Assignment 2: Insurance Logistic Regression Project

MSPA PREDICT 411-DL-SEC56

Darryl Buswell

1 Introduction

This document presents results of the second assignment for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to build a number of logistic regression models which are able to predict the probability that a customer will get into a car accident, and to also build a linear regression model which is able to predict the value of damages for those involved in an accident. In order to specify each model, we use a number of automated variable selection techniques and also manually select variables based on our interpretation of those which would be relevant for predicting both target variables. As a final step for this assessment, we present a SAS routine which is able to generate predictions of accident probability and accident value based on a withheld test set of data.

As a bonus for this assessment, we also present a SAS routine in the form of a coded decision tree which is able to generate a number of new variables and correct missing observations. Although the decision tree was not leveraged as part of the data preparation routine, it is hoped that the framework of this routine can be used for future assessments.

2 Data

The dataset contains 8,161 data records, with variables which characterize vehicle owners as well as the vehicles themselves. There are two target variables within the dataset. One which indicates whether the owner was involved in a car accident (TARGET_FLAG) and one which indicates the value of the accident for those who were involved in a car accident (TARGET_AMT). TARGET_FLAG is a categorical variable while TARGET_AMT is a continuous variable.

At a first pass, it seems the dataset has quite a large amount of scope. There are 23 variables tracking a number of attributes. Of the total variable count, 10 are formatted as character type and 13 are formatted as numeric type. The table below shows a list of variables included in the original dataset.

Table 2.1: Variable Descriptions

Original Variable	Renamed Variable	Format	Description
CAR_TYPE	C_CAR_TYPE	Char	Type of Car
CAR_USE	C_CAR_USE	Char	Vehicle Use
EDUCATION	C_EDUCATION	Char	Max Education Level
JOB	C_JOB	Char	Job Category
MSTATUS	C_MSTATUS	Char	Marital Status
PARENT1	C_PARENT1	Char	Single Parent
RED_CAR	C_RED_CAR	Char	A Red Car
REVOKED	C_REVOKED	Char	License Revoked (Past 7 Years)
SEX	C_SEX	Char	Gender
URBANICITY	C_URBANICITY	Char	Home/Work Area
AGE	N_AGE	Num	Age
BLUEBOOK	N_BLUEBOOK	Num	Value of Vehicle
CAR_AGE	N_CAR_AGE	Num	Vehicle Age
CLM_FREQ	N_CLM_FREQ	Num	#Claims(Past 5 Years)
HOMEKIDS	N_HOMEKIDS	Num	#Children @Home
HOME_VAL	N_HOME_VAL	Num	Home Value

Original Variable	Renamed Variable	Format	Description
INCOME	N_INCOME	Num	Income
KIDSDRIV	N_KIDSDRIV	Num	#Driving Children
MVR_PTS	N_MVR_PTS	Num	Motor Vehicle Record Points
OLDCLAIM	N_OLDCLAIM	Num	Total Claims(Past 5 Years)
TIF	N_TIF	Num	Time in Force
TRAVTIME	N_TRAVTIME	Num	Distance to Work
YOJ	N_YOJ	Num	Years on Job

For this assessment, we have renamed each variable according to its format type. Renamed variables can be seen in the table above with either the 'C_' prefix for those formatted as character type, or 'N_' prefix for those formatted as numeric type. Finally, note that while CLM_FREQ, HOMEKIDS and KIDSDRIV are of numeric format, these variables are arguably categorical in nature due to their low bin resolution.

The original dataset also included a data dictionary with a proposed theoretical effect for each variable. The table below summarizes the proposed effects.

Table 2.2: Proposed Effect of Variables

Variable	Theoretical Effect
C_CAR_TYPE	unknown effect on probability of collision but likely effect on payout
C_CAR_USE	commercial vehicles are driven more so may have higher likelihood of collision
C_EDUCATION	unknown effect however more educated people likely drive more safely
C_JOB	in theory white collar jobs tend to be safer
C_MSTATUS	in theory married people tend to drive more safely
C_PARENT1	unknown effect
C_RED_CAR	urban legend is that red cars are more likely to be in accidents
C_REVOKED	if license was revoked then is likely to be a more risky driver
C_SEX	urban legend is that women have less crashes than men
C_URBANICITY	unknown effect
N_AGE	young and old people tend to be risky
N_BLUEBOOK	unknown effect on probability of collision but likely effect on payout
N_CAR_AGE	unknown effect on probability of collision but likely effect on payout
N_CLM_FREQ	the more claims filed in the past the more likely to file in the future
N_HOMEKIDS	unknown effect
N_HOME_VAL	in theory home owners tend to drive more safely
N_INCOME	in theory wealthier people tend to get in fewer accidents
N_KIDSDRIV	teenagers driving the vehicle are more likely to get in crashes
N_MVR_PTS	if you get more tickets you are likely to get in more crashes
N_OLDCLAIM	if total payouts over last period was high then future payouts will likely be high
N_TIF	long time customers are usually more safe
N_TRAVTIME	longer drives to work indicate longer exposure to risk
N_YOJ	longer time spent in the workforce likely tend to drive more safely

We note that a number of variables have an 'unknown effect' and may have little justification for causality with our target variables. These variables will be noted and likely excluded from any manually specified predictive models.

3 Data Exploration

Prior to performing any model building, a number of data exploration routines are conducted. These routines allow us to gain an understanding of any potential limitations of the dataset including identifying variables which have missing observations, outlier observations, or those variables which may benefit from transformation.

3.1 Univariate Data Analysis

Summary statistics for each of the numeric variables is shown in the table below.

Table 3.1.1: Data Statistics

Variable	Minimum	Maximum	Mean	Std Dev	N Miss	N
TARGET_AMT	0	107586.14	1504.32	4704.03	0	8161
N_KIDSDRIV	0	4	0.1710575	0.5115341	0	8161
N_AGE	16	81	44.7903127	8.6275895	6	8155
N_HOMEKIDS	0	5	0.7212351	1.1163233	0	8161
N_YOJ	0	23	10.4992864	4.0924742	454	7707
N_INCOME	0	367030.26	61898.1	47572.69	445	7716
N_HOME_VAL	0	885282.34	154867.29	129123.78	464	7697
N_TRAVTIME	5	142.1206304	33.4887972	15.904747	0	8161
N_BLUEBOOK	1500	69740	15709.9	8419.73	0	8161
N_TIF	1	25	5.351305	4.1466353	0	8161
N_OLDCLAIM	0	57037	4037.08	8777.14	0	8161
N_CLM_FREQ	0	5	0.7985541	1.1584527	0	8161
N_MVR_PTS	0	13	1.695503	2.1471117	0	8161
N_CAR_AGE	-3	28	8.3283231	5.7007424	510	7651

We can see that a number of variables suffer from missing observations and will therefore benefit from some form of imputation. We can also see that N_CAR_AGE has a minimum value of -3. We will need to investigate this variable further to confirm whether this observation is a data error. Finally, we note that a number of variables are shown to have a minimum value of zero with a low mean yet high maximum value. These variables may potentially be zero-inflated and/or have non-normal distributions.

Visualization methods can also be used to gain a greater understanding of each variable. For this assessment, bar plots for each of the character variables were generated and reviewed and likewise, histogram and box plots were generated and reviewed for all numeric variables. We have selected a number of variables for further discussion below.

Figure 3.1.1 Bar Plot: Car Type

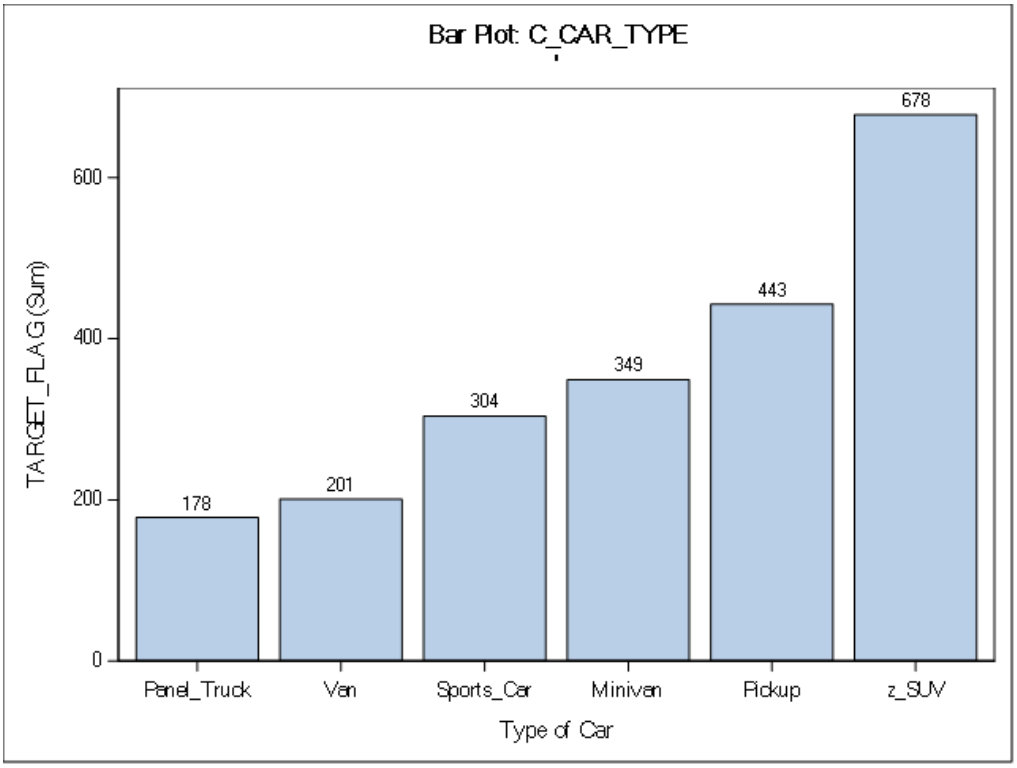


Figure 3.1.2 Bar Plot: Education

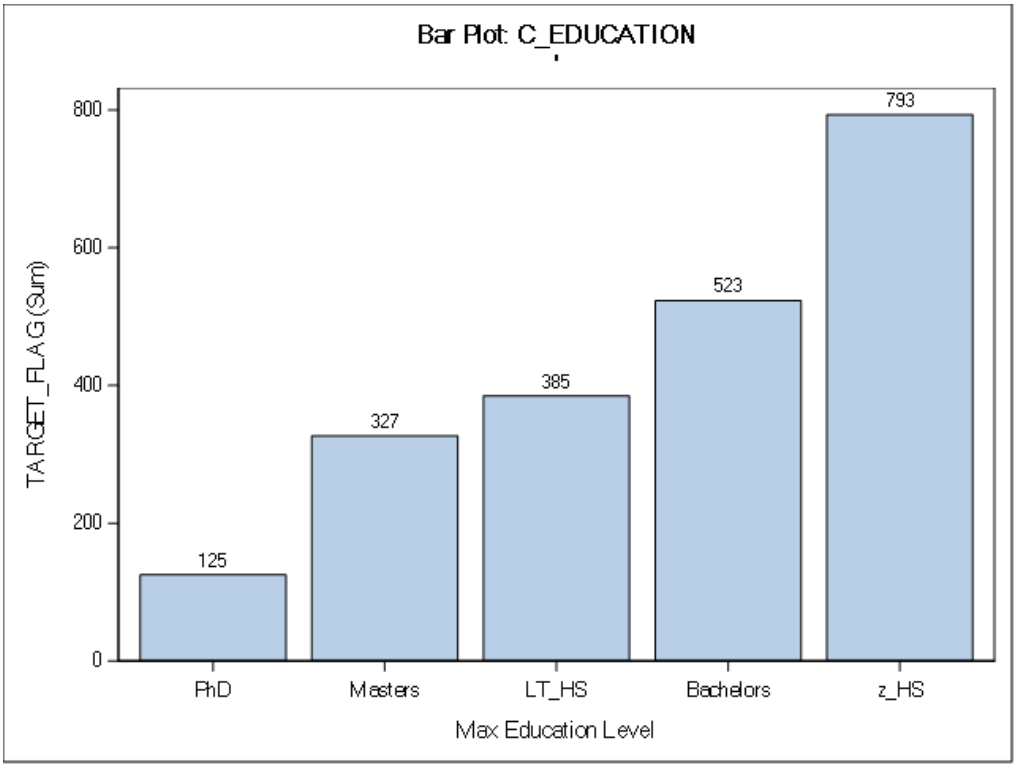


Figure 3.1.3 Bar Plot: Job

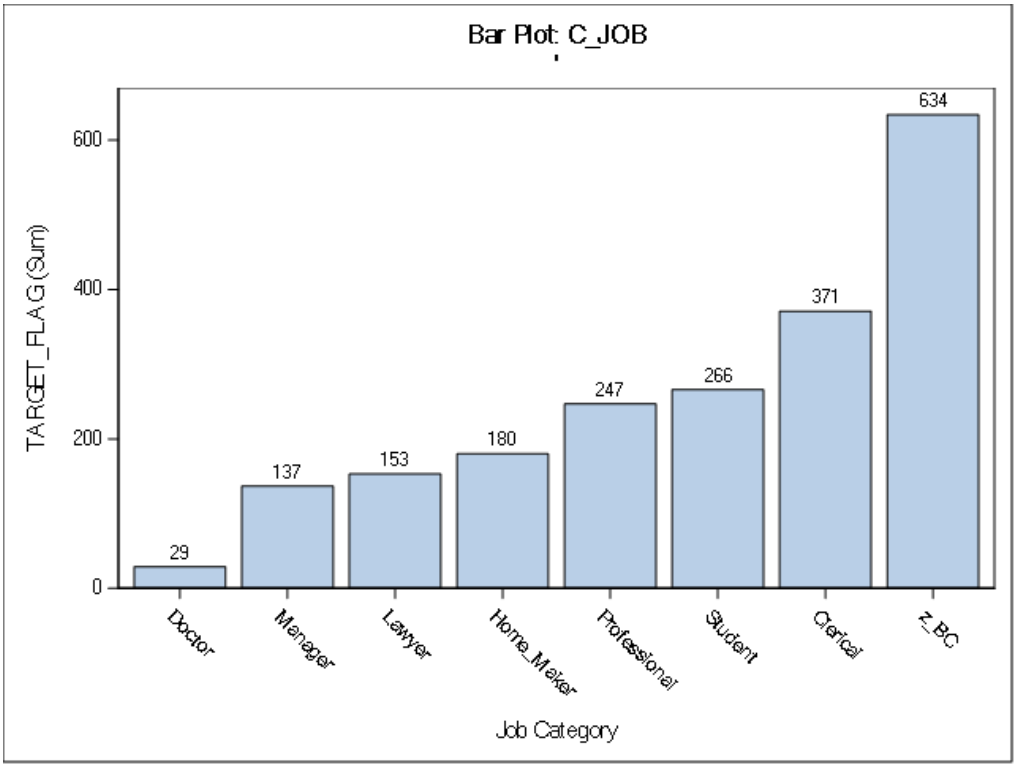


Figure 3.1.4 Bar Plot: Sex

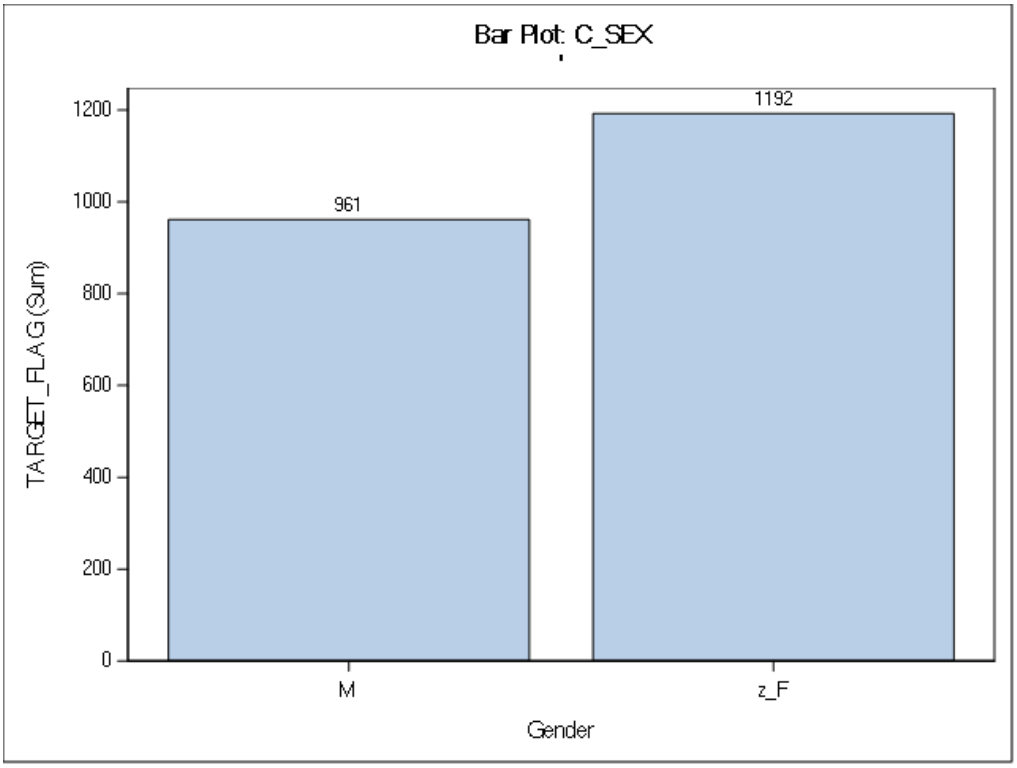


Figure 3.1.5 Histogram: Age

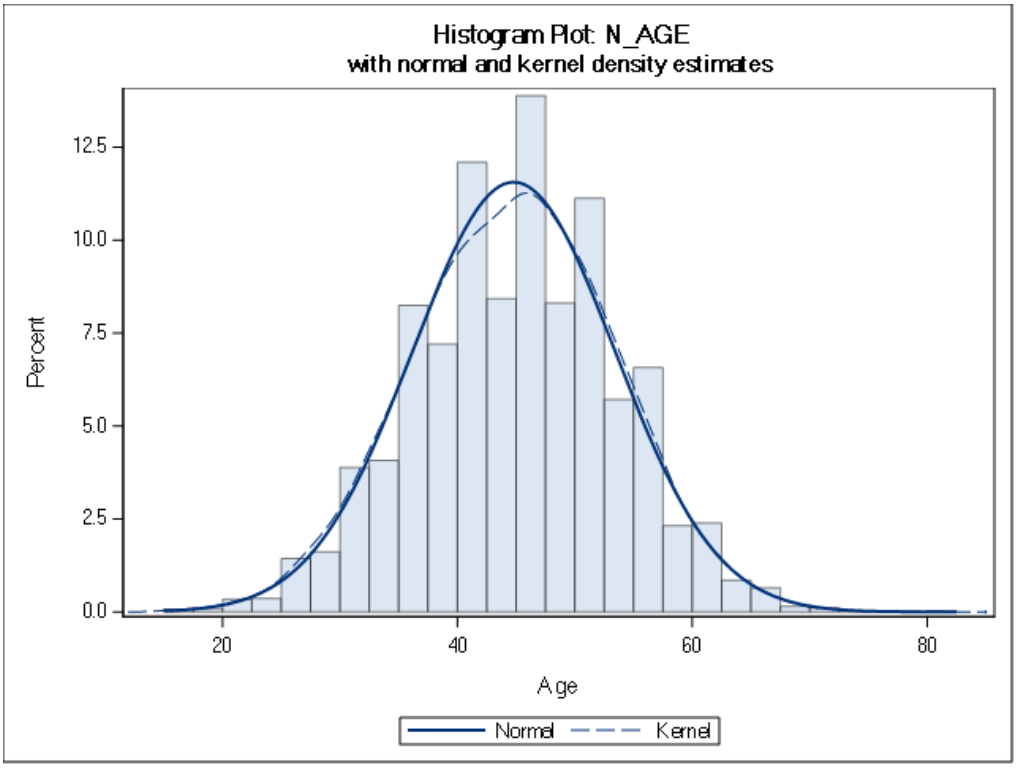


Figure 3.1.6 Box Plot: Age

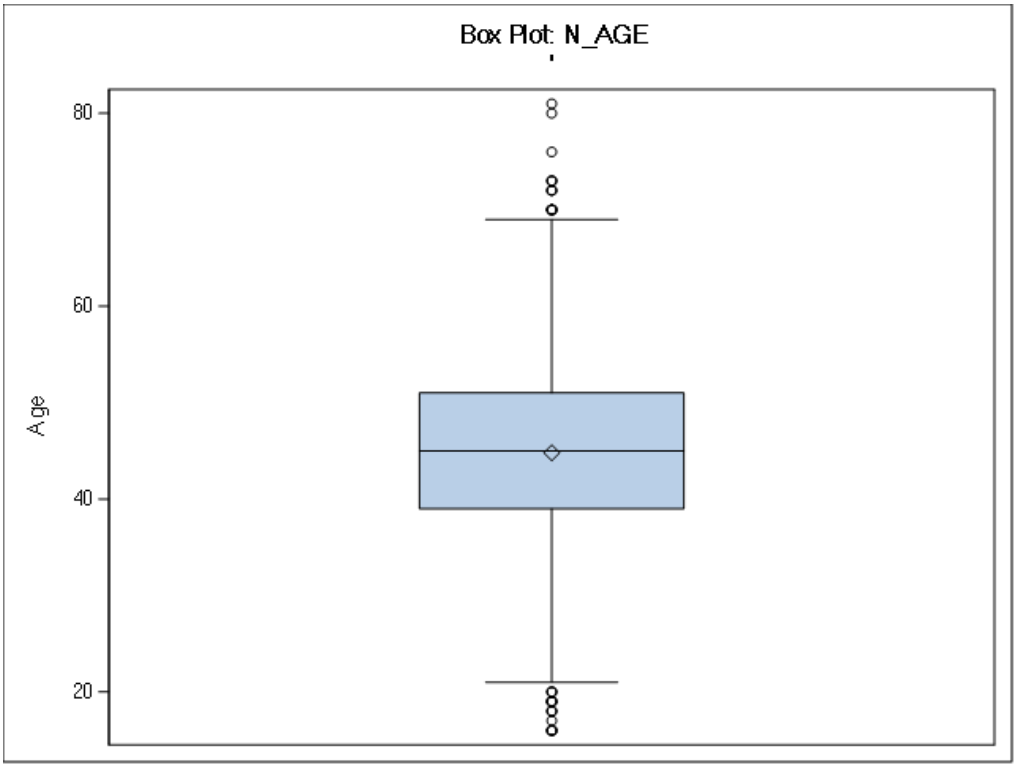


Figure 3.1.7 Histogram: Income

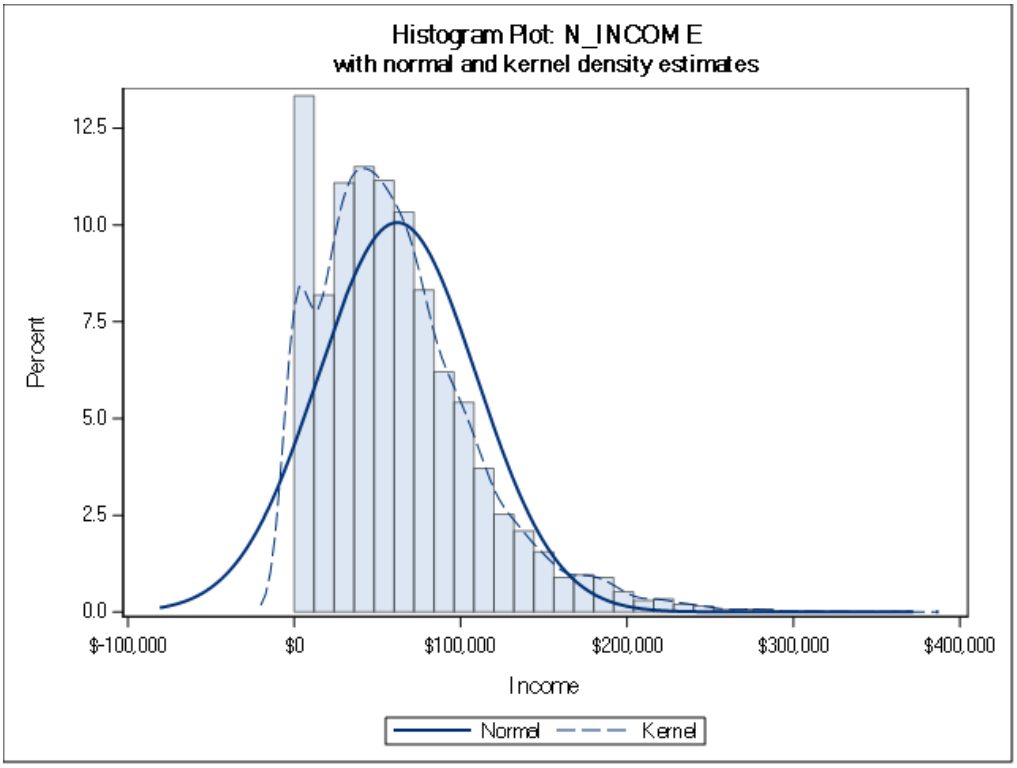
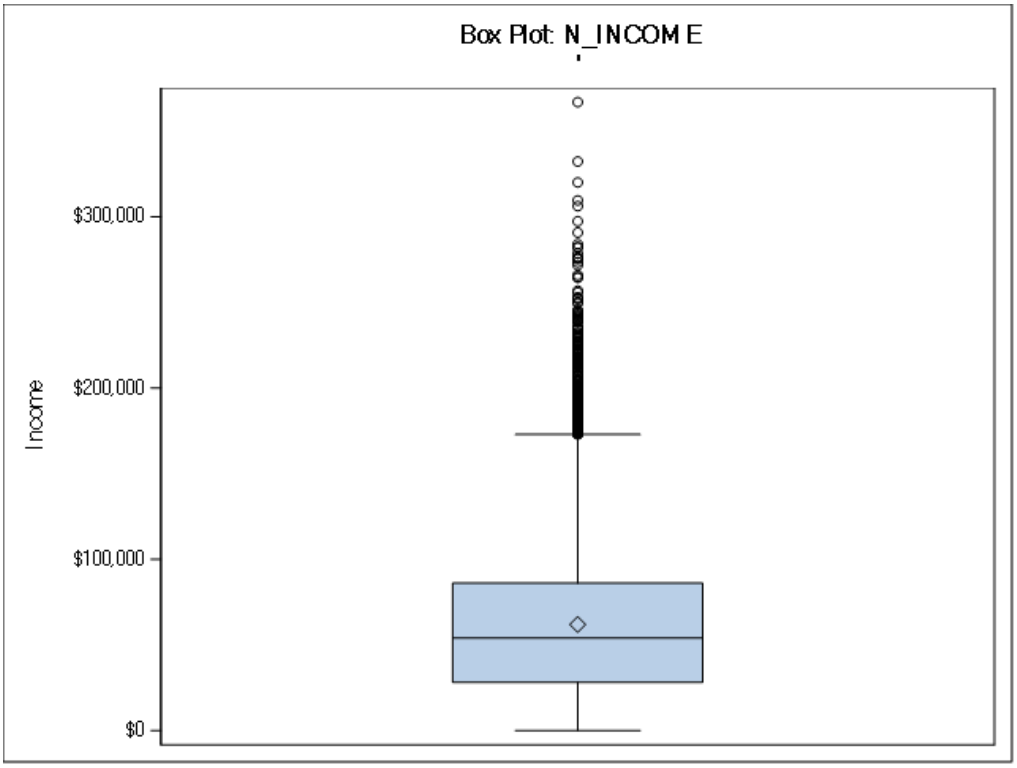


Figure 3.1.8 Box Plot: Income



In terms of vehicle characteristics, we can see that the dataset has a higher representation of those who drive a SUV or pickup vehicle compared to those who drive a van or panel truck. In terms of driver characteristics, we can see that the majority of drivers have a lower level of education, tend to be unemployed or employed in non-professional roles, and are generally female rather than male.

We note that few of the numeric variables exhibit a normal shaped distribution. In-fact, N_AGE was found to have the most normal shaped distribution, while the remaining variables were found to have varying degrees of skew. Our initial concerns of the presence of zero-inflated data were also confirmed from the above exploratory analysis, with many variables, including N_INCOME, having a greater representation of zero values.

3.2 Bivariate Data Analysis

Since we intend on building a prediction model for both the chance of a car accident as well as the value of accident, we have an interest in those variables which have explanatory power over these variables. As such, we use the SAS procedure 'corr' to see if any of the numeric variables have a high Pearson correlation coefficient in relation to accident value. We also generate frequency tables for each of the character variables against the flag which indicates whether there was a car accident.

The table below summarizes the correlation coefficients between accident value and each numeric variable.

Table 3.2.1: Correlations for Accident Value vs. Numeric Data

Variable	Correlation
N_KIDSDRIV	0.05539
N_AGE	-0.04173
N_HOMEKIDS	0.06199
N_YOJ	-0.02209
N_INCOME	-0.05831
N_HOME_VAL	-0.0856
N_TRAVTIME	0.02777
N_BLUEBOOK	-0.0047
N_TIF	-0.04648
N_OLDCLAIM	0.07095
N_CLM_FREQ	0.11642
N_MVR_PTS	0.13787
N_CAR_AGE	-0.05882

None of the continuous variables are reported to have a particularly strong positive or negative correlation coefficient with the response variable, with the greatest absolute correlation being reported by 'N_MVR_PTS' and 'N_CLM_FREQ' at 0.14 and 0.12 respectively. Interestingly, we see a very weak correlation between N_BLUEBOOK and accident value.

The tables below summarize the percentage of observations which coincided with an accident for each level of a number of chosen character variables.

Table 3.2.2: Freq Table for Accident Flag vs. Education

C_EDUCATION	<HighSchool	z_HighSchool	Bachelors	Masters	PhD
% which had accidents	32%	34%	23%	20%	17%

We can see that a greater share of drivers with a low level of education were involved in an accident, compared to drivers with a high level of education.

Table 3.2.3: Freq Table for Accident Flag vs. Job

C_JOB	z_BlueCollar	Clerical	Student	Professional	HomeMaker	Lawyer	Manager	Doctor
% which had accidents	33%	29%	37%	22%	28%	18%	14%	12%

Likewise, we see that a greater share of student drivers and drivers in non-professional employment were involved in an accident.

Table 3.2.4: Freq Table for Accident Flag vs. Urbanicity

C_URBANICITY	Highly Urban/ Urban	z_Highly Rural/Rural
% which had accidents	31%	7%

Perhaps not surprisingly, a greater proportion of accidents occurred in urban areas compared to rural areas.

Table 3.2.5: Freq Table for Accident Flag vs. Revoked

C_REVOKED	Yes	No
% which had accidents	44%	24%

And finally, we can see that those divers who had their licence revoked within the past seven years had a much greater share of accidents.

4 Data Preparation

The data preparation routine for this assessment follows a four step process for numeric variables. This includes 1) identifying and correcting any data errors, 2) trimming variables to account for outliers, 3) imputing variables to account for missing values, and finally, 4) performing a log transformation of all existing and newly created variables. Note that during this process, new dummy variables are created in order to reflect any identified outlier or missing observations.

4.1 Data Errors

We note that the only data error found within this dataset is the -3 observation for N_CAR_AGE. Hence, we have replaced this observation with its absolute value as a first step of the data preparation routine. The N_CAR_AGE variable will include this replaced observation for all subsequently derived variables.

4.2 Data Outliers

From the univariate and bivariate analysis, we have identified that the majority of numeric variables are in-fact zero-inflated and/or suffer from outlier observations. A review of the percentiles for each variable confirms this, with a rather large gap between the min, max and the 1st and 99th percentile for each variable. A summary of percentiles for each variable can be found in the table below.

Table 4.1: Quantiles Summary

	Min	0.01	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.99	Max
N_KIDSDRIV	0	0	0	0	0	0	0	1	1	2	4
N_AGE	16	25	30	34	39	45	51	56	59	64	81
N_HOMEKIDS	0	0	0	0	0	0	1	3	3	4	5
N_YOJ	0	0	0	5	9	11	13	15	15	17	23
N_INCOME	0	0	0	4362	28094	54028	86021	123217	152283	215536	367030
N_HOME_VAL	0	0	0	0	0	161160	238724	316587	374931	500309	885282
N_TRAVTIME	5	5	7	13	22	33	44	54	60	75	142
N_BLUEBOOK	1500	1500	4900	6000	9280	14440	20850	27460	31110	39090	69740
N_TIF	1	1	1	1	1	4	7	11	13	17	25
N_OLDCLAIM	0	0	0	0	0	0	4636	9583	27090	42820	57037
N_CLM_FREQ	0	0	0	0	0	0	2	3	3	4	5
N_MVR_PTS	0	0	0	0	0	1	3	5	6	8	13
N_CAR_AGE	-3	1	1	1	1	8	12	16	18	21	28

For this assessment, we have elected to generate trimmed copies of each numeric variable by their 1st/99th and 10th/90th percentiles. Trimmed variables include the suffix '_T99' and '_T90' respectively. Based on the quantile summary above, we note that the majority of variables which are trimmed by their 1st/99th percentiles will retain zero value observations, while the majority of variables which are trimmed by their 10th/90th percentiles will see an elimination of zero value observations. A new set of dummy variables are also created in order to capture those variables which were identified as having outlier observations according to the percentile threshold discussed above. These dummy variables include the suffix '_OF'.

4.3 Missing Data

Following introducing copies of trimmed numeric variables, we then look towards imputing values for missing observations. By recalculating statistical measures for each variable after trimming, we are able to avoid imputing skewed values for those variables which have been trimmed into each variable. For this assessment, we generate new imputed variables based on that variable's median value. Note that in order to simplify the SAS logic used for this assessment, all variables will include the suffix '_IME', however only those variables shown to have missing observations in the previous sections have actually received imputation. A new set of dummy variables are also created in order to capture those variables which were identified as having missing observations. These variables include the suffix '_MF'.

4.4 Data Transformation

We perform a natural logarithm transformation of each of the numeric variables. Variables which have been transformed include the suffix '_LN'. Such a transformation will help penalize extreme values and may provide an improved fit within subsequent regression models. This transformation is performed for all of the newly created non-dummy variables discussed above.

4.5 Dummy Variables

Finally, we create a number of dummy variables based on bins of numeric variables and add these to the prepared numeric variables discussed above. The dummy variables created for this assessment are detailed in the table below.

Table 4.2: Dummy Variable Summary

Variable	Criteria
N_AGE_Risk_Yes	(N_AGE <= 30
N_AGE_Risk_No	(N_AGE_Risk_Yes = 0)
N_BLUEBOOK_Hi	(N_BLUEBOOK >= 27000)
N_BLUEBOOK_Lo	(N_BLUEBOOK_Hi = 0)
N_CLM_FREQ_No	(N_CLM_FREQ = 0)
N_CLM_FREQ_Yes	(N_CLM_FREQ > 0)
N_CLM_FREQ_Hi	(N_CLM_FREQ >= 2)
N_CLM_FREQ_Lo	(N_CLM_FREQ_Hi = 0)
N_HOMEKIDS_No	(N_HOMEKIDS = 0)
N_HOMEKIDS_Yes	(N_HOMEKIDS > 0)
N_INCOME_No	(N_INCOME = 0)
N_INCOME_Yes	(N_INCOME > 0)
N_INCOME_Hi	(N_INCOME >= 85000)
N_INCOME_Lo	(N_INCOME_Hi = 0)
N_KIDSDRIV_No	(N_KIDSDRIV = 0)
N_KIDSDRIV_Yes	(N_KIDSDRIV > 0)
N_MVR_PTS_No	(N_MVR_PTS = 0)
N_MVR_PTS_Yes	(N_MVR_PTS > 0)
N_MVR_PTS_Hi	(N_MVR_PTS >= 4)
N_MVR_PTS_Lo	(N_MVR_PTS_Hi = 0)
N_OLDCLAIM_No	(N_OLDCLAIM = 0)
N_OLDCLAIM_Yes	(N_OLDCLAIM > 0)
N_OLDCLAIM_Hi	(N_OLDCLAIM >= 9500)
N_OLDCLAIM_Lo	(N_OLDCLAIM_Hi = 0)
N_RENTER_Yes	(N_HOME_VAL <= 14400)
N_RENTER_No	(N_RENTER_Yes = 0)
N_TRAVTIME_Hi	(N_TRAVTIME >= 50)
N_TRAVTIME_Lo	(N_TRAVTIME_Hi = 0)

Note that the newly created dummy variables include an appropriate suffix to reflect its criteria.

5 Model Development

For this section, we build two different classes of regression model. First, we build a linear regression model in order to predict the value of an accident. Second, we build three logistic regression models in order to predict the probability of a car accident.

5.1 Model 0: Linear Regression (Stepwise Selection Model)

From a pool of 13 numeric variables, only N_BLUEBOOK and N_CAR_AGE would seem to have reasonable justification for predicting accident value, with the remaining 11 variables either having a greater jusitification for predicting the probability of an accident or having little justification for predicting either the value or probability of an accident. As such, this assessment initially attempted to build a linear regression model using only these two variables. However the fit was found to be unsatisfactory both in terms of its goodness-of-fit and due to failure of a number of OLS assumptions. As an alternative, this assessment has elected to use an automated variable technique based on stepwise selection to specify the linear regression model (Model_LinR_S).

For the linear regression model with stepwise selection (Model_LinR_S), we elected to use a SLENTRY value of 0.10. This indicates that variables should only be added to the specification if they have a significance level

(p-value) of less than 10%. We also elected to use a SLSTAY value of 0.10, which indicates that variables should not be removed from the specification if they have a significance level (p-value) less than 10%.

Parameter estimates for the Model_LinR_S are shown below.

Table 5.1.1: Linear Regression (Stepwise Selection Model) Parameter Estimates

Variable	DF	Est.	S.E.	t Value	$Pr >	t
Intercept	1	607.45408	407.17719	1.49	0.1358	0
N_MVR_PTS_OF	1	1033.99599	587.13791	1.76	0.0783	1.1402
N_AGE_Risk_Yes	1	515.35611	180.30355	2.86	0.0043	1.05295
N_CLM_FREQ_No	1	-759.93781	119.31081	-6.37	<.0001	1.29417
N_HOMEKIDS_Yes	1	287.91233	126.02274	2.28	0.0224	1.38916
N_INCOME_Lo	1	363.19203	132.53969	2.74	0.0062	1.23457
N_KIDSDRIV_Yes	1	569.54126	181.63075	3.14	0.0017	1.33804
N_RENTER_Yes	1	548.74609	108.68295	5.05	<.0001	1.01356
N_BLUEBOOK_IME	1	0.01586	0.00645	2.46	0.0139	1.12956
N_CAR_AGE_T90_IME	1	-33.37866	9.94921	-3.35	0.0008	1.1553
N_MVR_PTS_IME	1	175.36712	28.57688	6.14	<.0001	1.44367
N_TIF_IME_LN	1	-271.31906	72.49254	-3.74	0.0002	1.00219
N_TRAVTIME_T99_IME_LN	1	248.96682	90.88666	2.74	0.0062	1.0027

For Model_LinR_S, the majority of coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis and conclude that each have non-zero coefficients. The only exception is the coefficient estimate for the outlier flag N_MVR_PTS_OF. However, it is difficult to assess the polarity of coefficient estimates, as many of the included variables have little justification for predicting the value of a car accident.

Goodness-of-fit information for Model_LinR_S is shown below.

Table 5.1.2: Linear Regression Stepwise Selection Model) Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	7178263642	598188637	28.11	<.0001
Error	8148	1.73385E+11	21279474		
Corrected Total	8160	1.80563E+11			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics for Model_LinR_S are shown below.

Table 5.1.3: Linear Regression (Stepwise Selection Model) Performance Metrics

Measure	Statistic	Measure	Statistic
MSE	21279474	R-Square	0.0398
MAE	2033.08	Adj R-Sq	0.0383
Root MSE	4612.96797	C(p)	13
Dependent Mean	1504.32465	AIC	137715.611
Coeff Var	306.6471	BIC	137717.653

The R-square value above suggests that Model_LinR_S explains only ~4% of the variability in TARGET_AMT using each of the included predictor variables. The adjusted R-squared value also indicates a similar level of explanatory power.

5.2 Model 1: Logistic Regression (Subjective Model)

For the first logistic regression model (Model_LogR_Subj), we include only those variables which have a reasonable justification for predicting the probability of a car accident. We build on this by including missing variable flags for retained variables which were shown to have missing variables and include outlier flags for retained variables which were shown to have a large amount of outliers. Finally, we removed any variables from this specification which were found to be highly insignificant.

Parameter estimates for the Model_LogR_Subj are shown below.

Table 5.2.1: Logistic Regression (Subjective Model) Parameter Estimates

Parameter		DF	Est.	S.E.	Wald ChiSq	Pr > ChiSq
Intercept		1	-1.5376	0.2464	38.9366	<.0001
C_CAR_USE	Commercial	1	0.7095	0.0779	82.9619	<.0001
C_EDUCATION	Bachelors	1	-0.4024	0.0832	23.4094	<.0001
C_EDUCATION	LT_HS	1	-0.0356	0.092	0.1497	0.6988
C_EDUCATION	Masters	1	-0.3458	0.1465	5.5672	0.0183
C_EDUCATION	PhD	1	-0.0547	0.2025	0.0729	0.7871
C_JOB	Clerical	1	0.0825	0.1034	0.6366	0.425
C_JOB	Doctor	1	-0.8406	0.2974	7.9879	0.0047
C_JOB	Home_Maker	1	0.0623	0.1441	0.1871	0.6653
C_JOB	Lawyer	1	-0.1798	0.185	0.9449	0.331
C_JOB	Manager	1	-0.8386	0.1343	39.014	<.0001
C_JOB	Professional	1	-0.1728	0.1139	2.3034	0.1291
C_JOB	Student	1	0.0464	0.1216	0.1454	0.7029
C_MSTATUS	Yes	1	-0.7483	0.0611	149.7778	<.0001
C_REVOKED	No	1	-0.9076	0.0934	94.462	<.0001
C_URBANICITY	Urban	1	2.2928	0.1116	422.4525	<.0001
N_AGE_IME		1	-0.00958	0.00359	7.1237	0.0076
N_CLM_FREQ_IME		1	0.2169	0.0294	54.5098	<.0001
N_INCOME_IME		1	-0.00000794	0.00000118	45.2549	<.0001
N_KIDSDRIV_IME		1	0.4864	0.0631	59.4666	<.0001
N_MVR_PTS_IME		1	0.1123	0.0149	56.7098	<.0001
N_OLDCLAIM_IME		1	-0.00002	0.000004046	14.4407	0.0001
N_TIF_IME		1	-0.0531	0.00754	49.6228	<.0001
N_TRAVTIME_IME		1	0.0142	0.00192	54.7967	<.0001
N_AGE_MF		1	2.4375	1.1897	4.1974	0.0405
N_INCOME_MF		1	-1.1339	0.5248	4.6686	0.0307
N_INCOME_OF		1	1.1128	0.5118	4.7269	0.0297
N_KIDSDRIV_OF		1	-0.5083	0.347	2.1456	0.143
N_MVR_PTS_OF		1	0.5927	0.3271	3.2841	0.07

We interpret coefficients as the variables' influence on the likelihood of a crash. That is, an increase in the value of any variable with a positive coefficient is estimated to result in an increase in likelihood of a crash. With this in mind, we are critical of a number of coefficient estimates for the subjective model. For instance, we would not expect an increase in the payout of previous claims (N_OLDCLAIM_IME) to result in a decrease in likelihood of a crash.

Goodness-of-fit information for Model_LogR_Subj is shown below.

Table 5.2.2: Logistic Regression (Subjective Model) Test of Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1870.7673	28	<.0001
Score	1673.6328	28	<.0001
Wald	1271.9264	28	<.0001

We compute the Kolmogorov-Smirnov (KS) statistic for the model by performing a random sampling for test and validation using the SAS NPAR1WAY procedure. The results, along with model performance criteria are shown below.

Table 5.2.3: Logistic Regression (Subjective Model) Performance Metrics

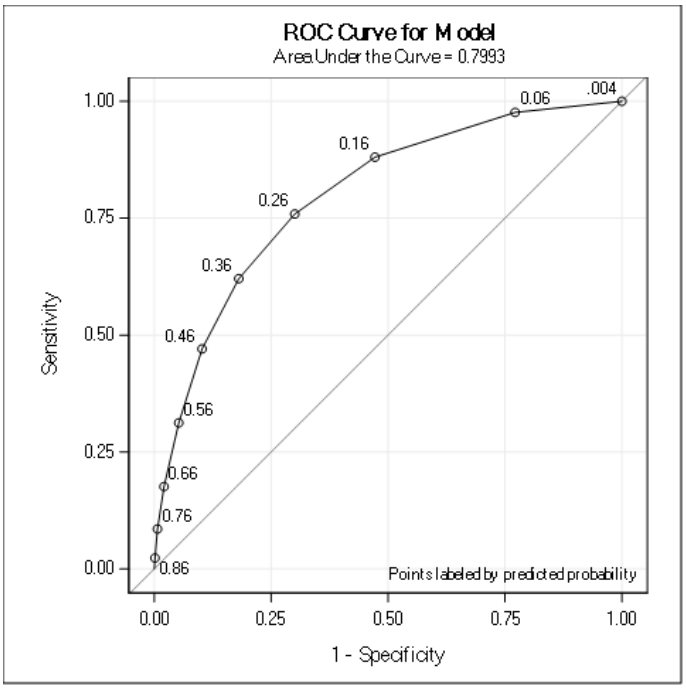
Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	7003.854
SC	8825.562	7205.129
-2 Log L	8816.622	6945.854

Measure	Statistic	Measure	Statistic
KS	0.203568	D	0.461715
KSa	17.787425	Pr > KSa	<.0001

These statistics will be helpful in drawing a comparison between alternative model estimations.

Finally, we can observe the ROC curve for this model.

Figure 5.2.1 ROC Curve (Subjective Model)



The ROC curve suggests that Model_LogR_Subj has 0.7993 coverage and lies significantly above the diagonal.

5.3 Model 2: Logistic Regression (Forward Selection Model)

We leverage automated variables selection techniques to specify the remaining two logistic regression models. For the first, we use the forward selection technique for variable selection (Model_LogR_F). Note that we elected to use a SLENTRY value of 0.15 as part of this technique, which indicates that variables should only be added to the specification if their coefficient estimation has a significance level (p-value) less than 15%.

Parameter estimates for Model_LogR_F are shown below.

Table 5.3.1: Logistic Regression (Forward Selection Model) Parameter Estimates

Parameter		DF	Est.	S.E.	Wald ChiSq	Pr > ChiSq
Intercept		1	2.2393	0.6232	12.9132	0.0003
C_PARENT1	No	1	-0.1633	0.1265	1.6672	0.1966
C_MSTATUS	Yes	1	-0.5654	0.0942	36.0183	<.0001
C_EDUCATION	Bachelors	1	-0.3997	0.0851	22.0644	<.0001
C_EDUCATION	LT_HS	1	-0.00154	0.0963	0.0003	0.9872
C_EDUCATION	Masters	1	-0.3613	0.1504	5.7738	0.0163
C_EDUCATION	PhD	1	-0.1091	0.2029	0.2892	0.5907
C_JOB	Clerical	1	0.0894	0.1081	0.684	0.4082
C_JOB	Doctor	1	-0.944	0.3028	9.7165	0.0018
C_JOB	Home_Maker	1	-0.3276	0.1679	3.8083	0.051
C_JOB	Lawyer	1	-0.1902	0.1919	0.9825	0.3216
C_JOB	Manager	1	-0.9099	0.1419	41.103	<.0001
C_JOB	Professional	1	-0.1775	0.1218	2.1233	0.1451
C_JOB	Student	1	-0.4303	0.1483	8.4246	0.0037
C_CAR_USE	Commercial	1	0.7771	0.0949	67.0621	<.0001
C_CAR_TYPE	Minivan	1	-0.7607	0.0882	74.3163	<.0001
C_CAR_TYPE	Panel_Truck	1	-0.12	0.174	0.4755	0.4905
C_CAR_TYPE	Pickup	1	-0.0981	0.0967	1.0278	0.3107
C_CAR_TYPE	Sports_Car	1	0.1848	0.1005	3.3811	0.0659
C_CAR_TYPE	Van	1	-0.1662	0.1304	1.6258	0.2023
C_REVOKED	No	1	-0.9284	0.0974	90.8318	<.0001
C_URBANICITY	Urban	1	2.3624	0.1141	428.3742	<.0001
N_AGE_MF		1	1.6736	1.2018	1.9392	0.1638
N_CAR_AGE_MF		1	0.2066	0.1237	2.7884	0.0949
N_AGE_Risk_Yes		1	0.5679	0.1005	31.91	<.0001
N_BLUEBOOK_Hi		1	-0.5269	0.1636	10.3686	0.0013
N_CLM_FREQ_Yes		1	0.6429	0.0823	61.0209	<.0001
N_HOMEKIDS_No		1	-0.3472	0.1399	6.1614	0.0131
N_INCOME_Hi		1	-0.3739	0.1067	12.2746	0.0005
N_MVR_PTS_Hi		1	-0.4519	0.1327	11.5988	0.0007
N_BLUEBOOK_T90_IME		1	0.000017	0.000007	5.0557	0.0245
N_HOMEKIDS_T99_IME		1	-0.1043	0.0551	3.5836	0.0584
N_HOME_VAL_T99_IME		1	-0.0000003	7.108E-07	0.2316	0.6304
N_MVR_PTS_IME		1	0.1617	0.0256	39.8646	<.0001
N_OLDCLAIM_IME		1	-0.00002	0.0000044	22.6314	<.0001
N_TRAVTIME_T99_IME		1	0.0169	0.00208	65.6788	<.0001
N_YOJ_T99_IME		1	0.0174	0.0113	2.4011	0.1213
N_BLUEBOOK_T99_IME_L		1	-0.3568	0.0656	29.6028	<.0001
N_HOME_VAL_T99_IME_L		1	-0.0238	0.0137	3.0299	0.0817
N_INCOME_T99_IME_LN		1	-0.08	0.0183	19.0906	<.0001

Parameter	DF	Est.	S.E.	Wald ChiSq	Pr > ChiSq
N_KIDSDRIV_IME_LN	1	0.7506	0.1129	44.2155	<.0001
N_TIF_IME_LN	1	-0.3262	0.0436	56.0911	<.0001

We can see that the high SLENTRY value has resulted in including a number of insignificant variables within the specification. We are again cautious in the interpretation of a number of variables noting their unexpected polarity.

Goodness-of-fit information for Model_LogR_F is shown below.

Table 5.3.2: Logistic Regression (Forward Selection Model) Test of Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2177.8624	41	<.0001
Score	1904.0449	41	<.0001
Wald	1388.3069	41	<.0001

As with the previous model, we show performance statistics for Model_LogR_F below. We can see an improvement in both the AIC and KS statistics for this model, even though it has a much higher variable count.

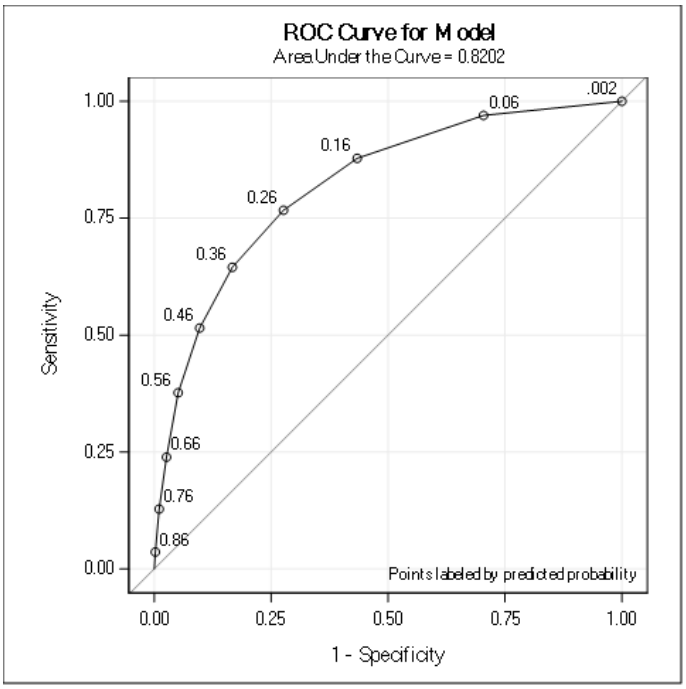
Table 5.3.3: Logistic Regression (Forward Selection Model) Performance Metrics

Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	6722.759
SC	8825.562	7014.26
-2 Log L	8816.622	6638.759

Measure	Statistic	Measure	Statistic
KS	0.219433	D	0.497699
KSa	19.173695	Pr > KSa	<.0001

Finally, we can observe the ROC curve for this model.

Figure 5.3.1 ROC Curve (Forward Selection Model)



The ROC curve suggests that Model_LogR_F has 0.8202 coverage and lies significantly above the diagonal. This is an improvement over the ROC curve for Model_LogR_Subj.

5.4 Model 3: Logistic Regression (Stepwise Selection Model)

For the final logistic regression model, we again use the stepwise technique for variable selection (Model_LogR_S). Note that we elected to use a SLENTRY value of 0.02 as part of this technique, which indicates that variables should only be added to the specification if they have a significance level (p-value) of less than 2%. We also elected to use a SLSTAY value of 0.02, which indicates that variables should not be removed from the specification if they have a significance level (p-value) less than 2%.

Parameter estimates for Model_LogR_S are shown below.

Table 5.4.1: Logistic Regression (Stepwise Selection Model) Parameter Estimates

Parameter		DF	Est.	S.E.	Wald ChiSq	Pr > ChiSq
Intercept		1	2.0478	0.6079	11.3476	0.0008
C_PARENT1	No	1	-0.3078	0.1	9.4681	0.0021
C_MSTATUS	Yes	1	-0.5291	0.0839	39.7965	<.0001
C_EDUCATION	Bachelors	1	-0.3933	0.0847	21.5486	<.0001
C_EDUCATION	LT_HS	1	-0.013	0.0958	0.0183	0.8923
C_EDUCATION	Masters	1	-0.3638	0.1499	5.8884	0.0152
C_EDUCATION	PhD	1	-0.1052	0.2028	0.2693	0.6038
C_JOB	Clerical	1	0.0749	0.1071	0.4899	0.484
C_JOB	Doctor	1	-0.9423	0.3028	9.6875	0.0019
C_JOB	Home_Maker	1	-0.3622	0.1652	4.8072	0.0283
C_JOB	Lawyer	1	-0.1861	0.1915	0.9436	0.3313
C_JOB	Manager	1	-0.912	0.1415	41.5138	<.0001
C_JOB	Professional	1	-0.1776	0.1215	2.1372	0.1438
C_JOB	Student	1	-0.3516	0.1441	5.9505	0.0147
C_CAR_USE	Commercial	1	0.7603	0.0945	64.7697	<.0001

Parameter		DF	Est.	S.E.	Wald ChiSq	Pr > ChiSq
C_CAR_TYPE	Minivan	1	-0.764	0.0881	75.2865	<.0001
C_CAR_TYPE	Panel_Truck	1	-0.1189	0.1735	0.4699	0.493
C_CAR_TYPE	Pickup	1	-0.1055	0.0964	1.1959	0.2741
C_CAR_TYPE	Sports_Car	1	0.1737	0.1003	3.0001	0.0833
C_CAR_TYPE	Van	1	-0.1629	0.1299	1.5726	0.2098
C_REVOKED	No	1	-0.9242	0.0972	90.3173	<.0001
C_URBANICITY	Urban	1	2.3598	0.114	428.3237	<.0001
N_AGE_Risk_Yes		1	0.6004	0.0994	36.4633	<.0001
N_BLUEBOOK_Hi		1	-0.5338	0.1636	10.6516	0.0011
N_CLM_FREQ_Yes		1	0.6365	0.0822	60.0212	<.0001
N_INCOME_Hi		1	-0.3263	0.1001	10.6235	0.0011
N_MVR_PTS_Hi		1	-0.4492	0.1324	11.5101	0.0007
N_BLUEBOOK_T90_IME		1	0.000018	0.0000075	5.7881	0.0161
N_HOME_VAL_T99_IME		1	-0.000001	3.784E-07	14.0791	0.0002
N_MVR_PTS_IME		1	0.162	0.0255	40.2604	<.0001
N_OLDCLAIM_IME		1	-0.00002	0.0000044	21.8101	<.0001
N_TRAVTIME_T99_IME		1	0.0168	0.00208	65.1491	<.0001
N_BLUEBOOK_T99_IME_L		1	-0.3661	0.0652	31.5458	<.0001
N_INCOME_T99_IME_LN		1	-0.0597	0.0143	17.4061	<.0001
N_KIDSDRIV_IME_LN		1	0.7976	0.1023	60.7396	<.0001
N_TIF_IME_LN		1	-0.3233	0.0435	55.3218	<.0001

Unlike the previous forward selection technique, the stepwise selection technique has resulted in the inclusion of less insignificant variables. Also unlike the manually specified model, the automated variable selection techniques clearly have a greater preference for inclusion of dummy variables created as part of the data preparation routine.

Goodness-of-fit information for Model_LogR_S is shown below.

Table 5.4.2: Logistic Regression (Stepwise Selection Model) Test of Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2162.1458	35	<.0001
Score	1888.787	35	<.0001
Wald	1381.8523	35	<.0001

Model performance statistics for Model_LogR_S are shown below. Performance statistics are quite similar to those of the forward selection technique.

Table 5.4.3: Logistic Regression (Stepwise Selection Model) Performance Metrics

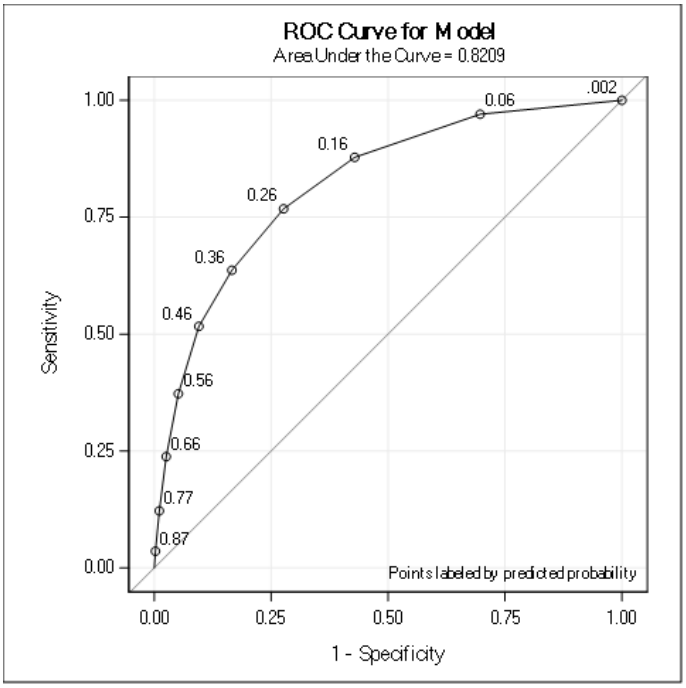
Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	6726.476
SC	8825.562	6976.334
-2 Log L	8816.622	6654.476

Measure	Statistic	Measure	Statistic
KS	0.218525	D	0.495639

Measure	Statistic	Measure	Statistic
KSa	19.094351	Pr > KSa	<.0001

Finally, we can observe the ROC curve for this model.

Figure 5.4.1 ROC Curve (Stepwise Selection Model)



The ROC curve suggests that Model_LogR_S has 0.8209 coverage and lies significantly above the diagonal. This is a slight improvement over the ROC curve for Model_LogR_F.

6 Model Selection

The subjective model resulted in a specification with the highest AIC score. This specification included 13 numeric variables which were hand picked, however, a few of these variables were found to have questionable polarity estimates. The forward selection technique resulted in a specification with the lowest AIC score. However, this specification included a number of insignificant variables. Finally, the stepwise selection technique resulted in a specification with an AIC only slightly higher than the forward selection technique, but avoided including the high number of insignificant numeric variables.

A summary of performance metrics over each model using the training set of data is shown below.

Table 6.1: Model Performance Metric Summary

Model	Pred	AIC	SC	KS	Ksa	ROC Coverage
Model_LogR_Subj	28	7003.854	7205.129	0.203568	17.787425	0.7993
Model_LogR_F	41	6722.759	7014.26	0.219433	19.173695	0.8202
Model_LogR_S	35	6726.476	6976.334	0.218525	19.094351	0.8209

Based on the results above, this assessment concludes that Model_LogR_S is the superior model. Its performance metrics were among the most favorable, yet it avoided the inclusion of insignificant variables.

7 Model Deployment Code

Please see Appendix A for the final deployment code.

8 Bonus

Please see Appendix B for the coded decision tree.

9 Conclusion

Each of the logistic regression models fitted as part of this assessment has its own definition of a 'best' model specification. However this assessment ultimately found Model_LogR_S to be the superior model for predicting the probability of an accident, due to its relatively low AIC score and since it retained only significant variables. It is important to note however, that no single statistical method can be relied on to identify the 'true' or 'best' model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model. This may be particularly relevant when considering the short-comings of the dataset used for this assessment. That is, we found little statistical relationship between many of the variables within the dataset, yet, we had a fundamental basis for including many variables and an expectation of appropriate coefficient polarity and model specification based on those fundamentals.

Appendix A: Final Deployment Code

```
%LET key = INDEX;
%LET response1 = TARGET_FLAG;
%LET response2 = TARGET_AMT;
%LET varname = name;

%LET data = logit_insurance;
%LET contents = &data._contents;


* Load the dataset;

libname mydata '/sscc/home/d/dgb2583/411/' access = readonly;

DATA &data.;
    *SET mydata.logit_insurance;
    SET mydata.logit_insurance_test;
RUN; QUIT;

PROC CONTENTS DATA = &data. OUT = &contents.;
RUN; QUIT;

*PROC PRINT DATA = &contents. (OBS=20);
*RUN; QUIT;

PROC MEANS DATA = &data. MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;


* Data rename;

%MACRO rename_num(varname);
    DATA &data_def.;
        SET &data_def. (RENAME = (&varname. = N_&varname.));
    RUN; QUIT;
%MEND;

%MACRO rename_cat(varname);
    DATA &data_def.;
        SET &data_def. (RENAME = (&varname. = C_&varname.));
    RUN; QUIT;
%MEND;

TITLE1 '';
TITLE2 '';

DATA &data._name;
    SET &data.;
RUN; QUIT;

PROC CONTENTS DATA = &data._name OUT = &contents._name;
RUN; QUIT;
```

```
DATA &contents._name;
    SET &contents._name;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
RUN; QUIT;

%LET data_def = &data._name;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%rename_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            WHERE type = 2;
                CALL EXECUTE('%rename_cat('||name||')');
    END;
RUN; QUIT;

DATA &data._name;
    SET &data._name;
    IF C_EDUCATION = "<High School"             then C_EDUCATION = "LT_HS";
    IF C_EDUCATION = "z_High School"            then C_EDUCATION = "z_HS";
    IF C_JOB = "z_Blue Collar"                  then C_JOB = "z_BC";
    IF C_JOB = "Home Maker"                     then C_JOB = "Home_Maker";
    IF C_CAR_TYPE = "Panel Truck"               then C_CAR_TYPE = "Panel_Truck";
    IF C_CAR_TYPE = "Sports Car"                then C_CAR_TYPE = "Sports_Car";
    IF C_RED_CAR = "no"                         then C_RED_CAR = "No";
    IF C_RED_CAR = "yes"                        then C_RED_CAR = "Yes";
    IF C_URBANICITY = "Highly Urban/ Urban"     then C_URBANICITY = "Urban";
    IF C_URBANICITY = "z_Highly Rural/ Rural"   then C_URBANICITY = "z_Rural";
RUN; QUIT;

PROC MEANS DATA = &data._name MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._name OUT = &contents._name;
RUN; QUIT;


* Data preparation;

%MACRO means(varname);
    PROC means DATA = &data_def. noprint;
    OUTPUT OUT = &varname. (DROP = _freq_ _type_)
        nmiss(&varname.)    = &varname._nmiss
        n(&varname.)        = &varname._n
```

```
        mean(&varname.)     = &varname._mean
        median(&varname.)   = &varname._median
        mode(&varname.)     = &varname._mode
        std(&varname.)      = &varname._std
        skew(&varname.)     = &varname._skew
        P1(&varname.)       = &varname._P1
        P5(&varname.)       = &varname._P5
        P10(&varname.)      = &varname._P10
        P25(&varname.)      = &varname._P25
        P50(&varname.)      = &varname._P50
        P75(&varname.)      = &varname._P75
        P90(&varname.)      = &varname._P90
        P95(&varname.)      = &varname._P95
        P99(&varname.)      = &varname._P99
        min(&varname.)      = &varname._min
        max(&varname.)      = &varname._max
        qrange(&varname.)   = &varname._qrange
        ;
    RUN; QUIT;
%MEND;

%MACRO transpose(varname);
    PROC transpose DATA = &varname. OUT = &varname._t;
        var _numeric_;
    RUN; QUIT;
%MEND;

%MACRO symputx_num(varname);
    DATA _null_;
        SET &varname._t;
            CALL symputx(_name_, strip(col1), 'g');
    RUN; QUIT;
%MEND;

%MACRO outlier(varname);
    DATA &data_def.;
        SET &data_def.;
            *IF (&varname. < &&&varname._P10) OR (&varname. > &&&varname._P90) THEN
            *   &varname._OF = 1.0; *ELSE &varname._OF = 0.0;

            *IF (&varname. < &&&varname._P5) OR (&varname. > &&&varname._P95) THEN
            *   &varname._OF = 1.0; *ELSE &varname._OF = 0.0;

            IF (&varname. < &&&varname._P1) OR (&varname. > &&&varname._P99) THEN
                &varname._OF = 1.0; ELSE &varname._OF = 0.0;
    RUN; QUIT;
%MEND;

%MACRO trim(varname);
    DATA &data_def.;
        SET &data_def.;
            &varname._T90 = &varname.;
            *&varname._T90 = max(min(&varname.,&&&varname._P90),&&&varname._P10);
```

```
            IF (&varname._T90 < &&&varname._P10) OR (&varname._T99 > &&&varname._P90) THEN
                &varname._T90 = '.';

            *&varname._T95 = &varname.;
            *&varname._T95 = max(min(&varname.,&&&varname._P95),&&&varname._P5);
            *IF (&varname._T95 < &&&varname._P5) OR (&varname._T95 > &&&varname._P95) THEN
            *   &varname._T95 = '.';

            &varname._T99 = &varname.;
            *&varname._T99 = max(min(&varname.,&&&varname._P99),&&&varname._P1);
            IF (&varname._T99 < &&&varname._P1) OR (&varname._T99 > &&&varname._P99) THEN
                &varname._T99 = '.';
    RUN; QUIT;
%MEND;

%MACRO missing(varname);
    DATA &data_def.;
        SET &data_def.;
            IF missing(&varname.) THEN
                &varname._MF = 1.0; ELSE &varname._MF = 0.0;
    RUN; QUIT;
%MEND;

%MACRO impute(varname);
    DATA &data_def.;
        SET &data_def.;
            *&varname._IMU = &varname.;
            *IF missing(&varname._IMU) THEN
            *   &varname._IMU = &&&varname._mean;

            *&varname._IMO = &varname.;
            *IF missing(&varname._IMO) THEN
            *   &varname._IMO = &&&varname._mode;

            &varname._IME = &varname.;
            IF missing(&varname._IME) THEN
                &varname._IME = &&&varname._median;
    RUN; QUIT;
%MEND;

%MACRO transform(varname);
    DATA &data_def.;
        SET &data_def.;
            &varname._LN = sign(&varname.) * log(abs(&varname.)+1);
            *&varname._SQ = (&varname.*&varname.);
            *&varname._RT = sqrt(&varname.);
    RUN; QUIT;
%MEND;

%MACRO drop(varname);
    DATA &data_def.;
        SET &data_def.;
            DROP &varname.;
```

```
    RUN; QUIT;
%MEND;

TITLE1 '';
TITLE2 '';

* Adhoc changes;

DATA &data._clean;
    SET &data._name;
RUN; QUIT;

DATA &data._clean;
    SET &data._clean;
    N_CAR_AGE = abs(N_CAR_AGE);
RUN; QUIT;

* Create new dataset of flags for continuous variables;

DATA &data._flag;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._flag OUT = &contents._flag;
RUN; QUIT;

DATA &contents._flag;
    SET &contents._flag;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
RUN; QUIT;

%LET data_def = &data._flag;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._flag NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._flag NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%missing('||name||')');
                CALL EXECUTE('%outlier('||name||')');
    END;
RUN; QUIT;
```

```
DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

DATA &data._flag;
    MERGE &data._flag &data.(KEEP = &key.);
RUN; QUIT;

PROC MEANS DATA = &data._flag MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._flag OUT = &contents._flag;
RUN; QUIT;

* Create dummy variables;

DATA &data._dum;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._dum OUT = &contents._dum;
RUN; QUIT;

DATA &contents._dum;
    SET &contents._dum;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
RUN; QUIT;

DATA &data._dum;
    SET &data._dum;
        N_AGE_Risk_Yes  =   (N_AGE <= 30 | N_AGE >= 60);
        N_AGE_Risk_No   =   (N_AGE_Risk_Yes = 0);

        N_BLUEBOOK_Hi   =   (N_BLUEBOOK >= 27000);
        N_BLUEBOOK_Lo   =   (N_BLUEBOOK_Hi = 0);

        N_CLM_FREQ_No   =   (N_CLM_FREQ = 0);
        N_CLM_FREQ_Yes  =   (N_CLM_FREQ > 0);
        N_CLM_FREQ_Hi   =   (N_CLM_FREQ >= 2);
        N_CLM_FREQ_Lo   =   (N_CLM_FREQ_Hi = 0);

        N_HOMEKIDS_No   =   (N_HOMEKIDS = 0);
        N_HOMEKIDS_Yes  =   (N_HOMEKIDS > 0);

        N_INCOME_No     =   (N_INCOME = 0);
        N_INCOME_Yes    =   (N_INCOME > 0);
        N_INCOME_Hi     =   (N_INCOME >= 85000);
        N_INCOME_Lo     =   (N_INCOME_Hi = 0);
```

```
        N_KIDSDRIV_No    =    (N_KIDSDRIV = 0);
        N_KIDSDRIV_Yes   =    (N_KIDSDRIV > 0);

        N_MVR_PTS_No     =    (N_MVR_PTS = 0);
        N_MVR_PTS_Yes    =    (N_MVR_PTS > 0);
        N_MVR_PTS_Hi     =    (N_MVR_PTS >= 4);
        N_MVR_PTS_Lo     =    (N_MVR_PTS_Hi = 0);

        N_OLDCLAIM_No    =    (N_OLDCLAIM = 0);
        N_OLDCLAIM_Yes   =    (N_OLDCLAIM > 0);
        N_OLDCLAIM_Hi    =    (N_OLDCLAIM >= 9500);
        N_OLDCLAIM_Lo    =    (N_OLDCLAIM_Hi = 0);

        N_RENTER_Yes     =    (N_HOME_VAL <= 14400);
        N_RENTER_No      =    (N_RENTER_Yes = 0);

        N_TRAVTIME_Hi    =    (N_TRAVTIME >= 50);
        N_TRAVTIME_Lo    =    (N_TRAVTIME_Hi = 0);
RUN; QUIT;

%LET data_def = &data._dum;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

DATA &data._dum;
    MERGE &data._dum &data.(KEEP = &key.);
RUN; QUIT;

PROC MEANS DATA = &data._dum MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._dum OUT = &contents._dum;
RUN; QUIT;

* Add trimmed series to original dataset;

DATA &data._trim;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._trim OUT = &contents._trim;
RUN; QUIT;

DATA &contents._trim;
    SET &contents._trim;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
```

```
RUN; QUIT;

%LET data_def = &data._trim;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trim NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trim NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%trim('||name||')');
    END;
RUN; QUIT;

* Impute all continuous series in original dataset;

DATA &data._imp;
    SET &data._trim;
RUN; QUIT;

PROC CONTENTS DATA = &data._imp OUT = &contents._imp;
RUN; QUIT;

DATA &contents._imp;
    SET &contents._imp;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
RUN; QUIT;

%LET data_def = &data._imp;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
```

```
            WHERE type = 1;
                CALL EXECUTE('%impute('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

* Transform all continuous series in original dataset;

DATA &data._trans;
    SET &data._imp;
RUN; QUIT;

PROC CONTENTS DATA = &data._trans OUT = &contents._trans;
RUN; QUIT;

DATA &contents._trans;
    SET &contents._trans;
        IF name = "&key." then DELETE;
        IF name = "&response1." then DELETE;
        IF name = "&response2." then DELETE;
RUN; QUIT;

%LET data_def = &data._trans;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trans NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%transform('||name||')');
    END;
RUN; QUIT;

PROC MEANS DATA = &data._trans MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._trans OUT = &contents._trans;
RUN; QUIT;

* Merge Datasets;

DATA &data._merged;
    MERGE &data._flag &data._dum &data._trans;
    *DROP where TYPE _CHARACTER_;
RUN; QUIT;

PROC CONTENTS DATA = &data._merged OUT = &contents._merged;
```

```
RUN; QUIT;

PROC MEANS DATA = &data._merged MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;


* Testing;

DATA &data._scored (KEEP = INDEX P_TARGET_FLAG P_TARGET_AMT);
    SET &data._merged;

    C_PARENT1_NO = 0;
    C_MSTATUS_Yes = 0;
    C_EDUCATION_Bachelors = 0;
    C_EDUCATION_LT_HS = 0;
    C_EDUCATION_Masters = 0;
    C_EDUCATION_PhD = 0;
    C_JOB_Clerical = 0;
    C_JOB_Doctor = 0;
    C_JOB_Home_Maker = 0;
    C_JOB_Lawyer = 0;
    C_JOB_Manager = 0;
    C_JOB_Professional = 0;
    C_JOB_Student = 0;
    C_CAR_USE_Commercial = 0;
    C_CAR_TYPE_Minivan = 0;
    C_CAR_TYPE_Panel_Truck = 0;
    C_CAR_TYPE_Pickup = 0;
    C_CAR_TYPE_Sports_Car = 0;
    C_CAR_TYPE_Van = 0;
    C_REVOKED_No = 0;
    C_URBANICITY_Urban = 0;

    IF C_PARENT1 = "No" THEN C_PARENT1_NO = 1;
    IF C_MSTATUS = "Yes" THEN C_MSTATUS_Yes = 1;
    IF C_EDUCATION = "Bachelors" THEN C_EDUCATION_Bachelors = 1;
    IF C_EDUCATION = "LT_HS" THEN C_EDUCATION_LT_HS = 1;
    IF C_EDUCATION = "Masters" THEN C_EDUCATION_Masters = 1;
    IF C_EDUCATION = "PhD" THEN C_EDUCATION_PhD = 1;
    IF C_JOB = "Clerical" THEN C_JOB_Clerical = 1;
    IF C_JOB = "Doctor" THEN C_JOB_Doctor = 1;
    IF C_JOB = "Home_Maker" THEN C_JOB_Home_Maker = 1;
    IF C_JOB = "Lawyer" THEN C_JOB_Lawyer = 1;
    IF C_JOB = "Manager" THEN C_JOB_Manager = 1;
    IF C_JOB = "Professional" THEN C_JOB_Professional = 1;
    IF C_JOB = "Student" THEN C_JOB_Student = 1;
    IF C_CAR_USE = "Commercial" THEN C_CAR_USE_Commercial = 1;
    IF C_CAR_TYPE = "Minivan" THEN C_CAR_TYPE_Minivan = 1;
    IF C_CAR_TYPE = "Panel_Truck" THEN C_CAR_TYPE_Panel_Truck = 1;
    IF C_CAR_TYPE = "Pickup" THEN C_CAR_TYPE_Pickup = 1;
    IF C_CAR_TYPE = "Sports_Car" THEN C_CAR_TYPE_Sports_Car = 1;
    IF C_CAR_TYPE = "Van" THEN C_CAR_TYPE_Van = 1;
    IF C_REVOKED = "No" THEN C_REVOKED_No = 1;
```

```
IF C_URBANICITY = "Urban" THEN C_URBANICITY_Urban = 1;

LOG_ODDS   =
2.0478   +
(C_PARENT1_No   *   -0.3078)   +
(C_MSTATUS_Yes   *   -0.5291)   +
(C_EDUCATION_Bachelors   *   -0.3933)   +
(C_EDUCATION_LT_HS   *   -0.013)   +
(C_EDUCATION_Masters   *   -0.3638)   +
(C_EDUCATION_PhD   *   -0.1052)   +
(C_JOB_Clerical   *   0.0749)   +
(C_JOB_Doctor   *   -0.9423)   +
(C_JOB_Home_Maker   *   -0.3622)   +
(C_JOB_Lawyer   *   -0.1861)   +
(C_JOB_Manager   *   -0.912)   +
(C_JOB_Professional   *   -0.1776)   +
(C_JOB_Student   *   -0.3516)   +
(C_CAR_USE_Commercial   *   0.7603)   +
(C_CAR_TYPE_Minivan   *   -0.764)   +
(C_CAR_TYPE_Panel_Truck   *   -0.1189)   +
(C_CAR_TYPE_Pickup   *   -0.1055)   +
(C_CAR_TYPE_Sports_Car   *   0.1737)   +
(C_CAR_TYPE_Van   *   -0.1629)   +
(C_REVOKED_No   *   -0.9242)   +
(C_URBANICITY_Urban   *   2.3598)   +
(N_AGE_Risk_Yes   *   0.6004)   +
(N_BLUEBOOK_Hi   *   -0.5338)   +
(N_CLM_FREQ_Yes   *   0.6365)   +
(N_INCOME_Hi   *   -0.3263)   +
(N_MVR_PTS_Hi   *   -0.4492)   +
(N_BLUEBOOK_T90_IME   *   0.000018)   +
(N_HOME_VAL_T99_IME   *   -0.00000142)   +
(N_MVR_PTS_IME   *   0.162)   +
(N_OLDCLAIM_IME   *   -0.00002)   +
(N_TRAVTIME_T99_IME   *   0.0168)   +
(N_BLUEBOOK_T99_IME_LN   *   -0.3661)   +
(N_INCOME_T99_IME_LN   *   -0.0597)   +
(N_KIDSDRIV_IME_LN   *   0.7976)   +
(N_TIF_IME_LN   *   -0.3233)
;

ODDS   =   EXP(LOG_ODDS);
P_TARGET_FLAG   =   ((ODDS) / (1 + ODDS));

P_TARGET_AMT   =
607.45408   +
(N_MVR_PTS_OF   *   1033.99599)   +
(N_AGE_Risk_Yes   *   515.35611)   +
(N_CLM_FREQ_No   *   -759.93781)   +
(N_HOMEKIDS_Yes   *   287.91233)   +
(N_INCOME_Lo   *   363.19203)   +
(N_KIDSDRIV_Yes   *   569.54126)   +
(N_RENTER_Yes   *   548.74609)   +
```

```
    (N_BLUEBOOK_IME   *   0.01586)   +
    (N_CAR_AGE_T90_IME   *   -33.37866)   +
    (N_MVR_PTS_IME   *   175.36712)   +
    (N_TIF_IME_LN   *   -271.31906)   +
    (N_TRAVTIME_T99_IME_LN   *   248.96682)
    ;

RUN; QUIT;

PROC PRINT DATA = &data._scored;
RUN; QUIT;

PROC EXPORT DATA = &data._scored
    OUTFILE = '/sscc/home/d/dgb2583/411/out.csv'
    DBMS = csv
    REPLACE;
RUN; QUIT;

DATA '/sscc/home/d/dgb2583/411/out';
    SET &data._scored;
RUN; QUIT;
```

Appendix B: Coded Decision Tree

```
DATA &data;
    SET &data;

    IF missing(N_INCOME) THEN DO;
        IF C_JOB = "Clerical" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 26000;
            IF C_EDUCATION = "Bachelors" THEN N_INCOME = 48000;
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 35000;
            ELSE N_INCOME = 35000;
        END;
        IF C_JOB = "Doctor" THEN DO;
            IF C_EDUCATION = "Phd" THEN N_INCOME = 130000;
            ELSE N_INCOME = 130000;
        END;
        IF C_JOB = "Home Maker" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 5000;
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 8000;
            IF C_EDUCATION = "Bachelors" THEN N_INCOME = 15000;
            IF C_EDUCATION = "Masters" THEN N_INCOME = 16000;
            IF C_EDUCATION = "PhD" THEN N_INCOME = 26000;
            ELSE N_INCOME = 15000;
        END;
        IF C_JOB = "Lawyer" THEN DO;
            IF C_EDUCATION = "Masters" THEN N_INCOME = 90000;
            IF C_EDUCATION = "PhD" THEN N_INCOME = 120000;
            ELSE N_INCOME = 100000;
        END;
        IF C_JOB = "Manager" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 50000;
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 60000;
            IF C_EDUCATION = "Bachelors"THEN N_INCOME = 80000;
            IF C_EDUCATION = "Masters" THEN N_INCOME = 90000;
            IF C_EDUCATION = "PhD" THEN N_INCOME = 140000;
            ELSE N_INCOME = 80000;
        END;
        IF C_JOB = "Professional" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 30000;
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 60000;
            IF C_EDUCATION = "Bachelors" THEN N_INCOME = 80000;
            IF C_EDUCATION = "Masters" THEN N_INCOME = 90000;
            IF C_EDUCATION = "PhD" THEN N_INCOME = 130000;
            ELSE N_INCOME = 80000;
        END;
        IF C_JOB = "Student" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 6000;
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 6000;
            IF C_EDUCATION = "Bachelors"THEN N_INCOME = 9000;
            ELSE N_INCOME = 6000;
        END;
        IF C_JOB = "z_Blue Collar" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_INCOME = 40000;
```

```
            IF C_EDUCATION = "z_High School" THEN N_INCOME = 60000;
            IF C_EDUCATION = "Bachelors" THEN N_INCOME = 80000;
            IF C_EDUCATION = "Masters" THEN N_INCOME = 70000;
            ELSE N_INCOME = 60000;
        END;
        ELSE N_INCOME = 60000;
    END;


    IF missing(N_CAR_AGE) THEN DO;
        IF C_JOB = "Clerical" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
            ELSE N_CAR_AGE = 4;
        END;
        IF C_JOB = "Doctor" THEN DO;
            IF C_EDUCATION = "Phd" THEN N_CAR_AGE = 10;
            ELSE N_CAR_AGE = 10;
        END;
        IF C_JOB = "Home Maker" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
            IF C_EDUCATION = "Masters" THEN N_CAR_AGE = 7;
            IF C_EDUCATION = "PhD" THEN N_CAR_AGE = 10;
            ELSE N_CAR_AGE = 5;
        END;
        IF C_JOB = "Lawyer" THEN DO;
            IF C_EDUCATION = "Masters" THEN N_CAR_AGE = 7;
            IF C_EDUCATION = "PhD" THEN N_CAR_AGE = 10;
            ELSE N_CAR_AGE = 8;
        END;
        IF C_JOB = "Manager" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
            IF C_EDUCATION = "Masters" THEN N_CAR_AGE = 7;
            IF C_EDUCATION = "PhD" THEN N_CAR_AGE = 10;
            ELSE N_CAR_AGE = 8;
        END;
        IF C_JOB = "Professional" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
            IF C_EDUCATION = "Masters" THEN N_CAR_AGE = 7;
            IF C_EDUCATION = "PhD" THEN N_CAR_AGE = 10;
            ELSE N_CAR_AGE = 8;
        END;
        IF C_JOB = "Student" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
```

```
            ELSE N_CAR_AGE = 4;
        END;
        IF C_JOB = "z_Blue Collar" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_CAR_AGE = 3;
            IF C_EDUCATION = "z_High School" THEN N_CAR_AGE = 4;
            IF C_EDUCATION = "Bachelors" THEN N_CAR_AGE = 5;
            IF C_EDUCATION = "Masters" THEN N_CAR_AGE = 7;
            ELSE N_CAR_AGE = 4;
        END;
        ELSE N_CAR_AGE = 5;
    END;

    IF missing(N_HOME_VAL) THEN DO;
        IF C_JOB = "Clerical" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 100000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 140000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 120000;
            ELSE N_HOME_VAL = 120000;
        END;
        IF C_JOB = "Doctor" THEN DO;
            IF C_EDUCATION = "Phd" THEN N_HOME_VAL = 240000;
            ELSE N_HOME_VAL = 240000;
        END;
        IF C_JOB = "Home Maker" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 70000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 100000;
            IF C_EDUCATION = "Masters" THEN N_HOME_VAL = 80000;
            IF C_EDUCATION = "PhD" THEN N_HOME_VAL = 110000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 80000;
            ELSE N_HOME_VAL = 90000;
        END;
        IF C_JOB = "Lawyer" THEN DO;
            IF C_EDUCATION = "Masters" THEN N_HOME_VAL = 200000;
            IF C_EDUCATION = "PhD" THEN N_HOME_VAL = 220000;
            ELSE N_HOME_VAL = 200000;
        END;
        IF C_JOB = "Manager" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 120000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 190000;
            IF C_EDUCATION = "Masters" THEN N_HOME_VAL = 200000;
            IF C_EDUCATION = "PhD" THEN N_HOME_VAL = 270000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 150000;
            ELSE N_HOME_VAL = 200000;
        END;
        IF C_JOB = "Professional" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 130000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 190000;
            IF C_EDUCATION = "Masters" THEN N_HOME_VAL = 210000;
            IF C_EDUCATION = "PhD" THEN N_HOME_VAL = 290000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 160000;
            ELSE N_HOME_VAL = 190000;
        END;
        IF C_JOB = "Student" THEN DO;
```

```
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 16000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 12000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 16000;
            ELSE N_HOME_VAL = 15000;
        END;
        IF C_JOB = "z_Blue Collar" THEN DO;
            IF C_EDUCATION = "< High School" THEN N_HOME_VAL = 130000;
            IF C_EDUCATION = "Bachelors" THEN N_HOME_VAL = 180000;
            IF C_EDUCATION = "Masters" THEN N_HOME_VAL = 90000;
            IF C_EDUCATION = "z_High School" THEN N_HOME_VAL = 160000;
            ELSE N_HOME_VAL = 160000;
        END;
        ELSE N_HOME_VAL = 150000;
    END;
RUN; QUIT;
```