A Bayesian Analysis of University of Michigan Law School Admissions

Chris Cheng, Joon Choi, Alyssa Yang

I. Introduction

This study endeavors to construct a predictive model for assessing admission outcomes at the University of Michigan Law School by analyzing a combination of demographic variables and academic indicators. Specifically, Bayesian analysis techniques will be used to incorporate factors such as LSAT scores, undergraduate GPAs, international student status, and underrepresented minority (URM) status. Drawing upon historical data from the University of Michigan Law School applicants, our objective is to develop a robust model capable of accurately projecting the admission status of future applicants.

This research seeks to deepen the understanding of the determinants influencing admission decisions, while offering practical insights for both prospective applicants and admissions committees. Through the application of statistical methodologies to institutional data, this study contributes to the broader conversation surrounding admissions practices within higher education institutions.

II. Background

Law school admissions are known to be extremely competitive, particularly at institutions like the University of Michigan which ranks among the nation's top law schools. Annually, the University of Michigan Law School receives over 6,000 applications, yet only approximately 600 applicants are admitted. In this highly competitive landscape, applicants face the challenge of building up a strong application and navigating between different qualifications. LSAT scores and undergraduate GPAs are deemed to be the most crucial academic factors that law schools take into account, while other factors like race, ethnicity, or visa status also influence admission status. Given the significance of these metrics in distinguishing candidates, prospective applicants often struggle with selecting the most suitable law schools to apply to, especially considering the financial investment required for application fees.

By delving into the statistical analysis of key admission factors, such as LSAT scores, GPAs, and demographic characteristics, this study aims to offer clarity amidst the ambiguity surrounding law school admissions. Through examining historical data and utilizing predictive Bayesian modeling techniques, we hope to create a statistical model that can assist applicants in their law school navigation process and optimize their chances of securing admission to the University of Michigan Law School.

III. Methodology

Initially, we were planning to utilize official data from the University of Michigan OBP (Office of Budget and Planning) which specifically contained demographic data of University of Michigan applicants such as their race/ethnicity, gender, standardized test scores, GPA, and their corresponding percentile, etc. However, in attempting to analyze the data, it became apparent that constructing a multiparameter Bayesian model necessitated individual-level data points rather than aggregated statistics – we needed heaps of specific demographic data for each individual, rather than generalized summary figures.

Thus, instead of using the data from the University of Michigan OBP, we utilized a dataset from a website called "Law School Numbers", which contains University of Michigan Law School applicants' individual statistics from 2003-2024 that were voluntarily supplied by the applicants. The statistics include application status, LSAT score, undergraduate GPA, and other demographic statistics and important dates relevant to the law school admission process.

However, a fundamental problem with utilizing this website was that the data was not in the form of a dataset that is readily available for statistical analysis software (i.e. a .csv file); the data was displayed in a website and was not publicly available in a desirable form. To mitigate this issue, automated data collection, processing, and cleaning was necessary, so the first step of the statistical analysis was to create an effective way of collecting the data from this website. We created our own web crawler in Python utilizing different Python libraries, specifically Selenium to automate web browsing and BeautifulSoup to parse through the raw HTML data on the website. As a result, we were able to read in and process the data into a .csv file, recording individual applicants' username, application status (Accepted, Rejected, Waitlisted, Waitlisted then Accepted, Waitlisted then Rejected), LSAT score, undergraduate GPA, whether they are a URM (Under-Represented Minority), whether they are international, and what year they applied in.

Then, for the data cleaning step, we first binarized the acceptance data into only Accepted and Rejected, and discarded the data with missing columns, disregarding the ambiguous – empty columns, or unconcluded waitlisted admission status – data. In total, there were 6,609 applicants' data over 21 years that were collected. From here, we were able to read the dataset into R and begin the process of creating our Bayesian model.

Due to the computational limitations regarding processing speed with the computers we had access to, we decided to slightly reduce the size of our dataset and take only the first 200 applicants' statistics from each year. However, some years had less than 200 applicants, so our dataset was then reduced to information on 3,472 applicants, which was a little bit over half of the size of the original dataset.

For the creation of our model, we decided to use STAN to implement our analysis, as STAN's framework facilitates probabilistic predictions and uncertainty quantification in a Bayesian standpoint, providing more comprehensive insights into acceptance probabilities based on applicant characteristics like LSAT scores and GPA. Our probability of acceptance predictions followed the form of the inverse logit function:

$$\frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}}$$

where α is the intercept, x_1 corresponds to LSAT score, x_2 corresponds to GPA, x_3 is an indicator of whether someone is a URM, x_4 is an indicator of whether someone is international, and β_1, β_2, β_3, and β_4 are their corresponding weights. Thus, the goal of our STAN model was to find final values for α, β_1, β_2, β_3, and β_4 given our historical data.

Originally we planned to create a Beta-Binomial model, but we soon realized that this type of model did not make much sense for making our predictions; while we aimed to impose robust Beta priors for α and βs, we realized that integrating this information into the inverse logit function to derive probabilities would nullify the significance of these priors. Therefore, we instead decided to use weakly informative priors

$$\alpha \sim N(0.284,\ 100) \text{ and } \beta \sim N(0, 100)$$

where the mean of α at 0.284 represents the acceptance rate of the University of Michigan Law School of 28.4%. We decided to use a large variance for both priors to provide more flexibility for our model to be influenced by the data points rather than the prior, regularize for overfitting, and allow our model to be more robust to any possible outliers.

The final implementation of our STAN model can be seen in the screenshot below. Our input data includes the number of observations and vectors holding the applicants' statistics. The parameters to be found are α and our β vector, and our final model utilizes the weakly informative priors as described above and a Binomial likelihood function.

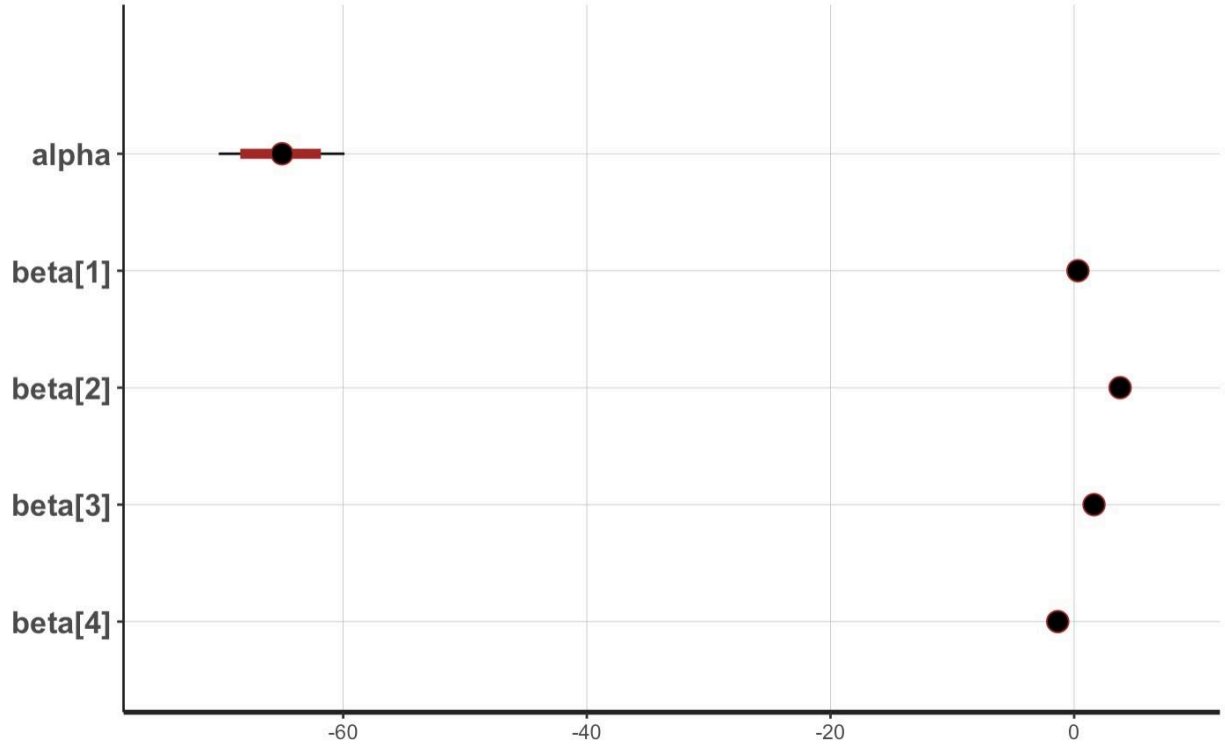
```
model_string <- "
data {
  int<lower=0> N;                  // Number of observations
  int<lower=0,upper=1> status[N];  // Accepted (1) or Rejected (0)
  vector[N] lsat;                  // LSAT
  vector[N] gpa;                   // GPA
  int<lower=0,upper=1> urm[N];     // URM indicator
  int<lower=0,upper=1> intl[N];    // International indicator
}

parameters {
  real alpha;                // Intercept
  vector[4] beta;            // Coefficients for predictors
}

model {
  alpha ~ normal(0.284, 100);         // Prior for intercept
  beta ~ normal(0, 100);           // Prior for coefficients

  // Likelihood
  for (i in 1:N) {
    real p;
    p = inv_logit(alpha + beta[1] * lsat[i] + beta[2] * gpa[i] + beta[3] * urm[i] + beta[4] * intl[i]);
    status[i] ~ binomial(1, p); // Likelihood
  }
}
"
```

In order to input our data into the STAN model, we created N = 3,472 long vectors for each of our predictors—LSAT, GPA, URM, and international—and also our response variable, their status. We still utilized a Binomial likelihood because their status is only one of two outcomes where 1 = accepted and 0 = rejected. Each entry in our status vector has a Binomial(1, p) distribution where p is the inverse logit function of each applicants' information.

The screenshot below demonstrates how we prepared the data in order to feed it to our STAN model. Specifically, we changed the status variable to either 1 or 0 to represent acceptance (as described above) and kept all others the same as given in our dataset.

```
# Prepare data
data_list <- list(
  N = nrow(data),
  status = as.integer(data$Status == "Accepted"),
  lsat = data$LSAT,
  gpa = data$GPA,
  urm = as.integer(data$URM),
  intl = as.integer(data$Intl)
)
```

After running our code on STAN, we obtained the below as our posterior distribution plot for α and each of the βs.



Thus, from our posterior samples, we obtained mean values of α = -65.07, β_1 = 0.3099, β_2 = 3.799, β_3 = 1.635, β_4 = -1.348 and our final model for predicting the probability of acceptance into the University of Michigan Law School is

$$\text{p(acceptance)} = \frac{e^{-65.07+0.3099x_1+3.799x_2+1.635x_3-1.348x_4}}{1+e^{-65.07+0.3099x_1+3.799x_2+1.635x_3-1.348x_4}}$$

where x_1, x_2, x_3, and x_4 are the same as described previously.

IV. Results

We applied our resulting model to a real-world situation by creating a hypothetical candidate with an LSAT score of 173, a GPA of 3.703, who is international but not an underrepresented minority. Plugging these statistics into the model, we obtained a probability of acceptance of around 76.69%.

To check if the results of our model makes sense, we used two approaches: comparing our hypothetical candidate's statistics to the average statistics of accepted candidates in our dataset,

and changing our hypothetical candidate's statistics to see if our posterior probability of acceptance changes according to our expectations.

For the first approach, we filtered our dataset to contain only those whose status was accepted. From there, we found that the mean LSAT score of these candidates was around 170.14, and the mean GPA was around 3.717. Thus, our relatively high probability of acceptance for our hypothetical candidate makes sense because their LSAT score and GPA are both higher than that of the means.

For our second approach, we first tried lowering the LSAT score to 160. This is much lower than the mean LSAT score of accepted individuals, so we would expect that the probability of acceptance drops drastically. Our model confirmed our expectations, outputting a probability of acceptance of just 5.51%. We then reset the LSAT score back to 173 and lowered the GPA to 3.5 which is also much lower than the mean, and we would again expect the probability to drop. Once more, our model confirmed our expectations, outputting a probability of acceptance of 60.33%. Then changing our hypothetical candidate to be non-international, our model outputted a probability of 92.69%, and changing them to be a URM outputted a probability of 94.37%. Both of these results were similar to our expectations; thus, we can conclude that our model outputs predictions aligning with our expectations and is reliable.

V. Conclusion

Through the development and implementation of our study, we were able to successfully create a model that predicts the probability of acceptance for a candidate applying to the University of Michigan Law School. Through the analysis of law school applicants in previous years, we were able to find academic and demographic trends utilizing a STAN model in R and inputting our findings into an inverse logit function to output a probability of acceptance given new and unseen data.

Analyzing the results of our model, we can see that changing the LSAT score had the most impact on the predicted probability, dropping it by around 70%. Thus, we can conclude that LSAT score is the most powerful predictor in our model, and candidates should focus on this aspect of their application the most when aspiring to gain admission to the University of Michigan Law School. With a lower LSAT score compared to the median LSAT score for accepted candidates, one should reconsider applying to the University of Michigan Law School unless their undergraduate GPA is extremely high, as the rate of acceptance will be substantially lower for them.

While our model demonstrated its reliability and accuracy within its predictions, given more time, we could have expanded the scope of our model and added more predictors. Factors such

as extracurricular activities, family history, and personal essays or letters of recommendation contribute greatly to university acceptance decisions, and by integrating factors such as these, we could have allowed our model to better reflect the multidimensionality of candidates and their qualifications.

In conclusion, our study highlights the importance of the utilization of statistical methods in order to contribute to the analysis of admissions practices within higher education institutions, and allows future students to predict their probability of acceptance into the University of Michigan Law School by comparing their demographic and academic records to those of previous applicants.