Insurance Logistic Regression

FY 2016

Author Eric Lewis

TABLE OF CONTENTS

Contents

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		8161	0	0.2638157	0.4407276
KIDSDRIV	#Driving Children	8161	0	0.1710575	0.5115341
AGE	Age	8155	6	44.7903127	8.6275895
HOMEKIDS	#Children @Home	8161	0	0.7212351	1.1163233
YOJ	Years on Job	7707	454	10.4992864	4.0924742
INCOME	Income	7716	445	61898.10	47572.69
HOME_VAL	Home Value	7697	464	154867.29	129123.78
TRAVTIME	Distance to Work	8161	0	33.4887972	15.9047470
BLUEBOOK	Value of Vehicle	8161	0	15709.90	8419.73
TIF	Time in Force	8161	0	5.3513050	4.1466353
OLDCLAIM	Total Claims (Past 5 Years)	8161	0	4037.08	8777.14
CLM_FREQ	#Claims (Past 5 Years)	8161	0	0.7985541	1.1584527
MVR_PTS	Motor Vehicle Record Points	8161	0	1.6955030	2.1471117
CAR_AGE	Vehicle Age	7651	510	8.3283231	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.

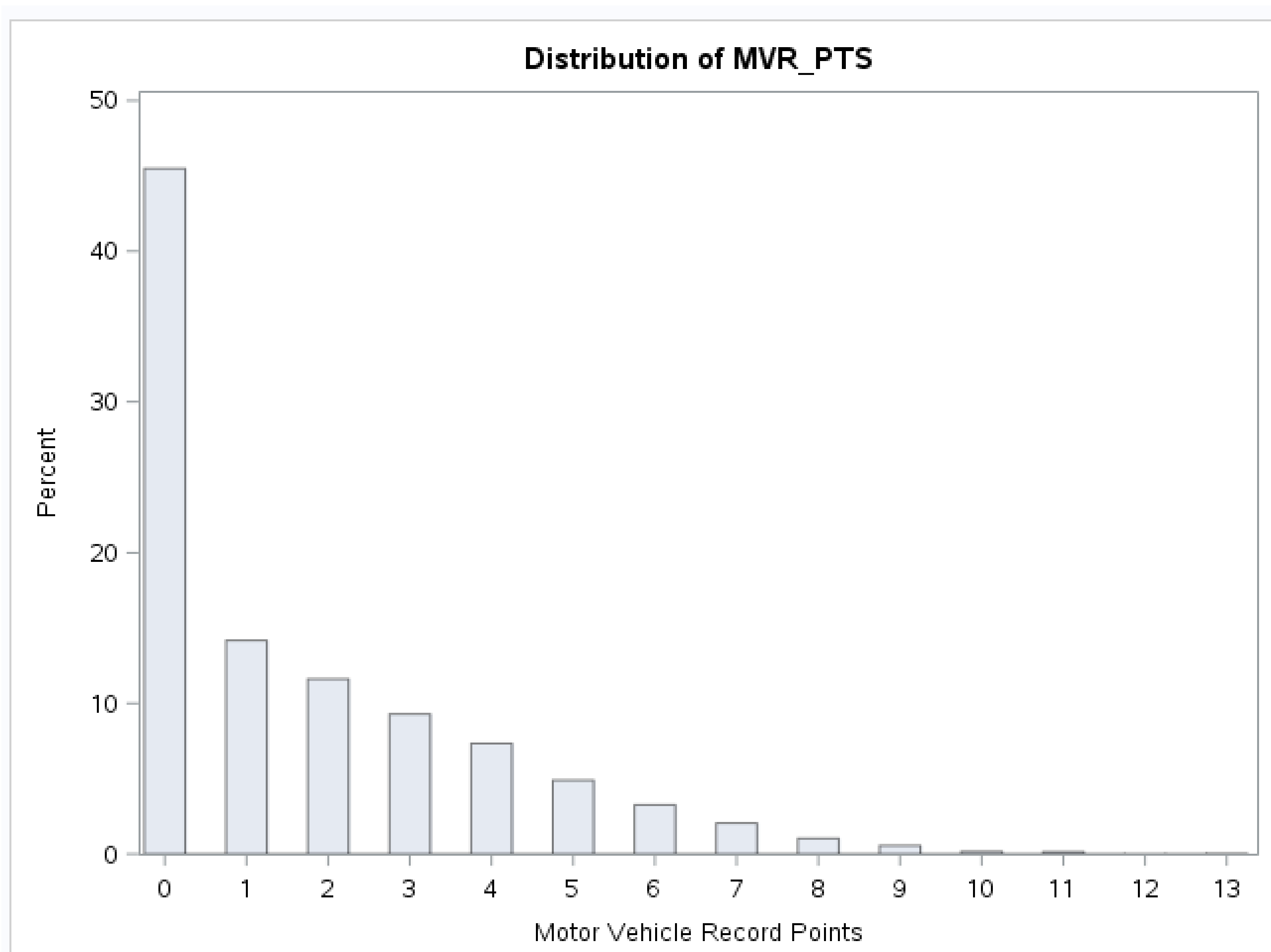

Figure 1: Distribution of MVR_PTS

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.

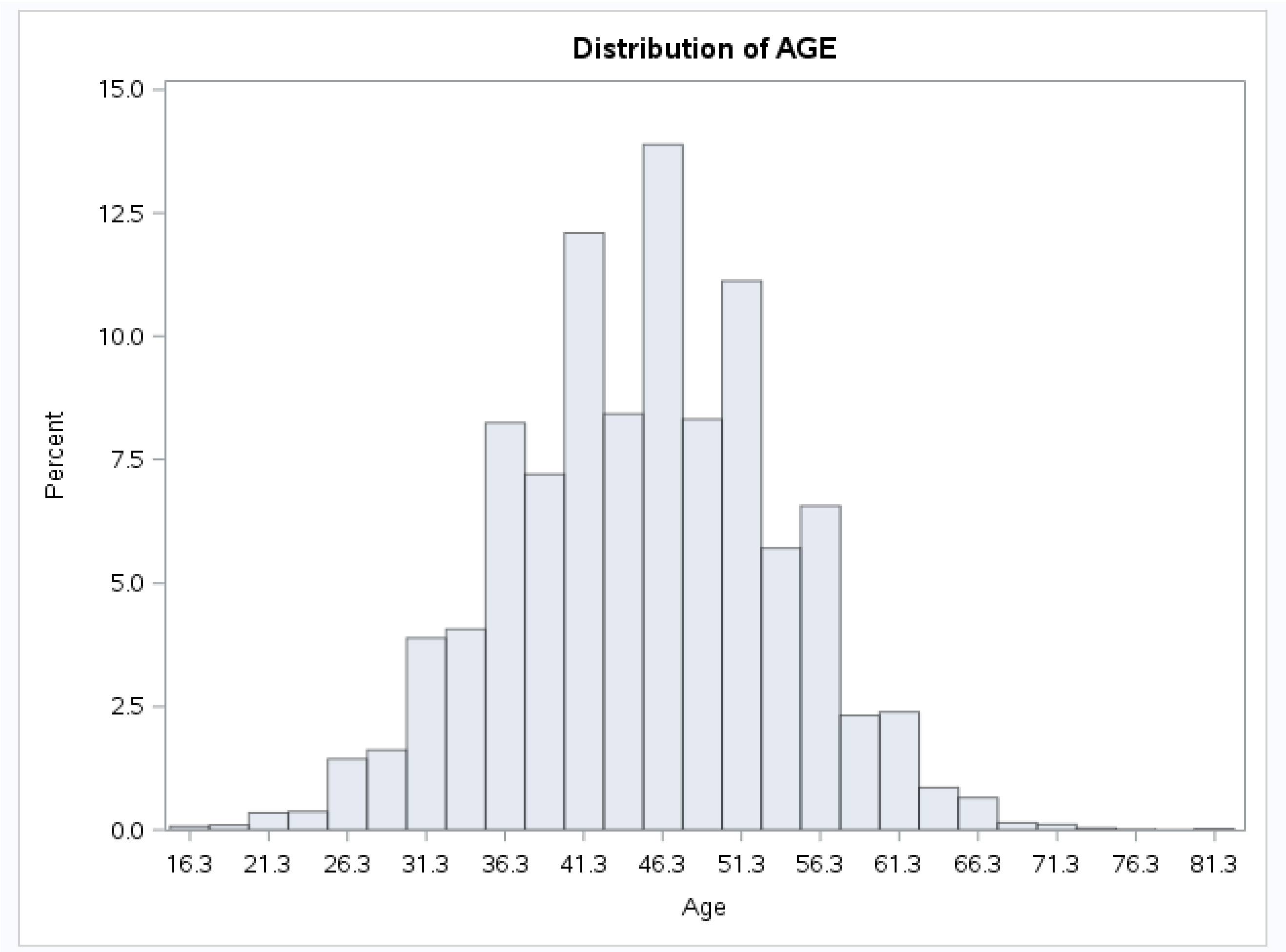

Figure 2: Distribution of AGE

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.

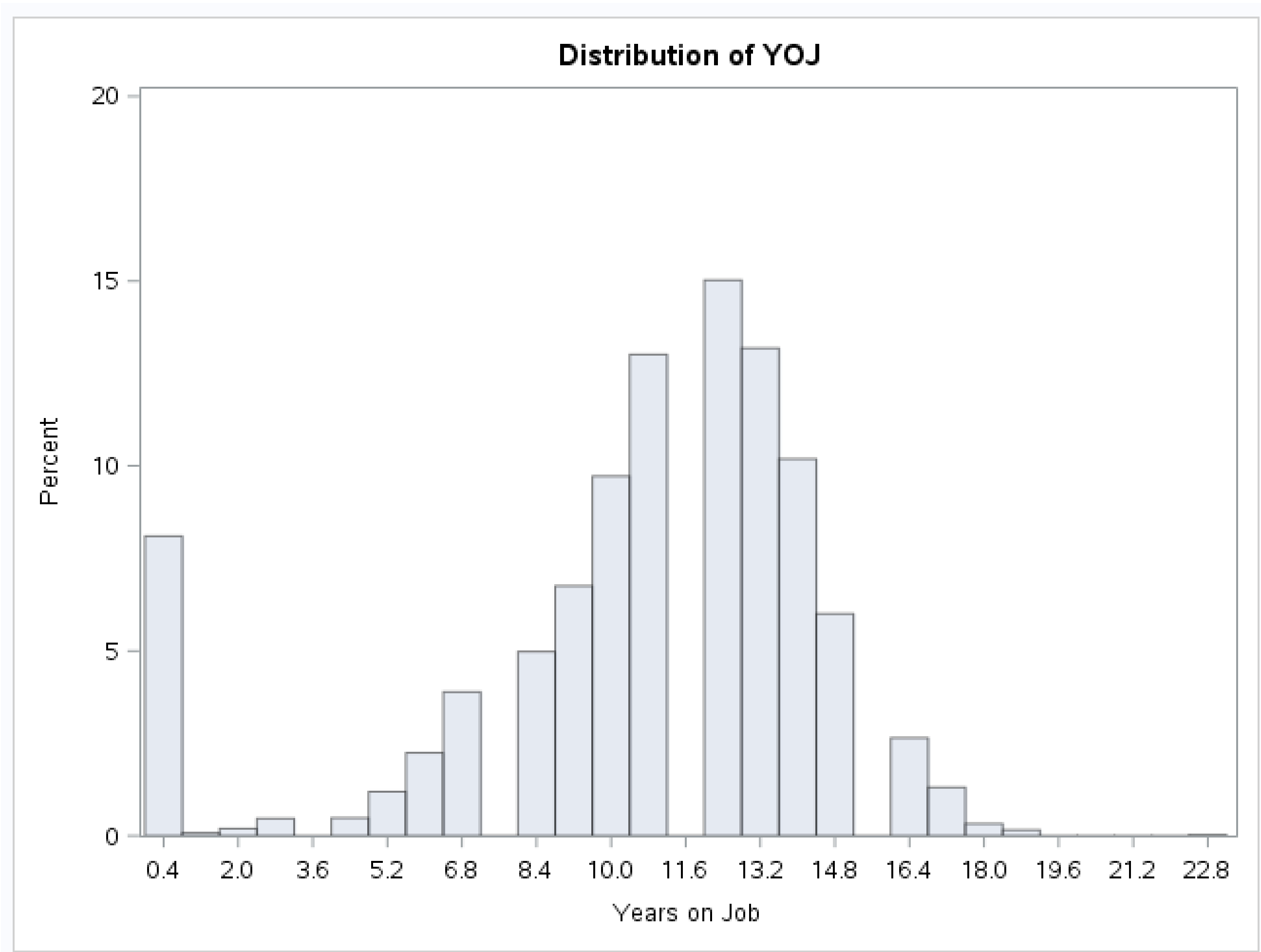

Figure 3: Distribution of YOJ

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.

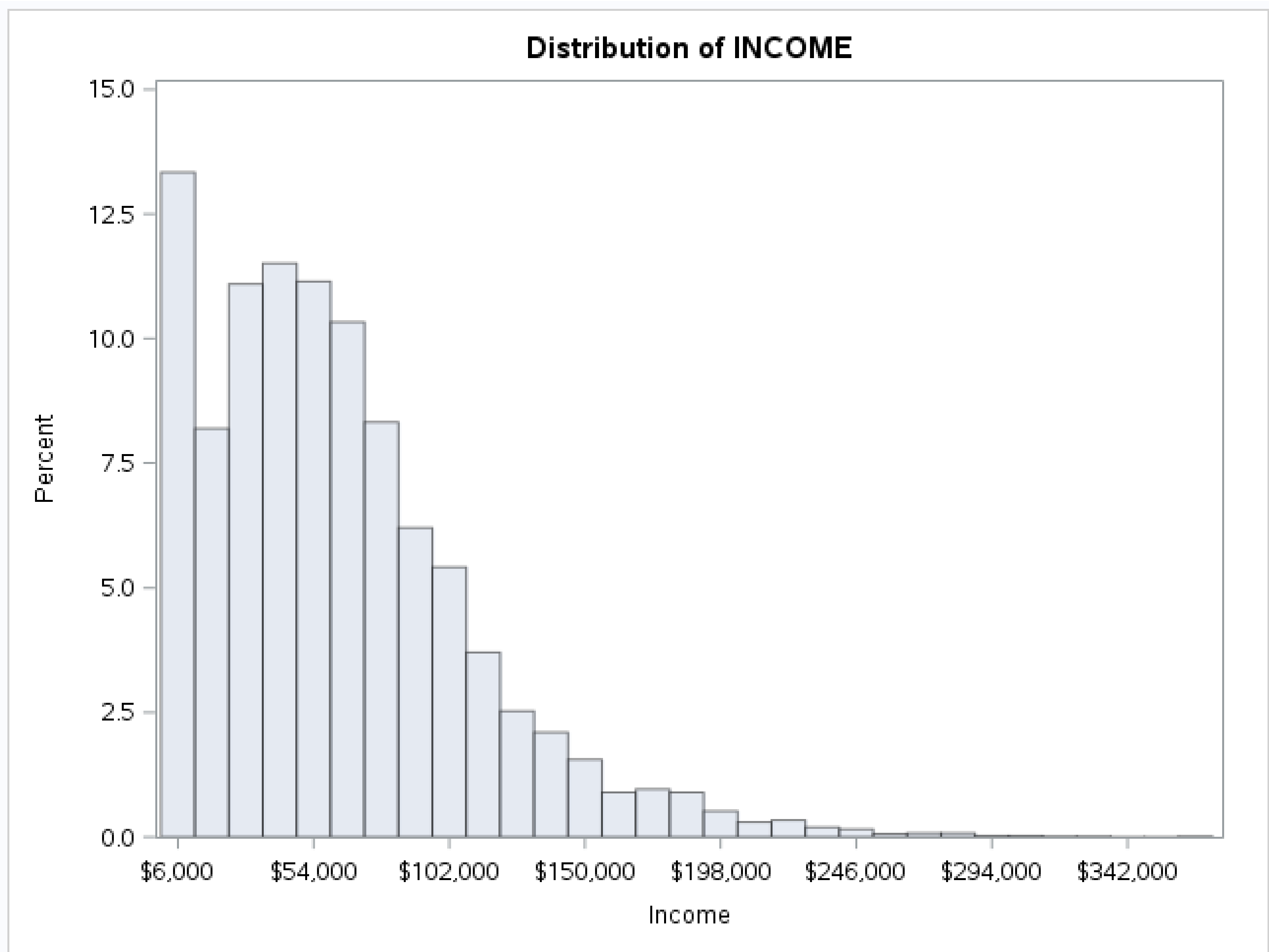

Figure 4: Distribution of INCOME

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.

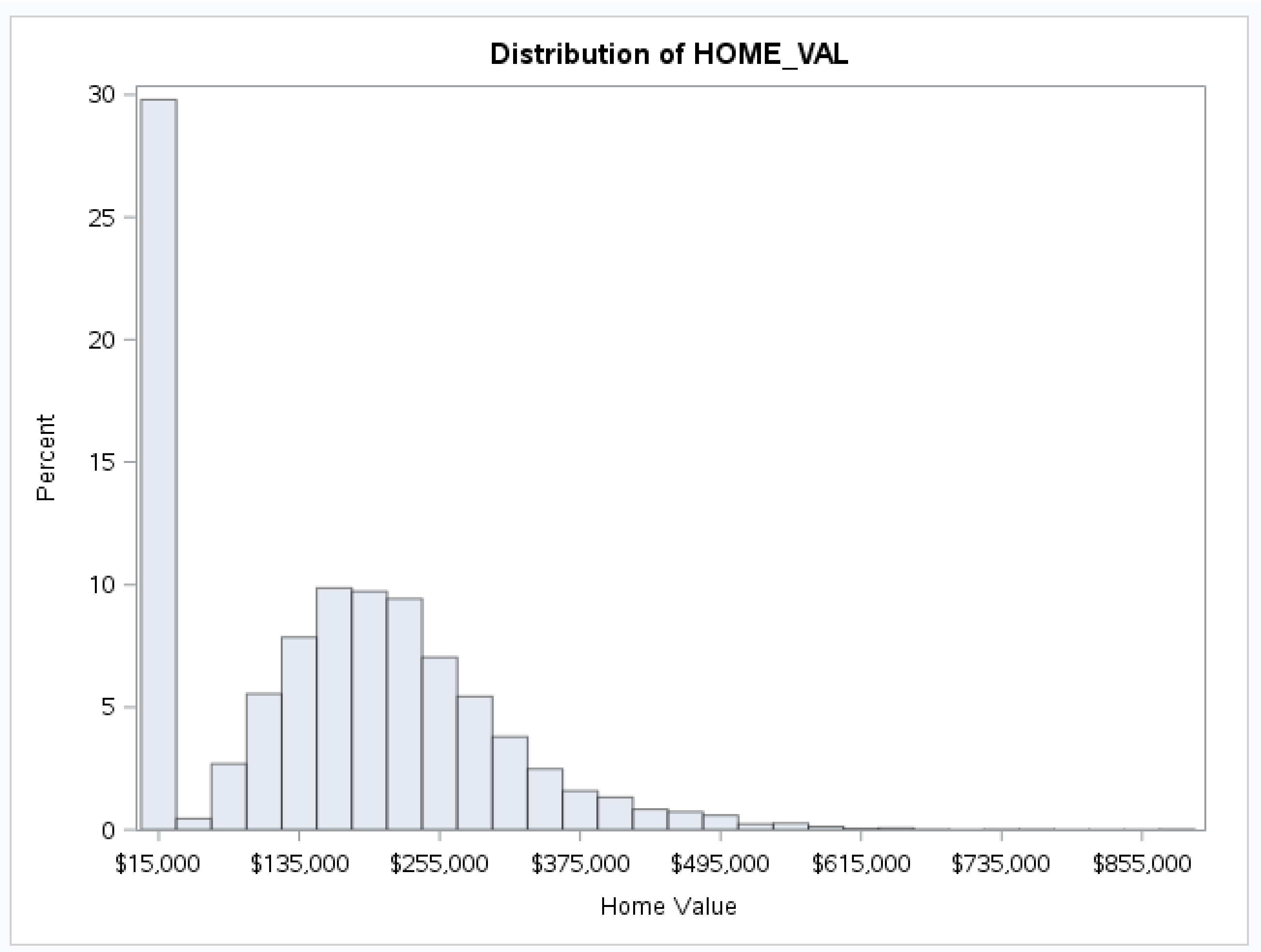

Figure 5: Distribution of HOME_VAL

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.

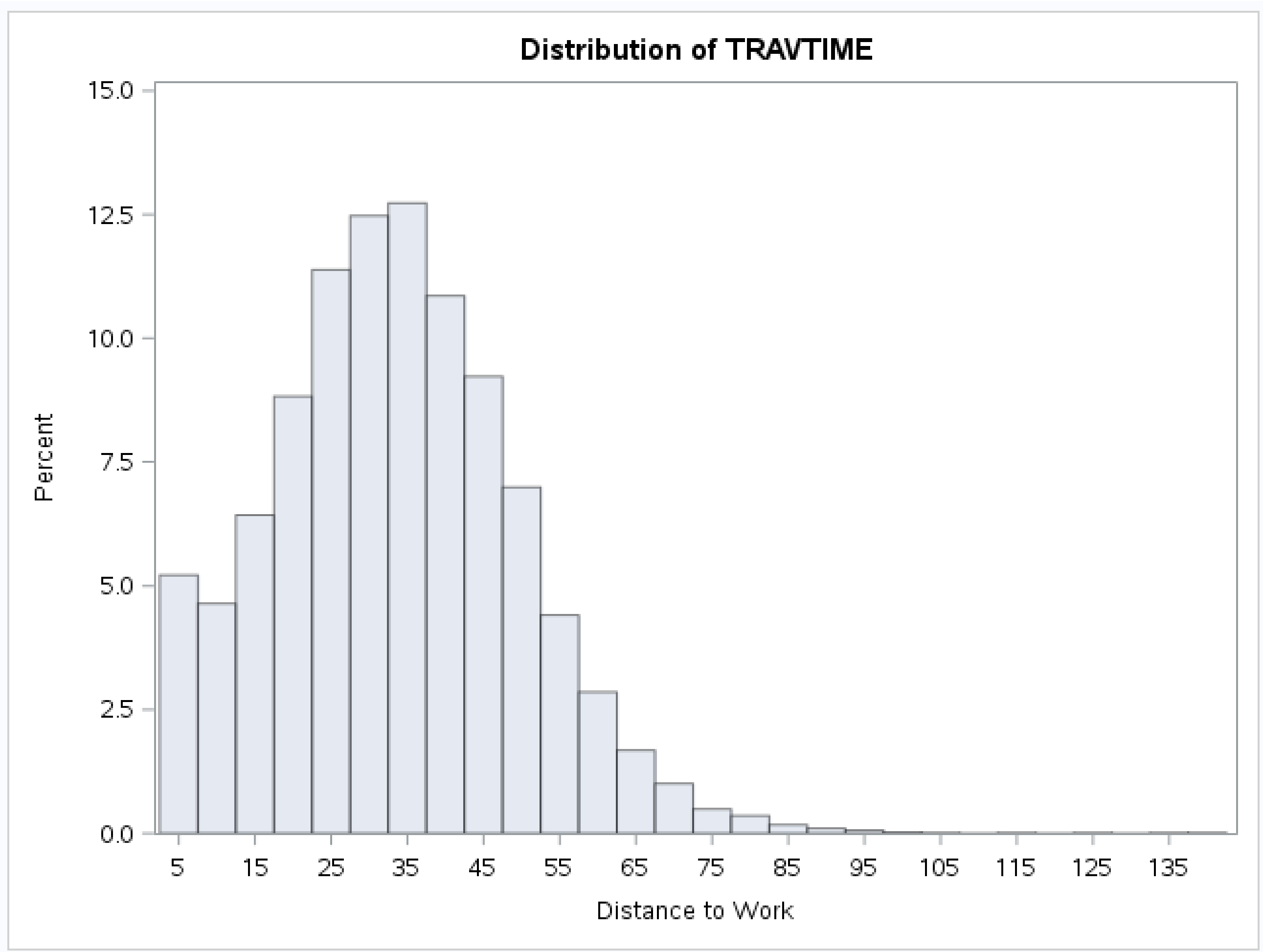

Figure 6: Distribution of TRAVTIME

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

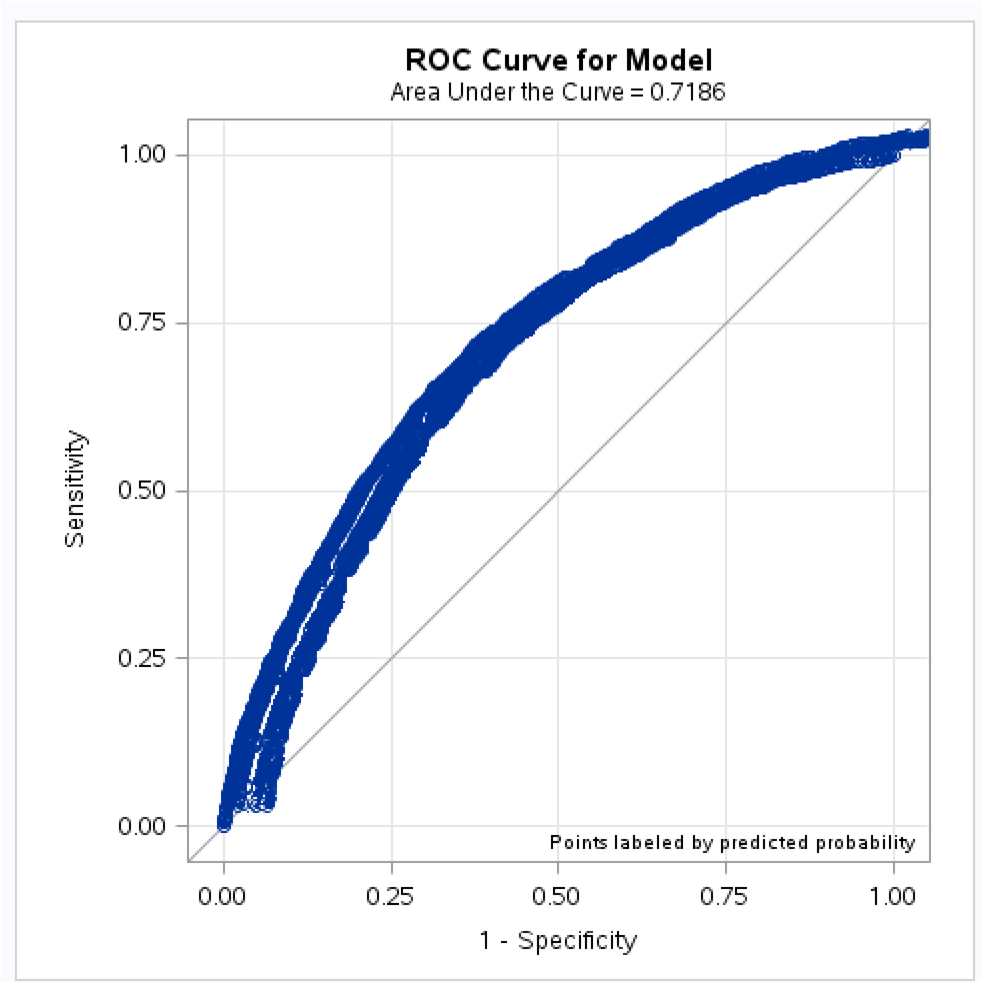

Figure 7: ROC Curve for Model 1, area {0.7186}

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.

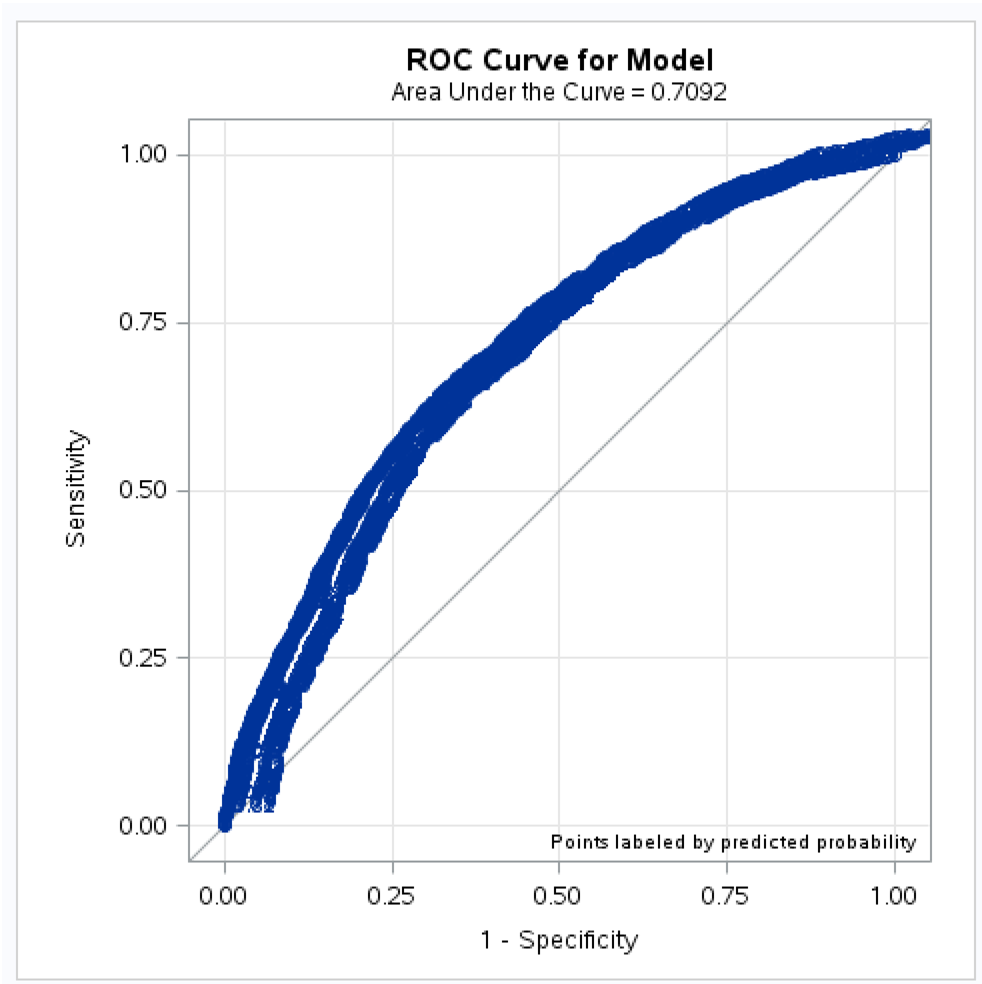

Figure 8: ROC Curve for Model 2, area {0.7902}

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.

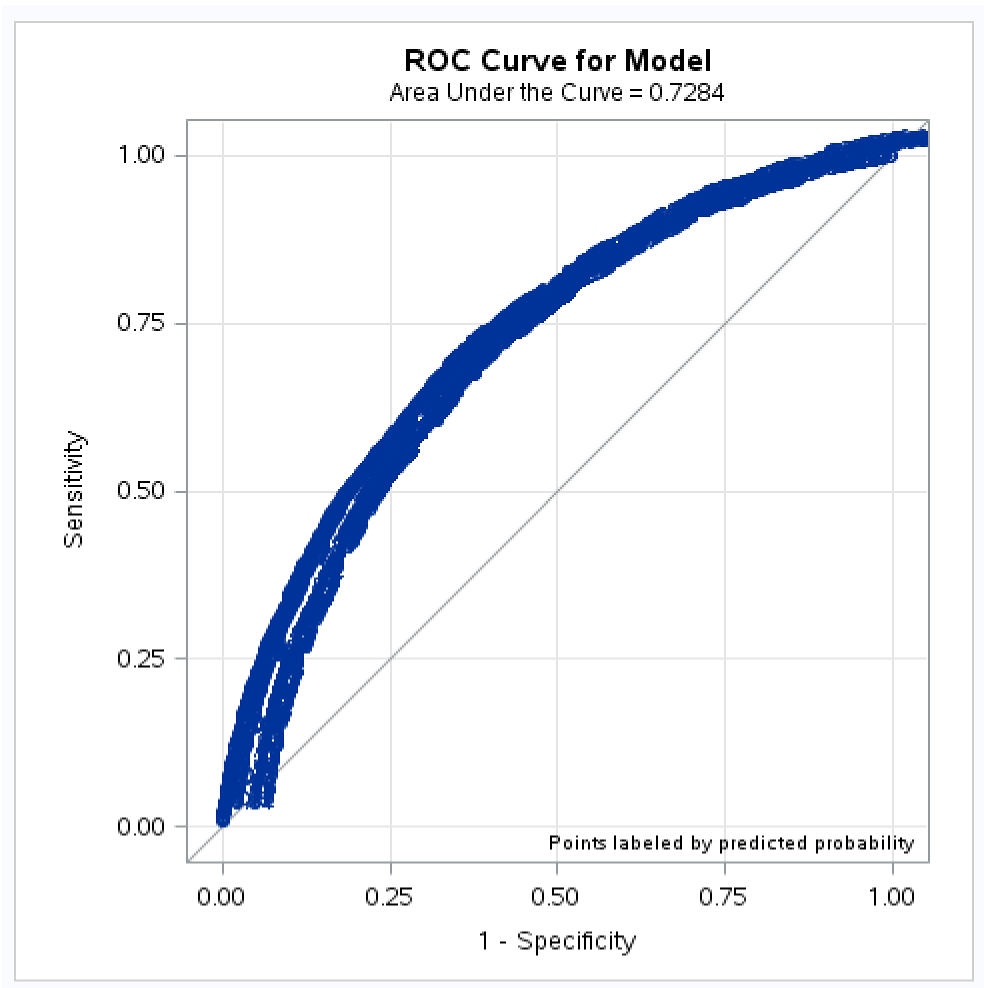

Figure 9: ROC Curve for Model 3, area {0.7284}

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.

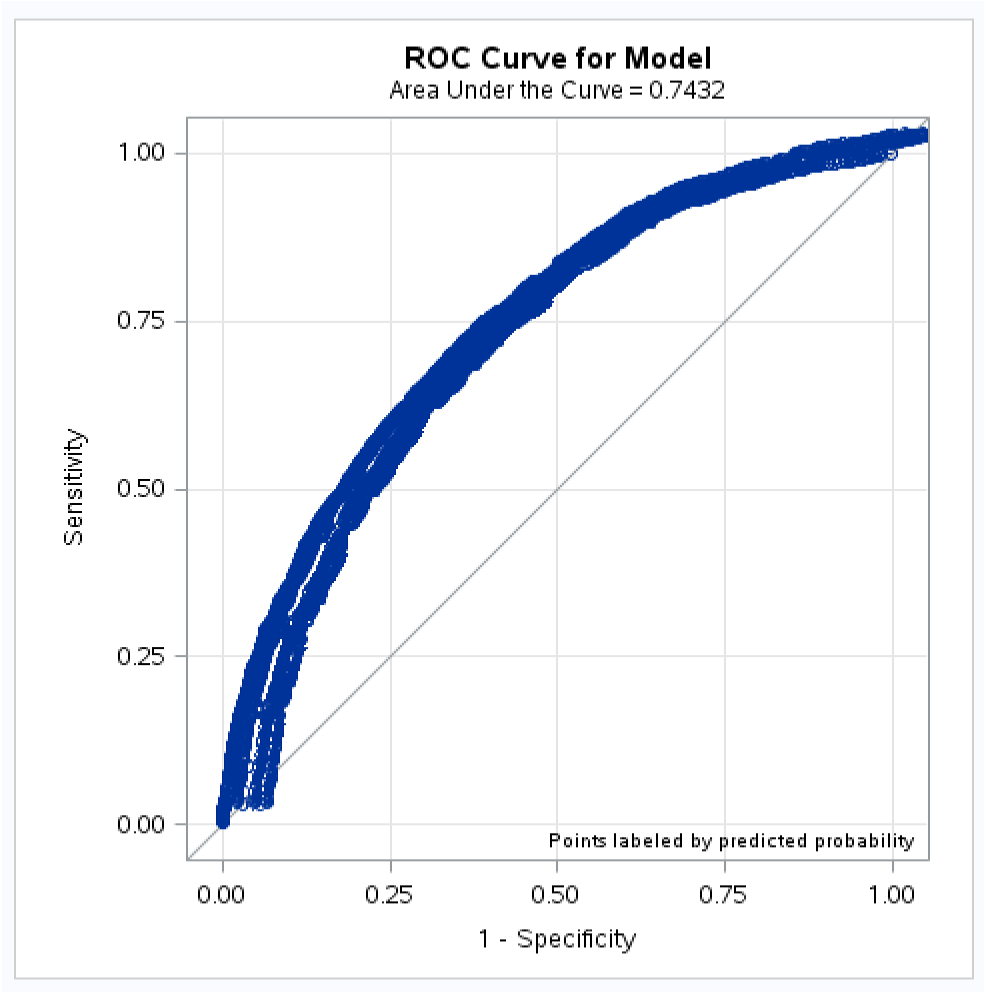

Figure 10: ROC Curve for Model 3, area {0.7432}

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
2. **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;
```