Homework 4

Justin Hsi Team 1

November 22, 2020

1. Data Exploration

The auto insurance training dataset has 26 variables and 8161 observations. Of the variables, 24 of them are predictors for two responses: TARGET_FLAG and TARGET_AMT is numerical.

To explore the training data: - used the summary function to see means, medians, and quartiles of predictors - used str function to see the data type of each predictor - explored TARGET_FLAG in relation to some other variables such as AGE and CAR_AGE - looked at distribution of some numerical variables such as AGE and MVR_PTS

From the summary function, the TARGET_FLAG is binary and 26% of the 8161 records were accidents.

2. Data Preparation

This data was prepared to build both a binary logistic model and a multiple linear regression model. The binary logisitc model was used to predict the TARGET_FLAG response variable and the multiple linear regression model was used to predict the TARGET_AMT variable.

We want to train the multiple linear regression model on records that actually have a valid TARGET_AMT variable, so its training dataset is a subset of the full dataset where TARGET_FLAG is 1.

We cleaned up INCOME, HOME_VAL, BLUEBOOK, and OLDCLAIM to be numerics instead of factors by stripping out dollar signs and commas.

We made dummy variable columns for all variables that had NA (AGE, YOJ, CAR_AGE) and then filled those columns with their median values.

The training dataset for the binary logistic regression model was labeled train_df. The training dataset for the multiple linear regression model was titled train_amt_df.

3. Build Models

First, we built two models using most predictors as numerics. Then we used the step AIC function to find the best variables for each model.

One model was a Binary Logistic Regression model for the TARGET_FLAG response titled step_BLR. The second model was a Multiple Linear Regression for the TARGET_AMT response titled MLR_all_vars.

4. Select Models

To finally select a model, we used Stepwise AIC (both backward and forward) to do model selection and ended with a Binary Logistic Regression AIC of 8718.2 and a Multiple Linear Regression Multiple R-squared of 0.003804.

Appendix

Import Libraries and Data

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## corrplot 0.84 loaded

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: lattice

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
# Loading the data
git_dir <- 'https://raw.githubusercontent.com/odonnell31/DATA621-HW4/main/data'
#class_data = read.csv(paste(git_dir, "/classification-output-data.csv", sep=""))
train_df = read.csv(paste(git_dir, "/insurance_training_data.csv", sep=""))
test_df = read.csv(paste(git_dir, "/insurance-evaluation-data.csv", sep = ""))
head(train_df, 2)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ  INCOME PARENT1
## 1     1           0          0        0  60        0  11 $67,349      No
## 2     2           0          0        0  43        0  11 $91,449      No
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
##   TIF CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11  Minivan     yes   $4,461        2      No       3      18
## 2   1  Minivan     yes       $0        0      No       0       1
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
```

Data Exploration & Preparation

See a summary of each column in the train_df set

```r
# view a summary of all columns
summary(train_df)
```

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME      PARENT1
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0      : 615   No :7084
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445   Yes:1077
##  Median :45.00   Median :0.0000   Median :11.0   $26,840 :   4
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   $48,509 :   4
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   $107,375:   3
##  NA's   :6                        NA's   :454    (Other) :7086
##     HOME_VAL    MSTATUS      SEX              EDUCATION
##  $0      :2294   Yes :4894   M  :3786   <High School :1203
##          : 464   z_No:3267   z_F:4375   Bachelors    :2242
##  $111,129:   3                          Masters      :1658
##  $115,249:   3                          PhD          : 728
##  $123,109:   3                          z_High School:2330
##  $153,061:   3
##  (Other) :5391
##              JOB          TRAVTIME            CAR_USE        BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   $1,500 : 157
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   $6,000 :  34
##  Professional :1117   Median : 33.00                     $5,800 :  33
##  Manager      : 988   Mean   : 33.49                     $6,200 :  33
##  Lawyer       : 835   3rd Qu.: 44.00                     $6,400 :  31
##  Student      : 712   Max.   :142.00                     $5,900 :  30
##  (Other)      :1413                                      (Other):7843
##       TIF                 CAR_TYPE    RED_CAR       OLDCLAIM       CLM_FREQ
```

```
##  Min.   : 1.000   Minivan    :2145   no :5783   $0     :5009   Min.   :0.0000
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :    4   1st Qu.:0.0000
##  Median : 4.000   Pickup     :1389              $1,391 :    4   Median :0.0000
##  Mean   : 5.351   Sports Car : 907              $4,263 :    4   Mean   :0.7986
##  3rd Qu.: 7.000   Van        : 750              $1,105 :    3   3rd Qu.:2.0000
##  Max.   :25.000   z_SUV      :2294              $1,332 :    3   Max.   :5.0000
##                                                 (Other):3134
##  REVOKED       MVR_PTS          CAR_AGE                     URBANICITY
##  No :7161   Min.   : 0.000   Min.   :-3.000   Highly Urban/ Urban  :6492
##  Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000   z_Highly Rural/ Rural:1669
##              Median : 1.000   Median : 8.000
##              Mean   : 1.696   Mean   : 8.328
##              3rd Qu.: 3.000   3rd Qu.:12.000
##              Max.   :13.000   Max.   :28.000
##                               NA's   :510
```

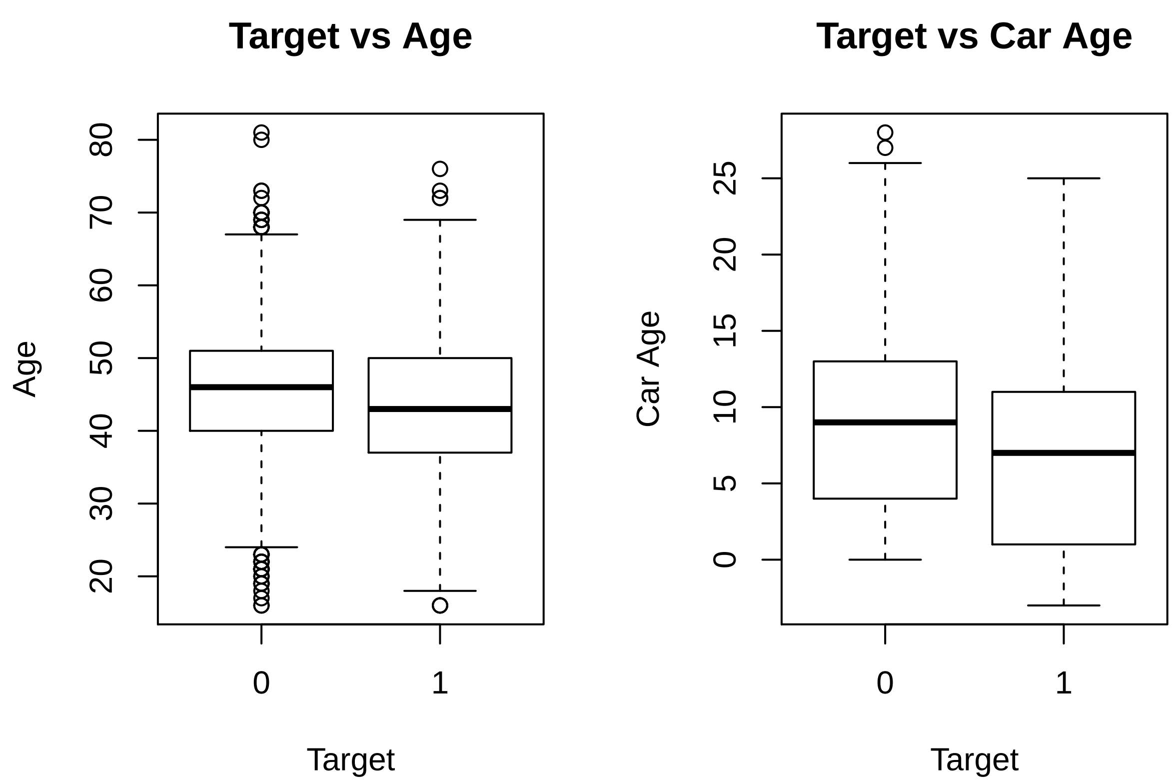
Look at the data type of each variable

```r
# data type of predictors
str(train_df)
```

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : Factor w/ 6613 levels "","$0","$1,007",..: 5033 6292 1250 1 509 746 1488 315 4765 28
##  $ PARENT1    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
##  $ HOME_VAL   : Factor w/ 5107 levels "","$0","$100,093",..: 2 3259 348 3917 3034 2 1 4167 2 2 ...
##  $ MSTATUS    : Factor w/ 2 levels "Yes","z_No": 2 2 1 1 1 2 1 1 2 2 ...
##  $ SEX        : Factor w/ 2 levels "M","z_F": 1 1 2 1 2 2 2 1 2 1 ...
##  $ EDUCATION  : Factor w/ 5 levels "<High School",..: 4 5 5 1 4 2 1 2 2 2 ...
##  $ JOB        : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : Factor w/ 2 levels "Commercial","Private": 2 1 2 2 2 1 2 1 2 1 ...
##  $ BLUEBOOK   : Factor w/ 2789 levels "$1,500","$1,520",..: 434 503 2212 553 802 746 2672 701 135 85
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...
##  $ RED_CAR    : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
##  $ OLDCLAIM   : Factor w/ 2857 levels "$0","$1,000",..: 1449 1 1311 1 432 1 1 510 1 1 ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 2 ...
```

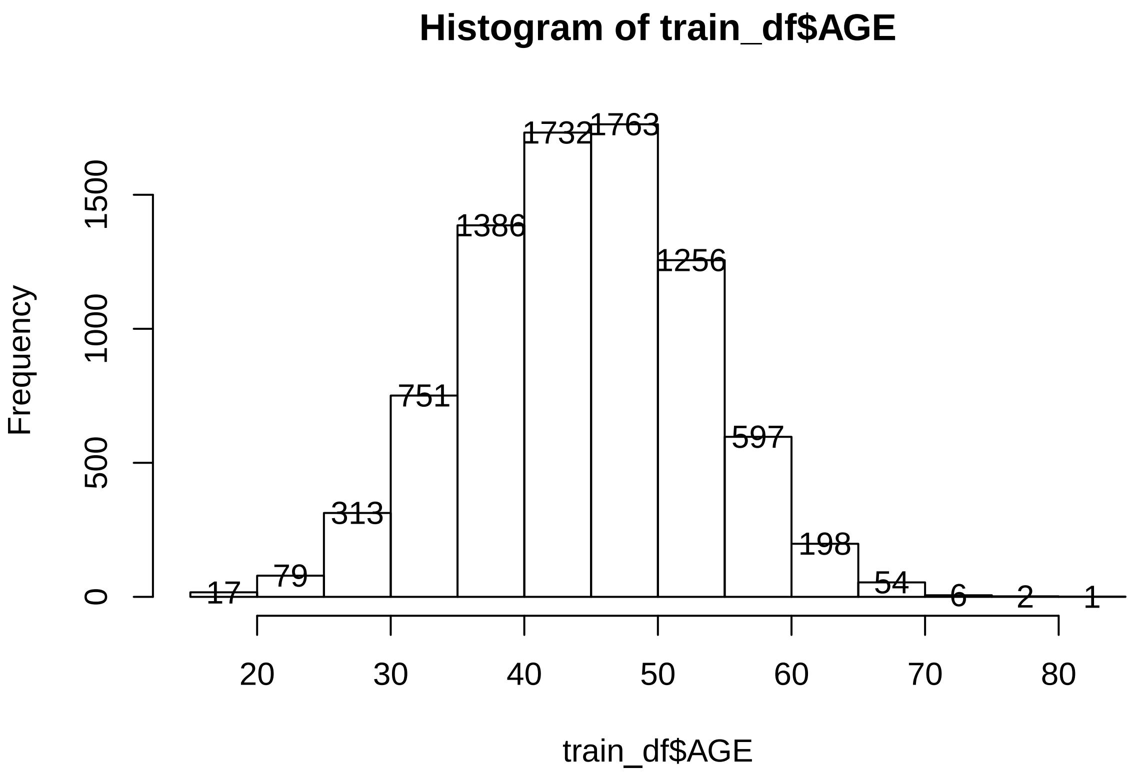
Look at the relationship between TARGET_FLAG and some of the numerical variables.

```r
par(mfrow=c(1,2))
# plot response variable "target" against predictor variable "age" and "car_age"
boxplot(AGE ~ TARGET_FLAG, train_df,
        main="Target vs Age",
        xlab="Target",
        ylab="Age")
boxplot(CAR_AGE ~ TARGET_FLAG, train_df,
        main="Target vs Car Age",
        xlab="Target",
        ylab="Car Age")
```

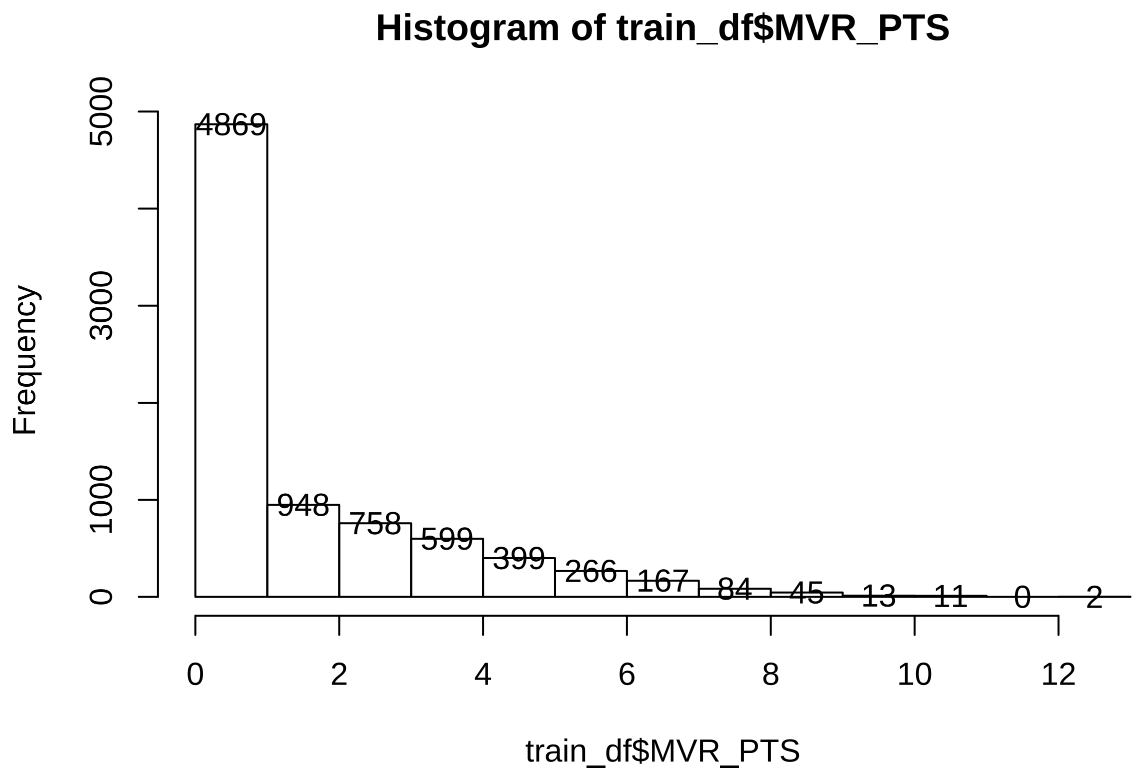


Look at the distribution of some numerical variables.

```r
h <- hist(train_df$AGE)
text(h$mids,h$counts,labels=h$counts)
```




```
h <- hist(train_df$MVR_PTS)
text(h$mids,h$counts,labels=h$counts)
```



Cleanup INCOME, HOME_VAL, BLUEBOOK, and OLDCLAIM to be numerics by stripping out dollar signs and commas.

```
numeric = function(input) {
  out = sub("\\$", "", input)
  out = as.numeric(sub(",", "", out))
  return(out)
}


train_df = as.tbl(train_df) %>%
  mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),
            numeric)

test_df= as.tbl(test_df) %>%
  mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),
            numeric)
```

Check for NA's

```
has_NA = names(which(sapply(train_df, anyNA)))
has_NA
```

```
## [1] "AGE"      "YOJ"      "INCOME"   "HOME_VAL" "CAR_AGE"
```

Check test_df for NA's

```
has_NA_test = names(which(sapply(test_df, anyNA)))
has_NA_test
```

```
## [1] "TARGET_FLAG" "TARGET_AMT"  "AGE"         "YOJ"         "INCOME"
## [6] "HOME_VAL"    "CAR_AGE"
```

Since we see our test_df has NAs for the same variables as test, we need to come up with a way to handle making predictions on records that have these values as NA. We will create an "_NA" columns as dummy variables for AGE, YOJ, and CAR_AGE, 1 marking them as NA and 0 if they have a value.

```
for (col in has_NA)
{
   new_col = (paste(col,"_NA", sep=""))
   train_df[,new_col] = as.numeric(is.na(train_df[,col]))
   test_df[,new_col] = as.numeric(is.na(test_df[,col]))
   # fill missing numerics with median value
   train_df[,col][is.na(train_df[,col])] = median(unlist(train_df[,col]), na.rm=TRUE)
   test_df[,col][is.na(test_df[,col])] = median(unlist(test_df[,col]), na.rm=TRUE)
}
```

Create train_amt_df dataframe for multiple linear regression model

```
train_amt_df <- subset(train_df, TARGET_AMT > 0)
summary(train_amt_df$TARGET_FLAG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Modeling

1) Binary Logistic Regression

```
# preliminary exploration with one predictor
model1 <- glm(formula = TARGET_FLAG ~ AGE, family = binomial(), data = train_df)
summary(model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0728  -0.8042  -0.7403   1.4313   2.0168
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.186818   0.131990   1.415    0.157
## AGE         -0.027373   0.002954  -9.265   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 9330.8  on 8159  degrees of freedom
## AIC: 9334.8
##
## Number of Fisher Scoring iterations: 4
```

Binary Logistic Regression Model with more variables

```r
BLR_all_vars = glm(TARGET_FLAG ~ AGE +
                   CAR_AGE +
                   MVR_PTS +
                   YOJ +
                   CLM_FREQ +
                   TIF, family = binomial(), data = train_df)
summary(BLR_all_vars)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8021  -0.7630  -0.6108   0.9899   2.4099
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.095985   0.153832   0.624    0.533
## AGE         -0.019810   0.003107  -6.376 1.82e-10 ***
## CAR_AGE     -0.035902   0.004949  -7.254 4.04e-13 ***
## MVR_PTS      0.147989   0.012363  11.971  < 2e-16 ***
## YOJ         -0.025942   0.006464  -4.013 5.99e-05 ***
## CLM_FREQ     0.293062   0.022906  12.794  < 2e-16 ***
## TIF         -0.045555   0.006689  -6.811 9.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 8704.2  on 8154  degrees of freedom
## AIC: 8718.2
##
## Number of Fisher Scoring iterations: 4
```

Step through AIC scores to find best model

```r
step_BLR = stepAIC(BLR_all_vars)
```

```
## Start:  AIC=8718.2
## TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ + TIF
##
##            Df Deviance    AIC
## <none>          8704.2 8718.2
## - YOJ       1   8720.1 8732.1
## - AGE       1   8745.2 8757.2
## - TIF       1   8752.2 8764.2
## - CAR_AGE   1   8757.7 8769.7
## - MVR_PTS   1   8847.9 8859.9
## - CLM_FREQ  1   8864.3 8876.3
```

```r
summary(step_BLR)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8021  -0.7630  -0.6108   0.9899   2.4099
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.095985   0.153832   0.624    0.533
## AGE         -0.019810   0.003107  -6.376 1.82e-10 ***
## CAR_AGE     -0.035902   0.004949  -7.254 4.04e-13 ***
## MVR_PTS      0.147989   0.012363  11.971  < 2e-16 ***
## YOJ         -0.025942   0.006464  -4.013 5.99e-05 ***
## CLM_FREQ     0.293062   0.022906  12.794  < 2e-16 ***
## TIF         -0.045555   0.006689  -6.811 9.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 8704.2  on 8154  degrees of freedom
## AIC: 8718.2
##
## Number of Fisher Scoring iterations: 4
```

2) Multiple Linear Regression

Multiple Linear Regression models with many variables

```r
MLR_all_vars = lm(TARGET_AMT ~ AGE +
                  CAR_AGE +
```

```r
                    MVR_PTS +
                    YOJ +
                    CLM_FREQ +
                    TIF, data = train_amt_df)
summary(MLR_all_vars)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, data = train_amt_df)
##
## Residuals:
##    Min      1Q Median     3Q    Max
##  -6311  -3111  -1579    160 101042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4117.71     892.22   4.615 4.16e-06 ***
## AGE            22.58      17.80   1.268   0.2048
## CAR_AGE       -23.46      31.64  -0.741   0.4586
## MVR_PTS       132.33      68.03   1.945   0.0519 .
## YOJ            56.31      38.36   1.468   0.1423
## CLM_FREQ      -64.15     140.39  -0.457   0.6478
## TIF            -7.77      42.47  -0.183   0.8549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7739 on 2146 degrees of freedom
## Multiple R-squared:  0.003804,   Adjusted R-squared:  0.001019
## F-statistic: 1.366 on 6 and 2146 DF,  p-value: 0.2248
```

Predictions on Evaluation Set

```r
# step_BLR prediction on test
test_preds_BLR = round(predict(step_BLR, newdata=test_df, type='response'))
test_df$TARGET_FLAG = test_preds_BLR
test_preds_MLR = predict(MLR_all_vars, newdata=test_df)
test_df$TARGET_AMT = test_preds_MLR

# write out evaluation data with predictions
write.csv(test_df, 'eval_with_preds.csv')
```