Auto Insurance R Report

Logan Strouse

5/14/2019

Bingo Bonus

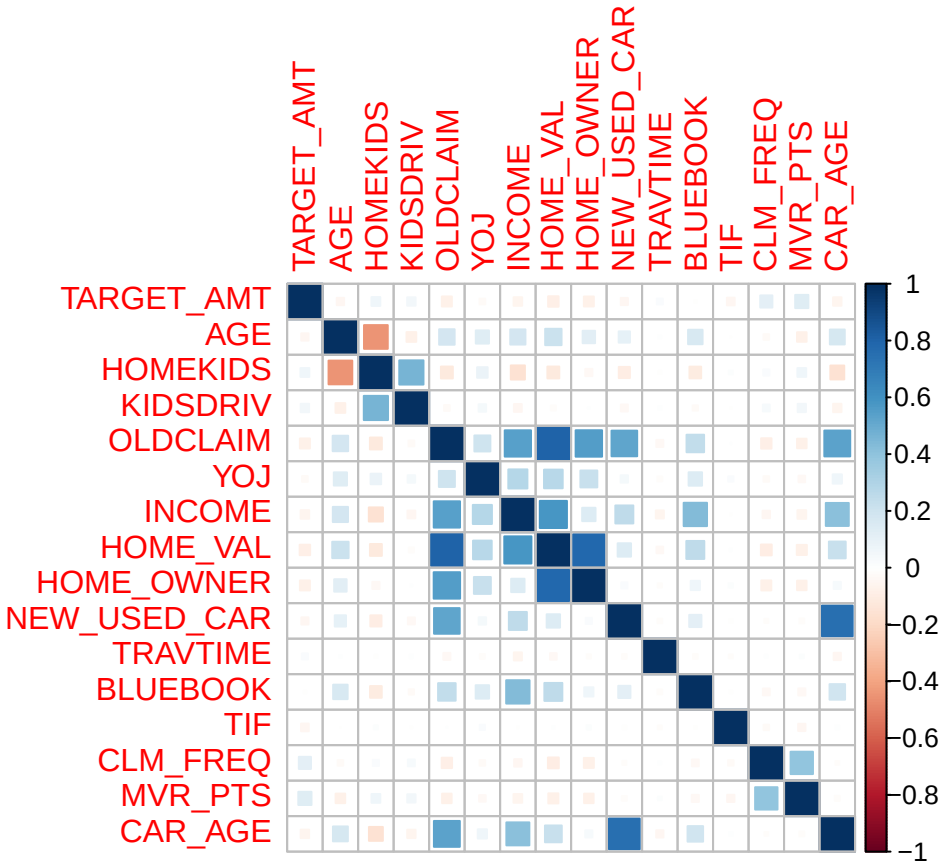
For this assignment I would like the opportunity to recieve 5-10 points for doing the writeup and execution within an RMD document. This was a good learning experience for myself due to my previous courses being in Python and not having much experience with Rstudio. This document seems to have a better functionality than Jupyter Notebooks with Python. I also used rpart decision trees to build a model, unfortunely that model wasn't chosen but the code is in the raw R file. Thanks in Advance!

Introduction

This assignment was a look at investigating different logistic regression models to predict a binary varaible value. There was also a continous variable that predicted the amount of a possible claim that also needed to be modeled. In order to successfully build a prediction model, I had to clear up missing values and also create a few new variables that could possibly aid to help increase the accuracy of the model. Multiple techniques were used to correct the missing values in both the training and testing data sets. These included manual imputation through limits as well as using auto-imputation through packages, like mice. Various statistics were also used to pick the winning model, amoung those chosen include AIC, BIC and the AUC.

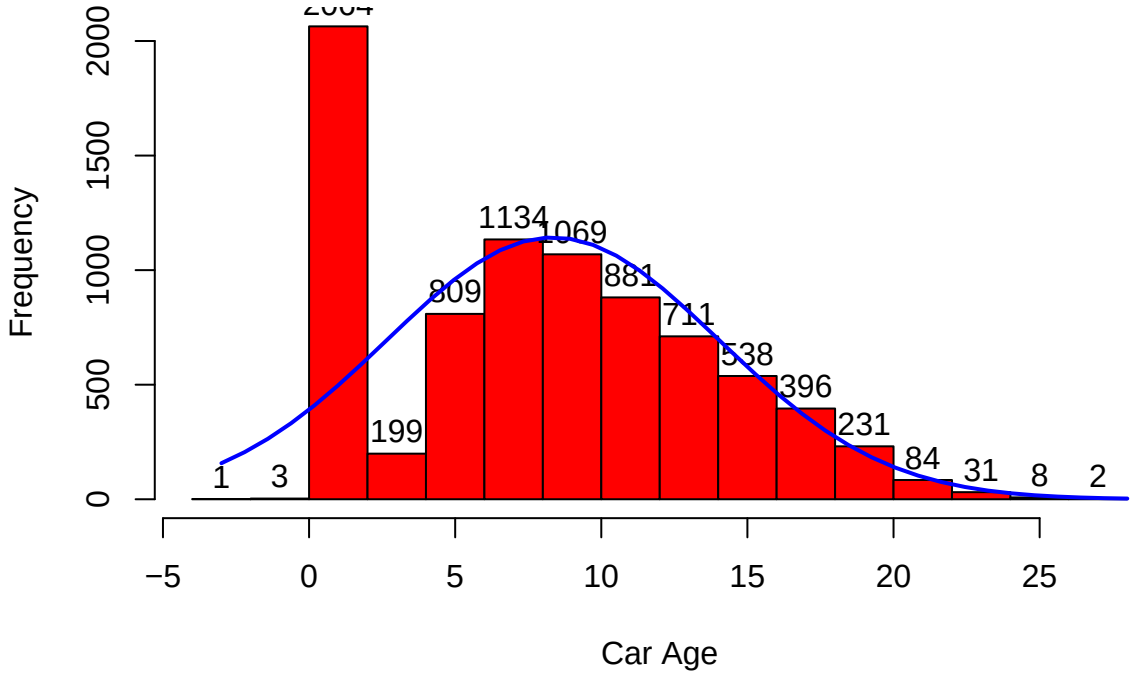
Data Exploration

Exploring the data was key to determining the relationships between the variables. It is also helpful to see where there are possible outliars and missing data. There were a total of 8,161 observations and 25 variables (not counting index) in the training data set. The test data set had 2,141 observations. Having 80% of the data available in the traiing dataset helped to develope a stronger model and the sample size was sufficient. Upon some basic charting, I was able to use the correlation plot below to see which variables had strong relationships amongst each other. I used this for part of my approach to builing the models and was mindful of some of the possible multicollinearity. Kids driving and Home kids were examples of something to be mindful of for example.



Data Preparation

The data preparation step of the project was one of the most time-intensive steps of the project. During this part of the project the data was adjusted to relect the correct type. This included telling R to look at the variables in the correct form of integer,numeric or factor. It was at this point that I Binned the income to 6 levels as well. I used the IQR range as the main reasons for breaking down the pay groups into semi-equal groups. I also created two other new variables for assessment. They were binary variables of Home_Owner and NEW_USED_CAR. I created the NEW_USED_CAR variable based upon the distribution of the car age data. Below is a plot I created with the normal curve to exemplify this. Based upon the heavily weighted 1 year age, I signified that as a prime candidate to be a new car. The thought behind this as well, was to prove that new cars might or might not be prone to more claims.



It was also during this stage that I used the mice package to fill missing values using predictive mean matching. The mice package did not work well on the Old Claim field, so I hand inputed that with the mean of the column for the missing values. Later in the project I also imputed negative values to zero on claim amounts. I did not see the idea of refunds as being applicable for this dataset. The sapply function was applied to check and confirm the values for both test and train were taken care of. An example of that is below.

```
##         INDEX   TARGET_FLAG    TARGET_AMT      KIDSDRIV           AGE
##             0             0             0             0             0
##      HOMEKIDS           YOJ        INCOME       PARENT1      HOME_VAL
##             0             0             0             0             0
##       MSTATUS           SEX     EDUCATION           JOB      TRAVTIME
##             0             0             0             0             0
##       CAR_USE      BLUEBOOK           TIF      CAR_TYPE       RED_CAR
##             0             0             0             0             0
##      OLDCLAIM      CLM_FREQ       REVOKED       MVR_PTS      CAR_AGE
##             0             0             0             0             0
##    URBANICITY DO_KIDS_DRIVE    INCOME_bin    HOME_OWNER  NEW_USED_CAR
##             0             0             0             0             0
```

Build Models

Throughout the course of the project I built multiple models. My three best are below. They included a standard full logistic regression model, backwards step selection model and a regsubsets model that selected variables based on variable numbers. I also tried some other minor variations to each model and was not able to improve thier AIC or AUC scores. Below are the three models that were eligible to be selected towards being chosen as a champion model. Overall the models all had similar AIC/BIC scores and AUC scores. The coeffients in the logistic regression model are harder to explain and with them being small numbers it was difficult to decyper the differences between the models based simply on that. They all seemed to show the same pattern, and that is why the scoring was important and mentioned in the next section.

CAR_AGE	freq
-3	1
0	3
1	2050
2	14
3	59
4	140
5	321
6	488
7	561
8	573
9	565
10	504
11	487
12	394
13	372
14	339
15	294
16	244
17	236
18	160
19	135
20	96
21	56
22	28
23	20
24	11
25	6
26	2
27	1
28	1

CAR_TYPE	freq
Minivan	2145
Panel Truck	676
Pickup	1389
Sports Car	907
Van	750
z_SUV	2294

Complete Model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + BLUEBOOK + TRAVTIME + KIDSDRIV +
##     SEX + URBANICITY + HOMEKIDS + INCOME + OLDCLAIM + DO_KIDS_DRIVE +
##     HOME_OWNER + NEW_USED_CAR + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + TIF + EDUCATION + MSTATUS + PARENT1 + RED_CAR +
##     CAR_USE + CAR_TYPE + YOJ + JOB + INCOME_bin + HOME_VAL, family = binomial(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5452  -0.7122  -0.3977   0.6139   3.1382
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.697e+00  3.670e-01  -7.349 2.00e-13 ***
## AGE                    -4.048e-03  4.097e-03  -0.988 0.323188
## BLUEBOOK               -2.057e-05  5.281e-06  -3.895 9.83e-05 ***
## TRAVTIME                1.495e-02  1.891e-03   7.906 2.65e-15 ***
## KIDSDRIV                2.013e-01  1.246e-01   1.616 0.105998
## SEXz_F                 -8.921e-02  1.122e-01  -0.795 0.426493
## URBANICITYUrban         2.411e+00  1.132e-01  21.302  < 2e-16 ***
## HOMEKIDS                3.002e-02  3.781e-02   0.794 0.427203
## INCOME                 -7.470e-07  1.907e-06  -0.392 0.695270
## OLDCLAIM                2.060e-07  5.438e-07   0.379 0.704776
## DO_KIDS_DRIVE1          3.602e-01  1.955e-01   1.842 0.065456 .
## HOME_OWNER             -3.248e-01  1.523e-01  -2.132 0.033004 *
## NEW_USED_CAR           -2.024e-01  1.185e-01  -1.708 0.087616 .
## CLM_FREQ                1.482e-01  2.560e-02   5.787 7.18e-09 ***
## REVOKEDYes              7.320e-01  8.058e-02   9.085  < 2e-16 ***
## MVR_PTS                 1.072e-01  1.362e-02   7.876 3.39e-15 ***
## CAR_AGE                 1.327e-02  1.115e-02   1.191 0.233801
## TIF                    -5.536e-02  7.367e-03  -7.514 5.74e-14 ***
## EDUCATIONBachelors     -3.753e-01  1.202e-01  -3.122 0.001796 **
## EDUCATIONMasters       -3.441e-01  1.869e-01  -1.841 0.065601 .
## EDUCATIONPhD           -2.604e-01  2.211e-01  -1.178 0.238937
## EDUCATIONz_High School  5.383e-03  9.809e-02   0.055 0.956234
## MSTATUSz_No             4.799e-01  8.889e-02   5.399 6.70e-08 ***
## PARENT1Yes              3.561e-01  1.111e-01   3.204 0.001355 **
## RED_CAR1               -1.172e-02  8.652e-02  -0.135 0.892228
## CAR_USEPrivate         -7.610e-01  9.219e-02  -8.254  < 2e-16 ***
## CAR_TYPEPanel Truck     5.708e-01  1.622e-01   3.519 0.000433 ***
## CAR_TYPEPickup          5.709e-01  1.011e-01   5.647 1.63e-08 ***
## CAR_TYPESports Car      1.019e+00  1.303e-01   7.824 5.11e-15 ***
## CAR_TYPEVan             6.324e-01  1.267e-01   4.992 5.97e-07 ***
## CAR_TYPEz_SUV           7.791e-01  1.115e-01   6.985 2.85e-12 ***
## YOJ                     1.304e-02  1.093e-02   1.194 0.232537
## JOBClerical             4.096e-01  1.984e-01   2.065 0.038967 *
## JOBDoctor              -4.106e-01  2.678e-01  -1.533 0.125224
## JOBHome Maker           1.009e-01  2.236e-01   0.451 0.651662
## JOBLawyer               1.216e-01  1.703e-01   0.714 0.475115
## JOBManager             -5.480e-01  1.717e-01  -3.191 0.001415 **
```

```
## JOBProfessional         1.719e-01  1.793e-01   0.958 0.337850
## JOBStudent             -1.821e-02  2.314e-01  -0.079 0.937287
## JOBz_Blue Collar        3.126e-01  1.865e-01   1.676 0.093758 .
## INCOME_binLow          -6.766e-01  1.753e-01  -3.859 0.000114 ***
## INCOME_binMedium       -7.383e-01  2.009e-01  -3.675 0.000238 ***
## INCOME_binHigh         -7.934e-01  2.252e-01  -3.523 0.000427 ***
## INCOME_binAffluent     -1.171e+00  2.821e-01  -4.151 3.31e-05 ***
## HOME_VAL               -3.054e-07  8.045e-07  -0.380 0.704195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7272.2  on 8116  degrees of freedom
## AIC: 7362.2
##
## Number of Fisher Scoring iterations: 5

## Complete Model AIC: 7362.198

## Complete Model BIC: 7677.518
```

Backwards Stepwise Model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ BLUEBOOK + TRAVTIME + KIDSDRIV +
##     URBANICITY + DO_KIDS_DRIVE + HOME_OWNER + CLM_FREQ + REVOKED +
##     MVR_PTS + TIF + EDUCATION + MSTATUS + PARENT1 + CAR_USE +
##     CAR_TYPE + JOB + INCOME_bin, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5408  -0.7130  -0.3999   0.6101   3.1316
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.887e+00  3.084e-01  -9.359  < 2e-16 ***
## BLUEBOOK               -2.316e-05  4.711e-06  -4.915 8.88e-07 ***
## TRAVTIME                1.482e-02  1.888e-03   7.849 4.20e-15 ***
## KIDSDRIV                2.197e-01  1.220e-01   1.801 0.071678 .
## URBANICITYUrban         2.416e+00  1.132e-01  21.341  < 2e-16 ***
## DO_KIDS_DRIVE1          3.699e-01  1.953e-01   1.894 0.058249 .
## HOME_OWNER             -3.606e-01  8.458e-02  -4.264 2.01e-05 ***
## CLM_FREQ                1.482e-01  2.556e-02   5.797 6.73e-09 ***
## REVOKEDYes              7.356e-01  8.045e-02   9.144  < 2e-16 ***
## MVR_PTS                 1.079e-01  1.359e-02   7.938 2.05e-15 ***
## TIF                    -5.474e-02  7.354e-03  -7.444 9.78e-14 ***
## EDUCATIONBachelors     -3.854e-01  1.128e-01  -3.417 0.000634 ***
## EDUCATIONMasters       -3.010e-01  1.636e-01  -1.840 0.065701 .
## EDUCATIONPhD           -2.530e-01  1.979e-01  -1.279 0.201068
## EDUCATIONz_High School  9.894e-03  9.742e-02   0.102 0.919102
## MSTATUSz_No             4.382e-01  8.447e-02   5.188 2.13e-07 ***
## PARENT1Yes              4.378e-01  9.574e-02   4.573 4.82e-06 ***
```

```
## CAR_USEPrivate         -7.579e-01  9.205e-02  -8.234  < 2e-16 ***
## CAR_TYPEPanel Truck     6.155e-01  1.511e-01   4.074 4.61e-05 ***
## CAR_TYPEPickup          5.664e-01  1.009e-01   5.613 1.99e-08 ***
## CAR_TYPESports Car      9.523e-01  1.079e-01   8.828  < 2e-16 ***
## CAR_TYPEVan             6.553e-01  1.222e-01   5.361 8.26e-08 ***
## CAR_TYPEz_SUV           7.156e-01  8.621e-02   8.300  < 2e-16 ***
## JOBClerical             4.326e-01  1.972e-01   2.194 0.028222 *
## JOBDoctor              -3.989e-01  2.658e-01  -1.501 0.133407
## JOBHome Maker           9.567e-02  2.195e-01   0.436 0.662995
## JOBLawyer               1.219e-01  1.690e-01   0.721 0.470749
## JOBManager             -5.493e-01  1.709e-01  -3.214 0.001308 **
## JOBProfessional         1.772e-01  1.781e-01   0.995 0.319596
## JOBStudent              8.976e-04  2.269e-01   0.004 0.996843
## JOBz_Blue Collar        3.233e-01  1.855e-01   1.743 0.081370 .
## INCOME_binLow          -5.556e-01  1.355e-01  -4.099 4.15e-05 ***
## INCOME_binMedium       -6.379e-01  1.586e-01  -4.023 5.76e-05 ***
## INCOME_binHigh         -7.167e-01  1.683e-01  -4.258 2.06e-05 ***
## INCOME_binAffluent     -1.133e+00  1.782e-01  -6.360 2.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7280.5  on 8126  degrees of freedom
## AIC: 7350.5
##
## Number of Fisher Scoring iterations: 5

## Backwards Model AIC: 7350.451

## Backwards Model BIC: 7595.701
```

Reg Subsets Model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ URBANICITY + INCOME + JOB + MSTATUS +
##     TIF + DO_KIDS_DRIVE + TRAVTIME + BLUEBOOK + REVOKED + MVR_PTS +
##     CAR_USE, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3753  -0.7376  -0.4265   0.6917   3.0781
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.969e+00  2.158e-01 -13.757  < 2e-16 ***
## URBANICITYUrban   2.392e+00  1.097e-01  21.812  < 2e-16 ***
## INCOME           -6.295e-06  9.698e-07  -6.491 8.51e-11 ***
## JOBClerical       4.476e-01  1.553e-01   2.882  0.00395 **
## JOBDoctor        -4.387e-01  2.441e-01  -1.797  0.07226 .
## JOBHome Maker     3.434e-01  1.821e-01   1.886  0.05931 .
## JOBLawyer        -1.073e-01  1.596e-01  -0.673  0.50126
## JOBManager       -6.486e-01  1.531e-01  -4.238 2.26e-05 ***
```

```
## JOBProfessional  -2.324e-02  1.435e-01  -0.162  0.87136
## JOBStudent        3.643e-01  1.735e-01   2.100  0.03570 *
## JOBz_Blue Collar  2.752e-01  1.321e-01   2.082  0.03730 *
## MSTATUSz_No       7.788e-01  5.821e-02  13.379  < 2e-16 ***
## TIF              -5.437e-02  7.202e-03  -7.549 4.37e-14 ***
## DO_KIDS_DRIVE1    8.048e-01  8.271e-02   9.730  < 2e-16 ***
## TRAVTIME          1.406e-02  1.842e-03   7.634 2.28e-14 ***
## BLUEBOOK         -2.787e-05  4.000e-06  -6.967 3.23e-12 ***
## REVOKEDYes        7.524e-01  7.885e-02   9.541  < 2e-16 ***
## MVR_PTS           1.478e-01  1.253e-02  11.801  < 2e-16 ***
## CAR_USEPrivate   -8.012e-01  7.575e-02 -10.577  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7519.3  on 8142  degrees of freedom
## AIC: 7557.3
##
## Number of Fisher Scoring iterations: 5

## Reg Subsets Model AIC: 7557.283

## Reg Subsets Model BIC: 7690.418
```

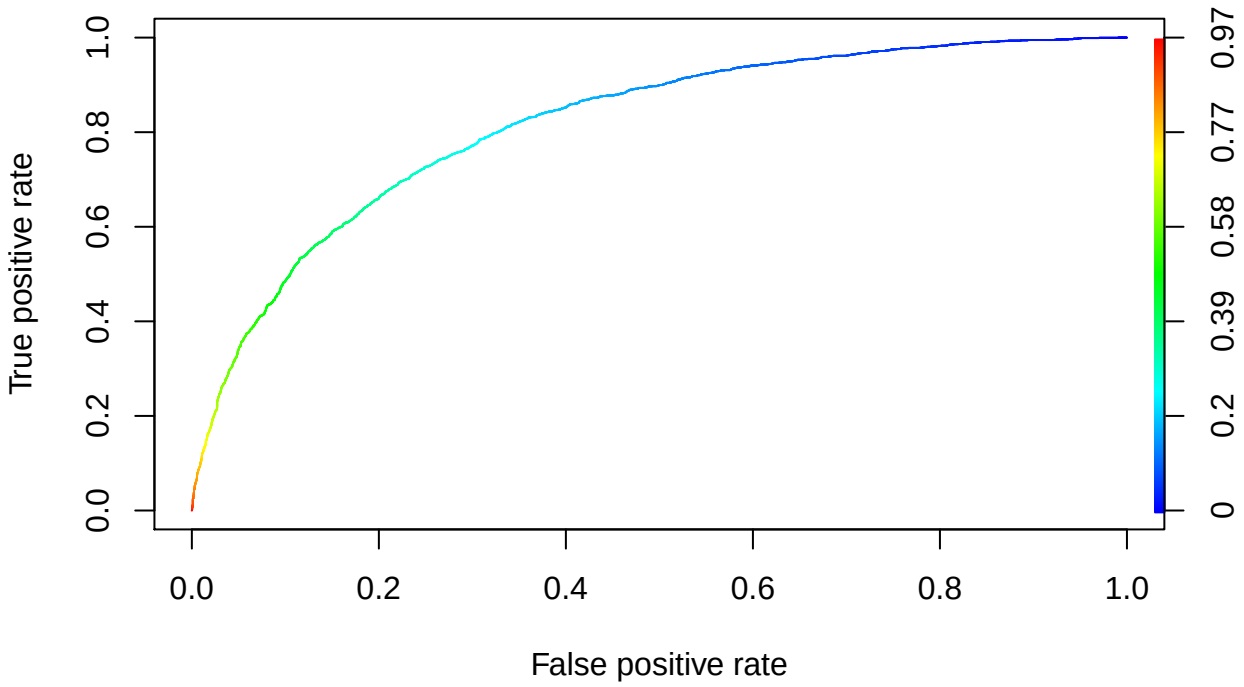
Select Model

The winning model that had the best AIC with the most area under the curve was the backwards stepwise model. The AUC was .8150246 and the AIC was 7347.265. It is shown below along with the other two models that were shown prior with their AUC charts KS statistics. I decided based on my readings and prior work that AIC and the AUC statistic were the most important, since MSE is not great for logistic models. For reference in the output below, the y.values are the AUC scores and the ks statistic is the last number mentioned with each model that begins with a .47.

Backwards Stepwise Model

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.8147124
##
##
```

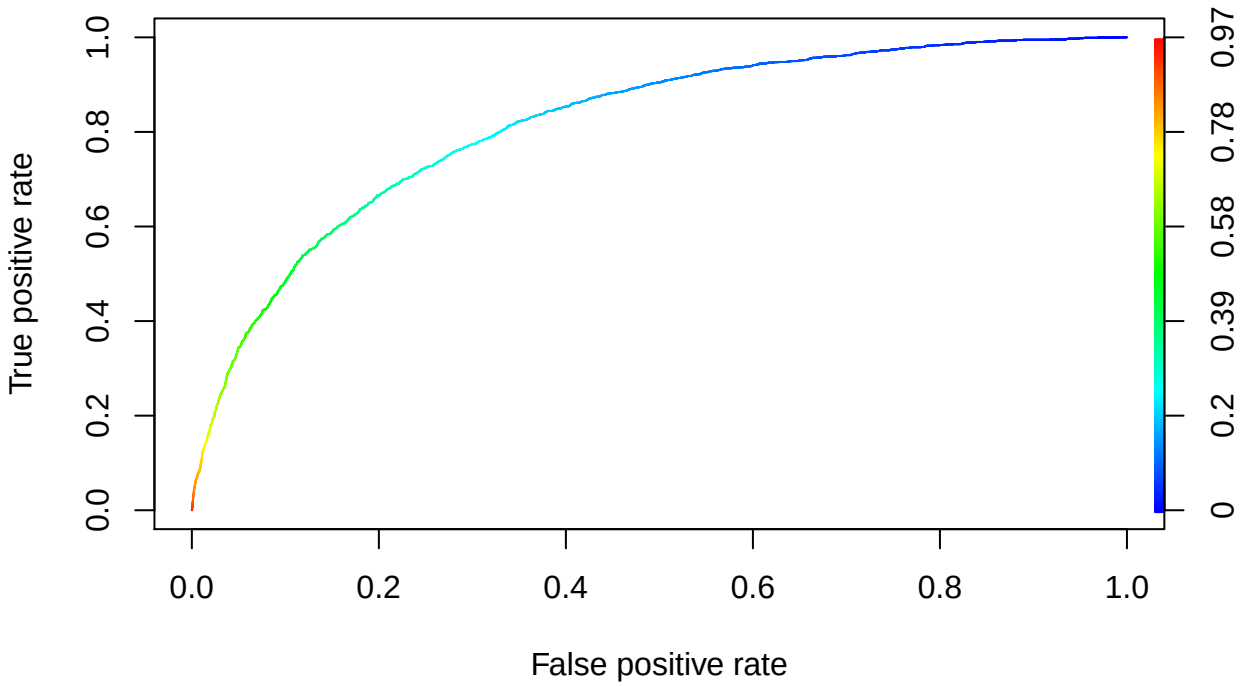
```
## Slot "alpha.values":
## list()
```




```
## [1] 0.4744
```

Complete Model

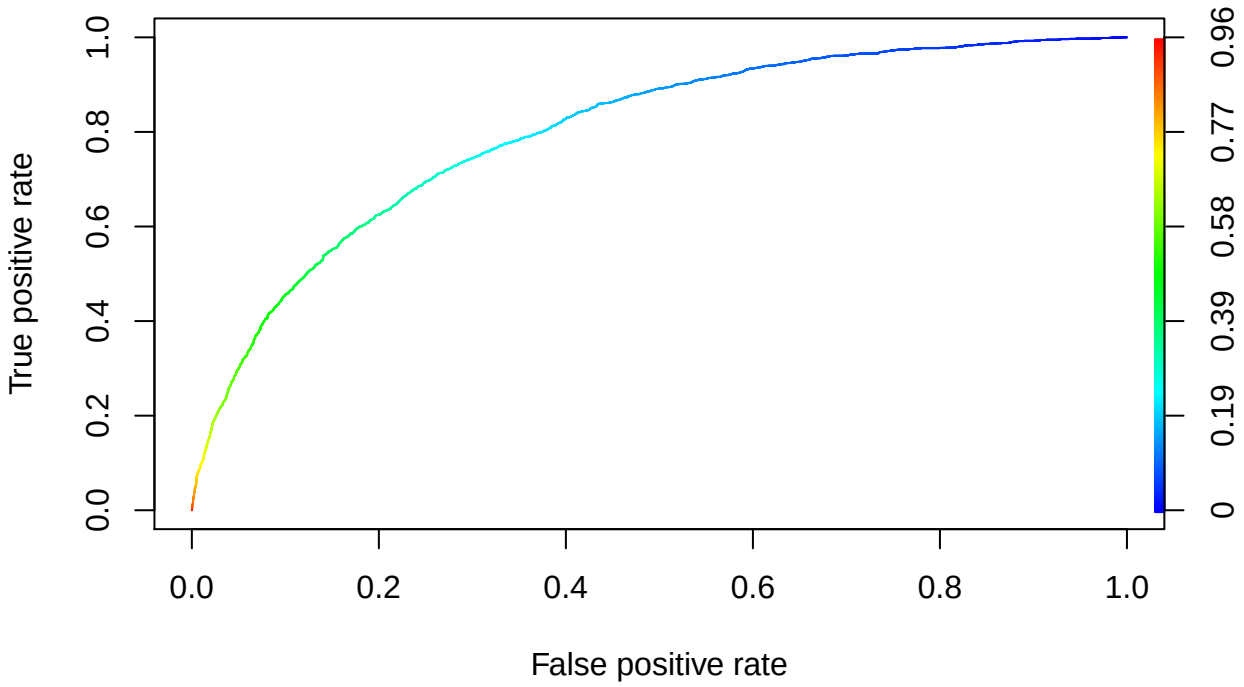
```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.8153596
##
##
## Slot "alpha.values":
## list()
```




```
## [1] 0.4763
```

Reg Subsets Model

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.7981166
##
##
## Slot "alpha.values":
## list()
```




```
## [1] 0.4466
```

Stand Alone Scoring Program

P_Target_AMT Scoring Program (includes correction for negative values) test$P_TARGET_AMT <- predict(subset_amt_model, newdata = test, type = "response")test$P_TARGET_AMT[(test$P_TARGET_AMT <= 0)] = 0

P_Target_Flag Scoring Program test$P_TARGET_FLAG <- predict(backwards_model, newdata = test, type = "response")

The formulas for both of these are derived off of the Model Coefficients in the summaries for both model's above.

Scored Data Set

This was submitted separately.

Conclusion

Overall, this assignment was a great look at the aspects that can be used to predict risk and potential outcome of events. I believe that a similar model could be deployed in many other industries like the credit industry or even potential employement opportunites. Unfortunetly, there are some ethical issues on the later option. There is other data that might have helped with this data set. If we could get our hands on some IOT data that contains speed at time of crash, brake use and other telemetrics we might be able to refine the accuracy of these models even further.