125 YEARS AT THE HEART OF ENGINEERING EDUCATION
ASEE
2018 ASEE ANNUAL CONFERENCE & EXPOSITION
JUNE 24–27, 2018
SALT PALACE CONVENTION CENTER
SALT LAKE CITY, UT

Paper ID #23648

Identifying the Best Admission Criteria for Data Science Using Machine Learning

Dr. Anahita Zarei, University of the Pacific

Dr. Anahita Zarei earned her PhD in Electrical Engineering from University of Washington, Seattle in 2007 and subsequently took up a faculty position at department of Computer and Electrical Engineering at University of the Pacific. In 2014, she joined the Data Science program where she has been teaching courses in Statistical Learning, Machine Learning, and Research Methods. Her research interests include signal processing and application of computational intelligence.

Richard Hutley, University of the Pacific

Director and Professor of Practice, Data Science at the University of the Pacific. CEO of Stratathought, and former Vice President of Innovation at Cisco Systems. Prior to joining Cisco, Mr. Hutley was the Chief Information Officer of Concert Communications, a division of British Telecom.

©American Society for Engineering Education, 2018

Identifying The Best Admission Criteria for Data Science Using Machine Learning

Anahita Zarei, Rick Hutley
azarei@pacific.edu, rhutley@pacific.edu
Data Science
University of the Pacific

Abstract

Utilization of analytics by a large array of industries has attracted many people from diverse academic backgrounds to pursue a degree in Analytics and Data Science. One of the challenges facing the admission committee over the past few years has been the selection of best criteria used for student admission. The objective of this study is to identify a set of rules based on previous admission decisions and achievement of admitted students to capture the characteristics of a successful admission. This study considers statistical and machine learning techniques to provide a better set of guidelines for future admission processes.

Introduction

Big data is taking the world by storm. What can be achieved today with the abundance of data and the available technology is extraordinary. As a result, data analytics has become a game changer in a growing number of industries: Healthcare analytics has the potential to reduce cost and improve the quality of patientcare; insurance companies use data analytics for risk assessment and fraud detection; legal analytics has made it possible for law professionals to gain deep insight from the outcome of past litigations and develop informed strategies for achieving the desired results. Due to such utilization of analytics, this field has attracted students from a variety of academic backgrounds to pursue a degree in Analytics and Data Science in recent years.

Considering this tremendous demand for data scientists, our institution launched a Masters degree in Data Science in 2014. This is a two-year program covering courses in rigorous Math and programming, as well as courses entailing soft skills such as visual storytelling and consulting skills.

One of the challenges for faculty on the admission committee in the past few years has been selecting the best criteria for student admission. Typically, in engineering disciplines the admission decision is based on students' performance on courses such as calculus, physics and pre-engineering topics [1]. However, due to the nature of Data Science field the applicants come from very diverse undergraduate programs. For instance, some of our top graduating students had an undergraduate degree in Creative Writing or Healthcare. We have witnessed many cases in which the admission criteria that are commonly used in other technical fields did not necessarily translate to identification of successful candidates for the Data Science program. Finding the right students who will be successful in this program is crucial both for the candidates and the university.

The objective of this study is to identify a set of rules based on previous admission decisions and achievements of admitted students to capture the characteristics of a successful admission. We apply statistical and machine learning techniques (using 4 cohorts' information for training and 1 cohort for test) to provide us with a better set of guidelines for future admission processes.

Data

The data comprised of information on 50 students that enrolled in the Data Science Masters program at University of the Pacific from 2014 to 2017. It consisted of their undergraduate major and GPA, age, gender, and their program GPA. Additionally, we extracted information such as math and programming skills, and success in the program based on the data on their transcripts. We rated their math competency prior to admission on a scale of 0 to 3. This rating was based on their undergraduate major and their performance on courses with mathematical content. The majority of admitted students were STEM majors (32% were Engineering/CS/Math and 26% were science majors). The remaining students were from a range of non-STEM backgrounds including Social Sciences, Business, Finance, and Healthcare fields. Their success in the program was measured by their post-enrollment GPA.

Table 1 provides a summary of descriptive statistics such as measures of central tendency and dispersion. Figure 1 shows a histogram of students' mathematical competency, incoming GPA, and program GPA. As it is depicted in the figure, most of the admitted students have an agreeable level of mathematical skills upon starting the program. We also note that many (60% of) students perform well (GPA $>$ 3.5) on the program. Figure 2 shows the distribution of age and gender. The age of our students has a wide range from 20 to 58 years old, 32% of which were female and 68% male.

Table 1: Summary Statistics

Variable	Mean	Median	Std	Min	Max
Age	30.36	27.0	8.77	20	58
Undergraduate GPA(0-4)	3.23	3.30	0.46	2.24	4.0
Program GPA(0-4)	3.59	3.66	0.47	2.97	3.99

Methods

Our methodology includes 3 parts:

- Statistical Analysis involving description of the data and performing hypothesis test.
- Unsupervised clustering of the data using the fuzzy c-means clustering technique to find natural groupings of students' data.
- An adaptive neuro-fuzzy inference system to create a set of admission rules that is best associated with the data.

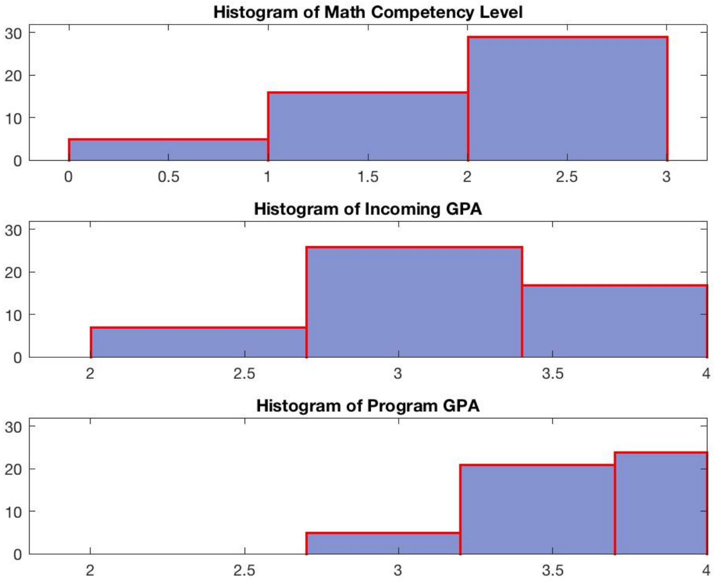

Figure 1: Histogram of Math Competency, Incoming GPA, and Program GPA

Statistical Analysis

One casual observation made by faculty was that students who were recent graduates seemed to perform better than those who had a longer gap between earning their bachelor's degree and enrolling in the Data Science program. We investigated if there was indeed an association between the number of gap years between graduation and our graduate program and success in the program. This analysis was conducted through the Wilcoxon rank-sum test.

Specifically, the investigation examined the median difference in gap years between those whose program GPA was in the bottom 10% and rest of the students. Table 2 shows the values of median and the p-value. Although students with a weaker performance had a longer gap than those who had a higher GPA, this difference was not statistically significant. Thus, since the p-value is larger than the threshold value of 0.05, we fail to reject the null hypothesis. We conclude that there is not enough evidence to state that those with a weaker performance in the program have a longer gap between graduation and starting the Data Science program.

Upon closer examination of the data, we noted that while those who were fresh graduates did indeed perform better in math and programming courses, this superior performance was not

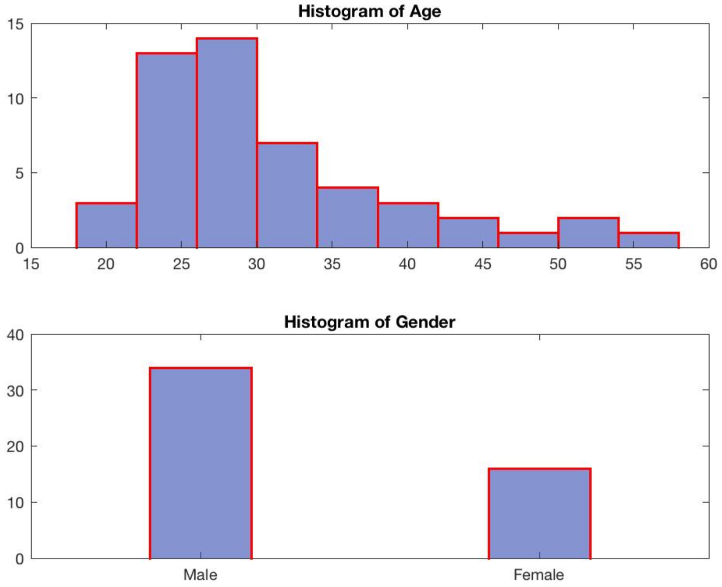

Figure 2: Histogram of Age and Gender

Table 2: Hypothesis Test Summary

	Bottom 10%	Top 90%
Median Gap Years	6 years	3 years
p-value		0.442

necessarily translated to a higher GPA. The Data Science program comprises a diverse set of courses some of which include mastery of soft skills such as project presentation and storytelling. Many students who had longer gap years were indeed seasoned professionals in their respective fields and thus had the opportunity to master such skills. Therefore, they performed very well on courses that assessed these skills, which improved their overall GPA.

Data Clustering

We applied fuzzy c-means (FCM) [2] clustering to identify "similar" groups of students based on incoming GPA, level of math competency, and their success in the program as reflected by their program GPA. The purpose of clustering is to find hidden patterns in the data and form groupings such that students within one group have a higher measure of similarity than students in any other group. Using FCM clustering a student belongs to every cluster to a certain degree, rather than having a rigid categorization. This provides great flexibility, as explained in the next section.

Figure 3 shows the clustered data and cluster centers. In this figure, we classified each student into the cluster with the largest membership value. Table 3 provides the coordinates of the cluster centers. This clustering sheds light on the underlying structure of data. All cluster centers are prototypical data points that illustrate a characteristic behavior of the data.

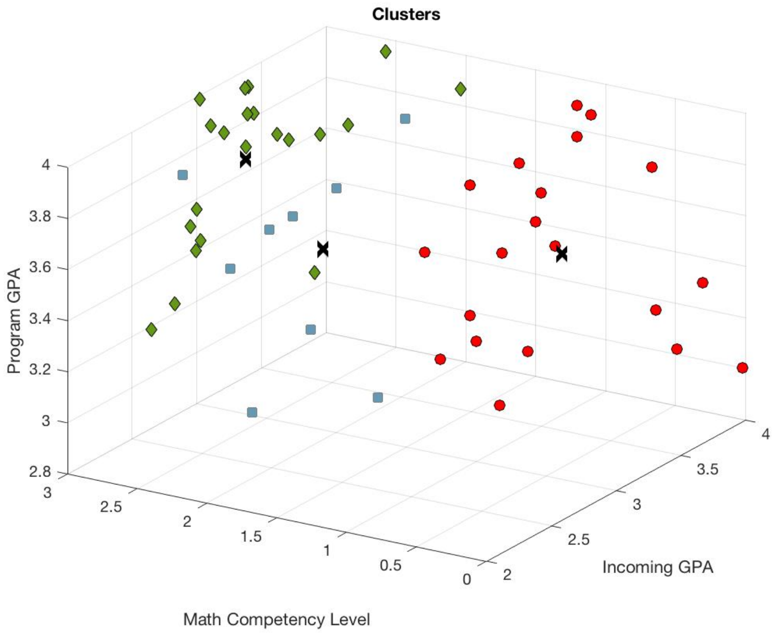

Figure 3: Data Clusters

Table 3: Cluster Centers

	Incoming GPA	Math Competency	Program GPA
Center 1	3.26	2.89	3.69
Center 2	2.84	1.94	3.56
Center 3	3.42	0.77	3.52

It is interesting to note that the cluster center with the highest program GPA is associated with the highest math competency level among the three cluster centers, but not the highest incoming GPA. We used these cluster centers to identify the rule base in the next section.

An adaptive neuro-fuzzy inference system (ANFIS)

The main objective of this study was to use data to improve the guidelines for student admission. We used a neuro-fuzzy approach to accomplish this task. We provide a brief overview of fuzzy inference below but encourage interested readers to consult one of these references [3]-[4] for a complete overview of neuro-fuzzy approach.

The motivation for using the fuzzy approach was the inherent imprecision and uncertainty in attributes such as competency, preparedness, success, etc.. Such features may not be best described by crisp values, but rather by fuzzy sets. Unlike crisp sets where a member either belongs to a group or not; in fuzzy sets, a member can partially belong to multiple groups. Fuzzy divides the data to different linguistic categories. Membership functions are then used to show the degree of membership of each member of the data set to these subjective linguistic concepts. They create a smooth transition between the members and non-members of the fuzzy set, unlike the binary membership in traditional logic.

We decomposed the universe of discourse of incoming GPA, math competency, and program GPA to the following set of linguistic terms: $\{$*low, average, high*$\}$. Figure 4 shows the membership functions for the input variables, namely incoming GPA and math competency. For example, the membership function for math competency reflects the degree of the membership of numbers in the [0,3] interval to each member of the fuzzy subset $\{$*low, average, high*$\}$. It should be noted that the choice for membership functions parameters is not based on heuristic observations by faculty in the admission committee. Rather, it is optimized by neural networks such that the overall system output, namely program GPA, is most consistent with the data.

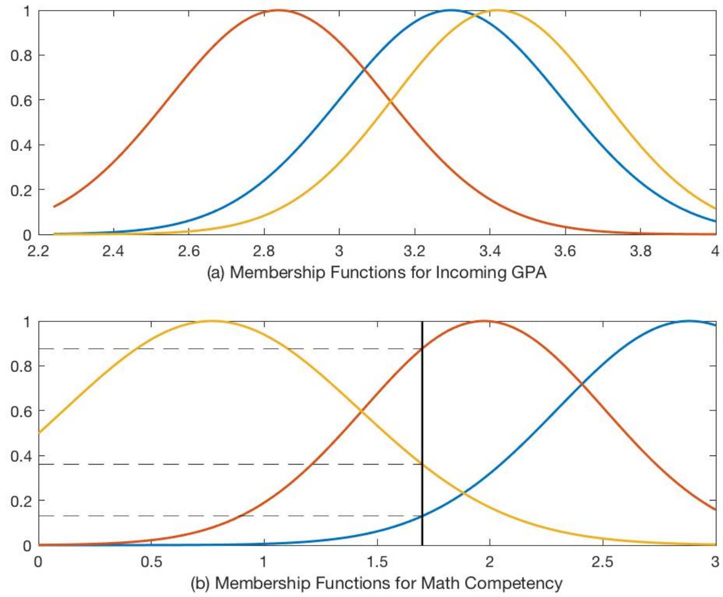

Figure 4: Membership Functions

Fuzzification is the process of converting the numerical values to the linguistic expressions by means of the membership function. As an example, we use membership function in figure 4b to fuzzify the crisp value of math competency level of 1.7. We obtain

$$\mu_{mathCompetency}(1.7) = \{\mu_{low}, \mu_{average}, \mu_{high}\} = \{0.36, 0.88, 0.13\}.$$

Therefore, the crisp value 1.7 is now fuzzifed to linguistic values ‘low’ with a membership value of 0.36, ‘average’ with a membership value of 0.88 and ‘high’ with a membership value of 0.13.

The fuzzified inputs are then inputted to a Sugeno Fuzzy Model [5] that generates fuzzy rules systematically from a given set of data. One of the important steps in fuzzy inference design is deriving the rule base and selection of membership function parameters. We used fuzzy c-means clustering as explained in the previous section to identify the rule base. The clusters found in the data identify groups of students that map into an associate class. Therefore, we translated each cluster center into a fuzzy rule and subsequently optimized the model through the adjustment of the parameters. This was done to achieve a set of rules tailored to students' data by means of Adaptive Neuro Fuzzy Inference System (ANFIS) [6].

Our system produced the following rule set:

1. If (incomingGPA is Average) and (mathCompetency is High) then (programGPA is High)
2. If (incomingGPA is Low) and (mathCompetency is Average) then (programGPA is Low)
3. If (incomingGPA is High) and (mathCompetency is Low) then (programGPA is Average

It is intriguing that the first derived rule based on the data indicates that the factor contributing to the highest performance in the program is a high level of math competency rather than a high incoming GPA. This rule guides committee members to put more weight on an applicants' major as well as their performance in their undergraduate math courses instead of their overall GPA. This rule is also consistent with the initial cluster centers that were found in data prior to training.

Figure 5 depicts an example of the inference process of the fuzzy Sugeno model for a student with an incoming GPA of 3.0 and math competency level of 2.2. The first two columns on the left show the fuzzification of each input variable to the corresponding linguistic terms. The last column depicts location and firing strength (dark blue) of the output spikes corresponding to the expressed rule. The firing strength for each rule is the product of membership values evaluated at input values of 3.0 and 2.2. The final output (program GPA) is found by taking the weighted average of each rule. The inference system predicts a GPA of 3.53 for this student.

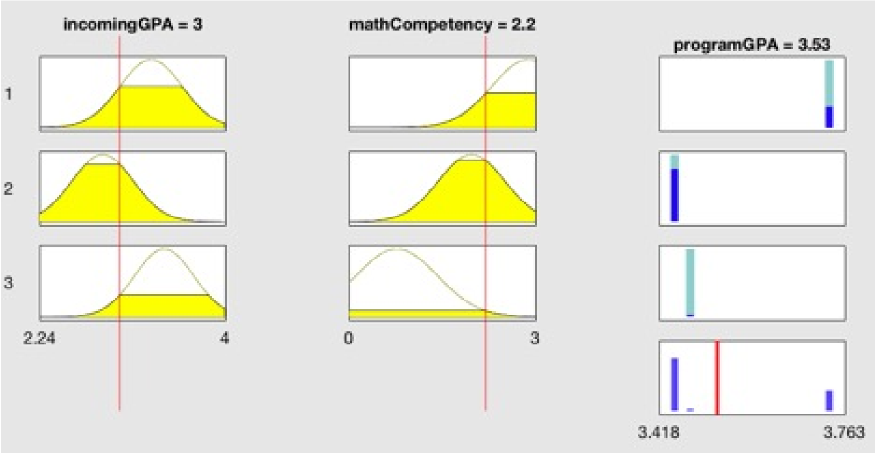

Figure 5: Inference Process Example

Figure 6 shows the nonlinear surface of the fuzzy inference system for the program GPA. As can be concluded from the figure, the lower values of GPA and math competency correspond to lower points on the surface. As these values increase, the surface smoothly transitions to higher values of program GPA.

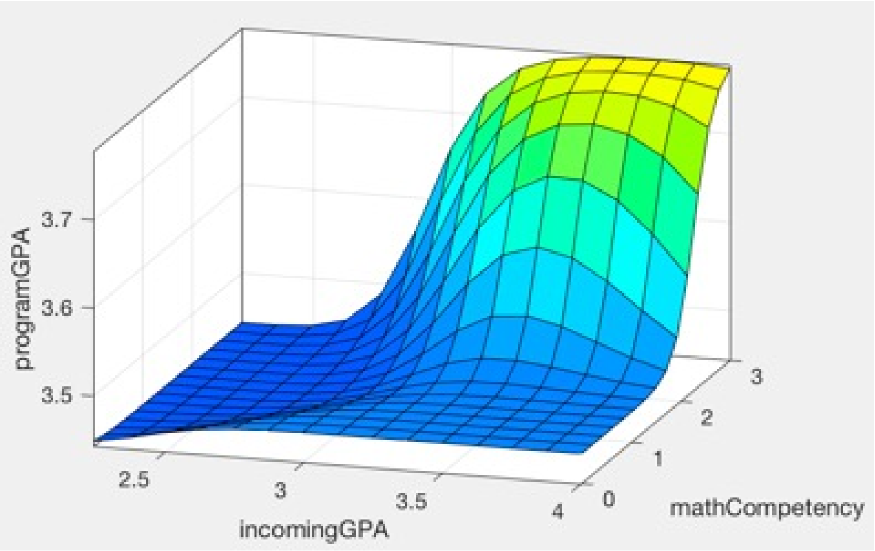

Figure 6: Program GPA as a function of Incoming GPA and Math Competency

Model validation is an important part of training. To monitor the training process, we employed 4/5 of the data for training and 1/5 for validation. Training the network for 10 epochs resulted in the minimum validation error. Table 4 shows the values of Mean Absolute Error (MAE) of models before and after training. We successfully reduced MAE from 0.26 to 0.15 for validation points by applying ANFIS, and optimized our Sugeno model.

Table 4: Mean Absolute Error (MAE) Before and After Applying ANFIS

	Initial Model	Optimized Model
Training	0.28	0.10
Validation	0.26	0.15

Conclusion

We employed students' data enrolled in Data Science program at University of the Pacific from 2014 to 2017 and created a fuzzy inference system that predicts their success in the program based on undergraduate GPA and level of math competency. While the exact prediction of performance of an individual in a program is complex, this system provides an extra resource for committee members to consult, following a careful review of the application data. The set of rules derived based on the data indicates that the main factor contributing to the highest performance in the program is the candidates' high level of math competency rather than their high incoming GPA. Hence, our results encourage committee members to place greater consideration on applicants' performance in their undergraduate math courses rather than their overall GPA and always consider the GPA along with the major when assessing an applicant.

Our statistical analysis indicated that while recent graduates had a better performance than those with longer gap years, the difference in the two groups' performance was not statistically significant. The Data Science program comprises a varied set of courses; some of which include mastery of soft skills such as project presentation, consultation skills and storytelling. A great number of students who had longer gap years also had the opportunity to master these skills while working in their respective industry. Therefore, this conclusion may persuade committee members to take a closer look at an applicant's work history. While competency in math and programming is the most important factor in student's success in this program, the value of industry experience and acquired soft skills should not be underestimated when making admission decisions.

Future analysis will investigate whether students admitted using the fuzzy model are statistically more successful than students admitted under the more traditional requirements.

References

[1] Aintablian and Ghirmai. Correlation of admission data to undergraduate student success in electrical engineering. In *Proceedings of 2017 ASEE Annual Conference Exposition*, 2017.

[2] A tutorial on clustering algorithms. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html, 2018. [Online; accessed 20-January-2018].

[3] Fuzzy logic tutorial. https://www.tutorialspoint.com/fuzzy_logic/index.html, 2018. [Online; accessed 1-January-2018].

[4] Robert Fullér. http://uni-obuda.hu/users/fuller.robert/nfs.html, 2018. [Online; accessed 1-January-2018].

[5] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Systems, Man. and Cybernetics*, 15:116–132, 1985.

[6] Jyh-Shing Roger Jang and Chuen-Tsai Sun. Neuro-fuzzy and control. In *Proceedings IEEE*, 1995.