Multiple Linear Regression with Applications in College Admission Rate

Peiyu Zuo*

Wuhan No. Seventeen High Schools, Wuhan, China

*Corresponding author: rulan@ldy.edu.rs

Abstract. The relevance of comprehending the factors influencing admission rates has increased due to the intensifying rivalry for college admissions. The goal of this research is to use multivariate linear regression analysis to identify and quantify the variables that significantly affect college admission rates. Using information from Kaggle, this study examines several variables, such as the cumulative grade point average (CGPA), the letter of recommendation, the statement of purpose, the university rating of the undergraduate school, the scores on the Graduate Record Examination (GRE) and the Test of English as a Foreign Language (TOEFL). This essay will use python to complete the multiple linear regression model, get the basic dataset as well as preliminary conclusions, and produce and analyze the residuals and fit tables, correlation heatmaps, plot each independent variable's relationship to the dependent variable can help to better understand how these factors affect people's response or the dependent variable, and evaluate the findings in the results. Ultimately, the results of this study indicate that the chance of admission was mostly predicted by CGPA, GRE, letter of reference, and research experience scores; the likelihood of admission was most significantly influenced by CGPA and GRE scores. By using this information to optimize their admission approach, the institution may utilize the model to assist them determine which factors are most essential in influencing an applicant's chances of being admitted.

Keywords: College admission rates, Multiple linear regression, Academic performance, Statistical analysis.

1. Introduction

Forecasting future events from available information and hints is the art of outcome prediction. A powerful tool for developing and assessing theories based on causal description and explanation is statistical modelling prediction. Using historical data to inform future projections or potential outcomes helps establish connections between the old and the new [1]. Predicting student performance is crucial for university to avoid student failure. One metric that can be used to assess student performance is the Average of cumulative grade points (CGPA), which is also a crucial factor that needs to be considered for institutional accreditation. Thus, university must develop appropriate instruments to track and evaluate students' performance and development [1].

The purpose of this analysis and study is to develop an analytical model that will enable the university to gain a competitive advantage over other universities by streamlining the first stage of their admissions process for international applicants. The university will select students with strong academic records who are likely to succeed and generate high-quality research during their studies by using this model. As a result, the school will have a competitive edge in drawing in new students and funding because of its increased graduation, retention, and reputation rates.

Here, this essay uses a series of virtual datasets found on Kaggle about the performance of undergraduate student, a total of 500 applicants who want to apply to postgraduate degrees in the United States. Based on various factors, this study will calculate each applicant's chance of being accepted into the university, which the author will refer to as the "explanatory" or "independent" variables in this case study. The various independent factors are the results of the following: (1) Graduate Record Examination (GRE); (2) Test of English as a Foreign Language (TOEFL); (3) university rating of the undergraduate school; (4) Statement of Purpose (SOP); (5) Letter of Recommendation (LOR); (6) Cumulative Grade Point Average (CGPA); and the research. Using a multiple linear regression model with python, the effects of these variables on the response or

dependent variable—referred to as "chance of admission" in the dataset in the methodology —will be examined, and the findings will be evaluated in the results.

By analyzing the OLS regression results, residuals plot, correlation heatmap, and each of the plotted variables, the results of the multiple regression analysis can be shown as CGPA, GRE scores, letter of recommendation scores, and research experience were the main predictors of likelihood of admission, with CGPA and GRE scores having the most significant impact on likelihood of admission. Based on this conclusion, more resources and guidance can be provided to applicants not only on how to improve their CGPA and GRE scores, but also to help them achieve better results in the application process. It will also help optimize admissions strategies, as CGPA and GRE scores have been found to have a significant impact on the likelihood of admission, and admissions committees can use these metrics as the main evaluation criteria. In addition, optimizing the scoring system to ensure that the weighting of these metrics is aligned with admissions decisions can also help improve the accuracy of admissions decisions.

2. Method and Model

2.1. Insight of Linear Regression

Bag examined and predicted the odds of candidates being admitted to graduate school using three different types of multiple regression models: random forests, decision trees, and linear regression. Furthermore, she evaluated these models' effectiveness using assessment metrics including mean square error and coefficient of determination. This study found that linear regression, with its lower mean square error and higher R-squared score, was the most effective of the three models in predicting a candidate's chances of acceptance [2]. The multiple linear regression model will be used in this presentation as it helps to determine the relative contribution relation between every independent variable and the dependent variable and the extent to which each variable influences the dependent variable [3].

It was proposed that exploratory data analysis or supervised machine learning may be used to accurately forecast the likelihood of admission. In order to choose and fit models, he employed machine learning methods based on exploratory data analysis approaches for cross-validation. Additionally, this author used Python to develop an output robust linear regression prediction model for evaluating graduate applicant admission probability [4]. In the findings of this paper, a linear regression predictive model will also be constructed using Python. Not only will all the collected data be statistically summarized, but the data will also be analyzed using residual plots and correlation heat maps and a linear regression study will be carried out for each of the independent variables based on the research data.

Four distinct machine learning techniques were employed by other author: random forest, decision tree regression, support vector regression, and linear regression. In addition, the performance of each algorithm was compared using a range of evaluation criteria in an effort to create the best possible predictive model that could examine the key factors affecting an applicant's chances of being admitted. The study also shows that cumulative GPA is frequently the most significant element influencing an applicant's odds of getting admitted, and that the most effective machine learning technique available for predicting graduate school acceptance rates at the moment is random forest regression [5]. The same results were obtained as in this paper: the highest correlation coefficient was found for GPA, with a p-value of zero, indicating that it has the greatest impact on the likelihood of college admission. Among the independent variables examined in this paper, there are two variables that are exactly the opposite of GPA, i.e., college ratings and SOP, which also have the smallest impact on the likelihood of college acceptance.

2.2. Benchmark Model

First, this study paper might consider building a benchmark model with full parameters that is an ordinary least-squared model. It is read as

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \epsilon_i. \quad (1)$$

Here, Y_i is the predicted "Chance of Admit" for the i-th observation. β_0 is the intercept of the model. $\beta_1, \beta_2, \ldots, \beta_7$ are the coefficients that represent the impact of every independent variable on the outcome. ϵ_i is the error term, representing the variation in the dependent variable that cannot be explained by the independent components. The dataset includes a number of parameters, as Table 1 displays [6].

Table 1. Specifications and Synopsis

Number	Variable symbols	Explanation
1	X1	GRE Scores (out of 340)
2	X2	TOEFL Scores (out of 120)
3	X3	University Rating (out of 5)
4	X4	Statement of Purpose Strength
5	X5	Letter of Recommendation Strength (out of 5)
6	X6	Undergraduate GPA (out of 10)
7	X7	Research Experience (either 0 or 1)
8	X8	Chance of Admit (ranging from 0 to 1)
9	β_0	This complete parameter multiple linear regression model's intercept.
10	β_i $(i = 1, 2, \cdots, 7)$	The multiple linear regression model's corresponding coefficients, including all parameters.

The applicant's overall undergraduate academic performance as measured by GPA indicators, the range of their GRE and TOEFL scores, the ranking of the university, the caliber of their purpose statements and the strength of their letters of recommendation, and whether they have research experience in an academic field relevant to their application are all described by every single column in the dataset's midsection. The response variable, which shows the likelihood of being admitted, is the final column. It is a continuous number between 0 and 1. And the descriptive statistics of these variables are shown in Table 2.

Table 2. The Variables' Descriptive Statistical Results

	Serial No.	GRE Score	TOEFL	University Rating	SOP	LOR	CGPA	Chance of Admit
count	500	500	500	500	500	500	500	500
mean	250.5	316.5	107.2	3.2	3.4	3.5	0.6	0.7
std	144.5	11.3	6.1	1.2	1	1	1.5	0.1
min	1	290	92	1	1	1	0	0.3
25%	125.8	308	103	2	2.5	3	0	0.6
50%	250.5	317	107	3	3.5	3.5	1	0.7
75%	375.3	325	112	4	4	4	1	0.8
max	500	340	120	5	5	5	1	1

3. Results

3.1. Regression analysis

The result of R-Squared (which is not shown in the table) in the regression model shows that the model can account for 82.2% of the variability in the dependent variable. The regression model fits the data well and is statistically significant, as the Table 3 demonstrates with a low p-value. There are

500 observations in the data source, and the model's standard error is quite small. For more details, see Table 3.

The statistical significance for every separate variable with respect to the response variable's result is indicated by the p-values [7]. A substantial influence of the variable on the one that is dependent is shown by a p-value of less than 0.05. The results of this analysis show that while University Rating and SOP are not statistically significant predictors of Chance of Admission, GRE, TOEFL, LOR, CGPA, and Research are, especially CGPA and GRE score. In summary the following model was developed

$$Y_i = -1.2757 + 0.0019X_1 + 0.0028X_2 + 0.0059X_3 + 0.0016X_4 + 0.0169X_5 + 0.1184X_6 + 0.0243X_7. \quad (2)$$

Table 3. Regression Results

	coef	std err	t	P>| t |	[0.025	0.975]
const	-1.2757	0.104	-12.232	0.000	-1.481	-1.071
GRE Score	0.0019	0.001	3.7	0.000	0.001	0.003
TOEFL Score	0.0028	0.001	3.184	0.002	0.001	0.004
University Rating	0.0059	0.004	1.563	0.119	-0.002	0.013
SOP	0.0016	0.05	0.348	0.728	-0.007	0.011
LOR	0.0169	0.004	4.074	0.000	0.009	0.025
CGPA	0.1184	0.01	12.198	0.000	0.099	0.137
Research	0.0243	0.007	3.68	0.000	0.011	0.037

3.2. Residual and Fitted Plot

If there are any non-linear patterns in the residuals of the outcome variables, it may be seen in the Fitted vs. Residual graphic, cf. Fig. 1. If the model is unable to capture the non-linear connection, this graphic might depict an implicit or projected non-linear relationship between independent factors and a response variable. If the residuals in this model are uniformly distributed around a horizontal line with no discernible patterns, then there are no nonlinearity relationships to be included in the model due to this flat trend [8].

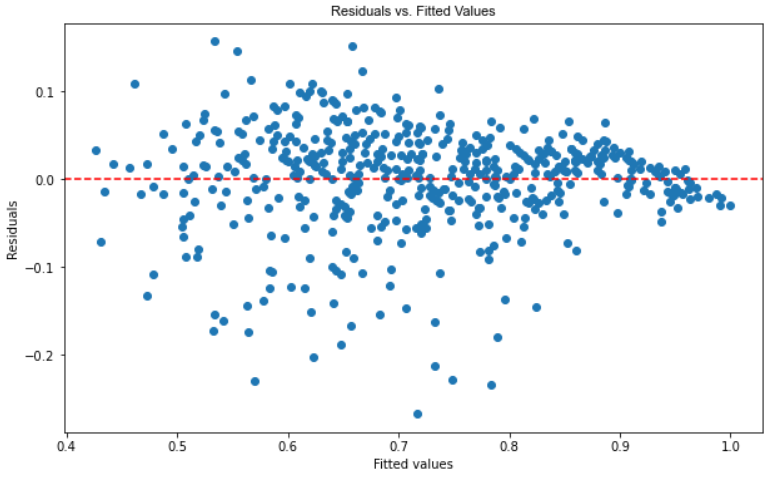

Fig. 1 Illustration of the residual versus fitted values, which used to assess the fit of the regression model.

The residual chart shows that the residuals are evenly distributed around the horizontal line at zero and that there is an obvious trend. This can be taken to indicate that the regression model fits the data well. By using this model, the university may more precisely estimate each applicant's likelihood of being admitted, which may lessen the likelihood of admitting students with a track record of subpar academic achievement. The institution will boost its reputation and retention and graduation rates by

accepting students who have a strong academic record and increased prospects of succeeding in their studies. Together, these will offer the institution a competitive advantage in luring new students and obtaining funds, which will raise its ranking among the best universities.

3.3. Correlation analysis

The correlation between the various variables is displayed in this coefficient heatmap shown in Fig. 2. As can be seen in the last column, the connection between the independent factors and the likelihood of Entrance will be the focus of people's focus. Higher values signify a greater association. The values range from 0.55 to 1. Research has the lowest correlation among the factors, while CGPA, GRE, and TOEFL scores have the highest correlations.

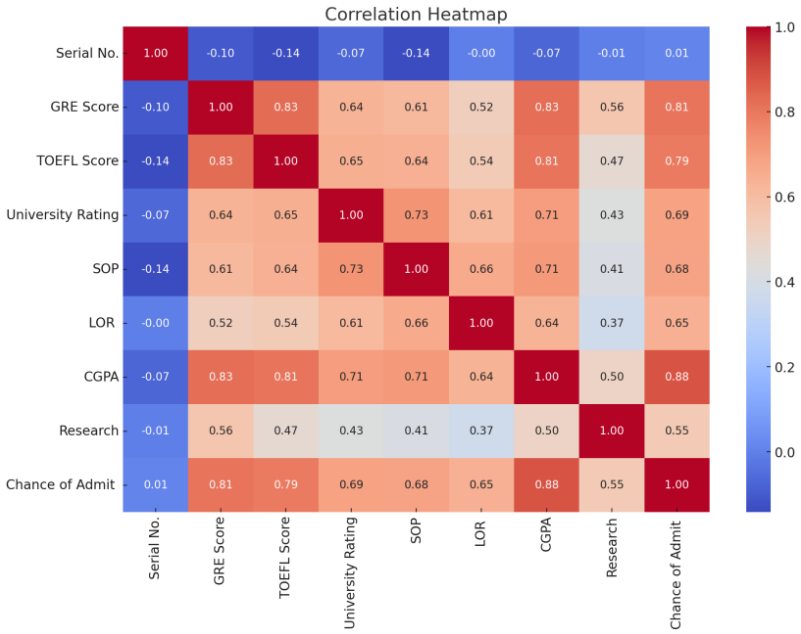

Fig. 2 This correlation heatmap visualises the interrelationships between the variables.

3.4. Plotted Variables

There is a noticeable increasing trend for every variable that is displayed, which indicates that the chance of admission rises with each increase in a single unit for that variable [9]. For instance, candidates who have written research papers before have a higher chance of being accepted than those who have not.

These scatterplots and fitted lines demonstrate the impact of several key factors on the likelihood of admission, all of which are positively correlated with the likelihood of admission. The slopes of the fitted lines show that CGPA and GRE scores have the most significant impact on the likelihood of admission, with letter of recommendation scores and research experience also having a significant impact, but to a lesser extent. These graphs shown in Fig. 3 visualize the varying degrees of influence of each factor on admissions outcomes.

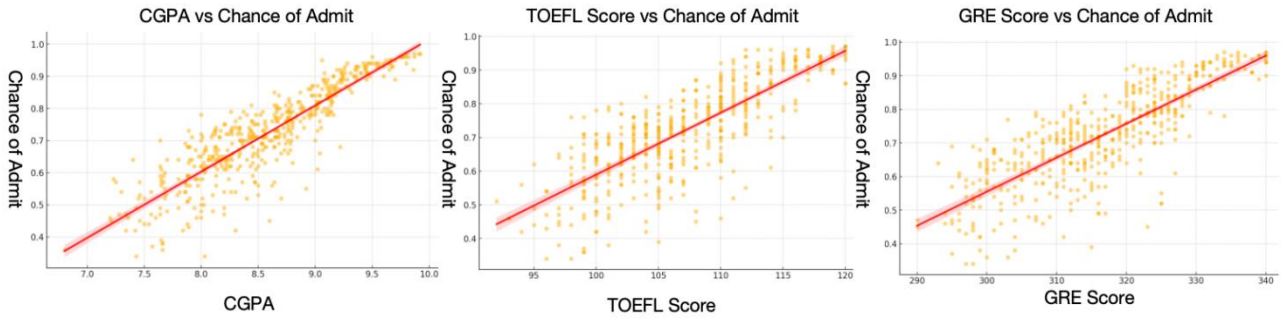

Fig. 3 These Plotted Variables shows the relationship between three variables and Chance of Admit.

4. Conclusion

Compared to its existing admissions procedure, the university will be able to make more objective and data-driven decisions thanks to the application of a multiple linear regression model in this study. This will assist the school in enhancing the equity and openness of its admissions procedure, which is crucial in the fiercely competitive academic world of today—especially given that it is in a place where thousands of students from all over the world vie for the opportunity to pursue higher education. This paper's primary finding is that the possibility of admission was mostly predicted by CGPA, GRE, letter of reference, and research experience scores; CGPA and GRE scores had the greatest influence on this likelihood. Employing this regression model allows the university to demonstrate its transparent and data-driven candidate screening process, which is in line with the growing demand from educational institutions for responsibility regarding their admissions procedures and equitable access to educational opportunities.

Future research could examine the use of multifactor models' equivalent regression forms, which are often employed in the study of financial behavior to quantify additional important both known and unknown admissions characteristics that are not addressed in this paper, such as the admissions officers' moods or emotions, and to assist prospective applicants in understanding graduate admissions factors more thoroughly and methodically. Furthermore, this study may quantify the level of unpredictability in the graduate admissions process using stochastic actuarial models. Finally, there are some factors that may affect college enrollment but are not included in the model, such as economic fluctuations, policy changes, or societal trends. These omitted variables may result in models that are not fully accurate. Future research could explore additional predictor variables such as extracurricular activities, leadership experiences, and personal statements to fully assess applicant potential.

References

[1] Rolly, T.D. Predicting Students' Academic Performance Using Regression Analysis. American Journal of Educational Research. 2022, 10(11), 640-646.

[2] Bo,L. Research on the Admission of Graduate Students Based on Multiple Regression Model. MSEA 2022, ACSR 101, pp. 681–689.

[3] Njokko, I. Linear Regression for Admission Prediction--- a Competitive Strategy Case Study. 2023, Medium. https://medium.com/@imanjokko/linear-regression-for-admission-prediction-a-competitive-strategy-case-study-ca4d1e9ed071

[4] Richard, R& Ayang, A. 2023, A Multiple Linear Regression on Factors Affecting Students' Academic Performance during Online Learning.

[5] Huang,Y. Research on the application of data mining in the analysis of predicted scores in university examinations. Anhui University. 2014, DOI:10.7666/d.Y2722871.

[6] Chari, Deepa, and Geoff Potvin. 'Understanding the Importance of Graduate Admissions Criteria According to Prospective Graduate Students'. Physical Review Physics Education Research, 2019, 15(2): 023101.

[7] Chen, Y.T.: Personality traits, emotional intelligence, and academic achievements of university students. Am. J. Applied. Psychol. 2015, 4(3): 39–44.

[8] Bibi, S., Saqlain, S.: Relationship between emotional intelligence and self-esteem among pakistani university students. Cell Dev. Biol. 2016, 6(4): 1000279

[9] Duru, E., Balkis, M.: Procrastination, self-esteem, academic performance, and well-being: a moderated mediation model. Int. J. Educ. Psychol. 2017, 6(2): 97-119.