# INSURANCE CLAIMS

- PRACHI (G12)

Table of Contents

# Introduction

## Problem statement

- Insurance industry is bleeding with losses with inside and outside of the system
- This has prompted need to set up a new anti-fraud department whose job is to identify the risk and loss to this frauds and to find ways to reduce fraudulent claims
- The department has to build a prediction model to find the claims which should be rejected but closed/Accepted in advance to help reduce frauds
- The department has to classify the claims based on their severity and would require more focus and analysis

## Need for the project

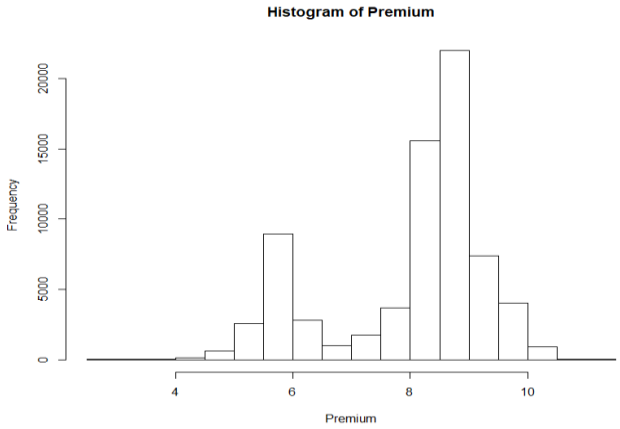
- India has a huge market for insurance and the number of frauds are increasing linearly
- Insurance sector alone results in 40,000 crores loss which is close to 8.5% revenue of insurance in India
- Anti-fraud department would eventually help in reducing the risks and losses of the insurance sector and also tighten the disbursement and payment security of the industry

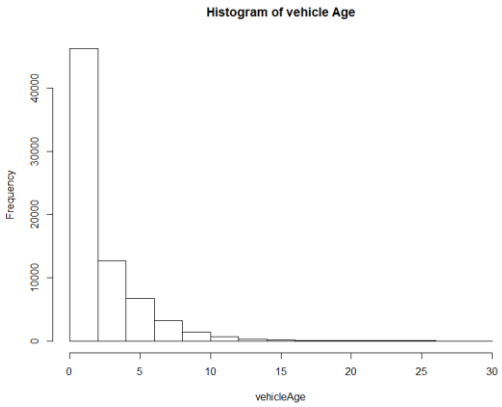
## Understanding business/Social Opportunity

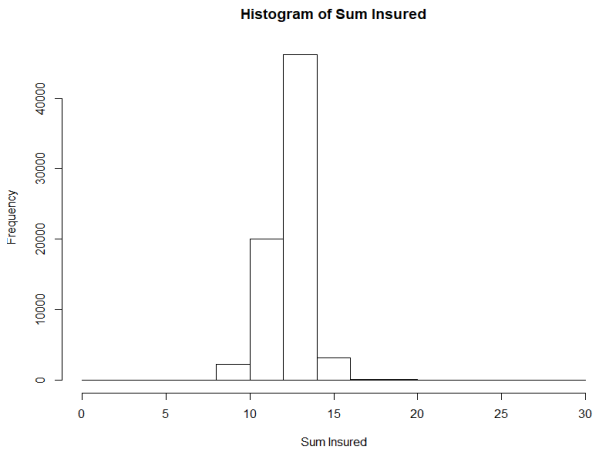
- This will have an impact in increasing the revenue of the insurance companies in India
- This will also help in preventing the money loss of innocent people

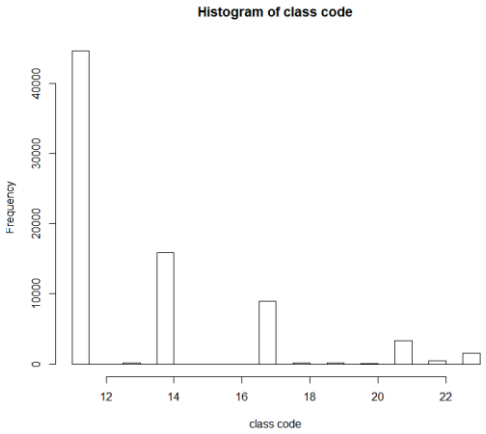
# Exploratory Data Analysis

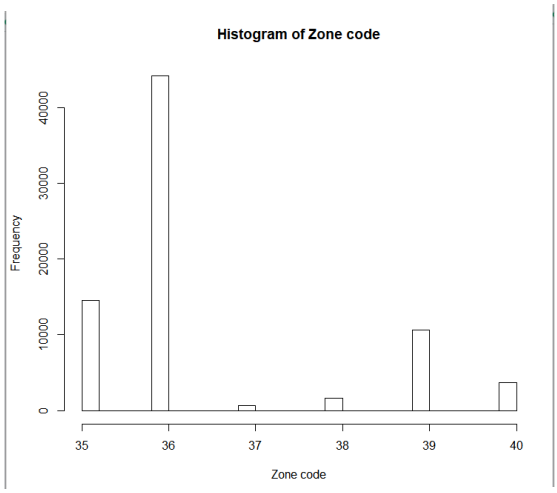
## Univariate Analysis

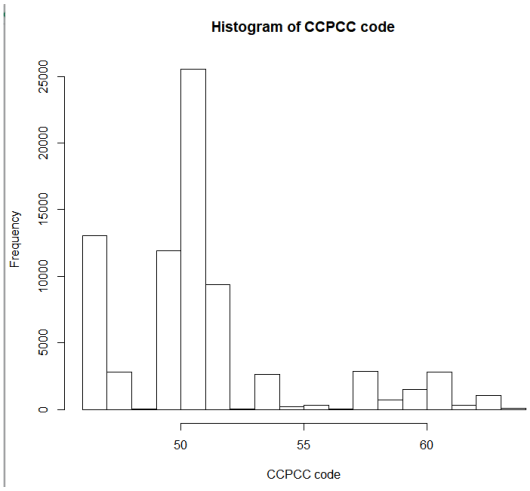


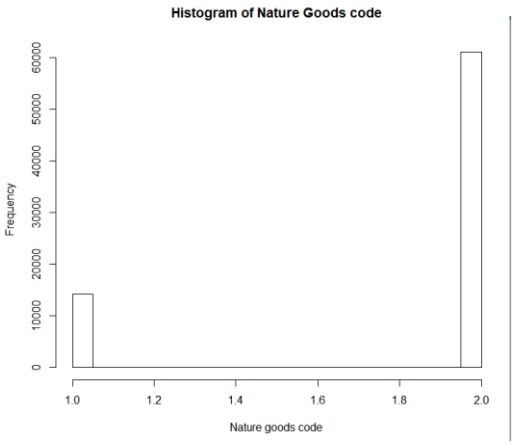



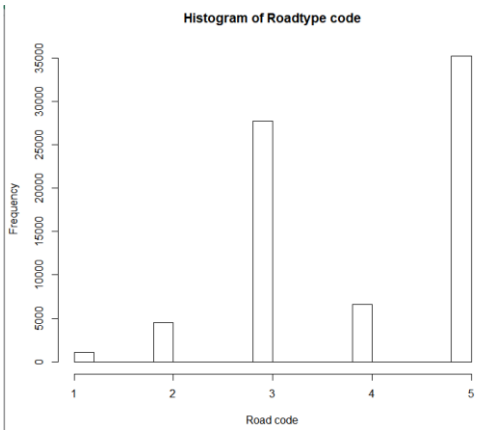



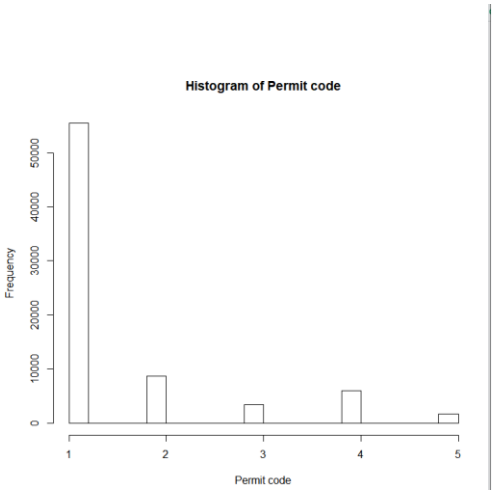



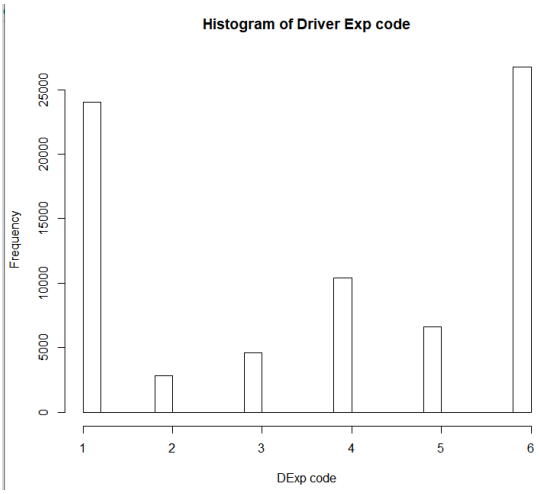

Histogram of Zone code
Frequency
0
10000
20000
30000
40000
35
36
37
38
39
40
Zone code

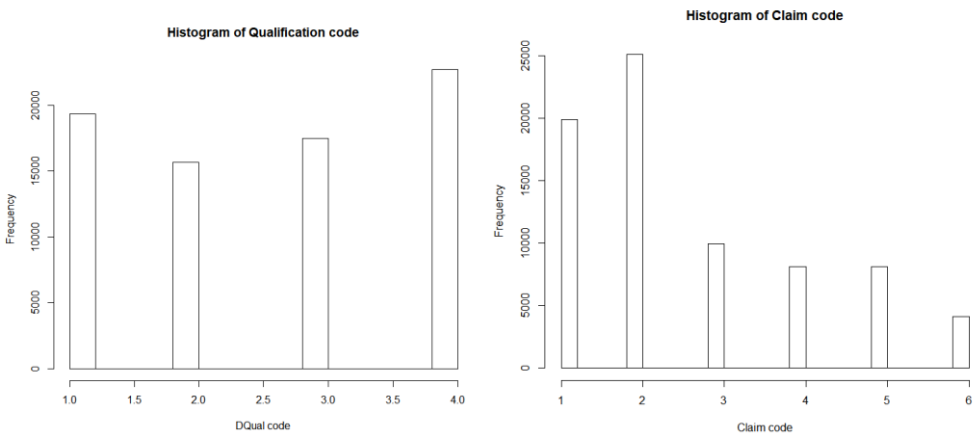

Histogram of CCPCC code
Frequency
0
5000
10000
15000
20000
25000
50
55
60
CCPCC code


Histogram of Nature Goods code
Frequency
0
10000
20000
30000
40000
50000
60000
1.0
1.2
1.4
1.6
1.8
2.0
Nature goods code


Histogram of Roadtype code
Frequency
0
5000
10000
15000
20000
25000
30000
35000
1
2
3
4
5
Road code

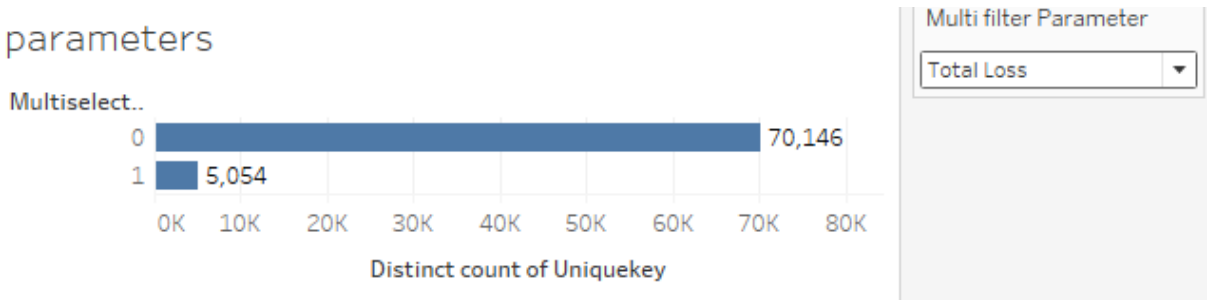

Histogram of Permit code
Frequency
0
10000
20000
30000
40000
50000
1
2
3
4
5
Permit code

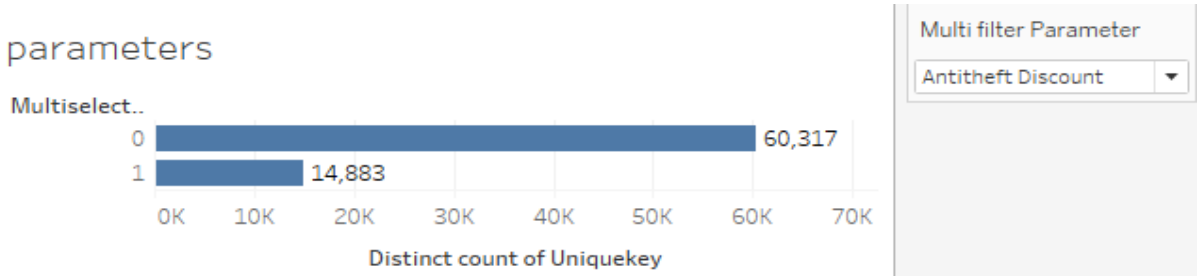

Histogram of Driver Exp code
Frequency
0
5000
10000
15000
20000
25000
1
2
3
4
5
6
DExp code

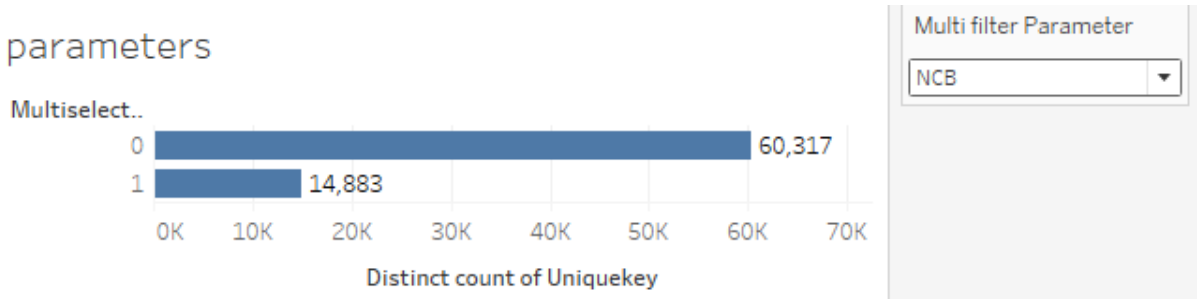


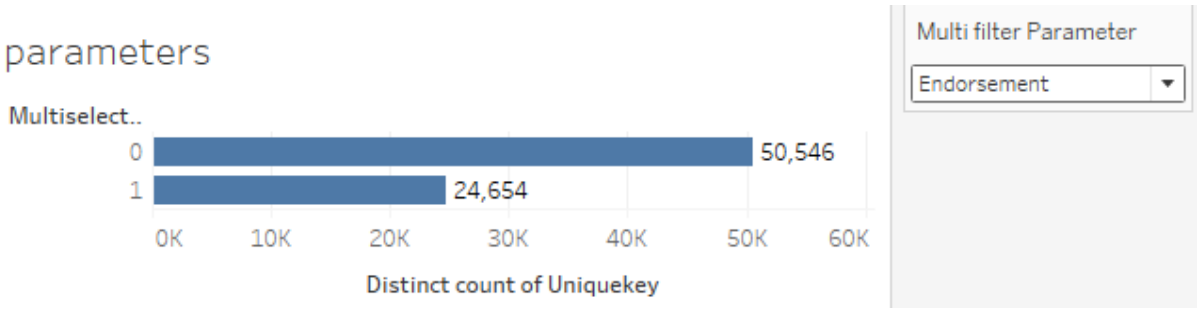

- Premium of the insurance company lies majorly between 8,000 to 10000 INR
- Vehicles which purchase high number of insurance policies from this company are in the age group of 0 to 5 years
- Sum insured of the policy sold by the company lies in between 10,000 to 15000 INR
- Total number of claims in the last 5 years is 1 for most of the insurance policy customers
- Majority of the drivers of the insured vehicles are either graduate /Post graduate
- Nature of the goods carried by the vehicles insured are non-hazardous
- Permit given by the government to most of the insured vehicles is local
- Road type of most the areas where the non-commercial vehicles are used is other than hilly, city / town, district roads
- CC not exceeding 1000 CC applicable for Private cars and Taxis are in majority as per the CC code histogram

# Bivariate Analysis



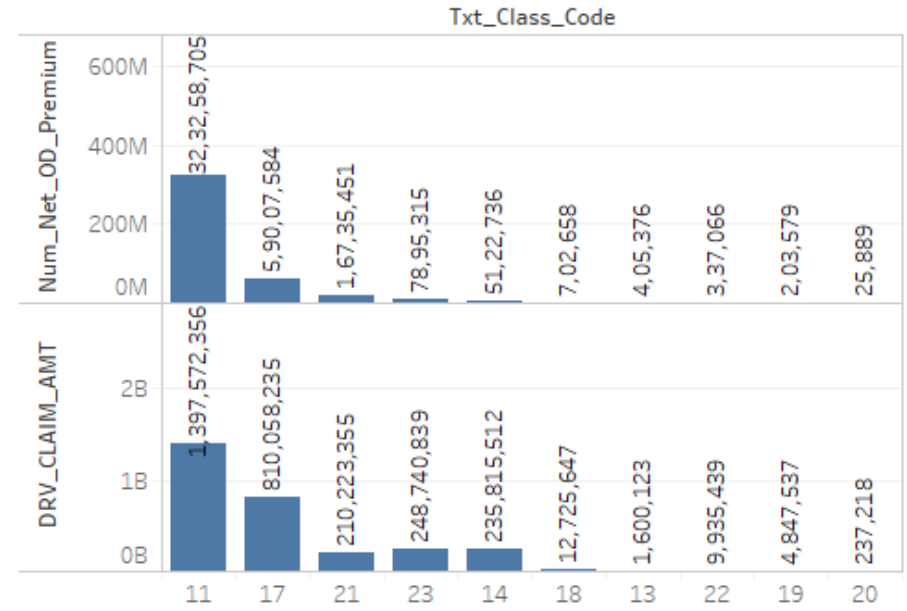


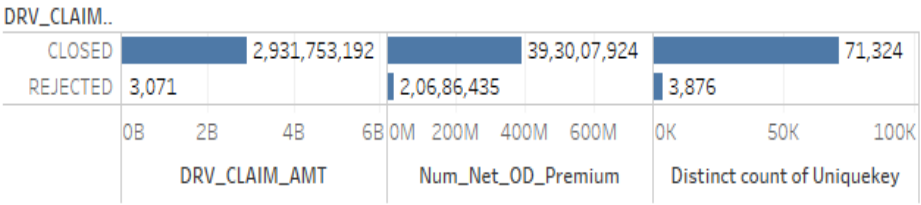




- In the given insurance claims dataset the total loss suffered by the company is less around 5,054 policies suffered total loss
- If the vehicle has the antitheft device in it then the company provides discount to the customer
- So 14883 policies of the company have been given antitheft discounts as well as NCB discounts
- Almost half of the policies sold are getting endorsed which is around 24654

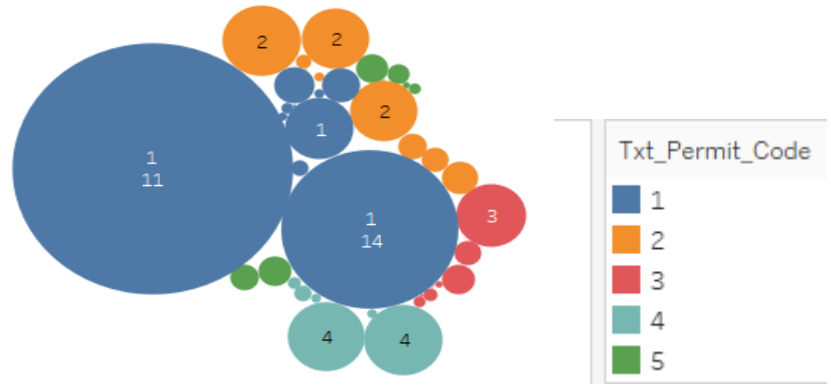
## carclassPremium

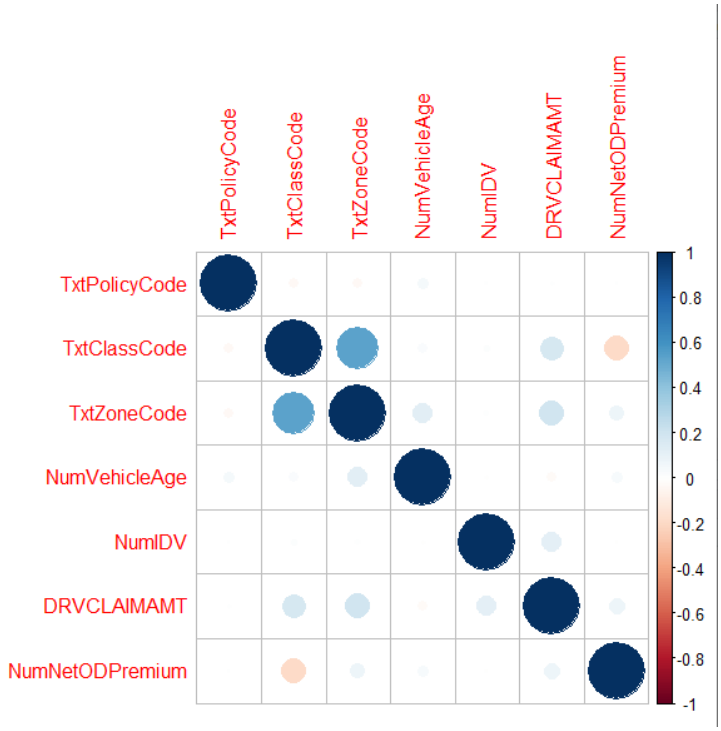



## status&Premium




## classbubblechart






- For the bar chart it is visible that the amount of claims accepted by the organization are very high as compared to the ones which are rejected
- Goods Carrying vehicles other than three wheelers –Public has the highest premium as well as claim amount
- Goods Carrying motorized three wheelers and pedal cycles – Private has the lowest claim amount and premium
- Goods Carrying vehicles other than three wheelers –Public, Two wheelers, Private Car has local permit code which contribute majority of policies sold
- Vehicles with permit code of hilly areas are the ones which are least sold since there is possibility of getting damage
- Vehicle zone code and Vehicle class code have good correlation

# Data Cleaning and Pre - processing

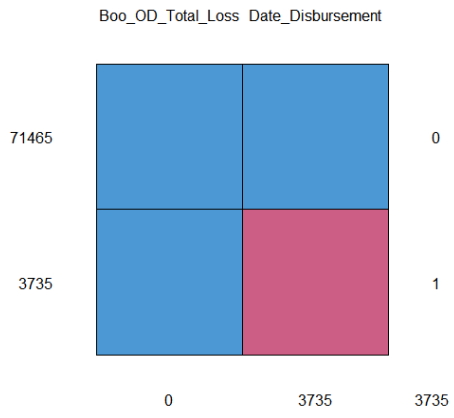
## Missing Value Treatment

- Date Disbursement column of the given dataset has 3735 missing values
- To exclude NA values from the column date disbursement na.omit() function can be used

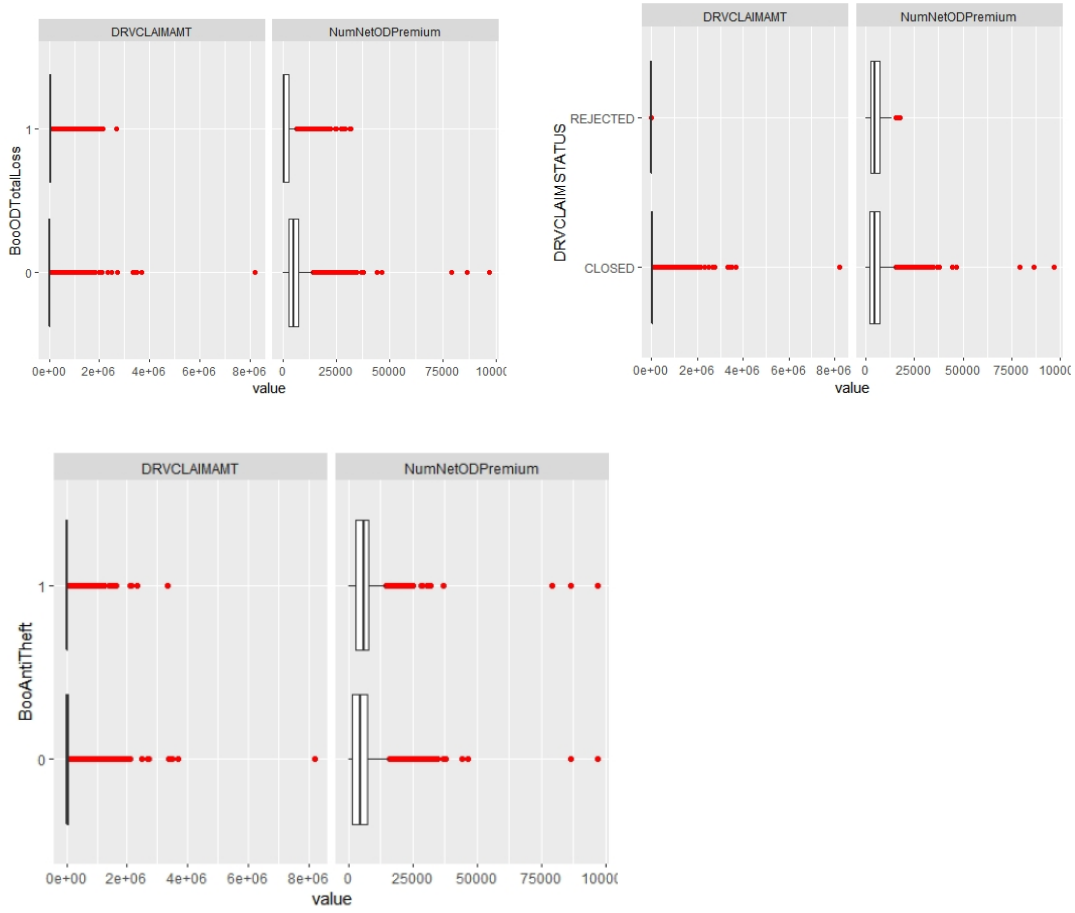
- Highlighted part in the chart shows that date disbursement column missing values
- Variables have been renamed by removing special characters like "-" to avoid errors

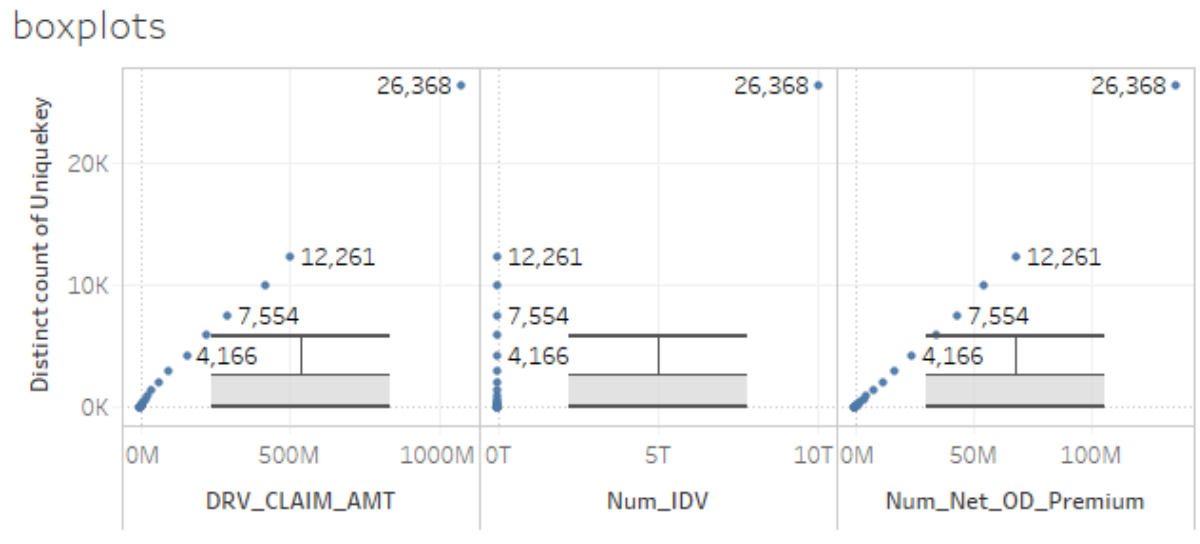
```
> colSums(is.na(InsuranceClaimsData))
                    Uniquekey               Txt_Policy_Year               Boo_Endorsement
                            0                             0                            0
            Txt_Location_RTA               Txt_Policy_Code                Txt_Class_Code
                            0                             0                            0
               Txt_Zone_Code               Num_Vehicle_Age          Txt_CC_PCC_GVW_Code
                            0                             0                            0
           Txt_Colour_Vehicle                       Num_IDV               Txt_Permit_Code
                            0                             0                            0
       Txt_Nature_Goods_Code            Txt_Road_Type_Code    Txt_Vehicle_Driven_By_Code
                            0                             0                            0
         Txt_Driver_Exp_Code       Txt_Claims_History_Code Txt_Driver_Qualification_Code
                            0                             0                            0
    Txt_Incurred_Claims_Code Boo_TPPD_Statutory_Cover_only                Txt_Claim_Year
                            0                             0                            0
          Date_Accident_Loss            Txt_Place_Accident         Date_Claim_Intimation
                            0                             0                            0
             Txt_TAC_NOL_Code              Date_Disbursement             Boo_OD_Total_Loss
                            0                          3735                            0
               DRV_CLAIM_AMT               DRV_CLAIM_STATUS                Boo_AntiTheft
                            0                             0                            0
                     Boo_NCB               Num_Net_OD_Premium
                            0                             0
```

```
> colnames(InsuranceClaimsData)[colSums(is.na(InsuranceClaimsData)) > 0]
[1] "Date_Disbursement"
> mean(InsuranceClaimsData$Date_Disbursement)
[1] NA
> data1<-na.omit(InsuranceClaimsData)
> sum(is.na(data1))
[1] 0
```



# Outlier Detection




- OD Total loss discount and net premium has all of the outliers in the range of 0 to 25000 INR
- Claim status rejected has very few outliers with respect to claim amount and Premium
- Claim status closed and net OD Premium has outliers in the range of 25000 to 50000 and later in the range of 75000 to 10000 INR

- Median Quartile of the net OD premium with claim status as closed lies around 14500 INR

## Outlier Treatment

- Interquartile Outlier treatment can be used for treating outliers
- 26368 policies have total premium greater than 100 million
- Claim amount have outliers in the range of 0 to 500 million
- Sum insured of the policies has outliers less than 2 trillion and later directly at 10 trillion with the count of 26368

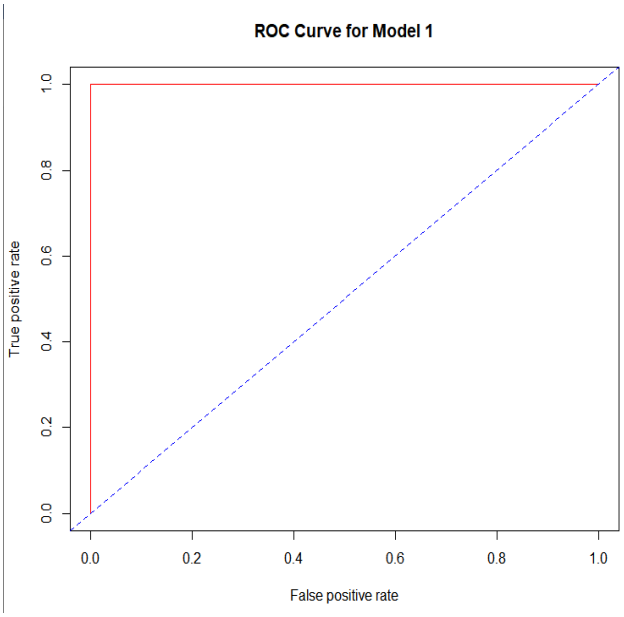
```
> quantile(data1$NumIDV, c(0.95))
    95%
1125297
> data1$NumIDV[which(data1$NumIDV>1125297)]<- 1125297
> quantile(data1$NumNetODPremium,c(0.95))
  95%
15291
> data1$NumNetODPremium[which(data1$NumNetODPremium>15291)] <- 15291
> quantile(data1$DRVCLAIMAMT,c(0.95))
   95%
155492
> data1$DRVCLAIMAMT[which(data1$DRVCLAIMAMT> 155492)] <- 155492
```

# Model Building

## Logistic Regression

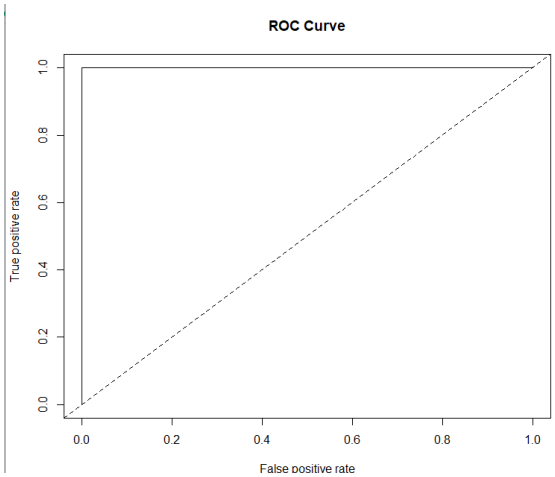
- The insurance claims dataset consists of a categorical dependent factor claim status which depends on various independent factors
- Claim amount of the policy is significant whereas vehicle driven by the driver and the experience of the driver, nature of goods carried by the vehicle are highly significant
- Driver's qualification is least significant among other factors used for the model
- Maximum likelihood of the model considering nature of goods carried by vehicle and experience of the driver is 9 used to best fit the data
- The logistic Regression model shows that the independent factors such as vehicle driver, driver experience ,claim amount and nature of goods carried by the vehicle should be majorly considered since the AIC of that model is 10.379 which is less as compared to other models

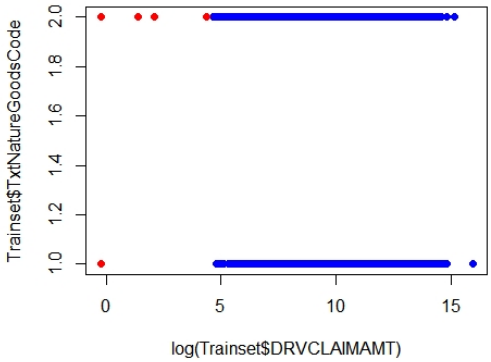
- Gini coefficient of the model is 0.1% very less as compared to the 60% criteria of the model
- Accuracy, sensitivity and specificity of the model is 1 which is prefect as compared to other models
- The model shows that the VIF of the vehicle driven by the driver is 1.04 highest as compared to the other model variables

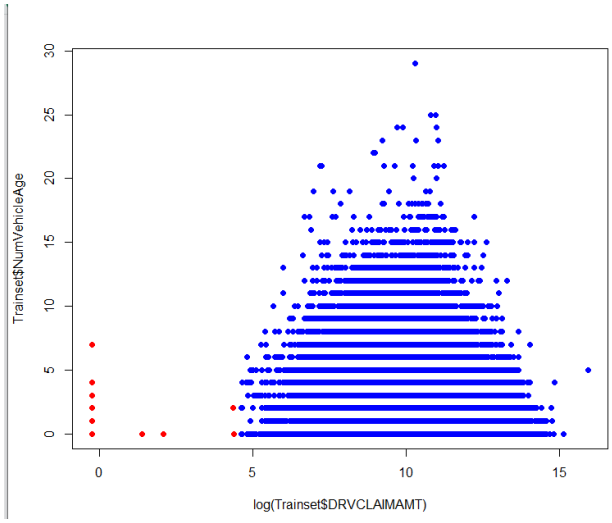


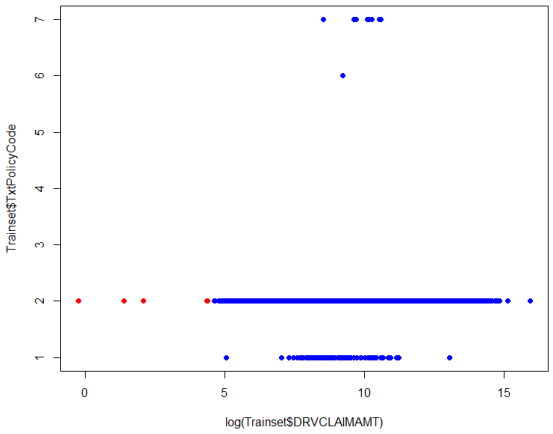
## CART

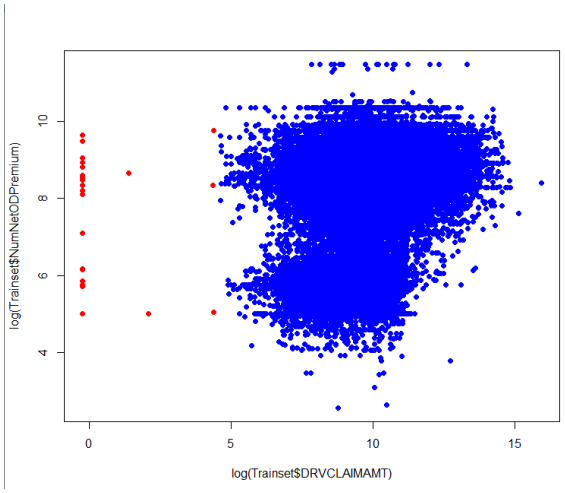
- In the left graph it is visible that the number of claims accepted by the company increases when the goods carried by the vehicle are non-hazardous
- Claims accepted are in the range of 5 to 15 lakhs when goods carried by the vehicle are hazardous whereas in non - hazardous the claim amount of the vehicle exceeds 15 lakh also
- As the Age of the vehicle increases the number of claims accepted by the company and the claim amount is reduced
- Rejected claims of the vehicle are less than overall 5 lakh claim amount and age of the vehicle between 0 to 5 are mostly rejected by the company having amount less than 5 lakh
- Higher the claim amount and higher the qualification of the driver greater are the chances of claim acceptance
- Graduate drivers are having lower probability of claim rejection
- Premium amount of the vehicle is directly proportional to claim amount
- Premium amount between 6,000 to 10,000 and claim amount less than 5 lakh has the highest probability of claim rejection
- 57061 policies are predicted to be rejected by the company and are actually accepted whereas 111 policies which are predicted to the rejected are actually rejected on the train dataset
- 14263 policies are predicted to be rejected by the company and are actually rejected whereas 30 policies which are predicted to the rejected are actually rejected on the test dataset
- Sensitivity ,specificity and accuracy of this model is 100%
- Vehicles driven on city / town and on other road types are having higher number of claim acceptance by the company
- Vehicles driven on district roads are least rejected when claimed by the customer
- Recall and precision of the CART model shows value as 1 which means it is the best model

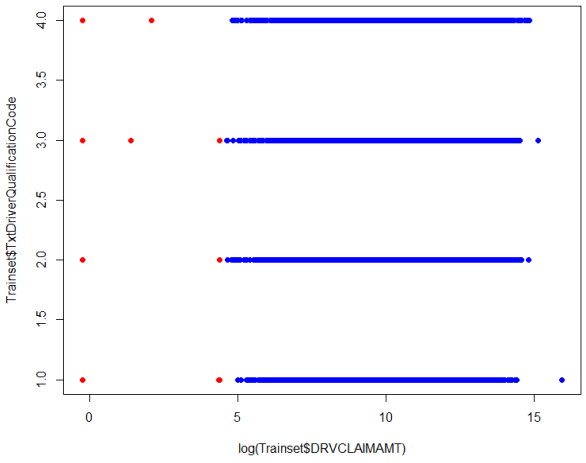

ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

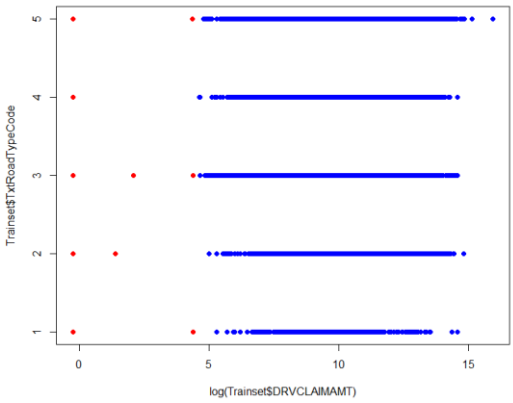

Trainset$TxtNatureGoodsCode
log(Trainset$DRVCLAIMAMT)
1.0
1.2
1.4
1.6
1.8
2.0
0
5
10
15

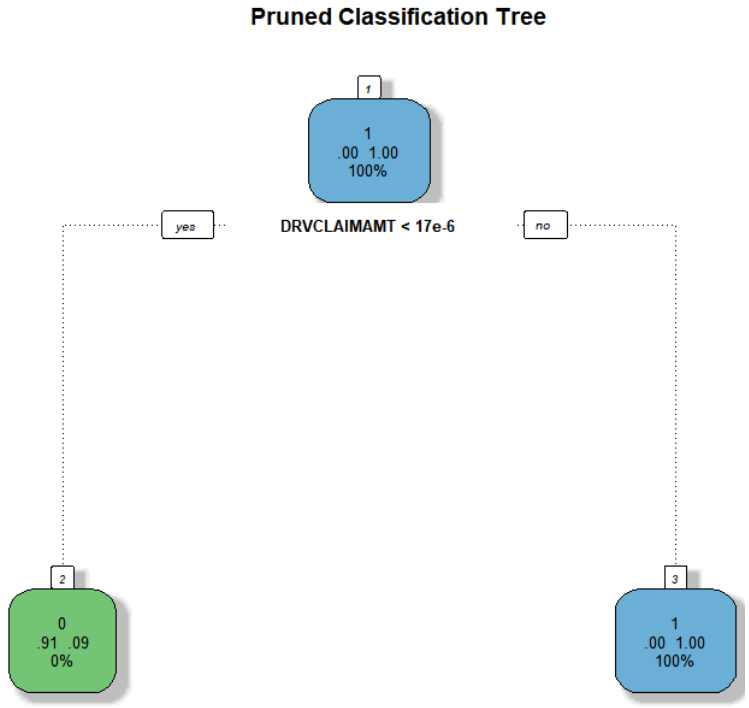

Trainset$NumVehicleAge
log(Trainset$DRVCLAIMAMT)
0
5
10
15
20
25
30


Trainset$TxtPolicyCode
log(Trainset$DRVCLAIMAMT)
1
2
3
4
5
6
7
0
5
10
15


log(Trainset$NumNetODPremium)
log(Trainset$DRVCLAIMAMT)
4
6
8
10
0
5
10
15


Trainset$TxtDriverQualificationCode
log(Trainset$DRVCLAIMAMT)
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
5
10
15

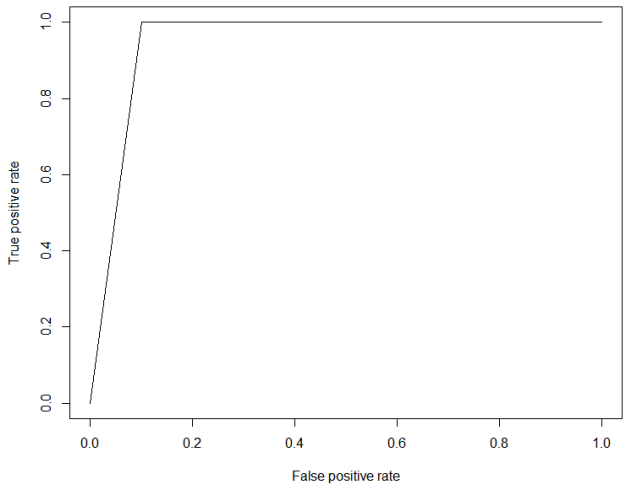



## Naive Bayes

- AUC of the curve plotted in Naive Bayes algorithm is 0.9499299 which implies 94.9% of the predictions of this model are correct
- Accuracy of this model is 99.5 %
- The number of correct predictions that the occurrence is positive is 0.995 whereas the number of correct predictions that the occurrence is negative is 0.991
- Recall and precision of the model is 0
- K test value of this model is 0

- Gini coefficient is  0.001013452 which is 0.1% which is very less as compared to 60% so the model is not good model



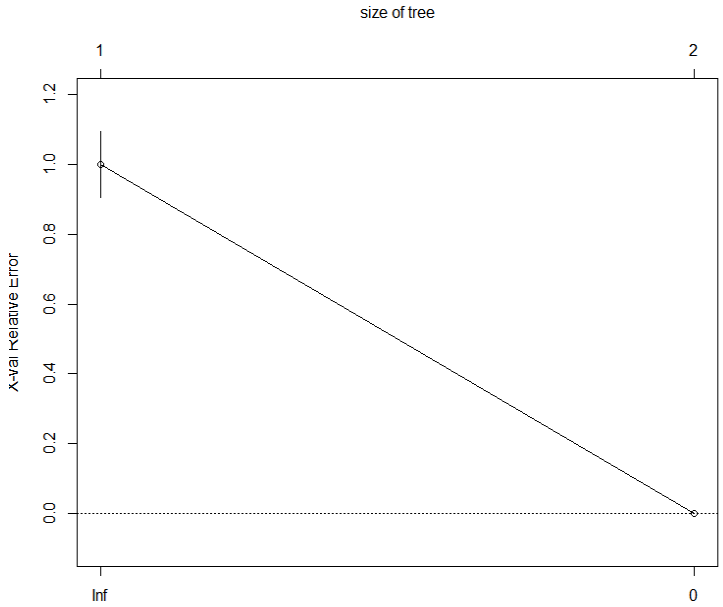

## KNN Algorithm

- Accuracy of the KNN model is 99.8 %
- The number of correct predictions that the occurrence is positive is 1 whereas the number of correct predictions that the occurrence is negative is 0
- Recall rate of the KNN model is 0
- Precision of the KNN model is 0
- AUC of the KNN model is 0.9931 which implies 99.3% of the predictions of this model are correct

## Ensemble modelling

### 1) Decision tree

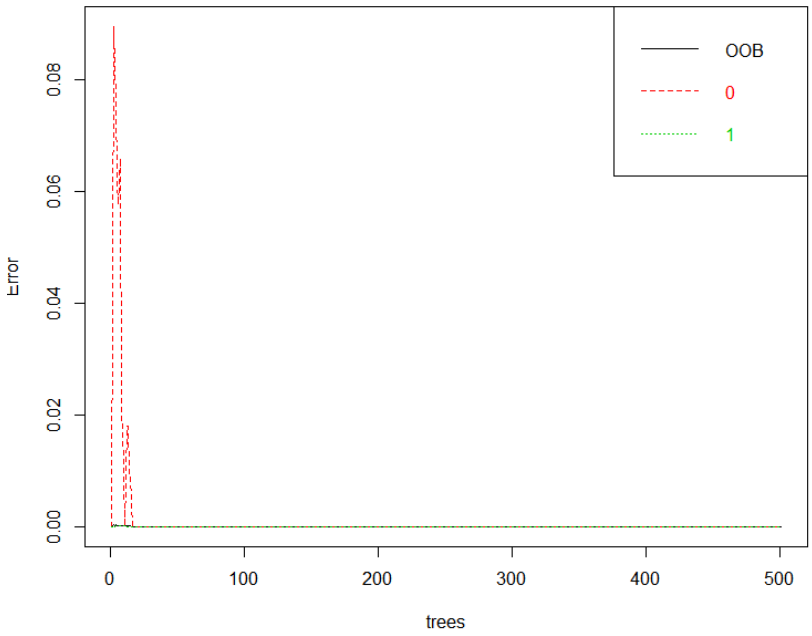
- The important variable as per the CART decision tree is claim amount
- The total number of records in the trainset is 57172 and threshold for the claim amount can be considered as 91.5 as per the tree
- The number of policies having claim amount less than 91.5 is 111 and the ones having claim amount greater than or equal to 91.5 is 57061
- As the size of the tree increases the root node error decreases to 0.001

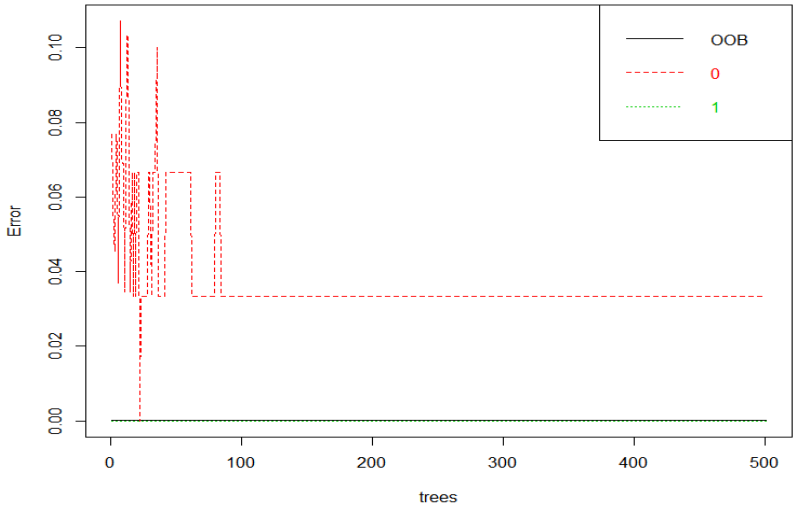



## 2) Random Forest

- Number of trees as per classification are 501 and number of variables each split is considered as 3
- Important variables that contribute highest to insurance claims are Claim amount, endorsement, antitheft discount and total OD loss of the policy as per training dataset and testing dataset
- Variables contributing least to the insurance claims are Net OD premium and age of the vehicle as per training dataset
- TPPDStatutorycoveronly variable is least important variable in test dataset
- As the number of trees increases in the training dataset the OOB error starts gradually decreasing and remains constant after 30 trees
- As the number of trees increases in the testing dataset the OOB error starts gradually decreasing and remains constant after 100 trees
- Recall and precision is also 1 for this model
- Correct predictions turned out to be positive are 1
- Correct predictions turned out to be negative are 1
- Since the random forest gives 0 % OOB estimate error rate so random forest tuning is not required for training dataset
- Since the random forest gives 0.06 % OOB estimate error rate so random forest tuning is not required for testing dataset

Since the tuning is not required for the model so the model is not overfit and is stable




# Model Validation

- The tabulated model comparison below shows that the random forest is the best model
- Decision trees can be used for tuning models like KNN and CART
- Random forest does not require tuning

| Models | Logistic Regression | CART | KNN | Naive Bayes | Random Forest |
|---|---|---|---|---|---|
| Gini Coefficient | 0.1 | 0.01 | 0.1 | 0.1 | 0 |
| Sensitivity | 1 | 1 | 1 | 0.995 | 1 |
| Specificity | 1 | 1 | 0 | 0.991 | 0.98 |
| Precision | 1 | 1 | 0 | 0 | 1 |
| Recall | 1 | 1 | 0 | 0 | 1 |

## Conclusion:

- Random Forest is the best fit model for the given dataset since the OOB rate is minimal and also the tree is stable and not overfitted with 98 % accuracy
- Company should keep a watch on those vehicles which are non - hazaradous and are running on city/town or other road types while accepting the claim
- Since this vehicles are having least probability of getting damaged
- Private Car (only three wheelers with 750 to 1000cc) and Goods Carrying vehicles other than three wheelers –Public are the class of the vehicle whose claims are accepted highly
- Goods Carrying motorised three wheelers and pedal cycles – Private is the class of the vehicle whose claims are least accepted
- Claim history of the policy is either no claim or 1 claim which are accepted by the insurance company
- Reviews from the internal employees regarding the claims accepted by the company should be conducted regularly so as to check whether the fraud is conducted by some internal employee

- The claims staff should be made aware of the current market conditions so that the claim would be accepted by the provided guidelines
- The package and liability only policies are the ones whose claims are majorly accepted
- So there should be a strict check conducted by the insurance company team for the damage caused to the vehicle either by fire, theft or accident

# Recommendations

- Company should keep a watch on those vehicles which are non - hazardous and are running on city/town or other road types while accepting the claim since this vehicles are having least probability of getting damaged
- Claim history of the policy is either no claim or 1 claim which are accepted by the insurance company
- Reviews from the internal employees regarding the claims accepted by the company should be conducted regularly so as to check whether the fraud is conducted by some internal employee
- The claims staff should be made aware of the current market conditions so that the claim would be accepted by the provided guidelines
- The package and liability only policies are the ones whose claims are majorly accepted
- Since the number of claims are increasingly getting closed so we should increase the policy term for car insurance

## Business Domain

- The dataset can be used in banking domain for giving loan and EMI facility for buying cars
- It can also be used in health care sectors since vehicles are prone to accidents
- It can also be used in transport facility start up companies like Ola/Uber