STAT474
Assignment 4

Forecasting Admissions to "Elite" University with Random Forest Classifier

James Wang
04/13/16

Problem Statement

A high school guidance counselor's primary functions include advising students on his/her academic trajectory, including assisting with the college admissions process. If a particular high school was interested in predicting which of its students would gain admittance to a particular "Elite University", it could analyze its students' historical admissions decision data along with its students' attributes.

When it comes to prediction, there is no tool that performs consistently better than the random forest algorithm. However, the superiority that random forest offers with its predictive power comes at a cost: it loses the interpretability of a decision tree, which is why it is referred to as a black-box algorithm. Thus, a trained random forest classifier will not be able to tell us how exactly our predictors are linked to the response variable.

Nonetheless, useful plots generated from the random forest classifier, such as the variable importance plot and the partial dependence plot(s) can clue us in on which predictors are the most important in terms of forecasting accuracy, and the average relationship between a given predictor and the response variable with all other predictors held constant, respectively.

This means that we can reasonably expect our following analysis with the random forest classifier to yield us: (1) a powerful predictive model, and (2) limited insight on the importance and effectiveness of the predictors according to the random forest classifier.

The purpose of building this predictive tool is to assist the guidance counselor in making better informed recommendations to her students as to what their chances are of getting into "Elite University". This can help many students mold their college admissions expectations closer to their true potential.

In addition, important predictors, that seem to have the most effect on Elite University's decision-making process when it comes to the kinds of students they admit, can be identified. Furthermore, given all other predictors held constant, the relationship between a particular predictor and the response variable (whether Elite University admits or rejects) can be studied to gain more insight on what kinds of attribute values need to be met by students in order to have the best chance of getting admitted to Elite University.

Data Description

The dataset contains 8700 instances of profiles of students. There are 8 predictors and 1 response variable. There are 829 profiles missing a value for the one of the race attributes. 1724 profiles are missing values for income. Additionally, about 24 values are missing from the sex attribute. It is unclear whether or not these missing values across predictors overlap, and more importantly, whether or not these missing values will influence our analysis. We will address this problem in due time.

The observations were taken from Elite University's applicant pool from the previous year. Almost immediately this violates one of our assumptions from above, since the applicant pool to Elite University includes students outside of this one particular high school. We can either relax our assumption and say that every applicant's high school education system/environment is identical to our particular high school's education system (which is highly unrealistic), or we can proceed with caution that we may be comparing apples to oranges in terms of probability of a student with a particular profile from this particular high school to be admitted into Elite University versus a student with a particular profile from another high

school to be admitted into Elite University. This will make a Level II Analysis hard to justify, and potentially the results from Level II Analysis unreliable.

Level I Analysis

Univariate Statistics…

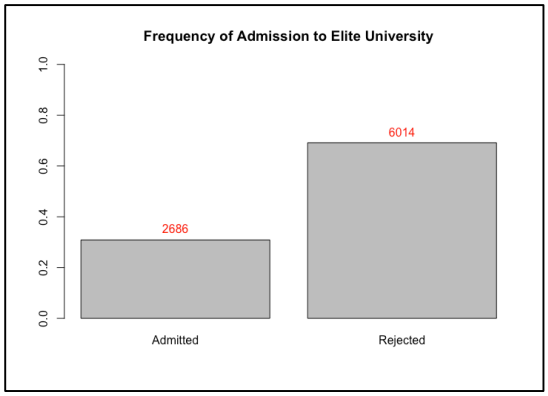

Figure 1.

From *Figure* 1, we see that almost 31% of the profiles in the dataset were admitted to Elite University. It does not matter whether or not this actually reflects Elite University's admissions rate (note: elite universities typically have admissions rates less than 20%) since we are not predicting admissions rates. In fact, it is probably good that we have more "Admitted" profiles in our dataset, since it will help the random forest algorithm identify classify "Admitted" applicants more accurately.

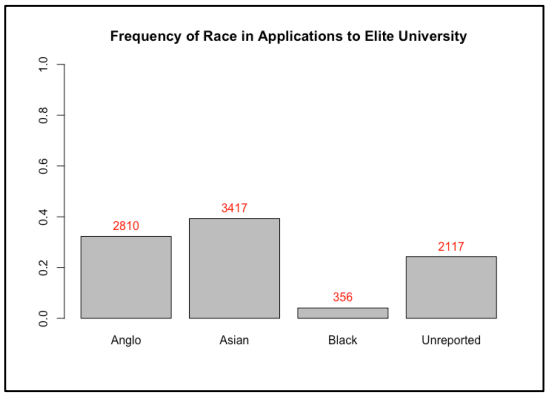

Figure 2.

From *Figure 2*, we can see that the majority of the applicants to Elite University seemed to be identify as Asian. Many applicants in the dataset are missing their race. These missing values may pose a problem when it comes to performing our analysis later on.

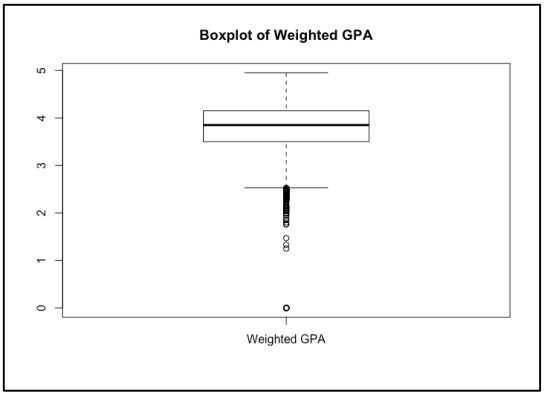

Figure 3.

Figure 3 seems to indicate that the median weighted GPA was just shy of 4.0. Taking AP classes allows students to reach a maximum of 5.0 weighted GPA. There seem to be a few outliers with value of 0, which is most likely incorrect. A GPA of 0 would mean a student is failing all their classes, and therefore would probably not apply to Elite University in the first place. These outliers with value of 0 could be counted as missing values and will be treated later.

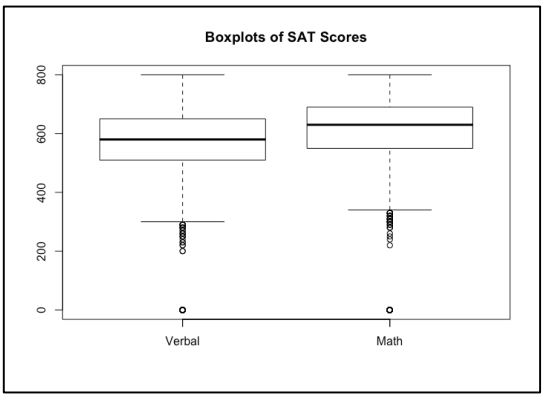

Figure 4.

Each SAT section has a score ranging from 200 to 800. The median score for both sections seems to sit around 600. Again, the 0 values are most likely errors (since the lowest possibly SAT score for a particular section is 200), so they will be treated as missing values.

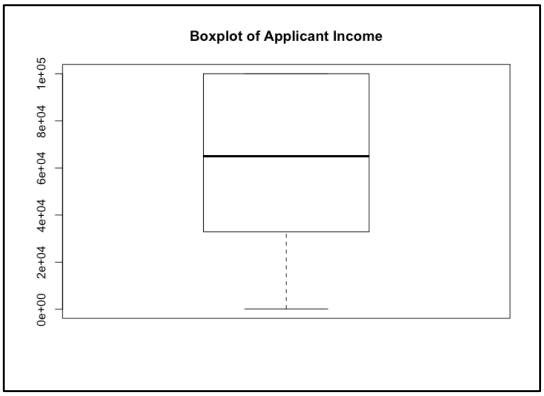

Figure 5.

The top 25% of the applicant pool reported an income of $100,000. The median income was $65,000. There is no income data for 1754 of the applicants.

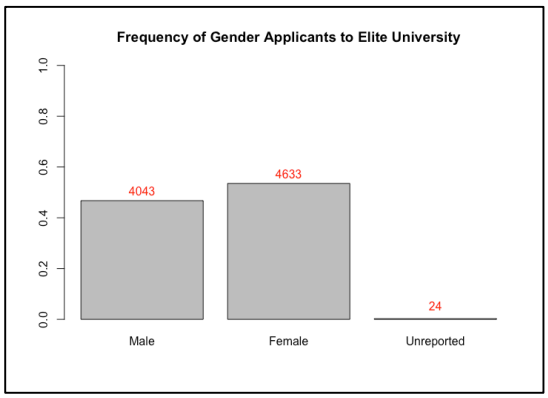

Figure 6.

The applicants to the university are fairly even, with slightly more females (4633) than males (4043).

Bivariate Statistics…

Now let's analyze how these attributes vary between those who were admitted and those who were rejected.

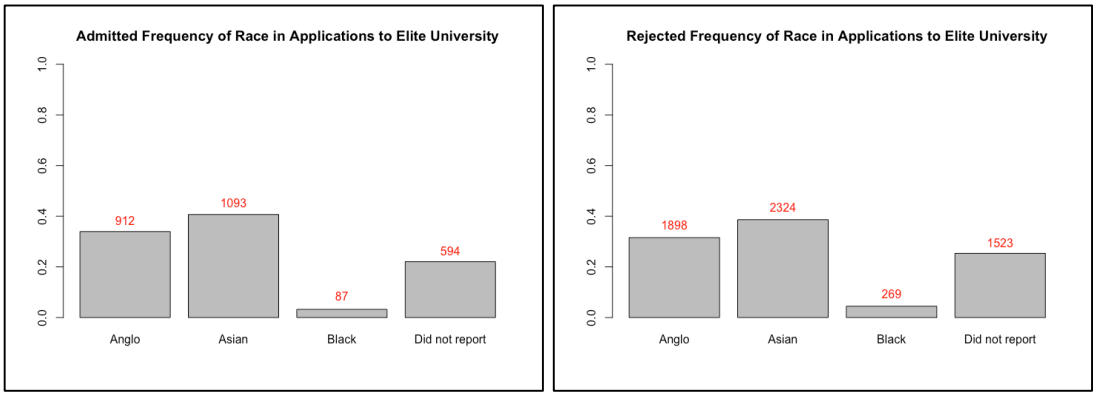

Figure 7a & 7b.

The relative frequencies do not seem to change when we subset by race. The admitted pool and rejected pool have almost the same racial proportions. The number of rejected applicants is at least twice as large as the number of accepted applicants across each race.

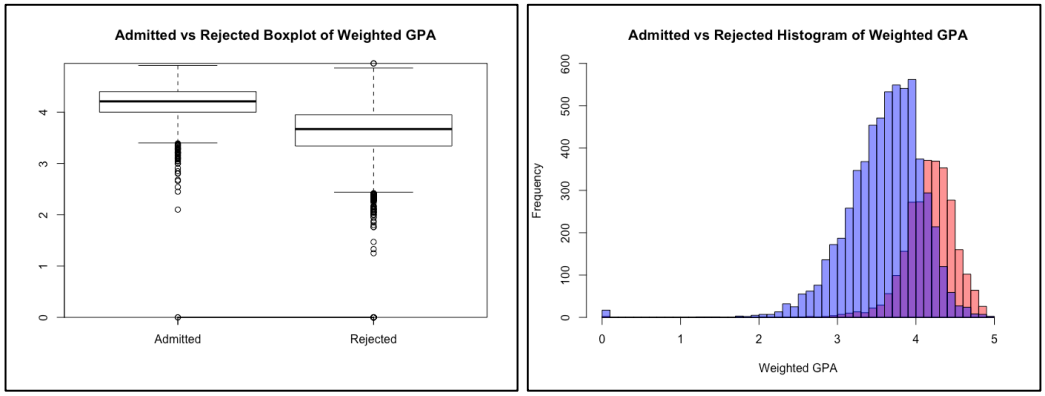

Figure 8a & 8b.

As expected, admitted applicants have a higher weighted GPA across every summary statistic compared to rejected applicants. Surprisingly, most of the top 25th percentile of rejected applicants had a weighted GPA above 4.0. Figure 8b gives a visual on the overlap between rejected and admitted applicants weighted GPAs. This is interesting because there are many rejected students who have just as high of a GPA as an admitted student. This indicates that there are other variables at play that influence Elite University's admissions decision-making.

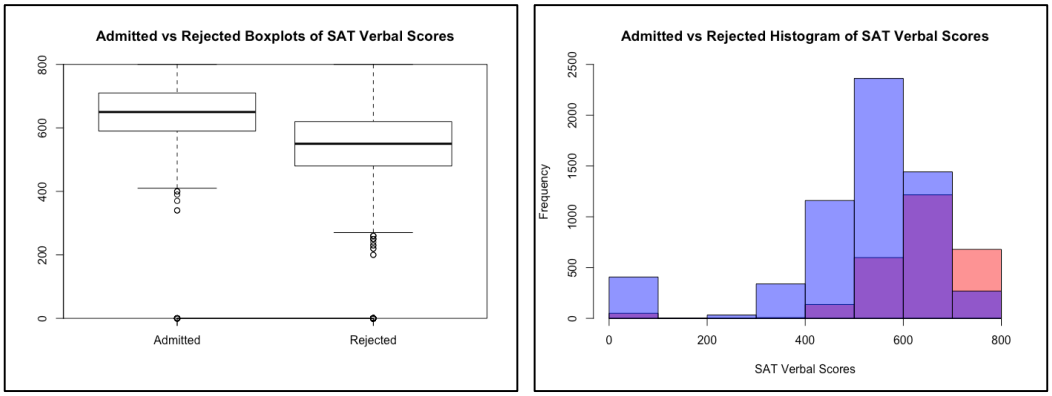

Figure 9a & 9b.

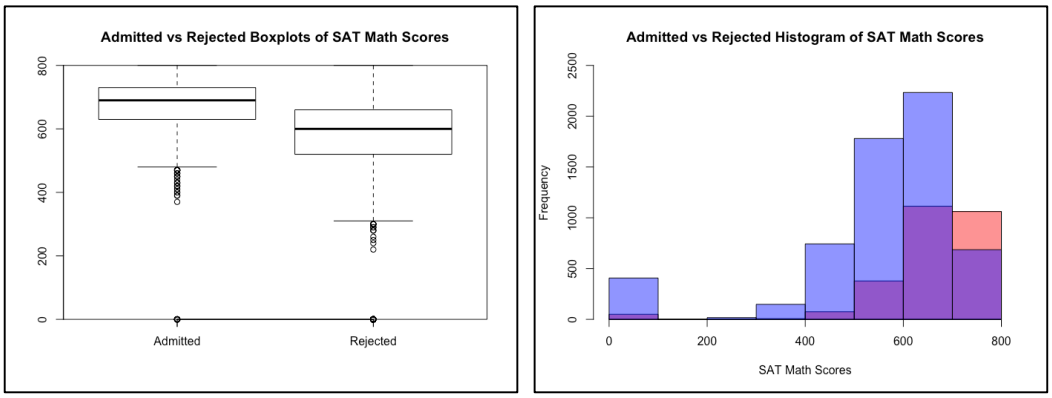

Figure 10a & 10b.

As expected, admitted applicants have higher SAT scores across both sections. Again, surprisingly, applicants with seemingly high SAT Verbal and Math scores were rejected. This suggests that there is more to consider when admitting an applicant to Elite University than just his/her GPA and SAT scores.

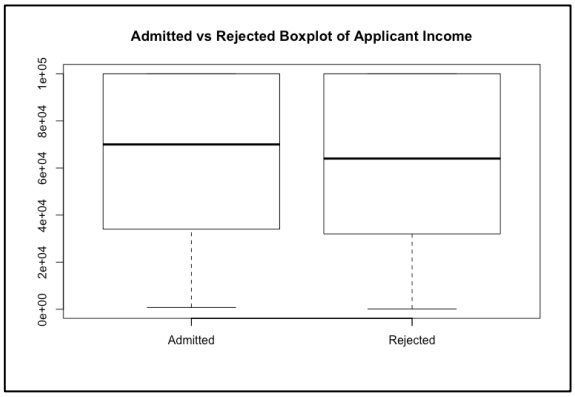

Figure 11.

While the variation seems to be similar between the incomes of the admitted and rejected applicants, the measures of central tendency seem to differ slightly. The median income of the admitted applicants ($70,000) is slightly higher than the median income of the rejected applicants ($64,000). While many universities make their admissions decisions via a "financial-aid-blind" process, a student's ability to pay must be taken into consideration somehow, or else the university does not make any money.

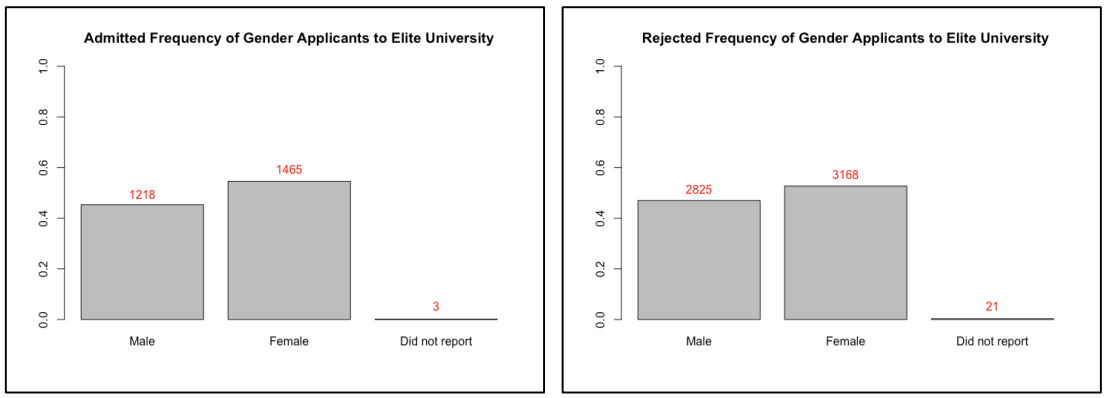

Figure 12.

The relative frequencies between male and female applicants do not change across the admitted pool and rejected pool of applicants. This undermines gender's importance as a variable when it comes to determining whether or not a student will get admitted.

Concluding Thoughts…

The Level I analysis above did not yield any compelling insight as to why some students are admitted, and others are rejected. While admitted students do tend to have higher median SAT scores and weighted GPAs, there seems to be other factors at play affecting Elite University's admissions, because there are many rejected applicants that also have SAT scores and weighted GPAs just as high. There seems to be no discrepancies in relative frequencies between the admitted pool and rejected pool in regards to race,

gender, and income. This is good news, because it shows that Elite University does not discriminate against race, gender, or income of its applicants. This is bad news, because it makes the guidance counselors job much harder in determining which students get admitted. We now turn to Level II analysis.

Level II Analysis

Assumptions

Before we dive into a Level II analysis where we try to build a random forest classifier for forecasting purposes, we'll need to defend if a Level II analysis is appropriate. The first assumption that we make is that the data are random, independent realizations from nature's joint probability distribution. It is unclear whether or not the data was a random sample from the total pool of applicants to Elite University in the previous year, or if the data is the total pool of applicants. We should proceed with caution, knowing that if the data was not randomly sampled, the insights we derive from this data may not be representative of the population. Additionally, there is some likelihood that some realizations in the data are not independent. For example, perhaps two applicants were twins, and the admissions board decided that if one of them got in, then they would both get in. Another example would be that the admissions board has set a quota for how many people they want to take from a particular high school, city, or region. Again, we will need to proceed with caution, knowing that independence in the realizations is not certain.

In this case, nature's joint probability distribution is the distribution between whether or not an applicant is admitted to Elite University and the attributes that describe the profile of the applicant. The key to a credible data generating process is that it is stable enough to draw inferences from. That means, we assume that the process that produces applicants applying to Elite University (within a reasonable time frame of several years), does not change. I think this is a credible assumption to make, especially if the time frame of making inferences is only a few years. After a few years, the model and joint probability distribution will probably need to be updated to account for general changes in the education system.

Out-of-bag (OOB) data will be used to obtain fitted values and an honest estimate of the out-of-sample error rate. Thus the OOB error rate is an estimate of the test error rate (had there been a test set), and the test error rate is an estimate of the true generalization error of the approximation of the true response surface in the joint probability distribution. Effectively, OOB data is an approximation for test data. Lastly, the term "honest" here refers to the fact that the OOB data was not used to determine the partitions in the random forest classifier, and thus confusion tables and error rates derived from OOB data are "honest" estimates.

If the guidance counselor plans to use this random forest model to help predict the admittance probabilities of her own high school's students, then this could spell trouble. The problem here would be that the data generating process is different between this particular high school's ability to generate applicants to Elite University (we don't have this data), and the aggregate of every other high school's ability to generate applicants to Elite University (we only have this data). Nonetheless, since this particular high school's data generating process is part of the larger data generating process as well, perhaps we can still make some reasonable inferences on this particular high school's student's ability to be admitted to Elite University.

Improvements over CART

Random forest makes a number of improvements upon the CART/decision tree algorithm. The first improvement is less bias through the ability to grow larger trees (pending a large sample size as well). This leads to interpolation of the data. The next improvement is less variance through the ability to grow many trees (500 at least) and averaging the fitted values across all the trees. Another improvement is one

that contributes to the diversity amongst the trees in the random forest: sampling from a subset of predictors at each split. The fitted values resulting from these trees are more independent of one another. Additionally, this gives more predictors, even weak ones, a chance to contribute to forecasting performance. The aggregation of the predictive power of many weak predictors can actually improve fit and forecasting ability greatly.

Missing Values

Missing values only become a problem when a) excluding them makes the dataset too small and/or b) there was a systematic error in the data collection process that interfered with the values from being collected. Thus we need to take a subject-matter approach into considering why there is missing data, and if there seems to be a pattern with how the data is missing. For example, when it comes to the SATs, I suspect that almost all applicants who didn't take the SAT took another standardized test instead, such as the ACT, which is not reported in this dataset. The people who did not take the SAT seem to appear at random. Additionally, other applicants with missing information could possibly be due to several reasons. One reason could be that, in the end, these applicants decided not to apply to Elite University, so they did not bother filling out the rest of their information. This can be either due to acceptance to another university, or lack of interest. Therefore, we can simply perform row-wise deletion and delete the rows with missing values (NAs). The only concern here is if the dataset gets too small – luckily, this is not the case and the dataset is still rather large even after removing the rows with missing values. The dataset now contains 6175 *complete* applicant profiles.

Asymmetric Costs

We need to use our subject matter expertise to decide how bad misclassifying a student as "Admitted" is relative to misclassifying a student as "Rejected". Since the purpose of this model is to determine the student profile that has a high chance of getting admitted into Elite University, then misclassifying a student as "Admitted" is worse than misclassifying a student as "Rejected". It is more important to know with a very high degree of certainty that a student labeled as "Admitted" will get in, because at the end of the day, any student regardless of label can still apply. Thus, we will choose an asymmetric loss matrix that minimizes the false positives. It is important to note that the costs are not parameters to be tuned. These asymmetric costs are a feature of the training process that is determined in advance, based on substantive or policy considerations. Without more information regarding the guidance counselor's needs or expectations, it indeed may seem a bit arbitrary to simply specify any type of asymmetric cost ratio. Perhaps we will aim for a misclassification costs of: False Positive = 10 x False Negative, since a 10 to 1 ratio is usually sufficient.

There is evidence in the area of random forest research that suggests that stratified bootstrap samples seem to work the best for introducing asymmetric costs. The bootstrap distribution will be determined via trial and error. The sample size for the less common response category should be equal to about two-thirds of the number of instances in the category. In other words, since there are 1925 admits and 4250 rejects in our dataset with missing values removed, "admits" is the less common response category. And thus, the sample size should be equal to 2/3 * 1925, or about 1284 for the stratified bootstrap sampling.

Final Model

Random Forest w/o Misclassification Costs…

Parameters:

- Number of trees = 500
- # of Predictors Sampled at Each Split = $\sqrt{\#\ of\ predictors} = \sqrt{8} \approx 2$

OOB Error Rate: 16.19%

Confusion Matrix:

Actual\Predicted	0	1	Model Error
0	3930	320	0.07529412
1	680	1245	0.35324675
Use Error	0.1475054	0.2044728	0.16194330

Table 1.

An OOB Error Rate of 16.19% is not too bad! The only issue here is that, as a high school guidance counselor, I want to minimize the number of false positives that I make (student will get rejected from Elite University but the algorithm says they will get admitted). Thus we need to implement misclassification costs.

Random Forest w/ Misclassification Costs…

Parameters:

- Number of trees = 500
- # of Predictors Sampled at Each Split = $\sqrt{\#\ of\ predictors} = \sqrt{8} \approx 2$
- sampsize = c(450,1284)
    - Stratified Sample Size of "Admit" = 1284
    - Stratified Sample Size of "Reject" = 450

OOB Error Rate: 26.01%

Confusion Matrix:

Actual\Predicted	0	1	Model Error
0	2789	1461	0.34376471
1	145	1780	0.07532468
Use Error	0.04942059	0.4507868	0.260081

Table 2.

The OOB Error Rate increased about 10% with the asymmetric costs. From the confusion table, we gather that a false positive is 10 times more costly than a false negative. The reason why this model is more valuable to the guidance counselor is because it predicts who will not get admitted to Elite University quite accurately (if the model predicts that the student will get rejected from Elite University, there is a 95.05% chance that the prediction was correct). This is more valuable to students because this can help them focus on applying to other schools that they have a more realistic chance of getting into. Simultaneously, this model also limits the number of students who actually will get admitted to Elite University but will be labeled as "rejected" by the model (that

is, 7.53% of students who will get admitted into Elite University will be incorrectly labeled as "rejected" by the model).

Lastly, if the model predicts that you will get admitted into Elite University, there is nearly a 50-50 chance that you do get in. The reason why I am comfortable with this is because in real life, getting admitted to a truly "elite" university really comes down to luck for many people. Admissions officers at Ivy League schools frequently say that they can build an equally strong class from their rejected applicant pool compared to their accepted applicant pool. For most applicants, there is really no guarantee that you will get into an elite university. Thus, it may be more useful to have a model that serves as more of a "reality check" for high school seniors looking to apply to Elite University – that is, it is more realistic to guarantee that you *will not* get into an elite university than it is to guarantee that you *will* get into an elite university.

Variable Importance Plot…

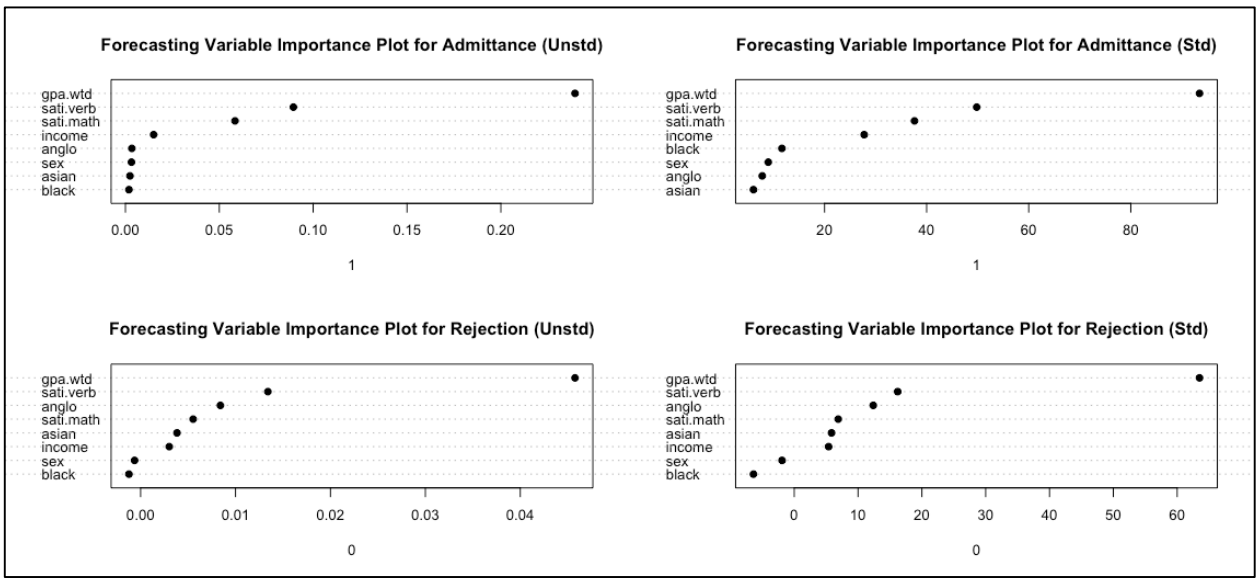

Figure 13. 1 = Admitted and 0 = Rejected

These variables are ranked by their contribution to the forecasting accuracy of the model. The reason why the variable rankings differ from plot to plot is because the variable importance plot is affected by which class label is being used (since class label is assigned by majority vote, which is determined by the margin and the number of actual class members) and whether or not the plots are standardized. It is advisable to study all the variations of the variable importance plot to get a holistic picture of what is going on. It seems that "weighted GPA" and "SAT Verbal score" are ranked 1st and 2nd for each of the four plots. This strongly suggests that these two variables the most important in terms of improving the model's forecasting accuracy. Important limitations to note are that (a) the variable importance plot does not tell us *how* a predictor is linked to the response variable and (b) these plot shouldn't be used to perform feature selection to use in other models, particularly because these variables are ranked by forecastability, not causality.

Partial Dependence Plots…

A partial dependence plot conveys the average relationship between a given predictor and the response within a fixed, joint distribution of the other predictors. The response functions displayed in the partial dependence plot reflect the relationship between a given predictor and the response, conditioning on the rest of the predictors.

Race:

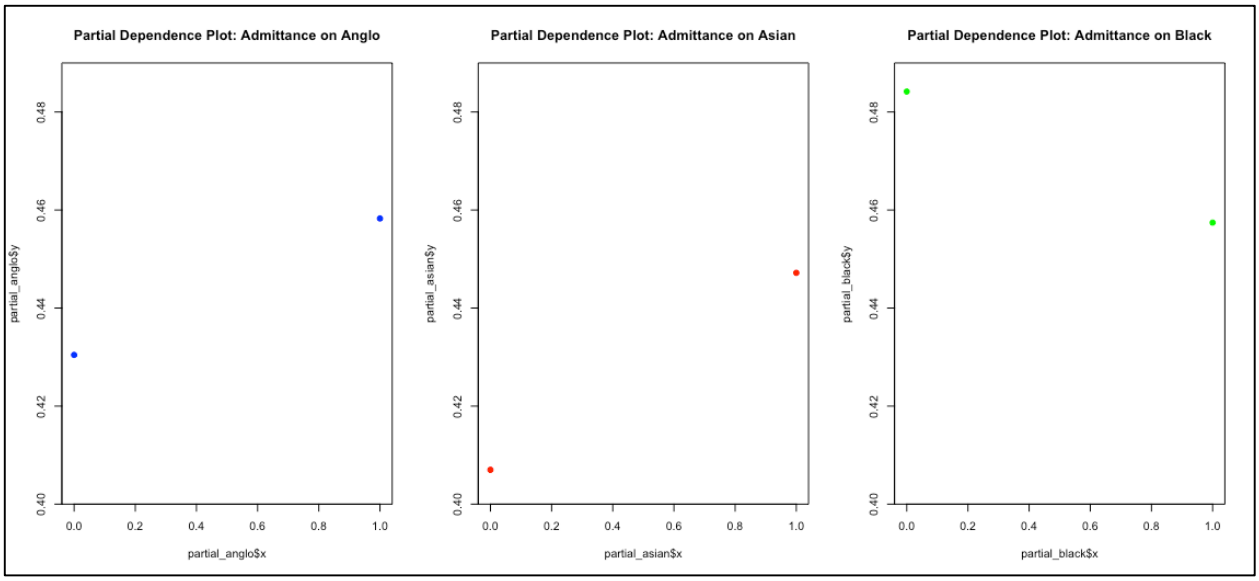

Figure 14. Anglo, Asian, and Black are all binary variables. Y-axis is in units of centered logits.

From Figure 14, it is possible to back out proportions for the centered logits using the formula: $\frac{e^{(2*logit)}}{1+e^{(2*logit)}}$. These proportions can then be used to evaluate whether or not a predictor is strongly associated with the response. For example, "Anglo" is not strongly associated with the response. The proportion of applicants who get admitted while being non-Anglo is 70.28%, while the proportion of applicants who get admitted while being Anglo is 71.43%. Thus, being Anglo did not affect your chances of getting admitted by too much. An interesting observation is that being black lowers your chances of getting admitted, but you still have a better chance of getting admitted than if you were anglo (by a little bit) or asian.

GPA:

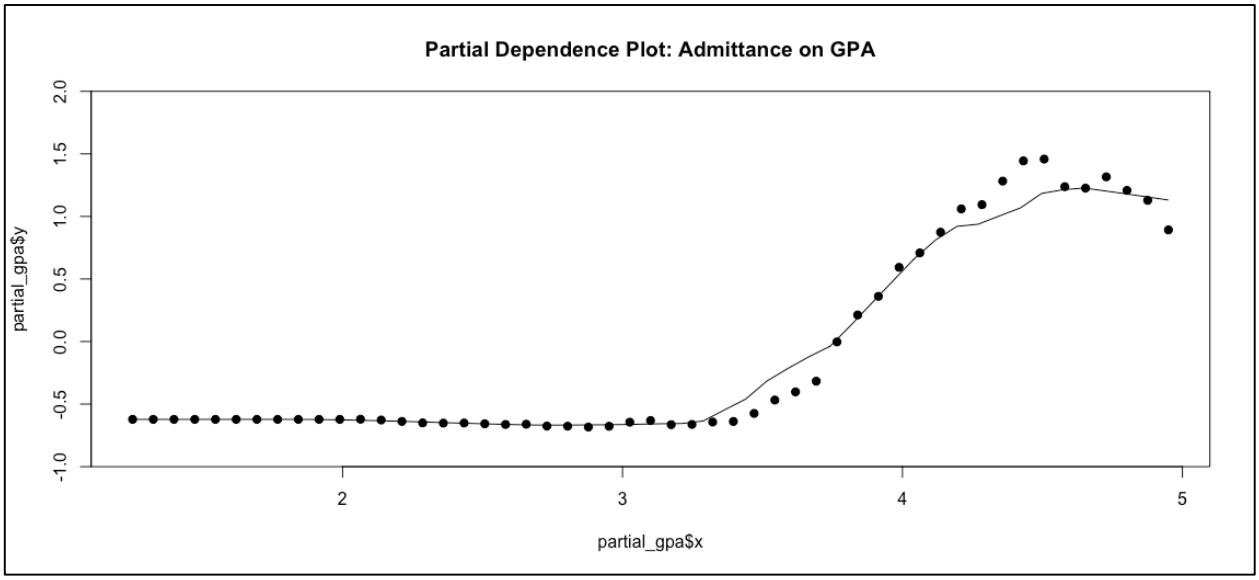

Figure 15.

Figure 15 tells us that up until a GPA of 3.5 your chances of getting admitted to Elite University are not increasing. It seems that a GPA of around 4.5 is the sweet spot of maximizing your chances of getting into Elite University. This may seem counter-intuitive at first, because shouldn't a GPA of 5.0 give you the greatest chance of getting into Elite University? Well, the reality is that at elite universities, they don't just look at your GPA. Many elite universities particularly value extracurricular activities and leadership roles. Perhaps students who attain a 5.0 GPA are often too preoccupied in maintaining their high GPA than engaging in extracurricular/leadership activities, and thus they end up hurting their chances of getting into Elite University.

SATs:

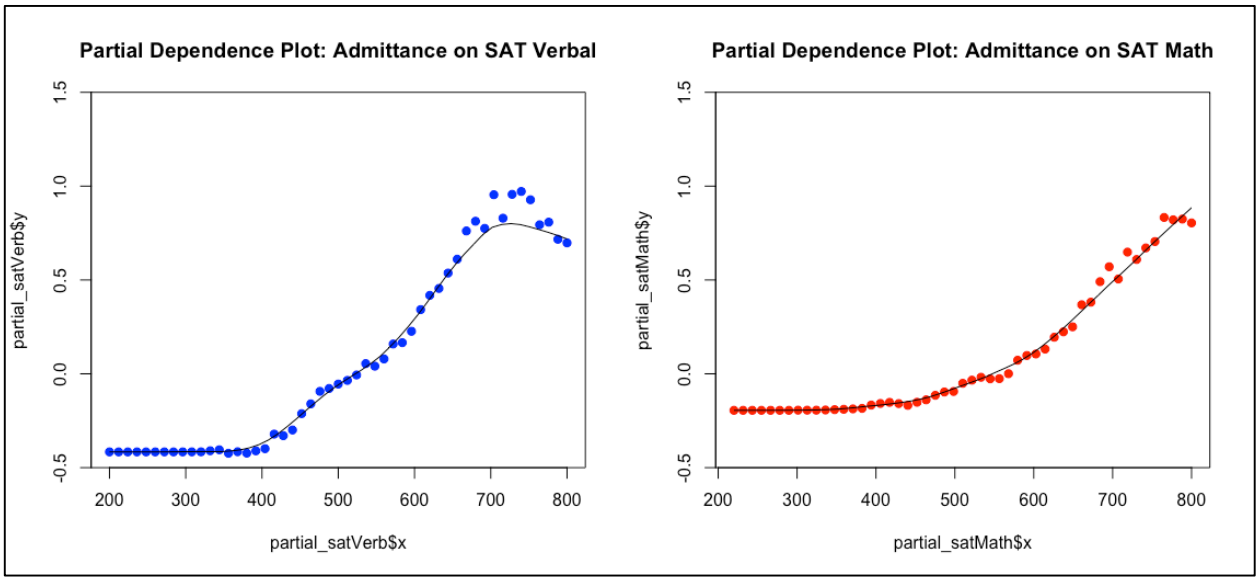

Figure 16.

The same phenomenon observed in the partial dependence plot for admittance on GPA is also evident in the SAT Verbal plot, and slightly evident in the SAT Math plot. This makes sense because perhaps the time it requires a student to attain an SAT score of higher than 750 in both sections is actually counter-productive because it takes away time that they could be spending on doing extracurricular activities, which are also factors when Elite University makes admissions decisions. The SAT Verbal plot seems to indicate that up until a score of 400, your chances of getting into Elite University will not be increasing. For the SAT Math plot, that threshold seems to be around 450.

Income:

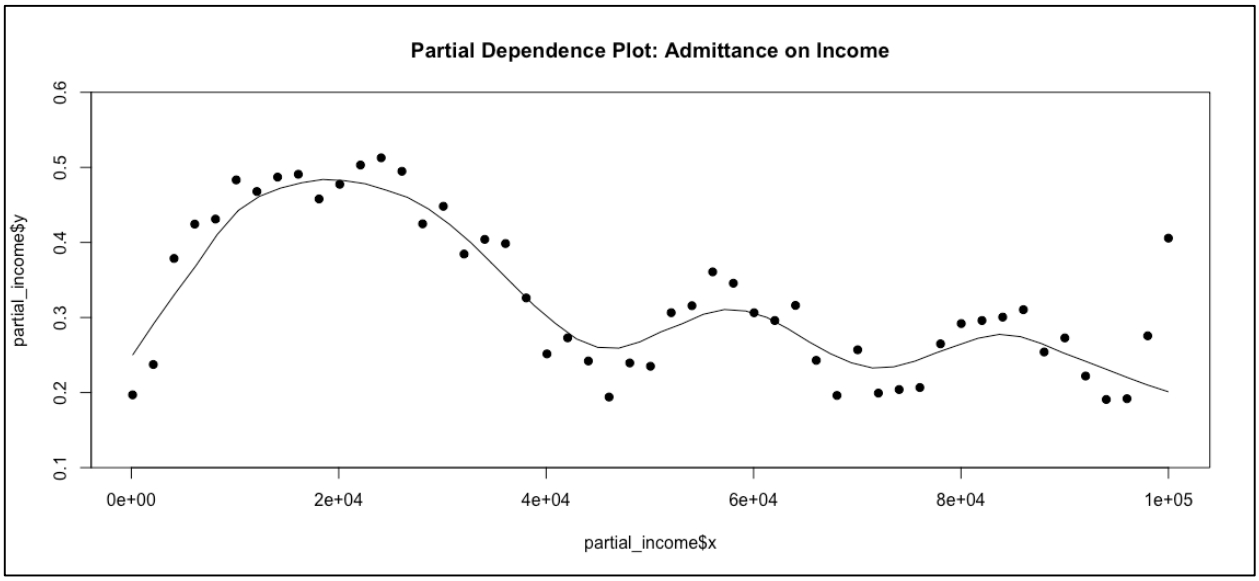

Figure 17.

The average relationship between income and admittance does seem a bit random. However, generally speaking, it does seem that applicants coming from a household income of around $20,000 seem to have the best chances at making it into Elite University. The crests and troughs of the relationship do seem odd. Perhaps there is a systematic idiosyncrasy in the way Elite University accepts its students (perhaps many come for the same city, state, high school or perhaps Elite University knowingly or unknowingly accepts its students based on a certain kind of profile(s)).

Empirical Margins…

These margins represent the reliability of the class assigned to a given observation. A large margin often means that the signal dominates the noise for that particular observation. It is important to realize that reliability does not equate to validity of the class label.

Here are some summary statistics after calculating the empirical margins for each one of our 6175 observations.

Min	1st Quantile	Median	Mean	3rd Quantile	Max
-1.00000	-0.03579	0.49440	0.32810	0.76690	1.00000

Table 3.

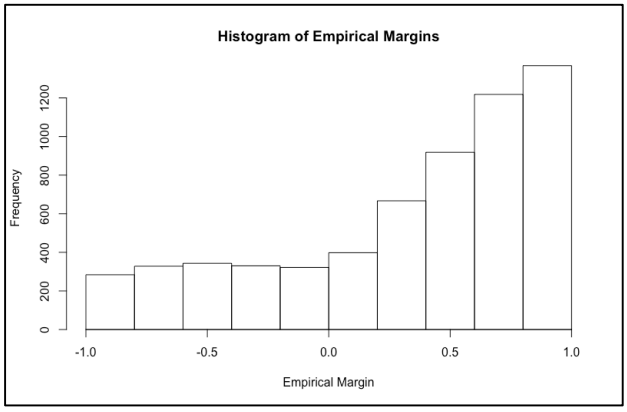

Figure 18.

73.97% of all observations have an empirical margin greater than 0. About 50% of all observations have an empirical margin greater than 0.5. In terms of evaluating the reliability of the labels made by the random forest classifier, I would say that it is not great. About a quarter of our observations have negative empirical margin, meaning they were misclassified. And another quarter of our observations were correctly classified, but have rather small empirical margins (less than 0.5), suggesting that observations with the same predictor values have a higher chance of being misclassified.

Conclusions

From our random forest classifier, we can conclude first and foremost that there must be other relevant predictors that contribute to Elite University's admissions decisions. This is supported by the fact that our Level 1 analysis yielded not a single predictor that demonstrated a strong relationship in separating the admitted students from the rejected students. Additionally, our Level II analysis saw slight dips in admittance chances towards the upper ends of GPA, SAT Verbal, and SAT Math. It doesn't make sense for people with the highest GPAs, SAT Verbal, and SAT Math not to get into Elite University, unless there were some other factors at play here.

Additionally, the asymmetric costs adjustment to the random forest classifier was necessary to make the algorithm more applicable and useful to the guidance counselor.

Next Steps

As for next steps, I would encourage the guidance counselor to collect more data on admitted and rejected students to Elite University, specifically more predictors (potentially ones that characterize extracurricular or leadership activities).

Furthermore, I would strongly encourage the guidance counselor to collect data on her *own* school's students and their track record of admittance/rejection to Elite University. This way the inferences she draws from any statistical analysis she performs on the data can more strongly characterize her population of interest (i.e. her own school's students).

R Code

```
data <- Admissions

dim(data)
str(data)
summary(data) # lots of NAs
attach(data)

###### Level 1 Analysis
### Univariate
admitBarplot <- barplot(c(length(admit[admit==1])[1]/8700,
                          length(admit[admit==0])[1]/8700),
                        main="Frequency of Admission to Elite University",
                        names.arg=c("Admitted","Rejected"),
                        ylim=c(0,1))
text(x=admitBarplot,y=c(0.31,0.69),label=c(2686,6014),pos=3,col="red")

raceBarplot <- barplot(c(sum(complete.cases(anglo[anglo==1]))/8700,
                         sum(complete.cases(asian[asian==1]))/8700,
                         sum(complete.cases(black[black==1]))/8700,
                         2117/8700),
                       main="Frequency of Race in Applications to Elite University",
                       names.arg=c("Anglo","Asian","Black","Unreported"),
                       ylim=c(0,1))
text(x=raceBarplot,y=c(0.32,0.39,0.04,0.24),label=c(2810,3417,356,2117),pos=3,col="red")

boxplot(gpa.wtd,main="Boxplot of Weighted GPA")
axis(side=1,at=c(1),labels=c("Weighted GPA"))

satBoxplot <- boxplot(sati.verb,sati.math,main="Boxplots of SAT Scores")
axis(side=1,at=c(1,2),labels=c("Verbal","Math"))

boxplot(income,main="Boxplot of Applicant Income")
summary(income)

sexBarplot <- barplot(c(dim(data[data$sex==1,])[1]/8700,
                        dim(data[data$sex==0,])[1]/8700,
                        24/8700),
                        main="Frequency of Gender Applicants to Elite University",
                        names.arg=c("Male","Female","Unreported"),
                        ylim=c(0,1))
text(x=sexBarplot,y=c(0.46,0.53,0.01),label=c(4043,4633,24),pos=3,col="red")

### Bivariate
admitYes <- data[data$admit==1,]
admitNo <- data[data$admit==0,]

raceBarplot <- barplot(c(sum(complete.cases(admitYes$anglo[admitYes$anglo==1]))/2686,
                         sum(complete.cases(admitYes$asian[admitYes$asian==1]))/2686,
                         sum(complete.cases(admitYes$black[admitYes$black==1]))/2686,
                         594/2696),
                       main="Admitted Frequency of Race in Applications to Elite University",
                       names.arg=c("Anglo","Asian","Black","Did not report"),
                       ylim=c(0,1))
text(x=raceBarplot,y=c(0.34,0.41,0.04,0.22),label=c(912,1093,87,594),pos=3,col="red")

raceBarplot <- barplot(c(sum(complete.cases(admitNo$anglo[admitNo$anglo==1]))/6014,
                         sum(complete.cases(admitNo$asian[admitNo$asian==1]))/6014,
                         sum(complete.cases(admitNo$black[admitNo$black==1]))/6014,
                         1523/6014),
                       main="Rejected Frequency of Race in Applications to Elite University",
```

```
                      names.arg=c("Anglo","Asian","Black","Did not report"),
                      ylim=c(0,1))
text(x=raceBarplot,y=c(0.32,0.39,0.05,0.25),label=c(1898,2324,269,1523),pos=3,col="red")

boxplot(admitYes$gpa.wtd,admitNo$gpa.wtd,main="Admitted vs Rejected Boxplot of Weighted GPA")
axis(side=1,at=c(1,2),labels=c("Admitted","Rejected"))

hist(admitYes$gpa.wtd,col=rgb(1,0,0,0.5),breaks=seq(0,5,by=.1),ylim=c(0,600),main="Admitted    vs
Rejected Histogram of Weighted GPA",xlab="Weighted GPA")
hist(admitNo$gpa.wtd,add=T,col=rgb(0,0,1,0.5),breaks=seq(0,5,by=.1))

satBoxplot <- boxplot(admitYes$sati.verb,admitNo$sati.verb,main="Admitted vs Rejected Boxplots of
SAT Verbal Scores")
axis(side=1,at=c(1,2),labels=c("Admitted","Rejected"))

hist(admitYes$sati.verb,col=rgb(1,0,0,0.5),breaks=seq(0,800,by=100),ylim=c(0,2500),main="Admitted
vs Rejected Histogram of SAT Verbal Scores",xlab="SAT Verbal Scores")
hist(admitNo$sati.verb,add=T,col=rgb(0,0,1,0.5),breaks=seq(0,800,by=100))

satBoxplot <- boxplot(admitYes$sati.math,admitNo$sati.math,main="Admitted vs Rejected Boxplots of
SAT Math Scores")
axis(side=1,at=c(1,2),labels=c("Admitted","Rejected"))

hist(admitYes$sati.math,col=rgb(1,0,0,0.5),breaks=seq(0,800,by=100),ylim=c(0,2500),main="Admitted
vs Rejected Histogram of SAT Math Scores",xlab="SAT Math Scores")
hist(admitNo$sati.math,add=T,col=rgb(0,0,1,0.5),breaks=seq(0,800,by=100))

boxplot(admitYes$income,admitNo$income,main="Admitted vs Rejected Boxplot of Applicant Income")
axis(side=1,at=c(1,2),labels=c("Admitted","Rejected"))
summary(admitYes$income)
summary(admitNo$income)

sexBarplot <- barplot(c(sum(admitYes$sex==1,na.rm=TRUE)/2686,
                        sum(admitYes$sex==0,na.rm=TRUE)/2686,
                        3/2686),
                      main="Admitted Frequency of Gender Applicants to Elite University",
                      names.arg=c("Male","Female","Did not report"),
                      ylim=c(0,1))
text(x=sexBarplot,y=c(0.45,0.55,0.01),label=c(1218,1465,3),pos=3,col="red")

sexBarplot <- barplot(c(sum(admitNo$sex==1,na.rm=TRUE)/6014,
                        sum(admitNo$sex==0,na.rm=TRUE)/6014,
                        21/6014),
                      main="Rejected Frequency of Gender Applicants to Elite University",
                      names.arg=c("Male","Female","Did not report"),
                      ylim=c(0,1))
text(x=sexBarplot,y=c(0.47,0.53,0.01),label=c(2825,3168,21),pos=3,col="red")

###### Level II Analysis
install.packages("randomForest")
library(randomForest)

### Remove missing values
dataNA <- na.omit(data)
summary(dataNA)
dataZero <- dataNA[dataNA$gpa.wtd!=0,]
summary(dataZero)
dataMath <- dataZero[dataZero$sati.math!=0,]
summary(dataMath)
dataRead <- dataMath[dataMath$sati.verb!=0,]
summary(dataRead)
dataFinal <- dataRead
```

```
dim(dataFinal) # n = 6175
dataFinal$admit <- as.factor(dataFinal$admit)

### Random Forest Classifier
# No Misclassification Costs
rf1 <- randomForest(admit~.,
                    data=dataFinal,
                    ntree=500,
                    importance=TRUE)

# With Misclassification Costs
rf2 <- randomForest(admit~.,
                    data=dataFinal,
                    ntree=500,
                    importance=TRUE,
                    sampsize=c(450,1284)) # 450,1284

### Variable Importance Plot
par(mfrow=c(2,2))
varImpPlot(rf2,type=1,scale=F,class=1,
           main="Forecasting Variable Importance Plot for Admittance (Unstd)",
           pch=19)
varImpPlot(rf2,type=1,scale=T,class=1,
           main="Forecasting Variable Importance Plot for Admittance (Std)",
           pch=19)
varImpPlot(rf2,type=1,scale=F,class=0,
           main="Forecasting Variable Importance Plot for Rejection (Unstd)",
           pch=19)
varImpPlot(rf2,type=1,scale=T,class=0,
           main="Forecasting Variable Importance Plot for Rejection (Std)",
           pch=19)

### Partial Dependence Plots
# race
partial_anglo <- partialPlot(rf2,pred.data=dataFinal,x.var=anglo,which.class=1)
partial_asian <- partialPlot(rf2,pred.data=dataFinal,x.var=asian,which.class=1)
partial_black <- partialPlot(rf2,pred.data=dataFinal,x.var=black,which.class=1)

par(mfrow=c(1,3))
scatter.smooth(partial_anglo$x,partial_anglo$y,span=.5,
               main="Partial Dependence Plot: Admittance on Anglo",
               pch=19,col="blue",
               ylim=c(0.40,0.49))
scatter.smooth(partial_asian$x,partial_asian$y,span=.5,
               main="Partial Dependence Plot: Admittance on Asian",
               pch=19,col="red",
               ylim=c(0.40,0.49))
scatter.smooth(partial_black$x,partial_black$y,span=.5,
               main="Partial Dependence Plot: Admittance on Black",
               pch=19,col="green",
               ylim=c(0.40,0.49))

# GPA
partial_gpa <- partialPlot(rf2,pred.data=dataFinal,x.var=gpa.wtd,which.class=1)

par(mfrow=c(1,1))
scatter.smooth(partial_gpa$x,partial_gpa$y,span=.25,
               main="Partial Dependence Plot: Admittance on GPA",
               pch=19,col="black",
               ylim=c(-1,2))

# SAT
```

```
partial_satVerb <- partialPlot(rf2,pred.data=dataFinal,x.var=sati.verb,which.class=1)
partial_satMath <- partialPlot(rf2,pred.data=dataFinal,x.var=sati.math,which.class=1)

par(mfrow=c(1,2))
scatter.smooth(partial_satVerb$x,partial_satVerb$y,span=.25,
               main="Partial Dependence Plot: Admittance on SAT Verbal",
               pch=19,col="blue",
               ylim=c(-0.5,1.5))
scatter.smooth(partial_satMath$x,partial_satMath$y,span=.25,
               main="Partial Dependence Plot: Admittance on SAT Math",
               pch=19,col="red",
               ylim=c(-0.5,1.5))

# income
partial_income <- partialPlot(rf2,pred.data=dataFinal,x.var=income,which.class=1)

par(mfrow=c(1,1))
scatter.smooth(partial_income$x,partial_income$y,span=.25,
               main="Partial Dependence Plot: Admittance on Income",
               pch=19,col="black",
               ylim=c(0.1,0.6))

### Empirical Margins
mr <- margin(rf2)
summary(mr)
sum(mr>0)/6175 # 0.7397571
hist(mr,
     main="Histogram of Empirical Margins",
     xlab="Empirical Margin")
```