HW 4

Team 2

April 24, 2019

Contents

OVERVIEW **1**

Dependencies . . . 2

Objective . . . 2

DATA PREPARATION **2**

DATA CLEANING . . . 3

DATA EXPLORATION **5**

Summary Statistics . . . 5

Distributions . . . 6

Scatter plot matrix . . . 9

Correlation . . . 10

PART 3. BUILD MODELS **12**

Multiple Linear Regression . . . 12

Binary Logistic Regression . . . 17

AIC, BIC, LOglik, pseudoR2 . . . 17

Choosing the best model and applying it on test data set . . . 17

Assessing the performance of the choosen model: ROC and AUC . . . 17

PART 4: SELECT MODELS **17**

MLR Evaluation . . . 17

BLR Evaluation . . . 17

Predictions . . . 18

OVERVIEW

In this homework assignment, we will explore, analyze and model a data set containing approximately 8000 records representing a customer 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 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

Dependencies

Replication of our work requires the following packages in Rstudio:

```
# analyze data
library(corrplot)
library(randomForest)
library(olsrr)
library(psych)

#organize data
library(dplyr)
library(tidyr)

#visualize data
library(reshape2)
library(ggplot2)
library(ggpubr)

# dummy variables
library(sjmisc)
```

Objective

Our objective is to build multiple linear regression and binary logistic regression models on the `training` data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car.

DATA PREPARATION

The following chart outlines the variable imported from the `training.csv` dataset.

VARIABLE 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 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 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

DATA CLEANING

During our initial review, we recognized that the dataset required cleaning before further analysis and exploration.

```
FALSE 'data.frame': 8161 obs. of  25 variables:
FALSE  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
FALSE  $ TARGET_AMT : num  0 0 0 0 0 ...
FALSE  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
FALSE  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
FALSE  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
FALSE  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
FALSE  $ INCOME     : Factor w/ 6613 levels "","$0 ","$1,007 ",..: 5033 6292 1250 1 509 746 1488 315 47
FALSE  $ PARENT1    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
FALSE  $ HOME_VAL   : Factor w/ 5107 levels "","$0 ","$100,093 ",..: 2 3259 348 3917 3034 2 1 4167 2 2
FALSE  $ MSTATUS    : Factor w/ 2 levels "Yes","z_No": 2 2 1 1 1 2 1 1 2 2 ...
FALSE  $ SEX        : Factor w/ 2 levels "M","z_F": 1 1 2 1 2 2 2 1 2 1 ...
FALSE  $ EDUCATION  : Factor w/ 5 levels "<High School",..: 4 5 5 1 4 2 1 2 2 2 ...
FALSE  $ JOB        : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...
FALSE  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
FALSE  $ CAR_USE    : Factor w/ 2 levels "Commercial","Private": 2 1 2 2 2 1 2 1 2 1 ...
FALSE  $ BLUEBOOK   : Factor w/ 2789 levels "$1,500 ","$1,520 ",..: 434 503 2212 553 802 746 2672 701 13
FALSE  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
FALSE  $ CAR_TYPE   : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...
FALSE  $ RED_CAR    : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
FALSE  $ OLDCLAIM   : Factor w/ 2857 levels "$0 ","$1,000 ",..: 1449 1 1311 1 432 1 1 510 1 1 ...
FALSE  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
FALSE  $ REVOKED    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...
FALSE  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
FALSE  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
FALSE  $ URBANICITY : Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 2 ...

FALSE Length  Class   Mode
FALSE      0   NULL   NULL
```

Regular Expression

We first convered the `INCOME`, `HOMEVAL`, `BLUEBOOK`, and `OLDCLAIM` varaibles from factors to numeric by removing currency characters from the dataset.

```
# Remove currency characters from dataset
training$HOME_VAL<-as.numeric(gsub(",","",(gsub("\\$","",as.character(training$HOME_VAL)))))
training$INCOME<-as.numeric(gsub(",","",(gsub("\\$","",as.character(training$INCOME)))))
training$BLUEBOOK<-as.numeric(gsub(",","",(gsub("\\$","",as.character(training$BLUEBOOK)))))
training$OLDCLAIM<-as.numeric(gsub(",","",(gsub("\\$","",as.character(training$OLDCLAIM)))))

# Remove "z_" characters from dataset
training$MSTATUS<-gsub("z_","",as.character(training$MSTATUS))
training$URBANICITY<-gsub("z_","",as.character(training$URBANICITY))
training$CAR_TYPE<-gsub("z_","",as.character(training$CAR_TYPE))
training$EDUCATION<-gsub("z_","",as.character(training$EDUCATION))
training$JOB<-gsub("z_","",as.character(training$JOB))
```

Imputation

For the following incomplete cases, we replaced the value of NULL data with the mean of the relevant data vector.

	key	value
4	AGE	6
6	YOJ	454
7	INCOME	445
9	HOME_VAL	464
24	CAR_AGE	510

```
for(i in c(4,6,7,9,24)){
  training[is.na(training[,i]), i] <- mean(training[,i], na.rm = TRUE)
}
```

We found that mean imputation of the HOME_VAL and INCOME values do not change the variance. However, this assumes that these people have both income and homes. If not, this would bias our models.

We also found one record in which car age was less than zero. We also choose to impute this value with its corresponding mean.

```
training$CAR_AGE[training$CAR_AGE<0] <- mean(training$CAR_AGE)
```

Binary Variables

We then converted PARENT1, MSTATUS, SEX, CAR_USE, RED_CAR, REVOKED, and URBANICITY into binary variables.

```
training <- training %>%
  mutate(PARENT1 = ifelse(PARENT1 == "No", 0, 1)) %>%
  mutate(PARENT1=as.factor(PARENT1)) %>%
  mutate(MSTATUS = ifelse(MSTATUS == "No", 0, 1)) %>%
  mutate(MSTATUS=as.factor(MSTATUS)) %>%
  mutate(SEX = ifelse(SEX == "M", 0, 1)) %>%
  mutate(SEX=as.factor(SEX)) %>%
  mutate(CAR_USE = ifelse(CAR_USE == "Commercial", 0, 1)) %>%
  mutate(CAR_USE=as.factor(CAR_USE)) %>%
  mutate(RED_CAR = ifelse(RED_CAR == "no", 0, 1)) %>%
  mutate(RED_CAR=as.factor(RED_CAR)) %>%
  mutate(REVOKED = ifelse(REVOKED == "No", 0, 1)) %>%
  mutate(REVOKED=as.factor(REVOKED)) %>%
  mutate(URBANICITY = ifelse(URBANICITY == "Highly Rural/ Rural", 0, 1)) %>%
  mutate(URBANICITY=as.factor(URBANICITY))
```

Dummy Variables

Lastly, we changed the factors in the CAR_TYPE, EDUCATION, and JOB variables to dummy variables. Note that we discovered JOB had four NULL values. We choose to drop these values from our analysis.

```
training <- training %>%
  to_dummy(CAR_TYPE, EDUCATION, JOB, suffix = "label") %>%
  bind_cols(training) %>%
  select(TARGET_FLAG, everything()) %>%
  select(-CAR_TYPE, -EDUCATION, -JOB, -JOB_) %>%
  mutate(URBANICITY=as.factor(URBANICITY))

training[,1:20] <- sapply(training[,1:20],as.factor)
```

DATA EXPLORATION

Summary Statistics

We look at summary of the data below. Note that the stars next to the variable names indicate which variables are factors in our new dataset.

	vars	n	mean	sd	min	max	range	se
TARGET_FLAG*	1	8161	0.26	0.44	0	1.0	1.0	0.00
CAR_TYPE_Minivan*	2	8161	0.26	0.44	0	1.0	1.0	0.00
CAR_TYPE_Panel Truck*	3	8161	0.08	0.28	0	1.0	1.0	0.00
CAR_TYPE_Pickup*	4	8161	0.17	0.38	0	1.0	1.0	0.00
CAR_TYPE_Sports Car*	5	8161	0.11	0.31	0	1.0	1.0	0.00
CAR_TYPE_SUV*	6	8161	0.28	0.45	0	1.0	1.0	0.00
CAR_TYPE_Van*	7	8161	0.09	0.29	0	1.0	1.0	0.00
EDUCATION_<High School*	8	8161	0.15	0.35	0	1.0	1.0	0.00
EDUCATION_Bachelors*	9	8161	0.27	0.45	0	1.0	1.0	0.00
EDUCATION_High School*	10	8161	0.29	0.45	0	1.0	1.0	0.00
EDUCATION_Masters*	11	8161	0.20	0.40	0	1.0	1.0	0.00
EDUCATION_PhD*	12	8161	0.09	0.29	0	1.0	1.0	0.00
JOB_Blue Collar*	13	8161	0.22	0.42	0	1.0	1.0	0.00
JOB_Clerical*	14	8161	0.16	0.36	0	1.0	1.0	0.00
JOB_Doctor*	15	8161	0.03	0.17	0	1.0	1.0	0.00
JOB_Home Maker*	16	8161	0.08	0.27	0	1.0	1.0	0.00
JOB_Lawyer*	17	8161	0.10	0.30	0	1.0	1.0	0.00
JOB_Manager*	18	8161	0.12	0.33	0	1.0	1.0	0.00
JOB_Professional*	19	8161	0.14	0.34	0	1.0	1.0	0.00
JOB_Student*	20	8161	0.09	0.28	0	1.0	1.0	0.00
TARGET_AMT	21	8161	1504.32	4704.03	0	107586.1	107586.1	52.07
KIDSDRIV	22	8161	0.17	0.51	0	4.0	4.0	0.01
AGE	23	8161	44.79	8.62	16	81.0	65.0	0.10
HOMEKIDS	24	8161	0.72	1.12	0	5.0	5.0	0.01
YOJ	25	8161	10.50	3.98	0	23.0	23.0	0.04
INCOME	26	8161	61898.09	46257.33	0	367030.0	367030.0	512.05
PARENT1*	27	8161	1.13	0.34	1	2.0	1.0	0.00
HOME_VAL	28	8161	154867.29	125398.88	0	885282.0	885282.0	1388.10
MSTATUS*	29	8161	1.60	0.49	1	2.0	1.0	0.01
SEX*	30	8161	1.54	0.50	1	2.0	1.0	0.01
TRAVTIME	31	8161	33.49	15.91	5	142.0	137.0	0.18
CAR_USE*	32	8161	1.63	0.48	1	2.0	1.0	0.01
BLUEBOOK	33	8161	15709.90	8419.73	1500	69740.0	68240.0	93.20
TIF	34	8161	5.35	4.15	1	25.0	24.0	0.05
RED_CAR*	35	8161	1.29	0.45	1	2.0	1.0	0.01
OLDCLAIM	36	8161	4037.08	8777.14	0	57037.0	57037.0	97.16
CLM_FREQ	37	8161	0.80	1.16	0	5.0	5.0	0.01
REVOKED*	38	8161	1.12	0.33	1	2.0	1.0	0.00
MVR_PTS	39	8161	1.70	2.15	0	13.0	13.0	0.02
CAR_AGE	40	8161	8.33	5.52	0	28.0	28.0	0.06
URBANICITY*	41	8161	1.80	0.40	1	2.0	1.0	0.00

Distributions

Below, we examine the distribution of variables using histograms, density plots, violin plots, and bar plots.

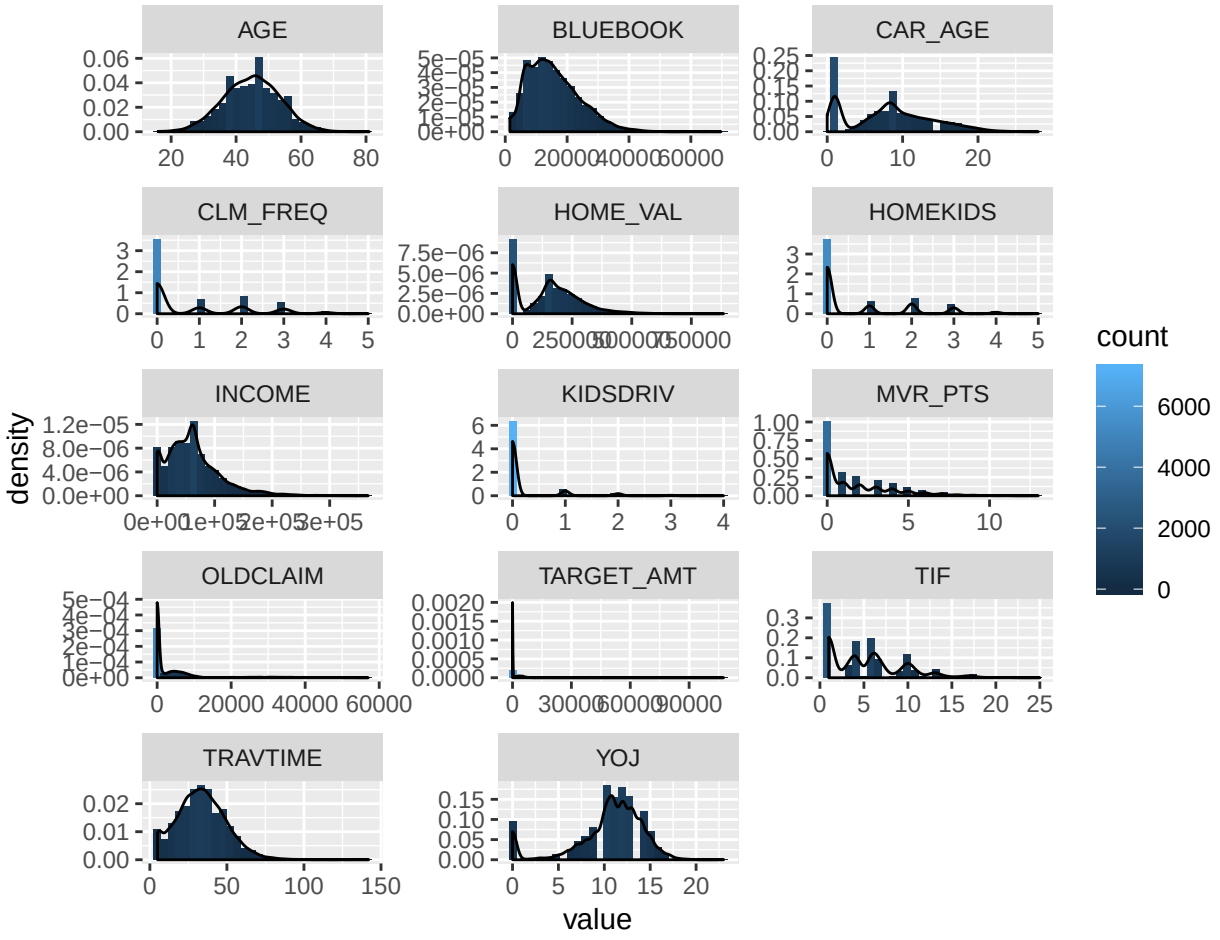
Numeric Variables Distributions

Our numeric variables are shown below using the following histograms and density plots.

From this output, we can tell that the variables, `AGE` and `YOJ`, appear to be fairly normally distributed.

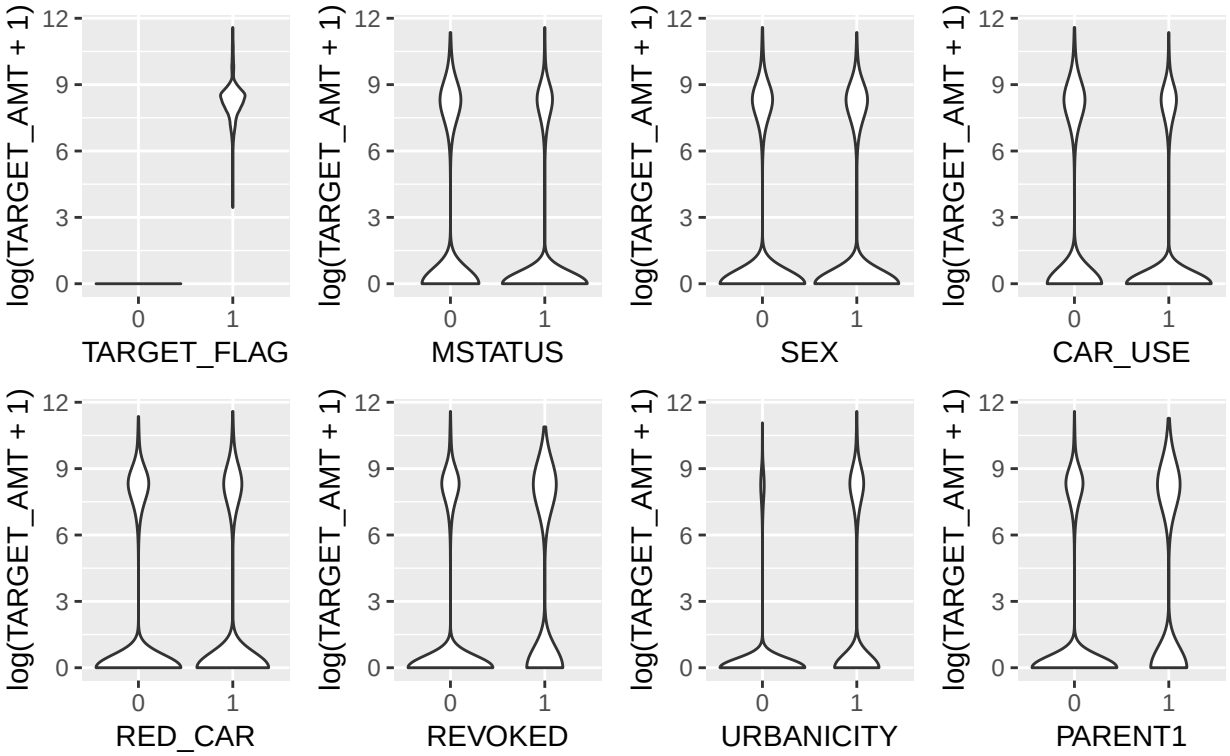
However, `CLM_FREQ`, `HOME_VAL`, `HOME_KIDS`, `KIDSDRIV`, and `MVR_PTS` appear to follow either a continuous or discrete quasi-Poisson process.

In any case, `CLM_FREQ`, `HOME_KIDS`, `KIDSDRIV`, `MVR_PTS`, `OLDCLAIM`, `TIF`, `TRAVTIME`, and `YOJ` are count values. Log transforms will be inappropriate and would skew this data.

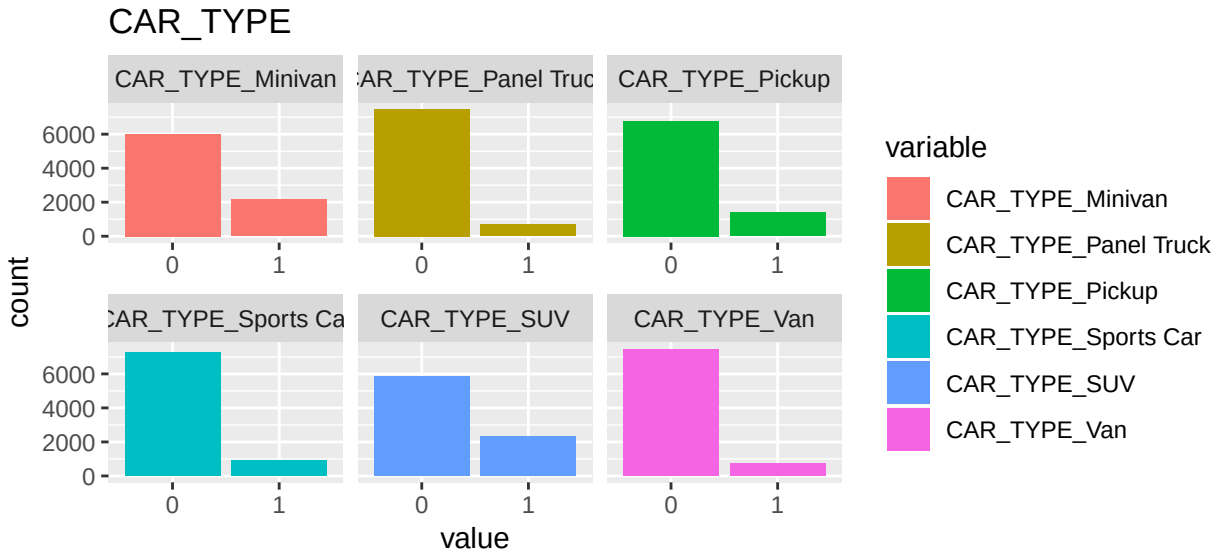


Binary Variable Distribution

The following violin plots show the distribution of our binary variables.

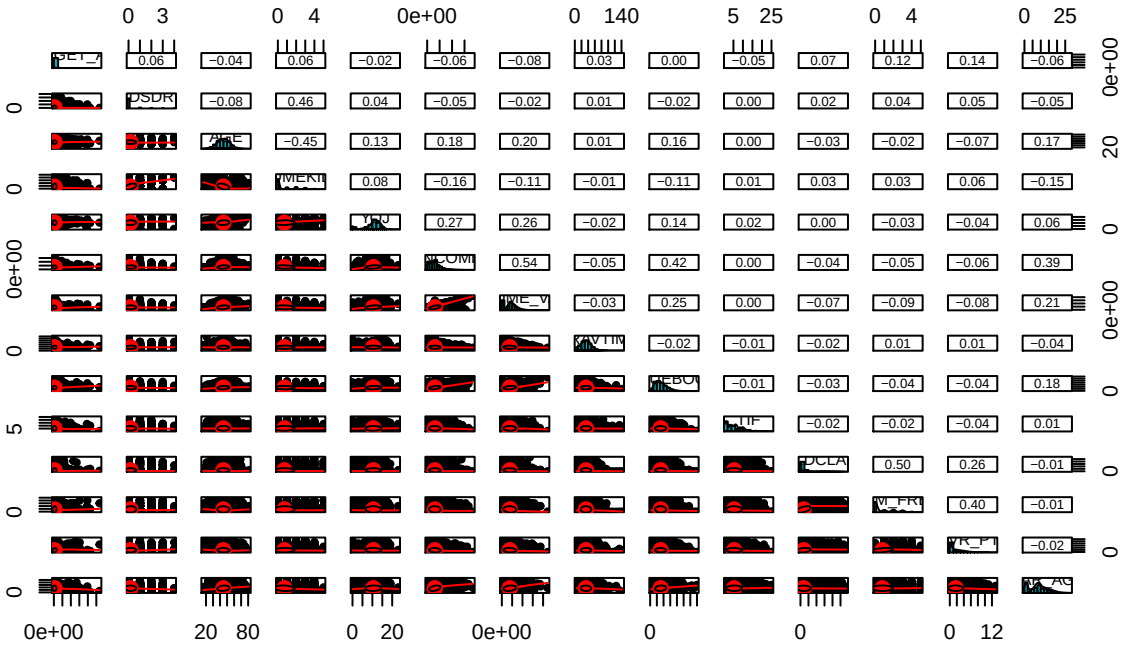


Dummy Variable Distributions



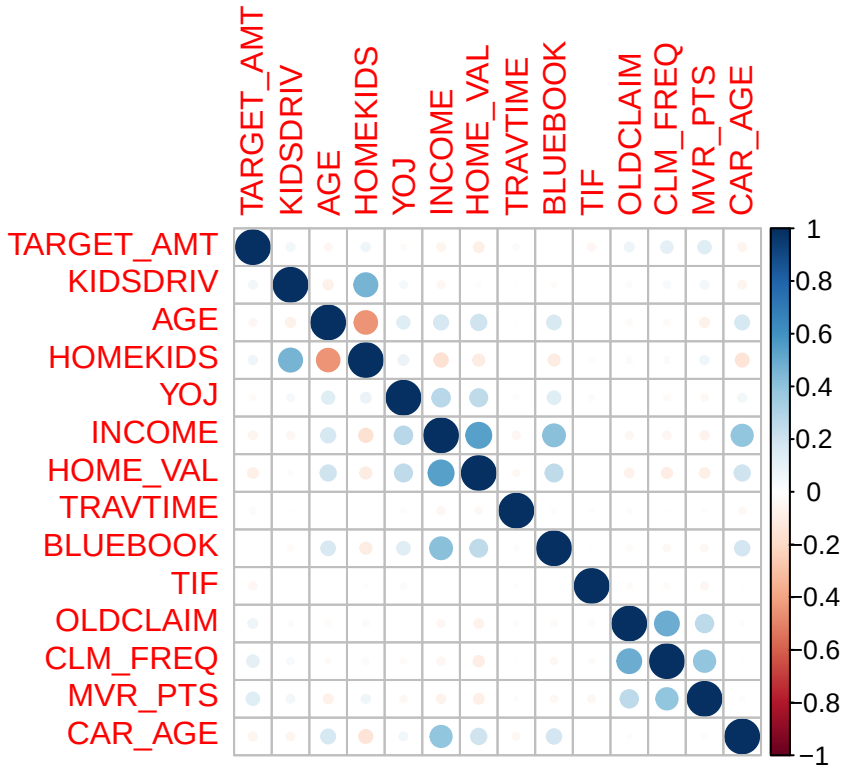
Scatter plot matrix

We then build scatter plot matrix for continious variables

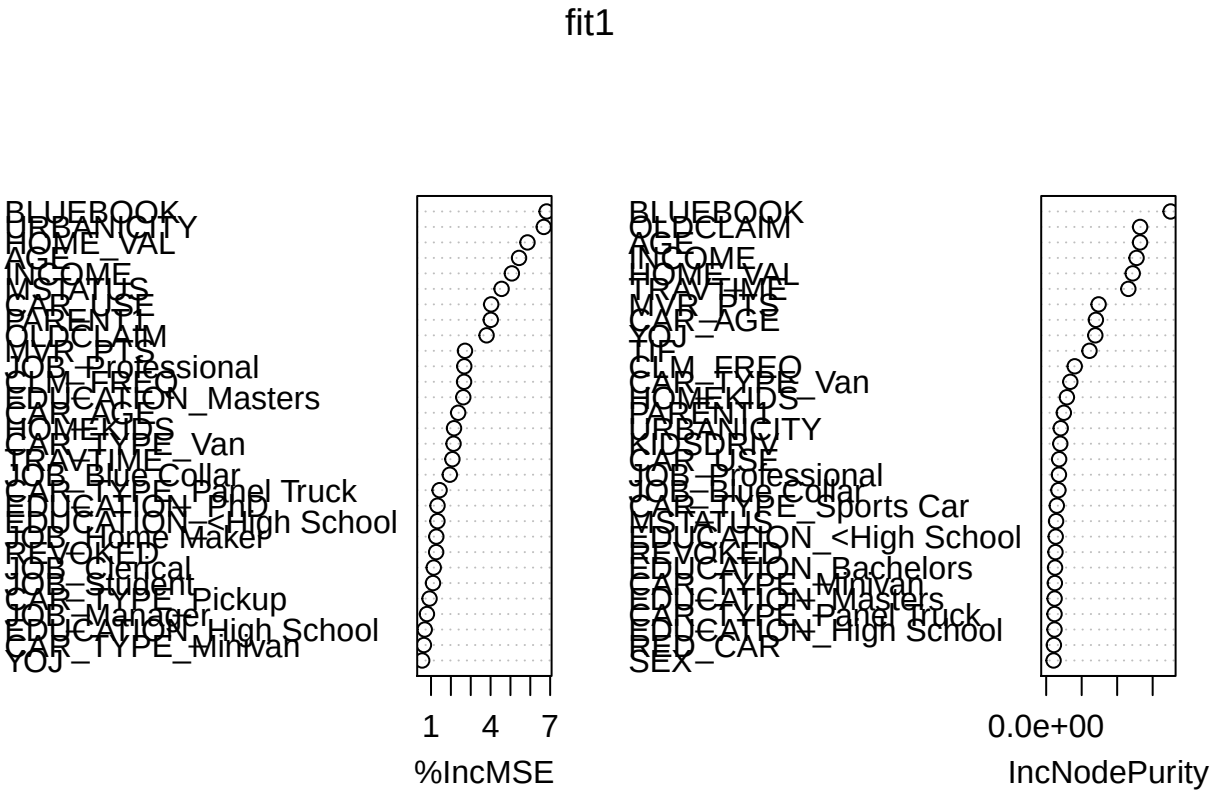


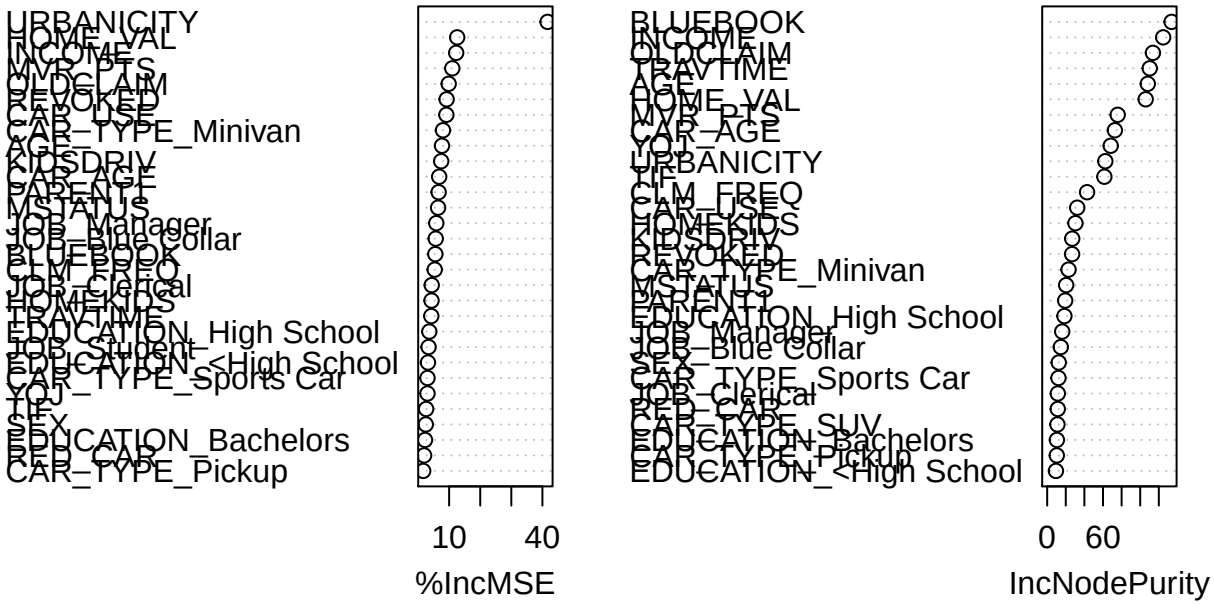
Correlation

We can see our correlation matrix below. A dark blue circle represents a strong positive relationship and a dark red circle represents a strong negative relationship between two variables.



Finally, we can use the `randomforest` package to verify our assumptions from the correlation plot.





PART 3. BUILD MODELS

Using the transformed data above, we developed two multiple linear regression and three binary logistic regression models. Through these models, we hope to predict **(1)** the probability that a person will crash their car and **(2)** the amount of money it will cost if the person does crash their car.

Multiple Linear Regression

MLR for `TARGET_AMT`.

MLR 1

The following MLR model seeks to predict `TARGET_AMT` using all variables.

```
FALSE
FALSE Call:
FALSE lm(formula = TARGET_AMT ~ ., data = lm_train)
FALSE
FALSE Residuals:
FALSE    Min     1Q Median     3Q    Max
FALSE  -5893  -1697   -761    345 103780
FALSE
```

```
FALSE Coefficients: (2 not defined because of singularities)
FALSE                              Estimate Std. Error t value Pr(>|t|)
FALSE (Intercept)                 7.781e+02  5.878e+02   1.324  0.18566
FALSE CAR_TYPE_Minivan1          -5.152e+02  2.132e+02  -2.416  0.01570 *
FALSE `CAR_TYPE_Panel Truck`1    -2.499e+02  2.632e+02  -0.949  0.34241
FALSE CAR_TYPE_Pickup1           -1.397e+02  2.211e+02  -0.632  0.52753
FALSE `CAR_TYPE_Sports Car`1      5.060e+02  2.954e+02   1.713  0.08678 .
FALSE CAR_TYPE_SUV1               2.363e+02  2.671e+02   0.885  0.37628
FALSE CAR_TYPE_Van1                      NA         NA      NA       NA
FALSE `EDUCATION_<High School`1  -2.865e+02  3.560e+02  -0.805  0.42090
FALSE EDUCATION_Bachelors1       -5.450e+02  2.888e+02  -1.887  0.05922 .
FALSE `EDUCATION_High School`1   -3.755e+02  3.224e+02  -1.165  0.24419
FALSE EDUCATION_Masters1         -2.637e+02  2.541e+02  -1.037  0.29955
FALSE EDUCATION_PhD1                     NA         NA      NA       NA
FALSE `JOB_Blue Collar`1          5.076e+02  3.219e+02   1.577  0.11483
FALSE JOB_Clerical1               5.292e+02  3.414e+02   1.550  0.12115
FALSE JOB_Doctor1                -5.011e+02  4.087e+02  -1.226  0.22023
FALSE `JOB_Home Maker`1           3.501e+02  3.646e+02   0.960  0.33683
FALSE JOB_Lawyer1                 2.310e+02  2.956e+02   0.781  0.43469
FALSE JOB_Manager1               -4.791e+02  2.885e+02  -1.661  0.09677 .
FALSE JOB_Professional1           4.564e+02  3.088e+02   1.478  0.13939
FALSE JOB_Student1                2.856e+02  3.739e+02   0.764  0.44504
FALSE KIDSDRIV                    3.142e+02  1.132e+02   2.776  0.00551 **
FALSE AGE                         5.274e+00  7.065e+00   0.747  0.45537
FALSE HOMEKIDS                    7.796e+01  6.535e+01   1.193  0.23295
FALSE YOJ                        -4.505e+00  1.509e+01  -0.298  0.76537
FALSE INCOME                     -4.399e-03  1.804e-03  -2.438  0.01478 *
FALSE PARENT11                    5.762e+02  2.020e+02   2.852  0.00435 **
FALSE HOME_VAL                   -5.620e-04  5.907e-04  -0.951  0.34138
FALSE MSTATUS1                   -5.689e+02  1.448e+02  -3.929 8.61e-05 ***
FALSE SEX1                       -3.683e+02  1.838e+02  -2.004  0.04512 *
FALSE TRAVTIME                    1.195e+01  3.222e+00   3.708  0.00021 ***
FALSE CAR_USE1                   -7.794e+02  1.644e+02  -4.741 2.17e-06 ***
FALSE BLUEBOOK                    1.430e-02  8.621e-03   1.659  0.09714 .
FALSE TIF                        -4.820e+01  1.218e+01  -3.958 7.63e-05 ***
FALSE RED_CAR1                   -4.800e+01  1.491e+02  -0.322  0.74743
FALSE OLDCLAIM                   -1.055e-02  7.436e-03  -1.418  0.15617
FALSE CLM_FREQ                    1.417e+02  5.503e+01   2.576  0.01002 *
FALSE REVOKED1                    5.495e+02  1.735e+02   3.167  0.00155 **
FALSE MVR_PTS                     1.752e+02  2.592e+01   6.758 1.50e-11 ***
FALSE CAR_AGE                    -2.687e+01  1.280e+01  -2.098  0.03590 *
FALSE URBANICITY1                 1.664e+03  1.394e+02  11.942  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE Residual standard error: 4544 on 8123 degrees of freedom
FALSE Multiple R-squared:  0.07104, Adjusted R-squared:  0.06681
FALSE F-statistic: 16.79 on 37 and 8123 DF,  p-value: < 2.2e-16
```

Model Analysis

The low f-statistic suggests a weak relationship when using all predicted and response variables. The Adjusted R^2 value of 0.06681 means that only 6.681% of the variance observed in the `TARGET_AMT` variable can be explained when using all response variables.

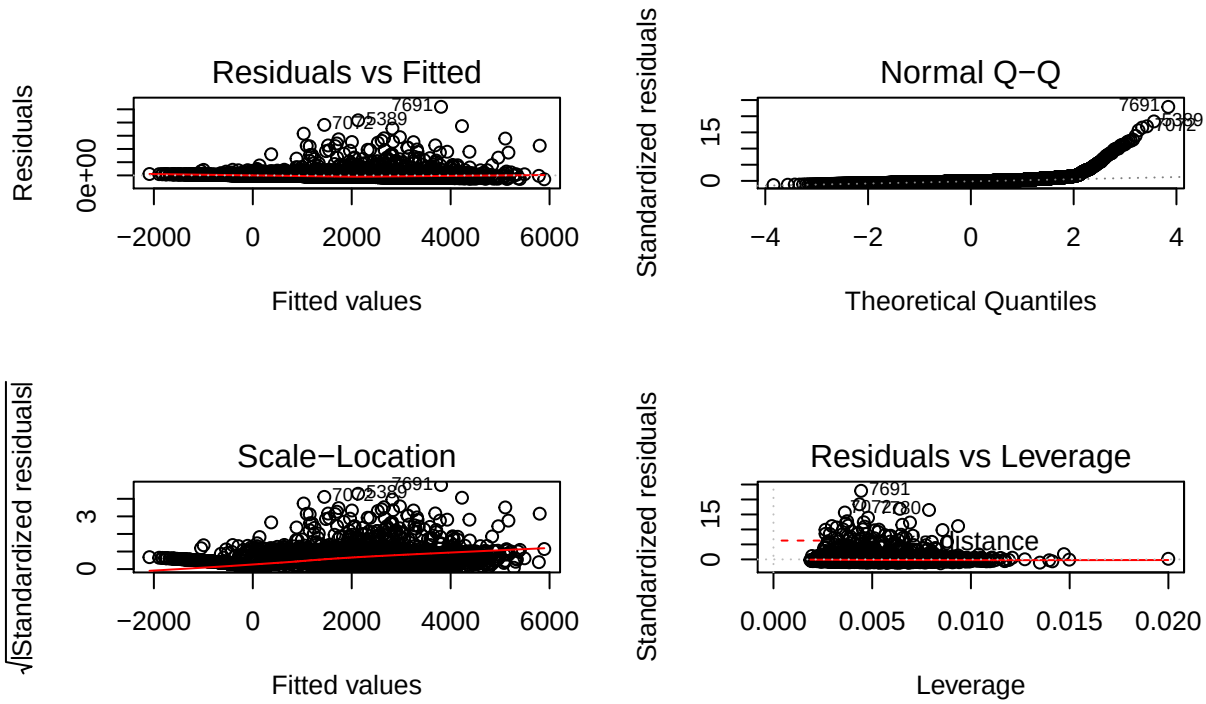
Additionally, all variables show small t-statistics, which means there is a higher degree of variability in the coefficient estimates for all variables with the exceptions of the following:

	tstat
MVR_PTS	6.757625
URBANICITY1	11.941701

From this model, only the following variables had significance levels less than 0.05%.

	pval
CAR_TYPE_Minivan1	0.0157
KIDSDRIV	0.0055
INCOME	0.0148
PARENT11	0.0044
MSTATUS1	0.0001
SEX1	0.0451
TRAVTIME	0.0002
CAR_USE1	0.0000
TIF	0.0001
CLM_FREQ	0.0100
REVOKED1	0.0015
MVR_PTS	0.0000
CAR_AGE	0.0359
URBANICITY1	0.0000

The residual plots below show that our data as is should not be used for linear modeling without futher data transformations. The normality assumption for linear regression is not met as the residuals do not follow a straight line and the data does not meet homoscedastic assumption. The line is not horizontal and the residuals are not randomly-distributed around it.



MLR 2

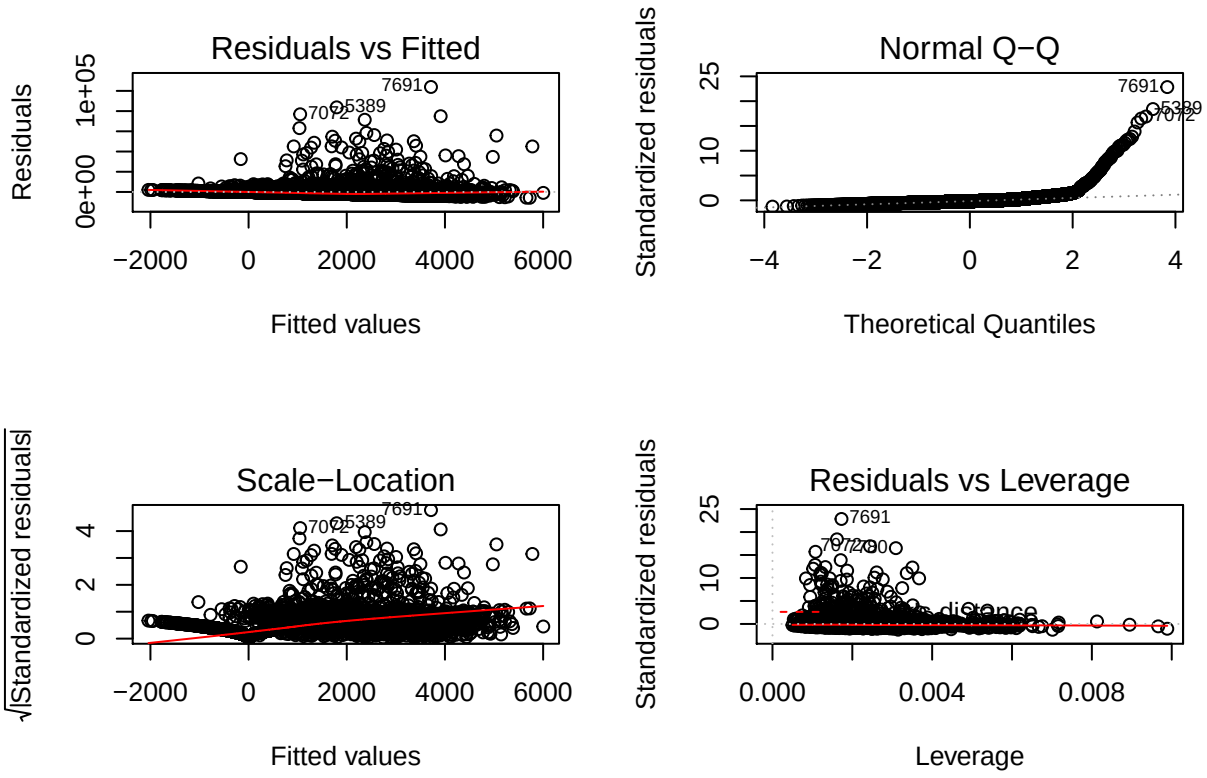
The following model is based off of the variables in model 1 with high t- and significant p-values.

```
FALSE
FALSE Call:
FALSE lm(formula = TARGET_AMT ~ CAR_TYPE_Minivan + KIDSDRIV + INCOME +
FALSE     PARENT1 + MSTATUS + SEX + TRAVTIME + CAR_USE + TIF + CLM_FREQ +
FALSE     REVOKED + MVR_PTS + CAR_AGE + URBANICITY, data = lm_train)
FALSE
FALSE Residuals:
FALSE    Min     1Q Median     3Q    Max
FALSE  -5727  -1688   -795    292 103871
FALSE
FALSE Coefficients:
FALSE                      Estimate Std. Error t value Pr(>|t|)
FALSE (Intercept)         1.205e+03  2.327e+02   5.177 2.31e-07 ***
FALSE CAR_TYPE_Minivan1 -5.435e+02  1.233e+02  -4.409 1.05e-05 ***
FALSE KIDSDRIV           3.754e+02  1.021e+02   3.677 0.000238 ***
FALSE INCOME            -5.955e-03  1.220e-03  -4.883 1.07e-06 ***
FALSE PARENT11           6.556e+02  1.764e+02   3.716 0.000204 ***
FALSE MSTATUS1          -5.863e+02  1.193e+02  -4.916 9.02e-07 ***
FALSE SEX1              -6.180e+01  1.114e+02  -0.555 0.579220
FALSE TRAVTIME           1.278e+01  3.219e+00   3.970 7.25e-05 ***
FALSE CAR_USE1          -7.751e+02  1.149e+02  -6.748 1.60e-11 ***
FALSE TIF               -4.763e+01  1.217e+01  -3.913 9.20e-05 ***
FALSE CLM_FREQ           1.159e+02  4.878e+01   2.375 0.017557 *
FALSE REVOKED1           4.673e+02  1.549e+02   3.016 0.002568 **
FALSE MVR_PTS            1.797e+02  2.578e+01   6.971 3.39e-12 ***
FALSE CAR_AGE           -3.615e+01  1.004e+01  -3.600 0.000320 ***
```

```
FALSE URBANICITY1        1.516e+03  1.356e+02  11.176  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE Residual standard error: 4552 on 8146 degrees of freedom
FALSE Multiple R-squared:  0.06522, Adjusted R-squared:  0.06361
FALSE F-statistic:  40.6 on 14 and 8146 DF,  p-value: < 2.2e-16
```

Model Analysis

Changing the coefficients increased our f-statistic, however the R-square value is still too low to make inference from this model. Our residual plots still show our model is not appropriate for linear regression.



Binary Logistic Regression

MODEl 1

MODEl 2

AIC, BIC, LOglik, pseudoR2

Choosing the best model and applying it on test data set

Assessing the performance of the choosen model: ROC and AUC

PART 4: SELECT MODELS

Select the best multiple linear regression model and the best binary logistic regression model. Discuss why you selected your models.

MLR Evaluation

Use a metric such as Adjusted R2, RMSE, etc. Explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output.

Using the training data set, evaluate the multiple linear regression model based on:

Mean Squared Error

R2

F-statistic

Residual plots.

BLR Evaluation

Use a metric such as log likelihood, AIC, ROC curve, etc. Using the training data set, evaluate the binary logistic regression model based on:

Accuracy

Classification Error Rate

Precision

Sensitivity

Specificity

F1 score

AUC

Confusion Matrix

Predictions

Make predictions using the evaluation data set.