Logistic Regression Model for Insurance Claims

Setu Madhavi Namburu

2/9/2019

BINGO BONUS ATTEMPTED

1. Use of rpart (Recursive Partitioning And Regression Trees) to impute data - 20
2. Use of Decision trees (GBM, tree, Random forest) for variable selection - 20
3. Use of interaction plots and correlation plots - 10

1 INTRODUCTION

An automotive insurance company accumulated years worth of car crash claims data from their customers. They are interested in predicting probability of a customer's car crash so they can estimate the losses from the crashes which they can use to plan budget or adjust premiums to at risk customers. A data scientist is assigned to build predictive models using historical claims data. This data set contains approximately 8000 records. Each record represents a customer at an auto insurance company. Each record has two target variables. The first target 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 target variable is TARGET_AMT. 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. While the data does not have claims with time stamps at each customer level, it does have accumulated quantities like historical total claim payments, time in flight (how long the customer has been insured) as potential predictor variables. The data scientist decided to build a two stage model, 1st one for predicting the probability of a crash and 2nd model for predicting the claim amount. Multiplying the probability of crash with claim amount gives estimated potential loss from the crash. Six logistic regression models are built and are compared against each other using several metrics. Best model is selected based on performance metrics as well as simplicity of variables selected.

The data scientist followed below steps to build and validate the models (typical CRISP-DM process):

- Business problem understanding
  - Clarify the purpose and intension of the model, ask questions about available data and gather theoretical knowledge
- Data understanding
  - Gather and examine the data and available variables
  - Perform data quality checks
  - Conduct exploratary data analysis
  - Note down the insights observed
- Data Preparation
  - Address outliers and missing values
  - Create transformed variables based on EDA or based on knowledge gathered from subject matter experts
- Model Building
  - Experiment with various models
  - Calculate various performance metrics including predictive metrics on hold out data set
  - Perform model diagnostic checks
  - Iterate with EDA to add or modify variables until feasible model is derived
- Model Evaluation
  - Perform cross-validation
  - Run the model in proof of concept mode to make practical sense from experts
  - Iterate with above steps if practical significance is not met

- Model Deployment
  - Prepare data pipelining for new data
  - Use the predictive model built above to score the new data

2 DATA EXPLORATION

The dataset consist of claims data for 8161 unique customers (observations/rows) and 27 variables. The first variable is index variable which can be used as a unique identifier for each customer. There are two target variables one for if the car was in a crash and another if yes what was the claim amount.

2.1 Data Description

Tabl1 show the list of variables, their definitions and theoretical effect on the target variable. While complete details about the data collection are not available, given that variables like TIF, OLDCLAIM are the aggregate values for each customer, it is assumed that the flag and amount represent the recent event for a given customer & car.

Table 1: Variable names and descriptions

VAR_NAME	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 also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	#Children @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 into crashes
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 crashes
OLDCLAIM	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
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 safe

2.2 Descriptive Statistics

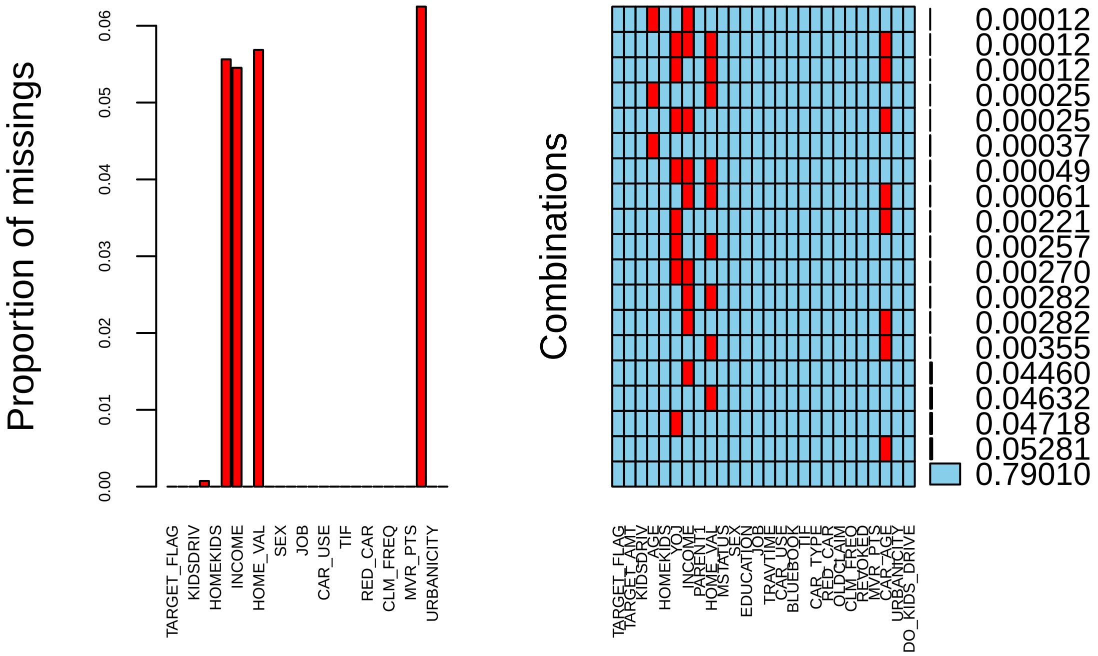
There are 11 categorical variables and 13 numerical variables and their descriptive statistics are shown in table2. The variables represent customer and claim attributes. Some data quality checks are performed while loading the data and corrected for their formats. Variable JOB has several missing values, so they are recoded as UNKNOWN as data scientist thinks imputing categorical type data is not a good idea.

2.3 Data Quality Checks

Figure 1 shows missing patterns in the data, as seen 79% of the data has complete observations so quality of the data set seems pretty good. AGE variable has only 6 missing values and other four variables have about 5% missing values with CAR_AGE being the highest. There is no significant pattern in the missingness of the variables.

Table 2: Descriptive statistics

type	variable	missing	complete	n	n_unique	top_counts	ordered	mean	sd	p0	p25	p50	p75	p100	hist
factor	CAR_TYPE	0	8161	8161	6	z_S: 2294, Min: 2145, Pic: 1389, Spo: 907	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	CAR_USE	0	8161	8161	2	Pri: 5132, Com: 3029, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	DO_KIDS_DRIVE	0	8161	8161	2	0: 7180, 1: 981, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	EDUCATION	0	8161	8161	5	z_H: 2330, Bac: 2242, Mas: 1658, <Hi: 1203	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	JOB	0	8161	8161	9	z_B: 1825, Cle: 1271, Pro: 1117, Man: 988	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	MSTATUS	0	8161	8161	2	Yes: 4894, z_N: 3267, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	PARENT1	0	8161	8161	2	No: 7084, Yes: 1077, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	RED_CAR	0	8161	8161	2	0: 5783, 1: 2378, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	REVOKED	0	8161	8161	2	No: 7161, Yes: 1000, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	SEX	0	8161	8161	2	z_F: 4375, M: 3786, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	TARGET_FLAG	0	8161	8161	2	0: 6008, 1: 2153, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
factor	URBANICITY	0	8161	8161	2	Urb: 6492, Rur: 1669, NA: 0	FALSE	NA	NA	NA	NA	NA	NA	NA	NA
integer	AGE	6	8155	8161	NA	NA	NA	44.79	8.63	16	39	45	51	81	
integer	CAR_AGE	510	7651	8161	NA	NA	NA	8.33	5.7	-3	1	8	12	28	
integer	CLM_FREQ	0	8161	8161	NA	NA	NA	0.8	1.16	0	0	0	2	5	
integer	HOMEKIDS	0	8161	8161	NA	NA	NA	0.72	1.12	0	0	0	1	5	
integer	KIDSDRIV	0	8161	8161	NA	NA	NA	0.17	0.51	0	0	0	0	4	
integer	MVR_PTS	0	8161	8161	NA	NA	NA	1.7	2.15	0	0	1	3	13	
integer	TIF	0	8161	8161	NA	NA	NA	5.35	4.15	1	1	4	7	25	
integer	TRAVTIME	0	8161	8161	NA	NA	NA	33.49	15.91	5	22	33	44	142	
integer	YOJ	454	7707	8161	NA	NA	NA	10.5	4.09	0	9	11	13	23	
numeric	BLUEBOOK	0	8161	8161	NA	NA	NA	15709.9	8419.73	1500	9280	14440	20850	69740	
numeric	HOME_VAL	464	7697	8161	NA	NA	NA	154867.29	129123.77	0	0	161160	238724	885282	
numeric	INCOME	445	7716	8161	NA	NA	NA	61898.09	47572.68	0	28097	54028	85986	367030	
numeric	OLDCLAIM	0	8161	8161	NA	NA	NA	4037.08	8777.14	0	0	0	4636	57037	
numeric	TARGET_AMT	0	8161	8161	NA	NA	NA	1504.32	4704.03	0	0	0	1036	107586.14	



values-1.bb

Figure 1: Pattern plot of missing observations

Table 3: Different quantile values of the numerical variables

vars	0.5%	1%	10%	50%	90%	99%	99.5%
TARGET_AMT	0	0	0.0	0	4904.0	19831.02	32235.34
KIDSDRIV	0	0	0.0	0	1.0	2.00	3.00
AGE	22	25	34.0	45	56.0	64.00	66.00
HOMEKIDS	0	0	0.0	0	3.0	4.00	4.00
YOJ	0	0	5.0	11	15.0	17.00	17.47
INCOME	0	0	4380.5	54028	123180.0	215519.80	237149.65
HOME_VAL	0	0	0.0	161160	316542.6	499336.52	545011.36
TRAVTIME	5	5	13.0	33	54.0	75.00	81.00
BLUEBOOK	1500	1500	6000.0	14440	27460.0	39000.00	42174.00
TIF	1	1	1.0	4	11.0	17.00	18.00
OLDCLAIM	0	0	0.0	0	9583.0	42801.40	46670.80
CLM_FREQ	0	0	0.0	0	3.0	4.00	4.00
MVR_PTS	0	0	0.0	1	5.0	8.00	9.00
CAR_AGE	1	1	1.0	8	16.0	21.00	22.00

The numerical data is also tested for spread and outliers. As shown in table 3, the lower and upper quantiles look reasonable except for TARGET_AMT where the right tail shows huge change from 5K to 32K. While <5K seem typical of insurance claims, higher dollar amounts are possible depending on the type of claim (due to uninsured, underinsured, expensive car etc). So the analyst decided to keep all the data for analysis.

2.4 Exploratory Data Analysis

EDA is performed using density plots and box plots for numerical variables as shown in figures 2&3. While densities look very similar for crash and non-crash events, box plots highlight some mean differences in those values for Flag=0,1 showing some potential for predictive ability to distiguish the crash.

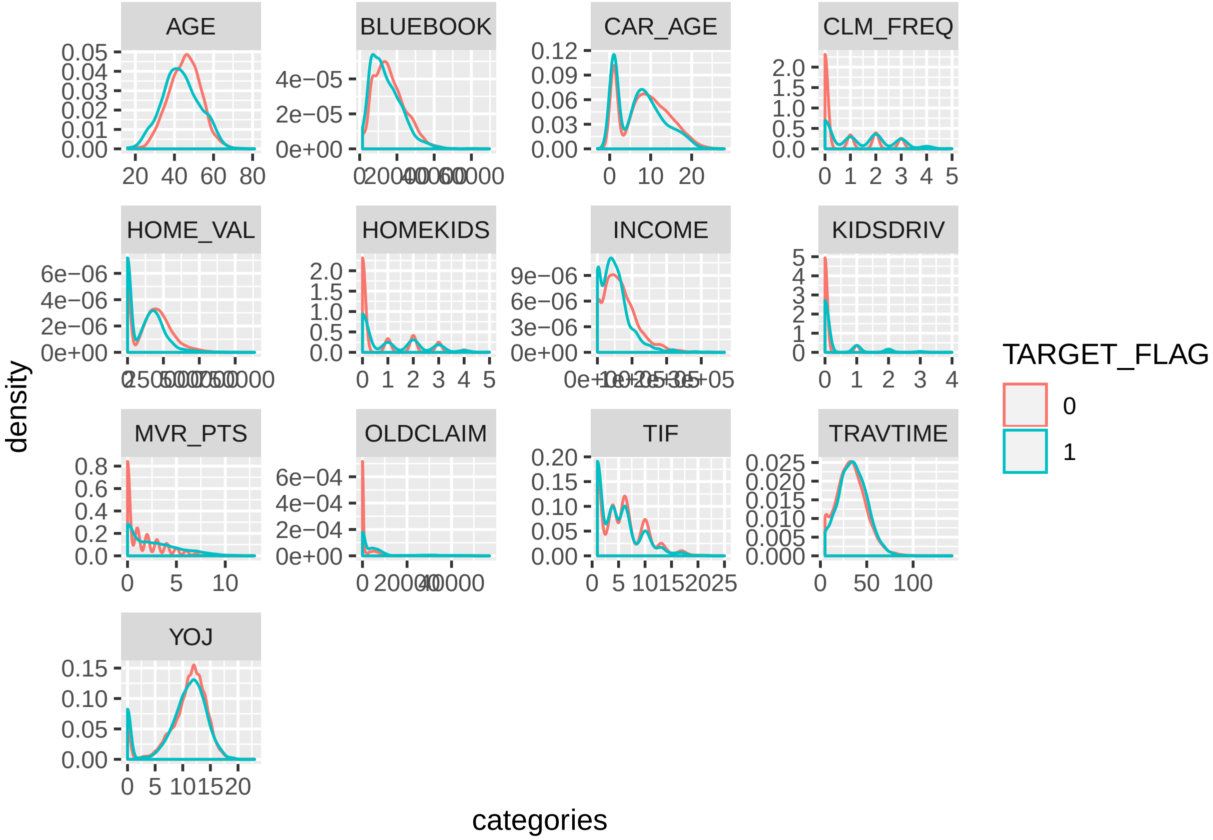


Figure 2: Density plots of numerical variables

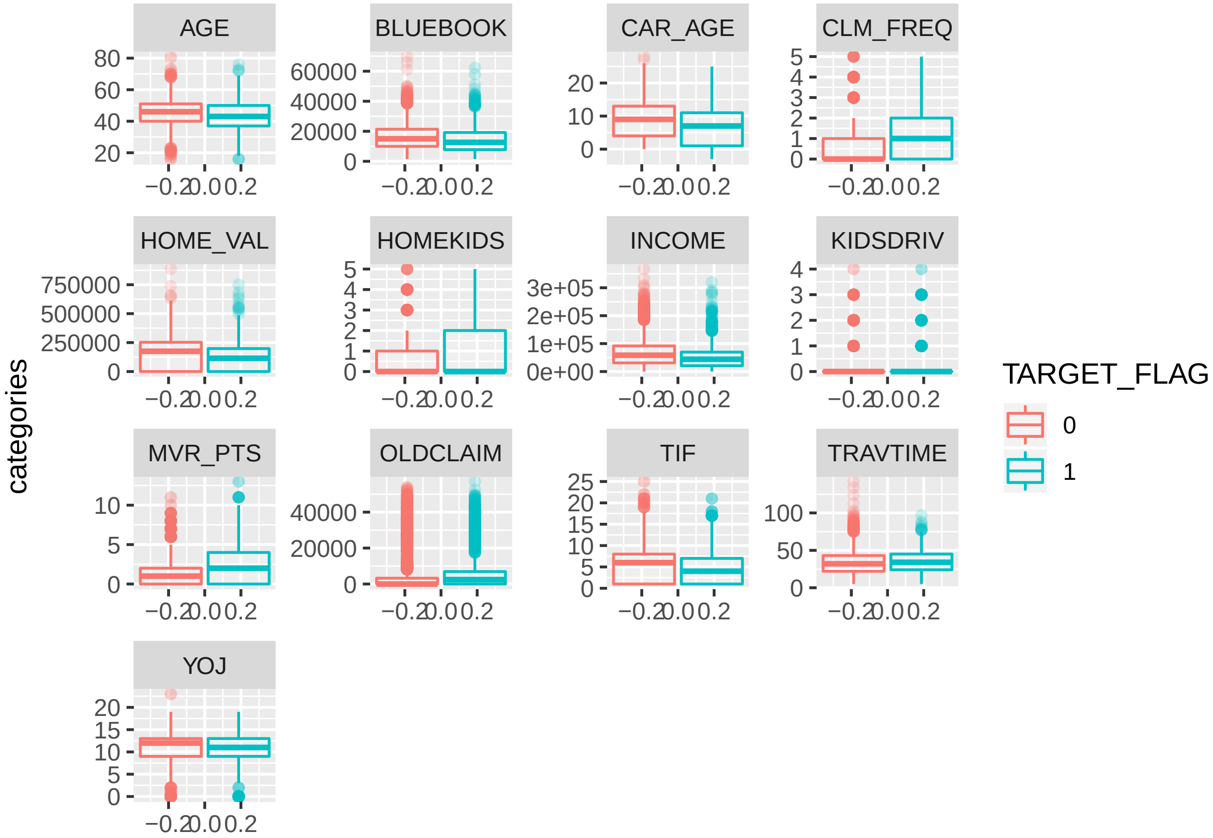


Figure 3: Boxplots of numerical variables

Correlations are another quick way to identify multi-collinearity between variables and if there is a correlation with target variable. Since our taget is binary variable, target amount is used for plotting correlations. As seen in figure 4, none of them seem to be strongly correlated with amount. The data has about 27% crash events (flag=1) and other non-crash events (flag=0) indicating imbalanced dataset. Moderate correlations between INCOME and HOME_VAL, BLUEBOOK (rich people tend to buy expensive homes, cars etc), OLDCLAIM and CLM_FREQ (more the claim frequency more historical claim amount) makes sense practically and may cause collinearity issues in moodelling.

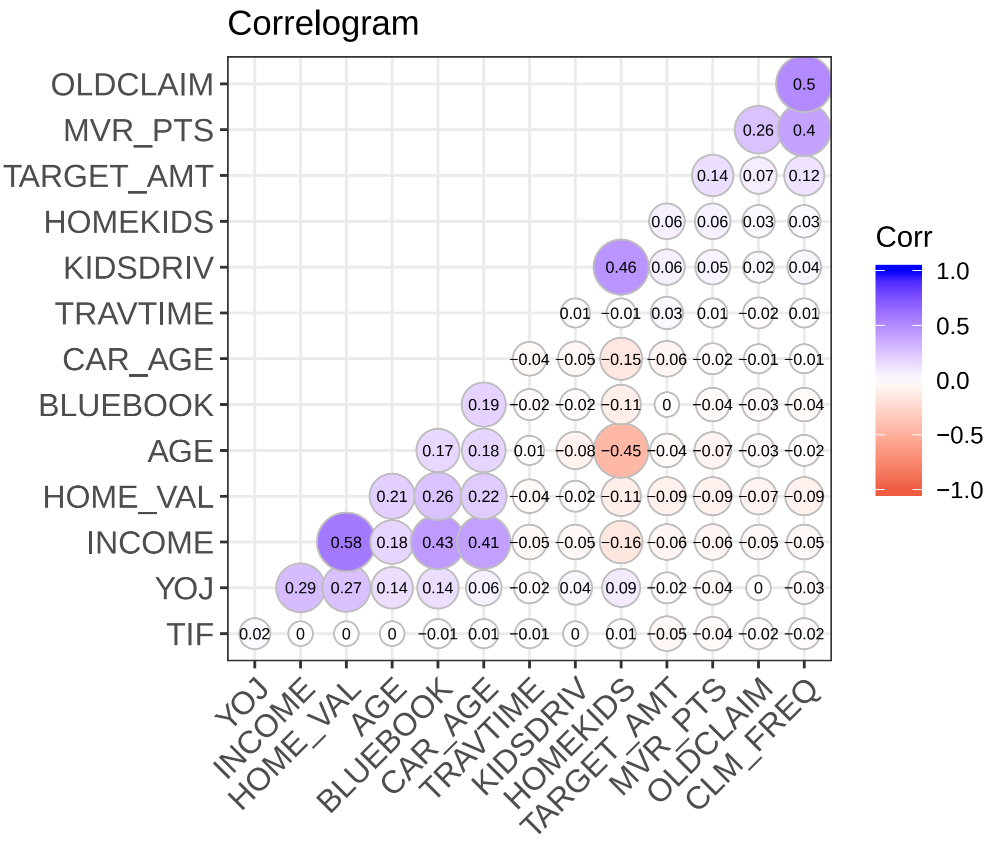


Figure 4: Correlation plot of numerical variables

For the categorical data bar plots are the best way to visualize quantities for quick view. As shown in figure 5, the data has good mix of 1,0s for all levels of the categories. Some levels are more dominant than others in some variables (e.g., There are more people where licence is not revoked vs where it is). In those situations it may make sense to build seperate models as less variation will make the variable insignificant.

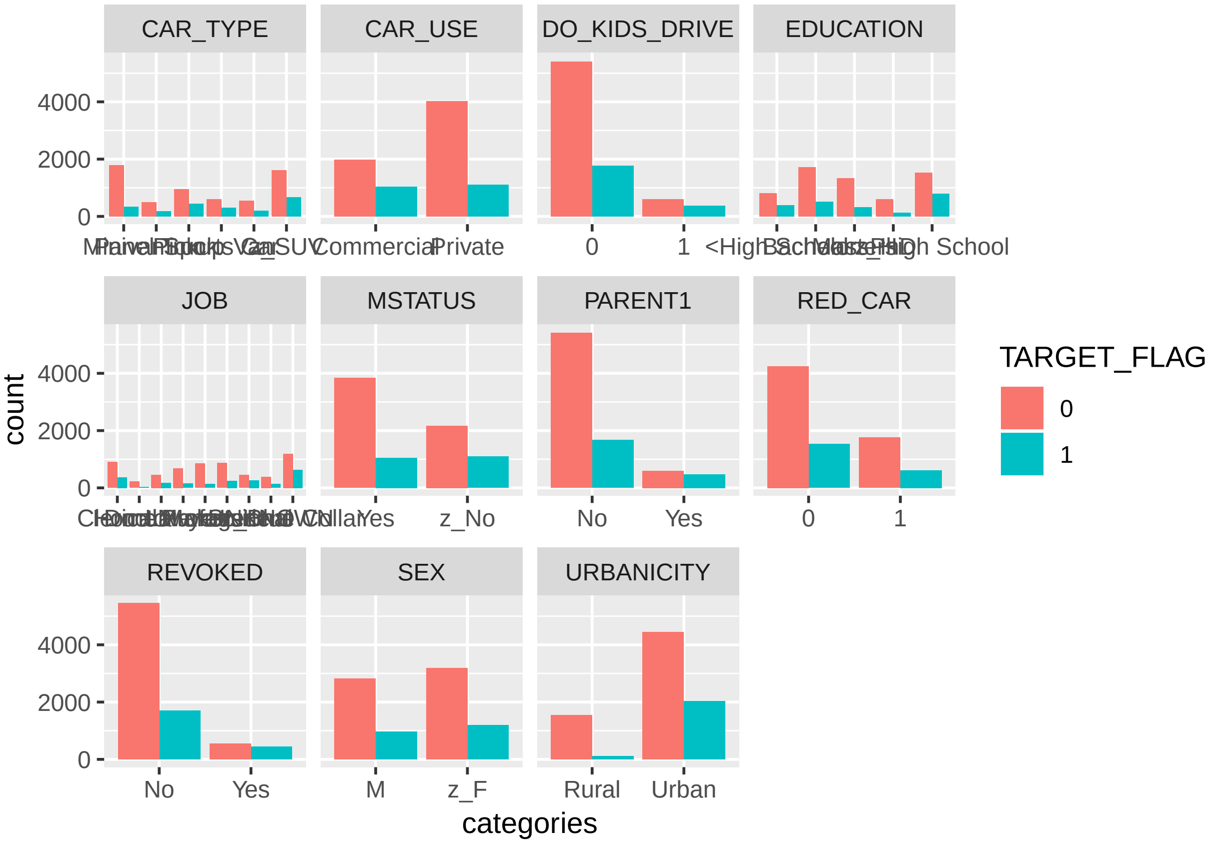


Figure 5: Bar plots of categorical variables

When we have many categorical variables and levels, the model may become complex to comprehend as model inferences are relative to a reference value. So studying categorical variables using interaction plots (** BONUS **), proportions in the data to create different sub models may be a better approach than building one master model. The categorical variables are converted into dummy variables for R. Table 4 shows variables with moderate correlations, the number for categorical variables is just a representation of data proportion (e.g., customers with masters degree tend to be more lawyers). Interaction plots in figure 6 show the impact of an interaction on claim amount. For example, for females the amount is drastically changing if they have a RED_CAR. These may be helpful to build different models based on practical feasibility and requirements or add interaction terms in the model. For this excercise we will not add any interaction terms to keep the models simple.

Table 4: Variables with more than moderate correlation

	Var1	Var2	Correlation
318	INCOME	HOME_VAL	0.5752443
729	TARGET_AMT	TARGET_FLAG_0	-0.5342461
781	TARGET_AMT	TARGET_FLAG_1	0.5342461
1262	CAR_AGE	EDUCATION_Masters	0.5074361
1482	EDUCATION_PhD	JOB_Doctor	0.5633243
1585	EDUCATION_Masters	JOB_Lawyer	0.5992751
2257	SEX_M	CAR_TYPE_z_SUV	-0.5346813
2258	SEX_z_F	CAR_TYPE_z_SUV	0.5346813
2309	SEX_M	RED_CAR_0	-0.6666207
2310	SEX_z_F	RED_CAR_0	0.6666207
2361	SEX_M	RED_CAR_1	0.6666207
2362	SEX_z_F	RED_CAR_1	-0.6666207
2602	KIDSDRIV	DO_KIDS_DRIVE_0	-0.9047358
2654	KIDSDRIV	DO_KIDS_DRIVE_1	0.9047358

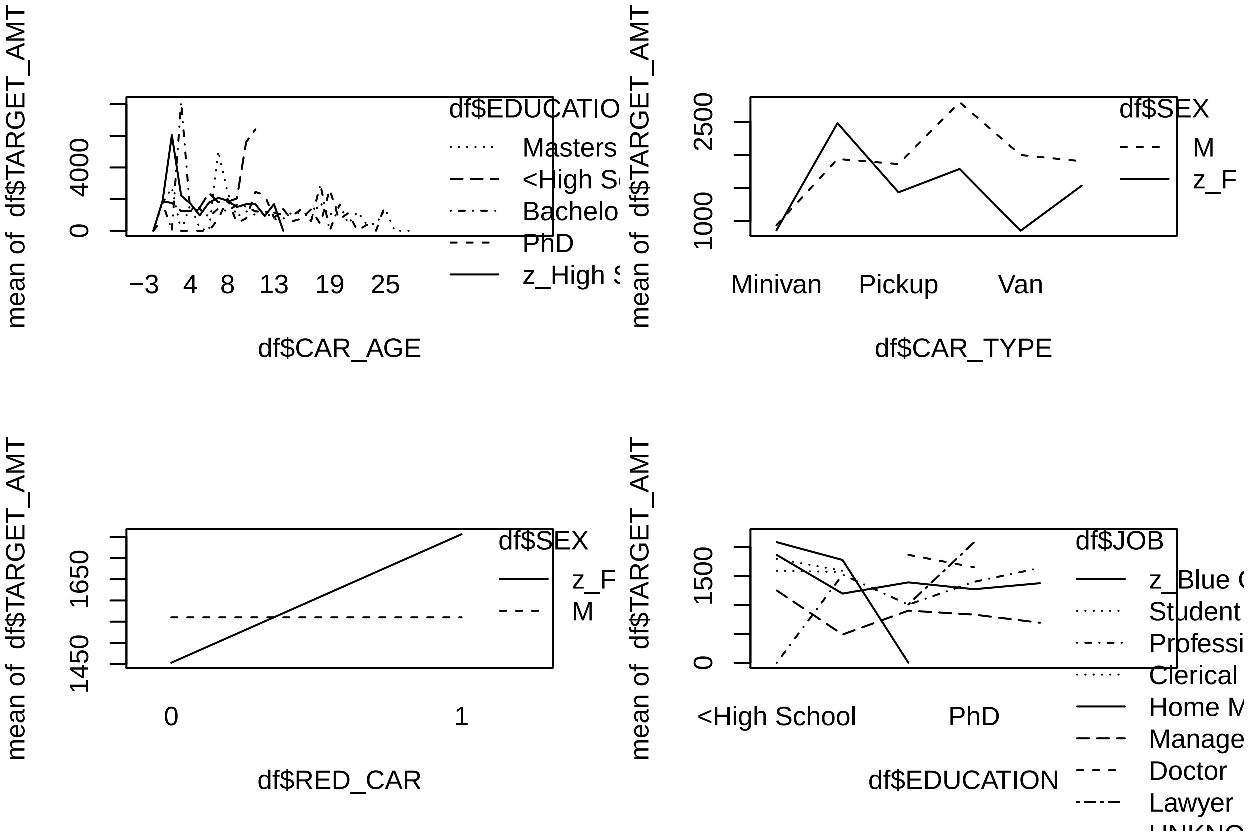


Figure 6: Interaction plots

3 DATA PREPARATION

The following sections describe missing value imputation schemes and variable transformations.

3.1 Missing Value Imputation

The five numerical variables are imputed for missing data. CAR_AGE is imputed with their mean values based on their car type. AGE, YOJ, INCOME,HOME_VAL are imputed using rpart trees (**BONUS**) using respective relevant variables using ANOVA method. Flag variables are created for the six variables with missing values to see if they have any significance for predictions.

3.2 Data Transformation

An indicator variable is added to represent if a customer owns home if HOME_VAL >0 (HOME_OWNER). As shown in figure 2, TRAVELTIME and BLUEBOOK have longer tails, so sqrt transformation is performed. INCOME variable is binned into four segments representing zero, loe, medium, high levels. Several other variables have bi-modality or multi-modality as seen in figure 2 but they are not transformed for keeping models simple. The resulting dataset has 33 predictor variables and 2 target variables which will be used in building 6 models in the following section.

4 MODEL BUILDING

While EDA provided some insights to pick few variables to build a model (along with intuition/SME), it is important to perform experiments with different variables so we can compare and contrast different models performance and not miss any unexpected insights from the data.

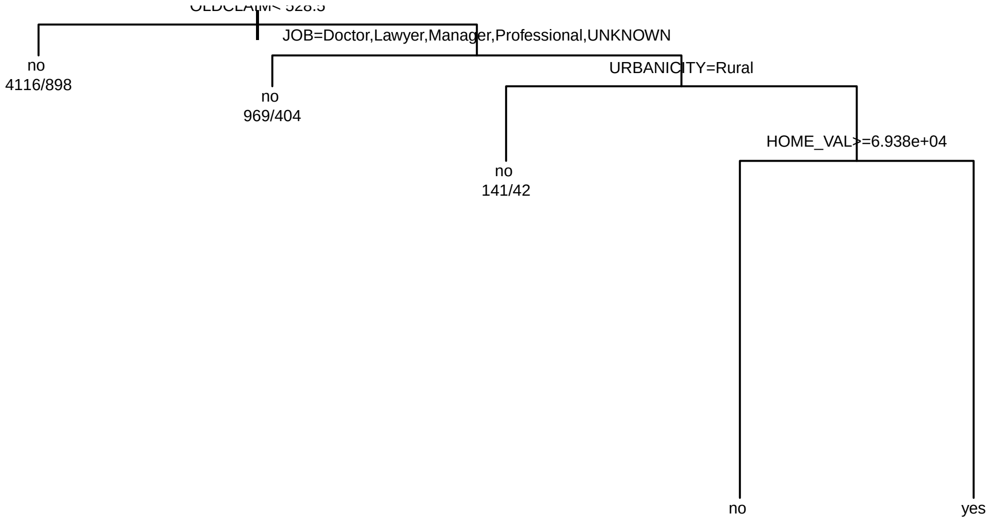
4.1 Variable Selection

Variables are selected using three different tree methods (GBM, simple decision tree, random forests) (** BONUS **). Table 5 shows top 20 variables selected by GBM based on relative influence. OLDCLAIM, MVR_PTS, REVOKED in the top 10 makes intuitive sense for being important variables to improve odds of the crash. URBANICITY is a little puzzling as theoretical effect is unknown but may be travelling from work/home if the location is in urban area, accidents are more prone to happen. The variables relative influence>1 are used to build a model in next section.

Table 5: Variable importance using GBM

var	rel.inf
JOB	14.586662
OLDCLAIM	14.259406
URBANICITY	12.290823
AGE	6.090356
MVR_PTS	6.026613
HOME_VAL	5.137779
REVOKED	4.980777
CAR_TYPE	4.810023
BLUEBOOK	4.570324
CAR_USE	3.908854
TRAVTIME	3.752616
PARENT1	3.158651
INCOME	2.941595
EDUCATION	2.566873
TIF	2.516207
KIDSDRIV	2.148219
MSTATUS	2.125128
DO_KIDS_DRIVE	1.961659
CAR_AGE	1.052900

Another simple decision tree is used to select most important variables. Customers with OLDCLAIM>528 and URBAN AREAs and HOME_VAL<70K seem to be more prone to crash. The three variables are used to build one model in next section.



tree-1.bb

Figure 7: Variable selection using decision tree

Random forest tree is used to select important variables with target as the crash (yes or no). This gave slightly different results based on Gini index than GBM but URBANICITY if removed may drastically effect the accuracy and deemed to be an important variable using all three methods. Variables with giniindex > 75 are carried forward to build a regression model in the next section.

selection using random forest tree-1.bb

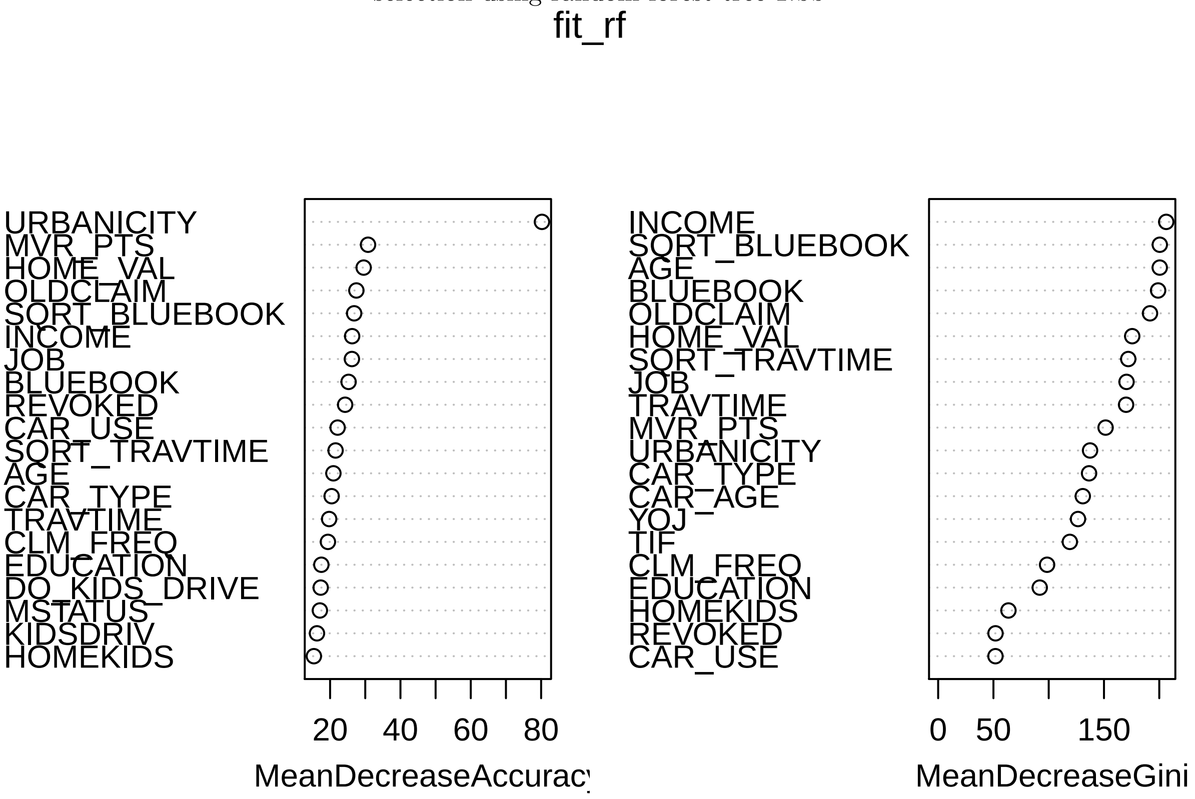


Figure 8: Variable importance using Random Forest

4.2 Various Logistic Regression Models

The dataset is divided into 80/20 train/validation splits to test the performance of the models on validation set/unseen data and six logistic regression models (where log odds of the target variable is regressed using bionomial link function) are built using variables selected using different methods. First model is built using all 33 variables to establish a baseline performance & step-wise regression model is the 2nd model.

JOB,BLUEBOOK,TIF,YOJ,EDUCATION,OLDCLAIM,HOME_VAL,INCOME_bin,CLM_FREQ,URBANICITY,KIDSDI are picked based on EDA insights and 3rd model is built using these variables. 4th model is based on variables selected using GBM, but OLDCLAIM,INCOME,HOME_VAL had coefficients that are nearly 0, so these variables are removed and the model is build again. The effects of these variables may be evident through other variables as removing these did not change BIC and ks.stat metrics much. Model 5 and 6 are built using variables selected using decision tree and random forest trees respectively. The model names, descriptions, various train test performance metrics along with number of terms in the model are shown in Table 6.

Table 6: Model fit metrics

logLik	AIC	BIC	deviance	ks.train	train.AUC	test.AUC	train.Accuracy	test.Accuracy	no_terms	model_name	model_description
-2902.576	5905.152	6244.506	5805.152	0.4702	0.8192027	0.8017780	0.7923347	0.7909429	50	full.model	All variables
-2905.951	5887.903	6145.811	5811.903	0.4702	0.8186366	0.8030849	0.7914185	0.7890819	38	step.model	variables selected by step-wise selection
-2996.141	6058.283	6282.256	5992.283	0.4608	0.8034131	0.8025638	0.7831730	0.7934243	33	eda.model	variables selected via EDA
-2978.725	6019.451	6229.850	5957.451	0.4608	0.8072387	0.8075329	0.7895862	0.7921836	31	gbm.model	variables selected based on gbm tree
-3296.017	6616.033	6697.478	6592.033	0.3780	0.7481887	0.7565448	0.7453046	0.7518610	12	tree.model	variables selected based on rpart tree
-3076.763	6217.527	6434.713	6153.527	0.4397	0.7895319	0.7887658	0.7738586	0.7834988	32	rf.model	variables selected based on rf tree

building-1.bb

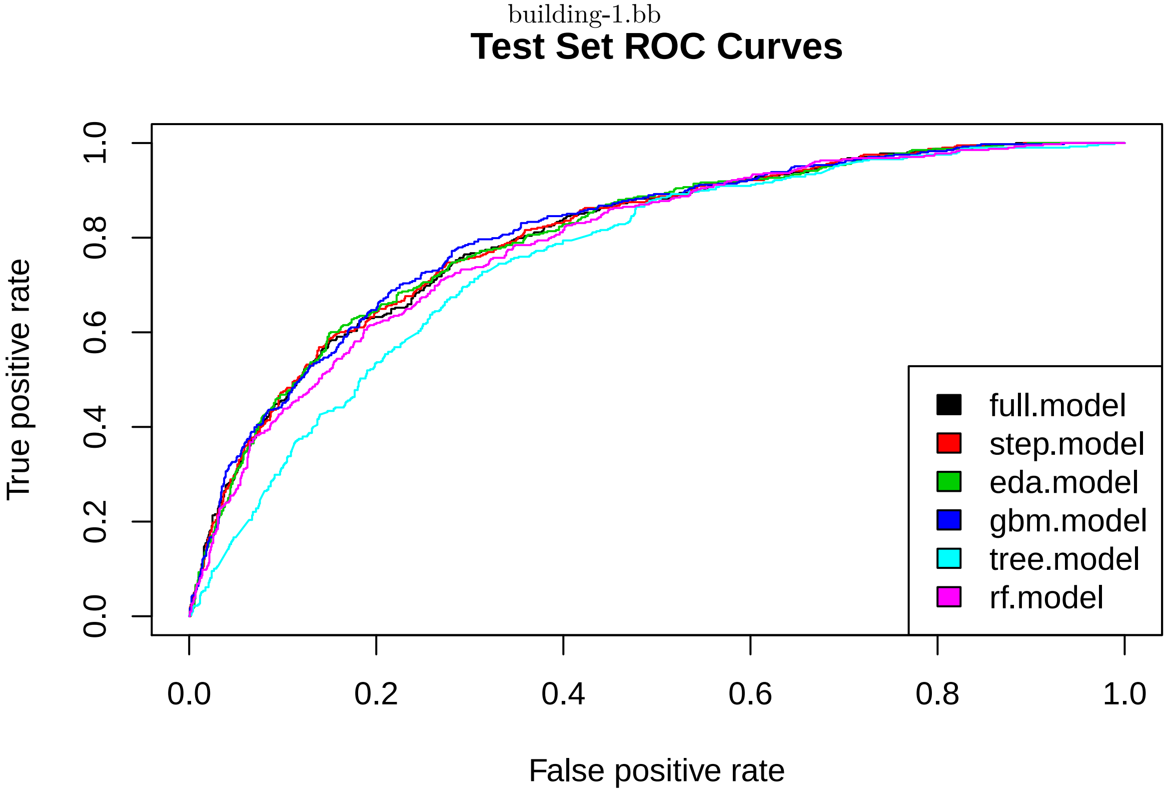


Figure 9: Validation set ROC curves for different models

Based on the observation of the metrics the models look fairly competitive. Figure 9 shows ROC curves of all six models based on test data, GBM seem to have highest area under the curve.

5 MODEL SELECTION

5.1 Model Ranking

Single metric alone cannot be used to select a champion model as that might give misleading results. Six models are ranked using all the metrics and the ranks are summed up for all metrics as well as selected metrics (BIS, KS test, test.AUC). Both aggregate ranks are taken into consideration along with number of terms (more weightage is given if simple terms are used in the model than transformations) to chose a champion model as gbm model (i.e., the model with variables selected via gbm insights) even though it has relatively low performance compared to full and step-wise models that are more complex.

Table 7: Model fit metrics - ranks

model_name	Rank.logLik	Rank.AIC	Rank.BIC	Rank.deviance	Rank.ks.train	Rank.train.AUC	Rank.test.AUC	Rank.train.Accuracy	Rank.test.Accuracy	no_terms	Rank_select_metrics	Rank_all
full.model	1	2	3	1	1	1	4	1	3	50	8	17
step.model	2	1	1	2	1	2	2	2	4	38	4	17
eda.model	4	4	4	4	2	4	3	4	1	33	9	30
gbm.model	3	3	2	3	2	3	1	3	2	31	5	22
tree.model	6	6	6	6	4	6	6	6	6	12	16	52
rf.model	5	5	5	5	3	5	5	5	5	32	13	43

5.2 Selected Model

After selecting gbm model as the winner, the model is retrained with all training data. The coefficients and their standard errors and significance values are shown in below table.

```
##
## The best model selected
## ==================================================
##                             Dependent variable:
##                         ---------------------------
##                                 TARGET_FLAG
## --------------------------------------------------
## JOBDoctor                         -0.953
##                                   (0.281)
##                                 t = -3.387
##                                p = 0.001***
##
## JOBHome Maker                      0.037
##                                   (0.132)
##                                  t = 0.281
##                                  p = 0.779
##
## JOBLawyer                         -0.450
##                                   (0.183)
##                                 t = -2.457
##                                 p = 0.014**
##
## JOBManager                        -1.090
##                                   (0.141)
##                                 t = -7.709
##                                p = 0.000***
##
## JOBProfessional                   -0.377
##                                   (0.122)
##                                 t = -3.097
##                                p = 0.002***
```

```
##
## JOBStudent                        0.071
##                                  (0.121)
##                                 t = 0.583
##                                 p = 0.560
##
## JOBUNKNOWN                       -0.593
##                                  (0.192)
##                                t = -3.089
##                               p = 0.003***
##
## JOBz_Blue Collar                 -0.198
##                                  (0.105)
##                                t = -1.893
##                                p = 0.059*
##
## URBANICITYUrban                   2.466
##                                  (0.111)
##                                t = 22.143
##                               p = 0.000***
##
## AGE                              -0.003
##                                  (0.004)
##                                t = -0.947
##                                 p = 0.344
##
## MVR_PTS                           0.143
##                                  (0.013)
##                                t = 11.293
##                               p = 0.000***
##
## REVOKEDYes                        0.741
##                                  (0.080)
##                                 t = 9.288
##                               p = 0.000***
##
## CAR_TYPEPanel Truck               0.613
##                                  (0.150)
##                                 t = 4.091
##                              p = 0.00005***
##
## CAR_TYPEPickup                    0.551
##                                  (0.100)
##                                 t = 5.507
##                              p = 0.00000***
##
## CAR_TYPESports Car                0.987
##                                  (0.107)
##                                 t = 9.217
##                               p = 0.000***
##
## CAR_TYPEVan                       0.623
##                                  (0.121)
##                                 t = 5.147
##                              p = 0.00000***
##
## CAR_TYPEz_SUV                     0.715
##                                  (0.086)
##                                 t = 8.366
##                               p = 0.000***
##
## BLUEBOOK                        -0.00003
##                                 (0.00000)
##                                t = -6.004
##                               p = 0.000***
##
## CAR_USEPrivate                   -0.777
##                                  (0.091)
##                                t = -8.520
##                               p = 0.000***
##
## TRAVTIME                          0.015
##                                  (0.002)
##                                 t = 7.898
##                               p = 0.000***
##
## PARENT1Yes                        0.397
##                                  (0.099)
##                                 t = 4.006
##                               p = 0.0001***
##
## EDUCATIONBachelors               -0.492
##                                  (0.113)
##                                t = -4.347
##                              p = 0.00002***
##
## EDUCATIONMasters                 -0.426
##                                  (0.176)
##                                t = -2.421
##                                p = 0.016**
##
## EDUCATIONPhD                     -0.453
##                                  (0.204)
##                                t = -2.216
##                                p = 0.027**
##
## EDUCATIONz_High School           -0.020
##                                  (0.094)
##                                t = -0.212
##                                 p = 0.833
##
## TIF                              -0.055
##                                  (0.007)
##                                t = -7.573
##                               p = 0.000***
##
## KIDSDRIV                          0.190
##                                  (0.121)
##                                t = 1.576
##                                p = 0.115
##
## MSTATUSz_No                      0.648
##                                (0.069)
##                                t = 9.427
##                              p = 0.000***
##
## DO_KIDS_DRIVE1                   0.422
##                                (0.194)
##                                t = 2.181
##                               p = 0.030**
##
## CAR_AGE                         -0.001
##                                (0.007)
##                               t = -0.151
##                                p = 0.880
##
## Constant                        -3.117
##                                (0.259)
##                               t = -12.016
##                              p = 0.000***
##
## ------------------------------------------------
## Observations                     8,161
## Log Likelihood                -3,693.013
## Akaike Inf. Crit.              7,448.026
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The selected model's equation is shown below. The reference value chosen here is arbitrary, but the signs of the coefficients that relate to log odds seem to be sensible theoretically. JOB unknown/missing seem to have large negative value which means people who did not provide job details seem to be safe drivers so gathering that data may be useful. While Bluebook has nearly 0 value, it is left in the model to keep as representative of richness of customers. As car age is increasing the risk seem to be reducing relatively which also makes sense as the drivers may be more experienced. Private CAR_USE to be less prone to crash compared to commercial use. Home makers, students seem to be more risky drivers compared to other job holders.

TARGET_FLAG = (-3.117) + (-0.953)*JOBDoctor + (0.037)*JOBHome Maker + (-0.45)*JOBLawyer + (-1.09)*JOBManager + (-0.377)*JOBProfessional + (0.071)*JOBStudent + (-0.593)*JOBUNKNOWN + (-0.198)*JOBz_Blue Collar + (2.466)*URBANICITYUrban + (-0.003)*AGE + (0.143)*MVR_PTS + (0.741)*REVOKEDYes + (0.613)*CAR_TYPEPanel Truck + (0.551)*CAR_TYPEPickup + (0.987)*CAR_TYPESports Car + (0.623)*CAR_TYPEVan + (0.715)*CAR_TYPEz_SUV + (0)*BLUEBOOK + (-0.777)*CAR_USEPrivate + (0.015)*TRAVTIME + (0.397)*PARENT1Yes + (-0.492)*EDUCATIONBachelors + (-0.426)*EDUCATIONMasters + (-0.453)*EDUCATIONPhD + (-0.02)*EDUCATIONz_High School + (-0.055)*TIF + (0.19)*KIDSDRIV + (0.648)*MSTATUSz_No + (0.422)*DO_KIDS_DRIVE1 + (-0.001)*CAR_AGE

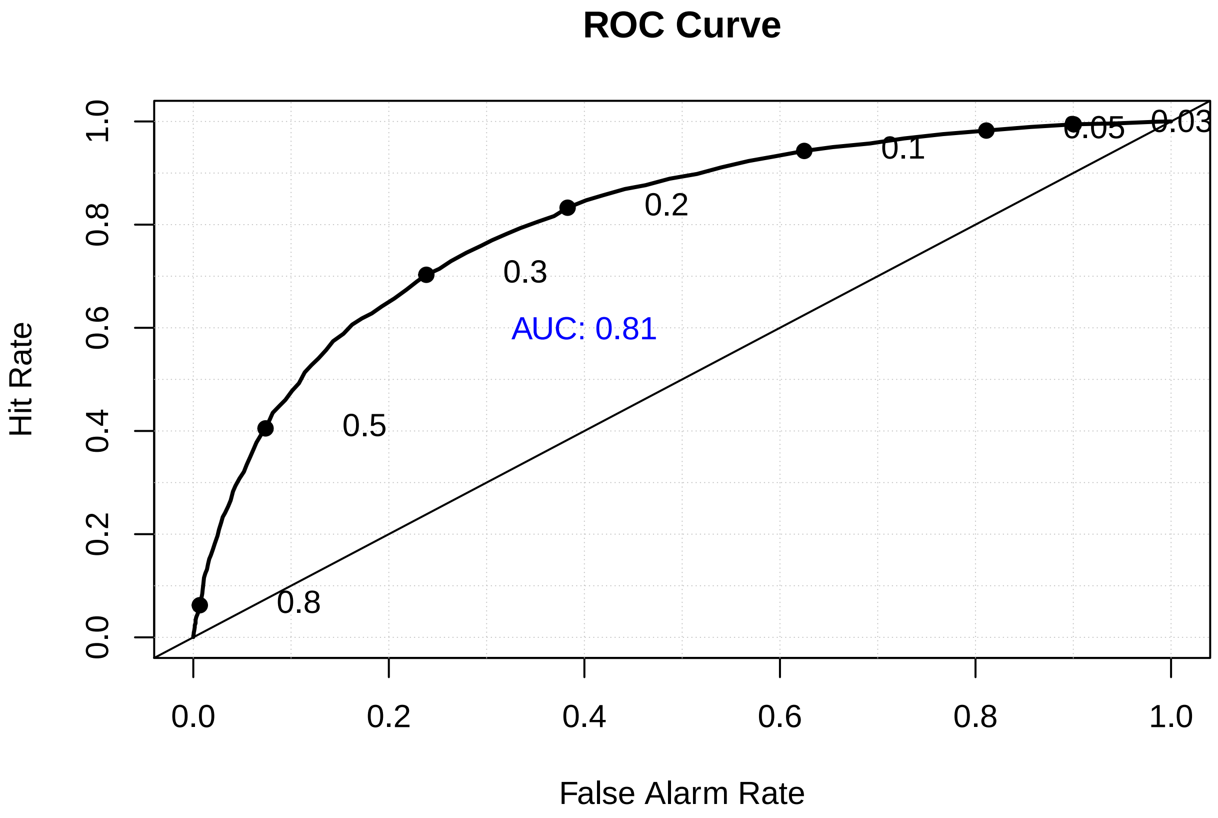


Figure 10: ROC curve of the selected model

Figure 10 shows the ROC curve based on full train data with an AUC of 81% and different probability threshold values to trade-off between true positive and false positives.

5.3 Multiple Linear Regression (MLR) Model

To estimate the losses, data is subsetted for vehicles that are in crash (TARGET_FLAG=1) and multiple linear regression model is build to predict target amount using BLUEBOOK and CAR_TYPE as the predictor variables. The model equation is given by:

TARGET_AMT = (4240.473) + (0.097)*BLUEBOOK + (371.854)*CAR_TYPEPanel Truck + (54.599)*CAR_TYPEPickup + (56.352)*CAR_TYPESports Car + (680.636)*CAR_TYPEVan + (-108.731)*CAR_TYPEz_SUV

6 STAND ALONE SCORING PROGRAM / MODEL DEPLOYMENT CODE

The test data is pre-processed using same steps that are perfomed on training data (mean values are stored from training data and used to impute test data as well). Below code shows the data processing and scoring steps of test data for producing crash probabilities and predicted claim amount. Multiplying the probability by claim amount gives estimated loss per each customer.

```
data1.original<-test
data1<-test
```

```r
library(rpart)
#age_impute <- rpart(AGE ~ JOB+INCOME, data1=data1[!is.na(data1$AGE), ], method="anova",na.action=na.om
ptratio_pred <- predict(age_impute, data1[is.na(data1$AGE), ])
data1$AGE[as.numeric(names(ptratio_pred))] <- ptratio_pred

#yoj_impute <- rpart(YOJ ~ JOB+AGE, data1=data1[!is.na(data1$YOJ), ], method="anova",na.action=na.omit)
ptratio_pred <- predict(yoj_impute, data1[is.na(data1$YOJ), ])
data1$YOJ[as.numeric(names(ptratio_pred))] <- ptratio_pred

#income_impute <- rpart(INCOME ~ JOB+YOJ, data1=data1[!is.na(data1$INCOME), ], method="anova",na.action
ptratio_pred <- predict(income_impute, data1[is.na(data1$INCOME), ])
data1$INCOME[as.numeric(names(ptratio_pred))] <- ptratio_pred

#homeval_impute <- rpart(HOME_VAL ~ JOB+INCOME, data1=data1[!is.na(data1$HOME_VAL), ], method="anova",n
ptratio_pred <- predict(homeval_impute, data1[is.na(data1$HOME_VAL), ])
data1$HOME_VAL[as.numeric(names(ptratio_pred))] <- ptratio_pred


data1$CAR_AGE <- na.aggregate(data1$CAR_AGE, data1$CAR_TYPE, mean, na.rm = TRUE)
data1$CAR_AGE[data1$CAR_AGE < 0 ] <- 0

flag_table <- (data1.original==data1)
flag_table[which(is.na(flag_table))] <- FALSE
flag.table<-as.data.frame(ifelse(flag_table == FALSE, 1,0))

##create flag variables
data1$AGE_FLAG <- flag.table$AGE
data1$INCOME_FLAG <- flag.table$INCOME
data1$YOJ_FLAG <- flag.table$YOJ
data1$HOME_VAL_FLAG <- flag.table$HOME_VAL
data1$CAR_AGE_FLAG <- flag.table$CAR_AGE

##create new variables
data1$HOME_OWNER <- ifelse(data1$HOME_VAL == 0, 0, 1)
data1$SQRT_TRAVTIME <- sqrt(data1$TRAVTIME)
data1$SQRT_BLUEBOOK <- sqrt(data1$BLUEBOOK)

# Bin Income
#data1$INCOME_bin[is.na(data1$INCOME)] <- "UNKNOWN"
data1$INCOME_bin[data1$INCOME == 0] <- "Zero"
data1$INCOME_bin[data1$INCOME >= 1 & data1$INCOME < 30000] <- "Low"
data1$INCOME_bin[data1$INCOME >= 30000 & data1$INCOME < 80000] <- "Medium"
data1$INCOME_bin[data1$INCOME >= 80000] <- "High"
data1$INCOME_bin <- factor(data1$INCOME_bin)
#data$INCOME_bin <- factor(data$INCOME_bin, levels=c("UNKNOWN","Zero","Low","Medium","High"))
```

7 SCORED DATA FILE

```r
newtest <- data1  %>% select(as.character(tree.vars$var))
data1<-within(data1,rm(TARGET_FLAG,TARGET_AMT))
```

Table 8: Summary of predicted targets

	P_TARGET_FLAG	P_TARGET_AMT
	Min. :0.0000	Min. : 0
	1st Qu.:0.0900	1st Qu.: 499
	Median :0.2200	Median :1276
	Mean :0.2704	Mean :1526
	3rd Qu.:0.4000	3rd Qu.:2275
	Max. :0.9400	Max. :6014

```
data1$P_TARGET_FLAG <- round(predict(gbm.model,newtest,type = "response"),2)

newtest <- data1 %>% select(BLUEBOOK,CAR_TYPE)
data1$P_TARGET_AMT <- round(predict(mlr.model,newtest)*data1$P_TARGET_FLAG)
predict.file <- data1 %>% select(INDEX,P_TARGET_FLAG,P_TARGET_AMT)
write.csv(predict.file,"LOGIT_INSURANCE_TEST_SETUNAMBURU.csv",row.names = FALSE )

predict.file %>% select(c(2:3)) %>% summary() %>% kableExtra:: kable(caption = "Summary of predicted ta
```

Table 8 shows summary statistics of predicted variables. Predicted TARGET_FLAG has mean value of 0.27 and TARGET_AMT has mean value of 1526 tallying well with training data.

8 CONCLUSION

Six logistic regression models are built using various variable selection methods to predict probability of crash for automotive insurance customers. Best model is selected using various fit statistics and predictive performance measures. The champion model selected seem to make sense intuitively as well based on the variables selected and signs of their coefficients. A multiple linear regression is used to build predictive model for estimating severity (claim amount) of the claim from a crash if it occurs. The estimated losses are calculated by multiplying probability of crash and severity. The average crash probability and estimated losses for new test data are in-line with training data statistics suggesting that chosen model is valid and can be deployed.