Insurance Logistic Regression

FY 2016

Author Eric Lewis

Contents

[**Bonus** 1](#_Toc450498448)

[**Introduction** 5](#_Toc450498449)

[**Exploratory Data Analysis** 6](#_Toc450498450)

[**Data Exploration** 9](#_Toc450498451)

[**Data Preparation** 18](#_Toc450498452)

[**Building Models** 19](#_Toc450498453)

[**Conclusion** 28](#_Toc450498454)

[**Appendix** 29](#_Toc450498455)

# **Bonus**

**Scored File as SAS Data Set**

Filename: insurance\_score\_04. sas7bdat

SAS Code on page 30

**PROC GENMOD**

The Kaggle score using PROC LOGISTIC is 0.79259 and the PRC GENMOD scored 0.78296 lower which is not highly significant; however, the GENMOD procedure does not provides additional insight into the variables. The key takeaway from using PROC GENMOD for this model does not appear to provide significant insight into the variables and it does not score better using this model. One key observation is that the AIC is lower using PROC GENMOD is 8503.58 and the PROC LOGISTIC is 9419.96, see table 1 and 2.

SAS Code on page 31

**PROC GENMOD**

* AIC is 8503.5830
* AICC is 8503.6763
* BIC is 8636.7183
* Kaggle is **0.78296**

**PROBIT Model**

One of the advantages is using the PROBIT model is that provides Association of Predicted Probabilities and Observed Responses, see table 3. This provides additional insight into the relationship between the predicted probabilities of this model, and the actual outcomes of the data.

SAS Code on page 45

**SAS Macro Use**

One of the advantages is using SAS macros is to create clear references towards file and data usage at the beginning of the SAS code, for example &INFILE., &TEMPFILE., and &SCRUBFILE. used in this model.

SAS Code on page 36

**Stand Alone Scoring Program for P\_Target\_AMT**

The analysis and scoring of the variable P\_Target\_AMT is the prediction of the insurance damage assuming the insured does get into a collision. A solid Kaggle score is 5424.78693, which is slightly greater than the decision tree model of 5386.32171, and lower than the baseline model of 5552.16599.

SAS Analysis Code on page 46

SAS Score Code on page 47

|  |  |
| --- | --- |
| The GLM Procedure  Model: MODEL 1  Dependent Variable: TARGET\_FLAG | |
| Number of Observations Read | 8161 |
| Number of Observations Used | 8161 |

|  |  |  |  |
| --- | --- | --- | --- |
| Criteria for Assessing Goodness of Fit | | | |
| Deviance | 8143 | 1348.2553 | 0.1656 |
| Scaled Deviance | 8143 | 8161.0000 | 1.0022 |
| Pearson Chi-Square | 8143 | 1348.2553 | 0.1656 |
| Scaled Pearson X2 | 8143 | 8161.0000 | 1.0022 |
| Log Likelihood |  | -4232.7915 |  |
| Full Log Likelihood |  | -4232.7915 |  |
| AIC (smaller is better) |  | 8503.5830 |  |
| AICC (smaller is better) |  | 8503.6763 |  |
| BIC (smaller is better) |  | 8636.7183 |  |

Table 1: PROC GENMOD

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Analysis of Maximum Likelihood Parameter Estimates | | | | | | | |
| Parameter | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept |  | 1 | 0.3687 | 0.0191 | 0.3312 | 0.4062 | 372.04 |
| KIDSDRIV |  | 1 | 0.0708 | 0.0089 | 0.0534 | 0.0881 | 63.81 |
| TRAVTIME |  | 1 | 0.0009 | 0.0003 | 0.0003 | 0.0015 | 10.08 |
| TIF |  | 1 | -0.0074 | 0.0011 | -0.0095 | -0.0053 | 46.31 |
| CLM\_FREQ |  | 1 | 0.0486 | 0.0043 | 0.0403 | 0.0570 | 130.55 |
| MVR\_PTS |  | 1 | 0.0266 | 0.0023 | 0.0221 | 0.0311 | 134.32 |
| IMP\_HOME\_VAL |  | 1 | -0.0000 | 0.0000 | -0.0000 | -0.0000 | 11.26 |
| IMP\_INCOME |  | 1 | -0.0000 | 0.0000 | -0.0000 | -0.0000 | 17.18 |
| USE\_P |  | 1 | -0.0973 | 0.0112 | -0.1192 | -0.0753 | 75.51 |
| MARRIED\_Y |  | 1 | -0.0876 | 0.0111 | -0.1094 | -0.0658 | 62.14 |
| REV\_L |  | 1 | 0.1638 | 0.0138 | 0.1367 | 0.1908 | 140.92 |
| IMP\_JOB | Clerical | 1 | -0.0059 | 0.0162 | -0.0376 | 0.0259 | 0.13 |
| IMP\_JOB | Doctor | 1 | -0.0501 | 0.0239 | -0.0969 | -0.0033 | 4.40 |
| IMP\_JOB | Home Maker | 1 | -0.0196 | 0.0207 | -0.0602 | 0.0210 | 0.89 |
| IMP\_JOB | Lawyer | 1 | -0.0456 | 0.0187 | -0.0824 | -0.0089 | 5.94 |
| IMP\_JOB | Manager | 1 | -0.1020 | 0.0175 | -0.1364 | -0.0677 | 33.89 |
| IMP\_JOB | Professional | 1 | -0.0463 | 0.0165 | -0.0787 | -0.0139 | 7.83 |
| IMP\_JOB | Student | 1 | -0.0079 | 0.0195 | -0.0461 | 0.0303 | 0.16 |
| Scale |  | 1 | 0.4065 | 0.0032 | 0.4003 | 0.4127 |  |

Table 2: PROC GENMOD Parameter Estimates

|  |  |  |  |
| --- | --- | --- | --- |
| Association of Predicted Probabilities and Observed Responses | | | |
| Percent Concordant | 74.7 | Somers' D | 0.495 |
| Percent Discordant | 25.3 | Gamma | 0.495 |
| Percent Tied | 0.0 | Tau-a | 0.192 |
| Pairs | 12935224 | c | 0.747 |

Table 3: PROBIT Model

# **Introduction**

The objective of this data analysis is to build a model to predict the probability that an auto insurance customer will get into a collision. Four models will be compared based upon the criteria of AIC, SC, and the area under the ROC curve. These are measures utilized in logistic regression to provide empirical for model comparison. The final determination will be determined by the highest probability scored using Kaggle, with 100% as the highest score possible though not probably. A final model will be selected and a short and long term recommendation will be delivered from this analysis.

# **Exploratory Data Analysis**

There are two main components in developing this predictive model:

* **Training data set** – utilized for exploratory data analysis, data preparation, building and selecting a predictive model. This data set contains 8,100 observations with the variables as shown in table 1 below.
* **Test data set** – utilized to score the model selected in the training phase of this analysis. The model results are being scored using Kaggle. This data set contains 2141 observations less the variable Each line item in the data set contains the specific data on the insured.

This analysis will determine which data elements are the highest correlated towards determining the probability of collision.

The following table provides the variable name, type, and definition as the initial step towards understanding the data.

|  |  |  |
| --- | --- | --- |
| VARIABLE NAME | TYPE | DEFINITION |
| INDEX |  | Identification Variable (do not use) |
| TARGET\_FLAG |  | Was Car in a crash? 1=YES 0=NO |
| TARGET\_AMT |  | If car was in a crash, what was the cost |
|  |  |  |
| 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 |

Table 4: Data Dictionary

The following table provides the variable name, and theoretical effect as an additional step towards understanding the data and how it relates towards building the predictive model.

|  |  |
| --- | --- |
| VARIABLE NAME | THEORETICAL EFFECT |
| INDEX | None |
| TARGET\_FLAG | None |
| TARGET\_AMT | None |
|  |  |
| AGE | Very young people tend to be risky. Maybe very old people also. |
| BLUEBOOK | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR\_AGE | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR\_TYPE | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR\_USE | Commercial vehicles are driven more, so might increase probability of collision |
| CLM\_FREQ | The more claims you filed in the past, the more you are likely to file in the future |
| EDUCATION | Unknown effect, but in theory more educated people tend to drive more safely |
| HOMEKIDS | Unknown effect |
| HOME\_VAL | In theory, home owners tend to drive more responsibly |
| INCOME | In theory, rich people tend to get into fewer crashes |
| JOB | In theory, white collar jobs tend to be safer |
| KIDSDRIV | When teenagers drive your car, you are more likely to get into crashes |
| MSTATUS | In theory, married people drive more safely |
| MVR\_PTS | If you get lots of traffic tickets, you tend to get into more crashes |
| OLDCLAIM | If your total payout over the past five years was high, this suggests future payouts will be high |
| PARENT1 | Unknown effect |
| RED\_CAR | Urban legend says that red cars (especially red sports cars) are riskier. Is that true? |
| REVOKED | If your license was revoked in the past 7 years, you probably are a riskier driver. |
| SEX | Urban legend says that women have less crashes then men. Is that true? |
| TIF | People who have been customers for a long time are usually more safe. |
| TRAVTIME | Long drives to work usually suggest greater risk |

Table 5: Data Dictionary {THEORETICAL EFFECT}

# **Data Exploration**

**Missing Data**

The training data set contains the following statistical variables with missing data. Depending on the next step, which is correlating the statistical variables with the probability of collision. The key observations from table 6 are the quantity of missing values per statistical variable along with the mean and standard deviation if we choose to impute the missing data elements.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Label | N | N Missing | Mean | Std Dev |
| TARGET\_FLAG  KIDSDRIV  AGE  HOMEKIDS  YOJ  INCOME  HOME\_VAL  TRAVTIME  BLUEBOOK  TIF  OLDCLAIM  CLM\_FREQ  MVR\_PTS  CAR\_AGE | #Driving Children  Age  #Children @Home  Years on Job  Income  Home Value  Distance to Work  Value of Vehicle  Time in Force  Total Claims (Past 5 Years)  #Claims (Past 5 Years)  Motor Vehicle Record Points  Vehicle Age | 8161  8161  8155  8161  7707  7716  7697  8161  8161  8161  8161  8161  8161  7651 | 0  0  6  0  454  445  464  0  0  0  0  0  0  510 | 0.2638157  0.1710575  44.7903127  0.7212351  10.4992864  61898.10  154867.29  33.4887972  15709.90  5.3513050  4037.08  0.7985541  1.6955030  8.3283231 | 0.4407276  0.5115341  8.6275895  1.1163233  4.0924742  47572.69  129123.78  15.9047470  8419.73  4.1466353  8777.14  1.1584527  2.1471117  5.7007424 |

Table 6: Missing & Mean

We will decide whether to impute or exclude the statistical variable from the predictive model depending on the results from correlating each statistic with the probability of collision.

* YOJ
* INCOME
* HOME\_VAL
* CAR\_AGE

**Variable Correlation to Target Flag**

Key observations from table 3 are that no variables are considered highly correlated to the probability of collision. This provides us with an early indication that the final model chosen will have to be scrutinized as to whether or not the predictability percentage of the model meets the business requirements for model usage. Motor Vehicle Record Points is the highest correlated statistic having approximately a 22% correlation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Label** | **Correlation** | **Target Flag** |
| TARGET\_FLAG | Probability of Collision |  |  |
| KIDSDRIV | #Driving Children | 0.10367 | increase |
| AGE | Age | -0.10322 | decrease |
| HOMEKIDS | #Children @Home | 0.11562 | increase |
| YOJ | Years on Job | -0.07051 | decrease |
| INCOME | Income | -0.14201 | decrease |
| HOME\_VAL | Home Value | -0.18374 | decrease |
| TRAVTIME | Distance to Work | 0.04815 | increase |
| BLUEBOOK | Value of Vehicle | -0.10338 | decrease |
| TIF | Time in Force | -0.08237 | decrease |
| OLDCLAIM | Total Claims (Past 5 Years) | 0.13808 | increase |
| CLM\_FREQ | #Claims (Past 5 Years) | 0.21620 | increase |
| MVR\_PTS | Motor Vehicle Record Points | 0.21920 | increase |
| CAR\_AGE | Vehicle Age | -0.10065 | decrease |

Table 7: Correlation with Target Flag

**Visual Representation of Variables**

The purpose of the visual or graphical representation of the distribution within the variables is to provide observations toward the predictive model variable selection to complement the correlation with target flag as shown in table 7 above. The histogram of the highest correlated variable Motor Vehicle Record Points is skewed-right indicating that more of the insured don’t have any points against their record or depending upon their location; the state may not have a point system, such as in Illinois. If the insured’s geographic location was available this variable could be cross-referenced with states that do have point systems in place.

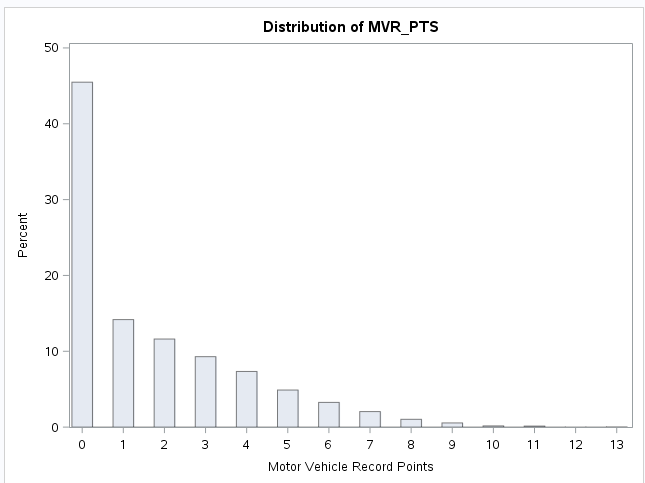


Figure 1: Distribution of MVR\_PTS

The histogram of age represents a normal distribution as one might expect from an auto insurance data set. While age only represents a 10% decrease in the probability of collision. Placing age into bins may prove itself very useful in developing a strong predictive model.

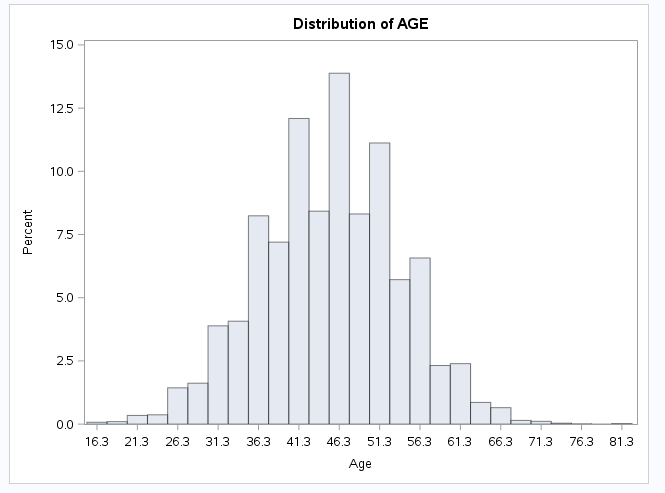


Figure 2: Distribution of AGE

The histogram of years on the job represents a normal distribution with outliers less than one-half year. Years on the job only represents a decrease of 7% in the probability of collision.

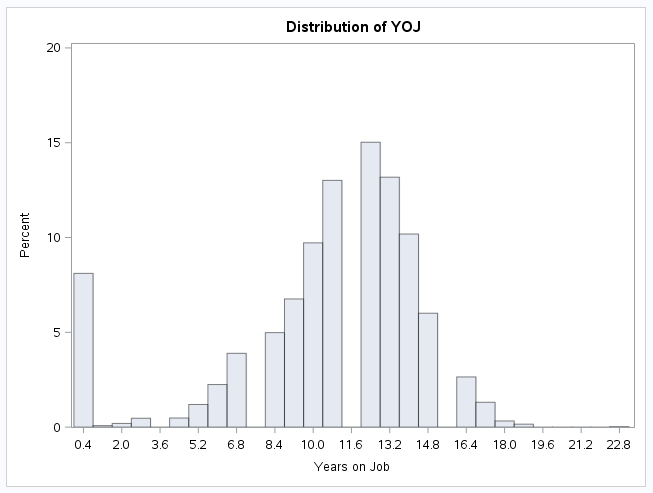


Figure 3: Distribution of YOJ

The histogram of income demonstrates a skewed-right distribution. Income only represents a decrease of 14% in the probability of collision. This would indicate that there are likely quit a few students who are working part-time as insured’s in this data set.

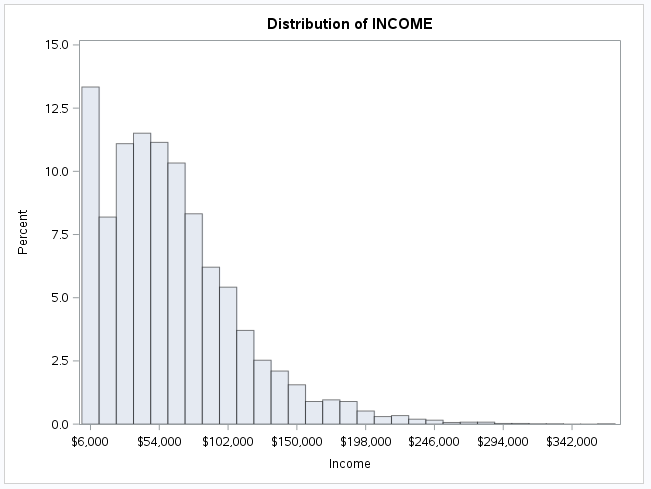


Figure 4: Distribution of INCOME

The histogram of home value demonstrates a skewed-right distribution. Home value represents a decrease of 19% in the probability of collision. There is a presence of a many outliers less than $15K, indicating that there are likely quit a few who are renting as insured’s in this data set.

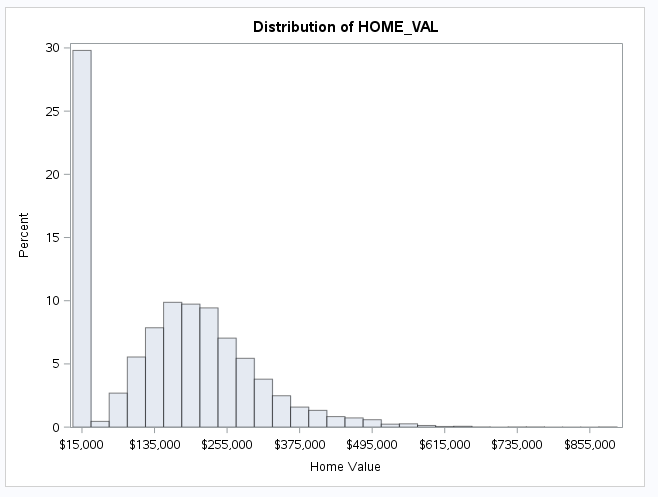


Figure 5: Distribution of HOME\_VAL

The histogram of home value demonstrates a slightly skewed-right to normal distribution. Travel time represents an increase of 5% in the probability of collision.

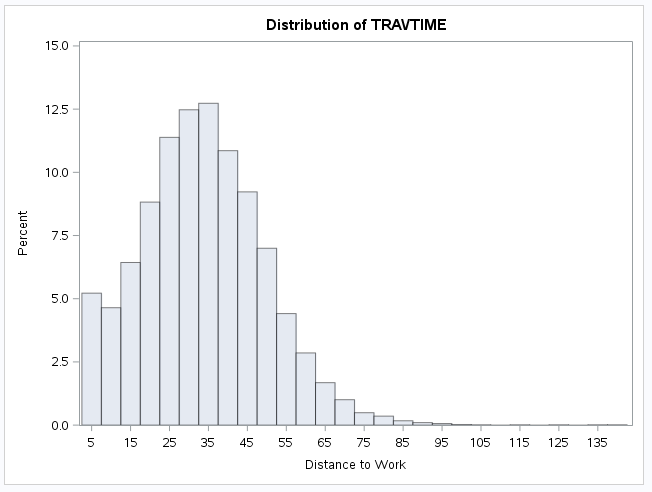


Figure 6: Distribution of TRAVTIME

# **Data Preparation**

The data preparation phase of this analysis encompasses preparing the data for modeling. Various techniques that will be reviewed are: imputing missing values, flagging missing variables, data transformation through combining variables and through the use of mathematical transformations.

**Missing Values**

The initial testing of the model will include imputing the following variable with missing values based upon their mean shown in table 2. The following variables are imputed using the mean in the case of missing data.

* YOJ – mean of 10.4992864
* INCOME – mean of 61898.10
* HOME\_VAL – mean of 154867.29
* CAR\_AGE – mean of 8.3283231

**Transforming Data into Buckets**

The following variables are transformed into buckets based partly upon the theoretical effect of decreasing the probability of collision and their analysis of maximum likelihood estimates.

* USE\_P = car use equal to private.
* MARRIED\_Y = married status equal to yes.
* REV\_L = revoked equal to yes.
* IMP\_INCOME = for missing values a doctor is equal to $100K, lawyer is equal to $80K, else Blue collar.

**Mathematical Data Transformations**

Attempts were made to transform variables mathematically for example Logarithm and square root data transformations were attempted; however, the predictive value of the model demonstrated no improvement.

**Combining Variables**

Attempts were made to combine variables to perform ratio analysis; however, the predictive value of the model demonstrated no improvement.

# **Building Models**

Five base models were utilized as comparison for this analysis using Logistic Regression. The primary basis for final model variable selection is based upon the lowest AIC and SC score, as well as the largest area under the ROC curve. Secondarily the variable selection process is based upon forward, backward, and stepwise variable selection where all three procedures yielded similar Chi Squared values when using all the variables in the data set and imputing the missing data with their perspective means. The initial stand-alone variable selection criteria are based upon a combination of the variables correlation to target flag in table 7: Correlation with Target Flag and the significance of <.0001 within the parameter estimates as show in table 9 below, summary forward variable selection.

**First Model**

This model is known as the base model. It is a model of all the numeric variables in the data set which is utilized as a baseline having an AIC of 9419.962, SC of 9426.96, area under the ROC curve of 0.7186, and a Kaggle score of 0.75741. All 11 numeric variables are selected in this model. The purpose of this model is to serve as a baseline model.

|  |  |  |
| --- | --- | --- |
| Model Fit Statistics | | |
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 9419.962 | 9053.599 |
| SC | 9426.969 | 9067.613 |
| -2 Log L | 9417.962 | 9049.599 |

Table 8: Model Fit Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Summary of Forward Selection | | | | | | |
| Step | **Effect Entered** | **DF** | **Number In** | **Score Chi-Square** | **Pr > ChiSq** | **Variable Label** |
| 1 | MVR\_PTS | 1 | 1 | 392.1144 | <.0001 | Motor Vehicle Record Points |
| 2 | IMP\_HOME\_VAL | 1 | 2 | 222.5610 | <.0001 | Home Value |
| 3 | CLM\_FREQ | 1 | 3 | 147.2044 | <.0001 | #Claims (Past 5 Years) |
| 4 | KIDSDRIV | 1 | 4 | 67.7076 | <.0001 | #Driving Children |
| 5 | TIF | 1 | 5 | 48.1064 | <.0001 | Time in Force |
| 6 | IMP\_CAR\_AGE | 1 | 6 | 34.9244 | <.0001 | Vehicle Age |
| 7 | BLUEBOOK | 1 | 7 | 23.3352 | <.0001 | Value of Vehicle |
| 8 | HOMEKIDS | 1 | 8 | 15.1269 | 0.0001 | #Children @Home |
| 9 | TRAVTIME | 1 | 9 | 15.0489 | 0.0001 | Distance to Work |
| 10 | IMP\_AGE | 1 | 10 | 5.7801 | 0.0162 | Age |
| 11 | OLDCLAIM | 1 | 11 | 3.9751 | 0.0462 | Total Claims (Past 5 Years) |

Table 9: Summary of Forward Variable Selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Maximum Likelihood Estimates | | | | | |
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -0.4866 | 0.1860 | 6.8409 | 0.0089 |
| KIDSDRIV | 1 | 0.2931 | 0.0549 | 28.5624 | <.0001 |
| HOMEKIDS | 1 | 0.0717 | 0.0294 | 5.9462 | 0.0147 |
| TRAVTIME | 1 | 0.00663 | 0.00167 | 15.8184 | <.0001 |
| BLUEBOOK | 1 | -0.00001 | 3.457E-6 | 18.5407 | <.0001 |
| TIF | 1 | -0.0469 | 0.00678 | 47.8964 | <.0001 |
| OLDCLAIM | 1 | 6.239E-6 | 3.133E-6 | 3.9668 | 0.0464 |
| CLM\_FREQ | 1 | 0.2621 | 0.0256 | 104.8606 | <.0001 |
| MVR\_PTS | 1 | 0.1397 | 0.0126 | 123.6236 | <.0001 |
| IMP\_AGE | 1 | -0.00854 | 0.00356 | 5.7367 | 0.0166 |
| IMP\_HOME\_VAL | 1 | -2.74E-6 | 2.419E-7 | 128.3662 | <.0001 |
| IMP\_CAR\_AGE | 1 | -0.0222 | 0.00512 | 18.7083 | <.0001 |

Table 10: Analysis of Maximum Likelihood Estimates

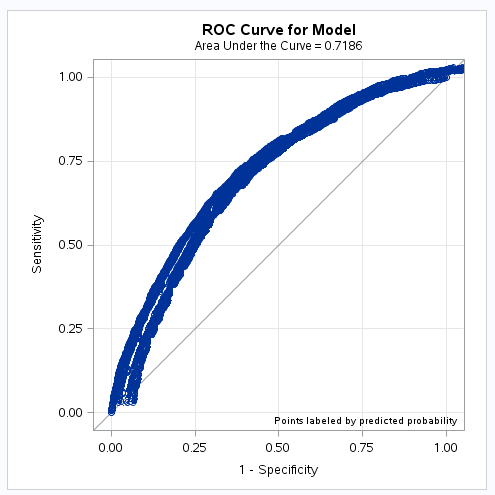


Figure 7: ROC Curve for Model 1, area {0.7186}

**Second Model**

This model is the base model using only the numeric variables having significance of <.0001 in the analysis of maximum likelihood estimates. It is a model six of the numeric variables in the data set having an AIC of 9419.962, SC of 9426.96, area under the ROC curve of 0.7092, and a Kaggle score of 0.75741. This model will serve as a building block model to incrementally include categorical variables for performance improvement.

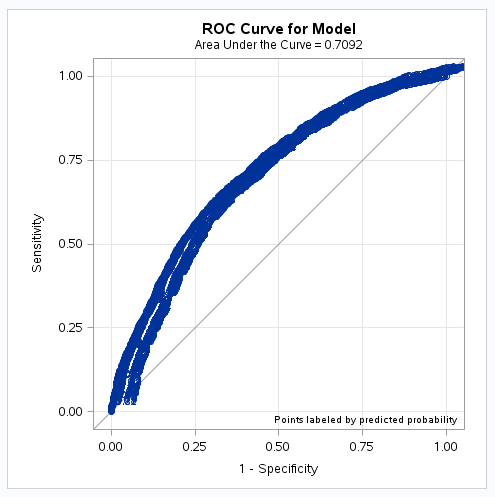


Figure 8: ROC Curve for Model 2, area {0.7902}

**Third Model**

This model is built based off the second model including categorical variables of car use, marital status, and revoked. It is a model of six variables in the data set having an AIC of 9419.962, SC of 9426.96, area under the ROC curve of 0.7284, and a Kaggle score of 0.79257. Through the addition of categorical variables to this model, the results have significantly improved over the first and second models. This model will continually be built upon incrementally adding categorical variables for performance improvement.

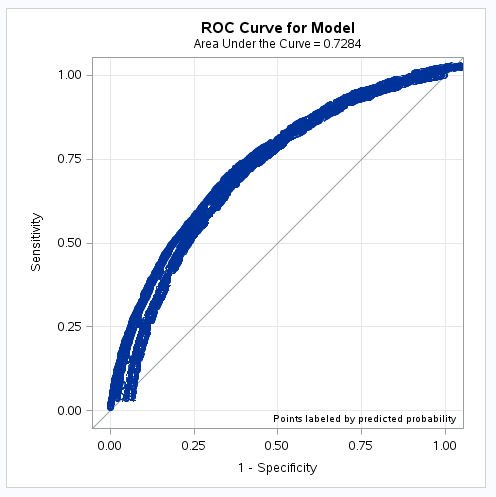


Figure 9: ROC Curve for Model 3, area {0.7284}

**Selected Final Model**

This model is built based off the third model including categorical variables as kids driving, travel time,

time in force, and imputed home value. It is a model of ten variables in the data set having an AIC of 9419.962, SC of 9426.969, area under the ROC curve of 0.7474, and a Kaggle score of 0.79257. Through the addition of categorical variables to this model, the results have significantly improved over the first thru third models. An important observation is the results from the analysis of maximum likelihood estimates for the variable imputed job. I decided to utilize doctor, lawyer, manager, and professional based upon the estimate, the percentage of the decreased probability of collision. The most significant is the position of manager, which has a 39% decrease in the probability of collision. The next step for further model improvement is to use decision tree analysis for numeric variable imputation and further categorical variable categorization. Due to time constraints for first model delivery, decision tree analysis will be included in phase two of this analysis.

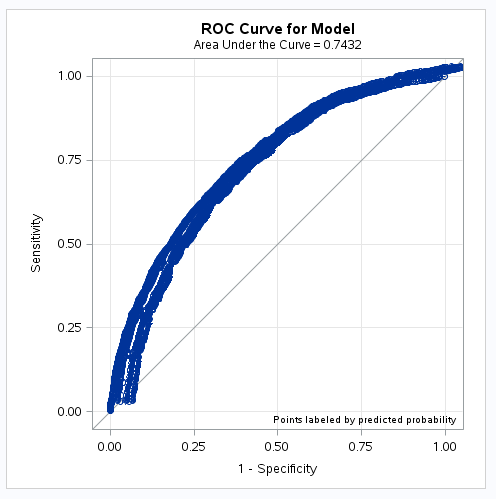


Figure 10: ROC Curve for Model 3, area {0.7432}

|  |  |  |
| --- | --- | --- |
| Model Fit Statistics | | |
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 9419.962 | 8180.159 |
| SC | 9426.969 | 8306.287 |
| -2 Log L | 9417.962 | 8144.159 |

Table 11: Model Fit Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| **Parameter** | **Label** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept |  | 1 | -0.2774 | 0.0673 | 17.0088 | <.0001 |
| KIDSDRIV |  | 1 | 0.2298 | 0.0294 | 61.2920 | <.0001 |
| TRAVTIME |  | 1 | 0.00361 | 0.000997 | 13.1403 | 0.0003 |
| TIF |  | 1 | -0.0280 | 0.00399 | 49.4915 | <.0001 |
| CLM\_FREQ |  | 1 | 0.1615 | 0.0142 | 129.0436 | <.0001 |
| MVR\_PTS |  | 1 | 0.0802 | 0.00765 | 109.8853 | <.0001 |
| IMP\_HOME\_VAL |  | 1 | -7.4E-7 | 1.88E-7 | 15.4939 | <.0001 |
| IMP\_INCOME |  | 1 | -2.79E-6 | 5.923E-7 | 22.2012 | <.0001 |
| USE\_P |  | 1 | -0.3402 | 0.0390 | 76.1980 | <.0001 |
| MARRIED\_Y |  | 1 | -0.2938 | 0.0387 | 57.7228 | <.0001 |
| REV\_L |  | 1 | 0.5026 | 0.0454 | 122.4587 | <.0001 |
| IMP\_JOB | Clerical | 1 | -0.0159 | 0.0552 | 0.0823 | 0.7742 |
| IMP\_JOB | Doctor | 1 | -0.1528 | 0.0893 | 2.9310 | 0.0869 |
| IMP\_JOB | Home Maker | 1 | -0.0742 | 0.0717 | 1.0705 | 0.3008 |
| IMP\_JOB | Lawyer | 1 | -0.1452 | 0.0677 | 4.5988 | 0.0320 |
| IMP\_JOB | Manager | 1 | -0.3902 | 0.0658 | 35.2183 | <.0001 |
| IMP\_JOB | Professional | 1 | -0.1459 | 0.0582 | 6.2932 | 0.0121 |
| IMP\_JOB | Student | 1 | -0.0796 | 0.0661 | 1.4498 | 0.2286 |

Table 12: Analysis of Maximum Likelihood Estimates

# **Conclusion**

This analysis is a comparison four identified models, including five scoring attempts utilized for model testing, and optimization. The models were compared based upon AIC, SC, area under the ROC curve, variable correlation to target flag, automated variable selection techniques, goodness-of-fit statistics, testing for multicollinearity, and the closest to 100% standalone scoring using Kaggle. The selected final model had an AIC of 9419.962, SC of 9426.969, and a Kaggle score of 0.79257. The Kaggle score is considered an excellent model score based upon the benchmark decision tree model score of 0.79553 and benchmark worst model score of 0.50000. The observations indicate the scoring of the eight chosen and imputed variables explains 79% of predictive accuracy towards an insured’s probability of collision.

**Near Term Recommendation**

We should begin assessing our insured’s collision risk based upon using the following 11 variables:

|  |
| --- |
| Key Insurance Variables |
| KIDSDRIV |
| TRAVTIME |
| TIF |
| CLM\_FREQ |
| MVR\_PTS |
| IMP\_HOME\_VAL |
| IMP\_INCOME |
| USE\_P |
| MARRIED\_Y |
| REV\_L |
| IMP\_JOB |

**Long Term Recommendation**

There is room for improvement upon this predictive model, with the long term goal to reach a model Kaggle score of greater than 0.85666, thus improving the predictive accuracy of the model. The methodology utilized to build a long term model will conduct variable selection based upon decision tree analysis using either R, Angoss, or SAS Enterprise Miner. 1

# **Appendix**

1. Decision Trees for Decision Making, HBR,

<https://hbr.org/1964/07/decision-trees-for-decision-making>

1. Decision Tree, Wikipedia,

<https://en.wikipedia.org/wiki/Decision_tree>

**SAS Utilized for Output of Scored File as SAS Data Set**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Unit 02: Insurance {Export .sas7bdat} \*;

\* Eric Lewis \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

proc import datafile='insurance\_score\_04.csv'

dbms=csv

out=scored

replace;

run;

\*proc print data=scored;

data 'insurance\_score\_04';

set scored;

run;

quit;

**SAS Utilized for PROC GLM & PROC GENMOD**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Unit 02: INSURANCE LOGISTIC REGRESSION PROJECT {Score} \*;

\* Eric Lewis Section 55 Spring 2016 \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

%let PATH = /folders;

%let NAME = INS;

%let LIB = &NAME..;

libname &NAME. "&PATH.";

%let INFILE = &LIB.LOGIT\_INSURANCE\_TEST;

%let TEMPFILE = TEMPFILE;

data &TEMPFILE.;

set &INFILE.;

\*libname score\_me '/folders';

\*data testing;

\* set score\_me.logit\_insurance\_test;

data validate;

set &TEMPFILE.;

IMP\_HOME\_VAL = HOME\_VAL;

I\_IMP\_HOME\_VAL = 0;

label IMP\_HOME\_VAL = 'Home Value';

label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

if missing(IMP\_HOME\_VAL) then do;

IMP\_HOME\_VAL = 154867.29;

I\_IMP\_HOME\_VAL = 1;

end;

IMP\_CAR\_AGE = CAR\_AGE;

I\_IMP\_CAR\_AGE = 0;

if missing(IMP\_CAR\_AGE) then do;

IMP\_CAR\_AGE = 8.3283231;

I\_IMP\_CAR\_AGE = 1;

end;

IMP\_INCOME = INCOME;

I\_IMP\_INCOME = 0;

if missing(IMP\_INCOME) then do;

IMP\_INCOME = 61898.10;

I\_IMP\_INCOME = 1;

end;

IMP\_JOB = JOB;

if missing(IMP\_JOB) then do;

if IMP\_INCOME > 100000 then

IMP\_JOB = "Doctor";

else if IMP\_INCOME > 80000 then

IMP\_JOB = "Lawyer";

else

IMP\_JOB = "z\_Blue Collar";

end;

if CAR\_USE in ('Commercial' 'Private') then do;

USE\_P = (car\_use eq 'Private');

end;

if MSTATUS in ('Yes' 'z\_No') then do;

MARRIED\_Y = (MSTATUS eq 'Yes');

end;

if REVOKED in ('No' 'Yes') then do;

REV\_L = (REVOKED eq 'Yes');

end;

Drop HOME\_VAL;

Drop CAR\_AGE;

Drop INCOME;

Drop JOB;

data score;

set validate;

YHAT =

0.0708 \* KIDSDRIV

+ 0.0009 \* TRAVTIME

- 0.0074 \* TIF

+ 0.0486 \* CLM\_FREQ

+ 0.0266 \* MVR\_PTS

- 0.0000 \* IMP\_HOME\_VAL

- 0.0000 \* IMP\_INCOME

- 0.0973 \* USE\_P

- 0.0876 \* MARRIED\_Y

+ 0.1638 \* REV\_L

- 0.0501 \* (IMP\_JOB in ("Doctor"))

- 0.0456 \* (IMP\_JOB in ("Lawyer"))

- 0.1020 \* (IMP\_JOB in ("Manager"))

- 0.0463 \* (IMP\_JOB in ("Professional"))

+ 0.3687;

P\_TARGET\_FLAG = exp(YHAT) / (1+exp(YHAT));

keep index P\_TARGET\_FLAG;

proc print data=score;

proc export data=score

outfile='/folders/insurance\_score\_GENMOD\_05.csv'

dbms=csv

replace;

run;

**SAS Utilized for Scoring**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Model Four {7.8537} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

libname four11 '/folders';

data testing;

set four11.logit\_insurance\_test;

data testing\_fixed;

set testing;

IMP\_HOME\_VAL = HOME\_VAL;

I\_IMP\_HOME\_VAL = 0;

label IMP\_HOME\_VAL = 'Home Value';

label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

if missing(IMP\_HOME\_VAL) then do;

IMP\_HOME\_VAL = 154867.29;

I\_IMP\_HOME\_VAL = 1;

end;

IMP\_CAR\_AGE = CAR\_AGE;

I\_IMP\_CAR\_AGE = 0;

if missing(IMP\_CAR\_AGE) then do;

IMP\_CAR\_AGE = 8.3283231;

I\_IMP\_CAR\_AGE = 1;

end;

IMP\_INCOME = INCOME;

I\_IMP\_INCOME = 0;

if missing(IMP\_INCOME) then do;

IMP\_INCOME = 61898.10;

I\_IMP\_INCOME = 1;

end;

if CAR\_USE in ('Commercial' 'Private') then do;

USE\_P = (car\_use eq 'Private');

end;

if MSTATUS in ('Yes' 'z\_No') then do;

MARRIED\_Y = (MSTATUS eq 'Yes');

end;

if REVOKED in ('No' 'Yes') then do;

REV\_L = (REVOKED eq 'Yes');

end;

data testing\_score;

set testing\_fixed;

wat =

0.3886 \* KIDSDRIV

+ 0.00672 \* TRAVTIME

- 0.0473 \* TIF

+ 0.2643 \* CLM\_FREQ

+ 0.1383 \* MVR\_PTS

- 0.00000135 \* IMP\_HOME\_VAL

- 0.00000661 \* IMP\_INCOME

- 0.6846 \* USE\_P

- 0.4871 \* MARRIED\_Y

+ 0.8497 \* REV\_L

- 0.4588;

P\_TARGET\_FLAG = exp(wat) / (1+exp(wat));

keep index P\_TARGET\_FLAG;

proc print data=testing\_score;

proc export data=testing\_score

outfile='/folders/insurance\_score\_02.csv'

dbms=csv

replace;

run;

**SAS Utilized for Analysis**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Unit 02: INSURANCE LOGISTIC REGRESSION PROJECT {Analysis} \*;

\* Eric Lewis Section 55 Spring 2016 \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

%let PATH = /folders;

%let NAME = INS;

%let LIB = &NAME..;

libname &NAME. "&PATH.";

%let INFILE = &LIB.LOGIT\_INSURANCE;

%let TEMPFILE = TEMPFILE;

%let SCRUBFILE = SCRUBFILE;

\*proc print data=&INFILE.(obs=5);

\*run;

data &TEMPFILE.;

set &INFILE.;

drop INDEX;

drop TARGET\_AMT;

run;

\*proc print data=&TEMPFILE.(obs=5);

\*run;

\*proc contents data=&TEMPFILE.;

\*run;

data &SCRUBFILE.;

set &TEMPFILE.;

\*proc print data=&SCRUBFILE.(obs=5);

\*run;

\*proc contents data=&SCRUBFILE.;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Find means, missing data \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*proc means data=&TEMPFILE. n nmiss mean std;

\*var \_numeric\_ ;

\*run;

\*if missing(YOJ) then YOJ = 10.4992864;

\*if missing(INCOME) then INCOME = 61898.10;

\*if missing(HOME\_VAL) then HOME\_VAL = 154867.29;

\*if missing(CAR\_AGE) then CAR\_AGE = 8.3283231;

\*proc corr data=&TEMPFILE. rank plots=all;

\* var KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME\_VAL TRAVTIME BLUEBOOK TIF OLDCLAIM CLM\_FREQ MVR\_PTS CAR\_AGE;

\* with TARGET\_FLAG;

\*run;

\*proc freq data=&TEMPFILE.;

\*table \_character\_ /missing;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Data Exploration: Visual Analysis \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*if missing(YOJ) then YOJ = 10.4992864;

\*if missing(INCOME) then INCOME = 61898.10;

\*if missing(HOME\_VAL) then HOME\_VAL = 154867.29;

\*if missing(CAR\_AGE) then CAR\_AGE = 8.3283231;

\* proc univariate data=&TEMPFILE. normal;

\* var KIDSDRIV;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var AGE;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var HOMEKIDS;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var YOJ;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var INCOME;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var HOME\_VAL;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var TRAVTIME;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var BLUEBOOK;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var TIF;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var OLDCLAIM;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var CLM\_FREQ;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var MVR\_PTS;

\* histogram;

\* proc univariate data=&TEMPFILE. normal;

\* var CAR\_AGE;

\* histogram;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Data Preparation: Variable Selection \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*if missing(YOJ) then YOJ = 10.4992864;

\*if missing(INCOME) then INCOME = 61898.10;

\*if missing(HOME\_VAL) then HOME\_VAL = 154867.29;

\*if missing(CAR\_AGE) then CAR\_AGE = 8.3283231;

\*proc reg data=&TEMPFILE.;

\*model TARGET\_FLAG = KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME\_VAL TRAVTIME BLUEBOOK TIF OLDCLAIM CLM\_FREQ MVR\_PTS CAR\_AGE;

\*/selection=forward;

\*/selection=backward;

\*/selection=stepwise;

\*run;

\*quit;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Impute missing data w/means \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* IMP\_AGE = AGE;

\* I\_IMP\_AGE = 0;

\* label IMP\_AGE = 'Age';

\* label I\_IMP\_AGE = 'Age Imp Flag';

\* if missing(IMP\_AGE) then do;

\* IMP\_AGE = 44.7903127;

\* I\_IMP\_AGE = 1;

\* end;

\* IMP\_CAR\_AGE = CAR\_AGE;

\* I\_IMP\_CAR\_AGE = 0;

\* label IMP\_CAR\_AGE = 'Vehicle Age';

\* label I\_IMP\_CAR\_AGE = 'Vehicle Age Imp Flag';

\* if missing(IMP\_CAR\_AGE) then do;

\* IMP\_CAR\_AGE = 8.3283231;

\* I\_IMP\_CAR\_AGE = 1;

\* end;

\* IMP\_HOME\_VAL = HOME\_VAL;

\* I\_IMP\_HOME\_VAL = 0;

\* label IMP\_HOME\_VAL = 'Home Value';

\* label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

\* if missing(IMP\_HOME\_VAL) then do;

\* IMP\_HOME\_VAL = 154867.29;

\* I\_IMP\_HOME\_VAL = 1;

\* end;

\* IMP\_INCOME = INCOME;

\* I\_IMP\_INCOME = 0;

\* label IMP\_INCOME = 'Income';

\* label I\_IMP\_INCOME = 'Income Imp Flag';

\* if missing(IMP\_INCOME) then do;

\* IMP\_INCOME = 61898.10;

\* I\_IMP\_INCOME = 1;

\* end;

\* IMP\_YOJ = YOJ;

\* I\_IMP\_YOJ = 0;

\* label IMP\_YOJ = 'Years on Job';

\* label I\_IMP\_YOJ = 'Years on Job Imp Flag';

\* if missing(IMP\_YOJ) then do;

\* IMP\_YOJ = 10.4992864;

\* I\_IMP\_YOJ = 1;

\* end;

\* Drop AGE;

\* Drop CAR\_AGE;

\* Drop HOME\_VAL;

\* Drop INCOME;

\* Drop YOJ;

\*proc means data=&SCRUBFILE. nmiss mean median;

\*var \_numeric\_ ;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Correlation of all numeric values \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*proc corr data=&SCRUBFILE.;

\* var TARGET\_FLAG

KIDSDRIV

HOMEKIDS

TRAVTIME

BLUEBOOK

TIF

OLDCLAIM

CLM\_FREQ

MVR\_PTS

IMP\_AGE

IMP\_YOJ

IMP\_INCOME

IMP\_HOME\_VAL

IMP\_CAR\_AGE;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Build categories etc \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* First Model {All Variables Numeric Values} [Base Model] {0.75741} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*proc logistic data=&SCRUBFILE.;

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

HOMEKIDS

TRAVTIME

BLUEBOOK

TIF

OLDCLAIM

CLM\_FREQ

MVR\_PTS

IMP\_AGE

IMP\_YOJ

IMP\_INCOME

IMP\_HOME\_VAL

IMP\_CAR\_AGE

/selection=forward;

\*run;

\*proc logistic data=&SCRUBFILE.;

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

HOMEKIDS

TRAVTIME

BLUEBOOK

TIF

OLDCLAIM

CLM\_FREQ

MVR\_PTS

IMP\_AGE

IMP\_YOJ

IMP\_INCOME

IMP\_HOME\_VAL

IMP\_CAR\_AGE;

\*run;

\*proc logistic data=&SCRUBFILE. plot(only)=(roc(ID=prob));

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

HOMEKIDS

TRAVTIME

BLUEBOOK

TIF

OLDCLAIM

CLM\_FREQ

MVR\_PTS

IMP\_AGE

IMP\_YOJ

IMP\_INCOME

IMP\_HOME\_VAL

IMP\_CAR\_AGE;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Second Model {Select Variables Numeric Values} {0.78537} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*proc logistic data=&SCRUBFILE.;

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL;

\*run;

\*proc logistic data=&SCRUBFILE. plot(only)=(roc(ID=prob));

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Third Model {Select Variables Numeric Values} {0.79257} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* IMP\_CAR\_AGE = CAR\_AGE;

\* I\_IMP\_CAR\_AGE = 0;

\* if missing(IMP\_CAR\_AGE) then do;

\* IMP\_CAR\_AGE = 8.3283231;

\* I\_IMP\_CAR\_AGE = 1;

\* end;

\* IMP\_INCOME = INCOME;

\* I\_IMP\_INCOME = 0;

\* if missing(IMP\_INCOME) then do;

\* IMP\_INCOME = 61898.10;

\* I\_IMP\_INCOME = 1;

\* end;

\* if CAR\_USE in ('Commercial' 'Private') then do;

\* USE\_P = (car\_use eq 'Private');

\* end;

\* if MSTATUS in ('Yes' 'z\_No') then do;

\* MARRIED\_Y = (MSTATUS eq 'Yes');

\* end;

\* if REVOKED in ('No' 'Yes') then do;

\* REV\_L = (REVOKED eq 'Yes');

\* end;

\*proc logistic data=&SCRUBFILE.;

\*model TARGET\_FLAG( ref="0" ) =

CLM\_FREQ

IMP\_INCOME

MVR\_PTS

USE\_P

MARRIED\_Y

REV\_L;

\*run;

\*proc logistic data=&SCRUBFILE. plot(only)=(roc(ID=prob));

\*model TARGET\_FLAG( ref="0" ) =

CLM\_FREQ

IMP\_INCOME

MVR\_PTS

USE\_P

MARRIED\_Y

REV\_L;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Fourth Model {Combined Models 2 & 3} {7.8537} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* IMP\_HOME\_VAL = HOME\_VAL;

\* I\_IMP\_HOME\_VAL = 0;

\* label IMP\_HOME\_VAL = 'Home Value';

\* label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

\* if missing(IMP\_HOME\_VAL) then do;

\* IMP\_HOME\_VAL = 154867.29;

\* I\_IMP\_HOME\_VAL = 1;

\* end;

\* IMP\_CAR\_AGE = CAR\_AGE;

\* I\_IMP\_CAR\_AGE = 0;

\* if missing(IMP\_CAR\_AGE) then do;

\* IMP\_CAR\_AGE = 8.3283231;

\* I\_IMP\_CAR\_AGE = 1;

\* end;

\* IMP\_INCOME = INCOME;

\* I\_IMP\_INCOME = 0;

\* if missing(IMP\_INCOME) then do;

\* IMP\_INCOME = 61898.10;

\* I\_IMP\_INCOME = 1;

\* end;

\* if CAR\_USE in ('Commercial' 'Private') then do;

\* USE\_P = (car\_use eq 'Private');

\* end;

\* if MSTATUS in ('Yes' 'z\_No') then do;

\* MARRIED\_Y = (MSTATUS eq 'Yes');

\* end;

\* if REVOKED in ('No' 'Yes') then do;

\* REV\_L = (REVOKED eq 'Yes');

\* end;

\*proc logistic data=&SCRUBFILE. plot(only)=(roc(ID=prob));

\*proc logistic data=&SCRUBFILE.;

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL

IMP\_INCOME

USE\_P

MARRIED\_Y

REV\_L;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Fifth Model {Model 4+} {0.79259} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

IMP\_HOME\_VAL = HOME\_VAL;

I\_IMP\_HOME\_VAL = 0;

label IMP\_HOME\_VAL = 'Home Value';

label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

if missing(IMP\_HOME\_VAL) then do;

IMP\_HOME\_VAL = 154867.29;

I\_IMP\_HOME\_VAL = 1;

end;

IMP\_CAR\_AGE = CAR\_AGE;

I\_IMP\_CAR\_AGE = 0;

if missing(IMP\_CAR\_AGE) then do;

IMP\_CAR\_AGE = 8.3283231;

I\_IMP\_CAR\_AGE = 1;

end;

IMP\_INCOME = INCOME;

I\_IMP\_INCOME = 0;

if missing(IMP\_INCOME) then do;

IMP\_INCOME = 61898.10;

I\_IMP\_INCOME = 1;

end;

IMP\_JOB = JOB;

if missing(IMP\_JOB) then do;

if IMP\_INCOME > 100000 then

IMP\_JOB = "Doctor";

else if IMP\_INCOME > 80000 then

IMP\_JOB = "Lawyer";

else

IMP\_JOB = "z\_Blue Collar";

end;

if CAR\_USE in ('Commercial' 'Private') then do;

USE\_P = (car\_use eq 'Private');

end;

if MSTATUS in ('Yes' 'z\_No') then do;

MARRIED\_Y = (MSTATUS eq 'Yes');

end;

if REVOKED in ('No' 'Yes') then do;

REV\_L = (REVOKED eq 'Yes');

end;

Drop HOME\_VAL;

Drop CAR\_AGE;

Drop INCOME;

Drop JOB;

\*proc logistic data=&SCRUBFILE.;

proc logistic data=&SCRUBFILE. plot(only)=(roc(ID=prob));

class IMP\_JOB /param=ref;

model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL

IMP\_INCOME

USE\_P

MARRIED\_Y

REV\_L

IMP\_JOB /link=probit;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* PROC GENMOD {0.78296} \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*PROC GENMOD data=&SCRUBFILE.;

\*class IMP\_JOB /param=ref;

\*model TARGET\_FLAG( ref="0" ) =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL

IMP\_INCOME

USE\_P

MARRIED\_Y

REV\_L

IMP\_JOB;

\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* END \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Model: Target Amount \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

proc reg data=&SCRUBFILE.;

model TARGET\_AMT =

KIDSDRIV

TRAVTIME

TIF

CLM\_FREQ

MVR\_PTS

IMP\_HOME\_VAL

IMP\_INCOME;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* END \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**SAS Utilized for Scoring P\_Target\_AMT {Bonus}**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Unit 02: INSURANCE LOGISTIC REGRESSION PROJECT {Score} \*;

\* Eric Lewis Section 55 Spring 2016 \*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

%let PATH = /folders;

%let NAME = INS;

%let LIB = &NAME..;

libname &NAME. "&PATH.";

%let INFILE = &LIB.LOGIT\_INSURANCE\_TEST;

%let TEMPFILE = TEMPFILE;

data &TEMPFILE.;

set &INFILE.;

\*libname score\_me '/folders';

\*data testing;

\* set score\_me.logit\_insurance\_test;

data validate;

set &TEMPFILE.;

IMP\_HOME\_VAL = HOME\_VAL;

I\_IMP\_HOME\_VAL = 0;

label IMP\_HOME\_VAL = 'Home Value';

label I\_IMP\_HOME\_VAL = 'Home Value Imp Flag';

if missing(IMP\_HOME\_VAL) then do;

IMP\_HOME\_VAL = 154867.29;

I\_IMP\_HOME\_VAL = 1;

end;

IMP\_CAR\_AGE = CAR\_AGE;

I\_IMP\_CAR\_AGE = 0;

if missing(IMP\_CAR\_AGE) then do;

IMP\_CAR\_AGE = 8.3283231;

I\_IMP\_CAR\_AGE = 1;

end;

IMP\_INCOME = INCOME;

I\_IMP\_INCOME = 0;

if missing(IMP\_INCOME) then do;

IMP\_INCOME = 61898.10;

I\_IMP\_INCOME = 1;

end;

IMP\_JOB = JOB;

if missing(IMP\_JOB) then do;

if IMP\_INCOME > 100000 then

IMP\_JOB = "Doctor";

else if IMP\_INCOME > 80000 then

IMP\_JOB = "Lawyer";

else

IMP\_JOB = "z\_Blue Collar";

end;

if CAR\_USE in ('Commercial' 'Private') then do;

USE\_P = (car\_use eq 'Private');

end;

if MSTATUS in ('Yes' 'z\_No') then do;

MARRIED\_Y = (MSTATUS eq 'Yes');

end;

if REVOKED in ('No' 'Yes') then do;

REV\_L = (REVOKED eq 'Yes');

end;

Drop HOME\_VAL;

Drop CAR\_AGE;

Drop INCOME;

Drop JOB;

data score;

set validate;

YHAT =

0.3832 \* KIDSDRIV

+ 0.00604 \* TRAVTIME

- 0.0480 \* TIF

+ 0.2703 \* CLM\_FREQ

+ 0.1365 \* MVR\_PTS

- 0.00000140 \* IMP\_HOME\_VAL

- 0.00000511 \* IMP\_INCOME

- 0.5871 \* USE\_P

- 0.4881 \* MARRIED\_Y

+ 0.8504 \* REV\_L

- 0.0168 \* (IMP\_JOB in ("Doctor"))

- 0.0197 \* (IMP\_JOB in ("Lawyer"))

- 0.4968 \* (IMP\_JOB in ("Manager"))

- 0.0144 \* (IMP\_JOB in ("Professional"))

- 0.6214;

P\_TARGET\_FLAG = exp(YHAT) / (1+exp(YHAT));

P\_TARGET\_AMT =

1281.92319

+ 418.60635 \* KIDSDRIV

+ 6.75812 \* TRAVTIME

- 45.73608 \* TIF

+ 273.94736 \* CLM\_FREQ

+ 221.35366 \* MVR\_PTS

- 0.00230 \* IMP\_HOME\_VAL

- 0.00110 \* IMP\_INCOME;

\* keep index P\_TARGET\_FLAG P\_TARGET\_AMT;

if P\_TARGET\_AMT > 0 then;

keep index P\_TARGET\_AMT;

proc print data=score;

proc export data=score

outfile='/folders/insurance\_score\_05.csv'

dbms=csv

replace;

run;