CUNY DATA 621 Homework 4

Raphael Nash

Exploration

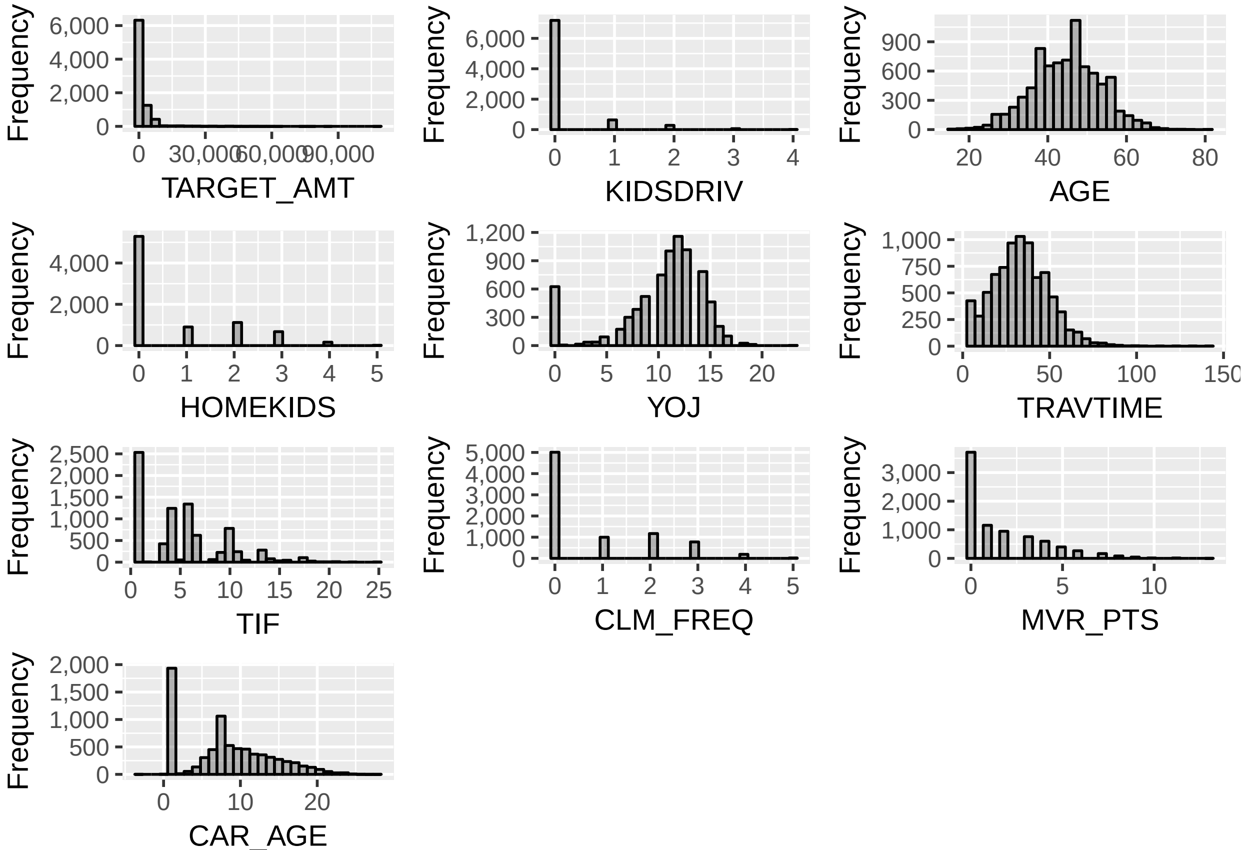
```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14               No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE
## 1       $0    z_No   M           PhD  Professional       14    Private
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial
##   BLUEBOOK TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS
## 1  $14,230  11    Minivan     yes   $4,461        2      No       3
## 2  $14,940   1    Minivan     yes       $0        0      No       0
## 3   $4,010   4      z_SUV      no  $38,690        2      No       3
## 4  $15,440   7    Minivan     yes       $0        0      No       0
## 5  $18,000   1      z_SUV      no  $19,217        2     Yes       3
## 6  $17,430   1 Sports Car      no       $0        0      No       0
##   CAR_AGE          URBANICITY
## 1      18 Highly Urban/ Urban
## 2       1 Highly Urban/ Urban
## 3      10 Highly Urban/ Urban
## 4       6 Highly Urban/ Urban
## 5      17 Highly Urban/ Urban
## 6       7 Highly Urban/ Urban
```

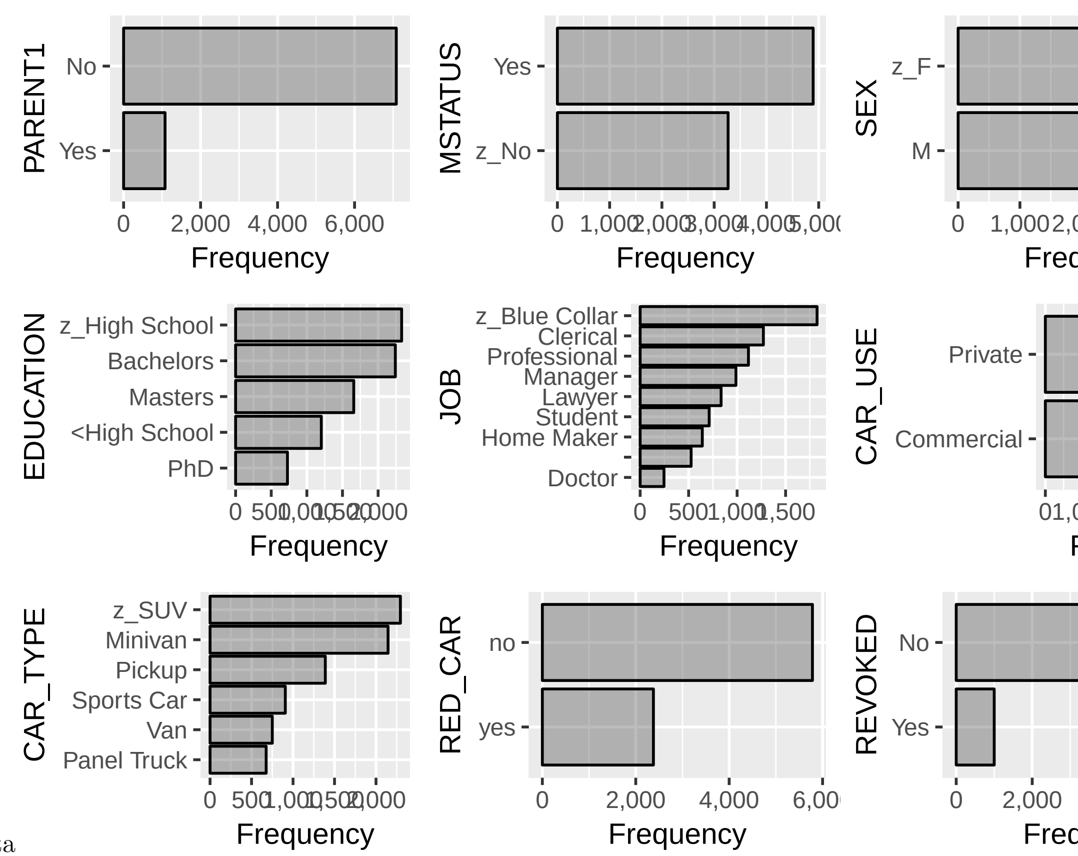
```
summary(training_df )
```

```
##      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
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0      : 615
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445
##  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
```

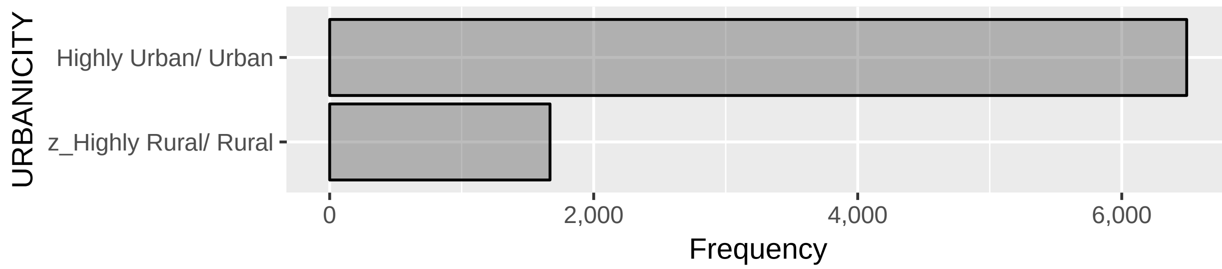
```
## NA's   :6                       NA's   :454    (Other) :7086
## PARENT1        HOME_VAL    MSTATUS      SEX             EDUCATION
## No :7084   $0      :2294   Yes :4894   M  :3786   <High School :1203
## Yes:1077           : 464   z_No:3267   z_F:4375   Bachelors    :2242
##            $111,129:   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
## Min.   : 1.000   Minivan    :2145   no :5783   $0      :5009
## 1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4
## Median : 4.000   Pickup     :1389              $1,391 :   4
## Mean   : 5.351   Sports Car : 907              $4,263 :   4
## 3rd Qu.: 7.000   Van        : 750              $1,105 :   3
## Max.   :25.000   z_SUV      :2294              $1,332 :   3
##                                                (Other):3134
##    CLM_FREQ      REVOKED       MVR_PTS         CAR_AGE
## Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
## 1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
## Median :0.0000              Median : 1.000   Median : 8.000
## Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
## 3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
## Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                              NA's   :510
##                 URBANICITY
## Highly Urban/ Urban  :6492
## z_Highly Rural/ Rural:1669
##
##
##
##
##
```

Plot histograms for continous data





Create Bar Plots for descrete data



Transformation

1) Car age should not be less than 0, so if it is 0 then make it 0
2) Make missing values for Job "Unknown"
3) Convert Currency to numberic for Income and Home Value

3)Fill missing values with median

Builid models

Logistical Regression.

Build Logistical Regression model to predict if person has a claim (TARGET_FLAG). Of course inorder to do theis we will have to remove TARGET_AMT and INDEX from the dataframe. INDEX truely has not value, and you only have a claim amoiunt if you have a claim. I will use backward selection to select the best logistical model.
Model 1

Model with all variables:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = trainng_logit_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5849  -0.7127  -0.3982   0.6265   3.1525
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.170e-01  2.715e-01  -1.904 0.056855 .
## KIDSDRIV                        3.862e-01  6.122e-02   6.308 2.82e-10 ***
## AGE                            -1.013e-03  4.020e-03  -0.252 0.800987
## HOMEKIDS                        4.965e-02  3.713e-02   1.337 0.181156
## YOJ                            -1.106e-02  8.582e-03  -1.288 0.197582
## INCOME                         -3.422e-06  1.082e-06  -3.164 0.001559 **
## PARENT1Yes                      3.820e-01  1.096e-01   3.485 0.000492 ***
## HOME_VAL                       -1.307e-06  3.420e-07  -3.821 0.000133 ***
## MSTATUSz_No                     4.938e-01  8.358e-02   5.908 3.46e-09 ***
## SEXz_F                         -8.247e-02  1.120e-01  -0.736 0.461666
## EDUCATIONBachelors             -3.799e-01  1.156e-01  -3.285 0.001020 **
## EDUCATIONMasters               -2.877e-01  1.788e-01  -1.609 0.107559
## EDUCATIONPhD                   -1.651e-01  2.139e-01  -0.772 0.440322
```

```
## EDUCATIONz_High School           1.790e-02  9.505e-02   0.188 0.850671
## JOBDoctor                       -8.565e-01  2.863e-01  -2.991 0.002780 **
## JOBHome Maker                   -1.783e-01  1.449e-01  -1.231 0.218394
## JOBLawyer                       -3.058e-01  1.856e-01  -1.648 0.099365 .
## JOBManager                      -9.680e-01  1.439e-01  -6.726 1.75e-11 ***
## JOBProfessional                 -2.488e-01  1.245e-01  -1.998 0.045673 *
## JOBStudent                      -1.946e-01  1.315e-01  -1.480 0.138853
## JOBUnknown                      -4.108e-01  1.967e-01  -2.089 0.036724 *
## JOBz_Blue Collar                -1.001e-01  1.067e-01  -0.938 0.348139
## TRAVTIME                         1.457e-02  1.883e-03   7.736 1.03e-14 ***
## CAR_USEPrivate                  -7.564e-01  9.172e-02  -8.247  < 2e-16 ***
## BLUEBOOK                        -2.084e-05  5.263e-06  -3.960 7.51e-05 ***
## TIF                             -5.546e-02  7.344e-03  -7.553 4.27e-14 ***
## CAR_TYPEPanel Truck              5.607e-01  1.618e-01   3.466 0.000529 ***
## CAR_TYPEPickup                   5.540e-01  1.007e-01   5.500 3.80e-08 ***
## CAR_TYPESports Car               1.025e+00  1.299e-01   7.892 2.97e-15 ***
## CAR_TYPEVan                      6.185e-01  1.265e-01   4.891 1.01e-06 ***
## CAR_TYPEz_SUV                    7.682e-01  1.113e-01   6.904 5.06e-12 ***
## RED_CARyes                      -9.702e-03  8.636e-02  -0.112 0.910553
## OLDCLAIM                        -1.389e-05  3.910e-06  -3.554 0.000380 ***
## CLM_FREQ                         1.960e-01  2.855e-02   6.865 6.67e-12 ***
## REVOKEDYes                       8.874e-01  9.133e-02   9.716  < 2e-16 ***
## MVR_PTS                          1.133e-01  1.361e-02   8.324  < 2e-16 ***
## CAR_AGE                         -9.825e-04  7.544e-03  -0.130 0.896378
## URBANICITYz_Highly Rural/ Rural -2.390e+00  1.128e-01 -21.181  < 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: 7297.6  on 8123  degrees of freedom
## AIC: 7373.6
##
## Number of Fisher Scoring iterations: 5
```

Model 2

To make Model 2 I will remove the non-signifigant variables from model 1. This Model will predict TARGET_FLAG based on INCOME, PARENT1, HOME_VAL, MSTATUS,EDUCATION,JOB, TRAVTIME ,CAR_USE, BLUEBOOK, TIF, CAR_TYPE, OLDCLAIM, CLM_FREQ, REVOKED_MVR_PTS, and URBANICITY. All

```
##
## Call:
## glm(formula = TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS +
##     EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY, family = "binomial",
##     data = trainng_logit_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2676  -0.7201  -0.4057   0.6488   3.1051
##
```

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -5.845e-01  1.896e-01  -3.084 0.002044 **
## INCOME                          -3.287e-06  1.071e-06  -3.070 0.002139 **
## PARENT1Yes                       6.332e-01  9.117e-02   6.945 3.78e-12 ***
## HOME_VAL                        -1.361e-06  3.397e-07  -4.007 6.16e-05 ***
## MSTATUSz_No                      3.796e-01  7.827e-02   4.850 1.23e-06 ***
## EDUCATIONBachelors              -3.981e-01  1.084e-01  -3.672 0.000240 ***
## EDUCATIONMasters                -2.939e-01  1.604e-01  -1.832 0.067019 .
## EDUCATIONPhD                    -1.904e-01  1.989e-01  -0.957 0.338342
## EDUCATIONz_High School           5.785e-03  9.422e-02   0.061 0.951045
## JOBDoctor                       -8.863e-01  2.842e-01  -3.119 0.001817 **
## JOBHome Maker                   -1.328e-01  1.351e-01  -0.983 0.325649
## JOBLawyer                       -3.432e-01  1.838e-01  -1.867 0.061944 .
## JOBManager                      -9.636e-01  1.422e-01  -6.775 1.25e-11 ***
## JOBProfessional                 -2.669e-01  1.235e-01  -2.162 0.030649 *
## JOBStudent                      -1.267e-01  1.248e-01  -1.015 0.310023
## JOBUnknown                      -4.350e-01  1.954e-01  -2.226 0.026011 *
## JOBz_Blue Collar                -8.427e-02  1.059e-01  -0.796 0.426175
## TRAVTIME                         1.434e-02  1.872e-03   7.660 1.86e-14 ***
## CAR_USEPrivate                  -7.365e-01  9.104e-02  -8.089 6.00e-16 ***
## BLUEBOOK                        -2.252e-05  4.690e-06  -4.801 1.58e-06 ***
## TIF                             -5.547e-02  7.326e-03  -7.571 3.70e-14 ***
## CAR_TYPEPanel Truck              6.165e-01  1.503e-01   4.101 4.12e-05 ***
## CAR_TYPEPickup                   5.478e-01  1.001e-01   5.472 4.45e-08 ***
## CAR_TYPESports Car               9.581e-01  1.070e-01   8.953  < 2e-16 ***
## CAR_TYPEVan                      6.350e-01  1.217e-01   5.219 1.80e-07 ***
## CAR_TYPEz_SUV                    7.198e-01  8.560e-02   8.409  < 2e-16 ***
## OLDCLAIM                        -1.474e-05  3.891e-06  -3.790 0.000151 ***
## CLM_FREQ                         2.016e-01  2.838e-02   7.104 1.21e-12 ***
## REVOKEDYes                       9.196e-01  9.056e-02  10.154  < 2e-16 ***
## MVR_PTS                          1.181e-01  1.354e-02   8.721  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.337e+00  1.115e-01 -20.963  < 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: 7358.7  on 8130  degrees of freedom
## AIC: 7420.7
##
## Number of Fisher Scoring iterations: 5
```

MODEL 3

In this iteration, I will remove all values from the model where all levels of that variable do not have a p value of <.001. This will get our model focusing on the mose signifigant values. This model will predict TARGET_FLAG based on HOME_VAL, MSTATUS, TRAVTIME, CAR_USE, BLUEBOOK, TIF, CAR_TYPE, OLDCLAIM, CLM_FREQ, REVOKED, MVR_PTS, URBANICITY

```
##
## Call:
## glm(formula = TARGET_FLAG ~ HOME_VAL + MSTATUS + TRAVTIME + CAR_USE +
```

```
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY, family = "binomial", data = trainng_logit_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3100  -0.7451  -0.4416   0.7307   3.0416
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.122e-01  1.414e-01  -4.331 1.49e-05 ***
## HOME_VAL                        -2.878e-06  2.789e-07 -10.317  < 2e-16 ***
## MSTATUSz_No                      3.414e-01  6.529e-02   5.229 1.70e-07 ***
## TRAVTIME                         1.379e-02  1.829e-03   7.537 4.79e-14 ***
## CAR_USEPrivate                  -8.929e-01  6.835e-02 -13.064  < 2e-16 ***
## BLUEBOOK                        -3.436e-05  4.466e-06  -7.694 1.42e-14 ***
## TIF                             -5.298e-02  7.166e-03  -7.393 1.43e-13 ***
## CAR_TYPEPanel Truck              4.391e-01  1.386e-01   3.167 0.001538 **
## CAR_TYPEPickup                   4.671e-01  9.567e-02   4.882 1.05e-06 ***
## CAR_TYPESports Car               9.475e-01  1.032e-01   9.178  < 2e-16 ***
## CAR_TYPEVan                      4.859e-01  1.177e-01   4.130 3.63e-05 ***
## CAR_TYPEz_SUV                    7.327e-01  8.281e-02   8.848  < 2e-16 ***
## OLDCLAIM                        -1.388e-05  3.793e-06  -3.658 0.000254 ***
## CLM_FREQ                         1.926e-01  2.773e-02   6.943 3.83e-12 ***
## REVOKEDYes                       9.162e-01  8.819e-02  10.389  < 2e-16 ***
## MVR_PTS                          1.283e-01  1.324e-02   9.695  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.057e+00  1.087e-01 -18.921  < 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: 7626.2  on 8144  degrees of freedom
## AIC: 7660.2
##
## Number of Fisher Scoring iterations: 5
```

The coefiencents of the model make a lok of sense. Travel Time, Claim Frequency , license points (MVR_PTS) all have positive coeficients and that makes sense to me that thoses variables would be positively correlated to if there is a Claim. It also makes sense that Bluebook would be negative related to if there is a Claim since people take better care of more expensive cars.
##Claim Amount Regression Model

The second model we will create will predict the claim amount, based on if the person had a claim.

Model 1

This model will contain all variables.

```
## 'data.frame':    2153 obs. of  24 variables:
##  $ TARGET_AMT: num  2946 4021 2501 6077 1267 ...
##  $ KIDSDRIV  : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ AGE       : int  34 37 34 53 53 45 28 43 32 40 ...
##  $ HOMEKIDS  : int  1 2 0 0 0 0 1 0 1 0 ...
##  $ YOJ       : int  12 11 10 14 11 0 13 13 9 11 ...
```

```
##  $ INCOME    : num  125301 107961 62978 77100 130795 ...
##  $ PARENT1   : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ HOME_VAL  : num  0 333680 0 0 0 ...
##  $ MSTATUS   : Factor w/ 2 levels "Yes","z_No": 2 1 2 2 2 1 1 1 1 2 ...
##  $ SEX       : Factor w/ 2 levels "M","z_F": 2 1 2 2 1 2 2 1 2 1 ...
##  $ EDUCATION : Factor w/ 5 levels "<High School",..: 2 2 2 3 4 1 5 1 5 1 ...
##  $ JOB       : Factor w/ 9 levels "Clerical","Doctor",..: 9 9 1 4 8 3 9 9 1 9 ...
##  $ TRAVTIME  : int  46 44 34 15 64 48 29 52 26 20 ...
##  $ CAR_USE   : Factor w/ 2 levels "Commercial","Private": 1 1 2 2 1 2 1 1 2 1 ...
##  $ BLUEBOOK  : num  17430 16970 11200 18300 28340 ...
##  $ TIF       : int  1 1 1 1 6 1 6 1 1 4 ...
##  $ CAR_TYPE  : Factor w/ 6 levels "Minivan","Panel Truck",..: 4 5 6 4 2 6 6 2 6 3 ...
##  $ RED_CAR   : Factor w/ 2 levels "no","yes": 1 2 1 1 2 1 1 2 1 2 ...
##  $ OLDCLAIM  : num  0 2374 0 0 0 ...
##  $ CLM_FREQ  : int  0 1 0 0 0 0 2 0 0 1 ...
##  $ REVOKED   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
##  $ MVR_PTS   : int  0 10 0 0 3 3 0 3 0 13 ...
##  $ CAR_AGE   : num  7 7 1 11 10 5 1 1 1 6 ...
##  $ URBANICITY: Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 1 ...

##
## Call:
## lm(formula = TARGET_AMT ~ ., data = regression_training_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8947  -3174  -1502    482  99585
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.707e+03  1.490e+03   2.489   0.0129 *
## KIDSDRIV                        -1.714e+02  3.166e+02  -0.541   0.5884
## AGE                              1.832e+01  2.124e+01   0.863   0.3885
## HOMEKIDS                         2.133e+02  2.071e+02   1.030   0.3033
## YOJ                              1.916e+01  4.918e+01   0.390   0.6968
## INCOME                          -9.004e-03  6.742e-03  -1.335   0.1819
## PARENT1Yes                       2.784e+02  5.873e+02   0.474   0.6355
## HOME_VAL                         2.191e-03  2.020e-03   1.084   0.2783
## MSTATUSz_No                      8.015e+02  4.935e+02   1.624   0.1045
## SEXz_F                          -1.401e+03  6.564e+02  -2.135   0.0329 *
## EDUCATIONBachelors               2.528e+02  6.416e+02   0.394   0.6936
## EDUCATIONMasters                 1.181e+03  1.083e+03   1.090   0.2757
## EDUCATIONPhD                     2.383e+03  1.312e+03   1.817   0.0693 .
## EDUCATIONz_High School          -3.972e+02  5.145e+02  -0.772   0.4402
## JOBDoctor                       -2.424e+03  1.867e+03  -1.298   0.1944
## JOBHome Maker                   -3.317e+02  8.395e+02  -0.395   0.6928
## JOBLawyer                        1.787e+01  1.158e+03   0.015   0.9877
## JOBManager                      -1.088e+03  9.318e+02  -1.168   0.2430
## JOBProfessional                  7.496e+02  7.196e+02   1.042   0.2977
## JOBStudent                      -1.955e+02  7.303e+02  -0.268   0.7890
## JOBUnknown                      -3.097e+02  1.203e+03  -0.257   0.7968
## JOBz_Blue Collar                 2.106e+02  5.877e+02   0.358   0.7201
## TRAVTIME                         7.294e-01  1.108e+01   0.066   0.9475
## CAR_USEPrivate                  -4.407e+02  5.216e+02  -0.845   0.3983
```

```
## BLUEBOOK                          1.245e-01  3.053e-02   4.078 4.71e-05 ***
## TIF                              -1.572e+01  4.252e+01  -0.370   0.7116
## CAR_TYPEPanel Truck              -6.425e+02  9.605e+02  -0.669   0.5036
## CAR_TYPEPickup                   -5.951e+01  5.968e+02  -0.100   0.9206
## CAR_TYPESports Car                1.064e+03  7.502e+02   1.418   0.1564
## CAR_TYPEVan                       6.238e+01  7.708e+02   0.081   0.9355
## CAR_TYPEz_SUV                     9.048e+02  6.668e+02   1.357   0.1749
## RED_CARyes                       -1.931e+02  4.965e+02  -0.389   0.6974
## OLDCLAIM                          2.494e-02  2.263e-02   1.102   0.2707
## CLM_FREQ                         -1.159e+02  1.580e+02  -0.733   0.4635
## REVOKEDYes                       -1.126e+03  5.166e+02  -2.179   0.0295 *
## MVR_PTS                           1.111e+02  6.853e+01   1.621   0.1052
## CAR_AGE                          -9.716e+01  4.400e+01  -2.208   0.0273 *
## URBANICITYz_Highly Rural/ Rural -9.722e+01  7.562e+02  -0.129   0.8977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7690 on 2115 degrees of freedom
## Multiple R-squared:  0.03055,    Adjusted R-squared:  0.01359
## F-statistic: 1.802 on 37 and 2115 DF,  p-value: 0.002246
```

Model 2

From regression model I am am going to drop all but the statisticall signifigant variables. This model will predict target amount based on BLUEBOOK, REVOKED, CAR_AGE, EDUCATION

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + REVOKED + CAR_AGE + EDUCATION,
##     data = regression_training_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8166  -3085  -1567    340 100623
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4921.06078  496.72975   9.907  < 2e-16 ***
## BLUEBOOK                  0.10795    0.02081   5.189 2.32e-07 ***
## REVOKEDYes             -671.57579  409.95236  -1.638   0.1015
## CAR_AGE                -105.67758   43.60157  -2.424   0.0154 *
## EDUCATIONBachelors      403.78114  565.85320   0.714   0.4756
## EDUCATIONMasters        777.96331  735.57245   1.058   0.2903
## EDUCATIONPhD           1101.78005  925.90221   1.190   0.2342
## EDUCATIONz_High School -317.00067  479.88225  -0.661   0.5090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7684 on 2145 degrees of freedom
## Multiple R-squared:  0.01843,    Adjusted R-squared:  0.01523
## F-statistic: 5.754 on 7 and 2145 DF,  p-value: 1.284e-06
```

Model 3

For this model I will again drop all but the most signifigant variabels. That will mean this model will predict claim amout based on car age and blue book.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + CAR_AGE, data = regression_training_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8010  -3105  -1557    350 101207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4429.9152   375.7643  11.789  < 2e-16 ***
## BLUEBOOK       0.1162     0.0203   5.725 1.18e-08 ***
## CAR_AGE      -51.9374    31.5013  -1.649   0.0993 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7688 on 2150 degrees of freedom
## Multiple R-squared:  0.01519,    Adjusted R-squared:  0.01427
## F-statistic: 16.58 on 2 and 2150 DF,  p-value: 7.147e-08
```

Model 4

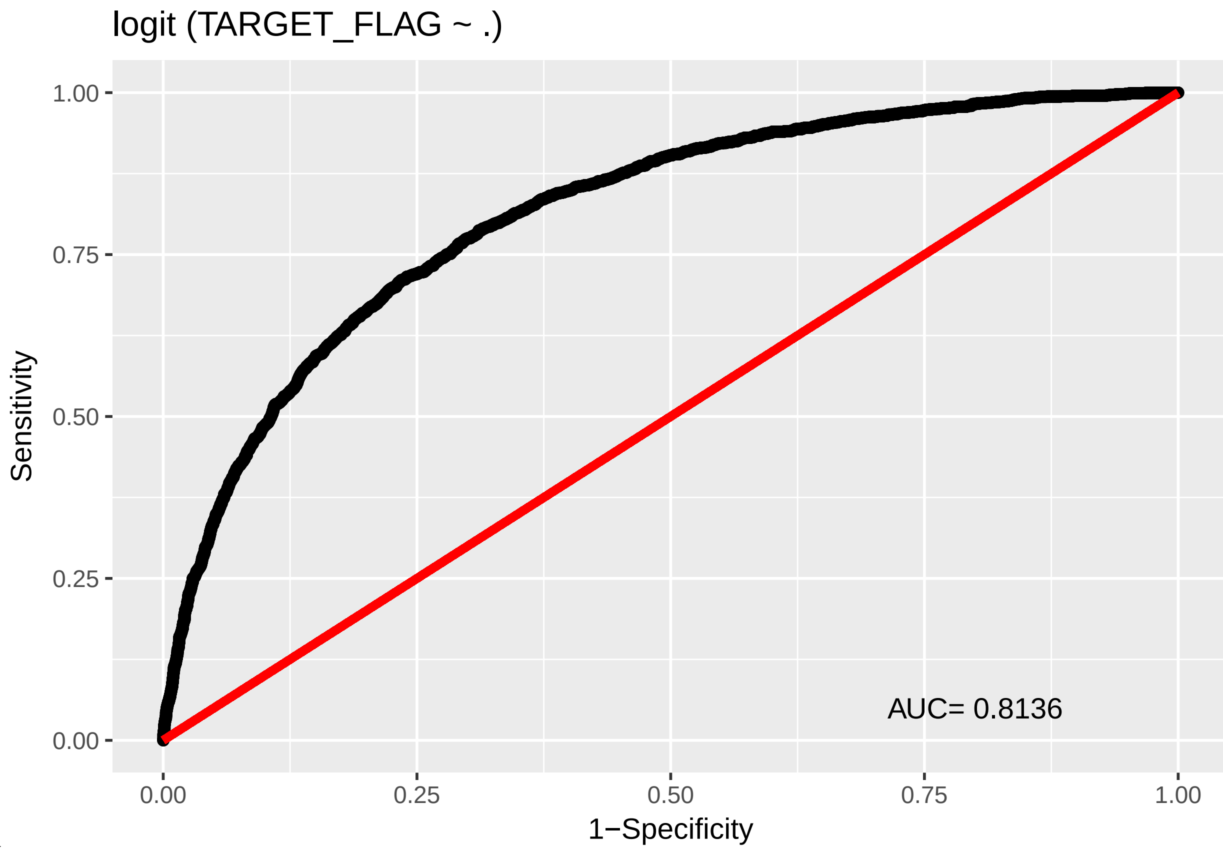
I am going to build one last model that will just predict claim amount from bluebook value. This is becasue car age is not signifigant in model3. This model makes a lot of sense, since the amount of payout is capped by the blue book value.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK, data = regression_training_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7757  -3083  -1541    295 101459
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.132e+03  3.295e+02  12.540  < 2e-16 ***
## BLUEBOOK    1.102e-01  1.997e-02   5.515  3.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7691 on 2151 degrees of freedom
## Multiple R-squared:  0.01394,    Adjusted R-squared:  0.01349
## F-statistic: 30.42 on 1 and 2151 DF,  p-value: 3.9e-08
```

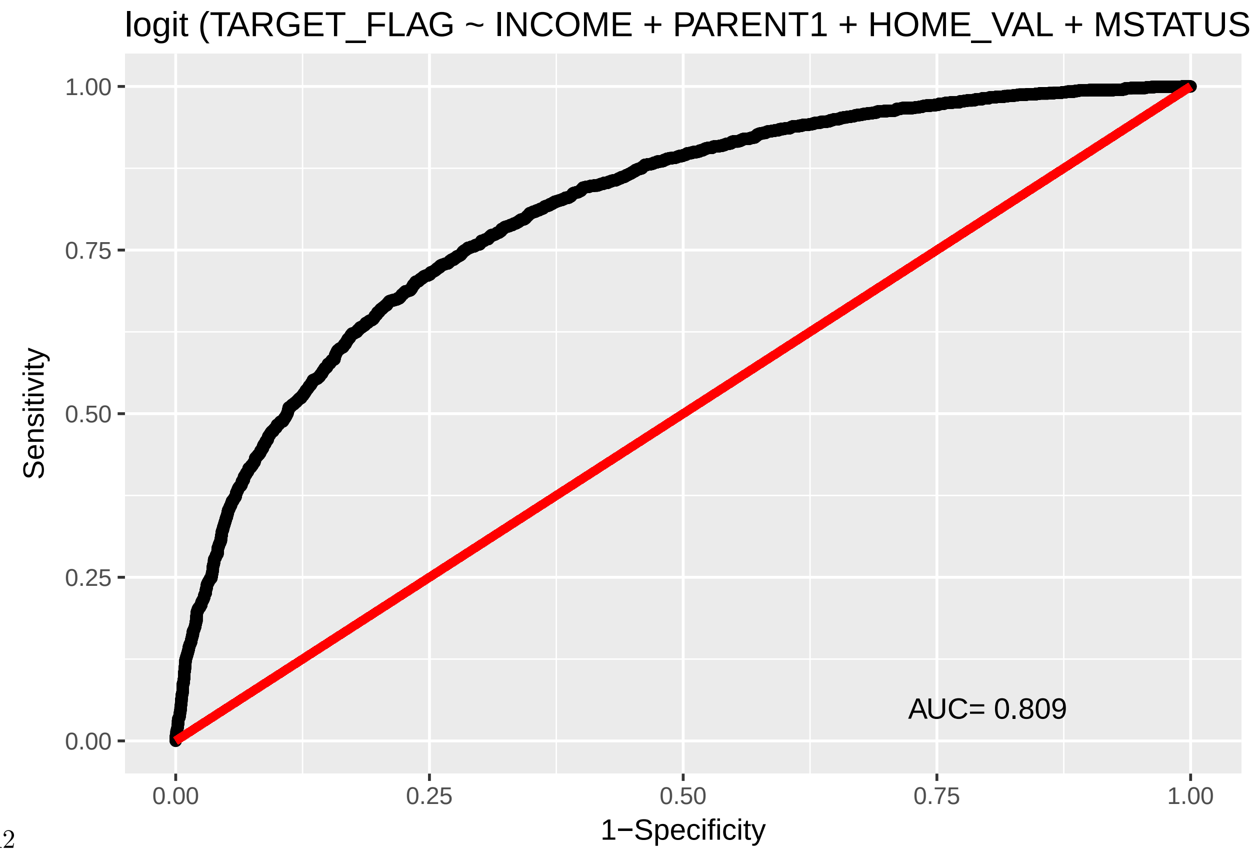
Model Selection

To Select Models I am going to look at the ROC curves and the area unver the curve.

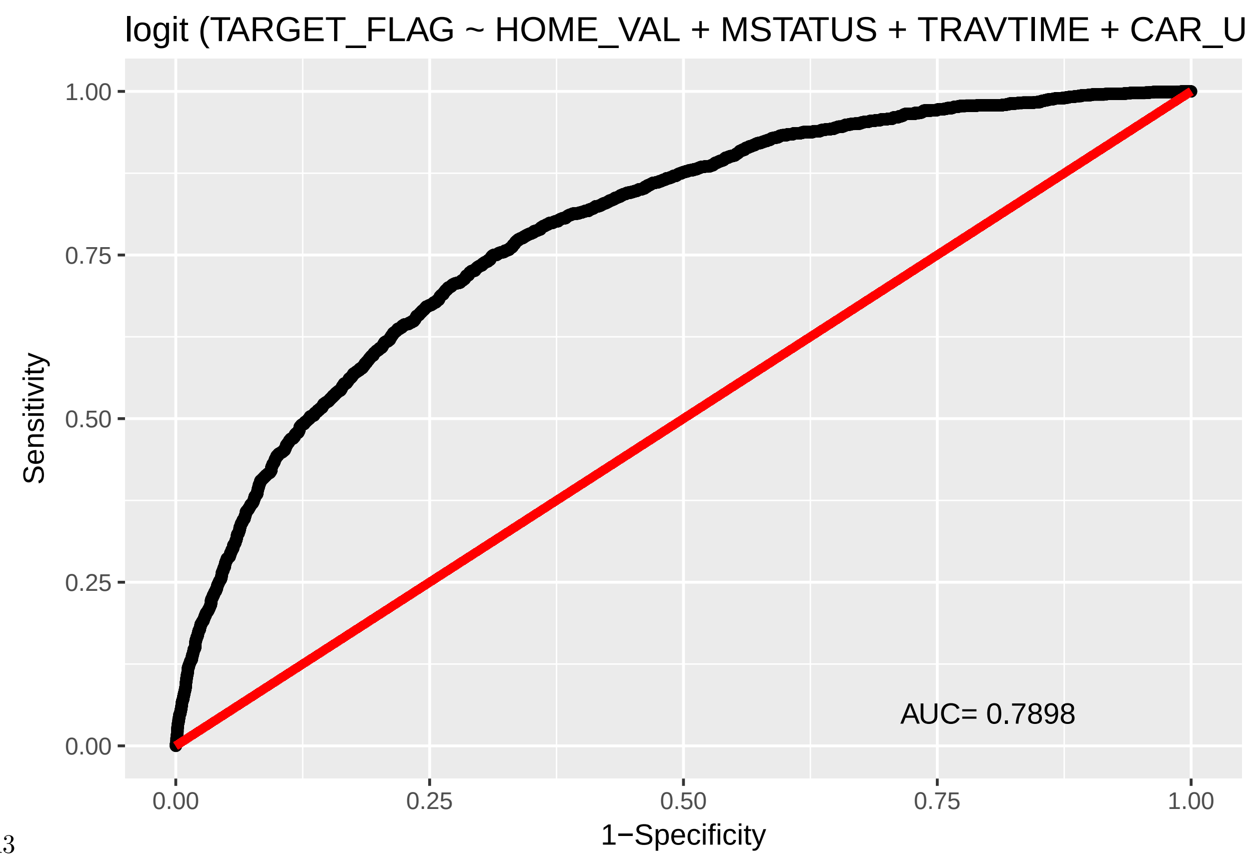
Model Selection Logistical Regression



Model1



Model2



Model3

I am going to select model3. All three models have about the same ROC, but model 3 is signifiganly simplier.

Claim Amount Regression Model Selection

I am going to pick the fourth model. This is becasue while all the models have about the same R^2 value, model 4 is signifigantly simplier.

Make Predications

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1
## 1     3           0   6552.017        0  48        0  11  52881      No
## 2     9           0   6217.110        1  40        1  11  50815     Yes
## 3    10           0   4781.638        0  44        2  12  43486     Yes
## 4    18           0   5148.493        0  35        2  11  21204     Yes
## 5    21           0   5830.425        0  59        0  12  87460      No
## 6    30           0   6958.532        0  46        0  14  51778      No
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE
## 1        0    z_No   M     Bachelors       Manager       26    Private
## 2        0    z_No   M z_High School       Manager       21    Private
## 3        0    z_No z_F z_High School z_Blue Collar       30 Commercial
## 4        0    z_No   M z_High School      Clerical       74    Private
## 5        0    z_No   M z_High School       Manager       45    Private
## 6   207519     Yes   M     Bachelors  Professional        7 Commercial
##   BLUEBOOK TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS
## 1    21970   1         Van     yes        0        0      No       2
## 2    18930   6     Minivan      no     3295        1      No       2
## 3     5900  10       z_SUV      no        0        0      No       0
## 4     9230   6      Pickup      no        0        0     Yes       0
## 5    15420   1     Minivan     yes    44857        2      No       4
## 6    25660   1 Panel Truck      no     2119        1      No       2
##   CAR_AGE           URBANICITY
## 1      10   Highly Urban/ Urban
## 2       1   Highly Urban/ Urban
## 3      10 z_Highly Rural/ Rural
## 4       4 z_Highly Rural/ Rural
## 5       1   Highly Urban/ Urban
## 6      12   Highly Urban/ Urban
```

Appendix R Code

library(ggplot2) library(reshape2) library(corrplot) library(forecast) library(dplyr) library(Deducer) library(tidyr) library(DataExplorer) library(speedglm)

Exploration

training_df <- read.csv (“insurance_training_data.csv”) eval_df <- read.csv(“insurance-evaluation-data.csv”) head(training_df)

summary(training_df)

histogram_df <- training_df[, c(“TARGET_AMT”, “KIDSDRIV”, “AGE”,“HOMEKIDS”, “YOJ”, “INCOME” , “HOME_VAL”, “TRAVTIME”, “BLUEBOOK”, “TIF”, “OLDCLAIM”, “CLM_FREQ”, “MVR_PTS”, “CAR_AGE”)] plot_histogram(histogram_df)

bar_df <- training_df[,c("PARENT1","MSTATUS","SEX","EDUCATION","JOB","CAR_USE","CAR_TYPE", "RED_CAR","REVOKED","URBANICITY")]

plot_bar(bar_df)

Transformation

training_df$CAR_AGE[training_df$CAR_AGE < 0] <- 0 training_df$JOB < -as.character(training_df$JOB) training_df$JOB[training_df$JOB == ""] <- "Unknown" training_df$JOB < -as.factor(training_df$JOB) training_df$INCOME < -as.numeric(gsub('[,]',',', training_df$INCOME))training_df$HOME_VAL <- as.numeric(gsub('[$,]',',', training_df$HOME_VAL))training_df$BLUEBOOK <- as.numeric(gsub('[$,]',',', training_df$BLUEBOOK))training_df$OLDCLAIM <- as.numeric(gsub('[$,]',',', training_df$OLDCLAIM))eval_df$CAR_AGE < 0] <- 0 eval_df$JOB < -as.character(eval_df$JOB) eval_df$JOB[eval_df$JOB == ""] <- "Unknown" eval_df$JOB < -as.factor(eval_df$JOB) eval_df$INCOME < -as.numeric(gsub('[,]', '', eval_df$INCOME))eval_df$HOME_VAL <- as.numeric(gsub('[$,]',',', eval_df$HOME_VAL))eval_df$BLUEBOOK <- as.numeric(gsub('[$,]',',', eval_df$BLUEBOOK))eval_df$OLDCLAIM <- as.numeric(gsub('[$,]',", eval_df$OLDCLAIM))

3)Fill missing values with median

training_df$AGE[is.na(training_df$AGE)] = median(training_df$AGE, na.rm = TRUE)training_df$CAR_AGE[is.na(training_ = median(training_df$CAR_AGE, na.rm = TRUE)training_df$INCOME[is.na(training_df$INCOME)] = median(training_df$INCOME, na.rm = TRUE)training_df$YOJ[is.na(training_df$YOJ)] = median(training_df$YOJ, na.rm TRUE)training_df$HOME_VAL[is.na(training_df$HOME_VAL)] = median(training_df$HOME_VAL, na.rm=TRUE)

eval_df$AGE[is.na(eval_df$AGE)] = median(eval_df$AGE, na.rm = TRUE)eval_df$CAR_AGE[is.na(eval_df$CAR_AGE)] = median(eval_df$CAR_AGE, na.rm = TRUE)eval_df$INCOME[is.na(eval_df$INCOME)] = median(eval_df$INCOME, na.rm TRUE)eval_df$YOJ[is.na(eval_df$YOJ)] = median(eval_df$YOJ, na.rm = TRUE)eval_df$HOME_VAL[is.na(eval_df$HOME_ = median(eval_df$HOME_VAL, na.rm=TRUE)

Builid models

Logistical Regression.

Model 1

trainng_logit_df <- training_df[, !names(training_df) %in% c("INDEX", "TARGET_AMT")]

logit_model1 <- glm (TARGET_FLAG ~ . , family = 'binomial', data=trainng_logit_df) summary(logit_model1)

Model 2

logit_model2 <- glm (TARGET_FLAG ~ INCOME+ PARENT1 + HOME_VAL+ MSTATUS+EDUCATION+JOB+ TRAVTIME +CAR_USE+ BLUEBOOK + TIF+ CAR_TYPE+OLDCLAIM+CLM_FREQ+REVOKED+MVR_PTS+URI , family = 'binomial', data=trainng_logit_df) summary(logit_model2)

MODEL 3

logit_model3 <- glm (TARGET_FLAG ~ HOME_VAL+ MSTATUS+ TRAVTIME +CAR_USE+ BLUEBOOK + TIF+ CAR_TYPE+OLDCLAIM+CLM_FREQ+REVOKED+MVR_PTS+URBANICITY , family = 'binomial', data=trainng_logit_df) summary(logit_model3)

Claim Amount Regression Model

Model 1

regression_training_df <- training_df[training_df$TARGET_FLAG == 1,] regression_training_df <- regression_training_df[, !names(regression_training_df) %in% c("INDEX", "TARGET_FLAG")]

reg_model1 <- lm(TARGET_AMT ~., data = regression_training_df) summary(reg_model1)

Model 2

reg_model2 <- lm(TARGET_AMT ~BLUEBOOK + REVOKED + CAR_AGE + EDUCATION, data = regression_training_df) summary(reg_model2)

Model 3

reg_model3 <- lm(TARGET_AMT ~BLUEBOOK + CAR_AGE , data = regression_training_df) summary(reg_model3)

Model 4

reg_model4 <- lm(TARGET_AMT ~BLUEBOOK , data = regression_training_df) summary(reg_model4)

Model Selection

Model Selection Logistical Regression

rocplot(logit_model1)

rocplot(logit_model2)

rocplot(logit_model3)

Claim Amount Regression Model Selection

Make Predications

probs <- predict(logit_model3,eval_df) prediction <- ifelse (probs > .5 ,1,0)

eval_df$TARGET_FLAG<-prediction

eval_df$TARGET_AMT < -with(eval_d f, ifelse(TARGET_F LAG == 0, BLUEBOOK * reg_m odel4$coefficients[2] + reg_model4$coefficients[1],0))

head(eval_df)