Research on the Factors Determining the Chance of Graduate Admission

Yansong Li*

Faculty of Arts and Sciences, University of Toronto, Ontario, Canada

*Corresponding author: ys.li@mail.utoronto.ca

Abstract. Nowadays, due to various reasons, more students are beginning to pursue higher academic degrees, such as master's degrees. This article discusses the factors that influence the chances of graduate admissions. The aim is to quantify each determinant by establishing a linear model, thus helping everyone understand how much each factor can affect the admission chance. The dataset used in this article comes from the Kaggle, which includes eight variables and 400 observations. This article establishes several multiple linear models using the smallest AIC selection, the smallest BIC selection, and the LASSO selection. Models are screened based on some indicative values such as R^2_{adj}, SS_{res}, and R^2. After establishing the final model, assumptions (Normality, homoscedasticity, multicollinearity, linear relationship) and prediction errors are checked to verify the model's effectiveness. The article ultimately finds that every predictor positively correlates with the admission chances. It means that the more achievements an applicant has, the higher the chance of admission. This conclusion is consistent with our initial hypothesis. The final results can help applicants understand the importance of each application material (predictor). By inputting their existing achievements for each predictor into the model, they can predict their chances of admission, identify deficiencies, and work on improvements.

Keywords: Linear model; chances of admissions; application for master's degrees.

1. Introduction

Postgraduate degrees are becoming very popular nowadays, with more and more undergraduates choosing to continue their studies after graduation rather than entering the job market directly. Liu et al. analyzed the relationship between higher education and employment in China and showed that educational attainment can directly impact a student's subsequent employment [1]. Specifically, the higher the level of education, the higher the chance of employment and the higher the starting salary. Bisma also concluded from an analysis of a British opinion poll that another reason for pursuing a postgraduate degree might be pressure from family and society [2]. Many undergraduates say they do not want to be inferior to their peers, friends, and family, most of whom have graduate degrees.

However, no matter the reason, it is undeniable that postgraduate qualifications are becoming increasingly popular nowadays. Against this backdrop, the purpose of this article is to analyze the factors that determine graduate school admissions and to grasp the extent to which each factor determines the chances of admission to graduate school, thus helping applicants to understand which specific factors play a crucial and irreplaceable role in graduate school applications.

Judith discussed the importance of the GRE and TOEFL scores for graduate school applications. She showed that the GRE and TOEFL scores play a significant role in measuring a student's academic ability [3]. Nathan discussed that letters of recommendation are an effective predictor of graduate school performance [4]. Additionally, Nathan pointed out that combining GMAT and UGPA can significantly improve the accuracy and effectiveness of student performance prediction [5, 6]. These articles analyze different factors in detail and show their accuracy and validity in predicting students' academic ability. However, none of the articles discusses the impact of a combination of different factors on the chance of admission to graduate school. For this reason, this paper will include all of the factors (predictors) mentioned above and add some other essential factors to provide a more comprehensive discussion of how different factors work together to affect the chances of admission to graduate school.

Guilherme et al. used the statistical modeling approach of LR to predict student grades or academic performance in their article, and their study showed that the LR model has excellent performance in prediction [7-9]. Therefore, this paper will also primarily use the statistical approach of MLR to predict the impact of each factor on the chances of admission to graduate school. Similar to the statistical approach used by Shaobo [10], this paper will also quantify the impact of each factor on output by comparing the predictive capabilities of different models to determine the final relative optimal model.

2. Methods

2.1. Data Source

The dataset we used in this paper is extracted from Kaggle, which includes several parameters that are considered necessary during the application of the master's program [11, 12]. The data is mainly collected in terms of an Indian perspective and inspired by the UCLA Graduate Dataset. Since the dataset's source is from a reliable website, the model built based on this database is relatively more precise. The sufficient observations in the database also prove that it is representative to minimize the bias caused by fewer data recordings.

2.2. Data Visualization

There are 400 observations and nine variables in total, which can prove the dataset is representative. The response variable is the chance of admission, and there are seven potential predictors, including three numerical predictors and four categorical predictors. Minimum value, Q1, median, Q3, and maximum value shown in Table 1 give us a rough idea of how data is distributed.

Table 1. Summary Table of Variables

Variables	Min	Q1	Median	Mean	Q3	Max	SD
Chance of Admission	0.34	0.64	0.73	0.73	0.83	0.97	0.14
GRE Score	290.00	309.00	318.00	317.20	325.00	340.00	11.60
TOEFL Score	92.00	103.00	107.00	107.50	112.00	120.00	6.28
University Rating	1.00	2.00	3.00	3.08	4.00	5.00	1.14
Strength of Statement of Purpose (SOP)	1.00	2.50	3.50	3.42	4.00	5.00	0.99
Strength of Reference Letter (LOP)	1.00	3.00	3.50	3.45	4.00	5.00	0.91
CGPA	6.80	8.18	8.63	8.60	9.08	9.22	0.61
Research	0.00	0.00	1.00	0.56	0.56	1.00	0.50

As shown in Figure 1, the histogram shows that the overall distribution of the response variable is left-skewed. It does not satisfy the linear assumption of normality. Thus, we need to do a Boxcox transformation to improve the normality and make the model more precise.

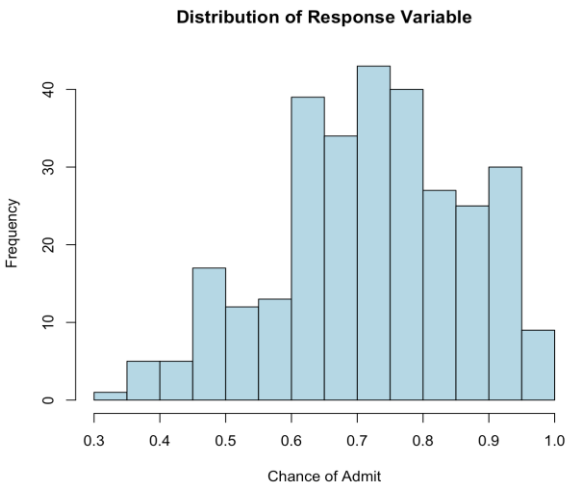

Fig. 1 Distribution of The Response Variable

Predictors 1, 2, and 3 are numerical, so we use scatterplot to present data. As shown in Figure 2, all three scatterplots show an apparent positive linear relationship, and the population of the response variable at any value of the predictor in each graph has almost the same spread. Therefore, they roughly satisfy the linear assumptions.

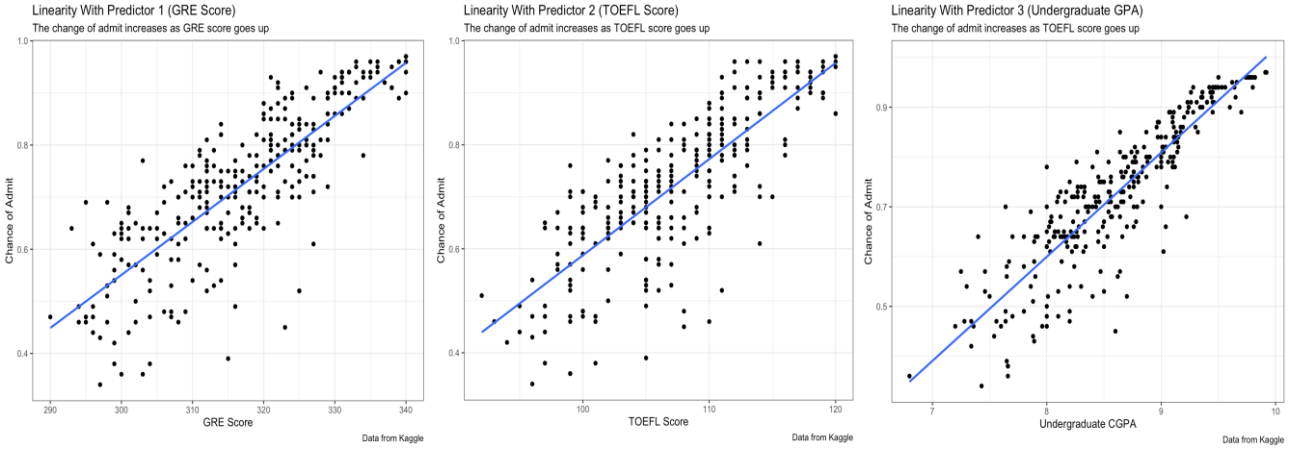

Fig. 2 Distribution of Numerical Variables

Predictors 4, 5, 6, and 7 are categorical, so we use a side-by-side boxplot to present. As shown in Figure 3, in these four boxplots, each category in each predictor has a different distribution regarding the response variable, which indicates that predictors are mutually independent. In addition, the vertical length of each boxplot in each graph is almost the same, which implies that the population of response at any category of the predictor has almost the same spread.

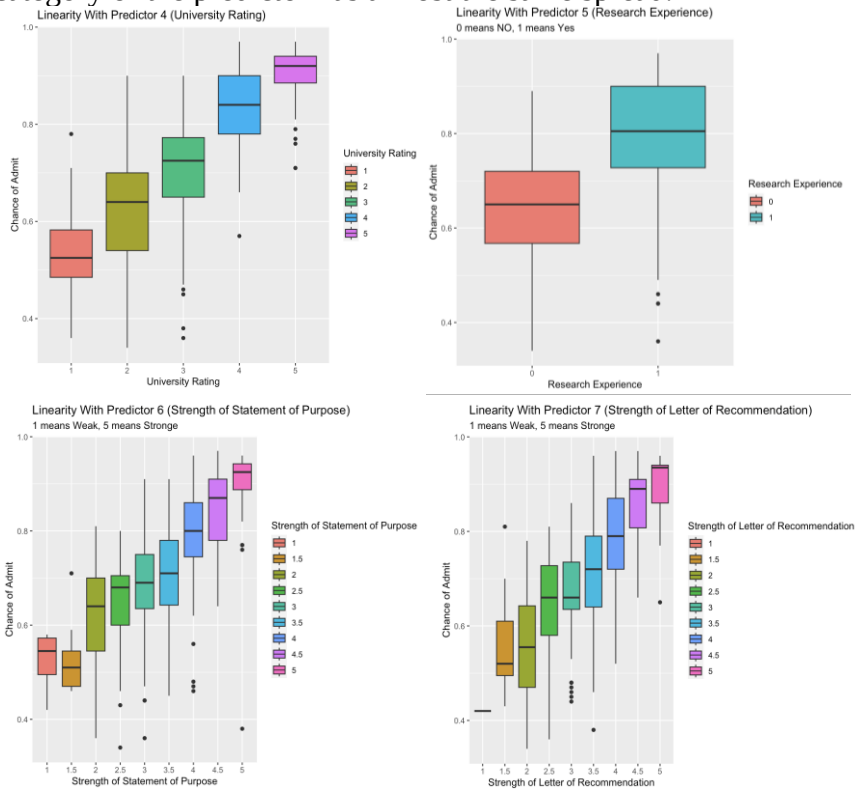

Fig. 3 Distribution of Categorical Variables

2.3. Methodology

The dataset is first loaded into Rstudio and performs the necessary data cleaning. The cleaned dataset is randomly divided into two parts, i.e., the training part (75% of the data) and the testing part (25% of the data). The training part is used to build the model and perform the analysis, while the testing part is used to check the validity and effectiveness of the model.

The model-building process is conducted based on the existing predictors. This paper will establish several models, and the optimal model will be selected by comparing the indicative values among models (such as R2, R2adj, AIC, and BIC). Specifically, four models will be created based on different methodologies. The first model contains all available predictors, the second is generated based on the rule of the smallest AIC selection, the third is based on the smallest BIC selection, and the fourth is based on the LASSO selection. A better model will be selected by comparing the indicative values of each model. To make the model more precise, it is necessary to check the significance of each predictor. If the model contains insignificant predictors, the partial F test will be an effective and reliable method to help finalize whether to keep these predictors.

Once the final model is generated, the assumptions of the model will be checked by observing the pattern of the residual plot and QQ plot. If it fails to satisfy the assumptions, some adjustments, such as boxcox transformation, should be applied to prove the model's accuracy. The outliers, leverage points, and influential points must also be analyzed in model goodness checking. The paper only needs to report these special points if they are caused by recording mistakes; no further adjustments are needed. The final step before checking validation is to calculate the VIF of each predictor, which can help to determine the multicollinearity.

In the model's validation process, the prediction error will be calculated by using the actual value of the response variable in the test part and the predicted value obtained from the final model. The magnitude of prediction error can indicate the effectiveness of the model. The final step is to change the dataset from train to test and do the same process as described above to check assumptions, outliers, leverage points, influential points, and VIF again. We expected that the characteristics based on the test part would be similar to those based on the training dataset, which implies that the final model is general instead of just specialized for the training dataset.

3. Results and Discussion

3.1. Model Building

Model 1 is fitted using all available predictors. The Table 2 shows S & SSres & R2 & R2adj & AIC & BIC for model 1. Table 3 shows P-value for each predictor in model 1.

MODEL 1:

$$\widehat{Chance\ of\ Admission} = -1.2935 + 0.0018 \times \boldsymbol{GRE\ Score} + 0.0028 \times \boldsymbol{TOEFL\ Score} + 0.0095 \times \boldsymbol{University\ Rating} - 0.0051 \times \boldsymbol{SOP} + 0.0172 \times \boldsymbol{LOP} + 0.1228 \times \boldsymbol{CPGA} + 0.02 \times I(\boldsymbol{Research}) \quad (1)$$

Model 2 is built based on stepwise selection with the smallest AIC. Table 2 shows some indicative values of this model and Table 3 shows relevant P-value of each predictor in model 2.

MODEL 2:

$$\widehat{Chance\ of\ Admission} = -1.291 + 0.0019 \times \boldsymbol{GRE\ Score} + 0.0027 \times \boldsymbol{TOEFL\ Score} + 0.008 \times \boldsymbol{University\ Rating} + 0.0153 \times \boldsymbol{LOP} + 0.1216 \times \boldsymbol{CPGA} + 0.0198 \times I(\boldsymbol{Research}) \quad (2)$$

Model 3 is built based on stepwise selection with the smallest BIC and the corresponding indicative values and p-value can refer to the table 2 and table 3 below.

MODEL 3:

$$\widehat{Chance\ of\ Admission} = -1.5419 + 0.0032 \times \boldsymbol{GRE\ Score} + 0.0204 \times \boldsymbol{LOP} + 0.138 \times \boldsymbol{CPGA} \quad (3)$$

Model 4 is generated with the LASSO selection and it has the same result as model 2. The corresponding indicative values and p-value are shown in the Table 2 and Table 3 below.

MODEL 4:

$$\widehat{Chance\ of\ Admission} = -1.291 + 0.0019 \times \boldsymbol{GRE\ Score} + 0.0027 \times \boldsymbol{TOEFL\ Score} + 0.008 \times \boldsymbol{University\ Rating} + 0.0153 \times \boldsymbol{LOP} + 0.1216 \times \boldsymbol{CPGA} + 0.0198 \times I(\boldsymbol{Research}) \quad (4)$$

R^2_{adj} indicates the percentage of variation in y (response variable) that can be explained by the regression line on average, so the higher R^2_{adj} is, the better the model is. Under other criteria, a better model can also be determined based on the lowest AIC and BIC. Comparing indicative values among the four models in Table 2, it is evident that model 2 and model 4 have the highest R^2 and R^2_{adj} and the smallest AIC. Since model 2 is precisely the same as model 4, we can finalize the frame of the best model.

Table 2. Summary Table of the Indicative Values for Models

Model No.	S	SSres	R2	R2adj	AIC	BIC
Model 1	0.0621	1.1245	0.8120	0.8156	-1661.93	-1624.60
Model 2	0.0620	1.1271	0.8196	0.8159	-1663.25	-1629.62
Model 3	0.0631	1.1768	0.8166	0.8097	-1656.29	-1633.77
Model 4	0.0620	1.1271	0.8196	0.8159	-1663.25	-1629.62

As shown in Table 3, the model shows that "University Rating" has a greater p-value based on the significance level of 0.05, so launching a partial F test to evaluate whether to keep this predictor is necessary. The result shows a p-value of 0.1367, indicating that "University Rating" is irrelevant for estimating the chance of admission, so removing this predictor can improve the model.

Table 3. Summary Table of the P-value of Each Predictor

Models	GRE Score	TOEFL Score	University Rating	SOP	LOP	CGPA	Reasearch
Model 1	0.0093	0.0229	0.0951	0.4168	0.0078	2×10^{-16}	0.0236
Model 2	0.0066	0.0301	0.1367	-	0.0110	2×10^{-16}	0.0253
Model 3	3.99×10^{-8}	-	-	-	0.0004	2×10^{-16}	-
Model 4	0.0066	0.0301	0.1367	-	0.0110	2×10^{-16}	0.0253

Note: "-" means the predictor does not include in the model.

As described in the methodology part, the histogram of the response variable tells that it does not perfectly satisfy the assumption of normality, so a Boxcox transformation should be applied to enhance the model's effectiveness. The power to the response variable under Boxcox transformation is 2. Therefore, the final model is shown below.
FINAL MODEL:

$$\widehat{Chance\ of\ Admission}^2 = -2.4558 + 0.0028 \times \boldsymbol{GRE\ Score} + 0.0046 \times \boldsymbol{TOEFL\ Score} + 0.0261 \times \boldsymbol{LOP} + 0.1749 \times \boldsymbol{CPGA} + 0.0317 \times I(\boldsymbol{Research}) \quad (5)$$

3.2. Assumption and Validation Checking

To check the assumptions of the final model. As shown in Figure 4, there is no symmetric pattern, no cluster, and no obvious fanning pattern in residual plot and scale-Location. Thus, linearity, independence, and homoscedasticity are satisfied. In addition, we can find more points in the QQ plot that are close to the QQ line, which implies that the normality of the model has been improved. Therefore, the final model satisfies the assumptions of the linear model. Cook's distance shows the influence of each observation on the fitted response values. Since several points are flagged, it indicates that some special points should be reported in the paper.

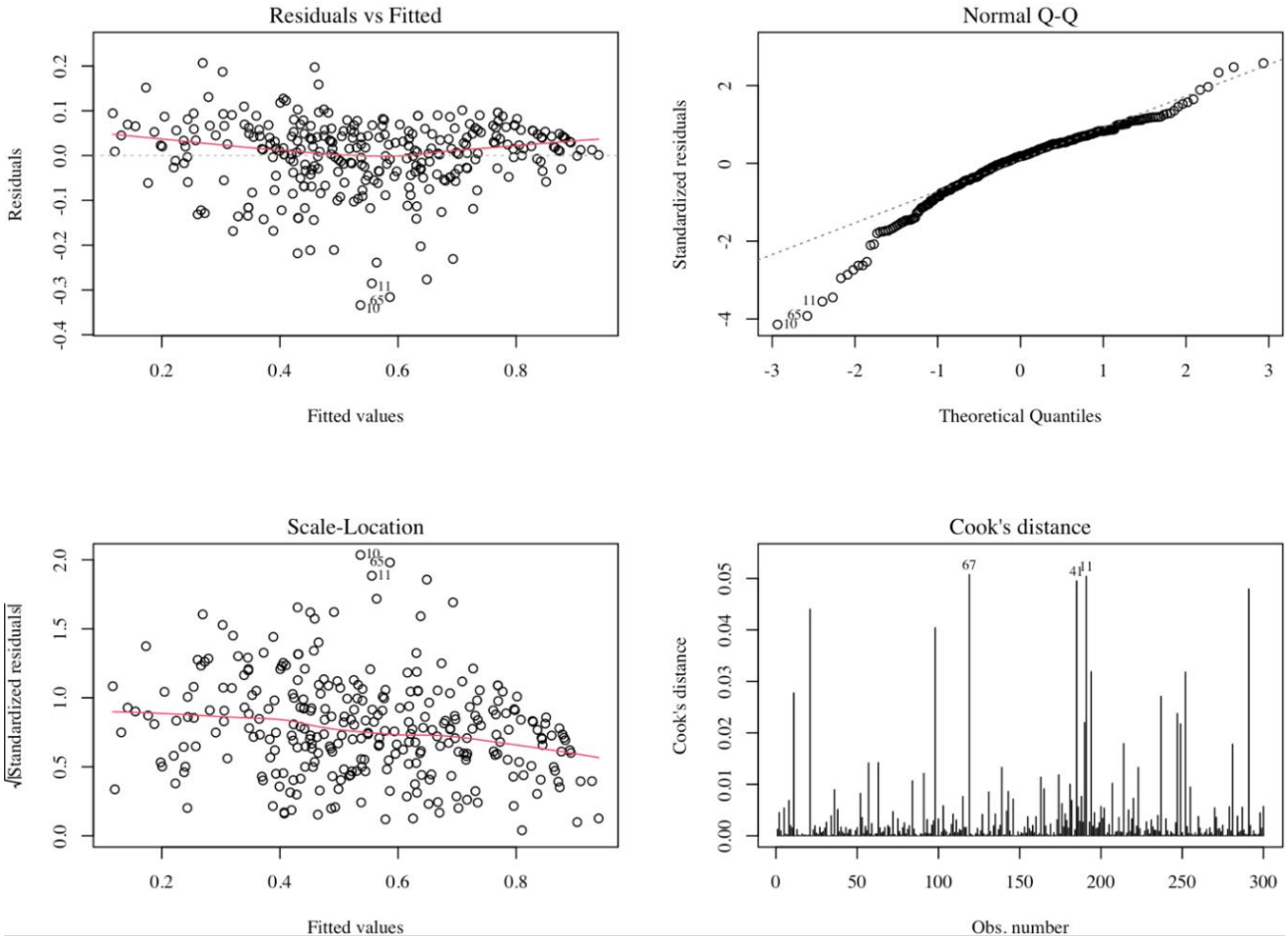

Fig. 4 Residual Plot & Scale-Location & QQ Plot & Cook's Distance for Final Model

The result shows one outlier, 14 leverage points, and 0 influential points in the final model. Since they are not caused by recording mistakes, we only need to report them, but no adjustment is needed. As shown in Table 4, the VIF of each predictor is smaller than five, implying no multicollinearity in the final model.

Table 4. VIF of Each Predictor for Final Model with Train Dataset

GRE Score	TOEFL Score	SOP	CGPA	Research
4.894127	4.484597	2.003245	4.624881	1.485359

The prediction error we got from R is 0.00788, almost equal to 0, indicating that it is a valid model. In addition, as described in the method parts, when we change the dataset from the train part to the test part, the characteristics we got from the test are similar to those in the train. Therefore, we can conclude that this model is for general, not just specialized for training dataset.

3.3. Model Interpretation

The final model has been shown in the result part, and the values in Table 5 help readers understand the model better. For example, this final model tells us that holding other variables constant, when the GRE Score increases by 1, the square of chance of admission will increase by 0.0028 on average. Similarly, holding other variables constant, the difference in the square of chance of admission for students who have research and who do not is 0.0317 on average. The way to interpret other variables is similar to what we described above. Based on Table 5, all predictors are positively related to the response variable, which means the more significant the applicant's effort (achievement), the higher the chance of admission. This final model meets the goal of this report. People can input their score of the determinants into the model to get a relatively precise estimation of the chance of admission.

Table 5. The Summary Table of Final Fitted Model

	Estimate	SD	P-value	If significant
(Intercept)	-2.4558	0.1719	2×10^{-16}	YES
GRE Score	0.0028	0.0009	0.0018	YES
TOEFL Score	0.0046	0.0015	0.0040	YES
Strength of Reference Letter	0.0261	0.0073	0.0004	YES
CGPA	0.1749	0.0165	2×10^{-16}	YES
Research	0.0317	0.0115	0.0062	YES
S = 0.08125		R2 = 0.8436		R2adj = 0.8409

4. Conclusion

In the current era, more students are pursuing higher academic degrees for various reasons, with the master's degree becoming the most sought-after. Against this social backdrop, this article discusses how several essential determinants during the application process can influence the admission chances for graduate programs. By constructing multiple linear models and referring to the indicative value from each model, the author eliminated predictors with weaker influence to enhance the model's predictive power. The final linear model indicates that all included predictors positively correlate with the dependent variable (chance of graduate admission). In other words, the greater the value of each predictor, the higher the chances of graduate admission. It suggests that students wishing to apply for graduate programs with higher scores (or more accomplishments) have a better chance of receiving an offer, which aligns with our initial hypothesis. This article can help students preparing for graduate applications to estimate their chances of admission based on their current performance, and the linear model constructed in the article can also guide students in identifying their weaknesses for targeted improvement.

Based on Figure 4 in the report, even if the author uses Boxcox to transform the response variable better to meet the normality requirement, some points still deviate from the QQ line. Thus, the final model is not perfectly normally distributed. In addition, when the author changes the dataset to testing and do the checking process again, one of the predictors (CGPA) in the testing dataset has a VIF greater than 5. The reason behind it could be overfitting. Since this paper did a Boxcox transformation for both the training and testing dataset, it may cause overfitting, leading to different VIF for predictors in training part and testing part. In future research, we can choose a larger sample size to reduce the consequences of the abovementioned limitations.

References

[1] Li F, John M W, Ding X. The expansion of higher education, employment and over-education in China. International Journal of Educational Development, 2008, 28(6): 687–697.

[2] Mowjee B. Are Postgraduate Students ‘Rational Choosers’? An Investigation of Motivation for Graduate Study Amongst International Students in England. Research in Comparative and International Education, 2013, 8(2): 193–213.

[3] Zimmermann J, von Davier A A, Buhmann J M, Heinimann H R. Validity of GRE General Test scores and TOEFL scores for graduate admission to a technical university in Western Europe. European Journal of Engineering Education, 2018, 43(1): 144–165.

[4] Kuncel N R, Kochevar R J, Ones D S. A Meta-analysis of Letters of Recommendation in College and Graduate Admissions: Reasons for hope. International Journal of Selection and Assessment, 2014, 22(1): 101–107.

[5] Kuncel N R, Credé M, Thomas L L. A Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for Graduate Student Academic Performance. Academy of Management Learning & Education, 2007, 6(1): 51–68.

[6] Esmeraldo G, et al. Using Genetic Programming and Linear Regression for Academic Performance Analysis. Industry and Innovation Tracks, Practitioners' and Doctoral Consortium, 2018, 174–179.

[7] Shawwa L A, et al. Factors potentially influencing academic performance among medical students. Advances in Medical Education and Practice, 2015, 65–75.

[8] Arsad P M, Buniyamin N, Manan J A. Prediction of engineering students' academic performance using Artificial Neural Network and Linear Regression: A comparison. 2013 IEEE 5th Conference on Engineering Education (ICEED), 2013, 43–48.

[9] Froud R, Hansen S H, Ruud H K, Foss J, Ferguson L, Fredriksen P M. Relative performance of machine learning and linear regression in predicting quality of life and academic performance of school children in Norway: Data analysis of a quasi-experimental study. Journal of Medical Internet Research, 2021, 23(7): e22021–e22021.

[10] Huang S, Fang N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers and Education, 2013, 61(1): 133–145.

[11] Acharya M S. Graduate admission 2. Kaggle, 2018.

[12] Mohan S Acharya, Asfia Armaan, Aneeta S Antony. A Comparison of Regression Models for Prediction of Graduate Admissions. IEEE International Conference on Computational Intelligence in Data Science, 2019.