**Documentation For Response Model**

**Summary:**

The question we would like to address here is whether a customer will respond to our $25 cashback offer. To do this, data from the 2014 debit card campaign is used for modeling to identify what differentiate a respondent from non-respondent. In 2014, we sent out 10394 promotion emails to customer and got 159 responses (1.5%).

**Assumptions Before modeling:**

It should be noted that there are few assumptions we need to make before modeling. The model will be built on the data from 2014, with people who responded as y=1 group and who didn’t respond as y=0 group. However, the list was already after selection by the model John built in 2014, which may not be a counterpart of the group we will apply our model to in August. So while we are making an assumption that the customer list in 2014 is an approximate to checking account customer we are targeting, we need to bear in mind that this many induces some biases.

Another assumption we need to make is aimed to deal with unbalanced dataset. As you may see, 159 vs. 10000+ indicates the dataset is really unbalanced and it also shows that the threshold we want to be beat is not 0.5 (SAS set 0.5 as a default cut-off point). Rather, if a person with a probability obviously higher than 1.5%, we may have confidence to believe he will be more likely to respond than other customers. The goal of this model is to outperform random selection rather than achieve a dramatically high response rate.

**Overview of Variables**

The following is a snapshot of some data points in the dataset, just to give an idea of how the data look like.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CIFKEY** | **Online Range** | **ATM Range** | **NASbuck** | **HOBucket** | **BillPay** | **Branch Range** | **SumBookBal** | **If.response** | **RMAge** | **DigitalInFlag** | **Age** |
| 14720 | No | High | 100-249k | Likely | FALSE | Medium | 6363.3 | No | 20.4 | TRUE | 61 |
| 14571 | No | Medium | 100-249k | NA | FALSE | High | 108539.02 | No | 22.8 | FALSE | 53 |
| 14481 | No | Medium | <25k | Likely | TRUE | High | 242.34 | No | 42.3 | TRUE | 76 |
| 13297 | No | No | 250-499k | NA | TRUE | Medium | 542.73 | No | 14.0 | TRUE | 50 |
| 12978 | High | High | 500-749k | Likely | TRUE | High | 4482.8 | No | 6.0 | TRUE | 58 |
| 12769 | High | No | 100-249k | Likely | FALSE | High | 4029.64 | No | 7.8 | TRUE | 66 |

**Modeling**

* Preparation Stage

1. Check Multicollinearity

If one variable is multicollinear with other variables, it means the variable can be represented as the linear combination of other variables and we need to exclude this variable. Fortunately, multicollinearity is not a problem for this model according to VIF test.

1. Data Cleaning and Imputation

There are two variables containing missing values: NASbucks and HOBucket.

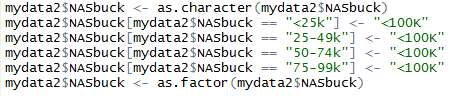
|  |  |  |
| --- | --- | --- |
| Variable Name | Missing | Rate |
| NASbuck | 242 | 2.3% |
| HOBucket | 1754 | 16.9% |

There are few ways to deal with missing variables, including deleting observations with missing values, deleting variables with missing values, or do data imputation. In this case, multiple imputation is chosen to handle missing values. To do that, we need to make another assumption about the missing pattern for data. The premise of applying imputation is that the probability of being missing is not related to the value of a variable itself. For example, if people with low income have the tendency to avoid reporting their income range than high incomer, applying data imputation in this case would be a problem. In this case, however, we have no idea of how Expedia collected data. So we assume that they collected NASbuck and HOBucket indirectly from customer, and thus missing rate is not related with data value itself.

* Modeling Process

1. Combine insignificant levels

I ran logistic regression on each variable to see how they are related to response variables individually. Then I combined some factor levels that appear not to be significant.



For example, I combined some levels for NASbuck to reduce variable complexity.

1. Identifying influential observations

After running a logistic model on all candidate variables, most of the variables appear to be insignificant, even variables like SumBookBal that have been proven to be useful. So I tried identifying influential observations (potential outliers) and fit the model again with these data points. Excluding influential observations reduce the variance and made some variable become significant.

1. Excluding insignificant varaibles

Some variables may be of little importance in predicting response. In this case, p-value of OnlineRange stays very high even I removed influential observations. So we decided to remove this variable.

1. Try second-order terms

The relationships between response variable and predictors are not necessarily linear. So it is necessary to add second-order terms for numeric predictor to test if they are significant. In this Case, none of three candidates second-order term are significant, so we don’t have to worry about that.

1. Interaction

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.

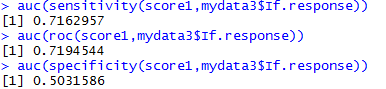
In this model, two interactions are found significant after testing. NASbuck and BranchRange as well as ATMRange and BranchRange. Thus, we add interaction terms into the model NASbuck\*BranchRange+ATMRange\*BranchRange.

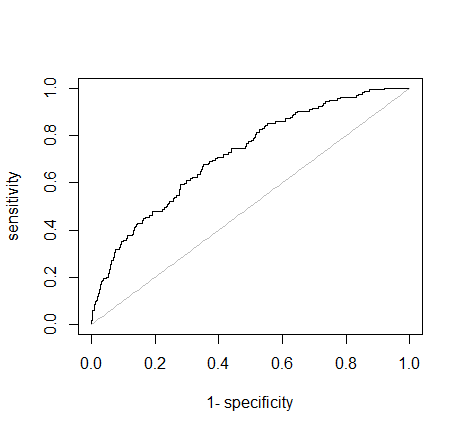
1. Evaluation

**Final Model:**

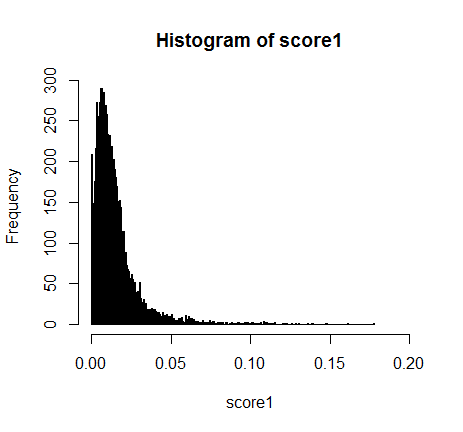
model17=glm(If.response~Age+ATMRange+NASbuck+HOBucket+BillPay+BranchRange+SumBookBal+RMAge+DigitalInFlag+NASbuck\*BranchRange+ATMRange\*BranchRange,data=mydata3,family = binomial)

Above is the final response model. However, this is not the end of model building, we still need 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 respond (email is costless!). So our goal becomes how to maximize the sensitivity given a certain number of email sent.





As I mentioned before, the predicted probability of response should be very low because of unbalanced sample. The plot of probability confirms my expectation. All the predicted probabilities are below 20%. If the cut-off point is set to 1.5%, we have following classification table.





If the cut-off point is set to 1.1%, we have classification table as follow

