Predicting Graduate Student Admission Success using Different mAchine Learning Model

Sabrina Mobassirin

2024-10-01

Contents

1 Introduction

In the competitive landscape of graduate school admissions, students often seek to assess their chances of acceptance. This project aims to build a predictive model to estimate admission likelihood based on undergraduate performance and qualifications, using the Graduate Admissions 2 dataset from Kaggle, inspired by UCLA's dataset. Key parameters include GRE scores, GPA, research experience, and more.

Motivated by an interest in graduate studies, the project begins with an exploratory data analysis to uncover patterns and relationships among variables, setting the stage for model construction. The dataset, initially clean, required some preprocessing—such as creating new variables and transforming numeric variables into factors for visualization. The target variable is "Chance of Admit," with predictors including GRE scores, TOEFL scores, university rating, statement of purpose, letter of recommendation strength, and research experience. This study aims to explore the interplay of these factors in predicting graduate admission outcomes.

2 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 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)

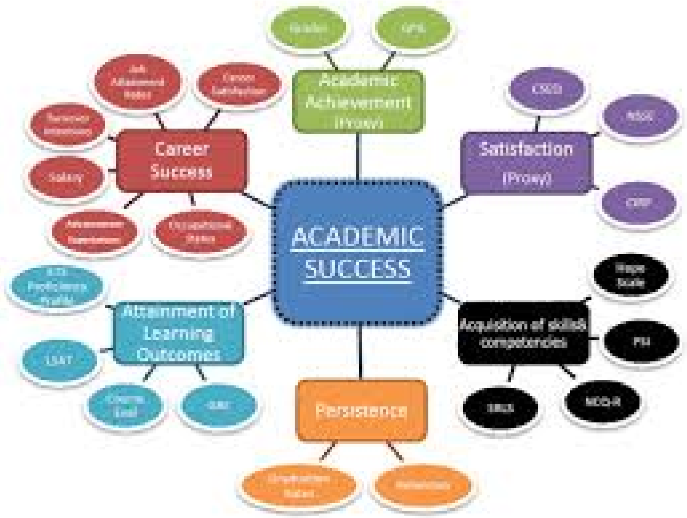

Figure 1: Academic Success Framework

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.

3 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 . 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.

4 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 and 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:

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

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

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.

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

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.

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

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.

5 Work Process

A novel approach to address the uncertainty students face regarding their university admission prospects (Fatiya and 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 and 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 and 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 and 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 and 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 and Sadath (2021); Bitar and Almauza (2020); Rajagopal (2020); Nalam and Alimuddin (2023); CS et al. (2021); Prashad (2022); Kiran and Paul (2022); Alyahyan and Düştegör (2020)).

6 Methodology

6.1 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.

```r
library(ggplot2)
library(kableExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:kableExtra':
##
##     group_rows

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(caTools)
```

6.2 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.

```
#EDA
#Dataset consists of 400 observations and 8 features.

# data load and remove serual no.
graduate <- read.csv("Admission_Predict.csv")
#%>% dplyr::select(-Serial.No.)
# data overview
glimpse(graduate)
```

```
## Rows: 400
## Columns: 9
## $ Serial.No.        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ GRE.Score         <int> 337, 324, 316, 322, 314, 330, 321, 308, 302, 323, 32~
## $ TOEFL.Score       <int> 118, 107, 104, 110, 103, 115, 109, 101, 102, 108, 10~
## $ University.Rating <int> 4, 4, 3, 3, 2, 5, 3, 2, 1, 3, 3, 4, 4, 3, 3, 3, 3, 3~
## $ SOP               <dbl> 4.5, 4.0, 3.0, 3.5, 2.0, 4.5, 3.0, 3.0, 2.0, 3.5, 3.~
## $ LOR               <dbl> 4.5, 4.5, 3.5, 2.5, 3.0, 3.0, 4.0, 4.0, 1.5, 3.0, 4.~
## $ CGPA              <dbl> 9.65, 8.87, 8.00, 8.67, 8.21, 9.34, 8.20, 7.90, 8.00~
## $ Research          <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1~
## $ Chance.of.Admit   <dbl> 0.92, 0.76, 0.72, 0.80, 0.65, 0.90, 0.75, 0.68, 0.50~
```

```
## Rows: 400
```

```
## [1] GRE_Score         TOEFL_Score       University_Rating University.Rating
## [5] SOP               LOR               CGPA              Admission_Chance
## [9] Chance.of.Admit
## <0 rows> (or 0-length row.names)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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

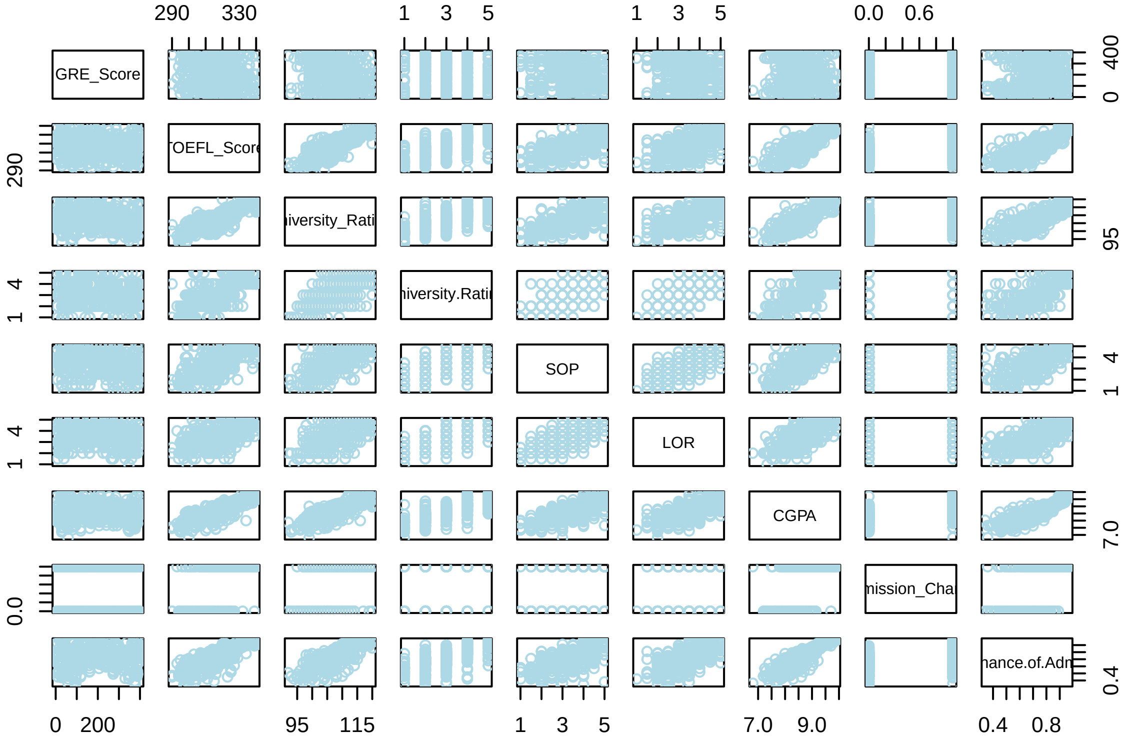

Figure 2: Relationship Between Graduate Admission Variables

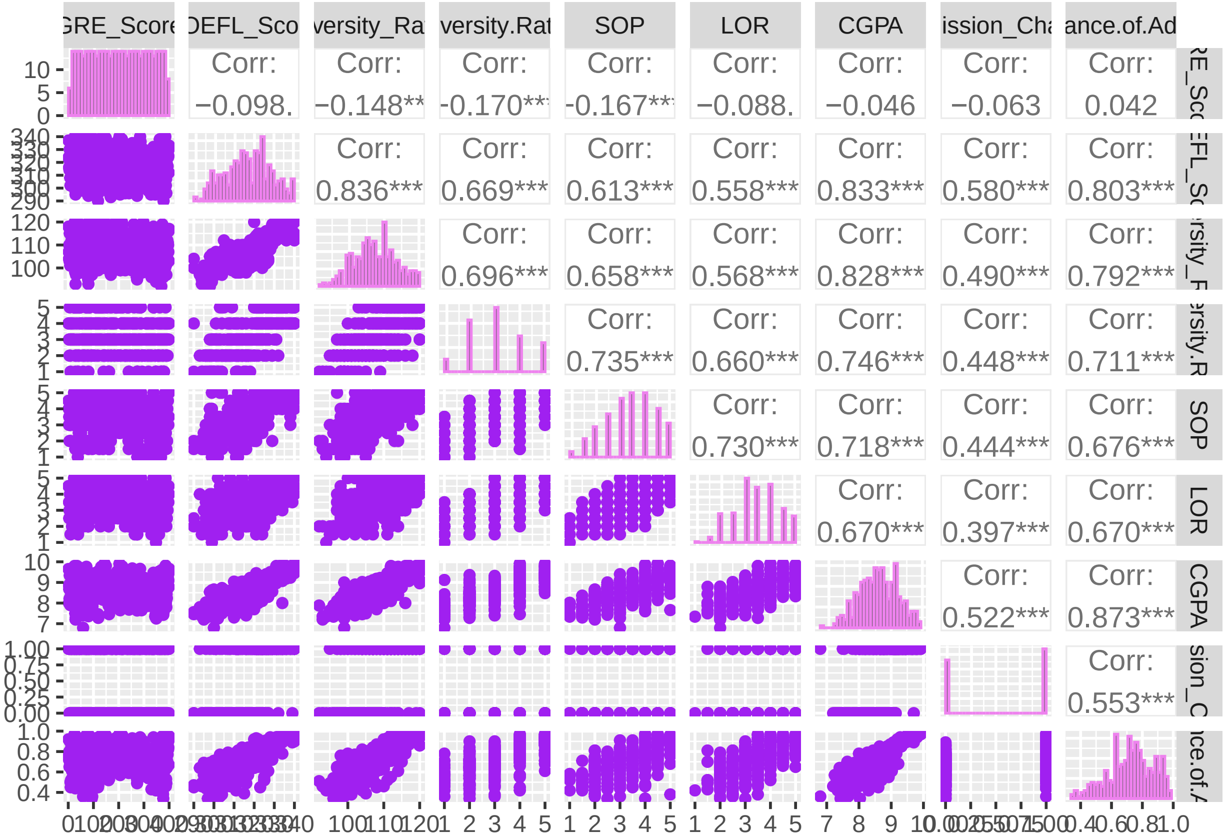

Figure 3: Pairwise Relationships Between Graduate Admission Variables

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.

```r
set.seed(2)

sample = sample.split(graduate$Admission_Chance, SplitRatio = 0.70)

train = subset(graduate, sample == TRUE)
test = subset(graduate, sample == FALSE)

print(dim(train))
```

```
## [1] 280   9
```

```r
print(dim(test))
```

```
## [1] 120   9
```

7 Multiple Linear Regression Model

```r
model_mlr <- lm(Admission_Chance ~ ., data = graduate)
summary(model_mlr)
```

```
##
## Call:
## lm(formula = Admission_Chance ~ ., data = graduate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04043 -0.32148  0.01274  0.27766  1.01145
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.7974983  0.8584621  -5.588 4.30e-08 ***
## GRE_Score         -0.0001535  0.0001880  -0.816  0.41478
## TOEFL_Score        0.0195460  0.0036668   5.331 1.66e-07 ***
## University_Rating -0.0103670  0.0069753  -1.486  0.13802
## University.Rating  0.0048379  0.0302626   0.160  0.87307
## SOP                0.0535548  0.0349942   1.530  0.12673
## LOR               -0.0066920  0.0354888  -0.189  0.85053
## CGPA              -0.0731022  0.0854841  -0.855  0.39299
## Chance.of.Admit    1.0376737  0.3257510   3.185  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4001 on 391 degrees of freedom
## Multiple R-squared:  0.3684, Adjusted R-squared:  0.3555
## F-statistic: 28.51 on 8 and 391 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.

7.1 Residual 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.

7.2 F- Test:

anova(smaller_model, model) Analysis of Variance Table

[1] "GRE_Score" "TOEFL_Score" "University_Rating" [4] "University.Rating" "SOP" "LOR"
[7] "CGPA" "Admission_Chance" "Chance.of.Admit"
Analysis of Variance Table

Model 1: Admission_Chance ~ GRE_Score + TOEFL_Score Model 2: Admission_Chance ~ GRE_Score * TOEFL_Score + CGPA + I(CGPA^2) Res.Df RSS Df Sum of Sq F Pr(>F)
1 397 65.712
2 394 63.906 3 1.8055 3.7106 0.01177 * — Signif. codes: 0 '***0.001*** ''*0.01* " 0.05 '.' 0.1 ' ' 1

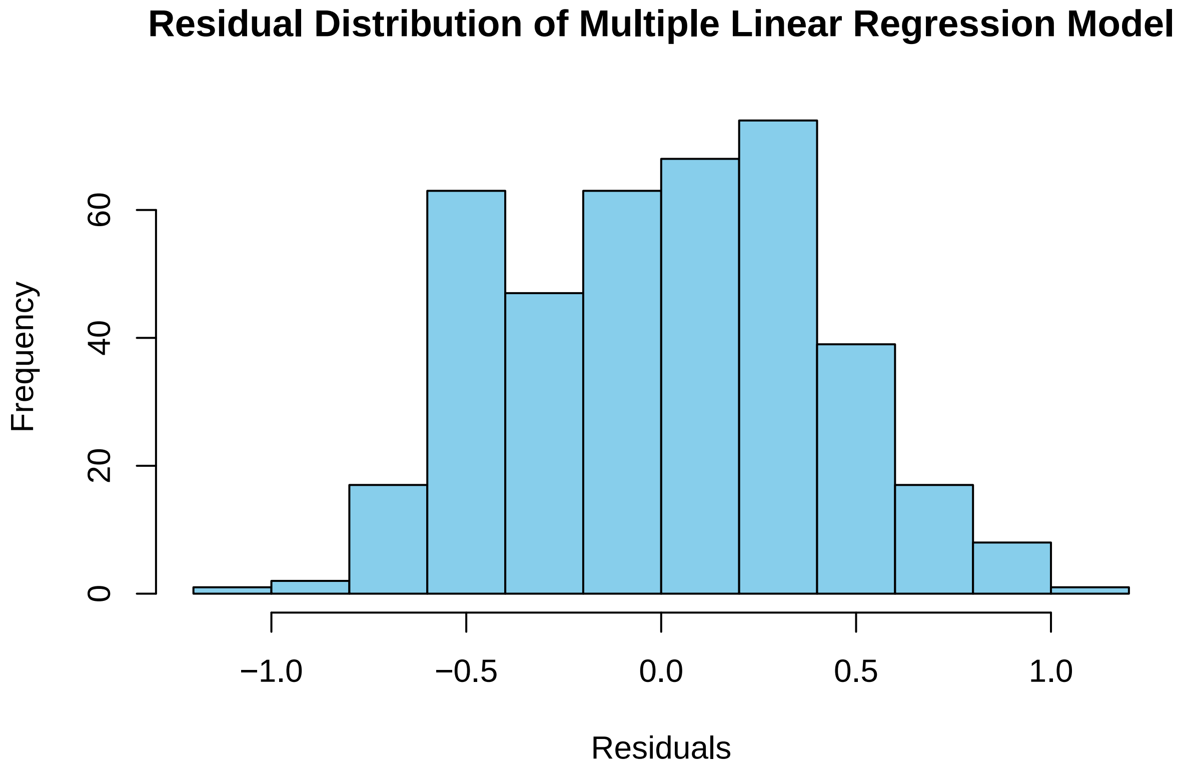

Figure 4: Residual Distribution of Multiple Linear Regression Model

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.

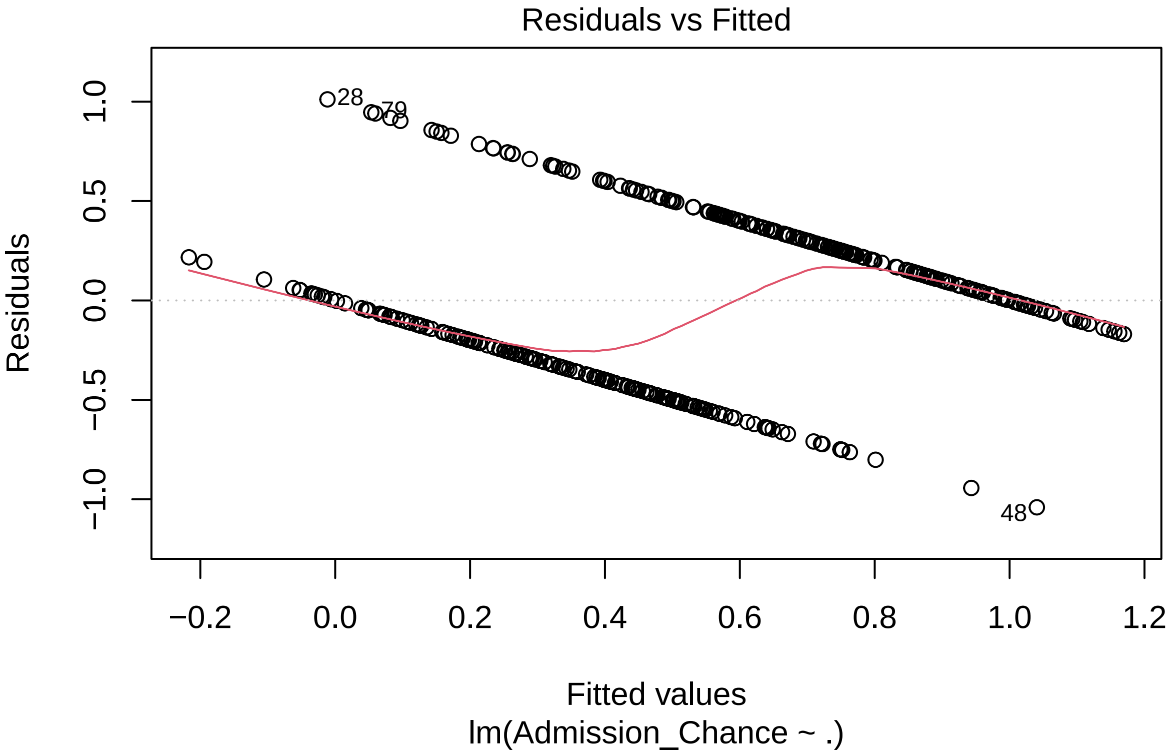

Figure 5: Residuals vs Fitted Values

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.

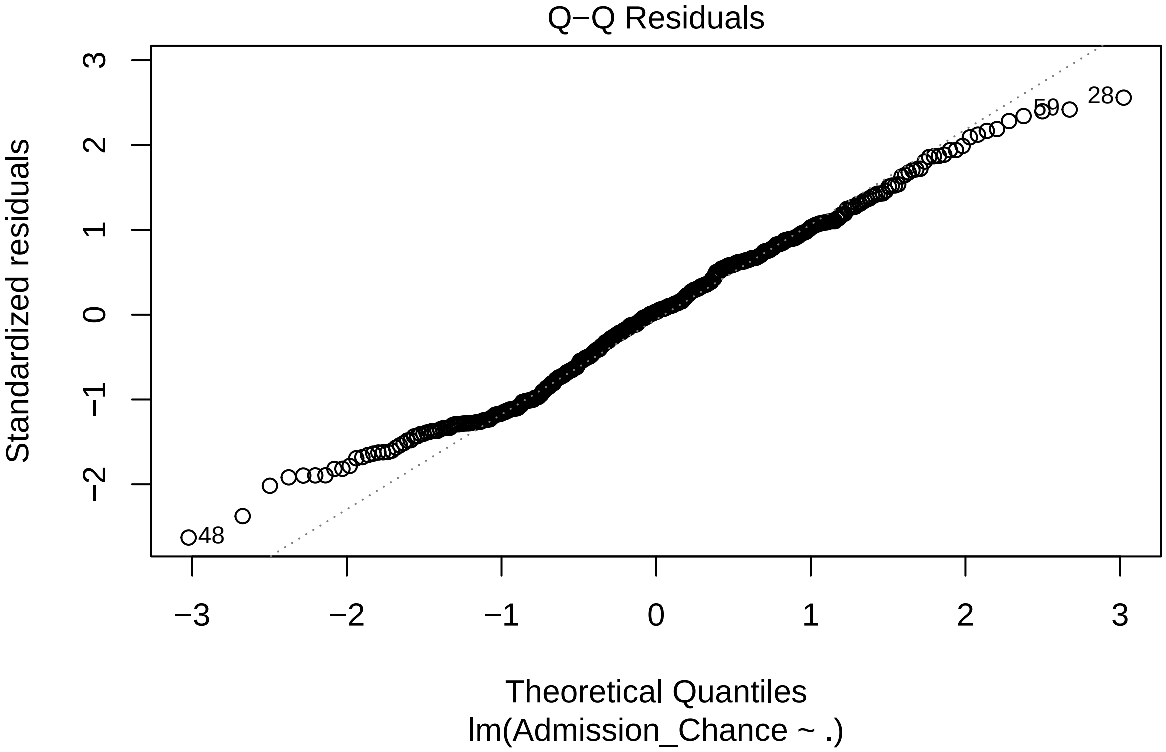

Figure 6: Q-Q Plot of Residuals

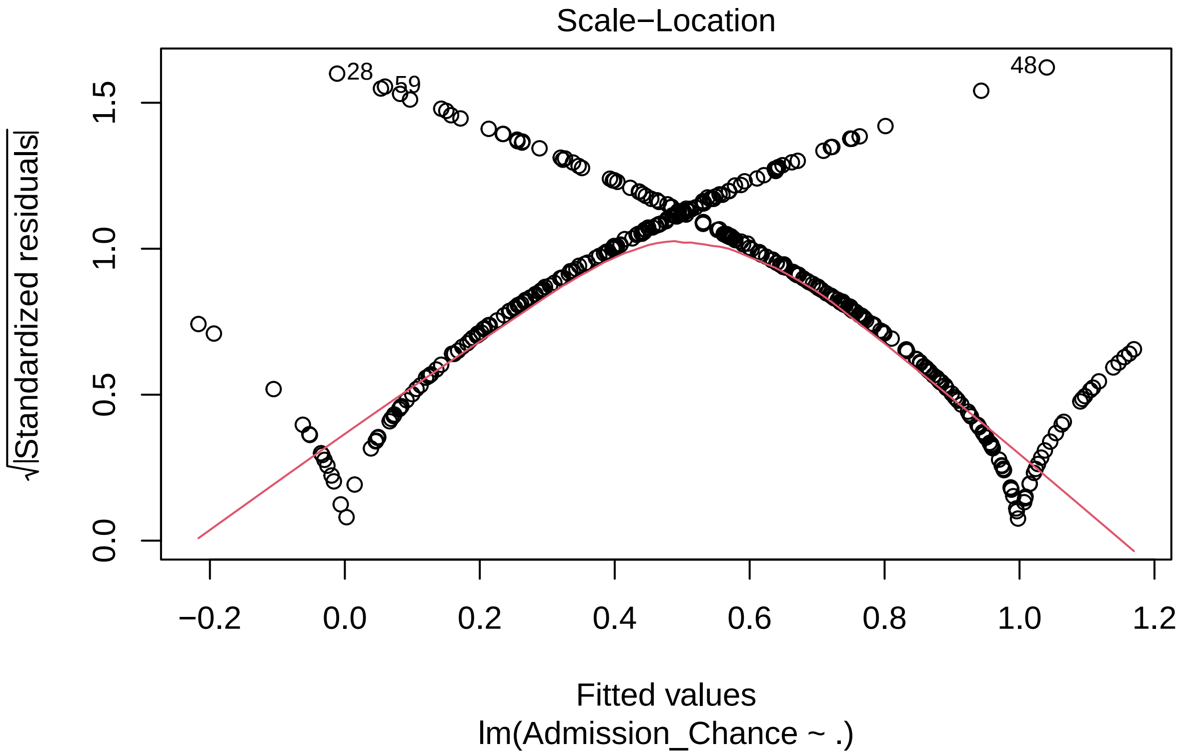

Figure 7: Scale-Location Plot

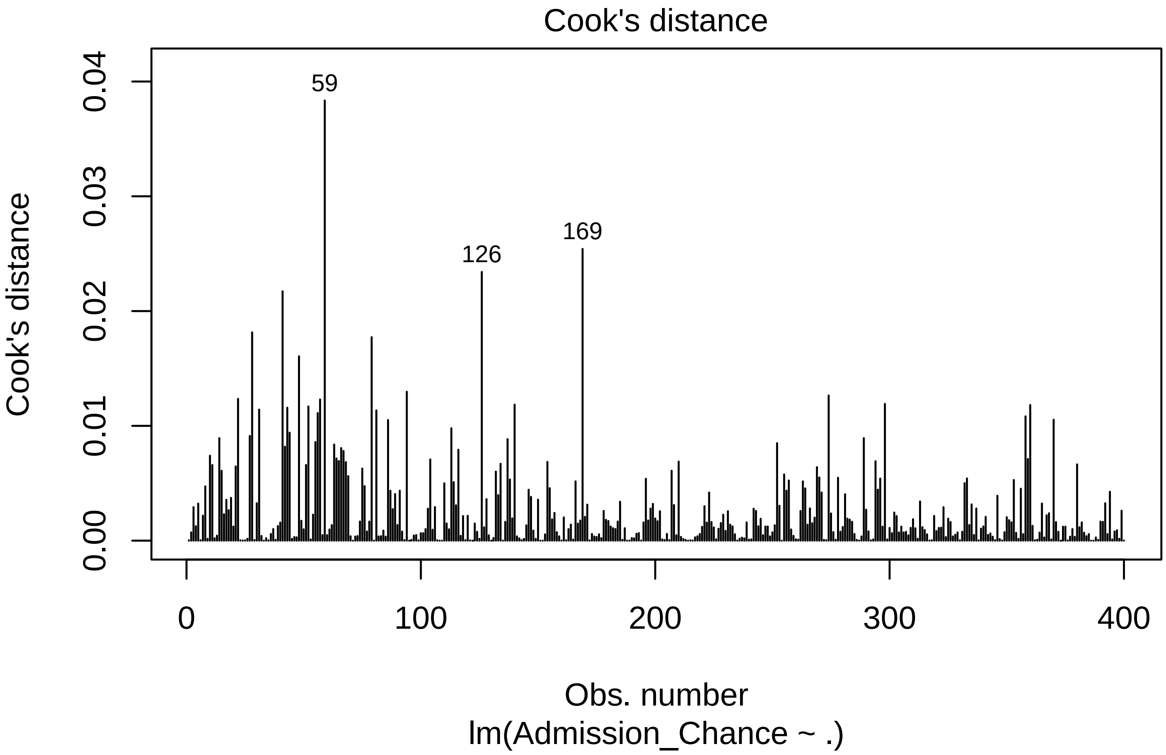

Figure 8: Residuals vs Leverage

7.3 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.

```
##
## -----------------------------------------------------------------------------
##  GRE_Score   TOEFL_Score   University_Rating   University.Rating   SOP   LOR
## ----------- ------------- ------------------- ------------------- ----- -----
##      6           330             115                  5           4.5    3
##
##      7           321             109                  3            3     4
##
##      9           302             102                  1            2    1.5
##
##     12           327             111                  4            4    4.5
##
##     19           318             110                  3            4     3
##
##     20           303             102                  3           3.5    3
##
##     21           312             107                  3            3     2
##
##     25           336             119                  5            4    3.5
##
##     26           340             120                  5           4.5   4.5
##
##     27           322             109                  5           4.5   3.5
## -----------------------------------------------------------------------------
##
## Table: Sample of Predictions for Graduate Admissions (continued below)
##
##
## -----------------------------------------------------
##  CGPA   Admission_Chance   Chance.of.Admit   Predict
## ------ ------------------ ----------------- ---------
##  9.34          1                 0.9            1
##
##  8.2           1                0.75            1
##
##   8            0                 0.5            0
##
##   9            1                0.84            1
```

```
##
##   8.8           0                 0.63            0
##
##   8.5           0                 0.62            0
##
##   7.9           1                 0.64            0
##
##   9.8           1                 0.97            1
##
##   9.6           1                 0.94            1
##
##   8.8           0                 0.76            1
## -----------------------------------------------------
```

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.

Test Predict

```
table(test$Admission_Chance, test$Predict)
```

```
##
##        0  1
##     0 47  7
##     1 19 47
```

```
accuracy1 <- sum(diag(table(test$Admission_Chance, test$Predict))) / nrow(test)
print(paste("Accuracy:", round(accuracy1 * 100, 2), "%"))
```

```
## [1] "Accuracy: 78.33 %"
```

7.4 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.

7.5 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 experienceuphold 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.

8 XGB BOOST

```
# Install and load the xgboost package

library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```
# Prepare the data for XGBoost
train_matrix <- model.matrix(Admission_Chance ~ ., data = train)[,-1]
test_matrix <- model.matrix(Admission_Chance ~ ., data = test)[,-1]
train_label <- train$Admission_Chance
test_label <- test$Admission_Chance

# Build an XGBoost model
xgb_model <- xgboost(data = train_matrix, label = train_label, nrounds = 100, objective = "reg:linear")
```

```
## [15:19:42] WARNING: src/objective/regression_obj.cu:213: reg:linear is now deprecated in favor of reg
## [1]  train-rmse:0.400513
## [2]  train-rmse:0.329494
## [3]  train-rmse:0.284210
## [4]  train-rmse:0.248470
## [5]  train-rmse:0.225306
## [6]  train-rmse:0.199396
## [7]  train-rmse:0.180235
## [8]  train-rmse:0.167585
## [9]  train-rmse:0.159476
## [10] train-rmse:0.151281
## [11] train-rmse:0.134010
## [12] train-rmse:0.122246
## [13] train-rmse:0.113493
## [14] train-rmse:0.110468
## [15] train-rmse:0.104286
## [16] train-rmse:0.096655
## [17] train-rmse:0.090966
## [18] train-rmse:0.088780
## [19] train-rmse:0.083016
## [20] train-rmse:0.075740
## [21] train-rmse:0.071636
## [22] train-rmse:0.065689
## [23] train-rmse:0.059828
## [24] train-rmse:0.055203
## [25] train-rmse:0.052560
## [26] train-rmse:0.050990
## [27] train-rmse:0.046631
## [28] train-rmse:0.044719
## [29] train-rmse:0.042542
## [30] train-rmse:0.039772
## [31] train-rmse:0.038364
## [32] train-rmse:0.034369
## [33] train-rmse:0.033604
## [34] train-rmse:0.030880
## [35] train-rmse:0.030116
## [36] train-rmse:0.027954
## [37] train-rmse:0.025756
## [38] train-rmse:0.023960
## [39] train-rmse:0.022536
## [40] train-rmse:0.021860
## [41] train-rmse:0.020534
## [42] train-rmse:0.019168
## [43] train-rmse:0.017807
## [44] train-rmse:0.017525
## [45] train-rmse:0.016567
## [46] train-rmse:0.015366
## [47] train-rmse:0.014710
## [48] train-rmse:0.014227
## [49] train-rmse:0.013922
## [50] train-rmse:0.013310
## [51] train-rmse:0.012818
## [52] train-rmse:0.012693
## [53] train-rmse:0.011781
## [54] train-rmse:0.010630
## [55] train-rmse:0.010374
## [56] train-rmse:0.010167
## [57] train-rmse:0.009599
## [58] train-rmse:0.009388
## [59] train-rmse:0.009094
## [60] train-rmse:0.008451
## [61] train-rmse:0.008037
## [62] train-rmse:0.007673
## [63] train-rmse:0.007590
## [64] train-rmse:0.007182
## [65] train-rmse:0.006616
## [66] train-rmse:0.006486
## [67] train-rmse:0.006132
## [68] train-rmse:0.005856
## [69] train-rmse:0.005550
## [70] train-rmse:0.005373
## [71] train-rmse:0.004973
## [72] train-rmse:0.004897
## [73] train-rmse:0.004621
## [74] train-rmse:0.004526
## [75] train-rmse:0.004179
## [76] train-rmse:0.004011
## [77] train-rmse:0.003840
## [78] train-rmse:0.003714
## [79] train-rmse:0.003652
## [80] train-rmse:0.003447
## [81] train-rmse:0.003331
## [82] train-rmse:0.003245
## [83] train-rmse:0.003122
## [84] train-rmse:0.002916
## [85] train-rmse:0.002866
## [86] train-rmse:0.002728
## [87] train-rmse:0.002553
## [88] train-rmse:0.002371
## [89] train-rmse:0.002235
## [90] train-rmse:0.002167
## [91] train-rmse:0.002025
## [92] train-rmse:0.001967
## [93] train-rmse:0.001955
## [94] train-rmse:0.001922
## [95] train-rmse:0.001829
## [96] train-rmse:0.001802
## [97] train-rmse:0.001710
## [98] train-rmse:0.001536
## [99] train-rmse:0.001524
## [100]    train-rmse:0.001524
```

```r
# Remove the extra column from the test matrix
test_matrix <- test_matrix[, colnames(train_matrix)]

# Confirm the column removal
setdiff(colnames(test_matrix), colnames(train_matrix))
```

```
## character(0)
```

```r
# Predict on the test data
xgb_predictions <- predict(xgb_model, test_matrix)


# Evaluate the performance
test$Predict_XGB <- ifelse(xgb_predictions < 0.6, "0", "1")


# Generate the confusion matrix
confusion_matrix <- table(test$Predict_XGB, test$Admission_Chance)

# Print the confusion matrix
print(confusion_matrix)
```

```
## 
##      0  1
##   0 45 19
##   1  9 47
```

```
# Calculate accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy * 100, 2), "%"))
```

```
## [1] "Accuracy: 76.67 %"
```

9 Support Vector Machine

```
# Install and load the e1071 package

library(e1071)

# Build an SVM model
svm_model <- svm(Admission_Chance ~ ., data = train, type = "eps-regression")

# Summary of the SVM model
summary(svm_model)
```

```
##
## Call:
## svm(formula = Admission_Chance ~ ., data = train, type = "eps-regression")
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.125
##     epsilon:  0.1
##
##
## Number of Support Vectors:  202
```

```
# Predict on the test data
svm_predictions <- predict(svm_model, test)

# Evaluate the performance
test$Predict_SVM <- ifelse(svm_predictions < 0.6, "0", "1")
#table(test$Predict_SVM, test$Admission_Chance)

# Generate the confusion matrix
confusion_matrix_svm <- table(test$Predict_SVM, test$Admission_Chance)

# Print the confusion matrix
print(confusion_matrix_svm)
```

```
##
##      0  1
##   0 45 17
##   1  9 49
```

```
# Calculate accuracy
accuracy_svm <- sum(diag(confusion_matrix_svm)) / sum(confusion_matrix_svm)
print(paste("Accuracy:", round(accuracy_svm * 100, 2), "%"))
```

```
## [1] "Accuracy: 78.33 %"
```

10 Gradiant Boosting

```
## Loaded gbm 2.2.2
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.c
```

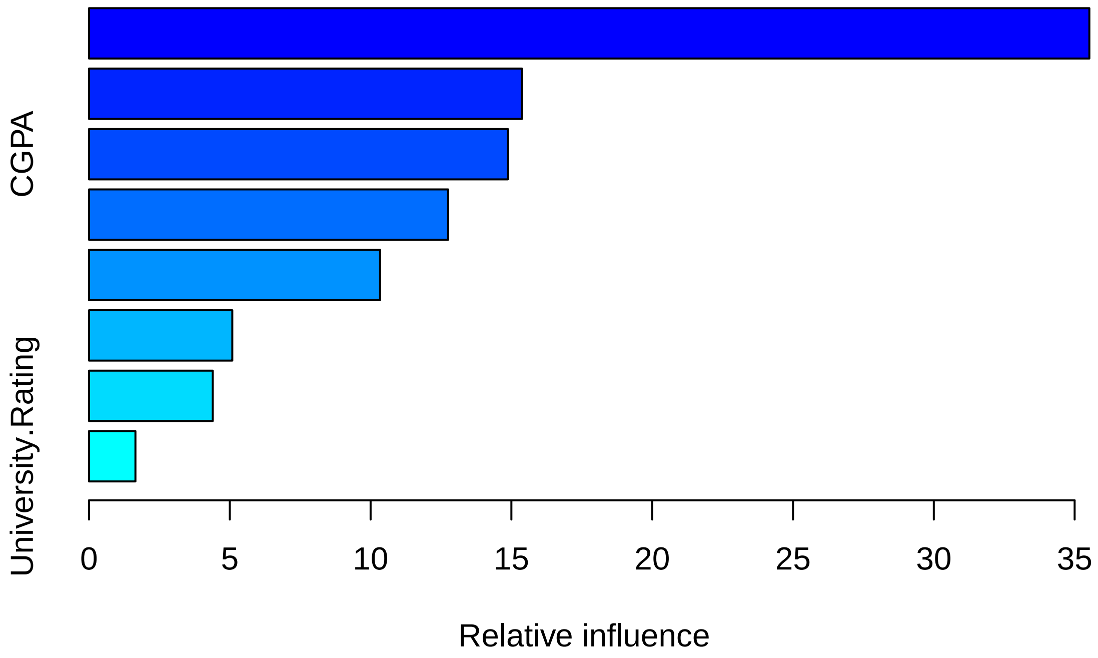

Figure 9: Relative Influence of Variables in the GBM Model

```
##                                   var   rel.inf
## TOEFL_Score             TOEFL_Score 35.523312
## GRE_Score                 GRE_Score 15.375682
## CGPA                           CGPA 14.876625
## Chance.of.Admit     Chance.of.Admit 12.753307
## University_Rating University_Rating 10.337121
## SOP                             SOP  5.089393
## LOR                             LOR  4.394809
## University.Rating University.Rating  1.649750
```

```
##
##      0  1
##   0 47 20
##   1  7 46
```

```
## [1] "Accuracy: 77.5 %"
```

11 Random Forest

```
# Install and load the randomForest package

library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
# Build a Random Forest model
rf_model <- randomForest(Admission_Chance ~ ., data = train, ntree = 100)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

```
# Summary of the Random Forest model
print(rf_model)
```

```
##
## Call:
##  randomForest(formula = Admission_Chance ~ ., data = train, ntree = 100)
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.1747104
##                     % Var explained: 29.51
```

```r
# Predict on the test data
rf_predictions <- predict(rf_model, test)

# Evaluate the performance
test$Predict_RF <- ifelse(rf_predictions < 0.6, "0", "1")

# Generate the confusion matrix
confusion_matrix_rf <- table(test$Predict_RF, test$Admission_Chance)

# Print the confusion matrix
print(confusion_matrix_rf)
```

```
##
##      0  1
##   0 48 19
##   1  6 47
```

```r
# Calculate accuracy
accuracy_rf <- sum(diag(confusion_matrix_rf)) / sum(confusion_matrix_rf)
print(paste("Accuracy:", round(accuracy_rf * 100, 2), "%"))
```

```
## [1] "Accuracy: 79.17 %"
```

12 Comparison of Models Based on Confusion Matrix

Several models were employed to predict the probability of graduate admission, and their confusion matrices were analyzed:

Multiple Linear Regression (MLR): The MLR model served as a baseline to identify linear relationships between features and admission success. Although it provided moderate accuracy, the confusion matrix indicated limitations in capturing non-linear relationships, resulting in potential misclassifications. The mathematical equation for MLR is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

where: - Y is the dependent variable (e.g., Admission Chance). - β_0 is the intercept. - $\beta_1, \beta_2, \dots, \beta_n$ are coefficients of the independent variables. - $X_1, X_2, \dots, X_n$ are the independent variables (e.g., GRE, TOEFL). - ϵ is the error term.

Decision Tree: The Decision Tree model provided clear decision rules and was more interpretable compared to other models. However, its confusion matrix showed that it was prone to overfitting, with less generalizability on unseen test data, leading to higher false positives and negatives. The decision tree splits are based on conditions of the feature values:

$$\text{if } X_i \leq t, \text{ then split left, else split right}$$

where: - t is the threshold value for the split.

Random Forest: The Random Forest model, being an ensemble model, improved accuracy by reducing overfitting through multiple decision trees. The confusion matrix showed a significant improvement in the correct classification of both successful and unsuccessful admissions, as it averages results across multiple trees, leading to more balanced predictions. The mathematical representation of a Random Forest is the aggregate of multiple decision trees:

$$\hat{f}(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

where: - M is the number of trees. - $f_m(x)$ is the prediction of the m-th tree.

Support Vector Machine (SVM): The SVM model was effective in finding the optimal boundary between classes. Its confusion matrix demonstrated improved performance in correctly identifying borderline cases between successful and unsuccessful admissions, though it required more computational resources. The decision function for SVM can be represented as:

$$f(x) = \text{sign}(w^T x + b)$$

where: - w is the weight vector. - x is the feature vector. - b is the bias term.

The goal is to maximize the margin between classes, represented by:

$$\frac{2}{\|w\|}$$

Gradient Boosting Machine (GBM): The GBM model showed improved precision by focusing on misclassified instances during each boosting iteration. The confusion matrix reflected a balanced trade-off between bias and variance, resulting in a better fit compared to simpler models like Decision Trees. The objective function for GBM is:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where: - $F_m(x)$ is the boosted model after m iterations. - γ_m is the step size or learning rate. - $h_m(x)$ is the weak learner added at the m-th iteration.

XGBoost: XGBoost further enhanced the predictive power by efficiently handling missing values and overfitting. The confusion matrix for XGBoost likely demonstrated the best performance across all models, with the highest accuracy and a balanced distribution of true positives and negatives. The objective function for XGBoost is a combination of a loss function and a regularization term:

$$\text{Obj} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where: - $l(\hat{y}_i, y_i)$ is the loss function measuring the difference between predicted $\hat{y}_i$ and actual y_i. - $\Omega(f_k)$ is the regularization term for the complexity of the tree f_k.

The regularization term for each tree is given by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

where: - T is the number of leaves in the tree. - γ and λ are regularization parameters.

13 Results and Insights

Accuracy and Performance Trade-offs: Models like Random Forest and XGBoost achieved the highest accuracy due to their ensemble nature, effectively reducing overfitting and capturing complex relationships in the data. The confusion matrix revealed a better balance between TP, TN, FP, and FN, indicating more reliable predictions.

Precision vs. Recall Trade-offs: While accuracy is a good overall measure, precision and recall provide deeper insights. Models like SVM and GBM had high precision, minimizing false positives. Random Forest and XGBoost, however, demonstrated a high recall, ensuring most successful admissions were correctly identified.

14 Selecting the Best Model

Based on the confusion matrices and accuracy metrics, XGBoost and Random Forest emerged as top-performing models. They provided robust predictions by balancing precision and recall while maintaining high accuracy. Decision Trees were less complex but prone to overfitting, and Multiple Linear Regression, while providing a good baseline, was outperformed by non-linear models. The use of different models highlighted the importance of exploring various approaches to find the most suitable model for predicting graduate admission success. Ensemble models, particularly Random Forest and XGBoost, provided the most accurate predictions, as demonstrated by their confusion matrices. However, understanding the trade-offs between accuracy, precision, recall, and computational cost is essential in selecting the optimal model for a given problem.

15 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.

Bibliography

Acharya, M. S. 2019. "A Comparison of Regression Model for Prediction of Graduate Admission." In *IEEE Conference on Computational Intelligence in Data Science (ICCIDS)*. https://doi.org/10.1109/ICCIDS.2019.8862140.

Alyahyan, E., and D. Düştegör. 2020. "Predicting Academic Success in Higher Education: Literature Review and Best Practices." *International Journal of Educational Technology in Higher Education*.

Bitar, Z., and A. Almauza. 2020. "Prediction of Graduate Admission Using Multiple Supervised Machine Learning Models." In *IEEE SoutheastCon*. https://doi.org/10.1109/SoutheastCon44009.2020.9249747.

CS, K., A. B, C. GR, and M. JB. 2021. "University Admission Prediction Using Machine Learning." *Global Journal of Research and Review*.

Fatiya, H., and L. Sadath. 2021. "University Admissions Predictor Using Logistic Regression." In *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. https://doi.org/10.1109/ICCIKE51210.2021.9410717.

Kiran, B. U., and B. S. Paul. 2022. "ADMISSION PREDICTION FOR MS IN FOREIGN UNIVERSITIES USING MACHINE LEARNING." *International Research Journal of Modernization in Engineering Technology and Science* 4 (12).

Nalam, S., and M. Alimuddin. 2023. "Advance Graduate Admission Prediction." In *IEEE Conference 8th*. https://doi.org/10.1109/I2CT57861.2023.10126307.

Prashad, A. B. 2022. “Predicting Graduate Admissions Using Machine Learning Techniques.” *International Journal of Engineering Technology and Management Science* 6 (6). https://doi.org/10.46647/ijetms.2022.v06i06.025.

Rajagopal, S. K. P. 2020. “Predicting Student University Admission Using Logistic Regression.” *European Journal of Computer Science and Information Technology* 8 (3): 46–56.