Yan Zhang

Assignment 2.

Advanced Data Mining and Analytics.

February 8, 2017

For this assignment, we are trying to find the best model for the variable median graduate debt, which is the GRAD\_DEBT\_MDN\_SUPP in the College Scorecard dataset. There are 122 variables in total.

The first step is clean the data, since we are going to build the model by liner regression, which means we need the variables with numeric. Hence I deleted the binary variables, which are the variables only contain 0 and 1. Also I deleted the variables are none numeric, such as “Institution name”, “City” and “Institution URL”. In the rest of the variables. I tried the find the variables that are related to the median graduate debt manually. For example, from my observation, PCIP (1-54) are not relate to the median graduate debt, because these means the percentage of degrees awarded in each two-digit CIP code field of study. And CIP provides a structure in which to track and report in fields of study. So these are not related to what we are trying to predict. For the variables like SAT score, I only kept the average score.

After the variable selection, there are 31 variables left. Then I replaced all the NA value with median, by calculate the correlation, we find the PREDDEG has the highest correlation with median graduate debt, which means how much debt you can get is very depend on your degree type. In the final model, there are 23 variables left. Also there are 8 variables are highly relate to the median graduate debt. We can find it from the AIC value, such as the SAT score.

The final model is:

GRAD\_DEBT\_MDN\_SUPP = PREDDEG + LOCALE + RELAFFIL +

UGDS + PPTUG\_EF + NPT4\_PUB + NPT4\_PRIV + NPT41\_PUB + NPT42\_PUB + NPT43\_PUB + NPT44\_PUB + NPT45\_PUB + NPT41\_PRIV + NPT43\_PRIV + NPT44\_PRIV + NPT45\_PRIV + PCTPELL + RET\_FT4 + RET\_FTL4 + RET\_PTL4 +PCTFLOAN + UG25ABV + GRAD\_DEBT\_MDN10YR\_SUPP