Springboard Data Science Capstone Project 1

Factors Influencing ACT Scores

Meral Balik

January 30, 2020

Contents

Capstone Project 1 - Milestone Report

Factors Influencing ACT Scores

1. Problem Statement

The ACT test is designed to assess high school students' general educational development and their ability to complete college-level work. ACT scores are also used to compare schools and districts on the basis of their average score and to evaluate state educational programs. Students can prepare for and take the test several times to attain their desired score. ACT test covers four subject areas: English, mathematics, reading, and science. Each subject area test receives a score ranging from 1 to 36.

In this study, I investigated the factors related to schools' performance on the ACT. In order to achieve this, I addressed the following questions.

1. What is the distribution of the ACT scores?
2. Which high schools received the highest and lowest overall ACT scores?
3. Which counties have the highest and lowest performing schools?
4. On which subjects are the schools performing better?
5. Is there an increase of the success level of schools with time?
6. What kind of factors influence success levels?
    - Ethnicity
    - Number of low income students
    - New teacher ratios
    - Population of the schools

6. Are the parents' and GreatSchools' ratings related to school success?

2. Dataset

The ACT data to be used is obtained from the California Department of Education (https://www.cde.ca.gov/ds/sp/ai/). Each dataset consists of a row for every accredited high school in California with its code number, school name, county name, enrollment based on the number of students in grade twelve, number of students, the average english, math, reading and science scores, the number of students who scored 21 points or more, the percent of students who scored 21 points. The data covers the period from 2014 through 2018.

The other two datasets showing the census information about the high schools in California are obtained from greatschools.org (https://www.greatschools.org/california) website. The census datasets include information about ID number, longitude, latitude, city, county, number of students enrolled, percentage of different ethnic backgrounds, percentage of teachers who are in their first or second years, percentage of low income students who have the right to receive reduced priced lunch of the high schools.

2.1. Data Wrangling

Data Wrangling Part 1:

For the ACT data, I had five different datasets belonging to different years ranging from 2014 to 2018. Before taking the cleaning steps, I added year columns to some of the datasets with no year column information.

1. **Cleaning steps:**
    a. To have consistent column names, I cleaned the extra white spaces in the column names and capitalized them to concatenate the datasets from all years.
    b. I constructed new columns by trimming the redundant ones.
    c. I formatted the year column and set it as index.
    d. I corrected the incompatible value types in columns.
    e. To eliminate empty spaces in the text of some columns, I used col.astype(str).str.strip() method.

2. **Missing Data:** I handled the missing data in two steps.
    a. In the data, the missing values were shown with ‘ * ’ sign. I changed them using na_values=['*'].
    b. Then, I dropped the rows with too many Nan values using .dropna(thresh=7) method.

3. **Outliers:** I applied the .describe() method on the data and inspected the outcomes. The max and min values of the columns were compatible with the other numbers.

Data Wrangling Part 2:

I had three csv files in this section: The ACT file, gsid file (school ID information used to retrieve other data from greatschools website) and the census data that I acquired using api request. The main process in this notebook was merging the datasets with as much common information as possible. The act data had school and county names, the gsid data had school and city names, and the census data had only the school names column that I could use in the merging process. Thus, I added county names to the gsid data with gsid['City'].map(city_cnty) method using the excel file which includes city names and their corresponding counties. Then, using the school ID numbers, I merged the census and gsid datasets and created census_id data with the command shown: census_id = pd.merge(census, gsid, left_on='Id', right_on='gsId', how ='inner', suffixes = ('','_gsid')).

Some school names were spelled differently in the ACT file and census_id data. To merge these files, I used the fuzzywuzzy library string matching. Another issue was that there were multiple schools with the same name in California. To fix this issue, I matched the schools names only if they were in the same county. I merged the ACT data with census_id data based on two columns (school names and county names) using the final_df = pd.merge(act, census_id, left_on=['Sname2', 'Cname'], right_on=['Sname', 'County'], how ='inner',suffixes = ('','3')) command.

1. **Cleaning steps:**

    a. I corrected the incompatible value types in columns (strings, integers, etc.).
    b. Then I removed the school names that caused duplicates in the final data.
    c. There were a few school names which were mismatched. Manually, I created a dictionary to correct the mismatched school names. Then, I dropped the redundant columns.

2. **Missing Data:** In the final data there wasn't much missing data because of the major cleaning steps that I've performed in the first data wrangling part. In the final data, I replaced the missing values with Nan and left them as is.

3. **Outliers:** I applied the .describe() method on the final data to be able to inspect the outliers. In the 'Enrollment' column, the maximum value seemed very high. So, setting up a threshold value for the number of students, I obtained the school names with high enrollment number using final_df[final_df['Enrollment']>4500] code. Those schools were the largest high schools in California according to the information on the website, and I kept them as is. The other columns didn't have any outliers.

3. Findings from Exploratory and Statistical Data Analysis

Through exploratory and statistical data analysis, significant relationships were found between several variables.

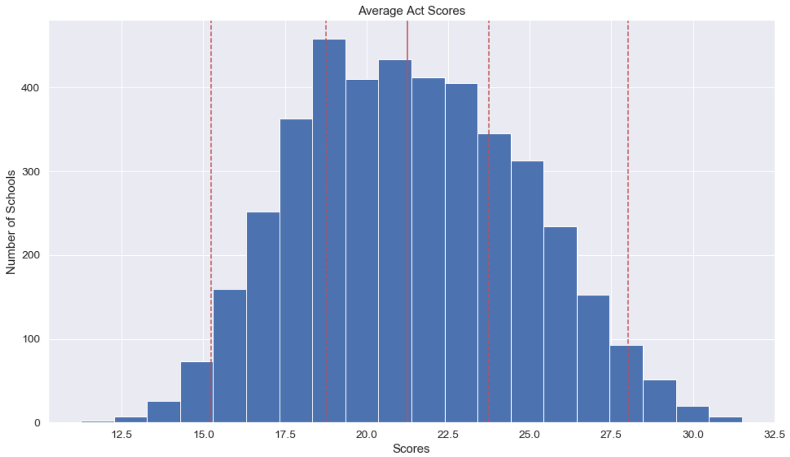


Figure 1. *Distribution of average ACT scores*

Figure 1 illustrates the distribution of the average ACT scores of the high schools. Red vertical lines represent the percentiles of the scores. Corresponding values for the 2.5, 25, 50, 75, 97.5 percentiles are: 15.25 18.75 21.25 23.75 28. It is clear that the scores are normally distributed.

About 4.5% of the schools received 28.00 (top 2.5 %) or above, and 6.3% of the schools received 15.00 (2.5%) or less at least once between 2014 and 2018. More detailed information about the calculations can be found in the Exploratory data analysis notebook.

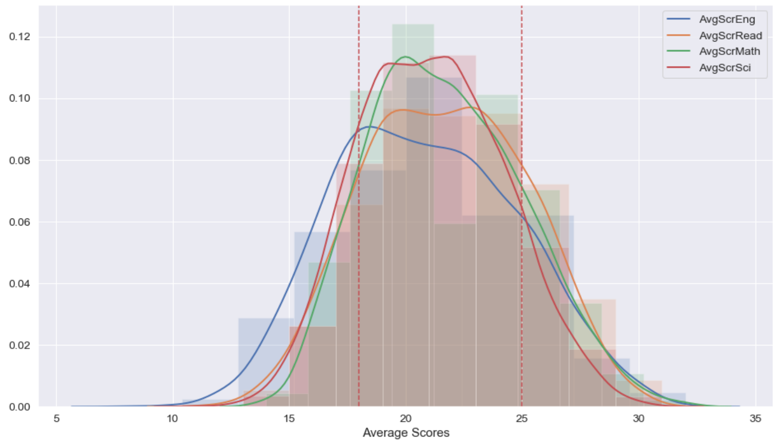
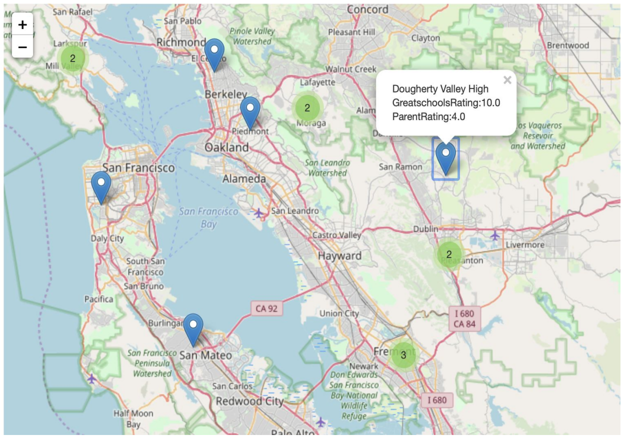


Figure 2. *Distribution of subject test scores*

Figure 2 illustrates the distribution of the scores in each subject test. The schools in the low score range (≤18) perform better in English. The schools in the middle score range (18-25) perform better on the Math and Science sections, worse in English. The schools in the high score range (>=25) perform worse on the Science section. About 60% of the schools received about 22 or less in English, and about 40% of the schools received 22.0 or more in Science. This number is less for the other subject tests.


Dougherty Valley High
GreatschoolsRating:10.0
ParentRating:4.0

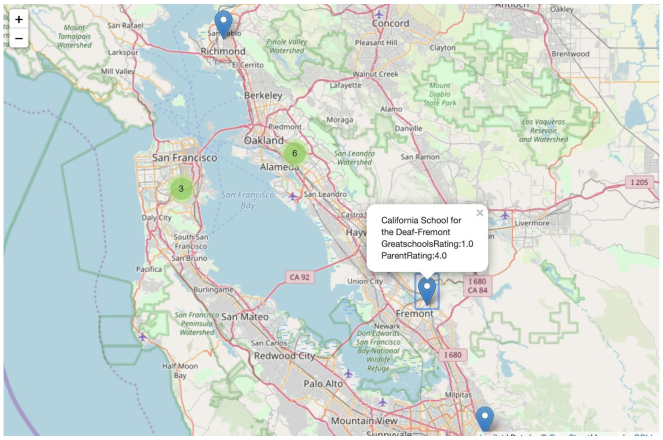

California School for
the Deaf-Fremont
GreatschoolsRating:1.0
ParentRating:4.0

***Figure 3.** Highest and lowest performing schools on map*

Figure 3 shows the highest and the lowest performing schools on California map. These maps did not show a specific pattern about the locations of the schools. The highest performing schools were in the following counties: Santa Cruz, Alameda, Contra Costa, Marin, Sacramento, San Francisco, San Diego, Orange, San Mateo, Santa Clara, Los Angeles, and Nevada. The lowest performing schools were in the following counties: San Joaquin, Santa Clara, Kings, Alameda, Contra Costa, Tulare, Riverside, Fresno, Sacramento, San Francisco, San Diego, San Bernardino, and Los Angeles.

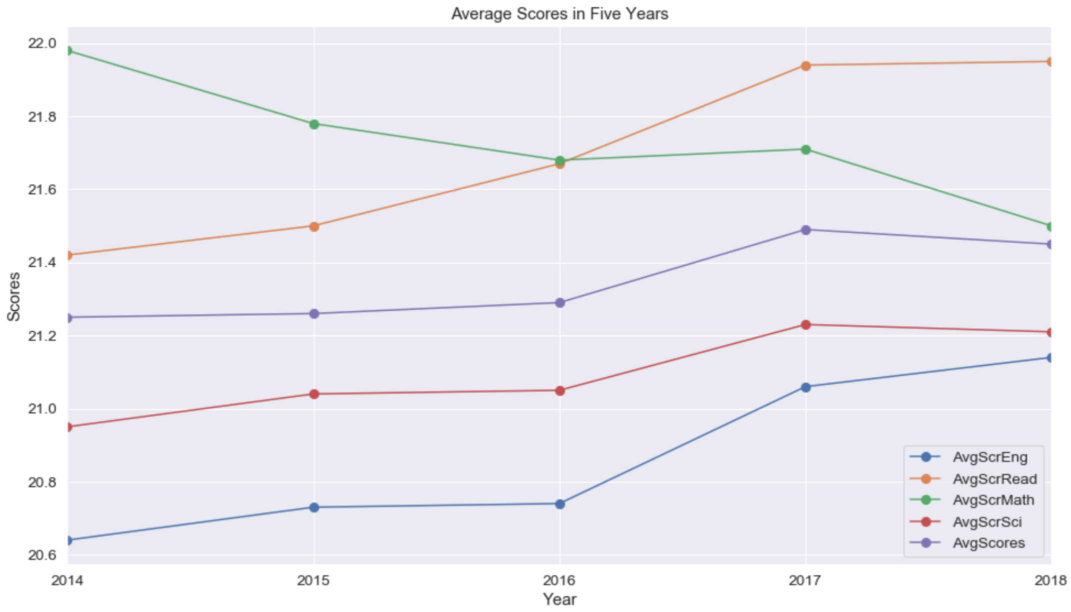


***Figure 4.** Average scores in five years*

Figure 4 summarizes the trend of the scores in each subject test with respect to the year. In general there is an increase in the majority of the subject scores . While average reading, english and science scores were increasing over the five years, the average math scores continued to decrease. If we average all of the scores, we can see the increase from 2014 to 2018.

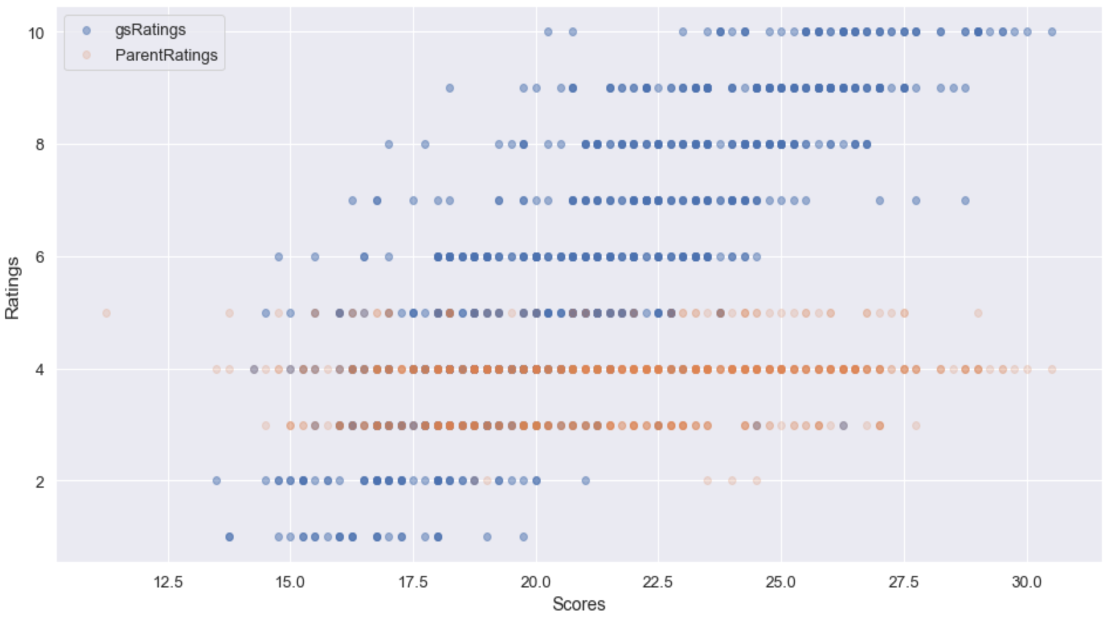


Figure 5. *Parent and Great Schools ratings*

Figure 5 illustrates the parent and greatschools.org ratings. There is a positive linear relationship between the GreatSchools ratings and the ACT scores. As the average ACT scores are increasing, the GreatSchools ratings are also increasing, but this is not the case for parent ratings. The parent ratings are in general ranging from 4 to 6. Even the schools with high scores received a relatively low rate from parents. So it can be concluded that there is no correlation between the parent ratings and the success level of a school.

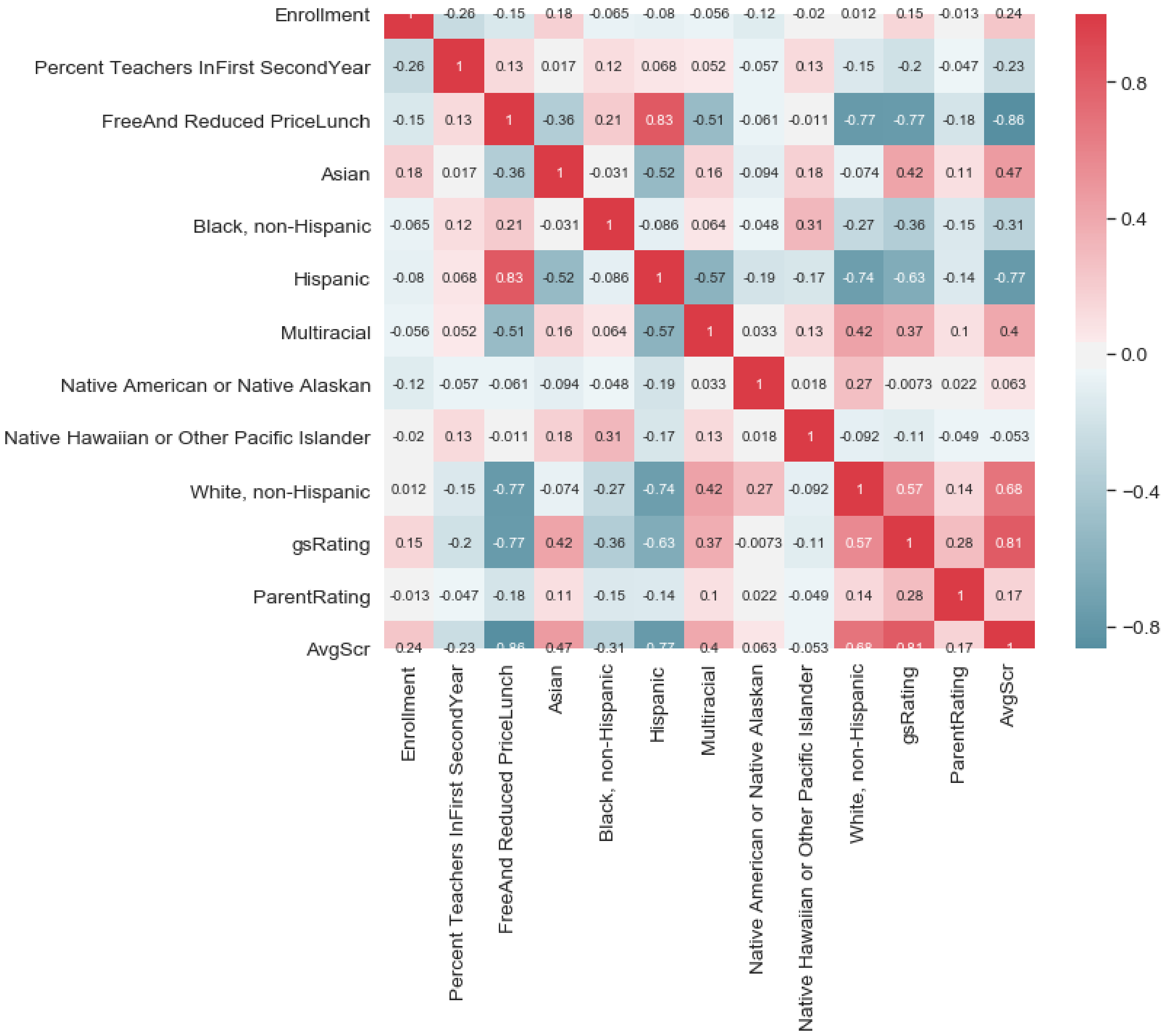


Figure 6. *Correlation Matrix*

Figure 6 is the correlation matrix that depicts the correlation between all possible pairs of values in a table. We can summarize the dataset and identify and visualize patterns. In the confusion matrix above, red color represents the positive correlation and green color represents the negative correlation between the variables. As the colors get darker, this is indicative of an increase in the correlation. Important insights from the above matrix is as follows:

The percentage of Hispanic students is positively correlated to the number of students who receive free and reduced price lunch.The percentage of White, non-Hispanic students is negatively correlated to the number of students who receive free and reduced price lunch.

The GreatSchools ratings are negatively correlated to the number of students who receive free and reduced price lunch, but positively correlated to the average Act scores.

Overall, average Act scores are negatively correlated to the number of students who receive free and reduced price lunch, and also are negatively correlated to the percentage of Hispanic students.

4. Machine Learning Analysis

The purpose of the machine learning analysis is using the several futures predicting the success level of a school. So the initial approach would be the determination of these success levels.

In the first section of the machine learning analysis, I approached the problem as an unsupervised learning problem. Using K-Means Clustering method on the average ACT scores from each section (English, Math, Reading, Science), I obtained classes that group the schools based on their ACT scores.

In the second section, I approached the problem as a supervised learning problem. The classes obtained from the first section constitutes the labels of the target . Using the other features in the data such as the percentage of the students who come from different ethnic backgrounds, the percentage of students who receive reduced and free meals, whether a school is public or charter, I try different machine learning algorithms to place the schools into the groups which were created based on the success level.

4.1. Unsupervised Learning

K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. The fundamental step for this unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. In order to find the best number of clusters, I used the elbow method and the Silhouette analysis.

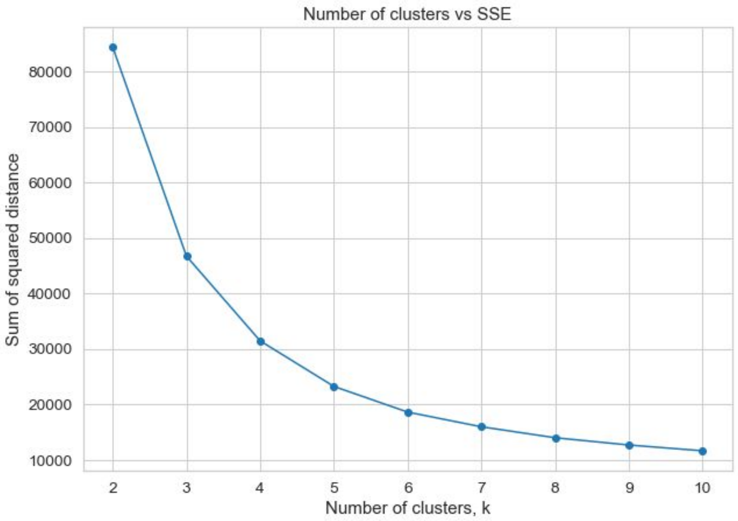


Figure 7. *Number of clusters using elbow method*

Figure 7 illustrates the elbow method which gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. This method generates a plot of the sum of squared distances for k in a specified range . If the plot looks like an arm, then the elbow on the arm is optimal k. The graph below showed that k= 3 or 4 might not be good choices. It's still hard to figure out a good number of clusters to use, because the curve is monotonically decreasing and not showing a good elbow point or has an obvious point where the curve starts flattening out. I also wanted to get a more concrete result using Silhouette analysis.

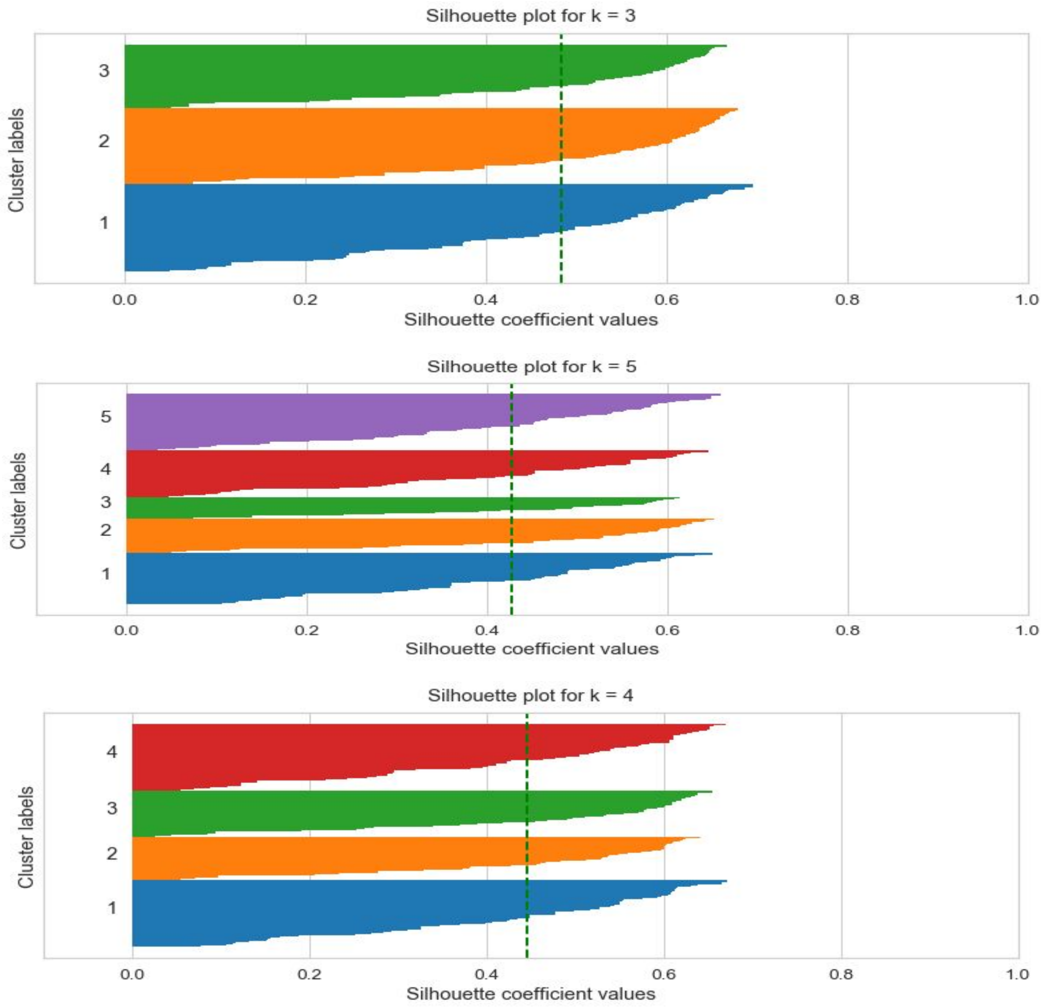


Figure 8. *Silhouette Analysis*

Figure 8 summarizes the Silhouette analysis which was used to determine the degree of separation between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters. I applied the Silhouette method to pick the best number of clusters among 3, 4 and 5. For k=3, Silhouette coefficient was 0.48, which was the highest among the other scores.

I applied the K-Means algorithm for n_clusters = 3, and obtained the classes that represent low, medium and high success levels of schools based on the ACT scores.

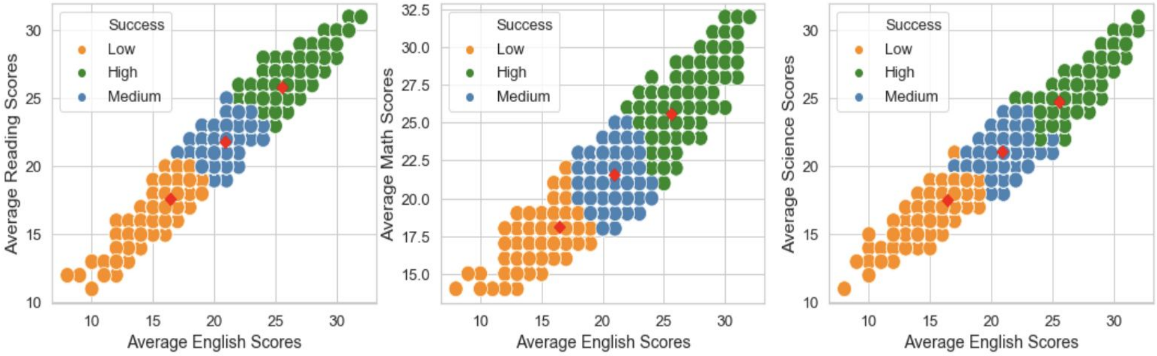


Figure 9. *Visualization of clusters*

I also visualized the schools according to their coordinates to see if there was a relation between the location and the success level of the school.

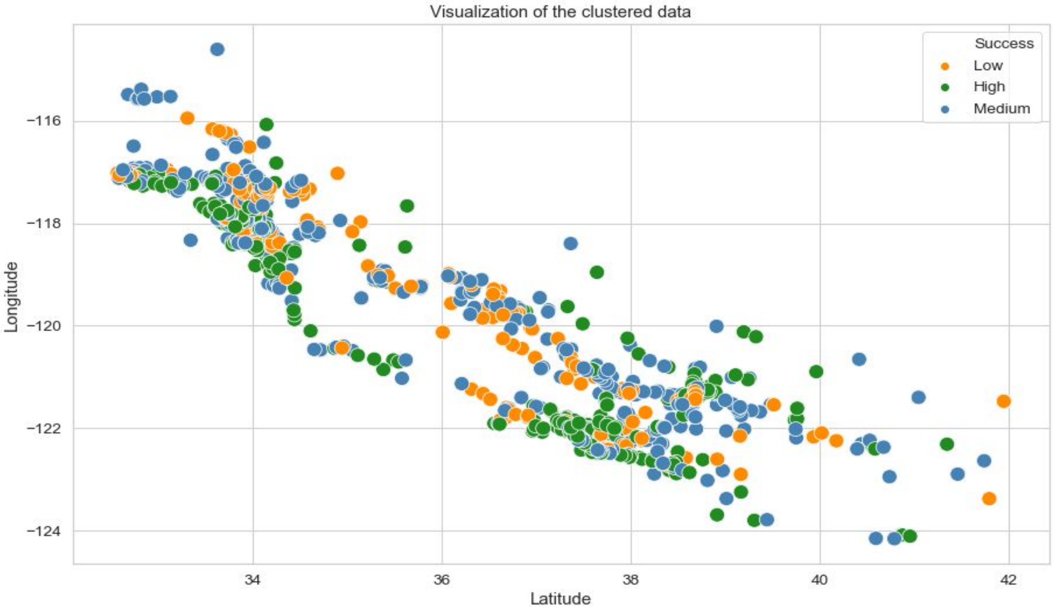


Figure 10. *Visualization of schools according to their coordinates*

Figure 10 shows the schools on the coordinate system. Orange, green and blue colors represent the schools with low, medium and high success levels respectively. It is hard to find a specific pattern looking at the schools' locations. For a more explicit visualization, I created stacked bar graphs to show the proportion of the schools with different success levels in each county for the SCT scores in 2018. This was more helpful in terms of seeing the counties with more or less successful schools.

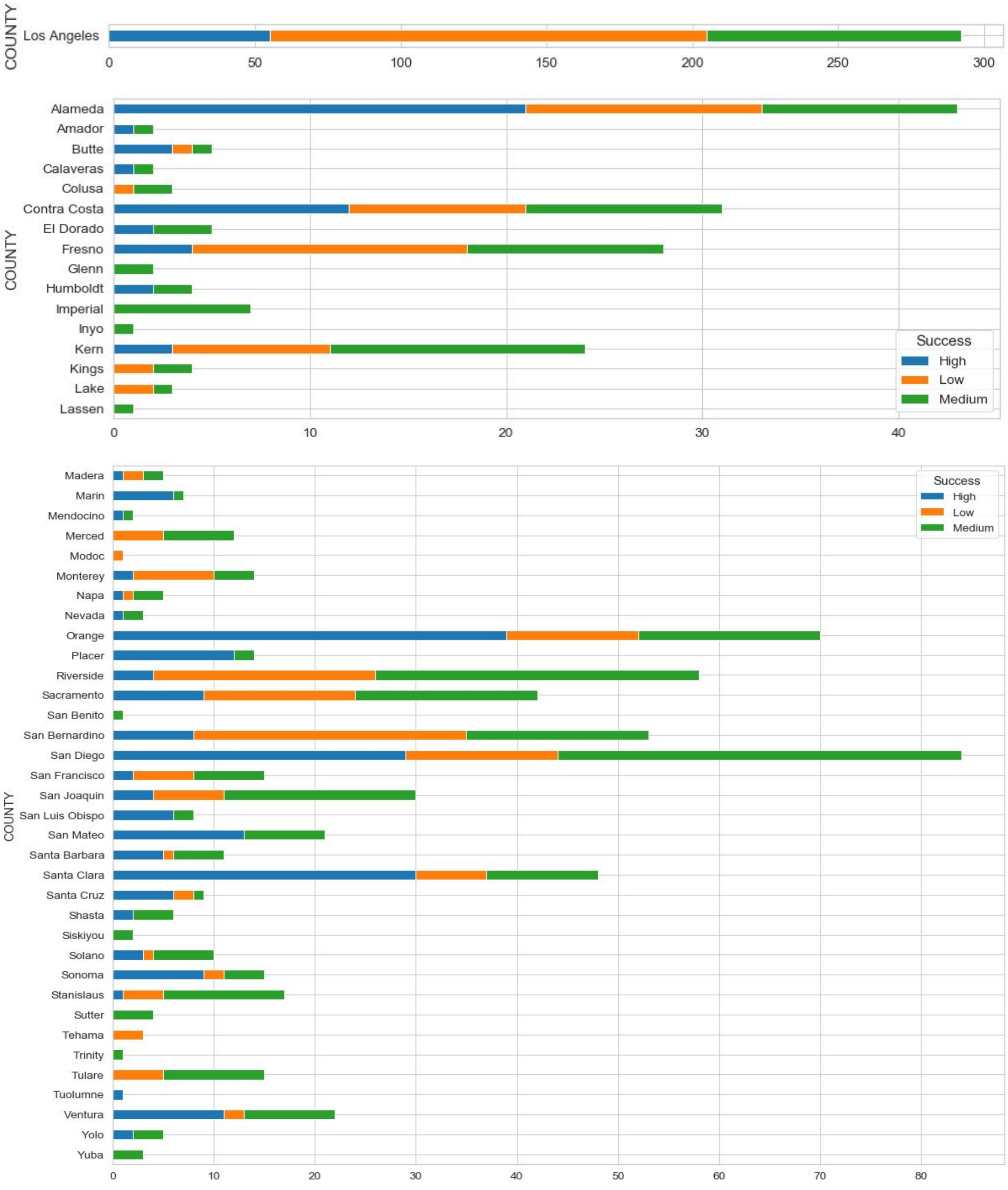


Figure 11. *Performance of schools in each county*

In the above graph, there are some counties which have high percentage of orange color meaning that they contain more schools with low success.

4.2. Supervised Learning

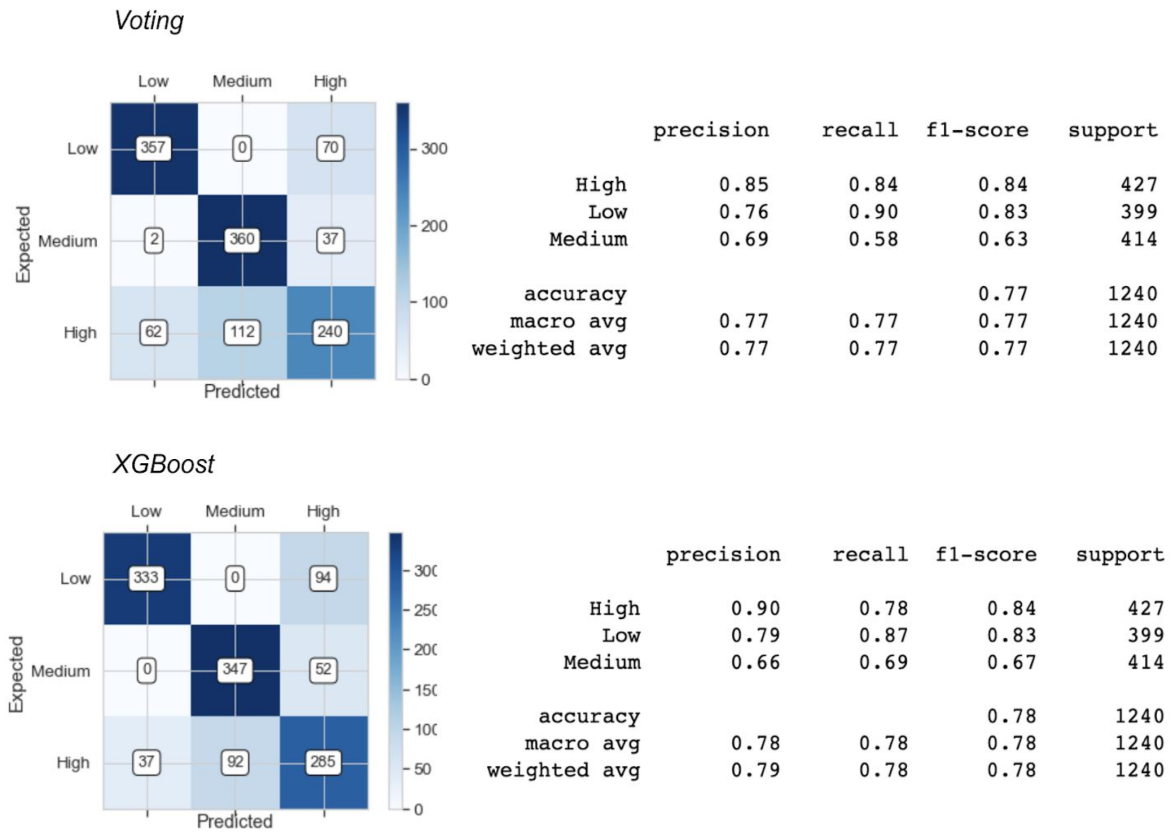
The aim of this section is to predict the success level of a school based on the percentage of the students who come from different ethnic backgrounds, the percentage of students who receive reduced and free meals, whether a school is public or charter. The first step for preparing the data for analysis was converting the values in the categorical column to numerical values. IThen I visualized the distribution of the classes by creating a categorical graph showing the number of schools in each class. Number of schools in each class was not equally distributed. This might cause a problem, because the classifier learns the classes with more samples better and remain weak on the smaller classes. So I used SMOTE technique to balance the number of schools in each class.

SMOTE (Synthetic Minority Oversampling Technique) consists of synthesizing elements for the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors. In each cluster, there was 2056 data points after applying this technique.

In order to begin applying machine learning algorithms data needs to be split into one train and one split set. In the data that I have, I had information about the schools starting from 2014 and ending in 2018. I split the data into the train set using the information from 2014 to 2017 and the test set using the information only from the year 2018.

Target and features are respectively success level of a school, the percentage of American Indian or Alaska Native, Asian, Pacific Islander, Filipino , Hispanic or Latino, African American, White , Two or More Raced students, the percentage of students who receive free or reduced lunch meals and whether a school is public or charter.

Since there are three classes for the machine learning algorithms to predict, I used the algorithms that approaches the problem as a multiclass problem. I applied Logistic Regression, KNeighbors Classifier and Decision Tree Classifier as a first approach. Then I used ensemble models Voting, Bagging, Random Forest Classifier and XGBoosting respectively. For most of the models, I used GridSearchCv for hyperparameter tuning. For some models, the parameters that I obtained from GridSearchCv resulted in a really high accuracy score on the train set and relatively lower score on the test set which was a sign of overfitting. Then I changed the parameters to be able to reduce the difference of the accuracy scores on the train and test set, but at the same time trying to keep the accuracy as high as possible. For each model, I calculated the accuracy score, obtained the confusion matrix and the classification report for a better comparison of the models. Calculating a confusion matrix can give a better idea of what the classification model is getting right and what types of errors it is making.



```
              precision    recall  f1-score   support

        High       0.85      0.84      0.84       427
         Low       0.76      0.90      0.83       399
      Medium       0.69      0.58      0.63       414

    accuracy                           0.77      1240
   macro avg       0.77      0.77      0.77      1240
weighted avg       0.77      0.77      0.77      1240
```

```
              precision    recall  f1-score   support

        High       0.90      0.78      0.84       427
         Low       0.79      0.87      0.83       399
      Medium       0.66      0.69      0.67       414

    accuracy                           0.78      1240
   macro avg       0.78      0.78      0.78      1240
weighted avg       0.79      0.78      0.78      1240
```

Figure 12. *Confusion Matrix and Classification Report*

Figure 12 shows the confusion matrix and the classification report of the best performing two models. In the first row of the confusion matrix, the first column indicates how many 'Low's were predicted correctly, and the second column, how many 'Medium's were predicted as 'Low's, and the third column, how many 'High's were predicted as 'Low's. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

For example in the confusion matrix for Voting model, there are 427 points in the first row, the model was successful in predicting 357 of those correctly as a low class, but 70 were marked as high class. There are a number of misclassifications in 'High'' class.

In the classification report, precision tells us, for all instances classified positive, what percent was correct. Recall is for all instances that were actually positive, what percent was classified correctly. The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

In the classification report for Voting model, there is a 90% precision for the high class meaning that among all positively classified instances, 90% of them were correct. Recall is 78% which means among all positive instances, 78% of those were classified correctly.

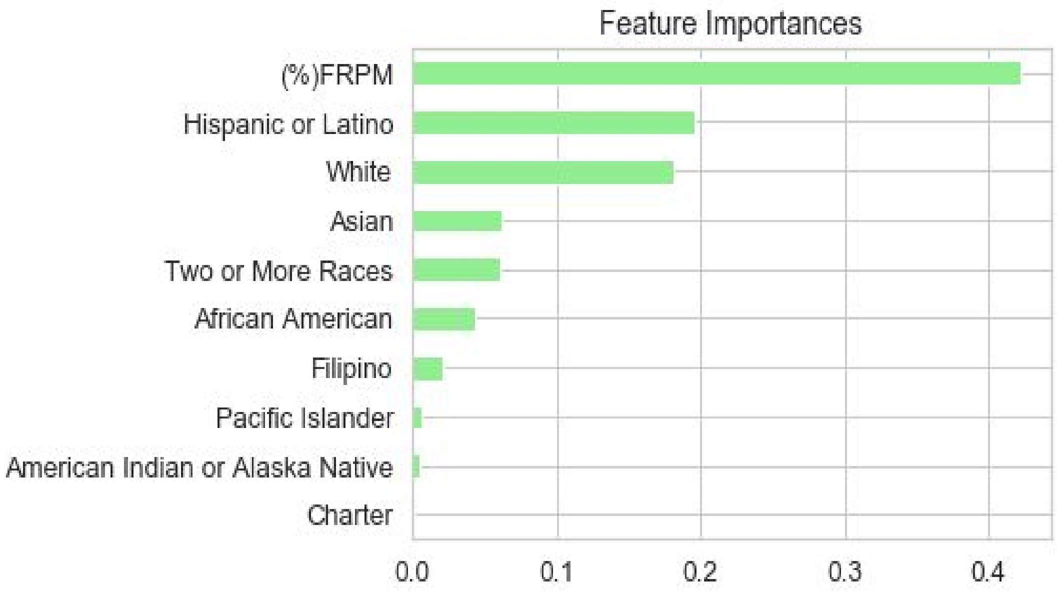


Figure 13. *Feature Importances*

Figure 13 shows the feature importances which represent how much including a particular variable improves the prediction based on the Random Forest Classifier. I wanted to determine which feature was most useful for discriminating between the classes to be learned. The graph showed that the best predictor of success level of a school is the percentage of the students who receive free and reduced priced meals. The second most important factor is the percentage of Hispanic students. It's followed by the percentage of White and Asian students. Whether the school is public or charter has no effect on the success level of a school.

As a result, the accuracy scores on the test sets are around 75-78%. The best performing models were Voting and XGBoost. Their accuracy scores on test set are the highest and the difference between the accuracy scores on train and test set is small. Future implementations of the models can be developed removing the variables that have no importance and the performance will not suffer.