Data Science II Midterm Report

Dayoung Yu (dry2115), Justin Hsie (jih2119), Christian Pascual (cbp2128)

3/18/2019

1 Introduction

Applying to graduate programs is a harrowing process for any student. Between the statement of purpose, letters of recommendation, GRE and GPA, there's a lot of factors that can influence the acceptance or rejection of a hopeful student. Given data on the application components, predicting a student's chance of poses an interesting regression problem. From a student perspective, the ability to predict chances of admission would allow them to ground their expectations and plan for the future.

The *Graduate Admissions* dataset from Kaggle contains data on about 500 Indian students applying for Master's programs in the United States. The response variable we hope to predict is `Chance of Admit`. The potential predictors are various Master's application components converted into continuous or categorical form, including: GRE score, TOEFL score, university rating, statement of purpose strength, letter of recommendations strength, cumulative GPA and the presence of research experience.

1.1 Data Cleaning

```
library(tidyverse)
library(glmnet) # LASSO
library(mgcv) # GAMs
library(modelr) # cross-validation help
library(earth)
library(caret) # cross-validation help
library(kableExtra)
library(corrplot)

# Loading in the dataset
admit.data = read.csv(file = "./admission_predict.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(
    gre.std = (gre_score - mean(gre_score)) / sd(gre_score),
    toefl.std = (toefl_score - mean(toefl_score)) / sd(toefl_score),
    cgpa.std = (cgpa - mean(cgpa)) / sd(cgpa),
    uni.rating = university_rating,
    sop.strength = sop,
    lor.strength = lor
  ) %>%
  dplyr::select(gre.std:lor.strength, research, chance_of_admit)
```

gre.std	toefl.std	cgpa.std	uni.rating	sop.strength	lor.strength	research	chance_of_admit
1.8174175	1.7770857	1.7750286	4	4.5	4.5	1	0.92
0.6664808	-0.0315692	0.4853733	4	4.0	4.5	1	0.76
-0.0417879	-0.5248388	-0.9530883	3	3.0	3.5	1	0.72
0.4894137	0.4617003	0.1546925	3	3.5	2.5	1	0.80
-0.2188550	-0.6892620	-0.6058734	2	2.0	3.0	0	0.65
1.1976824	1.2838162	1.2624733	5	4.5	3.0	1	0.90

In its raw form, the dataset only needs minimal formatting before we can start modeling. GRE, CGPA and TOEFL are already continuous and don't need coercing, but they will be centered and scaled. For university rating, statement of purpose, and letter of recommendation strength, these variables are seemingly categorical, but they will be treated as continuous for easier interpretation and to reduce the number of dummy predictors. The presence of research was recoded as a proper binary variable.

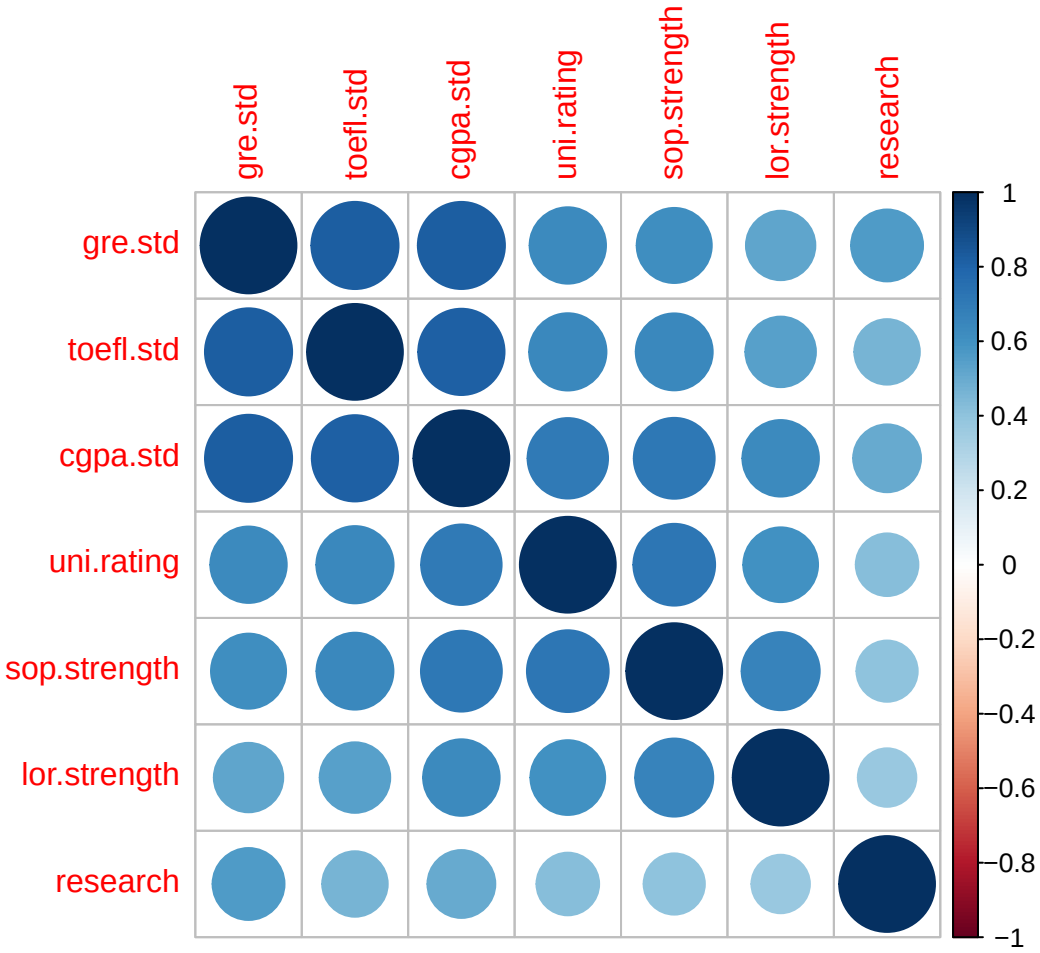
We will attempt to predict `Chance of Admit` using a total of 5 models: 3 linear models, ordinary linear regression, ridge regression and LASSO, and 2 non-linear models, a generalized additive and MARS model. We will be using the implementations in `lm`, `glmnet`, `gam` and `earth` to do the modeling.

This final product of seeks to create these models, compare them, and hopefully recommend one for use by future college hopefuls.

2 Exploratory Data Analysis

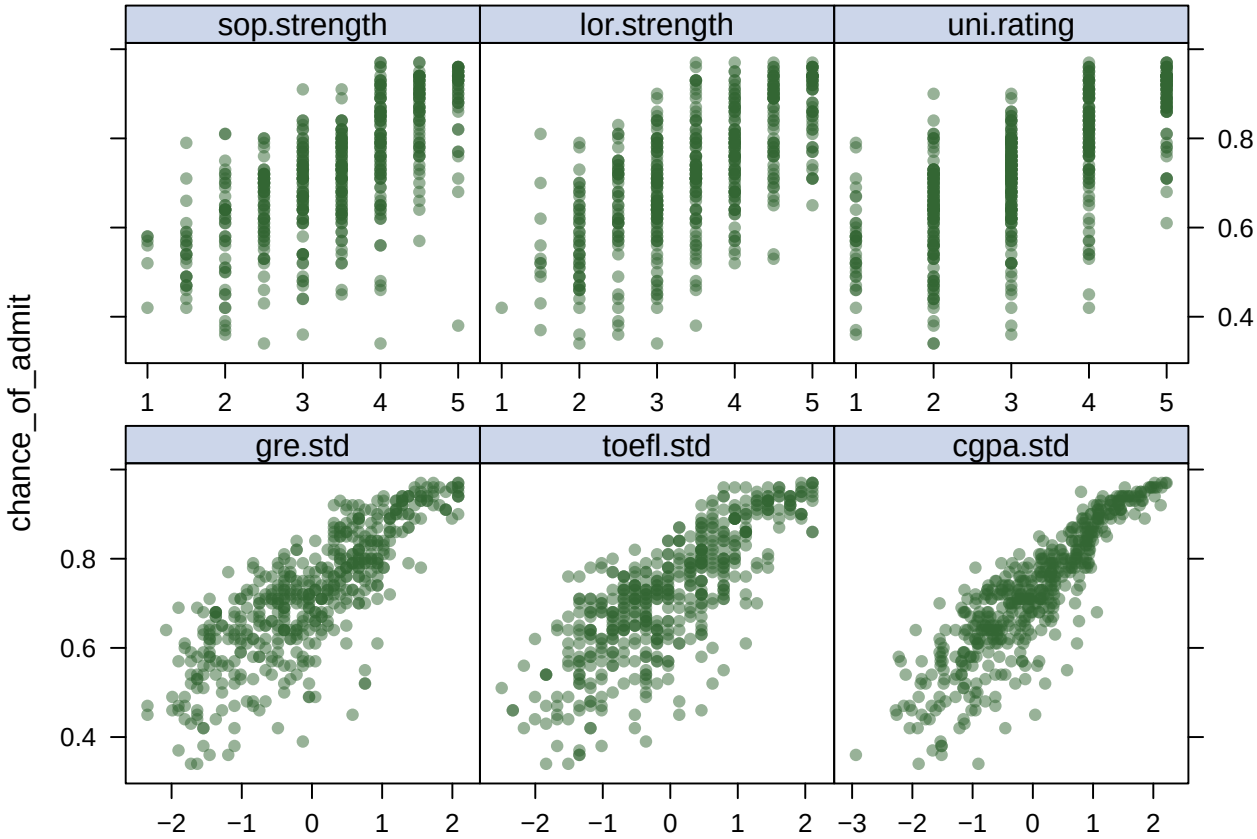
Before starting the modeling, we explored our dataset to see how each of the predictors related to the response and to each other. Our findings here motivated the inclusion of the models in our report.

2.1 Intercorrelation



The correlation matrix above shows that all predictors have moderately high correlation between each other ($\rho > 0.5$). The highest correlation occurs between `gre.std`, `toefl.std`, and `cgpa.std`. We expected this result

since applicants with high GPAs are also more likely to score well on other tests. Including all three `gre.std`, `toefl.std`, and `cgpa.std` predictors may result in multicollinearity and inflated coefficient estimates. In response to this, we decided to explore how a LASSO and ridge regression could curb the effect of this correlation and improve predictions.



The scatterplots above indicate the existence of a positive relationship between all predictors and `Chance of Admit`. We also saw that the the range of our desired response value only goes from about 40% to 95% in the dataset. Statement of purpose strength, letter of recommendation strength, and university rating all have approximately linear relationships with the chance of admission. With GPA, GRE, and TOEFL scores, we noticed a slight plateauing of admission chances at the higher end of the scores. With this non-linearity in the data, we hope that the use of a GAM and MARS model will capture this subtlety and improve prediction.

3 Models

Our dataset only contains 7 predictors, so we will incorporate all of them into each of our 4 models. Each of these factors are requested in most Master's program applications, so we will assume all will have an impact on predicting the chance of admission.

3.1 Linear Models

We plan to use ordinary linear regression as a *baseline* model to compare the other 4 models with. Although we suspect that all the variables will have an appreciable impact on the chance of admission, we still want to allow for the possibility that some predictors do not truly contribute. Furthermore, we also found high correlation between many of our predictors in our exploratory data analysis, so we also want to adjust for potential inflation of our regression coefficients. Therefore, we plan to use the ridge and LASSO models

Table 1: Average Test Fold MSE between the 5 candidate models

Model	Avg. Test MSE
LASSO	0.0036848
Linear	0.0036883
GAM	0.0036954
Ridge	0.0037299
MARS	0.0038899

to shrink the coefficients and perform variable selection to adjust for these findings and hopefully improve predictions.

3.2 Non-linear Models

In our exploratory data analysis, we also found that some of our predictors (CGPA and TOEFL) had a slight nonlinear relationship with the chance of admission. We hope to use a GAM model and a MARS model to better capture these nonlinear relationships and produce improved predictions as a result.

3.3 Model Tuning

We used `cv.glmnet` to find the optimal λ for both the ridge and LASSO models via 5-fold cross-validation. For the GAM, we found the optimal smoothing parameters via generalized cross-validation. In order to tune our MARS model, we created a tuning grid spanning from 1 to 3 degrees and from 1 to 50 terms. Then, we used `caret` to choose the best set from this grid.

3.4 Findings

For each of the models, we used 10-fold cross-validation to evaluate the average test fold MSE as a measurement of predictive ability. Table # below shows a comparison of the average test fold MSE for each of the 5 models.

The LASSO model performed the best out of the 5 models, although all of them are comparable since they have similar average test fold MSEs. The difference between the best and worst models is on the 1^{-5} scale, which would have a negligible difference in terms of predicting a student's chance of admission.

3.5 Important Variables

In our basic linear and LASSO model, CGPA and GRE score exert the strongest influences over a student's chance of admission. The ridge model resulted in coefficients that suggest that all variables affect the chance of admission on a similar scale; each of the variables increases the predicted chance of admission by about 1%. Conversely, the strength of a student's statement of purpose and the rating of the university have the least influence on the chance of admisison. This finding makes sense since we would expect factors that reflect a student's work and testing ability to have a greater impact on admission. The table below summarizes our findings with the linear models.

3.6 Limitations

The ridge model, GAM and MARS model were chosen to try to maximize the predictive ability, but each of these models is limited in interpretability. Attemptingt to explain what each coefficient means would be difficult, compared to the ordinary linear regression or LASSO model. Our data contained predictors that

Table 2: Coefficient estimates of the three linear models

Linear	LASSO	Ridge
0.6255	0.6281	0.5930
0.0210	0.0209	0.0163
0.0169	0.0166	0.0156
0.0716	0.0717	0.0191
0.0059	0.0058	0.0105
0.0016	0.0014	0.0118
0.0169	0.0165	0.0129
0.0243	0.0235	0.0204

had approximately linear relationships with chance of admission, so we believe that our models were flexible enough to capture the complexity in the data.

4 Conclusion