Mary Behnke

MBA Admissions - Final Project Report

Introduction

It is likely that a number of undergraduate business students or business professionals would like to apply to graduate school at some point. Given their demographic, professional, and performance attributes, it can be helpful for applicants to know the possibility of predicting whether they will be rejected, waitlisted, or admitted. The goal of this project is to discover if a model can be built to predict the probability of an applicant's admission status, what the most important factors are, and whether bias can be found in admission decisions, considering gender and international status. As an undergraduate business student myself, there's a very real possibility that I may one day apply to get my MBA, so this proposed model could give me a valuable tool to improve my chances of being admitted. Furthermore, as an individual, I am concerned about issues of bias in academic admission processes, and would be interested in discovering evidence for this in MBA admissions.

The predictive models will be built based on two outcome variables, one determining whether or not an applicant was rejected or not, and the other determining whether an applicant was admitted or waitlisted. The demographic features that will be taken into account are gender, international status, major, and race. The professional features that will be taken into account are work experience and work industry. The performance features that will be taken into account are gpa and gmat score. The gmat score is based on a test given to prospective MBA students.

Exploratory Data Analysis

This dataset was synthetically created, but reflects real-life admissions data. There are 6,194 rows and a total of 10 variables, including a unique identifier for each applicant. The other

variables are gender, international status, gpa, major, race, gmat score, work experience, work industry, and admission. Pre-processing included dealing with missing values, changing data types, and creating new variables to aid analysis. There were a number of missing values found in the race and admission columns. For race, the value “Not Provided” was imputed for each missing value, as this is usually an optional user-supplied field. For admission, “Rejected” was imputed for each missing value, as the previous values included only “Admit” and “Waitlist”. A number of variables were changed from character to factor data types in order to supply a clearer picture of the spread of the data (gender, international, major, work_industry, admission, and race). Finally, four new columns were created to aid future analysis: rejectID, admitID, genderID, and intID. The first two were used as outcome variables, and all are binary factors. RejectID gave a 1 for not rejected (admitted or waitlisted) and a 0 for reject, and admitID gave a 1 for admitted and 0 for waitlisted. GenderID and intID are simply binary versions of the gender and international columns, respectively.

First, a summary is supplied for every variable, showing a wide spread (Figure A.1). Given these statistics, it was decided to look more into the performance attributes of gpa and gmat score. For the gmat score, a histogram was created for each outcome variable (rejectID and admitID). For gpa, a box plot was created for each outcome variable. For the rejectID outcome variable, the gmat plot (Figure A.2) sees somewhat of a cutoff for admissions at the 655 mark, and there are many more rejected cases than not rejected. On the gpa plot (Figure A.3), you see a significantly higher average for not rejected applicants, but a larger spread for the rejected applicants. For the admitID outcome variable, the gmat plot (Figure A.4) shows a similar cutoff for the admitted versus waitlisted applications, and scores below 650 are quite muddied. The gpa plot (Figure A.5) shows much closer averages than the rejectID counterpart, this time with a

larger spread for admitted applicants. Given these plots, there is a consistent smaller difference between admitted and waitlisted applicants as opposed to rejected and not reject applicants, which seems intuitive. Furthermore, though there are observable cutoffs or averages for each metric, they are not rules that cannot be broken.

To prepare for analysis, a train-test split was performed, with 70% training and 30% test. Three models were created for each outcome variable, a logistic regression, decision tree, and an uplift model that took into account the binary variables of gender and international status. For the logistic regression and decision trees, eight variables and one interaction term were used: gender, international, gpa, major, race, gmat, work_exp, work_industry, and gpa:race.

Algorithm Training and Testing

rejectID - Logistic Regression

The goal of this logistic regression was to create a model able to predict one's chances of not being rejected as opposed to rejected. It was built with the previously identified eight variables and one interaction term. A few significant predictors were identified, those being genderMale and gmat (***), and raceOther, work_industryInvestment Management, gpa:raceOther (*). This model gave an accuracy of 83.85% on the test data.

rejectID - Decision Tree

The goal of this decision tree was to identify key predictors in predicting one's chances of not being rejected as opposed to rejected. It was built with the previously identified eight variables and one interaction term. The cp was set to 0.006, as that was the lowest value that saw a split, but the highest that retained interpretability. The top four predictors in the plot (Figure A.5) were very interesting, in order being gmat, gpa, race = Black, Hispanic, and gender = Male.

Further down, work industry also plays a significant role. This model gave an accuracy of 84.18% on the test data.

rejectID - Uplift Models

In the gender uplift model, with all observations set to *Male*, there was an average of a 10% decrease in percentage chance predicted by the logistic model for an applicant to not be rejected. In the international uplift model, with all observations set to *International*, there was an average of a 8% decrease in percentage chance predicted by the logistic model for an applicant to not be rejected.

admitID - Logistic Regression

The goal of this logistic regression was to create a model able to predict one's chances of being admitted as opposed to waitlisted. It was built with the previously identified eight variables and one interaction term. A number of significant predictors were identified, those being gmat (***), genderMale, work_industryFinancial Services, work_industryTechnology (*) and majorStem, work_industryInvestment Banking, work_industryInvestment Management, work_industryNonprofit/Gov (.) . This model gave an accuracy of 88% on the test data.

admitID - Decision Tree

The goal of this decision tree was to identify key predictors in predicting one's chances of being admitted as opposed to waitlisted. It was built with the previously identified eight variables and one interaction term. The cp was set to 0.001, as this was the highest value to produce any splits. The top predictors for this model were found to be gmat and race =

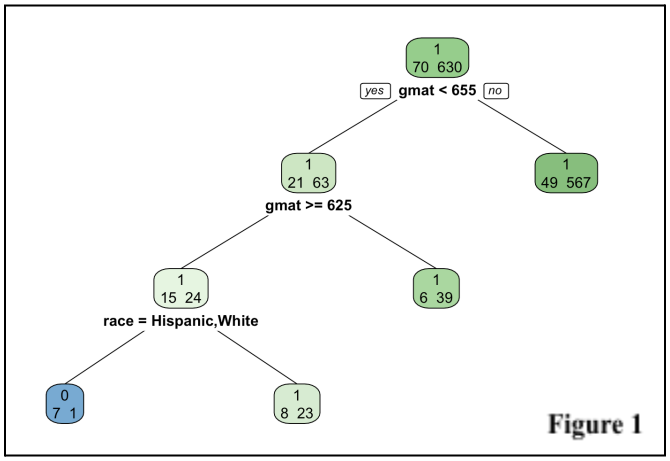

Figure 1

Hispanic, White, producing a very interesting set of splits (Figure 1). This model gave an accuracy of 90.33% on the test data.

admitID - Uplift Models

In the gender uplift model, with all observations set to *Male*, there was an average of a 5% decrease in percentage chance predicted by the logistic model for an applicant to not be rejected. In the international uplift model, with all observations set to *International*, there was an average of a 31% increase in percentage chance predicted by the logistic model for an applicant to not be rejected.

Discussion

Given these predictive models, there are a couple of findings. One is that it seems possible to create a somewhat accurate predictive model for MBA admissions based on demographic, professional, and performance metrics. This is evident based on the logistic regressions for rejectID and admitID with 83.85% and 88% accuracy, respectively. Furthermore, the decision trees created for each of these outcome variables showed that gmat in particular is an extremely important predictor for both. We also see that gpa, race, and gender become important in determining whether someone is not rejected, and that race is important for determining whether an applicant is admitted instead of waitlisted. The uplift models showed relative bias against applicants with gender = Male, and a notable difference in not rejected versus admitted/waitlisted for international applicants. These findings imply that there might be some evidence for bias in admissions based on these attributes, which is something that could be further looked into. On that note, a step after this could be to look into the attributes of race and work industry, as they both made interesting appearances in the models.

Appendix

Figure A.1 - Summary statistics

```
##  application_id    gender     international      gpa              major
##  Min.   :   1   Female:2251   False:4352    Min.   :2.650   Business  :1838
##  1st Qu.:1549   Male  :3943   True :1842    1st Qu.:3.150   Humanities:2481
##  Median :3098                               Median :3.250   STEM      :1875
##  Mean   :3098                               Mean   :3.251
##  3rd Qu.:4646                               3rd Qu.:3.350
##  Max.   :6194                               Max.   :3.770
##
##           race           gmat          work_exp
##  Asian       :1147   Min.   :570.0   Min.   :1.000
##  Black       : 916   1st Qu.:610.0   1st Qu.:4.000
##  Hispanic    : 596   Median :650.0   Median :5.000
##  Not Provided:1842   Mean   :651.1   Mean   :5.017
##  Other       : 237   3rd Qu.:680.0   3rd Qu.:6.000
##  White       :1456   Max.   :780.0   Max.   :9.000
##
##                          work_industry     admission
##  Consulting                     :1619   Admit   : 900
##  Private Equity/Venture Capital: 907   Rejected:5194
##  Technology                     : 716   Waitlist: 100
##  Nonprofit/Gov                  : 651
##  Investment Banking             : 580
##  Financial Services             : 451
##  (Other)                        :1270
```

Figure A.2 - rejectID distribution of gmat scores

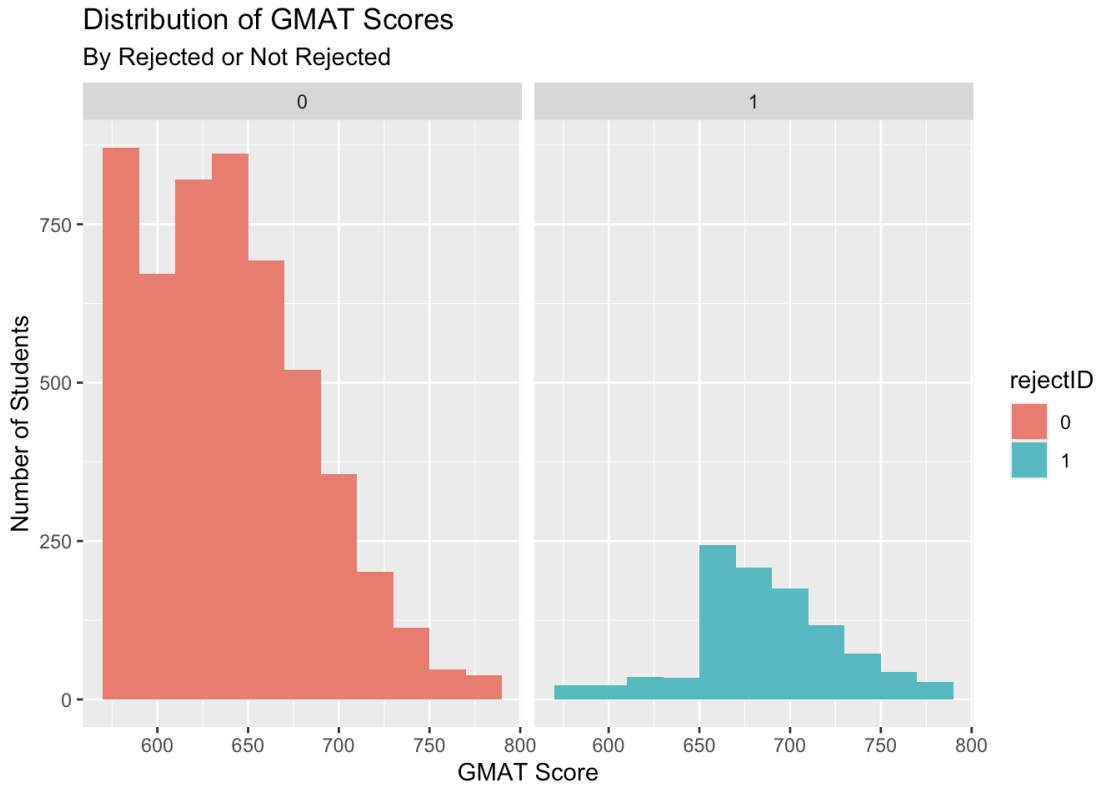

Figure A.3 - rejectID distribution of gpa

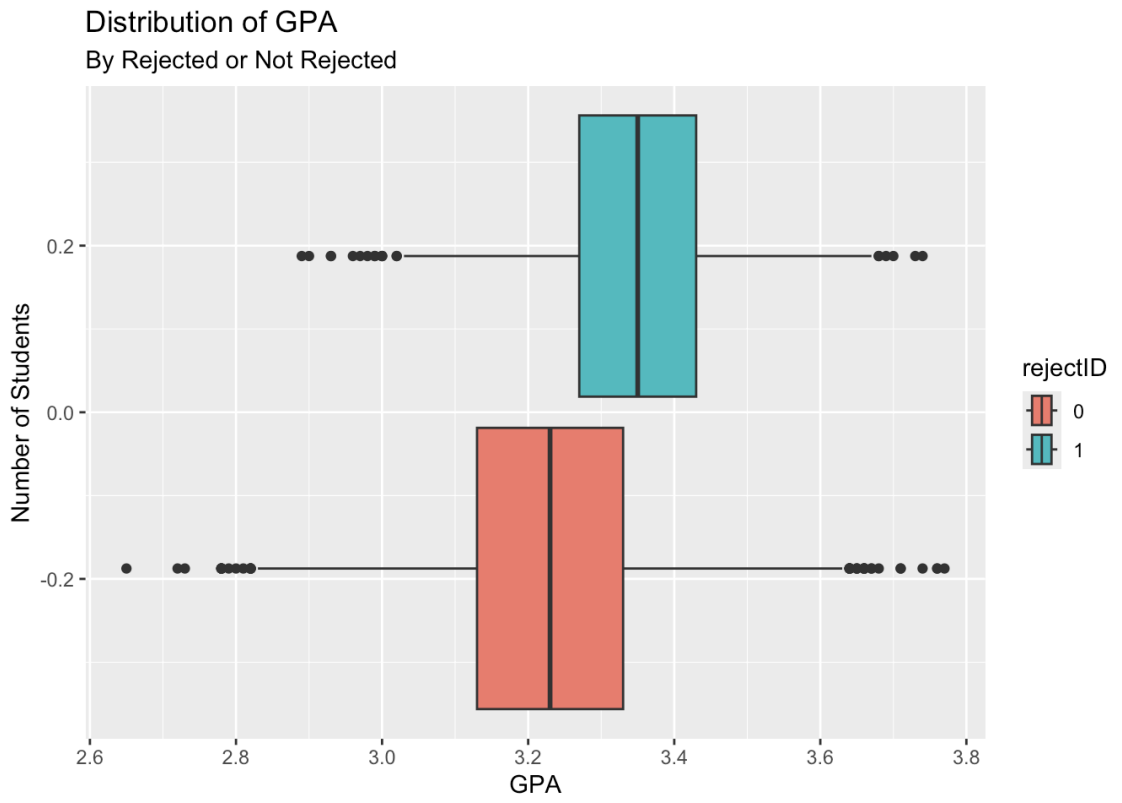

Figure A.4 - admitID distribution of gmat scores

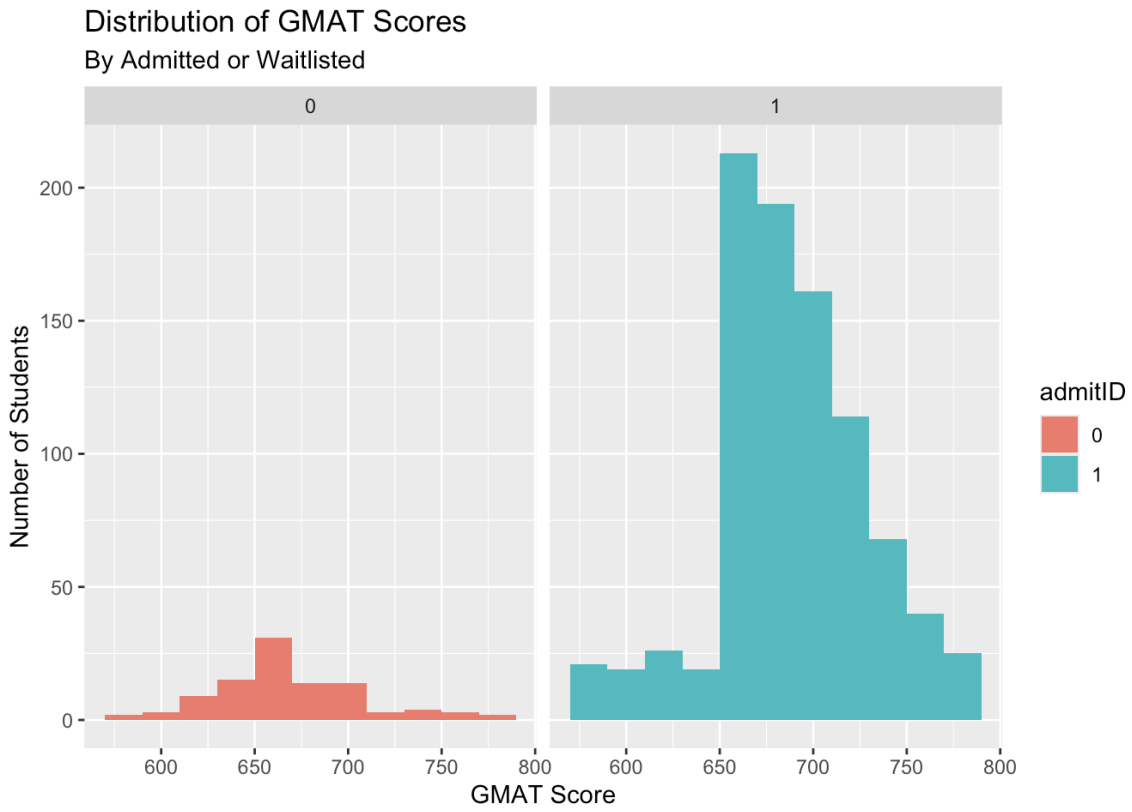

Figure A.4 - admitID distribution of gpa

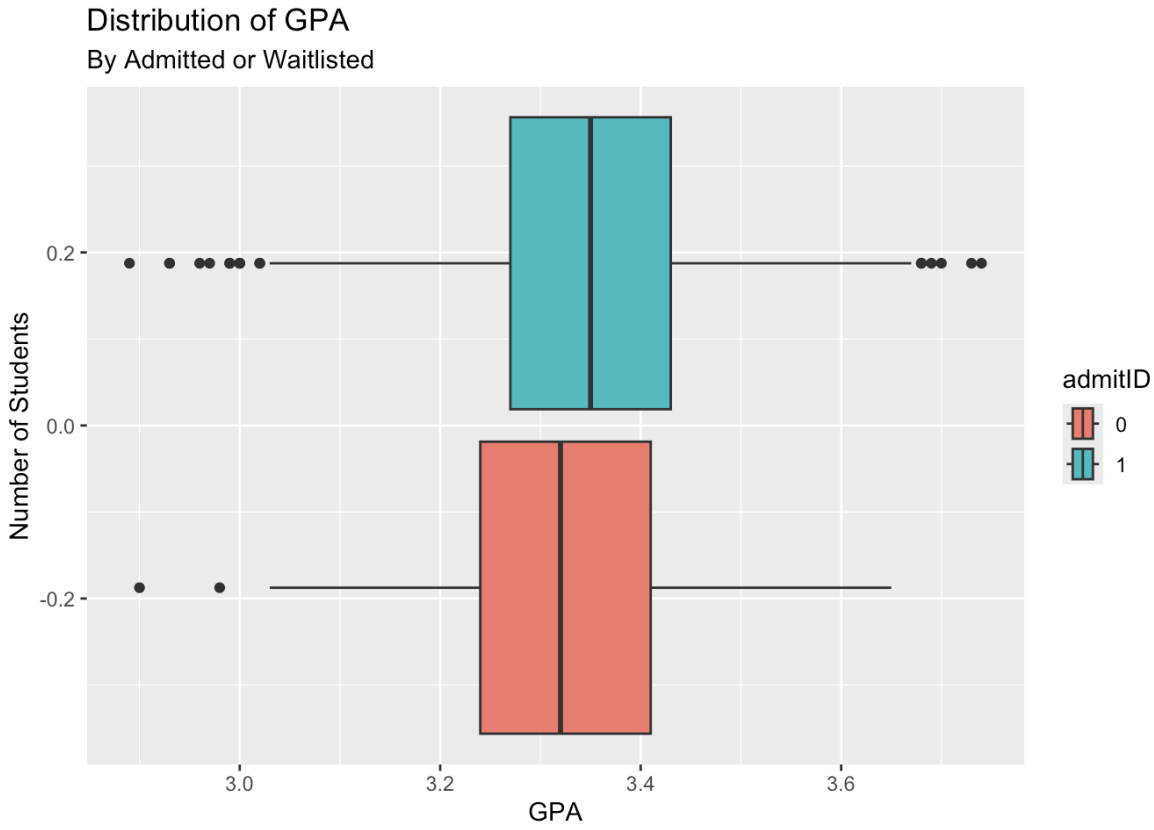

Figure A.5