Unit 02 Homework – Insurance

KAGGLE Name: *NikhilAgarwal*

Nikhil Agarwal
Northwestern University
PREDICT 411, Section 55

I am requesting a total of 80 bingo bonus points for the unit 2 homework. Please see my justification below.

Points Requested	**Category**	**Justification**
20	Decision Tree	I have used JMP to develop decision trees for almost all of the variables that had missing values
10	Macros	I have extensively used macros
20	R Code	Much of the homework has been rewritten in R. Please see the attached homework (NikhilAgarwal_HW2_RCode.r)
5	PROBIT Model	The chosen model was also reran using PROBIT
5	CLOGCLOG Model	The chosen model was also reran using CLOGLCOG
20	KS Statistic	I used the PROC NPAR1WAY to calculate the KS statistic – specifically looking for the D value

Table of Contents

INTRODUCTION

The intent of this assignment is to develop a logistic function that can be used to predict the likelihood of an individual becoming involved in a vehicular accident. Information on over 8000 customers was used to help construct the model. Prior to developing a single model, multiple logistic models were explored using primarily the stepwise and backward selection methods. Various model diagnostic parameters (e.g., AIC, SBC, and ROC curves) were used to determine the best model.

RESULTS

Data Exploration

The original dataset contains 23 distinct variables (outlined in Table 1[1]). These variables can be considered potential predictors. Note that the response variable will be TARGET_FLAG. The TARGET_FLAG is technically either a 0 or 1 (0 meaning customer not involved in accident and 1 meaning the customer involved in accident) in the data dictionary. However, the predicted response variable will simply indicate the likelihood (between 0 and 1) of the customer being involved in a vehicular accident.

Table 1: Brief description of default variables

Variable Name	Variable Type	Definition	Theoretical Effect
AGE	Continuous	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Continuous	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Continuous	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Categorical	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Categorical	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	Continuous	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Categorical	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	Continuous	#Children @Home	Unknown effect
HOME_VAL	Continuous	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Continuous	Income	In theory, rich people tend to get into fewer crashes
JOB	Categorical	Job Category	In theory, white collar jobs tend to be safer

[1] The information in this table were provided for this assignment

KIDSDRIV	Continuous	#Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Categorical	Marital Status	In theory, married people drive more safely
MVR_PTS	Continuous	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Continuous	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Categorical	Single Parent	Unknown effect
RED_CAR	Categorical	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	Categorical	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Categorical	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Continuous	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Continuous	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Categorical	Home/Work Area	Unknown
YOJ	Continuous	Years on Job	People who stay at a job for a long time are usually more safe

As the data are explored, it is wise to identify if any of the variables are missing. Figure 1 illustrates some basic statistics on the dataset for the continuous variables.

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum	1st Pctl	99th Pctl
INDEX		8161	0	5151.87	5133.00	1.0000000	10302.00	103.0000000	10197.00
TARGET_FLAG		8161	0	0.2638157	0	0	1.0000000	0	1.0000000
TARGET_AMT		8161	0	1504.32	0	0	107586.14	0	19866.59
KIDSDRIV	#Driving Children	8161	0	0.1710575	0	0	4.0000000	0	2.0000000
AGE	Age	8155	6	44.7903127	45.0000000	16.0000000	81.0000000	25.0000000	64.0000000
HOMEKIDS	#Children @Home	8161	0	0.7212351	0	0	5.0000000	0	4.0000000
YOJ	Years on Job	7707	454	10.4992864	11.0000000	0	23.0000000	0	17.0000000
INCOME	Income	7716	445	61898.10	54028.17	0	367030.26	0	215536.28
HOME_VAL	Home Value	7697	464	154867.29	161159.53	0	885282.34	0	500309.15
TRAVTIME	Distance to Work	8161	0	33.4887972	32.8709696	5.0000000	142.1206304	5.0000000	75.1443301
BLUEBOOK	Value of Vehicle	8161	0	15709.90	14440.00	1500.00	69740.00	1500.00	39090.00
TIF	Time in Force	8161	0	5.3513050	4.0000000	1.0000000	25.0000000	1.0000000	17.0000000
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	4037.08	0	0	57037.00	0	42820.00
CLM_FREQ	#Claims(Past 5 Years)	8161	0	0.7985541	0	0	5.0000000	0	4.0000000
MVR_PTS	Motor Vehicle Record Points	8161	0	1.6955030	1.0000000	0	13.0000000	0	8.0000000
CAR_AGE	Vehicle Age	7651	510	8.3283231	8.0000000	-3.0000000	28.0000000	1.0000000	21.0000000

Figure 1: Basic statistics on numerical variables

Note how five of the potential predictor variables have missing values. On a side note, the variable INDEX is simply a unique identifier that will not be used for any modelling purpose. Furthermore, the variable TARGET_FLAG is the key response variable for this assignment. TARGET_AMT is not documented nor modelled in this assignment.

Recall that there are ten categorical variables. Each one was explored in detail and it was discovered that the variable JOB had over 500 observations with missing values (see Figure 2).

For all variables (continuous and categorical) that have missing values, imputed values will be used in order to create a healthier model.

Job Category				
JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	526	6.45	526	6.45
Clerical	1271	15.57	1797	22.02
Doctor	246	3.01	2043	25.03
Home Maker	641	7.85	2684	32.89
Lawyer	835	10.23	3519	43.12
Manager	988	12.11	4507	55.23
Professional	1117	13.69	5624	68.91
Student	712	8.72	6336	77.64
z_Blue Collar	1825	22.36	8161	100.00

Figure 2: PROC FREQ output on JOB

Table 2 highlights the correlation of each continuous predictor variable to the response variable of TARGET_FLAG. The closer the value is to +1 or -1, the stronger the linear relationship. Unsurprisingly, no single predictor variable is highly correlated with TARGET_FLAG. Recall from Table 1 that certain continuous variables (such as TRAVTIME) could increase the likelihood of having an accident. In this case, if a variable (such as TRAVTIME) has a positive correlation, then it could be construed that it has a tendency to increase the TARGET_FLAG.

Table 2: PROC CORR output of TARGET_FLAG vs. numerical variables

Variable	Correlation
KIDSDRIV	0.10367
AGE	-0.10322
HOMEKIDS	0.11562
YOJ	-0.07051
INCOME	-0.14201
HOME_VAL	-0.18374
TRAVTIME	0.04815
BLUEBOOK	-0.10338
TIF	-0.08237
OLDCLAIM	0.13808
CLM_FREQ	0.2162
MVR_PTS	0.2192
CAR_AGE	-0.10065

Another key check is to determine if there are outliers. All 13 continuous predictor variables were checked for outliers. Figure 3 is an example of a histogram and boxplot for the variable INCOME. Note the many circles that are outside the whiskers in the boxplot. This is a strong indicator that outliers may be present.

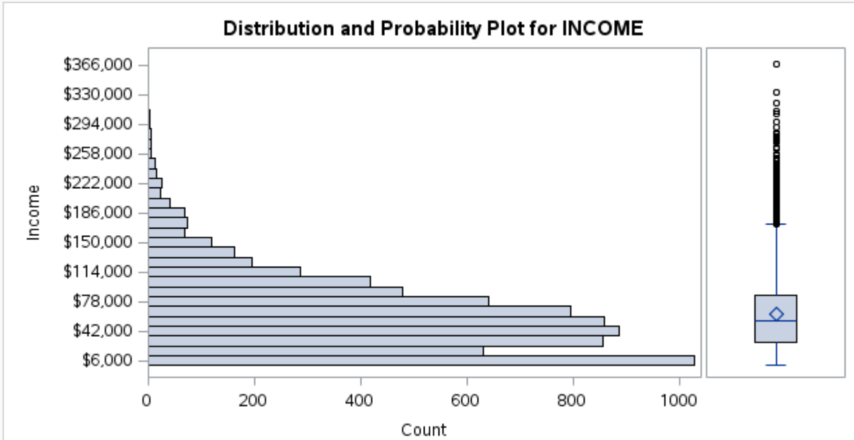

Figure 3: Histogram & Boxplot for INCOME

It was found that many of the continuous variables had outliers. Therefore, value caps will be deployed on each variable to ensure that outliers do not skew the model unnecessarily. The caps are explained in greater detail in the section, "Data Preparation".

Data Preparation

Imputation

Recall from the previous section, "Data Exploration", that there are six total variables (five continuous and one categorical) that have missing values. Table 3 summarizes the imputation method used for each variable. All six of these variables were removed from the modelling process and the imputed values were stored in different variables starting with the prefix "imp_". If there was a non-null value for any of the variables in Table 3, then that value was used in lieu of an imputed value. Furthermore, a flag variable was created (with the prefix "m_") to indicate if an imputed value was entered (indicated with a 1 meaning true) or if the original value was used (indicated with a 0 meaning false).

Table 3: Summary of imputation methods

Variable	Method of Imputation	Imputed Variable Name
AGE	Mean	imp_age
YOJ	Decision Tree	imp_yoj
INCOME	Decision Tree	imp_income
HOME_VAL	Decision Tree	imp_home_val
CAR_AGE	Decision Tree	imp_carage
JOB	Comparison of predictability	imp_job

Using JMP, decision trees were constructed for the appropriate variables (as listed in Table 3). An example decision tree for the variable YOJ is shown in Figure 4.

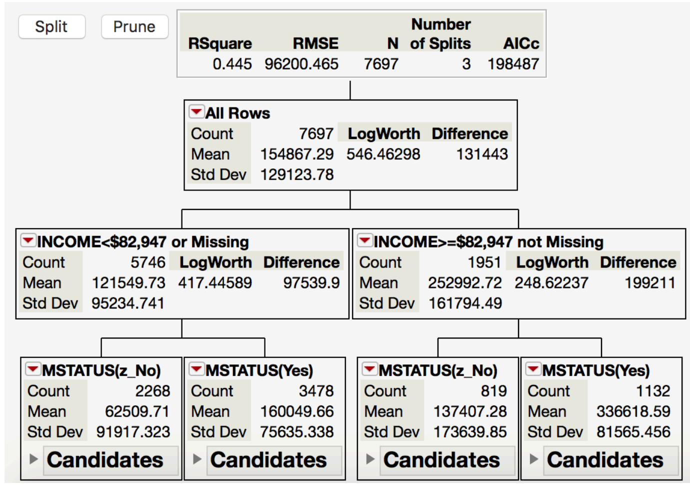

Figure 4: Example Decision Tree diagram from JMP

The following logic briefly describes (converting logic syntax to English) the decision tree for the variable HOME_VAL:

```
If INCOME is less than 82947 and then if MSTATUS is equal to NO, the imputed
value for HOME_VAL should be 62509.71 otherwise it should be 160049.66. If
INCOME is greater than or equal to 82947 and MSTATUS is NO then the imputed
value for HOME_VAL is 137407.28 otherwise it should be 336618.59.
```

For the variable AGE, it was found that the distribution was fairly normal (see Figure 5) and it was decided to simply impute the missing values of AGE with the mean. Note that there are some outliers and they will be identified. Recall from Figure 1 that the mean value of AGE is approximately 45.

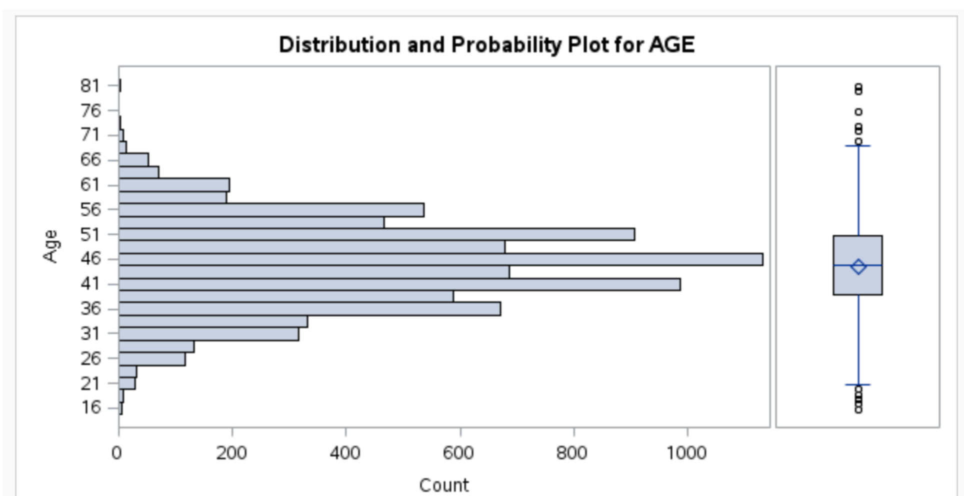

Figure 5: Histogram and Boxplot for AGE

For the variable JOB, a different approach was taken. Once all the other variables' missing values were imputed, a PROC MEANS statement comparing the imputed income to JOB was executed. The results are in Figure 6. The intent is to utilize the median income by job category to construct an imputed value if the JOB value was left blank and then use the imputed income value to determine the job.

Analysis Variable : imp_income						
Job Category	**N Obs**	**N**	**N Miss**	**Minimum**	**Maximum**	**Median**
Clerical	1271	1271	0	2954.14	113845.14	32200.31
Doctor	246	246	0	9619.42	215000.00	126677.71
Home Maker	641	641	0	17.9823292	185747.34	11000.00
Lawyer	835	835	0	1404.91	215000.00	82659.88
Manager	988	988	0	1021.68	215000.00	77825.66
Professional	1117	1117	0	8458.21	215000.00	70293.04
Student	712	712	0	5.2859290	93066.98	6000.00
z_Blue Collar	1825	1825	0	5014.93	198319.70	55506.87

Figure 6: PROC MEANS for imp_income Vs. JOB

Using the median values in Figure 6, the thresholds for each job were created. Table 4 summarizes the chosen values (which are slightly different from the values in Figure 6).

Table 4: Summary of minimum income threshold by JOB

Job	Minimum Income for Job
Doctor	128000
Lawyer	90000
Manager	80000
Professional	75000
z_Blue Collar	57000
Clerical	33000
Home Maker	11000
Student	Less than 11000

Outliers

In order to reduce the effects of outliers, an upper threshold value was employed for seven of the continuous variables based on the 99^{th} percentile. The intent is not to necessarily eliminate all outliers, but to reduce their effect on the overall model. These upper limits still allow for some 'peak' and 'valley' points. Table 5 highlights the upper thresholds used for the seven continuous variables. For each of these values, an outlier flag (with the syntax "out_") was used to indicate if the original value was an outlier.

Table 5: Outlier threshold summary

Variable	Upper Threshold
AGE	64
INCOME	215000
HOME_VAL	500000
TRAV_TIME	75
BLUEBOOK	40000
OLDCLAIM	43000
TIF	16

The minimum values were left alone for almost all of the variables. However, for the variable INCOME, many customers had reported a value of 0. This is somewhat peculiar, so a minimum threshold was identified for the variable INCOME. However, this minimum was designed to be dependent on the imputed job and was essentially the new median for each job (as seen in Figure 6).

Analysis Variable : imp_income						
imp_job	N Obs	N	N Miss	Minimum	Maximum	Median
Clerical	1325	1325	0	2954.14	113845.14	33055.77
Doctor	449	449	0	9619.42	215428.49	148595.01
Home Maker	655	655	0	17.9823292	185747.34	11000.00
Lawyer	975	975	0	1404.91	215000.00	90939.16
Manager	1021	1021	0	1021.68	215000.00	78582.21
Professional	1136	1136	0	8458.21	215000.00	70736.14
Student	713	713	0	5.2859290	93066.98	6000.00
z_Blue Collar	1887	1887	0	5014.93	198319.70	56848.52

Figure 7: PROC MEANS for imputed income with minimum threshold vs. imp_job

Table 6 summarizes the values coded into the algorithm for the minimum value of INCOME based on occupation. Recall that these values were only applied if and only if the original income reported by the customer was 0.

Table 6: Minimum Income Limit for Occupation

Occupation	Minimum Income Limit
Doctor	128000
Lawyer	90000
Manager	80000
Professional	75000
z_Blue Collar	57000
Clerical	33000
Home Maker	11000
Student	6000

Negative Value Handling

During the exploratory data analysis, it was discovered that the variable CAR_AGE had a negative value. None of the predictor variables should have a negative value. Therefore, all continuous variables were initially imputed with the absolute value of the original value. This will ensure that all values used in the model are correct in terms of having the wrong sign.

New Variables

Four new variables were created in order to better understand a customer's behavior. Table 7 illustrates the variables created and their intent.

Table 7: Summary of new variables constructed

Variable	Condition	Intent
f_renter	If HOME_VAL = 0 then 1 else 0	To understand if the customer is renting or owning a home
f_claimchk	If OLD_CLM > 0 then 1 else 0	To understand if the customer has had a claim in the past 5 years
f_speeder	If MVR_PTS > 0 then 1 else 0	To understand if the customer is a safe driver. Any points incurred may suggest he/she is not a safe driver
f_newjobchk	If YOJ > 0 then 1 else 0	If a customer has a new job, then it is possible he/she may be a more careful driver. 1 indicates that the customer does NOT have a new job
payout_per_year	OLDCLAIM/CLM_FREQ	To understand the average payout per instance if a payout occurred in the past 5 years

The intent of these variables is to understand a customer's 'behavior' at a high level with binary-like values (i.e., 0 meaning false and 1 meaning true) as well as understand the 'average' payout per instance if a payout has occurred.

Model Development

Over nine different models were constructed, but for the sake of but for the sake of succinctness, only three of them will be discussed in this report. Some of the models were constructed using either a stepwise selection method or a backward selection method. This approach enables a greater understanding of the statistical significance of a variable. For all models, the Log Odds and the corresponding probability calculation (i.e., the response variable, P_TARGET_FLAG) are all designed to derive the probability of a customer being involved in a vehicular accident. This was accomplished using the PROC LOGIT function in SAS with the reference set to 0.

As a primer, all of the models constructed have the following equation format:

$$Log\ Odds = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Using the derived log odds value, it is necessary to then convert this value to a probability using the following syntax:

$$P\_Target\_Flag = \frac{e^{LogOdds}}{1 + e^{LogOdds}}$$

This conversion is absolutely necessary in order to derive the probability of a customer being involved in a vehicular accident.

Model 1

The first model was designed to explore 41 total predictor variables including all imputation and outlier flags with the response variable being Log Odds. Table 8 illustrates a summary of the variables chosen and their respective coefficients.

Table 8: Model 1 Parameter Estimates

Variable	Variable Sub Element	Variable Code	Coefficient Notation	Coefficient Value
Intercept			β0	-0.9323
CAR_TYPE	Minivan	X1	β1	-0.6975
CAR_TYPE	Panel Truck	X2	β2	-0.1037
CAR_TYPE	Pickup	X3	β3	-0.1488
CAR_TYPE	Sports Car	X4	β4	0.2565
CAR_TYPE	Van	X5	β5	-0.0488
CAR_USE	Commercial	X6	β6	0.7574
EDUCATION	<High School	X7	β7	-0.0158
EDUCATION	Bachelors	X8	β8	-0.408
EDUCATION	Masters	X9	β9	-0.3716
EDUCATION	PhD	X10	β10	-0.3664
MSTATUS	Yes	X11	β11	-0.4437
PARENT1	No	X12	β12	-0.458
REVOKED	No	X13	β13	-0.9612
URBANICITY	Highly Urban/ Urban	X14	β14	2.3613
imp_job	Clerical	X15	β15	0.0555
imp_job	Doctor	X16	β16	-0.4323
imp_job	Home Maker	X17	β17	-0.3054
imp_job	Lawyer	X18	β18	-0.1344
imp_job	Manager	X19	β19	-0.8154
imp_job	Professional	X20	β20	-0.1388
imp_job	Student	X21	β21	-0.4019
KIDSDRIV		X22	β22	0.4174

MVR_PTS		X23	β23	0.0959
f_claimchk		X24	β24	0.6371
f_newjobchk		X25	β25	-0.5206
f_renter		X26	β26	0.3353
imp_bluebook		X27	β27	-0.00003
imp_income		X28	β28	-5.69E-06
imp_oldclaim		X29	β29	-0.00002
imp_tif		X30	β30	-0.0562
imp_travtime		X31	β31	0.0151
m_yoj		X32	β32	-0.4195
out_bluebook		X33	β33	0.6551
out_income		X34	β34	0.7914
out_oldclaim		X35	β35	8.8267

This model utilized a stepwise variable selection method and resulted in a total of 22 predictor variables (the table above shows many more variables, but note how the categorical variables are listed multiple times).

Figure 8 illustrates three key model fit statistics. Of key importance is to note that the values for these statistics is reduced when exploring both the intercept and covariates. Note that the lower the value of these statistics, the stronger the indicator that the model may be a better model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7328.283
SC	9426.969	7580.539
-2 Log L	9417.962	7256.283

Figure 8: Model 1 Fit Statistics

Figure 9 illustrates the ROC curve for this model. The ROC curve simply describes the amount true positive rate versus the false positive rate. Essentially, the closer the curve is to the left hand side and towards the top, the truer positive the rate. However, this would most likely be an indicator of an over-fitted model. In this case, the goal is to balance the curve without overfitting the model. For this model, the metric of the ROC curve (i.e., area under the curve) is 0.8163 (a perfect curve would be 1 – a potential indicator of an over-fitted model).

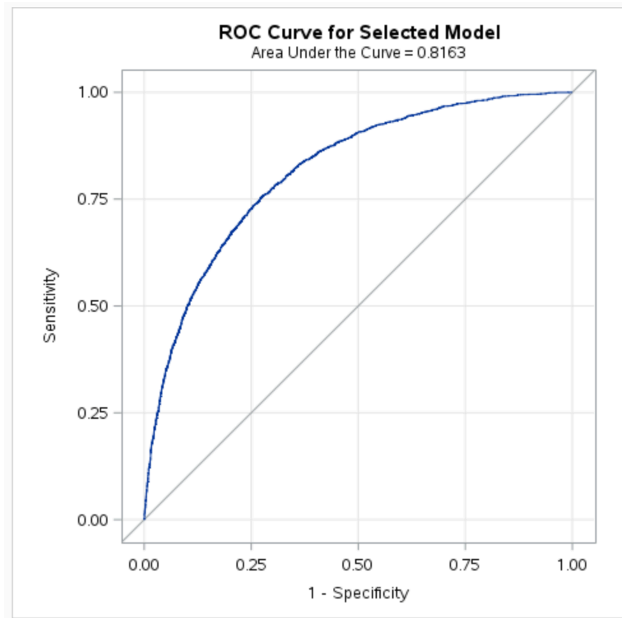

Figure 9: Model 1 ROC Curve

Model 2

For this model, a slightly different tactic was taken. Each variable's predictability was compared to the response variable of TARGET_FLAG. This enabled a greater understanding if a variable was predictable and useful to the overall modelling process. For instance, in this model, the variable RED_CAR or SEX were not provided as available variables for the model. It was found, during the EDA process, that these variables did not necessarily add predictability to the response variable. This model utilizes a backward variable selection method. Table 9 provides an overview of the variables chosen and their respective coefficients.

Table 9: Model 2 Parameter Estimates

Variable	Variable Subelement	Variable Code	Coefficient Notation	Coefficient Value
Intercept			β0	-0.8466
f_renter		X1	β1	0.3419
f_claimchk		X2	β2	0.6536
f_newjobchk		X3	β3	-0.2973
imp_income		X4	β4	-6.54E-06
imp_oldclaim		X5	β5	-0.00002
out_income		X6	β6	0.9073
out_oldclaim		X7	β7	9.0288
KIDSDRIV		X8	β8	0.4073
TIF		X9	β9	-0.0543
MVR_PTS		X10	β10	0.0987

EDUCATION	<High School	X11	β11	-0.00851
EDUCATION	Bachelors	X12	β12	-0.3985
EDUCATION	Masters	X13	β13	-0.3339
EDUCATION	PhD	X14	β14	-0.3684
imp_job	Clerical	X15	β15	0.0424
imp_job	Doctor	X16	β16	-0.4455
imp_job	Home Maker	X17	β17	-0.2172
imp_job	Lawyer	X18	β18	-0.2051
imp_job	Manager	X19	β19	-0.8784
imp_job	Professional	X20	β20	-0.1607
imp_job	Student	X21	β21	-0.2978
URBANICITY	Highly Urban/ Urban	X22	β22	2.2258
PARENT1	No	X23	β23	-0.4243
REVOKED	No	X24	β24	-0.9586
CAR_TYPE	Minivan	X25	β25	-0.7643
CAR_TYPE	Panel Truck	X26	β26	-0.4419
CAR_TYPE	Pickup	X27	β27	-0.1611
CAR_TYPE	Sports Car	X28	β28	0.2635
CAR_TYPE	Van	X29	β29	-0.2436
MSTATUS	Yes	X30	β30	-0.4397
CAR_USE	Commercial	X31	β31	0.7375

This model utilizes 27 predictor variables (in the table above, the categorical variables are listed multiple times) resulting in a chosen number of 18 predictor variables. This is in contrast to Model 1, which had 25 unique variables in the model. Looking at the various coefficients, there are some interesting conclusions that could be drawn by simply looking at the signs of each coefficient. For instance, a Minivan has a negative coefficient. This could be construed as Minivan is less likely to be involved in a vehicular accident since the driver may be transporting children and is driving safer. This is in contrast to the Sports Car which has a positive coefficient, suggesting that is more likely to be involved in a vehicular accident. An interesting peculiarity is the fact that a commercial car use significantly increases the probability of an accident, but the use of a panel truck has a significant reduction in the likelihood of an accident. A further point of investigation may be if a customer is using a vehicle for both personal and commercial use, but at the time of claim, the customer may have claimed the use on a commercial policy.

Figure 10 illustrates three key model fit statistics. Of key importance is to note that the values for these statistics is reduced when exploring both the intercept and covariates. Note that the lower the value of these statistics, the stronger the indicator that the model may be a better model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7418.277
SC	9426.969	7642.505
-2 Log L	9417.962	7354.277

Figure 10: Model 2 Fit Statistics

Figure 11 illustrates the ROC curve for this model. The ROC curve simply describes the amount true positive rate versus the false positive rate. Essentially, the closer the curve is to the left hand side and towards the top, the truer positive the rate. However, this would most likely be a true indicator of an over-fitted model. In this case, the goal is to balance the curve without overfitting the model. For this model, the metric of the ROC curve is 0.8101 (a perfect curve would be 1 – a potential indicator of an over-fitted model).

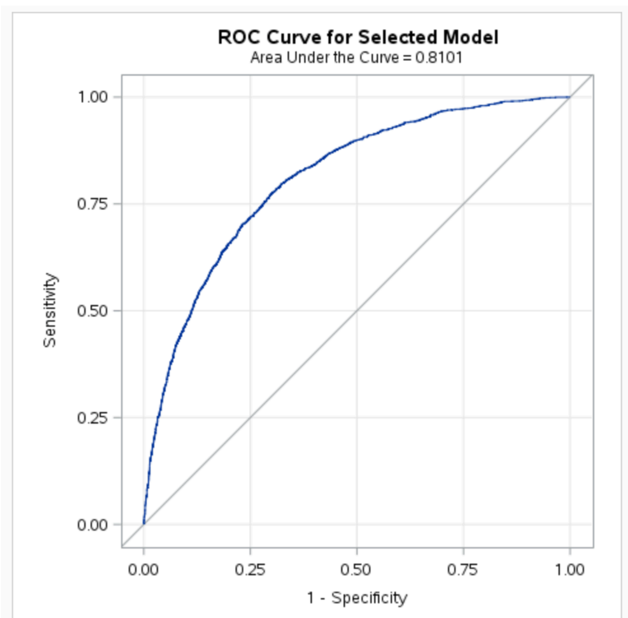

Figure 11: Model 2 ROC Curve

Model 3

For this model, all of the default variables were used without any imputation, outlier detection, missing value detection, etc. This model is most definitely wrong since it contains variables that have not been corrected for errors. As in the other models, the response variable continues to TARGET_FLAG. Table 10 summarizes the parameter estimates for this model.

Table 10: Model 3 Parameter Estimates

Variable	Variable Subelement	Variable Code	Coefficient Notation	Coefficient Value
Intercept			β0	-1.1937

EDUCATION	<High School	X1	β1	0.1709
EDUCATION	Bachelors	X2	β2	-0.2306
EDUCATION	Masters	X3	β3	-0.3229
EDUCATION	PhD	X4	β4	0.2179
JOB	Clerical	X5	β5	0.4064
JOB	Doctor	X6	β6	-0.5026
JOB	Home Maker	X7	β7	0.1165
JOB	Lawyer	X8	β8	0.2448
JOB	Manager	X9	β9	-0.6817
JOB	Professional	X10	β10	0.1125
JOB	Student	X11	β11	0.0933
URBANICITY	Highly Urban/ Urban	X12	β12	1.1533
PARENT1	No	X13	β13	-0.2363
REVOKED	No	X14	β14	-0.4272
CAR_TYPE	Minivan	X15	β15	-0.6216
CAR_TYPE	Panel Truck	X16	β16	0.118
CAR_TYPE	Pickup	X17	β17	-0.0705
CAR_TYPE	Sports Car	X18	β18	0.4339
CAR_TYPE	Van	X19	β19	-0.0212
MSTATUS	Yes	X20	β20	-0.2077
CAR_USE	Commercial	X21	β21	0.4148
KIDSDRIV		X22	β22	0.3292
INCOME		X23	β23	-3.50E-06
HOME_VAL		X24	β24	-1.44E-06
TRAVTIME		X25	β25	0.0156
BLUEBOOK		X26	β26	-0.00002
TIF		X27	β27	-0.0524
OLDCLAIM		X28	β28	-0.00001
CLM_FREQ		X29	β29	0.1995
MVR_PTS		X30	β30	0.1165

This model also utilizes 30 predictor variables, but note how not all values were used to create this model (see Figure 12).

Number of Observations Read	8161
Number of Observations Used	6045

Figure 12: Model 3 observations summary

Note how some of the coefficients are in contrast to what was seen earlier. In this model, the student has a higher likelihood of being involved in a vehicular accident as compared to what was seen in Model 2. The same is true for the occupation of Professional. It is possible that this could be attributed to the fact that an individual with a higher paying job may be more likely to purchase a more expensive vehicle or drive at a higher speed (thus increasing the likelihood of being involved in an accident).

In this model, over 2000 of the observations were excluded by SAS during the model creation. This is virtually unacceptable as it eliminates the model's ability to accurately predict. However, this model is included to highlight how its AIC value is far below that of the other values. Recall that the AIC calculation relies on the number of observations. As the number of observations are reduced, it is inherent that the overall AIC value will also be lower (see Figure 13 and compare to Figure 10).

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6992.858	5432.223
SC	6999.565	5640.140
-2 Log L	6990.858	5370.223

Figure 13: Model 3 Fit Statistics

Figure 14 illustrates the ROC curve for this model. The ROC curve simply describes the amount true positive rate versus the false positive rate. Essentially, the closer the curve is to the left hand side and towards the top, the truer positive the rate. However, this would most likely be a true indicator of an over-fitted model. In this case, the goal is to balance the curve without overfitting the model. For this model, the metric of the ROC curve is 0.8180 (a perfect curve would be 1 – a potential indicator of an over-fitted model).

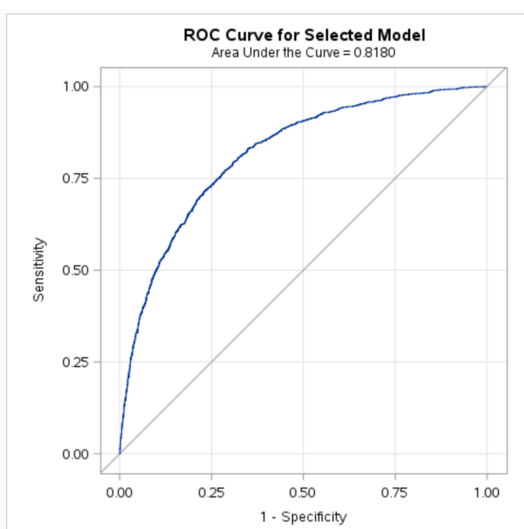

Figure 14: Model 3 ROC Curve

Model Selection

For the model selection process, three key metrics will be looked at: AIC, AUR (area under ROC curve), and the KS statistic. The KS statistic was determined using the PROC NPAR1WAY SAS procedure in order to develop the values. This value is not provided using PROC LOGISTIC. In the table below (Table 11), the D value is shown as it is the correct representation of the KS statistic. The AIC metric utilizes the intercept and covariate values. The D value is actually the two sample K-S test and analyzes the difference in both location and shape of the cumulative distribution functions of the two samples. The lower this number, the better the model (potentially). With the KS Statistic, the closer to 1, the more explanatory the model may be.

Table 11: Summary of fit statistics by model

Model	AIC	AUR	D Statistic (part of KS)	KS Statistic
Model 1	7328.283	0.8163	0.481161	0.212048
Model 2	7418.277	0.8101	0.474232	0.208994
Model 3	5432.223	0.8180	0.484227	0.213930

Based on these metrics alone, it would appear that Model 3 is the best model, followed by Model 1, and then by Model 2. However, at this point, it is important to point out that these metrics do not necessarily take into account the inherent cost of the model. By far, the parsimonious model is Model 2 as it only uses variables that have predictability against the response variable of TARGET_FLAG. Furthermore, as the D statistic is the maximum deviation between the two random samples, the minimal value would suggest that the model has better predictability. It is for this reason that Model 2 is the suggested model to use. This model is by no means the ‘perfect’ model as it does have some peculiarities. For instance, it is surprising to note that having an occupation of “clerical” tends to increase the likelihood of an individual being involved in a vehicular accident when compared to the other occupations. Another interesting case is how a commercial usage increases the likelihood of an accident, but a panel truck – which tends to be intuitively related to commercial usage – tends to have reduced likelihood of being involved in an accident.

Prior to deployment, it would be prudent to consult a subject matter expert to review each parameter’s sign. For instance, if a parameter is negative, does that make sense? Furthermore, it would also be wise to see if additional data sources or data points could be retrieved and potentially implemented into the model. Nevertheless, as this is the first model, implementing into production will enable the team to understand it’s performance. It is suggested that this model be monitored for a period of three months (at a minimum) to better understand it’s performance as new data are consumed.

CONCLUSION

Ten different models were developed (of which only three have been discussed in this analysis) to predict the likelihood of an individual being involved in a vehicular accident. The original dataset contained 23 distinct variables (a mixture of numeric and categorical types) with over 8000 observations. In the insurance world, there are much more data available that could potentially improve the overall model. One legal concern is the use of income to potentially penalize customers which may considered illegal. For this analysis, many of the data points have been adjusted for being outliers or imputed if missing. Although a model has been chosen, future work is required to better understand some of the non-intuitive sign issues. This will require further investigation which is outside the scope of this analysis. Finally, it would be prudent to monitor this chosen model’s performance for the next three months as new data are fed into it. Next steps may include (but not limited to) refining the model and focusing on improved data imputation.

BINGO BONUS: PROBIT MODEL

For the bingo bonus, it was asked to construct a model using the PROBIT option.

The chosen model for this assignment was reran using the PROBIT option. Note how the ROC curve (see Figure 15) has an AUC value of 0.81. This is roughly the same as what was seen in Model 2.

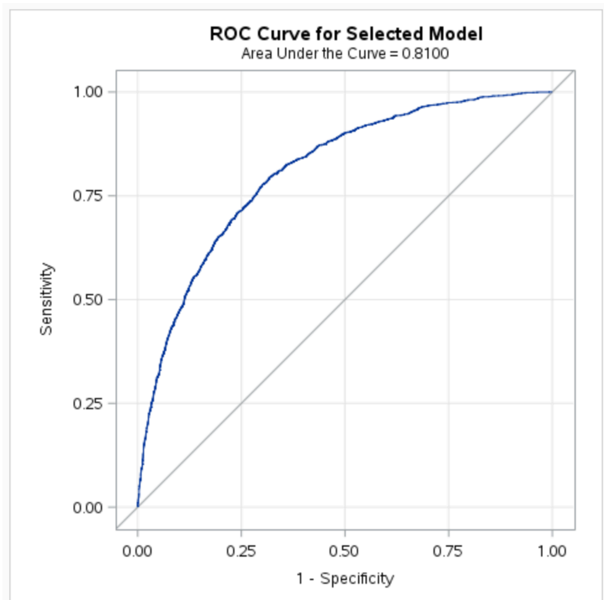

Figure 15: ROC Curve - PROBIT model

Note the AIC value of 7428.019 in Figure 16. This is a slightly higher value than what was seen for Model 2.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	9419.962	7428.019
SC	9426.969	7652.247
-2 Log L	9417.962	7364.019

Figure 16: Fit Statistics - PROBIT model

Table 12 highlights the parameter estimates for this model. Note how this model also has 31 parameter estimates – similar to Model 2.

Table 12: Parameter estimates - PROBIT model

Variable	Variable Subelement	Variable Code	Coefficient Notation	Coefficient Value
Intercept			β0	-0.4387
f_renter		X1	β1	0.1981
f_claimchk		X2	β2	0.3885
f_newjobchk		X3	β3	-0.1704
imp_income		X4	β4	-3.65E-06
imp_oldclaim		X5	β5	-0.00001
out_income		X6	β6	0.4913
out_oldclaim		X7	β7	5.248
KIDSDRIV		X8	β8	0.2336
TIF		X9	β9	-0.0321
MVR_PTS		X10	β10	0.0581
EDUCATION	<High School	X11	β11	-0.012
EDUCATION	Bachelors	X12	β12	-0.2352
EDUCATION	Masters	X13	β13	-0.19
EDUCATION	PhD	X14	β14	-0.1992
imp_job	Clerical	X15	β15	0.024
imp_job	Doctor	X16	β16	-0.2863
imp_job	Home Maker	X17	β17	-0.1362
imp_job	Lawyer	X18	β18	-0.134
imp_job	Manager	X19	β19	-0.4991
imp_job	Professional	X20	β20	-0.0985
imp_job	Student	X21	β21	-0.1588
URBANICITY	Highly Urban/ Urban	X22	β22	1.2191
PARENT1	No	X23	β23	-0.2427
REVOKED	No	X24	β24	-0.5559
CAR_TYPE	Minivan	X25	β25	-0.4315
CAR_TYPE	Panel Truck	X26	β26	-0.259
CAR_TYPE	Pickup	X27	β27	-0.0992
CAR_TYPE	Sports Car	X28	β28	0.1506
CAR_TYPE	Van	X29	β29	-0.1403
MSTATUS	Yes	X30	β30	-0.2555
CAR_USE	Commercial	X31	β31	0.4151

Finally, Figure 17 shows the KS statistic. Note how this D value (0.484227) is higher than the D value for Model 2 (0.474232). This is not a large difference, but it is noteworthy.

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.213930	D	0.484227
KSa	17.665778	Pr > KSa	<.0001

Figure 17: KS Statistic output

BINGO BONUS: CLOGCLOG MODEL

As an additional bingo bonus opportunity, Model 2 was reran using the CLOGCLOG option. Note how the area under the curve in Figure 18 is 0.8092. This is similar to what Model 2 had and similar to the ROC curve using the PROBIT option (see Figure 15).

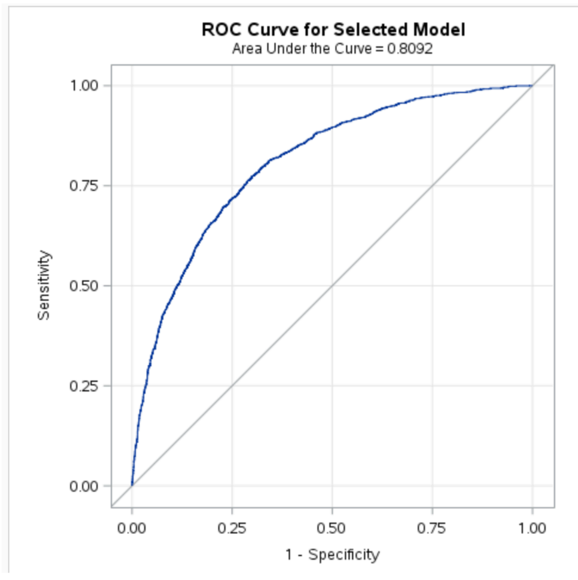

Figure 18: ROC curve - CLOGCLOG model

Note how the AIC value in Figure 19 is a bit higher than that of the PROBIT option (see Figure 16).

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	9419.962	7446.808
SC	9426.969	7671.036
-2 Log L	9417.962	7382.808

Figure 19: Fit Statistics - CLOGCLOG model

Table 1 highlights the parameter estimates for this model. Note how this model also has 31 parameter estimates – similar to Model 2 and the PROBIT option. Recall that this equation simply produces the Log Odds value. This value must then be converted to a probability using the following equation (which is different from the PROBIT & LOGIT options):

$$P\_TARGET\_FLAG \ = \ e^{-1*e^{LogOdds}}$$

Table 13: Parameter Estimates - CLOGCLOG model

Variable	Variable Subelement	Variable Code	Coefficient Notation	Coefficient Value
Intercept			β0	-1.0896
f_claimchk		X1	β1	0.4946
f_newjobchk		X2	β2	-0.1912
imp_income		X3	β3	-3.75E-06
imp_homeval		X4	β4	-1.23E-06
imp_oldclaim		X5	β5	-0.00002
out_income		X6	β6	0.7288
out_oldclaim		X7	β7	6.8258
KIDSDRIV		X8	β8	0.2629
TIF		X9	β9	-0.041
MVR_PTS		X10	β10	0.0741
EDUCATION	<High School	X11	β11	-0.0267
EDUCATION	Bachelors	X12	β12	-0.3067
EDUCATION	Masters	X13	β13	-0.2621
EDUCATION	PhD	X14	β14	-0.3112
imp_job	Clerical	X15	β15	-0.00189
imp_job	Doctor	X16	β16	-0.3998
imp_job	Home Maker	X17	β17	-0.176
imp_job	Lawyer	X18	β18	-0.177
imp_job	Manager	X19	β19	-0.7535
imp_job	Professional	X20	β20	-0.155
imp_job	Student	X21	β21	-0.2159
URBANICITY	Highly Urban/ Urban	X22	β22	1.8619
PARENT1	No	X23	β23	-0.2698
REVOKED	No	X24	β24	-0.7512
CAR_TYPE	Minivan	X25	β25	-0.5906
CAR_TYPE	Panel Truck	X26	β26	-0.2705
CAR_TYPE	Pickup	X27	β27	-0.0767
CAR_TYPE	Sports Car	X28	β28	0.2109
CAR_TYPE	Van	X29	β29	-0.1557
MSTATUS	Yes	X30	β30	-0.3387
CAR_USE	Commercial	X31	β31	0.5373

Finally, Figure 20 shows the KS statistic. Note how this D value (0.474107) is lower than the D for the PROBIT model (0.484227) and very similar to the D value for Model 2 (0.474232).

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.208939	D	0.474107
KSa	18.875194	Pr > KSa	<.0001

Figure 20: KS Statistic - CLOGCLOG model