Auto Insurance Data Assignment

Cesar Espitia

CUNY SPS Data 621

Table of Contents

[Abstract 3](#_Toc518824488)

[Auto Insurance Data Assignment 4](#_Toc518824489)

[Data Exploration 4](#_Toc518824490)

[Summary Statistics 4](#_Toc518824491)

[Data Preparation 6](#_Toc518824492)

[Model Building for Outcome Variable TARGET\_FLAG 8](#_Toc518824493)

[Model Building for Outcome Variable TARGET\_AMT 13](#_Toc518824494)

[Conclusion 20](#_Toc518824495)

[Appendix A: R Code 21](#_Toc518824496)

[Appendix B: CORRELATION MATRIX 23](#_Toc518824497)

Abstract

This assignment focused on analyzing data from an insurance auto company. The dataset contains over 8,000 records that encompass their policy holders. The data set has 26 variables, 2 outcome and 24 predictor, of different types such as continuous or factor type variables. The purpose for this assignment is to analyze the data, perform any data manipulation / clean-up and build three (3) binary logistic regression and three (3) multiple linear regression models using only the data (or derivatives thereof) to predict if the region is above or below the median crime rate. The chosen model provided an AIC **=** 7384.4 and R2 = 0.2879.

Keywords: insurance, data621

Auto Insurance Data Assignment

The following is the analysis and write-up based upon my interpretation of the data and predict if an individual is likely to have an accident, and then if they do, what the claim amount may be.

# Data Exploration

The purpose of this step is to get a ‘feel’ for the dataset. The following information describes the data from different angles including completeness, statistical summaries, visuals to determine the shape and effect of each variable and other items deemed pertinent.

## Summary Statistics

The first step is to look at the data to determine some items including completeness and the shape of each variable. The following are the results of summarizing the data in a table and the visualization of each variables density function (PDF).

Table 1

Summary Statistics for Moneyball Training Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VARIABLE | Min | 1Q | Median | Mean | 3Q | Max | NA |
| INDEX | 1 | 2559 | 5133 | 5152 | 7745 | 10302 |  |
| TARGET\_FLAG | 4.17222222 | 1.53680556 |  |  |  |  |  |
| TARGET\_AMT | 0 | 0 | 0 | 1504 | 1036 | 107586 |  |
| KIDSDRIV | 0 | 0 | 0 | 0.1711 | 0 | 4 |  |
| AGE | 16 | 39 | 45 | 44.79 | 51 | 81 | 6 |
| HOMEKIDS | 0 | 0 | 0 | 0.7212 | 1 | 5 |  |
| YOJ | 0 | 9 | 11 | 10.5 | 13 | 23 | 454 |
| INCOME | 0 | 28097 | 54028 | 61898 | 85986 | 367030 | 445 |
| PARENT1 | No:7084 | Yes:1077 |  |  |  |  |  |
| HOME\_VAL | 0 | 0 | 161160 | 154867 | 238724 | 885282 | 464 |
| MSTATUS | Yes:4894 | z\_No:3267 |  |  |  |  |  |
| SEX | M:3786 | z\_F:4375 |  |  |  |  |  |
| EDUCATION | <High\_School:1203 | Bachelors:2242 | Masters:1658 | PhD:728 | z\_High\_School:2330 | |  |
| JOB | z\_Blue\_Collar:1825 | Clerical:1271 | Professiol:1117 | Mager:988 | Lawyer:835 | Student:712 | (Other):1413 |
| TRAVTIME | 5 | 22 | 33 | 33.49 | 44 | 142 |  |
| CAR\_USE | Commercial:3029 | Private:5132 | |  |  |  |  |
| BLUEBOOK | 1500 | 9280 | 14440 | 15710 | 20850 | 69740 |  |
| TIF | 1 | 1 | 4 | 5.351 | 7 | 25 |  |
| CAR\_TYPE | Minivan:2145 | Panel\_Truck:676 | Pickup:1389 | Sports\_Car:907 | Van:750 | z\_SUV:2294 |  |
| RED\_CAR | no:5783 | yes:2378 |  |  |  |  |  |
| OLDCLAIM | 0 | 0 | 0 | 4037 | 4636 | 57037 |  |
| CLM\_FREQ | 0 | 0 | 0 | 0.7986 | 2 | 5 |  |
| REVOKED | No:7161 | Yes:1000 |  |  |  |  |  |
| MVR\_PTS | 0 | 0 | 1 | 1.696 | 3 | 13 |  |
| CAR\_AGE | -3 | 1 | 8 | 8.328 | 12 | 28 | 510 |
| URBANICITY | Highly\_Urban/Urban:6492 | z\_Highly\_Rural/Rural:1669 | | |  |  |  |

Note: Source: insurance-training-data.csv

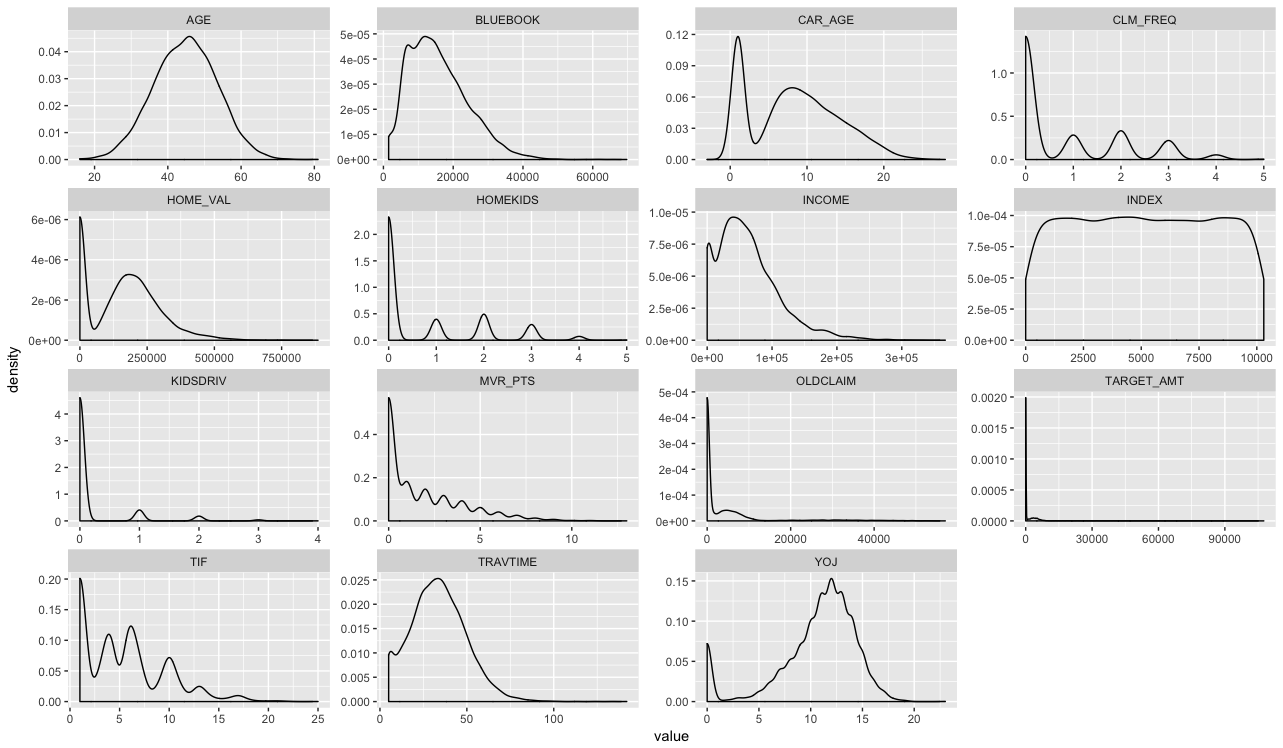
**

Figure 1. PDF for Each Dataframe Variable.

In looking at both, Table 1, Figure 1 and Appendix B (correlation matrix) together, we can note specific items that may skew our model building results.

*NA:* These incomplete cases will cause any correlation exercise to be incorrect or not possible. There are a few ways to deal with NAs including imputing the missing data or ignoring the variable altogether. For the purposes of this analysis, the variable CAR\_AGE, INCOME, HOME\_VALUE and AGE have missing information. The highest offender is CAR\_AGE with about 5% of the data missing while others are much lower than that.

*PDF:*Figure 1 shows the PDF of each variable, this allows us to see if the data is normal or not. For the numeric variables, four (4) variables (AGE, YOJ, TRAVTIME and INCOME) shows the typical normal density function but all others like ***CLM\_FREQ*** show left skewness and others show bimodality (CAR\_AGE). For the purposes of this analysis, the variables HOMEKIDS, MVR\_PTS, OLDCLAIM, TIF, KIDSDRIVE and CLM\_FREQ will be log transformed to remove the effects of skewness. All other variables were left as is because the shape didn’t warrant it.

*Correlation:* We look for correlated variables that we can make decisions on and determine which variable might be closely related to others either due to collinearity or other underlying factors that are visible at first glance in the dataset. Correlated variables bloat the model and don’t produce any more insight than ignoring one of the two that show correlation. In our data, none of the variables show any particular correlation that would be cause for alarm and would require removal in order to avoid collinearity.

# Data Preparation

The purpose of this step is to take the findings from the exploration and transform the data as needed. The following information describes the transformations done in order to prepare the data for model building and model selection.

*NA:* All missing values were imputed using the mean within each column even though it is not the most adequate for this data. The nearest neighbor method would have been more valid by using another variable to bin, but there was concern about causing bias due to calculating the mean on variables that have inherit bias in them (such as education). Therefore, the mean for the entire dataset (excluding NA values) was used for this analysis.

*Log Transformation:* For this dataset, six (6) variables were transformed that were deemed overtly skewed in comparison to other variables in the dataset. HOMEKIDS, MVR\_PTS, OLDCLAIM, TIF, KIDSDRIVE and CLM\_FREQ were the variables transformed using the log base 10 function, for example for the TIF column the transformation was *log(train$* TIF*+1)*. The value 1 was added in order to ensure that values of 0 continue to be 0 after the transformation (as log10(0) is not possible).

*Variable Creation:* For this dataset, no new variables were created. There was an option use binning to create a new variable such as the CAR\_AGE where flagging new cars as 1 vs those older than a specific amount of years as 0 could have been done but was decided against. There are enough varied variables in the data set to see how the models behave.

Correlation Check: Once these manipulations are done, a side-by-side comparison of the correlations matrix is done to ensure no inadvertent effects to the data.

|  |  |
| --- | --- |
|  |  |
| Before | After |

Figure 2. Correlation Comparison Before and After.

As can be noted, there were no real strong correlations before and no correlations after that warrant removal of any variables.

# Model Building for Outcome Variable TARGET\_FLAG

The purpose of this step is to take the modified dataset and begin exploring potential models that will be used on the final dataset provided. The following information describes the three (3) models built for this step and the relevant analysis to provide reasons for model selection in the next step.

**MODEL 1**

The first model takes in the data as manipulated in step two. In this first model, we have an AIC of 7384.4. The data in Table 2, shows that the model has an accuracy of 79.3%.

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5262 -0.7180 -0.3983 0.6545 3.1455

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -7.942e-01 3.293e-01 -2.412 0.015880 \*

KIDSDRIV 6.821e-01 1.103e-01 6.185 6.21e-10 \*\*\*

AGE 4.736e-05 4.078e-03 0.012 0.990734

HOMEKIDS 1.513e-01 8.300e-02 1.823 0.068320 .

YOJ -1.353e-02 8.578e-03 -1.577 0.114756

INCOME -3.457e-06 1.076e-06 -3.212 0.001317 \*\*

PARENT1Yes 3.295e-01 1.144e-01 2.881 0.003970 \*\*

HOME\_VAL -1.323e-06 3.419e-07 -3.871 0.000109 \*\*\*

MSTATUSz\_No 5.146e-01 8.493e-02 6.059 1.37e-09 \*\*\*

SEXz\_F -8.929e-02 1.120e-01 -0.797 0.425327

EDUCATIONBachelors -3.720e-01 1.154e-01 -3.223 0.001267 \*\*

EDUCATIONMasters -2.803e-01 1.785e-01 -1.570 0.116405

EDUCATIONPhD -1.496e-01 2.135e-01 -0.701 0.483401

EDUCATIONz\_High\_School 2.111e-02 9.487e-02 0.222 0.823945

JOBClerical 3.986e-01 1.963e-01 2.030 0.042359 \*

JOBDoctor -4.227e-01 2.662e-01 -1.588 0.112286

JOBHome\_Maker 2.049e-01 2.099e-01 0.976 0.328988

JOBLawyer 1.172e-01 1.693e-01 0.692 0.488652

JOBManager -5.616e-01 1.712e-01 -3.280 0.001038 \*\*

JOBProfessional 1.673e-01 1.782e-01 0.939 0.347724

JOBStudent 2.038e-01 2.140e-01 0.953 0.340799

JOBz\_Blue\_Collar 3.101e-01 1.853e-01 1.674 0.094190 .

TRAVTIME 1.483e-02 1.880e-03 7.890 3.02e-15 \*\*\*

CAR\_USEPrivate -7.604e-01 9.172e-02 -8.291 < 2e-16 \*\*\*

BLUEBOOK -2.079e-05 5.255e-06 -3.956 7.63e-05 \*\*\*

TIF -3.257e-01 4.138e-02 -7.869 3.56e-15 \*\*\*

CAR\_TYPEPanel\_Truck 5.701e-01 1.613e-01 3.533 0.000410 \*\*\*

CAR\_TYPEPickup 5.578e-01 1.007e-01 5.540 3.03e-08 \*\*\*

CAR\_TYPESports\_Car 1.031e+00 1.298e-01 7.942 2.00e-15 \*\*\*

CAR\_TYPEVan 6.158e-01 1.264e-01 4.872 1.10e-06 \*\*\*

CAR\_TYPEz\_SUV 7.787e-01 1.111e-01 7.007 2.43e-12 \*\*\*

RED\_CARyes -5.766e-03 8.631e-02 -0.067 0.946741

OLDCLAIM 6.763e-03 1.697e-02 0.398 0.690300

CLM\_FREQ 3.160e-01 1.277e-01 2.474 0.013363 \*

REVOKEDYes 7.242e-01 8.184e-02 8.850 < 2e-16 \*\*\*

MVR\_PTS 2.808e-01 4.202e-02 6.682 2.35e-11 \*\*\*

CAR\_AGE -1.807e-03 7.530e-03 -0.240 0.810372

URBANICITYz\_Highly\_Rural/ Rural -2.371e+00 1.130e-01 -20.989 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 9418.0 on 8160 degrees of freedom

Residual deviance: 7308.4 on 8123 degrees of freedom

AIC: 7384.4

Number of Fisher Scoring iterations: 5

Variable Interpretation:

|  |  |  |  |
| --- | --- | --- | --- |
| (Intercept) | -0.1158016 | JOBDoctor | -0.06163603 |
| KIDSDRIV | 0.09945167 | JOBHome\_Maker | 0.02986941 |
| AGE | 6.90481E-06 | JOBLawyer | 0.01709406 |
| HOMEKIDS | 0.02206017 | JOBManager | -0.08188439 |
| YOJ | -0.001972543 | JOBProfessional | 0.0243965 |
| INCOME | -5.04006E-07 | JOBStudent | 0.02972047 |
| PARENT1Yes | 0.04803969 | JOBz\_Blue\_Collar | 0.04522137 |
| HOME\_VAL | -1.92952E-07 | TRAVTIME | 0.00216305 |
| MSTATUSz\_No | 0.07503592 | CAR\_USEPrivate | -0.1108755 |
| SEXz\_F | -0.0130199 | BLUEBOOK | -3.03074E-06 |
| EDUCATIONBachelors | -0.05424472 | TIF | -0.04748519 |
| EDUCATIONMasters | -0.04086324 | CAR\_TYPEPanel\_Truck | 0.08311836 |
| EDUCATIONPhD | -0.02181469 | CAR\_TYPEPickup | 0.08132789 |
| EDUCATIONz\_High\_School | 0.003077558 | CAR\_TYPESports\_Car | 0.1502692 |
| JOBClerical | 0.05811519 | CAR\_TYPEVan | 0.0897826 |
| CAR\_TYPEz\_SUV | 0.1135413 | MVR\_PTS | 0.04094471 |
| RED\_CARyes | -0.000840661 | CAR\_AGE | -0.000263433 |
| OLDCLAIM | 0.000986117 | URBANICITYz\_Highly\_Rural/Rural | -0.3457765 |
| CLM\_FREQ | 0.04607979 |  |  |
| REVOKEDYes | 0.1055966 |  |  |

Table 2. Confusion Matrix Model 1

|  |  |  |
| --- | --- | --- |
| True \ Pred | 0 | 1 |
| 0 | 5,550 | 458 |
| 1 | 1,235 | 918 |

No variables seem peculiar expect for URBANCITY, this variable stands out with its coefficient at -0.35% chance that a policy holder will be in an accident (less likely) when they live in rural area (less cars, less opportunities for accidents). As this is a correct interpretation of the variable, for now, this variable will be left in.

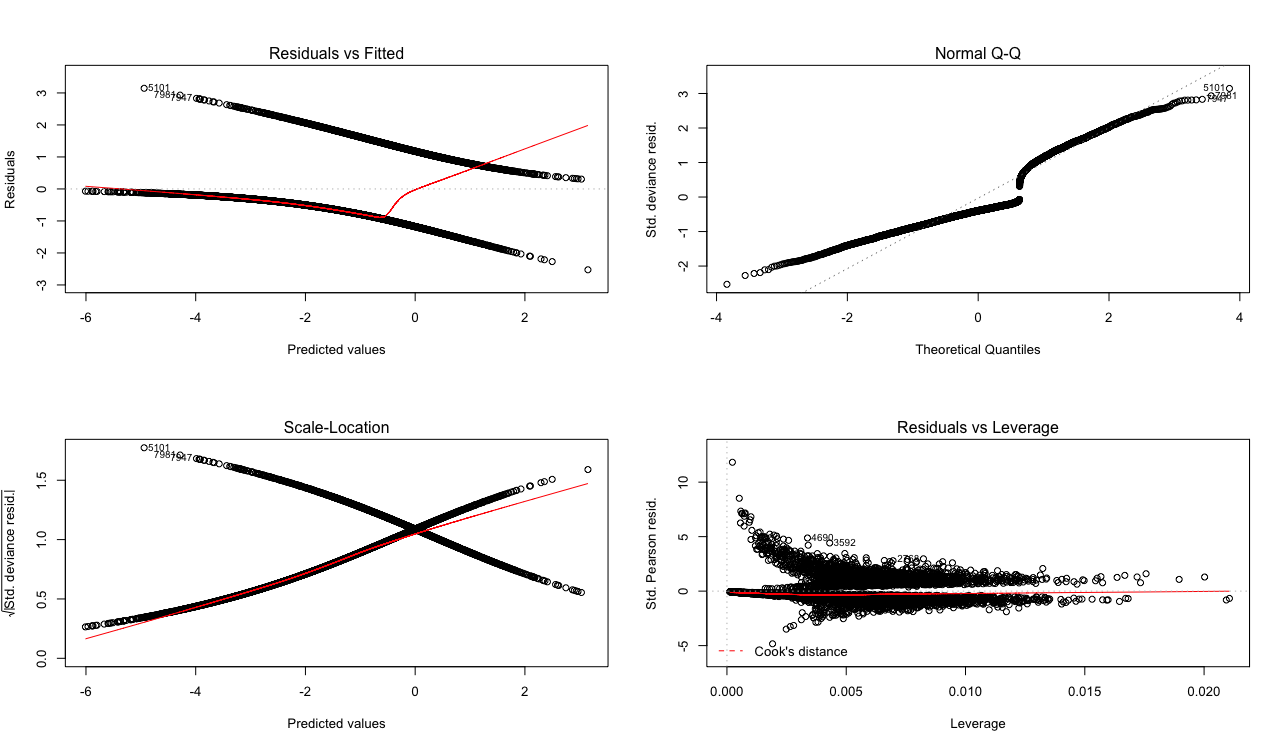


Figure 3. Model 1 (TARGET\_FLAG) Plots.

For this model, we can see that the Normal Q-Q plot shows a unique charactersis not only on the tails but also in the middle, this is due to the binary flag. However, in looking at the residucals we see heteroskedastic behavior.

**MODEL 2**

The second model only takes into account the variables noted of significance from Model 1 (p-value < 0.05). In this second model, we have an AIC of 7376.8. The data in Table 3, shows that the model has an accuracy of 79.0%.

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5523 -0.7190 -0.3985 0.6497 3.1365

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.728e-01 2.620e-01 -3.332 0.000863 \*\*\*

KIDSDRIV 7.664e-01 9.775e-02 7.841 4.48e-15 \*\*\*

INCOME -3.552e-06 1.071e-06 -3.317 0.000910 \*\*\*

PARENT1Yes 4.476e-01 9.451e-02 4.736 2.18e-06 \*\*\*

HOME\_VAL -1.367e-06 3.407e-07 -4.012 6.03e-05 \*\*\*

MSTATUSz\_No 4.766e-01 7.969e-02 5.981 2.22e-09 \*\*\*

EDUCATIONBachelors -3.839e-01 1.086e-01 -3.534 0.000409 \*\*\*

EDUCATIONMasters -3.062e-01 1.612e-01 -1.899 0.057514 .

EDUCATIONPhD -1.761e-01 1.997e-01 -0.882 0.377940

EDUCATIONz\_High\_School 1.682e-02 9.450e-02 0.178 0.858752

JOBClerical 4.011e-01 1.962e-01 2.044 0.040930 \*

JOBDoctor -4.251e-01 2.658e-01 -1.599 0.109770

JOBHome\_Maker 2.561e-01 2.038e-01 1.257 0.208790

JOBLawyer 1.091e-01 1.690e-01 0.646 0.518557

JOBManager -5.704e-01 1.711e-01 -3.335 0.000854 \*\*\*

JOBProfessional 1.578e-01 1.781e-01 0.886 0.375433

JOBStudent 2.732e-01 2.104e-01 1.299 0.194092

JOBz\_Blue\_Collar 3.064e-01 1.852e-01 1.654 0.098047 .

TRAVTIME 1.471e-02 1.877e-03 7.837 4.61e-15 \*\*\*

CAR\_USEPrivate -7.623e-01 9.158e-02 -8.324 < 2e-16 \*\*\*

BLUEBOOK -2.321e-05 4.715e-06 -4.922 8.56e-07 \*\*\*

TIF -3.257e-01 4.135e-02 -7.875 3.41e-15 \*\*\*

CAR\_TYPEPanel\_Truck 6.226e-01 1.505e-01 4.137 3.53e-05 \*\*\*

CAR\_TYPEPickup 5.528e-01 1.006e-01 5.497 3.86e-08 \*\*\*

CAR\_TYPESports\_Car 9.746e-01 1.074e-01 9.077 < 2e-16 \*\*\*

CAR\_TYPEVan 6.466e-01 1.220e-01 5.301 1.15e-07 \*\*\*

CAR\_TYPEz\_SUV 7.218e-01 8.585e-02 8.407 < 2e-16 \*\*\*

CLM\_FREQ 3.624e-01 5.464e-02 6.631 3.33e-11 \*\*\*

REVOKEDYes 7.349e-01 8.022e-02 9.161 < 2e-16 \*\*\*

MVR\_PTS 2.863e-01 4.138e-02 6.920 4.51e-12 \*\*\*

URBANICITYz\_Highly\_Rural/ Rural -2.373e+00 1.129e-01 -21.024 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 9418.0 on 8160 degrees of freedom

Residual deviance: 7314.8 on 8130 degrees of freedom

AIC: 7376.8

Number of Fisher Scoring iterations: 5

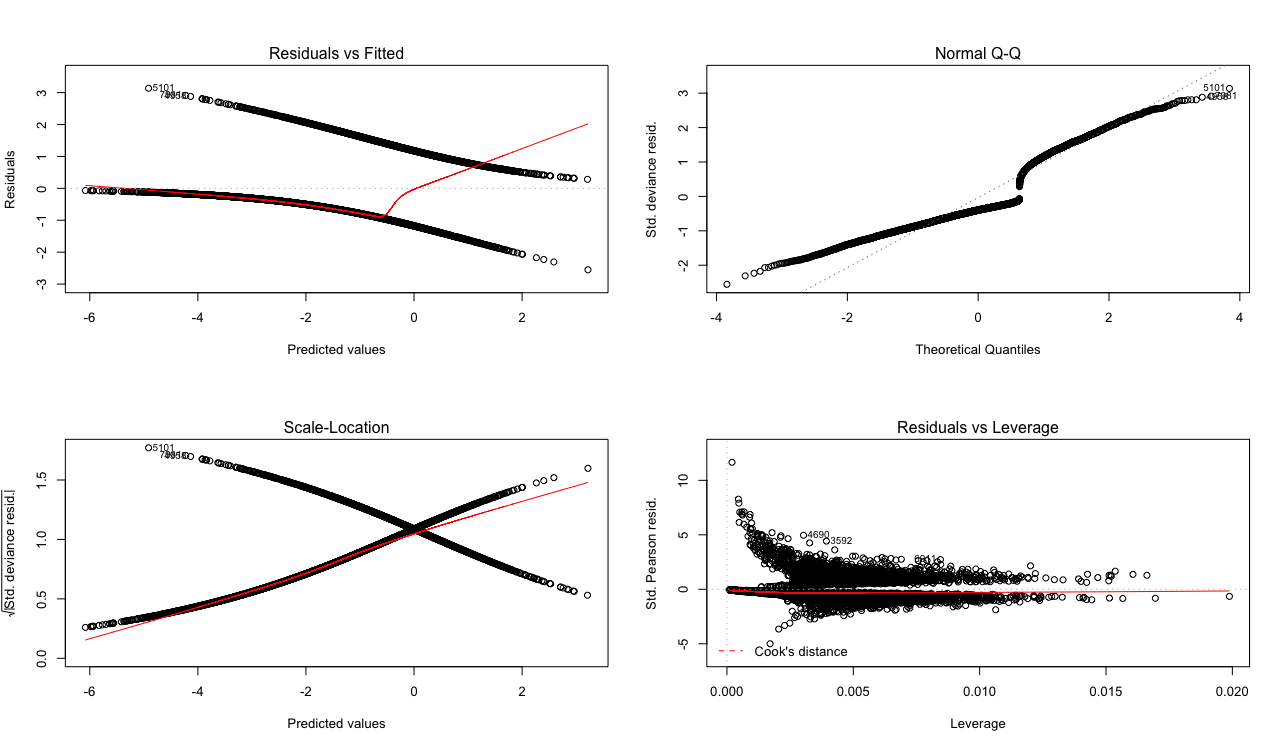
Variable Interpretation:

|  |  |  |  |
| --- | --- | --- | --- |
| (Intercept) | (0.13) | JOBHome\_Maker | 0.04 |
| KIDSDRIV | 0.11 | JOBLawyer | 0.02 |
| INCOME | (0.00) | JOBManager | (0.08) |
| PARENT1Yes | 0.07 | JOBProfessional | 0.02 |
| HOME\_VAL | (0.00) | JOBStudent | 0.04 |
| MSTATUSz\_No | 0.07 | JOBz\_Blue\_Collar | 0.04 |
| EDUCATIONBachelors | (0.06) | TRAVTIME | 0.00 |
| EDUCATIONMasters | (0.04) | CAR\_USEPrivate | (0.11) |
| EDUCATIONPhD | (0.03) | BLUEBOOK | (0.00) |
| EDUCATIONz\_High\_School | 0.00 | TIF | (0.05) |
| JOBClerical | 0.06 | CAR\_TYPEPanel\_Truck | 0.09 |
| JOBDoctor | (0.06) | CAR\_TYPEPickup | 0.08 |
| CAR\_TYPESports\_Car | 0.14 | REVOKEDYes | 0.11 |
| CAR\_TYPEVan | 0.09 | MVR\_PTS | 0.04 |
| CAR\_TYPEz\_SUV | 0.11 | URBANICITYz\_Highly\_Rural/ | (0.35) |

Table 3. Confusion Matrix Model 2

|  |  |  |
| --- | --- | --- |
| True \ Pred | 0 | 1 |
| 0 | 5,541 | 467 |
| 1 | 1,247 | 906 |

In this model, although the AIC dropped, the accuracy also dropped by 0.3% vs Model 1. No variables seem peculiar expect for URBANCITY, this variable stands out with its coefficient at -0.35% chance that a policy holder will be in an accident (less likely) when they live in rural area (less cars, less opportunities for accidents). As this is a correct interpretation of the variable, for now, this variable will be left in.

  
Figure 4. Model 2 (TARGET\_FLAG) Plots.

For this model, we can see that the Normal Q-Q plot shows unique characters is not only on the tails but also in the middle, this is due to the binary flag. However, in looking at the residuals we see heteroskedastic behavior. There is no major change to the first model.

**MODEL 3**

The third model is a reduced model. This dataset has a lot of variables, that from can be seen are secondary in nature to the true intent of the model. Is the policy holder in an accident and what was the value? This can be answered using the most basic items that are there. KIDSDRIVE and TRAVTIME affect who drives and the distance of travel, and then INCOME and HOME\_VAL to see if income and home value (higher risk of assets being seized in an accident) vs those who aren’t high income earners or homeowners. In this third model, we have an AIC of 9031.1. The data in Table 3, shows that the model has an accuracy of 73.6%.

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5299 -0.8217 -0.6749 1.2315 2.8090

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.876e-01 7.305e-02 -9.412 < 2e-16 \*\*\*

KIDSDRIV 7.266e-01 8.115e-02 8.953 < 2e-16 \*\*\*

INCOME -3.497e-06 6.826e-07 -5.123 3.01e-07 \*\*\*

HOME\_VAL -2.972e-06 2.499e-07 -11.895 < 2e-16 \*\*\*

TRAVTIME 5.880e-03 1.598e-03 3.679 0.000234 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 9418.0 on 8160 degrees of freedom

Residual deviance: 9021.1 on 8156 degrees of freedom

AIC: 9031.1

Number of Fisher Scoring iterations: 4

Variable Interpretation:

(Intercept) KIDSDRIV INCOME HOME\_VAL TRAVTIME

-1.271908e-01 1.344090e-01 -6.468917e-07 -5.498016e-07 1.087668e-03

Table 4. Confusion Matrix Model 3

|  |  |  |
| --- | --- | --- |
| True \ Pred | 0 | 1 |
| 0 | 5,937 | 71 |
| 1 | 2086 | 67 |

The increase in AIC is expected and shows that the model does no better than flipping a coin and therefore is not an appropriate model or this exercise. This means that other variables such as Job, Education and other policy holder specific items do impact the chances of an accident and thereby filing a claim. In this case, Model With this in mind, Model 1 so far seems the most appropriate so far.

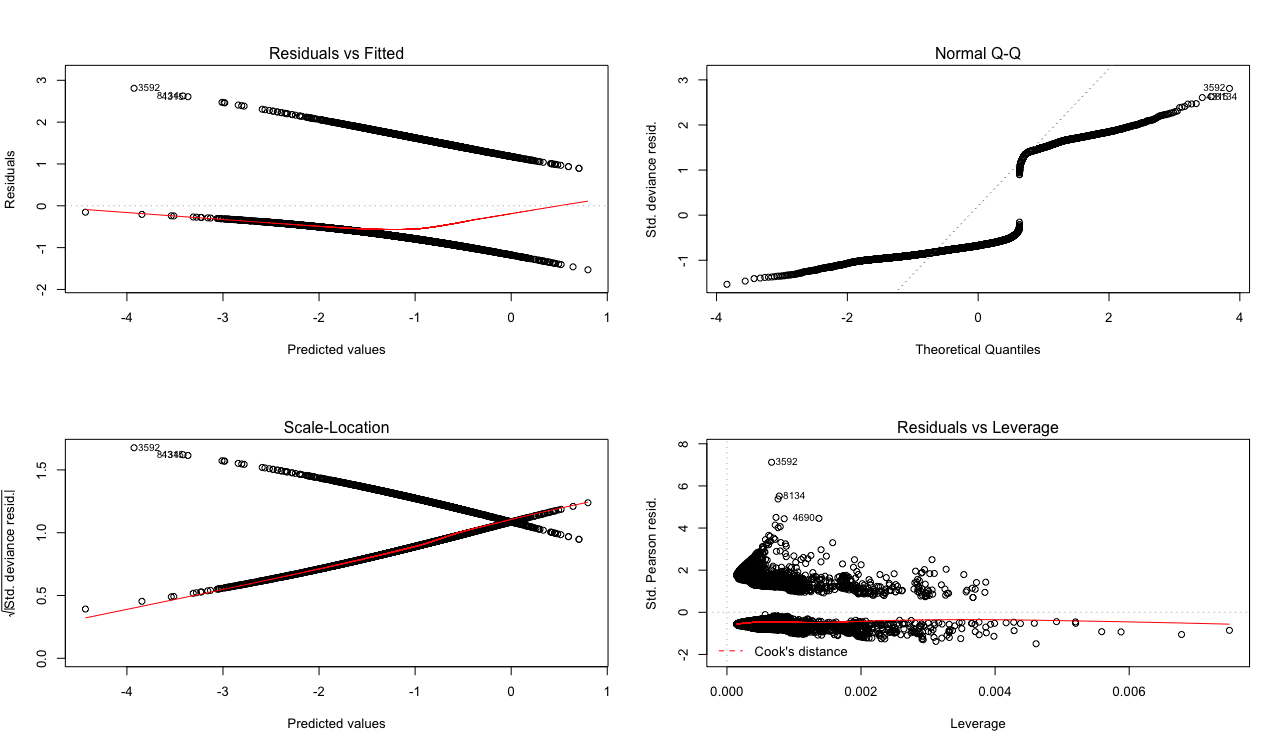


Figure 5. Model 3. (TARGET\_FLAG) Plots.

For this model, we can see that the Normal Q-Q plot is completely different than the others. This is due to the lack of variables to help explain the TARGET\_FLAG variable. This is also apparent in the unique shape of the residuals as they no longer sho any shape but cluster around -1 and 2.

# Model Building for Outcome Variable TARGET\_AMT

The purpose of this step is to take the modified dataset and begin exploring potential models that will be used on the final dataset provided. The following information describes the three (3) models built for this step and the relevant analysis to provide reasons for model selection in the next step.

**MODEL 1**

The first model takes in the data as manipulated in step two (with variables imputed and removed). In this first model, we have an R2 = 0.2879 and p-value < 0.05. The data in Figure 3, shows that there is not heteroscedastic and has a positive trend on the predicted vs fitted values.

Residuals:

Min 1Q Median 3Q Max

-6234 -465 -58 243 101178

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.975e+02 5.010e+02 -1.193 0.2331

TARGET\_FLAG1 5.707e+03 1.134e+02 50.329 < 2e-16 \*\*\*

KIDSDRIV -2.216e+01 1.781e+02 -0.124 0.9010

AGE 6.145e+00 6.271e+00 0.980 0.3272

HOMEKIDS 9.215e+01 1.256e+02 0.733 0.4633

YOJ 7.685e+00 1.319e+01 0.583 0.5601

INCOME -2.258e-03 1.577e-03 -1.431 0.1524

PARENT1Yes 1.209e+02 1.830e+02 0.661 0.5088

HOME\_VAL 3.864e-04 5.165e-04 0.748 0.4545

MSTATUSz\_No 1.770e+02 1.282e+02 1.381 0.1673

SEXz\_F -2.896e+02 1.606e+02 -1.804 0.0713 .

EDUCATIONBachelors 6.823e+01 1.790e+02 0.381 0.7031

EDUCATIONMasters 2.235e+02 2.620e+02 0.853 0.3937

EDUCATIONPhD 4.283e+02 3.110e+02 1.377 0.1685

EDUCATIONz\_High\_School -1.243e+02 1.502e+02 -0.828 0.4077

JOBClerical -8.406e+00 2.984e+02 -0.028 0.9775

JOBDoctor -2.812e+02 3.571e+02 -0.788 0.4310

JOBHome\_Maker -7.045e+01 3.185e+02 -0.221 0.8249

JOBLawyer 7.660e+01 2.582e+02 0.297 0.7667

JOBManager -1.265e+02 2.521e+02 -0.502 0.6158

JOBProfessional 1.733e+02 2.698e+02 0.642 0.5206

JOBStudent -1.306e+02 3.266e+02 -0.400 0.6892

JOBz\_Blue\_Collar 5.187e+01 2.813e+02 0.184 0.8537

TRAVTIME 5.682e-01 2.824e+00 0.201 0.8405

CAR\_USEPrivate -9.993e+01 1.443e+02 -0.693 0.4886

BLUEBOOK 2.944e-02 7.536e-03 3.906 9.45e-05 \*\*\*

TIF -1.653e+01 6.277e+01 -0.263 0.7922

CAR\_TYPEPanel\_Truck -5.880e+01 2.430e+02 -0.242 0.8088

CAR\_TYPEPickup -3.318e+01 1.493e+02 -0.222 0.8241

CAR\_TYPESports\_Car 2.098e+02 1.910e+02 1.099 0.2720

CAR\_TYPEVan 9.709e+01 1.865e+02 0.521 0.6026

CAR\_TYPEz\_SUV 1.621e+02 1.571e+02 1.032 0.3021

RED\_CARyes -2.696e+01 1.302e+02 -0.207 0.8360

OLDCLAIM 4.079e+00 2.908e+01 0.140 0.8884

CLM\_FREQ -8.551e+01 2.210e+02 -0.387 0.6989

REVOKEDYes -2.991e+02 1.385e+02 -2.160 0.0308 \*

MVR\_PTS 1.396e+02 6.716e+01 2.079 0.0376 \*

CAR\_AGE -2.520e+01 1.118e+01 -2.254 0.0242 \*

URBANICITYz\_Highly\_Rural/ Rural 2.987e+01 1.272e+02 0.235 0.8143

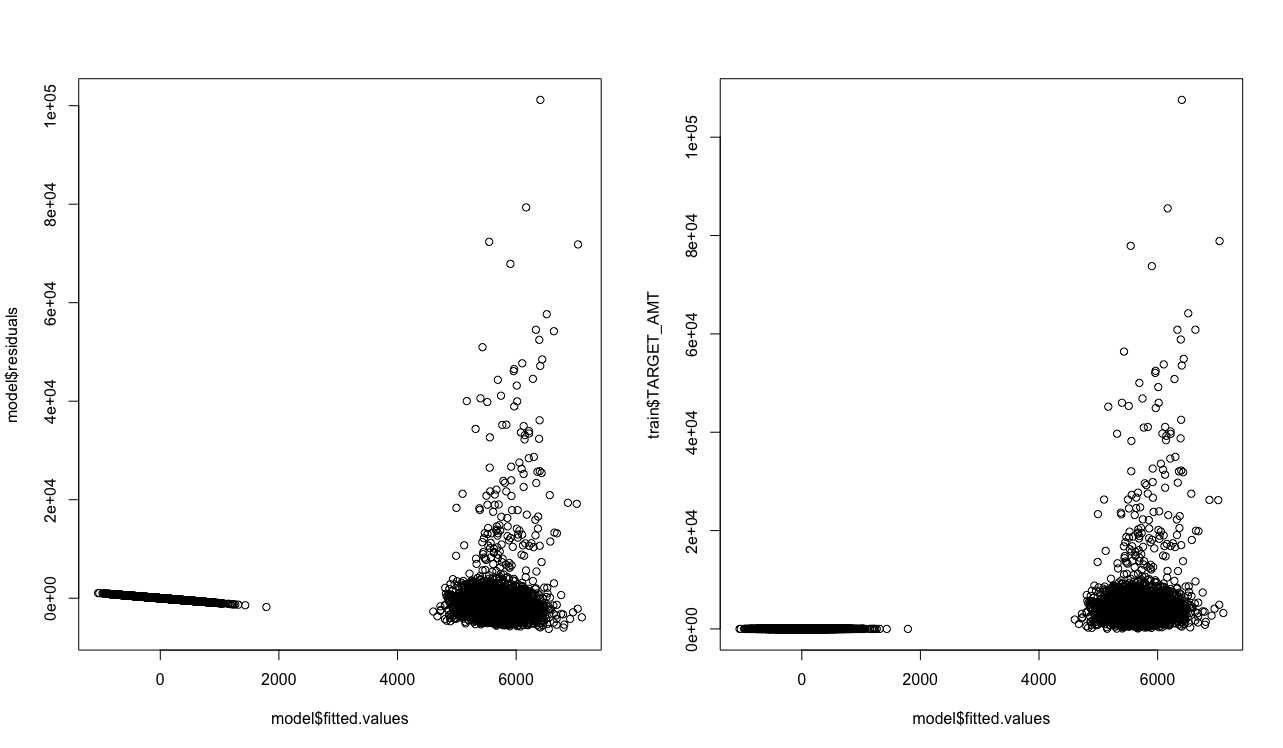
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3970 on 8122 degrees of freedom

Multiple R-squared: 0.2912, Adjusted R-squared: 0.2879

F-statistic: 87.8 on 38 and 8122 DF, p-value: < 2.2e-16

**Figure 7. Model Check for Residual Shape and Model vs. Actuals

What is peculiar in the results however, are that some variables have factor that is counterintuitive to the expected impact on TARGET\_AMT. As an example, JOB has varying signs for jobs that are high paying and would cause the thought that they have more expensive cars and therefore when they are in an accident the amount would be higher than others For now, all variablees will be left in.

**MODEL 2**

The second model only takes into account the variables noted of significance from Model 1 (p-value < 0.05). This means that BLUEBOOK, REVOKED, MVR\_PTS and CAR\_AGE would be the the only variables to be used in Model 2. In this second model, we have an R2 = 0.2886 and p-value < 0.05 which is only a marginal improvement in the model capability.

Residuals:

Min 1Q Median 3Q Max

-6269 -378 -34 192 101505

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.315e+02 1.206e+02 -3.579 0.000347 \*\*\*

TARGET\_FLAG1 5.735e+03 1.036e+02 55.334 < 2e-16 \*\*\*

BLUEBOOK 3.010e-02 5.328e-03 5.649 1.67e-08 \*\*\*

REVOKEDYes -2.874e+02 1.356e+02 -2.120 0.034021 \*

MVR\_PTS 1.309e+02 6.101e+01 2.145 0.031986 \*

CAR\_AGE -1.291e+01 8.122e+00 -1.590 0.111894

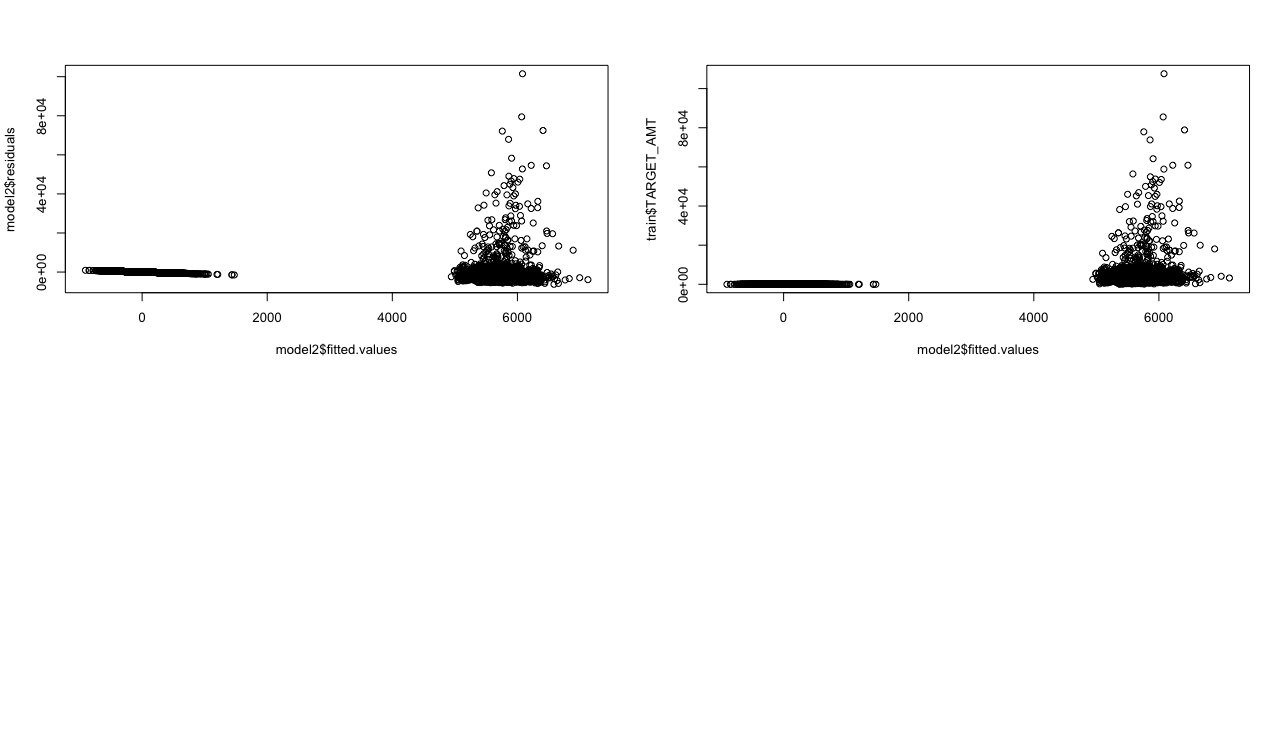
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3968 on 8155 degrees of freedom

Multiple R-squared: 0.289, Adjusted R-squared: 0.2886

F-statistic: 662.9 on 5 and 8155 DF, p-value: < 2.2e-16

**Figure 8. Model 2 Plots (Residuals vs Fitted and QQ)

**MODEL 3**

The third model takes the same variables used in the Model 3 from the TARGET\_FLAG model building. In this third model, we have an R2 = 0.1047 and p-value < 0.05 which is not an improvement in the model capability.

Residuals:

Min 1Q Median 3Q Max

-3610 -1652 -1239 -318 106277

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.680e+03 1.470e+02 11.426 < 2e-16 \*\*\*

KIDSDRIV 9.172e+02 1.789e+02 5.126 3.03e-07 \*\*\*

INCOME -1.242e-03 1.336e-03 -0.930 0.3522

HOME\_VAL -2.809e-03 4.920e-04 -5.710 1.17e-08 \*\*\*

TRAVTIME 7.234e+00 3.260e+00 2.219 0.0265 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4679 on 8156 degrees of freedom

Multiple R-squared: 0.01096, Adjusted R-squared: 0.01047

F-statistic: 22.59 on 4 and 8156 DF, p-value: < 2.2e-16

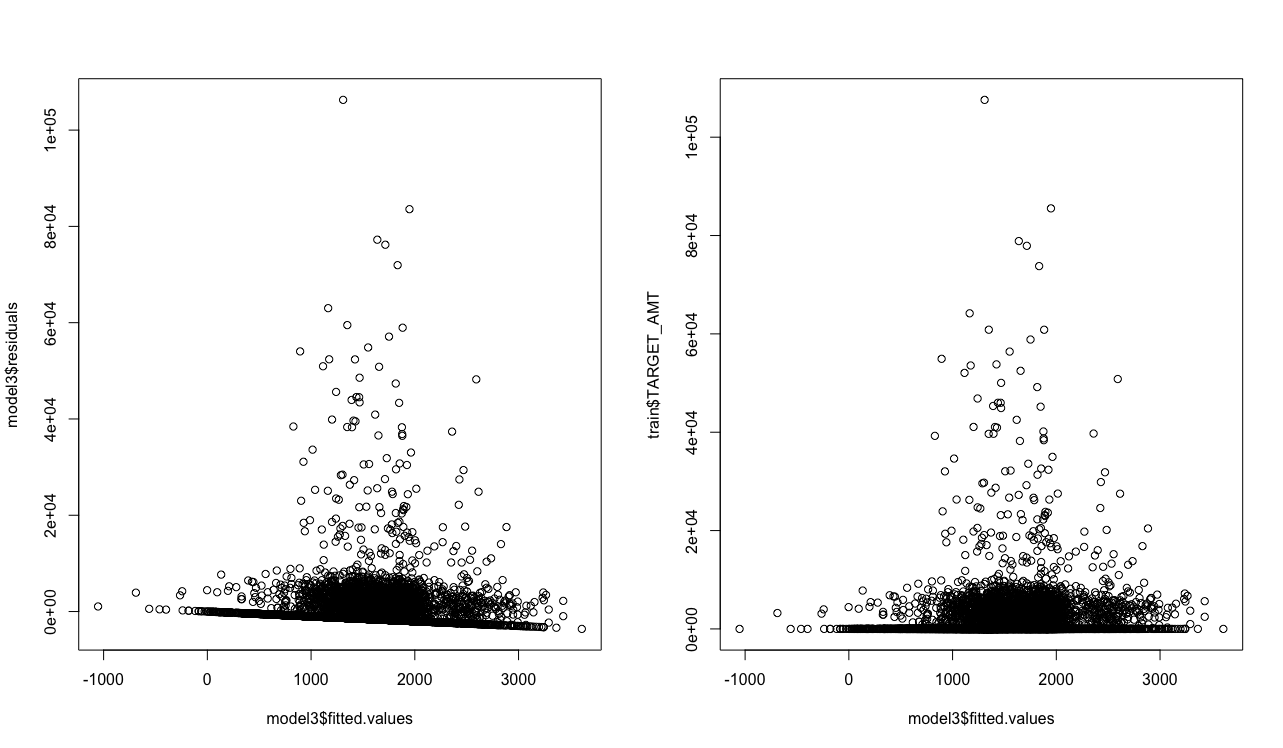
**

Figure 5. Model 3 Plots (Residuals vs Fitted and QQ)

Similar to the models from the TARGET\_FLAG exercise, the numerous amount of variables do have an impact in the value of the claim when there is an accident. This means that solely looking at variables that are correlated to driving behavior are not enough to explain the outcome variables. With this in mind, Model 1 is also the most appropriate for the TARGET\_AMT variable just like it was in the TARGET\_FLAG variable.

**METHODOLOGY**

Familiarity with the dataset subject is low and therefore the methodology will be more closely related to the statistical information presented. In this case, a combination of three (3) factors (AIC, Percent Accuracy, and ROC Curve) will be the criteria to select the model for the TARGET\_FLAG variable and one (1) factor (R2) for the TARGET\_AMT variable. The reason for this is that the significance of each variable is high in Model 1 through 3 as the adjustments for correlation and log transformations were already taken care of in Step 2 of the process. If Step 2 had not been done, then it would have been hidden in the model building and taken care of between Model 1 and Model 2. In addition, because this is a binary predictive exercise accuracy is also important for this exercise as seen in Table 5 below.

Table 5. Model Criteria Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| OUTCOME VARIABLE | Criteria | Model 1  (All Variables) | **Model 2**  **(Significant Variables Only)** | Model 3  (Politically Correct) |
| TARGET\_FLAG | AIC | 7384.4 | 7376.8 | 9031.1 |
| Accuracy % | 79.3% | 79.0% | 73.6% |
| TARGET\_AMT | R2 | 0.2879% | 0.2886% | 0.01047% |

Of importance also is the ROC Curves for each model which tell us if the model predictive capability is better than just chance (a coin-toss at 50/50). In looking at each curve blow in Figure 3, we can see that the ROC curve for model 1 has a better smoother transition in comparison to Model 1 and Model 2. Overall, the ROC curve for Model 1 trends to the upper left quadrant in a more evenly distributed manner versus the other two.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Model 1 | Model 2 | Model 3 |

Figure 6. ROC Curves for Each Model (Model 1 through 3) for TARGET\_FLAG.

With this in mind, Model 1 is best model with an AIC of 7384.4 for the TARGET\_FLAG variable and Model 1 is the best model for the TARGET\_AMT variable.

**TEST DATA**

The dataset had 2,141 entries and 26 columns and was modified to fit the final variables and scaling used in Model 1 from above. This means that the same process of adjustments and log transformations was done in order to be able to use the model correctly. The final predicted values are based upon a normalized value from the test data. The data is shown as follows with the corresponding summaries for the spread of the data.

Table 6. Predicted Statistics vs Summary of Model 1 Predicted Values for TARGET\_FLAG

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **Min.** |
| 0.0024 | 0.0774 | 0.2017 | 0.2638 | 0.4035 | 0.9589 | 0.0024 |
| 0.0031 | 0.0777 | 0.2183 | 0.2708 | 0.4102 | 0.9464 | 0.0031 |

Table 6 above is only meant as a comparison but it does highlight that the test data has a higher set of values that would be deemed 0 (that there is no claim). The spread of the data for test is also a lot tighter than the training values which may be a function of cases in the test data. The data might have more of TARGET\_FLAG = 0 or 1 which would skew the results.

Table 7. Predicted Statistics vs Summary of TARGET\_AMT in Training Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Predicted (Test) | | | Train |
|  | fit | lwr | upr | Actual |
| Min. | -1206.17 | -1870.4 | -542 | 0 |
| 1st | -255.615 | -782.6 | 256.4 | 0 |
| Median | -22.708 | -538.1 | 478.1 | 0 |
| Mean | -8.173 | -540.5 | 524.1 | 1,504 |
| 3rd | 223.762 | -303.8 | 774.3 | 1,036 |
| Max. | 1251.287 | 521.4 | 1998.7 | 107,586 |

Table 7 is only meant as a comparison but it does highlight that the training data doesn’t fall in the anywhere in the upper / lower limits except in the minimum values. The spread of the data for training is also a lot tighter than the predicted values which an issue in the method of normalizing the test data. This might indicate why the training and predicted values aren’t more closely aligned.

# Conclusion

Six (6) models were presented (3 for TARGET\_FLAG and 3 for TARGET\_AMT) after exploring and manipulating the data as necessary. With using a multi-criteria approach for this exercise, it became clear that the Model 1 was selected and provided an AIC of 7384.4 for TARGET\_FLAG and a R2=0.2879 for TARGET\_AMT which was basically using all the variables presented in the dataset. If more time were available, the creation of new variables would be explored to create more factored variables instead of continuous variables that were presented and could have provided better insight into the data set.

Appendix A: R Code

---

title: "Data 621"

author: 'Cesar Espitia HW #4'

date: "7/8/2018"

output: html\_document

---

## Abstract

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000

records representing a customer at an auto insurance company. Each record has two response variables. The

first response variable, TARGET\_FLAG, is a 1 or a 0. A “1” means that the person was in a car crash. A zero

means that the person was not in a car crash. The second response variable is TARGET\_AMT. This value is zero

if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

###Keywords: insurance, data621

## Data Exploration

```{r dataexploration}

knitr::opts\_chunk$set(echo = TRUE)

library(e1071)

library(dplyr)

library(purrr)

library(tidyr)

library(ggplot2)

library(corrplot)

library(FactoMineR)

library(VIF)

library(knitr)

library(kableExtra)

library(Hmisc)

library(pROC)

library(binr)

# read data

train = read.csv(file="data/insurance\_training\_data.csv")

dim(train)

#transform data

#this step is necessary in order to analyze data as it is not clean

currencyconv = function(input) {

out = sub("\\$", "", input)

out = as.numeric(sub(",", "", out))

return(out)

}

# Replace spaces with underscores

underscore = function(input) {

out = sub(" ", "\_", input)

return(out)

}

train = as.tbl(train) %>%

mutate\_at(c("INCOME","HOME\_VAL","BLUEBOOK","OLDCLAIM"),

currencyconv) %>%

mutate\_at(c("EDUCATION","JOB","CAR\_TYPE","URBANICITY"),

underscore) %>%

mutate\_at(c("EDUCATION","JOB","CAR\_TYPE","URBANICITY"),

as.factor) %>%

mutate(TARGET\_FLAG = as.factor(TARGET\_FLAG))

#check data

summary(train) %>% kable() %>% kable\_styling()

sapply(train, function(x) sum(is.na(x))) %>% kable() %>% kable\_styling()

# library(UpSetR)

#

# train %>% as\_shadow\_upset() %>% upset()

ntrain<-select\_if(train, is.numeric)

ntrain %>%

keep(is.numeric) %>% # Keep only numeric columns

gather() %>% # Convert to key-value pairs

ggplot(aes(value)) + # Plot the values

facet\_wrap(~ key, scales = "free") + # In separate panels

geom\_density()

#

# trainnum <- dplyr::select\_if(train, is.numeric)

#

# rcorr(as.matrix(trainnum))

# corrplot(cor(trainnum), method="square")

#

# # correlation test 1

# cor.test(trainnum$HOME\_VAL,trainnum$INCOME,method="pearson")

#

# #NOT significant ignore

```

## Data Preparation

```{r datapreparation}

# impute data for missing values

# use column mean for calculation

train$AGE[is.na(train$AGE)] <- mean(train$AGE, na.rm=TRUE)

train$YOJ[is.na(train$YOJ)] <- mean(train$YOJ, na.rm=TRUE)

train$HOME\_VAL[is.na(train$HOME\_VAL)] <- mean(train$HOME\_VAL, na.rm=TRUE)

train$CAR\_AGE[is.na(train$CAR\_AGE)] <- mean(train$CAR\_AGE, na.rm=TRUE)

train$INCOME[is.na(train$INCOME)] <- mean(train$INCOME, na.rm=TRUE)

#get complete cases

train <- train[complete.cases(train),]

train2<-train

# # transform data using log for skewed HOMEKIDS, MVR\_PTS, OLDCLAIM, TIF, KIDSDRIVE and CLM\_FREQ

train$HOMEKIDS <- log(train$HOMEKIDS+1)

train$MVR\_PTS <- log(train$MVR\_PTS+1)

train$OLDCLAIM <- log(train$OLDCLAIM+1)

train$TIF <- log(train$TIF+1)

train$KIDSDRIV <- log(train$KIDSDRIV+1)

train$CLM\_FREQ <- log(train$CLM\_FREQ+1)

#remove rad per correlation in prior section

train <- train[, !(colnames(train) %in% c("INDEX"))]

#

# #create variable

# train$new <- train$tax / (train$medv\*10)

#

trainnum <- dplyr::select\_if(train, is.numeric)

rcorr(as.matrix(trainnum))

corrplot(cor(trainnum), method="square")

cor.test(trainnum$HOMEKIDS,trainnum$AGE,method="pearson")

train2<-train

```

## Build Models LOGIT TARGET\_FLAG

```{r buildmodelslogit}

#MODEL 1

logit <- glm(formula = TARGET\_FLAG ~ . - TARGET\_AMT, data=train, family = "binomial" (link="logit"))

summary(logit)

exp(logit$coefficients)

logitscalar <- mean(dlogis(predict(logit, type = "link")))

logitscalar \* coef(logit)

confint.default(logit)

predlogit <- predict(logit, type="response")

train2$pred1 <- predict(logit, type="response")

summary(predlogit)

table(true = train$TARGET\_FLAG, pred = round(fitted(logit)))

#plots for Model 1

par(mfrow=c(2,2))

plot(logit)

data.frame(train2$pred1) %>%

ggplot(aes(x = train2.pred1)) +

geom\_histogram(bins = 50, fill = 'grey50') +

labs(title = 'Histogram of Predictions') +

theme\_bw()

plot.roc(train$TARGET\_FLAG, train2$pred1)

#extract variables that are significant and rerun model

sigvars <- data.frame(summary(logit)$coef[summary(logit)$coef[,4] <= .05, 4])

sigvars <- add\_rownames(sigvars, "vars")

colist<-dplyr::pull(sigvars, vars)

# colist<-colist[2:11]

colist<-c("KIDSDRIV","INCOME","PARENT1","HOME\_VAL","MSTATUS","EDUCATION","JOB","TRAVTIME","CAR\_USE","BLUEBOOK","TIF","CAR\_TYPE","CLM\_FREQ","REVOKED","MVR\_PTS","URBANICITY")

idx <- match(colist, names(train))

trainmod2 <- cbind(train[,idx], train2['TARGET\_FLAG'])

#MODEL 2

logit2 <- glm(TARGET\_FLAG ~ ., data=trainmod2, family = "binomial" (link="logit"))

summary(logit2)

exp(logit2$coefficients)

logit2scalar <- mean(dlogis(predict(logit2, type = "link")))

logit2scalar \* coef(logit2)

predlogit2 <- predict(logit2, type="response")

train2$pred2 <- predict(logit2, type="response")

summary(predlogit2)

table(true = train$TARGET\_FLAG, pred = round(fitted(logit2)))

#plots for Model 2

par(mfrow=c(2,2))

plot(logit2)

data.frame(train2$pred2) %>%

ggplot(aes(x = train2.pred2)) +

geom\_histogram(bins = 50, fill = 'grey50') +

labs(title = 'Histogram of Predictions') +

theme\_bw()

plot.roc(train$TARGET\_FLAG, train2$pred2)

#MODEL 3

#PC Model no racial bias

logit3 <- glm(TARGET\_FLAG ~ KIDSDRIV + INCOME + HOME\_VAL + TRAVTIME, data=train, family = "binomial" (link="logit"))

summary(logit3)

exp(logit3$coefficients)

predlogit3 <- predict(logit3, type="response")

train2$pred3 <- predict(logit3, type="response")

summary(predlogit3)

table(true = train$TARGET\_FLAG, pred = round(fitted(logit3)))

#plots for Model 3

par(mfrow=c(2,2))

plot(logit3)

data.frame(train2$pred3) %>%

ggplot(aes(x = train2.pred3)) +

geom\_histogram(bins = 50, fill = 'grey50') +

labs(title = 'Histogram of Predictions') +

theme\_bw()

plot.roc(train$TARGET\_FLAG, train2$pred3)

logit3scalar <- mean(dlogis(predict(logit3, type = "link")))

logit3scalar \* coef(logit3)

round(logitscalar \* coef(logit),2)

round(logit2scalar \* coef(logit2),2)

round(logit3scalar \* coef(logit3),2)

```

## Build Models GENERAL TARGET\_AMT

```{r buildmodels, include=TRUE}

#MODEL 1

model <- lm(TARGET\_AMT ~ ., data=train)

summary(model)

par(mfrow=c(1,2))

plot(model$residuals ~ model$fitted.values)

plot(model$fitted.values,train$TARGET\_AMT)

par(mfrow=c(2,2))

plot(model)

#extract variables that are significant and rerun model

sigvars <- data.frame(summary(model)$coef[summary(model)$coef[,4] <= .05, 4])

sigvars <- add\_rownames(sigvars, "vars")

colist<-dplyr::pull(sigvars, vars)

colist<-c("TARGET\_FLAG","BLUEBOOK","REVOKED","MVR\_PTS","CAR\_AGE")

idx <- match(colist, names(train))

trainmod2 <- cbind(train[,idx], train['TARGET\_AMT'])

#MODEL 2

model2<-lm(TARGET\_AMT ~ ., data=trainmod2)

summary(model2)

par(mfrow=c(2,2))

plot(model2$residuals ~ model2$fitted.values)

plot(model2$fitted.values,train$TARGET\_AMT)

par(mfrow=c(2,2))

plot(model2)

par(mfrow=c(1,2))

plot(model2$residuals ~ model2$fitted.values, main="New Reduced Var Model")

abline(h = 0)

plot(model$residuals ~ model$fitted.values, main="Orignal Model All Vars")

abline(h = 0)

#MODEL 3

#remove variables with opposite coefficients

model3<-lm(TARGET\_AMT ~ KIDSDRIV + INCOME + HOME\_VAL + TRAVTIME, data=train)

summary(model3)

par(mfrow=c(1,2))

plot(model3$residuals ~ model3$fitted.values)

plot(model3$fitted.values,train$TARGET\_AMT)

par(mfrow=c(2,2))

plot(model3)

```

## Select Models

```{r selectmodels}

test = read.csv(file="data/insurance-evaluation-data.csv")

test2<- test

dim(test)

test$TARGET\_AMT <- 0

test$TARGET\_FLAG <- 0

test = as.tbl(test) %>%

mutate\_at(c("INCOME","HOME\_VAL","BLUEBOOK","OLDCLAIM"),

currencyconv) %>%

mutate\_at(c("EDUCATION","JOB","CAR\_TYPE","URBANICITY"),

underscore) %>%

mutate\_at(c("EDUCATION","JOB","CAR\_TYPE","URBANICITY"),

as.factor) %>%

mutate(TARGET\_FLAG = as.factor(TARGET\_FLAG))

# impute data for missing values

# use column mean for calculation

test$HOMEKIDS <- log(test$HOMEKIDS+1)

test$MVR\_PTS <- log(test$MVR\_PTS+1)

test$OLDCLAIM <- log(test$OLDCLAIM+1)

test$TIF <- log(test$TIF+1)

test$KIDSDRIV <- log(test$KIDSDRIV+1)

test$CLM\_FREQ <- log(test$CLM\_FREQ+1)

# use column mean for calculation

test$AGE[is.na(test$AGE)] <- mean(test$AGE, na.rm=TRUE)

test$YOJ[is.na(test$YOJ)] <- mean(test$YOJ, na.rm=TRUE)

test$HOME\_VAL[is.na(test$HOME\_VAL)] <- mean(test$HOME\_VAL, na.rm=TRUE)

test$CAR\_AGE[is.na(test$CAR\_AGE)] <- mean(test$CAR\_AGE, na.rm=TRUE)

test$INCOME[is.na(test$INCOME)] <- mean(test$INCOME, na.rm=TRUE)

#get complete cases

#remove rad per correlation in prior section

test <- test[, !(colnames(test) %in% c("INDEX"))]

TARGET\_FLAG <- predict(logit, newdata = test, type="response")

y\_pred\_num <- ifelse(TARGET\_FLAG > 0.5, 1, 0)

y\_pred <- factor(y\_pred\_num, levels=c(0, 1))

summary(y\_pred)

rbind(round(summary(predlogit),4), round(summary(TARGET\_FLAG),4)) %>% kable()

test$TARGET\_FLAG <- as.factor(test$TARGET\_FLAG)

test2 <- test[, !(colnames(test) %in% c("TARGET\_FLAG"))]

TARGET\_AMT<- predict(model, newdata = test, interval='confidence') #data from scaling originally to get to actual wins

summary(TARGET\_AMT)

summary(model)

```

Appendix B: CORRELATION MATRIX

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | INDEX | TARGET\_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ | INCOME | HOME\_VAL | TRAVTIME | BLUEBOOK | TIF | OLDCLAIM | CLM\_FREQ | MVR\_PTS | CAR\_AGE |
| INDEX | 0 | 0.9572 | 0.1594 | 0.0022 | 0.9962 | 0.0189 | 0.4385 | 0.2881 | 0.0372 | 0.2089 | 0.4053 | 0.9091 | 0.0898 | 0.4765 | 0.9513 |
| TARGET\_AMT | 0.9572 | 0 | 0 | 0.0002 | 0 | 0.0525 | 0 | 0 | 0.0115 | 0.6712 | 0 | 0 | 0 | 0 | 0 |
| KIDSDRIV | 0.1594 | 0 | 0 | 0 | 0 | 0.0001 | 0 | 0.0825 | 0.4455 | 0.0516 | 0.8574 | 0.0653 | 0.0008 | 0 | 0 |
| AGE | 0.0022 | 0.0002 | 0 | 0 | 0 | 0 | 0 | 0 | 0.6342 | 0 | 0.9952 | 0.0082 | 0.0296 | 0 | 0 |
| HOMEKIDS | 0.9962 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5128 | 0 | 0.2859 | 0.0069 | 0.008 | 0 | 0 |
| YOJ | 0.0189 | 0.0525 | 0.0001 | 0 | 0 | 0 | 0 | 0 | 0.1369 | 0 | 0.0296 | 0.7936 | 0.0209 | 0.0009 | 0 |
| INCOME | 0.4385 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9276 | 0 | 0 | 0 | 0 |
| HOME\_VAL | 0.2881 | 0 | 0.0825 | 0 | 0 | 0 | 0 | 0 | 0.0018 | 0 | 0.8564 | 0 | 0 | 0 | 0 |
| TRAVTIME | 0.0372 | 0.0115 | 0.4455 | 0.6342 | 0.5128 | 0.1369 | 0 | 0.0018 | 0 | 0.1246 | 0.2945 | 0.0818 | 0.5535 | 0.3384 | 0.0008 |
| BLUEBOOK | 0.2089 | 0.6712 | 0.0516 | 0 | 0 | 0 | 0 | 0 | 0.1246 | 0 | 0.6242 | 0.0077 | 0.001 | 0.0004 | 0 |
| TIF | 0.4053 | 0 | 0.8574 | 0.9952 | 0.2859 | 0.0296 | 0.9276 | 0.8564 | 0.2945 | 0.6242 | 0 | 0.0473 | 0.0375 | 0.0002 | 0.4969 |
| OLDCLAIM | 0.9091 | 0 | 0.0653 | 0.0082 | 0.0069 | 0.7936 | 0 | 0 | 0.0818 | 0.0077 | 0.0473 | 0 | 0 | 0 | 0.2417 |
| CLM\_FREQ | 0.0898 | 0 | 0.0008 | 0.0296 | 0.008 | 0.0209 | 0 | 0 | 0.5535 | 0.001 | 0.0375 | 0 | 0 | 0 | 0.4151 |
| MVR\_PTS | 0.4765 | 0 | 0 | 0 | 0 | 0.0009 | 0 | 0 | 0.3384 | 0.0004 | 0.0002 | 0 | 0 | 0 | 0.0817 |
| CAR\_AGE | 0.9513 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0008 | 0 | 0.4969 | 0.2417 | 0.4151 | 0.0817 | 0 |