DATA 621 Assignment 4

Joby John,Jun Pan

April 18, 2019

In this, 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. We will 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.

#### 1. DATA EXPLORATION

Dataset contains 8161 rows and 25 variables. There are 6 observations missing data for Age, 454 missing data for YOK, 445 missing data for income and 464 missing data of home value. We will use mean or median of the variable to fill in any missing values. We will also tranform money and other formatted values to numerical values.

Below is the summary of the data.

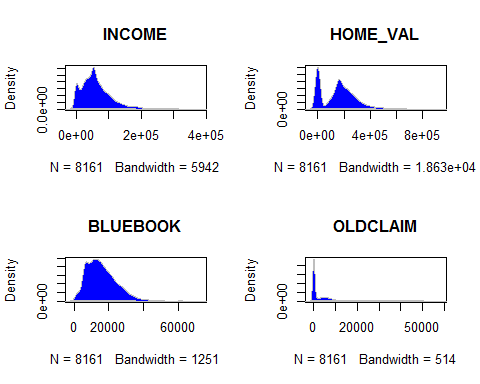
## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV   
## Min. : 1 Min. :0.0000 Min. : 0 Min. :0.0000   
## 1st Qu.: 2559 1st Qu.:0.0000 1st Qu.: 0 1st Qu.:0.0000   
## Median : 5133 Median :0.0000 Median : 0 Median :0.0000   
## Mean : 5152 Mean :0.2638 Mean : 1504 Mean :0.1711   
## 3rd Qu.: 7745 3rd Qu.:1.0000 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   
##   
##   
##   
##   
##

Data transformation

#### 2. DATA PREPARATION

Fix missing data with column median for income and home values; column mean for other columns. Convert indicator variables to 0s and 1s; 1 = Yes, Male for Sex, Commercial for Car Use, Red for RED\_CAR, and Highly Urban for URBANICITY Convert categorical predictor values to indicator variables - EDUCATION, CAR\_TYPE, JOB. Transform non normal variables like incone and house values using bobcox tranformation.

INCOME, HOME\_VAL, BLUEBOOK, and OLDCLAIM are represented as strings, so we will be extracting the numeric values for these.

Density plot of variables 

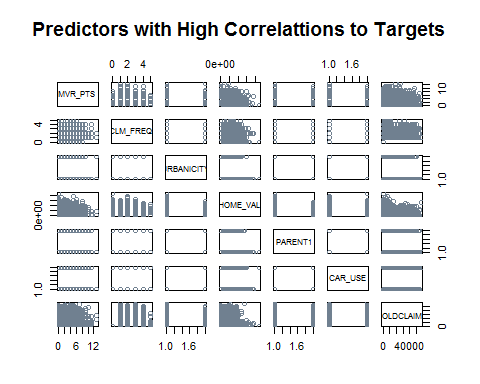
From the density plot, we can see that INCOME, HOME\_VAL, BLUEBOOK and OLDCLAIM are skewed. We are transforming these variables using ‘boxcoxfit’

## Fitted parameters:  
## lambda beta sigmasq   
## 0.4335048 265.9105875 6864.8508749   
##   
## Convergence code returned by optim: 0

## Fitted parameters:  
## lambda beta sigmasq   
## 0.1134775 26.3870095 2.8759681   
##   
## Convergence code returned by optim: 0

## Fitted parameters:  
## lambda beta sigmasq   
## 0.4610754 177.4257712 2217.4825612   
##   
## Convergence code returned by optim: 0

## Fitted parameters:  
## lambda beta sigmasq   
## -0.04511237 7.22517933 0.44041250   
##   
## Convergence code returned by optim: 0



#### 3. MODELS

#### Models for TARGET\_FLAG

##### Full Model

First model, we call it original\_full\_model, seeks the correlation between TARGET\_FLAG wiht original viarables. In this model, most of the variables are statistically significant with AIC score of 7373.6. We will compare the AIC values of other models to select a good model for predicting TARGET\_FLAG.

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ . - TARGET\_AMT, family = binomial(link = "logit"),   
## data = train\_flag)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5846 -0.7127 -0.3981 0.6263 3.1527   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.271e-01 3.215e-01 -2.884 0.003925 \*\*   
## KIDSDRIV 3.861e-01 6.122e-02 6.307 2.84e-10 \*\*\*  
## AGE -1.011e-03 4.020e-03 -0.251 0.801458   
## HOMEKIDS 4.976e-02 3.713e-02 1.340 0.180174   
## YOJ -1.121e-02 8.591e-03 -1.305 0.192013   
## INCOME -3.420e-06 1.082e-06 -3.162 0.001566 \*\*   
## PARENT1Yes 3.819e-01 1.096e-01 3.484 0.000493 \*\*\*  
## HOME\_VAL -1.306e-06 3.420e-07 -3.819 0.000134 \*\*\*  
## MSTATUSz\_No 4.937e-01 8.357e-02 5.908 3.46e-09 \*\*\*  
## SEXz\_F -8.256e-02 1.120e-01 -0.737 0.461177   
## EDUCATIONBachelors -3.803e-01 1.157e-01 -3.288 0.001007 \*\*   
## EDUCATIONMasters -2.885e-01 1.788e-01 -1.614 0.106502   
## EDUCATIONPhD -1.660e-01 2.139e-01 -0.776 0.437866   
## EDUCATIONz\_High\_School 1.790e-02 9.506e-02 0.188 0.850614   
## JOBClerical 4.109e-01 1.967e-01 2.089 0.036688 \*   
## JOBDoctor -4.458e-01 2.671e-01 -1.669 0.095115 .   
## JOBHome\_Maker 2.318e-01 2.102e-01 1.103 0.270060   
## JOBLawyer 1.049e-01 1.695e-01 0.619 0.535736   
## JOBManager -5.572e-01 1.716e-01 -3.248 0.001163 \*\*   
## JOBProfessional 1.619e-01 1.784e-01 0.907 0.364223   
## JOBStudent 2.155e-01 2.145e-01 1.005 0.315015   
## JOBz\_Blue\_Collar 3.108e-01 1.856e-01 1.675 0.093975 .   
## TRAVTIME 1.457e-02 1.883e-03 7.736 1.02e-14 \*\*\*  
## CAR\_USEPrivate -7.563e-01 9.172e-02 -8.246 < 2e-16 \*\*\*  
## BLUEBOOK -2.084e-05 5.263e-06 -3.959 7.52e-05 \*\*\*  
## TIF -5.546e-02 7.344e-03 -7.552 4.30e-14 \*\*\*  
## CAR\_TYPEPanel\_Truck 5.607e-01 1.618e-01 3.466 0.000529 \*\*\*  
## CAR\_TYPEPickup 5.540e-01 1.007e-01 5.500 3.79e-08 \*\*\*  
## CAR\_TYPESports\_Car 1.025e+00 1.299e-01 7.892 2.96e-15 \*\*\*  
## CAR\_TYPEVan 6.184e-01 1.265e-01 4.890 1.01e-06 \*\*\*  
## CAR\_TYPEz\_SUV 7.682e-01 1.113e-01 6.904 5.06e-12 \*\*\*  
## RED\_CARyes -9.684e-03 8.636e-02 -0.112 0.910718   
## OLDCLAIM -1.389e-05 3.910e-06 -3.553 0.000381 \*\*\*  
## CLM\_FREQ 1.959e-01 2.855e-02 6.864 6.70e-12 \*\*\*  
## REVOKEDYes 8.873e-01 9.133e-02 9.715 < 2e-16 \*\*\*  
## MVR\_PTS 1.133e-01 1.361e-02 8.323 < 2e-16 \*\*\*  
## CAR\_AGE -8.955e-04 7.541e-03 -0.119 0.905473   
## 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

##### Transformed Model: Including the transformed variables

Second model, we call it Transformed model, seeks the correlation between TARGET\_FLAG wiht original viarables. Skewed predictor variables are transformed in this model. In this model, most of the variables are statistically significant with AIC score of 7335.

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ . - TARGET\_AMT, family = binomial(link = "logit"),   
## data = train\_clean)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5688 -0.7106 -0.3929 0.6222 3.1687   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.635e-01 4.489e-01 1.032 0.301850   
## KIDSDRIV 3.954e-01 6.152e-02 6.427 1.30e-10 \*\*\*  
## AGE -2.074e-03 4.045e-03 -0.513 0.608084   
## HOMEKIDS 3.388e-02 3.754e-02 0.903 0.366708   
## YOJ 5.304e-03 9.871e-03 0.537 0.591049   
## INCOME 2.330e-06 2.432e-06 0.958 0.337855   
## PARENT1Yes 3.725e-01 1.102e-01 3.380 0.000725 \*\*\*  
## HOME\_VAL -3.300e-07 6.503e-07 -0.507 0.611893   
## MSTATUSz\_No 4.783e-01 8.683e-02 5.509 3.61e-08 \*\*\*  
## SEXz\_F -1.123e-01 1.126e-01 -0.998 0.318467   
## EDUCATIONBachelors -3.219e-01 1.174e-01 -2.741 0.006122 \*\*   
## EDUCATIONMasters -2.297e-01 1.801e-01 -1.275 0.202240   
## EDUCATIONPhD -2.107e-01 2.158e-01 -0.977 0.328776   
## EDUCATIONz\_High\_School 6.263e-02 9.629e-02 0.650 0.515415   
## JOBClerical 4.116e-01 1.971e-01 2.088 0.036755 \*   
## JOBDoctor -3.960e-01 2.659e-01 -1.489 0.136378   
## JOBHome\_Maker -1.141e-02 2.246e-01 -0.051 0.959483   
## JOBLawyer 1.171e-01 1.698e-01 0.690 0.490377   
## JOBManager -5.339e-01 1.713e-01 -3.117 0.001825 \*\*   
## JOBProfessional 1.903e-01 1.786e-01 1.066 0.286586   
## JOBStudent -1.316e-01 2.323e-01 -0.567 0.570890   
## JOBz\_Blue\_Collar 3.637e-01 1.860e-01 1.955 0.050597 .   
## TRAVTIME 1.472e-02 1.892e-03 7.778 7.35e-15 \*\*\*  
## CAR\_USEPrivate -7.493e-01 9.219e-02 -8.129 4.34e-16 \*\*\*  
## BLUEBOOK 4.183e-05 1.971e-05 2.122 0.033823 \*   
## TIF -5.468e-02 7.364e-03 -7.426 1.12e-13 \*\*\*  
## CAR\_TYPEPanel\_Truck 4.210e-01 1.658e-01 2.539 0.011106 \*   
## CAR\_TYPEPickup 5.625e-01 1.011e-01 5.565 2.63e-08 \*\*\*  
## CAR\_TYPESports\_Car 1.025e+00 1.304e-01 7.861 3.80e-15 \*\*\*  
## CAR\_TYPEVan 6.282e-01 1.265e-01 4.964 6.89e-07 \*\*\*  
## CAR\_TYPEz\_SUV 7.966e-01 1.122e-01 7.099 1.25e-12 \*\*\*  
## RED\_CARyes -4.836e-03 8.663e-02 -0.056 0.955486   
## OLDCLAIM -2.611e-05 4.775e-06 -5.468 4.56e-08 \*\*\*  
## CLM\_FREQ 4.481e-02 4.418e-02 1.014 0.310445   
## REVOKEDYes 9.524e-01 9.306e-02 10.235 < 2e-16 \*\*\*  
## MVR\_PTS 9.563e-02 1.411e-02 6.778 1.22e-11 \*\*\*  
## CAR\_AGE -4.036e-04 7.555e-03 -0.053 0.957395   
## URBANICITYz\_Highly\_Rural/ Rural -2.366e+00 1.140e-01 -20.750 < 2e-16 \*\*\*  
## INCOME\_MOD -7.703e-03 2.346e-03 -3.284 0.001025 \*\*   
## HOME\_VAL\_MOD -7.876e-02 4.661e-02 -1.690 0.091103 .   
## BLUEBOOK\_MOD -2.344e-02 7.161e-03 -3.274 0.001061 \*\*   
## OLD\_CLAIM\_MOD 6.943e-02 1.512e-02 4.593 4.38e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9418 on 8160 degrees of freedom  
## Residual deviance: 7251 on 8119 degrees of freedom  
## AIC: 7335  
##   
## Number of Fisher Scoring iterations: 5

##### Step Model

Third model, we call it Step model, seeks the correlation between TARGET\_FLAG with original viarables. This model uses “Stepwise Algorithm” on “Transformed Model”. In this model, most of the variables are statistically significant with AIC score of 7322. This model inludes only 32 predictor variables, other models have 42 predictors variables. This is a simpler model than the previous models with better AIC.

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + EDUCATION +   
## JOB + TRAVTIME + CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM +   
## REVOKED + MVR\_PTS + URBANICITY + INCOME\_MOD + HOME\_VAL\_MOD +   
## BLUEBOOK\_MOD + OLD\_CLAIM\_MOD, family = binomial(link = "logit"),   
## data = train\_clean)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5767 -0.7103 -0.3942 0.6202 3.1652   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.917e-01 4.036e-01 0.970 0.331875   
## KIDSDRIV 4.191e-01 5.522e-02 7.589 3.23e-14 \*\*\*  
## PARENT1Yes 4.388e-01 9.477e-02 4.630 3.66e-06 \*\*\*  
## MSTATUSz\_No 4.483e-01 8.302e-02 5.400 6.68e-08 \*\*\*  
## EDUCATIONBachelors -3.350e-01 1.101e-01 -3.041 0.002356 \*\*   
## EDUCATIONMasters -2.412e-01 1.622e-01 -1.488 0.136867   
## EDUCATIONPhD -2.033e-01 1.974e-01 -1.030 0.303004   
## EDUCATIONz\_High\_School 5.246e-02 9.560e-02 0.549 0.583183   
## JOBClerical 4.097e-01 1.962e-01 2.088 0.036810 \*   
## JOBDoctor -3.991e-01 2.651e-01 -1.505 0.132230   
## JOBHome\_Maker 2.704e-03 2.178e-01 0.012 0.990097   
## JOBLawyer 9.757e-02 1.692e-01 0.577 0.564154   
## JOBManager -5.475e-01 1.709e-01 -3.204 0.001358 \*\*   
## JOBProfessional 1.733e-01 1.781e-01 0.973 0.330611   
## JOBStudent -9.926e-02 2.258e-01 -0.440 0.660188   
## JOBz\_Blue\_Collar 3.543e-01 1.853e-01 1.912 0.055905 .   
## TRAVTIME 1.471e-02 1.890e-03 7.783 7.09e-15 \*\*\*  
## CAR\_USEPrivate -7.484e-01 9.206e-02 -8.129 4.32e-16 \*\*\*  
## BLUEBOOK 3.991e-05 1.908e-05 2.092 0.036465 \*   
## TIF -5.451e-02 7.358e-03 -7.408 1.28e-13 \*\*\*  
## CAR\_TYPEPanel\_Truck 4.848e-01 1.541e-01 3.146 0.001655 \*\*   
## CAR\_TYPEPickup 5.572e-01 1.009e-01 5.521 3.37e-08 \*\*\*  
## CAR\_TYPESports\_Car 9.442e-01 1.081e-01 8.732 < 2e-16 \*\*\*  
## CAR\_TYPEVan 6.594e-01 1.225e-01 5.381 7.42e-08 \*\*\*  
## CAR\_TYPEz\_SUV 7.177e-01 8.647e-02 8.299 < 2e-16 \*\*\*  
## OLDCLAIM -2.705e-05 4.689e-06 -5.768 8.03e-09 \*\*\*  
## REVOKEDYes 9.570e-01 9.299e-02 10.292 < 2e-16 \*\*\*  
## MVR\_PTS 9.563e-02 1.409e-02 6.788 1.14e-11 \*\*\*  
## URBANICITYz\_Highly\_Rural/ Rural -2.369e+00 1.140e-01 -20.786 < 2e-16 \*\*\*  
## INCOME\_MOD -5.794e-03 9.700e-04 -5.973 2.33e-09 \*\*\*  
## HOME\_VAL\_MOD -9.997e-02 2.430e-02 -4.114 3.89e-05 \*\*\*  
## BLUEBOOK\_MOD -2.372e-02 7.059e-03 -3.361 0.000778 \*\*\*  
## OLD\_CLAIM\_MOD 8.116e-02 9.828e-03 8.258 < 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 on 8160 degrees of freedom  
## Residual deviance: 7256 on 8128 degrees of freedom  
## AIC: 7322  
##   
## Number of Fisher Scoring iterations: 5

#### Models for Amount

All TARGET\_AMT modeling are done using the tranformed data. We are building models with and with out ‘TARGET\_FLAG’ to see its impact on models. Models with ‘TARGET\_FLAG’ have higher Adjusted R-squared values.

##### Full Model

First model includes all variables excluding index. This model has Adjusted R-squared of 0.2886 and only few variables are statistically significatint. A model with out ‘TARGET\_FLAG’ has low Adjusted R-squared of 0.06665. Model with ‘TARGET\_FLAG’ has fewer statistically significatint predictor. Since we are predicting claim amount, we believe we dont need to include ‘TARGET\_FLAG’ and any observation where ‘TARGET\_FLAG’ is zero. However, model with ‘TARGET\_FLAG’ has Adjusted R-squared value and needs further investigation. Both models are statistically significant comapred to the null model. We build models that includes TARGET\_FLAG and models with out TARGET\_FLAG.

##   
## Call:  
## lm(formula = TARGET\_AMT ~ ., data = train\_clean, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6156 -474 -79 229 101001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.077e+03 6.920e+02 -3.002 0.00269 \*\*   
## TARGET\_FLAG1 5.731e+03 1.139e+02 50.333 < 2e-16 \*\*\*  
## KIDSDRIV -3.979e+01 9.912e+01 -0.401 0.68812   
## AGE 7.055e+00 6.192e+00 1.139 0.25458   
## HOMEKIDS 5.175e+01 5.741e+01 0.901 0.36737   
## YOJ -5.062e+00 1.488e+01 -0.340 0.73373   
## INCOME -5.774e-03 3.586e-03 -1.610 0.10741   
## PARENT1Yes 1.568e+02 1.769e+02 0.887 0.37534   
## HOME\_VAL -5.119e-04 9.437e-04 -0.542 0.58749   
## MSTATUSz\_No 2.013e+02 1.310e+02 1.537 0.12439   
## SEXz\_F -2.657e+02 1.607e+02 -1.653 0.09841 .   
## EDUCATIONBachelors 1.249e+01 1.809e+02 0.069 0.94495   
## EDUCATIONMasters 1.760e+02 2.629e+02 0.669 0.50328   
## EDUCATIONPhD 4.605e+02 3.113e+02 1.479 0.13913   
## EDUCATIONz\_High\_School -1.597e+02 1.511e+02 -1.057 0.29033   
## JOBClerical -8.189e-01 2.983e+02 -0.003 0.99781   
## JOBDoctor -3.001e+02 3.570e+02 -0.841 0.40052   
## JOBHome\_Maker 1.212e+02 3.348e+02 0.362 0.71734   
## JOBLawyer 6.255e+01 2.583e+02 0.242 0.80864   
## JOBManager -1.405e+02 2.521e+02 -0.557 0.57744   
## JOBProfessional 1.542e+02 2.698e+02 0.571 0.56779   
## JOBStudent 1.930e+02 3.534e+02 0.546 0.58493   
## JOBz\_Blue\_Collar 2.050e+01 2.816e+02 0.073 0.94196   
## TRAVTIME 5.233e-01 2.823e+00 0.185 0.85296   
## CAR\_USEPrivate -9.743e+01 1.443e+02 -0.675 0.49945   
## BLUEBOOK -3.292e-02 2.907e-02 -1.132 0.25754   
## TIF -3.571e+00 1.067e+01 -0.335 0.73799   
## CAR\_TYPEPanel\_Truck 6.048e+01 2.476e+02 0.244 0.80703   
## CAR\_TYPEPickup -2.563e+01 1.494e+02 -0.172 0.86378   
## CAR\_TYPESports\_Car 2.335e+02 1.911e+02 1.222 0.22185   
## CAR\_TYPEVan 7.793e+01 1.865e+02 0.418 0.67612   
## CAR\_TYPEz\_SUV 1.424e+02 1.572e+02 0.905 0.36530   
## RED\_CARyes -2.562e+01 1.302e+02 -0.197 0.84400   
## OLDCLAIM 5.356e-03 7.867e-03 0.681 0.49601   
## CLM\_FREQ -1.379e+01 7.627e+01 -0.181 0.85649   
## REVOKEDYes -3.338e+02 1.543e+02 -2.163 0.03054 \*   
## MVR\_PTS 5.762e+01 2.343e+01 2.460 0.01392 \*   
## CAR\_AGE -2.556e+01 1.118e+01 -2.287 0.02219 \*   
## URBANICITYz\_Highly\_Rural/ Rural 3.004e+01 1.273e+02 0.236 0.81347   
## INCOME\_MOD 5.669e+00 3.487e+00 1.626 0.10400   
## HOME\_VAL\_MOD 8.547e+01 7.038e+01 1.214 0.22461   
## BLUEBOOK\_MOD 2.409e+01 1.082e+01 2.226 0.02604 \*   
## OLD\_CLAIM\_MOD -1.358e+01 2.552e+01 -0.532 0.59480   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3968 on 8118 degrees of freedom  
## Multiple R-squared: 0.2922, Adjusted R-squared: 0.2886   
## F-statistic: 79.8 on 42 and 8118 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = TARGET\_AMT ~ ., data = train1, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8194 -3200 -1477 475 99357   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.301e+02 2.517e+03 0.091 0.9272   
## KIDSDRIV -1.928e+02 3.175e+02 -0.607 0.5438   
## AGE 1.883e+01 2.128e+01 0.885 0.3765   
## HOMEKIDS 2.418e+02 2.089e+02 1.158 0.2472   
## YOJ -1.850e+01 5.647e+01 -0.328 0.7433   
## INCOME -2.033e-02 1.487e-02 -1.367 0.1718   
## PARENT1Yes 3.048e+02 5.886e+02 0.518 0.6046   
## HOME\_VAL -3.385e-04 3.975e-03 -0.085 0.9321   
## MSTATUSz\_No 8.846e+02 5.088e+02 1.738 0.0823 .  
## SEXz\_F -1.318e+03 6.617e+02 -1.992 0.0465 \*  
## EDUCATIONBachelors 1.371e+02 6.480e+02 0.212 0.8325   
## EDUCATIONMasters 1.013e+03 1.089e+03 0.930 0.3526   
## EDUCATIONPhD 2.548e+03 1.314e+03 1.939 0.0526 .  
## EDUCATIONz\_High\_School -4.982e+02 5.198e+02 -0.958 0.3379   
## JOBClerical 2.707e+02 1.204e+03 0.225 0.8222   
## JOBDoctor -2.268e+03 1.768e+03 -1.283 0.1996   
## JOBHome\_Maker 4.942e+02 1.325e+03 0.373 0.7093   
## JOBLawyer 2.596e+02 1.031e+03 0.252 0.8013   
## JOBManager -8.400e+02 1.068e+03 -0.787 0.4315   
## JOBProfessional 9.565e+02 1.133e+03 0.844 0.3988   
## JOBStudent 9.383e+02 1.369e+03 0.685 0.4932   
## JOBz\_Blue\_Collar 3.878e+02 1.152e+03 0.337 0.7364   
## TRAVTIME -3.967e-02 1.108e+01 -0.004 0.9971   
## CAR\_USEPrivate -4.098e+02 5.231e+02 -0.783 0.4335   
## BLUEBOOK -2.436e-02 1.153e-01 -0.211 0.8328   
## TIF -1.568e+01 4.251e+01 -0.369 0.7123   
## CAR\_TYPEPanel\_Truck -2.714e+02 1.000e+03 -0.271 0.7862   
## CAR\_TYPEPickup -9.873e+01 5.975e+02 -0.165 0.8688   
## CAR\_TYPESports\_Car 1.029e+03 7.506e+02 1.371 0.1706   
## CAR\_TYPEVan 2.776e+01 7.711e+02 0.036 0.9713   
## CAR\_TYPEz\_SUV 7.911e+02 6.711e+02 1.179 0.2387   
## RED\_CARyes -2.126e+02 4.975e+02 -0.427 0.6691   
## OLDCLAIM 3.624e-02 2.775e-02 1.306 0.1917   
## CLM\_FREQ -1.906e+01 2.366e+02 -0.081 0.9358   
## REVOKEDYes -1.185e+03 5.308e+02 -2.233 0.0257 \*  
## MVR\_PTS 1.273e+02 7.030e+01 1.810 0.0704 .  
## CAR\_AGE -9.674e+01 4.391e+01 -2.203 0.0277 \*  
## URBANICITYz\_Highly\_Rural/ Rural -1.126e+02 7.567e+02 -0.149 0.8817   
## INCOME\_MOD 1.624e+01 1.338e+01 1.214 0.2249   
## HOME\_VAL\_MOD 2.236e+02 2.642e+02 0.846 0.3974   
## BLUEBOOK\_MOD 5.441e+01 4.052e+01 1.343 0.1795   
## OLD\_CLAIM\_MOD -4.956e+01 8.318e+01 -0.596 0.5514   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7689 on 2111 degrees of freedom  
## Multiple R-squared: 0.03285, Adjusted R-squared: 0.01407   
## F-statistic: 1.749 on 41 and 2111 DF, p-value: 0.00241

##### Step

We used Stepwise Algorithm on above models and we got simliar results. We see from the results below, the full model has better Adjusted R-squared and more predictors are statistically significant. “model.step” which include TARGET\_AMT has better Adjusted R-squared than model with out. These models have fewer predictor varables and all the variables are statistically significant in the model with out TARGET\_AMT. These models are lot simpler than above models.

##   
## Call:  
## lm(formula = TARGET\_AMT ~ TARGET\_FLAG + INCOME + MSTATUS + SEX +   
## REVOKED + MVR\_PTS + CAR\_AGE + INCOME\_MOD + HOME\_VAL\_MOD +   
## BLUEBOOK\_MOD, data = train\_clean, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6204 -419 -75 202 101300   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.242e+03 2.427e+02 -5.119 3.15e-07 \*\*\*  
## TARGET\_FLAG1 5.729e+03 1.054e+02 54.369 < 2e-16 \*\*\*  
## INCOME -3.758e-03 2.465e-03 -1.525 0.1274   
## MSTATUSz\_No 2.406e+02 1.116e+02 2.155 0.0312 \*   
## SEXz\_F -1.660e+02 8.888e+01 -1.867 0.0619 .   
## REVOKEDYes -2.907e+02 1.355e+02 -2.145 0.0320 \*   
## MVR\_PTS 5.079e+01 2.098e+01 2.421 0.0155 \*   
## CAR\_AGE -1.330e+01 8.662e+00 -1.535 0.1247   
## INCOME\_MOD 3.363e+00 2.197e+00 1.531 0.1259   
## HOME\_VAL\_MOD 5.115e+01 3.597e+01 1.422 0.1551   
## BLUEBOOK\_MOD 1.196e+01 2.221e+00 5.387 7.37e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3964 on 8150 degrees of freedom  
## Multiple R-squared: 0.2906, Adjusted R-squared: 0.2897   
## F-statistic: 333.9 on 10 and 8150 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = TARGET\_AMT ~ MSTATUS + SEX + REVOKED + MVR\_PTS +   
## CAR\_AGE + HOME\_VAL\_MOD + BLUEBOOK\_MOD, data = train1, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7832 -3124 -1568 413 100209   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1966.524 775.159 2.537 0.0113 \*   
## MSTATUSz\_No 940.922 405.741 2.319 0.0205 \*   
## SEXz\_F -626.656 333.867 -1.877 0.0607 .   
## REVOKEDYes -668.298 409.207 -1.633 0.1026   
## MVR\_PTS 129.193 64.243 2.011 0.0445 \*   
## CAR\_AGE -52.243 31.509 -1.658 0.0975 .   
## HOME\_VAL\_MOD 208.787 120.953 1.726 0.0845 .   
## BLUEBOOK\_MOD 42.839 7.676 5.581 2.69e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7664 on 2145 degrees of freedom  
## Multiple R-squared: 0.02356, Adjusted R-squared: 0.02037   
## F-statistic: 7.393 on 7 and 2145 DF, p-value: 8.408e-09

##### Cross Validation

We see simliar results from this models as well. We see from the results below the full model has better Adjusted R-squared than model with out TARGET\_FLAG.

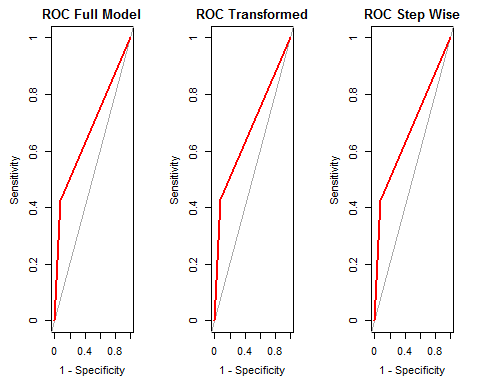
##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6156 -474 -79 229 101001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.077e+03 6.920e+02 -3.002 0.00269  
## TARGET\_FLAG1 5.731e+03 1.139e+02 50.333 < 2e-16  
## KIDSDRIV -3.979e+01 9.912e+01 -0.401 0.68812  
## AGE 7.055e+00 6.192e+00 1.139 0.25458  
## HOMEKIDS 5.175e+01 5.741e+01 0.901 0.36737  
## YOJ -5.062e+00 1.488e+01 -0.340 0.73373  
## INCOME -5.774e-03 3.586e-03 -1.610 0.10741  
## PARENT1Yes 1.568e+02 1.769e+02 0.887 0.37534  
## HOME\_VAL -5.119e-04 9.437e-04 -0.542 0.58749  
## MSTATUSz\_No 2.013e+02 1.310e+02 1.537 0.12439  
## SEXz\_F -2.657e+02 1.607e+02 -1.653 0.09841  
## EDUCATIONBachelors 1.249e+01 1.809e+02 0.069 0.94495  
## EDUCATIONMasters 1.760e+02 2.629e+02 0.669 0.50328  
## EDUCATIONPhD 4.605e+02 3.113e+02 1.479 0.13913  
## EDUCATIONz\_High\_School -1.597e+02 1.511e+02 -1.057 0.29033  
## JOBClerical -8.189e-01 2.983e+02 -0.003 0.99781  
## JOBDoctor -3.001e+02 3.570e+02 -0.841 0.40052  
## JOBHome\_Maker 1.212e+02 3.348e+02 0.362 0.71734  
## JOBLawyer 6.255e+01 2.583e+02 0.242 0.80864  
## JOBManager -1.405e+02 2.521e+02 -0.557 0.57744  
## JOBProfessional 1.542e+02 2.698e+02 0.571 0.56779  
## JOBStudent 1.930e+02 3.534e+02 0.546 0.58493  
## JOBz\_Blue\_Collar 2.050e+01 2.816e+02 0.073 0.94196  
## TRAVTIME 5.233e-01 2.823e+00 0.185 0.85296  
## CAR\_USEPrivate -9.743e+01 1.443e+02 -0.675 0.49945  
## BLUEBOOK -3.292e-02 2.907e-02 -1.132 0.25754  
## TIF -3.571e+00 1.067e+01 -0.335 0.73799  
## CAR\_TYPEPanel\_Truck 6.048e+01 2.476e+02 0.244 0.80703  
## CAR\_TYPEPickup -2.563e+01 1.494e+02 -0.172 0.86378  
## CAR\_TYPESports\_Car 2.335e+02 1.911e+02 1.222 0.22185  
## CAR\_TYPEVan 7.793e+01 1.865e+02 0.418 0.67612  
## CAR\_TYPEz\_SUV 1.424e+02 1.572e+02 0.905 0.36530  
## RED\_CARyes -2.562e+01 1.302e+02 -0.197 0.84400  
## OLDCLAIM 5.356e-03 7.867e-03 0.681 0.49601  
## CLM\_FREQ -1.379e+01 7.627e+01 -0.181 0.85649  
## REVOKEDYes -3.338e+02 1.543e+02 -2.163 0.03054  
## MVR\_PTS 5.762e+01 2.343e+01 2.460 0.01392  
## CAR\_AGE -2.556e+01 1.118e+01 -2.287 0.02219  
## `URBANICITYz\_Highly\_Rural/ Rural` 3.004e+01 1.273e+02 0.236 0.81347  
## INCOME\_MOD 5.669e+00 3.487e+00 1.626 0.10400  
## HOME\_VAL\_MOD 8.547e+01 7.038e+01 1.214 0.22461  
## BLUEBOOK\_MOD 2.409e+01 1.082e+01 2.226 0.02604  
## OLD\_CLAIM\_MOD -1.358e+01 2.552e+01 -0.532 0.59480  
##   
## (Intercept) \*\*   
## TARGET\_FLAG1 \*\*\*  
## KIDSDRIV   
## AGE   
## HOMEKIDS   
## YOJ   
## INCOME   
## PARENT1Yes   
## HOME\_VAL   
## MSTATUSz\_No   
## SEXz\_F .   
## EDUCATIONBachelors   
## EDUCATIONMasters   
## EDUCATIONPhD   
## EDUCATIONz\_High\_School   
## JOBClerical   
## JOBDoctor   
## JOBHome\_Maker   
## JOBLawyer   
## JOBManager   
## JOBProfessional   
## JOBStudent   
## JOBz\_Blue\_Collar   
## TRAVTIME   
## CAR\_USEPrivate   
## BLUEBOOK   
## TIF   
## CAR\_TYPEPanel\_Truck   
## CAR\_TYPEPickup   
## CAR\_TYPESports\_Car   
## CAR\_TYPEVan   
## CAR\_TYPEz\_SUV   
## RED\_CARyes   
## OLDCLAIM   
## CLM\_FREQ   
## REVOKEDYes \*   
## MVR\_PTS \*   
## CAR\_AGE \*   
## `URBANICITYz\_Highly\_Rural/ Rural`   
## INCOME\_MOD   
## HOME\_VAL\_MOD   
## BLUEBOOK\_MOD \*   
## OLD\_CLAIM\_MOD   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3968 on 8118 degrees of freedom  
## Multiple R-squared: 0.2922, Adjusted R-squared: 0.2886   
## F-statistic: 79.8 on 42 and 8118 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8194 -3200 -1477 475 99357   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.301e+02 2.517e+03 0.091 0.9272   
## KIDSDRIV -1.928e+02 3.175e+02 -0.607 0.5438   
## AGE 1.883e+01 2.128e+01 0.885 0.3765   
## HOMEKIDS 2.418e+02 2.089e+02 1.158 0.2472   
## YOJ -1.850e+01 5.647e+01 -0.328 0.7433   
## INCOME -2.033e-02 1.487e-02 -1.367 0.1718   
## PARENT1Yes 3.048e+02 5.886e+02 0.518 0.6046   
## HOME\_VAL -3.385e-04 3.975e-03 -0.085 0.9321   
## MSTATUSz\_No 8.846e+02 5.088e+02 1.738 0.0823 .  
## SEXz\_F -1.318e+03 6.617e+02 -1.992 0.0465 \*  
## EDUCATIONBachelors 1.371e+02 6.480e+02 0.212 0.8325   
## EDUCATIONMasters 1.013e+03 1.089e+03 0.930 0.3526   
## EDUCATIONPhD 2.548e+03 1.314e+03 1.939 0.0526 .  
## EDUCATIONz\_High\_School -4.982e+02 5.198e+02 -0.958 0.3379   
## JOBClerical 2.707e+02 1.204e+03 0.225 0.8222   
## JOBDoctor -2.268e+03 1.768e+03 -1.283 0.1996   
## JOBHome\_Maker 4.942e+02 1.325e+03 0.373 0.7093   
## JOBLawyer 2.596e+02 1.031e+03 0.252 0.8013   
## JOBManager -8.400e+02 1.068e+03 -0.787 0.4315   
## JOBProfessional 9.565e+02 1.133e+03 0.844 0.3988   
## JOBStudent 9.383e+02 1.369e+03 0.685 0.4932   
## JOBz\_Blue\_Collar 3.878e+02 1.152e+03 0.337 0.7364   
## TRAVTIME -3.967e-02 1.108e+01 -0.004 0.9971   
## CAR\_USEPrivate -4.098e+02 5.231e+02 -0.783 0.4335   
## BLUEBOOK -2.436e-02 1.153e-01 -0.211 0.8328   
## TIF -1.568e+01 4.251e+01 -0.369 0.7123   
## CAR\_TYPEPanel\_Truck -2.714e+02 1.000e+03 -0.271 0.7862   
## CAR\_TYPEPickup -9.873e+01 5.975e+02 -0.165 0.8688   
## CAR\_TYPESports\_Car 1.029e+03 7.506e+02 1.371 0.1706   
## CAR\_TYPEVan 2.776e+01 7.711e+02 0.036 0.9713   
## CAR\_TYPEz\_SUV 7.911e+02 6.711e+02 1.179 0.2387   
## RED\_CARyes -2.126e+02 4.975e+02 -0.427 0.6691   
## OLDCLAIM 3.624e-02 2.775e-02 1.306 0.1917   
## CLM\_FREQ -1.906e+01 2.366e+02 -0.081 0.9358   
## REVOKEDYes -1.185e+03 5.308e+02 -2.233 0.0257 \*  
## MVR\_PTS 1.273e+02 7.030e+01 1.810 0.0704 .  
## CAR\_AGE -9.674e+01 4.391e+01 -2.203 0.0277 \*  
## `URBANICITYz\_Highly\_Rural/ Rural` -1.126e+02 7.567e+02 -0.149 0.8817   
## INCOME\_MOD 1.624e+01 1.338e+01 1.214 0.2249   
## HOME\_VAL\_MOD 2.236e+02 2.642e+02 0.846 0.3974   
## BLUEBOOK\_MOD 5.441e+01 4.052e+01 1.343 0.1795   
## OLD\_CLAIM\_MOD -4.956e+01 8.318e+01 -0.596 0.5514   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7689 on 2111 degrees of freedom  
## Multiple R-squared: 0.03285, Adjusted R-squared: 0.01407   
## F-statistic: 1.749 on 41 and 2111 DF, p-value: 0.00241

1. SELECT MODELS (25 Points)

##### ROC and AUC

Sensitivity, Specificity ROC and Area under curver are very similar for all the models. We cant select one model over another based on these.



## Sensitivity Specificity Pos Pred Value   
## 0.8180077 0.6676364 0.9239348   
## Neg Pred Value Precision Recall   
## 0.4263818 0.9239348 0.8180077   
## F1 Prevalence Detection Rate   
## 0.8677505 0.8315157 0.6801863   
## Detection Prevalence Balanced Accuracy   
## 0.7361843 0.7428220

## Sensitivity Specificity Pos Pred Value   
## 0.8180879 0.6652205 0.9229361   
## Neg Pred Value Precision Recall   
## 0.4273107 0.9229361 0.8180879   
## F1 Prevalence Detection Rate   
## 0.8673549 0.8305355 0.6794510   
## Detection Prevalence Balanced Accuracy   
## 0.7361843 0.7416542

## Sensitivity Specificity Pos Pred Value   
## 0.8174591 0.6673977 0.9242676   
## Neg Pred Value Precision Recall   
## 0.4240595 0.9242676 0.8174591   
## F1 Prevalence Detection Rate   
## 0.8675885 0.8323735 0.6804313   
## Detection Prevalence Balanced Accuracy   
## 0.7361843 0.7424284

## Full Transformed Step Transformed  
## AUC 0.6751583 0.6751234 0.6741635

#### AIC

All three models have very simliar AIC. Transformed and Transformed Step models have slightly better AIC values.

## Full Transformed Transformed Step  
## AIC 7373.591 7335.034 7321.984

#### Deviance

Based on the ANOVA summary of the deviance we Transformed and Transformed Step models are slightly better than the full model.

## Analysis of Deviance Table   
## Resid. Df Resid. Dev Df Deviance  
## 1 8123 7297.6   
## 2 8119 7251.0 4 46.557  
## 3 8128 7256.0 -9 -4.950

We will choose the transform\_model\_step.glm model for predicting ‘TARGET\_FLAG’ because of its simplicity, fewer predictors, and slightly better AIC/Deviance over the other two models.

##### Models for predicting Target amount

Based on the table below we see that all models with ‘TARGET\_FLAG’ have higher Adjusted R-squared values. Models without TARGET\_FLAG and excluding TARGET\_FLAG1=0 observation have low Adjusted R-squared values, therefore doesnt explain the variablity in the data well. RMSE values for the models with TARGET\_FLAG are far better than the models without TARGET\_FLAG, suggesting models with TARGET\_FLAG1 are better fits the data. We believe FullOnlyFlag1Step is a better model for predicting TARGET\_AMOUNT. We choose this model based on its simplcity and by examing the predicted target amount on the train data. However, we are concerned about this model’s low Adjusted R-squared value.

## FullModel FullOnlyFlag1 FullModelStep   
## 0.288557224 0.014068488 0.289727934   
## FullOnlyFlag1Step CrossValidation CrossValidationFalg1   
## 0.020371268 0.304935460 0.009724781

## FullModel FullOnlyFlag1 FullModelStep   
## 3957.242 7613.161 3961.770   
## FullOnlyFlag1Step CrossValidation CrossValidationFalg1   
## 7649.656 3891.799 7661.976

# Test the reduced model on evaluation data

All three models have same prediction for ‘TARGET\_FLAG’. Different models predict different TARGET\_AMT. A better approach to predict ‘TARGET\_AMOUNT’ would be to predict TARGET\_FLAG and then run prediction on observation with TARGET\_FLAG=1; these predictions seem to be more realistic.

# R Code

library(car) library(caTools) library(caret) library(corrplot) library(data.table) library(dplyr) library(geoR) library(ggplot2) library(grid) library(gridExtra) library(kableExtra) library(knitr) library(MASS) library(naniar) library(nortest) library(pROC) library(pscl) library(psych) library(reshape) library(PerformanceAnalytics) library(ROCR) library(testthat) library(UpSetR) require(leaps) train <- read.csv(“<https://raw.githubusercontent.com/jjohn81/DATA621_Assignment4/master/insurance_training.csv>”) test <- read.csv(“<https://raw.githubusercontent.com/jjohn81/DATA621_Assignment4/master/insurance-evaluation.csv>”) test\_actual <- test

summary(train)

# remove “$" money = function(input) { out = sub("\\$”, “”, input)

out = as.numeric(sub(“,”, “”, out)) return(out) }

# Remove " “, replace with”*" underscore = function(input) { out = sub(" “,”*“, input) return(out) }

train = as.tbl(train) %>% mutate\_at(c(“INCOME”,“HOME\_VAL”,“BLUEBOOK”,“OLDCLAIM”), money) %>% mutate\_at(c(“EDUCATION”,“JOB”,“CAR\_TYPE”,“URBANICITY”), underscore) %>% mutate\_at(c(“EDUCATION”,“JOB”,“CAR\_TYPE”,“URBANICITY”), as.factor) %>% mutate(TARGET\_FLAG = as.factor(TARGET\_FLAG))

test = as.tbl(test) %>% mutate\_at(c(“INCOME”,“HOME\_VAL”,“BLUEBOOK”,“OLDCLAIM”), money) %>% mutate\_at(c(“EDUCATION”,“JOB”,“CAR\_TYPE”,“URBANICITY”), underscore) %>% mutate\_at(c(“EDUCATION”,“JOB”,“CAR\_TYPE”,“URBANICITY”), as.factor) %>% mutate(TARGET\_FLAG = as.factor(TARGET\_FLAG))

trainAGE)] <- mean(trainYOJ[is.na(train$YOJ)] <- mean(trainHOME\_VAL[is.na(train$HOME\_VAL)] <- median(trainCAR\_AGE[is.na(train$CAR\_AGE)] <- mean(trainINCOME[is.na(train$INCOME)] <- median(trainINDEX <- NULL

testAGE)] <- mean(testYOJ[is.na(test$YOJ)] <- mean(testHOME\_VAL[is.na(test$HOME\_VAL)] <- median(testCAR\_AGE[is.na(test$CAR\_AGE)] <- mean(testINCOME[is.na(test$INCOME)] <- median(test$INCOME, na.rm=TRUE) #test <- test[complete.cases(test),] test\_clean <- test test\_clean$INDEX <- NULL

train\_cleanINCOME) train\_cleanHOME\_VAL) train\_cleanBLUEBOOK) train\_cleanOLDCLAIM)

test\_cleanINCOME) test\_cleanHOME\_VAL) test\_cleanBLUEBOOK) test\_cleanOLDCLAIM)

ntrain<-select\_if(train\_clean, is.numeric) ntrain <- ntrain[,c(“INCOME”,“HOME\_VAL”,“BLUEBOOK”,“OLDCLAIM”)] ntrain <- as.data.frame((ntrain))

par(mfrow=c(2, 2)) colnames <- dimnames(ntrain)[[2]]

for(col in 1:4) { d <- density(na.omit(ntrain[,col])) plot(d, type=“n”, main=colnames[col]) polygon(d, col=“blue”, border=“gray”) }

boxcoxfit(train\_cleanINCOME >0]) train\_cleanINCOME ^0.443 boxcoxfit(train\_cleanHOME\_VAL > 0]) train\_cleanHOME\_VAL^0.102 boxcoxfit(train\_cleanBLUEBOOK\_MOD <- train\_cleanOLDCLAIM[train\_clean$OLDCLAIM>0]) train\_cleanOLDCLAIM + 1)

test\_cleanINCOME 0.443 test\_cleanHOME\_VAL0.102 test\_cleanBLUEBOOK^0.461 test\_cleanOLDCLAIM + 1)

pairs(~MVR\_PTS+CLM\_FREQ+URBANICITY+HOME\_VAL+PARENT1+CAR\_USE+OLDCLAIM, data=train\_clean, main=“Predictors with High Correlattions to Targets”, col=“slategrey”) train\_flag <- train[,-c(1)] full.model.glm <- glm(TARGET\_FLAG ~.-TARGET\_AMT , data = train\_flag, family = binomial(link=‘logit’)) summary(full.model.glm)

transform\_model.glm <- glm(TARGET\_FLAG ~.-TARGET\_AMT, data = train\_clean, family = binomial(link=‘logit’)) summary(transform\_model.glm) transform\_model\_step.glm <- step( transform\_model.glm,trace = 0, keep = NULL) summary(transform\_model\_step.glm)

full.model <- lm(TARGET\_AMT ~ . , data = train\_clean, na.action = na.exclude) summary(full.model)

train1 <- train\_clean[ which(train\_clean$TARGET\_FLAG==1),] train1 <- train1[,-c(1)]

full.model.onlyFlag1 <- lm(TARGET\_AMT ~ . ,data= train1, na.action = na.exclude)

summary(full.model.onlyFlag1)

model.step <- step(full.model, trace = 0, keep = NULL) summary(model.step) model.onlyFlga1.step <- step(full.model.onlyFlag1, trace = 0, keep = NULL) summary(model.onlyFlga1.step)

set.seed(123) train.control <- trainControl(method = “cv”, number = 10) # Train the model xModel <- train(TARGET\_AMT ~., data = train\_clean, method = “lm”, trControl = train.control)

summary(xModel)

train1 <- train\_clean[ which(train\_clean$TARGET\_FLAG==1),] train1 <- train1[,-c(1)] xModel\_onlyFlag1 <- train(TARGET\_AMT ~., data = train1, method = “lm”, trControl = train.control)

summary(xModel\_onlyFlag1)

par(mfrow=c(1, 3)) train\_flagTARGET\_FLAG, train\_flag$predict\_full\_glm, plot=T, asp=NA, legacy.axes=T, main = “ROC Full Model”, col=“red”, levels=c(0,1))

train\_clean$transform\_model\_glm <- ifelse(predict(transform\_model.glm, train\_clean, type=‘response’) > 0.5,1,0)

roc\_transform\_model\_glm <- roc(train\_cleantransform\_model\_glm, plot=T, asp=NA, legacy.axes=T, main = “ROC Transformed”, col=“red” ,levels=c(0,1))

train\_clean$transform\_model\_step\_glm <- ifelse(predict(transform\_model\_step.glm, train\_clean, type=‘response’) > 0.5,1,0)

roc\_transform\_model\_step\_glm <- roc(train\_cleantransform\_model\_step\_glm, plot=T, asp=NA, legacy.axes=T, main = “ROC Step Wise”, col=“red”, levels=c(0,1))

confusionMatrix(table(train\_flagpredict\_full\_glm))TARGET\_FLAG, train\_cleanbyClass confusionMatrix(table(train\_cleantransform\_model\_step\_glm))$byClass

auc.model <- matrix(c(auc(roc\_full\_glm),auc(roc\_transform\_model\_glm), auc(roc\_transform\_model\_step\_glm)), ncol = 3, nrow=1, byrow = T) colnames(auc.model)<-c(“Full”, “Transformed”, “Step Transformed”) rownames(auc.model)<- c(“AUC”) auc.model

aic <- matrix(c(full.model.glmaic,transform\_model\_step.glm$aic), ncol = 3, nrow=1, byrow = T) colnames(aic)<-c(“Full”, “Transformed”, “Transformed Step”) rownames(aic)<- c(“AIC”) aic

anova(full.model.glm, transform\_model.glm, transform\_model\_step.glm)

rsq <-c(summary(full.model)adj.r.squared, summary(model.step)adj.r.squared, xModelRsquared, xModel\_onlyFlag1Rsquared) modelnames <- c(‘FullModel’, ‘FullOnlyFlag1’, ‘FullModelStep’,‘FullOnlyFlag1Step’,‘CrossValidation’,‘CrossValidationFalg1’ ) names(rsq)<- modelnames rsq rmse <- c(sqrt(mean(full.modelresiduals2)), sqrt(mean(model.stepresiduals2)), xModelRMSE, xModel\_onlyFlag1RMSE)

names(rmse)<- modelnames rmse

test\_cleanTARGET\_AMT)

predicted <- predict(full.model.glm, newdata = test\_clean, type=“response”) test\_clean$TARGET\_FLAG <- as.factor(ifelse(predicted > 0.5,1,0))

test\_clean$TARGET\_FLAG\_Full <- as.factor(ifelse(predicted > 0.5,1,0))

preditcted <- predict(transform\_model.glm, newdata = test\_clean, type=“response”) test\_clean$TARGET\_FLAG\_Transform\_Model <- ifelse(predicted > 0.5,1,0) predicted <- predict(transform\_model\_step.glm, newdata = test\_clean, type="response") test\_clean$TARGET\_FLAG\_Transform\_Model\_Step <- ifelse(predicted > 0.5,1,0)

test\_cleanTARGET\_AMT\_Full\_Flag1 <- predict(full.model.onlyFlag1, newdata = test\_clean, type=“response”) test\_clean$TARGET\_AMT\_Step <- predict(model.step, newdata = test\_clean, type="response") test\_clean$TARGET\_AMT\_Step\_Flag1 <- predict(model.onlyFlga1.step, newdata = test\_clean, type=“response”) test\_cleanTARGET\_AMT\_xModel\_onlyFlag1 <- predict(xModel\_onlyFlag1, newdata = test\_clean) write.csv(test\_clean,“Data.csv”)