PROJECT REPORT

Regression Analysis on Graduate Admissions Data

STAT 6021: Linear Models for Data Science
Data Science Institute, University of Virginia

Charishma Ravoori	cr2st@virginia.edu
Charu Rawat	cr4zy@virginia.edu
Saurav Sengupta	ss4yd@virginia.edu
Sri Vaishnavi Vemulapalli	sv2fr@virginia.edu
Vaibhav Sharma	vs3br@virginia.edu

6 December, 2018

CONTENTS

1. Executive Summary

Each year, over 3 million college applications are filed in the US by about 750,000 students [1], an average of 4 applications per student. Each of them comes with a certain element of randomness or chance. The intended meritocracy inherent in college admissions gives way to uncertainty, doubt, and anxiety, even for students with exceptional credentials. Not all colleges are transparent about their admission processes and so it becomes tough for a student applicant to gauge whether he or she can get admission into an institution.

In this project, we demonstrate how regression analysis can be used on a sample dataset to ease this process. Universities can use similar analysis on their data to help with the admissions process, and students can use this analysis to determine how likely they are to get an admit given the strength of their profile.

2. Introduction

2.1. Problem Statement

There are many factors that influence admission decisions. And while colleges rely on more than quantitative data to make admissions decisions, quantitative data can show us in a concrete way many things that qualitative data cannot. Even though it's tough to understand and estimate how these factors are truly judged and filtered by colleges, we do know that some of the factors such as CGPA and GRE scores can weigh heavily on determining acceptance. Metrics such as these scores can be leveraged in form of data and be analyzed to gain insight into admission trends and can help students in shortlisting universities with their profiles saving time, effort and money that goes into the exhaustive application process. The predicted output can also give them a fair idea about their chances for admission to a particular university. The scope of such analysis can also be extended to help college institutions answer questions such as – "Do we know that standardized tests are a valid predictor of success in admission at our institution?"

2.2. Objective

In our analysis detailed in this report, we have adopted a data-driven approach towards quantifying the probability of successful admission or enrollment into college institutions dependent solely upon certain quantitative factors. Our objective with this analysis is two-fold. The first objective is to understand what factors are significant and relevant in determining enrollment and to what degree. In doing so, we draw inferences about relationships between the factors (variables) and identify any dependency that exists between them. We also seek to evaluate and prove certain hypotheses such as –

1. Is a student's CGPA a significant factor in determining his probability of enrollment?
2. Is a student's SOP rating significant in estimating his probability of enrollment given other strong factors?
3. Does university ranking play a significant role in determining enrollment success?

Evaluating these hypotheses will help us gain a better understanding of the admissions criteria. Our second goal is to utilize statistical methods to build a model that can predict the probability of enrollment success of an applicant given certain attributes. Such analysis can be immensely helpful to students as it can help them strategize their university shortlisting process.

3. Data

The data [2] was sourced from kaggle.com and is owned by user Mohan S Acharya on the Kaggle website. Originally, this data was extracted from the applicant's database of UCLA. Information was collected from students regarding what universities they submitted applications to and their respective CGPA, GRE, and TOEFL scores. This dataset consists of multiple quantitative measures of a student's performance GRE and TOEFL scores and is crucial in determining admissions into institutions. In addition to those measures, UCLA has provided ratings of the SOP and LOR which even though subjective have been rated by admissions officers who are experts in this area and have evaluated such documents for years. Additionally, there is an attribute provided by the university named *Chance.of.Admit* which gives an idea about the admission probability of an applicant. These features combined together and put in context, make the data ideal for us to study and analyze patterns that are relevant to achieving our objective.

The raw dataset consists of 400 records with 9 attributes where each row contains application information to a university. The dataset contains several parameters which are considered important during the application for graduate programs. The parameters included are:

1. Serial Number
2. GRE Score (out of 340)
3. TOEFL Score (out of 120)
4. University Ranking (out of 5)
5. Statement of Purpose Strength (Rating out of 5)
6. Letter of Recommendation Strength (Rating out of 5)
7. Undergraduate GPA (out of 10)
8. Research Experience (either 0 or 1) -
9. Chance of Admit (a probability ranging from 0 to 1)

3.1. Preparation and Exploration

The raw data was sourced from kaggle.com upon which pre-processing was done to obtain data in a desirable format to perform analysis upon. The sample of the raw data can be seen in Appendix-1.

We used the raw CSV file – "Admission_Predict.csv" provided on the Kaggle website.

The file was loaded in R as a data frame upon which the following preprocessing was performed to obtain the final format of the data.-

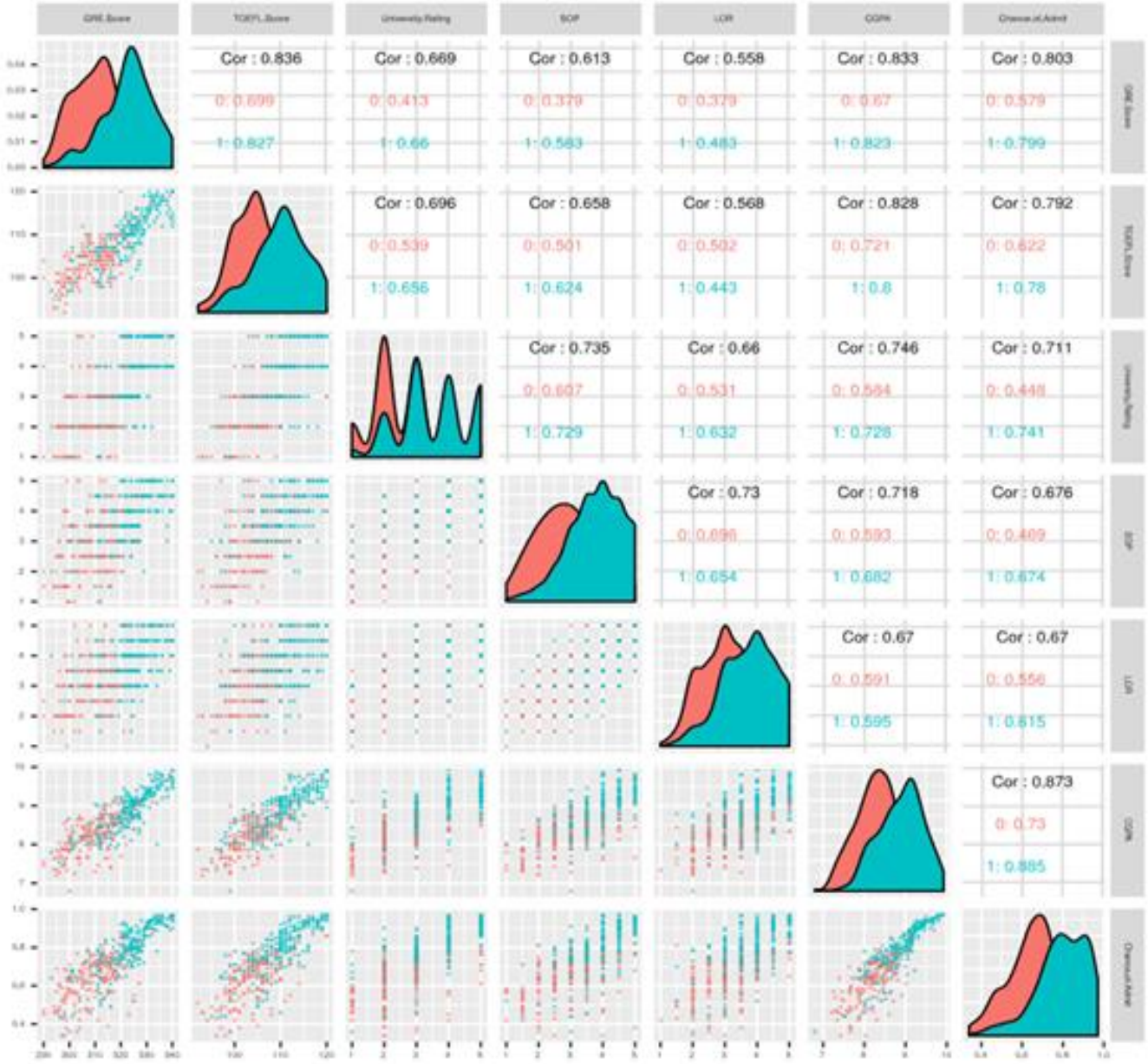
1. Dropping variables - We checked for duplicate rows in the data and found none. We dropped the variable *"Serial.No"* from the data since it only served as a primary key to the data and hence had no effect on *"Chance.of.Admit"* which is our response variable.
2. Missing value imputation - We checked our data for missing values but found none, hence no missing value imputation was required.
3. Variable type conversion - Variable "Research" in the data with values "0" or "1" was converted into a factor with 2 levels. All other variables remain numeric in type.

3.2 Summary Statistics

Summary statistics are presented in the table below. For all the attributes the mean and the median are quite close to each other in value.

GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
Min. :290.0	Min. : 92.0	Min. :1.000	Min. :1.0	Min. :1.000	Min. :6.800	0:181	Min. :0.3400
1st Qu.:308.0	1st Qu.:103.0	1st Qu.:2.000	1st Qu.:2.5	1st Qu.:3.000	1st Qu.:8.170	1:219	1st Qu.:0.6400
Median :317.0	Median :107.0	Median :3.000	Median :3.5	Median :3.500	Median :8.610	NA	Median :0.7300
Mean :316.8	Mean :107.4	Mean :3.087	Mean :3.4	Mean :3.453	Mean :8.599	NA	Mean :0.7244
3rd Qu.:325.0	3rd Qu.:112.0	3rd Qu.:4.000	3rd Qu.:4.0	3rd Qu.:4.000	3rd Qu.:9.062	NA	3rd Qu.:0.8300
Max. :340.0	Max. :120.0	Max. :5.000	Max. :5.0	Max. :5.000	Max. :9.920	NA	Max. :0.9700

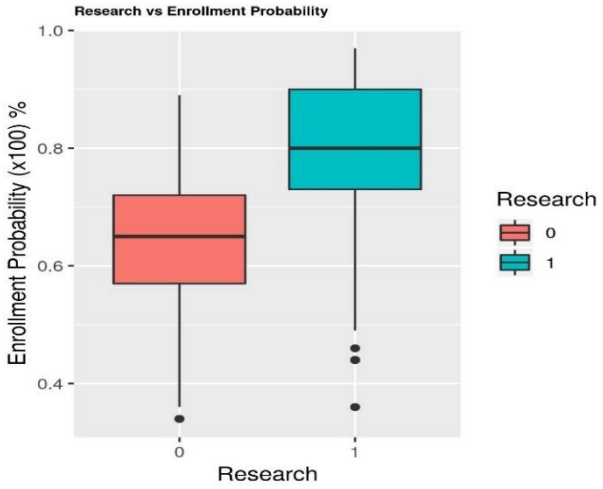
The pairwise plots between regressors and response are shown below. The variables are further grouped on the category "*Research*". It's observable that most of the variables in the data have an above 70% correlation with the response variable which is "*Chance.of.Admit*" or Enrollment probability. Though a caveat here is that there seems to be high multicollinearity in the data.



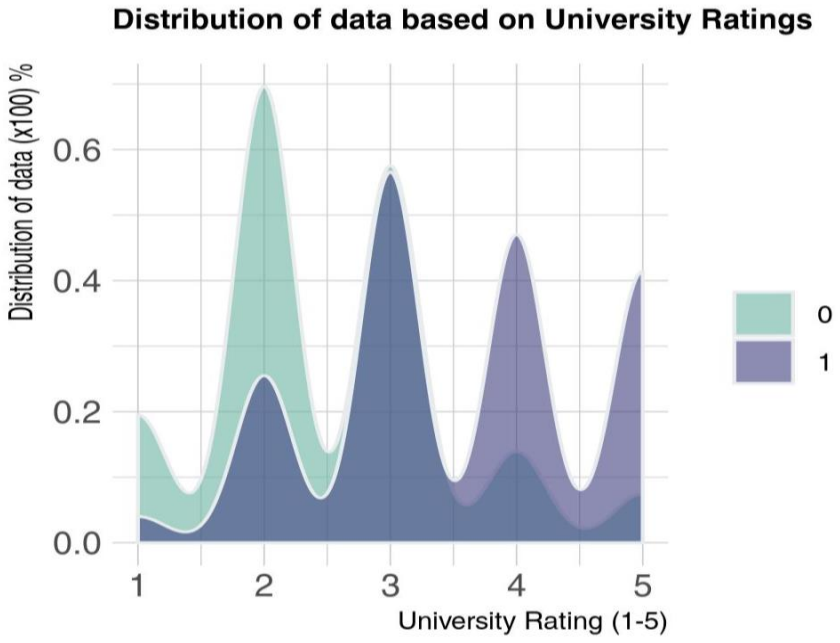
3.3 Exploratory Data Analysis

We were able to make certain observations about the data through exploratory data analysis. Some of our observations are summarized below –

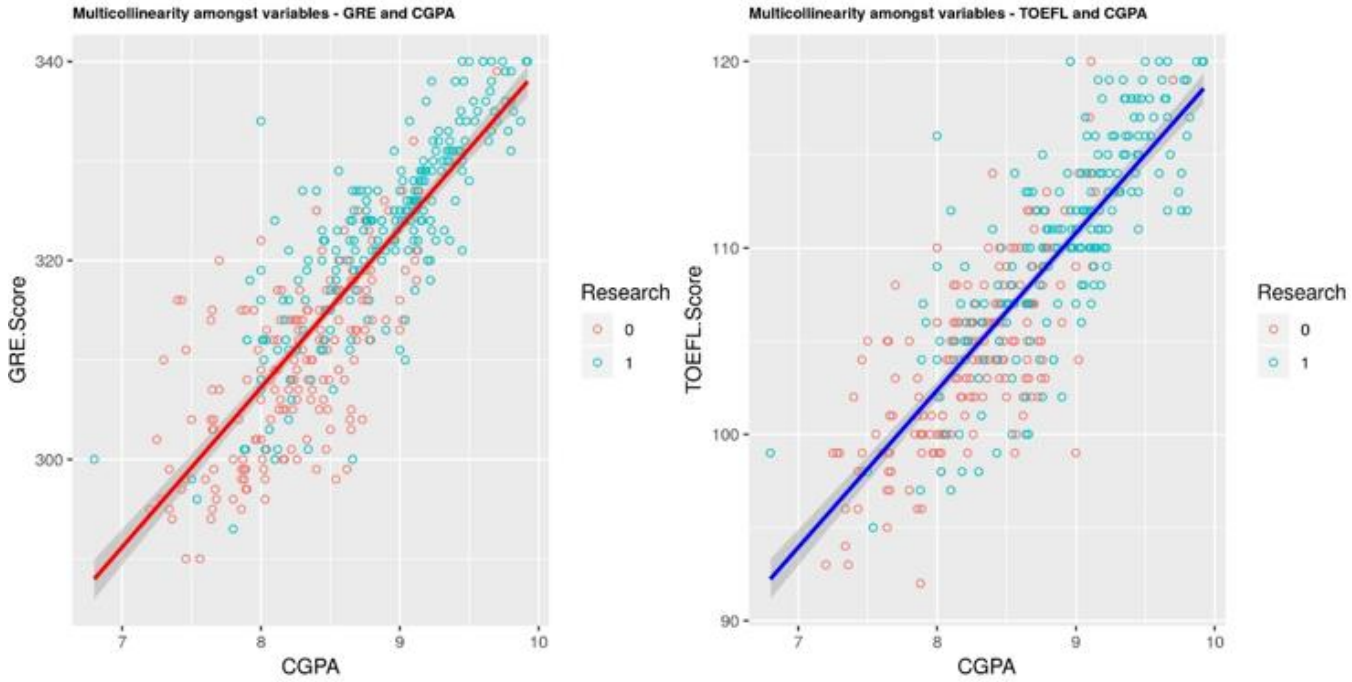
1. Students who have done research work in the past (Research = 1) have a higher chance of admission (~80% chance) on average as compared to students with no prior research experience (Research = 0) who have just a 60% chance of admission. However, there are certain exceptions in form of outliers to this trend that are visible in the boxplot.



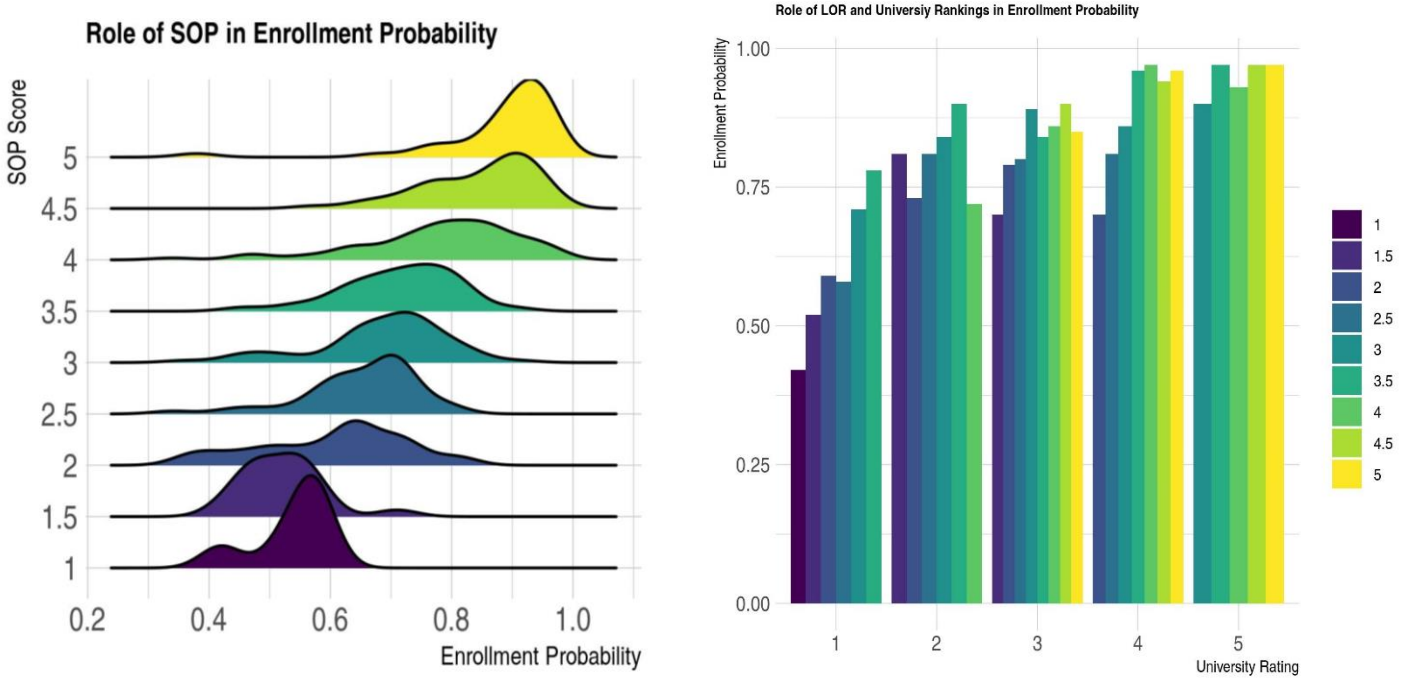
2. From the plot below, we can observe that in this data, students with no research experience typically apply to universities rated lower (rated 1 or 2) whereas the majority of the students with prior research experience apply to universities rated higher (4 or 5). There is a significant overlap in students with or without research experience that apply to mid-tier universities (rated 3).



3. The data does exhibit multicollinearity in form of linear relationships between certain variables. An example of that is visible in the plots below where we can see that both GRE and TOEFL score have a linear relationship with a student's CGPA. We addressed this issue further in our analysis detailed in this report.

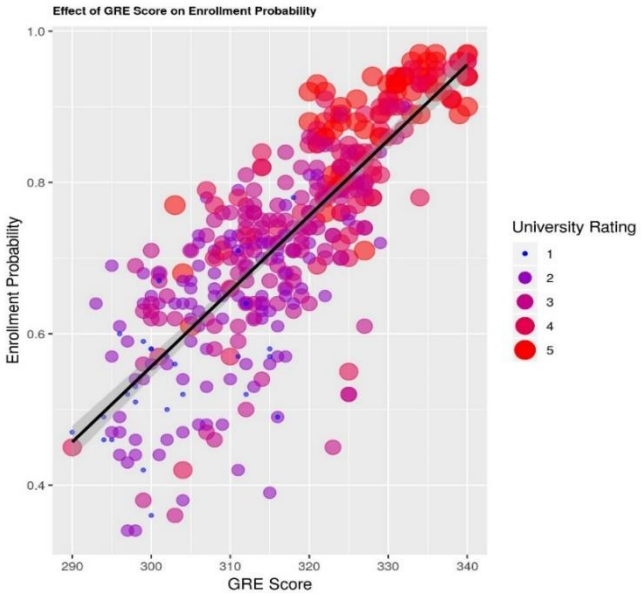


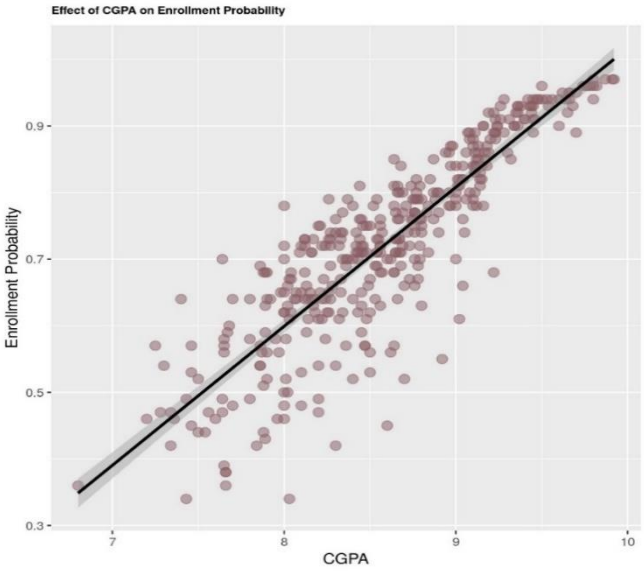
4. From the plot on the left below, we can see that the higher the SOP rating, the greater the chance of admission success. Typically, SOP's rated between 3 and 4.5 have a wider range of enrollment probability - roughly between 60 and 90% on enrollment success. The plot below on the right explore the effect that LOR and university ratings have on enrollment success. We observed that for applications submitted to universities rated lower (rated 1 or 2), there is more variation in the enrollment probabilities dependent upon LOR scores. Whereas for universities rated higher (rated 4 or 5), a LOR that's rated 3, 4 or 5 has roughly the same enrollment probability.



5. A student's GRE and TOEFL has a linear relationship with his or her chance of admission or enrollment success. The higher the score, the higher the enrollment probability which makes intuitive sense and is in accordance with our expectations.

   We can also observe from the plot below that applicants with a higher score tend to mostly apply to universities rated higher as we can see the larger sized points cluster towards the top end.





6. Lastly, we looked at the relationship between CGPA and enrollment probability. We could see a strong linear trend amongst these variables as well. Though an interesting observation here is that all applicants with CGPA > 9 have a success rate of 90% but when the CGPA decreases, we can see there are more data points that are not in complete keeping with the linear trend i.e. have a good CGPA but enrollment probability remains lower than expected. This could be due to other factors such as GRE Score, LOR rating etc.

4. Data Analysis and Statistical Inference

4.1 Methods in Model Building

From the exploratory data analysis, we could conclude that the variable "Chance of Admit" has a linear relationship with various other variables like CGPA, GRE Score, SOP rating, etc. Keeping this in mind, we decided to implement a multiple linear regression on the data as a starting point for our analysis, with response variable "*Chance.of.Admit*" and the other variables as regressors.

Formally, the fitted multiple linear regression equation for n observations is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ \ldots + \beta_n x_{in} \ for \ i = 1,2, \ \ldots \ n$$

where y_i is the i^{th} response variable and x_{ij} is the j^{th} regressor variable for the i^{th} observation.

4.2 Variable Selection

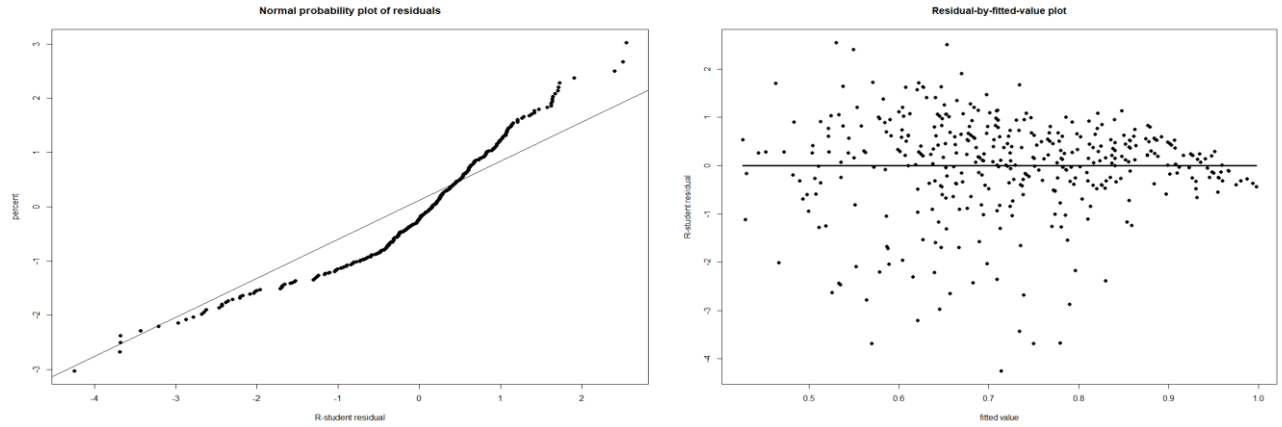
To select the best variables that would result in a good model fit to predict the *Chance.of.Admit*, we applied variable selection methods.

We fit multiple models using various combinations of the regressors. Then compared these models based on certain statistical measures like Mallow's stat (Cp), R-square, and BIC. The best model had the 5 variables- *CGPA, GRE.Score, LOR, Research,* and *TOEFL.Score.* It had a high R-square of ~80% and the lowest Cp. This model selection also matched what we got from using stepwise selection method which gives an ideal model selecting variables based on AIC for each combination of variables.

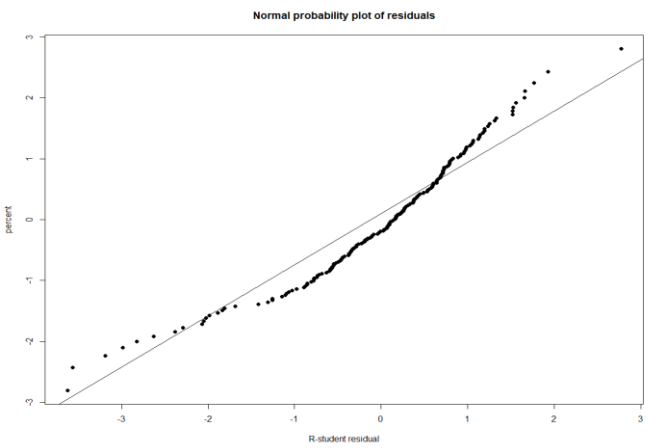
4.3 Model Adequacy

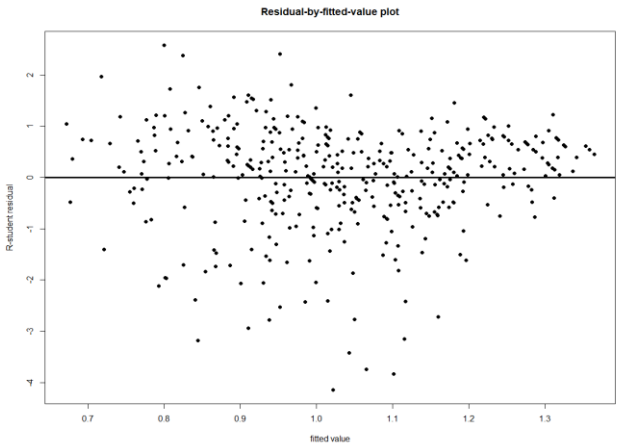
One of the basic assumptions of linear regression is that the variance of the errors is constant. When we checked our model for adequacy, we noticed that this assumption was violated. Upon further investigation, we decided to select a model based on a transformation of the *Chance.of.Admit* variable, especially since the range of this variable is between 0 and 1. The transformation of the variable stabilized the variances and gave us a satisfactory final model.

To determine if the normality assumption of the error variance held up, we plotted the Normal Probability Plot illustrated below. The points more or less lie on a straight line; however, we do see heavy tails that suggests there could potentially be a violation of the normality assumption. We investigated this further by examining more residual plots. The plot which has the residuals plotted against the fitted values shows a double bow pattern, which is expected since our outcome variable is a probability between 0 and 1.



In order to correct this, an arcsin transformation was applied to the square root of the response variable. After this transformation, we examine the residual plots and notice an improvement from the previous plots.



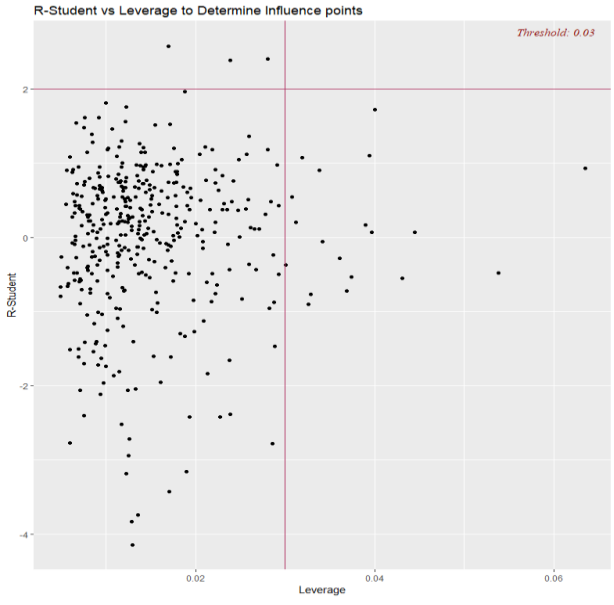


4.4 Diagnostics for Leverage and Influence

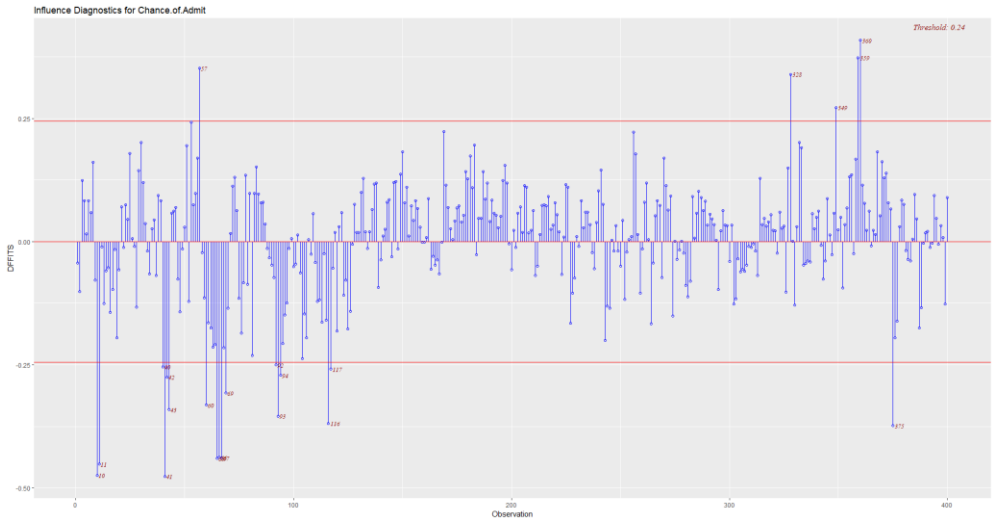
In order to determine if there are any leverage and influential points in the data that may control the model properties, we looked at some diagnostics for leverage and influence.

The **hat diagonal values** shown in Appendix-2 represent the distance of each observation from the center of x space. Large hat diagonal values indicate that the observations are remote in x-space from the rest of the sample. Leverage points do not necessarily influence regression, but a large value of hat diagonal and large residuals are likely to be influential. From the **R-Student residual vs Leverage** plot below, there are no leverage points that influence the regression. Re-fitting the model without these observations does not result in an increase in *MS.Res* which confirms that the leverage points are not influential.

	b.Intercept	b.CGPA	b.GRE.Score	b.LOR	b.TOEFL.Score	b.Research	MS.Res
full model	-1.44197911101656	0.146211906840582	0.00217618641429935	0.0254091216066022	0.00395741931574085	0.0287916464517809	0.00505132332425327
model without leverage points	-1.40368138840589	0.160569440870527	0.00172418532953567	0.0220309152968797	0.00388617736813345	0.0290588506818465	0.00515829483279646



The deletion influence of an i^{th} observation on the predicted value can be determined by examining the DFFITS statistic (Appendix-3). There are several influence points based on the cutoff value. Removing these values from the model, the MS.Res reduces indicating that the influence points were pulling the model in their direction.



	b.Intercept	b.CGPA	b.GRE.Score	b.LOR	b.TOEFL.Score	b.Research	MS.Res
full model	-1.44197911101656	0.146211906840582	0.00217618641429935	0.0254091216066022	0.00395741931574085	0.0287916464517809	0.00505132332425327
model without influence points	-1.42032699101047	0.14977850767188	0.00203712006547255	0.0256538270502602	0.0039420902293596	0.0270891067632231	0.00318568074780711

The **DFBETAs** statistic can be used to determine which observations have an influence on each regression coefficient. For DFBETAs plots see Appendix-4.

The **COVRATIO** statistic can also be used to determine influence points. (Appendix-5) This statistic gives information about the overall precision of regression. COVRATIO greater than 1 implies that the observation improves the precision of estimation while COVRATIO less than 1 implies that the observation degrades the precision of regression. As shown below, fitting the model without degrading points resulted in a reduction in MS.Res and fitting the model without enhancing points made the MS.Res higher confirming this.

	b.Intercept	b.CGPA	b.GRE.Score	b.LOR	b.TOEFL.Score	b.Research	MS.Res
full model	-1.44197911101656	0.146211906840582	0.00217618641429935	0.0254091216066022	0.00395741931574085	0.0287916464517809	0.00505132332425327
model without degrading points	-1.42032699101047	0.14977850767188	0.00203712006547255	0.0256538270502602	0.0039420902293596	0.0270891067632231	0.00475726810543991
model without enhancing points	-1.42790386289996	0.145734527010782	0.00217167824396226	0.025474727569511	0.00388100992296774	0.027954510372359	0.00510821943462651
model without influence points	-1.4852019503365	0.148807793463238	0.00238784925201732	0.0221542795228109	0.0037107282206924	0.0275520827541694	0.00333518010884464

While the removal of the influential data points does lower the *MS.Res* indicating a better model, after analyzing the points, we did not find any patterns nor were we able to discern any problems in the underlying data collection process that would justify the removal of such data points. Weighing this with the fact that removal of these points only marginally improves the model stats, we decided to not remove these influential data points.

4.5 Multicollinearity

On analyzing the relationships amongst the regressors, it was observed that there exists a linear relationship between CGPA, GRE.Score, and TOEFL.Score (refer to section 3.3). We sought out to correct this situation using another regression technique - Ridge Regression. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors and results in more reliable estimates. Hence, we used ridge regression to build a model and test our hypothesis, as discussed in the next session.

4.6 Validation of Model

We used cross-validation to compare different model iterations. The data was split using the **DUPLEX** algorithm for which the *Snee score* was 1.020237 indicating that the volumes of the prediction and validation regions are very similar. Our model variants showed continuous improvements. The prediction mean square error for the final model selected came out to be 0.031. This model involved *arcsin* transformation on *Chance.of.Admit* and then a ridge regression model to combat multicollinearity.

5. Hypothesis Testing and Results

The final model had the following regressors - *CGPA, GRE.Score, LOR, Research,* and *TOEFL.*Score which were regressed against *Chance.of.Admit.* After fitting the model using Ridge Regression, we tested out the following hypothesis keeping the confidence level at 95%. The significance of a variable, in this case, is determined if it's p-value is less than 0.05 and if its coefficient passes the t-test and its t-value is greater than the t-critical value calculated. (Refer Appendix-6 for the model summary.) The following hypothesis tests were performed -

1. Is a student's CGPA a significant factor in determining his probability of enrollment?
   The p-value of the coefficient associated with this variable (β) is < 0.05 and the associated t-value > t-critical value making it a significant factor.
2. Is a student's SOP rating significant in estimating his probability of enrollment given other strong factors?
   The p-value of the coefficient of SOP obtained from the t-test was 0.087, making it insignificant.
3. Does university ranking play a significant role in determining enrollment success?
   The p-value of the coefficient of University rating obtained from the t-test was 0.086, making it insignificant.

To test the entire model fit, we test for significance of regression which is a procedure often thought of as an overall test of model adequacy. We define a null and an alternate hypothesis –

$H0$: $\beta 1 = \beta 2 = \dots \beta k = 0$ *(Null Hypothesis)*

$H1$: $\beta j \neq 0$ for at least one j *(Alternate Hypothesis)*

The test for $H0$ is carried out using the following statistic-

$$F0 = MS\ \text{R} / MS\ \text{Res}$$

The null hypothesis, $H0$, is rejected if the calculated statistic, $F0$, is such that:

$$F0 > F\alpha, k, n\text{-}k\text{-}1$$

Our final model fit had an F-score ($F0$) of 260.72. The large F-score value implies that the model fit passed the significance of regression test. Further, the p-value associated with the coefficients was found to be less than 0.05 making all of them significant variables. The model has an adjusted R-square of 85.89%. All the 5 variables included in the model have a correlation of > 70% with the response variable. On using the model for prediction, we obtained a test error (MSE) of 3.1% on the predicted values. We also looked at the prediction intervals built around the predicted values (Appendix-8). Overall, the model fit seemed adequate and it's an improvement from the initial linear model fit.

6. Conclusion

After selecting the final model using cross-validation, we were able to test our hypothesis and were able to draw certain conclusions about the Graduate Admissions data. We validated our expectation that variables such as CGPA, GRE and TOEFL score are significant factors in predictive modeling of enrollment probability. Another takeaway from the analysis is that factors such as SOP rating and University Rankings aren't necessarily significant when it comes to predictive modeling even though they offer an insight into admission trends. There were two caveats during the data analysis – high multicollinearity in the data which was corrected by using variable penalizing techniques such as ridge regression and presence of influence points. The lack of additional knowledge about the data collection process makes it hard to justify the removal of influential data. However, even with the presence of these data points, the model quality is good. The fitted model on the complete data is accurate in predicting the probability of admission and nature of data makes the interpretation of the statistical analysis easy to understand.

7. References

1. https://bigfuture.collegeboard.org/get-in/applying-101/applying-to-college-faq
2. https://www.kaggle.com/mohansacharya/graduate-admissions
3. Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Introduction to Linear Regression Analysis (Fifth Edition) Wiley

8. Appendix

1. Raw Data Sample

	Serial.No.	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
1	1	337	118	4	4.5	4.5	9.65	1	0.92
2	2	324	107	4	4	4.5	8.87	1	0.76
3	3	316	104	3	3	3.5	8	1	0.72
4	4	322	110	3	3.5	2.5	8.67	1	0.8
5	5	314	103	2	2	3	8.21	0	0.65
6	6	330	115	5	4.5	3	9.34	1	0.9

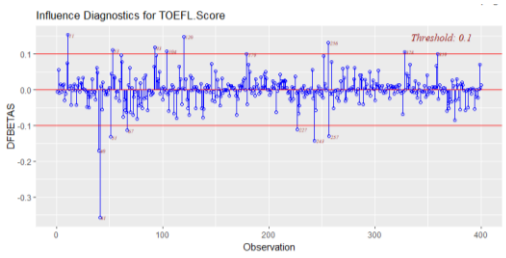
2. Leverage points

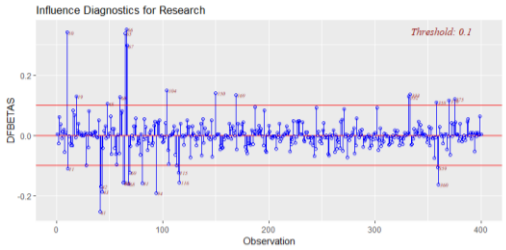
	Chance.of.Admit	CGPA	GRE.Score	LOR	TOEFL.Score	Research	Leverage
22	0.991156586431192	8.4	325	2	114	0	0.0341525052868918
32	1.03572551959974	8.3	327	4	103	1	0.0312416401691009
48	1.23273107201457	9.7	339	4	119	0	0.0328684830974102
51	1.05882363874542	8.3	313	4.5	98	1	0.0318771947218259
53	1.08259106339798	8	334	3	116	1	0.0634611393791657
55	0.991156586431192	8	322	3.5	110	0	0.030800956453124
56	0.927295218001612	7.7	320	3	103	0	0.0337886790595503
57	0.927295218001612	7.4	316	3	102	0	0.0400144196081922
59	0.643501108793284	6.8	300	2	99	1	0.0538338611706369
118	0.735314452816668	7.46	290	2.5	104	0	0.036071335162115
126	0.927295218001612	8.66	300	3	100	1	0.0368515498259439
167	0.937744490405147	8.33	302	5	102	0	0.0300664414078335
169	0.927295218001612	7.8	293	4	97	1	0.0393739165121798
252	0.991156586431192	9	316	3	99	0	0.0431011027606497
258	1.08259106339798	8.64	324	5	100	1	0.0444346277204759
259	1.07061671809741	8.76	326	5	102	1	0.0374292828243999
264	0.991156586431192	8.79	324	1.5	111	1	0.0325708776695078
298	1.18729932286396	9.11	320	4.5	120	0	0.0389771822227866
346	0.775397496610753	7.43	316	2	98	0	0.0397113431859026

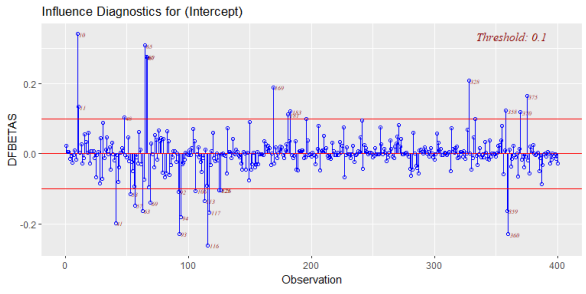
3. DFFITs

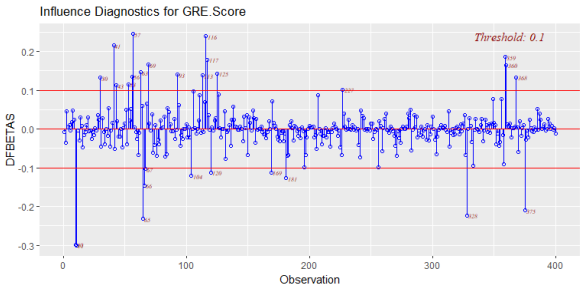
	Chance.of.Admit	CGPA	GRE.Score	LOR	TOEFL.Score	Research	DFFITS.Influence
10	0.735314452816668	8.6	323	3	108	0	-0.474257551039838
11	0.805403500574443	8.4	325	4	106	1	-0.451327045942299
40	0.765392826220454	7.7	307	3.5	108	0	-0.253718845174959
41	0.745355373380619	8	308	3	110	1	-0.476592462008499
42	0.775397496610753	8.2	316	2.5	105	1	-0.274835256089317
43	0.815416192620087	8.5	313	2	107	1	-0.340796471629158
57	0.927295218001612	7.4	316	3	102	0	0.351957562992999
60	0.705052836921493	8.3	311	2	104	0	-0.331293612492277
65	0.805403500574443	8.7	325	3.5	111	0	-0.440091658919767
66	0.835481873978228	8.92	325	3.5	112	0	-0.438109541438948
67	0.896305398645845	9.02	327	3	114	0	-0.438933957575871
69	0.969532110115768	9.22	318	4	109	1	-0.307082489505662
92	0.664215237877967	7.66	299	3.5	97	0	-0.250166659880451
93	0.622533419750133	8.03	298	3	98	0	-0.354087320875712
94	0.725253222200054	7.88	301	3	97	1	-0.271299584966802
116	0.948262907044763	9.04	310	4.5	106	1	-0.368636146110601
117	0.845543104594842	8.62	299	3.5	102	0	-0.258799689481242
328	0.980296311634579	7.86	295	2	101	0	0.339673168162188
349	0.855628870752376	7.25	302	2	99	0	0.272049526189832
359	0.991156586431192	7.64	314	2	105	0	0.373170914662064
360	1.11976951499863	8.44	321	1.5	107	0	0.408793923576471
375	0.674490928149051	7.65	315	2.5	105	0	-0.373310593597901

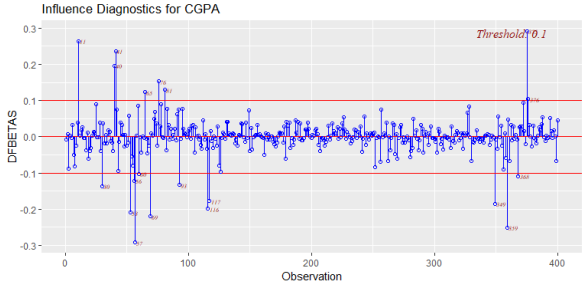
4. DFBETAs

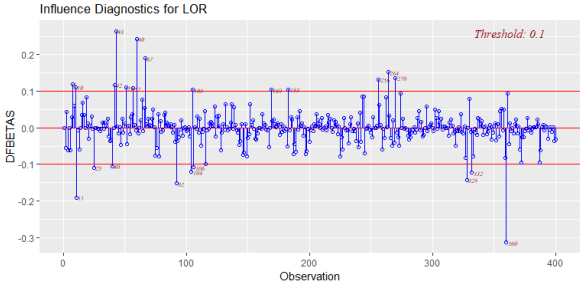












5. COVRATIO

	Chance.of.Admit	CGPA	GRE.Score	LOR	TOEFL.Score	Research	COVRATIO
10	0.735314452816668	8.6	323	3	108	0	0.796028874563457
66	0.835481873978228	8.92	325	3.5	112	0	0.825177807578345
65	0.805403500574443	8.7	325	3.5	111	0	0.834016778276623
11	0.805403500574443	8.4	325	4	106	1	0.866082216568118
93	0.622533419750133	8.03	298	3	98	0	0.882540680200426
67	0.896305398645845	9.02	327	3	114	0	0.890889028980356
60	0.705052836921493	8.3	311	2	104	0	0.90235624606867
68	0.855628870752376	8.64	316	3.5	107	1	0.909744860477446
69	0.969532110115768	9.22	318	4	109	1	0.919656102604607
41	0.745355373380619	8	308	3	110	1	0.930133543215149
42	0.7753974966107 53	8.2	316	2.5	105	1	0.932968157385291
328	0.980296311634579	7.86	295	2	101	0	0.93388204455393
64	0.845543104594842	8.5	315	3	107	1	0.937212553651078
43	0.815416192620087	8.5	313	2	107	1	0.947163027957645
116	0.948262907044763	9.04	310	4.5	106	1	0.9509964248996
375	0.674490928149051	7.65	315	2.5	105	0	0.954159342379797
359	0.991156586431192	7.64	314	2	105	0	0.954386196503193
126	0.927295218001612	8.66	300	3	100	1	1.04581424433494
32	1.03572551959974	8.3	327	4	103	1	1.04742459292375
259	1.07061671809741	8.76	326	5	102	1	1.05029510342907
22	0.991156586431192	8.4	325	2	114	0	1.05120826523375
118	0.735314452816668	7.46	290	2.5	104	0	1.05209117946201
298	1.18729932286396	9.11	320	4.5	120	0	1.05609396253476
252	0.991156586431192	9	316	3	99	0	1.05616430057826
346	0.7753974966107 53	7.43	316	2	98	0	1.05728017094119
258	1.08259106339798	8.64	324	5	100	1	1.06250793078906
59	0.643501108793284	6.8	300	2	99	1	1.06941493195157
53	1.08259106339798	8	334	3	116	1	1.06997305963546

6. Summary of Model with SOP

	Estimate	Std. Error	t - value	Pr(>|t|)	
(Intercept)	-1.4455322	0.1329807	-10.87	< 2e-16	***
CGPA	0.1466457	0.0134032	10.941	< 2e-16	***
GRE.Score	0.0021696	0.0006662	3.256	0.00123	**
LOR	0.0258557	0.0061043	4.236	2.84E-05	***
TOEFL.Score	0.0039889	0.0012068	3.305	0.00104	**
Research1	0.0289138	0.008878	3.257	0.00122	**
SOP	-0.0009054	0.0059107	-0.153	0.87834	

7. Summary of Model without SOP

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.3629892	0.1383111	-9.855	< 2e-16	***
CGPA	0.1406704	0.0134442	10.463	< 2e-16	***
GRE.Score	0.002145	0.0006626	3.237	0.00131	**
LOR	0.0222902	0.0056428	3.95	9.25E-05	***
TOEFL.Score	0.003612	0.0012016	3.006	0.00282	**
Research1	0.0280354	0.0088203	3.179	0.0016	**
University.Rating	0.0086835	0.005051	1.719	0.08637	.

8. Table showing prediction intervals and prediction values

	lower.prediction.interval	predicted.vals	upper.prediction.interval
1	-1.42188964939196	0.791506658187043	3.98996919855866
2	-1.66351777739893	1.07893729852126	3.78116505488977
3	-1.69229207730506	1.10744030517211	3.71868707749578
4	-1.57760334908169	1.02111996711071	3.79190078466987
5	-1.74403058006439	0.968649444722628	3.61951956087468
6	-1.43688710106181	1.05450201826546	3.93497864585832

9. Contributions

The names of the team members below along with their listed contributions serve as electronic signatures pledging that each member contributed towards this project. Additionally, the team collaborated to develop the data pipeline and framework for the project.

Charishma Ravoori – Contributed towards implementing statistical techniques used for diagnostics for leverage and influence points in the data. In addition to writing the code for that, also contributed to the interpretation of outlier analysis and its relevance in the context of meeting objectives of our data analysis. She also contributed to the formulation of the report.

Charu Rawat – Contributed to conducting the exploratory data analysis of the dataset to gain meaningful insights about the data that would serve as starting points to conduct data analysis and identify issues such as multicollinearity. She further built upon the preliminary insights gained to develop a framework for model selection criteria. She also contributed to the formulation of this report ensuring coherence in the integration of all ideas and findings of the analysis.

Saurav Sengupta – Contributed to conducting model adequacy checks which were integral to the statistical analysis. In addition to developing and applying techniques that could help check model adequacy, Saurav's interpretation of the analysis proved to be extremely valuable in enhancing the model by identifying areas of improvement such as variable transformations and multicollinearity and writing the code for it. He also contributed to the formulation of the report.

Sri Vaishnavi Vemulapalli – Contributed towards carrying out specialized diagnostics in detection of leverage and influence data points. Her insights from this study were essential to understanding the potential misgivings of the data analysis. She also contributed immensely in report writing for this project ensuring fluidity and cohesiveness.

Vaibhav Sharma – Built upon the takeaways from the exploratory data analysis to study and implement suitable methods for model building. This encompassed implementing techniques for variable selection as well as methods for cross-validation. This proved to be immensely crucial in developing and writing code for a predictive model of good fit. His interpretation of the model in the context of the potential issues that arose with the data was crucial in determining techniques that could be used for model enhancement such as ridge regression. He also contributed to the formulation of the report.