Predicting College Admissions: The Counsellor's Game

Victor Cheung

March 21, 2016

1 Introduction

In a country obsessed with educational attainment, attending college is a hallmark of accomplishment and a college degree an important milestone.

There is also a significant body of economic research that demonstrates the value of an elite college education and of the private gains thereof. The signaling effect to employers due to the specific alma mater implies that a graduate of an elite college is more likely to score interviews, since the candidate has already been pre-screened by the college.

However, applying to college is also a costly process. Money spent on college counseling and standardized test prep, effort and time on writing essays, and the opportunity costs thereof, can all add up quickly. It would be useful if we could anticipate college admission decisions, and have only students who are likely to be admitted apply. But first, we must find factors that strongly influence the admissions decision.

College admissions has been extensively studied from a sociological and historical perspective. A brief survey of the subject brings to the forefront much speculation with regards to how college choose from amongst its applicants. Willingham et al.

published a study in 1982 that investigated the importance of personal and academic factors for a number of selective liberal arts colleges, ranging from Williams to Colgate. They found that academic factors far outweigh personal ones; however, certain personal traits do associate weakly with admissions factors, including minority status and affiliation with the college. There is no indication in this study that higher socioeconomic classes receive preferential treatment in the admissions process.

We will investigate the same question armed with an extensive dataset from an elite college. In doing so, we will investigate the critical factors behind the admissions decision. We will build a forecasting model to evaluate the likelihood of admission by this college of future students.

Section 2 will provide a description of the data collected for applicants for this elite college for the past cycle, including univariate and bivariate statistics. Section 3 discusses problems associated with the data and how we attempted to resolve or at least ameliorate those problems. Section 4 discusses the validity of our forecasting model based on assumptions about the underlying data generation process, as well as why we emphasize models that encourage students to apply. Section 5 discusses the implications of our forecasting model, with coverage on the conditional probabilities of acceptance and the key factors in admissions. We will conclude the paper in section 6. In the appendix, we provide the `R` code for documentation and for duplicability of our results.

2 Data

Our data is for one university, presumably with an admissions rate and other factors that colloquially classify it as being "elite". We have data for applications during the past year, containing 8700 observations along with 9 variables. Our unit of

analysis is an applicant, or student, each with the following measurements.

Admit indicates whether an applicant was rejected (1) or rejected (0).

Anglo indicates whether an applicant was Anglo-Saxon (1) or not (0).

Asian indicates whether an applicant was Asian (1) or not (0).

Black indicates whether an applicant was black (1) or not (0).

GPA.weighted is the numeric measure of high school GPA weighted by AP courses. This means that AP courses are counted moreso than normal classes, such that the GPA can exceed some nominal bound, for example, a 5.0/4.0.

Sati.Verb is the applicant's SAT I verbal score, out of 800.

Sati.Math is the applicant's SAT I math score, out of 800.

Income is the applicant's household income. We note that all household income above $100,000 is binned at $99,999.

Sex is the applicant's sex, either male (1) or female (0).

2.1 Univariate Analysis

Immediately, we note there is missing data across the board. We will discuss the implications in the section "Problems with Data".

The summary statistics provides interesting information on the distribution of race and sex. 35.7% of applicants who reported race are anglo-saxon, 43.4% are Asian, and only 4.5% are black. As well, we see that only 46.6% of applicants are male. The mean and median of income is extremely misleading - the binning decision will drastically lower these measures of central tendency. We note also that

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
admit	8,700	0.309	0.462	0	0	1
anglo	7,871	0.357	0.479	0	0	1
asian	7,871	0.434	0.496	0	0	1
black	7,871	0.045	0.208	0	0	1
gpa.wtd	8,700	3.790	0.523	0.000	3.850	4.950
sati.verb	8,700	554.936	163.209	0	580	800
sati.math	8,700	592.276	168.476	0	630	800
income	6,976	63,245.140	32,826.860	120	65,000	99,999
sex	8,676	0.466	0.499	0	0	1

the average gpa is a 3.790, the average sat verbal score is 555 and the average sat math score is 592.

We also see that the admissions rate for this school is 30.9%, significantly higher than the most selective universities in the United States.

Consider now the histograms.

Figure 1: Histograms for gpa, sat and income

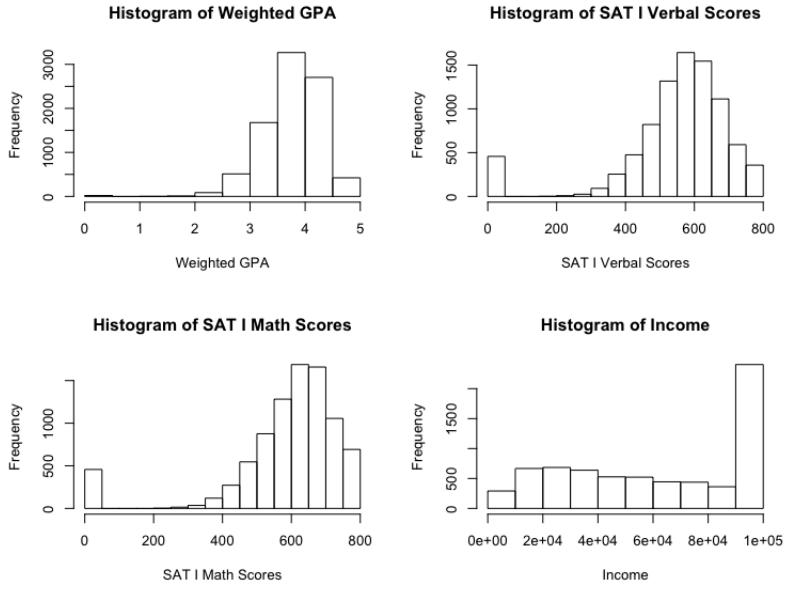


Figure 2: Histograms for Distribution of Admissions and Race

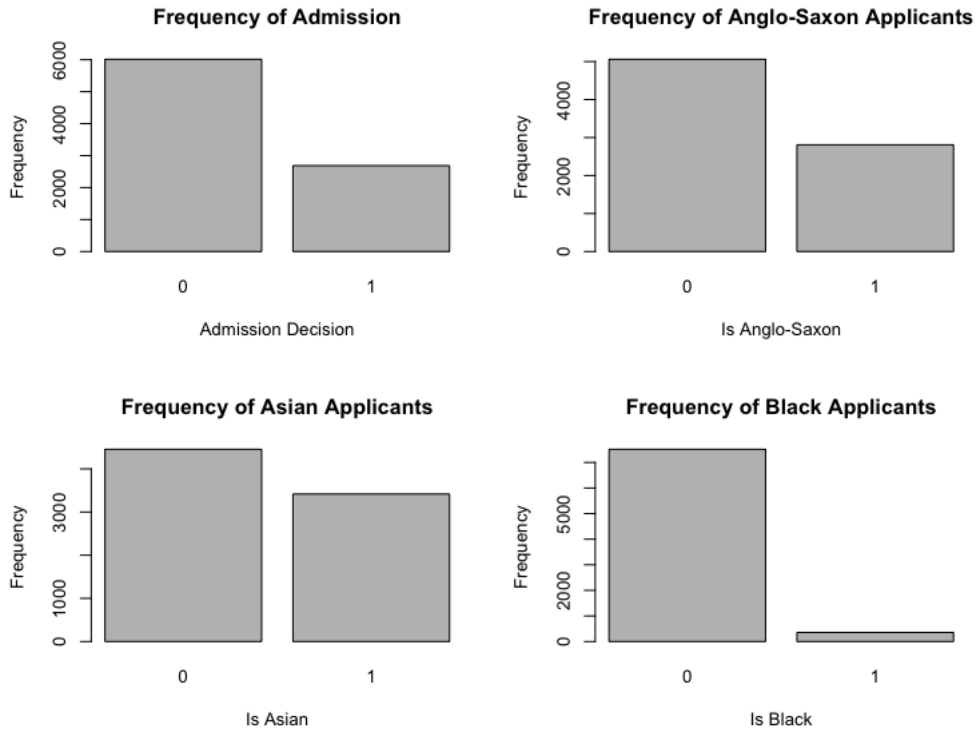
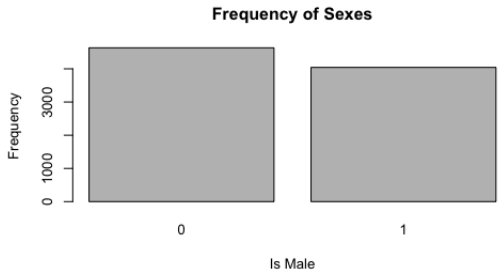


Figure 3: Histogram of Distribution of Sex



Weighted GPA GPA follows a left skewed distribution with a with median centered at 3.850. This GPA seems high on a 4.0 scale, but we recall it's weighted and thus 5.0 is the highest possible GPA. The distribution of GPAs also comports with our understanding of the distribution of GPAs. There are few students with extremely high GPAs, and a long left tail across lower GPAs. We note also that there are 19 students with 0 GPA.

SAT I Verbal Scores SAT I verbal scores follow a left skewed distribution with median at 580. The distribution conforms with our understanding of the distribution of test scores and there are values between all permissible scores between 200 and 800. We note there are 457 values at 0. This is most likely an artifact with data entry - a score of 0 is impossible since the minimum possible score is 200. Hence, 0 might have been entered for applicants who did not submit SAT I scores.

SAT I Math Scores SAT I math scores follow a left skewed distribution with median at 630. This histogram conforms with our understanding of the distribution of test scores and provides a good number of data points between the possible values of SAT I math scores between 200 and 800. We note that there are 457 values of SAT I math scores at 0. This implies there are 457 students who did not submit SAT I scores overall.

Income Income is the most problematic piece. Income is almost uniformly distributed amongst the bins everywhere below $100,000$. Indeed, a quick count shows that there are 2164 values of income at $99,999$. This suggests a very large tail to the right that has been cut off. This binning will affect our results by imposing artificial sparsity - the effects of extremely high income on admissions decisions cannot be captured. The income data is thus missing

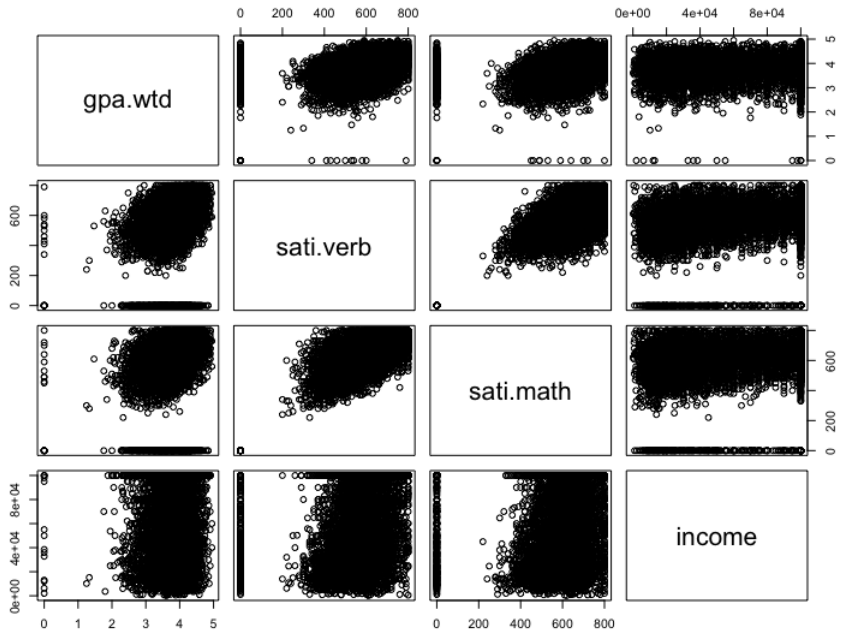
systematically.

Race Applicants are split into three different ethnic groups (or chose not to identify), of which Asians and Anglo-Saxons are the most frequent, with each at about 3000, and blacks the least, at less than 500. This comports with the proportions of applicants to elite colleges in the United States. [1]

Sex Applicants are roughly split equally between the two sexes. The slightly increased prevalence of female applicants comports with the trend of male underachievement against women in college education in the United States. [2]

2.2 Bivariate Statistics

Figure 4: Pairwise Correlations



[1] https://college.harvard.edu/admissions/admissions-statistics
[2] http://opinionator.blogs.nytimes.com/2013/02/02/the-boys-at-the-back/

An analysis of pairwise correlations shows the expected associations between gpa and sat scores. We see that, with the exception of entries for zero gpa or zero sat scores, that there is a clear positive correlation between sat math scores and gpa, and between sat verbal scores and gpa. There is also the obvious positive correlation between sat math and sat verbal scores. This shouldn't be a surprise. However, surprisingly there appears to be no correlation between income and sat scores or with GPA. The scatterplots are spread uniformly across the range of income, with bunching at $100,000 due to unfortunate binning decisions.

3 Problems with Data

In addition, brief look at the data in viewer as well as the summary statistics table indicates the presence of missing data.

Figure 5: Histograms for Observations with Missing Race

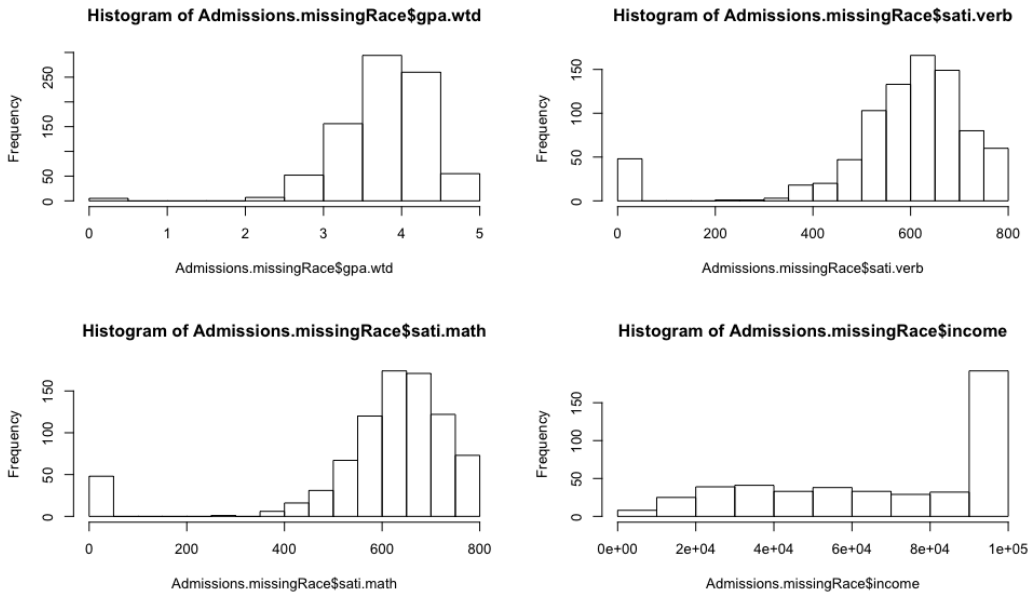


Figure 6: Histograms for Observations with Missing Income

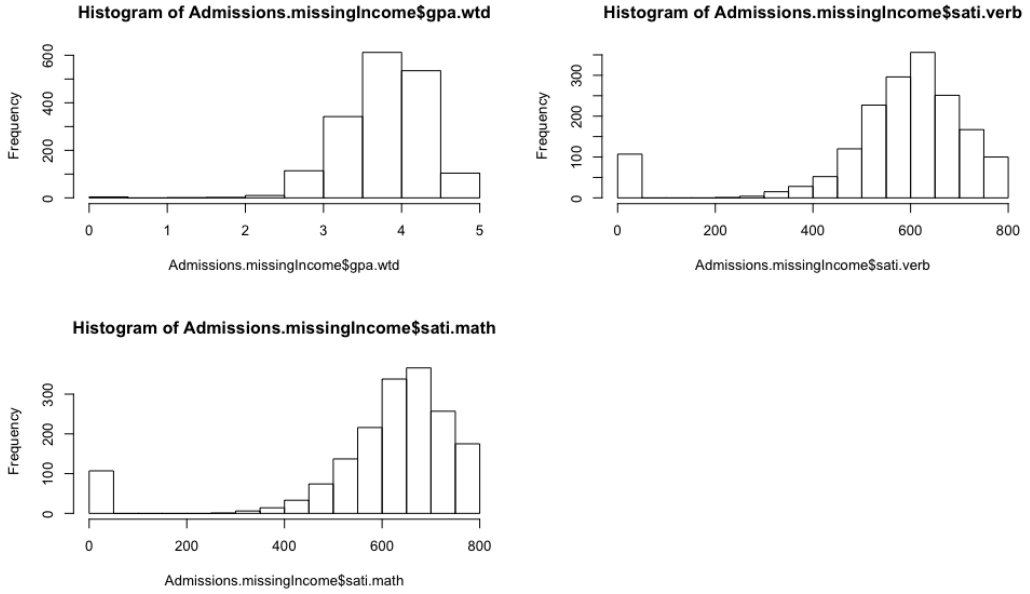


Figure 7: Histograms for Observations with Missing Sex

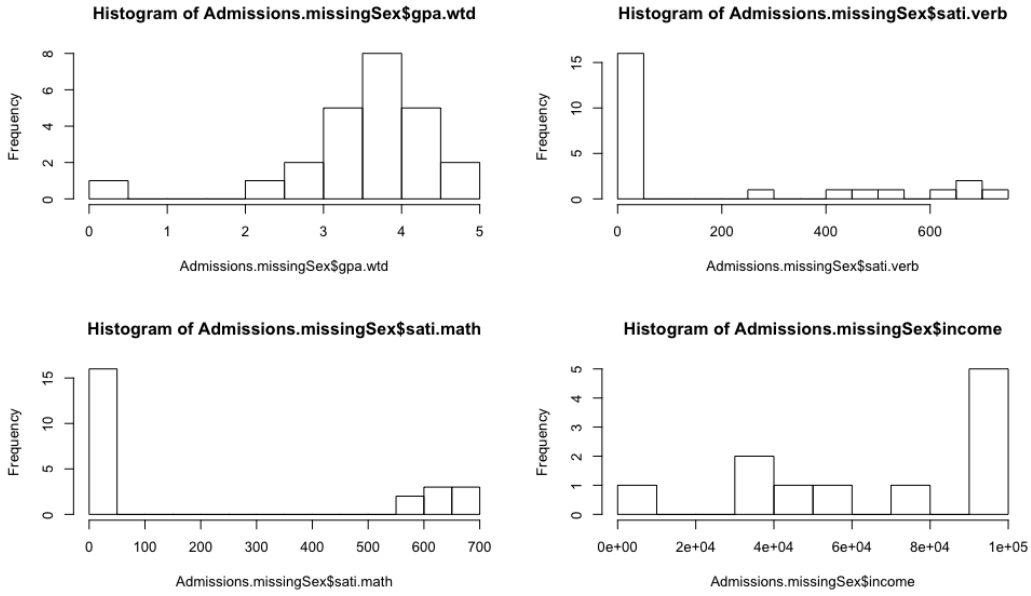


Figure 8: Histograms for Observations with Zero GPA

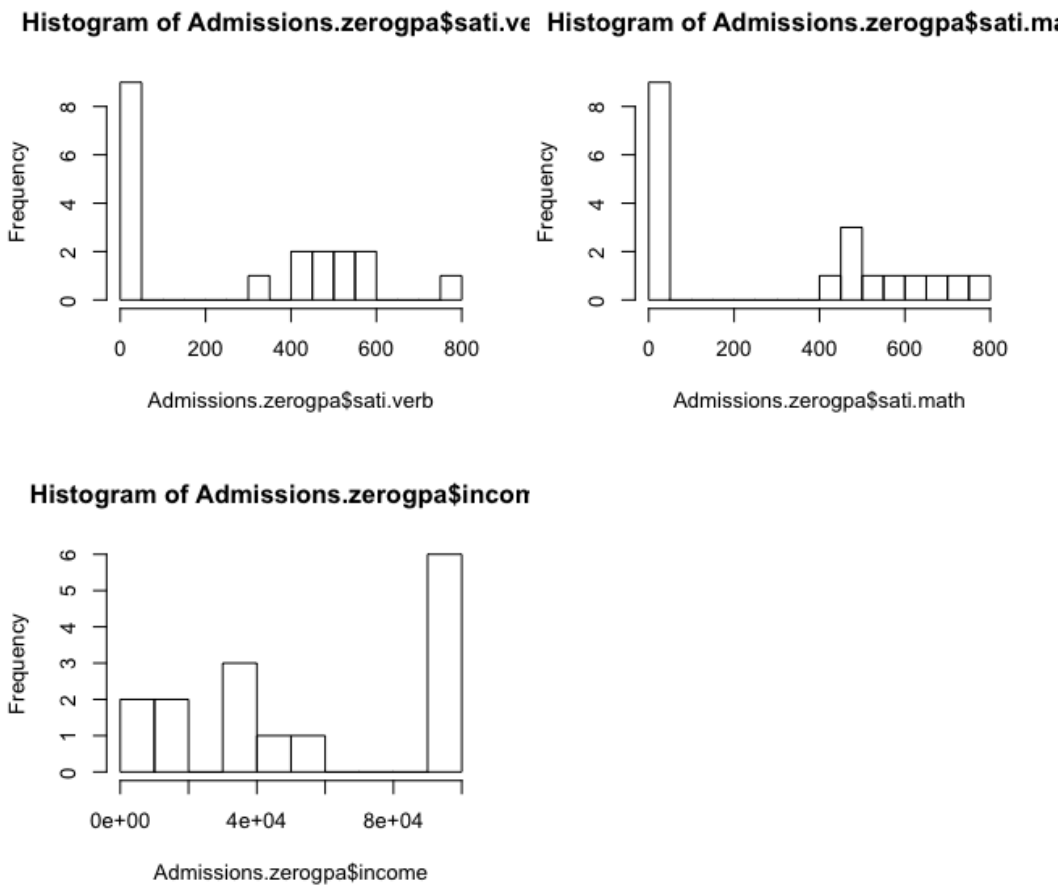
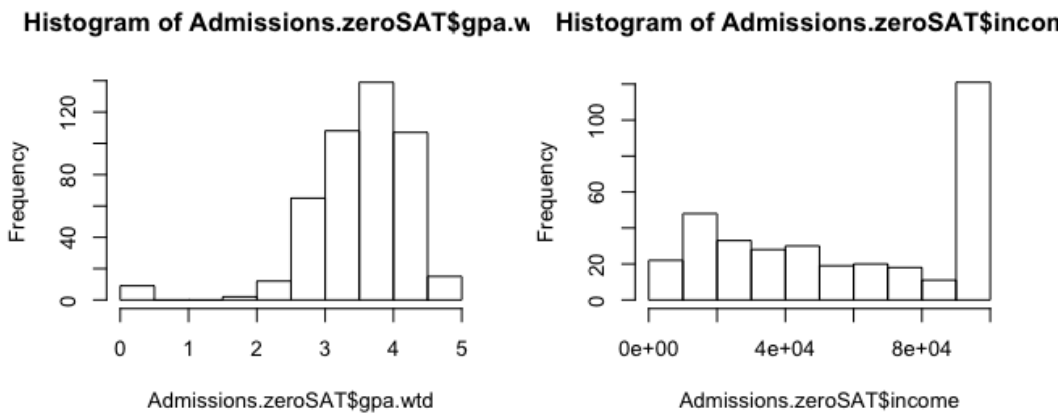


Figure 9: Histograms for Observations with Zero SAT



In general, we will assume where there is missing data or data is assumed be zero (gpa, sat), this is due to the applicant, not record keeping errors by the college.

Consider first income. Beyond the systematically missing data for income beyond $100,000, there are 1724 missing income values. This is likely due to general reluctance in disclosing salaries. From figure 3, the pairwise correlations show that no variables are strongly correlated with income regression to fill in the blanks is out of the picture. Consider Figure 6. Conditioned on missing income, the distributions of gpa and SAT scores are largely the same as in figure 1, the histograms for all data. This implies that those observations missing income are not systematically different from the rest of the observations. We would be safe in deleting observations with missing data; it's as if the school had randomly sampled the database and deleted entries.

Consider now SAT I scores. We effectively have missing data here. The students should have SAT I scores but they either did not submit it on time or did not take it. Moreover, we do not know if the college requires SAT I scores or not. Some colleges have recently moved to make standard testing optional, and that might be the case here. This implies that personal traits, recommendation letters and admissions essays play critical roles in the admissions decision - information that we do not have; however, since our college is "elite", we reasonably expect that SAT I scores are required and that these applicants were unclear on the application requirements. From figure 8, we see that for those with zero SAT scores, the distributions of gpa and of income are largely the same as the data as a whole. This suggests that the people who did not submit sat scores are not systematically different from everyone else. Again, we would then be safe in deleting observations with missing data.

Consider now race and sex data. In 1978 the Supreme Court ruled in Bakke v. Regents that public universities cannot have specific racial quotas when admitting

students but can use them when considering "goals" for a class. Private universities may still consider race when admitting. However, the Common Application makes the question optional, and students are not required to report it to the college. We believe that this is the case here. Consider figure 4. Again the distributions are largely similar to that in figure 1. This suggests that students who chose not to identify themselves as a specific race or are not systematically different from everyone else. Consider figure 6. Since there are only 24 applicants out of 8700 who chose not to identify as a specific sex, these histograms are unreliable. Nonetheless, we see that the distributions of gpa and of income are approximately the same as in figure 1. However, the distributions of sat scores are abnormal. There are far more observations with zero sat scores than would be suggested from figure 1. As well, examining the pattern of other missing data for observations missing sex shows that the prevalence of missing race and income to be abnormally high as well. This suggests a systematic pattern of missing data. Nonetheless, since there are so few observations that are missing sex, removing these few points from our thousands of observations should not dramatically affect our models.

Consider finally gpa. From figure 7 we see that those missing gpa also tend be missing gpa more often than average; income is roughly the same. This suggests data is systematically missing. Again, there are very few observations missing gpa.

Given the above, we must very careful then when applying CART from `rpart` and understand the default options for treating NAs. Where there is missing data, CART uses available data to decide on the primary split. Then, it constructs a set number of surrogate split decisions (default is 5 for `rpart`) when classifying a specific observation with a missing value for the primary split criterion; it goes through each of the surrogates until the observation has a value for one of them, at which point it's classified. Otherwise, `rpart` uses majority rule to decide on the

classification. However, this procedure is ad-hoc and arbitrary; we do not believe there is evidence that, for example, gpa and sat scores are good replacement criteria for income. There is little correlation between gpa and income or sat scores and income, and income may be important if the school doesn't allow for much financial aid.

Table 2: Summary of Deletion Decisions for Missing Data

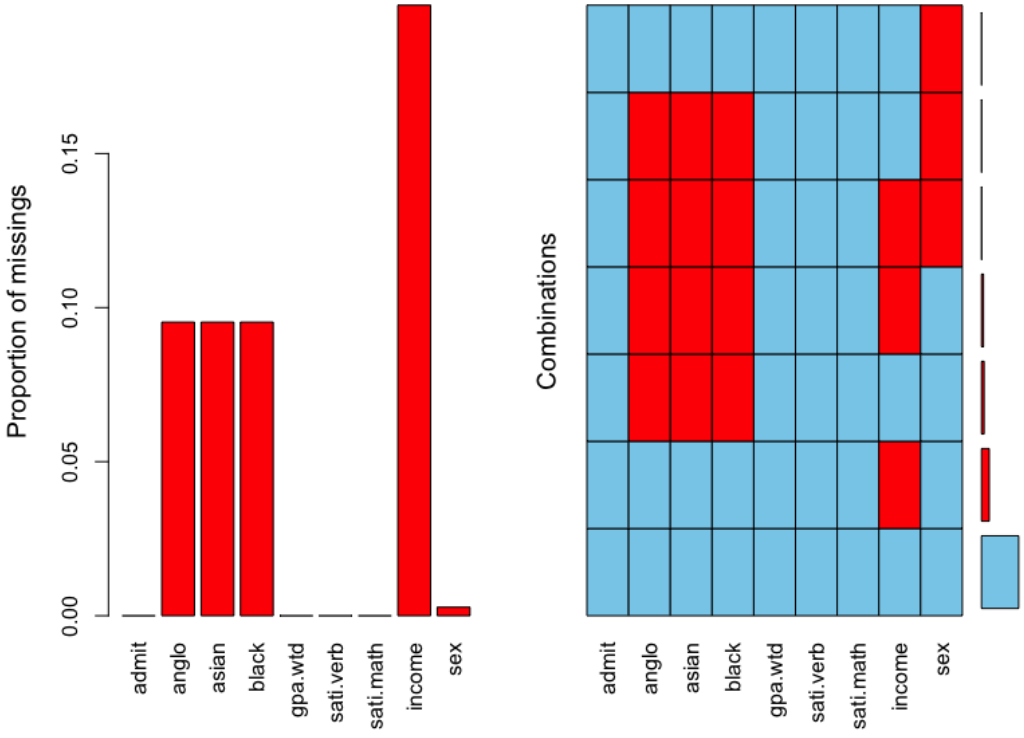
Statistic	Nature of Missingness	Delete Decision	Justification
income	Binning	No	Systematically missing and very many
income	NA	Yes	Not systematically missing data
race	NA	Yes	Not systematically missing data
sex	NA	Yes	Systematically missing but very few
gpa	0	Yes	Systematically missing but very few
sat scores	0	Yes	Not systematically missing data

4 Model Building

Since we have a relative large number of observations, we will proceed thus. From examination of the missing data, it appears that data points are missing largely completely randomly, despite the fact that students have opted not to provide race, sex or income. This is reflected in figure 10, which shows the prevalence of missing data for each variable, as much as the combinations of variables.

Thus, we will list wise delete all rows missing income, race, sex and gpa. The former appear as if to be randomly sampled from the data. The latter are justified by the fact that very few observations are actually missing sex and gpa, and hence

Figure 10: Histograms



deletion should not overly impact our models given our large sample size.

We will then partition the pre-processed data into three equal size random disjoint sets. This is for training, evaluation and testing purposes. We choose equal size partitions because our dataset is sufficiently large that sparsity is not an issue. And since we are building a forecasting model it's critical that our evaluation data gives us good feedback on the effects of tuning parameters, and that our testing data gives us a very honest assessment of generalization error. This requires that those datasets also be large. We are deleting close to 2700 observations, but given we started with 8700, this seems justified. An even split seems fair.

4.1 Justifying Level II

We must justify conducting a level II analysis since our goal is explicitly to construct a forecasting model. We want to estimate a tree based approximation to the true response surface with a focus on correct fitted values (classifications).

The joint probability distribution for admissions decisions draws on a well-defined and finite population of high school seniors. However, our sample was not randomly drawn from the population. Indeed, for our college, and for all colleges in generally, applicants are self-selected. For example, applicants to MIT and applicants to a state University probably differ in measurable factors such as SAT scores or GPA, and other nonmeasured factors. Thus, our sample of high school seniors for this "elite" university will not be representative of all high school seniors. Fortunately, future students who self-select to apply to this university will likely be similar to past students. We are thus justified in generalizing to the population of high school seniors who would self-select to apply to this elite college. We note this is a limited population and will present problems when applied to our high school students, as not all students will necessarily belong in to this population,

invalidating this necessary assumption for forecasting.

Another consideration is that the realizations of the response may not be independent. Interviews conducted with admissions officers indicate they seek to build "well-rounded" classes. For example, officers avoid having too many trombone players in a particular class. They want an orchestra instead. Therefore, the admission of one particular student may preclude the admission of another student in a way that is not tracked by our data.

The process underlying the admissions decision will most likely stay the same for the foreseeable future barring a drastic change in admissions policy either due to university policy, state or federal legislation or judicial challenges.

4.2 Procedure and Relative Costs

As CART conducts extensive data snooping as it relies on step functions. We therefore partition the data into three parts to prevent over-fitting: training data, evaluation data and testing data. We will train our model on the training data, tune `cp` and `priors` with the evaluation data to obtain an interpretable tree with the target relative costs, and finally obtain an honest performance estimate with the testing data. We will use confusion tables and performance measures therein for tuning and model selection.

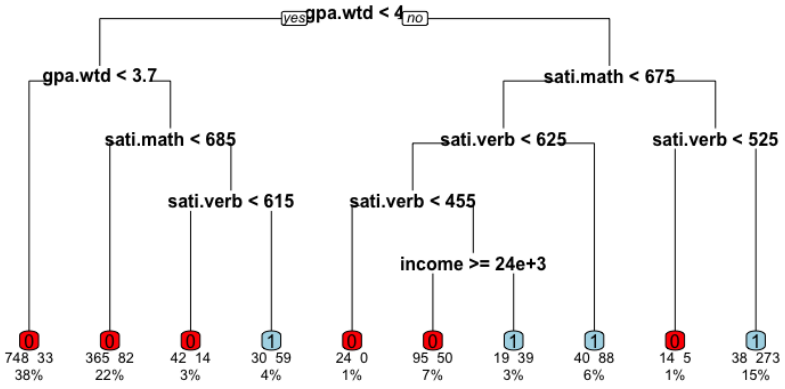
For relative costs, we assume that we are on the side of the student. The school has sufficient resources such that it does not need to prioritize between students. Assuming students consider the predictions from our model, i.e. that a predicted admit from our model will encourage them to apply and a predicted deny from our model will discourage them, we must then relax the statistical evidence necessary to classify a student as an admit. This does not mean that all students should apply to this elite college, or elite colleges in general. For students with drastically

unfavorable profiles and whose chance of admissions is very low, the opportunity costs of time and application fees are significant. Furthermore, there is an emotional cost associated with expecting to be admitted only to be denied. The costs to self-esteem may be lifelong and students may suffer from mental health issues as a result. These students should not apply. Model performance is therefore important.

Given the above, we allow that discouraging students from applying is more costly than encouraging students to apply. Thus, we want the cost of a false negative (predicted no admission when student is actually admitted) be five times the cost of a false positive (predict admission when student is actually denied). Thus, false positives should occur fives times more frequently than false negatives.

4.3 Models

Figure 11: Tree One with Symmetric Costs, cp = 0.01



For each terminal node, a zero (in red) indicates a reject classification. A one in

Table 3: Confusion Matrix for Tree One, Observed Relative Cost = 0.66

	Predicted Deny	Predicted Admit	Model Error
Denied	1259	153	10.8%
Admitted	230	417	35.5%
Use Error	15.4%	26.8%	Overall Error = 18.6%

blue indicates an admit classification.

Our first attempt shows an observed cost ratio of 0.66 with a very complex tree. SAT verbal and math scores show up multiple times at different levels of the tree, suggesting it's critical to the decision. However, interpretation is difficult. Furthermore, the number of observations at some terminal nodes, for example for the node for students with greater than 4 gpa, greater than 675 SAT math scores and less than 525 SAT verbal scores, has only 19 observations out of 2058.

We therefore step through different values of `prior` to hit our targeted relative cost of three and increase the complexity parameter in order to penalize for greater model complexity. We did this 10 times, using the evaluation data partition each time to construct a confusion table. The exact steps taken is available in the `R` code in the appendix. Our final tree is below.

Table 4: Confusion Matrix for Final Tree, Observed Relative Cost = 4.89

	Predicted Deny	Predicted Admit	Model Error
Denied	981	431	30.5%
Admitted	88	559	13.6%
Use Error	8.2%	43.5%	Overall Error = 25.2%

This last table is an asymptotically unbiased estimate of the population confusion table approximation.

Figure 12: Final Tree with Evaluation Data, cp = 0.01

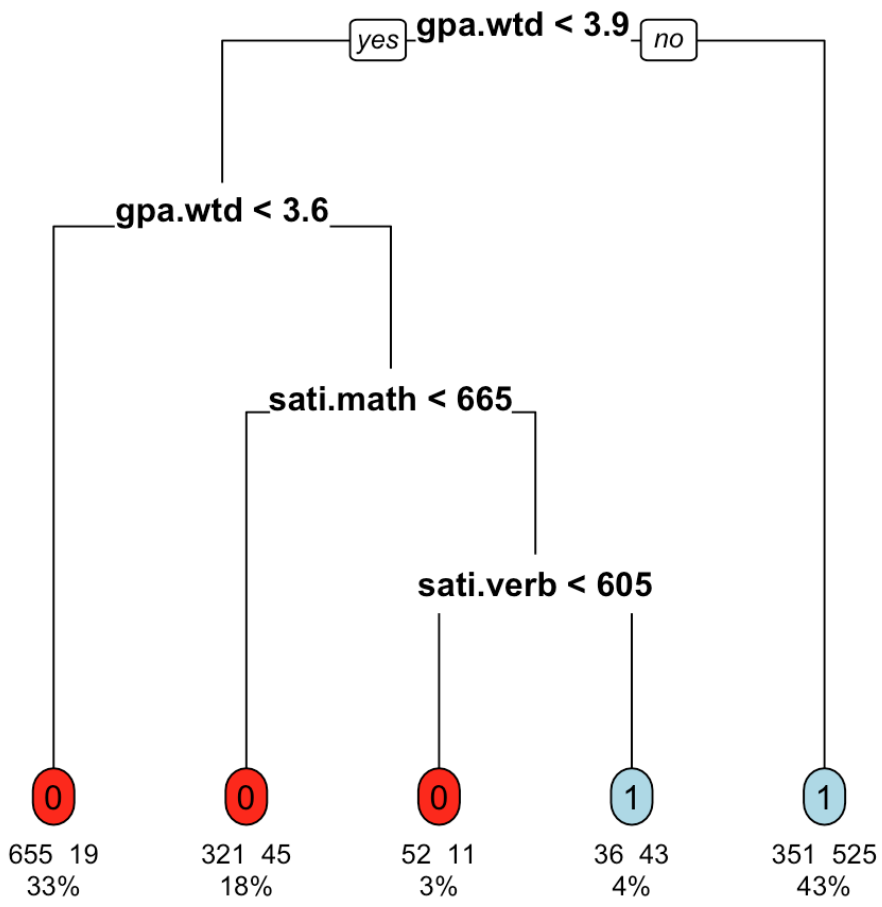


Table 5: Final Tree with Test Data, Observed Relative Cost = 5.53

	Predicted Deny	Predicted Admit	Model Error
Denied	975	448	31.5%
Admitted	81	554	12.7%
Use Error	76.7%	44.7%	Generalization Error = 25.7%

5 Discussion

Our final tree is relatively simple and interpretable. Students with low GPAs and low SAT scores are likely to be denied, and students with high GPAs and high SAT scores are likely to be admitted. Each terminal node has a reasonable number of observations, with the smallest terminal node containing 3% of the observations in the evaluation data partition. The conditional probability of admissions for each of terminal nodes are as follows

1. GPA < 3.9, GPA < 3.6: Admissions probability 2.8%.
2. GPA < 3.9, GPA > 3.6, SAT math < 665: Admissions probability 12.3%.
3. GPA < 3.9, GPA > 3.6, SAT math > 665, SAT verb < 605: Admissions probability 17.5%.
4. GPA < 3.9, GPA > 3.6, SAT math > 665, SAT verb > 605: Admissions probability 45.6%.
5. GPA > 3.9: Admissions probability 59.9%.

Dropping the test data down our final tree, we find that the tree does a decent job on the test data partition. This provides us with an estimate of the out of sample performance by simulating new realizations from the original joint probability distribution. The news is good. We find that the various performance measures have remained largely stable from the evaluation data. Observed relative cost has increased from 4.89 from 5.53, and overall error rate has stayed about the same, changing from 25.2% to 25.7%. We note that model errors and use errors have remained largely similar throughout as well, in Table 5 and Table 4.

As well, compare the confusion matrices (table 3, table 4) for the initial tree with symmetric cost assumptions and the final tree with asymmetric costs. We see

that as we increased the relative cost of a false negative, that our model has relaxed the statistical evidence necessary to classify someone as an admit, and thus that overall error has increased. This is to be expected. When the misclassification costs are drastically different, as in this case, we are willing to accept more forecasting errors on average in order to capture an increased number of potential admits. This is precisely what we have done. The number of admitted students captured have increased by close to 140 from the initial tree to our final tree.

We note the absence of race, income and sex from our model. Insofar as our elite college makes admissions decisions in a similar way to other colleges, which would comport with our expectation, this contradicts popular conception elite colleges discriminate against certain races in admissions, including a current lawsuit launched against the University of Texas for their use of race in admissions decisions. [3]

We note that main effects account for only one terminal node, whereas the other terminal nodes depend on interaction effects between gpa, sat verbal and sat math.

Our tree shows that weighted GPA is a tremendous lever in the admissions decision. It is the only main effect in the tree, and classifies those with weighted gpa below 3.6, equivalent to saying those with excellent GPAs but who have not taken AP classes and done well, to be denied. At the same time, those with high GPAs above 3.9 are likely to be admitted right away, without reference to SAT scores. Furthermore, even where students had course loads sufficiently heavy with AP courses and in which they did well, if they did not score high enough on SAT math (below 665), they are likely to be denied. Then, even if they did do well on SAT math, if they did poorly (below 605) on SAT verbal, then they are also likely to be denied. Finally, high SAT scores may save students from the perils of a low GPA only if their math is above 665 and their verbal is above 605.

[3] http://www.heritage.org/research/reports/2015/12/discriminatory-racial-preferences-in-college-admissions-return-to-the-supreme-court-fisher-v-university-of-texas-at-austin

SAT scores only affect the decision through interaction effects together with GPA. Indeed, its subordinate role to GPA reinforces the primacy of GPA in the determinants of the admissions decision. Applicants with high GPA and low SAT scores however, will still find themselves denied. As well, the SAT scores interact with each other: admissions officers ding applicants who do not score highly on both portions of the SAT.

6 Conclusion

Our main findings are thus. The predictor of utmost consequence is GPA. Without it, there is nothing to be done regarding admissions to this college. Even with a high enough GPA, applicants must look for high SAT math and verbal scores to maximize their likelihood of admissions. Otherwise, the likely decision is still a deny. Notably, factors such as sex, income and race are missing in whole from the tree. They do not play a statistically significant role in the decision, according to our model.

We also found and dealt with persistent patterns of missing data through listwise deletion. Consideration of the use of surrogates and its implications deterred us from allowing for NAs in the data. We extensively characterized the nature of missing data, and found that for income and race that data was missing as if randomly. We found that for sex data was missing systematically but there were very few missing data points. We found for observations with zero SAT scores that the data was missing as if randomly, and that those with zero GPA had data that was missing systematically but there were very dew of these. Finally, we acknowledge that binning decision caused systematic deletion of income beyond $100,000 and that we could not resolve this.

Our findings are confined to this one college. While elite colleges likely make

decisions in a similar way, there will still be differences in the process. Some elite colleges may prefer football players over physics whiz, and others pianists over trombone players. If the underlying admissions process is different, then the joint probability distribution will be different as well. Hence, we warn that model will likely not generalize well for the admissions decisions of other colleges or even elite colleges.

For this one college, we would also like to keep track of forecasting accuracy over time in order to assess shifting trends in the admissions process. For example, a comparison of forecasting accuracy five years from now vs. this year could provide insight into the stationarity of the college admissions process or the shifting pool of self-selected applicants.

Our study motivates a direction for future research. We would like to obtain tree approximations of the joint probability distribution for admissions decisions for all elite universities in general. A good start might be obtaining the admissions data for a university of similar standing for the past year. We could then build another forecasting model using CART, which we can use to compare to the one presented in this paper. Substantial differences would indicate a divergence in either the pool of applicants or the admissions decision process. Substantively similar models, with emphasis on gpa then SAT, might point to a sameness in the decisions for prospective students for all similar universities.

7 Appendix

```
load("~/Dropbox/University of Pennsylvania/S2016/STAT474/CART Project
    2/AdmissionsData.rdata")

#examining missing data
library(VIM)
summary(aggr(Admissions))
par(mfrow = c(2,2))
Admissions.missingRace <- subset(Admissions, is.na(Admissions$anglo))
Admissions.missingIncome <- subset(Admissions, is.na(Admissions$income
    ))
Admissions.missingSex <- subset(Admissions, is.na(Admissions$sex))
Admissions.zerogpa <- subset(Admissions, Admissions$gpa.wtd == 0)
Admissions.zeroSAT <- subset(Admissions, Admissions$sati.verb == 0) #
    if missing verb, also missing math.
par(mfrow = c(2,2))
hist(Admissions.missingRace$gpa.wtd, breaks = 10)
hist(Admissions.missingRace$sati.verb, breaks = 16)
hist(Admissions.missingRace$sati.math, breaks = 16)
hist(Admissions.missingRace$income, breaks = 10)
par(mfrow = c(2,2))
hist(Admissions.missingIncome$gpa.wtd, breaks = 10)
hist(Admissions.missingIncome$sati.verb, breaks = 16)
hist(Admissions.missingIncome$sati.math, breaks = 16)
par(mfrow = c(2,2))
hist(Admissions.missingSex$gpa.wtd, breaks = 10)
hist(Admissions.missingSex$sati.verb, breaks = 16)
hist(Admissions.missingSex$sati.math, breaks = 16)
hist(Admissions.missingSex$income, breaks = 10)
par(mfrow = c(2,2))
hist(Admissions.zerogpa$sati.verb, breaks = 16)
hist(Admissions.zerogpa$sati.math, breaks = 16)
hist(Admissions.zerogpa$income, breaks = 10)
par(mfrow = c(2,2))
hist(Admissions.zeroSAT$gpa.wtd, breaks = 10)
hist(Admissions.zeroSAT$income, breaks = 10)


#univariate statistics
library(stargazer)
stargazer(Admissions, median = T)
sum(na.omit(Admissions$income == 99999))
sum(na.omit(Admissions$sati.math == 0))
sum(na.omit(Admissions$sati.verb == 0))
sum(na.omit(Admissions$gpa.wtd == 0))


#histograms
par(mfrow = c(2,2))
hist(Admissions$gpa.wtd, main = "Histogram of Weighted GPA", xlab = "
```

```
    Weighted GPA", ylab = "Frequency")
hist(Admissions$sati.verb, main = "Histogram of SAT I Verbal Scores",
    xlab = "SAT I Verbal Scores", ylab = "Frequency")
hist(Admissions$sati.math, main = "Histogram of SAT I Math Scores",
    xlab = "SAT I Math Scores", ylab = "Frequency")
hist(Admissions$income, main = "Histogram of Income", xlab = "Income",
     ylab = "Frequency")
plot(as.factor(Admissions$admit), main = "Frequency of Admission",
    xlab = "Admission Decision", ylab = "Frequency")
plot(as.factor(Admissions$anglo), main = "Frequency of Anglo-Saxon
    Applicants", xlab = "Is Anglo-Saxon", ylab = "Frequency")
plot(as.factor(Admissions$asian), main = "Frequency of Asian
    Applicants", xlab = "Is Asian", ylab = "Frequency")
plot(as.factor(Admissions$black), main = "Frequency of Black
    Applicants", xlab = "Is Black", ylab = "Frequency")
plot(as.factor(Admissions$sex), main = "Frequency of Sexes", xlab = "
    Is Male", ylab = "Frequency")


#bivariate statistics
pairs(Admissions[5:8])

#CART
library(caret)
library(rpart)
library(rpart.plot)
attach(Admissions)

  #create 3 random disjoint splits of data
  #listwise deletion and reconstruction
  set.seed(1234)
  temp <- Admissions[complete.cases(Admissions),]
  temp <- subset(temp, temp$gpa.wtd != 0)
  temp <- subset(temp, temp$sati.verb != 0)
  temp <- subset(temp, temp$sati.math != 0)
  index <- sample(1:6175, 6175, replace = F)
  temp <- temp [index,]
  train <- temp[1:2058,]
  eval <- temp[2059:4117,]
  test <- temp[4118:6175,]

par(mfrow = c(1,1))

#cp = 0.01, default prior =c(0.6870748, 0.3129252)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train, method = "class", cp = .01)
prp(out, extra = 101, fallen.leaves = T, Margin = .1, uniform = T,
    faclen = 20, varlen = 20, box.col = c("red", "lightblue")[
    out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)
```

```
#cp = 0.05, prior = c(.5, .5)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c(.5,
    .5)), method = "class", cp = .05)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.025, prior = c(.5, .5)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c(.5,
    .5)), method = "class", cp = .025)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.01, prior = c(.5, 0.5)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c(0.5,
    0.5)), method = "class", cp = 0.01)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.01, prior = c(0.55, 0.45)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c(0.55,
     0.45)), method = "class", cp = 0.01)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.01, prior = c(0.525,  1- 0.525)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c
    (0.525, 1-0.525)), method = "class", cp = 0.01)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
```

```
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.025, prior = c(0.515, 1- 0.515)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c
    (0.515, 1 - 0.515)), method = "class", cp = 0.025)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#cp = 0.015, prior = c(0.515, 1- 0.515)
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c
    (0.515, 1 - 0.515)), method = "class", cp = .015)
prp(out, extra = 101, under = T, fallen.leaves = T, Margin = .1,
    uniform = T, faclen = 20, varlen = 20, box.col = c("red", "
    lightblue")[out$frame$yval])
pred <- predict(out, eval, type = "class")
tab <- table(eval$admit,pred)
print(tab)

#final tree
out <- rpart(admit ~ anglo + black + asian + income + sex + gpa.wtd +
    sati.math + sati.verb, data = train,  parms = list(prior = c(0.5,
    0.5)), method = "class", cp = 0.01)

#honest performance assessment

finalTree <- out
pred <- predict(finalTree, test, type = "class")
tab <- table(test$admit, pred)
print(tab)
```

References

Willingham, Warren W., Breland, Hunter M. 1983. "Personal Qualities and College Admissions". American Journal of Education. Vol. 91, No. 2 (Feb., 1983), pp. 279-282.