Summary

Blain Morin

May 12, 2018

First, I examined the training data:

```
Predict_NoShow_Train = read_csv("Predict_NoShow_Train.csv")
summary(Predict_NoShow_Train)
```

```
##        ID              Age            Gender         
##  Min.   :100003   Min.   :  0.00   Length:180000     
##  1st Qu.:324419   1st Qu.: 19.00   Class :character  
##  Median :548838   Median : 38.00   Mode  :character  
##  Mean   :549331   Mean   : 37.78                     
##  3rd Qu.:774096   3rd Qu.: 56.00                     
##  Max.   :999990   Max.   :113.00                     
##  DateAppointmentWasMade DateOfAppointment    DaysUntilAppointment
##  Min.   :2013-05-29     Min.   :2014-01-02   Min.   :  1.00      
##  1st Qu.:2014-04-09     1st Qu.:2014-04-24   1st Qu.:  3.00      
##  Median :2014-07-28     Median :2014-08-08   Median :  8.00      
##  Mean   :2014-07-23     Mean   :2014-08-05   Mean   : 13.17      
##  3rd Qu.:2014-10-30     3rd Qu.:2014-11-13   3rd Qu.: 19.00      
##  Max.   :2015-03-10     Max.   :2015-03-11   Max.   :398.00      
##     Diabetes         Alcoholism       Hypertension     Handicapped     
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000  
##  Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000  
##  Mean   :0.08062   Mean   :0.02533   Mean   :0.2228   Mean   :0.02114  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :4.00000  
##      Smoker         Scholarship       Tuberculosis       RemindedViaSMS  
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.00000   Median :0.0000000   Median :1.0000  
##  Mean   :0.05346   Mean   :0.09829   Mean   :0.0004778   Mean   :0.5281  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000000   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000000   Max.   :2.0000  
##  DayOfTheWeek          Status         
##  Length:180000      Length:180000     
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
## 
```

The summary allows us to identify the factor and continuous variables and also sheds light on the prevalance of certain conditions (which will come in handy later when we decide which variables should be used as predictors). We know right away that we can drop ID, the date the appointment was made, and the date of the appointment because these likely will not have predictive power.

Next, I cleaned the data by identifying factors, centering and scaling variables, and creating a model matrix.I also examined how many patients were in each factor bin, and eliminated the factors where there was a large discrepency.

I then ran a basic logistic regression using all the variables. Although its prediction accuracy was fairly good, the loss was greater than 1 (which was far behind the public leaderboard). I decided to try 4 algorithms that could improve the logistic loss: GBM, SVM, Neural Networks, and GAM. For each, I used the caret package to help tune the parameters. GBM gave the best logistic loss value. Using the results from GBM, I eliminated the predictors with low influence and reran the model. This edited GBM run is what I used in my predictNoshow function. Overall, my best logistic loss on the public leader board was .60875.