefecture","Oita Prefecture","Kochi Prefecture","Hiroshima Prefecture","Niigata Prefecture","Okayama Prefecture","Ehime Prefecture","Kagawa Prefecture","Tokushima Prefecture","Hyogo Prefecture","Gifu Prefecture","Miyazaki Prefecture","Nagasaki Prefecture","Ishikawa Prefecture","Yamagata Prefecture","Shizuoka Prefecture","Aomori Prefecture","Okinawa","Akita","Nagano Prefecture","Iwate Prefecture","Kumamoto Prefecture","Yamaguchi Prefecture","Saga Prefecture","Nara Prefecture","Mie","Gunma Prefecture","Wakayama Prefecture","Yamanashi Prefecture","Tottori Prefecture","Kagoshima prefecture","Fukui Prefecture","Shiga Prefecture","Toyama Prefecture","Shimane Prefecture","Ibaraki Prefecture","Saitama","Chiba","Shinjuku, Takadanobaba Nakano - Kichijoji","Kyoto","Ebisu, Meguro Shinagawa","Ginza Shinbashi, Tokyo, Ueno","Aichi","Kawasaki, Shonan-Hakone other","Fukuoka","Tochigi","Minami other","Shibuya, Aoyama, Jiyugaoka","Ikebukuro Kagurazaka-Akabane","Akasaka, Roppongi, Azabu","Yokohama","Miyagi","Fukushima","Much","Kochi","Tachikawa

Machida, Hachioji other","Hiroshima","Niigata","Okayama","Ehime","Kagawa","Northern","Tokushima","Hyogo","Gifu","Miyazaki","Nagasaki","Ishikawa","Yamagata","Shizuoka","Aomori","Okinawa","Akita","Nagano","Iwate","Kumamoto","Yamaguchi","Saga","Nara","Triple","Gunma","Wakayama","Yamanashi","Tottori","Kagoshima","Fukui","Shiga","Toyama","Shimane","Ibaraki"),

 stringsAsFactors = F)

# Append data with translated columns...

# COUPON\_LIST\_TRAIN.CSV

coupon\_list\_train = read.csv("../input/coupon\_list\_train.csv", as.is=T) # Read data file to translate

names(trans)=c("jp","en\_capsule") # Rename column

coupon\_list\_train=merge(coupon\_list\_train,trans,by.x="CAPSULE\_TEXT",by.y="jp",all.x=T) # Join translation onto original data

names(trans)=c("jp","en\_genre"); coupon\_list\_train=merge(coupon\_list\_train,trans,by.x="GENRE\_NAME",by.y="jp",all.x=T)

names(trans)=c("jp","en\_small\_area"); coupon\_list\_train=merge(coupon\_list\_train,trans,by.x="small\_area\_name",by.y="jp",all.x=T)

names(trans)=c("jp","en\_ken"); coupon\_list\_train=merge(coupon\_list\_train,trans,by.x="ken\_name",by.y="jp",all.x=T)

names(trans)=c("jp","en\_large\_area"); coupon\_list\_train=merge(coupon\_list\_train,trans,by.x="large\_area\_name",by.y="jp",all.x=T)

write.csv(coupon\_list\_train, "coupon\_list\_train\_en.csv", row.names = F)

# COUPON\_LIST\_TEST.CSV

coupon\_list\_test = read.csv("../input/coupon\_list\_test.csv", as.is=T) # Read data file to translate

names(trans)=c("jp","en\_capsule") # Rename column

coupon\_list\_test=merge(coupon\_list\_test,trans,by.x="CAPSULE\_TEXT",by.y="jp",all.x=T) # Join translation onto original data

names(trans)=c("jp","en\_genre"); coupon\_list\_test=merge(coupon\_list\_test,trans,by.x="GENRE\_NAME",by.y="jp",all.x=T)

names(trans)=c("jp","en\_small\_area"); coupon\_list\_test=merge(coupon\_list\_test,trans,by.x="small\_area\_name",by.y="jp",all.x=T)

names(trans)=c("jp","en\_ken"); coupon\_list\_test=merge(coupon\_list\_test,trans,by.x="ken\_name",by.y="jp",all.x=T)

names(trans)=c("jp","en\_large\_area"); coupon\_list\_test=merge(coupon\_list\_test,trans,by.x="large\_area\_name",by.y="jp",all.x=T)

write.csv(coupon\_list\_test, "coupon\_list\_test\_en.csv", row.names = F)

# USER\_LIST.CSV

user\_list = read.csv("../input/user\_list.csv", as.is=T)

names(trans)=c("jp","en\_pref\_name"); user\_list=merge(user\_list,trans,by.x="PREF\_NAME",by.y="jp",all.x=T)

write.csv(user\_list, "user\_list\_en.csv", row.names = F)

**PrepareData.R**

setwd("~/Documents/MIS620/Project/Input")

# Retrieve raw data

user\_list <- read.csv("user\_list\_en.csv", as.is = T)

coupon\_train <- read.csv("coupon\_list\_train\_en.csv", as.is = T)

coupon\_test <- read.csv("coupon\_list\_test\_en.csv", as.is = T)

visit <- read.csv("coupon\_visit\_train.csv", as.is = T)

# Clean & Explore data

# User data

user\_list$REG\_DATE = as.Date(user\_list$REG\_DATE)

user\_list$WITHDRAW\_DATE = as.Date(user\_list$WITHDRAW\_DATE)

user\_list$SEX\_ID <- as.factor(user\_list$SEX\_ID)

user\_list$en\_pref\_name <- as.factor(user\_list$en\_pref\_name)

str(user\_list)

summary(user\_list)

# Coupon data

to\_date <- c("DISPFROM", "DISPEND", "VALIDFROM", "VALIDEND")

for (a in to\_date) {

 coupon\_train[,names(coupon\_train)==a] <- as.Date(coupon\_train[,names(coupon\_train)==a])

 coupon\_test[(coupon\_test)==a] <- as.Date(coupon\_test[,names(coupon\_test)==a])

}

to\_factor <- c("USABLE\_DATE\_MON", "USABLE\_DATE\_TUE", "USABLE\_DATE\_WED", "USABLE\_DATE\_THU", "USABLE\_DATE\_FRI", "USABLE\_DATE\_SAT", "USABLE\_DATE\_SUN", "USABLE\_DATE\_HOLIDAY", "USABLE\_DATE\_BEFORE\_HOLIDAY", "en\_capsule", "en\_genre", "en\_small\_area", "en\_ken", "en\_large\_area")

for (a in to\_factor) {

 coupon\_train[,names(coupon\_train)==a] <- as.factor(coupon\_train[,names(coupon\_train)==a])

 coupon\_test[,names(coupon\_test)==a] <- as.factor(coupon\_test[,names(coupon\_test)==a])

}

str(coupon\_train)

summary(coupon\_train)

# Visit Data

visit$PURCHASE\_FLG <- as.factor(visit$PURCHASE\_FLG)

visit$I\_DATE <- as.Date(visit$I\_DATE)

str(visit)

summary(visit)

**Transform.R**

library(data.table)

attrs <- c("COUPON\_ID\_hash", "PRICE\_RATE", "CATALOG\_PRICE", "DISCOUNT\_PRICE", "en\_genre", "en\_ken")

coupon\_train <- coupon\_train[,names(coupon\_train) %in% attrs]

coupon\_test <- coupon\_test[,names(coupon\_test) %in% attrs]

# Number of visit and purchase of each coupon

attrs <- c("PURCHASE\_FLG", "VIEW\_COUPON\_ID\_hash", "USER\_ID\_hash")

visit <- visit[,names(visit) %in% attrs]

visit\_table <- data.table(visit)

group\_table <- visit\_table[,list(C\_VISIT = .N, C\_PURCH = sum(PURCHASE\_FLG==1)), by=VIEW\_COUPON\_ID\_hash]

coupon\_train <- merge(coupon\_train, group\_table, by.x = "COUPON\_ID\_hash", by.y = "VIEW\_COUPON\_ID\_hash", all.x = T)

coupon\_test <- merge(coupon\_test, group\_table, by.x = "COUPON\_ID\_hash", by.y = "VIEW\_COUPON\_ID\_hash", all.x = T)

coupon\_train[is.na(coupon\_train)] <- 0

coupon\_test[is.na(coupon\_test)] <- 0

plot(coupon\_train$C\_VISIT, coupon\_train$C\_PURCH) #Looks like there is an outlier

cor(coupon\_train$C\_VISIT, coupon\_train$C\_PURCH) #0.7789

# Number of visit and purchase of each user and coupon

group\_table <- visit\_table[,list(VISIT = .N, PURCH = sum(PURCHASE\_FLG==1)), by=list(VIEW\_COUPON\_ID\_hash, USER\_ID\_hash)]

visit\_user\_count <- group\_table

# Number of visit and purchanse of each genre

genre\_table <- data.table(coupon\_train)[,list(COUPON = .N, VISIT = sum(C\_VISIT), PURCH = sum(C\_PURCH)), by=en\_genre]

# Number of visit and purchanse of each user

group\_table <- visit\_table[,list(VISIT = .N, PURCH = sum(PURCHASE\_FLG==1)), by=USER\_ID\_hash]

user\_list <- merge(user\_list, group\_table, by="USER\_ID\_hash", all.x = T, all.y = F)

hist(user\_list$PURCH)

hist(user\_list$VISIT)

hist(user\_list$AGE)

# How many time user visit each genre

coupon.genre <- coupon\_train[,names(coupon\_train) %in% c("COUPON\_ID\_hash", "en\_genre")]

coupon.genre <- rbind(coupon.genre, coupon\_test[,names(coupon\_test) %in% c("COUPON\_ID\_hash", "en\_genre")])

visit.genre <- merge(visit, coupon.genre, by.x = "VIEW\_COUPON\_ID\_hash", by.y = "COUPON\_ID\_hash", all.x = T, all.y = F)

visit.genre <- visit.genre[!is.na(visit.genre$en\_genre),]

visit.genre <- data.table(visit.genre)

visit.genre.count <- visit.genre[,list(G\_VISIT = .N, G\_PURCH = sum(PURCHASE\_FLG==1)), by=list(USER\_ID\_hash, en\_genre)]

visit.genre.count <- as.data.frame(visit.genre.count)

rm(visit.genre)

# How many time user visit each cluster

visit.cluster <- merge(visit, coupon\_cluster, by.x = "VIEW\_COUPON\_ID\_hash", by.y = "COUPON\_ID\_hash", all.x = T, all.y = F)

visit.cluster <- visit.cluster[!is.na(visit.cluster$cluster)]

visit.cluster <- data.table(visit.cluster)

visit.cluster.count <- visit.cluster[,list(Cl\_VISIT = .N, Cl\_PURCH = sum(PURCHASE\_FLG==1)), by=list(USER\_ID\_hash, cluster)]

visit.cluster.count <- as.data.frame(visit.cluster.count)

**PrepareTable.R**

library(data.table)

# Merge user table

attrs <- c("SEX\_ID", "AGE", "USER\_ID\_hash", "en\_pref\_name")

users\_for\_merge <- user\_list[, names(user\_list) %in% attrs]

merge.visit.user <- merge(visit, users\_for\_merge, by = "USER\_ID\_hash", all.x = T, all.y = F)

# Merge coupon table

all\_coupon <- unique(rbind(coupon\_train, coupon\_test))

merge.visit.coupon <- merge(merge.visit.user, all\_coupon, by.x = "VIEW\_COUPON\_ID\_hash", by.y = "COUPON\_ID\_hash")

merge.visit.coupon <- as.data.frame(merge.visit.coupon)

# Merge genre Count

merge.genre <- merge(merge.visit.coupon, visit.genre.count, by = c("USER\_ID\_hash","en\_genre"))

merge.genre <- merge.genre[!is.na(merge.genre$en\_genre),]

# Merge cluster count

merge.cluster <- merge(merge.genre, coupon\_cluster, by.x = "VIEW\_COUPON\_ID\_hash", by.y = "COUPON\_ID\_hash")

merge.cluster <- merge(merge.cluster, visit.cluster.count, by = c("USER\_ID\_hash","cluster"))

# Compare prefecture

w\_table <- merge.cluster

w\_table$Same\_Pref <- 0

w\_table$Same\_Pref[merge.cluster$en\_ken == merge.cluster$en\_pref\_name] <- 1

w\_table$Same\_Pref <- as.factor(w\_table$Same\_Pref)

write.csv(w\_table, "w\_table.csv")

rm(merge.cluster)

rm(merge.genre)

rm(merge.visit.coupon)

rm(merge.visit.user)

rm(users\_for\_merge)

**Appendix C - R Code in Model Planning Phase**

**Coupon\_Clustering.R**

# Prepare data

setwd("~/Documents/MIS620/Project/Input")

coupon\_train <- read.csv("coupon\_list\_train\_en.csv", as.is = T)

coupon\_test <- read.csv("coupon\_list\_test\_en.csv", as.is = T)

coupon\_train$is\_train <- 1

coupon\_test$is\_train <- 0

all\_coupon <- rbind(coupon\_train, coupon\_test)

# Prepare for cluster

str(all\_coupon)

# Add attributes to be clustered here. Avoid categorical data.

attributes <- c("PRICE\_RATE", "CATALOG\_PRICE", "DISCOUNT\_PRICE", "DISPPERIOD")

coupon\_cluster <- all\_coupon[names(all\_coupon) %in% attributes]

wss <- vector(mode = "numeric", length = 10)

for (n in 1:10) {

 wss[n] <- sum(kmeans(coupon\_cluster, n)$withinss)

}

plot(1:10, wss)

coupon\_kmeans <- kmeans(coupon\_cluster, 4)

coupon\_kmeans$centers   # 1: low price low rate, 4: high rate high price.

all\_coupon$cluster <- coupon\_kmeans$cluster

# Discriptive statistics

all\_coupon$en\_capsule <- as.factor(all\_coupon$en\_capsule)

all\_coupon$en\_genre <- as.factor(all\_coupon$en\_genre)

all\_coupon$is\_train <- as.factor(all\_coupon$is\_train)

show\_attr <- c("en\_capsule", "en\_genre", "VISIT", "PURCH", "is\_train", "PRICE\_RATE", "CATALOG\_PRICE", "DISCOUNT\_PRICE")

c <- 1

summary(all\_coupon[all\_coupon$cluster==c, names(all\_coupon) %in% show\_attr])

c <- 2

summary(all\_coupon[all\_coupon$cluster==c, names(all\_coupon) %in% show\_attr])

c <- 3

summary(all\_coupon[all\_coupon$cluster==c, names(all\_coupon) %in% show\_attr])

c <- 4

summary(all\_coupon[all\_coupon$cluster==c, names(all\_coupon) %in% show\_attr])

# Create cluster table to be join later

coupon\_cluster <- all\_coupon[,names(all\_coupon) %in% c("COUPON\_ID\_hash", "cluster")]

write.csv(coupon\_cluster, "coupon\_cluster.csv")

# Remove unused variables to save memory

rm(all\_coupon)

rm(coupon\_test)

rm(coupon\_train)

rm(coupon\_kmeans)

rm(attributes)

rm(c)

rm(n)

rm(wss)

rm(show\_attr)

**Appendix D - R Code for Model Building Phase**

**Model.R**

# Model

library(data.table)

library(caret)

# Partition training vs validation (70:30, time series)

attrs <- c("PURCHASE\_FLG", "VISIT", "G\_VISIT", "G\_PURCH", "Cl\_VISIT", "Cl\_PURCH","PRICE\_RATE", "DISCOUNT\_PRICE", "CATALOG\_PRICE", "C\_PURCH", "Same\_Pref")

set.seed(1234)

rand <- runif(nrow(w\_table), 0, 1)

visit.train <- w\_table[rand < 0.7, names(w\_table) %in% attrs]

visit.val <- w\_table[rand >= 0.7, names(w\_table) %in% attrs]

rm(rand)

visit.train$G\_PURCH[visit.train$G\_PURCH > 0] <- 1

visit.train$Cl\_PURCH[visit.train$Cl\_PURCH > 0] <- 1

visit.train$G\_PURCH <- as.factor(visit.train$G\_PURCH)

visit.train$Cl\_PURCH <- as.factor(visit.train$Cl\_PURCH)

visit.val$G\_PURCH[visit.val$G\_PURCH > 0] <- 1

visit.val$Cl\_PURCH[visit.val$Cl\_PURCH > 0] <- 1

visit.val$G\_PURCH <- as.factor(visit.val$G\_PURCH)

visit.val$Cl\_PURCH <- as.factor(visit.val$Cl\_PURCH)

# Logistic Regression

lm\_model <- glm(PURCHASE\_FLG ~ G\_PURCH + Cl\_PURCH + PRICE\_RATE + DISCOUNT\_PRICE + C\_PURCH + Same\_Pref,

              data = visit.train,

              family = binomial(link = "logit"))

summary(lm\_model)

**Validation.R**

# Check the performance

library(ROCR)

fit.pr = predict.glm(lm\_model, newdata =  visit.val)

fit.pred = prediction(fit.pr, visit.val$PURCHASE\_FLG)

fit.perf = performance(fit.pred,"tpr","fpr")

plot(fit.perf,lwd=2,col="blue",

    main="ROC")

abline(a=0,b=1)

performance(fit.pred, "auc")

# Plot lift chart

fit.perf <- performance(fit.pred,"lift","rpp")

plot(fit.perf, main="lift curve", colorize=F)

**PrepareSubmissionFile.R**

# Prepare submission file

output <- data.frame(USER\_ID\_hash = character(22873), PURCHASED\_COUPONS = character(22873), stringsAsFactors = F)

batch\_size <- 1000

i <- 1

while (i <= 22873) {

 j <- min(c(i+batch\_size-1, 22873))

 a.user <- user\_list[i:j,]

 test\_data <- merge(a.user, coupon\_test, by = NULL)

 test\_data2 <- merge(test\_data, visit.genre.count, by = c("USER\_ID\_hash", "en\_genre"), all.x = T)

 test\_data2 <- merge(test\_data2, coupon\_cluster, by = "COUPON\_ID\_hash")

 test\_data2 <- merge(test\_data2, visit.cluster.count, by = c("USER\_ID\_hash", "cluster"), all.x = T)

 test\_data2[is.na(test\_data2$G\_PURCH),names(test\_data2)=="G\_PURCH"] <- 0

 test\_data2[is.na(test\_data2$Cl\_PURCH),names(test\_data2)=="Cl\_PURCH"] <- 0

 test\_data2$G\_PURCH[test\_data2$G\_PURCH > 0] <- 1

 test\_data2$Cl\_PURCH[test\_data2$Cl\_PURCH > 0] <- 1

 test\_data2$G\_PURCH <- as.factor(test\_data2$G\_PURCH)

 test\_data2$Cl\_PURCH <- as.factor(test\_data2$Cl\_PURCH)

 test\_data2[is.na(test\_data2$C\_PURCH),names(test\_data2)=="C\_PURCH"] <- 0

 test\_data2[is.na(test\_data2$VISIT),names(test\_data2)=="VISIT"] <- 0

 test\_data2$Same\_Pref <- 0

 test\_data2$Same\_Pref[as.character(test\_data2$en\_ken) == as.character(test\_data2$en\_pref\_name)] <- 1

 test\_data2$Same\_Pref <- as.factor(test\_data2$Same\_Pref)

 pr <- predict.glm(lm\_model, test\_data2)

 summary(pr)

 test\_data <- cbind(test\_data2[,names(test\_data2) %in% c("USER\_ID\_hash", "COUPON\_ID\_hash")],pr)

 test\_data <- data.table(test\_data, key = "USER\_ID\_hash")

 test\_data <- test\_data[order(test\_data$pr),]

 max\_pr <- test\_data[, tail(.SD, 10), by = USER\_ID\_hash]

 out\_coupon <- max\_pr[, paste(COUPON\_ID\_hash, collapse = " "), by = "USER\_ID\_hash"]

 output[i:j,1] <- out\_coupon$USER\_ID\_hash

 output[i:j,2] <- out\_coupon$V1

 rm(test\_data)

 rm(test\_data2)

 rm(pr)

 i <- i + batch\_size

 write.csv(output, "../Output/output.csv")

 cat(paste(c("...", i), collapse = ""))

}

**Appendix E- R Code for Data Visualizations**

#User\_list.csv

library(ggplot2)

#barplot for Gender

table\_SEX\_ID <- table(user\_list$SEX\_ID)

gender\_percent <- table\_SEX\_ID/sum(table\_SEX\_ID)

barplot(gender\_percent, main ="Gender Distribution", xlab = "Gender", las = 1,

       ylab = "No of Users", names.arg = c("Female","Male"),

       ylim = c(0,0.6), col = c("blue","red"))

       text(0.6,0.55, "48%")

       text(1.8,0.57, "52%")

#barplot for Age Groups

table\_USER\_AGE\_GROUP <- table(user\_list$AGE\_GROUPS)

age\_percent <- table\_USER\_AGE\_GROUP/sum(table\_USER\_AGE\_GROUP)

barplot(age\_percent, main = "Age Distribution", ylab = "No of Users", las = 1

       ,names.arg = c("14-23","24-33","34-43","44-53","54-63","64-73","74-83")

       ,xlab = "Age Groups", ylim = c(0,0.4),

       col = "aquamarine3")

#User Prefecture(residence)

table\_PREF\_NAME <- table(user\_list$PREF\_NAME\_en)

top\_10\_prefecture <- sort(table\_PREF\_NAME, decreasing = TRUE)[1:10]

top\_10\_prefecture[1]/sum(table\_PREF\_NAME) \* 100 #31.72% Missing Values

barplot(top\_10\_prefecture, main = "Top 10 User Prefecture", las = 1,

       xlab = "Prefecture Names", ylab = "No of Users",

       ylim = c(0,7500), col = "navyblue",

       names.arg = c("NA","Tokyo","Kanagawa","Osaka","Aichi","Hyogo",

                     "Saitama","Chiba","Fukuoka","Hokkaido"))

#Registration Date and Time

table\_DATE\_REG\_DATE <- table(user\_list$Date\_REG\_DATE)

popular\_dates <- sort(table\_DATE\_REG\_DATE, decreasing = TRUE)[1:30]

table\_TIME\_REG\_DATE <- table(user\_list$Time\_REG\_DATE)

popular\_time <- sort(table\_TIME\_REG\_DATE,decreasing = TRUE)[1:30]

plot(popular\_time,popular\_dates)

#############################################################################

#Coupon\_list\_train.csv

discount\_rate\_table <- table(coupon\_list\_train$PRICE\_RATE\_GROUPS)

discount\_percent <- discount\_rate\_table/sum(discount\_rate\_table) \* 100

barplot(discount\_percent, main = "Coupon Discount Rate",

       xlab = "Discount Rate Groups", ylab = "No of Coupons",

       ylim = c(0,100), col = "yellow")

text(0.7,6,"0.03%")

text(2.0,7,"0.41%")

text(3.2,74,"66.87%")

text(4.3,33,"27.07%")

text(5.5,14,"5.62%")

summary(coupon\_list\_train$DISPPERIOD)

coupon\_display\_days <- table(coupon\_list\_train$DISPPERIOD)

coupon\_display\_percent <- coupon\_display\_days/sum(coupon\_display\_days) \* 100

barplot(coupon\_display\_percent, main = "Coupon Display Period", las = 1,

       xlab = "No of Days", ylab = "Percentage of Coupons", ylim = c(0,40),

       xlim = c(0,36), col = "red")

summary(coupon\_list\_train$VALIDPERIOD\_RANGE)

Coupon\_Validity <- table(coupon\_list\_train$VALIDPERIOD\_RANGE)

coupon\_valid\_percent <- Coupon\_Validity / sum(Coupon\_Validity) \* 100

barplot(coupon\_valid\_percent, main = "Coupon Validity Period",las = 1,

       xlab = "Validity Days in Groups", ylab = "Percentage of Coupons",

       ylim = c(0,50), col = "black")

#############################################################################

#coupon\_visit\_train.csv

read.csv("coupon\_visit\_train\_en.csv", as.is = TRUE)

coupon\_purchase\_count <- table(coupon\_visit\_train$PURCHASE\_FLG)

count\_percent <- coupon\_purchase\_count/sum(coupon\_purchase\_count) \* 100

barplot(count\_percent, names.arg = c("Viewed","Purchased"), las = 1,

       ylab = "Percentage of Users", main = "No of Users Viewed vs Purchased"

       ,ylim = c(0,100), col = "red")

text(0.8,80,"95.7%")

text(1.9,20,"5.3%")

coupon\_popular\_day <- table(coupon\_visit\_train$PURCHASE\_FLG,

                           coupon\_visit\_train$Weekday)

coupon\_day\_df <- data.frame(c("Monday","Tuesday","Wednesday","Thursday",

                             "Friday","Saturday","Sunday"),

                           c(13.39,15.84,15.44,14.82,13.75,11.52,10.92),

                           c(0.65,0.71,0.71,0.67,0.62,0.49,0.48))

names(coupon\_day\_df) <- c("Weekday","Not Purchased","Purchased")

str(coupon\_day\_df)

plot(coupon\_day\_df)

#############################################################################

#Coupon\_detail\_train.csv

summary(coupon\_detail\_train$ITEM\_COUNT)#Possible Outliers

summary(coupon\_detail\_train$ITEM\_COUNT\_GROUP)

coupon\_count\_group <- table(coupon\_detail\_train$ITEM\_COUNT\_GROUP)

group\_percent <- coupon\_count\_group /sum(coupon\_count\_group) \* 100

barplot(group\_percent, main = "Percentage of Item Purchased",

       ylab = "No of coupons Purchased", las = 1,

       ylim = c(0,100), col = "red")

small\_area\_count <- table(coupon\_detail\_train$SMALL\_AREA\_NAME\_en,

     coupon\_detail\_train$ITEM\_COUNT\_GROUP)

popular\_area\_count <- sort(small\_area\_count, decreasing = TRUE)[1:10]

#############################################################################