Data Science II Midterm Report

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

3/18/2019

Contents

1 Introduction

Applying to graduate programs is a harrowing process for any student. Between the statement of purpose, letters of recommendation, GRE and grades, there's a lot of components 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. 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. This report seeks to create, evaluate and select a predictive model that takes these covariates and predicts a student's chance of admission. We hope to figure out what factors are the most crucial (or unhelpful) elements to improving one's chances of admissions.

1.1 Data Cleaning

```
library(tidyverse)
admit.data = read.csv(file = "./admission_predict.csv") %>%
  janitor::clean_names() %>%
  mutate(
    uni.rating = factor(university_rating, labels = c(1, 2, 3, 4, 5), ordered = TRUE),
    sop.strength = factor(sop, ordered = TRUE),
    lor.strength = factor(lor, ordered = TRUE),
    cgpa = cgpa * 100, # interpret unit increase as a hundredth increase
    has.research = ifelse(research == 1, TRUE, FALSE)
  )
```

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 transformation. For university rating, statement of purpose,

and letter of recommendation strength, these variables were converted into ordinal variables to properly characterize their categorial nature. The presence of research was recoded as a proper binary variable.

Since `chance_of_admit` represents a probability of admission, we will attempt to model it using standard logistic regression as well as logistic-LASSO. We will be using the implementations in `glm` and `glmnet` to do the modeling. Before we start this process, we will explore how each of our predictors are interrelated and related to the response.

2 Exploratory Data Analysis

3 Modeling

4 Conclusion

5 From Canvas:

~~Introduction~~

~~Describe your data set. Provide proper motivation for your work.~~ ~~What questions are you trying to answer? How did you prepare and clean the data?~~

Exploratory data analysis/visualization

Is there any interesting structure present in the data? What were your findings? Here you can use any techniques as long as they are adequately explained. If you cannot find anything interesting, then describe what you tried and show that there isn't much visible structure. Data science is NOT manipulating the data in some way until you get an answer.

Models

What predictor variables did you include? What technique did you use, and why did you choose it? What assumptions, if any, are being made by using this technique? If there were tuning parameters, how did you pick their values? How did you make your predictions? Discuss the training/test performance if you have a test data set. Which variables play important roles in predicting the response? What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

Conclusions

What were your findings? Are they what you expect? What insights into the data can you make?