What matters when applying to a graduate school in the United States?

College of Education

Benjamin Kweku Lugu

A data analysis project submitted for BER 540: Statistical Methods in Educational Research

Instructor

Dr. JoonHo Lee

November 22, 2021

Contents

1 Introduction

As the world emerges into a global village spearheaded by the recent technological advancements, people from all walks of life are judiciously investing in education. It comes with no surprise that the United States (US) is ranked first in education since 2020. The quality of education in the US coupled with her numerous opportunities attracts a lot students annually. According to open doors, China and India are two of the world's populous countries that send a large number of students to the US for studies.

To be a successful candidate in the admission recruitment, several factors are considered. First, standardized test scores such as the Graduate Record Examination (GRE) and Test of English as a Foreign Language (TOEFL) are required by most graduate programs. Second, letters of recommendation and statement of purpose provides admission committee information about the applicant's strengths and weaknesses, achievements and their ability to thrive in graduate school. Good undergraduate grade point average (GPA) and research experience are also essential. An applicant with a good GRE score, TOEFL score, high undergraduate GPA, has research experience with a strong recommendation letters and statement of purpose is likely to get admission easily.

In this study, I examined some of the factors mentioned above and their effect on an Indian applicant's ability to gain admission in a US university's graduate school.

1.1 Research question and statement

The current study, however, is guided by the following research question and statement.

1. Design a model for predicting an Indian applicant's chances of getting admission into a graduate school in the US.
2. Does
    a. GRE scores,
    b. TOEFL scores,
    c. University ranking,
    d. Letters of recommendation,
    e. Statement of purpose,
    f. Undergraduate GPA, and
    g. Research experience, influence an Indian applicant's chances of getting admission into a graduate school in the US?

2 Data

This data was created to predict graduate admission of Indian students into an American university. The data contain several influential variables necessary to trigger admission for master's programs. Table 1 summarizes the data characteristics. It has seven continuous and one discrete variable. Admission is measured as a probability with higher values indicating the posibility of getting an admission. Other variables such as GRE score or research experience (1 = applicant has research experience and 0 otherwise) and whether university ranking is important (1 = high ranking to 5 = low ranking). Because the data has been preclean, I proceeded with the analysis. I pulled this data set from Kaggle (www.kaggle.com), a data science competition website.

Table 1: Characteristics of the data

Name	Type	Scale	Coded.as
GRE score	Continuous	0 to 340	gre
TOEFL score	Continuous	0 to 120	toefl
University ranking	Continuous	1 to 5	uranking
Statement of purpose	Continuous	0 to 5	sop
Letters of recommendation	Continuous	0 to 5	lor
Undergraduate GPA	Continuous	0 to 10	cgpa
Research experience	Discrete	0 or 1	research
Chance of admission	Continuous	0 to 1	Admit_chance

3 Data Exploration

Data exploration provides important information about the data. But first, I loaded the data into R and preview the responses of the first six applicants (see Table 2). For example, the first applicant chose a low ranking university despite having high GRE and TOEFL score of 337 and 118 respectively, an undergraduate GPA of 9.65 (out of 10), has research experience and statement of purpose and letters of recommendation rated 4.5 (out of 5) each. From the result, this applicant has 92% chance of getting admission for a master's program in the US.

```
data <- read.csv('data.csv')
kbl(head(data), caption = "Sample data preview\\label{tab:tab2}", booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 2: Sample data preview

Serial	gre	toefl	uranking	sop	lor	cgpa	research	admit_chance
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4.0	4.5	8.87	1	0.76
3	316	104	3	3.0	3.5	8.00	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.80
5	314	103	2	2.0	3.0	8.21	0	0.65
6	330	115	5	4.5	3.0	9.34	1	0.90

Because the purpose of this study is to build a regression model for prediction and examine the factors that influence admission chances, I splitted the data into two: `train` and `test` data. The `train` data will be used to build and train the regression model, while the `test` data will validate the model's prediction accuracy. So, 85% of the data will be used for the training. This represents 425 applicants. The remaining 75 applicants' responses will be used for prediction on the `test` data.

```
# Sample size determination
nrow(data)
```

```
## [1] 500
```

```
# Splitting data into test and train
# 85% of the sample size for training
```

```
sample_size <- floor(.85 * nrow(data))
sample_size
```

```
## [1] 425
```

```
set.seed(558)  # make data reproducible
# generate random sample without replacement
train_ind <- sample(seq_len(nrow(data)), size = sample_size)
# generate data of responses according to the sample
train <- data[train_ind, ]
# generates the remaining 20% of the sample
test <- data[-train_ind, ]
```

Table 3 represents a preview of the training data set.

```
kbl(head(train), caption = "Sample of the training data\\label{tab:tab3}", booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 3: Sample of the training data

	Serial	gre	toefl	uranking	sop	lor	cgpa	research	admit_chance
426	426	323	111	5	4.0	5.0	9.86	1	0.92
193	193	322	114	5	4.5	4.0	8.94	1	0.86
13	13	328	112	4	4.0	4.5	9.10	1	0.78
461	461	319	105	4	4.0	4.5	8.66	1	0.77
109	109	331	116	5	5.0	5.0	9.38	1	0.93
334	334	319	108	3	3.0	3.5	8.54	1	0.71

Table 4 represents a preview of the testing data set.

```
kbl(head(test), caption = "Sample of the test data\\label{tab:tab4}", booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 4: Sample of the test data

	Serial	gre	toefl	uranking	sop	lor	cgpa	research	admit_chance
6	6	330	115	5	4.5	3.0	9.34	1	0.90
8	8	308	101	2	3.0	4.0	7.90	0	0.68
10	10	323	108	3	3.5	3.0	8.60	0	0.45
20	20	303	102	3	3.5	3.0	8.50	0	0.62
23	23	328	116	5	5.0	5.0	9.50	1	0.94
24	24	334	119	5	5.0	4.5	9.70	1	0.95

For the remaining part of the study, the training data will be used for further exploration and analysis.

3.1 Data description

The `describe` function from the `psych` package provides a summary information such as the sample size (n), mean, standard deviation (sd) among others. For example, in Table 5 the best applicant(s) has(have)

Table 5: Descriptive statistics of applicants

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Serial	1	425	256.6164706	143.8716711	252.00	257.3724340	183.842400	1.00	500.00	499.00	-0.0186561	-1.2062846	6.9788011
gre	2	425	316.6235294	11.2865512	317.00	316.6422287	11.860800	290.00	340.00	50.00	-0.0603292	-0.6854525	0.5474781
toefl	3	425	107.2776471	6.0014223	107.00	107.1906158	5.930400	92.00	120.00	28.00	0.0573542	-0.6187102	0.2911117
uranking	4	425	3.1364706	1.1265291	3.00	3.1319648	1.482600	1.00	5.00	4.00	0.0273562	-0.7608423	0.0546447
sop	5	425	3.3894118	0.9649547	3.50	3.4120235	0.741300	1.00	5.00	4.00	-0.2204804	-0.7102754	0.0468072
lor	6	425	3.5000000	0.9354143	3.50	3.5190616	0.741300	1.00	5.00	4.00	-0.1962012	-0.7442497	0.0453743
cgpa	7	425	8.5805412	0.6017517	8.60	8.5832845	0.681996	6.80	9.92	3.12	-0.0488100	-0.5465509	0.0291892
research	8	425	0.5741176	0.4950588	1.00	0.5923754	0.000000	0.00	1.00	1.00	-0.2987251	-1.9152532	0.0240139
admit_chance	9	425	0.7230118	0.1402907	0.73	0.7283284	0.148260	0.34	0.97	0.63	-0.2985998	-0.4806713	0.0068051

97% of getting admission and the average admission rate is 72%. Also, the average test scores obtained by these applicants is approximately 317 and 107 for GRE and TOEFL respectively. Similarly, the statement of purpose and letters of recommendation are above average with an undergraduate GPA exceeding 8.0 (on a scale of 10.0). Thus, to a large extent, the applicants possesses good qualities for admission.

```
des_stats <- psych::describe(train)
kbl(des_stats, caption = "Descriptive statistics of applicants\\label{tab:tab5}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "scale_down"))
```

Table 6 provides further information on the number and percentage of applicants who prioritize university ranking. About 34% of the applicants prefer averagely ranked university. Meanwhile, about twice the number of applicants who chose highly ranked universities prefer the least ranked ones to increase their admission intake.

```
urank <- count(train, uranking)
ud <- data.frame(
  University_ranking = urank$uranking,
  Number_of_applicants = urank$n,
  Percentage_of_applicants = round(urank$n/sum(urank$n)*100,2)
)
kbl(ud, caption = "Aplicants' choice of university ranking\\label{tab:tab6}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 6: Aplicants' choice of university ranking

University_ranking	Number_of_applicants	Percentage_of_applicants
1	29	6.82
2	98	23.06
3	143	33.65
4	96	22.59
5	59	13.88

In as much as university rankng is crucial, research experience can be pivotal and may give an applicant a competitive edge. It can be seen that more than half of the applicants have research experience (see Table 7).

```
res_exp <- count(train, research)
rexp <- data.frame(
  Research_experience = res_exp$research,
  Number_of_applicants = res_exp$n,
  Percentage_of_applicants = round(res_exp$n/sum(res_exp$n)*100,2)
)
kbl(rexp,
    caption = "Distribution of Aplicants by research experience\\label{tab:tab7}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 7: Distribution of Aplicants by research experience

Research_experience	Number_of_applicants	Percentage_of_applicants
0	181	42.59
1	244	57.41

Again, I explored the relationship between research experience and university ranking. The crosstabulation provides insightful information.

```
suv <- data.frame(train$research, train$uranking)
kbl(table(suv),
  caption = "Crosstabulation of unviersity ranking and research experience\\label{tab:tab8}",
  booktabs = T) %>%
  kable_styling() %>%
  pack_rows("Research Experience", 1,2) %>%
  add_indent(c(1,2), level_of_indent = 12) %>%
  add_header_above(c(" " = 1, "University Ranking" = 5))
```

It is surprising that there is an excess of 45 applicants with research experience who applied to a low ranked university. In contrast, 42 more applicants with no research experience chose the second most rated universities. Several reasons may account for this decisions ranging from program of choice, application documents to funding.

```
# Graphical representation
ggplot(train, aes(uranking)) + geom_histogram(bins = 10, binwidth = 0.5) +
  xlab("University ranking") + ylab("Number of students") +
```

Table 8: Crosstabulation of unviersity ranking and research experience

		University Ranking				
		1	2	3	4	5
Research Experience						
	0	19	70	63	22	7
	1	10	28	80	74	52

```
  facet_wrap(~research)
```

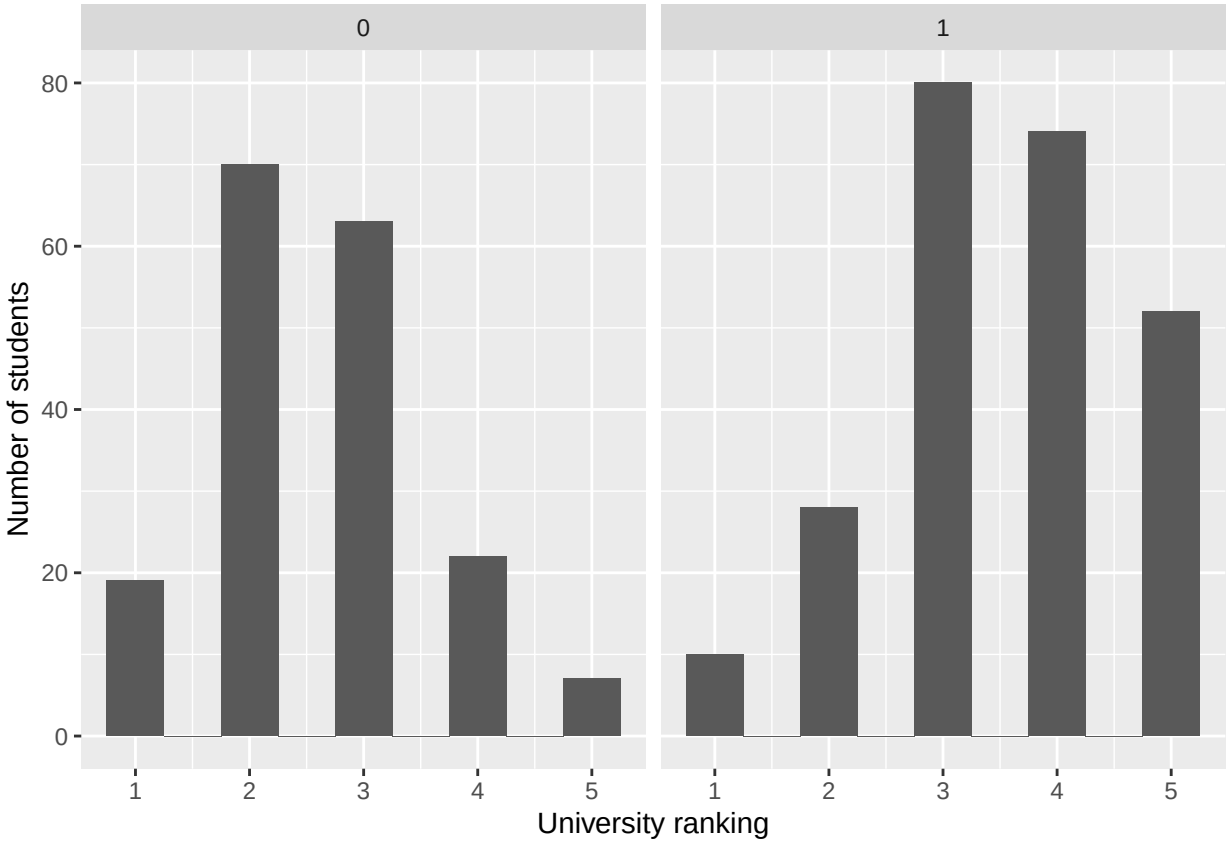


Figure 1: University ranking by level of research experience

3.2 Plots and Correlation analysis

Correlation analysis shows the strength of the relationship between variables. It ranges from -1 to 1 and correlation coefficients closer to -1 or 1 are signs of strong correlation. For this analysis, I am interested in exploring the correlation between admission chance and all the other variables. From Table 9, there exist a significant, positive and moderate to strong correlation coefficients ranging from 0.548 to 0.890. The highest of this is the undergraduate GPA, followed by GRE score and TOEFL score.

```
toyota <- rcorrst(train[, 2:9])
# correlation with p values
kbl(toyota,
    caption = "Correlation with p values\\label{tab:tab27}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position")) %>%
    footnote(general = "**** p < .0001",
    footnote_as_chunk = T
)
```

Table 9: Correlation with p values

	gre	toefl	uranking	sop	lor	cgpa	research
gre							
toefl	0.82****						
uranking	0.62****	0.64****					
sop	0.61****	0.63****	0.71****				
lor	0.53****	0.54****	0.60****	0.66****			
cgpa	0.83****	0.81****	0.70****	0.70****	0.63****		
research	0.57****	0.45****	0.41****	0.39****	0.37****	0.50****	
admit_chance	0.81****	0.79****	0.69****	0.69****	0.65****	0.89****	0.55****

Note: **** p < .0001

4 Building a regression model

In this section, I built a multiple regression model for predicting admission chance. Admission chance is the dependent variable while the other variables represent the independent variables. Equation represents the model.

$$y = \beta_0 + \sum_{i=1}^{7} \beta_i x_i + \epsilon$$

where y is the admission chance, β_0 is the intercept,β_i are the independent variables and ϵ represents the error term.

The R code below execute the model. Before viewing the results, I examined the assumptions of linear regression.

```
reg_model <- lm(admit_chance ~ gre + toefl + uranking + sop + lor + cgpa + research,
                data = train)
```

4.1 Regression diagnostics

It is very important to examine the regression diagnostic and address possible problems before making decisions. The following assumptions are examined. I used the `autoplot` function in the `ggfortify` package to generate the plots. For multicollinearity, I extracted the variance inflation factor function `(vif)` from the `car` package.

4.1.1 Linearity

From the residual vs fitted plot below, there is no distinct pattern. In other words, the data points are randomly and evenly dispersed about the reference line, this means the relationship is linear.

4.1.2 Normal Q-Q plot

The plot shows that the normality assumption has been met as the data points do not deviate extremely from the normal probability line. Also, the values of the skewness and kurtosis for the variables fall within the interval ±2, indicating the non-violation of normality assumption (see Table 5).

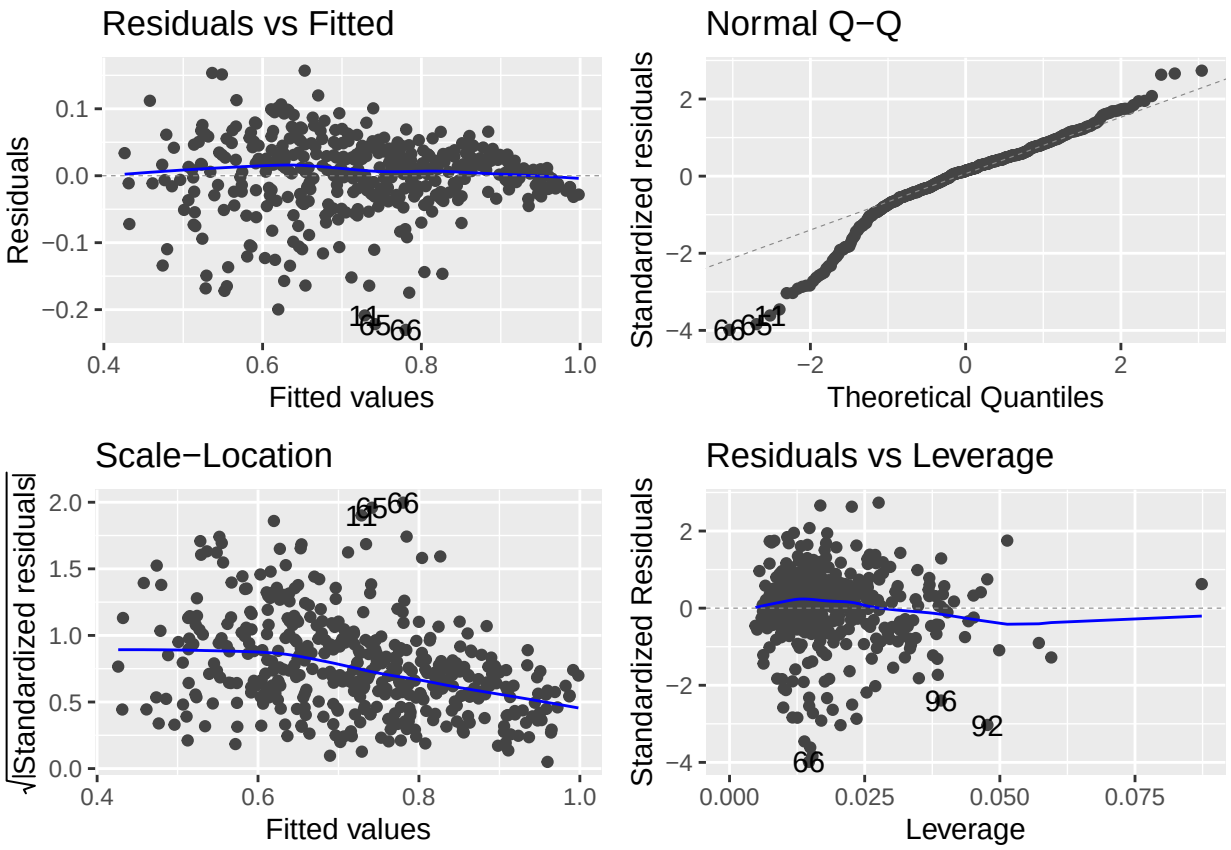
4.1.3 Homoscedasticity

The scale-location plot indicate spread of the data and it is used to check the homogeneity of variance of the residuals. The residuals are well spread but decreases slightly along the fitted values. Thus, homoscedasticity is satisfied.

4.1.4 Outliers

The residuals vs leverage provide a good information on the influential variables in the regression result. The absence of cook distance line on the plot is situation where outliers are not present in the study.

```
autoplot(reg_model)
```



4.1.5 Multicolinearity

This can be done by using the vif in the `car package`. I computed the variance inflation factor (VIF) and tolerance (1/VIF) to check if there is a high correlation between the independent variables. A rule of thumb for interpreting the variance inflation factor is:

- 1 = not correlated
- Between 1 and 5 = moderately correlated and
- Greater than 5 = highly correlated

The results show a moderate correlation between the independent variables (see Table 10).

```
vit <- data.frame(
  VIF = round(car::vif(reg_model),2),
  Tolerance = round(1/car::vif(reg_model),2)
)
kbl(vit, caption = "VIF and Tolerance values\\label{tab:tab10}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 10: VIF and Tolerance values

	VIF	Tolerance
gre	4.64	0.22
toefl	3.88	0.26
uranking	2.52	0.40
sop	2.68	0.37
lor	2.03	0.49
cgpa	4.91	0.20
research	1.52	0.66

From the above discourse, it is evident that the regression assumptions have been satisfied.

Now, the regression model is represented below.

```
summary(reg_model)
```

```
##
## Call:
## lm(formula = admit_chance ~ gre + toefl + uranking + sop + lor +
##     cgpa + research, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.230281 -0.024359  0.008006  0.032318  0.156955
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2001788  0.1111348 -10.799  < 2e-16 ***
## gre          0.0015605  0.0005391   2.895 0.003995 **
## toefl        0.0024503  0.0009273   2.642 0.008541 **
## uranking     0.0055134  0.0039833   1.384 0.167056
## sop          0.0069735  0.0047958   1.454 0.146678
## lor          0.0139331  0.0043040   3.237 0.001303 **
## cgpa         0.1237272  0.0104022  11.894  < 2e-16 ***
## research     0.0259628  0.0070336   3.691 0.000253 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.05818 on 417 degrees of freedom
## Multiple R-squared:  0.8309, Adjusted R-squared:  0.828
## F-statistic: 292.6 on 7 and 417 DF,  p-value: < 2.2e-16
```

Though the regression output produces substantive information about the R-squared and adjusted R-squared, I also examined the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) to ensure the data adequately fit the model.

```
# Model performance
mod <- data.frame(
  RMSE = round(rmse(reg_model, data = train),4),
  MAE = round(mae(reg_model, data = train),4)
)
kbl(mod, caption = "Model performance\\label{tab:tab11}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 11: Model performance

RMSE	MAE
0.0576	0.0419

4.2 Prediction

As stated in the introduction, the purpose of this study is to predict admission chance of a prospective applicant into a university in the US. At this point, I am going to use the `test` data for the prediction. The first six responses of the data is shown in Table 12.

```
kbl(head(test), caption = "Sample of test data\\label{tab:tab13}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 12: Sample of test data

	Serial	gre	toefl	uranking	sop	lor	cgpa	research	admit_chance
6	6	330	115	5	4.5	3.0	9.34	1	0.90
8	8	308	101	2	3.0	4.0	7.90	0	0.68
10	10	323	108	3	3.5	3.0	8.60	0	0.45
20	20	303	102	3	3.5	3.0	8.50	0	0.62
23	23	328	116	5	5.0	5.0	9.50	1	0.94
24	24	334	119	5	5.0	4.5	9.70	1	0.95

Table 13 and 14 provide information on the performance metrics for the predictive model.

```
prediction <- data.frame(predicted_admit_chance = predict(reg_model, test),
          admit_chance  = test$admit_chance
```

```
          )
pred <- data.frame(
  RMSE = RMSE(prediction$predicted_admit_chance, prediction$admit_chance),
  MAE = MAE(prediction$predicted_admit_chance, prediction$admit_chance),
  R_Square = R2(prediction$predicted_admit_chance, prediction$admit_chance)
)
kbl(round(pred,4), caption = "Model performance of predicted model\\label{tab:tab15}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 13: Model performance of predicted model

RMSE	MAE	R_Square
0.07	0.048	0.7704

```
# predictive accuracy
Correlat <-  rcorrst(prediction)
kbl(Correlat, caption = "Predictive Accuracy\\label{tab:tab16}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position")) %>%
    footnote(general = "**** p < .0001",
    footnote_as_chunk = T)
```

Table 14: Predictive Accuracy

	predicted_admit_chance
predicted_admit_chance	
admit_chance	0.88****

Note: **** p < .0001

Now, the predicted and actual rate of admission chance is shown in Table 15.

```
kbl(head(prediction), caption = "Model performance of predicted model\\label{tab:tab17}",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 15: Model performance of predicted model

	predicted_admit_chance	admit_chance
6	0.8788829	0.90
8	0.5930508	0.68
10	0.7152861	0.45
20	0.6570022	0.62
23	0.9293617	0.94
24	0.9638543	0.95

5 Discussions and limitations

The findings obtained from the analysis are intriguing. First, the model account for 81.5% of the variability in applicants' chance of getting admission into a US university for a graduate program. From the outcome of RMSE and MAE, the model resulted in respectively 5.8% and 4.2% error values. These are good values which validates the performance of the study model. Second, I observed that, among the independent variables, university ranking ($\beta = .0060$, $p > 0.05$) and statement of purpose ($\beta = .0061$, $p > 0.05$) have no significant effect on an applicant's admission chance. In other words, the university ranking and SOP does not influence an individual's chances of getting admission. However, GRE scores ($\beta = .0016$, $p < 0.01$), TOEFL score ($\beta = .0028$, $p < 0.01$), letters of recommendation ($\beta = .0141$, $p < 0.01$), undergraduate GPA ($\beta = .120$, $p < 0.001$) and research experience ($\beta = .0274$, $p < 0.001$) have positive and significant impact on an applicant's admission chance. This means applicants from India must prioritize these influential variables when considering graduate studies in the United States.

The model is also designed to predict the admission chances of applicants when it is fed with required information. For a group of 75 applicants, the model has an accuracy rate of 87.9% and accounted for 77.2% of the variability in the admission chance with low error rate.

Inspite of the relevant findings obtained, the study is not without limitations. First, insufficient information was provided on the scale used for rating statement of purpose, letters of recommendation, admission chance etc. Data was not collected on whether admissions into these university were funded or not. Though challenging, it is quite easier for an applicant with good grades to get admission into a US graduate school without funding. So, the findings of this study must be interpreted with caution.

6 Conclusion

I conclude by entreating Indians who desire to study in the US to have a very high CGPA in their undergraduate program, good research experience and they should never forget to rely on Professors who can provide efficient and convincing recommendation letters. A good GRE and/or TOEFL score will enhance their chances of gaining admission into a graduate school in the US.

Reference

Mohan S. Acharya, Asfia Armaan & Aneeta S. Antony. 2019. *A Comparison of Regression Models for Prediction of Graduate Admissions. IEEE International Conference on Computational Intelligence in Data Science.* https://www.kaggle.com/mohansacharya/graduate-admissions.