Data 621 - Homework 4

Group 4 Layla Quinones, Ian Costello, Dmitriy Burtsev & Esteban Aramayo

11/21/2021

Overview

In this homework assignment, we explore, analyze and model a data set containing approximately 8,000 records representing customers 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.

Our objective is to build a multiple linear regression model and a binary logistic regression model 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.

Libraries Used

We use the standard libraries such as `tidyverse`, `ggplot2`, and `caret`. We also make use of the `pROC` package to construct the curves.

Data Exloration

As usual, our data are stored on GitHub at our team's main repository for easy access across team members. With our initial glimpse of the data, we noticed that there was variables that should be doubles that are classed as character strings. These seem to be on values with currency, such as income, cars values, and home values. We'll have to do some data cleaning before we continue with the exploration.

There are 8,161 observations in this data set and 26 columns. We know that `INDEX`, `TARGET_FLAG` and `TARGET_AMT` are not predictor variables. This gives us **8,161 observations** with **23 predictors** that are a combination of int, double and character data types. We also see that the character variables will have to converted to factors in order for us to explore their distributions. Variables such and `INCOME`, `HOME_VAL`, `BLUEBOOK`, `OLDCLAIM` will be converted to numeric because they are numbers with values that have meaning in their hierarchy.

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
```

```
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

Missing Values

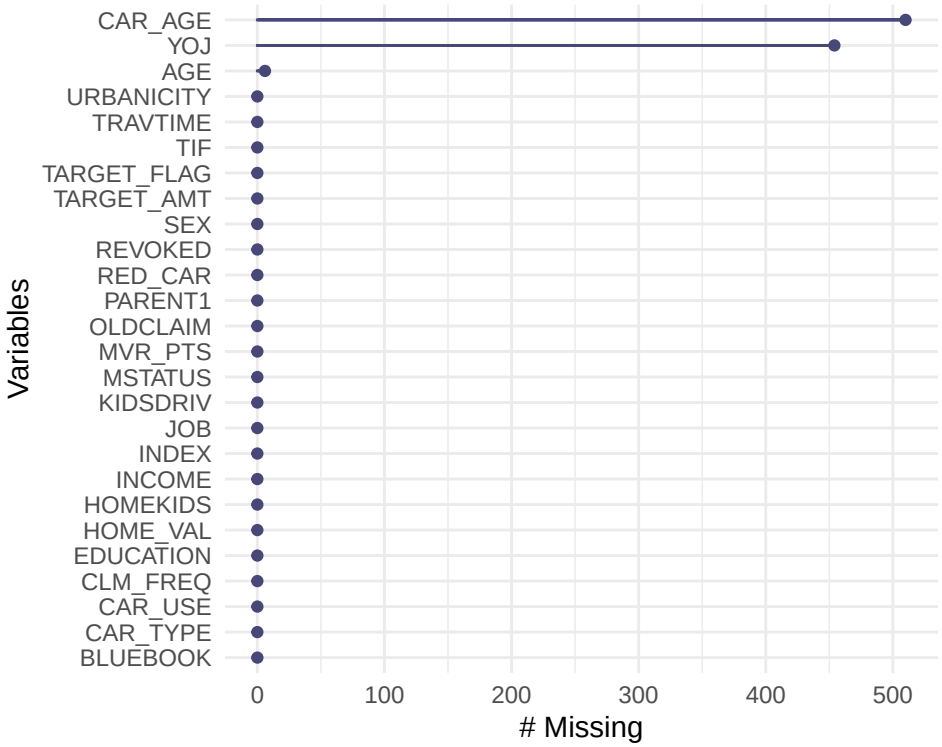

Figure 1. Plot of missing values

As shown in figure 1, there are missing variables in the columns `Car_AGE`, `AGE` and `YOJ`. None of these exceed the 10% missing data, so we will continue with all variables for now and not dropping any of them.

DATA CLEANING - CONVERTING DATA TYPES

Before getting too far ahead, using regular expressions, we'll have to do some standard cleaning to remove the `$`, `z_`, and `,` and put in a different variable name from numeric strings. We'll also change all other character variables into factors. We'll glimpse the data again to confirm these changes.

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
```

```
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0,~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
## $ EDUCATION   <fct> PhD, High School, High School, <High School, PhD, Bachelor~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

Summary Statistics

Displaying summary statistics again to confirm data cleaning and get a sense of the spread of the data elements.

Feature Histograms

For each of the variables, these histograms in figure 2 (Code Appendix 2.6) provide a nice overview of each feature, its variation, and paths for potential transformations later on for model construction. Histograms are a quick way to see the shape of the distributions for each feature. Of note is the only normally distributed variable, age. The other features appear skewed to some degree. We can also begin to see the affect of outliers that we'll have to account for later on.

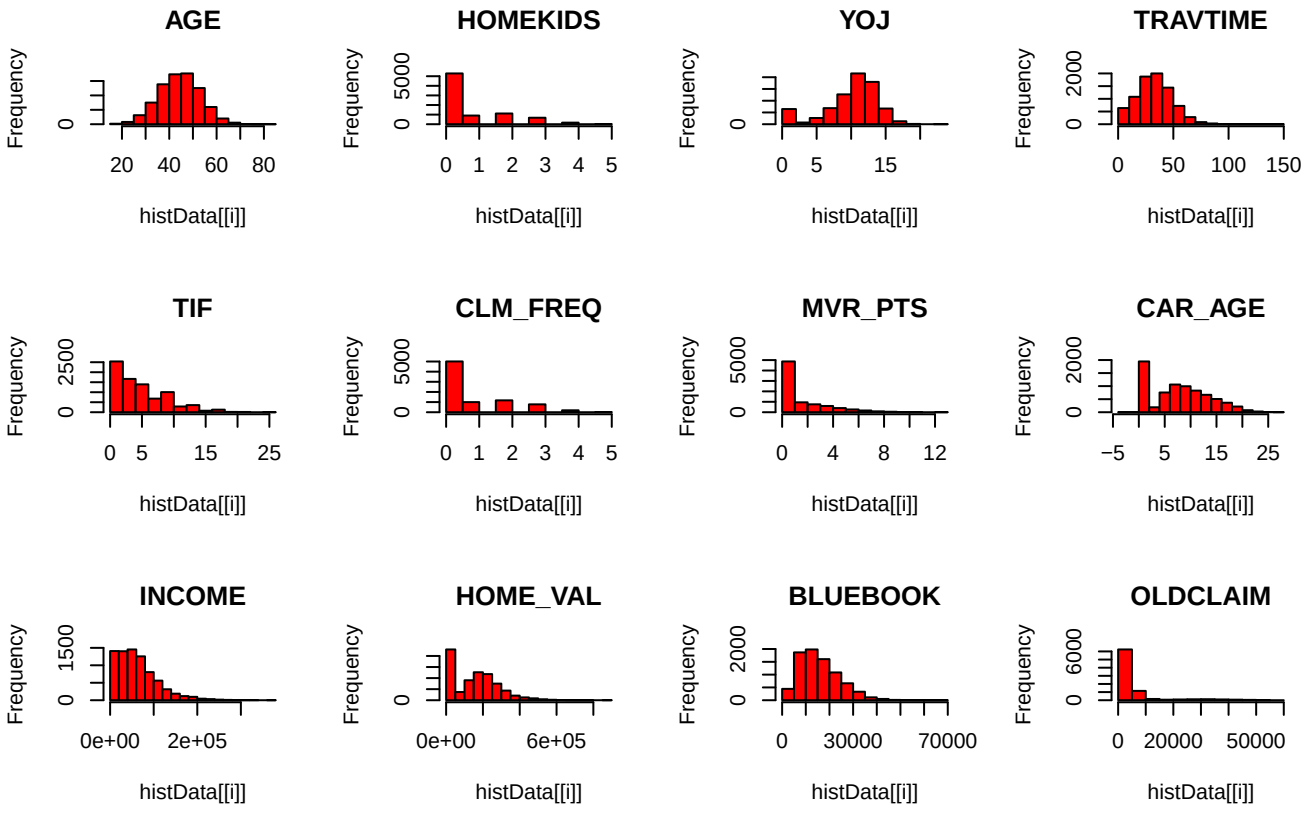

Figure 2. Feature histograms

Outlier Analysis

Feature Box Plots

Let's identify the variables with outlier values using boxplots. From these initial box plots we can see that there are some outliers. In particular, `TRAVTIME`, `INCOME`, and `HOME_VAL` have many outliers which are spread out more compared to the other variables.

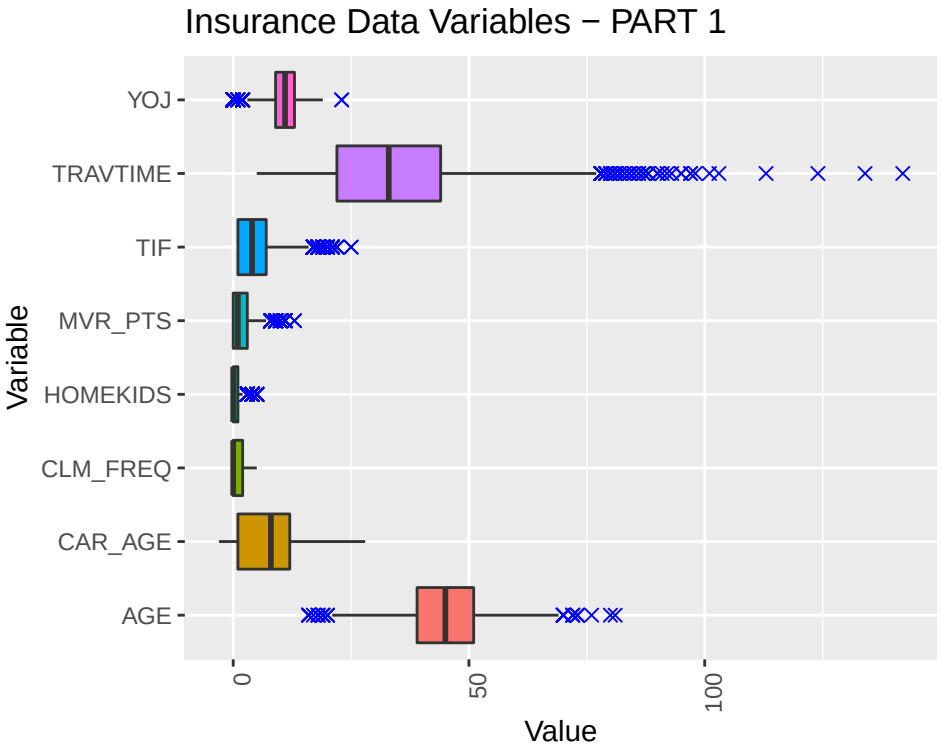

Figure 3. Feature box plot, part 1*

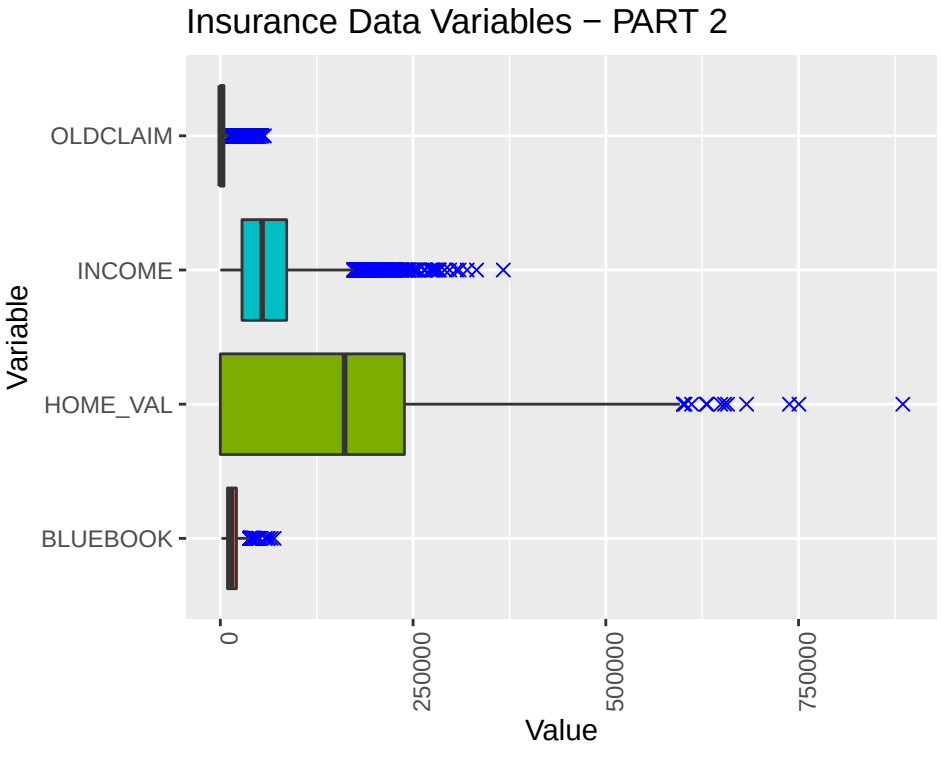

Figure 4. Feature box plot, part 2

Categorical Predictors - with target variable

The figures below show each categorical feature and the prevalence of each factor. The balance and imbalance of factors may affect how models are built and function in the later part of this project. In reviewing each factor, the imbalance may lead us to select or exclude features.

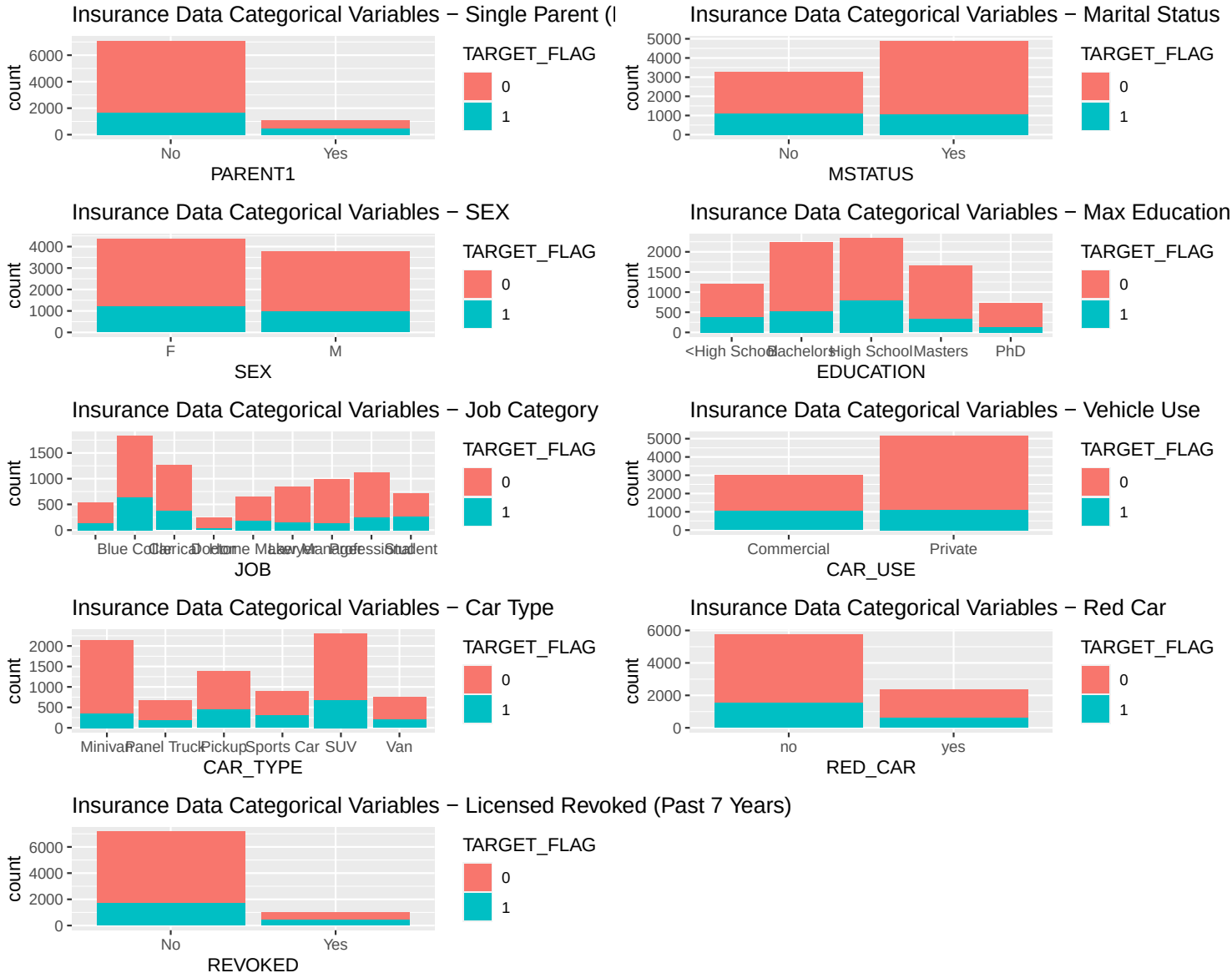

Figure 5. Categorical variables showing balance/imbalance

Numeric Data - Relationship to Target

These plot visualizations (Figure 6) **(Code Appendix)** gives us an idea of the outliers we have in each variable, but does not give us a good sense of the distribution. We can use the histograms (Figure 2) above to interpret shape. If the notches of two boxes do not overlap, then this suggests that the medians are significantly different.

For the features we see some outliers, we can decide to either throw out that variable out altogether and not consider it in our models or impute the outliers with median values. Before deciding on a course of action, we'll look at a few other things.

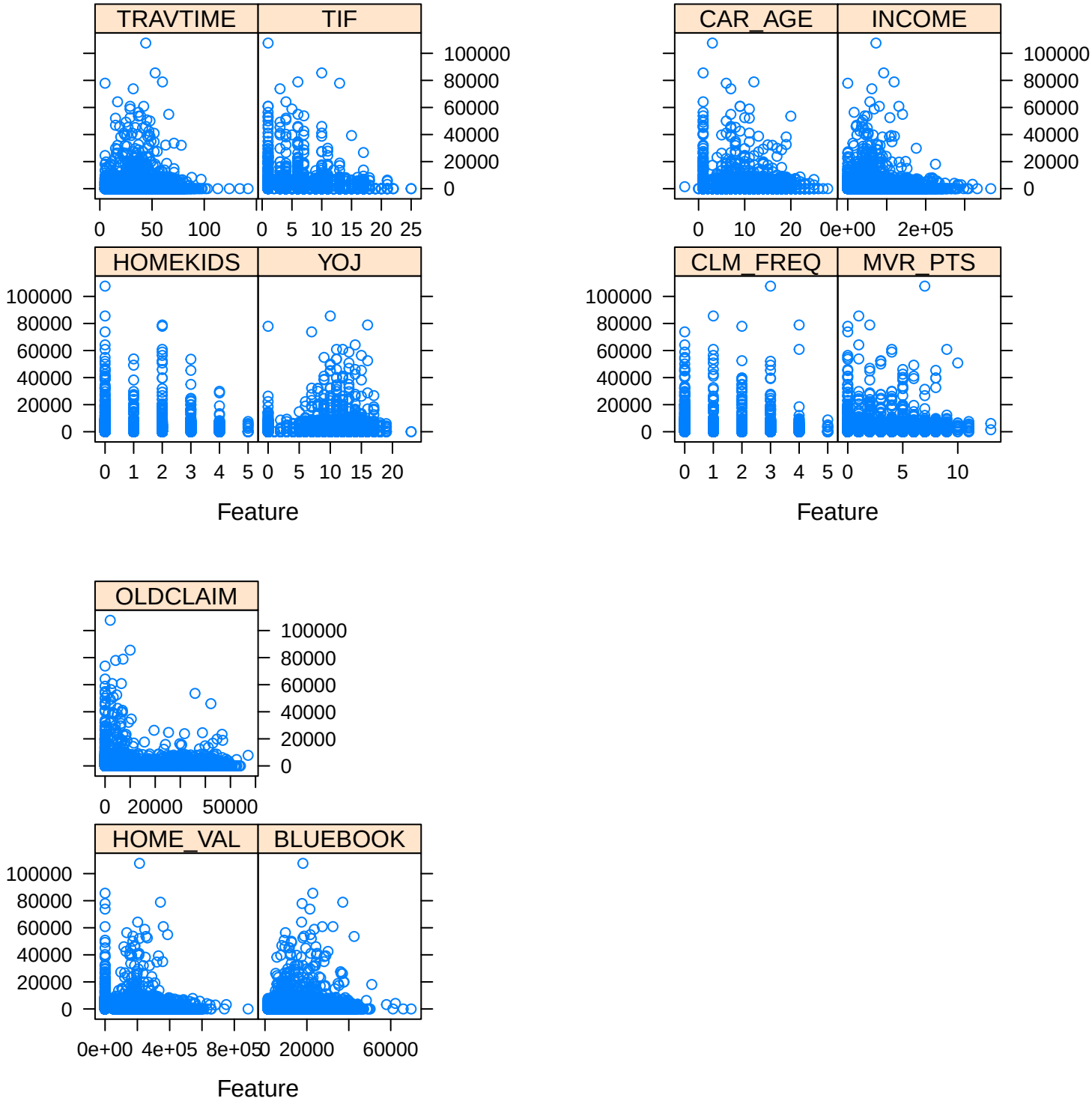

Figure 6. Feature plots

Correlation

Let's use a heat map to see the level of correlation of the numeric predictor variables. From the correlation matrix (figure 7) below and from the `findCorrelation()` function, there does not appear to be any multiple colinearity we have to account for.

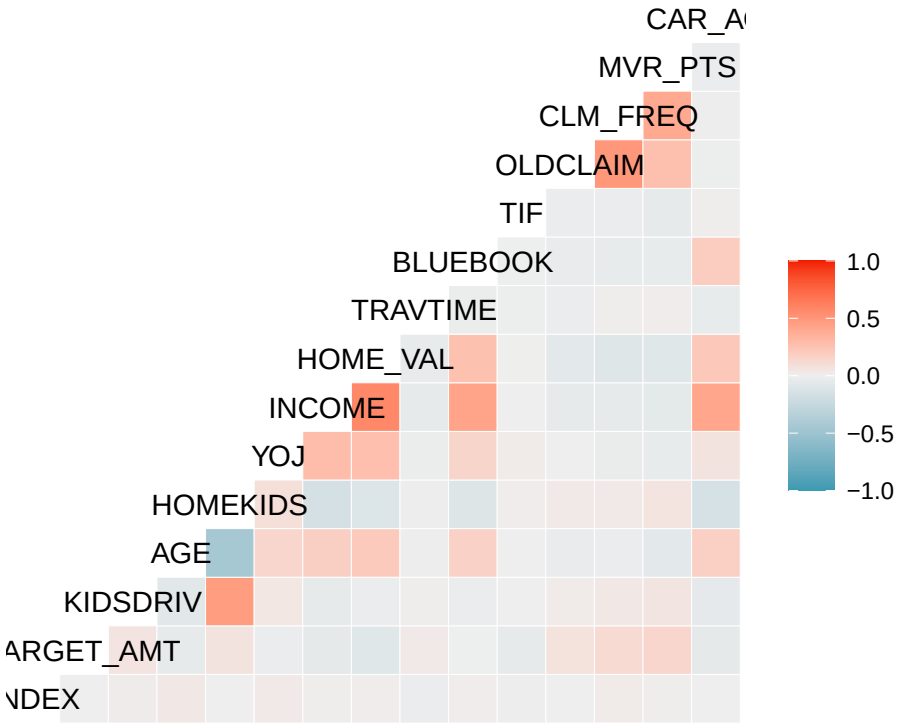

Figure 7. Correlation matrix

```
## All correlations <= 0.75
```

```
## character(0)
```

Data Preparation

Impute Missing Values

For the missing values in `Age` and `YOJ` we will impute with the mean, and for `HOME_VAL`, `INCOME`, and `CAR_AGE` we will use the median. We determine this by observing their distribution and if there is more skew we'll take the median, more normal and we'll use mean.

Variable Importance

To determine the variable importance the following steps were taken:

- A training data frame `prepTrainA` was prepared for the `TARGET_FLAG` response variable and its associated predictor variables.
- A training data frame `prepTrainB` was prepared for the `TARGET_AMT` response variable and its associated predictor variables.
- Using the `prepTrainA` data frame, a classification model `modelA` was trained using the `Learning Vector Quantization (lvq)` method. From it, the variable importance was summarized and plotted.

```
## ROC curve variable importance
##
##            Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
```

```
## HOME_VAL       0.6176
## URBANICITY     0.6026
## INCOME         0.5961
## CAR_USE        0.5782
## MSTATUS        0.5751
## BLUEBOOK       0.5750
## HOMEKIDS       0.5706
## AGE            0.5686
## CAR_AGE        0.5640
## CAR_TYPE       0.5632
## PARENT1        0.5605
## REVOKED        0.5565
## TIF            0.5543
## EDUCATION      0.5424
## JOB            0.5414
## KIDSDRIV       0.5387
## TRAVTIME       0.5371
## YOJ            0.5362
## SEX            0.5119
## RED_CAR        0.5036
```

Table 1. Feature importance - part A

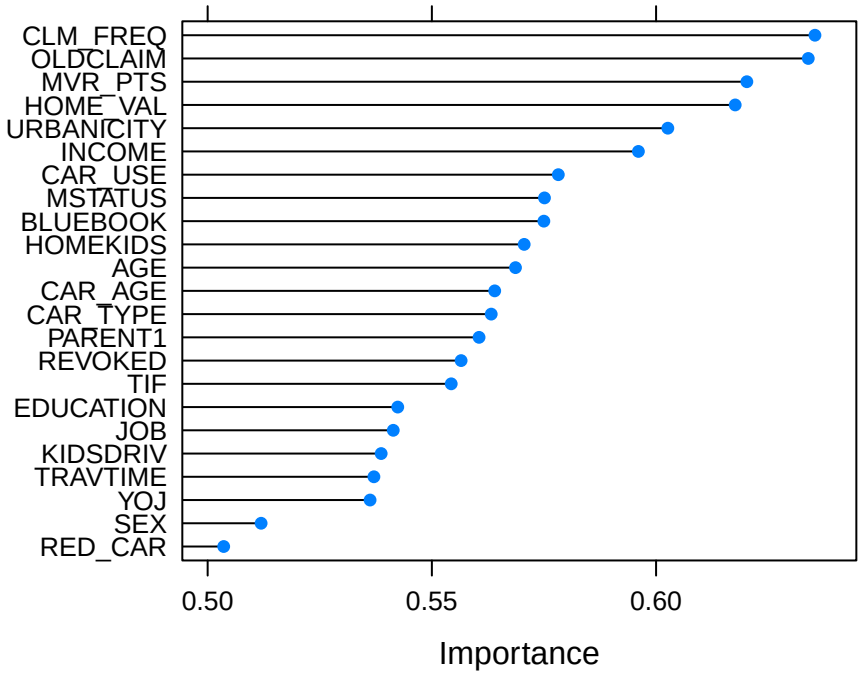

Figure 8. Plot of feature importance - part A

According to the plots above (figure 8), we can predict which variables would contribute best to the categorical predictions for `TARGET_FLAG`. We can use this to inform our data transformations.

Using the `prepTrainB` data frame, a classification/regression model `modelB` was trained using the `Generalized Linear Model (glm)` method. From it, the variable importance was summarized and plotted.

```
## glm variable importance
##
##   only 23 most important variables shown (out of 37)
##
##                                 Overall
```

```
## `URBANICITYHighly Urban/ Urban`  11.944
## MVR_PTS                           6.764
## CAR_USEPrivate                    4.741
## `CAR_TYPESports Car`              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
## KIDSDRIV                          2.776
## CLM_FREQ                          2.574
## INCOME                            2.441
## CAR_TYPEVan                       2.413
## CAR_TYPEPickup                    2.200
## CAR_AGE                           2.096
## SEXM                              2.007
## BLUEBOOK                          1.663
## JOBManager                        1.660
## `JOBBlue Collar`                  1.578
## JOBClerical                       1.550
## JOBProfessional                   1.478
## OLDCLAIM                          1.420
```

Table 2. Feature importance - part B

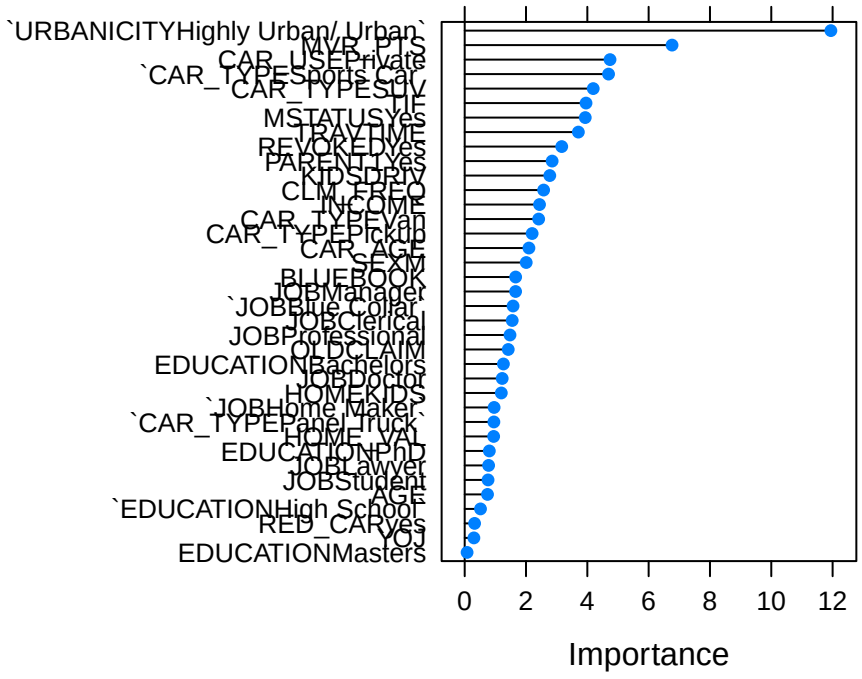

Figure 9. Plot of feature importance - part B

According to the plots above, we can predict which variables would contribute best to the numerical predictions for `TARGET_AMT`. We can use this to inform our data transformations.

Train Test Split

We partition the training data and set a seed in two data sets. One to be used for training purposes and one for validation/testing purposes.

Model Building

Binary Logistic Regression for dependent variable TARGET_FLAG

Binary Logistic Regression Model 1

For this model, we only include the predictor variables that have `theoretical effect on probability of collition`, which was provided as part of the definition of the variables.

Additionally, we remove the variables that were deemed as "urban legends," such as `RED_CAR` and `SEX`. From our importance variable model `importanceA`, we know that the variables `RED_CAR` and `SEX` ranked in the bottom 2 items of the importance list of 23 items. Hence, we don't include them. We also exclude variables having a theoretical "unknown effect" on probability of collision, such as `EDUCATION`.

$AIC = 6570.6$

Binary Logistic Regression Model 2

In order to improve on our first model, we use all the variables from Model 1, but we exclude the variables `YOJ`, which proved to be the least statistically significant for our Model 1.

Additionally, we include the variables `OLDCLAIM` and `URBANICITY`, which ranked 4th and 5th in our list of 23 predictor variable importance model `importanceA`.

$AIC = 6053.2$

```
## ROC curve variable importance
##
##   only 5 most important variables shown (out of 23)
##
##            Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
## HOME_VAL       0.6176
## URBANICITY     0.6026
```

Table 3. Feature importance - Binary model 2

We can see a significant improvement on the `residual deviance` and `AIC` values.

Binary Logistic Regression Model 3

In order to improve even more on our previous model, we add the variables `BLUEBOOK` and `HOMEKIDS`, which ranked 9th and 10th in our list of 23 predictor variable importance model `importanceA`.

At this point, the top 10 most statistically important of our set of 23 predictor variables are included in this model. This time, we can see an even more significant improvement on the `residual deviance` and `AIC` values.

$AIC = 6015$

```
## ROC curve variable importance
##
##   only 10 most important variables shown (out of 23)
##
##            Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
## HOME_VAL       0.6176
## URBANICITY     0.6026
```

```
## INCOME         0.5961
## CAR_USE        0.5782
## MSTATUS        0.5751
## BLUEBOOK       0.5750
## HOMEKIDS       0.5706
```

Table 4. Feature importance - Binary model 3

Binary Logistic Regression Model 4

Taking what we've learned from models 1-3, we add the variables `CAR_AGE`, `PARENT1` and `EDUCATION`, which ranked 12th, 14th and 17th in our list of 23 predictor variable importance model `importanceA`,

We also remove the variables `AGE` and `HOMEKIDS`, which from the previous models do not seem to contribute much, i.e., do not seem to be statistically significant for most of the models.

$AIC = 5897.7$

At this point, we can see most significant improvement on the `residual deviance` and `AIC` values.

Binary Logistic Regression Model 5

Just out of curiosity, what if we ignored all the statistical correlation and variable importance that we used for the previous four models. We use a model that includes all the predictor variables and the response variable `TARGET_FLAG`.

The results above show the best improvement so far.

Even after seeing the most significant improvement of all models, we still see that variables `AGE`, `HOMEKIDS`, `SEX`, and `RED_CAR (yes)` are not statistically significant. Which, lead us to believe that it might be true that deeming the variables `RED_CAR` and `SEX` as "urban legends" might be just urban legends. Those variable show little to no correlation to the probability of collision.

The variable `EDUCATION` seems to be statistically significant. At least for the values "Bachelors" and "Masters" we see that, based on the sign of their coefficients, they have a negative correlation to the theoretical probability of collision. So, it appears that people with higher education tend to have fewer accidents.

$AIC = 5903.2$

Linear Regression Models for dependent variable TARGET_AMT

Linear Regression Model 1

For our first model, we only include the predictor variables that have `theoretical probably of effecting the payout if there is a crash`, which was provided as part of the definition of the variables.

From the summary results we can see that we obtained low values for

Multiple R-squared: 0.01945 and **Adjusted R-squared: 0.01809**

Which, shows that using only predictor variables that have `theoretical probably of effecting the payout if there is a crash` is not a good way to go, for those variables do not seem to be enough to provide statistically significant results.

Linear Regression Model 2

For our second model, we only include the top 10 most important predictor variables that we gathered from our variable importance trained model `modelB`.

```
## glm variable importance
##
##   only 10 most important variables shown (out of 37)
##
```

```
##                                 Overall
## `URBANICITYHighly Urban/ Urban`  11.944
## MVR_PTS                           6.764
## CAR_USEPrivate                    4.741
## `CAR_TYPESports Car`              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
```

Table 5. Top 10 predictor variables from importance model B

Below are the results of applying our linear model 2 of TARGET_AMT vs Top 10 predictor variables.

From the summary results we can see that we obtained much better values for

Multiple R-squared: 0.05666, **Adjusted R-squared: 0.05478**

Linear Regression Model 3

We begin with a `baseline` model that includes all the predictor variables from Model 2 and the response variable `TARGET_AMT`. We remove the variables `CLM_FREQ` because its Pr value is 0.157159, which exceeds our requested 0.05 threshold. We will also add the next 6 variables from our variable importance model `modelB`. The added variables are: `KIDSDRIV`, `CLM_FREQ`, `INCOME`, `CAR_AGE`, `SEX`, and `BLUEBOOK`.

From the summary results above, we can see that the added variables have helped improve the values of our key indicators

Multiple R-squared: 0.06527, **Adjusted R-squared: 0.06254**

However, now we have to be skeptical about adding too many predictor variables for we do not want to end up with potential multi-colinearity issues.

Model Selection

Binary logistic regression

Confusion Matrices

We generate confusion matrices for our five models using a $p = 0.5$ threshold.

Confusion Matrix for Model 1:

Confusion Matrix for Model 2:

Confusion Matrix for Model 3:

Confusion Matrix for Model 4:

Confusion Matrix for Model 5:

ROC Curves

We generate the ROC curves for all of our models.

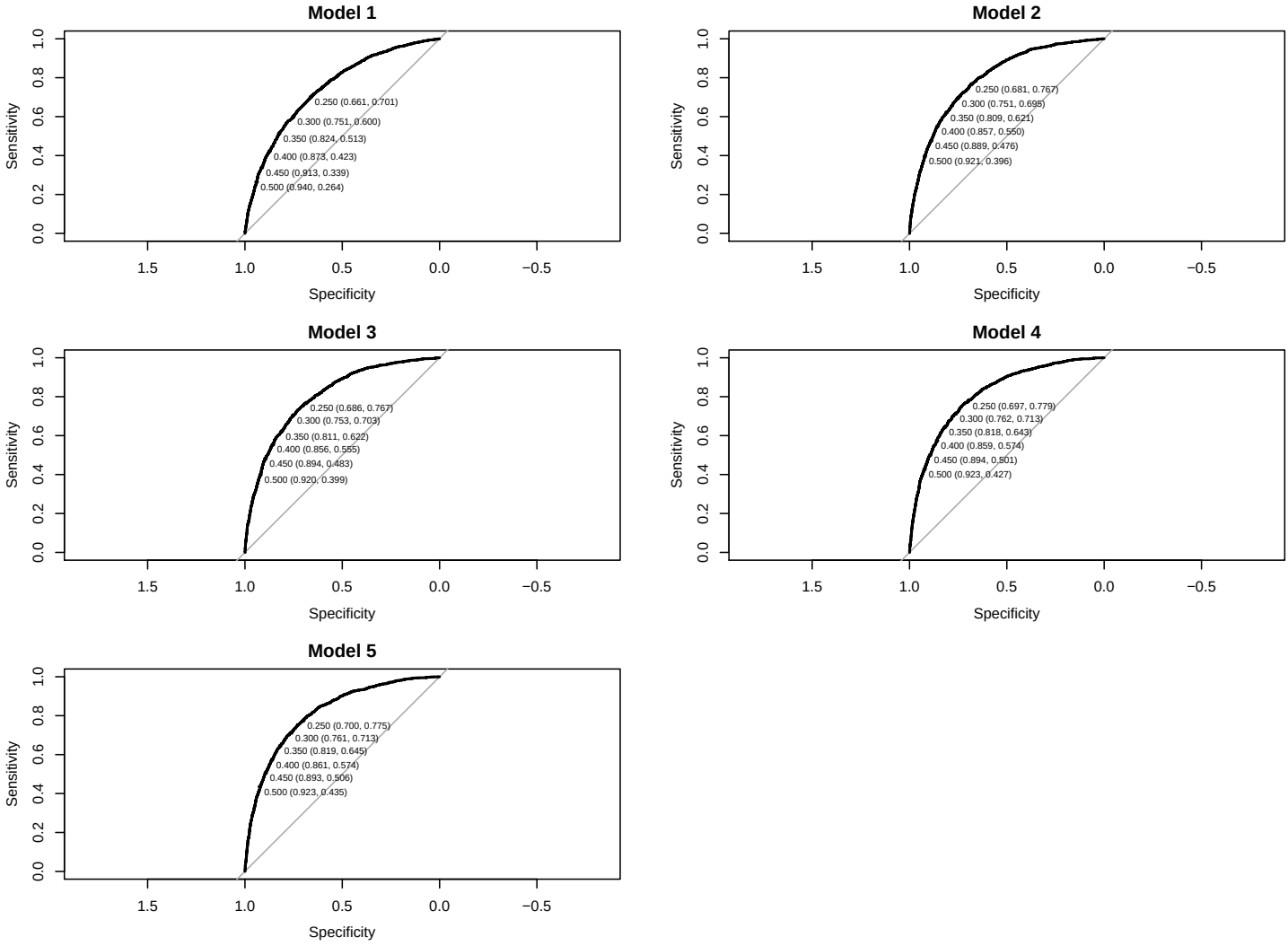

Figure 9. ROC curves for models 1-5

The ROC curves in figure 9, show how the different binary models perform. Generally, a lower AIC and a steeper curve are indicators of better models. Model 5, while having a slightly high AIC than model 4, has a much better ROC curve. So we will

likely select this for our binary model. ## Linear regression model option summary of TARGET_AMT

Model options summary

Below is a summary of the key indicators for all three models to help us decide which model is the best. Based on the summary, we will select model one, with the lowest R^2 value.

Model	F-statistic	p-value	Adjusted R-squared	Multiple R-squared
Model 1	14.36	< 2.2e-16	0.01809	0.01945
Model 2	30.10	< 2.2e-16	0.05478	0.05666
Model 3	23.92	< 2.2e-16	0.06254	0.06527

Table 6. Summary statistics of linear models

Residual Plots

We now compare the Residual plots to help us decide on the selection of the best model. The residual plots in figures 10 through 12, show that the distributions in the models are not looking that normal even after outlier imputation. Invoking the central limit theorem (CLT) we will do nothing further as we have sufficient data to understand that the variance of the errors in these models is likely finite.

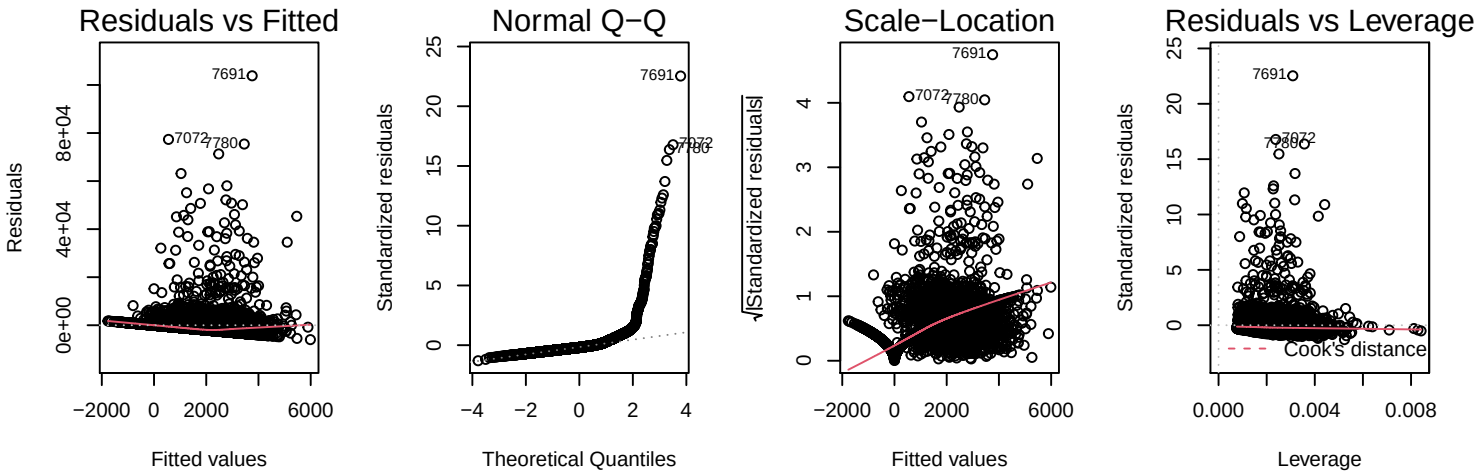

Figure 10. Model 1 plots

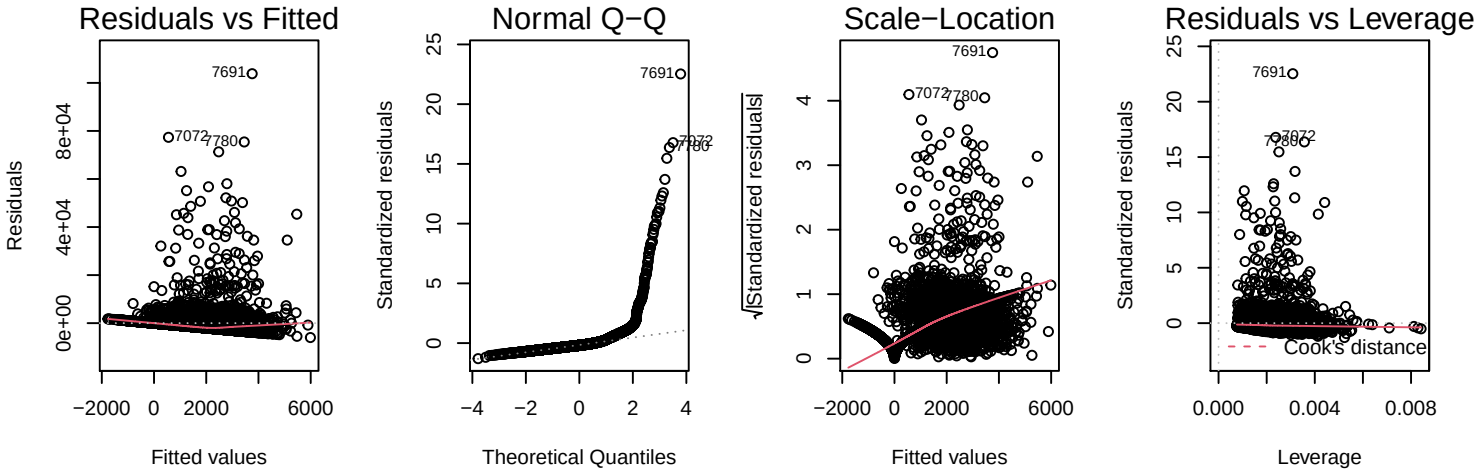

Figure 11. Model 2 plots

```
##
## Call:
## lm(formula = TARGET_AMT ~ URBANICITY + MVR_PTS + CAR_USE + CAR_TYPE +
##     CAR_TYPE + TIF + MSTATUS + TRAVTIME + REVOKED + PARENT1 +
##     KIDSDRIV + CLM_FREQ + INCOME + CAR_AGE + SEX + BLUEBOOK,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5748  -1683   -800    313 103642
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.247e+02  3.525e+02   0.921 0.357028
## URBANICITYHighly Urban/ Urban  1.581e+03  1.523e+02  10.382  < 2e-16 ***
## MVR_PTS                        1.805e+02  2.911e+01   6.199 6.03e-10 ***
## CAR_USEPrivate                -8.952e+02  1.403e+02  -6.378 1.92e-10 ***
## CAR_TYPEPanel Truck           -7.406e+01  2.958e+02  -0.250 0.802279
## CAR_TYPEPickup                 3.513e+02  1.861e+02   1.888 0.059115 .
## CAR_TYPESports Car             9.632e+02  2.440e+02   3.947 7.99e-05 ***
## CAR_TYPESUV                    7.759e+02  2.013e+02   3.855 0.000117 ***
## CAR_TYPEVan                    5.817e+02  2.346e+02   2.480 0.013173 *
## TIF                           -5.309e+01  1.362e+01  -3.897 9.82e-05 ***
## MSTATUSYes                    -5.762e+02  1.347e+02  -4.276 1.93e-05 ***
## TRAVTIME                       1.210e+01  3.621e+00   3.341 0.000839 ***
## REVOKEDYes                     3.298e+02  1.757e+02   1.877 0.060576 .
## PARENT1Yes                     7.486e+02  1.983e+02   3.775 0.000162 ***
## KIDSDRIV                       3.741e+02  1.162e+02   3.219 0.001291 **
## CLM_FREQ                       7.797e+01  5.511e+01   1.415 0.157159
## INCOME                        -6.445e-03  1.463e-03  -4.405 1.08e-05 ***
## CAR_AGE                       -3.480e+01  1.133e+01  -3.072 0.002135 **
## SEXM                           3.297e+02  1.798e+02   1.834 0.066718 .
## BLUEBOOK                       1.638e-02  9.656e-03   1.696 0.089849 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4597 on 6508 degrees of freedom
## Multiple R-squared:  0.06527,    Adjusted R-squared:  0.06254
## F-statistic: 23.92 on 19 and 6508 DF,  p-value: < 2.2e-16
```

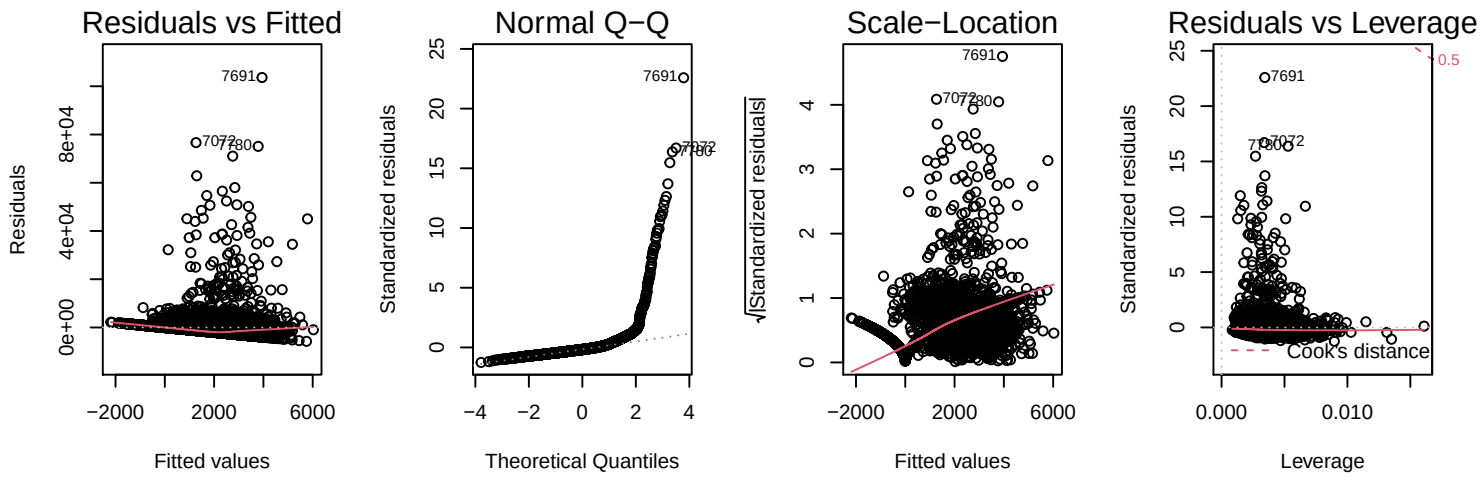

Figure 12. Model 3 plots

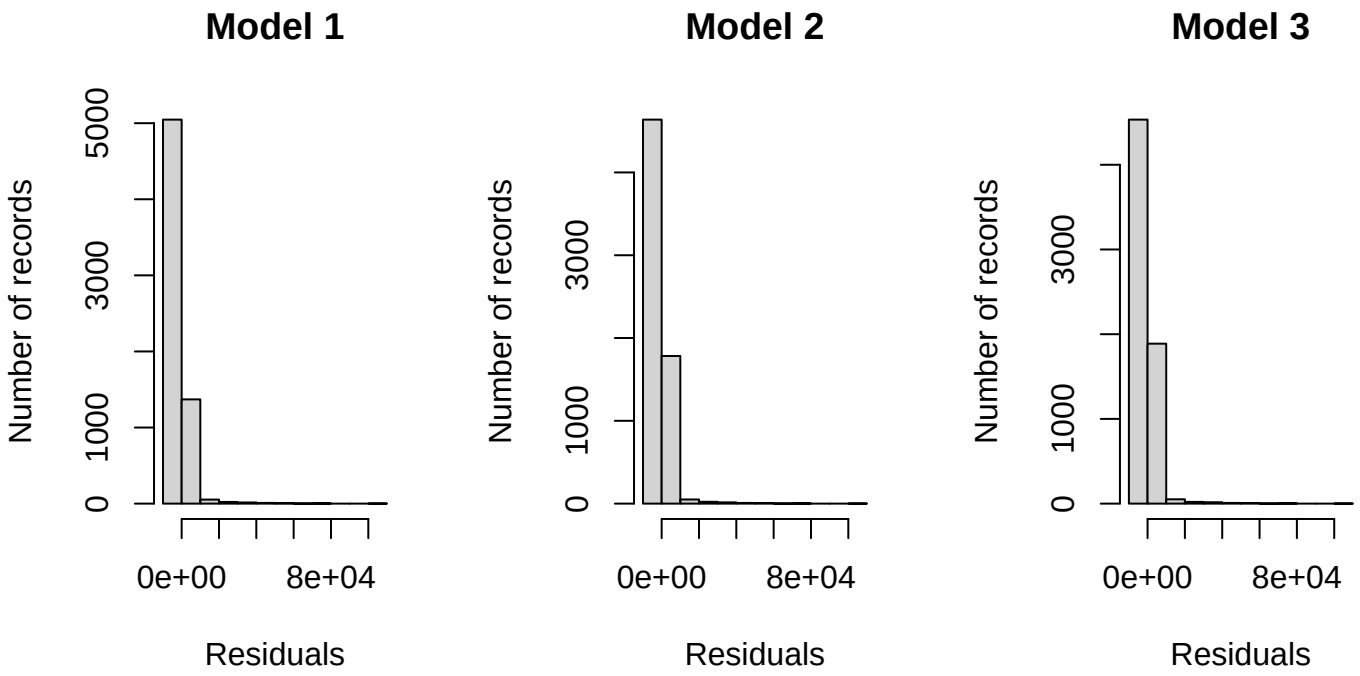

Figure 13. Plots of model residuals

Conclusions

Code Appendix

0.1 Overview

0.1.1 Libraries Used

library(tidyverse) library(ggplot2) library(VIM) library(GGally) library(caret) library(broom) library(naniar) library(stringr) library(pROC) library(ggpubr)

1.1 Data Exloration

1.2 Data Import

rawTrain <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW4/insurance_training_data.csv", header = TRUE, stringsAsFactors = FALSE) rawTest <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW4/insurance-evaluation-data.csv")

1.3 Glimpse on Training Data

```
glimpse(rawTrain)
```

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0,~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
## $ EDUCATION   <fct> PhD, High School, High School, <High School, PhD, Bachelor~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

1.4 Figure 1. Plot of missing values

gg_miss_var(rawTrain)

1.5 Data Cleaning - Converting Data Types

rawTrain <- rawTrain %>% mutate(INCOME = gsub("\$", "", INCOME), #Remove $ HOME_VAL = gsub("\", "", $HOME_VAL$), BL
$gsub($"
","", BLUEBOOK), OLDCLAIM = gsub("\$","", OLDCLAIM), MSTATUS = gsub("z_","", MSTATUS), SEX = gsub("z_","", SEX), EDUCATION= gsub("z_","", EDUCATION), JOB= gsub("z_","", JOB), CAR_TYPE= gsub("z_","", CAR_TYPE), URBANICITY= gsub("z_","", URBANICITY), INCOME = as.numeric(gsub(",","", INCOME)), #remove , and cast to numeric HOME_VAL = as.numeric(gsub(",","", HOME_VAL)), BLUEBOOK = as.numeric(gsub(",","", BLUEBOOK)), OLDCLAIM = as.numeric(gsub(",","", OLDCLAIM)), TARGET_FLAG = as.factor(TARGET_FLAG))

rawTrain[sapply(rawTrain, is.character)] <- lapply(rawTrain[sapply(rawTrain, is.character)], as.factor)

rawTest <- rawTest %>% mutate(INCOME = gsub("\$", "", INCOME), #Remove $ HOME_VAL = gsub("\", "", $HOME_VAL$), BLU
$gsub($"
","", BLUEBOOK), OLDCLAIM = gsub("\$","", OLDCLAIM), MSTATUS = gsub("z_","", MSTATUS), SEX = gsub("z_","", SEX), EDUCATION= gsub("z_","", EDUCATION), JOB= gsub("z_","", JOB), CAR_TYPE= gsub("z_","", CAR_TYPE), URBANICITY= gsub("z_","", URBANICITY), INCOME = as.numeric(gsub(",","", INCOME)), #remove , and cast to numeric HOME_VAL = as.numeric(gsub(",","", HOME_VAL)), BLUEBOOK = as.numeric(gsub(",","", BLUEBOOK)), OLDCLAIM = as.numeric(gsub(",","", OLDCLAIM)), TARGET_FLAG = as.factor(TARGET_FLAG), TARGET_AMT = as.numeric(TARGET_AMT))

rawTest[sapply(rawTest, is.character)] <- lapply(rawTest[sapply(rawTest, is.character)], as.factor)

1.6 Confirmation glimpse

```
# Let's glimpse the data to confirm the data cleaning.
glimpse(rawTrain)
```

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0,~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
## $ EDUCATION   <fct> PhD, High School, High School, <High School, PhD, Bachelor~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

1.7 Summary Statistics

```
#Display summary statistics again to confirm data cleaning.
summary(rawTrain)
```

```
##      INDEX       TARGET_FLAG   TARGET_AMT        KIDSDRIV           AGE
##  Min.   :    1   0:6008      Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.: 2559   1:2153      1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median : 5133               Median :     0   Median :0.0000   Median :45.00
##  Mean   : 5152               Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.: 7745               3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :10302               Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                                NA's   :6
##     HOMEKIDS           YOJ           INCOME       PARENT1       HOME_VAL
##  Min.   :0.0000   Min.   : 0.0   Min.   :     0   No :7084   Min.   :     0
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Yes:1077   1st Qu.:     0
##  Median :0.0000   Median :11.0   Median : 54028              Median :161160
##  Mean   :0.7212   Mean   :10.5   Mean   : 61898              Mean   :154867
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986              3rd Qu.:238724
##  Max.   :5.0000   Max.   :23.0   Max.   :367030              Max.   :885282
##                   NA's   :454    NA's   :445                NA's   :464
##  MSTATUS    SEX            EDUCATION              JOB          TRAVTIME
```

```
##  No :3267   F:4375   <High School:1203   Blue Collar :1825   Min.   :  5.00
##  Yes:4894   M:3786   Bachelors   :2242   Clerical    :1271   1st Qu.: 22.00
##                      High School :2330   Professional:1117   Median : 33.00
##                      Masters     :1658   Manager     : 988   Mean   : 33.49
##                      PhD         : 728   Lawyer      : 835   3rd Qu.: 44.00
##                                          Student     : 712   Max.   :142.00
##                                          (Other)     :1413
##        CAR_USE        BLUEBOOK          TIF               CAR_TYPE
##  Commercial:3029   Min.   : 1500   Min.   : 1.000   Minivan    :2145
##  Private   :5132   1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676
##                    Median :14440   Median : 4.000   Pickup     :1389
##                    Mean   :15710   Mean   : 5.351   Sports Car : 907
##                    3rd Qu.:20850   3rd Qu.: 7.000   SUV        :2294
##                    Max.   :69740   Max.   :25.000   Van        : 750
##
##  RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED       MVR_PTS
##  no :5783   Min.   :    0   Min.   :0.0000   No :7161   Min.   : 0.000
##  yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000
##             Median :    0   Median :0.0000              Median : 1.000
##             Mean   : 4037   Mean   :0.7986              Mean   : 1.696
##             3rd Qu.: 4636   3rd Qu.:2.0000              3rd Qu.: 3.000
##             Max.   :57037   Max.   :5.0000              Max.   :13.000
##
##     CAR_AGE                     URBANICITY
##  Min.   :-3.000   Highly Rural/ Rural:1669
##  1st Qu.: 1.000   Highly Urban/ Urban:6492
##  Median : 8.000
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

1.8 Figure 2. Feature histograms

histData <- rawTrain %>% select(AGE, HOMEKIDS, YOJ,TRAVTIME, TIF, CLM_FREQ, MVR_PTS, CAR_AGE, INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM)

par(mfrow = c(3,4)) for(i in 1:ncol(histData)) {#distribution of each variable hist(histData[[i]], main = colnames(histData[i]), col = "red")

}

1.9 Outlier Analysis

1.9.1 Figure 3. Feature box plot, part 1

longData <- histData %>% select(-HOME_VAL, -INCOME, -BLUEBOOK, -OLDCLAIM) %>% # remove this for scale issue will plot below gather(key = Variable, value = Value)

ggplot(longData, aes(Variable, Value, fill = Variable)) +geom_boxplot(outlier.colour="blue", outlier.shape=4, outlier.size=2, show.legend=FALSE) + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + coord_flip()+ labs(title="Insurance Data Variables - PART 1", y="Value")

1.9.2 Figure 4. Feature box plot, part 2

longData2 <- histData %>% select(HOME_VAL, INCOME, BLUEBOOK, OLDCLAIM) %>% # remove this for scale issue will plot below gather(key = Variable, value = Value)

ggplot(longData2, aes(Variable, Value, fill = Variable)) +geom_boxplot(outlier.colour="blue", outlier.shape=4, outlier.size=2, show.legend=FALSE) + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + coord_flip()+ labs(title="Insurance Data Variables - PART 2", y="Value")

1.10 Figure 5. Categorical variables showing balance/imbalance

p1 <- ggplot(rawTrain, aes(x = PARENT1, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Single Parent (Parent 1)")

p2 <- ggplot(rawTrain, aes(x = MSTATUS, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Marital Status")

p3 <- ggplot(rawTrain, aes(x = SEX, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - SEX")

p4 <- ggplot(rawTrain, aes(x = EDUCATION, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Max Education Level")

p5 <- ggplot(rawTrain, aes(x = JOB, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Job Category")

p6 <- ggplot(rawTrain, aes(x = CAR_USE, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Vehicle Use")

p7 <- ggplot(rawTrain, aes(x = CAR_TYPE, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Car Type")

p8 <- ggplot(rawTrain, aes(x = RED_CAR, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Red Car")

p9 <- ggplot(rawTrain, aes(x = REVOKED, fill = TARGET_FLAG)) + geom_bar() + labs(title="Insurance Data Categorical Variables - Licensed Revoked (Past 7 Years)")

ggarrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, ncol = 2, nrow = 5)

1.11 Numeric Data - Relationship to Target / Figure 6. Feature plots

par(mfrow = c(5,3))

histData <- rawTrain %>% select(TARGET_AMT, AGE, HOMEKIDS, YOJ,TRAVTIME, TIF, CLM_FREQ, MVR_PTS, CAR_AGE, INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM)

featurePlot(x= histData[3:6], y = histData[['TARGET_AMT']]) featurePlot(x= histData[7:10], y = histData[['TARGET_AMT']]) featurePlot(x= histData[11:13], y = histData[['TARGET_AMT']])

1.12 Correlation

1.12.1 Figure 7. Correlation matrix

ggcorr(rawTrain)

1.12.2 findCorrelation() function

findCorrelation(cor(histData),cutoff = 0.75, verbose = TRUE, names = TRUE)

2.1 Data Preparation

2.2 Impute Missing Values

```
#due to skew home_val, income  will be imputed with median
#Age YOJ with the mean

#new DF
prepTrain <- rawTrain %>%
```

```
  select(-INDEX)

#impute NAs
prepTrain$AGE[is.na(prepTrain$AGE)] <- mean(prepTrain$AGE, na.rm=TRUE)
prepTrain$YOJ[is.na(prepTrain$YOJ)] <- mean(prepTrain$YOJ, na.rm=TRUE)
prepTrain$HOME_VAL[is.na(prepTrain$HOME_VAL)] <- median(prepTrain$HOME_VAL, na.rm=TRUE)
prepTrain$INCOME[is.na(prepTrain$INCOME)] <- median(prepTrain$INCOME, na.rm=TRUE)
prepTrain$CAR_AGE[is.na(prepTrain$CAR_AGE)] <- mean(prepTrain$CAR_AGE, na.rm=TRUE)
```

Variable Importance

To determine the variable importance the following steps were taken:

- A training data frame `prepTrainA` was prepared for the `TARGET_FLAG` response variable and its associated predictor variables.
- A training data frame `prepTrainB` was prepared for the `TARGET_AMT` response variable and its associated predictor variables.
- Using the `prepTrainA` data frame, a classification model `modelA` was trained using the `Learning Vector Quantization (lvq)` method. From it, the variable importance was summarized and plotted.

```
## ROC curve variable importance
##
##             Importance
## CLM_FREQ        0.6354
## OLDCLAIM        0.6339
## MVR_PTS         0.6202
## HOME_VAL        0.6176
## URBANICITY      0.6026
## INCOME          0.5961
## CAR_USE         0.5782
## MSTATUS         0.5751
## BLUEBOOK        0.5750
## HOMEKIDS        0.5706
## AGE             0.5686
## CAR_AGE         0.5640
## CAR_TYPE        0.5632
## PARENT1         0.5605
## REVOKED         0.5565
## TIF             0.5543
## EDUCATION       0.5424
## JOB             0.5414
## KIDSDRIV        0.5387
## TRAVTIME        0.5371
## YOJ             0.5362
## SEX             0.5119
## RED_CAR         0.5036
```

Table 1. Feature importance - part A

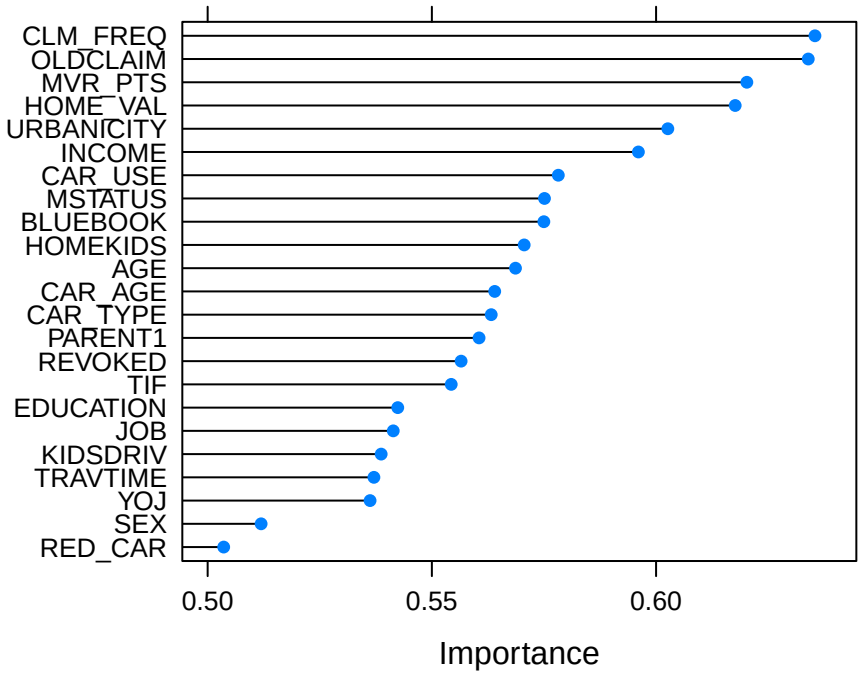

Figure 8. Plot of feature importance - part A

According to the plots above (figure 8), we can predict which variables would contribute best to the categorical predictions for `TARGET_FLAG`. We can use this to inform our data transformations.

Using the `prepTrainB` data frame, a classification/regression model `modelB` was trained using the `Generalized Linear Model (glm)` method. From it, the variable importance was summarized and plotted.

```
## glm variable importance
##
##   only 23 most important variables shown (out of 37)
##
##                                 Overall
## `URBANICITYHighly Urban/ Urban`  11.944
## MVR_PTS                           6.764
## CAR_USEPrivate                    4.741
## `CAR_TYPESports Car`              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
## KIDSDRIV                          2.776
## CLM_FREQ                          2.574
## INCOME                            2.441
## CAR_TYPEVan                       2.413
## CAR_TYPEPickup                    2.200
## CAR_AGE                           2.096
## SEXM                              2.007
## BLUEBOOK                          1.663
## JOBManager                        1.660
## `JOBBlue Collar`                  1.578
## JOBClerical                       1.550
## JOBProfessional                   1.478
## OLDCLAIM                          1.420
```

Table 2. Feature importance - part B

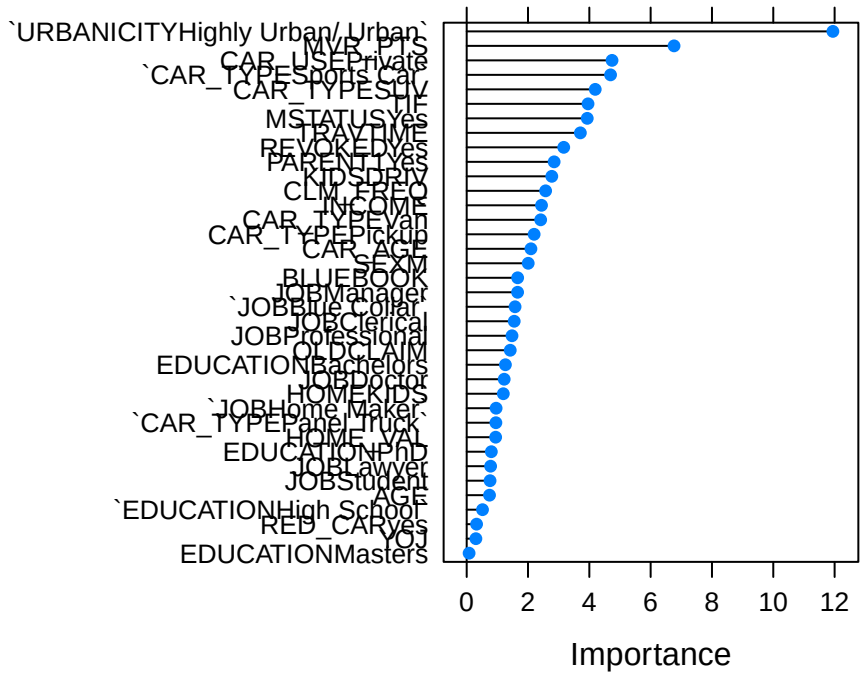

Figure 9. Plot of feature importance - part B

According to the plots above, we can predict which variables would contribute best to the numerical predictions for `TARGET_AMT`. We can use this to inform our data transformations.

Train Test Split

We partition the training data and set a seed in two data sets. One to be used for training purposes and one for validation/testing purposes.

Model Building

Binary Logistic Regression for dependent variable TARGET_FLAG

Binary Logistic Regression Model 1

For this model, we only include the predictor variables that have `theoretical effect on probability of collition`, which was provided as part of the definition of the variables.

Additionally, we remove the variables that were deemed as "urban legends," such as `RED_CAR` and `SEX`. From our importance variable model `importanceA`, we know that the variables `RED_CAR` and `SEX` ranked in the bottom 2 items of the importance list of 23 items. Hence, we don't include them. We also exclude variables having a theoretical "unknown effect" on probability of collision, such as `EDUCATION`.

$AIC = 6570.6$

Binary Logistic Regression Model 2

In order to improve on our first model, we use all the variables from Model 1, but we exclude the variables `YOJ`, which proved to be the least statistically significant for our Model 1.

Additionally, we include the variables `OLDCLAIM` and `URBANICITY`, which ranked 4th and 5th in our list of 23 predictor variable importance model `importanceA`.

$AIC = 6053.2$

```
## ROC curve variable importance
##
##   only 5 most important variables shown (out of 23)
##
##            Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
## HOME_VAL       0.6176
## URBANICITY     0.6026
```

Table 3. Feature importance - Binary model 2

We can see a significant improvement on the `residual deviance` and `AIC` values.

Binary Logistic Regression Model 3

In order to improve even more on our previous model, we add the variables `BLUEBOOK` and `HOMEKIDS`, which ranked 9th and 10th in our list of 23 predictor variable importance model `importanceA`.

At this point, the top 10 most statistically important of our set of 23 predictor variables are included in this model. This time, we can see an even more significant improvement on the `residual deviance` and `AIC` values.

$AIC = 6015$

```
## ROC curve variable importance
##
##   only 10 most important variables shown (out of 23)
##
##            Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
## HOME_VAL       0.6176
## URBANICITY     0.6026
## INCOME         0.5961
## CAR_USE        0.5782
## MSTATUS        0.5751
## BLUEBOOK       0.5750
## HOMEKIDS       0.5706
```

Table 4. Feature importance - Binary model 3

Binary Logistic Regression Model 4

Taking what we've learned from models 1-3, we add the variables `CAR_AGE`, `PARENT1` and `EDUCATION`, which ranked 12th, 14th and 17th in our list of 23 predictor variable importance model `importanceA`,

We also remove the variables `AGE` and `HOMEKIDS`, which from the previous models do not seem to contribute much, i.e., do not seem to be statistically significant for most of the models.

$AIC = 5897.7$

At this point, we can see most significant improvement on the `residual deviance` and `AIC` values.

Binary Logistic Regression Model 5

Just out of curiosity, what if we ignored all the statistical correlation and variable importance that we used for the previous four models. We use a model that includes all the predictor variables and the response variable `TARGET_FLAG`.

The results above show the best improvement so far.

Even after seeing the most significant improvement of all models, we still see that variables `AGE`, `HOMEKIDS`, `SEX`, and `RED_CAR (yes)` are not statistically significant. Which, lead us to believe that it might be true that deeming the variables `RED_CAR` and `SEX` as "urban legends" might be just urban legends. Those variable show little to no correlation to the probability of collision.

The variable `EDUCATION` seems to be statistically significant. At least for the values "Bachelors" and "Masters" we see that, based on the sign of their coefficients, they have a negative correlation to the theoretical probability of collision. So, it appears that people with higher education tend to have fewer accidents.

$AIC = 5903.2$

Linear Regression Models for dependent variable TARGET_AMT

Linear Regression Model 1

For our first model, we only include the predictor variables that have `theoretical probably of effecting the payout if there is a crash`, which was provided as part of the definition of the variables.

From the summary results we can see that we obtained low values for

Multiple R-squared: 0.01945 and **Adjusted R-squared: 0.01809**

Which, shows that using only predictor variables that have `theoretical probably of effecting the payout if there is a crash` is not a good way to go, for those variables do not seem to be enough to provide statistically significant results.

Linear Regression Model 2

For our second model, we only include the top 10 most important predictor variables that we gathered from our variable importance trained model `modelB`.

```
## glm variable importance
##
##   only 10 most important variables shown (out of 37)
##
##                                 Overall
## 'URBANICITYHighly Urban/ Urban'  11.944
## MVR_PTS                           6.764
## CAR_USEPrivate                    4.741
## 'CAR_TYPESports Car'              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
```

Table 5. Top 10 predictor variables from importance model B

Below are the results of applying our linear model 2 of TARGET_AMT vs Top 10 predictor variables.

From the summary results we can see that we obtained much better values for

Multiple R-squared: 0.05666, **Adjusted R-squared: 0.05478**

Linear Regression Model 3

We begin with a `baseline` model that includes all the predictor variables from Model 2 and the response variable `TARGET_AMT`. We remove the variables `CLM_FREQ` because its Pr value is 0.157159, which exceeds our requested 0.05 threshold. We will also add the next 6 variables from our variable importance model `modelB`. The added variables are: `KIDSDRIV`, `CLM_FREQ`, `INCOME`, `CAR_AGE`, `SEX`, and `BLUEBOOK`.

From the summary results above, we can see that the added variables have helped improve the values of our key indicators

Multiple R-squared: 0.06527, **Adjusted R-squared: 0.06254**

However, now we have to be skeptical about adding too many predictor variables for we do not want to end up with potential multi-colinearity issues.

Model Selection

Binary logistic regression

Confusion Matrices

We generate confusion matrices for our five models using a $p = 0.5$ threshold.

Confusion Matrix for Model 1:

Confusion Matrix for Model 2:

Confusion Matrix for Model 3:

Confusion Matrix for Model 4:

Confusion Matrix for Model 5:

ROC Curves

We generate the ROC curves for all of our models.

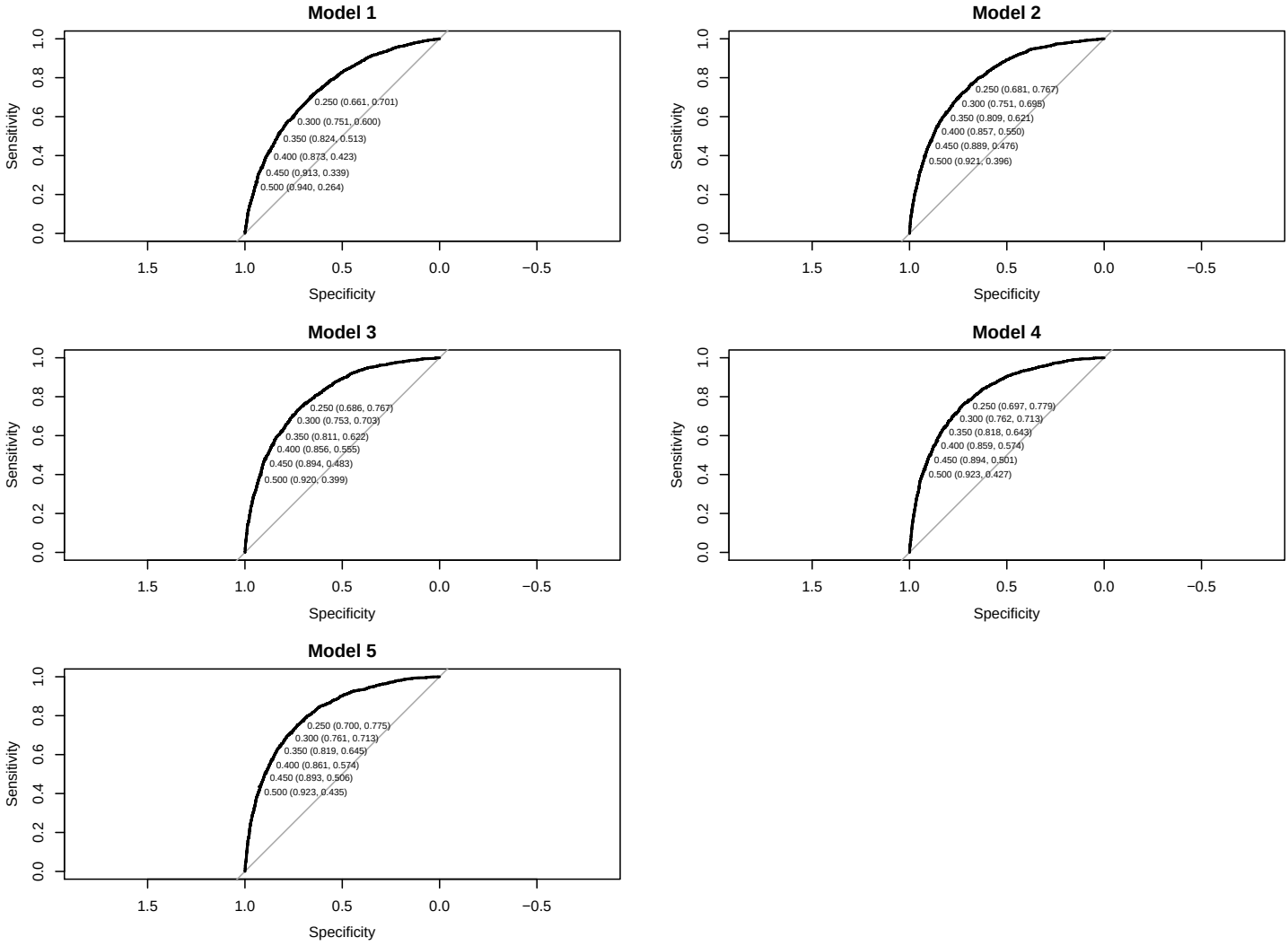

Figure 9. ROC curves for models 1-5

Linear regression model option summary of TARGET_AMT

Model options summary

Table 6. Summary statistics of linear models

Residual Plots

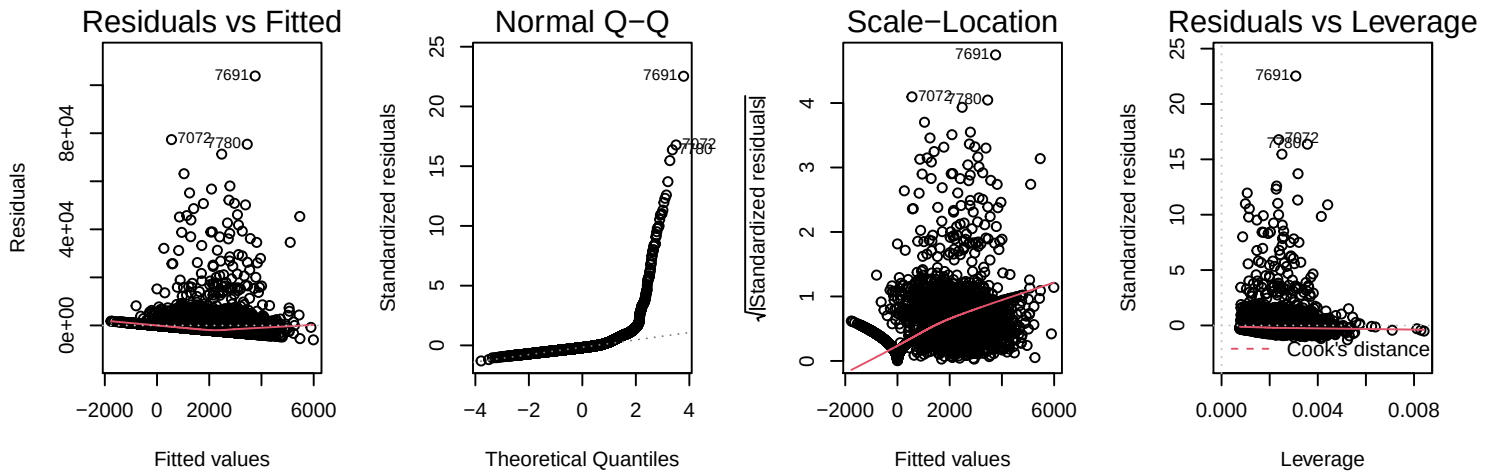

Figure 10. Model 1 plots

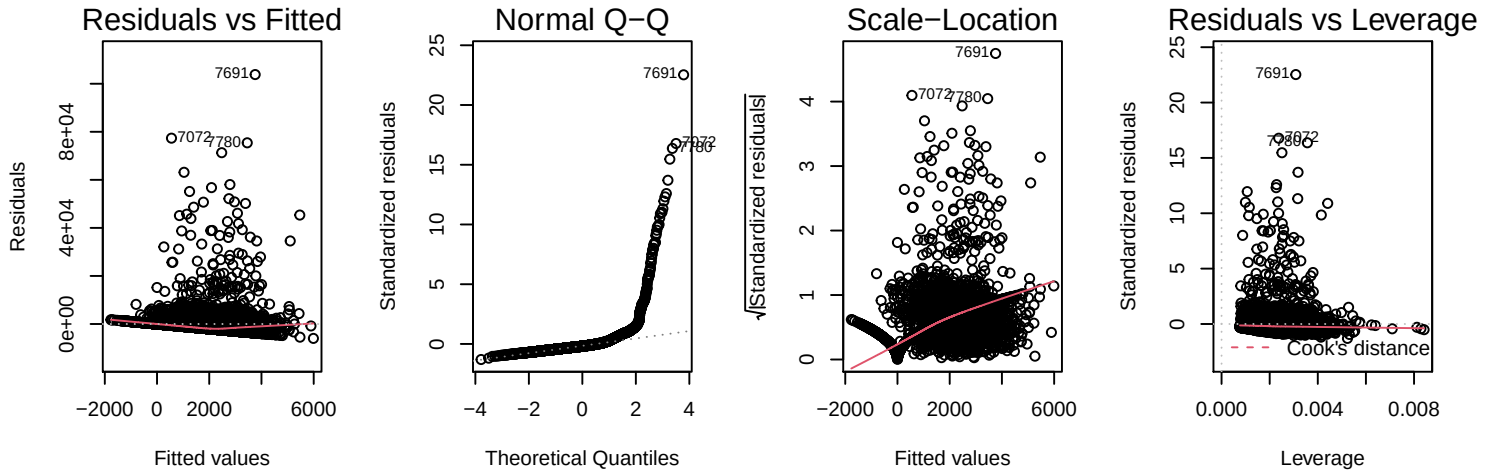

Figure 11. Model 2 plots

```
##
## Call:
## lm(formula = TARGET_AMT ~ URBANICITY + MVR_PTS + CAR_USE + CAR_TYPE +
##     CAR_TYPE + TIF + MSTATUS + TRAVTIME + REVOKED + PARENT1 +
##     KIDSDRIV + CLM_FREQ + INCOME + CAR_AGE + SEX + BLUEBOOK,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5748  -1683   -800    313 103642
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.247e+02  3.525e+02   0.921 0.357028
## URBANICITYHighly Urban/ Urban  1.581e+03  1.523e+02  10.382  < 2e-16 ***
## MVR_PTS                        1.805e+02  2.911e+01   6.199 6.03e-10 ***
## CAR_USEPrivate                -8.952e+02  1.403e+02  -6.378 1.92e-10 ***
## CAR_TYPEPanel Truck           -7.406e+01  2.958e+02  -0.250 0.802279
## CAR_TYPEPickup                 3.513e+02  1.861e+02   1.888 0.059115 .
## CAR_TYPESports Car             9.632e+02  2.440e+02   3.947 7.99e-05 ***
## CAR_TYPESUV                    7.759e+02  2.013e+02   3.855 0.000117 ***
## CAR_TYPEVan                    5.817e+02  2.346e+02   2.480 0.013173 *
## TIF                           -5.309e+01  1.362e+01  -3.897 9.82e-05 ***
## MSTATUSYes                    -5.762e+02  1.347e+02  -4.276 1.93e-05 ***
## TRAVTIME                       1.210e+01  3.621e+00   3.341 0.000839 ***
## REVOKEDYes                     3.298e+02  1.757e+02   1.877 0.060576 .
## PARENT1Yes                     7.486e+02  1.983e+02   3.775 0.000162 ***
## KIDSDRIV                       3.741e+02  1.162e+02   3.219 0.001291 **
## CLM_FREQ                       7.797e+01  5.511e+01   1.415 0.157159
## INCOME                        -6.445e-03  1.463e-03  -4.405 1.08e-05 ***
## CAR_AGE                       -3.480e+01  1.133e+01  -3.072 0.002135 **
## SEXM                           3.297e+02  1.798e+02   1.834 0.066718 .
## BLUEBOOK                       1.638e-02  9.656e-03   1.696 0.089849 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4597 on 6508 degrees of freedom
## Multiple R-squared:  0.06527,    Adjusted R-squared:  0.06254
## F-statistic: 23.92 on 19 and 6508 DF,  p-value: < 2.2e-16
```

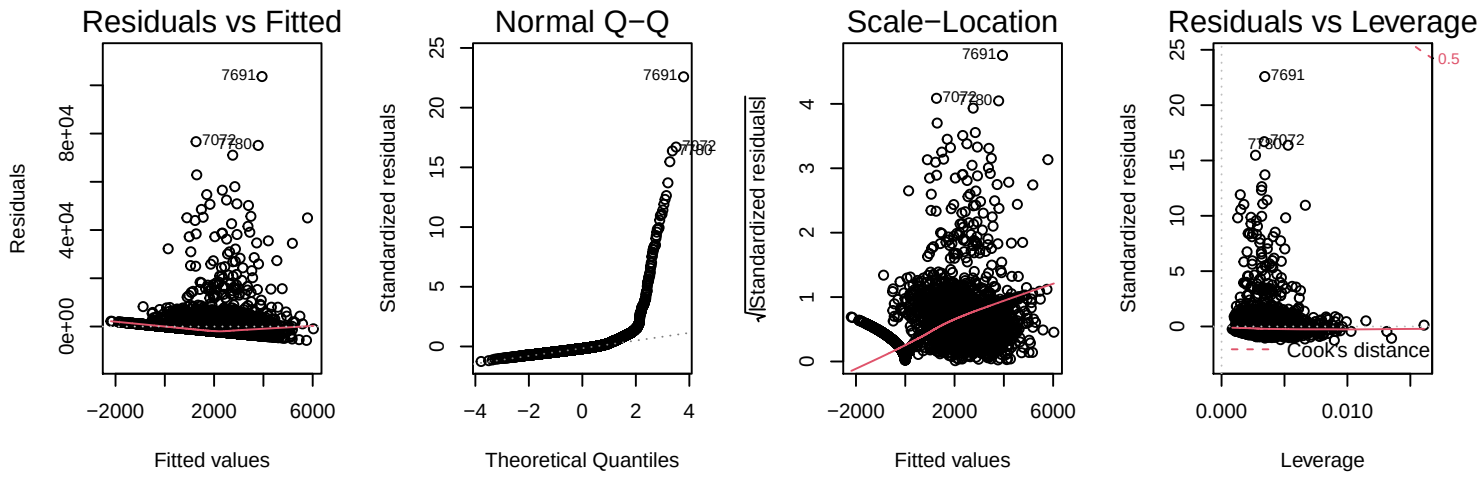

Figure 12. Model 3 plots

par(mfrow=c(1,2)) hist(linearRegModel1$residuals, xlab = "Residuals", ylab = "Number of records", main = "Model 1") hist(linearRe
xlab="Residuals",ylab="Number of records",main="Model 2") hist(linearRegModel3$residuals, xlab="Residuals",ylab="Number
of records",main="Model 3")

Figure 13. Plots of model residuals

Conclusions

rawTest$target_prob <- -predict(logRegModel5, newdata = rawTest)rawTest$TARGET_FLAG_PRED <- ifelse(rawTest$target_prob >=
-1.362578885, 1, 0)rawTest$TARGET_AMT_PRED <- ifelse(rawTest$TARGET_FLAG_PRED == 0, 0, predict(linearRegModel1,
newdata = rawTest))

rawTest %>% write.csv(., “insurance_pred.csv”, row.names = F)