San Diego Schools

The Health of Education is in the Data

10.7.2017

All data, code, analysis, and results can be found here:
https://github.com/IvansStus/SanDiegoSchools

Competition Category: Team of College Students with Individual College Student competing for internship (Ivan Stus)
Answering: Descriptive Questions, Correlational Questions, and Predictive Question

Team Name: San Diego Schools

Ivan Stus
SDSU Computer Science
ivan.stus.echs2017@gmail.com

Michael Galarnyk
UCSD Data Science
mgalarny@gmail.com

Jillian Jarrett
UCSD Data Science
jillianjarrett@gmail.com

Orysya Stus
UCSD Data Science
orysyastus2012@gmail.com

Team Member whose Dropbox folder contains the team's submission: Orysya Stus (orysyastus2012@gmail.com)

Aim

Examine features contributing to differing fail and pass rates across high schools through San Diego County via descriptive, statistical, and predictive methods.

Methodology

I. GreatSchools API Calls (https://www.greatschools.org/)

In addition to using the data supplied by the hackathon organizers, we used scraped data from greatschools.org. GreatSchools is a nonprofit organization that provides consumers with information about K-12 schools and education. The website has tools/metrics for evaluating and comparing schools, and allows for users to write reviews. The API methods used included school profile, city school profile, school census information, and test scores. The extracted information was used to complement the VOSD dataset and provide a more in depth of understanding of factors which affect AP test scores. Python packages BeautifulSoup, requests, and pandas were used.

II. Web Scraping

To access more granular data from GreatSchools, Python's Scrapy was used to access test scores by ethnicity and subject. The purpose of this was to pull data with metrics showing how students were testing relative to the state average in multiple subjects (Math, English, and Science). Five columns of data were scraped: Ethnicity (Categorical), Subject (Categorical: Math, English, Science), Test Scores (Percentile), State Average Test Score, gsid (to merge back with original dataset).

III. Data Integration between Data Sources

This project integrates multiple data sources as to provide additional perspectives on the health of education in San Diego. The 'VOSD San Diego County Schools Dataset 2017' (VOSD) dataset was used as the primary data source to which each of the other data sources can be joined. The 'ELA 2017 SD County Scale Scores gr 3,8,11' (ELA) and 'MATH 2017 SD County Scale Scores gr 3,8,11' (MATH) datasets can be merged to the VOSD dataset by combining the Country Code, District Code, and School Code together to be a primary key to the CDSCode in VOSD. GreatSchools (GS) provides a unique identifier for each school included, gsid. Both a programmatic and manual approach was employed to create a mapping between each VOSD school, primary_key in 'VOSD San Diego County Schools Dataset 2017 mapped' and 'gsid'.

Our final dataset (Final_dataset.xlsx is comprised of the following data):

Tab Name	**Origin**
VOSD SD County School Mapped	Hackathon provided, includes mapping to GreatSchools gsid for further utilization of GreatSchools API.
ELA 2017 SD County Scale Scores	Hackathon provided.
MATH 2017 SD County Scale Score	Hackathon provided.
GreatSchools Census Data	GreatSchools API Calls (https://www.greatschools.org/api/docs/schoolCensusData.page). Can be merged with VOSD by gsid.
GreatSchools Test Scores RTP	GreatSchools Web Scraped. Data shows score comparisons between state and school.Can be merged with VOSD by gsid
GreatSchools Test Scores	GreatSchools API Calls https://www.greatschools.org/api/docs/school-test-scores/

IV. Feature Examination and Transformation

Data Cleaning/Transformations

The following was completed on the VOSD data:

For the analysis, we dropped the following columns, 'CDSCode', 'District', 'Street', 'StreetAbr', 'City', '*', 'State', 'Phone', 'OpenDate', 'DocType', 'SOCType', 'US News', 'GSoffered', 'Latitude', 'Longitude', 'LastUpdate', 'primary_key', and 'gsid' as they have no predictive power (assumption). Additionally we dropped, '2015-2016 ELA Status Level', '2015-2016 ELA Status Level (Decode)', '2015-2016 ELA Change Level', '2015-2016 ELA Change Level (Decode)', '2015-2016 ELA Color', '2015-2016 ELA Color (Decode)', '2015-2016 ELA box', '2015-2016 ELA box (Decode)', '2015-2016 Math Status Level', '2015-2016 Math Status Level (Decode)', '2015-2016 Math Change Level', '2015-2016 Math Change Level (Decode)', '2015-2016 Math Color', '2015-2016 Math Color (Decode)', '2015-2016 Math box', and '2015-2016 Math box (Decode)' due to the sparsity of data.

The columns, '2015-2016 ELA Current Status (avg. distance from level 3)', '2015-2016 ELA Prior Status (age. Distance from level 3)' as well as '2015-2016 Math Current Status (avg. distance from level 3)', 2015-2016 Math Prior Status (age. Distance from level 3) were dropped because the 'Change.in.Difference.Between.Current.Prior.Status' and 'Math.Change.Difference.Between.Current.Prior.Status' columns were more interesting. Similarly we dropped, '2015-2016 AP Score = 1', '2015-2016 AP Score = 2', '2015-2016 AP Score = 3', ''2015-2016 AP Score = 4', '2015-2016 AP Score = 5', '2015-2016 Enrollment Grades 10-12', and '2015-2016 AP Number Tested'.

Transforming the Data

Replaced 'Y' and 'N' with the integers 1 and 0 respectively for the 'Charter' column to allow it to be incorporated into the model.

Similarly the values 'x' and 'X' in the 'AVID', 'Dual Language', 'Arts', 'IB', and 'Education' columns were replaced with the integer 1. The missing values were replaced with the integer 0.

V. Feature Engineering and Selection - Generating the Target Variable

The AP scores were considered as a composite, the overall value the school. The columns '% 2015-2016 AP Score = 1', '% 2015-2016 AP Score = 2', '% 2015-2016 AP Score = 3', '% 2015-2016 AP Score = 4', and '% 2015-2016 AP Score = 5' were used to create a composite average score.

Analysis

I Descriptive Questions: In order to determine the San Diego high school and city trends in regards to the AP composite scores, horizontal bar charts were generated to show which high schools and cities tended to excel or fall behind.

II. Both covariance and correlation heatmaps were generated using both VOSD and GreatSchools data to demonstrate the relations between attributes affected the composite AP score. Examination of these heatmaps along with evaluation of the predictive model provides an understanding as to what has the highest impact on the high schools' composite AP scores. The correlation heatmap will be discussed below.

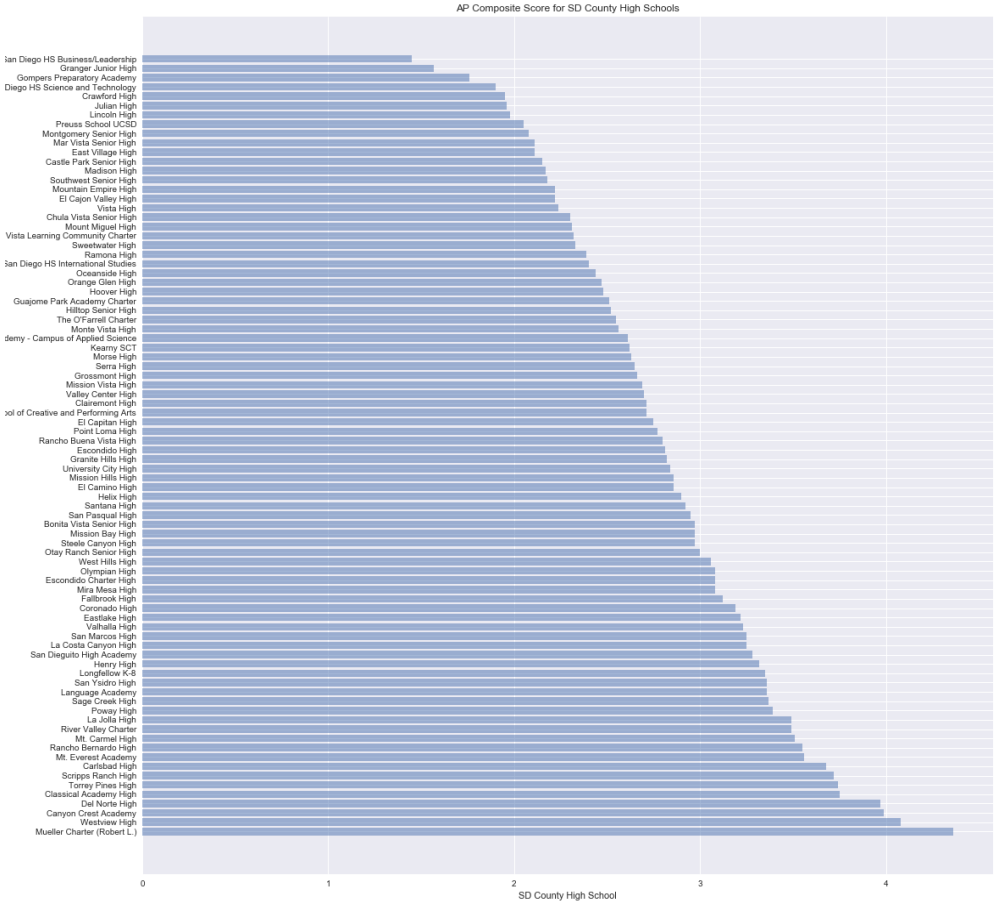
III. Both a decision tree regressor and linear regression were tested on a training set of about 80 high schools in San Diego county to determine the best model to predict the high school's' composite AP scores. The linear regression model had the lower mean squared error (MSE) of 0.153885892918 compared to the decision tree regressor's MSE of 0.21999375. Analysis of coefficients will be discussed below.

Results

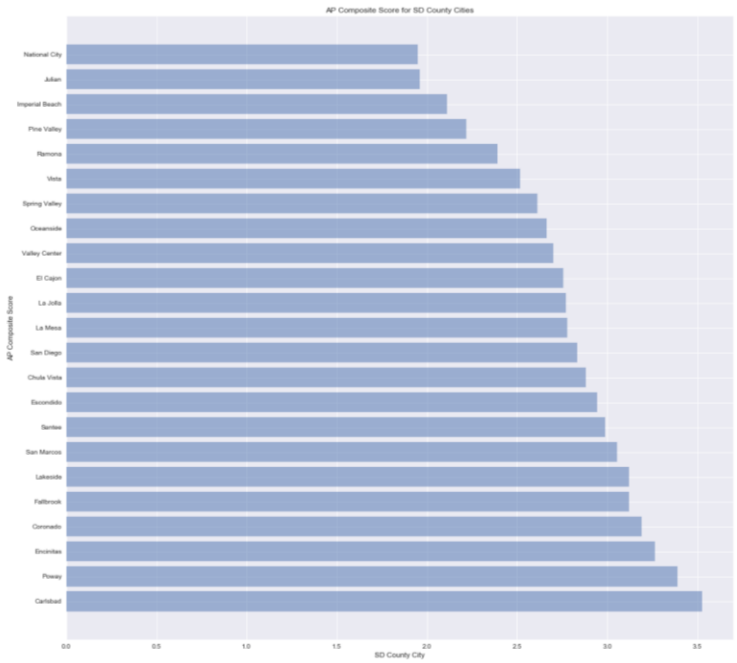
Note: All figures have corresponding .csv files for easier value interpretation.

I. Descriptive Questions

AP Composite scores vary across San Diego County High Schools, with Mueller Charter (Robert L.), Westview High, and Canyon Crest Academy earning the highest composite AP scores and San Diego HS Business/Leadership, Gompers Preparatory Academy and Granger Junior High earning the lowest composite AP scores.



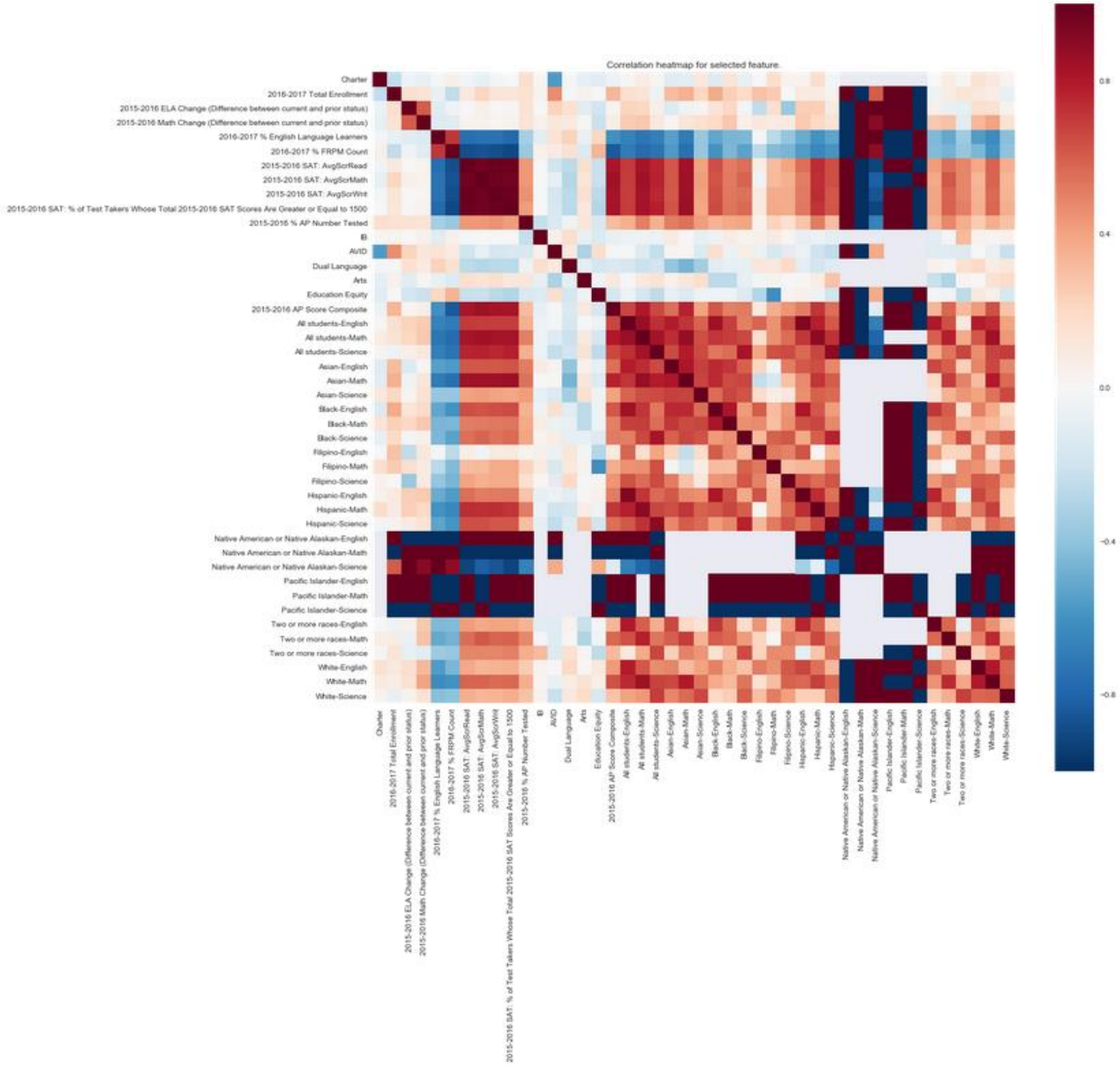
Additionally, high schools in Carlsbad earned the highest composite AP score while high schools in National City earned the lowest composite AP score.



II. Correlational Questions

The VOSD and GreatSchools Test Scores RTP datasets were merged, with correlation heatmaps analyzed. the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. Notable correlation relation include:

1. AP Composite Score is negatively correlated with English Language Learners and FRPM (free or reduced lunch meal) Count, showing that those that are learning English or have lower economic status do not have high AP composite scores.
2. AP Composite Score is positively correlated with any of the SAT scores, showing that success in one type of an exam in success in others.
3. FRPM was also negatively correlated with SAT scores, showing that this group of individuals do not score as high as their counterparts.
4. AVID and Charter and negatively correlated, meaning that there are less AVID programs in Charter schools.



III. Predictive Questions

Features of interest were extracted from the VOSD data and used within a linear regression model to determine the coefficients which influenced predictions of the composite AP score. Below are the results:

```
Coefficients:
Charter :  0.0138852593482
IB :  -0.215036212319
2015-2016 ELA Change (Difference between current and prior status) :  -0.00251213392298
2015-2016 Math Change (Difference between current and prior status) :  0.00421449204012
2016-2017 % English Language Learners :  -0.000966539660395
2015-2016 SAT: AvgScrWrit :  0.0088434623387
AVID :  -0.0361338621381
Arts :  -0.145489929854
2015-2016 SAT: % of Test Takers Whose Total 2015-2016 SAT Scores Are Greater or Equal to 1500 :  -0.0180519510478
2015-2016 SAT: AvgScrMath :  0.00697047826187
2015-2016 SAT: AvgScrRead :  -0.00274398894828
2016-2017 % FRPM Count :  -0.00663534728466
2015-2016 % AP Number Tested :  -0.00404794529971
2016-2017 Total Enrollment :  0.000229619182197
Dual Language :  0.0468746798779
Education Equity :  0.0168119807812

Intercept:  -2.71387028657
```

Notably, if the school is a Charter school and the schools 2015-2016 SAT: AvgScrWrit scores the Composite AP Score increases, while if the school has an IB or Arts program this decreases the Composite AP Score.

Conclusions

Through this project, our team accessed additional data, GreatSchools, which provided another perspective of the health of San Diego's education system. We highlighted the differences between high school composite AP scores, interpreted a correlation heatmap, and recommended a predictive model to be used for estimating future Composite AP Scores. In order to reinforce our findings, additional data can be used to further understand the year to year trends of San Diego high schools.