# Predicting Graduat Student Admission Success using Multiple Linear Regression Model

## Introduction

In the competitive realm of graduate school admissions, prospective students are often eager to assess their likelihood of acceptance. This project sets out to address this very question by creating a predictive model that estimates the chances of admission into a graduate program based on an applicant's undergraduate academic performance and qualifications. To achieve this goal, the dataset selected for this project is the Graduate Admissions 2 dataset, available on Kaggle and inspired by the UCLA graduate admissions dataset. This dataset encompasses various crucial parameters in the admissions process, including GRE scores, undergraduate GPA, research experience, and more.

Motivated by a personal interest in pursuing graduate studies, the choice of this dataset stems from the desire to delve into how various factors influence the likelihood of admission. The project kicks off with an exploratory data analysis, aimed at uncovering intriguing patterns and elucidating the relationships among different variables. This initial examination paves the way for the construction of a regression model.

Regarding the dataset itself, it was initially relatively clean, requiring some preprocessing steps. This involved creating new variables based on existing ones and eliminating unnecessary variables for the analysis. Numeric variables were also transformed into factors, with assigned levels for enhanced readability before visualization. For instance, the binary research experience column (0 or 1) could be converted into True/False or Yes/No as desired.

The dataset comprises a range of variables, including GRE scores, TOEFL scores, university rating, statement of purpose and letter of recommendation strength, undergraduate GPA, research experience, and the critical metric of "Chance of Admit," which serves as the target variable for prediction. With this array of information, this project endeavors to shed light on the nuanced

interplay of these factors and their impact on the graduate school admission process.

## Literature Review

### Success Definition:

Student success is a crucial component of higher education institutions because it is considered as an essential criterion for assessing the quality of educational institutions (National Commission for Academic Accreditation & Assessment Standards for Quality Assurance and Accreditation of Higher Education Institutions, 2015, 3).While this is a multi-dimensional definition, authors in (T et al., 2015, 5) gave an amended definition concentrating on the most important six components, that is to say *“Academic achievement, satisfaction, acquisition of skills and competencies, persistence, attainment of learning objectives, and career success”* .(Fig 1)

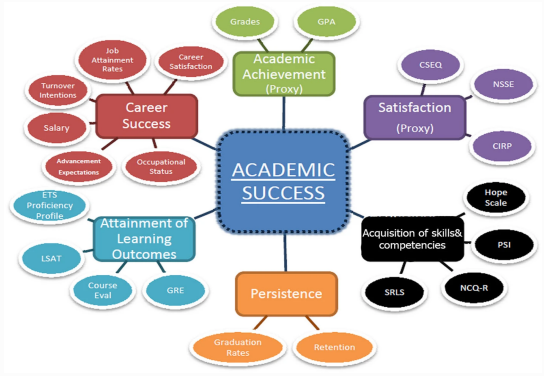


Fig 1: Defining Academic success and its measurement (T et al., 2015, 5)

In spite of recommendations advocating a more comprehensive conceptualization of the term, a significant portion of existing research primarily gauges academic success through a narrow lens, predominantly equating it with academic achievement. This academic achievement is primarily evaluated based on Grade Point Average (GPA) or Cumulative Grade Point Average (CGPA) . Another facet of academic success revolves around students' persistence, often referred to as academic resilience , which, in a circular manner, is also predominantly appraised through the use of grades and GPA – the most widespread and accessible metrics within educational institutions.

### Influential factor predicting academic Success:

A pivotal aspect in the context of predicting academic success among higher education students pertains to the unambiguous definition of what constitutes academic success. Once this fundamental premise is established, the focus shifts to identifying the potential influential factors, which, in turn, dictate the nature of data collection and mining efforts. Remarkably, two prominent factors, specifically prior-academic achievement and student demographics, feature prominently in a substantial 69% of the research studies examined. This alignment with the findings of prior literature reviews underscores the widespread use of internal assessment grades and Cumulative Grade Point Average (CGPA) as prevalent factors in Educational Data Mining (EDM) studies (Sahiri & Hossain, 2015, 414-422). Of particular significance, accounting for over 40% of the influence, is the factor of prior academic achievement, effectively representing students' academic history. This is frequently denoted by the grades or

analogous academic performance indicators that students have garnered in their preceding educational experiences, encompassing both pre-university and university data.

## Dataset Selection:

Before starting any of the projects everyone should select their dataset.It matters from where anyone gets their dataset.The attributes of the dataset also matters to what kind of model they are using for those attiributes.As i discussed above influential factors of predicting academic success, on following that in this review i only discussed with grad admission where some factors are most important for admission confirmation.There are two dataset most paper used which almost similar attributes like on UCLA dataset,which is collected from kaggle.This dataset contains parameters that are considered carefully by the admissions committee. First section contains scores including GRE, TOEFL and Undergraduate GPA. Statement of Purpose and Letter of Recommendation are two other important entities. Research Experience is highlighted in binary form. All the parameters are normalized before training to ensure that values lie between the specified range. A few profiles in the dataset contain values that have been previously obtained by students. A unique feature of this dataset is that it contains an equal number of categorical and numerical features. The data has been collected and prepared typically from an Indian student's perspective. However, it can also be used by other grading systems with minor modifications (Acharya, 2019).

The data of use (Bitar & Almauza, 2020) consisted of five-hundred instances with no null value entries nor any categorical attributes; each instance in the dataset represented an applicant. This dataset has been acquired from UCLA's admittance history data. The number of attributes given in dataset is eight where all attributes are numeric:

I. **GRE Score** (General Record Examinations); this score measures general knowledge in undergrad Math and English. This score ranges from a value of 260 to 340

II. **TOEFL Score** (Test of English as a Foreign Language); this score measures students' English abilities. This score ranges from a value 0 to 120.

III. **SOP** (Statement of Purpose); a letter written by the applicant explaining their purpose of the application. This is scored on a range from one to five.

IV. **LOR** (Letter of Recommendation); tests the weight of the recommendation provided by the applicant. This is scored on a range from one to five.

V. **CGPA** (Cumulative GPA); based on the academic performance of the applicant in undergraduate studies. This is scored on a range from one to ten.
**University Rating**; based on the reputation of the applicant's

previous university. This is scored on a range from one to five.

VI. **Research Experience**; binary value based on whether the applicant has any research familiarity. This value is either one or zero.

VII. **Chance of Admission**; the rate of admission into graduate school. This attribute is the targeted value which will be predicted as the rate from zero to one.

## Model Selection:

There we discussed several papers where some of them used different regression models and some used machine learning models. On the other hand in some papers they used both regression and machine learning models.

## Algorithms:

Predictive modeling in Machine Learning encompasses a diverse array of algorithms. This section provides an overview of the algorithms employed in crafting our predictive models, including Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), Random Forest Regression (RFR), the Logistic Regression Model (LRM) and the Naive Bayes(NB).

**Linear Regression (LR)** assumes paramount significance in the realm of Machine Learning, especially in the domain of supervised learning. It offers a means to establish a connection between a dependent variable and one or more independent variables by identifying a straight regression line that best encapsulates the relationship.

**Decision Tree (DT)** takes precedence as one of the most prevalent techniques for classification and prediction. Manifesting as a tree structure, each internal node with outgoing edges signifies a condition based on an attribute, while each branch represents an outcome of the test. Terminal nodes, or leaf nodes, are the ultimate bearers of class labels.

**Support Vector Regression (SVR)**, a well-known Machine Learning technique, serves in both classification and regression. It closely resembles Linear Regression with slight variations. SVR facilitates the definition of permissible error within the predictive model, discerning an optimal line to accommodate the data.

**Random Forest Regression (RFR)** is an ensemble learning method that assembles a multitude of decision trees during the training phase. It leverages the collective predictions of individual trees to enhance the overall predictive accuracy.

**Logistic Regression Model (LRM)** originating from the field of statistics, is instrumental in binary classification problems, typically those encompassing two distinct class values. It is extensively utilized to estimate the probability of a data sample belonging to a specific class. The logistic regression formula is encapsulated as:

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x)).

**Naive Bayes(NB):** is a classification technique widely used in tasks like text categorization and spam detection. It's based on Bayes' theorem and assumes feature independence. This model estimates class probabilities by considering the likelihood of features given a class, making it valuable for real-world applications.

Each of these algorithms plays a pivotal role in shaping predictive models, and their distinctions offer versatility in tackling a spectrum of real-world scenarios.

## Evaluation Method:

Model evaluation is a fundamental aspect of constructing robust Machine Learning models, and various evaluation methods are available. In the subsequent section, we will delve into the discussion of three primary evaluation metrics we will employ, which are R-squared (R2), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

**R-Squared**, often referred to as the coefficient of determination, is a pivotal indicator for assessing the quality of a simple linear regression model. It quantifies the goodness of fit between the model and the observed data, essentially gauging how effectively the regression equation describes the data point distribution. In simpler terms, when R2 is close to 0, it indicates that the data points are widely dispersed around the regression line. Conversely, an R2 value nearing 1 signifies that the data points tightly cluster around the regression line. The elusive perfect alignment of data points on the regression line results in R2 equaling 1.

**Mean Square Error (MSE)** calculates the mean of the squared differences between model predictions and actual observations. It is the primary target for minimization in single or multiple regression scenarios. The methodology hinges on the premise that the average of prediction residuals is typically non-zero, but the mean of their squares yields a valuable metric.

**Root Mean Square Error (RMSE)** is a standard measure to quantify model evaluation errors. It is essentially the square root of the mean of the squared errors. The RMSE metric provides a concise summary of the magnitude of errors in the model evaluation process.

## Work Process:

A novel approach to address the uncertainty students face regarding their university admission prospects (Fatiya & Sadath, 2021) utilizes logistic regression and machine learning to create an admissions predictor, aiding students in evaluating their competitiveness for different universities. This study acknowledges the paramount importance of student admissions in educational institutions, with various algorithms such as random forest, multiple linear regression, and k-nearest neighbor. Notably, previous models have employed various algorithms, such as random forest, multiple linear regression, and k-nearest neighbor. The results underscore that logistic regression outperforms these algorithms, indicating its superiority in predicting admission chances. This research endeavors to empower students with a more comprehensive understanding of their admission possibilities (Bitar & Almauza, 2020).

Moreover, a study revolves around predicting student admission to Master's degree programs in universities, primarily employing logistic regression and various admission criteria, including GRE scores, TOEFL scores, university ratings, Statements of Purpose (SOP), Letters of Recommendation (LOR), and CGPA. The research aims to identify the most influential variables that impact successful admissions (Rajagopal, 2020). It introduces models like SVM, Gaussian Naive Bayes, and Logistic Regression, with empirical evidence underscoring the superiority of the Logistic Regression model. This research endeavors to empower students with a more comprehensive understanding of their admission possibilities well in advance (Nalam & Alimuddin, 2023).

Furthermore, the paper (CS et al., 2021) addresses the critical issue of student admissions in higher education, employing AI models like Linear Regression, Decision Tree Regressor, and Random Forest Regressor to predict a student's likelihood of admission to a master's program. It highlights the application of machine learning in simplifying the complex and often opaque university admission process. Notably, Linear Regression outperforms other models, offering students an early insight into their acceptance prospects.

Another research article (Prashad, 2022) proposes a machine learning model that aids students in predicting their chances of admission to specific universities based on their test scores and relevant data. The study compares different machine learning methods, including Support Vector Machines (SVM), Random Forest, and linear regression, ultimately achieving an average accuracy of 79%. This model offers students a cost-effective and time-saving alternative to traditional advisory services and application fees. It mitigates the need for costly consultancy services or unreliable online resources.

Moreover, a comprehensive review paper (Kiran & Paul, 2022) focuses on the application of data mining techniques to predict student success in educational institutions. The study highlights the potential of knowledge mining in evaluating student admission trends and offers guidelines for efficient utilization of educational data mining methods. It employs machine learning techniques, including linear regression and random forest algorithms, with CatBoost showing the highest accuracy.

Lastly, a study (Alyahyan & Düştegör, 2020) provides a clear six-stage framework for educators to efficiently utilize educational data mining (EDM) methods. It underscores the relevance of factors such as prior academic achievement, student demographics, e-learning activity, and psychological attributes in predicting academic success. The significance of early student performance prediction for universities to take timely actions is highlighted.

This literature review presents a wide array of research efforts aimed at simplifying university admission processes, empowering students with data-driven insights, and enhancing decision-making in the pursuit of higher education (Fatiya & Sadath, 2021; Bitar & Almauza, 2020; Rajagopal, 2020; Nalam & Alimuddin, 2023; CS et al., 2021; Prashad, 2022; Kiran & Paul, 2022; Alyahyan & Düştegör, 2020).

## Methodology

**Data Preparation:**

Before applying regression model to the data, it is essential to prepare the data properly. This involves tasks such as standardizing, normalizing, or encoding categorical variables. In this dataset, there are no categorical variables to be encoded, but standardization or normalization techniques can be used to deal with outliers and ensure that the data is scaled appropriately. Moreover, it will be easier for our model to study the data in a normalized form than the provided syntax. Finally, the data will go through different stages in terms of cleaning and preparation beyond the standard terms to achieve the best possible result for this study. Noting that the "Chance of Admit" column has been transformed into a binary input of 1 and 0 to make it easier for the AI to predict acceptance or rejection. This was achieved by marking all students who have an 80% or greater chance of being accepted as 1, and those with less than 80% chance as 0.

**Exploratory Data Analysis:**

Started with the EDA i have done and checked null values and than analysis the relationship between variables where i got a plot(Fig 2),where it showcases a series of scatter plots and histograms that illustrate the relationships between graduate admission variables like GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research experience, and Admission Chance. Higher GRE and TOEFL scores generally correlate with higher university ratings, SOP, LOR, and CGPA, indicating a trend where higher academic and English proficiency align with better overall application strength. The concentration of data points at the higher end of CGPA and standardized test scores suggests that the applicant pool is academically strong. Research experience divides the applicant pool into two distinct groups, with those having research experience also showing higher scores in other variables. When it comes to the chance of admission, there is a clear positive relationship with all the variables, particularly with CGPA and standardized test scores. The plots do not suggest a need for non-linear transformations such as quadratic or logarithmic, as the trends appear mostly linear. This indicates that while all factors are relevant, academic performance and test scores are especially significant in determining admission chances.

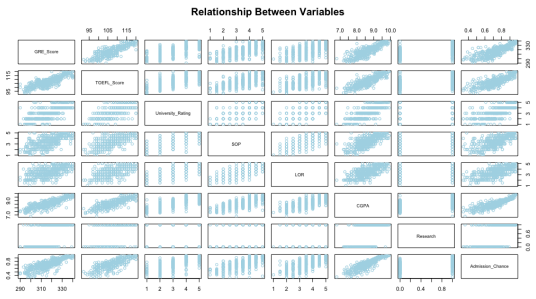


Fig 2: Relationship Between the variables

**Pairs Plot Analysis:**

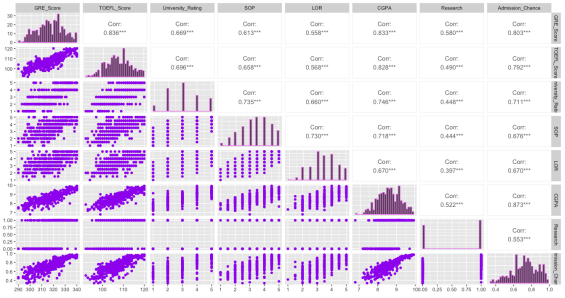

Fig 3: Correlation Between Explanatory and Response Variable

The pairs plot above(Fig 3) is a matrix of scatter plots, histograms, and correlation coefficients for various variables related to graduate admissions, including GRE Score, TOEFL Score, University Rating, Statement of Purpose (SOP), Letters of Recommendation (LOR), Cumulative Grade Point Average (CGPA), Research experience, and the chance of admission. Notably, GRE and TOEFL scores are highly correlated (0.836***), indicating that students who perform well on one test tend to perform well on the other. CGPA shows a very strong correlation with admission chances (0.873***), suggesting it is a critical factor in admission decisions. Research experience has a moderate correlation with admission chances (0.553**). While most relationships appear linear, indicating that higher scores and ratings generally correlate with better admission chances, the scatter plots suggest that at higher ranges of GRE and TOEFL scores, the increase in admission chances diminishes, which could warrant the inclusion of a quadratic term in the predictive model. However, a log transformation does not appear necessary as the relationships do not exhibit exponential characteristics. Both LOR and SOP have identical correlations with admission chances (0.670***), possibly reflecting their equal importance in the evaluation process. The triple asterisks denote a high level of statistical significance, reinforcing the reliability of these correlations.

## Multiple Linear Regression Model Fitting

**Beta-Coffeficiant:**

```
    Residuals:
     Min       1Q   Median       3Q      Max
-0.26259 -0.02103  0.01005  0.03628  0.15928

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.2594325  0.1247307 -10.097
< 2e-16 ***
GRE_Score         0.0017374  0.0005979
2.906  0.00387 **
TOEFL_Score       0.0029196  0.0010895
2.680  0.00768 **
University_Rating 0.0057167  0.0047704
1.198  0.23150
SOP              -0.0033052  0.0055616  -0.594
0.55267
LOR               0.0223531  0.0055415   4.034
6.6e-05 ***
CGPA              0.1189395  0.0122194   9.734
< 2e-16 ***
Research          0.0245251  0.0079598   3.081
0.00221 **
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06378 on 392 degrees of freedom
Multiple R-squared: 0.8035, Adjusted R-squared: 0.8
F-statistic: 228.9 on 7 and 392 DF, p-value: < 2.2e-16

The regression analysis summary reveals that the most significant predictor in the model is research experience, which significantly influences the dependent variable. While the model effectively explains a substantial amount of variance, the individual coefficients for GRE Score, TOEFL Score, CGPA, and their interaction are not statistically significant. This implies that, despite the model's overall effectiveness, these particular variables, as currently modeled, do not have a statistically significant impact on the dependent variable. The significant intercept indicates that the model is meaningful, but caution should be exercised in interpreting the practical significance of these findings, as statistical significance doesn't always equate to real-world impact.

The model has a Multiple R-squared of 0.8035 and an Adjusted R-squared of 0.8, suggesting a high level of explanatory power. These values imply that the model explains approximately 79% of the variance in the dependent variable, which is a strong fit in many contexts. However, while these high R-squared values are promising, they don't guarantee that the model will accurately describe the population. This is because R-squared values are sensitive to the sample data and might not capture all relevant predictors or the complexities of the relationships in the population. Additionally, a high R-squared does not imply causation, nor does it ensure that the model is free from bias or specification errors. It's also crucial to consider the significance of the individual coefficients: in this model, not all coefficients are statistically significant, suggesting that some predictors might not meaningfully contribute to the model. Therefore, while the model demonstrates a strong fit, cautious interpretation and validation with new data are advised.

**Residuals Analysis:**

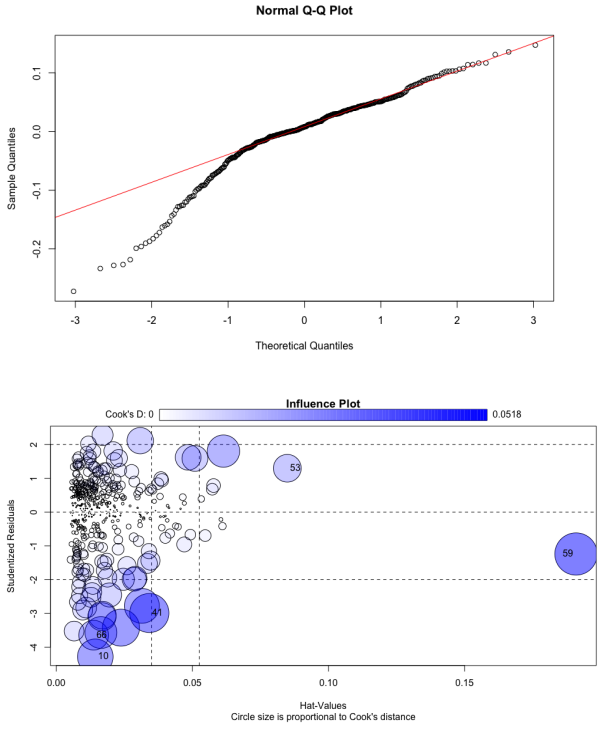


Fig 4: Residuals Model Analysis

The residuals model seems to perform reasonably well in terms of the assumptions of linearity, normality, and homoscedasticity. The influence plot, however, identifies a few points that could be disproportionately impacting the model. It would be prudent to investigate these points further to understand if they are data entry errors, outliers due to exceptional but valid circumstances, or influential points that are valid but have a large impact on the model. Re-fitting the model without these points and comparing the results would help ascertain their effect on the model's predictive power and determine if the original model is robust or if it's overly sensitive to these data points. The decision to remove any data points should be carefully considered and justified, as it can impact the model's validity and generalizability.

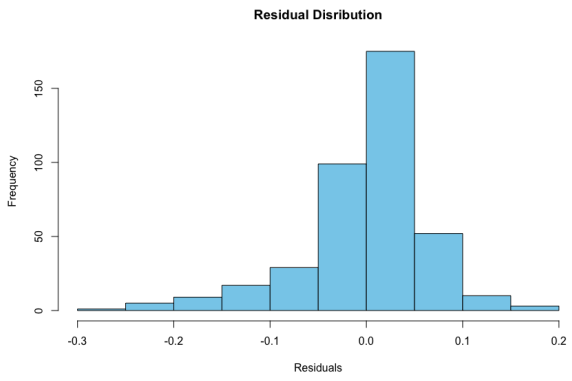


Fig 4: Histogram of Residuals Analysis

The histogram(Fig 4) of residuals suggests that the statistical model is performing reasonably well. The residuals are predominantly centered around zero, indicating no significant bias in the model's predictions. However, there's a slight skewness to the right, with a modest accumulation of small positive residuals. This skewness might hint at potential issues with the model's assumptions, such as non-normality of errors, or it may reflect an aspect of the underlying data distribution that the model does not capture. The range of residuals is fairly narrow, which generally indicates that the model's predictions are not far off from the actual values. Overall, while the presence of skewness warrants further investigation, the model appears to be reasonably accurate in its predictions, as indicated by the concentration of residuals around zero.

**F- Test:**

```
 anova(smaller_model, model)
Analysis of Variance Table

Model 1: graduate$Admission_Chance ~ graduate$GRE_Score +
graduate$TOEFL_Score
Model 2: graduate$Admission_Chance ~ graduate$GRE_Score *
graduate$TOEFL_Score +
    graduate$CGPA + I(graduate$CGPA^2) +
graduate$Research
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1    397 2.4952
2    393 1.6866  4   0.80856 47.1 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

The ANOVA comparison between the smaller and larger models indicates that the larger model, which includes additional variables and interactions (GRE Score * TOEFL Score, CGPA, CGPA squared, and Research), provides a significantly better fit to the data than the smaller model (which only includes GRE Score and TOEFL Score). This conclusion is drawn from the highly significant F-test ($p < 2.2e-16$) and the substantial reduction in Residual Sum of Squares (RSS) in the larger model. The addition of these variables and the interaction term significantly improves the model's ability to explain the variability in the graduate admission chances. Given these results, reporting the larger model to a boss would be advisable, as it not only captures more factors influencing admission chances but also does so with greater statistical significance, thereby providing a more comprehensive and accurate representation of the factors influencing graduate admissions.

## Summary

**Model Interpretation:**

In the regression model summary, 'Research' emerges as the only variable with statistical significance, likely indicating that research experience has a substantial and consistent impact on the model's outcome, which might be related to graduate admissions chances. The lack of significance in other variables such as GRE and TOEFL scores, and CGPA could stem from several factors: these predictors may not linearly correlate with the outcome, or there might be multicollinearity—where GRE and TOEFL scores, both academic performance indicators, are interrelated, potentially diluting each other's statistical impact. Additionally, the significant effect of 'Research' could overshadow other variables, suggesting that when evaluating graduate applications, research experience might be weighted more heavily than test scores or GPA. Moreover, the non-significant squared CGPA term suggests the relationship between CGPA and the outcome is not quadratic, or the model may not be capturing other forms of non-linearity. The absence of significant effects for these variables does not negate their practical importance but may reflect the model's specification, sample size, or the complexity of the relationships in the data, suggesting a reevaluation of the model's structure could be warranted to better understand the dynamics at play.

```
#Prediction Intervals for future values
Predict <- predict(model_mlr, test)
test$Predict <- ifelse(Predict < 0.6, "0", "1")
kable(test[1:10,]) %>%
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),

latex_options="scale_down")
#Confidence Interval
Predict2 <- predict(model_mlr, interval = "confidence")
View(Predict2)
```

| | GRE_Score | TOEFL_Score | University_Rating | SOP | LOR | CGPA | Research | Admission_Chance | Predict |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.80 | 1 |
| 5 | 314 | 103 | 2 | 2.0 | 3.0 | 8.21 | 0 | 0.65 | 1 |
| 6 | 330 | 115 | 5 | 4.5 | 3.0 | 9.34 | 1 | 0.90 | 1 |
| 7 | 321 | 109 | 3 | 3.0 | 4.0 | 8.20 | 1 | 0.75 | 1 |
| 9 | 302 | 102 | 1 | 2.0 | 1.5 | 8.00 | 0 | 0.50 | 0 |
| 10 | 323 | 108 | 3 | 3.5 | 3.0 | 8.60 | 0 | 0.45 | 1 |
| 12 | 327 | 111 | 4 | 4.0 | 4.5 | 9.00 | 1 | 0.84 | 1 |
| 15 | 311 | 104 | 3 | 3.5 | 2.0 | 8.20 | 1 | 0.61 | 1 |
| 16 | 314 | 105 | 3 | 3.5 | 2.5 | 8.30 | 0 | 0.54 | 1 |
| 19 | 318 | 110 | 3 | 4.0 | 3.0 | 8.80 | 0 | 0.63 | 1 |

Fig 5: Prediction Model

The 'Admission_Chance' column represents a model's estimated probability of each applicant's admission, while the 'Predict' column categorizes these chances into binary outcomes (1 for likely admission and 0 for unlikely admission), using 0.6 as the threshold.From the snapshot of the table, it is evident that the model is using these attributes to calculate the likelihood of admission and then applying a decision threshold to make a binary prediction. This classification could be used by admissions committees to quickly identify candidates who meet a certain probability threshold for admission. It also appears that applicants with higher GRE scores, TOEFL scores, CGPAs, and those with research experience tend to have higher admission chances. The table format, along with the threshold-based classification, provides a clear and practical way to interpret the model's continuous predictions in a binary, actionable manner for decision-making processes.

**Confidence Interval:**

Confidence intervals(fit,lower,upper) i have found from my analysis give an indication of the precision of the model's predictions and the uncertainty around these predictions. For example, in row 1, the model predicts an admission chance of 0.9515 with a 95% confidence interval ranging from approximately 0.8252 to 0.9770. This means we can be 95% confident that the true mean admission chance for applicants with a similar profile to the one represented in row 1 lies within this interval. The actual admission chance for any one individual could be different, as individual outcomes vary more than means.

**Hypothesis Testing:**

- Conducted hypothesis tests: 8 (denoted as `m`).
- Post-Bonferroni correction results:
  Significant variables: GRE Score, LOR, CGPA, Research.
  Non-significant variables: TOEFL Score, University Rating, SOP.
- Implications of the correction:
  GRE scores, letters of recommendation, CGPA, and research experience uphold statistical significance.
  TOEFL Score, University Rating, and SOP's significance is diminished after accounting for multiple testing.
- Consequences for analysis:
  Strongest relationships (smallest p-values) remain significant.

## Discussion

This multiple linear regression analysis conducted in this study provides valuable insights into the factors influencing graduate school admissions. The model's robust R-squared values suggest a strong explanatory power, indicating that a significant proportion of the variance in admission chances can be accounted for by the predictor variables used. Notably, 'Research' has emerged as a highly significant predictor, underscoring the importance of research experience in the admission process. In contrast, variables like TOEFL Score, University Rating, and SOP, which became non-significant after a Bonferroni correction, suggest their impact on admission chances is less clear-cut than initially believed.

The residual analysis indicates a good fit for the model, with residuals centered around zero and no apparent patterns that would suggest non-linear relationships or heteroscedasticity. However, the slight skewness observed in the residuals calls for cautious interpretation and suggests that further refinement of the model might be beneficial.

## Conclusion

This study's findings highlight the complex nature of graduate admissions and the multifaceted criteria used in decision-making processes. The significant variables in the model—GRE Score, LOR, CGPA, and Research—should be considered by prospective students as key areas to focus on when preparing their applications. However, it's important to recognize that the admissions process is inherently subjective and can be influenced by factors beyond those quantifiable in a regression model.

## Future Work

- Model Enhancement: To address the skewness in the residuals, future work could explore non-linear models or transformation of variables to achieve a more normal distribution of residuals.
- Data Expansion: Including more diverse data points from different universities and programs could improve the model's generalizability.
- Feature Engineering: Creating new variables, such as an aggregate score from SOP and LOR, might uncover more nuanced relationships.
- Algorithm Exploration: Employing other predictive modeling techniques, such as decision trees or neural networks, could yield different insights and potentially improve prediction accuracy.
- Qualitative Analysis: Incorporating qualitative factors, like personal statements, could provide a more holistic view of the admissions process.

- Time Series Analysis: Examining how admission trends change over time might offer predictive insights into future admission cycles.

By expanding the scope of the data and employing a broader range of analytical techniques, future research could build upon this foundation to provide an even richer understanding of the graduate admissions process.

## References


Acharya, M. S. (2019). A comparison of regression model for prediction of graduate admission. *IEEE conference paper*. 10.1109/ICCIDS.2019.8862140

Algamdhi, A., & Barsheed, A. (2020). A machine learning approach for graduate admission prediction. *IVSP 2020 2nd international conference*. 10.1145/3388818.3393716

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational technology in higher education*.

Bitar, Z., & Almauza, A. (2020, March). Prediction of Graduate Admission using Multiple Supervised Machine Learning Models. *IEEE southwest*. 10.1109/SoutheastCon44009.2020.9249747

CS, K., B, A., GR, C., & JB, M. (2021, October). University Admission Prediction Using Machine learning. *Global Journal of research and review*.

Fatiya, H., & Sadath, L. (2021). *University Admissions Predictor Using Logistic Regression*. https://doi.org/10.1109/ICCIKE51210.2021.9410717

Guabassi, I. a., & Bousalim, Z. (2021). A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithm. *International journal of online and biomedical institute*, *17*(02), 5.

https://doi.org/10.3991/ijoe.v17i02.20049

Iman, A., & Tiang, X. (2021). A Comparison of Classification Models in Predicting Graduate Admission Decision. *Journal of Higher Education Theory and practice*, *21*(7). https://doi.org/10.33423/jhetp.v21i7.4498

Kiran, B. U., & Paul, B. S. (2022, December). ADMISSION PREDICTION FOR MS IN FOREIGN UNIVERSITIES USING MACHINE LEARNING. *International Research Journal of Modernization in Engineering Technology and Science*, *04*(12).

Kumar, S.T.P. S., Anish, R., & Hariana, M. (2023). Prediction of foreigh Admission Using Data Science. *International Journal Of Engineering Research and Technology*, *11*(03). 10.17577/IJERTCONV11IS03032

Nalam, S., & Alimuddin, M. (2023). Advance Graduate admission Prediction. *IEEE Conference 8th*. 10.1109/I2CT57861.2023.10126307

National Commission for Academic Accreditation & Assessment Standards for Quality Assurance and Accreditation of Higher Education Institutions (Ed.). (2015). NAtional Commission for Academic Accreditation &amp. 3.