Improving Education in NJ Grade Schools with Digital Applications with Aviv Learning

Amy Le, Gilat Mandelbaum, Sakin Presswala

Introduction

Educational Data Mining (EDM) is an emerging field of Data mining where Machine Learning and AI tools can add much value by analyzing vast amounts of data generated from educational settings to uncover important patterns in student's learning behavior, effectiveness of different technology and learning tools and impact of demographic and other parameters on educational outcomes. The aim of our project is to use EDM to lay the groundwork for Aviv Learning, a new education company, to build an interactive digital application that can help students of New Jersey (NJ) K-12 public schools improve their performance on standardized tests. For this pilot study we plan to make use of NJ's 2016-17 public school performance report to answer the following questions:

A) What regions/ school districts in NJ are "underperforming"?

B) What subjects and grade levels are most in need of improvement?

C) What are the underlying reasons for variation in distribution of funds across districts and whether there is any meaningful distinction between "Charter Schools" vs "Non-Charter Schools"?

Motivation

Predicting the School Performance

Academic performance is considered the most important element of students' education. Going to a good school means student can have an education background which is highly respected, which will later help them enhance their learning ability. Obviously, it would be beneficial to identify schools at risk of underperforming in order to better implement intervention strategies. Our analysis will support Aviv Learning to identify which schools/ school areas are at risk of underperformance and which features have a significant impact on the school performance.

Subject Analysis

At a student level, knowing where there is need for improvement is essential for academic success. Some students spend hours studying and still not understand the material. Our analysis will pinpoint where majority of NJ grade school students are falling behind in and what they need to do to improve. This digital application will enhance student's ability to study more efficiently with engaging games and activities and our analysis will guide Aviv Learning to their first prototype.

Expenditure Analysis

NJ regularly features as a top State and ranks 2nd in the 2018 annual rankings by Education Week, which ranks US State's quality of K-12 education based on various parameters in categories such as chance for success, school finance and K-12 achievement (https://www.edweek.org/ew/collections/quality-counts-2018-state-grades/report-card-map-rankings.html). However, NJ also regularly features under top 5 States when it comes to educational spending per pupil and it is evident in census surveys that there no linear relationship between educational spending and positive educational outcomes

(https://www.census.gov/newsroom/press-releases/2015/cb15-98.html).For instance, States such as New-York and Alaska have a higher per pupil expenditure compared to NJ but rank much lower in educational quality rankings while States such as Massachusetts and Maryland spend comparative less but rank one spot above and below NJ in Quality counts respectively. Moreover, in our exploratory data analysis in Excel we found that there was a significant variance in the distribution of educational funds across different school districts in NJ and also a higher proportion of Charter schools were at the lower end of receiving State funds. These observations motivated us to research whether NJ has any scope of reducing or rebalancing its K-12 expenditure without compromising on quality and whether we could distinguish Charter School districts from Non-Charter public school districts based on different parameters. Our findings from this research might help us persuade the State educational board to cut down overspending in certain areas and instead invest in creative technological solutions such as our educational application to improve outcomes without straining their budget.

Literature review

Predicting the School Performance

The paper Aher, S. B., & Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems*, 51, 1-1 applied data mining techniques to build a Course Recommendation System in e-learning, which suggests the subjects to the students that they might be interested. The dataset collected from students includes 13 categories and 82 courses related to two branches i.e. Computer Science & Engineering (CSE) and Information Technology (IT). The authors then limited the data into 12 categories and 36 courses by considering only courses chosen by more than 100 students. To build the recommendation system, they used Association Rules and Clustering algorithm. Association rules are used to see the relationship among the features. In this case, Apriori association rule is applied because in course recommendation system, we only care about positive association rules. For example, if a student is interested in Deep Learning, he may want to learn Machine Learning. Therefore, Apriori association rule is chosen for recommending the course to the students. The result showed that if the support is increased then number of association rules are less. For support value of 0.6 they got only one association rule. Another approach is that they applied K-means clustering classify the data into different clusters. Then, Apriori rule is used on each cluster. The cluster with all association rules are positive containing "yes" is chosen as they recommend the course to the students. The result showed that there were 10 association rules for support value of 0.85, which means that the model can recommend 10 various combination of courses to the students. This approach increases the strength of the association rule. So this Course Recommendation System could help students to choose a proper course combination according to their interests. Future work of this paper is to investigate other combination of data mining techniques which may be the combination of classification and association rule or combination of classification, clustering and association rule.

Subject Analysis

Decision trees are the best classifiers for classification, especially for education data. "Data mining techniques can be used in education field to enhance our understanding of learning processes to focus on identifying, extracting, and evaluating variables related to the learning

process of students," (Priyam et. al., pp 334). Priyam et. al. compared some decision tree algorithms; C4.5, ID3, SPRINT, CART, and SLIQ. Each one has a unique way for classification and are useful in machine learning for education data (Priyam et. al. pp 334). Priyam et. al. found that C4.5's accuracy was the highest compared to ID3 and CART. Also C4.5 and ID3 algorithms have higher accuracy in terms of true positive rate than CART (Priyam et. al. pp 336). ID3 and C4.5 were able to identify which students were most likely to fail so those students can receive appropriate counseling (Priyam et. al., pp.336). According to Priyam et. al. the problem with CART, C4.5, and ID3 is they are more efficient with small sets of data. SLIQ and SPRINT are meant to use for larger sets of data. According to the paper, SPRINT is the best decision tree algorithm to use because it is fast, scalable, and removes data memory restrictions on data, making it efficient for very large data sets (Priyam et. al., pp.335). SPRINT would be the perfect contender as the classifier for determining target subjects for the application since this data set is going to be large with all fourteen years included. Another option is using C4.5 decision trees due to their high accuracy in identifying true positives.

Expenditure Analysis

Hochschild JL. Social Class in Public Schools. Journal of Social Issues. 2003, this paper outlines some important issues that were faced by American Public Schools more than a decade ago when this paper was published. The paper reviews different research studies conducted to study different causes that lead to disparities among schools at state level, district level and school level within district. Amongst the issues highlighted, the most significant is the distinct demographic patterns (socioeconomic, family, racial and ethnic backgrounds) of students attending disadvantageous schools (in terms of teaching quality, infrastructure, test scores on national and international assessments, dropout rates etc.) vs more advantageous schools. Though many reforms have been made to the educational structure of American public schools, since the paper was published in 2003, some of these causes might still be the underlying reasons for differences in performance within New Jersey's State schools that we intend to analyze through our project. This paper helped us gain insight into the larger policy issues that need to be addressed at national, state and district level in order to eliminate or at least reduce disparities and inequalities in our educational system and give all students an equal opportunity to succeed.

Another important paper that was reviewed was ***Educational data mining: A survey from 1995 to 2005***. This paper gave us an high level understanding of problems that EDM can answer using different machine learning algorithms and an overview of important studies that were carried out in EDM field since 1995 to 2005. Some of the papers that were referred to in this literature review were also very relevant to our project and were further reviewed by us in detail.

Approach

Predicting the School Performance

We investigate potential factors that may affect the performance of schools in New Jersey. The dataset we used is available on the New Jersey Department of Education website. We limited ourselves to data during 2016-2017 school year. The downloaded data includes 53 sheets describing school statistics, enrollment trends, racial groups, English and Math performance, teachers experience, Violence records, dropout rate, etc.

The first part of data cleaning process is to create a Primary key "School_ID" by combining the county code, district code, and school code. The reason is because some schools in the dataset share the same school code. With "School_ID" as the index, we merged the data from multiple worksheets into three main categories: **school characteristics** including school geographics, the number of teachers, teachers' average years of experience, school grade, whether Internet at school met requirement, the expenditure per pupil, rates of violence, weapons, HIB (Harassment, Intimidation, Bullying), the percentage of days faculty were present, school accountability percentile; **student performance** including the performance on Math and English, the dropout rate, and the suspension rate; **student demographics** including the number of students and the percentage of each racial group. Next, we obtained the student-teacher ratio based on the number of students and teachers. Then, based on Map of NJ, we divided 22 counties of NJ into 6 regions for our visualization purpose, which was also added as a feature to our dataset.

Region	**County**
Atlantic City	Atlantic County
Delaware	Mercer, Burlington, Camden, Gloucester, Salem County
Gateway	Middlesex, Union, Essex, Hudson, Bergen, Passaic County
Shore	Monmouth County and Ocean County
Skylands	Sussex, Morris, Warren Hunterdon, Somerset County
Southern Shore	Cumberland County and Cape May County.

Table A1: The Geographic Region of New Jersey

The next step of data cleaning process is handling with missing values. The average number of missing data in every column that has missing values is 13%. We used Python to perform County mean imputation in numeric columns and County mode imputation in categorical columns on those missing data. In order to apply machine learning algorithms with Python, we applied Label Encoding for categorical columns and 'StandardScaler' function to scale the data to zero mean and unit variance.

Our target label is "Under_Performing" with two value: Yes and No. An underperforming school is the school that has the School Accountability Percentile under 20%. According to NJ Education report, School Accountability Percentile is a rating (0~100%) given to a school based on its educational performance. I will refer to this column as School Performance.

The final dataset in this part has 20 attributes with 2516 instances (schools). Attributes such as the number of teachers was not included to avoid multicollinearity issues with the generated 'Student-Teacher ratio' feature. Attributes such as county code, district code, county name, district name, school name are also removed because they are combined to create the index 'School_ID'. Also, in the 'School_Accountability_Percentile', values under 20% were replaced

with ‘Yes’ while values above 20% were considered ‘No’. I also changed this column name to Under_Performance as our predicted column.

Geographic Region	The geographic location of NY grouped by counties	ELA performance	The student performance on English
StudentTeacher_ Ratio	The ratio of Students to Teachers	Math performance	The student performance on Math
TeacherAvgYears ExpInSchool	The average years of experience of teachers at schools	American Indian or Alaska Native	The percentage of students of American Indian or Alaska Native
Grade Type	E (PK-5); M (6-8); H(9-12); I(PK-8); T(6-12); A(PK-12)	Black or African American Hispanic	The percentage of Black or African American Hispanic students
Technology_Met _Requirement	Whether schools meet internet requirements (Yes or No)	Asian	The percentage of Asian students
Expenditure per Pupil	The expenditure given to a student	Native Hawaiian or Pacific Islander	The percentage of Native Hawaiian or Pacific Islander students
Violence_Weapo _HIB	The percentage of Violence, Weapons, HIB incidents	Two or More Races	The percentage of Two or More Races students
PercentDaysPrese nt	The percentage of days the Faculty were present	White	The percentage of White students
Dropout_Rate	The dropout rate of schools	Underperformance	Whether schools underperform (Yes or No)
Total_Enrollment	The total enrollment in a school	Suspension_Rate	The suspension rate of schools

Table A2: Attributed used in Final dataset

Using this dataset, classification techniques such as Logistic Regression, Support Vector Machine, Tree and Random Forest classifiers, some bagging and boosting classifiers such as Bootstrapped Decision Tree, Gradient Boosting, and AdaBoost are then run on the whole dataset and feature subsets to identify which schools are underperforming. We investigated several models to determine which factors have the greatest effect on the school’s performance, as represented by “Under_Performance” variable. Most algorithms were implemented with the defaults from scikit library in Python. The result can then be used to help schools improve their performance.

As for the evaluation metrics, we want to determine which of these classifiers is most effective at predicting school performance. Given the imbalance between our two classes (only ~7% of schools in our 2,516 school samples underperformed), overall accuracy is not a particularly useful metric; we could obtain a 93% accuracy simply by predicting schools would never underperform, regardless of how underperforming they are. Instead, we used precision-recall on

the minority class to evaluate the model performance. In this context, precision - recall can be interpreted as follows:

Precision: Proportion of schools that actually underperformed out of all predicted underperforming schools. A higher precision indicates that our model is only choose actual "at risk" schools, which would lead to more cost-efficient strategy.

Recall: Proportion of underperforming schools correctly identified. A higher recall indicates that our model will predict more of the at risk schools, and therefore leads to more effective strategy. Our goal is to maximize Recall without Precision being low.

Subject Analysis

Partnership of Assessment of Readiness of College and Careers, or PARCC, is an examination testing English Language Arts and Mathematics for grades 3 to 11. Its goal is to "... provide high-quality assessments to measure students' progress toward college and career readiness," (Score Report Interpretation Guide). The NJ Department of Education's data provided us with each school's average PARCC by subject, grade, and student group (homeless, disabled, etc,), gender, and race. The state's average PARCC score per instance was also presented in the data file. Each instances' average PARCC subject score was compared to the state's average to determine proficiency of English and Math subjects for grade levels 3 to 11. Analysis of proficiency will assist us in determining which subject needs improvement. This will push us in the right direction as to what subject Aviv Learning's application should target to improve students' test scores.

Using the same 53 data sheets provided by the NJ Department of Education, we used the following attributes in Table B1.

Attribute	Range	Type
ELA Average PARCC Score per Student Group	650-850	Numeric
Math Average PARCC Score per Student Group	650-850	Numeric
State Average Score per Student Group	650-850	Numeric
Students to Administration Ratio	1-500:1	Ratio
Students to Teacher Ratio	1-500:1	Ratio
Devices to Student Ratio	1-29:1	Ratio
Chronic Absences per Grade	0-30%	Numeric
Education of Teachers	Masters, Bachelors	Nominal
Student Group	14 Categories of Race, Gender, Homeless, and Disabilities	Nominal
Internet Speed Met	Y/N	Nominal
Teacher's Average Years of Experience	1 to 17	Numeric

Table B1: All attributes used for analysis with their range and data type.

English Language Arts and Mathematics performance on the PARCC will be the class to predict. The listed attributes were chosen because each of them contributes to how well a student does in class. The ratio of students per administration worker and teacher could impact student performance because there should be enough of each to help them succeed. The number of devices, such as computers, per student is another essential attribute because this determines how

much technological resources are given to student to study and do their assignments. Also internet connectivity is needed and should be provided at optimal level for students research for assignments and complete their homework. Teachers' level of education and experience could impact how well a student does on their PARCC exam because teachers should be providing them the knowledge and test taking skills to succeed on the exam. Those with a certain amount of experience and education could do a better job at preparation than others. Lastly, a school's number of chronic absences could cause a decrease in PARCC performance. If students are not participating in school for an extended period of time, they are not benefiting from in class exercises and lessons that will prepare them for the PARCC. All these attributes should contribute to predicting proficiency for Math and English sections of the PARCC.

All analysis was done in Python using their machine learning, data processing, and graph creating libraries such as pandas, sklearn, and matplotlib. All above attributes had a seperate excel sheet and were merged together, using pandas, based on their primary key, "SCHOOL ID", a combination of each schools' county, district, and school ID value. Then we separated the file into English and Math subjects so analysis can be done on a per subject basis. First we preprocessed the predicting class. Students who score above the state's average are considered high proficient and have the ability to succeed in college and beyond. Those below average are deemed as students who need improvements in their studies. We created a function that sorted each instances' PARCC average score into high and low proficiency based on if they scored above the state's average or below the state's average, respectively. This resulted in a binary variable which will be predicted in supervised classification machine learning techniques. Attributes that also included a state average were teachers' average years of experience and chronic absences per grade level were converted to categorical variables based on if each attribute did high, low, or average compared to the state's average. Teacher's education level, if internet speed was met, and student/racial groups were already categorical variables, but were converted to numerical variables so they can be interpreted in machine learning classifiers. The number of students per administrator and teacher, and number of devices per student were numerical values with a variety of ranges and no average to compare to. Standard Scalar is a function from sklearn which standardized each of the attributes' values so all values are of equal variance and can be comparable to the categorical values in the model. Instances that had missing values for any one attribute was eliminated from the dataset. If one attribute is missing for one school, then the whole school would be deleted. Out of 2,517 schools, only 9 schools were not used for analysis. After data has been processed and ready for analysis, the distribution of high and low proficiency, which is even and no need for more data processing, are displayed in Table B2. The number of instances from the Math and English subject datasets per grade is displayed in Figure B1. There are less instances as the grades increases so when comparing grades' proficiencies to each other we will use specific methods for comparison will be discussed in the results section.

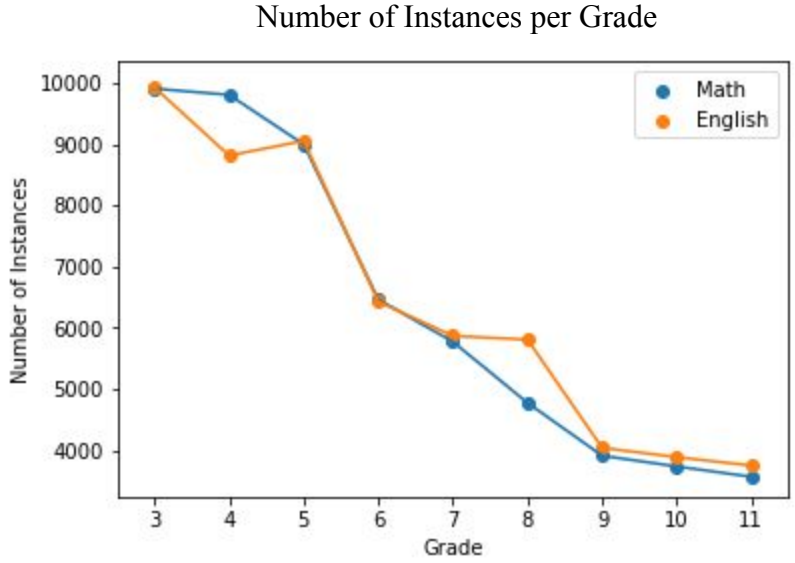


Figure B1. Line graph showing the number of instances in Math and English datasets per grade.

Proficiency	Math	English
Low Proficiency	53.14%	50.35%
High Proficiency	46.86%	49.65%

Table B2. Distribution of Low and High Proficiency for Math and English subjects from the PARCC exam.

The algorithms used for classification were Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression. Those classifiers were chosen because this problem is predicting a supervised binary outcome. Also our literature search suggested decision trees are a good method for predicting how well students would do in class (Priyam et. al.). First we created the response variables from the "Proficiency" column in each dataset and removed that column from the sets of features. English and Math feature and class datasets were each split into 30% test data and 70% training data using the cross validation function from the sklearn library. Each training set fitted into the classification models and predicted the testing data.

The metrics to determine which model will predict what subject is underperforming will be precision and recall. The model should be able to predict as many accurate true positives (low proficiency) as possible while minimizing its predictions of false negatives. Our priority is to find which subject students need the most support in. Inaccurately predicting subjects as low proficiency when students are actually doing well may lead to providing extra support. That would be a waste of money and efforts when there is a subject that students really need help with.

Expenditure Analysis

Districtwide performance report from year 2016-17 was downloaded from NJ's 2016-17 public school performance report . The detailed explanation of these reports can be obtained from https://rc.doe.state.nj.us/Documents/ReferenceGuide.html.The main tools used for this part of the project were Microsoft Excel and Weka.

The Data consisted of summary of districtwide data for 673 unique public school district codes including 88 Charter School Districts. The data reported can be broadly categorized into Demographic characteristics, Academic Achievement characteristics and School climate & environment characteristics. The data pertaining to 18 different attributes (table C1) were aggregated in a single Excel worksheet using Vlookup function on "DISTRICT_CODE". Null and missing data values were filtered and replaced by State mean/median values for some attributes and by zero for some others depending on type of attribute. Numeric attributes were discretized to test Apriori association rule mining algorithm and also to test improvement in classification algorithms using discretized version of attributes.

Attributes	Description	Attributes	Description
Per Pupil Expenditure	Total average expenditure (federal + state) per pupil in each district	Economically disadvantaged students	-
Teacher Avg. yrs experience	Teacher's Average number of year of experience in district	Disabled students	-
Admin avg, yrs experience	Administrators average number of year of experience in district	Hispanic students	-
Student/Teacher ratio	Average number of students per teacher in district	African American students	-
Student/Admin ratio	Average number of students per administrator in district	Asian students	-
ELA literacy performance	English language standardized test performance	White students	-
Math Performance	Math standardized test performance	ELA median growth rate	Calculated using the Student Growth Percentile (SGP) methodology.

Chronic Absenteeism rate	absentee rate is equal to or greater than 10%	Math median growth rate	Calculated using the Student Growth Percentile (SGP) methodology.
Violence Rate	Number of incidents per 100 students	Type of district	Charter or Non-Charter

Table C1: Attributes used in the analysis

For Exploratory data analysis top 10 and bottom 10 districts in terms of average total expenditure per pupil were studied in detail and line graphs were plotted for these 2 series for all attributes to visualize the underlying differences between them. Attributes were also visualized using bar graphs for percentage of Charter and Non-Charter public school districts that were above or below the third and first quartiles for that attribute respectively. These bar graphs helped identify the key attributes where the two classes significantly differed. Lastly, different excel worksheets were converted to csv files ready to be imported into Weka for testing classification algorithms.

Result:

Predicting the School Performance

Exploratory analysis

The final dataset in this part has 20 attributes with 2516 instances (schools) as described above. We first did some exploratory analysis to obtain the overall state performance.

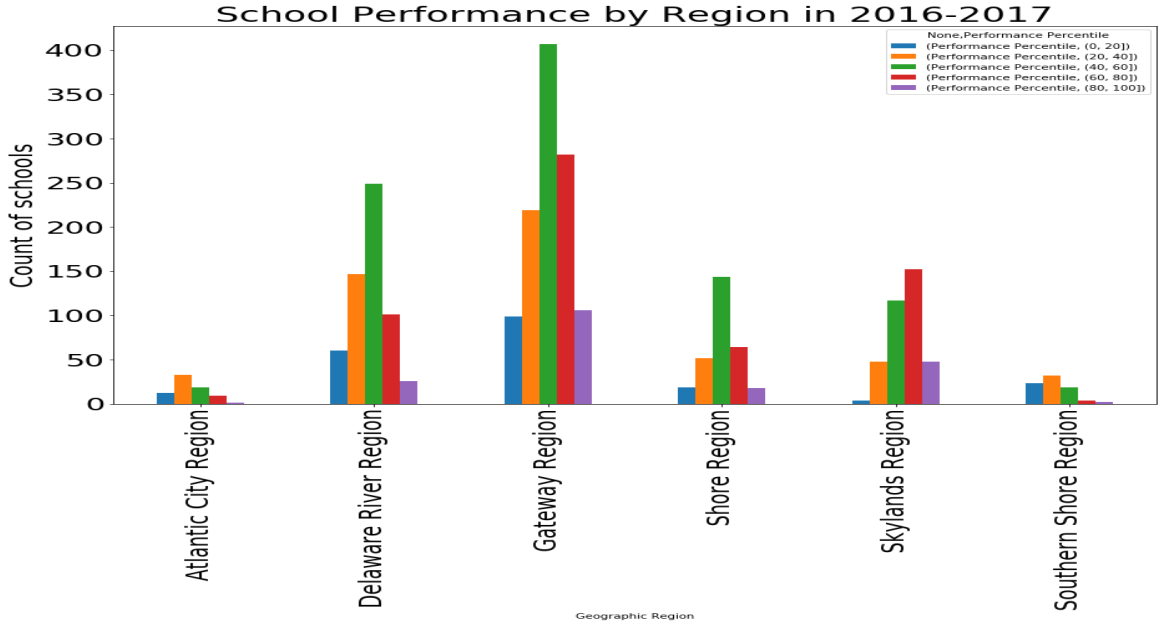


Figure A1: School Performance by Region in 2016-2017 in NJ

During the school year 2016-2017, NJ has 2516 schools with 1,373,521 students. The below graph shows the percentage of schools in a given region receiving a particular Accountability Percentile. It is worth noting that the Gateway Region has the highest percentage across different percentiles.

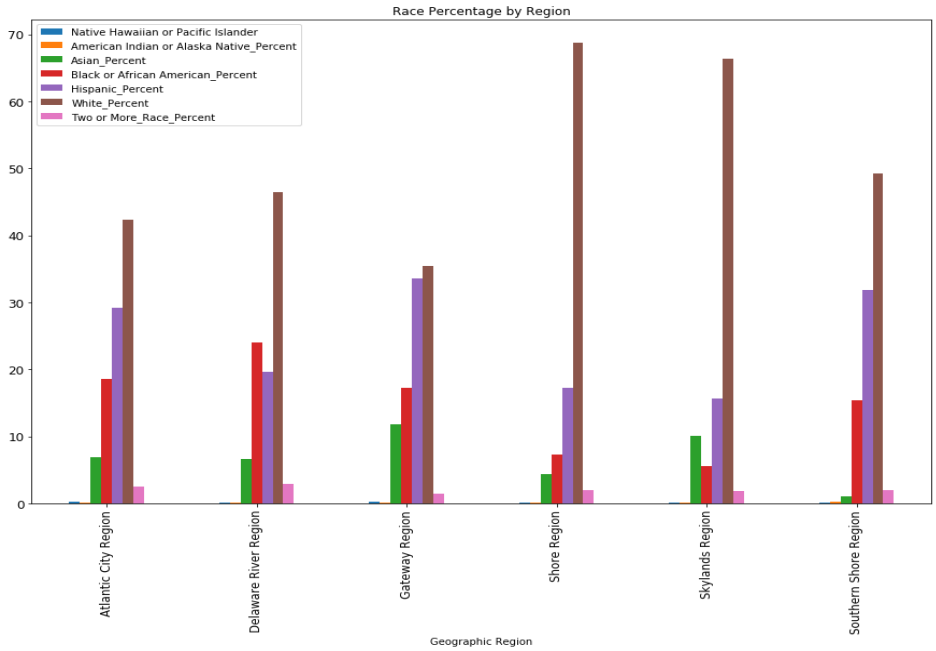


Figure A2: Race Percentage by Region in NJ

A closer look was taken at the percentages of the racial demographics of various regions. The plot showed that White students and Hispanic students are the most populous demographic in all regions. Students of Black or African descent are most numerous in Delaware and Atlantic Region. The Asian student population is highest in the Gateway and Skyland regions. Two or more race students are distributed evenly throughout all regions.

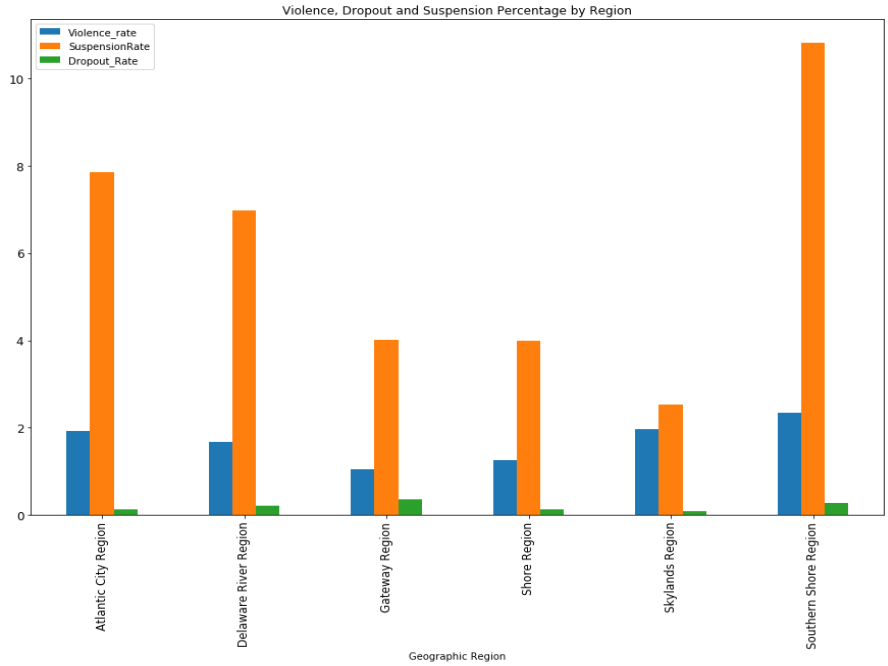


Figure A3: Violence, Dropout, and Suspension Percentage by Region in NJ

We did exploratory analysis concerned the rate of suspensions, violence-weapon-HIB, dropouts. The figures show that the suspension and dropout rates are lowest in the Skyland Regions while Gateway and Shore Regions have the lowest violence-weapon-HIB rate. Suspension rates are particularly high in the Atlantic City, Delaware and Southern Regions. All these rates are particularly highest in Southern Shore Region, which includes Cumberland County and Cape May County.

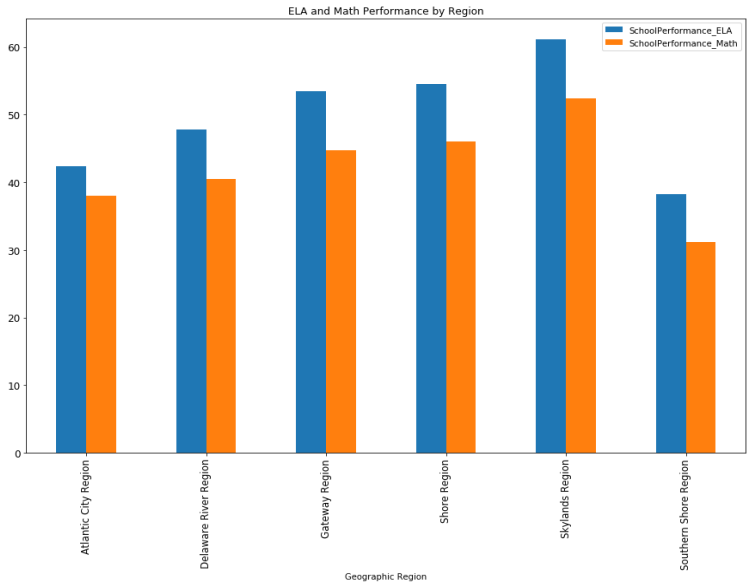


Figure A4: ELA and Math performance by Region in NJ

Next, we looked at the the average performance on Math and English of all regions. The Skyland Region has the highest percentage of its schools receiving a highest grade on both Math and English while Southern Shore Region's Performance is the lowest.

Model Building

Given the imbalance in the class distribution (only ~7% of schools in our 2,516 school samples underperformed), we tried undersampling the majority class (e.g. schools not underperformed) when training the model. Because the model will seek to maximize accuracy, it will often underestimate predictions for the less common observations. Therefore, to handle with this imbalance problem, we use undersampling, which is a common technique when dealing with imbalanced classes. As mentioned in the Approach part, we implemented the following models using Python's sklearn library: Logistic Regression, SVM, Random Forest, Bootstrapped, Decision Tree, Gradient Boosting Classifier, AdaBoost Classifier. For all models, we implemented cross validation over ten iterations, withholding 30% of the data as the test set.

Below are our precision-recall curve before and after undersampling technique.

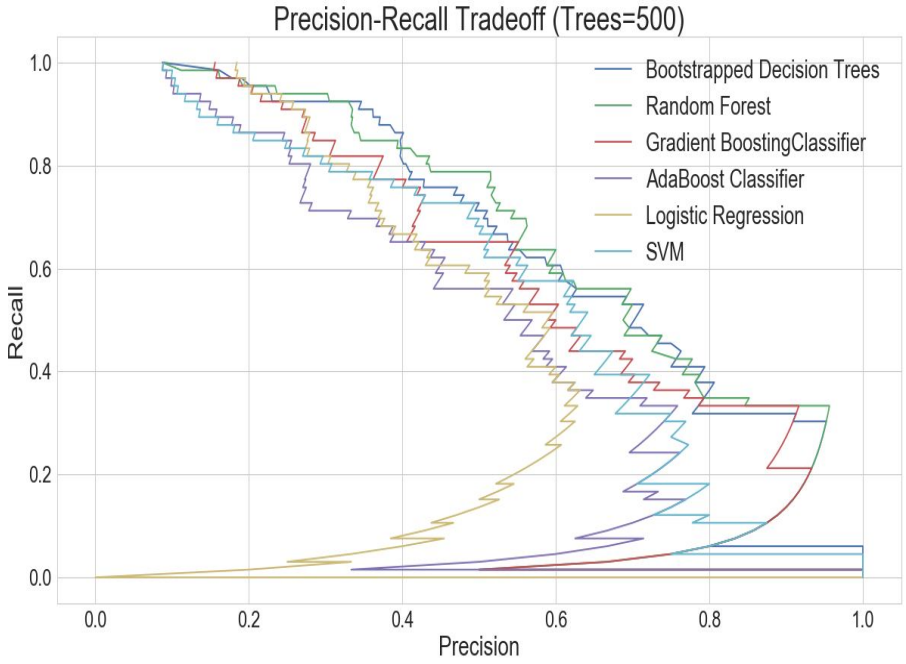


Figure A5: Precision-Recall curve before Undersampling

The curves above show the precision - recall tradeoffs for each classifier tested. Bootstrapped Decision Trees performed the best, but it still did not perform very well - only ~ 38% precision for ~80% recall. Could we do better?

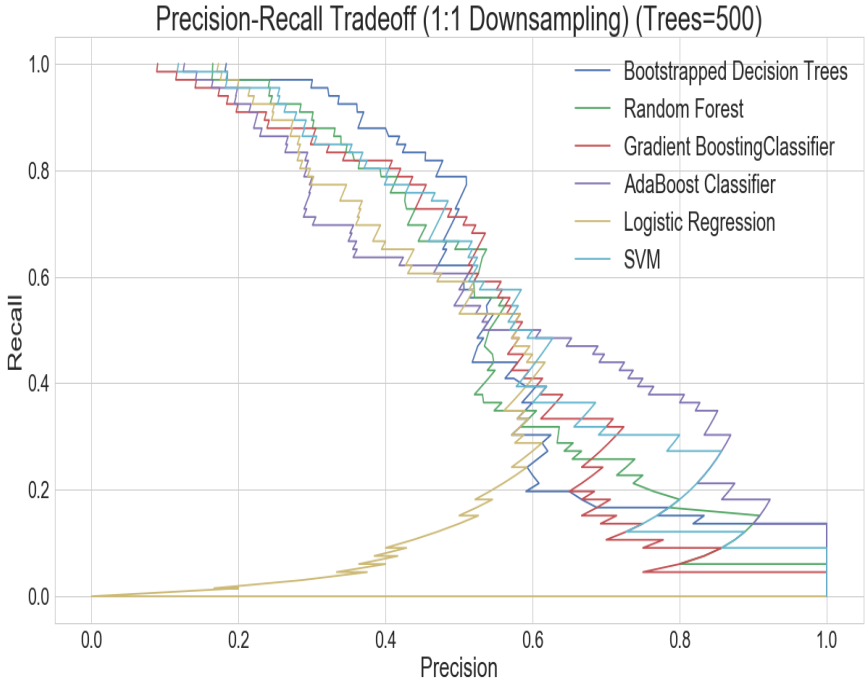


Figure A6: Precision-Recall curve after Undersampling

Next, after undersampling the majority class, Bagging Classifier or Bootstrapped Decision Trees still performed the best. We got an increase in the precision ~46% for ~80% recall.

We also looked at F1 score and Precision-Recall Area Under Curve. From the result, Bootstrapped Decision Trees performed better than other classifiers so we decided to choose it as our school performance predictor.

Model	Precision	Recall	F1 score	PR AUC
Logistic Regression	20	97	33	0.2
SVM	31	83	45	0.27
Random Forest	30	88	44	0.27
Boostrapped decision tree	**35**	**87**	**50**	**0.32**
Gradient Boosting Classifier	32	85	47	0.29
AdaBoost Classifier	33	86	47	0.29

Table A3: Classifiers' performance after Undersampling

We also explored various methods of feature selection for Bootstrapped Decision Trees algorithms to determine which features might be most useful in predicting school performance.

Univariate Feature Selection

Feature	Anova F-Value
SchoolPerformance_ELA	344.4
SchoolPerformance_Math	177.8
White(race)	117.4
Black or African American	85.7
Dropout_Rate	43.4

Table A4: Top 5 features by Univariate Feature Selection

We used the Anova F-value as the scoring function for the feature set, then selected the top 5 features with the highest score. The result showed that F-values for School Performance on English and Math were significantly higher than the rest.

Recursive Feature Elimination

Feature	Recursive feature support
SchoolPerformance_ELA	TRUE
SchoolPerformance_Math	TRUE
Suspension Rate	TRUE
White(race)	TRUE
Black or African American	TRUE

Table A5: Top 5 features by Recursive Feature Elimination

Recursive Feature Elimination works by assigning weights to each feature and prunes the one with smallest weight at each iteration. The table below is the result of Scikit's recursive feature elimination implemented in Python. It showed that the top 5 most important features are School Performance on Math, English, Suspension Rate, White race, Black or African American.

Tree-Based Feature Selection

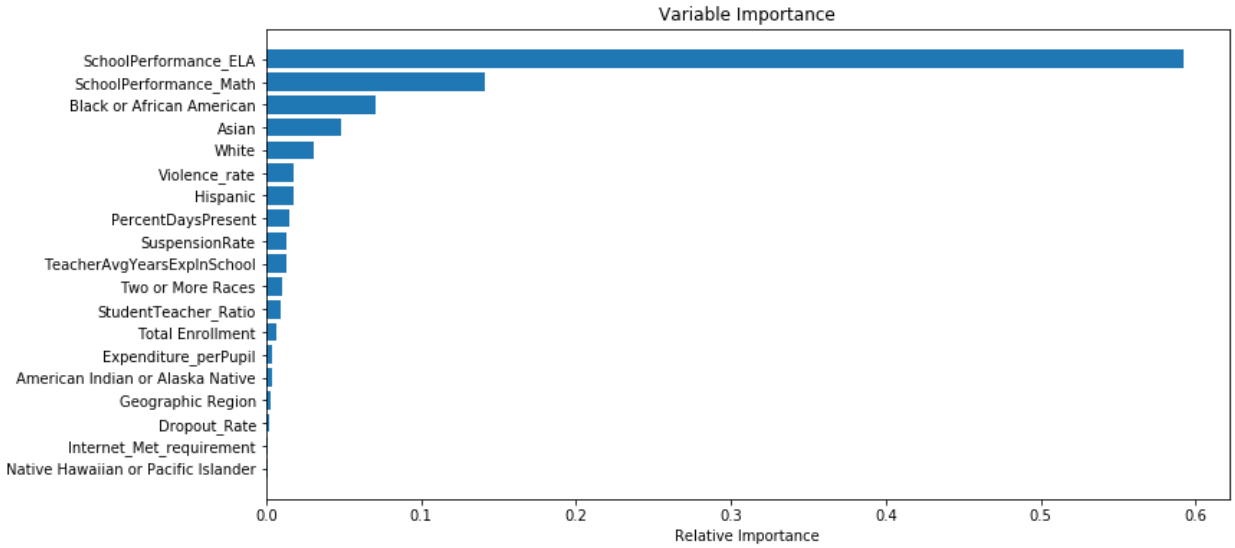


Figure A7: Feature importance with Bootstrapped Decision Trees

We used a tree-based feature selection to compute the feature importance in Bootstrapped Decision Trees model. Features with the highest importance score were selected. The below chart showed that top 5 features are school performance on Math, English, the percentage of Black or African American; Asian, White. Some features such as Internet_met_Requirement; Expenditure per student; Student-Teacher Ratio seem to have no impact on the overall school performance.

Subject Analysis

The resulting metrics are shown in Table B3.

Math	Precision	Recall	F-Measure	Accuracy
Random Forest	0.83	0.82	0.82	82.40%
SVM	0.69	0.69	0.69	68.63%
Decision Tree	0.82	0.82	0.82	82.39%
Logistic Regression	0.69	0.68	0.68	68.37%
English	Precision	Recall	F-Measure	Accuracy
Random Forest	0.83	0.83	0.83	83.26%
SVM	0.7	0.69	0.69	69.48%
Decision Tree	0.82	0.82	0.82	82.47%
Logistic Regression	0.69	0.68	0.68	67.61%

Table B3. Precision, recall, F-measure, and accuracy results from classification machine learning classifiers for Math and English datasets.

Performance of our model is presented in Figure B2, where accuracy is 82.4% and 83.26% for Math and English datasets, respectively. However, accuracy does not determine how well the model predicted the correct number of true positives. Precision of .83 for both Math and English is good compared to other models and feature selection. This means that 83% of true positive results, or the number of instances that were categorized as low proficient, were correctly predicted. Recall for Math is 82% and English is 83%. Table B4 compares the two highest precision and recall models' confusion matrices for both Math and English datasets, Random Forest and Decision Tree. Decision Tree does minimize the number of false negatives since it has the lowest compared to Random Forest. However, Random Forest predicted more true positives correctly, and ensuring that our model predicts the subject that requires the most help is our priority, hence Random Forest was the model chosen for analysis.

Math Confusion Matrices			
Random Forest		Decision Tree	
TP: 7855	FP: 1174	TP: 7523	FP: 1506
FN: 1830	TN: 6271	FN: 1500	TN: 6547
English Confusion Matrices			
Random Forest		Decision Tree	
TP: 7665	FP: 1219	TP: 7313	FP: 1571
FN: 1722	TN: 6963	FN: 1509	TN: 7176

Table B4: Confusion Matrices from Random Forest and Decision Tree classifiers.

Table B2 shows for both Math and English datasets Random Forest has the highest precision and recall results, hence that will be the model we use for analysis. Next we calculated which attributes are contributed most to the model. Tree Feature Selection function from sklearn calculated a weight of importance for each feature. Figure B2 demonstrates that for both Math and English datasets "Students per Admin, "Student Groups", "Grade", and the "Devices per Student" were the top four attributes the model depicted. Attributes below 0.1, which were "Internet Speed Met", "Teacher Experience Yrs", and "Teacher Education Level" were removed to see if precision increased. However, seen in Table B4, precision decreased so we will keep all the attributes in the model.

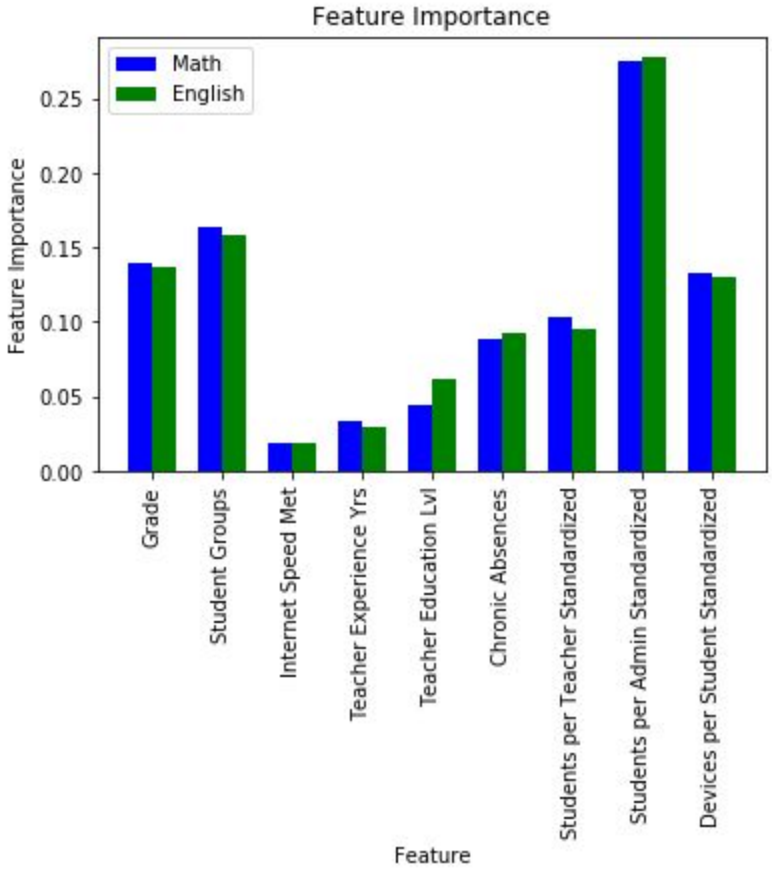


Figure B2. Feature importance levels for each attribute in the Math and English datasets.

After Feature Selection Results	Math	English
Precision	0.8	0.79
Recall	0.79	0.79
F-Measure	0.79	0.79
Accuracy	79.47%	79.04%

Table B4. Results after removing “Internet Speed Met”, “Teacher Experience Yrs”, and “Teacher Education Level” from the Random Forest model.

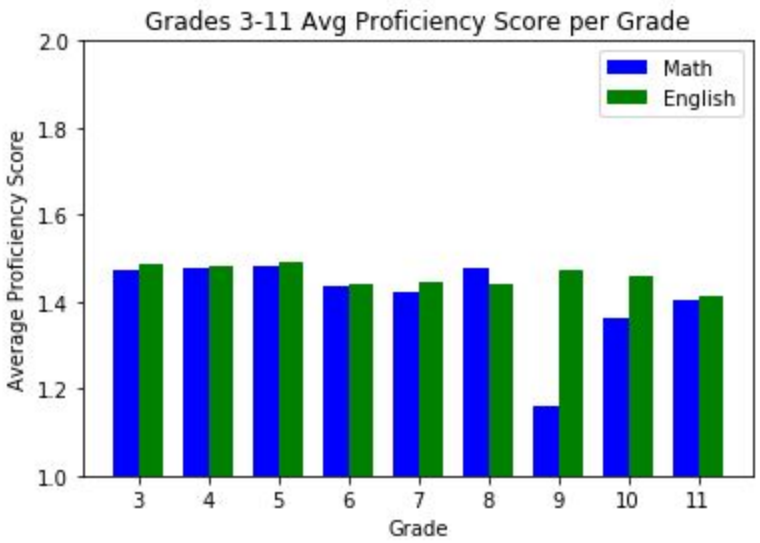


Figure B5: Average Random Forest model predicted proficiency class results per grade for Math and English datasets.

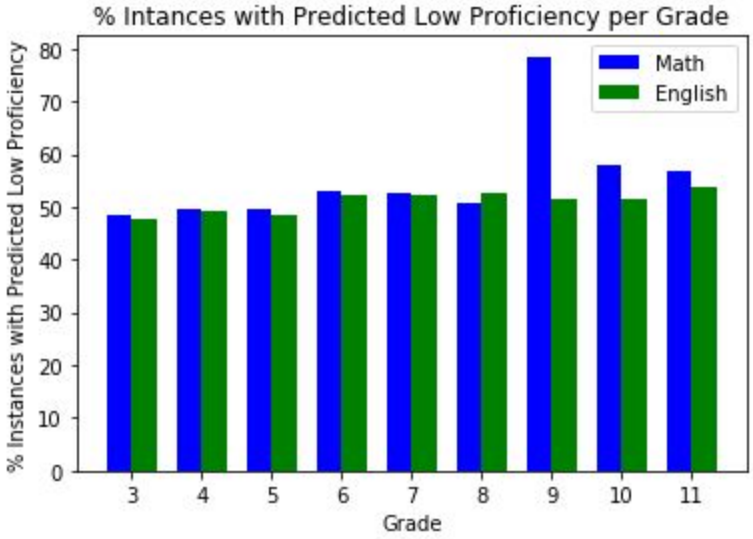


Figure B6: Percent Instances with Predicted Low Proficiency Class Results per grade for Math and English datasets.

Figure B5 shows that by the average amount of "Proficiency" categories predicted by the Random Forest model, 9th grade math has the lowest average score of 1.15, meaning they have more low proficiency categories than high proficiency. This method was chosen to neutralize effects of the wide range of instances per grade, as seen in Figure 1B. Taking the average value of proficiency categories per grade gives a better estimate as to which grade has the lowest proficiency so it can be comparable to other grades. Figure B6 also shows predictions from the Random Forest model that grade 9 math has 78.5% of instances that are in the low proficiency class. Grade 9 has the highest percentage of low proficiency compared to other grades.

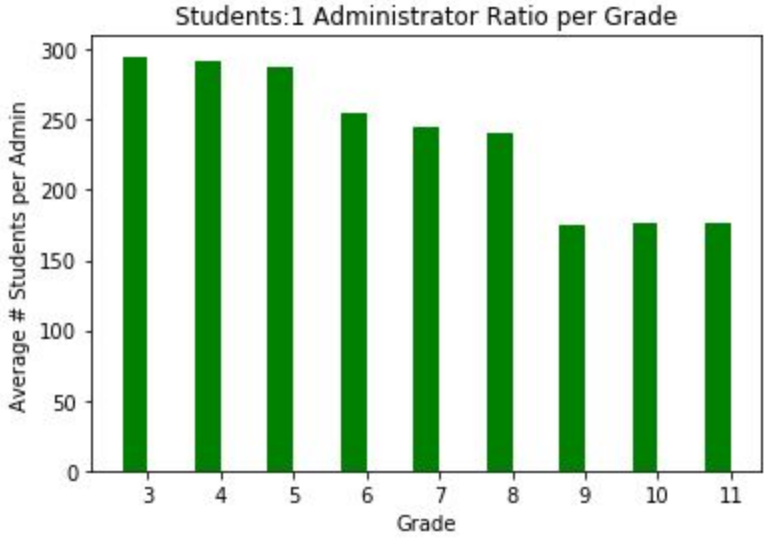


Figure B7: Average number of students per administrator personnel for each grade.

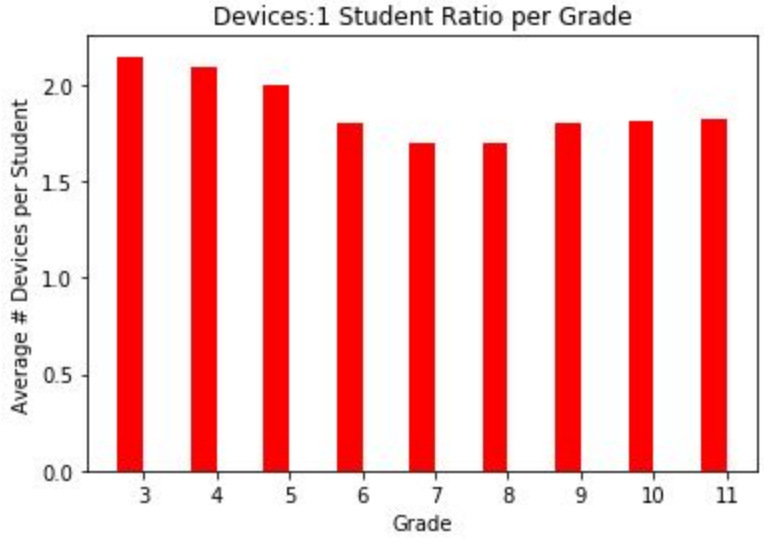


Figure B8: Average number of devices per school for each grade.

Figure B7 and B8 show the average number of students per administration personnel and average number of devices per student. High school grades (9-11) have the lowest number of students, with grade 9, at 175 students, being the lowest. The number of devices per student is highest at 3rd grade with 2.15 devices per student and lowest at 7th and 8th grade with 1.7 devices per student for both grades.

Expenditure Analysis

Before reporting our results for this section we would like to mention some limitations of our analysis as stated below:

Many important variables such as parents' educational level and family income level that have been shown to be important predictors of student success in may research studies were not factored in due to unavailability of data from our primary source.

Many other variables such as several exam scores for testing students' college readiness were not considered for this analysis since these exams are administered only for higher grades (8 and above) and the 673 districts considered for this analysis all serve different grade levels.

Standardized Math and ELA scores were the only measures used to determine academic outcomes since these tests are administered from grade 3-12 and data was available for most districts.

Variables such as graduation rates and dropout rates that could also be important predictors of quality were also overlooked since the data on these was only available for districts catering to 9-12th grade only.

Due to presence of outliers in the data and imbalanced class distribution of Charter vs Non-Charter school districts, comparisons between attributes was made using quartiles which are more robust to outliers and percentages of Charter and Non- Charters to address imbalanced classes.

Exploratory Data Analysis - part 1

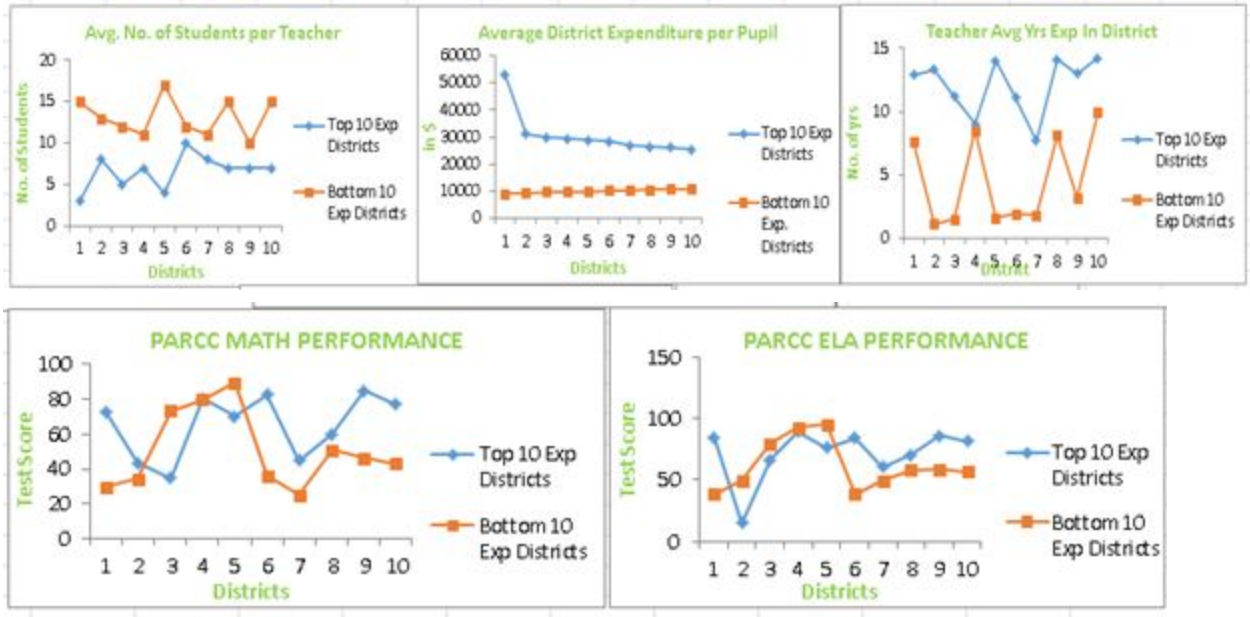


Fig 1.c

Observations

We get a glimpse of disparity in funds received by different districts NJ by looking at the above graphs of top 10 and bottom 10 districts by expenditure. Although the statewide average expenditure per pupil stands at $15585, the top 10 districts seem to be receiving an average of approximately $30,000 (excluding the outlier at $52,873) and bottom 10 districts receive an

average of approximately $10,000. Also, 7 out 10 bottom 10 districts expenditure wise are Charter districts.
The two clear reasons for this difference in expenditure were found to be student vs staff ratio which is much lower in general for top 10 districts and Teacher average experience in district which is generally higher for top 10 districts.
On the other hand no significant difference can be observed between the two categories when the PARCC Math and ELA English literacy performance scores are compared.
Although we cannot conclusively say this due to the limitations of our analysis mentioned above but the above graphs suggests that the State might be spending more on smaller class sizes (which means lower student vs staff ratio) and higher salaries for more experienced staff without significantly improving the quality of education.

Exploratory Data Analysis - part 2

Fig 2.c

Observations

The above bar graphs compares percentage of Charter and Non Charter districts based on the 17 attributes that were selected. Significant differences can be observed between the two classes based on the following attributes:
Teacher average years of experience in district (2nd from left)
Economically disadvantaged students
Students with Disability
Hispanic Students
African American Students
White Students
Administrators average experience in district
The above mentioned attributes (along with number of Asian students) were selected as a smaller subset of attributes to test classification accuracy in Weka in addition to the full set of attributes.

Weka Classification Algorithm Results:

The entire range of classification algorithms from simplest Zero R to sophisticated ones such as Neural Networks were run on numeric and discretized version of attributes. Full set and reduced set of attributes were also tested and cost sensitive classifier was also used. Moreover, Sample

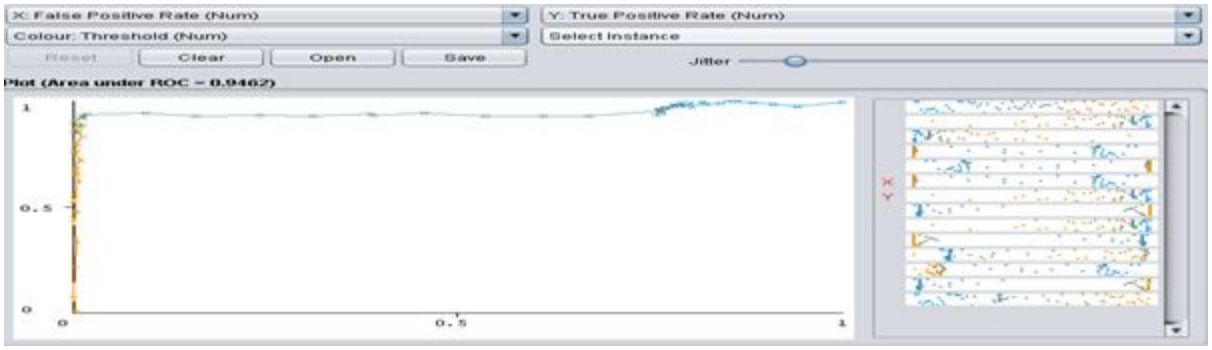
downsizing was also done to achieve a more balanced class distribution by randomly selecting 150 Non-charter Districts in Excel along with 88 Charter districts to get a sample of 238 instances. 10 fold cross validation was used for all iterations.
False Positive rate (Positive class = Non-Charters) was the main evaluation metric that was used to select top 3 algorithms namely Random forest, J48 decision tree and Logistic Regression. Other metrics such as ROC curves, Kappa Statistic, F measure and Accuracy Rate were also compared. Discretized version of attributes did not lead to any improvement in performance of algorithms except for Naïve Bayes, and therefore the results have not been reported here. Apriori algorithm was also run on discretized attributes to study association rules but did not give any meaningful outcome and hence has not been reported.

Random Forest: We see from the table below that using a more balanced downsized sample and cost sensitive classifier significantly improves the FP rate. The algorithm performance also improves significantly from the base accuracy of 63.0252% to 94.5378% when downsized sample is used. Also, using a smaller subset of attributes does not impact the metrics much. The area under ROC curve is around 95% which means the two classes "Charters" and "Non-Charters" have been very well distinguished by the algorithm.

ALGORITHMS	Full Sample Results : 673 instances , 18 attributes	Downsized Sample Results : 238 instances , 18 attributes
1. Random Forest	a. Accuracy = 97.474 %	a. Accuracy = 94.5378 %
	b. Weighted Avg. F-measure = 0.975	b. Weighted Avg. F-measure = 0.945
	c. Weighted Avg. ROC Area = 0.986	c. Weighted Avg. ROC Area = 0.949
	d. FP rate = 10/88 = 11.36 %	d. FP rate = 9/88 = 10.27%
	e.Kappa Statistic = 0.8872	e.Kappa Statistic = 0.8814
	f. ZeroR accuracy = 86.9242 %	f. ZeroR accuracy = 63.0252 %
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 attributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 94.1176 %	a. Accuracy = 94.1176 %
	b. Weighted Avg. F-measure = 0.941	b. Weighted Avg. F-measure = 0.941
	c. Weighted Avg. ROC Area = 0.959	c. Weighted Avg. ROC Area = 0.948
	d. FP rate = 5/88 = 5.68 %	d. FP rate = 6/88 = 6.82 %
	e. Kappa Statistic = 0.875	e. Kappa Statistic = 0.8744
	f. Zero R accuracy = 36.9748 %	f. Zero R accuracy = 36.9748 %

Table 1.c

Random Forest ROC curve for "Charter" class – Downsized Sample

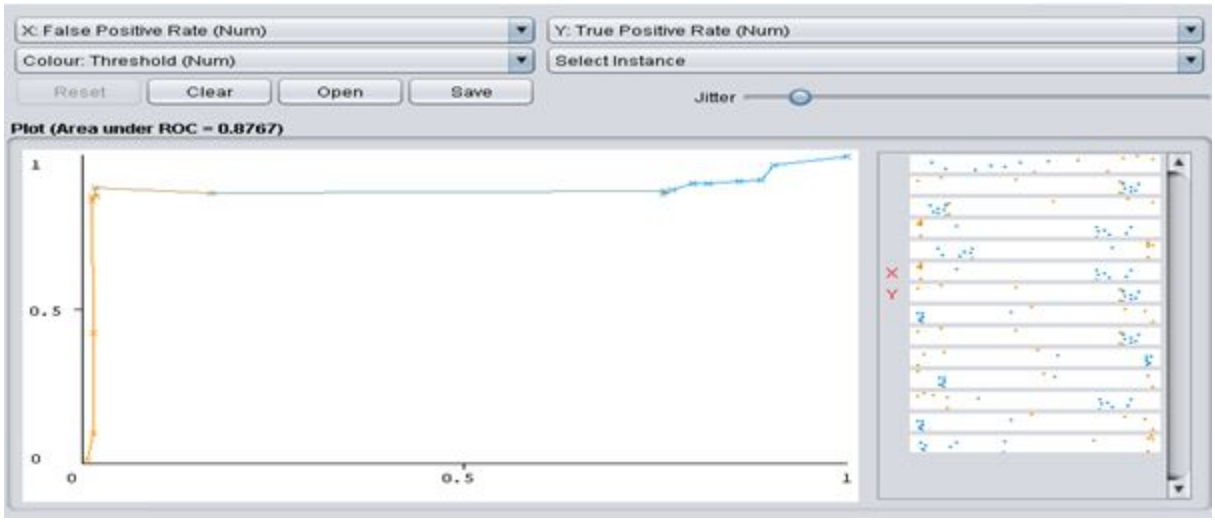


J48 : We see from the table below that using a more balanced downsized sample does not lead to any improvement in FP rate, on the contrary it negatively impacts all metrics. But when cost sensitive classifier is used it slightly improves the FP rate. Using a smaller subset of attributes does not impact the metrics much. The area under ROC curve is around 88% which is considerably less compared to Random Forest.

ALGORITHMS	Full Sample Results : 673 instances , 18 attributes	Downsized Sample Results : 238 instances , 18 attributes
2. J48	a. Accuracy = 96.2853 %	a. Accuracy = 90.7563 %
	b. Weighted Avg. F-measure = 0.963	b. Weighted Avg. F-measure = 0.908
	c. Weighted Avg. ROC Area = 0.924	c. Weighted Avg. ROC Area = 0.903
	d. FP rate = 11/88 = 12.5 %	d. FP rate = 11/88 = 12.5%
	e. Kappa Statistic = 0.8389	e. Kappa Statistic = 0.8017
	f. ZeroR accuracy = 86.9242 %	f. ZeroR accuracy = 63.0252 %
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 attributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 89.4958 %	a. Accuracy = 89.4958 %
	b. Weighted Avg. F-measure = 0.896	b. Weighted Avg. F-measure = 0.896
	c. Weighted Avg. ROC Area = 0.898	c. Weighted Avg. ROC Area = 0.899
	d. FP rate = 10/88 = 11.36%	d. FP rate = 10/88 = 11.36%
	e. Kappa Statistic = 0.7772	e. Kappa Statistic = 0.7772
	f. Zero R accuracy = 36.9748 %	f. Zero R accuracy = 36.9748 %

Table 2.c

J48 ROC curve for “Charter” class – Downsized Sample

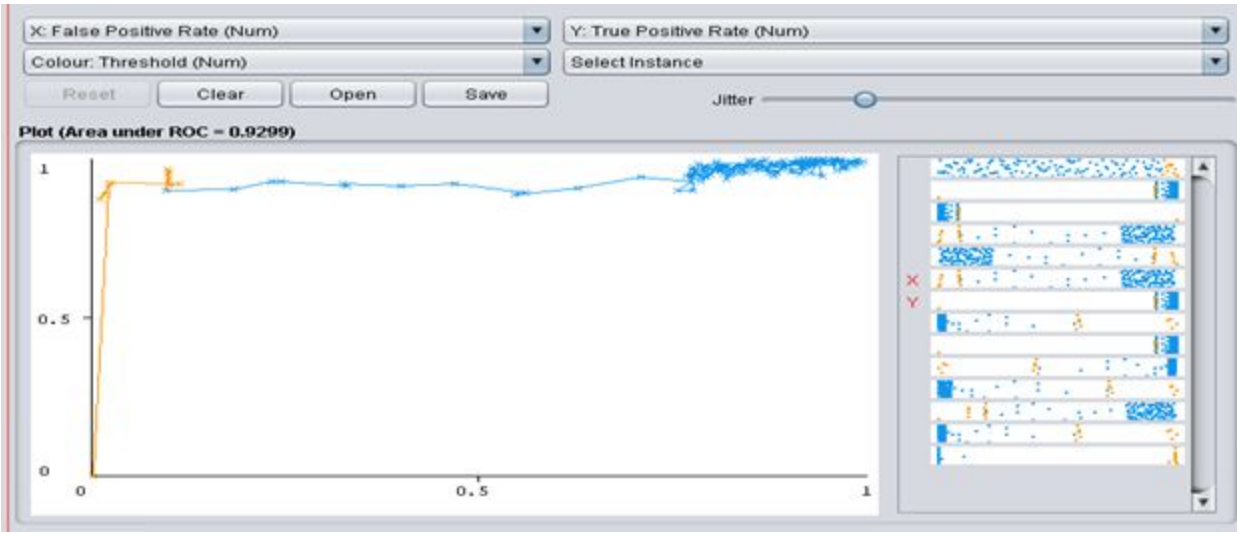


Logistic Regression: Here we see a significant drop in FP rate when a more balanced downsized sample is used, even though the overall accuracy decreases.However, using cost sensitive classifier or a subset of attributes does not improve classification performance any further.The area under ROC is around 93% which is again quite good.

ALGORITHMS	Full Sample Results : 673 instances, 18 attributes	Downsized Sample Results : 238 instances, 18 attributes
3. Logistic Regression	a. Accuracy = 96.8796 %	a. Accuracy = 92.0168 %
	b. Weighted Avg. F-measure = 0.969	b. Weighted Avg. F-measure = 0.921
	c. Weighted Avg. ROC Area = 0.977	c. Weighted Avg. ROC Area = 0.921
	d. FP rate = 11/88 = 12.5%	d. FP rate = 7/88 = 7.95%
	e.Kappa Statistic = 0.8621	e.Kappa Statistic =0.8307
	f. ZeroR accuracy = 86.9242 %	f. ZeroR accuracy = 63.0252 %
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 atributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 92.0168 %	a. Accuracy = 91.5966 %
	b. Weighted Avg. F-measure = 0.921	b. Weighted Avg. F-measure = 0.916
	c. Weighted Avg. ROC Area = 0.837	c. Weighted Avg. ROC Area = 0.939
	d. FP rate = 7/88 = 7.95%	d. FP rate = 7/88 = 7.95%
	e. Kappa Statistic = 0.8307	e. Kappa Statistic = 0.8222
	f. Zero R accuracy = 36.9748 %	f. Zero R accuracy = 36.9748 %

Table 3.c

Logistic Regression ROC curve for “Charter” class – Downsized Sample



Discussion

Predicting the school Performance

Across undersampling technique and multiple classification models, we observed a precision ~46% for ~80% recall from Bootstrapped Decision Tree. With feature selection methods, the most important features are school performance on English, Math, Suspension rate. This means that if a school improves upon those attributes, they should be able to improve their performance percentile. Surprisingly, other features like the the expenditure per student, whether schools meet Internet requirement, the student-teacher ratio do not have much impact on the school performance.

From this part of our analysis, we would like to market our product to **Southern Shore Region** (Cumberland County and Cape May County), where schools have low performance on Math and English. The area also has the highest percentage of Violence, Dropout, and Suspension rate.

Given that a large variety of algorithms are used in this analysis, future work will include investigating new features related to poverty, unemployment, parent education data when further investigating region-wise differences in school performance. Access to data on school performance in previous years should also be considered when drawing conclusions and making recommendations.

Subject Analysis

PARCC is an online assessment for grades 3-11 that tests students' knowledge on reading comprehension and critical thinking in both English Language and Math subjects. Figure B5 shows that 9th grade math, or Algebra I, has the lowest average Proficiency class score based on average score per instance for 9th grade on the Math section of the PARCC. This means that there were more low proficiency categorical values (which were valued at 1, high proficiency valued at 2) compared to other grades. Also 9th grade math has the highest percentage of students in the low proficiency category from the Math dataset at 78.5%, so 78.5% of students in 9th grade math are performing less than the state's average Math PARCC score. Examination of Algebra I PARCC Online Practice Test Answer Document, the test involves multiple choice and

open ended questions that test students in three types of ways. “Type 1: conceptual, understanding, fluency, and application...Type II: written arguments/justifications, critique of reasoning… Type III: modeling/application of real world context…,” (Informational Guide to PARCC Math Summative Assessment Algebra I). Open ended questions have a scale from levels 1 to 5, where 5 is the maximum score per question, while 1 is the lowest. Students who answer questions at a 5 level answer completely by justifying their work and getting the correct answer. However, level 1 answers are not complete and only explain a minor portion of the question and is incorrect. If most 9th grade students perform poorly on the exam, it could be because they are not learning how to interpret their answers and complete their work. If Aviv Learning were to make a mobile game, it should incorporate understanding the methods of Algebra I and allow students to justify how they got to their answer and what it means in a creative and engaging way.

Figure B8 does not show obvious significant results for 9th grade for the number of devices per student. Figure B7 shows that for 9-11th grade has the lowest amount of students per administrator personnel on average, with 9th grade being the lowest. Compared to the other grade levels, 9th grade should have enough administrators to help them achieve better PARCC Algebra I scores if they seeked help. However, it is possible that those high schools have too many administrators who are unsure how to engage students to learn and study more effectively and are focusing on other areas of the school. For example, “...47 percent of [school administrators] believe they already spend a disproportionate amount of time managing disciplinary issues...” (School Administrators: An Occupational Overview). They should cut that percentage and spend some of their time improving students’ school performance. Administrators can then focus on those students who need improvement on their PARCC scores, particularly Algebra I, using Aviv Learning’s gaming application.

Expenditure Analysis

From this part of our analysis we would like to make following recommendations to the NJ State education department:

Study cost vs benefits of small class sizes and if possible decide on an optimum class size based on research recommendations since small class sizes significantly increase the expenditure per pupil in terms of salaries paid to staff.

Link salaries paid to staff (Teachers, Administrators, Counsellors etc) based on some weighted average of factors such as experience, qualifications and students overall positive outcomes rather than just increasing salaries linearly with number of years of experience.

Since there is an ongoing debate of whether Charter schools add any value over common public schools, more data should be collected and evaluated from past years and also from other Charter Schools from different States to conduct a detailed cost-benefit analysis. Also hiring more experienced staff for Charter schools and funding equivalent to that of other public schools might improve performance outcomes. It should also be noted that Charter schools have more economically disadvantaged and minority students which might pose additional challenges for them.

Outlier districts in terms of various parameters should be identified and the underlying issues should be addressed in order to have a more balanced and equitable distribution of resources across all Districts in NJ and to provide every child an equal opportunity to succeed.

Lastly, NJ should evaluate its public educational system in relation to States that have similar characteristics (Demographic, cost of living etc) but consistently rank amongst top 15 States in terms of quality of education with a much lower outgo per student. Also, when determining quality of education within a State the variance of outcomes should receive an equal weightage as the mean outcomes.

We might have to look at more data across different years to check if we can spot similar patterns in the funding structure of NJ. If we have sufficient evidence to validate our claims regarding overspending in some areas we might be able to convince NJ State to cut down expenditure in these areas and instead invest in digital learning tools such as our educational application to improve performance outcomes at a reasonable cost.

Conclusion

EDM is a growing field that can beneficially impact schools at a student, school, and district level. Our analysis provides Aviv Learning with a starting point on how and where to begin their business journey in NJ. They can sell their product to the Southern Shore Region where those schools need the most improvement overall, create their product to help students taking Algebra I to improve their PARCC scores, and we provided the company with data to convince schools to spread their spending into successful learning digital applications. We were able to use a variety of supervised classification machine learning techniques on NJ Department of Education data that will make a positive difference for Aviv Learning and NJ public schools across the state.

To enhance our analysis, we would like to add more years to the dataset to see if there has been any trends overtime in performance of schools, PARCC test scores, and expenditures. NJ Department of Education provides another 11 years of historic data on their website which we would like to use to find those trends. Analysis on student's grades in school subjects would be another step to take to see if improvements can be made and if their grades reflect their standardized test scores. Also more data about school expenditures could be useful to determine a set price per student for Aviv Learning's product.

Primary and secondary state schools are a crucial part of a students' life of all backgrounds and ethnicities. Using EDM to find solutions to problems in the education system will provide teachers and administration a path to improving all student's academic careers and futures.

Work Cited

Aher, S. B., & Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. Knowledge-Based Systems, 51, 1-14

Hochschild, JL. *Social Class in Public Schools. Journal of Social Issues. 2003;59(4) :821-840* (Retrieved from: https://scholar.harvard.edu/jlhochschild/publications/social-classpublic-schools)

*Informational Guide to PARCC Math Summative Assessment Algebra I. (*2016). Retrieved from http://www.nj.gov/education/assessment/parcc/guides/math/AlgebraI.pdf.

Priyam, A., Abhijeet, Gupta, R., Ratheeb, A., Srivastavab, S. (2013).Comparative Analysis of Decision Tree Classification Algorithms.*International Journal of Current Engineering and Technology*, *3(2),* 334-337.

Romero, Cristóbal & Ventura, Sebastian. *(2007). Educational data mining: A survey from 1995 to 2005.Expert Systems with Applications. 33. 135-146. 10.1016/j.eswa.2006.04.005.*

School Administrators: An Occupational Overview. (2016). Retreived from http://dpeaflcio.org/programs-publications/issue-fact-sheets/school-administrators-an-occupational-overview/.

Score Report Interpretation Guide. (2016)/ Retrieved from http://www.nj.gov/education/archive/assessment/parcc/scores/Spring16ScoreReportInterpretationGuide.pdf.

Map of New Jersey. Retrieved from https://www.funnewjersey.com/upload_user/new_jersey_state_info/map_of_new_jersey_counties.htm

Work Log

Activities	Amy Le	Gilat Mandelbaum	Sakina Presswala
Literature Search	2/21/18: 3	2/21/18: 2:30pm - 5:30pm: 3	2/21/18: 10 am-1.30pm: 3.5
Project Proposal	2/22/18: 3 2/23/18: 7	2/21/18:7pm - 9pm: 2 2/22/18: 2:40pm - 5:40pm, 8pm-10:30pm: 6.5 8.5	2/21/18 : 10am - 12am: 2 2/24/18:11 pm-1 pm: 2
Proposal ppt	2/25/18: 2	2/25/18: 10:40am - 12:40pm : 2	2/15/18 : 11am-5pm: 6
Data Cleaning & Preprocessing	3/10-4/16: 25.5	4/9/18: 4:40pm-5:40pm:1 4/13/18: 4pm-7:30pm: 3.5 4/14/18: 7pm-11pm: 4 4/17/18: 2pm-3pm: 1 4/18/18: 11am-11:30am, 2:30-9pm: 6.5 4/21/18:9:30pm-11pm: 1.5 4/22/18: 12pm-6pm: 6 4/23/18: 10am-1:30pm, 3pm-8pm: 5.5 29.5	3/1/18-4/1/18: 20
Data Analysis	4/17-4/25: 20.5	4/24/18: 11am-1pm, 3pm-5pm: 4 4/25/18: 10am-12pm,1pm-6pm,7pm-12am: 12 4/26/18: 12pm-5pm:	4/1/18 - 4/5/18: 5 4/10/18 - 4/26/18 : 15

		5 27	
Poster Presentation	4/18-4/19: 5	4/14/18: 7:40am-8:40am: 1 4/16/18: 6pm-6:30pm: .5 4/19/18: 12pm-3pm: 3 4.5	4/10/18: 3
Final ppt	4/20-4/26: 3	4/25/18: 8am-9am:2 4/26/18: 11am-12pm: 1 3	4/20/18: 3
Final Report	4/28/18: 6 4/29/18: 4	4/28/18: 5pm-12am: 7 4/29/18: 10am-12pm, 4pm-9pm: 7 14	2/1/18-4/25/18 : 5
Total	79	91.5	64.5

Final Project Presentation

Improving Education in NJ Grade Schools with Digital Applications

with Aviv Learning

Data Analytics Final Presentation
Spring 2018

△

Amy Le
Gilat Mandelbaum
Sakina Presswala

RUTGERS
Professional Science Master's Program

Master of Business and Science Degree

Aviv Learning - Educational Mobile Gaming

Goal: *Engage students to study more efficiently and effectively*

Analysis Questions to market his product

1. ***Where?***
2. ***What?***
3. ***How much?***

CAN WE IDENTIFY SCHOOLS AT RISK OF UNDERPERFORMING BY PERFORMANCE AND OTHER CHARACTERISTICS ?

***Data**: NJ Department of Education (2016-2017)*

Class: Under_Performing

Yes: School_Accountability_Percentile: 0~20%
No : School_Accountability_Percentile: 21~100%

School Characteristics Demographics

- Geographic location
- Student-Teacher Ratio
- Teachers Average years_Experience
- Grade Type
- Technology_Met_Requirement
- Expenditure per Pupil
- Violence, Weapons, HIB
- Percentage of days Faculty were present

Student Performance

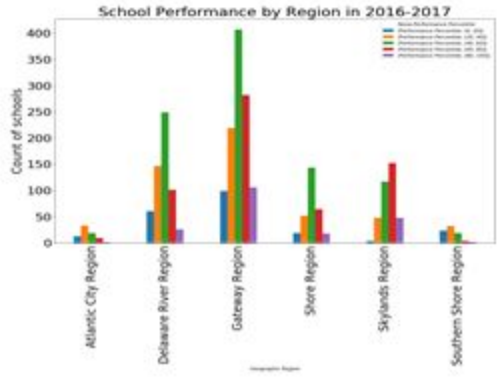
- Drop-out Rate
- Suspension Rate
- ELA performance
- Math performance

Student

- Total enrollment
- Racial and Ethnic Group

Overall State Summary(2016-2017)

Regions	Counties	Districts	Schools	Students
6	22	658	2516	1,373,521



Counties divided by regions

Atlantic City	Atlantic County
Delaware	Mercer, Burlington, Camden, Gloucester, Salem County
Gateway	Middlesex, Union, Essex, Hudson, Bergen, Passaic County
Shore	Monmouth County and Ocean County
Skylands	Sussex, Morris, Warren Hunterdon, Somerset County
Southern Shore	Cumberland County and Cape May County.

	Performance Percentile				
Performance Percentile	(0, 20]	(20, 40]	(40, 60]	(60, 80]	(80, 100]
Geographic Region					
Atlantic City Region	12	33	19	9	1
Delaware River Region	60	147	249	101	26
Gateway Region	99	219	407	282	106
Shore Region	19	52	144	64	18
Skylands Region	4	48	117	152	48
Southern Shore Region	23	32	19	4	2

DATA PREPROCESSING

- Created PK, Merged "school characteristics"; "student race", and "student performance" data with School_ID as index
- Mean/Mode imputation on missing data (13%/column has missing data) group by **the County**
- Label Encoding for Categorical attributes
- StandardScaler for Numerical attributes
- Create target class: 'Under_Performing'

  If 'School_Accountability_Percentile' < 20.0, 'Yes', 'No'
- df.shape() -> (2516, 20)

```
Data columns (total 20 columns):
Geographic Region                  2516 non-null object
Total Enrollment                   2516 non-null int64
Dropout_Rate                       2516 non-null float64
Violence_rate                      2516 non-null float64
SuspensionRate                     2516 non-null float64
Internet_Met_requirement           2516 non-null object
Expenditure_perPupil               2516 non-null int64
PercentDaysPresent                 2516 non-null float64
TeacherAvgYearsExpInSchool         2516 non-null float64
StudentTeacher_Ratio               2516 non-null float64
American Indian or Alaska Native   2516 non-null float64
Asian                              2516 non-null float64
Black or African American          2516 non-null float64
Hispanic                           2516 non-null float64
Native Hawaiian or Pacific Islander 2516 non-null float64
Two or More Races                  2516 non-null float64
White                              2516 non-null float64
SchoolPerformance_ELA              2516 non-null float64
SchoolPerformance_Math             2516 non-null float64
Under_Performance                  2516 non-null object
dtypes: float64(15), int64(2), object(3)
```

MODEL BUILDING

- Class ('Under_Performing') distribution:
  93% No
  7% Yes

=> highly imbalanced class.

- Split the dataset into training data and testing data (test_size: 0.3, stratified by class)
- Applied **Undersampling** the majority class to balance the class(Yes and No) distribution (1:1) in training data
- Classifiers:
  - Logistic Regression
  - SVM
  - Random Forest
  - Bootstrapped Decision Tree
  - Gradient Boosting Classifier
  - AdaBoost Classifier

EVALUATION : Precision - Recall Curve

RECALL

- Proportion of underperforming schools correctly identified
- High Recall=> Models focus on minimizing False Negative
  =>Models will predict more of the 'at risk' schools

=> more effective policy.

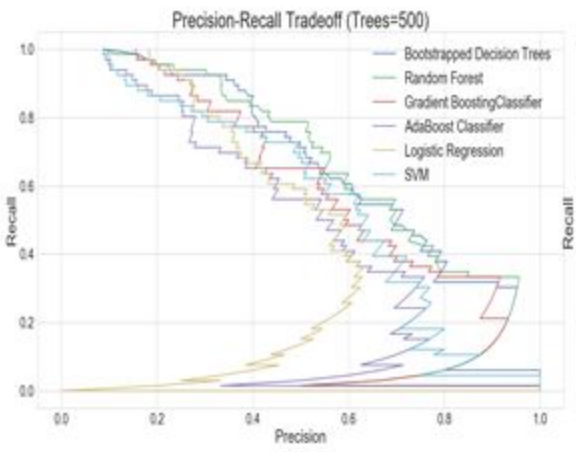
PRECISION

- Proportion of schools that actually underperformed out of all predicted underperforming schools
- High Precision => Models only choose the actual 'at risk' schools

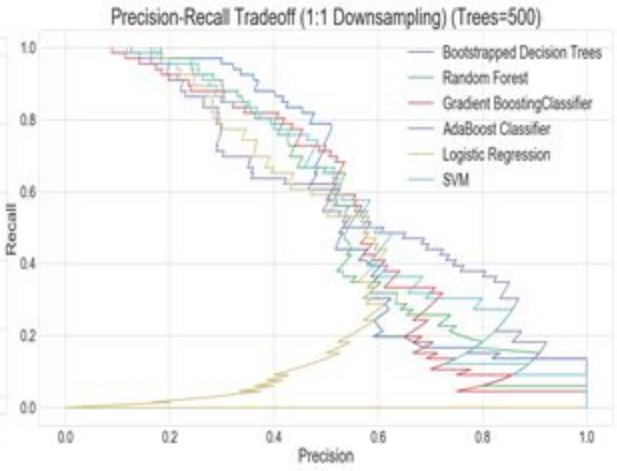
=> more cost-efficient policy

EVALUATION : Precision - Recall Curve

Before Undersampling



After Undersampling

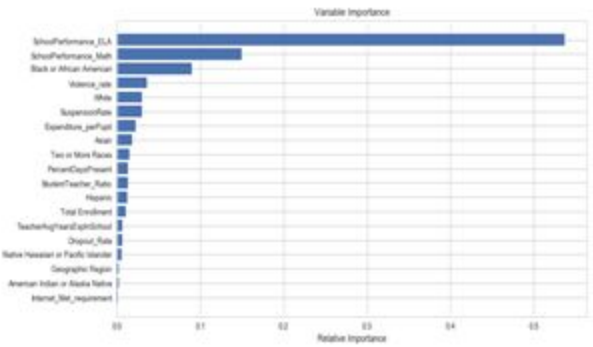


EVALUATION :

Model	Precision	Recall	F1 score	PR AUC
Logistic Regression	20	97	33	0.2
SVM	31	83	45	0.27
Random Forest	30	88	44	0.27
Boostrapped decision tree	35	87	50	0.32
Gradient Boosting Classifier	32	85	47	0.29
AdaBoost Classifier	33	86	47	0.29

FEATURE IMPORTANCE

- Tree-based Feature Selection

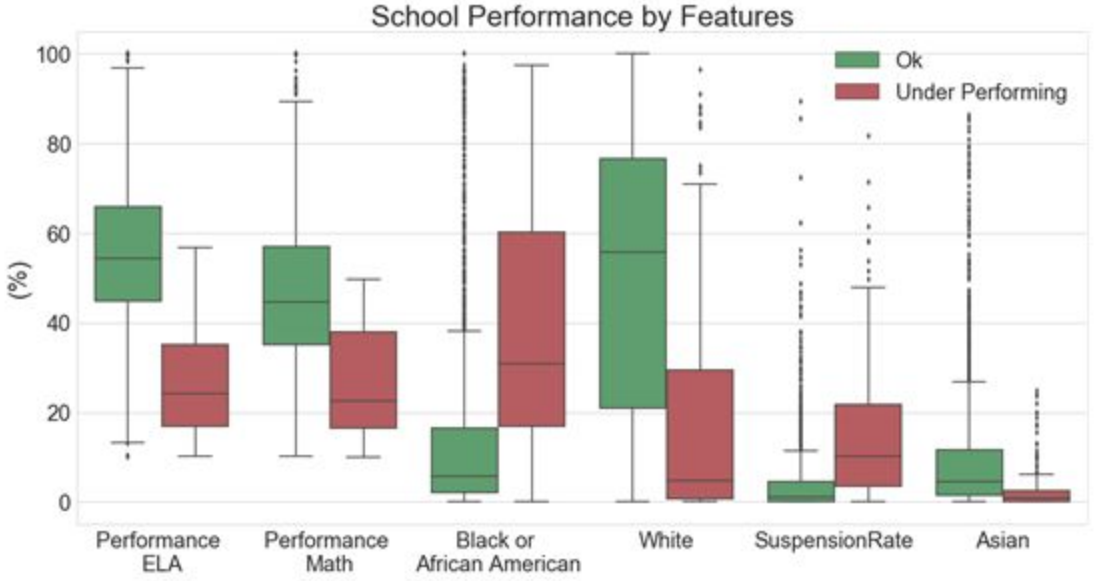


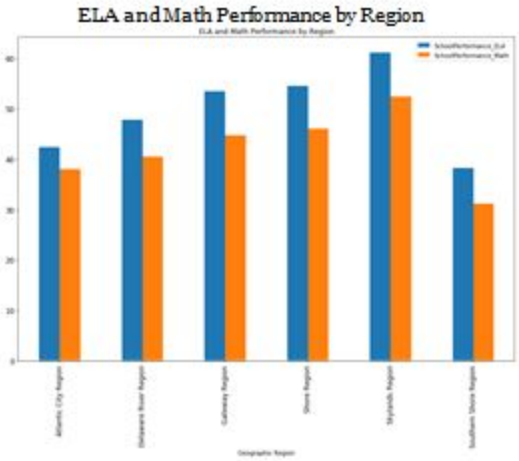
- Recursive Feature Elimination

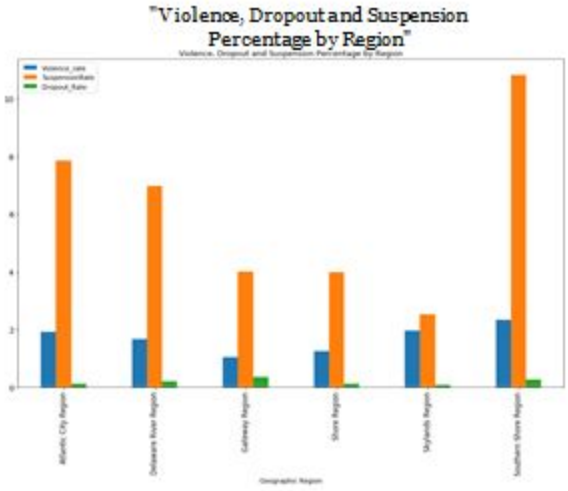
Feature	Recursive feature support
SchoolPerformance_ELA	TRUE
SchoolPerformance_Math	TRUE
Suspension Rate	TRUE
White(race)	TRUE
Black or African American	TRUE

- Univariate Feature Selection

Feature	Anova F-Value
SchoolPerformance_ELA	344.4
SchoolPerformance_Math	177.8
White(race)	117.4
Black or African American	85.7
Dropout Rate	43.4







=>> **Southern Shore**: Cumberland County and Cape May County.

***Data**: NJ Department of Education (2016-2017)*

Class: Proficiency

High: Student Mean > State Mean (Negative)
Low : Student Mean < State Mean (Positive)

School Characteristics Demographics

- ❑ Student-Teacher Ratio
- ❑ Student-Admin Ratio
- ❑ Teachers Avg Yrs Experience
- ❑ Grade Type
- ❑ Internet Speed Met
- ❑ Devices-Student Ratio
- ❑ Chronic Absences

Student Performance**

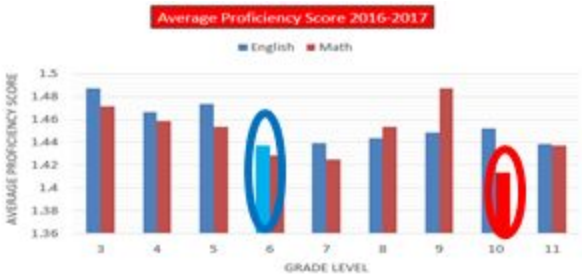
- ❑ ELA performance (3-11)
- ❑ Math performance (3-11)

Student

- ❑ Racial/Ethnic/Student Group

PARCC
Partnership for Assessment of Readiness for College and Careers

DATA PREPROCESSING

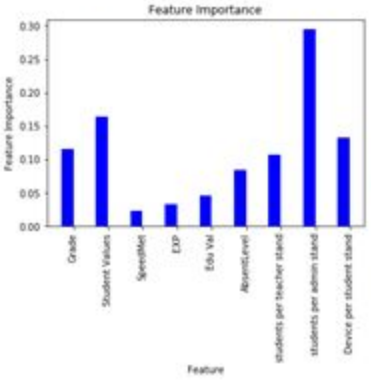


- Merged sheets using School ID as Primary Key
- Instances with missing values were removed
- Label Encoding for Categorical attributes
- StandardScaler for Numerical attributes
- Split data into Math and English, then into Primary and Secondary School
- Create target class: 'Proficiency'
  - Distribution:

Primary Schools				Secondary Schools	
	Low	High		Low	High
Math	50.30%	49.70%		64.70%	35.20%
English	49.90%	50.10%		52.20%	48%

Cross Validation: Split the dataset into training data and testing data (test_size: 0.3)

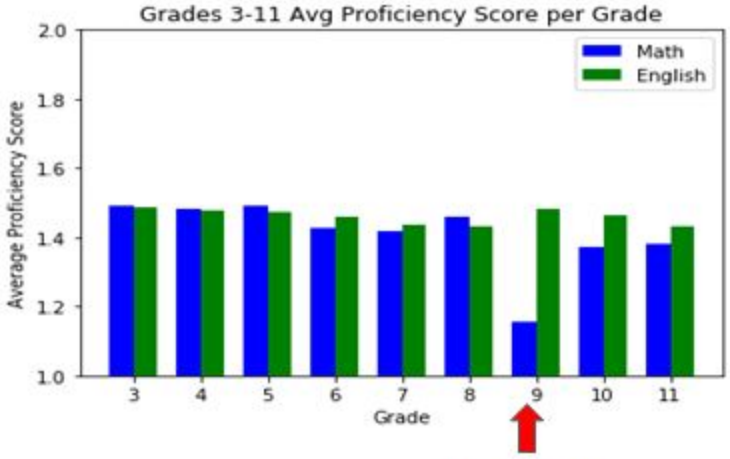
We want high precision and



Math Primary	Precision	Recall	F-Measure	Accuracy%
Random Forest	0.82	0.81	0.81	81.44
SVM	0.68	0.68	0.67	67.61
Decision Tree	0.81	0.81	0.81	81.4
Logistic Regression	0.68	0.68	0.68	67.62
Math Secondary	Precision	Recall	F-Measure	Accuracy%
Random Forest	0.84	0.84	0.84	84.24
SVM	0.72	0.73	0.71	72.82
Decision Tree	0.83	0.83	0.83	82.96
Logistic Regression	0.71	0.72	0.71	71.84
English Primary	Precision	Recall	F-Measure	Accuracy%
Random Forest	0.82	0.82	0.82	82.31
SVM	0.68	0.68	0.67	67.6
Decision Tree	0.81	0.81	0.81	81.5
Logistic Regression	0.68	0.68	0.67	67.51
English Secondary	Precision	Recall	F-Measure	Accuracy%
Random Forest	0.84	0.84	0.84	84.36
SVM	0.72	0.71	0.71	71.31
Decision Tree	0.84	0.84	0.84	83.76
Logistic Regression	0.72	0.71	0.71	70.91

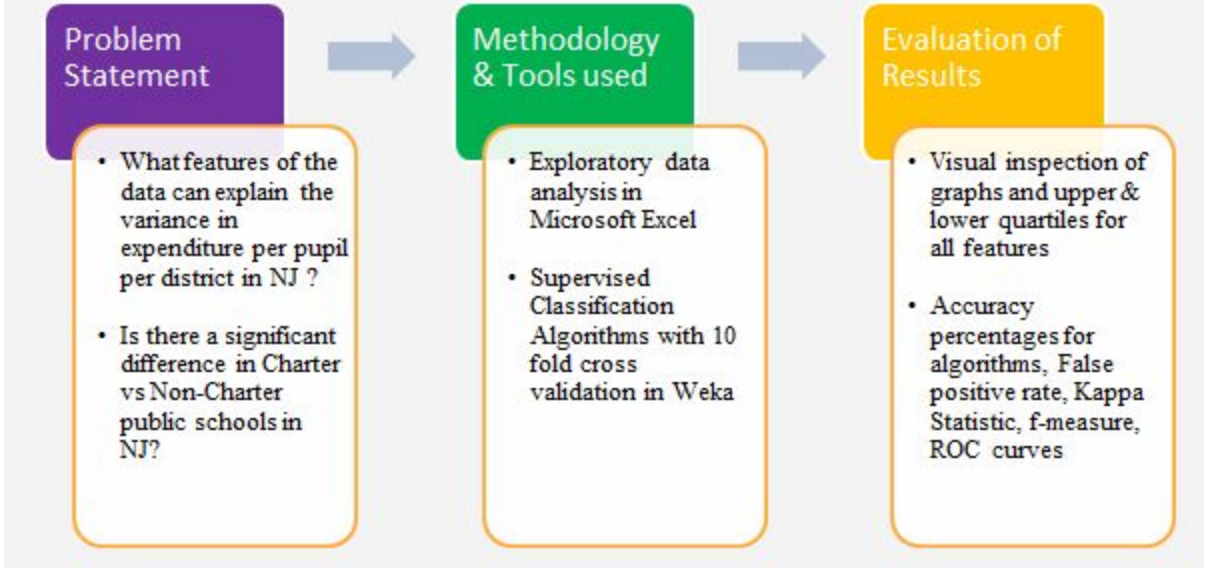
Predicted Results with Random Forest: By Grade

78.5% 9th Graders Classified as **Low** Proficiency

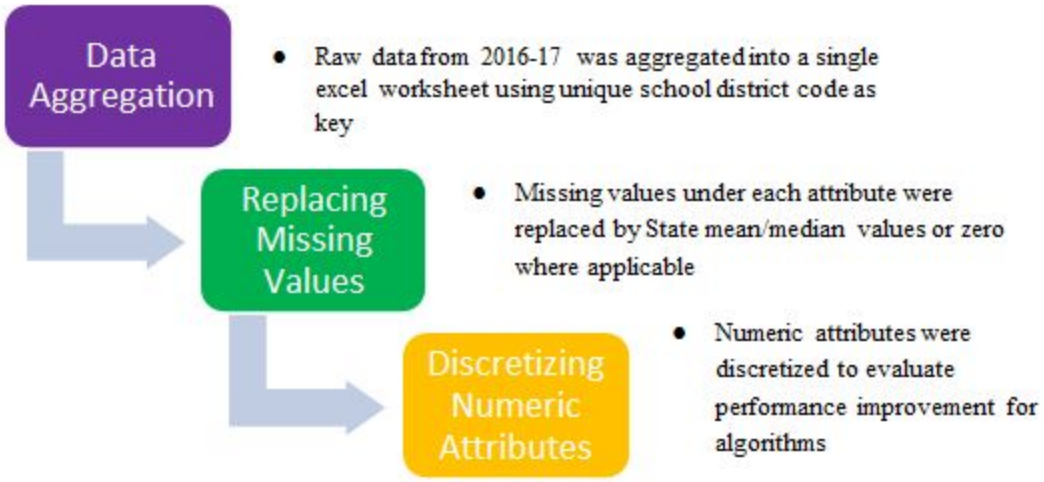


9th Grade Math = Algebra I Lowest Average Proficiency Score

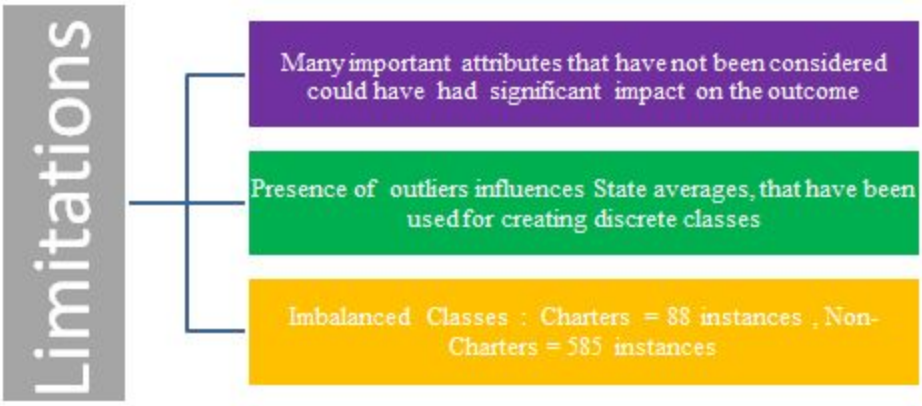
Overview



Data Preprocessing Steps

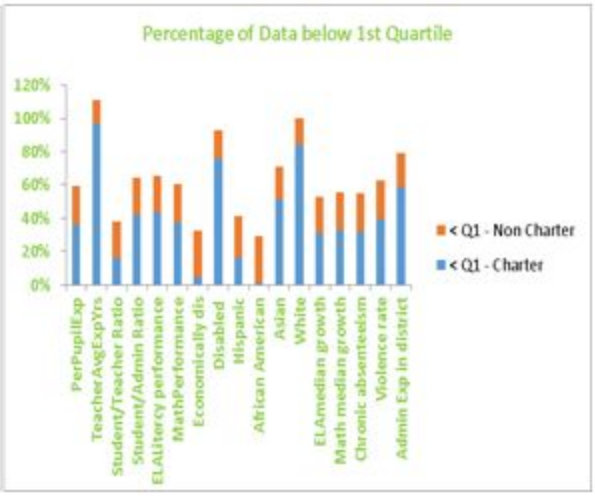


Limitations



Exploratory Data Analysis

Exploratory Data Analysis



Classifier Performance

ALGORITHMS	Full Sample Results : 673 instances , 18 attributes	Downsized Sample Results : 238 instances , 18 attributes
1. Random Forest	a. Accuracy = 97.474 %	a. Accuracy = 94.5378 %
	b. Weighted Avg. F-measure = 0.975	b. Weighted Avg. F-measure = 0.945
	c. Weighted Avg. ROC Area = 0.986	c. Weighted Avg. ROC Area = 0.949
	d. FP rate = 10/88 = 11.36 %	d. FP rate = 9/88 = 10.27%
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 attributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 94.1176 %	a. Accuracy = 94.1176 %
	b. Weighted Avg. F-measure = 0.941	b. Weighted Avg. F-measure = 0.941
	c. Weighted Avg. ROC Area = 0.959	c. Weighted Avg. ROC Area = 0.948
	d. FP rate = 5/88 = 5.68 %	d. FP rate = 6/88 = 6.82 %

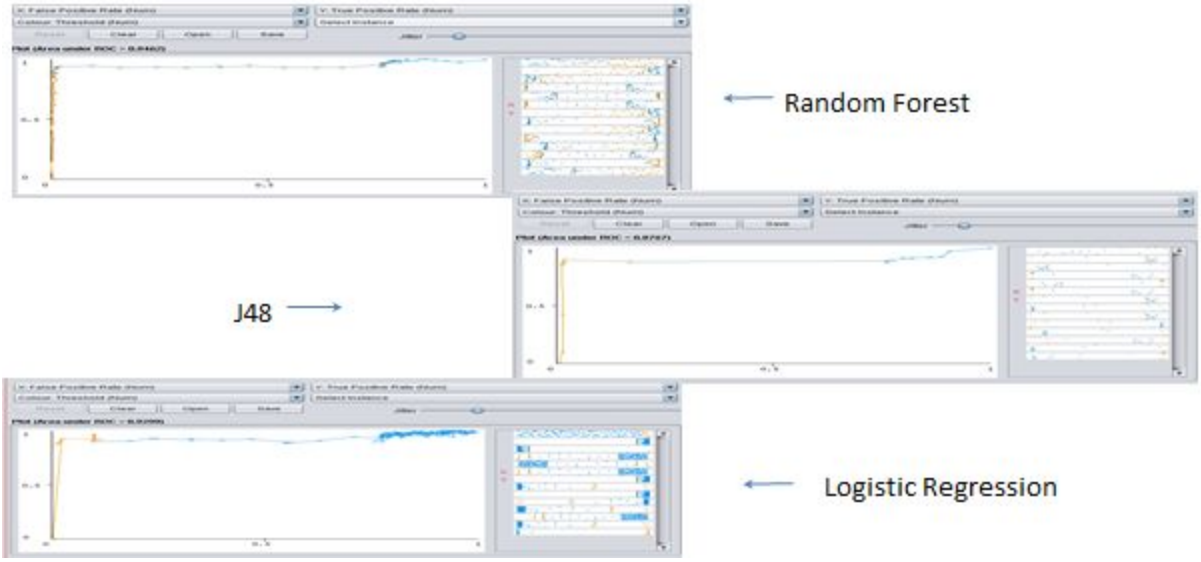
Classifier Performance

ALGORITHMS	Full Sample Results : 673 instances , 18 attributes	Downsized Sample Results : 238 instances , 18 attributes
2. J48	a. Accuracy = 96.2853 %	a. Accuracy = 90.7563 %
	b. Weighted Avg. F-measure = 0.963	b. Weighted Avg. F-measure = 0.908
	c. Weighted Avg. ROC Area = 0.924	c. Weighted Avg. ROC Area = 0.903
	d. FP rate = 11/88 = 12.5 %	d. FP rate = 11/88 = 12.5%
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 attributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 89.4958 %	a. Accuracy = 89.4958 %
	b. Weighted Avg. F-measure = 0.896	b. Weighted Avg. F-measure = 0.896
	c. Weighted Avg. ROC Area = 0.898	c. Weighted Avg. ROC Area = 0.899
	d. FP rate = 10/88 = 11.36%	d. FP rate = 10/88 = 11.36%

Classifier Performance

ALGORITHMS	Full Sample Results : 673 instances , 18 attributes	Downsized Sample Results : 238 instances , 18 attributes
3. Logistic Regression	a. Accuracy = 96.8796 %	a. Accuracy = 92.0168 %
	b. Weighted Avg. F-measure = 0.969	b. Weighted Avg. F-measure = 0.921
	c. Weighted Avg. ROC Area = 0.977	c. Weighted Avg. ROC Area = 0.921
	d. FP rate = 11/88 = 12.5%	d. FP rate = 7/88 = 7.95%
	Downsized Sample + Cost Sensitive Classifier : 238 instances , 18 attributes , cost FP : cost FN = 3:1	**Downsized Sample + Cost Sensitive Classifier + Reduced attributes : 238 instances , 9 attributes , cost FP : cost FN = 3:1**
	a. Accuracy = 92.0168 %	a. Accuracy = 91.5966 %
	b. Weighted Avg. F-measure = 0.921	b. Weighted Avg. F-measure = 0.916
	c. Weighted Avg. ROC Area = 0.837	c. Weighted Avg. ROC Area = 0.939
	d. FP rate = 7/88 = 7.95%	d. FP rate = 7/88 = 7.95%

ROC curves for "Charter" class



Key Observations & Suggestions

1. Study cost vs benefits of small class sizes
2. Compare Charter vs Non Charter districts on different parameters using historical data
3. Identify outlier districts and aim for a more equitable distribution of resources across all Districts in the State
4. Compare and contrast different performance measures with State's that rank high on positive educational outcomes but spend considerably less than NJ.

Work Log

Activities	Amy Le	Gilat Mandelbaum	Sakina Presswala
Literature Search	2/21/18: 3	2/21/18: 3	2/21/18: 10 am-1.30pm: 3.5
Project Proposal	2/22/18: 3 2/23/18: 7	2/21/18: 2 2/22/18: 7.5	2/24/18:11pm-1 pm: 2
Proposal ppt	2/25/18: 2	2/25/18: 2	2/15/18 : 11am-5pm:6
Data Cleaning & Preprocessing	3/10-4/16: 25.5	4/9/18-4/19/18: 33.5	3/1/18-4/1/18:20
Data Analysis	4/17-4/25: 20.5	4/19/18-4/25/16: 26.5	4/1/18 - 4/5/18: 5
Poster Presentation	4/18-4/19: 5	4/17/18-4/19/18: 4.5	4/10/18: 2
Final ppt	4/20-4/26: 3	4/19/18-4/25/18: 3	4/20/18:2
Miscellaneous	N/A	N/A	2/1/18-4/25/18 : 10

Questions / Comments ?