Capstone Project: Predicting Educational Achievement Gaps

Orlando Trejo, Ph.D.
February 23, 2017

Define: Domain Background

The educational achievement and income gaps among white, black, and Hispanic students across our nation increasingly jeopardizes the future welfare of our society and economy. The education achievement gap due to income inequality costs the United States about $3.5 billion dollars per day in gross domestic product (GDP) growth [2]. A 2016 study done through Stanford's Center for Educational Analysis (CEPA) shows the persistent existence of the achievement gap and that parent income gap remains a strong predictor of students' academic success [1]. This study also points out that in many cases white students outperform black and Hispanic students despite having the same socioeconomic background. Therefore, it is important to identify additional factors to parent's socioeconomic status and student race that are predictive of the achievement gap.

Additional factors, like parent involvement, have been shown to have strong correlation with student's academic success [5]. However, predictors of this nature are difficult to quantify and measure so large datasets are not readily available for analysis. Fortunately, CEPA's datasets contains variables in addition to parent's income level that may serve as a starting to point to measure the relative importance of variables in predicting the achievement gap. The scope of this project is to use classification algorithms on CEPA's datasets to determine what are key variables, also referred to as features and covariates, which are indicative of the achievement gap for black and Hispanic students. The achievement gap is defined as the grade level difference in school performance, as measured by exams, between white and black students and white and Hispanic students.

Define: Problem Statement

Before closing the achievement gap, insight gaps concerning the relevance of variables in addition to income inequality need to be addressed first. Closing such gaps poses a challenging task that can be strategically addressed through high-volume data analytics enabled by machine learning strategies. To start closing insight gaps, this project takes a two-step approach: (1) separately determine the features with highest importance when predicting the achievement gap for black and Hispanic students and (2) using these features to compare the performance of different classification models in order to assess the validity of the features selected. The features are the covariates, while the labels are the achievement gaps provided by CEPA [3]. As described below, the achievement gaps from CEPA are binned in labels for classification purposes. Furthermore, by understanding the benefits and shortcomings of CEPA's dataset on student achievement gaps using machine learning (ML) algorithms, it may be possible to suggest new types of data to be collected in future studies.

Define: Datasets

The datasets for this project can be obtained through Stanford's Center for Education Policy and Analysis (CEPA) website [3]. The data on achievement gaps will be used to obtain the labels for the ML algorithms. This dataset can be decomposed along three

main dimensions: race, subject, and grade level. Data on the covariates gathered by CEPA will be used to obtain features for the models. The dataset on covariates contains 169 variables. However, the covariates dataset is not entirely decomposed by the same three dimensions as the achievement gap labels. Therefore, some of the features will have to be reused when analyzing by race. Data is available for both the features and labels for the following academic years: 2008-2009, 2009-2010, 2010-2011, 2011-2012, and 2012-2013.

Given the fact that the achievement gap varies significantly along socioeconomic and racial divides, it will be important to have the flexibility to measure and compare the importance of the top achievement gap predictors from the covariate dataset. Predictors for the achievement gap of white students may consist of parents' education and school district funding. However, for black students, predictors may consist more of racial diversity in the school district and number of members in the household. Similarly, for Hispanic students, the top predictive covariates may differ. Overall, the covariates and their respective importance to the prediction of achievement gaps may change as a function of race.

Define: Solution Statement

Given the high degree of detail in the datasets, it will be necessary to explore different types of machine leaning algorithms. The solution is to find the ML model that best predicts the achievement gap by race. Moreover, as described earlier, it is also important to quantify the top predictors along different dimensions in order to understand what are the contributing covariates for a given achievement gap breakdown of race.

Define: Benchmark Model

Since the performances of classification algorithms are going to be compared, the dummy classifier is the benchmark model for this work. This classifier is commonly used as baseline for actual classification algorithms.

Define: Evaluation Metrics

The evaluation metric is to compare the weighted F1 score of using three types of classifiers to predict the achievement gap. In our case, the weighted F1 score is a good choice because it takes into account the average F1 score for each label. The F1 score is applicable for our purpose because the goal is to simultaneously maximize the precision and recall of the achievement gap, which balances the cost of false positives and false negatives. Since the F1 score is the harmonic mean of precision and recall, then it is a good choice for evaluating and comparing the performance of the classification models. Thus, the weighted F1 score helps determine the validity of the selected covariates.

Models that have more accurate predictions on the test data will be deemed as better performance models. Moreover, models that sustain the weighted F1 score across grade levels, subjects, and academic years for a given racial group will be preferred over models that do well only in specific situations. However, given the nature of how cultural variations may impact each racial group differently, it is not expected for models to be

general across racial groups. Moreover, different covariates are used for different racial groups in order to reduce dimensionality as explained below.

Analysis: Data Exploration

For this project, the features for the machine learning (ML) models are referred to as covariates and the labels as achievement gaps. The goal is to optimize the prediction accuracy of the achievement gap using the covariates data. As the covariates dataset shows, there are multiple years under consideration (2009-2013) and more than a 150 columns containing data that may be used in the ML models. Therefore, it is necessary to go through the column names and descriptions in the codebook provided by CEPA in order to determine what columns are relevant. The codebook is contained in the zip file for this project.

By going through the names of the covariate column names, it is clear that some columns are transformation of other columns and some covariates are race specific. Therefore, a subset of covariate column names is identified for the ML models and stored in the "covariates_subset" CSV file. Ten covariates for black and Hispanic students are shown in Table 1 as an example. This manual selection reduces the number of features used for the ML to 41 for both of the race groups. More detail on the meaning of the covariates is contained in the "codebook_covariates_v1_1" Excel file. In the interest of space, the covariates in Table 1 can be described as follows:

0. ‘leaid’ is the identification number for a school district
1. ‘year’ is the academic year in which the covariate data corresponds to
2. ‘urban’ denotes if a school district is in an urban area (0: no, 1:yes)
3. ‘perblk’ and ‘perhsp’ is the percentage of black and Hispanic students in a school district, respectively
4. ‘perell’ is the percentage of English Language Learners in a school district
5. ‘perspeced’ is the percentage of Special Education students in a school district
6. ‘perfrl’ is the percentage of students with free lunch in a school district
7. ‘stutch_blk’ and ‘stutch_hsp’ is the average student to teacher ratio for black and Hispanic students, respectively
8. ‘percharter_blk’ and ‘percharter_hsp’ is the percent of public school black and Hispanic students in Charter schools, respectively
9. ‘hswhtblk’ and ‘hswhthsp’ is the average of the ratio of each schools’ racial diversity to the district-wide racial diversity for black and Hispanic students, respectively.

Table 1. An example of the features manually selected for the black and Hispanic student populations. Some of the features are common while other features are race specific.

	cvrts_blk	cvrts_hsp
0	leaid	leaid
1	year	year
2	urban	urban
3	perblk	perhsp
4	perell	perell
5	perspeced	perspeced
6	perfrl	perfrl
7	stutch_blk	stutch_hsp
8	percharter_blk	percharter_hsp
9	hswhtblk	hswhthsp

A scatter plot matrix is used to visualize the distribution of multiple randomly chose covariates at the same time. As shown in Figures 1 and 2, the distributions for the covariates describing the black and Hispanic students have similar distributions. Given the shape of the distributions of many of the covariates resemble normal distributions, than not many feature transformations are necessary.

Now that the covariates dataset has been reduced to more meaningful features for black and Hispanic students, the next step is to go through the achievement gaps dataset. The achievement gap dataset captures gaps between the black and white populations, as well as gaps between the white and Hispanic students. Essential common columns in the covariates and gaps dataset are the school district identification number (leaid) and the academic year. These common columns are used for matching covariates entries to the achievement gaps.

Grade level, subject and race also split the achievement gaps dataset. This means that the covariates data entries need to be used multiple times during the matching between the covariates and gaps. Furthermore, the gaps dataset is restructured in order to put all achievement gaps, subjects, and races in individual columns.

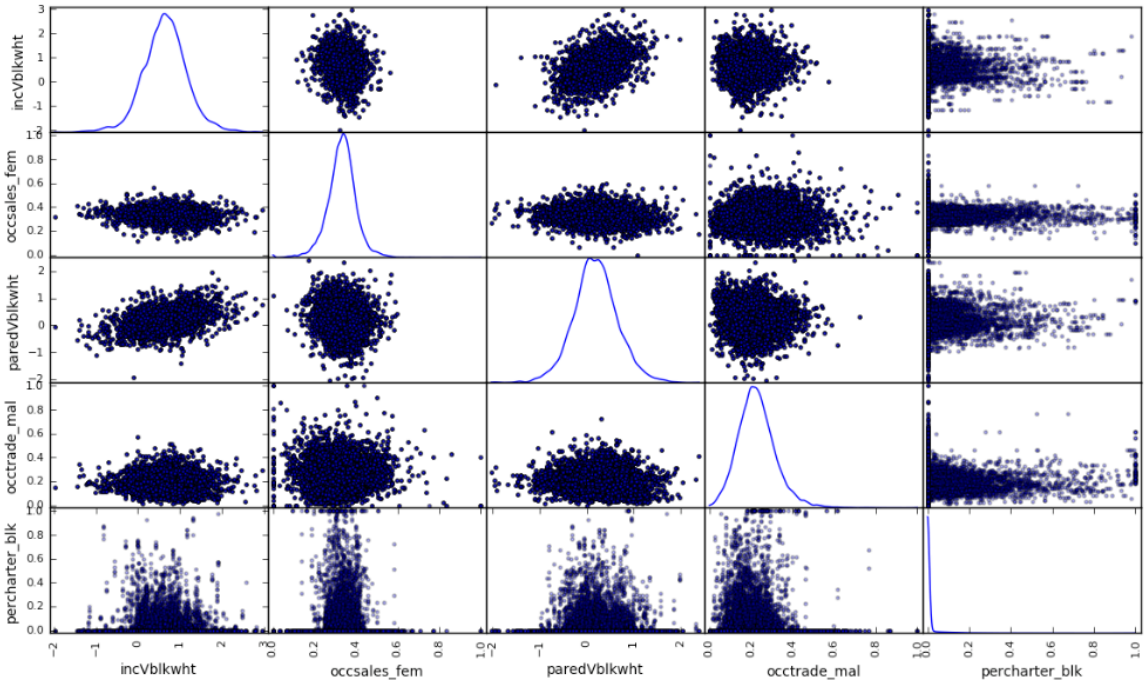

Figure 1. Scatter plot matrix of randomly selected covariates for the black student population.

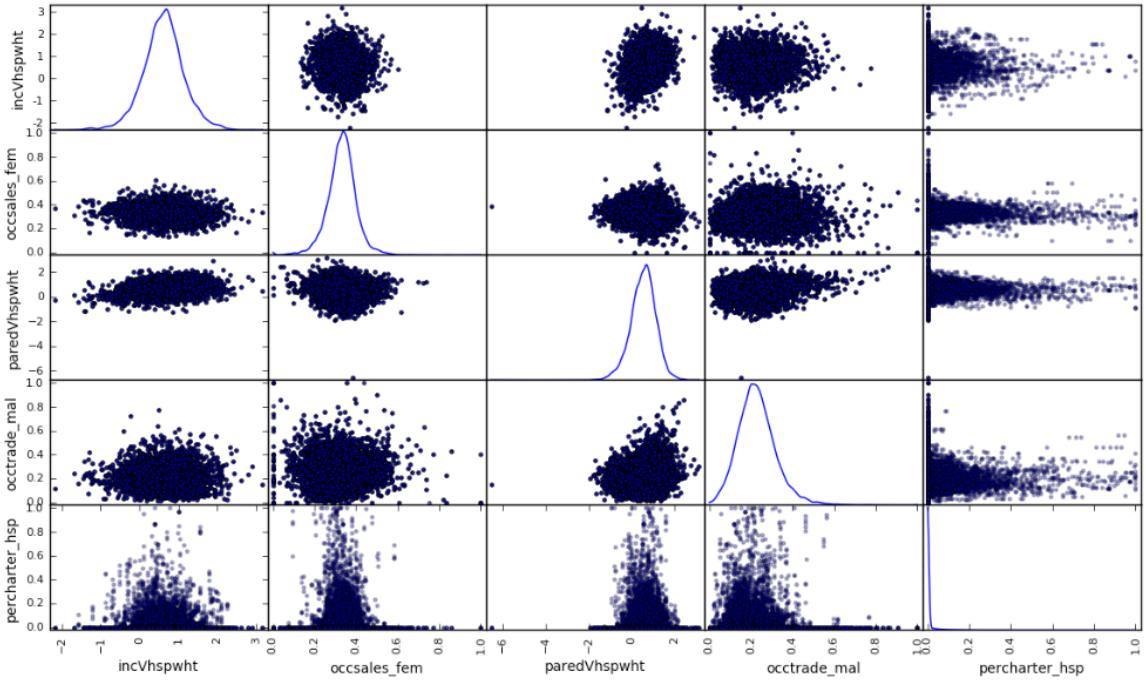

Figure 2. Scatter plot matrix of randomly selected covariates for the Hispanic student population.

Table 2. Table showing the initial ten rows for the re-structured achievement gaps dataset.

	leaid	year	grade	subject	race	gap	gapse
0	100005.0	2009.0	3.0	ela	hsp	0.495028	0.144482
1	100005.0	2009.0	3.0	math	hsp	0.359011	0.142677
2	100005.0	2009.0	4.0	ela	hsp	0.645853	0.131344
3	100005.0	2009.0	4.0	math	hsp	0.385385	0.128722
4	100005.0	2009.0	5.0	ela	hsp	0.877928	0.141010
5	100005.0	2009.0	5.0	math	hsp	0.537650	0.144171
6	100005.0	2009.0	6.0	ela	hsp	0.662599	0.165949
7	100005.0	2009.0	6.0	math	hsp	0.457464	0.158139
8	100005.0	2009.0	7.0	ela	hsp	1.062303	0.158241
9	100005.0	2009.0	7.0	math	hsp	0.849210	0.162989

From the re-structured achievement gaps dataset the statistics for each race group is calculated. There are 120460 achievement gap entries for black students and 140258 entries for the Hispanic students. The mean for each group differ by approximately a tenth of a grade level in the achievement gap and have a similar standard deviation of approximately three tenths of a grade level. Both black and Hispanic students are on average half a grade level behind their white peers. The application of t-test also determines that the difference between black and Hispanic is statistically significant. This means that on average black students are expected to underperform white and Hispanic students. Other studies have also seen this trend of black student trailing behind in academic achievement for generations [6]. Figure 3 captures the probability histogram and normal distribution fits for black and Hispanic students.

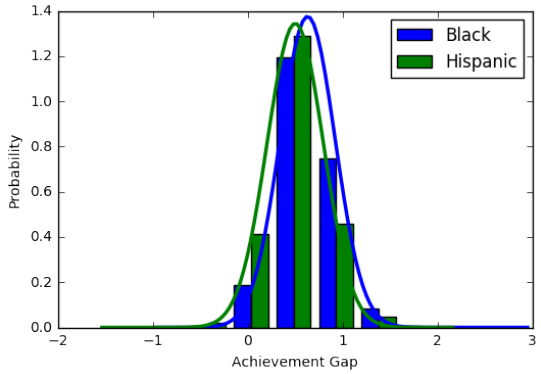

Figure 3. Histograms and normal distribution fits for the achievement gaps for black (green) and Hispanic (blue) students.

Analysis: Algorithms and Techniques

Until now, the main algorithms and techniques used have been scatter plots, histograms, normal distribution fitting, and t-test. The use of this first set of tools is appropriate because the data's characteristics do not exhibit any noticeable abnormalities. Therefore, the next step is to implement another set of ML models to gain more accuracy in prediction. This next set of models consists of:

- Dummy Classification (for baseline/benchmark)
- Decision Tree Classification
- Logistic Regression
- Support Vector Machines (SVMs)

The Dummy Classifier is a commonly used baseline model for ML classification algorithms. In this case, the default parameters are kept for the Dummy Classification model. The parameter strategy has a default value for 'stratified,' which means that the model generates predicts based on the distribution of the labels. This is an appropriate strategy given the similarity of the normal distribution of the achievement gaps (Fig. 3) This classifier is not tuned in this project.

Decision trees are an excellent choice for this problem because they require little data preparation and can lead to interpretable results. Moreover, they provide information on feature importances in the weighted F1 score that can be used for feature reduction. Therefore, the top features after applying decision tress on these dataset will the only ones that remain for the logistic regression and SVMs models. The main thing to look out for when using decision trees is the tendency of over fitting the data. This can be addressed by tuning the maximum depth of the tree, which is done in the refinement section of this project. Another parameter of interest is the criterion, which is can be determined by the Gini impurity (I_G) or entropy (I_H).

The Gini impurity is described by the following equation:

$$I_G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

Where, p(i|t) is the fraction of samples that belong to a class c for a particular note t.

The entropy is described by the following equation:

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2 p(i|t)$$

Logistic regression is another common ML model used on datasets similar to the ones being used in this project. It works efficiently with large number of observations and can also help provide probabilities for interpretation. Also, logistic regression is not sensitive to small noises, which can be expected in the achievement gaps data as testing procedure vary across school districts. Furthermore, like decision tress, logistic regression can

handle well multi-collinearity among feature. This is especially important since the key feature of the income gap is likely to be highly correlated with other features contained in the covariates dataset. One way to tune the logistic regression model is to choose the optimal regularization algorithm between l1-regularization and l2-regularization. This is part of the model tuning in the refinement section. The inverse of the regularization parameter also needs to be tuned, C.

The regularized cost function of logistic regression can be described by the following equation:

$$J(\boldsymbol{w}) = C\left[\sum_{i=1}^{n}\left(-\log\left(\emptyset\left(z^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\right)\left(-\log\left(1 - \emptyset(z)\right)\right)\right] + \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$$

Where, **w** is a vector containing the weights, z is the linear combination of the features, and y are the labels.

Finally, SVMs are also a promising ML model for this project because their approach is based on decision boundaries, which makes them less sensitive to missing data and non-linear feature interactions. Therefore, SVMs provide a complementary approach to decision tress. The main drawbacks of SVMs are that the results are not intuitive and they do not handle well large number of observation, which means results from SVMs may not yield high interpretability and given our large number observation may take a long time to train and tune. For SVMs the main parameters to tune are the kernel, cost, and gamma. Tuning of SVMs is explored in the refinement section.

For SVMs, the following expression needs to be minimized:

$$\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\left(\sum_i \zeta^{(i)}\right)$$

Where, ζ is the slack variable which allows for convergence in the presence of misclassifications.

Methodology: Data Preprocessing

For the implementation of the decision tree classifiers, it is necessary to bin the achievement gaps into classification labels. To do so and without losing much achievement gap resolution, the gaps are binned as following:

- Gaps below -1.25 are labeled -3
- Gaps greater or equal to -1.25 and below -0.75 are labeled -2
- Gaps greater or equal to -0.75 and below -0.25 are labeled -1
- Gaps greater or equal to -0.25 and below 0.25 are labeled 0
- Gaps greater or equal to 0.25 and below 0.75 are labeled 1
- Gaps greater or equal to 0.75 and below 1.25 are labeled 2
- Gaps greater or equal to 1.25 and below 1.75 are labeled 3
- Gaps great or equal to 1.75 are labeled 4

In essence, this labeling scheme bins student into groups that span half a grade level, where black and Hispanic students are assumed to be at grade levels as their white peers in the group labeled “0.” Since the average achievement gap of black and Hispanic students compared with white students is at least 0.5 of a grade level, then the binning needs to maintain students with an achievement gap of 0.25 of a grade level and above as separate from the grade level group labeled “0.”

One last step of data preprocessing is to do feature scaling on all the columns of the covariates datasets. Feature scaling entails taking the difference between the maximum value per column and each element, followed by dividing the difference by the column range. This results in columns with values between 0 and 1.

Methodology: Implementation

For the benchmark model, the weighted F1 score on the black student population is 0.46 and the weighted F1 score for the Hispanic student population is 0.45.

The achievement gap labeling function is applied to find the weighted F1 score of the Decision Tree Classification models. It is expected that the weighted F1 score will increase since the achievement gap labeling itself takes care of reducing the room for incorrect predictions. Nevertheless, the labeling scheme is the same for all classification models and their relative performance contains information that may help understand the nature of the dataset better. That is, by knowing which classification algorithm works best, then it may be possible to understand what covariates are the most relevant in this study and what necessary information needs to be gathered to further increase predictive performance.

The results from the Decision Tree Classification models show at least a 20% increase in weighted F1 score. For black students, the tree classification model has a predictive score of 0.67 while the model yields 0.66 prediction score on Hispanic students. The top predictors for these models and their importance are shown in Figures 4 and 5 for black and Hispanic students, respectively. Both race groups show that the education gap in the parents is the top predictor of the achievement gap. The second top predictors are related the percentage of males or females with a Bachelors degree or higher for the black and Hispanic student populations, respectively. It is important to note that the education gap and percentage of higher education degrees is an instance of multi-collinearity and needs to be accounted for during the tuning of the ML models as noted above.

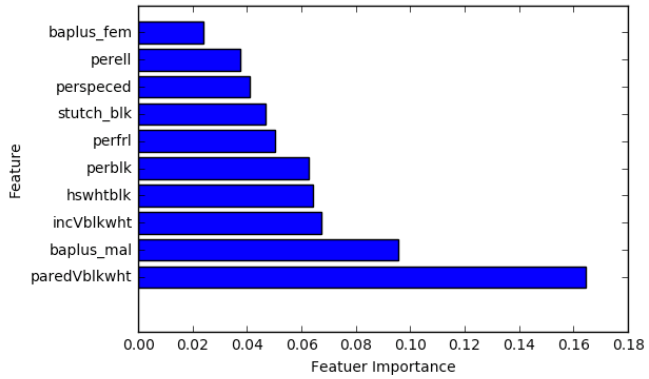

Figure 4. Decision Tree Classification model for black students yielding a weighted F1 score of 0.66 on the validation set.

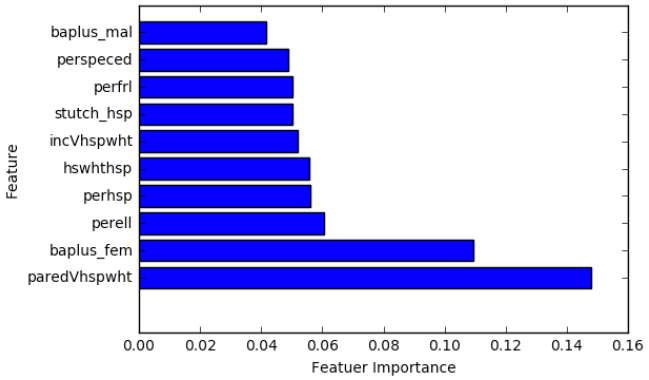

Figure 5. Decision Tree Classification model for Hispanic students yielding a prediction weighted F1 score of 0.65 on the validation set.

From the results of the Decision Tree Classification model, the top five covariates for each race are chosen to subset the training and validation sets for following Logistic Regression and SVM models. This is helpful for feature and dimensionality reduction. For the black students, the top covariates are: parent education gap (paredVblkwht), percentage of males with Bachelors degree and above (baplus_mal), income gap (incVblkwht), information index (hswhtblk), and the percentage of black students (perblk). The covariates for the Hispanic students are: parent education gap (paredVhspwht), percentage of females with Bachelors degree and above (baplus_fem), percentage of English Language Learners (perell), and percentage of Hispanic students (perhsp). All of these covariates are measured in a per district basis. After feature reduction, the decision tree classification weighted F1 score is similar to pre-feature reduction. For the black student population, the weighted F1 score drops to 0.65 from 0.66 and for Hispanic students the weighted F1 score increases from 0.65 to 0.66. Therefore, it is safe to assume that the top five features contain enough information to run the Logistic Regression and SVM models. Figures 6 and 7 capture the feature importance of the covariates used from this point forward using the Decision Tree Classification model.

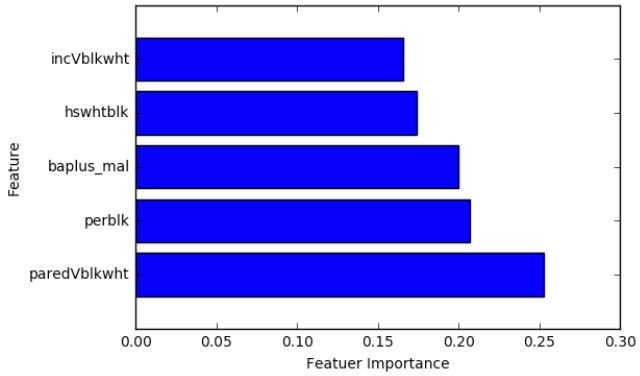

Figure 6. Decision Tree Classification model on the subset of covariates for black students yielding a weighted F1 score of 0.67 on the validation set.

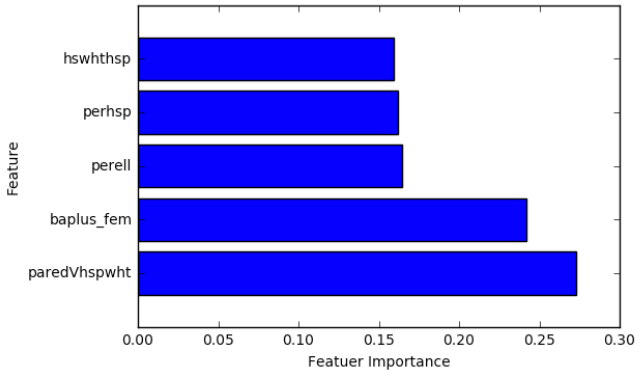

Figure 7. Decision Tree Classification model on the subset of covariates for Hispanic students yielding a prediction weighted F1 score of 0.66 on the validation set.

The next ML models tested are Logistic Regression and SVMs using the top five covariates determined by the Decision Tree Classification models in the above section. As summarized in Table 3, the decision tree models outperform the logistic regression and SVM models by at least 6 percent. The next step is to refine each of the models and determine the optimal weighted F1 score possible from the current set of datasets, transformations, feature reductions, and ML models.

The next step is now to refine the models by doing a grid search on critical parameter for ML model. In the case of the Decision Tree Classification, the important parameters are: (1) criterion, which determines if quality of split is determined by the Gini impurity or information gain and the (2) maximum depth, which determines the maximum depth of the tree. The values considered for the criterion parameter are ‘gini’ and ‘entropy’. The maximum depth is explored from 15 to 40 by increments of 1. Given number of samples, this range for the maximum depth was chosen to enable convergence without risking over fitting the data. From running the tuning functions, the optimal values for the parameters used to describe the black population are ‘entropy’ for the criterion with a maximum depth of 16. The parameter values for the Hispanic population are ‘gini’ for the criterion

and 16 for the maximum depth. Tuning of the parameters does not result in gain of prediction accuracy (Table 4).

Table 3. Weighted F1 score comparison of the Decision Tree Classification, Logistic Regression, and SVMs models on the black and Hispanic validation sets.

	Machine Learning Model	Training Time (sec)	F1 Training	F1 Validation
0	Decision Tree Classifier - Black	0.21	0.73	0.65
1	Decision Tree Classifier - Hispanic	0.24	0.74	0.66
2	Logistic Regression - Black	0.80	0.59	0.59
3	Logistic Regression - Hispanic	0.93	0.54	0.54
4	Support Vector Machines - Black	106.55	0.57	0.57
5	Support Vector Machines - Hispanic	123.54	0.53	0.52

The parameters of interest for the Logistic Regression models are: (1) penalty, which determines between the L1 and L2 type of regularization and (2) C, which is the inverse of the regularization strength. In this study, the values tested for the C parameter are 1, 10, 100, and 200. The C parameters were chosen in order to span different orders of magnitude. For both black and Hispanic datasets, the optimal penalty is L2 penalized logistic regression. While the optimal C parameter for the black student population is 200 and 100 for the Hispanic group. Similar to the Decision Tree Classification, tuning of the parameters does not result in gain in prediction accuracy (Table 4).

For the SVMs models, the parameters to tune are: (1) C, which is similar to the penalty parameter of Logistic Regression and (2) gamma, which is the kernel coefficient for the radial basis function. The radial basis function (rbf) is kept for as the kernel for this ML model since results using the linear kernel yielded lower accuracy score. Moreover, the computational time of running of SVMs on this dataset is significantly larger so it is not practical to tune multiple parameters on the SVM models. Therefore, a small parameter window is explored for the C and gamma parameter. For the C parameter, the values searched are 0.1, 1, 10 to gauge different orders of magnitude. While for the gamma parameter, the parameters are kept near the recommend 1/number of features criteria resulting in searching 0.3 and 1.3. From the tuning, the optimal parameters for the black students and Hispanic students are 10 for the C parameter and 1.3 for the gamma parameter. There is a slight increase in the weighted F1 score for SVMs (Table 4). Given the high computation cost and the low performance score compared to Decision Tree Classification, further optimization of the SVM models is not pursued.

Table 4. Weighted F1 score comparison of the tuned Decision Tree Classification, Logistic Regression, and SVMs models on the black and Hispanic validation sets.

	Machine Learning Model	Training Time (sec)	F1 Training	F1 Validation
0	Decision Tree Classifier - Black	0.20	0.73	0.65
1	Decision Tree Classifier - Hispanic	0.21	0.72	0.66
2	Logistic Regression - Black	0.89	0.59	0.59
3	Logistic Regression - Hispanic	1.07	0.54	0.54
4	Support Vector Machines - Black	301.36	0.60	0.60
5	Support Vector Machines - Hispanic	469.67	0.54	0.54

Results: Model Evaluation and Validation

The final model evaluation and validation is done on the test set that was not used during the training and tuning of the ML models. This final test set consists of the covariate and achievement gaps data pertaining to the academic year 2013. The tuned ML models are ran on the original training sets with the 2013 data as the test set. After tuning the best performing model consists of the Decision Tree Classification ML models. There is a slight loss in prediction accuracy, which means the models generalized well to the final test set. That is choosing a maximum depth of 16 allows retaining prediction accuracy while not over fitting to the training set. While the optimal criterion selection differs for the black and Hispanic datasets, it is possible that similar results are obtained using the same criterion.

Results: Justification

As compared to the dummy classification model, all of the classification models have better weighted F1 scores. This is in part, as mentioned earlier, that by binning the achievement gap into labels automatically increases weighted F1 score due to reduction in the range of values that need to be predicted. Nevertheless, by employing the ML models, it is possible to see what are the top covariates that predict the achievement gap for each race.

Conclusion: Free-Form Visualization

As shown in Figure 8, employing classification ML models results in significant improvement in the prediction capacity of the achievement gap relative to the baseline dummy classification model. The best performing models are based on Decision Tree Classification, while Logistic Regression and SVMs have similar performance. Both race groups result in similar model performance despite having slightly different covariates used for features. As shown in Figures 6 and 7, the common features in the top 5 covariates for black and Hispanic students are the parent education gap, the respective race percentage, and the information index. The last two features have strong cultural ties. For instance, the race percentage contains information on the diversity of the population in the school district. Similarly, the information index is the computed average deviation of the school diversity from the school district racial diversity. Therefore, the top covariates used to determine the classification performance of the ML models in Figure 8 have implications that are more than income inequality.

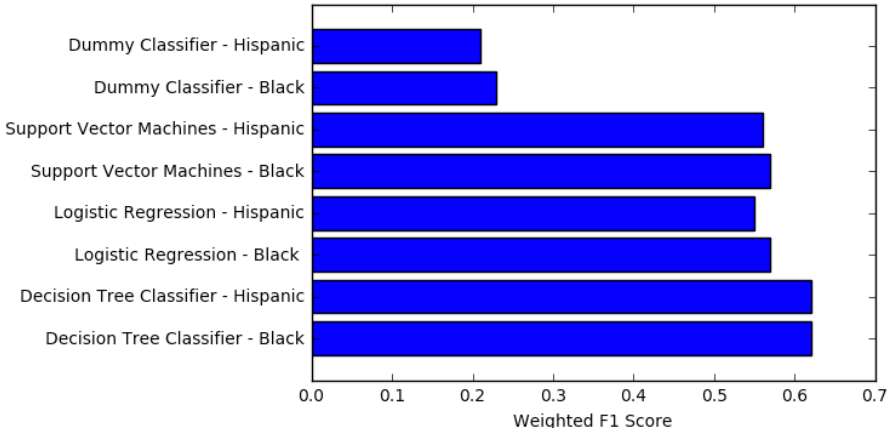

Figure 8. Summary plot of weighed F1 scores for the models used to predict the achievement gap classification for black and Hispanic students on the final test set.

Conclusion: Reflection

This project was helpful in understanding the subtleties in defining a problem statement and finding a dataset that can help solve it. In the case of predicting the achievement gap, it is difficult to quantify many of the relevant features that play out in the classroom on a day-to-day basis. So this was a useful exercise to understand the extent of the potential stored in the information that can be quantified with current methods and procedures. As it was shown, by running a classification algorithm it was possible to identify features that can trend well with overall status of our education system but fail to capture a significant portion of academic performance. This made it difficult to fully exploit the tuning process of ML models because the ML models with their default parameters already provided close to the optimal weighted F1 score. Moreover, despite the relative simplicity of the dataset, it was also clear how it is difficult to tune SVMs when the number of samples is large. The computational time in itself using SVMs made it difficult to properly span the parameter space appropriately.

Moreover, as can be seen in the income gaps, education gaps, and percentage of degree holders, the covariates can be highly collinear. While the ML models can deal with collinearity, it is not possible to extract more information from such features. Therefore, the datasets are limited in the amount of information that can be used to predict the achievement gap. This is not surprising, as it is known that many students can succeed despite coming from low-income background from parents without a formal education. The opposite is also a known phenomena, where educated parents may not have the time to focus on the educational development of their children.

Conclusion: Improvement

A key way to improve this project is to complement the covariates dataset with other potential features that may help predict the achievement gap. This may be possible by using the school district identification numbers and cross-referencing them with counties and Census lots. By doing so, it may be possible to link datasets from other sources and

studies that capture trends that may help understand the cultural makeup of a given district. For instance, information on parents working hours, number of jobs, commuting time, and disposable income close to the resolution of the school district level may provide a boost in achievement gap prediction. If parents to do not have the time to spend time with their kids, then it is possible that the kids will not be focused at home or at school. Other relevant information consists of the racial make up of the teachers and schools themselves. That is, is the diversity of student population reflected by the diversity in the teachers? Cultural alignment between teachers and students is critical in establishing communication and accountability among students, parents, teachers, and administration. Datasets that capture more of the cultural underpinning is necessary to get a better understanding and prediction of the achievement gap.

In terms of technical improvement, it will be beneficial to dig into the spatial distribution of the achievement gaps. In this project, I did not account the location of the school districts. That is, it is possible to introduce another set of columns that account for the proximity of neighboring school districts and their performance. Moreover, it is also possible to split datasets again between the school districts that trend well with the covariates and the school districts that are difficult to predict. From the latter dataset, it may be possible that other covariates are more predictive of their performance or they share common characteristics that can benefit from a specific form of data augmentation. Fortunately, there are many different ways to take this project to the next step.

References

1. Sean F. Reardon, *School District Socioeconomic Status, Race and Academic Achievement*, (2016).
2. McKinsey & Company, *The Economic Impact of The Achievement Gap in America's Schools*, (2009).
3. Sean F. Reardon, Demetra Kalogrides, Andrew Ho, Ben Shear, Kenneth Shores, Erin Fahle. (2016). Stanford Education Data Archive. http://purl.stanford.edu/db586ns4974.
4. Motoko Rich, Amanda Cox, and Matthew Bloch, *Money, Race, and Success: How Your School District Compares*, The New York Times, (2016).
5. William H. Jeynes, *A META-ANALYSIS The Effects of Parental Involvement on Minority Children's Academic Achievement,* (2003)
6. National Center for Education Statistics, *Trends in Academic Progress: Reading 1971-2012 / Mathematics 1973-2012,* (2013)