A CASE OF FINANCIAL DATA

SCORING PROJECT

Thanh-Tam NGUYEN

Khanh TRUONG

"Data! Data! Data! I can't make bricks without clay!"
— **Sir Arthur Conan Doyle**

TABLE OF CONTENTS

1. INTRODUCTION

The context of the project is based on a new Fund product introduced by a financial institution. A test campaign was conducted on a group of representative agency in order to build an economically optimal fundraising campaign.

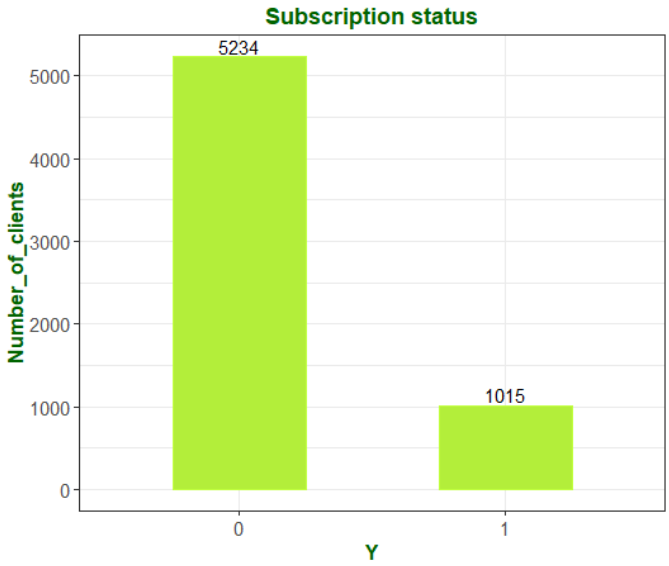
The objective of this project is to identify the most potential customers who have the most tendancy to subscribe to this Fund.

2. DESCRIPTION OF THE DATASET

2.1. Target variable Y:

Variable Y represents the choices of the clients in subscribing the Fund. For simplicity during analysis, we set Y=0 (instead of Y=9) representing the choice of a client who does not subscribe the fund. Hence, Y is a categorical variable. In total, out of 6249 clients, only 1015 clients subscribing the fund (16.24% of clients).

Statistical Descript. for Y	
n	6249
mean	0.16
sd	0.37
trimmed	0.08
min	0
max	1
skew	1.83
kurtosis	1.35



2.2. Explained variables:

2.2.1. Distribution of all explained variables:

Some important points are drawn from the statistical description for all explained variables in the table:

- There are 13 out of 23 variables which are continuous variables. For each categorical variables, the level is 0 and 1.
- There are four variables contain missing values including SLDTIT, SLDPEA, SLDBLO and SLDLIQ with respectively missing rate (44.55%, 11.89%, 7.67%, 22.37% out of 6249 observations)
- Based on the skewness measure, except for the variable AGE and ANCCDD of which distribution are fairly symmetric, 20 variables have positive skew which means the distribution concentrates on the left of the firgue, especially SOLMOY and FLUCRE.

This implies, for examples, most of the clients in the group of representative agency have small or medium credit loan.

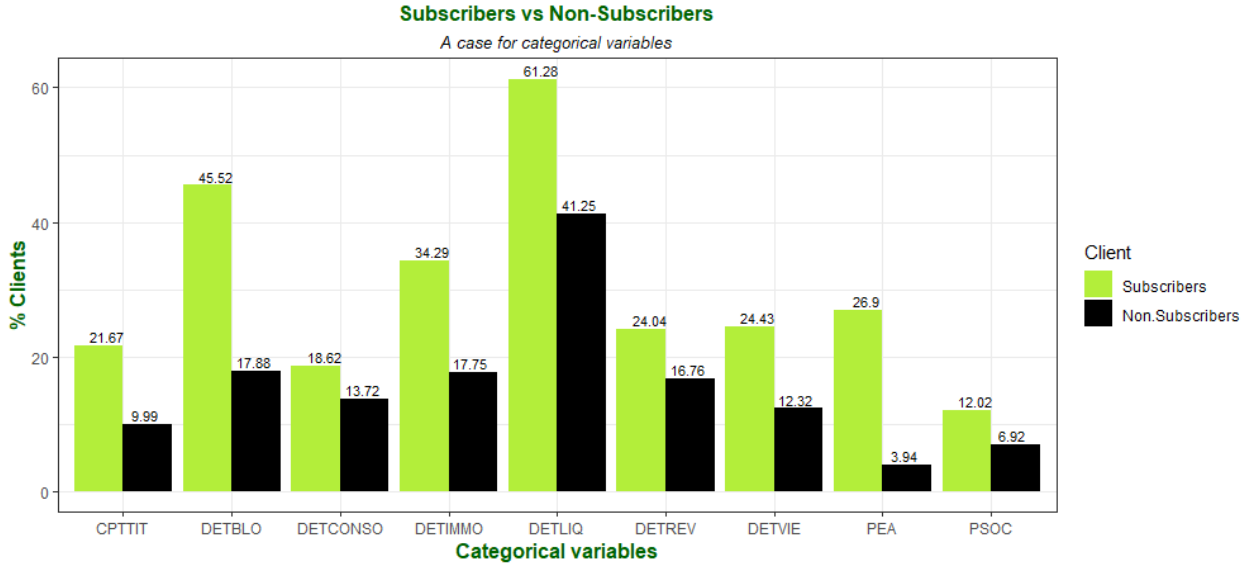
Variable	n	mean	sd	median	min	max	skew	Type
AGE	6249	43.32	16.67	41	19	100	0.62	Continuous
ANCCDD	6249	10.42	5.65	10	0	22	0.07	Continuous
SURFIN	6249	14088.65	32035.34	2626.92	0	654089.3	6.33	Continuous
SOLMOY	6249	2009.21	12495.71	768.71	-27295.2	632365.9	43.03	Continuous
FLUCRE	6249	1784.79	5648.18	1167.71	0	266520.5	36.9	Continuous
FLUDEE	6249	-1844.27	3313.26	-1143.42	-89927.2	75.42	-11.79	Continuous
NBCRE	6249	3.25	2.59	2.67	0	20	1.34	Continuous
NBDEE	6249	18.15	16.14	14.67	0	110	1.14	Continuous
NBJDE	6249	5.13	8.6	0	0	56.33	1.88	Continuous
DETVIE	6249	0.14	0.35	0	0	1	2.04	Categorical
DETIMMO	6249	0.2	0.4	0	0	1	1.47	Categorical
DETCONSO	6249	0.15	0.35	0	0	1	2.01	Categorical
DETREV	6249	0.18	0.38	0	0	1	1.67	Categorical
SLDLIQ	2784	4919.24	7791.77	2908.37	0	204349.6	10.86	Continuous
DETLIQ	6249	0.45	0.5	0	0	1	0.22	Categorical
CPTTIT	6249	0.12	0.32	0	0	1	2.35	Categorical
PEA	6249	0.08	0.27	0	0	1	3.18	Categorical
PSOC	6249	0.08	0.27	0	0	1	3.16	Categorical
SLDTIT	743	18858.87	31284.44	10081.5	0	381868.5	5.24	Continuous
SLDPEA	479	12291.24	22083.44	3600.11	0	202050.6	3.81	Continuous
DETBLO	6249	0.22	0.42	0	0	1	1.33	Categorical
SLDBLO	1398	14296.06	29366.85	2335.1	0	381868.5	4.73	Continuous

2.2.2. Comparison between subscribers and non- subscribers

a. Group of categorical variables:

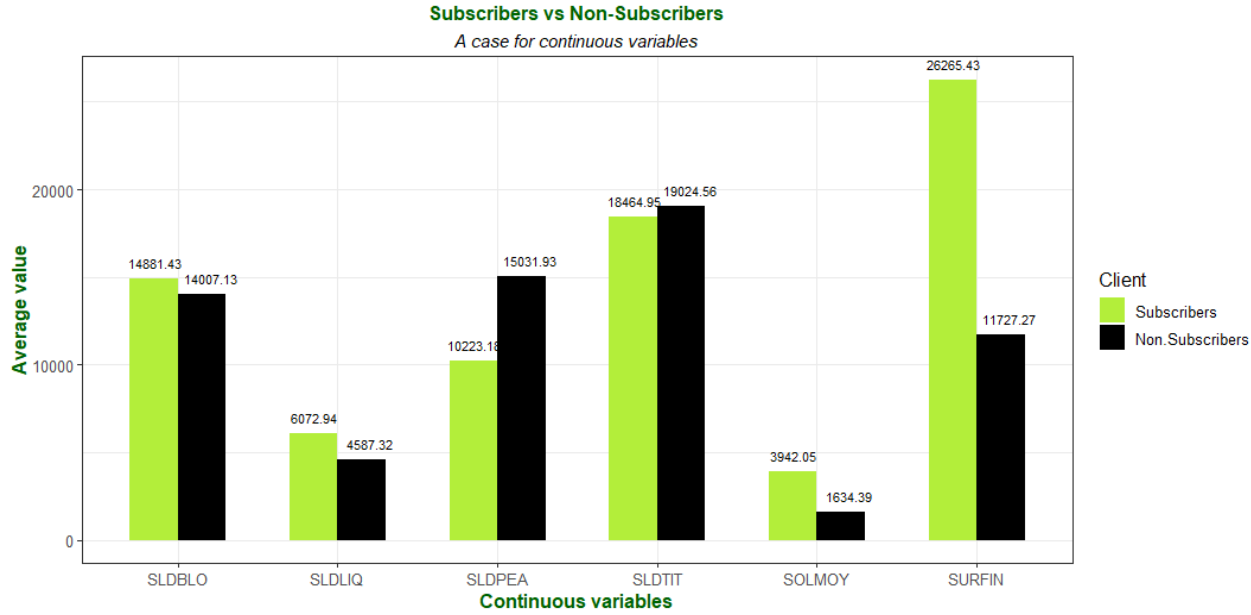
From the graph, it can be seen that on average, the percentage of clients using certain types of products in group of subscribers is always higher than the group of non- subscribers. The biggest difference is for the locked saving account (DETBLO) where 45.54% of subscribers use the product while that rate in the counterpart group is only 17.88%. Such difference is also observed in the life insurance contract (PEA) where the corresponding rate of users in two

groups are respectively 26.9% and 3.94%. Hence, it could imply that the more clients use products like DETBLO and PEA, the more likely subscription occurs.



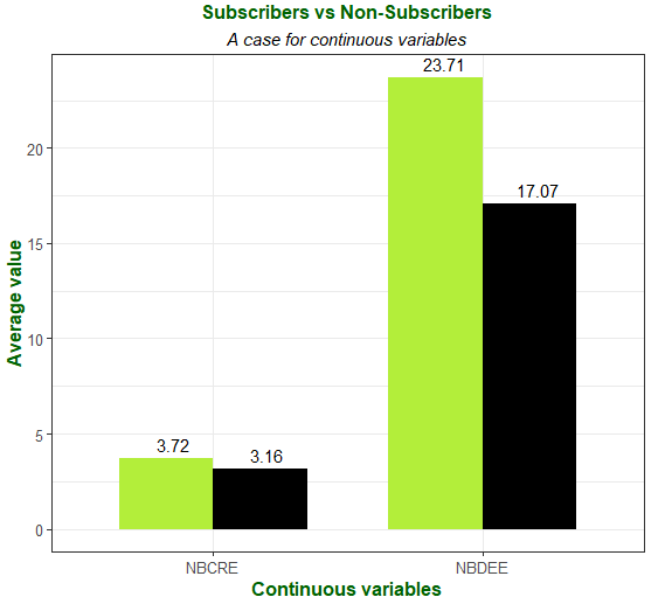
b. Group of continuous variables:

For the continuous variables SLDBLO, SLDIQ, SLDPEA, SLDTIT, SOLMOY, on average there is not much difference between group of subscription and group of non- subscription. That is not the case for SUFRIN (Total saving account) where the average of saving in group of subscribers is more than 26,000 while the figure in the other group is below 12,000 – a half of the counterparts. This might imply clients have higher saving could likely subscribe to the fund.



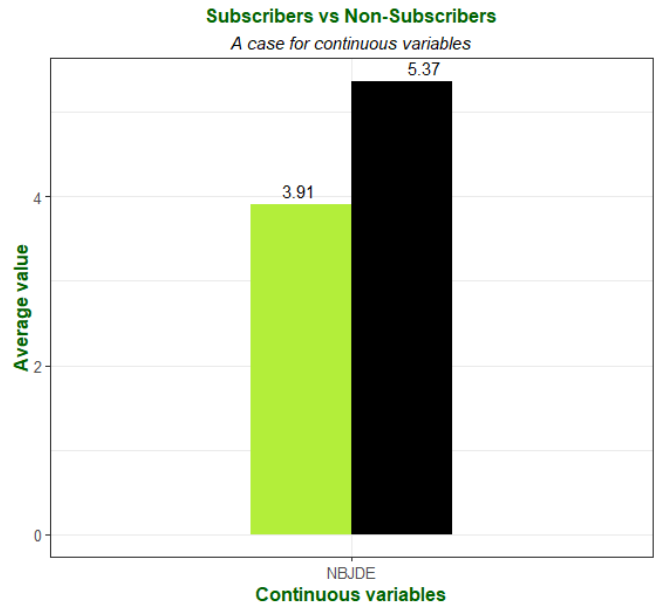
Little difference is also observed in variables NBCRE, NBDEE, AGE and ANGCDD between two groups. So the total number of transactions, whether it is debit or credit, the customer age

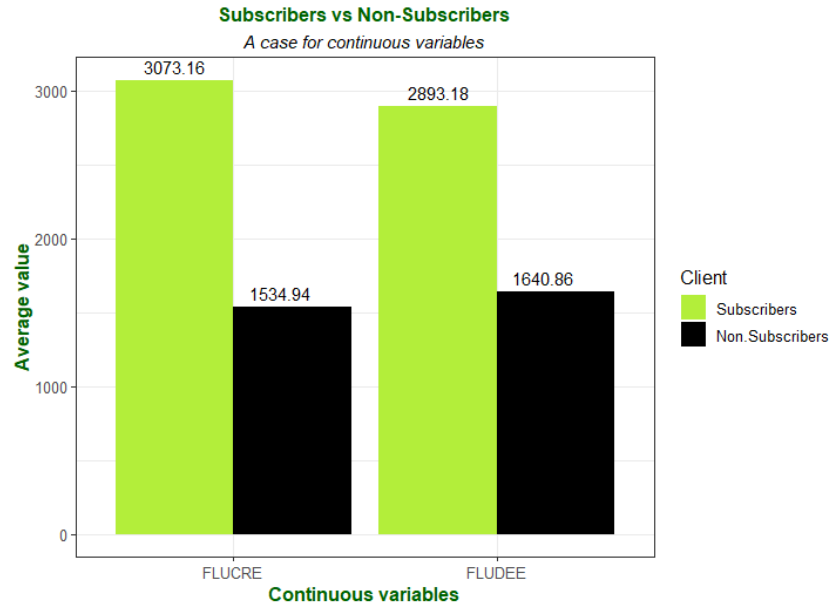
or the length of customer relationship might have little impact on the subscription decision of clients.



For the case of NBJDE (Volume of day debtor), non-subscribers have 5.37 days of debt while subscribers only have 3.91 days, almost a half. So the more the day debtors one client has, the more unlikely he/she subscribes to the fund.

Last but not least, there is a clear difference for variables FLUCRE and FLUDEE between two groups. In both cases, subscribers on average have double cash flow more than the other group. So higher cash flow encourages clients to subscribe the fund.



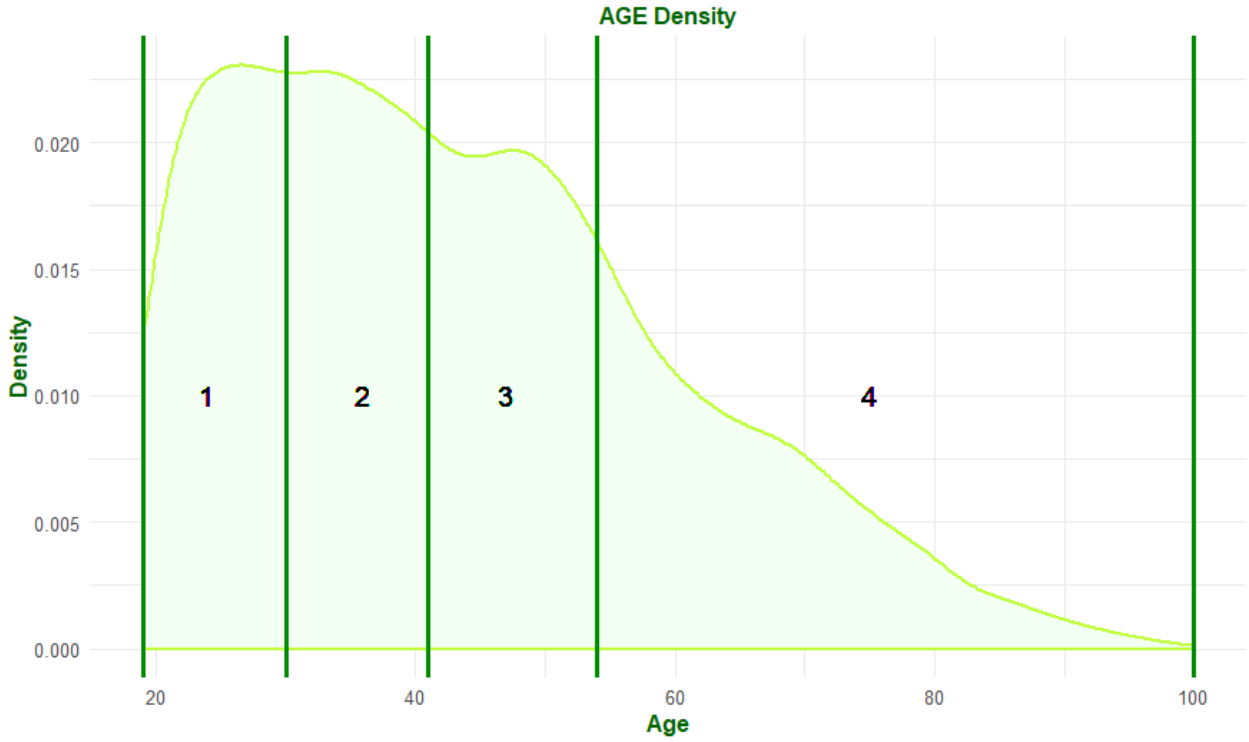


3. UNIVARIATE ANALYSIS

3.1. Discretize continuous (numeric) variables

Every continuous variable is transformed in to several categories in order to adapt logistic regression - one of our models. Generally, one continuous variable is divided into four groups: first quarter, second quarter, third quarter and fourth quarter.

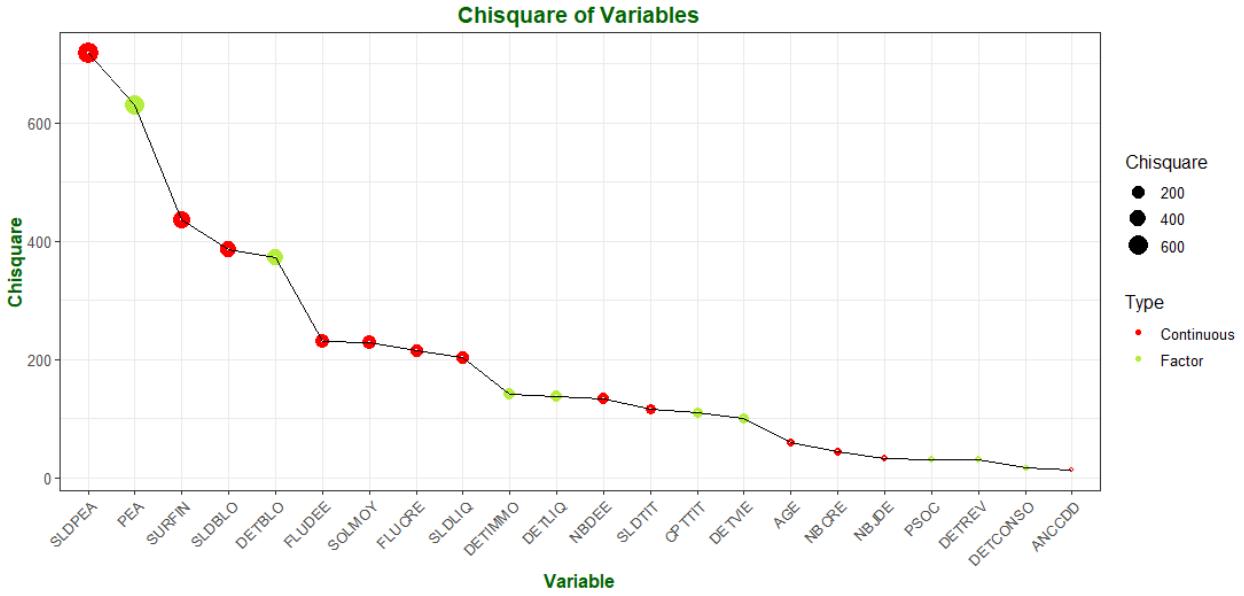
For example, variable AGE of a client is converted into group 1 if the client is from 19 (min) to 30 (first quantile) years old. Similarly, variable AGE of a client is assigned to group 2 if age of the client is from 31 (first quantile) to 41 (second quantile, median).



Apart from categories 1, 2, 3 and 4 respectively to four quarters, category "NA" is introduced when there is missing value. Moreover, variables SLDPEA, SLDBLO and NBJDE have zero as the majority value. For those three variables, zero will be the first category (group 0). The remaining values are split in to four categories respecting to the quarters (not taking zero into account when divide quarters).

Variable	% Zero (Ignore NA)	Categories
SLDPEA	53%	0, 1, 2, 3, 4, NA
SLDBLO	27%	0, 1, 2, 3, 4, NA
NBJDE	29%	0, 1, 2, 3, 4, NA
Others		1, 2, 3, 4, NA

3.2. Importance of variables

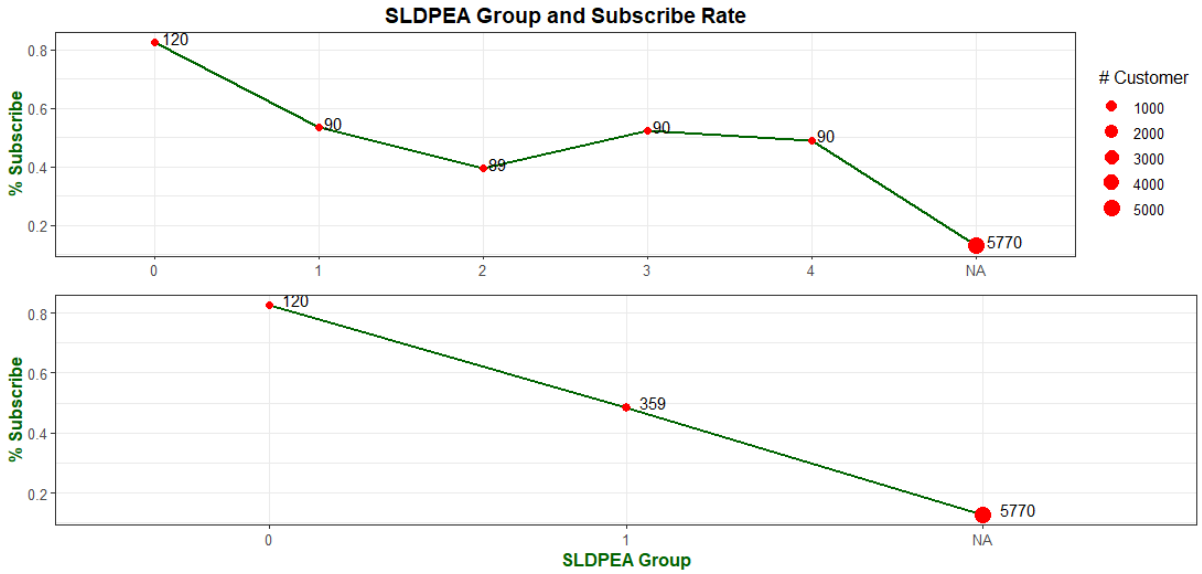


Chi-square statistic is employed to measure the effect of one variable to the subscribe rate. Accordingly, SLDPEA (amount of PEA) has the strongest effect on subscribe rate of clients. The following significant features are PEA (a PEA customer), SURFIN (Total saving), SLDBLO (Amount of locked saving), etc.

3.3. Examine more on important continuous variables

Based on chi-square statistics, variables having highest chi-square values will be more examined in order to diminish the information loss. Particularly SLDPEA, SURFIN and SLDBLO will be detail investigated.

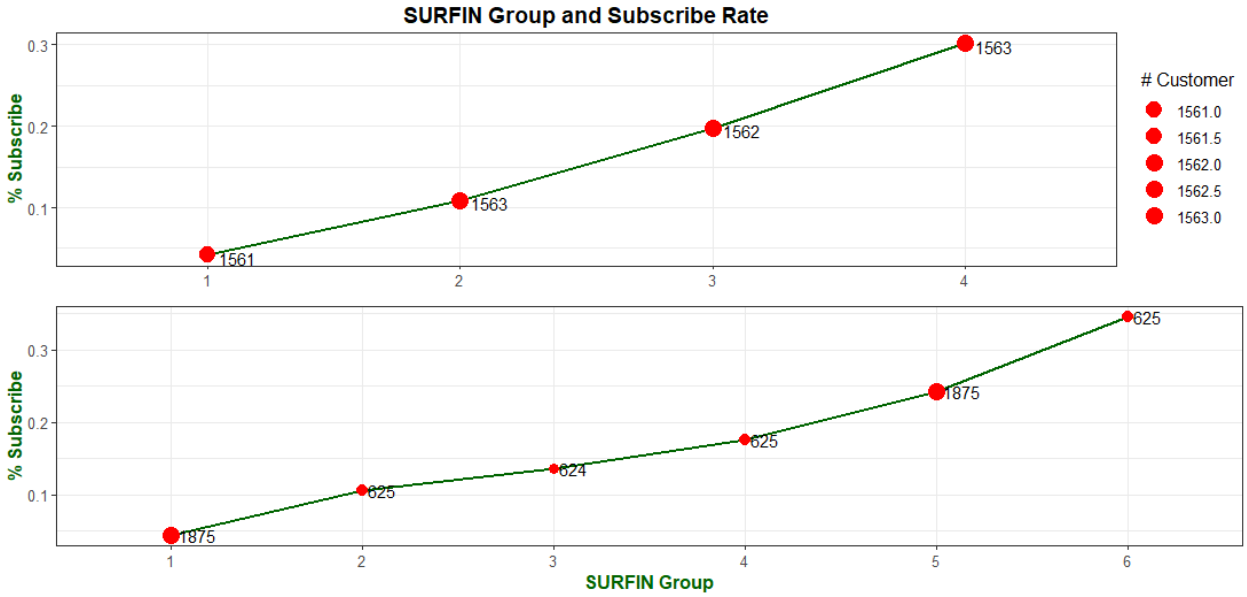
3.3.1. SLDPEA



The first figure demonstrates the evolution of subscribe rate in different quarter groups – the first approach of categorizing. Subscribe rate of clients in groups 1, 2, 3 and 4 seem to be indifferent. Moreover, each of these group contains only around 90 clients. Therefore, we decided to join group 1, 2, 3 and 4 together. As a result, the second figure shows the final categories of variable SLDPEA:

- **Category 0**: SLDPEA = 0
- **Category 1:** SLDPEA ≥ 1
- **Category NA**: SLDPEA missing

3.3.2. SURFIN

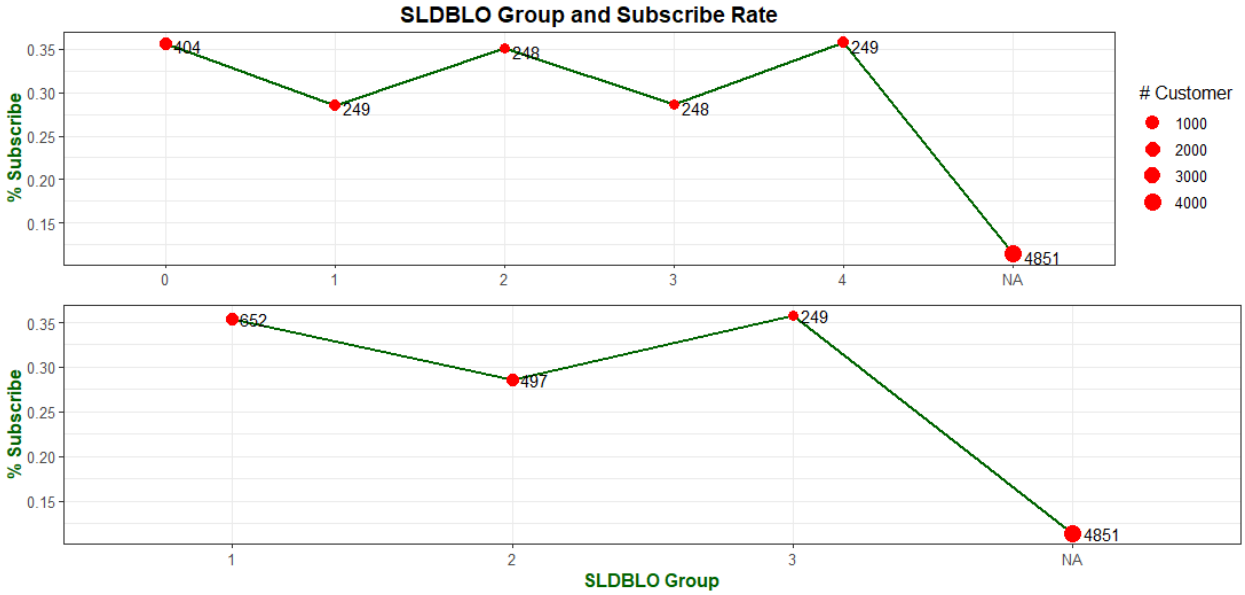


Unlike SLDPEA, variable SURFIN needs to be broken in more categories since the four quarter groups seems be too robustly distinguished. From four quarters groups at the beginning, SURFIN is split into six smaller groups as following:

- **Category 1:** 0 ≤ SURFIN < 81 (0.3-quantile)
- **Category 2**: 81 ≤ SURFIN < 882 (0.4-quantile)
- **Category 3:** 882 ≤ SURFIN < 2,627 (0.5-quantile)
- **Category 4:** 2,627 ≤ SURFIN < 5,891 (0.6-quantile)
- **Category 5:** 5,891 ≤ SURFIN < 37,986 (0.9-quantile)
- **Category 6:** 37,986 ≤ SURFIN < 654,089 (max)

3.3.3. *SLDBLO*

Recall that "group zero" are clients who have zero of SLDBLO. Group 1, 2, 3 and 4 are four quarters respectively, ignoring zero and missing value.

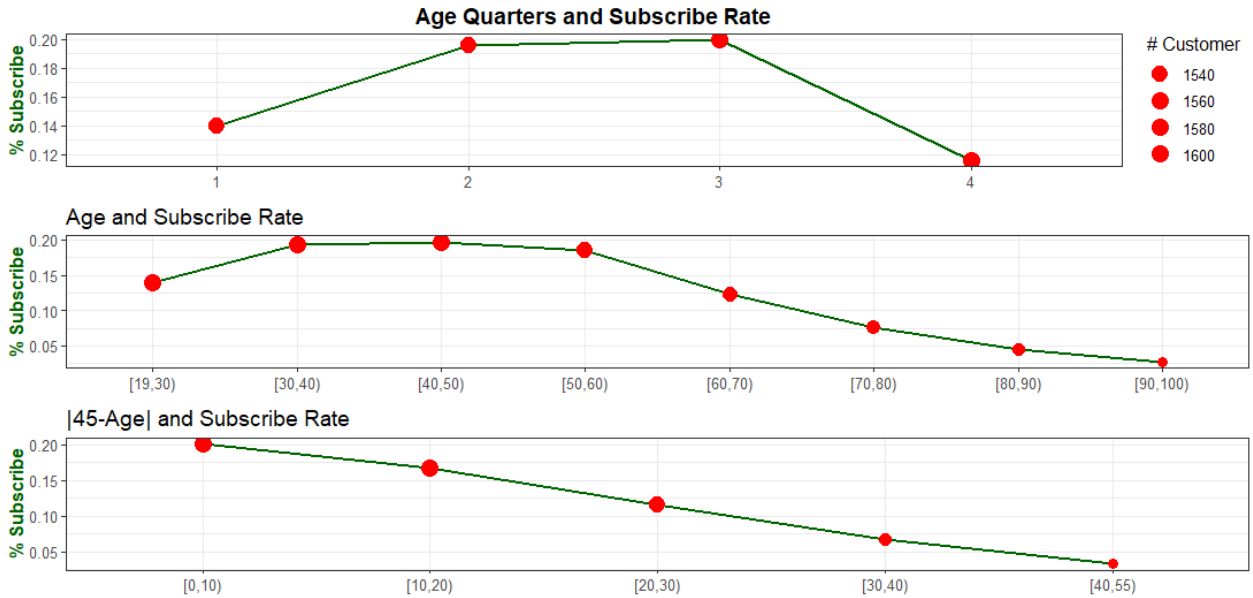


We put clients in old "group 0" and old "group 2" of SLDBLO together (becoming new "group 1") since subscribe rates of these two group are similar. Likewise, new "group 2" are composed by old "group 1" and old "group 3". Whereas old "group 4" and old "group NA" are kept unchanged, they are just renamed into "group 3" and NA respectively.

- **Category 1:** SLDBLO = 0 | 2,085 ≤ SLDBLO < 7,052
- **Category 2:** 20 ≤ SLDBLO < 2,085 | 7,051 ≤ SLDBLO < 19,665
- **Category 3**: 19,665 ≤ SLDBLO < 202,050
- **Category NA:** SLDBLO missing

3.3.4. *AGE*

Although AGE is not among top important variables, we find it as an interesting illustration of data transformation before applying statistical model.



The first plot demonstrates four groups according to four quarters of variable AGE. The second plot shows a different type of AGE grouping, which has 10 years at each step width. Apart from the groups AGE>70, this second plot seems to be symmetric at AGE=45. Therefore, we create a third way of grouping which takes the |45 - AGE| then splits into 10-year-steps. As a result, we get a satisfactory linear relationship between subscribe rate and |45 - AGE| groups.

3.4. Feature selection

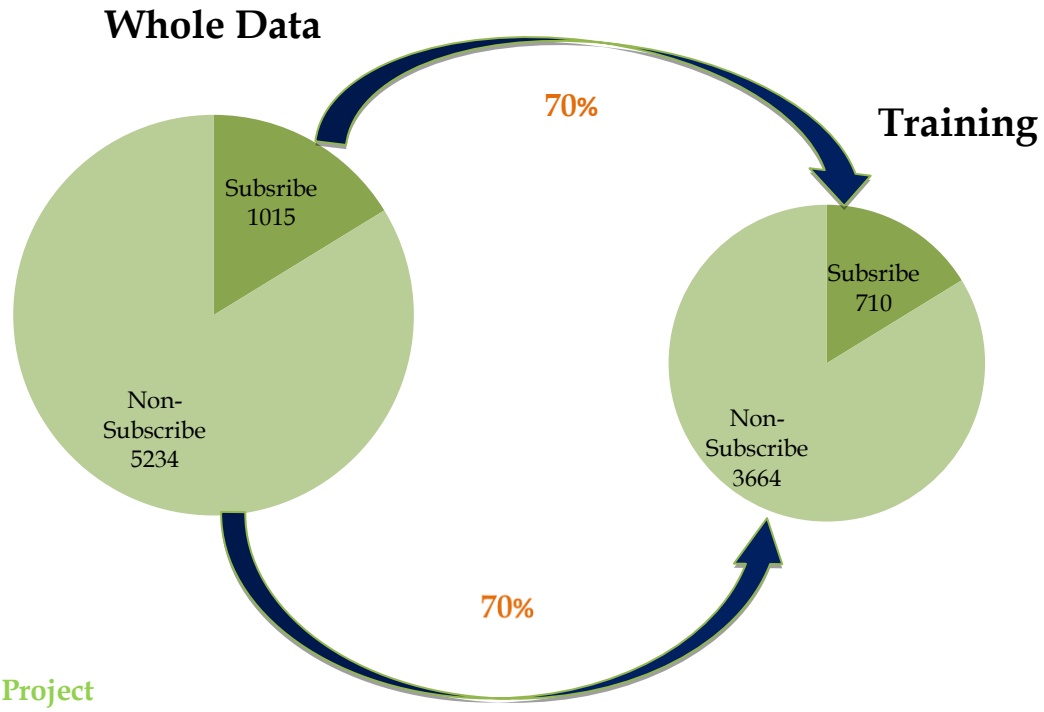
Step-wise feature selection is used in order to select the best model, in term of AIC criteria. Accordingly, eight continuous variables (which are already transformed to multiple modalities factor) and one factor variable are chosen.

Variable	Description
SLDPEA_group	Amount of PEA
SURFIN_group	Total saving
NBDEE_group	Volume of crediting transaction
DETIMMO	Customer with Housing Loan (1/0)
AGE_group	Age
SOLMOY_group	Average position of checking account
ANCCDD_group	Seniority Customer
SLDTIT_group	Amount of trading account
NBJDE_group	Volume of day debtor

4. STATISTICAL MODELS

4.1. Split training set and testing set

Before performing the analysis, in order to ensure the subscribe rate in the training set and testing set are equivalent, we take 70% observations which have Y=1 and 70% observations which have Y=0. As a result, the training set (testing set) has 70% (30%) size of whole data set and obtains subscribe rate exactly identical with the whole data set. We are going to train the model on the training set, then evaluate and compare models based on testing set.



4.2. Theoretical background

In this section, we are going to apply three models for the prediction.

Random Forest: An ensemble learning method for classification. This algorithm generates multiple bootstrap training sets from the original training set and uses each of them to generate a classifier for inclusion in the ensemble. In general, bagging does more to reduce the variance in the base models than the bias, so bagging performs best relative to its base models when the base models have high variance and low bias. Decision trees are unstable, which explains why bagged decision trees often outperform individual decision trees.

Logistic Regression: An appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

XGBoost (eXtreme Gradient Boosting): This algorithm involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified. Boosting does more to reduce bias than variance. For this reason, boosting tends to improve upon its base models most when they have high bias and low variance. Examples of such models are Naive Bayes classifiers and decisions tumps. Boosting's bias reduction comes from the way it adjusts its distribution over the training set.

4.3. Model Construction

4.3.1. Logistic Regression

Details of logistic model is presented as follows:

```
Call:
glm(formula = Y ~ SLDPEA_group + SURFIN_group + NBDEE_group +
    DETIMMO + AGE_group + SOLMOY_group + ANCCDD_group + SLDTIT_group +
    NBJDE_group, family = "binomial", data = train_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5442  -0.5908  -0.3752  -0.2245   3.0865

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.11634    0.47220  -0.246 0.805393
SLDPEA_group1  -2.05350    0.35818  -5.733 9.86e-09 ***
SLDPEA_groupNA -3.34179    0.32929 -10.149  < 2e-16 ***
```

```
SURFIN_group2   0.79943    0.21462   3.725 0.000195 ***
SURFIN_group3   1.06243    0.20207   5.258 1.46e-07 ***
SURFIN_group4   1.12864    0.19987   5.647 1.63e-08 ***
SURFIN_group5   1.64405    0.16430  10.006  < 2e-16 ***
SURFIN_group6   2.06263    0.20954   9.844  < 2e-16 ***
NBDEE_group2    0.53366    0.15645   3.411 0.000647 ***
NBDEE_group3    0.51017    0.15871   3.214 0.001307 **
NBDEE_group4    0.69437    0.16543   4.197 2.70e-05 ***
DETIMMO1        0.51915    0.10910   4.758 1.95e-06 ***
AGE_group2     -0.15980    0.10807  -1.479 0.139245
AGE_group3     -0.22596    0.14175  -1.594 0.110929
AGE_group4     -1.52446    0.38895  -3.919 8.88e-05 ***
AGE_group5     -2.54494    1.02504  -2.483 0.013037 *
SOLMOY_group2   0.43083    0.20514   2.100 0.035713 *
SOLMOY_group3   0.78610    0.21736   3.617 0.000298 ***
SOLMOY_group4   0.82895    0.22592   3.669 0.000243 ***
ANCCDD_group2  -0.01606    0.14505  -0.111 0.911816
ANCCDD_group3  -0.10433    0.14375  -0.726 0.467978
ANCCDD_group4  -0.35451    0.14477  -2.449 0.014331 *
SLDTIT_group2  -0.58400    0.31365  -1.862 0.062610 .
SLDTIT_group3  -0.71837    0.31252  -2.299 0.021526 *
SLDTIT_group4  -0.96682    0.32634  -2.963 0.003051 **
SLDTIT_groupNA -0.60482    0.21518  -2.811 0.004942 **
NBJDE_group1    0.11740    0.15009   0.782 0.434092
NBJDE_group2    0.07455    0.16566   0.450 0.652695
NBJDE_group3    0.10337    0.20320   0.509 0.610959
NBJDE_group4    0.70498    0.24141   2.920 0.003497 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3879.8  on 4373  degrees of freedom
Residual deviance: 3113.6  on 4344  degrees of freedom
AIC: 3173.6

Number of Fisher Scoring iterations: 6
```

4.3.2. *Random Forest*

Details of Random Forest model is as follows:

```
Call:
 randomForest(formula = Y ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 13.97%
Confusion matrix:
     0   1 class.error
0 3587  77  0.02101528
1  534 176  0.75211268
```

4.3.3. *XGBoost*

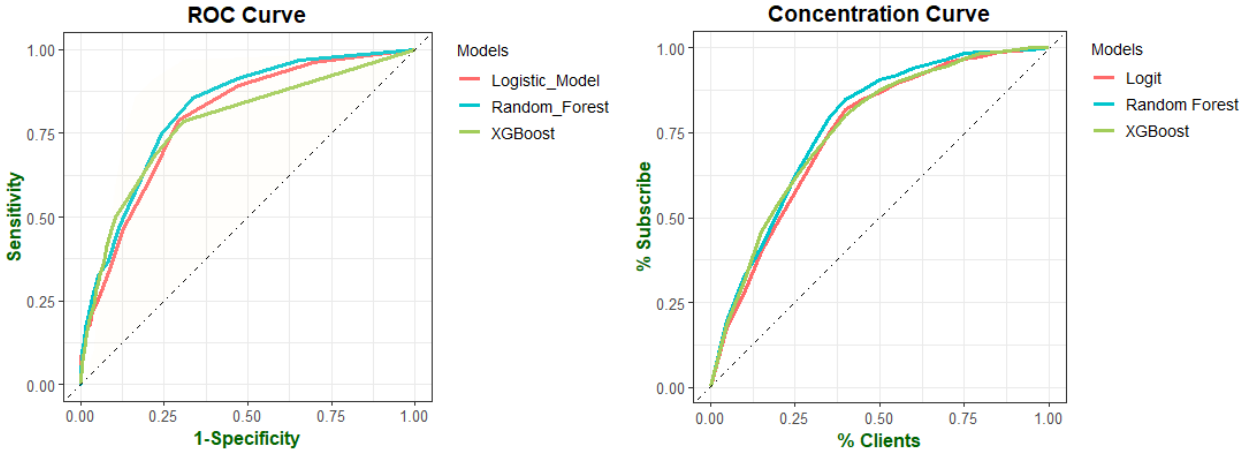
Details of XGBoost model is as follows:

```
call:
  xgb.train(params = params, data = dtrain, nrounds = nrounds,
    watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
    early_stopping_rounds = early_stopping_rounds, maximize = maximize,
    save_period = save_period, save_name = save_name, xgb_model = xgb_model,
    callbacks = callbacks)
params (as set within xgb.train):
  objective = "1", eta = "0.15", max_depth = "5", colsample_bytree = "0.8", min_child_weight = "3",
subsample = "1", silent = "1"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 22
niter: 500
nfeatures : 22
evaluation_log:
    iter train_error
       1    0.137403
       2    0.133059
---
```

```
499  0.000457
500  0.000457
```

4.4. Evaluation and Model Selection

In order to avoid over-fitting, the evaluation would be carried on the test set instead of the training set. We evaluate the model based on *(1)* ROC Curve and *(2)* Concentration Curve.



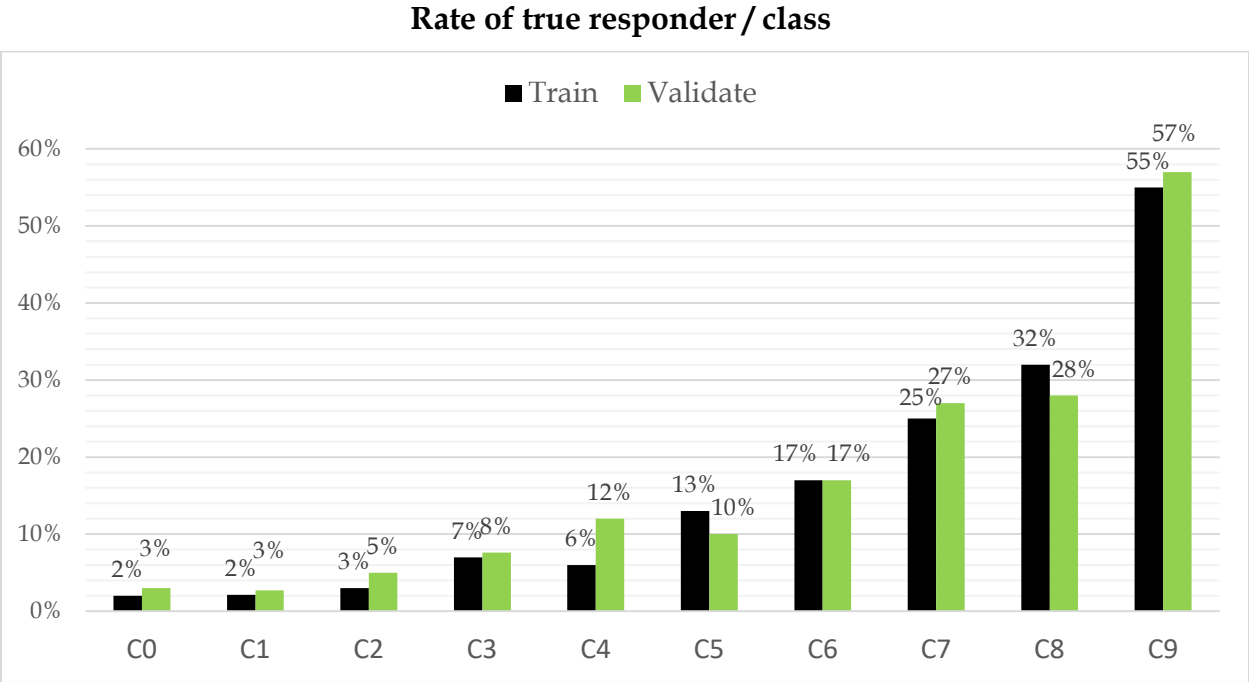
Regarding ROC curve, the graph shows that Random Forest performs the best as it is the closet curve that follows the left-hand border and then the top border of the ROC space for most of the time. Indeed, we obtain the AUC as follows:

Model	AUC
Logistic Regression	0.7048856
Random Forest	0.7469692
XGBoost	0.6972765

Regarding the Concentration Curve, if we look at the left side of the graph, with the same number of clients being contacted, Random Forest deliver the highest subscription rate for most of the time compared to other models.

So we could conclude that Random Forest is the best model to predict the customer behavior for this dataset.

4.5. Model Application



Training set and Validating set are split into to 10 classes which are corresponding to 10% size of the data. Subscribe rate of every class is presented in the "Rate of true responder" bar chart above. Accordingly, if the company decides to run the marketing campaign on 10% of the clients, approximately 57% of them will purchase the fund. The next 10% of the clients are expected to have subscribe rate of 28%, etc. As a result, a Marketing Department Manager should base on the resources of the company and the chart above to decide how many clients the department should cover on the campaign. As consultant party, if the marketing campaign is not big, we recommend the company can approach 10% of the clients because the subscribe rate of this class is far higher than the following groups.

5. CONCLUSION

It can be inferred from the aforementioned analysis that in general discretized groups from original variables have better discrimination effect. And hence, they are better variables to be input in to the classification model. Selected variables include: SLDPEA_group, SURFIN_group, NBDEE_group, DETIMMO, AGE_group, SOLMOY_group, ANCCDD_group, SLDTIT_group and NBJDE_group.

Three models are tested and based on both ROC curve and Concentration Curve, Random Forest outperforms the rest. Although choosing how much the market size for the marketing campaign is totally based on the strategy of the company, rate of true responder chart is a significant guideline to the manager to make decision.