Assignment #2: Insurance

PREDICT 411 Section 55

Scott M. Morgan

KAGGLE: ScottMorgan

Scott Morgan
KAGGLE: ScottMorgan

DATE: July 30, 2017

TO: Dr. Donald Wedding, President of Wedding Insurance

FROM: Mr. Scott Morgan, Senior Analytics Associate

SUBJECT: Enhanced Predictive Capabilities

EXECUTIVE SUMMARY

The US is facing a challenging landscape given the new regime in the White House and continued struggle for top and bottom line growth in a slowly recovering economy. Coupled with volatility in international economies and mercurial oil prices, the only thing that is certain is uncertainty. Short-term, auto-insurers could feel these growth struggles by way of fewer vehicle sales. Longer-term, the move toward driverless vehicles and ride-sharing threatens to eliminate millions (if not billions) of insurable risks. The battle for market share and profitability may depend on the expanding the use of predictive analytics to overcome growth challenges and threats of transformational technology.

It has come to the attention of senior leadership the need to devise a more robust system of identifying risk among its customer base. The factors that increase customers' propensity to wreck their vehicles is a source of great interest to our business as it influences the premiums we charged which, in turn, impacts profitability. The knowledge of demographic backgrounds and behavioral tendencies that identifies risky factors, and the development of predictive models based on those metrics, is particularly important to stakeholders charged with policy decisions.

Overall, this exercise is intended to ensure Wedding Insurance is utilizing forward-looking analysis in our pricing structure to protect from possible asymmetric risks caused by antiquated valuing methodologies. Phrased differently, we want to ensure customers are being charged appropriately given their perceived level of risk. The following report is a detailed account of the predictive modeling process that aims to accomplish this.

To summarize, our basic managerial recommendation is to be particularly mindful when establishing policies with customers that have the following characteristics as they represent higher-risk profiles:

- Customers in urban areas (i.e. cities or densely populated areas)
- Customers who have frequently filed claims in the last 5 years
- Customers who frequently get traffic violations (i.e. motor vehicle record points)
- Customers who have had their drivers license revoked in the last 7 years
- Customers who are currently in high school or college

Scott Morgan
KAGGLE: ScottMorgan

INTRODUCTION

The purpose of this report is to build logistic regression models to predict the probability of a customer crashing their automobile. Logistic regression is a binary statistical technique considered powerful when the purpose is to determine the likelihood of an event or the probability of its occurrence (Schwert, 2000). To accomplish this objective, I use historical insurance data provided by the course instructor and functionality within SAS Studio and SAS Enterprise Miner to produce an end-to-end predictive modeling process which utilizes an assortment of numerical and categorical variables to predict the propensity for customers to get into automobile accidents. While the analysis uses several robust, enterprise quality analytics systems, our primary focus remains simplicity and interpretability.

RESULTS

In the subsequent sections, we generate and evaluate a series of predictive models. We first use the functionality within SAS Studio and SAS Enterprise Miner to perform a brief exploratory data analysis (EDA) to build an understanding of potential predictor variables and their relationship to the response. Following this, we examine the variables for deficiencies such as missing data and outliers as a precursor to preparing the data set for modeling through imputation and elimination. Finally, we construct a series of predictive models. The first two models serve as base lines to gain an understanding of how continuous and categorical variables impact the results of logistical regression by looking at them separately (i.e. one model containing strictly continuous variables, the other containing strictly categorical variables). The third iteration uses all of the available variables and generates a model using automated variable selection techniques. The fourth model leverages SAS Enterprise Miner and decision trees to identify variables of importance. Once identified, these variables are used in a final regression analysis through SAS Studio. The final model uses probit regression instead of logistic to serve as a comparison of the two methods. The primary quantitative metrics we will be using for comparison are Akaike Information Criterion (AIC), Somer's D, Gamma, Area Under the Curve and Kolmogorov-Smirnov (K-S) statistic. We then recommend the most logical, effective solution for use by management.

EXPLORATORY DATA ANALYSIS (EDA). We will be using an assortment of tools to perform the initial EDA before cleaning the data and ultimately constructing the predictive models. I begin the analysis by examining the variables in the data set using SAS content procedure. The data set contains 8161 observations and 24 variables; one variable is a unique identifier and thus is excluded from the analysis. There are a total of 13 continuous variables and 10 categorical variables. Each record represents a customer at an auto insurance company. Each record has two target variables. The target variable, TARGET_FLAG, is a 1 or a "0". A "1" indicates that the person was in a car crash. A zero indicates that the person was not in a car crash. The data set primarily consist of non-negative discrete variables as well as categorical variables.

While the naming convention of the variables is relatively effective, the data dictionary is a helpful resource. The data set itself has a wide variety of statistics that appear analytically interesting and appropriate to build a model with to predict the propensity of customer to get into car accidents. Table 1 below provides an alphabetic list of the possible predictor variables, descriptions, data types and correlations of the continuous variables with the TARGET_FLAG (red represents negatively correlated while green is positively correlated).

Table 1: Alphabetic List of Variables and Correlations

Variable	Description	Data Type	Correlations
AGE	Age of Driver	Continuous	-0.10
BLUEBOOK	Value of Vehicle	Continuous	-0.10
CAR_AGE	Vehicle Age	Continuous	-0.10
CAR_TYPE	Type of Car	Categorical	N/A
CAR_USE	Vehicle Use	Categorical	N/A
CLM_FREQ	#Claims(Past 5 Years)	Continuous	0.22
EDUCATION	Max Education Level	Categorical	N/A
HOMEKIDS	#Children @Home	Continuous	0.12
HOME_VAL	Home Value	Continuous	-0.18
INCOME	Income	Continuous	-0.14
JOB	Job Category	Categorical	N/A
KIDSDRIV	#Driving Children	Continuous	0.10
MSTATUS	Marital Status	Categorical	N/A
MVR_PTS	Motor Vehicle Record Points	Continuous	0.22
OLDCLAIM	Total Claims(Past 5 Years)	Continuous	0.14
PARENT1	Single Parent	Categorical	N/A
RED_CAR	A Red Car	Categorical	N/A
REVOKED	License Revoked (Past 7 Years)	Categorical	N/A
SEX	Gender	Categorical	N/A
TIF	Time in Force	Continuous	-0.08
TRAVTIME	Distance to Work	Continuous	0.05
URBANICITY	Home/Work Area	Categorical	N/A
YOJ	Years on Job	Continuous	-0.07

We also include the data dictionary and theoretical effects for reference (Table 2). The theoretical effects are particularly important as we will reference this logic during the examination of coefficients in the model building process.

Table 2: Data Dictionary and Theoretical Effects

VARIABLE NAME	THEORETICAL EFFECT
AGE	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Unknown effect, but probably effect the payout if there is a crash.
CAR_AGE	Unknown effect, but probably effect the payout if there is a crash.
CAR_TYPE	Unknown effect, but probably effect the payout if there is a crash .
CAR_USE	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	The more claims you filed in the past, the more you are likely to file in the future.
EDUCATION	Unknown effect, but in theory more educated people tend to drive more safely.
HOMEKIDS	Unknown effect, though more kids at home could mean more teenager drivers.
HOME_VAL	Home owners tend to drive more responsibly.
INCOME	Rich people tend to get into fewer crashes.
JOB	White collar jobs tend to be safer.
KIDSDRIV	When teenagers drive your car, you are more likely to get into crashes.
MSTATUS	Married people drive more safely.
MVR_PTS	If you get lots of traffic tickets, you tend to get into more crashes.
OLDCLAIM	If your total payout over the past five years was high, future payouts could be high.
PARENT1	Single parents are the primary driver which could lead to more accidents.
RED_CAR	Urban legend says that red cars (especially red sports cars) are more risky.
REVOKED	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Urban legend says that women have less crashes then men.
TIF	People who have been customers for a long time are usually more safe.
TRAVTIME	Long drives to work usually suggest greater risk.
URBANICITY	People who drive in cities are more likely to get into accidents.
YOJ	People who stay at a job for a long time are usually more safe.

Correlations for Continuous Variables. Using the PROC CORR function we begin our analysis of the continuous variables. From the table above, we can see that Motor Vehicle Record Points, #Claims(Past 5 Years), Total Claims(Past 5 Years)are mildly positively correlated with TARGET_FLAG. Intuitively these make sense as more claims and tickets are likely predictive of a driver's propensity to get into accidents. Conversely, only Home Value and Income are mildly negatively correlated with TARGET_FLAG. These are also logical and in the predicted direction. The magnitude of these findings is somewhat disappointing as we hope for a correlation of +0.5 (-0.5) or more (less). We reserve EDA of the categorical variables for a later section.

Mean Analysis for Continuous Variables. Using the PROC MEANS function with the CLASS set to TARGET_FLAG, we are able to decompose the data set further and compare the mean values for all continuous variables by zero and 1. The objective here is to corroborate the findings from the correlation analysis and possibly identify additional predictors. As a rule of thumb, a meaningful difference in average values could signal predictive power. The abbreviated version of this output is provided in Table 3 below. While the relative change needs to be considered with regard to the respective variables, several insights can be extracted from the table. There appears to noteworthy differences between #Driving Children and #Children@Home; which is because the two are likely related. Drivers who make more money are also more susceptible to getting into crashes, which is contrary to expectations. More expensive vehicles also tend to get into more accidents; this raises suspicion that these are higher-end

recreational vehicles. Total Claims(Past 5 Years), #Claims(Past 5 Years) and Motor Vehicle Record Points, the variables identified in the prior section, also have meaningful differences between those who get into accidents and those who don't. All of these variables of interest are bolded in Table 3 below.

Table 3: Mean Analysis of Continuous Variables by TARGET_FLAG

Variable	Label	No Crash (0)	Crash (1)	Delta
KIDSDRIV	**#Driving Children**	**-**	**0.26**	**(0.26)**
AGE	Age	43.00	43.30	(0.30)
HOMEKIDS	**#Children @Home**	**-**	**0.94**	**(0.94)**
YOJ	Years on Job	11.00	10.02	0.98
INCOME	**Income**	**43,971.50**	**50,641.30**	**(6,669.80)**
HOME_VAL	Home Value	114,564.82	115,256.55	(691.73)
TRAVTIME	Distance to Work	34.44	34.77	(0.33)
BLUEBOOK	**Value of Vehicle**	**12,600.00**	**14,255.90**	**(1,655.90)**
TIF	Time in Force	4.00	4.78	(0.78)
OLDCLAIM	**Total Claims(Past 5 Years)**	**2,448.00**	**6,061.55**	**(3,613.55)**
CLM_FREQ	**#Claims(Past 5 Years)**	**1.00**	**1.22**	**(0.22)**
MVR_PTS	**Motor Vehicle Record Points**	**2.00**	**2.48**	**(0.48)**
CAR_AGE	Vehicle Age	7.00	7.37	(0.37)

Categorical Variables of Interest. As we have established a relatively firm idea of what continuous variables we will include in the models, we shift our attention to identifying categorical variables of interest. Using the PROC FREQ function, we are able to view the various categorical variables by TARGET_FLAG and examine which categories get into more car accidents (i.e. "PROC EYEBALL"). Table 4 provides a summary of this output. For the sake of brevity, we only include those variables that are analytically interesting and bold them for ease of viewing. RED_CAR(A Red Car) and SEX(Gender) were excluded from the analysis as there doesn't appear to be meaningful relationships between those who get into crashes and those who don't. To summarize:

- ***PARENT1(Single Parent):*** Single parents appear to crash their vehicles more frequently. This is perhaps because they are the primary drivers in the family.
- ***MSTATUS(Marital Status):*** Single people crash their vehicles more frequently. This logic is similar to above where they are the primary drivers.
- ***EDUCATION(Max Education Level):*** Drivers in high school (or before) crash their vehicles more frequently than people who are more educated.
- ***JOB(Job Category):*** Students and blue collar workers crash their vehicles more frequently. We posit for students that this is due to their lack of experience driving. For blue collar workers this could be because they are frequently required to drive as part of their job.
- ***CAR_USE(Vehicle Use):*** Vehicles being driven for commercial use crash more frequently. This is likely due to the vehicles being on the road for more time.
- ***REVOKED(License Revoked (Past 7 Years)):*** Drivers who have had their licenses revoked crash their vehicles more frequently than those who haven't. These individuals are possibly less responsible drivers in general.
- ***URBANICITY(Home/Work Area):*** Drivers in highly urban areas (i.e. cities) crash their vehicles much more frequently than those people who drive in rural areas. This makes sense as there are more obstacles and driving distractions in densely populated areas.

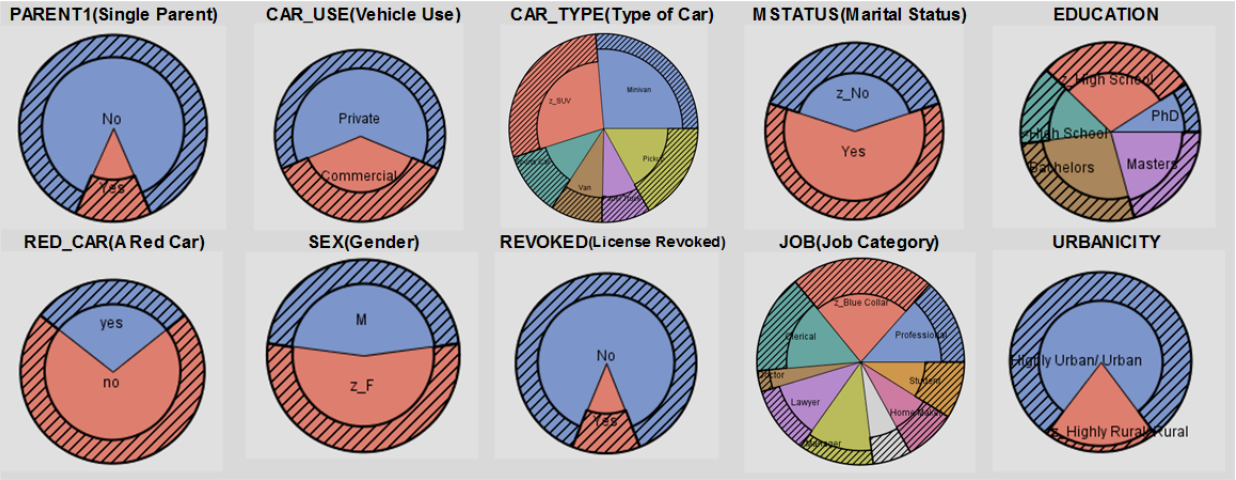
Table 4: PROC Frequency (i.e. PROC "Eyeball") – Frequency of Accidents

Variable	No Crash (0)	Crash (1)	Variable	No Crash (0)	Crash (1)
PARENT1(Single Parent)			**JOB(Job Category) - CONTINUED**		
No	76.33	23.67	Lawyer	81.68	18.32
Yes	**55.8**	**44.2**	Manager	86.13	13.87
			Professional	77.89	22.11
MSTATUS(Marital Status)			**Student**	**62.64**	**37.36**
Yes	78.48	21.52	**z_Blue Collar**	**65.26**	**34.74**
z_No	**66.33**	**33.67**			
			CAR_USE(Vehicle Use)		
EDUCATION			**Commercial**	**65.43**	**34.57**
<High School	68.00	32.00	Private	78.45	21.55
Bachelors	76.67	23.33			
Masters	80.28	19.72	**REVOKED(License Revoked (Past 7 Years))**		
PhD	82.83	17.17	No	76.12	23.88
z_High School	**65.97**	**34.03**	**Yes**	**55.7**	**44.3**
JOB(Job Category)			**URBANICITY(Home/Work Area)**		
Clerical	70.81	29.19	**H. Urban/ Urban**	**68.61**	**31.39**
Doctor	88.21	11.79	**z_H. Rural/ Rural**	**93.11**	**6.89**
Home Maker	71.92	28.08			

Lastly, we utilize functionality in SAS Enterprise Miner to complete the analysis of the categorical variables. This is done in a SAS Code Node. Exhibit 1 below provides the proportion of crashes ("1") in each categorical variable. These are in line with our observations from Table 4 and anecdotally a more visually appealing way to view the information. A noteworthy observation is that URBANCITY and JOB appear potentially predictive as a large portion of their respective pie charts are shaded.

Scott Morgan
KAGGLE: ScottMorgan

(BINGO BONUS #1) Exhibit 1: Frequency of Car Accidents ("1") by Categorical Variable



Outliers. As we begin to transition into the data preparation portion of the modeling process, we take a step back to identify potential outliers in the continuous variables. Table 5 below provides a detailed statistical summary of all the continuous variables in the data set. Variables that appear to have possible outliers are INCOME and HOME_VAL. There is also a negative number for CAR_AGE which suggests poor data quality. There could be a number reasons for the presence outliers and poor data quality. First, there are data integrity issues when dealing with publicly available data sets. The original proprietors of the data altered and scaled the data differently than what was initially measured. Additionally, having not collected these data ourselves we are forced to rely on the original practitioners in terms of collection techniques and accuracy. We discuss how to handle the presence of outliers in the subsequent section on data preparation.

Missing Data. Similar to identifying outliers, Table 5 assists in finding continuous variables where data is missing. In the insurance data set, there are several variables with missing data: AGE, YOJ, INCOME, HOME_VAL and CAR_AGE. We again reserve discussing the resolution of this issue to the following discussion on data preparation. Lastly, using the PROC FREQ function (not shown) we can see that in terms of categorical variables, 526 records are missing in the JOB classification, or 6.45% of the total data set.

Table 5: The MEANS Procedure (Raw Data Set - Continuous Variables)

Variable	N Miss	Mean	Min	5th Pctl	Median	95th Pctl	Max
KIDSDRIV	0	0.17	0.00	0.00	0.00	1.00	4.00
AGE	6	44.79	16.00	30.00	45.00	59.00	81.00
HOMEKIDS	0	0.72	0.00	0.00	0.00	3.00	5.00
YOJ	454	10.50	0.00	0.00	11.00	15.00	23.00
INCOME	445	61,898.10	0.00	0.00	54,028.17	152,283.24	367,030.26
HOME_VAL	464	154,867.29	0.00	0.00	161,159.53	374,931.16	885,282.34
TRAVTIME	0	33.49	5.00	7.02	32.87	60.41	142.12
BLUEBOOK	0	15,709.90	1,500.00	4,900.00	14,440.00	31,110.00	69,740.00
TIF	0	5.35	1.00	1.00	4.00	13.00	25.00
OLDCLAIM	0	4,037.08	0.00	0.00	0.00	27,090.00	57,037.00
CLM_FREQ	0	0.80	0.00	0.00	0.00	3.00	5.00
MVR_PTS	0	1.70	0.00	0.00	1.00	6.00	13.00
CAR_AGE	510	8.33	-3.00	1.00	8.00	18.00	28.00

The EDA uncovered several interesting elements in the data set. A number of variables traditionally thought of to be associated with higher crash propensities where found not to be relevant. These variables include the type of vehicle, gender and the color red. Overall, we successfully identified several continuous and categorical variables to include in the regression models discussed in subsequent sections. Below is a summary of the variables we consider predictive and will keep under consideration throughout the model building process:

- ***Continuous Variables:*** OLDCLAIM, CLM_FREQ and MVR_PTS

- ***Categorical Variables:*** URBANICITY(Home/Work Area), JOB and REVOKED(License Revoked (Past 7 Years))

DATA PREPARATION. In this section, we prepare the data by changing the original values as well as create several new variables. Before moving on to the model portion of the report, we present a discussion of the new data set.

Outlier Resolution. In attempt to create a better fit for our models, we first modify the data set to eliminate possible outliers as they can significantly influence regression results. To accomplish this, we impute outlier values, replacing them with the 5^{th} and 95^{th} percentile breakpoints for each respective variable. The 95^{th} percentile breakpoint is applied to INCOME and HOME_VAL as they have extreme high values. CAR_AGE has a negative minimum, which is impossible and we therefore impute any value less than 1 with the 5^{th} percentile breakpoint.

Missing Data Resolution. Similar to the outlier remedy, we impute missing values in the data set by replacing them with the median values of the distribution. While there are more sophisticated replacement techniques, median imputation is used because it is a number that is already present in the data set and is less susceptible to outlier errors as compared to mean imputation. In addition to median imputation, a missing flag for each variable is generated to determine whether there is a difference in outcomes associated with missing versus complete data. We apply this methodology to AGE, YOJ, INCOME and CAR_AGE. Lastly, we derive the missing JOB data points by using imputed salary ranges for Doctors and Lawyers.

The resultant structure of the new data set (not shown) has no missing values, 1 imputed categorical variable and 5 additional continuous variables in the form of missing data flags. Outlier deletion and imputation are intended to improve the robustness of models and can have powerful effects on fit. While the effects are sometimes not in the desired direction, these analytical activities can be beneficial if they improve the relationship between two or more variables.

BUILD MODELS. In this section, we generate 5 regression models and discuss the findings. All models use the cleaned data set described above and reference coding. Note that for categorical variables the PROC LOGISTIC function automatically assigns a basis of interpretation.

Model 1: Continuous Variables. For our initial model, we use all of the continuous variables in the data set. Using the SAS logistic regression procedure, we generate the following results (Table 6):

Table 6: Analysis of Maximum Likelihood Estimates for Model 1

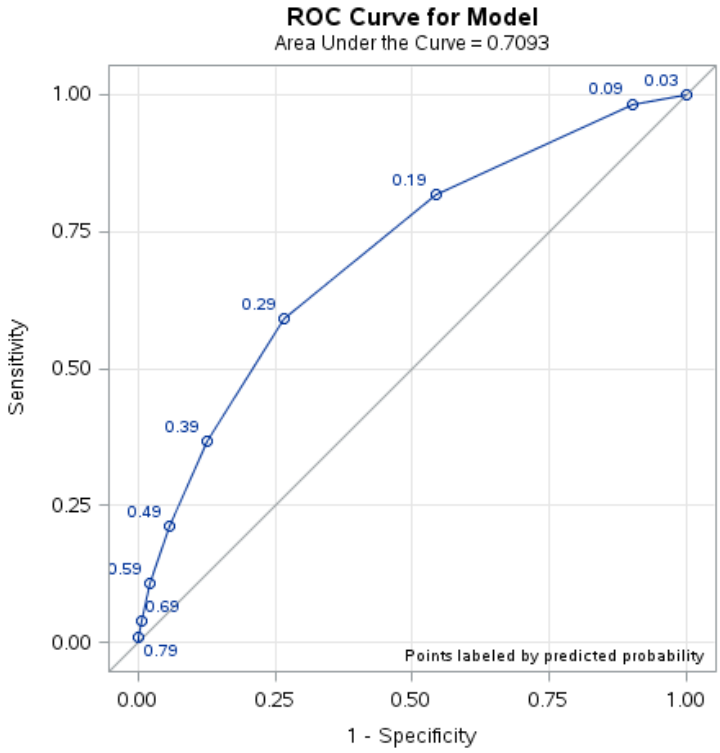
Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-0.4318	0.1894	5.1997	0.0226
KIDSDRIV	0.299	0.0551	29.4592	<.0001
HOMEKIDS	0.0677	0.0301	5.0558	0.0245
TRAVTIME	0.00665	0.00167	15.8703	<.0001
BLUEBOOK	-0.00001	3.64E-06	11.9836	0.0005
TIF	-0.0468	0.00679	47.4786	<.0001
OLDCLAIM	6.33E-06	3.14E-06	4.0779	0.0434
CLM_FREQ	0.2632	0.0256	105.3009	<.0001
MVR_PTS	0.1386	0.0126	121.0722	<.0001
IMP_INCOME	-1.68E-06	9.01E-07	3.4928	0.0616
IMP_AGE	-0.00838	0.00361	5.3986	0.0202
M_IMP_AGE	2.0657	1.1122	3.4497	0.0633
M_INCOME	-0.0767	0.1194	0.4124	0.5208
IMP_HOME_VAL	-2.71E-06	2.75E-07	96.9112	<.0001
M_HOME_VAL	0.0119	0.1151	0.0106	0.9178
IMP_YOJ	-0.00436	0.00702	0.3857	0.5345
M_YOJ	0.0586	0.1157	0.2569	0.6123
IMP_CAR_AGE	-0.0189	0.00546	11.9507	0.0005
M_CAR_AGE	0.0832	0.1088	0.5847	0.4445

The interpretation of Model 1 is a relatively straightforward. For the sake of brevity, we only interpret the first 3 coefficients. For every 1 unit increase in the number of teenagers driving a customer's car, the odds of getting into an accident increases by 0.299. For every 1 unit increase in the number of children at home, the odds of getting into an accident increases by 0.0677. Lastly, for every one-unit increase in the time it takes to travel to work, the odds of getting into an accident increases by 0.0066.

Model 1 includes 8 original variables, 5 imputed variables and 5 flags for missing variables. The majority of the coefficients are in the predicted direction and statistically significant, which is encouraging. KIDSDRIV, CLM_FREQ, MVR_PTS and M_IMP_AGE all appear to be meaningful. It makes sense as the number of teenager drivers, frequency of claims and traffic tickets increase the propensity to get into accidents increases, but it is also interesting that the missing age variable appears predictive of car crashes (though not statistically significant). This could be because higher risk drivers might leave this information out of policy applications because they think it is beneficial somehow. The claims of predictive power regarding number of children driving, claim frequency and motor vehicles recorded points during the EDA are corroborated in these results.

We next examine the ROC (Exhibit 2) and see that Model 1 is performing at 70.93% coverage, which is a good start.

Exhibit 2: Model 1 ROC



Model 2: Categorical Variables. For the second model, we use all of the categorical variables in the data set. Using the SAS logistic regression procedure, we generate the following results (Table 7):

Table 7: Analysis of Maximum Likelihood Estimates for Model 2

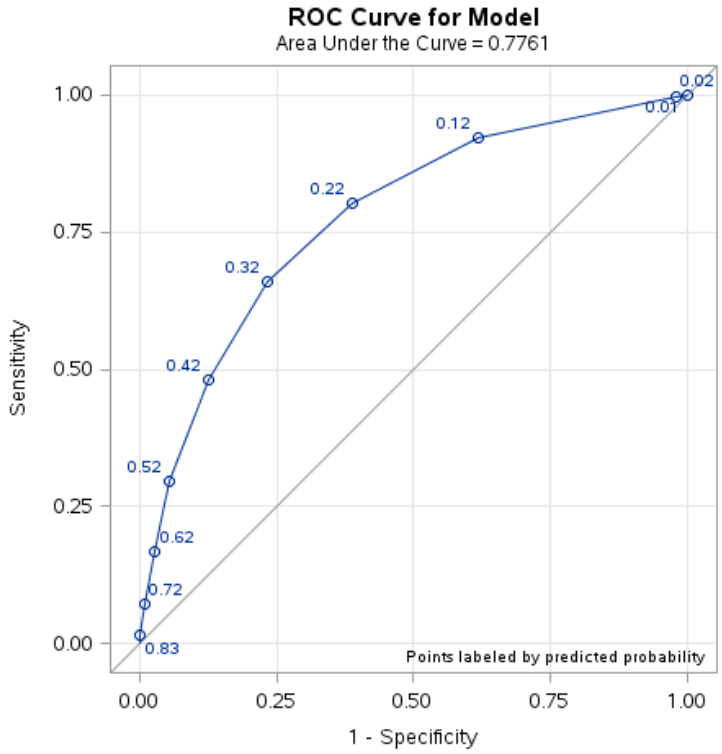
Analysis of Maximum Likelihood Estimates					
Parameter		**Estimate**	**Standard Error**	**Wald Ch-Sq**	**Pr > ChiSq**
Intercept		-1.2628	0.1827	47.7972	<.0001
CAR_TYPE	**Minivan**	-0.9811	0.0993	97.5716	<.0001
CAR_TYPE	**Panel Truck**	-0.7845	0.1506	27.141	<.0001
CAR_TYPE	**Pickup**	-0.3859	0.107	13.0166	0.0003
CAR_TYPE	**Sports Car**	0.2757	0.0946	8.4936	0.0036
CAR_TYPE	**Van**	-0.5455	0.1331	16.805	<.0001
CAR_USE	**Commercial**	0.7656	0.085	81.1234	<.0001
EDUCATION	**<High School**	0.0454	0.0908	0.2505	0.6167
EDUCATION	**Bachelors**	-0.4748	0.0795	35.6802	<.0001
EDUCATION	**Masters**	-0.4822	0.1154	17.4692	<.0001
EDUCATION	**PhD**	-0.5834	0.1538	14.3858	0.0001
IMP_JOB	**Clerical**	0.2122	0.099	4.5972	0.032
IMP_JOB	**Doctor**	-0.654	0.1993	10.7641	0.001
IMP_JOB	**Home Maker**	0.3049	0.1261	5.8422	0.0156
IMP_JOB	**Lawyer**	-0.3192	0.1416	5.0805	0.0242
IMP_JOB	**Manager**	-0.9637	0.1231	61.248	<.0001
IMP_JOB	**Professional**	-0.1782	0.1084	2.7013	0.1003
IMP_JOB	**Student**	0.3625	0.1075	11.3752	0.0007
MSTATUS	**Yes**	-0.5368	0.0654	67.3749	<.0001
PARENT1	**No**	-0.6391	0.0886	51.9825	<.0001
RED_CAR	**no**	-0.0293	0.0835	0.123	0.7258
REVOKED	**No**	-0.7669	0.0775	98.012	<.0001
SEX	**M**	0.2428	0.0996	5.9461	0.0148
URBANICITY	**Highly Urban/ Urban**	2.3952	0.1069	502.1873	<.0001

The interpretation of Model 2 is slightly different than Model 1 due to the inclusion of categorical variables. For the sake of brevity, we only interpret the first coefficients for CAR_TYPE, EDUCATION and IMP_JOB. If the customer is driving a Minivan, the odds of getting into an accident decreases by -0.9811. If the customer is a high school dropout, the odds of getting into an accident increase by 0.0454. Lastly, if the customer has a clerical job, the odds of getting into an accident increase by 0.2122.

Model 2 includes 9 of the original categorical variables and 1 imputed variable. Once again, most of the coefficients are in the predicted direction and make sense intuitively. Sports cars and commercial use vehicles all increase the propensity of getting into accidents, as does being a clerical worker and a student. Being a male also increases the likelihood of accidents, though this is not statistically significant. Having a red car surprisingly decreases the likelihood of accidents, though this was also not statistically significant. The key finding from this model is that living in a highly urban area strongly increases the chances of accidents. This and the job classification supports our findings from the EDA.

We next examine the ROC (Exhibit 3) and see that Model 2 is performing at 77.61% coverage, a notable improvement from Model 1.

Exhibit 3: Model 2 ROC



Model 3: Automated Selection. For the third model, we use automated variable selection to generate 3 different models. Automated variable selection methods are useful when there are many variables to evaluate and manual selection becomes cumbersome. For this exercise we use 3 common computational techniques (Stepwise, Forward and Backward) for generating subset regression models to arrive at Model 3. The main criteria we will be evaluating on are the AIC and ROC (not shown). All continuous and categorical variables are available as candidates for inclusion. Using the SAS logistic regression procedure, we generate the following results (Table 8) for the best subset model which uses Forward Selection:

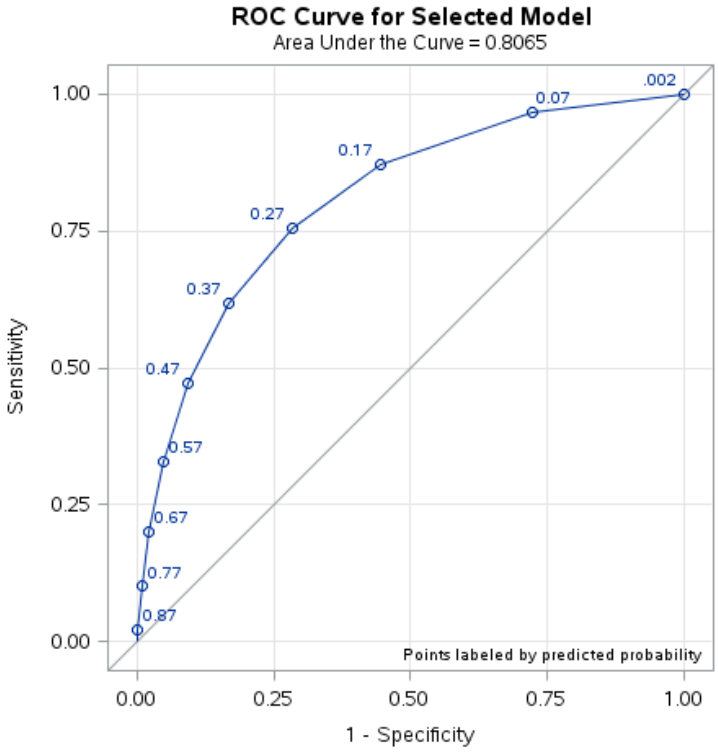
Table 8: Analysis of Maximum Likelihood Estimates for Model 3 (FORWARD)

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		-1.3134	0.2064	40.4844	<.0001
KIDSDRIV		0.4238	0.0552	58.9271	<.0001
TRAVTIME		0.0146	0.0019	59.9320	<.0001
BLUEBOOK		0.0000	0.0000	22.7576	<.0001
TIF		-0.0557	0.0074	57.3324	<.0001
OLDCLAIM		0.0000	0.0000	13.0952	0.0003
CLM_FREQ		0.1961	0.0285	47.2306	<.0001
MVR_PTS		0.1129	0.0136	68.9348	<.0001
IMP_INCOME		0.0000	0.0000	11.9386	0.0005
M_IMP_AGE		2.2038	1.2371	3.1734	0.0748
IMP_HOME_VAL		0.0000	0.0000	18.5477	<.0001
CAR_TYPE	Minivan	-0.7155	0.0860	69.1658	<.0001
CAR_TYPE	Panel Truck	-0.1422	0.1500	0.8983	0.3433
CAR_TYPE	Pickup	-0.1824	0.0933	3.8224	0.0506
CAR_TYPE	Sports Car	0.2554	0.0980	6.7875	0.0092
CAR_TYPE	Van	-0.0867	0.1201	0.5205	0.4706
CAR_USE	Commercial	0.7836	0.0885	78.4841	<.0001
EDUCATION	<High School	-0.0123	0.0943	0.0171	0.8961
EDUCATION	Bachelors	-0.3990	0.0837	22.7344	<.0001
EDUCATION	Masters	-0.3923	0.1205	10.5955	0.0011
EDUCATION	PhD	-0.3751	0.1632	5.2844	0.0215
IMP_JOB	Clerical	0.1146	0.1040	1.2130	0.2707
IMP_JOB	Doctor	-0.3510	0.2177	2.5988	0.1069
IMP_JOB	Home Maker	-0.0315	0.1383	0.0521	0.8195
IMP_JOB	Lawyer	-0.0999	0.1494	0.4473	0.5036
IMP_JOB	Manager	-0.7895	0.1283	37.8918	<.0001
IMP_JOB	Professional	-0.0914	0.1133	0.6512	0.4197
IMP_JOB	Student	-0.0797	0.1236	0.4155	0.5192
MSTATUS	Yes	-0.4524	0.0802	31.8132	<.0001
PARENT1	No	-0.4569	0.0944	23.4039	<.0001
REVOKED	No	-0.8955	0.0913	96.2021	<.0001
URBANICITY	Highly Urban/ Urban	2.3888	0.1129	447.3708	<.00

The interpretation of Model 3 is similar to Model 1 and Model 2 as it contains both continuous and categorical variables and, therefore, we elect not to repeat the exercise. Model 3 includes 9 continuous variables, 1 missing continuous variable flag and 8 categorical variables. Most coefficients are in the predicted direction. Interestingly, the algorithm selected several continuous variables that have very marginal impacts. These include BLUEBOOK, OLDCLAIM, IMP_INCOME and IMP_HOME_VAL. M_IMP_AGE again had a large coefficient though it is not statistically significant. In terms of categorical variables, the various car types were all in the predicted direction, though only the Minivan was statistically significant. EDUCATION is an interesting case where the coefficient for drivers with below a high school education switched from positive to negative. This is not in line with expectations. Finally, living in a highly urban area was again found to strongly increase the chances of accidents. This variable appears to be highly predictive.

We next examine the ROC (Exhibit 4) and see that Model 3 is performing at 80.65% coverage, the best thus far. The inclusion of both continuous and categorical variables appears to create a more robust predictive model.

Exhibit 4: Model 3 ROC



Overall, the results for Model 3 are encouraging but the changing of coefficient directionality is an area of concern.

Model 4: Decision Tree Imputation and Decision Tree Selection(BINGO BONUS #2). For our fourth model, we leverage functionality within SAS Enterprise to create a decision tree to identify variable importance and then use SAS Studio to run a regression model incorporating those variables selected. Table 9 below provides the top 10 most important variables identified by SAS Enterprise Miner.

Table 9: Variable Importance

Variable Name	Label	Splitting Rules	Importance
CLM_FREQ	#Claims(Past 5 Years)	1	1.0000
IMP_JOB	Imputed: Job Category	2	0.8913
URBANICITY	Home/Work Area	2	0.7385
KIDSDRIV	#Driving Children	2	0.5209
MVR_PTS	Motor Vehicle Record Points	3	0.5031
REVOKED	License Revoked (Past 7 Years)	4	0.4287
IMP_HOME_VAL	Imputed: Home Value	1	0.4025
MSTATUS	Marital Status	2	0.3615
CAR_TYPE	Type of Car	2	0.3604
OLDCLAIM	Total Claims(Past 5 Years)	1	0.2977
TRAVTIME	Distance to Work	2	0.2795
BLUEBOOK	Value of Vehicle	1	0.2360

Using the SAS logistic regression procedure, we generate the following results (Table 10):

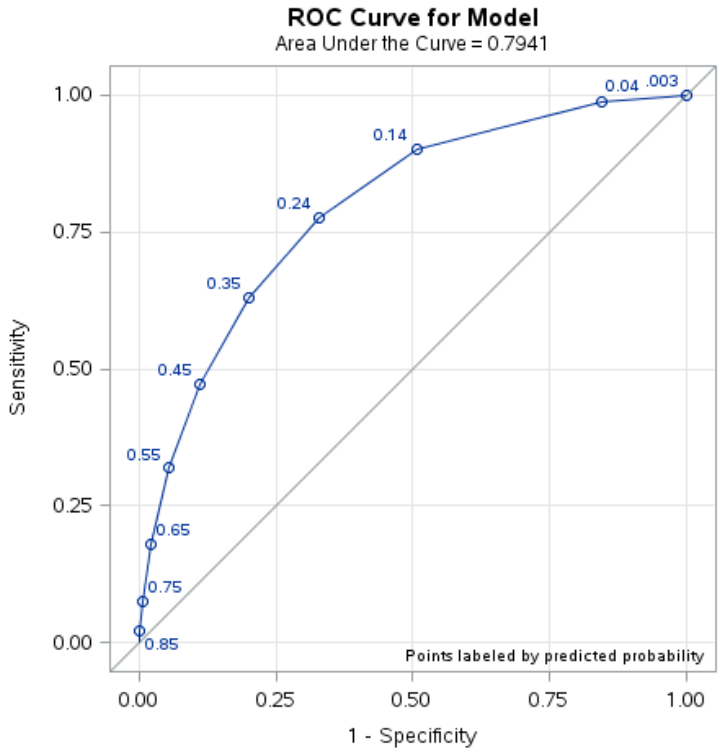
Table 10: Analysis of Maximum Likelihood Estimates for Model 4

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		-1.6555	0.1766	87.9225	<.0001
CLM_FREQ		0.2024	0.0281	51.9978	<.0001
IMP_JOB	Clerical	-0.1364	0.09	2.2958	0.1297
IMP_JOB	Doctor	-1.2087	0.1603	56.8658	<.0001
IMP_JOB	Home Maker	-0.336	0.1187	8.0088	0.0047
IMP_JOB	Lawyer	-0.8813	0.1026	73.8368	<.0001
IMP_JOB	Manager	-1.3957	0.1116	156.4196	<.0001
IMP_JOB	Professional	-0.6689	0.0959	48.6773	<.0001
IMP_JOB	Student	0.00763	0.114	0.0045	0.9466
URBANICITY	Highly Urban/ Urban	2.2644	0.1107	418.2943	<.0001
KIDSDRIV		0.4633	0.0524	78.2837	<.0001
MVR_PTS		0.1235	0.0133	85.6155	<.0001
REVOKED	No	-0.913	0.0893	104.617	<.0001
IMP_HOME_VAL		-2.00E-06	3.31E-07	36.5512	<.0001
MSTATUS	Yes	-0.5319	0.0689	59.5712	<.0001
CAR_TYPE	Minivan	-0.7022	0.0842	69.4701	<.0001
CAR_TYPE	Panel Truck	0.3388	0.1346	6.3412	0.0118
CAR_TYPE	Pickup	0.0802	0.0846	0.8984	0.3432
CAR_TYPE	Sports Car	0.25	0.096	6.7765	0.0092
CAR_TYPE	Van	0.1545	0.1128	1.8754	0.1709
OLDCLAIM		-0.00001	3.83E-06	14.0415	0.0002
TRAVTIME		0.0135	0.00185	53.4037	<.0001
BLUEBOOK		-0.00002	4.57E-06	27.7743	<.0001

The interpretation of Model 4 is similar to Model 1 and Model 2 as it contains both continuous and categorical variables and, therefore, we elect not to repeat the exercise. Model 4 includes 7 continuous variables, 4 of the original categorical variables and 1 imputed categorical variable. All are generally in the predicted direction with a few exceptions. The coefficients for driving panel trucks, pickups and vans have all switched from negative to positive, though none are statistically significant. This raises some concern but we decide not to remove the variables at risk of overfitting the model.

We examine the ROC (Exhibit 5) and see that Model 4 is performing at 79.41% coverage, slightly lower than Model 3 but superior to Model 1 and Model 2. At this juncture, it is evident that any additional interations will likely only generate incremental improvement to the ROC.

Exhibit 5: Model 4 ROC



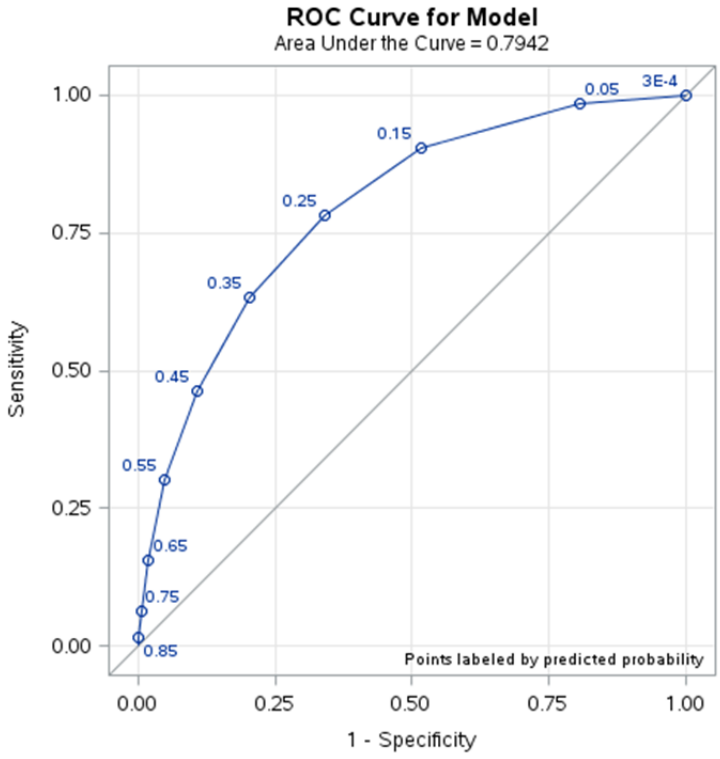
Model 5: Probit Regression Model (BINGO BONUS #3). For the fifth and final model to predict TARGET_FLAG, we present a probit regression model, a technique which is very similar to logistical regression. The difference between the two methods lies in the assumption about the distribution of errors. Logistical assumes standard logistic distribution of errors while probit assumes a normal distribution of errors. Using the SAS logistic regression procedure and link=probit option to fit a probit model, we generate the following results (Table 11):

Table 11: Analysis of Maximum Likelihood Estimates for Model 3

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		-0.9183	0.0984	87.0045	<.0001
CLM_FREQ		0.1208	0.0166	52.9692	<.0001
IMP_JOB	Clerical	-0.0813	0.0529	2.3595	0.1245
IMP_JOB	Doctor	-0.704	0.09	61.1268	<.0001
IMP_JOB	Home Maker	-0.2007	0.0696	8.3104	0.0039
IMP_JOB	Lawyer	-0.5072	0.0591	73.7581	<.0001
IMP_JOB	Manager	-0.8004	0.0623	164.8148	<.0001
IMP_JOB	Professional	-0.3846	0.0558	47.4758	<.0001
IMP_JOB	Student	0.0155	0.067	0.0534	0.8172
URBANICITY	Highly Urban/ Urban	1.2478	0.0571	477.3029	<.0001
KIDSDRIV		0.2673	0.0307	75.9002	<.0001
MVR_PTS		0.0726	0.00789	84.6108	<.0001
REVOKED	No	-0.5298	0.0525	101.6496	<.0001
IMP_HOME_VAL		-1.14E-06	1.91E-07	35.3926	<.0001
MSTATUS	Yes	-0.3115	0.04	60.7555	<.0001
CAR_TYPE	Minivan	-0.4025	0.0481	69.9554	<.0001
CAR_TYPE	Panel Truck	0.1913	0.0778	6.0512	0.0139
CAR_TYPE	Pickup	0.042	0.0495	0.7191	0.3964
CAR_TYPE	Sports Car	0.1465	0.0563	6.7765	0.0092
CAR_TYPE	Van	0.0846	0.0656	1.6623	0.1973
OLDCLAIM		-7.95E-06	2.26E-06	12.4275	0.0004
TRAVTIME		0.0077	0.00106	52.3425	<.0001
BLUEBOOK		-0.00001	2.61E-06	27.0964	<.0001

Model 5 uses the same variables as Model 4 to make for an easier comparison. The coefficient directionality is largely identical but the effects of the variables are more muted in aggregate for Model 4. We examine the ROC (Exhibit 6) and see that Model 5 is performing at 79.42% coverage, nearly identical to Model 4. At this juncture, it is evident that any additional iterations will likely only generate incremental improvement to the ROC.

Exhibit 6: Model 5 ROC



MODEL SELECTION. We have presented 5 different predictive models and now must chose the "best model." For purposes of this exercise we use a combination of quantitative and qualitive measures to assess the best model. Table 11 below provides the quantitative comparison for each of the 5 models. The main quantitative criteria we are evaluating on are the Akaike Information Criterion (AIC), Somer's D, Gamma, Area Under the Curve (AUC) and Kolmogorov-Smirnov (K-S) statistic. Our guiding rules are to minimize the AIC while the Somer's D, Gamma and AUC should be as close to 1.0 as possible. In regards to the K-S Statistic, it is a relative indicator of curve fit and if the value goes below 0.05 the lack of fit is considered significant.

Table 12: Model Comparison Statistics

Criteria	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	8482.341	7797.95	7357.059	7552.289	7559.017
Somer's D	0.419	0.552	0.613	0.588	0.588
Gamma	0.505	0.624	0.676	0.655	0.653
AUC	0.7093	0.7761	0.8065	0.7941	0.229
K-S Statistic	0.148562	0.192942	0.210695	0.197581	0.197246

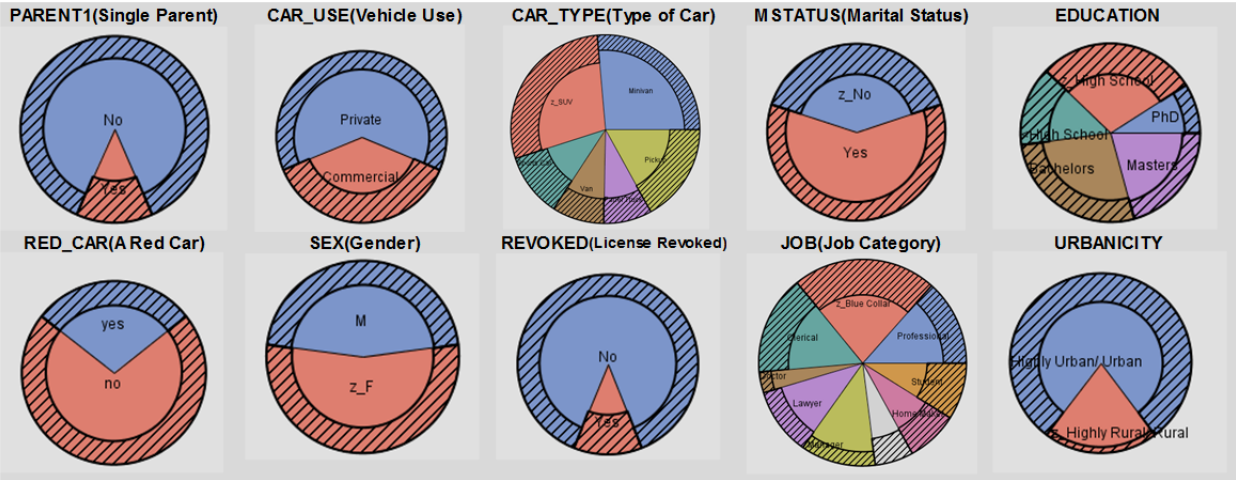
All of the models are relatively sound, however the results markedly improved as we used a confluence of continuous and categorical variables in the later iterations (Models 3-5). There remains a mild level of concern over the coefficients switching directions in Model 3 and Model 4, but most of their impacts were not statistically significant and could likely be removed from the models completely. Both Model 3 and Model 4 contain all of the variables identified in the EDA as being important. While Model 3 is statistically superior by a small margin, Model 4 is preferred as it has fewer predictors which makes its interpretation and eventual application of the results easier for business users. Model 5 is nearly identical to Model 4 and since we are more comfortable with logistic regression, Model 4 is the recommendation.

CONCLUSION

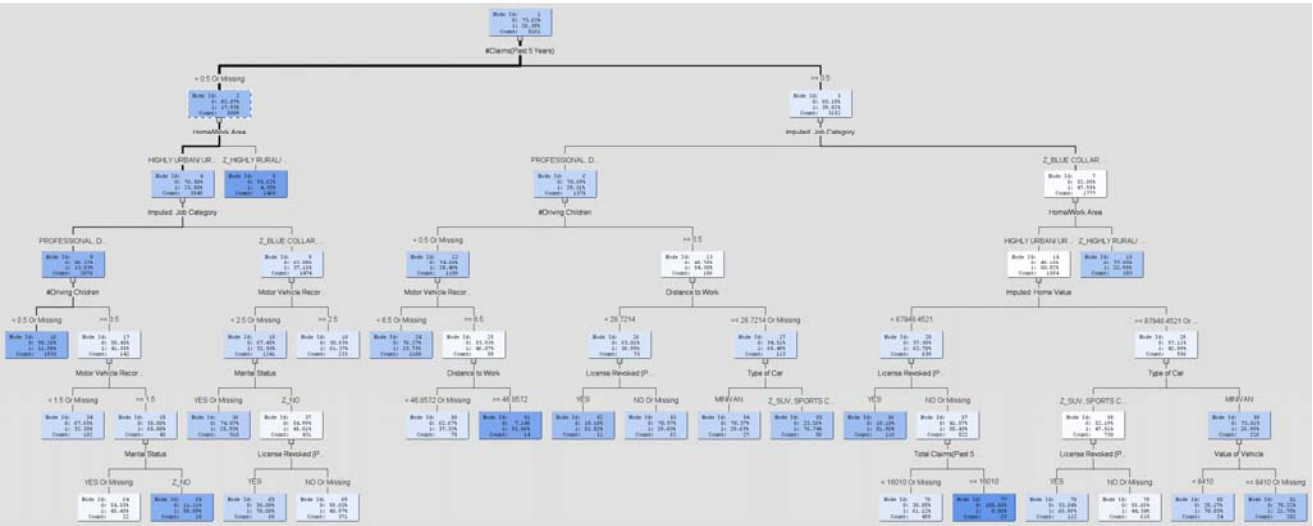
Building sound regression models is a cornerstone of predictive analytics and to creating a culture of proactive data-driven decision making. For this exercise, a historical insurance data set was used to construct logistic regression models to predict car crashes and a final model was selected using a variety of criteria. The potential value of enhanced predictive forecasting for the insurance industry is evident. There are a host of sophisticated tools in the marketplace capable of digesting a tremendous volume of information every minute and while my basic managerial recommendation is to stress simplicity, it is also important to recognize the need to make business critical decisions using imperfect information. The data set for this exercise was diverse and the selected model is not flawless, but I am comfortable with the insight it generates as it incorporates the variables identified as having predictive power and am confident in it's deployment in a production environment.

Scott Morgan
KAGGLE: ScottMorgan

BINGO BONUS #1 (10 pts?): Frequency of Car Accidents ("1") by Categorical Variable; Enterprise Miner. Part of the EDA



BINGO BONUS #2 (20 pts): Decision Tree for Variable Selection for Model 4 using SAS Enterprise Miner



BINGO BONUS #3 (5 pts): Probit regression for Model 5

Table 11: Analysis of Maximum Likelihood Estimates for Model 3

Parameter		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		-0.9183	0.0984	87.0045	<.0001
CLM_FREQ		0.1208	0.0166	52.9692	<.0001
IMP_JOB	Clerical	-0.0813	0.0529	2.3595	0.1245
IMP_JOB	Doctor	-0.704	0.09	61.1268	<.0001
IMP_JOB	Home Maker	-0.2007	0.0696	8.3104	0.0039
IMP_JOB	Lawyer	-0.5072	0.0591	73.7581	<.0001
IMP_JOB	Manager	-0.8004	0.0623	164.8148	<.0001
IMP_JOB	Professional	-0.3846	0.0558	47.4758	<.0001
IMP_JOB	Student	0.0155	0.067	0.0534	0.8172
URBANICITY	Highly Urban/ Urban	1.2478	0.0571	477.3029	<.0001
KIDSDRIV		0.2673	0.0307	75.9002	<.0001
MVR_PTS		0.0726	0.00789	84.6108	<.0001
REVOKED	No	-0.5298	0.0525	101.6496	<.0001
IMP_HOME_VAL		-1.14E-06	1.91E-07	35.3926	<.0001
MSTATUS	Yes	-0.3115	0.04	60.7555	<.0001
CAR_TYPE	Minivan	-0.4025	0.0481	69.9554	<.0001
CAR_TYPE	Panel Truck	0.1913	0.0778	6.0512	0.0139
CAR_TYPE	Pickup	0.042	0.0495	0.7191	0.3964
CAR_TYPE	Sports Car	0.1465	0.0563	6.7765	0.0092
CAR_TYPE	Van	0.0846	0.0656	1.6623	0.1973
OLDCLAIM		-7.95E-06	2.26E-06	12.4275	0.0004
TRAVTIME		0.0077	0.00106	52.3425	<.0001
BLUEBOOK		-0.00001	2.61E-06	27.0964	<.0001

BINGO BONUS #4 (10 pts): You'll notice that I used SAS Macros in my file titled Homework_02_Scott_Morgan_ANALYSIS_Code.sas. I even included the periods to look like a pro.

BINGO BONUS #5 (20 pts): I present the PROC GENMOD model below next to the Model 4 (Logistic) and Model 5 (Probit); this code is located at the end of the file titled Homework_02_Scott_Morgan_ANALYSIS_Code.sas. Each of these models use the same variables to make comparison easy.

PROC GENMOD appears to produce slightly different output (not shown). For one, PROC GENMOD includes all categorical variables were PROC LOGISTIC removes one for a basis of interpretation. This does not make a difference when using PROC GENMOD however as the parameter estimates for these variables are 0.00. The independent variables coefficients are generally in the same direction, though their influence is less pronounced. The intercept sign has also switched. By the AIC metric (below), the Logistic model still seems superior.

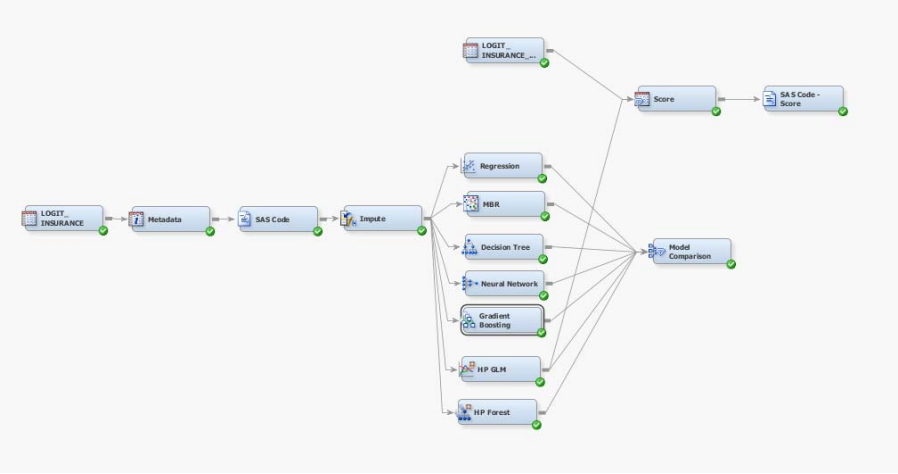
Scott Morgan
KAGGLE: ScottMorgan

PROC GENMOD Comparison

	Logistic	Probit	Generalized
AIC	7559.017	7552.289	7925.2322
K-S Statistic	0.197246	0.197581	0.194943

BINGO BONUS #6 (20 pts?): Using SAS Enterprise Miner to predict P_TARGET_AMT

The second target variable is TARGET_AMT in the data set. This value is zero if the person did not crash their car. But if the customers did crash their car, this number will be a value greater than zero. I use SAS Enterprise Miner to build a model that predicts the insurance premium to charge customers if they do crash their cars. Below is a screenshot of the diagram I created; as you can see I included several different models. The model comparison node chose the neural network but I was a bit skeptical of using this iteration given your warning about those overfitting with smaller data sets. The output also looked pretty funky as well. The model I ultimately submitted uses gradient boosting with a decision tree as the assessment measurement. I'm curious how, if I understand this correctly, an approach that uses iterative decisions trees will perform in the larger Kaggle data set. The variable importance the model identified is provided below as well.



Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	BLUEBOOK	Value of Vehicle	36	1.00000
2	IMP_HOME_VAL	Imputed: Home Value	11	0.46017
3	IMP_JOB	Imputed: Job Category	9	0.45502
4	MVR_PTS	Motor Vehicle Record Points	9	0.39719
5	TRAVTIME	Distance to Work	8	0.37737
6	IMP_CAR_AGE	Imputed: Vehicle Age	7	0.34387
7	MSTATUS	Marital Status	4	0.27483
8	CAR_TYPE	Type of Car	5	0.27295
9	OLDCLAIM	Total Claims(Past 5 Years)	3	0.26760
10	IMP_AGE	Imputed: Age	4	0.26687
11	EDUCATION	Max Education Level	4	0.26173
12	IMP_INCOME	Imputed: Income	4	0.23750
13	REVOKED	License Revoked (Past 7 Years)	3	0.20773
14	IMP_YOJ	Imputed: Years on Job	2	0.18590
15	SEX	Gender	1	0.17437
16	PARENT1	Single Parent	1	0.15629
17	HOMEKIDS	#Children @Home	1	0.14234

Scott Morgan
KAGGLE: ScottMorgan

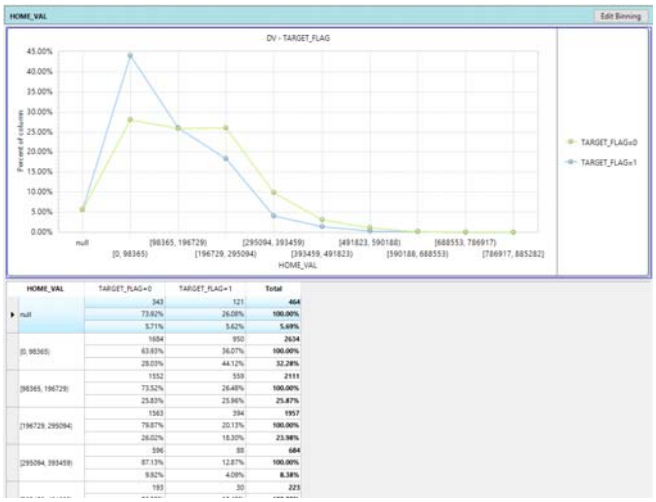
BINGO BONUS #7 (+20 pts): Using Angoss to make a Decision tree

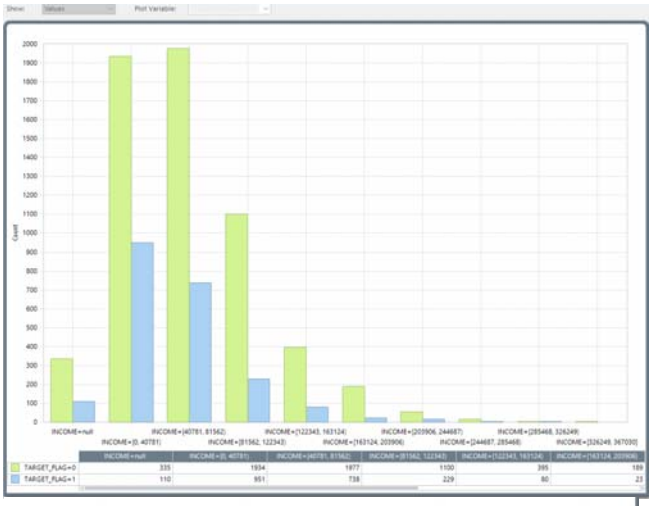
Angoss is a great tool; especially its ability to perform on-screen EDA. Below are a below examples:

Getting a view of the raw data set

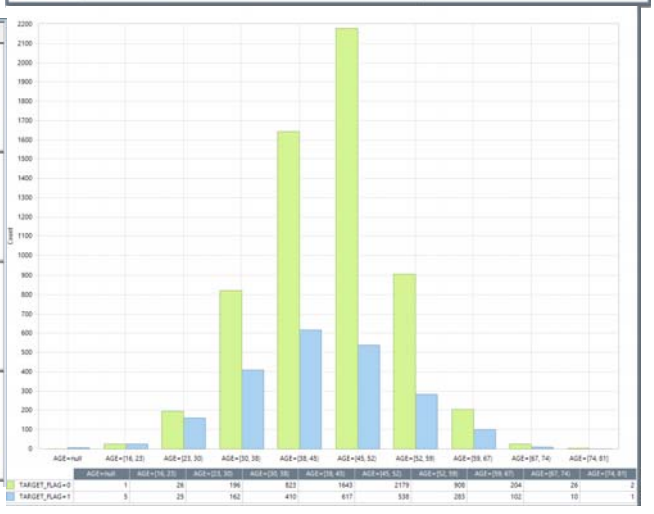
#	Field Name	Field Label	Data Type	Cardinality	Unique Count	# of Missing Values	% of Missing Value	Minimum	Maximum	Mean	Standard Deviation
1	TARGET_FLAG	TARGET_FLAG	Number	2	0	0	0.00 %	0.00	1.00	0.26	0.44
2	KIDSDRIV	KIDSDRIV	Number	5	0	0	0.00 %	0.00	4.00	0.17	0.51
4	HOMEKIDS	HOMEKIDS	Number	6	0	0	0.00 %	0.00	5.00	0.72	1.12
7	PARENT1	PARENT1	String	2	0	0	0.00 %	No	Yes		
9	MSTATUS	MSTATUS	String	2	0	0	0.00 %	Yes	z_No		
10	SEX	SEX	String	2	0	0	0.00 %	M	z_F		
11	EDUCATION	EDUCATION	String	5	0	0	0.00 %	<High School	z_High School		
13	TRAVTIME	TRAVTIME	Number	97	10	0	0.00 %	5.00	142.00	33.49	15.91
14	CAR_USE	CAR_USE	String	2	0	0	0.00 %	Commercial	Private		
15	BLUEBOOK	BLUEBOOK	String	2789	900	0	0.00 %	$1,500	$9,990		
16	TIF	TIF	Number	23	0	0	0.00 %	1.00	25.00	5.35	4.15
17	CAR_TYPE	CAR_TYPE	String	6	0	0	0.00 %	Minivan	z_SUV		
18	RED_CAR	RED_CAR	String	2	0	0	0.00 %	no	yes		
19	OLDCLAIM	OLDCLAIM	String	2857	2592	0	0.00 %	$0	$9999		
20	CLM_FREQ	CLM_FREQ	Number	6	0	0	0.00 %	0.00	5.00	0.80	1.16
21	REVOKED	REVOKED	String	2	0	0	0.00 %	No	Yes		
22	MVR_PTS	MVR_PTS	Number	13	0	0	0.00 %	0.00	13.00	1.70	2.15
24	URBANICITY	URBANICITY	String	2	0	0	0.00 %	Highly Urban/ Urban	z_Highly Rural/ Rural		
3	AGE	AGE	Number	61	4	6	0.07 %	16.00	81.00	44.79	8.63
6	INCOME	INCOME	Number	6613	6136	445	5.43 %	0.00	367,030.00	61,898.09	47,572.68
5	YOJ	YOJ	Number	22	0	454	5.56 %	0.00	23.00	10.50	4.09
8	HOME_VAL	HOME_VAL	Number	5107	4819	464	5.69 %	0.00	885,282.00	154,867.29	129,123.77
23	CAR_AGE	CAR_AGE	Number	31	3	510	6.25 %	-3.00	28.00	8.33	5.70
12	JOB	JOB	String	9	0	526	6.45 %	Clerical	z_Blue Collar		

Other EDA tools





Row Number	Field Name	Information Value	Entropy Explained	Whole Dataset	TARGET_FLAG=0	TARGET_FLAG=1
1	INCOME	0.12457	0.02045			
2	PARENT1	0.11349	0.0197			
3	HOMEKIDS	0.08381	0.01421			
4	HOME_VAL	0.18841	0.03055			



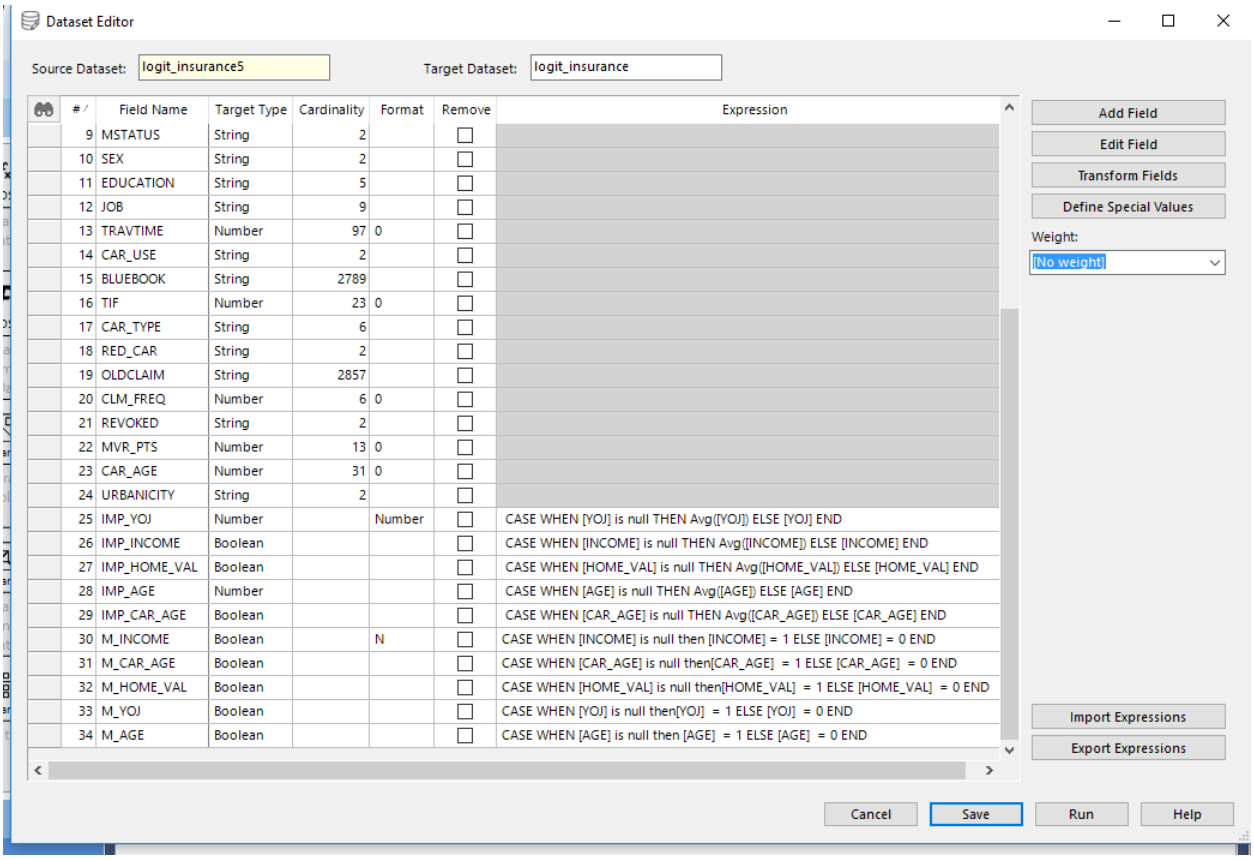
Scott Morgan
KAGGLE: ScottMorgan

Correlation Matrix

	TARGET_FLAG	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	TIF	CLM_FREQ	MVR_PTS	CAR_AGE
TARGET_FLAG	1	0.10367	-0.10322	0.11562	-0.07051	-0.14201	-0.18374	0.04837	-0.08237	0.2162	0.2192	-0.10065
KIDSDRIV	0.10367	1	-0.07518	0.46402	0.0433	-0.04713	-0.01979	0.00845	-0.00199	0.03706	0.05357	-0.05399
AGE	-0.10322	-0.07518	1	-0.44544	0.13607	0.18097	0.20998	0.00527	-7E-05	-0.02409	-0.07158	0.17622
HOMEKIDS	0.11562	0.46402	-0.44544	1	0.08683	-0.15933	-0.11068	-0.00725	0.01181	0.02935	0.0606	-0.15215
YOJ	-0.07051	0.0433	0.13607	0.08683	1	0.28607	0.26992	-0.01695	0.02479	-0.02631	-0.03786	0.06141
INCOME	-0.14201	-0.04713	0.18097	-0.15933	0.28607	1	0.57524	-0.04708	-0.00103	-0.04775	-0.06316	0.41424
HOME_VAL	-0.18374	-0.01979	0.20998	-0.11068	0.26992	0.57524	1	-0.03553	0.00206	-0.09405	-0.08539	0.21747
TRAVTIME	0.04837	0.00845	0.00527	-0.00725	-0.01695	-0.04708	-0.03553	1	-0.0116	0.00656	0.0106	-0.03823
TIF	-0.08237	-0.00199	-7E-05	0.01181	0.02479	-0.00103	0.00206	-0.0116	1	-0.02302	-0.04105	0.00777
CLM_FREQ	0.2162	0.03706	-0.02409	0.02935	-0.02631	-0.04775	-0.09405	0.00656	-0.02302	1	0.39664	-0.00932
MVR_PTS	0.2192	0.05357	-0.07158	0.0606	-0.03786	-0.06316	-0.08539	0.0106	-0.04105	0.39664	1	-0.0199
CAR_AGE	-0.10065	-0.05399	0.17622	-0.15215	0.06141	0.41424	0.21747	-0.03823	0.00777	-0.00932	-0.0199	1

Data Cleaning

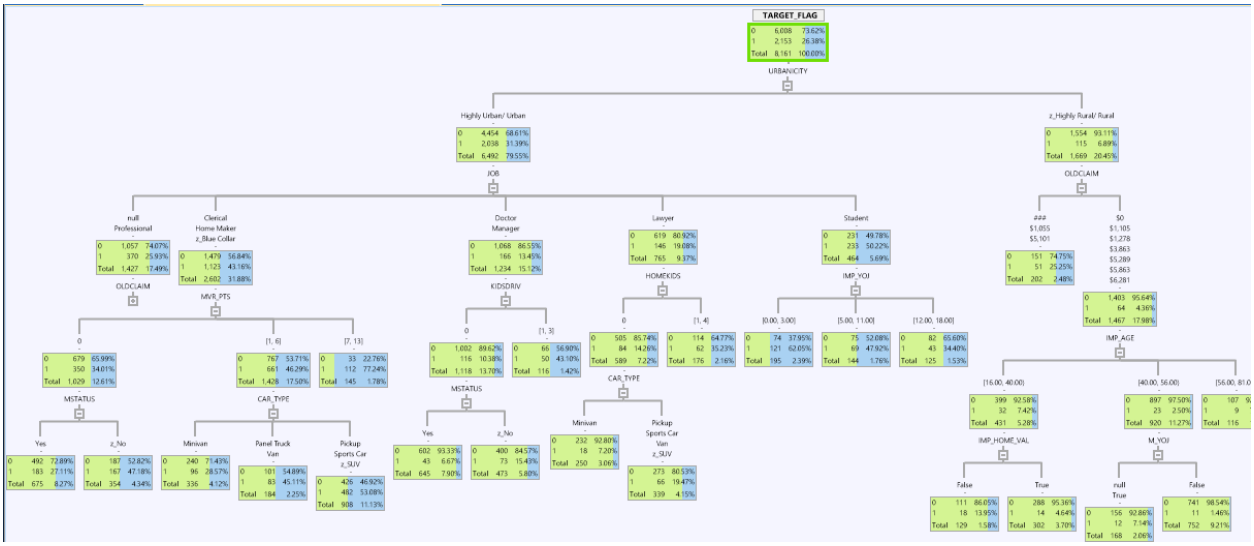
This was not the easiest as I had to do a lot of the cleaning by hand. I may have missed the functionality to do this in a more automated fashion like SAS Enterprise Miner.



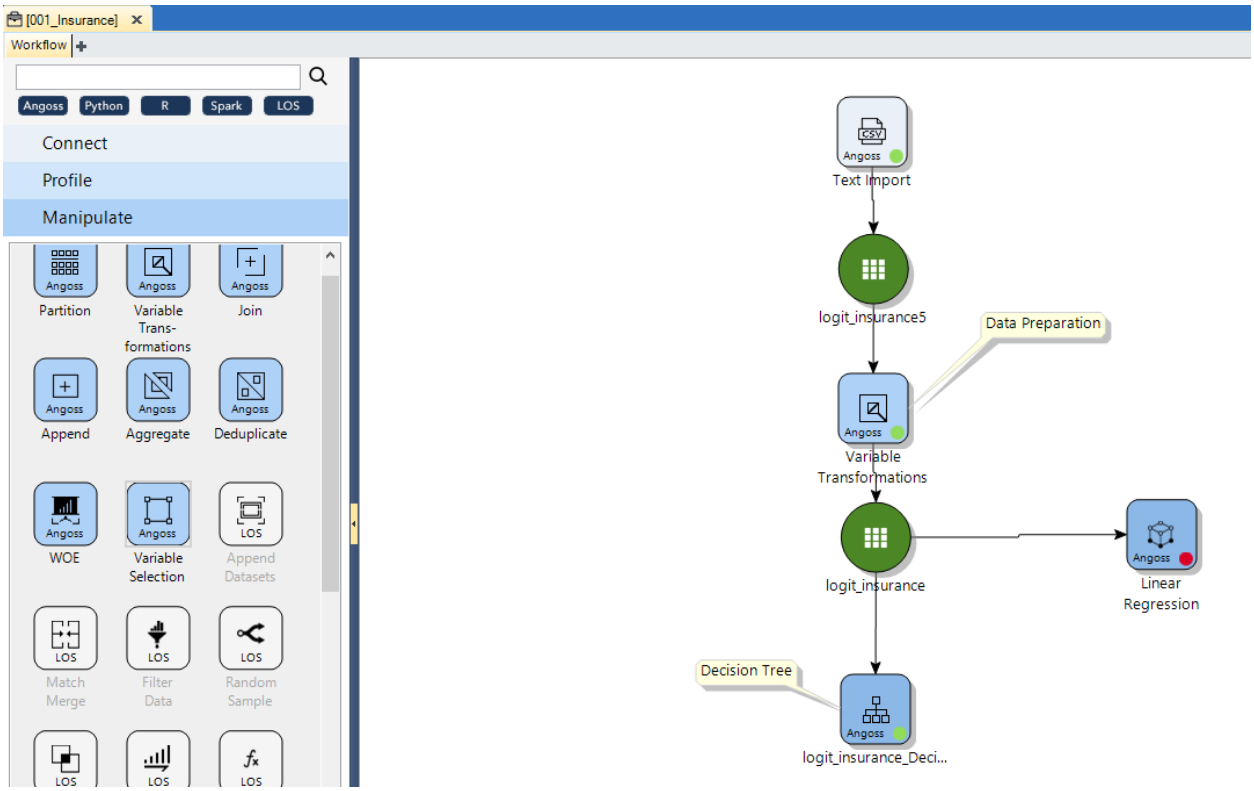
#	Field Name	Target Type	Cardinality	Format	Remove	Expression
9	MSTATUS	String	2			
10	SEX	String	2			
11	EDUCATION	String	5			
12	JOB	String	9			
13	TRAVTIME	Number	97	0		
14	CAR_USE	String	2			
15	BLUEBOOK	String	2789			
16	TIF	Number	23	0		
17	CAR_TYPE	String	6			
18	RED_CAR	String	2			
19	OLDCLAIM	String	2857			
20	CLM_FREQ	Number	6	0		
21	REVOKED	String	2			
22	MVR_PTS	Number	13	0		
23	CAR_AGE	Number	31	0		
24	URBANICITY	String	2			
25	IMP_YOJ	Number		Number		CASE WHEN [YOJ] is null THEN Avg([YOJ]) ELSE [YOJ] END
26	IMP_INCOME	Boolean				CASE WHEN [INCOME] is null THEN Avg([INCOME]) ELSE [INCOME] END
27	IMP_HOME_VAL	Boolean				CASE WHEN [HOME_VAL] is null THEN Avg([HOME_VAL]) ELSE [HOME_VAL] END
28	IMP_AGE	Number				CASE WHEN [AGE] is null THEN Avg([AGE]) ELSE [AGE] END
29	IMP_CAR_AGE	Boolean				CASE WHEN [CAR_AGE] is null THEN Avg([CAR_AGE]) ELSE [CAR_AGE] END
30	M_INCOME	Boolean		N		CASE WHEN [INCOME] is null then [INCOME] = 1 ELSE [INCOME] = 0 END
31	M_CAR_AGE	Boolean				CASE WHEN [CAR_AGE] is null then[CAR_AGE] = 1 ELSE [CAR_AGE] = 0 END
32	M_HOME_VAL	Boolean				CASE WHEN [HOME_VAL] is null then[HOME_VAL] = 1 ELSE [HOME_VAL] = 0 END
33	M_YOJ	Boolean				CASE WHEN [YOJ] is null then[YOJ] = 1 ELSE [YOJ] = 0 END
34	M_AGE	Boolean				CASE WHEN [AGE] is null then [AGE] = 1 ELSE [AGE] = 0 END

Scott Morgan
KAGGLE: ScottMorgan

Decision Tree

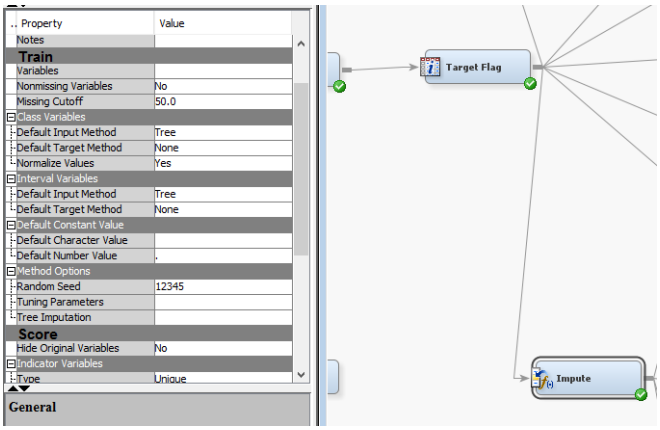


Complete Workflow



Scott Morgan
KAGGLE: ScottMorgan

BINGO BONUS #8 (+10 pts?): Last but not least, I used a decision tree in Enterprise Miner for variable selection and ran a regression model. I include screen shot below and also a supplementary scoring code program for you to run. I thought it was useful to run my models through the manual data cleaning code for good practice but I really wanted to include something from Enterprise Miner because it's such a power tool. Great video.



REFERENCE(S)

Schwert, G.W. 2000. Hostility in takeovers: In the eyes of the beholder, Journal of Finance 55: 2599–2640.