

Home Work Assignment - 04

Critical Thinking Group 5

Contents

1 Overview	1
2 Data Exploration and Cleanup / Common Transformations	2
2.1 Variable Identification	2
2.2 Data Cleanup	3
3.2 Data Preparation	10
3.3 Build Models	16
3.4 Model Evaluation Using VALID Data	20
3.5 Final Logistic Model Selection Summary	21
4 Linear Regression for TARGET_AMT	27
4.1 Data Summary and Correlation Analysis	27
4.2 Data Preparation	32
4.3 Build Models	37
4.4 Final Linear Model Selection Summary	41
5 Prediction Using Evaluation Data	48
5.1 Transformation of Evaluation Data	48
5.2 Model Output for Logistic Regression	48
5.3 Model Output for Linear Regression	48
5.4 Conclusion	49

newpage

1 Overview

The data set contains approximately 8161 records. Each record represents a customer profile at an auto insurance company. Each record has two response variables.

The first response variable, TARGET_FLAG, is a 1 or a 0. A “1” means that the person was in a car crash. A zero means that the person was not in a car crash.

The second response variable is TARGET_AMT. This is the amount spent on repairs if there was a crash. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

We will be exploring, analyzing, and modeling the training data. Since there are 2 different predictions we have to work with, we will deal with each prediction independently. The following are the 2 predictions we will be modeling for:

1. TARGET_FLAG - This dependent variable tells whether there was a crash or not. This is a binary variable and as such we will be using a Logistic Regression Model to predict this.
2. TARGET_AMT - This dependent variable gives the amount / cost of repairs if there was a crash. This is a continuous variable and we will be using a Linear Regression Model to predict this.

Each of the above models will be built and evaluated separately. In the first section of this document we will deal with the Logistic Model for TARGET_FLAG and in the second section we will deal with Linear Model for the TARGET_AMT

Out of the many models for each task, we will go ahead and shortlist one model that works the best. We will then use these models (one for each task) on the test / evaluation data.

To attain our objective, we will be following the below steps for each modeling exercise:

1 -Data Exploration 2 -Data Preparation 3 -Build Models 4 -Select Models

Model Selection Strategy: As a strategy, we will split the train dataset into 2 parts - TRAIN and VALID. In the VALID dataset, we will hold out some values to validate how well the model is trained using the TRAIN dataset. We will then use the Model that performs the best on the EVALUATION data to give the required output. We will split the TRAIN / VALID data after the **Data Exploration / Preparation** before the **Build Models**.

Please Note:

- There are some common clean-up and transformations that we will carry out initially that will serve all the models.
- While working on the Linear Models for the TARGET_AMT, we will be using only a subset of the data where the TARGET_FLAG = 1. This will give us all the records where there was a crash and subsequently a repair amount.
- While Predicting the TARGET_AMT with the given Evaluation dataset, We will take the output of the TARGET_FLAG predictions on the Evaluation dataset and use only those rows that were classified as a “Crash” and use it as the input to the TARGET_AMT prediction. So this is a two step prediction, one for the TARGET_FLAG and using the output to predict TARGET_AMT.

newpage

2 Data Exploration and Cleanup / Common Transformations

In this section we go ahead and perform some common cleanup and create additional variables that will be used for modeling both the logistic as well as the linear regressions. We will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

- Variable Identification / Relationships
- Data Clean-up
- Common Transformations
- Create Missing Flags / Impute Missing Values

2.1 Variable Identification

First let's display and examine the data dictionary or the data columns as shown in below table:

Table 1: Variable Description

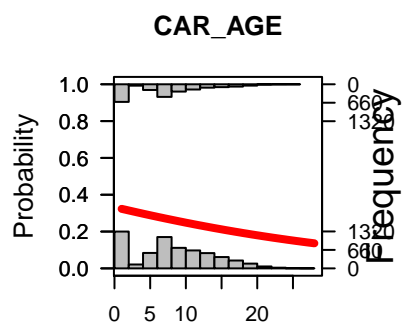
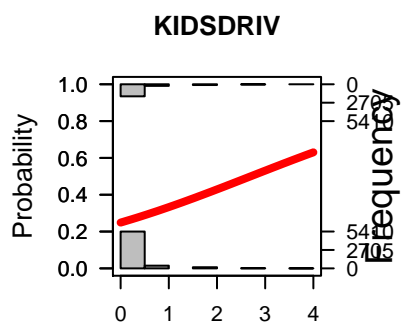
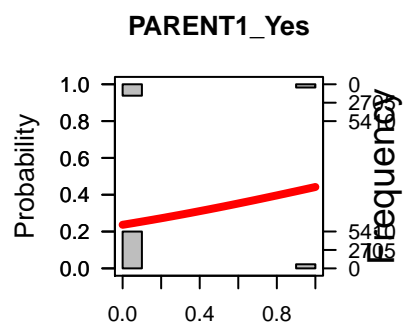
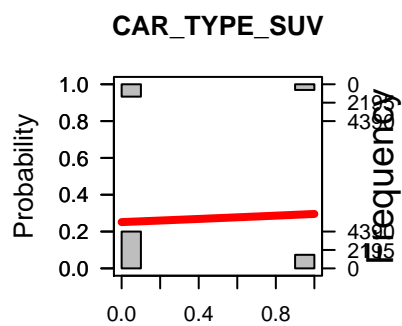
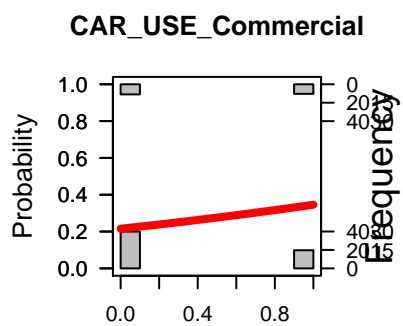
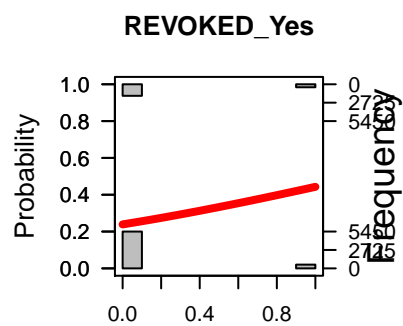
VARIABLE	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people a
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effec
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effec
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effec
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase prob
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likel
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get int
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more cra
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this su
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) a
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably
SEX	Gender	Urban legend says that women have less crashes then men. I
TIF	Time in Force	People who have been customers for a long time are usually
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more saf

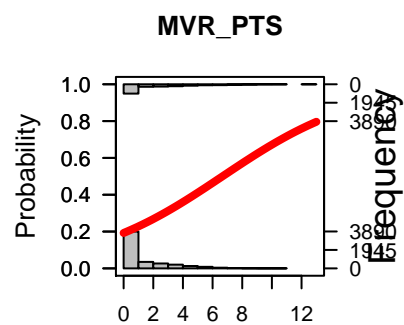
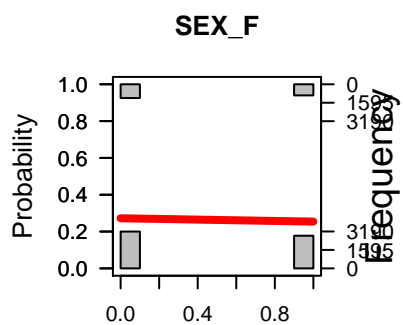
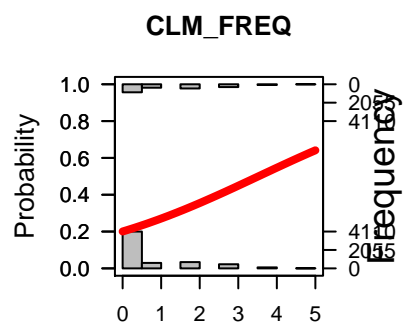
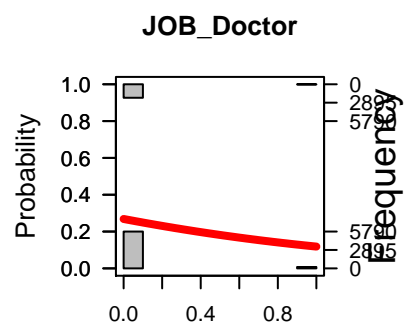
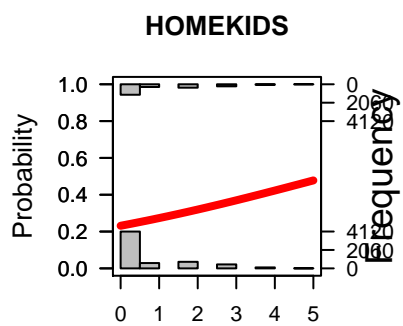
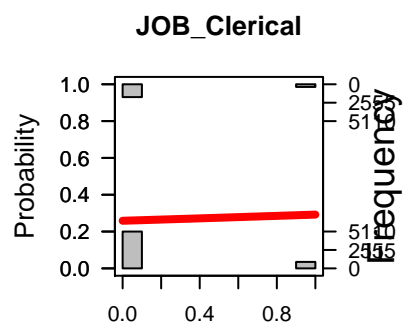
We notice that there are 2 dependent variables - TARGET_FLAG and TARGET_AMT. Apart from these 2 dependent variables, we have 23 independent or predictor variables.

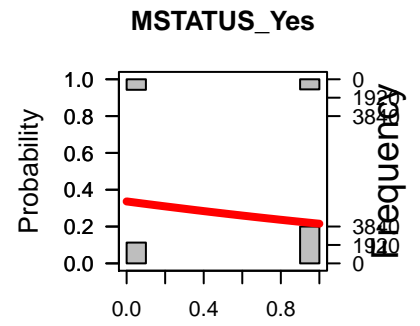
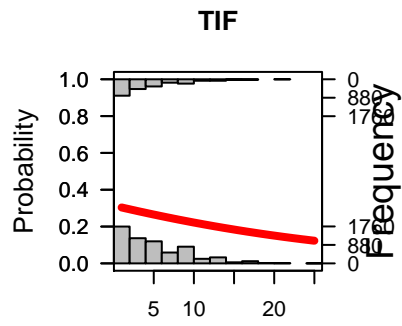
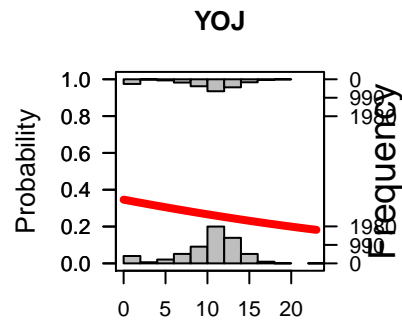
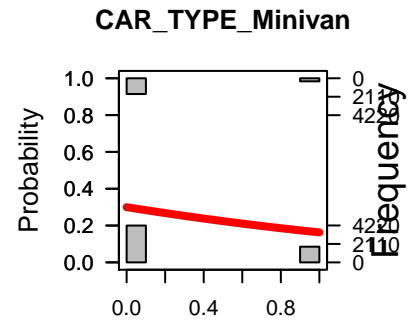
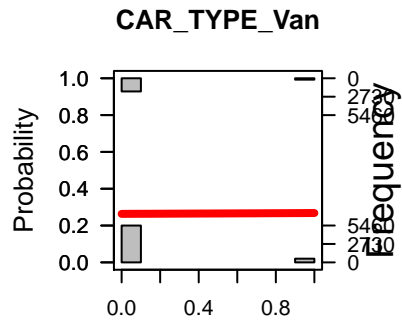
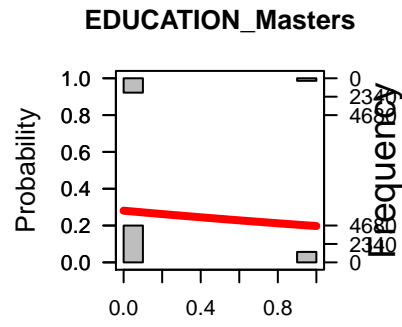
2.2 Data Cleanup

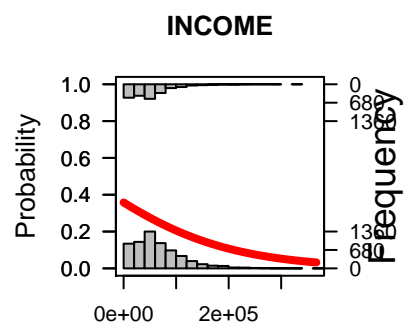
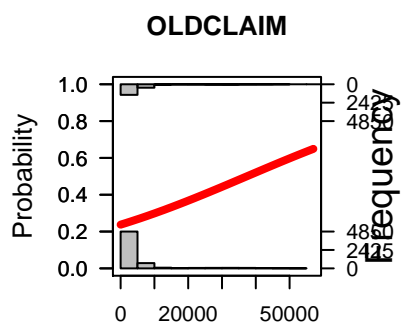
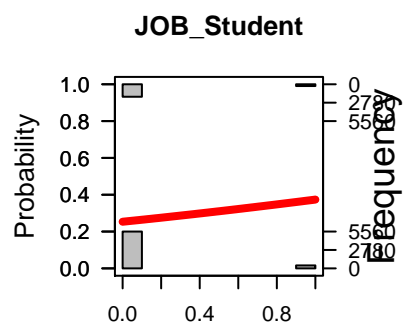
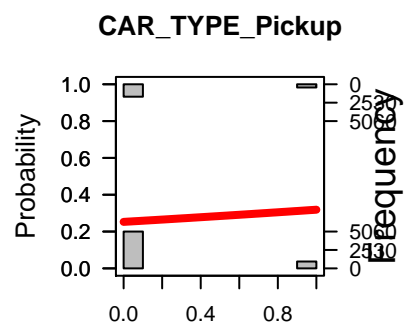
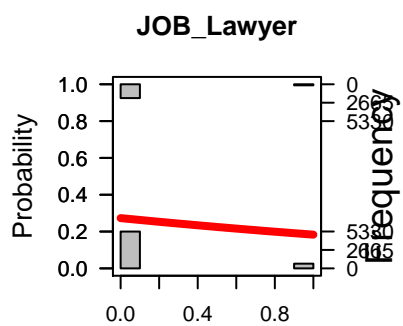
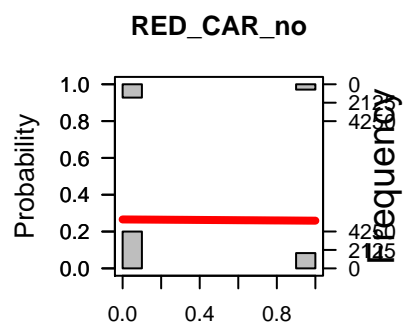
[1] “Yes” “z_No” [1] “M” “z_F” [1] “= 0) It must always be less than equals to 1 (since $p \leq 1$).

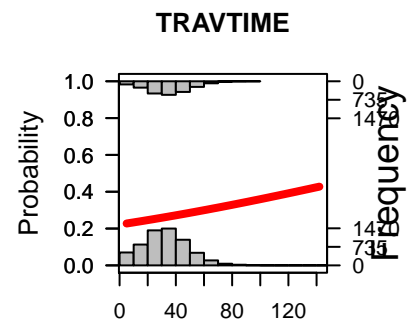
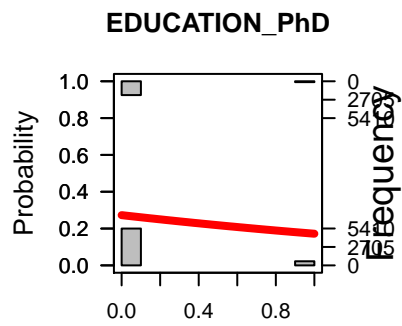
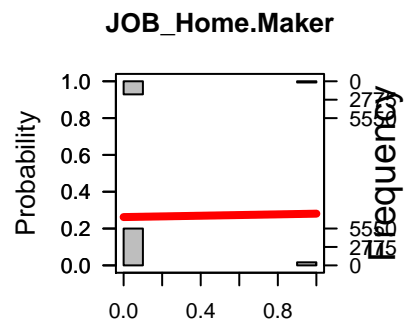
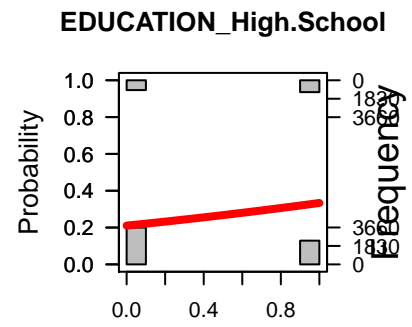
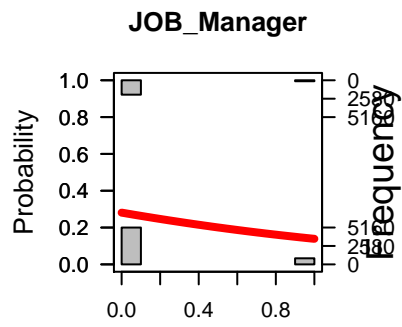
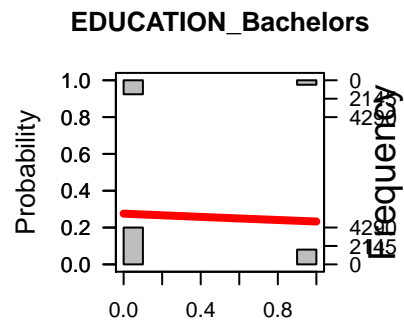
Now let’s investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:

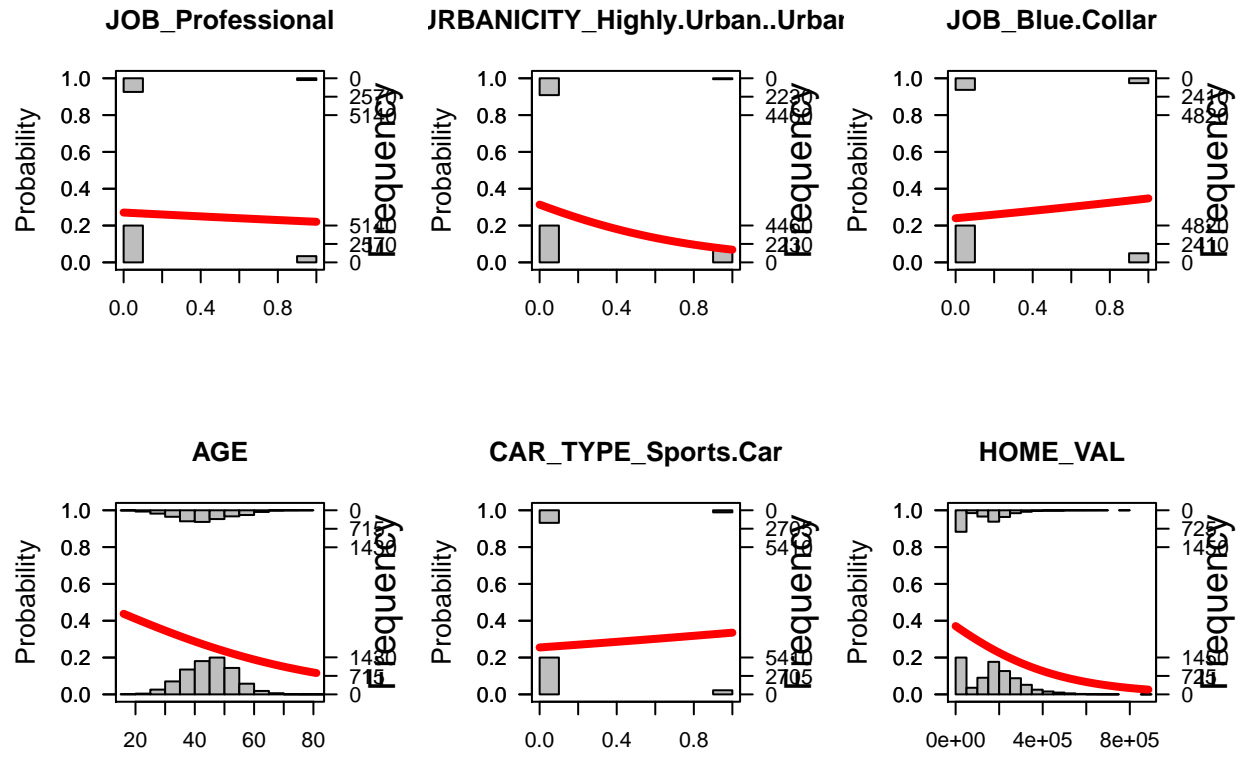


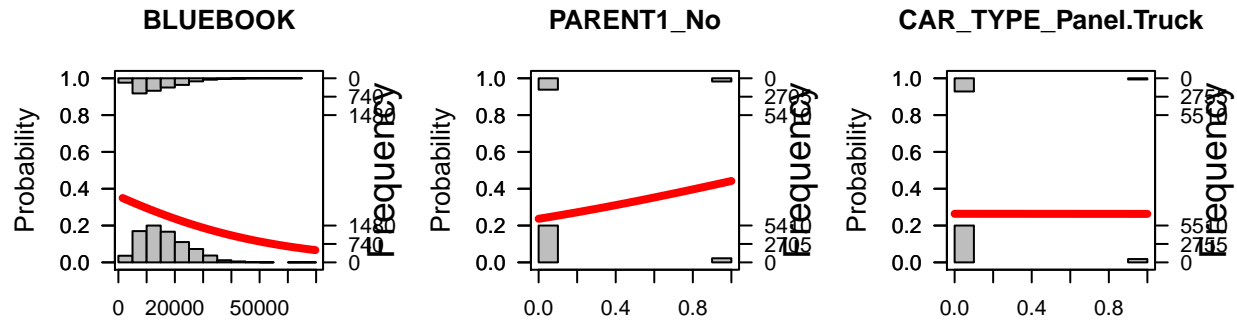












3.1.5.1 Interpretation You can see that the probability of crashing increases as we get closer to the “1” classification for the CAR_TYPE_Van, RED_CAR_no, JOB_Home.Maker, SEX_F, JOB_Clerical, CAR_TYPE_SUV, TRAVTIME, BLUEBOOK, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, JOB_Student, KIDSDRIV, JOB_Blue.Collar, HOMEKIDS, MSTATUS_No, EDUCATION_High.School, CAR_USE_Commercial, REVOKED_Yes, PARENT1_Yes, OLDCLAIM, CLM_FREQ, MVR_PTS, URBANICITY_Highly.Urban..Urban variables.

You can see that the probability of crashing decreases as we get closer to the “1” classification for the HOME_VAL, CAR_TYPE_Minivan, MSTATUS_Yes, JOB_Manager, AGE, CAR_AGE, TIF, EDUCATION_Masters, YOJ, EDUCATION_PhD, JOB_Lawyer, JOB_Doctor, EDUCATION_Bachelors, JOB_Professional, INCOME, SEX_M, RED_CAR_yes, CAR_TYPE_Panel.Truck variables.

3.2 Data Preparation

Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

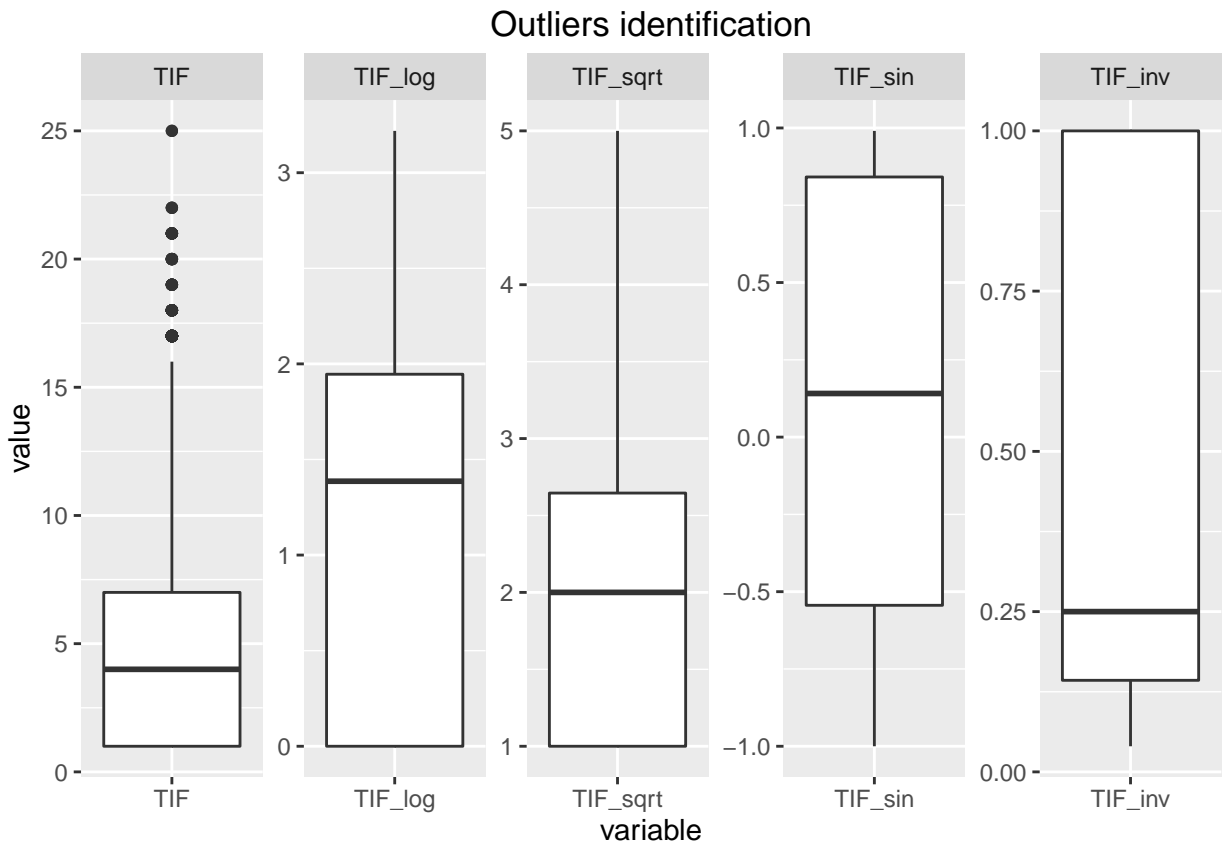
We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

3.2.1 Outliers treatment

In this sub-section, we will check different transformations for AGE, BLUEBOOK and TIF to create the appropriate outlier-handled / transformed variables.

Transformations for TIF

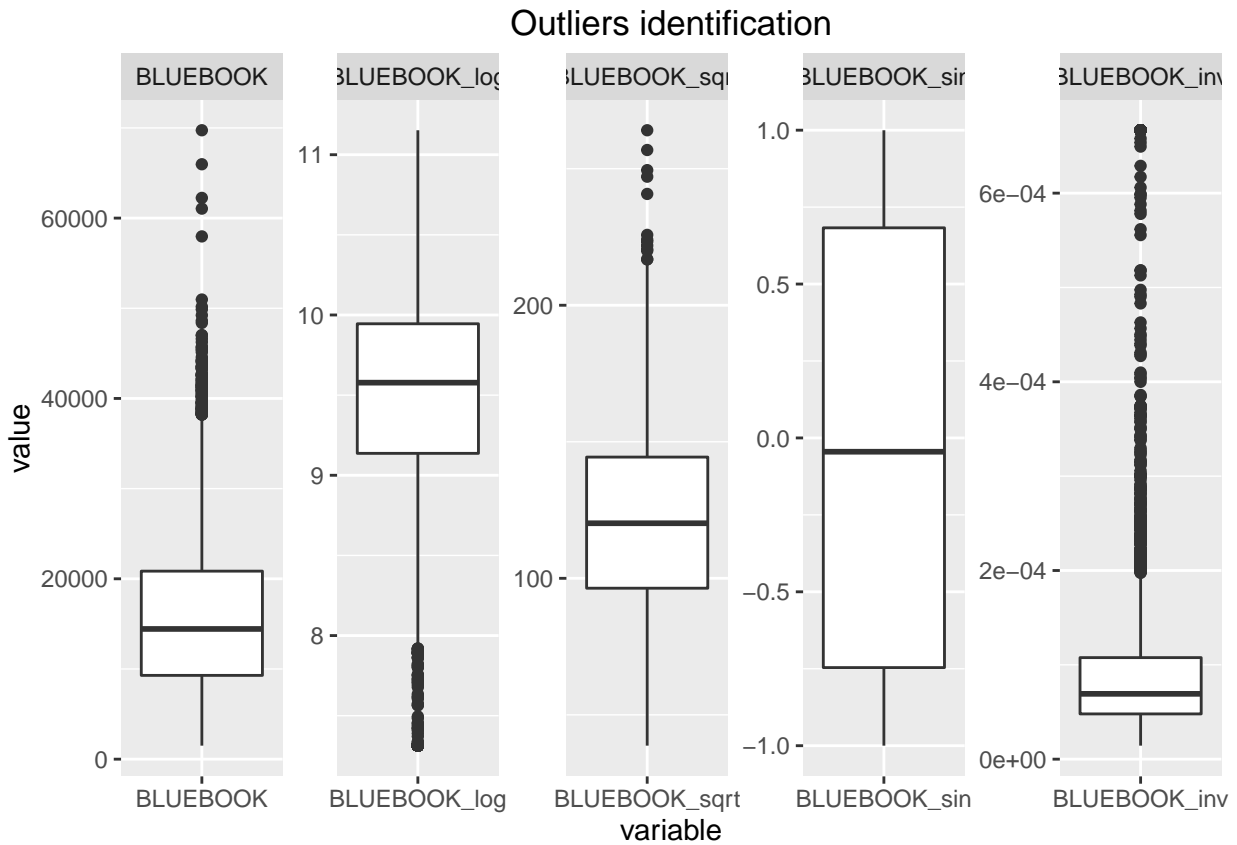
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	1.00	1.00	4.00	5.35	7.00	25.00



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, the sin transformation seems to be better distributed. Hence, We will create this variable.

Transformations for BLUEBOOK

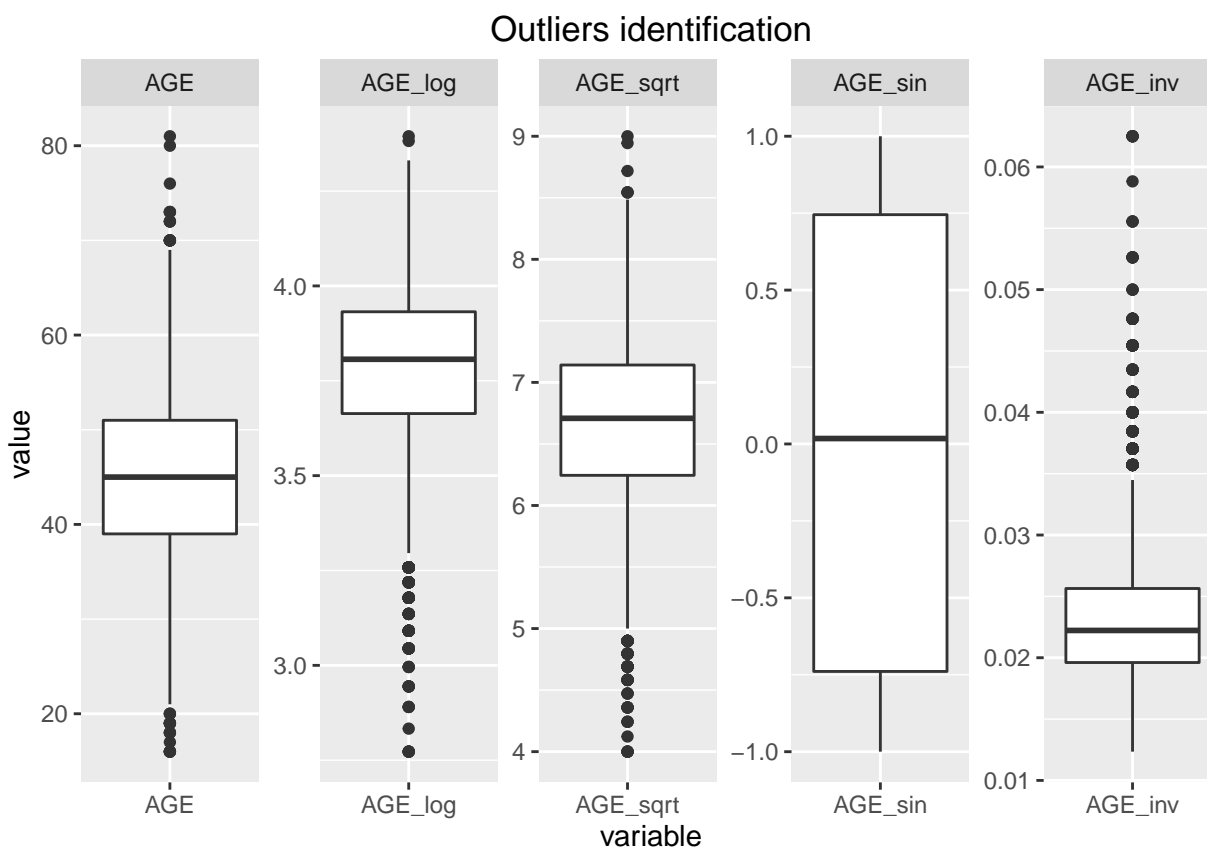
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	1500	9290	14440	15710	20850	69740



From the above charts we can see that a sin transformation works well. Hence, We will create this variable.

Transformations for AGE

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	16.00	39.00	45.00	44.79	51.00	81.00



From the above charts we can see that a sin works well for AGE. Hence, We will create this variable.

3.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_FLAG

CAR_TYPE - If you drive Minivans and Panel Trucks you have lesser chance of being in a crash as against Pickups, Sports, SUVs and Vans. Since the distinction is clear, we believe that binning this variable accordingly will help strengthen the correlation. Accordingly, we will bin this variable as below:

CAR_TYPE_FLAG_BIN :

- 1 : if CAR_TYPE is Minivans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Vans

EDUCATION - If you have only a high school education then you are more likely to crash than if you have a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation:

EDUCATION_FLAG_BIN :

- 0 : if EDUCATION is High School
- 1 : if EDUCATION is Bachelors, Masters or Phd

JOB - If you are a Student, Homemaker, or in a Blue Collar or Clerical job, you are more likely to be in a crash against Doctor, Lawyer, Manager or professional. Again binning this variable will strengthen the correlation:

JOB_TYPE_FLAG_BIN :

- 1 : if JOB_TYPE is Student, Homemaker, or in a Blue Collar or Clerical
- 0 : if JOB_TYPE is Doctor, Lawyer, Manager, professional, Unknown
- INCOME - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this at zero value.

INCOME_FLAG_BIN :

- 1 : if INCOME \leq 0
- 0 : if INCOME $>$ 0
- YOJ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

YOJ_FLAG_BIN :

- 1 : if YOJ \leq 0
- 0 : if YOJ $>$ 0
- HOME_VAL - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

HOME_VAL_FLAG_BIN :

- 1 : if HOME_VAL \leq 0
- 0 : if HOME_VAL $>$ 0
- OLDCLAIM- There is a huge difference in the correlation when we transform this variable. Binning this variable seems like a good idea.

OLDCLAIM_FLAG_BIN :

- 1 : if OLDCLAIM \leq 0
- 0 : if OLDCLAIM $>$ 0
- CLM_FREQ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

CLM_FREQ_FLAG_BIN :

- 1 : if CLM_FREQ \leq 0
- 0 : if CLM_FREQ $>$ 0
- MVR_PTS - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

MVR_PTS_FLAG_BIN :


```
## $ CAR_USE_Commercial: num 0 1 0 0 0 1 0 1 0 1 ...
## $ MSTATUS_Yes : num 0 0 1 1 1 0 1 1 0 0 ...
## $ PARENT1_Yes : num 0 0 0 0 0 1 0 0 0 0 ...
## $ RED_CAR_yes : num 1 1 0 1 0 0 0 1 0 0 ...
## $ REVOKED_Yes : num 0 0 0 0 1 0 0 1 0 0 ...
## $ SEX_M : num 1 1 0 1 0 0 0 1 0 1 ...
## $ URBANICITY_Rural : num 0 0 0 0 0 0 0 0 0 1 ...
## $ YOJ_MISS : num 0 0 0 0 1 0 1 1 0 0 ...
## $ INCOME_MISS : num 0 0 0 1 0 0 0 0 0 0 ...
## $ HOME_VAL_MISS : num 0 0 0 0 0 0 1 0 0 0 ...
## $ CAR_AGE_MISS : num 0 0 0 0 0 0 0 0 0 0 ...
## $ TIF_sin : num -1 0.841 -0.757 0.657 0.841 ...
## $ BLUEBOOK_sin : num -0.988 -0.988 0.971 0.8 -0.97 ...
## $ AGE_sin : num -0.305 -0.832 -0.428 0.67 -0.262 ...
## $ CAR_TYPE_FLAG_BIN : num 1 1 0 1 0 0 0 0 0 0 ...
## $ EDUCATION_FLAG_BIN: num 1 0 0 0 1 1 0 1 1 1 ...
## $ JOB_TYPE_FLAG_BIN : num 0 1 1 1 0 1 1 1 1 0 ...
## $ INCOME_FLAG_BIN : num 0 0 0 0 0 0 0 0 0 0 ...
## $ YOJ_FLAG_BIN : num 0 0 0 0 0 0 0 0 0 0 ...
## $ HOME_VAL_FLAG_BIN : num 1 0 0 0 0 1 0 0 1 1 ...
## $ OLDCLAIM_FLAG_BIN : num 0 1 0 1 0 1 1 0 1 1 ...
## $ CLM_FREQ_FLAG_BIN : num 0 1 0 1 0 1 1 0 1 1 ...
## $ MVR_PTS_FLAG_BIN : num 0 1 0 1 0 1 1 0 1 0 ...
## $ CAR_AGE_FLAG_BIN : num 0 1 0 0 0 0 1 0 1 0 ...
## $ TRAVTIME_FLAG_BIN : num 1 0 1 0 0 0 0 0 0 0 ...
```

3.3 Build Models

In this section, we will create 3 models. Aside from using original and transformed data, we will also using different methods and functions such as Linear Discriminant Analysis, step function, and logit function to enhance our models. Below is our model definition: -Model 1- This model will be created using all the variables in train data set with logit function GLM. -Model 2: This model step function will be used to enhance the model 1. -Model 3- This model will be created using classification and regression tree.

3.3.1 Prepare TRAIN and VALID datasets

However, prior to that, we hold out a subset of data as a validation dataset to check model performance. This will be useful when we select a model.

```
smp_size <- floor(0.80 * nrow(DS_TARGET_FLAG))

## set the seed to make your partition reproducible
set.seed(123)

train_index <- sample(seq_len(nrow(DS_TARGET_FLAG)), size = smp_size)

DS_TARGET_FLAG_TRAIN<- DS_TARGET_FLAG[train_index, ]
DS_TARGET_FLAG_VALID <- DS_TARGET_FLAG[-train_index, ]
```


3.3.2 Model 1 and enahncement of Model 1 with step function (Model 2)

In this model, we will be using all the given variables in train data set. We will create model using logit function. We will then step thru the model to remove unnecessary variables and generate the refined model. We will highlight the summary of the refined model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.988e-01  3.403e-01  -1.172  0.241181
## KIDSDRIV       3.979e-01  6.763e-02   5.883  4.03e-09 ***
## AGE          -4.759e-03  4.458e-03  -1.067  0.285793
## HOMEKIDS       1.818e-02  4.169e-02   0.436  0.662766
## YOJ           1.598e-02  1.378e-02   1.159  0.246305
## INCOME        -3.058e-06  1.260e-06  -2.427  0.015231 *
## HOME_VAL      -8.986e-07  6.553e-07  -1.371  0.170277
## TRAVTIME       1.334e-02  2.786e-03   4.786  1.70e-06 ***
## BLUEBOOK      -1.680e-05  4.762e-06  -3.527  0.000420 ***
## TIF           -5.153e-02  9.256e-03  -5.567  2.59e-08 ***
## OLDCLAIM      -2.124e-05  4.735e-06  -4.487  7.23e-06 ***
## CLM_FREQ       7.001e-02  4.953e-02   1.413  0.157567
## MVR_PTS        1.042e-01  2.119e-02   4.918  8.73e-07 ***
## CAR_AGE        5.177e-03  1.077e-02   0.481  0.630837
## CAR_USE_Commercial 7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes    -5.337e-01  9.590e-02  -5.565  2.62e-08 ***
## PARENT1_Yes     3.787e-01  1.219e-01   3.106  0.001898 **
## RED_CAR_yes    -1.602e-02  9.656e-02  -0.166  0.868237
## REVOKED_Yes     1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M          -7.591e-03  9.676e-02  -0.078  0.937467
## URBANICITY_Rural -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS       -9.088e-02  1.503e-01  -0.605  0.545393
## INCOME_MISS    -8.443e-02  1.491e-01  -0.566  0.571275
## HOME_VAL_MISS  -1.280e-02  1.414e-01  -0.091  0.927878
## CAR_AGE_MISS    2.667e-01  1.351e-01   1.975  0.048291 *
## TIF_sin        2.893e-02  5.475e-02   0.528  0.597165
## BLUEBOOK_sin   -2.722e-02  4.562e-02  -0.597  0.550670
## AGE_sin        1.864e-02  4.599e-02   0.405  0.685238
## CAR_TYPE_FLAG_BIN -5.584e-01  8.259e-02  -6.760  1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923  8.73e-05 ***
## JOB_TYPE_FLAG_BIN  3.225e-01  9.760e-02   3.304  0.000953 ***
## INCOME_FLAG_BIN  4.796e-01  3.508e-01   1.367  0.171569
## YOJ_FLAG_BIN    8.043e-02  3.797e-01   0.212  0.832256
## HOME_VAL_FLAG_BIN  1.107e-02  1.552e-01   0.071  0.943148
## OLDCLAIM_FLAG_BIN -4.899e-01  1.371e-01  -3.572  0.000354 ***
## CLM_FREQ_FLAG_BIN      NA         NA         NA         NA
## MVR_PTS_FLAG_BIN  2.785e-02  9.490e-02   0.293  0.769170
## CAR_AGE_FLAG_BIN  8.810e-02  1.170e-01   0.753  0.451457
```

```
## TRAVTIME_FLAG_BIN -9.989e-02 1.104e-01 -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7551.7 on 6524 degrees of freedom
## Residual deviance: 5886.1 on 6487 degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Interpretation for the TF_Model1 and TF_Model1_ref *newline*

TF_Model1:

From model 1 summary we can find following important points-

- (i) Variable URBANICITY_Rural has most significant association with lowest p value. negative value of log odd function indicates that chances of accidents are higher in Urbanicity areas compare to rural area.
- (ii) For MSTATUS_Yes variable log odd is negative which indicates married people tend to drive slowly and have less number of accidents.
- (iii) Sex variable has no significant association which means driving patterns does not depend on men and women.
- (iv) variable REVOKED_Yes has strong association which indicates if person's license has been revoked in last 7 years then chance of end up in accidents are much higher with log odds value of 0.809090.
- (v) If person has a claim in last 5 years then chances of more claims are higher. Variable OLDCLAIM_FLAG_BIN indicates that with negative log odds value(1 is here no claim -0.559409).
- (vi) AIC value of the model is AIC: 6078.7 and number of iteration was 5.

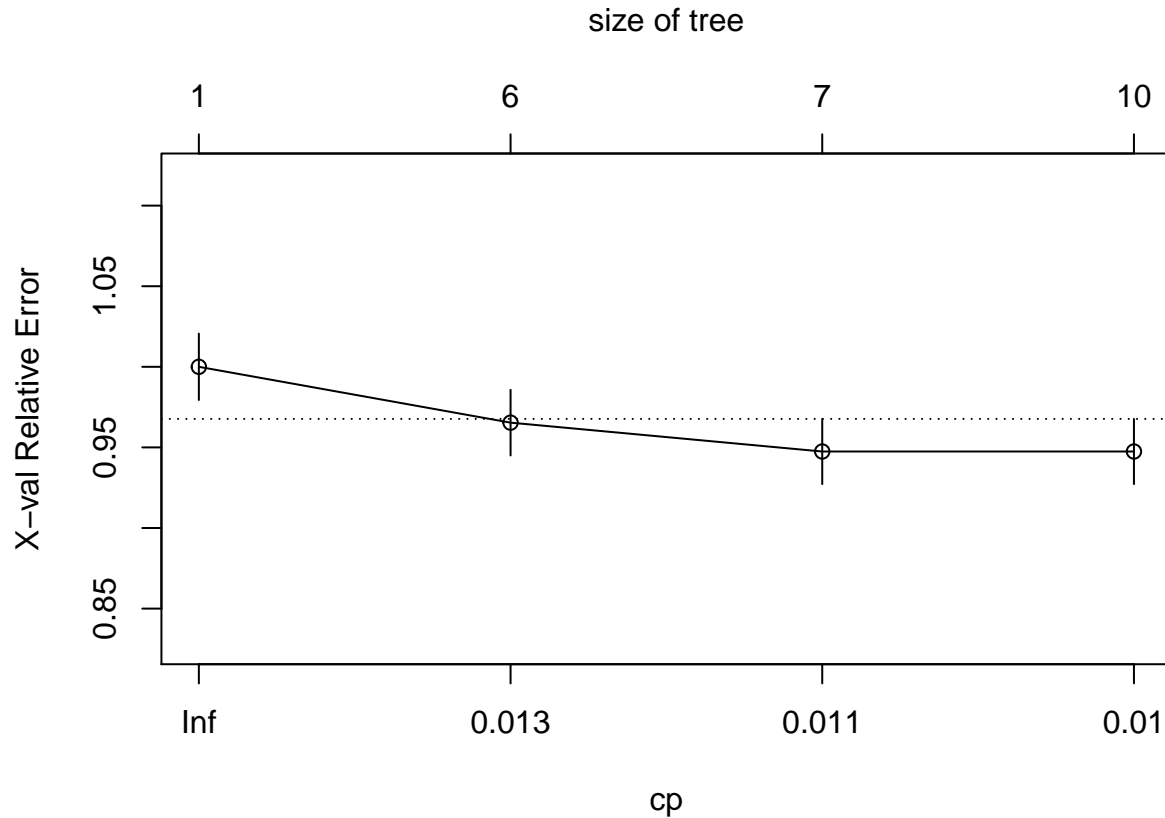
TF_Model1_ref:

Model 1 has AIC value 6078.7 and enhanced model TF_Model1_ref has AIC value 6064.9 slightly better compared to model1. We will look into more details on model1_ref below-

- (i) Based on the outcome from model_ref, it can be seen that following variables KIDSDRIV, PARENT1_Yes, MSTATUS_Yes, CAR_USE_Commercial, REVOKED_Yes, TIF_sin, CAR_TYPE_FLAG_BIN, EDUCATION_FLAG_BIN, JOB_TYPE_FLAG_BIN, INCOME_FLAG_BIN, HOME_VAL_FLAG_BIN, OLDCLAIM_FLAG_BIN, URBANICITY_Rural are only statistically significant. Most of the variables are having similar association as above model 1.
- (ii) As for the statistically significant variables, URBANICITY_Rural has the lowest p-value suggesting a strong association of the URBANICITY_Rural to the target variable. Implication is also same negative value indicate lower chances of accidents in rural areas.
- (iii) One interesting outcome is when childrens are driving your car then more chances of accidents with log odd value of 0.41327 for variable KIDSDRIV.
- (iv) For variable CAR_TYPE_FLAG_BIN there is high negative correlation is there with log odds value of -0.65867 that means Minivan and Panel truck has higher chance of getting into an accident.
- (v) Variable EDUCATION_FLAG_BIN has negative log odds value of -0.46755 indicating that people with higher education above high school has less chance of an accident compared to the other group.
- (vi) No. of iterations are 5 before lowest value of AIC was derived for this model.

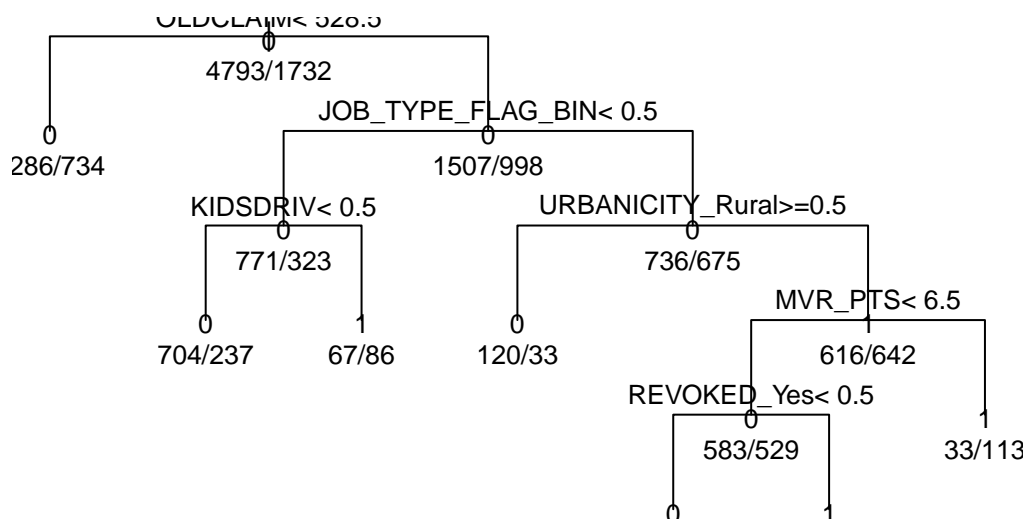
3.3.3 Model 3

In this model, we will be using original variables; however we use the CART (Classification and Regression Trees) algorithm to train the model. We will then Prune the tree and have a look at the summary of this pruned model.



```
##
## Classification tree:
## rpart(formula = TARGET_FLAG ~ ., data = DS_TARGET_FLAG_TRAIN,
##       method = "class")
##
## Variables actually used in tree construction:
## [1] JOB_TYPE_FLAG_BIN KIDSDRIV      MVR_PTS      OLDCLAIM
## [5] REVOKED_Yes      URBANICITY_Rural
##
## Root node error: 1732/6525 = 0.26544
##
## n= 6525
##
##      CP nsplit rel error  xerror   xstd
## 1 0.016051    0  1.00000 1.00000 0.020594
## 2 0.010970    5  0.91975 0.96536 0.020360
## 3 0.010778    6  0.90878 0.94746 0.020235
```

Pruned Classification Tree for TARGET_FLAG



Interpretation for Model 3 *newline*

Following analysis can be drawn from this model: (i) The following variables have been used for classification - OLD_CLAIM, JOB_TYPE_FLAG_BIN, URBANICITY_Rural, KIDSDRIV, MVR_PTS, REVOKED_Yes.

(ii) lowest Cp value and Xerror occurred on split 7.

(iii) OLDCLAIM_FLAG_BIN is the first variable used to split the classification based on its value 0 and 1. When there is claim (1 in above variable) branch is further split to other branches by variable JOB_TYPE_FLAG_BIN (based on value 0 and 1). Based on value of JOB_TYPE_FLAG_BIN (0 and 1) there is two different routes in classification. one Split (774/323) is based on variable KIDSDRIV and the other one (738/675) is based on URBANICITY_Rural variable. Using the above variable total 7 splits have been performed for classification.

3.4 Model Evaluation Using VALID Data

Lets go ahead and apply the above models to the VALID dataset that we had held out. Below is the table of predictions for each of the models:

3.4.1 Evaluation of Model 1

Table 2: Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.447619	0.6460481	0.8269948	0.3926341	0.8079817

Model 1 has good accuracy value close to 78.3%. sensitivity value is lower than the specificity value.

3.4.2 Evaluation of Model 2

Table 3: Model 2 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
2	0.7830882	0.2169118	0.2690476	0.70625	0.7914402	0.2537439	0.8079817

Model 2 has good accuracy value close to 77.3% and very close to model1. sensitivity value is lower than the specificity value.

3.4.3 Evaluation of Model 3

Table 4: Model 3 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.4476190	0.6460481	0.8269948	0.3926341	0.8079817
2	0.7830882	0.2169118	0.2690476	0.7062500	0.7914402	0.2537439	0.8079817
3	0.7561275	0.2438725	0.1666667	0.5932203	0.7688243	0.1451789	0.6733272

This model has accuracy value of 75.4%. AUC for this model is 67.4 % and less compared to the other two models.

3.5 Final Logistic Model Selection Summary

Following is the comparison of various metrics for above 3 models

Table 5: Model Performance Metrics Comparison

Model_No	Accuracy	Error_Rate	AUC	Precision	sensitivity	specificity	F1_Score
1	0.7947304	0.2052696	0.8079817	0.4476190	0.6460481	0.8269948	0.3926341
2	0.7830882	0.2169118	0.8079817	0.2690476	0.7062500	0.7914402	0.2537439
3	0.7561275	0.2438725	0.6733272	0.1666667	0.5932203	0.7688243	0.1451789

From the comparison table, we see that Model 1 is quite superior from the accuracy and AUC perspective.

The AUC provides the best score on probability of correctly identifying the patterns at various cut off values. The Accuracy, on the other hand, is calculated as specific cut off value. For this assignment we will go with cut off value of 0.5 and choose the Model 1 based on Accuracy value for further prediction on evaluation data set.

3.5.1 Detailed Inference for Final Model

The following analysis will be carried out on the final model:

- (i) Relevant variables in the model
- (ii) Estimate confidence interval for coefficient
- (iii) odds ratios and 95% CI
- (iv) AUC curve
- (v) Distribution of prediction

3.5.2 Most important variables in the model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.988e-01  3.403e-01  -1.172  0.241181
## KIDSDRIV       3.979e-01  6.763e-02   5.883  4.03e-09 ***
## AGE           -4.759e-03  4.458e-03  -1.067  0.285793
## HOMEKIDS       1.818e-02  4.169e-02   0.436  0.662766
## YOJ            1.598e-02  1.378e-02   1.159  0.246305
## INCOME        -3.058e-06  1.260e-06  -2.427  0.015231 *
## HOME_VAL      -8.986e-07  6.553e-07  -1.371  0.170277
## TRAVTIME       1.334e-02  2.786e-03   4.786  1.70e-06 ***
## BLUEBOOK      -1.680e-05  4.762e-06  -3.527  0.000420 ***
## TIF           -5.153e-02  9.256e-03  -5.567  2.59e-08 ***
## OLDCLAIM      -2.124e-05  4.735e-06  -4.487  7.23e-06 ***
## CLM_FREQ       7.001e-02  4.953e-02   1.413  0.157567
## MVR_PTS        1.042e-01  2.119e-02   4.918  8.73e-07 ***
## CAR_AGE        5.177e-03  1.077e-02   0.481  0.630837
## CAR_USE_Commercial 7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes    -5.337e-01  9.590e-02  -5.565  2.62e-08 ***
## PARENT1_Yes     3.787e-01  1.219e-01   3.106  0.001898 **
## RED_CAR_yes    -1.602e-02  9.656e-02  -0.166  0.868237
## REVOKED_Yes     1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M          -7.591e-03  9.676e-02  -0.078  0.937467
## URBANICITY_Rural -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS       -9.088e-02  1.503e-01  -0.605  0.545393
## INCOME_MISS    -8.443e-02  1.491e-01  -0.566  0.571275
## HOME_VAL_MISS  -1.280e-02  1.414e-01  -0.091  0.927878
## CAR_AGE_MISS    2.667e-01  1.351e-01   1.975  0.048291 *
## TIF_sin        2.893e-02  5.475e-02   0.528  0.597165
## BLUEBOOK_sin   -2.722e-02  4.562e-02  -0.597  0.550670
## AGE_sin        1.864e-02  4.599e-02   0.405  0.685238
## CAR_TYPE_FLAG_BIN -5.584e-01  8.259e-02  -6.760  1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923  8.73e-05 ***
## JOB_TYPE_FLAG_BIN  3.225e-01  9.760e-02   3.304  0.000953 ***
```

```
## INCOME_FLAG_BIN      4.796e-01  3.508e-01  1.367 0.171569
## YOJ_FLAG_BIN         8.043e-02  3.797e-01  0.212 0.832256
## HOME_VAL_FLAG_BIN    1.107e-02  1.552e-01  0.071 0.943148
## OLDCLAIM_FLAG_BIN   -4.899e-01  1.371e-01  -3.572 0.000354 ***
## CLM_FREQ_FLAG_BIN      NA         NA         NA         NA
## MVR_PTS_FLAG_BIN     2.785e-02  9.490e-02  0.293 0.769170
## CAR_AGE_FLAG_BIN     8.810e-02  1.170e-01  0.753 0.451457
## TRAVTIME_FLAG_BIN   -9.989e-02  1.104e-01  -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 7551.7  on 6524  degrees of freedom
## Residual deviance: 5886.1  on 6487  degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Following are the most relevant variables for the model: CAR_USE_Commercial, REVOKED_Yes, URBANICITY_Rural, CAR_TYPE_FLAG_BIN, TRAVTIME, OLDCLAIM_FLAG_BIN, MVR_PTS, TIF_SIN, KIDSDRIV, PARENT1_Yes, EDUCATION_FLAG_BIN, MSTATUS_Yes, BLUEBOOK, OLDCLAIM, JOB_TYPE_FLAG_BIN, HOME_VAL, INCOME_FLAG_BIN.

we can write the equation of the Model 1 as:

$$\log(y) = -0.4015 + 0.3431 * KIDSDRIV - 0.000001027 * HOME_VAL + 0.01557 * TRAVTIME - 0.00001858 * BLUEBOOK - 0.05103 * TIF - 0.00001873 * OLDCLAIM + 0.1043 * MVR_PTS + 0.7632 * CAR_USE_Commercial - 0.4066 * MSTATUS_Yes + 0.5246 * PARENT1_Yes + 1.025 * REVOKED_Yes - 2.221 * URBANICITY_Rural - 0.5662 * CAR_TYPE_FLAG_BIN - 0.3996 * EDUCATION_FLAG_BIN + 0.3584 * JOB_TYPE_FLAG_BIN + 0.311 * INCOME_FLAG_BIN - 0.627 * OLDCLAIM_FLAG_BIN$$

3.5.3 Analysis of odds ratios of variables 95% CI

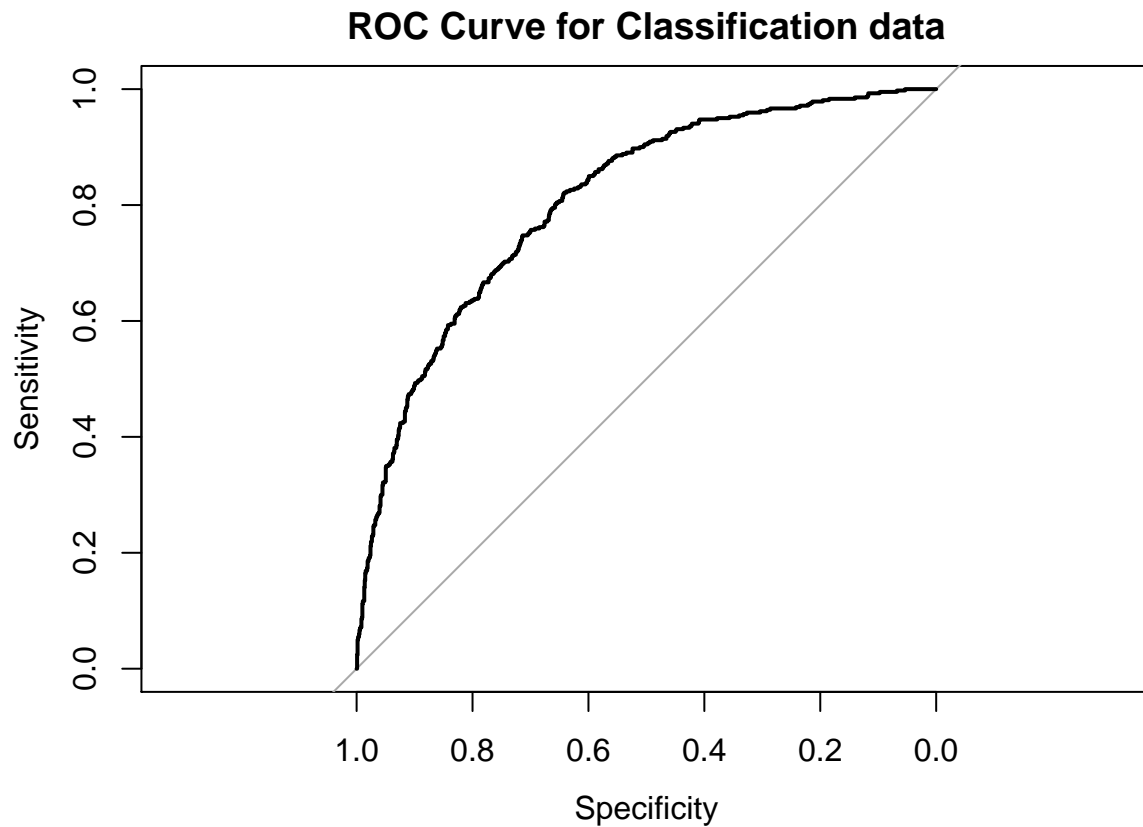
##	OR	2.5 %	97.5 %
## (Intercept)	0.67113058	0.34449416	1.3074714
## KIDSDRIV	1.48862259	1.30382382	1.6996140
## AGE	0.99525237	0.98659334	1.0039874
## HOMEKIDS	1.01834688	0.93844701	1.1050495
## YOJ	1.01610813	0.98902600	1.0439318
## INCOME	0.99999694	0.99999447	0.9999994
## HOME_VAL	0.99999910	0.99999782	1.0000004
## TRAVTIME	1.01342437	1.00790509	1.0189739
## BLUEBOOK	0.99998320	0.99997387	0.9999925
## TIF	0.94977489	0.93269993	0.9671624
## OLDCLAIM	0.99997876	0.99996948	0.9999880
## CLM_FREQ	1.07251612	0.97328473	1.1818647
## MVR_PTS	1.10986959	1.06470918	1.1569455
## CAR_AGE	1.00519015	0.98418928	1.0266391
## CAR_USE_Commercial	2.11945792	1.82459644	2.4619701
## MSTATUS_Yes	0.58643036	0.48593986	0.7077019
## PARENT1_Yes	1.46035120	1.14992968	1.8545705
## RED_CAR_yes	0.98410869	0.81442639	1.1891436
## REVOKED_Yes	2.86451319	2.33977796	3.5069293

## SEX_M	0.99243744	0.82099263	1.1996844
## URBANICITY_Rural	0.09894387	0.07738989	0.1265009
## YOJ_MISS	0.91313153	0.68015614	1.2259085
## INCOME_MISS	0.91903450	0.68611221	1.2310296
## HOME_VAL_MISS	0.98728640	0.74837344	1.3024707
## CAR_AGE_MISS	1.30565907	1.00200553	1.7013335
## TIF_sin	1.02935564	0.92462226	1.1459523
## BLUEBOOK_sin	0.97314416	0.88991133	1.0641617
## AGE_sin	1.01881446	0.93100276	1.1149085
## CAR_TYPE_FLAG_BIN	0.57214143	0.48662962	0.6726796
## EDUCATION_FLAG_BIN	0.68635498	0.56871853	0.8283239
## JOB_TYPE_FLAG_BIN	1.38053840	1.14018363	1.6715608
## INCOME_FLAG_BIN	1.61542333	0.81225230	3.2127856
## YOJ_FLAG_BIN	1.08375386	0.51487406	2.2811839
## HOME_VAL_FLAG_BIN	1.01113208	0.74588194	1.3707103
## OLDCLAIM_FLAG_BIN	0.61268865	0.46828247	0.8016259
## CLM_FREQ_FLAG_BIN	NA	NA	NA
## MVR_PTS_FLAG_BIN	1.02824027	0.85372369	1.2384312
## CAR_AGE_FLAG_BIN	1.09209197	0.86830795	1.3735506
## TRAVTIME_FLAG_BIN	0.90493543	0.72885274	1.1235577

The following points can be made for the important variables in the model:

In keeping all other variables same, the odds of an accident increases as follow: 1.8449962 for per unit change in CAR_USE_Commercial, 2.2458626 per unit change in REVOKED_Yes, 0.1104633 for per unit change in URBANICITY_Rural, etc. Any value which is less than 1, it means that there is less chance of an event with the per unit increase of the variable.

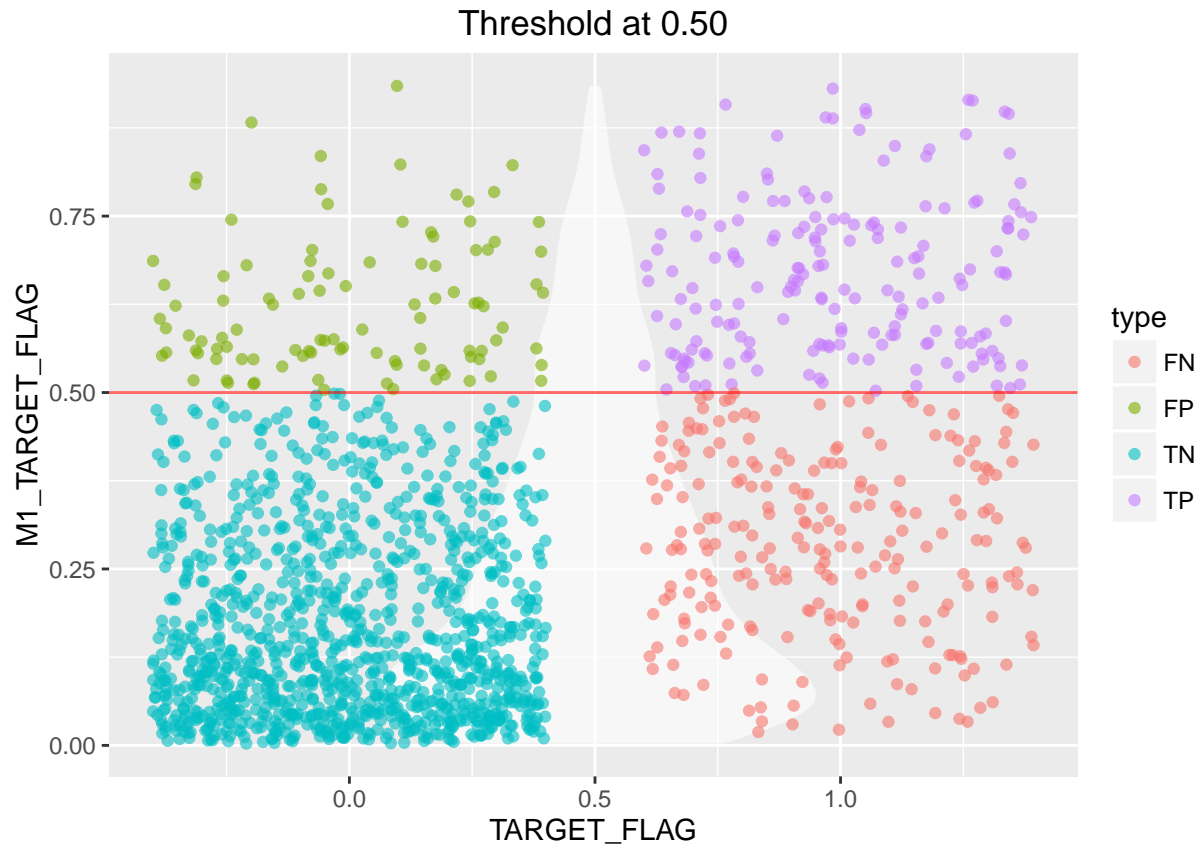
3.5.4 ROC curve for the selected model



```
##  
## Call:  
## roc.formula(formula = DS_TARGET_FLAG_VALID$TARGET_FLAG ~ DS_TARGET_FLAG_VALID$M1_TARGET_FLAG,      da  
##  
## Data: DS_TARGET_FLAG_VALID$M1_TARGET_FLAG in 1212 controls (DS_TARGET_FLAG_VALID$TARGET_FLAG 0) < 42  
## Area under the curve: 0.808
```

3.5.5 Distribution of the Predictions

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```



Considering the target has value 1 (accident occurs) and 0 when no accident, then the above plot illustrates the tradeoff of choosing a reasonable threshold. In other words, if the threshold is increased, the number of false positive (FP) results is lowered; while the number of false negative (FN) results increases.

4 Linear Regression for TARGET_AMT

In this section we will use Linear regression to model the TARGET_AMT. We will first start with the Data Exploration. We will be using only those records where the TARGET_FLAG is 1. This indicates that the vehicle crashed. In such a scenario, we will be modeling the cost of repair using Linear Regression. First, lets create the required data set for the “Crashed” data from the existing “clean” full data and look at the structure of the resulting dataset. We will remove from the new “crashed” dataset all those variables that were created specifically for predicting TARGET_FLAG. We will be creating these variables separately for predicting TARGET_AMT.

We notice that the dependent variable here is TARGET_AMT. Apart from the dependent variables, we have 49 independent or predictor variables.

Also, since we created this dataset from the “Clean” full dataset, we already have taken care of the missing values.

However, we may need to look into the outliers and correlations again since we have a new target variable to correlate against.

4.1 Data Summary and Correlation Analysis

4.1.1 Data Summary

In this section, we will create summary data to better understand the relationship each of the variables have with our dependent variables using correlation, central tendency, and dispersion as shown below:

4.1.2 Correlations

Now we will produce the correlation table between the independent variables and the dependent variable - TARGET_AMT

Table 6: Correlation between TARGET_AMT and predictor variables

	Correlation_TARGET_AMT
TARGET_AMT	1.000000
BLUEBOOK	0.1181297
CAR_TYPE_Panel.Truck	0.0682806
SEX_M	0.0513430
CAR_TYPE_Van	0.0499290
CAR_USE_Commercial	0.0496142
INCOME	0.0454604
JOB_Professional	0.0406747
JOB_Unknown	0.0402613
MVR_PTS	0.0396710
YOJ	0.0343807
EDUCATION_PhD	0.0294767
HOME_VAL	0.0291859
AGE	0.0279828
RED_CAR_yes	0.0271768
PARENT1_Yes	0.0238302
JOB_Blue.Collar	0.0155259
EDUCATION_Masters	0.0143267
EDUCATION_Bachelors	0.0136662

	Correlation_TARGET_AMT
JOB_Lawyer	0.0102382
TRAVTIME	0.0053657
CLM_FREQ	0.0023251
INDEX	0.0010044
HOMEKIDS	0.0002698
KIDSDRIV	-0.0000869
URBANICITY_Rural	-0.0048888
OLDCLAIM	-0.0049723
CAR_TYPE_Minivan	-0.0058234
TIF	-0.0060620
JOB_Doctor	-0.0122018
CAR_AGE	-0.0135348
JOB_Clerical	-0.0151891
CAR_TYPE_Sports.Car	-0.0152654
CAR_TYPE_Pickup	-0.0174060
JOB_Manager	-0.0256129
JOB_Home.Maker	-0.0293974
JOB_Student	-0.0331511
MSTATUS_Yes	-0.0351848
EDUCATION_High.School	-0.0359529
REVOKED_Yes	-0.0365018
CAR_TYPE_SUV	-0.0405600

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_AMT.

However, BLUEBOOK, CAR_TYPE_Panel.Truck, SEX_M, CAR_TYPE_Van, CAR_USE_Commercial, INCOME, INCOME_IMPUTE, JOB_Professional, JOB_Unknown, MVR_PTS, YOJ, YOJ_IMPUTE, HOME_VAL_IMPUTE, EDUCATION_PhD, HOME_VAL, AGE, AGE_IMPUTE, RED_CAR_yes, PARENT1_Yes, YOJ_MISS, HOME_VAL_MISS, JOB_Blue.Collar, EDUCATION_Masters, EDUCATION_Bachelors, JOB_Lawyer, TRAVTIME, CLM_FREQ, HOMEKIDS have a positive correlation.

Similarly, KIDSDRIV, TRAVTIME_FLAG_BIN, INCOME_MISS, URBANICITY_Rural, OLDCLAIM, CAR_TYPE_Minivan, TIF, CAR_AGE_MISS, JOB_Doctor, CAR_AGE, CAR_AGE_IMPUTE, JOB_Clerical, CAR_TYPE_Sports.Car, CAR_TYPE_Pickup, JOB_Manager, JOB_Home.Maker, JOB_Student, MSTATUS_Yes, EDUCATION_High.School, REVOKED_Yes, CAR_TYPE_SUV have a negative correlation.

Lets now see how values in some of the variable affects the correlation:

CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

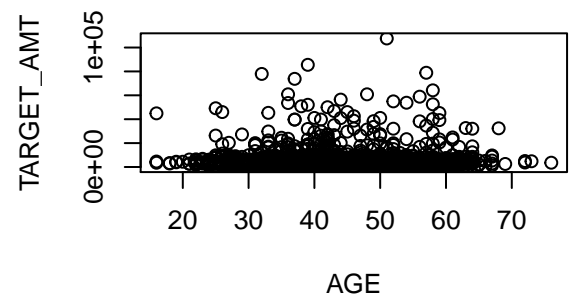
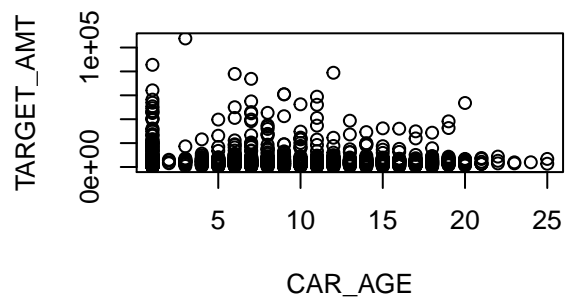
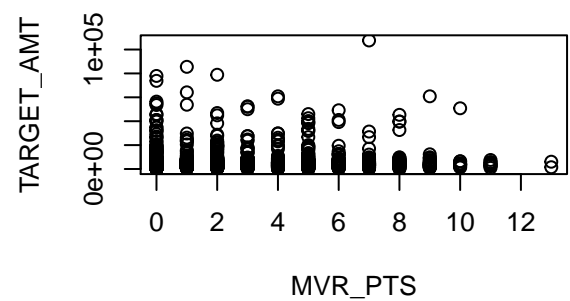
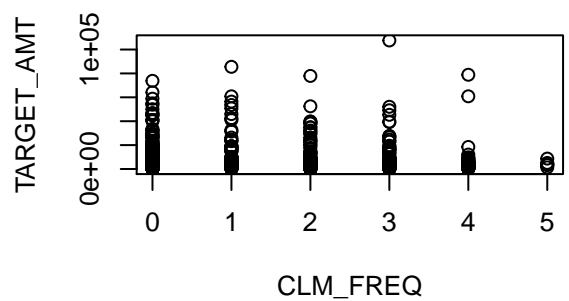
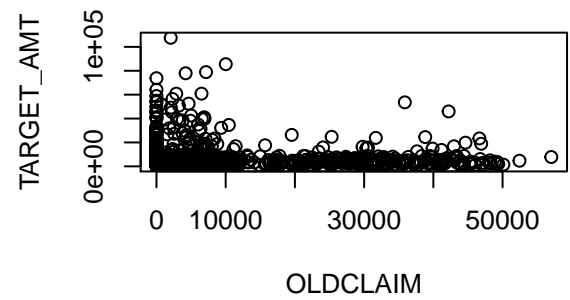
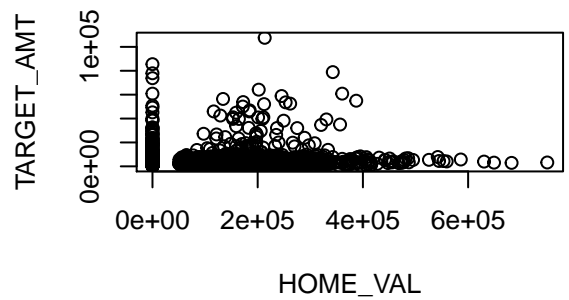
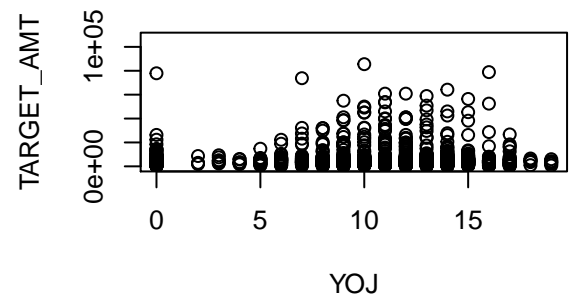
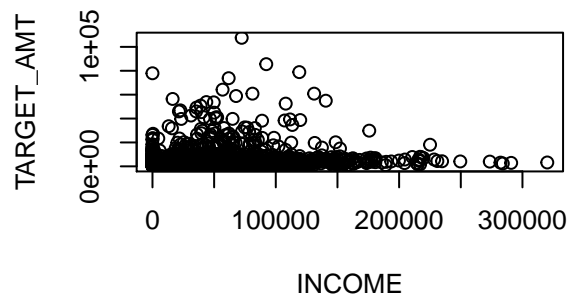
EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

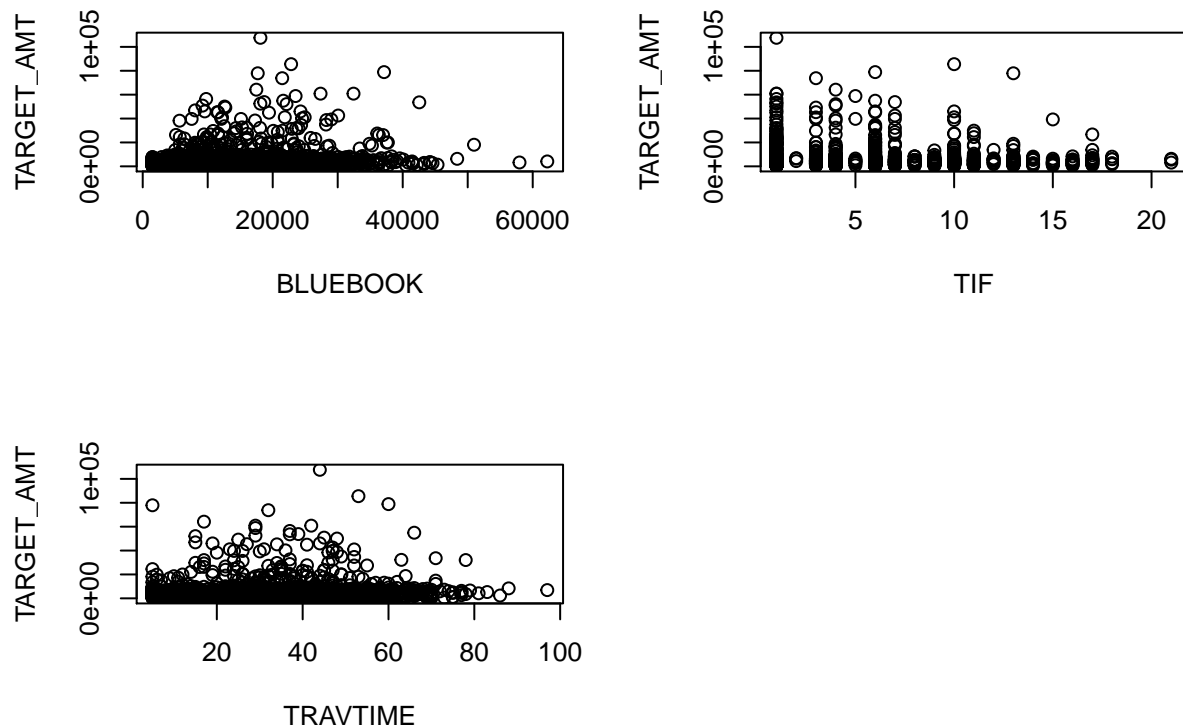
JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation.

4.1.3 Binning of Variables

Lets have a look at the following numeric variables that have 0 as one of their values: INCOME, YOJ, HOME_VAL, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, AGE, BLUEBOOK, TIF, TRAVTIME.

The goal here is to see if we can bin these variables into zero and non-zero bin values and check the correlations. While doing that we will also see how the variables are distributed vis-a-vis TARGET_AMT.





From the outputs above, we can come to the following conclusions:

- INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.
- OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, > 10000, We will use these values to create 2 bins.
- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original variable as is.
- AGE - There is no specific pattern that emerges in AGE. We will retain the variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain the variable as is.
- TIF - Looking at the plot we can conclude that this is not a good variable for binning. We will retain this variable as is.

- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

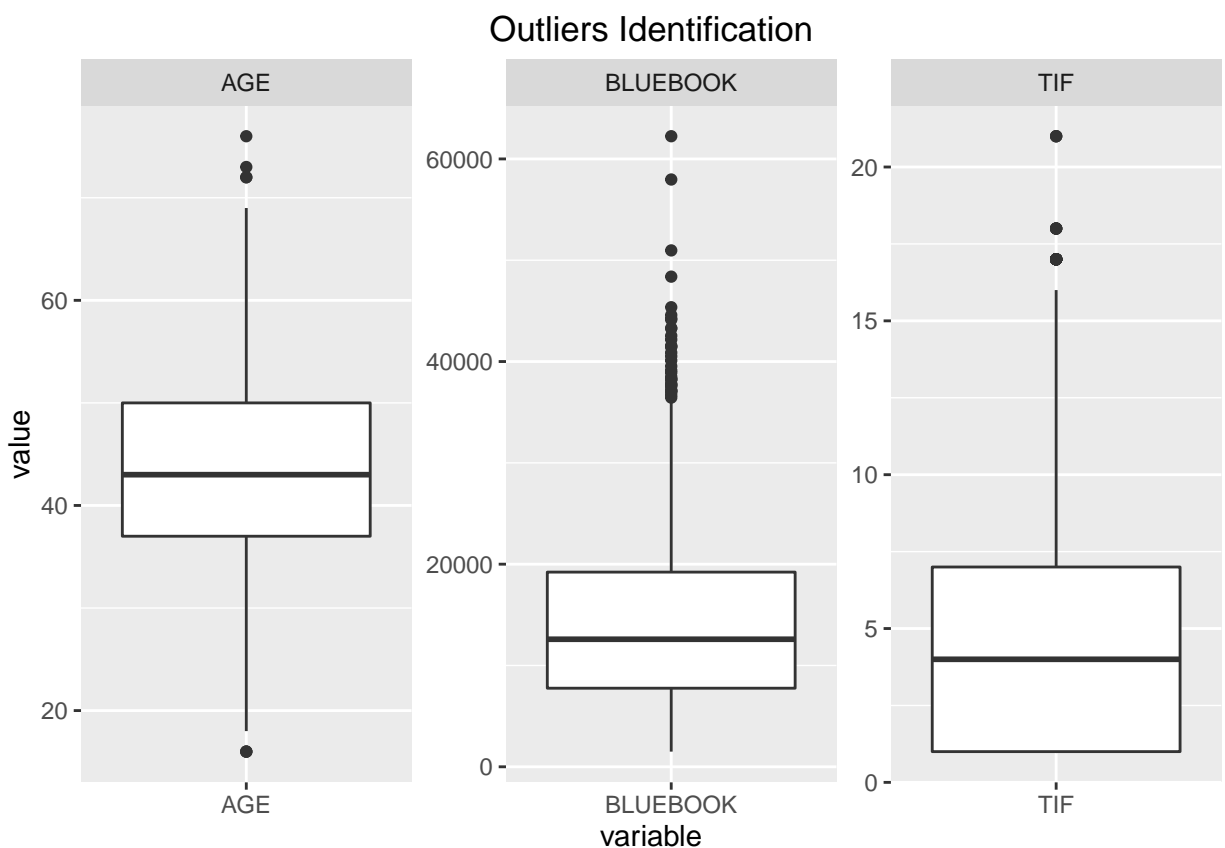
We will carry out the above transformations in the Data Preparation phase.

4.1.4 Outliers identification

In this sub-section, we will look at the boxplots and determine the outliers in variables and decide on whether to act on the outliers.

We will do the outliers only on the numeric variables: AGE, BLUEBOOK and TIF. The other variables will be binned and would not need outlier handling.

Below are the plots:



From the “Outliers identification” plot above, we see that we have few outliers that we need to treat.

We see that all the 3 variables need to be treated when we do the data preparation for modeling the TARGET_AMT.

4.2 Data Preparation

Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

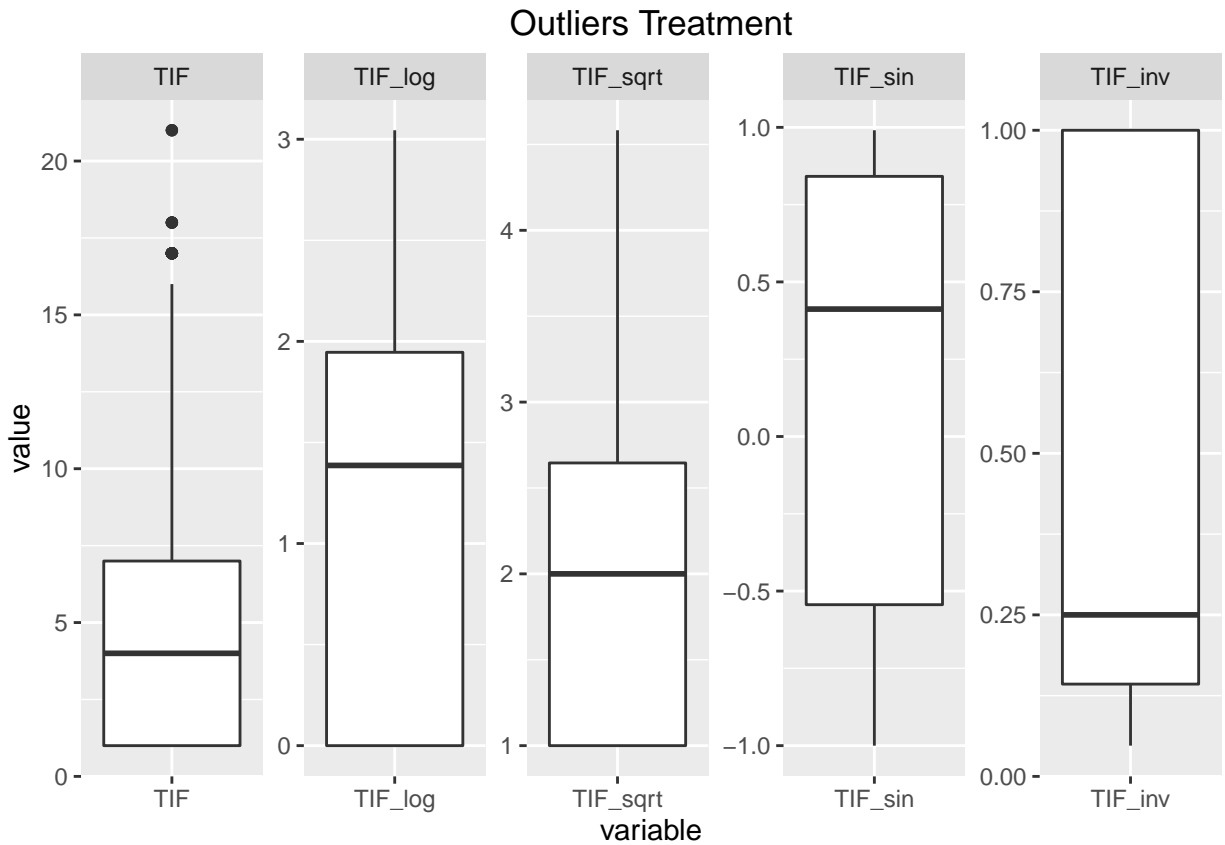
We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

In this sub-section, we will check different transformations for each of the variables - AGE, BLUEBOOK, TIF - and create the appropriate outlier-handled / transformed variables.

The figure displays five box plots for the variable AGE, each representing a different transformation: AGE, AGE_log, AGE_sqrt, AGE_sin, and AGE_inv. The y-axis is labeled 'value' and the x-axis is labeled 'variable'. Each plot shows the median, quartiles, and outliers.

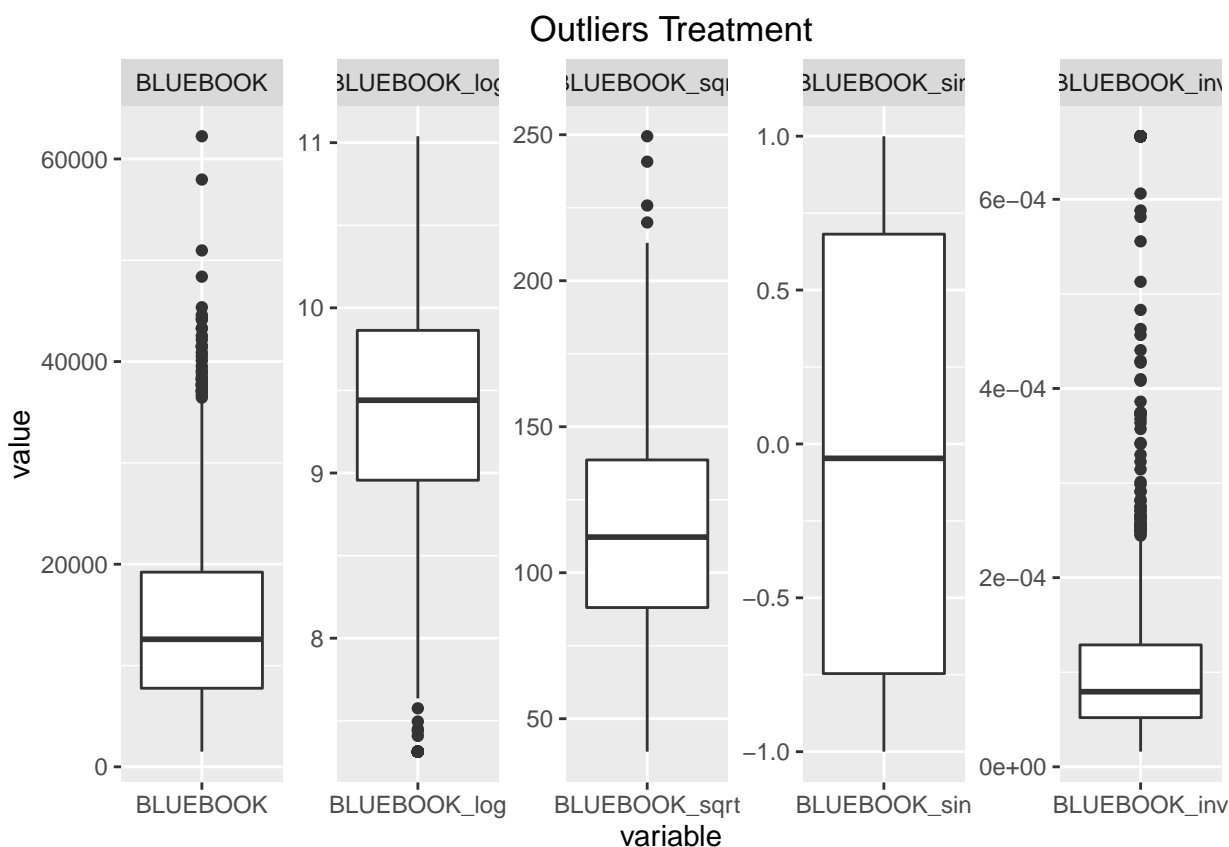
- AGE:** The median is approximately 43. The interquartile range (IQR) is from about 37 to 50. Whiskers extend from approximately 18 to 68. Outliers are present at approximately 15, 75, 78, and 80.
- AGE_log:** The median is approximately 3.7. The IQR is from about 3.6 to 3.9. Whiskers extend from approximately 3.2 to 4.2. Outliers are present at approximately 2.8, 2.9, 3.0, 3.1, 3.2, and 3.3.
- AGE_sqrt:** The median is approximately 6.5. The IQR is from about 6.1 to 7.1. Whiskers extend from approximately 4.8 to 8.5. Outliers are present at approximately 4.0, 4.2, 4.3, 4.4, 4.5, and 8.8.
- AGE_sin:** The median is 0.0. The IQR is from about -0.6 to 0.8. Whiskers extend from approximately -1.0 to 1.0. There are no outliers.
- AGE_inv:** The median is approximately 0.023. The IQR is from about 0.020 to 0.027. Whiskers extend from approximately 0.018 to 0.035. Outliers are present at approximately 0.038, 0.040, 0.042, 0.044, 0.046, 0.048, 0.050, 0.052, 0.054, 0.056, and 0.060.

Transformations for TIF



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, a sin transformation seems to be more appropriate as it is well centered. Hence, We will create these variables.

Transformations for BLUEBOOK



From the above charts we can see that a sin transformation works well for BLUEBOOK. We will create these variables.

4.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_AMT.

The following were some of the observations we made during the data exploration phase for TARGET_AMT

CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distinction is clear, we believe that binning this variable accordingly will help strengthen the correlation. Accordingly, we will bin these variables as below:

CAR_TYPE_AMT_BIN :

- 1 : if CAR_TYPE is Vans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Minivans

EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation. Accordingly, we will bin these variables as below:

EDUCATION_AMT_BIN :

- 1 : if EDUCATION is High School
- 0 : if EDUCATION is Bachelors, Masters or Phd

JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation. Accordingly, we will bin these variables as below:

JOB_TYPE_AMT_BIN :

- 1 : if JOB_TYPE is Lawyer, Professional, Unknown or in a Blue Collar
- 0 : if JOB_TYPE is Doctor, Manager, Home Maker, Student, or Clerical

INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.

INCOME_AMT_BIN :

- 1 : if INCOME \leq 125000
- 0 : if INCOME $>$ 125000
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.

YOJ_AMT_BIN :

- 1 : if YOJ \geq 7 and YOJ \leq 17
- 0 : ELSE 0
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.

HOME_VAL_AMT_0_10K_BIN :

- 1 : if HOME_VAL \geq 0 and HOME_VAL \leq 10000
- 0 : ELSE 0

HOME_VAL_AMT_60K_400K_BIN :

- 1 : if HOME_VAL \geq 60000 and HOME_VAL \leq 400000
- 0 : ELSE 0

OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, $>$ 10000, We will use these values to create 2 bins.

OLDCLAIM_AMT_0_2K_BIN :

- 1 : if OLDCLAIM \geq 0 and OLDCLAIM \leq 2000
- 0 : ELSE 0

OLDCLAIM_AMT_2K_10K_BIN :

- 1 : if OLDCLAIM \geq 2000 and OLDCLAIM \leq 10000
- 0 : ELSE 0

- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.

CLM_FREQ_AMT_BIN :

- 1 : if CLM_FREQ < 4
- 0 : if CLM_FREQ >= 4
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.

MVR_PTS_AMT_BIN :

- 1 : if MVR_PTS <=2
- 0 : if MVR_PTS > 0
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original variable as is.

CAR_AGE_AMT_BIN :

- 1 : if CAR_AGE <= 1
- 0 : if CAR_AGE > 0
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

TRAVTIME_AMT_BIN :

- 1 : if TRAVTIME <= 20
- 0 : if TRAVTIME > 0

4.3 Build Models

Now that we have the dataset in a shape that can be modeled, we will go ahead and train the model for TARGET_AMT. We will train 2 models and select the best among these 2 models. The following will be the model specifications:

- Model1 (All Variables in Linear Dataset) - This will use the standard lm for building the model. We will use all available variables.
- Model2 - (A few selected variables) - This will use the standard lm for building the model. However, we will use only a few selected variables that seemed to have a good correlation with TARGET_AMT.

4.3.1 Model 1

In this model, we will be using the standard lm modeling technique. We will use the entire set of variables from the Linear dataset.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9677   -3249   -1387    792   75391
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.005e+03  6.472e+03   0.310  0.75674
## KIDSDRIV      -1.683e+01  1.782e+03  -0.009  0.99247
## AGE           1.114e+01  2.565e+01   0.434  0.66415
## HOMEKIDS       3.102e+02  4.101e+02   0.756  0.44956
## YOJ           -1.288e+02  1.189e+02  -1.083  0.27895
## INCOME        -1.149e-02  1.083e-02  -1.061  0.28897
## HOME_VAL       1.941e-03  4.389e-03   0.442  0.65832
## TRAVTIME       5.253e+00  1.641e+01   0.320  0.74890
## BLUEBOOK       1.528e-01  3.399e-02   4.496 7.41e-06 ***
## TIF           -4.696e+01  1.116e+02  -0.421  0.67391
## OLDCLAIM       1.300e-02  4.408e-02   0.295  0.76809
## CLM_FREQ      -5.938e+02  3.286e+02  -1.807  0.07095 .
## MVR_PTS        6.051e+02  2.594e+02   2.332  0.01980 *
## CAR_AGE       -1.689e+02  7.621e+01  -2.217  0.02678 *
## CAR_USE_Commercial  6.193e+01  5.625e+02   0.110  0.91235
## MSTATUS_Yes    -1.815e+03  6.199e+02  -2.928  0.00346 **
## PARENT1_Yes    -7.486e+02  7.574e+02  -0.988  0.32309
## RED_CAR_Yes    -1.314e+02  5.557e+02  -0.236  0.81308
## REVOKED_Yes    -1.130e+03  6.178e+02  -1.830  0.06749 .
## SEX_M          1.924e+03  7.219e+02   2.666  0.00776 **
## URBANICITY_Rural -2.500e+02  8.220e+02  -0.304  0.76106
## EDUCATION_Bachelors -3.466e+03  1.323e+03  -2.620  0.00887 **
## EDUCATION_High.School -4.125e+03  1.420e+03  -2.904  0.00373 **
## EDUCATION_Masters -2.236e+03  1.083e+03  -2.065  0.03911 *
## JOB_Blue.Collar  8.884e+02  1.311e+03   0.677  0.49824
## JOB_Clerical    -1.907e+02  1.376e+03  -0.139  0.88981
## JOB_Doctor      -3.153e+03  1.885e+03  -1.673  0.09460 .
## JOB_Home.Maker  -1.200e+02  1.503e+03  -0.080  0.93638
## JOB_Lawyer      -3.472e+02  1.156e+03  -0.300  0.76396
## JOB_Manager     -9.609e+02  1.227e+03  -0.783  0.43367
## JOB_Professional  1.427e+03  1.289e+03   1.107  0.26827
## JOB_Student      2.542e+01  1.563e+03   0.016  0.98703
## CAR_TYPE_Minivan  1.140e+02  8.588e+02   0.133  0.89438
## CAR_TYPE_Panel.Truck -6.953e+01  9.576e+02  -0.073  0.94213
## CAR_TYPE_Pickup  5.865e+02  8.258e+02   0.710  0.47769
## CAR_TYPE_Sports.Car  2.203e+03  1.115e+03   1.975  0.04839 *
## CAR_TYPE_SUV     1.972e+03  1.047e+03   1.882  0.05998 .
## AGE_sin        -1.392e+01  2.635e+02  -0.053  0.95788
## TIF_sin        -2.987e+02  3.791e+02  -0.788  0.43086
```

```
## BLUEBOOK_sin          1.310e+02  2.608e+02   0.502  0.61557
## INCOME_AMT_BIN        -5.870e+02  1.278e+03  -0.459  0.64598
## YOJ_AMT_BIN           1.765e+03  1.017e+03   1.735  0.08296 .
## HOME_VAL_AMT_0_10K_BIN 2.353e+03  1.959e+03   1.201  0.22985
## HOME_VAL_AMT_60K_400K_BIN 2.907e+03  1.496e+03   1.944  0.05212 .
## OLDCLAIM_AMT_0_2K_BIN  -2.182e+03  1.453e+03  -1.501  0.13344
## OLDCLAIM_AMT_2K_10K_BIN -4.936e+02  1.139e+03  -0.434  0.66466
## CLM_FREQ_AMT_BIN      -1.331e+03  1.183e+03  -1.125  0.26082
## MVR_PTS_AMT_BIN        1.740e+03  7.970e+02   2.183  0.02918 *
## CAR_AGE_AMT_BIN        -6.829e+02  7.036e+02  -0.971  0.33190
## TRAVTIME_AMT_BIN        7.007e+01  6.639e+02   0.106  0.91596
## KIDSDRIV_AMT_BIN_0     -9.347e+02  1.928e+03  -0.485  0.62787
## KIDSDRIV_AMT_BIN_1      2.079e+03  2.355e+03   0.883  0.37742
## HOMEKIDS_AMT_BIN_0     -5.618e+02  9.971e+02  -0.563  0.57318
## HOMEKIDS_AMT_BIN_3      8.051e+02  1.428e+03   0.564  0.57285
## YOJ_AMT_BIN_0_AND_9To14 6.615e+02  7.114e+02   0.930  0.35259
## INCOME_AMT_BIN_MISS_0  -2.171e+02  1.778e+03  -0.122  0.90283
## TIF_AMT_BIN_6          -1.557e+02  8.834e+02  -0.176  0.86015
## OLDCLAIM_AMT_BIN_MISS_0 4.083e+02  1.026e+03   0.398  0.69059
## MVR_PTS_AMT_BIN_0      4.155e+02  6.288e+02   0.661  0.50888
## MVR_PTS_AMT_BIN_5      1.819e+03  1.083e+03   1.680  0.09312 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7577 on 1642 degrees of freedom
## Multiple R-squared:  0.05812,    Adjusted R-squared:  0.02428
## F-statistic: 1.717 on 59 and 1642 DF,  p-value: 0.0006848
```

Interpretation of the Model

Based on the Model output, below are the characteristics of the refined model :

- The Residual standard error is 7577
- Multiple R-squared: 0.05812
- Adjusted R-squared: 0.02428
- F-statistic: 1.717 on 59 and 1642 DF
- p-value: < 0.0006848

4.3.2 Model 2

In this model, we will be using the standard lm modeling technique. We will use only those variables that seemed to have a good correlation with TARGET_AMT.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = DS_SELECTED_VARS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9665  -3182  -1427    758   75206
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

## (Intercept)	-8.686e+02	6.126e+03	-0.142	0.88725	
## AGE	1.359e+01	2.550e+01	0.533	0.59414	
## BLUEBOOK	1.533e-01	3.261e-02	4.702	2.8e-06	***
## CAR_AGE	-1.077e+02	7.287e+01	-1.478	0.13958	
## CAR_AGE_AMT_BIN	-2.792e+02	6.877e+02	-0.406	0.68485	
## CAR_USE_Commercial	-1.467e+01	5.460e+02	-0.027	0.97857	
## CLM_FREQ	-6.803e+02	2.613e+02	-2.603	0.00932	**
## CLM_FREQ_AMT_BIN	-1.586e+03	1.128e+03	-1.406	0.15978	
## HOME_VAL	1.886e-03	4.376e-03	0.431	0.66661	
## HOME_VAL_AMT_0_10K_BIN	2.213e+03	1.955e+03	1.132	0.25778	
## HOME_VAL_AMT_60K_400K_BIN	2.762e+03	1.492e+03	1.851	0.06430	.
## HOMEKIDS	2.804e+02	4.087e+02	0.686	0.49275	
## HOMEKIDS_AMT_BIN_0	-5.455e+02	9.930e+02	-0.549	0.58280	
## HOMEKIDS_AMT_BIN_3	6.281e+02	1.424e+03	0.441	0.65922	
## INCOME	-8.304e-03	1.067e-02	-0.778	0.43654	
## INCOME_AMT_BIN	-1.080e+03	1.259e+03	-0.858	0.39089	
## KIDSDRIV	8.219e+01	1.779e+03	0.046	0.96315	
## KIDSDRIV_AMT_BIN_0	-8.472e+02	1.925e+03	-0.440	0.65983	
## KIDSDRIV_AMT_BIN_1	2.178e+03	2.351e+03	0.926	0.35447	
## MSTATUS_Yes	-1.760e+03	6.180e+02	-2.848	0.00445	**
## MVR_PTS	6.149e+02	2.590e+02	2.374	0.01769	*
## MVR_PTS_AMT_BIN	1.798e+03	7.952e+02	2.261	0.02387	*
## MVR_PTS_AMT_BIN_0	3.765e+02	6.280e+02	0.600	0.54890	
## MVR_PTS_AMT_BIN_5	1.901e+03	1.082e+03	1.757	0.07910	.
## OLDCLAIM	6.387e-03	4.365e-02	0.146	0.88368	
## OLDCLAIM_AMT_0_2K_BIN	-2.130e+03	1.385e+03	-1.538	0.12431	
## OLDCLAIM_AMT_2K_10K_BIN	-6.152e+02	1.135e+03	-0.542	0.58796	
## PARENT1_Yes	-6.825e+02	7.537e+02	-0.905	0.36538	
## RED_CAR_yes	-1.771e+02	5.530e+02	-0.320	0.74888	
## REVOKED_Yes	-1.065e+03	6.139e+02	-1.734	0.08303	.
## SEX_M	1.898e+03	7.037e+02	2.697	0.00707	**
## TIF	2.202e+01	8.168e+01	0.270	0.78754	
## TIF_AMT_BIN_6	3.152e+02	7.370e+02	0.428	0.66892	
## TRAVTIME	6.230e+00	1.640e+01	0.380	0.70399	
## TRAVTIME_AMT_BIN	5.334e+01	6.636e+02	0.080	0.93594	
## URBANICITY_Rural	-1.737e+02	8.191e+02	-0.212	0.83210	
## YOJ	-1.253e+02	8.558e+01	-1.464	0.14351	
## YOJ_AMT_BIN	1.847e+03	1.012e+03	1.824	0.06828	.
## YOJ_AMT_BIN_0_AND_9To14	6.231e+02	5.531e+02	1.127	0.26003	
## EDUCATION_Masters	-2.763e+02	7.656e+02	-0.361	0.71822	
## EDUCATION_High.School	-7.452e+02	5.639e+02	-1.322	0.18646	
## JOB_Clerical	-1.080e+03	6.458e+02	-1.672	0.09478	.
## JOB_Doctor	-1.692e+03	1.709e+03	-0.990	0.32210	
## JOB_Home.Maker	-6.548e+02	1.021e+03	-0.641	0.52151	
## JOB_Lawyer	-2.863e+02	1.029e+03	-0.278	0.78096	
## JOB_Professional	3.888e+02	7.117e+02	0.546	0.58492	
## JOB_Manager	-1.396e+03	9.149e+02	-1.526	0.12711	
## JOB_Student	-7.835e+02	9.216e+02	-0.850	0.39537	
## CAR_TYPE_Panel.Truck	-2.730e+01	8.979e+02	-0.030	0.97575	
## CAR_TYPE_Pickup	5.834e+02	5.965e+02	0.978	0.32815	
## CAR_TYPE_Sports.Car	2.118e+03	8.203e+02	2.582	0.00992	**
## CAR_TYPE_SUV	1.814e+03	7.294e+02	2.487	0.01297	*
## ---					
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					


```
##
## Residual standard error: 7577 on 1650 degrees of freedom
## (450 observations deleted due to missingness)
## Multiple R-squared:  0.05333,    Adjusted R-squared:  0.02407
## F-statistic: 1.823 on 51 and 1650 DF,  p-value: 0.0004045
```

Interpretation of the Model

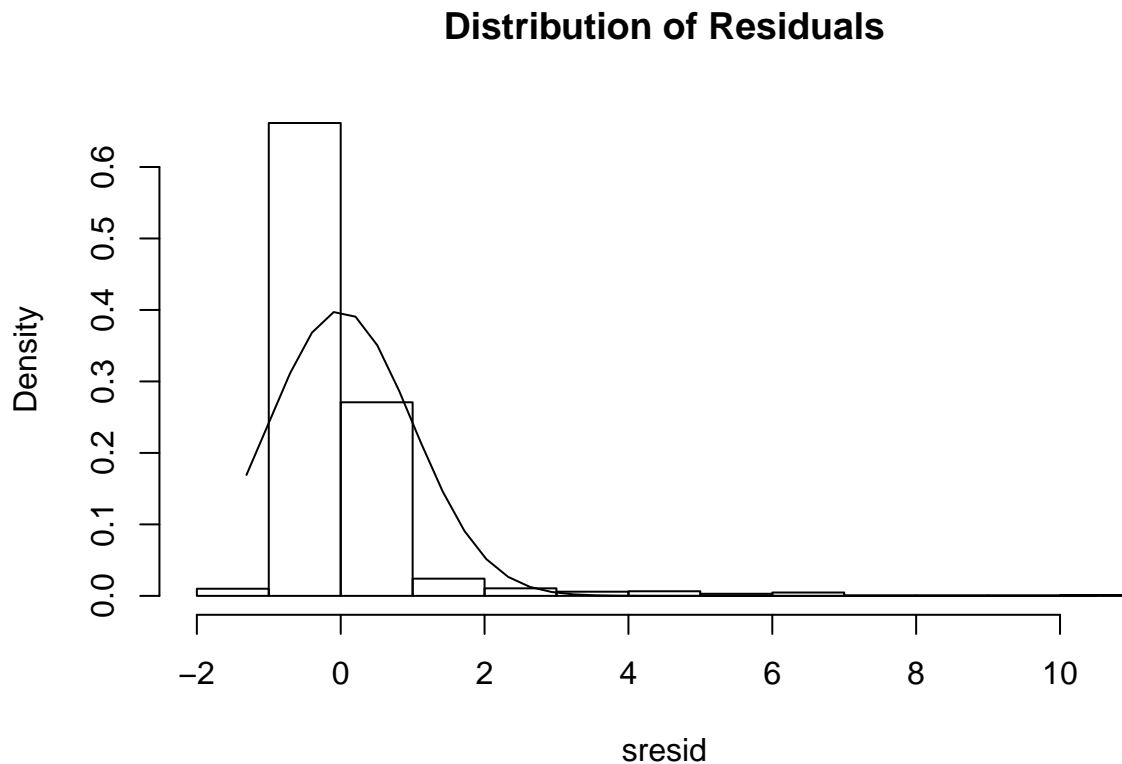
Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 7577
- Multiple R-squared: 0.05333
- Adjusted R-squared: 0.02407
- F-statistic: 1.823 on 51 and 1650 DF
- p-value: < 0.0004045

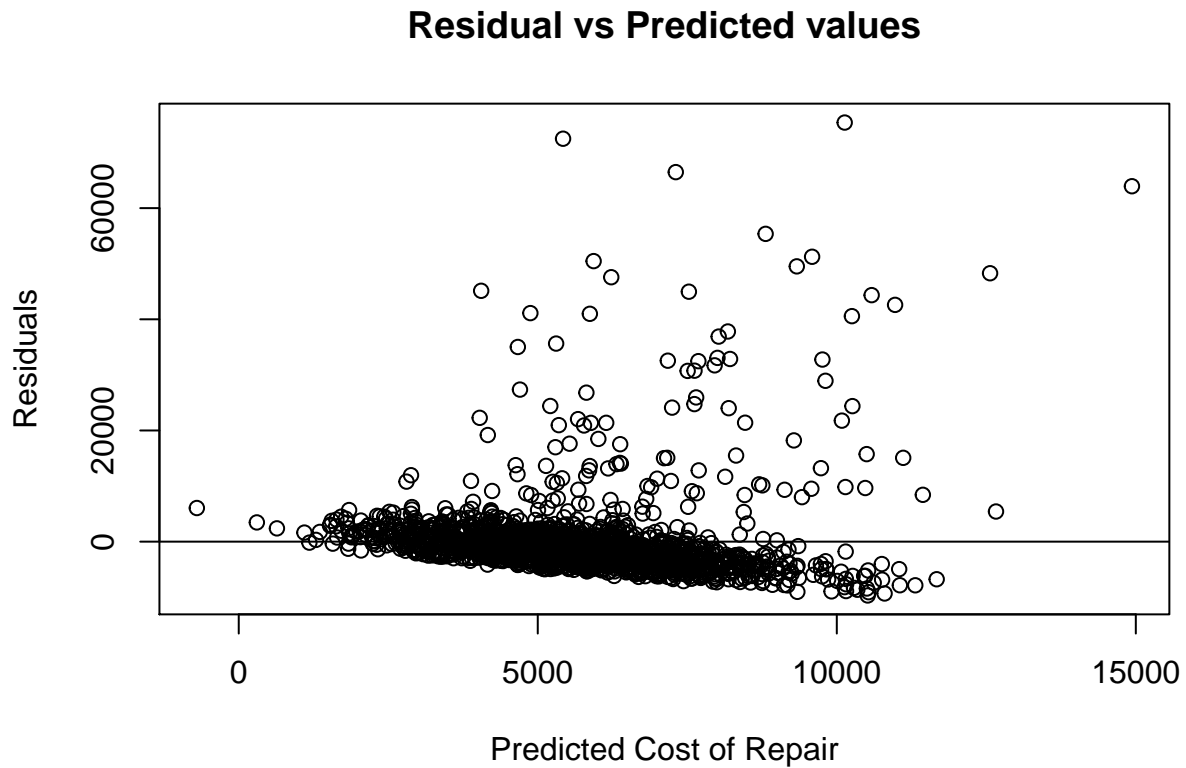
4.4 Final Linear Model Selection Summary

Based on the Models above, it is clear that Model 1 performs slightly better than Model 2. We will now use this model to validate further. We will create plots to validate the assumption of Linear Regression:

Normality check of residual values:



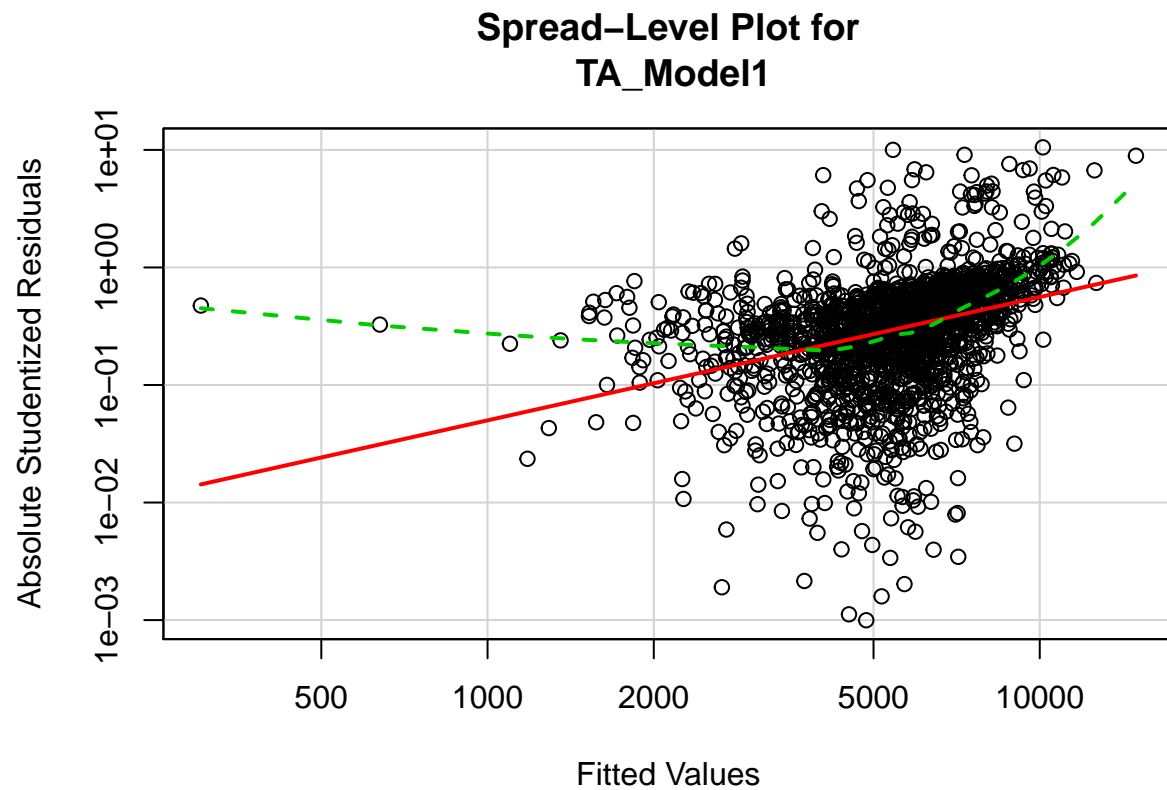
Based on the normality plot it appears that residual distribution is nearly normal. This indicates the mean of the difference between our predictions.



Distribution of residual values are random around base line and do not show any pattern around base line.

Evaluate homoscedasticity:

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1173.713    Df = 1    p = 3.149406e-257
```



```
##
## Suggested power transformation:  -0.0493236
```

The test confirms the non-constant error variance test. It also has a p-value higher than a significance level of 0.05.

Analysis of collinearity:

##	KIDSDRIV	AGE
##	5.929505	1.359778
##	HOMEKIDS	YOJ
##	2.659453	2.901429
##	INCOME	HOME_VAL
##	2.532339	2.860978
##	TRAVTIME	BLUEBOOK
##	1.354146	1.549806
##	TIF	OLDCLAIM
##	2.404457	2.435889
##	CLM_FREQ	MVR_PTS
##	2.240646	3.678998
##	CAR_AGE	CAR_USE_Commercial
##	2.285783	1.531233
##	MSTATUS_Yes	PARENT1_Yes
##	1.687018	1.725320
##	RED_CAR_yes	REVOKED_Yes

##	1.358885	1.350861
##	SEX_M	URBANICITY_Rural
##	1.954711	1.032607
##	EDUCATION_Bachelors	EDUCATION_High.School
##	3.056577	3.841265
##	EDUCATION_Masters	JOB_Blue.Collar
##	2.083557	3.278455
##	JOB_Clerical	JOB_Doctor
##	2.898793	1.282385
##	JOB_Home.Maker	JOB_Lawyer
##	2.196223	1.654125
##	JOB_Manager	JOB_Professional
##	1.563833	2.204776
##	JOB_Student	CAR_TYPE_Minivan
##	2.740883	1.731236
##	CAR_TYPE_Panel.Truck	CAR_TYPE_Pickup
##	1.446472	1.813465
##	CAR_TYPE_Sports.Car	CAR_TYPE_SUV
##	2.134942	2.646477
##	AGE_sin	TIF_sin
##	1.019210	1.398977
##	BLUEBOOK_sin	INCOME_AMT_BIN
##	1.015772	1.643573
##	YOJ_AMT_BIN	HOME_VAL_AMT_0_10K_BIN
##	2.047651	5.255209
##	HOME_VAL_AMT_60K_400K_BIN	OLDCLAIM_AMT_0_2K_BIN
##	4.044083	3.953793
##	OLDCLAIM_AMT_2K_10K_BIN	CLM_FREQ_AMT_BIN
##	3.007588	1.279461
##	MVR_PTS_AMT_BIN	CAR_AGE_AMT_BIN
##	2.144426	1.784942
##	TRAVTIME_AMT_BIN	KIDSDRIV_AMT_BIN_0
##	1.353599	3.930334
##	KIDSDRIV_AMT_BIN_1	HOMEKIDS_AMT_BIN_0
##	3.165772	2.705844
##	HOMEKIDS_AMT_BIN_3	YOJ_AMT_BIN_0_AND_9To14
##	1.299757	1.793634
##	INCOME_AMT_BIN_MISS_0	TIF_AMT_BIN_6
##	3.111211	2.116778
##	OLDCLAIM_AMT_BIN_MISS_0	MVR_PTS_AMT_BIN_0
##	2.753036	1.608030
##	MVR_PTS_AMT_BIN_5	
##	2.083625	

Variables have been tested with variance inflation factors (VIF). If any variable has value which is greater than 3 then the highest value variable been removed from model and model performance has been evaluated. Following are the out comes from this assessment steps-

Pass 1- Based on that variance inflation factors (VIF) following variable “KIDSDRIV” has highest value > 3 and is removed from model, and model is evaluated without that variable. Adjusted R² value has changed to 0.02487 due to removal of this variable. Hence this variable is not adding lot of value to the model and can be removed.

##

```
## Call:
## lm(formula = TARGET_AMT ~ . - KIDSDRIV, data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9677  -3249  -1393    792   75391
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.968e+03  5.166e+03   0.381  0.70325
## AGE           1.114e+01  2.564e+01   0.435  0.66392
## HOMEKIDS      3.099e+02  4.091e+02   0.758  0.44879
## YOJ          -1.288e+02  1.189e+02  -1.084  0.27868
## INCOME        -1.149e-02  1.083e-02  -1.061  0.28876
## HOME_VAL      1.941e-03  4.388e-03   0.442  0.65822
## TRAVTIME      5.252e+00  1.640e+01   0.320  0.74884
## BLUEBOOK      1.528e-01  3.398e-02   4.498 7.33e-06 ***
## TIF          -4.693e+01  1.115e+02  -0.421  0.67389
## OLDCLAIM      1.299e-02  4.406e-02   0.295  0.76811
## CLM_FREQ     -5.938e+02  3.285e+02  -1.808  0.07083 .
## MVR_PTS       6.051e+02  2.593e+02   2.333  0.01976 *
## CAR_AGE       -1.689e+02  7.617e+01  -2.218  0.02670 *
## CAR_USE_Commercial 6.238e+01  5.604e+02   0.111  0.91138
## MSTATUS_Yes   -1.815e+03  6.197e+02  -2.929  0.00345 **
## PARENT1_Yes   -7.485e+02  7.570e+02  -0.989  0.32296
## RED_CAR_yes   -1.314e+02  5.555e+02  -0.237  0.81302
## REVOKED_Yes   -1.130e+03  6.175e+02  -1.831  0.06734 .
## SEX_M         1.924e+03  7.216e+02   2.667  0.00773 **
## URBANICITY_Rural -2.499e+02  8.217e+02  -0.304  0.76107
## EDUCATION_Bachelors -3.466e+03  1.322e+03  -2.621  0.00884 **
## EDUCATION_High.School -4.125e+03  1.420e+03  -2.905  0.00372 **
## EDUCATION_Masters -2.236e+03  1.082e+03  -2.065  0.03905 *
## JOB_Blue.Collar  8.878e+02  1.309e+03   0.678  0.49790
## JOB_Clerical   -1.910e+02  1.375e+03  -0.139  0.88957
## JOB_Doctor     -3.153e+03  1.884e+03  -1.673  0.09450 .
## JOB_Home.Maker -1.202e+02  1.502e+03  -0.080  0.93624
## JOB_Lawyer     -3.472e+02  1.156e+03  -0.300  0.76392
## JOB_Manager    -9.610e+02  1.227e+03  -0.783  0.43348
## JOB_Professional 1.427e+03  1.287e+03   1.108  0.26786
## JOB_Student    2.459e+01  1.560e+03   0.016  0.98743
## CAR_TYPE_Minivan 1.142e+02  8.585e+02   0.133  0.89422
## CAR_TYPE_Panel.Truck -6.954e+01  9.573e+02  -0.073  0.94210
## CAR_TYPE_Pickup  5.863e+02  8.254e+02   0.710  0.47760
## CAR_TYPE_Sports.Car 2.203e+03  1.115e+03   1.976  0.04829 *
## CAR_TYPE_SUV    1.972e+03  1.047e+03   1.883  0.05990 .
## AGE_sin       -1.383e+01  2.633e+02  -0.053  0.95810
## TIF_sin       -2.986e+02  3.789e+02  -0.788  0.43072
## BLUEBOOK_sin   1.310e+02  2.608e+02   0.502  0.61551
## INCOME_AMT_BIN -5.872e+02  1.277e+03  -0.460  0.64574
## YOJ_AMT_BIN    1.765e+03  1.016e+03   1.737  0.08265 .
## HOME_VAL_AMT_0_10K_BIN 2.352e+03  1.958e+03   1.202  0.22971
## HOME_VAL_AMT_60K_400K_BIN 2.906e+03  1.495e+03   1.944  0.05202 .
## OLDCLAIM_AMT_0_2K_BIN -2.182e+03  1.453e+03  -1.502  0.13329
## OLDCLAIM_AMT_2K_10K_BIN -4.937e+02  1.138e+03  -0.434  0.66450
```

```
## CLM_FREQ_AMT_BIN          -1.331e+03  1.183e+03  -1.125  0.26068
## MVR_PTS_AMT_BIN           1.740e+03  7.962e+02   2.186  0.02899 *
## CAR_AGE_AMT_BIN           -6.830e+02  7.034e+02  -0.971  0.33172
## TRAVTIME_AMT_BIN          7.013e+01  6.637e+02   0.106  0.91587
## KIDSDRIV_AMT_BIN_0        -9.177e+02  7.034e+02  -1.305  0.19217
## KIDSDRIV_AMT_BIN_1         2.099e+03  9.813e+02   2.139  0.03257 *
## HOMEKIDS_AMT_BIN_0        -5.624e+02  9.949e+02  -0.565  0.57194
## HOMEKIDS_AMT_BIN_3         8.055e+02  1.426e+03   0.565  0.57237
## YOJ_AMT_BIN_0_AND_9To14    6.616e+02  7.110e+02   0.930  0.35228
## INCOME_AMT_BIN_MISS_0     -2.171e+02  1.778e+03  -0.122  0.90281
## TIF_AMT_BIN_6             -1.556e+02  8.831e+02  -0.176  0.86016
## OLDCLAIM_AMT_BIN_MISS_0    4.081e+02  1.025e+03   0.398  0.69058
## MVR_PTS_AMT_BIN_0         4.154e+02  6.286e+02   0.661  0.50879
## MVR_PTS_AMT_BIN_5         1.819e+03  1.082e+03   1.681  0.09302 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7574 on 1643 degrees of freedom
## Multiple R-squared:  0.05812,    Adjusted R-squared:  0.02487
## F-statistic: 1.748 on 58 and 1643 DF,  p-value: 0.0005054
```

Pass 2- Based on that variance inflation factors (VIF) following variable “HOME_VAL_AMT_0_10K_BIN” has highest value < 3 and is removed from model, and model is evaluated without that variable. Adjusted R² values changed to 0.02461. Hence this variable is not adding lot of value to the model and can be removed. We stopped at this point as further removal of variables led to a rapid deterioration of Adjusted R².

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - KIDSDRIV - HOME_VAL_AMT_0_10K_BIN,
##     data = na.omit(DS_TARGET_AMT))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9838  -3162  -1404    806   75299
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.867e+03  4.919e+03   0.786  0.43182
## AGE            1.254e+01  2.562e+01   0.489  0.62457
## HOMEKIDS       2.883e+02  4.087e+02   0.705  0.48064
## YOJ           -1.294e+02  1.189e+02  -1.088  0.27662
## INCOME        -8.605e-03  1.056e-02  -0.815  0.41525
## HOME_VAL      -1.617e-03  3.238e-03  -0.499  0.61753
## TRAVTIME       4.999e+00  1.640e+01   0.305  0.76061
## BLUEBOOK       1.532e-01  3.398e-02  4.510 6.96e-06 ***
## TIF           -5.043e+01  1.115e+02  -0.452  0.65107
## OLDCLAIM       1.527e-02  4.403e-02   0.347  0.72881
## CLM_FREQ      -5.833e+02  3.284e+02  -1.776  0.07589 .
## MVR_PTS        5.892e+02  2.590e+02   2.275  0.02306 *
## CAR_AGE       -1.687e+02  7.618e+01  -2.214  0.02698 *
## CAR_USE_Commercial  6.536e+01  5.604e+02   0.117  0.90718
## MSTATUS_Yes    -1.890e+03  6.167e+02  -3.064  0.00222 **
## PARENT1_Yes    -6.960e+02  7.559e+02  -0.921  0.35732
## RED_CAR_yes    -1.426e+02  5.555e+02  -0.257  0.79743
```

```

## REVOKED_Yes          -1.119e+03  6.175e+02  -1.812  0.07010 .
## SEX_M                1.930e+03  7.216e+02   2.675  0.00755 **
## URBANICITY_Rural     -2.570e+02  8.218e+02  -0.313  0.75456
## EDUCATION_Bachelors  -3.421e+03  1.322e+03  -2.588  0.00973 **
## EDUCATION_High.School -4.093e+03  1.420e+03  -2.883  0.00398 **
## EDUCATION_Masters    -2.156e+03  1.081e+03  -1.996  0.04613 *
## JOB_Blue.Collar      9.015e+02  1.310e+03   0.688  0.49133
## JOB_Clerical         -2.158e+02  1.375e+03  -0.157  0.87530
## JOB_Doctor           -3.241e+03  1.883e+03  -1.721  0.08539 .
## JOB_Home.Maker       -2.327e+02  1.499e+03  -0.155  0.87670
## JOB_Lawyer           -3.148e+02  1.156e+03  -0.272  0.78533
## JOB_Manager          -9.840e+02  1.227e+03  -0.802  0.42254
## JOB_Professional     1.463e+03  1.287e+03   1.137  0.25571
## JOB_Student          2.080e+02  1.553e+03   0.134  0.89348
## CAR_TYPE_Minivan     1.520e+02  8.580e+02   0.177  0.85939
## CAR_TYPE_Panel.Truck -2.704e+01  9.568e+02  -0.028  0.97745
## CAR_TYPE_Pickup       6.040e+02  8.254e+02   0.732  0.46439
## CAR_TYPE_Sports.Car  2.226e+03  1.115e+03   1.997  0.04599 *
## CAR_TYPE_SUV         1.999e+03  1.047e+03   1.909  0.05641 .
## AGE_sin              -7.305e+00  2.632e+02  -0.028  0.97787
## TIF_sin              -3.045e+02  3.789e+02  -0.804  0.42164
## BLUEBOOK_sin         1.315e+02  2.608e+02   0.504  0.61430
## INCOME_AMT_BIN       -2.323e+02  1.243e+03  -0.187  0.85172
## YOJ_AMT_BIN          1.773e+03  1.017e+03   1.744  0.08132 .
## HOME_VAL_AMT_60K_400K_BIN 1.341e+03  7.330e+02   1.830  0.06751 .
## OLDCLAIM_AMT_0_2K_BIN -2.134e+03  1.452e+03  -1.470  0.14185
## OLDCLAIM_AMT_2K_10K_BIN -4.602e+02  1.138e+03  -0.404  0.68598
## CLM_FREQ_AMT_BIN     -1.371e+03  1.182e+03  -1.159  0.24658
## MVR_PTS_AMT_BIN       1.703e+03  7.957e+02   2.140  0.03253 *
## CAR_AGE_AMT_BIN      -6.791e+02  7.035e+02  -0.965  0.33450
## TRAVTIME_AMT_BIN      5.845e+01  6.637e+02   0.088  0.92984
## KIDSDRIV_AMT_BIN_0    -9.284e+02  7.034e+02  -1.320  0.18707
## KIDSDRIV_AMT_BIN_1     2.062e+03  9.810e+02   2.102  0.03570 *
## HOMEKIDS_AMT_BIN_0    -5.557e+02  9.950e+02  -0.559  0.57656
## HOMEKIDS_AMT_BIN_3     7.544e+02  1.426e+03   0.529  0.59684
## YOJ_AMT_BIN_0_AND_9To14 6.725e+02  7.111e+02   0.946  0.34438
## INCOME_AMT_BIN_MISS_0 -2.521e+02  1.778e+03  -0.142  0.88723
## TIF_AMT_BIN_6        -1.820e+02  8.830e+02  -0.206  0.83674
## OLDCLAIM_AMT_BIN_MISS_0 4.252e+02  1.025e+03   0.415  0.67841
## MVR_PTS_AMT_BIN_0     3.963e+02  6.285e+02   0.631  0.52840
## MVR_PTS_AMT_BIN_5     1.761e+03  1.082e+03   1.628  0.10374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7575 on 1644 degrees of freedom
## Multiple R-squared:  0.05729,    Adjusted R-squared:  0.02461
## F-statistic: 1.753 on 57 and 1644 DF,  p-value: 0.000521

```

Final model was derived after number of iterations of variable eliminations were carried out. VIF values in the final model among variables < 3. In this scenario a model with slightly less performance was selected to avoid collinearity effect among variables and reduced complexity.

5 Prediction Using Evaluation Data

Now that we have selected the final models for both the TARGET_FLAG and the TARGET_AMT, we will go ahead and use these models to predict the results for the evaluation dataset. After transforming the data to meet the needs of the trained models, we will apply the models in 2 steps.

Step 1 - Here we use the transformed evaluation dataset to predict for the TARGET_FLAG using the requisite predictors.

Step 2 - Once we have the prediction for the TARGET_FLAG, we will filter this data for only those rows that were predicted for a CRASH. We then use this smaller dataset to predict for the TARGET_AMT.

5.1 Transformation of Evaluation Data

First we need to transform the evaluation dataset to account for all the predictors that were used in both the models.

5.2 Model Output for Logistic Regression

We now apply the final Logistic regression model that was trained for predicting the TARGET_FLAG. Below is a table of predictions.

Count for Crash / No Crash

Table 7: Predicted Crash Counts

Crash Predicted?	Counts
0	1410
1	304

5.3 Model Output for Linear Regression

Next we filter for the “predicted” crashes and we apply the final linear model to this smaller dataset to predict the TARGET_AMT. Below are the results for ONLY the “Crashed” records.

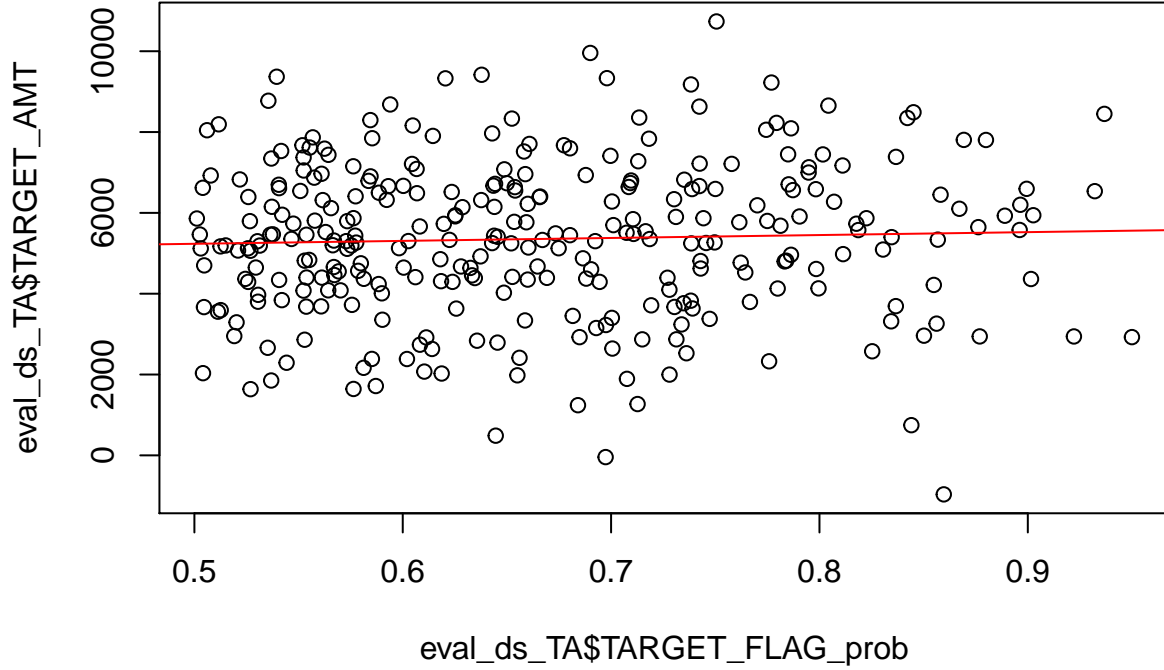
Top 10 Records by TARGET_AMT for TARGET_FLAG = 1

Table 8: Linear Regression Results

INDEX	TARGET_FLAG	TARGET_FLAG_prob	TARGET_AMT
2839	1	0.8596743	-964.90059
9764	1	0.6973731	-42.85442
598	1	0.6446238	486.94479
5361	1	0.8441118	746.90664
2889	1	0.6840995	1241.52457
8672	1	0.7127093	1269.64559
6447	1	0.5269622	1638.14701
748	1	0.5763727	1643.47425
6329	1	0.5870784	1715.89161

INDEX	TARGET_FLAG	TARGET_FLAG_prob	TARGET_AMT
4584	1	0.5368640	1852.60160

5.4 Conclusion



Outcome from regression model and outcome from linear model was plotted in the chart above. It can be seen from the chart above that probability associated with classification and predicted amount from linear model are linearly related though those two outcomes came from two completely different models. Relationship between those two are linear which indicates with the increase in probability for accident outcome, expense amount for repair also goes up with given data set. One key reason for that is there are number of variables in given data set which is common and playing important roles both the models such as- URBANICITY_Rural, MSTATUS_Yes, PARENT1_Yes, REVOKED_Yes, TRAVTIME, KIDSDRIV.