Finding Meaning in New York City Public School Data

CS229: Machine Learning, Fall 2012

Sophia Westwood
Computer Science Department
Stanford University
sophia@cs.stanford.edu

Daniel Jackoway
Computer Science Department
Stanford University
jackoway@cs.stanford.edu

ABSTRACT

In this paper, we apply machine-learning techniques on data from over five hundred New York City schools and examine the factors behind individual school performance in 2011 and 2006 based on the demographic, test performance, and survey data collected in 2006. We uncover major factors shaping school improvement and decline; provide a predictive model for schools' future trajectories; and rigorously evaluate common assumptions about school performance. We measure school performance with the New York State English Language Arts student test results. We find that a school's starting position in 2006 is extremely powerful in an unexpected inverse direction. Schools above the mean tend to get worse, and schools below the mean get better. Yet, initial state test score is only one window into the data. We explore key features beyond score in predicting both current performance and future improvement, and relate the two learning objectives. While demographic data allows us to accurately classify schools as above- or below-average, we make strong predictions of future improvement based on just two questions on parent involvement. Finally, we provide machine-learning evidence of the strong and nuanced web of interrelated factors that characterize schools.

1. INTRODUCTION

The New York City Department of Education is the largest system of public schools in America, covering approximately 1700 schools and 1.1 million students[6]. In addition, New York City serves a large population of low-income and minority students. The school system collects and publishes extensive data on their schools, including detailed parent and teacher surveys ranging back to 2006. Yet, there exists little machine-learning analysis based on this data.

This combination – a large and diverse student body, an extensive and detailed dataset on schools, and the dearth of existing rigorous analyses – motivates our analysis of the New York City school system to identify key factors in school improvement and decline and predict future school performance. We aim to apply machine-learning techniques to better understand what shapes primary school success and failure; provide a predictive model for schools' future trajectories; and rigorously evaluate common assumptions about school performance.

2. BACKGROUND

Previous work in the area of education and data has often emphasized performance predictions for individual students. For example, Lloyd predicted student failure from third-grade data based on a regression model that labels students as Graduating or Failing through current grades, family characteristics, and test scores [1]. In addition, a 2008 study on the California High School Exit Exam investigated factors in fourth grade – including student behavior as well as student performance – that predict later student success and failure on the test. The study pinpointed classroom behavior, student absences, ethnicity, test scores, and English Learner status [11].

Additional studies consider the effects of particular intervention techniques on school performance. An analysis of No Child Left Behind's school accountability requirements found that school poverty and district size were the best predictors of whether underperforming schools would be able to achieve the necessary improvement, where high-poverty schools in large districts were the least likely to succeed. Adopting strategies such as dedicating resources to align the curriculum with standards were also predictors of success [10]. Other studies investigate the success of particular intervention programs such as Success For All in inner-city elementary schools [2].

Our project differs from these approaches in that it does not look at the performance of individual students, nor at the success of particular intervention programs. Rather, we analyze the factors that affect general school improvement or decline across a broad range of schools in the New York City school district. We examine factors including both quantitative achievement results and demographic information, as well as extensive official surveys of parents and teachers. The surveys include, for example, ratings of perceived teacher quality, parent-teacher-student interaction, perceived school environment, and perceived gang influence.

3. DATA

We obtain our data set by merging publicly-available spreadsheets from the New York City Department of Education. Focusing on third grade, we combine surveys of 216,914 parents [4]; surveys of 31,592 teachers [4]; school demographics [7]; and New York State English Language Arts test scores [9]. [1]. In all, we have 524 features for 554 schools.

4. METHODOLOGY & RESULTS

We now present our methodology and results. Sections 4.1 and 4.2 develop a model based on the strong inverse rela-

[1] See [8] for further information on survey methodology and questions.

Table 1: Classifying improvement. Columns: Train accuracy, test accuracy, precision for Did Not Improve, recall for those Did Not Improve, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	1.00	.773	.77	.79	.78	.76
Linear SVM	1.00	.737	.71	.81	.77	.66
Gaussian SVM	.930	.743	.78	.68	.71	.80
Random Forest	1.00	.784	.79	.78	.75	.84

tionship between current score and future improvement; Section 4.3 looks at predicting this model's misclassifications; Section 4.4 analyzes key features for future improvement beyond current score; Section 4.5 analyzes key features for predicting current score; and Section 4.6 teases apart the relationship between the problem of predicting current score and predicting future improvement with regard to the key features for each. Finally, Section 4.7 presents additional mathematical insight into the structure of the featurespace through Principal Components Analysis.

We normalize all features to have a mean of 0 and variance of 1. In addition, we evaluate binary classification results with hold-out cross validation, randomly placing 70% of our 554 examples in the training set, and the remaining 30% in the testing set.

4.1 Initial: predict school improvement

As an initial approach, we pose the following binary classification problem: "Based on 2006 data, predict whether individual schools' third-grade New York State English Language Test results were better in 2011, or not." The 524 features consist of the 2006 English state test score, 2006 demographic data, and 2006 parent and teacher survey responses. The labels we predict are the 2011 New York State English Test results. The data points are individual public schools in the New York City school district.

We train logistic regression, linear and Gaussian SVMs, and a Random Forest Classifier and achieve high precision, recall, and accuracy. See Table 1 for results. Logistic regression and the random forest classifier perform the best on the test set, achieving 77.3% and 78.4% test accuracy respectively, with reasonable precision and recall on both classes.

We next explore naive classifiers to estimate how well one can predict this problem without complex machine learning. One might assume the model "Good schools get better, bad schools get worse." Good schools generally have involved parents, engaged students, and strong teachers; thus, one might expect to see a continued upward trend. On the other side, bad schools often have less-involved parents, more disconnected students, and weaker teachers, which might make them continue to get worse barring intervention. Our first naive classifier, then, labels schools as "improving" if their 2006 English score is above the mean 2006 English score across all schools, and as "not improving" otherwise. Our second naive classifier inverts these labelings. The results in Table 2 show that, contrary to intuition, a naive classifier that outputs "Good schools get worse, bad schools get better" produces even higher performance than complicated machine learning classifiers. The strong performance of this naive classifier suggests a dominating trend: regression to the mean.

4.2 Multiclass classification

We proceed to explore this naive result in more depth.

Table 2: Naive classifier: Columns: Train accuracy, test accuracy, precision for Did Not Improve, recall for those Did Not Improve, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Good improve	.116	.150	.19	.21	.09	.09
Bad improve	.884	.850	.91	.79	.81	91

Table 3: Multiclass classification: Columns: Train accuracy, test accuracy, precision for Decline, recall for Decline, precision for Neutral, recall for Neutral, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec	2 pr	2 rec
Logistic	1.00	.613	.50	.59	.67	.72	.60	.35
Linear	1.00	.570	.46	.54	.64	.69	.44	.24
Gaussian	.887	.616	.68	.33	.61	.94	.00	.00
Rand Forest	.995	.651	.66	.49	.65	.87	.64	.21
Naive	.791	.785	.79	.77	.79	.85	.75	.62

The average 2006 score for all schools is 660 points (within a rough range of 500 to 800). The change in English score between 2006 and 2011 follows a normal distribution with mean .089 points and standard deviation 13.7 points. Based on this distribution, we expand our binary classification problem to a multiclass problem with three classes: Significant Improve, Neutral, Significant Decline. Neutral consists of ±10.5 points from the mean ($\approx \pm 0.75\sigma$), Significant Improve consists of > 10.5 points from the mean ($\approx > 0.75\sigma$), and Significant Decline consists of < 10.5 points from the mean ($\approx < -0.75\sigma$).

Table 3 presents the performance results of the algorithms. Note that Logistic Regression and the Linear SVM involve one-vs-all multiclass classification, the Gaussian SVM involve one-vs-one multiclass classification, and the Random Forest classifier is inherently multiclass. The three-class naive classifier labels "Decline" if the school's 2006 English score is below (mean - σ), "Neutral" if the 2006 score is between (mean - σ) and (mean + σ) and "Better" if the 2006 score is above (mean + σ). Inspecting the test accuracy, precision, and recall of the different algorithms supports the same conclusion as the binary classification: a naive prediction of "bad schools get better, good schools get worse" outperforms more advanced machine learning classifiers.

4.3 Insight into misclassifications

Now that multiclass classification has verified the validity of the naive model, we return to the binary classification problem and investigate the reasons for success of the "bad schools get better, good schools get worse" model. Then, we delve into its points of failure by modeling naive misclassifications as a prediction problem.

Figure 1 presents the linear relationship between the base 2006 English score, and the 2006 to 2011 change in English score. The Pearson correlation coefficient is -0.88, demonstrating a powerful inverse correlation. This strong linear correlation between 2006 score and score improvement explains the success of the naive classification model, as the regression line roughly maps a net change of 0 to a 2006 score of 660, the mean starting score.

We now examine the schools that the naive model misclassifies. These schools are either above-the-mean schools that are getting better, or below-the-mean schools that are getting worse. As judged by third-grade English state test scores, just 8.0% of above-mean schools improved and just 15.3% of below-mean schools got worse. Because the dataset for good schools that improve is so small, we focus here on

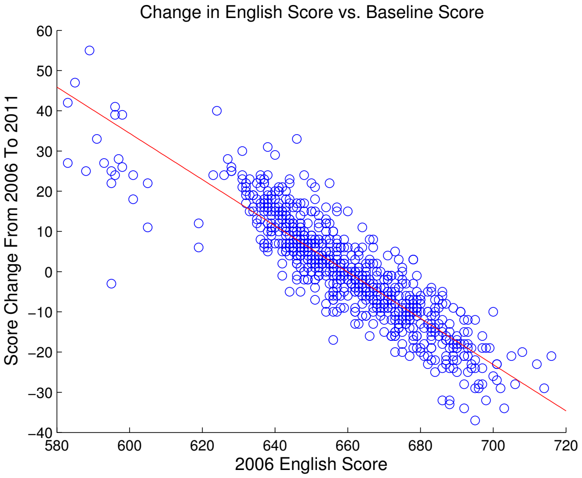

Figure 1: There is a strong negative correlation between a school's 2006 score and its score increase from 2006 to 2011.

Table 4: Classifying improvement of below-mean schools. Columns: Train accuracy, test accuracy, precision for Did Not Improve, recall for those Did Not Improve, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	1.00	.564	.08	.57	.95	.56
Linear SVM	1.00	.546	.08	.57	.95	.54
Gaussian SVM	.985	.691	.08	.36	.94	.71
Random Forest	1.00	.656	.10	.57	.96	.66

the bad schools classification problem.

We train logistic regression, SVMs, and Random Forest classifiers, where the features and labels are the same as the prediction problem in Section 4.1, but the data points are limited to schools with 2006 test scores below the mean. Note that the binary naive classification would predict "Improve" for every data point in this set. Similarly, the SVM and Random Forest predictions initially output "Improve" for all the data points, reflecting the skewed training data. We thus resample the data to formulate a training set with even representation from both classes; concretely, there are 34 training examples from each class. Table 4 shows the results of training learning algorithms to predict whether a bad school will improve or not.

Even with resampling, the learning algorithms struggle on this second classification problem. While each achieves high training accuracy, test accuracy is low; in addition precision is extremely low for the "Did Not Improve" class. We attempt to address the high variance by trimming down the featurespace (by running an SVM with L1 norm penalty; see Section 4.4 for details on feature pruning), but to little success. The high variance suggests that we may need more training examples – particularly bad schools that got worse – in order to make better predictions for below-average schools.

4.4 Key features for predicting improvement

So far, we have included the 2006 English score as a feature, and the success of the naive model proves it is a strong predictor. We now remove the starting score feature and investigate other key features for making predictions of future performance. We focus on feature pruning via an SVM with L1 norm penalty [3] to determine feature importance in the binary classification problem described in Section 4.1, with the 2006 English score feature removed.

We present lists of chosen features when the SVM's penalty is such that it chooses first two and then five features, together with each feature's weight. Positive weights mean the algorithm is more likely to select "Improved" when this feature's value is high, while negative weights mean the algorithm is more likely to select "Did Not Improve" when this feature's value is high.

Top Two Features For Predicting Improvement

1. Percentage of parents who reported talking to a teacher about how to help their children learn better once or twice this school year. (-0.0439)
2. Percentage of parents who reported talking to a teacher about their children's academic progress at least once a week. (6.01650e-05)

Top Five Features For Predicting Improvement

1. Percentage of parents who reported talking to a teacher about how to help their children learn better once or twice this school year. (-0.0810)
2. Percentage of parents who reported talking to a teacher about their children's academic progress at least once a week. (0.0362)
3. Percentage of teachers who reported that 76-100% of their students had a parent attend a conference (-0.0340)
4. Percentage of teachers who "strongly disagree" that adults at their school are often disrespectful to students (-0.0122)
5. Percent of students who are Asian. (-0.0070)

Many of the presumed "strong-performance" features (such as parent improvement) have negative weights, meaning that the algorithm is likely to select "Did Not Improve" when these "good" features are high. This apparent paradox stems from the strong inverse relationship between 2006 English score and improvement. Strangely, having a high percentage of parents who talk "to a teacher about their children's academic progress at least once a week" has a positive weight – this suggests that high values will predict an "Improve" label, implying a bad school. We propose that a high percentage of parents and teachers talking extremely often about children's progress perhaps suggests that children are struggling. It is also notable that the best few features for prediction on the data are primarily survey responses, not demographic data. We posit that the many complex factors affecting school performance and improvement, and the variance within demographic categories, might mean that a survey answer may capture more information. (See Sections 4.5 and 4.6 for more analysis of predictions based on demographic data.)

While we saw above how 2006 English scores predict improvement well, we can also make strong predictions with the top survey question features. Training on just the top two features for predicting improvement, we achieve the results in Table 5, a performance close to the initial machine-learning results from when training on all the data in Section 4.1. Training on just the top five features for predicting improvement, we achieve even stronger results (see Table 6). We do not attempt to draw causality from these results, or even a root cause; but, it is notable how much information about a school's future trajectory can be gleaned from so

Table 5: Classifying improvement using only two features. Columns: Train accuracy, test accuracy, precision for Did Not Improve, recall for those Did Not Improve, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Regression	.741	.694	.73	.71	.65	.68
Linear SVM	.741	.700	.74	.71	.65	.69
Gaussian SVM	.759	.711	.77	.68	.65	.75
Random Forest	.874	.722	.77	.71	.67	.74

Table 6: Classifying improvement using only five features. Columns: Train accuracy, test accuracy, precision for Did Not Improve, recall for those Did Not Improve, precision for Improved, recall Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Regression	.791	.756	.82	.72	.70	.80
Linear SVM	.794	.756	.82	.72	.70	.80
Gaussian SVM	.802	.750	.82	.70	.68	.81
Random Forest	.997	.744	.79	.73	.69	.76

little information on the school, centering mostly around parent-teacher interactions.

4.5 Key features for predicting current performance

To follow up our analysis of key features for predicting improvement, we now investigate the key features for predicting current performance. We model the classification problem as follows: the feature space consists of all survey and demographic data, the labels are "Above" if the 2006 score is above the mean and "Not Above" otherwise, and the data points are the schools. We prune features as in Section 4.4 and present lists of the top features with their scores:

Top Two Features For Predicting Current Performance

1. Number of Black students (-.000348)
2. Number of White students (.000209)

Top Five Features For Predicting Current Performance

1. Number of Asian students (0.00272)
2. Number of White students (0.00203)
3. Number of Black students (-0.00109)
4. Number of Hispanic students (-0.000598)
5. Teachers report that foreign language is not offered in any form at the school (-0.000424)

Depressingly, these feature lists demonstrate the strength of school demographic features – likely also reflective of socioeconomic status and school resources – in predicting the current performance on state tests. Tables 7 and 8 show the performance of learning algorithms trained using just the top two features and just the top five features, respectively.

Table 7: Classifying 2006 score (above mean/not above) using only two features. Columns: Train accuracy, test accuracy, precision for Not Above, recall for those Not Above, precision for Above, recall for Above.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	.812	.776	.71	.94	.90	.61
Linear SVM	.807	.776	.71	.94	.90	.61
Gaussian SVM	.816	.789	.72	.95	.92	.62
Random Forest	.990	.795	.76	.89	.85	.70

Table 8: Classifying 2006 score (above mean/note above) using only five features. Columns: Train accuracy, test accuracy, precision for Not Above, recall for those Not Above, precision for Above, recall for Above.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	.880	.861	.84	.91	.89	.80
Linear SVM	.883	.866	.84	.93	.91	.79
Gaussian SVM	.888	.882	.85	.95	.93	.80
Random Forest	.997	.882	.89	.89	.87	.87

Table 9: Classifying improvement using five features most predictive of current score. Columns: Train accuracy, test accuracy, precision for Not Improved, recall for Did Not Improve, precision for Improved, recall for Improved.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	.782	.744	.86	.66	.66	.86
Linear SVM	.782	.732	.86	.63	.64	.86
Gaussian SVM	.803	.750	.88	.65	.66	.89
Random Forest	.997	.707	.75	.72	.65	.69

The high test accuracy, precision, and recall of these algorithms demonstrates an unfortunate degree of success in predicting current score based primarily on current demographic data.

4.6 Relationship between improvement and current score

We now connect the key features for improvement identified in Section 4.6 and the key features for performance identified in Section 4.5 to the strong inverse relationship between current score and future improvement analyzed in Sections 4.1 and 4.2. We investigate how closely tied the key features for predicting improvement (from Section 4.6) are to the key features for predicting current score (from Section 4.5) by examining how each set of five key features performs on the opposite classification problem. Table 9 presents results for classifying school improvement using just the five features most predictive of current score; Table 10 presents results for classifying current performance using just the five features most predictive of school improvement.

The relatively strong performance in both tables confirm that current score and improvement are closely linked, as seen in Sections 4.1 and 4.2. Yet, the slightly weaker performance of the score key features on classifying future improvement as compared to the improvement key features (see Section 4.6) suggests that the machine-learning approach to improvement does more than simply approximate current score and use the score-improvement inverse correlation; other features lend additional insight. Finally, such strong performance on limited sets of features also suggest the interrelatedness of the featurespace.

4.7 Analysis of feature space with PCA

We now use Principal Components Analysis on the normalized data to gain additional mathematical insight into

Table 10: Classifying 2006 score (above mean/not above) using five features most predictive of improvement. Columns: Train accuracy, test accuracy, precision for Not Above, recall for those Not Above, precision for Above, recall for Above.

	Train	Test	0 pr	0 rec	1 pr	1 rec
Logistic Reg.	.868	.887	.88	.89	.89	.88
Linear SVM	.868	.887	.88	.89	.89	.88
Gaussian SVM	.878	.887	.87	.92	.91	.86
Random Forest	.100	.857	.88	.83	.84	.88

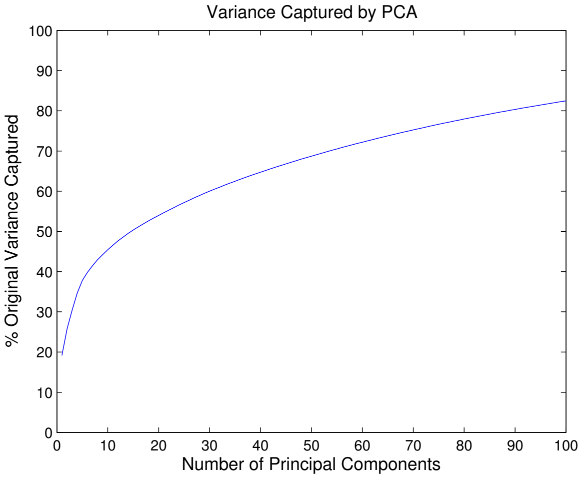

Figure 2: The first principal component captures 19% of the data's variance, and a long tail of further principal components slowly approach capturing the rest.

the structure of the featurespace. While two principal components together account for 25% of the variability in the data, one hundred principal components account for just 82% of the variability. Figure 2 presents how the total original variability accounted for changes with the number of principal components. While there is moderate frontloading in the first ten features, there exists an extremely long tail. One interpretation suggests that there are a few powerful underlying factors – perhaps all related to the idea of socioeconomic status or parent involvement – that affect a substantial amount of the variance. The long tail, however, suggests that much of the variance in the data cannot be accounted for by only a few dimensions, speaking to the richness of the dataset and the nuances of measuring schools with statistical data. These PCA results also suggest that there likely exists much variance within each binary label, as just a few features – accounting for just a fraction of the variance in the data – enabled us to make strong distinctions between the classes (see Sections 4.6, 4.5, and 4.6). This hypothesis makes intuitive sense – above-average or below-average schools still have wide variability. It is also possible that PCA's inability to represent the variability of the data with a small number of principal components suggests that features in the data – parent involvement and well-trained teachers, or gang violence and jaded teachers, for example – might have nonlinear interactions that PCA struggles to fully capture.

5. CONCLUSIONS AND FUTURE WORK

We have analyzed the strong inverse correlation between current score and future improvement (Sections 4.1 and 4.2) and attempted to predict which schools would buck this trend (Section 4.3). Deriving real-world insight from the data, we have determined the top features that predict future improvement (Section 4.4), as well as the key features for predicting current score (Section 4.5). We have related the two classification problems to better understand both (Section 4.6), and have applied Principal Components Analysis to provide further mathematical insight into the structure of the featureset (Section 4.7).

We find that "Bad schools get better, good schools get worse" describes the strong inverse correlation between current score and future improvement. Lack of sufficient counterexamples inhibits further analysis of why some good schools do get better and some bad schools do not get better. After removing current score as a feature, key features centering around parent involvement best predict future improvement or decline. For classification of schools' current scores as above- or below-average, demographic features appear the most predictive. With a set of five or even just two key features, we can classify future improvement or decline with an accuracy of 70 to 80%, and current performance with an accuracy of 80 to 90%. The trimmed featureset for improvement performs admirably in predicting score, and vice versa. Despite our ability to make strong classifications from few features, analysis of the featurespace suggests a nuanced, deep web of interrelated factors that characterize schools, opening up space for future work with larger datasets to better understand the differences within each of our classes ("Improve" and "Did Not Improve", and "Above-Average" and "Below-Average").

Preliminary analysis of math scores within the same framework produces the similar results. Future work in this domain might generalize results to other grades and other school districts, with particular focus on differences within below-average or above-average schools.

6. REFERENCES

[1] D. N. Lloyd. Prediction of school failure from third-grade data. *Educational and Psychologocial Measurement*, 38:1193–1200, 1978.

[2] N. A. Madden, R. E. Slavin, N. L. Karweit, L. J. Dolan, and B. A. Wasik. Success for all: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal*, 30.1:123–148, 1993.

[3] H. T. Nguyen, K. Franke, , and S. Petrovi'c. On general definition of l1-norm support vector machines for feature selection. *International Journal of Machine Learning and Computing*, 1.3:279–283, August 2011.

[4] N. D. of Education. 2007 learning environment survey, 2007.

[5] N. D. of Education. A new view of new york city school performance, 2002-2009, 2009.

[6] N. D. of Education. About us, 2012.

[7] N. D. of Education. Demographic snapshot 2012, 2012.

[8] N. D. of Education. Learning environment survey report 2006-07, 2012.

[9] N. D. of Education. New york city results on the new york state english language arts (ela) & mathematics tests grades 3 - 8, 2012.

[10] C. Padilla, K. Woodworth, and K. Laguarda. Evaluation of title i accountability systems: School improvement efforts and assistance to identified schools. Technical report, American Educational Research Association, 2006.

[11] A. C. Zau and J. R. Betts. Predicting success, preventing failure: An investigation of the california high school exit exam. Technical report, Public Policy Institute of California, 2008.