Socio-economic Status and the Impact on Academic Performance

Vuong Tran
Georgia Institute of Technology

Kshitij Srivastava
Georgia Institute of Technology

Ben Spivey
Georgia Institute of Technology

Nick Orangio
Georgia Institute of Technology

Dave Dyer
Georgia Institute of Technology

1 INTRODUCTION & PROBLEM DEFINITION

The impact of socio-economic status (SES) on students' academic performance (AP) is a well-studied area of research. Academic performance is often positively correlated with higher resources/funding [1]and negatively correlated with higher needs, such as free and reduced lunch programs, but the magnitude of the correlation is not widely agreed-upon[2]. In this project, we model the impact of SES on AP using New York state school data. We will produce a working model and a visual interface to explore the model interactively. We improve on status quo by investigating additional confounding factors (e.g., teacher education), evaluating Machine Learning (ML) models compared to state of the art fixed-effect & classical statistic-based models [3], and creating insightful, interactive data visualizations.

2 SURVEY

The relationship between SES and academic achievement has been demonstrated earlier in Benner et al. [5], Zwick [6], Farooq [4], and Battle et al. [7]. All prior work agree in differing measures that SES is an important predictor of academic achievement in combination with other factors like parental involvement [5], parental education level[4], systematic test scores[2], ethnic group [6] and race [7]. We've also seen that [4–7], most of prior research relies on parametric models where there is an attempt to quantify said relationships by evaluating relative goodness of fits using ANOVA and T-Tests. While this is a good practice to understand relative relationships, it fails to capture random effects and confounding factors as in [6, 7]. We learned that linear fixed effects models are commonly used to account for confounding factors for educational studies [1, 13, 14], but these may not account for any non-linear relationships as a machine learning model would. Typically, SES is measured by a few different measures – Free & Reduced Lunch Ratio (FRL)[2], family education levels[4], occupation, and income. Academic performance, meanwhile, is typically measured by GPA or some form of standardized testing [2]. We intend to explore the relationship between other, more novel, aggregate measures of both SES and academic performance.

3 PROPOSED METHOD

3.1 Intuition & Innovation

Previous NYSED studies based on these data considered relationships between N/RC (Needs to Resource Ratio) and academic performance at a district level but have not evaluated SES correlations to academic performance [17]. Our method evaluates SES factors at the school level and includes confounding factors like staff experience and school type (public/charter). We evaluate the NYSED and Cornell datasets for SES factors related to school student population (e.g., economically disadvantaged students) and school financial statistics (e.g., total revenue). Compared to similar works [1, 6, 7, 13, 14], we plan to use not just parametric models but also machine learning (ML) models to consider non-linear relationships without holding data distribution assumptions. Some clear differences and innovation in our approach are

- Evaluate socio-economic factors in addition to N/RC which only considers Poverty and Free lunches to assess performance.
- Use machine learning to model non-linear relationships hitherto unexplored in this kind of study.
- Use data visualizations not only at the EDA stage to explore covariate relationships but also to depict Actual vs Expected across county/school level.

3.2 Approach

This analysis uses data from Cornell's New York Education Data Hub and from the NYSED Graduation Rate and Report Card databases. We first identify proxies for both SES and Academic Performance at the school, district, and county level over the past five years. The proxy we use for academic performance at a school, district, county level is Graduation Rate. The NYSED SES factors and covariates include student subgroup percentage per school: ethnicity, sex, disability, economic disadvantage, English proficiency, homelessness, in foster care, migrant status, and parents in armed forces. The NYSED dataset also includes confounding factors that can affect school AP: charter/public school status and inexperienced staff (less than four years) and teachers out of certification. Cornell SES factors include school factors such as school revenue, expenditures, and salaries and district/student factors such as property value, gross income, and pupil wealth ratio.

Initial exploratory data analysis involves data cleaning, identifying correlations between the SES factors, confounding factors, and Academic Performance, identifying whether confounding factors (e.g. charter schools) associate with Academic Performance, and fitting ML models to estimate feature importance on NYSED and Cornell datasets. The data cleaning step removes factors that are highly correlated before assessing feature importance, resulting in independent variables suitable for analysis. Given our choice of independent variables (SES and confounding factors) and dependent variables (Graduation Rate), we use bivariate analysis to examine relationships in the data with school and district level observations within the state. We investigate the non-parametric correlation between the independent variables and Graduation Rate by calculating the Rank Spearman Correlation and comparing with out-of-bag impurity feature and permutation importance derived from random forest modeling. Heat-maps and bar charts are used to visualize the correlation coefficients and feature importances and identify patterns and interactions that exist. A Mann Whitney U Test is used to evaluate if graduation rate distributions differ when subsetted by factors such as charter schools.

Additional exploratory data analysis was performed using dimensionality reduction with Principal Component Analysis due to the many Cornell independent variables and linear regression and random forest models to predict graduation rate. The predictions are used to validate the model accuracy with mean squared error.

Visualization includes a data analytics dashboard built with D3 and JavaScript that plots county level statistics in a choropleth showing all counties in New York state colored by the SES factor, confounding factor, or graduation rate. The choropleth also serves as a geographic fit assessment and identification of outlier data across the counties - some counties were found to have missing data for some factors.

4 EVALUATION

4.1 Objectives & Design of experiments

Our end objective is to learn relationships between SES and Academic Performance. The below is a list of objectives with their associated experiments -

- **Objective**: Identify proxies for SES and Academic Performance **Experiment**: Examine Academic Performance at various levels of aggregation and evaluate in terms of richness of SES variables.
- **Objective**: Determine which variables correlate with academic performance and quantify their degree of association **Experiment**: Assess univariate statistics by plotting correlations for both continuous and categorical variables.
- **Objective**: Assess how each of the socio-economic variables are geographically distributed across counties **Experiment**: On a choropleth, show the distribution of socio-economic variables through color palettes.
- **Objective**: Fit models to test relationship between Socio-economic variables and Academic Performance **Experiment**: Fit exploratory models, assess fit through error metrics or explained variance. Infer most important predictors.

4.2 Observations & Experiments

Identify proxies for SES and Academic Performance - For this objective, we took a multipronged approach - the grad rate database from NYSED contains graduation rate and drop out rates as proxies for Academic Performance along with a small set of SES features. The SES metrics for the NYSED dataset include school percentage of economic disadvantage, homelessness, and foster care. Other covariates such as ethnicity, sex, disabled status, and English proficiency are also included. The Cornell dataset includes many more SES features including school revenue, expenditures, salaries, district property value, gross income, and other financial metrics. This allows a more granular analysis of the factors that can be used as proxies for SES. The Cornell dataset contains student cohort information such as class size and total students graduated, which were used to calculate graduation rate as the proxy for Academic Performance.

Determine which variables correlate with academic performance and quantify their degree of association

We performed analysis elucidating the distributions of our Academic Performance estimate. Based on the histogram for graduation rate (1), we have an idea that it has a left skew and if we use a parametric model, we may consider transforming the distribution. The Cornell dataset shows a similar distribution, providing confidence that the two datasets are comparable and can be used in separate analyses to reach valid conclusions. We examined the correlation matrix for the Cornell dataset (3) .

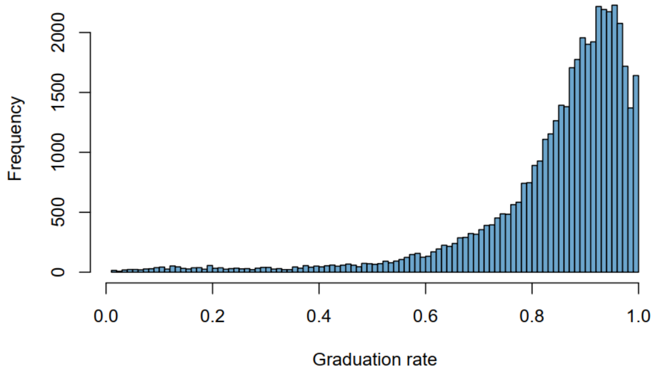


Figure 1: Graduation rate at a school level using NYSED

In the Cornell dataset, there are no strong correlations between graduation rate and the selected socioeconomic variables; however graduation rate has a relatively high negative correlation with *staid* and *boc* variables which represent State Aid and BOCES instructional expenditure. The SES variables show high correlations with one another. For example, teacher salaries *tsal* and Pupil instructional expenditure *pps* seem to be highly correlated. High correlations among independent variables suggest that machine learning models may do better in predicting academic performance compared to parametric models.

The NYSED dataset is used to analyze correlations between SES indicators and graduation rate. The SES

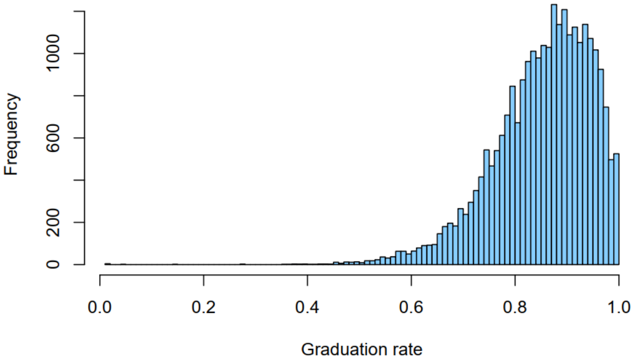


Figure 2: Graduation rate at a school level using Cornell

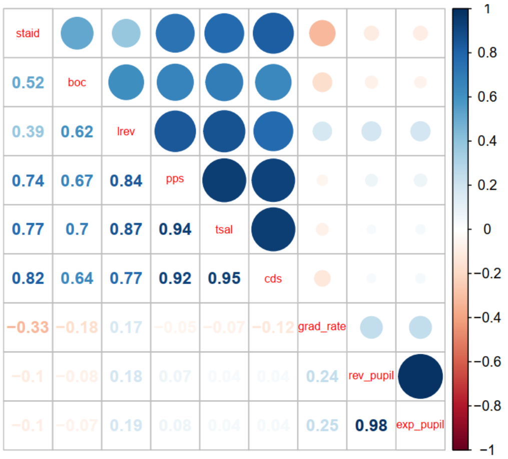


Figure 3: Correlation heatmap for Cornell dataset

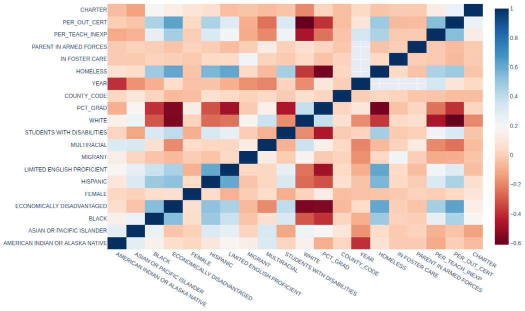


Figure 4: Correlation heatmap using NYSED dataset

indicators are calculated as the percent student population for each school where each school across the state serves as a data point. Confounding variables included are school type (public or charter) and teacher and principal experience and certifications status. The Rank Spearman Correlation coefficient is calculated between all SES indicators, confounding variables, and the graduation rate.

The features that are most negatively correlated with the graduation rate in order are: Homelessness (-0.58), Economically disadvantaged (-0.56), Limited English proficiency (-0.49), Students with disabilities (-0.46), and Percentage out of certification (-0.39). Ethnicity also had correlations with graduation rate to a lesser degree with the strongest negative correlations as Black (-0.39) and Hispanic (-0.32). It must be noted that these are bivariate correlations which do not imply causation and additionally, by definition, do not take into account the impact of other variables on graduation rate.

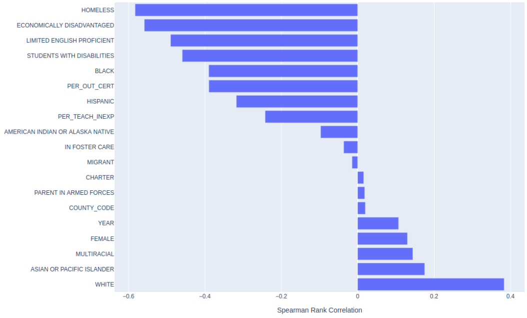


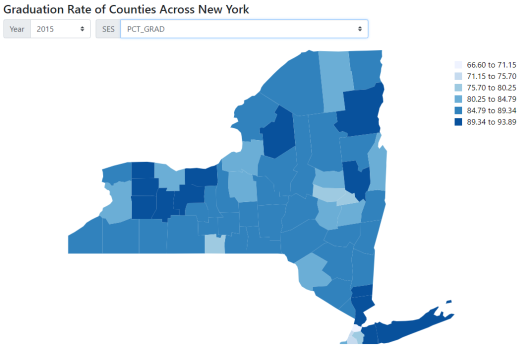
Figure 5: Rank Spearman correlations for graduation rate using NYSED

The Spearman plot shows the correlations with graduation rate in (5). The centered bar graph provides an alternative method to view correlation strength compared to a heatmap.

We notice similar observations in the correlation graphs for both the datasets. Socio-economic variables that imply economic disadvantage correlate highly with a low graduation rate. Further, a several socio-economic variables are correlated with one another.

Objective: Assess how each of the socio-economic variables are geographically distributed across counties

As described in our references, the southern counties have high graduation rate, but there is no discernible pattern here especially when compared to the poverty rate by county in (7).


Graduation Rate of Counties Across New York

Year 2015

SES PCT_GRAD

Figure 6: Graduation rate by county

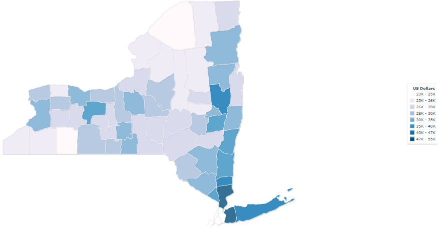


Figure 7: Poverty rate by county

We do see in (8) that counties that are economically disadvantaged have lower graduation rates compared to counties that are not economically disadvantaged. Further, in the next section we will perform hypothesis tests to check the statistical validity of the claim that economically disadvantaged schools/students have lower graduation rates.

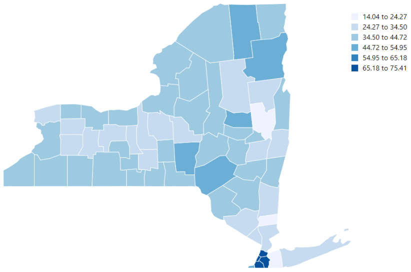


Figure 8: Economically disadvantaged by county

Objective: Fit models to test relationship between socio-economic variables and Academic Performance

Within the correlation and geographical distribution sections, we have worked towards identifying factors that may be considered predictors in a predictive model for Academic Performance. In this section, we will evaluate these assumptions and then fit some models to verify/reject hypotheses.

In our first set of experiments we tried to determine whether graduation rate differences for certain SES variables identified in the NYSED dataset are statistically significant. For this exercise we will compare distributions visually and use a parametric test to validate or disprove the null hypothesis which is - Graduation rate aren't any different for SES variables in consideration.

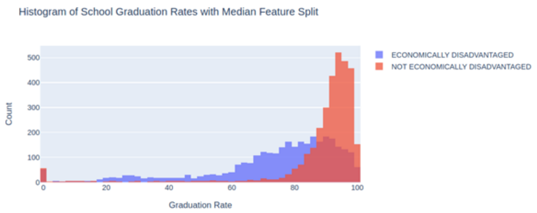


Figure 9: Count of schools by graduation rate for high and low economic disadvantage

In (9), schools are grouped as being above or below the median for economically disadvantaged (ED) student population. Schools with above median percent of ED students demonstrate a statistically significant lower mean graduation rate. The distributions are compared with a one-side Mann-Whitney U Test whether the ED is less than not ED. The test p-value (p<0.001) supports rejecting the null hypothesis that these distributions are equivalent.

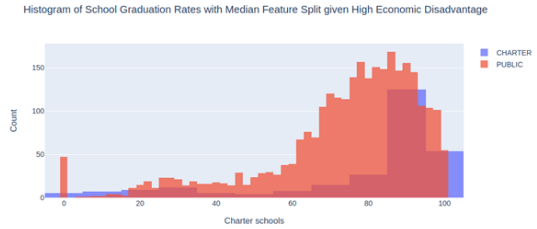


Figure 10: Graduation rates for high economically disadvantaged population for charter and public schools

The distribution of graduate rates for charter schools for students of lower socio-economic status is evaluated by selecting schools having a percentage greater than the median for the economically disadvantaged (ED) factor. These schools are then split between charter and public schools. The distributions are compared with a one-sided Mann-Whitney U Test whether the distribution of Charter graduation rates differs from Public graduation rates. The test p-value (p<0.001) supports rejecting the null hypothesis that these distributions are equivalent.

The hypothesis tests above indicate that school type and economic status of the school may be associated with graduation rates. However, these analysis approaches do not control for the effect of other variables, an area where machine learning provides additional insight. To evaluate feature importance using a multivariate approach, we fit a random forest model to determine variable importance measures for Graduation Rate using the NYSED dataset.

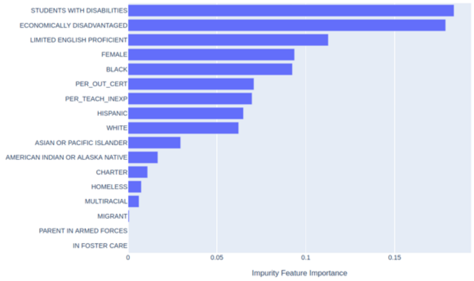


Figure 11: Impurity feature importance

The impurity-based feature importance produces a similar ranking of top features as what we saw in (5) except that the Homeless feature is much lower and the Female feature is much higher. Some factors that have high correlation with graduation on a univariate basis are seen more clearly with the Spearman Rank Correlation than with the impurity feature importance. Random forest impurity feature importances can have a tendency to prefer features with high cardinality and can dismiss one feature over another if they are correlated. However, random forests are a multivariate machine learning method and generally represent a multivariate system more accurately than univariate correlations.

We also used a random forest regressor to model graduation rates for the Cornell dataset to examine feature importance in a more feature rich and granular dataset.

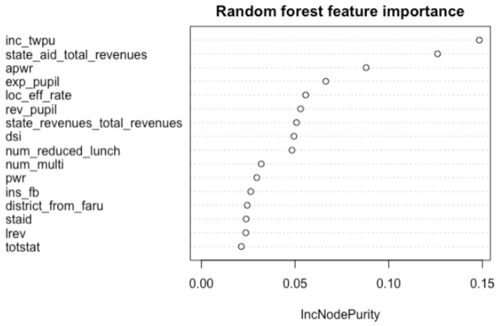
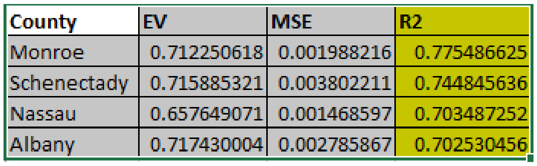


Figure 12: Feature importances on the Cornell dataset

The Cornell dataset strongly shows the effect of economic disadvantage and also identifies specific economic metrics that are most closely associated with graduation rate. The top four variables with the highest feature importances are *inc_twpu* which implies NYS Adjusted Gross Income / Total Wealth Per Pupil (TWPU) , *state_aid_total_revenues* which implies the level of State Aid received, *apwr* which implies Pupil wealth ratio and *exp_pupil* which implies pupil expenditure. All these variables cement the value of economic advantage in determination of graduation rate. Both Random Forest models (on the NYSED and Cornell datasets) explain around 65 percent of the variance in graduation rates, measured on the training dataset.

We also modeled county level variations in the Cornell dataset, to evaluate if the significant SES variables differ across counties when modeling graduation rates. On the top four county with credible sample sizes, we noticed that economic advantage stays the most important predictor (13).



County	EV	MSE	R2
Monroe	0.712250618	0.001988216	0.775486625
Schenectady	0.715885321	0.003802211	0.744845636
Nassau	0.657649071	0.001468597	0.703487252
Albany	0.717430004	0.002785867	0.702530456

Figure 13: Model fit metrics across 4 counties with most credible sample sizes

5 CONCLUSION & DISCUSSION

Our initial objective was to elucidate relationships between SES and Academic Performance. The below is a list of objectives and associated conclusions:

- **Objective**: Identify proxies for SES and Academic Performance **Conclusion**: SES metrics identified from two datasets include: economic disadvantage, disabilities, (NYSED dataset), and school revenue, expenditures, salaries, district property value, district average gross income, and pupil wealth ratio (Cornell dataset). The primary Academic Performance metric identified was graduation rate.
- **Objective**: Determine which variables correlate with academic performance and quantify their degree of association **Conclusion**: In our analysis, we noticed that features that imply economic advantage - state aid, instructional expenditure are negatively correlated (weak correlations) with academic performance in the Cornell dataset. For the NYSED dataset, we noticed that homelessness, economic disadvantage, limited English proficiency and disabled status have high negative correlations with graduation rate.
- **Objective**: Assess how each of the socio-economic variables are geographically distributed across counties **Conclusion**: Geographically, using NYSED data, it is observed that counties that are economically disadvantaged have lower graduation rates compared to counties that are not economically disadvantaged.
- **Objective**: Fit models to test relationship between socio-economic variables and Academic Performance **Conclusion**: We performed hypothesis tests which indicate that graduation rates are statistically lower for economically disadvantaged schools and that graduation rates differ for charter schools compared to public schools. We fit machine learning models indicating economic disadvantage is the biggest predictor of graduation rates across both datasets, while identifying specific measures of economic disadvantage that associate most closely with graduation rates.

6 DISTRIBUTION OF TEAM MEMBER EFFORT

In entirety, team effort has been equally distributed equally among team members. Our team members skill sets were broad and thus all of us have contributed to each areas more/less in an equal amount. We adopted an agile approach, and remained flexible to re-distribute work load as new information/requirements arose.

REFERENCES

[1] Jinnai, Y. (2014). "Direct and indirect impact of charter schools' entry on traditional public schools: New evidence from North Carolina." *Economic Letters*, 124, 452-456.

[2] Sirin, S. R. (2005). "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research." *Review of Educational Research*, 75(3), 417–453.

[3] Hearn, J. C. (1988). "Attendance at Higher-Cost Colleges: Ascribed, Socioeconomic, and Academic Influences on Student Enrollment Patterns." *Economics of Education Review*,7(1), 65-76.

[4] Farooq, M.S., Chaudhry, A.H., Shafiq, M., Berhanu, G. (2011). "Factors Affecting Students' Quality of Academic Performance: A Case of Secondary School Level." *Journal of Quality and Technology Management*, VII(II)01-04.

[5] Benner, A. D., Boyle, A. E., Sadler, S. (2016). "Parental Involvement and Adolescents' Educational Success: The Roles of Prior Achievement and Socioeconomic Status." *Journal of Youth and Adolesence*, 45, 1053-1064.

[6] Zwick, R. (2012). "The Role of Admissions Test Scores, Socioeconomic Status, and High School Grades in Predicting College Achievement." *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, 49(2), 23-30.

[7] Battle, J., Lewis, M. (2008). "The Increasing Significance of Class: The Relative Effects of Race and Socioeconomic Status on Academic Achievement." *Journal of Poverty*, 6(2), 21-35.

[8] Lee, J. O., Kosterman, R., Jones, T.M., Herrenkohl, T.I., Rhew, I.C., Catalano, R.F., Hawkins, J.D. (2016) "Mechanisms linking high school graduation to health disparities in young adulthood: a longitudinal analysis of the role of health behaviours, psychosocial stressors, and health insurance." *Public Health*, 139, 61-69.

[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L. (2018). *Explaining Explanations: An Overview of Interpretability of Machine Learning.* IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 80-89.

[10] Domanico, R. (2020). "NYC Student Achievement: What State and National Test Scores Reveal." *The Manhattan Institute for Policy Research* https://eric.ed.gov/?id=ED604331.

[11] Singla, K., Bose, J., Naik, C. (2018) *Analysis of Software Engineering for Agile Machine Learning Projects.* 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 1-5, doi: 10.1109/INDICON45594.2018.8987154.

[12] Berkowitz, R., Moore, H., Astor, R.A., Benbenishty, R. (2017). "A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement" *Review of Educational Research*, 87(2), 425–469.

[13] Winters, M. (2012). "Measuring the effect of charter schools on public school student achievement in an urban environment: Evidence from New York City." *Economics of Education Review*, 31, 293-301.

[14] Harris, D.N., Sass, T.R. (2011). "Teacher training, teacher quality, and student achievement." *Journal of Public Economics*, 95(7-8), 798-812.

[15] Domingos, P. (1998). *How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications.* 15th National Conference Artificial Intelligence (AAAI-98) and International Conference on Machine Learning (ICML-98): Workshop on AI Approaches to Time Series Problems.

[16] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). *Introduction to Statistical Learning: with Applications in R.* Springer. 147, 354-355.

[17] http://www.p12.nysed.gov/repcrd2004/information/similar-schools/guide.shtml *Which schools are similar*