IE University
Team Report

Group
Juan Gil
Karl Juhl
Joaquin Maroto
Alvaro Pernia
Oriol Vall

Research Question

What Variables Affect the Acceptance of Students into Masters' Degrees Within the United States?

Table of Contents

Abstract

This study attempts to create a model for the purpose of predicting chance of admission to a master's program at university based on three academic variables; GRE score, TOEFL score, and cumulative GPA. The data was extracted from Kaggle and processed and used to develop the model with R Studio. The model was determined to explain roughly 78% of the data using the model which is proficient enough for the goals set. GPA was found to be by far the most relevant variable when predicting chance of admission.

KEY WORDS

GRE Scores / TOEFL Scores / University Rating / Statement of Purpose and Letter of Recommendation Strength / Undergraduate GPA / Research Experience / Chance of Admission

Introduction

To begin with, the top universities name their admission process as "holistic". Looking at all the components of the application, such as essays, recommendations and extracurricular activities and giving a major importance to grades and test scores. An admissions' survey was carried out on the top 75 colleges within the United States (based on their admission rate) and evidence was found which showing there are several unknown patterns and factors which alter the entrance approval. Many universities mainly focus on qualifications and later, take into account the rest of the qualities of the applicants. Nevertheless, there is a minor group which

focuses on the students who fit best the college needs and culture, specially in degrees such as liberal arts.

In most cases, these universities try to support students coming from minority groups and those of them who can't have enough money for paying all the costs of their bachelor's degree. By offering scholarships and promoting diversity within their campuses. Rachel B. Rubin, a PhD scholar in education at Harvard, carried out a research to reveal the variables affecting college acceptance. In her research, she carried out an experiment where she asked universities and colleges about the value that universities take most into account. Most universities stated that their main focus is on grades (GPA, IB, A Levels) and external exams (SAT, TOEFL, GRE) for both domestic and international students (Gupta et al. 2016). The best institutions of the country admit to have created algorithms and formulas to choose those students whose academic performance outstand from the majority. In this research, we will look at the different variables which determine the acceptance in Masters Programs within the top universities in the United States. The parameters included are : 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1). This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

But high competence between applicants leads to take more variables into account. Rubin conducted a survey which was later followed by interviews, asking 63 out of the 75 top universities, which variables they take into account for the applications further considerations to

covet slots. The responses were anonymous to ensure honesty and decrease bias. The responses stated that even-though most of these schools call themselves “holistic”, each university takes more into account one variable. Rubin et al. (2002) stated that “Contrary to public opinion, selective institutions are highly systematic with regard to their admissions processes and practices within individual institutions. However, there is a great deal of inconsistency across institutions, potentially creating the illusion that student selection is arbitrary.”

Most of the universities (76% of the responses) look for some type of academic outperformance (e.g. Higher Honour Roll). Which varies between the difficulty of the courses attended, external test scores… On the other hand, there exists a minor group of these universities (24% of the responses) which focus their attention on instructors’ recommendations and personal statement (1,5000 essay), where they look for qualities in the applicant which could fit the university’s culture. The Harvard student specified that this minority group does not ignore academic excellence but it is mostly taken into account for the first cut. These minor groups, which rely their acceptance on institutional fit, are represented by two variables: remarkable talents (sports and music) and minority status. Which is extremely controversial as the Supreme Court imposed a law preventing racial discrimination in advanced education. Founded on her research, Rubin indicated (Aquino, 2017):

> "When an applicant has an exceptional talent (e.g. music, athletics) or is part of a severely underrepresented group at the institution, the applicant may not compete for admission against the larger applicant pool. Instead, he/she may compete only among those with the same talent or within the same group. In these circumstances, sets of

> applications are considered separately based on a university's institutional needs. As a result, disparities may arise between the levels of academic merit of certain subgroups of students. One private university dean noted, 'The hardest part is that everyone [in the school community] wants more of something and it's a balancing act -- it's a zero sum game. Size [of the school] is fixed, but faculty, trustees, etc., want more students of color, more athletes, more great pianists.... But who will you cut out to have more of those people? We get so many of those really strong kids who don't have that extra something.... It's starting to make the world angry with us.'"

Rubin's findings face a Supreme Court's law imposed by Justice Sandra Day O'Connor in 2003. She banned racial inequality in the admission process case of the University of Michigan (UM) who stated that "to be narrowly tailored, a race-conscious admissions program cannot use a quota system – it cannot 'insulat[e] each category of applicants with certain desired qualifications from competition with all other applicants.' *Bakke, supra*, at 315 (opinion of Powell, J.). Instead, a university may consider race or ethnicity only as a 'plus in a particular applicant's file,' without 'insulat[ing] the individual from comparison with all other candidates for the available seats.' *Id.,* at 317. In other words, an admissions program must be 'flexible enough to consider all pertinent elements of diversity in light of the particular qualifications of each applicant, and to place them on the same footing for consideration, although not necessarily according them the same weight."

Finally, Rubin (2001) indicated in an interview conducted in 2011, that her findings corroborates that minority students have higher chances in a similar based condition of getting

into the university. The issue could lead to legal sanctions towards the top ranked universities. But she says that “it happens most often not for minority students but for students who can pay full tuition”. Universities reward families who make a bigger monetary effort to literate their children and compete in the full pond. In 2014, *Higher Ed* survey of admissions directors corroborated her findings, which have been taken into account as the basis for future studies. Therefore, we hypothesise that the most significant parameter for the acceptance rate of students in Master Programs within the US is going to be the Undergraduate GPA obtained.

Methodology

First of all we needed to find a reliable source to acquire our data. We used Kaggle is globally recognised source for datasets where you can different types of datasets (from GDP Per Capita between countries to Blood Pressure of Kids under 12). Data scientists use machine learning technology to upload these datasets. Furthermore, their reliability lies on unbiased datasets. To begin with, we organized our data to make it more clear and easily work with it. Then, after analyzing variables, some of them needed to be changed as some countries had “NA” data. Similarly, others couldn't be analyzed because even though, the data was given in numerical values, Rstudio could not read them as numbers, and we had to give the instruction to do so.

After changing and conditioning the data, we used the function “describe”. to get an overview of the data. After, we created boxplots to check for outliers or if there was any high skew data. There weren’t any outliers and therefore, data didn’t have to be quantized between 0 and 1, to be analyze. After the data was ready to be sued, we correlate all the possible variables and created graphs to visualize the data to easily understand the obtained results.

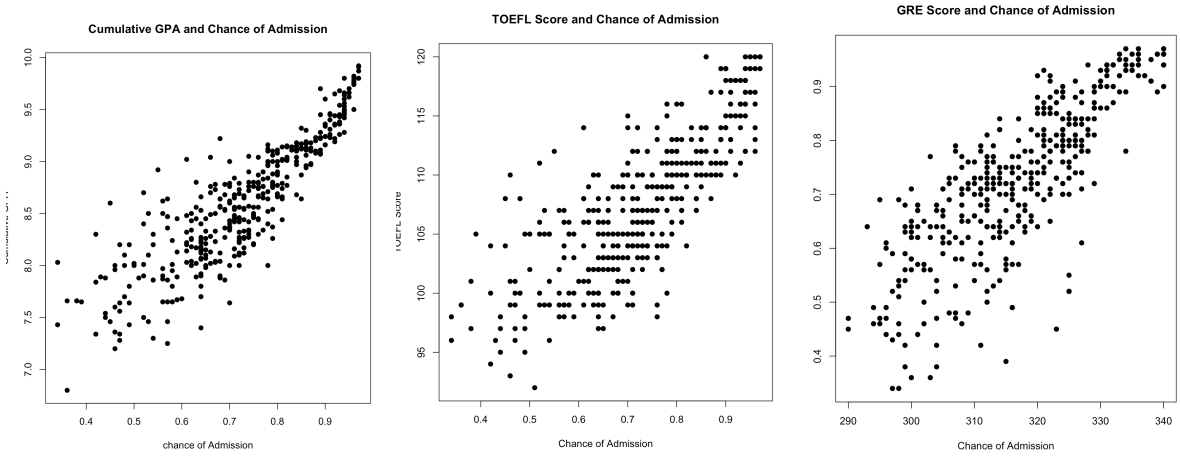
After correlating each variable one by one, we took all the numeric variables and created one correlation plot containing all the variables. The correlation value and plot showed us that there is a strong correlation relating the chance of admissions and the CGPA (cumulative grade point average). We used cgpa, gre, and toefl variables to predict chance of admission which are the ones with the highest correlation. We didn't take into account the rest of the other variables as they became part of the second plane. Once that we obtained these results, we made our regression model only taking into account the variables mentioned before. We calculate how much influantibility the variables have, known as the betas and coefficients. We found the adjusted R-squared to be 0.783 which means that it is good model, 78.3% of the data is explained through the new model.

One of the last things made to this test was to k-fold cross validation. We divide the data in two datasets, where one dataset was used for training and the other for testing. After, we analyzed and visualized the results from the estimation of the importance of the variables. Then, we created a predictive model estimation. Finally we validated the model that was just created by repeating the method mentioned before. Once that became validated we were ready to make estimations (later provided in the results and discussion). Our final step in Rstudio was to check for multicollinearity, we used the the code "vif()" command in order to check for this.

Results

As explained in the methodology, in order to correctly analyze our data we used several techniques which allowed us to have better knowledge about our variables and results. The cross validation table was essential to obtain the information used for the outcome. We found out that

there is a strong and positive correlation between all the independent variables studied, and the chance of acceptance.



Figures 1,2 & 3 (Acceptance chance and GPA, GRE, TOEFL)

To begin with, the root mean square was of 0.069, very close to 0. It tells us that there is no significant spread in values of the dependent variable around the regression line. With a 0.78 of Rsquared value showing us our model to be accurate, and therefore, reliable. This is backed up by the mean absolute error which was 0.0512. We predicted that if a student scored full marks in all exams, he would have a 100.9% chance of being accepted in university. This shows us that our model is not perfect because it is impossible to have 100.9% chance.

Scoring the lowest grades in all 3 exams, we predicted a 35% chance of being accepted in university with a 95% confidence interval. Scoring average grades in exams, being 316.81/340 for the GRE, 107.4/120 for the TOEFL and a cumulative GPA of 8.5/10 there is a 72% chance of being accepted, also with 95% confidence.

Discussion

The findings of this study show that when referring to college acceptance the independent variables with the highest importance, that is, with the bigger correlation where the GPA, TOEFL, and GRE. scores and therefore when applying to a university those are the factors taken more into account by colleges. These results are the ones predicted previous to the experiment in our hypothesis where we stated that the gpa is be the most effective predictor of chance of admission, which was the one between the three variables with the highest correlation to the chance of acceptance. Also, we created and predicted model for the acceptance chance using those three variables combining them and using the mean, maximum and minimum values of the dataset to check the probability of being accepted depending on the values of these three factors. The results showed that, as being the highest correlation factors, the models using the mean and maximum values resulted in high percentage and the one model with the minimum values rather low percentage but not compared to how high the previous ones where, in other words, the prediction for high values had a much bigger percentage accordingly than the lower values had a low percentage. Not only that, but then when creating our model for prediction in acceptance we found out that it was pretty accurate although it had some flaws that could be improved on further recreations of this one. This was due to the fact that when predicting the model we only used these three variables with the highest correlation whereas if we had used those variables with a lower correlation the percentage of acceptance predicted would have been inferior to 100%. In addition, some limitations and improvement to be done would be to mix up the maximum and minimum values of the variables in the model as the predictions made only use either the maximum for all three factors or the mean or minimum instead of mixim them into

using the mean,maximum and minimum one of each. This could be done to check the acceptance chance in case of having a bad toefl exam or GRE. score.

Work Cited and Data Sets (Not APA Reference Yet)

Keller, M. C., Garver-Apgar, C. E., Wright, M. J., Martin, N. G., Corley, R. P., Stallings, M. C., . . . Zietsch, B. P. (n.d.). The genetic correlation between height and IQ: Shared genes or assortative mating? Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3617178/

Average height of men and women worldwide. (n.d.). Retrieved from https://www.worlddata.info/average-bodyheight.php

IQ by Country. (n.d.). Retrieved from https://www.worlddata.info/iq-by-country.php

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5041406/#CR4

Harris, M. A., Brett, C. E., Deary, I. J., & Starr, J. M. (2016, September 29). Associations among height, body mass index and intelligence from age 11 to age 78 years. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5041406/#CR4

(n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/7836726