STA363SecAProj2

Rickey Huang

3/16/2021

Abstract/ Executive Summary

Section 1: Data Cleaning

In this section, the data set is explored and cleaned in order to improve the quality of the data for a good prediction.

Section 1.1: Cleaning the missing data

First of all, any missing data are inspected. However there is no missing data in this college data set. Since the header for the first column which shows the names of colleges are missed, I add a header (*"College"*) to a copy of the original data set I created and use this copy for the analysis below.

Section 1.2: Adjusting variables

From the infromation provided, I learn that the number of student enrolled in the colleges are usually not easily to be collected, the column stored this information is deleted from our data set. Also, since the acceptance rate is a more appropriate varaible than the numebr of acceptance, a new collumn named *"Rate"* is created using the existing variables acceptance *"Accept"* to be divided by the number of applications per academic year *"Apps"*. After adding this new variable to our data, since it is perfectly correlated to the variable *"Accept"*, the old and incomparable variable *"Accept"* is deleted. After arranging the variables, I got the data set for the analysis in this project, which has 777 observations and 18 variables, and among them only the variable *"Private"* is a categorical variable with two levels. Since the goal of this project is to predict the number of applications received during an academic year, the variable *"Apps"* would be the response variable in this project, and all other variables except the names of universities would be the exploratory variables. For the convenience of the analysis, I removed the column storing the college names.

Section 2: Selection Only

In order to have a comparatively precise prediction in the end, several models are fitted compared in this project. This section focuses on the selection-only Least Square Linear Regression (LSLR) model, which also implements the Best Subset Selection (BSS) technique to refine the variables we have.

Section 2.1: Best Subset Selection - Stage 1

In the first stage of the BSS, all possible models containing 1 variable, 2 variables, and all the way to the full models (with 16 exploratory variables here) is created. R^2 is used to determine the best models among the models using the same amount of variables.

Section 2.2: Best Subset Selection - Stage 2

Proceeding to the second stage, I compared how well models created in the stage 1 are. Since our goal for this project is prediction, the AIC would be a good metric for comparing.

```
##  [1] 380.265419 134.241699  46.963788  33.035074  20.953791  15.058922
##  [7]   9.339390   7.322785   7.608040   7.463554   7.936910   9.563652
## [13]  11.260814  13.024936  15.012856  17.000000
```

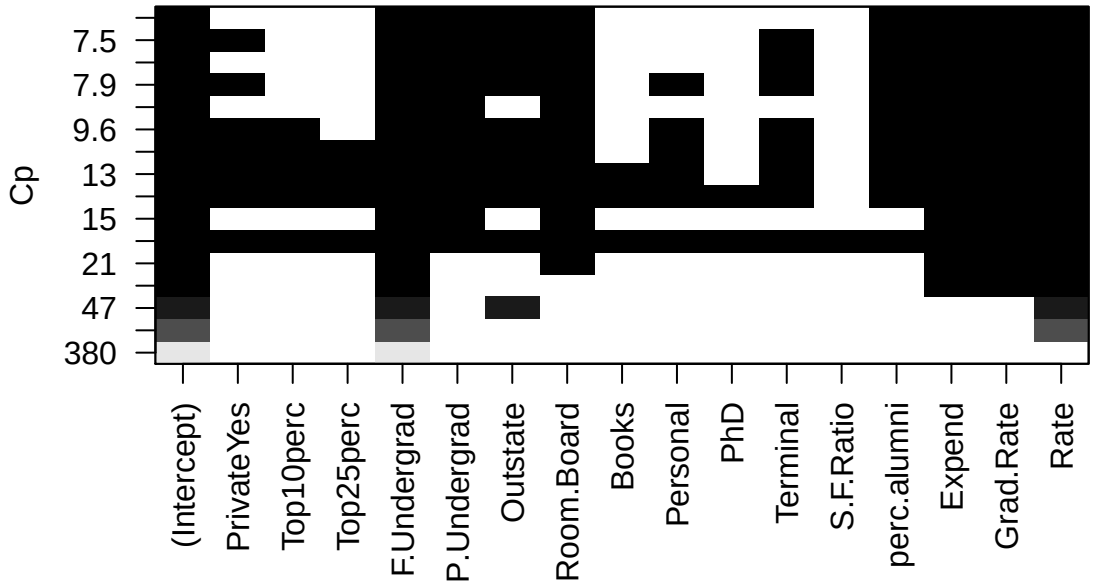

Figure 1: AIC's for the models created from the stage 1

Gathering the AIC's computed from the models, the Figure @ref(fig:BSSAIC) is created to visualize the result. Since we want a model with a smaller AIC and less variables, the model with features *"F.undergrad", "P.Undergrad", "Outside", "Room.Board", "perc.alumni", "Expend", "Grad.Rate", "rate",* and the intercept with an AIC of 7.322785 is the best fit.

Section 3: Shrinkage Only

Section 4: Selection and Shrinkage

Section 5: Elastic Net

Section 6: Conclusion