An Analysis of the Insurance Industry

Applying Predictive Models with Logistic Regression

Scott Herman

February 8th, 2017

Prepared for Predict-411: Generalized Linear Models
Northwestern University Masters in Science, Predictive Analytics
Kaggle File Submission: Insurance_Scored_STH.csv

Introduction

The purpose of this analysis is to examine the customer records of an auto insurance company in order to predict the probability of risk that an individual has of getting into an automobile accident, as well as the predicted severity of damages of those who are involved in a car crash, in order to ultimately predicted the expected losses for the insurance company. The data set for this analysis contains 8,161 observations based upon a set of customer-specific characteristics and behaviors which we will utilize in developing our predictive models.

There are two target variables within this data set which will be used in developing our predictive models. The first target variable, Target Flag, is binary in nature and signifies whether or not a customer has been involved in an auto accident. Target Amount, the second target variable, represents the amount of coverage provided to the customer paid by the insurer in the event of a car crash. Our initial model for Target Flag will be built using Logistic Regression methods estimating the probability of a customer getting into a car crash. Linear Regression will be used in developing the model to predict the Target Amount of damages paid by the insurance company in the event that a customer is involved in an accident. We will utilize a number of variable selection techniques in order to devleop this our model The results from each will compared and adjusted in order to identify the best combination of predictor variables which appear to maximize the accuracy within the model results.

Background

In terms of existing research on the subject, the amount of information released by major insurance firms related to how they specifically predict an individual's risk of getting into an auto accident is fairly limited, as these firms are in the business of maximizing their profits while minimizing cost. This makes them reluctant to share any proprietary trade secrets. However, there are studies that do examine how major insurers align their prices to their customers based upon a selected criteria of personal characteristics demonstrated by each individual customer, including a September 2016 report by the Consumer Federation of America. This results found that major insurers tend to charge high income drivers more than their lower-income customers, even when the lower-income individual has a perfect driving record. In fact, the study found that that 85% Progressive charges moderate-income drivers with perfect driving records more for their insurance premiums than their upper-Income counter-part with multiple violations on their record1. This will be in interesting relationship to examine as we get into the second phase of modeling development to predict the Target Amount.

Additionally, the study further goes on to provide a generalized list some of the typical variables that are likely to increase one's probability of getting into an accident. Among these characteristics, age, credit score, and driving history were cited as the most important. Younger drivers are typically involved in more accidents than older ones, and those with lower credit scores tend to file higher than average insurance claims. These initial examinations will provide some initial direction in our exploratory analysis of the insurance data set and it will be wise to keep these factors in mind as we further examine the relationships between the variables within our data set.

As a preface to our analysis, Table 1 below includes the variable defintions for each variable analyzed in our research. These definitions describe the theoretical effect each of our predictor variables may have on our target responses. These assumptions will tested along the way, and ether validated or denied based upon the results of our statistical analysis.

Table 1: Variable Definitions

Variable	Theoretical Correlation
AGE	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Unknown effect on probability of collision
CAR_AGE	Unknown effect on probability of collision
CAR_TYPE	Unknown effect on probability of collision
CAR_USE	Commerical Verhicles are driven more which might increase likelhood of crash
CLM_FREQ	In thoeory, those who have filed a claim in the past are more likely to do so in the future.
EDUCATION	Unknown effect on probability of collision
HOMEKIDS	Unknown effect on probability of collision
HOME_VAL	In theory, home owners tend to drive more responsibly.
INCOME	In theory, rich people tend to get into fewer accidents.
JOB	White collar workers genereally tend to crash less.
KIDSDRIV	When parents have kids in the vehicle, they tend to drive much safer.
MSTATUS	Unknown effect.
MVR_PTS	If you get lots of traffic tickets, you tend to get into more crashes.
OLDCLAIM	If you've submitted past claims, you tend to get into more crashes.
PARENT_1	Unknown effect.
RED_CAR	Urban legend says that red cars, (especially red sports cars) are more risky. Is that true?
REVOKED	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Urban legend says that women have less crashes then men. Is that true?
TARGET_AMT	Response 2: If a person crashes, how much did it cost the insurance agency
TARGET_FLAG	Response 1: The number of times a person crashes their car; 1=YES,2=NO
TIF	People who have been customers for a long time are usually more safe.
TRAVTIME	Long drives to work usually suggest greater risk
URBANICITY	Unknown effect
YOJ	People who stay at the same job for a long time are less likely to crash

Data Exploration

This analysis begins by understanding the variables within our data set along with their corresponding observations in an attempt to identify the structure of our data. Our initial data set contains 25 total variables; 23 predictor variables which will be utilized to accurately predict the probability and amount of our two response variables. In this initial phase of our reserch, it was also determined that data set is made up of both categorial and numeric variables. In order to accurately evaluate these two varaiable types, we will explore their relationships to our target variables separately, as this information will be critically important to pay attention to as we begin building our models. First, we will generate some visuals to undertand the initial relationships contained within this data set.

Figure 1 shows us the distribution of Target Flag and indicates that there are far more individuals who have not been involved in a car crash than those who have. The percentage of all records within our data set possessing a Target Flag value of 1 is 26.38%. This will be important to note when attempting to determine the factors that have the greatest impact on an individuals' liklihood of getting into a crash.

Figure 1: Distribution of Target Flag

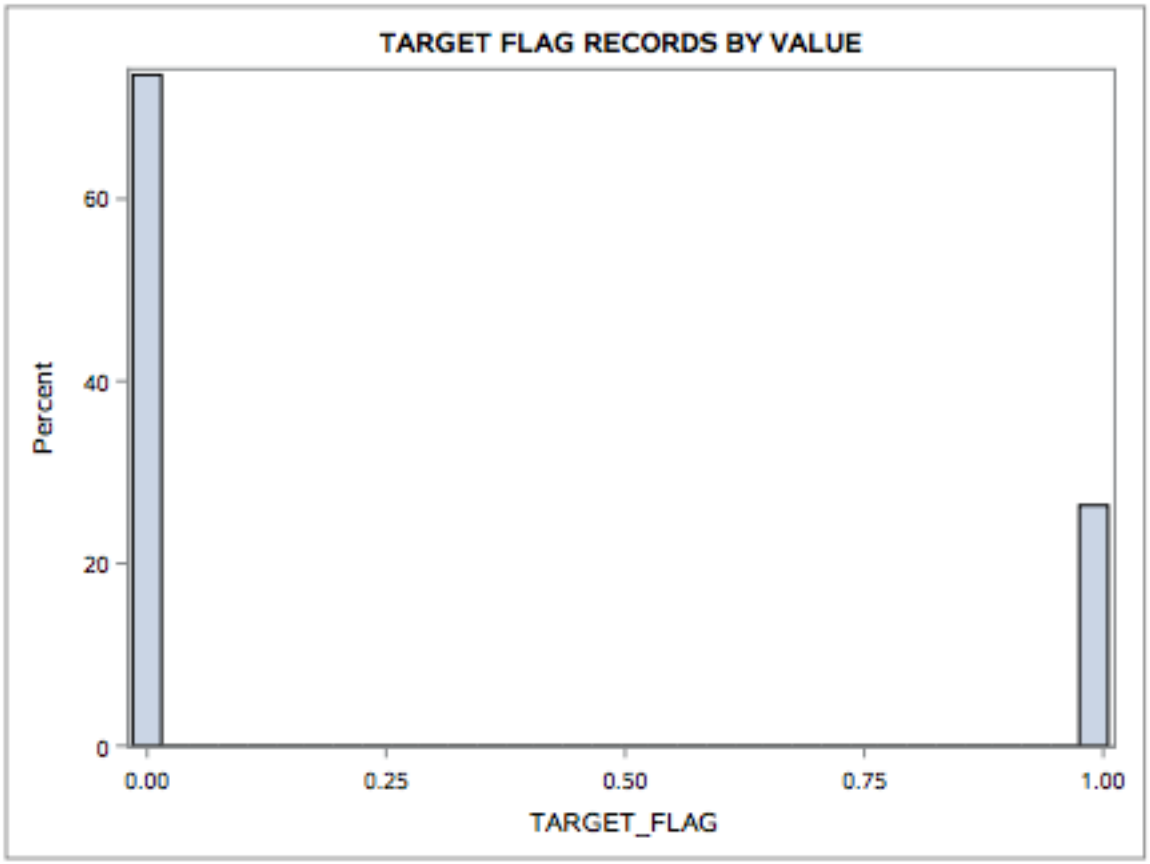

Figure 1: Distribution of Target Flag

Next, let's analyze the distributions of our numeric response variables. Figure 2 and Figure 3 below show a scatterplot matrix of these relationships. It appears we have two separate types of numeric variables here. One set, given within the first visual, indicates a set of continous variables.

Figure 2: Continous Variable Distributions

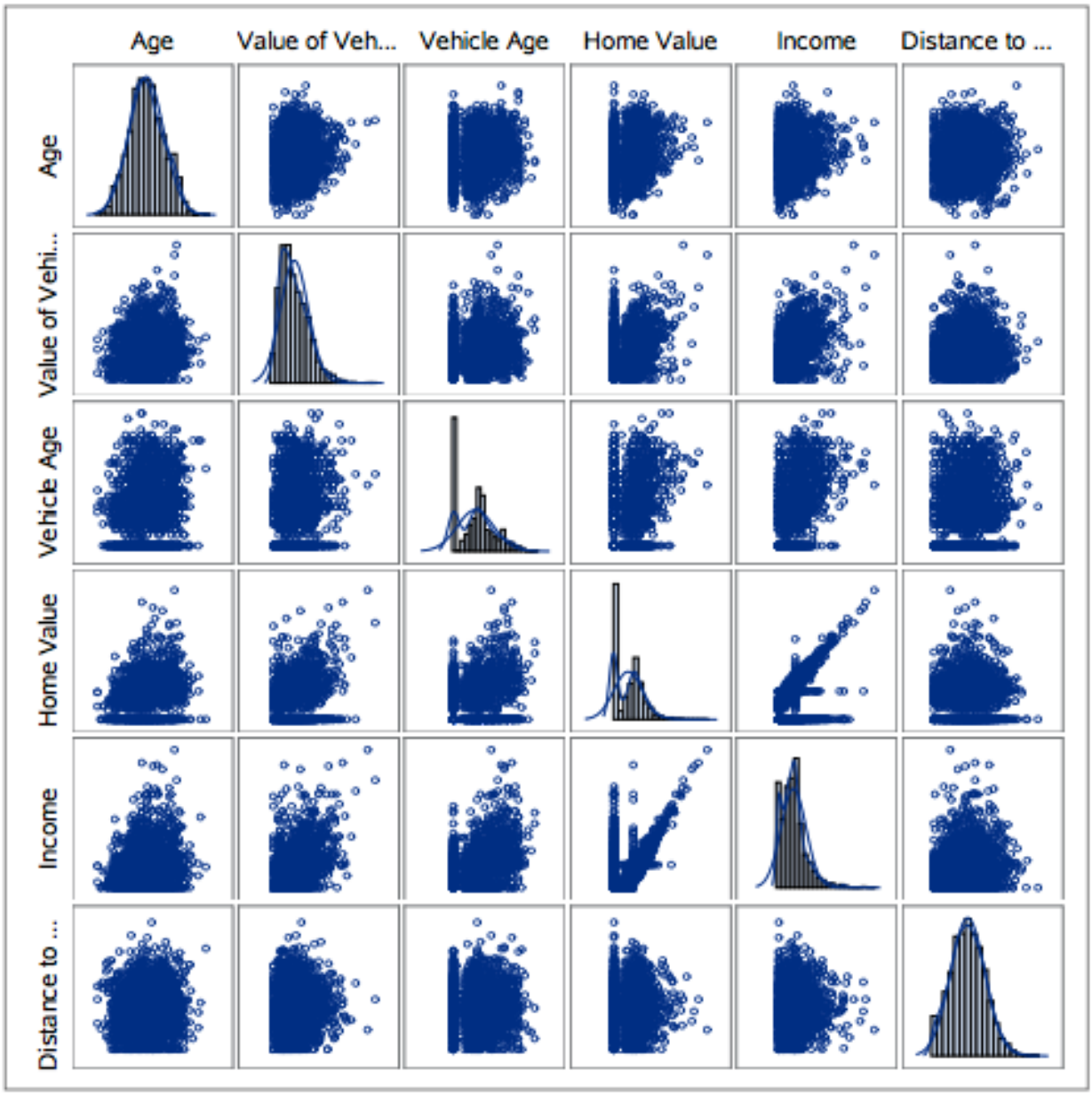

Figure 2: Continuous Variable Distributions

Figure 2 shows us that we also have some variables that appear continuous in nature, but are actually composed of interval level values. These will needed to be treated separately when we reach our model building stage, as these variables will impact the results of our predicted Target Flag response.

Figure 3: Interval Variable Distributions

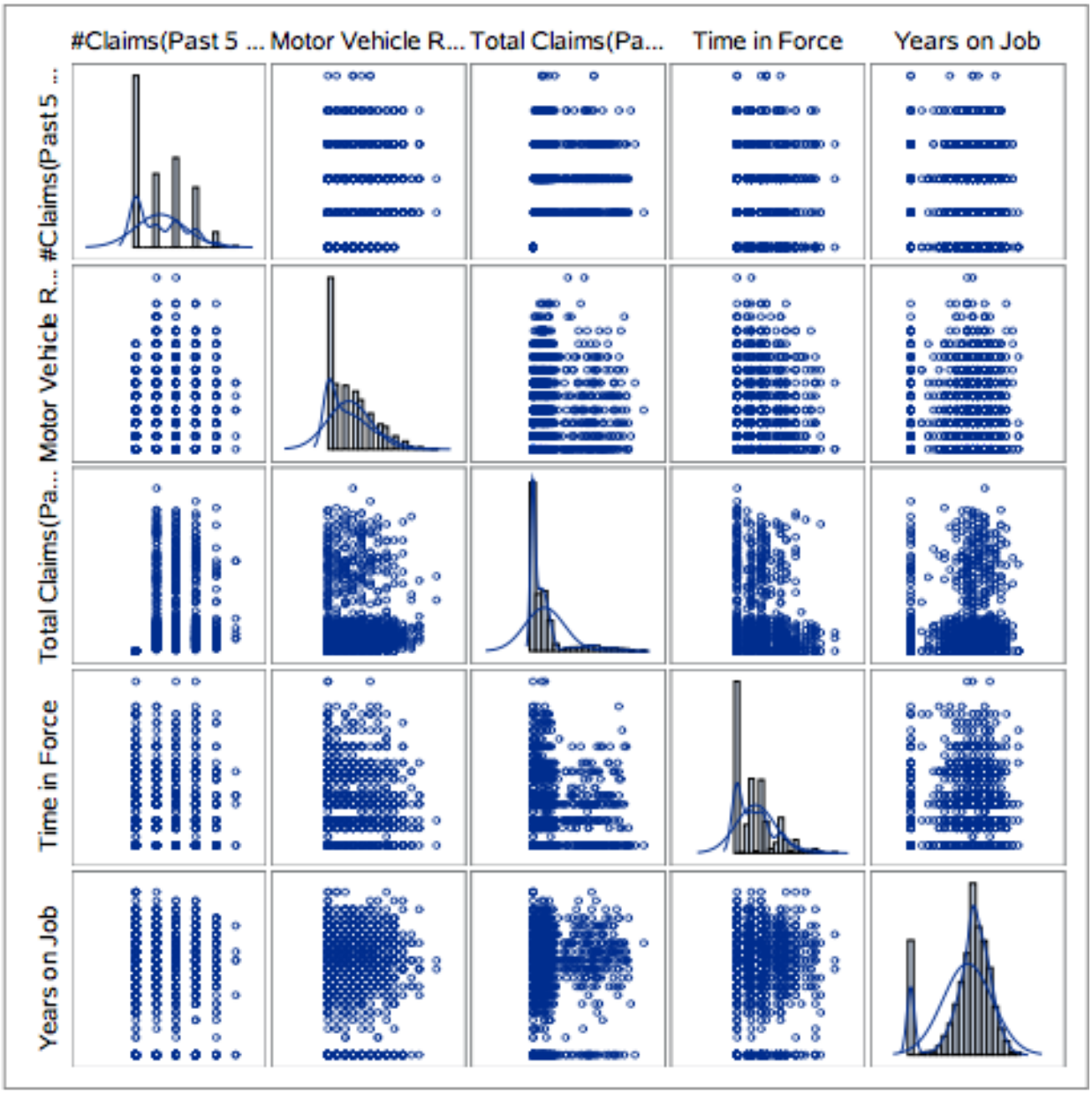

Figure 3: Interval Distributions

Table 2 below gives the summary results for each of the numeric variable within our data set.

Table 2: Numeric Variable Summary Results

Variable	N	Mean	Std Dev
AGE	8155	39	8.628
BLUEBOOK	8161	15710	8419.734
CAR_AGE	7651	8.328	5.701
CLM_FREQ	8161	0.799	1.158
HOME_VAL	7707	154867.290	129123.777
INCOME	7716	28093.883	47572.687
MVR_PTS	8161	1.696	2.147
OLDCLAIM	8161	4037.076	877.139
TARGET_AMT	8161	1504.325	4704.027
TIF	8161	5.351	4.147
TRAVTIME	8161	33.489	15.905
YOJ	7707	10.499	4.092

These results reveal that there are a number of missing values within our data set that will need to be corrected. Prior to making these adjustments, we'll analyze the remaining variables within our data set. Next, we'll analyze the values contained by our cateogrial variables. These results are given in Table 2.

Table 2: Categorical Variable Levels

Variable	N	N Levels
KIDSDRIV	8155	4
HOMEKIDS	8161	5
PARENT_1	7651	2
MSTATUS	8161	2
SEX	7707	2
EDUCATION	7716	5
JOB	8161	9
CAR_USE	8161	2
CAR_TYPE	8161	6
RED_CAR	8161	2
REVOKED	8161	2
URBANICITY	8161	2

Based upon the summary results above, it appears that our data set possesses a total of 8,161 observations recorded for each variable. However, this also reveals that we have a number of variables with missing values that will need to be accounted for. After reviewing the records across all of our variables, there are a total of six variables with missing values that will need to be corrected prior to our model development. The varaibles identified with missing records are highlighted below, along with the number of missing values for each:

Table 3: Numeric Variables with Missing Values

Variable	nMissing Values
AGE	6
YOJ	454
INCOME	445
HOME_VAL	464
CAR_AGE	510
JOB	526

Additionally, a potential input error was also spotted in this step. Car Age, appears to have a minimum value for listed as "-3 ".Since it is not possible to have a car with this age, it will be important to assess and adjust this when we reach the data preparation stage.

In order to make the proper adjustments to these missing values, we'll also need to understand their distributions. In this step, a number of different graphical procedures were used to visualize the distributions for each of our variables which enables us to understand the correlation to Target Flag, as well as in identifying potential outlier values. Additionally, this will help us determine if thre are additional ways we can group these records that will promote simplicity and increased predictive accuracy.

Data Corrections and Preparation

After uncovering the initial quality issues possessed by our data set, we'll proceed with making the proper adjustments in order to clean and prep our data set. Again, this step of the process is critical in promoting accuracy within our final model results. In moving forward with this step, the identified variables with observations that will need to be adjusted are; AGE, YOJ, INCOME, HOME VAL, CAR AGE, and JOB. Rather that deleting these missing observations from our data set, we'll create and include a set of imputed variables that will replace these null values. After understanding these distributions, we decided to utilize the median values for each variable in place of these missing observations. These changes are documented below in Table 4.

Table 4: Variable Imputations

Imputed Variable	Imputation Values
IMP_AGE	45
IMP_YOJ	11
IMP_INCOME	54028
IMP_HOME_VAL	161160
IMP_CAR_AGE	8
IMP_JOB	Blue Collar

The figures above revealed a few variables with uneven and skewed distributions, with a large number of values at or below zero. In order to address this, we've created a set of design variables that will group these observations into separate values based upon their probability of a car crash. This should promote increased accuracy of our probability model results.

Table 5: Design Variable Levels

Design Variable	Reference Levels
HOMEOWN	2
AGE_GROUP	4
BLUEBOOK_GROUP	4
CAR_AGE_GROUP	2
INCOME_GROUP	4
OLDCLAIM_GROUP	4
COMMUTE	4
POINTS	3
DRIVE_KIDS	2
KIDS_HOME	2

Our newly created design variables, along with our imputed variables will be utilized in our final set of predictor variables when we begin building our logistic models. The final set of predictor variables is listed below in the table below.

Table 6: Final Set of Predictor Variables

Variable	N
IMP_AGE	8161
BLUEBOOK	8161
IMP_CAR_AGE	8161
CLM_FREQ	8161
IMP_HOME_VAL	8161
IMP_INCOME	8161
MVR_PTS	8161
OLDCLAIM	8161
TRAVTIME	8161
IMP_YOJ	8161
JOB	8161
EDUCATION	8161
CAR_TYPE	8161
SEX	8161
MSTATUS	8161
PARENT1	8161
RED_CAR	8161
REVOKED	8161
URBANICITY	8161
HOMEOWN	8161
AGE_GROUP	8161
BLUE_BOOK_GROUP	8161
CAR_AGE_GROUP	8161
HOMEVAL_GROUP	8161
INCOME_GROUP	8161
OLDCLAIM_GROUP	8161
COMMUTE	8161
POINTS	8161
DRIVE_KIDS	8161
KIDS_HOME	8161

Variable Selection and Model Development Process

After successfully cleaning and prepping our data set, we'll move onto the model development process. In devleping our logistic model, three different variable selection techniques were utilized; Backward, Stepwise and Manual Selection. The results from each of these procedures are summarized below, and were used to compare how well each model fit the actual results. These comparisons were evaluated using each model's resulting AIC Score, the Maximum Likelihood Estimates of the variables selected within each, as well as the ROC Curves in determine the probabilities accounted for within each. Additional consideration was also given to the interpretability and simplicity of the model itself.

In addition, the data set still contains a number of categorical variables that contain more than two levels. This will also need to be considered and accounted for when developing our Logistic Model. This procedure requires us to use a parameter reference. These parameters should be based upon the level-value indicating the lowest correlation to a car-crash for each of these variables. Based upon the SAS CORR procedure, we've identified the parameters for each of these variables, below:

Table 8: Categorical Variables and Parameter References

Categorical Variable	Parameter Reference
SEX	Female
JOB	Doctor
EDUCATION	PhD
CAR_TYPE	Minivan
CAR_USE	Commercial
PARENT1	No
REVOKED	No
HOMEOWN	Yes
POINTS	0
COMMUTE	0

Logistic Regression Summary 1: Backward Selection

The first model was developed using Backward Selection, and consisted of 38 total variables. The AIC Score produced by this technique was 9419.962 which will serve as our initial baseline in determining the predictive strength of our following models. The Backward Selection Fit Statistics are listed below.

The ROC Curve produced by this procedure resulted in a predicted probability score of 0.7666. This appears to be a strong initial score, but again the results seen with this technique will be utilized in comparision with our results to follow. The Backward Selection ROC Curve is plotted below.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7983.621
SC	9426.969	8249.891
-2 Log L	9417.962	7907.621

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1510.3416	37	<.0001
Score	1423.3201	37	<.0001
Wald	1134.3974	37	<.0001

Figure 4: Backward Selection Fit Statistics

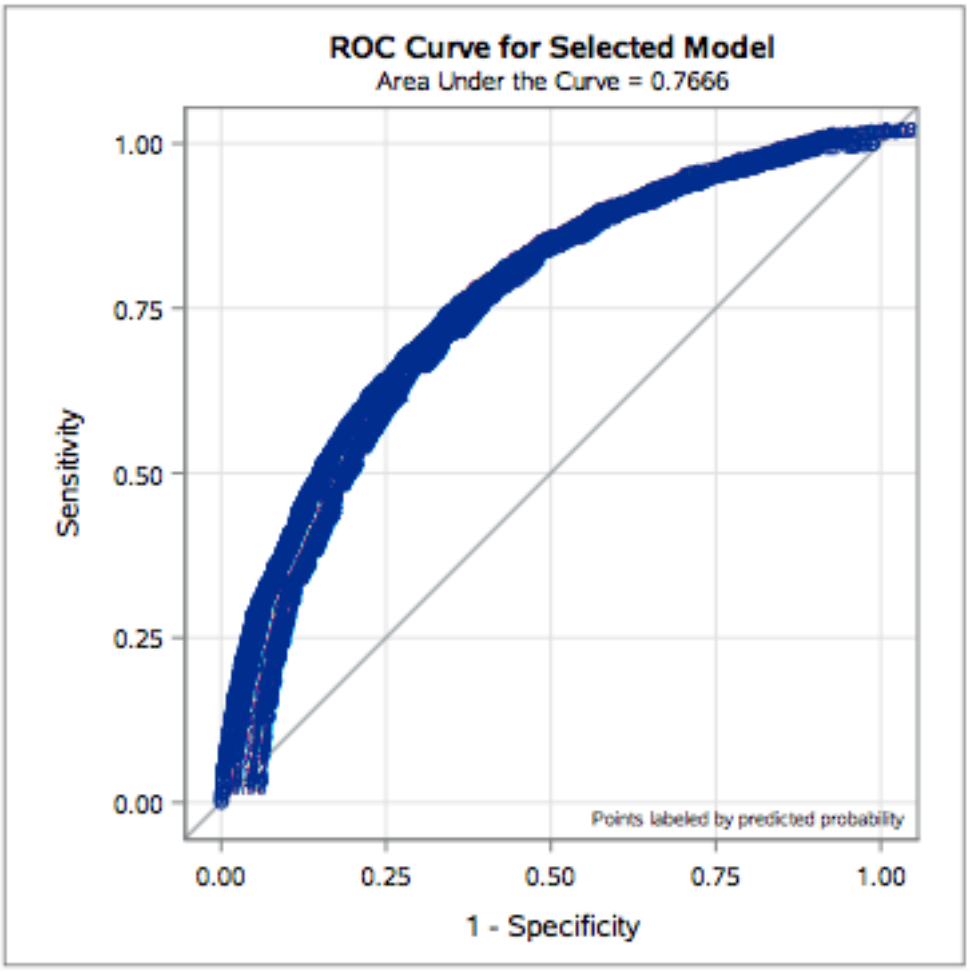

Figure 5: ROC Curve: Backward Selection

Logistic Regression Summary 2: Stepwise Selection

The next procedure utilized the Stepwise procedure, which selected 32 total variables and appears to improve upon our initial model. The AIC Score achieved by this technique is equivalent to the score produced by Backward Selection, however, our Chi-Score Likelihood Ratio improved to 1939.9748 We'll want to compare these results, along with the ROC Score achieved by this procedure in order to determine the strength of this developed model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7541.987
SC	9426.969	7766.215
-2 Log L	9417.962	7477.987

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1939.9748	31	<.0001
Score	1717.9599	31	<.0001
Wald	1325.2115	31	<.0001

Figure 6: Stepwise Selection Fit Statistics

In reviewing the ROC Curve visualized below, it appears that our Stepwise Selection produdure has improved upon our initial model, as this technique produced a predicted probability score of 0.8015. Although this is a considerable improvement, we'll want to assess this with the results attained from our third and final seleciton technique.

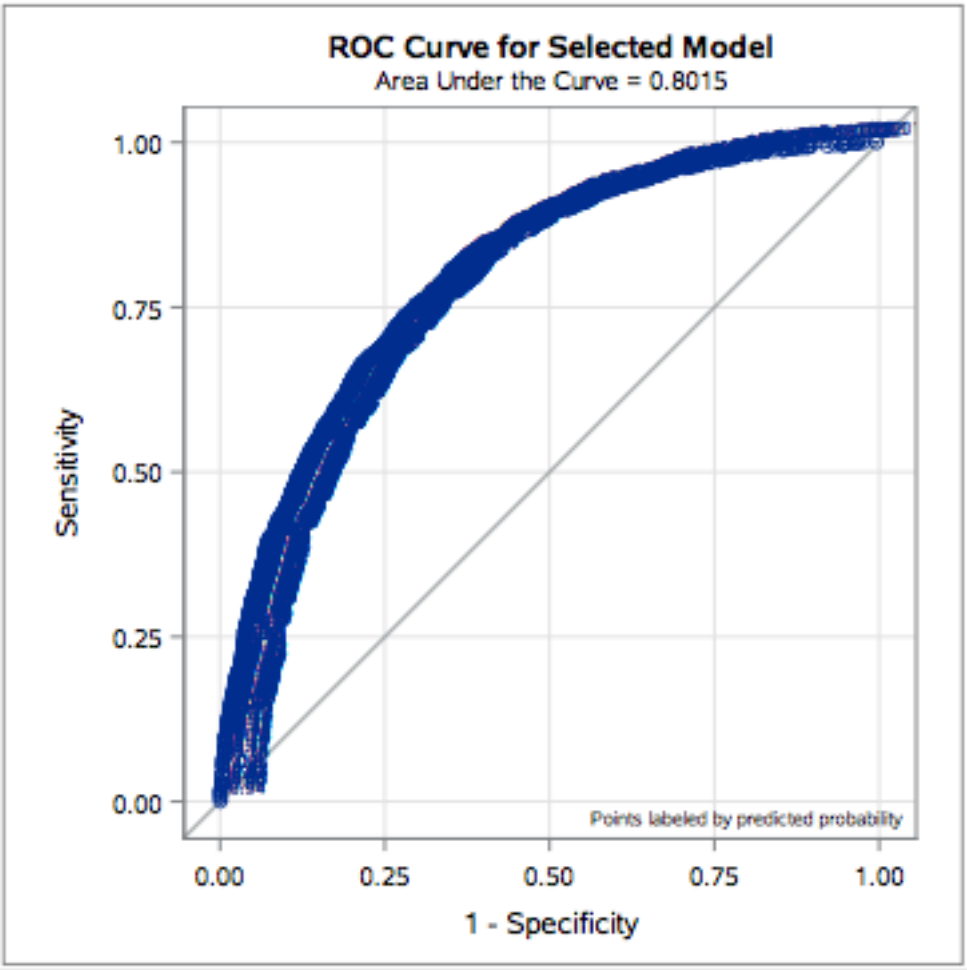

Figure 7: ROC Curve: Stepwise Selection

Logistic Regression Summary 3: Manual Selection

Although it was initially hypothesized that we could improve the accuracy of our results with a manual selection procedure, this approach proved to yield the results with the lowest level of accruacy. This technique used a comparable 34 variables, however, the Likelihood Ratio and ROC Score were considerably lower than what we were able to attain with the Stepwise and Backward technques. These results are shown below.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	8193.794
SC	9426.969	8432.036
-2 Log L	9417.962	8125.794

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1292.1685	33	<.0001
Score	1234.3076	33	<.0001
Wald	1003.1875	33	<.0001

Figure 8: Manual Selection Fit Statistics

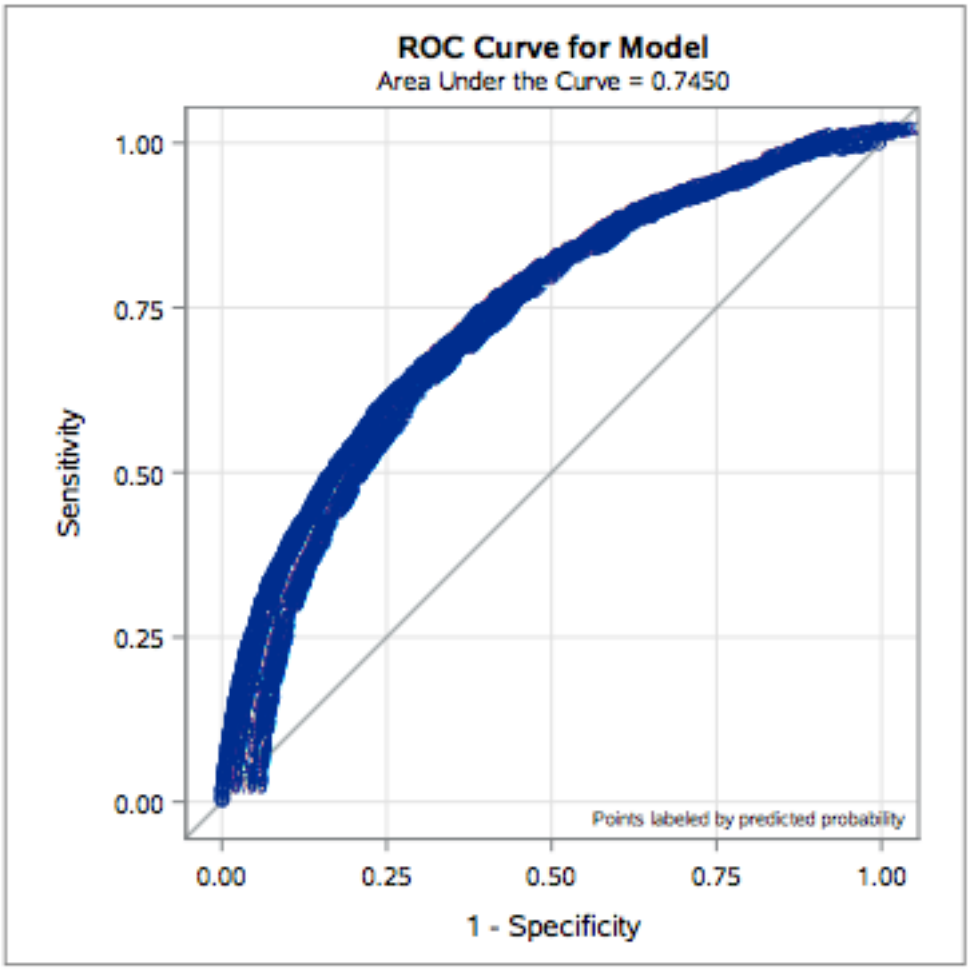

Figure 9: ROC Curve: Manual Selection

Final Predictive Model Formula

Given the results discussed above, the selection procedure yielding the results with the highest level of predictive accruacy were produced by the Stepwise procedure. The formula for this model is listed below and will be used to score our predicted results versus the actual.

YHAT = 0.8143
- (AGE_GROUP in("0"))*0.8887*
*+ (BLUEBOOK_GROUP in("0"))*0.4224
+ (BLUEBOOK_GROUP in("1"))*0.1804*
*+ (BLUEBOOK_GROUP in("2"))*0.4121
- (HOMEOWN in("0"))*0.3667*
*- (INCOME_GROUP in("0"))*0.4458
+ (INCOME_GROUP in("1"))*0.1862*
*+ (INCOME_GROUP in("2"))*0.0233
+ (KIDS_HOME in("0"))*0.4954*
*- (OLD_CLAIM_GROUP in("0"))*0.5326
- (POINTS in("0"))*0.3220*
*+ (JOB in("Clerical"))*0.4484
+ (JOB in("Home Maker"))*0.4799*
*+ (JOB in("Lawyer"))*0.4102
- (JOB in("Manager"))*0.3757*
*+ (JOB in("Professional"))*0.3268
+ (JOB in("Student"))*0.2426*
*+ (JOB in("z_Blue Collar"))*0.3856
+ (EDUCATION in("<High School"))*0.4873*
*+ (EDUCATION in("Bachelors"))*0.0850
+ (EDUCATION in("Masters"))*0.0357*
*+ (EDUCATION in("z_High School"))*0.4685
+ (CAR_TYPE in ("Panel Truck"))*0.4473*
*+ (CAR_TYPE in ("Pickup"))*0.5175
+ (CAR_TYPE in ("Sports Car"))*0.9403*
*+ (CAR_TYPE in ("Van"))*0.5519
+ (CAR_TYPE in ("z_SUV"))*0.6990*
*- (CAR_USE in ("Private"))*0.7817
- (URBANICITY in ("0"))*2.1804*
*- (REVOKED in ("0"))*0.7335
- (MSTATUS in ("0"))*0.5459

APPENDIX

Appendix Figure A.1

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
INDEX		8161	0	1.000	5151.868	5133.000	10302.000	2978.894
TARGET_FLAG		8161	0	0.000	0.264	0.000	1.000	0.441
KIDSDRIV	#Driving Children	8161	0	0.000	0.171	0.000	4.000	0.512
HOMEKIDS	#Children @Home	8161	0	0.000	0.721	0.000	5.000	1.116
BLUEBOOK	Value of Vehicle	8161	0	1500.000	15709.900	14440.000	69740.000	8419.734
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	0.000	4037.076	0.000	57037.000	8777.139
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0.000	0.799	0.000	5.000	1.158
MVR_PTS	Motor Vehicle Record Points	8161	0	0.000	1.696	1.000	13.000	2.147
IMP_AGE		8161	0	16.000	44.790	45.000	81.000	8.624
IMP_CAR_AGE		8161	0	0.000	8.309	8.000	28.000	5.519
IMP_HOMEVAL		8161	0	0.000	155225.067	161160.000	885282.345	125407.351
IMP_INCOME		8161	0	0.000	61468.960	54028.000	367030.262	46291.838
IMP_YOJ		8161	0	0.000	10.527	11.000	23.000	3.979
HOMEOWN		8161	0	0.000	0.281	0.000	1.000	0.450
AGE_GROUP		8161	0	0.000	0.012	0.000	1.000	0.108
BLUEBOOK_GROUP		8161	0	1.000	2.501	3.000	4.000	1.118
CAR_AGE_GROUP		8161	0	0.000	0.000	0.000	1.000	0.019
INCOME_GROUP		8161	0	1.000	2.446	2.000	4.000	1.117
OLDCLAIM_GROUP		8161	0	0.000	0.386	0.000	1.000	0.487
COMMUTE		8161	0	1.000	2.527	3.000	4.000	1.105
POINTS		8161	0	0.000	0.949	1.000	2.000	0.925
DRIVE_KIDS		8161	0	0.000	0.120	0.000	1.000	0.325
KIDS_HOME		8161	0	0.000	0.352	0.000	1.000	0.478

Figure 10: Predictor Variable Summary Statistics

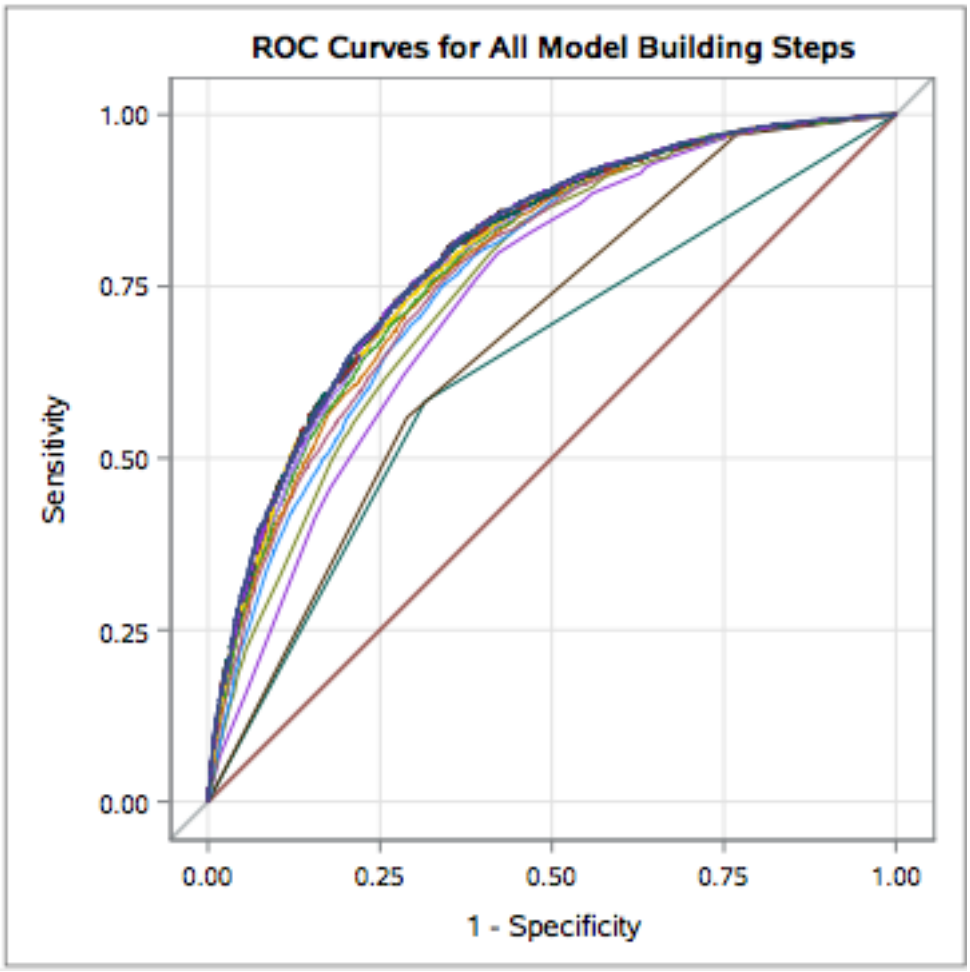

Figure 11: ROC Curve for All Model Building Steps: Stepwise Selection

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4055	0.3827	13.4890	0.0002
AGE_GROUP	0	1	-0.9476	0.2370	15.9882	<.0001
BLUEBOOK_GROUP	0	1	0.1811	0.0968	3.4974	0.0615
BLUEBOOK_GROUP	1	1	0.4075	0.0990	16.9373	<.0001
BLUEBOOK_GROUP	2	1	0.2007	0.0983	4.1718	0.0411
HOMEOWN	0	1	0.5834	0.1892	9.5086	0.0020
INCOME_GROUP	0	1	-0.4168	0.0947	19.3721	<.0001
INCOME_GROUP	1	1	0.1736	0.1127	2.3734	0.1234
INCOME_GROUP	2	1	0.0306	0.0841	0.1328	0.7155
KIDS_HOME	0	1	0.3495	0.0675	26.8030	<.0001
OLDCLAIM_GROUP	0	1	-0.7948	0.0583	185.5396	<.0001
POINTS	0	1	-0.3578	0.0598	35.7832	<.0001
JOB	Clerical	1	0.2673	0.2616	1.0443	0.3068
JOB	Home Maker	1	0.2744	0.2688	1.0421	0.3073
JOB	Lawyer	1	0.3835	0.2654	2.0871	0.1486
JOB	Manager	1	-0.3004	0.2590	1.3448	0.2462
JOB	Professional	1	0.2921	0.2605	1.2570	0.2622
JOB	Student	1	0.2191	0.2880	0.5790	0.4467
JOB	z_Blue Collar	1	0.3532	0.2542	1.9312	0.1646
EDUCATION	<High School	1	0.3397	0.1608	4.4616	0.0347
EDUCATION	Bachelors	1	0.0292	0.1459	0.0401	0.8413
EDUCATION	Masters	1	0.0508	0.1474	0.1190	0.7301
EDUCATION	z_High School	1	0.3558	0.1485	5.7405	0.0166
CAR_TYPE	Panel Truck	1	0.4909	0.1454	11.4021	0.0007
CAR_TYPE	Pickup	1	0.5169	0.0983	27.6667	<.0001
CAR_TYPE	Sports Car	1	0.8553	0.1042	67.3605	<.0001
CAR_TYPE	Van	1	0.5287	0.1197	19.5028	<.0001
CAR_TYPE	z_SUV	1	0.6719	0.0836	64.5940	<.0001
CAR_USE	0	1	0.6898	0.0958	51.7950	<.0001
CAR_USE	1	0	0	.	.	.
CAR_USE	Private	1	-0.6601	0.1201	30.1902	<.0001
PARENT1	0	1	-0.3575	0.1318	7.3546	0.0067

Figure 12: Analysis of Maximum Likelihood Estimates: Backward Selection

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.8143	0.3578	5.1810	0.0228
AGE_GROUP	0	1	-0.8887	0.2422	13.4623	0.0002
BLUEBOOK_GROUP	0	1	0.1804	0.0988	3.3298	0.0680
BLUEBOOK_GROUP	1	1	0.4121	0.1014	16.5159	<.0001
BLUEBOOK_GROUP	2	1	0.2176	0.1008	4.6637	0.0308
HOMEOWN	0	1	-0.3667	0.0822	19.8798	<.0001
INCOME_GROUP	0	1	-0.4458	0.0962	21.4886	<.0001
INCOME_GROUP	1	1	0.1862	0.1170	2.5346	0.1114
INCOME_GROUP	2	1	0.0233	0.0865	0.0728	0.7873
KIDS_HOME	0	1	0.4954	0.0607	66.7240	<.0001
OLDCLAIM_GROUP	0	1	-0.5326	0.0606	77.2230	<.0001
POINTS	0	1	-0.3220	0.0617	27.2666	<.0001
JOB	Clerical	1	0.4484	0.2616	2.9373	0.0866
JOB	Home Maker	1	0.4799	0.2701	3.1562	0.0756
JOB	Lawyer	1	0.4102	0.2636	2.4218	0.1197
JOB	Manager	1	-0.3757	0.2575	2.1280	0.1446
JOB	Professional	1	0.3268	0.2595	1.5854	0.2080
JOB	Student	1	0.2426	0.2809	0.7458	0.3878
JOB	z_Blue Collar	1	0.3856	0.2532	2.3184	0.1279
EDUCATION	<High School	1	0.4873	0.1633	8.9006	0.0029
EDUCATION	Bachelors	1	0.0850	0.1470	0.3348	0.5629
EDUCATION	Masters	1	0.0357	0.1474	0.0587	0.8086
EDUCATION	z_High School	1	0.4685	0.1500	9.7621	0.0018
CAR_TYPE	Panel Truck	1	0.4473	0.1486	9.0629	0.0026
CAR_TYPE	Pickup	1	0.5175	0.1007	26.4252	<.0001
CAR_TYPE	Sports Car	1	0.9403	0.1072	76.9779	<.0001
CAR_TYPE	Van	1	0.5519	0.1223	20.3467	<.0001
CAR_TYPE	z_SUV	1	0.6990	0.0853	67.1130	<.0001
CAR_USE	Private	1	-0.7817	0.0900	75.3889	<.0001
URBANICITY	0	1	-2.1804	0.1098	394.0526	<.0001
REVOKED	0	1	-0.7335	0.0792	85.7332	<.0001
MSTATUS	0	1	-0.5459	0.0737	54.9174	<.0001

Figure 13: Analysis of Maximum Likelihood Estimates: Stepwise Selection

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3372	0.3989	11.2381	0.0008
AGE_GROUP	0	1	-1.0893	0.2384	20.8777	<.0001
BLUEBOOK_GROUP	0	1	0.1813	0.0959	3.5761	0.0586
BLUEBOOK_GROUP	1	1	0.4115	0.0990	17.2881	<.0001
BLUEBOOK_GROUP	2	1	0.2038	0.0979	4.3387	0.0373
HOMEOWN	0	1	0.6145	0.2176	7.9726	0.0047
INCOME_GROUP	0	1	-0.4595	0.0934	24.2225	<.0001
INCOME_GROUP	1	1	0.1669	0.1110	2.2618	0.1326
INCOME_GROUP	2	1	0.0312	0.0828	0.1423	0.7060
POINTS	0	1	-0.5915	0.0566	109.1792	<.0001
JOB	Clerical	1	0.3174	0.2576	1.5182	0.2179
JOB	Home Maker	1	0.2698	0.2647	1.0388	0.3081
JOB	Lawyer	1	0.3667	0.2614	1.9677	0.1607
JOB	Manager	1	-0.2754	0.2549	1.1672	0.2800
JOB	Professional	1	0.2957	0.2565	1.3288	0.2490
JOB	Student	1	0.2465	0.2838	0.7547	0.3850
JOB	z_Blue Collar	1	0.3707	0.2503	2.1935	0.1386
EDUCATION	<High School	1	0.3597	0.1582	5.1688	0.0230
EDUCATION	Bachelors	1	0.0412	0.1439	0.0819	0.7747
EDUCATION	Masters	1	0.0603	0.1453	0.1724	0.6780
EDUCATION	z_High School	1	0.3621	0.1464	6.1192	0.0134
CAR_TYPE	Panel Truck	1	0.4992	0.1441	12.0009	0.0005
CAR_TYPE	Pickup	1	0.5141	0.0971	28.0379	<.0001
CAR_TYPE	Sports Car	1	0.9447	0.1061	79.2755	<.0001
CAR_TYPE	Van	1	0.5414	0.1186	20.8395	<.0001
CAR_TYPE	z_SUV	1	0.7589	0.0870	76.1754	<.0001
CAR_USE	0	1	0.7363	0.0951	59.9093	<.0001
CAR_USE	1	0	0	.	.	.
CAR_USE	Private	1	-0.6721	0.1206	31.0880	<.0001
SEX	0	1	0.2403	0.1494	2.5885	0.1076
PARENT1	0	1	-0.6849	0.1136	36.3807	<.0001
URBANICITY	0	1	-2.2136	0.1577	196.9598	<.0001
REVOKED	0	1	-0.7867	0.1338	34.5823	<.0001

Figure 14: Analysis of Maximum Likelihood Estimates: Manual Selection