HW 02 INSURANCE
Taylor Vender
PREDICT 411 SEC 60

Kaggle Name: Taylor V
BINGO BONUS : Decision trees for variable replaces for IMP_JOB and IMP_INCOME (20 points)
SAS Macros (10 points)(below)

```
%let PATH	=
/home/taylorvender20180/my_courses/donald.wedding/c_8888/PRED411/UNIT02/HW;
%let NAME	= LR;
%let LIB		= &NAME..;

libname &NAME. "&PATH.";


%let INFILE	=		&LIB.logit_insurance;
%let TEMPFILE = TEMPFILE;
%let SCRUBFILE		= SCRUBFILE;
%let BUCKETFILE = BUCKETFILE;
%LET LASTFILE = LASTFILE;
```

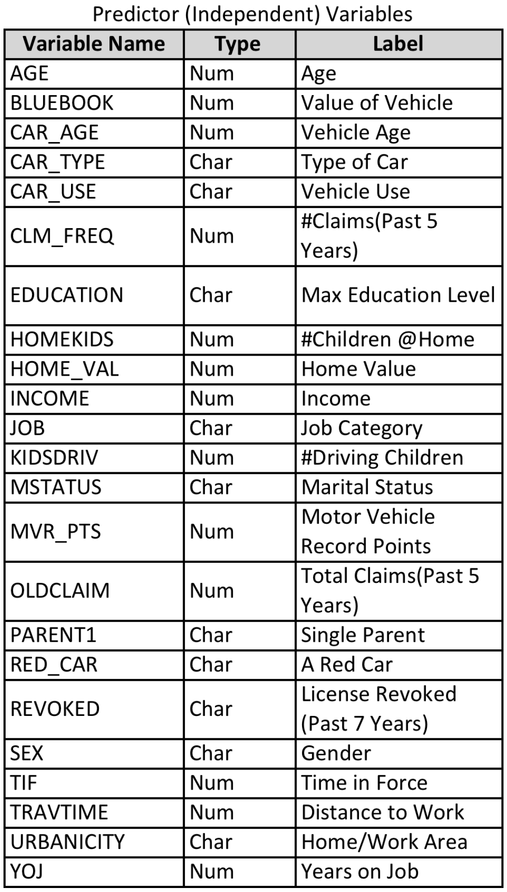
INTRODUCTION:

For this assignment, we will analyze a data set (logit_insurance) that contains 8161 records and use it as the training set for building our model. Each record represents a particular customer at an auto insurance company. Each record contains data on the customer, and we will use this data to develop a logistic regression model to predict the probability that a customer will crash their car.

RESULTS:

DATA EXPLORATION:

The logit_insurance dataset that we are using to train our model contains 8161 observations. Each observation is a record that contains information on a customer at an auto insurance company. The records have two target variables. The first target variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash and a "0" means that the person was not in a car crash. The second target variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero. The TARGET_AMT won't be used in this assignment.



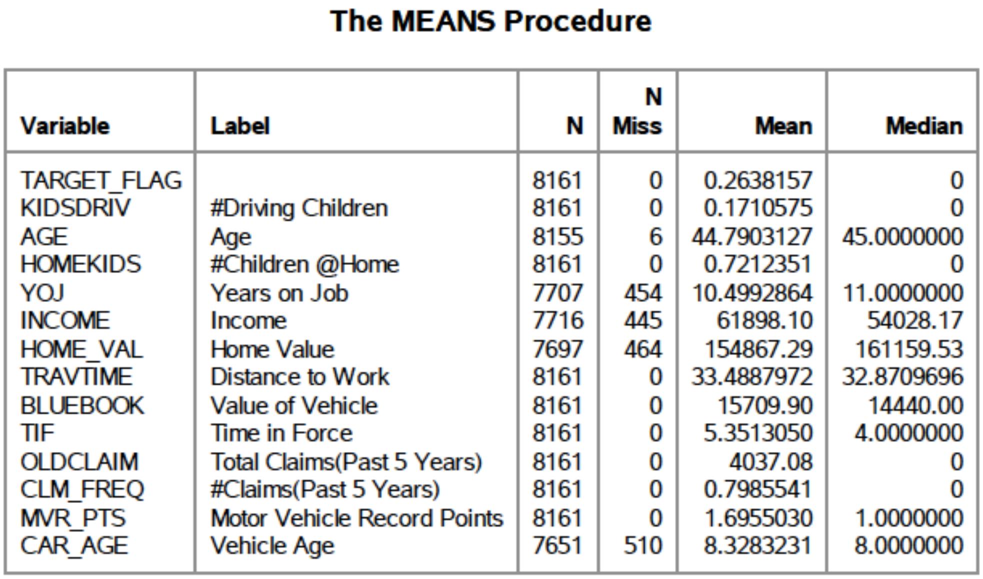
Predictor (Independent) Variables

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

The records contain 10 character variables and 13 numeric variables. Right off the bat we are given some information about the theoretical effect that each variable has on one or both of the independent variables. At the onset, we will give each variable equal consideration. As we do more analysis of the dataset, we will have more information to decide whether or not we think a variable is predictive and if it should be dropped from our model.

VARIABLE NAME	THEORETICAL EFFECT
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 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?
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
YOJ	People who stay at a job for a long time are usually more safe

First we will examine the numeric variables to see if the variables have any missing values. In the next section of this assignment, we will replace any missing values with the mean or median of the variable or by using a decision tree.

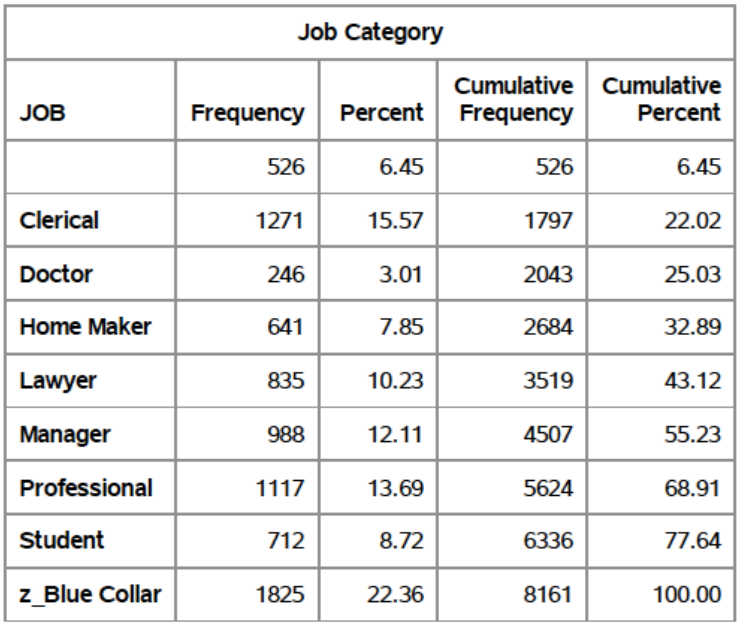


The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median
TARGET_FLAG		8161	0	0.2638157	0
KIDSDRIV	#Driving Children	8161	0	0.1710575	0
AGE	Age	8155	6	44.7903127	45.0000000
HOMEKIDS	#Children @Home	8161	0	0.7212351	0
YOJ	Years on Job	7707	454	10.4992864	11.0000000
INCOME	Income	7716	445	61898.10	54028.17
HOME_VAL	Home Value	7697	464	154867.29	161159.53
TRAVTIME	Distance to Work	8161	0	33.4887972	32.8709696
BLUEBOOK	Value of Vehicle	8161	0	15709.90	14440.00
TIF	Time in Force	8161	0	5.3513050	4.0000000
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	4037.08	0
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0.7985541	0
MVR_PTS	Motor Vehicle Record Points	8161	0	1.6955030	1.0000000
CAR_AGE	Vehicle Age	7651	510	8.3283231	8.0000000

While AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE all have missing values, it is important to notice that the number of missing values isn't an overwhelmingly large proportion of the total number of observations for any of the variables. Just based on this observation, I don't find it necessary to remove any of these variables from the model. I am comfortable imputing the variables with their medians or using a decision tree to fix the missing values.

Since the character variables themselves don't have a mean or median, we have to use a different procedure to look for missing values in out character variables. Thus, the table below looks quite different from the table above.

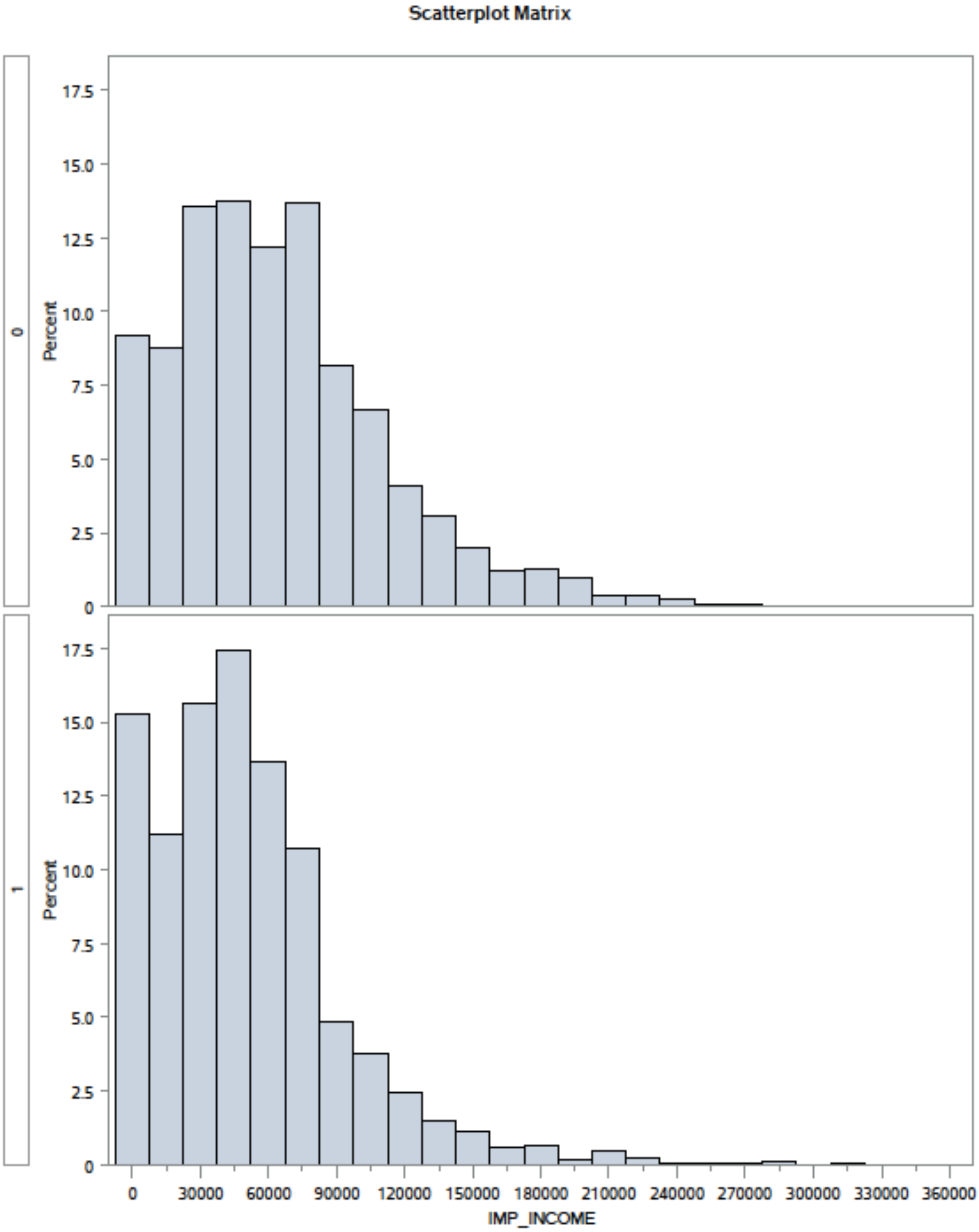


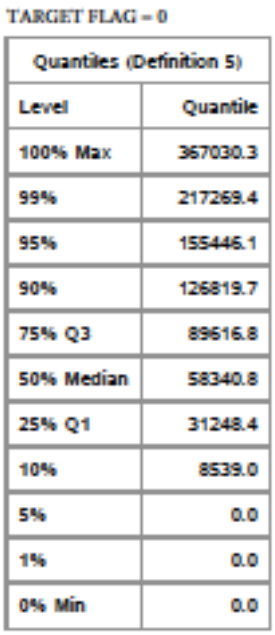
Job Category

JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	526	6.45	526	6.45
Clerical	1271	15.57	1797	22.02
Doctor	246	3.01	2043	25.03
Home Maker	641	7.85	2684	32.89
Lawyer	835	10.23	3519	43.12
Manager	988	12.11	4507	55.23
Professional	1117	13.69	5624	68.91
Student	712	8.72	6336	77.64
z_Blue Collar	1825	22.36	8161	100.00

For each character variable, I looked at a table similar to the one above. As you can see, the cell in the first row under job is blank. This means that there are 526 observations that have a missing value for the variable JOB. Since this is the only character variable with missing values, I only included this table in my report. The other tables do not have any blank category values.

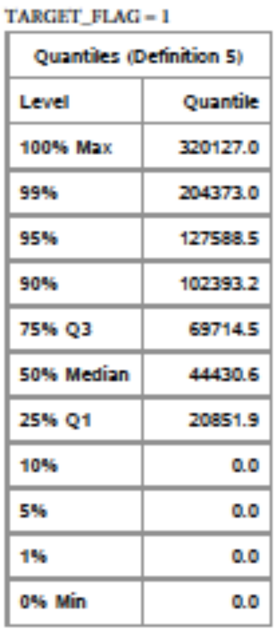
After I fixed all of the missing values (as described in the next section), I looked at the histograms and quantiles of the numeric variables to check for outliers. In order to make sure I didn't cut off any values that could be predictive, I examined the data split up by the people who got in an accident and those who didn't. For brevity, I will not share my analysis for all of the variables. Instead, I will show my analysis for IMP_INCOME, which is the only variable in my model that I decided to put a cap on due to large outliers.


Scatterplot Matrix
0
1
Percent
Percent
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
0
30000
60000
90000
120000
150000
180000
210000
240000
270000
300000
330000
360000
IMP_INCOME



TARGET FLAG = 0

Quantiles (Definition 5)	
Level	Quantile
100% Max	367030.3
99%	217269.4
95%	155446.1
90%	126819.7
75% Q3	89616.8
50% Median	58340.8
25% Q1	31248.4
10%	8539.0
5%	0.0
1%	0.0
0% Min	0.0



TARGET_FLAG = 1

Quantiles (Definition 5)	
Level	Quantile
100% Max	320127.0
99%	204373.0
95%	127588.5
90%	102393.2
75% Q3	69714.5
50% Median	44430.6
25% Q1	20851.9
10%	0.0
5%	0.0
1%	0.0
0% Min	0.0

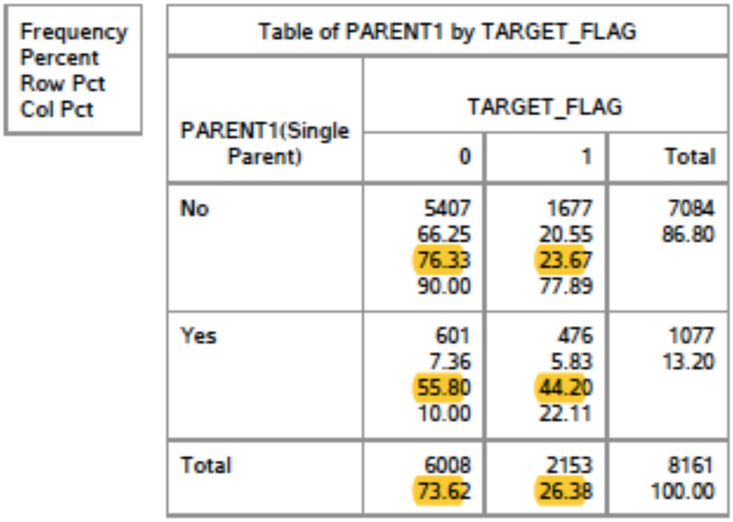
As you can see there is quite a long tail on the right, so I decided to cap the values for the imputed variable income. Since I didn't want to cut the values off at a point that could possible be predictive, I decided to put the cap at the highest value of the two 99th percentiles.

In order to make sure that we have a robust model, it is useful to go through the variables and get rid of variables that don't look like they are predictive of the target variable.

The tables that directly follow compare break the categorical variables into two groups that represent the target variable (TARGET_FLAG = 0 and TARGET_FLAG = 1). As stated before, 0 is people who don't crash car, 1 is people who do crash car.

For our purpose, only care about the third row in each box (the numbers are marked by a red box). The total at the bottom of each table gives you the total % that does crash their car and those that don't. Comparing the total percent to the percent of each category that crashes and the percent that doesn't tells you if something is predictive or not. For instance, if blue-collar people don't crash their car less than the total (average person), we would assume that this variable is predictive.

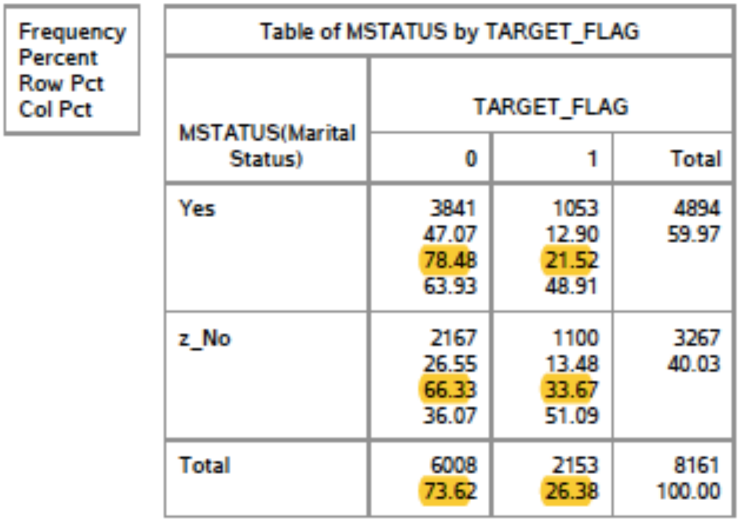
The FREQ Procedure



Frequency
Percent
Row Pct
Col Pct

Table of PARENT1 by TARGET_FLAG

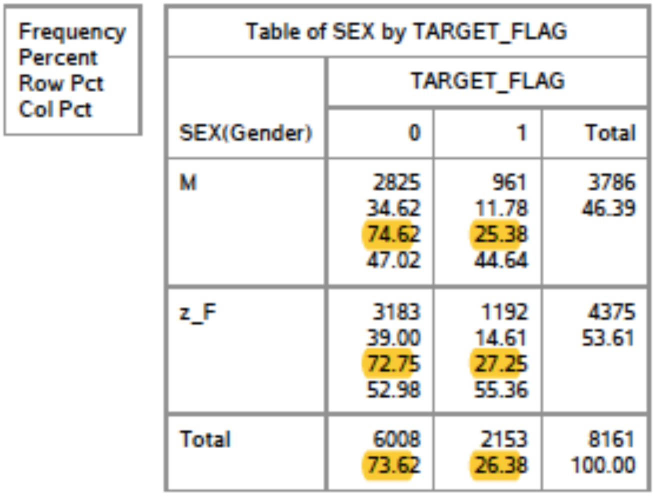
PARENT1(Single Parent)	TARGET_FLAG 0	1	Total
No	5407 66.25 76.33 90.00	1677 20.55 23.67 77.89	7084 86.80
Yes	601 7.36 55.80 10.00	476 5.83 44.20 22.11	1077 13.20
Total	6008 73.62	2153 26.38	8161 100.00



Frequency
Percent
Row Pct
Col Pct

Table of MSTATUS by TARGET_FLAG

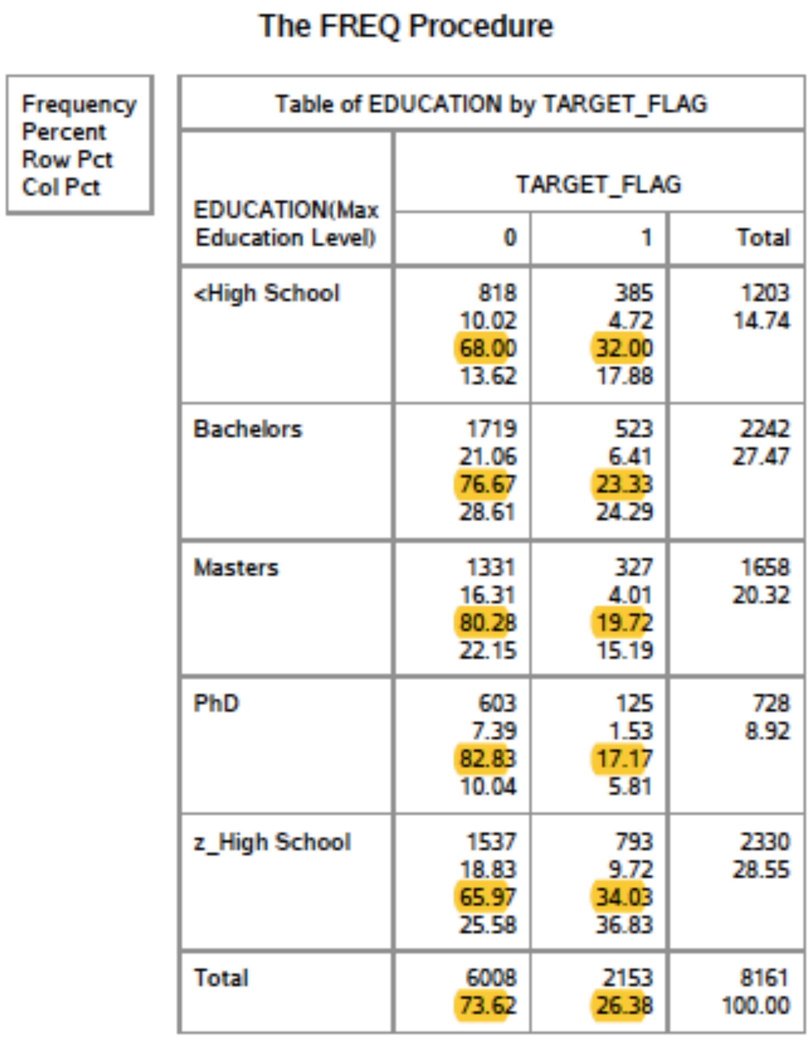
MSTATUS(Marital Status)	TARGET_FLAG 0	1	Total
Yes	3841 47.07 78.48 63.93	1053 12.90 21.52 48.91	4894 59.97
z_No	2167 26.55 66.33 36.07	1100 13.48 33.67 51.09	3267 40.03
Total	6008 73.62	2153 26.38	8161 100.00



Frequency
Percent
Row Pct
Col Pct

Table of SEX by TARGET_FLAG

SEX(Gender)	TARGET_FLAG 0	1	Total
M	2825 34.62 74.62 47.02	961 11.78 25.38 44.64	3786 46.39
z_F	3183 39.00 72.75 52.98	1192 14.61 27.25 55.36	4375 53.61
Total	6008 73.62	2153 26.38	8161 100.00

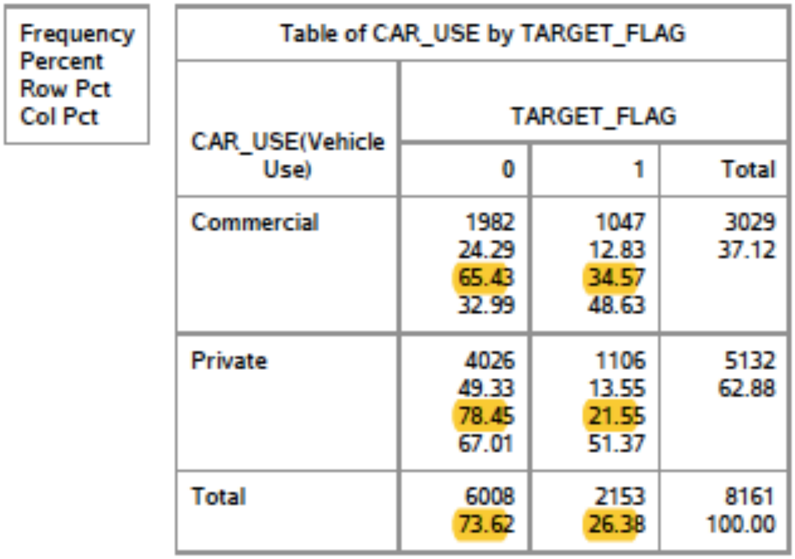


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of EDUCATION by TARGET_FLAG

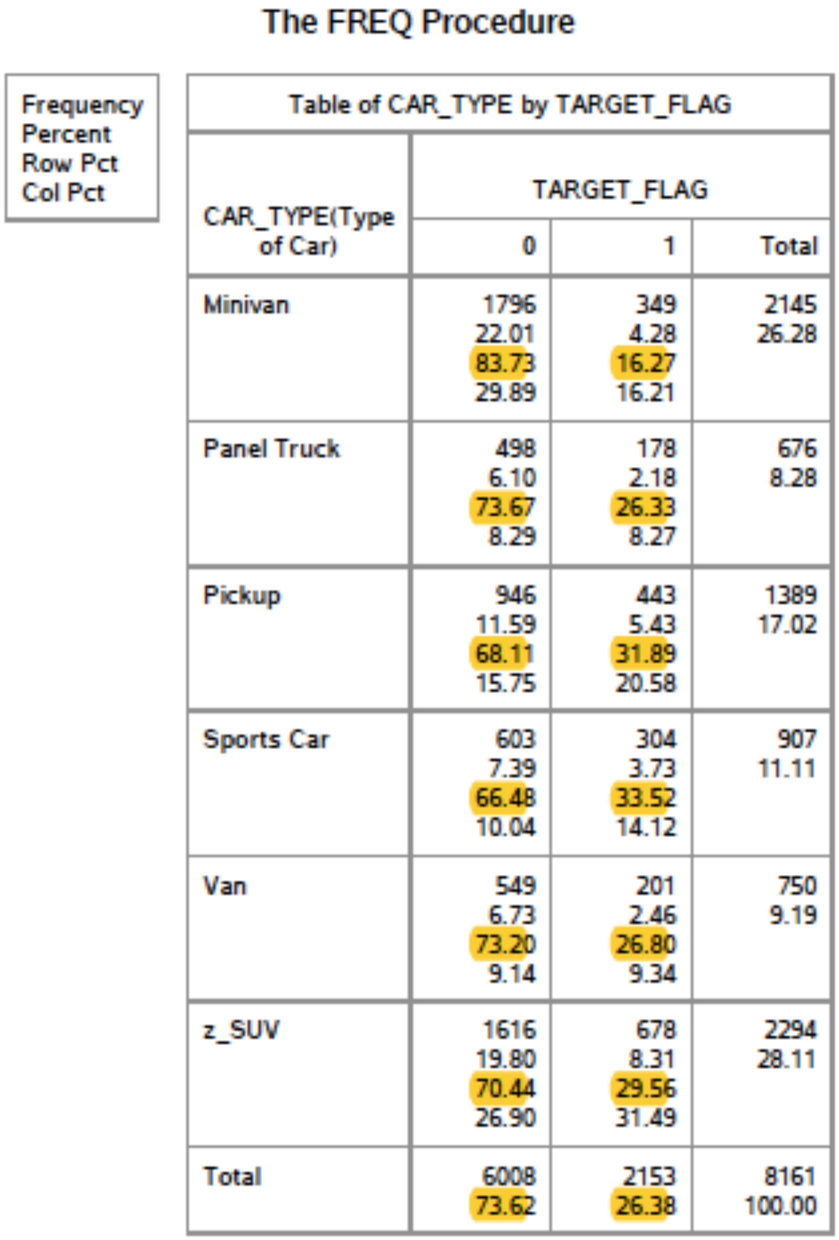
EDUCATION(Max Education Level)	TARGET_FLAG 0	TARGET_FLAG 1	Total
<High School	818 10.02 68.00 13.62	385 4.72 32.00 17.88	1203 14.74
Bachelors	1719 21.06 76.67 28.61	523 6.41 23.33 24.29	2242 27.47
Masters	1331 16.31 80.28 22.15	327 4.01 19.72 15.19	1658 20.32
PhD	603 7.39 82.83 10.04	125 1.53 17.17 5.81	728 8.92
z_High School	1537 18.83 65.97 25.58	793 9.72 34.03 36.83	2330 28.55
Total	6008 73.62	2153 26.38	8161 100.00



Frequency
Percent
Row Pct
Col Pct

Table of CAR_USE by TARGET_FLAG

CAR_USE(Vehicle Use)	TARGET_FLAG 0	TARGET_FLAG 1	Total
Commercial	1982 24.29 65.43 32.99	1047 12.83 34.57 48.63	3029 37.12
Private	4026 49.33 78.45 67.01	1106 13.55 21.55 51.37	5132 62.88
Total	6008 73.62	2153 26.38	8161 100.00

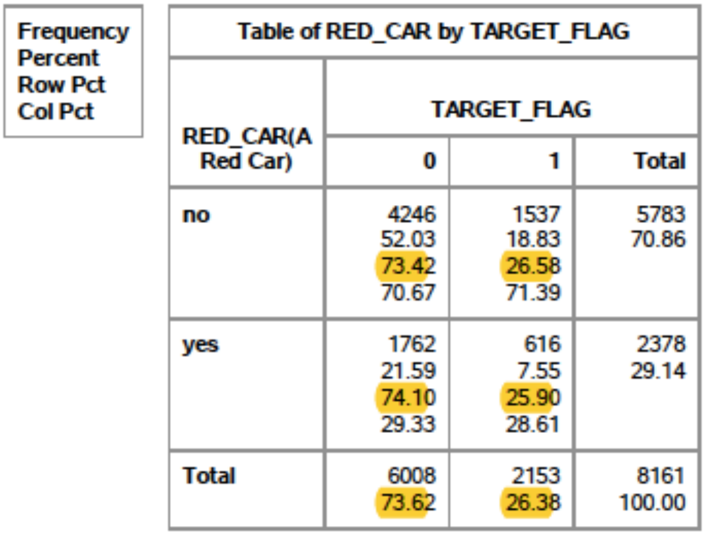


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of CAR_TYPE by TARGET_FLAG

CAR_TYPE(Type of Car)	TARGET_FLAG 0	1	Total
Minivan	1796 22.01 83.73 29.89	349 4.28 16.27 16.21	2145 26.28
Panel Truck	498 6.10 73.67 8.29	178 2.18 26.33 8.27	676 8.28
Pickup	946 11.59 68.11 15.75	443 5.43 31.89 20.58	1389 17.02
Sports Car	603 7.39 66.48 10.04	304 3.73 33.52 14.12	907 11.11
Van	549 6.73 73.20 9.14	201 2.46 26.80 9.34	750 9.19
z_SUV	1616 19.80 70.44 26.90	678 8.31 29.56 31.49	2294 28.11
Total	6008 73.62	2153 26.38	8161 100.00

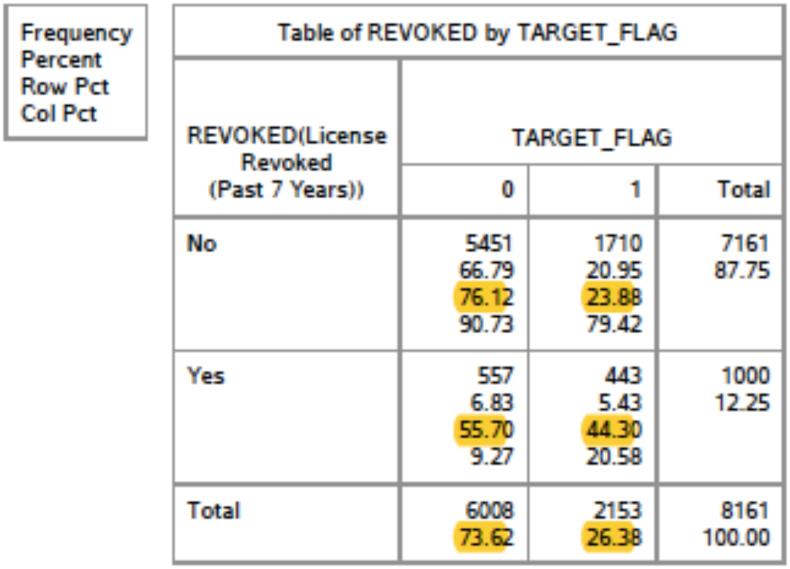


Frequency
Percent
Row Pct
Col Pct

Table of RED_CAR by TARGET_FLAG

RED_CAR(A Red Car)	TARGET_FLAG		
	0	1	Total
no	4246 52.03 73.42 70.67	1537 18.83 26.58 71.39	5783 70.86
yes	1762 21.59 74.10 29.33	616 7.55 25.90 28.61	2378 29.14
Total	6008 73.62	2153 26.38	8161 100.00

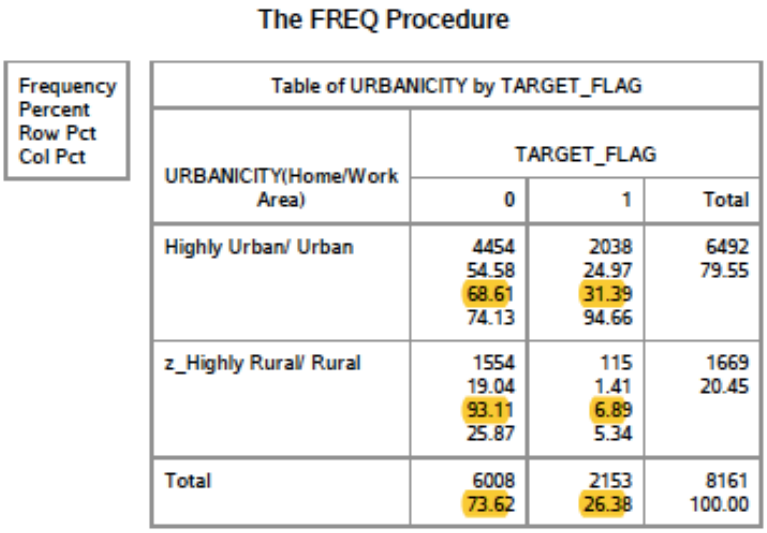
From the total we see that 26.38% of the people crash their car. Looking at the third row under the column that says 1, which indicates that the people do crash their cars, we see that if you do have a red car it is 25.90% and if you don't have a red car it is 26.58% that crashes. Since there is very little difference between the % of people with red cars that crash and people that do not have red cars that crash, this variable does not seem predictive in predicting our target variable TARGET_FLAG. If you have enough data, there might be some sort of interaction such as a red car is more likely to be a sports car and the driver is more likely to crash. However we don't see that here and don't have enough data to look at that, so we will remove RED_CAR from our model completely.



Frequency
Percent
Row Pct
Col Pct

Table of REVOKED by TARGET_FLAG

REVOKED(License Revoked (Past 7 Years))	TARGET_FLAG		
	0	1	Total
No	5451 66.79 76.12 90.73	1710 20.95 23.88 79.42	7161 87.75
Yes	557 6.83 55.70 9.27	443 5.43 44.30 20.58	1000 12.25
Total	6008 73.62	2153 26.38	8161 100.00

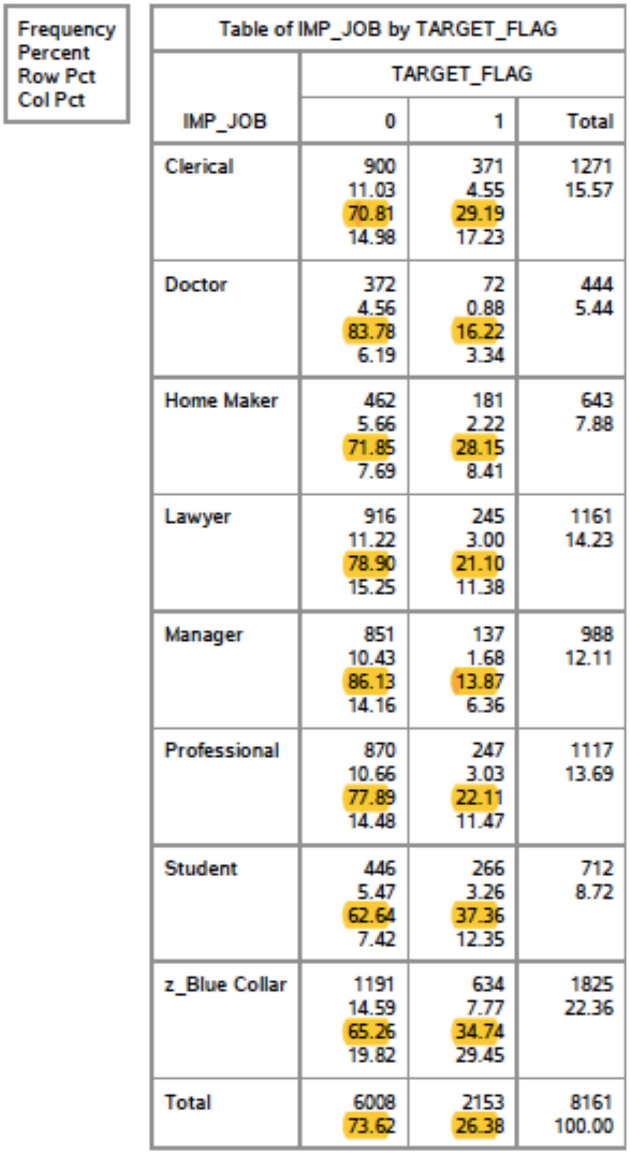


The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of URBANICITY by TARGET_FLAG

URBANICITY(Home/Work Area)	TARGET_FLAG		
	0	1	Total
Highly Urban/ Urban	4454 54.58 68.61 74.13	2038 24.97 31.39 94.66	6492 79.55
z_Highly Rural/ Rural	1554 19.04 93.11 25.87	115 1.41 6.89 5.34	1669 20.45
Total	6008 73.62	2153 26.38	8161 100.00



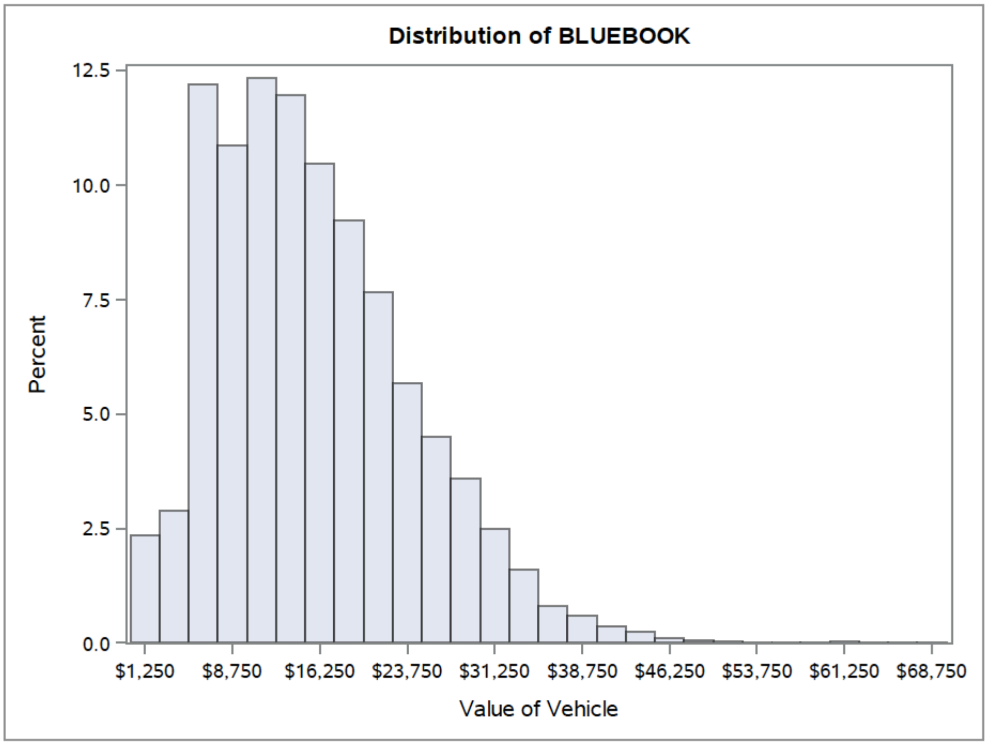
Frequency Percent Row Pct Col Pct

Table of IMP_JOB by TARGET_FLAG

IMP_JOB	TARGET_FLAG		
	0	1	Total
Clerical	900 11.03 70.81 14.98	371 4.55 29.19 17.23	1271 15.57
Doctor	372 4.56 83.78 6.19	72 0.88 16.22 3.34	444 5.44
Home Maker	462 5.66 71.85 7.69	181 2.22 28.15 8.41	643 7.88
Lawyer	916 11.22 78.90 15.25	245 3.00 21.10 11.38	1161 14.23
Manager	851 10.43 86.13 14.16	137 1.68 13.87 6.36	988 12.11
Professional	870 10.66 77.89 14.48	247 3.03 22.11 11.47	1117 13.69
Student	446 5.47 62.64 7.42	266 3.26 37.36 12.35	712 8.72
z_Blue Collar	1191 14.59 65.26 19.82	634 7.77 34.74 29.45	1825 22.36
Total	6008 73.62	2153 26.38	8161 100.00

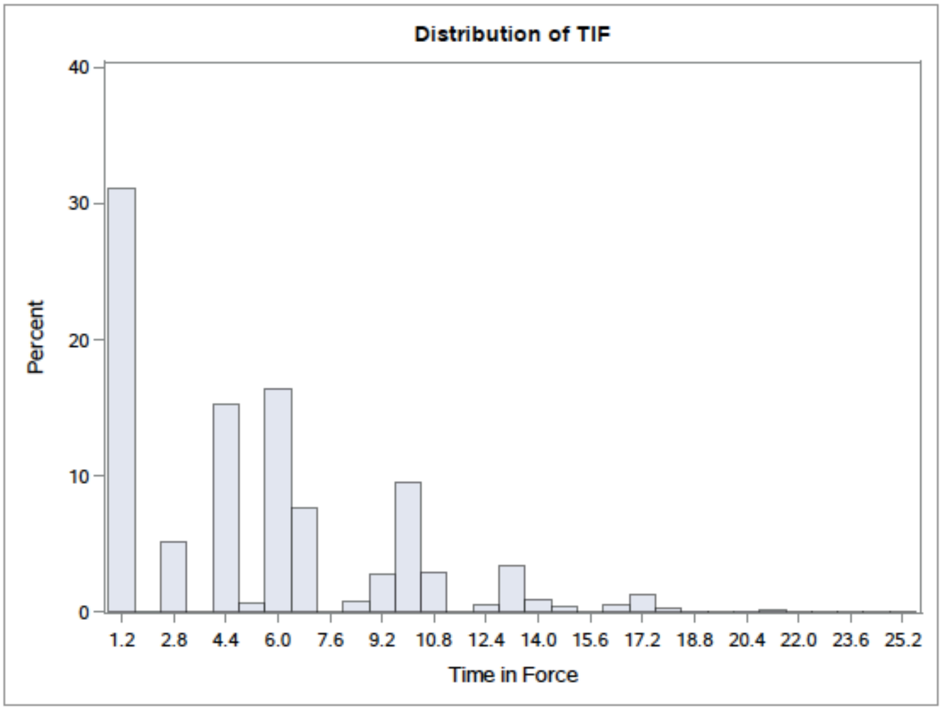
In order to try to improve the model I got using stepwise variable selection with the variables as given (with the exception of necessary imputations to get rid of missing values), I decided to create new variables from a few of the old variables by grouping the values together and creating "buckets" or categories.

After looking at the distributions of the numerical variables using histograms, I thought that my model would benefit from transforming four of the variables that had particular large ranges by grouping the values into four buckets with cutoffs at the 25th, 50th, and 75th percentiles.

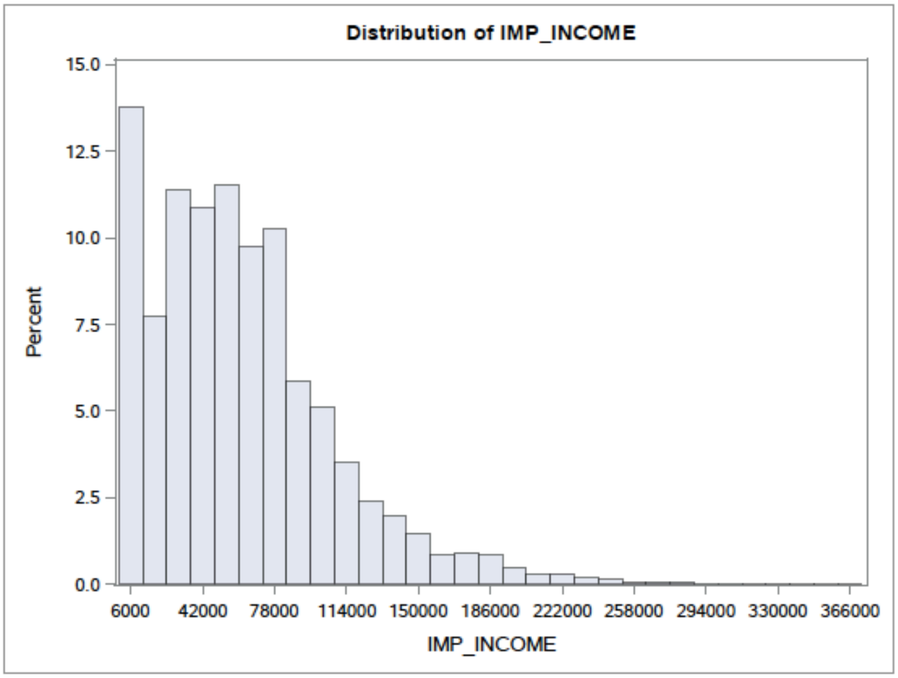


Saturday, February 11, 2017 10:01:47 P

The UNIVARIATE Procedure



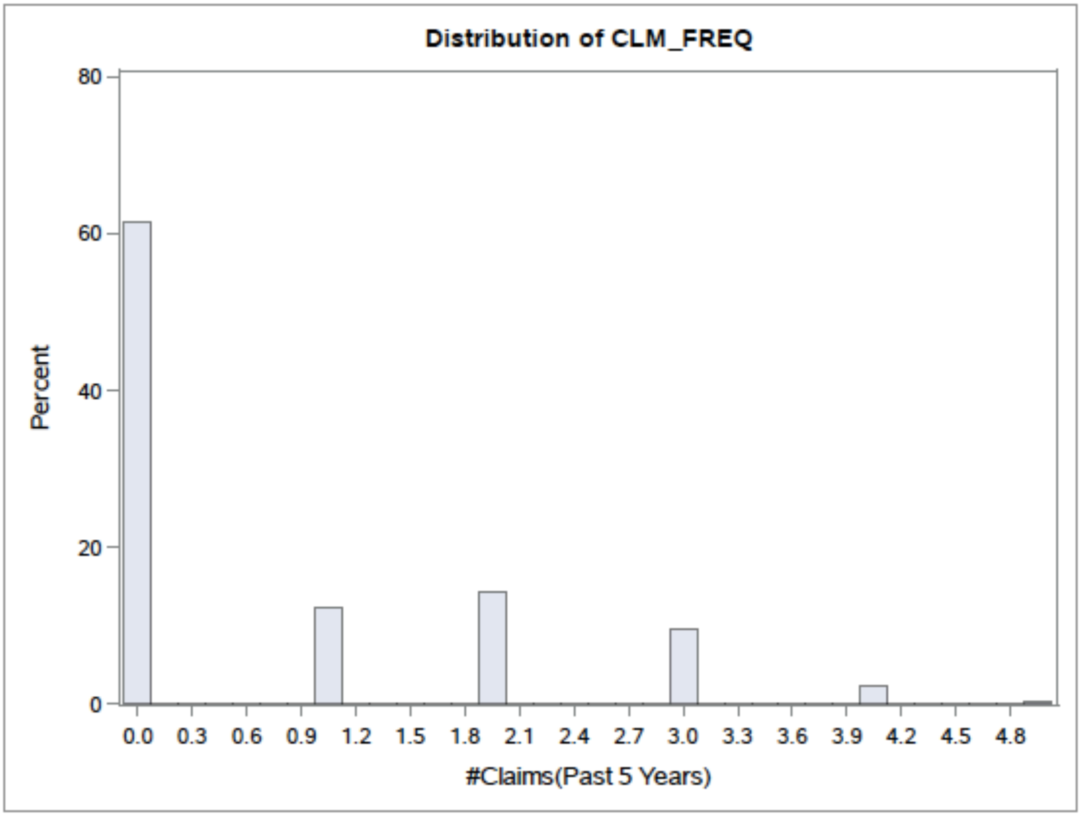
The UNIVARIATE Procedure



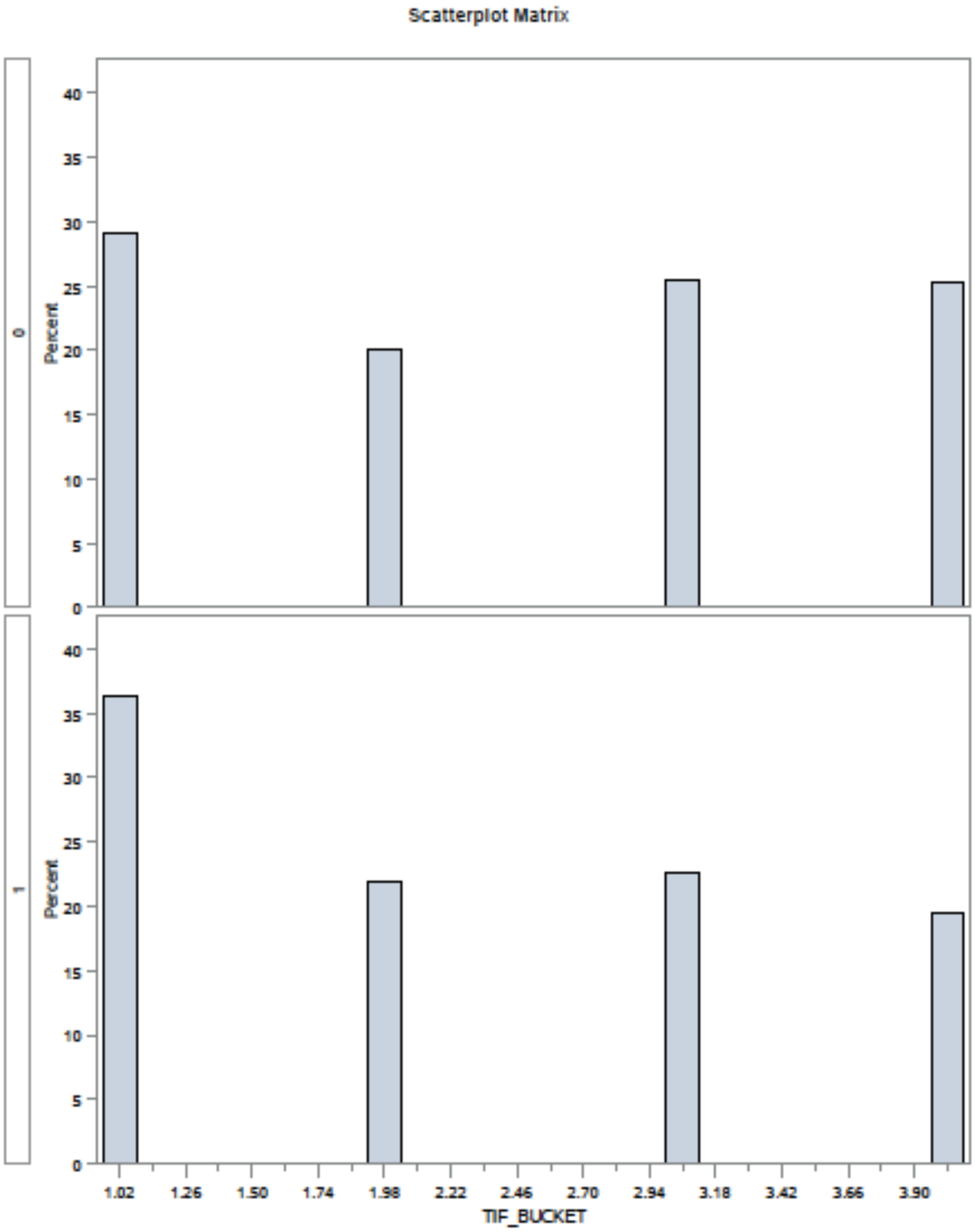
After seeing the distributions shown above, I thought we could improve the model by consolidating some of the data by using four buckets instead of basing the predictions on single values from such a wide range. The histogram below shows the distribution of the #

of claims each person (represented by a single record) in the logit_insurance data set had in the last 5 years. Since an overwhelmingly majority had 0 claims, I decided to create a new variable called HAS_CLMD which would consolidate the values into two groups with HAS_CLMD = 0 representing that the person had 0 claims in the last 5 years and HAS_CLMD = 1 representing that the person has had one or more claims in the last 5 years.

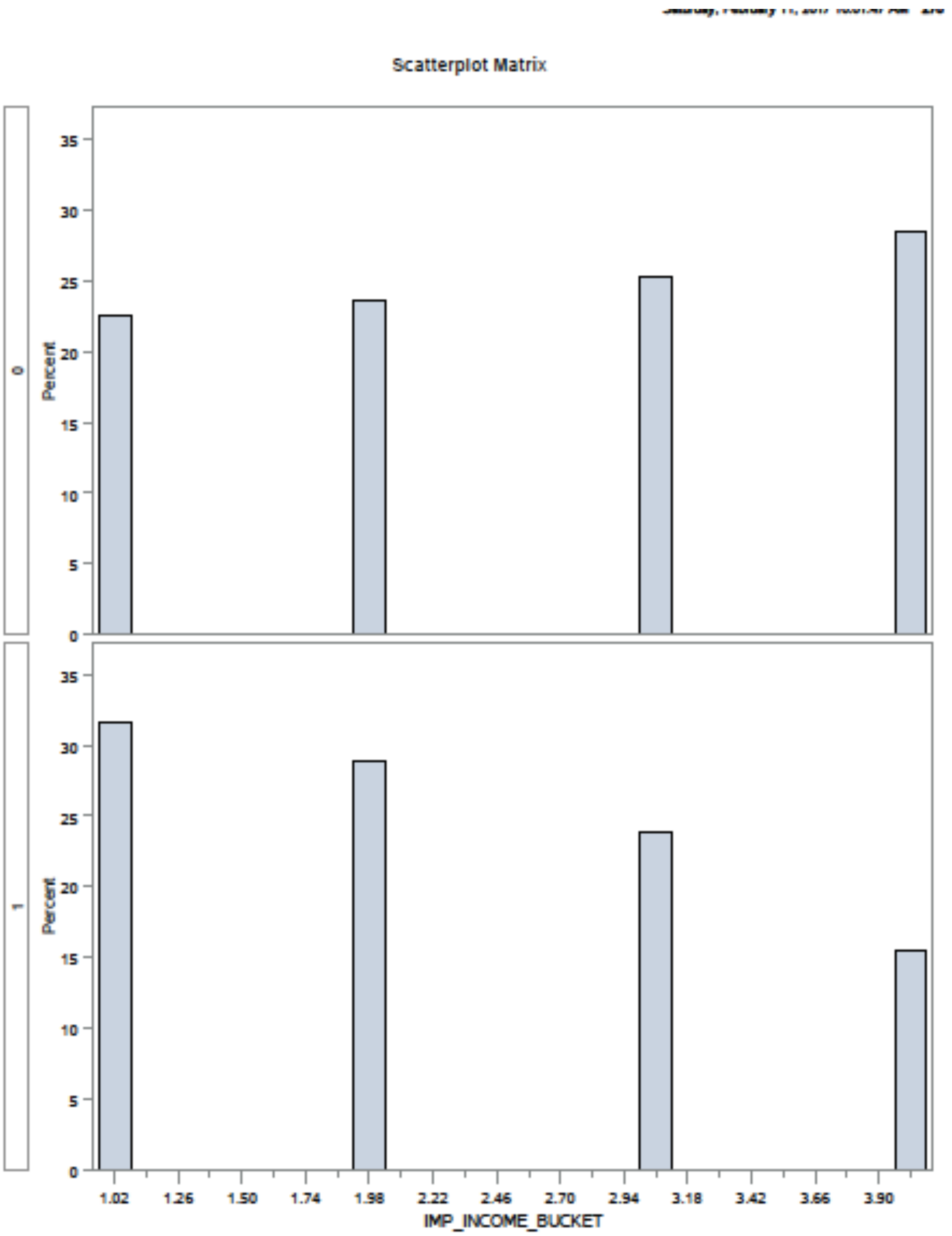
The UNIVARIATE Procedure



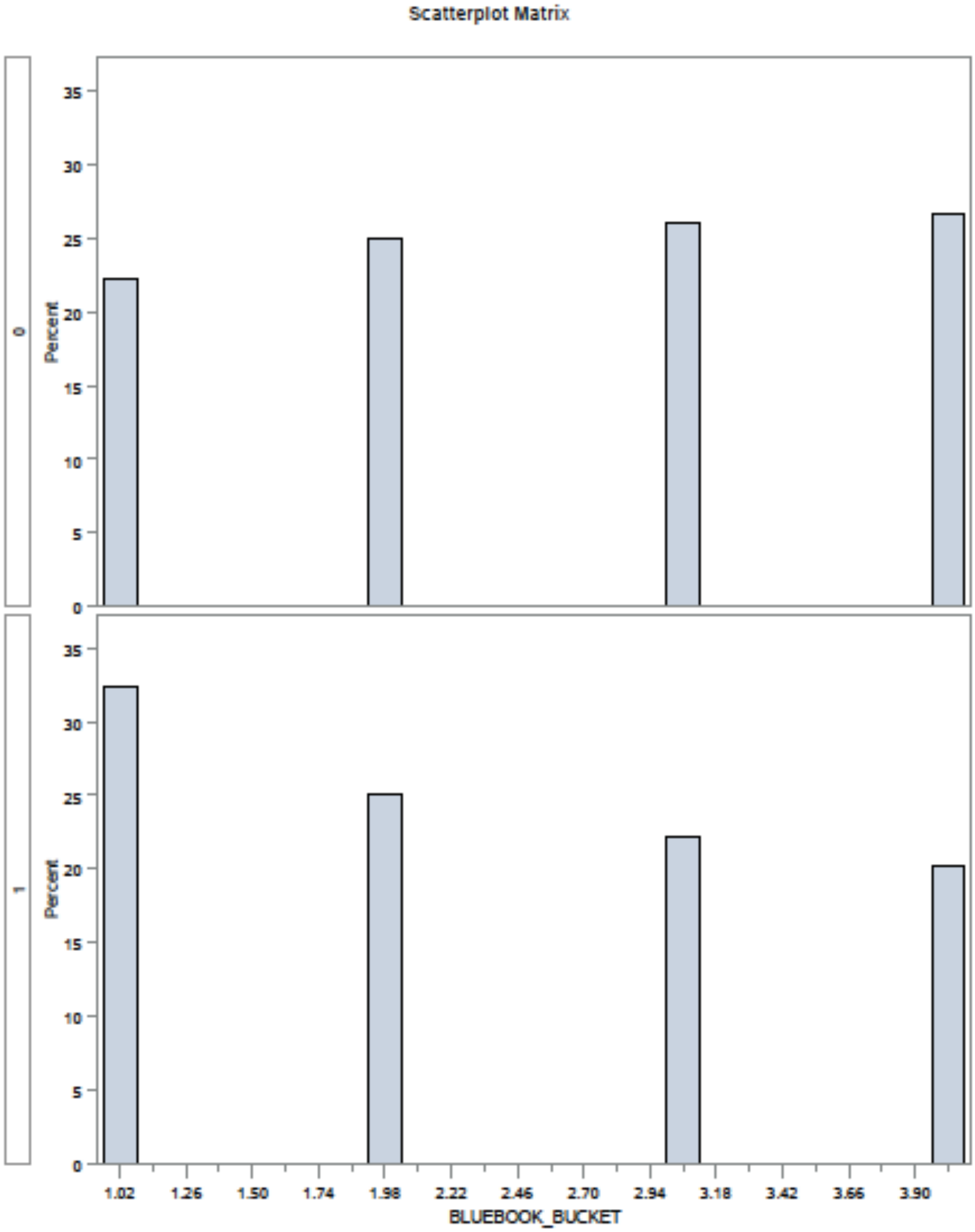
After I created the new variables BLUEBOOK_BUCKET, TIF_BUCKET, IMP_INCOME_BUCKET and HAS_CLMD, I was interested in whether or not these new variables are predictive of TARGET_FLAG. Therefore, I plotted a comparative display of the histograms of the variables group by TARGET_FLAG = 0 and TARGET_FLAG = 1 to see if there were any noticeable difference between the two groups that would imply that the variable is in fact predictive.



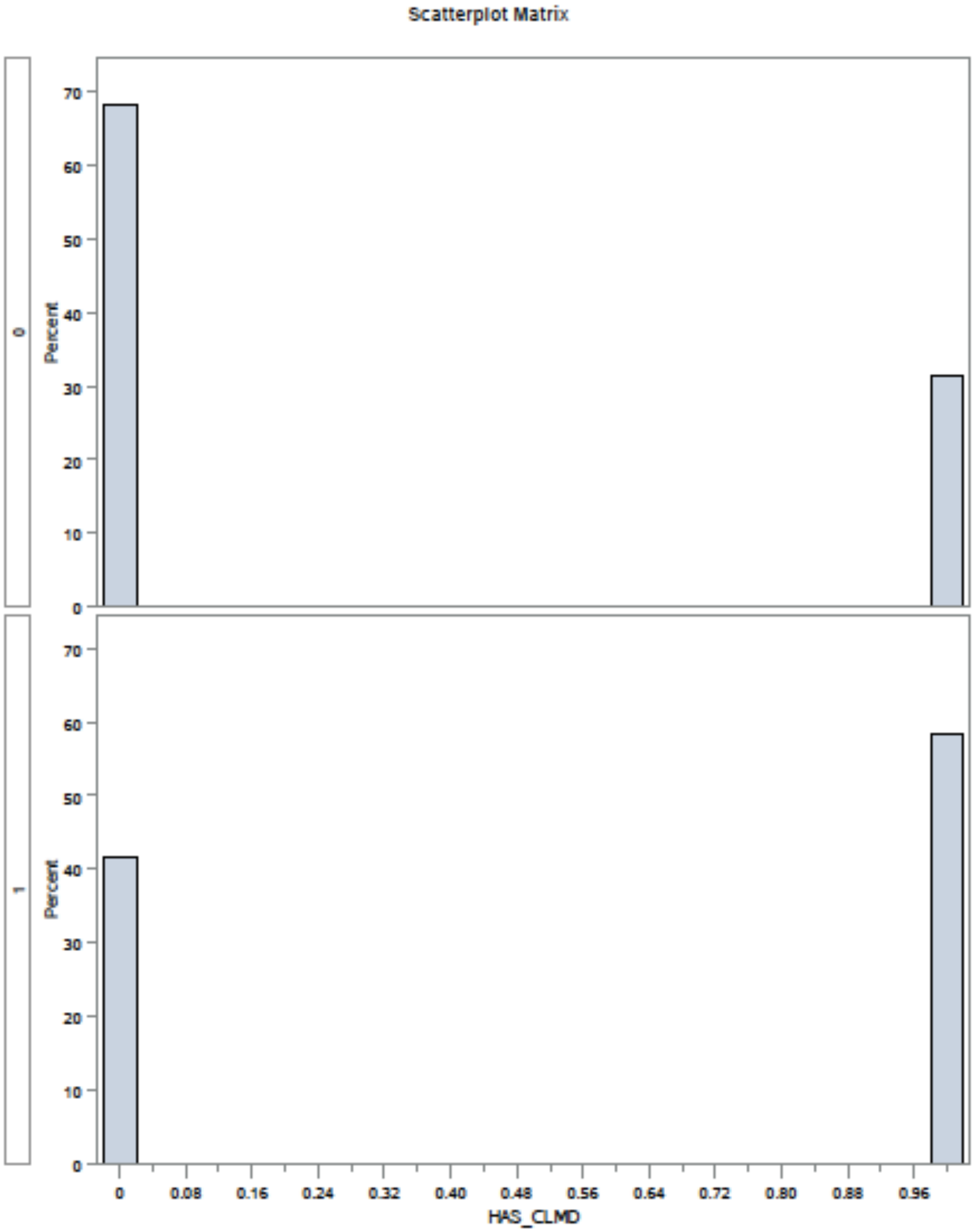
Using Donald Wedding's PROC EYEBALL procedure, we see that in the first bucket, which represents the people who have been customers for the shortest amount of time i.e. whose TIF (time in force) equal to a year, more people get in car crashes (TARGET_FLAG = 1). For the middle two buckets there is less of a difference between the people who get into a crash and those who do not, but in the last bucket which represents the people who have been customers of the insurance company for over 7 years, we see that more people do not get in crashes (TARGET_FLAG = 0). Based off these observations we see that TIF_BUCKET is likely predictive and it follows the theoretical effect of the original variable TIF (time in force) from the data dictionary which is that people who have been customers for a long time are usually more safe.



Again, using PROC EYEBBALL as before, as IMP_INCOME_BUCKET, or the values of the ranges of the variable INCOME that we used for the buckets, increases there are more people that do not crash than the people in the same income bucket that do crash. So, putting all other factors aside, based on our data, we see that the theoretical effect that rich people tend to get into fewer crashes of the original variable INCOME from the data dictionary is still reasonable.

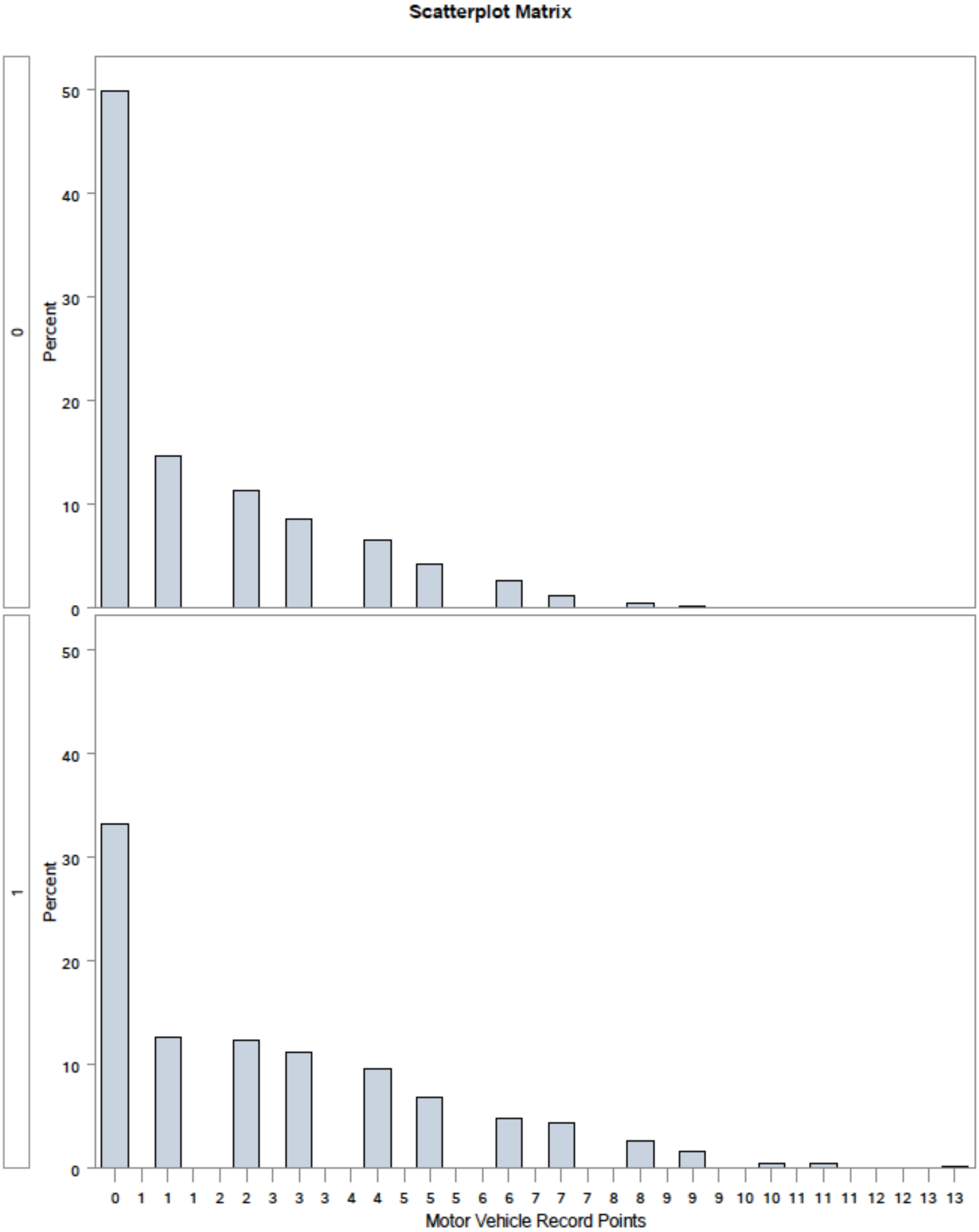


Similar to the buckets for income, we see that as the buckets for BLUEBOOK which is the value of the vehicle increase, the number of people in the same bucket that crash is lower than those that do not. While the theoretical effect was on our TARGET_FLAG variable that tells whether or not the person crashes is unknown according to the data dictionary, it seems reasonable to believe that, disregarding all other factors, as the value of the vehicle increases, the people crash less.



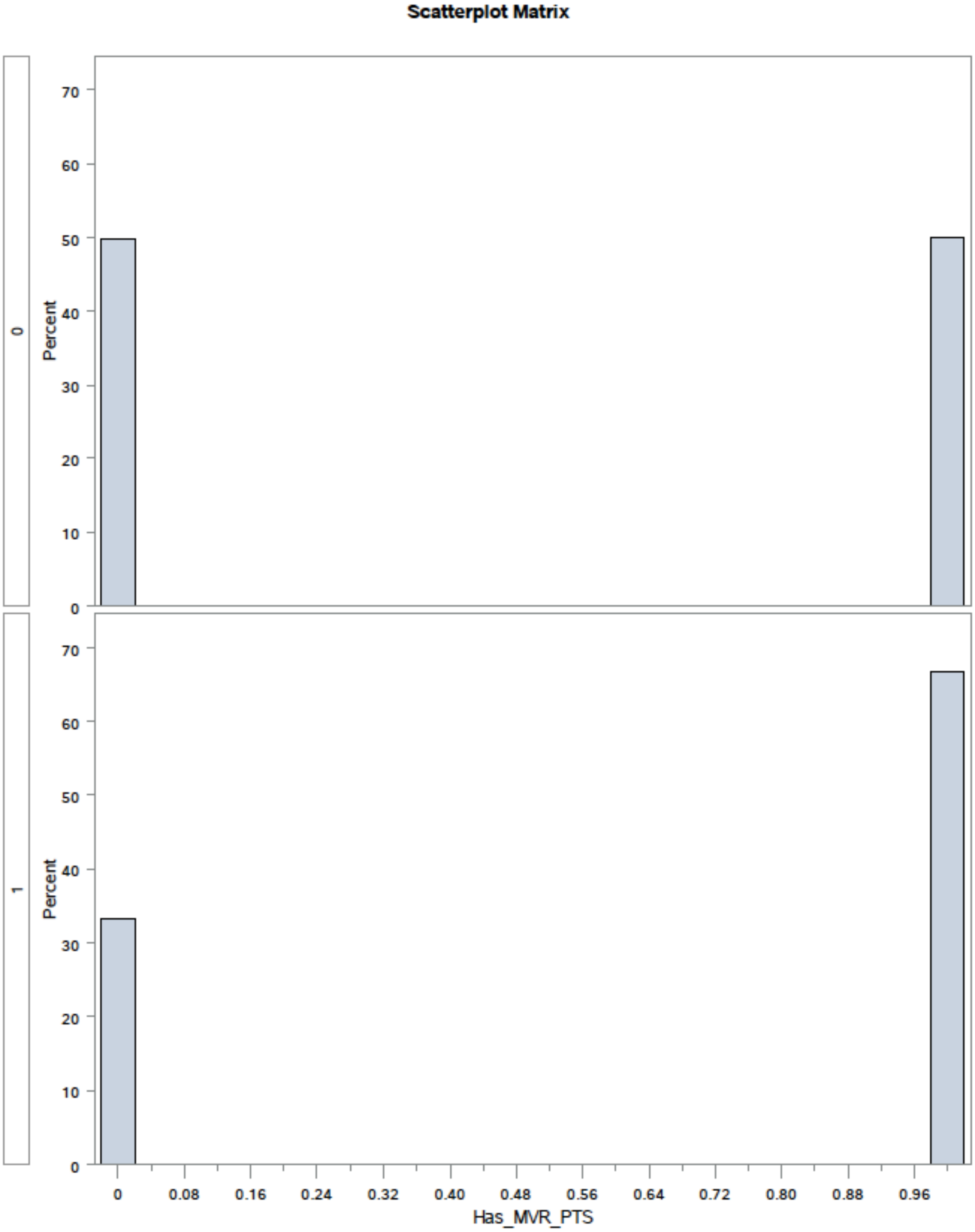
With a quick look at the histograms for HAS_CLMD, it is obvious that of the people who have not had a previous claim in the last five years, more of them do not crash. Conversely, in the group of people who have had at least one claim in the last five years, more of them crash than do not. This makes sense as well as follows the theoretical effect of the original variable CLM_FREQ in that people who have had a claim in the last five years are more likely to file a claim (for a car crash) in the future.

After creating a model with these new variables I still wanted to improve my model. Looking so I took another look at the histograms of the variables like I did above and saw something interesting about MVR_PTS.



Looking at the distribution for motor vehicle record points, or the number of driving tickets the person has received, we see that the largest difference between the number of people that do crash and do not crash are the people who have 0 motor vehicle record points (or 0 tickets). This shows that of people who do not have any tickets, more people do not crash.

Because the there isn't an overwhelming difference between the rest of the groups of people by number of motor vehicle record points, I decided to create a new variable called HAS_MVR_PTS where HAS_MVR_PTS = 0 if the person does not have any motor vehicle record points and HAS_MVR_PTS = 1 if the person has one or more motor vehicle record points.

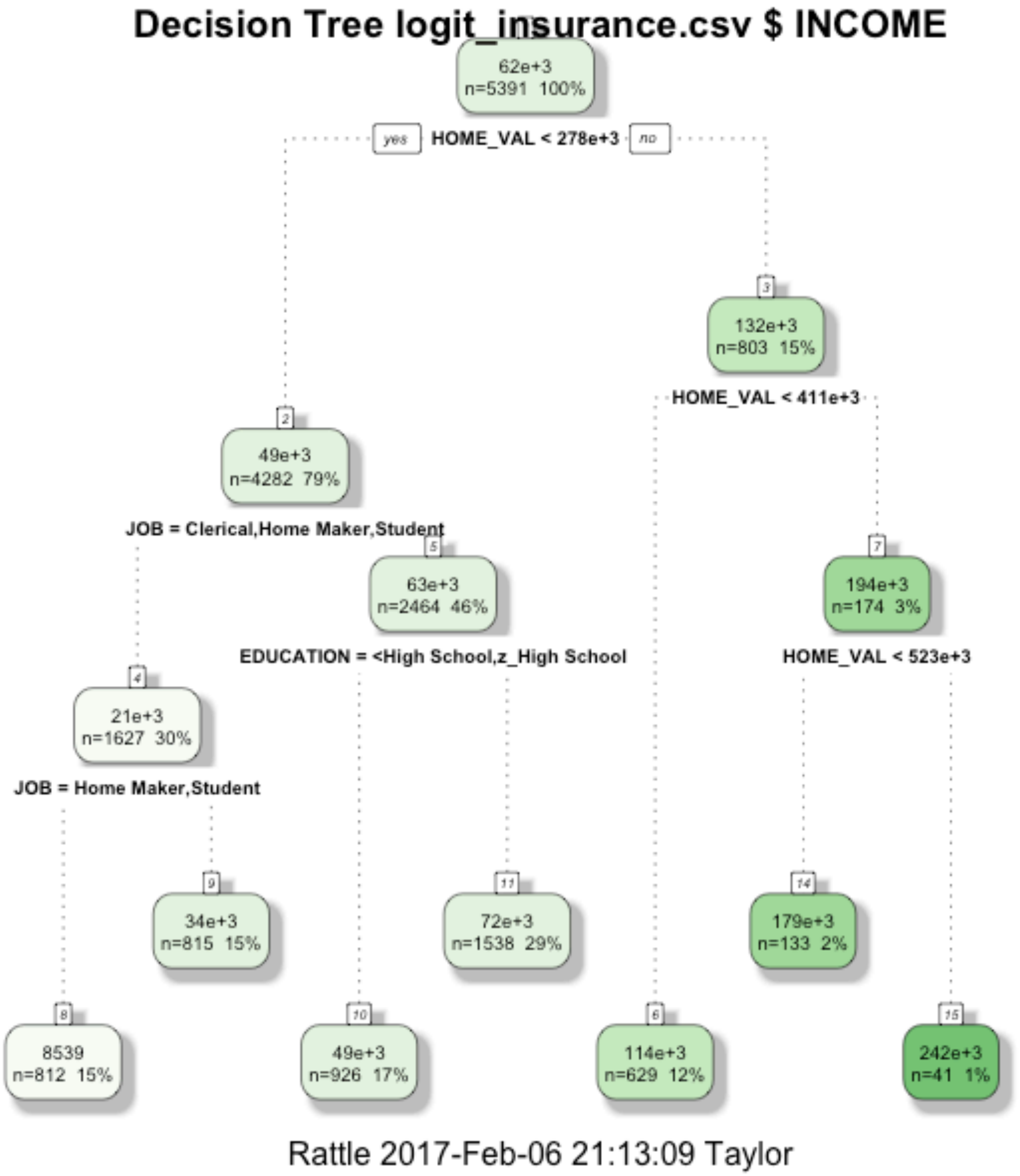


Using PROC EYEBALL, we see the Has_MVR_PTS is predictive since of the people who have gotten a traffic ticket, more of them crash than do not and of the people who have not gotten a ticket more of them do not crash than those that do.

DATA PREPARATION:

As stated in the previous section, AGE, YOJ, INCOME, HOME_VAL, CAR_AGE and JOB all have missing values that must be fixed. Since decision trees are often a good way to predict missing values, I am going to use a decision tree to impute the missing values for the variables INCOME and JOB.

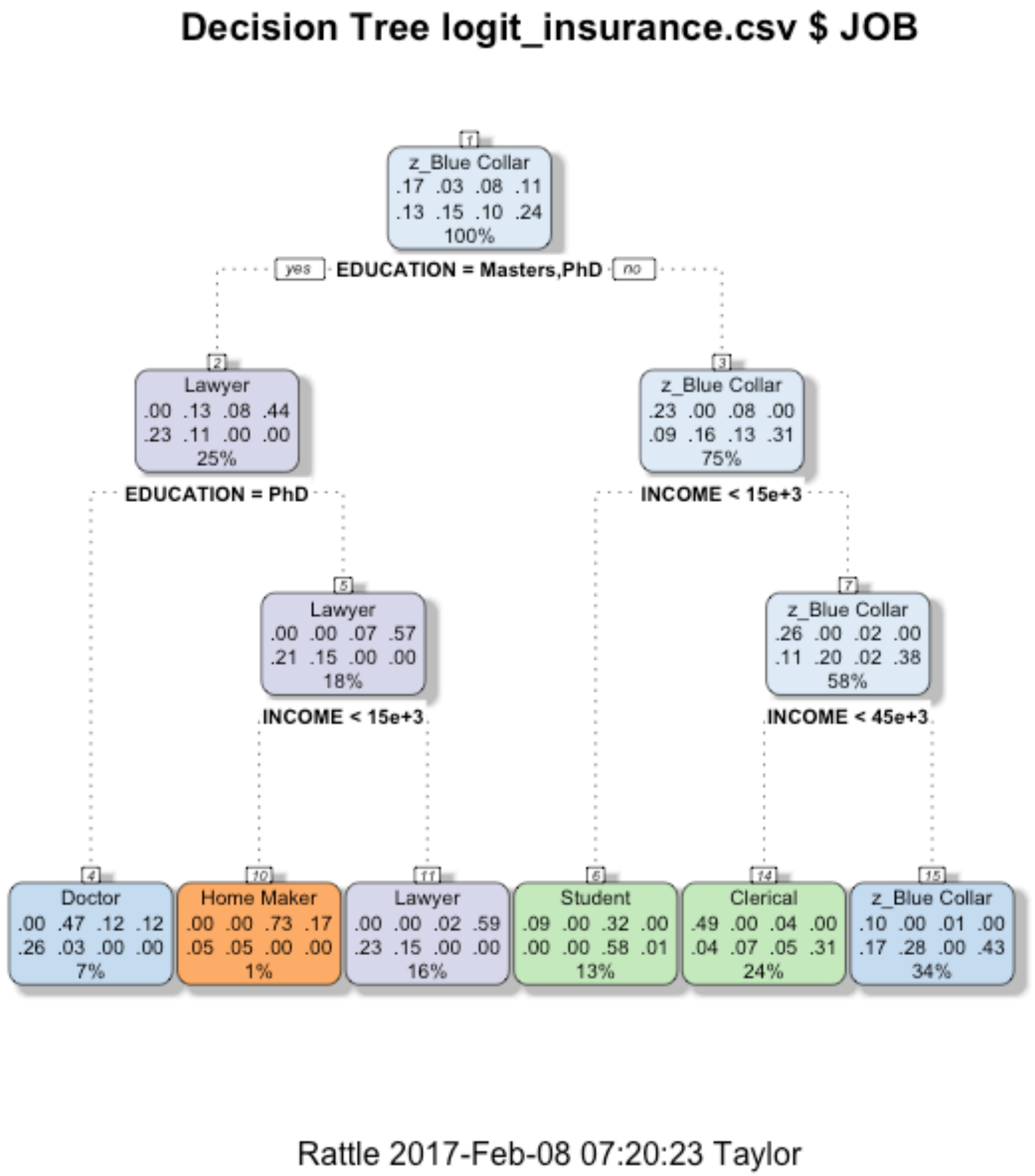
Below is a decision tree with three levels that I created in R in order to predict the missing income values based on the other predictor variable values in the record. I used the rules from the tree to create the imputed variable IMP_INCOME. After I replaced the missing values using this tree, I capped IMP_INCOME and created the variable CAP_IMP_INCOME, using the approach described in the data exploration section.



Decision Tree logit_insurance.csv $ INCOME

Rattle 2017-Feb-06 21:13:09 Taylor

Similarly, I used the tree below to predict the missing JOB values based on the values of the other predictor variables in the record. I used the rules from the tree to create the imputed variable IMP_JOB.



I imputed the missing values of the remaining variables AGE, YOJ, HOME_VAL, and CAR_AGE using the median values of each variable. I dropped the original variables and created the new variables IMP_AGE, IMP_YOJ, IMP_HOME_VAL, and IMP_CAR_AGE respectively.

The rest of the transformations that I made to the data are described in the previous section, Data Exploration, because many of these changes were made off of observations I made from the exploration or affected the next steps of exploration. Therefore, I thought that these transformations fit better in the previous section.

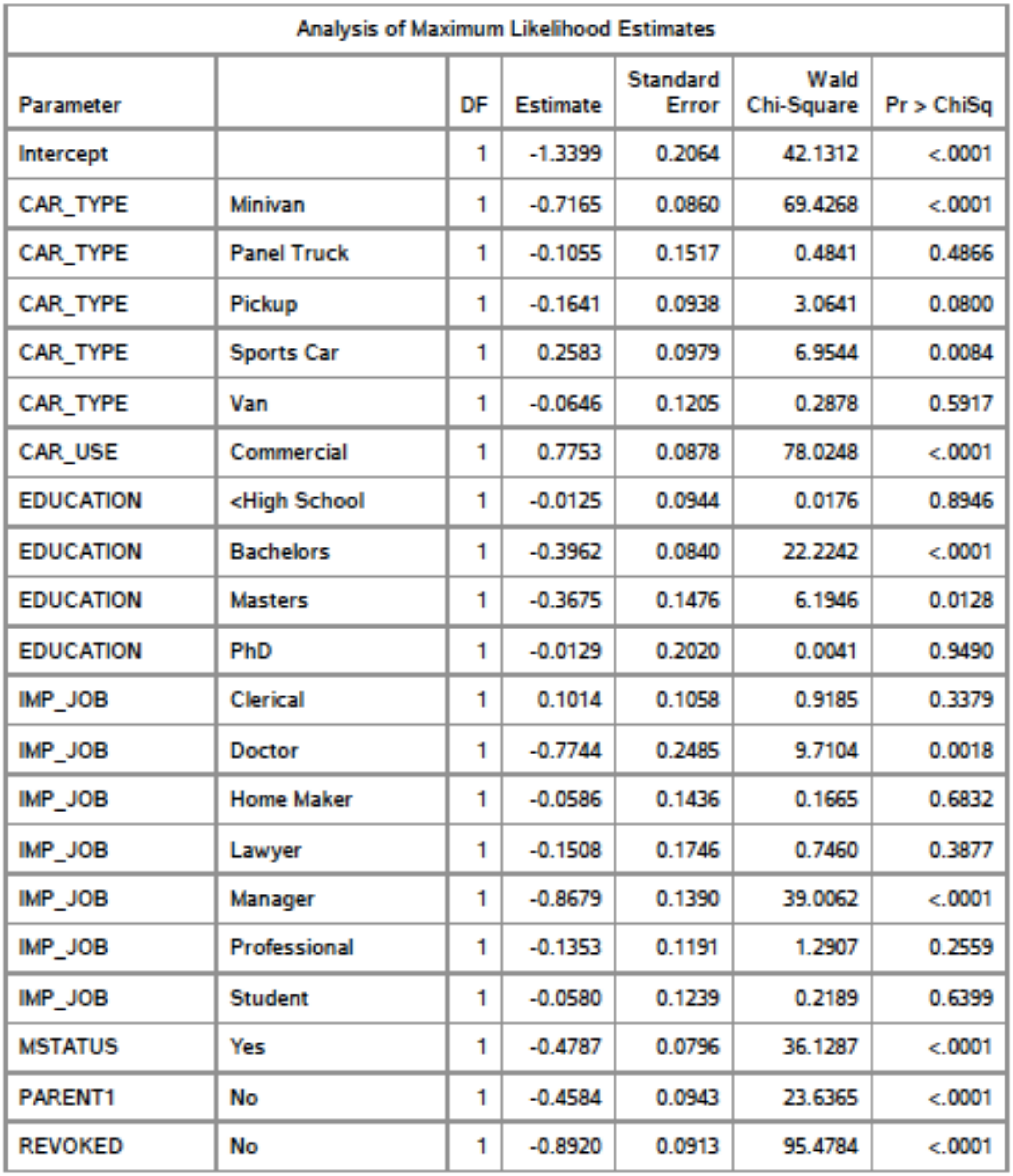
BUILD MODELS:

Unlike linear regression, the estimated coefficients of a logistic regression model cannot be easily interpreted directly. For instance, a logit coefficient of 0.4196 for KIDSDRV implies that the log-odds of the person getting in a crash increases by 0.4196 as the number of kids driving increases by 1. For categorical variables such as CAR_TYPE, each coefficient represents the log-odds of a car crash for the corresponding value of CAR_TYPE, such as Minivan, Pickup, etc. Since an increase or decrease in log-odds isn't something easily interpretable, most people just look at the signs of the coefficients to see if they are make sense and are intuitive. For our models, a negative coefficient implies that the insurance customer is "safer" and a positive coefficient implies that the customer is riskier, i.e. more likely to have a car crash.

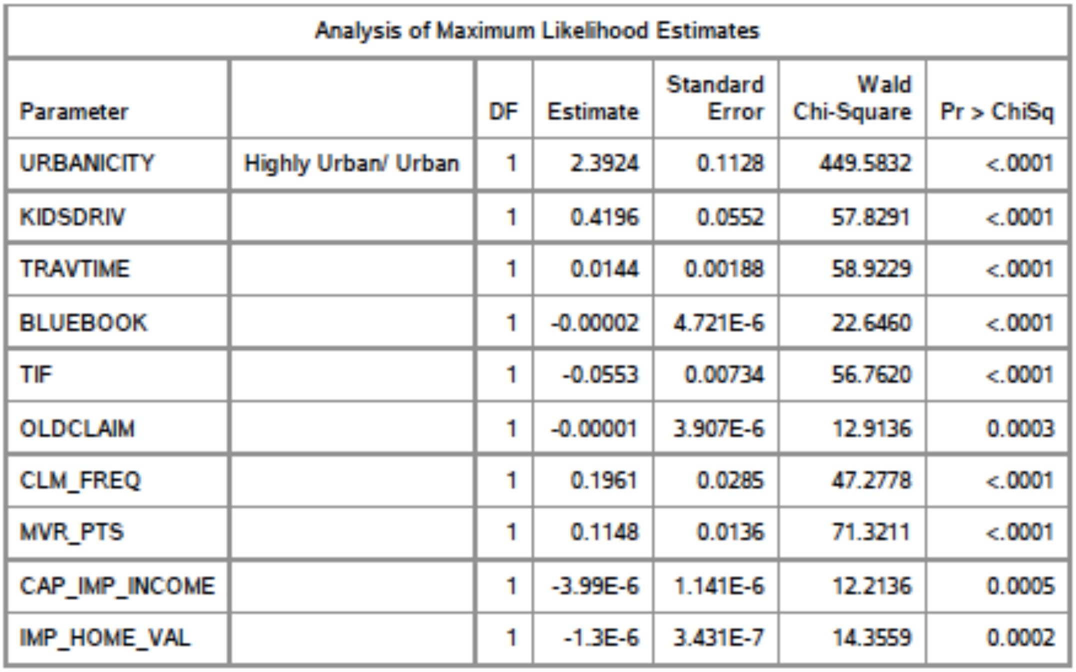
In the Analysis of Maximum Likelihood Estimates tables below, we have coefficient estimates, the estimated standard error of these estimates, and test statistics that test for the null hypothesis that the corresponding coefficient is equal to 0 (implying that the variable is not significant). The test statistic we are given is the Wald Chi-Square, which is calculated by taking the square of each coefficient divided by its standard error. A small p-value implies that the variable (or particular value of a class variable) is significant, rejecting the null hypothesis that the coefficient is equal to 0.

MODEL 1: STEPWISE SELECTION

My first model was built using stepwise selection. Stepwise selection adds variables into the model one by one at each step. Once a new variable is added, SAS will look at all of the variables in the next step to see if, even with the addition of this new variable, the existing variables still have significance in predicting the dependent variable. A benefit of stepwise selection is that once a variable is added, it can be removed in the next step if it is no longer significant. The Analysis of Maximum Likelihood Estimates for this stepwise variable selection model is directly below.



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3399	0.2064	42.1312	<.0001
CAR_TYPE	Minivan	1	-0.7165	0.0860	69.4268	<.0001
CAR_TYPE	Panel Truck	1	-0.1055	0.1517	0.4841	0.4866
CAR_TYPE	Pickup	1	-0.1641	0.0938	3.0641	0.0800
CAR_TYPE	Sports Car	1	0.2583	0.0979	6.9544	0.0084
CAR_TYPE	Van	1	-0.0646	0.1205	0.2878	0.5917
CAR_USE	Commercial	1	0.7753	0.0878	78.0248	<.0001
EDUCATION	<High School	1	-0.0125	0.0944	0.0176	0.8946
EDUCATION	Bachelors	1	-0.3962	0.0840	22.2242	<.0001
EDUCATION	Masters	1	-0.3675	0.1476	6.1946	0.0128
EDUCATION	PhD	1	-0.0129	0.2020	0.0041	0.9490
IMP_JOB	Clerical	1	0.1014	0.1058	0.9185	0.3379
IMP_JOB	Doctor	1	-0.7744	0.2485	9.7104	0.0018
IMP_JOB	Home Maker	1	-0.0586	0.1436	0.1665	0.6832
IMP_JOB	Lawyer	1	-0.1508	0.1746	0.7460	0.3877
IMP_JOB	Manager	1	-0.8679	0.1390	39.0062	<.0001
IMP_JOB	Professional	1	-0.1353	0.1191	1.2907	0.2559
IMP_JOB	Student	1	-0.0580	0.1239	0.2189	0.6399
MSTATUS	Yes	1	-0.4787	0.0796	36.1287	<.0001
PARENT1	No	1	-0.4584	0.0943	23.6365	<.0001
REVOKED	No	1	-0.8920	0.0913	95.4784	<.0001



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
URBANICITY	Highly Urban/ Urban	1	2.3924	0.1128	449.5832	<.0001
KIDSDRIV		1	0.4196	0.0552	57.8291	<.0001
TRAVTIME		1	0.0144	0.00188	58.9229	<.0001
BLUEBOOK		1	-0.00002	4.721E-6	22.6460	<.0001
TIF		1	-0.0553	0.00734	56.7620	<.0001
OLDCLAIM		1	-0.00001	3.907E-6	12.9136	0.0003
CLM_FREQ		1	0.1961	0.0285	47.2778	<.0001
MVR_PTS		1	0.1148	0.0136	71.3211	<.0001
CAP_IMP_INCOME		1	-3.99E-6	1.141E-6	12.2136	0.0005
IMP_HOME_VAL		1	-1.3E-6	3.431E-7	14.3559	0.0002

The logistic regression equation for this model is

LOG_P_TARGET = -1.3399

+ (BLUEBOOK * -0.00002)

+ ((CAR_TYPE in ("Minivan")) * -0.7165)

+ ((CAR_TYPE in ("Panel Truck")) * -0.1055)

+ ((CAR_TYPE in ("Pickup")) * -0.1641)

+ ((CAR_TYPE in ("Sports Car")) * 0.2583)

+ ((CAR_TYPE in ("Van")) * -0.0646)

+ ((CAR_USE in ("Commercial")) * 0.7753)

+ ((EDUCATION in ("<High School")) * -0.0125)

+ ((EDUCATION in ("Bachelors")) * -0.3962)

+ ((EDUCATION in ("Masters")) * -0.3675)

+ ((EDUCATION in ("PhD")) * -0.0129)

+ ((MSTATUS in ("Yes")) * -0.4787)

+ ((PARENT1 in ("No")) * -0.4585)

```
   + ((REVOKED in ("No")) * -0.8920)

   + ((URBANICITY in ("Highly Urban/ Urban")) * 2.3924)

   + ((IMP_JOB in ("Clerical")) * 0.1014)

   + ((IMP_JOB in ("Doctor")) * -0.7744)

   + ((IMP_JOB in ("Home Maker")) * -0.0586)

   + ((IMP_JOB in ("Lawyer")) * -0.1508)

   + ((IMP_JOB in ("Manager")) * -0.8679)

   + ((IMP_JOB in ("Professional")) * -0.1353)

   + ((IMP_JOB in ("Student")) * -0.0580)

   + (KIDSDRIV * 0.4196)

   + (TRAVTIME * 0.0144)

   + (TIF * -0.05533)

   + (OLDCLAIM * -0.00001)

   + (CLM_FREQ * 0.1961)

   + (MVR_PTS * 0.1148)

   + (CAP_IMP_INCOME * -0.00000399)

   + (IMP_HOME_VAL * -0.0000013)
```

This equation predicts the LOGIT (log-odds) of the target. So in order to get our P_TARGET_FLAG that we are interested, first we have to convert LOGIT to odds using the exponential transform, and then convert the odds to the probabilities as follows.

```
if LOG_P_TARGET > 20 then LOG_P_TARGET = 20;

if LOG_P_TARGET < -20 then LOG_P_TARGET = -20;

exp_transform = exp(LOG_P_TARGET);

P_TARGET_FLAG = (exp_transform / ( 1+exp_transform));
```

The check for the values of LOG_P_TARGET is there to ensure that we do not get a number that is too large for the computer to handle.

First we will start by looking at the estimates for the class variables. For the variable CAR_TYPE, the coefficients for each particular value of CAR_TYPE are negative, implying that the customer is safer, except for CAR_TYPE equal to Sports Car. This seems pretty intuitive because I would expect someone driving a Sports Car to partake in riskier driving behaviors, such as speeding, making them more likely to get in a car crash. However, looking at the p-values for these variables, CAR_TYPE equal to Minivan has a really small p-value <.0001 and Sports Car has the second smallest p-value of .0084. This implies that CAR_TYPE equal to Sports Car and Minivan are the only statistically significant categories of the variable CAR_TYPE at a 5% significance level and that the other categories aren't necessarily predictive since we fail to reject the null hypothesis that the coefficients for each particular value of CAR_TYPE (that is not Sports Car or Minivan) is equal to 0.

For CAR_USE, Commercial has a positive coefficient. This could be because commercial vehicles are driven more, so might increase probability of collision.

For all of the values of EDUCATION, the coefficients are negative. According to the Wald Chi-Square test, EDUCATION equal to Bachelors is the only significant one with a p-value <.0001, so we reject the null hypothesis that the coefficient for EDUCATION equals to Bachelors is equal to 0 and the negative sign implies the customer is safer.

For IMP_JOB, all of the particular values of the variable have a negative coefficient, implying that the person is safer, except for clerical. While I do not know why someone with a clerical job would be riskier, it is important to note that the only statistically significant job is Manager. Since we failed to reject the null hypothesis that the coefficient for IMP_JOB equal to Clerical is equal to zero, it is possible that the positive coefficient for a clerical job could very well be due to chance. However, Manager has a negative coefficient, which implies that managers are safer customers. This could have to do with a certain level of responsibility that a manager has.

MSTATUS equal to Yes has a positive coefficient, which goes along with the theoretical assumption that married people drive more safely.

Parent1 equal to No has a negative coefficient. This implies that a customer is predicted as safer if they are not a single parent. This could be that people in a steady relationship tend to have a higher sense of responsibility and are more careful. It follows the assumptions of MSTATUS.

Revoked equal to No has a negative coefficient. This is intuitive because if your license was revoked in the past 7 years, you probably are a more risky driver. So, if it wasn't revoked, you are probably safer.

URBANICITY equal to Highly Urban/ Urban has a positive coefficient. This makes sense because if your Home/Work Area is in an urban area, there is likely more traffic, which would increase your chances of getting in an accident.

KIDSDRV has a positive coefficient. This is what we would expect since when teenagers drive your car, you are more likely to get into crashes.

TRAVTIME has a positive coefficient. This makes sense because we would expect long drives to work to suggest greater risk.

BLUEBOOK has a negative coefficient, which implies that as the value of the vehicle increases, the driver is safer and less likely to crash. This makes sense because drivers with more expensive cars are probably much more careful driving.

OLDCLAIM has a negative coefficient, which we wouldn't necessarily expect. OLDCLAIM is Total Claims in the past 5 years and we would expect that if the total claims are high, then it's likely the customer is a risky driver. However, the negative coefficient could be do to an interaction between CLM_FREQ, which has a positive coefficient implying that having had a claim in the last 5 years implies riskier driver, and TIF, which has a negative coefficient and we would assume if claims are high that the person has been a customer for a long time (long enough to have several claims).

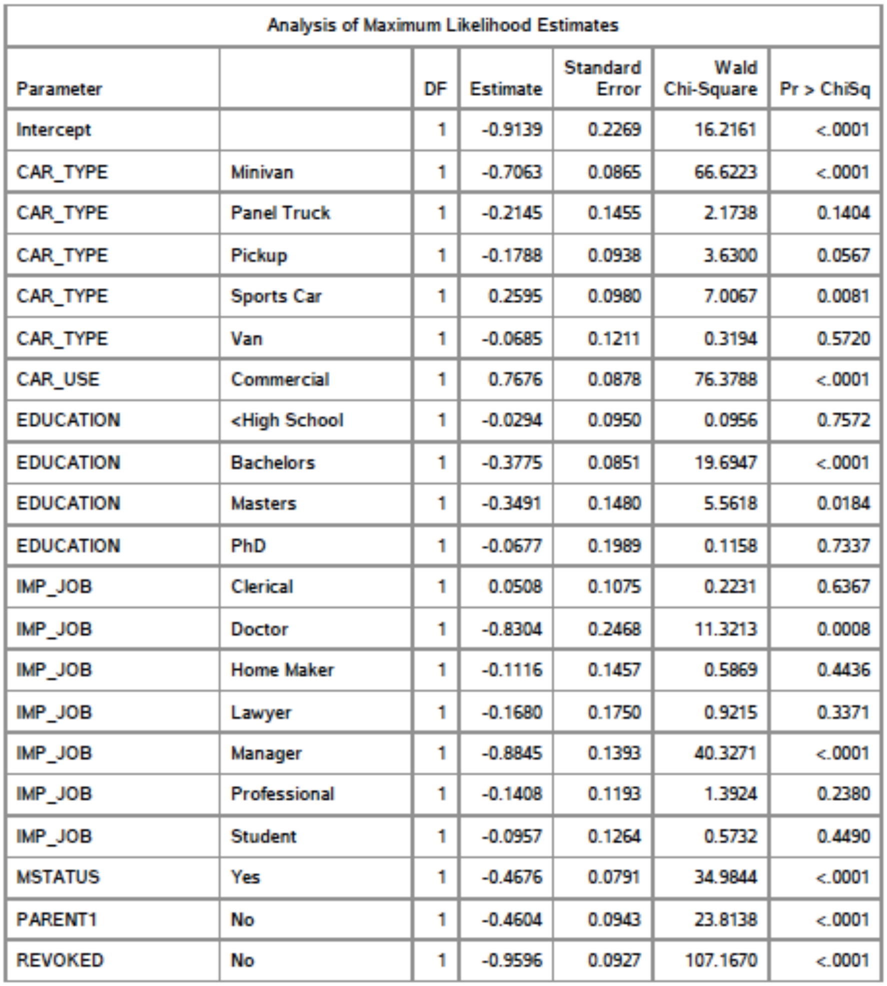
MVR_PTS has a positive coefficient, which we would expect because if you get lots of traffic tickets, you tend to get into more crashes.

CAP_IMP_INCOME and IMP_HOME_VAL both have negative coefficients that imply that people with more money tend to be safer. This makes sense based on our assumptions that people who have good jobs are safer.

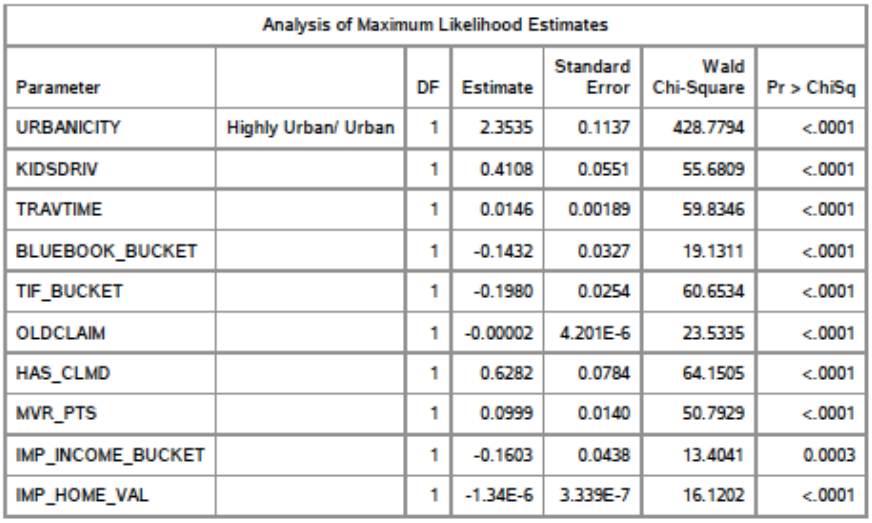
MODEL 2: STEPWISE SELECTION WITH BUCKET VARIABLES (AS NUMERIC)

In order to try to improve the model I got using stepwise variable selection with the variables as given (with the exception of necessary imputations to get rid of missing values), I decided to create new variables from a few of the old variables by grouping the values together and creating "buckets" or categories.

After looking at the distributions of the numerical variables using histograms, I thought that my model would benefit from transforming four of the variables that had particular large ranges by grouping the values into four buckets with cutoffs at the 25th, 50th, and 75th percentiles. The steps that I took to do this, as well as why I decided to group the data into buckets, are explained in the Data Exploration section.



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.9139	0.2269	16.2161	<.0001
CAR_TYPE	Minivan	1	-0.7063	0.0865	66.6223	<.0001
CAR_TYPE	Panel Truck	1	-0.2145	0.1455	2.1738	0.1404
CAR_TYPE	Pickup	1	-0.1788	0.0938	3.6300	0.0567
CAR_TYPE	Sports Car	1	0.2595	0.0980	7.0067	0.0081
CAR_TYPE	Van	1	-0.0685	0.1211	0.3194	0.5720
CAR_USE	Commercial	1	0.7676	0.0878	76.3788	<.0001
EDUCATION	<High School	1	-0.0294	0.0950	0.0956	0.7572
EDUCATION	Bachelors	1	-0.3775	0.0851	19.6947	<.0001
EDUCATION	Masters	1	-0.3491	0.1480	5.5618	0.0184
EDUCATION	PhD	1	-0.0677	0.1989	0.1158	0.7337
IMP_JOB	Clerical	1	0.0508	0.1075	0.2231	0.6367
IMP_JOB	Doctor	1	-0.8304	0.2468	11.3213	0.0008
IMP_JOB	Home Maker	1	-0.1116	0.1457	0.5869	0.4436
IMP_JOB	Lawyer	1	-0.1680	0.1750	0.9215	0.3371
IMP_JOB	Manager	1	-0.8845	0.1393	40.3271	<.0001
IMP_JOB	Professional	1	-0.1408	0.1193	1.3924	0.2380
IMP_JOB	Student	1	-0.0957	0.1264	0.5732	0.4490
MSTATUS	Yes	1	-0.4676	0.0791	34.9844	<.0001
PARENT1	No	1	-0.4604	0.0943	23.8138	<.0001
REVOKED	No	1	-0.9596	0.0927	107.1670	<.0001



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
URBANICITY	Highly Urban/ Urban	1	2.3535	0.1137	428.7794	<.0001
KIDSDRIV		1	0.4108	0.0551	55.6809	<.0001
TRAVTIME		1	0.0146	0.00189	59.8346	<.0001
BLUEBOOK_BUCKET		1	-0.1432	0.0327	19.1311	<.0001
TIF_BUCKET		1	-0.1980	0.0254	60.6534	<.0001
OLDCLAIM		1	-0.00002	4.201E-6	23.5335	<.0001
HAS_CLMD		1	0.6282	0.0784	64.1505	<.0001
MVR_PTS		1	0.0999	0.0140	50.7929	<.0001
IMP_INCOME_BUCKET		1	-0.1603	0.0438	13.4041	0.0003
IMP_HOME_VAL		1	-1.34E-6	3.339E-7	16.1202	<.0001

The logistic regression equation for this model is

```
LOG_P_TARGET = -0.9139
            + ((CAR_TYPE in ("Minivan")) * -0.7063)
   + ((CAR_TYPE in ("Panel Truck")) * -0.2145)
   + ((CAR_TYPE in ("Pickup")) * -0.1788)
   + ((CAR_TYPE in ("Sports Car")) * 0.2595)
   + ((CAR_TYPE in ("Van")) * -0.0685)
   + ((CAR_USE in ("Commercial")) * 0.7676)
   + ((EDUCATION in ("<High School")) * -0.0294)
   + ((EDUCATION in ("Bachelors")) * -0.3775)
   + ((EDUCATION in ("Masters")) * -0.3491)
   + ((EDUCATION in ("PhD")) * -0.0677)
   + ((IMP_JOB in ("Clerical")) * 0.0508)
   + ((IMP_JOB in ("Doctor")) * -0.8304)
   + ((IMP_JOB in ("Home Maker")) * -0.1116)
   + ((IMP_JOB in ("Lawyer")) * -0.1680)
   + ((IMP_JOB in ("Manager")) * -0.8845)
   + ((IMP_JOB in ("Professional")) * -0.1408)
   + ((IMP_JOB in ("Student")) * -0.0957)
   + ((MSTATUS in ("Yes")) * -0.4676)
   + ((PARENT1 in ("No")) * -0.4604)
   + ((REVOKED in ("No")) * -0.9596)
   + ((URBANICITY in ("Highly Urban/ Urban")) * 2.3535)
   + (KIDSDRIV * 0.4108)
   + (TRAVTIME * 0.0146)
   + (BLUEBOOK_BUCKET * -0.1432)
   + (TIF_BUCKET * -0.1980)

   + (OLDCLAIM * -0.00002)

   + (HAS_CLMD * 0.6282)

   + (MVR_PTS * 0.0999)

   + (IMP_INCOME_BUCKET * -0.1603)

   + (IMP_HOME_VAL * -0.00000134)


if LOG_P_TARGET > 20  then LOG_P_TARGET = 20;

if LOG_P_TARGET < -20 then LOG_P_TARGET = -20;

exp_transform = exp(LOG_P_TARGET);

P_TARGET_FLAG = (exp_transform / ( 1+exp_transform));
```

First we will start by looking at the estimates for the class variables. For the variable CAR_TYPE, the coefficients for each particular value of CAR_TYPE are negative, implying that the customer is safer, except for CAR_TYPE equal to Sports Car. This seems pretty intuitive because I would expect someone driving a Sports Car to partake in riskier driving behaviors, such as speeding, making them more likely to get in a car crash. However, looking at the p-values for these variables, CAR_TYPE equal to Minivan has a really small p-value <.0001 and Sports Car has the second smallest p-value of .0081. This implies that CAR_TYPE equal to Sports Car and Minivan are the only statistically significant categories of the variable CAR_TYPE at a 5% significance level and that the other categories aren't necessarily predictive since we fail to reject the null hypothesis that the coefficients for each particular value of CAR_TYPE (that is not Sports Car or Minivan) is equal to 0.

For CAR_USE, Commercial has a positive coefficient. This could be because commercial vehicles are driven more, so might increase probability of collision.

For all of the values of EDUCATION, the coefficients are negative. According to the Wald Chi-Square test, EDUCATION equal to Bachelors is the only significant one with a p-value <.0001, so we reject the null hypothesis that the coefficient for EDUCATION equals to Bachelors is equal to 0 and the negative sign implies the customer is safer.

For IMP_JOBS, all of the particular values of the variable have a negative coefficient, implying that the person is safer, except for clerical. While I do not know why someone with a clerical job would be riskier, it is important to note that the only statistically significant job is Manager, so the positive coefficient for a clerical job could very well be due to chance.

However, Manager has a negative coefficient, which implies that managers are safer customers. This could have to do with a certain level of responsibility that a manager has.

MSTATUS equal to Yes has a positive coefficient which goes along with the theoretical assumption that married people drive more safely.

Parent1 equal to No has a negative coefficient. This implies that a customer is predicted as safer if they are not a single parent. This could be that people in a steady relationship tend to have a higher sense of responsibility and are more careful. It follows the assumptions of MSTATUS.

Revoked equal to No has a negative coefficient. This is intuitive because if your license was revoked in the past 7 years, you probably are a more risky driver. So, if it wasn't revoked, you are probably safer.

URBANICITY equal to Highly Urban/ Urban has a positive coefficient. This makes sense because if your Home/Work Area is in an urban area, there is likely more traffic, which would increase your chances of getting in an accident.

KIDSDRV has a positive coefficient. This is what we would expect since when teenagers drive your car, you are more likely to get into crashes.

TRAVTIME has a positive coefficient. This makes sense because we would expect long drives to work to suggest greater risk.

BLUEBOOK_BUCKET has a negative coefficient, which implies that as the value of the vehicle increases, the driver is safer and less likely to crash. This makes sense because drivers with more expensive cars are probably much more careful driving.

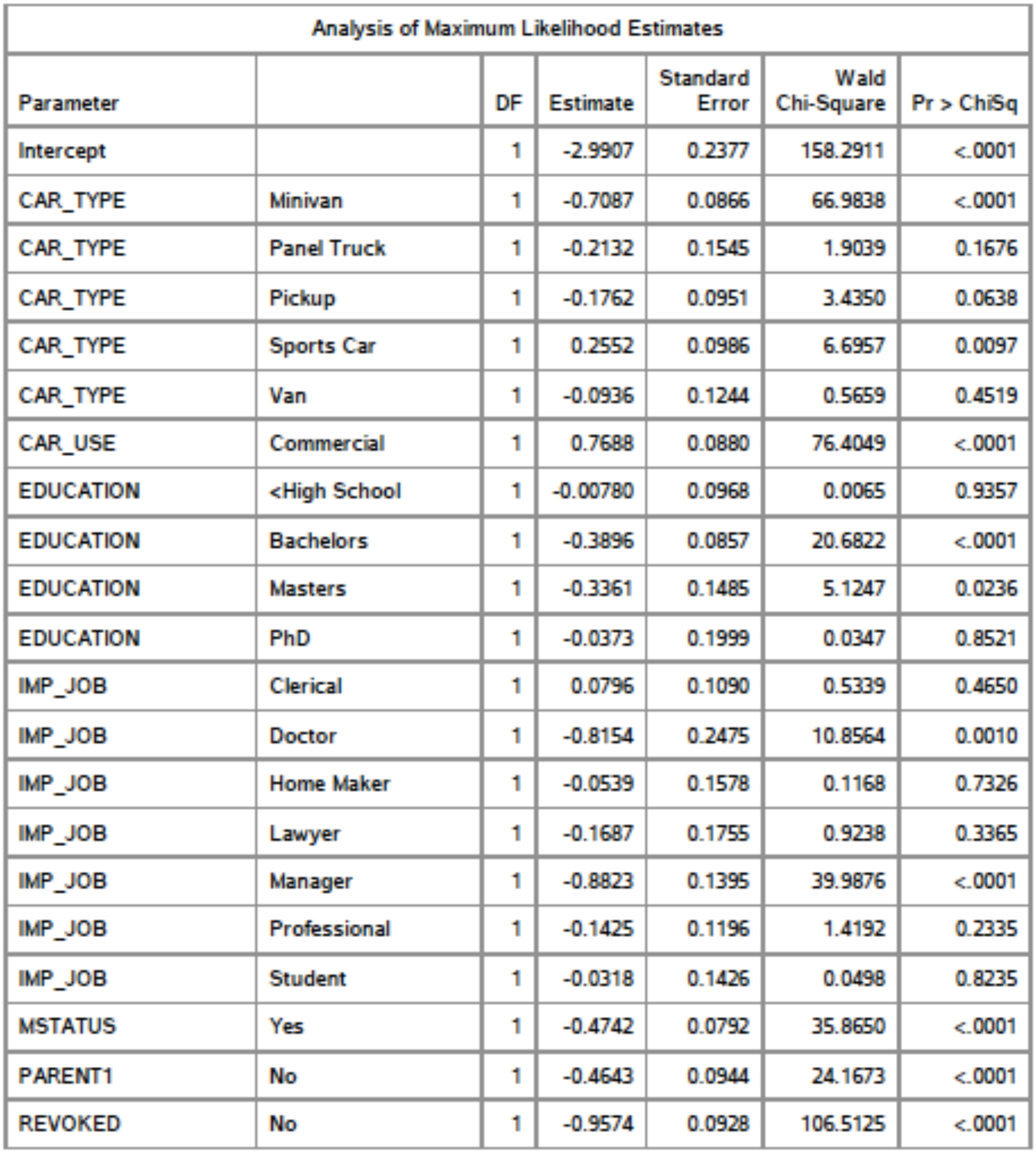
OLDCLAIM has a negative coefficient, which we wouldn't necessarily expect. OLDCLAIM is Total Claims in the past 5 years and we would expect that if the total claims are high, then it's likely the customer is a risky driver. However, the negative coefficient could be do to an interaction between HAS_CLMD, which has a positive coefficient implying that having had a claim in the last 5 years implies riskier driver, and TIF, which has a negative coefficient and we would assume if claims are high that the person has been a customer for a long time.

MVR_PTS has a positive coefficient, which we would expect because if you get lots of traffic tickets, you tend to get into more crashes.

IMP_INCOME_BUCKET and IMP_HOME_VAL both have negative coefficients that imply that people with more money tend to be safer. This makes sense based on our assumptions that people who have good jobs are safer.

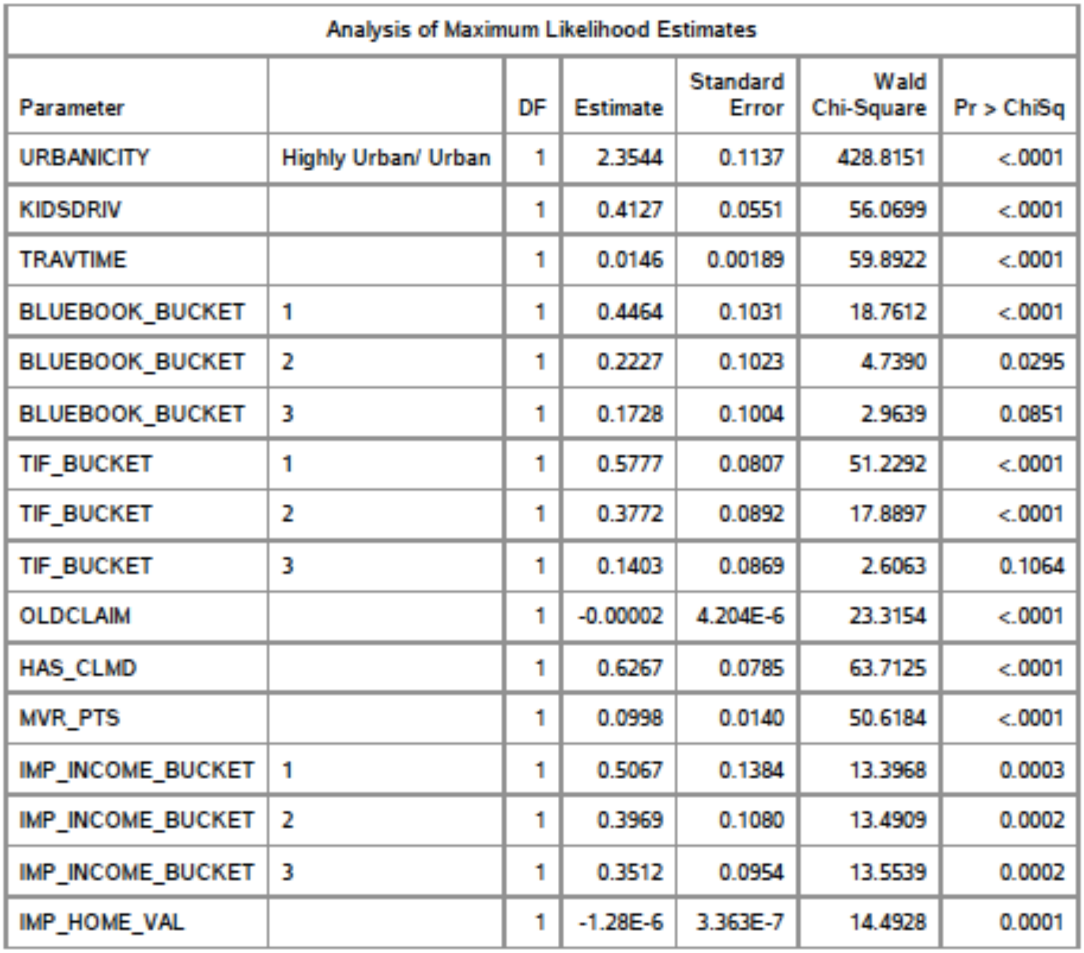
MODEL 3: STEPWISE SELECTION WITH BUCKET VARIABLES AS CLASS VARIABLES

For my next model, I decided to treat the bucket variables as CLASS variables with four categories instead of integers with values 1 through 4. For example, my new data set that I created by adding the bucketed variables contains a variable called IMP_INCOME_BUCKET. IMP_INCOME_BUCKET has integer values 1 to 4 (where 1 denotes the lowest income and 4 denotes the highest income). While in the previous model we treated this variable as a numeric variable, I wanted to try to treat it as a set of categories and see if it improved the model.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.9907	0.2377	158.2911	<.0001
CAR_TYPE	Minivan	1	-0.7087	0.0866	66.9838	<.0001
CAR_TYPE	Panel Truck	1	-0.2132	0.1545	1.9039	0.1676
CAR_TYPE	Pickup	1	-0.1762	0.0951	3.4350	0.0638
CAR_TYPE	Sports Car	1	0.2552	0.0986	6.6957	0.0097
CAR_TYPE	Van	1	-0.0936	0.1244	0.5659	0.4519
CAR_USE	Commercial	1	0.7688	0.0880	76.4049	<.0001
EDUCATION	<High School	1	-0.00780	0.0968	0.0065	0.9357
EDUCATION	Bachelors	1	-0.3896	0.0857	20.6822	<.0001
EDUCATION	Masters	1	-0.3361	0.1485	5.1247	0.0236
EDUCATION	PhD	1	-0.0373	0.1999	0.0347	0.8521
IMP_JOB	Clerical	1	0.0796	0.1090	0.5339	0.4650
IMP_JOB	Doctor	1	-0.8154	0.2475	10.8564	0.0010
IMP_JOB	Home Maker	1	-0.0539	0.1578	0.1168	0.7326
IMP_JOB	Lawyer	1	-0.1687	0.1755	0.9238	0.3365
IMP_JOB	Manager	1	-0.8823	0.1395	39.9876	<.0001
IMP_JOB	Professional	1	-0.1425	0.1196	1.4192	0.2335
IMP_JOB	Student	1	-0.0318	0.1426	0.0498	0.8235
MSTATUS	Yes	1	-0.4742	0.0792	35.8650	<.0001
PARENT1	No	1	-0.4643	0.0944	24.1673	<.0001
REVOKED	No	1	-0.9574	0.0928	106.5125	<.0001



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
URBANICITY	Highly Urban/ Urban	1	2.3544	0.1137	428.8151	<.0001
KIDSDRIV		1	0.4127	0.0551	56.0699	<.0001
TRAVTIME		1	0.0146	0.00189	59.8922	<.0001
BLUEBOOK_BUCKET	1	1	0.4464	0.1031	18.7612	<.0001
BLUEBOOK_BUCKET	2	1	0.2227	0.1023	4.7390	0.0295
BLUEBOOK_BUCKET	3	1	0.1728	0.1004	2.9639	0.0851
TIF_BUCKET	1	1	0.5777	0.0807	51.2292	<.0001
TIF_BUCKET	2	1	0.3772	0.0892	17.8897	<.0001
TIF_BUCKET	3	1	0.1403	0.0869	2.6063	0.1064
OLDCLAIM		1	-0.00002	4.204E-6	23.3154	<.0001
HAS_CLMD		1	0.6267	0.0785	63.7125	<.0001
MVR_PTS		1	0.0998	0.0140	50.6184	<.0001
IMP_INCOME_BUCKET	1	1	0.5067	0.1384	13.3968	0.0003
IMP_INCOME_BUCKET	2	1	0.3969	0.1080	13.4909	0.0002
IMP_INCOME_BUCKET	3	1	0.3512	0.0954	13.5539	0.0002
IMP_HOME_VAL		1	-1.28E-6	3.363E-7	14.4928	0.0001

The logistic regression equation for this model is

LOG_P_TARGET = -2.9907

+ ((CAR_TYPE in ("Minivan")) * -0.7087)

+ ((CAR_TYPE in ("Panel Truck")) * -0.2132)

+ ((CAR_TYPE in ("Pickup")) * -0.1762)

+ ((CAR_TYPE in ("Sports Car")) * 0.2552)

+ ((CAR_TYPE in ("Van")) * -0.0936)

+ ((CAR_USE in ("Commercial")) * 0.7688)

+ ((EDUCATION in ("<High School")) * -0.00780)

+ ((EDUCATION in ("Bachelors")) * -0.3896)

```
+ ((EDUCATION in ("Masters")) * -0.3361)
+ ((EDUCATION in ("PhD")) * -0.0373)
+ ((IMP_JOB in ("Clerical")) * 0.0796)
+ ((IMP_JOB in ("Doctor")) * -0.8154)
+ ((IMP_JOB in ("Home Maker")) * -0.0539)
+ ((IMP_JOB in ("Lawyer")) * -0.1687)
+ ((IMP_JOB in ("Manager")) * -0.88423)
+ ((IMP_JOB in ("Professional")) * -0.1425)
+ ((IMP_JOB in ("Student")) * -0.0318)
+ ((MSTATUS in ("Yes")) * -0.4742)
+ ((PARENT1 in ("No")) * -0.4643)
+ ((REVOKED in ("No")) * -0.9574)
+ ((URBANICITY in ("Highly Urban/ Urban")) * 2.3544)
+ (KIDSDRIV * 0.4127)
+ (TRAVTIME * 0.0146)
+ ((BLUEBOOK_BUCKET in ("1")) * 0.4464)
+ ((BLUEBOOK_BUCKET in ("2")) * 0.2227)
+ ((BLUEBOOK_BUCKET in ("3")) * 0.1728)
+ ((TIF_BUCKET in ("1")) * 0.5777)
+ ((TIF_BUCKET in ("2")) * 0.3772)
+ ((TIF_BUCKET in ("3")) * 0.1403)
+ (OLDCLAIM * -0.00002)
+ (HAS_CLMD * 0.6267)
+ (MVR_PTS * 0.0998)
+ ((IMP_INCOME_BUCKET in ("1"))* 0.5067)
+ ((IMP_INCOME_BUCKET in ("1"))* 0.3969)
```

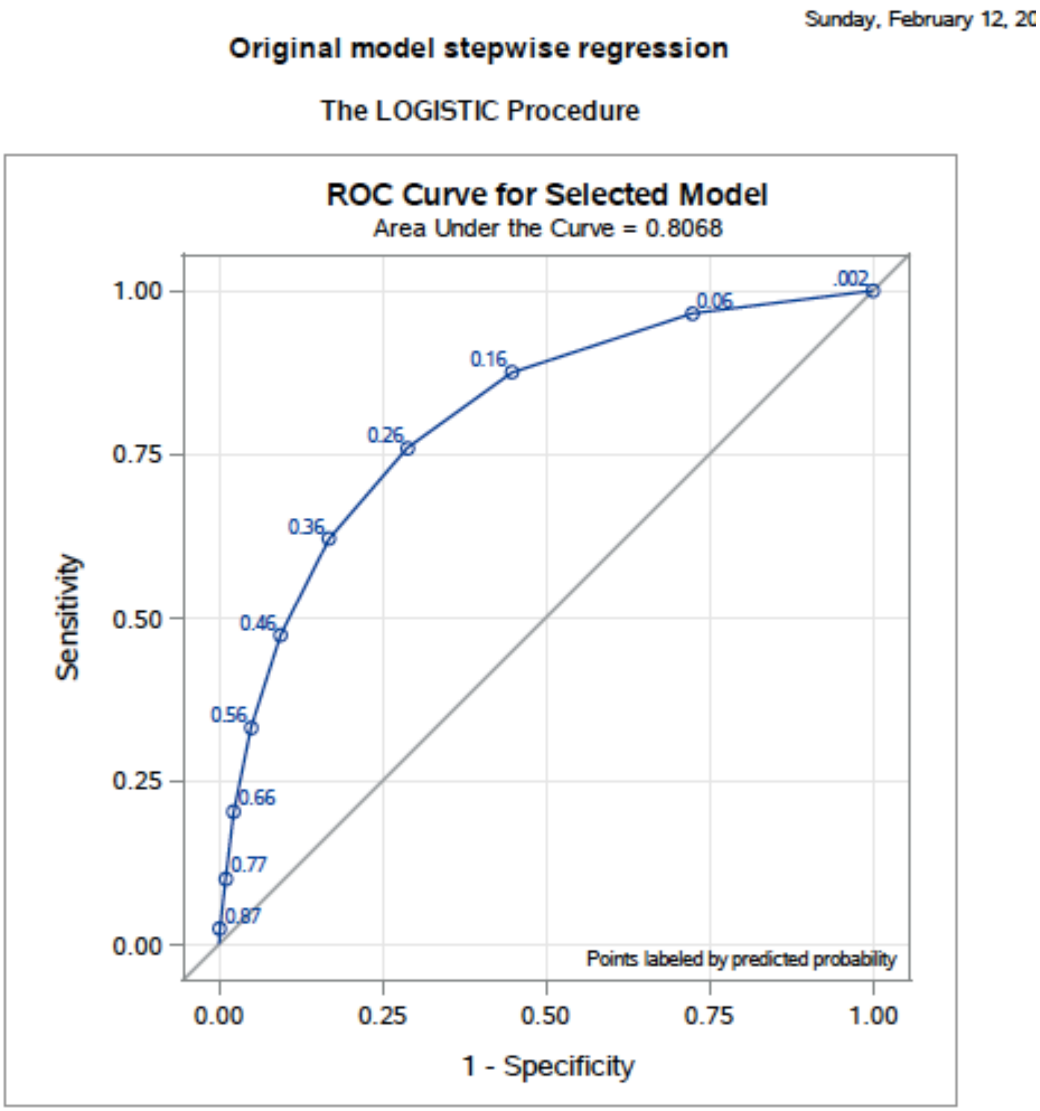
```
   + ((IMP_INCOME_BUCKET in ("1"))* 0.3512)

   + (IMP_HOME_VAL * -0.00000128)


if LOG_P_TARGET > 20  then LOG_P_TARGET = 20;

if LOG_P_TARGET < -20 then LOG_P_TARGET = -20;

exp_transform = exp(LOG_P_TARGET);

P_TARGET_FLAG = (exp_transform / ( 1+exp_transform));
```

While I hoped that making the buckets categories would improve the model, the coefficients are actually pretty confusing. For all of the variables besides the bucket variables, the signs of the coefficients stayed the same as in Model 2 and actually only varied slightly. On the other hand, all if the bucket variable categories have positive coefficients, while in model 2 the bucket variables (not by category) all had negative coefficients. As I explained in the previous model, the negative coefficients made more sense. Therefore, I do not like this model and will not go any further with it.

SELECT MODELS:

For the ROC curves, sensitivity is the proportion of crashes that are correctly predicted, and specificity is the proportion of non-crashes that are correctly predicted. Thus, the y-axis represents the Cumulative True Positive Rate and the x-axis (1-Specificity) represents the Cumulative False Positive Rate. The 45-degree line represents the expected ROC curve for a randomly generated model without any predictive power. So, the more the ROC curve bows out from the 45-degree line, the greater the predictive power of the model. Note that the labels on the ROC curve are the predicted probabilities for the cutpoints.

The Area Under the Curve (AUC) is another metric that is commonly used with the ROC curve. A perfect model (one with 100% accuracy) would have an ROC curve that is the shape of an upside down L that goes straight from the origin to the top left corner and then across the top to the top right corner.



Sunday, February 12, 20

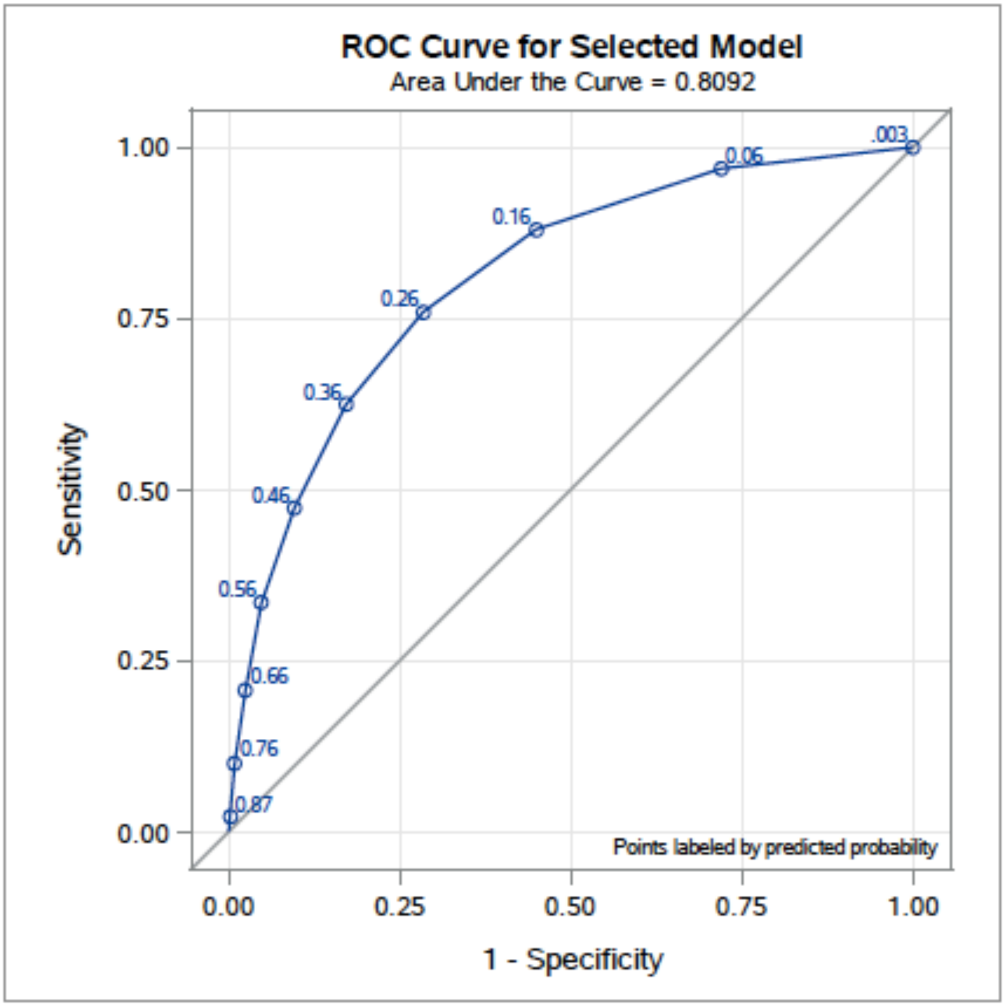
Original model stepwise regression

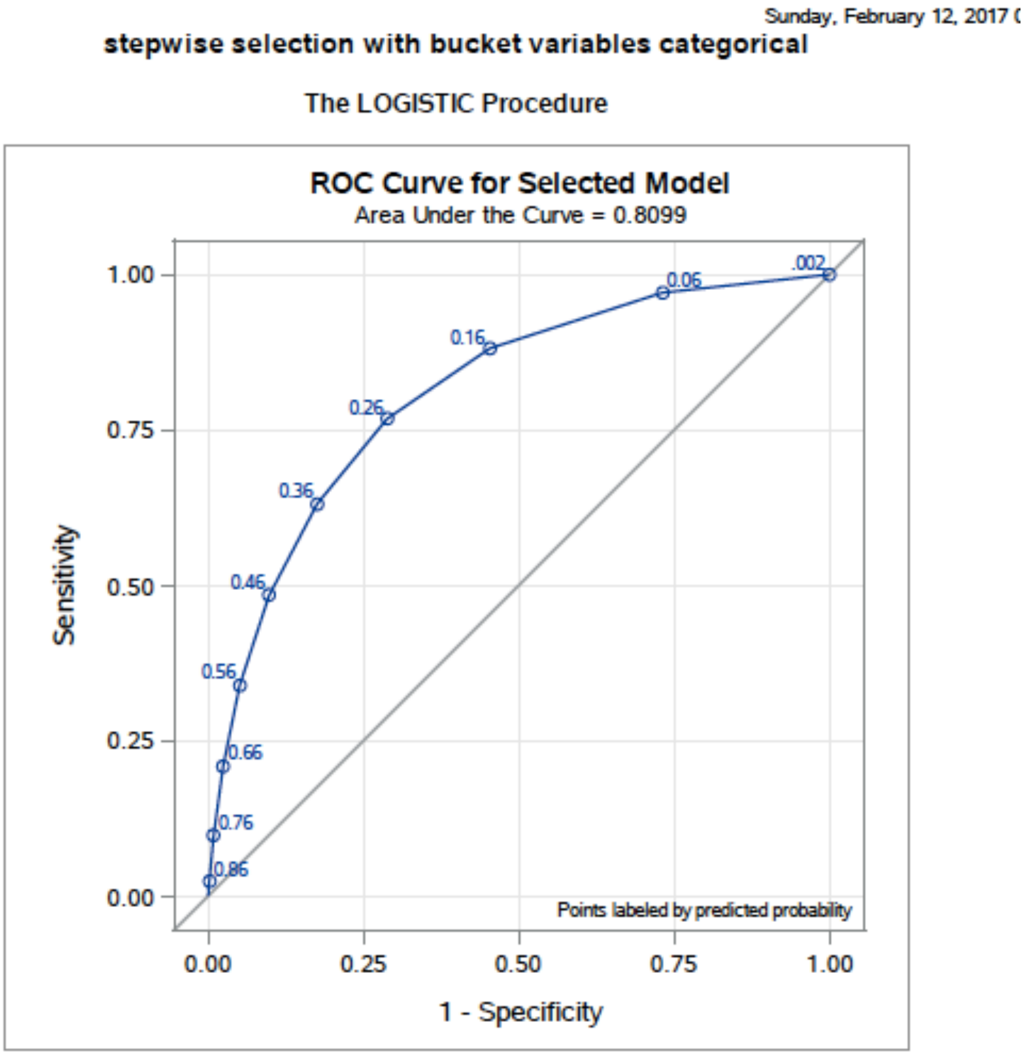
The LOGISTIC Procedure

Sunday, February 12, 2

stepwise selection with bucket variables numerical

The LOGISTIC Procedure





Sunday, February 12, 2017 (

stepwise selection with bucket variables categorical

The LOGISTIC Procedure

For all three models, the ROC curve represents a "Good" model. Each curve has a nice bow out rom the 45-degree line with lift at the beginning, which implies that each model is good at rank ordering and that there is a high chance of predicting true positives. The AUC for the first model is the smallest. Model 3 has the largest AUC at 0.8099, but the AUC for Model 2 is not that far behind at 0.8092.

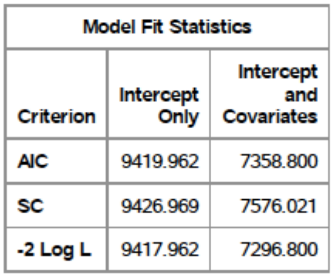
PROC LOGISTIC does not provide the KS test or a straightforward way to get the KS Statistic. The KS Value is the maximum difference between the cumulative true positive rate and the cumulative true negative rate. When comparing two models, the model with the highest KS value is usually preferred. As stated in the homework instructions, we may use GAIN as a good approximation of the KS value. Gain is greatest difference between the predictive model and the random model, which we can approximate visually by the maximum vertical distance between our model ROC curve and the random model 45-degree line. Based on the similarity of the curves, it seems as though there is little to no difference in the Gain between the models.

In each table below, Model Fit Statistics are displayed for two different models, a model with an intercept and no predictor variables (covariates) and a model that includes all of the predictors (covariates) included in the model of interest. Since we built three different models and are comparing their relative goodness of fit, there are three tables, one for each model.

For -2 Log L criterion, a lower value implies better fit. So using this criterion, we see that our second model, which is the stepwise selection model with bucket variables as numeric, is our best model. It is important to note that the value of this statistic is highly dependent on the number of observations. For this example, all of our models were fit to the same data, with the same number of observations, so this is not an issue. However, with -2 Log L, a model with more predictor variables is often found as a better fit. Due to this bias, -2 Log L is not my fit statistic of choice.

The other two statistics shown in the tables solve the problem that I mentioned above by penalizing models that have too many predictor variables. The formula for AIC is AIC = -2 Log L + 2k where k is the number of parameters. Since we want the AIC to be small, the 2k factor penalizes for the number of predictors. Similarly, the formula for SC is SC = -2 Log L + k Log n. So, the SC has an even greater penalty for additional parameters. Both of these statistics are good for comparing models with different numbers of predictor variables. We see that model 2 has the lowest SC and AIC of the three models, so based on these Model Fit Statistics, I choose Model 2 as the best model.

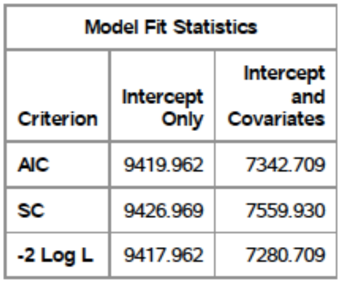
Model 1:



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7358.800
SC	9426.969	7576.021
-2 Log L	9417.962	7296.800

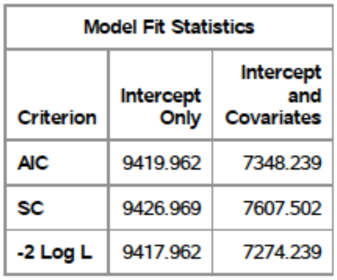
Model 2:



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7342.709
SC	9426.969	7559.930
-2 Log L	9417.962	7280.709

Model 3:



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7348.239
SC	9426.969	7607.502
-2 Log L	9417.962	7274.239

While these statistics give us an informal way to compare models, it is also important to use your judgment when using a best model. With that said, I would definitely select a model with slightly worse performance if it makes more sense or is more parsimonious, as long as it checks out(within reason) based on other criterion. However, it turns out that Model 2 is my best performing model and is also the model I think makes the most sense. The original stepwise model was a little too loose because I gave SAS the freedom to choose from any of the variables in the data set. With the bucket variables, I grouped values for 4 of the variables into buckets that I thought made more sense for the rest of the data to build a predictive model and created 4 more, less value specific, variables. While I did the same thing for the Model 3, I did not like how the coefficients for the bucket variables came out when I used those variables as class variables, so I definitely do not prefer Model 3.

STAND ALONE SCORING PROGRAM:

Below is the scoring program for Model 2, which is our “best” model.

```
data BUCKETSCORE;

set &LIB.logit_insurance_test;


IMP_AGE = AGE;

if missing( IMP_AGE ) then IMP_AGE = 45.0;


IMP_YOJ = YOJ;

if missing( YOJ ) then IMP_YOJ = 11.0;


IMP_INCOME = INCOME;

if missing( INCOME ) then do;

        IMP_INCOME = 62000.00;
        if IMP_HOME_VAL < 278000.00 then IMP_INCOME = 49000.00;
        if IMP_HOME_VAL < 278000.00 AND (JOB = 'Home Maker' or JOB = 'Student')
        then IMP_INCOME = 8539.00;
        if IMP_HOME_VAL < 278000.00 AND JOB = 'Clerical' then IMP_INCOME = 34000.00;
        if IMP_HOME_VAL < 278000.00 AND JOB not in('Home Maker','Student','Clerical')
AND
        EDUCATION in('<High School','z_High School') then IMP_INCOME = 49000.00;
        if IMP_HOME_VAL < 278000.00 AND JOB not in('Home Maker','Student','Clerical')
        AND EDUCATION not in('<High School','z_High School') then IMP_INCOME =
72000.00;


        if 41000.00>IMP_HOME_VAL >= 278000.00 then IMP_INCOME = 114000.00;
        if 523000.00> IMP_HOME_VAL >= 411000.00 then IMP_INCOME = 179000.00;
        if IMP_HOME_VAL>=523000.00 then IMP_INCOME = 242000.00;
end;




IMP_HOME_VAL = HOME_VAL;
if missing ( HOME_VAL ) then IMP_HOME_VAL = 161159.53;


IMP_CAR_AGE = CAR_AGE;
if missing ( CAR_AGE ) then IMP_CAR_AGE = 8.0;



IMP_JOB = JOB;
if missing(IMP_JOB) then do;
```

```
        IMP_JOB = "z_Blue Collar";

        if EDUCATION = "PhD" then IMP_JOB = "Doctor";

        if EDUCATION = "Masters" and IMP_INCOME < 15000.00 then

                IMP_JOB = "Home Maker";

        if EDUCATION = "Masters" and IMP_INCOME > 15000.00 then

                IMP_JOB = "Lawyer";

        if EDUCATION not in("Masters", "PhD") and IMP_INCOME < 15000.00 then

                IMP_JOB = "Student";

        if EDUCATION not in("Masters", "PhD") and 15000.00< IMP_INCOME < 45000.00
then

                IMP_JOB = "Clerical";

        if EDUCATION not in("Masters", "PhD") and IMP_INCOME > 45000.00 then

                IMP_JOB = "z_Blue Collar";

end;


        CAP_IMP_INCOME = IMP_INCOME;

        IF IMP_INCOME > 217269.4 then CAP_IMP_INCOME = 217269.4;




        BLUEBOOK_BUCKET = BLUEBOOK;

        if BLUEBOOK < 9280 then BLUEBOOK_BUCKET = 1; *25th;

    else if BLUEBOOK < 14440 then BLUEBOOK_BUCKET = 2; *50th;

    else if BLUEBOOK < 20850 then BLUEBOOK_BUCKET = 3; *75th;

    else BLUEBOOK_BUCKET = 4;
```

```
 TIF_BUCKET = TIF;

 if TIF = 1 then TIF_BUCKET = 1; *25th;

        else if TIF <= 4 then TIF_BUCKET = 2; *50th;

        else if TIF <= 7 then TIF_BUCKET = 3; *75th;

        else TIF_BUCKET = 4;


        IMP_INCOME_BUCKET = IMP_INCOME;

if IMP_INCOME < 28300.00 then IMP_INCOME_BUCKET = 1; *25th;

   else if IMP_INCOME < 53600.00 then IMP_INCOME_BUCKET = 2; *50th;

   else if IMP_INCOME < 83300.00 then IMP_INCOME_BUCKET = 3; *75th;

   else IMP_INCOME_BUCKET = 4;


        HAS_CLMD = CLM_FREQ;

if CLM_FREQ = 0 then HAS_CLMD = 0;

                   else HAS_CLMD = 1;


                   drop BLUEBOOK;

                   drop TIF;

                   drop IMP_INCOME;

                   drop CLM_FREQ;

                   drop AGE;

                   drop YOJ;

                   drop INCOME;

                   drop HOME_VAL;

                   drop CAR_AGE;

                   drop JOB;
```

```
                drop RED_CAR;

LOG_P_TARGET = -0.9139
        + ((CAR_TYPE in ("Minivan")) * -0.7063)
  + ((CAR_TYPE in ("Panel Truck")) * -0.2145)
  + ((CAR_TYPE in ("Pickup")) * -0.1788)
  + ((CAR_TYPE in ("Sports Car")) * 0.2595)
  + ((CAR_TYPE in ("Van")) * -0.0685)
  + ((CAR_USE in ("Commercial")) * 0.7676)
  + ((EDUCATION in ("<High School")) * -0.0294)
  + ((EDUCATION in ("Bachelors")) * -0.3775)
  + ((EDUCATION in ("Masters")) * -0.3491)
  + ((EDUCATION in ("PhD")) * -0.0677)
  + ((IMP_JOB in ("Clerical")) * 0.0508)
  + ((IMP_JOB in ("Doctor")) * -0.8304)
  + ((IMP_JOB in ("Home Maker")) * -0.1116)
  + ((IMP_JOB in ("Lawyer")) * -0.1680)
  + ((IMP_JOB in ("Manager")) * -0.8845)
  + ((IMP_JOB in ("Professional")) * -0.1408)
  + ((IMP_JOB in ("Student")) * -0.0957)
  + ((MSTATUS in ("Yes")) * -0.4676)
  + ((PARENT1 in ("No")) * -0.4604)
  + ((REVOKED in ("No")) * -0.9596)
  + ((URBANICITY in ("Highly Urban/ Urban")) * 2.3535)
  + (KIDSDRIV * 0.4108)
  + (TRAVTIME * 0.0146)
```

```
    + (BLUEBOOK_BUCKET * -0.1432)
    + (TIF_BUCKET * -0.1980)
    + (OLDCLAIM * -0.00002)
    + (HAS_CLMD * 0.6282)
    + (MVR_PTS * 0.0999)
    + (IMP_INCOME_BUCKET * -0.1603)
    + (IMP_HOME_VAL * -0.00000134)
    ;

if LOG_P_TARGET > 20  then LOG_P_TARGET = 20;
if LOG_P_TARGET < -20 then LOG_P_TARGET = -20;

exp_transform = exp(LOG_P_TARGET);
P_TARGET_FLAG = (exp_transform / ( 1+exp_transform));



KEEP INDEX;
KEEP P_TARGET_FLAG;

run;



proc print data=BUCKETSCORE(obs=5);
run;

proc means data=BUCKETSCORE n nmiss;
```

```
run; quit;

libname scrlib "/home/taylorvender20180/my_courses/PRED411/UNIT02";

data scrlib.VENDERinsurance_BUCKETSCORE2;
set BUCKETSCORE;
run;

proc export data=scrlib.VENDERinsurance_BUCKETSCORE2 DBMS=csv
outfile='/home/taylorvender20180/my_courses/PRED411/UNIT02/VENDERinsurance_BUCKETSCORE2.csv' replace;
run;
```

CONCLUSION:

From this assignment, it is very clear that created a logistic regression model is a lot more involved than building a linear regression model. Especially since there were some many potential predictor variables in the dataset, we had to do much more EDA, which gets even more complicated since a logistic model assumes a nonlinear relationship between the probability and explanatory variables. While a lot of similar statistics are used in logistic regression as were used in linear regression, the interpretations are slightly more complicated and take a couple extra steps to really get a good analysis. For instance, the coefficients have no intuitive meaning until you start thinking about them in terms of odds rather than probabilities. Or, to take a simpler and less in-depth approach, you can check the coefficients by seeing if the signs make reasonable sense.

In the end, we could have done a lot ore to transform the variables, which could have improved our models greatly. While I did transform some variables from examining their distributions using histograms, the model hardly improved with the addition of these transformed variables. In the future, I would definitely be interested in exploring the data even further to try to improve the performance of my models.