---

title: "Project4_621"

author: "Nnaemezue Obi-Eyisi"

date: "April 21, 2018"

output: html_document

BUILDING A MULTIPLE AND BINARY LOGISTIC REGRESSION MODEL TO PREDICT THE PROBABILITY THAT A PERSON WILL CRASH THEIR CAR AND THE AMOUNT OF MONEY IT WILL COST IF THE PERSON DOES CRASH THEIR CAR

INTRODUCTION

The goal of this project is to explore, analyze and model the data set containing information about a customer in an auto insurance company, then we would create a binary logistic regression model to predict whether the customer was in a car crash or not. If the customer was in a car crash we would then predict the cost of the crash using a linear regression model. This dataset contains about 23 predictor variables, for two target dependent variable.

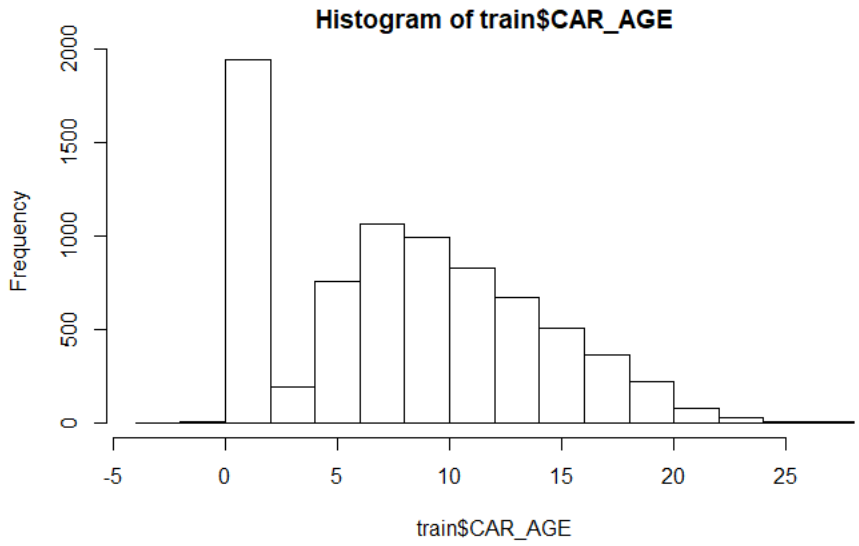
The plan is to use a Logit model for regression prediction then use linear regression for cost of crash.

DATA EXPLORATION

I did some summary statistics on the dataset, below is a table describing the mean, median etc of the predictors. We have 8161 observations where 26% are of cars in a crash. We see a lot of skew in various variables for example Home_Val, OLDCLAIMS, REVOKED, etc. There is some missing values in certain variables that need to be analyzed. This table translates categorical variables to numeric which is not ideal, a summary statistics function would display the true summary of our data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
INDEX	1	8161	5151.87	2978.89	5133	5151.93	3841.42	1	10302.0	10301.0	0.00	-1.20	32.97
TARGET_FLAG	2	8161	0.26	0.44	0	0.20	0.00	0	1.0	1.0	1.07	-0.85	0.00
TARGET_AMT	3	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107586.1	8.71	112.29	52.07
KIDSDRIV	4	8161	0.17	0.51	0	0.03	0.00	0	4.0	4.0	3.35	11.78	0.01
AGE	5	8155	44.79	8.63	45	44.83	8.90	16	81.0	65.0	-0.03	-0.06	0.10
HOMEKIDS	6	8161	0.72	1.12	0	0.50	0.00	0	5.0	5.0	1.34	0.65	0.01
YOJ	7	7707	10.50	4.09	11	11.07	2.97	0	23.0	23.0	-1.20	1.18	0.05
INCOME*	8	8161	2875.55	2090.68	2817	2816.95	2799.15	1	6613.0	6612.0	0.11	-1.29	23.14
PARENT1*	9	8161	1.13	0.34	1	1.04	0.00	1	2.0	1.0	2.17	2.73	0.00
HOME_VAL*	10	8161	1684.89	1697.38	1245	1516.50	1842.87	1	5107.0	5106.0	0.52	-1.18	18.79
MSTATUS*	11	8161	1.40	0.49	1	1.38	0.00	1	2.0	1.0	0.41	-1.83	0.01
SEX*	12	8161	1.54	0.50	2	1.55	0.00	1	2.0	1.0	-0.14	-1.98	0.01
EDUCATION*	13	8161	3.09	1.44	3	3.11	1.48	1	5.0	4.0	0.12	-1.38	0.02
JOB*	14	8161	5.69	2.68	6	5.81	2.97	1	9.0	8.0	-0.31	-1.22	0.03
TRAVTIME	15	8161	33.49	15.91	33	33.00	16.31	5	142.0	137.0	0.45	0.66	0.18
CAR_USE*	16	8161	1.63	0.48	2	1.66	0.00	1	2.0	1.0	-0.53	-1.72	0.01
BLUEBOOK*	17	8161	1283.62	893.51	1124	1259.57	1132.71	1	2789.0	2788.0	0.25	-1.36	9.89
TIF	18	8161	5.35	4.15	4	4.84	4.45	1	25.0	24.0	0.89	0.42	0.05
CAR_TYPE*	19	8161	3.53	1.97	3	3.54	2.97	1	6.0	5.0	0.00	-1.52	0.02
RED_CAR*	20	8161	1.29	0.45	1	1.24	0.00	1	2.0	1.0	0.92	-1.16	0.01
OLDCLAIM*	21	8161	552.27	862.20	1	380.32	0.00	1	2857.0	2856.0	1.31	0.25	9.54
CLM_FREQ	22	8161	0.80	1.16	0	0.59	0.00	0	5.0	5.0	1.21	0.28	0.01
REVOKED*	23	8161	1.12	0.33	1	1.03	0.00	1	2.0	1.0	2.30	3.30	0.00
MVR_PTS	24	8161	1.70	2.15	1	1.31	1.48	0	13.0	13.0	1.35	1.38	0.02
CAR_AGE	25	7651	8.33	5.70	8	7.96	7.41	-3	28.0	31.0	0.28	-0.75	0.07
URBANICITY*	26	8161	1.20	0.40	1	1.13	0.00	1	2.0	1.0	1.46	0.15	0.00

Summary stats output discovery



We also noticed a set of predictors with dollar signs which needed to be converted to numeric for easier analysis

We should convert the currency data in the below columns to numeric

INCOME

HOME_VAL

BLUEBOOK

OLDCLAIM

We used a replace and convert function to make then suitable for regression modelling.

Then we notice that we have a lot of missing values in the monetary predictors.

1.INCOME

2.HOME_VAL

3.JOB

The tactic used to impute these values was by inferring the INCOME from the average income for the missing values occupation.

Occupation type	**Average income**
	118852.938
Clerical	33861.186
Doctor	128679.697
Home Maker	12073.336
Lawyer	88304.8
Manager	87461.555
Professional	76593.096
Student	6309.657
z_Blue Collar	58957.012

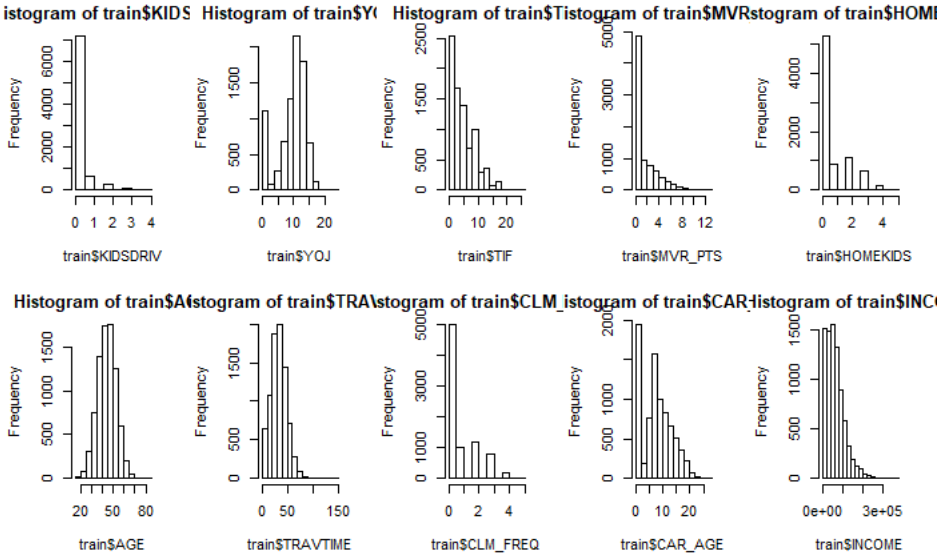
For records with blank Home_Val values we would impute the missing values to zero, because we assume these are people renting

The JOB predictor was left as it is.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
INDEX	1	8161	5151.87	2978.89	5133	5151.93	3841.42	1	10302.0	10301.0	0.00	-1.20	32.97
TARGET_FLAG	2	8161	0.26	0.44	0	0.20	0.00	0	1.0	1.0	1.07	-0.85	0.00
TARGET_AMT	3	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107586.1	8.71	112.29	52.07
KIDSDRIV	4	8161	0.17	0.51	0	0.03	0.00	0	4.0	4.0	3.35	11.78	0.01
AGE	5	8161	44.79	8.62	45	44.83	8.90	16	81.0	65.0	-0.03	-0.06	0.10
HOMEKIDS	6	8161	0.72	1.12	0	0.50	0.00	0	5.0	5.0	1.34	0.65	0.01
YOJ	7	8161	9.92	4.65	11	10.44	2.97	0	23.0	23.0	-1.04	0.20	0.05
INCOME	8	8161	61602.56	46931.04	54877	56714.85	41927.93	0	367030.0	367030.0	1.18	2.18	519.50
PARENT1*	9	8161	1.13	0.34	1	1.04	0.00	1	2.0	1.0	2.17	2.73	0.00
HOME_VAL	10	8161	146062.19	130426.72	151957	133608.69	177081.74	0	885282.0	885282.0	0.55	-0.06	1443.76
MSTATUS*	11	8161	1.40	0.49	1	1.38	0.00	1	2.0	1.0	0.41	-1.83	0.01
SEX*	12	8161	1.54	0.50	2	1.55	0.00	1	2.0	1.0	-0.14	-1.98	0.01
EDUCATION*	13	8161	3.09	1.44	3	3.11	1.48	1	5.0	4.0	0.12	-1.38	0.02
JOB*	14	8161	5.69	2.68	6	5.81	2.97	1	9.0	8.0	-0.31	-1.22	0.03
TRAVTIME	15	8161	33.49	15.91	33	33.00	16.31	5	142.0	137.0	0.45	0.66	0.18
CAR_USE*	16	8161	1.63	0.48	2	1.66	0.00	1	2.0	1.0	-0.53	-1.72	0.01
BLUEBOOK	17	8161	15709.90	8419.73	14440	15036.89	8450.82	1500	69740.0	68240.0	0.79	0.79	93.20
TIF	18	8161	5.35	4.15	4	4.84	4.45	1	25.0	24.0	0.89	0.42	0.05
CAR_TYPE*	19	8161	3.53	1.97	3	3.54	2.97	1	6.0	5.0	0.00	-1.52	0.02
RED_CAR*	20	8161	1.29	0.45	1	1.24	0.00	1	2.0	1.0	0.92	-1.16	0.01
OLDCLAIM	21	8161	4037.08	8777.14	0	1719.29	0.00	0	57037.0	57037.0	3.12	9.86	97.16
CLM_FREQ	22	8161	0.80	1.16	0	0.59	0.00	0	5.0	5.0	1.21	0.28	0.01
REVOKED*	23	8161	1.12	0.33	1	1.03	0.00	1	2.0	1.0	2.30	3.30	0.00
MVR_PTS	24	8161	1.70	2.15	1	1.31	1.48	0	13.0	13.0	1.35	1.38	0.02
CAR_AGE	25	8161	8.31	5.52	8	7.96	5.93	0	28.0	28.0	0.30	-0.60	0.06
URBANICITY*	26	8161	1.20	0.40	1	1.13	0.00	1	2.0	1.0	1.46	0.15	0.00

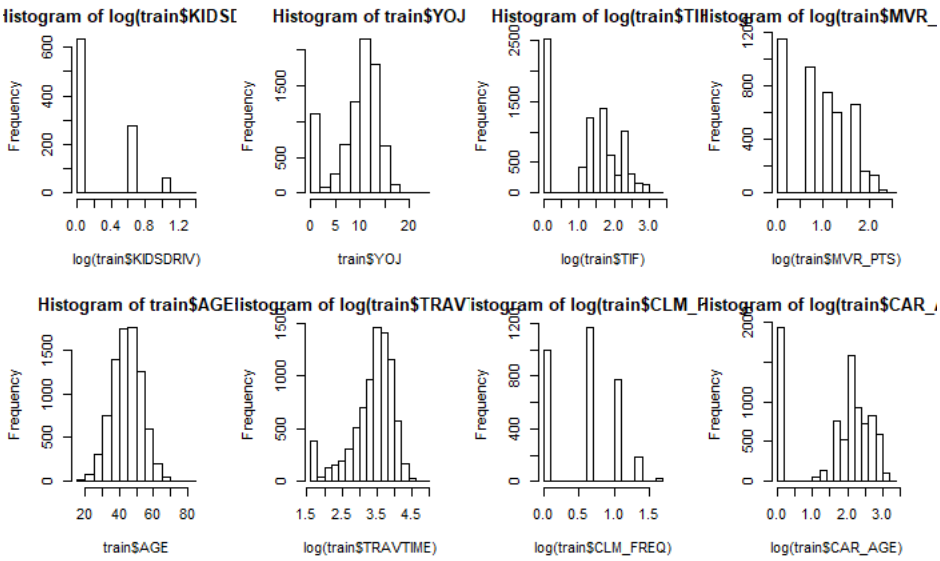
For the multiple linear regression model, one of the assumptions to build in a regression model is the normality of predictors

Investigate normality of predictors



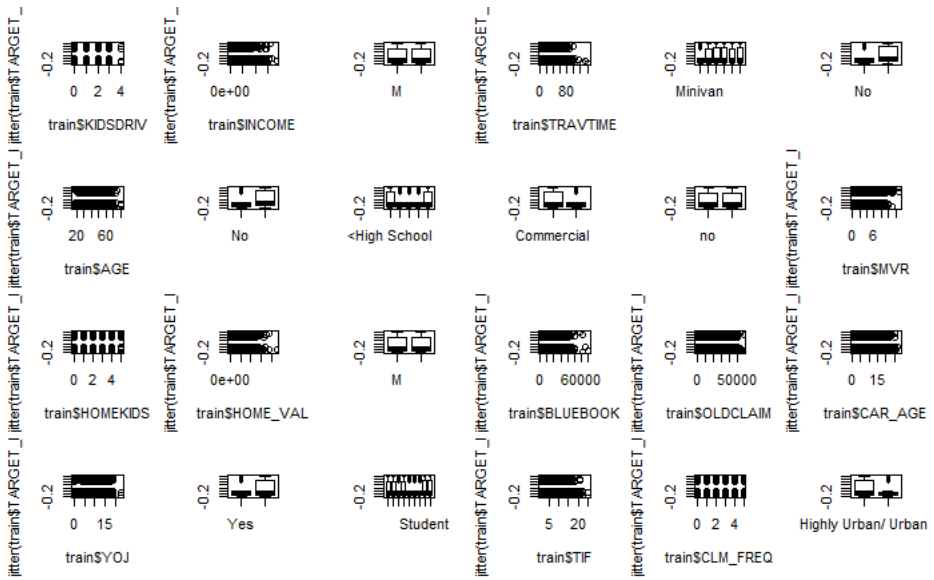
We notice some of the predictors are normally distributed while others are skewed.

I decided to try transforming some variable to using log to see how their distribution will play out. I noticed that only one predictor really had a decent normal distribution after log transformation

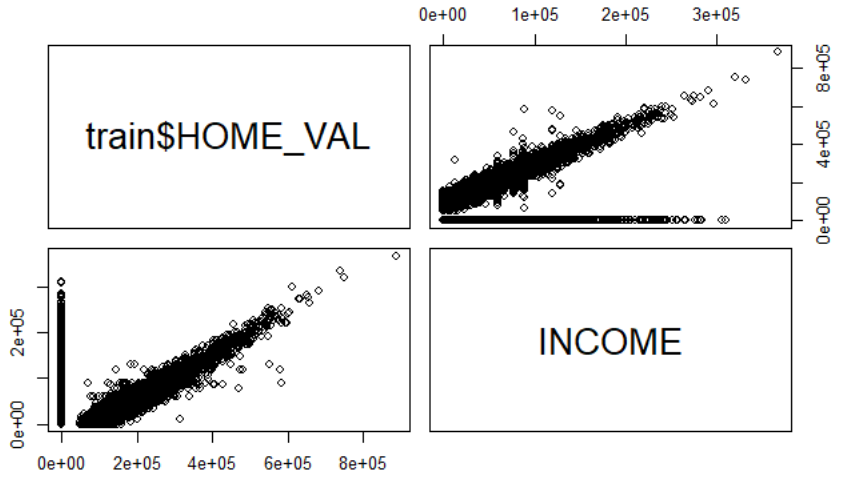


Let's study correlation between variables and predictors

All predictors versus Target flag



We noticed that some variables are highly correlated with each other. We will try to analyze it further



DATA PREPARATION

To prepare the data for analysis my first step was to convert the monetary predictors to numeric, since the model was interpreting it as factors

We should convert the currency data in the below columns to numeric

1. INCOME
2. HOME_VAL
3. BLUEBOOK
4. OLDCLAIM

We used a replace and convert function to make then suitable for regression modelling.

Then we notice that we have a lot of missing values in the monetary predictors.

1.INCOME

2.HOME_VAL

3.JOB

The tactic used to impute these values was by inferring the INCOME from the average income for the missing values occupation.

Occupation type	**Average income**
	118852.938
Clerical	33861.186

Doctor	128679.697
Home Maker	12073.336
Lawyer	88304.8
Manager	87461.555
Professional	76593.096
Student	6309.657
z_Blue Collar	58957.012

For records with blank Home_Val values we would impute the missing values to zero, because we assume these are people renting

We have a couple of predictors with missing values like below

1. AGE

2. YOJ

3. CAR_AGE

For AGE predictor, we will use the Hmisc package Impute function to impute the missing AGE with mean from the data set. We also round it off to the nearest whole number to be consistent with the numeric precision of the dataset.

For YOJ predictor, we will replace it with 0 because when we analyzed the dataset we noticed that the occupation of customers with a missing 'Years on Job' predictor is 'Student'. I also guessed that most likely it will be missing for any individual that is currently unemployed or took up a new position in less than a year

For CAR_AGE predictor, we analyzed the histogram and noticed a negative value. I then proceeded to revert to the absolute value of that number. For observation where the CAR_AGE is missing we imputed the missing value with the median

To decide which variable needed transformation we decided to run a logit regression model on all predictors to see what was significant and what wasn't

```
Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + INCOME + PARENT1 +
    HOME_VAL + MSTATUS + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
    OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + BLUEBOOK + CAR_AGE +
    EDUCATION + JOB + YOJ + RED_CAR + URBANICITY + SEX, family =
binomial(link = "logit"),
    data = train)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.6371  -0.7156  -0.3952   0.6372   3.1572

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -8.725e-01  3.595e-01  -2.427 0.015216 *
KIDSDRIV                         4.363e-01  6.458e-02   6.755 1.43e-11 ***
AGE                             -4.326e-03  4.313e-03  -1.003 0.315874
INCOME                          -3.657e-06  1.279e-06  -2.860 0.004240 **
PARENT1Yes                       4.863e-01  1.140e-01   4.267 1.98e-05 ***
HOME_VAL                        -8.428e-07  3.672e-07  -2.295 0.021709 *
MSTATUSz_No                      5.807e-01  9.125e-02   6.364 1.97e-10 ***
TRAVTIME                         1.581e-02  2.195e-03   7.201 5.98e-13 ***
CAR_USEPrivate                  -7.254e-01  1.063e-01  -6.821 9.07e-12 ***
TIF                             -5.431e-02  8.435e-03  -6.439 1.20e-10 ***
CAR_TYPEPanel Truck              6.397e-01  1.862e-01   3.435 0.000593 ***
CAR_TYPEPickup                   5.676e-01  1.169e-01   4.855 1.21e-06 ***
CAR_TYPESports Car               1.117e+00  1.509e-01   7.405 1.31e-13 ***
CAR_TYPEVan                      6.229e-01  1.478e-01   4.215 2.50e-05 ***
CAR_TYPEz_SUV                    8.460e-01  1.297e-01   6.525 6.79e-11 ***
OLDCLAIM                        -1.470e-05  4.562e-06  -3.222 0.001271 **
CLM_FREQ                         1.872e-01  3.333e-02   5.616 1.96e-08 ***
REVOKEDYes                       8.810e-01  1.066e-01   8.264  < 2e-16 ***
MVR_PTS                          1.181e-01  1.581e-02   7.471 7.93e-14 ***
BLUEBOOK                        -2.391e-05  6.065e-06  -3.943 8.06e-05 ***
CAR_AGE                          1.829e-03  8.680e-03   0.211 0.833128
EDUCATIONBachelors              -3.573e-01  1.336e-01  -2.674 0.007488 **
EDUCATIONMasters                -3.597e-01  2.058e-01  -1.748 0.080452 .
EDUCATIONPhD                    -1.850e-01  2.470e-01  -0.749 0.453765
EDUCATIONz_High School           2.142e-02  1.103e-01   0.194 0.846037
JOBClerical                      3.575e-01  2.283e-01   1.566 0.117458
JOBDoctor                       -4.819e-01  3.034e-01  -1.588 0.112209
JOBHome Maker                    1.875e-01  2.425e-01   0.773 0.439355
JOBLawyer                        1.190e-01  1.943e-01   0.612 0.540269
JOBManager                      -5.660e-01  1.986e-01  -2.851 0.004364 **
JOBProfessional                  1.129e-01  2.070e-01   0.546 0.585331
JOBStudent                       2.400e-01  2.478e-01   0.968 0.332879
JOBz_Blue Collar                 2.613e-01  2.147e-01   1.217 0.223597
YOJ                             -1.031e-02  8.097e-03  -1.274 0.202722
RED_CARyes                       1.882e-02  9.923e-02   0.190 0.849540
URBANICITYz_Highly Rural/ Rural -2.414e+00  1.307e-01 -18.475  < 2e-16 ***
SEXz_F                          -1.656e-01  1.300e-01  -1.274 0.202754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7100.8  on 6120  degrees of freedom
Residual deviance: 5470.1  on 6084  degrees of freedom
AIC: 5544.1

Number of Fisher Scoring iterations: 5
```

We noticed that the below predictors are not significant and could you some transformation

1. AGE
2. CAR_AGE
3. EDUCATION Categorical
4. JOB Categorical
5. YOJ

6. RED_CAR
7. SEX

We decided to LOG transform the AGE, CAR_AGE and YOJ because I noticed some skewness during data exploration. With the transformation applied I noticed some slight improved for the McFadden Rsquared of the Logistic model.

BUILD MODELS

We would first build a logistic regression model to predict whether the customer's car will be in a car crash, if it is in a crash we would then predict how much the crash will cost with a linear regression model.

LOGISTIC MODEL 1

We would use all variables without transformation.

```
glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + INCOME + PARENT1 +
    HOME_VAL + MSTATUS + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
    OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + BLUEBOOK + CAR_AGE +
    EDUCATION + JOB + YOJ + RED_CAR + URBANICITY + SEX, family =
binomial(link = "logit"),
    data = train)
```

When we analyze the coefficients, we see that some of the most influential predictors to car crash are CAR_TYPESports Car with a coefficient of 1.117. This means that sports cars are more likely to have a car crash. Another predictor with weight is URBANICITYz_Highly Rural/ Rural with a -2.414 coefficient this means that cars in rural areas are less likely to get into a car accident. Looking at the other predictors their coefficients seem to make theoretical sense. I am not sure if I agree with females being less likely to be involved in a car crash

LOGISTIC MODEL 2

We used all the variables again except that we transformed the AGE and CAR_AGE predictors with log transform

```
Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + log(AGE) + INCOME + PARENT1 +
    HOME_VAL + MSTATUS + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
    OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + BLUEBOOK + EDUCATION +
    log(CAR_AGE + 1) + JOB + YOJ + RED_CAR + URBANICITY + SEX,
    family = binomial(link = "logit"), data = train)
```

After transforming the AGE predictor, I noticed that the significance level got better and it still had the same interpretation as the older people are less likely to be in a car crash

Another predictor that I transformed is CAR_AGE to log(CAR_AGE + 1). This was not really effective because the significance level of the predictor was not really improved. However, I got a lower AIC and improved Rsquared.

LOGISTIC MODEL 3

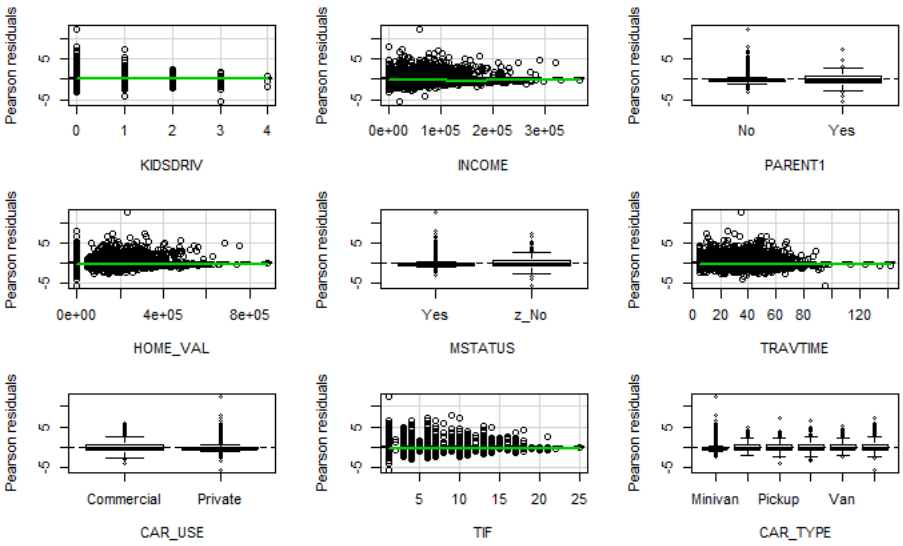
We used all the variables again but transformed additional variables

```
glm(formula = TARGET_FLAG ~ KIDSDRIV + log(AGE) + INCOME + PARENT1 +
    HOME_VAL + MSTATUS + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
    OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + BLUEBOOK + EDUCATION +
    log(CAR_AGE + 1) + JOB + log(YOJ + 1) + URBANICITY + SEX,
    family = binomial(link = "logit"), data = train)
```

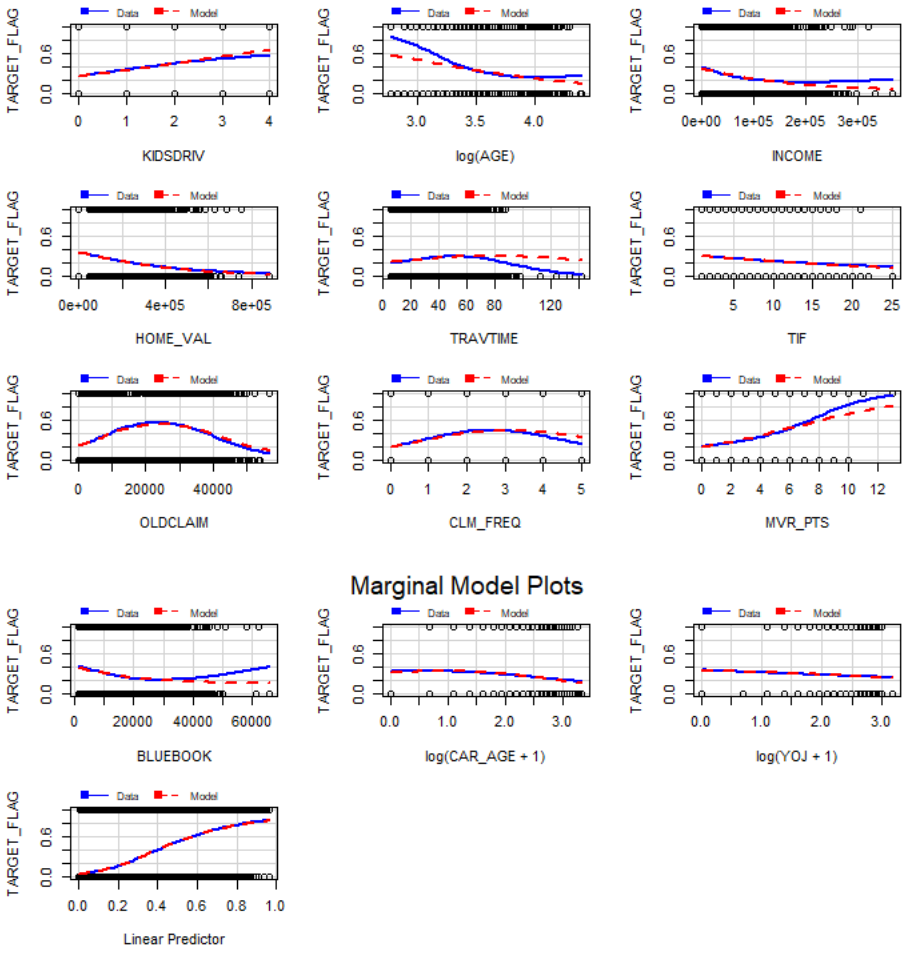
In this case we transformed YOJ predictor (Years on Job) with log transform and we noticed some improvement with the significance level up to 0.07, the coefficient was still negative implying that the more years on a job then less likely the customer will have a car crash.

We also noticed some improvement in the McFadden Rsquared to 0.2302

Analyzing some model diagnostics we could visualize the errors like below



We also looked at the Marginal Model plots



We see that certain predictors are not properly fitted with the model.

MULTIPLE LINEAR REGRESSION MODEL 1

```
lm(formula = TARGET_AMT ~ TARGET_FLAG + KIDSDRIV + log(AGE) +
    INCOME + PARENT1 + log(HOME_VAL + 1) + MSTATUS + TRAVTIME +
    CAR_USE + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
    MVR_PTS + BLUEBOOK + EDUCATION + CAR_AGE + JOB + YOJ + URBANICITY +
    SEX, data = train)
```

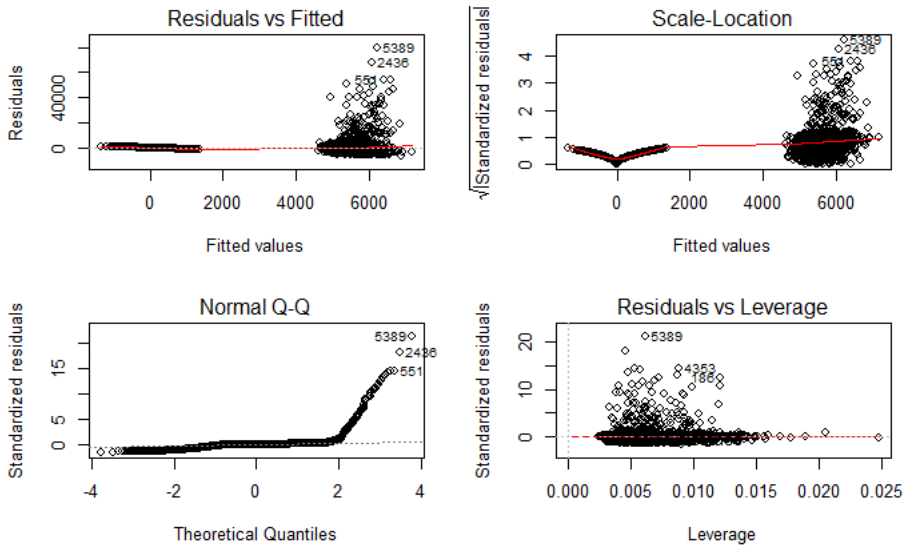
Building on the logistic regression model, I added the Target_Flag as a predictor for logistic model because we assume that we can only predict the car crash cost if only there was a crash in the first place. I additionally transformed some variables like the log of the HOME_VAL predictor.

Looking at the significant levels they are not significant according to the alpha of 0.05. While I analyzed the coefficients of some predictors, they seemed to correspond to the theoretical assumption.

I also noted that the most significant predictors are the BLUEBOOK and Target_FLAG, in determining the cost of the crash. Other significant predictors are

CLM_FREQ
CAR_AGE
REVOKEDYes
MVR_PTS
SEX

I didn't really understand why the payout of the crash is less when the license has been revoked. Also, the CLM_FREQ implies that the more the customer has filled claims in the past the less the crash payout. Though it makes sense. Finally, if it is a female car owner the crash payout will be less.



MULTIPLE LINEAR REGRESSION MODEL 5

In this case we include all the predictors and transformed certain predictors and added interactions between predictors

```
lm(formula = TARGET_AMT ~ TARGET_FLAG + INCOME + MVR_PTS + CLM_FREQ +
    OLDCLAIM + CLM_FREQ * OLDCLAIM + REVOKED + CAR_USE + TIF +
    BLUEBOOK * CAR_AGE + CAR_AGE + AGE * PARENT1 + KIDSDRIV +
    TRAVTIME + HOME_VAL * INCOME + EDUCATION * JOB + MSTATUS +
    CAR_TYPE * CAR_USE + JOB + YOJ + RED_CAR + URBANICITY + SEX,
    data = train)
```

The interactions are

CLM_FREQ * OLDCLAIM

BLUEBOOK * CAR_AGE

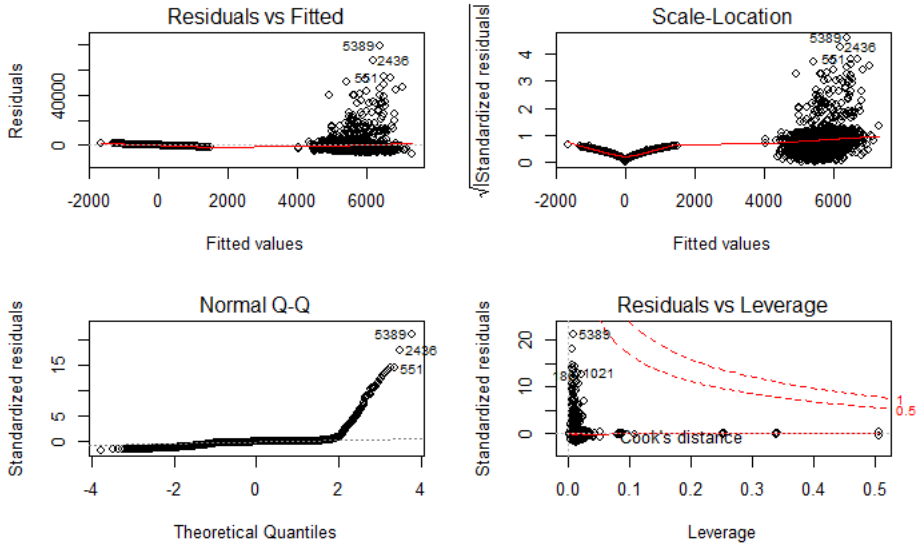
AGE * PARENT1

HOME_VAL * INCOME

CAR_TYPE * CAR_USE

These interactions were added because of my suspicion in the correlation between the variable even though most of them were not significant.

This led to an improvement in adjusted Rsquared
Adjusted R-squared: 0.3087



SELECT MODELS

Evaluation of Logistic Regression Model

After deliberation I decided to go with the Model #3 for my logistic regression model. I chose this model because it had the maximum McFadden Rsquared 0.2302 with most predictors with significance level.

```
    Null deviance: 7100.8  on 6120  degrees of freedom
Residual deviance: 5465.7  on 6085  degrees of freedom
AIC: 5537.7
Number of Fisher Scoring iterations: 5
```

Using the test data set, I tried to validate the logistic model

Confusion Matrix and Statistics

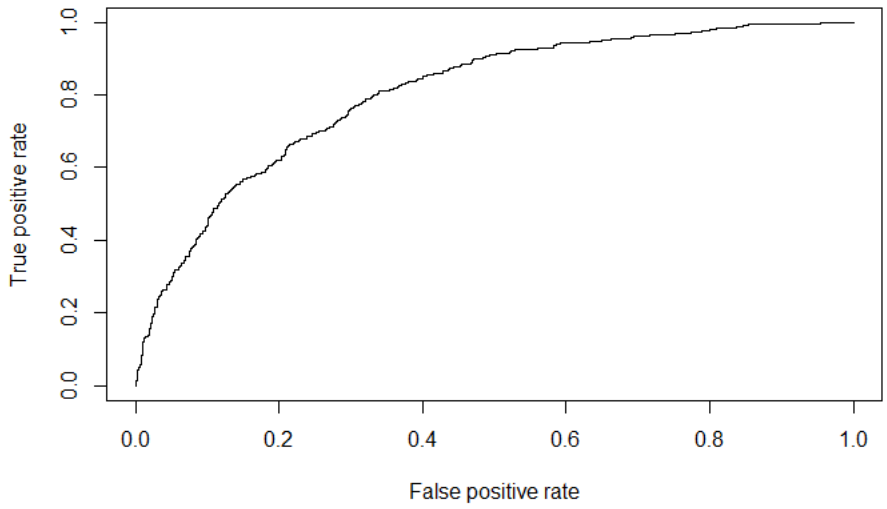
```
          Reference
Prediction    0    1
         0 1391  312
         1  129  208

               Accuracy : 0.7838
                 95% CI : (0.7653, 0.8015)
    No Information Rate : 0.7451
    P-Value [Acc > NIR] : 2.502e-05

                  Kappa : 0.3564
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4000
            Specificity : 0.9151
         Pos Pred Value : 0.6172
         Neg Pred Value : 0.8168
             Prevalence : 0.2549
         Detection Rate : 0.1020
   Detection Prevalence : 0.1652
      Balanced Accuracy : 0.6576

       'Positive' Class : 1
```

Area under the curve is 0.8042244

Precision =0.4

F1 score is 2*precision*sensitivity/(sensitivity+precision) =2*0.4*0.4/(0.8) =0.4

The plot of the ROC curve is shown below



Evaluation of Multiple Linear Regression Model

After considering the various Linear regression models, I decided to go with model 5. It has the highest F-statistic. F-statistics we can reject null hypothesis and say that overall addition of variables is significantly improving the model.

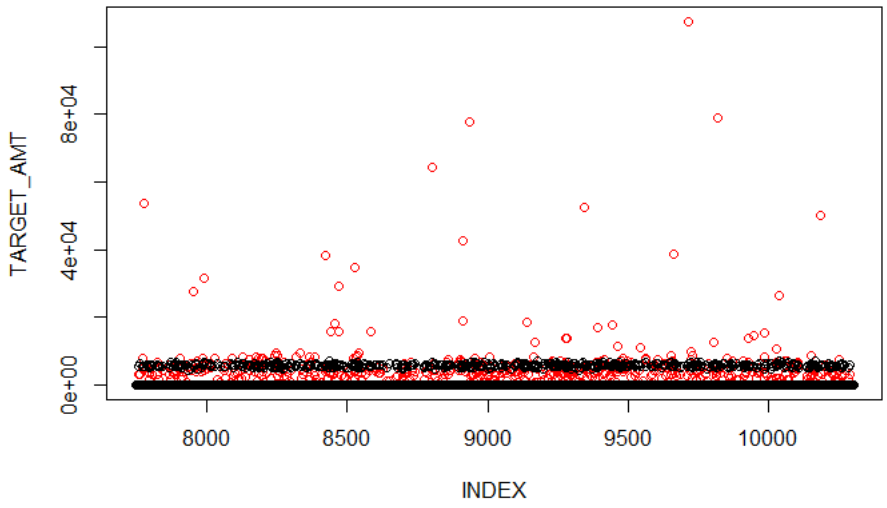
We got an Adjusted R-squared of 0.3087

```
Residual standard error: 3763 on 6058 degrees of freedom
Multiple R-squared:  0.3157,   Adjusted R-squared:  0.3087
F-statistic: 45.08 on 62 and 6058 DF,  p-value: < 2.2e-16
```

After splitting my dataset into test set. I ran it with the model and could get a

RMSE of 4537.195

A plot below shows the difference between the Observed Car crash cost (in Red) and predicted car crash cost



The model was evaluated with evalutation data set attached.

APPENDIX

```
data <- read.csv("https://raw.githubusercontent.com/nobieyi00/CUNY_DATA621/master/insurance_training_data.csv")
#summary(data)
train <- data
#train <- data[1:6121,]
#test <- data[6122:8161,]
if(!(c("psych") %in% rownames(installed.packages()))) {install.packages('psych')}
library(psych)

psych::describe(train)
summary(train)
```

```
train$INCOME <- as.numeric(gsub('[$,]', '', train$INCOME))
train$HOME_VAL <- as.numeric(gsub('[$,]', '', train$HOME_VAL))
train$BLUEBOOK <- as.numeric(gsub('[$,]', '', train$BLUEBOOK))
train$OLDCLAIM <- as.numeric(gsub('[$,]', '', train$OLDCLAIM))
library(Hmisc)



# impute with mean value
train$AGE <- round(with(train ,Hmisc::impute(AGE, mean)), digits = 0)

train$YOJ[is.na(train$YOJ)==TRUE] <- 0
hist(train$CAR_AGE)
train$CAR_AGE[(train$CAR_AGE==-3)==TRUE & is.na(train$CAR_AGE==-3)==FALSE ] <- 3
train$CAR_AGE <- round(with(train ,Hmisc::impute(CAR_AGE, median)), digits = 0)
aggregate(train$INCOME, list(train$JOB), mean,na.rm=TRUE)

train$INCOME[is.na(train$INCOME)==TRUE & train$JOB ==''] <- 118852.938
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Clerical'] <-33861.186
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Doctor'] <-128679.697
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Home Maker']<-12073.336
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Lawyer']<-88304.800
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Manager']<-87461.555
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Professional']<-76593.096
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='Student']<-6309.657
train$INCOME[is.na(train$INCOME)==TRUE & train$JOB =='z_Blue Collar'] <-58957.012
train$HOME_VAL[is.na(train$HOME_VAL)==TRUE] <- 0
summary(train)

psych::describe(train)
layout(matrix(c(1,2,3,4,5,6,7,8,9,10),2,5))
hist(train$KIDSDRIV )#highly skew
hist(train$AGE)#skew to the right
hist(train$YOJ)#highly skewed
hist(train$TRAVTIME)
hist(train$TIF)
hist(train$CLM_FREQ)#skewed
```

```
hist(train$MVR_PTS)#skewed
hist(train$CAR_AGE)#skewed
hist(train$HOMEKIDS )
hist(train$INCOME)
layout(matrix(c(1,2,3,4,5,6,7,8),2,4))
hist(log(train$KIDSDRIV) )#highly skew
hist(train$AGE)
hist(train$YOJ)
hist(log(train$TRAVTIME))
hist(log(train$TIF))
hist(log(train$CLM_FREQ))#skewed
hist(log(train$MVR_PTS))#skewed
hist(log(train$CAR_AGE))#skewed
library(ggplot2)
#install.packages('gridExtra')
library(gridExtra)
# Basic box plot


layout(matrix(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24),4,6) )
plot(train$KIDSDRIV,jitter(train$TARGET_FLAG ))
plot(train$AGE,jitter(train$TARGET_FLAG ))
plot(train$HOMEKIDS,jitter(train$TARGET_FLAG ))
plot(train$YOJ ,jitter(train$TARGET_FLAG ))
plot(train$INCOME,jitter(train$TARGET_FLAG ))
plot(train$PARENT1 ,jitter(train$TARGET_FLAG ))
plot(train$HOME_VAL,jitter(train$TARGET_FLAG ))
plot(train$MSTATUS ,jitter(train$TARGET_FLAG ))
plot(train$SEX ,jitter(train$TARGET_FLAG ))
plot(train$EDUCATION  ,jitter(train$TARGET_FLAG ))
plot(train$SEX ,jitter(train$TARGET_FLAG ))
plot(train$JOB  ,jitter(train$TARGET_FLAG ))
plot(train$TRAVTIME ,jitter(train$TARGET_FLAG ))
plot(train$CAR_USE ,jitter(train$TARGET_FLAG ))
plot(train$BLUEBOOK ,jitter(train$TARGET_FLAG ))
plot(train$TIF ,jitter(train$TARGET_FLAG ))
plot(train$CAR_TYPE ,jitter(train$TARGET_FLAG ))
plot(train$RED_CAR ,jitter(train$TARGET_FLAG ))
```

```
plot(train$OLDCLAIM ,jitter(train$TARGET_FLAG ))
plot(train$CLM_FREQ ,jitter(train$TARGET_FLAG ))
plot(train$REVOKED,jitter(train$TARGET_FLAG ))
plot(train$MVR ,jitter(train$TARGET_FLAG ))
plot(train$CAR_AGE ,jitter(train$TARGET_FLAG ))
plot(train$URBANICITY ,jitter(train$TARGET_FLAG ))


a<- table("Target"=train$TARGET_FLAG, "Urbanicity"=train$URBANICITY)
a[c(1,2),1]=a[c(1,2),1]/sum(a[c(1,2),1])
a[c(1,2),2]=a[c(1,2),2]/sum(a[c(1,2),2])
a



 ggplot(train, aes(x=TARGET_FLAG, y= AGE )) +
  geom_boxplot(aes(group = cut_width(TARGET_FLAG, 0.5)))

 ggplot(train, aes(x=TARGET_FLAG, y= HOMEKIDS )) +
  geom_boxplot(aes(group = cut_width(TARGET_FLAG, 0.5)))

ggplot(train, aes(x=TARGET_FLAG, y= YOJ )) +
  geom_boxplot(aes(group = cut_width(TARGET_FLAG, 0.5)))
plot(train$KIDSDRIV, train$TARGET_AMT)

plot(train$AGE, train$TARGET_AMT)
plot(train$YOJ, train$TARGET_AMT)


plot(train$INCOME, train$TARGET_AMT)
plot(train$PARENT1, train$TARGET_AMT)
plot(train$HOME_VAL, train$TARGET_AMT)
plot(train$MSTATUS, train$TARGET_AMT)
plot(train$EDUCATION, train$TARGET_AMT)
plot(train$SEX, train$TARGET_AMT)
plot(train$JOB, train$TARGET_AMT)
plot(train$TRAVTIME, train$TARGET_AMT)
plot(train$CAR_USE, train$TARGET_AMT)
```

```
plot(train$BLUEBOOK, train$TARGET_AMT)

plot(train$TIF, train$TARGET_AMT)


plot(train$CAR_TYPE, train$TARGET_AMT)

plot(train$RED_CAR, train$TARGET_AMT)

plot(train$OLDCLAIM, train$TARGET_AMT)

plot(train$CLM_FREQ, train$TARGET_AMT)

plot(train$REVOKED, train$TARGET_AMT)

plot(train$MVR_PTS, train$TARGET_AMT)

plot(train$CAR_AGE, train$TARGET_AMT)

plot(train$URBANICITY, train$TARGET_AMT)

pairs(~~train$AGE+TIF, data= train)

pairs(~~train$YOJ+INCOME, data= train)

pairs(~~train$HOME_VAL+INCOME, data= train)

pairs(~~train$BLUEBOOK+INCOME, data= train)

pairs(~~train$BLUEBOOK+CAR_AGE, data= train)

pairs(~~train$OLDCLAIM+CLM_FREQ, data= train)

dataset1<- train

train <- dataset1[1:6121,]

test <- dataset1[6122:8161,]


model1<- glm(formula = TARGET_FLAG~  KIDSDRIV+AGE+INCOME+PARENT1+HOME_VAL+MSTATUS+TRAVTIME
+CAR_USE+TIF+CAR_TYPE+OLDCLAIM
+CLM_FREQ+REVOKED+MVR_PTS+BLUEBOOK+CAR_AGE+EDUCATION+JOB+YOJ+RED_CAR+ URBANICITY+SEX,data =
train,family = binomial(link = "logit")  )

summary(model1)

library(pscl)

pR2(model1)

model2<- glm(formula = TARGET_FLAG~  KIDSDRIV+log(AGE)+INCOME+PARENT1+HOME_VAL+MSTATUS+TRAVTIME
+CAR_USE+TIF+CAR_TYPE+OLDCLAIM
+CLM_FREQ+REVOKED+MVR_PTS+BLUEBOOK+EDUCATION+log(CAR_AGE+1)+JOB+YOJ+RED_CAR+
URBANICITY+SEX,data = train,family = binomial(link = "logit")  )

summary(model2)


pR2(model2)

model3<- glm(formula = TARGET_FLAG~  KIDSDRIV+log(AGE)+INCOME+PARENT1+HOME_VAL+MSTATUS+TRAVTIME
+CAR_USE+TIF+CAR_TYPE+OLDCLAIM
+CLM_FREQ+REVOKED+MVR_PTS+BLUEBOOK+EDUCATION+log(CAR_AGE+1)+JOB+log(YOJ+1)+ URBANICITY+SEX,data
= train,family = binomial(link = "logit")  )

summary(model3)


pR2(model3)
```

```
library('rms')


library(car)

residualPlots(model3)

marginalModelPlots(model3)

influenceIndexPlot(model3)

plot(model3)

outlierTest(model3)

influencePlot(model3,col='red',id.n = 3)

model4<- lm(formula = TARGET_AMT ~
TARGET_FLAG+KIDSDRIV+log(AGE)+INCOME+PARENT1+log(HOME_VAL+1)+MSTATUS+TRAVTIME
+CAR_USE+TIF+CAR_TYPE+OLDCLAIM
+CLM_FREQ+REVOKED+MVR_PTS+BLUEBOOK+EDUCATION+CAR_AGE+JOB+YOJ+ URBANICITY+SEX,data = train )

summary(model4)

layout(matrix(c(1,2,3,4),2,2) )

plot(model4)

model5<- lm(formula = TARGET_AMT ~
TARGET_FLAG+KIDSDRIV+log(AGE)+INCOME+PARENT1+HOME_VAL+MSTATUS+TRAVTIME
+CAR_USE+TIF+CAR_TYPE+OLDCLAIM
+CLM_FREQ+REVOKED+MVR_PTS+BLUEBOOK+EDUCATION+log(CAR_AGE+1)+JOB+YOJ+RED_CAR+
URBANICITY+SEX,data = train )

summary(model5)

plot(model5)

model5<- lm(formula = TARGET_AMT ~TARGET_FLAG+INCOME+MVR_PTS+CLM_FREQ+OLDCLAIM+CLM_FREQ*OLDCLAIM
+REVOKED+CAR_USE+TIF+BLUEBOOK*CAR_AGE+CAR_AGE

      +AGE*PARENT1
+KIDSDRIV+TRAVTIME+HOME_VAL*INCOME+EDUCATION*JOB+MSTATUS+CAR_TYPE*CAR_USE+JOB+YOJ+RED_CAR+
URBANICITY+SEX,data = train )

summary(model5)

layout(matrix(c(1,2,3,4),2,2) )

plot(model5)

suppressMessages(library(dplyr))

suppressMessages(library(ggplot2))

suppressMessages(library(caret))

suppressMessages(library(e1071))

suppressMessages(library(pROC))


fitted.results <- predict(model3,newdata=test,type='response')

fitted.results <- ifelse(fitted.results > 0.5,1,0)

test['predicted'] <- fitted.results

confusionMatrix(test$predicted, test$TARGET_FLAG, positive = "1")
```

```
#install.packages('ROCR')
library(ROCR)
p <- predict(model3, newdata=test, type="response")
pr <- prediction(p, test$TARGET_FLAG)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
test['fitted.results_lm'] <-ifelse(test$TARGET_FLAG  == 0,0,predict(model5,test))
test[,c('INDEX','fitted.results_lm',"TARGET_AMT",'TARGET_FLAG')]
#plot(test[,c('INDEX','fitted.results_lm',"TARGET_AMT")])
plot(test$INDEX, test$TARGET_AMT)
plot(test$INDEX, test$fitted.results_lm)
rmse <- function(error)
{
   sqrt(mean(error^2))
}
error <- test$TARGET_AMT - test$fitted.results_lm
rmse(error)

with(test,plot(INDEX,TARGET_AMT,col='red'))
with(test,points(INDEX,fitted.results_lm))
summary(test)
evaluation <- read.csv('https://raw.githubusercontent.com/nobieyi00/CUNY_DATA621/master/insurance-evaluation-data.csv')
summary(evaluation)
evaluation$INCOME <- as.numeric(gsub('[$,]', '', evaluation$INCOME))
evaluation$HOME_VAL <- as.numeric(gsub('[$,]', '', evaluation$HOME_VAL))
evaluation$BLUEBOOK <- as.numeric(gsub('[$,]', '', evaluation$BLUEBOOK))
evaluation$OLDCLAIM <- as.numeric(gsub('[$,]', '', evaluation$OLDCLAIM))


evaluation$AGE[is.na(evaluation$AGE)==TRUE] <-round( mean(evaluation$AGE,na.rm=TRUE), digits = 0)
evaluation$YOJ[is.na(evaluation$YOJ)==TRUE] <- 0
evaluation$CAR_AGE[is.na(evaluation$CAR_AGE)==TRUE] <- round(median(evaluation$CAR_AGE,na.rm = TRUE), digits = 0)
aggregate(evaluation$INCOME, list(evaluation$JOB), mean,na.rm=TRUE)
```

```
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB ==''] <- 110314.30
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Clerical'] <-32906.17
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Doctor'] <-123812.38
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Home Maker']<-13415.93
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Lawyer']<-88473.91
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Manager']<-90295.68
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Professional']<-73548.99
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='Student']<-6086.00
evaluation$INCOME[is.na(evaluation$INCOME)==TRUE & evaluation$JOB =='z_Blue Collar'] <-58639.85
evaluation$HOME_VAL[is.na(evaluation$HOME_VAL)==TRUE] <- 0



fitted.results_ev <- predict(model3,newdata=evaluation,type='response')
fitted.results_ev <- ifelse(fitted.results_ev > 0.5,1,0)
evaluation['TARGET_FLAG'] <- fitted.results_ev

#fit Linear model

evaluation['TARGET_AMT'] <-ifelse(evaluation$TARGET_FLAG   == 0,0,predict(model5,evaluation))

head(evaluation)

write.csv(evaluation, file = "C:/Users/Mezu/Documents/Data621/insurance_eval_results.csv")
```