DATA621_Homework4_JR

Jeyaraman Ramalingam

5/5/2021

Contents

Overview

In this homework assignment, you 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. Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided).

Data Exploration

Insurance Training Data

Sample

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PA
1	0	0	0	60	0	11	$67,349	No
2	0	0	0	43	0	11	$91,449	No
4	0	0	0	35	1	10	$16,039	No
5	0	0	0	51	0	14		No
6	0	0	0	50	0	NA	$114,986	No
7	1	2946	0	34	1	12	$125,301	Yes

Input Dataset Summaries

```
##      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   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION              JOB              TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF          CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM            CLM_FREQ        REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

Missing Data Check

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1
3	NA	NA	0	48	0	11	$52,881	No
9	NA	NA	1	40	1	11	$50,815	Yes
10	NA	NA	0	44	2	12	$43,486	Yes
18	NA	NA	0	35	2	NA	$21,204	Yes
21	NA	NA	0	59	0	12	$87,460	No
30	NA	NA	0	46	0	14		No

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         510           0
```

Insurance Evaluation Data

Sample

Input Dataset Summaries

```
##      INDEX       TARGET_FLAG    TARGET_AMT        KIDSDRIV          AGE
##  Min.   :    3   Mode:logical   Mode:logical   Min.   :0.0000   Min.   :17.00
##  1st Qu.: 2632   NA's:2141      NA's:2141      1st Qu.:0.0000   1st Qu.:39.00
##  Median : 5224                                 Median :0.0000   Median :45.00
##  Mean   : 5150                                 Mean   :0.1625   Mean   :45.02
##  3rd Qu.: 7669                                 3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :10300                                 Max.   :3.0000   Max.   :73.00
##                                                                 NA's   :1
##     HOMEKIDS            YOJ           INCOME            PARENT1
##  Min.   :0.0000   Min.   : 0.00   Length:2141        Length:2141
##  1st Qu.:0.0000   1st Qu.: 9.00   Class :character   Class :character
##  Median :0.0000   Median :11.00   Mode  :character   Mode  :character
##  Mean   :0.7174   Mean   :10.38
##  3rd Qu.:1.0000   3rd Qu.:13.00
##  Max.   :5.0000   Max.   :19.00
##                   NA's   :94
##    HOME_VAL           MSTATUS              SEX             EDUCATION
##  Length:2141        Length:2141        Length:2141        Length:2141
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      JOB              TRAVTIME       CAR_USE            BLUEBOOK
##  Length:2141        Min.   :  5.00   Length:2141        Length:2141
##  Class :character   1st Qu.: 22.00   Class :character   Class :character
##  Mode  :character   Median : 33.00   Mode  :character   Mode  :character
##                     Mean   : 33.15
##                     3rd Qu.: 43.00
##                     Max.   :105.00
##
##       TIF           CAR_TYPE           RED_CAR            OLDCLAIM
##  Min.   : 1.000   Length:2141        Length:2141        Length:2141
##  1st Qu.: 1.000   Class :character   Class :character   Class :character
##  Median : 4.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5.245
##  3rd Qu.: 7.000
##  Max.   :25.000
##
##     CLM_FREQ       REVOKED             MVR_PTS          CAR_AGE
##  Min.   :0.000   Length:2141        Min.   : 0.000   Min.   : 0.000
##  1st Qu.:0.000   Class :character   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.000   Mode  :character   Median : 1.000   Median : 8.000
##  Mean   :0.809                      Mean   : 1.766   Mean   : 8.183
##  3rd Qu.:2.000                      3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.000                      Max.   :12.000   Max.   :26.000
##                                                      NA’s   :129
##   URBANICITY
##  Length:2141
##  Class :character
##  Mode  :character
##
##
##
##
```

Missing Data Check

```
##        INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##            0        2141        2141           0           1           0
##          YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##           94           0           0           0           0           0
##    EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##            0           0           0           0           0           0
##     CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##            0           0           0           0           0           0
##      CAR_AGE  URBANICITY
##          129           0
```

Findings

The findings from Data Exploration on Training and Evaluation dataset are below.

1. Some of the character columns are prefixed by ‘z_’ which needs to be corrected
2. Numeric format for Dollar Amount fields need to be fixed.
3. Imputation needs to be done for the missing values.

We will perform all of these exercises in the Data Preparation step.

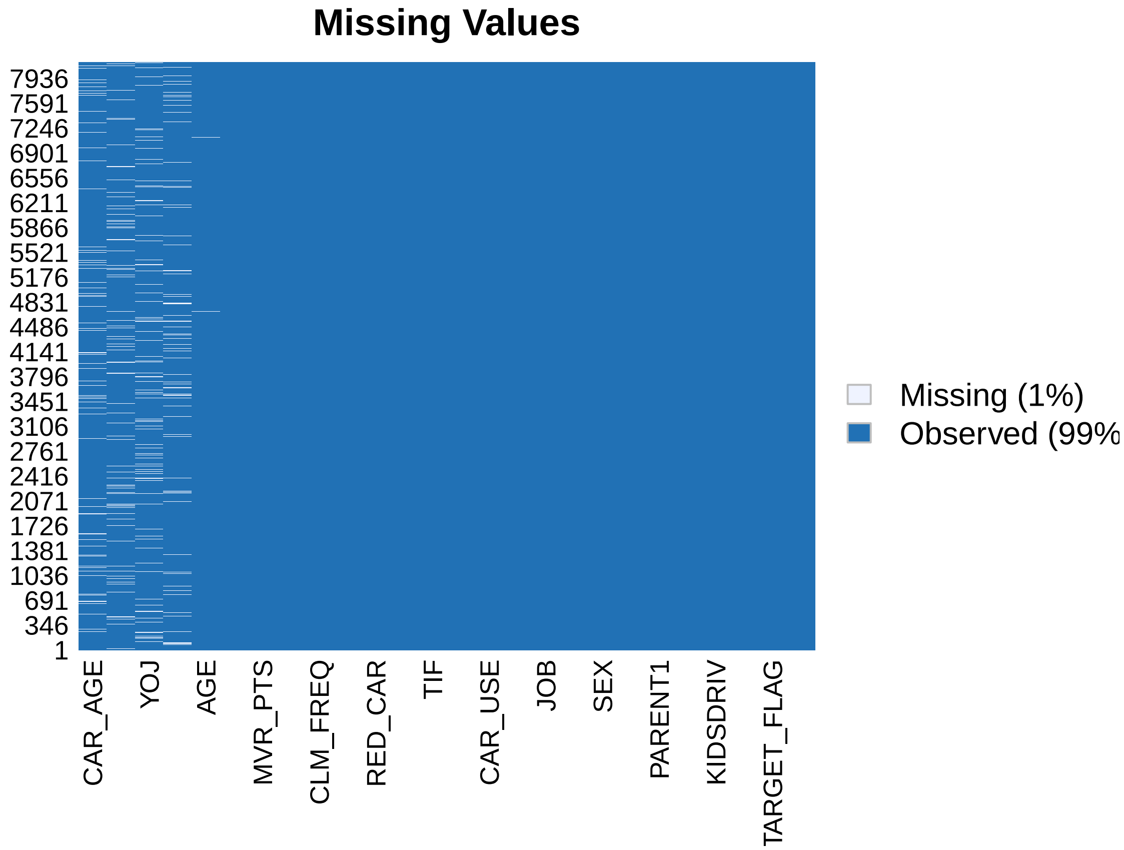
INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1
1	0	0	0	60	0	11	67349	No
2	0	0	0	43	0	11	91449	No
4	0	0	0	35	1	10	16039	No
5	0	0	0	51	0	14	NA	No
6	0	0	0	50	0	NA	114986	No
7	1	2946	0	34	1	12	125301	Yes

Data Preparation

Training Data - Fix Formatting

Training Data - Missing Data Check

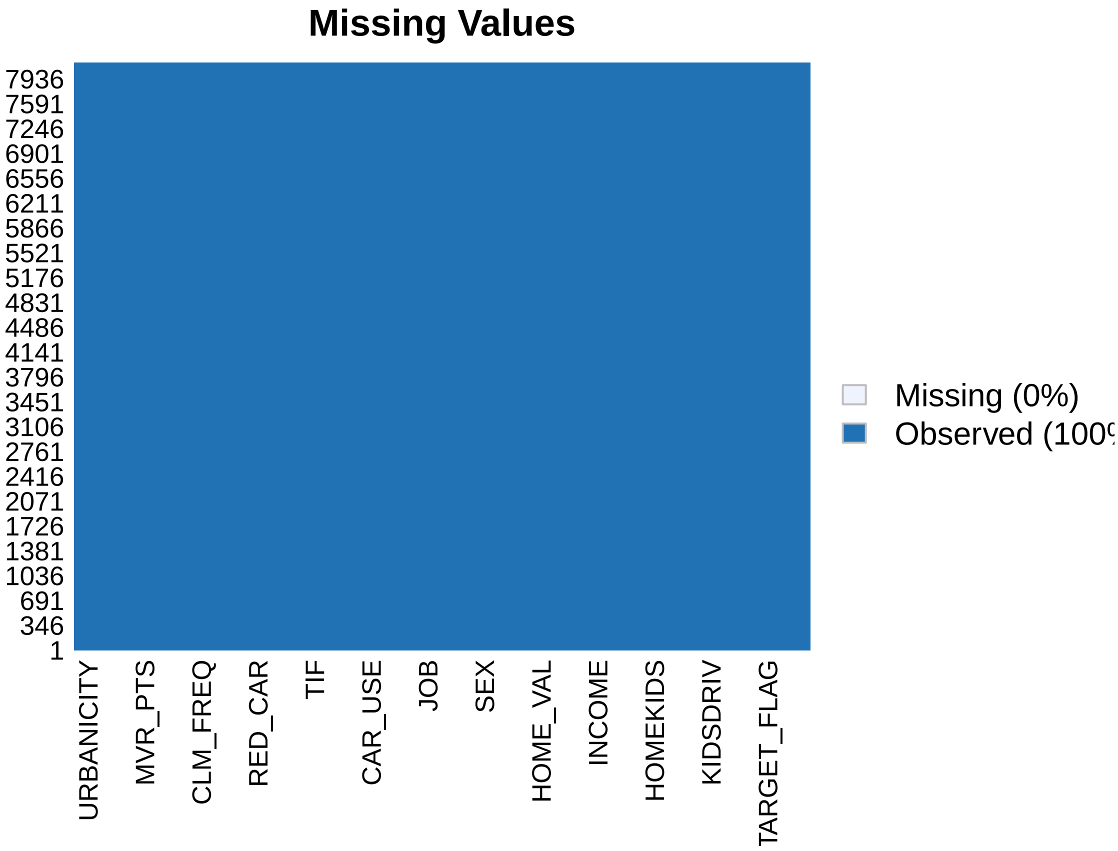
```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454         445           0         464           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         510           0
```



	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
	Min. : 1	Min. :0.0000	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.
	1st Qu.: 2559	1st Qu.:0.0000	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.:
	Median : 5133	Median :0.0000	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :
	Mean : 5152	Mean :0.2638	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212	Mean :10
	3rd Qu.: 7745	3rd Qu.:1.0000	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1
	Max. :10302	Max. :1.0000	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23

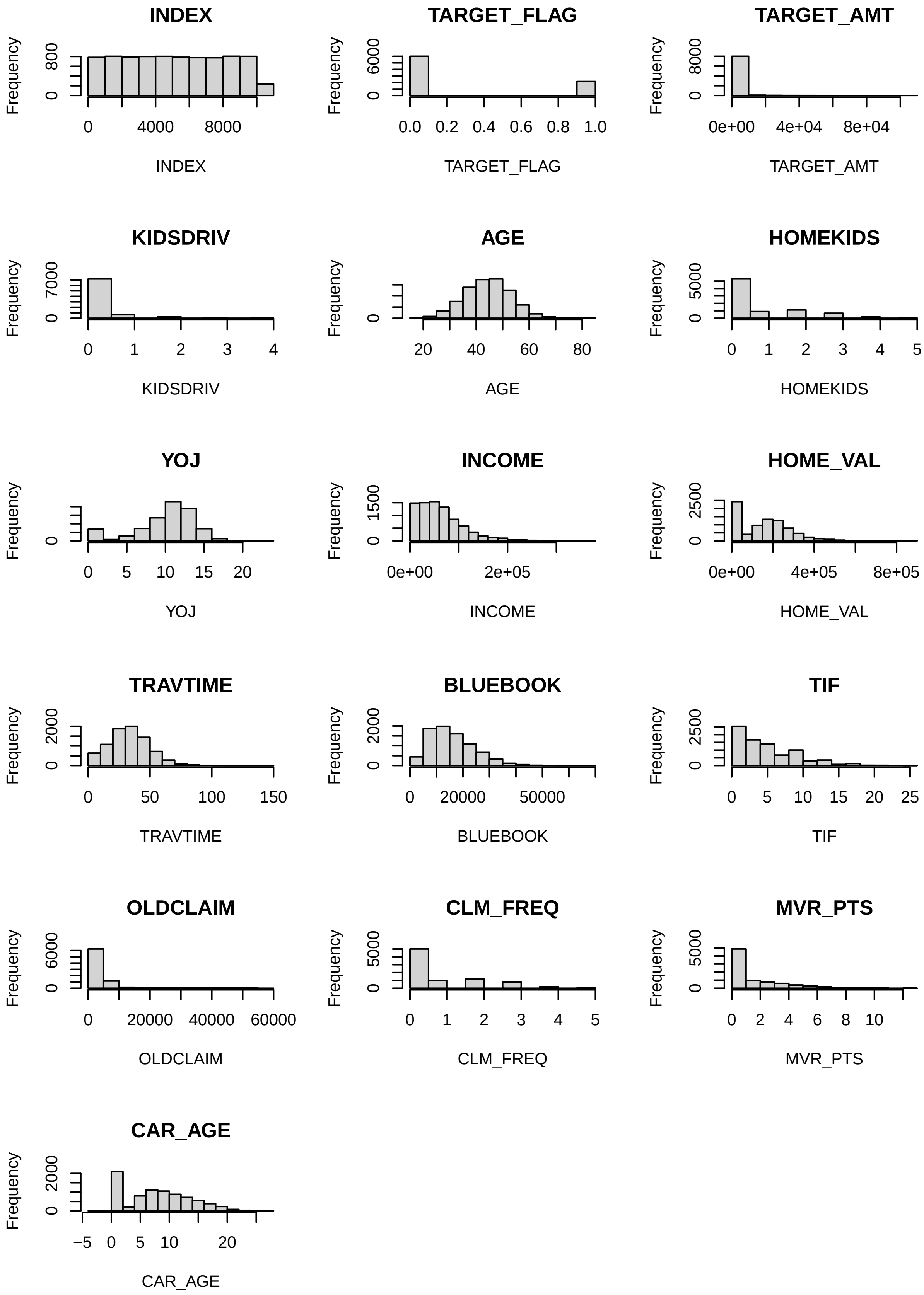
Training Data - Missing Data Re-test

```
##        INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##            0           0           0           0           0           0
##          YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##            0           0           0           0           0           0
##    EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##            0           0           0           0           0           0
##     CAR_TYPE     RED_CAR    OLDCLAIM   CLM_FREQ     REVOKED     MVR_PTS
##            0           0           0           0           0           0
##      CAR_AGE  URBANICITY
##            0           0
```

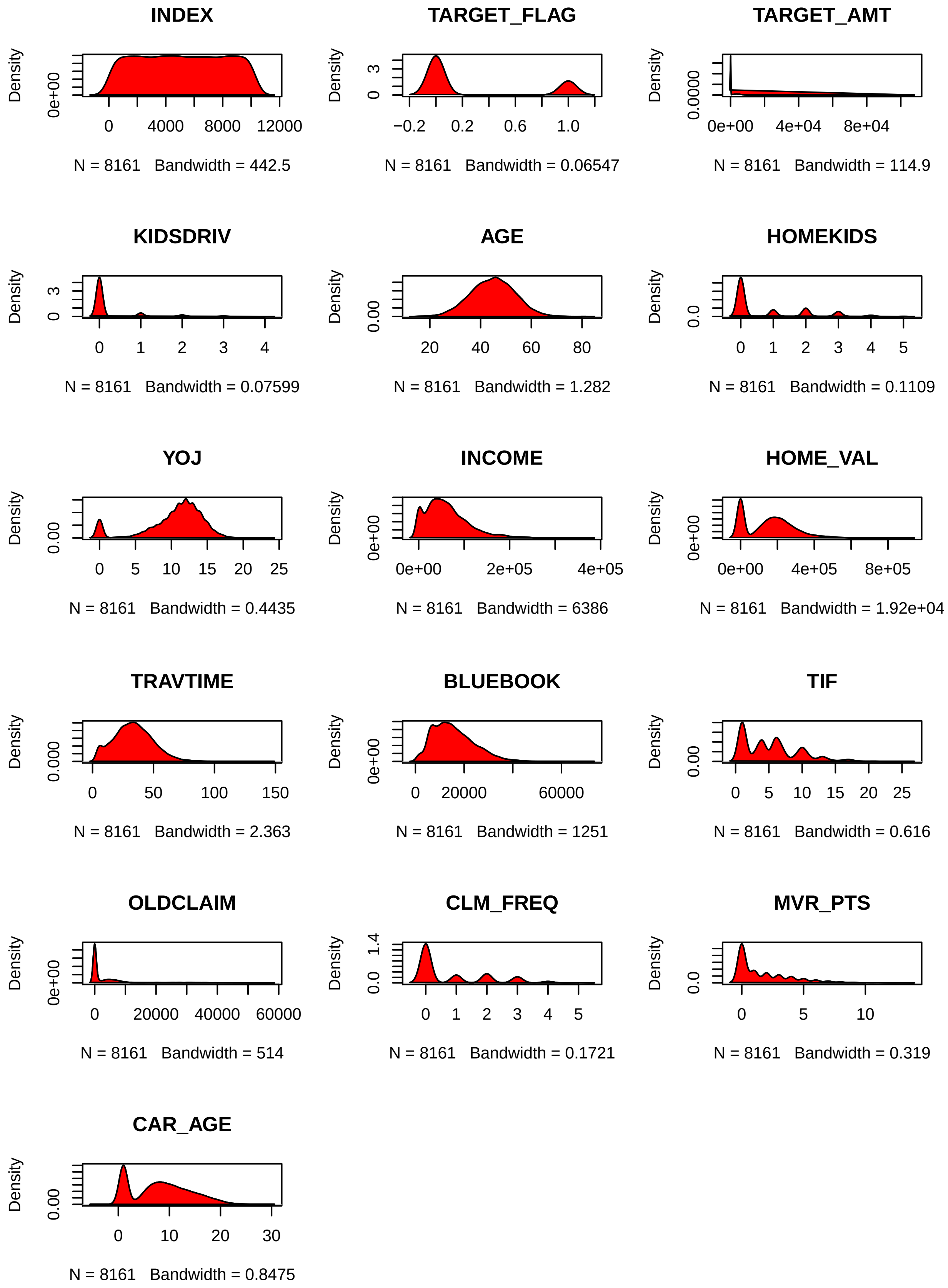


Training Data - Summary

Training Data - Histograms



Training Data - Box Plots

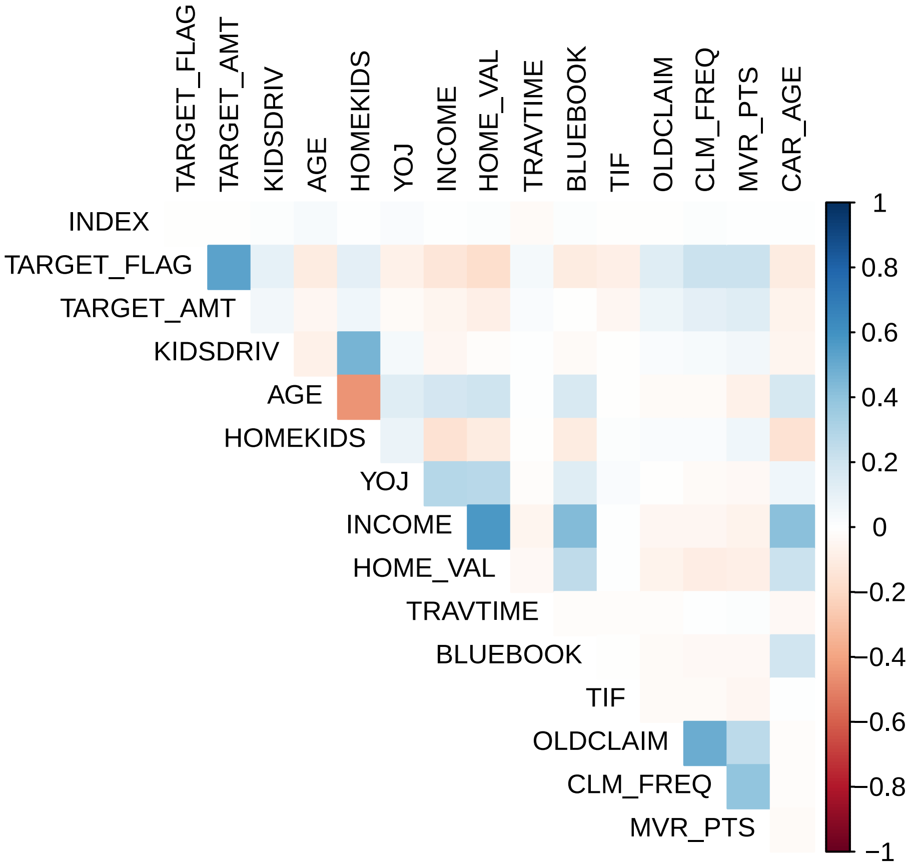


INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1
3	NA	NA	0	48	0	11	52881	No
9	NA	NA	1	40	1	11	50815	Yes
10	NA	NA	0	44	2	12	43486	Yes
18	NA	NA	0	35	2	NA	21204	Yes
21	NA	NA	0	59	0	12	87460	No
30	NA	NA	0	46	0	14	NA	No

Training Data - Skewness Report

```
##        INDEX  TARGET_FLAG   TARGET_AMT     KIDSDRIV          AGE     HOMEKIDS
##  0.002003877  1.071661372  8.706303371  3.351837433 -0.028603110  1.341127092
##          YOJ       INCOME     HOME_VAL     TRAVTIME     BLUEBOOK          TIF
## -1.205346319  1.192166999  0.498080519  0.446817389  0.794214109  0.890812001
##     OLDCLAIM     CLM_FREQ      MVR_PTS      CAR_AGE
##  3.119039986  1.208798507  1.347840258  0.277045900
```

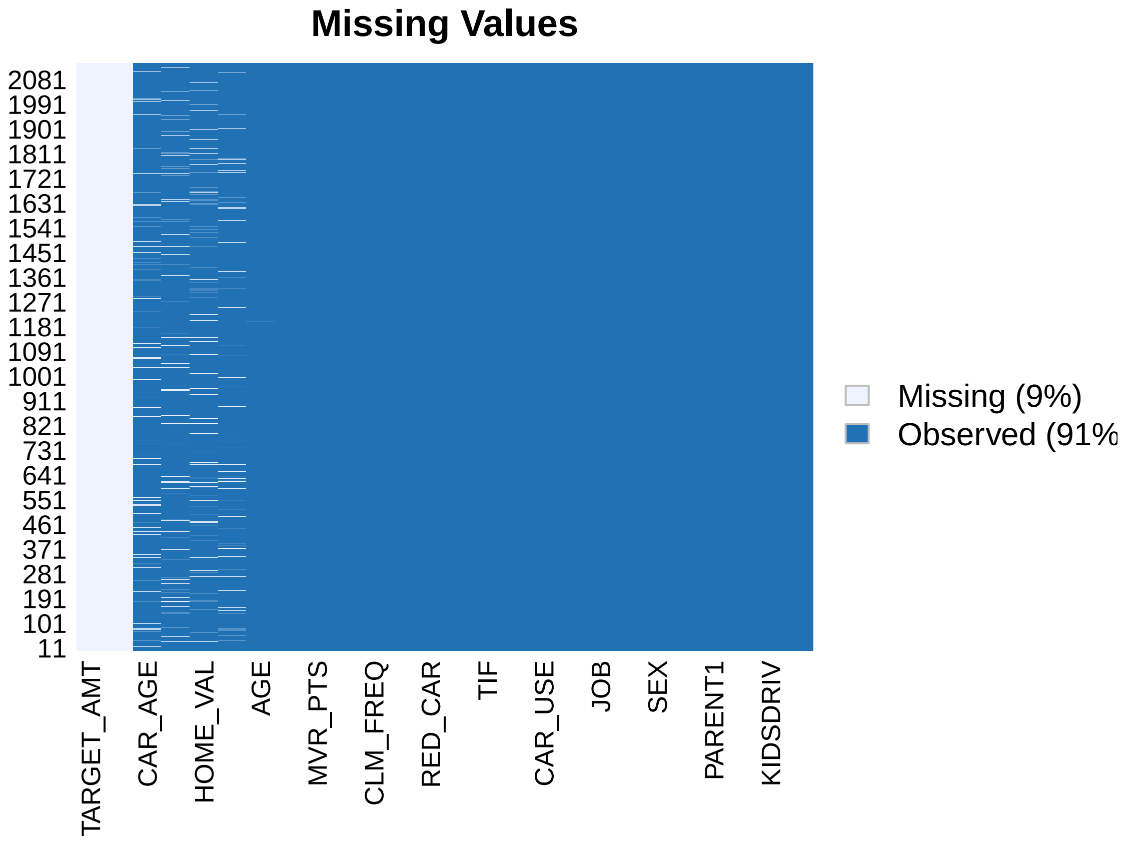
Training Data - Correlation Report



Evaluation Data - Fix Formatting

Evaluation Data - Missing Data Check

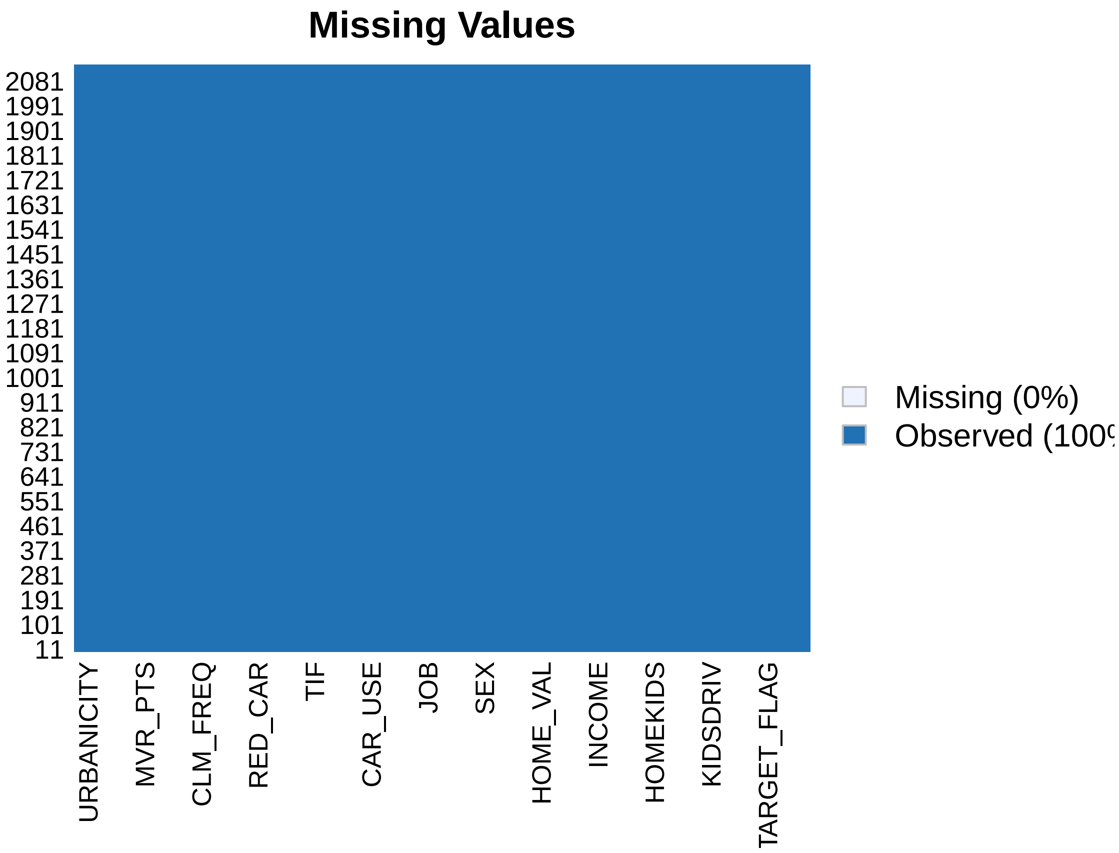
```
##        INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0        2141        2141           0           1           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##          94         125           0         111           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         129           0
```



Evaluation Data - Missing Data Re-test

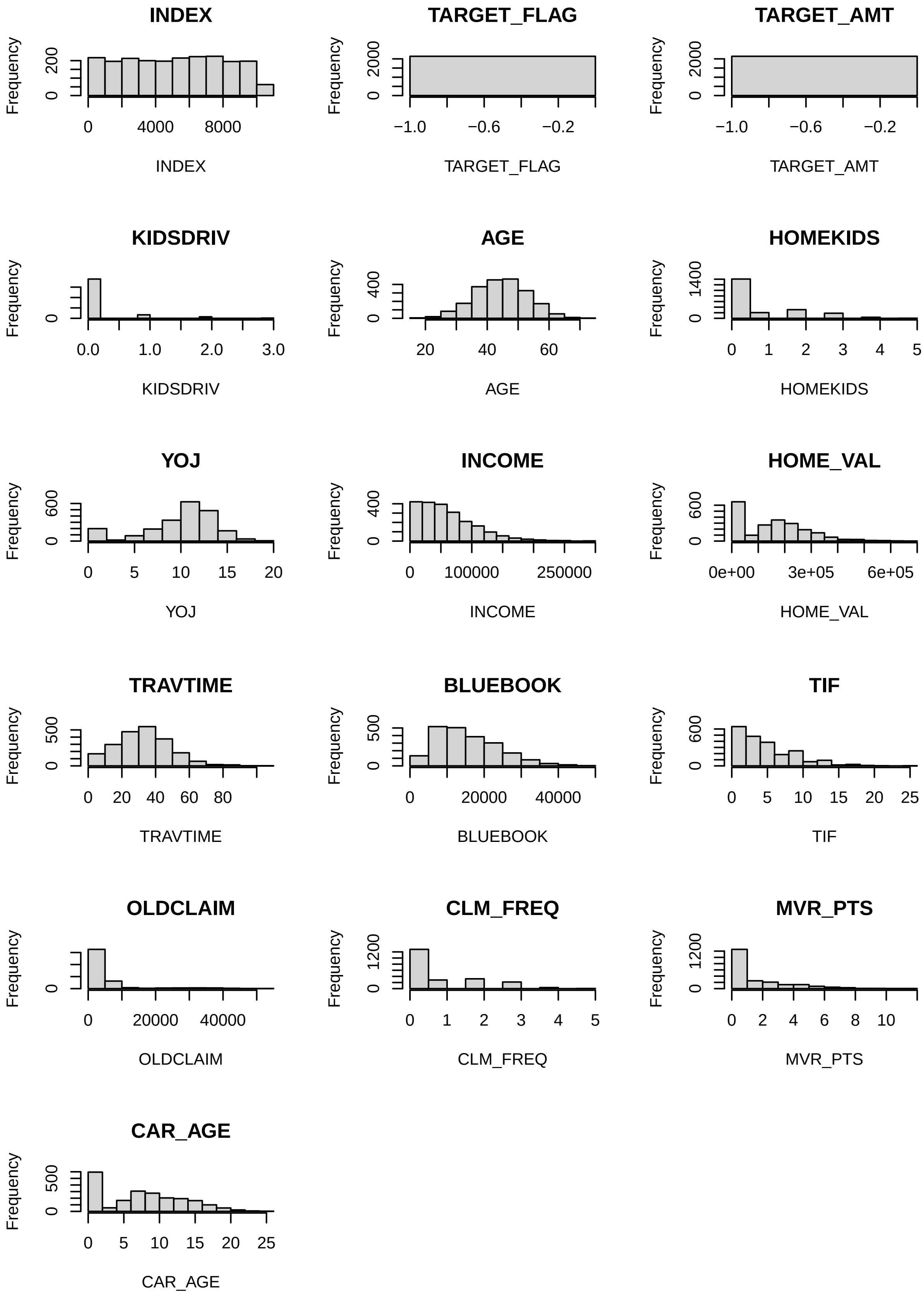
```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           0           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##           0           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##           0           0
```

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
	Min. : 3	Min. :0	Min. :0	Min. :0.0000	Min. :17.00	Min. :0.0000	Min. : 0.
	1st Qu.: 2632	1st Qu.:0	1st Qu.:0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.:
	Median : 5224	Median :0	Median :0	Median :0.0000	Median :45.00	Median :0.0000	Median :
	Mean : 5150	Mean :0	Mean :0	Mean :0.1625	Mean :45.01	Mean :0.7174	Mean :10
	3rd Qu.: 7669	3rd Qu.:0	3rd Qu.:0	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1
	Max. :10300	Max. :0	Max. :0	Max. :3.0000	Max. :73.00	Max. :5.0000	Max. :19

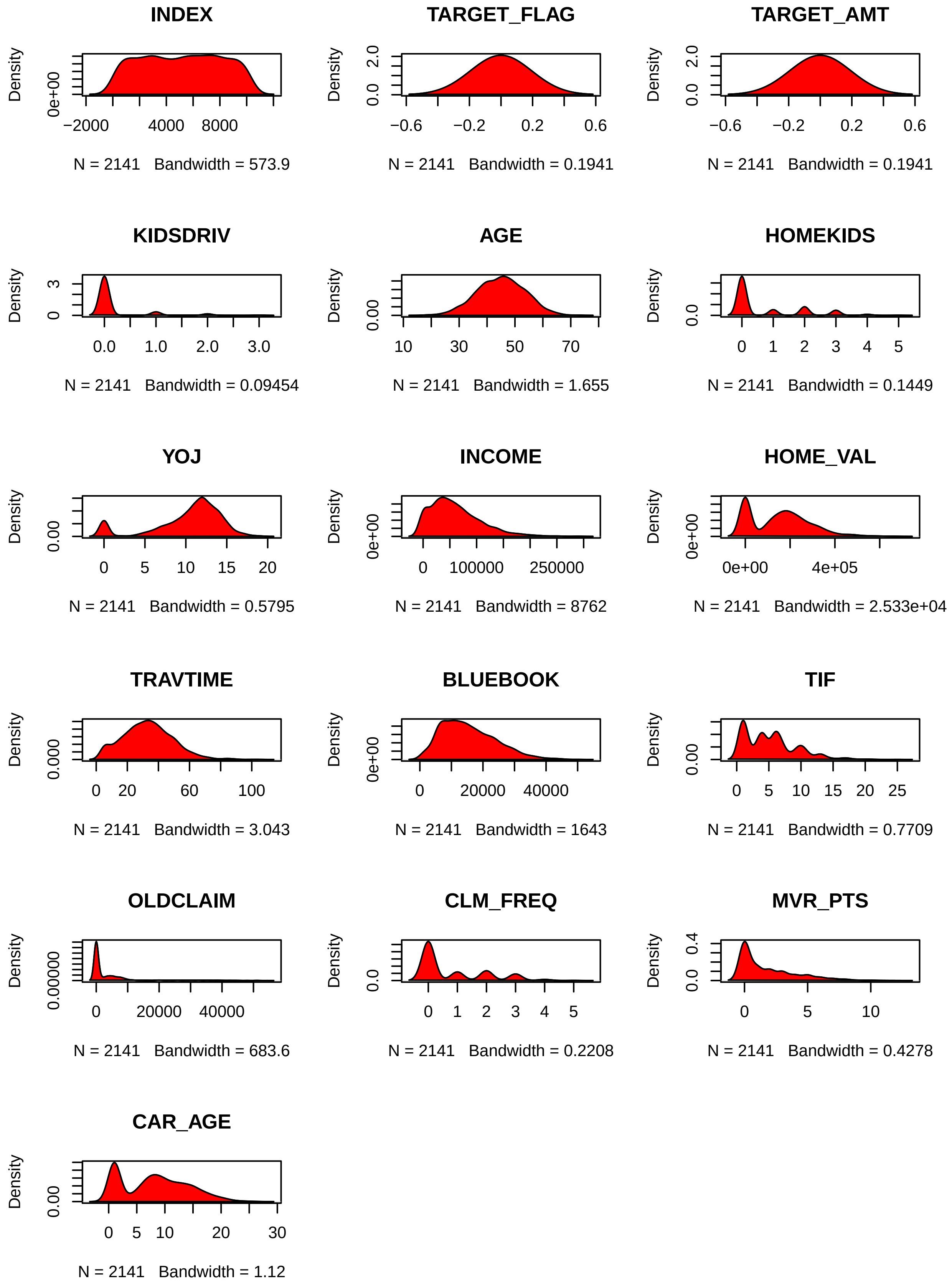


Evaluation Data - Summary

Evaluation Data - Histograms



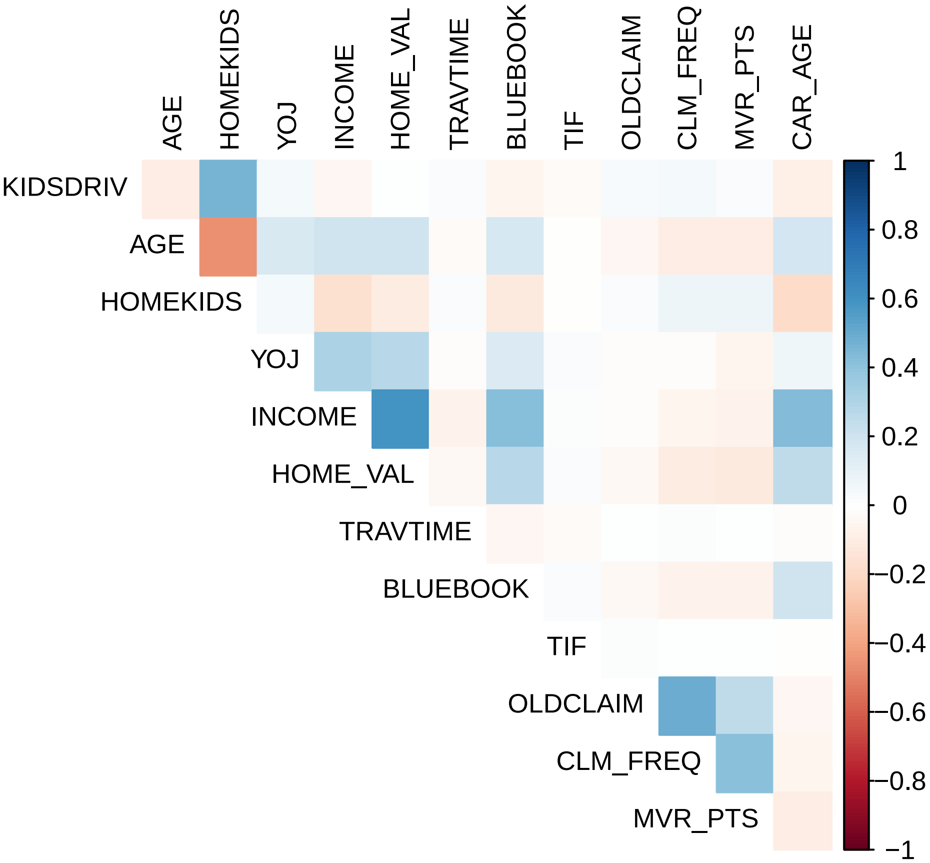
Evaluation Data - Box Plots



Evaluation Data - Skewness Report

```
##        INDEX  TARGET_FLAG   TARGET_AMT     KIDSDRIV          AGE     HOMEKIDS 
## -0.007836838          NaN          NaN  3.282091571 -0.054270193  1.316940480 
##          YOJ       INCOME     HOME_VAL     TRAVTIME     BLUEBOOK          TIF 
## -1.179883078  1.048837555  0.511585570  0.388959497  0.675922236  0.927584339 
##     OLDCLAIM     CLM_FREQ      MVR_PTS      CAR_AGE 
##  3.113317818  1.132576010  1.308706012  0.265656260
```

Evaluation Data - Correlation Report



Data Models

Model Preparation

The Training Insurance data is chosen and the train test split is created with 80% as factor. After the dataset split the plan is to create following models and predict evaluation dataset using the best model.

1. Logistic Regression Model 1 - > TARGET FLAG and TARGET AMOUNT
2. Logistic Regression Model 2 - > TARGET FLAG and Other Columns
3. Logistic Regression Model 3 - > Stepwise regression

Logistic Regression Model - 1

The model glm is similar to Generalized Linear Model but it has ability to find confidence set of models (best models) from the list of all possible models (candidate models). Models are fitted with the specified fitting function (glm) and are ranked with the criterion 'aic'

The model takes training dataset and linear regression is calculated for response variable (TARGET_FLAG) and other explanatory variables. Summary of the model is displayed on the output and AUC (Area under the curve) is calcluated

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial,
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3323  -0.7079  -0.3886   0.6132   3.2090
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.834e+00  3.252e-01  -8.715  < 2e-16 ***
## INDEX                         7.973e-06  1.102e-05   0.723 0.469533
## KIDSDRIV                      4.237e-01  6.836e-02   6.198 5.72e-10 ***
## AGE                          -1.073e-03  4.528e-03  -0.237 0.812654
## HOMEKIDS                      3.638e-02  4.177e-02   0.871 0.383806
## YOJ                          -2.012e-02  9.424e-03  -2.134 0.032808 *
## INCOME                       -3.168e-06  1.212e-06  -2.614 0.008937 **
## HOME_VAL                     -1.103e-06  3.774e-07  -2.924 0.003458 **
## TRAVTIME                      1.566e-02  2.122e-03   7.377 1.62e-13 ***
## BLUEBOOK                     -2.240e-05  5.912e-06  -3.789 0.000151 ***
## TIF                          -5.233e-02  8.274e-03  -6.325 2.54e-10 ***
## OLDCLAIM                     -1.354e-05  4.375e-06  -3.094 0.001972 **
## CLM_FREQ                      1.955e-01  3.201e-02   6.108 1.01e-09 ***
## MVR_PTS                       1.143e-01  1.528e-02   7.483 7.24e-14 ***
## CAR_AGE                      -1.343e-03  8.198e-03  -0.164 0.869870
## PARENT1Yes                    4.354e-01  1.232e-01   3.533 0.000411 ***
## MSTATUSYes                   -4.972e-01  9.321e-02  -5.335 9.58e-08 ***
## SEXM                          6.302e-02  1.269e-01   0.497 0.619424
## EDUCATIONBachelors           -4.135e-01  1.296e-01  -3.190 0.001425 **
## EDUCATIONHigh School          2.755e-02  1.072e-01   0.257 0.797140
## EDUCATIONMasters             -3.030e-01  1.990e-01  -1.523 0.127843
## EDUCATIONPhD                 -1.957e-01  2.386e-01  -0.820 0.412139
## JOBClerical                   1.497e-01  1.204e-01   1.243 0.213815
## JOBDoctor                    -8.363e-01  3.299e-01  -2.535 0.011244 *
## JOBHome Maker                -1.042e-01  1.707e-01  -0.611 0.541362
## JOBLawyer                    -1.334e-01  2.082e-01  -0.641 0.521807
## JOBManager                   -7.948e-01  1.563e-01  -5.084 3.71e-07 ***
## JOBProfessional              -9.675e-02  1.343e-01  -0.720 0.471404
## JOBStudent                   -3.649e-02  1.448e-01  -0.252 0.801026
## JOBUnknown                   -2.391e-01  2.064e-01  -1.158 0.246721
## CAR_USEPrivate               -7.056e-01  1.027e-01  -6.868 6.49e-12 ***
## CAR_TYPEPanel Truck           6.108e-01  1.832e-01   3.333 0.000858 ***
## CAR_TYPEPickup                5.812e-01  1.145e-01   5.078 3.81e-07 ***
## CAR_TYPESports Car             1.081e+00  1.466e-01   7.372 1.68e-13 ***
## CAR_TYPESUV                    8.507e-01  1.252e-01   6.797 1.07e-11 ***
## CAR_TYPEVan                    7.024e-01  1.434e-01   4.897 9.73e-07 ***
## RED_CARyes                     5.749e-04  9.805e-02   0.006 0.995322
## REVOKEDYes                     8.918e-01  1.017e-01   8.767  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.507e+00  1.316e-01  19.055  < 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: 7517.2  on 6528  degrees of freedom
## Residual deviance: 5786.6  on 6490  degrees of freedom
## AIC: 5864.6
##
## Number of Fisher Scoring iterations: 5
```

AIC of the Model 1 is 5865.2

Logistic Regression Model - 1 Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  189   91
##          0  250 1102
##
##                Accuracy : 0.7911
##                  95% CI : (0.7705, 0.8105)
##     No Information Rate : 0.731
##     P-Value [Acc > NIR] : 1.182e-08
##
##                   Kappa : 0.4
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4305
##             Specificity : 0.9237
##          Pos Pred Value : 0.6750
##          Neg Pred Value : 0.8151
##              Prevalence : 0.2690
##          Detection Rate : 0.1158
##    Detection Prevalence : 0.1716
##       Balanced Accuracy : 0.6771
##
##        'Positive' Class : 1
##
```

Accuracy of the Model 1 is 79.4%

Logistic Regression Model - 2

The model glm is similar to Generalized Linear Model but it has ability to find confidence set of models (best models) from the list of all possible models (candidate models). Models are fitted with the specified fitting function (glm) and are ranked with the criterion ‘aic’

The model takes training dataset and linear regression is calculated for response variable (TARGET_FLAG) and other explanatory variables. Summary of the model is displayed on the output and AUC (Area under the curve) is calcluated

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, family = binomial, data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2887  -0.7200  -0.4015   0.6215   3.2046
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.762e+00  2.196e-01 -12.577  < 2e-16 ***
## KIDSDRIV                        4.106e-01  6.646e-02   6.178 6.48e-10 ***
## HOMEKIDS                        3.818e-02  3.761e-02   1.015 0.310005
## INCOME                         -4.453e-06  1.100e-06  -4.047 5.18e-05 ***
## PARENT1Yes                      4.400e-01  1.217e-01   3.615 0.000300 ***
## HOME_VAL                       -1.157e-06  3.655e-07  -3.165 0.001550 **
## MSTATUSYes                     -5.069e-01  9.205e-02  -5.507 3.65e-08 ***
## EDUCATIONBachelors             -5.835e-01  1.174e-01  -4.970 6.71e-07 ***
## EDUCATIONHigh School           -5.133e-02  1.037e-01  -0.495 0.620698
## EDUCATIONMasters               -5.450e-01  1.475e-01  -3.696 0.000219 ***
## EDUCATIONPhD                   -5.961e-01  1.838e-01  -3.243 0.001181 **
## TRAVTIME                        1.605e-02  2.103e-03   7.633 2.30e-14 ***
## CAR_USEPrivate                 -7.834e-01  8.217e-02  -9.534  < 2e-16 ***
## BLUEBOOK                       -2.482e-05  5.277e-06  -4.704 2.55e-06 ***
## TIF                            -5.232e-02  8.216e-03  -6.367 1.92e-10 ***
## CAR_TYPEPanel Truck             6.104e-01  1.613e-01   3.785 0.000154 ***
## CAR_TYPEPickup                  5.368e-01  1.112e-01   4.826 1.39e-06 ***
## CAR_TYPESports Car              1.008e+00  1.198e-01   8.414  < 2e-16 ***
## CAR_TYPESUV                     8.042e-01  9.580e-02   8.395  < 2e-16 ***
## CAR_TYPEVan                     6.891e-01  1.353e-01   5.095 3.48e-07 ***
## CLM_FREQ                        1.472e-01  2.846e-02   5.172 2.32e-07 ***
## REVOKEDYes                      7.498e-01  8.898e-02   8.427  < 2e-16 ***
## MVR_PTS                         1.148e-01  1.510e-02   7.600 2.95e-14 ***
## CAR_AGE                        -1.031e-04  8.130e-03  -0.013 0.989877
## URBANICITYHighly Urban/ Urban  2.437e+00  1.308e-01  18.638  < 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: 7517.2  on 6528  degrees of freedom
## Residual deviance: 5849.8  on 6504  degrees of freedom
## AIC: 5899.8
##
## Number of Fisher Scoring iterations: 5
```

AIC of the Model 2 is 5900

Logistic Regression Model - 2 Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  177   96
##          0  262 1097
##
##                Accuracy : 0.7806
##                  95% CI : (0.7598, 0.8005)
##     No Information Rate : 0.731
##     P-Value [Acc > NIR] : 2.263e-06
##
##                   Kappa : 0.3665
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4032
##             Specificity : 0.9195
##          Pos Pred Value : 0.6484
##          Neg Pred Value : 0.8072
##              Prevalence : 0.2690
##          Detection Rate : 0.1085
##    Detection Prevalence : 0.1673
##       Balanced Accuracy : 0.6614
##
##        'Positive' Class : 1
##
```

Accuracy of the Model 2 is 78.4%

Logistic Regression Model - 3

The stepwise regression takes the predictors and adds/removes based on the significance of the predictors. At first the model is run with 0 predictors and the predictors are added in sequence based on its significance. Since the model chooses the predictors by itself all predictors (explanator variables) are considered for model against target variable.

Adding to the stepwise regression we are also considering the transformed dataset with new variables derived from the existing variables.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF +
##     CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY, family = binomial,
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3110  -0.7223  -0.4028   0.6259   3.1979
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.747e+00  2.167e-01 -12.678  < 2e-16 ***
## KIDSDRIV                       4.380e-01  6.084e-02   7.198 6.09e-13 ***
## INCOME                        -4.454e-06  1.098e-06  -4.057 4.98e-05 ***
## PARENT1Yes                     5.025e-01  1.050e-01   4.787 1.69e-06 ***
## HOME_VAL                      -1.181e-06  3.647e-07  -3.238 0.001205 **
## MSTATUSYes                    -4.795e-01  8.794e-02  -5.453 4.96e-08 ***
## EDUCATIONBachelors            -5.925e-01  1.096e-01  -5.405 6.49e-08 ***
## EDUCATIONHigh School          -5.656e-02  1.034e-01  -0.547 0.584248
## EDUCATIONMasters              -5.589e-01  1.221e-01  -4.578 4.69e-06 ***
## EDUCATIONPhD                  -6.092e-01  1.656e-01  -3.680 0.000233 ***
## TRAVTIME                       1.600e-02  2.102e-03   7.611 2.72e-14 ***
## CAR_USEPrivate                -7.851e-01  8.215e-02  -9.556  < 2e-16 ***
## BLUEBOOK                      -2.502e-05  5.275e-06  -4.743 2.11e-06 ***
## TIF                           -5.206e-02  8.212e-03  -6.340 2.30e-10 ***
## CAR_TYPEPanel Truck            6.092e-01  1.612e-01   3.780 0.000157 ***
## CAR_TYPEPickup                 5.331e-01  1.111e-01   4.796 1.62e-06 ***
## CAR_TYPESports Car             1.009e+00  1.198e-01   8.424  < 2e-16 ***
## CAR_TYPESUV                    8.058e-01  9.577e-02   8.414  < 2e-16 ***
## CAR_TYPEVan                    6.879e-01  1.352e-01   5.089 3.59e-07 ***
## CLM_FREQ                       1.473e-01  2.845e-02   5.179 2.23e-07 ***
## REVOKEDYes                     7.520e-01  8.896e-02   8.454  < 2e-16 ***
## MVR_PTS                        1.151e-01  1.510e-02   7.622 2.50e-14 ***
## URBANICITYHighly Urban/ Urban  2.436e+00  1.308e-01  18.628  < 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: 7517.2  on 6528  degrees of freedom
## Residual deviance: 5850.9  on 6506  degrees of freedom
## AIC: 5896.9
##
## Number of Fisher Scoring iterations: 5
```

AIC of the Model 3 is 5897.4

Logistic Regression Model - 3 Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL
3	0	0	0	48	0	11	52881	0
9	0	0	1	40	1	11	50815	0
10	0	0	0	44	2	12	43486	0
18	0	0	0	35	2	6	21204	0
21	0	0	0	59	0	12	87460	0
30	0	0	0	46	0	14	90213	207519

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  180   97
##          0  259 1096
##
##                Accuracy : 0.7819
##                  95% CI : (0.761, 0.8017)
##     No Information Rate : 0.731
##     P-Value [Acc > NIR] : 1.286e-06
##
##                   Kappa : 0.3721
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4100
##             Specificity : 0.9187
##          Pos Pred Value : 0.6498
##          Neg Pred Value : 0.8089
##              Prevalence : 0.2690
##          Detection Rate : 0.1103
##    Detection Prevalence : 0.1697
##       Balanced Accuracy : 0.6644
##
##        'Positive' Class : 1
##
```

Accuracy of the Model 3 is 78.3%

Model Selection

While comparing three models the best performing model is Logistic Regression Model 3 with stepwise regression. The below parameters are considered for choosing the model 3 as best model.

1. AIC value Based on the AIC value we can say Model 2 is performing better.
2. AUC Based on the AUC value we can say Model 2 is performing better.
3. Accuracy Based on the Accuracy value we can say Model 3 is performing better.

Evaluation Data Prediction

Conclusion and Output

```
## NULL
```

Overall we found that Model 2 (Logistic Regression with all explanatory variables) performs better in predicting the TARGET FLAG and TARGET AMOUNT for the evaluation data set.