Prediction of Policyholders' Interests on Cross-Sell Product -- Vehicle Insurance

by

M.S., Wenxiao Zhou, Yisha Zhou, Jiayi Zhou, Yuxin Tang

University of Connecticut, 2020

Final Project

Math 5637 -- Statistics for Actuarial Modeling

Advisor: Emiliano A. Valdez

Fall 2020

Executive Summary

Insurance is a way of managing risks which protects against the financial risks that are present at all stages of people's lives and businesses. From the aspect of insurance companies, finding the target customers by their characteristics will make the sellings efficient. From the aspect of customers, they give priority to programs with a trusted insurance company and reasonable premiums. In this paper, we aim to build a predict model to detect our clients' interests in our new product-Vehicle Insurance when they already have enrolled in the Health insurance in our company.

Firstly, according to the comparisons between several methods of imbalanced problem solving, we take advantage of the SMOTE algorithm to get a balanced training group with people who are interested in the new insurance and not. Further fitted models are basing on this sample. And then, we construct the Logistic regression model to estimate whether a customer is interested in the insurance with AUC = 0.785 under ROC curve. After that, the Balanced Random Forest algorithm helps to significantly improve the model accuracy with AUC = 0.851 under ROC curve. Finally, after combining information from clustering analysis, we make some conclusions to our target customers: with a driver's license, pays high premiums for health insurance, mostly people aged between 40 and 60, the vehicle age is 1 - 2 years, and the car got damaged in the past.

Section 1. Introduction

As an industry, insurance regards as a slow-growing, safe sector for investors. Insurance companies base their business models around assuming and diversifying risk. The essential insurance model involves pooling risk from individual payers and redistributing it across a bigger portfolio. Most insurance companies generate revenue in two ways: Charging premiums in exchange for insurance coverage, then reinvesting those premiums into other interest-generating assets.

The most common types of personal insurance policies are auto, health, homeowners, and life. Most individuals in the United States have at least one of these types of insurance, and car insurance is required by law. Auto insurance is an important protection for not just policyholder's car, but their financial liability as well. If you get into an accident without insurance, you could potentially be stuck paying for hundreds of thousands of dollars in damages and injuries.

For an Insurance company that has provided Health Insurance to its customers, when the company plan to launch new auto insurance, finding the most valued customers and sell products to them is important for the company. These health insurance policyholders are the most direct ones that the company can get in touch with, so if we can summarize the characteristics of target customers, it will save time to sell new auto insurance products. On the other hand, based on these policyholders have already buy health insurance, this company is reliable for them, so they may be more interested in enrolling in this auto insurance.

Therefore, building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue.

Section 2. Characteristic

2.1 Introduction of Original Data

In order to construct a model to predict whether a policyholder who previously owned the health insurance would be interested in a new vehicle insurance, we have the information about demographics, vehicle and policy from Kaggle which includes 12 variables. The original train data sample size is over 300,000 which is sufficient and close to the real situation, the proportion of people interested in the new insurance is much lower than the people who are not interested in, numerically, the percentage of interested is approximately 12.3 and the percentage of not interested is 87.7. We reselect sample from it to form a new train dataset for modeling, however, the test dataset doesn't require to fix the imbalanced problem, thus we sample a subset by randomly and get response = 0 and 1 are respectively 66879 and 9342. We both have predictors that are continuous and categorical: Continuous variables as customer age, region code, annual premium and vintage, and categorical variables includes gender, whether has a driving license, previously insured or not, vehicle age, vehicle damage, and channel of outreaching to the customer. The response variable for our topic is defined as a dummy variable which would be equal to 1 if the customer is interested, and 0 otherwise.

2.2 Analysis of Original Data

- ID -- Unique ID for the customer;
- Gender -- Gender of the customer;

- Age -- Age of the customer;
- Driving License -- Whether the customer has a driving license, we let 0 as the customer do not have a driving license, and 1 as the customer has a driving license;
- Region Code -- Unique code for the region of the customer;
- Previously Insured -- Whether the customer insured or not before, we let 0 as the customer do not have vehicle insurance before, and 1 as the customer has vehicle insurance before;
- Vehicle Age -- The age of the vehicle;
- Vehicle Damage -- Whether the customer damaged their vehicle before, we let 0 as the customer did not damage their vehicles before, and 1 as the customer damaged their vehicles before;
- Annual Premium -- The amount customers pay as premium in the year;
- Policy Sales Channel -- Anonymized code for the channel of outreaching to the customer ie. Different agents, over mail, over phone, in person, etc;
- Vintage -- Number of days the number has been associated with the company;
- Response -- Whether the customer is interested in the insurance, we let 0 as the customer is not interested in the insurance, and 1 as the customer interested in the insurance. In the following model, we use abbreviation of variables for convenience.

In the following, we will adopt the abbreviation notations for variables in all subsequent analyses. According to above variables, the order is: ID, Gender, Age, Drive, Code, Insured, Vage, Damage, Premium, Channel, Vintage, y.

Because the distribution of Premium is discrete and its values are much larger than other variables, Log(Premium) is used for analysis instead of Premium for the model.

Continuous variables summary:

```
> summary(data.frame(data$Age,data$logPremium,data$Vintage))
    data.Age     data.logPremium   data.Vintage
 Min.   :20.00   Min.   : 7.875   Min.   : 10.0
 1st Qu.:25.00   1st Qu.:10.103   1st Qu.: 82.0
 Median :36.00   Median :10.363   Median :154.0
 Mean   :38.82   Mean   :10.015   Mean   :154.3
 3rd Qu.:49.00   3rd Qu.:10.582   3rd Qu.:227.0
 Max.   :85.00   Max.   :13.200   Max.   :299.0
```

Figure 2.1: Continuous Variable Summary

We selected three continuous variables to discuss, i.e. Age, Log Premium and Vintage, result is shown in Figure 2.1. Age has a range from 20 to 85 with mean 38.82. From the analysis of the response variable, we observe Log Premium ranges from 7.875 to 13.2 with mean 10.015. For Vintage, the range is from 10 to 229 with mean 154.3.

2.3 Correlation

Through our correlation plot, which is shown in Figure 2.2, we discover some variables including Age, Insured, Vehicle Age, Damage, and Channel have relatively strong correlation with response variable.

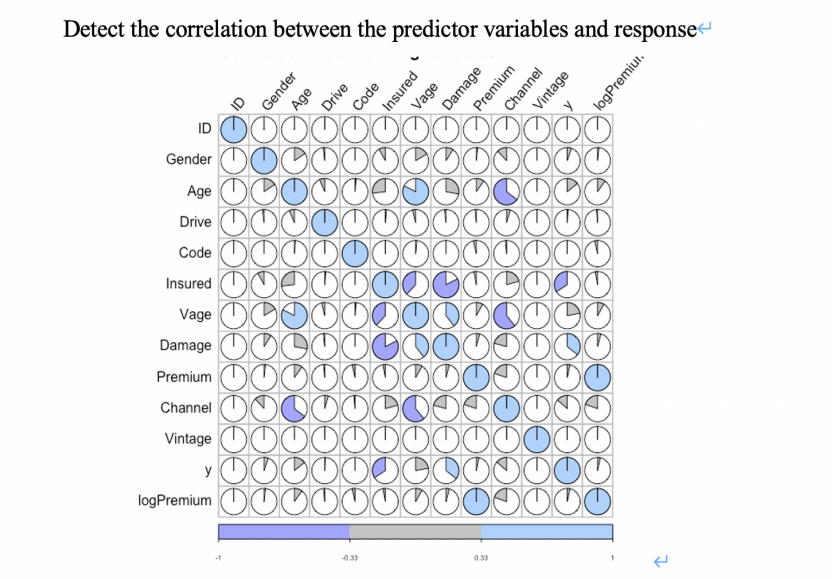

Figure 2.2: Correlation Plot

Basing on the legend on the x-axis, purple represents negative correlation between variables, while blue represents positive correlation between variables. Locked the position at Age column and Vage row, we have blue area over 3/4, which means Age has a high positive correlation with the Vage. Similar for Insured column and Damage row, we consider a high negative correlation between two variables. Also, Insured has some negative correlation with Vage. These variables may be considered as collinearity problems, further analysis about the variable selection in Logistics regression will be discussed in Section 3.

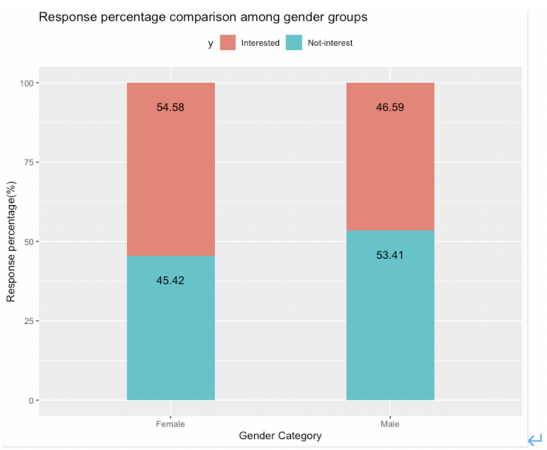

Figure 2.3 (a): Response for Gender

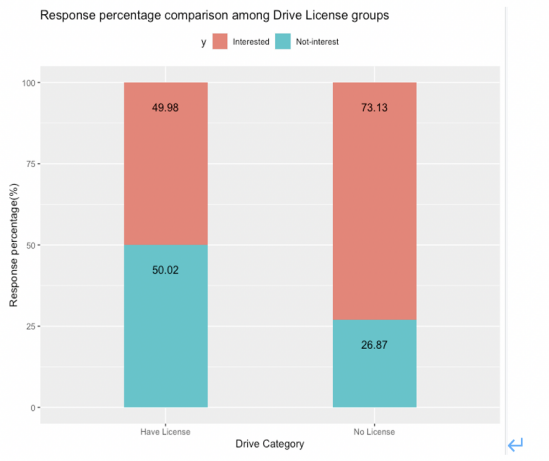

Figure 2.3 (b): Response for Drive License

From the Figure 2.3 above, we find the distributions of interests in insurance among gender groups are similar, however, when detecting the relationship of Drive License and response variable, it shows significant differences among groups. In that way, we should focus on the effect this variable leads to response in the following analyses.

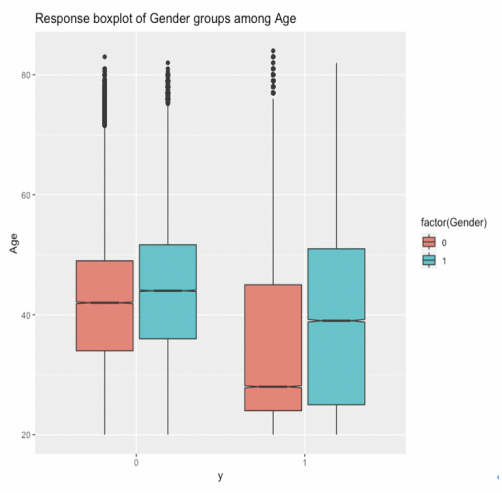

Figure 2.4 (a): Gender Boxplot (Age)

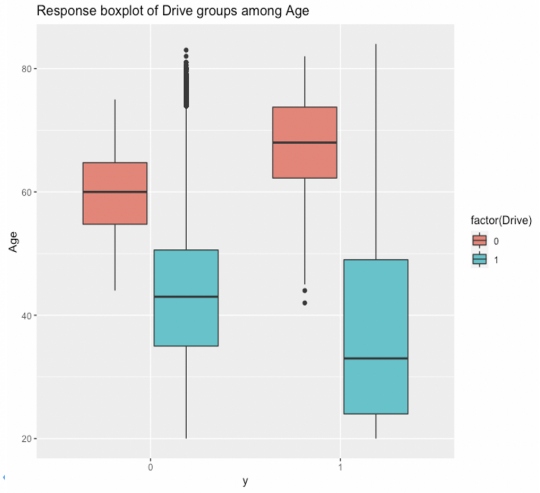

Figure 2.4 (b): Drive Boxplot (Age)

For the logistic regression model, in addition to analyses of categorical and continuous variables, the intersection term between them needs to be detected as well. In the boxplot of gender groups among age, shown in Figure 2.4 (a), it should be noticed that there is no significant intersection between gender and age because of the similar distribution of response. On the contrary, in the drive and age intersection boxplot, shown in 2.4 (b), we noticed that when a fixed category of y, drive equals to 1 has greater spread than the gender equals to 0, which means there is a

significant intersection between Drive and Age. Damage and Log-Premium have the intersection relationship as similar to drive and age, which is shown in the Appendix.

2.4 Imbalance Problem

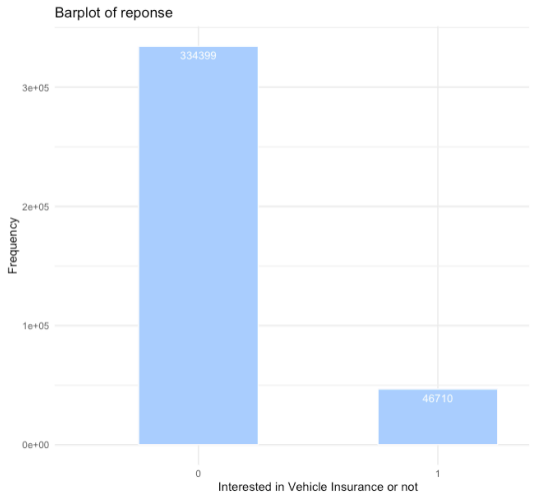

Figure 2.5: Bar plot of response

By the Figure 2.5, it is obvious that the responses are not represented equally. For our model, we plan to apply the logistic regression model for prediction which prefers balanced data for utilized prediction capability, hence using SMOTE algorithm to solve the imbalance problem is necessary which will be explained in the Model Selection section.

Section 3. Model Selection and Interpretation

3.1 SMOTE Algorithm for Imbalanced Problem

The preceding data characteristics section established that the numbers of responses ($y = 1\ or\ 0$, i.e. whether a customer is interested in the vehicle insurance) in two categories are significantly different, which leads us to solving it by SMOTE algorithm.

3.1.1 Algorithm Theory

SMOTE (synthetic minority oversampling technique) is used to solving the imbalance problem by oversampling. It is an improved method basing on random sampling, which adopts a strategy of simply copying samples to increase minority samples, is not prone to the problem of model overfitting. Thus, according to analyses of the minority class, SMOTE artificially synthesize new samples more scientifically.

3.1.2 Procedures

- Step 1: Set the minority class set A, for each $x \in A$, find the k-nearest neighbors of x by calculating the Euclidean distance between x and every other sample in set A;
- Step 2: the sampling rate N is set according to the imbalanced proportion. For every x in the minority sample, select $x_1 \dots x_n$ through its' k-nearest neighbors as a small set A_1;
- Step 3: For every randomly selected $x_i \in A_1 (\forall i = 1, \dots n)$, generate a new example by the formula: $x' = x + rand(0,1) * |x - x_i|$.

3.1.3 Application

We use DMwR::SMOTE function in R by setting parameters as Figure 3.1 below, i.e. the average number of generating minority class sample is 1, we use 5-nearest neighbors when using over-sampling, and for majority class, we choose 2N/100 samples corresponding to the minority, where N represents the original sample size in minority class.

```
train<-SMOTE(y~.,trainsplit,perc.over=100,perc.under=200)
```

Figure 3.1: Setting of SMOTE Function

After applying the algorithm, we have the samples in train groups with response = 0 and response = 1 are 74703 and 74736 respectively, which can be considered as a balanced dataset.

3.2 Logistics Regression Model

3.2.1 Modeling

In order to detect the factors affecting the response of whether a customer will purchase vehicle insurance, we should introduce the Logistics Regression Model. Before applying it, we should detect the correlation between variables as well as determine the adequate intersection terms.

Since the correlation plot in Figure 2.2 shows some high-correlated variables, Vage and Age, Damage and Insured, we only keep Age and Damage predictors. As for the coefficients between continuous variables (Age, Log-Annual Premium and Vintage) are small, we consider no multicollinearity problems here. After detecting the Boxplots of categorical variables (Gender, Driving, Insured and Vehicle-age) by groups among three continuous variables (see details in Figure 2.4 and Appendix 5.1), we keep two intersection variables: Drive*Age, Damage*LogPremium. After performing the Type III error test on all the variables, we keep significant variables to get a reduced model without Vintage term, the detailed R codes are in Appendix 5.3 and 5.4.

Thus, we find the "best" Logistic regression model for estimate the response, the model can be written as the formula

$$y = -6.875 - 0.135 * Gender + 6.449 * Drive + 0.522 * Damage + 0.120 * Age$$

$$+0.403 * logPremium - 0.126 * Drive_{Age} - 0.463 * Damage_{logPremium}$$

Notes: Gender, Drive Damage corresponding to the dummy variables when the values equal 1 represent the category 1;

$Drive_{Age}, Damage_{logPremium}$ corresponding to the intersection terms we introduce to the model fitting.

3.2.2 Interpretation

We interpretate these significant coefficients below,

(Note: every time we discuss a single coefficient, we assume that all the other variables are fixed)

β_{Gender}: the odds ratio of a male customer's interest in insurance over a female customer's interest is $exp(-0.135) = 0.874$.

β_{Drive}: the odds ratio of insurance interest of a customer with drive license over one without license is $exp(6.449) = 632.070$.

β_{Damage}: the odds ratio of insurance interest of a customer with car got damaged before over one without damaged is $\exp(0.522) = 1.685$

β_{age}: the odds of a customer's interest in insurance increase multiplicatively by $exp(0.120) = 1.127$ for every unit increase in age.

$\beta_{logPremium}$: the odds of a customer's interest in insurance increase multiplicatively by $exp(0.403) = 1.496$ for every unit increase in Log-Premium.

$\beta_{Drive_{Age}}$: the odds ratio of insurance interest of a customer with drive license multiplicatively by $exp(-0.126) = 0.882$ for every unit increase in age.

$\beta_{Damage_{logPremium}}$: the odds ratio of insurance interest of a customer once had car damage multiplicatively by $exp(-0.463) = 0.629$ for every unit increase in Log-Premium.

Therefore, according to this logistics regression model, in general we find that:

- For categorical predictors: The customer with a driver's license is highly interested in vehicle insurance, the customer who got the car damaged is highly interested in vehicle insurance.

This conclusion makes sense, since car owners are more concerned about buying car insurance, and those who have experienced car repairs are also more concerned about having good insurance for the future. Also, we get the result that female customers having more interest in this insurance than males. Since car damage is negative correlated with a customer's current stage of having insured, we should focus on investigating people who do not have insured in any vehicle insurance.

- For continuous predictors: The influence of a one-unit increase in LogPremium is larger than Age increase. The coefficient for Vintage is not significant that we can pay less attention to this factor for future study. The more the customer needs to pay as a premium for health insurance, the more interest they have for the new vehicle insurance. It is reasonable, since the more one customer pays for his or her health insurance, the more he or she trusts the insurance company, and the more likely he or she is to try this new auto insurance.

3.2.3 Prediction

We apply the "best" logistics model towards the test samples and get the ROC Curve below, shown in Figure 3.2. The area under the Curve of ROC = 0.785. We should find better model for prediction.

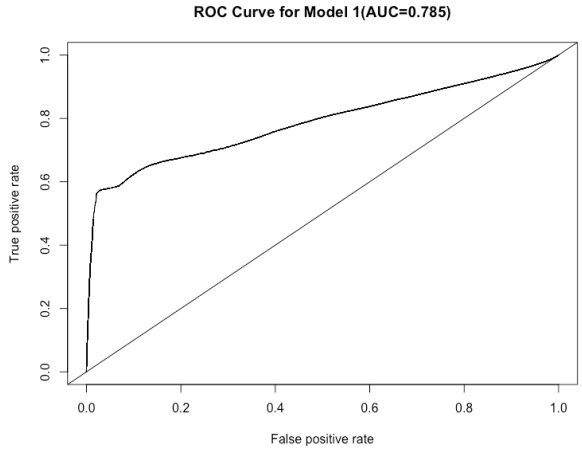

Figure 3.2: Roc Curve of Logistic Model

3.3 Random Forest

3.3.1 Modeling

Random forest is an ensemble learning method of classification by generating bootstrap samples from the training data (Breiman, 2001). However, the generalized Random forest method suffer from the curse of learning from extremely imbalanced training data set. Basing on the methods

discussed for random forest with imbalanced data (Chao Chen et al., 2004), we introduce balanced random forest method (BRF).

Balanced random forest algorithm is used to ensemble trees induced from balanced down-sampled data without data loss:

- Step 1: A large number of bootstrap samples are taken from the training data and create separate unpruned tree for each set;
- Step 2: Randomly samples a subset of predictors at each split to encourage diversity of the resulting trees.
- Step 3: Combining all the prediction results in every tree to generate a single prediction for an individual sample.

Under the down-sampling method for BRF, our random forest can take random samples of size $c * nmin$, where c is the number of categories of response, $nmin$ is the number of samples in the minority class. In our example, we have $c = 2, nmin = 37368$.

3.3.2 Prediction

When using this down-sampled data set to the cross-validation procedure of random forest, we have a better predicted performance with AUC = 0.8511. This accuracy is much better than the logistics regression model. The ROC Curve of random forest plot is shown as Figure 3.3.

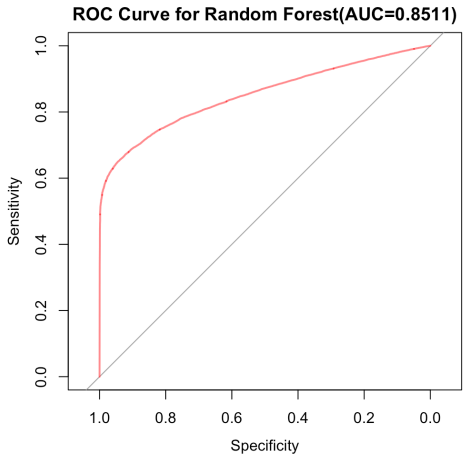

Figure 3.3: ROC Curve of Random Forest

3.4 Cluster Analysis

3.4.1 Modeling

Here, we apply one of the most widely used centroid-based clustering algorithm method: K-Means clustering, which can organize the data into non-hierarchical clusters. K-means algorithm is basing on the distances between samples to determine the clusters. Our goal is to minimum the square errors.

When suppose we separate data as clusters $(C_1, C_2, \dots, C_k)$

$$\min E = \sum_{i=1}^{k} \sum_{x \in C_i} \left|\left|x - \mu_i\right|\right|_2^2$$

Where $\mu_i \; is \; the \; mean \; of \; cluster \; C_i, \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x.$

However, this method is sensitive with outliers and imbalanced data problem, thus, we perform resampling method again to take a random sample from the data with size 6000 to perform the algorithm.

3.4.2 Analysis

Although data clustering is of the unsupervised learning methods, here, we focus on detecting the relationship between people's desire of insurance with their potential characteristics. In that way, we don't pay much attention.to the results of classification for each sample under data clustering comparing with their response of interests.

Since the original data is so large, we randomly select 3000 samples each from response = 1 and response = 0 and summary the general characteristics of the potential customers. Below is only one of the results from our repeated randomized trials.

After determining the optimal number of clusters is 3, we can cluster all the samples into three groups. At first, we can determine the relationship between categories and the aspiration of customers' interests in vehicle insurance by the following Table 3.1.

Table 3.1: The Relationship between Clusters and Response

Clusters \ Response	Y = 1	Y = 0
Cluster 1	1059	0
Cluster 2	1110	0
Cluster 3	831	3000

It's clear that Cluster 3 contains all the samples with response = 0, so we only detect the characteristics that cluster 1 and 2 have and consider these characteristics of our potential customers who are interested in the vehicle insurance.

To visualize the clustering process, we show three clusters below. From the clustering plot, Figure 3.4, cluster 2 and cluster 3 are significantly different, cluster 1 is the potential customer group that has both common characteristics with cluster 1 and cluster 3.

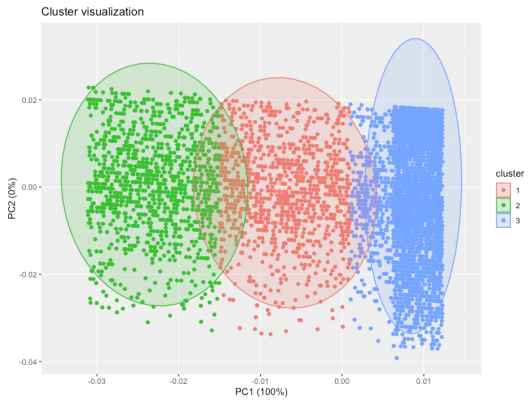

Figure 3.4: Clustering Plot

After combining the Cluster 1 and 2 as one cluster with Cluster 3 as one cluster, we detect the differences between these two groups.

From Figure 3.5, our potential customers are corresponding to ones are most among age of 40-60 years old. For people who are age less than 30, they may be less possibilities to be the most "valuable" customers that we should focus on.

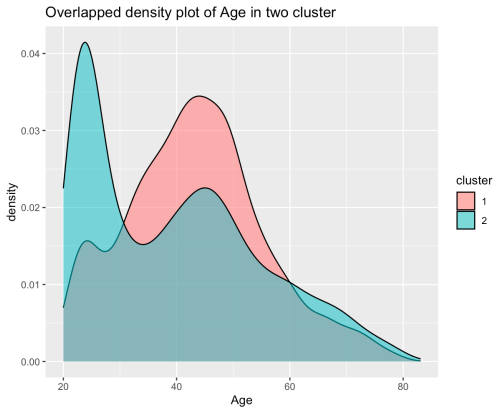

Figure 3.5: Overlapped Density Plot

We calculate the relative frequencies of each category for four variables comparison in the following Table 3.2, there are some distinct features with our potential customers: the vehicle age is 1 - 2 years; the car got damaged in the past; the customers do not have the vehicle insurance.

Table 3.2: Relative Frequency for Variable Comparison

Predictor Cluster	Category	Cluster 1	Cluster 2
Gender	0	0.389	0.454
	1	0.611	0.546
Vage	1	0.162	0.392
	2	0.733	0.562
	3	0.105	0.046
Damage	0	0.018	0.454
	1	0.982	0.546
Insured	0	0.997	0.592
	1	0.003	0.408

Section 4. Summary and Conclusion

In this article, we aim to solve a practical problem: predict whether our existing health-insurance customers will be interested in the new vehicle insurance based on their information provided about themselves and their vehicle conditions. Our first problem is dealing with extremely imbalanced data. After comparing two methods: (1) using a random sample subset to replace the whole data; (2) using SMOTE algorithm to synthesis minority group data, we have a method (1) is easy to get overfitting problem with information loss in the further prediction process. Thus, we keep the SMOTE algorithm and get a training group with response = 1/0 74703 and 74736 respectively. This is a large sample for further prediction.

After applying Exploratory Data Analysis (EDA), we detect the correlation between predictors, some multicollinearity problems among predictors as well as determine the intersection terms in the model. The “best” fitted model is with predictors Gender, Drive, Damage, Age, LogPremium, Drive*Age, Damage*LogPremium. The predicted accuracy is diagnosed by ROC curve, we have AUC = 0.785 which is not a very good prediction accuracy. Balanced Random Forest algorithm helps to improve the accuracy with AUC = 0.8511 since the method is used to ensemble trees

induced from balanced down-sampled data without data loss.

The data clustering method is used for additional interpretation and summary of our potential customers characteristics. Some of the features in clusters that customers who are interested in the insurance are similar as the interpretation of logistics regression model. Here, we summarize the features they have and make suggestions towards the company.

1. When consider the conditions of the cars, our potential customers are those with drive license, and got car damaged. What's more, investigating with people who do not have insured in any vehicle insurance will be more effective.

2. When consider the basic information of customers, we have female customers show higher interest than males, so if our marketing is based on a family unit, marketing to women may be more effective. It's shown that customers age within 40-60 are more likely to interested in the insurance than the age within 20-30. These properties are useful when we contact with target customers.

3. For our clients who have enrolled in other kinds insurances for a long period, we should focus on the ones who pay high premium every year. Since these clients have long period enrolling in the company, they have confidence with the company and are more willing to try new insurance projects.

Section 6. References

[1]. Data source: https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv

[2]. Breiman, L. (2001). Random forest. Machine Learning, 45, 5–32.

[3]. Chen, Chao & Breiman, Leo. (2004). Using Random Forest to Learn Imbalanced Data. University of California, Berkeley.

Section 7. Appendix

5.1 Boxplot

From the boxplot, it can be notified that there is no significant intersection between gender and vintage because of the similar distribution of response.

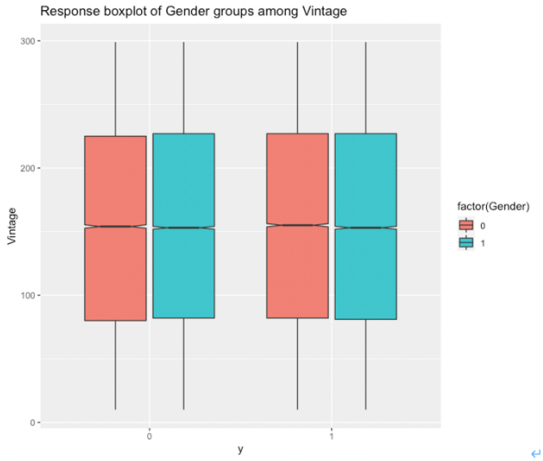

Figure 5.1.1: Boxplot of Gender groups and Vintage

Below, we posted all the other intersection detection plots from Figure 5.1.2 to Figure 5.1.12, the interpretations are the same. Only the boxplot of Vehicle age and Age shows some significant difference between three categories, so we consider intersections between two variables. However, we may consider collinearity between these two variables, and delete one of them in the Logistic regression analysis.

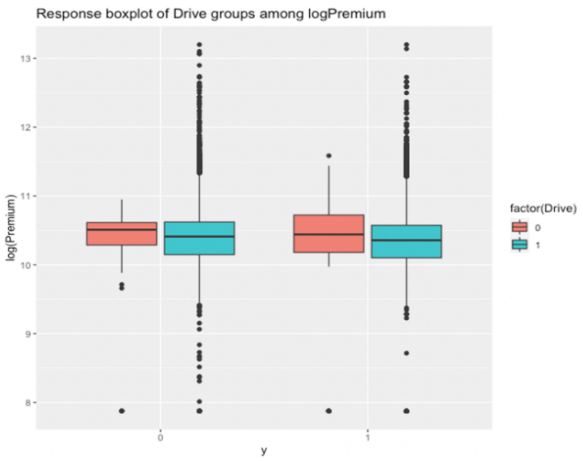

Figure 5.1.2: Boxplot of Drive and Log Premium

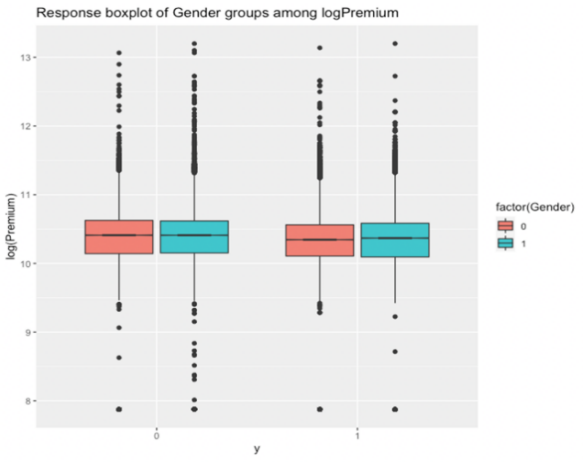

Figure 5.1.3: Boxplot of Gender and Log Premium

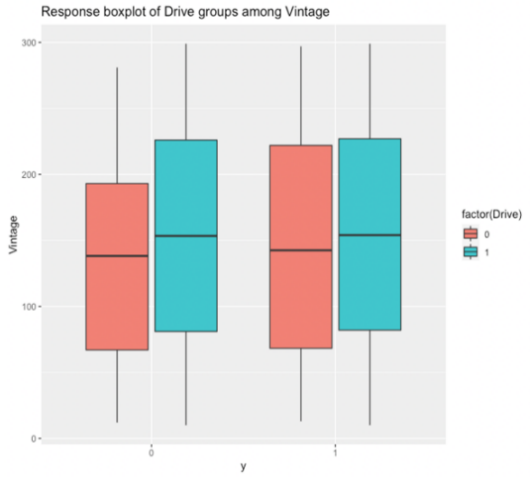

Figure 5.1.4: Boxplot of Drive and Vintage

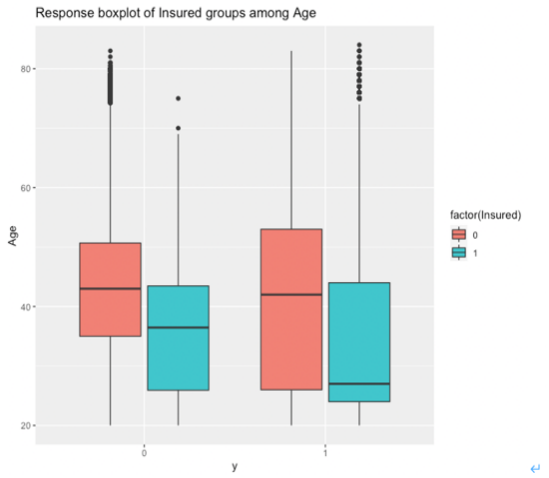

Figure 5.1.5: Boxplot of Insured groups and age

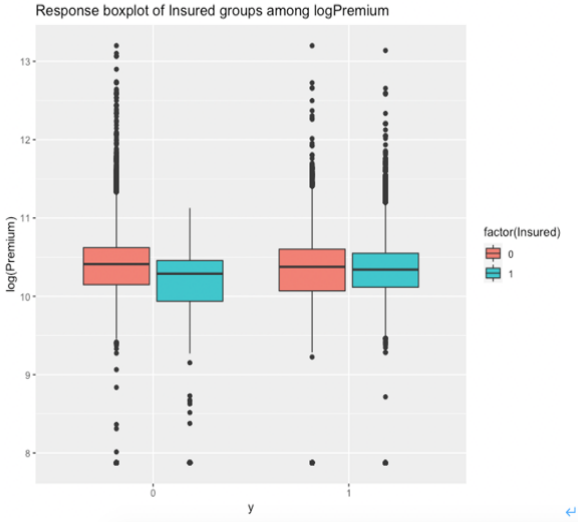

Figure 5.1.6: Boxplot of Insured and Log Premium

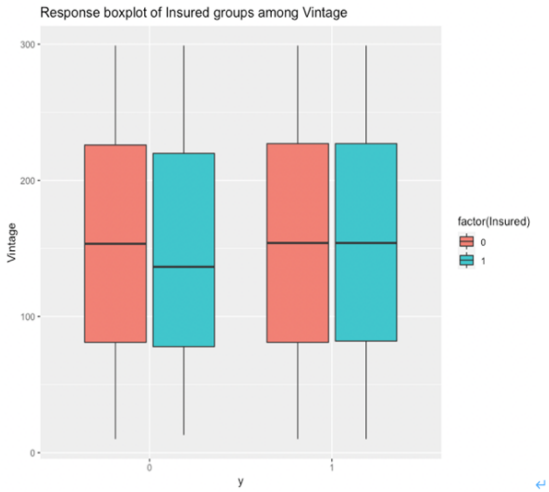

Figure 5.1.7: Boxplot of Insured and Vintage

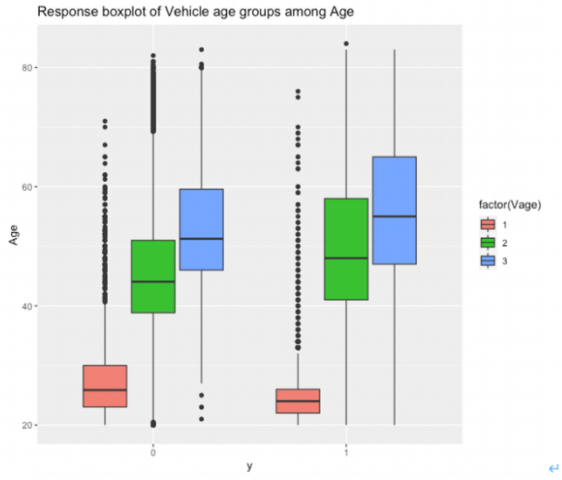

Figure 5.1.8: Boxplot of Vehicle age and Age

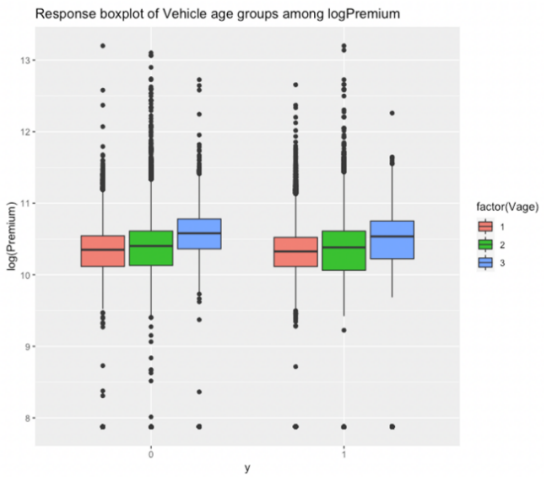

Figure 5.1.9: Boxplot of Vehicle age and Log Premium

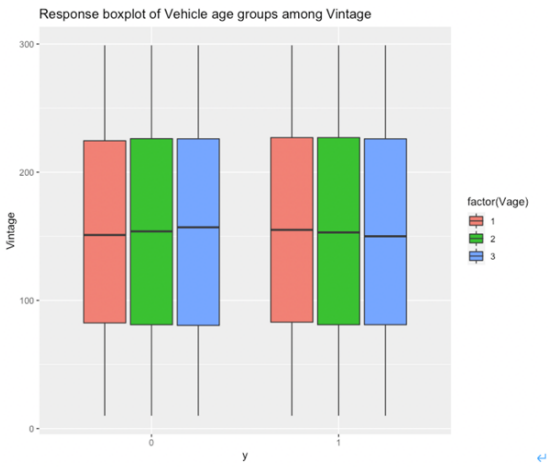

Figure 5.1.10: Boxplot of Vehicle age and Vintage

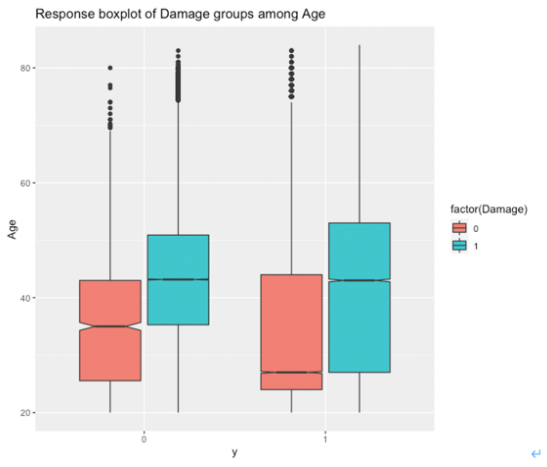

Figure 5.1.11: Boxplot of Damage and Age

From the boxplot, it can be noticed that when fixed category of y, damage equals to 0 has greater spread than the gender equals to 1, which means there is a significant intersection between damage and age.

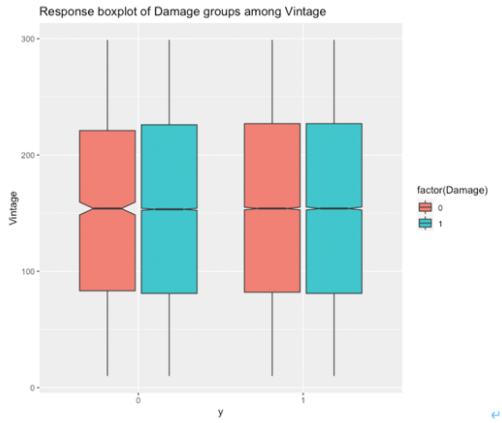

Figure 5.1.12: Boxplot of Damage and Vintage

From the boxplot, it can be notified that there is no significant intersection between damage and vintage because of the similar distribution of response.

5.2 Age Histogram

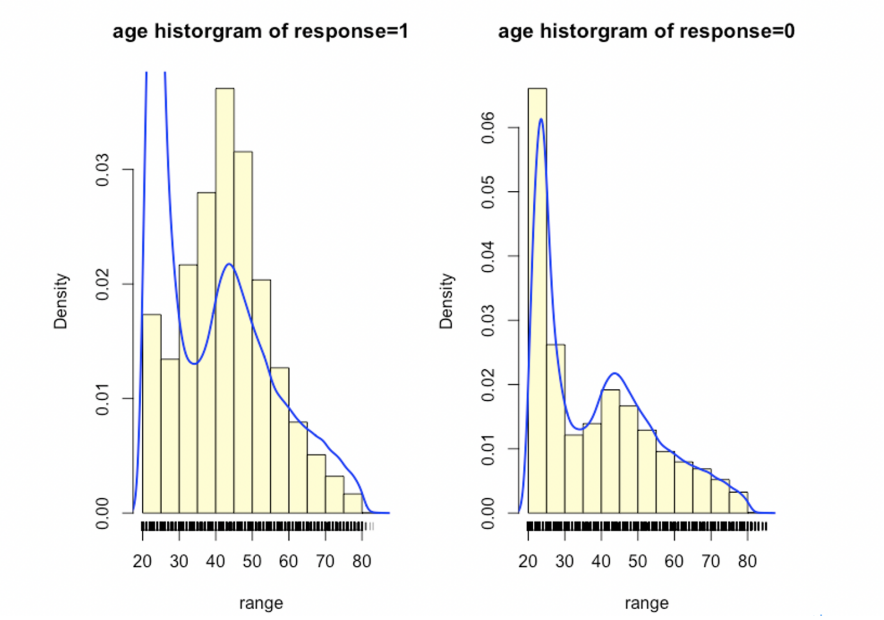

Figure 5.2.1: Histogram of Age

5.3 The Full Logistic Regression Model

```
Call:
glm(formula = y ~ Gender + Drive + Damage + Age + logPremium +
    Vintage + Drive * Age + Damage * logPremium, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.84381  -0.85654   0.07443   0.37959   2.08486

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.880e+00  1.474e+00  -4.668 3.05e-06 ***
Gender1            -1.353e-01  1.316e-02 -10.278  < 2e-16 ***
Drive1              6.447e+00  1.459e+00   4.418 9.94e-06 ***
Damage1             5.222e-01  2.206e-01   2.367   0.0179 *
Age                 1.203e-01  2.313e-02   5.201 1.99e-07 ***
logPremium          4.033e-01  2.183e-02  18.469  < 2e-16 ***
Vintage             4.508e-05  7.722e-05   0.584   0.5594
Drive1:Age         -1.255e-01  2.314e-02  -5.425 5.80e-08 ***
Damage1:logPremium -4.630e-01  2.273e-02 -20.373  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207166  on 149438  degrees of freedom
Residual deviance: 143709  on 149430  degrees of freedom
AIC: 143727

Number of Fisher Scoring iterations: 6
```

Figure 5.3.1: Fitted Full Model Result

After detecting the nonsignificant predictors, we should delete the predictor Vintage. Then, we can apply hypothesis test for nested model and get the “best” model as below.

5.4 The Reduced Logistic Regression Model

```
Call:
glm(formula = y ~ Gender + Drive + Damage + Age + logPremium +
    Drive * Age + Damage * logPremium, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8425  -0.8565   0.0746   0.3798   2.0850

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.87531    1.47378  -4.665 3.08e-06 ***
Gender1           -0.13530    0.01316 -10.277  < 2e-16 ***
Drive1             6.44902    1.45893   4.420 9.85e-06 ***
Damage1            0.52237    0.22063   2.368   0.0179 *
Age                0.12032    0.02313   5.202 1.97e-07 ***
logPremium         0.40329    0.02183  18.470  < 2e-16 ***
Drive1:Age        -0.12554    0.02313  -5.426 5.75e-08 ***
Damage1:logPremium -0.46302    0.02273 -20.374  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207166  on 149438  degrees of freedom
Residual deviance: 143709  on 149431  degrees of freedom
AIC: 143725

Number of Fisher Scoring iterations: 6
```

Figure 5.4.1: Fitted Reduced Model

From the results in Deviance test, we assume the hypothesis

$$H_0: \beta_6 = 0 \ \ vs. \ \ H_1: at\ least\ one \ \ \beta_8 \neq 0$$

$where\ \beta_6\ corresponding\ to\ coefficient\ of\ Vintage$

$G^2 = D_0 - D_1 = 0.341 < 3.841, where\ \chi^2_{0.975,1} = 3.841, we\ cannot\ reject\ H_0.$

Thus, we keep this reduced model as the “best” Logistic regression model for estimate the response.

5.5 Data Clustering with determining cluster number

Here, we can determine the optimal numbers of clusters by Elbow method, which is to define clusters such that the total intra-cluster variation (or total within-cluster sum of square (WSS)) is minimized. From the Figure 5.4 below, we should determine k by comparing the slope of decreasing of the total within sum of square. The adequate k is one that the slope becoming flatter than the previous ones. Thus, k = 3 is adequate.

```
library(factoextra)
#scale the data first
df<-scale(newclus)
fviz_nbclust(df,FUNcluster=kmeans,method="wss")+geom_vline(xintercept=3,linetype=2)
```

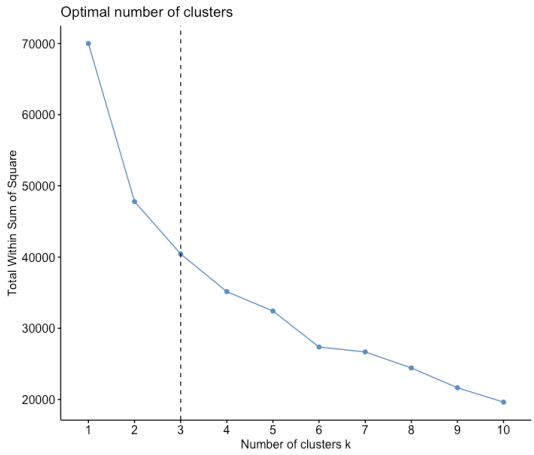

Figure 5.5.1: Total Within Sum of Square

5.6 Summarize characteristics of Cluster 1 and 2

For comparing the differences of categorical variables in two clusters, we calculate the frequencies in each group and for each of the variable.

```
> G1<-table(d1$Gender)
> V1<-table(d1$Vage)
> D1<-table(d1$Damage)
> I1<-table(d1$Insured)
> prop.table(G1)

        0         1
0.3891194 0.6108806
> prop.table(V1)

        1         2         3
0.1622868 0.7325957 0.1051176
> prop.table(D1)

          0          1
0.01751959 0.98248041
> prop.table(I1)

           0           1
0.997233748 0.002766252
> G2<-table(d2$Gender)
> V2<-table(d2$Vage)
> D2<-table(d2$Damage)
> I2<-table(d2$Insured)
> prop.table(G2)

        0         1
0.4539285 0.5460715
> prop.table(V2)

         1          2          3
0.39232576 0.56173323 0.04594101
> prop.table(D2)

        0         1
0.4536674 0.5463326
> prop.table(I2)

        0         1
0.5922736 0.4077264
```

Figure 5.6.1: Calculated Frequency