Predicting College Admissions: The Counsellor's Game

Random Forest Edition

Victor Cheung

March 21, 2016

1 Introduction

We return once again to the college admissions data, but this time with random forest as our preferred method of analysis. Much of the discussions concerning missing data carries over to this work, so we have attached it in the appendix. In this paper we will focus primarily on the discussions pertinent to random forests.

We will examine college admissions 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 examination of variable importance plots, partial dependence plots and the empirical margin. We will conclude the paper in section 6. In the appendix, we discuss the exact nature of the missing data (covered in the previous paper) and 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

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

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

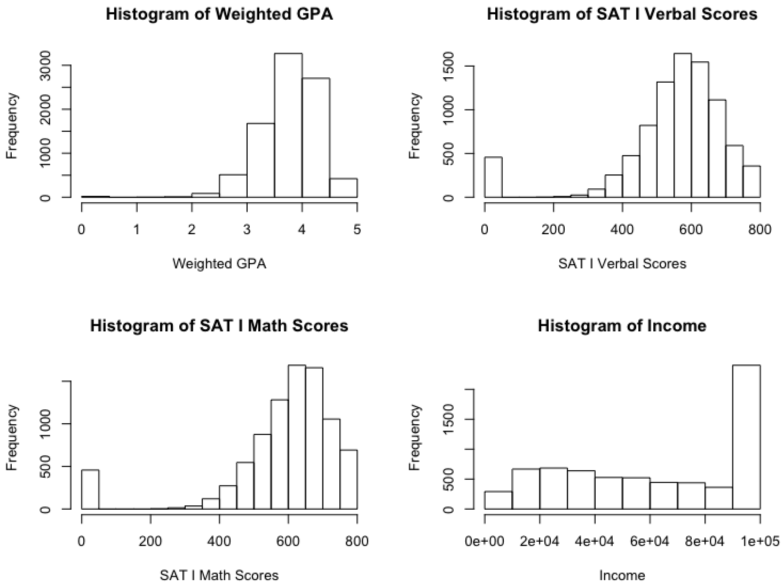
The summary statistics provides interesting information on the distribution of race and sex. 35.7% of applicants who reported race are anglo-saxon, 43.4% are Asian, and only 4.5% are black. As well, we see that only 46.6% of applicants are male. The mean and median of income is extremely misleading -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

Weighted GPA GPA follows a left skewed distribution with a with median centered at 3.850. 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.

Figure 1: Histograms for gpa, sat and income

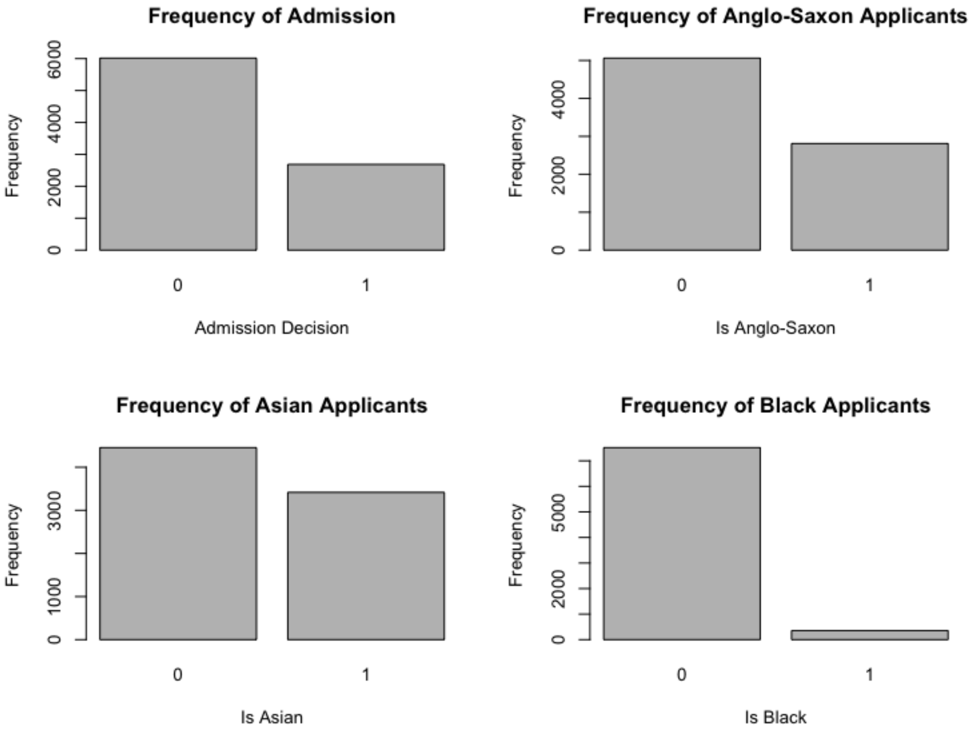

Histogram of Weighted GPA

Histogram of SAT I Verbal Scores

Histogram of SAT I Math Scores

Histogram of Income

Figure 2: Histograms for Distribution of Admissions and Race

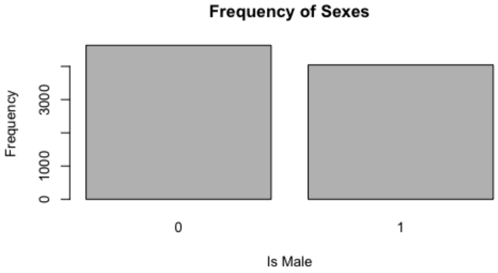

Frequency of Admission

Frequency of Anglo-Saxon Applicants

Frequency of Asian Applicants

Frequency of Black Applicants

Figure 3: Histogram of Distribution of Sex



SAT I Verbal Scores SAT I verbal scores follow a left skewed distribution with median at 580. 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. 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.

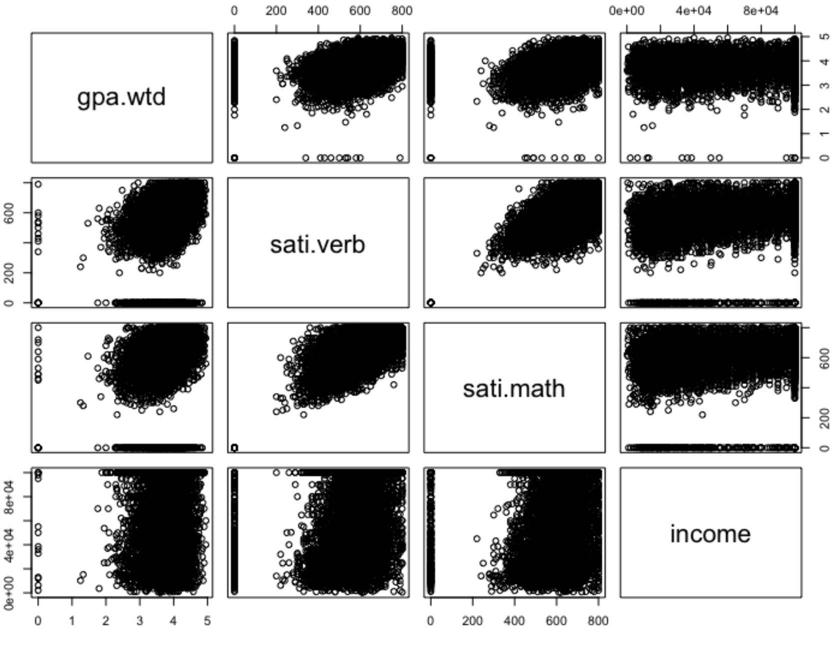
Income Income is the most problematic piece. Income is almost uniformly distributed amongst the bins everywhere below $100,000$. Indeed, a quick count shows that there are 2164 values of income at $99,999$. This suggests a very large tail to the right that has been cut off. This binning will affect our 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.

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. [1]

2.2 Bivariate Statistics

Figure 4: Pairwise Correlations



An analysis of pairwise correlations shows the expected associations between gpa

[1] http://opinionator.blogs.nytimes.com/2013/02/02/the-boys-at-the-back/

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. This implies that cannot use regression to fill in missing income data.

3 Problems with Data

The analysis is largely the same as in the previous submission with CARTs; and we leave the already covered portions in the appendix.

We did not include observations with missing values for predictors for CART. We may do so this time if random forests handle missingness better than CART. Random forests, in dealing with categorical missing data, employs two approaches. The first uses the modal values for the predictor in place of missing values. The second first uses modal values to compute a random forest, constructs a proximity matrix, and then gives heavier weights to cases most like the ones with missing data for approximating missing values.

There are multiple reasons why these tricks should not be employed in this case. First, from the literature we know that imputing values makes OOB measures of fit far too optimistic - the random forest algorithm attributes more information to these observations than there actually is. Furthermore, that a single weighted value is imputed for each observation distorts level II analysis by overriding the variations in imputation. The nature of the data presents problems as well, most exemplified by income. Due to binning, the modal value of income is $100,000. However, there is no reason to believe that this is the correct modal value, were income not capped. The distribution of income is probably approximately uniform with a right skew, which makes the use of modal values even more suspect.

Once again, since we have a large number of observations, and since most of the missingness resembles deletion at random, we listwise delete all observations with missing data to preserve the integrity of level II analysis.

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

Random Forests present a number of benefits over decision trees. In ideal situations, it allows for decreases in both the variance and the bias of fit by growing trees as large as possible and then voting over trees. As well, growing a forest of trees allows us to use random bootstrap samples of data for each tree and random subsets of predictors for each split, thus increasing independence in fitted values between trees. This helps moderate the effects of extremely influential observations and predictors. In this way, we also allow weak predictors that would have been ignored in a single decision tree to contribute towards a more complex fit.

We did not have to partition our dataset into training, evaluating and testing datasets this time. The algorithm already uses OOB data to compute fitted values, and we use the confusion tables produced by the forests to evaluate the results of tuning.

In building the random forests, there are a number of tuning parameters we will consider. As we will see, it turns out that default settings for `randomForest()` in `R` work well.

In terms of the node size, which was an important tool for pruning in CART, for random forests we allow the terminal node to be of size 1. This happens to be the default option in `R`. Overfitting is an obvious concern, but this is alleviated by the fact that each tree is trained with a subset of the data, and that we are voting over trees. In fact, for random forests we want each tree to be as large as possible. In the statistical literature, this is known as interpolation where each terminal node is perfectly homogeneous. It introduces the benefit of local robustness, where observations that vacillate between realizations is isolated in small terminal nodes, and where averaging over trees reduces the generalization error that would've resulted had we only used one tree.

For the number of trees, the lore is that 500 trees is a good balance between having enough trees and incurring too much computational time. This is the default in `R`. We will start with 500 trees when evaluating our models with OOB data. Then we verify our results from tuning by running a random forest with tuned parameters and 3000 trees. We need not worry about overfitting due to the increased number of trees - Breiman has already proved that random forests do not overfit as more trees are grown.

For the number of predictors sampled at each split, the `R` default is 2. Again, our concern is that each predictor has sufficient opportunity to compete and contribute to reducing impurity. Given our goal of maximizing tree size, and of allowing for 500 trees, sampling 2 predictors should be sufficient at each split. We are also cautious against increasing this parameter, since from the bivariate analysis we know there is a high correlation between certain predictors (SAT math, verbal and weighted GPA).

4.1 Justifying Level II

Random forest is concerned with level II analysis by design. The algorithm already uses OOB data to obtain fitted values for each tree, to measure out of sample performance, and to compute predictor importance. However, the subtleties behind the algorithm makes a formal justification difficult, and work is underway to understand statistical inference using random forests. We produce a forest based approximation to the true response surface, with an estimated generalization error that converges to the true generalization error of that approximation.

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 non-measured 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 Relative Costs

For random forests, there are four ways of tuning for the target relative cost: changing the priors for each tree, stratified sampling, changing the weight of the votes and changing the margins of victory for voting. The last two works through mechanisms easily understood, but they leave the trees themselves unchanged and still assumes that the costs of misclassification are equivalent. We prefer to change the trees themselves to reflect those costs. Furthermore, the response is 30% admit and 70% deny, which is not heavily unbalanced. Since about 4000 observations are sampled for each tree in the random forest, this alleviates concerns about inadequate representation of admitted students. Therefore, we will use priors to adjust the structures of the trees themselves, and in that way obtain our target relative cost.

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.

Our target relative cost for false negatives to false positives is again 5 to 1. This implies false negatives are five times more costly than false positives, hence false negatives should be five times less prevalent than false positives. This is equivalent to asking for more stringent statistical evidence before classifying as denied, and asking for laxer statistical evidence before classifying as admitted. We anticipate that overall error rates will increase relative to when the priors are are just the empirical ratios.

4.3 Models

Our first attempt produced the confusion table below.

Table 3: First Random Forest, Observed Relative Cost = 0.458

	Predicted Admitted	Predicted Denied	Model Error
Admitted	1235	690	35.8%
Denied	316	3934	7.43%
Use Error	20.4%	14.9%	Generalization Error = 16.3%

We avoid estimating too many cost ratios, since doing so taints the OOB "test" data. However, since we estimated <10 times and that our sample size is large at around 6000 observations, that OOB assessment of out of sample performance should remain valid. Our final random forest, found by stepping through different values for `classwt`, produced the following confusion tables, variable importance plots, and partial dependence plots. A histogram of the empirical margins for the final random forest is also presented.

Table 4: Final Random Forest, Observed Relative Cost = 5.28

	Predicted Admitted	Predicted Denied	Model Error
Admitted	1701	224	11.6%
Denied	1182	3068	27.8%
Use Error	41.0%	6.80%	Generalization Error = 22.8%

5 Discussion

Moving from our first model, with symmetric cost assumptions, to our final model, with the target cost ratio of 5 to 1, we observe that the actual relative cost is now 5.28, decently close to the target. As well, generalization error increased from 16.3% to 22.8%, reflecting the fact that we want to capture more of the potential admits at the cost of higher generalization error. Furthermore, the number of predicted denials has decreased while the number of predicted admits have increased - this is what we would expect, given we have increased statistical evidence necessary for the former and decreased that for the latter.

From our CART paper we learnt that SAT math and verbal scores and gpa were the most important predictors. That story is confirmed with our random forest model. In `R`, these predictors were randomly shuffled such that they can no longer contribute to forecasts, which produced figures 5, the variable importance plots. These are measures of the average difference in prediction error on OOB

Figure 5: Variable Importance Plot

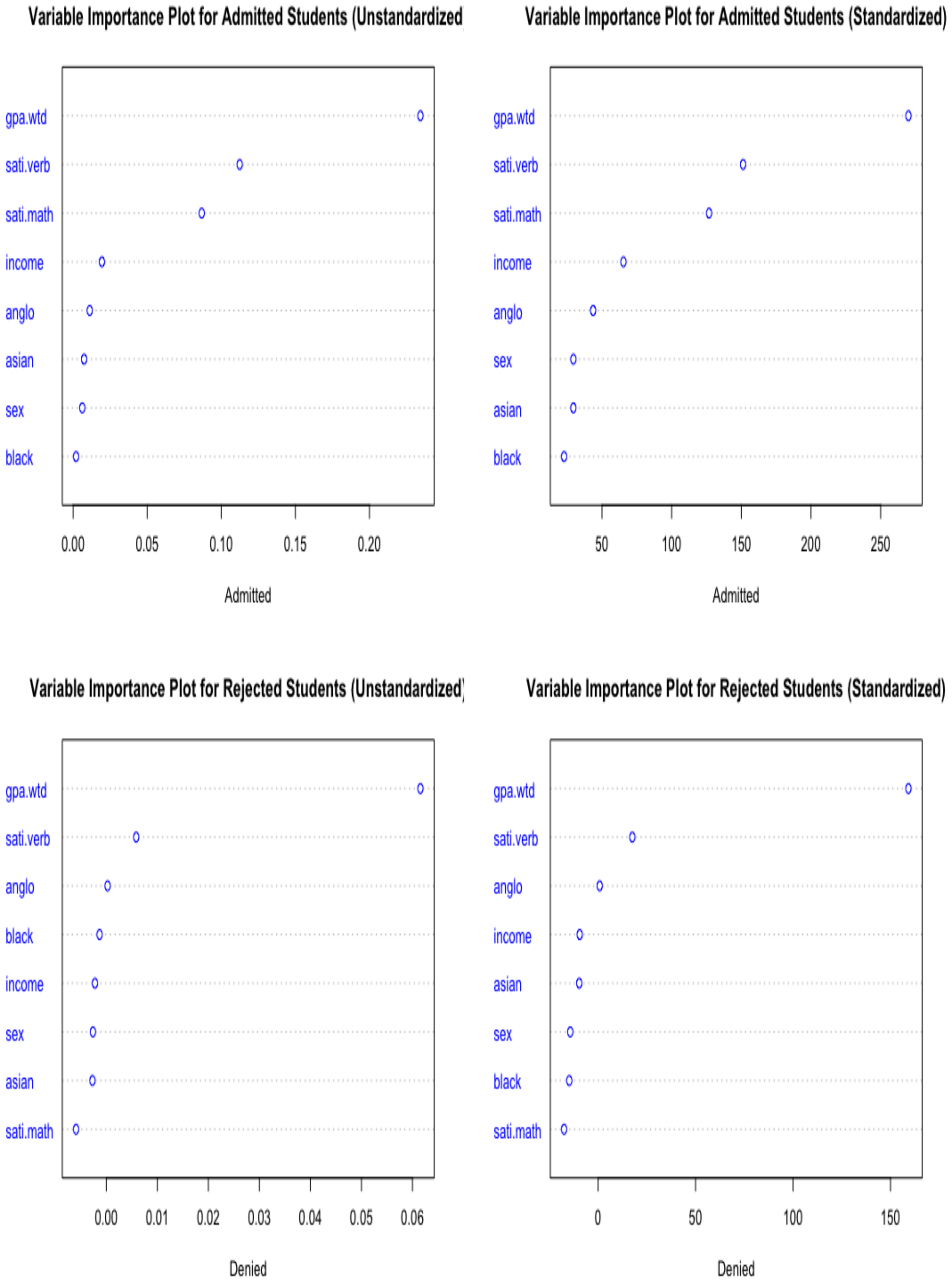


Figure 6: Partial Dependence Plot

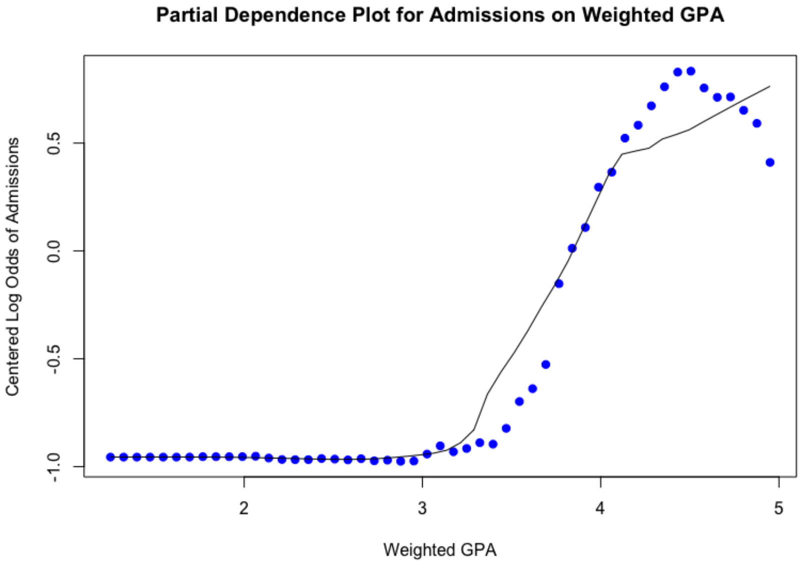


Figure 7: Partial Dependence Plot

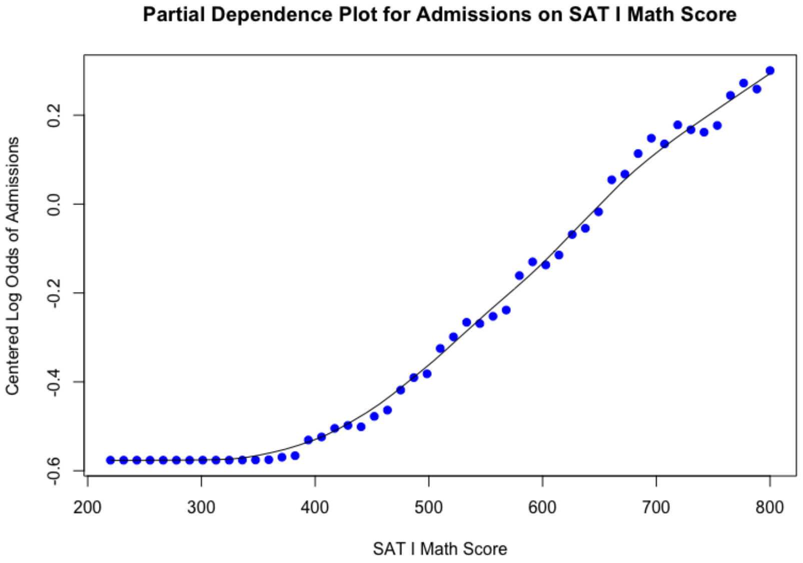


Figure 8: Partial Dependence Plot

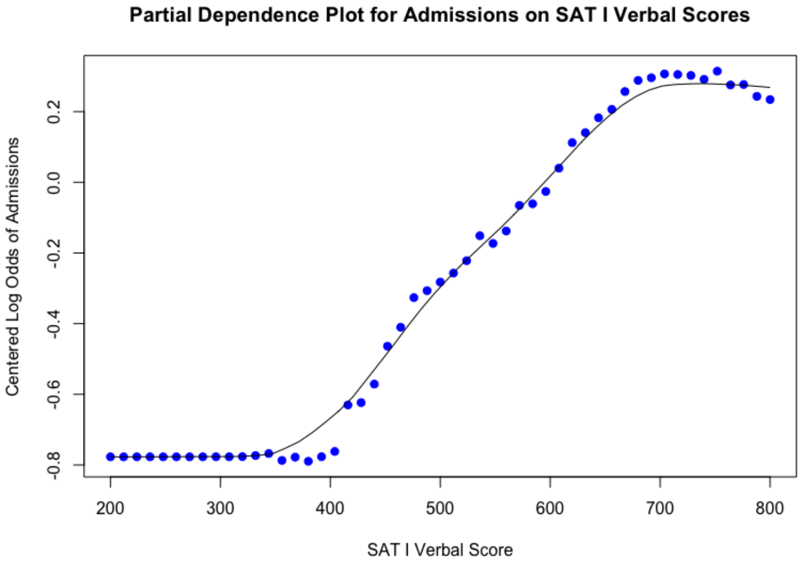


Figure 9: Partial Dependence Plot

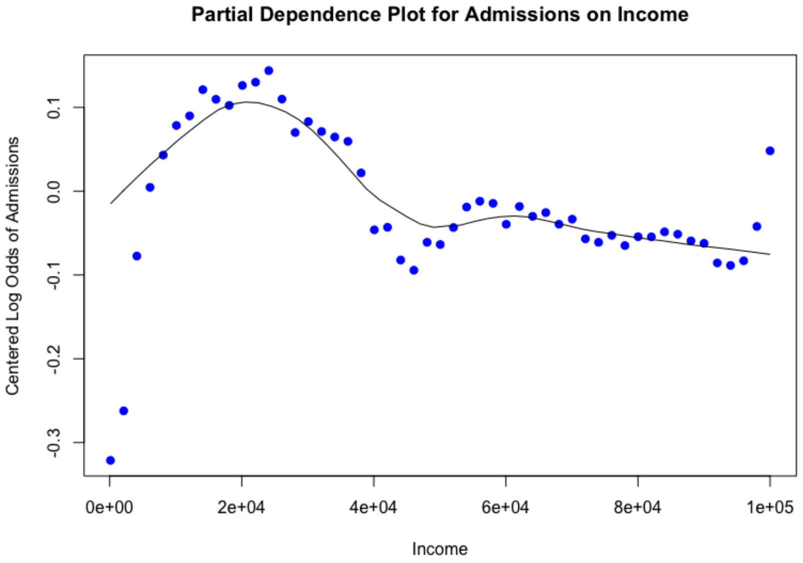


Figure 10: Partial Dependence Plot

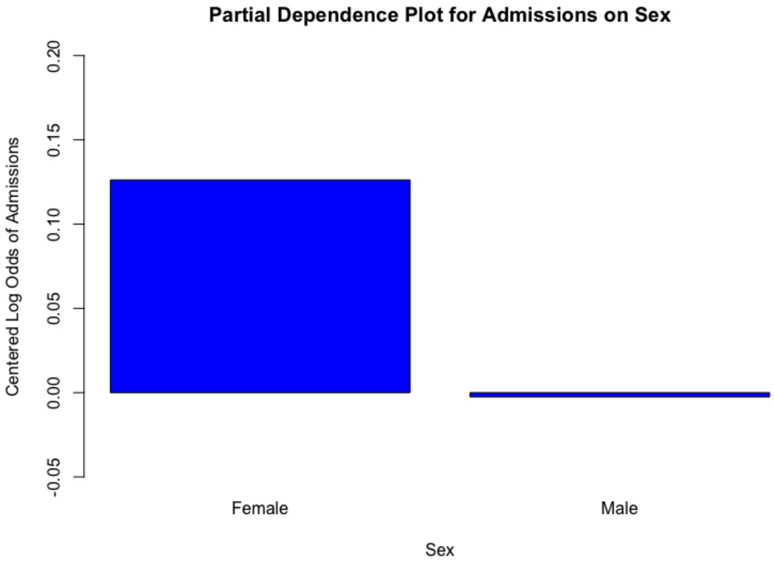


Figure 11: Partial Dependence Plot

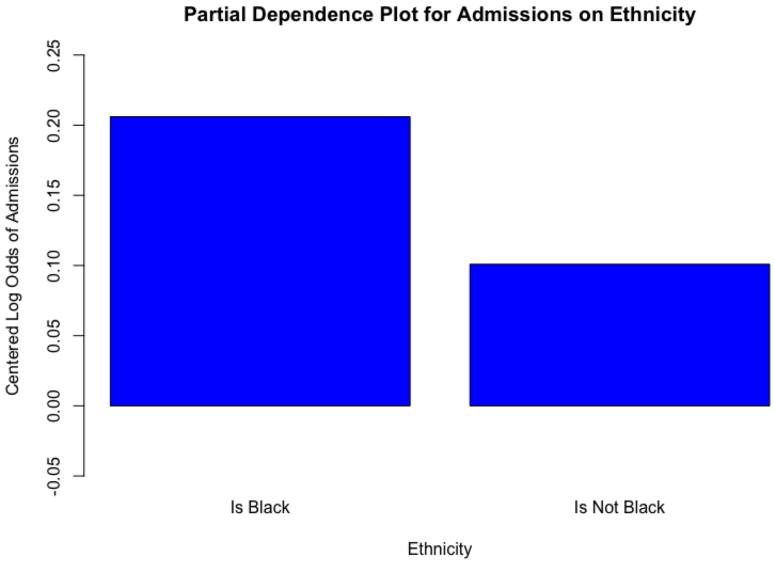


Figure 12: Partial Dependence Plot

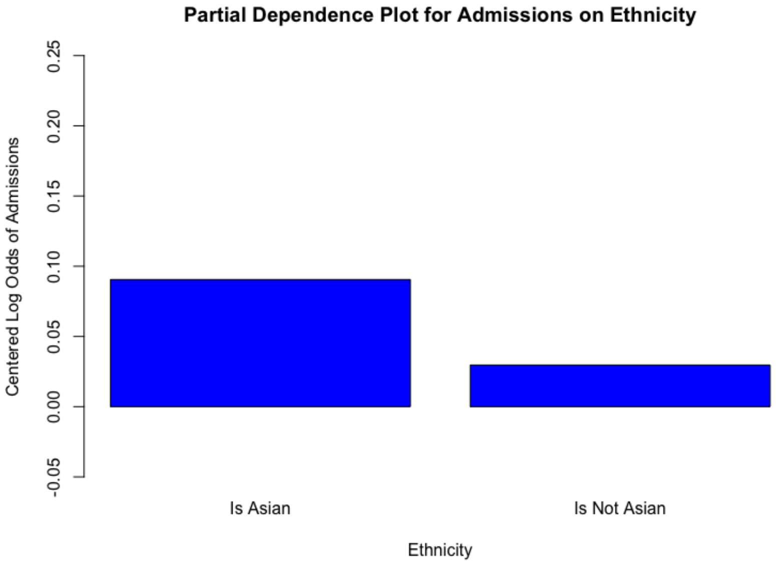


Figure 13: Partial Dependence Plot

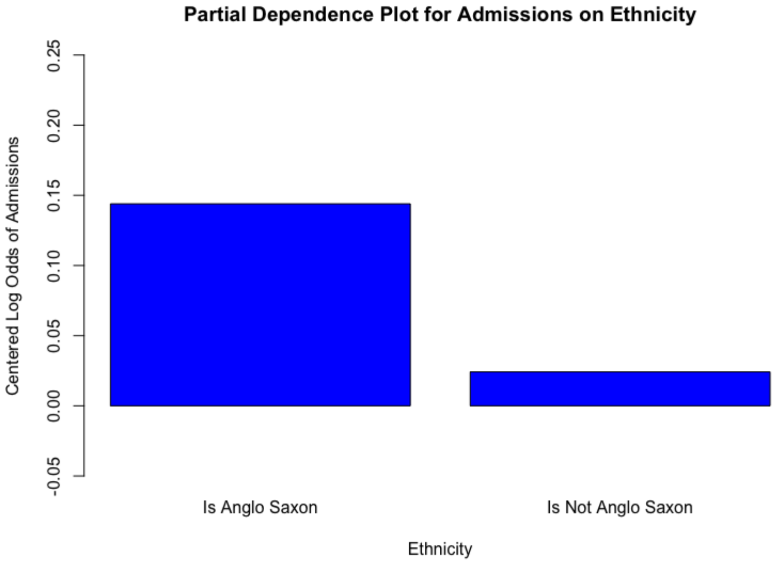
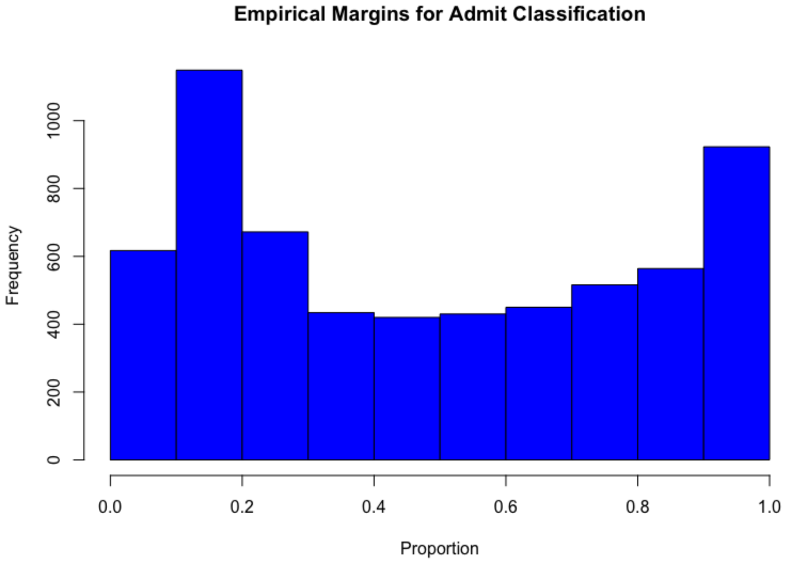


Figure 14: Histogram of Empirical Margins



observations when predictors are shuffled vs. not shuffled. Our models remain fixed between shuffles.

Consider first admitted students. From figure 5, we see that when gpa is shuffled, that forecasting error increases from 17.6% to about 40%. When verbals scores are shuffled, the error increases to about 30%, and for math, to about 27%. These are sizable increases in forecasting error, and given our problem domain, these predictors clearly matter for us as college counselors. GPA is by far the most important predictor in terms of forecasting accuracy, followed by verbal score and math score. We caution against inferring causal effects of GPA or sat scores on the admissions decision, as this can only derive from the data collection process. We also have no information on the functional relationship between these predictors and the response.

The percentage effects on forecasting accuracy for students denied admissions is different due to differences in number of observations and margins, but the standardized plots from figures 5 show that the effects from GPA is similar. However, SAT scores have dropped drastically in importance; SAT math is now negligible, and SAT verbal has an effect on accuracy just above 1%.

All other predictors have very small effects on accuracy. The sex, ethnicity and income of the applicants have effects on forecasting accuracy that are effectively zero. In particular, for denied students, shuffling on the `income, sex, asian, sati.math` variables produced negative decreases (gains) in forecasting accuracy. This is pure noise.

Note for the partial dependence plots, since our response variable is binary, that we need only plot one of the responses; the other will be a mirror image. The partial dependence plots consider the effect on admissions when the predictor being examined changes, while all other predictors are held constant. The vertical axis is center log odds where the odds are proportions of votes from the forest.

From figure 6, increasing GPA below 3 has no effect on the the prospect of admissions; from 3 onwards, the prospect of admission rapidly increases. We note that the rate of increase appears to slow down when GPA is past 4. A reasonable explanation is that admissions officers at this elite college automatically discard applications with extremely low GPAs. Past a certain threshold, the officers will consider your application, and the better your GPA, the more favorably they look upon your application. However, once your GPA is extremely high, there are no long additional points awarded for having near-perfect GPA. It's likely then that past some high threshold for GPA, and given some unexplained criteria is met, applicants are likely admitted. The same story holds for SAT verbal scores in figure 8. Past the threshold at around 350, odds of admissions is increasing; however, the effect on admissions hits a plateau around 700. Increased SAT verbal scores past this point does not increase the prospect of admissions.

From figure 7, SAT math scores tells a slightly different story to GPA and SAT Verbal. Prospects for admissions is unaffected by SAT scores below approximately 400; however, it increases approximately linearly towards the full score of 800. The effect on prospect of admissions does not diminish past some upper bound.

Finally, for income in figure 9, there's an odd pattern. Odds of admissions is increasing from 0 to $20,000, at which point it decreases and slowly plateaus towards higher incomes. This is probably explainable by some confounding and hidden factor operating in the background.

Consider now the effects of sex and ethnicity in figures 10 through 14. Females are substantially more favored in admissions than males. And out of the three ethnicities, it appears that blacks have the best prospects of admissions relative to nonblacks, followed by whites relative to nonwhites, and finally asians relative to nonasians.

Since we intend to use the final random forest for forecasting, we are concerned with the reliability of our forecasts. This is hinted at by the empirical margins. In figure 15, we see that the number of observations where the vote is close to evenly split dips relative to where votes are highly skewed. There are only around 800 out of 6175 cases for which the proportions hover between 0.4 and 0.6. This implies that for the majority of cases, the margins by which students are classified is large. Given our large sample size, the abundance of admitted and rejected students, as well as the scarcity of ambiguous cases with low margins, we need not be overly concerned that dropping predictors would result in dramatically altered forecasting accuracy. In short, our model is reliable.

6 Conclusion

We built a random forest of 3,000 trees, using a target cost ratio of 5 to 1, and ended up with an overall error rate of 17.6% and a observed cost ratio of 5.28.

From the variable importance plots, we found that the three most important variables for forecasting accuracy were GPA, SAT verbal and SAT math scores, in that order. Notably, factors such as sex, income and ethnicity were largely noisy or inconsequential. They do not play a statistically significant role in the decision, according to our model.

From the partial dependence plots, we saw that GPA and SAT scores had no effect on prospect of admissions until a certain threshold, at which point the odds increase largely linearly. For GPA is a diminishing effect short of 5.0/5.0. For SAT verbal, there is a plateau beginning at 700. There is no plateau for SAT math. The effect of income on admissions is strange, initially rising then slowly tapering off, with a mode around $20,000. We believe this is due to some unmeasured variable or policy acting in the background. For the gender of the applicant, we found a substantial differential effect on admissions prospect that favors females over males. Finally, in terms of ethnicity, prospects of admission appear best for blacks, better for anglo-saxons, and decent for asians.

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.

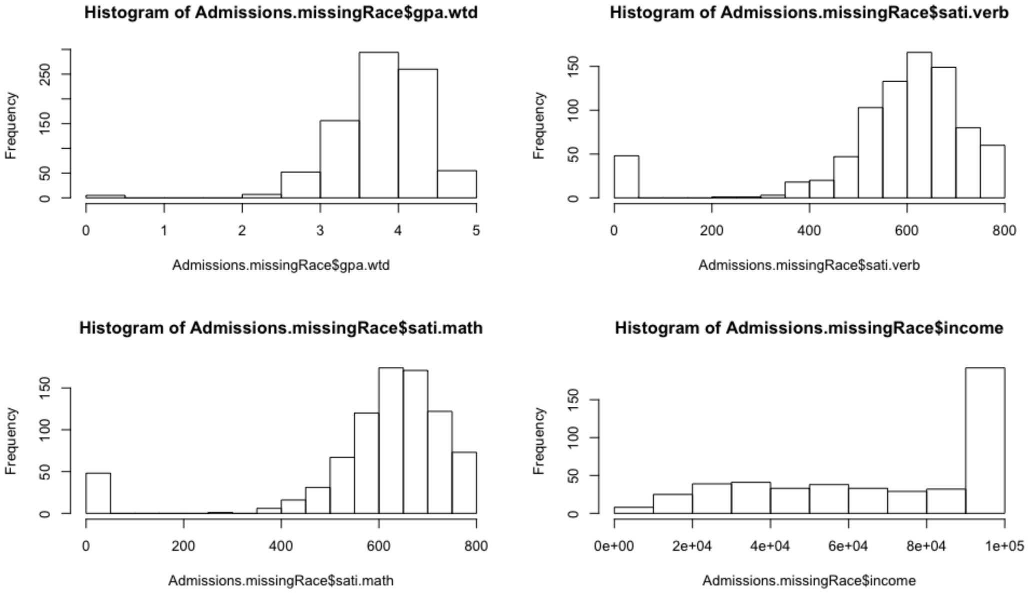
For this one college, we would also like to keep track of forecasting accuracy over time in order to assess shifting trends in the admissions process. For example, a comparison of forecasting accuracy five years from now vs. this year could provide insight into the stationarity of the college admissions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

Missing Data Analysis from CART

Figure 15: Histograms for Observations with Missing Race



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

Figure 16: Histograms for Observations with Missing Income

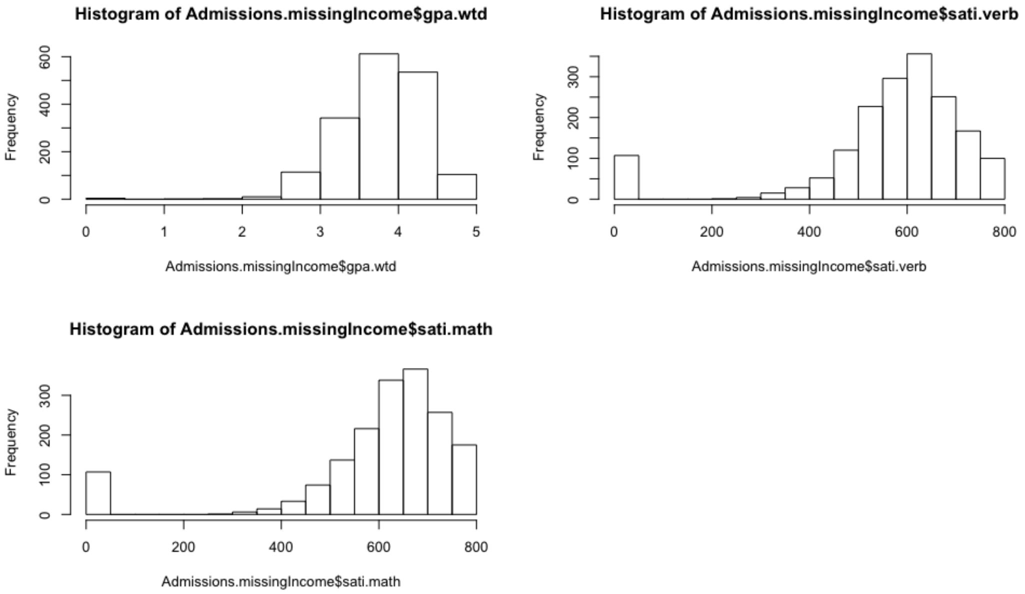


Figure 17: Histograms for Observations with Missing Sex

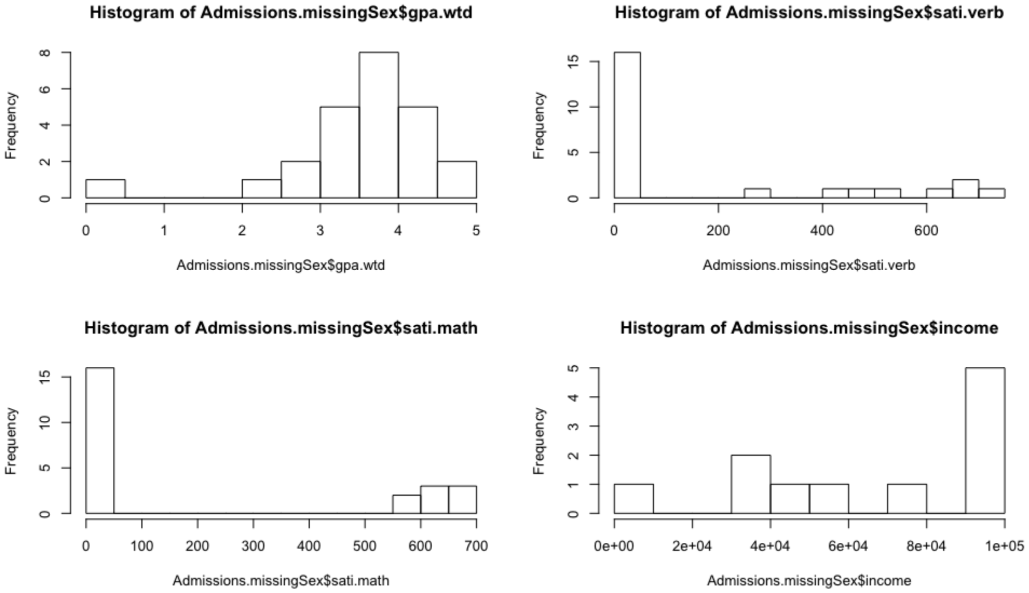


Figure 18: Histograms for Observations with Zero GPA

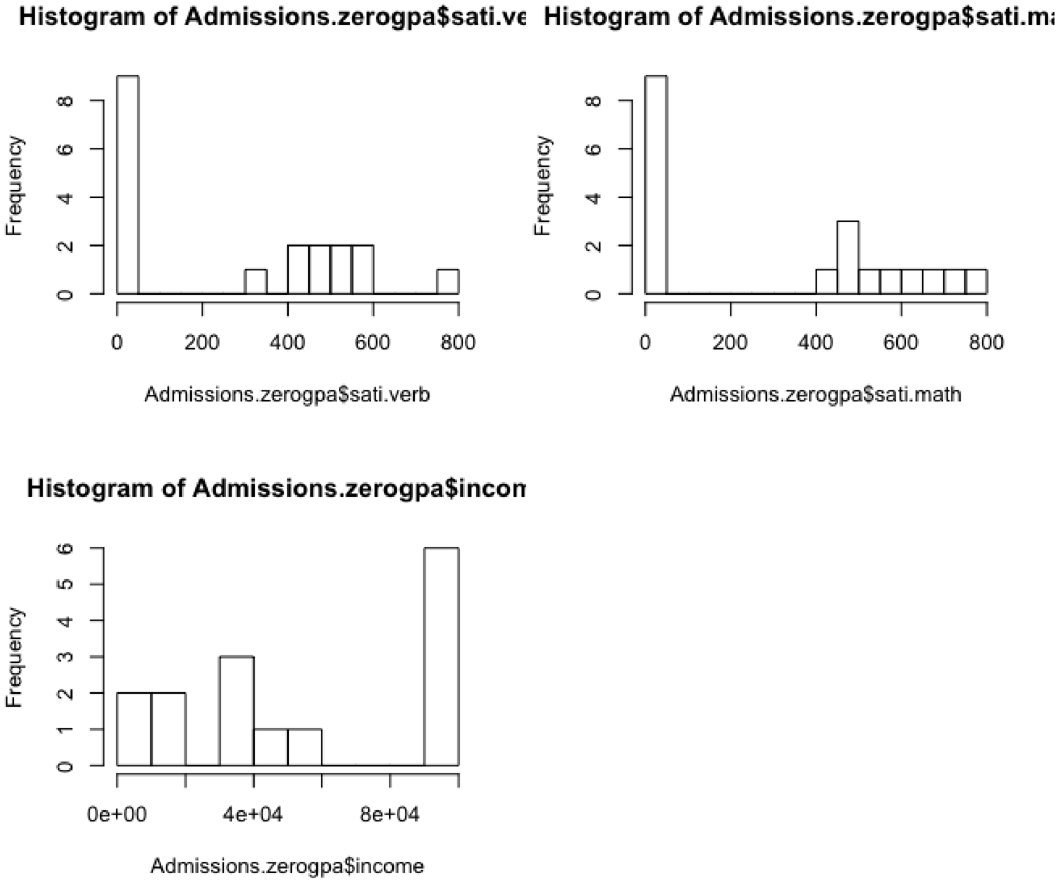
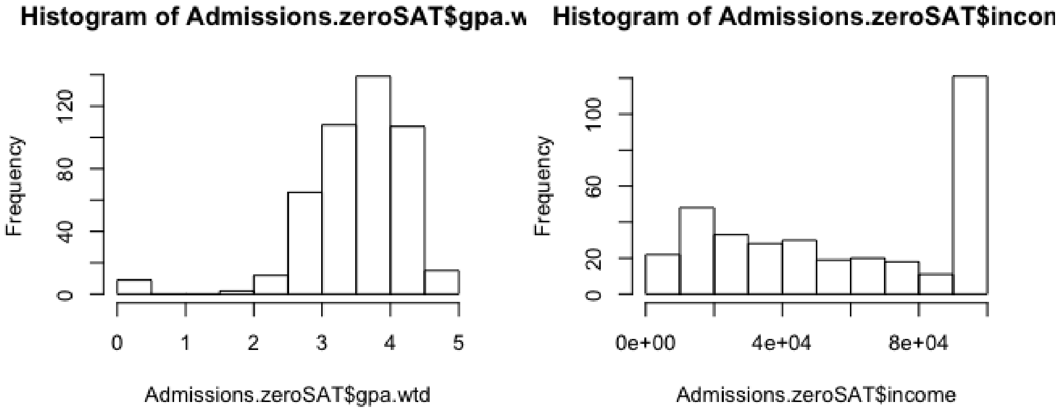


Figure 19: Histograms for Observations with Zero SAT



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.

R Code

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

#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$sati.math != 0)
temp$sex <- ifelse(temp$sex == ’1’, ’Male’, ifelse(temp$sex == ’0’, ’
    Female’, NA))
temp$admit <- ifelse(temp$admit == ’1’, ’Admitted’, ifelse(temp$admit
    == ’0’, ’Denied’, NA))
temp$anglo <- ifelse(temp$anglo == ’1’, ’Is Anglo Saxon’, ifelse(
    temp$anglo == ’0’, ’Is Not Anglo Saxon’, NA))
temp$black <- ifelse(temp$black == ’1’, ’Is Black’, ifelse(temp$black
    == ’0’, ’Is Not Black’, NA))
temp$asian <- ifelse(temp$asian == ’1’, ’Is Asian’, ifelse(temp$asian
    == ’0’, ’Is Not Asian’, NA))
temp$sex <- as.factor(temp$sex)
temp$admit <- as.factor(temp$admit)
temp$anglo <- as.factor(temp$anglo)
temp$black <- as.factor(temp$black)
temp$asian <- as.factor(temp$asian)
attach(temp)


#emprirical distribution
rf1 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T)
rf2 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.5, 0.5))
rf3 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.4, 0.6))
rf4 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.3, 0.7))
rf5 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.1, 0.9))
rf6 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.2, 0.8))
rf7 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.15, 0.85))
rf8 <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.175, 1- 0.175))
#verify with ntree = 3000
final <- randomForest(admit ~ anglo + black + asian + gpa.wtd + sati.
    verb + sati.math + income + sex, data = temp, importance=T,
    classwt = c(0.15, 0.85), ntree = 3000)

#confusion matrix
print(final)

#variable importance plots
par(mfrow = c(1,1))
varImpPlot(final,type=1,scale=F, class = 'Admitted', main = "Variable
    Importance Plot for Admitted Students (Unstandardized)", col = "
    blue")
varImpPlot(final,type=1,scale=T, class = 'Admitted', main = "Variable
    Importance Plot for Admitted Students (Standardized)", col = "blue
    ")
varImpPlot(final,type=1,scale=F, class = 'Denied', main = "Variable
    Importance Plot for Rejected Students (Unstandardized)", col = "
    blue")
varImpPlot(final,type=1,scale=T, class = 'Denied', main = "Variable
    Importance Plot for Rejected Students (Standardized)", col = "blue
    ")

#partial dependence plots
part1 <- partialPlot(final, pred.data = temp, x.var = gpa.wtd, rug = T
     , which.class = 'Admitted')
scatter.smooth(part1$x,part1$y,span=1/3,xlab="Weighted GPA",ylab="
    Centered Log Odds of Admissions", main="Partial Dependence Plot
    for Admissions on Weighted GPA",col="blue",pch=19)
part2 <- partialPlot(final, pred.data = temp, x.var = sati.math, rug =
     T, which.class = 'Admitted')
scatter.smooth(part2$x,part2$y,span=1/3,xlab="SAT I Math Score",ylab="
    Centered Log Odds of Admissions", main="Partial Dependence Plot
    for Admissions on SAT I Math Score",col="blue",pch=19)
part3 <- partialPlot(final, pred.data = temp, x.var = sati.verb, rug =
     T, which.class = 'Admitted')
scatter.smooth(part3$x, part3$y, span = 1/3,xlab="SAT I Verbal Score",
     ylab="Centered Log Odds of Admissions", main="Partial Dependence
    Plot for Admissions on SAT I Verbal Scores",col="blue",pch=19)
part4 <- partialPlot(final, pred.data = temp, x.var = income, rug = T,
     which.class = 'Admitted')
scatter.smooth(part4$x, part4$y, span = 1/3,xlab="Income", ylab="
    Centered Log Odds of Admissions", main="Partial Dependence Plot
    for Admissions on Income",col="blue",pch=19)
part5 <- partialPlot(final, pred.data = temp, x.var = sex, rug = T,
    which.class = 'Admitted', xlab="Sex", ylab="Centered Log Odds of
    Admissions", main="Partial Dependence Plot for Admissions on Sex",
     ylim = c(-2.5, 0.25))
part6 <- partialPlot(final , pred.data = temp , x.var = asian , rug = T,
    which.class = 'Admitted', xlab="Ethnicity", ylab="Centered Log
    Odds of Admissions", main="Partial Dependence Plot for Admissions
    on Ethnicity", ylim = c(-2.5, 0.25))
part7 <- partialPlot(final , pred.data = temp , x.var = black , rug = T,
    which.class = 'Admitted', xlab="Ethnicity", ylab="Centered Log
    Odds of Admissions", main="Partial Dependence Plot for Admissions
    on Ethnicity", ylim = c(-2.5, 0.25))
part8 <- partialPlot(final , pred.data = temp , x.var = anglo , rug = T,
    which.class = 'Admitted', xlab="Ethnicity", ylab="Centered Log
    Odds of Admissions", main="Partial Dependence Plot for Admissions
    on Ethnicity", ylim = c(-2.5, 0.25))

#empirical margins
hist(final$votes[,1], xlab = "Proportion", main = "Empirical Margins
    for Admit Classification", col = "blue")
summary(final$votes[,1])
```

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.