

DATA 621 - Homework 4

2022-11-16

Problem Statement and Goals

In this report, we generate two different models; a multiple linear regression model and a binary logistic regression model. The multiple linear regression model contains a target variable called **TARGET_AMT**, which is the amount of money it will cost if the person crashes their car. The binary logistic regression model target variable, **TARGET_FLAG** consists of 0's and 1's. 1 represents that the person was in a car crash, and zero indicates that the person was not in a car crash. The analysis detailed in this report shows the testing of several models from which a best multiple linear regression model and a best binary logistic regression model were selected based on model performance and various metrics.

Data Exploration

The following is a summary of the variables provided within the data to generate the binary logistic regression and multiple linear regression models.

Variable Name	Definition	Theoretical Effect
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect

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

A summary of the variables is shown below. The INDEX variable has been removed. The summary below reveals that AGE, VOJ, INCOME, HOME_VAL, and CAR_AGE have missing values.

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS
0:6008	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000
1:2153	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000
	Median : 0	Median :0.0000	Median :45.00	Median :0.0000
	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212
	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000
	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000
			NA's :6	
YOJ	INCOME	PARENT1	HOME_VAL	MSTATUS
Min. : 0.0	Min. : 0	No :7084	Min. : 0	No :3267
1st Qu.: 9.0	1st Qu.: 28097	Yes:1077	1st Qu.: 0	Yes:4894
Median :11.0	Median : 54028		Median :161160	
Mean :10.5	Mean : 61898		Mean :154867	
3rd Qu.:13.0	3rd Qu.: 85986		3rd Qu.:238724	

Max.	:23.0	Max.	:367030	Max.	:885282		
NA's	:454	NA's	:445	NA's	:464		
SEX		EDUCATION		JOB			
F:4375	<High School:1203	Blue Collar	:1825	Min.	: 5.00		
M:3786	Bachelors :2242	Clerical	:1271	1st Qu.:	22.00		
	High School :2330	Professional:	1117	Median	: 33.00		
	Masters :1658	Manager	: 988	Mean	: 33.49		
	PhD : 728	Lawyer	: 835	3rd Qu.:	44.00		
		Student	: 712	Max.	:142.00		
		(Other)	:1413				
CAR_USE		BLUEBOOK		TIF			
Commercial:3029	Min.	: 1500	Min.	: 1.000	Minivan	:2145	
Private :5132	1st Qu.:	9280	1st Qu.:	1.000	Panel Truck:	676	
	Median	:14440	Median	: 4.000	Pickup	:1389	
	Mean	:15710	Mean	: 5.351	Sports Car	: 907	
	3rd Qu.:	20850	3rd Qu.:	7.000	SUV	:2294	
	Max.	:69740	Max.	:25.000	Van	: 750	
RED_CAR		OLDCLAIM		CLM_FREQ			
no :5783	Min.	: 0	Min.	:0.0000	REVOKED		
yes:2378	1st Qu.:	0	1st Qu.:	0.0000	No :7161	Min.	: 0.000
	Median	: 0	Median	:0.0000	Yes:1000	1st Qu.:	0.000
	Mean	: 4037	Mean	:0.7986		Median	: 1.000
	3rd Qu.:	4636	3rd Qu.:	2.0000		Mean	: 1.696
	Max.	:57037	Max.	:5.0000		3rd Qu.:	3.000
						Max.	:13.000
CAR_AGE		URBANICITY					
Min.	:-3.000	Highly Rural/	Rural:1669				
1st Qu.:	1.000	Highly Urban/	Urban:6492				
Median	: 8.000						
Mean	: 8.328						
3rd Qu.:	12.000						
Max.	:28.000						
NA's	:510						

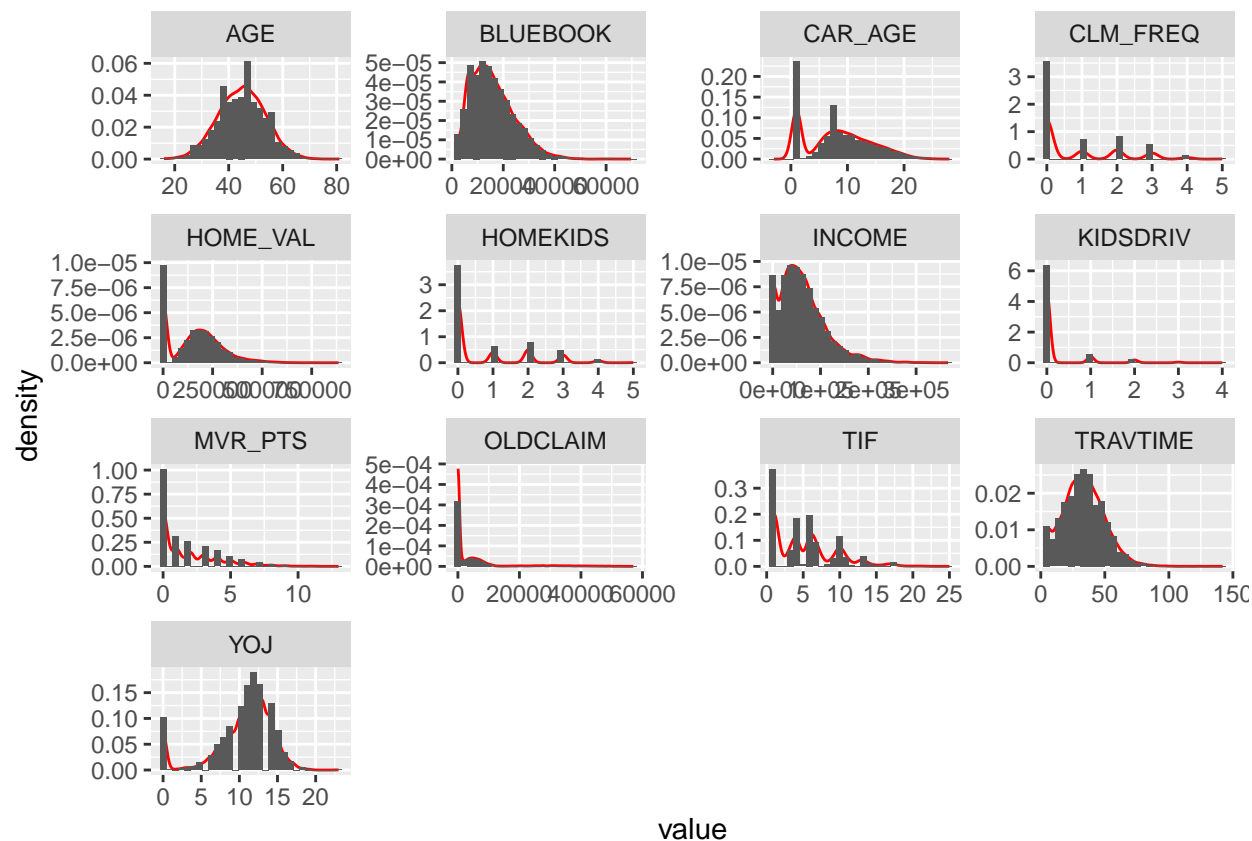


Figure XX: Histograms for all the variables.

The density plots above show that BLUEBOOK, INCOME, and TRAVTIME could be transformed in order to fit the normal distribution assumption of a linear regression model. The variables with a bimodal distribution were dealt with and an explanation of the process is provided in the “Dealing with Bimodal Variables” section.

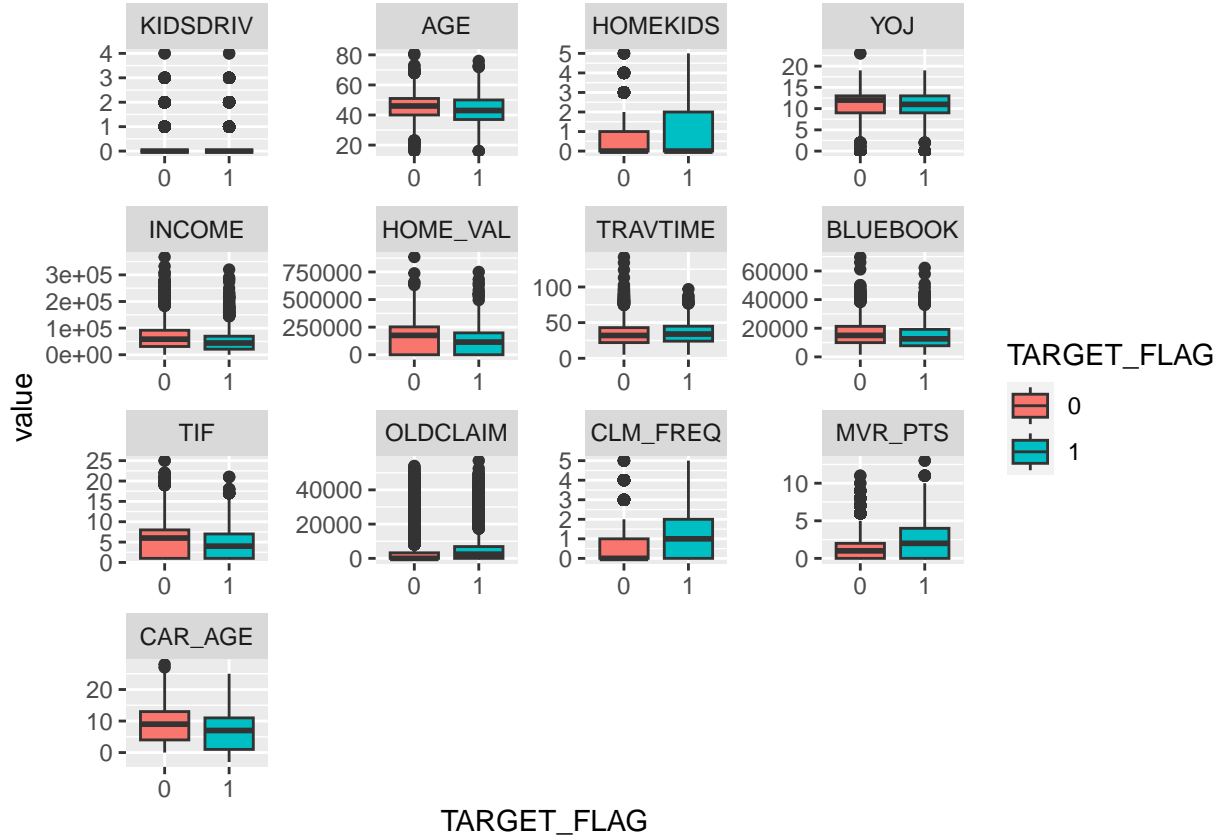


Figure XX: Boxplots for the dataset

We can see some findings that support the theoretical effects for some of the variables using the boxplots in Figure xx. It seems that younger cars are more likely to get into crashes as opposed to older cars as shown in the `CAR_AGE` boxplot. The theoretical effect of the `CLM_FREQ` (The more claims you filed in the past, the more you are likely to file in the future) is supported by the `CLM_FREQ` boxplot. The theoretical effect of `MVRPTS` (If you get lots of traffic tickets, you tend to get into more crashes) is supported by the `MVRPTS` boxplot. It would also seem that the theoretical effects of `INCOME` and `TIF` are also supported by the data.

Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features. The correlation plot is only able to determine the correlation for continuous variables. There are methodologies to determine correlations for categorical variables (tetrachoric correlation). However there is only one binary predictor variable which is why the multicollinearity will only be considered for the continuous variables.

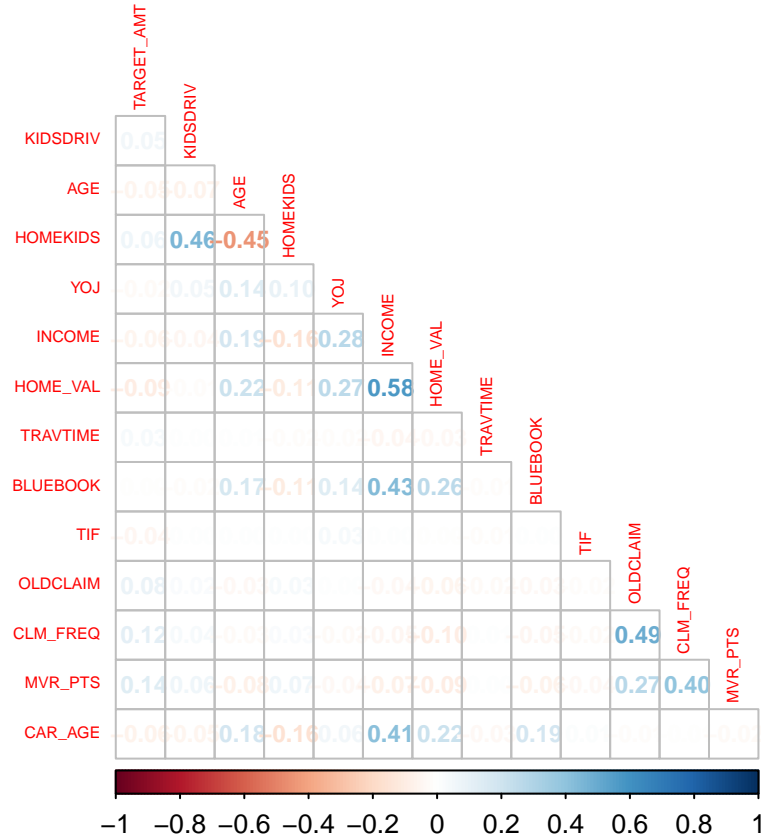


Figure xx: Multicollinearity plot for continuous predictor variables

The figure above shows that there isn't much multicollinearity between the variables. There is a moderately positive correlation of 0.58 between INCOME and HOME_VAL.

NA exploration

As can be seen below, some of the columns have missing values. Contextually, this can be possible because not every metric must have a value- for example it is possible that an entire season can be played without a batter being hit by the pitch. However it is less likely that an entire season can be played without any strikeouts by batters. We did some research and came up with ways to address each of these issues- more on that later.

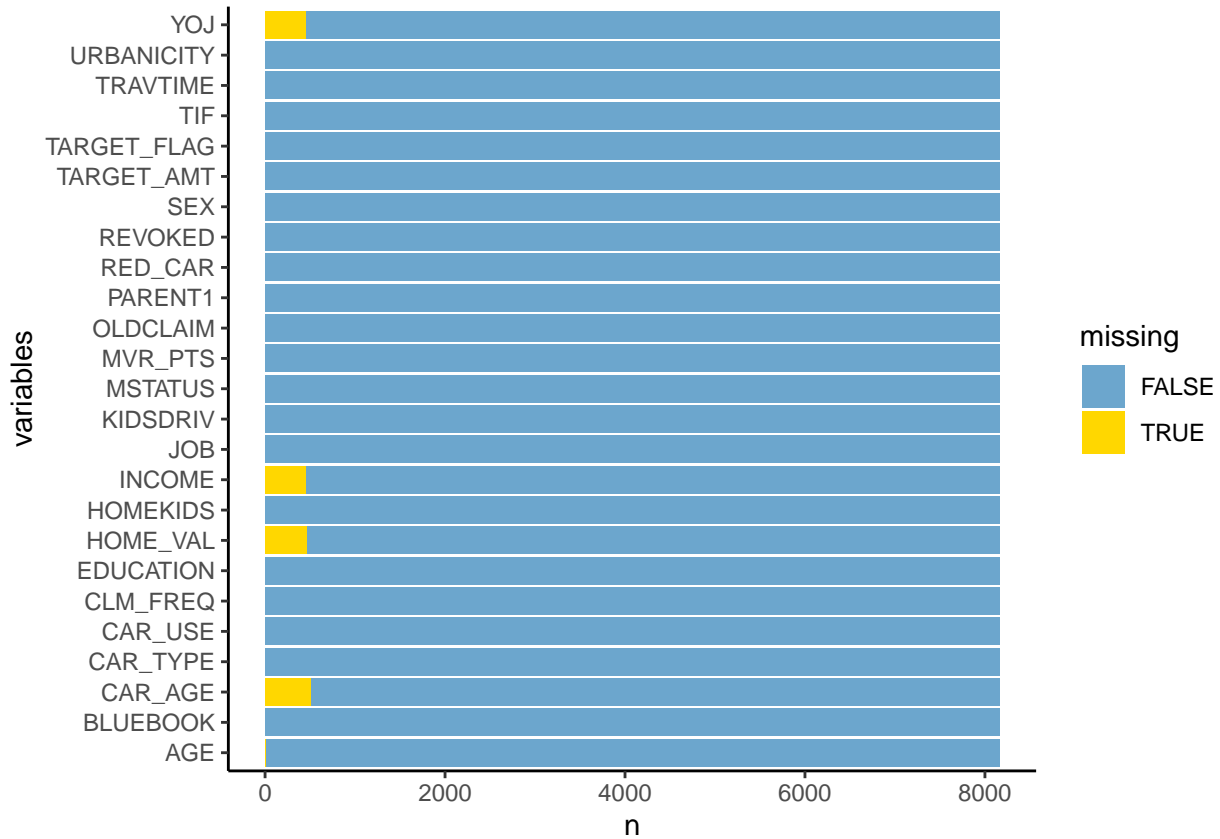


Figure xx: Barplot of number of missing values for each predictor.

The barplot above shows that Y0J, INCOME, HOME_VAL, AGE, and CAR_AGE were missing some data values. However, the amount of missing data for each variable is less than 10%. Therefore, imputing can be done on the missing data.

Data Preparation

Dealing with Missing Values

In general, imputations by the means/medians is acceptable if the missing values only account for 5% of the sample. Peng et al.(2006) However, should the degree of missing values exceed 20% then using these simple imputation approaches will result in an artificial reduction in variability due to the fact that values are being imputed at the center of the variable's distribution.

Our team decided to employ another technique to handle the missing values: Multiple Regression Imputation using the MICE package.

The MICE package in R implements a methodology where each incomplete variable is imputed by a separate model. Alice points out that plausible values are drawn from a distribution specifically designed for each missing datapoint. Many imputation methods can be used within the package. The one that was selected for the data being analyzed in this report is PMM (Predictive Mean Matching), which is used for quantitative data.

Van Buuren explains that PMM works by selecting values from the observed/already existing data that would most likely belong to the variable in the observation with the missing value. The advantage of this is that it selects values that must exist from the observed data, so no negative values will be used to impute missing data. Not only that, it circumvents the shrinking of errors by using multiple regression models. The variability between the different imputed values gives a wider, but more correct standard error. Uncertainty

is inherent in imputation which is why having multiple imputed values is important. Not only that. Marshall et al. 2010 points out that:

“Another simulation study that addressed skewed data concluded that predictive mean matching ‘may be the preferred approach provided that less than 50% of the cases have missing data...’

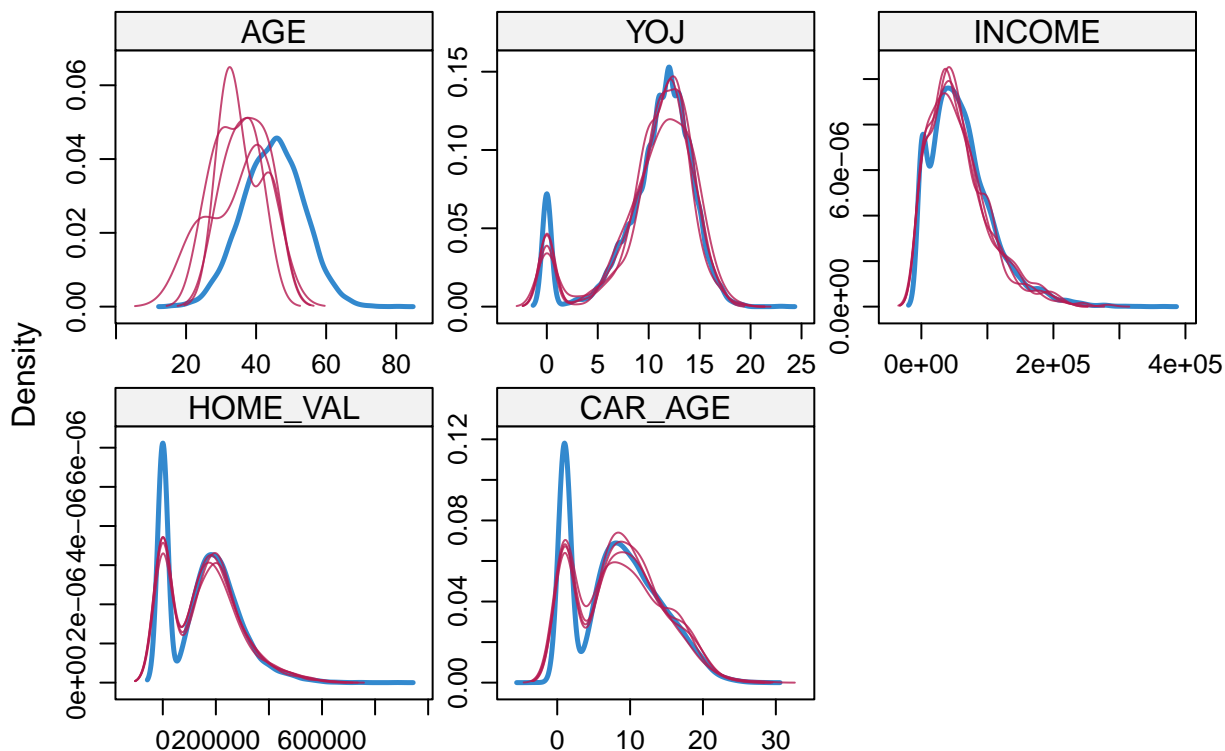


Figure xx: Density plots for variables containing missing data. The number of multiple imputations was set to 4. Each of the red lines represents the distribution for each imputation.

The blue lines for each of the graphs above represent the distributions the non-missing data for each of the variables while the red lines represent the distributions for the imputed data. Note that the distributions for the imputed data for each of the iterations closely matches the distributions for the non-missing data, which is ideal. If the distributions did not match so well, than another imputing method would have had to have been used.

Feature Manipulation based on Multicollinearity Plot

There is a significant amount of observations for INCOME with a value of 0. Therefore, we reasoned that we could create a new dummy variable based on the INCOME, where 0 was unemployed and any positive value for income would be employed. Then, we could effectively be rid of the INCOME variable while still having some sort of distinction that represents this variable that does not have a high correlation with any of the other variables.

Dealing with Bimodal Variables

Bimodal distributions in data are interesting, in that they represent features which actually contain multiple (2) inherent systems resulting in separated distributional peaks. While a Box-Cox transformation could have been undertaken in order to transform the bimodal variables to a normal distribution. However, this throws away important information that is inherent in the bimodal variable itself. The fact that the variable is

bimodal in the first place is essentially ignored, and the predicted values in the linear multiple regression model will not reflect this bimodality.

For variables that displayed bimodality, new variables were created; `bi_CAR_AGE`, `bi_CLM_FREQ`, `bi_HOME_VAL`, `bi_KIDSDRIV`, `bi_YOJ`. For many of these variables, there are a significant number of 0 values, which results in the bimodal distributions shown above, so 0 will represent observations with a value of 0 and 1 will represent any observations with a value greater than 0. For `CAR_AGE`, many cars are 1 years old, so 0 represents observations where the `CAR_AGE` is 1, while 1 represents any observations with a value greater than 1.

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS
0:6008	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000
1:2153	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000
	Median : 0	Median :0.0000	Median :45.00	Median :0.0000
	Mean : 1504	Mean :0.1711	Mean :44.78	Mean :0.7212
	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000
	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000

YOJ	PARENT1	HOME_VAL	MSTATUS	SEX
Min. : 0.00	No :7084	Min. : 0	No :3267	F:4375
1st Qu.: 9.00	Yes:1077	1st Qu.: 0	Yes:4894	M:3786
Median :11.00		Median :161549		
Mean :10.51		Mean :154921		
3rd Qu.:13.00		3rd Qu.:238630		
Max. :23.00		Max. :885282		

EDUCATION	JOB	TRAVTIME	CAR_USE
<High School:1203	Blue Collar :1825	Min. : 5.00	Commercial:3029
Bachelors :2242	Clerical :1271	1st Qu.: 22.00	Private :5132
High School :2330	Professional:1117	Median : 33.00	
Masters :1658	Manager : 988	Mean : 33.49	
PhD : 728	Lawyer : 835	3rd Qu.: 44.00	
	Student : 712	Max. :142.00	
	(Other) :1413		

BLUEBOOK	TIF	CAR_TYPE	RED_CAR	OLDCLAIM
Min. : 1500	Min. : 1.000	Minivan :2145	no :5783	Min. : 0
1st Qu.: 9280	1st Qu.: 1.000	Panel Truck: 676	yes:2378	1st Qu.: 0
Median :14440	Median : 4.000	Pickup :1389		Median : 0
Mean :15710	Mean : 5.351	Sports Car : 907		Mean : 4037
3rd Qu.:20850	3rd Qu.: 7.000	SUV :2294		3rd Qu.: 4636
Max. :69740	Max. :25.000	Van : 750		Max. :57037

CLM_FREQ	REVOKED	MVR_PTS	CAR_AGE
Min. :0.0000	No :7161	Min. : 0.000	Min. : -3.000
1st Qu.:0.0000	Yes:1000	1st Qu.: 0.000	1st Qu.: 1.000
Median :0.0000		Median : 1.000	Median : 8.000
Mean :0.7986		Mean : 1.696	Mean : 8.331
3rd Qu.:2.0000		3rd Qu.: 3.000	3rd Qu.:12.000
Max. :5.0000		Max. :13.000	Max. :28.000

URBANICITY	EMPLOYMENT	bi_CAR_AGE	bi_CLM_FREQ	bi_HOME_VAL
Highly Rural/ Rural:1669	Min. :0.0000	0:2059	0:5009	0:2436
Highly Urban/ Urban:6492	1st Qu.:1.0000	1:6102	1:3152	1:5725
	Median :1.0000			
	Mean :0.9188			
	3rd Qu.:1.0000			

Max. :1.0000

```
bi_KIDSDRIV bi_Y0J
0:7180      0: 660
1: 981      1:7501
```

Box-Cox Transformation for Skewed Variables

Based on the previous distribution plot (using histograms) we noticed that a select group of columns exhibited non-normal skew. In order to address this skewness and attempt to normalize these features for future modeling, we will employ box-cox transformations. Because some of these values include 0, we will need to replace any zero values with infinitesimally small, non-zero values.

The λ 's that were used to transform the skewed variables are shown on Table 2.

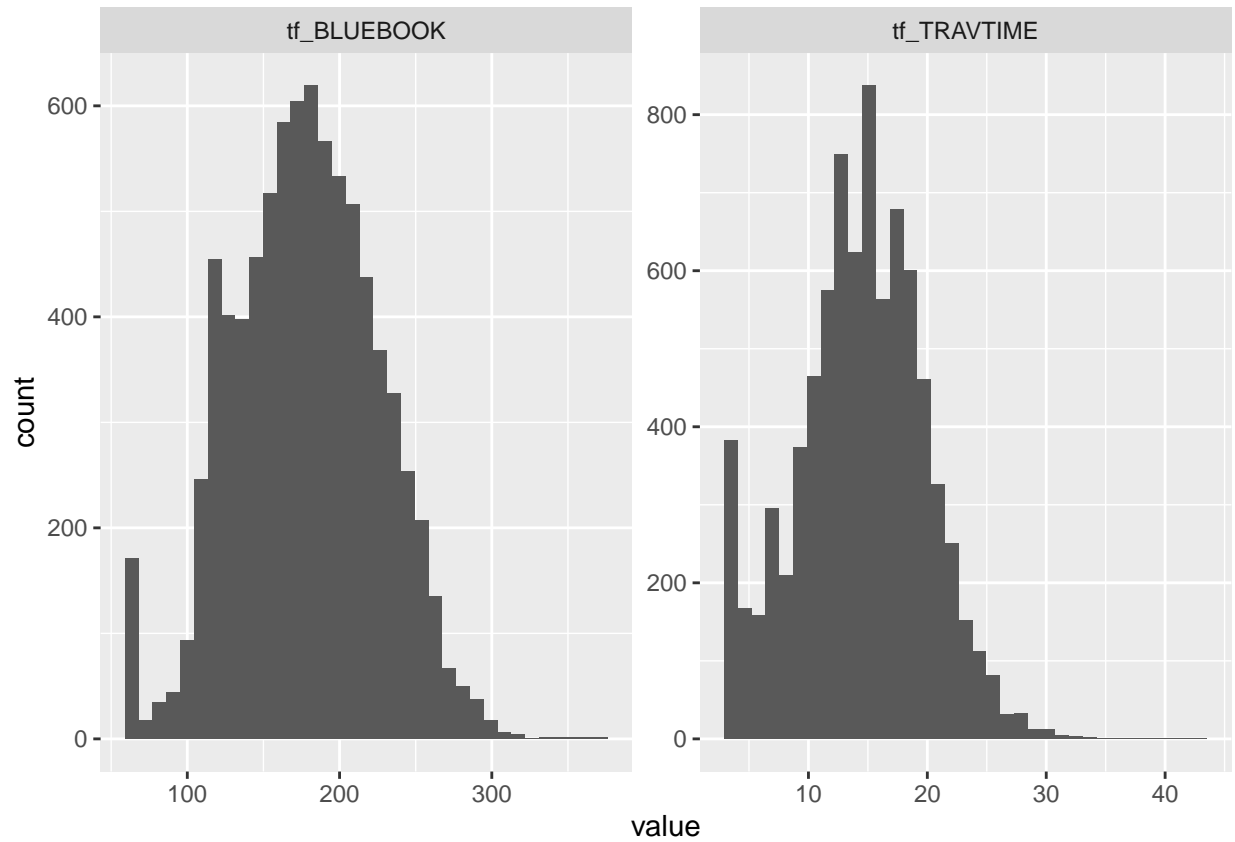


Figure 9: Histograms for transformed variables.

Column Name	λ
BLUEBOOK	0.461
TRAVTIME	0.687

Table 2: λ 's for skewed variables.

Split Data Into Testing and Training

The data was into testing and training subsets such that 60% of it will be used to train, and 40% to test. The first row shows the split for the testing data while the second row shows the split for the training data.

```
0    1
1202 431
```

```
0    1
4806 1722
```

Build Models

Binary Logistic Regression Model with Original Variables

Call:

```
glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
     data = original_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.5666	-0.7165	-0.4059	0.6457	3.1729

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.914e+00	4.235e-01	-6.882	5.91e-12 ***
KIDSDRIV	3.419e-01	7.628e-02	4.482	7.41e-06 ***
AGE	-8.291e-03	5.079e-03	-1.632	0.102631
HOMEKIDS	3.082e-02	4.632e-02	0.665	0.505782
YOJ	2.398e-03	1.078e-02	0.222	0.823952
INCOME	-3.379e-06	1.426e-06	-2.370	0.017792 *
PARENT1Yes	3.832e-01	1.371e-01	2.796	0.005181 **
HOME_VAL	-9.392e-07	4.380e-07	-2.144	0.032021 *
MSTATUSYes	-3.896e-01	1.081e-01	-3.604	0.000314 ***
SEX	2.633e-01	1.368e-01	1.925	0.054227 .
EDUCATIONBachelors	-3.822e-01	1.470e-01	-2.600	0.009318 **
EDUCATIONHigh School	-7.199e-02	1.200e-01	-0.600	0.548562
EDUCATIONMasters	-3.745e-01	2.300e-01	-1.628	0.103502
EDUCATIONPhD	-8.301e-02	2.742e-01	-0.303	0.762122
JOBBlue Collar	3.301e-01	2.357e-01	1.400	0.161432
JOBclerical	5.640e-01	2.489e-01	2.266	0.023459 *
JOBDoctor	-1.766e-01	3.184e-01	-0.555	0.579141
JOBHome Maker	3.696e-01	2.683e-01	1.377	0.168380
JOBLawyer	4.619e-01	2.118e-01	2.181	0.029213 *
JOBManager	-5.162e-01	2.212e-01	-2.334	0.019601 *
JOBProfessional	2.335e-01	2.267e-01	1.030	0.303096
JOBStudent	2.969e-01	2.762e-01	1.075	0.282419
TRAVTIME	1.509e-02	2.367e-03	6.374	1.85e-10 ***
CAR_USEPrivate	-8.155e-01	1.156e-01	-7.058	1.70e-12 ***
BLUEBOOK	-2.024e-05	6.528e-06	-3.100	0.001936 **
TIF	-4.928e-02	9.245e-03	-5.330	9.82e-08 ***
CAR_TYPEPanel Truck	5.404e-01	2.021e-01	2.673	0.007508 **

CAR_TYPEPickup	5.430e-01	1.251e-01	4.340	1.43e-05	***
CAR_TYPESports Car	1.174e+00	1.613e-01	7.278	3.40e-13	***
CAR_TYPESUV	9.089e-01	1.371e-01	6.631	3.33e-11	***
CAR_TYPEVan	6.673e-01	1.583e-01	4.216	2.48e-05	***
RED_CARyes	-1.483e-01	1.076e-01	-1.379	0.168044	
OLDCLAIM	-1.377e-05	4.828e-06	-2.851	0.004351	**
CLM_FREQ	2.278e-01	3.549e-02	6.419	1.37e-10	***
REVOKEDYes	8.722e-01	1.142e-01	7.637	2.23e-14	***
MVR_PTS	1.140e-01	1.706e-02	6.687	2.28e-11	***
CAR_AGE	-8.623e-03	9.414e-03	-0.916	0.359687	
URBANICITYHighly Urban/ Urban	2.312e+00	1.413e-01	16.359	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5964.0 on 5155 degrees of freedom
 Residual deviance: 4636.3 on 5118 degrees of freedom
 (1372 observations deleted due to missingness)
 AIC: 4712.3

Number of Fisher Scoring iterations: 5