STAT 474 Assignment 1: University Admissions

Arjun Lal

This analysis will attempt to forecast admissions for future applicants to a particular "elite" university using data points from previous applicants ranging across race to GPA to income. Furthermore, this analysis will venture to determine exactly how each of the predictors present in the data are related to an admission decision. The conclusions from this analysis will better inform the current high school's principal and current students on what features of an application will be most important when applying to this particular university, and to what degree an individual's admission decision can be forecasted.

The data itself were taken from the aforementioned university. Across a total of 50,000 previous applicants to the university, 8700 were randomly sampled. The response variable is 'admit', and the predictors are 'asian', 'anglo', 'black', 'income', 'gpa.wtd', 'sex', 'sati.verb', and 'sati.math'. The predictors were selected because they were thought to be the most important factors in determining admission.

The analysis will be Level II. As stated in the description of the data, the data were drawn randomly. It is not clear what region or high school these applicants were selected from, which might have skewed the underlying distribution. For example, it is possible that this particular university favors a particular set of high schools when determining admissions, as it is well known that some top universities accept a large number of applicants from several "elite" private high schools. For this reason, it is possible that some of the observations are not entirely independent and only subject matter knowledge would be able to shed light on this possible discrepancy. However, given the large size of the dataset coupled with the randomness of the sampling, we can infer that the observations are independent for the most part. Thus, we can operate with the assumption that the data is independently and identically distributed, meaning that we can proceed with a Level II analysis.

We approach this analysis fully knowing that it will be taken within the "wrong model" perspective. It is clear that not all of the factors that determine an applicants admission or rejection are present in the data. Anything from the strength of a candidate's interviews to the temperament of an admissions officer could play key roles in deciding a potential admission. Furthermore, an additive model may very well not be the way that the data is actually generated. Nevertheless, the model we are using as an approximation (the generalized additive model) is quite interpretable and may be able to elucidate important factors in admissions at this university.

Cleaning and Transforming the Data

Upon examining the data in aggregate, it becomes immediately clear that there are several problems that require cleaning and wrangling of the data. A large amount of data is missing across several of the predictors. This is most significant in the income category, in which there are 1,724 missing observations. We have three options in dealing with this problem. One is to drop income as a predictor altogether, the second is to impute the missing values, and the last is

to drop rows in our dataset with missing observations here. We will utilize the third. Income is likely related in some way to admission, whether it be via donations to the university or more resources in preparing a stronger application for a candidate, and so dropping it altogether would likely remove important information. We also know that the household income in the data was capped at $100,000 and so imputing the data would be nonsensical as the data is already not representative of the true distribution of income and we do not know exactly why the data is missing. Thus, we drop the 1,724 observations corresponding to NAs in income, leaving us with 6,976 observations. Though that was a considerable loss, we still have a large enough dataset to partition into training, evaluation, and test datasets, as we will see later on. For similar reasons, we will also drop the 11 missing observations in the sex category, leaving us with 6,965 observations.

We now attend to the race-oriented features in the data: anglo, black, and asian. Upon dropping missing observations in income and sex, there are 461 missing observations in each of the three categories. Further examination reveals that these missing observations happen to occur in the *exact* same rows in our dataset. We can speculate as to why: anglo, black, and asian surely do not cover the whole gamut of race when it comes to applicants. Left out are those with less represented backgrounds such as Hispanics or Native-Americans. Looking even closer at the data allows us to see that there are also 1,740 rows in the data where all three of anglo, black, and asian are coded to 0. It is possible that during the data collection process, these 0s and NAs were coded in when there was confusion over an applicants race. Thus, we will introduce a new feature in our dataset, race_other, that is coded to 1 whenever all three of anglo, black, and asian are NA or 0, and 0 otherwise. Hopefully, this will help capture some of the missing information that was left out when the data was being collected. In line with this, we will transform instances of NAs in anglo, black, and asian to 0. To make our data easier to work with and our results more interpretable, we will convert all of our categorical variables from numeric to more distinguishable names (eg. asian goes from 0 and 1 to “non-asian” and “asian”).

Univariate Statistics

We will now examine the univariate statistics of each of the predictors, starting with the categorical ones. Looking at the admit category first, we can see that out of our 6,965 observations, 2,107 were admitted, about a 30.2% empirical acceptance rate. Though we do not know what the true acceptance rate is, this will be used as our baseline when evaluating the accuracy of our forecasting later on. The anglo and asian categories yield similar proportions. About 31% of the sampled applicants were anglo and 41% were asian. Additionally, almost 23% of the applicants were not accounted for by anglo, asian, or black, implying that there was a significant lack of representation in the initial data. What is concerning is the fact that only 4% of the remaining applicants are black, meaning that the ability of our model to better inform black applicants is weak. There is too disproportionately small a representation of black applicants in our data to draw conclusions of the same strength as those that we can about asian or anglo applicants. Additionally, there is a very even spread in terms of sex, with about 53% female and 47% male.

Looking at the income variable, we can see a large range. At the minimum is a reported household income of $120, surely an abnormal amount. However concerning this may

be, we keep it because we truly do not know what the financial circumstances of that particular applicant were and because we do not know what the spread in income of future applicants will be either. The median income reported is $65,000, which is a reasonable number commensurate with the median income in the US. Looking at a Figure 1, we can see that there are the largest number of observations at the far right tail of the plot. Knowing that income was capped at $100,000, this would imply that there are a decent number of observations with income above that threshold. This upper bound is a double-edged sword: though it prevents us from working with the true distribution of income, it also prevents outliers such as very well-off applicants from significantly affecting our results. The rest of the plot is quite uniform.

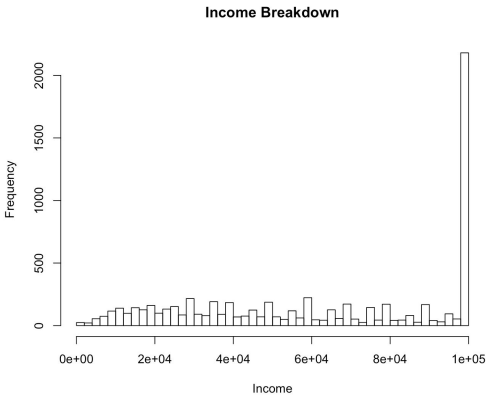


Figure 1

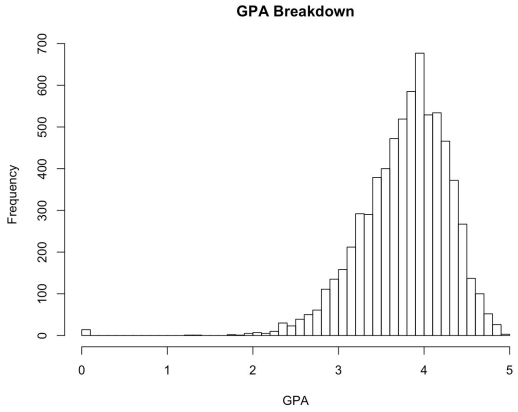


Figure 2

We now examine the quantitative metrics that seem to overwhelmingly define a candidates background, weighted GPA and SAT scores. Figure 2 belies an outlier at the very minimum, a weighted GPA of 0. This is indeed quite strange, as this would necessitate failing all of one's classes at most high schools. Nevertheless, we will keep it as it is not entirely nonsensical. The rest of the plot shows that weighted GPAs are quite normal, with a median around 3.85 and a max around 4.95. One thing to keep in mind during this analysis is that no matter how normal and reasonably distributed these GPAs seem to be, we do not know the distributions of GPAs and class difficulties at the various high schools that these applicants attended. A 3.7 for one applicant at high school 'A' could very well translate to a 4.2 at high school 'B'. Only subject matter expertise could really describe what a standardized dataset of weighted GPAs would look like. This is not an issue that we can readily address and so we will still proceed. The verbal and math SAT scores each have minimums of 0. This is quite an inconsistency in our data, as scores for the verbal and math sections are lower bounded by 200, meaning that there is no way that these scores of 0 were achieved if the tests were actually taken. This implies that there was either an error in the data collection process or that these applicants simply did not take the SAT. The latter option is more probable. Without further knowledge of the university's application policies, we must speculate that there were indeed applicants who either took another standardized test, such as the ACT, or just did not take a test at all but still applied. Unfortunately, we have no way of discerning which applicants fell into either of these categories. We will keep these scores of 0 in our data but will also keep in mind that they will most likely skew our results. The histograms of verbal and math scores are shown below. Apart from the 0s,

both are quite normal in distribution, though the math scores are left-tailed and have a local peak at a score of 800. Additionally, it would appear that the distribution of math scores encompasses more high scores than does the distribution of verbal scores. This, along with the peak in the math scores at 800, might imply that the math section of the SAT is easier to score well on and that scoring well on the verbal section could make an applicant "stand out" more and thus bolster his or her application.

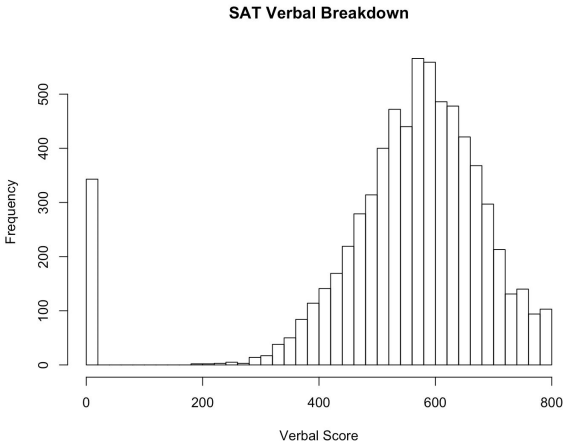


Figure 3

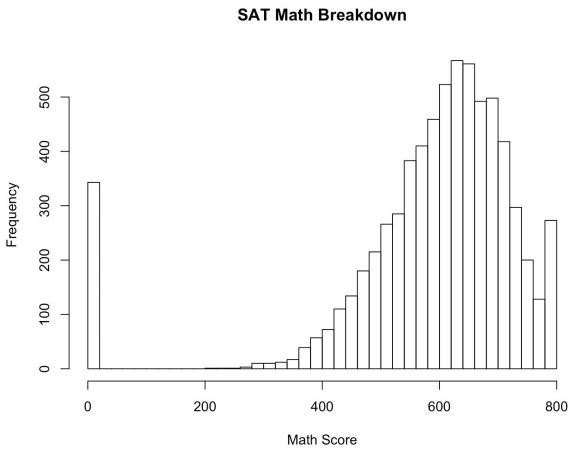
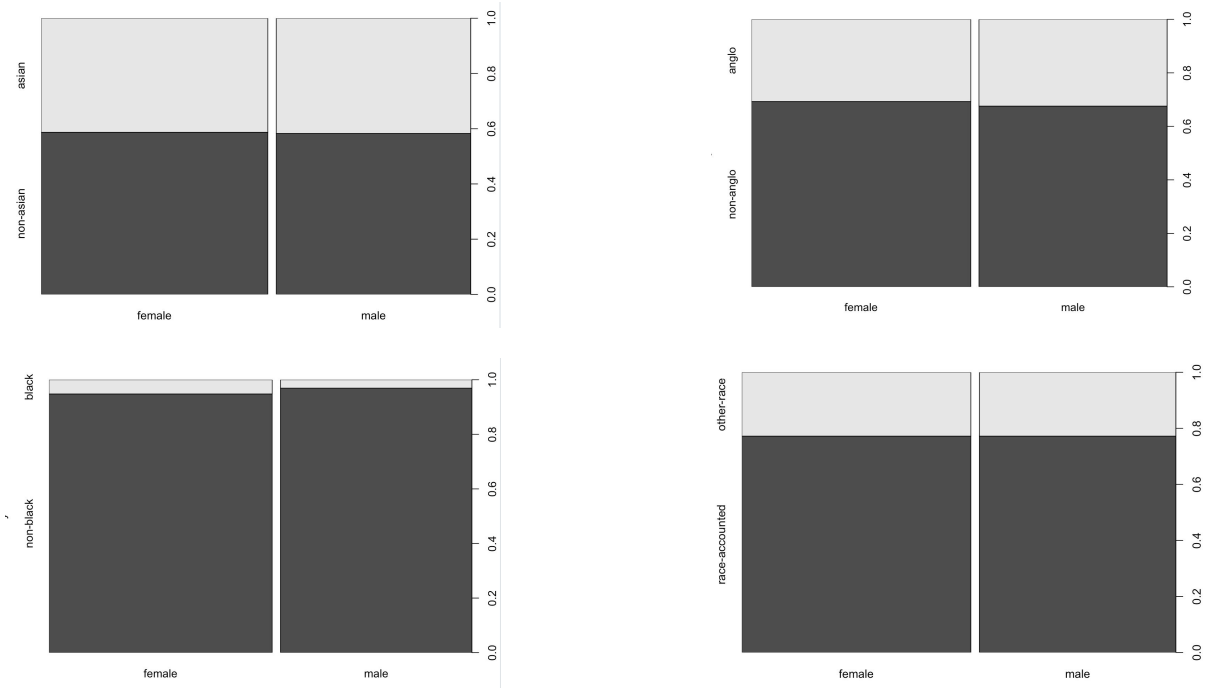


Figure 4

Bivariate Statistics

We will now examine the bivariate statistics of the predictors. We will not consider bivariate statistics of our race-oriented features as a few quick queries documented in the Appendix reveal that they are mutually exclusive (eg. no applicant was both black and asian).



Figures 5, 6, 7, & 8

Looking at race and sex, we can see that there are very even distributions shown in Figures 5, 6, 7, and 8. The only statistic that jumps out with a closer look is that out of the black applicants, there were almost twice as many females as males, at 195 and 99, respectively. This is a byproduct of the small number of black applicants in total, but further suggests that our model will not forecast well to black applicants, males in particular.

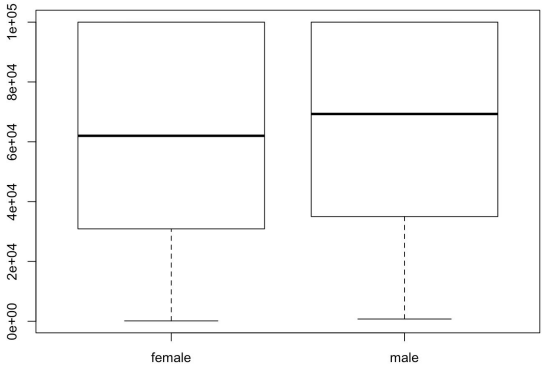
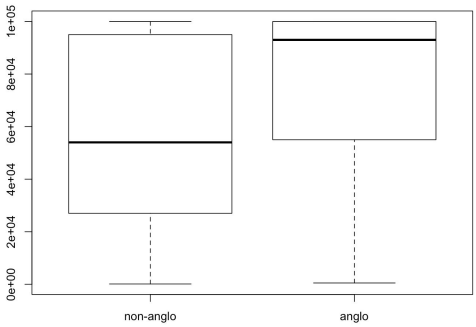
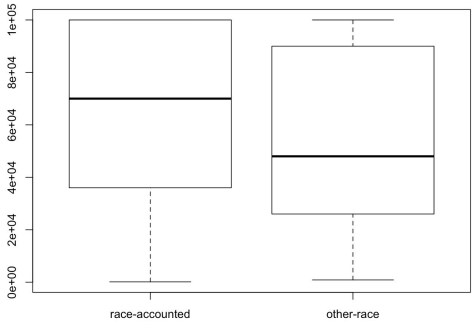


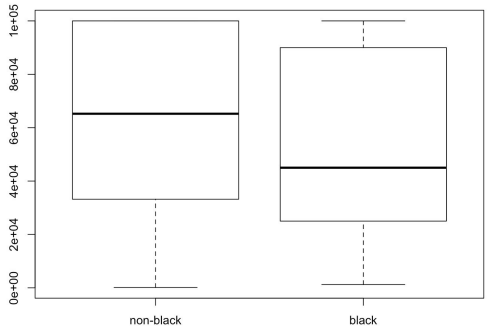
Figure 9

Shown in the Figure 9, male and female applicants came from similar distributions of income, though the median income of males was a bit higher. Figures 10, 11, 12, and 13, shown below, show much more serious issues in the backgrounds of applicants. The median incomes for applicants who were black, asian, and “other races” were much lower than the median incomes of those who were not, the issue being most apparent with black and “other race” applicants.

However, applicants who were anglo (Figure 10) came from incomes much higher than that of their peers. This is not a surprising observation, as there has been a great deal of discussion surrounding the financial backgrounds of college applicants of different races, but it is still an issue that may become evident in forecasting. More income means more resources to prepare for classes and tests and thus stronger applications. If income is indeed a key factor in determining admission, then it is likely that these blacks, asians, and other races may be negatively affected.







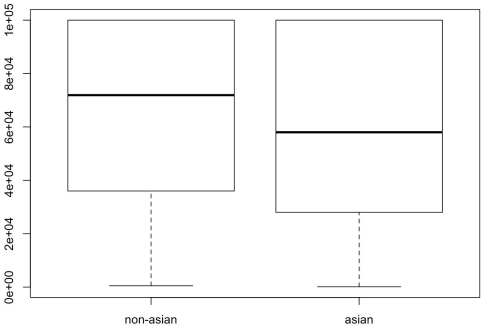


Figure 10 (top left), *Figure 11* (top right), *Figure 12* (bottom right), *Figure 13* (bottom left)

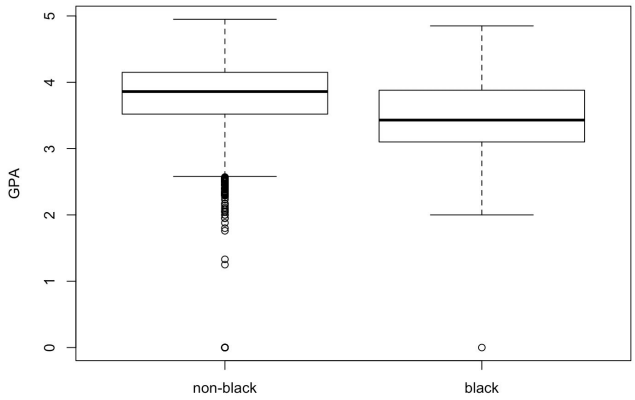
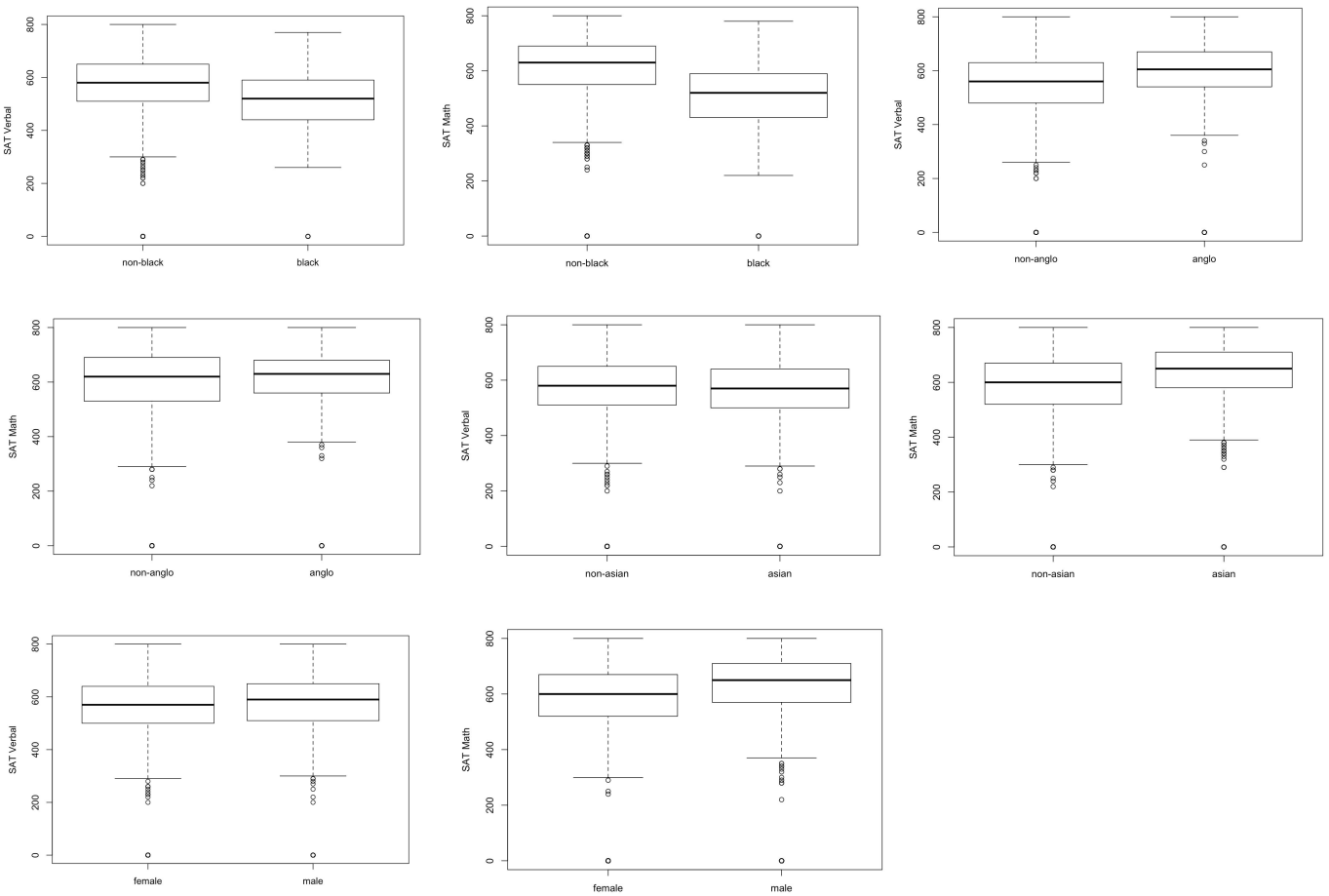


Figure 14

Looking at plots of race and sex vs. GPA reveals nothing of importance except for the fact that the GPAs of black applicants (Figure 14) were smaller than those of non-black applicants,

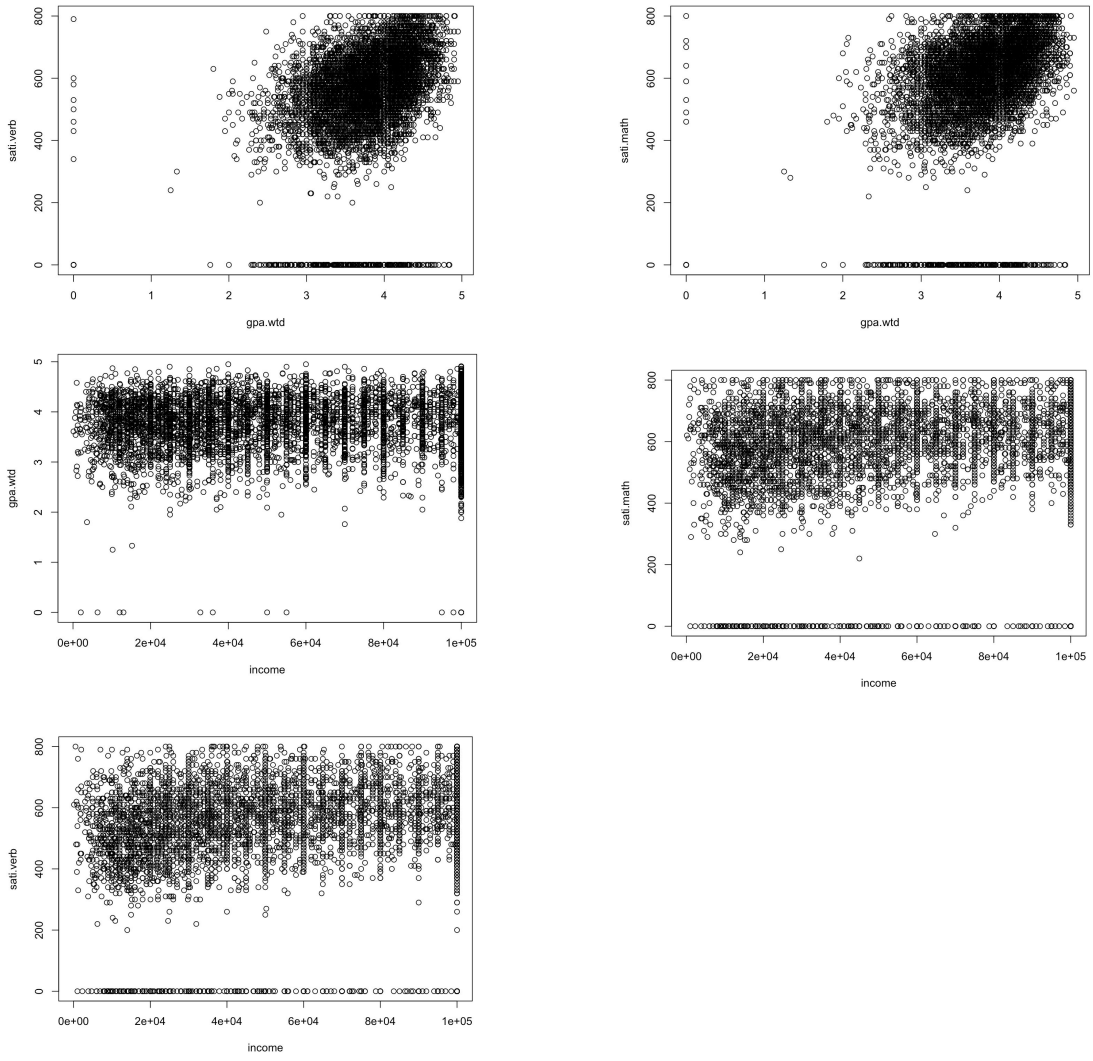
corroborating our claims from above that our model might be skewed negatively towards black applicants. The GPAs of black applicants *do* appear to be more concentrated with less variance, but this is likely because of the small number of black applicants. Plots of race and sex vs. SAT scores (both verbal and math) again reveal somewhat similar findings for black and other-race applicants, as they scored significantly lower than other applicants (Figures 15 and 16). There were some differences between the verbal and math sections for other groups of applicants, however. Though anglo applicants scored higher than their peers on the verbal section, they were roughly equal in the math section, though there was less variance in scores in the math section (Figures 17 and 18). Furthermore, even though asian applicants were on par with their peers on the verbal section, they scored significantly higher on the math section (Figures 19 and 20). The same comparison can be made between male and female applicants; though they scored similarly on the verbal section, male applicants scored higher on the math section (Figures 21 and 22).



Top Row (left to right): *Figures 15, 16, 17*; Middle Row (left to right): *Figures 18, 19, 20*; Bottom Row (left to right): *Figures 21, 22*

GPA and SAT scores seemed to be somewhat correlated, shown in Figures 23 and 24. The correlation between GPA and verbal scores was about 0.32, and the correlation between GPA and math scores was about 0.31. This is probably explained by the fact that students who perform well in school are more likely to perform well on standardized tests, and vice-versa.

There was really no correlation (0.065) between income and GPA (Figure 25). This again is indicative of the fact that the GPAs have their own underlying distributions depending on high school, independent of income. Income was also somewhat correlated with SAT scores, as the correlation with verbal scores was 0.21 and that with math scores was 0.18 (Figures 26 and 27). This reveals that students with higher incomes may have had more resources to prepare for standardized tests.



Top Row (left to right): *Figures 23, 24*; Middle Row (left to right): *Figures 25, 26*; Bottom: *Figure 27*

Applying the Generalized Additive Model

We will now apply the generalized additive model to our data, keeping in mind that we are operating in a Level II analysis. At a dataset size of 6,965, we are easily able to split our data up into three approximately equal size datasets for training, evaluation, and testing. We will use our training set to construct several models, the best of which we will choose using our evaluation set. We will then evaluate forecasting performance via our test set. As we are in Level II, we will

operate with probabilities instead of proportions. Before we proceed, we note that a Bayes classifier would yield a 69.7% misclassification rate on our data. This will be our baseline when determining how useful our forecasting really is.

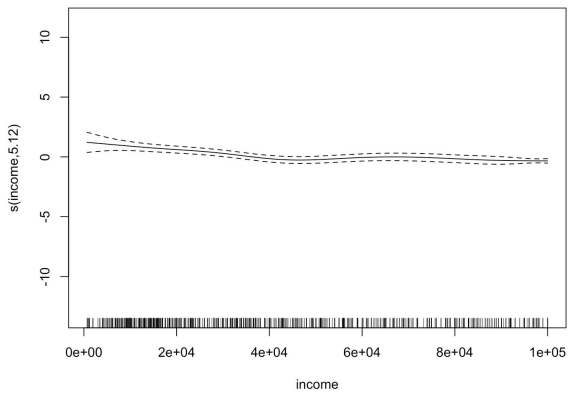
In constructing our models as part of our tuning process, we will be smoothing the numeric predictors (income, gpa.wtd, sati.verb, and sati.math). The categorical predictors will remain untouched. We construct models with sp values set equal to 0.01, 0.05, 0.1, and 1.

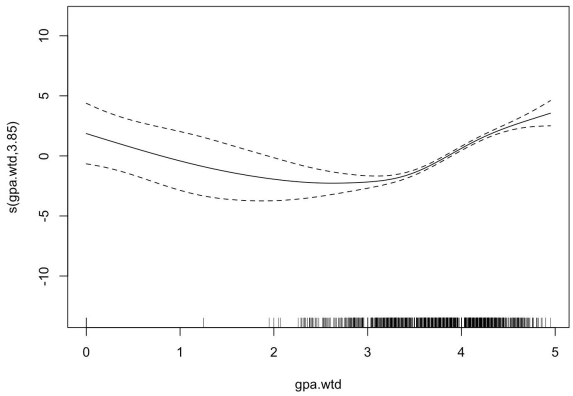
The first thing that stands out from the summary statistics of our model is that the estimate for race_other is 0, and the z-values and p-values are both NA, signifying that race_other is linearly dependent on the other race predictors in the dataset and therefore not useful. Sadly, it would appear that the creation of this new variable was not in fact fruitful in our analysis, and so we will drop it here. Nevertheless, we will still keep in mind the observations of the bivariate and univariate statistics surrounding this variable discovered above. We will also continue working with the same transformed data that we have been analyzing, as all of our findings thus far have been based on it.

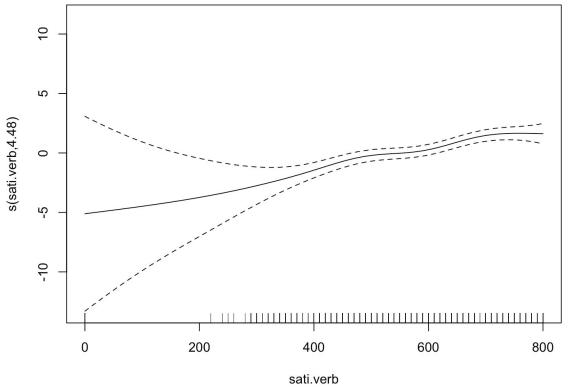
We observe the following statistics for various values of sp in our tuning process.

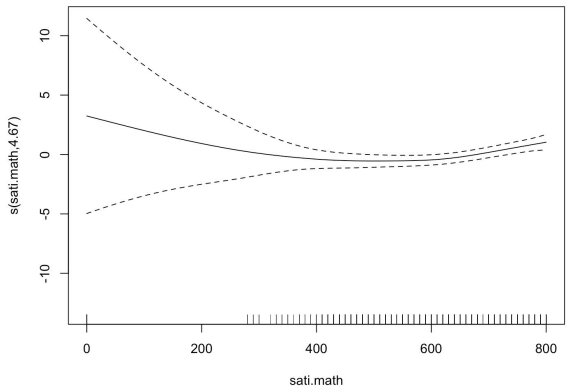
Value of sp	**Train Misclassification**	**Eval Misclassification**	**Deviance Explained**
0.01	0.148	0.159	41.4%
0.05	0.148	0.156	41.1%
0.1	0.149	0.157	40.9%
1	0.150	0.160	39.9%

From these metrics, we can see that a model with sp equal to 0.05 yields the best performance on the evaluation data. Therefore, we will proceed with this model.









(Top Left) *Figure 28*, (Top Right) *Figure 29,* (Bottom Right) *Figure 30*, (Bottom Left) *Figure 31*

We will now see what can be learned from this model's performance on the test data. We will first examine the partial dependence plots (Figures 28 to 31) of our smooth predictors. Income appears to have almost no relationship with admission (Figure 28). Though we did speculate earlier that income might have a relationship with admission, this should not come as too much of a surprise. It is a fact that more income allows a student more access to resources that would better prepare his or her college application. However, whether or not these resources are capitalized on depends entirely on the student at hand and is a factor that is not measured in the data. Thus, we cannot expect income to have such a cut and dry relationship with admission. GPA has a relationship with admission that steadily increases from around 2.0 until its maximum value (the range at which we have the highest density of observations) (Figure 29). The same observation can be made about the sati.verb score (Figure 30). These observations are not surprising. This is an elite university and so we would indeed expect there to be a relationship between higher quantitative metrics and admission. There are two different observations that require a bit of explanation, however. Firstly, the strength of the relationship between sati.verb and admission is much weaker than that between GPA and admission. We can interpret this as meaning that GPA is a much better metric to measure when determining the strength of an applicant. This might be because high SAT scores are quite common amongst applicants to elite institutions. They serve as much more of a "checkbox" on an application than something that sets an applicant apart from the rest. GPA, on the other hand, really is dependent on high school, as we noted above. For example, achieving a high GPA via AP courses at a high school in a relatively impoverished neighborhood is a far greater feat than is doing so at a high school in a well off neighborhood. Strong students with high potential but a lack of resources might not be able to score well on the SAT. However, they are able to stand out amid their high school peers, something that this university might take notice of. These claims are corroborated by the sati.math partial dependence plot (Figure 31). The relationship shown here is quite negligible. This is most likely due to the aforementioned reasons and also the fact that, as discussed above, it appears that performing well on the verbal section of the SAT is less common than acing the math section.

We will now look at the direction size of our regression coefficients, converting log-odds terms to odds for greater interpretability. It would appear that if a candidate is anglo or asian, his or her odds at admission are multiplied by 0.66 and 0.49, respectively. However, if a candidate is black, his or her odds at admission multiply by 2.15. Though we did acknowledge that black candidates were severely underrepresented in our data, we can attribute this to diversification efforts made by the university. It is well known that white and asian students are traditionally the biggest racial groups on college campuses, while black students are among the severe minority. These statistics may reflect efforts to mitigate this issue. Additionally, we can see that if a candidate is male, his odds at admission multiply by 0.75. This may also reflect efforts to balance gender inequalities on this campus.

Let us now take a look at the actual forecasting done on the test set. The misclassification on the test set was 0.157, which reflects very good overall performance of the model. However, we can see that if the model predicts an acceptance, the probability it is a false positive is around 25%. This confirms our worries from before that the model would focus in too much on the quantitative aspects of an application rather than the more qualitative ones. These false positives are likely attributed to students who had high GPAs and SAT scores but simply did not have compelling enough extracurriculars, essays, or interviews to be admitted. We can also see that if an applicant is predicted to be rejected, the probability it is a false negative is 17%. We can explain this by the reverse: even though these students did not have GPAs or SAT scores on par with those usually accepted, they had other factors that made up for it.

Our test set has a very large number of observations assumedly drawn from a joint probability distribution, enough to allow usage of the nonparametric bootstrap. This will allow us to compute an asymptotic estimate of the 95% confidence interval around estimates of misclassification in the approximation to the true response surface. We calculate proportions of misclassifications across 500 bootstrap samples. The estimates range from 0.1283 to 0.1774, and the 95% confidence interval here is 0.1373 to 0.1666. Even the maximum misclassification here is worlds better than that of the Bayes classifier.

Conclusion

The data here yielded rewarding results, despite the large quantity of missing data which may have skewed the results. We were able to observe several features of the predictors that set them apart. For example, it would appear that GPA plays a more important role in determining admission than SAT scores (this fact may relax the many high schoolers who are stressed over acing the SAT). Additionally, it seems that belonging to some racial categories might hurt one's chances at admission, while belonging to others may drastically help. An important observation that was quite surprising is that income seems to have almost no relationship with admission, implying there is a good amount of balance when it comes to considering financial facts in admission at this university. Though this model was able to forecast remarkably well, it is important to lay out a few concerns should another analysis of this kind be taken. First and foremost, there needs to be better quality of data because the missing values that were dropped or transformed may have affected the results. Secondly, there needs to be some sort of solution to the problem that GPAs are not the same for every high school, meaning that conclusions drawn from GPA should be taken with a grain of salt. Lastly, there should be more predictors that

encompass an applicant's overall application. These could include information such as scores on other tests, metrics about extracurriculars, interview strengths, etc. Hopefully doing so could remove some of the false positives and false negatives observed above. In conclusion, this was a useful analysis that can be used to inform the principal and future applicants at this particular high school.

Appendix

```
library(mgcv)
setwd('/Users/arjunlal/Desktop/STAT-474-HW-1')
load("/Users/arjunlal/Desktop/STAT-474-HW-1/admissions.rdata")
summary(admissions)
attach(admissions)

# dropping na in income
admissions <- subset(admissions, !is.na(admissions$income))
admissions <- subset(admissions, !is.na(admissions$sex))
subset(admissions, admissions$black == 0 & admissions$anglo == 0 & admissions$asian ==
0)
admissions$black <- as.numeric(admissions$black)
admissions$asian <- as.numeric(admissions$asian)
admissions$anglo <- as.numeric(admissions$anglo)
admissions$admit <- as.numeric(admissions$admit)

# recoding the other race as race_other
admissions$race_other <- 0
admissions$race_other[is.na(admissions$black) & is.na(admissions$anglo) &
is.na(admissions$asian)]<-1
admissions$race_other[(admissions$black == 0) & (admissions$anglo == 0) &
(admissions$asian == 0)]<-1
admissions$race_other <- as.numeric(admissions$race_other)
admissions$anglo[is.na(admissions$anglo)] <- 0 # repeat for the other races
admissions$asian[is.na(admissions$asian)] <- 0
admissions$black[is.na(admissions$black)] <- 0

sum(is.na(admissions$black))

# MAKE INTO FACTORS FIRST
admissions$admit <- as.factor(admissions$admit)
admissions$anglo <- as.factor(admissions$anglo)
admissions$asian <- as.factor(admissions$asian)
admissions$black <- as.factor(admissions$black)
admissions$sex <- as.factor(admissions$sex)
admissions$race_other <- as.factor(admissions$race_other)
admissions$income <- as.numeric(admissions$income)

# SECOND CONVERT FACTORS INTO THEIR ACTUAL NAMES
levels(admissions$admit)[1] <-"rejected"
levels(admissions$admit)[2] <-"accepted"
levels(admissions$anglo)[1] <-"non-anglo"
levels(admissions$anglo)[2] <-"anglo"
levels(admissions$asian)[1] <-"non-asian"
levels(admissions$asian)[2] <-"asian"
levels(admissions$black)[1] <-"non-black"
levels(admissions$black)[2] <-"black"
levels(admissions$sex)[1] <-"female"
levels(admissions$sex)[2] <-"male"
levels(admissions$race_other)[1] <-"race-accounted"
```

```
levels(admissions$race_other)[2] <-"other-race"

attach(admissions)
summary(admisisons)
hist(income, breaks = 50, main = "Income Breakdown", xlab = "Income", ylab =
"Frequency")
hist(gpa.wtd, breaks = 50, main = "GPA Breakdown", xlab = "GPA", ylab = "Frequency")
hist(sati.verb, breaks = 50, main = "SAT Verbal Breakdown", xlab = "Verbal Score",
ylab = "Frequency")
hist(sati.math, breaks = 50, main = "SAT Math Breakdown", xlab = "Math Score", ylab =
"Frequency")

sum((asian == "asian") & (black == "black"))
sum((asian == "asian") & (anglo == "anglo"))
sum((black == "black") & (anglo == "anglo"))

plot(sex, asian)
plot(sex, anglo)
plot(sex, black)
plot(sex, race_other)

sum((black == "black") & (sex == "male"))
sum((black == "black") & (sex == "female"))
sum((anglo == "anglo") & (sex == "male"))
sum((anglo == "anglo") & (sex == "female"))

plot(sex, income)
plot(asian, income)
plot(black, income)
plot(anglo, income)
plot(race_other, income)
plot(sex, gpa.wtd, ylab = "GPA")
plot(asian, gpa.wtd, ylab = "GPA")
plot(black, gpa.wtd, ylab = "GPA")
plot(anglo, gpa.wtd, ylab = "GPA")
plot(race_other, gpa.wtd, ylab = "GPA")
plot(sex, sati.verb, ylab = "SAT Verbal")
plot(asian, sati.verb, ylab = "SAT Verbal")
plot(black, sati.verb, ylab = "SAT Verbal")
plot(anglo, sati.verb, ylab = "SAT Verbal")
plot(race_other, sati.verb, ylab = "SAT Verbal")
plot(sex, sati.math, ylab = "SAT Math")
plot(asian, sati.math, ylab = "SAT Math")
plot(black, sati.math, ylab = "SAT Math")
plot(anglo, sati.math, ylab = "SAT Math")
plot(race_other, sati.math, ylab = "SAT Math")
plot(gpa.wtd, sati.verb)
cor(gpa.wtd, sati.verb)
plot(gpa.wtd, sati.math)
cor(gpa.wtd, sati.math)
plot(income, gpa.wtd)
cor(income, gpa.wtd)
plot(income, sati.math)
plot(income, sati.verb)
plot(admit, asian)
plot(admit, black)
plot(admit, anglo)
plot(admit, race_other)
sum((admit == "admitted") & (race_other == "other-race"))
sum((admit == "admitted") & (race_other == "race-accounted"))
sum((admit == "admitted") & (anglo == "anglo"))
```

```
sum((admit == "admitted") & (anglo == "non-anglo"))
sum((admit == "admitted") & (asian == "asian"))
sum((admit == "admitted") & (asian == "non-asian"))
plot(admit, sex)
plot(admit, income)
plot(admit, gpa.wtd)
plot(admit, sati.verb)
plot(admit, sati.math)

# GAM

# Construct 3 random disjoint splits
index<-sample(1:6965,6965,replace=F) # shuffle row numbers
temp2<-admissions[index,] # put in random number
Train<-temp2[1:2321,] #training data
Eval<-temp2[2322:4642,] #evaluation data
Test<-temp2[4643:6965,] #test data
# SP EQUAL TO 0.01
out1<-gam(admit~s(income,sp=0.01)+s(gpa.wtd,sp=0.01)
+s(sati.verb,sp=0.01)+s(sati.math,sp=0.01)+anglo+asian+black+sex+race_other,data=Train
,family=binomial)
summary(out1)
plot(out1)
Tab<-table(out1$fitted.values>.5,
           Train$admit) # Confusion table
(Tab[1,2]+Tab[2,1])/sum(Tab) # Proportion Misclassified
#Evaluate
predsE=predict(out1,Eval,type="response")
Tab<-table(predsE>.5,Eval$admit) #Confusion Table
Tab
(Tab[1,2]+Tab[2,1])/sum(Tab)
# SP EQUAL TO 0.05
out.5<-gam(admit~s(income,sp=0.05)+s(gpa.wtd,sp=0.05)
+s(sati.verb,sp=0.05)+s(sati.math,sp=0.05)+anglo+asian+black+sex,data=Train,family=bin
omial)
summary(out.5)
plot(out.5)
Tab<-table(out.5$fitted.values>.5,
           Train$admit) # Confusion table
(Tab[1,2]+Tab[2,1])/sum(Tab) # Proportion Misclassified

#Evaluate
predsE=predict(out.5,Eval,type="response")
Tab<-table(predsE>.5,Eval$admit) #Confusion Table
Tab
(Tab[1,2]+Tab[2,1])/sum(Tab)
# SP EQUAL TO 0.1
out2<-gam(admit~s(income,sp=.1)+s(gpa.wtd,sp=.1)
+s(sati.verb,sp=.1)+s(sati.math,sp=.1)+anglo+asian+black+race_other+sex,data=Train,fam
ily=binomial)
summary(out2)
plot(out2)
Tab<-table(out2$fitted.values>.5,
           Train$admit) # Confusion table
(Tab[1,1]+Tab[2,2])/sum(Tab) # Proportion Misclassified
#Evaluate
predsE=predict(out2,Eval,type="response")
Tab<-table(predsE>.5,Eval$admit) #Confusion Table
Tab
(Tab[1,1]+Tab[2,2])/sum(Tab)
# SP EQUAL TO 1
```

```
out3<-gam(admit~s(income,sp=1)+s(gpa.wtd,sp=1)
+s(sati.verb,sp=1)+s(sati.math,sp=1)+anglo+asian+black+race_other+sex,data=Train,famil
y=binomial)
summary(out3)
plot(out3)
Tab<-table(out3$fitted.values>.5,
           Train$admit) # Confusion table
(Tab[1,1]+Tab[2,2])/sum(Tab) # Proportion Misclassified
#Evaluate
predsE=predict(out3,Eval,type="response")
Tab<-table(predsE>.5,Eval$admit) #Confusion Table
Tab
(Tab[1,1]+Tab[2,2])/sum(Tab)
#Estimate using Test data with tuning paramaters determined
out4<-gam(admit~s(income,sp=.05)+s(gpa.wtd,sp=.05)
+s(sati.verb,sp=.05)+s(sati.math,sp=.05)+anglo+asian+black+race_other+sex,data=Test,fa
mily=binomial)
summary(out4)
plot(out4)
predsT=predict(out4,Test,type="response")
Tab<-table(predsT>.5, Test$admit) #Confusion Table
Tab
(Tab[1,2]+Tab[2,1])/sum(Tab) #Generalization error
# Bootstrap to get 95% CI for generalization error
Error<-rep(NA, 500)
for(i in 1:500)
{
  index<-sample(1:2323,2323,replace=T)
  TestB<-Test[index,]
  predsB=predict(out4,TestB,type="response")
  Tab<-table(predsB>.5,TestB$admit) #Confusion table
  Error[i]<-(Tab[1,2]+Tab[2,1])/sum(Tab)
}
summary(Error)
hist(Error,breaks=20)
quantile(Error,probs=c(.025,.975))
```