Assignment #2
Predict 411 Fall 2017 Section 55

Introduction

The purpose of this assignment is to analyse an insurance data set in order to build a model that predicts if a person will be involved in a car crash and then another model to determine the severity of the crash with respect to costs. Prior to building the predictive models, the data will be explored to identify missing values, outliers and potential variable transformations. The transformed data set will then be used in three different logistic regression models. The three models will be compared against each other and the best model will be selected based on model validation metrics.

1.0 Data Exploration

The insurance data set contains 8161 observations of 26 variables. Each record represents a customer at an auto insurance company. Each record has two target variables. The first, TARGET_FLAG, has been coded with a "1" if the person was in a car crash. A "0" means that the person was not in a car crash. The second target variable, TARGET_AMT, is "0" if the person was not in a crash and a continuous scale value greater than zero if they were in a crash.

The variables in the insurance data set can be viewed in Table 1 as well as a brief definition for each variable. The data set is a mix of categorical and continuous variables. The theoretical effect of each variable on driving behaviours is described in Appendix 1.

Table 1 - Variables in the insurance data set.

VARIABLE	DEFINITION	VARIABLE	DEFINITION
INDEX	Identification Variable	JOB	Job Category
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	KIDSDRIV	#Driving Children
TARGET_AMT	If car was in a crash, what was the cost	MSTATUS	Marital Status
AGE	Age of Driver	MVR_PTS	Motor Vehicle Record Points
BLUEBOOK	Value of Vehicle	OLDCLAIM	Total Claims(Past 5 Years)
CAR_AGE	Vehicle Age	PARENT1	Single Parent
CAR_TYPE	Type of Car	RED_CAR	A Red Car
CAR_USE	Vehicle Use	REVOKED	License Revoked (Past 7 Years)
CLM_FREQ	#Claims(Past 5 Years)	SEX	Gender
EDUCATION	Max Education Level	TIF	Time in Force
HOMEKIDS	#Children @Home	TRAVTIME	Distance to Work
HOME_VAL	Home Value	URBANICITY	Home/Work Area
INCOME	Income	YOJ	Years on Job

1.1 Summary Statistics

Table 2 shows the summary statistics for the continuous variables in the data set. The "miss" column represents the number of missing records for the variable. The variables AGE, CAR_AGE, HOME_VAL, INCOME and YOJ have missing values. The missing values will be further investigated and treated in Section 2 as well as exploring the categorical variables individually.

Table 2 - Summary statistics for the numerical variables.

Variable	n	miss	mean	sd	skew	krts	min	max	IQR
TARGET_AMT	8161	0	1504	4704	9	112	0	107586	1036
AGE	8155	6	44.79	8.63	-0.03	-0.06	16	81	12
BLUEBOOK	8161	0	15709.9	8419.73	0.79	0.79	1500	69740	11570
CAR_AGE	7651	510	8.33	5.7	0.28	-0.75	-3	28	11
CLM_FREQ	8161	0	0.8	1.16	1.21	0.29	0	5	2
HOMEKIDS	8161	0	0.72	1.12	1.34	0.65	0	5	1
HOME_VAL	7,697	464	154,867	129,124	0	0	0	885,282	238,724
INCOME	7,716	445	61,898	47,573	1	2	0	367,030	57,889
KIDSDRIV	8161	0	0.17	0.51	3.35	11.79	0	4	0
MVR_PTS	8161	0	1.7	2.15	1.35	1.38	0	13	3
OLDCLAIM	8161	0	4037	8777	3	10	0	57037	4636
TIF	8161	0	5.35	4.15	0.89	0.42	1	25	6
TRAVTIME	8161	0	33	16	0	1	5	142	21
YOJ	7707	454	10.5	4.09	-1.2	1.18	0	23	4

A review of the maximum and minimum columns reveal two values that stand out. The CAR_AGE variable has a minimum value of "-3". This appears to be an error as the age of a car should not be negative. The other value is the maximum of 142 for TRAVTIME. Assuming that the unit of measurement is in minutes, a commute of over two hours appears unusual. However, if the vehicle is used for work purposes then it may be realistic.

The minimum value of zero for HOME_VAL and INCOME are worth mentioning. It is possible for someone not to own a house, therefore the value could be zero. The missing values could represent not being able to obtain that item of data from a customer. An income of zero could mean that the customer is unemployed.

2.0 Data Preparation

The exploratory data analysis conducted in Section 1 identified missing values and potential outliers in the data set. For the variables with missing values, new variables with the prefix IMP were created to highlight the fact that the missing values were imputed using an average value. Flag variables were also created using the prefix M, which used a '0' if it was the original value or a '1' if it had been imputed. The following sections detail how these values were treated so that an effective analysis could be completed. The categorical variables are also explored and are recoded into indicator variables.

2.1 AGE

Six missing values were identified in the AGE variable. The missing values were replaced with the mean value and the variables IMP_AGE and M_AGE were created. As the theoretical effect suggests that young and old people may be riskier drivers, it may be useful to divided into age categories.

In the third regression model, the age categories will be introduced. The age categories were broken into AGE_L, AGE_M, AGE_H with the cutoffs being <=25, 26-65 and greater than 65 respectively. AGE_M will be treated as the base category and won't be included in the regression.

2.2 CAR_AGE

A negative number had been included in the data set for car age. Values below zero were treated as a missing value. Missing values were replaced with the mean value and the variables IMP_CAR_AGE and M_CAR_AGE were created.

2.3 HOME_VAL

There were 464 missing home values. The missing values were replaced with the mean value and the variables IMP_HOME_VAL and M_HOME_VAL were created.

2.4 INCOME

There were 445 missing income values. The missing values were replaced with the mean value and the variables IMP_INCOME and M_INCOME were created.

2.5 YOJ

There were 454 missing values for how long people stay at one job. The missing values were replaced with the mean value and the variables IMP_YOJ and M_YOJ were created.

2.6 CAR_TYPE

The variable CAR_TYPE is a categorical variable. The frequency of each category is shown in Table 3 and it can been seen that there were no missing values. The variable was coded into several indicator variables. The indicator variables created were CT_MVAN, CT_PICKUP, CT_SPORTS, CT_VAN and CT_SUV with Panel Truck being the base variable.

Table 3 - Frequency table for the categorical variable CAR_TYPE.

CAR_TYPE	Minivan	Panel Truck	Pickup	Sports Car	Van	z_SUV	Total
Frequencies:	2145	676	1389	907	750	2294	8161
Proportions:	0.263	0.083	0.17	0.111	0.092	0.281	1

2.7 CAR_USE

The variable CAR_USE is a categorical variable. The frequency of each category is shown in Table 4. Private car use was treated as the base and the indicator variable CU_Commercial was created.

Table 4 - Frequency table for the categorical variable CAR_USE.

CAR_USE	Commercial	Private	Total
Frequencies:	3029	5132	8161
Proportions:	0.371	0.629	1.000

2.8 EDUCATION

The variable EDUCATION had categorical data and the frequency of each category is shown in Table 5**.** z_High School was chosen as the base variable and ED_NOHS, ED_BACH, ED_MAST, ED_PHD were created as indicator variables.

Table 5 - Frequency table for the categorical variable EDUCATION.

	<High School	Bachelors	Masters	PhD	z_High School	Total
Frequencies:	1203	2242	1658	728	2330	8161
Proportions:	0.147	0.275	0.203	0.089	0.286	1

2.9 JOB

The variable JOB had categorical data and it had 526 missing values. The frequency of each category is shown in Table 6. The missing values will be treated as the base variable and JOB_CLERK, JOB_DOC, JOB_HOME, JOB_LAW, JOB_MAN, JOB_PROF, JOB_STU, JOB_BLUE were created as the indicator variables.

Table 6 - Frequency table for the categorical variable JOB.

	Blank	Clerical	Doctor	Home Maker	Lawyer	Manager	Professional	Student	z_Blue Collar	Total
Freq:	526	1271	246	641	835	988	1117	712	1825	8161
Prop:	0.064	0.156	0.03	0.079	0.102	0.121	0.137	0.087	0.224	1

2.10 MSTATUS

The variable MSTATUS, which represents marriage status is a categorical variable. Table 7 shows the frequency for each category. z_No was treated as the base and the indicator variable MS_YES was created.

Table 7 - Frequency table for the categorical variable MSTATUS.

MSTATUS	Yes	z_No	Total
Frequencies:	4894	3267	8161
Proportions:	0.600	0.400	1.000

2.11 PARENT1

The variable PARENT1, which represents the customer being a single parent is a categorical variable. Table 8 shows the split between single parents (YES) and non-single parents. The indicator variable PA_1 was created and was coded “1” for yes and “0” for no.

Table 8 - Frequency table for the categorical variable PARENT1.

	No	Yes	Total
Frequencies:	7084	1077	8161
Proportions:	0.868	0.132	1.000

2.12 RED_CAR

The variable RED_CAR represents if the customer has a red coloured car. There is an urban legend that says that red cars (especially red sports cars) are more risky. Table 9 shows the split of red cars versus non-red cars. The indicator variable RED_YES was created.

Table 9 - Frequency table for the categorical variable RED_CAR.

	no	yes	Total
Frequencies:	5783	2378	8161
Proportions:	0.709	0.291	1.000

2.13 REVOKED

The variable REVOKED represents if a customer's license had been revoked in the past seven years. Table 10 shows the number of customers that have had their license revoked. The indicator variable RVOK_YES was created and was coded with a "1' if it had been revoked.

Table 10 - Frequency table for the categorical variable REVOKED.

	No	Yes	Total
Frequencies:	7161	1000	8161
Proportions:	0.877	0.123	1.000

2.14 SEX

The variable SEX represents if the customer is a male or female. Table 11 shows the gender split of the customers in the data set. The indicator variable SEX_M was created and was coded a "1" if the customer is a male.

Table 11 - Frequency table for the categorical variable SEX.

	M	z_F	Total
Frequencies:	3786	4375	8161
Proportions:	0.464	0.536	1.000

2.15 URBANICITY

The variable URBANICITY represents the home/work area as being urban or rural. Table 12 shows the split between urban and rural customers. The categorical variable RURAL was created and was coded with a "1" if the value was z_HighlyRural/Rural.

Table 12 - Frequency table for the categorical variable URBANICITY.

	Highly Urban/ Urban	z_HighlyRural/Rural	Total
Frequencies:	6492	1669	8161
Proportions:	0.795	0.205	1.000

2.16 Treated Variables Summary

Table 13 lists the 44 variables that were used in the logistic regression models in Section 3.0. The original variables that had missing values have been removed and replaced with the imputed variables. The categorical variables have been removed and replaced with the indicator variables. IMP_AGE will be replaced with AGE_L and AGE_H in the third model. TARGET_AMT or TARGET FLAG will not be used as a predictor variables.

Table 13 - List of treated variables to be used in regression models.

TARGET_FLAG	CT_VAN	M_INCOME	PA_1
TARGET_AMT	CT_SUV	JOB_CLERK	RED_YES
M_AGE	CU_COMMERCIAL	JOB_DOC	RVOK_YES
IMP_AGE	CLM_FREQ	JOB_HOME	SEX_M
IMP_YOJ	ED_NOHS	JOB_LAW	TIF
M_YOJ	ED_BACH	JOB_MAN	TRAVTIME
BLUEBOOK	ED_MAST	JOB_PROF	RURAL
IMP_CAR_AGE	ED_PHD	JOB_STU	OLDCLAIM
M_CAR_AGE	HOMEKIDS	JOB_BLUE	
CT_MVAN	IMP_HOME_VAL	KIDSDRIV	
CT_PICKUP	M_HOME_VAL	MS_YES	
CT_SPORTS	IMP_INCOME	MVR_PTS	

2.17 Correlation Matrix

A correlation matrix was created for the variables and can be seen in Appendix 2. From the matrix it can be seen that INCOME is positively correlated with HOME_VAL, higher levels of education (ED_PHD, ED_MAST) and BLUEBOOK value while being negatively correlated with lower education (ED_NOHS).

The most positively correlated variables with TARGET_FLAG (excluding TARGET_AMT) appear to be OLDCLAIM, PA_1, CU_COMMERCIAL, CLM_FREQ, RVOK_YES and JOB_BLUE. The most negatively correlated variables with TAGET_FLAG appear to be IMP_INCOME, IMP_AGE, RURAL, MS_YES and IMP_HOME_VAL. These variables will be included the manually selected logistic regression model discussed in Section 3.

3.0 Building Models

The overall goal of the model is to predict the expected losses of a customer if they crash their car. It is difficult to do this with just one model due to the fact that it is a zero inflated target. This is due to the payout amount being a continuous target but has a high frequency of records at zero. If the amount is not zero, it is distributed around a larger value.

The approach that has been used to resolve this issue has been to break it into two separate models. Firstly, a logistic regression model that uses TARGET_FLAG as the response variable to calculate the probability that a person will crash their car. The second model is an ordinary least squares linear regression model that calculates the payout amount (severity) if a person does crash their car. The expected losses is calculated by multiplying the probability by the severity.

Three different logistic regression models were experimented to calculate the probability that a person will crash their car. These models are discussed and analysed in the following sections.

3.1 Model 1 Manual Selection

The first model was created using the variables identified by reviewing the correlation matrix discussed in Section 2.17. The highest VIF score was 1.79 for the variable MS_YES therefore multicollinearity was not considered to be an issue. The Logit regression coefficients are shown in Table 14. By taking the exponential of the regression coefficients, the odds ratio can be calculated. The probability can then be calculated using the odds ratio. The two formulas below show how this is done.

- Odds_Y = exp(Logit_Y)
- Prob_Y = Odds_Y/(1 + Odds_Y)

Table 14 – Logit coefficient estimates for the manual variable selection model.

Coefficients:	Estimate	Std. Error	z value
(Intercept)	-0.6277857177	0.0780734015	-8.041
OLDCLAIM	-0.0000120929	0.0000037136	-3.256
PA_1	0.6747496815	0.0866838886	7.784
CU_COMMERCIAL	0.6772250204	0.0630688189	10.738
CLM_FREQ	0.2898930114	0.0261570752	11.083
RVOK_YES	0.9154599432	0.0869506297	10.529
JOB_BLUE	0.3179920279	0.0707596768	4.494
IMP_INCOME	-0.0000091604	0.0000008298	-11.039
RURAL	-2.0345857885	0.1068084249	-19.049
MS_YES	-0.2532439673	0.0748935105	-3.381
IMP_HOME_VAL	-0.0000017435	0.0000003248	-5.368

Null deviance: 9418.0 on 8160 degrees of freedom
Residual deviance: 7843.9 on 8150 degrees of freedom
AIC: 7865.9

The associations of the coefficients for Model 1 were reviewed against the theoretical driving behaviour associations for the variables. A positive association would increase the probability of a crash while a negative associate would reduce the probability of a crash. OLDCLAIM had a negative association which was not expected as it was thought that the higher claims amount made in the past would potentially positively effect the probability of a crash. As expected this is the case for claim frequency (CLM_FREQ).

The variable PA_1 which represents being a single parent has a positive association. It was unknown what affect this variable would have. The variable RURAL has a negative association. This would mean that people living in rural areas have a lower probability of crashing compared to those living in an urban area (assuming other variables are held constant). All other variables had associations that lined up with the theoretical effect.

Metrics to determine the overall fit for generalized linear models include the McFadden adjusted R-squared metric and Akaike's Information Criertion (AIC). The formulas to calculate these two metrics are shown below.

$$R^2_{McF} = 1 - \frac{logLik(M_u) - (k+1)}{logLik(M_c)}$$

$$AIC = Deviance(M_u) + 2C$$

Where:

- $logLik(M_u)$ is the log likelihood of the unconstrained model (full)
- $logLik(M_c)$ is the log likelihood of the constrained model (null)
- k is the number of predictor variables
- C is the number of coefficients (predictor variables + intercept)
- $Deviance(M_u) = -2 * logLik(M_u)$.

The calculated metrics for Model 1 are shown in Table 15. These metrics will be compared against the other models in Section 4.

Table 15 - Model 1 goodness of fit validation metrics.

	$logLik(M_c)$	$logLik(M_u)$	K	R^2_{McF}	$Deviance(M_u)$	AIC
Model 1	4708.981	-3921.933	10	0.1648	7843.87	7865.9

Receiver operating characteristic (ROC) curves can be used to compare the relative performance among different classifiers. It measures the tradeoff between selecting as many true positives as possible while avoiding false positive. The Kolmogorov-Smirnov (KS) statistic represents the maximum difference between the cumulative true positive and cumulative false positive rate. It can be used to determine the cutoff point for selecting as many true positives as possible while avoiding false positives. Figure 1 displays the ROC curve for Model 1. The KS statistic was calculated to be 0.415.

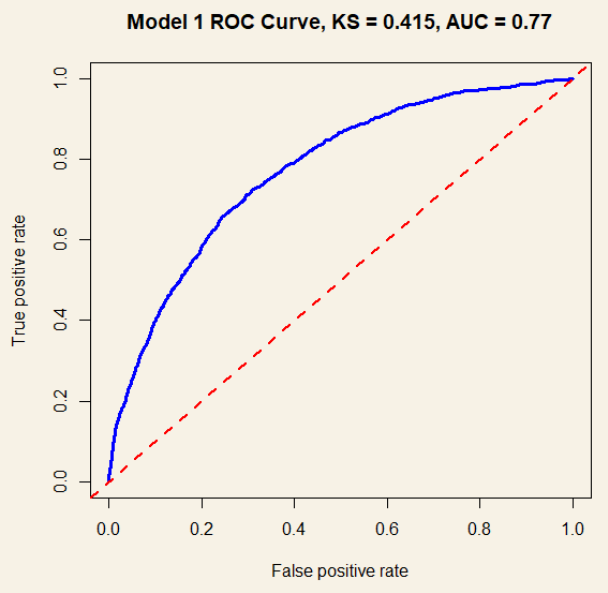

Figure 1 - ROC Curve for Model 1.

3.2 Model 2 Backward Selection

The second model was created using the backward automated variable selection technique. The technique starts with an initial model that contains all the predictor variables. The subsequent cycles of the algorithm remove variables that do not make a statistical significant contribution to the model. It continues until the remaining variables in the model cannot be removed without affecting the predictive power of the model.

Table 16 – Logit coefficient estimates for the backward variable selection model.

Coefficients:	Estimate	Std. Error	z value
(Intercept)	0.086936216	0.151381078	0.574
IMP_INCOME	-0.000003865	0.000001009	-3.83
OLDCLAIM	-0.000011208	0.000003861	-2.903
CT_SUV	-0.151391408	0.088381242	-1.713
JOB_CLERK	0.23350298	0.09655733	2.418
PA_1	0.397360863	0.108280431	3.67
M_AGE	2.251071469	1.20505509	1.868
CU_COMMERCIAL	0.7154922	0.077226699	9.265
JOB_DOC	-0.511644384	0.248890197	-2.056
CLM_FREQ	0.260075777	0.027220808	9.554
RVOK_YES	0.861631766	0.090756253	9.494
SEX_M	-0.104333665	0.073584077	-1.418
ED_BACH	-0.383509481	0.078454862	-4.888
JOB_MAN	-0.717960565	0.109553476	-6.554
TIF	-0.05734727	0.007294186	-7.862
BLUEBOOK	-0.000028174	0.00000426	-6.613
ED_MAST	-0.367272376	0.105573064	-3.479
TRAVTIME	0.014844455	0.001872042	7.93
ED_PHD	-0.265617253	0.160477152	-1.655
RURAL	-2.429673109	0.112220224	-21.651
HOMEKIDS	0.054617319	0.033947254	1.609
JOB_BLUE	0.16907774	0.091143236	1.855
IMP_YOJ	-0.013775282	0.007987779	-1.725
CT_MVAN	-0.796854425	0.088521742	-9.002
IMP_HOME_VAL	-0.000001382	0.000000334	-4.137
KIDSDRIV	0.39725101	0.059770159	6.646
CT_PICKUP	-0.224643055	0.089452465	-2.511
MS_YES	-0.480644713	0.08272727	-5.81

```
Null deviance: 9418.0  on 8160  degrees of freedom
Residual deviance: 7370.9  on 8133  degrees of freedom
AIC: 7426.9
```

The backward selection technique resulted in 27 predictor variables being included in the model. Table 16 shows the logit coefficient estimates. The lower educated type jobs (JOB_BLUE, JOB_CLERK) have positive associations while the higher education jobs (JOB_DOC, JOB_MAN) have negative associations as expected. This is also reflected in the education variables (ED_BACH, ED_MAST, ED_PHD). Having kids at home and kids driving both result in a positive association.

Travel time (TRAVTIME) and commercial use of vehicles (CU_COMMERICAL) have positive associations. This makes sense as the vehicles are driven more may increase the probability of a crash. The other variable associations are as expected similar to Model 1. However, of particular note would be the vehicle type category. SUVs and minivans were included in the model and have negative associations. The other car type indicator variables were not included in the model. The highest VIF score for Model 2 was 2.19 for the variable IMP_INCOME therefore multicollinearity was not considered to be an issue.

Table 17 - Model 2 goodness of fit validation metrics.

	$logLik(M_c)$	$logLik(M_u)$	K	R^2_{McF}	$Deviance(M_u)$	AIC
Model 2	4708.981	-3685.441	27	0.2114	7370.882	7426.9

The calculated metrics for Model 2 are shown in Table 17. These metrics will be compared against the other models in Section 4. Figure 2 displays the ROC curve for Model 2. The KS statistic was calculated to be 0.47.

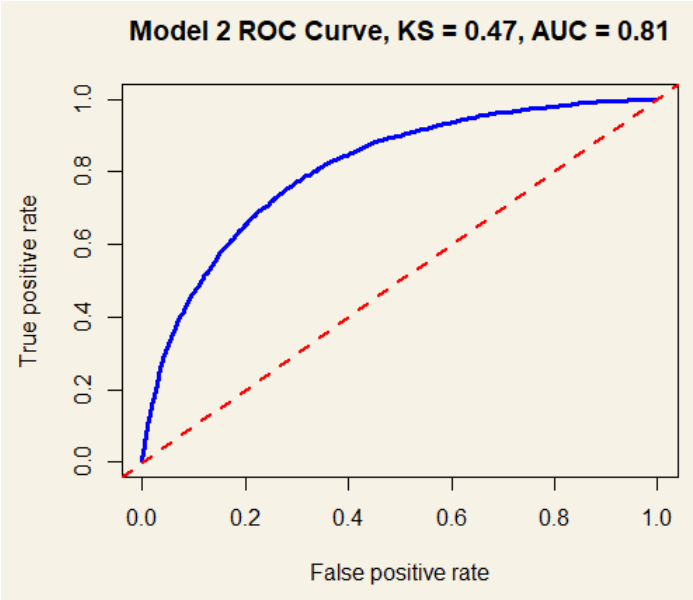

Figure 2 - ROC Curve for Model 2

3.3 Model 3 Backward Selection with Age Brackets

The third logistic regression model was created using the backward automated variable selection technique similar to Model 2. The difference between the two models is that the AGE variable was changed into age brackets as discussed in Section 2.1 of the report. The logit coefficient estimates are shown in Table 19. The automated variable selection technique resulted in including 24 variables in the model.

The lower age bracket between of customers of age 25 and below has been included in the model and has a positive association. This is aligned to theoretical effect as it is thought younger drivers have riskier driving behaviours. The associations for the rest of the variables is as expected and similar to Model 2. The highest VIF score for Model 3 was 2.03 for the variable MS_YES therefore multicollinearity was not considered to be an issue.

The calculated metrics for Model 3 are shown in Table 18. These metrics will be compared against the other models in Section 4. Figure 3 displays the ROC curve for Model 3. The KS statistic was calculated to be 0.48.

Table 18 - Model 3 goodness of fit validation metrics.

	$logLik(M_c)$	$logLik(M_u)$	K	R^2_{McF}	$Deviance(M_u)$	AIC
Model 3	4708.981	-3680.417	24	0.2131	7360.834	7410.8

Table 19 – Logit coefficient estimates for the backward variable selection model with age brackets.

Coefficients:	Estimate	Std. Error	z value
(Intercept)	-0.0803885926	0.1334457566	-0.602
IMP_INCOME	-0.0000043505	0.0000009340	-4.658
OLDCLAIM	-0.0000109999	0.0000038548	-2.854
JOB_CLERK	0.2594073624	0.0937736182	2.766
PA_1	0.3763134543	0.1085544860	3.467
AGE_L	1.0946490791	0.2422160975	4.519
CU_COMMERCIAL	0.6952036363	0.0726917194	9.564
JOB_DOC	-0.6883424477	0.2268269645	-3.035
CLM_FREQ	0.2585946715	0.0272015748	9.507
RVOK_YES	0.8568861700	0.0907458562	9.443
ED_BACH	-0.3419352446	0.0726353100	-4.708
JOB_MAN	-0.7191106228	0.1094985255	-6.567
TIF	-0.0578797334	0.0073013472	-7.927
BLUEBOOK	-0.0000262141	0.0000041040	-6.387
ED_MAST	-0.2886862957	0.0935223570	-3.087
TRAVTIME	0.0147496273	0.0018725147	7.877
RURAL	-2.4211769557	0.1121730380	-21.584
HOMEKIDS	0.0511172454	0.0339674333	1.505
JOB_BLUE	0.2302712468	0.0850282757	2.708
IMP_YOJ	-0.0140239496	0.0079662703	-1.76
CT_MVAN	-0.7712495231	0.0760402316	-10.143
IMP_HOME_VAL	-0.0000013913	0.0000003336	-4.171
KIDSDRIV	0.4153405129	0.0600052226	6.922
CT_PICKUP	-0.1823802266	0.0800696108	-2.278
MS_YES	-0.4735555009	0.082704712	-5.726

Null deviance: 9418.0 on 8160 degrees of freedom
Residual deviance: 7360.8 on 8136 degrees of freedom
AIC: 7410.8

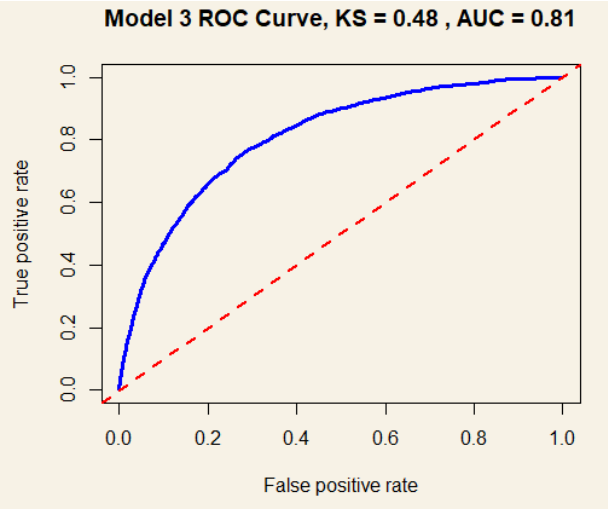

Figure 3 - ROC Curve for Model 3.

3.4 Severity Model

A severity model was required to be built to estimate the expected losses of a customer. To build the model, an ordinary least squares regression was created to model the variable TARGET_AMT. The correlation matrix revealed that there were not many variables that had a strong correlation with TARGET_AMT.

The dictionary for the data set was reviewed and it was determined that the variables BLUEBOOK and CAR_AGE would probably effect the payout if there was a crash. The model was only built using records where the payout amount was greater than zero. Table 20 shows the regression coefficients for the severity model. It makes sense that the bluebook coefficient positively effects the payout size and the age of the car decreases the size of the payout. The model does not account for very much of the variability in the target variable. It only has an adjusted R-squared value of 0.014.

Table 20 - Regression coefficients for the severity model.

	Estimate	Std Err	t-value	p-value
(Intercept)	4435.929904	376.165213	11.793	0
bluebook	0.116327	0.020303	5.73	0
car_age	-52.728785	31.495207	-1.674	0.094

The equation used to create the variable P_TARGET_AMOUNT is the following:

P_TARGET_AMOUNT = 4435.9299 + 0.116327*bluebook-52.728785*car_age.

4.0 Model Comparison

This section of the report compares the model validation metrics of the three different logistic regression models with the aim of selecting the best model for predicting the probability of a customer being involved with a crash.

When comparing two models using McFadden's adjusted R-squared, the model with the larger value may indicate a better model. AIC provides a method to compare models of different sizes. It is based on deviance and has a penalty term for more complicated models (larger number of predictor variables). When looking at AIC, the model with the smaller value is deemed the better model.

Table 21 - Comparison of the three logistic regression model key validation statistics.

	R2MCF	AIC	AUC
Model 1	0.1648	7865.9	0.77
Model 2	0.2114	7426.9	0.81
Model 3	0.2131	7410.8	0.81

From Table 21 it can be seen that Model 3 has the highest McFadden adjusted R-squared value and the smallest AIC value. Model 2 & 3 had the largest area under the ROC Curve (AUC) with values of 0.81. The area under a ROC Curve can be used to compare classifiers. The larger the AUC, the better the classifier. AUC greater than 0.8 can be taken to indicate good discrimination potential.

Conclusion

Based on the metrics discussed in Section 4 and the fact that the associations of the logit coefficients were logical and interpretable, Model 3 was selected as the best model to predict the probability of a customer being involved in a car crash.

Interestingly, the indicator variable RED_CAR was not deemed to have a statistically significant contribution to the model and was excluded by the automated variable selection algorithm. This would suggest that drivers of red cars may not be risker than drivers of other coloured cars.

References

DS. (n.d.). *Guide to Credit Scoring in R*. Retrieved from Cran.R-Project: https://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf

Kabacoff, R. (2015). *R in Action, Data analysis and graphics with R, II Edition.* Shelter Island: Manning.

University of Victoria. (n.d.). *ROC Curves*. Retrieved from https://web.uvic.ca/~maryam/DMSpring94/Slides/9_roc.pdf

Appendix 1

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

Appendix 2 – Correlation Matrix