Midterm Project Check In Report

1. Specific project goals

The overall objective is to design and evaluate prediction of high poverty stricken schools from the publicly available data across the U.S., and to evaluate the value of Title 1 school designation. The resulting model might grant insight into the factors critical to the success of school districts.

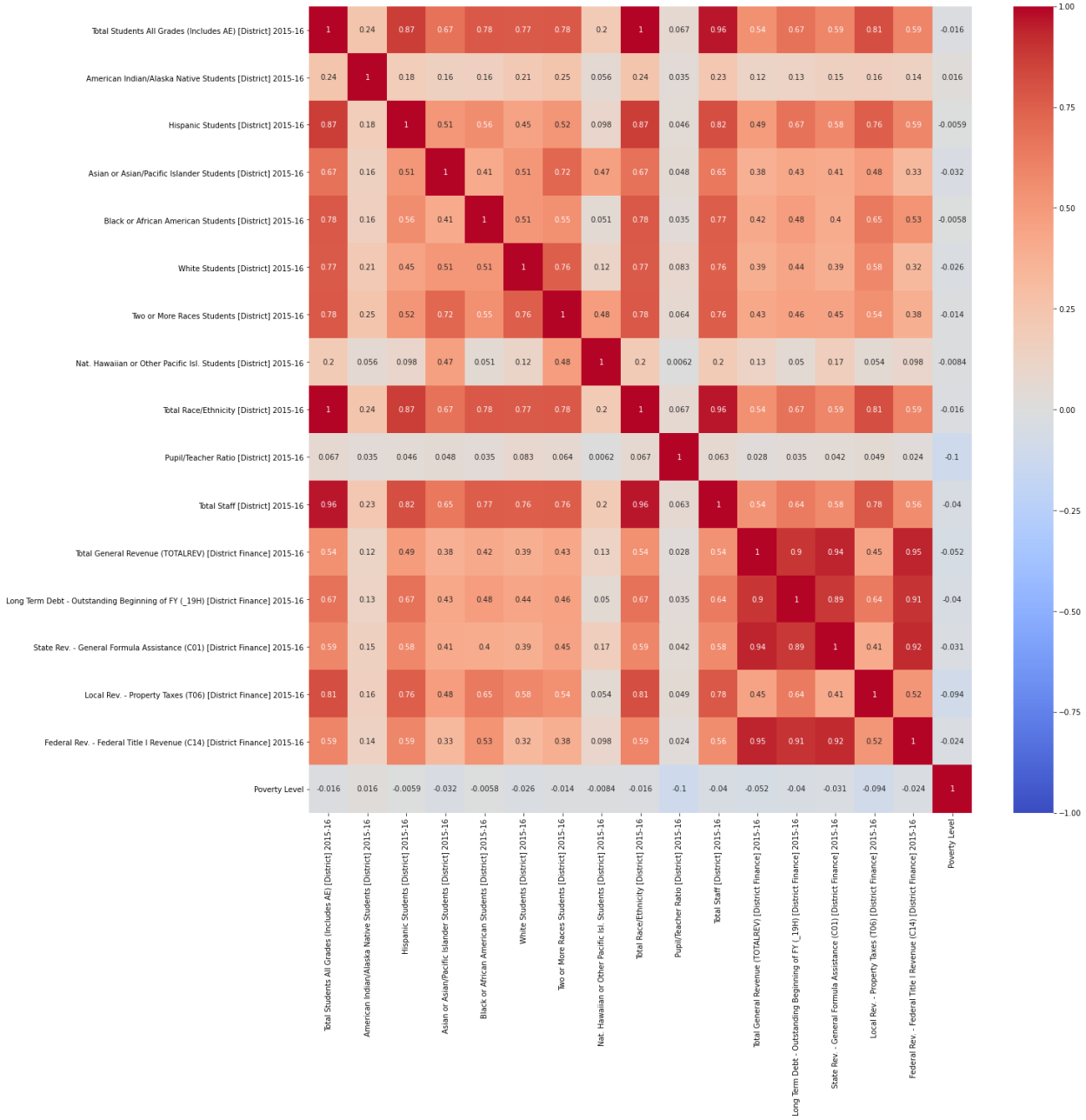
2. Data gathering (what data do you have, figure on label distribution in training and test set)

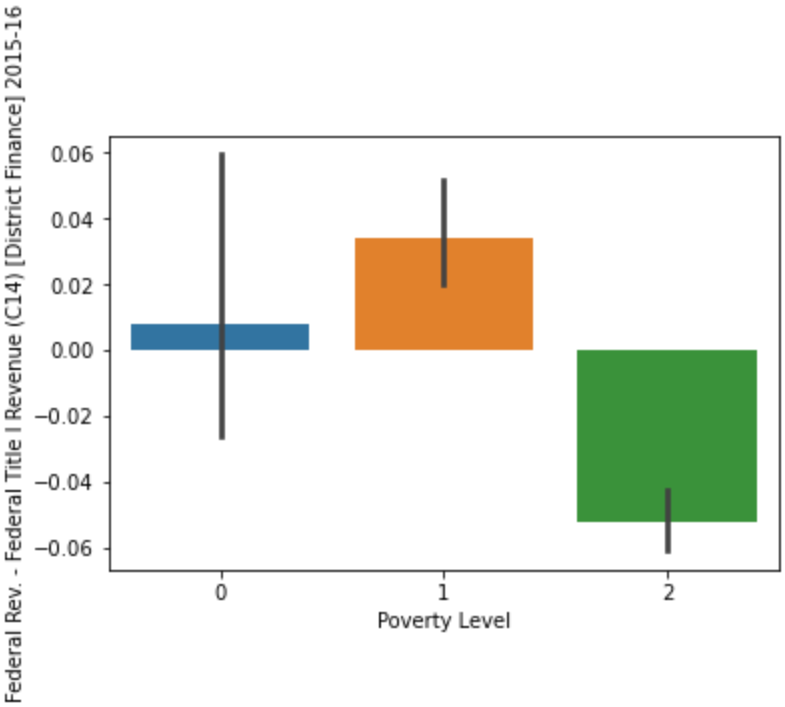
We have data for 3 school years from 2015 to 2018. Schools began reporting more varieties of data in 2015 than previous years, and 2018 is the latest year fiscal data is available. Thus, this selection allows for the greatest range of data.

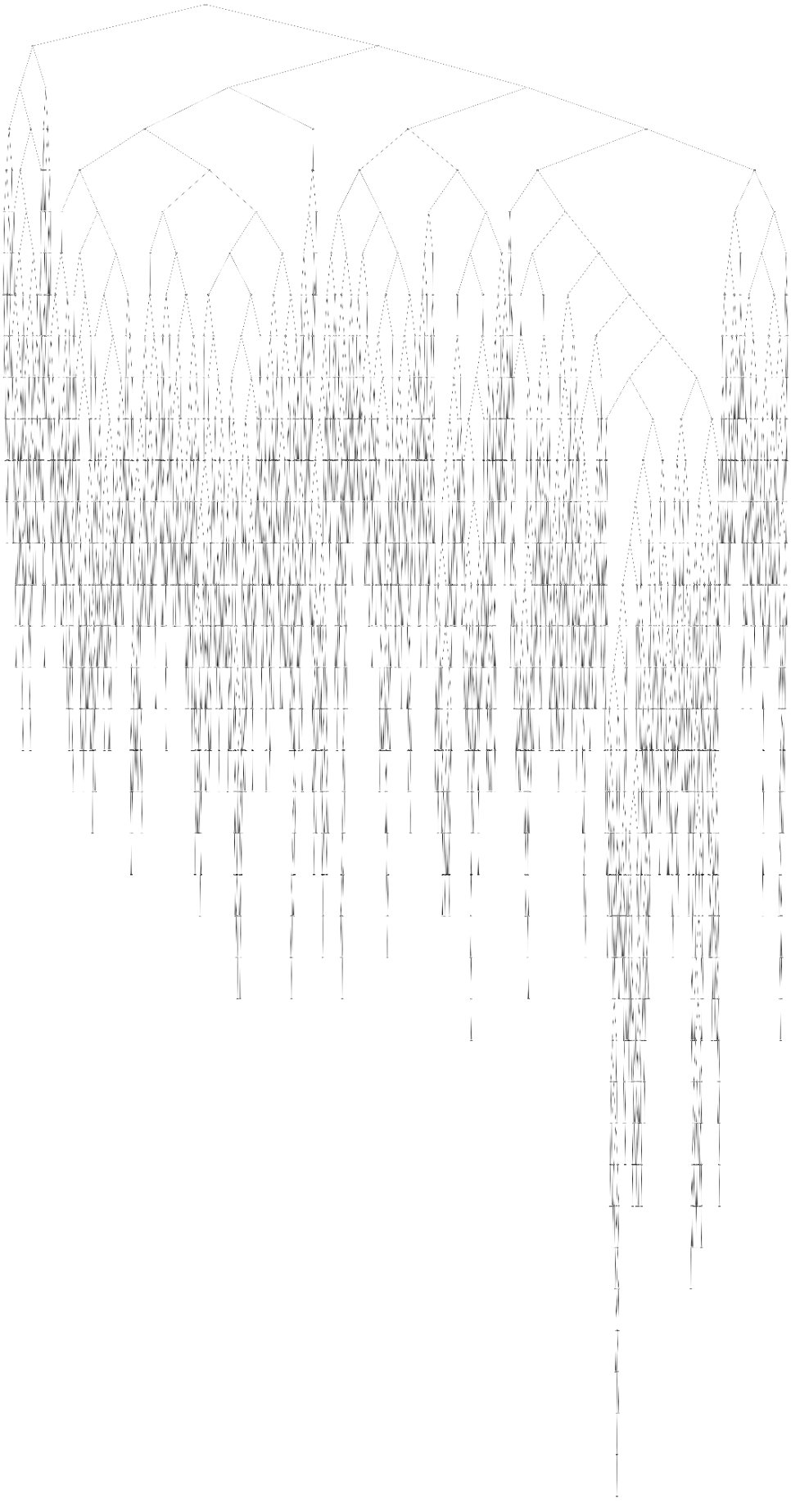
3. Data Cleaning (what do you do with NaN values)
The pandas isna.sum() method was used to check for any existing NA values. And the existing NA values were replaced with 0 using the fillna() method.

4. Feature Normalization and Selection: What features are you using? What normalization feature selection will you employ?
Features: Total students all grades, Total General Revenue, Total Race/Ethnicity

5. Figures and Exploratory Data Analysis
Figures: Heatmap to analyze correlation


Total Students All Grades (Includes AE) [District] 2015-16
American Indian/Alaska Native Students [District] 2015-16
Hispanic Students [District] 2015-16
Asian or Asian/Pacific Islander Students [District] 2015-16
Black or African American Students [District] 2015-16
White Students [District] 2015-16
Two or More Races Students [District] 2015-16
Nat. Hawaiian or Other Pacific Isl. Students [District] 2015-16
Total Race/Ethnicity [District] 2015-16
Pupil/Teacher Ratio [District] 2015-16
Total Staff [District] 2015-16
Total General Revenue (TOTALREV) [District Finance] 2015-16
Long Term Debt - Outstanding Beginning of FY (_19H) [District Finance] 2015-16
State Rev. - General Formula Assistance (C01) [District Finance] 2015-16
Local Rev. - Property Taxes (T06) [District Finance] 2015-16
Federal Rev. - Federal Title I Revenue (C14) [District Finance] 2015-16
Poverty Level
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Federal Rev. - Federal Title I Revenue (C14) [District Finance] 2015-16
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
1
2
Poverty Level



6. Attempt at Data Modeling.

We used a basic linear regressor as a baseline model. It performed poorly, but better than random with a testing score of 0.38147570792025276 (⅓ chance is random). We also trained a decision tree classifier which scored perfectly on the training sample and 0.7160850170558909 on the testing sample. The score of 1 on training data might indicate that some feature is directly correlated with our label, or that this model overfits. The idea that it overfits is supported by the image of the tree above, which shows a massive set of decision nodes.