Unit 2 Homework

Insurance Logistic Regression Project

Kevin Wong

Spring 2016

Kaggle name: kevinwong

Bingo Bonus Points Attempted: <u>110</u> (section at end of report)

Introduction

The objective of this assignment is to determine the probability of crashing for customers at an auto insurance company using logistic regression. In order to accomplish this, a dataset of approximately 8000 records will be used. This dataset will be examined in detail and then will be prepared for use in building several predictive models. The models will be compared and as a result a final "best" model will be chosen.

Data Exploration

The first step is to get a good understanding of the data before performing any transformations or model building. The dataset contains 8161 customer records with 26 variables as shown in Table 1. The dataset contains an INDEX variable that denotes the identity of the customer. The TARGET_FLAG variable indicates whether the customer did not crash their car represented by the value "0" or did crash their car represented by the value "1". TARGET_AMT indicates how much money is paid if the customer crashed their car.

The other 23 variables contain information on each customer that include personal information such as age and income, vehicle information such as vehicle age and use, and driving record information such as claims frequency and motor vehicle points. The dataset contains both numeric and categorical values. Each variable has some theoretical impact on the probability of crashing a car. Variables such as KIDSDRIV, CLM_FREQ, TRAVTIME, and MVR_PTS are assumed to have a negative impact by increasing the risk of crashing. Variables such as YOJ, INCOME, and HOME_VAL are assumed to have a positive impact by decreasing the risk of crashing.

Table 1 shows summary statistics for all numeric variables. This table gives a sense of the distribution of the variables. For example, the average age of the customers is about 45 years and the average income of the customers is $61,898. The minimum age is 16 and the maximum age is 81. The lowest income is $0 and the highest income is $367,030.

It also appears that the numeric variables AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE have missing values as indicated by the "N Miss" column. AGE is missing 6 records, YOJ is missing 454 records, INCOME is missing 445 records, HOME_VAL is missing 464 records, and CAR_AGE is missing 510 records. These will be dealt with in the subsequent section called Data Preparation where those missing values will be imputed using the mean and decision trees.

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
INDEX		8161	0	5151.87	5133.00	2978.89	1.0000000	10302.00
TARGET_FLAG		8161	0	0.2638157	0	0.4407276	0	1.0000000
TARGET_AMT		8161	0	1504.32	0	4704.03	0	107586.14
KIDSDRIV	#Driving Children	8161	0	0.1710575	0	0.5115341	0	4.0000000
AGE	Age	8155	6	44.7903127	45.0000000	8.6275895	16.0000000	81.0000000
HOMEKIDS	#Children @Home	8161	0	0.7212351	0	1.1163233	0	5.0000000
YOJ	Years on Job	7707	454	10.4992864	11.0000000	4.0924742	0	23.0000000
INCOME	Income	7716	445	61898.10	54028.17	47572.69	0	367030.26
HOME_VAL	Home Value	7697	464	154867.29	161159.53	129123.78	0	885282.34
TRAVTIME	Distance to Work	8161	0	33.4887972	32.8709696	15.9047470	5.0000000	142.1206304
BLUEBOOK	Value of Vehicle	8161	0	15709.90	14440.00	8419.73	1500.00	69740.00
TIF	Time in Force	8161	0	5.3513050	4.0000000	4.1466353	1.0000000	25.0000000
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	4037.08	0	8777.14	0	57037.00
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0.7985541	0	1.1584527	0	5.0000000
MVR_PTS	Motor Vehicle Record Points	8161	0	1.6955030	1.0000000	2.1471117	0	13.0000000
CAR_AGE	Vehicle Age	7651	510	8.3283231	8.0000000	5.7007424	-3.0000000	28.0000000

Table 1

The categorical (or character) variables are PARENT1, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, RED_CAR, REVOKED, and URBANICITY. These variables contain several levels such as PARENT1 is whether the customer is a single parent or not denoted by the “Yes” or “No” levels.

It appears that the JOB variable is the only categorical variable that contains missing values shown below in Table 2. There are 526 missing values that are not categorized. Again, these missing values will be imputed in the next section.

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

Table 2

Aside from business intuition, it is useful to examine the correlation of the response variable, TARGET_FLAG, with all other variables. The correlations give a sense of which independent variables may be good predictors of the response variable as shown in Table 3 below. Keep in mind that Pearson Correlation Coefficients table can only show numeric variables. The INDEX, TARGET_FLAG, and TARGET_AMT variables can be ignored here as they are reference variables. The criteria for variable consideration will be a p-value of 0.05. All independent variables appear to be statistically significant with the response variable TARGET_FLAG with p-values below 0.0001. In summary, all these numeric variables are statistically viable inputs to our model.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS	CAR_AGE
TARGET_FLAG	-0.00167 0.8801 8161	1.00000 8161	0.53425 <.0001 8161	0.10367 <.0001 8161	-0.10322 <.0001 8155	0.11562 <.0001 8161	-0.07051 <.0001 7707	-0.14201 <.0001 7716	-0.18374 <.0001 7697	0.04815 <.0001 8161	-0.10338 <.0001 8161	-0.08237 <.0001 8161	0.13808 <.0001 8161	0.21620 <.0001 8161	0.21920 <.0001 8161	-0.10065 <.0001 7651

Table 3

It's important to also examine the correlation between independent variables in order to tease out any issues with multicollinearity or if the number of independent variables can be reduced or combined to form interactions. Based on analysis of the matrix of scatterplots between each independent variable not shown here, there are no relationships that are near perfect that would warrant creating an interaction terms.

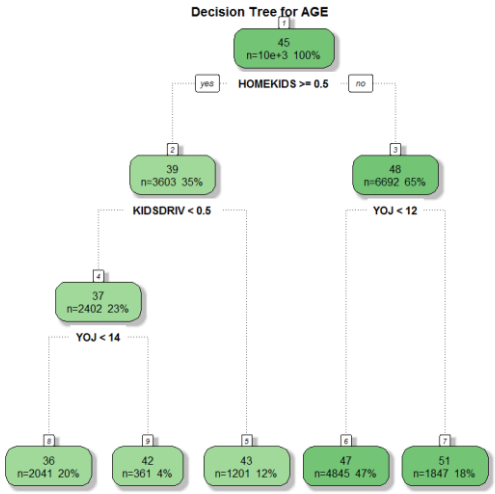
Data Preparation

Imputing Missing Values

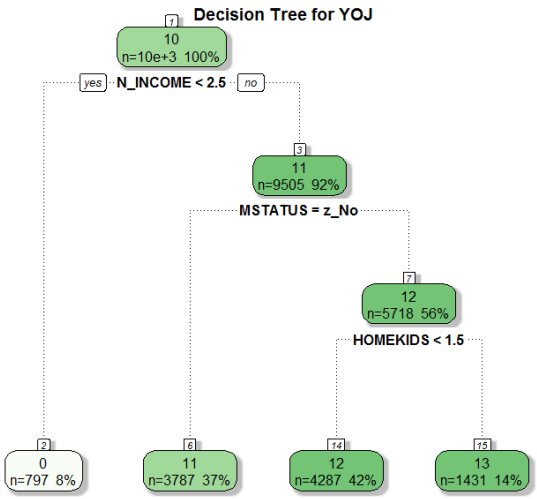
Earlier, it was discovered that six variables contain missing values. Those variables are AGE, YOJ, INCOME, HOME_VAL, CAR_AGE, and JOB. Data can be missing because the customer may have forgotten or chosen to forgo filling in an answer or simply the data was unavailable. Regardless, it is critical to deal with missing values otherwise the models cannot be built.

If independent variables have too many missing values, say greater than 50%, it may be best to remove the entire variable. Fortunately, this is not the case as AGE is only missing 6 records and the other five variables are missing about 500 records out 8161 records which make them candidates for imputation. The method for imputation used is a decision tree. Although it might be good enough to impute the variables with missing values with the mean, the goal is to maximize prediction accuracy so decision trees are a good choice. The following decision trees are generated in the R programming language using the "rpart" and "rattle" packages which help to create nice looking decision trees.

The first decision tree used is for the AGE variable. The interpretation of the tree is if a customer has more than 0.5 kids at home and less than 0.5 kids driving and years on the job is less than 14 then the customer's age is 36. With the same conditions except that if years on job is greater than 14 then the customer's age is 42. While detailed, it does not make sense that someone can have 0.5 kids at home or 0.5 kids driving a car. The tree will end up selecting customers who have 1 or more kids at home and 1 or less kids driving a car. Everything else makes sense as the more years on the job, the customer is likely older in age.

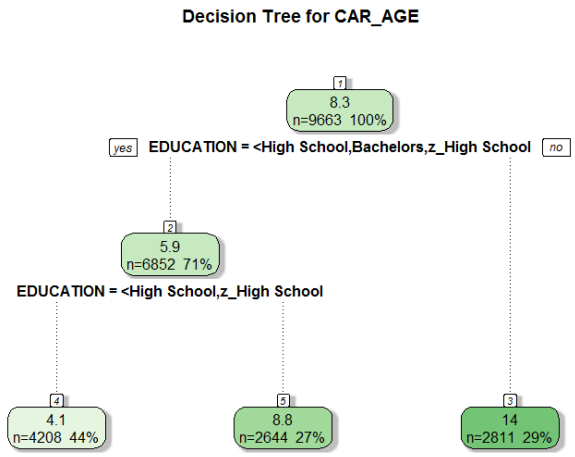


The next decision tree used is for the YOJ variable. This tree can be interpreted as if a customer has less than $2.50 then they probably have worked 0 years. Alternatively, if the customer makes more than $2.50 and they are not married then they've probably worked 11 years on the job. If the customer makes more than $2.50 and is married and has less than 1.5 kids at home then they've been on the job for 12 years. If they have more than 1.5 kids then they've worked 13 years on the job. This makes sense as more income, married status, and more kids likely means more years on the job.

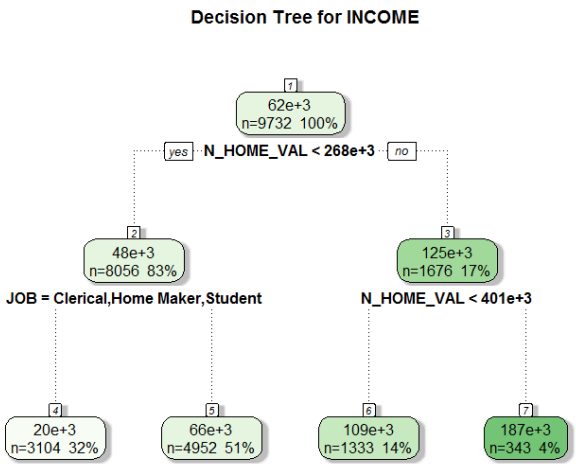


The next decision tree is for the CAR_AGE variable. According to this tree, if a customer's education level is bachelors or less then the age of their car is likely 4.1 years or 8.8 years depending on whether they are education level is high school or lower or at the bachelors level. Alternatively, if the customer's education level is greater than bachelors then the age of their car is likely 14 years. This tree is not entirely intuitive as there can be people who have had cars for longer despite education level. At the same time, this is not entirely counter-intuitive as it's

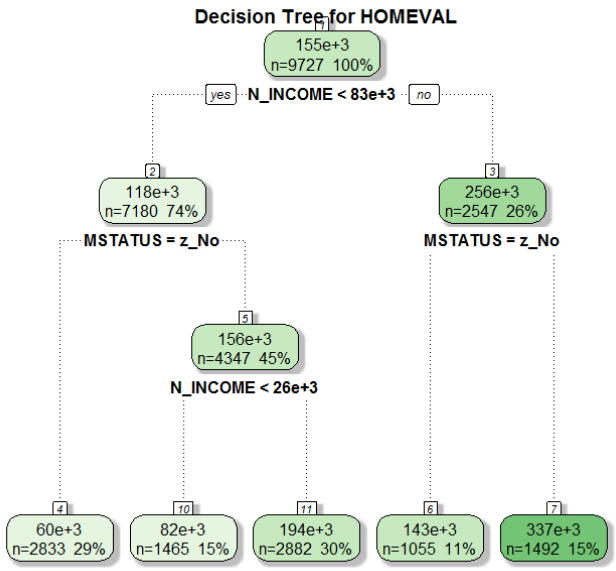
possible someone with higher education level can take better care of a car, so this tree will be used.



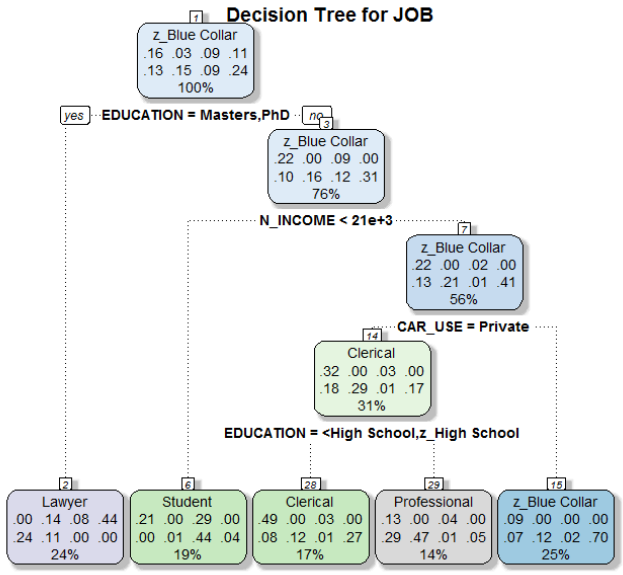
The next decision tree is for the INCOME variable. It is interpreted as if a customer has a home valued less than $268k and they are a clerical, home maker, or student then they make $20k per year. On the other hand, if their home is valued less than $268k and they are not a clerical, home maker, or student then they make $66k per year. Alternatively, if a customer's home is valued more than $268k and less than $401k then they make $109k per year. Lastly, if their home value is greater than $401k then they likely make $187k per year. This decision tree makes perfect sense. Someone who owns a more expensive home tends to make more income. Also, clerical, home maker, and student professions tend to make less than the other professions.



The next decision tree is for HOMEVAL. The tree is interpreted as if a customer makes less than $83k and is not married then their home value is $60k. If the customer makes less than $83k but is married and happens to make less than $26k then their home value is $82k. If the customer makes less than $83k, is married, and makes more than $26k then the home value is $194k. Alternatively, if a customer makes more than $83k and is not married then their home value is $143k. If they make more than $83k, but are married then their home value is $337k. For the most part this decision tree makes sense as higher income individuals tend to have more expensive homes. This tree puts some emphasis in marital status with married individuals having higher valued homes, which generally makes sense.



The last decision tree is for JOB. The decision tree is interpreted as if a customer has a Masters or PhD, then they are likely a lawyer. If the customer does not have a Masters or PhD and makes less than $21k then they are likely a student. If a customer does not have a Masters or PhD and makes more than $21k and has a car for private use and education level is high school or less then their profession is clerical. But if their education level is greater than high school then they are likely a professional. Lastly, if a customer does not have a Masters or PhD and their income is greater than $21k and they use their car for commercial purposes then they are likely a blue collar worker. This decision tree does make sense in general. Higher educated people tend to have a higher job profession like lawyer. People that make little income are likely to be students so the less than $21k income branch makes sense. Also, people who often use their cars for commercial reasons tend to be blue collar workers which again, makes sense.



When variables contain missing data, sometimes the fact that the data is missing can be a good predictor. Flag variables are created for all the variables above where if the data was missing then that flag variable will be "1" and if it is not missing then the flag variable will be given a "0". These flag variables (e.g. AGE_FLAG, YOJ_FLAG, INCOME_FLAG, JOB_FLAG, CAR_AGE_FLAG, and HOME_VAL_FLAG) will be included as part of the model development process.

After imputing all the variables with missing values, the missing values from before are now represented with "0" values in the "N Miss" column. The missing 526 values in the JOB variable from earlier are also non-existent. This indicates that all missing values have been successfully imputed.

Variable	Label	N	N Miss	Mean	Median
TARGET_FLAG		8161	0	0.2638157	0
KIDSDRIV	#Driving Children	8161	0	0.1710575	0
HOMEKIDS	#Children @Home	8161	0	0.7212351	0
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
IMP_AGE		8161	0	44.7851979	45.0000000
AGE_FLAG		8161	0	0.000735204	0
IMP_YOJ		8161	0	10.4799657	11.0000000
YOJ_FLAG		8161	0	0.0556304	0
IMP_CAR_AGE		8161	0	8.3435118	8.0000000
CAR_AGE_FLAG		8161	0	0.0624923	0
IMP_INCOME		8161	0	61663.61	54909.10
INCOME_FLAG		8161	0	0.0545276	0
IMP_HOME_VAL		8161	0	154637.49	160202.35
HOME_VAL_FLAG		8161	0	0.0568558	0
JOB_FLAG		8161	0	0.0644529	0

IMP_JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Clerical	1271	15.57	1271	15.57
Doctor	444	5.44	1715	21.01
Home Maker	641	7.85	2356	28.87
Lawyer	876	10.73	3232	39.60
Manager	1275	15.62	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

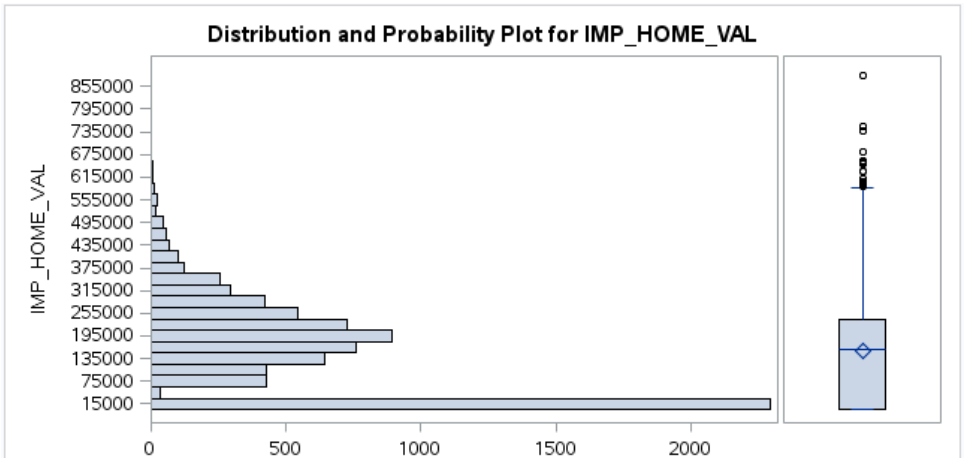
Extreme Values

There are several variables that have extremely large values or odd values that may be of concern. For instance, if we go back to Table 1, the CAR_AGE variable has a minimum value of -3 years. This does not make any sense because one cannot own a car for negative years. There

are two ways to fix this value: 1) Make that value “0” or 2) make that value ”3”. It seems more likely that this could be a typo where the customer or insurance company accidentally put a negative sign. Option 2 appears to be a sensible choice here.

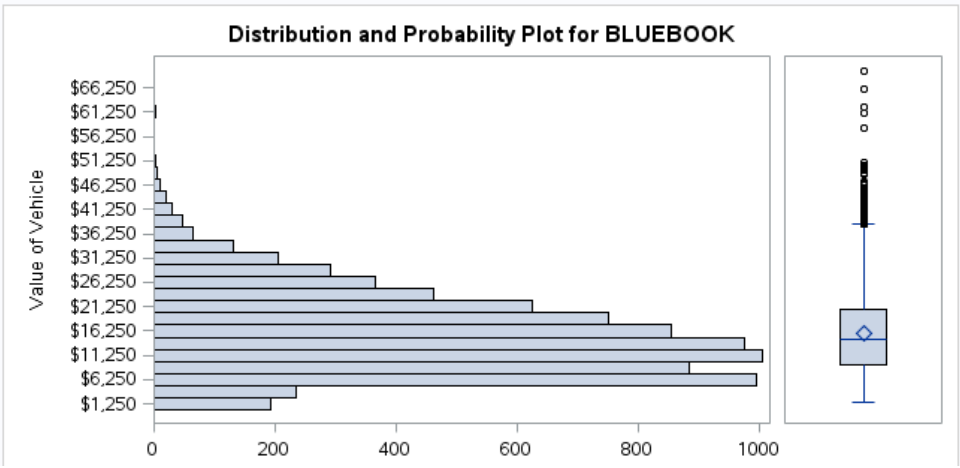
A few variables have a wide range of values and contain extremely large values. Those variables are IMP_HOME_VAL, BLUEBOOK, and TRAVTIME. From Table 1 the maximum values are spread out very far from the majority of the data points. This causes the distribution of the data to be right skewed which violates the assumption of normality. Even if this is the case, regression is often robust enough where models are still quite accurate. With that said, it is viable to cap the values at 99% quantile on the IMP_HOME_VAL, BLUEBOOK, and TRAVTIME variables to reduce the large values. The following graphs show the distribution of the original variables and the values in the quantiles to the left.

Quantiles (Definition 5)	
Level	Quantile
100% Max	885282
99%	497746
95%	370038
90%	318758
75% Q3	236446
50% Median	160202
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0



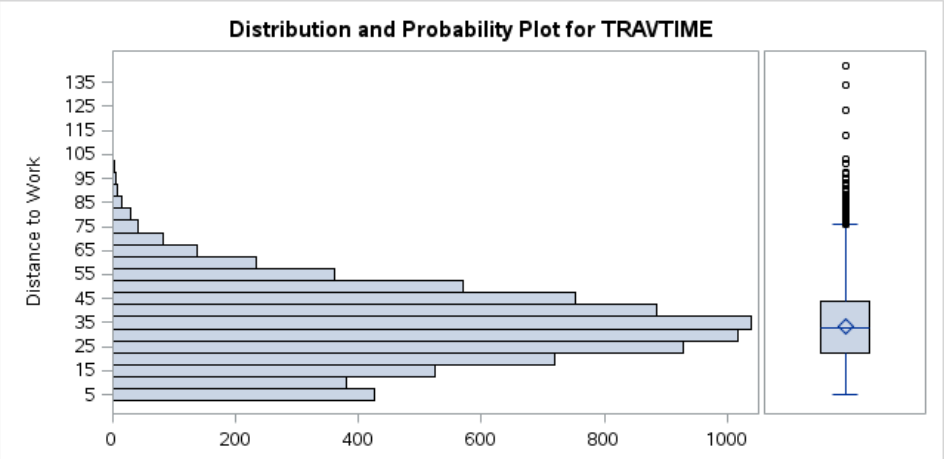
For IMP_HOME_VAL variable, if values are greater than 497746 then cap at 497746.

Quantiles (Definition 5)	
Level	Quantile
100% Max	69740
99%	39090
95%	31110
90%	27460
75% Q3	20850
50% Median	14440
25% Q1	9280
10%	6000
5%	4900
1%	1500
0% Min	1500



For BLUEBOOK, if values are greater than 39090 then cap at 39090.

Quantiles (Definition 5)	
Level	Quantile
100% Max	142.12063
99%	75.14433
95%	60.41450
90%	53.72325
75% Q3	43.80707
50% Median	32.87097
25% Q1	22.45170
10%	12.59378
5%	7.02184
1%	5.00000
0% Min	5.00000



For TRAVTIME, if values are greater than 75 then cap at 75.

The original variables had quite a bit of skewing as revealed by the long tails in the distribution and boxplot. The transformed variables have less skewing and will be used in the model development process.

Model Building

Now that the dataset is better understood and prepared for model development, there are several approaches for creating an optimal model. Several variable selection techniques will be used to create a variety of models such as manual inclusion of all variables, stepwise selection, and decision trees. The techniques for generating each model will be discussed and the parameters will be analyzed to determine if inclusion in the model makes intuitive sense. Metrics will also be displayed for each model for consideration when the best model is selected.

<u>Model 1</u>
Variable selection method: All variables
AIC = 7363.235
Log Likelihood = 7277.235

The first model includes decision tree imputation, flag variables, fixed negative value, and extreme value capping. The model was produced using logistic regression.

Because this is the first model, including all the variables is helpful to see which variables are statistically significant in predicting the response variable. This model will be used as a baseline for the next few models to see if there is any improvement when changes are made in the model development process.

The model contains 29 variables, which include flag variables. On top of that, the categorical variables contain several levels which make this model rather difficult to interpret.

Nevertheless, the model makes intuitive sense. Most of the variables have signs that we would expect. For example, an increase in the KIDSDRIV variable will increase the odds of crashing by 0.3879. This makes sense as the more kids a customer has driving, the higher the chance of a vehicle crash.

An example of a variable that is counterintuitive is the "PhD" category of the EDUCATION variable. According to the model, having a PhD increases the odds of crashing by 0.00951. It is normally expected that the higher the education level, the less risky the driver.

Model 1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3289	0.2917	20.7540	<.0001
IMP_AGE		1	-0.00009	0.00403	0.0005	0.9820
BLUEBOOK		1	-0.00002	5.345E-6	17.8697	<.0001
IMP_CAR_AGE		1	-0.00371	0.00782	0.2246	0.6355
CAR_TYPE	Minivan	1	-0.7579	0.1115	46.2301	<.0001
CAR_TYPE	Panel Truck	1	-0.1765	0.2003	0.7760	0.3784
CAR_TYPE	Pickup	1	-0.2033	0.1168	3.0297	0.0818
CAR_TYPE	Sports Car	1	0.2543	0.0984	6.6795	0.0098
CAR_TYPE	Van	1	-0.1226	0.1594	0.5916	0.4418
CAR_USE	Commercial	1	0.7766	0.0921	71.0903	<.0001
CLM_FREQ		1	0.1960	0.0286	47.1098	<.0001
EDUCATION	<High School	1	-0.0155	0.0953	0.0265	0.8706
EDUCATION	Bachelors	1	-0.3899	0.0910	18.3768	<.0001
EDUCATION	Masters	1	-0.2958	0.1617	3.3454	0.0674
EDUCATION	PhD	1	0.00951	0.2163	0.0019	0.9649
HOMEKIDS		1	0.0522	0.0372	1.9709	0.1603
IMP_HOME_VAL		1	-1.45E-6	3.576E-7	16.4250	<.0001
IMP_INCOME		1	-3.13E-6	1.124E-6	7.7329	0.0054
IMP_JOB	Clerical	1	0.0999	0.1075	0.8635	0.3528
IMP_JOB	Doctor	1	-1.0814	0.2692	16.1363	<.0001
IMP_JOB	Home Maker	1	-0.1373	0.1544	0.7906	0.3739
IMP_JOB	Lawyer	1	-0.1973	0.1830	1.1618	0.2811
IMP_JOB	Manager	1	-0.8442	0.1381	37.3848	<.0001
IMP_JOB	Professional	1	-0.1394	0.1199	1.3513	0.2451
IMP_JOB	Student	1	-0.1526	0.1309	1.3595	0.2436
KIDSDRIV		1	0.3879	0.0613	40.0469	<.0001
MSTATUS	Yes	1	-0.4692	0.0861	29.6959	<.0001
MVR_PTS		1	0.1130	0.0137	68.4446	<.0001
OLDCLAIM		1	-0.00001	3.915E-6	12.6429	0.0004
PARENT1	No	1	-0.3721	0.1099	11.4647	0.0007
RED_CAR	no	1	0.0266	0.0867	0.0941	0.7590
REVOKED	No	1	-0.8878	0.0915	94.1895	<.0001
SEX	M	1	0.0791	0.1121	0.4973	0.4807
TIF		1	-0.0557	0.00736	57.3872	<.0001
TRAVTIME		1	0.0152	0.00192	62.4654	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3907	0.1128	449.2466	<.0001
IMP_YOJ		1	-0.0160	0.00845	3.5910	0.0581
AGE_FLAG		1	2.1420	1.2369	2.9990	0.0833
CAR_AGE_FLAG		1	0.1475	0.1182	1.5572	0.2121
HOME_VAL_FLAG		1	-0.0623	0.1236	0.2542	0.6141
INCOME_FLAG		1	-0.0579	0.1308	0.1960	0.6579
JOB_FLAG		1	0.5099	0.1603	10.1246	0.0015
YOJ_FLAG		1	0.0543	0.1269	0.1828	0.6690

This is a large model and it appears several variables are not statistically significant at the 0.05 level. Those variables are IMP_CAR_AGE, CAR_TYPE (Panel Truck, Pickup, Van), EDUCATION (<High School, Masters, PhD), HOMEKIDS, IMP_JOB (Clerical, Home Maker, Lawyer,

Professional, Student), RED_CAR (no), SEX (M), IMP_YOJ, AGE_FLAG, CAR_AGE_FLAG, HOME_VAL_FLAG, INCOME_FLAG, and YOJ_FLAG. For now, these variables will be kept in the model because some are categorical variables that have several levels so if some are statistically significant at 0.05 then all levels should be kept in the model. Because this is a full model, even variables that do not meet the 0.05 threshold will be used. Later in Model 3, stepwise selection is employed where variables not statistically significant will be dropped.

Despite this model having a variable that is counterintuitive, all other variables make sense. Also, there are non-statistically significant variables, but those will be included in the model to keep as reference model.

Model 2
Variable selection method: All variables using probit link function
AIC = 7374.686
Log Likelihood = 7288.686

Model 2 mirrors Model 1 in that it includes decision tree imputation, flag variables, fixed negative value, and extreme value capping. The only difference is that this model was generated using the probit link function. Model 1 was generated using the logit link function.

Model 2 contains the same number of variables as Model 1. All variables are manually included in order to compare whether the change in the link function reveals any differences. A quick look at the AIC and Log Likelihood metrics, there is a minimal difference. The AIC and Log Likelihood are slightly higher in this model.

When examining the coefficient signs, they are exactly the same as Model 1. Most of the signs of the variables make sense. Again, the “PhD” category in EDUCATION variable has a positive sign which indicates that having a PhD increases the odds of crashing by 0.000802. Even though this variable is counterintuitive, it is very small. It is expected that this issue will have minimal effect on the outcome of the model. Therefore, this model is still under consideration. However, because this model does not improve over Model 1, Model 1 is still considered the better model at this point.

Model 2

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept		1	-0.7355	0.1665	19.5223	<.0001
IMP_AGE		1	0.000198	0.00233	0.0072	0.9324
BLUEBOOK		1	-0.00001	3.047E-6	17.0957	<.0001
IMP_CAR_AGE		1	-0.00224	0.00449	0.2502	0.6169
CAR_TYPE	Minivan	1	-0.4348	0.0632	47.2832	<.0001
CAR_TYPE	Panel Truck	1	-0.1161	0.1151	1.0185	0.3129
CAR_TYPE	Pickup	1	-0.1253	0.0671	3.4880	0.0618
CAR_TYPE	Sports Car	1	0.1486	0.0573	6.7260	0.0095
CAR_TYPE	Van	1	-0.0823	0.0913	0.8126	0.3673
CAR_USE	Commercial	1	0.4404	0.0535	67.7489	<.0001
CLM_FREQ		1	0.1171	0.0168	48.5754	<.0001
EDUCATION	<High School	1	-0.0155	0.0556	0.0778	0.7804
EDUCATION	Bachelors	1	-0.2302	0.0527	19.0523	<.0001
EDUCATION	Masters	1	-0.1648	0.0920	3.2072	0.0733
EDUCATION	PhD	1	0.000802	0.1225	0.0000	0.9948
HOMEKIDS		1	0.0337	0.0216	2.4374	0.1185
IMP_HOME_VAL		1	-7.81E-7	2.057E-7	14.4189	0.0001
IMP_INCOME		1	-1.78E-6	6.423E-7	7.6568	0.0057
IMP_JOB	Clerical	1	0.0540	0.0627	0.7423	0.3889
IMP_JOB	Doctor	1	-0.6072	0.1508	16.2122	<.0001
IMP_JOB	Home Maker	1	-0.0876	0.0897	0.9538	0.3287
IMP_JOB	Lawyer	1	-0.1239	0.1044	1.4077	0.2354
IMP_JOB	Manager	1	-0.4839	0.0786	37.9291	<.0001
IMP_JOB	Professional	1	-0.0805	0.0698	1.3301	0.2488
IMP_JOB	Student	1	-0.0723	0.0761	0.9026	0.3421
KIDSDRIV		1	0.2210	0.0357	38.2550	<.0001
MSTATUS	Yes	1	-0.2796	0.0496	31.7384	<.0001
MVR_PTS		1	0.0661	0.00802	67.8638	<.0001
OLDCLAIM		1	-7.68E-6	2.291E-6	11.2345	0.0008
PARENT1	No	1	-0.2082	0.0641	10.5452	0.0012
RED_CAR	no	1	0.0101	0.0500	0.0406	0.8403
REVOKED	No	1	-0.5129	0.0535	92.0591	<.0001
SEX	M	1	0.0497	0.0639	0.6049	0.4367
TIF		1	-0.0327	0.00421	60.4975	<.0001
TRAVTIME		1	0.00872	0.00110	62.3147	<.0001
URBANICITY	Highly Urban/ Urban	1	1.3119	0.0583	506.2705	<.0001
IMP_YOJ		1	-0.00891	0.00489	3.3226	0.0683
AGE_FLAG		1	1.3105	0.6887	3.6208	0.0571
CAR_AGE_FLAG		1	0.0869	0.0683	1.6169	0.2035
HOME_VAL_FLAG		1	-0.0468	0.0718	0.4240	0.5150
INCOME_FLAG		1	-0.0310	0.0753	0.1691	0.6809
JOB_FLAG		1	0.2728	0.0916	8.8652	0.0029
YOJ_FLAG		1	0.0295	0.0735	0.1608	0.6884

<u>Model 3</u>
Variable selection method: Stepwise
AIC = 7353.309
Log Likelihood = 7289.309

The model includes decision tree imputation, flag variables, fixed negative value, and extreme value capping. Model 3 is the exact same as Model 1 except that instead of manually including all variables into the model, variables are selected using the stepwise technique. By using the

stepwise selection method, the model yields 18 variables compared to 29 in the previous two models. This simpler model increases interpretability.

In this model, the parameters are very intuitive. For example, in the previous two models, having a PhD slightly increased the odds of crashing. In this case, having a PhD slightly decreases the odds of crashing by 0.0211.

Notice despite using stepwise selection, some variables are still not statistically significant at the 0.05 level. Those variables are CAR_TYPE (Panel Truck, Pickup, Van), EDUCATION (<High School, PhD), and IMP_JOB (Clerical, Home Maker, Lawyer, Professional, Student). This is because certain levels in the variables are statistically significant so the entire variable is included with all its categories.

Model 3

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3507	0.2084	42.0221	<.0001
BLUEBOOK		1	-0.00002	4.795E-6	26.3836	<.0001
CAR_TYPE	Minivan	1	-0.7124	0.0861	68.5349	<.0001
CAR_TYPE	Panel Truck	1	-0.0837	0.1537	0.2967	0.5860
CAR_TYPE	Pickup	1	-0.1633	0.0939	3.0274	0.0819
CAR_TYPE	Sports Car	1	0.2588	0.0980	6.9703	0.0083
CAR_TYPE	Van	1	-0.0473	0.1211	0.1529	0.6958
CAR_USE	Commercial	1	0.7720	0.0918	70.7602	<.0001
CLM_FREQ		1	0.1962	0.0285	47.2844	<.0001
EDUCATION	<High School	1	-0.0128	0.0947	0.0183	0.8924
EDUCATION	Bachelors	1	-0.4032	0.0839	23.0825	<.0001
EDUCATION	Masters	1	-0.3361	0.1419	5.6073	0.0179
EDUCATION	PhD	1	-0.0211	0.2031	0.0108	0.9174
IMP_HOME_VAL		1	-1.48E-6	3.562E-7	17.1745	<.0001
IMP_INCOME		1	-3.29E-6	1.119E-6	8.6432	0.0033
IMP_JOB	Clerical	1	0.0978	0.1071	0.8334	0.3613
IMP_JOB	Doctor	1	-1.0909	0.2686	16.4993	<.0001
IMP_JOB	Home Maker	1	-0.0524	0.1452	0.1305	0.7179
IMP_JOB	Lawyer	1	-0.1972	0.1827	1.1653	0.2804
IMP_JOB	Manager	1	-0.8491	0.1376	38.0825	<.0001
IMP_JOB	Professional	1	-0.1462	0.1197	1.4922	0.2219
IMP_JOB	Student	1	-0.0582	0.1239	0.2210	0.6383
KIDSDRIV		1	0.4197	0.0552	57.9179	<.0001
MSTATUS	Yes	1	-0.4520	0.0820	30.3711	<.0001
MVR_PTS		1	0.1145	0.0136	70.8363	<.0001
OLDCLAIM		1	-0.00001	3.909E-6	13.1336	0.0003
PARENT1	No	1	-0.4540	0.0943	23.1734	<.0001
REVOKED	No	1	-0.8937	0.0913	95.7940	<.0001
TIF		1	-0.0555	0.00735	57.1485	<.0001
TRAVTIME		1	0.0150	0.00192	61.3772	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3887	0.1126	449.7108	<.0001
JOB_FLAG		1	0.5222	0.1598	10.6825	0.0011

<u>Model 4</u>
Variable selection method: Decision Tree
AIC = 7977.859
Log Likelihood = 7943.859

This model includes decision tree imputation, fixed negative value, and extreme value capping. The model was generated using logistic regression. However, the variables were selected using a decision tree produced in the RStudio software. Compared to the previous three models, Model 4 is by far the simplest one. There are only 6 variables in the model.

```
Variables actually used in tree construction:
[1] CAR_TYPE    JOB         N_HOME_VAL N_OLDCLAIM TRAVTIME    URBANICITY
```

As you can see above, by using the "rpart" package in the R programming language, a decision tree was fitted to predict TARGET_FLAG based on all the variables. Ultimately, the decision tree determined six variables to be significant. Those variables are CAR_TYPE, JOB, HOME_VAL, OLDCLAIM, TRAVTIME, and URBANICITY. These variables do make sense in predicting the odds of crashing.

In terms of the signs of the parameters estimates, they do meet the common sense gauge. For example, all categories of jobs decrease the odds of crashing relative to blue collar professions. Also, variables like TRAVTIME and URBANICITY increase the odds of crashing as expected.

Most variables are statistically significant at less than 0.0001. Again, some variables do not meet that threshold like CAR_TYPE (Panel Truck, Pickup, Van) and IMP_JOB (Student). These variables are included because the other categories are significant within the respective variables so all categories are included.

Model 4

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept		1	-2.4802	0.1440	296.7784	<.0001
CAR_TYPE	Minivan	1	-0.8213	0.0804	104.2105	<.0001
CAR_TYPE	Panel Truck	1	0.1798	0.1121	2.5725	0.1087
CAR_TYPE	Pickup	1	0.0998	0.0817	1.4940	0.2216
CAR_TYPE	Sports Car	1	0.2150	0.0925	5.4076	0.0200
CAR_TYPE	Van	1	0.0411	0.1044	0.1551	0.6937
IMP_JOB	Clerical	1	-0.2478	0.0879	7.9523	0.0048
IMP_JOB	Doctor	1	-1.2727	0.1468	75.1883	<.0001
IMP_JOB	Home Maker	1	-0.5199	0.1147	20.5626	<.0001
IMP_JOB	Lawyer	1	-0.9393	0.1067	77.5157	<.0001
IMP_JOB	Manager	1	-1.2899	0.0962	179.8489	<.0001
IMP_JOB	Professional	1	-0.7253	0.0942	59.3037	<.0001
IMP_JOB	Student	1	-0.1689	0.1086	2.4197	0.1198
IMP_HOME_VAL		1	-3.5E-6	2.609E-7	179.4944	<.0001
OLDCLAIM		1	0.000019	2.86E-6	44.1743	<.0001
TRAVTIME		1	0.0141	0.00182	59.6099	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3620	0.1063	493.8088	<.0001

Model Selection

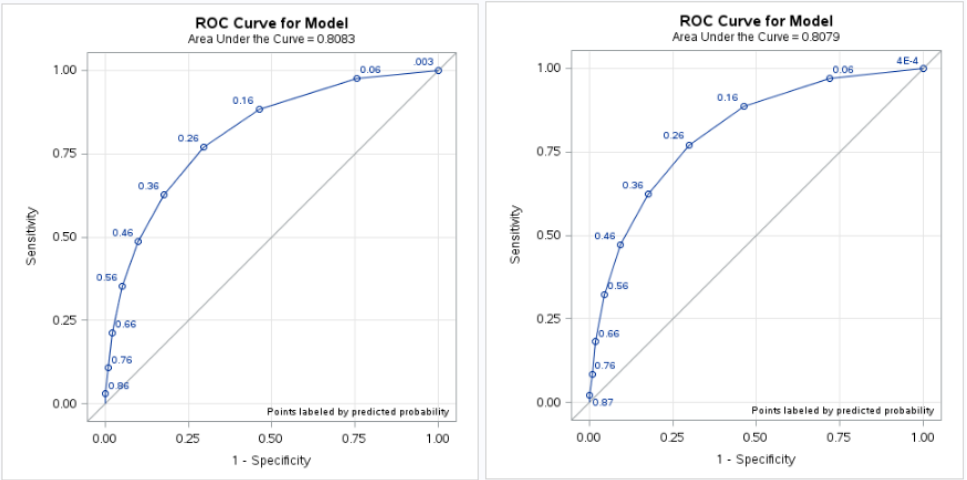
The best model will be selected on several criteria. The AIC will be the primary metric of interest as it is used for model validation and generally penalizes overfitting models. A lower AIC value is preferable.

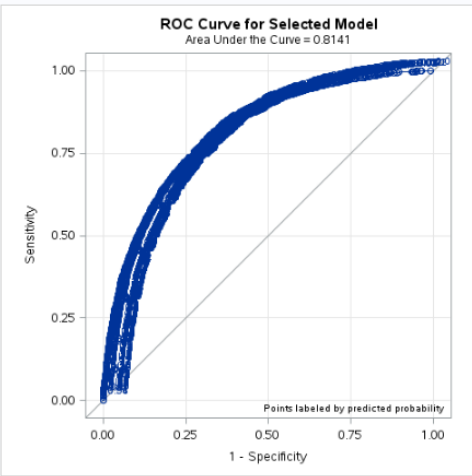
Other metrics that will be considered are the Log Likelihood, the Area Under the ROC curve (AUC), and the Kolmogorov-Smirnov statistic. A lower Log Likelihood value is better and a higher AUC value and K-S statistic are better. The Log Likelihood metric is used extensively in logistic regression and is used when models produce probabilities for binary outcomes. The ROC curve reveals the true positive rate versus the false positive rate. If the line is more to the left and upper region, the better the accuracy and generally the greater the AUC. For the K-S statistic, it measures the correct classification of positive and negative groups.

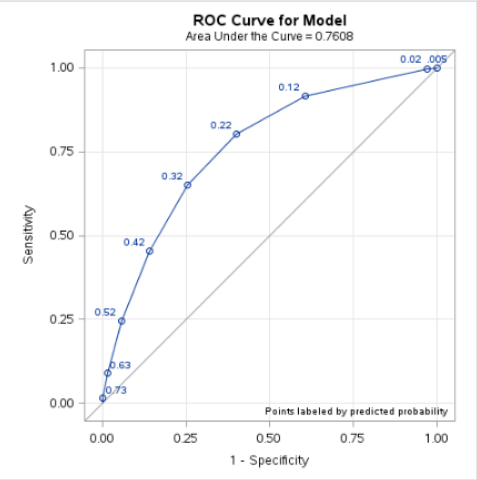
Comparing the AIC values for each model, Model 1 scored 7363.235, Model 2 scored 7374.686, Model 3 scored 7353.309, and Model 4 scored 7977.859. The best performer here is Model 3 while Model 4 appears to have perform the worst. It makes sense that Model 3 came out on top because the AIC metric penalizes overfitting. Although many of the variables in the dataset were predictive of the TARGET_FLAG response variable, Model 3 only had 18 variables compared to Model 1 and 2's 29 variables. There was enough predictive power without overfitting which gave Model 3 the edge here. Model 4 likely performed the worst because it lacked predictive accuracy by using too few variables. Based on the AIC metric, Model 3 wins here.

The next metric to examine is the Log Likelihood. Model 1 scored 7277.235, Model 2 scored 7288.686, Model 3 scored 7289.309, and Model 4 scored 7943.859. Model 1, 2, and 3 were very close with Model 1 coming out ahead. Again, Model 4 was the worst here.

The following four plots show the ROC curves for each of the four models. Model 1 is the upper left, Model 2 is the upper right, Model 3 is the lower left, and Model 4 is the lower right.







Model 1 had an AUC of 0.8083. Model 2 had an AUC of 0.8079. Model 3 had an AUC of 0.8141. Model 4 had an AUC of 0.7608. Once again, Model 3 outperforms the rest. The use of stepwise selection was able to find the optimal set of variables that maximized the AUC value as seen in the lower left ROC curve above. All models did perform better than a random model represented by the straight line.

The following are the K-S statistics for Model 1, 2, 3, and 4: 0.211037, 0209875, 0.21014, and 0.1817. Model 1 has the best K-S statistic of the bunch despite a marginal victory. Model 3 comes in second place. All the models are correctly classifying positive and negative groups about 20% better than a random model would.

After consideration of all the above criteria, the final model is Model 3. The model had the lowest AIC metric and the highest AUC value out of the four models. It came in second behind Model 1 in the Log Likelihood value and K-S statistic. Aside from performing well on the metrics, Model 3 had the best predictive accuracy and interpretability tradeoff. With only 18 variables, it had as much prediction accuracy as Model 1 while still coming ahead on two out of the four metrics.

Conclusion

Prior to building models to predict the probability of an auto insurance customer crashing, the dataset containing 8161 records were examined for size, distribution, correlations, and missing/extreme values. The dataset was prepared by fixing odd values, imputing and creating flag variables for missing values, and setting a cap for variables with extreme values.

Four models were built using different techniques and compared. They were analyzed and judged on common sense and several metrics, i.e. AIC, Log Likelihood, AUC, and K-S statistic. The triumphant model was Model 3. It performed very well on metrics and contained far less variables than similar performing models. Model 3 was chosen as the best model for accuracy and parsimony.

Bingo Bonus

Proc GENMOD

The champion model selected from the report is Model 3. Model 3 utilizes decision tree imputation, capping extreme values and stepwise variable selection.

First off, the results are not the same. The GENMOD results produce a higher AIC value at 7729.5685 while Model 3 still wins with 7353.309. That is a dramatic difference despite using the same variables and techniques.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8118	1218.7694	0.1501
Scaled Deviance	8118	8161.0000	1.0053
Pearson Chi-Square	8118	1218.7694	0.1501
Scaled Pearson X2	8118	8161.0000	1.0053
Log Likelihood		-3820.7843	
Full Log Likelihood		-3820.7843	
AIC (smaller is better)		7729.5685	
AICC (smaller is better)		7730.0564	
BIC (smaller is better)		8037.8819	

The results of the PROC GENMOD code is also slightly different. Metrics are given for AICC and BIC while PROC LOGISTIC does not give us these stats. We also get the Wald 95% confidence limits in regards to the parameter estimates.

Examining the parameter estimates below, PROC GENMOD has included all the variables which differ from our champion model. With that said, the signs of the coefficients do pass the common sense filter. However, several variables have a 0.000 value which is odd. These would be better removed from the model if that is the case. There is also an added parameter called scale of 0.3664 which I am not sure what value this statistic adds. It needs to be further investigated.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3427	0.0425	0.2593	0.4260	64.93	<.0001
IMP_AGE		1	-0.0000	0.0008	-0.0012	0.0012	0.00	0.9837
BLUEBOOK		1	-0.0000	0.0000	-0.0000	-0.0000	15.22	<.0001
IMP_CAR_AGE		1	-0.0008	0.0011	-0.0030	0.0014	0.52	0.4725
CAR_TYPE	Minivan	1	-0.1016	0.0153	-0.1316	-0.0717	44.32	<.0001
CAR_TYPE	Panel Truck	1	-0.0426	0.0290	-0.0994	0.0143	2.15	0.1421
CAR_TYPE	Pickup	1	-0.0303	0.0169	-0.0635	0.0029	3.20	0.0738
CAR_TYPE	Sports Car	1	0.0388	0.0152	0.0089	0.0686	6.49	0.0109
CAR_TYPE	Van	1	-0.0255	0.0228	-0.0702	0.0191	1.25	0.2628
CAR_USE	Commercial	1	0.1218	0.0140	0.0944	0.1493	75.55	<.0001
CLM_FREQ		1	0.0323	0.0047	0.0231	0.0414	47.47	<.0001
EDUCATION	<High School	1	-0.0062	0.0146	-0.0348	0.0225	0.18	0.6741
EDUCATION	Bachelors	1	-0.0615	0.0136	-0.0881	-0.0349	20.49	<.0001
EDUCATION	Masters	1	-0.0367	0.0227	-0.0812	0.0079	2.60	0.1066
EDUCATION	PhD	1	-0.0054	0.0301	-0.0644	0.0536	0.03	0.8570
HOMEKIDS		1	0.0068	0.0056	-0.0041	0.0177	1.51	0.2189
IMP_HOME_VAL		1	-0.0000	0.0000	-0.0000	-0.0000	12.66	0.0004
IMP_INCOME		1	-0.0000	0.0000	-0.0000	-0.0000	4.52	0.0335
IMP_JOB	Clerical	1	0.0144	0.0165	-0.0179	0.0467	0.76	0.3818
IMP_JOB	Doctor	1	-0.1585	0.0362	-0.2294	-0.0876	19.20	<.0001
IMP_JOB	Home Maker	1	-0.0145	0.0230	-0.0596	0.0307	0.39	0.5299
IMP_JOB	Lawyer	1	-0.0517	0.0261	-0.1028	-0.0007	3.94	0.0471
IMP_JOB	Manager	1	-0.1378	0.0198	-0.1766	-0.0989	48.28	<.0001
IMP_JOB	Professional	1	-0.0291	0.0181	-0.0647	0.0064	2.58	0.1081
IMP_JOB	Student	1	-0.0147	0.0201	-0.0541	0.0247	0.53	0.4656
KIDSDRIV		1	0.0606	0.0096	0.0418	0.0795	39.66	<.0001
MSTATUS	Yes	1	-0.0662	0.0127	-0.0910	-0.0413	27.25	<.0001
MVR_PTS		1	0.0212	0.0022	0.0169	0.0255	92.37	<.0001
OLDCLAIM		1	-0.0000	0.0000	-0.0000	-0.0000	14.34	0.0002
PARENT1	No	1	-0.0746	0.0172	-0.1083	-0.0408	18.79	<.0001
RED_CAR	no	1	0.0057	0.0127	-0.0192	0.0306	0.20	0.6550
REVOKED	No	1	-0.1536	0.0148	-0.1825	-0.1246	108.32	<.0001
SEX	M	1	0.0132	0.0156	-0.0174	0.0439	0.71	0.3978
TIF		1	-0.0080	0.0010	-0.0100	-0.0059	59.03	<.0001
TRAVTIME		1	0.0021	0.0003	0.0016	0.0027	56.69	<.0001
URBANICITY	Highly Urban/ Urban	1	0.2975	0.0118	0.2743	0.3208	630.87	<.0001
IMP_YOJ		1	-0.0028	0.0013	-0.0052	-0.0003	4.83	0.0280
AGE_FLAG		1	0.3762	0.1588	0.0650	0.6875	5.61	0.0178
CAR_AGE_FLAG		1	0.0198	0.0177	-0.0149	0.0545	1.25	0.2642
HOME_VAL_FLAG		1	-0.0147	0.0185	-0.0510	0.0216	0.63	0.4271
INCOME_FLAG		1	-0.0086	0.0189	-0.0457	0.0285	0.21	0.6481
JOB_FLAG		1	0.0513	0.0234	0.0054	0.0971	4.81	0.0283
YOJ_FLAG		1	0.0151	0.0187	-0.0216	0.0518	0.65	0.4204
Scale		1	0.3864	0.0030	0.3806	0.3924		

Recreate code in R

```
# Import and load data files and libraries
setwd("../Users/kphas_000/Desktop/PREDICT 411/Unit02/Insurance")
train <- read.csv("logit_insurance.csv",header=TRUE,na.strings=c("","NA"))
test <- read.csv("logit_insurance_test.csv",header=TRUE,na.strings=c("","NA"))

library(rpart)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
library(randomForest)
library(useful)
library(party)

# combine train and test datasets for data preparation
combi <- rbind(train,test)

# check data
str(combi)
summary(combi)

# convert dollar columns to numeric
cur2num <- function(x) {as.numeric(sub(',','',(sub('\\$','',as.character(x)))))}
combi$N_INCOME <- cur2num(combi$INCOME)
combi$N_HOME_VAL <- cur2num(combi$HOME_VAL)
combi$N_BLUEBOOK <- cur2num(combi$BLUEBOOK)
combi$N_OLDCLAIM <- cur2num(combi$OLDCLAIM)
```

```
head(combi)

# drop the old columns
combi$TARGET_AMT <- NULL
combi$INCOME <- NULL
combi$HOME_VAL <- NULL
combi$BLUEBOOK <- NULL
combi$OLDCLAIM <- NULL

summary(combi)

# there is a negative value which makes no sense so we fix it
combi$CAR_AGE[6941] <- 3  # most likely the -3 was typo and should have been 3
summary(combi)

## impute columns with missing values using decision tree
imp_age <- rpart(AGE ~ . - TARGET_FLAG, data=combi[!is.na(combi$AGE),], method="anova")
combi$AGE[is.na(combi$AGE)] <- predict(imp_age, combi[is.na(combi$AGE),])
fancyRpartPlot(imp_age, main="Decision Tree for AGE",sub="")

imp_yoj <- rpart(YOJ ~ . - TARGET_FLAG, data=combi[!is.na(combi$YOJ),], method="anova")
combi$YOJ[is.na(combi$YOJ)] <- predict(imp_yoj, combi[is.na(combi$YOJ),])
fancyRpartPlot(imp_yoj, main="Decision Tree for YOJ",sub="")

imp_car <- rpart(CAR_AGE ~ . - TARGET_FLAG, data=combi[!is.na(combi$CAR_AGE),],
method="anova")
combi$CAR_AGE[is.na(combi$CAR_AGE)] <- predict(imp_car, combi[is.na(combi$CAR_AGE),])
```

```
fancyRpartPlot(imp_car, main="Decision Tree for CAR_AGE",sub="")


imp_income <- rpart(N_INCOME ~ . - TARGET_FLAG, data=combi[!is.na(combi$N_INCOME),],
method="anova")

fancyRpartPlot(imp_income)  # too many levels to code so let's prune the tree

plotcp(imp_income)

printcp(imp_income)

imp_income2 <- prune(imp_income,cp=0.05)

fancyRpartPlot(imp_income2, main="Decision Tree for INCOME",sub="")

combi$N_INCOME[is.na(combi$N_INCOME)] <- predict(imp_income2,
combi[is.na(combi$N_INCOME),])


imp_homeval <- rpart(N_HOME_VAL ~ . - TARGET_FLAG,
data=combi[!is.na(combi$N_HOME_VAL),], method="anova")

fancyRpartPlot(imp_homeval)

plotcp(imp_homeval)

printcp(imp_homeval)

imp_homeval2 <- prune(imp_homeval,cp=0.05)

fancyRpartPlot(imp_homeval2, main="Decision Tree for HOMEVAL",sub="")

combi$N_HOME_VAL[is.na(combi$N_HOME_VAL)] <- predict(imp_homeval2,
combi[is.na(combi$N_HOME_VAL),])


imp_job <- rpart(JOB ~ . - TARGET_FLAG, data=combi[!is.na(combi$JOB),], method="class")

fancyRpartPlot(imp_job)

plotcp(imp_job)

printcp(imp_job)

imp_job2 <- prune(imp_job,cp=0.05)

fancyRpartPlot(imp_job2, main="Decision Tree for JOB",sub="")
```

```
combi$JOB[is.na(combi$JOB)] <- predict(imp_job2, combi[is.na(combi$JOB),],type="class")

# check to make sure NAs are gone
summary(combi)
anyNA(combi[,!names(combi) %in% "TARGET_FLAG"])

# cap extreme values
summary(combi$N_HOME_VAL)
combi$N_HOME_VAL[combi$N_HOME_VAL > 497746] <- 497746
summary(combi$N_HOME_VAL)

summary(combi$N_BLUEBOOK)
combi$N_BLUEBOOK[combi$N_BLUEBOOK > 39090] <- 39090
summary(combi$N_BLUEBOOK)

summary(combi$TRAVTIME)
combi$TRAVTIME[combi$TRAVTIME > 75] <- 75
summary(combi$TRAVTIME)

# split out the imputed datasets back to training and test
train <- combi[1:8161,]
test <- combi[8162:10302,]
anyNA(train)
dim(train)
dim(test)

# fit full model
```

```
fit <- rpart(TARGET_FLAG ~ . - INDEX, data=train, method="class") # will get prob two columns for 0 (no crash) and 1 (crash)

summary(fit)

fancyRpartPlot(fit)


# write score file

Prediction <- predict(fit, test)

submit <- data.frame(INDEX = test$INDEX, P_TARGET_FLAG = Prediction)

colnames(submit)[2] <- "P_TARGET_FLAG"

submit$P_TARGET_FLAG.1 <- NULL

head(submit)

write.csv(submit, file = "model.csv", row.names = FALSE)




## Other imputation/modeling methods such as K-Nearest Neighbors and Random Forests

# impute using mice

library(mice)

miceImp <- mice(combi[,!names(combi) %in% "TARGET_FLAG"],method="rf")

miceOut <- complete(miceImp)

anyNA(miceOut)


miceFull <- cbind(miceOut,TARGET_FLAG=combi$TARGET_FLAG) # bring TARGET_FLAG back into data so we can run model


train <- miceFull[1:8161,]

test <- miceFull[8162:10302,]

anyNA(train)
```

```
anyNA(test)

set.seed(222)
fit <- cforest(TARGET_FLAG ~ . - INDEX, data=train, controls=cforest_unbiased(ntree=2000))
fit

Prediction <- predict(fit, test, OOB=TRUE, type = "response")
submit <- data.frame(INDEX = test$INDEX, P_TARGET_FLAG = Prediction)
colnames(submit)[2] <- "P_TARGET_FLAG"
write.csv(submit, file = "miceForestCap.csv", row.names = FALSE)

## impute missing values using kNN
library(DMwR)
knnOut <- knnImputation(combi[,!names(combi) %in% "TARGET_FLAG"])
anyNA(knnOut)
summary(knnOut)
```