Noé Flores
Predict Auto Insurance

Introduction:
The auto insurance industry operates on a very basic premise of issuing insurance to drivers for a premium. That premium is determined by several factors, inclusive of the driver's age, sex, employment status, etc. However, one of the most important factors to consider is how many, if any, accidents an individual might have in the future. If an individual doesn't have any accidents now, could we predict how many they are likely to have based on certain variables? Answering this question will be the main objective of this analysis. We will perform an exploratory analysis of our data set to gather information regarding which variables we will be incorporating into this research and follow through with any data preparation needed to ensure we have an optimal model. Finally, using logistic regression, we will build and compare several models to determine which has the greatest predictive value, interpretability, and practicality based on AIC, log likelihood, Area Under the Curve (AUC) and Ks Statistic. These models will be built and compared with our end goal in mind to select the best possible fitting and most intuitive model.

Data Exploration:
A quick examination of the contents of our data set produces the results below. There are 8161 customer records with 26 variables. Each record represents a customer at an auto insurance company, and each record contains two target variables. The first variable is TARGET_FLAG, which indicates whether the customer was in a car crash and is expressed by the value of "0" if they did not crash their car, and a value of "1" if they did. The TARGET_AMT indicates how much money is paid if the customer did in fact crash their car and will have a value of "0" if there was no crash.

Figure 1: Summary table of Data Contents.

Insurance Data	Contents
Observations	8161
Variables	26
Indexes	0
Observation Length	216
Deleted Observations	0
Compressed	NO
Sorted	NO

Figure.1 gives us a summary breakdown of the information we previously discussed about our data set.

Figure 2: Summary Data Dictionary and Variable Theoretical Effects.

VARIABLE	DEFINITION	THEORETICAL EFFECT
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	#Children @Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	#Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Figure .2 contains a breakdown of the observations and shows the 23 measurable variables and their theoretical effect or impact on the price of auto insurance. The theoretical effects are good reference points to have when we begin building our models. The variables in Figure .2 contain personal information such as age, income, and sex and information on the vehicle itself, such as the type of car, vehicle age, and the value. There are also some variables that perhaps at first glance we wouldn't consider to be pertinent to determining the cost of insurance, such as Home Value and Number of Children at home. What's clear, is that this dataset contains both numeric and categorical variables to work into our models.

In Figure.3 below, we have some basic summary statistics of our data set. We use this summary to make a quick examination of the data set to look for possible issues, including missing data. The variables in this table contain a breakdown of measures of central tendency and information with respect to the possible missing values we just mentioned. This is useful to get a sense of how the data is spread. The mean and median help us understand the distribution of our variables. If they are relatively close to each other, it's indicative that the data is likely to be closely divided around the mean, or close to symmetrical. If the data is spread further apart, it's indicative that the distribution for that particular variable is not normally distributed. We can see that there is a pretty broad range when it comes to the age of drivers, which ranges from 16 to 81. We can also observe that the distance to work varies a great deal when considering the average commute time possibly pointing out the potential for outliers. There are some instances of missing values for some of our variables, chief amongst stem the CAR_AGE which has 510 missing values, followed by HOME_VALUE with 464. In these cases of missing or erroneous values, we will need to take steps to impute or fix the data in order to produce a stable predictive model.

Figure 3: Summary Statistics.

Variable	Label	Minimum	Maximum	Mean	Median	Range	N	N Miss
KIDSDRIV	#Driving Children	0	4	0.17	0	4	8161	0
AGE	Age	16	81	44.79	45	65	8155	6
HOMEKIDS	#Children @Home	0	5	0.72	0	5	8161	0
YOJ	Years on Job	0	23	10.50	11	23	7707	454
INCOME	Income	0	367030.26	61898.10	54028.17	367030	7716	445
HOME_VAL	Home Value	0	885282.34	154867.29	161159.53	885282	7697	464
TRAVTIME	Distance to Work	5	142.12	33.49	32.87	137	8161	0
BLUEBOOK	Value of Vehicle	1500	69740	15709.90	14440	68240	8161	0
TIF	Time in Force	1	25	5.35	4	24	8161	0
OLDCLAIM	Total Claims(Past 5 Years)	0	57037	4037.08	0	57037	8161	0
CLM_FREQ	#Claims(Past 5 Years)	0	5	0.80	0	5	8161	0
MVR_PTS	Motor Vehicle Record Points	0	13	1.70	1	13	8161	0
CAR_AGE	Vehicle Age	-3	28	8.33	8	31	7651	510

The Frequency tables check for any possible missing data in our categorical variables. The categorical variables in this data set are (PARENT1), (MSTATUS), (SEX), (EDUCATION), (JOB), (CAR_USE), (CAR_TYPE), (RED_CAR), (REVOKED), and (URBANICITY). It appears the only issue present is within the JOB variable which looks to be missing 526 values.

Figure 4: Jobs Variable Frequency Summary.

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

Recall that if (TARGET_FLAG) has a value of "1" it indicates that the person was in a car crash. With that in mind, we can run a simple correlation against our numeric predictor variables.

Figure 5: Summary Correlation matrix to TARGET_WINS.

Correlation to TARGET_FLAG Variable	Correlation	P-Value	Observations
KIDSDRIV	0.10367	<.0001	8161
AGE	-0.10322	<.0001	8155
HOMEKIDS	0.11562	<.0001	8161
YOJ	-0.07051	<.0001	7707
INCOME	-0.14201	<.0001	7716
HOME_VAL	-0.18374	<.0001	7697
TRAVTIME	0.04815	<.0001	8161
BLUEBOOK	-0.10338	<.0001	8161
TIF	-0.08237	<.0001	8161
OLDCLAIM	0.13808	<.0001	8161
CLM_FREQ	0.21620	<.0001	8161
MVR_PTS	0.21920	<.0001	8161
CAR_AGE	-0.10065	<.0001	7651

Motor Vehicle Record (MVR_PTS) and Number of Claims the Past 5 years (CLM_FREQ) definitely stand out from the field, but there aren't any variables present with excessively high correlations to a positive (TARGET_FLAG) indicating an accident. The P-Values would appear to suggest that they are all statistically significant since they are all well below the.05 threshold. It would be wise to note that age of driver (AGE) shows a negative correlation to the likelihood of an accident. This is interesting because of the extensive range present in the age category. Young drivers tend to be more dangerous, as do older drivers. Perhaps age is dominated by the sweet spot of the mean/median age of the observations.

Data Preparation:
Our exploratory analysis revealed that we had six variables in our dataset with missing variables that need to be addressed before we proceed with modeling.

Figure 6: Summary of Variables with missing data.

Variable	Label	Minimum	Maximum	Mean	N	N Miss
AGE	Age	16	81	44.79	8155	6
YOJ	Years on Job	0	23	10.50	7707	454
INCOME	Income	0	367030	61898.10	7716	445
HOME_VAL	Home Value	0	885282	154867.29	7697	464
CAR_AGE	Vehicle Age	-3	28	8.33	7651	510
JOBS	Job Category	n/a	n/a	n/a	7635	526

Figure .6 provides a quick summary of the variables we previously identified that will require attention. In some cases, it would be prudent to remove variables with missing observations from our dataset completely. In this particular analysis, that doesn't appear to be a necessity and we will replace any missing observations with the variable's mean value. Given the large number of observations for each variable, (greater than 7600) and the relatively small number of missing values (less than or equal to 526) for each variable, simple imputation should suffice. A decision tree could get more precise answers for our missing values, but it shouldn't be necessary here since at a maximum 6% of the variables information is missing. Those adjustments are on display below.

Figure 7: Variable Imputation.

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
TARGET_FLAG		8161	0	0.26	0.44	0	1	1
KIDSDRIV	#Driving Children	8161	0	0.17	0.51	0	4	4
HOMEKIDS	#Children @Home	8161	0	0.72	1.12	0	5	5
TIF	Time in Force	8161	0	5.35	4.15	1	25	24
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	4037.08	8777.14	0	57037	57037
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0.80	1.16	0	5	5
MVR_PTS	Motor Vehicle Record Points	8161	0	1.70	2.15	0	13	13
IMP_AGE		8161	0	44.79	8.62	16	81	65
IMP_INCOME		8161	0	61605.67	45080.66	0	215536	215536
M_INCOME		8161	0	0.05	0.23	0	1	1
IMP_HOME_VAL		8161	0	154309.55	123561.60	0	500309	500309
M_HOME_VAL		8161	0	0.06	0.23	0	1	1
IMP_YOJ		8161	0	10.53	3.98	0	23	23
IMP_CAR_AGE		8161	0	8.33	5.52	0	28	28
M_CAR_AGE		8161	0	0.06	0.24	0	1	1
R_TRAVTIME		8161	0	33.39	15.57	5	75.14	70.14
M_TRAVTIME		8161	0	0.01	0.10	0	1	1
R_BLUEBOOK		8161	0	15661.10	8248.47	1500	39090	37590
M_BLUEBOOK		8161	0	0.01	0.10	0	1	1
JOB_WHITE_COLLAR		8161	0	0.18	0.38	0	1	1

With respect to outlying and extreme data points, we don't feel it's a necessity to adjust for many of those data points at this time. Our data set has enough observations (8161) that we can feel confident to include most of the data points. There are some variables that need to be addressed. We do have some values that could possibly be erroneous, such as CAR_AGE having a minimum age of -3 (Impossible). In the case of income and home value, it is possible for a student to select "0" for the income answer, or for renters to list home value as "0". This is possibly an issue with data gathering and survey building and not outliers. With that said, some variable ranges appears to indicate that there is a high chance that they are not normally distributed around the mean which could hinder our models. A total of five variables were capped in order to bring their values more in line with the other observations. INCOME was capped at $215,536, which is at the 99% level for that variable. HOME_VAL was capped at $500,309 also at the 99% value. TRAVTIME was capped at 75.14 minutes, the 99% level for that variable. BLUEBOOK was capped at $39,090, the 99% level for that variable. Finally, CAR_AGE went through a different adjustment. We mentioned the fact that a value of -3 was impossible for a car's age. What we did was cap any CAR_AGE variable that was less than "0" to a value equal to "0".

The final change to our continuous data comes in the form of a logistic transformation for some of our variables. We typically incorporate logistic transformations for independent variables in an effort to scale it and make it more normally distributed. Total Claims, Number of Claims (Past 5 years), Motor Vehicle record points, Number of driving children, Number of Children at home, and Time in Force will be log transformed to improve these concentrated values. These adjustments are displayed below in Figure.8

Figure 8: Variable adjustments.

Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
IMP_INCOME	8161	0	61605.67	45080.66	0	215536.00	215536.00
IMP_HOME_VAL	8161	0	154309.55	123561.60	0	500309.00	500309.00
R_TRAVTIME	8161	0	33.39	15.57	5	75.14	70.14
R_BLUEBOOK	8161	0	15661.10	8248.47	1500	39090.00	37590.00
IMP_CAR_AGE	8161	0	8.33	5.52	0	28.00	28.00
log_OLDCLAIM	8161	0	3.39	4.31	0	10.95	10.95
log_MVR_PTS	8161	0	0.71	0.74	0	2.64	2.64
log_TIF	8161	0	1.62	0.71	0.69	3.26	2.56
log_KIDSDRIV	8161	0	0.10	0.29	0	1.61	1.61
log_CLM_FREQ	8161	0	0.42	0.56	0	1.79	1.79
log_HOMEKIDS	8161	0	0.38	0.54	0	1.79	1.79

To correct the JOB variable, we could not follow the same rule set for fixing missing values. By definition, these variables aren't numeric, and there is no "Mean" to put into the missing job field. We created a simple rule stating that if income was greater than $125,000 we would attribute the job to a Doctor. If it was greater than $80,000, we would attribute the job to a Lawyer. Any values below that would be considered a "Blue Collar" Job. Below is the comparison of the JOB variable before, and after imputation.

Figure 9: Summary of Imputed Job Variable.

Before

Job	Freq	Percent	Frequency	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

After

IMP_JOB	Freq	Percent	Frequency	Percent
Clerical	1271	15.57	1271	15.57
Doctor	455	5.58	1726	21.15
Home Maker	641	7.85	2368	29.00
Lawyer	978	11.98	3345	40.99
Manager	988	12.11	4333	53.09
Professional	1117	13.69	5450	66.78
Student	712	8.72	6162	75.51
z_Blue Collar	1999	24.49	8161	100

Modeling:

The models we will be developing in this analysis are logistic regression models which allow us to model the relationship between our dependent variable, TARGET_FLAG and one or more independent or predictor variables.

Model#1 Imputed with Mean Values for missing Data: Normal Selection:

This model contains all of the variables in our dataset. In the case of continuous variables, we imputed any missing observations with the mean of each variable, and we followed a simple decision tree to correct the lone categorical variable with missing values. In addition to this, we also have all the log-transformed variables included in this model. No automated selection process was incorporated into this model. This is a very robust and full model. We have 36 total variables involved here, and the categorical variables have different levels within each of them making this model a lot to deal with. Also of note, is that many of these intercepts aren't statistically significant and some exceed our .05 threshold by a considerable amount.

Figure.10 Parameter estimates for logistic Model #1

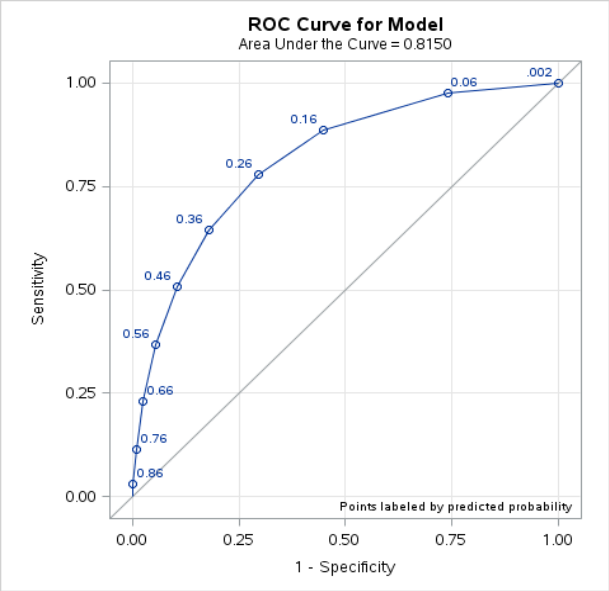
Analysis of Maximum Likelihood Estimates						
Parameter	Label	DF	Estimate	Std Error	Chi-Square	Pr>Chi-Square
Intercept		1	0.275	1.528	0.032	0.857
IMP_AGE		1	0.001	0.004	0.056	0.812
R_BLUEBOOK		1	0.000	0.000	15.700	<.0001
IMP_CAR_AGE		1	0.001	0.008	0.021	0.886
CAR_TYPE	Minivan	1	-0.773	0.114	46.317	<.0001
CAR_TYPE	Panel Truck	1	-0.169	0.213	0.629	0.428
CAR_TYPE	Pickup	1	-0.189	0.120	2.477	0.116
CAR_TYPE	Sports Car	1	0.274	0.099	7.699	0.006
CAR_TYPE	Van	1	-0.201	0.166	1.466	0.226

CAR_USE	**Commercial**	1	0.764	0.094	66.264	<.0001
CLM_FREQ		1	0.237	0.314	0.568	0.451
EDUCATION	**<High School**	1	-0.015	0.096	0.026	0.872
EDUCATION	**Bachelors**	1	-0.391	0.091	18.437	<.0001
EDUCATION	**Masters**	1	-0.348	0.167	4.361	0.037
EDUCATION	**PhD**	1	0.021	0.216	0.009	0.923
HOMEKIDS		1	0.117	0.359	0.106	0.745
HIGH_OLDCLAIM		1	0.198	0.133	2.196	0.138
IMP_HOME_VAL		1	0.000	0.000	11.923	0.001
IMP_INCOME		1	0.000	0.000	11.994	0.001
IMP_JOB	**Clerical**	1	0.064	0.108	0.354	0.552
IMP_JOB	**Doctor**	1	-0.917	0.304	9.128	0.003
IMP_JOB	**Home Maker**	1	-0.147	0.156	0.880	0.348
IMP_JOB	**Lawyer**	1	-0.180	0.191	0.886	0.347
IMP_JOB	**Manager**	1	-0.875	0.141	38.405	<.0001
IMP_JOB	**Professional**	1	-0.144	0.121	1.413	0.235
IMP_JOB	**Student**	1	-0.143	0.133	1.167	0.280
KIDSDRIV		1	0.049	1.181	0.002	0.967
MSTATUS	**Yes**	1	-0.561	0.091	37.753	<.0001
MVR_PTS		1	0.280	0.079	12.630	0.000
OLDCLAIM		1	0.000	0.000	10.959	0.001
PARENT1	**No**	1	-0.201	0.125	2.581	0.108
RED_CAR	**no**	1	0.089	0.093	0.923	0.337
REVOKED	**No**	1	-0.954	0.097	96.819	<.0001
SEX	**M**	1	0.139	0.117	1.415	0.234
TIF		1	-0.004	0.024	0.032	0.857
R_TRAVTIME		1	0.015	0.002	59.014	<.0001
URBANICITY	**Highly Urban/ Urban**	1	2.339	0.113	425.042	<.0001
IMP_YOJ		1	-0.011	0.009	1.621	0.203
log_OLDCLAIM		1	0.022	0.113	0.037	0.847
log_MVR_PTS		1	-0.823	0.342	5.782	0.016
log_TIF		1	-0.300	0.139	4.698	0.030
log_KIDSDRIV		1	0.551	3.203	0.030	0.864
log_CLM_FREQ		1	-0.551	0.962	0.329	0.566
log_HOMEKIDS		1	-0.552	1.097	0.253	0.615
No_CLM_FREQ		1	-0.514	0.981	0.275	0.600
No_HOMEKIDS		1	-0.596	0.470	1.606	0.205
No_KIDSDRIV		1	0.005	1.089	0.000	0.997
No_PTS		1	-0.523	0.227	5.330	0.021

For the most part, Model#1 looks ok. In most cases, the parameter estimates displayed in Figure.9 coincide with the effect we believe they should have on the probability of a crash. This is highlighted in the professional fields, such as Doctor, Lawyer, Manager, Professional, all reducing the likelihood of being in an accident. High Old Claims, and being a Male is shown to increase the

possibility of a crash which makes perfect sense. There are some peculiarities. Being a Ph.D. apparently increases the chances of being in an accident. Perhaps too smart for their own good.

Figure 11: Model #1 ROC Curve and Fit Diagnostics.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	6823.212
SC	8825.562	7156.356
-2 Log L	8816.622	6727.212

The ROC Curve for our first model indicates a decent fit. The curve shows the tradeoff between sensitivity and specificity. The closer our curve is to the left-hand border and the top border of the ROC space, the more accurate our model is, and by contrast the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the model. The area under the curve (AUC) is equal to 81.5%. The closer AUC is to 100%, the better the overall diagnostic performance of the test. At 82% this should be considered a good fit. The AIC is 8818.62. We generally want that number to be as low as possible, but it's a comparative measurement, so we will use it to see how well our other models fit. The log likelihood is 8816.62 which is the criterion for selecting parameters in logistic regression. It serves a similar function to R-Squared and F-value. We want to maximize log likelihood so we generally want to see this number high. These two metrics will serve a part in our final model selection criterion.

Model#2 Imputed with Mean Values for missing Data: Forward Selection:

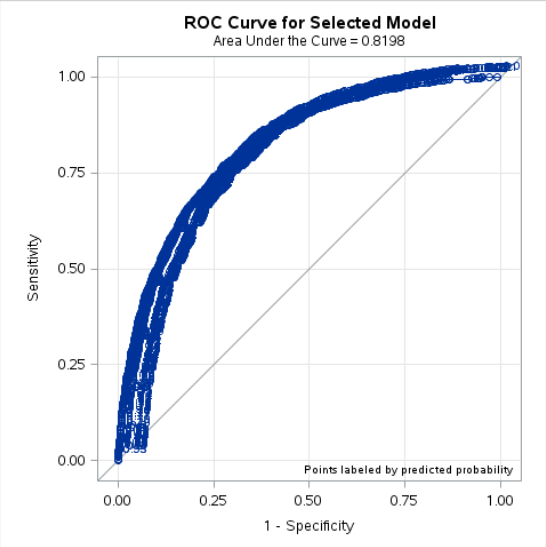
This model contains all of the variables and data adjustments in our dataset, but we incorporated Forward selection, an automated variable selection process. At each step, a variable that is not already in the model is tested for inclusion into the model. The most significant of these variables is added to the model based on the variable's overall impact on the statistical significance of the model.

Figure.12 Parameter estimates for logistic Model #2

Analysis of Maximum Likelihood Estimates						
Parameter	**Label**	**DF**	**Estimate**	**Std Error**	**Chi-Square**	**Pr>Chi-Square**
Intercept		1	-0.426	0.242	3.085	0.079
R_BLUEBOOK		1	0.000	0.000	22.944	<.0001
CAR_TYPE	**Minivan**	1	-0.714	0.087	67.548	<.0001
CAR_TYPE	**Panel Truck**	1	-0.072	0.169	0.180	0.671
CAR_TYPE	**Pickup**	1	-0.133	0.096	1.947	0.163
CAR_TYPE	**Sports Car**	1	0.276	0.098	7.888	0.005
CAR_TYPE	**Van**	1	-0.115	0.128	0.813	0.367
CAR_USE	**Commercial**	1	0.774	0.094	68.294	<.0001
EDUCATION	**<High School**	1	-0.021	0.095	0.047	0.829
EDUCATION	**Bachelors**	1	-0.385	0.085	20.737	<.0001
EDUCATION	**Masters**	1	-0.333	0.149	4.989	0.026
EDUCATION	**PhD**	1	0.037	0.204	0.032	0.858
IMP_HOME_VAL		1	0.000	0.000	12.082	0.001
IMP_INCOME		1	0.000	0.000	13.270	0.000
IMP_JOB	**Clerical**	1	0.063	0.108	0.343	0.558
IMP_JOB	**Doctor**	1	-0.877	0.302	8.455	0.004
IMP_JOB	**Home Maker**	1	-0.103	0.147	0.486	0.486
IMP_JOB	**Lawyer**	1	-0.164	0.190	0.743	0.389
IMP_JOB	**Manager**	1	-0.857	0.141	37.207	<.0001
IMP_JOB	**Professional**	1	-0.135	0.121	1.244	0.265
IMP_JOB	**Student**	1	-0.108	0.126	0.735	0.391
MSTATUS	**Yes**	1	-0.569	0.091	39.377	<.0001
MVR_PTS		1	0.100	0.015	47.010	<.0001
OLDCLAIM		1	0.000	0.000	22.637	<.0001
PARENT1	**No**	1	-0.216	0.125	2.995	0.084
REVOKED	**No**	1	-0.944	0.096	95.795	<.0001
R_TRAVTIME		1	0.015	0.002	59.800	<.0001
URBANICITY	**Highly Urban/ Urban**	1	2.334	0.113	424.631	<.0001
log_TIF		1	-0.324	0.043	56.541	<.0001
log_KIDSDRIV		1	0.608	0.109	31.178	<.0001
No_CLM_FREQ		1	-0.616	0.082	57.094	<.0001
No_HOMEKIDS		1	-0.218	0.091	5.744	0.017

Immediately we see that Model#2 is much easier to conceptualize than the first model. For the most part, the parameter estimates displayed in Figure.11 coincide with the effect we believe they should have on the probability of a crash and more intercepts are statistically significant now.

Figure 13: Model #2 ROC Curve and Fit Diagnostics.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	6808.241
SC	8825.562	7030.337
-2 Log L	8816.622	6744.241

The ROC Curve for our second model is slightly better than our first. This ROC curve shows all the different models built in the selection process. The area under the curve (AUC) is equal to ~82% overall. Once again, the closer AUC is to 100%, the better the overall diagnostic performance of the test. At 82% this should be considered a good fit. The AIC is 8818.62. and the log likelihood is 8816.62.

Model#3 Imputed with Mean Values for missing Data: Stepwise Selection:

We used the same variable inputs in Model#3, but we incorporated a Stepwise automated variable selection procedure, which adds or removes predictor variables in steps based on the impact to the statistical significance of the model

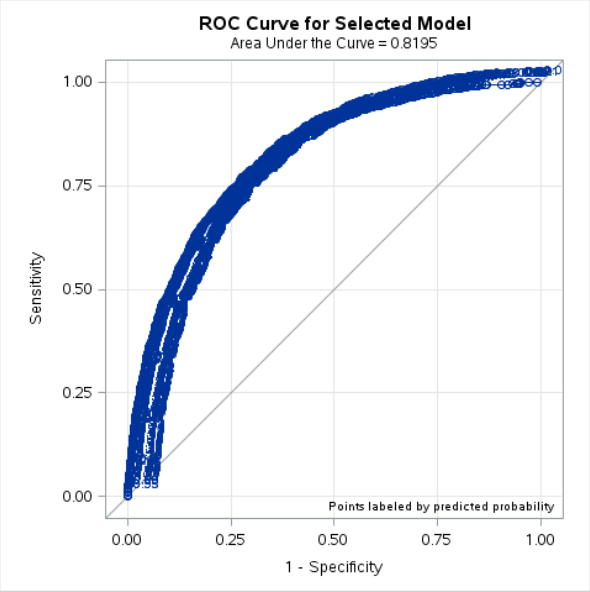
Figure.14 Parameter estimates for logistic Model #3

Analysis of Maximum Likelihood Estimates						
Parameter	Label	DF	Estimate	Std Error	Chi-Square	Pr>ChiSquare
Intercept		1	-0.492	0.239	4.234	0.040
R_BLUEBOOK		1	0.000	0.000	22.585	<.0001
CAR_TYPE	Minivan	1	-0.713	0.087	67.463	<.0001
CAR_TYPE	Panel Truck	1	-0.075	0.169	0.196	0.658
CAR_TYPE	Pickup	1	-0.134	0.096	1.954	0.162
CAR_TYPE	Sports Car	1	0.279	0.098	8.080	0.005
CAR_TYPE	Van	1	-0.118	0.128	0.856	0.355
CAR_USE	Commercial	1	0.775	0.094	68.490	<.0001
EDUCATION	<High School	1	-0.022	0.095	0.052	0.821
EDUCATION	Bachelors	1	-0.383	0.085	20.538	<.0001
EDUCATION	Masters	1	-0.333	0.149	5.000	0.025
EDUCATION	PhD	1	0.040	0.204	0.039	0.844
IMP_HOME_VAL		1	0.000	0.000	11.548	0.001
IMP_INCOME		1	0.000	0.000	13.900	0.000
IMP_JOB	Clerical	1	0.063	0.108	0.337	0.561
IMP_JOB	Doctor	1	-0.884	0.302	8.589	0.003
IMP_JOB	Home Maker	1	-0.104	0.147	0.498	0.481
IMP_JOB	Lawyer	1	-0.162	0.190	0.725	0.395
IMP_JOB	Manager	1	-0.852	0.141	36.757	<.0001
IMP_JOB	Professional	1	-0.132	0.121	1.200	0.273

IMP_JOB	**Student**	1	-0.111	0.126	0.774	0.379
MSTATUS	**Yes**	1	-0.658	0.074	78.396	<.0001
MVR_PTS		1	0.101	0.015	47.329	<.0001
OLDCLAIM		1	0.000	0.000	22.969	<.0001
REVOKED	**No**	1	-0.946	0.096	96.295	<.0001
R_TRAVTIME		1	0.015	0.002	59.581	<.0001
URBANICITY	**Highly Urban/ Urban**	1	2.328	0.113	424.612	<.0001
log_TIF		1	-0.325	0.043	56.824	<.0001
log_KIDSDRIV		1	0.599	0.108	30.523	<.0001
No_CLM_FREQ		1	-0.615	0.082	56.952	<.0001
No_HOMEKIDS		1	-0.315	0.071	19.675	<.0001

Model#3 appears to be very similar to Model#2. Coming out of their respective automatic selection processes, they each produced a similar amount of intercepts and ROC curves. The intercepts, for the most part, agree with our understanding of what would be considered an increased risk for accidents, such as (CAR_USE-Commercial) and (URBANCITY), but it still doesn't make sense why a Ph.D. would put you more at risk. There is also an issue with the fact that despite the inclusion of an automated selection process, we still see intercepts that are not significant.

Figure 15: Model #3 ROC Curve and FitDiagnostics.



Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	8818.622	6808.239
SC	8825.562	7024.395
-2 Log L	8816.622	6747.239

The ROC Curve for our third model is slightly better than our first but worse than our second model. In either case, the difference is entirely negligible. The ROC curve shows all the different models built into the selection process. The area under the curve (AUC) is equal to ~82%. We know the closer AUC is to 100%, the better the overall diagnostic performance of the test. At ~82% this should be considered a good fit. The AIC is 8818.62. and the log likelihood is equal to 8816.62.

Model#4 Imputed with Mean Values for missing Data: PROBIT regression.

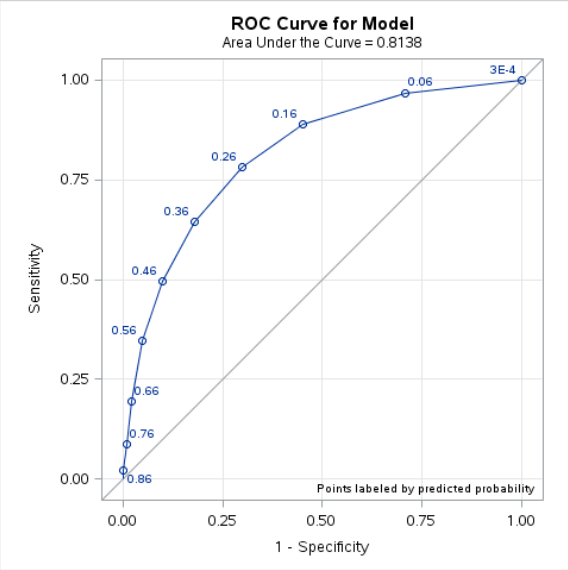
In Model #4. we used the same variable inputs as Model #1, but this time we ran a PROBIT Regression model. PROBIT regression is used to model binary outcome variables, such as the ones in our data set.

Figure.16 Parameter estimates for logistic PROBIT Model #4

Analysis of Maximum Likelihood Estimates						
Parameter	**Label**	**DF**	**Estimate**	**Std Error**	**Chi-Square**	**Pr>ChiSquare**
Intercept		1	0.248	0.902	0.075	0.784
IMP_AGE		1	0.001	0.002	0.124	0.725
R_BLUEBOOK		1	0.000	0.000	15.633	<.0001
IMP_CAR_AGE		1	0.000	0.005	0.001	0.972
CAR_TYPE	**Minivan**	1	-0.441	0.064	46.894	<.0001
CAR_TYPE	**Panel Truck**	1	-0.106	0.123	0.740	0.390
CAR_TYPE	**Pickup**	1	-0.120	0.069	3.016	0.082
CAR_TYPE	**Sports Car**	1	0.159	0.058	7.678	0.006
CAR_TYPE	**Van**	1	-0.125	0.095	1.722	0.190
CAR_USE	**Commercial**	1	0.432	0.055	62.671	<.0001
CLM_FREQ		1	0.151	0.186	0.659	0.417
EDUCATION	**<High School**	1	-0.015	0.056	0.073	0.787
EDUCATION	**Bachelors**	1	-0.231	0.053	19.250	<.0001
EDUCATION	**Masters**	1	-0.187	0.094	3.906	0.048
EDUCATION	**PhD**	1	0.010	0.123	0.006	0.936
HOMEKIDS		1	0.062	0.210	0.088	0.767
HIGH_OLDCLAIM		1	0.119	0.079	2.260	0.133
IMP_HOME_VAL		1	0.000	0.000	10.604	0.001
IMP_INCOME		1	0.000	0.000	11.282	0.001
IMP_JOB	**Clerical**	1	0.036	0.063	0.318	0.573
IMP_JOB	**Doctor**	1	-0.535	0.168	10.176	0.001
IMP_JOB	**Home Maker**	1	-0.090	0.091	0.988	0.320
IMP_JOB	**Lawyer**	1	-0.122	0.108	1.272	0.259
IMP_JOB	**Manager**	1	-0.496	0.080	38.428	<.0001
IMP_JOB	**Professional**	1	-0.085	0.071	1.463	0.226
IMP_JOB	**Student**	1	-0.062	0.077	0.652	0.420
KIDSDRIV		1	-0.014	0.692	0.000	0.983
MSTATUS	**Yes**	1	-0.330	0.052	39.667	<.0001
MVR_PTS		1	0.165	0.046	12.988	0.000
OLDCLAIM		1	0.000	0.000	9.310	0.002
PARENT1	**No**	1	-0.111	0.073	2.352	0.125
RED_CAR	**no**	1	0.0456	0.0534	0.7306	0.3927
REVOKED	**No**	1	-0.5459	0.0564	93.6662	<.0001
SEX	**M**	1	0.0836	0.0666	1.5736	0.2097
TIF		1	-0.00514	0.0138	0.1384	0.7098
R_TRAVTIME		1	0.00886	0.00115	59.701	<.0001
URBANICITY	**Highly Urban/ Urban**	1	1.2867	0.059	475.4354	<.0001
IMP_YOJ		1	-0.00612	0.00509	1.4418	0.2298
log_OLDCLAIM		1	0.00394	0.0673	0.0034	0.9533
log_MVR_PTS		1	-0.4833	0.1988	5.908	0.0151
log_TIF		1	-0.1623	0.0795	4.1692	0.0412
log_KIDSDRIV		1	0.4059	1.8768	0.0468	0.8288
log_CLM_FREQ		1	-0.3607	0.5694	0.4012	0.5265
log_HOMEKIDS		1	-0.3026	0.6409	0.2229	0.6369
No_CLM_FREQ		1	-0.3891	0.5841	0.4439	0.5053
No_HOMEKIDS		1	-0.3451	0.2744	1.5809	0.2086
No_KIDSDRIV		1	0.0208	0.6375	0.0011	0.974
No_MVR_PTS		1	-0.2984	0.1311	5.1816	0.0228

Model#4 looks very similar to Model#1. There are a lot of variables here to work through and once again, many of them do not appear to bear any statistical significance to our model. However, the intercepts are mostly intuitive and in line with the exception of PhD.

Figure 17: Model #4 ROC Curve and FitDiagnostics.



Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	8818.622	6834.638
SC	8825.562	7167.782
-2 Log L	8816.622	6738.638

The ROC Curve for our fourth model is very similar to our first model. This isn't surprising, given that all of the same predictor variables are in this model. The ROC curve shows the area under the curve (AUC) is equal to 81.4% overall. At ~81% this is considered a good fit. The AIC is 8818.62. and the log likelihood is 8816.62.

Model Selection:
We were able to run four different models based on variable adjustments, different compositions, manual selection procedures and automated selection processes. The table below provides some goodness of fit metrics we previously discussed during the modeling process.

Figure.18 Model Comparison

Model	Number of Variables	AIC	Log-Likelihood	AUC	Ks
(1) Mean Imp Variable adjustment all Variables	47	8818.62	8816.62	0.815	0.211399
(2) Mean Imp Variable adjustment Forward Selection	31	8818.62	8816.62	0.8198	0.211774
(3) Mean Imp Variable adjustment Stepwise selection	30	8818.62	8816.62	0.8195	0.211577
(4) Mean Imp Variable adjustment Probit Regression	47	8818.62	8816.62	0.8138	0.211688

When we first started this process, we mentioned that we would base our model selection on AIC, AUC, Log Likelihood, and Ks Statistic. The Ks-Statistic provides us with a useful metric when dealing with continuous distributions, discrete values, and non-classified data. These values appear to indicate that each of them is performing better than a randomly selected model. The AUC is best when it approaches 100%. In each of these models the AUC is similar, but Model #2, built with the Forward selection, has a slight edge over the other models. We want the lowest possible AIC values when comparing the goodness of fit. All four models have the same AIC, so that metric is negligible for our selection criteria. Log Likelihood is similar to AIC, and once again, we see that we have the same values across all of our models.

Taking into consideration the information we just presented and considering the need for choosing a model that not only satisfies our predictive requirements but also delivers ease of use, we believe Model #3 to be the best fit. Model #3 is a logistic regression model that incorporates Stepwise automated selection. This model had the least amount of variables to consider and had the second highest Area Under the Curve (AUC). For these reasons, we feel this model will be the most productive moving forward with our analysis.

Conclusion:

After going through and analyzing the data with various modeling techniques, we generated three different logistic regression and one Probit regression model. We examined our data for inconsistencies, such as missing observations or extreme outliers, and made imputations that were in line with the size of our data set and our goals for the analysis. As is standard, the modeling process was challenging. We had models with very good and similar goodness of fit attributes that made our decisions challenging. In this particular exercise, we decided that a model with fewer variables would be more intuitive to the end user, which is the ultimate goal of data science. Perhaps looking forward we can incorporate different modeling techniques, such as Factor Analysis, and Principal Component Analysis in an effort to streamline our model and make it more efficient.

BONUS

PROC GENMOD:

I ran PROCGENMOD using the same parameters as my best Model, which was Model #3. Below are the outputs.

Figure.19Proc GenMod

Analysis of Maximum Likelihood Estimates								
Parameter	Label	DF	Estimate	Std Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept		1	0.5753	0.2498	0.0857	1.0649	5.3	0.0213
IMP_AGE		1	0.0001	0.0006	-0.0012	0.0013	0.01	0.9246
R_BLUEBOOK		1	0	0	0	0	13.66	0.0002
IMP_CAR_AGE		1	0	0.0011	-0.0023	0.0022	0	0.9834
CAR_TYPE	Minivan	1	-0.1018	0.0154	-0.1319	-0.0717	43.83	<.0001
CAR_TYPE	Panel Truck	1	-0.0425	0.0307	-0.1027	0.0178	1.91	0.1673
CAR_TYPE	Pickup	1	-0.027	0.0172	-0.0608	0.0068	2.45	0.1176
CAR_TYPE	Sports Car	1	0.041	0.0152	0.0112	0.0707	7.29	0.0069
CAR_TYPE	Van	1	-0.0359	0.0234	-0.0818	0.0099	2.36	0.1242
CAR_USE	Commercial	1	0.1204	0.0142	0.0925	0.1482	71.65	<.0001
CLM_FREQ		1	0.0385	0.0525	-0.0644	0.1413	0.54	0.4639
EDUCATION	<High School	1	-0.0053	0.0146	-0.0339	0.0233	0.13	0.7149
EDUCATION	Bachelors	1	-0.0634	0.0135	-0.0898	-0.037	22.14	<.0001
EDUCATION	Masters	1	-0.0485	0.0229	-0.0934	-0.0036	4.49	0.0342
EDUCATION	PhD	1	-0.0067	0.0298	-0.0651	0.0517	0.05	0.8214
HOMEKIDS		1	0.0176	0.0554	-0.0911	0.1263	0.1	0.7508
HIGH_OLDCLAIM		1	0.0336	0.0225	-0.0104	0.0776	2.24	0.1349
IMP_HOME_VAL		1	0	0	0	0	8.28	0.004
IMP_INCOME		1	0	0	0	0	7.26	0.0071
IMP_JOB	Clerical	1	0.0119	0.0164	-0.0203	0.0441	0.52	0.4689
IMP_JOB	Doctor	1	-0.1367	0.0387	-0.2126	-0.0608	12.46	0.0004
IMP_JOB	Home Maker	1	-0.0116	0.0231	-0.0568	0.0337	0.25	0.616
IMP_JOB	Lawyer	1	-0.0459	0.0266	-0.098	0.0061	2.99	0.0839
IMP_JOB	Manager	1	-0.1405	0.02	-0.1796	-0.1014	49.59	<.0001
IMP_JOB	Professional	1	-0.0275	0.0181	-0.063	0.0081	2.29	0.13
IMP_JOB	Student	1	-0.0091	0.0201	-0.0485	0.0303	0.2	0.6519
KIDSDRIV		1	-0.0005	0.1907	-0.3743	0.3733	0	0.9979
MSTATUS	Yes	1	-0.0755	0.0131	-0.1012	-0.0499	33.33	<.0001
MVR_PTS		1	0.0619	0.0122	0.0381	0.0858	25.86	<.0001
OLDCLAIM		1	0	0	0	0	10.56	0.0012
PARENT1	No	1	-0.0545	0.0189	-0.0916	-0.0175	8.32	0.0039
RED_CAR	no	1	0.0162	0.0134	-0.01	0.0424	1.46	0.2262

REVOKED	**No**	1	-0.159	0.0153	-0.1891	-0.129	107.45	<.0001
SEX	**M**	1	0.0203	0.0161	-0.0112	0.0519	1.6	0.2065
TIF		1	0.0001	0.0034	-0.0065	0.0067	0	0.9725
R_TRAVTIME		1	0.0021	0.0003	0.0015	0.0027	53.69	<.0001
URBANICITY	**Highly Urban/ Urban**	1	0.2932	0.0121	0.2696	0.3168	590.99	<.0001
IMP_YOJ		1	-0.0021	0.0013	-0.0047	0.0004	2.73	0.0987
log_OLDCLAIM		1	0.0027	0.0191	-0.0347	0.0402	0.02	0.8858
log_MVR_PTS		1	-0.1787	0.0522	-0.2811	-0.0764	11.72	0.0006
log_TIF		1	-0.0482	0.0198	-0.0869	-0.0095	5.95	0.0147
log_KIDSDRIV		1	0.1193	0.5155	-0.8911	1.1296	0.05	0.817
log_CLM_FREQ		1	-0.0809	0.1607	-0.3959	0.234	0.25	0.6144
log_HOMEKIDS		1	-0.0825	0.1696	-0.4148	0.2498	0.24	0.6265
No_CLM_FREQ		1	-0.0752	0.1654	-0.3994	0.249	0.21	0.6494
No_HOMEKIDS		1	-0.0807	0.0725	-0.2227	0.0614	1.24	0.2657
No_KIDSDRIV		1	0.0202	0.1743	-0.3213	0.3618	0.01	0.9076
No_MVR_PTS		1	-0.094	0.0337	-0.1601	-0.0279	7.76	0.0053
Scale		1	0.3843	0.0031	0.3783	0.3904		

Figure.20 Goodness of fit Criterion

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	7587	1127.6061	0.1486
Scaled Deviance	7587	7635	1.0063
Pearson Chi-Square	7587	1127.6061	0.1486
Scaled Pearson X2	7587	7635	1.0063
Log Likelihood		-3532.0693	
Full Log Likelihood		-3532.0693	
AIC (smaller is better)		7162.1386	
AICC (smaller is better)		7162.7846	
BIC (smaller is better)		7502.223	

Model Comparisons:

GENMOD

Analysis of Maximum Likelihood Estimates			
Parameter	Label	DF	Estimate
Intercept		1	0.5753
IMP_AGE		1	0.0001
R_BLUEBOOK		1	0
IMP_CAR_AGE		1	0
CAR_TYPE	Minivan	1	-0.1018
CAR_TYPE	Panel Truck	1	-0.0425
CAR_TYPE	Pickup	1	-0.027
CAR_TYPE	Sports Car	1	0.041
CAR_TYPE	Van	1	-0.0359
CAR_USE	Commercial	1	0.1204
CLM_FREQ		1	0.0385
EDUCATION	<High School	1	-0.0053
EDUCATION	Bachelors	1	-0.0634
EDUCATION	Masters	1	-0.0485
EDUCATION	PhD	1	-0.0067
HOMEKIDS		1	0.0176
HIGH_OLDCLAIM		1	0.0336
IMP_HOME_VAL		1	0
IMP_INCOME		1	0
IMP_JOB	Clerical	1	0.0119
IMP_JOB	Doctor	1	-0.1367
IMP_JOB	Home Maker	1	-0.0116
IMP_JOB	Lawyer	1	-0.0459
IMP_JOB	Manager	1	-0.1405
IMP_JOB	Professional	1	-0.0275
IMP_JOB	Student	1	-0.0091
KIDSDRIV		1	-0.0005
MSTATUS	Yes	1	-0.0755
MVR_PTS		1	0.0619
OLDCLAIM		1	0
PARENT1	No	1	-0.0545
RED_CAR	no	1	0.0162
REVOKED	No	1	-0.159
SEX	M	1	0.0203
TIF		1	0.0001
R_TRAVTIME		1	0.0021
URBANICITY	Highly Urban/	1	0.2932
IMP_YOJ		1	-0.0021
log_OLDCLAIM		1	0.0027
log_MVR_PTS		1	-0.1787
log_TIF		1	-0.0482
log_KIDSDRIV		1	0.1193
log_CLM_FREQ		1	-0.0809
log_HOMEKIDS		1	-0.0825
No_CLM_FREQ		1	-0.0752
No_HOMEKIDS		1	-0.0807
No_KIDSDRIV		1	0.0202
No_MVR_PTS		1	-0.094
Scale		1	0.3843

LOGISTIC

Analysis of Maximum Likelihood Estimates			
Parameter	Label	DF	Estimate
Intercept		1	-0.492
R_BLUEBOOK		1	0
CAR_TYPE	Minivan	1	-0.713
CAR_TYPE	Panel Truck	1	-0.075
CAR_TYPE	Pickup	1	-0.134
CAR_TYPE	Sports Car	1	0.279
CAR_TYPE	Van	1	-0.118
CAR_USE	Commercial	1	0.775
EDUCATION	<High School	1	-0.022
EDUCATION	Bachelors	1	-0.383
EDUCATION	Masters	1	-0.333
EDUCATION	PhD	1	0.04
IMP_HOME_VAL		1	0
IMP_INCOME		1	0
IMP_JOB	Clerical	1	0.063
IMP_JOB	Doctor	1	-0.884
IMP_JOB	Home Maker	1	-0.104
IMP_JOB	Lawyer	1	-0.162
IMP_JOB	Manager	1	-0.852
IMP_JOB	Professional	1	-0.132
IMP_JOB	Student	1	-0.111
MSTATUS	Yes	1	-0.658
MVR_PTS		1	0.101
OLDCLAIM		1	0
REVOKED	No	1	-0.946
R_TRAVTIME		1	0.015
URBANICITY	Highly Urban/	1	2.328
log_TIF		1	-0.325
log_KIDSDRIV		1	0.599
No_CLM_FREQ		1	-0.615
No_HOMEKIDS		1	-0.315

Proc GENMOD appears to include every variable in the analysis. The intercepts are slightly different and that's something of interest. What's also interesting is that it produced better goodness of fit metrics. The AIC in PROCGENMOD = 7162.13 while the AIC in my third model came in at 8818.62. The lower the AIC, the better fitting model.

Appendix: SAS Code

```
/****Load data***/
%let ME =  noeflores20160;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT02/HW;
%let NAME = HW;
%let LIB = &NAME.;
libname &NAME. "&PATH.";
%let INFILE = HW.LOGIT_INSURANCE;
%let INFILE2 = HW.LOGIT_INSURANCE_TEST;

/***Check contents of data***/
proc contents data=&INFILE order=varnum;
run;

proc print data=&INFILE.(obs=6);
run;

data mydata;
set &INFILE.;
drop INDEX;
drop TARGET_AMT;
run;

proc print data= mydata(obs=7);
run;

/*****proc means on numeric variables******/
proc means data= mydata MIN MAX MEAN MEDIAN MODE std N NMISS;
var _numeric_ ;
run;

/*****proc means without TARGET_FLAG********/
proc means data= mydata MIN MAX MEAN MEDIAN RANGE N NMISS;
var KIDSDRIV
AGE
HOMEKIDS
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
CLM_FREQ
MVR_PTS
CAR_AGE
;
run;

/***proc freq for char variables***/
proc freq data= mydata;
table _character_ /missing;
run;

/***Proc Corr to check for initial correlations***/
```

```
proc corr data=mydata;
var _numeric_;
with TARGET_FLAG;
run;

proc univariate normal plot data=mydata;
var _numeric_;
RUN;


/*****proc means on numeric variables******/
proc means data= mydata MIN MAX MEAN N NMISS;
var AGE YOJ INCOME HOME_VAL CAR_AGE;
run;

/********Impute missing variables****/
data SCRUBFILE;
set mydata;

/********Income******/

IMP_AGE = AGE;
if missing( IMP_AGE ) then IMP_AGE = 44.8;

IMP_INCOME    = INCOME;
M_INCOME      = 0;
if missing(IMP_INCOME) then do;
IMP_INCOME = 61898;
M_INCOME = 1;
end;
if IMP_INCOME > 215536
then IMP_INCOME = 215536;

IMP_HOME_VAL = HOME_VAL;
M_HOME_VAL = 0;
if missing(IMP_HOME_VAL) then do;
IMP_HOME_VAL = 154867;
M_HOME_VAL = 1;
end;
if IMP_HOME_VAL > 500309
then IMP_HOME_VAL = 500309;

IMP_YOJ = YOJ;
if missing( IMP_YOJ )
then IMP_YOJ = 11;

IMP_CAR_AGE = CAR_AGE;
M_CAR_AGE     = 0;
if missing(IMP_CAR_AGE) then do;
IMP_CAR_AGE = 8.33;
M_CAR_AGE = 1;
end;
if IMP_CAR_AGE < 0
then IMP_CAR_AGE = 1;

R_TRAVTIME = TRAVTIME;
```

```
M_TRAVTIME  = 0;
if R_TRAVTIME > 75.14 then do;
R_TRAVTIME = 75.14;
M_TRAVTIME = 1;
end;

R_BLUEBOOK = BLUEBOOK;
M_BLUEBOOK  = 0;
if R_BLUEBOOK>39090 then do;
R_BLUEBOOK = 39090;
M_BLUEBOOK = 1;
end;


IMP_JOB = JOB;
if missing(IMP_JOB) then do;
if IMP_INCOME > 125000 then
IMP_JOB = "Doctor";
else if IMP_INCOME > 80000 then
IMP_JOB = "Lawyer";
        else
IMP_JOB = "z_Blue Collar";
end;

JOB_WHITE_COLLAR = IMP_JOB in ("Doctor","Lawyer");

drop AGE;
drop INCOME;
drop HOME_VAL;
drop YOJ;
drop CAR_AGE;
drop BLUEBOOK;
drop TRAVTIME;

log_OLDCLAIM = log(OLDCLAIM+1);
HIGH_OLDCLAIM = 0;
if OLDCLAIM > 6000 then HIGH_OLDCLAIM = 1;

log_TIF = log(TIF+1);

log_MVR_PTS = log(MVR_PTS+1);
No_MVR_PTS = 0;
if MVR_PTS = 0 then No_MVR_PTS = 1;

log_KIDSDRIV = log(KIDSDRIV+1);
No_KIDSDRIV = 0;
if KIDSDRIV = 0 then No_KIDSDRIV = 1;

log_CLM_FREQ = log(CLM_FREQ+1);
No_CLM_FREQ = 0;
if CLM_FREQ = 0 then No_CLM_FREQ = 1;

log_CLM_FREQ = log(CLM_FREQ+1);
No_CLM_FREQ = 0;
if CLM_FREQ = 0 then No_CLM_FREQ = 1;
```

```
log_HOMEKIDS = log(HOMEKIDS+1);
No_HOMEKIDS = 0;
if HOMEKIDS = 0 then No_HOMEKIDS = 1;
run;


/*****************************************************************************/
proc means data=SCRUBFILE N nmiss Mean std MIN MAX RANGE;
var _numeric_;
run;

/*****************************************************************************/
proc means data=SCRUBFILE N nmiss Mean std MIN MAX RANGE;
var IMP_INCOME IMP_HOME_VAL R_TRAVTIME R_BLUEBOOK IMP_CAR_AGE log_OLDCLAIM
log_MVR_PTS
log_TIF log_KIDSDRIV log_CLM_FREQ ;
run;

/*****************************************************************************/
proc freq data=SCRUBFILE;
table _character_ /missing;
run;

proc univariate normal plot data=SCRUBFILE;
var _numeric_;
RUN;


/*****************************************************************************/
proc print data=SCRUBFILE(obs=8);
run;

/*****************************************************************************/
proc freq data=SCRUBFILE;
table ( _character_ ) * TARGET_FLAG /missing;
run;

/*****************************************************************************/
proc means data=SCRUBFILE n nmiss mean median;
var _numeric_;
run;

/*****************************************************************************/
proc univariate data=SCRUBFILE;
class TARGET_FLAG;
var IMP_INCOME;
histogram;
run;

/*******proc contents****/
proc contents data=scrubfile;
run;

/*******Model 1 Logictic all variables *******/
proc logistic data=SCRUBFILE plot(only)=(roc(ID=prob));
class _character_ /param=ref;
```

```
model TARGET_FLAG( ref="0" ) =
                                        IMP_AGE
                                        R_BLUEBOOK
                                        IMP_CAR_AGE
                                        CAR_TYPE
                                        CAR_USE
                                        CLM_FREQ
                                        EDUCATION
                                        HOMEKIDS
                                        HIGH_OLDCLAIM
                                        IMP_HOME_VAL
                                        IMP_INCOME
                                        IMP_JOB
                                        KIDSDRIV
                                        MSTATUS
                                        MVR_PTS
                                        OLDCLAIM
                                        PARENT1
                                        RED_CAR
                                        REVOKED
                                        SEX
                                        TIF
                                        R_TRAVTIME
                                        URBANICITY
                                        IMP_YOJ
                                        log_OLDCLAIM
                                        log_MVR_PTS
                                        log_TIF
                                        log_KIDSDRIV
                                        log_CLM_FREQ
                                        log_HOMEKIDS
                                        No_CLM_FREQ
                                        No_HOMEKIDS
                                        No_KIDSDRIV
                                        No_MVR_PTS
                                  /roceps=0.1
                                        ;
/*score data=scrubfile*/output out=SCORED p=p_target;
run;


/*********Model 2 Logistic Forward selection***********/
proc logistic data=scrubfile plot(only)=(roc(ID=prob));
class _character_ /param=ref;
model TARGET_FLAG( ref="0" ) =
                                        IMP_AGE
                                        R_BLUEBOOK
                                        IMP_CAR_AGE
                                        CAR_TYPE
                                        CAR_USE
                                        CLM_FREQ
                                        EDUCATION
                                        HOMEKIDS
                                        HIGH_OLDCLAIM
                                        IMP_HOME_VAL
                                        IMP_INCOME
```

```
                                        IMP_JOB
                                        KIDSDRIV
                                        MSTATUS
                                        MVR_PTS
                                        OLDCLAIM
                                        PARENT1
                                        RED_CAR
                                        REVOKED
                                        SEX
                                        TIF
                                        R_TRAVTIME
                                        URBANICITY
                                        IMP_YOJ
                                        log_OLDCLAIM
                                        log_MVR_PTS
                                        log_TIF
                                        log_KIDSDRIV
                                        log_CLM_FREQ
                                        log_HOMEKIDS
                                        No_CLM_FREQ
                                        No_HOMEKIDS
                                        No_KIDSDRIV
                                        No_MVR_PTS
                                        /selection=Forward
                                        ;
/*score data= scrubfile*/output out=SCORED p=p_target;
run;


/****Model 3 PROC LOGISTIC STEPWISE ****/
proc logistic data=scrubfile plot(only)=(roc(ID=prob));
class _character_ /param=ref;
model TARGET_FLAG( ref="0" ) =
                                        IMP_AGE
                                        R_BLUEBOOK
                                        IMP_CAR_AGE
                                        CAR_TYPE
                                        CAR_USE
                                        CLM_FREQ
                                        EDUCATION
                                        HOMEKIDS
                                        HIGH_OLDCLAIM
                                        IMP_HOME_VAL
                                        IMP_INCOME
                                        IMP_JOB
                                        KIDSDRIV
                                        MSTATUS
                                        MVR_PTS
                                        OLDCLAIM
                                        PARENT1
                                        RED_CAR
                                        REVOKED
                                        SEX
                                        TIF
                                        R_TRAVTIME
                                        URBANICITY
```

```
                                        IMP_YOJ
                                        log_OLDCLAIM
                                        log_MVR_PTS
                                        log_TIF
                                        log_KIDSDRIV
                                        log_CLM_FREQ
                                        log_HOMEKIDS
                                        No_CLM_FREQ
                                        No_HOMEKIDS
                                        No_KIDSDRIV
                                        No_MVR_PTS/selection=stepwise
                                        ;
/*score data= scrubfile*/output out=SCORED p=p_target;
run;


/*****Model 4******/
proc logistic data=SCRUBFILE plot(only)=(roc(ID=prob));
class _character_ /param=ref;
model TARGET_FLAG( ref="0" ) =
                                        IMP_AGE
                                        R_BLUEBOOK
                                        IMP_CAR_AGE
                                        CAR_TYPE
                                        CAR_USE
                                        CLM_FREQ
                                        EDUCATION
                                        HOMEKIDS
                                        HIGH_OLDCLAIM
                                        IMP_HOME_VAL
                                        IMP_INCOME
                                        IMP_JOB
                                        KIDSDRIV
                                        MSTATUS
                                        MVR_PTS
                                        OLDCLAIM
                                        PARENT1
                                        RED_CAR
                                        REVOKED
                                        SEX
                                        TIF
                                        R_TRAVTIME
                                        URBANICITY
                                        IMP_YOJ
                                        log_OLDCLAIM
                                        log_MVR_PTS
                                        log_TIF
                                        log_KIDSDRIV
                                        log_CLM_FREQ
                                        log_HOMEKIDS
                                        No_CLM_FREQ
                                        No_HOMEKIDS
                                        No_KIDSDRIV
                                        No_MVR_PTS
                                    /link=probit roceps=0.1
                                        ;
```

```
/*score data=scrubfile*/output out=SCORED p=p_target;
run;


* ROC AND KS STAT *;
proc npar1way data=SCORED edf;
   class target_flag;
   var p_target;
run;

/**********************Stand alone scoring code*************************************/
/*****************************************************************************************/
data test;
set &INFILE2.;


data scorefile;
set test;

/********Variable Adjustments*****/
/********Income******/

IMP_AGE = AGE;
if missing( IMP_AGE ) then IMP_AGE = 44.8;

IMP_INCOME    = INCOME;
M_INCOME      = 0;
if missing(IMP_INCOME) then do;
IMP_INCOME = 61898;
M_INCOME = 1;
end;
if IMP_INCOME > 215536
then IMP_INCOME = 215536;

IMP_HOME_VAL = HOME_VAL;
M_HOME_VAL = 0;
if missing(IMP_HOME_VAL) then do;
IMP_HOME_VAL = 154867;
M_HOME_VAL = 1;
end;
if IMP_HOME_VAL > 500309
then IMP_HOME_VAL = 500309;

IMP_YOJ = YOJ;
if missing( IMP_YOJ )
then IMP_YOJ = 11;

IMP_CAR_AGE = CAR_AGE;
M_CAR_AGE    = 0;
if missing(IMP_CAR_AGE) then do;
IMP_CAR_AGE = 8.33;
M_CAR_AGE = 1;
end;
if IMP_CAR_AGE < 0
then IMP_CAR_AGE = 1;
```

```
R_TRAVTIME = TRAVTIME;
M_TRAVTIME   = 0;
if R_TRAVTIME > 75.14 then do;
R_TRAVTIME = 75.14;
M_TRAVTIME = 1;
end;

R_BLUEBOOK = BLUEBOOK;
M_BLUEBOOK  = 0;
if R_BLUEBOOK>39090 then do;
R_BLUEBOOK = 39090;
M_BLUEBOOK = 1;
end;


IMP_JOB = JOB;
if missing(IMP_JOB) then do;
if IMP_INCOME > 125000 then
IMP_JOB = "Doctor";
else if IMP_INCOME > 80000 then
IMP_JOB = "Lawyer";
            else
IMP_JOB = "z_Blue Collar";
end;

JOB_WHITE_COLLAR = IMP_JOB in ("Doctor","Lawyer");

drop AGE;
drop INCOME;
drop HOME_VAL;
drop YOJ;
drop CAR_AGE;
drop BLUEBOOK;
drop TRAVTIME;

log_OLDCLAIM = log(OLDCLAIM+1);
HIGH_OLDCLAIM = 0;
if OLDCLAIM > 6000 then HIGH_OLDCLAIM = 1;

log_TIF = log(TIF+1);

log_MVR_PTS = log(MVR_PTS+1);
No_MVR_PTS = 0;
if MVR_PTS = 0 then No_MVR_PTS = 1;

log_KIDSDRIV = log(KIDSDRIV+1);
No_KIDSDRIV = 0;
if KIDSDRIV = 0 then No_KIDSDRIV = 1;

log_CLM_FREQ = log(CLM_FREQ+1);
No_CLM_FREQ = 0;
if CLM_FREQ = 0 then No_CLM_FREQ = 1;

log_CLM_FREQ = log(CLM_FREQ+1);
No_CLM_FREQ = 0;
if CLM_FREQ = 0 then No_CLM_FREQ = 1;
```

```
log_HOMEKIDS = log(HOMEKIDS+1);
No_HOMEKIDS = 0;
if HOMEKIDS = 0 then No_HOMEKIDS = 1;


TEMP = -.492
-0.00002*R_BLUEBOOK
-0.7124*(CAR_TYPE in ("Minivan"))
-0.075*(CAR_TYPE in ("Panel Truck"))
-0.134*(CAR_TYPE in ("Pickup"))
+0.279*(CAR_TYPE in ("Sports Car"))
-0.118*(CAR_TYPE in ("Van"))
+0.775*(CAR_USE in ("Commercial"))
-0.022*(EDUCATION in ("<High School"))
-0.383*(EDUCATION in ("Bachelors"))
-0.333*(EDUCATION in ("Masters"))
+0.040*(EDUCATION in ("PhD"))
-1.48E-6*IMP_HOME_VAL
-3.29E-6*IMP_INCOME
+0.063*(IMP_JOB in ("Clerical"))
-0.884*(IMP_JOB in ("Doctor"))
-0.104*(IMP_JOB in ("Home Maker"))
-0.162*(IMP_JOB in ("Lawyer"))
-0.852*(IMP_JOB in ("Manager"))
-0.132*(IMP_JOB in ("Professional"))
-0.111*(IMP_JOB in ("Student"))
-0.658*(MSTATUS in ("Yes"))
+0.101*MVR_PTS
-0.00001*OLDCLAIM
-0.946*(REVOKED in  ("No"))
+0.0150*R_TRAVTIME
+2.328*(URBANICITY in ("Highly Urban/ Urban"))
-0.325*Log_TIF
+0.599*Log_KIDSDRIV
-0.615*No_CLM_FREQ
-0.315* No_HOMEKIDS;
TEMP = exp( TEMP );
TEMP = TEMP / (1.0+TEMP);
P_TARGET_FLAG = TEMP;
drop TEMP;

P_TARGET_AMT = 1500+0.1*R_BLUEBOOK       - 41.75664 * IMP_CAR_AGE + 464.81029 * CLM_FREQ;


keep index P_TARGET_FLAG P_TARGET_AMT;

run;

/*******Please print******/
proc print data=scorefile;
run;

/*******Save to Folder********/
libname NOELIB "/home/noeflores20160/sasuser.v94";
```

```
data NOELIB.NOEFILE;
set scorefile;
run;

proc print data=NOELIB.NOEFILE;
run;

proc print data=scorefile;
var P_TARGET_FLAG P_TARGET_AMT;
run;



/******BINGO BONUS*********************/
proc genmod data=SCRUBFILE;
class _character_ /param=ref;
model TARGET_FLAG( ref="0" ) =
                                        IMP_AGE
                                        R_BLUEBOOK
                                        IMP_CAR_AGE
                                        CAR_TYPE
                                        CAR_USE
                                        CLM_FREQ
                                        EDUCATION
                                        HOMEKIDS
                                        HIGH_OLDCLAIM
                                        IMP_HOME_VAL
                                        IMP_INCOME
                                        IMP_JOB
                                        KIDSDRIV
                                        MSTATUS
                                        MVR_PTS
                                        OLDCLAIM
                                        PARENT1
                                        RED_CAR
                                        REVOKED
                                        SEX
                                        TIF
                                        R_TRAVTIME
                                        URBANICITY
                                        IMP_YOJ
                                        log_OLDCLAIM
                                        log_MVR_PTS
                                        log_TIF
                                        log_KIDSDRIV
                                        log_CLM_FREQ
                                        log_HOMEKIDS
                                        No_CLM_FREQ
                                        No_HOMEKIDS
                                        No_KIDSDRIV
                                        No_MVR_PTS
                                        ;
run;
```

References:

(1) Hoffmann, J. P., (2004). *Generalized Linear Models: An Applied Approach*. Boston, MA: Pearson Education, Inc.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). *Introduction to Linear Regression Analysis* Hoboken, NJ: Wiley.

(3) Cody, R. (2011). *SAS: Statistics by Example*. Carey, NC: SAS Institute Inc.

(4) Evaluating forecast Accuracy. https://www.otexts.org/fpp/2/5 (accessed May 9, 2017)