Kaggle Name: Christopher Jose

Introduction

The objective of this report is to determine the likelihood of an individual crashing their vehicle, given other characteristics of that individual. These individuals are insureds at a particular auto insurance company. Each record in the dataset represents information about a particular insured, such as their age, income, total claims in the past 5 years, etc.

Within this report, I perform the following functions with the above goal in mind:
1. Exploratory Data Analysis (EDA)
2. Data Preparation
3. Model Building
4. Model Selection

I do EDA to know how to best prepare the data in the Data Preparation step. In the Model Building step, I go over four models that I have built off data prepared in the Data Preparation step. I then compare the four models in the Model Selection step to select the optimal model. The optimal model can then be used to predict the likelihood of an individual crashing their vehicle, based on the characteristics of that individual.

Three out of the four models are logistic regression models, which predict binary response variables and can spit out the probability that a response equals 1 or 0. Linear regression should not be used for binary response variables. I also incorporate a Decision Tree as a benchmark to get a sense of how well my logistic models are doing.

EXPLORATORY DATA ANALYSIS

I examine the data, in particular, the predictors, to answer the following questions:

1. Which predictors will likely be predictive of crashing a car?
2. Are there missing values that need to be imputed?
3. Are there outliers that need to be capped?
4. Are there variables that might benefit from being redefined?

The dataset contains 8161 observations and 26 variables. 24 variables will be considered in predicting the dependent variable, TARGET_FLAG, a binary variable where a value of 1 indicates a crash and a value of 0 does not. TARGET_AMT will not be considered as a predictor, since an amount above zero guarantees a crash and thus TARGT_FLAG=1. 14 of the 24 variables are numeric. 10 of the 24 variables are categorical.

The data is as follows:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
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

1. Which predictors will likely be predictive of crashing a car?

For a given numeric predictor, I examine whether there is a sizable difference between its mean for insureds who have crashed their car versus those who have not crashed their car. About half of the numeric variables appear to have a noticeable difference and suggest greater predictability of TARGET_FLAG.

Numeric predictors that seem predictive:
kidsdriv, income, home_val, oldclaim, clm_freq, mvr_pts
Numeric predictors that do not seem predictive:
age, homekids, yoj, travtime, bluebook, TIF, car_age

For example, the mean income for insureds who did not crash is $66,000, while the mean income for insureds who did crash is $50,000. Since this difference seems sizable, income appears to be predictive of crashing the car.

TARGET_FLAG	N Obs	Variable	Label	Mean
0	6008	INDEX		5154.84
		TARGET_FLAG		0
		TARGET_AMT		0
		KIDSDRIV	#Driving Children	0.1393142
		AGE	Age	45.3227901
		HOMEKIDS	#Children @Home	0.6439747
		YOJ	Years on Job	10.6718337
		INCOME	Income	65951.97
		HOME_VAL	Home Value	169075.41
		TRAVTIME	Distance to Work	33.0303446
		BLUEBOOK	Value of Vehicle	16230.95
		TIF	Time in Force	5.5557590
		OLDCLAIM	Total Claims(Past 5 Years)	3311.59
		CLM_FREQ	#Claims(Past 5 Years)	0.6486352
		MVR_PTS	Motor Vehicle Record Points	1.4137816
		CAR_AGE	Vehicle Age	8.6709220
1	2153	INDEX		5143.56
		TARGET_FLAG		1.0000000
		TARGET_AMT		5702.18
		KIDSDRIV	#Driving Children	0.2596377
		AGE	Age	43.3012104
		HOMEKIDS	#Children @Home	0.9368323
		YOJ	Years on Job	10.0167488
		INCOME	Income	50641.30
		HOME_VAL	Home Value	115256.55
		TRAVTIME	Distance to Work	34.7681203
		BLUEBOOK	Value of Vehicle	14255.90
		TIF	Time in Force	4.7807710
		OLDCLAIM	Total Claims(Past 5 Years)	6061.55
		CLM_FREQ	#Claims(Past 5 Years)	1.2169066
		MVR_PTS	Motor Vehicle Record Points	2.4816535
		CAR_AGE	Vehicle Age	7.3674789

For a given categorical variable, I inspect whether there is a sizable difference between each level's proportion of observations that had a car crash. If each level has around the same proportion of observations that had car crashes, then this suggests the variable will be a poor predictor. If at least one level's proportion of observations with car crashes differs sizably from other levels, then this suggests the variable is predictive of a car crash.

Below, I examine the proportion of married parents who had car crashes versus the proportion of single parents who had car crashes. 44.20% of insureds who are single parents had car crashes, while only 23.67% of insureds who are married parents had car crashes. This is a sizable difference. A married parent seems less likely to have a car crash.

Examining the **Row Pct** row of levels *No* and *Yes*, for TARGET_FLAG=1, observe that 23.67% of married parents had car crashes, while 44.20% of single parents had car crashes.

Frequency
Percent
Row Pct
Col Pct

Table of PARENT1 by TARGET_FLAG

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

Repeating the above process,
Categorical predictors that seem predictive:
parent1, mstatus, education, job, car_use, car_type, revoked, urbanicity.
Categorical predictors that do not seem predictive: sex, red_car.

2. Are there missing values that need to be imputed?

Missing values are a problem for logistic regression. Six numeric variables have missing values: age, yoj, income, home_val, car_age. Only one categorical variable has missing values, job, and has 526 missing values.

Numeric variables with missing values are shown below:

Variable	Label	N Miss
INDEX		0
TARGET_FLAG		0
TARGET_AMT		0
KIDSDRIV	#Driving Children	0
AGE	Age	6
HOMEKIDS	#Children @Home	0
YOJ	Years on Job	454
INCOME	Income	445
HOME_VAL	Home Value	464
TRAVTIME	Distance to Work	0
BLUEBOOK	Value of Vehicle	0
TIF	Time in Force	0
OLDCLAIM	Total Claims(Past 5 Years)	0
CLM_FREQ	#Claims(Past 5 Years)	0
MVR_PTS	Motor Vehicle Record Points	0
CAR_AGE	Vehicle Age	510

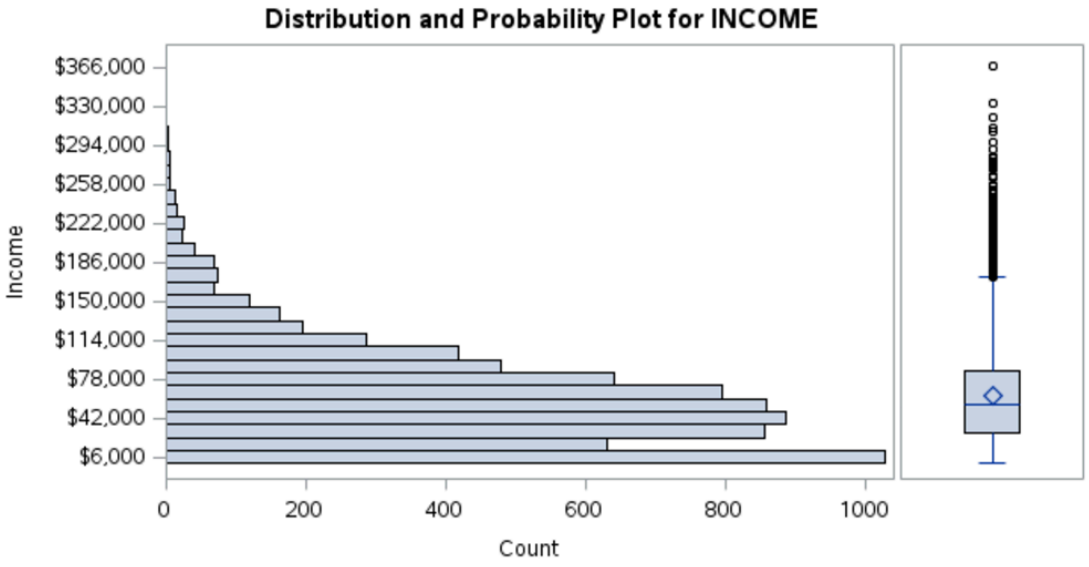
Upon inspection of the categorical variable *job* by education level, we observe that only Masters and PhD degree holders have missing *job* values. This category (of missing values) seems likely to represent white-collar jobs with higher income levels.

JOB(Job Category)	EDUCATION(Max Education Level)					
	<High School	Bachelors	Masters	PhD	z_High School	Total
	0	0	328	198	0	526
	0.00	0.00	4.02	2.43	0.00	6.45
	0.00	0.00	62.36	37.64	0.00	
	0.00	0.00	19.78	27.20	0.00	

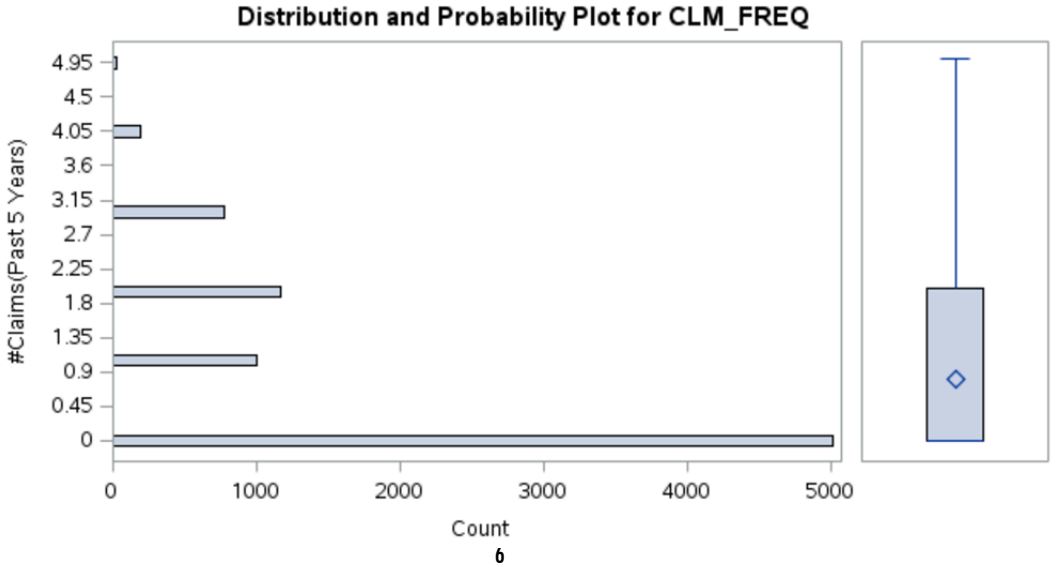
3. Are there outliers that need to be capped?

Numeric predictors that have skewed distributions may have influential outliers that can affect regression results by making the model less predictive. Most variables in this dataset are right skewed. Also, some outliers can result from incorrectly recorded information. In the case of this dataset, one observation has a negative value for car_age, which is clearly incorrect.

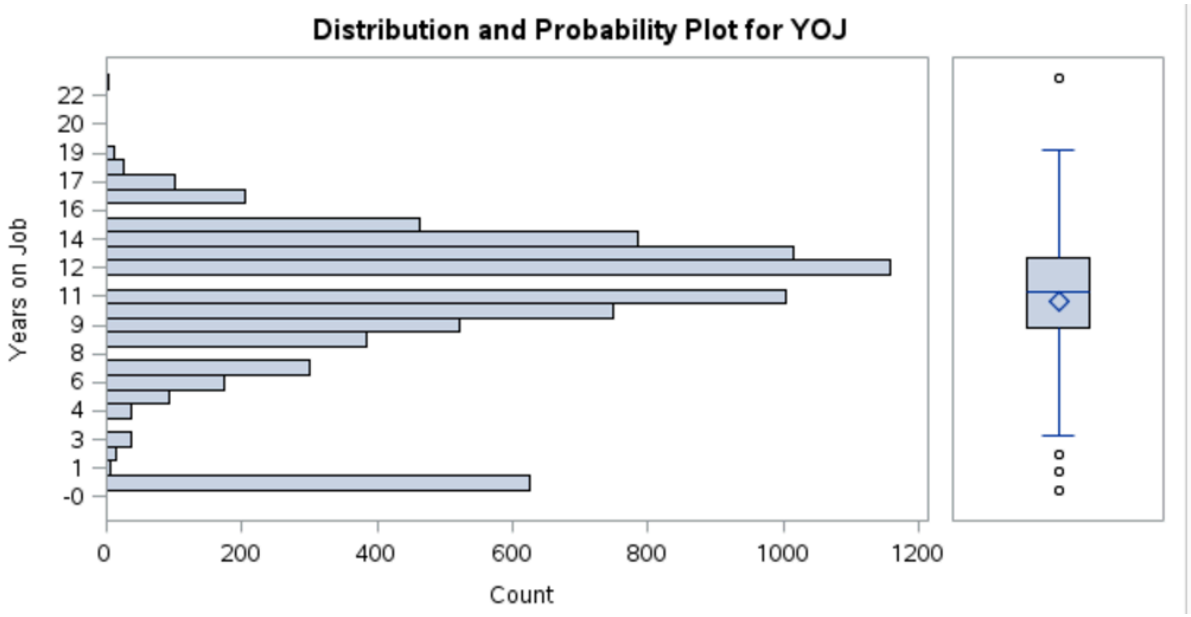
Predictors that have right skewed distributions and with large ranges include:
income, home_val, travtime, bluebook, oldclaim. Constraining these predictors could benefit the model.



Predictors with right skewed distributions, but with small ranges (range=max-min), include: homekids, tif, clm_freq, mvr_pts, car_age. These predictors have such small ranges that reducing the range of their data would result in too much information loss, in my opinion.



Age and yoj are the only predictors that appear normally distributed, though yoj has a high incidence of 0 values.



Variable car_age, which represents a vehicle's age, has a value of -3, which does not make sense.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3	6941	25	5145

DATA PREPARATION

After exploring the data, I do the following:
1. Impute missing values of each numeric variable with its median.
2. Trim variables with large ranges that have right-skewed distributions to reduce the impact of outliers on the model.
3. Log transform certain variables
4. Consider redefining the age variable, which I found to be not very predictive.

1. Impute missing values with the median

I "fill in" missing values with each variable's median (instead of the mean) to reduce the effect of each variable's outliers on the imputed values. Recall that the following variables have missing values: age, yoj, income, home_val, car_age. The mean of income is about $8000 more than the median due to income's outliers.

Flag variables are made for every variable and are labeled with the original variable's name, but with an M in front, where a 1 for an observation indicates that it had a missing value which was median imputed for the variable associated with the flag variable. As shown below, all variables have zero missing values (where *N Miss* indicates the number of missing values). Imputed values are shown in the red box.

Variable	N Miss
INDEX	0
M_KIDSDRIV	0
kidsdriv	0
M_AGE	0
age	0
M_HOMEKIDS	0
homekids	0
M_YOJ	0
yoj	0
M_INCOME	0
income	0
M_HOME_VAL	0
home_val	0
M_TRAVTIME	0
travtime	0
M_BLUEBOOK	0
bluebook	0
M_TIF	0
tif	0
M_OLDCLAIM	0
oldclaim	0
M_CLM_FREQ	0
clm_freq	0
M_MVR_PTS	0
mvr_pts	0
M_CAR_AGE	0
car_age	0

Imputation Results

Variable	Imputation Indicator	Imputed Variable	N Missing	Type of Imputation	Imputation Value (Seed)
KIDSDRIV	M_KIDSDRIV	IM_KIDSDRIV	0	Pseudo Median	0
AGE	M_AGE	IM_AGE	6	Pseudo Median	45.00000
HOMEKIDS	M_HOMEKIDS	IM_HOMEKIDS	0	Pseudo Median	0
YOJ	M_YOJ	IM_YOJ	454	Pseudo Median	11.00000
INCOME	M_INCOME	IM_INCOME	445	Pseudo Median	54028
HOME_VAL	M_HOME_VAL	IM_HOME_VAL	464	Pseudo Median	161160
TRAVTIME	M_TRAVTIME	IM_TRAVTIME	0	Pseudo Median	32.87097
BLUEBOOK	M_BLUEBOOK	IM_BLUEBOOK	0	Pseudo Median	14440
TIF	M_TIF	IM_TIF	0	Pseudo Median	4.00000
OLDCLAIM	M_OLDCLAIM	IM_OLDCLAIM	0	Pseudo Median	0
CLM_FREQ	M_CLM_FREQ	IM_CLM_FREQ	0	Pseudo Median	0
MVR_PTS	M_MVR_PTS	IM_MVR_PTS	0	Pseudo Median	1.00000
CAR_AGE	M_CAR_AGE	IM_CAR_AGE	510	Pseudo Median	8.00000

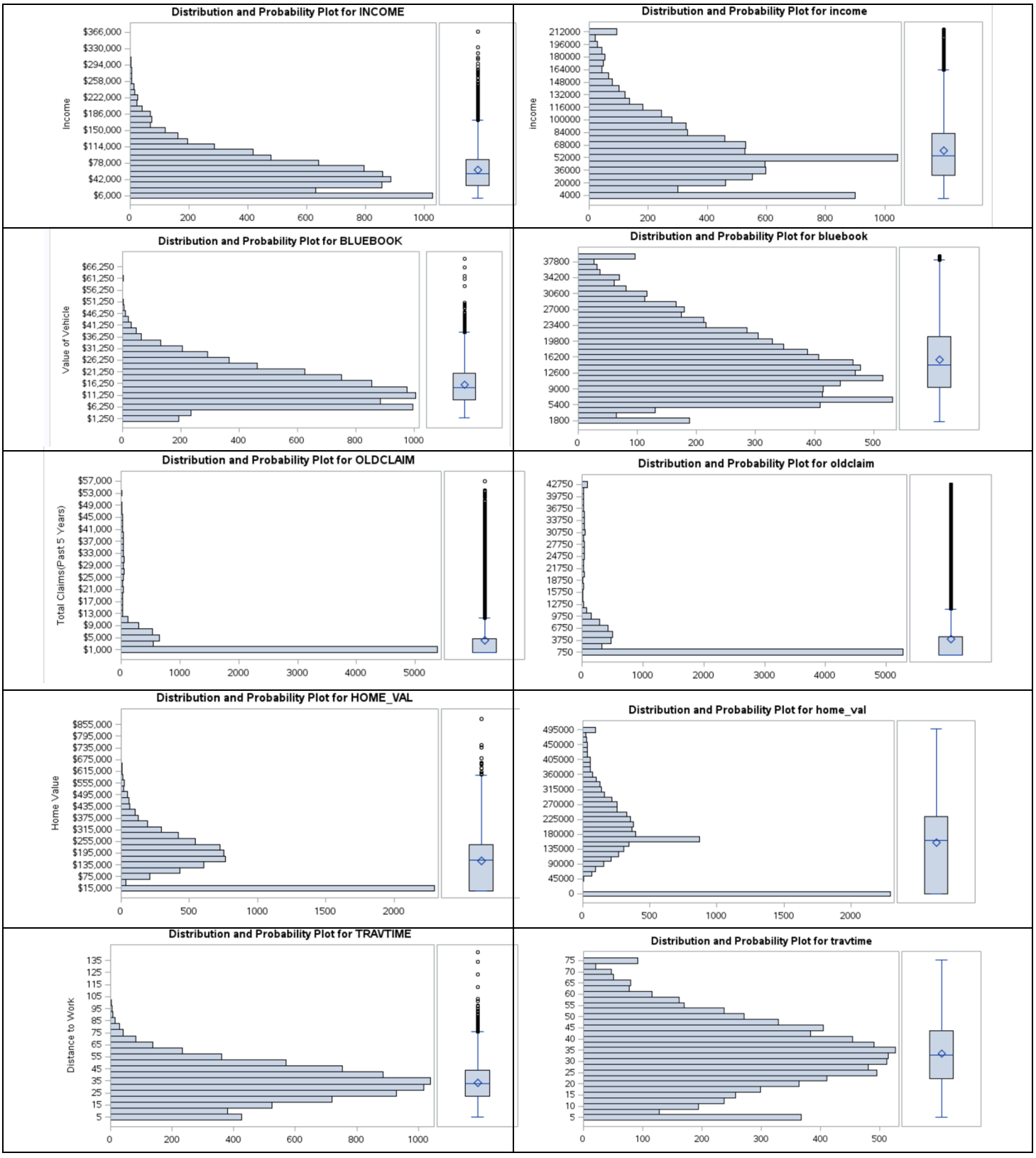
2. Trimming variables with large ranges and skewed distributions

For the variables I mentioned in the EDA section that have large ranges and right-skewed distributions (income, home_val, travtime, bluebook, oldclaim), I've trimmed their data to stay within the range of the 1st and 99th percentiles (P1 and P99) of their original distributions.

Before and after displays of their distributions shown on the next page. The trimmed distributions for the below variables have smaller ranges, as expected. Income, for instance, is now capped at $215,000.

Before Trimming to P1 and P99

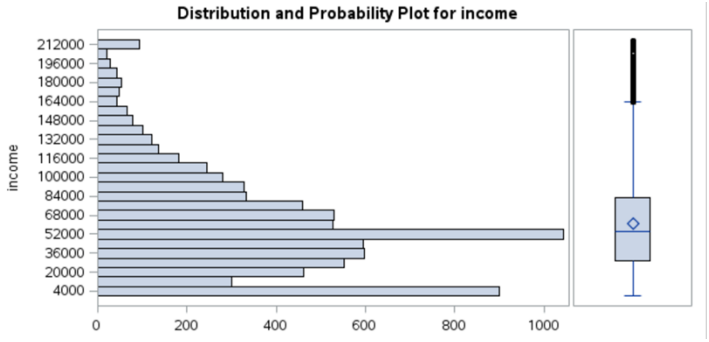
After Trimming to P1 and P99

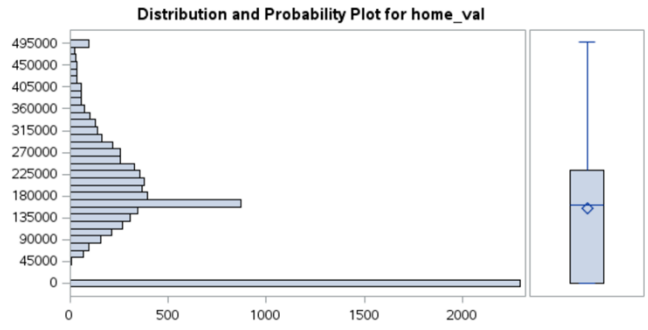


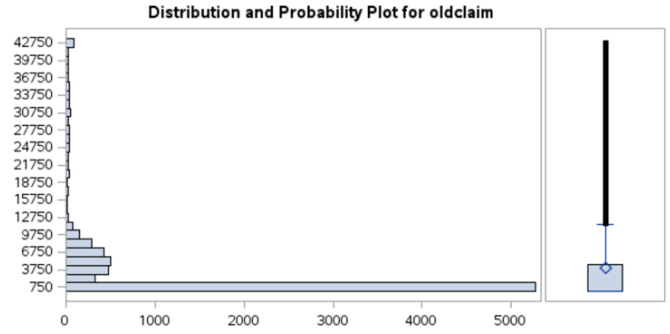
3. Log transform certain variables

Income, home_val, and oldclaim are still right skewed and have a wide range even after trimming their distributions. I will apply a natural log transformation to the variables, which further reduces their ranges as well as the effect of outliers. Missing values after the transformation are set equal to zero (since a value of zero caused them to be missing in the first place).

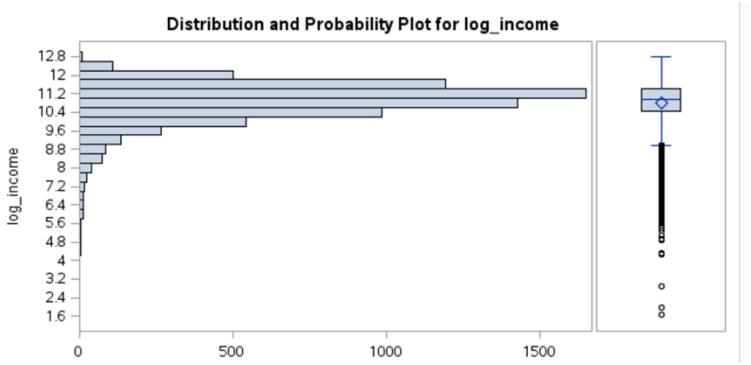
Before log transformation

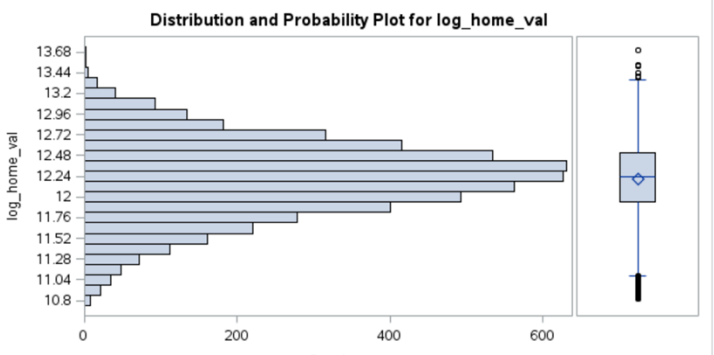


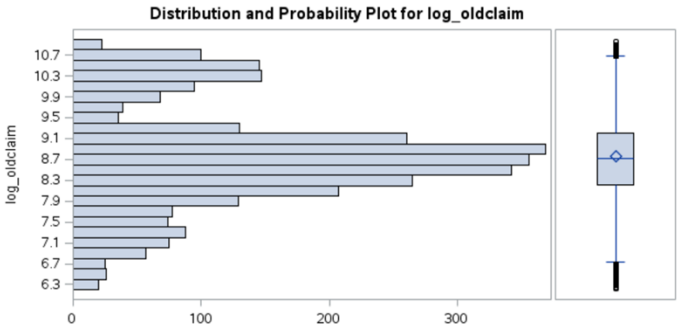




After log transformation







4. Consider redefining the age variable

The age variable did not appear that predictive of TARGET_FLAG, which is surprising because younger people are regarded as riskier in the industry. I consider making two dummy variables, such that

age1 = 1 if age<=20
age2 = 1 if age>52
where age1=age2=0 indicates 21<=age<=51

The proportion of insureds under 21 who got into a crash is 23.53%, while the proportion of insureds over 21 who got into a crash is 26.39%. The proportion of insureds over 52 who got into a crash is 28.19%, while the proportion of insureds under 52 who got into a crash is 25.96%. There is not a noticeable difference between each of these pairs of proportions, so I will leave the age variable as is.

Table of age2 by TARGET_FLAG

age2	TARGET_FLAG 0	1	Total
0	4905 60.10 74.04 81.64	1720 21.08 25.96 ~~78.89~~	6625 81.18
1	1103 13.52 71.81 18.36	433 ~~5.31~~ 28.19 20.11	1536 18.82
Total	6008 73.62	2153 26.38	8161 100.00

Table of age1 by TARGET_FLAG

age1	TARGET_FLAG 0	1	Total
0	5995 73.46 73.61 99.78	2149 26.33 26.39 ~~99.81~~	8144 99.79
1	13 0.16 76.47 0.22	4 ~~0.05~~ 23.53 0.19	17 0.21
Total	6008 73.62	2153 26.38	8161 100.00

MODEL BUILDING

Model 1– Logistic Regression

This method runs a logistic regression model off all predictors, including flag variables created from median imputation, several trimmed numeric variables, as well as 3 log transformed variables (where the original variables are excluded).

Model 2– Logistic Regression with Insignificant Variables Removed

This method runs a logistic regression model off the predictors in model 1, except those that were insignificant are not included. Thus, it keeps the following variables: kidsdriv, tif, bluebook, car_type, education, job, mstatus, parent1, revoked, sex, and urbanicity.

Model 3– Logistic Regression with Stepwise Variable Selection

This method iteratively adds significant predictors to the model and removes insignificant predictors from the model. I use all predictors that were in model 1.

Model 4– Decision Tree

This method uses all predictors from model 1 and builds a classification tree. I incorporate this model as a benchmark to get a sense of how well the logistic models are doing.

All models are built off a 70% train split to increase generalizability.

Model 1 – Logistic Regression

I run a logistic regression model off all predictors, including flag variables created from median imputation, several trimmed numeric variables, as well as 3 log transformed variables (where the original variables are excluded).

The Wald Test, Likelihood Ratio Test (LRT), and Score Test, are asymptotically equivalent and test the global null hypothesis that all coefficients are zero. If they are in alignment, then this is evidence that the model offers greater predictability than the intercept-only (null) model. All tests reject the null at the .0001 level, so we are certain this model does better than the null model.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1234.5287	42	<.0001
Score	1097.9136	42	<.0001
Wald	862.5395	42	<.0001

The following statistics, Akaike's Information Criterion (AIC), Schwarz Bayesian Information Criterion (SC), and Deviance (-2 Log L), are useful for comparing this model to other models. The smaller these values, the better the model fits the data. AIC is the deviance adjusted with a penalty for model complexity (more predictors). SC is the deviance adjusted with a greater penalty than that of AIC for model complexity. Deviance is based on the model's likelihood value in that the smaller the deviance, the greater the model's likelihood.

As we can see below, the model comparison statistics show that the hypothesized model is better than the intercept-only model:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6600.702	5450.174
SC	6607.353	5736.145
-2 Log L	6598.702	5364.174

Estimated coefficients from the model are as follows:

Coefficients with low p-values are accentuated with a red box.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7780	1.5415	1.3304	0.2487
bluebook		1	-6.13E-6	4.363E-6	1.9746	0.1600
travtime		1	0.00230	0.00220	1.0948	0.2954
log_income		1	-0.00006	0.0170	0.0000	0.9974
log_home_val		1	-0.00903	0.00647	1.9511	0.1625
kidsdriv		1	0.1881	0.0737	6.5174	0.0107
age		1	0.00216	0.00457	0.2225	0.6372
homekids		1	-0.0164	0.0405	0.1636	0.6859
yoj		1	0.0157	0.0124	1.6137	0.2040
log_oldclaim		1	0.0117	0.0160	0.5361	0.4640
tif		1	-0.0244	0.00834	8.5477	0.0035
clm_freq		1	-0.0607	0.0576	1.1134	0.2914
mvr_pts		1	0.0197	0.0175	1.2591	0.2618
car_age		1	0.00175	0.00645	0.0737	0.7861
CAR_TYPE	Minivan	1	-1.0783	0.1195	81.4607	<.0001
CAR_TYPE	Panel Truck	1	-0.8973	0.1830	24.0343	<.0001
CAR_TYPE	Pickup	1	-0.4400	0.1286	11.7039	0.0006
CAR_TYPE	Sports Car	1	0.2380	0.1143	4.3367	0.0373
CAR_TYPE	Van	1	-0.6083	0.1612	14.2345	0.0002
CAR_USE	Commercial	1	0.8931	0.1065	70.3431	<.0001
EDUCATION	<High School	1	0.0886	0.1098	0.6501	0.4201
EDUCATION	Bachelors	1	-0.5073	0.0967	27.5007	<.0001
EDUCATION	Masters	1	-0.4652	0.1660	7.8574	0.0051
EDUCATION	PhD	1	-0.5718	0.2079	7.5634	0.0060
JOB	Clerical	1	0.2291	0.1229	3.4760	0.0623
JOB	Doctor	1	-0.6595	0.3310	3.9691	0.0463
JOB	Home Maker	1	0.2936	0.1603	3.3542	0.0670
JOB	Lawyer	1	-0.2731	0.2175	1.5766	0.2093
JOB	Manager	1	-0.9770	0.1631	35.8899	<.0001
JOB	NA	1	-0.4163	0.2131	3.8179	0.0507
JOB	Professional	1	-0.1347	0.1381	0.9524	0.3291
JOB	Student	1	0.2670	0.1327	4.0454	0.0443
MSTATUS	Yes	1	-0.4906	0.0788	38.7855	<.0001
PARENT1	No	1	-0.5887	0.1078	29.8454	<.0001
RED_CAR	no	1	-0.0313	0.0999	0.0984	0.7538
REVOKED	No	1	-0.7621	0.0923	68.1362	<.0001
SEX	M	1	0.2815	0.1190	5.5939	0.0180
URBANICITY	Highly Urban/ Urban	1	2.5387	0.1328	365.3638	<.0001
M_AGE	0	1	0.2344	1.4771	0.0252	0.8739
M_YOJ	0	1	0.1310	0.1537	0.7268	0.3939
M_INCOME	0	1	-0.1534	0.1461	1.1016	0.2939
M_HOME_VAL	0	1	-0.0116	0.1506	0.0059	0.9387
M_CAR_AGE	0	1	0.00617	0.1388	0.0020	0.9645

Model results are to be interpreted as follows:

Log-odds of getting into a car accident = Sum(Estimated Coefficient j x Predictor j) across all j
Odds = e^(log-odds)
Estimated Probability = odds / (1+odds)
where the above is calculated for a given observation's values.

Interpreting the Coefficients:

e^(estimated coefficient j) can be interpreted as the multiplicative change in the odds for a one-unit increase in predictor j and is called the odds ratio estimate. These values are shown below.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
bluebook	1.000	1.000	1.000
travtime	1.002	0.998	1.007
log_income	1.000	0.967	1.034
log_home_val	0.991	0.979	1.004
kidsdriv	1.207	1.045	1.394
age	1.002	0.993	1.011
homekids	0.984	0.909	1.065
yoj	1.016	0.992	1.041
log_oldclaim	1.012	0.981	1.044
tif	0.976	0.960	0.992
clm_freq	0.941	0.841	1.053
mvr_pts	1.020	0.985	1.055
car_age	1.002	0.989	1.014
CAR_TYPE Minivan vs z_SUV	0.340	0.269	0.430
CAR_TYPE Panel Truck vs z_SUV	0.408	0.285	0.584
CAR_TYPE Pickup vs z_SUV	0.644	0.501	0.829
CAR_TYPE Sports Car vs z_SUV	1.269	1.014	1.587
CAR_TYPE Van vs z_SUV	0.544	0.397	0.747
CAR_USE Commercial vs Private	2.443	1.983	3.010
EDUCATION <High School vs z_High School	1.093	0.881	1.355
EDUCATION Bachelors vs z_High School	0.602	0.498	0.728
EDUCATION Masters vs z_High School	0.628	0.454	0.869
EDUCATION PhD vs z_High School	0.564	0.376	0.848
JOB Clerical vs z_Blue Collar	1.257	0.988	1.600
JOB Doctor vs z_Blue Collar	0.517	0.270	0.989
JOB Home Maker vs z_Blue Collar	1.341	0.980	1.836
JOB Lawyer vs z_Blue Collar	0.761	0.497	1.166
JOB Manager vs z_Blue Collar	0.376	0.273	0.518
JOB NA vs z_Blue Collar	0.659	0.434	1.001
JOB Professional vs z_Blue Collar	0.874	0.667	1.146
JOB Student vs z_Blue Collar	1.306	1.007	1.694
MSTATUS Yes vs z_No	0.612	0.525	0.714
PARENT1 No vs Yes	0.555	0.449	0.686
RED_CAR no vs yes	0.969	0.797	1.179
REVOKED No vs Yes	0.467	0.389	0.559
SEX M vs z_F	1.325	1.049	1.673
URBANICITY Highly Urban/ Urban vs z_Highly Rural/ Rural	12.663	9.761	16.429
M_AGE 0 vs 1	1.264	0.070	22.861
M_YOJ 0 vs 1	1.140	0.843	1.541
M_INCOME 0 vs 1	0.858	0.644	1.142
M_HOME_VAL 0 vs 1	0.988	0.736	1.328
M_CAR_AGE 0 vs 1	1.006	0.767	1.321

I will evaluate the coefficients on predictors with low p-values (are statistically significant). The odds ratio estimate for the statistically significant predictor, kidsdriv, is 1.207, which means when number of children driving the car increases by one, the odds of an accident for the insured increases by 20.7%.

All coefficients on significant predictors appear to make sense. Positive coefficients indicate the exponentiated value (estimated odds ratio) will be over 1, and negative values indicate the estimated odds ratio will be under 1. The coefficient on tif is negative, which makes sense, since we'd expect that insureds that stay longer with the insurer are less risky. The coefficients on the car type variables also make sense. Bigger cars, like minivans and panel trucks, have negative coefficients, while the sports car has a positive coefficient. We'd expect the sports car drivers to be riskier. Bachelors degree holders have a negative coefficient, meaning that going from a high school education level to a bachelor's level offers a predicted decrease in the odds of getting into an accident (since high school education is the base dummy variable level). The jobs dummy variable coefficients make sense. We see negative coefficients on white collar jobs, like lawyer, and positive coefficients on jobs like clerical and student.

Some coefficients on insignificant predictors do not appear to make sense. Log income and log home value are not statistically significant, which is surprising. The coefficient on log income does not make sense, since it is positive and we'd expect wealthier people would get into less accidents. The coefficient on yoj is also counterintuitive. It is positive (though small), meaning the odds ratio is over 1 (=1.016), and more years on the job means a higher estimated chance of an accident, when we would think it would mean less chance of an accident.

Additional Goodness of Fit Metrics

The percentage of concordant pairs is 79.1%, which is high. This indicates that observations with actual TARGET_FLAG values of 1 will, in most cases, have higher predicted probabilities than observations with actual TARGET_FLAG values of 0.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.1	Somers' D	0.582
Percent Discordant	20.9	Gamma	0.582
Percent Tied	0.0	Tau-a	0.226
Pairs	6346530	c	0.791

Model 2 - Logistic Regression with Insignificant Variables Removed

I remove insignificant variables from Model 1 to make Model 2, which consists of predictors: kidsdriv, tif, bluebook, car_type, education, job, mstatus, parent1, revoked, sex, and urbanicity. I've left in bluebook, whose p-value was insignificant in Model 1, but not very high. Thus, this model excludes the flag variables, the two log transformed variables, etc. Most variables are statistically significant in this model. Odds Ratios are on the right, indicating the multiplicative change in the odds for a one-unit increase in each variable.

All odds ratios make sense. Longer term customers who have less kids driving, higher education, a better job, are married, are not a single parent, do not drive a sports car, have not had their license revoked recently, are not male, and don't live in an urban area, are expected to have lower odds of getting into an accident.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3130	0.2142	37.5580	<.0001
kidsdriv		1	0.1786	0.0631	8.0004	0.0047
tif		1	-0.0245	0.00831	8.7256	0.0031
bluebook		1	-5.16E-6	4.126E-6	1.5630	0.2112
CAR_TYPE	Minivan	1	-1.0763	0.1192	81.4963	<.0001
CAR_TYPE	Panel Truck	1	-0.8912	0.1822	23.9138	<.0001
CAR_TYPE	Pickup	1	-0.4397	0.1283	11.7457	0.0006
CAR_TYPE	Sports Car	1	0.2460	0.1140	4.6570	0.0309
CAR_TYPE	Van	1	-0.6013	0.1606	14.0098	0.0002
CAR_USE	Commercial	1	0.8852	0.1062	69.4786	<.0001
EDUCATION	<High School	1	0.0776	0.1095	0.5023	0.4785
EDUCATION	Bachelors	1	-0.5106	0.0964	28.0798	<.0001
EDUCATION	Masters	1	-0.4599	0.1652	7.7505	0.0054
EDUCATION	PhD	1	-0.5721	0.2066	7.6649	0.0056
JOB	Clerical	1	0.2298	0.1224	3.5260	0.0604
JOB	Doctor	1	-0.6625	0.3301	4.0275	0.0448
JOB	Home Maker	1	0.2639	0.1590	2.7559	0.0969
JOB	Lawyer	1	-0.2768	0.2170	1.6273	0.2021
JOB	Manager	1	-0.9833	0.1626	36.5869	<.0001
JOB	NA	1	-0.4263	0.2124	4.0285	0.0447
JOB	Professional	1	-0.1363	0.1377	0.9801	0.3222
JOB	Student	1	0.2681	0.1322	4.1121	0.0426
MSTATUS	Yes	1	-0.4969	0.0782	40.3584	<.0001
PARENT1	No	1	-0.5919	0.1072	30.5045	<.0001
REVOKED	No	1	-0.7624	0.0920	68.7150	<.0001
SEX	M	1	0.3005	0.1029	8.5264	0.0035
URBANICITY	Highly Urban/ Urban	1	2.5384	0.1326	366.5268	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
kidsdriv	1.196	1.056	1.353
tif	0.976	0.960	0.992
bluebook	1.000	1.000	1.000
CAR_TYPE Minivan vs z_SUV	0.341	0.270	0.431
CAR_TYPE Panel Truck vs z_SUV	0.410	0.287	0.586
CAR_TYPE Pickup vs z_SUV	0.644	0.501	0.828
CAR_TYPE Sports Car vs z_SUV	1.279	1.023	1.599
CAR_TYPE Van vs z_SUV	0.548	0.400	0.751
CAR_USE Commercial vs Private	2.423	1.968	2.984
EDUCATION <High School vs z_High School	1.081	0.872	1.340
EDUCATION Bachelors vs z_High School	0.600	0.497	0.725
EDUCATION Masters vs z_High School	0.631	0.457	0.873
EDUCATION PhD vs z_High School	0.564	0.376	0.846
JOB Clerical vs z_Blue Collar	1.258	0.990	1.600
JOB Doctor vs z_Blue Collar	0.516	0.270	0.985
JOB Home Maker vs z_Blue Collar	1.302	0.953	1.778
JOB Lawyer vs z_Blue Collar	0.758	0.496	1.160
JOB Manager vs z_Blue Collar	0.374	0.272	0.514
JOB NA vs z_Blue Collar	0.653	0.431	0.990
JOB Professional vs z_Blue Collar	0.873	0.666	1.143
JOB Student vs z_Blue Collar	1.307	1.009	1.694
MSTATUS Yes vs z_No	0.608	0.522	0.709
PARENT1 No vs Yes	0.553	0.448	0.683
REVOKED No vs Yes	0.467	0.390	0.559
SEX M vs z_F	1.351	1.104	1.652
URBANICITY Highly Urban/ Urban vs z_Highly Rural/ Rural	12.659	9.762	16.416

Model 3 - Logistic Regression with Stepwise Variable Selection

This method iteratively adds significant predictors to the model and removes insignificant predictors from the model. I use all predictors that were in model 1. The stepwise model starts with urbanicity, which has the greatest chi-square value out of all predictors, then adds job, then car_type, etc, until the model adds 11 predictors in total. No predictors end up being removed during this process. All predictors have intuitive odds ratio estimates.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	URBANICITY		1	1	311.3263		<.0001	Home/Work Area
2	JOB		8	2	351.3044		<.0001	Job Category
3	CAR_TYPE		5	3	128.9728		<.0001	Type of Car
4	MSTATUS		1	4	109.0411		<.0001	Marital Status
5	REVOKED		1	5	76.2036		<.0001	License Revoked (Past 7 Years)
6	CAR_USE		1	6	61.1170		<.0001	Vehicle Use
7	PARENT1		1	7	30.7175		<.0001	Single Parent
8	EDUCATION		4	8	36.3288		<.0001	Max Education Level
9	tif		1	9	8.5201		0.0035	
10	SEX		1	10	8.4844		0.0036	Gender
11	kidsdriv		1	11	8.2005		0.0042	

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.4046	0.0982	204.6216	<.0001
kidsdriv		1	0.1801	0.0631	8.1478	0.0043
tif		1	-0.0244	0.00830	8.6465	0.0033
CAR_TYPE	Minivan	1	-0.6160	0.0788	61.0672	<.0001
CAR_TYPE	Panel Truck	1	-0.4411	0.1210	13.2947	0.0003
CAR_TYPE	Pickup	1	0.0231	0.0763	0.0919	0.7617
CAR_TYPE	Sports Car	1	0.7126	0.1070	44.3719	<.0001
CAR_TYPE	Van	1	-0.1445	0.1031	1.9670	0.1608
CAR_USE	Commercial	1	0.4412	0.0531	69.1060	<.0001
EDUCATION	<High School	1	0.3671	0.1044	12.3702	0.0004
EDUCATION	Bachelors	1	-0.2169	0.0797	7.4030	0.0065
EDUCATION	Masters	1	-0.1661	0.1082	2.3571	0.1247
EDUCATION	PhD	1	-0.2777	0.1449	3.6731	0.0553
JOB	Clerical	1	0.4247	0.1126	14.2370	0.0002
JOB	Doctor	1	-0.4753	0.2628	3.2713	0.0705
JOB	Home Maker	1	0.4577	0.1265	13.0980	0.0003
JOB	Lawyer	1	-0.0932	0.1483	0.3949	0.5297
JOB	Manager	1	-0.7944	0.1123	50.0389	<.0001
JOB	NA	1	-0.2335	0.1538	2.3046	0.1290
JOB	Professional	1	0.0553	0.1032	0.2875	0.5919
JOB	Student	1	0.4633	0.1310	12.5081	0.0004
MSTATUS	Yes	1	-0.2476	0.0391	40.1295	<.0001
PARENT1	No	1	-0.2962	0.0536	30.5445	<.0001
REVOKED	No	1	-0.3804	0.0460	68.5283	<.0001
SEX	M	1	0.1540	0.0514	8.9930	0.0027
URBANICITY	Highly Urban/ Urban	1	1.2681	0.0662	366.4374	<.0001

Odds Ratio Estimates

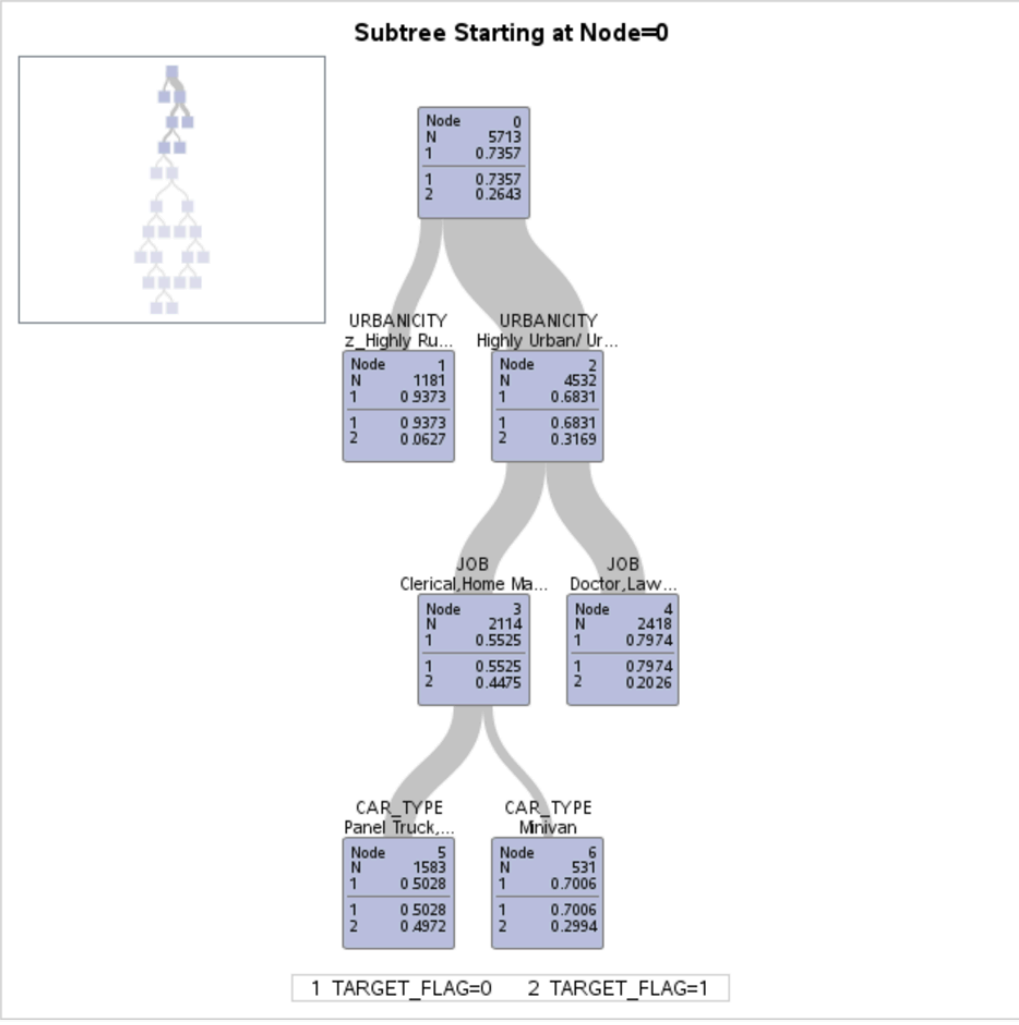
Effect	Point Estimate	95% Wald Confidence Limits	
kidsdriv	1.197	1.058	1.355
tif	0.976	0.960	0.992
CAR_TYPE Minivan vs z_SUV	0.339	0.268	0.428
CAR_TYPE Panel Truck vs z_SUV	0.404	0.283	0.576
CAR_TYPE Pickup vs z_SUV	0.642	0.500	0.826
CAR_TYPE Sports Car vs z_SUV	1.280	1.024	1.600
CAR_TYPE Van vs z_SUV	0.543	0.397	0.744
CAR_USE Commercial vs Private	2.416	1.963	2.975
EDUCATION <High School vs z_High School	1.076	0.869	1.334
EDUCATION Bachelors vs z_High School	0.600	0.497	0.725
EDUCATION Masters vs z_High School	0.632	0.457	0.873
EDUCATION PhD vs z_High School	0.565	0.377	0.847
JOB Clerical vs z_Blue Collar	1.258	0.990	1.599
JOB Doctor vs z_Blue Collar	0.511	0.268	0.977
JOB Home Maker vs z_Blue Collar	1.300	0.952	1.775
JOB Lawyer vs z_Blue Collar	0.749	0.490	1.146
JOB Manager vs z_Blue Collar	0.372	0.270	0.511
JOB NA vs z_Blue Collar	0.651	0.429	0.988
JOB Professional vs z_Blue Collar	0.869	0.664	1.139
JOB Student vs z_Blue Collar	1.307	1.009	1.694
MSTATUS Yes vs z_No	0.609	0.523	0.710
PARENT1 No vs Yes	0.553	0.448	0.682
REVOKED No vs Yes	0.467	0.390	0.560
SEX M vs z_F	1.361	1.113	1.664
URBANICITY Highly Urban/ Urban vs z_Highly Rural/ Rural	12.631	9.742	16.376

Model 4 – Decision Tree

This method uses all predictors from Model 1 and builds a classification tree that ends up consisting of 11 variables: job, urbanicity, car_type, revoked, mstatus, car_use, education, tif, log_home_val, m_yoj, and car_age. The most important variables in the tree, meaning the most predictive variables, are given by variable importance values, as shown below. If an observation is dropped down the tree, then it's urbanicity value is checked first, as shown below. If highly rural, then the process ends and the predicted value of P(Y=1) is .0627. If highly urban, then the process continues until it hits the other terminal nodes, each with their own P(Y=1) predicted values.

Variable Importance

Variable	Variable Label	Training Relative	Training Importance	Count
JOB	Job Category	1.0000	11.6290	1
URBANICITY	Home/Work Area	0.9462	11.0034	1
CAR_TYPE	Type of Car	0.5126	5.9606	2
REVOKED	License Revoked (Past 7 Years)	0.4091	4.7577	1
MSTATUS	Marital Status	0.2803	3.2594	1
CAR_USE	Vehicle Use	0.2387	2.7753	1
EDUCATION	Max Education Level	0.2344	2.7262	1
tif		0.1849	2.1507	1
log_home_val		0.1797	2.0892	1
M_YOJ		0.1556	1.8100	1
car_age		0.1434	1.6678	1



MODEL SELECTION

Below, I evaluate the best model based on multiple criterion. I compare models' AIC, SC and deviance in round 1, models' KS in round 2, and models' AUC in round 3. A ">" sign means the model on the left side of the sign performed better than the model on the right side.

Results	Metrics Compared
Round 1: Model 2 = Model 3 > Model 1	AIC, SC, Deviance
Round 2: Model 1 > Model 2 > Model 3	KS
Round 3: Model 1 > Model 2 > Model 3	AUC

Selected Model

I select Model 2. It has about the same AIC, SC, and Deviance as Model 3, but a better KS and AUC than model 3. It's SC is over 100 points lower than Model 1's SC because Model 1 has more predictors and is penalized for having greater model complexity. While Model 1 has a better KS and AUC than Model 2, these values are not much better, while Model 2's SC is much better than Model 1's SC. While I only had the ROC metric available for the Decision Tree, its AUC was far too low compared to the other models (.74 vs .78+).

Explanation of Metrics Used

Deviance (-2 Log L) is based on the model's likelihood value in that the smaller the deviance, the greater the model's likelihood. AIC is the deviance adjusted with a penalty for model complexity (more predictors). SC is the deviance adjusted with a greater penalty than that of AIC for model complexity. We want to minimize deviance, AIC, and SC.

The Kolmogorov-Smirnov (KS) statistic, which is to be maximized, is the maximum difference between each associated sensitivity and specificity values.

The area under the curve (AUC) of the ROC curve measures the model's ability to accurately classify Y=1 and Y=0 across all potential thresholds, as shown by the ROC curve.

Round 1

Model 1 – All Variables

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6600.702	5450.174
SC	6607.353	5736.145
-2 Log L	6598.702	5364.174

Model 2- Variables Removed (tie)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6600.702	5428.607
SC	6607.353	5608.171
-2 Log L	6598.702	5374.607

Model 3- Stepwise (tie)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6600.702	5428.175
SC	6607.353	5601.088
-2 Log L	6598.702	5376.175

Model 4 – Decision Tree

NA

Round 2

Model 1 – All Variables (best)

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.197445	D	0.447757
KSa	14.923760	Pr > KSa	<.0001

Model 2 – Variables Removed

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.195585	D	0.443538
KSa	14.783140	Pr > KSa	<.0001

Model 3 – Stepwise

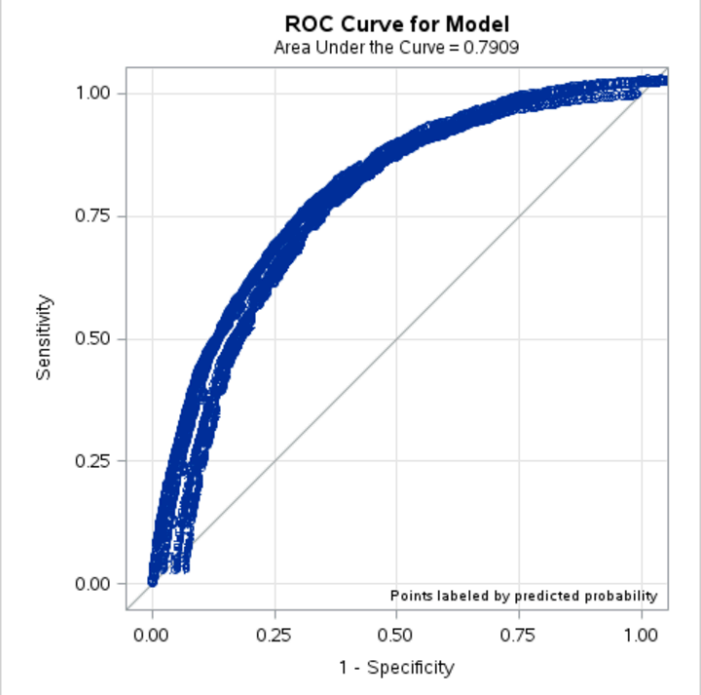
Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.194731	D	0.441603
KSa	14.718639	Pr > KSa	<.0001

Model 4 – Decision Tree

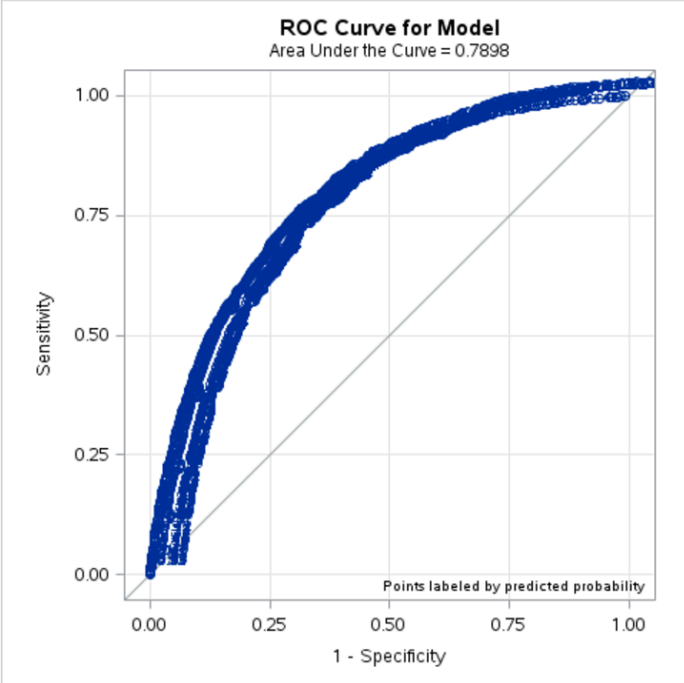
NA

Round 3

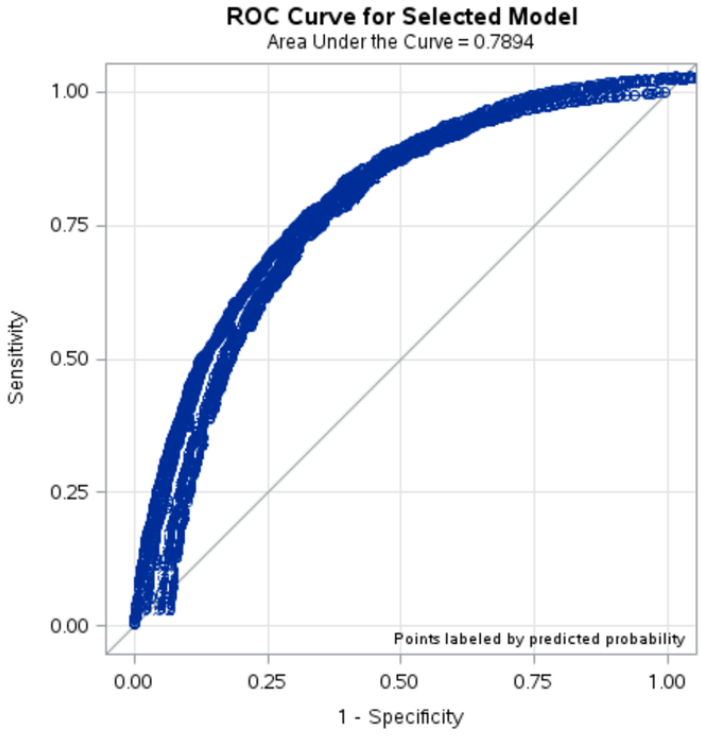
Model 1 – Logistic (best)



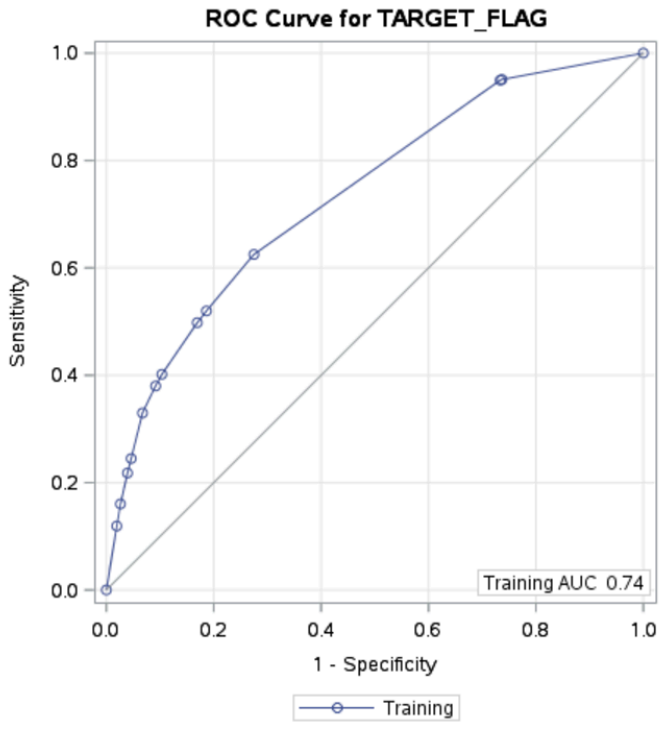
Model 2 – Variables Removed



Model 3 – Stepwise



Model 4 – Decision Tree



CONCLUSION

The selected logistic model (Model 2) offers a consider improvement above an intercept-only model in predicting whether or not an insured will crash, based upon that insured's characteristics. It offers a low SC without much reduction in KS or AUC. This model, which determines the likelihood of a claim, can be used with a model that predicts claim severity, the amount of a claim, given that there is a claim. Together, frequency x severity = expected claim amount. These kinds of models are absolutely vital in the insurance industry, thus illustrating the need for data scientists as well as actuaries.

It is important to note that selection of predictors in this analysis have excluded any consideration to ethical, legal, or societal concerns. State laws may prohibit the inclusion of certain variables in ratemaking, where a model like this would undoubtedly contribute towards ratemaking at an insurance company. Additionally, if the public became aware of certain variables used in an insurer's ratemaking activities, as evidenced by certain subgroups' premiums going up, and if this awareness caused an uproar over the fact that a certain variable was used, then the damage to the insurer's reputation might not be worth the competitive advantage gained from using the rating variable.