**Ji Jung**

**Predict 411 SEC 56**

# Auto Insurance Problem

## Introduction

Auto Insurance Problem data set contains approximately 8,100 records presenting customers at an auto insurance company. Each record has two target variables, TARGET_FLAG which indicates if a customer has been involved in a car crash and the second variable of TARGET_AMT which would be zero if a customer has not crashed his/her vehicle, or be greater than zero if there has been a crash.

For this assignment, the goal is to build components for a probability/severity model by constructing a linear regression to estimate the probability that a customer will be involved in an accident, and a model to estimate the cost in the event of a crash.

## Introduction to Exploratory Data Analysis and Data Preparation

Some background will be presented followed by an initial examination. Following steps will occur after;

- Obtain histograms for all continuous variables. (seeking variability)
- Examine the bi-serial correlations for each continuous variable and the dependent variable. (seeking higher correlation for better independent variable)
- Discretize each continuous variable with cut-points. (seeking dichotomous relationship to dependent variable)
- Examine the cross tabulation of the discretized variables. (seeking non-linear relationship to dependent variable)
- If there are any differences in the proportion profiles the originating variable will be incorporated into the model.
- May consider transformation of the originating variable.
- Obtain frequency counts for all categorical variables. (seeking variability)
- Obtain crosstabs of our dependent variable by each categorical variable. (seeking proportional variation)
- Code each categorical variable with dummy variables. (seeking individual explanation over dependent variable)

With these procedures, I will identify variables with a high variability for inclusion in the model. Not all diagnostics graphics will be included but I will attempt to convey the thorough nature of exploration.

## Data Background

A data dictionary is provided in this assignment which describes variables and provides an estimate at the theoretical effect.

*Table 1 Data Dictionary, with Data Types*

| VARIABLE NAME | TYPE | DEFINITION |
|---|---|---|
| AGE | continuous | Age of Driver |
| BLUEBOOK | continuous | Value of Vehicle |
| CAR_AGE | continuous | Vehicle Age |
| CAR_TYPE | categorical | Type of Car |
| CAR_USE | categorical | Vehicle Use |
| CLM_FREQ | continuous | #Claims(Past 5 Years) |
| EDUCATION | categorical | Max Education Level |
| HOMEKIDS | continuous | #Children @Home |
| HOME_VAL | continuous | Home Value |
| INCOME | continuous | Income |
| JOB | categorical | Job Category |
| KIDSDRIV | categorical | #Driving Children |
| MSTATUS | categorical | Marital Status |
| MVR_PTS | continuous | Motor Vehicle Record Points |
| OLDCLAIM | continuous | Total Claims (Past 5 Years) |
| PARENT1 | categorical | Single Parent |
| RED_CAR | categorical | A Red Car |
| REVOKED | categorical | License Revoked (Past 7 Years) |
| SEX | categorical | Gender |
| TIF | continuous | Time in Force |
| TRAVTIME | continuous | Distance to Work |
| URBANICITY | categorical | Home/Work Area |
| YOJ | continuous | Years on Job |

From Table 1, some variables which seem to be continuous from to their initial data type could be categorical instead, such as the number of children at home. While all variables will be equally attended to, if low correlation between the dependent variable and an indicative variable based on description, considerations will be made on how manipulations of such variable will turn out.

**Exploratory Data Analysis on Continuous Variables**

First step is an examination of continuous variable means with respect to the dependent variable, shown under Table 2.

*Table 2 Means with respect to the Dependent Variable (Target Flag)*

| TARGET_FLAG | N Obs | Variable | Label | N | N Miss | Mean | Std Dev | Range | 5th Pctl | 95th Pctl |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 6008 | AGE | Age | 6007 | 1 | 45.322791 | 8.2022705 | 65.000000 | 32.000000 | 58.000000 |
| | | BLUEBOOK | Value of Vehicle | 6008 | 0 | 16230.95 | 8401.95 | 68240.00 | 5300.00 | 31500.00 |
| | | CAR_AGE | Vehicle Age | 5640 | 368 | 8.6709220 | 5.7201267 | 28.000000 | 1.0000000 | 18.000000 |
| | | CLM_FREQ | #Claims(Past 5 Years) | 6008 | 0 | 0.6486352 | 1.0860488 | 5.0000000 | 0.0000000 | 3.0000000 |
| | | HOMEKIDS | #Children @Home | 6008 | 0 | 0.6439747 | 1.0762090 | 5.0000000 | 0.0000000 | 3.0000000 |
| | | HOME_VAL | Home Value | 5665 | 343 | 169075.41 | 129938.83 | 885282.34 | 0.0000000 | 390219.07 |
| | | INCOME | Income | 5673 | 335 | 65951.97 | 48552.20 | 367030.26 | 0.0000000 | 159516.58 |
| | | MVR_PTS | Motor Vehicle Record Points | 6008 | 0 | 1.4137816 | 1.8916611 | 11.000000 | 0.0000000 | 5.0000000 |
| | | OLDCLAIM | Total Claims(Past 5 Years) | 6008 | 0 | 3311.59 | 8143.61 | 53986.00 | 0.0000000 | 22642.00 |
| | | TIF | Time in Force | 6008 | 0 | 5.5557590 | 4.2020970 | 24.000000 | 1.0000000 | 13.000000 |
| | | TRAVTIME | Distance to Work | 6008 | 0 | 33.030344 | 16.131290 | 137.12063 | 6.3550660 | 60.506403 |
| | | YOJ | Years on Job | 5677 | 331 | 10.671833 | 3.9175259 | 23.000000 | 0 | 15.000000 |

| TARGET_FLAG | N Obs | Variable | Label | N | N Miss | Mean | Std Dev | Range | 5th Pctl | 95th Pctl |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2153 | AGE | Age | 2148 | 5 | 43.3012104 | 9.5646287 | 60.0000000 | 27.0000000 | 60.0000000 |
| | | BLUEBOOK | Value of Vehicle | 2153 | 0 | 14255.90 | 8299.81 | 60740.00 | 4100.00 | 29330.00 |
| | | CAR_AGE | Vehicle Age | 2013 | 142 | 7.3674789 | 5.5353874 | 28.0000000 | 1.0000000 | 18.0000000 |
| | | CLM_FREQ | #Claims(Past 5 Years) | 2153 | 0 | 1.2169066 | 1.2483641 | 5.0000000 | 0 | 3.0000000 |
| | | HOMEKIDS | #Children @Home | 2153 | 0 | 0.9368323 | 1.1954470 | 5.0000000 | 0 | 3.0000000 |
| | | HOME_VAL | Home Value | 2033 | 121 | 115256.55 | 118150.14 | 750455.22 | 0 | 311529.40 |
| | | INCOME | Income | 2043 | 110 | 50641.30 | 42782.04 | 320126.98 | 0 | 130380.98 |
| | | MVR_PTS | Motor Vehicle Record Points | 2153 | 0 | 2.4816535 | 2.5791851 | 13.0000000 | 0 | 8.0000000 |
| | | OLDCLAIM | Total Claims(Past 5 Years) | 2153 | 0 | 6061.55 | 10071.09 | 57037.00 | 0 | 32025.00 |
| | | TIF | Time in Force | 2153 | 0 | 4.7807710 | 3.9329017 | 20.0000000 | 1.0000000 | 13.0000000 |
| | | TRAVTIME | Distance to Work | 2153 | 0 | 34.768120 | 15.185385 | 91.6143255 | 9.6941973 | 60.340343 |
| | | YOJ | Years on Job | 2030 | 123 | 10.016748 | 4.5122598 | 19.0000000 | 0 | 15.0000000 |

From Table 2, several variables have missing records. More importantly, there are noticeable differences in means between 0/1 target flag although they are within the variable standard deviation.

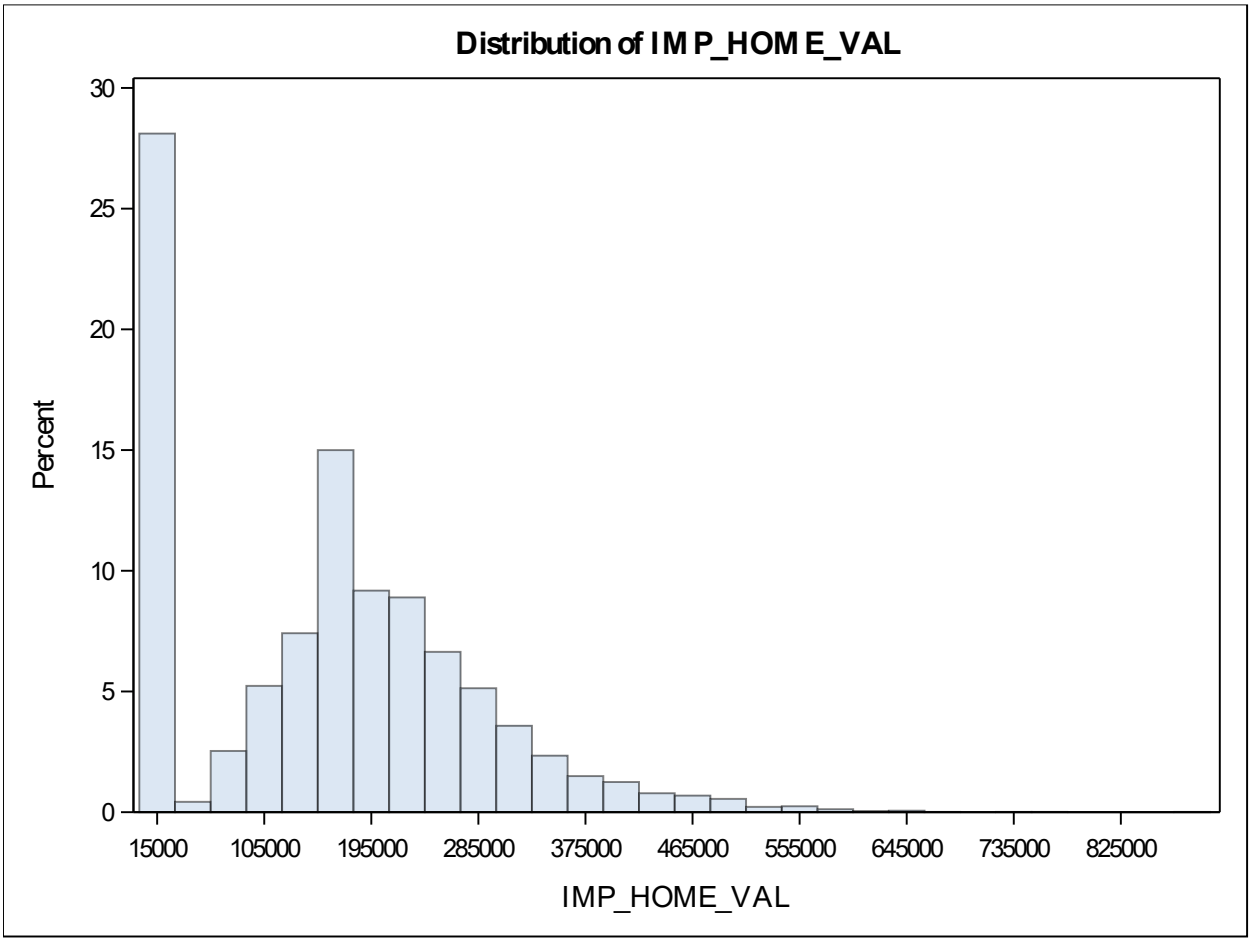
Of missing records, no variables are exceedingly large proportions within the observed data. Therefore, a blanket approach will be utilized in imputation of the mean. While there may be deeper consequences in assigning the mean to imputed values, it may be better than removing the observations from the data set.

Histograms of variables will be made as tests for normality. Some variables appear to resemble categorical variables, notably CLM_FREQ, HOMEKIDS, MVR_PTS. It seems reasonable to pull them to the categorical side of the analysis. However, they are much easier to interpret if they are used as continuous variables. It is easier to understand what one unit increase will affect in these variables.

From the histograms, three variables have very strong presences to the left side; IMP_CAR_AGE, IMP_HOME_VAL, and OLDCLAIM. For the variable IMP_CAR_AGE, there appears to be little data prior to

one year of car age. For IMP_HOME_VAL, it can be argued that survey method may be behind in creating this significant left side presence.

*Figure 1Distribution of IMP_HOME_VAL*




It is highly likely that a driver without a house would report the value as zero, which will create a distortion in mean. Thus, it is an indicator variable for the home ownership. A pre-qualifier question as whether a driver owns a house may have deterred this distortion, followed by an approximation of home values.

This applies to the variable OLDCLAIM. A significant left presence in the variable likely indicates the lack of divided questions in the method of data collection instead of the measured phenomena. Thus, a potential transformation is to be considered only if there are promises for inclusion into the model.

The simple correlation between the continuous variables and the dependent variable is showed under Table 3.

*Table 3 Continuous Variable Correlation to TARGET_FLAG*

| Pearson Correlation Coefficients, N = 8161<br>Prob > \|r\| under H0: Rho=0 | | | | | |
|---|---|---|---|---|---|
| | **TARGET_FLAG** | **IMP_AGE** | **BLUEBOOK** | **IMP_CAR_AGE** | **CLM_FREQ** |
| **IMP_AGE** | -0.10313<br><.0001 | 1.00000 | 0.16492<br><.0001 | 0.17080<br><.0001 | -0.02408<br>0.0296 |
| **BLUEBOOK**<br>Value of Vehicle | -0.10338<br><.0001 | 0.16492<br><.0001 | 1.00000 | 0.18337<br><.0001 | -0.03634<br>0.0010 |
| **IMP_CAR_AGE** | -0.09734<br><.0001 | 0.17080<br><.0001 | 0.18337<br><.0001 | 1.00000 | -0.00902<br>0.4151 |
| **IMP_HOME_VAL** | -0.17848<br><.0001 | 0.20422<br><.0001 | 0.25181<br><.0001 | 0.20487<br><.0001 | -0.09143<br><.0001 |
| **IMP_INCOME** | -0.13824<br><.0001 | 0.17614<br><.0001 | 0.41849<br><.0001 | 0.39093<br><.0001 | -0.04655<br><.0001 |
| **OLDCLAIM**<br>Total Claims(Past 5 Years) | 0.13808<br><.0001 | -0.02928<br>0.0082 | -0.02952<br>0.0077 | -0.01300<br>0.2402 | 0.49513<br><.0001 |
| **TIF**<br>Time in Force | -0.08237<br><.0001 | -0.00007<br>0.9952 | -0.00542<br>0.6242 | 0.00752<br>0.4971 | -0.02302<br>0.0375 |
| **TRAVTIME**<br>Distance to Work | 0.04815<br><.0001 | 0.00560<br>0.6131 | -0.01680<br>0.1292 | -0.03680<br>0.0009 | 0.00641<br>0.5625 |
| **IMP_YOJ** | -0.06849<br><.0001 | 0.13198<br><.0001 | 0.13961<br><.0001 | 0.05760<br><.0001 | -0.02554<br>0.0210 |

Variables will move forward into the model construction phase if it exceeds the 0.10 threshold. The travel time has low correlation. While the concept of longer travel may seem to increase the likelihood of a crash, variables TIF, TRAVTIME and YOJ will not be considered for usage.

From the cross correlations, there are some potential issues with variables. Decisions will be held off until the model construction stage and after computing the variance inflation factor.

Next step is calculation of point bi-serial correlations. First, creation of indicator variable families for each of the continuous variables will be made. Points will be made along the Q1, Mean and Q3 quantiles. For some variables, such as IMP_HOME_VAL, other quartiles must be considered due to skewness.

Examining the point bi-serial correlations show the proportion profiles. Each continuous variable is broken into four points. From observations, there are different proportion profiles in every single continuous variable. In addition, there is a change in proportion profile across points which indicates that there is a non-linear relationship between the dependent variable and the continuous variables. This likely indicates that transformations are needed to the variable during model incorporation. Of the

continuous variables, IMP_HOME_VAL cannot be used raw. An indicator variable called I_HOMEOWN will be made with zero value if IMP_HOME_VAL is zero, and one if the variable reports a value. Result is that 28% of observations include the missing values that have been imputed are not home owners.

**Categorical Variables**

A bi-serial correlation will be performed the proportion profiles will be examined. From the observations, the dataset is pre-processed with variables with the z_* prefix. This made easier to consider the categorical variable such as MSTATUS as both married and unmarried status can be examined. Every single categorical variable has differences in proportion profiles except for URBANICITY. This variable will not be considered further based on this criterion. Considering MVR_PTS as a categorical variable may have been a mistake since the variable has 13 bins. Instead, MVR_PTS will be considered as a continuous variable.

Next step is to codify each categorical variable as a family of dummy variables. To avoid the dummy variable trap, variables of references are chosen with the smallest category in each variable. The only exception will be the 'Yes'/'No' variables and they will have dummy variables to be one value with Yes answer.

**Build Models (Logistics, Probability of Crash)**

A logistics regression model will be made with the default scoring method, which is the equivalent to fitting by iteratively reweighted least squares. This method results in a large grid of the Chi-Square score for each model, incrementing by the number of incorporated variables. Table 4 examines the best scoring for the two top models with a single variable through seven variables.

*Table 4 Logistics Regression by Fisher Score*

| N Variables | Chi-Square Score | Incorporated Variables |
|---|---|---|
| 1 | 360.5259 | MVR_PTS |
| 1 | 349.4074 | CLM_FREQ |
| 2 | 508.7027 | CLM_FREQ MVR_PTS |
| 2 | 504.413 | MVR_PTS REV_L |
| 3 | 641.5337 | CLM_FREQ MVR_PTS REV_L |
| 3 | 639.9598 | MVR_PTS USE_P REV_L |
| 4 | 775.7635 | IMP_INCOME MVR_PTS USE_P REV_L |

| N Variables | Chi-Square Score | Incorporated Variables |
|---|---|---|
| 4 | 763.1352 | CLM_FREQ IMP_INCOME USE_P REV_L |
| 5 | 888.9681 | CLM_FREQ IMP_INCOME MVR_PTS USE_P REV_L |
| 5 | 880.2277 | IMP_INCOME MVR_PTS USE_P MARRIED_Y REV_L |
| 6 | 983.9883 | CLM_FREQ IMP_INCOME MVR_PTS USE_P MARRIED_Y REV_L |
| 6 | 968.6079 | CLM_FREQ IMP_INCOME MVR_PTS I_HOMEOWN USE_P REV_L |
| 7 | 1044.1525 | CLM_FREQ IMP_INCOME MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L |
| 7 | 1028.586 | CLM_FREQ IMP_INCOME MVR_PTS I_HOMEOWN TYPE_MINI USE_P REV_L |

While using a single selection criteria, produced models are optimistic. The top performing models will be constructed with these number of variables. Four variables will be first due to the increase in Chi-Square score as more models are included. No more than seven variables will be considered as there will be a need to include the use of each TYPE_* for every vehicle in the data set. It makes the explanation less complicated.

IMP_INCOME is incorporated into each of the models. Its histogram showed a heavy left. While It passed the tests for normality, a log transformation was performed which resulted in decrease performance in ROC, concordant and discordant. Therefore, IMP_INCOME will be continuously used even with its skewed histogram.

**Select Model (Logistics, Probability of Crash)**

Details for three selected models will be listed and the results are interpreted.

$$Logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

*Table 5 Model with Four Variables*

| In Model | In Data | Label |
|---|---|---|
| $Y$ is | TARGET_FLAG | Crashes |
| $X_1$ is | IMP_INCOME | Imputed Income |

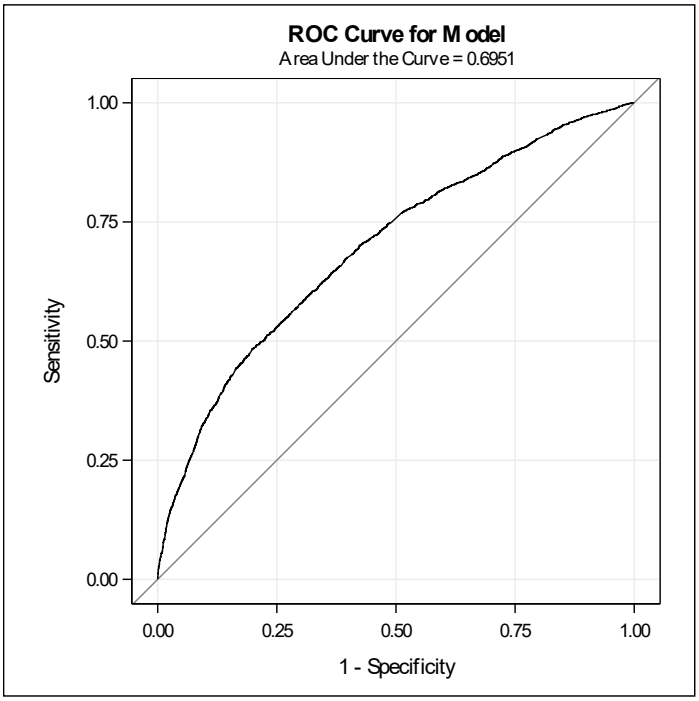
| | | |
|---|---|---|
| $X_2$ is | MVR_PTS | Motor Vehicle Record Points |
| $X_3$ is | USE_P | Vehicle Use (Personal/Commercial) |
| $X_4$ is | REV_L | Licensed Revoked |

*Table 6 Coefficient Values on Model with Four Variables*

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.6500 | 0.0624 | 108.6164 | <.0001 |
| **IMP_INCOME** | 1 | -8.2E-6 | 6.583E-7 | 155.0936 | <.0001 |
| **MVR_PTS** | 1 | 0.2013 | 0.0117 | 295.9227 | <.0001 |
| **USE_P** | 1 | -0.6941 | 0.0540 | 165.2281 | <.0001 |
| **REV_L** | 1 | 0.9122 | 0.0732 | 155.2338 | <.0001 |

The interpretation of the above model, all variables being held, is that: For a one unit increase in income there will be a 0.000065% decrease in the likelihood of an accident. For a one unit increase in the amount of motor vehicle record points there is a 0.0117% increase in the likelihood of a crash. If one is using a personal vehicle there is a 5.4% increase of an accident. Finally, if the driver has had a license revoked in the past, there is a 7.32% increase in the likelihood of a crash.

*Figure 2 ROC Curve for Model*



From examining ROC, the model performs at 0.6951 coverage. The curve above the diagonal line. While the IMP_INCOME variable is difficult in interpretability, scaling after parameter estimate is not acceptable and scaling before parameter estimate seems unreasonable, or illogical.

$$Logit\ (Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

*Table 7 Model with Five Variables*

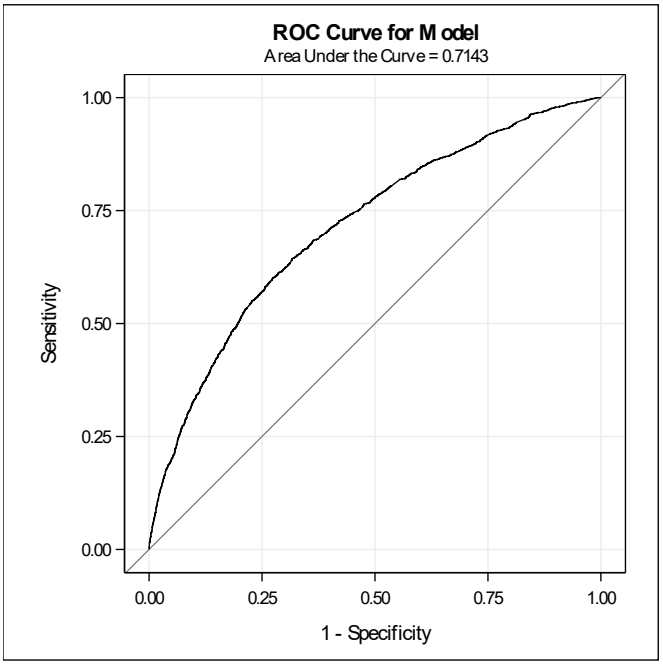
| In Model | | In Data | Label |
|---|---|---|---|
| *Y* | is | TARGET_FLAG | Crashes |
| $X_1$ | is | CLM_FREQ | #Claims (Past 5 Years) |
| $X_2$ | is | IMP_INCOME | Imputed Income |
| $X_3$ | is | MVR_PTS | Motor Vehicle Record Points |
| $X_4$ | is | USE_P | Vehicle Use (Personal/Commercial) |
| $X_5$ | is | REV_L | Licensed Revoked |

*Table 8 Coefficient Values on Model with Five Variables*

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.8074 | 0.0646 | 156.3133 | <.0001 |
| **CLM_FREQ** | 1 | 0.2709 | 0.0233 | 135.2113 | <.0001 |
| **IMP_INCOME** | 1 | -8.19E-6 | 6.654E-7 | 151.5979 | <.0001 |
| **MVR_PTS** | 1 | 0.1463 | 0.0126 | 135.3884 | <.0001 |
| **USE_P** | 1 | -0.6694 | 0.0545 | 150.7663 | <.0001 |
| **REV_L** | 1 | 0.8967 | 0.0739 | 147.2318 | <.0001 |

For one unit increase in the claim frequency (another claim within the last five-year period) the likelihood of a crash increases by 2.33%. For a unit increase in income the likelihood of a crash increases by 0.000665%. For one unit increase in motor vehicle record points the likelihood of a crash increases by 1.26%. Use of a personal vehicle results in 5.45% more likelihood of a crash as opposed to a commercial vehicle. Finally, if the driver has their licensed revoked in the past there is 7.39% increase in likelihood of a crash.

*Figure 3 ROC Curve for Model #2*




Five-variable model increased area under the curve from 0.6951 to 0.7143 with the inclusion of CLM_FREQ into the model.

$$Logit\ (Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + e$$

*Table 9 Model with Six Variables*

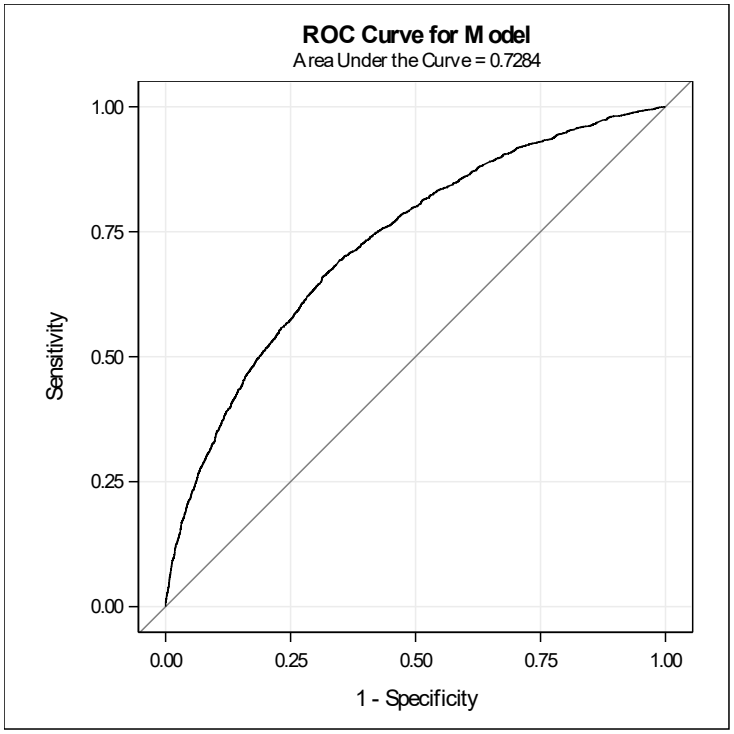
| In Model | In Data | Label |
|---|---|---|
| $Y$ is | TARGET_FLAG | Crashes |
| $X_1$ is | CLM_FREQ | #Claims(Past 5 Years) |
| $X_2$ is | IMP_INCOME | Imputed Income |
| $X_3$ is | MVR_PTS | Motor Vehicle Record Points |
| $X_4$ is | USE_P | Vehicle Use (Personal/Commercial) |
| $X_5$ is | MARRIED_Y | Marital Status (Yes Married) |
| $X_6$ is | REV_L | Licensed Revoked |

*Table 10 Coefficient Values on Model with Six Variables*

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.4333 | 0.0727 | 35.5139 | <.0001 |
| **CLM_FREQ** | 1 | 0.2631 | 0.0235 | 124.9811 | <.0001 |
| **IMP_INCOME** | 1 | -8.56E-6 | 6.705E-7 | 162.8418 | <.0001 |
| **MVR_PTS** | 1 | 0.1449 | 0.0127 | 130.5962 | <.0001 |
| **USE_P** | 1 | -0.6733 | 0.0550 | 149.9436 | <.0001 |
| **MARRIED_Y** | 1 | -0.5995 | 0.0543 | 121.8612 | <.0001 |
| **REV_L** | 1 | 0.8764 | 0.0746 | 138.1145 | <.0001 |

For one unit increase in the claim frequency (another claim within the last five-year period) the likelihood of a crash increases by 2.35%. For one unit increase in income the likelihood of a crash decreases by 0.000067%. For one unit increase in motor vehicle record points the likelihood of a crash increases by 1.27%. Use of a personal vehicle results in a 5.5% more likelihood of a crash as opposed to a commercial vehicle. Being married results in a 5.43% more likelihood of a crash. Finally, if the driver has their licensed revoked in the past there is a 7.46% increase in likelihood of a crash.

*Figure 4 ROC Curve for Model #3*




This model, with six variables, provides the highest ROC coverage of 0.7284.

For each of these models, increases are made in the ROC coverage values with additions of independent variables. Each of the included independent variables are simple to explain in their interpretations. For the model deployment, six variable model will be used to predict the likelihood of a crash.

As an experiment, the six-variable model was incorporated with a variable with an imputed with the mean. An indicator variable for the observations was made. Thus, a seven-variable model was created to examine whether the ROC curve improved, but the result was not improved. Thus, the six-variable model is the best choice for reduced complexity during deployment and interpretation.

**Select Model #2 (Linear Regression, Estimation of Cost)**

For this step of analysis, a logistics regression will be utilized as the selection criteria be based on the Adjusted R-Square metric. The models will be culled down to smaller numbers of parameters to support interpretability. The selection of a model is;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3$$

*Table 11 Three Variable model for TARGET_AMT*

| In Model | In Data | Label |
|---|---|---|
| $Y$ is | TARGET_AMT | Cost after Crash |
| $X_1$ is | HOMEKIDS | Value of Vehicle |
| $X_2$ is | IMP_CAR_AGE | Vehicle Age |
| $X_3$ is | CLM_FREQ | #Claims (Past 5 Years) |

*Table 12 Three Variable Model Coefficient Values*

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | Intercept | 1 | 1324.64489 | 109.49862 | 12.10 | <.0001 |
| **HOMEKIDS** | #Children @Home | 1 | 216.66500 | 46.73093 | 4.64 | <.0001 |
| **IMP_CAR_AGE** | | 1 | -41.75664 | 9.44730 | -4.42 | <.0001 |
| **CLM_FREQ** | #Claims(Past 5 Years) | 1 | 464.81029 | 44.54290 | 10.44 | <.0001 |

If all variables are held equal, the intercept provides us with an indication of the target amount crash cost. For every one unit increase in the number of kids at home, the cost increases by approximately $217 on average. For an increase of age of the vehicle is that cost of crash decreases by approximately $42 on average. For every one unit increase in the number of claims filed over the last five years, the cost increases approximately $465 on average.

However, the model is far from the best model that can be found with the variables within this data set.

*Table 13 Thee Variable Model Goodness-of-Fit*

| Source | |
|---|---|
| Root MSE | 4659.18377 |
| R-Square | 0.0193 |
| Adj R-Square | 0.0190 |
| F Value | 53.61 |

Several models using automatic variable selection were made. The models that scored highest with goodness-of-fit selection criteria incorporated more than 15 variables, including dummy variables. Despite the poor performance, a simpler model for implementation is chosen instead.

**Conclusion**

The construction of logistics models requires a lot of preparatory data manipulation. Within the dataset, given variables have been prepared which assisted in analysis. However, there are a lot of iterative work to examine variables that would be acceptable to take forward into the model construction and selection. In the case of using variation in ratio, almost every single variable was qualified to be taken forward. As this modeling methodology was being utilized, it will be interesting to examine how early observation of variation in ratio will be indicative of need for incorporation into the model, and likely more importantly how the variable needs to be transformed.

All exhaustive transformation of variables was not explored and incorporated into the model and it likely will impact the overall performance of the models constructed. Some of the transformed variables early on due to examination of their nature (via histograms) did not perform any better after transformation.

Interestingly, few of the dummy variable families was selected during the automated variable selection techniques. In this analysis, when these variables were not considered to immediate model incorporations, it was viewed as acceptable due to their increase of the interpretation. The group membership dummy variables seemed to be more popular during selection, as well as provided a simpler form for interpretation.

The logistics model is more naturally interpretable than previous models constructed with the OLS method.