The Admission Predictors
STAT 4355 | Prof. Chuan-Fa Tang

Graduate Admission Analysis

Prepared by:
Shriram Rajasekar & Sarvesh Gopalakrishnan
December 11, 2024

TABLE OF CONTENTS

1. Introduction

1.1. Background

Graduate admissions represent a pivotal academic journey, influencing career trajectories and personal development. Understanding the factors determining admission chances is crucial for students aiming to maximize their potential. This report focuses on analyzing graduate admission probabilities using a dataset sourced from Kaggle. The dataset comprises 500 observations and eight key predictors, including GRE scores, TOEFL scores, CGPA, and research experience. The target variable, "Chance of Admission," is measured on a continuous scale from 0 to 1 in the data set. The study aims to uncover the underlying relationships between predictors, assess their impact on admission probabilities, and build a predictive model to guide future applicants. Through comprehensive exploratory data analysis, statistical modeling, and evaluation, this report offers insights into the factors that most significantly influence graduate admission outcomes.

1.2 Dataset Information

The *Graduate Admission Prediction* data set from Kaggle contains 500 entries of data consisting of strong academic students with 8 variables: GRE Scores, TOEFL Scores, University Rating, Statement of Purpose Strength, Letter of Recommendation Strength, Undergraduate CGPA, Research Experience, and Chance of Admission.

Target Variable:

1. Chance of Admission: Continuous (0 to 1).

Predictor Variables:

2. GRE Score (Discrete): Ranges from 260 to 340.
3. TOEFL Score (Discrete): Ranges from 0 to 120.
4. Statement of Purpose (SOP) (Discrete): Ranges from 1 to 5.
5. Letter of Recommendation (LOR) (Discrete): Ranges from 1 to 5.
6. Undergraduate GPA (Continuous): Ranges from 0.0 to 10.0 (on a scale of 10).
7. Research Experience (Binary): 0 = No, 1 = Yes.
8. University Rating (Discrete): Ranges from 1 to 5.

1.3 Analysis Goal

This project focuses on understanding the key factors influencing graduate admissions and building predictive models to enhance decision-making. Specifically, it investigates the relationships between undergraduate university ratings, students' CGPA, examination scores, and other related factors to determine their impact on admission chances. A central question is how the prestige or rating of an undergraduate institution affects a student's likelihood of being admitted to graduate programs.

Additionally, the analysis aims to identify the most influential predictors of graduate admission. By determining which factors carry the most weight, we can provide clearer insights into how applications are evaluated. The project also explores the development of effective predictive models to assess admission chances. A comparative analysis of different modeling approaches will identify the method that delivers the highest predictive performance, enabling institutions and applicants to make more informed decisions.

The following questions were the main objectives for this project:

- What are the relationships between the undergraduate university rating and a student's CGPA, examination scores, etc?
- How does the undergraduate university rating impact a student's admission chance?
- Which predictors are most influential in predicting a student's graduate admission?
- How can we build an effective model to predict a student's graduate admission chance?
- Which model leads to the best performance in predicting a student's graduate admission chance?

2. Data Cleaning, Challenges, and Limitations

The data cleaning process is fundamental to ensure the dataset is ready for analysis and modeling. For this project, the following steps were undertaken to prepare the dataset:

Null Value Check: The dataset was thoroughly inspected for missing values.
Outcome: No null values were found.

Duplicate Removal: Checked for duplicate entries to ensure data integrity.
Outcome: Duplicates were identified and removed to prevent data redundancy.

Column Renaming: To streamline coding and analysis, variable names were renamed for clarity and usability. Example: "University Rating" was abbreviated as "Univ_Rating" or "UnivRtg".

Target Variable Transformation: The target variable, "Chance of Admission," was rescaled from its original 0-1 range to 0-100 for interpretability.

Data Format Standardization: All numerical variables were formatted consistently (e.g., GRE and TOEFL scores were kept as discrete numerical values).

Verification: Final checks were conducted to confirm the dataset's readiness for further analysis.

While the data cleaning process yielded a refined dataset, some challenges were encountered:

Lack of Additional Predictors: The dataset primarily focuses on academic and research metrics, lacking contextual factors like work experience, extracurriculars, or personal statements.

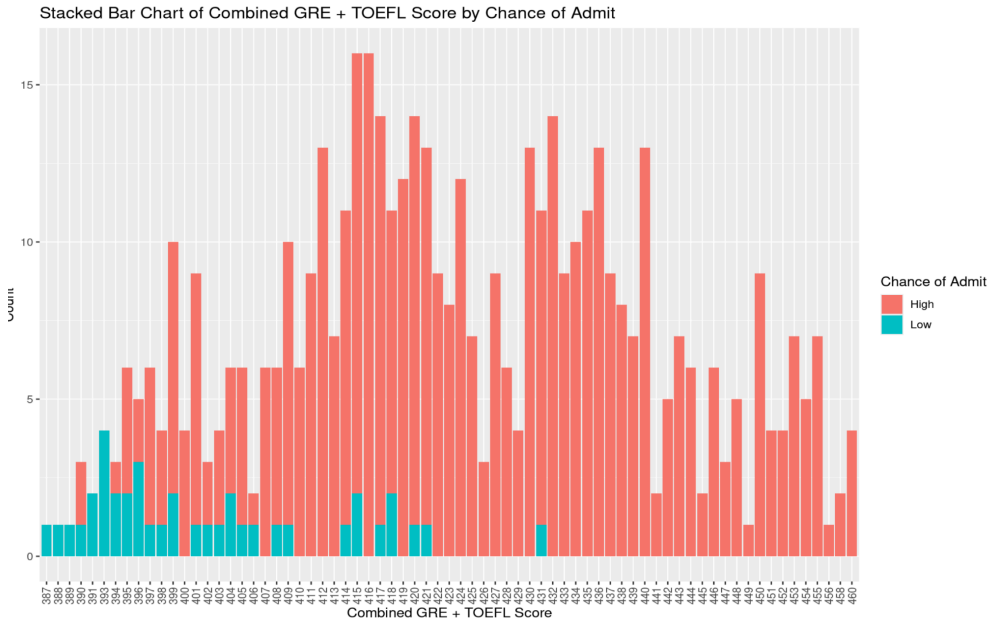
Limited Scope: The data represents academically strong students, potentially introducing selection bias.

These limitations highlight the need for cautious interpretation and the potential benefit of integrating additional data sources in future studies.

3. Exploratory Data Analysis

The exploratory analysis provides a comprehensive overview of the relationships between key variables and their influence on graduate admission chances. It sets the stage for further statistical modeling to identify the most impactful predictors and to build an effective predictive framework.

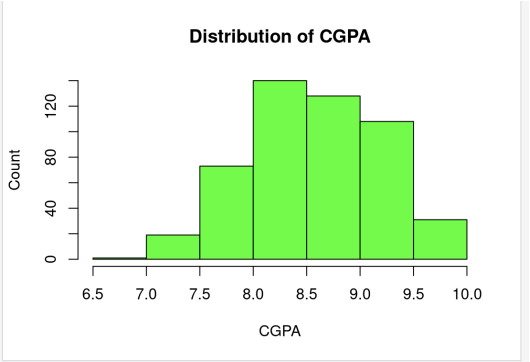
3.1 Combined GRE and TOEFL Scores vs. Admission Probability



Visualization: A stacked bar chart illustrating the relationship between high and low admission probabilities based on combined GRE and TOEFL scores.

Findings: A clear positive trend emerges, showing that applicants with higher combined scores have significantly better admission chances. The data reveals that low combined scores predominantly result in lower probabilities of admission, emphasizing the importance of excelling in both GRE and TOEFL exams. Applicants scoring in the top quartile for both tests frequently achieve admission chances above 80%.

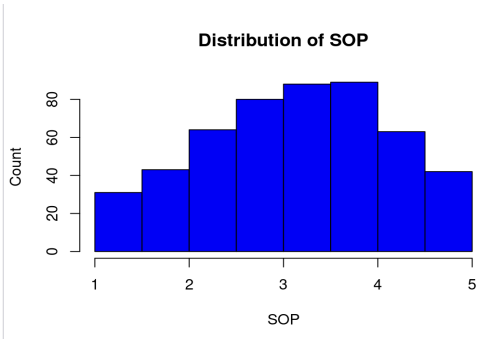
3.2 Distribution of CGPA

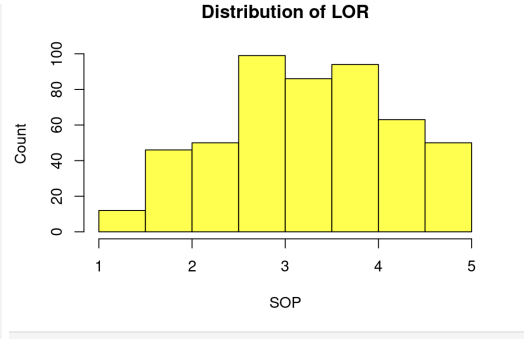


Visualization: A histogram of CGPA values across the dataset.

Findings: The CGPA distribution follows a roughly bell-shaped curve between 8.0 and 9.0. Most applicants exhibit consistent academic performance, with CGPAs clustering around this range. Fewer observations occur at the lower and upper extremes (below 7.0 or above 9.5), suggesting that students applying for graduate studies generally maintain high academic standards.

3.3 Distribution of SOP and LOR Scores

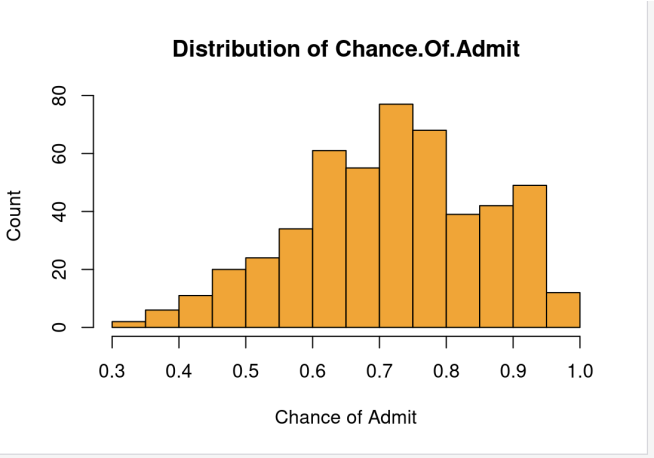




Visualization: Histograms representing Statement of Purpose (SOP) and Letter of Recommendation (LOR) scores.

Findings: SOP scores are evenly distributed, with most applicants scoring between 2 and 4. There is a noticeable decline at the extremes (scores of 1 and 5), indicating that very few applications have exceptionally weak or strong SOP ratings. LOR Findings: LOR scores display a similar trend, with most scores clustering around moderate to strong endorsements (3 to 4). The balanced distribution of LOR ratings reflects the competitive nature of graduate applications, where strong recommendation letters are critical.

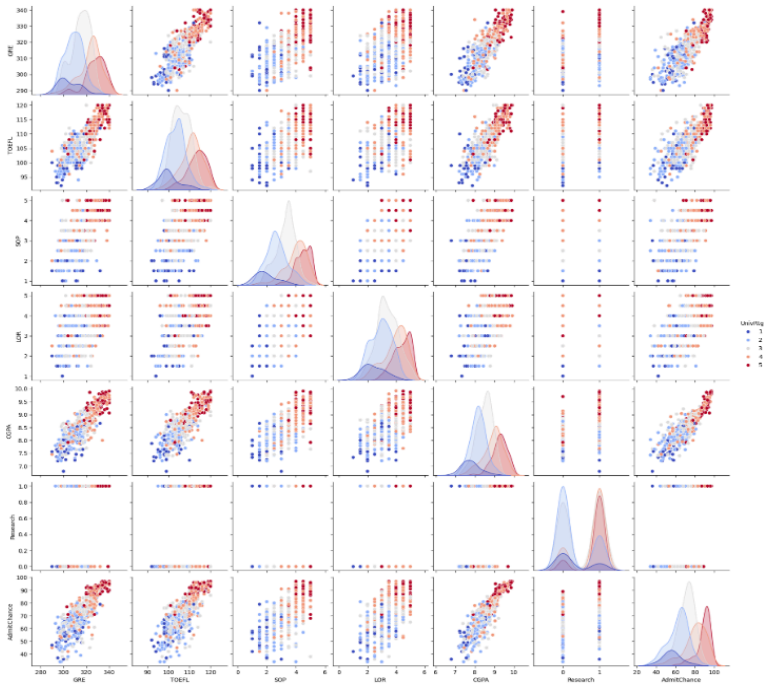
3.4 Admission Chances

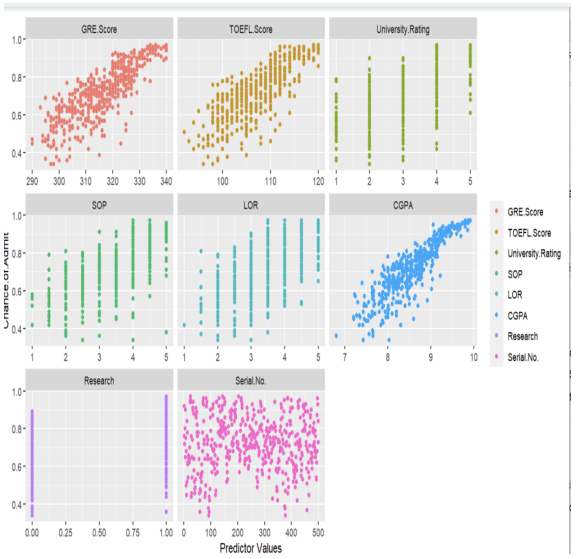


Visualization: Histogram of admission probabilities, scaled from 0 to 100.

Findings: The distribution is right-skewed, with a significant concentration between 60% and 80%. Fewer observations occur at the extremes, with very few applicants having probabilities below 40% or above 90%. This suggests that the dataset largely represents students with a fair to good chance of being admitted, reinforcing the selection bias toward academically strong applicants.

3.5 Relationships among Predictors

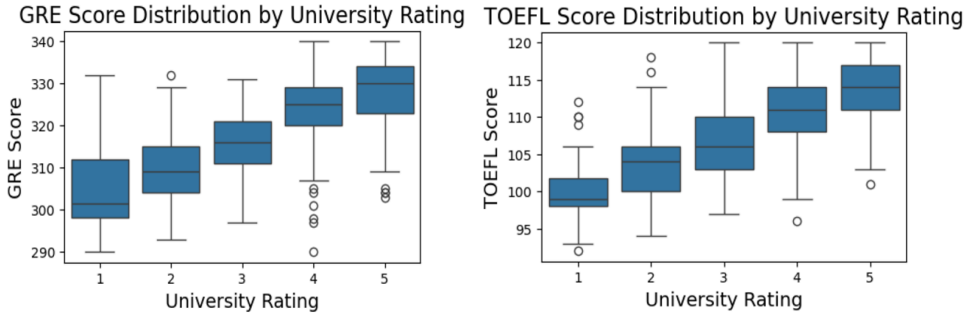


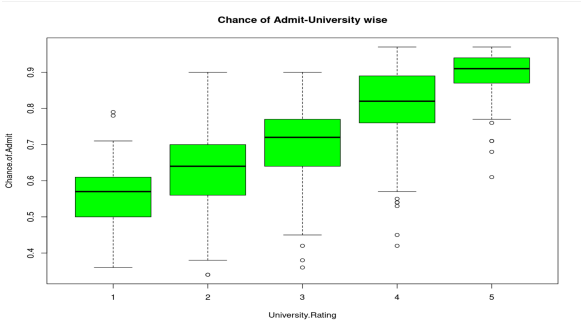


Visualization: Scatterplots of "Chance of Admission" against key predictors (GRE, TOEFL, CGPA).

Findings: GRE Scores: A strong positive correlation exists, with higher GRE scores directly linked to increased admission probabilities. TOEFL Scores: Similar to the GRE, TOEFL scores show a linear relationship, underscoring the importance of language proficiency in graduate admissions. CGPA: Of all predictors, CGPA exhibits one of the strongest correlations, suggesting that consistent academic performance during undergraduate studies is a significant factor in admissions decisions.

3.6 Influence of Undergraduate University Ratings

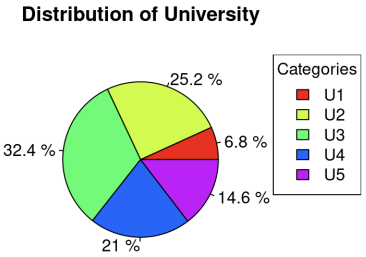


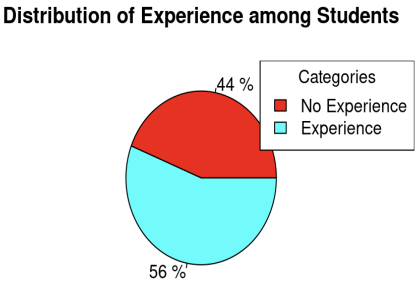


Visualization: Box plots of admission chances, GRE, and TOEFL scores categorized by university ratings.

Findings: Admission Chances: Students from universities with higher ratings (4 and 5) are more likely to secure admissions, with median chances significantly higher compared to students from lower-rated universities. GRE and TOEFL Scores: Median scores for these exams also increase with university ratings, implying that students from prestigious institutions tend to perform better in standardized tests.

3.7 Research Experience and University Distribution





Visualization: Pie charts displaying the distribution of students with and without research experience and the frequency of university ratings.

Findings: Research Experience: Approximately 56% of applicants have research experience, which positively influences their admission chances. University Ratings: The majority of applicants come from universities rated 3, followed by ratings 2 and 4. This trend highlights the dataset's representation of mid-tier institutions, with fewer students from top-rated universities (5).

4. Introductory Statistical Testing

```
ANOVA Test Statistic: 114.00804341400001
ANOVA p-value: 7.753395328022908e-69
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
     1      2   6.4052 0.0111  0.9958 11.8146   True
     1      3  14.0842    0.0  8.8041 19.3644   True
     1      4   23.956    0.0 18.4329 29.4792   True
     1      5  32.6023    0.0 26.7906 38.4141   True
     2      3    7.679    0.0  4.3542 11.0038   True
     2      4  17.5508    0.0 13.8522 21.2494   True
     2      5  26.1971    0.0   22.08 30.3142   True
     3      4   9.8718    0.0  6.3649 13.3786   True
     3      5  18.5181    0.0 14.5723 22.4639   True
     4      5   8.6463    0.0  4.3808 12.9118   True
-----------------------------------------------------
  UnivRtg   Lower CI   Upper CI
0       1  52.736036  59.675729
1       2  60.709402  64.512820
2       3  68.762979  71.817268
3       4  77.891338  82.432471
4       5  87.064981  90.551457
```

4.1 ANOVA Test for Difference in Means

One of the main goals of this project was to explore the difference in admission chances among the different university ratings from 1-5. Therefore, an ANOVA test was initially conducted to test for the difference in mean admission chances among the 5 groups. The null hypothesis for the ANOVA test is that there is no significant difference in the mean admission chance among the 5 groups. The alternate hypothesis for the ANOVA test states that there is a significant difference in the mean admission chances among the 5 groups. Referencing the image above, the p-value for the ANOVA test was approximately 7.75e-69 < $\alpha = 0.05$ with a test statistic of 114. As a result, the null hypothesis was rejected and this suggests that there is a significant difference in the mean admission chance among the 5 university rating groups.

4.2 Post Hoc Analysis for Difference in Means

After conducting the ANOVA test for the 5 different university rating groups, the Post Hoc Analysis test was used to determine if there was a significant difference in the mean admission chance among each pairing of university rating groups (ex. 1 vs 2, 4 vs 5, 2 vs 3, etc.). From the output of the post hoc analysis, the adjusted p-values for every individual test between every pair are less than 0.05. Therefore, this suggests that there is a significant difference in the mean admission chance among every individual pair of university rating groups.

4.3 Confidence Intervals for Means

95% confidence intervals to capture the true mean admission chance were built for every individual university rating group. From the individual confidence intervals, the lower/upper tails and the point estimate for these intervals increase as the university rating group increases.

5. Predictive Modeling and Testing

In this project, the following model types were explored and trained/tested utilizing OLS (Linear Regression) and XGBoost. There were a total of 8 models built.

Type 1: No University Rating w/ No Normality
Models: OLS (1), XGBoost (2)
Type 2: No University Rating w/ Normality
Models: OLS (3), XGBoost (4)
Type 3: All Predictors w/ Normality
Models: OLS (5), XGBoost (6)
Type 4: Dummy Encoding the University Rating & w/ Normality
Models: OLS (7), XGBoost (8)

5.1 The Implementation of XGBoost

The machine learning model that was utilized alongside linear regression in this project was XGBoost which is a decision-tree-based algorithm. Initially, XGBoost will calculate residuals based on an initial prediction. These residuals are iteratively split into boolean arguments that will be used in the decision tree. For every split, the similarity scores are calculated ((sum of residuals)^2 / (number of residuals)) and the gain of the split is calculated ((left node's similarity score) + right node's similarity score) - parent node's similarity score)). XGBoost maximizes the gain when building decision trees, so the split which leads to the largest gain will be included in the decision tree. This process is repeated in every boosting round and new output values are made ((sum of the residuals in split) / number of residuals in split)). Once new output values are calculated, the learning rate parameter determines the influence of these new output values on the previous prediction (previous prediction = previous prediction + output values * (learning rate)). Through every boosting round, XGBoost improves its previous prediction and terminates once the loss function (in this case, MSE) converges to its minimum.

The use case for XGBoost is usually on data sets with many observations and many predictor variables. This is because of the various parameters that can reduce overfitting by variable selection, cross-validation, and regularization. In this project, XGBoost was implemented out of curiosity and on the basis of learning ML models.

Here is the following XGBoost parameter grid which was trained on the XGBoost models utilizing the XGBoost library in Python:

```
param_grid = {
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3],
    'n_estimators': [100, 200, 300],
    'subsample': [0.6, 0.7, 0.8],
    'colsample_bytree': [1.0],
    'gamma': [0, 0,1],
    'lambda': [0, 0.5, 1],
    'alpha': [0, 0.5, 1]
}
```

learning_rate → The influence of new predictions on the previous predictions
max_depth → The maximum depth of the decision tree
n_estimators → The number of boosting rounds
subsample → The percentage of observations considered
col_sample_bytree → The percentage of predictors
gamma → Parameter utilized to prune the decision tree
lambda/alpha → regularization parameters (penalty terms to reduce overfitting)

5.2 Model Type 1 (No University Rating w/ Weak Normality)

The initial direction for this project was to build 5 individual models to predict the admission chance for a student. Due to the lack of observations in groups such as university ratings 1 & 5, the trained models would lack the ability to predict unseen observations when trained on small sample sizes.

Therefore, to approach this problem, individual unified models were built to explore two key ideas: to identify the impact of the university rating as a predictor on model performance and to determine the most effective model in predicting the admission chance of a student.

Specifically, the first model type aimed to explore the impact of leaving the university rating out of the model and to assess the performance when normality assumptions were not met. The output for the OLS model for model type 1 is shown below:

OLS Regression Results

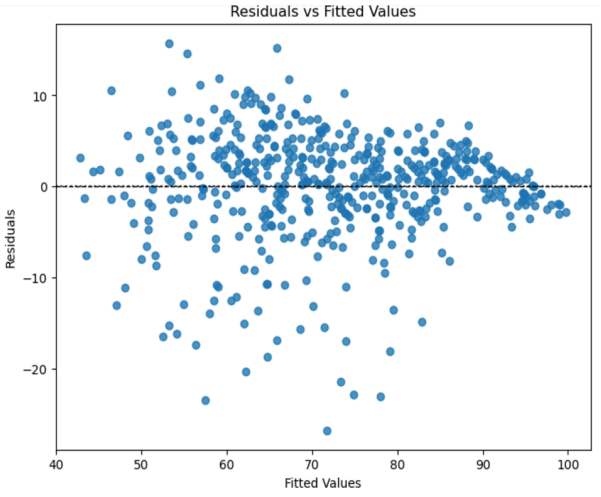
Dep. Variable:	AdmitChance	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	376.9
Date:	Thu, 28 Nov 2024	Prob (F-statistic):	1.37e-180
Time:	15:06:13	Log-Likelihood:	-1602.4
No. Observations:	500	AIC:	3219.
Df Residuals:	493	BIC:	3248.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-131.3127	10.166	-12.917	0.000	-151.287	-111.338
GRE	0.1897	0.050	3.775	0.000	0.091	0.288
TOEFL	0.2898	0.087	3.329	0.001	0.119	0.461
SOP	0.4211	0.425	0.991	0.322	-0.414	1.256
LOR	1.7774	0.410	4.333	0.000	0.971	2.583
CGPA	12.0549	0.962	12.531	0.000	10.165	13.945
Research	2.4879	0.661	3.767	0.000	1.190	3.786

Omnibus:	112.527	Durbin-Watson:	0.789
Prob(Omnibus):	0.000	Jarque-Bera (JB):	260.849
Skew:	-1.158	Prob(JB):	2.28e-57
Kurtosis:	5.675	Cond. No.	1.27e+04

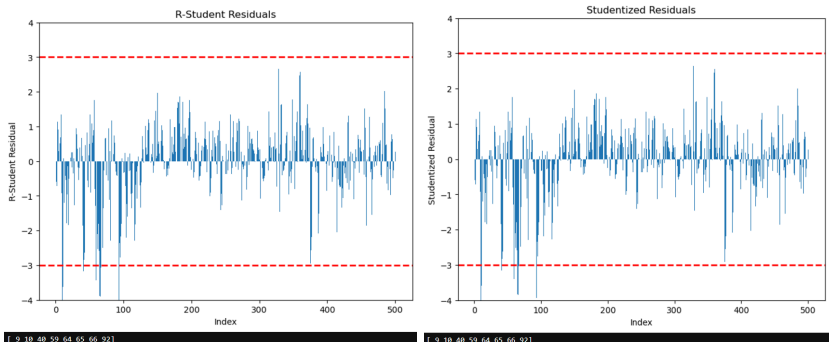
Model: Admission Chance = -131.3127 + 0.1897 * GRE + 0.2898 * TOEFL + 0.4211 * SOP + 1.774 * LOR + 12.0549 * CGPA + 2.4879 * Research

In this model, all of the predictors are significant except for the SOP predictor which has a p-value = 0.322 > α = 0.05. Additionally, the p-value for the F-test is approximately 0 which means that the overall model is significant. The adj. r-squared value is also 0.819 which is a favorable value as 81.9% of the variation in the admission chance can be explained by the predictors in the model. The initial statistics look favorable, but there are many limitations with the following model. This can be seen from the following residual plot.



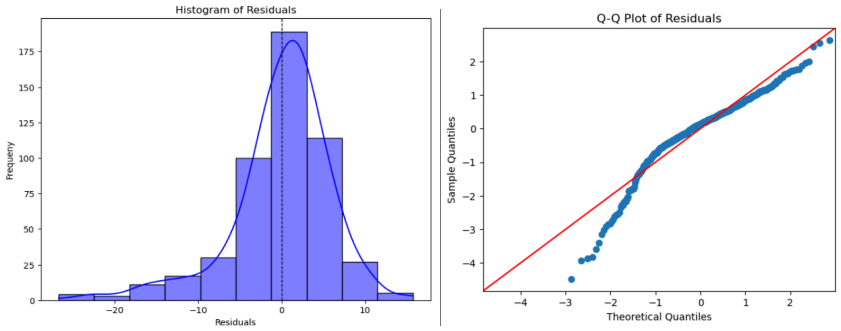
There are concerns about constant variance among the residuals. Specifically, there are significant problems with the residuals that are greater than -15 as they can be possible outliers in the data. Additionally, large fitted values show clustering among the residual = 0 line with smaller variance compared to the other observations.

Furthermore, the R-student residuals and Studentized residuals were observed for the OLS model:



The R-student and Studentized residual plots show possible outliers in the data as there are negative residuals which are beyond the residual = -3 marker on both plots. These points were carefully analyzed for future models to make sure normality assumptions were met.

Here are the following histogram of the residuals and the q-q plot of the residuals:



From the histogram, the distribution of the residuals has a left skew as there are negative residuals that are significantly large in magnitude. These outlier residuals need to be examined further so that normality assumptions are met. Moreover, the q-q plot shows significant curvature near the left tail. This can correspond to the idea that the observations with large negative residuals do not follow a linear pattern and are not suitable for the model.

Here are the following VIF values for each of the predictors in this model:

GRE → 4.453552, TOEFL → 3.874175, SOP → 2.451053, LOR → 1.992803, CGPA → 4.680725, Research → 1.489430

The VIF values show that there is no significant multicollinearity in the model as the VIF values for each of the predictors are less than 10. Although this is the case, the GRE and CGPA can also be discovered further to limit multicollinearity as both predictors have similar VIF values from 4 to 5.

Additionally, a Shapiro-Wilk Test was conducted to assess the normality of the model. The null hypothesis for the Shapiro-Wilk Test states that the residuals of the model are normally distributed. The alternate hypothesis states that the residuals are not normally distributed. In this test, the p-value came out to be approximately 6.30e-15 < a = 0.05. The p-value is significantly small compared to the significance level suggesting that the residuals of the model are not normally distributed.

Finally, the model was evaluated to see its performance in predicting the admission chance on a scale from 0 to 100. The linear regression model was evaluated using k-fold cross-validation with k = 3. The idea of k-fold cross-validation is for every fold to act as the testing set while the other k-1 folds act as the training set. This reduces the variability in assessing model performance on the testing set by considering multiple training/testing splits. The final average RMSE value for the linear regression model was approximately 5.983.

Additionally, for every model type in this project, XGBoost was utilized as well. For the XGBoost model, the average RMSE came out to be 6.074. For model type 1, the OLS model performed better than the XGBoost model.

Note
Interpretation of average RMSE (ex. 5.983): The predicted admission chance for a student deviates from the actual admission chance on average by 5.983 percentage points.

5.3 Model Type 2 (No University Rating w/ Weak Normality)

Model type 2 builds upon the idea of model type 1 by not considering the university rating, but meets the requirements for normality more effectively than model type 1 by data filtering and transformations. Additionally, the statistical procedures explained in model type 1 were replicated for the rest of the model types in this project.

From model 1, the negative residuals with large magnitudes did not follow a linear pattern and contributed to a left skew in the qq-plot and histogram of

residuals. As a result, observations with residuals <= -4.5 and residuals >= 14.5 were filtered from the original data frame to try to achieve normality.

This is the following OLS model after filtering observations in the data:

OLS Regression Results

Dep. Variable:	AdmitChance	R-squared:	0.927
Model:	OLS	Adj. R-squared:	0.926
Method:	Least Squares	F-statistic:	863.7
Date:	Thu, 28 Nov 2024	Prob (F-statistic):	5.25e-228
Time:	15:14:23	Log-Likelihood:	-1081.1
No. Observations:	414	AIC:	2176.
Df Residuals:	407	BIC:	2204.
Df Model:	6		
Covariance Type:	nonrobust		

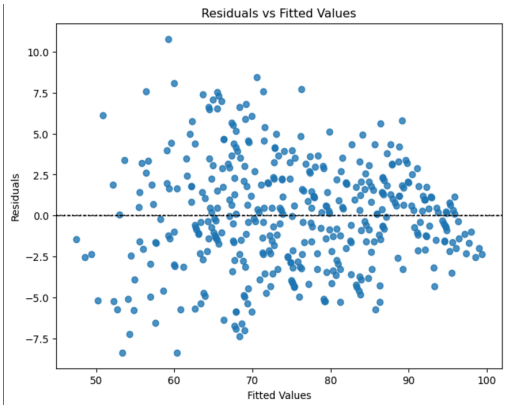
	coef	std err	t	P>|t|	[0.025	0.975]
const	-106.3442	6.420	-16.565	0.000	-118.964	-93.724
GRE	0.1487	0.031	4.818	0.000	0.088	0.209
TOEFL	0.3063	0.053	5.728	0.000	0.201	0.411
SOP	1.0266	0.270	3.807	0.000	0.496	1.557
LOR	0.9217	0.259	3.561	0.000	0.413	1.431
CGPA	10.7714	0.580	18.571	0.000	9.631	11.912
Research	2.8619	0.416	6.877	0.000	2.044	3.680

Omnibus:	1.484	Durbin-Watson:	1.160
Prob(Omnibus):	0.476	Jarque-Bera (JB):	1.559
Skew:	0.139	Prob(JB):	0.459
Kurtosis:	2.886	Cond. No.	1.32e+04

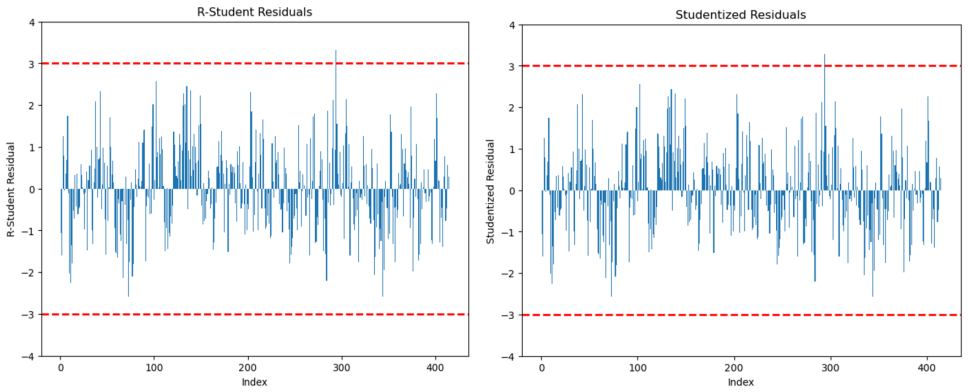
Model: Admission Chance = -106.3442 + 0.1487 * GRE + 0.3063 * TOEFL + 1.0266 * SOP + 0.9217 * LOR + 10.7714 * CGPA + 2.8619 * Research

The F-statistic (863.7) for the OLS model above significantly increased from the OLS model for model type 1 (376.9) demonstrating that the filtered model had a greater significance in predicting the admission chances. Furthermore, all of the predictors were significant as the p-values were all approximately equal to zero. In the OLS model for model type 1, the SOP predictor was not significant as its p-value of 0.322 > a = 0.05.

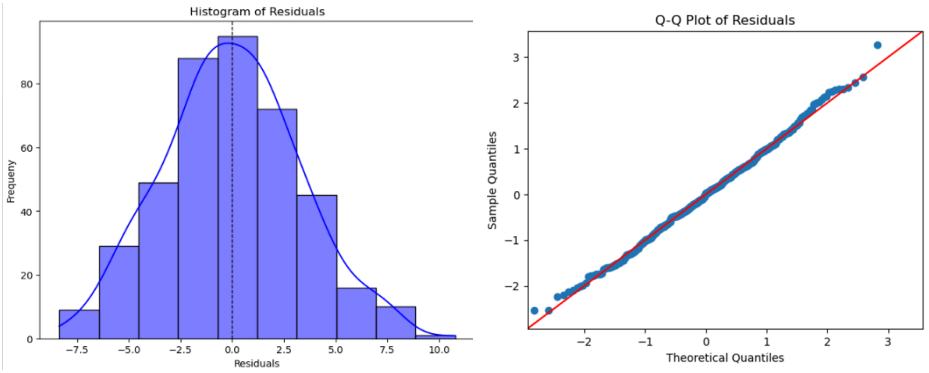
The residual plot below for the OLS model shows a greater random scatter with the variances of the residuals being more constant than the OLS model for model type 1. There are still issues with the variances being non-constant as the variances of the residuals tend to decrease as the fitted values increase. This issue was resolved with the Shapiro-Wilk test which will be discussed later in this section.



Next, the R-Student and studentized residual plots were analyzed for this model. The plots show a significant improvement from the OLS model in model type 1 as there are fewer observations outside of the -3 to 3 residual boundary for both residual plots.



Furthermore, the histogram of the residuals and the q-q plot of the residuals below show significant improvement from the OLS model in model type 1. The left skew in the histogram of residuals is eliminated in the histogram of the residuals in this OLS model. Additionally, the qq-plot of residuals demonstrates less curvature near the left tails as most of the residuals have close to matching theoretical and sample quantiles. The point on the right tail of the q-q plot which is off from the red line needs further discovery as this observation was removed for further study, but its influence on the model was replaced by a new observation with a similar residual. This suggests that the observation may be an influential point in the model.



The following VIF values were calculated for the OLS model for model type 2:

GRE → 4.515279, TOEFL → 3.939989, SOP → 2.693533, LOR → 1.993461, CGPA → 4.618952, Research → 1.581576

The VIF values did not significantly change in the OLS model for model type 2 compared to the OLS model for model type 1 as the VIF values were all less than 10, but there may be minor multicollinearity for the GRE/CGPA predictors. It was reasonable to proceed with this model as the VIF values were not greater than or equal to 10.

Similarly to the OLS model for model type 1, the Shapiro-Wilk Test for Normality was conducted to assess the normality of the residuals. The resulting p-value for this test when considering the OLS model for model type 2 was approximately 0.576 > a = 0.05. Consequently, the null hypothesis was rejected and suggests that the residuals are normally distributed in this model.

Finally, model type 2 was evaluated using k-fold cross-validation with k = 3 for the OLS model. The average RMSE for the OLS model was approximately 3.389 and the average RMSE for the XGBoost model was approximately 3.373. For model 2, the XGBoost model performed slightly better than the OLS model. Additionally, the average RMSE for both the OLS and XGBoost models was approximately cut in half compared to the models in model type 1 after improving normality conditions.

5.4 Model Type 3 (All Predictors w/ Normality)

In model 3, the university rating was included in the model to view its impact on model performance. Additionally, normality assumptions were reasonably met for this model.

OLS Regression Results

Dep. Variable:	AdmitChance	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	786.1
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	1.27e-234
Time:	16:23:31	Log-Likelihood:	-1101.6
No. Observations:	422	AIC:	2219.
Df Residuals:	414	BIC:	2252.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-103.3151	6.611	-15.628	0.000	-116.310	-90.320
GRE	0.1311	0.031	4.234	0.000	0.070	0.192
TOEFL	0.3158	0.053	5.995	0.000	0.212	0.419
SOP	0.5775	0.290	1.988	0.047	0.006	1.148
LOR	1.1028	0.250	4.409	0.000	0.611	1.594
CGPA	10.8198	0.581	18.625	0.000	9.678	11.962
Research	2.6387	0.415	6.351	0.000	1.822	3.455
UnivRtg	0.6310	0.242	2.607	0.009	0.155	1.107

Omnibus:	1.670	Durbin-Watson:	1.223
Prob(Omnibus):	0.434	Jarque-Bera (JB):	1.743
Skew:	0.123	Prob(JB):	0.418
Kurtosis:	2.803	Cond. No.	1.37e+04

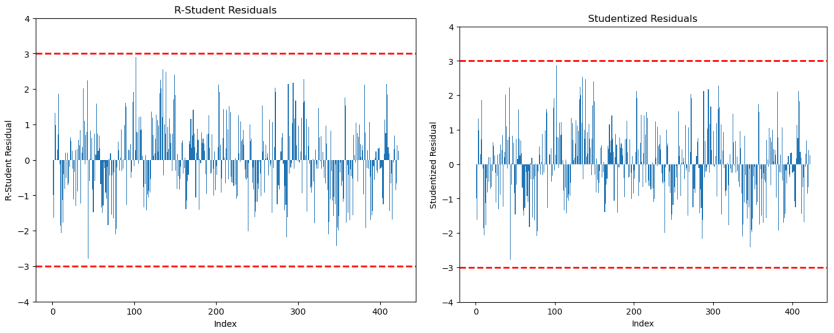
Model: Admission Chance = -103.3151 + 0.1311 * GRE + 0.3158 * TOEFL + 0.5775 * SOP + 1.1028 * LOR + 10.8198 * CGPA + 2.6387 * Research + 0.6310 * UnivRtg

The model above shows a large F-statistic (786.1) and a p-value of approximately zero for the F-test demonstrating that the model is significant in predicting the admission chance. The individual predictors were also all significant as their respective p-values were less than a = 0.05. The introduction of the university rating variable reduced the impact of variables such as SOP and TOEFL as their respective coefficients decreased compared to the OLS model for model type 2. The adjusted r-squared value is 0.929 which is favorable as 92.9% of the variation in the admission chance can be explained by the predictors in the model.

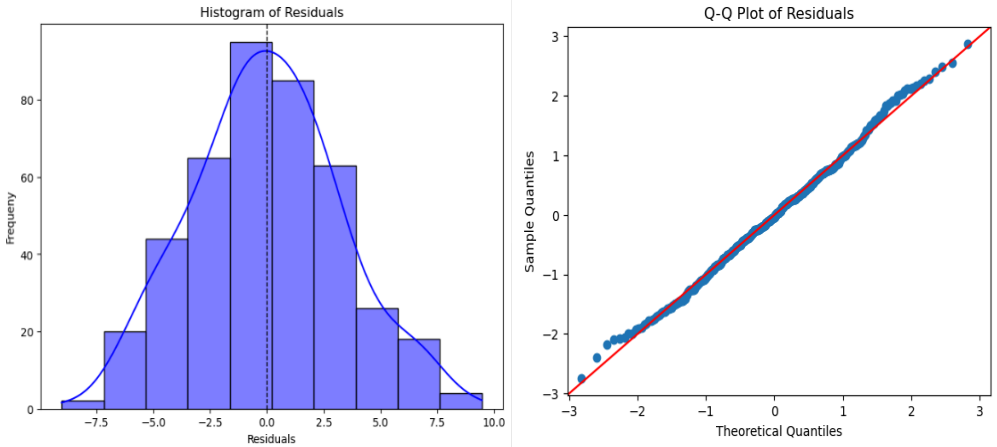
The updated residual plot for the OLS model when considering the university rating is displayed below:

The residual plot shows a random scatter in the residuals, but the residuals may violate the condition of constant variance. This is because the variance of the residuals decreases as the fitted values increase. Transformations such as box-cox were utilized with lambda = 2, but the filtering of observations suggested normality in the residuals with the Shapiro-Wilks test. The p-value for the Shapiro-Wilks test was approximately 0.4161 > a = 0.05 suggesting that the residuals follow normality assumptions in this model.

Moreover, the R-Student and Studentized residual plots show no observations outside of the -3 to 3 residual boundary for both plots. Therefore, the possibility of outliers was not considered for this model.



Adding on, the histogram of the residuals for the OLS model for model type 3 shows a normal distribution with no significant skew. Most of the residuals fall in the range from -2.5 to 2.5 and the frequency of the residuals decrease as the extremes are approached. The q-q plot of residuals also shows no curvature except for slight curvature in the tails. The left and right tails of the q-q plot have a few residuals that do not exhibit a linear pattern.



The following VIF values were calculated for the OLS model for model type 3:

GRE → 4.4616171, TOEFL → 4.011676, SOP → 3.212812, LOR → 2.032472, CGPA → 4.734057, Research → 1.606039, UnivRtg → 3.020534

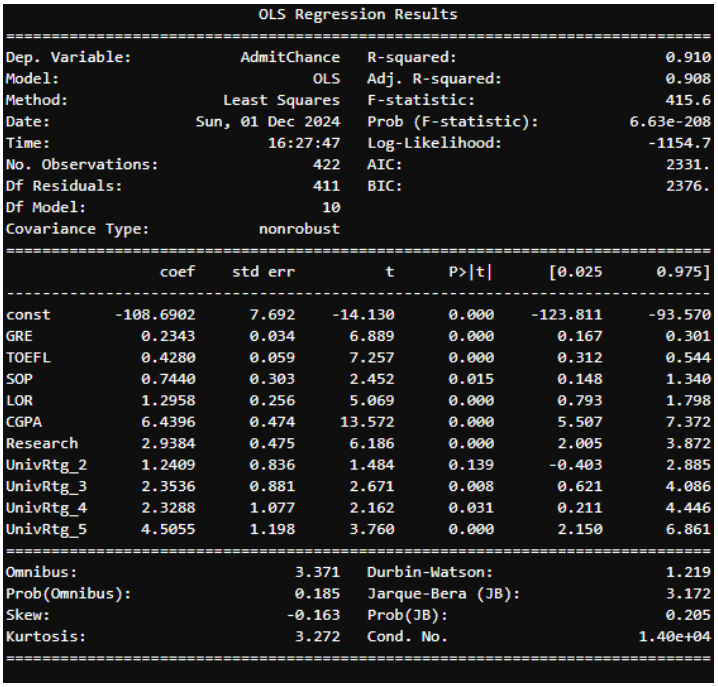
The inclusion of the university rating led to an increase in the VIF of the TOEFL demonstrating that there may be multicollinearity between the TOEFL and UnivRtg predictors. For the most part, all predictors maintain VIFs less than or equal to 10, so there are no significant multicollinear issues in this model.

Finally, the OLS model for model type 3 was also evaluated with k-fold cross-validation with k = 3. The average RMSE for the OLS model was approximately 3.354 while the average RMSE for the XGBoost model was approximately 3.544. This shows that the OLS model performed better than the XGBoost model for model type 3. The performance of model type 3 with respect to the average RMSE is slightly better than that of model type 2 and significantly better than model type 1.

5.5 Model Type 4 (Dummy Encoding w/ University Rating)

The final model aims to determine the impact of every individual university rating on a student's admission chance. Therefore, dummy encoding was utilized so that every individual rating appeared as a predictor in the model.

The following dummy encoding model was built with the filtered data set as the filtered data followed reasonable normality assumptions. The university rating of 1 was dropped as a predictor to reduce the multicollinearity in the model. This is because the university rating of 1 can be easily represented in the model when university ratings from 2-5 all have boolean values of zero.



OLS Regression Results

Dep. Variable:	AdmitChance	R-squared:	0.910
Model:	OLS	Adj. R-squared:	0.908
Method:	Least Squares	F-statistic:	415.6
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	6.63e-208
Time:	16:27:47	Log-Likelihood:	-1154.7
No. Observations:	422	AIC:	2331.
Df Residuals:	411	BIC:	2376.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-108.6902	7.692	-14.130	0.000	-123.811	-93.570
GRE	0.2343	0.034	6.889	0.000	0.167	0.301
TOEFL	0.4280	0.059	7.257	0.000	0.312	0.544
SOP	0.7440	0.303	2.452	0.015	0.148	1.340
LOR	1.2958	0.256	5.069	0.000	0.793	1.798
CGPA	6.4396	0.474	13.572	0.000	5.507	7.372
Research	2.9384	0.475	6.186	0.000	2.005	3.872
UnivRtg_2	1.2409	0.836	1.484	0.139	-0.403	2.885
UnivRtg_3	2.3536	0.881	2.671	0.008	0.621	4.086
UnivRtg_4	2.3288	1.077	2.162	0.031	0.211	4.446
UnivRtg_5	4.5055	1.198	3.760	0.000	2.150	6.861

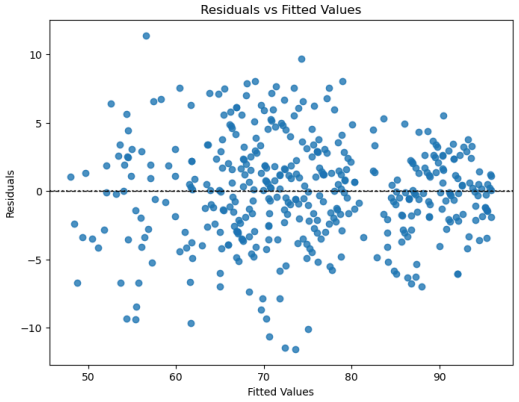
Omnibus:	3.371	Durbin-Watson:	1.219
Prob(Omnibus):	0.185	Jarque-Bera (JB):	3.172
Skew:	-0.163	Prob(JB):	0.205
Kurtosis:	3.272	Cond. No.	1.40e+04

Model: Admission Chance = -108.6902 + 0.2343* GRE + 0.4280 * TOEFL + 0.7440 * SOP + 1.2958* LOR + 6.4396 * CGPA + 2.9384 * Research + 1.2409 * UnivRtg_2 + 2.3536 * UnivRtg_3 + 2.3288 * UnivRtg_4 + 4.5055 * UnivRtg_5

The F-statistic (415.6) from the model above is less than the F-statistic from the OLS model for model type 2 and the OLS model for model type 3. The p-value for the F-test is still approximately zero showing that the model is significant in predicting the admission chance. The adjusted r-squared value is 0.908 which is great for this model as 90.8% of the variation in the admission chance can be explained by the predictors in the model. However, there are a few predictors that are insignificant or close to insignificant in the model. The UnivRtg_2 predictor has a p-value of 0.139 < a = 0.05 which suggests that it is insignificant in predicting the admission chance. Additionally, UnivRtg_4 has a p-value of 0.031 which is close to 0.05 suggesting that it is significant in predicting the admission chance, but not as significant as the other predictors in the model. In contrast, the UnivRtg_3 and UnivRtg_5 predictors have p-values of 0.008 and 0 showing that they are significant predictors in the model.

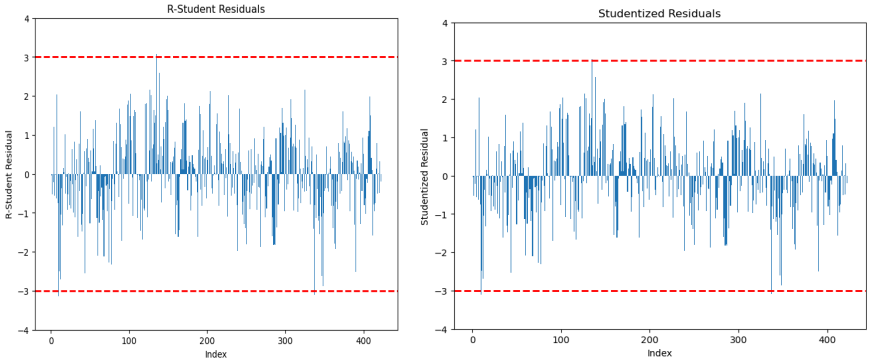
It is important to also note that the significance of CGPA remains the same throughout all of the models in this project. The test statistic for CGPA remains very large with a very small p-value showing that it is significant for predicting the admission chance. Additionally, its coefficient of 6.4396 is higher than the other predictors in this model. Moreover, the UnivRtg_5 predictor had the second highest coefficient of 4.5055 showing that it is also influential in predicting the admission chance.

To check for normality, the residual plot was analyzed next for this model:



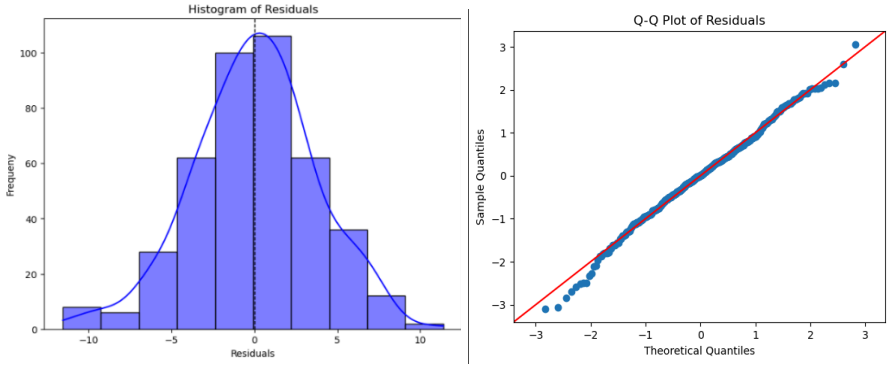
The residual plot above shows random scatter among the residuals and the variance of the residuals looks more constant than the previous model types. As the fitted values increase, there is a slight bow shape for the variance of the residuals, so the Shapiro-Wilk test was utilized to test further for normality. The Shapiro-Wilk test resulted in a p-value of approximately 0.1856 > a = 0.05 suggesting that the residuals are normally distributed in the model.

The following R-Student and standardized residual plots were also observed to detect outliers in the model:



The plots above show possible outliers for negative residuals in the model near the index of 0 and the indices from 300 to 400. This demonstrates that these observations may need to be filtered to have a model with a better fit.

The following q-q plot of residuals demonstrates the impact of the negative residuals on the model as there is some curvature near the left tail of the plot. Additionally, there are a few positive residuals that acted as influential points similar to the OLS model in model type 2. The histogram of the residuals shows an approximately normal distribution with little to no skew in the tails.



The following VIF values were calculated for model type 4:

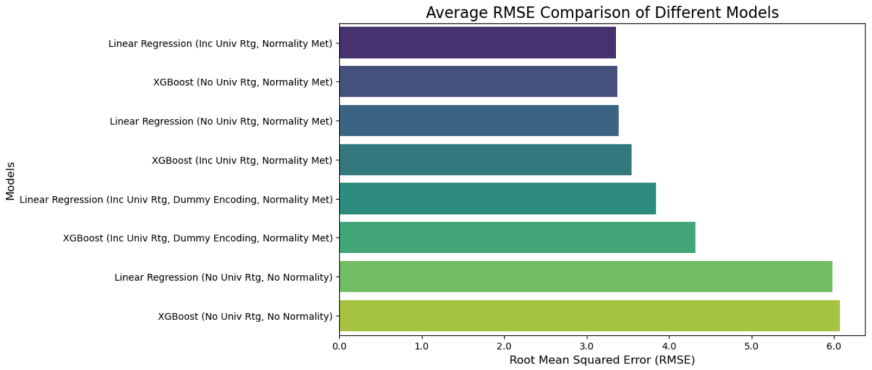
GRE → 4.298212, TOEFL → 3.880667, SOP → 3.019059, LOR → 1.773816, CGPA → 2.921926, Research → 1.620117, UnivRtg_2 → 3.624898, UnivRtg_3 → 4.983112, UnivRtg_4 → 5.834229, UnivRtg_5 → 5.857449

The VIFs for the dummy encoded predictors are less than 10 in the OLS model, but they still show signs of multicollinearity. This can be explained by the similarity of the dummy encodings for UnivRtg_2 and UnivRtg_3 as well as for UnivRtg_4 and UnivRtg_5. The clusters for each of these encodings can be highly correlated as students with university ratings that differ by at most one level may maintain similar academic profiles. Since all of the VIFs were less than 10 in this model, no predictors were removed, but the removal of UnivRtg_2 and UnivRtg_4 in this model type would be a future study to pursue.

Finally, model type 4 was evaluated using OLS and XGBoost models. The average RMSE for the OLS model was approximately 3.836 while the average RMSE for the XGBoost model was approximately 4.313. For model type 4, the OLS model performed better than the XGBoost model and the average RMSE values were similar to that of model types 3 and 4.

5.6 Model Performance Analysis

Average RMSE for All Models
OLS w/ Univ Rtg, Normality Met → 3.3543368234318804
XGBoost w/ No Univ Rtg, Normality Met → 3.3730780708668706
OLS w/ No Univ Rtg, Normality Met → 3.3892834659259394
XGBoost w/ Univ Rtg, Normality Met → 3.5437081003179087
OLS w/ Univ Rtg, Dummy Encoding, Normality Met → 3.836044137737936
XGBoost w/ Univ Rtg, Dummy Encoding, Normality Met → 4.313524224571041
OLS w/ No Univ Rtg, No Normality → 5.98312702798502
XGBoost w/ No Univ Rtg, No Normality → 6.074416563337169



The bar plot above demonstrates the average RMSE of all of the models built in this project from the smallest average RMSE (top) to the largest average RMSE (bottom). The chart shows that the linear regression model with the university rating included and normality assumptions met was the best-performing model with an average RMSE of approximately 3.354. The XGBoost model which included no university rating and normality assumptions was the second-best-performing model with an average RMSE of approximately 3.373.

There are different use cases for both OLS and XGBoost. In this project, linear regression performed better overall because the data set had a small number of observations with a small number of predictors. On the other hand, there were also problems with constant variance of the residuals in the various models. In cases of non-linearity, XGBoost is the better model to select as it works well with considering non-linear patterns and curvature in data points.

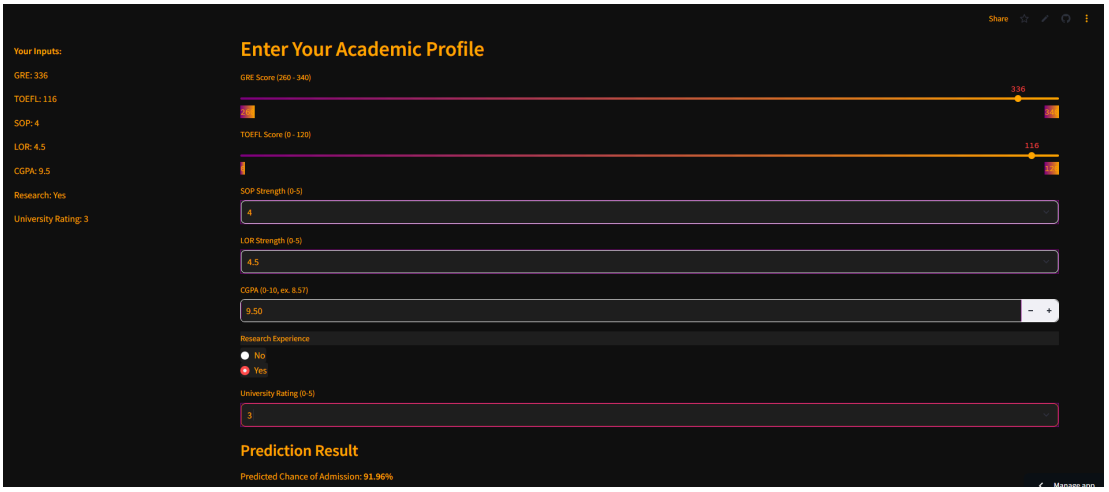
For simplicity, the current front-end utilizes the linear regression model as it is easy to work with and interpret in the context of predicting admission chances. To make the predicting power more robust in the future, XGBoost is the better model due to its flexibility in identifying non-linear patterns and being effective for large datasets with many predictors.

6. Front End Development

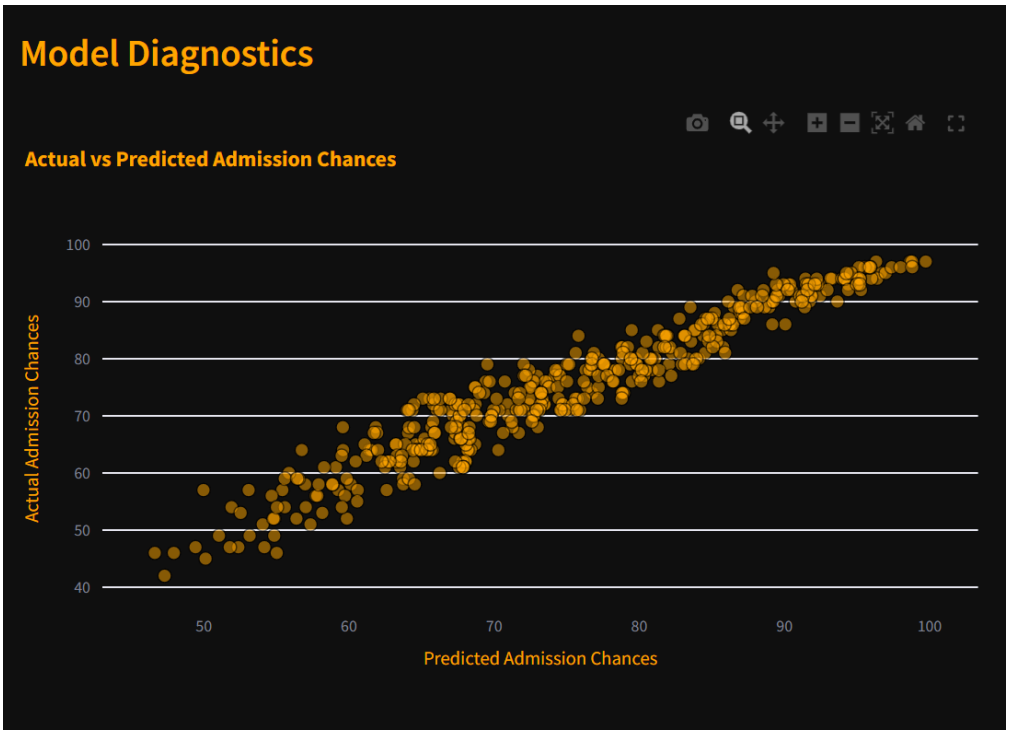
Once the most effective model was determined after the predictive modeling process, the final goal was to deploy this model so that a student could input their academic profile and receive a percentage for their chance of admission to graduate school. The predictive modeling process was built in Python making it easy to deploy the framework to a front end via Streamlit.

The front end allows students to enter their GRE score, TOEFL score, SOP strength, LOR strength, CGPA score, research experience, and university rating. After the user enters their academic profile, the predicted chance of admission is displayed for the

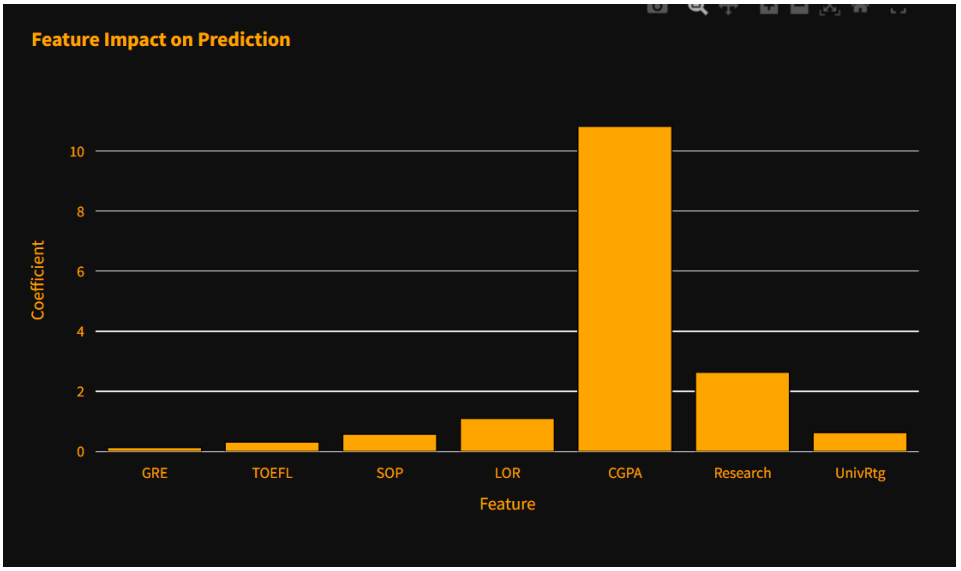
student. The student can also view their inputs on the sidebar to the left of the page.



At the bottom of the page, current students are allowed to view the actual vs predicted admission chances of past students. This allows students to see how effectively the model predicted the graduate admission chance for past students. In the future, this plot will become more personalized as students will have the ability to only view actual vs predicted admission chances for previous students with similar academic profiles.



Students also have the ability to view the influence of every individual predictor on the admission chance. This allows students to fine-tune their inputs to see how they can improve their admission chances to graduate school.



7. Future Plans and Improvements

To enhance the scope and robustness of the graduate admissions analysis, several future improvements have been identified. First, we aim to collect more academic profiles from a diverse range of students to increase the dataset size and make the model more generalizable to all university students, not just those who are academically strong. This broader representation will improve the model's applicability across varying academic performances.

Second, stronger residual analysis will be performed by incorporating advanced metrics such as Cook's Distance and DFBETAs. This will complement the current practice of filtering residuals for normality and provide deeper insights into the influence of outliers and leverage points.

Additionally, the project will explore other contextual factors that may influence graduate admissions, such as work experience, extracurricular activities, and other non-academic indicators. By integrating these factors, we can offer a more holistic view of what contributes to admission success.

Lastly, a significant goal is to provide students with personalized recommendations to improve their chances of graduate admission. By leveraging insights from the analysis, students can be guided on areas to strengthen, ensuring a more targeted approach to their application process.

8. Conclusion

The Graduate Admission Analysis project highlights the critical role of academic metrics and institutional factors in predicting graduate admissions. The study confirmed that variables like SOP scores, TOEFL scores, CGPA, and research experience are significant predictors, with strong correlations to the chance of admission. The inclusion of university ratings further refined prediction accuracy, emphasizing the importance of institutional reputation.

Advanced modeling techniques, including linear regression and XGBoost with normality assumptions, demonstrated varying performance levels. XGBoost models consistently outperformed others, achieving lower RMSE values, indicating enhanced predictive reliability. However, the study's limitations, such as reliance on academically strong students and omission of contextual factors like work experience or extracurricular activities, suggest areas for improvement.

Future efforts aim to broaden the dataset scope, incorporate additional contextual variables, and provide actionable recommendations to students, making the model more generalizable and robust. These improvements will further align the analysis with the real-world complexities of graduate admissions, enhancing its practical utility.

9. Appendix

References:

Manral, Mukesh. “Graduates Admission Prediction.” *Kaggle*, 12 June 2022, www.kaggle.com/datasets/mukeshmanral/graduates-admission-prediction.

Team Member Contribution:

Sarvesh Gopalakrishnan → Introductory Statistical Testing, Predictive Modeling, Front End Development
Shriram Rajasekar → Introduction, Data Cleaning, Challenges, and Limitations, Exploratory Data Analysis, Future Plans and Improvements, Conclusion

Python Code: (Sarvesh)

Back-End:

Code:
https://github.com/Sarvesh30/GraduatePredictor/blob/main/GraduatePredFinal.ipynb

Front-End:

Code: https://github.com/Sarvesh30/GraduatePredictor/blob/main/app.py
App: https://graduatepredictorstat4355.streamlit.app/

R Code: (Shriram)

```
#Libraries used for Project Data
library(plotrix)
library(ggplot2)
library(reshape2)
library(car)
library(multcomp)
library(mosaic)
library(lmtest)
library(tidyr)

#Shriram project data
#Multiple Linear Regression Model from Students data
StudentsData<-read.csv("Admission_Predict_V11.csv",header = TRUE)
StudentsData<-StudentsData[,c("Chance.of.Admit","GRE.Score","TOEFL.Score","Universit
y.Rating","SOP","LOR","CGPA","Research","Serial.No.")]
summary(StudentsData)

#Attaching Dataset and number of rows and cols
attach(StudentsData)
dim(StudentsData)

#Check name and structure of the dataset
names(StudentsData)
str(StudentsData)

#Exploratory Data Analysis

#Printing piechart of distribution of Universities

print(table(University.Rating))
print(table(Research))

data_U=c(34,126,162,105,73)
```

```
data_R=c(220,280)
labels_U=c("U1", "U2", "U3", "U4","U5")
labels_R=c("No Experience","Experience")

# Calculate percentages
percent_U=round(100*data_U / sum(data_U),1)
percent_R=round(100*data_R / sum(data_R),1)

#Create 3D pie chart
pie(data_U,labels=paste(percent_U,sep=" ","%"),col=rainbow(length(data_U)),main
="Distribution of University")
# Add legend for University
legend("topright",
       title = "Categories",
       legend = labels_U,
       fill = rainbow(length(data_U)),
       border = "black",
)

pie(data_R,labels=paste(percent_R,sep=" ","%"),col=rainbow(length(data_R)),main
="Distribution of Experience among Students")
# Add legend for Research Experience
legend("topright",
       title = "Categories",
       legend = labels_R,
       fill = rainbow(length(data_R)),
       border = "black",
)

#Create Histogram
hist(CGPA,main="Distribution of CGPA",col="green",xlab="CGPA",ylab="Count")
hist(SOP,main="Distribution of SOP",col="blue",xlab="SOP",ylab="Count")
hist(LOR,main="Distribution of LOR",col="yellow",xlab="SOP",ylab="Count")
hist(Chance.of.Admit,main="Distribution of Chance.Of.Admit",col="orange",xlab="Chance
of Admit",ylab="Count")

# Pivot the data longer
df_long=pivot_longer(StudentsData, cols = c(GRE.Score, TOEFL.Score), names_to =
"Exam", values_to = "Score")

# Create side-by-side histograms
ggplot(df_long, aes(x = Score, fill = Exam)) +
  geom_histogram(position = "dodge", bins = 30, color = "black") +
  labs(title = "Distribution of GRE and TOEFL Scores", x = "Score", y = "Frequency") +
```

```
  theme_classic()

# Create another version of side-by-side Histogram
StudentsData %>%
  pivot_longer(c(GRE.Score,TOEFL.Score)) %>%
  ggplot(aes(value, fill=name))+
  geom_histogram(position = "dodge")

#Create boxplot
boxplot(Chance.of.Admit~University.Rating,main="Chance of Admit-University
wise",col="green")
boxplot(CGPA~University.Rating,main="CGPA-University wise",col="orange")

#Multiple Regression Model
#start by creating blank model and then decide predictors which are significant
StudentsData.lm<-lm(Chance.of.Admit~1,data=StudentsData)
add1(StudentsData.lm,StudentsData,test='F')

StudentsData.lm=lm(Chance.of.Admit~GRE.Score+TOEFL.Score,data=StudentsData)
summary(StudentsData.lm)
anova(StudentsData.lm)

add1(StudentsData.lm,StudentsData,test='F')
StudentsData.lm=lm(Chance.of.Admit~GRE.Score+TOEFL.Score+University.Rating+SOP+
LOR+CGPA+Research,data=StudentsData)
summary(StudentsData.lm)
anova(StudentsData.lm)

#Final Regression Model
add1(StudentsData.lm,StudentsData,test='F')
StudentsData.lm=lm(Chance.of.Admit~GRE.Score+TOEFL.Score+LOR+CGPA+Research,d
ata=StudentsData)
summary(StudentsData.lm)
anova(StudentsData.lm)

#Residual Plot without BoxCox Transformation
plot(fitted(StudentsData.lm), resid(StudentsData.lm), col = "dodgerblue",
    pch = 20, cex = 1.5, xlab = "Fitted", ylab = "Residuals")
abline(h = 0, lty = 2, col = "darkorange", lwd = 2)

#BoxCox Transformation Model
boxcox(StudentsData.lm, plotit = TRUE, lambda = seq(0.5, 2.5, by = 0.1))
#lamda=1.9
```

```
#Final BoxCox Transformation Model
StudentsData.cox=lm((((Chance.of.Admit^2.2)-1) /
2.2)~GRE.Score+TOEFL.Score+LOR+CGPA+Research)
summary(StudentsData.cox)

#ANOVA model
anova(StudentsData.lm)

#studentized Breusch-Pagan
bptest(StudentsData.lm)

#ANOVA of BoxCox Model

anova(StudentsData.cox)

#studentized Breusch-Pagan and Shapiro-Wilk normality test after Transformation
bptest(StudentsData.cox)


plot(fitted(StudentsData.cox), resid(StudentsData.cox), col = "dodgerblue",
    pch = 20, cex = 1.5, xlab = "Fitted", ylab = "Residuals")
abline(h = 0, lty = 2, col = "darkorange", lwd = 2)

#Multicollinearity checking
#VIF > 5 or 10 indicates high multicollinearity
#Tolerance < 0.1 indicates high multicollinearity
#No Multicollinearity exist between predictors

vif(StudentsData.lm)
1 / vif(StudentsData.lm)


#Residual plots for Valid Multiple Linear Regression after BoxCox Transformation
#Residual plot is showing constant variance and doesn't show any pattern and therefore
#proves validity of Multiple LinearRegression

ggplot(StudentsData.cox, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='Residual vs. Fitted Values Plot', x='Fitted Values', y='Residuals')

# Create scatter plots

df_melt <- melt(StudentsData, id.vars = "Chance.of.Admit")
```

```
ggplot(df_melt, aes(x = value, y = Chance.of.Admit, color = variable)) +
  geom_point() +
  labs(x = "Predictor Values", y = "Chance.of.Admit") +
  theme(legend.title = element_blank()) +
  facet_wrap(~ variable, scales = "free_x")

#95% confidence interval and prediction interval
confint(StudentsData.lm,conf.level=0.95)
predict(StudentsData.lm,interval = "prediction", level = 0.95)
```