Auto Insurance Claims

Daniel Craig, John Cruz, Shaya Engelman, Noori Selina, Gavriel Steinmetz-Silber

2024-04-02

Required Libraries

```
library(janitor)
library(kableExtra)
library(latex2exp)
library(psych)
library(scales)
library(stringr)
library(ggcorrplot)
library(tidyverse)
library(mice)
library(ggmice)
library(caret)
library(bestNormalize)
library(e1071)
library(car)
library(glmnet)
library(pROC)
library(Metrics)
```

Introduction

In this project, we work with a dataset containing approximately 8000 records representing customers at an auto insurance company. Each record has two response variables. The first response variable, `TARGET_FLAG`, is a `1` or a `0` (zero). A `1` means that the person was in a car crash. A `0` means that the person was not in a car crash. The second response variable is `TARGET_AMT`. This value is zero if the person did not crash their car, however, if they did crash their car, this number will be a value greater than zero.

We begin by exploring the data, with an emphasis on the shape of the dataset as well as its variables. This exploration will include investigating missing values, skewness, and correlations between variables. We then move to data preparation, where we will transform the data to address concerns found in the exploration phase. We will then build models of two kinds. First, we will build binary logistic regression models to predict whether or not a person will crash their car. Second, we build multiple linear regression models to predict the amount of money it will cost if the person crashes their car. Finally, we evaluate the models, select the best ones, and make predictions using an evaluation dataset.

VARIABLE	DEFINITION	THEORETICAL EFFECT
`INDEX`	Identification Variable (do not use)	None

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

Data Exploration

Import Data

When we import the training and evaluation dataset, we have 26 columns representing each variable we have defined above. We also have 8,161 total rows for the training set and 2,141 rows for the evaluation set. As we glance through the values in each column, we can see there is some data wrangling that will needs to be performed prior to evaluating any summary statistics.

Table 2: Training Set

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	
1	0	0	0	60	0	11	$67,349	No	
2	0	0	0	43	0	11	$91,449	No	
4	0	0	0	35	1	10	$16,039	No	
5	0	0	0	51	0	14		No	
6	0	0	0	50	0	NA	$114,986	No	
7	1	2946	0	34	1	12	$125,301	Yes	

Dimensions:

8161 x 26

Table 3: Evaluation Set

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	
3	NA	NA	0	48	0	11	$52,881	No	
9	NA	NA	1	40	1	11	$50,815	Yes	
10	NA	NA	0	44	2	12	$43,486	Yes	
18	NA	NA	0	35	2	NA	$21,204	Yes	
21	NA	NA	0	59	0	12	$87,460	No	
30	NA	NA	0	46	0	14		No	

Dimensions:

2141 x 26

Data Wrangling

In this subsection, we make the first alterations to the training dataset. **Any changes applied to the training set will be similarly applied to the evaluation set, unless otherwise noted.**

First, we can drop the `INDEX` column as it provides no value to our analysis.

Table 4: Training Set

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_V
0	0	0	60	0	11	$67,349	No	$0
0	0	0	43	0	11	$91,449	No	$257,252
0	0	0	35	1	10	$16,039	No	$124,191
0	0	0	51	0	14		No	$306,251
0	0	0	50	0	NA	$114,986	No	$243,925
1	2946	0	34	1	12	$125,301	Yes	$0

Note:

Dropped 'INDEX' column:

Next, we note that the `INCOME`, `HOME_VAL`, `BLUEBOOK` and `OLDCLAIM` columns are in a currency string format and need to be changed to a numeric value we can work with.

Table 5: Training Set: Before

INCOME	HOME_VAL	BLUEBOOK	OLDCLAIM
$67,349	$0	$14,230	$4,461
$91,449	$257,252	$14,940	$0
$16,039	$124,191	$4,010	$38,690
	$306,251	$15,440	$0
$114,986	$243,925	$18,000	$19,217
$125,301	$0	$17,430	$0

Table 6: Training Set: After

INCOME	HOME_VAL	BLUEBOOK	OLDCLAIM
67349	0	14230	4461
91449	257252	14940	0
16039	124191	4010	38690
NA	306251	15440	0
114986	243925	18000	19217
125301	0	17430	0

Next, we observe that some of the values in the columns`MSTATUS`, `SEX`, `EDUCATION`, `JOB`, `CAR_TYPE`, `URBANICITY` has extra characters `z_` that need to be removed.

Table 7: Training Set: Before

MSTATUS	SEX	EDUCATION	JOB	CAR_TYPE	URBANICITY
z_No	M	PhD	Professional	Minivan	Highly Urban/ Urban
z_No	M	z_High School	z_Blue Collar	Minivan	Highly Urban/ Urban
Yes	z_F	z_High School	Clerical	z_SUV	Highly Urban/ Urban
Yes	M	<High School	z_Blue Collar	Minivan	Highly Urban/ Urban
Yes	z_F	PhD	Doctor	z_SUV	Highly Urban/ Urban
z_No	z_F	Bachelors	z_Blue Collar	Sports Car	Highly Urban/ Urban

Table 8: Training Set: After

MSTATUS	SEX	EDUCATION	JOB	CAR_TYPE	URBANICITY
No	M	PhD	Professional	Minivan	Highly Urban/ Urban
No	M	High School	Blue Collar	Minivan	Highly Urban/ Urban
Yes	F	High School	Clerical	SUV	Highly Urban/ Urban
Yes	M	<High School	Blue Collar	Minivan	Highly Urban/ Urban
Yes	F	PhD	Doctor	SUV	Highly Urban/ Urban
No	F	Bachelors	Blue Collar	Sports Car	Highly Urban/ Urban

Now that we've fixed the values within the dataset, we also want to ensure that the types within the datasets are correct. Specifically, we will change some of our variables' values into factors. These variables and values are:

- `PARENT1`: Yes/No

- `MSTATUS`: Yes/No
- `SEX`: M/F
- `RED_CAR`: Yes/No (Fix capital punctuation of these values)
- `REVOKED`: Yes/No
- `EDUCATION`: High School, Bachelors, Masters, PhD (Ordered Factor as each level has an ordered precedence of completing it.)

Table 9: Training Set: Before

PARENT1	MSTATUS	SEX	RED_CAR	REVOKED	EDUCATION
No	No	M	yes	No	PhD
No	No	M	yes	No	High School
No	Yes	F	no	No	High School
No	Yes	M	yes	No	<High School
No	Yes	F	no	Yes	PhD
Yes	No	F	no	No	Bachelors

Table 10: Training Set: After

PARENT1	MSTATUS	SEX	RED_CAR	REVOKED	EDUCATION
No	No	M	Yes	No	PhD
No	No	M	Yes	No	High School
No	Yes	F	No	No	High School
No	Yes	M	Yes	No	<High School
No	Yes	F	No	Yes	PhD
Yes	No	F	No	No	Bachelors

Summary Statistics

With the dataset in good shape, we are now ready to take a deeper look at the data within.

Table 11: Summary Statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TARGET_FLAG	1	8161	0.26	0.44	0	0.20	0.00	0	1.0	1.0	1.07	-0.85	0.00
TARGET_AMT	2	8161	1504.32	4704.03	0	593.71	0.00	0	107586.1	107586.1	8.71	112.29	52.07
KIDSDRIV	3	8161	0.17	0.51	0	0.03	0.00	0	4.0	4.0	3.35	11.78	0.01
AGE	4	8155	44.79	8.63	45	44.83	8.90	16	81.0	65.0	-0.03	-0.06	0.10
HOMEKIDS	5	8161	0.72	1.12	0	0.50	0.00	0	5.0	5.0	1.34	0.65	0.01
YOJ	6	7707	10.50	4.09	11	11.07	2.97	0	23.0	23.0	-1.20	1.18	0.05
INCOME	7	7716	61898.09	47572.68	54028	56840.98	41792.27	0	367030.0	367030.0	1.19	2.13	541.58
HOME_VAL	9	7697	154867.29	129123.77	161160	144032.07	147867.11	0	885282.0	885282.0	0.49	-0.02	1471.79
TRAVTIME	14	8161	33.49	15.91	33	33.00	16.31	5	142.0	137.0	0.45	0.66	0.18
BLUEBOOK	16	8161	15709.90	8419.73	14440	15036.89	8450.82	1500	69740.0	68240.0	0.79	0.79	93.20
TIF	17	8161	5.35	4.15	4	4.84	4.45	1	25.0	24.0	0.89	0.42	0.05
OLDCLAIM	20	8161	4037.08	8777.14	0	1719.29	0.00	0	57037.0	57037.0	3.12	9.86	97.16
CLM_FREQ	21	8161	0.80	1.16	0	0.59	0.00	0	5.0	5.0	1.21	0.28	0.01
MVR_PTS	23	8161	1.70	2.15	1	1.31	1.48	0	13.0	13.0	1.35	1.38	0.02
CAR_AGE	24	7651	8.33	5.70	8	7.96	7.41	-3	28.0	31.0	0.28	-0.75	0.07

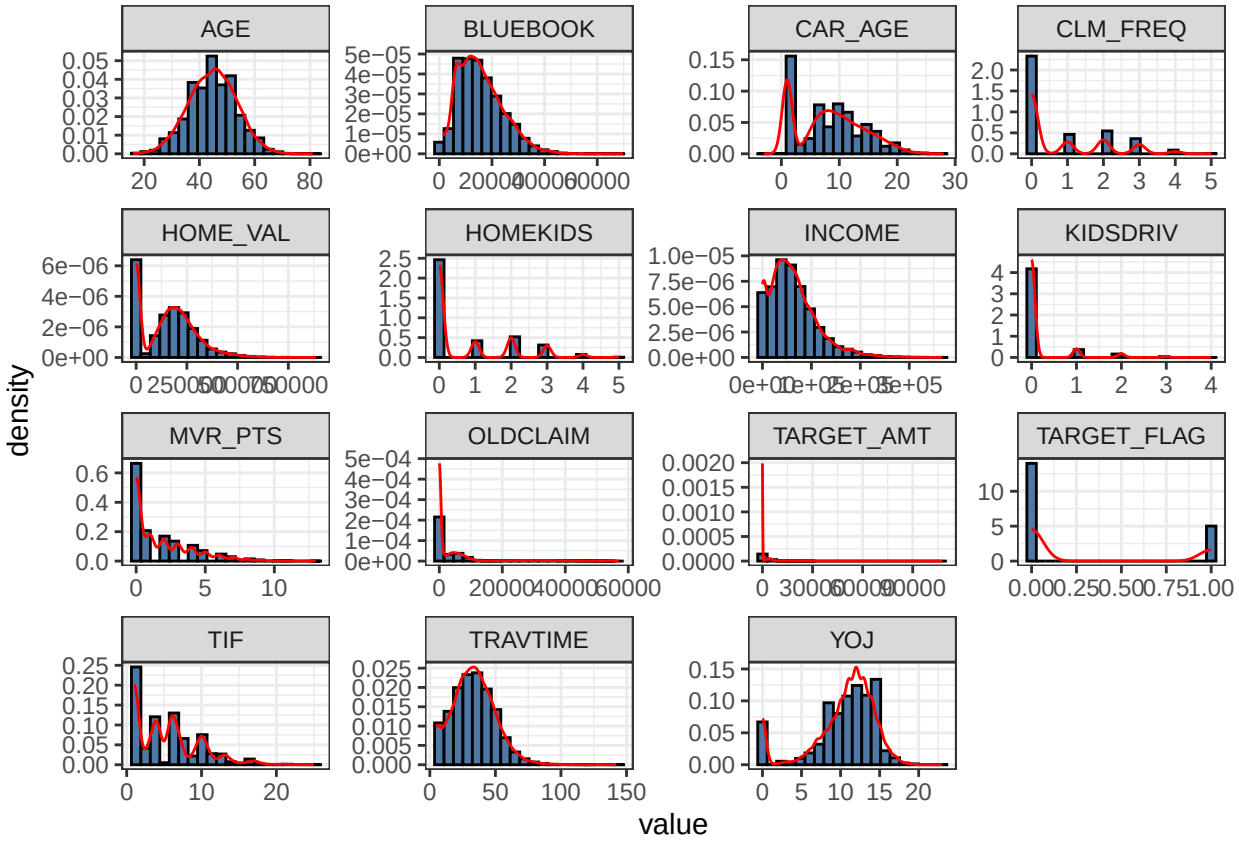
We have an average customer age of 44.79. Their average income is almost $62k while their home value is approximately $155k. For cars in a crash there is an average cost of $1500.

Visualizations

We move now to visualize the data. Some of the data is continuous and some of it is categorical; naturally, we will visualize those different types of data in different manners.

Density

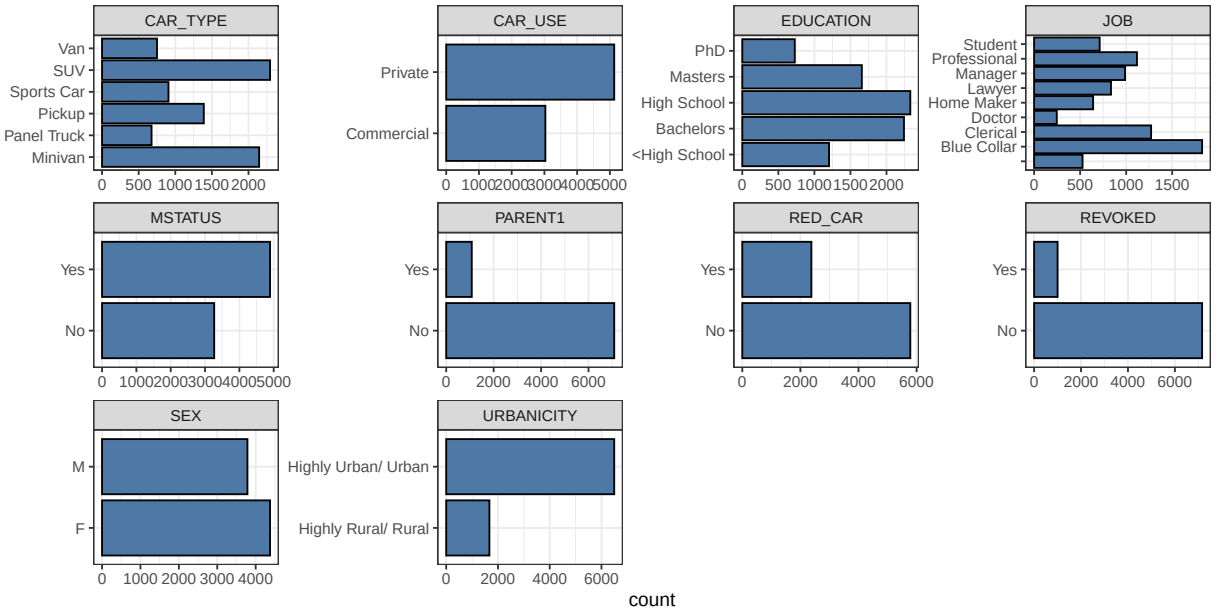
We can get a better idea of the distributions and skewness by plotting our continuous variables:



We have a normal distribution for `AGE`. As for our first response variable `TARGET_FLAG`, it clearly shows the logit function between zero and one. Other plots show significant right skewness for `BLUEBOOK`, `INCOME`, `MVR_PTS`, `OLDCLAIM`, `TARGET_AMT`, `TIF` and `TRAVTIME`. This is quite intuitive; values for these columns cannot be negative, and so they are bounded only on the left side. We also have some bimodal distributions for `CAR_AGE`, `HOME_VAL` and `YOJ`. We would need to perform some transformations on these variables, and possibly consider grouping the bimodal variables.

Bar Plots

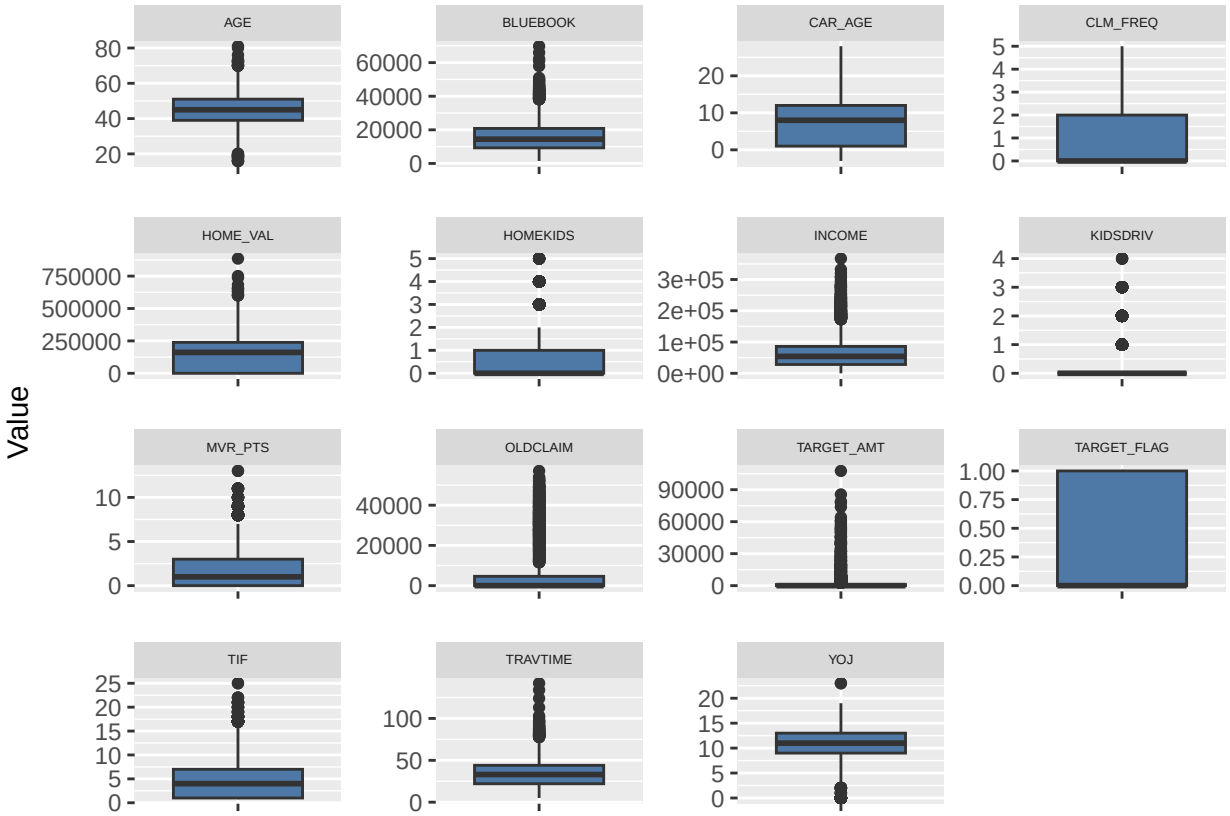
Our bar plots show us how our categorical data is divided up.



Among other observations, we can see that most of the car types we have are either `SUV` or `Minivan`. Additionally, most drivers' highest education is `High School` or `Bachelors`. The drivers predominately live/work in `Highly Urban/Urban` areas. Their car use is mainly in a private (and not commercial) capacity.

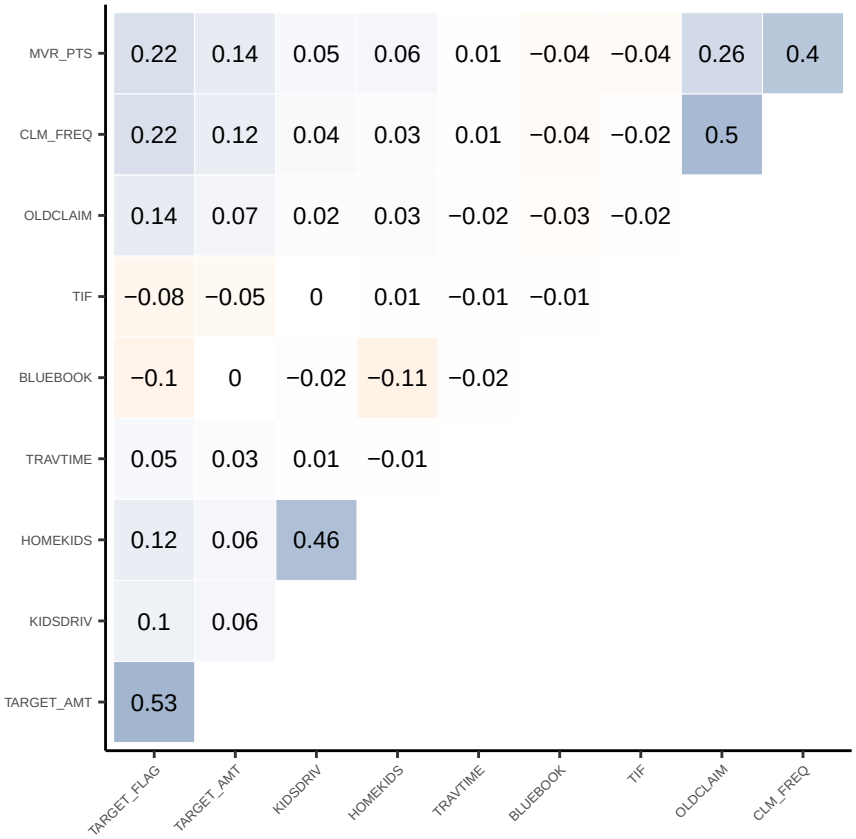
Box Plots

Visualizations can also display the presence of outliers. We expect quite a few outliers, especially when it comes to the value of cars, income of drivers, and home values.



And indeed, our box plots show us there are some outliers to be dealt with. We can see the `BLUEBOOK` value of cars have some quite pricey vehicles being insured. Unsurprisingly, `HOMEKIDS` and `KIDSDRIV` also have outliers (many drivers don't have children let alone those that drive). Looking at those two box plots, it's clear that the middle half of the data for `HOMEKIDS` is higher than the middle half of the data for `KIDSDRIV`. Yet again this is intuitive; only a subset of the drivers' children are driving. Still, we'd expect some correlation between these two columns, which takes us to the next subsection.

Correlation Matrix



As expected, we have some moderately strong correlations between some of our variables. This will have to be addressed with when we build our models.

- `KIDSDRIV` and `HOMEKIDS`: As discussed, we expect multicollinearity as if you have children, they may be of age to drive already
- `MVR_PTS` and `CLM_FREQ`: The multicollinearity is intuitive as, if you have higher motor vehicle points accumulated from negative driving habits, you may be more likely to have accidents and require to file more claims than the average driver.
- `CLM_FREQ` and `OLDCLAIM`: There would be some multicollinearity since those that file more claims are likely to have a higher total claim value over the past 5 years.

- `TARGET_AMT` and `TARGET_FLAG`: Perhaps most obviously of all, since we expect TARGET_AMT to be zero if the person did not crash their car, but greater than zero if they did crash their car.

Missing Values

Table 12: Missing Values Count

AGE	YOJ	INCOME	HOME_VAL	CAR_AGE
6	454	445	464	510

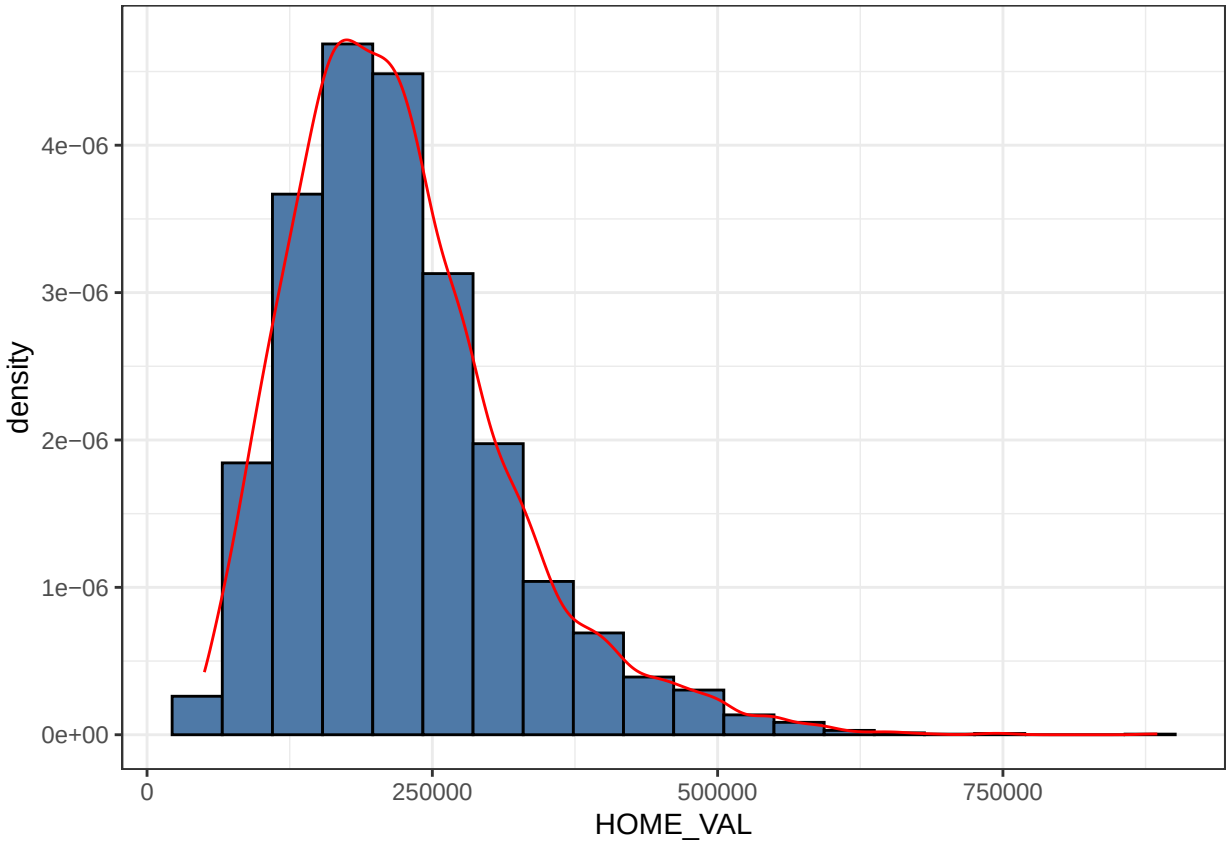
We can see we have some columns missing values.

- `AGE`: This column is only missing a few values and, given that it is a normally distributed variable, we have many options to impute them

- `YOJ`: We are missing a lot of values for how many year people have been at their job
- `INCOME`: We don't have how much money they are making in a year. It could be that they are not working.
- `HOME_VAL`: These missing values may be under the assumption they don't own a home and possibly renting. We return to this point in a moment.
- `CAR_AGE`: The highest amount of values we don't have is how old the car is.

There is a nuanced point about `HOME_VAL`. As aforementioned, these plausibly represent rentals. However, recall the density plots earlier; there were many 0s for `HOME_VAL`. There cannot realistically be that many houses actually valued at \$0. It is possible, then, that the 0s *also* represent rentals. In that case, we should convert the 0s to missing values, and impute them as we will the other missing values for this column.

Plotting the `HOME_VAL` data without the 0s, we get:



And we can see a much more normal distribution than before.

We now check if there are any other suspect 0s:

Table 13: Zero Counts in Training Dataset

	Zero.Count
TARGET_FLAG	6008
TARGET_AMT	6008
KIDSDRIV	7180
AGE	0
HOMEKIDS	5289
YOJ	625
INCOME	615
PARENT1	0
HOME_VAL	0
MSTATUS	0
SEX	0
EDUCATION	0
JOB	0
TRAVTIME	0
CAR_USE	0
BLUEBOOK	0
TIF	0
CAR_TYPE	0
RED_CAR	0
OLDCLAIM	5009
CLM_FREQ	5009
REVOKED	0
MVR_PTS	3712
CAR_AGE	3
URBANICITY	0

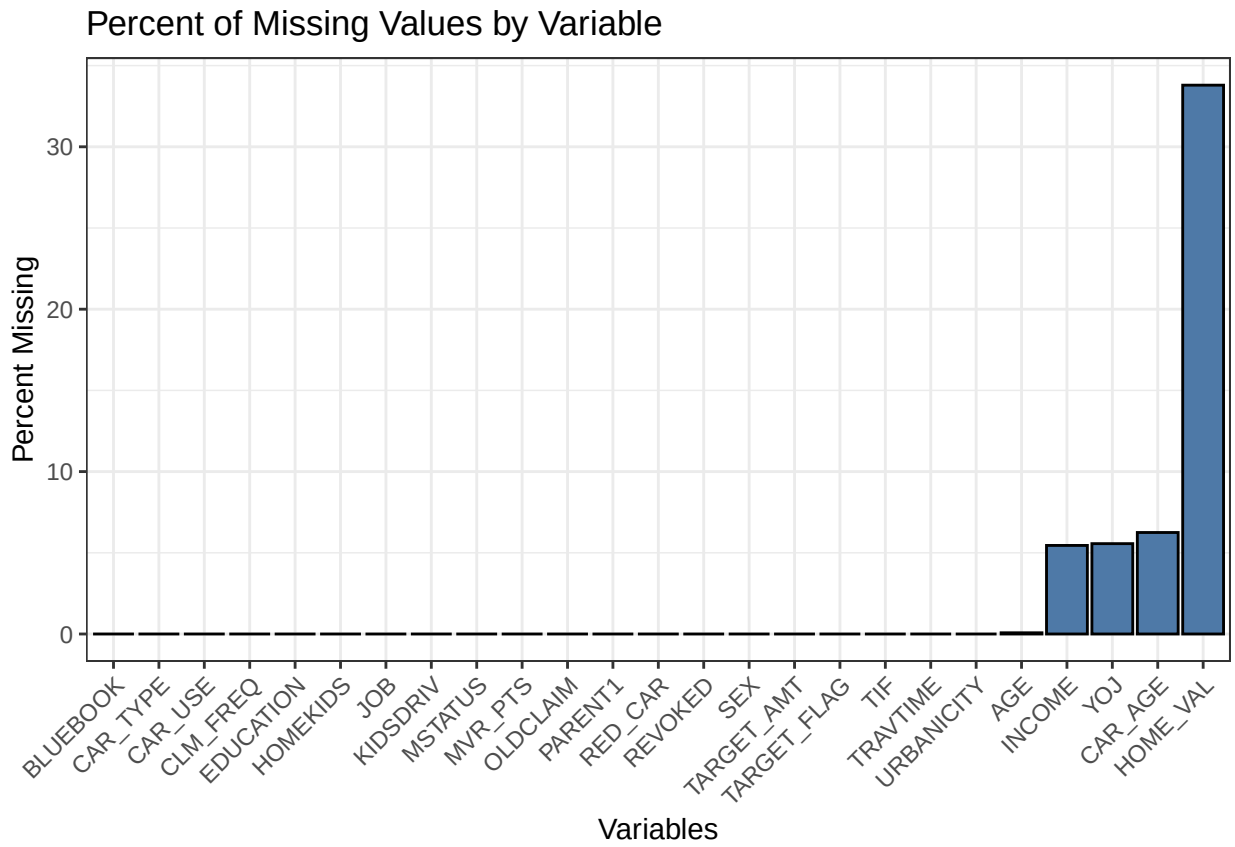
There are no other columns containing suspect 0 values. This concludes the largely exploratory phase of our analysis; we move now to consider broader transformations of the data.

Data Preparation

While some basic data preparation steps have been taken in the previous section, we will now work to more comprehensively transform the data to ensure the multiple linear regression and logistic regression models perform as best as possible.

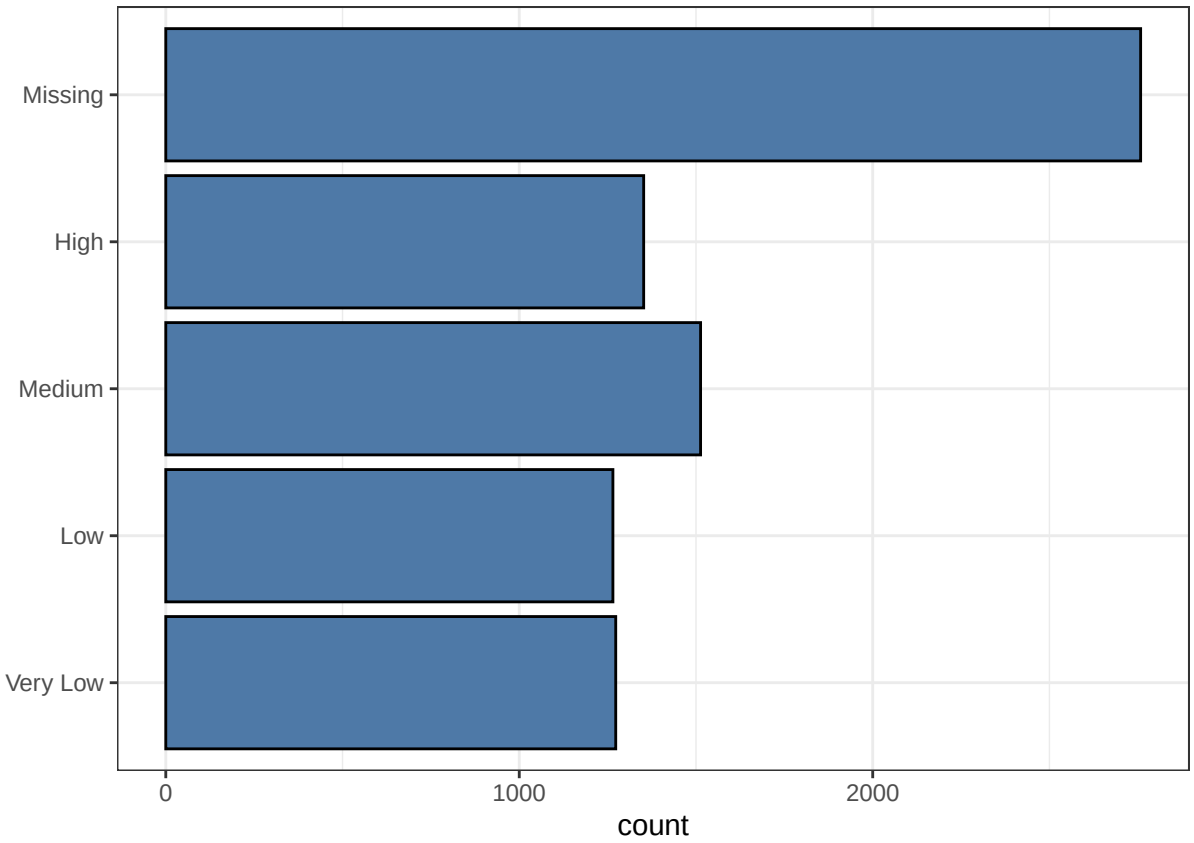
Missing Values

We begin where we left off in the last section: missing values. As noted, the only variables missing data are `CAR_AGE`, `HOME_VAL`, `YOJ`, `INCOME`, and `AGE`. Recall, that we recently *increased* the number of missing values in `HOME_VAL` by converting 0s to NAs. We can see the severity of missing values for all columns here:



Aside from `HOME_VAL`, there really are not too many missing values–always under 6%. Shortly, we will impute values for those other columns. But it's worth thinking deeper about `HOME_VAL` yet again. Recall, we are assuming that the missing values are associated with rentals (and converted 0s to NAs accordingly). But then, we surely don't want to impute missing values for `HOMEVAL`; the very fact that they are missing is meaningful. Instead, then, let's convert the variable to a categorical one, and include the missing values as a category.

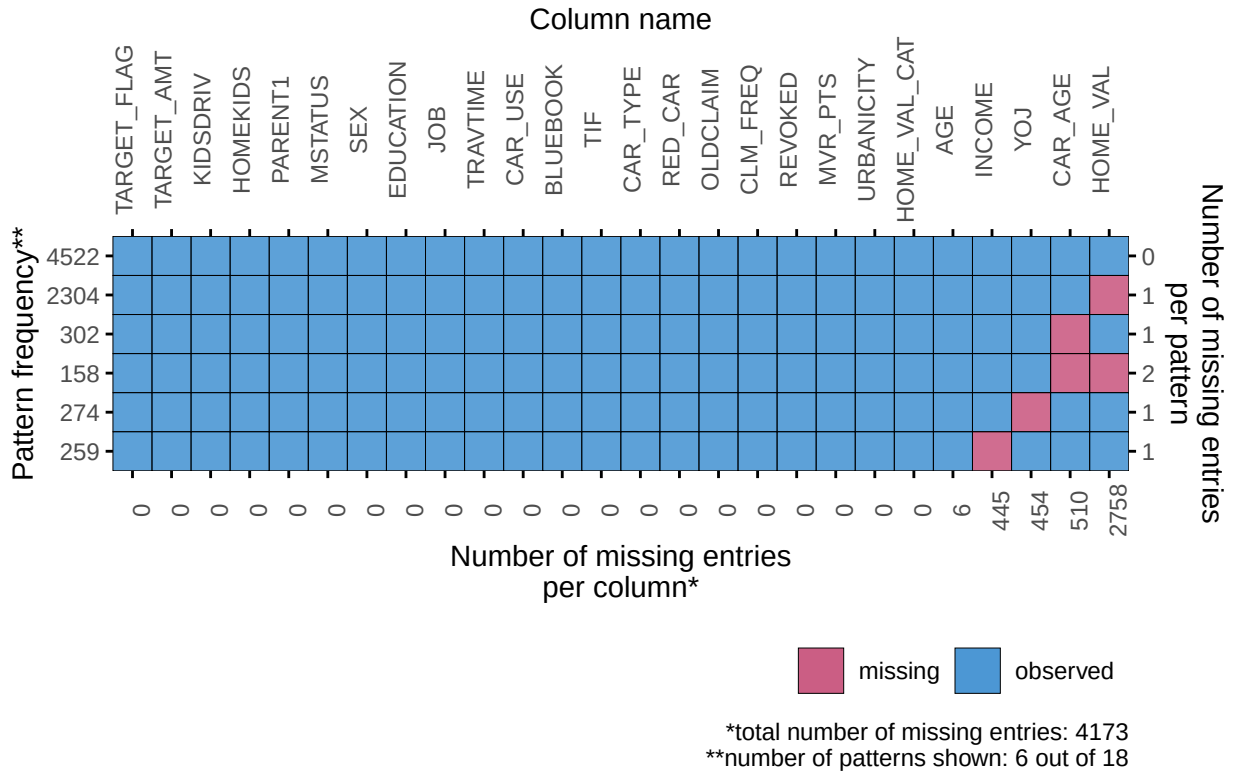
After converting, we can do another bar plot:



And we see that the data is divided fairly evenly, although "Missing" is the largest category.

Let's investigate patterns potentially underlying the missing values:

```
plot_pattern(train, square = TRUE, rotate = TRUE, npat = 6)
```



First, we note that the fact that a missing pattern primarily impacts HOME_VAL supports the hypothesis we had about renters. This suggests that we can feel comfortable with the categorical version of the variable, and do away with the continuous `HOME_VAL` variable.

Second, and more broadly, the patterns encourage us to use MICE to impute values. Missing values can co-occur, suggestive of potential relationships between the variables. Still, other columns are missing values in an isolated manner. A one-size-fits-all method of imputation is thus inappropriate, and we prefer MICE as it will impute each variable's missing data in a way that addresses its missingness.

Imputations

Before we can impute missing values, we perform the train-test split to avoid data leakage:

We then use MICE to impute. Critically, we ignore the test values when imputing for both sets, to avoid data leakage.

```
## Warning: Number of logged events: 4
```

```
## Warning: Number of logged events: 4
```

Table 14: Summary Statistics Comparison Across Datasets

Variable_Stat	Dataset (Pre-Imputations)	Train Imputed	Test Imputed
CAR_AGE_min	-3.000000	-3.000000	0.000000
CAR_AGE_q1	1.000000	1.000000	1.000000
CAR_AGE_median	8.000000	8.000000	8.000000
CAR_AGE_mean	8.347639	8.365517	8.295752
CAR_AGE_q3	12.000000	12.000000	12.000000
CAR_AGE_max	27.000000	27.000000	28.000000
YOJ_min	0.000000	0.000000	0.000000
YOJ_q1	9.000000	9.000000	9.000000
YOJ_median	11.000000	11.000000	11.000000
YOJ_mean	10.499536	10.505654	10.515441
YOJ_q3	13.000000	13.000000	13.000000
YOJ_max	23.000000	23.000000	19.000000
INCOME_min	0.000000	0.000000	0.000000
INCOME_q1	28127.500000	27907.000000	27786.750000
INCOME_median	54007.000000	53598.000000	53841.000000
INCOME_mean	62215.300443	62089.682514	60849.925980
INCOME_q3	85865.500000	85752.000000	85472.000000
INCOME_max	332339.000000	332339.000000	367030.000000
AGE_min	16.000000	16.000000	17.000000
AGE_q1	39.000000	39.000000	39.000000
AGE_median	45.000000	45.000000	45.000000
AGE_mean	44.781573	44.781612	44.810733
AGE_q3	51.000000	51.000000	51.000000
AGE_max	81.000000	81.000000	76.000000

These summary statistics are highly encouraging. With a few exceptions, the values for each statistic are consistent across the three datasets. This consistency implies that the distribution of the datasets with imputed values mirror the original distribution. This is true about the means and medians as well. Edge cases likewise appear to be handled effectively.

Outliers and Transformations

In the data exploration section, we took a cursory look at outliers. All outliers appeared to be reasonable values, but outliers introduce heavier skew into distributions which negatively impact logistic and linear models. To handle outliers, several types of transformations were performed. Skewness with an absolute value greater than 1 is considered heavily skewed, if within .5 to 1 or -.5 to -1 it is moderately skewed, and between 0 to .5 or 0 to -.5 is considered lightly skewed.

We can assess the most appropriate transformations for each variable using the bestNormalize function.

Table 15: Best Transformations

Variable	Transformation
BLUEBOOK	orderNorm
INCOME	orderNorm
MVR_PTS	sqrt_x
OLDCLAIM	center_scale
TIF	yeojohnson
TRAVTIME	boxcox
YOJ	sqrt_x
CLM_FREQ	sqrt_x
CAR_AGE	yeojohnson

Again, we must be quite careful to avoid data leakage, calculating the parameters for these transformations using only the training set, and then applying the transformations to our other sets using the same parameters.

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(...)`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

Table 16: Pre and Post Transformation Skewness Comparison

Variable	Pre-Transformation Skew	Post-Transformation Skew
BLUEBOOK	0.791	0.042
INCOME	1.195	0.146
MVR_PTS	1.337	0.393
OLDCLAIM	3.190	3.190
TIF	0.890	-0.034
TRAVTIME	0.471	-0.043
YOJ	-1.208	-2.253
CLM_FREQ	1.217	0.711
CAR_AGE	0.268	-0.188

For the most part, we see that the transformations (for the most part) reduced skewness to much more acceptable levels. Unfortunately the worst offender, `OLDCLAIM` was not improved. But overall, the transformed dataset is much more normal. This will likely improve the performance of our models down the line.

Outliers

Even with the transformations, we have some outliers. We will handle them with outlier replacement. Again, it is critical that we leverage parameters from the training set to guide outlier replacement for all sets. This allows us to avoid data leakage.

Encoding, Center/Scale/NearZeroVariance

The final part of the data preparation involves encoding categorical data with one-hot encoding (OHC). Also, we center and scale (CS) all continuous data to avoid extreme values distorting the scale (again using parameters estimated from only the training set) and check continuous data for near-zero variance (NZV). We treat ordinal data as continuous since the distances between values are consistent and meaningful. We can accomplish CS and NZV in one step in a preprocessing pipeline.

The dataframes are now fully processed. We are ready to move on to the modeling phase.

Modeling

With the preprocessing behind us, we are ready to model our data. In the first part of the modeling, we will build logistic regression models, predicting whether the person got into a car crash. Then, in the second part, we will build multiple linear regression models predicting the payout if a person *did* crash their car.

Logistic regression

Again, in the first step, we aim to predict whether or not a driver got into a crash. Given that we have both our original variables and their transformations, we need to be careful not to include the same variables twice. It is worth explicitly categorizing them:

Table 17: Variables Summary

Category	Variables
Encoded	SEX.F, SEX.M, EDUCATION.L, EDUCATION.Q, EDUCATION.C, JOBBlue Collar, JOBClerical, JOBD
Original	KIDSDRIV, AGE, HOMEKIDS, YOJ, INCOME, PARENT1, MSTATUS, SEX, EDUCATION, JOB, TRA
Transformed	BLUEBOOK_transformed, INCOME_transformed, MVR_PTS_transformed, OLDCLAIM_transformed,

We also have to be cautious since we used One-Hot Encoding (OHC) to transform the categorical variables. Specifically, we need to worry about multicollinearity between those columns. For the variables with only two features, we can simply drop one of them as the second adds no extra knowledge. For other collinear variables, we need to be more careful and employ different strategies, as we will soon discuss.

Model 1 Let's start by building a rather simple model just to get a basic idea of what we are dealing with. The idea here is to use the transformed variables, as they're easy to work with (all continuous), and we anticipate they'll perform better than their non-transformed counterparts.

```
simple_model1 <- glm(TARGET_FLAG ~ ., data = train_final[, c(transformed_columns, encoded_columns, "TARG
summary(simple_model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(transformed_columns, encoded_columns, "TARGET_FLAG")])
##
## Coefficients: (9 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.05574    0.13818   0.403  0.68666
## BLUEBOOK_transformed          -0.21664    0.02212  -9.793  < 2e-16 ***
```

```
## INCOME_transformed              -0.22395    0.03350  -6.684 2.32e-11 ***
## MVR_PTS_transformed              0.24224    0.01780  13.608  < 2e-16 ***
## OLDCLAIM_transformed            -0.11981    0.01905  -6.288 3.21e-10 ***
## TIF_transformed                 -0.23220    0.01574 -14.755  < 2e-16 ***
## TRAVTIME_transformed             0.26369    0.01626  16.219  < 2e-16 ***
## YOJ_transformed                 -0.08354    0.01940  -4.306 1.67e-05 ***
## CLM_FREQ_transformed             0.36401    0.02708  13.443  < 2e-16 ***
## CAR_AGE_transformed             -0.06006    0.02170  -2.767  0.00565 **
## SEX.F                            0.03524    0.05865   0.601  0.54796
## SEX.M                                 NA         NA      NA       NA
## EDUCATION.L                     -0.20571    0.08609  -2.390  0.01687 *
## EDUCATION.Q                      0.27093    0.04915   5.512 3.55e-08 ***
## EDUCATION.C                      0.14289    0.04169   3.427  0.00061 ***
## `JOBBlue Collar`                 0.48525    0.09379   5.174 2.29e-07 ***
## JOBClerical                      0.55794    0.10098   5.525 3.29e-08 ***
## JOBDoctor                       -0.18215    0.13389  -1.361  0.17367
## `JOBHome Maker`                  0.27626    0.11381   2.427  0.01521 *
## JOBLawyer                        0.40011    0.09066   4.413 1.02e-05 ***
## JOBManager                      -0.35423    0.08924  -3.970 7.20e-05 ***
## JOBProfessional                  0.29441    0.09082   3.242  0.00119 **
## JOBStudent                       0.06058    0.11568   0.524  0.60051
## CAR_USECommercial                0.80164    0.04921  16.289  < 2e-16 ***
## CAR_USEPrivate                        NA         NA      NA       NA
## CAR_TYPEMinivan                 -0.66597    0.06764  -9.846  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.09080    0.07718  -1.176  0.23943
## CAR_TYPEPickup                  -0.14177    0.06721  -2.109  0.03492 *
## `CAR_TYPESports Car`             0.23356    0.09199   2.539  0.01111 *
## CAR_TYPESUV                      0.05328    0.08380   0.636  0.52493
## CAR_TYPEVan                           NA         NA      NA       NA
## `URBANICITYHighly Rural/ Rural` -2.48282    0.06146 -40.399  < 2e-16 ***
## `URBANICITYHighly Urban/ Urban`       NA         NA      NA       NA
## `HOME_VAL_CAT.Very Low`         -0.39760    0.05783  -6.875 6.18e-12 ***
## HOME_VAL_CAT.Low                -0.37895    0.05656  -6.700 2.08e-11 ***
## HOME_VAL_CAT.Medium             -0.35455    0.05319  -6.665 2.64e-11 ***
## HOME_VAL_CAT.High               -0.38809    0.06561  -5.915 3.32e-09 ***
## HOME_VAL_CAT.Missing                  NA         NA      NA       NA
## RED_CAR.No                      -0.01745    0.04639  -0.376  0.70686
## RED_CAR.Yes                           NA         NA      NA       NA
## REVOKED.No                      -0.88408    0.04934 -17.918  < 2e-16 ***
## REVOKED.Yes                           NA         NA      NA       NA
## PARENT1.No                      -0.71558    0.04898 -14.610  < 2e-16 ***
## PARENT1.Yes                           NA         NA      NA       NA
## MSTATUS.No                       0.35077    0.04281   8.193 2.54e-16 ***
## MSTATUS.Yes                           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33055  on 28564  degrees of freedom
## Residual deviance: 25417  on 28528  degrees of freedom
## AIC: 25491
##
## Number of Fisher Scoring iterations: 5
```

Based on these results, we will remove variables causing multicollinearity problems. For the columns with only two options, we simply remove any one of them. For the other variables, we will remove the variable with the least correlation to the target variable as we assume they would have the least impact on the final model.

Table 18: Correlations with TARGET FLAG

	x
SEX.F	0.0318414
SEX.M	-0.0318414
EDUCATION.L	-0.1316931
EDUCATION.Q	0.0172592
EDUCATION.C	0.0841712
JOBBlue Collar	0.1043799
JOBClerical	0.0261655
JOBDoctor	-0.0530153
JOBHome Maker	0.0123769
JOBLawyer	-0.0629416
JOBManager	-0.1006736
JOBProfessional	-0.0337543
JOBStudent	0.0734056
CAR_USECommercial	0.1427534
CAR_USEPrivate	-0.1427534
CAR_TYPEMinivan	-0.1339852
CAR_TYPEPanel Truck	-0.0020186
CAR_TYPEPickup	0.0481774
CAR_TYPESports Car	0.0574468
CAR_TYPESUV	0.0515462
CAR_TYPEVan	0.0018546
URBANICITYHighly Rural/ Rural	-0.2342684
URBANICITYHighly Urban/ Urban	0.2342684
HOME_VAL_CAT.Very Low	0.0047065
HOME_VAL_CAT.Low	-0.0167182
HOME_VAL_CAT.Medium	-0.0648757
HOME_VAL_CAT.High	-0.1241068
HOME_VAL_CAT.Missing	0.1585339
RED_CAR.No	0.0093489
RED_CAR.Yes	-0.0093489
REVOKED.No	-0.1433965
REVOKED.Yes	0.1433965
PARENT1.No	-0.1697927
PARENT1.Yes	0.1697927
MSTATUS.No	0.1473048
MSTATUS.Yes	-0.1473048
TARGET_FLAG	1.0000000

Again, with the binary columns, we can just remove any one them. Thus, we'll remove:

- SEX.M
- CAR_USEPrivate

- `URBANICITYHighly Urban/ Urban`

- RED_CAR.Yes
- REVOKED.Yes
- PARENT1.Yes
- MSTATUS.Yes

Of the categorical, but not binary, variables we will remove the least informative:

- EDUCATION.Q
- JOBHome Maker
- CAR_TYPEVan
- HOME_VAL_CAT.Low

We now rerun the simple model without these columns to get a look at some of the coefficients.

```
simple_model2 <- glm(TARGET_FLAG ~ ., data = train_final[, c(transformed_columns, encoded_columns_filte
summary(simple_model2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(transformed_columns, encoded_columns_filtered, "TARGET_FLAG")])
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.14454    0.12461  -1.160  0.24609
## BLUEBOOK_transformed             -0.21551    0.02211  -9.745  < 2e-16 ***
## INCOME_transformed               -0.22659    0.03194  -7.094 1.30e-12 ***
## MVR_PTS_transformed               0.24349    0.01779  13.685  < 2e-16 ***
## OLDCLAIM_transformed             -0.12331    0.01903  -6.480 9.16e-11 ***
## TIF_transformed                  -0.23204    0.01571 -14.766  < 2e-16 ***
## TRAVTIME_transformed              0.26208    0.01625  16.129  < 2e-16 ***
## YOJ_transformed                  -0.09174    0.01909  -4.806 1.54e-06 ***
## CLM_FREQ_transformed              0.36370    0.02707  13.438  < 2e-16 ***
## CAR_AGE_transformed              -0.06561    0.02163  -3.034  0.00242 **
## SEX.F                             0.04996    0.05844   0.855  0.39260
## EDUCATION.L                      -0.36568    0.08224  -4.446 8.73e-06 ***
## EDUCATION.C                       0.10844    0.04026   2.693  0.00707 **
## `JOBBlue Collar`                  0.28450    0.07093   4.011 6.05e-05 ***
## JOBClerical                       0.33886    0.07042   4.812 1.50e-06 ***
## JOBDoctor                        -0.07637    0.12329  -0.619  0.53565
## JOBLawyer                         0.23378    0.07558   3.093  0.00198 **
## JOBManager                       -0.55033    0.07095  -7.757 8.72e-15 ***
## JOBProfessional                   0.04378    0.06672   0.656  0.51174
## JOBStudent                       -0.16478    0.08038  -2.050  0.04036 *
## CAR_USECommercial                 0.74458    0.04784  15.563  < 2e-16 ***
## CAR_TYPEMinivan                  -0.68327    0.06746 -10.128  < 2e-16 ***
## `CAR_TYPEPanel Truck`            -0.07561    0.07698  -0.982  0.32603
## CAR_TYPEPickup                   -0.14146    0.06716  -2.106  0.03519 *
## `CAR_TYPESports Car`              0.20999    0.09178   2.288  0.02215 *
## CAR_TYPESUV                       0.02517    0.08356   0.301  0.76326
## `URBANICITYHighly Rural/ Rural` -2.47380    0.06135 -40.321  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`           0.02691    0.05947   0.453  0.65084
## HOME_VAL_CAT.Medium               0.01552    0.05772   0.269  0.78799
```

```
## HOME_VAL_CAT.High                0.01103    0.07376   0.150  0.88109
## HOME_VAL_CAT.Missing             0.39769    0.05633   7.060 1.67e-12 ***
## RED_CAR.No                      -0.01573    0.04631  -0.340  0.73411
## REVOKED.No                      -0.88511    0.04926 -17.967  < 2e-16 ***
## PARENT1.No                      -0.71243    0.04892 -14.563  < 2e-16 ***
## MSTATUS.No                       0.34353    0.04276   8.035 9.38e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33055  on 28564  degrees of freedom
## Residual deviance: 25449  on 28530  degrees of freedom
## AIC: 25519
##
## Number of Fisher Scoring iterations: 5
```

We see that we no longer have any singularities in the model and, notably, the Deviance and AIC of these two models are identical since we haven't actually improved the model. We just improved the accuracy of the individual coefficients by removing perfect multicollinearity. However, we still have insignificant variables that likely are only making the model worse. Additionally, while we no longer have perfect multicollinearity, there is still definitely some collinearity in the categorical variables. We will use a combination of the vif() function and the correlation matrix from above to try to improve on this.

Table 19: VIF Values simple model 2

	x
BLUEBOOK_transformed	1.996188
INCOME_transformed	3.785523
MVR_PTS_transformed	1.194844
OLDCLAIM_transformed	1.772831
TIF_transformed	1.011940
TRAVTIME_transformed	1.037135
YOJ_transformed	1.646669
CLM_FREQ_transformed	1.668053
CAR_AGE_transformed	1.931736
SEX.F	3.509678
EDUCATION.L	3.768989
EDUCATION.C	1.457579
‘JOBBlue Collar‘	3.971995
JOBClerical	2.765126
JOBDoctor	1.379093
JOBLawyer	1.993747
JOBManager	1.699790
JOBProfessional	2.082224
JOBStudent	2.422258
CAR_USECommercial	2.319272
CAR_TYPEMinivan	3.058352
‘CAR_TYPEPanel Truck‘	1.974441
CAR_TYPEPickup	2.854481
‘CAR_TYPESports Car‘	3.725190
CAR_TYPESUV	6.081278
‘URBANICITYHighly Rural/ Rural‘	1.133181
‘HOME_VAL_CAT.Very Low‘	1.959488
HOME_VAL_CAT.Medium	1.984954
HOME_VAL_CAT.High	2.342268
HOME_VAL_CAT.Missing	3.149562
RED_CAR.No	1.832713
REVOKED.No	1.328290
PARENT1.No	1.367305
MSTATUS.No	1.878037

Given this information, we opt to remove CAR_TYPESUV which has a VIF value of over 6. We will then fit a new model.

```
encoded_columns_filtered2 <- setdiff(encoded_columns_filtered, "CAR_TYPESUV")

transformed_model <- glm(TARGET_FLAG ~ ., data = train_final[, c(transformed_columns, encoded_columns_fi

summary(transformed_model)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(transformed_columns, encoded_columns_filtered2, "TARGET_FLAG")])
```

```
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.13486    0.12038  -1.120 0.262606
## BLUEBOOK_transformed            -0.21846    0.01983 -11.016  < 2e-16 ***
## INCOME_transformed              -0.22707    0.03190  -7.118 1.09e-12 ***
## MVR_PTS_transformed              0.24338    0.01779  13.682  < 2e-16 ***
## OLDCLAIM_transformed            -0.12339    0.01903  -6.485 8.85e-11 ***
## TIF_transformed                 -0.23197    0.01571 -14.763  < 2e-16 ***
## TRAVTIME_transformed             0.26220    0.01624  16.141  < 2e-16 ***
## YOJ_transformed                 -0.09158    0.01908  -4.800 1.59e-06 ***
## CLM_FREQ_transformed             0.36389    0.02706  13.449  < 2e-16 ***
## CAR_AGE_transformed             -0.06551    0.02163  -3.029 0.002452 **
## SEX.F                            0.05975    0.04855   1.231 0.218432
## EDUCATION.L                     -0.36447    0.08214  -4.437 9.11e-06 ***
## EDUCATION.C                      0.10905    0.04021   2.712 0.006683 **
## `JOBBlue Collar`                 0.28776    0.07010   4.105 4.04e-05 ***
## JOBClerical                      0.34011    0.07030   4.838 1.31e-06 ***
## JOBDoctor                       -0.07424    0.12309  -0.603 0.546425
## JOBLawyer                        0.23564    0.07533   3.128 0.001758 **
## JOBManager                      -0.54900    0.07081  -7.753 8.99e-15 ***
## JOBProfessional                  0.04506    0.06658   0.677 0.498533
## JOBStudent                      -0.16294    0.08014  -2.033 0.042038 *
## CAR_USECommercial                0.74176    0.04691  15.812  < 2e-16 ***
## CAR_TYPEMinivan                 -0.69796    0.04657 -14.988  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.08034    0.07533  -1.066 0.286207
## CAR_TYPEPickup                  -0.15544    0.04853  -3.203 0.001359 **
## `CAR_TYPESports Car`             0.18730    0.05244   3.572 0.000355 ***
## `URBANICITYHighly Rural/ Rural` -2.47381    0.06136 -40.320  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`          0.02674    0.05946   0.450 0.652928
## HOME_VAL_CAT.Medium              0.01535    0.05772   0.266 0.790247
## HOME_VAL_CAT.High                0.01053    0.07375   0.143 0.886487
## HOME_VAL_CAT.Missing             0.39714    0.05630   7.054 1.74e-12 ***
## RED_CAR.No                      -0.01522    0.04629  -0.329 0.742239
## REVOKED.No                      -0.88532    0.04926 -17.973  < 2e-16 ***
## PARENT1.No                      -0.71215    0.04891 -14.561  < 2e-16 ***
## MSTATUS.No                       0.34373    0.04275   8.041 8.94e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33055  on 28564  degrees of freedom
## Residual deviance: 25449  on 28531  degrees of freedom
## AIC: 25517
##
## Number of Fisher Scoring iterations: 5
```

Now let's remove SEX and RED_CAR, both of which are not significant predictors.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(transformed_columns, encoded_columns_filtered3, "TARGET_FLAG")])
```

```
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.09091    0.11204  -0.811 0.417150
## BLUEBOOK_transformed            -0.21442    0.01955 -10.965  < 2e-16 ***
## INCOME_transformed              -0.22798    0.03189  -7.149 8.73e-13 ***
## MVR_PTS_transformed              0.24358    0.01778  13.696  < 2e-16 ***
## OLDCLAIM_transformed            -0.12355    0.01902  -6.495 8.31e-11 ***
## TIF_transformed                 -0.23195    0.01571 -14.767  < 2e-16 ***
## TRAVTIME_transformed             0.26232    0.01624  16.150  < 2e-16 ***
## YOJ_transformed                 -0.09194    0.01908  -4.820 1.44e-06 ***
## CLM_FREQ_transformed             0.36392    0.02706  13.450  < 2e-16 ***
## CAR_AGE_transformed             -0.06465    0.02162  -2.991 0.002781 **
## EDUCATION.L                     -0.36410    0.08214  -4.433 9.31e-06 ***
## EDUCATION.C                      0.11331    0.04006   2.828 0.004683 **
## `JOBBlue Collar`                 0.28869    0.07008   4.120 3.80e-05 ***
## JOBClerical                      0.33431    0.07015   4.766 1.88e-06 ***
## JOBDoctor                       -0.08245    0.12286  -0.671 0.502173
## JOBLawyer                        0.23140    0.07523   3.076 0.002099 **
## JOBManager                      -0.55354    0.07061  -7.840 4.52e-15 ***
## JOBProfessional                  0.03995    0.06643   0.601 0.547566
## JOBStudent                      -0.16463    0.08006  -2.056 0.039756 *
## CAR_USECommercial                0.73046    0.04600  15.879  < 2e-16 ***
## CAR_TYPEMinivan                 -0.72056    0.04293 -16.785  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.11359    0.07050  -1.611 0.107166
## CAR_TYPEPickup                  -0.17425    0.04605  -3.784 0.000155 ***
## `CAR_TYPESports Car`             0.19852    0.05170   3.840 0.000123 ***
## `URBANICITYHighly Rural/ Rural` -2.47207    0.06133 -40.309  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`          0.02883    0.05942   0.485 0.627578
## HOME_VAL_CAT.Medium              0.01525    0.05768   0.264 0.791497
## HOME_VAL_CAT.High                0.00832    0.07372   0.113 0.910138
## HOME_VAL_CAT.Missing             0.39809    0.05628   7.074 1.51e-12 ***
## REVOKED.No                      -0.88599    0.04925 -17.989  < 2e-16 ***
## PARENT1.No                      -0.71671    0.04877 -14.696  < 2e-16 ***
## MSTATUS.No                       0.34244    0.04272   8.016 1.09e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33055  on 28564  degrees of freedom
## Residual deviance: 25450  on 28533  degrees of freedom
## AIC: 25514
##
## Number of Fisher Scoring iterations: 5
```

At this point, all the predictors in our model are significant and there isn't much collinearity between them. However, since we used the transformed variables to create this model, interpreting this model and applying it to new data can be difficult. The new data would first have to undergo the same transformations. This takes us to our next model.

Model 2 Now that we have found the predictors we would like to include in our model, we can try using the non-transformed versions of the variables and recreating the model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(original_non_cat, encoded_columns_filtered3, "TARGET_FLAG")])
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.205525   0.098698  -2.082 0.037310 *
## KIDSDRIV                         0.234527   0.017074  13.736  < 2e-16 ***
## AGE                             -0.064419   0.018835  -3.420 0.000626 ***
## HOMEKIDS                        -0.001323   0.023109  -0.057 0.954338
## YOJ                              0.036972   0.017055   2.168 0.030168 *
## INCOME                          -0.232389   0.026699  -8.704  < 2e-16 ***
## TRAVTIME                         0.275148   0.015947  17.254  < 2e-16 ***
## BLUEBOOK                        -0.214624   0.019445 -11.037  < 2e-16 ***
## TIF                             -0.212673   0.016063 -13.240  < 2e-16 ***
## OLDCLAIM                         0.195236   0.018516  10.544  < 2e-16 ***
## CLM_FREQ                         0.092819   0.018992   4.887 1.02e-06 ***
## MVR_PTS                          0.113073   0.015692   7.206 5.78e-13 ***
## CAR_AGE                         -0.033059   0.022812  -1.449 0.147282
## EDUCATION.L                     -0.494919   0.081152  -6.099 1.07e-09 ***
## EDUCATION.C                      0.112578   0.040256   2.797 0.005166 **
## `JOBBlue Collar`                 0.142905   0.068588   2.084 0.037204 *
## JOBClerical                      0.191983   0.067864   2.829 0.004670 **
## JOBDoctor                       -0.124792   0.122001  -1.023 0.306366
## JOBLawyer                        0.212103   0.074933   2.831 0.004647 **
## JOBManager                      -0.651884   0.070061  -9.304  < 2e-16 ***
## JOBProfessional                 -0.013433   0.065758  -0.204 0.838132
## JOBStudent                      -0.162008   0.079937  -2.027 0.042694 *
## CAR_USECommercial                0.793163   0.046358  17.109  < 2e-16 ***
## CAR_TYPEMinivan                 -0.708268   0.043001 -16.471  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.183841   0.070609  -2.604 0.009223 **
## CAR_TYPEPickup                  -0.184491   0.046126  -4.000 6.34e-05 ***
## `CAR_TYPESports Car`             0.252429   0.051729   4.880 1.06e-06 ***
## `URBANICITYHighly Rural/ Rural` -2.517245   0.061708 -40.793  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`          0.048042   0.059496   0.807 0.419384
## HOME_VAL_CAT.Medium              0.045126   0.057828   0.780 0.435180
## HOME_VAL_CAT.High               -0.007048   0.071965  -0.098 0.921987
## HOME_VAL_CAT.Missing             0.429493   0.056278   7.632 2.32e-14 ***
## REVOKED.No                      -0.857060   0.045375 -18.888  < 2e-16 ***
## PARENT1.No                      -0.487039   0.058667  -8.302  < 2e-16 ***
## MSTATUS.No                       0.471772   0.045261  10.423  < 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: 33055  on 28564  degrees of freedom
## Residual deviance: 25396  on 28530  degrees of freedom
## AIC: 25466
##
## Number of Fisher Scoring iterations: 5
```

HOMEKIDS does not seem very predictive at all; we remove it:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(non_cat2, encoded_columns_filtered3, "TARGET_FLAG")])
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.207135   0.094608  -2.189 0.028568 *
## KIDSDRIV                         0.234099   0.015348  15.253  < 2e-16 ***
## AGE                             -0.063996   0.017330  -3.693 0.000222 ***
## YOJ                              0.036693   0.016343   2.245 0.024754 *
## INCOME                          -0.232420   0.026694  -8.707  < 2e-16 ***
## TRAVTIME                         0.275162   0.015944  17.258  < 2e-16 ***
## BLUEBOOK                        -0.214623   0.019445 -11.037  < 2e-16 ***
## TIF                             -0.212702   0.016055 -13.248  < 2e-16 ***
## OLDCLAIM                         0.195228   0.018515  10.544  < 2e-16 ***
## CLM_FREQ                         0.092817   0.018992   4.887 1.02e-06 ***
## MVR_PTS                          0.113067   0.015692   7.205 5.79e-13 ***
## CAR_AGE                         -0.033055   0.022812  -1.449 0.147328
## EDUCATION.L                     -0.494804   0.081126  -6.099 1.07e-09 ***
## EDUCATION.C                      0.112538   0.040250   2.796 0.005175 **
## `JOBBlue Collar`                 0.142938   0.068586   2.084 0.037153 *
## JOBClerical                      0.191958   0.067862   2.829 0.004675 **
## JOBDoctor                       -0.124814   0.122001  -1.023 0.306280
## JOBLawyer                        0.212142   0.074930   2.831 0.004638 **
## JOBManager                      -0.651785   0.070040  -9.306  < 2e-16 ***
## JOBProfessional                 -0.013336   0.065736  -0.203 0.839233
## JOBStudent                      -0.162252   0.079823  -2.033 0.042088 *
## CAR_USECommercial                0.793166   0.046358  17.109  < 2e-16 ***
## CAR_TYPEMinivan                 -0.708184   0.042976 -16.479  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.183779   0.070601  -2.603 0.009239 **
## CAR_TYPEPickup                  -0.184368   0.046076  -4.001 6.30e-05 ***
## `CAR_TYPESports Car`             0.252378   0.051721   4.880 1.06e-06 ***
## `URBANICITYHighly Rural/ Rural` -2.517268   0.061706 -40.794  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`          0.047954   0.059476   0.806 0.420089
## HOME_VAL_CAT.Medium              0.045197   0.057815   0.782 0.434359
## HOME_VAL_CAT.High               -0.007028   0.071964  -0.098 0.922203
## HOME_VAL_CAT.Missing             0.429444   0.056272   7.632 2.32e-14 ***
## REVOKED.No                      -0.856993   0.045360 -18.893  < 2e-16 ***
## PARENT1.No                      -0.485556   0.052639  -9.224  < 2e-16 ***
## MSTATUS.No                       0.472360   0.044084  10.715  < 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: 33055  on 28564  degrees of freedom
## Residual deviance: 25396  on 28531  degrees of freedom
## AIC: 25464
##
## Number of Fisher Scoring iterations: 5
```

And we can safely remove CAR_AGE as well:

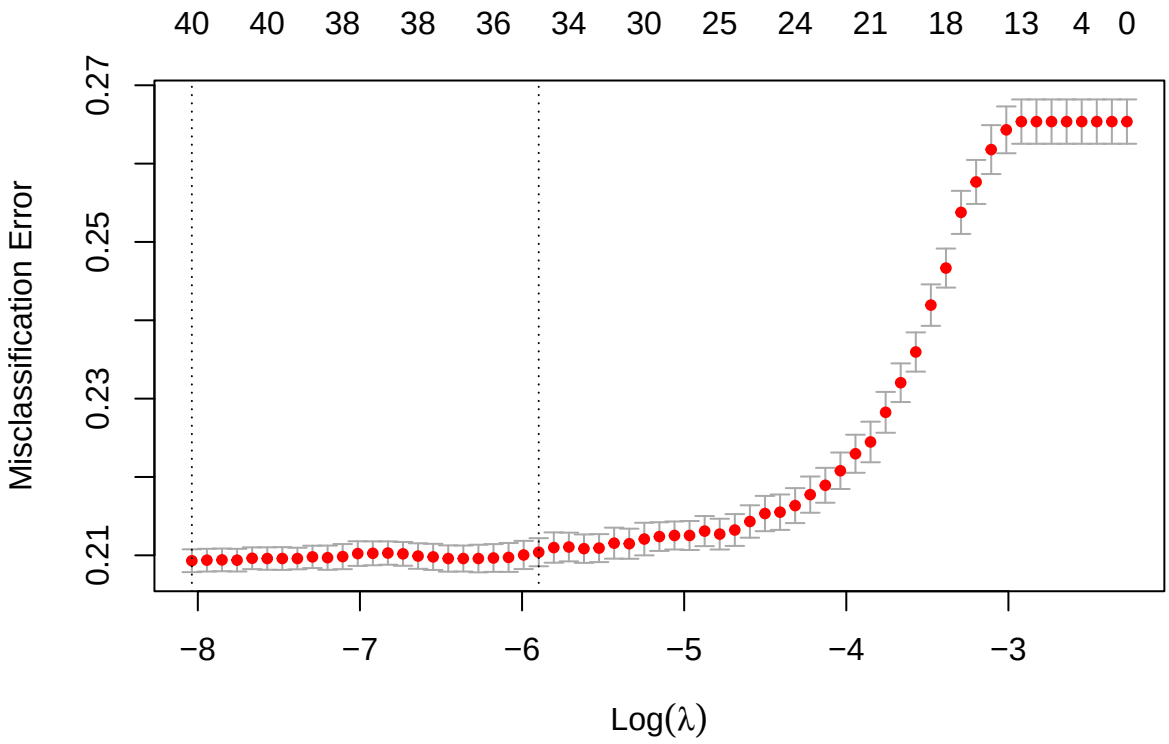
```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = train_final[,
##     c(non_cat3, encoded_columns_filtered3, "TARGET_FLAG")])
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.210877   0.094575  -2.230 0.025765 *
## KIDSDRIV                         0.234221   0.015344  15.265  < 2e-16 ***
## AGE                             -0.064003   0.017330  -3.693 0.000222 ***
## YOJ                              0.036937   0.016342   2.260 0.023806 *
## INCOME                          -0.234588   0.026655  -8.801  < 2e-16 ***
## TRAVTIME                         0.275327   0.015945  17.267  < 2e-16 ***
## BLUEBOOK                        -0.214138   0.019443 -11.014  < 2e-16 ***
## TIF                             -0.213139   0.016049 -13.280  < 2e-16 ***
## OLDCLAIM                         0.195577   0.018518  10.561  < 2e-16 ***
## CLM_FREQ                         0.092674   0.018994   4.879 1.07e-06 ***
## MVR_PTS                          0.112396   0.015686   7.166 7.74e-13 ***
## EDUCATION.L                     -0.549977   0.071694  -7.671 1.70e-14 ***
## EDUCATION.C                      0.127315   0.038934   3.270 0.001075 **
## `JOBBlue Collar`                 0.143051   0.068580   2.086 0.036987 *
## JOBClerical                      0.192950   0.067852   2.844 0.004459 **
## JOBDoctor                       -0.123855   0.122017  -1.015 0.310075
## JOBLawyer                        0.210671   0.074911   2.812 0.004919 **
## JOBManager                      -0.652421   0.070041  -9.315  < 2e-16 ***
## JOBProfessional                 -0.012739   0.065726  -0.194 0.846323
## JOBStudent                      -0.165111   0.079802  -2.069 0.038544 *
## CAR_USECommercial                0.792679   0.046363  17.097  < 2e-16 ***
## CAR_TYPEMinivan                 -0.708507   0.042972 -16.488  < 2e-16 ***
## `CAR_TYPEPanel Truck`           -0.185209   0.070593  -2.624 0.008700 **
## CAR_TYPEPickup                  -0.184729   0.046078  -4.009 6.10e-05 ***
## `CAR_TYPESports Car`             0.251993   0.051723   4.872 1.10e-06 ***
## `URBANICITYHighly Rural/ Rural` -2.517142   0.061697 -40.798  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`          0.049015   0.059467   0.824 0.409805
## HOME_VAL_CAT.Medium              0.048687   0.057766   0.843 0.399323
## HOME_VAL_CAT.High               -0.000682   0.071831  -0.009 0.992424
## HOME_VAL_CAT.Missing             0.431009   0.056263   7.661 1.85e-14 ***
## REVOKED.No                      -0.857398   0.045361 -18.902  < 2e-16 ***
## PARENT1.No                      -0.486373   0.052638  -9.240  < 2e-16 ***
## MSTATUS.No                       0.471786   0.044080  10.703  < 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: 33055  on 28564  degrees of freedom
## Residual deviance: 25398  on 28532  degrees of freedom
## AIC: 25464
##
## Number of Fisher Scoring iterations: 5
```

At first pass, the non-transformed data performs almost as well as the transformed data. We will come back to analyze them later.

Thus far, we have created two models, using largely the same variables. In both cases, we used a lot of

predictors in our models. We also relied on our own intuition of which variables to include and which not to. Now, we will try an automated variable selection method.

Model 3 Again, we would now like to leverage a method that automates variable selection. We will try using Lasso regression to this end–a method that does both feature selection and avoids overfitting the data.




```
## 46 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                  -3.582686e-01
## BLUEBOOK_transformed         -2.136373e-01
## INCOME_transformed           -2.225871e-01
## MVR_PTS_transformed           2.411654e-01
## OLDCLAIM_transformed         -1.137712e-01
## TIF_transformed              -2.296326e-01
## TRAVTIME_transformed          2.600776e-01
## YOJ_transformed              -8.220575e-02
## CLM_FREQ_transformed          3.577001e-01
## CAR_AGE_transformed          -5.891536e-02
## SEX.F                         2.577099e-02
## SEX.M                        -7.631074e-13
## EDUCATION.L                  -2.319275e-01
## EDUCATION.Q                   2.413108e-01
## EDUCATION.C                   1.392531e-01
## JOBBlue Collar                4.185922e-01
## JOBClerical                   4.744570e-01
```

```
## JOBDoctor                      -1.989135e-01
## JOBHome Maker                   1.945416e-01
## JOBLawyer                       3.270194e-01
## JOBManager                     -4.096388e-01
## JOBProfessional                 2.179706e-01
## JOBStudent                      .
## CAR_USECommercial               7.774901e-01
## CAR_USEPrivate                 -3.665167e-12
## CAR_TYPEMinivan                -5.725101e-01
## CAR_TYPEPanel Truck             .
## CAR_TYPEPickup                 -4.678947e-02
## CAR_TYPESports Car              3.149258e-01
## CAR_TYPESUV                     1.373379e-01
## CAR_TYPEVan                     7.984963e-02
## URBANICITYHighly Rural/ Rural -2.459208e+00
## URBANICITYHighly Urban/ Urban  1.977084e-12
## HOME_VAL_CAT.Very Low          -9.471697e-03
## HOME_VAL_CAT.Low                .
## HOME_VAL_CAT.Medium             1.179678e-02
## HOME_VAL_CAT.High              -1.294599e-02
## HOME_VAL_CAT.Missing            3.728283e-01
## RED_CAR.No                     -2.054552e-03
## RED_CAR.Yes                     .
## REVOKED.No                     -8.701874e-01
## REVOKED.Yes                     9.215208e-13
## PARENT1.No                     -7.113841e-01
## PARENT1.Yes                     .
## MSTATUS.No                      3.472215e-01
## MSTATUS.Yes                    -3.861648e-12

## [1] "Best Lambda:  0.000323309244257453"
```

The model is quite interesting, with a few especially interesting coefficients. CAR_USECommercial is strongly correlated with more crashes. This makes sense; perhaps drivers are less careful when it is not their own car at risk. Additionally, these drivers may be driving more often. URBANICITYHighly Rural/ Rural is strongly negative, which again is intuitive. There are fewer drivers in rural areas, so it's less likely that another car will drive into yours, for example. Unsurprisingly, REVOKED.No (not having a revoked license) is associated with a lower likelihood of a crash.

Model Comparison We now run our three models on the test data and compare the results.

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1
```

```
## Warning in roc.default(test_final$TARGET_FLAG, lasso_model_predictions):
## Deprecated use a matrix as predictor. Unexpected results may be produced,
## please pass a numeric vector.

## Setting direction: controls < cases
```

Table 20: Comparison of Logistic Models

Model	RMSE	AIC	Accuracy	Precision	Recall	F1_Score	ROC_AUC
Model_Transformed	0.4653869	25514.47	0.7834150	0.3915228	0.6362245	0.4847425	0.7942114
Model_Untransformed	0.4638042	25464.04	0.7848856	0.4128728	0.6328200	0.4997150	0.7984942
Lasso Model	0.4638923	NA	0.7848039	0.3946625	0.6403464	0.4883450	0.7929631

It's worth noting that we did not include the AIC of the Lasso model because the regularization complicates assumptions about AIC. That aside, something that is really striking is how similar the metrics are for the three models! The RMSE for the untransformed model is just slightly lowest, indicating that this model reduces error best. AIC for the untransformed model is also lowest, which means it does best at minimizing information loss. The statistics we're most interested are accuracy, precision, and recall. The untransformed model has a very small edge for accuracy and precision; the Lasso model has a slight edge for recall (this is noteworthy if catching true positives is especially important). The second model also has the highest F1 score and ROC-AUC score (it best discriminates the classes under different thresholds).

Overall, the differences are so slight. However, the second model performs best for nearly all metrics. It does not perform best for recall, which plausibly is quite important in our context. However, it under-performs the Lasso model by only 0.0072004–this is not a major concern at all. Most importantly, though, the second model is highly interpretable. This fact alone would be a tiebreaker, even if it didn't outperform the other models. As such, we select the second model: "final_non-transformed_model."

Multiple Linear Regression

Thus far, we built models predicting whether or not a car was in a crash, and we selected the best model. We move now to predict the payout if, in fact, a car was in a crash. We begin by manually selecting variables based on correlation and intuition, followed by refining the model using the stepwise regression approach. Subsequently, we'll compare the predictive performance of the manually selected model with the stepwise regression model to determine which one better predicts the TARGET_AMT. We will focus on metrics such as RMSE and R-squared to evaluate predictive accuracy.

Model 1 Now, for variable selection, it stands to reason that if variables were significant in predicting *whether* there was a crash, they would also be significant in predicting the payout after a crash. Let's test that intuition.

```
#have to add back ticks or it just won't work
predictors <- c("KIDSDRIV", "AGE", "YOJ", "INCOME", "TRAVTIME", "BLUEBOOK", "TIF",
                "OLDCLAIM", "CLM_FREQ", "MVR_PTS", "EDUCATION.L", "EDUCATION.C",
                "`JOBBlue Collar`", "JOBClerical", "JOBDoctor", "JOBLawyer",
                "JOBManager", "JOBProfessional", "JOBStudent", "CAR_USECommercial",
                "CAR_TYPEMinivan", "`CAR_TYPEPanel Truck`", "CAR_TYPEPickup",
                "`CAR_TYPESports Car`", "`URBANICITYHighly Rural/ Rural`",
                "`HOME_VAL_CAT.Very Low`", "HOME_VAL_CAT.Medium", "HOME_VAL_CAT.High",
                "HOME_VAL_CAT.Missing", "REVOKED.No", "PARENT1.No", "MSTATUS.No")
```

```r
formula <- as.formula(paste("TARGET_AMT ~", paste(predictors, collapse=" + ")))
mlr1 <- lm(formula, data = train_final)
summary(mlr1)
```

```
##
## Call:
## lm(formula = formula, data = train_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5661  -1775   -784    381 103949
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2810.494    177.476  15.836  < 2e-16 ***
## KIDSDRIV                            217.511     29.002   7.500 6.58e-14 ***
## AGE                                  35.855     31.173   1.150 0.250067
## YOJ                                   4.837     29.072   0.166 0.867849
## INCOME                             -180.845     46.505  -3.889 0.000101 ***
## TRAVTIME                            234.589     28.267   8.299  < 2e-16 ***
## BLUEBOOK                             27.282     33.866   0.806 0.420490
## TIF                                -215.067     27.989  -7.684 1.59e-14 ***
## OLDCLAIM                             43.167     37.785   1.142 0.253280
## CLM_FREQ                            136.293     38.112   3.576 0.000349 ***
## MVR_PTS                             208.682     30.575   6.825 8.96e-12 ***
## EDUCATION.L                        -583.498    126.675  -4.606 4.12e-06 ***
## EDUCATION.C                          40.731     70.617   0.577 0.564087
## `JOBBlue Collar`                    -31.867    125.246  -0.254 0.799164
## JOBClerical                          49.615    122.464   0.405 0.685377
## JOBDoctor                          -376.141    196.140  -1.918 0.055156 .
## JOBLawyer                           110.513    131.855   0.838 0.401957
## JOBManager                         -924.749    119.735  -7.723 1.17e-14 ***
## JOBProfessional                     111.068    117.848   0.942 0.345960
## JOBStudent                         -239.272    147.475  -1.622 0.104715
## CAR_USECommercial                   954.853     85.347  11.188  < 2e-16 ***
## CAR_TYPEMinivan                    -577.629     72.456  -7.972 1.62e-15 ***
## `CAR_TYPEPanel Truck`              -509.313    132.321  -3.849 0.000119 ***
## CAR_TYPEPickup                     -209.171     85.318  -2.452 0.014226 *
## `CAR_TYPESports Car`                114.244     97.386   1.173 0.240765
## `URBANICITYHighly Rural/ Rural` -1881.873     76.195 -24.698  < 2e-16 ***
## `HOME_VAL_CAT.Very Low`            -267.017    107.529  -2.483 0.013026 *
## HOME_VAL_CAT.Medium                -234.421    102.376  -2.290 0.022040 *
## HOME_VAL_CAT.High                  -196.478    123.380  -1.592 0.111294
## HOME_VAL_CAT.Missing                116.660    103.714   1.125 0.260673
## REVOKED.No                         -312.336     89.595  -3.486 0.000491 ***
## PARENT1.No                         -838.618    101.260  -8.282  < 2e-16 ***
## MSTATUS.No                          404.501     79.407   5.094 3.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4717 on 28532 degrees of freedom
## Multiple R-squared:  0.06831,    Adjusted R-squared:  0.06726
## F-statistic: 65.37 on 32 and 28532 DF,  p-value: < 2.2e-16
```

The model as a whole is significant, although we note that the Adjusted R-squared is quite low. One thing that's crucial to note at this point, however, is that we really only want to predict payouts for when the TARGET_FLAG is 1; otherwise the payout is 0. So let's see how well we can do if we just multiply everything by the flag!

```
##
## Call:
## lm(formula = formula, data = train_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9772   -407    -50    182  98182
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             1970.4638   158.0469  12.468
## KIDSDRIV:TARGET_FLAG0                    -28.4489    32.5642  -0.874
## KIDSDRIV:TARGET_FLAG1                    144.0430    40.8194   3.529
## TARGET_FLAG0:AGE                          51.2883    32.7545   1.566
## TARGET_FLAG1:AGE                         176.7488    48.5476   3.641
## TARGET_FLAG0:YOJ                         -21.6436    29.0844  -0.744
## TARGET_FLAG1:YOJ                          83.1716    51.9167   1.602
## TARGET_FLAG0:INCOME                       87.3026    46.4044   1.881
## TARGET_FLAG1:INCOME                     -292.3788    82.4195  -3.547
## TARGET_FLAG0:TRAVTIME                     14.0717    28.6553   0.491
## TARGET_FLAG1:TRAVTIME                     70.4384    49.3442   1.427
## TARGET_FLAG0:BLUEBOOK                     -0.4141    34.0365  -0.012
## TARGET_FLAG1:BLUEBOOK                    774.2379    60.2575  12.849
## TARGET_FLAG0:TIF                          -2.1602    28.3742  -0.076
## TARGET_FLAG1:TIF                        -211.0707    48.6265  -4.341
## TARGET_FLAG0:OLDCLAIM                     34.4627    42.4430   0.812
## TARGET_FLAG1:OLDCLAIM                   -612.3448    53.2366 -11.502
## TARGET_FLAG0:CLM_FREQ                    -24.1086    42.3402  -0.569
## TARGET_FLAG1:CLM_FREQ                    323.1709    54.6081   5.918
## TARGET_FLAG0:MVR_PTS                     -11.1591    33.2745  -0.335
## TARGET_FLAG1:MVR_PTS                     231.6953    45.1832   5.128
## TARGET_FLAG0:EDUCATION.L                -115.9451   129.0330  -0.899
## TARGET_FLAG1:EDUCATION.L                  43.1633   215.6166   0.200
## TARGET_FLAG0:EDUCATION.C                   1.2804    72.4887   0.018
## TARGET_FLAG1:EDUCATION.C                -357.6296   118.3050  -3.023
## TARGET_FLAG0:`JOBBlue Collar`           -419.4723   126.7119  -3.310
## TARGET_FLAG1:`JOBBlue Collar`            492.8868   199.1576   2.475
## TARGET_FLAG0:JOBClerical                -495.8922   122.8803  -4.036
## TARGET_FLAG1:JOBClerical                 858.8200   190.6683   4.504
## TARGET_FLAG0:JOBDoctor                  -366.2016   187.5902  -1.952
## TARGET_FLAG1:JOBDoctor                  -805.2224   428.7391  -1.878
## TARGET_FLAG0:JOBLawyer                  -435.4200   128.7557  -3.382
## TARGET_FLAG1:JOBLawyer                  1208.2596   236.9814   5.099
## TARGET_FLAG0:JOBManager                 -455.1447   115.1203  -3.954
## TARGET_FLAG1:JOBManager                 -925.9004   229.4867  -4.035
## TARGET_FLAG0:JOBProfessional            -451.0646   116.0287  -3.888
## TARGET_FLAG1:JOBProfessional            1763.1868   195.9964   8.996
## TARGET_FLAG0:JOBStudent                 -314.3454   155.4414  -2.022
## TARGET_FLAG1:JOBStudent                  508.7804   226.4465   2.247
```

```
## TARGET_FLAG0:CAR_USECommercial               -100.4509    88.8942  -1.130
## TARGET_FLAG1:CAR_USECommercial               1252.4709   139.2479   8.995
## TARGET_FLAG0:CAR_TYPEMinivan                  -85.6880    71.6873  -1.195
## TARGET_FLAG1:CAR_TYPEMinivan                  258.2139   138.6647   1.862
## TARGET_FLAG0:`CAR_TYPEPanel Truck`           -109.2627   136.5449  -0.800
## TARGET_FLAG1:`CAR_TYPEPanel Truck`           -785.3666   217.8553  -3.605
## TARGET_FLAG0:CAR_TYPEPickup                   -93.7736    88.5081  -1.059
## TARGET_FLAG1:CAR_TYPEPickup                   -60.0866   138.0882  -0.435
## TARGET_FLAG0:`CAR_TYPESports Car`            -125.2667   103.3371  -1.212
## TARGET_FLAG1:`CAR_TYPESports Car`              66.6815   147.6361   0.452
## TARGET_FLAG0:`URBANICITYHighly Rural/ Rural`  -73.8745    74.7444  -0.988
## TARGET_FLAG1:`URBANICITYHighly Rural/ Rural` -734.6768   220.3300  -3.334
## TARGET_FLAG0:`HOME_VAL_CAT.Very Low`         -354.4719   107.0368  -3.312
## TARGET_FLAG1:`HOME_VAL_CAT.Very Low`         -271.6505   176.7837  -1.537
## TARGET_FLAG0:HOME_VAL_CAT.Medium             -342.4793   100.6463  -3.403
## TARGET_FLAG1:HOME_VAL_CAT.Medium              -22.1083   177.4310  -0.125
## TARGET_FLAG0:HOME_VAL_CAT.High               -448.6537   119.8643  -3.743
## TARGET_FLAG1:HOME_VAL_CAT.High                737.6524   231.0058   3.193
## TARGET_FLAG0:HOME_VAL_CAT.Missing            -329.6272   104.8792  -3.143
## TARGET_FLAG1:HOME_VAL_CAT.Missing            -250.2335   166.2395  -1.505
## TARGET_FLAG0:REVOKED.No                      -482.8913    97.2306  -4.966
## TARGET_FLAG1:REVOKED.No                      2409.2792   122.3963  19.684
## TARGET_FLAG0:PARENT1.No                      -684.8673   106.7114  -6.418
## TARGET_FLAG1:PARENT1.No                       457.7510   131.8742   3.471
## TARGET_FLAG0:MSTATUS.No                      -254.6131    78.9919  -3.223
## TARGET_FLAG1:MSTATUS.No                      1114.8123   133.7245   8.337
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## KIDSDRIV:TARGET_FLAG0                       0.382330
## KIDSDRIV:TARGET_FLAG1                       0.000418 ***
## TARGET_FLAG0:AGE                            0.117398
## TARGET_FLAG1:AGE                            0.000272 ***
## TARGET_FLAG0:YOJ                            0.456784
## TARGET_FLAG1:YOJ                            0.109162
## TARGET_FLAG0:INCOME                         0.059936 .
## TARGET_FLAG1:INCOME                         0.000390 ***
## TARGET_FLAG0:TRAVTIME                       0.623383
## TARGET_FLAG1:TRAVTIME                       0.153449
## TARGET_FLAG0:BLUEBOOK                       0.990294
## TARGET_FLAG1:BLUEBOOK                        < 2e-16 ***
## TARGET_FLAG0:TIF                            0.939315
## TARGET_FLAG1:TIF                            1.43e-05 ***
## TARGET_FLAG0:OLDCLAIM                       0.416812
## TARGET_FLAG1:OLDCLAIM                        < 2e-16 ***
## TARGET_FLAG0:CLM_FREQ                       0.569088
## TARGET_FLAG1:CLM_FREQ                       3.30e-09 ***
## TARGET_FLAG0:MVR_PTS                        0.737352
## TARGET_FLAG1:MVR_PTS                        2.95e-07 ***
## TARGET_FLAG0:EDUCATION.L                    0.368890
## TARGET_FLAG1:EDUCATION.L                    0.841337
## TARGET_FLAG0:EDUCATION.C                    0.985908
## TARGET_FLAG1:EDUCATION.C                    0.002506 **
## TARGET_FLAG0:`JOBBlue Collar`               0.000933 ***
## TARGET_FLAG1:`JOBBlue Collar`               0.013335 *
```

```
## TARGET_FLAG0:JOBClerical                     5.46e-05 ***
## TARGET_FLAG1:JOBClerical                     6.69e-06 ***
## TARGET_FLAG0:JOBDoctor                       0.050932 .
## TARGET_FLAG1:JOBDoctor                       0.060375 .
## TARGET_FLAG0:JOBLawyer                       0.000721 ***
## TARGET_FLAG1:JOBLawyer                       3.44e-07 ***
## TARGET_FLAG0:JOBManager                      7.72e-05 ***
## TARGET_FLAG1:JOBManager                      5.48e-05 ***
## TARGET_FLAG0:JOBProfessional                 0.000102 ***
## TARGET_FLAG1:JOBProfessional                  < 2e-16 ***
## TARGET_FLAG0:JOBStudent                      0.043157 *
## TARGET_FLAG1:JOBStudent                      0.024660 *
## TARGET_FLAG0:CAR_USECommercial               0.258484
## TARGET_FLAG1:CAR_USECommercial                < 2e-16 ***
## TARGET_FLAG0:CAR_TYPEMinivan                 0.231979
## TARGET_FLAG1:CAR_TYPEMinivan                 0.062593 .
## TARGET_FLAG0:`CAR_TYPEPanel Truck`           0.423604
## TARGET_FLAG1:`CAR_TYPEPanel Truck`           0.000313 ***
## TARGET_FLAG0:CAR_TYPEPickup                  0.289385
## TARGET_FLAG1:CAR_TYPEPickup                  0.663470
## TARGET_FLAG0:`CAR_TYPESports Car`            0.225440
## TARGET_FLAG1:`CAR_TYPESports Car`            0.651516
## TARGET_FLAG0:`URBANICITYHighly Rural/ Rural` 0.322984
## TARGET_FLAG1:`URBANICITYHighly Rural/ Rural` 0.000856 ***
## TARGET_FLAG0:`HOME_VAL_CAT.Very Low`         0.000929 ***
## TARGET_FLAG1:`HOME_VAL_CAT.Very Low`         0.124396
## TARGET_FLAG0:HOME_VAL_CAT.Medium             0.000668 ***
## TARGET_FLAG1:HOME_VAL_CAT.Medium             0.900840
## TARGET_FLAG0:HOME_VAL_CAT.High               0.000182 ***
## TARGET_FLAG1:HOME_VAL_CAT.High               0.001408 **
## TARGET_FLAG0:HOME_VAL_CAT.Missing            0.001674 **
## TARGET_FLAG1:HOME_VAL_CAT.Missing            0.132269
## TARGET_FLAG0:REVOKED.No                      6.86e-07 ***
## TARGET_FLAG1:REVOKED.No                       < 2e-16 ***
## TARGET_FLAG0:PARENT1.No                      1.40e-10 ***
## TARGET_FLAG1:PARENT1.No                      0.000519 ***
## TARGET_FLAG0:MSTATUS.No                      0.001269 **
## TARGET_FLAG1:MSTATUS.No                       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4107 on 28500 degrees of freedom
## Multiple R-squared:  0.2945, Adjusted R-squared:  0.2929
## F-statistic: 185.9 on 64 and 28500 DF,  p-value: < 2.2e-16
```

This model is *significantly* better. And it also strikes us as quite an intuitive approach. It guarantees that only observations predicted as crashes will have a positive TARGET_AMT predicted. Still, let's see if we can stick with the intuition behind that approach, but improve our predictors.

Model 2 We again intend to leverage the TARGET_FLAG value, but this time we use an automated stepwise approach. This iterative method continuously adds or removes predictors based on their level of significance, aiming to achieve the best-fit model possible:

From the stepwise process, we learn that our best model is:

TARGET_AMT ~ AGE:TARGET_FLAG + TARGET_FLAG:INCOME + TARGET_FLAG:TIF + TARGET_FLAG:YOJ + TARGET_FLAG:HOMEKIDS + TARGET_FLAG:CLM_FREQ + TARGET_FLAG:CAR_AGE + TARGET_FLAG:OLDCLAIM + TARGET_FLAG:BLUEBOOK_transformed + TARGET_FLAG:INCOME_transformed + TARGET_FLAG:MVR_PTS_transformed + TARGET_FLAG:TIF_transformed + TARGET_FLAG:YOJ_transformed + TARGET_FLAG:CLM_FREQ_transformed + TARGET_FLAG:CAR_AGE_transformed + TARGET_FLAG:SEX.F + TARGET_FLAG:SEX.M + TARGET_FLAG:EDUCATION.L + TARGET_FLAG:EDUCATION.Q + TARGET_FLAG:EDUCATION^4 + TARGET_FLAG:JOBBlue Collar + TARGET_FLAG:JOBClerical + TARGET_FLAG:JOBDoctor + TARGET_FLAG:JOBHome Maker + TARGET_FLAG:JOBLawyer + TARGET_FLAG:JOBManager + TARGET_FLAG:JOBStudent + TARGET_FLAG:CAR_USECommercial + TARGET_FLAG:CAR_TYPEPanel Truck + TARGET_FLAG:CAR_TYPEPickup + TARGET_FLAG:CAR_TYPESports Car + TARGET_FLAG:CAR_TYPESUV + TARGET_FLAG:URBANICITYHighly Rural/ Rural + TARGET_FLAG:HOME_VAL_CAT.Low + TARGET_FLAG:HOME_VAL_CAT.Medium + TARGET_FLAG:HOME_VAL_CAT.High + TARGET_FLAG:REVOKED.No + TARGET_FLAG:PARENT1.No + TARGET_FLAG:MSTATUS.No

Now, this seems quite involved. But it's worth repeating–the constant presence of `TARGET_FLAG` is simply to guarantee that only observations with a `TARGET_FLAG` of 1 will have a non-zero value for `TARGET_AMT`. Let us now print the summary:

```
##
## Call:
## lm(formula = best_stepwise_formula, data = train_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9680      0      0      0  98577
##
## Coefficients: (1 not defined because of singularities)
##                                            Estimate Std. Error t value
## (Intercept)                               4.134e+03  3.963e+02  10.431
## AGE:TARGET_FLAG0                          3.526e-11  3.534e+01   0.000
## AGE:TARGET_FLAG1                          2.918e+02  5.137e+01   5.682
## TARGET_FLAG0:INCOME                       7.154e-12  5.735e+01   0.000
## TARGET_FLAG1:INCOME                       3.948e+02  1.247e+02   3.166
## TARGET_FLAG0:TIF                         -1.241e-11  6.672e+01   0.000
## TARGET_FLAG1:TIF                         -6.371e+02  1.329e+02  -4.794
## TARGET_FLAG0:YOJ                          6.131e-12  3.503e+01   0.000
## TARGET_FLAG1:YOJ                         -1.881e+02  5.945e+01  -3.164
## TARGET_FLAG0:HOMEKIDS                    -2.727e-12  3.864e+01   0.000
## TARGET_FLAG1:HOMEKIDS                     3.126e+02  6.182e+01   5.057
## TARGET_FLAG0:CLM_FREQ                    -1.942e-12  1.389e+02   0.000
## TARGET_FLAG1:CLM_FREQ                    -6.856e+02  1.735e+02  -3.951
## TARGET_FLAG0:CAR_AGE                     -2.095e-11  2.042e+02   0.000
## TARGET_FLAG1:CAR_AGE                     -3.420e+03  3.594e+02  -9.518
## TARGET_FLAG0:OLDCLAIM                    -4.013e-12  4.605e+01   0.000
## TARGET_FLAG1:OLDCLAIM                    -6.453e+02  5.971e+01 -10.806
## TARGET_FLAG0:BLUEBOOK_transformed         2.973e-12  3.846e+01   0.000
## TARGET_FLAG1:BLUEBOOK_transformed         9.770e+02  6.426e+01  15.205
## TARGET_FLAG0:INCOME_transformed           8.900e-12  7.218e+01   0.000
## TARGET_FLAG1:INCOME_transformed          -9.076e+02  1.502e+02  -6.044
## TARGET_FLAG0:MVR_PTS_transformed          8.522e-12  3.570e+01   0.000
## TARGET_FLAG1:MVR_PTS_transformed          2.356e+02  5.059e+01   4.657
## TARGET_FLAG0:TIF_transformed              8.488e-12  6.719e+01   0.000
## TARGET_FLAG1:TIF_transformed              4.480e+02  1.310e+02   3.419
```

```
## TARGET_FLAG0:YOJ_transformed                  -1.099e-11  4.389e+01   0.000
## TARGET_FLAG1:YOJ_transformed                   2.321e+02  6.503e+01   3.569
## TARGET_FLAG0:CLM_FREQ_transformed             -7.629e-12  2.115e+02   0.000
## TARGET_FLAG1:CLM_FREQ_transformed              1.554e+03  2.761e+02   5.629
## TARGET_FLAG0:CAR_AGE_transformed               1.997e-11  2.009e+02   0.000
## TARGET_FLAG1:CAR_AGE_transformed               3.124e+03  3.348e+02   9.330
## TARGET_FLAG0:SEX.F                            -4.134e+03  4.778e+02  -8.652
## TARGET_FLAG1:SEX.F                            -1.095e+03  1.617e+02  -6.770
## TARGET_FLAG0:SEX.M                            -4.134e+03  4.744e+02  -8.714
## TARGET_FLAG1:SEX.M                                    NA         NA      NA
## TARGET_FLAG0:EDUCATION.L                      -4.836e-11  1.468e+02   0.000
## TARGET_FLAG1:EDUCATION.L                       1.451e+03  2.653e+02   5.470
## TARGET_FLAG0:EDUCATION.Q                      -2.636e-11  8.260e+01   0.000
## TARGET_FLAG1:EDUCATION.Q                       1.682e+03  1.426e+02  11.793
## TARGET_FLAG0:`EDUCATION^4`                    -1.692e-12  6.223e+01   0.000
## TARGET_FLAG1:`EDUCATION^4`                    -4.282e+02  1.036e+02  -4.133
## TARGET_FLAG0:`JOBBlue Collar`                 -2.070e-12  1.133e+02   0.000
## TARGET_FLAG1:`JOBBlue Collar`                 -1.581e+03  1.808e+02  -8.743
## TARGET_FLAG0:JOBClerical                       9.629e-12  1.200e+02   0.000
## TARGET_FLAG1:JOBClerical                      -1.696e+03  1.989e+02  -8.530
## TARGET_FLAG0:JOBDoctor                         7.652e-11  1.894e+02   0.000
## TARGET_FLAG1:JOBDoctor                        -3.889e+03  4.418e+02  -8.801
## TARGET_FLAG0:`JOBHome Maker`                   2.536e-11  1.537e+02   0.000
## TARGET_FLAG1:`JOBHome Maker`                  -1.801e+03  2.776e+02  -6.487
## TARGET_FLAG0:JOBLawyer                         4.249e-11  1.193e+02   0.000
## TARGET_FLAG1:JOBLawyer                        -1.027e+03  2.443e+02  -4.202
## TARGET_FLAG0:JOBManager                        2.341e-11  9.928e+01   0.000
## TARGET_FLAG1:JOBManager                       -2.727e+03  2.215e+02 -12.312
## TARGET_FLAG0:JOBStudent                        1.440e-11  1.609e+02   0.000
## TARGET_FLAG1:JOBStudent                       -1.539e+03  2.575e+02  -5.977
## TARGET_FLAG0:CAR_USECommercial                 1.316e-11  8.833e+01   0.000
## TARGET_FLAG1:CAR_USECommercial                 7.899e+02  1.417e+02   5.573
## TARGET_FLAG0:`CAR_TYPEPanel Truck`             1.988e-11  1.368e+02   0.000
## TARGET_FLAG1:`CAR_TYPEPanel Truck`            -1.693e+03  2.227e+02  -7.604
## TARGET_FLAG0:CAR_TYPEPickup                    1.216e-11  8.910e+01   0.000
## TARGET_FLAG1:CAR_TYPEPickup                   -3.087e+02  1.501e+02  -2.056
## TARGET_FLAG0:`CAR_TYPESports Car`              9.077e-12  1.221e+02   0.000
## TARGET_FLAG1:`CAR_TYPESports Car`              5.371e+02  2.119e+02   2.534
## TARGET_FLAG0:CAR_TYPESUV                       8.935e-12  9.599e+01   0.000
## TARGET_FLAG1:CAR_TYPESUV                       6.841e+02  1.846e+02   3.706
## TARGET_FLAG0:`URBANICITYHighly Rural/ Rural`  -2.260e-12  7.443e+01   0.000
## TARGET_FLAG1:`URBANICITYHighly Rural/ Rural`  -7.216e+02  2.170e+02  -3.324
## TARGET_FLAG0:HOME_VAL_CAT.Low                  5.760e-12  9.213e+01   0.000
## TARGET_FLAG1:HOME_VAL_CAT.Low                  1.293e+03  1.587e+02   8.147
## TARGET_FLAG0:HOME_VAL_CAT.Medium               7.784e-12  9.033e+01   0.000
## TARGET_FLAG1:HOME_VAL_CAT.Medium               4.766e+02  1.589e+02   2.999
## TARGET_FLAG0:HOME_VAL_CAT.High                 9.023e-12  1.106e+02   0.000
## TARGET_FLAG1:HOME_VAL_CAT.High                 7.987e+02  2.184e+02   3.657
## TARGET_FLAG0:REVOKED.No                       -2.135e-12  9.999e+01   0.000
## TARGET_FLAG1:REVOKED.No                        1.895e+03  1.305e+02  14.517
## TARGET_FLAG0:PARENT1.No                       -1.404e-11  1.219e+02   0.000
## TARGET_FLAG1:PARENT1.No                       -3.940e+02  1.663e+02  -2.369
## TARGET_FLAG0:MSTATUS.No                       -1.373e-11  7.713e+01   0.000
## TARGET_FLAG1:MSTATUS.No                        7.713e+02  1.363e+02   5.659
```

```
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## AGE:TARGET_FLAG0                         1.000000
## AGE:TARGET_FLAG1                         1.35e-08 ***
## TARGET_FLAG0:INCOME                      1.000000
## TARGET_FLAG1:INCOME                      0.001546 **
## TARGET_FLAG0:TIF                         1.000000
## TARGET_FLAG1:TIF                         1.64e-06 ***
## TARGET_FLAG0:YOJ                         1.000000
## TARGET_FLAG1:YOJ                         0.001556 **
## TARGET_FLAG0:HOMEKIDS                    1.000000
## TARGET_FLAG1:HOMEKIDS                    4.28e-07 ***
## TARGET_FLAG0:CLM_FREQ                    1.000000
## TARGET_FLAG1:CLM_FREQ                    7.80e-05 ***
## TARGET_FLAG0:CAR_AGE                     1.000000
## TARGET_FLAG1:CAR_AGE                      < 2e-16 ***
## TARGET_FLAG0:OLDCLAIM                    1.000000
## TARGET_FLAG1:OLDCLAIM                     < 2e-16 ***
## TARGET_FLAG0:BLUEBOOK_transformed        1.000000
## TARGET_FLAG1:BLUEBOOK_transformed         < 2e-16 ***
## TARGET_FLAG0:INCOME_transformed          1.000000
## TARGET_FLAG1:INCOME_transformed          1.52e-09 ***
## TARGET_FLAG0:MVR_PTS_transformed         1.000000
## TARGET_FLAG1:MVR_PTS_transformed         3.23e-06 ***
## TARGET_FLAG0:TIF_transformed             1.000000
## TARGET_FLAG1:TIF_transformed             0.000629 ***
## TARGET_FLAG0:YOJ_transformed             1.000000
## TARGET_FLAG1:YOJ_transformed             0.000359 ***
## TARGET_FLAG0:CLM_FREQ_transformed        1.000000
## TARGET_FLAG1:CLM_FREQ_transformed        1.83e-08 ***
## TARGET_FLAG0:CAR_AGE_transformed         1.000000
## TARGET_FLAG1:CAR_AGE_transformed          < 2e-16 ***
## TARGET_FLAG0:SEX.F                        < 2e-16 ***
## TARGET_FLAG1:SEX.F                       1.31e-11 ***
## TARGET_FLAG0:SEX.M                        < 2e-16 ***
## TARGET_FLAG1:SEX.M                             NA
## TARGET_FLAG0:EDUCATION.L                 1.000000
## TARGET_FLAG1:EDUCATION.L                 4.55e-08 ***
## TARGET_FLAG0:EDUCATION.Q                 1.000000
## TARGET_FLAG1:EDUCATION.Q                  < 2e-16 ***
## TARGET_FLAG0:`EDUCATION^4`               1.000000
## TARGET_FLAG1:`EDUCATION^4`               3.59e-05 ***
## TARGET_FLAG0:`JOBBlue Collar`            1.000000
## TARGET_FLAG1:`JOBBlue Collar`             < 2e-16 ***
## TARGET_FLAG0:JOBClerical                 1.000000
## TARGET_FLAG1:JOBClerical                  < 2e-16 ***
## TARGET_FLAG0:JOBDoctor                   1.000000
## TARGET_FLAG1:JOBDoctor                    < 2e-16 ***
## TARGET_FLAG0:`JOBHome Maker`             1.000000
## TARGET_FLAG1:`JOBHome Maker`             8.91e-11 ***
## TARGET_FLAG0:JOBLawyer                   1.000000
## TARGET_FLAG1:JOBLawyer                   2.65e-05 ***
## TARGET_FLAG0:JOBManager                  1.000000
## TARGET_FLAG1:JOBManager                   < 2e-16 ***
```

```
## TARGET_FLAG0:JOBStudent                       1.000000
## TARGET_FLAG1:JOBStudent                       2.30e-09 ***
## TARGET_FLAG0:CAR_USECommercial                1.000000
## TARGET_FLAG1:CAR_USECommercial                2.53e-08 ***
## TARGET_FLAG0:`CAR_TYPEPanel Truck`            1.000000
## TARGET_FLAG1:`CAR_TYPEPanel Truck`            2.97e-14 ***
## TARGET_FLAG0:CAR_TYPEPickup                   1.000000
## TARGET_FLAG1:CAR_TYPEPickup                   0.039787 *
## TARGET_FLAG0:`CAR_TYPESports Car`             1.000000
## TARGET_FLAG1:`CAR_TYPESports Car`             0.011276 *
## TARGET_FLAG0:CAR_TYPESUV                      1.000000
## TARGET_FLAG1:CAR_TYPESUV                      0.000211 ***
## TARGET_FLAG0:`URBANICITYHighly Rural/ Rural`  1.000000
## TARGET_FLAG1:`URBANICITYHighly Rural/ Rural`  0.000887 ***
## TARGET_FLAG0:HOME_VAL_CAT.Low                 1.000000
## TARGET_FLAG1:HOME_VAL_CAT.Low                 3.89e-16 ***
## TARGET_FLAG0:HOME_VAL_CAT.Medium              1.000000
## TARGET_FLAG1:HOME_VAL_CAT.Medium              0.002709 **
## TARGET_FLAG0:HOME_VAL_CAT.High                1.000000
## TARGET_FLAG1:HOME_VAL_CAT.High                0.000256 ***
## TARGET_FLAG0:REVOKED.No                       1.000000
## TARGET_FLAG1:REVOKED.No                        < 2e-16 ***
## TARGET_FLAG0:PARENT1.No                       1.000000
## TARGET_FLAG1:PARENT1.No                       0.017840 *
## TARGET_FLAG0:MSTATUS.No                       1.000000
## TARGET_FLAG1:MSTATUS.No                       1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4053 on 28487 degrees of freedom
## Multiple R-squared:  0.3132, Adjusted R-squared:  0.3114
## F-statistic: 168.7 on 77 and 28487 DF,  p-value: < 2.2e-16
```

Model Comparison We thus have two models, with a major similarity and a major difference. The major similarity is that they both leverage TARGET_FLAG in making a prediction for TARGET_AMT. This seems quite appropriate, which is why we retained this approach throughout. They are different, however, in that the first w selected variables manually (based on our best logistic model), and for the second we leveraged a stepwise regression approach. Needless to say, it's quite important to compare how they actually perform!

We summarize key metrics below:

Table 21: Multiple Linear Regression Models Comparison

Model	RMSE	MAE	R_squared
Best Stepwise	3637.063	988.030	0.2676403
MLR2	3598.572	1129.521	0.2830594

This table is helpful in guiding our model selection. MLR2 has a slightly lower RMSE value than Best Stepwise. It also has a slightly higher R-squared value. Now, in this particular instance, RMSE and MAE are more important than R-squared. This is because we're specifically interested in the financial impact of prediction errors–lower prediction errors mean more accurate payouts. But while Best Stepwise performs slightly worse with respect to RMSE and R-squared, it performs *far* better with respect to MAE. This metric

tells us the average error to expect, and so we find it meaningful that Best Stepwise performs better. It's also worth noting that predictive performance is extremely important in this regard, arguably more so than interpretability–there is no need to prioritize a much simpler model. Thus, despite the fact that MLR2 does slightly better with two metrics, because Best Stepwise has a far lower MAE value, we select that model.

Predictions

As we have now selected our models, we are ready to make predictions on the evaluation set. This is a slightly complicated process because our second model is dependent on our first one. We complete this process below:

Table 22: Sample 10 Predictions for Evaluation Dataset

	Index	TARGET_FLAG	TARGET_AMT
11	11	0	0.000
12	12	1	4906.043
13	13	1	6117.107
14	14	0	0.000
15	15	0	0.000
16	16	1	6965.201
17	17	1	6707.136
18	18	0	0.000
19	19	1	4674.479
20	20	1	3595.127

Conclusion

In this project, we aimed to predict whether or not crashes occurred, as well as the payout if they did. The construction of these models entailed vast data exploration and transformation. We predicted whether or not crashes would occur using binary logistic regression models, and we selects the best one. We then leveraged our predictions in our predictions of payouts, by creating interaction terms. We also utilized stepwise regression for our highest performing model. Finally, we generated predictions using these two models in order. Using these models would assist insurance companies in offering policies that are both fair as well as profitable and risk-averse.