Predicting SAT Scores Using High School Quality Reviews

Ryan Chandra, Arafat Mohammed, Jacob Wagner
Department of Computer Science
Stanford University
{ryandc, arafatm, jtwagner}@stanford.edu

Overview

Motivation

The American education system is one that attracts significant attention in both politics and media. Heated debates about Common Core, teachers unions, and charter schools have roiled policymakers, with little resulting consensus. The passion with which these conversations are undertaken is understandable – at stake, after all, are the hearts and minds of future generations.

Problem Statement

With all the possibilities of investment for schools – enrollment, teaching, diversity, etc. – it can be hard for administrators to best identify areas of impact. With that in mind, we sought to apply machine learning techniques to school reports to discover what most impacts student performance. By using SAT scores as a proxy for student outcomes, we hoped to provide actionable insights for decision-makers in our education system.

Data Set and Features

- Our data came from Data.gov, the United States government's repository for open data.
- Our data set is entitled "2013-2014 School Quality Reports Results for High Schools," and was published by the New York City Department of Education.[1]
- From the initial data set, we took the relevant information and wrote it to a CSV as a row entry with 25 school specific features.
- These 25 features include Quality Ratings, Enrollment, % ELL, % Students with Disabilities, % Free Lunch, 8th Grade School Proficiency, Ethnic Breakdown, Quality Review Questions, Teacher Attendance, Years of Principal Experience at This School, and Average SAT Score.
- The training data was a conglomeration of information from 265 schools, while the validation set and test set had 33 and 34 school examples, respectively.

Results

Figure 3. Table of model results

Model	Data	MSE	Variance Score
Linear Regression	Basic data	0.001897	0.91
Support Vector Regression	Basic data	0.002513	0.89
Random Forests	Basic data	0.002192	0.9
Neural network	Basic data	0.0019	
Linear Regression	Bias data	0.001603	0.94
Support Vector Regression	Bias data	0.002259	0.91
Random Forests	Bias data	0.002951	0.88
Neural network	Bias data	0.002337	
Linear Regression	RF Data	0.001981	0.91
Support Vector Regression	RF Data	0.002436	0.89
Random Forests	RF Data	0.002175	0.9
Neural network	RF Data	0.002315	
Linear Regression	PCA Data	0.004872	0.81
Support Vector Regression	PCA Data	0.005354	0.8
Random Forests	PCA Data	0.006057	0.77
Neural network	PCA Data	0.004715	
Linear Regression	Trimmed Data	0.006131	0.42
Support Vector Regression	Trimmed Data	0.005745	0.46
Random Forests	Trimmed Data	0.007527	0.29
Neural network	Trimmed Data	0.007329	

- For reference…
  - *Basic Data*: All-inclusive
  - *Bias Data*: No demographic info
  - *RF Data*: Only the most influential features, as given by random forests
  - *PCA Data*: Reduce 25 features down to 5 inclusive ones
  - *Trimmed Data*: No demographic or middle school information
- The lowest average Mean Squared Error (MSE) came from the full data set.
- The highest variance score came from the "Bias data" where we removed all racial and demographic information.
- The most important feature in predicting SAT scores is middle school performance.
- Linear regression and neural networks generally outperformed SVMs and random forests.

Models

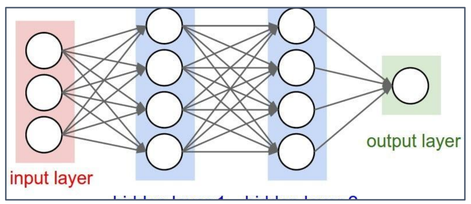

Figure 1. Neural Network Representation[2]

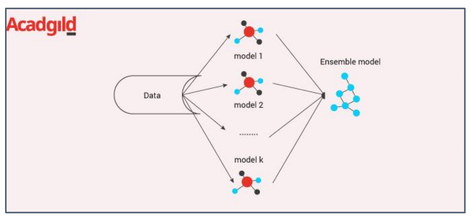

Figure 2. Random Forest Representation[3]

To tackle our problem, we wanted to implement a series of diverse algorithms. With that in mind, we picked four algorithms – linear regression, neural networks, random forests, and support vector machines – that aligned with the aforementioned goal and was additionally supported by our literature review.

Models

- Linear regression involves predicting an output y given a vector of inputs, x, using a set of weights, w.
- A neural network put simply, is a more advanced regression model that utilizes hidden layers.
- Random forests are an ensemble learning method that constructs decision trees and outputs the mean prediction of the individual trees.[4]
- Support vector machines are supervised algorithms that find a hyperplane to minimize the loss of the data.

Discussion

Given the importance of past performance, it is important for educators to understand that high school performance is holistic, and depends as much on prior experience as it does on current exposure. The challenge, then, is to integrate an education system that at times may seem siloed or otherwise disconnected. To further understand this system, we plan to interview and survey teachers to learn from their expertise. Another difficulty we faced is acknowledging that bias may exist in our data set. Our hope was that, by experimenting with the removal demographic information, we could still produce accurate results. Unfortunately, we found that our models performed approximately three times worse (see Figure 3 with "Trimmed Data"). These issues of fairness and bias are emblematic of larger problems in artificial intelligence, and are difficulties with which we continue to grapple.

Future Work

- Extend our research to more school systems.
- Rerun our analysis with earlier academic data.
- Experiment with new algorithms (e.g. k-nearest neighbors).
- Complete intense case studies to design plans to optimize SAT scores for high schools.

References

[1]https://catalog.data.gov/dataset/2013-2014-school-quality-reports-results-for-high-schools
[2]https://icdn6.digitaltrends.com/image/digitaltrends/artificial_neural_network_1-791x388.jpg
[3]https://acadgild.com/blog/wp-content/uploads/2018/09/Random-Forest.jpg
[4]https://en.wikipedia.org/wiki/Random_forest