Documentation For profitability

**Summary**

The question answered by profitability model is how likely is the customer going to spend $1000 in the first three months after activation. Debit usage data in the first three months for new customers in 2016 is used to model the behavior of customers who are going to activate their card this summer.

**Assumptions before modeling**

To build the model, we are making an assumption that the spending behavior in the first three months for targeted customers will be similar to customers activating and using their cards without reward. It’s very easy for customer to spend more than $250 in the two and half months, so our reward is more aimed to encourage activation rather than spend.

Another assumption we are making is that spending $1000 in the first three months by each customer is a reasonable goal for us to break even. To get this number, a simple break-even test was conducted. For every $100 spent by customer, we get about $1.5 revenue, which means that 1600 (25÷1.5%) is the number we need to break even in six months if we don’t consider time value of money and other benefits gained from this campaign like enhanced brand awareness. We can further split the goal into spending $1000 in the first three months and $600 in the last three months since people usually spend more when they just activate their cards and setting a higher goal would help reduce the risk of not getting the investment back.

**Overview of dataset**

The demographic data of the new customers opening their account in January, February, March and April are extracted from data base along with their debit usage data in the first months after their activation. A new label was created to indicate if the person spent over $1000 in the first three months after activation. For example, for customers who activated their debit card at the beginning of January, if the total spend of this person in January, February, March is over 1000, then we label him as 1, if not, his label would be 0.

The raw data for the profitability model can be found here G:\Analytics\Consumer Debit Card\2016 Campaign\Question2\Combined list

**Modeling**

**Preparation Stage**

1. Recode the variables and data types

Some binary data are represented in 0, 1 format. To avoid R treating these variables as numeric variables, it’s necessary to recode them into categorical variables.

Some of other variables are defined as a wrong type. Before modeling, we need to make sure the data type of every variable is correct.

1. Manually variable screening

R provides many packages and function to automate the variable screening process. However, building a model is not only about finding statistically significant relationship underlying the data, but also using the business judgement and business rule to valid the relationship in the model.

Pre-selection of variables is important process and it helps us to exclude some irrelevant variables at the first stage. Creating plots and doing individual logistic regression are two methods I used to conduct pre-selection.

After manual selection, estimated\_income, SafeDopsitFlag, DebitCard Flag and ClosestBranchDistance are removed from our candidate variables list. Removal of estimated\_income is due to the insignificance of the variable itself and also the overlapping information provided by NASbuck and Homeowner variable.

Some variables appeared to be insignificant, but we still retained them in the model, including Deposit.Balance and Loan.Balance.

1. Check Multicollinearity

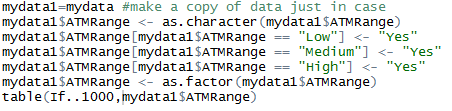
If one variable is multicollinear with other variables, it means that the variable can be represented as the linear combination of other variables and we need to exclude this variable. LoanFlag was deleted according to VIF test.

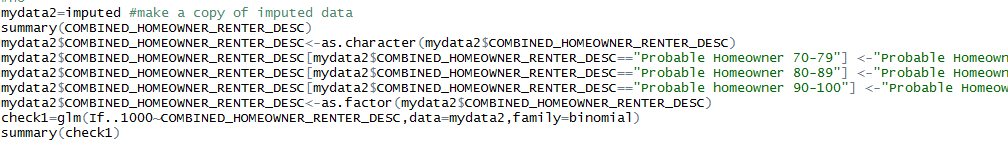
1. Data Cleaning and Imputation

There are three variables having missing values in the raw data: NASbuck (13%), HO (20%) and Education (13%). Imputation is used to handle the missing values.

**Modeling process**

1. Combining some insignificant levels are necessary to reduce model complexity.





As the example shows above, some insignificant levels of ATMRange and HO are combined.

1. Identifying influential observations

Outliers in the raw data will cause inaccurate estimation of coefficients. It’s necessary to remove the influential observations and fit the model to once more see the difference. After removing influential observations, some variables that were insignificant before became significant, for example, deposit balance.

1. Remove RMage

RMage appeared to be extremely insignificant although we have removed the influential observations. So I decided to remove this variable.

1. Try second-order terms

If interaction exists between two variables, then we can say these two variables exert influence on each other. For example, we consider two variable ATMRange and Age, for people whose ATM usage are low, Age is required to decrease 5 for probability of response to increase by 1%. But for high usage group, we may only need a decrease of 3 in age to yield an increase of 1%. Interaction effect is not considered by linear model, so we need to test if any interaction terms are significant and add them into model.

In this model, three interaction terms are found significant after testing. CustomerAge\*BranchRange+DigitalIndflag\*Deposit.Balance+Ann.Pre.Tax.Profit\*BillPay

1. Evaluation

Final model:

model11=glm(If..1000~CustomerAge+Ann.Pre.Tax.Profit+ATMRange+BillPay+BranchRange+DigitalIndflag+NASBcuk+Deposit.Balance+Loan.Balance+COMBINED\_HOMEOWNER\_RENTER\_DESC+EDUCATION+I(CustomerAge^2)+CustomerAge\*BranchRange+DigitalIndflag\*Deposit.Balance+Ann.Pre.Tax.Profit\*BillPay,data=mydata4,family = binomial)

Above is the final profitability model. However, this is not the end of modeling process, the last step would be to evaluate how effective this model is. So I calculate auc value of the model. Roc means the overall classification rate for the model, sensitivity (true positive) means the percentage the true 1 are classified as 1 in the model, and specificity (true negative) means the percentage of the true 0 are classified as 0 in the model. We are supposed to pay more attention on sensitivity rate in this case because the cost of missing a potential respondent is much larger than the cost of sending one more email to a person who is not going to spend over 1000 dollars in the first three months after promotion (email is costless!).

