Predicting Graduation Rates and What Impacts Them

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the course of

Masters

in

Department of Computer science

by

Nicholas Yahr, (712107197)

Premkumar Vankudoth, (104766296)

Likhita Pampana ,(464105177)

Sri Rasagna Yadlapally ,(692228773)

WESTERN MICHIGAN
UNIVERSITY

to

DEPARTMENT OF COMPUTER SCIENCE

WESTERN MICHIGAN UNIVERSITY

KALAMAZOO - 49006, USA

April 2020

DECLARATION

We, **Nicholas Yahr, (712107197),Premkumar Vankudoth, (104766296)**, **Likhita Pampana ,(464105177),Sri Rasagna Yadlapally ,(692228773)** hereby declare that, this report entitled **"Predicting Graduation Rates and What Impacts Them"** submitted to Western Michigan university towards partial requirement of **Master of Data science/ Computer Science** in **Department of computer science** is an original work carried out by us under the supervision of **Dr.Alvis Fong** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kalamazoo 49006
April 2020

ABSTRACT

The main aim of the project was to Predict and interpret what affects the graduation rates of the schools,Our data set is a collection of performance data collected from Massachusetts school districts. Our project seeks to predict how the various features impact each high school's graduation rates and if we can begin to understand what features have the most impact. This could then be used by school districts to see the warning signs of possible graduation rate drops so that it can be addressed before it becomes a problem and students have the greatest chance at success. Applying various Machine learning techniques to make prediction accuracy better.

Contents

List of Figures **v**

List of Tables **vi**

1 Introduction **1**

1.1 Data Set Description 2

2 Exploratory Data Analysis and Visualization **4**

2.1 Data Visualization 4

2.2 Data Pre-processing Techniques Used 8

2.2.1 Processing Techniques Used 8

3 Algorithms Used and Workflow **11**

3.1 Methods Used 12

3.2 Final Outcome 15

4 Results and observations **16**

5 Conclusion **20**

Bibliography **21**

List of Figures

1.1 Dataset Sample image 3

2.1 Barplot for the data against type of school 5

2.2 Distribution of students 6

2.3 Linearity for expenditure vs Teacher salary 7

2.4 Distribution of Y column ”% Graduated” 7

3.1 Features Extracted From The Pre processing 15

List of Tables

4.1 Features and Confidence Intervals 17

Chapter 1

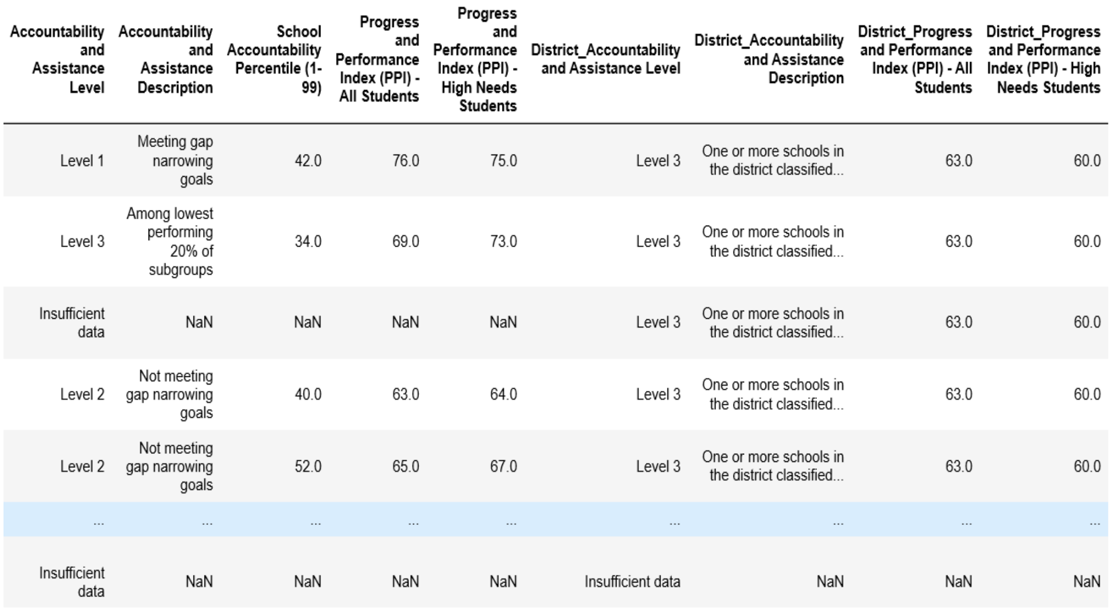
Introduction

The data we used for this project is Massachusetts public schools data from kaggle.Our data set is a collection of performance data collected from Massachussets school districts. As from the name it says schools data so the main aim of the data is to predict the graduation rate basing on the different factors effect the grad rate of the colleges.The importance that is placed on graduation rates as a measure of the success of institutions of higher education warrant the ongoing research into understanding the determinants of these educational outcomes. This project mainly concentrates on what factors mainly impact on the graduation rates. Feature selection methods were used to extract the meaningful features which also reduces if there is any over fitting of data. Main tools used to build this project are Jupyter notebook, python programming language, latex for report and excel for data analysis.

1.1 Data Set Description

This data set compiles data from the following Massachusetts Department of Education reports,the data set has 1861 schools, 302 different features including, **y** variable graduation rate.The data has been collected from various sources like, Enrollment by Grade Enrollment by Selected Population, Enrollment by Race/Gender, Class Size by Gender and Selected Populations,Teacher Salaries,Per Pupil Expenditure,Graduation Rates,Graduates Attending Higher Ed,Advanced Placement Participation,Advanced Placement Performance,SAT Performance,MCAS Achievement Results, Accountability Report.Reports from all these sources form to Massachusetts Public Schools Data.

As we mentioned data set has 302 features out of which some of them have the significant effect towards the dependent variable, the data has the features like,School Code ,School Name,School Type,Function,Contact Name,Address-1,Address-2,Town,State Zip,Phone,Fax,Grade,District Name, District Code, PK_Enrollment . . . and data set is mixture various kinds of data types like numerical, categorical, factors and characters. The data is collected during the year 2016-2017.



Accountability and Assistance Level	Accountability and Assistance Description	School Accountability Percentile (1-99)	Progress and Performance Index (PPI) - All Students	Progress and Performance Index (PPI) - High Needs Students	District_Accountability and Assistance Level	District_Accountability and Assistance Description	District_Progress and Performance Index (PPI) - All Students	District_Progress and Performance Index (PPI) - High Needs Students
Level 1	Meeting gap narrowing goals	42.0	76.0	75.0	Level 3	One or more schools in the district classified...	63.0	60.0
Level 3	Among lowest performing 20% of subgroups	34.0	69.0	73.0	Level 3	One or more schools in the district classified...	63.0	60.0
Insufficient data	NaN	NaN	NaN	NaN	Level 3	One or more schools in the district classified...	63.0	60.0
Level 2	Not meeting gap narrowing goals	40.0	63.0	64.0	Level 3	One or more schools in the district classified...	63.0	60.0
Level 2	Not meeting gap narrowing goals	52.0	65.0	67.0	Level 3	One or more schools in the district classified...	63.0	60.0
...	...	...	...	...	...	...	...	...
Insufficient data	NaN	NaN	NaN	NaN	Insufficient data	NaN	NaN	NaN

Figure 1.1: Dataset Sample image

As the Y variable of the data is numerical so the regression analysis would be the first thought, in order make things bit understandable the Y variable has been encoded in to the categorical feature, into ”Below Average” and ”Above Average” groups based on the mean of ”% Graduated” (73.4%).

Chapter 2

Exploratory Data Analysis and Visualization

Exploratory Data Analysis refers to the critical process of performing starting investigations on data so as to discover patterns.Exploratory knowledge Analysis (EDA) is that the start in your knowledge analysis method Here,you create sense of the information you've got then understand what queries you wish to raise and the way to border them, still as however best to manipulate your obtainable knowledge sources to urge the answers you wish. For instance checking the missing values from the dataset or Discovering the patterns from the dataset, listing anomalies and outliers.

2.1 Data Visualization

For every data set before preceding to further works the first thing we carry out is Exploratory data analysis by this we try to under stand the data and can work efficiently with the data set.

Data visualization techniques can be used to compare between the feature variables and target attribute.We need data visualization because a visible summary of data makes it easier to spot patterns and trends than ransacking through thousands of rows on a spreadsheet. It's the way the human brain works. Since the aim of information analysis is to achieve insights, data is far more valuable when it's visualized. whether or not an information analyst can pull insights from data without visualization, it'll be tougher to speak the meaning without visualization. Charts and graphs make communicating data findings easier whether or not you'll be able to identify the patterns without them.

We got some interesting plots for the better understanding of the data.

Plots to Understand Data Features

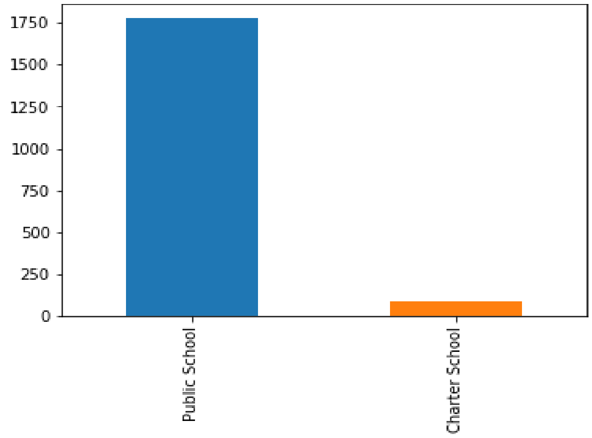


Figure 2.1: Barplot for the data against type of school

The above bar plot describes the distribution of type of school against the whole data set.
Interesting insight is that percentage of High needs and economic disadvantaged show similar distribution. can see in the below plot.

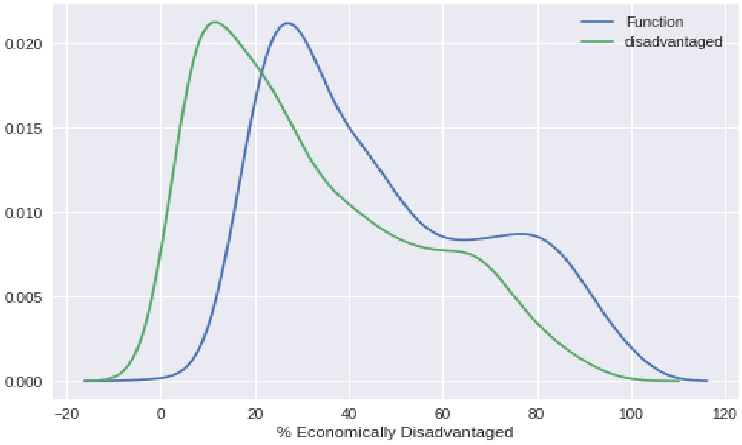


Figure 2.2: Distribution of students

Expenditures per pupil are highly correlated with teacher salary, but not correlated at all with class size. Seems like teacher salary affects the public budget a lot more than class size does.

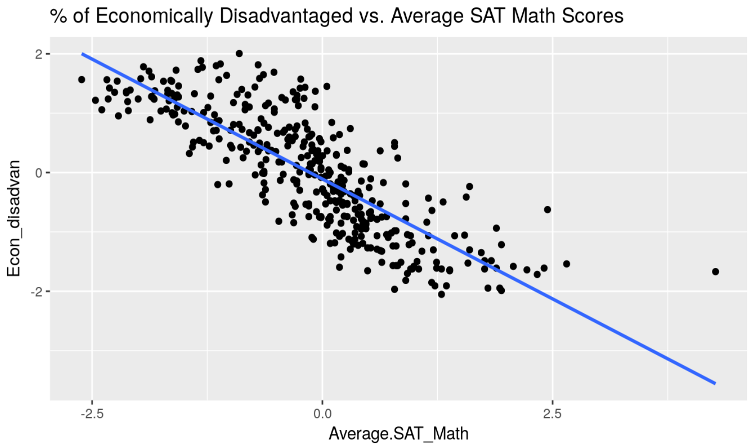


Figure 2.3: Linearity for expenditure vs Teacher salary

Now lets see how our Y variable is distributed, the below plot shows the exact how our Y is distributed and skewed.

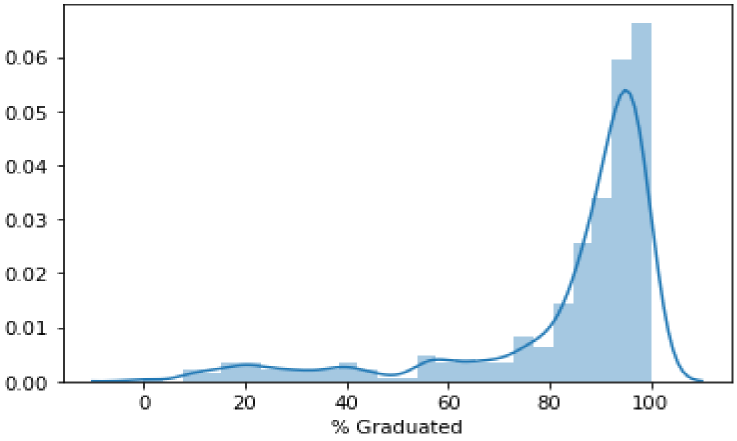


Figure 2.4: Distribution of Y column ”% Graduated”

2.2 Data Pre-processing Techniques Used

Before applying a algorithm to the data set its better to understand the patterns of the data. In order to make efficient analysis data cleaning is the necessary step we need to follow before concluding to any results.
The key things we need to concentrate while cleaning data are :

- check weather there are any Na's in the data .
- Find the variables which required hot encoding. like conversion of character variables to numerical
- Filling or removing the missing values
- Removing outliers

These are some of the key terms we need to concentrate while data preprocessing.

2.2.1 Processing Techniques Used

The main preparatory techniques used in the project are

Encoding

One-hot encoding appropriate features like **school type** which is a factor variable with 2 levels , which is either "Public School" or "Charter School",The other feature is function type of data with level as principal. These two variables are of type character. using encoding technique these are transformed to numerical values so that further can be used for data analysis.

Features like Accountability and Assistance Level, "District_Accountability and Assistance Level has the values like LEVEL1, LEVEL2, its better we turn them to some numerical.

Float Columns

There are some numerical variables which has values like 1,000 there are deli-meters in the numericals so removing them makes the data much more better.

Splitting the required columns

Splitting up the grades into separate features - They start off as strings looking like "K,1,2,3" indicating the grades that this school serves, and I turn them into features "Grade K", "Grade 1", "Grade 2", "Grade 3", etc. with a value 0 or 1 indicating whether or not that school serves that grade level.

The other data cleaning done for this data set are

- Replacing NA's with 0's where appropriate (such as in the "AP_Test Takers" feature)
- Removing a few features that make no sense to the data (such as the principal's name) or that are directly related to other features (remove "High School Graduates ()" because we already have "

We calculate the percentage of students who a graduating high school and drop any schools that have missing values for this graduation rate. Then we

split the data into two classes based on this graduation rate, ”Below Average” and ”Above Average” based on the mean of the graduation rates.
Removing unnecessary columns like School name, district name and other character columns.Now out of , 302 columns after all the cleaning work we left with 274 features,Next we removed the schools that did not report graduation rates, as those schools are invalid for the exploration we are performing. Now we have 376 schools left.Now we encode the ”% Graduated” column into ”Below Average” and ”Above Average” groups based on the mean of ”% Graduated” (73.4Now we are ready to build the model.

Chapter 3

Algorithms Used and Workflow

In order to build a good model we must know which all features are the best. There are many ways we can find significance level of variables to the dependent variables, we still have 274 features which is quite high for the data with 1500 rows.

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.Feature selection and Data cleaning should be the first and most important step of your model designing. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested.

3.1 Methods Used

Iterative Imputer

When the decision on how to handle missing values in your data, there are three possible options: remove the observations with the missing data, leave the missing values in place or impute values the practical choices were the mean, median or mode. This would be sufficient if there are few missing values and/or the variance of the data is not significant. And the one more option we have is to Iterative imputer pacakage from scikit learn.The IterativeImputer package allows the flexibility to choose a pre-loaded sci-kit learn model to iterate through the data to impute missing values. Replaces any remaining NA's with values based on repeated fits of the rest of the data if we miss any during the data cleaning processes.

Robust Scaler

Scale generally means to change the range of the values. The shape of the distribution doesn't change. Think about how a scale model of a building has the same proportions as the original, just smaller.Many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and/or close to normally distributed.

- Standardize generally means changing the values so that the distribution standard deviation from the mean equals one
- Normalize can be used to mean. Make the data much more convenient for the model.

Robust Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

For the data we used robust method so that it Scales down all the values to a much smaller range maintaining variance (this keeps all of the information that each value gives to the fit while making the classifier more accurate. It's also robust to outliers!)

Variance Threshold Selector

VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.
This part cuts out any features that are extremely similar to each other, drastically reducing the number of features in the model while maintaining model accuracy.

RFECV/Model Selector

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model's coef_ or feature_importances_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

First we used variance based feature selection method,This second selector does a more fine-tuned selection based upon the classification model that you are using. This cuts out the features that are unused by the model making this perform the task of Feature Selection algorithm such as LASSO

SGD Classifier with Logisitic Loss

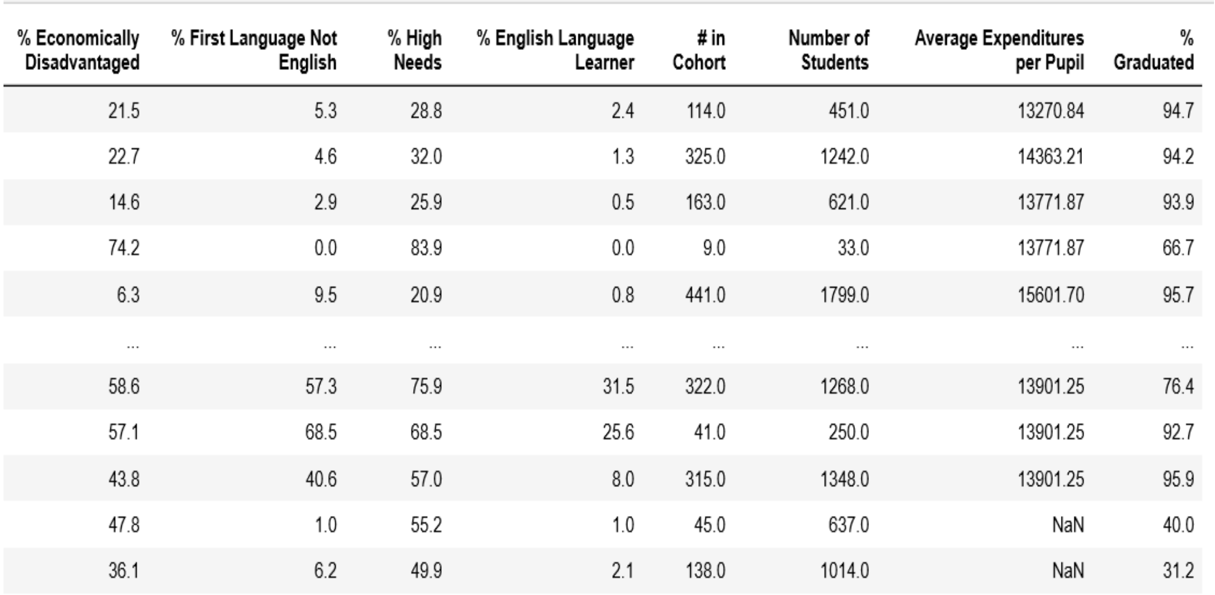
SGD Classifier implements regularised linear models with Stochastic Gradient Descent.Stochastic gradient descent considers only 1 random point while changing weights unlike gradient descent which considers the whole training data. As such stochastic gradient descent is much faster than gradient descent when dealing with large data sets. Logistic Regression by default uses Gradient Descent and as such it would be better to use SGD Classifier on larger data sets.By default, the SGD Classifier does not perform as well as the Logistic Regression. It requires some hyper parameter tuning to be done.SGDClassifier for general classification problems (like logistic regression) specifying a loss and penalty.

Actually fit the final classifier that gives us values. This classifier is a logistic classifier that implements feature reduction/selection so that we can have a very simple, very accurate model.

3.2 Final Outcome

This fit was performed 100 times in a bootstrap (with random samplings for the training and test sets) in order to give us accurate predictions of scores, probabilities, and coefficients as well as confidence intervals for all of these. Doing this bootstrap also helped to select the features that were truly important so that a simpler model could be built. Out of the 274 features it turned out that only 7 of them were truly predictive of the graduation rates! (The final model fits only these 7 and therefore removes the need for the two selectors but they are included above because they were used to find the 7 in the first place).

The below figure 3.1 shows the sample of the data used for modelling after all the Pre-processing techniques were applied.



% Economically Disadvantaged	% First Language Not English	% High Needs	% English Language Learner	# in Cohort	Number of Students	Average Expenditures per Pupil	% Graduated
21.5	5.3	28.8	2.4	114.0	451.0	13270.84	94.7
22.7	4.6	32.0	1.3	325.0	1242.0	14363.21	94.2
14.6	2.9	25.9	0.5	163.0	621.0	13771.87	93.9
74.2	0.0	83.9	0.0	9.0	33.0	13771.87	66.7
6.3	9.5	20.9	0.8	441.0	1799.0	15601.70	95.7
...	...	...	...	...	...	...	...
58.6	57.3	75.9	31.5	322.0	1268.0	13901.25	76.4
57.1	68.5	68.5	25.6	41.0	250.0	13901.25	92.7
43.8	40.6	57.0	8.0	315.0	1348.0	13901.25	95.9
47.8	1.0	55.2	1.0	45.0	637.0	NaN	40.0
36.1	6.2	49.9	2.1	138.0	1014.0	NaN	31.2

Figure 3.1: Features Extracted From The Pre processing

Chapter 4

Results and observations

The generated model shows all the qualifications of a good fit to the data. The probabilities of each classification was quite high (estimated at 100% with a confidence interval (CI) of 66.3%-100% for High Graduation Rate and 100% with a CI of 61.4-100% for Low Graduation Rate). Accuracy scores were also consistently high across three different scoring techniques. Basic accuracy scoring was estimated at 88.2% with CI 82.9-96.1%, the balanced accuracy score (correcting for a smaller selection of Low Graduation Rate schools) was 88.6% with CI 77.9-96%, and a Jaccard score of 84.6% and CI 77.6-94.2%. These scores all show that the model consistently gave strong results that can be trusted.

Due to the nature of logistic models the coefficients output cannot be directly interpreted, however they are still useful in their relative magnitude and direction. For each fit of the data larger coefficients have more impact on the output than lower coefficients, so in order to compare coefficients between each bootstrap they were scaled using sci-kit learn's Robust Scaler (without any centering, which would have changed directions on the resulting values). Direction lets us see which class the feature predicts, specifically a positive coefficient indicates that increases in this feature will lead to a high graduation rate, and a negative coefficient indicates that increases in this feature will lead to a lower graduation rate. The seven features and their confidence intervals are as follows in the form "Feature: Estimate (Lower Confidence Interval, Upper Confidence Interval)":

Feature	Estimate	Confidence_Interval (Lower_CI,Upper_CI)
% High Needs	-1.0180	(-2.5972, -0.3356)
% First Language Not English	0.7498	(0.4963, 1.3443)
% English Language Learner	-0.6345	(-0.9928, -0.2760)
Number of Students	0.5945	(0.1137, 1.1787)
# in Cohort	-0.5287	(-0.8648, -0.0845)
% Economically Disadvantaged	-0.4755	(-1.2229, 0.0252)
Average Expenditures per Pupil	0.1955	(0.0293, 0.4541)

Table 4.1: Features and Confidence Intervals

As these numbers show schools high needs students are at greatest risk of having lower graduation rates, as well as those schools that serve the economically disadvantaged, though the economic status is not as strong of a prediction as some of the other features as indicated by its confidence interval. Additionally increasing the size of student cohorts decreases the graduation rate as students will not get the one-on-one attention that they need. Schools with larger student bodies and more money to spend per student tend to have better graduation rates as well as those schools who serve students whose first language is not English.

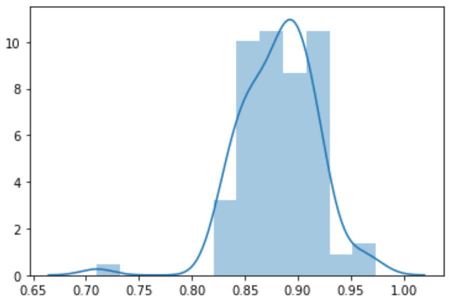
Accuracy Scores

- Accuracy Score: 0.882 (0.829, 0.961)
- Balanced Accuracy Score (accounts for different sized groups: 0.886 (0.779, 0.960)
- Jaccard Score: 0.846 (0.776, 0.942)

The Probabilities (confidence of prediction) of the groups are as follows:

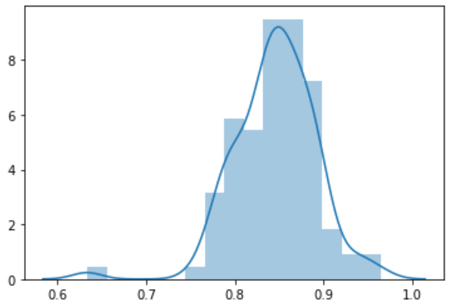
- High Graduation Rate: 1.000 (0.663, 1.000)
- Low Graduation Rate: 1.000 (0.614, 1.000)

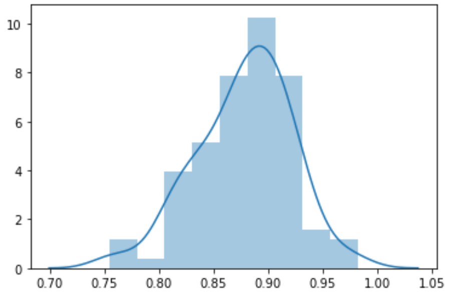
These plots show the above results Graphically



.

Accuray Scores Distrubution





Jaccard Score

Balanced Accuracy Score

Chapter 5

Conclusion

We can confidently say that the graduation rates of schools in the Manhattan school districts have been accurately modeled using the process described above. The logistic modeling has allowed us to interpret the findings of our model and isolate a small number of features that indicate schools that should get additional support in order to increase their graduation rates, including those with high needs students and English language learners. Further investigations into these particular types of schools should be done in order to discover the exact needs to be addressed within these schools.

Bibliography

[1] Gintare Karolina Dziugaite and Daniel M Roy. Entropy-sgd optimizes the prior of a pac-bayes bound: Generalization properties of entropy-sgd and data-dependent priors. *arXiv preprint arXiv:1712.09376*, 2017.

[2] Donald W Irvine. Multiple prediction of college graduation from preadmission data. *The Journal of Experimental Education*, 35(1):84–89, 1966.

[3] Yushan Liu and Steven D Brown. Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Intelligent Laboratory Systems*, 120:106–115, 2013.

[4] Terence J Tracey and William E Sedlacek. Prediction of college graduation using noncognitive variables by race. *Measurement and Evaluation in Counseling and Development*, 19(4):177–184, 1987.