Homework 2: Insurance Logistic Regression Project

Jack Doig
PREDICT 411 Section 59

Contents

Introduction 2
1. Data Exploration 3
2. Data Preparation 10
3. Build Models 11
4. Select Models 13
Conclusion 14

Introduction

The purpose of this paper is to analyse and model historical insurance data so as to build a logistic model to predict whether someone was in a car crash, and secondly to predict the expect cost of the crash, where it has occurred.

The data comes from an insurance company, with each record representing a customer, and a variety of demographic information such as education level, and information about the insured vehicle and its occupants, such as the age of the car. The predictor variables include both categorical and continuous variables.

This task will involve firstly exploring and optimizing the data such that it offers as much predictive intelligence as possible, and then developing and scoring several logistic regression models.

In order to determine the best model, a variety of different methods will be employed, including Stepwise, Forward and backward variable selection models, while each model judged on a balance of in-sample and out-of-sample measures, interpretability and simplicity.

The final model will then be analysed so as to understand its accuracy in predicting insurance claims, and further analysis undertaken if required.

1. Data Exploration

a) Data Summary

There are approximately 8200 rows in the data set. There are a 13 predictor variables, including both categorical and continuous variables.

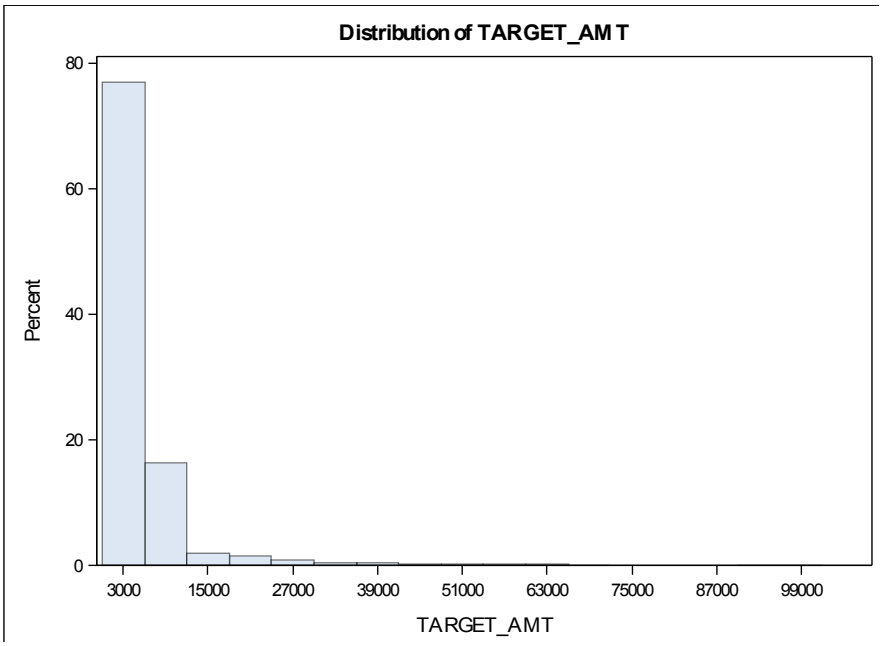
The predictor data can be described in 3 broad categories. There is a category of information about the driver, for example whether they have children, age, income, and house value. Secondly, there is information about the vehicle, for example the value, whether it is red, and its age, and finally there is a category of information about past claims of the customer.

b) Descriptive Statistics: Target Variables

There are no missing values among the two target variables, representing whether a customer had an accident and the amount of the claim. The first of these is a binary (Yes/No) field, while the second is a continuous variable in dollars.

Regarding the flag where the customer accident flag, 26.3% of the values are positive (1), with no missing data.

The cost of the accident variable has a maximum of $108,000, and where there was an accident, a mean of $5,702.17. The distribution of this variable, filtered again to where there was an accident, is below:



As can be seen in the distribution above, the value of the claims is less than $10,000 for 90% of the cases. There is however a long right tail of values, stretching to a maximum of 108,000. This is quite plausible data given my research on the severity of automobile accidents: The distribution is skewed towards minor accidents.

c) Descriptive Statistics: Understanding the Predictor variables

There are 13 predictor variables. The table below summarizes the continuous predictor variables

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
TARGET_FLAG		8161	0.263816	0.440728	0	1
TARGET_AMT		8161	1504.32	4704.03	0	107586.14
KIDSDRIV	#Driving Children	8161	0.171058	0.511534	0	4
AGE	Age	8155	44.79031	8.62759	16	81
HOMEKIDS	#Children @Home	8161	0.721235	1.116323	0	5
YOJ	Years on Job	7707	10.49929	4.092474	0	23
INCOME	Income	7716	61898.1	47572.69	0	367030.26
HOME_VAL	Home Value	7697	154867.3	129123.8	0	885282.34
TRAVTIME	Distance to Work	8161	33.4888	15.90475	5	142.1206304
BLUEBOOK	Value of Vehicle	8161	15709.9	8419.73	1500	69740
TIF	Time in Force	8161	5.351305	4.146635	1	25
OLDCLAIM	Total Claims(Past 5 Years)	8161	4037.08	8777.14	0	57037
CLM_FREQ	#Claims(Past 5 Years)	8161	0.798554	1.158453	0	5
MVR_PTS	Motor Vehicle Record Points	8161	1.695503	2.147112	0	13
CAR_AGE	Vehicle Age	7651	8.328323	5.700742	-3	28

This table highlights the missing variables (discussed further in the next section). It also highlights possible unrealistic values which may need to be imputed.

There is only one continuous predictor which has an unrealistic extreme value when considered alone in the context of car insurance. This is the age of the car of -3, which is impossible. Assuming the 0-value car ages are rounded to zero, this is the only extrema which will need to be imputed.

In terms of the categorical variables, here is a summary of the information contained within the binary (True/false) categorical predictors:

Predictor/label	% in category
Vehicle use - commercial	37.12 %
Vehicle use - private	62.88 %
Marital Status - Married	59.97%
Marital Status – Not married	40.03%
Single Parent – No	86.8%
Single Parent - Yes	13.2%
Gender - Male	46.39%
Gender - Female	53.61%
Red Car - Yes	29.14%
Red Car – No	70.86%
Licence Revoked – Yes	12.25%
Licence Revoked – No	87.75%
Home/Work area: highly Urban/Urban	79.55%
Home/Work area: highly rural/rural	20.45%

The data above appears to be reasonable when considered against the business context: The gender breakdown, single parent, marital status predictors fall roughly into line with the demographics of the community, as one would generally expect.

Max Education Level (%)	
<High School	14.74
Bachelors	27.47
Masters	20.32
PhD	8.92
z_High School	28.55

JOB	Percent	CAR TYPE	Percent
Clerical	16.65	Minivan	26.28
Doctor	3.22	Panel Truck	8.28
Home Maker	8.4	Pickup	17.02
Lawyer	10.94	Sports Car	11.11
Manager	12.94	Van	9.19
Professional	14.63	z_SUV	28.11
Student	9.33		
z_Blue Collar	23.9		

There are a few interesting observations in the above data: the absence of a normal car is irregular, this leads me to believe this is an insurance data set comprising special vehicles only. Instead, the highest type of car occurring is the minivan, at 26%.

The largest job type is blue collar (24%), and the proportion of highest education level is high school (high school and less accounts for approximately 43% of the data). This is indicative of a lower-income, less educated segment of the community in the data set.

d) Missing variables

There are number of missing variables which will need to be examined and possibly corrected. These are summarized below:

Variable	Number Missing (Percentage)
Age	6 (0.07%)
Years On Job	454 (5.5%)
Income	445 (5.5%)
Home Value	464 (5.7%)
Vehicle Age	510 (6.3%)
Job Category	526 (6.4%)

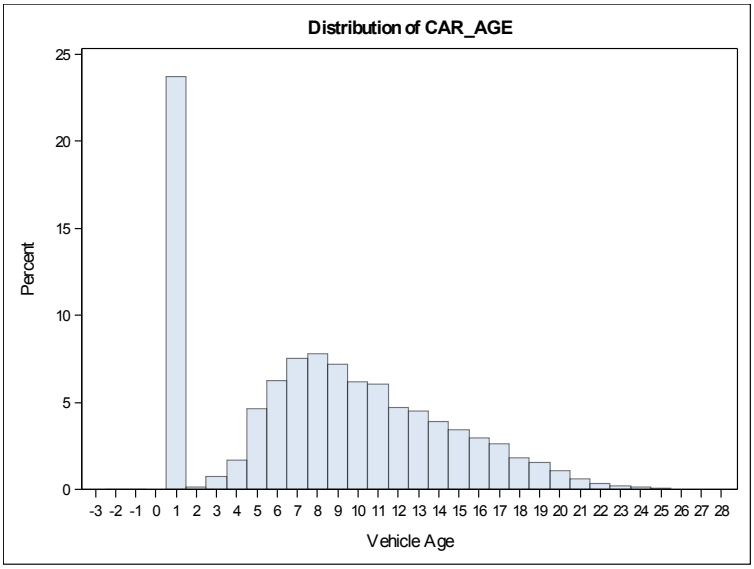
There are a number of approaches to take in order to correct these missing variables. They can be removed as rows or columns, they can form a separate category, or they can be imputed using an average, a distribution, or a more sophisticated approach such as a regression on another predictor. The best approach will be determined by looking at the correlations and descriptive statistics and then executed in the data preparation section.

e) Unrealistic Outliers

The quality of the data, when considered as isolated columns, is very good. Values for things such as home value, income, number of children being driven, ages and historical claims appear to all be within the realms of reality.

This is true for all the variables with the exception of one car age which is listed as -3. It is impossible to have a negative car age, so this will need be corrected.

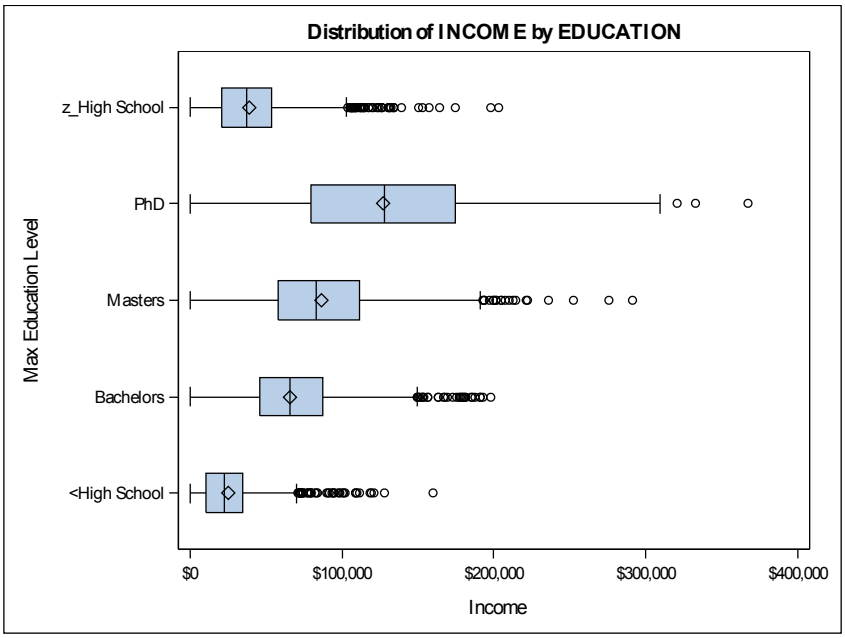
Secondly, the distribution of car ages is unrealistic which is shown in the histogram below. With car insurance virtually mandatory in every US state, I would assume a normal distribution of car ages in such a population. On the other hand, this could be explained by a 1-year bundled insurance with new cars, and the location for the insurance data is unknown. It is therefore my view that the 1-year spike in car ages should not be corrected.



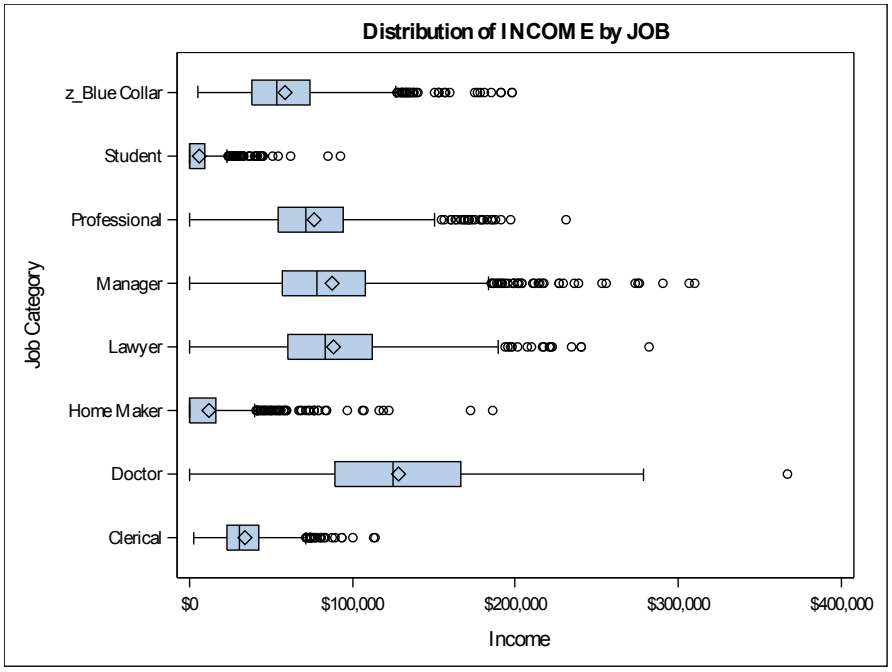
f) Data Inconsistency

There is some co-dependency of the variables, however, which may uncover some unrealistic data.

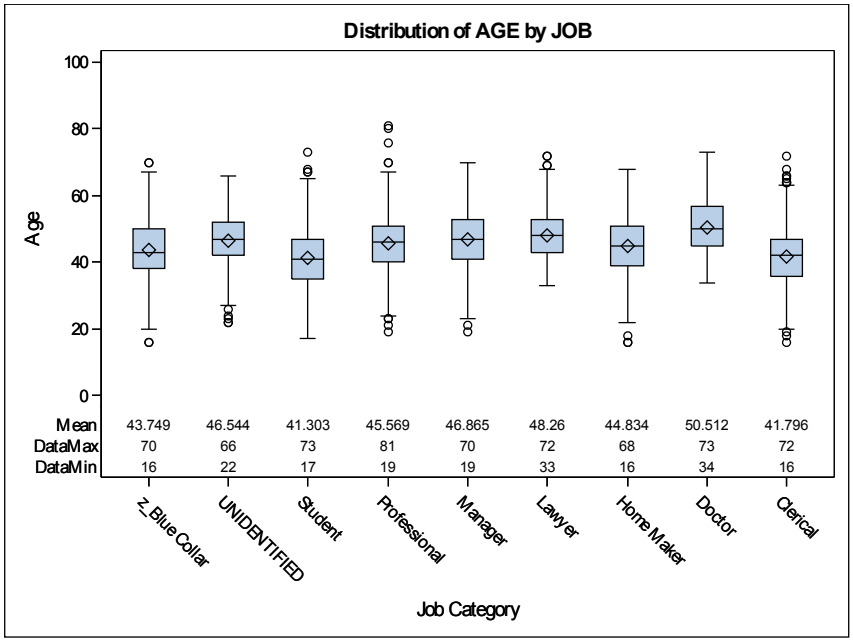
For example I would expect the income and job category or education level to be correlated. Below is a box plot for the education level against income:



What these boxplots show is a general correlation, as one would expect, between level of education and job function, and income. This analysis serves to validate the data in a general sense, although it would not be sensible to draw any firm conclusions about unrealistic data as there would naturally be exceptions in all categories. There may be professionals with zero salary as they are between jobs, for example.



There is a less expected distribution of ages by job function, however. I would expect that Students would generally be younger than the rest of the population, however the average age is 41.3 years.



g) Correlations in the data

There are two kinds of correlations to explore, firstly continuous variables where the correlations between variables might be useful for dimensionality reduction or for imputing missing data, and secondly correlations between the categorical variables and the target variable might indicate a relationship.

There are some interesting inferences we can make, which may help later in validating the model, checking it for consistency with real-world logic, suggest ways to come up with derived variables or reduce dimensionality and possibly helping in making it more parsimonious.

Below are the correlations between the continuous variables in the data:

	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS	CAR_AGE
TARGET_FLAG	-0.07051	-0.14201	-0.18374	0.04815	-0.10338	-0.08237	0.13808	0.2162	0.2192	-0.10065
TARGET_AMT	-0.02209	-0.05831	-0.0856	0.02777	-0.0047	-0.04648	0.07095	0.11642	0.13787	-0.05882
KIDSDRIV	0.0433	-0.04713	-0.01979	0.00856	-0.02155	-0.00199	0.0204	0.03706	0.05357	-0.05399
AGE	0.13607	0.18097	0.20998	0.0056	0.16503	-0.00007	-0.02929	-0.02409	-0.07158	0.17622
HOMEKIDS	0.08683	-0.15933	-0.11068	-0.00741	-0.10789	0.01181	0.02991	0.02935	0.0606	-0.15215
YOJ	1	0.28607	0.26992	-0.01693	0.14346	0.02479	-0.00298	-0.02631	-0.03786	0.06141
INCOME	0.28607	1	0.57524	-0.04704	0.42928	-0.00103	-0.04544	-0.04775	-0.06316	0.41424
HOME_VAL	0.26992	0.57524	1	-0.03537	0.25953	0.00206	-0.06919	-0.09405	-0.08539	0.21747
TRAVTIME	-0.01693	-0.04704	-0.03537	1	-0.0168	-0.01145	-0.01905	0.00641	0.01021	-0.03818
BLUEBOOK	0.14346	0.42928	0.25953	-0.0168	1	-0.00542	-0.02952	-0.03634	-0.03913	0.18976
TIF	0.02479	-0.00103	0.00206	-0.01145	-0.00542	1	-0.02196	-0.02302	-0.04105	0.00777
OLDCLAIM	-0.00298	-0.04544	-0.06919	-0.01905	-0.02952	-0.02196	1	0.49513	0.26449	-0.01339
CLM_FREQ	-0.02631	-0.04775	-0.09405	0.00641	-0.03634	-0.02302	0.49513	1	0.39664	-0.00932
MVR_PTS	-0.03786	-0.06316	-0.08539	0.01021	-0.03913	-0.04105	0.26449	0.39664	1	-0.0199
CAR_AGE	0.06141	0.41424	0.21747	-0.03818	0.18976	0.00777	-0.01339	-0.00932	-0.0199	1

There are some interesting findings (in red):

- There is a correlation of income and car value, but in addition with car age and years on job, so richer people tend to have longer-serving, more expensive cars, as well as staying at a job for a long time
- There is a correlation between the number of claims and the total value of the payout, while these are also correlated with the motor vehicle record points.

The below tables show the change in the proportion of the data, for the whole population versus the subset involved in accidents, for the categorical variables:

Max Education Level (%)	% population	% of accidents	change
<High School	14.74	17.88	3.14
Bachelors	27.47	24.29	-3.18
Masters	20.32	15.19	-5.13
PhD	8.92	5.81	-3.11
z_High School	28.55	36.83	8.28

JOB	% population	% of accidents	change
Clerical	16.65	18.39	1.74
Doctor	3.22	1.44	-1.78
Home Maker	8.4	8.92	0.52
Lawyer	10.94	7.59	-3.35
Manager	12.94	6.79	-6.15
Professional	14.63	12.2	-2.38
Student	9.33	13.19	3.86
z_Blue Collar	23.9	31.43	7.53

CAR TYPE	% population	% of accidents	change
Minivan	26.28	16.21	-10.07
Panel Truck	8.28	8.27	-0.01
Pickup	17.02	20.58	3.56
Sports Car	11.11	14.12	3.01
Van	9.19	9.34	0.15
z_SUV	28.11	31.49	3.38

Predictor/label	% population	% of accidents	change
Vehicle use - commercial	37.12	48.63	11.51
Vehicle use - private	62.88	51.37	-11.51
Marital Status - Married	59.97	48.91	-11.06
Marital Status – Not married	40.03	51.09	11.06
Single Parent – No	86.80	77.89	-8.91
Single Parent - Yes	13.20	22.11	8.91
Gender - Male	46.39	44.64	-1.75
Gender - Female	53.61	55.36	1.75
Red Car - Yes	29.14	28.61	-0.53
Red Car – No	70.86	71.39	0.53
Licence Revoked – Yes	12.25	20.58	8.33
Licence Revoked – No	87.75	79.42	-8.33
Home/Work area: highly Urban/Urban	79.55	94.66	15.11
Home/Work area: highly rural/rural	20.45	5.34	-15.11

The most significant findings are:

- High school and less in terms of education is over-represented in the accidents data, suggesting a strong predictor of being in an accident (+11.5%)
- A blue collar worker has a disproportionately higher chance of being in an accident relative to their representation in the population (+7.5%)
- A Minivan driver has a disproportionately lower chance of being in an accident (-10%)
- Living in Urban areas increases the chance of being in an accident significantly (+15%)
- Being unmarried, commercial vehicles, and being a single parent are all overrepresented relative to the proportion of the population.

2. Data Preparation

All imputation of missing or unrealistic variables have been rounded to the nearest whole number.

a) Missing variables

The list of missing variables and my approach to imputing them is outlined below

Variable	Number Missing (Percentage)	Approach
Age	6 (0.07%)	Impute via average age by job type
Years On Job	454 (5.5%)	Impute via bucket averages with from the income predictor, a
Income	445 (5.5%)	Impute via regression with closely correlated Home Value. Where Home Value not available use next best correlation which is Bluebook
Home Value	464 (5.7%)	If income =0, set home value = 0, consistent with 90% of the existing data where income=0, otherwise Impute via regression with closely correlated Income. Where income not available use next best correlation which is Bluebook
Vehicle Age	510 (6.3%)	Impute via regression with closely correlated income
Job Category	526 (6.4%)	Create a separate category, there may be a reason why these jobs were not disclosed, and there is a high number of PhD/Masters among them, suggesting some kind of special group.

b) Correcting variables

There is only one variable which I believe needs to be corrected, which is the car age of -3 and 0. Given the correlation with income, I will replace this variable with a regression imputation with the income predictor.

c) Dummy variables

In order for the regression to function, I will create dummy variables for all the categorical variables, with one dummy variable per category. This allows the categorical variables to be included as predictors in the regression.

For example, there are 9 new dummy variables for the JOB predictor, including the new category outlined above for unknown job types.

d) Derived variables

There seems to be two possibilities for derived variables. The purpose of this would be to reduce dimensionality, making any model easier to understand, and to possibility reduce mutli-colinearity and therefore reduce it side effects during model building.

Firstly with the correlation of home value, income, and years on the job, there is the possibility of creating a general measure of wealth.

The second interesting area is a correlation between motor vehicle record points, number of claims and the total claims. These variables can be combined to achieve a general measure of 'insurance claiming behaviour', with a high number being worse from the insurance company perspective.

In both cases, since these predictor variables are differently scaled, I have done this by reducing the average to zero and summing using the SAS procedure PROC STANDARD.

e) Mathematical transformations

I have explored square roots and natural logs of the predictor and the response variables in order to see if there was a noticeable increase in correlation with the response variable. The most promising transformation was the log transformation, however there was little noticeable difference in the correlations.

Given there is a sacrifice of interpretability using these transforms and little gain in correlation (if any), I will not pursue these mathematical transformations further.

3. Build Models

I have chosen to use the following ways of deriving a set of possible models:

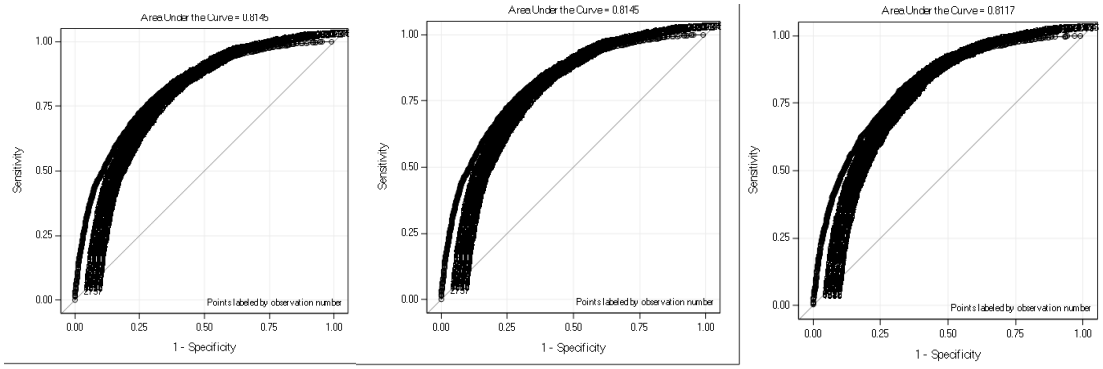
a) All original variables in a logistic regression, and then compared to a stepwise, forward and backward variable selection
b) Derived variables in a logistic regression, and then compared to a stepwise, forward and backward variable selection
c) All original variables in a probit regression, and then compared to a stepwise, forward and backward variable selection
d) Derived variables in a probit regression, and then compared to a stepwise, forward and backward variable selection

The selection criteria for these models I intend to look at the AIC, SC, concordance tests and the log likelihood, the ROC curve, as well as performance on an out of sample test.

I chose 20% of the data (randomly) to make up the out-of-sample test.

Model type	Selection method	# variables in model	AIC	SC	-2LogL	% concordant	% discordant	Out-of-sample test error rate
Original Variables – logistic	None	47	5860.185	6117.972	5784.185	81.5%	18.3%	21.92%
	Forward	20	5841.160	5983.621	5799.160	81.3%	18.5%	21.86%
	Backward	**17**	**5841.160**	**5983.621**	**5799.160**	**81.3%**	**18.5%**	**21.86%**
	Stepwise	20	5841.160	5983.621	5799.160	81.3%	18.5%	21.86%
Derived Variables replace original - logistic	None	47	5911.873	6142.524	5843.873	81.1%	18.7%	22.60%
	Forward	21	5892.920	6042.165	5848.920	81.0%	18.8%	22.35%
	Backward	20	5892.028	6034.489	5850.028	81.0%	18.8%	22.47%
	Stepwise	21	5892.028	6034.489	5850.028	81.0%	18.8%	22.47%
Original Variables – Probit	None	47	5865.397	6123.184	5789.397	81.5%	18.3%	22.05%
	Forward	20	5846.930	5989.391	5804.930	81.3%	18.5%	21.98%
	Backward	20	5846.930	5989.391	5804.930	81.3%	18.5%	21.98%
	Stepwise	20	5846.930	5989.391	5804.930	81.3%	18.5%	21.98%
Derived Variables replace original - Probit	None	47	5913.952	6144.603	5845.952	81.1%	18.7%	22.66%
	Forward	21	5895.414	6044.659	5851.414	81.0%	18.8%	22.54%
	Backward	21	5894.431	6036.892	5852.431	81.0%	18.8%	22.72%
	Stepwise	21	5894.431	6036.892	5852.431	81.0%	18.8%	22.72%

A ROC curve for three selected models is below, firstly the chosen model, secondly for a stepwise selection with all variables and finally for a probit regression for the derived set of variables.



The KS statistics for these models, using the NPAR1WAY procedure are 0.209898, 0.209898, and 0.209295 respectively for the models above.

Estimating the cost of the accidents

Finally in order to estimate the TARGET_AMT, I did a simple OLS regression with the most correlated predictor variable: Bluebook, however this resulted in an intercept of 4131, which is well above 75% of the cases. Instead I assigned an average value, using the median of $1469.

4. Select Models

My objective in choosing the model are as follows:

- The model should be meaningful, the predictor variables should contribute as one would expect for the real-world meaning of the predictors
- The model should perform well in in-sample testing based on AIC, SC and the concordance results
- In a choice between similarly performing complex versus parsimonious, the parsimonious model is preferred
- The model should perform well on out-of-sample testing

The best performing model, on the basis of the measures in the section above, is the model selected by the forward selection process with the original predictors. The reason for this it is score best on the following criteria: AIC, SC, and the out of sample set. It also only have 17 variables so is slightly easier to interpret than the other models.

The betas in the model generally make sense and are consistent with the findings of the data step:

Analysis of Maximum Likelihood Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	-1.7127
KIDSDRIV	1	0.4858
INCOME	1	-0.2529
HOME_VAL	1	-0.1712
TRAVTIME	1	0.0159
BLUEBOOK	1	-0.00003
TIF	1	-0.0566
OLDCLAIM	1	-0.1151
CLM_FREQ	1	0.202
MVR_PTS	1	0.2474
PARENT1_yes	1	0.432
MSTATUS_yes	1	-0.4919
EDUCATION_bach	1	-0.4222
EDUCATION_masters	1	-0.4208
JOB_Manager	1	-0.727
JOB_Doctor	1	-0.79
CAR_USE_private	1	-0.7734
CAR_TYPE_Minivan	1	-0.5958
CAR_TYPE_SportsCar	1	0.3298
URBANICITY_urban	1	2.353
REVOKED_yes	1	0.8939

The betas make sense, as reconciled with the learnings during the data exploration. The most significant predictors are the wealth, with a higher home value, income, higher education, better job, and car value reducing the probability of a claim. Similarly, private cars are less likely to be

involved in an accident, this is consistent with the moral hazard issue of driving someone else's vehicle.

Driving in an urban environment is likely to increase the probability of a crash, as is previous bad driving history, both of these make perfect real-world sense.

The only counter-intuitive part of the model is the OLD_CLAIM variable, representing historical insurance claims. I suspect this is a multicolinearity issue which I was trying to address with the derived variables.

Conclusion

In doing this analysis, I have explored and the insurance data set in the context of the real-world domain. Generally the quality of the data was good, however there were some interesting patterns in the data (for example the prevalence of 1-year old cars that are insured) which would certainly violate the assumptions for OLS regression that I needed to be conscious of while modelling.

There appeared to be little value in pursuing mathematical transformations of the data. There appeared to be little gain in correlation, however there is a trade-off in terms of interpretability which I was unwilling to make.

Probit regression seemed to perform more poorly than logistic regression on this dataset, although the difference between the models was marginal. The final model was a clear winner because of the mode limited number of predictors needed to achieve the same, or slightly better, performance.

The final model had an error rate of 21.86% on the out of sample test, contained 17 variables and all but one was consistent with the real-world intuition and findings in the data set during data exploration.

Therefore the model is reasonably accurate, a number of steps have been explored to arrive at this accuracy, and the model has a high level of interpretability, consistent with real-world intuition.