**STAT 6021: Linear Models for Data Science**

**Project – Using statistical modelling methods to predict probability of enrollment success in graduate admissions**

Project URL: <insert GitHub link>

|  |  |
| --- | --- |
| Charishma Ravoori | [cr2st@virginia.edu](mailto:cr2st@virginia.edu) |
| Charu Rawat | [cr4zy@virginia.edu](mailto:cr4zy@virginia.edu) |
| Saurav Sengupta | [ss4yd@virginia.edu](mailto:ss4yd@virginia.edu) |
| Sri Vaishnavi Vemulapalli | [sv2fr@virginia.edu](mailto:sv2fr@virginia.edu) |
| Vaibhav Sharma | [vs3br@virginia.edu](mailto:vs3br@virginia.edu) |

**Project Report**

**M.S. Data Science**

**University of Virginia**

**December 2018**

**CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Introduction | |  |
| 1.1 | | Problem Statement | 1 |
| 1.2 | | Objective | 1 |
| 2. | The Dataset | |  |
|  | 2.1 Source and Relevance | | 2 |
| 2.2 | | Data Description | 2 |
| 3. | Data Preparation and Exploration | |  |
| 3.1 | | Data Pre-processing | 4 |
|  | 3.2 | Summary Statistics | 4 |
| 3.3 | | Exploratory Data Analysis | 5 |
| 4. | Data Analysis and Statistical Inference | |  |
| 4.1 | | Methods in Model Building | 7 |
| 4.2 | | Variable Selection | 7 |
|  | 4.3 | Model Adequacy | 8 |
|  | 4.4 | Issues - Outlier Diagnostics and Multicollinearity | 8 |
|  | 4.5 | Validation of Model | 9 |
| 5. | Hypothesis Testing and Results | |  |
| 5.1 | | Results | 10 |
| 6. | Conclusion | |  |
| 6.1 | | Conclusions | 12 |
| 6.2 | | Future Enhancements | 13 |
| Appendix | | | 14 |

1. **Introduction**

Each year, over 3 million college applications are filed in the US by about 750,000 students, 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.

**1.1 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 into 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?”

**1.2 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 –

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.

1. **The Dataset**

**2.1 Source and Relevance**

This data was sourced from Kaggle and is owned by user Mohan S Acharya on the Kaggle website. Originally, this data was abstracted 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, some of which like GRE and TOEFL score are 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 admission 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 admission which gives an idea about admission probability of a particular 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.

**2.2 Data Description**

The raw dataset consists of 400 records with 9 attributes wherein each record pertains to an application to a particular university. The dataset contains of several parameters which are considered important during the application for graduate programs. The parameters included are:

1. Serial Number
2. GRE Scores (out of 340)
3. TOEFL Scores (out of 120)
4. University Rating (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)
10. **Data Preparation and Exploration**

The raw data was sourced from Kaggle upon which pre-processing was done to obtain data in a desirable format to perform analysis upon. The resultant data is attached in the appendix in form of a table.

* 1. **Data pre-processing**

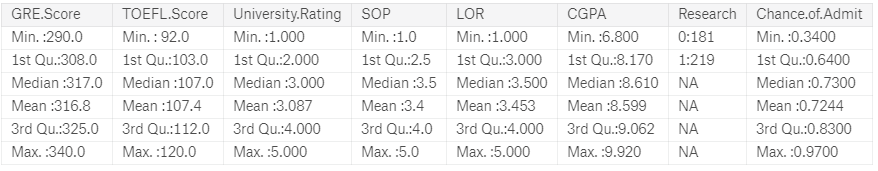
We used the raw csv file – “Admission\_Predict.csv” provided on the Kaggle website.

The file was loaded in R as a data frame upon which the following pre-processing was leveraged to obtain the final format of the data –

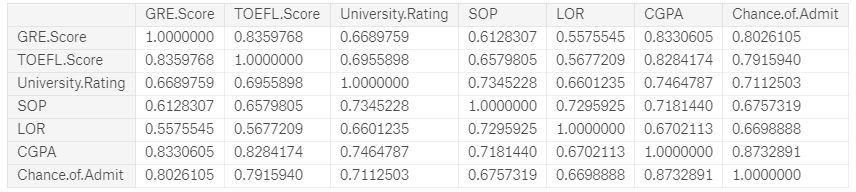
* Dropping variables - We checked for duplicate rows in the data and found none. We dropped the variable *“Serial.No”* from the data since it only served as a primary key to the data and hence had no effect on *“Chance.of.Admit”* which is our response variable.
* Missing value imputation - We checked our data for missing values but found none, hence no missing value imputation was required.
* 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.
  1. **Summary Statistics**

Below is a table detailing the summary statistics on the variables in the data. The mean GRE score of the applicants in the data is 316 and the mean TOEFL score is 107. The mean SOP and LOR rating is 3.4. For all the attributes the mean and the median are quite close to each other in value. The average probability of enrollment of applicants in this data is 0.72 or 72% with the minimum probability being 34% and the maximum probability being 97% for an applicant.

Summary Statistics Table:



Below is a table detailing the correlation summary stats amongst all the numerical variables. It’s observable that most of the variables in the data have an above 70% correlation with the response variable which is “Chance of Admission” or Enrolment probability. Though a caveat here is that there seems to be high multicollinearity in the data.

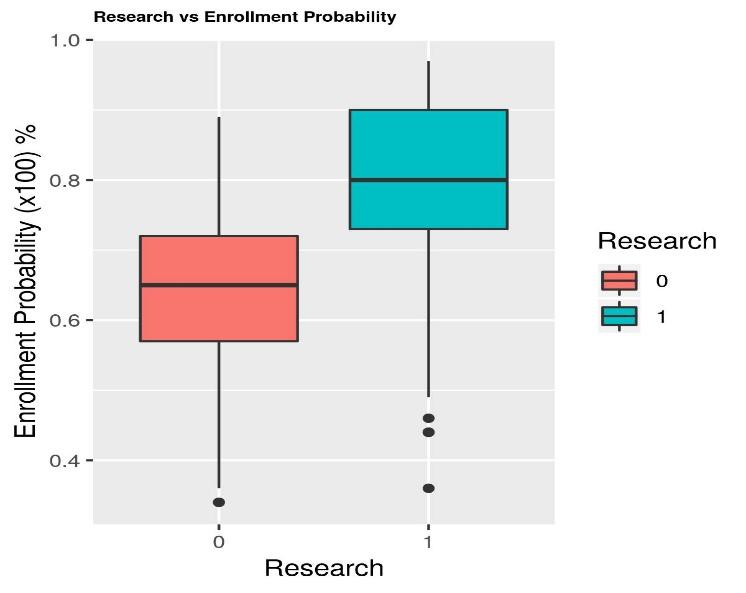


For easier visualisation purpose, below is the correlation matrix that exhibits the relationships amongst the variables. The variables are further grouped on the category “Research” –



* 1. **Exploratory Data Analysis**

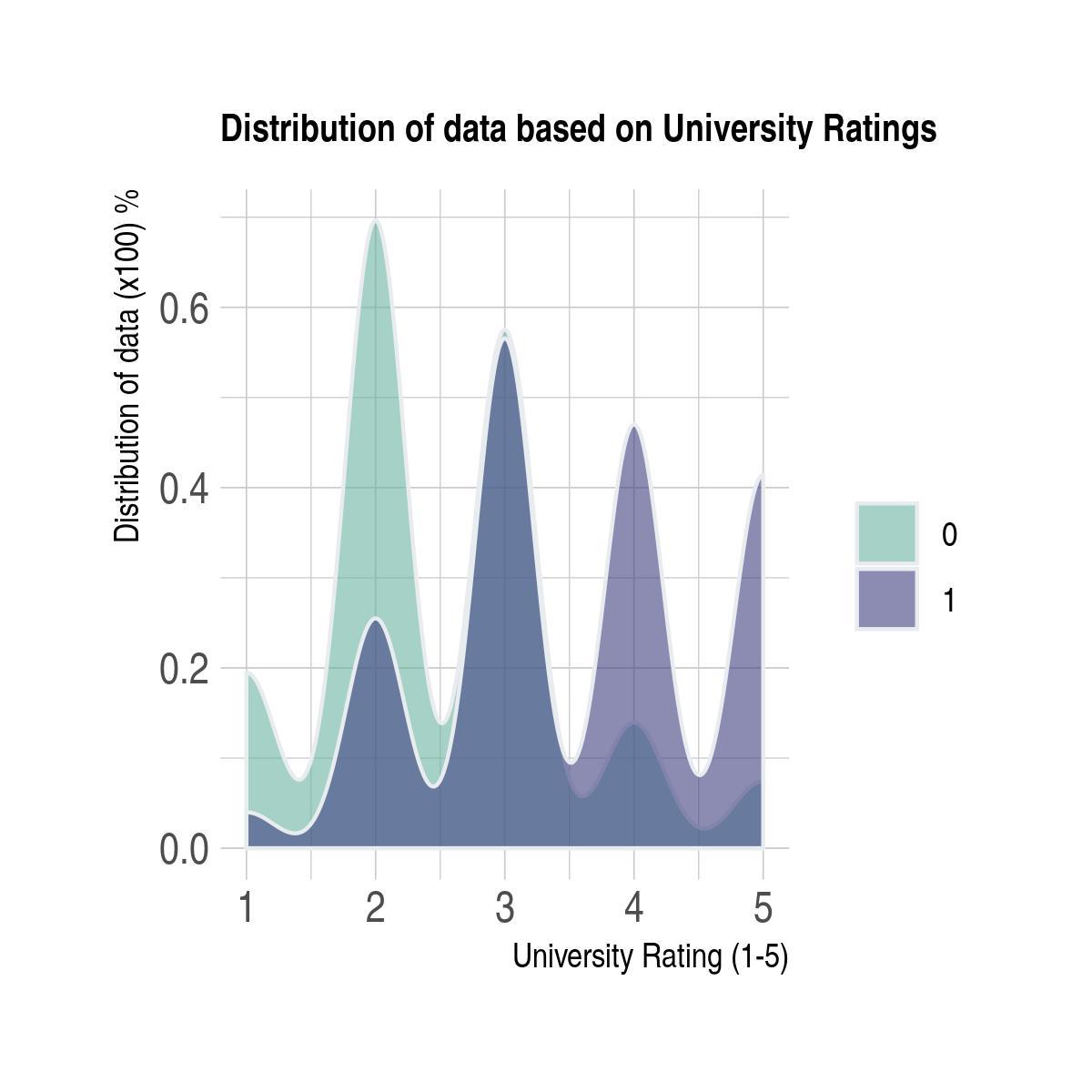
We were able make certain observations about the data through exploratory data analysis. Some of our observations are summarized below –

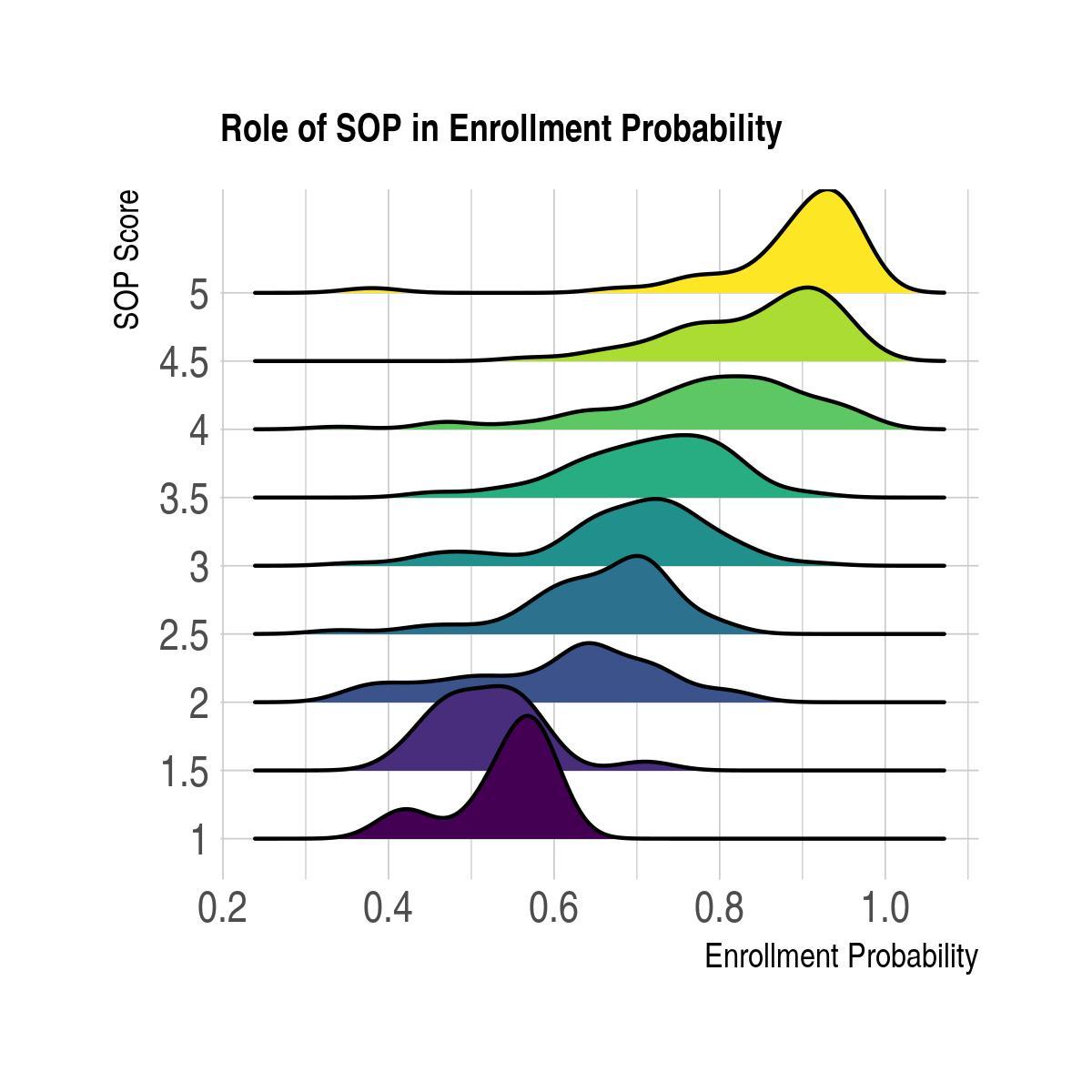


* Students who have done research work

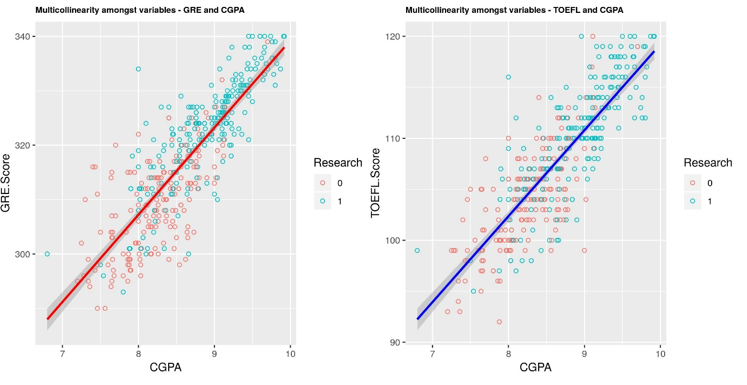
In the past (Research = 1) have a higher chance of admission (~80% chance) on average as compared to students with no prior research experience (Research = 0) who have just a 60% chance of enrolment. Having said that, there are certain exceptions in form of outliers to this trend that are visible in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

* Each university has a rating assigned from 1 to 5 (where 5 is the highest a university can get) to which a student applies. From the plot below, we can observe that in this data, students with no research experience typically exclusively apply to universities rated lower (rated 1 or 2) whereas 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).

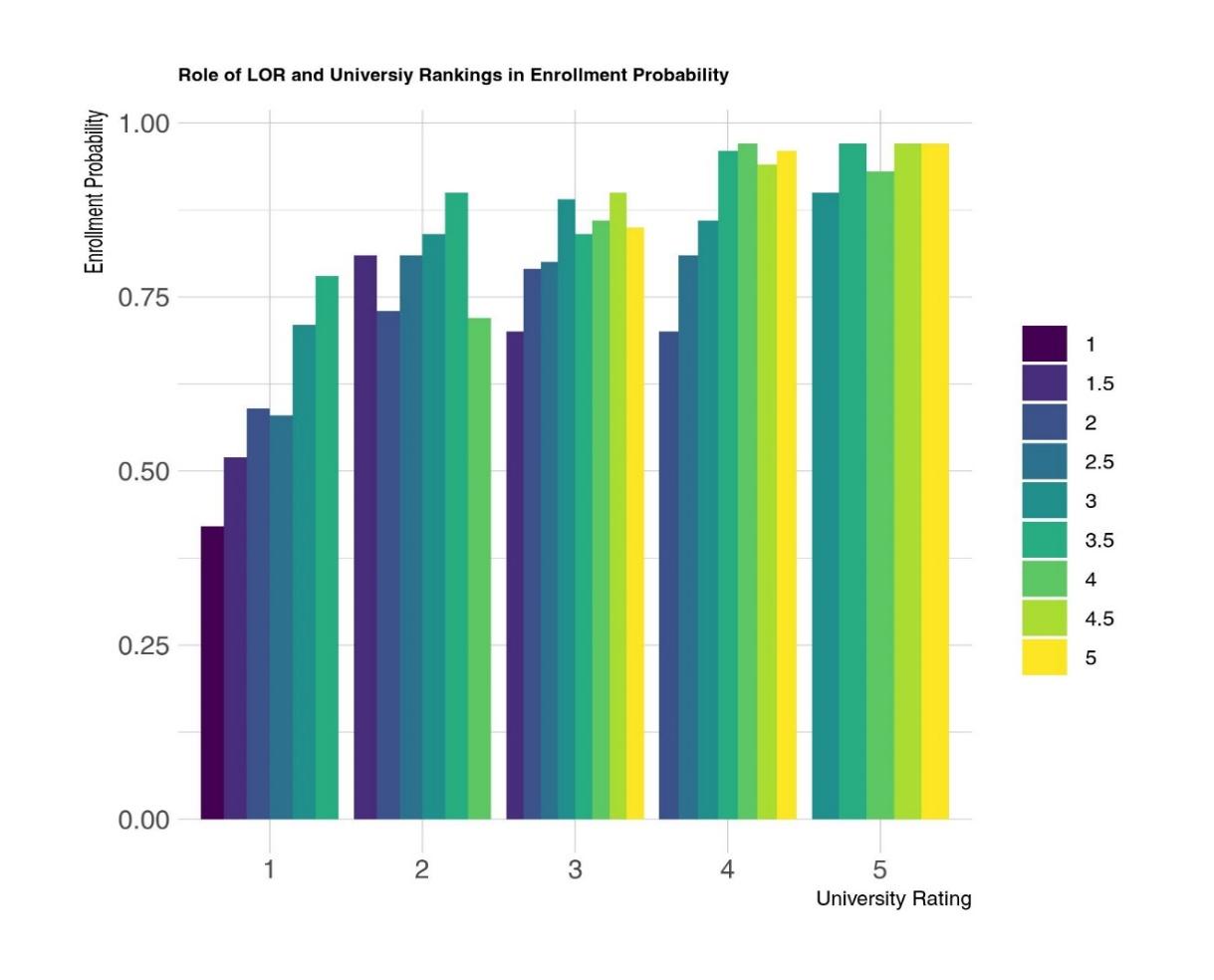


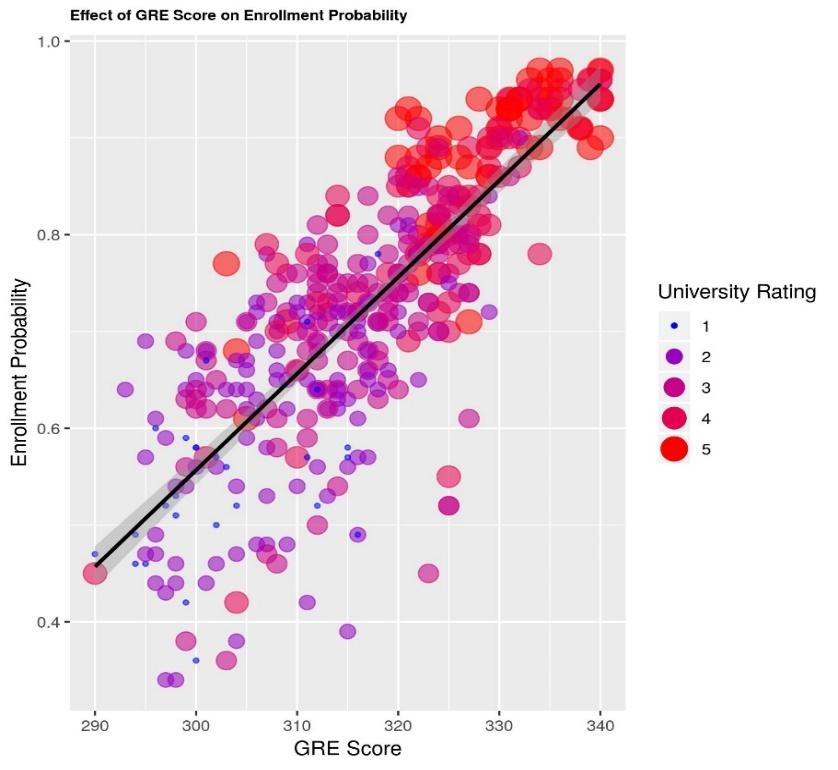


* SOP (Statement of Purpose) score plays an important role in determining enrolment success. A we can see from the plot below, the higher the SOP rating, the greater the chance of enrolment or admission success. Typically, SOP’s rated between 3 and 4.5 have a wider range of enrolment probability - roughly between 60 and 90% on enrolment success.
* 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 have addressed handling this issue further in our analysis detailed in this report.

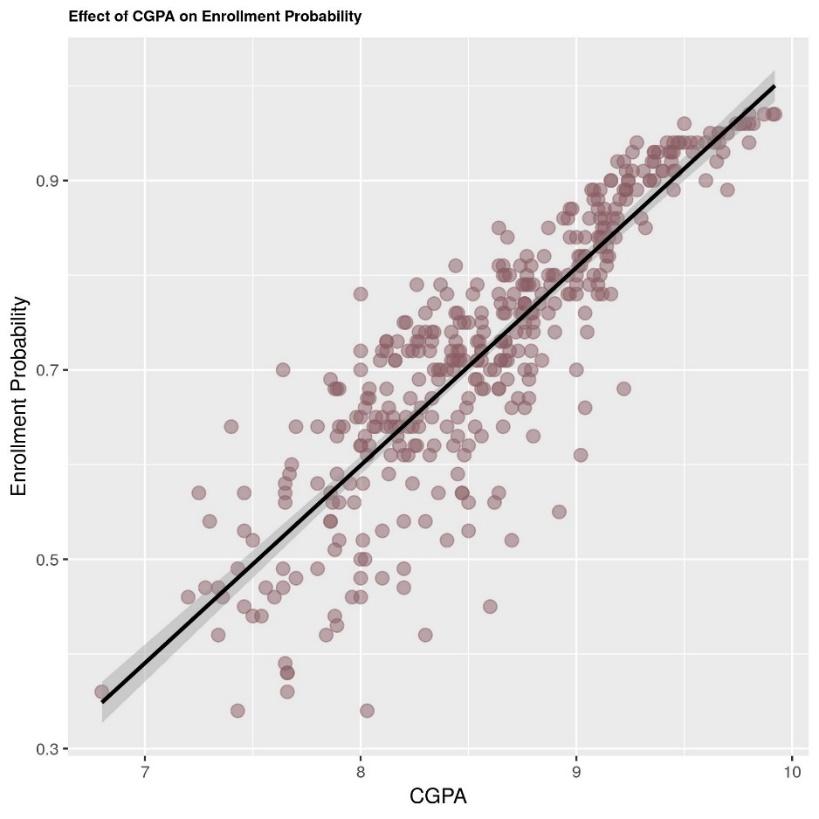


* Another metric similar to SOP score is the LOR rating that each student submits along with their application. We explored the effect that LOR and university ratings have on enrolment success. We observed that for applications submitted to universities rated lower (rated 1 or 2), there is more variation in the enrolment 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 chance of enrolment success as can be seen from the plot below. We also noticed that students with LOR’s rated 4 and above only tend to apply to universities rated above 3 which is somewhat expected.



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

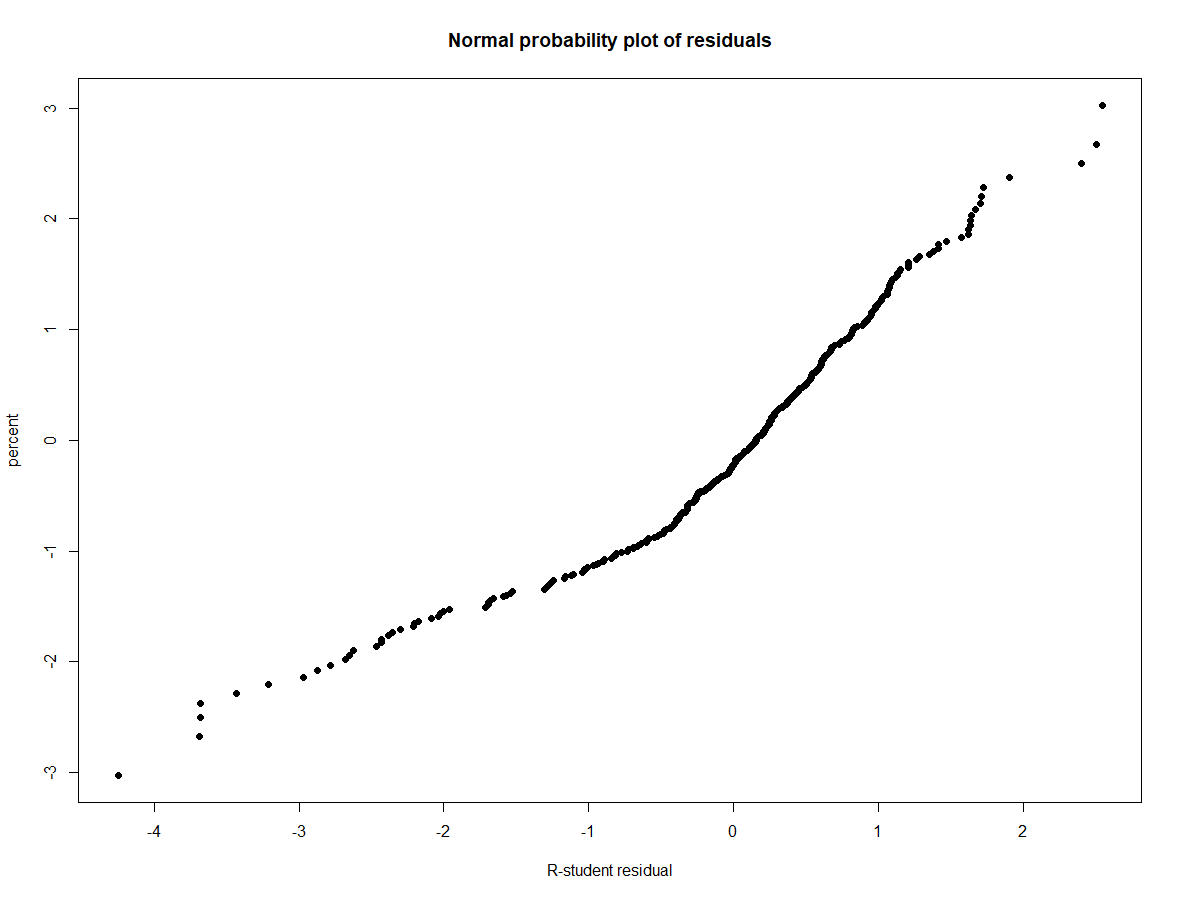
We can also observe from the plot below that applicants with a higher score tend to mostly apply to universities rated higher as we can see the larger sized points cluster towards the top end.



* Lastly, we looked at the relationship between CGPA and enrolment 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 enrolment probability remains lower than expected. This could be due to other factors such as GRE Score, LOR rating etc.

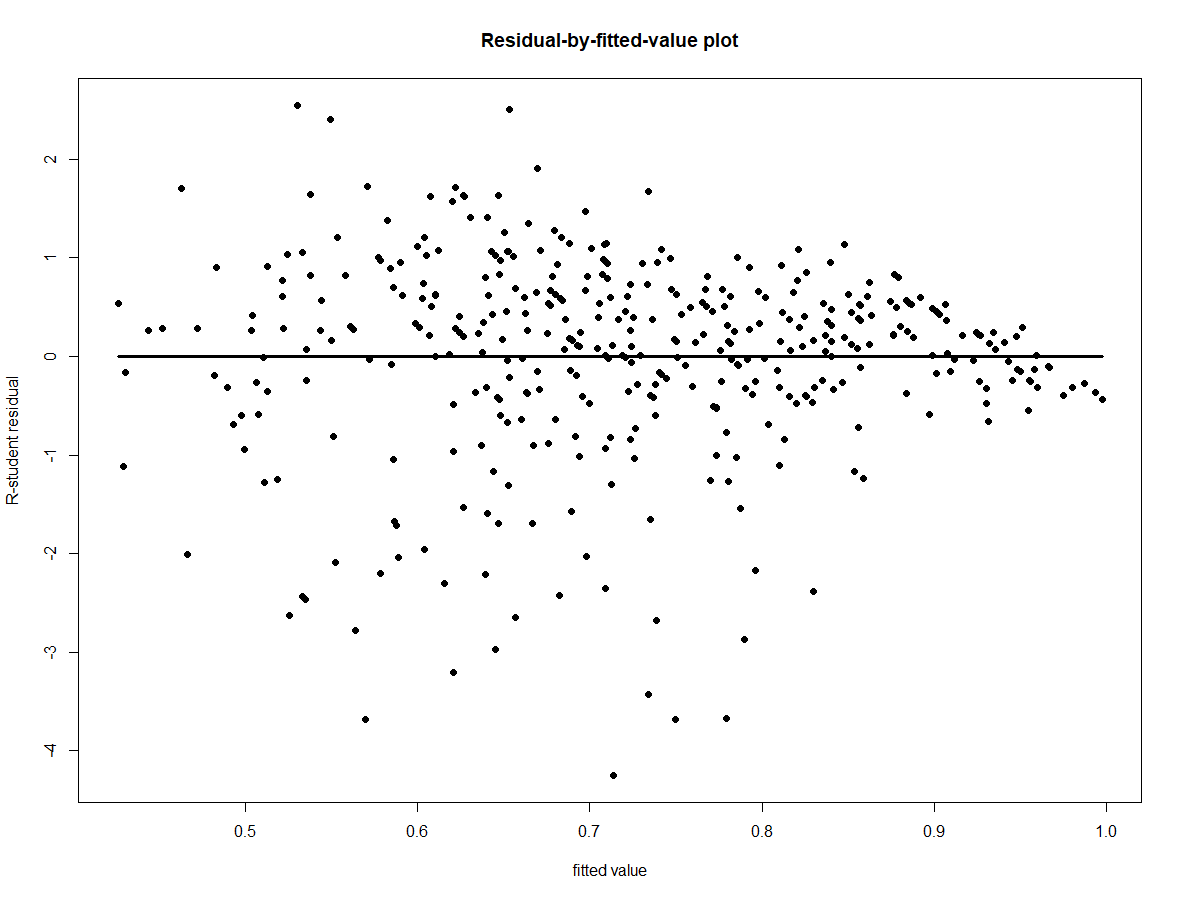
Model adequacy checking

We are trying to determine if the normality assumption of the error variance holds up during our analysis. For that we will build the **Normal Probability Plot** that will help us check that assumption.



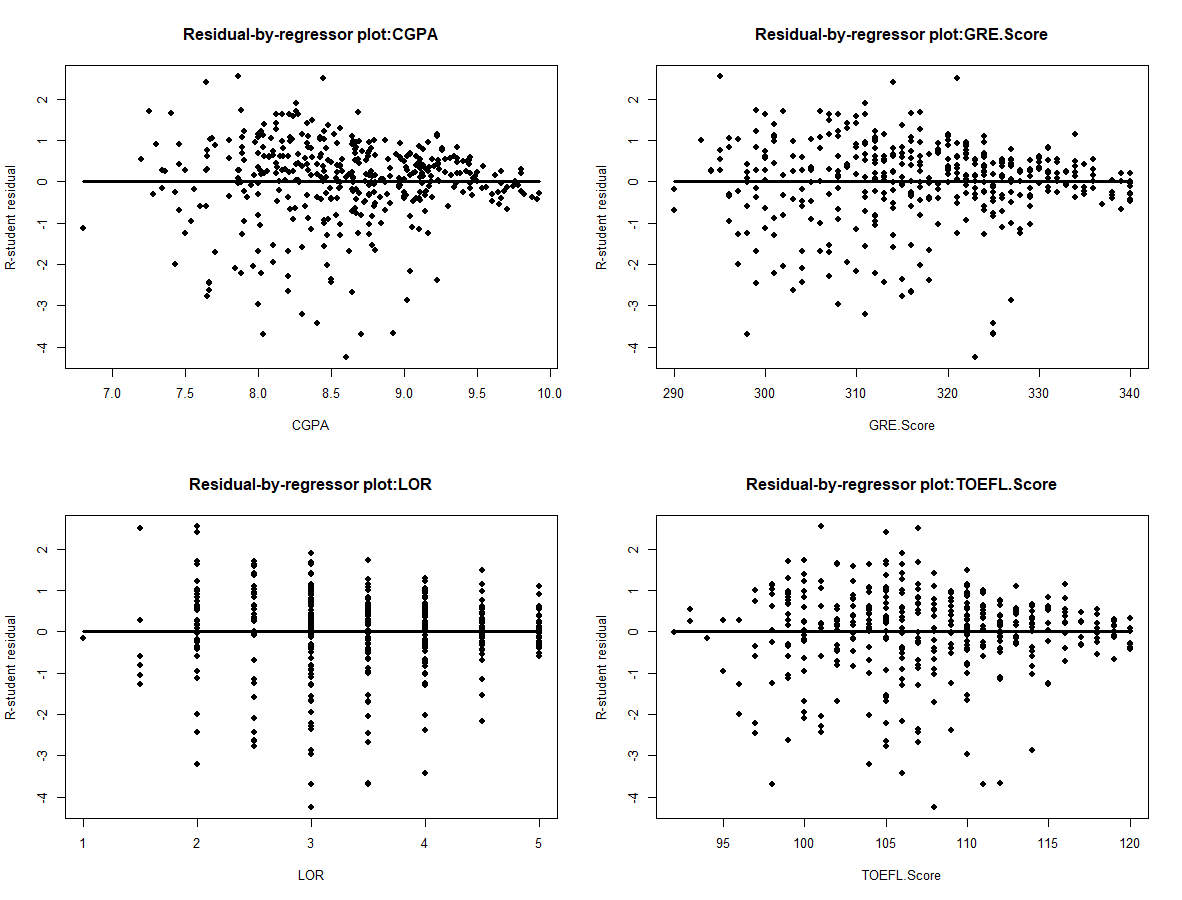
We can see that the points mostly lie on the straight line but the plot does show some heavy tails, that suggests that the normality assumption might not be completely accurate. However the tails are slight and not really prominent, but we need to further investigate different residual plots to test the normality assumption.

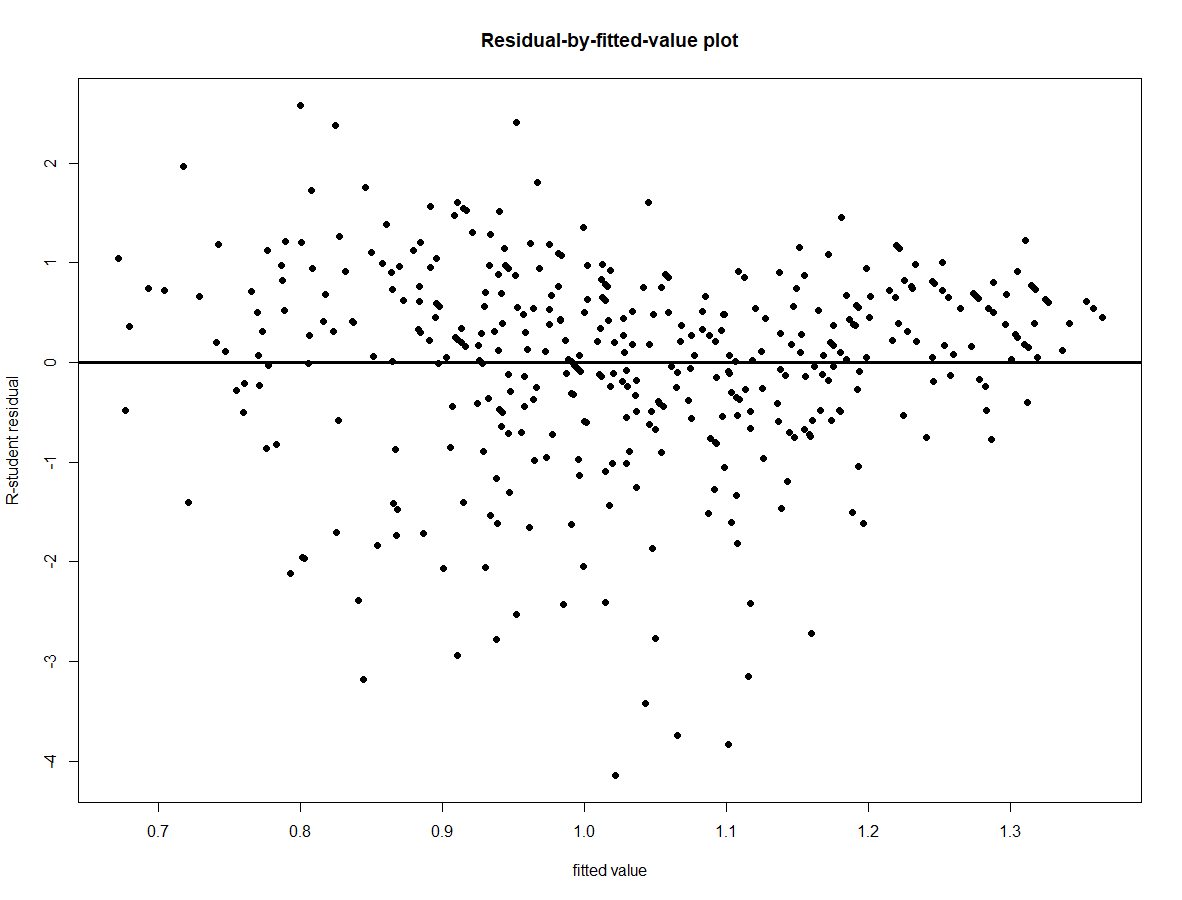
**Residual-by-fitted-value plot:**



This plot shows a double bow pattern, which is expected since our outcome variable is a proportion between 0 and 1, indicated the chances of getting an admit.

We investigate this further with more plots with **residuals vs regressors plots:**



All of them show the double bow pattern and hence points us towards variance stabilizing transformations. The recommended transformation is to convert the outcome to the arcsin of its square root. We investigate the residuals vs fitted value plots to check if the normality assumption is preserved.

The residual vs fitted value plot looks better than the previous plots since we the points lie more or less on a horizontal band.