HMW 4- Data 621

OMER OZEREN

Table of Contents

## Summary

The goal of this project is to build two models - one binary regression model to predict whether a vehicle will be involved in a crash and a linear regression model to predict possible payout.

Binary model resulted in good results. With the accuracy of close to ~80% it is noticeably better than just randomly guessing the outcome.

Linear model proved highly problematic. With value of 0.005 this model barely explains any variance in the outcome variable. Adding other variables that should influence the payout amount, such as CAR\_AGE, did not have a significant effect.

The model was created using only onservations for vehicles involved in a crash (TARGET\_FLAG==1). It is possible to create a linear regression model on all data, without this limitation. is significantly improved. In some tests, was about 0.27. However, I believe this explains the fit for observations with $0 payout (vehicles not involved in a crash) and does not help fit the actual payouts any better.

Looking at the relationship between Blue Book value and payout amount, there is a high amount of variability. There are many observations with low value, but high payout (too many to discount as outliers). This leads me to believe that some critical data is missing in order to accurately predict the payout.

## Data Exploration

The data set includes 8,161 observations with 24 variables (excluding the target variable).

#### Summary of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Type | Description | Comments |
| KIDSDRIV | Integer | No of children driving. | Ranges from 0 to 4. |
| AGE | Integer | Age of driver. | Ranges from 16 to 81. Contains 6 NAs. |
| HOMEKIDS | Integer | No of children at home. | Ranges from 0 to 5. |
| YOJ | Integer | Years on the job. | Ranges from 0 to 23. Contains 454 NAs (about 5.56% of all observations). |
| INCOME | Numeric | Income. | Ranges from $0 to $367,000. Contains 445 NAs (about 5.45% of all observations). Was converted to numeric by removing dollar signs and commas. |
| PARENT1 | Factor | Single parent flag. | Values: No, Yes. |
| HOME\_VAL | Numeric | Home value. | Ranges from $0 to $885,000. Contains 464 NAs (about 5.69% of all observations). Was converted to numeric by removing dollar signs and commas. |
| MSTATUS | Factor | Married flag. | Values: Yes, No. |
| SEX | Factor | Gender. | Values: M, F. |
| EDUCATION | Factor | Maximum education level. | Values: <High School, High School, Bachelors, Masters, PhD. |
| JOB | Factor | Job category. | Values: [Blank], Clerical, Doctor, Home Maker, Lawyer, Manager, Professional, Student, Blue Collar. |
| TRAVTIME | Integer | Distance to work. | Ranges from 5 to 142. |
| CAR\_USE | Factor | Vehicle use. | Values: Commercial, Private. |
| BLUEBOOK | Numeric | Vehicle value. | Ranges from $1,500 to $69,740. Was converted to numeric by removing dollar signs and commas. |
| TIF | Integer | Time in force. | Ranges from 1 to 25. |
| CAR\_TYPE | Factor | Vehicle type. | Values: Minivan, Panel Truck, Pickup, Sports Car, Van, SUV. |
| RED\_CAR | Factor | Red car flag. | Values: No, Yes |
| OLDCLAIM | Numeric | Total payout of claims. | Ranges from $0 to $57,040. Was converted to numeric by removing dollar signs and commas. |
| CLM\_FREQ | Integer | No of claims (past 5 years). | Ranges from 0 to 5. |
| REVOKED | Factor | Revoked license flag. | Values: No, Yes. |
| MVR\_PTS | Integer | Motor vehicle record points. | Ranges from 0 to 13. |
| CAR\_AGE | Integer | Vehicle age. | Ranges from -3 to 28. Contains 510 NAs (about 6.25% of all observations). |
| URBANICITY | Factor | Home/work area. | Values: Urban, Rural. |

The table below shows summary of numeric variables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | Median | Mean | SD | Max | Num of NAs | Num of Zeros |
| KIDSDRIV | 0 | 0 | 0.1711 | 0.5115 | 4 | 0 | 7180 |
| AGE | 16 | 45 | 44.79 | 8.628 | 81 | 6 | 0 |
| HOMEKIDS | 0 | 0 | 0.7212 | 1.116 | 5 | 0 | 5289 |
| YOJ | 0 | 11 | 10.5 | 4.092 | 23 | 454 | 625 |
| INCOME | 0 | 54028 | 61898 | 47573 | 367030 | 445 | 615 |
| HOME\_VAL | 0 | 161160 | 154867 | 129124 | 885282 | 464 | 2294 |
| TRAVTIME | 5 | 33 | 33.49 | 15.91 | 142 | 0 | 0 |
| BLUEBOOK | 1500 | 14440 | 15710 | 8420 | 69740 | 0 | 0 |
| TIF | 1 | 4 | 5.351 | 4.147 | 25 | 0 | 0 |
| OLDCLAIM | 0 | 0 | 4037 | 8777 | 57037 | 0 | 5009 |
| CLM\_FREQ | 0 | 0 | 0.7986 | 1.158 | 5 | 0 | 5009 |
| MVR\_PTS | 0 | 1 | 1.696 | 2.147 | 13 | 0 | 3712 |
| CAR\_AGE | -3 | 8 | 8.328 | 5.701 | 28 | 510 | 3 |

In the table above we can see a significant number of observations with value of 0. There is a logical explanation that these observations are valid:

* KIDSDRIV and HOMEKIDS: households without children
* YOJ and INCOME: unemployed individuals
* HOME\_VAL: renters
* OLDCLAIM, CLM\_FREQ and MVR\_PTS: safe drivers with no claims or DMV points

#### Hangling *NAs*

Several variables - AGE, YOJ, INCOME, HOME\_VAL, CAR\_AGE - contained some NA values. The number seemed significant; however, given the large number of observations it was safer to remove incomplete cases.

Additionally, JOB contained some blank values. It is possible that the blank value represented a certain category, for example *Unemployed*; however, without any indication whether it is missing value or specific category, it was treated similarly to NA values.

Additionally, CAR\_AGE had 4 values that appeared to be errors. One observations had CAR\_AGE as -3 and three observations had CAR\_AGE as 0. The negative value is clearly wrong; however, it is possible to make a case that 0 is a valid value (new car). Given low number of 0 values and high number of 1 for that variable, it is likely that 0 is an error and 1 represents a new car. The 4 affected records have were removed.

After these steps 2,120 observations were removed leaving 6,041 observations. As stated above, although the number of remove observations is significant, there are enough observations left to perform the analysis. For categorical variables, there are enough examples for each category.

#### Additional Details

* Boxplots and histograms were inspected for all variable in order to see any anomalies.
* Multiple values were prefixed with text ‘z\_’, which was removed from all observations. Affected variables are MSTATUS, EDUCATION, JOB, CAR\_TYPE and URBANICITY.
* Index column present in the data set has been removed.
* Levels for EDUCATION have been re-ordered to follow the most common order: *<High School*, *High School*, *Bachelors*, *Masters*, and *PhD*.
* Levels for JOB have been re-ordered to follow general order from low-paying to high-paying occupations: *Student*, *Blue Collar*, *Home Maker*, *Clerical*, *Professional*, *Manager*, *Lawyer*, and *Doctor*.
* Counts of observations for possible values of PARENT1:

|  |  |
| --- | --- |
| No | Yes |
| 5219 | 822 |

* Counts of observations for possible values of MSTATUS:

|  |  |
| --- | --- |
| Yes | No |
| 3594 | 2447 |

* Counts of observations for possible values of SEX:

|  |  |
| --- | --- |
| M | F |
| 2683 | 3358 |

* Counts of observations for possible values of EDUCATION:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| <High School | High School | Bachelors | Masters | PhD |
| 955 | 1871 | 1739 | 1060 | 416 |

* Counts of observations for possible values of JOB:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student | Blue Collar | Home Maker | Clerical | Professional | Manager | Lawyer | Doctor |
| 537 | 1476 | 484 | 1030 | 867 | 778 | 670 | 199 |

* Counts of observations for possible values of CAR\_USE:

|  |  |
| --- | --- |
| Commercial | Private |
| 2040 | 4001 |

* Counts of observations for possible values of CAR\_TYPE:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Minivan | Panel Truck | Pickup | Sports Car | Van | SUV |
| 1699 | 347 | 1012 | 714 | 488 | 1781 |

* Counts of observations for possible values of RED\_CAR:

|  |  |
| --- | --- |
| No | Yes |
| 4350 | 1691 |

* Counts of observations for possible values of REVOKED:

|  |  |
| --- | --- |
| No | Yes |
| 5297 | 744 |

* Counts of observations for possible values of URBANICITY:

|  |  |
| --- | --- |
| Urban | Rural |
| 4742 | 1299 |

* Counts of observations for possible values of KIDSDRIV:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 |
| 5309 | 473 | 207 | 50 | 2 |

* Counts of observations for possible values of HOMEKIDS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 |
| 3872 | 667 | 837 | 525 | 129 | 11 |

* Counts of observations for possible values of CLM\_FREQ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 |
| 3758 | 715 | 848 | 568 | 141 | 11 |

* As shown above there are several values for KIDSDRIV, HOMEKIDS and CLM\_FREQ with few observations.

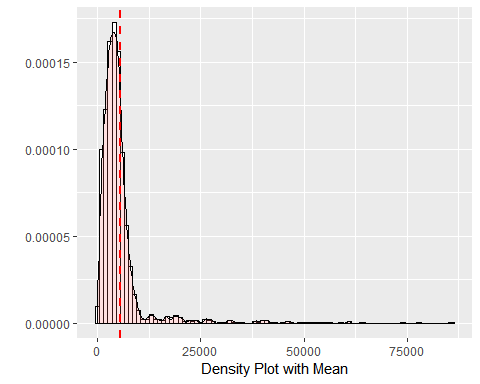
#### Target/Dependent Variable TARGET\_FLAG

Target variable for the binary regression model represents a flag whether a vehicle was involved in a crash. There are 4,440 observation with TARGET\_FLAG value of 0 and 1,601 observations with TARGET\_FLAG value of 1 making it about 75/25 split, or more precisely there are **73.5% of 0s and 26.5% of 1s**.

#### Target/Dependent Variable TARGET\_AMT

Target variable for the linear regression model represents the cost if a vehicle was involved in a crash. The value is presented only for observations with TARGET\_FLAG of 1. Distribution of TARGET\_AMT has a long right tail. There are no missing values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min | Median | Mean | SD | Max | Num\_NAs | Num\_Zeros |
| 30.28 | 4136 | 5586 | 7440 | 85524 | 0 | 0 |



## Modelling: Generalized Linear Model

The first model will be used to predict whether a vehicle will be involved in a crash. The dependent variable, TARGET\_FLAG, is binary. For this project it is assumed that observations are independent of each other as there is no reason to believe otherwise.

To test the accuracy of the model the main data set is split into training and testing data sets. The training set includes 75% of randomly chosen observations (4,531) while the testing set includes remaining 25% (1,510).

The starting point is a model that includes all independent variables. It has AIC value of 4085.5 and accuracy of 79.47%. Summary is below.

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ . - TARGET\_AMT, family = binomial(link = "logit"),   
## data = insTRAIN)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2532 -0.7047 -0.3874 0.6212 2.9097   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.057e-01 3.474e-01 -0.880 0.378824   
## KIDSDRIV 3.254e-01 8.162e-02 3.987 6.70e-05 \*\*\*  
## AGE -6.524e-03 5.457e-03 -1.195 0.231914   
## HOMEKIDS 1.095e-02 4.955e-02 0.221 0.825173   
## YOJ -4.818e-03 1.137e-02 -0.424 0.671742   
## INCOME -3.061e-06 1.632e-06 -1.876 0.060660 .   
## PARENT1Yes 3.940e-01 1.454e-01 2.709 0.006749 \*\*   
## HOME\_VAL -1.383e-06 4.914e-07 -2.815 0.004879 \*\*   
## MSTATUSNo 4.710e-01 1.165e-01 4.043 5.28e-05 \*\*\*  
## SEXF -1.745e-01 1.504e-01 -1.160 0.245995   
## EDUCATIONHigh School -4.663e-02 1.224e-01 -0.381 0.703169   
## EDUCATIONBachelors -5.190e-01 1.524e-01 -3.405 0.000662 \*\*\*  
## EDUCATIONMasters -6.353e-01 2.499e-01 -2.542 0.011006 \*   
## EDUCATIONPhD -6.920e-02 3.049e-01 -0.227 0.820482   
## JOBBlue Collar 9.702e-02 1.762e-01 0.551 0.581935   
## JOBHome Maker -8.012e-02 2.029e-01 -0.395 0.692862   
## JOBClerical 2.432e-01 1.762e-01 1.380 0.167584   
## JOBProfessional -9.445e-03 2.109e-01 -0.045 0.964277   
## JOBManager -6.364e-01 2.306e-01 -2.760 0.005782 \*\*   
## JOBLawyer 2.617e-01 2.815e-01 0.930 0.352593   
## JOBDoctor -5.717e-01 4.068e-01 -1.405 0.159912   
## TRAVTIME 1.420e-02 2.536e-03 5.599 2.15e-08 \*\*\*  
## CAR\_USEPrivate -8.163e-01 1.224e-01 -6.669 2.57e-11 \*\*\*  
## BLUEBOOK -2.233e-05 7.034e-06 -3.174 0.001503 \*\*   
## TIF -5.337e-02 9.914e-03 -5.383 7.32e-08 \*\*\*  
## CAR\_TYPEPanel Truck 6.422e-01 2.271e-01 2.827 0.004694 \*\*   
## CAR\_TYPEPickup 5.283e-01 1.322e-01 3.995 6.47e-05 \*\*\*  
## CAR\_TYPESports Car 1.104e+00 1.696e-01 6.512 7.42e-11 \*\*\*  
## CAR\_TYPEVan 4.579e-01 1.759e-01 2.603 0.009232 \*\*   
## CAR\_TYPESUV 7.862e-01 1.456e-01 5.398 6.74e-08 \*\*\*  
## RED\_CARYes -1.989e-01 1.199e-01 -1.658 0.097220 .   
## OLDCLAIM -1.796e-05 5.177e-06 -3.468 0.000524 \*\*\*  
## CLM\_FREQ 2.075e-01 3.832e-02 5.416 6.10e-08 \*\*\*  
## REVOKEDYes 8.732e-01 1.218e-01 7.167 7.67e-13 \*\*\*  
## MVR\_PTS 1.147e-01 1.851e-02 6.197 5.75e-10 \*\*\*  
## CAR\_AGE 7.534e-04 1.030e-02 0.073 0.941712   
## URBANICITYRural -2.360e+00 1.463e-01 -16.129 < 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: 5240.4 on 4530 degrees of freedom  
## Residual deviance: 4011.5 on 4494 degrees of freedom  
## AIC: 4085.5  
##   
## Number of Fisher Scoring iterations: 5

Running this model through the stepwise algorithm (stepAIC from the MASS package), removed AGE, HOMEKIDS, YOJ, SEX, RED\_CAR and CAR\_AGE. AIC is reduced slightly to 4078.2 and accuracy is improved very slightly to 79.54%.

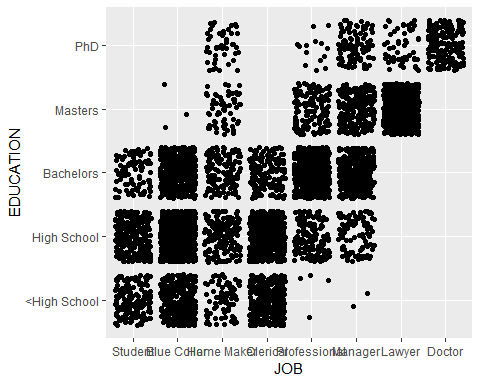
INCOME and HOME\_VAL are very right-skewed. To make results more normal, they are log-transformed (adding 1 to make sure that log-transformation is possible for 0 values). The new model again has slightly lower AIC of 4073.2 and slightly higher accuracy at 79.73%.

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ KIDSDRIV + log(INCOME + 1) + PARENT1 +   
## log(HOME\_VAL + 1) + MSTATUS + EDUCATION + JOB + TRAVTIME +   
## CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM + CLM\_FREQ +   
## REVOKED + MVR\_PTS + URBANICITY, family = binomial(link = "logit"),   
## data = insTRAIN)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3062 -0.7025 -0.3932 0.6137 2.9590   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.484e-01 2.494e-01 -1.397 0.162383   
## KIDSDRIV 3.248e-01 7.318e-02 4.439 9.05e-06 \*\*\*  
## log(INCOME + 1) -6.371e-02 1.901e-02 -3.351 0.000805 \*\*\*  
## PARENT1Yes 4.522e-01 1.255e-01 3.604 0.000313 \*\*\*  
## log(HOME\_VAL + 1) -3.312e-02 9.474e-03 -3.496 0.000473 \*\*\*  
## MSTATUSNo 3.924e-01 1.145e-01 3.426 0.000613 \*\*\*  
## EDUCATIONHigh School -5.933e-02 1.221e-01 -0.486 0.626878   
## EDUCATIONBachelors -5.946e-01 1.387e-01 -4.288 1.80e-05 \*\*\*  
## EDUCATIONMasters -7.595e-01 2.179e-01 -3.485 0.000492 \*\*\*  
## EDUCATIONPhD -3.272e-01 2.696e-01 -1.214 0.224811   
## JOBBlue Collar 3.822e-01 2.018e-01 1.894 0.058233 .   
## JOBHome Maker 6.055e-02 2.130e-01 0.284 0.776192   
## JOBClerical 6.048e-01 2.050e-01 2.950 0.003175 \*\*   
## JOBProfessional 2.447e-01 2.315e-01 1.057 0.290386   
## JOBManager -4.020e-01 2.480e-01 -1.621 0.105038   
## JOBLawyer 5.227e-01 2.983e-01 1.752 0.079716 .   
## JOBDoctor -3.706e-01 4.133e-01 -0.897 0.369957   
## TRAVTIME 1.422e-02 2.540e-03 5.599 2.16e-08 \*\*\*  
## CAR\_USEPrivate -8.122e-01 1.225e-01 -6.632 3.32e-11 \*\*\*  
## BLUEBOOK -2.595e-05 6.304e-06 -4.116 3.85e-05 \*\*\*  
## TIF -5.211e-02 9.911e-03 -5.258 1.45e-07 \*\*\*  
## CAR\_TYPEPanel Truck 6.821e-01 2.146e-01 3.179 0.001478 \*\*   
## CAR\_TYPEPickup 5.412e-01 1.321e-01 4.097 4.18e-05 \*\*\*  
## CAR\_TYPESports Car 1.057e+00 1.392e-01 7.594 3.10e-14 \*\*\*  
## CAR\_TYPEVan 4.567e-01 1.709e-01 2.673 0.007516 \*\*   
## CAR\_TYPESUV 7.558e-01 1.120e-01 6.750 1.48e-11 \*\*\*  
## OLDCLAIM -1.792e-05 5.186e-06 -3.455 0.000551 \*\*\*  
## CLM\_FREQ 2.081e-01 3.827e-02 5.439 5.37e-08 \*\*\*  
## REVOKEDYes 8.647e-01 1.219e-01 7.092 1.32e-12 \*\*\*  
## MVR\_PTS 1.148e-01 1.853e-02 6.196 5.79e-10 \*\*\*  
## URBANICITYRural -2.391e+00 1.471e-01 -16.254 < 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: 5240.4 on 4530 degrees of freedom  
## Residual deviance: 4011.2 on 4500 degrees of freedom  
## AIC: 4073.2  
##   
## Number of Fisher Scoring iterations: 5

Checking this model using various evaluation methods, a potential issue was discovered with variance-inflation (using vif in the car package).

## GVIF Df GVIF^(1/(2\*Df))  
## KIDSDRIV 1.088467 1 1.043296  
## log(INCOME + 1) 2.530758 1 1.590836  
## PARENT1 1.427433 1 1.194752  
## log(HOME\_VAL + 1) 1.965195 1 1.401854  
## MSTATUS 2.118875 1 1.455636  
## EDUCATION 7.054761 4 1.276616  
## JOB 29.531678 7 1.273563  
## TRAVTIME 1.034921 1 1.017311  
## CAR\_USE 2.336829 1 1.528669  
## BLUEBOOK 1.575884 1 1.255342  
## TIF 1.010050 1 1.005012  
## CAR\_TYPE 2.221595 5 1.083095  
## OLDCLAIM 1.679788 1 1.296066  
## CLM\_FREQ 1.449881 1 1.204110  
## REVOKED 1.341382 1 1.158181  
## MVR\_PTS 1.174200 1 1.083605  
## URBANICITY 1.142851 1 1.069042

The value for JOB is very high (although when adjusted for degrees of freedom is does not appear to be problematic). The value for EDUCATION is also fairly high. Comparing these two variables we can see possible correlation between them.



This is quite logical. Some occupations, such as doctor or lawyer, require a graduate level degree. At the same time lower levels of education are more connect to blue collar work rather than professional work. JOB was removed from the model.

#### Final Model

The formula for the final generalized linear model is TARGET\_FLAG ~ KIDSDRIV + log(INCOME+1) + PARENT1 + log(HOME\_VAL+1) + MSTATUS + EDUCATION + TRAVTIME + CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM + CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY. AIC is 4107.

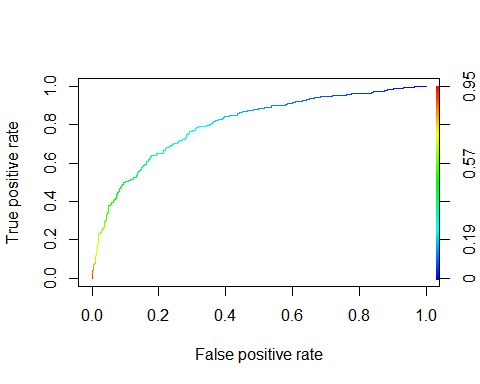
##   
## Call:  
## glm(formula = TARGET\_FLAG ~ KIDSDRIV + log(INCOME + 1) + PARENT1 +   
## log(HOME\_VAL + 1) + MSTATUS + EDUCATION + TRAVTIME + CAR\_USE +   
## BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM + CLM\_FREQ + REVOKED +   
## MVR\_PTS + URBANICITY, family = binomial(link = "logit"),   
## data = insTRAIN)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3080 -0.7092 -0.4102 0.6195 2.9691   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.012e-01 2.367e-01 -0.850 0.395436   
## KIDSDRIV 3.163e-01 7.239e-02 4.369 1.25e-05 \*\*\*  
## log(INCOME + 1) -4.046e-02 1.317e-02 -3.073 0.002119 \*\*   
## PARENT1Yes 4.630e-01 1.241e-01 3.730 0.000192 \*\*\*  
## log(HOME\_VAL + 1) -2.605e-02 8.541e-03 -3.050 0.002291 \*\*   
## MSTATUSNo 4.317e-01 1.092e-01 3.952 7.75e-05 \*\*\*  
## EDUCATIONHigh School -1.621e-01 1.183e-01 -1.371 0.170482   
## EDUCATIONBachelors -8.241e-01 1.235e-01 -6.673 2.50e-11 \*\*\*  
## EDUCATIONMasters -8.845e-01 1.426e-01 -6.202 5.59e-10 \*\*\*  
## EDUCATIONPhD -8.916e-01 1.931e-01 -4.617 3.89e-06 \*\*\*  
## TRAVTIME 1.438e-02 2.515e-03 5.717 1.09e-08 \*\*\*  
## CAR\_USEPrivate -8.525e-01 9.843e-02 -8.661 < 2e-16 \*\*\*  
## BLUEBOOK -2.722e-05 6.234e-06 -4.367 1.26e-05 \*\*\*  
## TIF -5.137e-02 9.859e-03 -5.211 1.88e-07 \*\*\*  
## CAR\_TYPEPanel Truck 5.620e-01 2.039e-01 2.757 0.005841 \*\*   
## CAR\_TYPEPickup 4.841e-01 1.287e-01 3.761 0.000169 \*\*\*  
## CAR\_TYPESports Car 1.006e+00 1.371e-01 7.335 2.21e-13 \*\*\*  
## CAR\_TYPEVan 4.125e-01 1.692e-01 2.438 0.014771 \*   
## CAR\_TYPESUV 7.359e-01 1.106e-01 6.655 2.83e-11 \*\*\*  
## OLDCLAIM -1.776e-05 5.131e-06 -3.461 0.000538 \*\*\*  
## CLM\_FREQ 2.059e-01 3.794e-02 5.428 5.71e-08 \*\*\*  
## REVOKEDYes 8.704e-01 1.209e-01 7.199 6.07e-13 \*\*\*  
## MVR\_PTS 1.218e-01 1.840e-02 6.618 3.64e-11 \*\*\*  
## URBANICITYRural -2.332e+00 1.471e-01 -15.855 < 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: 5240.4 on 4530 degrees of freedom  
## Residual deviance: 4059.0 on 4507 degrees of freedom  
## AIC: 4107  
##   
## Number of Fisher Scoring iterations: 5

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| **0** | 1046 | 64 |
| **1** | 241 | 159 |

fitting null model for pseudo-r2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| llh | llhNull | G2 | McFadden | r2ML | r2CU |
| -2029 | -2620 | 1181 | 0.2254 | 0.2295 | 0.3349 |

Accuracy is 79.80% which is noticeably higher than a flip of a coin. McFadden is 0.2254. Area under the curve (see plot below) is 0.8022.



Results below are for K-fold cross validation using 10 iterations. They hold up against results observed with simple training.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1046 64  
## 1 241 159  
##   
## Accuracy : 0.798   
## 95% CI : (0.7769, 0.818)  
## No Information Rate : 0.8523   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3959   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.8127   
## Specificity : 0.7130   
## Pos Pred Value : 0.9423   
## Neg Pred Value : 0.3975   
## Prevalence : 0.8523   
## Detection Rate : 0.6927   
## Detection Prevalence : 0.7351   
## Balanced Accuracy : 0.7629   
##   
## 'Positive' Class : 0   
##

Analyzing deviance (using anova) shows that all terms are significant although significance for MSTATUS and TRAVTIME are less significant than all other terms.

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: TARGET\_FLAG  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 4530 5240.4   
## KIDSDRIV 1 31.64 4529 5208.8 1.860e-08 \*\*\*  
## log(INCOME + 1) 1 45.45 4528 5163.3 1.570e-11 \*\*\*  
## PARENT1 1 99.30 4527 5064.1 < 2.2e-16 \*\*\*  
## log(HOME\_VAL + 1) 1 45.03 4526 5019.0 1.938e-11 \*\*\*  
## MSTATUS 1 3.56 4525 5015.5 0.05907 .   
## EDUCATION 4 85.47 4521 4930.0 < 2.2e-16 \*\*\*  
## TRAVTIME 1 5.70 4520 4924.3 0.01697 \*   
## CAR\_USE 1 81.67 4519 4842.6 < 2.2e-16 \*\*\*  
## BLUEBOOK 1 38.64 4518 4804.0 5.089e-10 \*\*\*  
## TIF 1 27.91 4517 4776.1 1.272e-07 \*\*\*  
## CAR\_TYPE 5 67.15 4512 4708.9 4.014e-13 \*\*\*  
## OLDCLAIM 1 49.73 4511 4659.2 1.766e-12 \*\*\*  
## CLM\_FREQ 1 109.12 4510 4550.1 < 2.2e-16 \*\*\*  
## REVOKED 1 61.82 4509 4488.2 3.762e-15 \*\*\*  
## MVR\_PTS 1 69.84 4508 4418.4 < 2.2e-16 \*\*\*  
## URBANICITY 1 359.42 4507 4059.0 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Variance inflation factors are low for all variables, so correlation is not high.

## GVIF Df GVIF^(1/(2\*Df))  
## KIDSDRIV 1.082636 1 1.040498  
## log(INCOME + 1) 1.214555 1 1.102068  
## PARENT1 1.418362 1 1.190950  
## log(HOME\_VAL + 1) 1.622557 1 1.273796  
## MSTATUS 1.954360 1 1.397984  
## EDUCATION 1.345798 4 1.037821  
## TRAVTIME 1.033747 1 1.016734  
## CAR\_USE 1.527886 1 1.236077  
## BLUEBOOK 1.565014 1 1.251005  
## TIF 1.008322 1 1.004152  
## CAR\_TYPE 1.935421 5 1.068261  
## OLDCLAIM 1.677986 1 1.295371  
## CLM\_FREQ 1.449134 1 1.203800  
## REVOKED 1.339632 1 1.157425  
## MVR\_PTS 1.171422 1 1.082322  
## URBANICITY 1.135802 1 1.065740

#### Coefficient Analysis

Looking at the model coefficients, the following variables make the crash more likely:

* Higher number of children driving
* Single parents and unmarried individuals
* Higher commuting distance
* License revocation within the past 7 years
* Higher number of claims within the past 5 years
* Higher number of DMV points

Additional points:

* Commercial vehicle use is more likely to result in a crash than private.
* Rural environment is less likely to result in a crash than urban.
* Individuals who have been customers longer are less likely to have a crash.
* Higher Blue Book value makes it less likely to result in a crash.
* Education makes it less likely to result in a crash.
* Higher income and home value make it less likely to result in a crash.
* Car type is hard to evaluate, but generally various types make it more likely to have a crash (to different degrees).
* Interestingly, larger previous payout make it less likely to result in a crash.

Most of these findings are in line with theoretical effect. Some coefficients help clarify it.

## Modelling: Linear Model

Linear modelling is used to predict the amount of the payout in case of a crash (TARGET\_AMT). **Only observations where a crash has occurred (TARGET\_FLAG==1) are used in training the linear model.** If there is no crash, payout will not be needed. As such if observations without a crash are included in the model, they may skew the results.

Isolating all observations with a reported crash, leaves 1,601 observations.

Again the data is divided into a training set (75%; 1,200 observations) and a testing set (25%; 401 observations).

The following 5 models were built and compared.

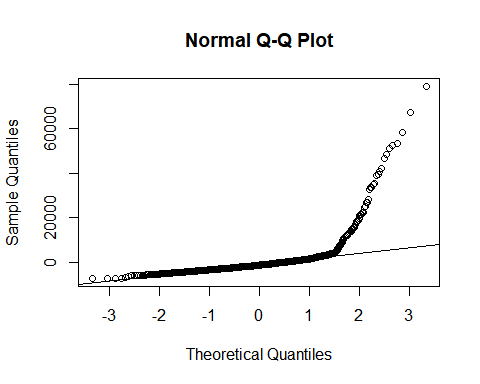
* **Model 1**: All independent variables.
* **Model 2**: Model 1 optimized using stepwise algorithm. This dropped all variables except PARENT1, MSTATUS, BLUEBOOK, OLDCLAIM, CLM\_FREQ, REVOKED and JOB.
* **Model 3**: Some variables from model 2 dropping less signficant ones. This model includes BLUEBOOK, OLDCLAIM, CLM\_FREQ and REVOKED.
* **Model 4**: Variables that theoretically should have an effect: BLUEBOOK, CAR\_AGE, CAR\_TYPE.
* **Model 5**: Only BLUEBOOK. This variable seems the most significant.

|  |  |  |
| --- | --- | --- |
| Model | Adjusted R^2 | Root-Mean Square Error |
| Model 1 | 0.01808 | 8416 |
| Model 2 | 0.02218 | 8367 |
| Model 3 | 0.01366 | 8344 |
| Model 4 | 0.01351 | 8373 |
| Model 5 | 0.01202 | 8336 |

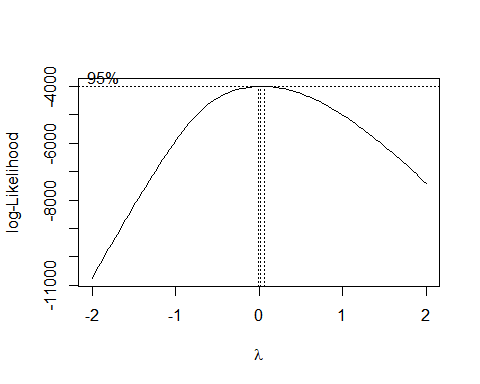
Even though Model 5 has the worst adjusted , it has the best RMSE value (accuracy in prediction using the testing set). It is also the simplest model using only Blue Book value. Further analysis will be based on this model.

##   
## Call:  
## lm(formula = TARGET\_AMT ~ BLUEBOOK, data = insLMtrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7307 -2864 -1349 483 79170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.041e+03 4.023e+02 10.044 < 2e-16 \*\*\*  
## BLUEBOOK 1.013e-01 2.566e-02 3.948 8.35e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7063 on 1198 degrees of freedom  
## Multiple R-squared: 0.01284, Adjusted R-squared: 0.01202   
## F-statistic: 15.58 on 1 and 1198 DF, p-value: 8.349e-05

Looking at the Q-Q plot, it is clear that there is a significant number of outliers at the higher range. This looks very problematic.



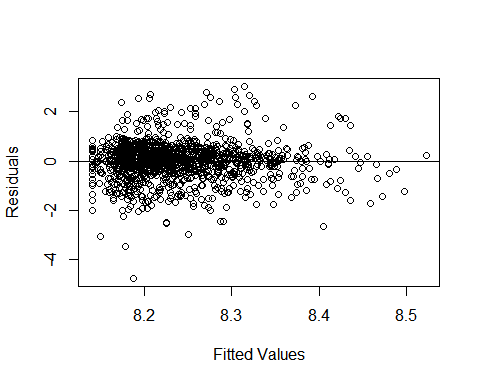
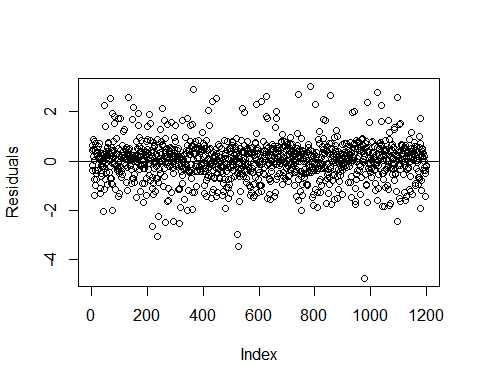
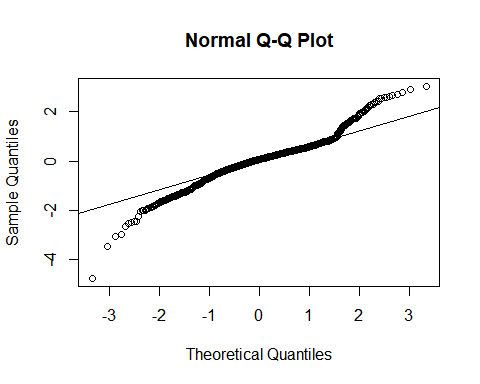
Consider Box-Cox transformation (plot below is generated using boxcox from the MASS package).



if lambda is picked to be 0, then target variable should be log-transformed.

##   
## Call:  
## lm(formula = log(TARGET\_AMT) ~ BLUEBOOK, data = insLMtrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7770 -0.3877 0.0599 0.4177 3.0424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.128e+00 4.592e-02 176.993 < 2e-16 \*\*\*  
## BLUEBOOK 8.166e-06 2.929e-06 2.788 0.00539 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8062 on 1198 degrees of freedom  
## Multiple R-squared: 0.006446, Adjusted R-squared: 0.005617   
## F-statistic: 7.772 on 1 and 1198 DF, p-value: 0.005389

This drastically reduced already low adjusted . However, Q-Q plot is improved. It is still not without issues, but noticeably better. Additionally, looking at the residual plots, there seems to be independence of observations. Another problem area is constant variance when plotting fitted values vs residuals.

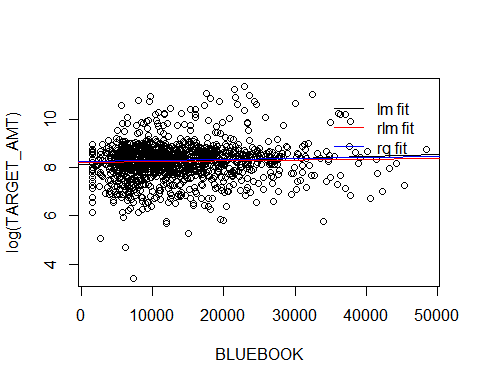


#### Robust/Quantile Regression

Looking at the scatterplot of BLUEBOOK vs log(TARGET\_AMT), there is a lot of variance and a lot of outliers. It is possible that some points are leverage points that interfere with the model. Two additional model were built in an effort to account for that.

The first model was created using robust regression (rlm in the MASS package). The second model was created using quantile regression (rq in the quantreg package).

Original model has RMSE of 0.7815. RLM model very slightly improves it to 0.7811. Finally, RQ model also very slightly improves it to 0.7771. Fits for all three models are presented in the scatterplot below.



Variation between model is small, so this approach did not generate significant improvement. Please see the Summary section in the beginning of the report for more analysis of linear model results.

## Evaluation Data Set

For illustration purposes, this report only includes the first 100 entries in the insurance-evaluation-data.csv file.

Evalution data is missing some INCOME and HOME\_EVAL values that are used in the binary regression model. Rather than trying to impute those values by replacing them with 0 or mean or media values or by building a model to predict them, these values are left as they are and the model cannot be used to predict the outcome for corresponding observations. Approach to imputing these values may change based on the situation and for this report, this part was not assumed. If any of these variables are likely to be missing in real-world use it may be worthwhile investigating omitting these variables from the model altogether.

|  |  |  |  |
| --- | --- | --- | --- |
| Index | TARGET\_FLAG Prob. | TARGET\_FLAG | TARGET\_AMT |
| 3 | 0.1816 | 0 | NA |
| 9 | 0.3604 | 0 | NA |
| 10 | 0.1493 | 0 | NA |
| 18 | 0.2319 | 0 | NA |
| 21 | 0.2756 | 0 | NA |
| 30 | NA | NA | NA |
| 31 | 0.4432 | 0 | NA |
| 37 | 0.4418 | 0 | NA |
| 39 | 0.0289 | 0 | NA |
| 47 | 0.1904 | 0 | NA |
| 60 | 0.0279 | 0 | NA |
| 62 | 0.5549 | 1 | 3812 |
| 63 | NA | NA | NA |
| 64 | 0.0792 | 0 | NA |
| 68 | 0.0261 | 0 | NA |
| 75 | NA | NA | NA |
| 76 | 0.671 | 1 | 3429 |
| 83 | 0.1779 | 0 | NA |
| 87 | 0.5449 | 1 | 3651 |
| 92 | 0.3123 | 0 | NA |
| 98 | 0.1521 | 0 | NA |
| 106 | 0.395 | 0 | NA |
| 107 | 0.0851 | 0 | NA |
| 113 | 0.3603 | 0 | NA |
| 120 | 0.2745 | 0 | NA |
| 123 | 0.4411 | 0 | NA |
| 125 | 0.4386 | 0 | NA |
| 126 | 0.3961 | 0 | NA |
| 128 | 0.1665 | 0 | NA |
| 129 | 0.1554 | 0 | NA |
| 131 | 0.221 | 0 | NA |
| 135 | 0.3263 | 0 | NA |
| 141 | 0.079 | 0 | NA |
| 147 | 0.1926 | 0 | NA |
| 148 | 0.1651 | 0 | NA |
| 151 | 0.0361 | 0 | NA |
| 156 | 0.1378 | 0 | NA |
| 157 | 0.1199 | 0 | NA |
| 174 | 0.0691 | 0 | NA |
| 186 | 0.5531 | 1 | 3668 |
| 193 | 0.2295 | 0 | NA |
| 195 | 0.5546 | 1 | 3831 |
| 212 | 0.0136 | 0 | NA |
| 213 | 0.5128 | 1 | 3920 |
| 217 | 0.0041 | 0 | NA |
| 223 | 0.2552 | 0 | NA |
| 226 | 0.1033 | 0 | NA |
| 228 | 0.2615 | 0 | NA |
| 230 | 0.0138 | 0 | NA |
| 241 | 0.5685 | 1 | 3610 |
| 243 | 0.1487 | 0 | NA |
| 249 | 0.3335 | 0 | NA |
| 281 | 0.7619 | 1 | 3691 |
| 288 | 0.1072 | 0 | NA |
| 294 | NA | NA | NA |
| 295 | 0.2027 | 0 | NA |
| 300 | 0.4442 | 0 | NA |
| 302 | 0.3606 | 0 | NA |
| 303 | 0.1077 | 0 | NA |
| 308 | 0.5506 | 1 | 3594 |
| 319 | 0.0118 | 0 | NA |
| 320 | 0.0944 | 0 | NA |
| 324 | 0.3479 | 0 | NA |
| 331 | 0.2167 | 0 | NA |
| 343 | 0.0511 | 0 | NA |
| 347 | 0.3246 | 0 | NA |
| 348 | 0.8159 | 1 | 3531 |
| 350 | 0.568 | 1 | 3967 |
| 357 | 0.1465 | 0 | NA |
| 358 | 0.058 | 0 | NA |
| 360 | NA | NA | NA |
| 366 | 0.1974 | 0 | NA |
| 367 | 0.7401 | 1 | 3580 |
| 368 | 0.2771 | 0 | NA |
| 376 | 0.7287 | 1 | 3645 |
| 380 | 0.3893 | 0 | NA |
| 388 | 0.3897 | 0 | NA |
| 396 | 0.252 | 0 | NA |
| 398 | 0.1245 | 0 | NA |
| 403 | 0.0448 | 0 | NA |
| 410 | 0.5654 | 1 | 3598 |
| 412 | 0.34 | 0 | NA |
| 420 | 0.2981 | 0 | NA |
| 434 | 0.0345 | 0 | NA |
| 440 | 0.4766 | 0 | NA |
| 450 | 0.5612 | 1 | 3887 |
| 453 | 0.2757 | 0 | NA |
| 464 | 0.2841 | 0 | NA |
| 465 | 0.0522 | 0 | NA |
| 466 | NA | NA | NA |
| 473 | 0.0847 | 0 | NA |
| 476 | 0.0903 | 0 | NA |
| 478 | NA | NA | NA |
| 479 | 0.1898 | 0 | NA |
| 493 | 0.0347 | 0 | NA |
| 497 | 0.2245 | 0 | NA |
| 503 | 0.0068 | 0 | NA |
| 504 | 0.3887 | 0 | NA |
| 505 | 0.3087 | 0 | NA |
| 507 | 0.2996 | 0 | NA |

## APPENDIX : R Script

# Required libraries  
library(knitr)  
library(kableExtra)  
library(gridExtra)  
library(ggplot2)  
library(dplyr)  
library(caTools)  
library(pscl)  
library(ROCR)  
library(MASS)  
library(caret)  
library(car)  
library(Metrics)  
library(quantreg)  
# Import data  
ins <- read.csv(url(paste0("https://raw.githubusercontent.com/omerozeren/DATA621/master/insurance\_training\_data.csv")),  
 na.strings=c("","NA"))  
# Basic statistic  
nrow(ins); ncol(ins)  
summary(ins)  
# TARGET\_FLAG - 6008 are 0, 2153 are 1  
table(ins$TARGET\_FLAG)  
class(ins$TARGET\_FLAG)   
# Integer (0/1)  
# TARGET\_AMT  
summary(ins[ins$TARGET\_FLAG==0, 'TARGET\_AMT'])  
summary(ins[ins$TARGET\_FLAG==1, 'TARGET\_AMT'])  
class(ins$TARGET\_AMT)   
# Only available for TARGET\_FLAG=1  
# Numeric: Ranges from 30.28 to 107600  
# KIDSDRIV - No of Driving Children  
table(ins$KIDSDRIV)  
class(ins$KIDSDRIV)  
# Integer - Ranges from 0 to 4  
# AGE  
summary(ins$AGE)  
class(ins$AGE)  
# Integer - Ranges from 16 to 81  
# 6 NAs  
# HOMEKIDS  
table(ins$HOMEKIDS)  
class(ins$HOMEKIDS)  
# Integer - Ranges from 0 to 5  
# YOJ - Years on Job  
summary(ins$YOJ)  
class(ins$YOJ)  
# Integer - Ranges from 0 to 23  
# 454 NAs - 5.56% of observations  
# INCOME  
class(ins$INCOME)  
summary(ins$INCOME)  
# Convert to Numeric - Ranges from $0 to $367,000  
# 445 NAs - 5.45% of observations  
# PARENT1 - Single Parent?  
table(ins$PARENT1)  
class(ins$PARENT1); levels(ins$PARENT1)  
# Factor - No, Yes  
# HOME\_VAL - Home Value  
class(ins$HOME\_VAL)  
summary(ins$HOME\_VAL)  
# Converted to Numeric - Ranges from $0 to $885,300  
# 464 NAs - 5.69% of observations  
# MSTATUS  
table(ins$MSTATUS)  
class(ins$MSTATUS); levels(ins$MSTATUS)  
# Factor - Yes, No  
# SEX  
table(ins$SEX)  
class(ins$SEX); levels(ins$SEX)  
# Factor - M, F  
# EDUCATION  
table(ins$EDUCATION)  
class(ins$EDUCATION); levels(ins$EDUCATION)  
# Factor - <HS, HS, BA, MA, PhD  
# JOB  
table(ins$JOB)  
class(ins$JOB); levels(ins$JOB)  
# Factor - [Blank], Clerical, Doctor, Home Maker, Lawyer, Manager,   
# Professional, Student, Blue Collar  
# TRAVTIME - Distance to work  
summary(ins$TRAVTIME)  
class(ins$TRAVTIME)  
# Integer - Ranges from 5 to 142  
# CAR\_USE  
class(ins$CAR\_USE); levels(ins$CAR\_USE)  
table(ins$CAR\_USE)  
# Factor - Commercial, Private  
# BLUEBOOK  
class(ins$BLUEBOOK)  
summary(ins$BLUEBOOK)  
# Numeric - Ranges from $1,500 to $69,740  
# TIF - Time in Force  
class(ins$TIF)  
summary(ins$TIF)  
# Integer - Ranges from 1 to 25  
# CAR\_TYPE  
class(ins$CAR\_TYPE); levels(ins$CAR\_TYPE)  
table(ins$CAR\_TYPE)  
# FActor - Minivan, Panel Truck, Pickup, Sports Car, Van, SUV  
# RED\_CAR  
class(ins$RED\_CAR); levels(ins$RED\_CAR)  
table(ins$RED\_CAR)  
# Factor - No, Yes  
# OLDCLAIM  
class(ins$OLDCLAIM)  
summary(ins$OLDCLAIM)  
# Numeric - Ranges from $0 to $57,040  
# CLM\_FREQ  
class(ins$CLM\_FREQ)  
summary(ins$CLM\_FREQ)  
# Integer - Ranges from 0 to 5  
# REVOKED  
class(ins$REVOKED); levels(ins$REVOKED)  
table(ins$REVOKED)  
# Factor - No, Yes  
# MVR\_PTS  
class(ins$MVR\_PTS)  
summary(ins$MVR\_PTS)  
# Integer - Ranges from 0 to 13  
# CAR\_AGE  
class(ins$CAR\_AGE)  
summary(ins$CAR\_AGE)  
nrow(ins[ins$CAR\_AGE<0 & !is.na(ins$CAR\_AGE), ])  
nrow(ins[ins$CAR\_AGE==0 & !is.na(ins$CAR\_AGE), ])  
nrow(ins[ins$CAR\_AGE==1 & !is.na(ins$CAR\_AGE), ])  
# Integer - Ranges from -3 to 28  
# 1 observation of -3 - invalid  
# 3 observations of 0 - likely invalid  
# 1,934 observations of 1 - reasonable (new car)  
# 510 NAs - 6.25% of observations  
# URBANICITY  
class(ins$URBANICITY); levels(ins$URBANICITY)  
table(ins$URBANICITY)  
# Factor - Urban, Rural  
ins$INCOME <- as.numeric(gsub('[$,]', '', ins$INCOME))  
ins$HOME\_VAL <- as.numeric(gsub('[$,]', '', ins$HOME\_VAL))  
levels(ins$MSTATUS)[match("z\_No",levels(ins$MSTATUS))] <- "No"  
levels(ins$SEX)[match("z\_F",levels(ins$SEX))] <- "F"  
levels(ins$EDUCATION)[match("z\_High School",  
 levels(ins$EDUCATION))] <- "High School"  
ins$EDUCATION <- factor(ins$EDUCATION,levels(ins$EDUCATION)[c(1,5,2:4)])  
levels(ins$JOB)[match("z\_Blue Collar",levels(ins$JOB))] <- "Blue Collar"  
ins$BLUEBOOK <- as.numeric(gsub('[$,]', '', ins$BLUEBOOK))  
levels(ins$CAR\_TYPE)[match("z\_SUV",levels(ins$CAR\_TYPE))] <- "SUV"  
levels(ins$RED\_CAR)[match("no",levels(ins$RED\_CAR))] <- "No"  
levels(ins$RED\_CAR)[match("yes",levels(ins$RED\_CAR))] <- "Yes"  
ins$OLDCLAIM <- as.numeric(gsub('[$,]', '', ins$OLDCLAIM))  
levels(ins$URBANICITY)[match("Highly Urban/ Urban",  
 levels(ins$URBANICITY))] <- "Urban"  
levels(ins$URBANICITY)[match("z\_Highly Rural/ Rural",  
 levels(ins$URBANICITY))] <- "Rural"  
ins$JOB <- factor(ins$JOB,levels(ins$JOB)[c(7, 8, 3, 1, 6, 5, 4, 2)])  
# Drop index column  
ins <- ins[-c(1)]  
insFull <- ins  
ins <- ins[complete.cases(ins), ]  
ins[ins$CAR\_AGE<1,'CAR\_AGE'] <- NA  
ins <- ins[complete.cases(ins), ]  
# Cuts down from 8,161 to 6,045  
# Get only complete cases  
nrow(ins[complete.cases(ins), ])  
nrow(ins)  
insBackup <- ins  
# Summary table  
sumIns = data.frame(Variable = character(),  
 Min = integer(),  
 Median = integer(),  
 Mean = double(),  
 SD = double(),  
 Max = integer(),  
 Num\_NAs = integer(),  
 Num\_Zeros = integer())  
for (i in c(3:7,9,14,16,17,20,21,23,24)) {  
 sumIns <- rbind(sumIns, data.frame(Variable = colnames(ins)[i],  
 Min = min(ins[,i], na.rm=TRUE),  
 Median = median(ins[,i], na.rm=TRUE),  
 Mean = mean(ins[,i], na.rm=TRUE),  
 SD = sd(ins[,i], na.rm=TRUE),  
 Max = max(ins[,i], na.rm=TRUE),  
 Num\_NAs = sum(is.na(ins[,i])),  
 Num\_Zeros = length(which(ins[,i]==0)))  
 )  
}  
colnames(sumIns) <- c("", "Min", "Median", "Mean", "SD", "Max", "Num of NAs",   
 "Num of Zeros")  
sumIns  
# Proportion of target variable  
table(ins$TARGET\_FLAG)  
table(ins$TARGET\_FLAG)/sum(table(ins$TARGET\_FLAG))  
# Exploratory plots (repeated for each variable)  
# Get descriptive plots:  
# Variables:   
# INDEX, TARGET\_FLAG, TARGET\_AMT, KIDSDRIV, AGE, HOMEKIDS, YOJ, INCOME,   
# PARENT1, HOME\_VAL, MSTATUS, SEX, EDUCATION, JOB, TRAVTIME, CAR\_USE,   
# BLUEBOOK, TIF, CAR\_TYPE, RED\_CAR, OLDCLAIM, CLM\_FREQ, REVOKED, MVR\_PTS,   
# CAR\_AGE, URBANICITY,   
v <- "TARGET\_AMT" # Variable to view  
pd <- as.data.frame(cbind(ins[, v], ins$TARGET\_FLAG))  
colnames(pd) <- c("X", "Y")  
# Boxplot  
bp <- ggplot(pd, aes(x = 1, y = X)) +   
 stat\_boxplot(geom ='errorbar') + geom\_boxplot() +   
 xlab("Boxplot") + ylab("") + theme(axis.text.x=element\_blank(),   
 axis.ticks.x=element\_blank())  
# Density plot  
hp <- ggplot(pd, aes(x = X)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="white") +  
 geom\_density(alpha=.2, fill="#FF6666") +   
 ylab("") + xlab("Density Plot with Mean") +  
 geom\_vline(aes(xintercept=mean(X, na.rm=TRUE)), color="red",   
 linetype="dashed", size=1)  
# Scatterplot  
sp <- ggplot(pd, aes(x=X, y=Y)) +   
 geom\_point() +   
 stat\_smooth(method="glm", method.args=list(family="binomial"), se=FALSE) +  
 xlab("Scatterplot with Logistic Regression Line")  
grid.arrange(bp, hp, sp, layout\_matrix=rbind(c(1,2,2),c(1,3,3)))  
# Correlation matrix  
cm <- cor(ins, use="pairwise.complete.obs")  
cm <- round(cm, 2)  
cmout <- as.data.frame(cm) %>% mutate\_all(function(x) {  
 cell\_spec(x, "html", color = ifelse(x>0.5 | x<(-0.5),"blue","black"))  
 })  
rownames(cmout) <- colnames(cmout)  
cmout %>%  
 kable("html", escape = F, align = "c", row.names = TRUE) %>%  
 kable\_styling("striped", full\_width = F)  
pairs(ins)  
# Split into train and validation sets  
set.seed(88)  
split <- sample.split(ins$TARGET\_FLAG, SplitRatio = 0.75)  
insTRAIN <- subset(ins, split == TRUE)  
insTEST <- subset(ins, split == FALSE)  
# BINARY REGRESSION MODEL  
# Modelling - Basic model  
model <- glm (TARGET\_FLAG ~ .-TARGET\_AMT, data = insTRAIN,   
 family = binomial(link="logit"))  
summary(model)  
pred <- predict(model, newdata=subset(insTEST, select=c(1:25)),   
 type='response')  
cm <- confusionMatrix(as.factor(insTEST$TARGET\_FLAG),   
 as.factor(ifelse(pred > 0.5,1,0)))  
cm$table  
cm$overall['Accuracy']  
pR2(model) # McFadden R^2  
# Stepwise approach  
model <- stepAIC(model, trace=FALSE, direction='both')  
# Model tweaking  
model <- glm(formula = TARGET\_FLAG ~ KIDSDRIV + log(INCOME+1) + PARENT1 +   
 log(HOME\_VAL+1) + MSTATUS + EDUCATION + JOB + TRAVTIME +   
 CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM + CLM\_FREQ +   
 REVOKED + MVR\_PTS + URBANICITY,   
 family = binomial(link = "logit"), data = insTRAIN)  
# ROC  
pr <- prediction(pred, insTEST$TARGET\_FLAG)  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf, colorize = TRUE, text.adj = c(-0.2,1.7))  
auc <- performance(pr, measure = "auc")  
(auc <- auc@y.values[[1]])  
# K-Fold cross validation  
ctrl <- trainControl(method = "repeatedcv", number = 10,   
 savePredictions = TRUE)  
model\_fit <- train(TARGET\_FLAG ~ KIDSDRIV + log(INCOME+1) + PARENT1 +   
 log(HOME\_VAL+1) + MSTATUS + EDUCATION + TRAVTIME +   
 CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM +   
 CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY,   
 data=insTRAIN, method="glm", family="binomial",  
 trControl = ctrl, tuneLength = 5)  
pred <- predict(model\_fit, newdata=insTEST)  
confusionMatrix(as.factor(insTEST$TARGET\_FLAG),  
 as.factor(ifelse(pred > 0.5,1,0)))  
# Deviance residuals  
anova(model, test="Chisq")  
# VIF  
vif(model)  
# Take out JOB   
ggplot(data = ins, aes(JOB, EDUCATION)) +  
 geom\_jitter()  
model <- glm(formula = TARGET\_FLAG ~ KIDSDRIV + log(INCOME+1) + PARENT1 +   
 log(HOME\_VAL+1) + MSTATUS + EDUCATION + TRAVTIME + CAR\_USE +   
 BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM + CLM\_FREQ + REVOKED +   
 MVR\_PTS + URBANICITY,   
 family = binomial(link = "logit"), data = insTRAIN)  
# LINEAR MODEL  
insLM <- ins  
insLM <- ins[ins$TARGET\_FLAG==1,]  
# Split into training and testing sets  
split <- sample.split(insLM$TARGET\_AMT, SplitRatio = 0.75)  
insLMtrain <- subset(insLM, split == TRUE)  
insLMtest <- subset(insLM, split == FALSE)  
# Initial models  
lmModel <- lm(TARGET\_AMT ~ .-TARGET\_FLAG,data = insLMtrain)  
summary(lmModel)  
lmModel <- stepAIC(lmModel, trace=FALSE, direction='both')  
summary(lmModel)  
lmModel <- lm(TARGET\_AMT ~ PARENT1 + MSTATUS + BLUEBOOK + OLDCLAIM +   
 CLM\_FREQ + REVOKED + JOB,data = insLMtrain)  
summary(lmModel)  
lmModel <- lm(TARGET\_AMT ~ BLUEBOOK + OLDCLAIM + CLM\_FREQ + REVOKED,  
 data = insLMtrain)  
summary(lmModel)  
lmModel <- lm(TARGET\_AMT ~ BLUEBOOK + CAR\_AGE + CAR\_TYPE,data = insLMtrain)  
summary(lmModel)  
lmModel <- lm(TARGET\_AMT ~ BLUEBOOK,data = insLMtrain)  
summary(lmModel)  
# Calculate RMSE  
pred <- predict(lmModel, newdata=insLMtest)  
rmse(insLMtest$TARGET\_AMT, pred)  
# Model plots  
plot(lmModel$residuals, ylab="Residuals")  
abline(h=0)  
plot(lmModel$fitted.values, lmModel$residuals,   
 xlab="Fitted Values", ylab="Residuals")  
abline(h=0)  
qqnorm(lmModel$residuals)  
qqline(lmModel$residuals)  
boxcox(lmModel)  
lmModel <- lm(log(TARGET\_AMT) ~ BLUEBOOK, data = insLMtrain)  
summary(lmModel)  
pred <- predict(lmModel, newdata=insLMtest)  
rmse(log(insLMtest$TARGET\_AMT), pred)  
lmModel2 <- rlm(log(TARGET\_AMT) ~ BLUEBOOK, data = insLMtrain)  
summary(lmModel2)  
pred <- predict(lmModel2, newdata=insLMtest)  
rmse(log(insLMtest$TARGET\_AMT), pred)  
lmModel3 <- rq(log(TARGET\_AMT) ~ BLUEBOOK, data = insLMtrain)  
summary(lmModel3)  
pred <- predict(lmModel3, newdata=insLMtest)  
rmse(log(insLMtest$TARGET\_AMT), pred)  
plot(log(TARGET\_AMT) ~ BLUEBOOK, data = insLMtrain)  
abline(lmModel)  
abline(lmModel2, col="red")  
abline(lmModel3, col="blue")  
legend("topright", inset=0.05, bty="n",  
 legend=c("lm fit", "rlm fit", "rq fit"),  
 lty=c(1,1,1),  
 col=c("black", "red", "blue"))  
# Prediction  
eval <- read.csv(url(paste0("https://raw.githubusercontent.com/omerozeren/DATA621/master/insurance-evaluation-data.csv")),  
 na.strings=c("","NA"))  
results <- eval[,1]  
eval$INCOME <- as.numeric(gsub('[$,]', '', eval$INCOME))  
eval$HOME\_VAL <- as.numeric(gsub('[$,]', '', eval$HOME\_VAL))  
levels(eval$MSTATUS)[match("z\_No",levels(eval$MSTATUS))] <- "No"  
levels(eval$SEX)[match("z\_F",levels(eval$SEX))] <- "F"  
levels(eval$EDUCATION)[match("z\_High School",  
 levels(eval$EDUCATION))] <- "High School"  
eval$EDUCATION <- factor(eval$EDUCATION,levels(eval$EDUCATION)[c(1,5,2:4)])  
levels(eval$JOB)[match("z\_Blue Collar",levels(eval$JOB))] <- "Blue Collar"  
eval$BLUEBOOK <- as.numeric(gsub('[$,]', '', eval$BLUEBOOK))  
levels(eval$CAR\_TYPE)[match("z\_SUV",levels(eval$CAR\_TYPE))] <- "SUV"  
levels(eval$RED\_CAR)[match("no",levels(eval$RED\_CAR))] <- "No"  
levels(eval$RED\_CAR)[match("yes",levels(eval$RED\_CAR))] <- "Yes"  
eval$OLDCLAIM <- as.numeric(gsub('[$,]', '', eval$OLDCLAIM))  
levels(eval$URBANICITY)[match("Highly Urban/ Urban",  
 levels(eval$URBANICITY))] <- "Urban"  
levels(eval$URBANICITY)[match("z\_Highly Rural/ Rural",  
 levels(eval$URBANICITY))] <- "Rural"  
eval$JOB <- factor(eval$JOB,levels(eval$JOB)[c(7, 8, 3, 1, 6, 5, 4, 2)])  
eval <- eval[-c(1)]  
pred <- predict(model, newdata=eval, type="response")  
results <- cbind(results, prob=round(pred,4))  
results <- cbind(results, predict=round(pred,0))  
pred <- predict(lmModel, newdata=eval, type="response")  
results <- cbind(results, exp(pred))  
results <- as.data.frame(results)  
results[results$predict==0 & !is.na(results$predict),'V4'] <- NA  
results[is.na(results$predict),'V4'] <- NA  
colnames(results) <- c("Index", "TARGET\_FLAG Prob.",   
 "TARGET\_FLAG", "TARGET\_AMT")  
pander(head(results, 100))