Graduate School Admissions

Maggie Russell, Steven Lu, Claire Luo
March 16, 2021

1 Introduction

As an undergraduate student, admission into graduate school can be difficult to predict. In preparation for applying, students must prepare a noteworthy application and select a shortlist of graduate school which fit their respective academic interests. However, in selecting a list of potential graduate schools and programs, students often face difficulty in selecting schools which align with the quality of their application and thus, give them a high probability of admission. As such, here, we present a model to (1) investigate which factors may influence graduate school admission the most in an effort to help guide students in preparing their applications and (2) predict the probability of admission given a student's application profile.

2 Methods

2.1 About the Data

The data set used to build our model is available at (1). This data set contains parameters which are considered to be part of a standard application profile for a graduate school admissions committee. It was trained by a machine learning based approach, and a few values were previously obtained by students. These data contain 400 observations and 7 independent variables. The response variable is given as a probability of admission into the University of California Los Angeles Graduate School (`ADMIT`). However, with additional data, this data set and our forthcoming model could be expanded to include data from additional graduate schools. Among the remaining six variables are several scores including GRE, TOEFL, and GPA. The GRE is a standardized test that is an admission requirement for graduate schools. The total score of the test is 340. The TOEFL is a standardized test to measure the English language ability of non-native speakers. The total score of the test is 120. These data were collected and prepared from an Indian student perspective. Thus, the undergraduate cumulative GPA is on the 10-point GPA system. In addition to these scores, the data set contains additional factor variables for statement of purpose (SOP), letter of recommendation (LOR), research experience (RE), and undergraduate university ranking (UR). The statement of purpose variable is a categorical variable which indicates the strength of the statement of purpose on a scale of 1-5 with margin 0.5. The letter of recommendation variable is a categorical variable that indicates the strength of the letter of recommendation on a scale of 1-5 with margin 0.5. Research experience is given as a binary variable which indicates whether the student has no research experience (0) or has research experience (1). The undergraduate university ranking is also a categorical variable which indicates the university rating on a scale of 1-5 with margin 1. As such, with these data, we will build a logistic regression model to predict graduate school admission given an application profile.

2.2 Data Processing

All rows were kept since there were no null values or invalid data points (e.g. having a GRE score above 340).

The original data contained a chance of admission variable (`ADMIT`) is a continuous variable between 0 and 1. We transformed that variable into a binary categorical variable (`TRANSFORMED_ADMIT`) consisting of only 0 and 1 where 1 represents that the student was admitted to graduate school. We chose the cutoff to be 0.7, so all chance of admission greater than 0.7 are assigned the value 1. This new binary variable now becomes our response variable. This cutoff was chosen using the procedure described in the "Cutoff selection" section below.

The data is split into a training set (300 observations) and a test set (100 observations) so that we can have an unbiased evaluation of a final model fit on the training data set. We used the `createDataPartition` function in R with $p = 0.75$ to split the data into two parts while ensuring that each data partition contained near equal proportions of response variables (0 and 1).

2.3 Cutoff selection

In order to decide on a cutoff value for transforming the ADMIT response variable into a binary categorical variable (TRANSFORMED_ADMIT) consisting of only 0 and 1 (as described in the Data processing section above), we conducted the steps described above and below (data processing, model selection, model diagnostics, and model testing) using various cutoff values. We chose to test the following cutoff values: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9. We can see the results of testing different cutoff values below:

Cutoff value	Model accuracy	Variable segregation by binary response?
0.5	0.94	Yes
0.55	0.89	Yes
0.6	0.89	Yes
0.65	0.79	No
0.7	0.80	No
0.75	0.91	Yes
0.8	0.87	Yes
0.85	0.93	Yes
0.9	0.96	Yes

Ultimately, we chose the cutoff value of 0.7 since it maximized the model prediction accuracy while not segregating a categorical variable with a single response (i.e. overfitting).

2.4 Model Selection

We have chosen to use logistic regression analysis to build our model since our (transformed) response variable is binary. As such, we wanted to use logistic regression to build a model to predict the probability of graduate school admission. We started by including all predictor variables in our model using the `glm` function in R. As such, we began by building the model below using our training data set:

```
model <- glm(TRANSFORMED_ADMIT ~ GRE + TOEFL + GPA + SOP + LOR + RE + UR, data)
```

We used the `step` function in R to select our model using forward and backward step-wise model selection. For each of these tests, we measured the AIC and BIC of each model. In the case where different models minimized BIC and AIC, we implemented a chi-squared test for nested model comparison to compare the two models. With this test, we can select the model which fits better to our data. For example, if the chi-squared difference value is significant, then the "larger" model with more freely estimated parameters will fit the data better than the "smaller" model.

2.5 Model Diagnostics

After selecting a model, we checked for model mis-specifications using residual plots. However, residuals for logistic regression are difficult to interpret (compared to residual plots in linear regression) since the expected distribution of the data changes with the fitted values. Even re-weighting the residuals with the expected dispersion, as with deviance residuals, does not allow for easily interpretable residual plots. As such, here, we used the DHARMa package in R which employs a simulation-based approach to transform logistic regression residuals to a standardized scale. Plotting these transformed residuals allowed for us to interpret the residual plots, intuitively. For a correctly specified logistic regression model, we would expect the following:

- a uniform distribution of the transformed, scaled residuals
- constant residual variance in residual versus fitted value plot.

If these assumptions were satisfied, we continued on to test for highly influential data points.

Using the deviance residuals (instead of the transformed residuals from the DHARMa package), we analyzed the deviance residuals versus leverage plot. With this plot, we assessed whether any highly influential data points were present in the training data set. We considered any data points which have Cook's distance value above 0.5 to be "highly influential." If all data points fell below the Cook's distance value above 0.5, we proceeded.

Lastly, we used the `vif` function from the `car` package to measure variance inflation factor scores for each variable in the model. If all variance inflation scores were below 5, we proceeded with model testing.

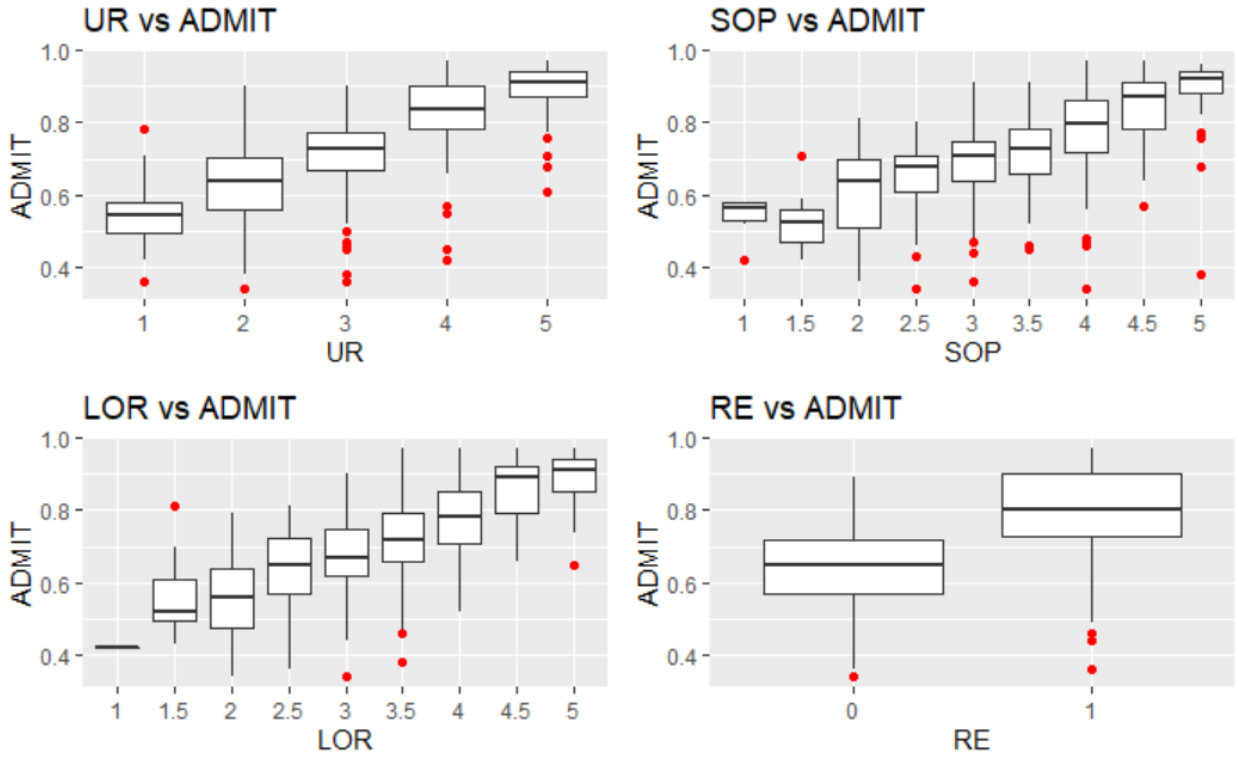
2.6 Model Testing

We used the model we built using our training data set to test how well it performs using our test data set. Using the R function `predict`, we constructed a confusion matrix. From here, we computed the accuracy by summing the number of correct classifications (admit or no admit, again, using a cutoff of 0.7) and dividing the sum by 100. The resulting score gave us the percent accuracy of our model in values in predicting graduate student admissions.

3 Results

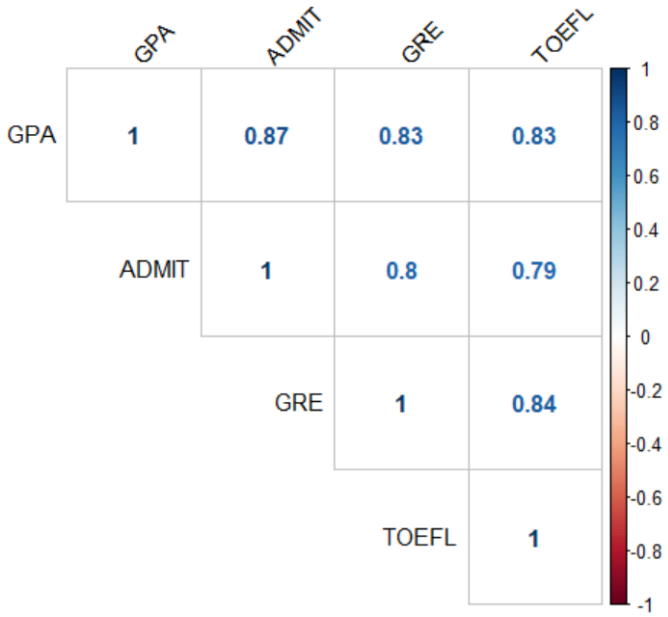
3.1 Exploratory data analysis reveals strong correlations between predictor and response variables

Prior to building our logistic regression model, we conducted exploratory data analysis to assess correlations between predictor variables. The plots below illustrate the strong, positive correlations between the categorical variables (UR, SOP, LOR, RE) and the chance of admission (`ADMIT`).



From the top-left plot, it is clear that the chance of admission is high when a candidate belongs to a high ranking undergraduate university. Based on the top-right plot, having a strong Statement of Purpose means high probability of getting admission. There are some rare cases where a candidate has a very strong SOP, but they have a below average chance to get admission. From the bottom left plot, we can see that Letter of Recommendation also has great influence towards getting admission in University. Lastly, based on the bottom right plot, we can see that having research experience generally improves the chance of admission. In general, these variables all have positive correlation with the response variable.

Next, we assessed the correlations between the numerical variables predictors (GRE, TOEFL, GPA) and the chance of admission.



Based on the correlation plot, we noted that GRE score, TOEFL score, and GPA, are highly correlated to each other. One potential explanation for these linear dependencies in our predictors is that higher GRE scores, TOEFL scores, GPA, and other factors generally result in greater chance of admission (as we discussed above). With these correlations in mind, we proceeded to build a logistic regression model.

3.2 Model selection procedures suggest a model with a subset of predictor variables

As discussed in the Methods section, we used the `step` function in R to select our model. We began with a full logistic model to predict the probability of admission using all possible predictor variables (GRE, TOEFL, GPA, statement of purpose, letter of recommendation, research experience, and undergraduate university rating). We will refer to this model as the "full model". Using backward step-wise model selection, we obtained a model containing the following predictors: GRE, statement of purpose, GPA, and research experience. We will refer to this model as the "GRE + SOP + GPA + RE" model. Lastly, using forward step-wise model selection, we obtained a model containing the following predictors: GRE, research experience, and GPA. We will refer to this model as the "GRE + RE + GPA" model. These three models have the following selection criteria scores:

Model	Residual Deviance	Df	AIC	BIC
`Full Model`	143.7503	275	193.7505	286.3448
GRE + SOP + GPA + RE	153.3688	288	177.3688	221.8141
GRE + RE + GPA	173.2073	296	181.2073	196.0224

Clearly, the "GRE + SOP + GPA + RE" model has the lowest AIC score and the "GRE + RE + GPA" model has the lowest BIC score. Because of this discrepancy, we utilized a chi-squared test for nested model comparison to compare these two models to decide which model is a better fit to the data. The chi-squared difference value had a significant p-value of 0.01. As such, we can conclude that the larger model, which is the "GRE + SOP + GPA + RE" model, is a better fit to the data. Thus, we will proceed in using the "GRE + SOP + GPA + RE" model for our remaining analyses. The output from the chi-squared test for nested model comparison is shown below:

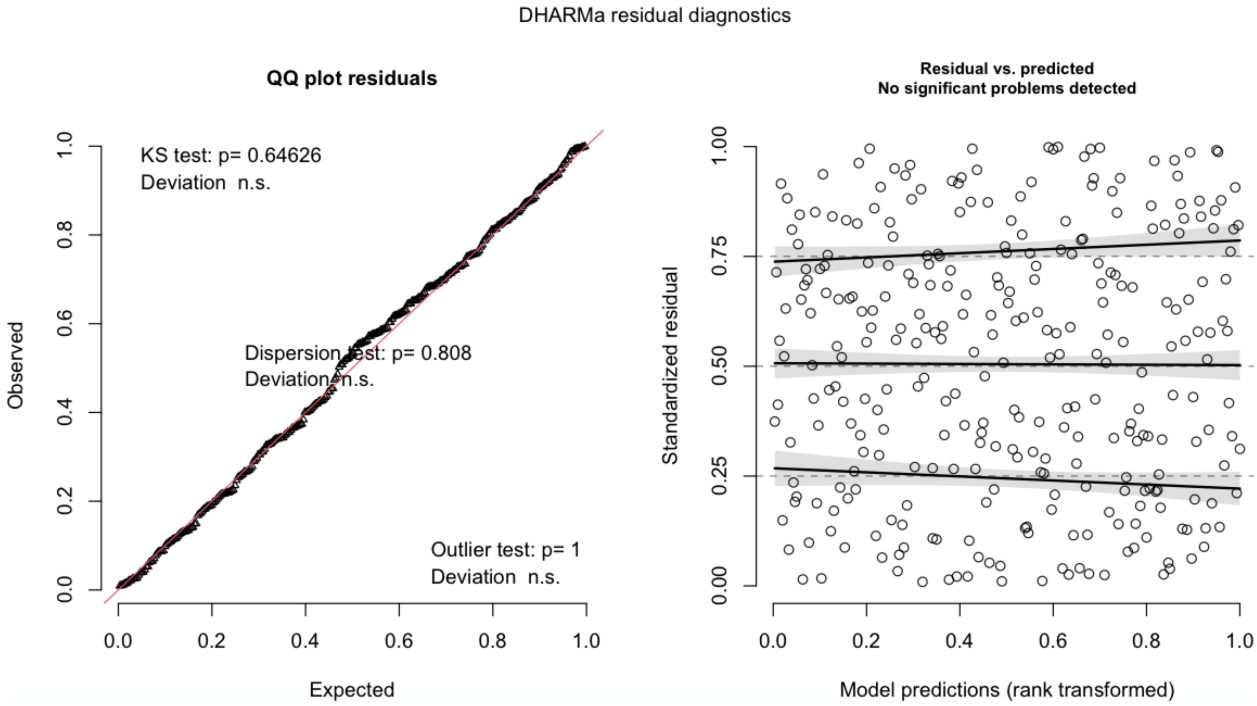
```
Analysis of Deviance Table

Model 1: ADMIT2 ~ GRE + as.factor(SOP) + GPA + as.factor(RE)
Model 2: ADMIT2 ~ GRE + GPA + as.factor(RE)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       288     153.37
2       296     173.21 -8  -19.838  0.01096 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 The selected model satisfies logistic regression assumptions

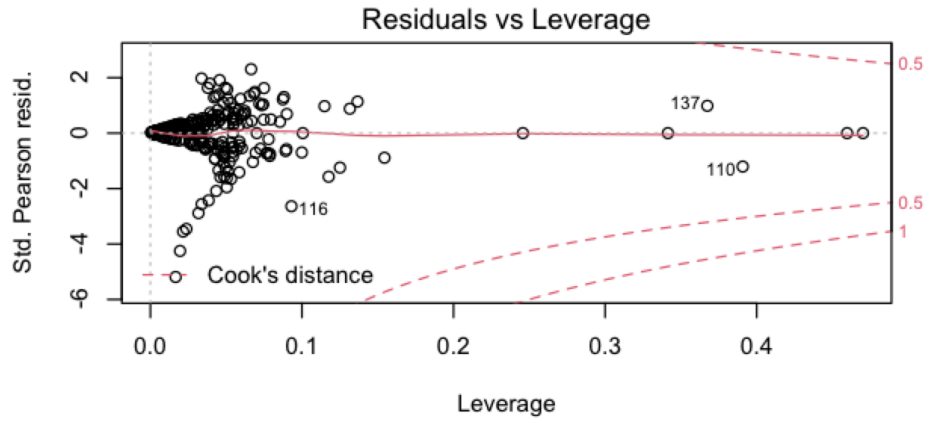
With the "GRE + SOP + GPA + RE" model, we wanted to check whether it satisfied the assumptions of logistic regression. Logistic regression requires the response variable to be binary. Using a cutoff value of 0.7 (selected as described in Methods), we transformed the continuous response variable (chance of admission) into a binary response (0 and 1) in order to satisfy this regression requirement prior to fitting the model. As such, the model satisfies this requirement.

Next, using the DHARMa package (as described in Methods), we transformed the logistic regression residuals to a standardized scale. With these transformed residuals, the DHARMa package allowed us to test the interpretable model assumptions: (1) whether the transformed residuals were uniformly distributed and (2) whether there was constant residual variance in the residual versus fitted value plot. We can check these assumptions in the plots below:



The figure on the left shows a qq-plot comparing the transformed residuals to a uniform distributed. With this figure, we can conclude that the transformed residuals follow the expected uniform distributed (KS test pvalue = 0.64626). The figure on the right shows a transformed residual versus model prediction plot. The model plot shows constant residual variance across prediction values. As such, we can conclude that the "GRE + SOP + GPA + RE" model is correctly specified.

Next, we looked for outliers in the data which, if present, could alter the model fit. To check for influential data outliers, we can plot the leverage versus standardized residuals plot below:



Because all points lie within the 0.5 Cook's distance line, we can conclude that there are no influential data outliers.

Lastly, logistic regression requires there to be little or no multi-collinearity among the independent variables. As such, we measured the variance inflation factors of all of the independent variables in the final model ("GRE + SOP + GPA + RE" model). For each variable, we got the following variance inflation factor scores:

Predictor Variable	VIF Score
GRE	1.342
SOP	1.273
GPA	1.414
RE	1.135

Because each variance inflation score above is below five, we can conclude that there is little to no multi-collinearity between the independent, predictor variables in the model.

3.4 GRE is the most important predictor of graduate school admission

Using the final "GRE + SOP + GPA + RE" model to predict the binary response binary variable of admission using GRE, GPA, research experience, and statement of purpose as predictors, the response binary variable of admission, we got the following results:

Variable	Estimate	Standard Error	z value	P value
Intercept	-77.24	3604.21	-0.02	0.9829
GRE	0.13	0.03	4.18	2.9e-05
SOP (1.5)	-0.32	3893.77	0	.9999
SOP (2)	15.23	3604.20	0	.9966
SOP (2.5)	15.72	3604.20	0	.9965
SOP (3)	15.59	3604.20	0	.9966
SOP (3.5)	15.27	3604.20	0	.9966
SOP (4)	15.80	3604.20	0	.9965
SOP (4.5)	15.73	3604.20	0	.9965
SOP (2.5)	16.15	3604.20	0	.9964
GPA	2.39	0.71	3.37	0.0007
RE (1)	0.63	0.41	1.55	0.1207

The p-values for each coefficients suggests that GRE is the most important predictor of graduate school admission chance, followed by GPA.

3.5 The model can predict admission with good accuracy

With the final "GRE + SOP + GPA + RE" model fit using the training data set, we wanted to test how well the model could predict chance of admission using the testing data set. Using the methods described in the Methods section, we constructed the following confusion matrix:

	Predicted: not admitted	Predicted: admitted
Actual: not admitted	38	6
Actual: admitted	14	42

With this matrix, we computed a model accuracy of 0.8.

4 Discussion

In the above analysis, we were able to address our original two research questions. First, we discovered that GRE is the most important predictor of graduate school admission chance, followed by GPA. However, from our exploratory data analysis, we also observed that each predictor (even those not included in the final model) has positive correlation with admission. Thus, we can conclude that while all predictors could affect the admission result, GRE score and GPA appear to be the more dominant predictors.

As shown above, our model has an 80% accuracy for predicting a candidate's chance of admission, which is fairly convincing. However, we must consider that this data set, and the resulting models, are more suited to international students, mainly Indian students. For example, the TOEFL test is only required for international students who did not study in English-based universities. Students who study at English-based universities are not required to take a TOEFL exam and, thus, their TOEFL scores are not considered in their graduate school admission. Additionally, the GPA scale for Indian universities on a ten-point scale instead of the commonly used 4.0 scale in the United States. As such, this model was trained and tested using data containing ten-point-scale GPAs. In order to apply this model more broadly, we would need to keep these limitations in mind. However, with further training and testing using additional data, we may be able to extend this model to a more broad audience. Nevertheless, this current model can still be used as a reference for students who want to know the most important predictors of graduate school admission and, perhaps, predict their chance of admission before they actually applied to a graduate school.

References

[1] Mohan S Acharya, Asfia Armaan, Aneeta S Antony. *A Comparison of Regression Models for Prediction of Graduate Admissions*, IEEE International Conference on Computational Intelligence in Data Science 2019.