Insurance

Nathan Belete

5/11/2018

Contents

Introduction

The purpose of this report is to analyze data collected from an auto insurance company. The data contains information on customers and weather they've been in a car accident and the claim amount. In this paper I will perform an exploratory data analysis, prepare data for modeling, explore various models from a predictive modeling prespective, quantify the predictive accuracy of the models using both in-sample and out-of-sample data to create an optimal model for deployment.

Sample Population

The insuarnce claims data set has 8,161 observations and 26 variables. For this paper, I am interested in using Logistic Regression Models to predict the number of whether a customer will be invovled in a car accident, and if so, the expected claim amount. For the purpose of exploratory data analysis, I will use the data as is. Later on, I will split my data 70/30 for the purpose of quantifying the predictive accuracy of the models.

Listed below are the descriptions of the 26 variables in the insurance dataset.

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

Data Preparation

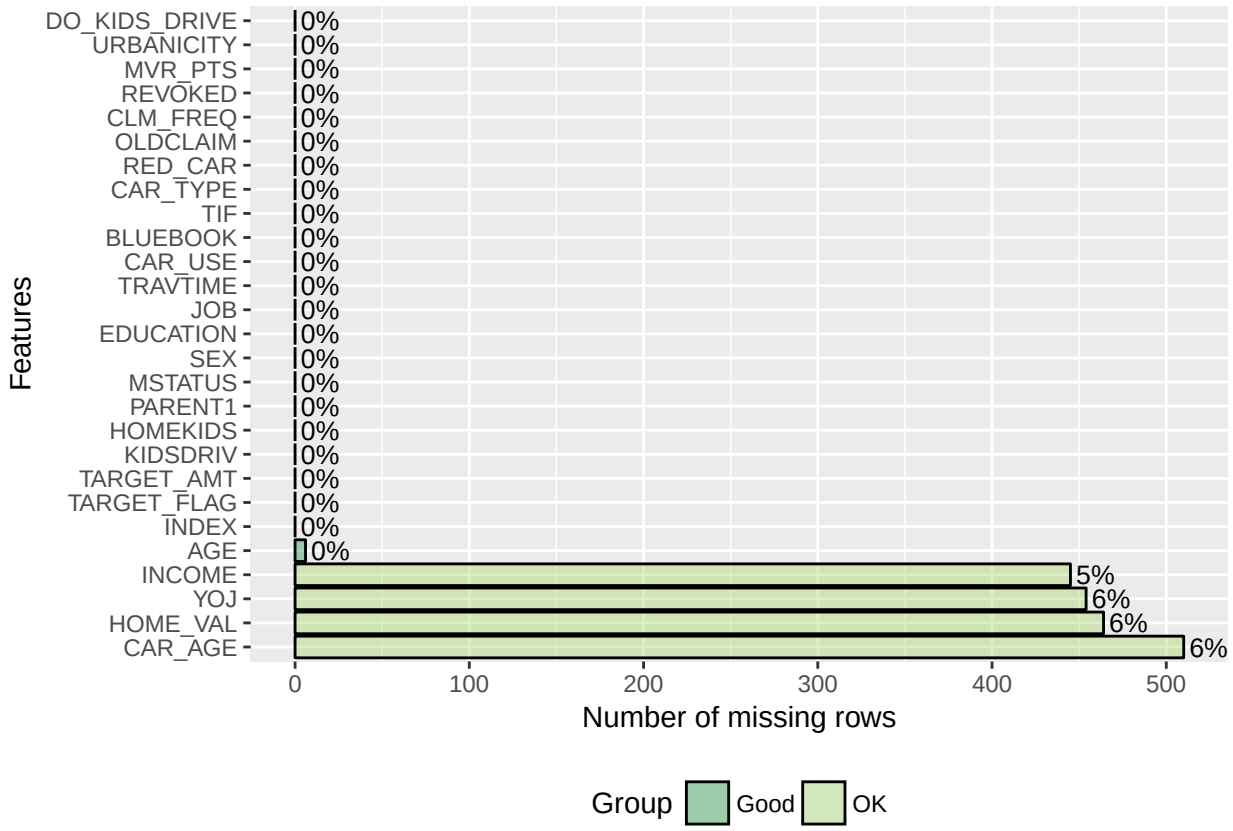
Taking a quick look at the summary statistics, we see that some variables in our dataset have missing values; we'll want to take care of missing values. To gain a deeper understanding of our data, lets go ahead and create a summary statistics of our variables before we visually inspect variables with missing values.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
TARGET_AMT	0	0	0	1504	1036	107586	0
AGE	16	39	45	45	51	81	6
YOJ	0	9	11	10	13	23	454
INCOME	0	28097	54028	61898	85986	367030	445
HOME_VAL	0	0	161160	154867	238724	885282	464
TIF	1	1	4	5	7	25	1
OLDCLAIM	0	0	0	4037	4636	57037	0
CLM_FREQ	0	0	0	1	2	5	0
MVR_PTS	0	0	1	2	3	13	0
CAR_AGE	-3	1	8	8	12	28	510

Missing Variables

Below we have a plot and a table of the missing values in our dataset. As you can see below, we have about four variables that have atleast 5% of their observation missing; we'll want to estimate these missing values.

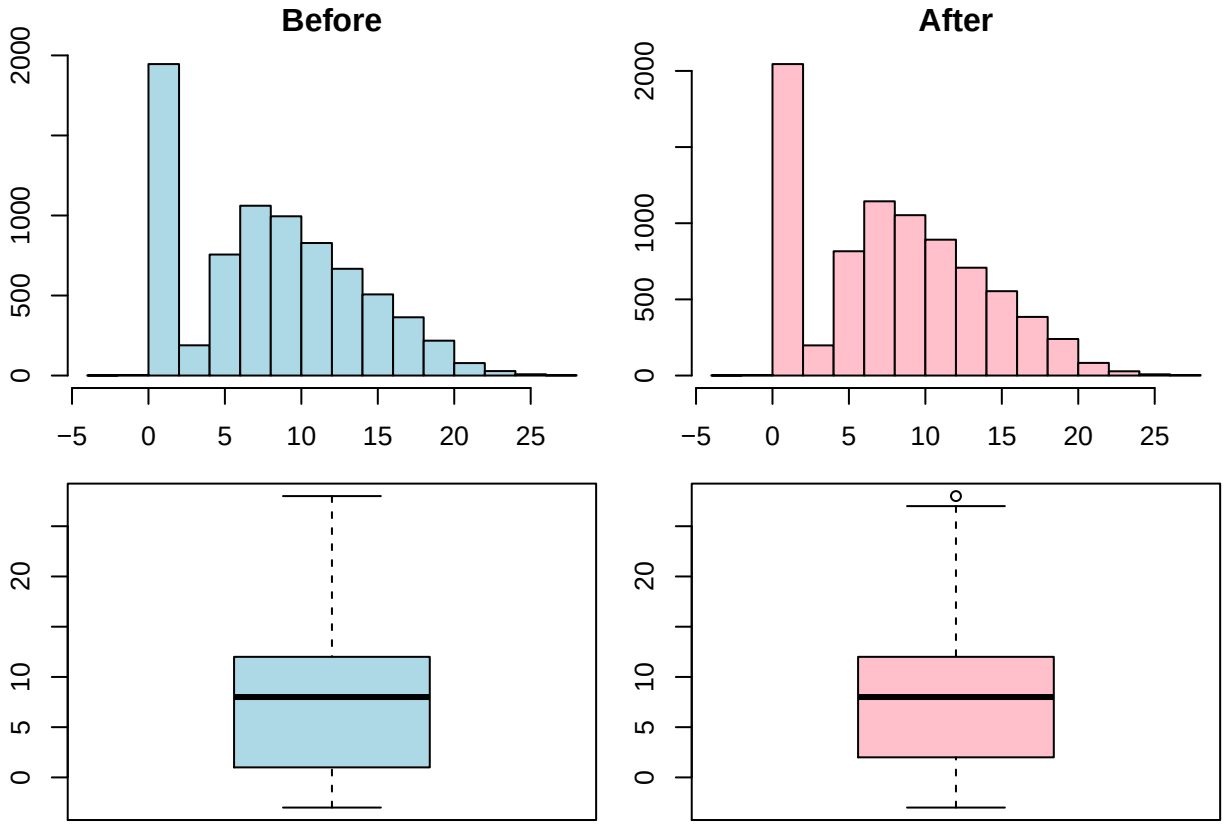
	feature	num_missing	pct_missing	group
25	CAR_AGE	510	0.0624923	OK
10	HOME_VAL	464	0.0568558	OK
7	YOJ	454	0.0556304	OK
8	INCOME	445	0.0545276	OK
5	AGE	6	0.0007352	Good
1	INDEX	0	0.0000000	Good
2	TARGET_FLAG	0	0.0000000	Good
3	TARGET_AMT	0	0.0000000	Good
4	KIDSDRIV	0	0.0000000	Good
6	HOMEKIDS	0	0.0000000	Good
9	PARENT1	0	0.0000000	Good
11	MSTATUS	0	0.0000000	Good
12	SEX	0	0.0000000	Good
13	EDUCATION	0	0.0000000	Good
14	JOB	0	0.0000000	Good
15	TRAVTIME	0	0.0000000	Good
16	CAR_USE	0	0.0000000	Good
17	BLUEBOOK	0	0.0000000	Good
18	TIF	0	0.0000000	Good
19	CAR_TYPE	0	0.0000000	Good
20	RED_CAR	0	0.0000000	Good
21	OLDCLAIM	0	0.0000000	Good
22	CLM_FREQ	0	0.0000000	Good
23	REVOKED	0	0.0000000	Good
24	MVR_PTS	0	0.0000000	Good
26	URBANICITY	0	0.0000000	Good
27	DO_KIDS_DRIVE	0	0.0000000	Good


DO_KIDS_DRIVE
URBANICITY
MVR_PTS
REVOKED
CLM_FREQ
OLDCLAIM
RED_CAR
CAR_TYPE
TIF
BLUEBOOK
CAR_USE
TRAVTIME
JOB
EDUCATION
SEX
MSTATUS
PARENT1
HOMEKIDS
KIDSDRIV
TARGET_AMT
TARGET_FLAG
INDEX
AGE
INCOME
YOJ
HOME_VAL
CAR_AGE
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
0%
5%
6%
6%
6%
Features
0
100
200
300
400
500
Number of missing rows
Group
Good
OK

Estimate Missing Values

Now that we have found the variables with missing values, lets go ahead and estimate these missing values. To predict these missing values, I am going to use predictive mean matching method (PMM). PMM, according to SAS, is an an imputation method that imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model - that was a mouthful. Basically, it calculates the predicted value using a regression model and picks the closest 'real' value to the predicted value.

Now that we've estimated the missing values, lets show a before and after picture of the 'Car Age' variable.



As you can see above, the imputation process did not change the orginal skewness nor did it change the lower or upper bound of the original data.

Now that we've imputed the missing values, lets go ahead and recreate the summary statistics table just to confirm that we've accounted for all missing values before moving forward.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TARGET_AMT	0	0	0	1504	1036	107586
AGE	16	39	45	45	51	81
YOJ	0	9	11	11	13	23
INCOME	0	27693	53698	61728	85752	367030
HOME_VAL	0	0	161166	155183	239022	885282
TRAVTIME	5	22	33	33	44	142
BLUEBOOK	1500	9280	14440	15710	20850	69740
TIF	1	1	4	5	7	25
OLDCLAIM	0	0	0	4037	4636	57037
CLM_FREQ	0	0	0	1	2	5
MVR_PTS	0	0	1	2	3	13
CAR_AGE	0	2	8	8	12	28

Great, we have no missing value!

Additional Features

Before moving on to doing Exploratory Data Analysis, lets go ahead and create dummary variables to identify the imputed/missing values. While we are at it, we'll also create some dummary variables for our logistic regression models. Listed below are some of the variables that were created.

Variable Name	Definition
CAR_AGE_NA_IND	Car Age Imputted/Missing value indicator
HOME_VAL_NA_IND	Home Value Imputted/Missing value indicator
INCOME_NA_IND	Income Imputted/Missing value indicator
YOJ_NA_IND	YOJ Imputted/Missing value indicator
AGE_NA_IND	Age Imputted/Missing value indicator
MSTATUS_Yes	Married
SEX_M	Male

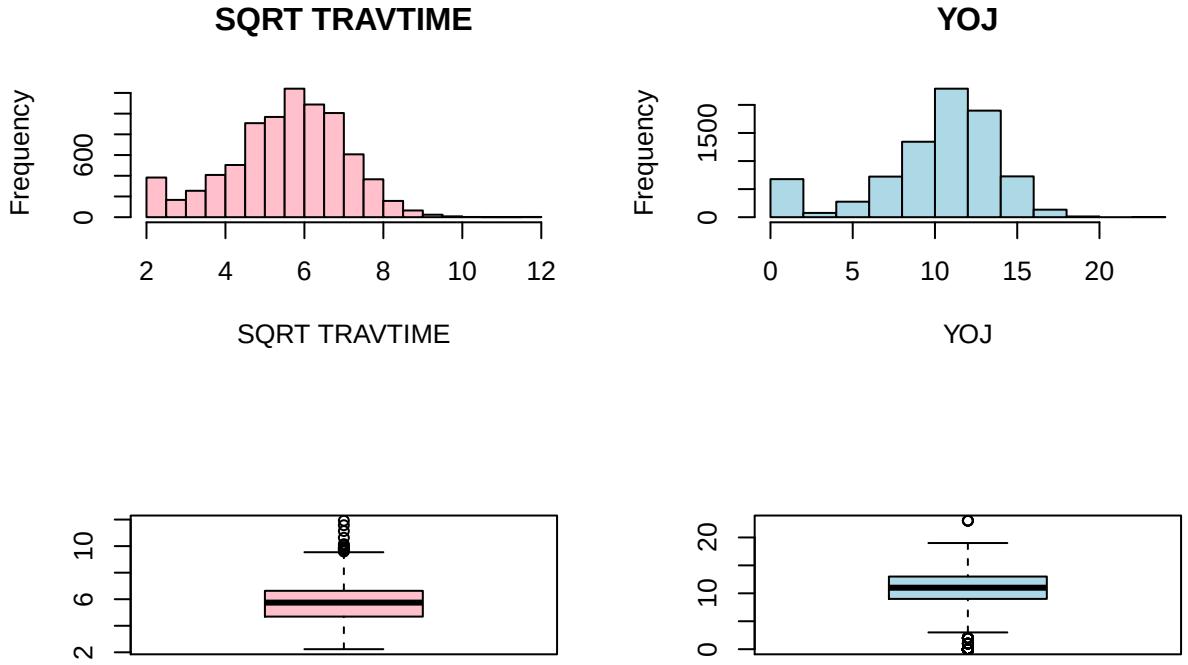
Exploratory Data Analysis

Target Variables

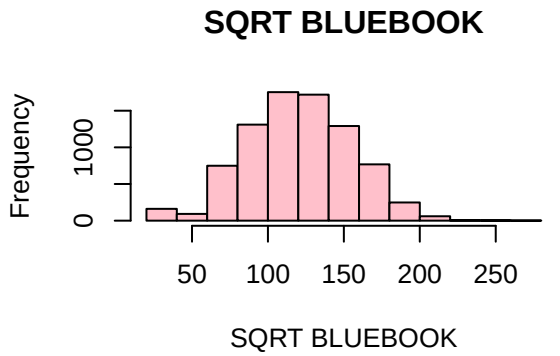
Now that we've taken care of the missing values, lets start our exploratory data analysis by looking at what we want to predict, the 'Target Flag' variable. As you can see below, the summary statistics shows that the variable has a mean of .264. Taking a look at the Cost Amount' variable we see that the average cost is around $1,504 and the median is around $0. Further, when look at the cost with respect to customer that experenced car accidents, we see that the median is around $4,104 and the average is around $5,702 with a max and min of $107,586 and $30, respectively.

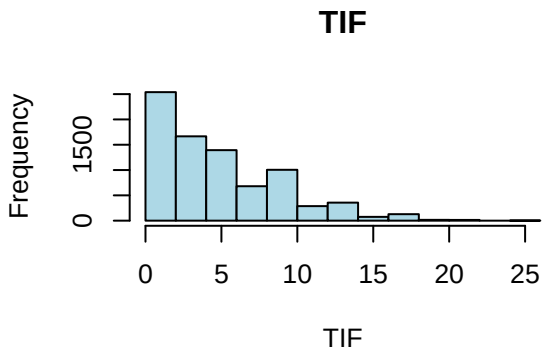
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Insuarnce Data Set: Target Flag	0	0	0	0.264	1	1
Insuarnce Data Set: Cost Amount	0	0	0	1504.000	1036	107586
Insuarnce Data Set: Cost Amount given Car Accident	30	2610	4104	5702.000	5787	107586
Insuarnce Data Set: Cost Amount given No Car Accident	0	0	0	0.000	0	0

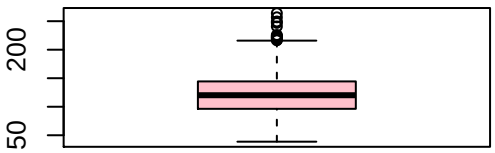
Next, we have a plot of travel time and years on the job.

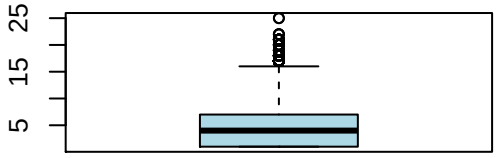


Next, we have a plot of Bluebook and Time in Force.

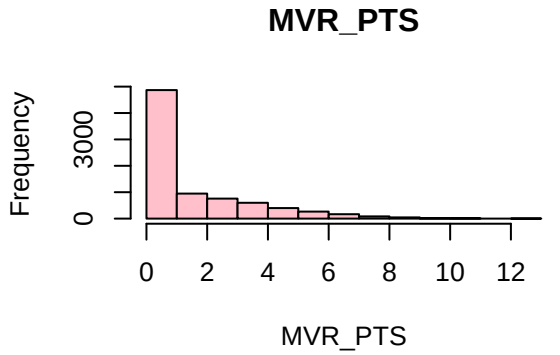


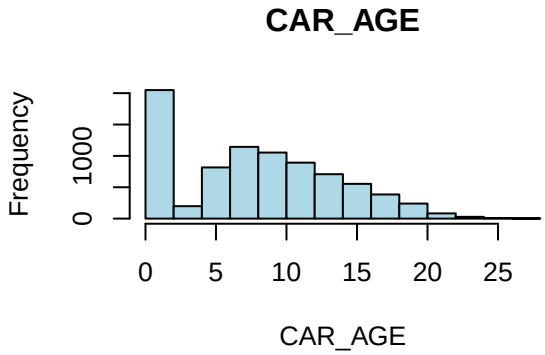


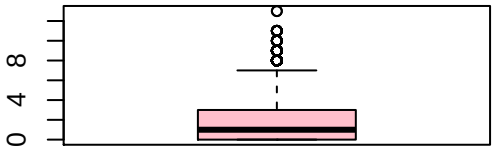


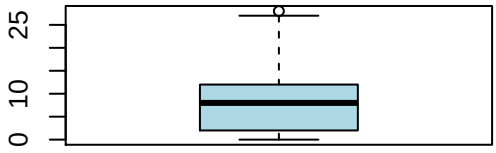


Finally we have a plot of Motor Vehicle Record Points and Car Age.





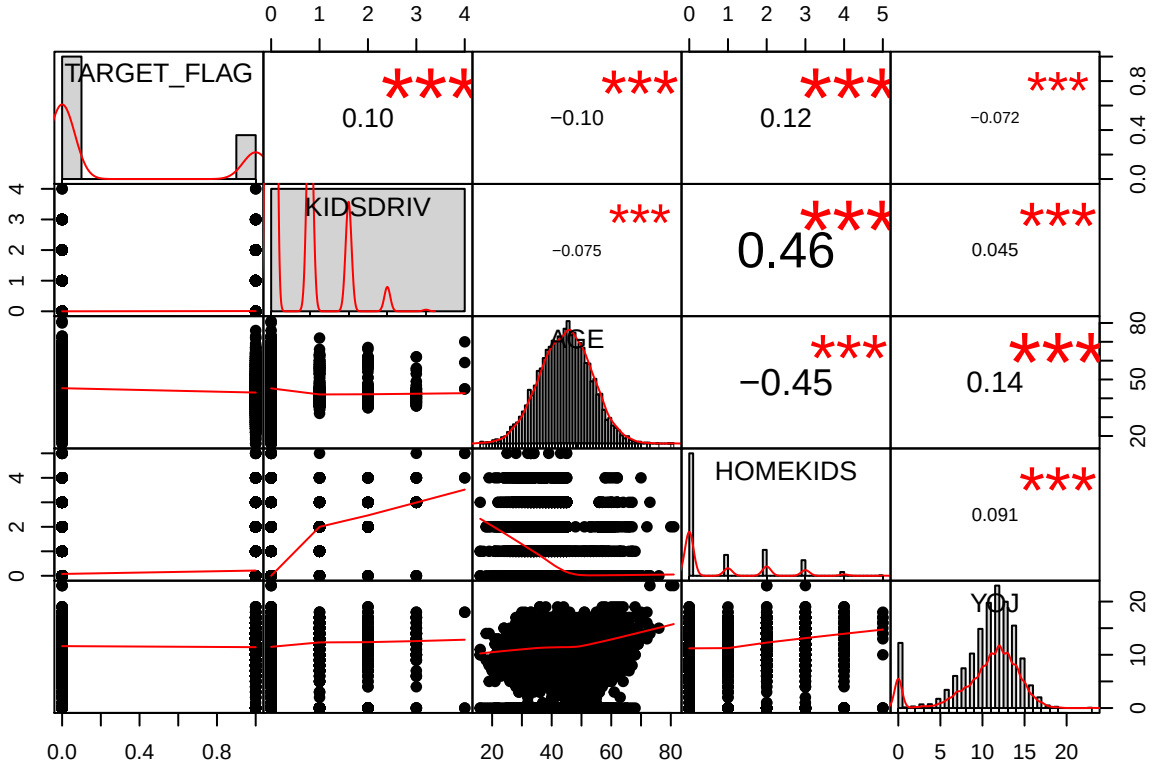




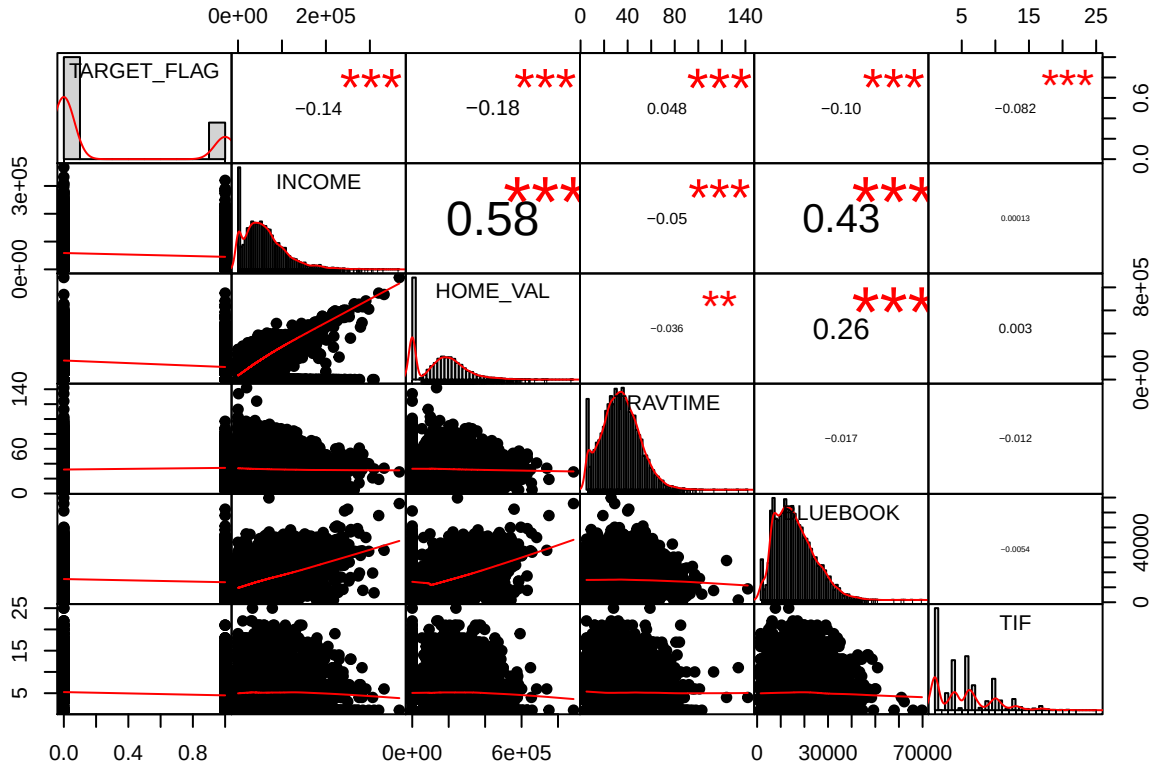
Correlation

Now that we have visually explored our dataset, lets now see how these variables correlate to our response variable and to each other.

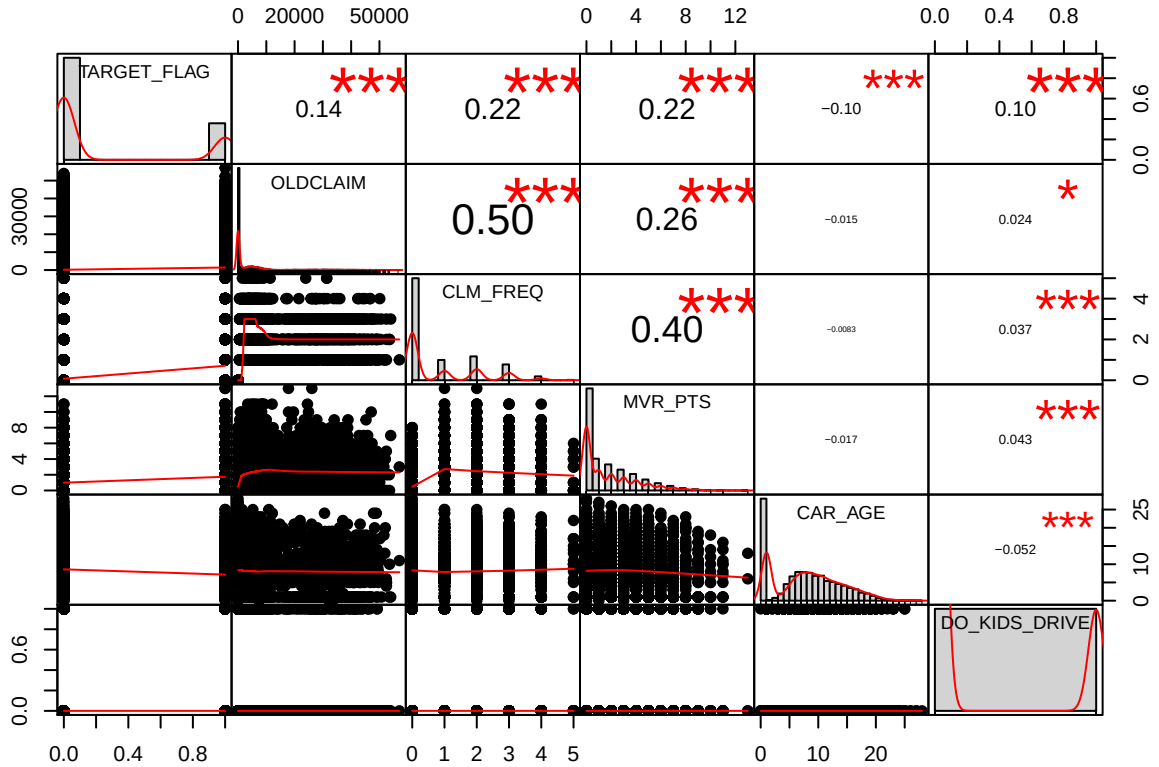
Taking a look at the correlation plot below, we see that our Target Flag variable doesn't have much correlation with the Kids Driving, Age, Kids at Home, and Years on the Job variables. However, we see that Kids at Home has a strong absolute correlation with Age and Kids Drving. We'll want to keep this in mind to aviod multicollinearity.



Next, we have a correlation plot of the Traget Flag variables and a few explanatory variables. As you can see, our Target Flag variable doesn't correlate well with these variables either. Interestingly, we see that Income correlates highly with Home Value and the Bluebook value variables.



Finally, we have another correlation plot. Once again we see that our Target Flag variable doesn't correlate well with our explantory variables. Old claim and Clain freqency seem to show some correlation, however, that pretty much it.



Overall, there doesn't seem to be variables that correlete well with out Target Flag variable.

Model Development

Model Prepartaion

Normalize data

Before moving on the developing our models, I'm going to first normalize all my data such that each variable has values between 0 and 1. I do this because I want to remove the impact of very large and very small variables.

Here is a quick snapshot of what the insuarnce dataset looks like now.

AGE	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK
0.677	0.478	0.183	0.000	0.066	0.187
0.415	0.478	0.249	0.291	0.124	0.197
0.292	0.435	0.044	0.140	0.000	0.037
0.538	0.609	0.078	0.346	0.197	0.204
0.523	0.609	0.313	0.276	0.226	0.242
0.277	0.522	0.341	0.000	0.299	0.233

Train / Test Split

To assess the Predictive Accuracy and Validity of our models, our insurance dataset was split into two groups; training and testing datasets. The spilt was done using a random uniform distribution such that roughly 70% of our data is used to train our models and the remaining 30% is used to test our models out-of-sample accuracy. The number of observations in each table is shown below.

Data	Count
Insurance	8,161
Training	5,733
Testing	2,428

Ok, now that I have normalized my insurance dataset and created a training and testing datasets, lets do a quick high level summary statistics of my Target Flag variable.

As you can see below, the training and testing datasets have a the Target Flag variable mean of 0.267 and 0.257, respectively. From here on, we'll use these values as a proxy to assess the accuracy of our models.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Training Data Set	0	0	0	0.267	1	1
Testing Data Set	0	0	0	0.257	1	1

Variable Selection

In this paper, I will use the backwards, forwards, and stepwise variable selection methods to build regression models.

Backward Variable Selection

I will start building my first model using Backward Variable Selection (BVS). With BVS, we start off with a current model that includes all of the explanatory variables. Next, we consider a candidate model that is different than the current model by the removal of one explanatory variable. If the candidate model has an AIC value that is less than our current model, we accept the candidate model as our current model. We then repeat the same steps until all explanatory variables are tested such that the current model has the lowest AIC value relative to all candidate models.

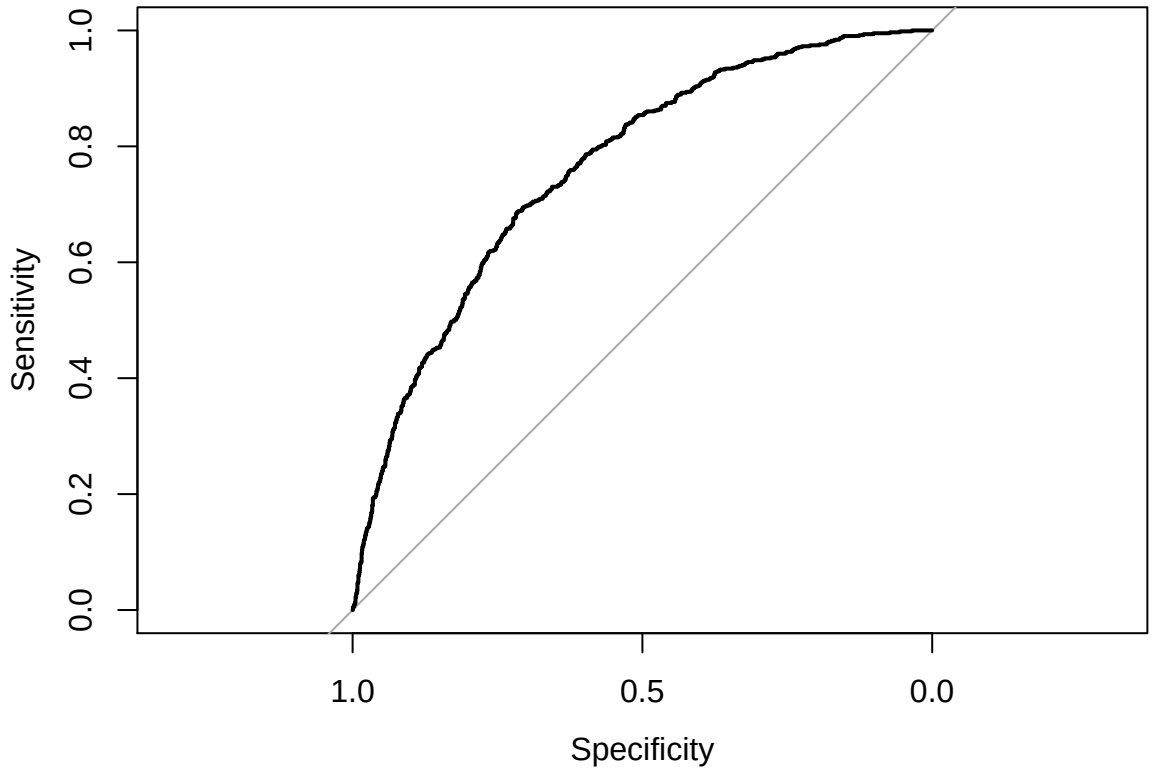
As you can see below, the optimal cutoff point for the BVS model is 0.53. Further, using this cutoff, we see that the model correctly classifies observations 76.7% of the time. This model also has an AIC value of 5716 and an AUC value of 0.763.

Model	Optimal Cutoff	AIC	AUC	Prob of Correct Classification	Prob of Incorrect Classification
Backward Variable Selection	0.53	5716	0.763	0.767	0.233

Further, taking a look at our probability distribution, we see that our BVS model has mean of 0.264.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Backward Variable Selection	0.011	0.121	0.219	0.264	0.368	0.879

Finally, taking a look at our PUC plot, we see that the plot is close to the upper left corner, implying that our model is reliable. As noted earlier, the AUC value is 0.763



Forward Variable Selection

Next, we will build our second model using Forward Variable Selection (FVS). With FVS, we start off with a current model that includes no explanatory variables but just the y-intercept. We then consider a candidate model that is different than the current model by the addition of one explanatory variable. If the candidate model has an AIC value that is less than our current model, we accept the candidate model as our current model. We then repeat the same steps until all explanatory variables are tested such that the current model has the lowest AIC value relative to all candidate models.

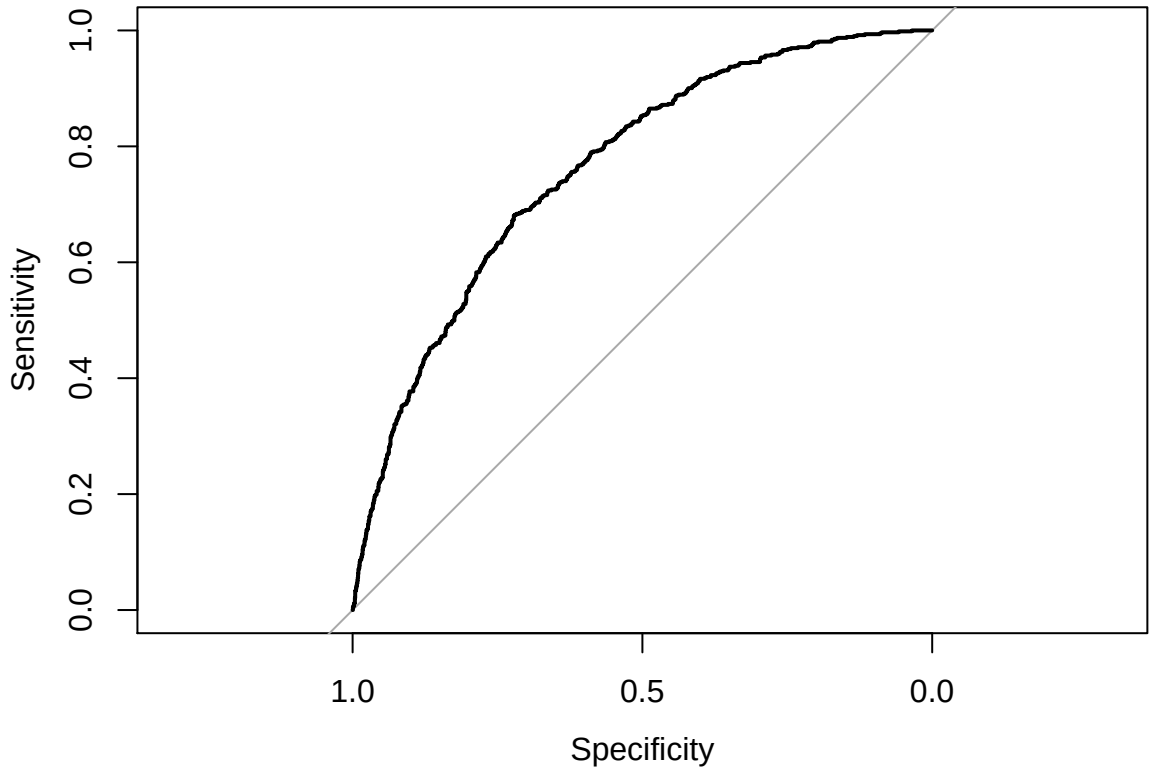
As you can see below, the optimal cutoff point for the FVS model is 0.53. Further, using this cutoff, we see that the model correctly classifies observations 76.6% of the time. This model also has an AIC value of 5716 and an AUC value of 0.762.

Model	Optimal Cutoff	AIC	AUC	Prob of Correct Classification	Prob of Incorrect Classification
Forward Variable Selection	0.53	5716	0.762	0.766	0.234

Further, taking a look at our probability distribution, we see that our FVS model has mean of 0.263.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Forward Variable Selection	0.011	0.121	0.22	0.263	0.369	0.87

Finally, taking a look at our PUC plot, we see that the plot is close to the upper left corner, implying that our model is reliable. As noted earlier, the AUC value is 0.762



Stepwise Variable Selection

Next, we will build our third model using Stepwise Variable Selection (SVS). With SVS, we start off with a current model that includes no explanatory variables but just the y-intercept. We then consider a candidate model that is different than the current model by the addition and subtraction of one explanatory variable. If the candidate model has an AIC value that is less than our current model, then we accept the candidate model as our current model. We then repeat the same steps until all explanatory variables are tested such that the current model has the lowest AIC value relative to all candidate models.

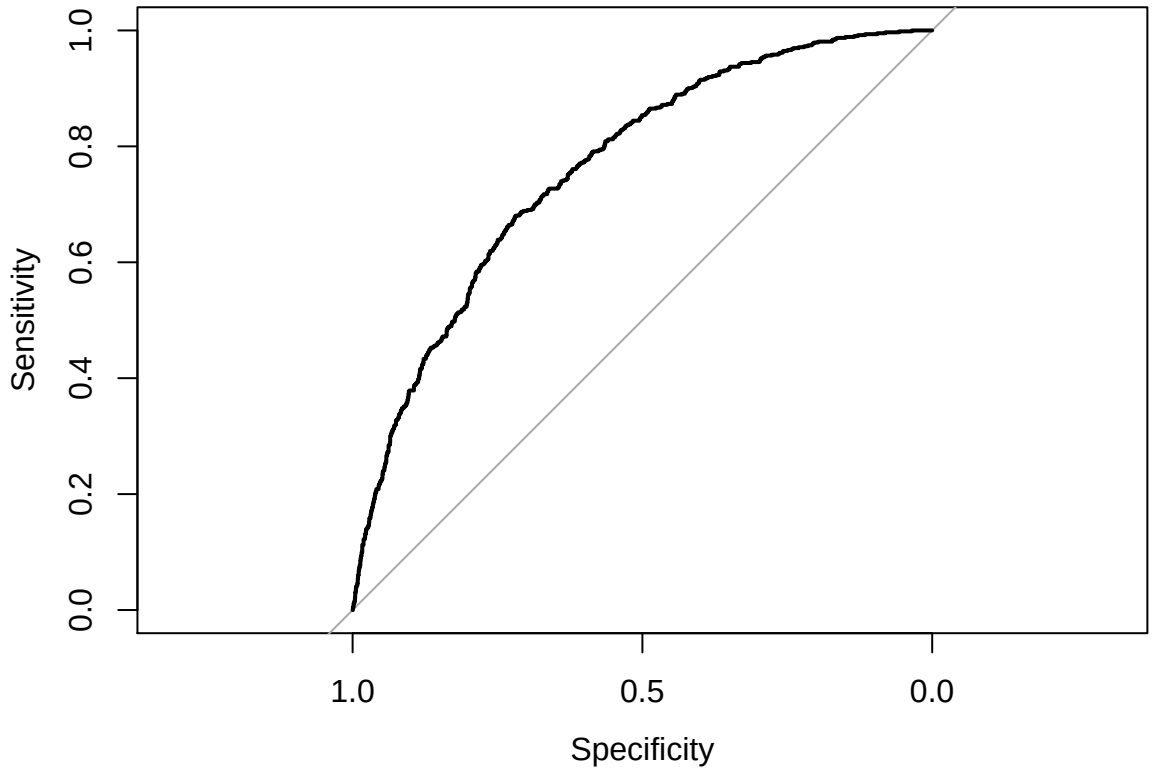
As you can see below, the optimal cutoff point for the SVS model is 0.5. Further, using this cutoff, we see that the model correctly classifies observations 77.1% of the time. This model also has an AIC value of 5715 and an AUC value of 0.762.

Model	Optimal Cutoff	AIC	AUC	Prob of Correct Classification	Prob of Incorrect Classification
Stepwise Variable Selection	0.5	5715	0.762	0.771	0.229

Further, taking a look at our probability distribution, we see that our SVS model has mean of 0.263.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Stepwise Variable Selection	0.011	0.121	0.22	0.263	0.369	0.87

Finally, taking a look at our PUC plot, we see that the plot is close to the upper left corner, implying that our model is reliable. As noted earlier, the AUC value is 0.762.



Model Comparison

Now that we've built all four of our models, lets see how they compare relative to each other using our training data set.

Take a look at the AIC values, it seems like the SVS model has the lowest AIC value. Looking at the AUC values, it seems like the BVS model has the highest AUC value. Looking at the probability of correct classification at the optimal cutoff value, we see that SVS model performed best.

Given how the SVS model had the highest probability of correct classification and lowest AIC value, SVS model might be the perferred model. Before moving forward with the SVS model, lets examine the distribution of the probability values.

Model	Optimal Cutoff	AIC	AUC	Prob of Correct Classification	Prob of Incorrect Classification
BVS	0.53	5716	0.763	0.767	0.233
FVS	0.53	5716	0.762	0.766	0.234
SVS	0.5	5715	0.762	0.771	0.229

Below we have the distribution of the target variable for the training, testing, and our logistic regression models. As you can see below, the mean distribution of our BVS, FVS, and SVS models are all very close to our testing and training dataset. Given what we've learned about our SVS model, this pretty much confirms that the SVS model is the perferred model.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Training Data Set	0.000	0.000	0.000	0.267	1.000	1.000
Testing Data Set	0.000	0.000	0.000	0.257	1.000	1.000
Backward Variable Selection	0.011	0.121	0.219	0.264	0.368	0.879
Forward Variable Selection	0.011	0.121	0.220	0.263	0.369	0.870
Stepwise Variable Selection	0.011	0.121	0.220	0.263	0.369	0.870

Model Selection

As mentioned above, I've selected the SVS model as the model. With this in mind, lets see coefficients of the model below.

Variables	Coefficients
Intercept	- 0.77172
MVR_PTS	+ 1.72122
HOME_VAL	- 1.34378
CAR_USE_Commercial	+ 0.59597
CLM_FREQ	+ 1.08632
Parent_Yes	+ 0.25199
JOB_Manager	- 0.58213
CAR_TYPE_Minivan	- 0.50878
TIF	- 1.18014
BLUEBOOK	- 0.93475
AGE_Btwn_45_and_55	- 0.47548
AGE_Btwn_35_and_45	- 0.34225
TRAVTIME	+ 1.03690
EDUCATION_Bachelors	- 0.27808
INCOME_High	- 0.32489
MSTATUS_Yes	- 0.39595
CAR_TYPE_Sports_Car	0.26686
OLDCLAIM	0.51097
INCOME_Zero	0.33800
JOB_z_Blue_Collar	0.26310
AGE_Btwn_65_and_75	- 0.79341
AGE_NA_IND	1.83307
DO_KIDS_DRIVE	0.63539
AGE_Btwn_21_and_25	0.58461

Appendix

BVS

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + HOME_VAL + TRAVTIME + BLUEBOOK +
##     TIF + OLDCLAIM + CLM_FREQ + MVR_PTS + DO_KIDS_DRIVE + AGE_NA_IND +
##     Parent_Yes + MSTATUS_Yes + EDUCATION_Bachelors + JOB_z_Blue_Collar +
##     JOB_Manager + CAR_USE_Commercial + CAR_TYPE_Minivan + CAR_TYPE_Sports_Car +
##     INCOME_Zero + INCOME_Low + INCOME_Medium + AGE_Btwn_25_and_35 +
##     AGE_Btwn_35_and_45 + AGE_Btwn_45_and_55 + AGE_Btwn_65_and_75,
##     family = binomial(), data = insurance.train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1192  -0.7619  -0.5139   0.8046   2.7549
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.36176    0.34420  -1.051 0.293254
## AGE                 -1.15917    0.48757  -2.377 0.017433 *
## HOME_VAL            -1.34299    0.33026  -4.066 4.77e-05 ***
## TRAVTIME             1.04383    0.28172   3.705 0.000211 ***
## BLUEBOOK            -0.92583    0.30351  -3.050 0.002285 **
## TIF                 -1.17660    0.19964  -5.894 3.78e-09 ***
## OLDCLAIM             0.51287    0.21944   2.337 0.019428 *
## CLM_FREQ             1.08765    0.15772   6.896 5.35e-12 ***
## MVR_PTS              1.70269    0.20294   8.390  < 2e-16 ***
## DO_KIDS_DRIVE        0.64090    0.10292   6.227 4.76e-10 ***
## AGE_NA_IND           1.85380    1.10766   1.674 0.094205 .
## Parent_Yes           0.23246    0.11457   2.029 0.042465 *
## MSTATUS_Yes         -0.39884    0.09044  -4.410 1.03e-05 ***
## EDUCATION_Bachelors -0.28412    0.07818  -3.634 0.000279 ***
## JOB_z_Blue_Collar    0.25703    0.09060   2.837 0.004556 **
## JOB_Manager         -0.58309    0.12984  -4.491 7.10e-06 ***
## CAR_USE_Commercial   0.59548    0.07987   7.456 8.95e-14 ***
## CAR_TYPE_Minivan    -0.50855    0.08676  -5.862 4.58e-09 ***
## CAR_TYPE_Sports_Car  0.27085    0.10393   2.606 0.009159 **
## INCOME_Zero          0.65496    0.14944   4.383 1.17e-05 ***
## INCOME_Low           0.29900    0.12187   2.453 0.014150 *
## INCOME_Medium        0.33298    0.09480   3.512 0.000444 ***
## AGE_Btwn_25_and_35  -0.44979    0.22056  -2.039 0.041420 *
## AGE_Btwn_35_and_45  -0.64196    0.15525  -4.135 3.55e-05 ***
## AGE_Btwn_45_and_55  -0.61986    0.11599  -5.344 9.08e-08 ***
## AGE_Btwn_65_and_75  -0.61134    0.45008  -1.358 0.174373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6651.7  on 5732  degrees of freedom
## Residual deviance: 5663.9  on 5707  degrees of freedom
## AIC: 5715.9
```

```
##
## Number of Fisher Scoring iterations: 5
```

FVS

```
##
## Call:
## glm(formula = TARGET_FLAG ~ MVR_PTS + HOME_VAL + CAR_USE_Commercial +
##     CLM_FREQ + Parent_Yes + JOB_Manager + CAR_TYPE_Minivan +
##     TIF + KIDSDRIV + BLUEBOOK + AGE_Btwn_45_and_55 + AGE_Btwn_35_and_45 +
##     TRAVTIME + EDUCATION_Bachelors + INCOME_High + MSTATUS_Yes +
##     CAR_TYPE_Sports_Car + OLDCLAIM + INCOME_Zero + JOB_z_Blue_Collar +
##     AGE_Btwn_65_and_75 + AGE_NA_IND + DO_KIDS_DRIVE + AGE_Btwn_21_and_25,
##     family = binomial(), data = insurance.train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1206  -0.7633  -0.5146   0.7955   2.7334
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.77022    0.13917  -5.534 3.12e-08 ***
## MVR_PTS              1.71453    0.20264   8.461  < 2e-16 ***
## HOME_VAL            -1.34405    0.31923  -4.210 2.55e-05 ***
## CAR_USE_Commercial   0.59417    0.07982   7.444 9.80e-14 ***
## CLM_FREQ             1.08698    0.15759   6.897 5.30e-12 ***
## Parent_Yes           0.25474    0.11016   2.312 0.020755 *
## JOB_Manager         -0.58568    0.12864  -4.553 5.29e-06 ***
## CAR_TYPE_Minivan    -0.50954    0.08678  -5.872 4.31e-09 ***
## TIF                 -1.18078    0.19964  -5.915 3.33e-09 ***
## KIDSDRIV             0.15948    0.13254   1.203 0.228879
## BLUEBOOK            -0.93246    0.29957  -3.113 0.001854 **
## AGE_Btwn_45_and_55  -0.47516    0.08860  -5.363 8.18e-08 ***
## AGE_Btwn_35_and_45  -0.34133    0.08771  -3.892 9.96e-05 ***
## TRAVTIME             1.03426    0.28130   3.677 0.000236 ***
## EDUCATION_Bachelors -0.27530    0.07747  -3.554 0.000380 ***
## INCOME_High         -0.32487    0.09332  -3.481 0.000499 ***
## MSTATUS_Yes         -0.39558    0.08898  -4.446 8.77e-06 ***
## CAR_TYPE_Sports_Car  0.26808    0.10392   2.580 0.009885 **
## OLDCLAIM             0.51197    0.21938   2.334 0.019609 *
## INCOME_Zero          0.33553    0.11640   2.883 0.003943 **
## JOB_z_Blue_Collar    0.26021    0.08892   2.926 0.003430 **
## AGE_Btwn_65_and_75  -0.79329    0.43792  -1.811 0.070066 .
## AGE_NA_IND           1.83117    1.10754   1.653 0.098258 .
## DO_KIDS_DRIVE        0.40556    0.21698   1.869 0.061608 .
## AGE_Btwn_21_and_25   0.58452    0.38046   1.536 0.124458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6651.7  on 5732  degrees of freedom
## Residual deviance: 5665.8  on 5708  degrees of freedom
## AIC: 5715.8
```

```
##
## Number of Fisher Scoring iterations: 5
```

SVS

```
##
## Call:
## glm(formula = TARGET_FLAG ~ MVR_PTS + HOME_VAL + CAR_USE_Commercial +
##     CLM_FREQ + Parent_Yes + JOB_Manager + CAR_TYPE_Minivan +
##     TIF + BLUEBOOK + AGE_Btwn_45_and_55 + AGE_Btwn_35_and_45 +
##     TRAVTIME + EDUCATION_Bachelors + INCOME_High + MSTATUS_Yes +
##     CAR_TYPE_Sports_Car + OLDCLAIM + INCOME_Zero + JOB_z_Blue_Collar +
##     AGE_Btwn_65_and_75 + AGE_NA_IND + DO_KIDS_DRIVE + AGE_Btwn_21_and_25,
##     family = binomial(), data = insurance.train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1225  -0.7637  -0.5159   0.8075   2.7351
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.77172    0.13914  -5.546 2.92e-08 ***
## MVR_PTS              1.72122    0.20261   8.495  < 2e-16 ***
## HOME_VAL            -1.34378    0.31926  -4.209 2.56e-05 ***
## CAR_USE_Commercial   0.59597    0.07981   7.467 8.17e-14 ***
## CLM_FREQ             1.08632    0.15757   6.894 5.41e-12 ***
## Parent_Yes           0.25199    0.11010   2.289 0.022099 *
## JOB_Manager         -0.58213    0.12859  -4.527 5.98e-06 ***
## CAR_TYPE_Minivan    -0.50878    0.08675  -5.865 4.49e-09 ***
## TIF                 -1.18014    0.19957  -5.913 3.35e-09 ***
## BLUEBOOK            -0.93475    0.29957  -3.120 0.001807 **
## AGE_Btwn_45_and_55  -0.47548    0.08859  -5.367 7.99e-08 ***
## AGE_Btwn_35_and_45  -0.34225    0.08770  -3.902 9.52e-05 ***
## TRAVTIME             1.03690    0.28119   3.688 0.000226 ***
## EDUCATION_Bachelors -0.27808    0.07745  -3.591 0.000330 ***
## INCOME_High         -0.32489    0.09332  -3.481 0.000499 ***
## MSTATUS_Yes         -0.39595    0.08899  -4.450 8.60e-06 ***
## CAR_TYPE_Sports_Car  0.26686    0.10388   2.569 0.010197 *
## OLDCLAIM             0.51097    0.21935   2.330 0.019832 *
## INCOME_Zero          0.33800    0.11634   2.905 0.003668 **
## JOB_z_Blue_Collar    0.26310    0.08889   2.960 0.003077 **
## AGE_Btwn_65_and_75  -0.79341    0.43823  -1.810 0.070219 .
## AGE_NA_IND           1.83307    1.10755   1.655 0.097912 .
## DO_KIDS_DRIVE        0.63539    0.10205   6.226 4.78e-10 ***
## AGE_Btwn_21_and_25   0.58461    0.38056   1.536 0.124497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6651.7  on 5732  degrees of freedom
## Residual deviance: 5667.3  on 5709  degrees of freedom
## AIC: 5715.3
##
```

```
## Number of Fisher Scoring iterations: 5
```