FROM:	John Hokkanen, 411-Section 60
TO:	Dr. Joseph Wilck
DATE:	October 21, 2016
PROJECT:	Create, analyze, select and document a logistic regression prediction algorithm to estimate insurance claims and provide a list of estimates and claim amounts.

EXECUTIVE SUMMARY

The following document explains the methodology used to evaluate the insurance dataset and prepare a prediction method. The supplied data set has a handful of variables that have missing data values and one variable that appears to have mixed information. Some variables were continuous or count variables and others were categorical variables. The categorical variables were converted to binary flags for current or future use in the mode. Some continuous variables were discretized where segments of the continuity had different linear relationships to the insurance claim response, and then those segments had binary flags created for modeling use. A variety of matrixed binary flags were created to explore possible predictive value.

Analysis confirmed that the data set has a substantial amount of noise. About 75% of the data is easily modeled, and about 15-20% appears to be discordant with the majority of the data. The challenge is to create model criteria which will correctly segregate and categorize this discordant data. In general the easily modeled data has approximately 25% of targets as claims that are correctly identified. For at least some of this discordant data with claims, they appear to have much in common with the no-claim data as though someone randomly changed the target claim variable from a zero to a one.

Multiple techniques were attempted to develop a model to handle this troublesome data, and, for the most part, those techniques failed to develop any breakthrough model. However, the examination of those results did lead to a better model. The final model was validated with a segregated section of training data set aside for validation. The model performed similarly, if not better, than it did with its training data, and the analyst believes that given similar data, the model will have similar performance.

DATA EXPLORATION

Missing Values

To identify missing values among numerical variables, proc means generated counts:

Variable	Label	N Miss	Action Taken
AGE	**Age**	6	Use median age
YOJ	**Years on Job**	403	Years on job was found to be highly correlated with Age. Young people had few years, with older people having much more. However, unlike age, there was considerable variance because people might change jobs after many years working at a job.

			Because of the collinearity question with Age, the decision was made not to invest a great deal of time on this variable and simply set missing values to the mean of 10.49.
INCOME	**Income**	406	In correlations this variable appeared to be a high-value variable. Therefore, effort was taken to impute missing values to maximize its use and minimize negative effects from using mean or median. With extensive knowledge home loans, the very strong relationship of salary to home values was already known. Consequently, home values were broken into categories: $50K-100K, $100-200K, $200-300K, $400-500K and >$500K, and mean income values were identified for these homeowner values and assigned to the missing income values. For people with homes valued at 0 or missing home values, the missing income values were assigned the mean for the person's job category; if the job category was nevertheless missing (which at this point contained only Masters and PhD persons), the mean for this education category was assigned. Finally, for non-student, non-homemaker persons whose income was set at 0 (see below), the mean value for the job category as assigned. This process resulted in a highly relevant variable with no missing values.
HOME_VAL	**Home Value**	419	Because Home Value was less significant as a variable, the decision to use it to impute income was made; to avoid additional collinearity between these two variables, missing Home Value variables were assigned the median home value after the income imputation occurred.
CAR_AGE	**Vehicle Age**	472	Initially, this variable was of interest, but exploration showed it to be problematic. In many cases, the values appear to be actual vehicle age; this could be confirmed by comparing Bluebook, individual Age and TimeInForce values. However, many values did not make sense, revealing that many values represented years of ownership. Of course, these two items are very different which raised concern about a model based on them. There was concern that because it might be impossible to know whether it was car age or years of ownership, the relationship of this combined variable could be the result of chance. Instead of spending considerable time imputing the variable, a binary flag representing the fact this item was missing and the variable was set to zero instead. As it turned out, the regression of the continuous variable was not used.

In addition, there were a few individuals who were not students and homemakers whose income was zero:

Variable	$0, excluding Students and Homemakers
INCOME	**15**

Because there was a small number and the $0 value was troubling, these 15 records were treated as missing values. The students and homemakers with zero income was left at $0.

For categorical variables, inspection of the data with an excel sheet revealed the following:

Variable	N Blanks
JOB	**488**

Because the categorical values for each job would be converted to binary flags for the job that they represent, a flag of F_JOB_UNKNOWN was created and populated.

Checking Variable Distributions/Scaling for Ease in Understanding

Though the distribution of the dependent variable is critical for model selection, the GLM models do not make assumptions about the independent variables. Nevertheless, several of the continuous variables (income, home value) were transformed by taking the square root to reduce the long tails to the right. However, when these transformed variables were tested in regressions, the area under the prediction curve went down. This is probably related to the fact that differing segments of the distribution have differing relationships to the target variable. Consequently, rather than try to identify the best transformation, the focus shifted to creating appropriate discretization of the continuous variables so that key segments could be adjusted mathematically and creating flag variables if one of these items was in the 99th percentile. Not transforming the data also had the benefit of easier interpretation of the coefficients, and, in fact, for income, bluebook value, and home value, the variables were scaled down to make the coefficients easier to understand. For example, income became income_imputed_100K (i.e., the imputed income had been divided by 100,000, and bluebook value become bluebook_10K (i.e., that bluebook value was divided by 10,000).

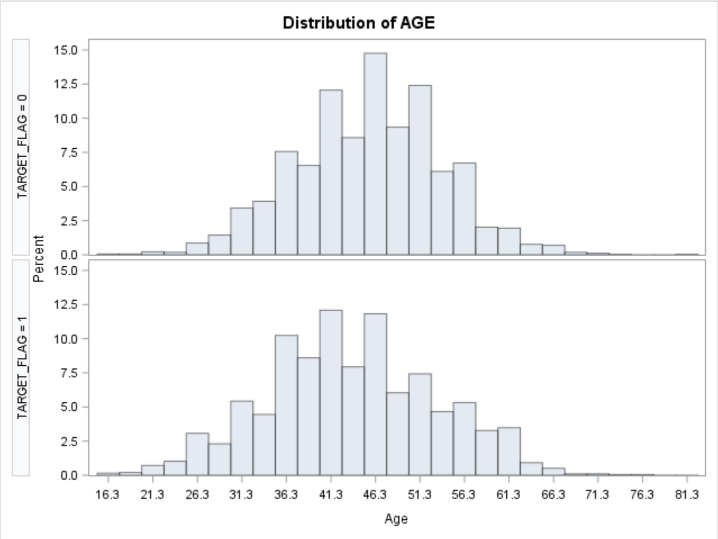
Correlations of Raw Numerical Variables

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

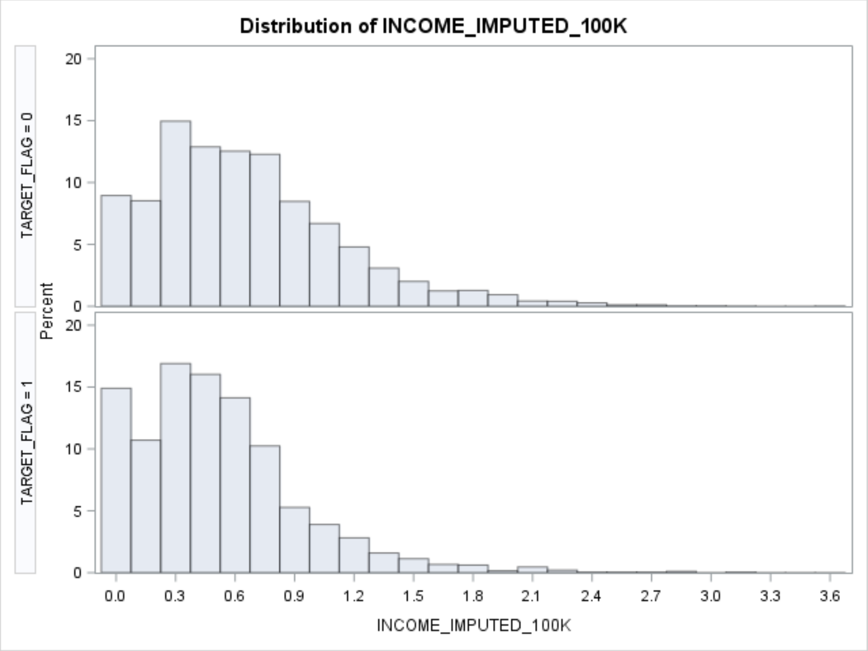
VARIABLE	TARGET_FLAG	TARGET_AMT	NOTES
KIDSDRIV	0.10331	0.0546	Driving kids affect claims in an upward direction
#Driving Children	<.0001	<.0001	
AGE	**-0.09884**	-0.04542	
Age	<.0001	<.0001	
HOMEKIDS	0.11107	0.06237	Kids at home also affect claims
#Children @Home	<.0001	<.0001	
YOJ	-0.06656	-0.0199	
Years on Job	<.0001	0.0951	

INCOME	-0.13849	-0.06228	Significantly related
Income	<.0001	<.0001	
HOME_VAL	-0.17902	-0.08715	Significantly related
Home Value	<.0001	<.0001	
TRAVTIME	0.04389	0.0246	
Distance to Work	0.0002	0.0338	
BLUEBOOK	-0.10212	-0.00716	
Value of Vehicle	<.0001	0.5366	
TIF	-0.08014	-0.04733	
Time in Force	<.0001	<.0001	
OLDCLAIM	0.13332	0.07375	
Total Claims(Past 5 Years)	<.0001	<.0001	
CLM_FREQ	0.20488	0.11366	2nd highest correlation – people who make claims are prone to making claims.
#Claims(Past 5 Years)	<.0001	<.0001	
MVR_PTS	0.21251	0.13712	Highest correlation
Motor Vehicle Record Points	<.0001	<.0001	
CAR_AGE	-0.09687	-0.05557	
Vehicle Age	<.0001	<.0001	

All of the variables showed up as significant, with a few being particularly important. To further the analysis, comparative histograms were used to look at the distributions of the numerical variables where target=0 and target=1. For example, for age (shown below), the target=0 histogram shows what one might expect for the general population, but the claims target=1 shows that youth does relate to claims.

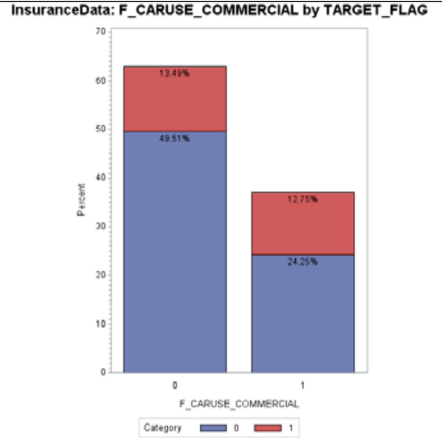


As will be shown later, the age variable is a bit more complicated when factoring in other variables, but the histograms create an easy visual for identifying relationships. For income, the relationship is a little less distinct, and yet one can see that higher incomes have a much higher no-claim to claim ratio, and the zero income is the reverse.

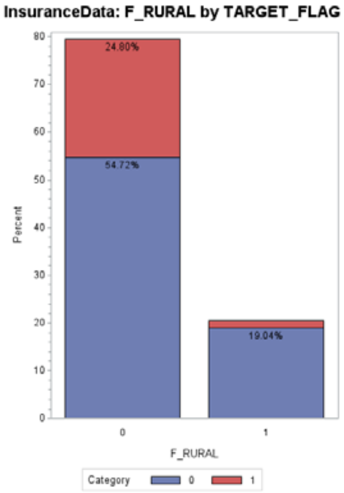


The discussion of the remaining histograms will be limited as they were used for an introduction to the data. More detailed analysis was conducted via cross-tabulations as discussed later.

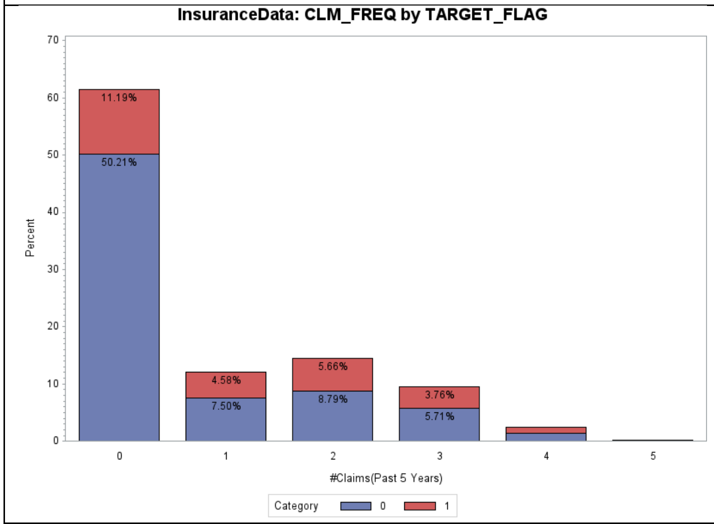
To assess the value of the categorical variables, several strategies were used. Some categoricals were binary values (e.g., Rural/Urban) while others had multiple categories (e.g., Job). All of them could be viewed with comparative barcharts that simultaneously allowed one to visually compare categories against relative portions of the claim target. Here are some examples with the kind of analysis that was performed:

	One easily sees from this comparative bar chart that the ratio of claim to no claim (relative sizes of red to blue) is much smaller for private use than commercial use. Though the red portion is approximately the same size, the blue no-claim portion is about half for commercial vehicles, meaning that there are twice as many claims in the commercial vehicle context. This makes sense because

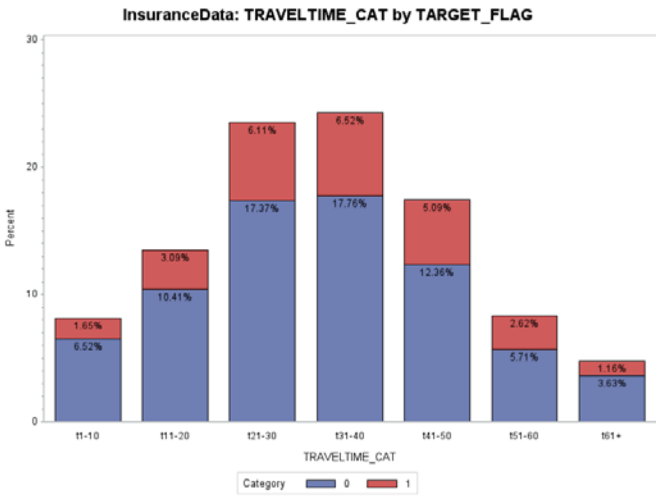
commercial vehicles are driven continuously and during the busiest times of the day.



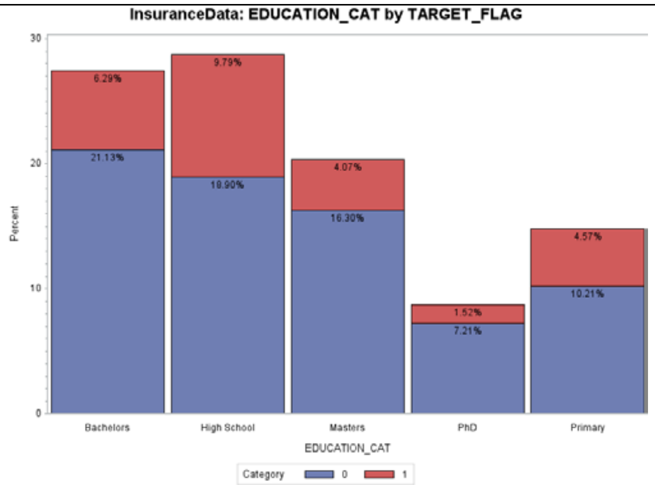
The most profound difference which is clearly visible in this chart is the difference between rural and urban driving. For rural (right side bar), the odds of a claim to no claim are a fraction of the odds for urban drivers. In fact, as would be discovered in the cross tabulations, most other categories that had relevance (e.g., age) disappeared from significance in the rural context. Simply driving where there are fewer cars meant that drivers who had high risk were much lower risk. This makes a lot of intuitive sense as most accidents are dual-car accidents rather than single-car accidents like driving into a fixed object.



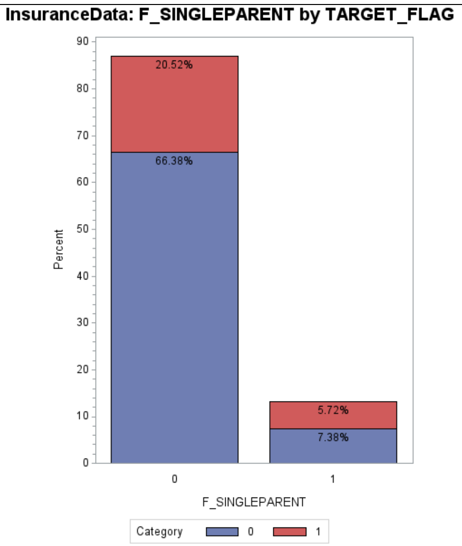
The claim frequency bar charts confirmed the much lower odds of having a claim for zero claims. It also shows that the odds of having a claim do not radically differ if one has one or three claims. The primary fact of importance is the existence of a prior claim.



Comparative bar charts were also created for discretized (i.e., binned) continuous variables like travel time. One can easily see that the odds of a claim are very similar whether one is travelling 11-20 minutes (3/10) or 31-40 minutes (6.5/18). However, two categories are somewhat different: the 1-10 minute has smaller odds as does the 61+ category. This makes sense because if you short distances mean less opportunity for an accident, and longer travel times suggest high-speed highway travel.



Education was an interesting variable. It showed that one category, PhD, was notably low in claims. Another category, Masters, was second. High school education was higher than mere primary education, but this was a general graph and did not control for rural individuals who probably have a higher number of primary-only education.



Being a single parent raises the odds of a claim from 20/66 to 5.7/7.4. This makes sense because single parents may do more driving as the driving is not shared and they may have greater distractions inside the car more of the time than people who are not single parents.

The prior diagrams illustrate the value for high-level analysis. Many more diagrams were reviewed over the course of the data discovery.

Cross-Tabulations

To acquire highly detailed analysis, cross tabulation results showed what was happening by breaking down the counts into categories. Many different kinds of multi-variable cross tabulations were created so that variables could be reviewed while controlling for other variables. The most notable of these was controlling for the rural/urban distinction, and separate cross tabulations were run for each and then compared. Other categorical variables were created that combined multiple variables like Job and Sex and Age categories.

Days of review were conducted looking at the cross tabulations whose discussion is not warranted. Nevertheless, a few will be discussed to illustrate the type of examination that was performed.

For example, the in reviewing the histogram for previous claims (above) indicated that there was not a great deal of difference for claims>1. The cross tabulation below for the previous claim flag (1=any number of claims) shows the magnitude if any claim is filed. The respective row records show that the probability of having a claim rises from 18% to 39% if any number of previous claim has been filed.

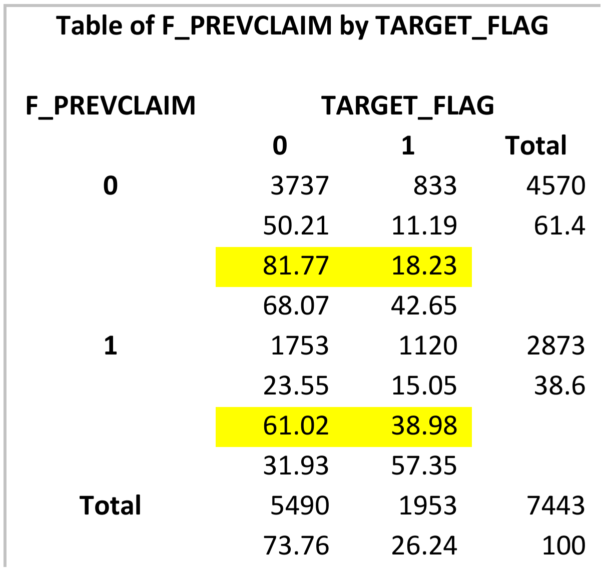


Table of F_PREVCLAIM by TARGET_FLAG

F_PREVCLAIM	TARGET_FLAG		
	0	1	Total
0	3737	833	4570
	50.21	11.19	61.4
	81.77	18.23	
	68.07	42.65	
1	1753	1120	2873
	23.55	15.05	38.6
	61.02	38.98	
	31.93	57.35	
Total	5490	1953	7443
	73.76	26.24	100

These cross tabulations can reveal insights that might otherwise be difficult to detect, and is easily illustrated with a Rural/Urban-AgeCategory cross tabulation. To begin, we should examine the impact of age category cross tab not controlling for Rural/Urban (partial cross tabulation shown):

Table of AGE_CAT by TARGET_FLAG

AGE_CAT	**TARGET_FLAG**		
	0	**1**	**Total**
a10_21	15	11	26
	0.2	0.15	0.35
	57.69	42.31	
	0.27	0.56	
a22_26	33	67	100
	0.44	0.9	1.34
	33	67	
	0.6	3.43	
a27_28	61	49	110
	0.82	0.66	1.48
	55.45	44.55	
	1.11	2.51	
a29_32	233	125	358
	3.13	1.68	4.81
	65.08	34.92	
	4.24	6.4	
a33_39	989	455	1444
	13.29	6.11	19.4
	68.49	31.51	
	18.01	23.3	
a40_45	1387	480	1867
	18.63	6.45	25.08
	74.29	25.71	
	25.26	24.58	
a46_48	832	211	1043
	11.18	2.83	14.01
	79.77	20.23	
	15.15	10.8	
Total	5490	1953	7443
	73.76	**26.24**	**100**

Data for ages above 48 were deleted to save space. The last line indicates that the general background rate for a claim is 26%. If one looks at the claims rates for the youth, there is a distinct increase for the three youngest categories. If one stopped there, one would have an insight, but would not fully understand the subtlety of the data. By looking at age, controlling for Rural/Urban, the relationship becomes much, much clearer. As is shown in the cross tab below, age is nowhere near as important in the rural location as it is in the urban locale, and it is the urban drivers that are having the accidents. In fact, for the youngest rural drivers, there is only a 15% accident rate which is much lower than the general rate. The next category has only a 20% accident rate, and the third has only a 5% accident rate, while for the urban category it exceeds 50%!

RURAL	TARGET_FLAG			URBAN	TARGET_FLAG			Diff in%
	0	1	Total		0	1	Total	
Rural_a10_21	11	2	13	Urban_a10_21	4	9	13	
	0.15	0.03	0.17		0.05	0.12	0.17	
	84.62	15.38			30.77	69.23		53.85
	0.2	0.1			0.07	0.46		
Rural_a22_26	12	3	15	Urban_a22_26	21	64	85	
	0.16	0.04	0.2		0.28	0.86	1.14	
	80	20			24.71	75.29		55.29
	0.22	0.15			0.38	3.28		
Rural_a27_28	19	1	20	Urban_a27_28	42	48	90	
	0.26	0.01	0.27		0.56	0.64	1.21	
	95	5			46.67	53.33		48.33
	0.35	0.05			0.77	2.46		
Rural_a29_32	71	10	81	Urban_a29_32	162	115	277	
	0.95	0.13	1.09		2.18	1.55	3.72	
	87.65	12.35			58.48	41.52		29.17
	1.29	0.51			2.95	5.89		
Rural_a33_39	320	34	354	Urban_a33_39	669	421	1090	
	4.3	0.46	4.76		8.99	5.66	14.64	
	90.4	9.6			61.38	38.62		29.02
	5.83	1.74			12.19	21.56		
Rural_a40_45	388	20	408	Urban_a40_45	999	460	1459	
	5.21	0.27	5.48		13.42	6.18	19.6	
	95.1	4.9			68.47	31.53		26.63
	7.07	1.02			18.2	23.55		
Rural_a46_48	184	7	191	Urban_a46_48	648	204	852	
	2.47	0.09	2.57		8.71	2.74	11.45	
	96.34	3.66			76.06	23.94		20.28
	3.35	0.36			11.8	10.45		
Rural_a49_56	326	15	341	Urban_a49_56	1198	329	1527	
	4.38	0.2	4.58		16.1	4.42	20.52	
	95.6	4.4			78.45	21.55		17.15

	5.94	0.77			21.82	16.85		
Rural_a57_62	70	13	83	**Urban_a57_62**	246	164	410	
	0.94	0.17	1.12		3.31	2.2	5.51	
	84.34	15.66			60	40		24.34
	1.28	0.67			4.48	8.4		
Rural_a63_99	16	2	18	**Urban_a63_99**	84	32	116	
	0.21	0.03	0.24		1.13	0.43	1.56	
	88.89	11.11			72.41	27.59		16.48
	0.29	0.1			1.53	1.64		

	0	**1**	**Total**
Total for both	5490	1953	7443
categories	73.76	26.24	100

This cross tabulation tells us that without the information about rural/urban, the age information will become very noisy because the fact of being in a rural location is far more important than one's age.

Looking at the existence of children is also insightful.

Table of HOMEKIDS by TARGET_FLAG

HOMEKIDS(#Children @Home)	**TARGET_FLAG**		
	0	**1**	**Total**
0	3750	1069	4819
	50.38	14.36	64.75
	77.82	22.18	
	68.31	54.74	
1	542	275	817
	7.28	3.69	10.98
	66.34	33.66	
	9.87	14.08	
2	677	342	1019
	9.1	4.59	13.69
	66.44	33.56	
	12.33	17.51	
3	415	210	625
	5.58	2.82	8.4
	66.4	33.6	
	7.56	10.75	
4	100	52	152
	1.34	0.7	2.04
	65.79	34.21	

	1.82	2.66	
5	6	5	11
	0.08	0.07	0.15
	54.55	45.45	
	0.11	0.26	
Total	5490	1953	7443
	73.76	26.24	100

As one can see, having kids at home increases the base probability of having a claim from 22% to (approximately) 33%. But we also know that married people are negatively correlated with claims, and married people often have children. Consequently, attention focuses on single parentage and kids who drive:

Table of F_SINGLEPARENT by TARGET_FLAG

F_SINGLEPARENT	**TARGET_FLAG**		
	0	**1**	**Total**
0	4941	1527	6468
	66.38	20.52	86.9
	76.39	23.61	
	90	78.19	
1	549	426	975
	7.38	5.72	13.1
	56.31	43.69	
	10	21.81	
Total	5490	1953	7443
	73.76	26.24	100

The cross tabulation reveals that being a single parent raises the probability of a claim from 23% to 43%! This suggests that the real issue with young kids may occur for single parents more than married persons.

Table of KIDSDRIV by TARGET_FLAG

KIDSDRIV(#Driving	**TARGET_FLAG**		
Children)	**0**	**1**	**Total**
0	4938	1608	6546
	66.34	21.6	87.95
	75.44	24.56	

	89.95	82.33	
1	362	212	574
	4.86	2.85	7.71
	63.07	36.93	
	6.59	10.86	
2	160	101	261
	2.15	1.36	3.51
	61.3	38.7	
	2.91	5.17	
3	28	30	58
	0.38	0.4	0.78
	48.28	51.72	
	0.51	1.54	
4	2	2	4
	0.03	0.03	0.05
	50	50	
	0.04	0.1	
Total	5490	1953	7443
	73.76	26.24	100

The driving kids tabulation shows that one or two driving kids jumps the claims rate from 25% to 37% and to 50% if one has three or more. Thus, we might infer that we need to identify where there are youthful urban drivers and single-parent drivers (who may be distracted or drive more than married parents) rather than focus on the fact of children.

Additional insights were teased out over tens of pages of cross tabulation examination, but the foregoing tables illustrate the process by which this data was analyzed. By segmenting the data and comparing the claims rates and controlling for variables, one can develop a reasonable mental model of what is happening in the driving public.

<u>INITIAL MODELING EFFORT</u>

The impact of the rural/urban difference as evidenced by the cross tabulations was compelling. Consequently, the decision was made to break the data along this fracture, and then try to break it into the most relevant couple of pieces for both of the rural and urban categories. The SAS code is included though it is now commented out because it completely failed to produce any improvement on the general model. In fact, for unknown reasons (which may be related to the sample size reduction for each group), the c values went down for the segmented regressions.

<u>SECOND MODELING EFFORT</u>

Once committed to modeling based on the entire sample, strategies then shifted to modeling a complex set of matrix variables by creating binary variables to implement the complex cross tabulations that been performed. In short, the thought was to let the automated intelligence determine what was the most relevant by providing it with an abundance of cross-related information (e.g., Age_Sex_Job, or MarriageStatus_Education_Job). These efforts also failed to produce any variables that were deemed important.

As part of that process, two types of variables were discovered to be relevant. The first was the supplementation of continuous variables with discretized bin flags. This allowed the regression algorithms to obtain the benefit of a finely tuned slope based on all the continuous data, but to adjust it for any specific segment with a discretized bin flag. Thus, while the general continuous variable Motor Vehicle Points was important, a kicker for a flag of Motor Vehicle Points > 7 could enhance the regression.

The second type of combination variable was a simple two dimensional matrix for variables that might have middling value by themselves. Combined, they could establish an important signal. For example the Job categorical variable could be combined with a discretized Age bin to yield a significant binary bin flag. For example, lawyers and doctors between the ages of 30 and 55 yielded flag variables with potential significance, but most others did not.

In addition to the above, new kinds of variables were created and tested to identify whether more general information offered a better signal. For example, HOMEKIDS is a counter that indicates how many kids are at home. But the binary flag of F_KIDSATHOME which indicates if HOMEKIDS>0 may be a more useful indicator. Likewise, F_PREVCLAIM seeks only to indicate whether there has been any claim filed rather than the count as in CLM_FREQ, and the former may also offer a better signal. The same is true for F_KIDSDRIVING which is a binary flag for KIDSDRIV>0.

More specific variables were also tested. For example, NONDRIVINGKIDS is a counter to indicate how many children are non-drivers as opposed to the driving kids (i.e,. KIDSDRIV). These more specific variables did not necessarily end up having any value.

<u>THIRD MODELING EFFORT</u>

Having settled on a strategy of a single, well-formed dataset with straightforward logistic regressions, the goal then moved to trying to develop multiple differing models and then averaging the two results. In this instance, I segregated the predictor variables into two major buckets – personal information and automobile/driving/claim information. From the outset, I realized that the claim information would likely yield a more predictive model. The hope was that by excluding all of that information from the personal data regression, I might tease out additional signal from those elements if I ran them separately. Since the two regressions were coming at the problem from very different perspectives, it was hoped that the triangulation of the two models would provide a better overall rank sequencing of the dataset.

Unfortunately, this effort also proved to not bear results upon which I was willing to submit for the final model. First, with less information, each model was understandably less predictive. Nevertheless, the combined model did eventually outperform the single regression, but only in the last 10% of the ranking. The concern was that for the models with less information that more noise was being modeled and that the outcome that the average did better in the end was simply a matter of chance. Had the results been more dramatic, then the analyst may have concluded that additional signal was being teased out in the combined model. This effort yielded two conclusions:

1) That wildly different models using triangulation definitely had potential.
2) Extra effort would need to be invested before the results warranted going this route;
3) Additional knowledge was needed about how to combine the two models. For example, should a simple mean, weighted average, or multiplication be used to combine them?

<u>FURTHER ANALYSIS - REGRESSION ON DISCORDANT DATA</u>

Before the final modeling effort began, an attempt was made to understand why 75% of the data was so easy to model and why 20% of the data was so difficult. To investigate this question, a simple model was developed that had a c-value of over 75%. I segregated a portion of the discordant data by exporting where target claim=0 and prediction >.5 and target claim =1 prediction<.5); I specifically did not worry about the target=0 where prediction was less than .5 because my main concern was trying to identify additional signal to inprove the rankings for actual target claims. A regression was run on this discordant dataset to examine what was important and how it was being modelled. The new regression coefficients were nearly the opposite of the successful model! For example, motor vehicle points were negatively associated with claims. This informed me that the discordant data, specifically the target claims, was remarkably noisy.

In a further effort to see if ANYTHING could be discerned from this data, the data was thoroughly examined in Microsoft Excel. After a few hours, one important data insight was discovered. Because the data was primarily claims that had been improperly predicted as no-claims, the goal was to find what, if any warning signs were available that a claim would occur. And I found one.

It was discovered that there are records where motor vehicle points are zero, but that a previous claim had been filed, and the intersection of these two items was of significance. Because all insurance claims typically require a police report, this suggests one of two situations: 1) The owner's vehicle was vandalized or damaged while being parked, which suggests that the person lives in a bad area of town; or 2) The vehicle was involved in an accident causing vehicle damage where no one was hurt (which in some states results in no investigative action by police and no tickets for the party causing the accident). In either situation, the critical pre-claim indicator of motor vehicle points was disabled. Consequently new binary flags were constructed to reflect a past claim with 0 or 1 motor vehicle points and a past claim. This would allow the regression to determine if either of these categories warranted an increase in the odds of a claim. This variable did show up in some of the earlier forms of the third model, but in the end, a recipe was selected where this variable was not included. It was a worthwhile finding for future model improvement.

FINAL MODELING EFFORT

With a well-formed dataset, an extensive list of variables, and a lot of knowledge about the important relationships in the data, the final general-purpose modeling effort began. Three models were developed: one with a very short list of the most important variables, a somewhat longer list, and an extensive list. The results were compared and then validated against validation data that had been removed from the training set.

Based on knowledge of the data and all of the previous modeling work that had been done to create segmented models, triangulated models, etc., I used the stepwise regression procedure with three different sets of variables with the goal of letting the computer begin the selection. Upon completion, I would examine it and remove any variables that I wanted removed. The goal was to create three levels of complexity and evaluate their performance. If the validation effort supported the best of the initial models, then that model would be submitted.

The three models that were developed are as follows. Terms associated negatively with are colored in red. As you can see, all the variables are significant for all three models. Discriptions of each variable are included with the variable name.

Model 1 - Very Short List

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Point Estimate	95% Wald Confidence Limits	
Intercept	1	-0.8339	0.1045	63.6398	<.0001	.4343 or 30.28%		
HOME_VAL_100K Home value divided by 100000.	1	-0.1145	0.0347	10.9163	0.001	0.892	0.833	0.954
INCOME_IMPUTED_100K Income as imputed above divided by 100000	1	-0.8332	0.0878	89.9796	<.0001	0.435	0.366	0.516
TRAVTIME Commute time in minutes	1	0.0142	0.00192	54.5151	<.0001	1.014	1.01	1.018
F_MVR_PTS_GT6 Flag variable, 1=More than 6 MVR points, 0=no	1	1.0057	0.1394	52.0706	<.0001	2.734	2.08	3.593
F_PREVCLAIM Flag variable, 1=Previous claim filed, 0=none	1	0.4765	0.062	59.1215	<.0001	1.61	1.426	1.818
F_RURAL Flag variable, 1=Rural, 0=Urban	1	-2.1935	0.1142	368.6127	<.0001	0.112	0.089	0.14
F_CARUSE_COMMERCIAL Flag var, 1=commercial vehicle, 0=private	1	0.6635	0.0622	113.9197	<.0001	1.942	1.719	2.193
F_KIDSATHOME Flag variable, 1=any number of kids at home	1	0.5738	0.062	85.5537	<.0001	1.775	1.572	2.005

F_REVOKED Flag variable, 1=license has been revoked	1	0.7491	0.0819	83.6662	<.0001	2.115	1.801	2.483
F_MARRIED Flag variable, 1=Married, 0=Not married	1	-0.6319	0.0728	75.4344	<.0001	0.532	0.461	0.613
F_VEH_MINIVAN Flag variable, 1=Auto is a minivan, 0=other	1	-0.6826	0.0759	80.7816	<.0001	0.505	0.435	0.586
F_JOB_MANAGER Flag variable, 1=Job is manager, 0=other	1	-0.7885	0.1083	52.9747	<.0001	0.455	0.368	0.562

Model 2 - Short List

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Point Estimate	95% Wald Confidence Limits	
Intercept	1	-0.5779	0.0858	45.4069	<.0001	.5611 or 35.94%		
INCOME_IMPUTED_100K Income as imputed above divided by 100000	1	-0.8395	0.0887	89.6501	<.0001	0.432	0.363	0.514
HOME_VAL_100K Home value divided by 100000.	1	-0.1193	0.0348	11.7619	0.0006	0.888	0.829	0.95
MVR_PTS Number of Motor Vehicle Record Points	1	0.071	0.0175	16.4077	<.0001	1.074	1.037	1.111
OLDCLAIM Numerical amount in dollars of previous claims	1	-0.00002	4.26E-06	22.9129	<.0001	1	1	1
F_PREVCLAIM Flag variable, 1=Previous claim filed, 0=none	1	0.6035	0.0807	55.9292	<.0001	1.828	1.561	2.142
F_RURAL Flag variable, 1=Rural, 0=Urban	1	-2.0492	0.1132	327.9291	<.0001	0.129	0.103	0.161
F_MVR_PTS_GT6 Flag variable, 1=More than 6 MVR points, 0=no	1	0.6226	0.1682	13.7022	0.0002	1.864	1.34	2.591
F_REVOKED Flag variable, 1=license has been revoked	1	0.9704	0.0938	106.9814	<.0001	2.639	2.196	3.172
F_CARUSE_COMMERCIAL Flag var, 1=commercial vehicle, 0=private	1	0.6135	0.0724	71.8989	<.0001	1.847	1.603	2.128
F_VEH_MINIVAN Flag variable, 1=Auto is a minivan, 0=other	1	-0.6933	0.0761	82.8978	<.0001	0.5	0.431	0.58
F_MARRIED Flag variable, 1=Married, 0=Not married	1	-0.6139	0.0727	71.3924	<.0001	0.541	0.469	0.624

F_KIDSATHOME Flag variable, 1=any number of kids at home	1	0.5719	0.0621	84.9167	<.0001	1.772	1.569	2.001
F_A5_J2 Flag var, 1=Age between 56 and 62 and job is professional, 0=any other	1	0.9607	0.4194	5.2476	0.022	2.614	1.149	5.946
F_A4_J6 Flag var, 1=Age between 34 and 55 and job is manager, 0=any other	1	-1.1175	0.2246	24.7486	<.0001	0.327	0.211	0.508
F_A5_J7 Flag var, 1=Age between 56 and 62 and job is blue collar, 0=any other	1	0.8356	0.2623	10.148	0.0014	2.306	1.379	3.856
F_A4_E2 Flag var, 1=Age between 34 and 55 and job is professional, 0=any other	1	0.3744	0.1035	13.075	0.0003	1.454	1.187	1.781

Model 3 - Long List

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Point Estimate	95% Wald Confidence Limits	
Intercept	1	-0.4518	0.2511	3.2374	0.072	.6365 or 38.9%		
INCOME_IMPUTED_100K Income as imputed above divided by 100000	1	-0.6527	0.0869	56.3544	<.0001	0.521	0.439	0.617
MVR_PTS Number of Motor Vehicle Record Points	1	0.0564	0.018	9.8407	0.0017	1.058	1.021	1.096
AGE Age of Person	1	-0.0222	0.00483	21.1556	<.0001	0.978	0.969	0.987
OLDCLAIM Numerical amount in dollars of previous claims	1	-0.00002	4.37E-06	19.624	<.0001	1	1	1
TRAVTIME Commute time in minutes	1	0.0175	0.00228	58.7566	<.0001	1.018	1.013	1.022
F_MVR_PTS_GT6 Flag variable, 1=More than 6 MVR points, 0=no	1	0.5925	0.1719	11.8807	0.0006	1.809	1.291	2.533
F_TRAVELTIME_61 Commute time exceeds 60 minutes	1	-0.4152	0.1701	5.9548	0.0147	0.66	0.473	0.922
F_a22_26 Flag var, age is between 22 and 26	1	1.1542	0.2695	18.3447	<.0001	3.172	1.87	5.378
F_a57_62 Flag var, age is between 57 and 62	1	1.106	0.136	66.0889	<.0001	3.022	2.315	3.946

F_a63_99 Flag var, age is between 63 and 99	1	0.8221	0.2504	10.7757	0.001	2.275	1.393	3.717
F_REVOKED Flag variable, 1=license has been revoked	1	0.9478	0.0964	96.7056	<.0001	2.58	2.136	3.116
F_PREVCLAIM Flag variable, 1=Previous claim filed, 0=none	1	0.5842	0.0823	50.3647	<.0001	1.794	1.526	2.108
F_RURAL Flag variable, 1=Rural, 0=Urban	1	-2.299	0.117	386.4456	<.0001	0.1	0.08	0.126
F_CARUSE_COMMERCIAL Flag var, 1=commercial vehicle, 0=private	1	0.741	0.0701	111.8229	<.0001	2.098	1.829	2.407
F_SINGLEPARENT Flag var, person is single parent	1	0.2436	0.1059	5.2963	0.0214	1.276	1.037	1.57
F_MARRIED Flag variable, 1=Married, 0=Not married	1	-0.5832	0.0833	49.0055	<.0001	0.558	0.474	0.657
F_KIDSDRIVING Flag var, 1=at least 1 child driving, 0=none	1	0.7508	0.0913	67.6238	<.0001	2.119	1.772	2.534
F_HOMEOWNER Flag var, 1=homeowner, 0=unknown or not	1	-0.2453	0.0745	10.8508	0.001	0.782	0.676	0.905
F_JOB_MANAGER Flag variable, 1=Job is manager, 0=other	1	-0.7273	0.1115	42.5416	<.0001	0.483	0.388	0.601
F_ED_HS Flag var, person graduated high school	1	0.496	0.0781	40.3686	<.0001	1.642	1.409	1.914
F_ED_NOHS Flag var, education is primary only	1	0.4254	0.0988	18.5336	<.0001	1.53	1.261	1.857
F_VEH_SPORTSCAR Flag var, car is sports car	1	0.3124	0.0976	10.2407	0.0014	1.367	1.129	1.655
F_VEH_MINIVAN Flag variable, 1=Auto is a minivan, 0=other	1	-0.671	0.0803	69.8943	<.0001	0.511	0.437	0.598
F_VEH_PANELTRUCK Flag var, vehicle is panel truck	1	-0.3103	0.1196	6.7288	0.0095	0.733	0.58	0.927

The models have many variables and the above table warrants some interpretation. The estimate is the coefficient for the particular variable. These coefficients create a linear equation that results in an output that is the log of the odds for the prediction. Thus, the coefficients are much more difficult to understand, which is why the third column from the right was added which is a point estimate for the particular variable which is obtained by exponentiating the coefficient. The point estimate is the change in the odds for a one unit change in the variable, or, in the case of the intercept, the value of the intercept in odds terms. (The probability associated with the intercept is calculated and is adjacent to the point estimate.)

It is easier to understand with an example. For Model 1, if all predictors are zero, the Intercept is -0.8339, which after exponentiating is .4343 (the odds), yielding a probability of 30.28%. Therefore, the base probability for a claim is 30.28%, and the model then adjusts that probability up or down depending upon the predictors. If F_PREVCLAIM is one (i.e., the person has a previous insurance claim), then that adds .6035 to the equation. After exponentiating the coefficient, that increases the odds of a claim by 1.61. To give an example, if the odds were 1 to 1 (i.e., 50% probability that a claim would be filed based on the rest of the equation), then having F_PREVCLAIM would raise the probability from 50% to 72%. All other variables work in the same manner, and so, INCOME_IMPUTED_100K would be multiplied by its coefficient which represents the change in the log-odds per $100,000 of income to create the relevant log-odds change.

The first two models do not warrant a complete explanation of them because they are not to be used; however, the following comments highlight important elements:

Model 1

- The model starts slightly above the overall mean for claims, and then works its way down.
- All of the negative coefficient variables have, accordingly, high impact. Increasing income, living in a rural location, being married, and driving a minivan all reduce the odds of a claim substantially. Home value per $100,000 has only a modest affect.
- All of that increase the odds are very sensible. Increasing commute times increases the relative odds of a claim. Having kids at home, driving a commercial vehicle, or having had a revoked licenses also increases those odds.
- On the factors that drive up the odds of a claim, the flag of motor vehicle points greater than 6, F_MVRPTS_GT6, is notable. If this flag is set, the log odds goes up by more than 1 which increases the odds by more than 2.7.
- As will be seen in the performance, this model, with only twelve variables, has good performance because the variables were selected carefully.

Model 2

- For the second model, the flag for job as a manager was replaced with the continuous variable of the dollar amount of previous claims. Travel time was dropped, and motor vehicle record points was added. Finally, additional employment-age flags were offered to the stepwise regression, and it selected four of them as significant:
    - F_A5_J2: Flag var, 1=Age between 56 and 62 and job is professional, 0=any other
    - F_A4_J6: Flag var, 1=Age between 34 and 55 and job is manager, 0=any other
    - F_A5_J7: Flag var, 1=Age between 56 and 62 and job is blue collar, 0=any other
    - F_A4_E2: Flag var, 1=Age between 34 and 55 and job is professional, 0=any other

- This model starts with an even higher intercept, thus meaning that it has to go down further. This was a concern about how it would fare.
- Most of the variables had similar coefficients to Model 1, though they were slightly adjusted to account for the new variables and higher intercept value.

Model 3

This final model has the form of the following equation:

```
-0.4518+F_PREVCLAIM*0.5842+F_RURAL*-2.299+INCOME_IMPUTED_100K*-0.6527+
F_CARUSE_COMMERCIAL*0.741+F_SINGLEPARENT*0.2436+F_REVOKED*0.9478+
F_VEH_MINIVAN*-0.671+F_JOB_MANAGER*-0.7273+F_MVR_PTS_GT6*0.5925+TRAVTIME*0.0175+
F_MARRIED*-0.5832+F_KIDSDRIVING*0.7508+F_a57_62*1.106+F_a22_26*1.1542+F_ED_HS*0.496+
F_ED_NOHS*0.4254+OLDCLAIM*-0.00002+AGE*-0.0222+F_a63_99*0.8221+
F_HOMEOWNER*-0.2453+F_VEH_SPORTSCAR*0.3124+MVR_PTS*0.0564+
F_VEH_PANELTRUCK*-0.3103+F_TRAVELTIME_61*-0.4152;
```

The following items should be noted about the final model:

- It has 5 continuous variables: income (-.65), amount of previous claim (.00002), travel time(-.0175), the number of motor vehicle points (.06), and person's age (-.02). These are all impactful variables based on the fact that the coefficients multiply by their continuous value. For example, for every $100,000 in income the log-odds is reduced by .65. For every minute of travel time, the log-odds are increased by .02. At fifty years old, one gets a full value of 1 subtracted from the log-odds. The previous claim amount is a smaller item, requiring a value of $25,000 to increase the log-odds by .5.

- It has some "adjuster" flags to enhance the continuous variable odds for specific categories. Coefficients are in parentheses to note impact:
    - F_MVR_PTS_GT6: Flag variable, 1=More than 6 MVR points, 0=no (+.59)
    - F_TRAVELTIME_61: Commute time exceeds 60 minutes (-.4)
    - Age category adjusters
        - F_a22_26: Flag var, age is between 22 and 26 (+1.15)
        - F_a57_62: Flag var, age is between 57 and 62 (+1.1)
        - F_a63_99: Flag var, age is between 63 and 99 (+.82)

- Driving-related attributes. All of these have highly impactful coefficients with rural at -2.3, revoked license at .94 , previous claim at .58, and commercial car use at .74.
    - F_RURAL: Flag variable, 1=Rural, 0=Urban
    - F_REVOKED: Flag variable, 1=license has been revoked
    - F_PREVCLAIM: Flag variable, 1=Previous claim filed, 0=none
    - F_CARUSE_COMMERCIAL: Flag var, 1=commercial vehicle, 0=private

- Car-specific adjustments. For sports car (+.3), and panel truck (-.3), these are minor, but minivan receives a more impactful coefficient (-.6)
    - F_VEH_SPORTSCAR: Flag var, car is sports car
    - F_VEH_MINIVAN: Flag variable, 1=Auto is a minivan, 0=other
    - F_VEH_PANELTRUCK: Flag var, vehicle is panel truck

- Personal attributes. Kids driving is impactful (+.75) as is Marriage (-.58), but the others are modest. Homeowner at -.25 and Single Parent -.24.
  - F_SINGLEPARENT: Flag var, person is single parent
  - F_MARRIED: Flag variable, 1=Married, 0=Not married
  - F_KIDSDRIVING: Flag var, 1=at least 1 child driving, 0=none
  - F_HOMEOWNER: Flag var, 1=homeowner, 0=unknown or not

- Job/education adjustments. These are listed in descending importance with job manager at -.73, high school education at .49 and primary-only education at .42.
  - F_JOB_MANAGER: Flag variable, 1=Job is manager, 0=other
  - F_ED_HS: Flag var, person graduated high school
  - F_ED_NOHS: Flag var, education is primary only

As you can see, there are no surprises. The factors that one would expect to increase the odds (e.g., driving a sports car or having kids) have positive coefficients. Factors that should reduce the odds (e.g., driving a minivan or panel truck—the opposite of a sports car, living in a rural area, or being married) all have negative coefficients. Even those elements that might be a surprise on one level like the age kickers, are not a surprise upon consideration. The age continuous variable gives a benefit to experience, but older drivers have slower reflexes and some of the age benefit disappears.

It is interesting to see what variables did not make it into the third model. The age kicker variables replaced the job-age kickers, and there was only one job variable. Two other car types were added to the minivan flag variable. The travel time kicker also fine-tuned the continuous variable from the same data.

The job and education categories are data-driven insights, though it is not exactly clear why someone who has no college is a worse driver. It may be that these people are the commercial drivers or have longer commutes due to lower incomes, and that they are, essentially, kickers to the income variable and commercial driving variable.

MODEL PERFORMANCE COMPARISON

First, the AIC numbers for these models are extremely large (i.e., >5000) because the number of combination pairs (10721970) is so great and the number of misclassifications so large. The example in the Powerpoint examples had only 100 data points, and so the AIC numbers had a significant change. In this case, the numbers were not different to use them as a discrimination tool. Therefore, the focus for performance will be on ROC curves, c values (area under the ROC curve), and other metrics below:

	Model 1	**Model 2**	**Model 3**
Percent Concordant	73.9	73.6	75.9
Percent Discordant	15.9	16	14.7

Percent Tied	10.2	10.4	9.5
Pairs	10721970	10721970	10721970
Somers' D	0.58	0.576	0.612
Gamma	0.645	0.643	0.676
Tau-a	0.224	0.223	0.237
C	0.79	0.788	0.806

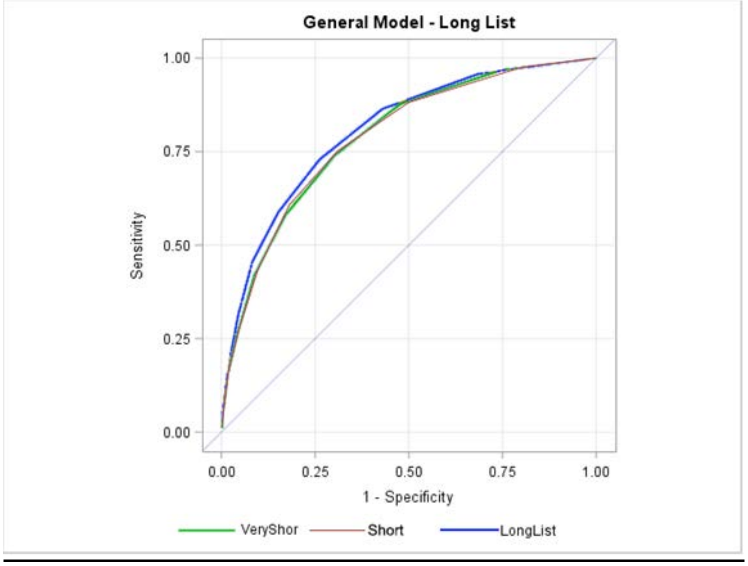
As one can see from the table above, Model 3 outperforms the first two models for every metric. As you can see from the c-values, these models are very close and have fine distinctions. The interesting item is that Model 1 with fewer terms performed better than Model 2, though a couple of the terms were different and this is surely what caused the decrease in performance. Model 2 had just the barest of increases in mis-predictions (as shown by percentage discordant and an increased number of ties), which in turn reduced the c-value. Meanwhile, Model 3's concordant matches went up, discordant matches went down, and so the total area under the curve went up. The Somer's D statistic (which subtracts off the discordant matches) is higher for the third model as well as the Gamma statistic which applies no penalty for ties.

K-S STATISTICS AND ROC CURVE ANALYSIS

The ROC curve reveals visually how Model 3 (blue) outperforms the other two models. First, as you can see, Model 1 (green) and Model 2 (red) are nearly on top of each other jostling for the better position. Model 3 starts out stronger and establishes a lead and then maintains that position. For the bottom half of the predictions, they are all very close in terms of their error rates, but the total area under the curve for Model 3 is slightly better as stated by the c statistic. This performance increase is reflected in the K-S statistic which shows the difference between the correct and error; Model 3 also has the best K-S statistic. One of the risks in using the K-S statistic at a fixed interval is shown by the differences between the max K-S statistic at the bottom and the green highlighted K-S in the decile table. The max statistic may not be accurately revealed by the decile numbers as is shown by the differences in the numbers.

K-S DATA STATISTICS

Decile	Mdl1_KS	Mdl2_KS	Mdl3_KS
1	0.219	0.219	0.241
2	0.359	0.350	0.391
3	0.419	0.425	0.445
4	0.435	0.442	0.471
5	0.440	0.429	0.457
6	0.387	0.382	0.405
7	0.317	0.320	0.324
8	0.226	0.226	0.233
9	0.121	0.117	0.123
10	0.000	0.000	0.000
Max	0.447	0.446	0.472



TARGET AMOUNT REGRESSION MODELING

The final task was to model the target claim amount. This was a very noisy model and the regression model is not particularly accurate because none of the variables correlated very well with the target claim amount variable. Many of the variables that determine probability of a claim also predict the amount. That comes as no surprise since claim amount and existence of a claim are necessarily correlated. The regression equation is as follows:

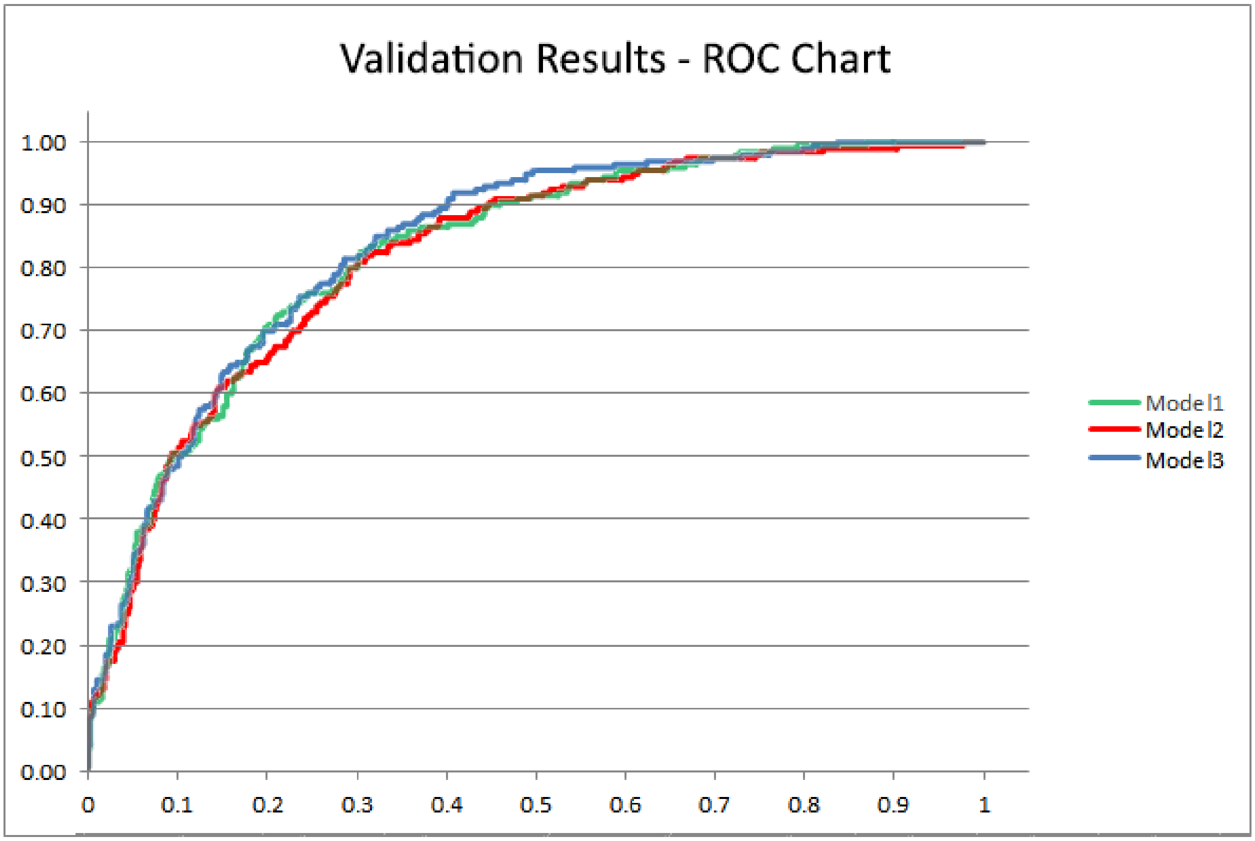
Regression Term	Analysis
1368.75956+	Base claim amout is $1369, and the other factors adjust it up or down.
INCOME_IMPUTED_100K*-742.01321+	Increasing income lowers the amount of the claim. This may relate to the fact that often wealthier people will cover expenses out of pocket even before making a claim.
BLUEBOOK_10K*104.19201+	Increasing car value increases claim amount which is no surprise given that collision insurance covers the cost of car damage.
MVR_PTS*163.22578+	The more reckless the driver, the higher the claim. That comes as no surprise.
F_MVR_PTS_GT7*1398.72296+	There is a $1400 kicker if your motor vehicle record has more than 7 points.
F_MVR0_CLAIM*463.15569	**This factor was interesting. It shows that drivers with zero points on their record have higher claims. This may relate to the fact that these are young drivers who made a mistake and totaled the car. Very interesting.**
TRAVTIME*11.16716+	Higher costs for higher commute times. This is no surprise as higher commute times probably correlate to higher speeds, which would in turn cause more damage.
F_RURAL*-1537.00823+	Being in the rural areas reduces total cost. Cars may be worth less, and certainly fewer distractions mean more

	opportunity to correct for errors and so accidents are less damaging as well as fewer.
F_REVOKED*543.07443+	A person with a revoked driver's license also has higher costs. This may simply relate to the fact that they are correlated with accidents and so part of the cost has been placed with this variable.
F_CARUSE_COMMERCIAL*562.11314+	Commercial cars have higher costs. That may relate to the fact that businesses often lease new vehicles and so the vehicles are worth more.
F_VEH_SPORTSCAR*317.4155+	Sports cars are worth more and travel at faster speeds. Any accident will likely cost more.
F_VEH_MINIVAN*-580.74619+	Minivans go slower and are driven by parents with kids. No surprise here.
F_JOB_MANAGER*-687.46026+	The job categories, as discussed earlier, may simply be proxies for other factors.
F_JOB_BLUECOLLAR*448.54071+	The job categories, as discussed earlier, may simply be proxies for other factors.
F_MARRIED*-637.20824+	This is likely a correlation that married people have fewer accidents.
F_KIDSDRIVING*633.88939+	Though this may relate simply to the higher correlation of kids driving, it is also reasonable to think that new drivers can make worse mistakes that cost more.
F_SINGLEPARENT*684.08701+	Single parents are highly correlated to accidents, so it may relate to that; that said, single parents may also be more distracted thus causing more expensive accidents.
F_SEX_MALE*209.63191+	**This is very interesting. Gender did not appear for purposes of whether a claim would occur, but it does appear that male drivers cause more expensive accidents. Male drivers may be more aggressive. Very interesting.**
F_ED_BA*-260.78415	Having a bachelor's degree (but not a Masters or PhD) makes one accidents less costly? This would seem to exist only due to the correlation with fact of a claim.

Because this regression is not a graded part of the project, discussion of the fit and validation results will be avoided. Suffice it to say that the variables are significant and it makes a prediction.

VALIDATION TESTING

To validate the models against unknown data, 718 random records containing 200 true claim records were set aside before the training and model selection began. The three models were calculated and performance data was exported. The claim amount model was not evaluated. The ROC curves for the three claim models are as follows:



The validation testing confirms the superiority of Model 3 with more area under the curve for Model 3 than the other two models. The K-S data was exported along with records that showed the percentage of true claims detected by the three models, and a portion of this data is included to give a more detailed feel for how Model 3 performs with unknown data.

As the data below shows, the three models ranking 10% of the true claims within the top 3+ percent of the records. The models jostle about (as one can see in the ROC curves above), with one different models falling and rising in performance. All three of them hit their max K-S scores (highlighted in orange brown below) at about 45% of the data. It was at that point, with the highest K-S score, Model 3 established a dominant lead and the other two models never caught up. In fact, Model 3 had identified all of the true claims nearly ten percent of the records earlier than Model 2 and about two percent ahead of Model 1. ***Overall, these are very encouraging validation results. In the training data set, Model 3 established a dominant position early on rather than later. However, the validation results suggest that Model 3 may have better performance in the lower rankings of the dataset where many errors may occur.***

Record		**K-S VALUES**		**% of Trues**			**Notes**
Percent	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	
3.1%	0.089	0.096	0.089	9.5%	**10.0%**	9.5%	All three models start out the same.
3.2%	0.094	0.101	0.094	**10.0%**	10.5%	**10.0%**	
7.2%	0.177	0.156	0.170	**20.0%**	18.5%	19.5%	

7.4%	0.182	0.161	0.175	20.5%	19.0%	**20.0%**	
7.5%	0.187	0.159	0.180	21.0%	19.0%	20.5%	
7.7%	0.185	0.164	0.185	21.0%	19.5%	21.0%	
7.8%	0.183	0.162	0.190	21.0%	19.5%	21.5%	
7.9%	0.181	0.167	0.195	21.0%	**20.0%**	22.0%	
10.0%	0.221	0.208	0.228	26.0%	25.0%	26.5%	
11.6%	0.256	0.242	0.249	**30.0%**	29.0%	29.5%	Model 2 has fallen behind.
11.7%	0.261	0.240	0.254	30.5%	29.0%	**30.0%**	
11.8%	0.266	0.245	0.259	31.0%	29.5%	30.5%	
12.0%	0.271	0.250	0.257	31.5%	**30.0%**	30.5%	
15.9%	0.327	0.321	0.334	39.5%	39.0%	**40.0%**	Model 3 pulls ahead.
16.0%	0.326	0.319	0.339	39.5%	39.0%	40.5%	
16.2%	0.331	0.324	0.344	**40.0%**	39.5%	41.0%	
16.3%	0.336	0.329	0.349	40.5%	**40.0%**	41.5%	
19.9%	0.396	0.396	0.389	48.5%	48.5%	48.0%	
20.5%	0.402	0.409	0.388	49.5%	**50.0%**	48.5%	But Model 3 falls back.
20.6%	0.400	0.407	0.387	49.5%	50.0%	48.5%	
20.8%	0.405	0.412	0.385	**50.0%**	50.5%	48.5%	
20.9%	0.410	0.410	0.390	50.5%	50.5%	49.0%	
21.0%	0.408	0.408	0.395	50.5%	50.5%	49.5%	
21.2%	0.407	0.407	0.400	50.5%	50.5%	**50.0%**	
26.9%	0.424	0.459	0.452	57.5%	**60.0%**	59.5%	
27.0%	0.429	0.457	0.457	58.0%	60.0%	**60.0%**	Model 3 catches up.
27.2%	0.427	0.462	0.462	58.0%	60.5%	60.5%	Note we are only 27% of
27.3%	0.426	0.460	0.460	58.0%	60.5%	60.5%	Total dataset.
27.4%	0.431	0.465	0.458	58.5%	61.0%	60.5%	
27.6%	0.436	0.463	0.463	59.0%	61.0%	61.0%	
27.7%	0.441	0.461	0.461	59.5%	61.0%	61.0%	
27.9%	0.446	0.459	0.466	**60.0%**	61.0%	61.5%	
30.1%	0.463	0.463	0.484	63.5%	63.5%	65.0%	
33.6%	0.505	0.463	0.505	**70.0%**	67.0%	**70.0%**	
35.9%	0.514	0.472	0.493	73.0%	**70.0%**	71.5%	
40.0%	0.499	0.486	0.513	76.0%	75.0%	77.0%	
42.6%	0.497	0.490	0.518	78.5%	78.0%	**80.0%**	At 40% of dataset, Model3
42.8%	0.502	0.495	0.523	79.0%	78.5%	80.5%	really gets ahead. It main-
42.9%	0.507	0.493	0.521	79.5%	78.5%	80.5%	tains that position.
43.0%	0.505	0.498	0.519	79.5%	79.0%	80.5%	
43.2%	0.510	0.503	0.524	**80.0%**	79.5%	81.0%	
43.3%	0.508	0.508	0.529	80.0%	80.0%	81.5%	
43.5%	0.507	0.507	0.527	80.0%	**80.0%**	81.5%	
44.3%	0.509	0.509	0.516	81.0%	81.0%	81.5%	
44.8%	0.522	0.501	0.515	82.5%	81.0%	82.0%	
46.8%	0.516	0.495	0.530	84.0%	82.5%	85.0%	

50.0%	0.499	0.478	0.513	86.0%	84.5%	87.0%	
53.9%	0.459	0.473	0.500	87.0%	88.0%	**90.0%**	
57.1%	0.449	0.456	0.491	89.5%	**90.0%**	92.5%	
57.2%	0.454	0.454	0.489	**90.0%**	90.0%	92.5%	
60.0%	0.429	0.429	0.464	91.0%	91.0%	93.5%	
61.7%	0.413	0.413	0.462	91.5%	91.5%	**95.0%**	
68.9%	0.361	0.347	0.375	**95.0%**	94.0%	96.0%	
70.3%	0.349	0.342	0.363	95.5%	**95.0%**	96.5%	
80.1%	0.255	0.241	0.248	98.5%	97.5%	98.0%	
82.3%	0.224	0.224	0.224	98.5%	98.5%	98.5%	
88.2%	0.157	0.150	0.164	99.5%	99.0%	**100.0%**	Model 3 finishes early.
90.0%	0.132	0.125	0.139	99.5%	99.0%	100.0%	
90.5%	0.131	0.117	0.131	**100.0%**	99.0%	100.0%	
98.3%	0.023	0.023	0.023	100.0%	**100.0%**	100.0%	
100.0%	0.000	0.000	0.000	100.0%	100.0%	100.0%	

CONCLUSION

This project was started a week in advance, but took almost nearly the entire allotted time. The reason for the delay in completion was simple; approach after approach was taken to attempt to find some additional signal in the noisy data, and approach after approach failed. The final model is a good model, and should provide similar performance given similar data. It is not, in my opinion, an exceptional one and still has room for improvement. My estimation is that proper analysis might improve the c statistic by 2-5%, though getting more performance than that may prove impossible. With additional time, the following approaches are the ones that I would take to try to improve the model:

- Ensemble modeling based on different strategies. There is definitely promise in this approach as marginal benefits were obtained by taking this strategy. Averaging multiple models may yield the insights of all the models (e.g., the wisdom of foolish models).

- Analysis of weaknesses of the current model. In analyzing the weakness of the simple model by looking at discordant data, the strange fact that there were people who had claims but no motor vehicle points records was discovered. Additional analysis of discordant data is likely to yield other insights.

- Additional investigation of high-value variables like motor vehicle points. Are there categories where the general relationship of points to claims break down? For example, are there doctors who have points who do not have claims? This would suggest that there may be people who, generally speaking, are excellent drivers but who get an errant speeding ticket. This sort of analysis may provide additional lift to the model.

Appendix

Extra Credit Remarks

To ensure that extra credit items are fully considered, I thought it might be useful to identify the following areas that may be a consideration for extra credit:

- SAS programming methodologies, including the use of macros and structured programming techniques like the use of include files.

- Extensive exploration of the data including matrixed categorical and discretized variables.

- Extensive attempts to improve the model by altering approaches including data segmentation and a triangulation approach based on wildly differing models.

- Analysis of the discordant data to determine what other variables might be created to further boost the signal. That effort located a case where persons had prior claims, but no motor vehicle points, suggesting that these persons could be missed because they do not get the bump that someone who had both claims and points.

- Use of WEKA for data exploration and data segmentation. Unfortunately, this work did not go very far because WEKA broke under the weight of the variable count and data records.

- Use of segregated validation data and performance checking to test against non-trained data. Export of the validation data so that an in-depth analysis of performance could be conducted.