INTRODUCTION

The purpose of this report is to present the analytical model created to predict automotive crashes for an insurance company. The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology (figure 1) is followed during the model building process.

This report is organized according to the steps in the diagram, from *Business Understanding* to *Evaluation*. Exploratory Data Analysis (EDA) is performed first which leads to certain data preparation steps prior to linear regression model building. Multiple types of modeling techniques are investigated before a final model is selected according to certain predetermined performance criteria.

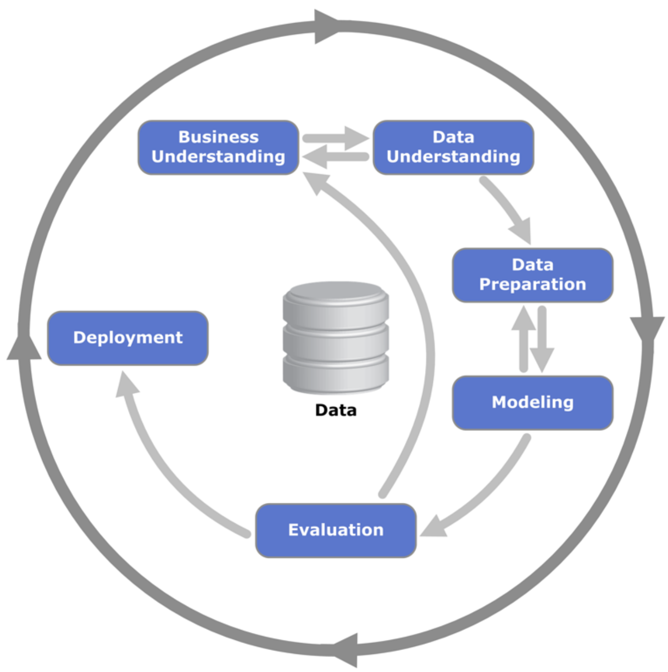


FIGURE 1. CRISP-DM METHODOLOGY

DATA UNDERSTANDING

The input dataset contained 8000 records, one for each customer. The predictor and response variables are indicated in table 1. The response variable TARGET_FLAG is 1 when the person was involved in a car crash, else the variable equals 0.

TABLE 1 INPUT DATA VARIABLES AND CATEGORIES

Category	Name	Description	Expected Effect on Response
RESPONSE	TARGET_FLAG	Car Crash Flag	
PREDICTOR	AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
	BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
	CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
	CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
	CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
	CLM_FREQ	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
	EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
	HOMEKIDS	#Children @Home	Unknown effect
	HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
	INCOME	Income	In theory, rich people tend to get into fewer crashes
	JOB	Job Category	In theory, white collar jobs tend to be safer
	KIDSDRIV	#Driving Children	When teenagers drive your car, you are more likely to get into crashes
	MSTATUS	Marital Status	In theory, married people drive more safely
	MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
	OLDCLAIM	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
	PARENT1	Single Parent	Unknown effect
	RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
	REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
	SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
	TIF	Time in Force	People who have been customers for a long time are usually more safe.
	TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
	URBANICITY	Home/Work Area	Unknown
	YOJ	Years on Job	People who stay at a job for a long time are usually more safe

MISSING DATA, INCORRECT DATA

Four of the original continuous variables have a small percentage of missing data. This will be addressed via data imputation described in the next section. Car age also has atleast one negative value, which will have to be addressed.

TABLE 2 MISSING DATA VALUES IN CONTINUOUS VARAIABLES

Variable	Label	N	N Miss	Minimum	Maximum	Mean
CAR_AGE	Vehicle Age	7651	510	-3	28	8.33
HOME_VAL	Home Value	7697	464	0	885282.34	154867.29
YOJ	Years on Job	7707	454	0	23	10.49
INCOME	Income	7716	445	0	367030.26	61898.10
AGE	Age	8155	6	16	81	44.79
KIDSDRIV	#Driving Children	8161	0	0	4	0.17
HOMEKIDS	#Children @Home	8161	0	0	5	0.72
TRAVTIME	Distance to Work	8161	0	5	142.12	33.48
BLUEBOOK	Value of Vehicle	8161	0	1500	69740	15709.90
TIF	Time in Force	8161	0	1	25	5.35
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	0	57037	4037.08
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0	5	0.79
MVR_PTS	Motor Vehicle Record Points	8161	0	0	13	1.69

One of the categorical variables also has missing values: JOB. This variable will also have to be treated before further modeling is performed.

TABLE 3 MISSING DATA VALUES IN CATEGORICAL VARAIABLES

JOB	Frequency	Percent
MISSING	526	6.45
Clerical	1271	15.57
Doctor	246	3.01
Home Maker	641	7.85
Lawyer	835	10.23
Manager	988	12.11
Professional	1117	13.69
Student	712	8.72
Blue Collar	1825	22.36

DATA PREPARATION

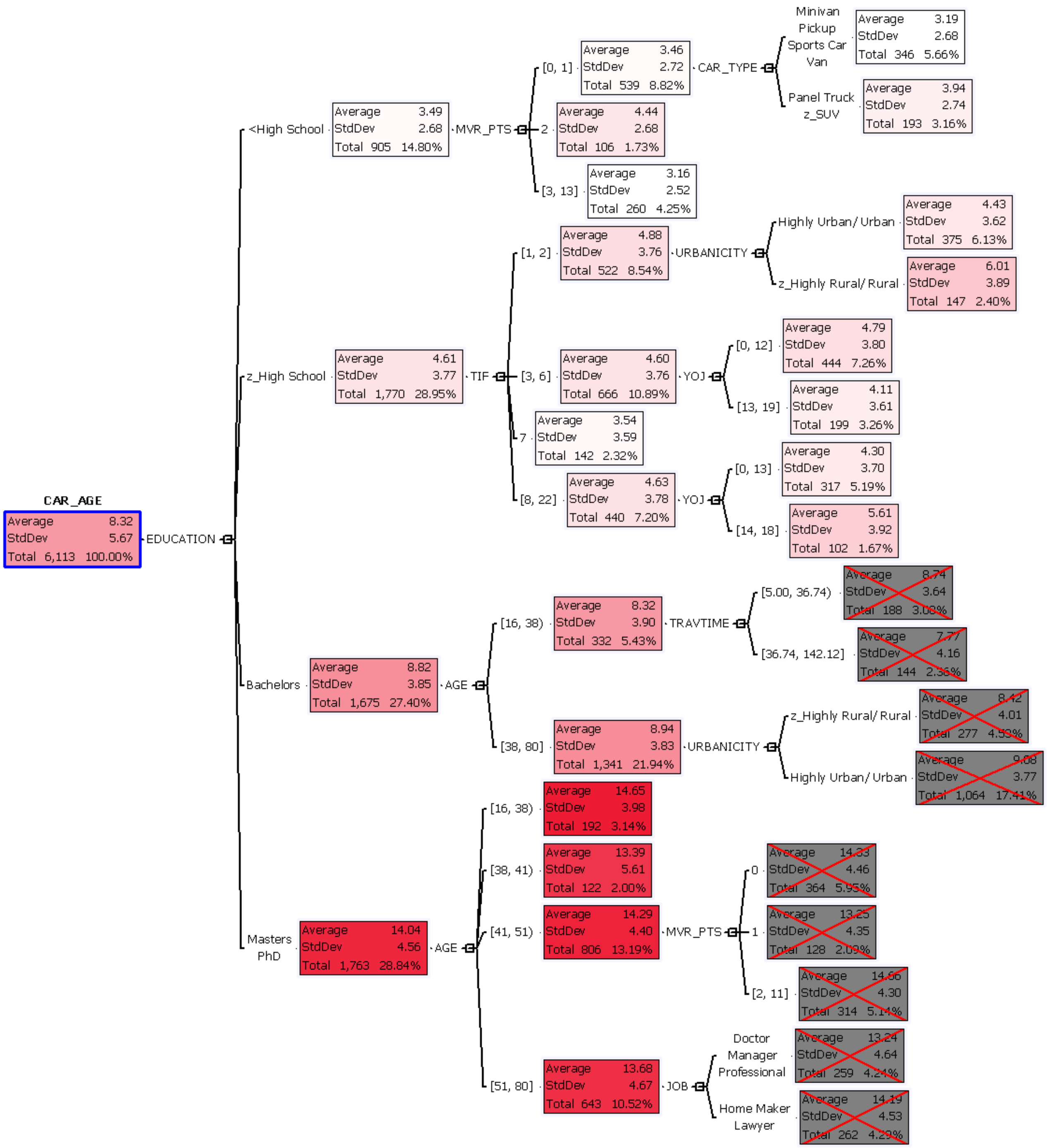
Based off of the EDA, three sets of data preparation steps are taken:

1. Data Imputation – Replace missing values with values computed by analyzing the remainder of the dataset
2. Outlier Elimination – Remove outliers based off of distribution of the data and performance of the regression models
3. Engineered Variables – Derive predictor variables based off of the original variables in the data set
4. Incorrect Categorical Variable – Recategorization of some variables
5. Simplified Categorical Variables – Simplification of some categorical variables

DATA IMPUTATION

CAR_AGE, HOME_VALUE, YOJ, and INCOME all have missing values. These have been addressed by performing data imputation using decision trees run using Angoss KnowledgeStudio. The imputed values are stored as IMP_CAR_AGE, IMP_HOME_VALUE, IMP_YOJ, and IMP_INCOME.

This is an example of a decision tree used to create imputed values for CAR_AGE variable.



To ensure missing values in test datasets are addressed, a few more rules are created based off of the robust average and standard deviations. The robust average is calculated by eliminating the top and bottom 5% of extreme outliers before calculating the mean value. This makes the mean robust against extreme values.

TABLE 4 IMPUTED VALUES FOR VARIABLES

Variable	Imputed Value
Travel Time	32.99 minutes
Sq Root Bluebook Value	120.35
Log Old Claim Value	0

OUTLIER ELIMINATION

Outliers adversely affect the fit of the regression model. Addressing outliers is an iterative process between the Data Preparation and Modeling steps. The process adopted is to eliminate data points that lie greater than 99% bounds for the variable. The cutoffs are shown in the table below.

TABLE 5 99% CUTOFF VALUES VALUES FOR CONTINOUS VARAIABLES

Variable	99% Cutoff
Travel Time	75 min
Sq Root Bluebook Value	200
Imputed Income	$220,000
Imputed Years on the Job	17 years
Imputed Car Age	21 years
Imputed Home Value	$511,660

ENGINEERED VARIABLES

The performance of the models is improved by creating new variables based off of the original dataset. These variables were evaluated in the modeling phase and retained if they added value to the final selected model.

TABLE 6 ENGINEERED VARIABLES

Engineered Variables	Formula	Interpretation
FLAG_HASOLDCLAIM	Claim Frequency > 0	True/False flag if the customer had a previous claim
FLAG_HAVEKIDS	Home Kids > 0	True/False flag if the customer has kids
FLAG_KIDSDRIV	Driving Kids > 0	True/False flag if the customer has kids who drive

FLAG_RENTAL	Home Rental = 0	True/False flag if the customer rents a place instead of owns a home
AMT_PER_CLAIM_LOG	Log (Old Claim $ / Claim Frequency)	How expensive were the previous repairs?
CLM_PER_TIF	Claim Frequency / Time in Force	How many claims did the customer have in the time they were with the insurance company? Higher values indicate a higher risk customer.
AMT_PER_TIF	Old Claim Amount / Time in Force	How expensive were the customer in the time they were with the insurance company? Higher values indicate a high risk customer.
BLUEBOOK_SQRT	Square Root of Bluebook Value	Unknown effect on probability of collision. Square root helps improve the normality of the variable which may increase its predictiveness
OLDCLAIM_LOG	Log of Old Claim Value	Log transformation may increase predictiveness of variable
TIF_BINNED	Binned TIF	Time in Force variable binned
STD_BLUEBOOK	Standardized Bluebook	Bluebook values standardized
STD_IMP_INCOME	Standardized Income	Income values standardized
STD_IMP_HOME_VAL	Standardized Home Value	Home Value standardized
IMP_AGE_BIN	Binned Age	Age binned. New drivers and old drivers are potentially more risky

INCORRECT CATEGORICAL VARIABLE

Investigation of the Car Usage variable split by Car Type shows the original dataset has ~2% Sports Cars classified as Commercial usage. These could be misclassified observations. These are corrected by reclassifying to Private usage in the CAR_USE variable.

TABLE 7 CAR_TYPE DIVIDED BY CAR_USE

Table of CAR_TYPE by CAR_USE			
	CAR_USE(Vehicle Use)		
CAR_TYPE(Type of Car)	Commercial	Private	Total
Minivan	441 5.40 20.56 14.56	1704 20.88 79.44 33.20	2145 26.28
Panel Truck	676 8.28 100.00 22.32	0 0.00 0.00 0.00	676 8.28
Pickup	850 10.42 61.20 28.06	539 6.60 38.80 10.50	1389 17.02
Sports Car	160 1.96 17.64 5.28	747 9.15 82.36 14.56	907 11.11
Van	454 5.56 60.53 14.99	296 3.63 39.47 5.77	750 9.19
z_SUV	448 5.49 19.53 14.79	1846 22.62 80.47 35.97	2294 28.11
Total	3029 37.12	5132 62.88	8161 100.00

SIMPLIFIED CATEGORICAL VARIABLES

Through some of the iterations between model evaluation and data exploration, a few categorical variables were simplified based off the p-values in the model's ANOVA tables. This is discussed further in the modeling section.

The categorical variables simplified are as follows:

TABLE 8 SIMPLIFIED VARIABLES

Variables	Simplification
IMP_JOB	'Other Jobs' = 'Clerical', 'Home Maker', 'Student', 'z_Blue Collar', or 'Lawyer'
EDUCATION	'<HS HS or PhD' = '<High School', 'z_High School', or 'PhD'
CAR_TYPE	'Other Cars' = 'Panel Truck', 'Van', or 'z_SUV'

MODELING & EVALUATION

MODEL BUILDING

The model building phase consisted of exploratory modeling work, multiple models being built and evaluated. The summary of the builds is as follows.

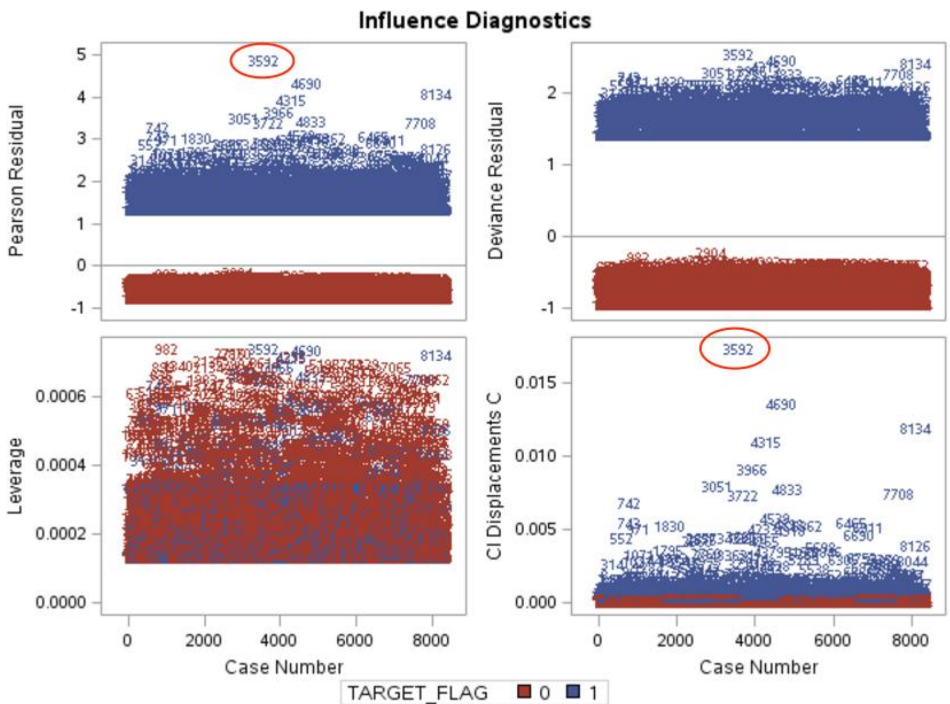
STEP 1: UNIVARIATE EXPLORATORY WORK

As a first step, the original and engineered response variables are regressed on the predictor variables by running univariate logistical regression models. A quick look at the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve, and the confidence interval for the odds are used to sort the potential for the response variables to explain the variation in the Target_Flag variable. The results of this study are shown below.

Variable	Categorical	Univariate Results	ROC AUC	Odds Point Estimate
URBANICITY	X	Very Strong	0.6	6
CAR_TYPE	X	Varied	0.58	0.8
IMP_JOB	X	Varied	0.6	0.7
MVR_PTS		Strong	0.62	1.2
FLAG_HASOLDCLAIM	X	Strong	0.63	0.3
CLM_FREQ		Strong	0.63	1.4
IMP_HOME_VAL		Strong	0.62	1
CAR_USE	X	Good	0.57	2
EDUCATION	X	Good	0.59	0.75
STD_IMP_INCOME		Good	0.57	0.73
FLAG_RENTAL	X	Good	0.57	0.5
MSTATUS	X	Good	0.57	0.5
PARENT1	X	Good	0.56	0.4
CLM_PER_TIF		Good	0.63	1.6
OLDCLAIM_LOG		Good	0.6	1.13
IMP_AGE		Good	0.57	0.9
FLAG_HAVEKIDS	X	Good	0.57	0.55
STD_BLUEBOOK		Decent	0.57	0.7
REVOKED	X	Decent	0.56	0.4
AMT_PER_CLAIM_LOG		Decent	0.63	1.2
IMP_INCOME		Decent	0.6	1
HOMEKIDS		Decent	0.57	1
IMP_CAR_AGE		Decent	0.56	0.95
KIDSDRIV		Poor	0.53	1
TRAVTIME		Poor	0.53	1
TIF		Poor	0.55	0.9
JOB_WHITE_COLLAR	X	Poor	0.53	1.6

TIF_BINNED	X	Poor	0.55	1
TRAVTIME_SQRT		Poor	0.53	1
SEX	X	Poor	0.51	1
RED_CAR	X	Poor	0.5	1
STD_IMP_HOME_VAL		Poor	0.5	1
IMP_YOJ		Poor	0.53	0.96
FLAG_KIDSDRIV	X	Poor	0.53	0.5

For each of these regression models, the outliers are studied using Leverage plots. As a result, some outliers were observed and addressed in the data preparation stage. For example, in the Home Value variable, there was one home (observation # 3592) with an outlier value of $750,455, as observed in the Pearson Residual plot shown below.



STEP 2: FULL MODEL - BASELINE

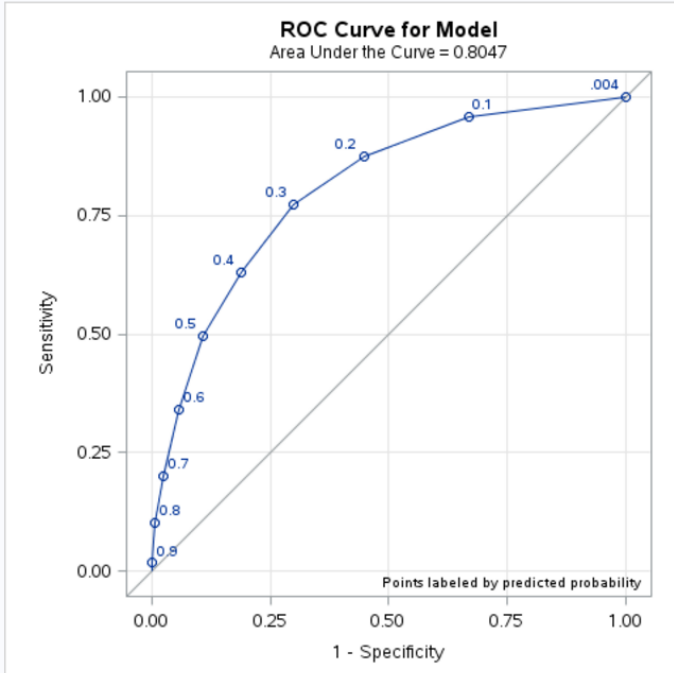
The first model run is performed with all the variables included. This will give us a good estimate of how much variability can be explained with all the variables included, as well as which variables are statistically significant to the model.

A total of 36 variables are input into the model. 18 of the 36 variables have P-values < 0.1 of the Wald Chi-Square test, which measures how relevant the variable is to the model. This gives us an indication of the potential variables important to the final model. Models like HOMEKIDS and RED_CAR are significant at the 0.15 level. These could potentially be important too.

Effect	DF	Wald	Pr > ChiSq
CAR_TYPE	3	60.7592	0.0001
CAR_USE	1	51.4281	0.0001
EDUCATION	2	23.5659	0.0001
IMP_JOB	3	38.2256	0.0001
MVR_PTS	1	40.6523	0.0001
REVOKED	1	36.1639	0.0001
STD_BLUEBOOK	1	14.6662	0.0001
TRAVTIME	1	27.5044	0.0001
URBANICITY	1	242.0758	0.0001
TIF	1	11.6819	0.0006
MSTATUS	1	10.211	0.0014
PARENT1	1	6.6996	0.0096
IMP_AGE_BIN	5	14.4946	0.0128
KIDSDRIV	1	5.6115	0.0178
FLAG_HASOLDCLAIM	1	4.8668	0.0274
SEX	1	3.8	0.0513
STD_IMP_INCOME	1	3.7208	0.0537
FLAG_RENTAL	1	3.1415	0.0763
HOMEKIDS	1	2.3515	0.1252
RED_CAR	1	2.2407	0.1344
CLM_PER_TIF	1	2.2095	0.1372
IMP_CAR_AGE	1	1.15	0.2835
MISS_CAR_AGE	1	1.0269	0.3109
MISS_INCOME	1	0.5527	0.4572
FLAG_KIDSDRIV	1	0.3371	0.5615
MISS_HOME_VAL	1	0.3317	0.5647
FLAG_HAVEKIDS	1	0.3214	0.5708
OLDCLAIM_LOG	1	0.2854	0.5932
MISS_JOB	1	0.1716	0.6787
CLM_FREQ	1	0.0893	0.765
AMT_PER_TIF	1	0.0824	0.7741
AMT_PER_CLAIM_LOG	1	0.0669	0.7958
MISS_YOJ	1	0.0416	0.8383
IMP_YOJ	1	0.0283	0.8665
MISS_AGE	1	0.0023	0.9618
STD_IMP_HOME_VAL	1	0.0022	0.9625

The performance of the model can be quantified by looking the Percent Concordant (76% agreement) and AUC of 80.5%.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.0	Somers' D	0.609
Percent Discordant	15.1	Gamma	0.669
Percent Tied	8.9	Tau-a	0.258
Pairs	3321836	c	0.805



since all the variables are not significantly, automated variable selection methods are used to narrow down the pool of variables.

STEP 3: AUTOMATED VARIABLE SELECTION METHODS

Three automated variable selection methods are attempted – Stepwise, Forward and Backward selection. Multiple versions of a multilinear regression model are built and examined at this step.

For each model, the model is fit to 80% of the data randomly selected. The resulting model is fit to the remaining 20% validation dataset. A summary of all the models evaluated is shown in the table below. The models can be compared using AUC, KS Statistic and % agreement in validation data set.

Model	All Var	Stepwise	Forward	Backward
Num of Params	36	17	17	17
Interpretability	Lowest	Medium	Medium	High
Intercept Only - SC		4875.31		
Int & Var - SC	4143.191	4008.688	4014.911	4013.544

Intercept Only - 2LogL	4867.026			
Int & Var - 2LogL	3762.092	3784.999	3782.938	3781.57
# P-val > 0.05	16	0	0	0
AUC	0.805	0.8023	0.8034	0.8026
KS Statistic	0.439025	0.439025	0.439025	0.439025
Validation Set %Agreement	0.760	0.757	0.760	0.762

Although the All Var model has the highest AUC, the other three models have much lower number of parameters with higher scores in competing performance parameters of SC, -2Log L and the % agreement in the validation set. The backward selection model has the lowest -2LogL score (3781) amongst the three auto selection models, and the highest % agreement score (76.2%). The interpretability is also the highest. The Stepwise and Forward selection methods rely on CLM_PER_TIF, AMT_PER_CLAIM_LOG which require special handling for observations to avoid division by zero. The backward selection does not rely on such a metric.

These are the variables selected by each method are shown here:

Parameters	All Var	Stepwise	Forward	Backward
CAR_TYPE	X	X	X	X
CAR_USE	X	X	X	X
EDUCATION	X	X	X	X
IMP_JOB	X	X	X	X
MVR_PTS	X	X	X	X
REVOKED	X	X	X	X
STD_BLUEBOOK	X	X	X	X
TRAVTIME	X	X	X	X
URBANICITY	X	X	X	X
TIF	X	X	X	X
MSTATUS	X	X	X	X
PARENT1	X	X	X	X
IMP_AGE_BIN	X	X		X
KIDSDRIV	X	X	X	X
FLAG_HASOLDCLAIM	X	X	X	X
STD_IMP_HOME_VAL	X	X	X	
OLDCLAIM_LOG	X		X	X
STD_IMP_INCOME	X			X
CLM_PER_TIF	X		X	
AMT_PER_CLAIM_LOG	X	X		
SEX	X			
FLAG_RENTAL	X			

HOMEKIDS	X
RED_CAR	X
IMP_CAR_AGE	X
MISS_CAR_AGE	X
MISS_INCOME	X
FLAG_KIDSDRIV	X
MISS_HOME_VAL	X
FLAG_HAVEKIDS	X
MISS_JOB	X
CLM_FREQ	X
AMT_PER_TIF	X
MISS_YOJ	X
IMP_YOJ	X
MISS_AGE	X

FINAL MODEL

The final logistic regression model selected to predict the team victories is as follows:

Log Odds = 0.2184
+ REVOKED in "No" * -0.7814
+ CAR_USE in "Commercial" * 0.7811
+ FLAG_RENTAL in "0" * -0.2547
+ FLAG_HASOLDCLAIM in "0" * -1.8107
+ EDUCATION in "<HS HS or PhD" * 0.4919
+ EDUCATION in "Bachelors" * 0.1362
+ MSTATUS in "Yes" * -0.4765
+ PARENT1 in "No" * -0.37
+ URBANICITY in "Highly Urban/ Urban" * 2.3071
+ IMP_JOB in "Doctor" * -0.6887
+ IMP_JOB in "Manager" * -0.661
+ IMP_JOB in "Other Jobs" * 0.1089
+ CAR_TYPE in "Minivan" * -1.1089
+ CAR_TYPE in "Other Cars" * -0.5232
+ CAR_TYPE in "Pickup" * -0.5453
+ STD_BLUEBOOK * -0.188
+ OLDCLAIM_LOG * -0.1672
+ STD_IMP_INCOME * -0.1789
+ TRAVTIME * 0.0135
+ TIF * -0.0516
+ KIDSDRIV * 0.4062
+ MVR_PTS * 0.1057

+ IMP_AGE_BIN in "GE20" * 0.2476
+ IMP_AGE_BIN in "GE25" * -0.3848
+ IMP_AGE_BIN in "GE35" * -0.6043
+ IMP_AGE_BIN in "GE55" * -0.2392
+ IMP_AGE_BIN in "GE65" * -0.6777

This equation does make very intuitive sense for the variables selected in the model.

WHAT INCREASES THE ODDS OF A CRASH?

The highest coefficient is for Urban vs Rural, with a point coefficient of 10.39, which indicates that drivers living in the city have a 10 times higher odds of getting into a car accident than those living in the rural parts of the country.

Those operating a car in the commercial capacity have higher chances of an accident, given the larger amount of time spent on the road. Commercial drivers are twice as likely to get into an accident than Private drivers.

Furthermore, if education is < High School, or High School, the odds of crashing are 1.74 vs those with a Masters only. What's interesting and unexplained is why this is also true for PhD customers.

Having kids who drive increase the odds of a crash by 1.5, while those customers who have points: each additional point increases the odds by 1.12.

If the customer has the license revoked before, it also increases the chances of repeated crashes.

The model also shows that customers driving Sports Cars have the highest chances of crashes compared to those driving Minivans, or other cars. Professionals are also have higher chances of crashes – Being a doctor halfs the risk of a crash, like Managers

CONCLUSION

Many models were successfully evaluated and a top performing model was selected using the Backward selection criteria combined with the performance of a 80-20 validation split. The model performs well with a low error score on the test dataset, and it does not violate any assumptions required for logistic regression modeling.

BINGO BONUS

SAS MACROS:

Main.sas:

```
%let PATH = /folders/myfolders/Assignment2;

%let NAME = unit02;

%let LIB = &NAME..;

LIBNAME &NAME. "&PATH.";

%let INFILE = &LIB.LOGIT_INSURANCE;

%let COMPCASEFILE = &LIB.COMPCASEFILE;

%let TESTFILE = &LIB.logit_insurance_test;

%let TESTCOMPCASEFILE = &LIB.TESTCOMPCASEFILE;

%let FINALSCORES = &LIB.FINALSCORESFILE;

ods graphics on;

%include "/folders/myfolders/Assignment2/processdata.sas";

%include "/folders/myfolders/Assignment2/model_full.sas";
```

FULL_MODEL.SAS:

```
%RemoveNegatives(&INFILE.,&COMPCASEFILE.);


%AgeTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);


%HomeValTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);


%IncomeTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);


%YOJTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);
```

```
%CarAgeTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);

%JobTreeRule(&COMPCASEFILE.,&COMPCASEFILE.);

%DropMissing(&COMPCASEFILE.,&COMPCASEFILE.);

%EngineeredVar(&COMPCASEFILE.,&COMPCASEFILE.);

%CleanCarUsage(&COMPCASEFILE.,&COMPCASEFILE.);

%ClipExtremesAddressMissing(&COMPCASEFILE.,&COMPCASEFILE.);

%AdjustOutliers(&COMPCASEFILE.,&COMPCASEFILE.);

%SimplifyJob(&COMPCASEFILE.,&COMPCASEFILE.);

%SimplifyEducation(&COMPCASEFILE.,&COMPCASEFILE.);

%SimplifyCarType(&COMPCASEFILE.,&COMPCASEFILE.);
```