Insurance Logistic Regression

Thomas Payne

Northwestern University

Predict 411 Section 59

Dr. Melvin Ott

May 8, 2016

Introduction

The business of insuring drivers can be a tricky one. You need to be able to answer questions like how much of a risk are we taking to insure this driver? How much cash reserve will I need in a fiscal year to ensure there is enough funding to pay claims? One tactic would be to insure as many people as possible and hope that the income received from the good drivers outweighs the cost of insuring the bad drivers, but that is not a sound business practice. This paper will attempt to use the insurance dataset to determine: first, what the probability is of a customer getting into a car accident, and second, what would the amount of an accident cost the insurance company. This approach is commonly referred to as the probability/severity model and requires the development of two separate models: a logistic model to determine the probability of a customer getting into a car crash and a model using linear regression to determine the estimated cost of a claim. The main focus of this paper will be on logistic regression and the use of maximum likelihood to estimation using the logit function. Models found in the paper will be selected using a stepwise procedure.

Data Exploration

Overview of Data

The insurance data set consists of biographical information from 8161 customers. Each customer is represented by an anonymized index number and a row of data with 24 predictor variables:

Predictor Variable			
Variable Name	**Variable Type**	**Label**	**Theoretical Effects**
CAR_TYPE	Categorical	Type of Car	Likely related to cost of repairs
EDUCATION	Categorical	Education Level	More educated people are safer drivers
JOB	Categorical	Job Category	White Collar Workers are safer drivers
CAR_USE	Categorical (dichotomous)	Vehicle Use	Commercial vehicles more likely to be in crash
MSTATUS	Categorical (dichotomous)	Marital Status	Married people are safer drivers
PARENT1	Categorical (dichotomous)	Single Parent	Single parents drive more often and more likely to get into classes
RED_CAR	Categorical (dichotomous)	A Red Car	Red cars get into more accidents
REVOKED	Categorical (dichotomous)	License Revoked in the Past 7 Years	Drivers who have had their licenses revoked are riskier drivers
SEX	Categorical (dichotomous)	Gender	Women are less likely to get into a crash
URBANICITY	Categorical (dichotomous)	Home/Work Area	Urban drivers are more likely to get into a crash
BLUEBOOK	Continuous	Bluebook Value of Vehicle	Likely related to cost of repairs
HOME_VAL	Continuous	Home Value	Home owners are safer drivers

Predictor Variable			
Variable Name	**Variable Type**	**Label**	**Theoretical Effects**
INCOME	Continuous	Income	Rich people tend to get in fewer crashes
OLDCLAIM	Continuous	Cost of Total Claims Made in the Past 5 Years	Higher claims are likely to result in higher future claims
TRAVTIME	Continuous	Distance to Work	Long commutes lead to greater risk of getting into a crash
AGE	Discrete	Age of Driver	Old and young drives are more likely to get in a crash
CAR_AGE	Discrete	Vehicle Age	Likely related to cost of repairs
CLM_FREQ	Discrete	Number of Claims in the Past 5 Years	People with more claims are more likely to get in accidents
HOMEKIDS	Discrete	Number of Children	Parents are safer drivers
KIDSDRIV	Discrete	Number of Children Who Drive	Teenage drivers who borrow a car are more likely to get in a crash.
MVR_PTS	Discrete	Motor Vehicle Record Points	Drivers with more points get into more crashes
TIF	Discrete	Years as a Customer	Long term customers are safer drivers
YOJ	Discrete	Years on Job	More years the job mean a safer driver

As is depicted above, the predictor variables measure: a customer's driving record, previous claims, income, education history, family history, and customer history through 10 categorical variables (3 are dichotomous), 5 continuous, and 8 discrete variables. The rest of the exploratory data analysis section will be used to assess the quality of the data and its ability to predict the likelihood of a customer getting into an accident and the cost of such an accident. It will also be used for an initial exploration of relationships between predictor variables and response variables.

Data Quality Check

Target Variables

Given the data provided, there are some inherent limitations. For example, there is not a variable given for the cause of car crash. The absence of such a variable could make low-risk customers (those unlikely to get into an accident) falsely appear to be high-risk customers (a customer with a strong likelihood of getting in an accident). Therefore, it is reasonable to assume that there would be a number of false positives in any model developed from the insurance dataset. Despite the fact that these occurrences should be infrequent, they could occur in values that lie far outside the pattern of the data. This is more of a concern for the continuous and discrete variables than for the categorical variables. The odds and probabilities should pull the values towards the more frequent occurrences. Having provided that disclaimer we can now begin the process of determining the quality of the data.

We begin investigating our two response variables, Target_Amt and Target_Flag. Target_Flag is a dichotomous variable with two categories to represent if a customer was in a crash. When the Target-Flag variable has a value of "0" this means that the customer has not been in a crash and "1" represents

customers who have been in an accident. Figure 1 shows the number of customers in the insurance dataset who have been in an accident, and also shows those who have not.

Figure 1: Target_Flag Frequency

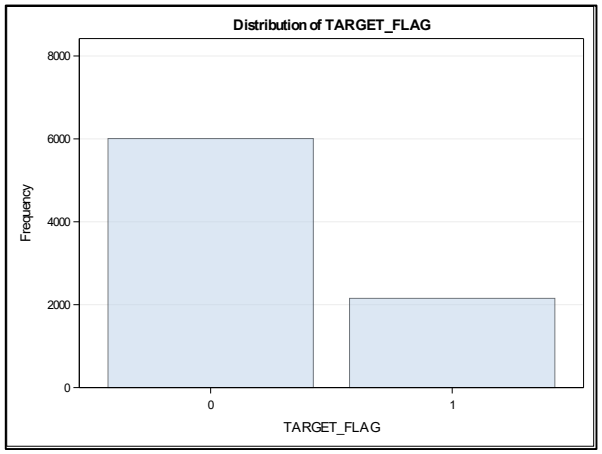


Table one provides detail on the prevalence of customers who were in a crash. Not surprisingly, getting into a car crash is a rare event -customers were three times more likely not to get into a car accident. Of the 8161 customers in the data set only 26% were involved in an accident.

Table 1: Target_Flag Frequency

TARGET_FLAG	Frequency	Percent	Odds
0	6008	73.62%	2.791
1	2153	26.38%	.358

Given the dissolution, the use of dummy variables, and the fact that the goal is to determine the probability of a customer being in a crash, it would be inappropriate, or at least inadvisable, to use ordinary least squares regression to build a model. It makes more sense to use logistic regression to model probabilities. One of the benefits of using logistic regression is that we are not bound by the need for parameters to enter the model linearly, or specific distribution of the error term. For logistic regression to be successful each observation needs to be statistically independent, the response variable has to have a value from zero to one, and there need to be enough observations in each category to have predictive power. Target_Flag meets all of these criteria.

The other predictor variable Target-Amt is a continuous target with a high frequency of records at zero, the effects of which are harmful in ordinary least square regression. Figure 2 shows just how harmful they can be to the distribution of the claim amounts. In the box chart, the effects of the preponderance of zero create a large number of outliers. The histogram in Figure 2 shows how those outliers highly skewed distributing.

Figure 2: Target_Amt Distribution

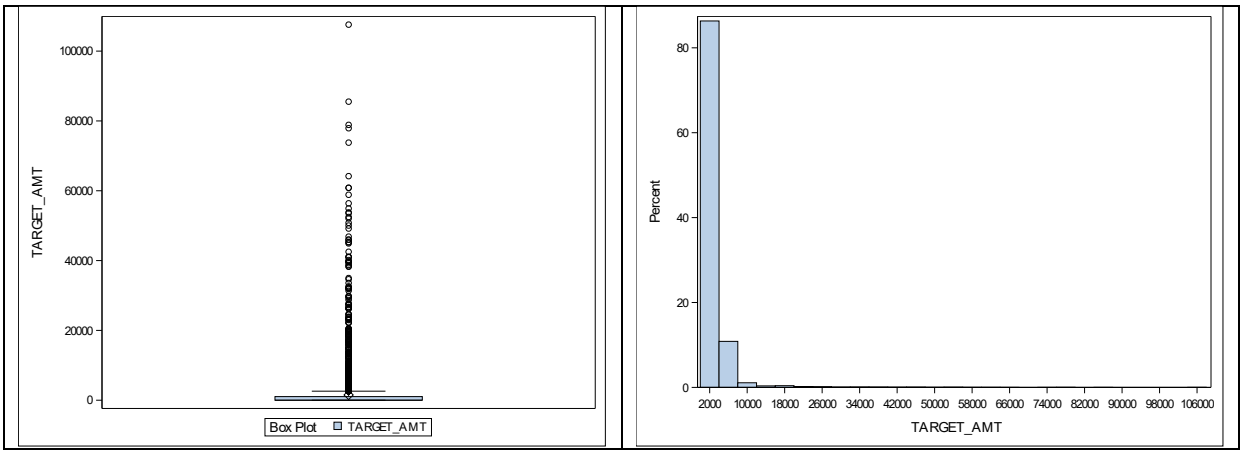
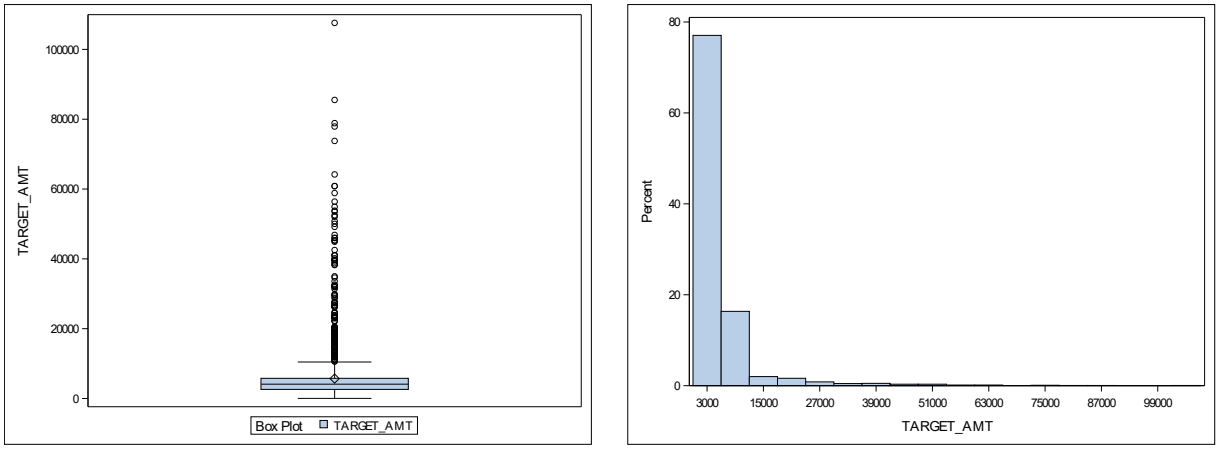


Figure 2 makes a strong case for the need to use the probability/severity modeling process. In doing so, we will first find the probability of a customer getting into a crash and will only predict the claim amount for customers who are going to be in a crash because we know the value will be zero for those who are not in a crash. Unfortunately, even after removing the zero values, Figure 3 shows that the distribution is still highly skewed with a large number of outliers. This could be a problem in ordinary least squares regression as we deal with concerns over the distribution, variance, and covariance of the error term, and the way the parameters of the independent variables enter the equation. A transformation will likely be needed during data preparation.

Figure 3: Target_Amt with Zeros Removed.



Predictor Variables

Another concern that we will need to address in data preparation is missing values. Regardless of the modeling methodology employed, missing values can be troublesome. Table two highlights the variables with missing values.

Table 2: Missing Values

Variable	Label	N Miss	Minimum

AGE	Age	6	0.07%
YOJ	Years on Job	454	5.56%
INCOME	Income	445	5.45%
HOME_VAL	Home Value	464	5.69%
CAR_AGE	Vehicle Age	510	6.25%
JOB	Job Category	526	6.45%

Of the six variables missing values, Job, and CAR_Age have the most missing values. Age is only missing six. The number of missing values is very manageable and something that will be taken care of during data preparation.

We now move onto exploring the categorical variables, starting first with the binary variables in table three:

Table 3: Binary Variables Probabilities and Odds Ratios

Variable	Category	Probability	Odds Ratio
CAR_USE	Commercial	35%	1.923
	Private	22%	0.520
MSTATUS	Yes	22%	0.540
	z_No	34%	1.852
PARENT1	Yes	44%	2.551
	No	24%	0.392
RED_CAR	yes	26%	0.966
	no	27%	1.035
REVOKED	Yes	44%	2.538
	No	24%	0.394
SEX	M	25%	0.908
	z_F	27%	1.101
URBANICITY	Highly Urban/ Urban	31%	6.183
	z_Highly Rural/ Rural	7%	0.162

In table three we can see that each of the variable has some predictive ability associated with it. Highlighted in yellow are the variables with the largest odds ratios, and predictive probabilities. For instance, a person living in an urban area is six times more likely to get into a crash than a person in a rural area. Because our primary objective is to determine the probability of a customer being in a crash we are not concerned with distribution and can concentrate on measures of prediction.

Next, we turn to the categorical variables with more than two categories to look at their predicative ability. In Table four we see variables Job, Car_Type and Education impact the probability of getting into an accident. We are able to reject the null hypothesis for each variable and conclude that when combined each of the coefficient that comprise an individual variable are statistically different from zero. The high Wald Chi-Scores also found in Table four leads us to the assumption that there will be

differences around the statistical effectiveness of the classes of each of these variables. Luckily, there is no convergence and less predictive classes can be removed during the variable selection process.

Table 4: Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
JOB	7	248.6530	<.0001
CAR_TYPE	5	164.9946	<.0001
EDUCATION	4	167.0529	<.0001

The last variable types in our hodgepodge of data variables are continuous and discrete variables. Table five outlines key descriptive statistic for the continuous and discrete variables – mean, median, minimum value, maximum values, the standard error the standard deviation.

Table 5: Descriptive Statistics for

Predictor Variable	Description	Mean	Median	Max	Min	Std Error	Std Dev
KIDSDRIV	#Driving Children	0	0	4	0	0	1
AGE	Age	45	45	81	16	0	9
HOMEKIDS	#Children @Home	1	0	5	0	0	1
YOJ	Years on Job	10	11	23	0	0	4
INCOME	Income	61898	54028	367030	0	542	47573
HOME_VAL	Home Value	154867	161160	885282	0	1472	129124
TRAVTIME	Distance to Work	33	33	142	5	0	16
BLUEBOOK	Value of Vehicle	15710	14440	69740	1500	93	8420
TIF	Time in Force	5	4	25	1	0	4
OLDCLAIM	Total Claims(Past 5 Years)	4037	0	57037	0	97	8777
CLM_FREQ	#Claims(Past 5 Years)	1	0	5	0	0	1
MVR_PTS	Motor Vehicle Record Points	2	1	13	0	0	2
CAR_AGE	Vehicle Age	8	8	28	-3	0	6

Notice that the variables highlighted in green have large standard errors and standard deviations, which reveal a large degree of dispersion between, and within, the data. Additionally, the minimum value for Car_Age is negative and likely a data entry mistake that will be addressed in data preparations. Otherwise, the continuous and discrete variables look fine and we can look to see if they have any predictive ability.

The predictive ability of continuous variables is not as strong as the categorical variables. Table six includes all of these variables, their odds ratios, predictive probabilities, and r-squared values. The r-squared values in logistic regression do not account for variance, but help determine goodness-of-fit.

The higher the value the better, so the low values in Table 6 describe models that individually do not fit the data well.

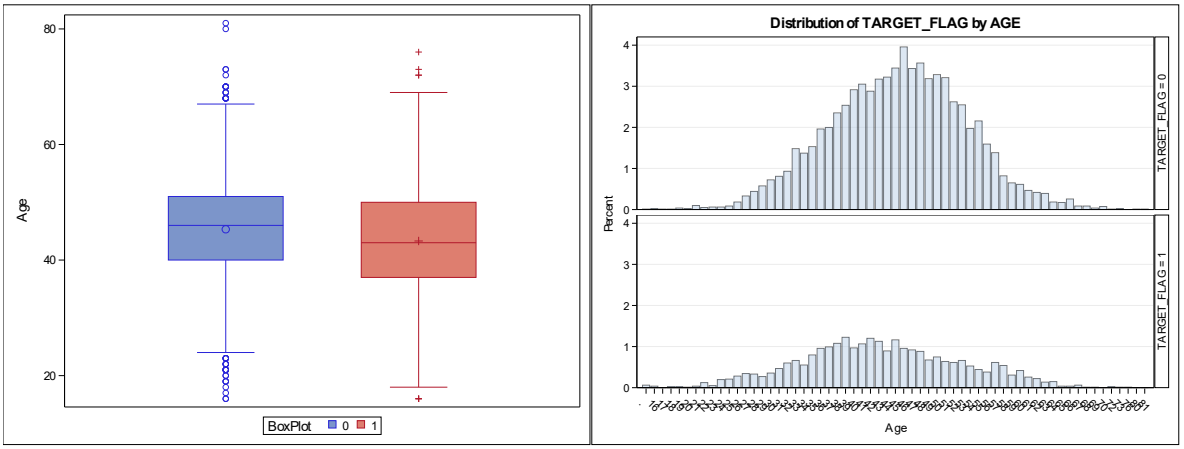
Table 6: Predictive Measures for Continuous and Discrete Variables

Variables	R^2	Max-rescaled R^2	Odds Ratio	Predictive Probability
TRAVTIME	0.0023	0.0034	1.007	22.2%
YOJ	0.0048	0.0071	0.963	33.7%
TIF	0.0070	0.0102	0.954	30.4%
KIDSDRIV	0.0098	0.0143	1.504	33.2%
CAR_AGE	0.0102	0.0150	0.960	32.3%
AGE	0.0106	0.0155	0.973	54.0%
BLUEBOOK	0.0110	0.0161	1.000	36.0%
HOMEKIDS	0.0127	0.0186	1.248	27.3%
OLDCLAIM	0.0171	0.0249	1.000	23.7%
INCOME	0.0216	0.0315	1.000	36.0%
HOME_VAL	0.0349	0.0509	1.000	37.1%
CLM_FREQ	0.0431	0.0629	1.482	27.1%
MVR_PTS	0.0441	0.0645	1.240	22.8%

Three variables of further interest in Table 6 include Age, Clm_Freq, and Mvr_Pts. These variables have the largest predictive probability, and the two largest odds ratios. They present an opportunity to look for ways to strengthen them and improve the productive ability of the model.

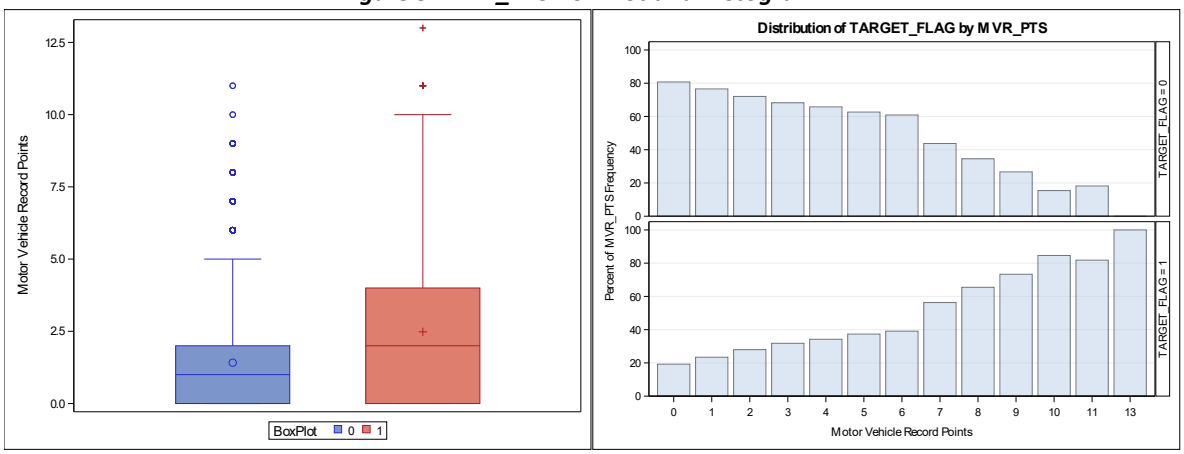
Not surprisingly, Age, has the highest predictive probability. There is a long held belief that young drivers are higher risks than older drivers. The box plot in Figure 4 is misleading in that it looks like it does not support this belief - the spread of the data in the box plot for customers who were in a crash and those who were not is very similar. However, the histogram of age shows that there are far more middle-aged people that do not get into accidents than young and old people. This is a factor that should be considered in data preparation.

Figure 4: Age Box Plot and Histograms



The next variable, MVR_PTS, has the largest r-squared and adjusted r-square values suggesting that the number of points on a driver's record can help to predict the probability of that person being in an accident. One confounding factor with this variable is the point system for each state. In most states points acquired in a short time span results in suspension, clearly it would be impossible to get into and accident if you are not driving. Additionally, the points could be the result of a crash and prevent any forward looking predictions.

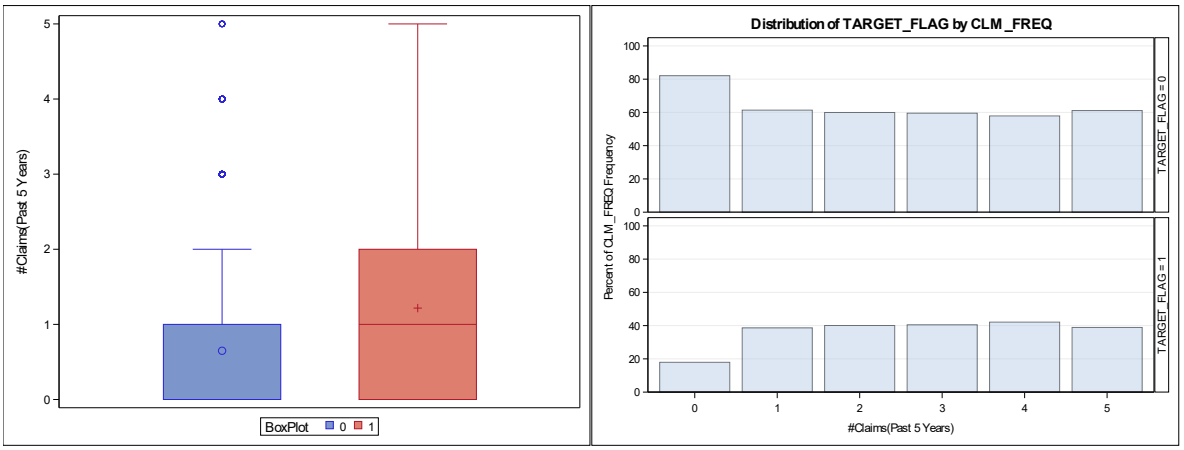
Figure 5: MVR_PTS Box Plot and Histogram



The histogram in Figure 5 quickly dispels any concerns for confounding effect. There is a positive correlation between the number of points on a person's driving record and the likelihood that they will get into an accident. The charts in Figure 5 notes that 7 points is the turning point- customers with 7 points or more tend to get into accidents.

The third variable selected for examination is Clm_Freq. Clm_Freq which represents the number of claims filed in the last five years. This variable has the highest odds ratio indicating that as the number of claims increase so does the probability of getting into a car crash. The histogram in Figure 6 is relatively flat after two. The box plot in the same figure has the cut off for the fourth quartile there as well. This variable may more useful as a binary categorical variable, something that again will be addressed in data preparation.

Figure 6: CLM_FREQ



The variables look usable, and aside from a few alternations in data preparation, we can move forward. Before we move forward, there is one last concern: multicollinearity and how target-amt relates to the discrete and continuous variables.

Multicollinearity in logistic regression can render estimates of coefficients unusable, based on an analysis of the correlations between variables. Surprisingly, just two continuous or discrete variables have Pearson 's R-values close to, or above, .5. Captured in Table 7, we see income and home values are correlated, as well as the amount paid to claims in the last five years, and the number of claims made in the last five years. It is reasonable to assume that if you have a larger income you would spend more on your home, and if you filed a lot of claims you would have received more money.

Table 7: Correlated Variables

Predictor Variable	Highest Correlation
HOME_VAL Home Value	INCOME 0.57524 <.0001
OLDCLAIM Total Claims(Past 5 Years)	CLM_FREQ 0.49513 <.0001

As mentioned before, the linear regression segment of this paper is not of great importance due partly to the inability of the variables in the database to explain variance. Table 8 shows that the R values for each of the continuous and discrete variables are extremely low. This means that the r-squared values would be even lower. A few transformations will be performed in data preparation, though they will likely have no drastic improvement on the model.

Table 8: TARGET_AMT Correlations

Predictor Variable	R Value with TARGET_AMT
BLUEBOOK	0.11809
INCOME	0.04547
MVR_PTS	0.03981
YOJ	0.03426
HOME_VAL	0.02896
AGE	0.02788
TRAVTIME	0.00494
CLM_FREQ	0.00196
HOMEKIDS	0.00047
KIDSDRIV	0.00002
OLDCLAIM	-0.00566
TIF	-0.00601
CAR_AGE	-0.01302

Data Preparation

Data exploration presents no reason not to move forward. The variables look as though they should have good predictive ability. A few notes remain from data exploration, such as negative value, new variables, missing values and transformation, and the easiest way to deal with the negative value sin the CAR_AGE variable. There are several options when dealing with the negative values, and the preferred method in this case is removal. Clearly any negative value was made in error and only one such variable (-3) existed. It can easily be removed without decreasing the predictive ability of our model. The methods for dealing with incorrect values are similar to those used in deal with missing values.

This time we can simply remove the observations without impacting the adequacy of the model. For that reason, missing values from the six variables: AGE, YOJ, INCOME, HOME_VAL, CAR_AGE, and JOB, will be replaced with the median value. The median was chosen because it is not influenced as easily by the extreme values that were discussed in the last section. After inputting the missing values, the last set of transformations performed is the creation of new variables.

The new variables are meant to increase the predictive ability of the model. The first set is transformations of the age variable.

AGE_2 = Age*Age
AGE_3 = Age*Age*Age
AGE_4= Age*Age* Age*Age

Age_2, Age_3 and Age_4 respectively represent age squared, age cubed, and age to the fourth power. The purpose of this transformation is to account for a polynomial distribution of age. The next set of variables involves transformations of continuous and discrete variables into binary categorical variables.

Table 9: New Binary Variables

Variable	Category	Probability	Odds Ratio
BIN_HOMEKIDS	Has Kids	22.2%	0.550
	No Kids	34.1%	1.819
BIN_OLDCLAIM	More than $4,637	22.1%	0.440
	Less than $4,637	39.2%	2.273
BIN_ MVR_PTS	Less than 6 points	24.9%	0.191
	More than 6 points	63.4%	5.227
BIN_CLM_FREQ	More than 1 claim	39.8%	3.026
	Less than 1 claim	17.9%	0.330

The new variables described in Table 9 have large odds rations and reasonable predictive probability. BIN_ MVR_PTS has the largest odds ratio and probability of any variable in the data set. The determination for how to categorize the variables in Table 9 was based off of shifts in the data. For

example, MVR_PTS showed that after 6 points customers were more likely to have been in a car crash. There was a similar difference in HOMEKIDS. People who had kids tended to get into more accidents, so a new variable was created to specify whether a customer had children. The last set of new variables was categorical variables that had more than two categories that individually had p-values about .05.

Table 10: New Dichotomous Variables

Variable	Category	Probability	Odds Ratio
BIN_JOB	Esq, Dr, or Mangr	15.4%	0.423
	Not	30.1%	2.362
BIN_EDUCATION	High school o	33.3%	1.876
	>High school	21.1%	0.533
BIN_CAR	SUV/Sports Car	30.7%	1.433
	Not	23.6%	0.698

Table 10 describes the new dichotomous variables and how they were combined. In the JOBS variables the classes of Lawyer, Doctor and Manager had low statistical significance. They were combined in an effort to have a variable that was significantly different from zero. Two smaller and less significant transformations were also done.

LOG_TARGET_AMT= log10(TARGET_AMT)
LN_BLUEBOOK = log(BLUEBOOK)

The transformations aim to make all the records equally reliable for the ordinary least squares regression that will be performed at the end.

Build Models

Using the transformed, and new, variables three models are constructed: a control model, a model with new variables, and a model that can use all the available variables. Each of the three models is determined using stepwise selection. The stepwise method works by successively adding or removing variables based on a partial test statistic being larger than the preselected test statistic value. The end result is the “best” model, but it should be noted that this is not necessarily the best overall model given the set of variables, and is not even a good model at that. The models will also use the corrected data where imputed values were used for missing observations. In the end, the following model was derived

Model 1

The first model is as follows, and was run using a cutoff and entry p-value of .01 and allowed for this model with 31 parameters.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3592	0.2073	42.9794	<.0001
CAR_TYPE	Minivan	1	-0.7132	0.0859	68.8897	<.0001
CAR_TYPE	Panel Truck	1	-0.1619	0.1500	1.1652	0.2804
CAR_TYPE	Pickup	1	-0.1883	0.0932	4.0758	0.0435
CAR_TYPE	Sports Car	1	0.2578	0.0979	6.9359	0.0084
CAR_TYPE	Van	1	-0.0953	0.1199	0.6319	0.4266
CAR_USE	Commercial	1	0.7801	0.0909	73.5783	<.0001
EDUCATION	<High School	1	-0.00297	0.0944	0.0010	0.9750
EDUCATION	Bachelors	1	-0.4163	0.0834	24.9242	<.0001
EDUCATION	Masters	1	-0.4597	0.1155	15.8558	<.0001
EDUCATION	PhD	1	-0.3579	0.1579	5.1413	0.0234
JOB	Clerical	1	0.1377	0.1049	1.7241	0.1892
JOB	Doctor	1	-0.5478	0.2595	4.4565	0.0348
JOB	Home Maker	1	0.0270	0.1395	0.0375	0.8465
JOB	Lawyer	1	-0.0265	0.1513	0.0306	0.8612
JOB	Manager	1	-0.7641	0.1227	38.8003	<.0001
JOB	Professional	1	-0.0809	0.1107	0.5340	0.4649
JOB	Student	1	-0.0121	0.1220	0.0098	0.9209
MSTATUS	Yes	1	-0.4689	0.0796	34.7409	<.0001
PARENT1	No	1	-0.4608	0.0942	23.9208	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3850	0.1129	446.6268	<.0001
REVOKED	No	1	-0.8913	0.0912	95.4113	<.0001
BLUEBOOK		1	-0.00002	4.717E-6	23.8592	<.0001
CLM_FREQ		1	0.1958	0.0285	47.1731	<.0001
HOME_VAL		1	-1.35E-6	3.407E-7	15.7156	<.0001
INCOME		1	-3.63E-6	1.073E-6	11.4204	0.0007
KIDSDRIV		1	0.4201	0.0551	58.1773	<.0001
MVR_PTS		1	0.1143	0.0136	70.8587	<.0001
OLDCLAIM		1	-0.00001	3.91E-6	13.2830	0.0003

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
TIF		1	-0.0555	0.00734	57.0904	<.0001
TRAVTIME		1	0.0145	0.00188	59.4415	<.0001

The final model is:

P_TARGET _FLAG = -0.7132*CAR_TYPE(Minivan) -0.1619*CAR_TYPE(Panel Truck) -0.1883*CAR_TYPE(Pickup) +0.2578*CAR_TYPE(Sports Car) -0.0953*CAR_TYPE(Van) +0.7801*CAR_USE(Commercial) -0.00297*EDUCATION(<High School) -0.4163*EDUCATION(Bachelors) -0.4597*EDUCATION(Masters) -0.3579*EDUCATION(PhD) + 0.1377*JOB(Clerical) -0.5478*JOB(Doctor) + 0.027*JOB(Home Maker) -0.0265*JOB(Lawyer) -0.7641*JOB(Manager) -0.0809*JOB(Professional) -0.0121*JOB(Student) -0.4689*MSTATUS(Yes) -0.4608*PARENT1(No) +2.385*URBANICITY(Highly Urban/ Urban) -0.8913*REVOKED(No) -0.00002*BLUEBOOK + 0.1958*CLM_FREQ -0.00000135*HOME_VAL -0.00000363*INCOME + 0.4201*KIDSDRIV +0.1143*MVR_PTS -0.00001*OLDCLAIM -0.0555*TIF + 0.0145*TRAVTIME

The model has a number of parameters that are not significantly different from zero (highlighted in yellow above). Additionally, the model also has some questionable coefficients. The model shows that the higher the amount paid to a customer in the past five years, the less likely they are to get into a crash. Unless the company has dropped a customer after a certain amount of money is collected this would not make sense. We see that the more claims filed by a customer in the last five years increases the odds of getting into an accident. Arguably those claims were associated with some request for funds making this outcome puzzling. The high number of dimensions, questionable coefficients, and low variables with low p-values make this model not ideal.

Model 2

Attempting to find a more parsimonious model the variables BIN_MVR_PTS, BIN_OLDCLAIM, BIN_HOMEKIDS, BIN_CLM_FREQ, BIN_JOB, BIN_EDUCATION, and BIN_CAR were used in place of the variables' representation of their names. The following model was created using .01 stepwise selection method:

Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.3633	0.4451	9.3787	0.0022
CAR_USE	**Commercial**	1	0.8825	0.0668	174.3902	<.0001
URBANICITY	**Highly Urban/ Urban**	1	2.3437	0.1115	442.0044	<.0001
MSTATUS	**Yes**	1	-0.6578	0.0699	88.6602	<.0001
AGE		1	-0.0935	0.0134	48.8997	<.0001
Age_3		1	0.000015	2.005E-6	55.6896	<.0001
BLUEBOOK		1	-0.00002	4.061E-6	26.9350	<.0001

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
HOME_VAL		1	-1.27E-6	3.263E-7	15.2296	<.0001
INCOME		1	-3.1E-6	9.367E-7	10.9697	0.0009
KIDSDRIV		1	0.4599	0.0596	59.6429	<.0001
TRAVTIME		1	0.0146	0.00185	61.6707	<.0001
BIN_MVR_PTS	More than 6 points	1	0.9177	0.1344	46.6551	<.0001
BIN_HOMEKIDS	Has Kids	1	0.2238	0.0790	8.0267	0.0046
BIN_CLM_FREQ	More than one claim	1	0.5235	0.0594	77.7012	<.0001

This model has 15 parameters each that are all significantly different form zero and there are no usual coefficients.

The final model is:

P_TARGET_FLAG = -1.3633*Intercept + 0.8825*CAR_USE(Commercial) + 2.3437*URBANICITY(Highly Urban/ Urban) -0.6578*MSTATUS(Yes) -0.0935*AGE + 0.000015*Age_3 -0.00002*BLUEBOOK - 0.00000127*HOME_VAL -0.0000031*INCOME+ 0.4599*KIDSDRIV +0.0146*TRAVTIME + 0.9177*BIN_MVR_PTS(More than 6 points) + 0.2238*BIN_HOMEKIDS(Has Kids) + 0.5235*BIN_CLM_FREQ(More than one claim) -0.7641*JOB(Manager)

Model 3

Our third model combines old and new variables to try and increase the predictive power of the model. The selection procedure was run using a higher .001 exit and entry cutoff p-value that allowed for a model with 26 parameters:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1451	0.2070	107.4282	<.0001
CAR_TYPE	Minivan	1	-0.6982	0.0855	66.7568	<.0001
CAR_TYPE	Panel Truck	1	-0.1399	0.1447	0.9349	0.3336
CAR_TYPE	Pickup	1	-0.1788	0.0912	3.8444	0.0499
CAR_TYPE	Sports Car	1	0.2642	0.0973	7.3754	0.0066
CAR_TYPE	Van	1	-0.0835	0.1177	0.5043	0.4776
CAR_USE	Commercial	1	0.7714	0.0859	80.6392	<.0001
JOB	Clerical	1	0.1172	0.1037	1.2792	0.2581
JOB	Doctor	1	-0.4530	0.2311	3.8428	0.0500
JOB	Home Maker	1	0.0297	0.1363	0.0474	0.8276

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
JOB	Lawyer	1	-0.0398	0.1330	0.0897	0.7645
JOB	Manager	1	-0.7709	0.1207	40.8218	<.0001
JOB	Professional	1	-0.0687	0.1060	0.4198	0.5170
JOB	Student	1	-0.00646	0.1212	0.0028	0.9575
MSTATUS	Yes	1	-0.4594	0.0793	33.5673	<.0001
PARENT1	No	1	-0.4617	0.0935	24.3772	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3465	0.1128	432.7157	<.0001
REVOKED	No	1	-0.7361	0.0798	85.0113	<.0001
BLUEBOOK		1	-0.00002	4.699E-6	24.1370	<.0001
HOME_VAL		1	-1.35E-6	3.394E-7	15.7499	<.0001
INCOME		1	-3.47E-6	1.023E-6	11.5099	0.0007
KIDSDRIV		1	0.4309	0.0543	62.9144	<.0001
TRAVTIME		1	0.0148	0.00187	63.1965	<.0001
BIN_MVR_PTS	More than 6 points	1	0.9620	0.1342	51.3719	<.0001
BIN_CLM_FREQ	More than one claim	1	0.5210	0.0599	75.5568	<.0001
BIN_EDUCAITON	High school or below	1	0.4169	0.0748	31.0557	<.0001

Though this model does not have any unusual coefficients it does have a number of individual variables that are not statistically significant. Highlighted in yellow are nine parameters from two variables JOB and CAR_TYPE. Simply removing the JOB and CAR_TYPE variable could improve the model. In the end the final model is:

P_Target_FLAG = -2.1451*Intercept -0.6982*CAR_TYPE(Minivan) -0.1399*CAR_TYPE(Panel Truck) -0.1788*CAR_TYPE(Pickup) + 0.2642*CAR_TYPE(Sports Car) -0.0835*CAR_TYPE(Van) + 0.7714*CAR_USE(Commercial) + 0.1172*JOB(Clerical) -0.453*JOB(Doctor) + 0.0297*JOB(Home Maker) -0.0398*JOB(Lawyer) -0.7709*JOB(Manager) -0.0687*JOB(Professional) -0.00646*JOB(Student) -0.4594*MSTATUS(Yes) -0.4617*PARENT1(No) + 2.3465*URBANICITY(Highly Urban/ Urban) -0.7361*REVOKED(No) -0.00002*BLUEBOOK -0.00000135*HOME_VAL -0.00000347*INCOME + 0.4309*KIDSDRIV + 0.0148*TRAVTIME+ 0.962*BIN_MVR_PTS(More than 6 points) + 0.521*BIN_CLM_FREQ(More than one claim) + 0.4169*BIN_EDUCAITON(High school or below)

Selecting Models

In determining which of the three models is best, it is important to not to rely on any one measure of goodness-of-fit. Logistics regression has a number of measures that evaluate the goodness such as the AIC, BIC, r-squared, -2 Log L, Kolmogorov-Smirnov (KS) statistic, the ROC cruve, and the area under the

curve (AUC). AIC is the Akaike Information Criterion and looks to maximize the expected entropy of a model. The BIC, or the Bayesian Information Criterion, assesses the overall fit of a model and allows the comparison of both nested and non-nested models. Free from concerns of multicollinearity, lower AIC, -2 Log L**,** and BIC values indicate that a model closer to the true relationship between predictor variables and the probability of getting into a car crash.

The r-squared value shows the goodness-of-fit even though it is no longer a measure of explained variability. We will be looking for the model with the largest r-squared value as well as the largest KS statistic and the AUC values. The KS value is the maximum distance is the maximum difference between the cumulative true positive rate and the cumulative false positive rate. This statistic helps discriminate true predictions from false positives. The AUC can be interpreted as the average ability of the rating model to accurately classify predictions from false positives. A higher AUC and KS denote higher degrees of predictive accuracy.

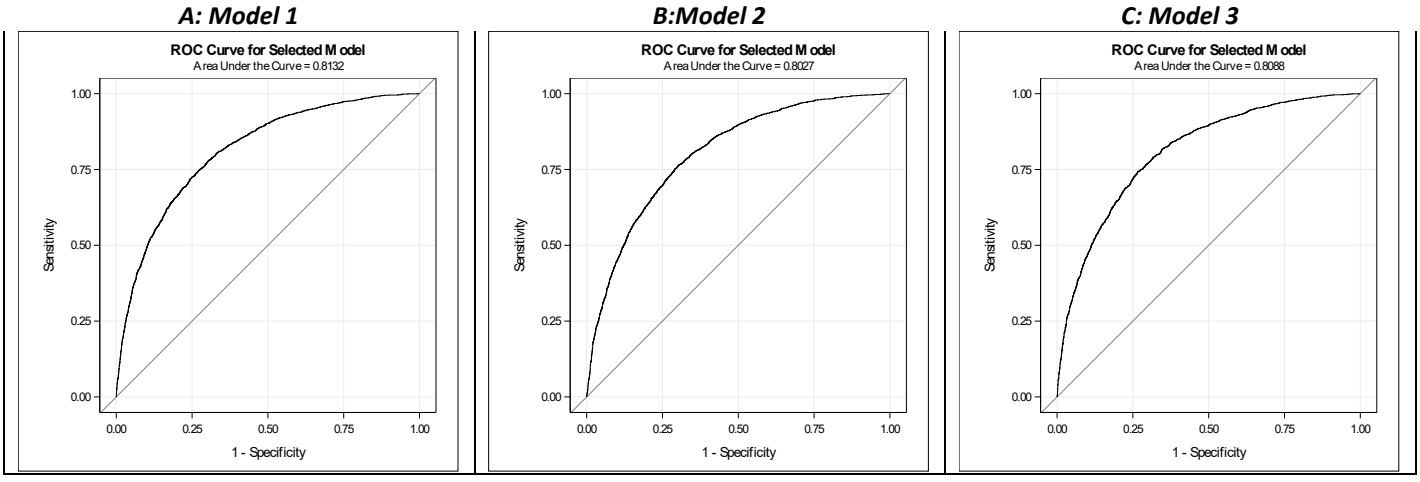
Therefore, in our analysis, we are looking for the model with the smallest -2 Log L, AIC, and BIC values, and the largest r-squared, KS, and AUC value. Table 11 holds the statistics for evaluating the three models.

Table 11: Model Fit Statistics

	AIC	BIC	-2 Log L	R-Squared	AUC	KS
Model 1	7365.161	7582.378	7303.161	0.2281	0.8132	47%
Model 2	7513.635	7632.754	7479.635	0.2112	0.8027	46%
Model 3	7425.762	7607.944	7373.762	0.2213	0.8088	47%

If we were to go off of sheer number, Model 1 is the best model, Model 2 is the second best, and Model 3 ranks third. Yet the values are so close to each other that there is no statistically significant reason to prefer one model of the other. Even the ROC curves are almost identical, not an earth shattering revelation given the proximity of the AUC values.

Figure 7: ROC Curves



ROC curve measures how well a model can differentiate between true positives and false positives. Good bow (as seen in Figure 7) indicates that when the model says it is positive the value is in fact

positive. The goodness-of-fit tests ultimately will not be of much use in determining the best model. The determination came down to parsimony and component significance. Model two had the smallest number of dimensions, 15, and each of the variables had a p-value less than .05.

Given the strong predictive ability of the model and perceived usefulness the final step is to add create the severity model. As stated before this model has poor predictive ability. As a quick summary the model was constructed using backward selection and a cutoff value of .05. It created the following model.

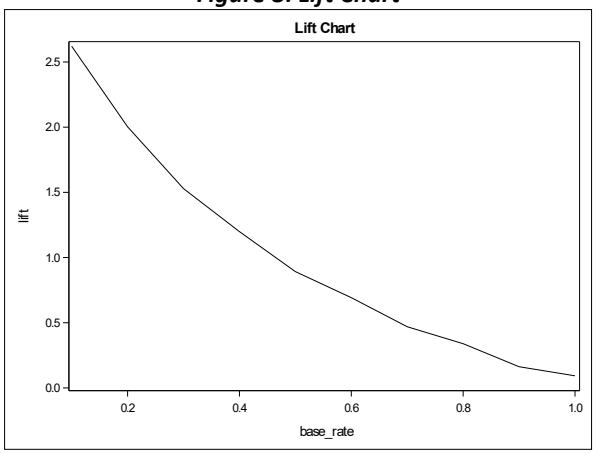
P_TARGET_AMT= 2.92118 + 0.07019* ln_Bluebook + 0.00608*MVR_PTS

The model used transformed variables to help meet the assumptions of linear regression. The transformations did little to improve the explained variance bringing it up slightly to 1.9 percent.

Conclusion

In most cases you can't come up with the best model, but one can come up with a good model. There are conceivably exists a large number of good models. In determining the best model for predicting the probability of a customer getting into a crash, we were able to devise three good models. Similarity in the fit statistics required some additional thought, but in the end we chose the model with the fewest number of parameters where each parameter was significantly different than zero. The model surmises that customers in urban areas, who drive commercially, have small incomes, have children that drive, filed multiple claims in the last five years, have children in general, are young, and single are more likely to get into a crash.

Figure 8: Lift Chart



Based on the Lift chart in Figure 8, the advantage to using the model drops off around 90%, indicating that the model is better than random selection for 90% of the data. The strong performance of this logistic regression model gives reason to believe the company could predict whether or not a customer might be involved in a crash. In determining the cost associated with that crash, more data should be gathered. A claim price could be impacted by the severity of the crash, restitution to victims of at fault drivers, and property damage. To make a more accurate model, an r-value greater than .019, like information, needs to be collected. It is also possible to use business rules to devise a good estimate for the claim value.

APPENDEX 1: All SAS Used

```
libname mydata '/home/thomaspayne20160/my_courses/mott/c_7416/SAS_Data' access=readonly;
libname mywork '/home/thomaspayne20160/ecourse/411/Assn_2';



********************************************************************************
***:
* Section 1A Do some EDA;
********************************************************************************
***;

proc contents data=mydata.logit_insurance;
run;

proc print data=mydata.logit_insurance(obs=10);
run;

proc means data=mydata.logit_insurance(drop=INDEX) nmiss mean median max min  stderr stddev
ndec=0;
run;



proc freq data=mydata.logit_insurance(drop=INDEX )  ;
  tables car_type car_use education homekids job kidsdriv mstatus mvr_pts
  parent1 red_car revoked sex tif travtime urbanicity car_age / missing;
run;

********************************************************************************
***:
```

```
* Section 1B Look At Targer_Flag
*******************************************************************************
***;
data temp1;
set mydata.logit_insurance;
If target_amt >0 then log_target_amt= log10(target_amt);
else log_target_amt=0;
run;

ods graphics on;
proc freq data=mydata.logit_insurance(drop=INDEX ) ;
  tables target_flag / plots(only)=freqplot;
  run;
ods graphics off;

Proc univariate data=temp1;
where target_amt>0;
Var target_amt;
run;

proc sgplot data=temp1;
 vbox target_amt;
 xaxis label='Cost';
 keylegend / title="Box Plot";
run;

proc sgplot data=temp1;
 histogram target_amt/SHOWBINS;
run;
```

```
proc sgplot data=temp1;
  Where target_amt > 0;
  vbox target_amt;
  xaxis label='Cost';
  keylegend / title="Box Plot";
run;

proc sgplot data=temp1;
  Where target_amt > 0;
  histogram target_amt/SHOWBINS;
run;

proc sgplot data=temp1;
  Where log_target_amt > 0;
  vbox log_target_amt;
  xaxis label='Log10 Cost';
  keylegend / title="Box Plot";
run;

proc sgplot data=temp1;
  Where target_amt > 0;
  histogram log_target_amt/SHOWBINS;
run;

*********************************************************************************
***:
* Section 1C Preditor Vraibles EDA;
```

```
*************************************************************************************
***;
ods graphics on;
proc freq data=mydata.logit_insurance(drop=INDEX )  ;
  tables car_type*target_flag
        car_use*target_flag
        education*target_flag
        job*target_flag
        mstatus*target_flag
        parent1*target_flag
        red_car*target_flag
        revoked*target_flag
        sex*target_flag
        urbanicity*target_flag / nocum plots(only)=freqplot(scale=Percent) missing;
run;
ods graphics off;

*///logistic models for categorical vriables;
proc logistic  data=temp1;
class car_type / PARAM=reference;
model target_flag(event='1')= CAR_TYPE ;
ESTIMATE 'z_SUV' CAR_TYPE -1 -1 -1 -1 -1;
run;

proc logistic  data=temp1;
class EDUCATION / PARAM=reference;
model target_flag(event='1')= EDUCATION;
ESTIMATE 'z_High Schoo' EDUCATION -1 -1 -1 -1;
run;
```

```
proc logistic  data=temp1;
class JOB / PARAM=reference;
model target_flag(event='1')= JOB;
ESTIMATE 'z_Blue Collar' JOB -1 -1 -1 -1 -1 -1 -1;
run;

proc logistic  data=temp1;
class CAR_USE / PARAM=reference;
model target_flag(event='1')= CAR_USE/ expb;
run;

proc logistic  data=temp1;
class MSTATUS / PARAM=reference;
model target_flag(event='1')= MSTATUS/ expb;
run;

proc logistic  data=temp1;
class PARENT1 / PARAM=reference;
model target_flag(event='1')= PARENT1/ expb;
run;

proc logistic  data=temp1;
class RED_CAR / PARAM=reference;
model target_flag(event='1')= RED_CAR / expb;
run;

proc logistic  data=temp1;
class REVOKED / PARAM=reference;
```

```
model target_flag(event='1')= REVOKED / expb;
run;

proc logistic  data=temp1;
class SEX / PARAM=reference;
model target_flag(event='1')= SEX / expb;
run;

proc logistic  data=temp1;
class URBANICITY / PARAM=reference;
model target_flag(event='1')= URBANICITY / expb;
run;

*/////// Continuous and Discrete variables  ;

proc sgplot data=temp1;
 vbox  BLUEBOOK  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  HOME_VAL  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  INCOME /group=target_flag;
```

```
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  OLDCLAIM  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  TRAVTIME  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  AGE  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  CAR_AGE  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
```

```
 vbox  CLM_FREQ  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  HOMEKIDS  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  KIDSDRIV  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  MVR_PTS  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

proc sgplot data=temp1;
 vbox  TIF  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;
```

```
proc sgplot data=temp1;
 vbox  YOJ  /group=target_flag;
 xaxis label="Crash";
 keylegend / title="BoxPlot";
run;

*// Logistic Single Varaible Models;
proc logistic  data=temp1;
model target_flag(event='1')= BLUEBOOK / expb rsq ;
Effectplot FIT(x=BLUEBOOK) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= HOME_VAL / expb rsq ;
Effectplot FIT(x=HOME_VAL) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= INCOME / expb rsq ;
Effectplot FIT(x=INCOME) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= OLDCLAIM / expb rsq ;
Effectplot FIT(x=OLDCLAIM) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= TRAVTIME / expb rsq ;
```

```
Effectplot FIT(x=TRAVTIME) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= AGE / expb rsq ;
Effectplot FIT(x=AGE) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= CAR_AGE / expb rsq ;
Effectplot FIT(x=CAR_AGE) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= CLM_FREQ / expb rsq ;
Effectplot FIT(x=CLM_FREQ) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= HOMEKIDS / expb rsq ;
Effectplot FIT(x=HOMEKIDS) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= KIDSDRIV / expb rsq ;
Effectplot FIT(x=KIDSDRIV) / Link;
run;

proc logistic  data=temp1;
```

```
model target_flag(event='1')= MVR_PTS / expb rsq ;
Effectplot FIT(x=MVR_PTS) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= TIF / expb rsq ;
Effectplot FIT(x=TIF) / Link;
run;

proc logistic  data=temp1;
model target_flag(event='1')= YOJ/ expb rsq ;
Effectplot FIT(x=YOJ) / Link;
run;

*///histogram of Age, MVR_PTS, and CLM_FREQ;

proc freq data=mydata.logit_insurance(drop=INDEX )  ;
  tables
       target_flag*YOJ
       target_flag*tif
       target_flag*homekids
       target_flag*car_Age
       target_flag*Age
       target_flag*MVR_PTS
       target_flag*CLM_FREQ / nocum nopercent norow plots(only)=freqplot(scale=percent) missing;
run;

proc corr data=mydata.logit_insurance(drop=INDEX ) ;
var BLUEBOOK HOME_VAL INCOME OLDCLAIM TRAVTIME AGE CAR_AGE
```

```
   CLM_FREQ HOMEKIDS KIDSDRIV MVR_PTS TIF YOJ;
Run;

proc corr data=mydata.logit_insurance(drop=INDEX ) noprint outp=CORFILE;
where target_amt>0;
var BLUEBOOK HOME_VAL INCOME OLDCLAIM TRAVTIME AGE CAR_AGE
   CLM_FREQ HOMEKIDS KIDSDRIV MVR_PTS TIF YOJ target_amt;
Run;

data CORFILE1;
set CORFILE;
if _TYPE_ in ("CORR");
keep _NAME_ TARGET_amt;
run;

proc sort data=CORFILE1;
by descending TARGET_amt;
run;

proc sort data=CORFILE1;
by descending TARGET_amt;
run;

proc print data=CORFILE1;
run;



data one;
  set temp1;
```

```
 if (oldclaim = 0) then claim=1;
 else if (oldclaim < 4637) then claim=2;
 else if (oldclaim < 27091) then  claim=3;
 else  claim=5;
 run;

proc univariate data= temp1;
var oldclaim;
run;

proc sgplot data=one;
 vbox claim/group=target_flag grouporder=ascending;
 xaxis label="claim";
 keylegend / title="BoxPlot";
 run;

proc sgplot  data= temp1;
histogram oldclaim;
run;

*************************************************************************************
***:
* SECTION 2: DATA PREP;
*************************************************************************************
***;

*/// Fix missing values with replacement;
data temp2;
  set temp1;
```

```
  if age = "." then age=45;
  if car_age = "." then car_age = 8;
  if home_val = "." then home_val = 161160;
  if income = "." then income = 54028;
  if job = "." or job = " " then job = "z_Blue Collar";
  if yoj = "." then yoj = 11;
run;

data temp3;
   set temp2;
*//cars can not have a negative age;
   if car_age < 0 then delete;
*//this is to help account for age if it is a polynomial distrobution;
   Age_2= Age*age;
   Age_3=age*age*age;
   Age_4=age*age*age*age;
*//create a new binomal variable;
   if CLM_FREQ < 1 then BIN_CLM_FREQ = 0;
     else BIN_CLM_FREQ =1;
*//create binomial variable for havinge kids;
   if homekids = 0 then BIN_HOMEKIDS = 0;
     else BIN_HOMEKIDS=1;
*//create a new binomal vriable for claims paid in last 5 years;
   if oldclaim < 4637 then BIN_OLDCLAIM = 0;
     else BIN_OLDCLAIM = 1;
*//create a new binomal based on driving record points;
   if MVR_PTS >6 then BIN_MVR_PTS = 1;
     else BIN_MVR_PTS = 0;
*//creat binary job variable;
```

```
   if job in ('Doctor', 'Lawyer', 'Manager') then BIN_JOB = 0;
      else BIN_JOB = 1;
*//create binary education vraible;
   if education in ('<High School', 'z_High School') then BIN_education = 1;
      else BIN_Education = 0;
   if car_type in ('z_SUV', 'Sports Car') then BIN_CAR = 1;
      else BIN_CAR = 0;
*//transform vraibles for OLS;
ln_Bluebook = log(bluebook);
run;

 *//check new vriables;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_HOMEKIDS/ expb;
run;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_CLM_FREQ/ expb;
run;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_OLDCLAIM/ expb;
run;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_MVR_PTS/ expb;
run;
```

```
proc logistic  data=temp3;
model target_flag(event='1')= BIN_CAR/ expb;
run;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_JOB/expb;
run;

proc logistic  data=temp3;
model target_flag(event='1')= BIN_EDUCATION/ expb;
run;

proc freq data=temp3 (drop=INDEX) ;
  tables
       target_flag*BIN_MVR_PTS
       target_flag*BIN_OLDCLAIM
       target_flag*BIN_HOMEKIDS
       target_flag*BIN_CLM_FREQ
       target_flag*BIN_CAR
       target_flag*BIN_EDUCATION
       target_flag*BIN_JOB / nocum nopercent norow plots(only)=freqplot(scale=percent) missing;
run;

*******************************************************************************
***:
* SECTION 3: Model Building;
*******************************************************************************
***;
*//model 1 using just fixed data;
```

```
proc logistic data=temp3 plots( unpack label ) = (roc );
class car_type(param=ref) car_use(param=ref) education(param=ref) job(param=ref)
    mstatus(param=ref) parent1(param=ref) red_car(param=ref) urbanicity(param=ref)
    REVOKED(param=ref) sex(param=ref);
model target_flag(event='1')=car_type car_use education job mstatus parent1 red_car urbanicity
    REVOKED sex age bluebook car_age clm_freq homekids home_val income kidsdriv mvr_pts oldclaim
tif
    travtime yoj  / selection=stepwise slentry=.01 slstay=.01 link=logit rsq;
output out=pred1 p=phat;
run;

proc rank data=pred1 out=training_scores1 descending groups=10;
var phat;
ranks score_decile;
run;

proc means data=training_scores1 sum;
class score_decile;
var target_flag;
output out=pm_out1 sum(target_flag)=Y_Sum;
run;

proc print data=pm_out1; run;

data ks_chart;
        set pm_out1 (where=(_type_=1));
        by _type_;
        Nobs=_freq_;
        score_decile = score_decile+1;
```

```
        if first._type_ then do;
                cum_obs=Nobs;
                model_pred=Y_Sum;
        end;
        else do;
                cum_obs=cum_obs+Nobs;
                model_pred=model_pred+Y_Sum;
        end;
        retain cum_obs model_pred;

        FP=NOBS-Y_SUM;
        Total_FP=CUM_OBS-MODEL_PRED;
        Cum_TP=model_pred/2152;
        CUM_FP=TOTAL_FP/6008;
        Cum_Diff = CUM_TP-CUM_FP;
        drop _freq_ _type_ ;
run;

proc print data=ks_chart; run;

*//model 2 using just fixed data and new variables;
proc logistic data=temp3 plots( unpack label ) = (roc );
class car_use (param=ref) parent1 (param=ref) red_car (param=ref)
    urbanicity (param=ref) mstatus (param=ref) REVOKED(param=ref)
    sex(param=ref);
model target_flag(event='1')= car_use parent1 red_car urbanicity
    mstatus  age age_2 age_3 age_4 bluebook car_age clm_freq home_val
    income kidsdriv travtime yoj BIN_MVR_PTS BIN_OLDCLAIM BIN_HOMEKIDS
```

```
    BIN_CLM_FREQ BIN_JOB BIN_EDUCATION BIN_CAR
    / selection=stepwise slentry=.01 slstay=.01 link=logit rsq;
output out=pred2 p=phat;
run;

proc rank data=pred2 out=training_scores2 descending groups=10;
var phat;
ranks score_decile;
run;

proc means data=training_scores2 sum;
class score_decile;
var target_flag;
output out=pm_out2 sum(target_flag)=Y_Sum;
run;

proc print data=pm_out2; run;

data ks_chart2;
        set pm_out2 (where=(_type_=1));
        by _type_;
        Nobs=_freq_;
        score_decile = score_decile+1;

        if first._type_ then do;
                cum_obs=Nobs;
                model_pred=Y_Sum;
        end;
        else do;
```

```
            cum_obs=cum_obs+Nobs;
            model_pred=model_pred+Y_Sum;
        end;
        retain cum_obs model_pred;

        FP=NOBS-Y_SUM;
        Total_FP=CUM_OBS-MODEL_PRED;
        Cum_TP=model_pred/2152;
        CUM_FP=TOTAL_FP/6008;
        Cum_Diff = CUM_TP-CUM_FP;
        drop _freq_ _type_ ;
run;

proc print data=ks_chart2; run;

*//model 3 all the varaibles;
proc logistic data=temp3 plots( unpack label ) = (roc );
class car_type(param=ref) car_use(param=ref) education(param=ref) job(param=ref)
    mstatus(param=ref) parent1(param=ref) red_car(param=ref) urbanicity(param=ref)
    REVOKED(param=ref) sex(param=ref);
model target_flag(event='1')=car_type car_use education job mstatus parent1 red_car
    urbanicity REVOKED sex age age_2 age_3 age_4 bluebook car_age clm_freq home_val
    income kidsdriv travtime yoj BIN_MVR_PTS BIN_MVR_PTS BIN_OLDCLAIM BIN_HOMEKIDS
    BIN_CLM_FREQ BIN_JOB BIN_EDUCATION BIN_CAR
    / selection=stepwise slentry=.001 slstay=.001 link=logit rsq;
output out=pred3 p=phat;
run;

proc rank data=pred3 out=training_scores3 descending groups=10;
```

```
var phat;
ranks score_decile;
run;

proc means data=training_scores3 sum;
class score_decile;
var target_flag;
output out=pm_out3 sum(target_flag)=Y_Sum;
run;

proc print data=pm_out3; run;

data ks_chart3;
        set pm_out3 (where=(_type_=1));
        by _type_;
        Nobs=_freq_;
        score_decile = score_decile+1;

        if first._type_ then do;
                cum_obs=Nobs;
                model_pred=Y_Sum;
        end;
        else do;
                cum_obs=cum_obs+Nobs;
                model_pred=model_pred+Y_Sum;
        end;
        retain cum_obs model_pred;

        FP=NOBS-Y_SUM;
```

```
        Total_FP=CUM_OBS-MODEL_PRED;
        Cum_TP=model_pred/2152;
        CUM_FP=TOTAL_FP/6008;
        Cum_Diff = CUM_TP-CUM_FP;
        drop _freq_ _type_ ;
run;

proc print data=ks_chart3; run;

*//OLS model to predict the claim amount;
proc reg data=temp3 plots=diagnostics(stats=(default aic));
where target_amt>0;
model log_target_amt = ln_BLUEBOOK BLUEBOOK HOME_VAL INCOME
   OLDCLAIM TRAVTIME AGE CAR_AGE CLM_FREQ HOMEKIDS KIDSDRIV
   MVR_PTS TIF YOJ/ selection=backward sls=.05 vif aic adjrsq sbc bic;
run;



*************************************************************************************
***:
* SECTION 5: DATA STEP;
*************************************************************************************
***;

data temp1;
set mydata.logit_insurance;

  If target_amt >0 then log_target_amt= log10(target_amt);
    else log_target_amt=0;
```

```
ln_Bluebook = log(bluebook);

if age = "." then age=45;
if car_age = "." then car_age = 8;
if home_val = "." then home_val = 161160;
if income = "." then income = 54028;
if job = "." or job = " " then job = "z_Blue Collar";
if yoj = "." then yoj = 11;

if car_age < 0 then delete;

Age_2= Age*age;
Age_3=age*age*age;
Age_4=age*age*age*age;

if CLM_FREQ < 1 then BIN_CLM_FREQ = 0;
  else BIN_CLM_FREQ =1;
if homekids = 0 then BIN_HOMEKIDS = 0;
  else BIN_HOMEKIDS=1;
if oldclaim < 4637 then BIN_OLDCLAIM = 0;
  else BIN_OLDCLAIM = 1;
if MVR_PTS >6 then BIN_MVR_PTS = 1;
  else BIN_MVR_PTS = 0;
if job in ('Doctor', 'Lawyer', 'Manager') then BIN_JOB = 0;
  else BIN_JOB = 1;
if education in ('<High School', 'z_High School') then BIN_education = 1;
  else BIN_Education = 0;
if car_type in ('z_SUV', 'Sports Car') then BIN_CAR = 1;
```

```
    else BIN_CAR = 0;

run;

proc logistic data=temp1 outmodel=insmodel plots(only)=roc  ;
class car_use (param=ref) parent1 (param=ref) red_car (param=ref)
    urbanicity (param=ref) mstatus (param=ref) REVOKED(param=ref)
    sex(param=ref);
model target_flag(event='1')= car_use parent1 red_car urbanicity
    mstatus  age age_2 age_3 age_4 bluebook car_age clm_freq home_val
    income kidsdriv travtime yoj BIN_MVR_PTS BIN_OLDCLAIM BIN_HOMEKIDS
    BIN_CLM_FREQ BIN_JOB BIN_EDUCATION BIN_CAR
    / selection=stepwise slentry=.01 slstay=.01 link=logit rsq;
output out=pred p=phat  ;
run;

data testdata;
  set mydata.logit_insurance_test;

  ln_Bluebook = log(bluebook);

  if age = "." then age=45;
  if car_age = "." then car_age = 8;
  if home_val = "." then home_val = 161160;
  if income = "." then income = 54028;
  if job = "." or job = " " then job = "z_Blue Collar";
  if yoj = "." then yoj = 11;

  if car_age < 0 then delete;
```

```
  Age_2= Age*age;
  Age_3=age*age*age;
  Age_4=age*age*age*age;

  if CLM_FREQ < 1 then BIN_CLM_FREQ = 0;
    else BIN_CLM_FREQ =1;
  if homekids = 0 then BIN_HOMEKIDS = 0;
    else BIN_HOMEKIDS=1;
  if oldclaim < 4637 then BIN_OLDCLAIM = 0;
    else BIN_OLDCLAIM = 1;
  if MVR_PTS >6 then BIN_MVR_PTS = 1;
    else BIN_MVR_PTS = 0;
  if job in ('Doctor', 'Lawyer', 'Manager') then BIN_JOB = 0;
    else BIN_JOB = 1;
  if education in ('<High School', 'z_High School') then BIN_education = 1;
    else BIN_Education = 0;
  if car_type in ('z_SUV', 'Sports Car') then BIN_CAR = 1;
    else BIN_CAR = 0;
run;

proc logistic inmodel=insmodel;
  score data=testdata out=testscore;
run;

data mywork.payne_ins_prob;
  set testscore;
  p_target_flag= p_1;
  keep index p_target_flag;
```

```
run;

proc print;
run;



*/// model here for p_target_amt///;
data mywork.payne_ins_amt;
  set testscore;

  log_p_target_amt =  +2.92118
  +ln_Bluebook                *0.07019
  +MVR_PTS              *0.00608;

  p_target_amt = 10**(log_p_target_amt);

  keep index p_target_amt;
run;



data mywork.payne_ins;
merge mywork.payne_ins_prob(in=ina) mywork.payne_ins_amt(in=inb);
by INDEX;
if ina;
run;



proc rank data=pred out=training_scores descending groups=10;
var phat;
```

```
ranks score_decile;
run;

proc means data=training_scores sum;
class score_decile;
var target_flag;
output out=pm_out sum(target_flag)=Y_Sum;
run;

proc print data=pm_out; run;

data lift_chart;
        set pm_out (where=(_type_=1));
        by _type_;
        Nobs=_freq_;
        score_decile = score_decile+1;

        if first._type_ then do;
                cum_obs=Nobs;
                model_pred=Y_Sum;
        end;
        else do;
                cum_obs=cum_obs+Nobs;
                model_pred=model_pred+Y_Sum;
        end;
        retain cum_obs model_pred;

        * 2148 represents the number of successes;
        * This value will need to be changed with different samples;
```

```
        pred_rate=model_pred/2152;
        base_rate=score_decile*0.1;
        Cum_Diff = pred_rate-base_rate;
        lift = y_sum / ((2152/8160)*816);
        drop _freq_ _type_ ;
run;

proc print data=lift_chart; run;

title 'Cumlative Response Curve';
proc sgplot data=lift_chart;
  series x=base_rate y=base_rate;
  series x=base_rate y=pred_rate;
run;

*///Univ Ed will not do proc gplot///;
title 'Lift Chart';

proc sgplot data=lift_chart;
  series x=base_rate y=lift;
run;
```