2020

Health Insurance Cross Sell Prediction

FINAL PROJECT – INTRODUCTION TO DATA SCIENTIST

HAIYEN KAMEL

Table of Contents

1 Section 1: Introduction

Data has always played a critical role in the insurance industry and today, the insurance companies have access to more data than ever before. They are taking advantage of recent advances in AI and Machine Learning to solve business challenges including underwriting, product pricing, fraud detection, sales and customer experiences. One of the challenges we are facing working with data in Insurance industry is imbalanced data. This is a situation where the number of observations belonging to one class is significantly higher than those on other classes. Training a machine learning model on an imbalanced dataset can affect the accuracy of the model as well as creating learning problem.

1.1 Objective:

The purpose of this project is to build a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the Insurance company. The goal of this study is also to review the different methods that can be used to tackle classification problems with imbalanced dataset.

1.2 Data Source

Data were provided by Kaggle consists of 2 datasets including train set and test set. There are 381,109 observations and 12 columns in train set and 127,037 observations in test set with only 11 columns, missing Response column. After some investigations, we found that this is an imbalance dataset with approximately 12.3% of the response that customers were interested in Vehicle Insurance while 87.7% of the response that customers were not interested in buying Vehicle Insurance.

1.3 Data Description:

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer

Variable	Definition
Driving_License	0: Customer does not have DL 1: Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1: Customer already has Vehicle Insurance 0: Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1: Customer got his/her vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to 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, Customer has been associated with the company
Response	1: Customer is interested 0: Customer is not interested

2 Section 2: Data Transformation &Visualization

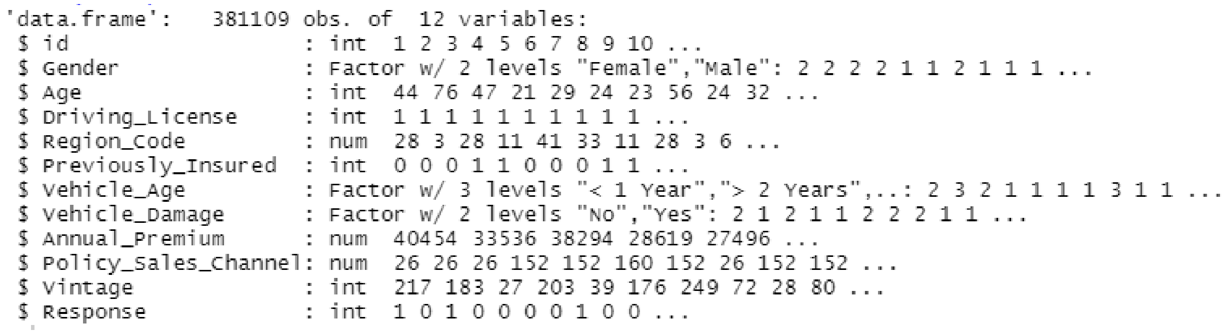
2.1 Data Explore & Pre-processing

For this project, train set will be used to build the model and we will use the predictive model to predict the missing Response in test set.

By taking a quick glance at the train set, “Gender”, “Vehicle Age” and “Vehicle Damage” are categorized as factor while our target variable “Response” is categorized as int. The train set consists of different type of data (integer, factor, numerical). In this case, we can work on converting integer to numerical values.

The Response column is the target variable and it helps to answer whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided or not. The goal of this project is to predict the Customer response (either 1 if a customer is interested or 0 if not) based on some features such as “Age”, “Driving_License”, “Previously_Insured”, etc. In this case, we will use both categorical and continuous variables.

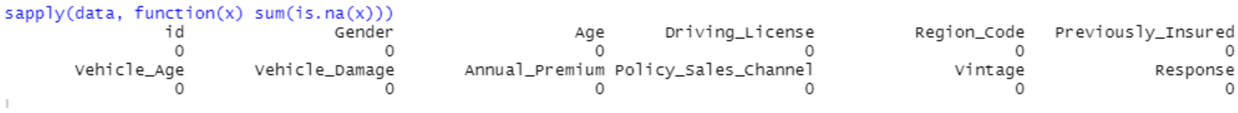
Let’s check the structure of the training dataset.



```
'data.frame':	381109 obs. of  12 variables:
 $ id                  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender              : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 2 1 1 1 ...
 $ Age                 : int  44 76 47 21 29 24 23 56 24 32 ...
 $ Driving_License     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Region_Code         : num  28 3 28 11 41 33 11 28 3 6 ...
 $ Previously_Insured  : int  0 0 0 1 1 0 0 0 1 1 ...
 $ Vehicle_Age         : Factor w/ 3 levels "< 1 Year","> 2 Years",..: 2 3 2 1 1 1 1 3 1 1 ...
 $ Vehicle_Damage      : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 2 2 2 1 1 ...
 $ Annual_Premium      : num  40454 33536 38294 28619 27496 ...
 $ Policy_Sales_Channel: num  26 26 26 152 152 160 152 26 152 152 ...
 $ Vintage             : int  217 183 27 203 39 176 249 72 28 80 ...
 $ Response            : int  1 0 1 0 0 0 0 1 0 0 ...
```

2.2 Exploratory Data Analysis

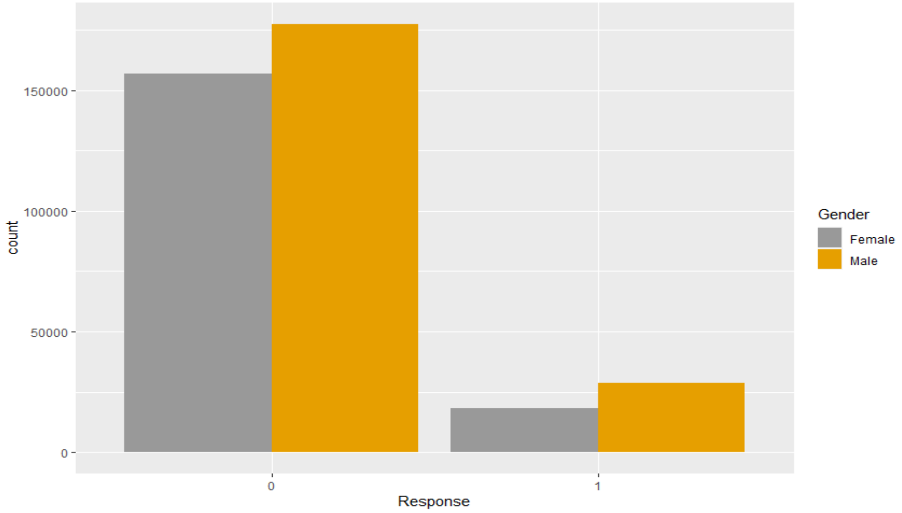
Using sapply() function can help us to identify missing values for each variables which were not found in the train set. All variable names appear to be in good format.



```
sapply(data, function(x) sum(is.na(x)))
                  id               Gender                  Age      Driving_License          Region_Code   Previously_Insured 
                   0                    0                    0                    0                    0                    0 
         Vehicle_Age       Vehicle_Damage       Annual_Premium Policy_Sales_Channel              Vintage             Response 
                   0                    0                    0                    0                    0                    0 
```

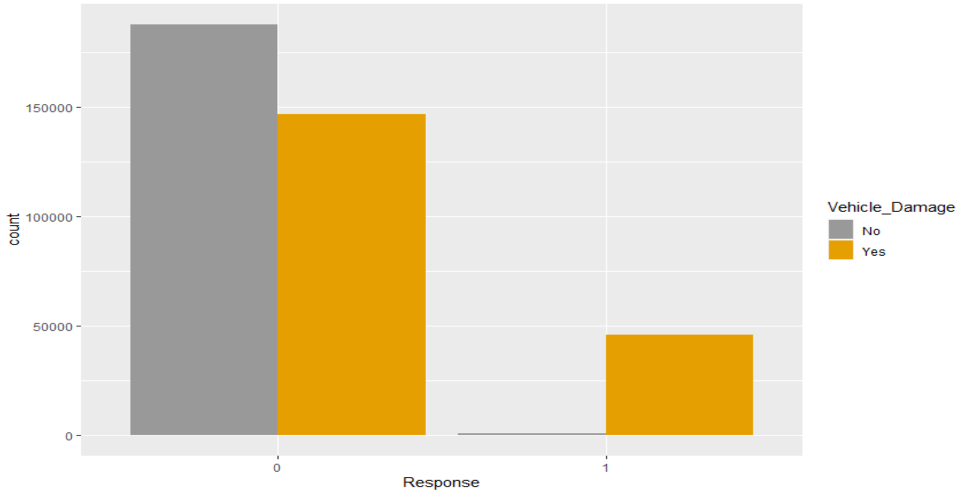
In order to have a better understanding of this dataset, we need to study the distribution of each variable. Figure 1 shows us that we have more Male policyholders than Female in both Response categories. The Chi-square test was used to evaluates whether there is a significant association between Response and Gender (χ^2 = 1047.7, df = 1, p-value < 0.00000000000000022). We can conclude that the two variables are in fact dependent since p-value is less than 0.05.

Figure 1 - Customer Response by Gender



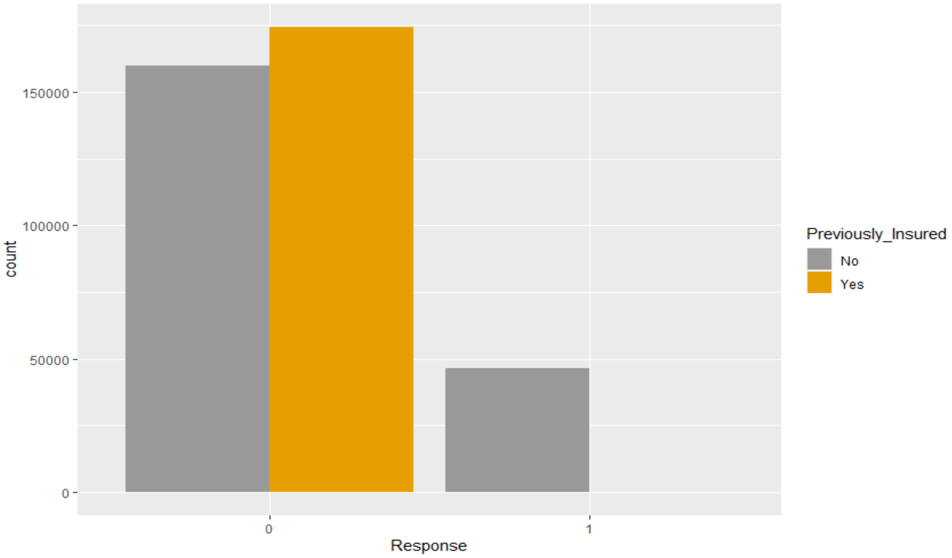
In the figure 2, the chart shows that customers who got his/her vehicle damaged in the past would be interested in buying additional Vehicle Insurance. We can see that there are more customers with no vehicle damaged in the past not interested in buying insurance than customers got vehicle damaged. However, we have more customers who got vehicle damaged are interested in getting Vehicle Insurance than customers who did not. The Chi-square test was used to evaluates whether there is a significant association between Response and Vehicle Damage = 47865, df = 1, p-value < 0.00000000000000022). In this case, since p-value is less than 0.05 so we believe Response & Vehicle Damage are dependent.

Figure 2 - Vehicle Damage Response



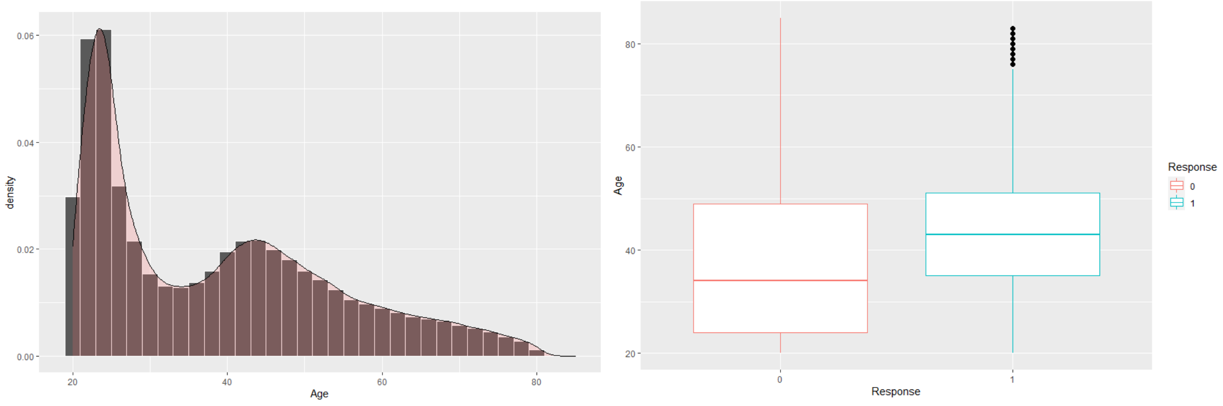
In contrast, we have more customers who were previously insured not interested in getting Vehicle Insurance than who were not and only customers who were int previously insured are interested in getting Vehicle Insurance. Based on both graphs above, the company should focus on customers who experienced vehicle damaged in the past along with uninsured customers. We also performed the same Chi-square test for variables “Previously Insured”, “Driving License” & “Vehicle Aged” and found that all p-values are less than significant level 0.5. We can conclude there are association between Response with all categorical variables.

Figure 3 - Previously Insured Customer Response

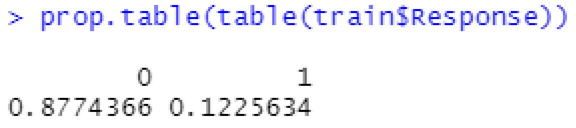


Most of the customers in this dataset are in the age range of 20-30 and the second highest age group is among ages of 40 – 50. Since the younger customers are the most active online, the company should focus their effort on digital marketing. On the other hand, the ages of both Response group 0 are more variable than other. The median Age for both groups seems to be different as well.

Figure 4 – Plot of Age by Response



We clearly can see that this dataset is greatly imbalanced with 87.7% of customer not interest and 12.3% of customer interested. Therefore, various approaches were implemented to deal with this problem. In this project, various methods were looking into such as Undersampling method, Oversampling Method, ROSE method.

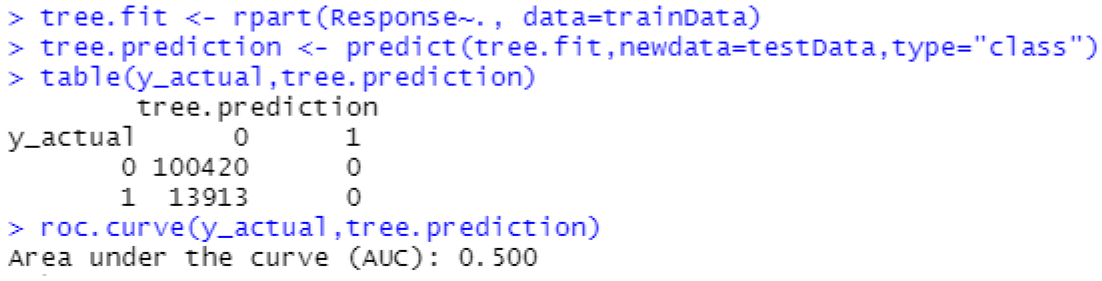


```
> prop.table(table(train$Response))

        0         1 
0.8774366 0.1225634 
```

3 Section 3: Machine Learning Algorithm

3.1 Train/Test Split & Imbalanced Classification

Before proceeding to the fitting process, we split the train set data into two chunks: 70% of observations will be included in training set & 30% of observations will be in testing set. The training set will be used to fit the predictive model which will be used testing over the testing set. As we see, this data set contains on 12.3% of Yes response and 87.7% of No response. This dataset is severely imbalanced. In order to address this problem, a decision tree algorithm is used for modeling purposes.



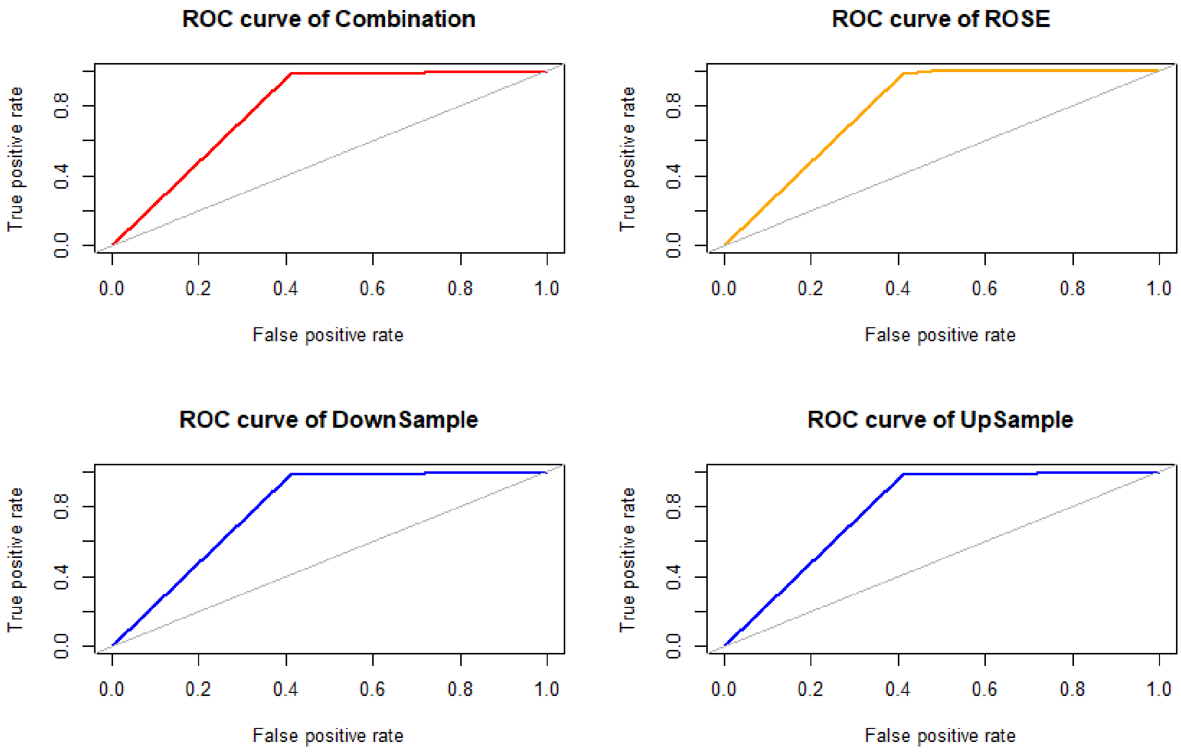
```
> tree.fit <- rpart(Response~., data=trainData)
> tree.prediction <- predict(tree.fit,newdata=testData,type="class")
> table(y_actual,tree.prediction)
        tree.prediction
y_actual      0      1
       0 100420      0
       1  13913      0
> roc.curve(y_actual,tree.prediction)
Area under the curve (AUC): 0.500
```

A decision tree model was built using all data in training set which is highly imbalanced. After fitting the model using decision tree, we check the final accuracy score using ROC AUC. It looks like the ACU = 0.500 which is extremely low score. Therefore, this model is not worth to use, and the data need to be balanced before applying a machine learning algorithm.

In this study, we used 3 methods to balance the data. Below are the methods used to treat imbalanced datasets:

1. UnderSampling: remove observations from the training dataset that belong to the majority class in order to better balance the class distribution. This method is also useful when we have a large dataset and it helps to reduce the number of training samples in order to improve run time and storage troubles. However, using this method may cause training dataset to lose important information pertaining to majority class.
2. OverSampling: this method is an opposite of UnderSampling where it works with smaller group. This method is used to randomly oversampling the minority class. However, this might lead to overfitting problem.
3. Combination of over- and under-sampling.
4. ROSE - Random Over-Sampling Examples.

Figure 5 - AUC ROC for Decision Trees

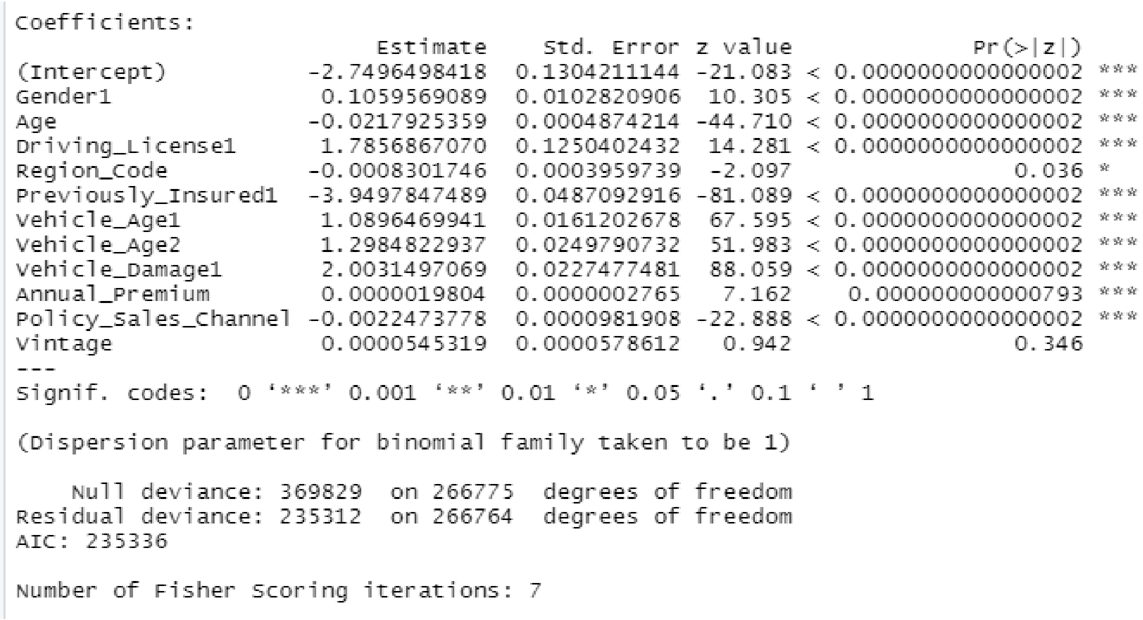


It shows that the data generated from ROSE, Combination, UpSampling, and DownSampling method provide better accuracy scores compared to Imbalanced data with AUC = 0.784 for DownSampling, UpSampling & Combination method. ROSE has the highest accuracy score of 0.789. Therefore, ROSE method dataset will be chosen to build Logistic Regression Model.

3.2 Logistic Regression Model

Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The predictors can be continuous, categorical or a mix of both. Continuous variables are Age, Annual Premium, Policy Sales Channel, and Vintage in this study, while other variables are categorical. Chi-square test was used to identify differences between groups. Since the target variable Response is binary, we will fit a binary logistic regression model using ROSE dataset. After fitting the model, we can interpret what the model is telling us based on Figure 6 table.

Figure 6 - Logistic Regression Model

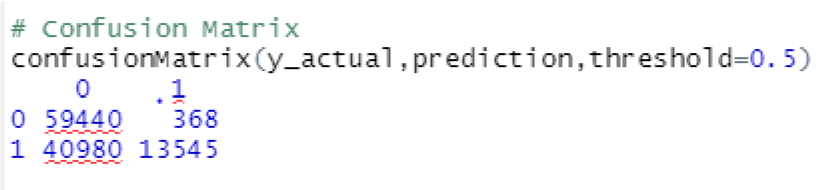


```
Coefficients:
                        Estimate    Std. Error z value               Pr(>|z|)    
(Intercept)         -2.7496498418  0.1304211144 -21.083 < 0.0000000000000002 ***
Gender1              0.1059569089  0.0102820906  10.305 < 0.0000000000000002 ***
Age                 -0.0217925359  0.0004874214 -44.710 < 0.0000000000000002 ***
Driving_License1     1.7856867070  0.1250402432  14.281 < 0.0000000000000002 ***
Region_Code         -0.0008301746  0.0003959739  -2.097                0.036 *  
Previously_Insured1 -3.9497847489  0.0487092916 -81.089 < 0.0000000000000002 ***
Vehicle_Age1         1.0896469941  0.0161202678  67.595 < 0.0000000000000002 ***
Vehicle_Age2         1.2984822937  0.0249790732  51.983 < 0.0000000000000002 ***
Vehicle_Damage1      2.0031497069  0.0227477481  88.059 < 0.0000000000000002 ***
Annual_Premium       0.0000019804  0.0000002765   7.162    0.000000000000793 ***
Policy_Sales_Channel -0.0022473778  0.0000981908 -22.888 < 0.0000000000000002 ***
Vintage              0.0000545319  0.0000578612   0.942                0.346    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 369829  on 266775  degrees of freedom
Residual deviance: 235312  on 266764  degrees of freedom
AIC: 235336

Number of Fisher Scoring iterations: 7
```

We can see that “Vintage” is not statistically significant because its p value greater than 0.05. All variables have extremely low p-values suggesting a strong association of Response with all variables except Region Code. The negative coefficient for predictors “Age”, “Previously Insured” suggests customers who are in this group are less likely interested in Vehicle Insurance. The older the customers, the less likely they will be interested in Vehicle Insurance. The same goes with customers who were previously insured. On the other hand, customers who got vehicle damaged I the past and/or have older Vehicle Age will be more likely interested in Vehicle Insurance.

3.3 Evaluate Logistic Model Accuracy

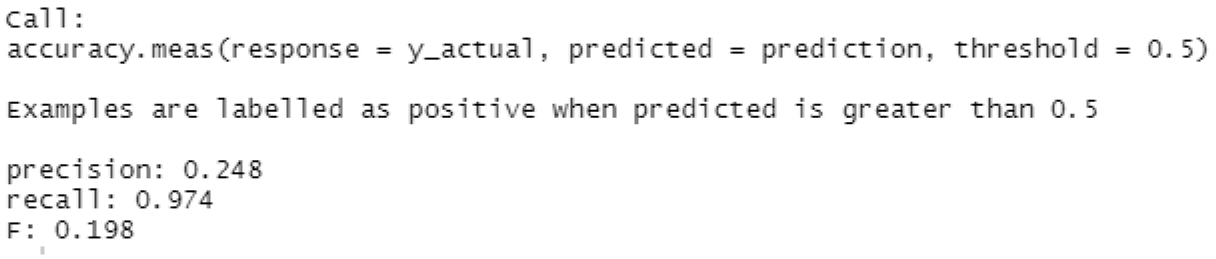
The precision of 0.248 gives us the percentage of Correctly Predicted Response from the pool of total Response. Low precision relates to the high false positive rate. We have 0.248 which is not good in this case. The model is not doing well with predicting customer interested in Vehicle Insurance



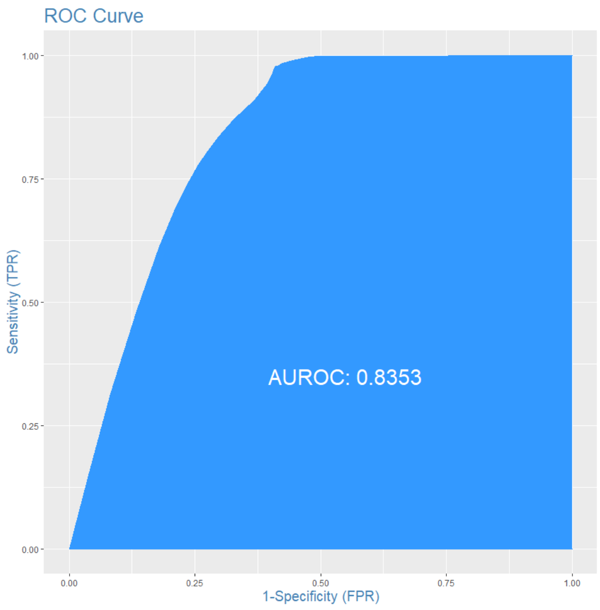
```
# Confusion Matrix
confusionMatrix(y_actual,prediction,threshold=0.5)
      0     1
0 59440   368
1 40980 13545
```

Recall is the ratio of correctly predicted positive observation. Here, we have 0.974 which is pretty good for this model.

With the Area under the curve (AUC) is 0.836 and high Sensitivity score, this model is better at predicting customers who are not interested in Vehicle Insurance than customers who interested.



```
Call:
accuracy.meas(response = y_actual, predicted = prediction, threshold = 0.5)

Examples are labelled as positive when predicted is greater than 0.5

precision: 0.248
recall: 0.974
F: 0.198
```



4 Section 4: Conclusion

There are many interesting findings while working with this project. The dataset that was provided by Kaggle is highly imbalanced which can be bias toward the majority class. Of all predictors available in the dataset, there is only 1 variable that is not associated with the target Response. Since the dataset is highly imbalanced, 4 methods were applied to find the best-balanced data.

Limitation- While the AUC ROC score for this model is good, but the model has high misclassification error of 36.16%. In order to solve this issue, different classification Machine

Learning Models should be examined for an imbalanced dataset such as Random Forest, XGBoost, KNN.