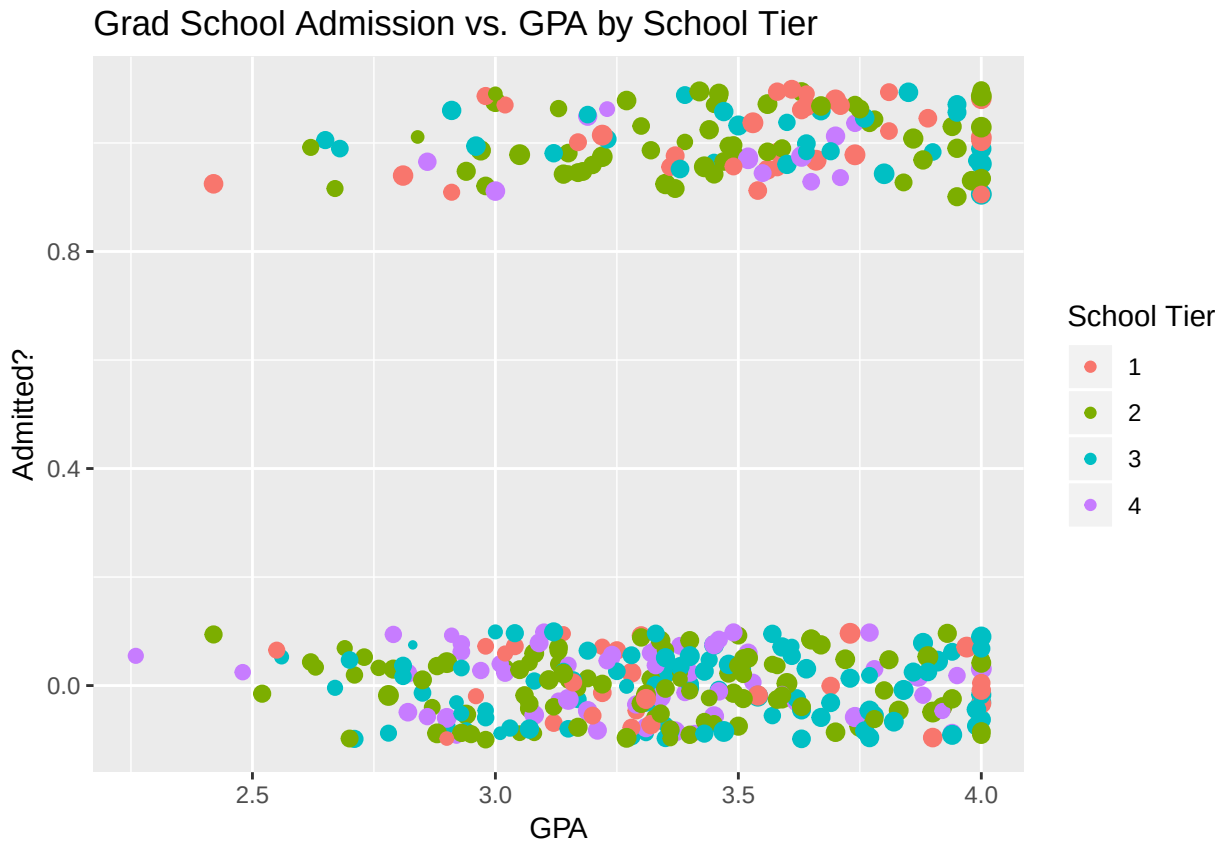
Final Project

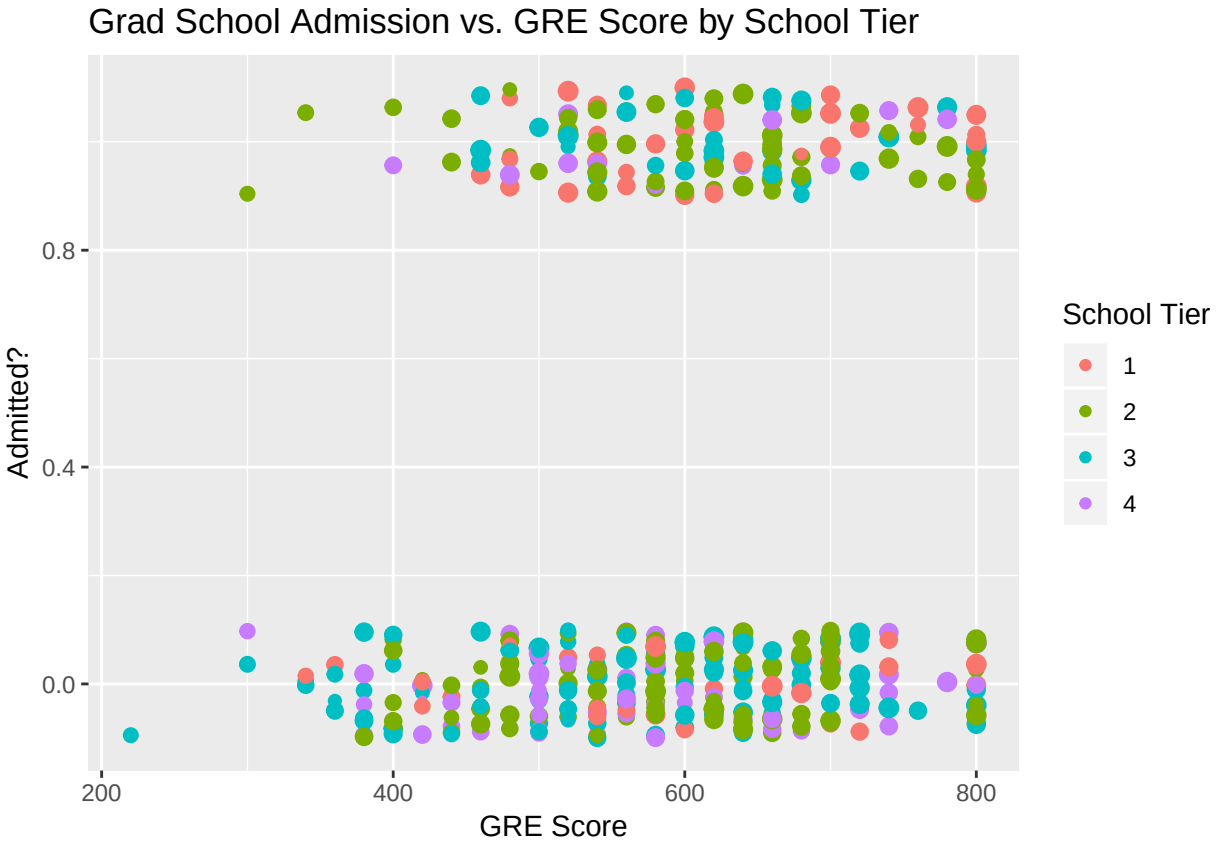
Francisco Robles

April 27, 2019

The topic dataset that I chose was a Grad School Admissions data set that had each student's GRE Score, their college GPA, and the School Tier that they went to (1 being the best and 4 being the worse). The data was collected from 400 students. I'm not sure how the data was collected, as I had this data set from a previous class that I found and it was just given to us. The question that I want to answer is can we use someone's GRE Score, GPA, and School Tier to predict whether or not they got accepted into grad school. I will be using a logistic regression model to see if we can come up with a way to try and precit the admission to grad school or not.

Lets try and see if we can visualize this data well.

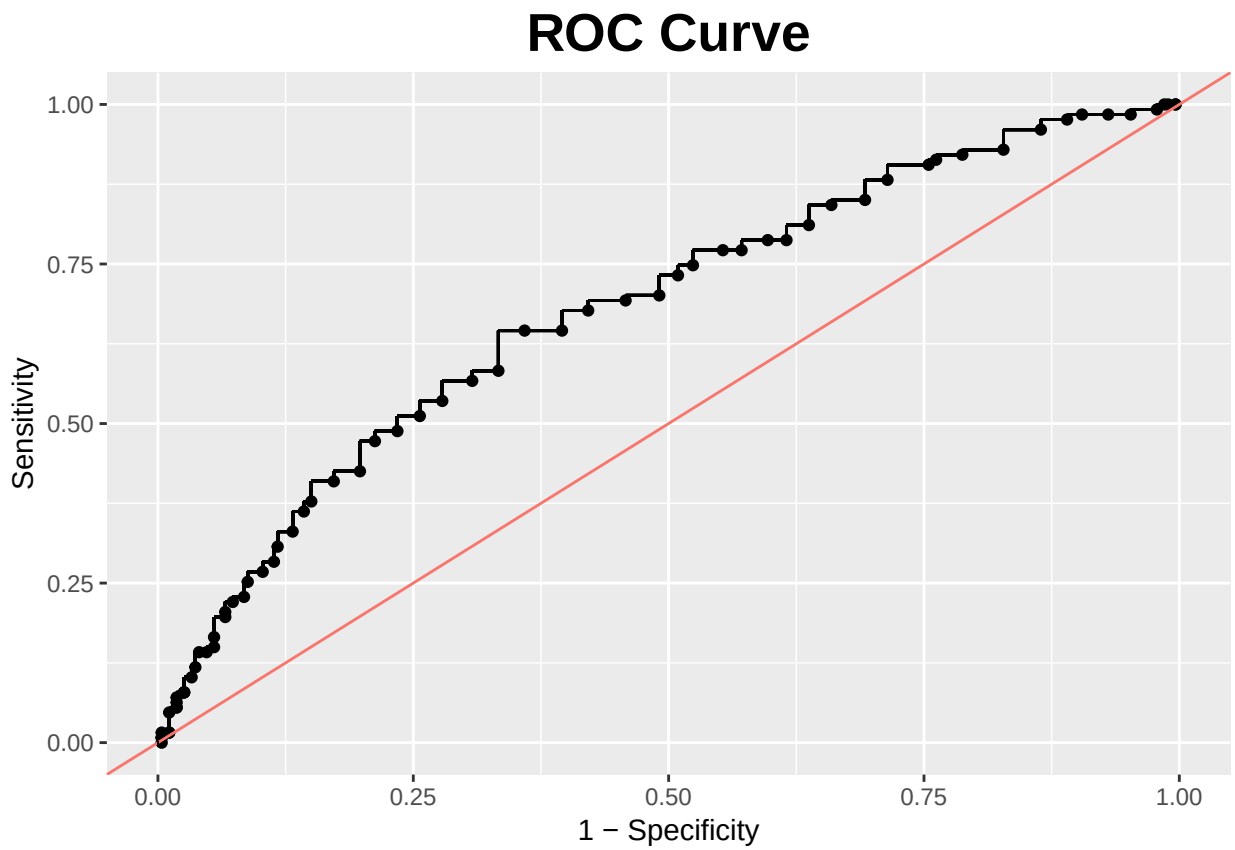




It's tough to see just visually how the different variables effect whether you get into grad school or not. What would be the best is if we can see almost a clear point where they went from not being accepted (0 for admit) to being admitted (1 for admit). It does just seem to show that the higher GRE and higher GPA seem to have a better chance at getting in, which makes sense.

Lets run our first model and see how well it does

```
grad_mod = fun.LogReg(df, 'admit', c('gre', 'gpa', 'rank'), .5)
grad_mod$Plots$ROC
```



So for an ROC plot, we are looking for a place where we can see a long vertical line, and we want to choose the point at the top of that vertical line. This one isn't super clear, but the point we want to choose is where the point is at .64 Sensitivity and .65 for Specificity. The .64 Sensitivity tells us that we can identify the people who did get into grad school 64% of the time, and the .65 for Specificity tells us that we can identify the people that got rejected from grad school 65% of the time. This is better than the 50/50 of us just randomly guessing.

We are then going to use that point that we just identified to come up with a cut-off point for our probability. It may seem like you would just say if our model says that you have a 50% chance of gettting in to grad school, we will predict that you do, but that isn't always the best case. That point actually corresponds to a cut-off of .33, meaning that if our model predicts you have over a 33% chance of being admitted to grad school, we will predict that you do make it in.

```
grad_mod2 = fun.LogReg(df, 'admit', c('gre', 'gpa', 'rank'), .33)
```

This model is now using the .33 cut-off. Lets see how our model ended up turning out

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5802  -0.8848  -0.6382   1.1575   2.1732
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.449548   1.132846  -3.045  0.00233 **
```

```
## gre          0.002294   0.001092   2.101  0.03564 *
## gpa          0.777014   0.327484   2.373  0.01766 *
## rank        -0.560031   0.127137  -4.405 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 459.44  on 396  degrees of freedom
## AIC: 467.44
##
## Number of Fisher Scoring iterations: 4
```

The main thing we need to focus on is the 'Estimate' column. Since each of our rows has at least 1 asterik next to it, it means that each of our predictors (GRE Score, GPA, and School Tier) are important in trying to predict whether students get admitted. If you look at the 'Estimate' column, since GRE and GPA are both positive numbers, it means that people are more likely to get admitted with higher GRE or higher GPA, which makes sense. Since the rank estimate is negative, it means that studnet from lower tier schools have a tougher time getting admitted to grad school.

Let's take a look at how well our model predicted the 400 students from our data set.

```
##
##       Pred No Pred Yes
##   No      175       98
##   Yes      45       82
```

We mainly want to be on that diagonal line from top left to bottom right (each one on the diagonal means that we predicted correctly). The ones that are not on the diagonal line are ones we didn't predict right.

```
##    Accuracy Sensitivity Specificity
##   0.6425000   0.6456693   0.6410256
```

This shows us our Sensitivity and Specificity, both of which we talked about above. It also shows us our accuracy as well (how often we predicted correctly).

To get a better idea about how our model can predict, we can try a couple of examples. The first example we will look at is the worst case scenario for getting in. I will use getting a GRE score of 220 and a GPA of 2.26, since those were the minimums from our data, and have them come from a very low tier school.

```
##            [,1]
## [1,] 0.03140279
```

This means that the lowest probability that this model will give us (with still using observations with our data), is a 3% chance of getting in.

The best case scenario would be getting a GRE score of 800, a GPA of 4.0 , and coming from a very high tier school.

```
##           [,1]
## [1,] 0.7178136
```

This means that the best that this model can predict is that you have a 72% of getting admitted to grad school.

To sum up, I think this model can work really well in trying to predict if a student will get in to grad school or not. It seems that all three of these variables really do matter and how much they do end up mattering. You can also see things like that 1 point of your GPA is worth more than going to a school tier lower (A 3.0

at a Tier 1 School has a less of a chance of getting accepted than a 4.0 at a Tier 2 School) or that 1 point of GPA is worth about as much as 340 points on your GRE.