VEHICLE INSURANCE FRAUD DETECTION

AAMD REPORT

Submitted to:
Sir Hassaan Khalid

Submitted By Group 19
Dua Mahboob - 19296
Maaz Siddiqui - 22302
Muhammad Atif - 19227
Rida Fatima - 19343

Contents

Acknowledgments

We would like to thank Sir Hassaan Khalid for his mentorship and guidance throughout the project. In addition, we would also like to thank a fellow batchmate, Hammad Shehryar, who guided us in model building and limitations as he had worked on a similar report earlier.

Executive Summary

This report addresses the increasing threats of vehicle insurance fraud for insurance companies. The goal is to develop a model that can help companies in detecting insurance fraud. For that purpose, an extensive dataset was used to train and test a model. Initially, data was cleaned and transformed so that the model could be trained efficiently. It was found that the dataset was highly skewed, and countermeasures were taken to mitigate the biases. Different strategies and techniques were used to train different models. Subsequently, they were analyzed based on major KPIs to determine which one is the most effective. Based on the evaluation, policy recommendations were made for efficient implementation of the model keeping the cost-benefit analysis in mind. Lastly, the limitations of the research are discussed.

Introduction

Vehicle insurance fraud refers to the deliberate act of deceiving an insurance company in order to obtain undeserved benefits/financial gains related to a vehicle insurance policy. It involves various fraudulent activities that are aimed at manipulating the insurance system for personal gain. There are different types of vehicle insurance fraud including staged accidents, exaggerated/false claims, application fraud, fraudulent injury claims, phantom vehicles, etc.

With the vehicle industry and the disposable income of consumers rapidly growing, studying vehicle insurance fraud detection becomes extremely important for us. Studying vehicle insurance fraud detection is crucial for safeguarding the financial interests of insurers and policyholders, maintaining fairness and trust in the industry, enhancing customer satisfaction, complying with regulatory requirements, and upholding the reputation of insurance companies.

Problem Statement

Insurance fraud imposes a substantial burden on insurance companies, policyholders, and society as a whole. A high number of fraudulent insurance claims can lead to high financial losses and financial instability for the firm, an increased risk resulting in higher premiums for customers, and high scrutiny and tedious processes leading to lower customer satisfaction and trust.

By effectively identifying fraudulent activities through robust models, insurers can minimize the financial impact caused by fraudulent claims. This, in turn, ensures the stability and sustainability of the insurance industry. It will also help in maintaining premium stability. By actively detecting fraud, insurers can help keep premium rates stable and prevent honest policyholders from shouldering the burden of increased costs. Moreover, it will increase customer satisfaction. Fraudulent claims can result in delays in claim settlements and investigations, causing frustration and dissatisfaction among genuine policyholders. By proactively detecting and addressing fraud, insurance companies can enhance customer satisfaction by providing faster and more efficient claim processing. Lastly, effective fraud detection contributes to the reputation of insurance companies within the industry. A strong reputation for fraud detection and prevention enhances an insurer's standing within the industry and fosters confidence among stakeholders.

Dataset Description

The Vehicle Insurance Fraud Detection Dataset was obtained from Kaggle. It has 15.4k rows and 33 columns displaying demographics, timeline, claim details, and fault details as X variables and Fraud activity found as Y Variable.

Data Exploration

Data Cleaning

To clean our data, we used the following steps:

1. Identifying duplicates: the dataset had only one index variable i.e policy number so no duplicates were found because all were unique values (duplicate values were checked through conditional formatting).
2. Checking all columns for anomalies: the ‘day of the week’ and ‘month’ claimed for policyholder 1517 were 0 so we deleted them.
3. Removed repetitive and unnecessary columns:
    a. Since there were two columns for age, we removed the column with age as a numerical variable and kept the categorical variable “age of policyholder”.
    b. Removed policy number as it was an index variable.
    c. Removed policy type which was the combination of vehicle category and base policy.

We now have 31 columns with no missing values.

Data Transformation

Converting into Factor Variables

7 columns, Year, week of month,Fraudfound_P, PolicyNumber, RepNumber, Deductible, and DriverRating were identified as numeric variables. The rest were categorical. Fraud found is the binary Y variable and the rest which were identified as numeric had a categorical nature as well. (See Appendix for variable identification and definitions). Therefore, we converted these variables into factor variables using R.

Feature Engineering

We checked the summary to identify variables with the minimum count. Subsequently, we converted the sub-categories with minimum count into ‘others’ on R so that it does not give skewed results. We made the ‘others’ category for Make and Marital Status.

The data was quite extensive and rich in terms of explanatory variables; therefore, we did not feel the need to create any more variables.

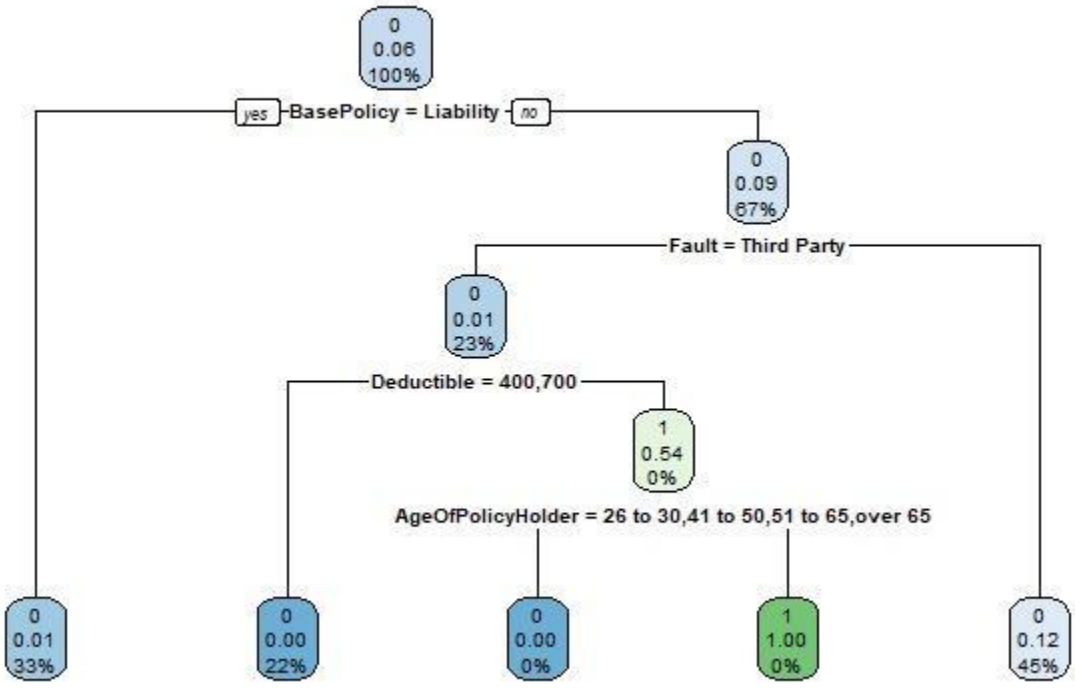
Model Building

Data Splitting

Since only 6% of our Y variable had Outcome=1, our data set was extremely skewed. Hence, we used stratified random sampling to ensure that the outcome variable is evenly distributed in training and testing. Our data split into training and testing was 70:30 subsequently.

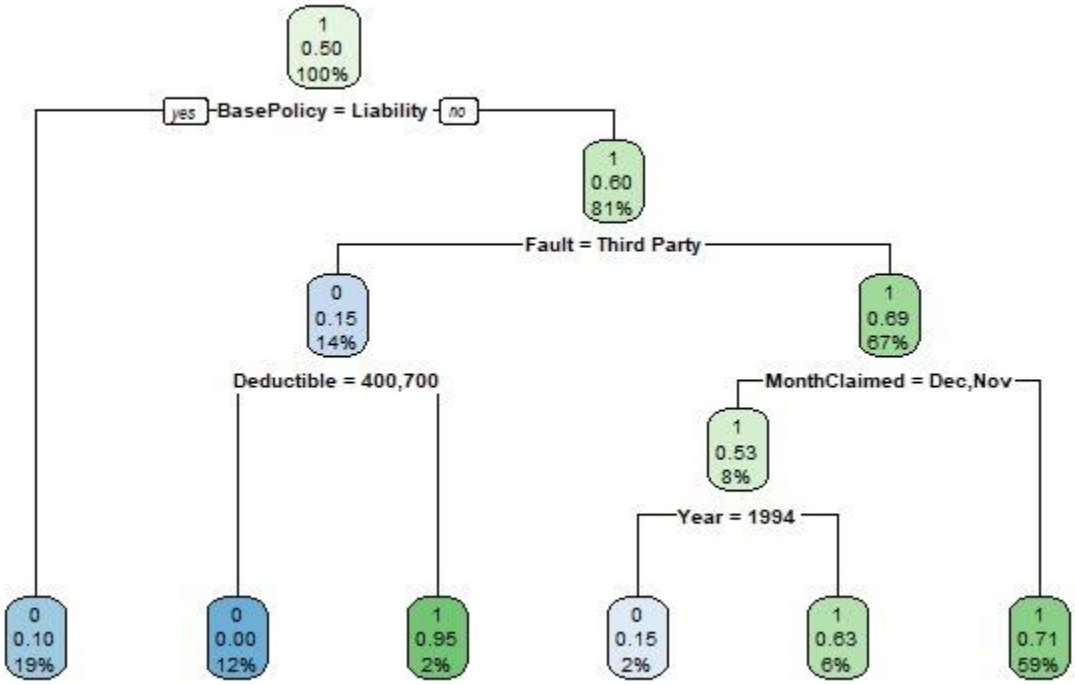
Model Training and Testing

We initially trained our model based on stratified random data splitting. The results were skewed and inconclusive as 94% of the values of the binary Y variable were 0, as is evident by the Decision Tree below.

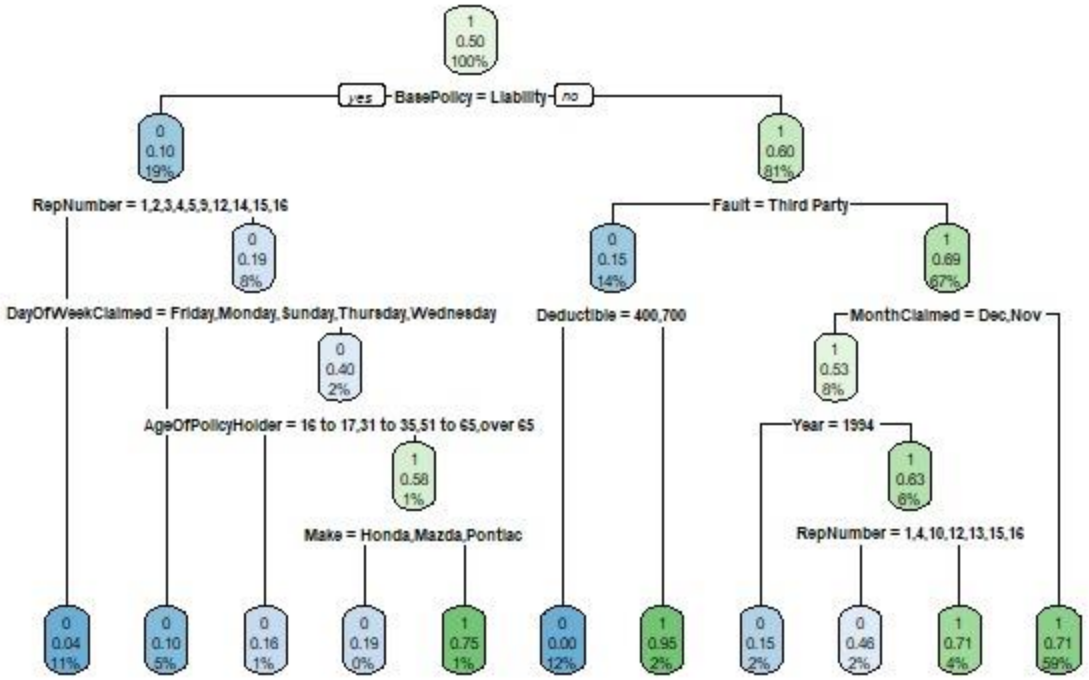


In such cases, the model may be biased towards the majority class and have difficulty learning patterns and making accurate predictions for the minority class. Then we assigned weights to address the class imbalance in the dataset i.e. 94% of values having outcome=0 and 6% of values

having outcome=1. The weights were assigned to the ratio of 1:16 for values 0,1 respectively. This aided in mitigating the issue by giving more importance to the minority class during model training. Then we trained all our models based on assigned weights. Following is the Decision Tree after the weighted approach.



Since the decision tree consisted of very few variables, we prompted it to further break down using minbucket and max depth function in R to maintain its readability and also generate valuable insights from it.



We also performed logistic regression on our datasets based on assigned weights. Then we calculated Accuracy, Precision, Recall, F-1 Score, AUC ROC, and AUC PR to compare and evaluate our models so that we are able to ground our recommendations to it.

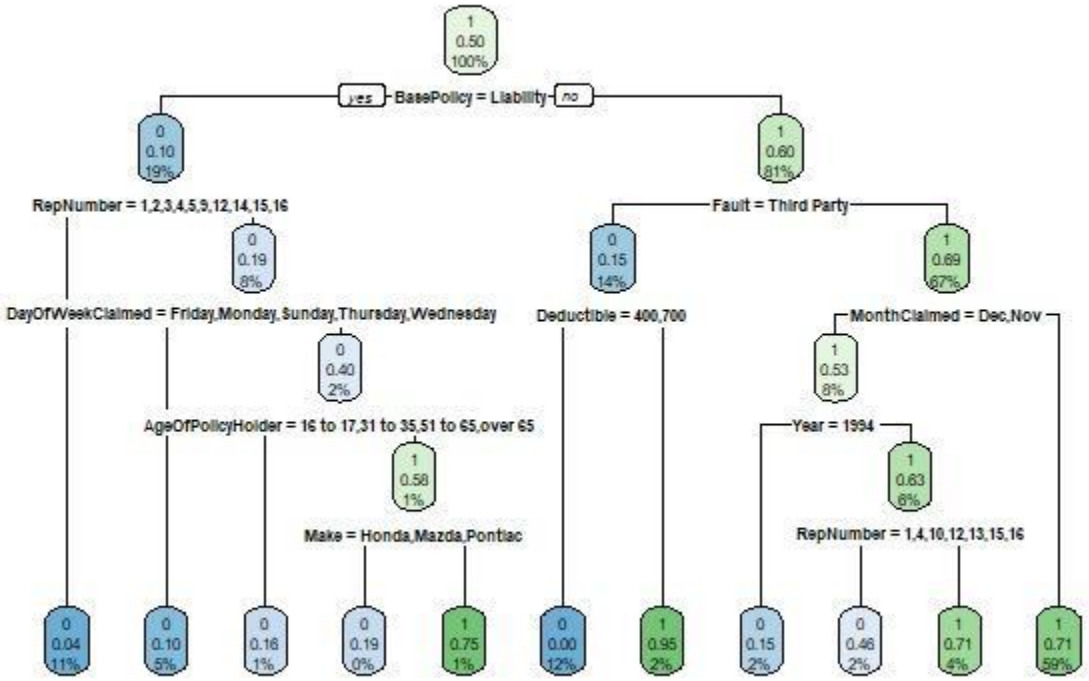
Model Analysis

Comparison of different models

Decision tree model	Accuracy	Precision	Recall	F1 score	AUC ROC	AUC PR
Model LR Train	0.77180	0.13784	0.90417	0.23921	0.83596	0.19003
Model LR Test	0.75300	0.12988	0.88043	0.22636	0.79617	0.14100
Model 2 Train	0.78221	0.13457	0.95672	0.23594	0.80078	0.16796
Model 2 Test	0.77273	0.13043	0.94565	0.22925	0.78997	0.16608
Model 3 Train	0.78881	0.13949	0.95209	0.24333	0.80632	0.17075
Model 3 Test	0.76615	0.13086	0.92029	0.22914	0.78623	0.15789

As observed in the table above, there is little variation between Model 2 and Model 3 which are decision tree models. We can also observe the trade-off between Precision and Recall as Model 3's Precision improved but Recall dropped compared to Model 2. However, F1 Score, AUC ROC and AUC PR all slightly increased for Model 3 which suggests that Model 3 is a better predictor than previous Models. Similarly, while Precision slightly increased for Model 3, Recall dropped slightly. This also highlights a trade-off and makes contextual preferences that much more important for us. Weighing in all the KPIs and the trade-offs, we decided to go with Model 3 because while there is little variation between the KPI values, this model gives a clearer in-depth view of the predictor variables. With the help of this model, we can identify meaningful relationships between more variables and make policy recommendations accordingly.

Analysis of the best model



The chosen model demonstrates encouraging performance in detecting auto insurance fraud. With an Area Under the ROC Curve (AUC-ROC) of 0.8063209, the model distinguishes between positive and negative classes effectively. However, the Precision-Recall Curve has a relatively low area under the curve (AUC-PR) of 0.1707505, indicating that there is room for improvement in simultaneously attaining high precision and recall. The confusion matrix reveals that the model accurately predicts non-fraud instances in the majority of instances (6,348 true negatives) but struggles with false positives (3,800 instances). While the model's high recall (0.95209) indicates effective fraud detection, its low precision (0.13949) results in a substantial number of false positives. Therefore, careful consideration is required to align the model's performance with the specific objectives and requirements of the auto insurance company, and additional optimization may be required to strike a balance between precision and accuracy.

Recommendations

Ideal fraud detection model would entail a model that accurately identifies all instances of actual fraud while avoiding false positives for legitimate transactions. However, it is rare to achieve this ideal equilibrium. In the context of fraud detection, it is essential to accomplish both a high recall, capturing majority of fraud cases, and reasonable precision, minimizing the costs associated with investigating cases that have been flagged.

The chosen model for this endeavor has a high recall rate of 0.95, indicating its ability to detect a vast majority of fraud incidents. However, it has only 0.14 percent accuracy for the positive class, resulting in a trade-off. This indicates the presence of a significant number of false-positive instances. Therefore, this strategy can be viable if the cost of human resources required to investigate the cases of fraud flagged by the model is significantly less than the cost incurred by actual cases of fraud that were ignored.

The Decision Tree (DT) indicates that fraud is less probable when the BasePolicy is "Liability" as opposed to "All Perils" or "Collision." This indicates that the company's policies regarding the latter category require revision and improvement. DT also emphasizes that "Third Party" or "Policy Holder" factor plays an important role in fraud. When a Third Party is at fault and the deductible is "300" or "500," fraud is highly probable. Similarly, fraud is more likely when the policyholder is at fault and the claim is filed in December or November. In these situations, the insurance company should pay special attention and consider implementing stricter rules and regulations, as well as requesting additional documentation and proof for verification.

Limitations

There are certain limitations with the dataset used for this project.

The most apparent limitation is the recency of data as it has been recording fraud claims from 1994 to 1996, which is quite outdated and obsolete today. The behavior, trends, and even the vehicles in the market have changed with time; hence this model might be inefficient for present-day situations for the purpose of detecting fraud. In order to efficiently detect fraud, working on present-day data will help in building a more accurate model.

Another limitation is that the model is not universally applicable and can only be used with this company because variables such as representative numbers are only specific to the company, so the model won't be usable with other companies' fraud detection. However, it does give us a framework and baseline to work with other models in the future for similar problem statements/objectives.

In case the cost of undetected fraud is less than the cost of investigation, the company might be inclined to rather let fraudulent instances occur rather than build models to detect the fraudulent activities and then set further inquiries. This model largely depends on the companies’ top and bottom lines.

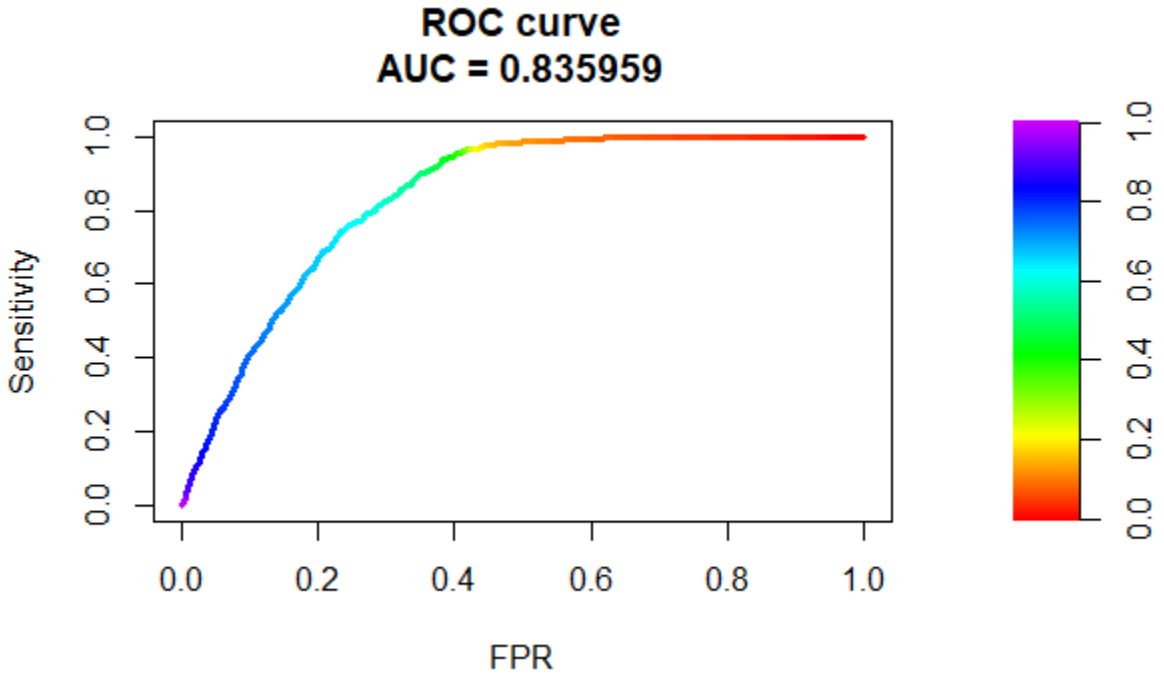
Appendix

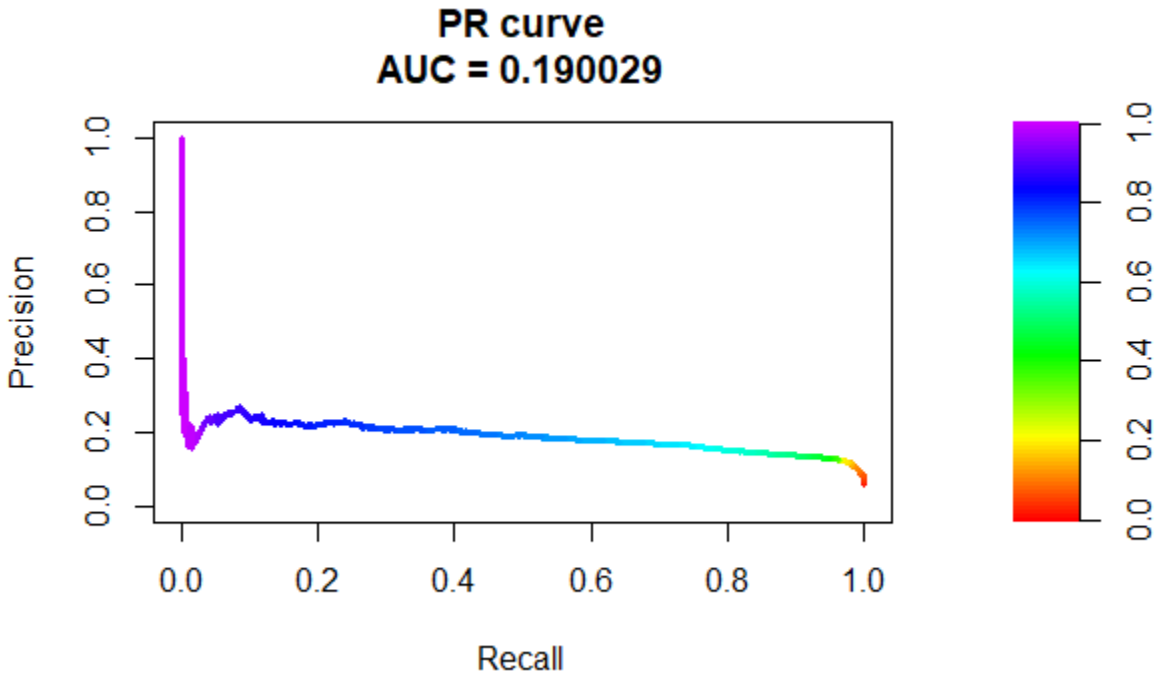
Variable Identification and Definition

Variable	Type	Definition
Month	Categorical	The month in which the accident happened
WeekOfMonth	Categorical	The week of the month in which the accident happened
DayOfWeek	Categorical	The day of the week in which the accident happened
Make	Categorical	The brand of the car
AccidentArea	Categorical	The area where the accident occurred (urban/rural)
DayOfWeekClaimed	Categorical	The day of the week in which the accident was claimed
MonthClaimed	Categorical	The month in which the accident was claimed
WeekOfMonthClaimed	Categorical	The week in which the accident was claimed
Sex	Categorical	Gender (Male/Female)
MaritalStatus	Categorical	Marital Status of the policy holder (Married/Single/Divorced/Widow)
Age	Numerical	Age of the policy holder
Fault	Categorical	Whose fault it was (Policy holder/third party)
PolicyType	Combination	Combination of Vehicle Type and Base Policy
VehicleCategory	Categorical	The category of the car among Sports, Utility, Sedan
VehiclePrice	Categorical	The price of the vehicle
FraudFound_P	Categorical	Was there a fraud found
PolicyNumber	Index	
RepNumber	Categorical	Representative of the Company
Deductible	Categorical	The initial amount paid by the policy owner
DriverRating	Categorical	What was the rating of the driver from 1-4
Days_Policy_Accident	Categorical	No. of days between the policy purchase and accident
Days_Policy_Claim	Categorical	No. of days between the policy purchased and the claim
PastNumberOfClaims	Categorical	No. of claims in the best
AgeOfVehicle	Categorical	Age of the vehicle at the time of the accident
AgeOfPolicyHolder	Categorical	Age of the policy holder
PoliceReportFiled	Categorical	Was their a police report filed after the accident
WitnessPresent	Categorical	Was there a witness present at the time of the accident
AgentType	Categorical	Was the policy sold under internal agent or external
NumberOfSuppliments	Categorical	Additional repair funds claimed to cover the damages
AddressChange_Claim	Categorical	Duration between the claim filed and change in address
NumberOfCars	Categorical	No. of cars that were involved in the accident
Year	Numerical	Year of accident
BasePolicy	Categorical	The type of damage covered by the policy

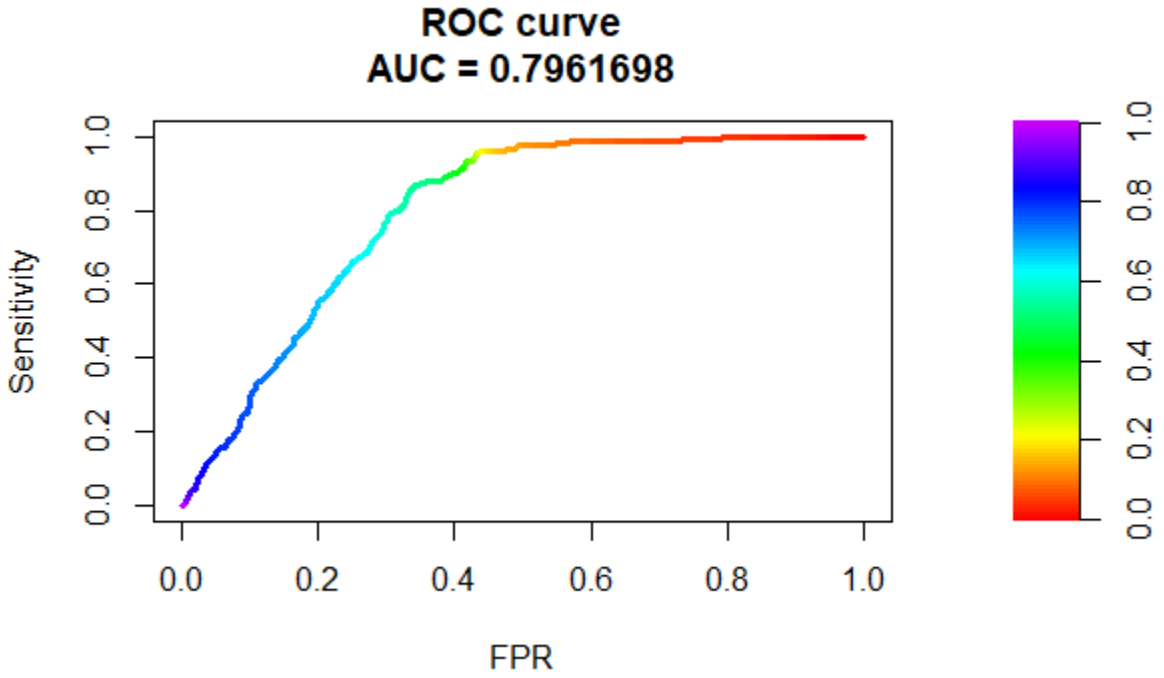
Model 1: Logistic Regression

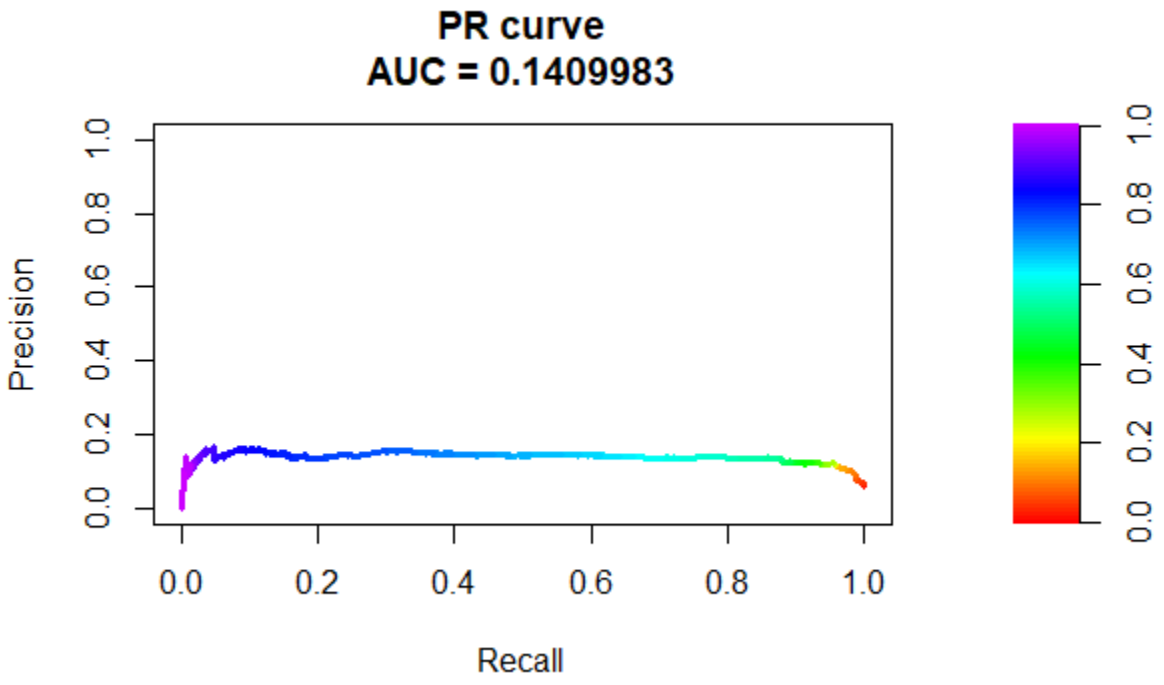
ROC & PR Plots - Training





ROC & PR Plots – Testing





Confusion Matrix –Training

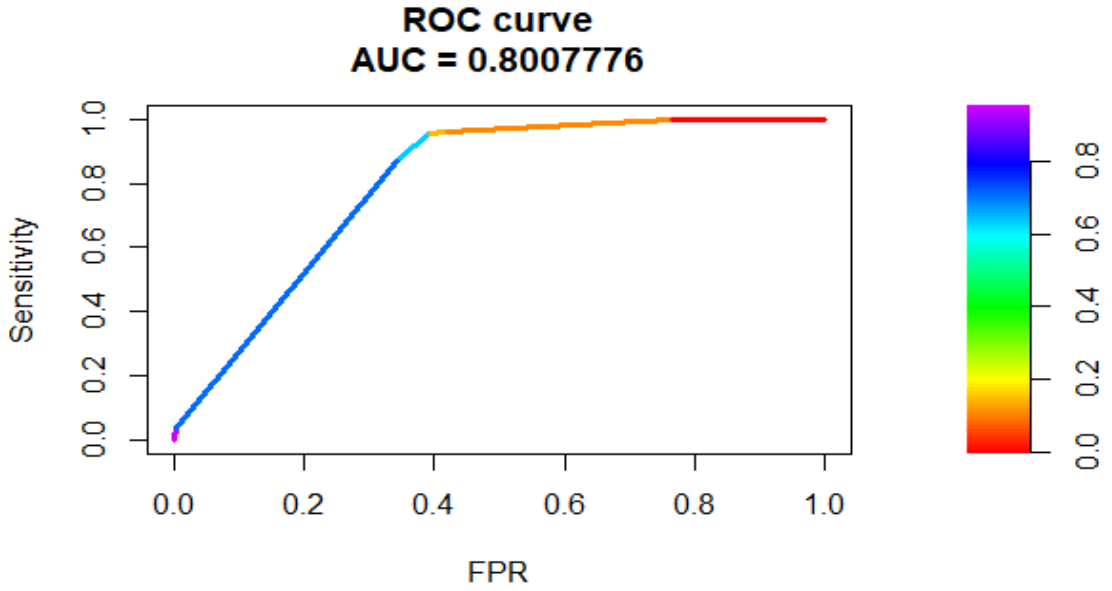
```
trainpred_c1    0    1
           0 6489   62
           1 3659  585

               Accuracy : 0.6553
                 95% CI : (0.6463, 0.6643)
    No Information Rate : 0.9401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1509

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6394
            Specificity : 0.9042
         Pos Pred Value : 0.9905
         Neg Pred Value : 0.1378
             Prevalence : 0.9401
         Detection Rate : 0.6011
   Detection Prevalence : 0.6069
      Balanced Accuracy : 0.7718
```

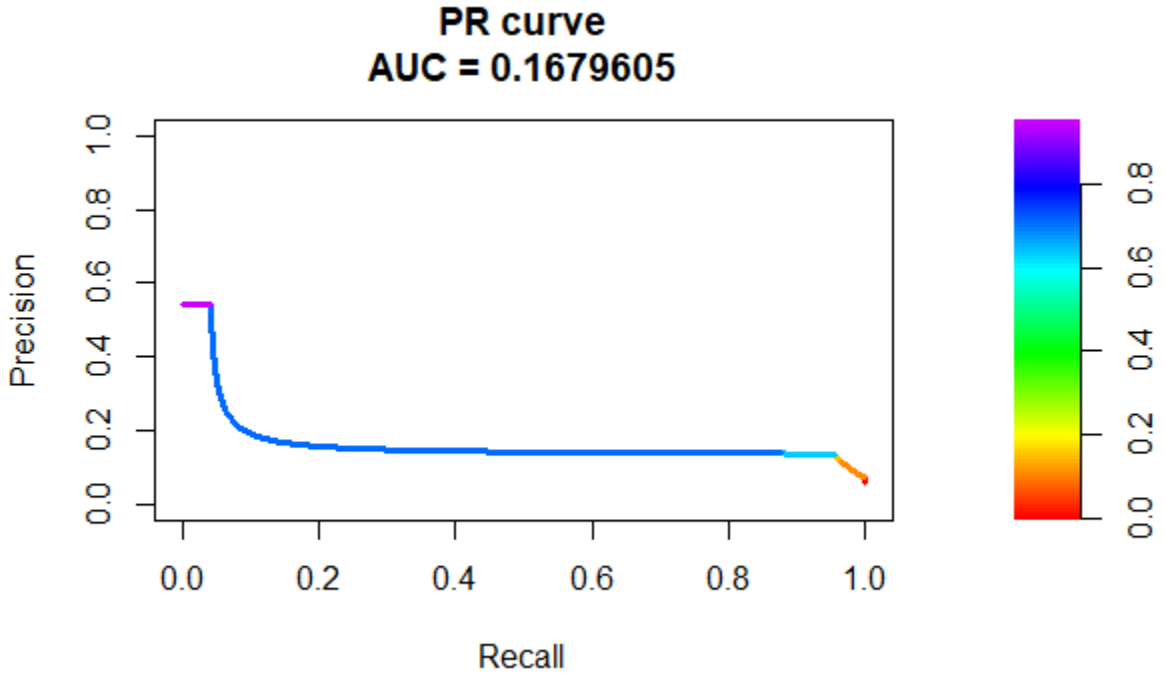
Confusion Matrix -Testing

```
testpred_c1    0    1
          0 2720   33
          1 1628  243

               Accuracy : 0.6408
                 95% CI : (0.6268, 0.6546)
    No Information Rate : 0.9403
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1365

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6256
            Specificity : 0.8804
         Pos Pred Value : 0.9880
         Neg Pred Value : 0.1299
             Prevalence : 0.9403
         Detection Rate : 0.5882
   Detection Prevalence : 0.5954
      Balanced Accuracy : 0.7530
```

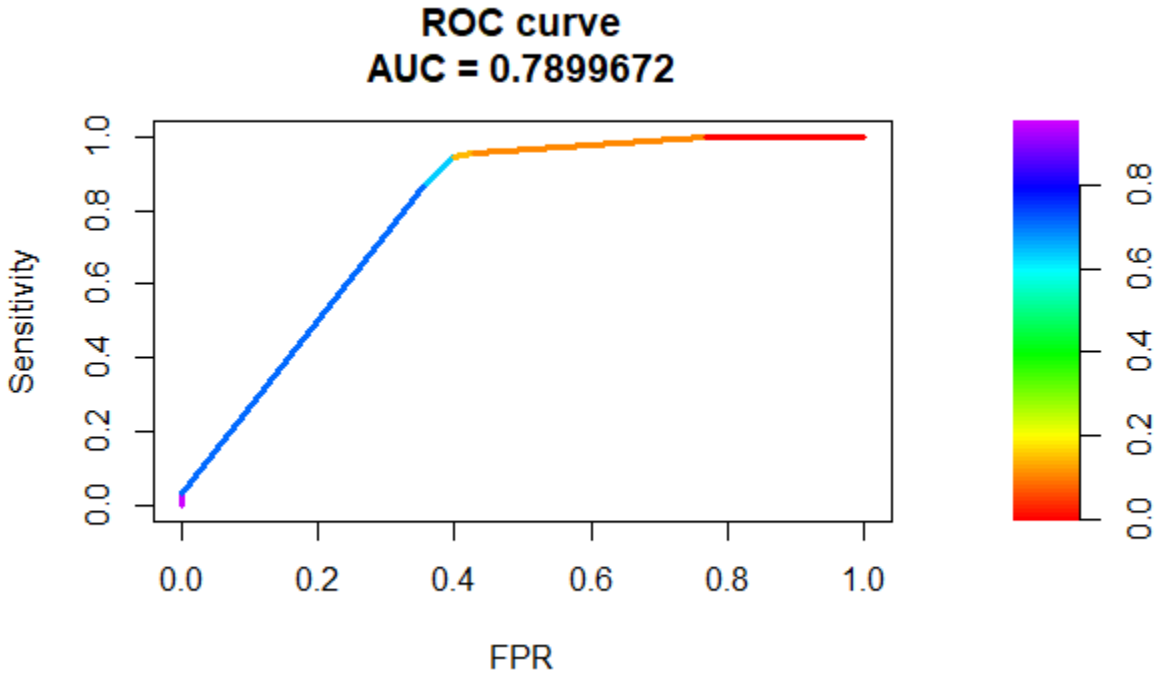
Model 2: Decision Tree without controls

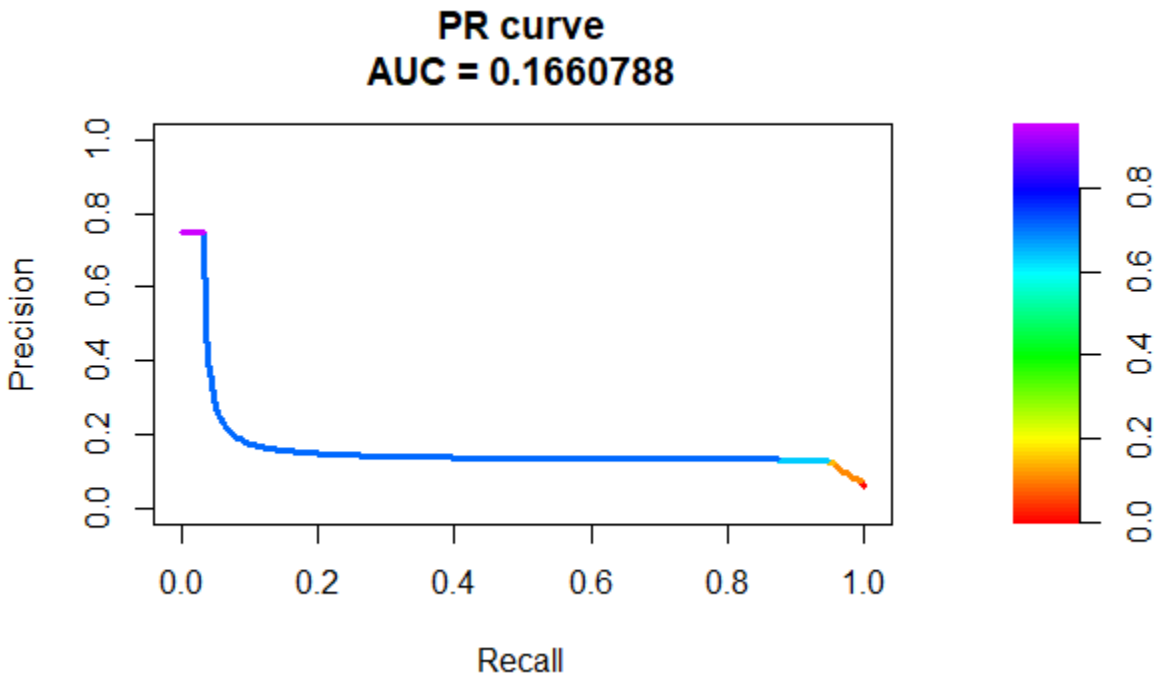
ROC & PR Plots – Training





ROC & PR Plots - Testing





Confusion Matrix – Training

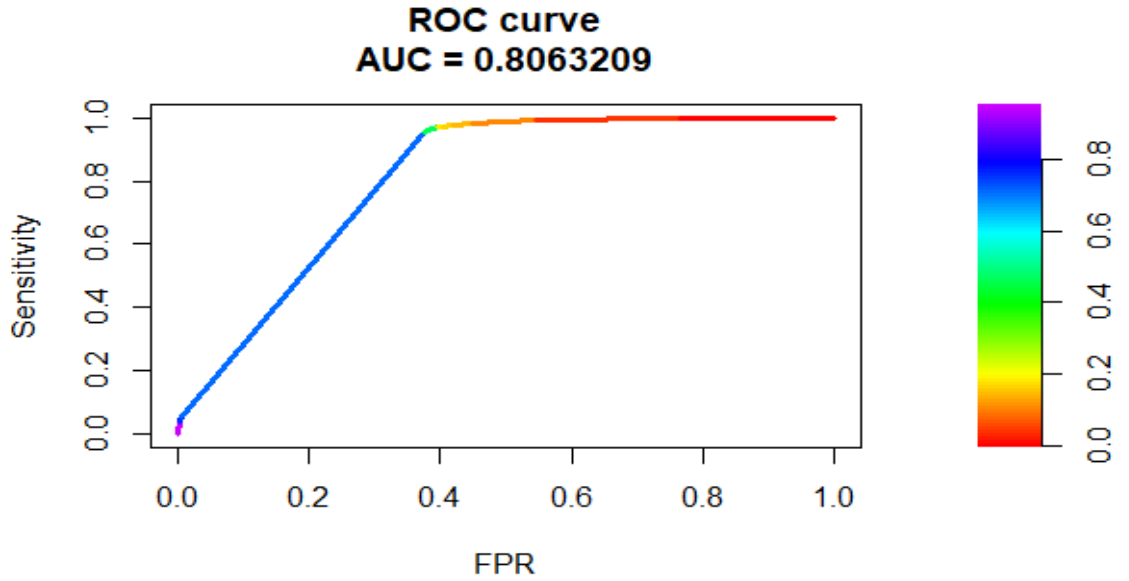
```
          Reference
Prediction    0    1
         0 6167   28
         1 3981  619

               Accuracy : 0.6286
                 95% CI : (0.6194, 0.6377)
    No Information Rate : 0.9401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1462

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.13457
                 Recall : 0.95672
                     F1 : 0.23594
             Prevalence : 0.05994
         Detection Rate : 0.05734
   Detection Prevalence : 0.42612
      Balanced Accuracy : 0.78221
```

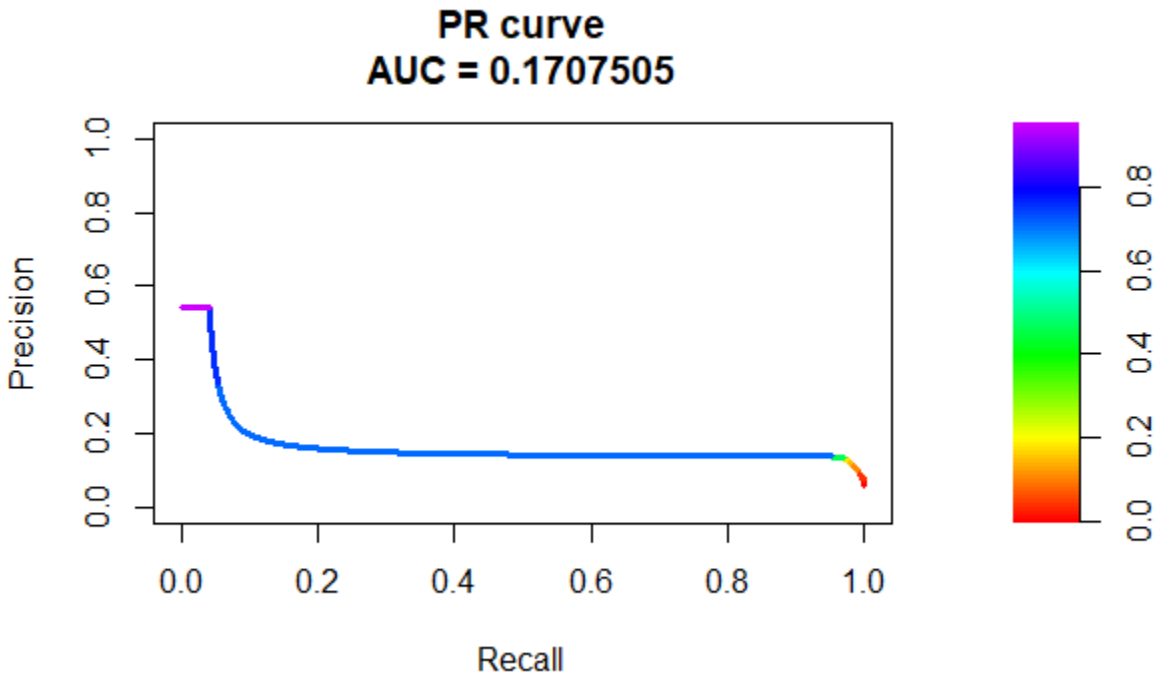
Confusion Matrix- Testing

```
Prediction    0    1
         0 2608   15
         1 1740  261

               Accuracy : 0.6205
                 95% CI : (0.6063, 0.6345)
    No Information Rate : 0.9403
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1389

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.13043
                 Recall : 0.94565
                     F1 : 0.22925
             Prevalence : 0.05969
         Detection Rate : 0.05644
   Detection Prevalence : 0.43274
      Balanced Accuracy : 0.77273

       'Positive' Class : 1
```

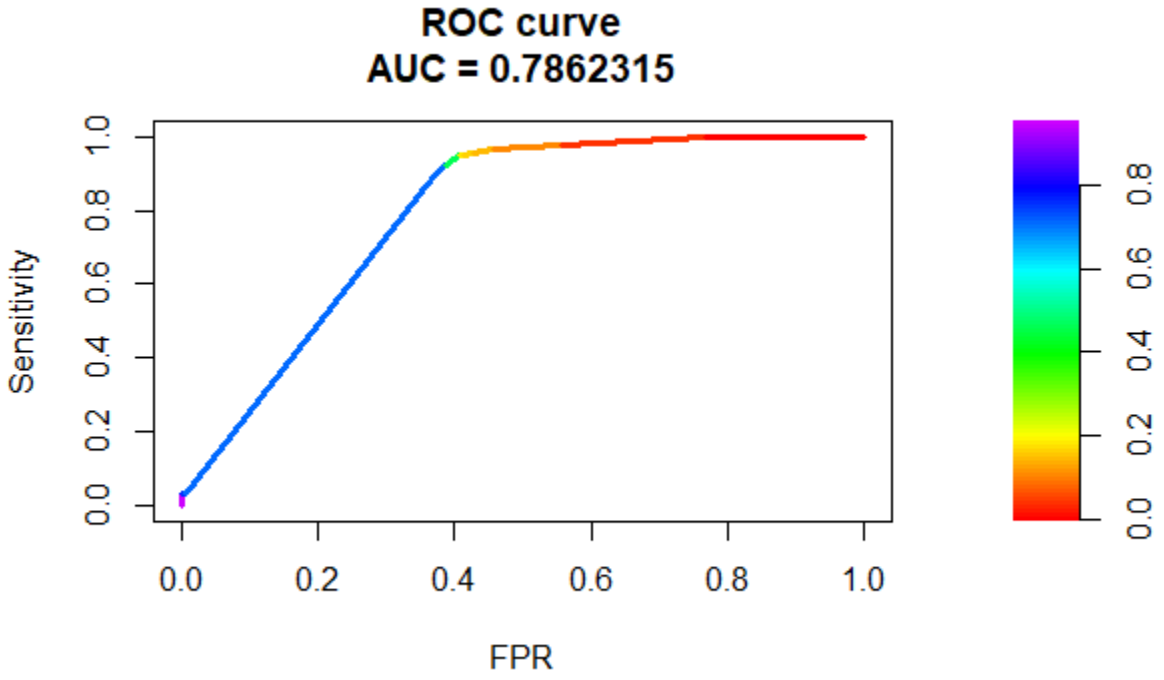
Model 3: Decision Tree With Controls

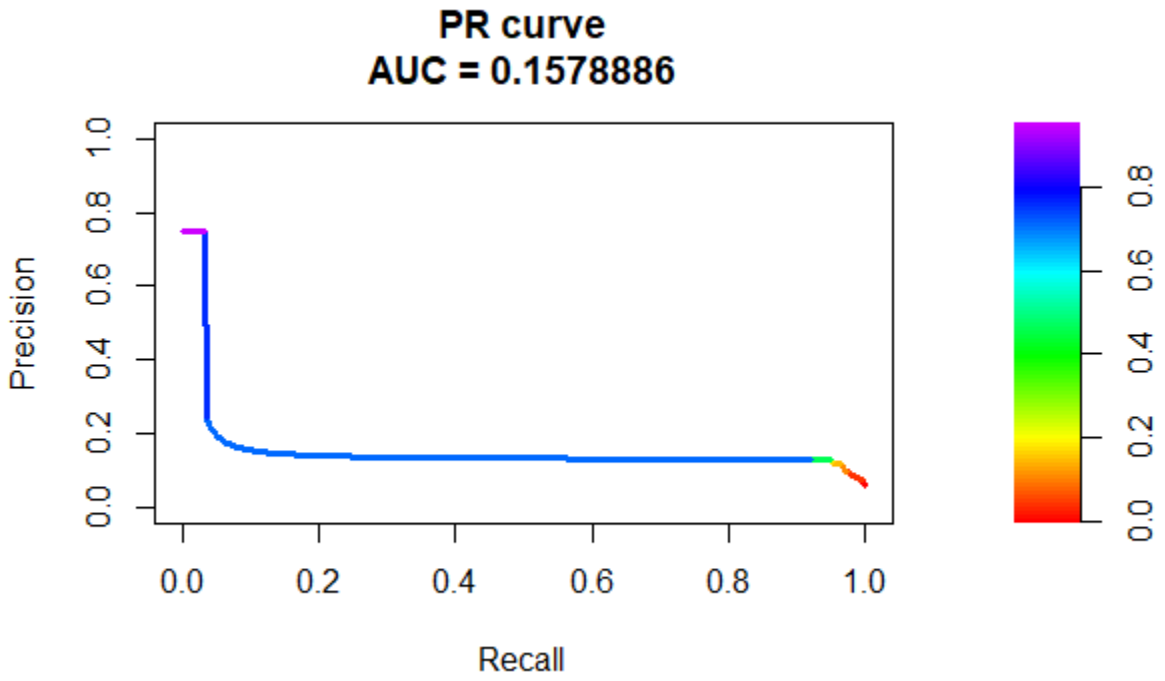
ROC & PR – Train





ROC & PR Plots- Testing





Confusion Matrix- Training

```
Prediction    0    1
         0 6348   31
         1 3800  616

               Accuracy : 0.6451
                 95% CI : (0.636, 0.6541)
    No Information Rate : 0.9401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.155

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.13949
                 Recall : 0.95209
                     F1 : 0.24333
             Prevalence : 0.05994
         Detection Rate : 0.05706
   Detection Prevalence : 0.40908
      Balanced Accuracy : 0.78881

       'Positive' Class : 1
```

Confusion Matrix- Testing

```
Prediction    0    1
         0 2661   22
         1 1687  254

               Accuracy : 0.6304
                 95% CI : (0.6163, 0.6443)
    No Information Rate : 0.9403
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1392

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.13086
                 Recall : 0.92029
                     F1 : 0.22914
             Prevalence : 0.05969
         Detection Rate : 0.05493
   Detection Prevalence : 0.41977
      Balanced Accuracy : 0.76615

       'Positive' Class : 1
```

R Script and Console

```
##DATA EXPLORATION
? rpart

## No documentation for 'rpart' in specified packages and libraries:
## you could try '??rpart'

#install.packages("vctrs")
library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

Rdata <- read_excel("Rdata.xlsx")
View(Rdata)
data <- Rdata
str(data)

## tibble [15,419 × 31] (S3: tbl_df/tbl/data.frame)
##  $ Month               : chr [1:15419] "Dec" "Jan" "Oct" "Jun" ...
##  $ WeekOfMonth         : num [1:15419] 5 3 5 2 5 4 1 1 4 3 ...
##  $ DayOfWeek           : chr [1:15419] "Wednesday" "Wednesday" "Fri
day" "Saturday" ...
##  $ Make                : chr [1:15419] "Honda" "Honda" "Honda" "Toy
ota" ...
```

```
##  $ AccidentArea        : chr [1:15419] "Urban" "Urban" "Urban" "Rur
al" ...
##  $ DayOfWeekClaimed    : chr [1:15419] "Tuesday" "Monday" "Thursday
" "Friday" ...
##  $ MonthClaimed        : chr [1:15419] "Jan" "Jan" "Nov" "Jul" ...
##  $ WeekOfMonthClaimed  : num [1:15419] 1 4 2 1 2 1 3 4 5 3 ...
##  $ Sex                 : chr [1:15419] "Female" "Male" "Male" "Male
" ...
##  $ MaritalStatus       : chr [1:15419] "Single" "Single" "Married"
"Married" ...
##  $ Fault               : chr [1:15419] "Policy Holder" "Policy Hold
er" "Policy Holder" "Third Party" ...
##  $ VehicleCategory     : chr [1:15419] "Sport" "Sport" "Sport" "Spo
rt" ...
##  $ VehiclePrice        : chr [1:15419] "more than 69000" "more than
69000" "more than 69000" "20000 to 29000" ...
##  $ FraudFound_P        : num [1:15419] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PolicyNumber        : num [1:15419] 1 2 3 4 5 6 7 8 9 10 ...
##  $ RepNumber           : num [1:15419] 12 15 7 4 3 12 14 1 7 7 ...
##  $ Deductible          : num [1:15419] 300 400 400 400 400 400 400
400 400 400 ...
##  $ DriverRating        : num [1:15419] 1 4 3 2 1 3 1 4 4 1 ...
##  $ Days_Policy_Accident: chr [1:15419] "more than 30" "more than 30
" "more than 30" "more than 30" ...
##  $ Days_Policy_Claim   : chr [1:15419] "more than 30" "more than 30
" "more than 30" "more than 30" ...
##  $ PastNumberOfClaims  : chr [1:15419] "none" "none" "1" "1" ...
##  $ AgeOfVehicle        : chr [1:15419] "3 years" "6 years" "7 years
" "more than 7" ...
##  $ AgeOfPolicyHolder   : chr [1:15419] "26 to 30" "31 to 35" "41 to
50" "51 to 65" ...
##  $ PoliceReportFiled   : chr [1:15419] "No" "Yes" "No" "Yes" ...
##  $ WitnessPresent      : chr [1:15419] "No" "No" "No" "No" ...
##  $ AgentType           : chr [1:15419] "External" "External" "Exter
nal" "External" ...
##  $ NumberOfSuppliments : chr [1:15419] "none" "none" "none" "more t
han 5" ...
##  $ AddressChange_Claim : chr [1:15419] "1 year" "no change" "no cha
nge" "no change" ...
##  $ NumberOfCars        : chr [1:15419] "3 to 4" "1 vehicle" "1 vehi
cle" "1 vehicle" ...
##  $ Year                : num [1:15419] 1994 1994 1994 1994 1994 ...
##  $ BasePolicy          : chr [1:15419] "Liability" "Collision" "Col
lision" "Liability" ...
```

```
summary(data)
```

```
##     Month              WeekOfMonth     DayOfWeek              Make
##  Length:15419       Min.   :1.000   Length:15419       Length:15419
##  Class :character   1st Qu.:2.000   Class :character   Class :chara
cter
##  Mode  :character   Median :3.000   Mode  :character   Mode  :chara
cter
##                     Mean   :2.789
##                     3rd Qu.:4.000
##                     Max.   :5.000
##  AccidentArea       DayOfWeekClaimed   MonthClaimed       WeekOfMon
thClaimed
##  Length:15419       Length:15419       Length:15419       Min.   :1
.000
##  Class :character   Class :character   Class :character   1st Qu.:2
.000
##  Mode  :character   Mode  :character   Mode  :character   Median :3
.000
##                                                           Mean   :2
.694
##                                                           3rd Qu.:4
.000
##                                                           Max.   :5
.000
##      Sex            MaritalStatus         Fault           VehicleCa
tegory
##  Length:15419       Length:15419       Length:15419       Length:15
419
##  Class :character   Class :character   Class :character   Class :ch
aracter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :ch
aracter
##
##
##
##  VehiclePrice        FraudFound_P      PolicyNumber     RepNumber
##  Length:15419       Min.   :0.00000   Min.   :    1   Min.   : 1.00
0
##  Class :character   1st Qu.:0.00000   1st Qu.: 3856   1st Qu.: 5.00
0
##  Mode  :character   Median :0.00000   Median : 7711   Median : 8.00
0
##                     Mean   :0.05986   Mean   : 7711   Mean   : 8.48
3
##                     3rd Qu.:0.00000   3rd Qu.:11566   3rd Qu.:12.00
0
##                     Max.   :1.00000   Max.   :15420   Max.   :16.00
```

```
##    Deductible     DriverRating   Days_Policy_Accident Days_Policy_C
laim
##  Min.   :300.0   Min.   :1.000   Length:15419         Length:15419
##  1st Qu.:400.0   1st Qu.:1.000   Class :character     Class :charac
ter
##  Median :400.0   Median :2.000   Mode  :character     Mode  :charac
ter
##  Mean   :407.7   Mean   :2.488
##  3rd Qu.:400.0   3rd Qu.:3.000
##  Max.   :700.0   Max.   :4.000
##  PastNumberOfClaims AgeOfVehicle       AgeOfPolicyHolder  PoliceRep
ortFiled
##  Length:15419       Length:15419       Length:15419       Length:15
419
##  Class :character   Class :character   Class :character   Class :ch
aracter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :ch
aracter
##
##
##
##  WitnessPresent      AgentType         NumberOfSuppliments AddressC
hange_Claim
##  Length:15419       Length:15419       Length:15419        Length:1
5419
##  Class :character   Class :character   Class :character    Class :c
haracter
##  Mode  :character   Mode  :character   Mode  :character    Mode  :c
haracter
##
##
##
##  NumberOfCars            Year       BasePolicy
##  Length:15419       Min.   :1994   Length:15419
##  Class :character   1st Qu.:1994   Class :character
##  Mode  :character   Median :1995   Mode  :character
##                     Mean   :1995
##                     3rd Qu.:1996
##                     Max.   :1996

colnames(data)

##  [1] "Month"                "WeekOfMonth"          "DayOfWeek"
##  [4] "Make"                 "AccidentArea"         "DayOfWeekClaime
d"
```

```
##  [7] "MonthClaimed"         "WeekOfMonthClaimed"   "Sex"
## [10] "MaritalStatus"        "Fault"                "VehicleCategory
"
## [13] "VehiclePrice"         "FraudFound_P"         "PolicyNumber"
## [16] "RepNumber"            "Deductible"           "DriverRating"
## [19] "Days_Policy_Accident" "Days_Policy_Claim"    "PastNumberOfCla
ims"
## [22] "AgeOfVehicle"         "AgeOfPolicyHolder"    "PoliceReportFil
ed"
## [25] "WitnessPresent"       "AgentType"            "NumberOfSupplim
ents"
## [28] "AddressChange_Claim"  "NumberOfCars"         "Year"
## [31] "BasePolicy"
```

```
#Checking for missing values
colSums(is.na(data))
```

```
##                Month          WeekOfMonth            DayOfWeek
##                    0                    0                    0
##                 Make         AccidentArea     DayOfWeekClaimed
##                    0                    0                    0
##         MonthClaimed   WeekOfMonthClaimed                  Sex
##                    0                    0                    0
##        MaritalStatus                Fault      VehicleCategory
##                    0                    0                    0
##         VehiclePrice         FraudFound_P         PolicyNumber
##                    0                    0                    0
##            RepNumber           Deductible         DriverRating
##                    0                    0                    0
## Days_Policy_Accident    Days_Policy_Claim   PastNumberOfClaims
##                    0                    0                    0
##         AgeOfVehicle    AgeOfPolicyHolder    PoliceReportFiled
##                    0                    0                    0
##       WitnessPresent           AgentType  NumberOfSuppliments
##                    0                    0                    0
##  AddressChange_Claim         NumberOfCars                 Year
##                    0                    0                    0
##           BasePolicy
##                    0
```

```
##DATA CLEANING & FEATURE REENGINEERING

#Removing Policy Number Column
column_name <- "PolicyNumber"
column_index <- which(colnames(data) == column_name)
print(column_index)
```

```
## [1] 15
```

```
data <- data[, -c(15)]
```

```
#Changing Minimum Count variables into "Others"
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df1 <- data %>%
  count(Make)
```

```
#Total 19 Makes of cars with only 5 having more than 1500 entries
#We are clubbing the rest in "Others"
data2 <- data
unique(data2$Make)
```

```
##  [1] "Honda"     "Toyota"    "Ford"      "Mazda"     "Chevrolet" "P
ontiac"
##  [7] "Accura"    "Dodge"     "Mercury"   "Jaguar"    "Nisson"    "V
W"
## [13] "Saab"      "Saturn"    "Porche"    "BMW"       "Mecedes"   "F
errari"
## [19] "Lexus"
```

```
#Removing minimum categories into others before changing into factor
#for "Make" Variable
data2$Make[data2$Make== "Ferrari"] <- "Others"
data2$Make[data2$Make== "Lexus"] <- "Others"
data2$Make[data2$Make== "Mercedes"] <- "Others"
data2$Make[data2$Make== "Porche"] <- "Others"
data2$Make[data2$Make== "Jaguar"] <- "Others"
data2$Make[data2$Make== "BMW"] <- "Others"
data2$Make[data2$Make== "Nisson"] <- "Others"
data2$Make[data2$Make== "Saturn"] <- "Others"
data2$Make[data2$Make== "Mercury"] <- "Others"
data2$Make[data2$Make== "Saab"] <- "Others"
data2$Make[data2$Make== "Dodge"] <- "Others"
```

```
data2$Make[data2$Make== "VW"] <- "Others"
data2$Make[data2$Make== "Ford"] <- "Others"
data2$Make[data2$Make== "Accura"] <- "Others"
unique(data2$Make)
```

```
## [1] "Honda"     "Toyota"    "Others"    "Mazda"     "Chevrolet" "Po
ntiac"
## [7] "Mecedes"
```

```
data2$Make[data2$Make== "Mecedes"] <- "Others" #Typo in Mercedes


#MaritalStatus
df2 <- data %>%
  count(MaritalStatus)
#We clubbed divorced and Widow into single as tehy were less than 100
data2$MaritalStatus[data2$MaritalStatus== "Widow"] <- "Single"
data2$MaritalStatus[data2$MaritalStatus== "Divorced"] <- "Single"
unique(data2$MaritalStatus)
```

```
## [1] "Single"  "Married"
```

```
#Converting into factor variable
df <- as.data.frame(lapply(data2, factor)) #All of remaining data is c
ategorical
str(df)
```

```
## 'data.frame':    15419 obs. of  30 variables:
##  $ Month               : Factor w/ 12 levels "Apr","Aug","Dec",..:
3 5 11 7 5 11 4 10 3 1 ...
##  $ WeekOfMonth         : Factor w/ 5 levels "1","2","3","4",..: 5 3
5 2 5 4 1 1 4 3 ...
##  $ DayOfWeek           : Factor w/ 7 levels "Friday","Monday",..: 7
7 1 3 2 1 3 1 3 6 ...
##  $ Make                : Factor w/ 6 levels "Chevrolet","Honda",..:
2 2 2 6 2 2 2 2 2 4 ...
##  $ AccidentArea        : Factor w/ 2 levels "Rural","Urban": 2 2 2
1 2 2 2 2 2 2 ...
##  $ DayOfWeekClaimed    : Factor w/ 7 levels "Friday","Monday",..: 6
2 5 1 6 7 2 6 7 7 ...
##  $ MonthClaimed        : Factor w/ 12 levels "Apr","Aug","Dec",..:
5 5 10 6 4 10 4 8 3 1 ...
##  $ WeekOfMonthClaimed  : Factor w/ 5 levels "1","2","3","4",..: 1 4
2 1 2 1 3 4 5 3 ...
##  $ Sex                 : Factor w/ 2 levels "Female","Male": 1 2 2
2 1 2 2 2 2 2 ...
##  $ MaritalStatus       : Factor w/ 2 levels "Married","Single": 2 2
```

```
1 1 2 2 1 2 2 1 ...
##  $ Fault               : Factor w/ 2 levels "Policy Holder",..: 1 1
1 2 2 2 2 1 1 1 ...
##  $ VehicleCategory     : Factor w/ 3 levels "Sedan","Sport",..: 2 2
2 2 2 2 2 2 2 3 ...
##  $ VehiclePrice        : Factor w/ 6 levels "20000 to 29000",..: 6
6 6 1 6 6 6 6 6 6 ...
##  $ FraudFound_P        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 1 ...
##  $ RepNumber           : Factor w/ 16 levels "1","2","3","4",..: 12
15 7 4 3 12 14 1 7 7 ...
##  $ Deductible          : Factor w/ 4 levels "300","400","500",..: 1
2 2 2 2 2 2 2 2 2 ...
##  $ DriverRating        : Factor w/ 4 levels "1","2","3","4": 1 4 3
2 1 3 1 4 4 1 ...
##  $ Days_Policy_Accident: Factor w/ 5 levels "1 to 7","15 to 30",..:
4 4 4 4 4 4 4 4 4 4 ...
##  $ Days_Policy_Claim   : Factor w/ 3 levels "15 to 30","8 to 15",..
: 3 3 3 3 3 3 3 3 3 3 ...
##  $ PastNumberOfClaims  : Factor w/ 4 levels "1","2 to 4","more than
4",..: 4 4 1 1 4 4 1 1 4 2 ...
##  $ AgeOfVehicle        : Factor w/ 8 levels "2 years","3 years",..:
2 5 6 7 4 4 6 8 5 7 ...
##  $ AgeOfPolicyHolder   : Factor w/ 9 levels "16 to 17","18 to 20",.
.: 4 5 7 8 5 3 6 1 5 6 ...
##  $ PoliceReportFiled   : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1
1 1 1 1 ...
##  $ WitnessPresent      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1
1 1 2 1 ...
##  $ AgentType           : Factor w/ 2 levels "External","Internal":
1 1 1 1 1 1 1 1 1 1 ...
##  $ NumberOfSuppliments : Factor w/ 4 levels "1 to 2","3 to 5",..: 4
4 4 3 4 2 1 4 2 2 ...
##  $ AddressChange_Claim : Factor w/ 5 levels "1 year","2 to 3 years"
,..: 1 4 4 4 4 4 4 4 4 4 ...
##  $ NumberOfCars        : Factor w/ 5 levels "1 vehicle","2 vehicles
",..: 3 1 1 1 1 1 1 1 1 1 ...
##  $ Year                : Factor w/ 3 levels "1994","1995",..: 1 1 1
1 1 1 1 1 1 1 ...
##  $ BasePolicy          : Factor w/ 3 levels "All Perils","Collision
",..: 3 2 2 3 2 2 2 2 2 1 ...

summary(df)

##      Month      WeekOfMonth     DayOfWeek           Make      Accid
entArea
```

```
##  Jan    :1411   1:3187      Friday   :2445   Chevrolet:1681   Rural
: 1597
##  May    :1367   2:3557      Monday   :2615   Honda    :2800   Urban
:13822
##  Mar    :1360   3:3640      Saturday :1982   Mazda    :2354
##  Jun    :1321   4:3398      Sunday   :1745   Others   :1626
##  Oct    :1305   5:1637      Thursday :2173   Pontiac  :3837
##  Dec    :1285               Tuesday  :2300   Toyota   :3121
##  (Other):7370               Wednesday:2159
##   DayOfWeekClaimed  MonthClaimed  WeekOfMonthClaimed     Sex
##  Friday   :2497    Jan    :1446   1:3449             Female: 2420
##  Monday   :3757    May    :1411   2:3720             Male  :12999
##  Saturday : 127    Mar    :1348   3:3583
##  Sunday   :  52    Oct    :1339   4:3433
##  Thursday :2660    Jun    :1293   5:1234
##  Tuesday  :3375    Feb    :1287
##  Wednesday:2951    (Other):7295
##  MaritalStatus             Fault       VehicleCategory          Veh
iclePrice
##  Married:10625   Policy Holder:11229   Sedan  :9670    20000 to 290
00 :8079
##  Single : 4794   Third Party  : 4190   Sport  :5358    30000 to 390
00 :3533
##                                        Utility: 391    40000 to 590
00 : 461
##                                                        60000 to 690
00 :  87
##                                                        less than 20
000:1096
##                                                        more than 69
000:2163
##
##  FraudFound_P   RepNumber    Deductible  DriverRating   Days_Policy
_Accident
##  0:14496      7      :1069   300:    8   1:3944       1 to 7      :
14
##  1:  923      9      : 999   400:14837   2:3800       15 to 30    :
49
##               1      : 987   500:  263   3:3884       8 to 15     :
55
##               5      : 987   700:  311   4:3791       more than 30:
15246
##               10     : 986                            none        :
55
##               12     : 977
##               (Other):9414
```

```
##     Days_Policy_Claim   PastNumberOfClaims      AgeOfVehicle  AgeOf
PolicyHolder
##  15 to 30    :   56   1          :3573     7 years    :5807   31 to
35:5593
##  8 to 15     :   21   2 to 4     :5485     more than 7:3981   36 to
40:4043
##  more than 30:15342   more than 4:2010     6 years    :3448   41 to
50:2828
##                       none       :4351     5 years    :1357   51 to
65:1392
##                                             new        : 372   26 to
30: 613
##                                             4 years    : 229   over
65 : 508
##                                             (Other)    : 225   (Othe
r) : 442
##  PoliceReportFiled WitnessPresent    AgentType      NumberOfSupplim
ents
##  No :14991         No :15332      External:15178   1 to 2     :2489
##  Yes:  428         Yes:   87      Internal:  241   3 to 5     :2017
##                                                    more than 5:3867
##                                                    none       :7046
##
##
##
##      AddressChange_Claim      NumberOfCars     Year           BaseP
olicy
##  1 year        :  170    1 vehicle  :14315   1994:6141   All Perils
:4448
##  2 to 3 years  :  291    2 vehicles :  709   1995:5195   Collision
:5962
##  4 to 8 years  :  631    3 to 4     :  372   1996:4083   Liability
:5009
##  no change     :14323    5 to 8     :   21
##  under 6 months:    4    more than 8:    2
##
##
```

```
# Importing libraries
#install.packages("caret")
#updating for caret
#install.packages("tibble")
#install.packages("purrr")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: ggplot2

## Loading required package: lattice

#Splitting Data
#For equal division of our target variable, using stratified
set.seed(123)

?createDataPartition

## starting httpd help server ...

##  done

library(caret)
#CreateDataPartition does stratified by default
set.seed(123)
trainIndex <-
  createDataPartition(df$FraudFound_P,
                      p = 0.7,
                      list = FALSE,
                      times = 1)

train <- df[trainIndex,]
test <- df[-trainIndex,]


#Decision Tree Model: Without Weights
library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.2.3

str(df)

## 'data.frame':    15419 obs. of  30 variables:
##  $ Month               : Factor w/ 12 levels "Apr","Aug","Dec",..:
3 5 11 7 5 11 4 10 3 1 ...
##  $ WeekOfMonth         : Factor w/ 5 levels "1","2","3","4",..: 5 3
5 2 5 4 1 1 4 3 ...
##  $ DayOfWeek           : Factor w/ 7 levels "Friday","Monday",..: 7
7 1 3 2 1 3 1 3 6 ...
##  $ Make                : Factor w/ 6 levels "Chevrolet","Honda",..:
2 2 2 6 2 2 2 2 2 4 ...
##  $ AccidentArea        : Factor w/ 2 levels "Rural","Urban": 2 2 2
1 2 2 2 2 2 2 ...
##  $ DayOfWeekClaimed    : Factor w/ 7 levels "Friday","Monday",..: 6
2 5 1 6 7 2 6 7 7 ...
```

```
##  $ MonthClaimed        : Factor w/ 12 levels "Apr","Aug","Dec",..:
5 5 10 6 4 10 4 8 3 1 ...
##  $ WeekOfMonthClaimed  : Factor w/ 5 levels "1","2","3","4",..: 1 4
2 1 2 1 3 4 5 3 ...
##  $ Sex                 : Factor w/ 2 levels "Female","Male": 1 2 2
2 1 2 2 2 2 2 ...
##  $ MaritalStatus       : Factor w/ 2 levels "Married","Single": 2 2
1 1 2 2 1 2 2 1 ...
##  $ Fault               : Factor w/ 2 levels "Policy Holder",..: 1 1
1 2 2 2 2 1 1 1 ...
##  $ VehicleCategory     : Factor w/ 3 levels "Sedan","Sport",..: 2 2
2 2 2 2 2 2 2 3 ...
##  $ VehiclePrice        : Factor w/ 6 levels "20000 to 29000",..: 6
6 6 1 6 6 6 6 6 6 ...
##  $ FraudFound_P        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 1 ...
##  $ RepNumber           : Factor w/ 16 levels "1","2","3","4",..: 12
15 7 4 3 12 14 1 7 7 ...
##  $ Deductible          : Factor w/ 4 levels "300","400","500",..: 1
2 2 2 2 2 2 2 2 2 ...
##  $ DriverRating        : Factor w/ 4 levels "1","2","3","4": 1 4 3
2 1 3 1 4 4 1 ...
##  $ Days_Policy_Accident: Factor w/ 5 levels "1 to 7","15 to 30",..:
4 4 4 4 4 4 4 4 4 4 ...
##  $ Days_Policy_Claim   : Factor w/ 3 levels "15 to 30","8 to 15",..
: 3 3 3 3 3 3 3 3 3 3 ...
##  $ PastNumberOfClaims  : Factor w/ 4 levels "1","2 to 4","more than
4",..: 4 4 1 1 4 4 1 1 4 2 ...
##  $ AgeOfVehicle        : Factor w/ 8 levels "2 years","3 years",..:
2 5 6 7 4 4 6 8 5 7 ...
##  $ AgeOfPolicyHolder   : Factor w/ 9 levels "16 to 17","18 to 20",.
.: 4 5 7 8 5 3 6 1 5 6 ...
##  $ PoliceReportFiled   : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1
1 1 1 1 ...
##  $ WitnessPresent      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1
1 1 2 1 ...
##  $ AgentType           : Factor w/ 2 levels "External","Internal":
1 1 1 1 1 1 1 1 1 1 ...
##  $ NumberOfSuppliments : Factor w/ 4 levels "1 to 2","3 to 5",..: 4
4 4 3 4 2 1 4 2 2 ...
##  $ AddressChange_Claim : Factor w/ 5 levels "1 year","2 to 3 years"
,..: 1 4 4 4 4 4 4 4 4 4 ...
##  $ NumberOfCars        : Factor w/ 5 levels "1 vehicle","2 vehicles
",..: 3 1 1 1 1 1 1 1 1 1 ...
##  $ Year                : Factor w/ 3 levels "1994","1995",..: 1 1 1
1 1 1 1 1 1 1 ...
```

```
##  $ BasePolicy          : Factor w/ 3 levels "All Perils","Collision
",..: 3 2 2 3 2 2 2 2 2 1 ...
```

```
summary(df)
```

```
##      Month      WeekOfMonth     DayOfWeek           Make      Accid
entArea
##  Jan    :1411   1:3187      Friday   :2445   Chevrolet:1681   Rural
: 1597
##  May    :1367   2:3557      Monday   :2615   Honda    :2800   Urban
:13822
##  Mar    :1360   3:3640      Saturday :1982   Mazda    :2354
##  Jun    :1321   4:3398      Sunday   :1745   Others   :1626
##  Oct    :1305   5:1637      Thursday :2173   Pontiac  :3837
##  Dec    :1285               Tuesday  :2300   Toyota   :3121
##  (Other):7370               Wednesday:2159
##   DayOfWeekClaimed  MonthClaimed  WeekOfMonthClaimed     Sex
##  Friday   :2497    Jan    :1446   1:3449             Female: 2420
##  Monday   :3757    May    :1411   2:3720             Male  :12999
##  Saturday : 127    Mar    :1348   3:3583
##  Sunday   :  52    Oct    :1339   4:3433
##  Thursday :2660    Jun    :1293   5:1234
##  Tuesday  :3375    Feb    :1287
##  Wednesday:2951    (Other):7295
##  MaritalStatus             Fault       VehicleCategory          Veh
iclePrice
##  Married:10625   Policy Holder:11229   Sedan  :9670    20000 to 290
00 :8079
##  Single : 4794   Third Party  : 4190   Sport  :5358    30000 to 390
00 :3533
##                                        Utility: 391    40000 to 590
00 : 461
##                                                        60000 to 690
00 :  87
##                                                        less than 20
000:1096
##                                                        more than 69
000:2163
##
##  FraudFound_P   RepNumber    Deductible  DriverRating   Days_Policy
_Accident
##  0:14496      7      :1069   300:    8   1:3944       1 to 7      :
14
##  1:  923      9      : 999   400:14837   2:3800       15 to 30    :
49
##               1      : 987   500:  263   3:3884       8 to 15     :
```

```
55
##               5      : 987   700:  311   4:3791       more than 30:
15246
##               10     : 986                            none        :
55
##               12     : 977
##               (Other):9414
##     Days_Policy_Claim   PastNumberOfClaims      AgeOfVehicle  AgeOf
PolicyHolder
##  15 to 30    :   56   1          :3573     7 years    :5807   31 to
35:5593
##  8 to 15     :   21   2 to 4     :5485     more than 7:3981   36 to
40:4043
##  more than 30:15342   more than 4:2010     6 years    :3448   41 to
50:2828
##                       none       :4351     5 years    :1357   51 to
65:1392
##                                             new        : 372   26 to
30: 613
##                                             4 years    : 229   over
65 : 508
##                                             (Other)    : 225   (Othe
r) : 442
##  PoliceReportFiled WitnessPresent    AgentType      NumberOfSupplim
ents
##  No :14991         No :15332      External:15178   1 to 2     :2489
##  Yes:  428         Yes:   87      Internal:  241   3 to 5     :2017
##                                                    more than 5:3867
##                                                    none       :7046
##
##
##
##      AddressChange_Claim      NumberOfCars     Year           BaseP
olicy
##  1 year        :  170    1 vehicle  :14315   1994:6141   All Perils
:4448
##  2 to 3 years  :  291    2 vehicles :  709   1995:5195   Collision
:5962
##  4 to 8 years  :  631    3 to 4     :  372   1996:4083   Liability
:5009
##  no change     :14323    5 to 8     :   21
##  under 6 months:    4    more than 8:    2
##
##
```

```
model <- rpart(FraudFound_P ~ ., data = train)
rpart.plot(model)

print(model)

## n= 10795
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 10795 647 0 (0.940064845 0.059935155)
##    2) BasePolicy=Liability 3525  25 0 (0.992907801 0.007092199) *
##    3) BasePolicy=All Perils,Collision 7270 622 0 (0.914442916 0.085
557084)
##      6) Fault=Third Party 2439  26 0 (0.989339893 0.010660107)
##       12) Deductible=400,700 2391   0 0 (1.000000000 0.000000000) *
##       13) Deductible=300,500 48  22 1 (0.458333333 0.541666667)
##         26) AgeOfPolicyHolder=26 to 30,41 to 50,51 to 65,over 65 22
0 0 (1.000000000 0.000000000) *
##         27) AgeOfPolicyHolder=31 to 35,36 to 40 26   0 1 (0.0000000
00 1.000000000) *
##      7) Fault=Policy Holder 4831 596 0 (0.876630097 0.123369903) *

#This DT gives majority of the predictions for non fraudulent cases wh
ich is not very helpful
#Hence using weights to have a better distribution

weights <- ifelse(train$FraudFound_P == "0", 1, 16)


#Decision Tree Model: With weights
model_2 <- rpart(FraudFound_P ~ ., data = train,weights = weights)
rpart.plot(model_2)

print(model_2)

## n= 10795
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 10795 10148 1 (0.49502439 0.50497561)
##    2) BasePolicy=Liability 3525   400 0 (0.89743590 0.10256410) *
##    3) BasePolicy=All Perils,Collision 7270  6648 1 (0.40048193 0.59
951807)
##      6) Fault=Third Party 2439   416 0 (0.85295157 0.14704843)
##       12) Deductible=400,700 2391     0 0 (1.00000000 0.00000000) *
```

```
##       13) Deductible=300,500 48    22 1 (0.05022831 0.94977169) *
##      7) Fault=Policy Holder 4831  4235 1 (0.30753032 0.69246968)
##       14) MonthClaimed=Dec,Nov 811   757 1 (0.46699568 0.53300432)
##         28) Year=1994 279    48 0 (0.85185185 0.14814815) *
##         29) Year=1995,1996 532   481 1 (0.37085582 0.62914418) *
##       15) MonthClaimed=Apr,Aug,Feb,Jan,Jul,Jun,Mar,May,Oct,Sep 4020
3478 1 (0.28625514 0.71374486) *

#Decision Tree Model: With Weights and Controls
#To do indepth analysis
model_3 <- rpart(FraudFound_P ~ .,
                 data = train,weights = weights,
                 control = rpart.control(cp=0.001,minbucket = 30,maxde
pth = 5))
rpart.plot(model_3)

print(model_3)

## n= 10795
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 10795 10148 1 (0.49502439 0.50497561)
##    2) BasePolicy=Liability 3525   400 0 (0.89743590 0.10256410)
##      4) RepNumber=1,2,3,4,5,9,12,14,15,16 2227    96 0 (0.95856711
0.04143289) *
##      5) RepNumber=6,7,8,10,11,13 1298   304 0 (0.80795957 0.1920404
3)
##       10) DayOfWeekClaimed=Friday,Monday,Sunday,Thursday,Wednesday
1002   112 0 (0.89882565 0.10117435) *
##       11) DayOfWeekClaimed=Saturday,Tuesday 296   192 0 (0.59663866
0.40336134)
##         22) AgeOfPolicyHolder=16 to 17,31 to 35,51 to 65,over 65 16
9    32 0 (0.83919598 0.16080402) *
##         23) AgeOfPolicyHolder=26 to 30,36 to 40,41 to 50 127   117
1 (0.42238267 0.57761733)
##           46) Make=Honda,Mazda,Pontiac 71    16 0 (0.81395349 0.186
04651) *
##           47) Make=Chevrolet,Others,Toyota 56    47 1 (0.24607330 0
.75392670) *
##    3) BasePolicy=All Perils,Collision 7270  6648 1 (0.40048193 0.59
951807)
##      6) Fault=Third Party 2439   416 0 (0.85295157 0.14704843)
##       12) Deductible=400,700 2391     0 0 (1.00000000 0.00000000) *
##       13) Deductible=300,500 48    22 1 (0.05022831 0.94977169) *
```

```
##      7) Fault=Policy Holder 4831  4235 1 (0.30753032 0.69246968)
##       14) MonthClaimed=Dec,Nov 811   757 1 (0.46699568 0.53300432)
##         28) Year=1994 279    48 0 (0.85185185 0.14814815) *
##         29) Year=1995,1996 532   481 1 (0.37085582 0.62914418)
##           58) RepNumber=1,4,10,12,13,15,16 240   192 0 (0.54285714
0.45714286) *
##           59) RepNumber=2,3,5,6,7,8,9,11,14 292   253 1 (0.28848347
0.71151653) *
##       15) MonthClaimed=Apr,Aug,Feb,Jan,Jul,Jun,Mar,May,Oct,Sep 4020
3478 1 (0.28625514 0.71374486) *
```

```
#Logistic Regression Model
weights <- ifelse(train$FraudFound_P == "0", 1, 16)

LRmodel <- glm(formula=FraudFound_P~.,
               data = train,
               family="binomial",weights =weights)

summary(LRmodel)
```

```
##
## Call:
## glm(formula = FraudFound_P ~ ., family = "binomial", data = train,
##     weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4500  -1.3716  -0.4830  -0.2908   9.2918
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(
>|z|)
## (Intercept)                         6.546940   1.250701   5.235 1.6
5e-07 ***
## MonthAug                           -0.590868   0.171389  -3.448 0.0
00566 ***
## MonthDec                            0.230218   0.162745   1.415 0.1
57187
## MonthFeb                            0.389969   0.139344   2.799 0.0
05132 **
## MonthJan                            0.473956   0.153393   3.090 0.0
02003 **
## MonthJul                           -0.692557   0.155912  -4.442 8.9
1e-06 ***
## MonthJun                            0.039312   0.138278   0.284 0.7
76185
```

```
## MonthMar                          0.432680   0.116761   3.706 0.0
00211 ***
## MonthMay                          0.075076   0.112967   0.665 0.5
06314
## MonthNov                          0.129661   0.169073   0.767 0.4
43145
## MonthOct                          0.419444   0.163676   2.563 0.0
10388 *
## MonthSep                         -0.095729   0.166284  -0.576 0.5
64820
## WeekOfMonth2                      0.169842   0.058311   2.913 0.0
03583 **
## WeekOfMonth3                      0.039439   0.061492   0.641 0.5
21288
## WeekOfMonth4                     -0.047830   0.061632  -0.776 0.4
37711
## WeekOfMonth5                     -0.290430   0.074836  -3.881 0.0
00104 ***
## DayOfWeekMonday                  -0.085472   0.064954  -1.316 0.1
88212
## DayOfWeekSaturday                 0.069518   0.067804   1.025 0.3
05236
## DayOfWeekSunday                   0.015531   0.070958   0.219 0.8
26742
## DayOfWeekThursday                -0.152952   0.068228  -2.242 0.0
24977 *
## DayOfWeekTuesday                 -0.110951   0.066874  -1.659 0.0
97094 .
## DayOfWeekWednesday               -0.220244   0.068475  -3.216 0.0
01298 **
## MakeHonda                        -0.210162   0.076032  -2.764 0.0
05707 **
## MakeMazda                        -0.209648   0.074358  -2.819 0.0
04811 **
## MakeOthers                        0.045950   0.078377   0.586 0.5
57694
## MakePontiac                      -0.224099   0.067840  -3.303 0.0
00955 ***
## MakeToyota                       -0.026966   0.070767  -0.381 0.7
03163
## AccidentAreaUrban                -0.299340   0.055524  -5.391 7.0
0e-08 ***
## DayOfWeekClaimedMonday           -0.211426   0.059763  -3.538 0.0
00404 ***
## DayOfWeekClaimedSaturday          0.529906   0.202455   2.617 0.0
08860 **
```

```
## DayOfWeekClaimedSunday              -0.519727   0.380334  -1.366 0.1
71782
## DayOfWeekClaimedThursday            -0.215246   0.063553  -3.387 0.0
00707 ***
## DayOfWeekClaimedTuesday             -0.172547   0.061515  -2.805 0.0
05032 **
## DayOfWeekClaimedWednesday           -0.203605   0.062468  -3.259 0.0
01117 **
## MonthClaimedAug                      1.055998   0.170317   6.200 5.6
4e-10 ***
## MonthClaimedDec                     -0.438452   0.160930  -2.724 0.0
06440 **
## MonthClaimedFeb                     -0.209265   0.137536  -1.522 0.1
28129
## MonthClaimedJan                     -0.238508   0.150855  -1.581 0.1
13869
## MonthClaimedJul                      0.583495   0.154539   3.776 0.0
00160 ***
## MonthClaimedJun                      0.118948   0.138333   0.860 0.3
89864
## MonthClaimedMar                     -0.097534   0.113453  -0.860 0.3
89961
## MonthClaimedMay                      0.312761   0.118971   2.629 0.0
08566 **
## MonthClaimedNov                     -0.633910   0.166067  -3.817 0.0
00135 ***
## MonthClaimedOct                     -0.102487   0.161488  -0.635 0.5
25663
## MonthClaimedSep                      0.416666   0.166809   2.498 0.0
12494 *
## WeekOfMonthClaimed2                 -0.111331   0.057016  -1.953 0.0
50863 .
## WeekOfMonthClaimed3                 -0.056834   0.061270  -0.928 0.3
53613
## WeekOfMonthClaimed4                  0.070206   0.060029   1.170 0.2
42190
## WeekOfMonthClaimed5                 -0.286609   0.081684  -3.509 0.0
00450 ***
## SexMale                              0.258718   0.055402   4.670 3.0
1e-06 ***
## MaritalStatusSingle                 -0.140857   0.047193  -2.985 0.0
02839 **
## FaultThird Party                    -3.902023   0.096015 -40.640  <
2e-16 ***
## VehicleCategorySport                 0.785776   0.137891   5.699 1.2
1e-08 ***
```

```
## VehicleCategoryUtility              -0.302235   0.111165  -2.719 0.0
06552 **
## VehiclePrice30000 to 39000           0.004380   0.050223   0.087 0.9
30509
## VehiclePrice40000 to 59000           0.521551   0.112773   4.625 3.7
5e-06 ***
## VehiclePrice60000 to 69000           0.154106   0.276425   0.557 0.5
77189
## VehiclePriceless than 20000          0.073103   0.068980   1.060 0.2
89243
## VehiclePricemore than 69000          0.023438   0.072383   0.324 0.7
46083
## RepNumber2                          -0.092220   0.099798  -0.924 0.3
55453
## RepNumber3                           0.002516   0.100009   0.025 0.9
79928
## RepNumber4                          -0.204543   0.100729  -2.031 0.0
42294 *
## RepNumber5                          -0.286981   0.100481  -2.856 0.0
04289 **
## RepNumber6                           0.272655   0.097491   2.797 0.0
05163 **
## RepNumber7                           0.412402   0.096161   4.289 1.8
0e-05 ***
## RepNumber8                          -0.259816   0.103212  -2.517 0.0
11825 *
## RepNumber9                          -0.062142   0.099614  -0.624 0.5
32739
## RepNumber10                         -0.029235   0.099760  -0.293 0.7
69484
## RepNumber11                         -0.151246   0.102177  -1.480 0.1
38809
## RepNumber12                         -0.580735   0.104488  -5.558 2.7
3e-08 ***
## RepNumber13                          0.251735   0.102927   2.446 0.0
14454 *
## RepNumber14                         -0.041361   0.102180  -0.405 0.6
85633
## RepNumber15                         -0.275487   0.101859  -2.705 0.0
06839 **
## RepNumber16                         -0.134981   0.100283  -1.346 0.1
78302
## Deductible400                       -4.905222   0.747249  -6.564 5.2
3e-11 ***
## Deductible500                       -1.647820   0.672169  -2.451 0.0
14226 *
```

```
## Deductible700                        -4.961860   0.756178  -6.562 5.3
2e-11 ***
## DriverRating2                         0.091479   0.052714   1.735 0.0
82676 .
## DriverRating3                         0.216932   0.051554   4.208 2.5
8e-05 ***
## DriverRating4                         0.139197   0.051786   2.688 0.0
07189 **
## Days_Policy_Accident15 to 30         -0.377205   0.833098  -0.453 0.6
50712
## Days_Policy_Accident8 to 15          -0.384434   0.784456  -0.490 0.6
24088
## Days_Policy_Accidentmore than 30      0.773185   0.736251   1.050 0.2
93642
## Days_Policy_Accidentnone              1.415127   0.737849   1.918 0.0
55123 .
## Days_Policy_Claim8 to 15             -0.403956   0.629236  -0.642 0.5
20888
## Days_Policy_Claimmore than 30        -1.631239   0.393564  -4.145 3.4
0e-05 ***
## PastNumberOfClaims2 to 4              0.174935   0.049264   3.551 0.0
00384 ***
## PastNumberOfClaimsmore than 4         0.140126   0.072320   1.938 0.0
52675 .
## PastNumberOfClaimsnone                0.135795   0.049286   2.755 0.0
05864 **
## AgeOfVehicle3 years                   1.163933   0.377597   3.082 0.0
02053 **
## AgeOfVehicle4 years                   2.357872   0.390160   6.043 1.5
1e-09 ***
## AgeOfVehicle5 years                   1.654074   0.378444   4.371 1.2
4e-05 ***
## AgeOfVehicle6 years                   1.553707   0.378779   4.102 4.1
0e-05 ***
## AgeOfVehicle7 years                   1.288567   0.378989   3.400 0.0
00674 ***
## AgeOfVehiclemore than 7               1.246295   0.381295   3.269 0.0
01081 **
## AgeOfVehiclenew                       0.441287   0.544729   0.810 0.4
17880
## AgeOfPolicyHolder18 to 20            -1.217985   0.763588  -1.595 0.1
10694
## AgeOfPolicyHolder21 to 25            -0.706799   0.476503  -1.483 0.1
37994
## AgeOfPolicyHolder26 to 30            -1.500929   0.414830  -3.618 0.0
00297 ***
```

```
## AgeOfPolicyHolder31 to 35          -1.259418   0.401444  -3.137 0.0
01706 **
## AgeOfPolicyHolder36 to 40          -1.249414   0.404705  -3.087 0.0
02020 **
## AgeOfPolicyHolder41 to 50          -1.391360   0.406120  -3.426 0.0
00613 ***
## AgeOfPolicyHolder51 to 65          -1.546501   0.408285  -3.788 0.0
00152 ***
## AgeOfPolicyHolderover 65           -1.597592   0.417065  -3.831 0.0
00128 ***
## PoliceReportFiledYes               -0.322691   0.115696  -2.789 0.0
05285 **
## WitnessPresentYes                  -0.096299   0.257036  -0.375 0.7
07921
## AgentTypeInternal                  -1.459442   0.238716  -6.114 9.7
3e-10 ***
## NumberOfSuppliments3 to 5          -0.249189   0.068017  -3.664 0.0
00249 ***
## NumberOfSupplimentsmore than 5      0.025731   0.058483   0.440 0.6
59957
## NumberOfSupplimentsnone             0.021382   0.052603   0.406 0.6
84385
## AddressChange_Claim2 to 3 years     0.792084   0.425293   1.862 0.0
62541 .
## AddressChange_Claim4 to 8 years    -0.589250   0.265459  -2.220 0.0
26436 *
## AddressChange_Claimno change        0.286863   0.201991   1.420 0.1
55557
## AddressChange_Claimunder 6 months   3.219174   1.052685   3.058 0.0
02228 **
## NumberOfCars2 vehicles              0.653445   0.243270   2.686 0.0
07229 **
## NumberOfCars3 to 4                  0.038881   0.120084   0.324 0.7
46105
## NumberOfCars5 to 8                  0.877411   0.522929   1.678 0.0
93371 .
## NumberOfCarsmore than 8           -11.975700 119.468182  -0.100 0.9
20152
## Year1995                            0.077091   0.045489   1.695 0.0
90125 .
## Year1996                           -0.013316   0.048436  -0.275 0.7
83380
## BasePolicyCollision                -0.480863   0.043160 -11.142  <
2e-16 ***
## BasePolicyLiability                -4.046892   0.155711 -25.990  <
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: 28417  on 10794  degrees of freedom
## Residual deviance: 19337  on 10673  degrees of freedom
## AIC: 19581
##
## Number of Fisher Scoring iterations: 9
```

```
#Stepwise Regression
SRmodel <- step(LRmodel)
```

```
## Start:  AIC=19581.35
## FraudFound_P ~ Month + WeekOfMonth + DayOfWeek + Make + AccidentAre
a +
##     DayOfWeekClaimed + MonthClaimed + WeekOfMonthClaimed + Sex +
##     MaritalStatus + Fault + VehicleCategory + VehiclePrice +
##     RepNumber + Deductible + DriverRating + Days_Policy_Accident +
##     Days_Policy_Claim + PastNumberOfClaims + AgeOfVehicle + AgeOfPo
licyHolder +
##     PoliceReportFiled + WitnessPresent + AgentType + NumberOfSuppli
ments +
##     AddressChange_Claim + NumberOfCars + Year + BasePolicy
##
##                         Df Deviance   AIC
## - WitnessPresent         1    19338 19580
## <none>                        19337 19581
## - Year                   2    19341 19581
## - NumberOfCars           4    19351 19587
## - PoliceReportFiled      1    19345 19587
## - MaritalStatus          1    19346 19588
## - PastNumberOfClaims     3    19351 19589
## - DriverRating           3    19356 19594
## - DayOfWeek              6    19362 19594
## - VehiclePrice           5    19361 19595
## - Days_Policy_Claim      2    19356 19596
## - WeekOfMonthClaimed     4    19362 19598
## - Days_Policy_Accident   4    19362 19598
## - NumberOfSuppliments    3    19361 19599
## - DayOfWeekClaimed       6    19368 19600
## - Sex                    1    19359 19601
## - Make                   5    19369 19603
## - AccidentArea           1    19367 19609
## - AgeOfPolicyHolder      8    19381 19609
```

```
## - AddressChange_Claim   4    19373 19609
## - WeekOfMonth           4    19374 19610
## - AgentType             1    19380 19622
## - VehicleCategory       2    19386 19626
## - Deductible            3    19405 19643
## - Month                11    19429 19651
## - MonthClaimed         11    19445 19667
## - AgeOfVehicle          7    19442 19672
## - RepNumber            15    19507 19721
## - BasePolicy            2    20120 20360
## - Fault                 1    23277 23519
##
## Step:  AIC=19579.49
## FraudFound_P ~ Month + WeekOfMonth + DayOfWeek + Make + AccidentAre
a +
##     DayOfWeekClaimed + MonthClaimed + WeekOfMonthClaimed + Sex +
##     MaritalStatus + Fault + VehicleCategory + VehiclePrice +
##     RepNumber + Deductible + DriverRating + Days_Policy_Accident +
##     Days_Policy_Claim + PastNumberOfClaims + AgeOfVehicle + AgeOfPo
licyHolder +
##     PoliceReportFiled + AgentType + NumberOfSuppliments + AddressCh
ange_Claim +
##     NumberOfCars + Year + BasePolicy
##
##                        Df Deviance   AIC
## <none>                      19338 19580
## - Year                  2    19342 19580
## - NumberOfCars          4    19351 19585
## - PoliceReportFiled     1    19346 19586
## - MaritalStatus         1    19346 19586
## - PastNumberOfClaims    3    19351 19587
## - DriverRating          3    19356 19592
## - DayOfWeek             6    19363 19593
## - VehiclePrice          5    19361 19593
## - Days_Policy_Claim     2    19357 19595
## - WeekOfMonthClaimed    4    19362 19596
## - Days_Policy_Accident  4    19363 19597
## - NumberOfSuppliments   3    19361 19597
## - DayOfWeekClaimed      6    19368 19598
## - Sex                   1    19359 19599
## - Make                  5    19369 19601
## - AccidentArea          1    19367 19607
## - AgeOfPolicyHolder     8    19381 19607
## - AddressChange_Claim   4    19374 19608
## - WeekOfMonth           4    19375 19609
## - AgentType             1    19380 19620
```

```
## - VehicleCategory       2    19386 19624
## - Deductible            3    19405 19641
## - Month                11    19429 19649
## - MonthClaimed         11    19445 19665
## - AgeOfVehicle          7    19442 19670
## - RepNumber            15    19507 19719
## - BasePolicy            2    20120 20358
## - Fault                 1    23283 23523

summary(SRmodel)

##
## Call:
## glm(formula = FraudFound_P ~ Month + WeekOfMonth + DayOfWeek +
##     Make + AccidentArea + DayOfWeekClaimed + MonthClaimed + WeekOfM
onthClaimed +
##     Sex + MaritalStatus + Fault + VehicleCategory + VehiclePrice +
##     RepNumber + Deductible + DriverRating + Days_Policy_Accident +
##     Days_Policy_Claim + PastNumberOfClaims + AgeOfVehicle + AgeOfPo
licyHolder +
##     PoliceReportFiled + AgentType + NumberOfSuppliments + AddressCh
ange_Claim +
##     NumberOfCars + Year + BasePolicy, family = "binomial", data = t
rain,
##     weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4504  -1.3713  -0.4831  -0.2908   9.2913
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(
>|z|)
## (Intercept)                      6.552459   1.250242   5.241 1.6
0e-07 ***
## MonthAug                        -0.592508   0.171332  -3.458 0.0
00544 ***
## MonthDec                         0.230202   0.162756   1.414 0.1
57244
## MonthFeb                         0.388013   0.139240   2.787 0.0
05325 **
## MonthJan                         0.473897   0.153406   3.089 0.0
02007 **
## MonthJul                        -0.694626   0.155813  -4.458 8.2
7e-06 ***
## MonthJun                         0.038552   0.138265   0.279 0.7
```

```
80377
## MonthMar                          0.430273   0.116599   3.690 0.0
00224 ***
## MonthMay                          0.074969   0.112968   0.664 0.5
06926
## MonthNov                          0.129301   0.169074   0.765 0.4
44416
## MonthOct                          0.418286   0.163647   2.556 0.0
10588 *
## MonthSep                         -0.096761   0.166261  -0.582 0.5
60579
## WeekOfMonth2                      0.169115   0.058278   2.902 0.0
03709 **
## WeekOfMonth3                      0.039063   0.061482   0.635 0.5
25193
## WeekOfMonth4                     -0.048324   0.061614  -0.784 0.4
32864
## WeekOfMonth5                     -0.289920   0.074822  -3.875 0.0
00107 ***
## DayOfWeekMonday                  -0.084886   0.064934  -1.307 0.1
91126
## DayOfWeekSaturday                 0.069662   0.067801   1.027 0.3
04209
## DayOfWeekSunday                   0.015765   0.070960   0.222 0.8
24184
## DayOfWeekThursday                -0.152413   0.068211  -2.234 0.0
25454 *
## DayOfWeekTuesday                 -0.110497   0.066863  -1.653 0.0
98416 .
## DayOfWeekWednesday               -0.219981   0.068472  -3.213 0.0
01315 **
## MakeHonda                        -0.209459   0.076013  -2.756 0.0
05859 **
## MakeMazda                        -0.208564   0.074307  -2.807 0.0
05003 **
## MakeOthers                        0.046359   0.078377   0.591 0.5
54189
## MakePontiac                      -0.224334   0.067846  -3.307 0.0
00945 ***
## MakeToyota                       -0.026239   0.070745  -0.371 0.7
10718
## AccidentAreaUrban                -0.299487   0.055516  -5.395 6.8
7e-08 ***
## DayOfWeekClaimedMonday           -0.211405   0.059764  -3.537 0.0
00404 ***
## DayOfWeekClaimedSaturday          0.528450   0.202365   2.611 0.0
```

```
09018 **
## DayOfWeekClaimedSunday             -0.519822   0.380314  -1.367 0.1
71681
## DayOfWeekClaimedThursday           -0.214815   0.063544  -3.381 0.0
00723 ***
## DayOfWeekClaimedTuesday            -0.172960   0.061507  -2.812 0.0
04923 **
## DayOfWeekClaimedWednesday          -0.203504   0.062468  -3.258 0.0
01123 **
## MonthClaimedAug                     1.057884   0.170240   6.214 5.1
6e-10 ***
## MonthClaimedDec                    -0.438186   0.160937  -2.723 0.0
06475 **
## MonthClaimedFeb                    -0.209251   0.137539  -1.521 0.1
28160
## MonthClaimedJan                    -0.238491   0.150868  -1.581 0.1
13926
## MonthClaimedJul                     0.583792   0.154533   3.778 0.0
00158 ***
## MonthClaimedJun                     0.119691   0.138320   0.865 0.3
86861
## MonthClaimedMar                    -0.095209   0.113302  -0.840 0.4
00733
## MonthClaimedMay                     0.312862   0.118973   2.630 0.0
08546 **
## MonthClaimedNov                    -0.633390   0.166064  -3.814 0.0
00137 ***
## MonthClaimedOct                    -0.101461   0.161463  -0.628 0.5
29749
## MonthClaimedSep                     0.417933   0.166772   2.506 0.0
12210 *
## WeekOfMonthClaimed2                -0.111525   0.057011  -1.956 0.0
50443 .
## WeekOfMonthClaimed3                -0.056543   0.061262  -0.923 0.3
56021
## WeekOfMonthClaimed4                 0.070597   0.060018   1.176 0.2
39489
## WeekOfMonthClaimed5                -0.289382   0.081353  -3.557 0.0
00375 ***
## SexMale                             0.258368   0.055394   4.664 3.1
0e-06 ***
## MaritalStatusSingle                -0.141055   0.047189  -2.989 0.0
02798 **
## FaultThird Party                   -3.902753   0.095998 -40.654  <
2e-16 ***
## VehicleCategorySport                0.785209   0.137863   5.696 1.2
```

```
3e-08 ***
## VehicleCategoryUtility             -0.301661   0.111153  -2.714 0.0
06649 **
## VehiclePrice30000 to 39000          0.003411   0.050158   0.068 0.9
45785
## VehiclePrice40000 to 59000          0.521273   0.112766   4.623 3.7
9e-06 ***
## VehiclePrice60000 to 69000          0.154585   0.276406   0.559 0.5
75978
## VehiclePriceless than 20000         0.073605   0.068968   1.067 0.2
85867
## VehiclePricemore than 69000         0.023279   0.072378   0.322 0.7
47730
## RepNumber2                         -0.092611   0.099794  -0.928 0.3
53398
## RepNumber3                          0.001515   0.099989   0.015 0.9
87915
## RepNumber4                         -0.204892   0.100724  -2.034 0.0
41931 *
## RepNumber5                         -0.287045   0.100480  -2.857 0.0
04280 **
## RepNumber6                          0.271499   0.097443   2.786 0.0
05332 **
## RepNumber7                          0.411942   0.096150   4.284 1.8
3e-05 ***
## RepNumber8                         -0.259948   0.103209  -2.519 0.0
11781 *
## RepNumber9                         -0.062038   0.099611  -0.623 0.5
33409
## RepNumber10                        -0.029320   0.099760  -0.294 0.7
68831
## RepNumber11                        -0.151411   0.102173  -1.482 0.1
38364
## RepNumber12                        -0.580891   0.104484  -5.560 2.7
0e-08 ***
## RepNumber13                         0.249566   0.102754   2.429 0.0
15150 *
## RepNumber14                        -0.041217   0.102179  -0.403 0.6
86666
## RepNumber15                        -0.275727   0.101859  -2.707 0.0
06791 **
## RepNumber16                        -0.135444   0.100274  -1.351 0.1
76779
## Deductible400                      -4.905750   0.747257  -6.565 5.2
0e-11 ***
## Deductible500                      -1.647517   0.672184  -2.451 0.0
```

```
14246 *
## Deductible700                       -4.962747   0.756184  -6.563 5.2
8e-11 ***
## DriverRating2                        0.091515   0.052713   1.736 0.0
82546 .
## DriverRating3                        0.215975   0.051490   4.195 2.7
3e-05 ***
## DriverRating4                        0.139171   0.051785   2.687 0.0
07200 **
## Days_Policy_Accident15 to 30        -0.373913   0.832280  -0.449 0.6
53242
## Days_Policy_Accident8 to 15         -0.382158   0.783691  -0.488 0.6
25806
## Days_Policy_Accidentmore than 30     0.776479   0.735310   1.056 0.2
90973
## Days_Policy_Accidentnone             1.417934   0.737024   1.924 0.0
54372 .
## Days_Policy_Claim8 to 15            -0.406627   0.629362  -0.646 0.5
18219
## Days_Policy_Claimmore than 30       -1.634885   0.393631  -4.153 3.2
8e-05 ***
## PastNumberOfClaims2 to 4             0.175319   0.049254   3.560 0.0
00372 ***
## PastNumberOfClaimsmore than 4        0.140924   0.072287   1.950 0.0
51234 .
## PastNumberOfClaimsnone               0.135980   0.049282   2.759 0.0
05794 **
## AgeOfVehicle3 years                  1.160452   0.377373   3.075 0.0
02104 **
## AgeOfVehicle4 years                  2.353969   0.389905   6.037 1.5
7e-09 ***
## AgeOfVehicle5 years                  1.649916   0.378159   4.363 1.2
8e-05 ***
## AgeOfVehicle6 years                  1.549991   0.378527   4.095 4.2
3e-05 ***
## AgeOfVehicle7 years                  1.284629   0.378722   3.392 0.0
00694 ***
## AgeOfVehiclemore than 7              1.243099   0.381079   3.262 0.0
01106 **
## AgeOfVehiclenew                      0.436942   0.544525   0.802 0.4
22306
## AgeOfPolicyHolder18 to 20           -1.220358   0.763557  -1.598 0.1
09987
## AgeOfPolicyHolder21 to 25           -0.707899   0.476497  -1.486 0.1
37376
## AgeOfPolicyHolder26 to 30           -1.503823   0.414756  -3.626 0.0
```

```
00288 ***
## AgeOfPolicyHolder31 to 35          -1.259992   0.401441  -3.139 0.0
01697 **
## AgeOfPolicyHolder36 to 40          -1.250938   0.404685  -3.091 0.0
01994 **
## AgeOfPolicyHolder41 to 50          -1.392522   0.406108  -3.429 0.0
00606 ***
## AgeOfPolicyHolder51 to 65          -1.547735   0.408272  -3.791 0.0
00150 ***
## AgeOfPolicyHolderover 65           -1.598567   0.417056  -3.833 0.0
00127 ***
## PoliceReportFiledYes               -0.329959   0.114019  -2.894 0.0
03805 **
## AgentTypeInternal                  -1.460605   0.238652  -6.120 9.3
4e-10 ***
## NumberOfSuppliments3 to 5          -0.248245   0.067969  -3.652 0.0
00260 ***
## NumberOfSupplimentsmore than 5      0.026803   0.058411   0.459 0.6
46329
## NumberOfSupplimentsnone             0.022196   0.052558   0.422 0.6
72803
## AddressChange_Claim2 to 3 years     0.791947   0.425286   1.862 0.0
62582 .
## AddressChange_Claim4 to 8 years    -0.589313   0.265451  -2.220 0.0
26416 *
## AddressChange_Claimno change        0.286438   0.201996   1.418 0.1
56179
## AddressChange_Claimunder 6 months   3.219067   1.052733   3.058 0.0
02230 **
## NumberOfCars2 vehicles              0.653320   0.243283   2.685 0.0
07244 **
## NumberOfCars3 to 4                  0.039164   0.120084   0.326 0.7
44318
## NumberOfCars5 to 8                  0.878903   0.522890   1.681 0.0
92791 .
## NumberOfCarsmore than 8           -11.973771 119.468181  -0.100 0.9
20165
## Year1995                            0.077773   0.045450   1.711 0.0
87046 .
## Year1996                           -0.012503   0.048388  -0.258 0.7
96105
## BasePolicyCollision                -0.480823   0.043161 -11.140  <
2e-16 ***
## BasePolicyLiability                -4.045919   0.155675 -25.990  <
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: 28417  on 10794  degrees of freedom
## Residual deviance: 19337  on 10674  degrees of freedom
## AIC: 19579
##
## Number of Fisher Scoring iterations: 9
```

```
##PREDICTING PROBABILITIES, ROC, PR, AND CONFUSION MATRIX FOR EACH MOD
EL

# Predict probabilities for training data Logistic Regression
trainpreds_1 <- LRmodel$fitted.values

# Predict probabilities for test data Logistic Regression
testpreds_1 <- predict(LRmodel, newdata = test, type = "response")

#Roc Curve Train Logistic Regression
library(PRROC)
```

```
## Warning: package 'PRROC' was built under R version 4.2.3
```

```
roc_1 <- roc.curve(scores.class0 = LRmodel$fitted.values,
                   weights.class0 = as.numeric(as.character(train$Frau
dFound_P)),
                   curve = T)
print(roc_1)
```

```
##
##   ROC curve
##
##     Area under curve:
##      0.835959
##
##     Curve for scores from  2.577634e-05  to  0.9977226
##     ( can be plotted with plot(x) )
```

```
plot(roc_1)

#PR Curve Train Logistic Regression
prcurve_1 <- pr.curve(scores.class0 = LRmodel$fitted.values,
                      weights.class0 = as.numeric(as.character(train$F
raudFound_P)),
                      curve = T)
print(prcurve_1)
```

```
##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.190029
##
##     Area under curve (Davis & Goadrich):
##      0.1900153
##
##     Curve for scores from  2.577634e-05  to  0.9977226
##     ( can be plotted with plot(x) )
```

```
plot(prcurve_1)
```

```
#ROC Curve Test Logistic Regression
roctest_1 <- roc.curve(testpreds_1,
                       weights.class0 = as.numeric(as.character(test$F
raudFound_P)),
                       curve = T)
print(roctest_1)
```

```
##
##   ROC curve
##
##     Area under curve:
##      0.7961698
##
##     Curve for scores from  8.123895e-08  to  0.9983899
##     ( can be plotted with plot(x) )
```

```
plot(roctest_1)
```

```
#PR Curve Test Logistic Regression
prcurvetest_1 <- pr.curve(testpreds_1,
                          weights.class0 = as.numeric(as.character(tes
t$FraudFound_P)),
                          curve = T)
print(prcurvetest_1)
```

```
##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.1409983
##
##     Area under curve (Davis & Goadrich):
##      0.1409851
```

```
##
##     Curve for scores from  8.123895e-08  to  0.9983899
##     ( can be plotted with plot(x) )

plot(prcurvetest_1)

#Confusion Matrix on Train Logistic Regression
pred_train_p1 <-
  predict(LRmodel, newdata = train, type = "response")

trainpred_c1 <- ifelse(pred_train_p1 > 0.5, 1, 0)

confusion_matrix_train <- table(trainpred_c1, train$FraudFound_P)

confusionMatrix(confusion_matrix_train)

## Confusion Matrix and Statistics
##
##
## trainpred_c1    0    1
##            0 6489   62
##            1 3659  585
##
##                Accuracy : 0.6553
##                  95% CI : (0.6463, 0.6643)
##     No Information Rate : 0.9401
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1509
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6394
##             Specificity : 0.9042
##          Pos Pred Value : 0.9905
##          Neg Pred Value : 0.1378
##              Prevalence : 0.9401
##          Detection Rate : 0.6011
##    Detection Prevalence : 0.6069
##       Balanced Accuracy : 0.7718
##
##        'Positive' Class : 0
##

#Confusion Matrix on Test Logistic Regression
pred_test_p1 <-
  predict(LRmodel, newdata = test, type = "response")
```

```r
testpred_c1 <- ifelse(pred_test_p1 > 0.5, 1, 0)

confusion_matrix_test <- table(testpred_c1, test$FraudFound_P)

confusionMatrix(confusion_matrix_test)
```

```
## Confusion Matrix and Statistics
##
##
## testpred_c1    0    1
##           0 2720   33
##           1 1628  243
##
##                Accuracy : 0.6408
##                  95% CI : (0.6268, 0.6546)
##     No Information Rate : 0.9403
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1365
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6256
##             Specificity : 0.8804
##          Pos Pred Value : 0.9880
##          Neg Pred Value : 0.1299
##              Prevalence : 0.9403
##          Detection Rate : 0.5882
##    Detection Prevalence : 0.5954
##       Balanced Accuracy : 0.7530
##
##        'Positive' Class : 0
##
```

```r
# Predict probabilities for training data DT Model 2
trainpreds_2 <- predict(model_2, newdata = train, type = "prob")

# Predict probabilities for test data DT Model 2
testpreds_2 <- predict(model_2, newdata = test, type = "prob")

#Roc Curve Train DT Model 2
library("PRROC")

roc_2 <- roc.curve(trainpreds_2[,2], #predicted probabilities
                   weights.class0 = as.numeric(as.character(train$Frau
dFound_P)), #actual flag,
```

```
                     curve = T)
print(roc_2)
```

```
##
##   ROC curve
##
##     Area under curve:
##      0.8007776
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )
```

```
plot(roc_2)
```

```
#PR Curve Train DT Model 2
prcurve_2 <- pr.curve(scores.class0 = trainpreds_2[,2],
                      weights.class0 = as.numeric(as.character(train$F
raudFound_P)),
                      curve = T)
print(prcurve_2)
```

```
##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.1679605
##
##     Area under curve (Davis & Goadrich):
##      0.1679689
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )
```

```
plot(prcurve_2)
```

```
#ROC Curve Test DT Model 2
roctest_2 <- roc.curve(testpreds_2[,2],
                       weights.class0 = as.numeric(as.character(test$F
raudFound_P)),
                       curve = T)
print(roctest_2)
```

```
##
##   ROC curve
##
##     Area under curve:
##      0.7899672
##
```

```
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(roctest_2)

#PR Curve Test DT Model 2
prcurvetest_2 <- pr.curve(scores.class0 = testpreds_2[,2],
                          weights.class0 = as.numeric(as.character(tes
t$FraudFound_P)),
                          curve = T)
print(prcurvetest_2)

##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.1660788
##
##     Area under curve (Davis & Goadrich):
##      0.1661989
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(prcurvetest_2)

#Confusion Matrix on Train Model 2
trainpred_c2 <- predict(model_2, newdata = train, type = "class")
pred_train_p2 <-
  predict(model_2, newdata = train, type = "prob")[, 2]

confusionMatrix(trainpred_c2,
                train$FraudFound_P,
                positive = "1",
                mode = "prec_recall")

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6167   28
##          1 3981  619
##
##                Accuracy : 0.6286
##                  95% CI : (0.6194, 0.6377)
##     No Information Rate : 0.9401
##     P-Value [Acc > NIR] : 1
```

```
##
##                  Kappa : 0.1462
##
##  Mcnemar's Test P-Value : <2e-16
##
##               Precision : 0.13457
##                  Recall : 0.95672
##                      F1 : 0.23594
##              Prevalence : 0.05994
##          Detection Rate : 0.05734
##    Detection Prevalence : 0.42612
##       Balanced Accuracy : 0.78221
##
##        'Positive' Class : 1
##
```

```
#Confusion Matrix on Test Model 2
testpred_c2 <- predict(model_2, newdata = test, type = "class")
pred_test_p2 <-
  predict(model_2, newdata = test, type = "prob")[, 2]

confusionMatrix(testpred_c2,
                test$FraudFound_P,
                positive = "1",
                mode = "prec_recall")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2608   15
##          1 1740  261
##
##                Accuracy : 0.6205
##                  95% CI : (0.6063, 0.6345)
##     No Information Rate : 0.9403
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1389
##
##  Mcnemar's Test P-Value : <2e-16
##
##               Precision : 0.13043
##                  Recall : 0.94565
##                      F1 : 0.22925
##              Prevalence : 0.05969
```

```
##          Detection Rate : 0.05644
##    Detection Prevalence : 0.43274
##       Balanced Accuracy : 0.77273
##
##        'Positive' Class : 1
##

#In depth Analysis Predictions
#Decision Tree with Controls
# Predict probabilities for training data DT Model 3
trainpreds_3 <- predict(model_3, newdata = train, type = "prob")

# Predict probabilities for test data DT Model 3
testpreds_3 <- predict(model_3, newdata = test, type = "prob")

#Roc Curve Train DT Model 3
library("PRROC")

roc_3 <- roc.curve(trainpreds_3[,2], #predicted probabilities
                   weights.class0 = as.numeric(as.character(train$Frau
dFound_P)), #actual flag,
                   curve = T)
print(roc_3)

##
##   ROC curve
##
##     Area under curve:
##      0.8063209
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(roc_3)

#PR Curve Train DT Model 3
prcurve_3 <- pr.curve(scores.class0 = trainpreds_3[,2],
                      weights.class0 = as.numeric(as.character(train$F
raudFound_P)),
                      curve = T)
print(prcurve_3)

##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.1707505
```

```
##
##     Area under curve (Davis & Goadrich):
##      0.1707579
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(prcurve_3)

#ROC Curve Test DT Model 3
roctest_3 <- roc.curve(testpreds_3[,2],
                       weights.class0 = as.numeric(as.character(test$F
raudFound_P)),
                       curve = T)
print(roctest_3)

##
##   ROC curve
##
##     Area under curve:
##      0.7862315
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(roctest_3)

#PR Curve Test DT Model 3
prcurvetest_3 <- pr.curve(scores.class0 = testpreds_3[,2],
                          weights.class0 = as.numeric(as.character(tes
t$FraudFound_P)),
                          curve = T)
print(prcurvetest_3)

##
##   Precision-recall curve
##
##     Area under curve (Integral):
##      0.1578886
##
##     Area under curve (Davis & Goadrich):
##      0.1582565
##
##     Curve for scores from  0  to  0.9497717
##     ( can be plotted with plot(x) )

plot(prcurvetest_3)
```

```r
#Confusion Matrix on Train DT Model 3
trainpred_c3 <- predict(model_3, newdata = train, type = "class")
pred_train_p3 <-
  predict(model_3, newdata = train, type = "prob")[, 2]

confusionMatrix(trainpred_c3,
                train$FraudFound_P,
                positive = "1",
                mode = "prec_recall")

## Confusion Matrix and Statistics
##
##            Reference
## Prediction    0    1
##          0 6348   31
##          1 3800  616
##
##                Accuracy : 0.6451
##                  95% CI : (0.636, 0.6541)
##     No Information Rate : 0.9401
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.155
##
##  Mcnemar's Test P-Value : <2e-16
##
##               Precision : 0.13949
##                  Recall : 0.95209
##                      F1 : 0.24333
##              Prevalence : 0.05994
##          Detection Rate : 0.05706
##    Detection Prevalence : 0.40908
##       Balanced Accuracy : 0.78881
##
##        'Positive' Class : 1
##

#Confusion Matrix on Test DT Model 3
testpred_c3 <- predict(model_3, newdata = test, type = "class")
pred_test_p3 <-
  predict(model_3, newdata = test, type = "prob")[, 2]

confusionMatrix(testpred_c3,
                test$FraudFound_P,
                positive = "1",
                mode = "prec_recall")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2661   22
##          1 1687  254
##
##                Accuracy : 0.6304
##                  95% CI : (0.6163, 0.6443)
##     No Information Rate : 0.9403
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1392
##
##  Mcnemar's Test P-Value : <2e-16
##
##               Precision : 0.13086
##                  Recall : 0.92029
##                      F1 : 0.22914
##              Prevalence : 0.05969
##          Detection Rate : 0.05493
##    Detection Prevalence : 0.41977
##       Balanced Accuracy : 0.76615
##
##        'Positive' Class : 1
##

write.csv(df, file = "Aamd Final.csv")
```

Contribution Sheet

Name	Contribution
Rida Fatima	Data exploration in R Data cleaning and feature engineering in R Data splitting Models training (Logistic Regression and Decision Trees)
Maaz Siddiqui	Predictions ROC PR Plots Confusion Matrices Comparison of Different Models Appendices
Muhammad Atif	Problem Identification Data Exploration Models Comparison and Analysis Policy Recommendations Limitations
Dua Mahboob	Initial Data Exploration in Excel Executive Summary, Introduction, Data Exploration, Data Transformation, Model Training detailed in the report Models Comparison Proofreading, Editing, Formatting