

Final Project

Bhavana Vaddi

Department of Engineering and Design, Johnson and Wales University

DATA 5400: Introduction to Predictive Modeling

Casey Godzyk



The project is related to the prediction of graduate admissions into university based on different parameters. The dataset includes 8 attributes and 400 samples (Graduate Admission 2, n.d.). A detailed description of the dataset is given below.

1. GRE scores (out of 340)
2. TOEFL scores (out of 120)
3. University rating (out of 5)
4. Statement of purpose (SOP) (out of 5)
5. Letter of recommendation (LOR) (out of 5)
6. Undergraduate GPA (CGPA) (out of 10)
7. Research experience (Research) (either 0 or 1)
8. Chance of admit (COA) (ranging from 0 to 1)

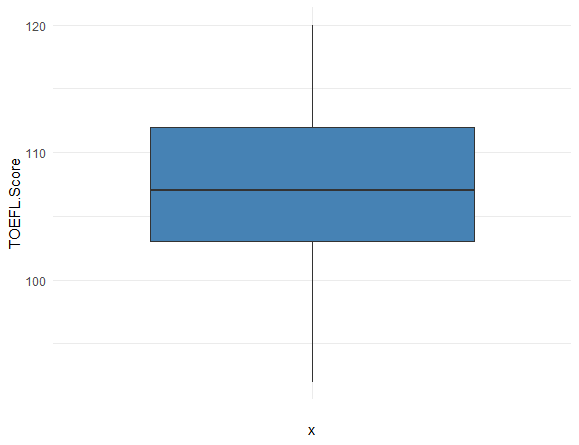
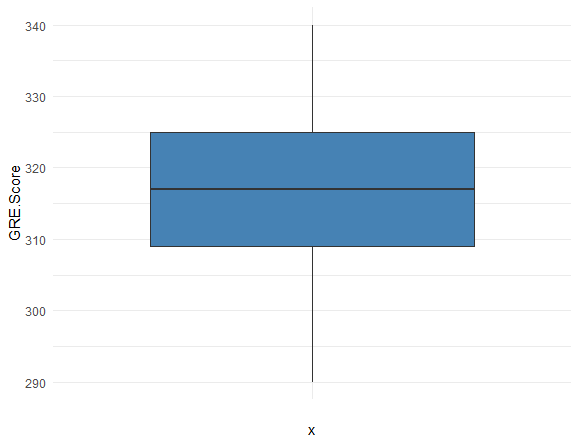
# Exploratory Data Analysis

## Data Preprocessing

The dataset had no missing values. Using boxplot, outliers were detected in three features LOR, CGPA and Chance of admit, which were dropped out as they might change the results. Two attributes, University rating and Research were found to be discrete data and the rest of the data was continuous. It was observed from the boxplot that the distribution of all attributes was normal except SOP, which was slightly left skewed as it was closer to the third quartile, shown in Figure 3.

Figure 1 Figure 2 Figure 3

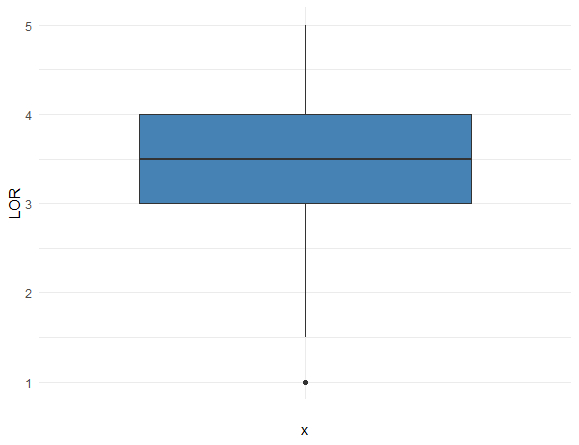
Boxplot of GRE Score Boxplot of TOEFL Score Boxplot of SOP

A blue rectangle with black lines

Description automatically generated

Figure 4 Figure 5 Figure 6

Boxplot of LOR Boxplot of CGPA Boxplot of COA

A blue rectangular object with black lines

Description automatically generatedA blue rectangular object with black lines

Description automatically generated

Figure 7 Figure 8

Boxplot of Research Boxplot of University rating

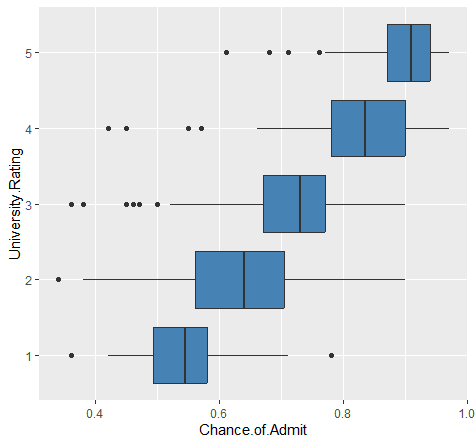
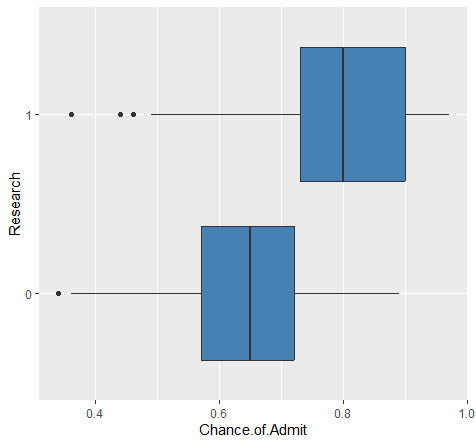


Table 1

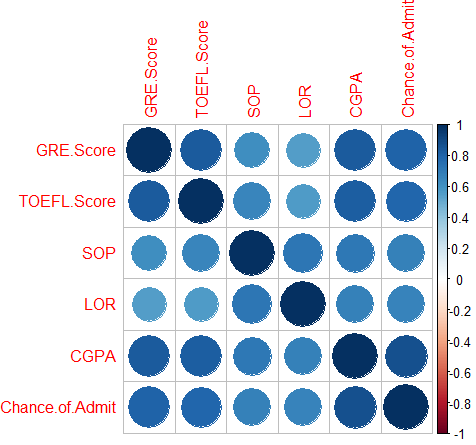
Descriptive statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| Min | 290.00 | 92.00 | 1.00 | 1.00 | 1.00 | 6.80 | 0.00 | 0.34 |
| 1st quartile | 308.00 | 103.00 | 2.00 | 2.50 | 3.00 | 8.17 | 0.00 | 0.64 |
| Median | 317.00 | 107.00 | 3.00 | 3.50 | 3.50 | 8.61 | 1.00 | 0.73 |
| Mean | 316.00 | 107.00 | 3.08 | 3.40 | 3.45 | 8.59 | 0.54 | 0.72 |
| 3rd quartile | 325.00 | 112.00 | 4.00 | 4.00 | 4.00 | 9.06 | 1.00 | 0.83 |
| Max | 340.00 | 120.00 | 5.00 | 5.00 | 5.00 | 9.92 | 1.00 | 0.97 |

From Table 1, it is evident that there is a low variance in data for SOP, LOR and CGPA.

Figure 9

Correlation



From Figure 9, correlation map shows that GRE, TOEFL score and CGPA are highly correlated with COA.

# Model:

Regression was selected to predict the chance of admission into the university based on different features. As the dependent variable is continuous, and linear relationship exists between chance of admittance and all the independent variables, linear regression would be the best option to test the hypothesis.

## Multiple linear regression

Multiple linear regression (MLR) was chosen to evaluate the relationship between target and predictor variables as there is more than one predictor variable. Basic equation for MLR is as follows.

Y = β0 + β1X1 + β2X2 +β3X3 +…..+ βnXn + e

The response variable or dependent variable in this dataset is chance of admit and the rest are independent variables (GRE scores, TOEFL scores, University rating, SOP, LOR, CGPA, research variables). The null and alternative hypothesis are as given below.

Null hypothesis: There is no relationship between independent variables and dependent variables or there is no effect of independent variables on dependent variable.

Alternative hypothesis: There is an effect of independent variables on dependent variable (chance of admit).

All the analysis was performed in R. Using createDataPartition() function from tidyverse package, Data was randomly split into two sets, 80% data into training set for building a model and 20% into testing set for evaluating the model. The Caret package was used to fit the MLR model, and the results were given in Figure 10.

Figure 10

Fitting Multiple linear regression Model

Call:

lm(formula = Chance.of.Admit ~ ., data = train.data)

Residuals:

Min 1Q Median 3Q Max

-0.228900 -0.022540 0.007878 0.033294 0.157764

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.402446 0.956146 -5.650 3.64e-08 \*\*\*

GRE.Score 0.583489 0.197853 2.949 0.003430 \*\*

TOEFL.Score 0.357566 0.118473 3.018 0.002755 \*\*

University.Rating2 -0.009176 0.015835 -0.579 0.562700

University.Rating3 -0.008880 0.016991 -0.523 0.601592

University.Rating4 -0.012198 0.020479 -0.596 0.551868

University.Rating5 0.003199 0.022444 0.143 0.886766

SOP -0.003697 0.005991 -0.617 0.537606

LOR 0.022047 0.005681 3.881 0.000127 \*\*\*

CGPA 0.119658 0.012766 9.373 < 2e-16 \*\*\*

Research1 0.023488 0.008595 2.733 0.006640 \*\*

---

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

Residual standard error: 0.06025 on 309 degrees of freedom

Multiple R-squared: 0.8221, Adjusted R-squared: 0.8164

F-statistic: 142.8 on 10 and 309 DF, p-value: < 2.2e-16

Since the P value for university rating and SOP was >0.05, these two parameters were considered as not significant and were removed to get better predictions. The results of the MLR model after removing University rating and SOP attributes were shown in Figure 11.

Figure 11

Fitting Multiple Linear regression model after removing two not statistically significant features

Call:

lm(formula = Chance.of.Admit ~ ., data = train.data)

Residuals:

Min 1Q Median 3Q Max

-0.23408 -0.02257 0.00808 0.03515 0.15620

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.393126 0.936702 -5.758 2.03e-08 \*\*\*

GRE.Score 0.598267 0.195705 3.057 0.00243 \*\*

TOEFL.Score 0.333569 0.114614 2.910 0.00387 \*\*

LOR 0.020315 0.004904 4.142 4.43e-05 \*\*\*

CGPA 0.120166 0.012094 9.936 < 2e-16 \*\*\*

Research1 0.022570 0.008490 2.658 0.00826 \*\*

---

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

Residual standard error: 0.06002 on 314 degrees of freedom

Multiple R-squared: 0.8206, Adjusted R-squared: 0.8177

F-statistic: 287.2 on 5 and 314 DF, p-value: < 2.2e-16

The call section shows fitting the regression model. The residuals are the difference between actual and predicted values. As median is almost centered at zero, the distribution is symmetrical. This validates that MLR model is predicting well in both high and low chances of admission. Figure 12 shows the outliers on both ends of the chart, but those on the lower end seem more severe than the upper end. Overall, the residuals look to have a normal distribution. Based on the coefficients in Figure 11, regression equation can be built for this model.

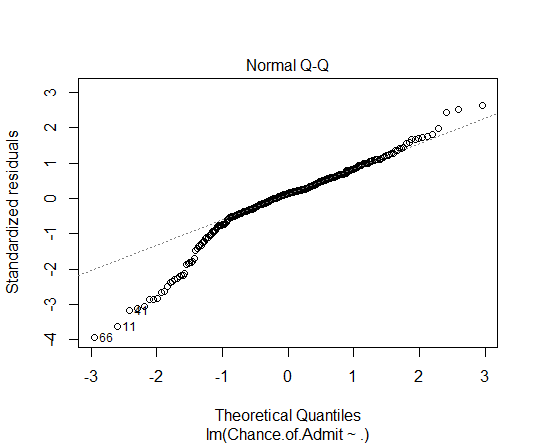
Y= -1.3223 + 0.0019 GRE Score + 0.0030 TOEFL Score + 0.0203 LOR + 0.1201 CGPA + 0.0225 Research1.

From this equation, it’s evident that if all the predictor variables are zero, chance of admission would be -1.3223, which can be inferred as no chances of admission. For every one unit increase in GRE score, COA increases by 0.0019 and for every one unit increase in TOEFL score COA increases by 0.0030 and for every unit increase in LOR, COA increases by 0.0203 and for every unit increase in CGPA, COA increases by 0.1201. For every unit increase in research, COA increases by 0.0225.

Residual standard error for this model is 0.06, which tells the average chance of admit the actual values of Y differ from predicted values. As the residual error is small in this model, which means that the model’s prediciton line is very close to the actual values, on average. Multiple r-squared value of 0.8203 means that independent variables explains 82.03% of variation in chance of admission. As the F-statistic is higher and p-value is <0.05, null hypothesis is rejected.

Figure 12

Q-Q plot



The Chance of admission was predicted on test data, using the trained model and the model performance was evaluated with regression metrics.

## Evaluation metrics

To evaluate the regression model, Mean Absolute Error (MAE), Mean Squared error (MSE), and Root Mean Squared Error (RMSE) were used (Linear Regression Essentials in R - Articles - STHDA, 2018).

Table 2

Regression metrics

|  |  |
| --- | --- |
| Metric | Value |
| MAE | 0.0535 |
| MSE | 0.0058 |
| RMSE | 0.0765 |

## Findings

An MAE of 0.0535 means that, on average, the model’s predictions are approximately 0.0535 away from the true Chance of admission into the university. An MSE of 0.0058 means that, on average, the squared prediction errors are approximately 0.0058. An RMSE of 0.0765 indicates that, on average, the model’s predictions have an error of approximately 0.0765 in the same units as the chance of admission.

# Conclusion

The model was concluded as effective as the MAE, MSE and RMSE were low. Adjusted R2 of 81.77% of variability in chance of admission is explained by the predictor variables. Thus, null hypothesis can be rejected and concluded that there is a relationship between a few independent variables on Chance of admission.

# References

Graduate Admission 2. (n.d.). Www.kaggle.com. <https://www.kaggle.com/datasets/mohansacharya/graduate-admissions>

https://datascienceplus.com/how-to-apply-linear-regression-in-r/

Linear Regression Essentials in R - Articles - STHDA. (2018, March 11). Www.sthda.com. <http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/#:~:text=A%20simple%20workflow%20to%20build>