Data 621 Group 2 HW 4: Insurance

Members: Omar Pineda, Jeff Littlejohn, Sergio Ortega Cruz, Chester Poon, Simon Ustoyev

11/15/2019

Problem Definition

The objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car.

Dataset Definition

VARIABLE.NAME	DEFINITION	THEORETICAL.EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

DATA EXPLORATION

Let's start with a glimpse of the data

```
##   TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1 HOME_VAL MSTATUS SEX     EDUCATION           JOB
## 1           0          0        0  60        0  11  $67,349      No       $0    z_No   M           PhD  Professional
## 2           0          0        0  43        0  11  $91,449      No $257,252    z_No   M z_High School z_Blue Collar
## 4           0          0        0  35        1  10  $16,039      No $124,191     Yes z_F z_High School      Clerical
## 5           0          0        0  51        0  14                No $306,251     Yes   M  <High School z_Blue Collar
```

```
## 6           0          0        0  50        0  NA $114,986      No $243,925     Yes z_F           PhD         Doctor
## 7           1       2946        0  34        1  12 $125,301     Yes        $0    z_No z_F     Bachelors z_Blue Collar
##   TRAVTIME    CAR_USE BLUEBOOK TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE          URBANICITY
## 1       14    Private  $14,230  11    Minivan     yes   $4,461        2      No       3      18 Highly Urban/ Urban
## 2       22 Commercial  $14,940   1    Minivan     yes       $0        0      No       0       1 Highly Urban/ Urban
## 4        5    Private   $4,010   4      z_SUV      no  $38,690        2      No       3      10 Highly Urban/ Urban
## 5       32    Private  $15,440   7    Minivan     yes       $0        0      No       0       6 Highly Urban/ Urban
## 6       36    Private  $18,000   1      z_SUV      no  $19,217        2     Yes       3      17 Highly Urban/ Urban
## 7       46 Commercial  $17,430   1 Sports Car      no       $0        0      No       0       7 Highly Urban/ Urban
```

And, here's the summary for all the variables in the dataset:

```
##   TARGET_FLAG     TARGET_AMT        KIDSDRIV        AGE         HOMEKIDS          YOJ           INCOME       PARENT1
##  Min.   :0.00   Min.   :     0   Min.   :0.0   Min.   :16   Min.   :0.0   Min.   : 0    $0        : 615   No :7084
##  1st Qu.:0.00   1st Qu.:     0   1st Qu.:0.0   1st Qu.:39   1st Qu.:0.0   1st Qu.: 9              : 445   Yes:1077
##  Median :0.00   Median :     0   Median :0.0   Median :45   Median :0.0   Median :11    $26,840 :   4
##  Mean   :0.26   Mean   :  1504   Mean   :0.2   Mean   :45   Mean   :0.7   Mean   :10    $48,509 :   4
##  3rd Qu.:1.00   3rd Qu.:  1036   3rd Qu.:0.0   3rd Qu.:51   3rd Qu.:1.0   3rd Qu.:13    $61,790 :   4
##  Max.   :1.00   Max.   :107586   Max.   :4.0   Max.   :81   Max.   :5.0   Max.   :23    $107,375:   3
##                                                NA's   :6                  NA's   :454   (Other) :7086
##      HOME_VAL    MSTATUS      SEX               EDUCATION                JOB          TRAVTIME         CAR_USE
##  $0      :2294   Yes :4894   M  :3786   <High School :1203   z_Blue Collar:1825   Min.   :  5   Commercial:3029
##          : 464   z_No:3267   z_F:4375   Bachelors    :2242   Clerical     :1271   1st Qu.: 22   Private   :5132
##  $111,129:   3                          Masters      :1658   Professional :1117   Median : 33
##  $115,249:   3                          PhD          : 728   Manager      : 988   Mean   : 33
##  $123,109:   3                          z_High School:2330   Lawyer       : 835   3rd Qu.: 44
##  $153,061:   3                                               Student      : 712   Max.   :142
##  (Other) :5391                                               (Other)      :1413
##     BLUEBOOK          TIF                CAR_TYPE    RED_CAR      OLDCLAIM        CLM_FREQ     REVOKED       MVR_PTS
##  $1,500 : 157   Min.   : 1.0   Minivan    :2145   no :5783   $0      :5009   Min.   :0.0   No :7161   Min.   : 0.0
##  $6,000 :  34   1st Qu.: 1.0   Panel Truck: 676   yes:2378   $1,310 :   4   1st Qu.:0.0   Yes:1000   1st Qu.: 0.0
##  $5,800 :  33   Median : 4.0   Pickup     :1389              $1,391 :   4   Median :0.0              Median : 1.0
##  $6,200 :  33   Mean   : 5.4   Sports Car : 907              $4,263 :   4   Mean   :0.8              Mean   : 1.7
##  $6,400 :  31   3rd Qu.: 7.0   Van        : 750              $1,105 :   3   3rd Qu.:2.0              3rd Qu.: 3.0
##  $5,900 :  30   Max.   :25.0   z_SUV      :2294              $1,332 :   3   Max.   :5.0              Max.   :13.0
##  (Other):7843                                                (Other):3134
##     CAR_AGE                   URBANICITY
##  Min.   :-3    Highly Urban/ Urban  :6492
##  1st Qu.: 1    z_Highly Rural/ Rural:1669
##  Median : 8
```

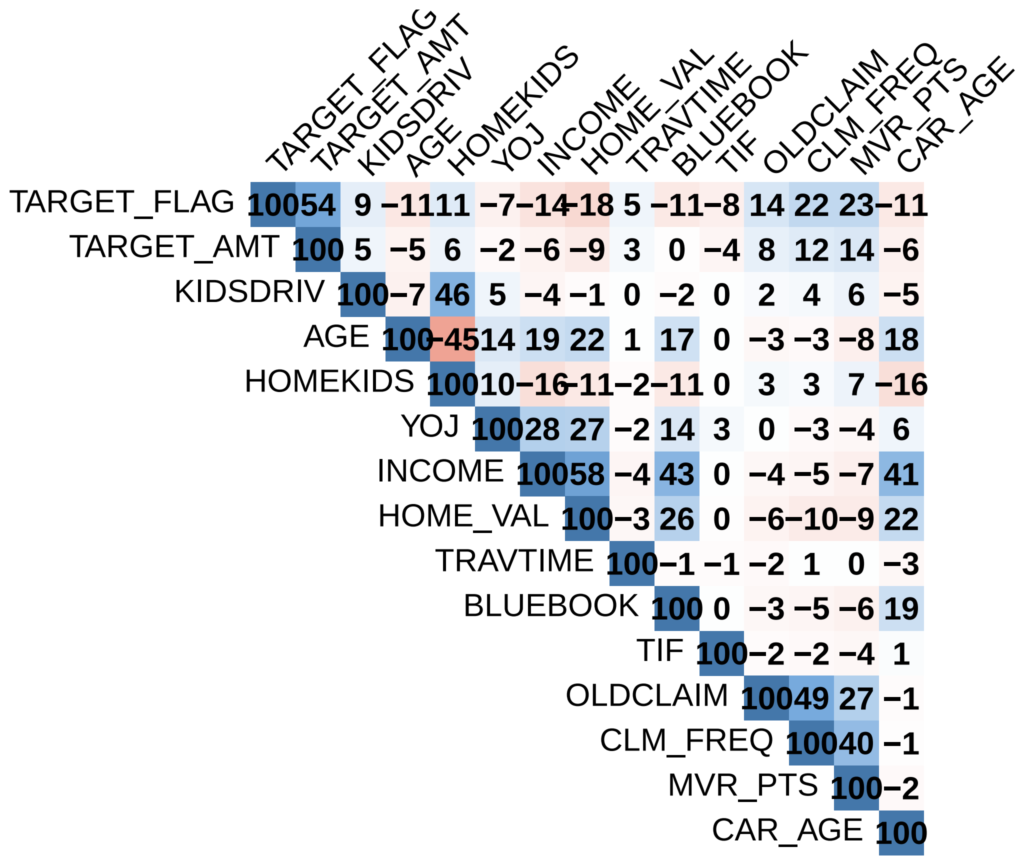
```
##  Mean   : 8
##  3rd Qu.:12
##  Max.   :28
##  NA's   :510
```

The summary on the data identifies the following variables with missing values (and counts)
1. AGE (6)
2. YOJ (454)
3. INCOME (445)
4. HOME_VAL (464)
5. CAR_AGE (510)

Also, based on the summary and the ranges for `Min` and `Max`, the data seems to be pretty clean and valid with no invalid outliers (except for some negative values in `CAR_AGE`). The currency data for variables, `INCOME`, `HOME_VAL`, `BLUEBOOK`, `OLDCLAIM`, got loaded as factors instead of numeric and therefore needs to be *"fixed"*. After the conversion to numeric values, the summary for these variables, below, also shows that the data seems valid, having appropriate ranges.

```
##      INCOME          HOME_VAL         BLUEBOOK        OLDCLAIM
##  Min.   :     0   Min.   :     0   Min.   : 1500   Min.   :    0
##  1st Qu.: 28097   1st Qu.:     0   1st Qu.: 9280   1st Qu.:    0
##  Median : 54028   Median :161160   Median :14440   Median :    0
##  Mean   : 61898   Mean   :154867   Mean   :15710   Mean   : 4037
##  3rd Qu.: 85986   3rd Qu.:238724   3rd Qu.:20850   3rd Qu.: 4636
##  Max.   :367030   Max.   :885282   Max.   :69740   Max.   :57037
##  NA's   :445      NA's   :464
```

Now let's see how numerical data is correlated to the target variables and to each other, based on the chart below.



Based on the chart, there are some cases with significant percentage of correlation. However such parings of correlated values are expected. For example, `KIDSDRIV` is expected to be correlated to `HOMEKIDS` and high `INCOME` would correlate with higher values of `HOME_VAL` and `BLUEBOOK`. Such correlation may not be addressed right away as we still need to prepare and possibly transform the data. Also, some of the correlated values may fall off during model selection.

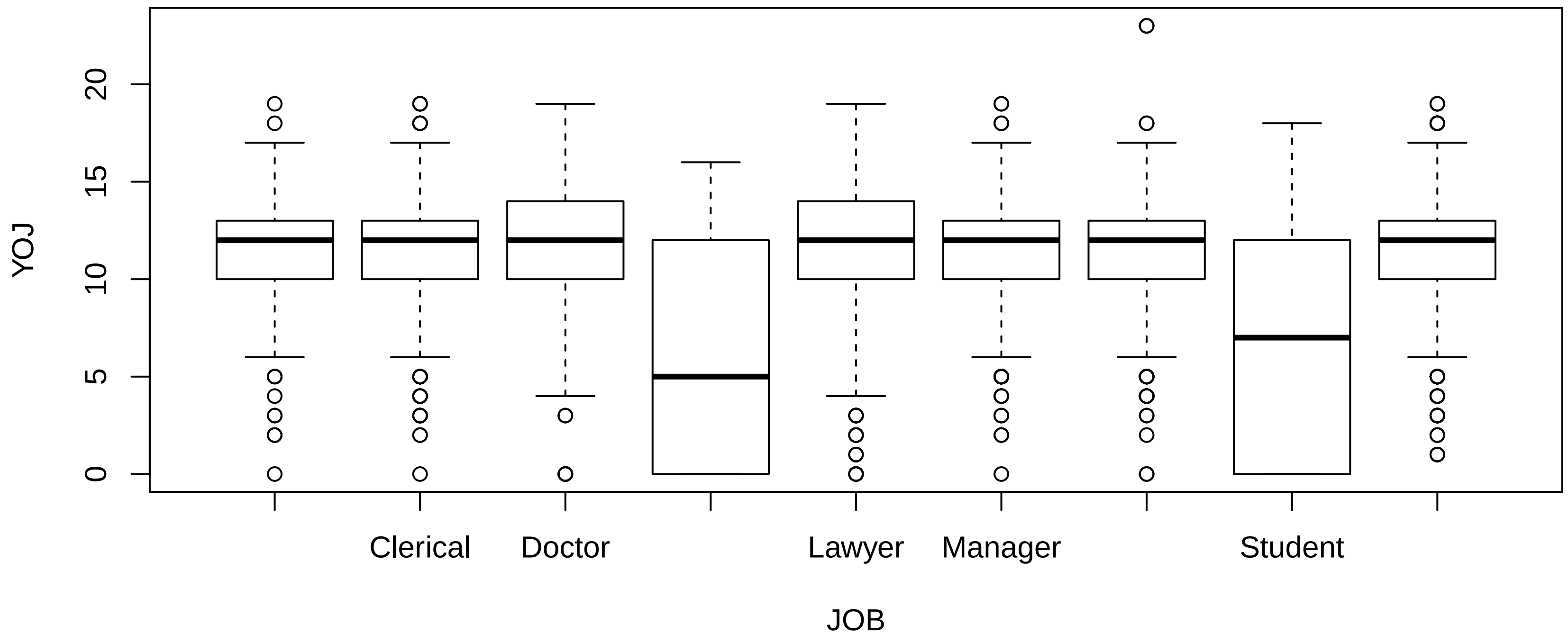
DATA PREPARATION

`AGE` Variable

Assigning a medium age would be appropriate given that there are only 6 records with missing values. Also those records either indicate having kids at home and/or being married and so assigning median age of 45 would seem reasonable.

YOJ (Years on Job) Variable

For the YOJ variable it would make sense to see how it is distributed accross different job types. Below the boxplot and aggregation table, against the JOB variable, show that the median values may be drastically different among different jobs. Therefore, assigning median values per job type rather than just the single, overall median value would be more appropriate.




```
##            JOB YOJ
## 1               12
## 2     Clerical  12
## 3       Doctor  12
## 4   Home Maker   5
## 5       Lawyer  12
```

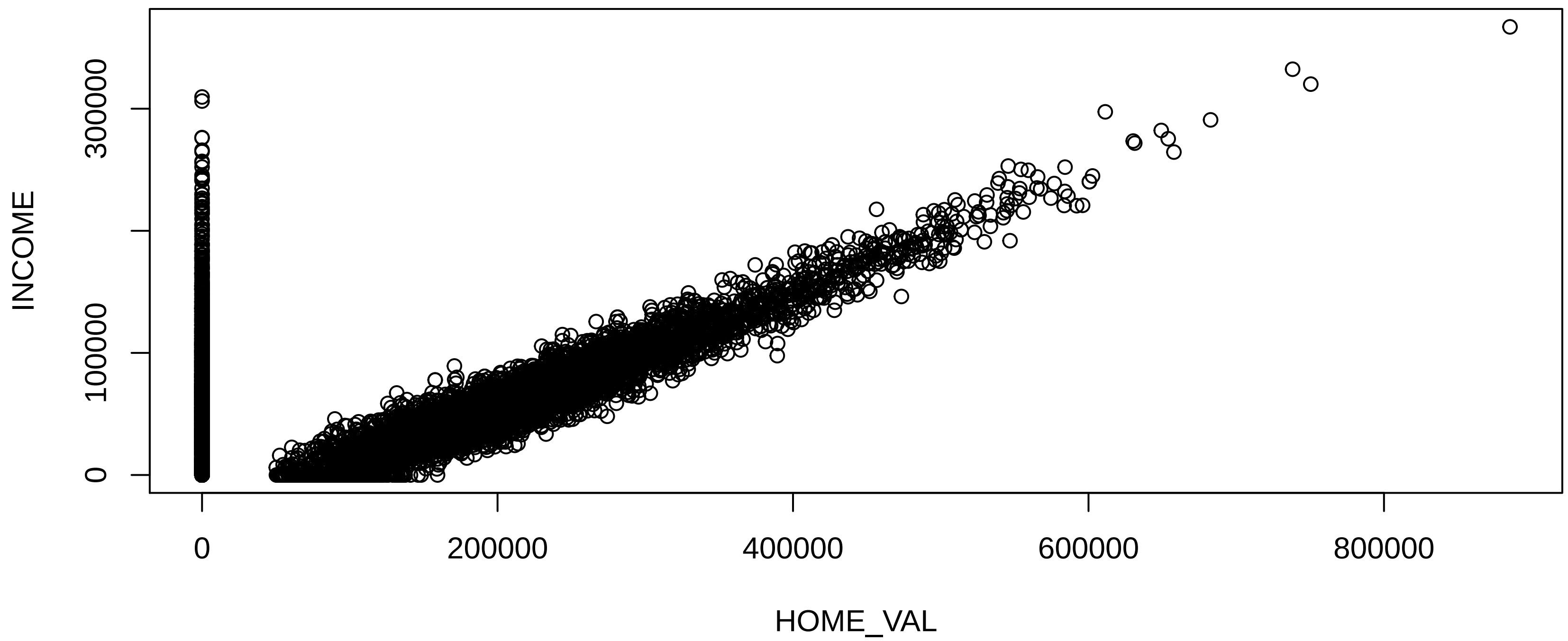
```
## 6       Manager  12
## 7  Professional  12
## 8       Student   7
## 9 z_Blue Collar  12
```

CAR_AGE Variable

Car age has some invalid negative values. We can assign them to NA and then deal with them as missing values. To deal with missing values of CAR_AGE, it may be a good idea to find a correlation with BLUEBOOK value and derive approximate values for the age. However, for this we would require knowing the make and model of the cars. Given that this information is not available to us and that it is considerable number of rows with the missing values, it may be best to simply asign median age.

INCOME and HOME_VAL Variables

Both the INCOME and the HOME_VAL variables have missing values. However there are only 33 instances where both variables jointly are missing values. Otherwise, individually, these variables have over 400 missing values. It would be no surprise, however, that the two variables are positively correlated, because the higher the income, the more expesive a home value can be expected. The plot below does show this correlation indeed.



Given such correlation, it may be possible to come up with an impute strategy where the two variables can help each other. We will be making an assumption here that the `HOME_VAL` variable with value of 0 is considered to indicate that someone is not a home owner. Therefore, we can design to execute the following strategy for imputing these two variables:

1. For the 33 instances where both are missing, randomly assign a value to `HOME_VAL` variable choosing between 0 and median home value.
2. Build a simple linear model to predict income values based on the home value (i.e. where home value $>$ 0). Any predicted negative amounts should be changed to 0.
3. Use median income for the remaining missing income values.
4. Finally, to avoid having two highly correlated variables, replace `HOME_VAL` variable with a new variable called, `HOME_OWN`, by transforming the `HOME_VAL` variable to a 0 or 1 binary indicator (0=*not a home owner*). Any missing values are to be randomly assigned to 0 or 1.

Before moving on, it would also make sense to create a new variable, `INCOME_CLASS`, by transforming the `INCOME` variable from being a continuous numeric variable into a categorical 3 level **(LOW, MID, HIGH)** variable. Using `INCOME` variable with exact numerical values, would not make sense as a predictor for the kind of responses we want to predict. Also, it would help us to deal with cases where income is entered as 0 value.

To create the 3 category levels, we used Inter-Quartile ranges, where below 25% would rank as `LOW`, above 75% would rank as `HIGH` and the rest is `MID`.

Before, moving on to building models, let's take the final look and validate the summary of the data. Note, that `INCOME` and `HOME_VAL` were replaced by `INCOME_CLASS` and `HOME_OWN` variables, respectively.

```
##   TARGET_FLG     TARGET_AMT        KIDSDRIV        AGE        HOMEKIDS          YOJ        PARENT1    MSTATUS
##  Min.   :0.00   Min.   :     0   Min.   :0.0   Min.   :16   Min.   :0.0   Min.   : 0.0   No :7084   Yes :4894
##  1st Qu.:0.00   1st Qu.:     0   1st Qu.:0.0   1st Qu.:39   1st Qu.:0.0   1st Qu.: 9.0   Yes:1077   z_No:3267
##  Median :0.00   Median :     0   Median :0.0   Median :45   Median :0.0   Median :12.0
##  Mean   :0.26   Mean   :  1504   Mean   :0.2   Mean   :45   Mean   :0.7   Mean   :10.5
##  3rd Qu.:1.00   3rd Qu.:  1036   3rd Qu.:0.0   3rd Qu.:51   3rd Qu.:1.0   3rd Qu.:13.0
##  Max.   :1.00   Max.   :107586   Max.   :4.0   Max.   :81   Max.   :5.0   Max.   :23.0
##
##   SEX               EDUCATION               JOB         TRAVTIME          CAR_USE         BLUEBOOK          TIF
##  M  :3786   <High School :1203   z_Blue Collar:1825   Min.   :  5   Commercial:3029   Min.   : 1500   Min.   : 1.0
##  z_F:4375   Bachelors    :2242   Clerical     :1271   1st Qu.: 22   Private   :5132   1st Qu.: 9280   1st Qu.: 1.0
##             Masters      :1658   Professional :1117   Median : 33                     Median :14440   Median : 4.0
##             PhD          : 728   Manager      : 988   Mean   : 33                     Mean   :15710   Mean   : 5.4
##             z_High School:2330   Lawyer       : 835   3rd Qu.: 44                     3rd Qu.:20850   3rd Qu.: 7.0
##                                  Student      : 712   Max.   :142                     Max.   :69740   Max.   :25.0
##                                  (Other)      :1413
##        CAR_TYPE    RED_CAR      OLDCLAIM         CLM_FREQ     REVOKED       MVR_PTS         CAR_AGE
##  Minivan    :2145   no :5783   Min.   :    0   Min.   :0.0   No :7161   Min.   : 0.0   Min.   : 0.0
##  Panel Truck: 676   yes:2378   1st Qu.:    0   1st Qu.:0.0   Yes:1000   1st Qu.: 0.0   1st Qu.: 4.0
##  Pickup     :1389              Median :    0   Median :0.0              Median : 1.0   Median : 8.0
##  Sports Car : 907              Mean   : 4037   Mean   :0.8              Mean   : 1.7   Mean   : 8.3
##  Van        : 750              3rd Qu.: 4636   3rd Qu.:2.0              3rd Qu.: 3.0   3rd Qu.:12.0
##  z_SUV      :2294              Max.   :57037   Max.   :5.0              Max.   :13.0   Max.   :28.0
##
##                  URBANICITY      HOME_OWN      INCOME_CLASS
##  Highly Urban/ Urban  :6492   Min.   :0.00   HIGH:2040
##  z_Highly Rural/ Rural:1669   1st Qu.:0.00   LOW :2040
##                               Median :1.00   MID :4081
##                               Mean   :0.69
##                               3rd Qu.:1.00
##                               Max.   :1.00
##
```

BUILD MODELS

To model prediction of the quantitative variable, `TARGET_AMT`, we started off with a simple linear model including all the variables. Progressing with stepwise, backward elimination, we arrived at our first model with reduced set of variables which are statistically significant. Here's the summary of this `LM` model.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG - RED_CAR - YOJ - AGE -
##     HOMEKIDS - EDUCATION - HOME_OWN - OLDCLAIM - BLUEBOOK - SEX,
##     data = m1.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5763  -1697   -756    341 103683
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       701.99     340.69    2.06  0.03938 *
## KIDSDRIV                          376.06     102.08    3.68  0.00023 ***
## PARENT1Yes                        639.02     176.51    3.62  0.00030 ***
## MSTATUSz_No                       596.81     119.32    5.00  5.8e-07 ***
## JOBClerical                       387.52     290.98    1.33  0.18298
## JOBDoctor                        -322.84     376.10   -0.86  0.39070
## JOBHome Maker                     275.73     335.53    0.82  0.41122
## JOBLawyer                         202.46     286.09    0.71  0.47915
## JOBManager                       -647.04     266.98   -2.42  0.01539 *
## JOBProfessional                   196.57     266.23    0.74  0.46031
## JOBStudent                        303.66     329.97    0.92  0.35746
## JOBz_Blue Collar                  326.47     266.05    1.23  0.21983
## TRAVTIME                           11.95       3.22    3.71  0.00021 ***
## CAR_USEPrivate                   -729.91     156.98   -4.65  3.4e-06 ***
## TIF                               -46.91      12.17   -3.85  0.00012 ***
## CAR_TYPEPanel Truck               565.05     243.07    2.32  0.02012 *
## CAR_TYPEPickup                    382.00     168.03    2.27  0.02303 *
## CAR_TYPESports Car                775.83     182.90    4.24  2.2e-05 ***
## CAR_TYPEVan                       671.03     204.20    3.29  0.00102 **
## CAR_TYPEz_SUV                     509.08     138.86    3.67  0.00025 ***
## CLM_FREQ                          106.91      48.83    2.19  0.02858 *
## REVOKEDYes                        447.98     154.91    2.89  0.00384 **
```

```
## MVR_PTS                          172.16      25.80    6.67  2.7e-11 ***
## CAR_AGE                          -28.16      11.23   -2.51  0.01220 *
## URBANICITYz_Highly Rural/ Rural -1659.46     139.33  -11.91  < 2e-16 ***
## INCOME_CLASSLOW                  460.24     206.89    2.22  0.02614 *
## INCOME_CLASSMID                  413.54     139.27    2.97  0.00299 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4550 on 8134 degrees of freedom
## Multiple R-squared:  0.0693, Adjusted R-squared:  0.0663
## F-statistic: 23.3 on 26 and 8134 DF,  p-value: <2e-16
```

Following similar progression for predicting the binary outcome of the `TARGET_FLAG` variable, here's the summary of the binomial logistic regression model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT - RED_CAR - CAR_AGE -
##     AGE - SEX - YOJ - HOMEKIDS, family = "binomial", data = m1.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.652  -0.714  -0.400   0.616   3.121
##
## Coefficients:
##                                  Estimate  Std. Error z value Pr(>|z|)
## (Intercept)                    -1.64727805  0.27099684   -6.08  1.2e-09 ***
## KIDSDRIV                        0.41921643  0.05510880    7.61  2.8e-14 ***
## PARENT1Yes                      0.46927171  0.09458331    4.96  7.0e-07 ***
## MSTATUSz_No                     0.45818984  0.08099350    5.66  1.5e-08 ***
## EDUCATIONBachelors             -0.38803071  0.11086150   -3.50  0.00047 ***
## EDUCATIONMasters               -0.29468590  0.16261314   -1.81  0.06996 .
## EDUCATIONPhD                   -0.24447866  0.19715020   -1.24  0.21495
## EDUCATIONz_High School          0.01430863  0.09702104    0.15  0.88275
## JOBClerical                     0.44312055  0.19572891    2.26  0.02358 *
## JOBDoctor                      -0.38539517  0.26530244   -1.45  0.14632
## JOBHome Maker                   0.34686591  0.20731896    1.67  0.09431 .
## JOBLawyer                       0.12064837  0.16897613    0.71  0.47523
## JOBManager                     -0.54503073  0.17089034   -3.19  0.00143 **
## JOBProfessional                 0.17462363  0.17804308    0.98  0.32669
```

```
## JOBStudent                        0.25322413  0.21737713    1.16  0.24406
## JOBz_Blue Collar                  0.33004455  0.18514993    1.78  0.07465 .
## TRAVTIME                          0.01451641  0.00188272    7.71  1.3e-14 ***
## CAR_USEPrivate                   -0.76349103  0.09177710   -8.32  < 2e-16 ***
## BLUEBOOK                         -0.00002361  0.00000469   -5.03  4.9e-07 ***
## TIF                              -0.05510047  0.00734730   -7.50  6.4e-14 ***
## CAR_TYPEPanel Truck               0.59382771  0.15106679    3.93  8.5e-05 ***
## CAR_TYPEPickup                    0.55035626  0.10070483    5.47  4.6e-08 ***
## CAR_TYPESports Car                0.96823935  0.10756994    9.00  < 2e-16 ***
## CAR_TYPEVan                       0.66067051  0.12229298    5.40  6.6e-08 ***
## CAR_TYPEz_SUV                     0.70686831  0.08612268    8.21  2.3e-16 ***
## OLDCLAIM                         -0.00001391  0.00000391   -3.56  0.00037 ***
## CLM_FREQ                          0.19554013  0.02854170    6.85  7.3e-12 ***
## REVOKEDYes                        0.88902639  0.09131311    9.74  < 2e-16 ***
## MVR_PTS                           0.11276894  0.01360639    8.29  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.40193782  0.11305849  -21.25  < 2e-16 ***
## HOME_OWN                         -0.30855951  0.07951630   -3.88  0.00010 ***
## INCOME_CLASSLOW                   0.64146723  0.12659960    5.07  4.0e-07 ***
## INCOME_CLASSMID                   0.45652770  0.08834417    5.17  2.4e-07 ***
## ---
## 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: 7292.9  on 8128  degrees of freedom
## AIC: 7359
##
## Number of Fisher Scoring iterations: 5
```

We'd like to see if we can possibly enhance and build additional models. Looking at both models, it appears that having a job as a Manager has the most statistical signfinicanse for our predictions. In both cases, the coefficients are negative, which seems to suggest that if you're a manager, then you're more likely to be a more responsible and a less risky driver. This made for an unanticipated, but a reasonable discovery, nevertheless. So, it may be a good idea to simplify the `JOB` predictor into a binary category of "Not Manager" and "Manager".

This resulted in `LM` [`TARGET_AMT`] model, where all the remaing variables are being significant as shown in the summary below.

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + PARENT1 + MSTATUS + JOB +
```

```
##     TRAVTIME + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY + INCOME_CLASS, data = m2.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5737  -1705   -763    346 103586
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      954.56     244.44    3.91  9.5e-05 ***
## KIDSDRIV                         382.80     101.97    3.75  0.00018 ***
## PARENT1Yes                       662.28     175.92    3.76  0.00017 ***
## MSTATUSz_No                      580.94     119.01    4.88  1.1e-06 ***
## JOBManager                      -836.38     161.65   -5.17  2.3e-07 ***
## TRAVTIME                          12.09       3.22    3.76  0.00017 ***
## CAR_USEPrivate                  -767.31     127.11   -6.04  1.6e-09 ***
## TIF                              -46.57      12.16   -3.83  0.00013 ***
## CAR_TYPEPanel Truck              495.98     226.91    2.19  0.02886 *
## CAR_TYPEPickup                   359.65     165.09    2.18  0.02940 *
## CAR_TYPESports Car               770.42     181.77    4.24  2.3e-05 ***
## CAR_TYPEVan                      635.81     200.32    3.17  0.00151 **
## CAR_TYPEz_SUV                    505.31     137.87    3.67  0.00025 ***
## CLM_FREQ                         106.01      48.78    2.17  0.02981 *
## REVOKEDYes                       455.72     154.77    2.94  0.00324 **
## MVR_PTS                          172.80      25.78    6.70  2.2e-11 ***
## CAR_AGE                          -35.43      10.00   -3.54  0.00040 ***
## URBANICITYz_Highly Rural/ Rural -1618.30     136.98  -11.81  < 2e-16 ***
## INCOME_CLASSLOW                  577.11     162.44    3.55  0.00038 ***
## INCOME_CLASSMID                  503.20     132.03    3.81  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4540 on 8141 degrees of freedom
## Multiple R-squared:  0.0687, Adjusted R-squared:  0.0666
## F-statistic: 31.6 on 19 and 8141 DF,  p-value: <2e-16
```

When applied to the binomial model, the newly transformed `JOB` variable resulted in higher significance for the `EDUCATION` variable for levels higher than "High School". Here's the summary of the model illustrating this point.

```
##
```

```
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT - RED_CAR - CAR_AGE -
##     AGE - SEX - YOJ - HOMEKIDS, family = "binomial", data = m2.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.627  -0.716  -0.404   0.621   3.086
##
## Coefficients:
##                                    Estimate  Std. Error z value Pr(>|z|)
## (Intercept)                     -1.36531671  0.19488940   -7.01  2.5e-12 ***
## KIDSDRIV                         0.42234080  0.05504577    7.67  1.7e-14 ***
## PARENT1Yes                       0.47857988  0.09426037    5.08  3.8e-07 ***
## MSTATUSz_No                      0.46811874  0.07896011    5.93  3.1e-09 ***
## EDUCATIONBachelors              -0.46227899  0.10036479   -4.61  4.1e-06 ***
## EDUCATIONMasters                -0.50466757  0.11042270   -4.57  4.9e-06 ***
## EDUCATIONPhD                    -0.60280313  0.14450345   -4.17  3.0e-05 ***
## EDUCATIONz_High School          -0.01249518  0.09343938   -0.13  0.89362
## JOBManager                      -0.73903516  0.10688093   -6.91  4.7e-12 ***
## TRAVTIME                         0.01459000  0.00188010    7.76  8.5e-15 ***
## CAR_USEPrivate                  -0.77367365  0.07415300  -10.43  < 2e-16 ***
## BLUEBOOK                        -0.00002344  0.00000467   -5.02  5.2e-07 ***
## TIF                             -0.05459858  0.00733614   -7.44  9.9e-14 ***
## CAR_TYPEPanel Truck              0.56928353  0.14358998    3.96  7.4e-05 ***
## CAR_TYPEPickup                   0.54223581  0.09858920    5.50  3.8e-08 ***
## CAR_TYPESports Car               0.97461278  0.10657818    9.14  < 2e-16 ***
## CAR_TYPEVan                      0.64772440  0.11993475    5.40  6.6e-08 ***
## CAR_TYPEz_SUV                    0.71207599  0.08540574    8.34  < 2e-16 ***
## OLDCLAIM                        -0.00001375  0.00000391   -3.52  0.00043 ***
## CLM_FREQ                         0.19488606  0.02848421    6.84  7.8e-12 ***
## REVOKEDYes                       0.88841422  0.09120179    9.74  < 2e-16 ***
## MVR_PTS                          0.11254815  0.01357604    8.29  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.38612901  0.11288604  -21.14  < 2e-16 ***
## HOME_OWN                        -0.27563005  0.07355723   -3.75  0.00018 ***
## INCOME_CLASSLOW                  0.70197581  0.10762896    6.52  6.9e-11 ***
## INCOME_CLASSMID                  0.48967278  0.08755130    5.59  2.2e-08 ***
## ---
## 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: 7305.1  on 8135  degrees of freedom
## AIC: 7357
##
## Number of Fisher Scoring iterations: 5
```

Interestingly, and is likely to be expected, the higher the education level, the more negative the coefficients' trend is. This again suggests that more educated people tend to be less likely to end up with a car accident. Therefore, similar to how we transformed the `JOB` variable, it made sense to transform `EDUCATION` to just two values, "Lower" and "Higher" ("Higher" standing for Bachelors and above). And again we ended up with a model where all the remaing variables ended up being significant.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + EDUCATION +
##     JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + URBANICITY + HOME_OWN + INCOME_CLASS,
##     family = "binomial", data = m2.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.627  -0.715  -0.403   0.619   3.093
##
## Coefficients:
##                                Estimate  Std. Error z value Pr(>|z|)
## (Intercept)                 -1.39442924  0.17815005   -7.83  5.0e-15 ***
## KIDSDRIV                     0.42365545  0.05504047    7.70  1.4e-14 ***
## PARENT1Yes                   0.48591470  0.09398333    5.17  2.3e-07 ***
## MSTATUSz_No                  0.46607963  0.07892966    5.90  3.5e-09 ***
## EDUCATIONHigher             -0.47840221  0.06734622   -7.10  1.2e-12 ***
## JOBManager                  -0.73515374  0.10651970   -6.90  5.1e-12 ***
## TRAVTIME                     0.01465646  0.00187898    7.80  6.2e-15 ***
## CAR_USEPrivate              -0.78183466  0.07079454  -11.04  < 2e-16 ***
## BLUEBOOK                    -0.00002374  0.00000466   -5.10  3.5e-07 ***
## TIF                         -0.05464261  0.00733540   -7.45  9.4e-14 ***
## CAR_TYPEPanel Truck          0.55937446  0.14206413    3.94  8.2e-05 ***
## CAR_TYPEPickup               0.53670833  0.09796521    5.48  4.3e-08 ***
## CAR_TYPESports Car           0.97139500  0.10650310    9.12  < 2e-16 ***
## CAR_TYPEVan                  0.64381021  0.11940622    5.39  7.0e-08 ***
```

```
## CAR_TYPEz_SUV                      0.71037810  0.08534269    8.32  < 2e-16 ***
## OLDCLAIM                          -0.00001369  0.00000390   -3.51  0.00045 ***
## CLM_FREQ                           0.19420069  0.02845580    6.82  8.8e-12 ***
## REVOKEDYes                         0.88858936  0.09117146    9.75  < 2e-16 ***
## MVR_PTS                            0.11263626  0.01357363    8.30  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.38394727  0.11289827  -21.12  < 2e-16 ***
## HOME_OWN                          -0.27306903  0.07351143   -3.71  0.00020 ***
## INCOME_CLASSLOW                    0.73107243  0.10375650    7.05  1.8e-12 ***
## INCOME_CLASSMID                    0.51779570  0.08364954    6.19  6.0e-10 ***
## ---
## 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: 7306.3  on 8138  degrees of freedom
## AIC: 7352
##
## Number of Fisher Scoring iterations: 5
```

SELECT MODELS

For both types of models (linear and logistic), the selection came down to the last versions of the models generated after all of the variable reductions and tranformations took place. In case of `LM` model the *Adjusted R-squared* value was slightly improved in the latest model. The bottom line is that the selection was mainly due to favoring more of a simpler model, with less variables, rather than due to statistical evaluations as those were very similar between the model versions.

APPENDIX - R statistical programming code

```r
library(knitr)
library(kableExtra)
library(plyr)
library(tidyverse)
library(corrplot)
library(reshape2)
library(ggplot2)

# Load dataset definition
url <- 'insurance_dataset_definition.csv'
```

```
ds <- read.csv(url, header = TRUE);
ds

# Load training dataset
url <- 'insurance_training_data.csv'
df <- read.csv(url, header = TRUE, row.names = 'INDEX')
head(df)
summary(df)

# Parse Numerical Data
# INCOME
df$INCOME <- parse_number(as.character(df$INCOME))
# HOME_VAL
df$HOME_VAL <- parse_number(as.character(df$HOME_VAL))
# BLUEBOOK
df$BLUEBOOK <- parse_number(as.character(df$BLUEBOOK))
# OLDCLAIM
df$OLDCLAIM <- parse_number(as.character(df$OLDCLAIM))
df %>% select(INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM) %>% summary()

# Show Correlation
cor.data <- df %>% select(TARGET_FLAG,TARGET_AMT,KIDSDRIV,AGE,HOMEKIDS,
                          YOJ,INCOME,HOME_VAL,TRAVTIME,BLUEBOOK,TIF,OLDCLAIM,
                          CLM_FREQ,MVR_PTS,CAR_AGE) %>% na.omit() %>%  cor()
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor.data, method = "shade", shade.col = NA, tl.col = "black",
         tl.srt = 45, col = col(200), addCoef.col = "black", cl.pos = "n",
         order = "original", type = "upper", addCoefasPercent = T)

# impute AGE with median value
median_age <- summary(df$AGE)[['Median']]
df[is.na(df$AGE),]['AGE'] <- median_age

# Box Plot of YOJ over JOB
plot(YOJ ~ JOB, df)
aggregate(YOJ ~ JOB, df, median)
# Imputing YOJ with median value per job
df_tmp <- df %>% group_by(JOB) %>%
```

```r
  mutate(NEW_YOJ = median(YOJ, na.rm = TRUE)) %>%
  select(JOB, YOJ, NEW_YOJ)
df[is.na(df$YOJ),]$YOJ <- df_tmp[is.na(df_tmp$YOJ),]$NEW_YOJ

# Impute `CAR_AGE` missing values
df$CAR_AGE[which(df$CAR_AGE < 0)] <- NA
median_car_age <- summary(df$CAR_AGE)[['Median']]
df[is.na(df$CAR_AGE),]['CAR_AGE'] <- median_car_age

# Transform INCOME and HOME_VAL
nrow_na <- nrow(df[is.na(df$INCOME) & is.na(df$HOME_VAL),])
plot(INCOME~HOME_VAL, df)
# 1
median_home_val <- summary(df$HOME_VAL)[['Median']]
df[is.na(df$INCOME) & is.na(df$HOME_VAL),]$HOME_VAL <- sample(c(0, median_home_val),
                                                              size=nrow_na, replace = T)
# 2
lm_data <- df[df$HOME_VAL > 0,]
lm1 <- lm(INCOME~HOME_VAL, data = lm_data)
lm1.predict <- predict(lm1, newdata = df[is.na(df$INCOME) & df$HOME_VAL > 0,]['HOME_VAL'])
df[is.na(df$INCOME) & df$HOME_VAL > 0,]$INCOME <- lm1.predict
rm(lm_data, lm1)
# deal with negative values
df[!is.na(df$INCOME) & df$INCOME < 0,]$INCOME <- 0
# 3
median_income <- summary(df$INCOME)[['Median']]
df[is.na(df$INCOME),]$INCOME <- median_income
# 4
df$HOME_OWN <- ifelse(df$HOME_VAL > 0, 1, 0)
# deal with missing values
nrow_na <- nrow(df[is.na(df$HOME_OWN),])
df[is.na(df$HOME_OWN),]$HOME_OWN <- sample(c(0, 1), size=nrow_na, replace = T)

# create INCOME_CLASS
sum_income <- summary(df$INCOME)
low_income_ub <- sum_income[['1st Qu.']]
high_income_lb <- sum_income[['3rd Qu.']]
rm(sum_income)
```

```
df$INCOME_CLASS <- as.factor(case_when(
  df$INCOME < low_income_ub ~ 'LOW',
  df$INCOME > high_income_lb ~ 'HIGH',
  TRUE ~ 'MID'))

# validate new model summary
df_train <- select(df, -'INCOME', -'HOME_VAL')
summary(df_train)

## Build Models
# Build first LM
m1.data <- df_train
m1.lm <- lm(TARGET_AMT ~ . -TARGET_FLAG-RED_CAR-YOJ-AGE-HOMEKIDS-EDUCATION-HOME_OWN
            -OLDCLAIM-BLUEBOOK-SEX, data = m1.data)
summary(m1.lm)

# Build first Logistic Model
b1.lm <- glm(formula = TARGET_FLAG ~ . -TARGET_AMT-RED_CAR-CAR_AGE-AGE-SEX-YOJ-HOMEKIDS,
             family = "binomial", data = m1.data)
summary(b1.lm)

# Transform JOB variable
m2.data = m1.data
m2.data$JOB <- factor(ifelse(m2.data$JOB != "Manager", "Not Manager", "Manager"),
                      levels = c("Not Manager", "Manager"))
# Build second LM
m2.lm <- lm(TARGET_AMT ~ . -TARGET_FLAG-RED_CAR-YOJ-AGE-HOMEKIDS-EDUCATION-HOME_OWN
            -OLDCLAIM-BLUEBOOK-SEX, data = m2.data)
m2.lm <- update(m2.lm, .~. -TARGET_FLAG-RED_CAR-YOJ-AGE-HOMEKIDS-EDUCATION-HOME_OWN
                -OLDCLAIM-BLUEBOOK-SEX, data = m2.data)
summary(m2.lm)

# Build second Logistic Model
b2.lm <- glm(formula = TARGET_FLAG ~ . -TARGET_AMT-RED_CAR-CAR_AGE-AGE-SEX-YOJ-HOMEKIDS,
             family = "binomial", data = m2.data)
summary(b2.lm)

# Tranform EDUCATION variable
```

```r
m2.data$EDUCATION <- mapvalues(m2.data$EDUCATION,
                               c("<High School", "Bachelors", "Masters",
                                 "PhD", "z_High School"),
                               c("Lower", "Higher", "Higher", "Higher", "Lower"))

# Build third Logistic Model
b2.lm <- glm(formula = TARGET_FLAG ~ . -TARGET_AMT-RED_CAR-CAR_AGE-AGE-SEX-YOJ-HOMEKIDS,
             family = "binomial", data = m2.data)
b2.lm <- update(b2.lm, .~. -TARGET_AMT-RED_CAR-CAR_AGE-AGE-SEX-YOJ-HOMEKIDS,
                data = m2.data)
summary(b2.lm)
```

PREDICTIONS - R statistical programming code

```r
# Load Data
url <- './insurance-evaluation-data.csv'
df.fin <- read.csv(url, header = TRUE, row.names = 'INDEX')
df <- df.fin

## Prepare Data
# Transform EDUCATION
df$EDUCATION <- mapvalues(df$EDUCATION,
                          c("<High School", "Bachelors", "Masters", "PhD", "z_High School"),
                          c("Lower", "Higher", "Higher", "Higher", "Lower"))

# Transform JOB
df$JOB <- factor(ifelse(df$JOB != "Manager", "Not Manager", "Manager"),
                 levels = c("Not Manager", "Manager"))
levels(df$JOB)

# Parse INCOME
df$INCOME <- parse_number(as.character(df$INCOME))

# Parse HOME_VAL
df$HOME_VAL <- parse_number(as.character(df$HOME_VAL))

# Parse BLUEBOOK
df$BLUEBOOK <- parse_number(as.character(df$BLUEBOOK))
```

```
# Parse OLDCLAIM
df$OLDCLAIM <- parse_number(as.character(df$OLDCLAIM))

# Impout missing CAR_AGE
df[is.na(df$CAR_AGE),]['CAR_AGE'] <- median_car_age

# Impute missing INCOME data
# 1
nrow_na <- nrow(df[is.na(df$INCOME) & is.na(df$HOME_VAL),])
df[is.na(df$INCOME) & is.na(df$HOME_VAL),]$HOME_VAL <- sample(
  c(0, median_home_val), size=nrow_na, replace = T)

# 2
lm_data <- df[df$HOME_VAL > 0,]
lm1.predict <- predict(lm1, newdata = df[is.na(df$INCOME) & df$HOME_VAL > 0,]['HOME_VAL'])
df[is.na(df$INCOME) & df$HOME_VAL > 0,]$INCOME <- lm1.predict
# deal with negative values
df[!is.na(df$INCOME) & df$INCOME < 0,]$INCOME <- 0

# 3
df[is.na(df$INCOME),]$INCOME <- median_income

# 4
df$HOME_OWN <- ifelse(df$HOME_VAL > 0, 1, 0)
# deal with missing values
nrow_na <- nrow(df[is.na(df$HOME_OWN),])
df[is.na(df$HOME_OWN),]$HOME_OWN <- sample(c(0, 1), size=nrow_na, replace = T)
summary(df$HOME_OWN)

# Create INCOME_CLASS
df$INCOME_CLASS <- as.factor(case_when(
  df$INCOME < low_income_ub ~ 'LOW',
  df$INCOME > high_income_lb ~ 'HIGH',
  TRUE ~ 'MID'))

# str(df)
# summary(df)
```

```r
m.predict <- predict(m2.lm, newdata = df)
b.predict <- predict(b2.lm, newdata = df)

df.fin$TARGET_FLAG <- ifelse(b.predict > .5, 1, 0)
df.fin$TARGET_AMT <- m.predict
df.fin[df.fin$TARGET_FLAG == 0,]$TARGET_AMT <- ''
write.csv(df.fin, "insurance-evaluation-data-completed.csv")
```