Impact of Public K-12 School Features on Student Academic Performance

Sophia Kuen Cheng, Meixun Zheng, Chih-Pin Cho

Introduction

This project examines how school characteristics influence academic performance in U.S. public K-12 schools. The **supervised learning** component focuses on building models to predict school-level proficiency scores in **math** for elementary and middle schools, as well as classifying schools as above or below national average. The **unsupervised learning** component explores the distribution of Advanced Placement **(AP) course** resources and offerings across public high schools nationwide.

We were motivated to identify gaps in student academic performance and the availability of educational resources, along with the key factors contributing to these disparities. The insights from this study could inform **educational policies**, ultimately working toward reducing achievement gaps and fostering more equitable learning environments.

Methods and novel contributions

For supervised learning, we developed regression and classification models. The regression model utilized school characteristics to predict the percentage of elementary and middle school students achieving grade-level proficiency in math. The classification model used school features to categorize schools as above or below national average based on their math performance. **For unsupervised learning**, we developed principal component analysis and gaussian mixture models to learn which school features impact AP resource availability.

While educators have explored factors influencing student academic performance, with a focus on a subset of school characteristics (e.g. student-teacher ratio), our project went a step further to build and evaluate different machine learning models using a comprehensive list of school characteristics and national datasets. Our project contributes to the existing literature by offering a machine-learning based approach to understanding school resources and its impact on academic performance at a national scale.

Main findings

For supervised learning, our results show that key school features are good predictors of students' math performance in elementary and middle schools. However, we also found that the models could be biased towards minority students. **For unsupervised learning**, our modeling shows that the average salary of support staff is twice as correlated to student performance than the average salary of teachers. Overall, results show that it is possible to estimate the availability of AP resources to a school based on key school features.

Related Work

Studies have shown that individual school features such as teacher wages and class sizes play a role in student performances. Loeb and Page (2000) found that raising teacher salaries could reduce high school dropout ratings. Antoniou and colleagues (2024) built linear and non-linear models to study the impact of school and class size on student performance, with results showing that the ideal cut-off was 801 for school enrollment and 27 students per class. The author used an international dataset rather than the U.S. data set. Also, the authors did not use advanced modeling approaches as we did in this project. Regarding AP course enrollment, Warne (2017) noted that students from higher-income families, those attending suburban schools, and white students are more likely to participate in AP courses. The study also highlighted that the academic and financial benefits of AP participation remain inconclusive. However, Warne's research is only a literature review and theoretical discussion without model developments.

Related work from milestone 1

One of our Milestone 2 team members (MZ) conducted an exploratory analysis of U.S. K-12 school retention rates for her Milestone 1 project. However, our Milestone 2 project only incorporated the *"school characteristics"* dataset from Milestone 1. All other datasets used in Milestone 2 are unique, as the project has different goals and scope. Additionally, the Milestone 1 project analyzed school retention rates, whereas Milestone 2 examined math performance and AP course enrollments. Additionally, Milestone 1 relied on exploratory and statistical analysis, with no machine learning components. In contrast, Milestone 2 developed machine learning models.

Data Sources

Supervised learning

For supervised learning, we focused on school-level performance in standardized math assessment. It is important to note that standardized math assessments are only administered at the elementary and middle school levels. We used these datasets, all downloaded as CSV:

- School characteristics
- School internet data
- School finance data
- Math performance

(1) School characteristics data

This dataset includes all public K-12 schools in the U.S. from the National Center for Education Statistics (NCES) for the 2020-21 reporting year. It provides a unique school ID and key school attributes. The raw dataset has the shape of **(100722, 79)**.

- **Examples of categorical variables**: state, school level (elementary, middle, high), locale code (rural, suburban, town, city), and whether the school is a magnet, charter, or title 1 school (eligible for additional funding to reduce achievement gap among racial groups).
- **Examples of numeric variables**: total school enrollment, including breakdowns by gender and race; other features such as the number of students receiving free lunch, student-teacher ratio, and number of full-time teachers.

(2) School internet and finance data

These datasets provide additional school features from the 2020-21 reporting year.

- **Internet data**: Information on school internet access, such as whether the school has Wi-Fi and the number of Wi-Fi-enabled devices. The original shape is **(97575, 10)**.
- **Finance data**: Financial details such as total salaries for teachers, instructional support, and other school staff. The original shape is **(385134, 15)**.

(3) Math performance data

This dataset includes all K-12 schools from the 2020-21 reporting year and provides additional school characteristics and academic performance metrics. The original shape is **(3850524, 26)**.

- It contains **school features** such as the number of students who are homeless, have limited English proficiency, or come from economically disadvantaged families.
- It also contains the percentage of students meeting **grade-level proficiency** in math in each school.

Unsupervised learning

For unsupervised learning, we focused on AP course offerings and enrollment in high schools. We used the ***same school features*** datasets described earlier as input features, while the AP course enrollment data served as the target variable.

(4) AP course enrollment data

This dataset contains information on AP courses offered at each school from the 2017-18 reporting year (the most recent available). It includes three categories of AP courses: AP Science (Biology, Chemistry, Physics, Environmental Science), AP Math courses (calculus and statistics), and AP Computer Science courses. The raw data has the shape of **(97575, 94)**.

- **Examples of categorical variables**: Did the school have any students enrolled in any AP courses? Did the school have any students enrolling in AP math, science, or computer science courses?
- **Examples of numeric variables**: the total number of different AP courses offered at each school; the total number of students enrolled in each type of AP courses, with enrollment breakdowns by demographic groups.

Data Pre-processing

Initial data cleaning

- **Irrelevant columns:** We removed columns that we don’t need.
- **Column naming:** We renamed columns for readability and standardized the school ID column across all files.
- **Reserved codes:** Some columns contained reserved codes representing "Not Applicable" or other special cases (e.g., -1, -2, -3, -9, -11). We treated these as missing values.
- **Missing values:** During initial data preprocessing, we dropped missing values in key columns of interest. For modeling, we applied additional strategies to handle missing values, including imputation using the mean, median, or the most frequent category.
- **Invalid values**: Some columns contained inconsistent values, such as "Yes," "No," "N," and "M." We retained only rows where the value was either "Yes" or "No."
- **Data errors:** We identified schools with a reported total enrollment of 0. To maintain data accuracy, we dropped these schools.
- **Inconsistent data format:** In some files, the school ID started with a leading "0", while others did not. We removed all leading zeros across all files.
- **Inconsistent file formats:** We converted all datasets to a wide format, ensuring each unique school appears in one row with its corresponding features.

Data merging

After initial cleaning, we merged the datasets using the school ID column (NCES_SCH_ID).

- For supervised learning, we merged the school features dataset with the math performance dataset.
- For unsupervised learning, we merged the school features data with the AP course enrollment dataset.

Exploratory Analysis and Feature Engineering

We conducted Exploratory data analysis (EDA) on the two merged datasets to identify key patterns.

EDA for school features

For **categorical** variables, we performed frequency counts and visualizations to understand variable distributions. For numeric variables, we computed descriptive statistics and used boxplots to detect **extreme outliers**. For instance, we found that some schools reported an average teacher salary as high as $5 million per year, which was likely a data entry error. We dropped schools with extreme outliers and retained only those with an average teacher salary of $200,000 or below per year. As an example, the two figures below 2 illustrate the distributions of average teacher salary and student-teacher ratio after extreme outlier removal.

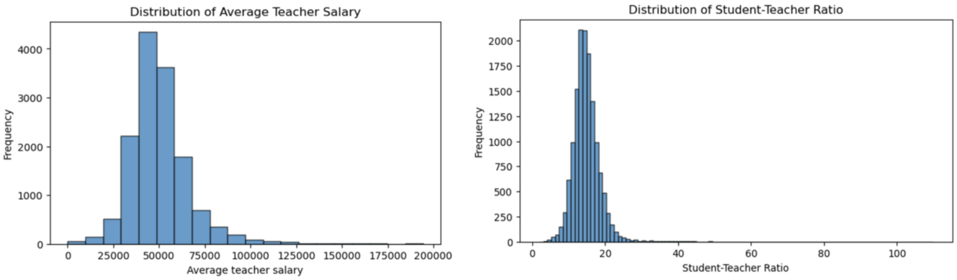


Figure 1. Distribution of average teacher salary (left) and student-teacher ratio (right).

EDA for target variables

For supervised learning

The EDA results show that, on average, **34%** of students per school meet grade-level proficiency in math. The distribution of the percent of students meeting grade-level proficiency math is presented in the figures below.

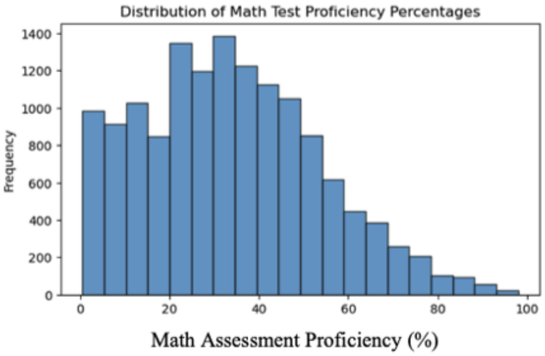


Figure 2. Distribution of the percentage of students reaching grade-level proficiency in math.

Unsupervised learning

EDA shows that across the nation's high schools, the percentage of students at each school enrolled in at least one AP course, when such courses are offered, ranged from 0 to 60%. The percentage of students enrolled in AP math, science, and computer science courses range from 0-20%, 0-20%, and 0-10% respectively.

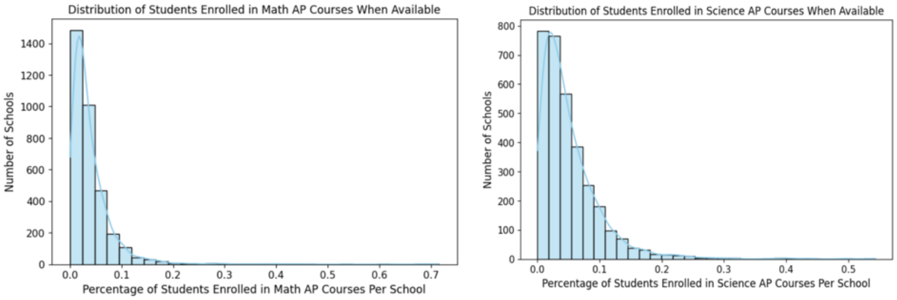


Figure 3. Distribution of students enrolled in AP math (left) and science (right) courses.

Computing new features

As part of feature engineering, we also created additional columns to refine school characteristics. To prevent the models from favoring schools with larger enrollments, we computed the **percentage of students enrolled** in each school by demographic features, including gender, race, and socioeconomic status, rather than using raw enrollment counts.

For AP courses (unsupervised learning), we summed up the **total number of AP courses** offered at each school and the **percentage of students** at each school who are enrolled in AP math, science, and computer science courses.

Final datasets

Our final dataset for supervised learning has the shape of **(14149, 31)**, whereas the final dataset for unsupervised learning has the shape of **(6012, 32)**. See Appendix A and B for the list of key variables.

Supervised Learning

For supervised learning, we developed **regression and classification** models. For the classifier, we predicted whether schools are “risky” on performance (“normal” and ”risky”) based on whether it is at/above or below the national average (34%) while for regression we predicted the percentage of students meeting grade-level proficiency at each school (0-100%).

We split the data into 70%, 15%, and 15% for training, validation, and test. We then transformed the data, including one-hot encoding for categorical variables and normalization for numeric variables.

(1) Classification model

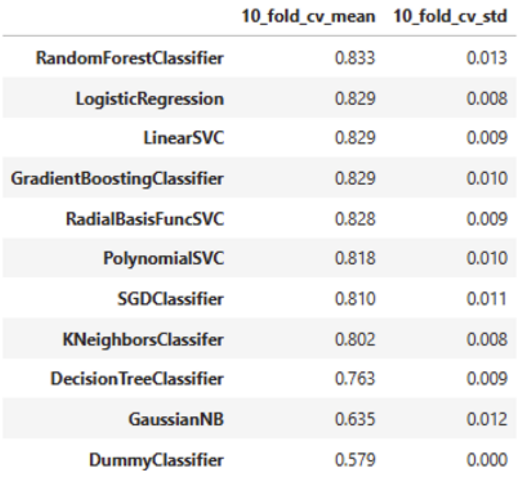
Model development

For classification, we tried methods such as KNeighborsClassifier, LogisticRegression, Support Vector Machine (SVM), and Gradient Boosting. We experimented with each approach and refined them to identify the best performing model. We used 10-fold cross validation to ensure model robustness. To optimize model parameters, we applied both GridSearchCV and RandomizedSearchCV to identify the optimal hyperparameters.

Model evaluation

- **Cross validation:** RandomForestClassifier and LogisticRegression had the highest accuracy score (0.833 and 0.829 respectively; see Table 1).
- **GridSearchL:** The RandomForestClassifier yielded a slightly higher AUC score (0.911) than Logistic Regression (0.905).

- **RandomizedSearchCV:** The best parameters obtained were n_estimators = 400, max_features = 6, and min_samples_leaf = 1.
- **Learning curve analysis:** we got 0.03 increase in AUC score as we increased the data size.



	10_fold_cv_mean	10_fold_cv_std
RandomForestClassifier	0.833	0.013
LogisticRegression	0.829	0.008
LinearSVC	0.829	0.009
GradientBoostingClassifier	0.829	0.010
RadialBasisFuncSVC	0.828	0.009
PolynomialSVC	0.818	0.010
SGDClassifier	0.810	0.011
KNeighborsClassifer	0.802	0.008
DecisionTreeClassifier	0.763	0.009
GaussianNB	0.635	0.012
DummyClassifier	0.579	0.000

Table 1. Cross validation results of different classification models.

For the best performing model (random forest), the following are the performance data on the test set.

- **Precision, recall, and f1-score**: accuracy =0.838, precision = 0.849, recall = 0.751, F1 = 0.797, and AUC = 0.827.
- **Feature importance**: We use a basic estimator-based method to retrieve and rank important features followed by a more robust permutation importance approach. While the permutation importance is more flexible and model-agnostic,it has a higher computational cost. Both methods produced similar feature importance.

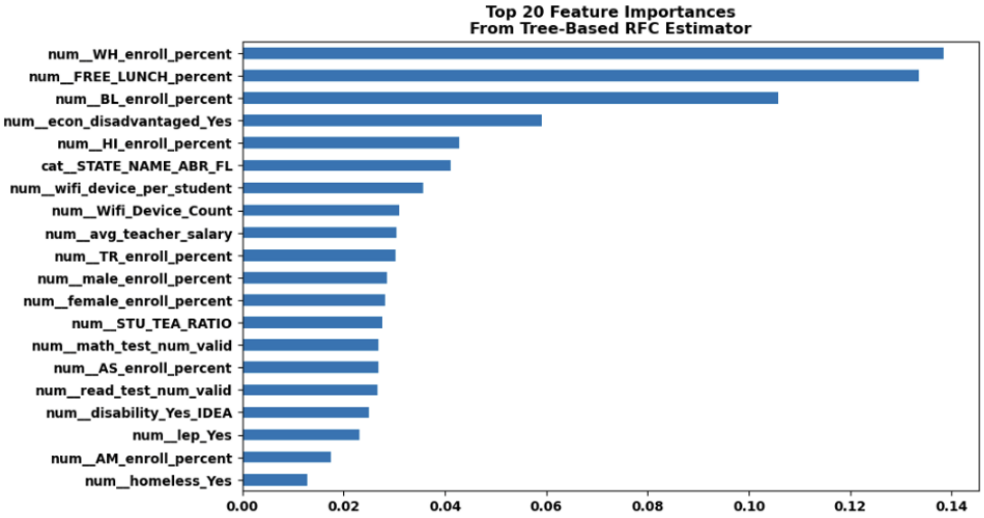


Figure 4. Top 20 features for the random forest classifier.

Failure Analysis: Analysis of the following three cases suggested that the model could be biased based on demographic features.

(1) School indexed 12692: incorrectly classified as “risky”. The school’s percentage of white student enrollment is lower than the national average while black enrollment is above national average – so was the percentage of students receiving free lunch.
(2) School indexed 13356: incorrectly classified as “risky”: The school’s percentage of white student enrollment is lower than the national average while Hawaiian student enrollment is higher than average. These contributed to the prediction error.
(3) School indexed 1111: incorrectly classified as “normal”. The school has a lower than average percentage of students receiving free lunch and lower black student enrollment. These factors contributed to the model incorrectly predicting it as “normal” why in fact it was risky.

We could be more careful during feature transformation such as scaling, and one hot encoding, as well as doing more robust grid search support, hyperparameter optimization and regularization to avoid model overfitting. Overfitting is a critical issue especially if some demographic groups are under-represented, which is the case with educational data.

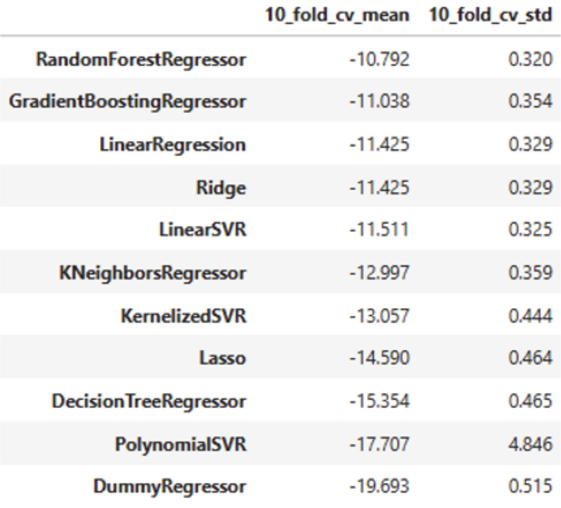
(2) Regression model

Model development

We tried a few regression models and gradually refined them to identify the best performing model, e.g., KneighborsRegressor, linear regression, Ridge and Lasso regression, Support Vector Machine, RandomForestRegressor and GradientBoostRegressor. We used 10-fold cross validation to ensure model robustness and applied GridSearchCV and RandomizedSearchCV to identify the optimal hyperparameters.

Model evaluation

- **Cross-validation**: We used Sklean’s “neg_root_mean_squared_error”. Since the values are negative, the lower the absolute value, the better the model performance. Table x shows that our top two performing models are random forest and ridge regression models.
- **GridSearchCV:** The random forest regressor achieved the highest r square value (0.696).
- **RandomizedSearchCV:** n_estimators = 500, max_features = 8 and min_samples_leaf = 1.
- **Learning curve analysis:** With increased sample size, we increased r square by about 0.08.



	10_fold_cv_mean	10_fold_cv_std
RandomForestRegressor	-10.792	0.320
GradientBoostingRegressor	-11.038	0.354
LinearRegression	-11.425	0.329
Ridge	-11.425	0.329
LinearSVR	-11.511	0.325
KNeighborsRegressor	-12.997	0.359
KernelizedSVR	-13.057	0.444
Lasso	-14.590	0.464
DecisionTreeRegressor	-15.354	0.465
PolynomialSVR	-17.707	4.846
DummyRegressor	-19.693	0.515

Table 2. Cross validation results of different regression models.

For the best performing model (random forest), we obtained these final performance data: MAE = 8.445, MSE =119.399, RMSE = 10.927, R square = 0.709.

Feature importance: The results are similar to the classification model.

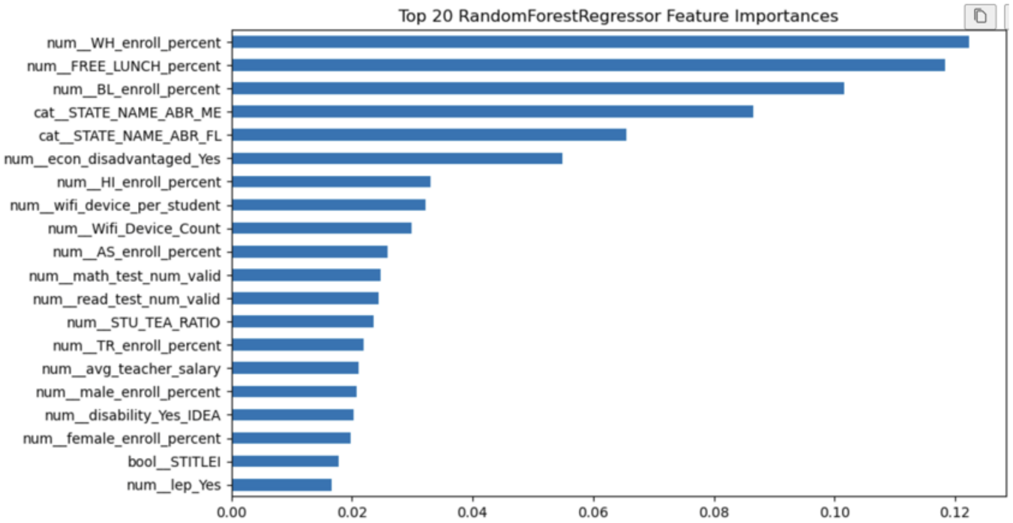


Figure 5. Top 20 features of the random forest regressor.

Failure analysis: As the target variable is a numeric (continuous) variable, we examined the difference between the predicted and the actual value. We focused on the schools with the highest value of prediction residuals:

(1) School indexed (6657): Actual value = 10, predicted value = 45. This is mainly because the percentage of white students in the school is much higher than the national average, and slightly lower black enrollment.
(2) School indexed (7422): actual value = 57, predicted value = 20. This is mainly because the school has very high black enrollment but low white enrollment.

As can be seen, the model makes errors based on students' demographic features. This is not surprising as we know the model output reflects bias in the training data. As we discussed during the introduction section, there were a lot of missing values and data entry errors in the raw dataset. The poor data quality might also have contributed to model errors. Additionally, we might need other features not in the current dataset to make more accurate predictions. Students' academic performance is impacted by many factors, including but are not limited to school features, teacher quality/dedication, family support, and students' own motivation and dedication. It would be helpful to use some of these features for future modeling.

Unsupervised Learning

Model development

For our unsupervised learning, we conducted **principal component analysis (PCA)** on a set of 19 school features to gain insight on how each of these features contribute to a student's success. We used the availability of AP courses and the percentage of students enrolled as a proxy for successful students. We developed **three** different PCA models, using two different flavors - vanilla PCA and KernelPCA. Given that we did not know what relationship exists, we used KernelPCA in case our data is nonlinear.

We first tuned the number of components to use for the PCA using GridSearchCV. For the KernelPCA, we additionally tuned the hyperparameters for the kernel and alpha values.

Given the diversity of school features, we suspect that they will form clusters of features. For example, a charter school is more likely to be in a city than a rural area, or an “alternative education” school is less likely to have WiFi than a regular school. To investigate this, we developed an additional **gaussian mixture model (GMM)** to identify these clusters with the goal of performing PCA on a cluster level.

In addition to tuning the hyper parameters, we also compared three different scoring methods for the grid search, including: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Negative Log Likelihood (*default for GMM*). Both AIC and BIC are variations on the default negative log likelihood scoring, by imposing a penalty for model complexity. All three scoring methods returned the same hyper parameters.

Model evaluation

We begin our model evaluation by verifying how representative it is of our actual data by performing an inverse transform of our input data and calculating the root mean squared error (RMSE) between the result and the original data.

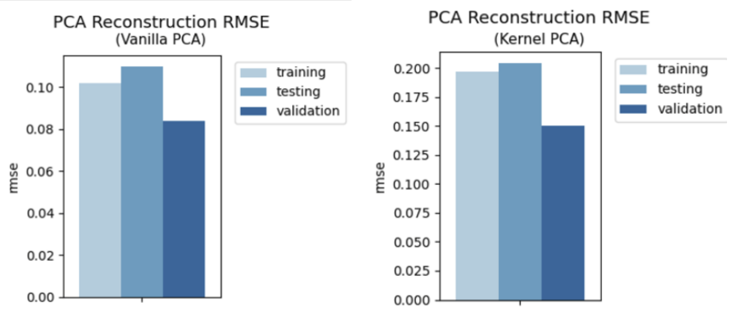


Figure 6. Root mean squared error (RMSE) of Vanilla PCA (left) and Kernel PCA (right).

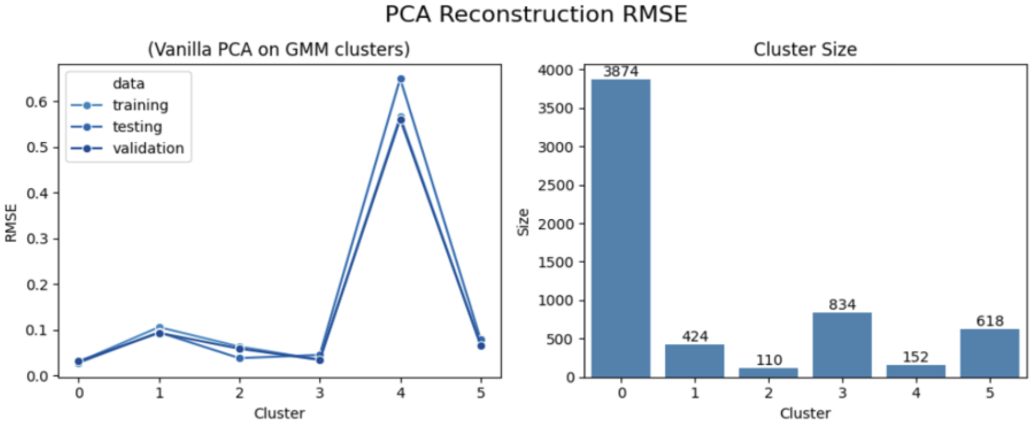


Figure 7. Root mean squared (RMSE) of Vanilla PCC on GMM clusters.

The difference in the reconstruction error between the vanilla and kernel PCA on the un-clustered data may indicate that the data has a more linear tendency than non-linear. Satisfyingly, the reconstruction error on the largest cluster is the lowest, indicating that cluster is well represented by PCA.

For the GMM, we additionally performed a bivariate analysis on each of the clusters and compared them with the initial bivariate analysis we performed as part of our exploratory data analysis. While the EDA bivariate showed rough shapes, some of which had directionality, the analysis on the clusters showed distinct shapes and patterns. For example, cluster 0 shows a clear bimodal kernel density estimate for

"NORM_TOT_FREE_LUNCH" (students receiving free lunch – normalized), while cluster 3 shows a possible 3 modes, but not nearly as well defined.

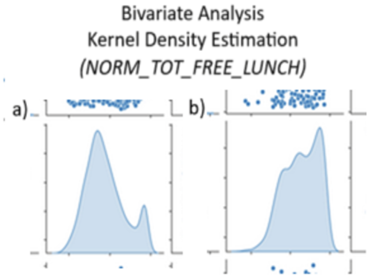
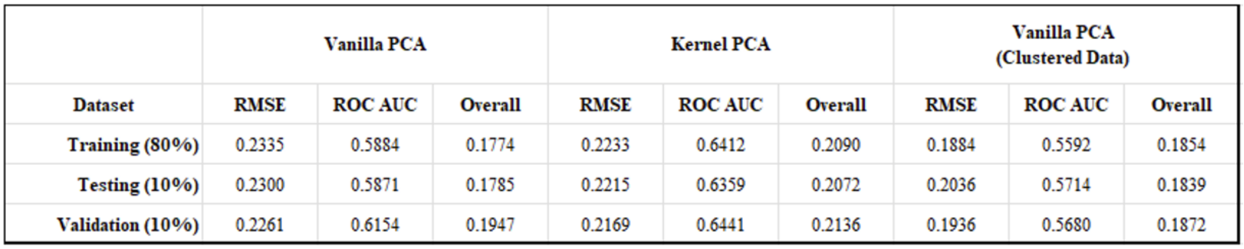


Figure 8. Bivariate analysis - Kernel dentistry estimation.
a) Cluster 0; b) Cluster 3

To analyze the effectiveness of the models, we fit each with our training data and then use them to predict our target features. **For the categorical target features**, we used a Ridge regression as our data is multicollinear. We performed hyperparameter tuning on the Ridge regression to identify the optimal parameters. Since this is a classification task, we chose the receiver operating characteristic curve (ROC) as the categorical target evaluation metric. **For the numeric targets**, we chose linear regression for result interpretability and used RMSE as the evaluation metric. To obtain an overall metric, we combined these two scores and took the average, where a higher overall metric means the model performs better. Given that the two metrics have opposite directionality, we combined them by subtracting the RMSE (lower is better) from the ROC AUC (higher is better).



	Vanilla PCA			Kernel PCA			Vanilla PCA (Clustered Data)		
Dataset	**RMSE**	**ROC AUC**	**Overall**	**RMSE**	**ROC AUC**	**Overall**	**RMSE**	**ROC AUC**	**Overall**
Training (80%)	0.2335	0.5884	0.1774	0.2233	0.6412	0.2090	0.1884	0.5592	0.1854
Testing (10%)	0.2300	0.5871	0.1785	0.2215	0.6359	0.2072	0.2036	0.5714	0.1839
Validation (10%)	0.2261	0.6154	0.1947	0.2169	0.6441	0.2136	0.1936	0.5680	0.1872

Table 3. Performance of three PCA methods.

Of our three models, the kernel PCA performs the best overall. When considering each target feature type, the clustered data performs better on numeric target features, but barely better than random on the binary target features. Both of the un-clustered data models performed similarly for numeric targets, with the kernel PCA performing slightly better on predicting binary targets.

We examined the principal component covariances reported for the vanilla PCA and observed that the school features most correlated to the top three principal components are NORM_salaries_support (salary for support staff-normalized), MAGNET_No (not a magnet school), and avg_teacher_salary (average teacher salary).

Evaluation of each cluster's PCA

We also evaluated each cluster's PCA individually, with results visualized in the figures below.

- We compared the cluster that performed the best for binary target prediction (cluster 2, when looking at validation results) with the clusters that performed the worst (clusters 1, 3 and 4). We found that **cluster 2 is less likely to have Title I schools** as well as **less likely to be in a city** when compared to clusters 1, 3, and 4.

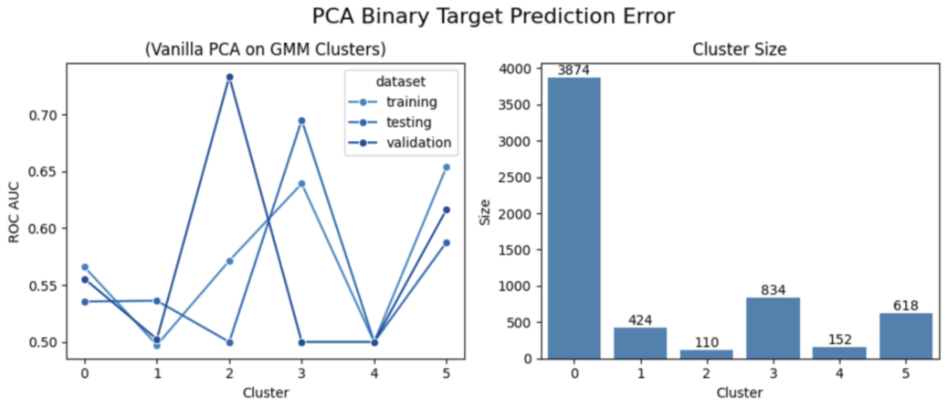


Figure 9. PCA binary target prediction error.

- Unsurprisingly, the smallest cluster (2) has the best predictive power, but it is also the smallest cluster. Interestingly, cluster 1, which is over eight times smaller than cluster 0, has similar predictive power.

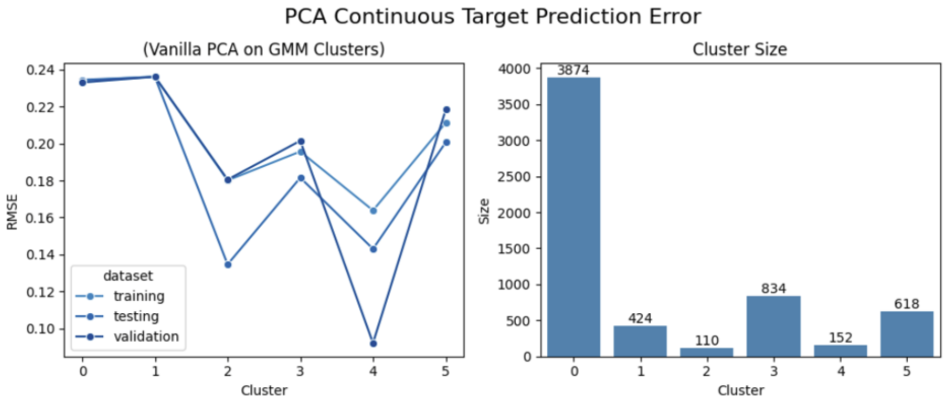


Figure 10. PCA continuous target prediction error.

- We also looked at the principal component covariances reported for each of the clustered schools to understand how each cluster looks. Interestingly, a distinguishing characteristic in these clusters is **whether the school is in a town.**

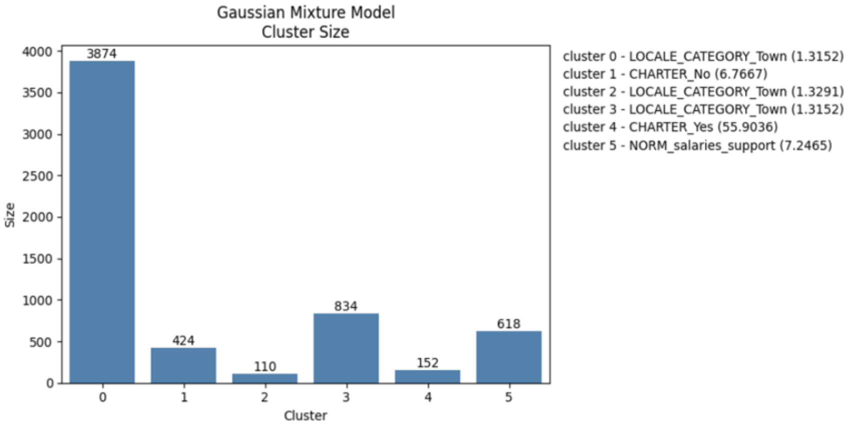


Figure 11. Gaussian mixture model cluster size. Feature that explains the most variance is shown on the right for each cluster.

Sensitivity analysis

We performed a sensitivity analysis on our best performing model, the kernel PCA with un-clustered data. For each feature, we dropped it from our data, re-fit the PCA with data-minus-a-feature, and calculated the root mean square error. The training data is most to the amount of salary spent on instructional aides and support staff per student. A hypothesis from this that is worth further exploring is if this implies that student performance may be impacted by the adults that are available to support them. Surprisingly, features that have negligible impact include the salary spent on teachers per student and student teacher ratio.

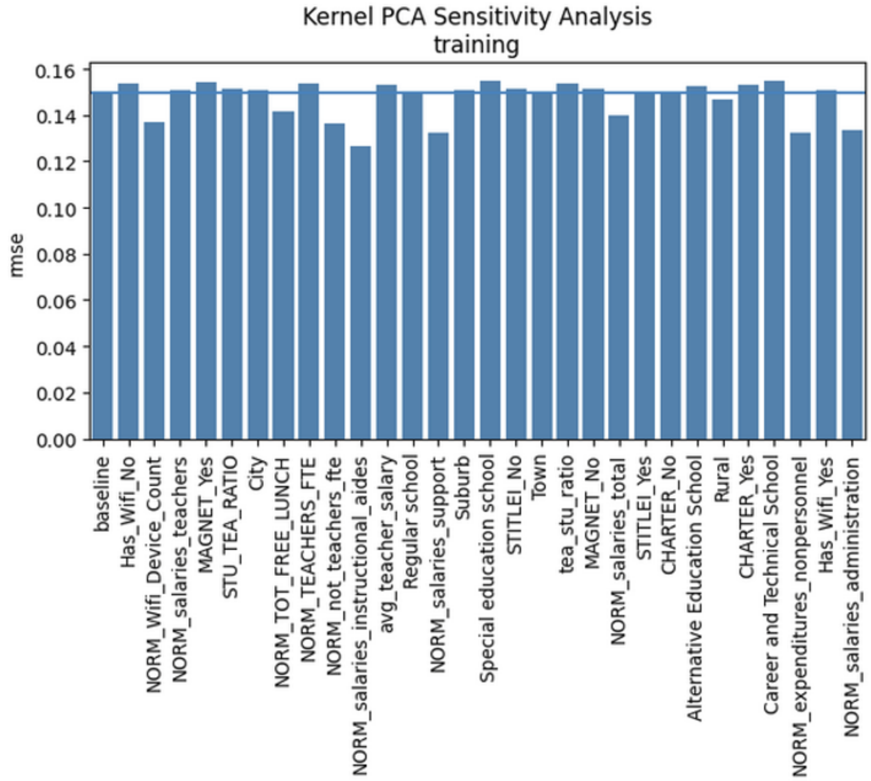


Figure 12. Kernel PCA sensitivity analysis result.

Clustered school performance

We examined how each cluster performs on specific target features. We looked at the average number of AP courses offered, students enrolled per cluster, as well as the percentage of schools in each cluster that offer AP courses. Interestingly, cluster 0 and cluster 3 have similar actual performance despite being quite different. Cluster 0 is over four times larger than cluster 3. Cluster 3 is more likely to have magnet and charter schools than cluster 0. In fact, cluster 0 does not have any magnet or charter schools, compared to cluster 3 which has 30% charter schools and 20% magnet schools.

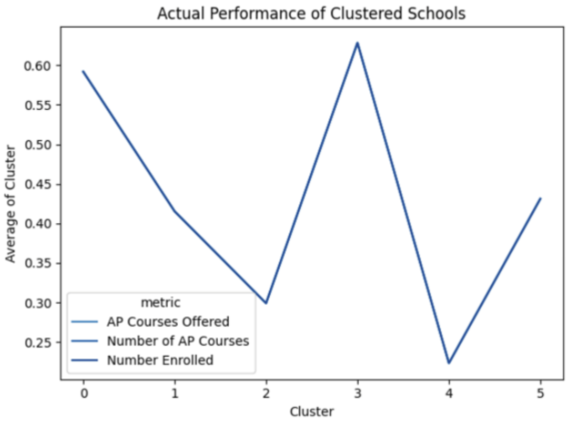


Figure 13. Actual performance of clustered schools.

Discussion

Supervised learning

It surprised us that our models are biased towards certain demographic groups. While we learned in the courses that models can be biased towards certain groups based on gender, race, and other features, it is still a bit surprising to see the results in the model developed by ourselves. One of the key challenges we encountered, as we also briefly reported in the introduction setting, is the messiness of the raw data. There are many missing values and incorrect data entries in the data that each school reported to the U.S. Department of Education. As we have learned, garbage in, garbage out. To ensure high model performance and fairness in model performance, it is important to check for data quality before building any model. In our project, we were able to detect and handle these challenges by computing descriptive data, visualizing the data distribution, and conducting EDA. If time allows and if additional data is available, we would be interested in examining other input features, such as students' family factors (e.g., their parents' educational level, etc.).

Unsupervised learning

We were initially surprised that there were four times more high schools in towns and rural areas (cluster 0) then there were in cities (cluster 3). We realized this was due to our own personal biases as city people ourselves. It was also surprising that the location of a school did not have a significant impact, after the data was normalized for student enrollment. Additionally, we had expected more impact from the categorical features of a school, but it turned out that continuous features have a larger impact on our analysis.

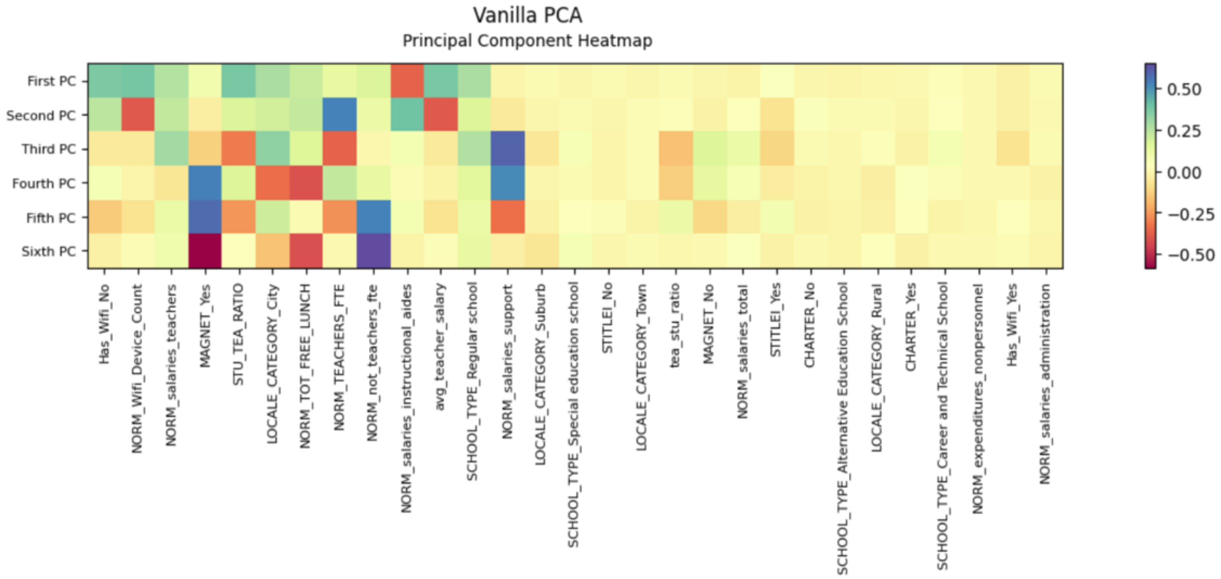


Figure 14. Heatmap of the vanilla PCA.

In addition to the same data quality issue discussed above, when developing our GMM, we initially had a fourth scoring method for choosing a model. Given that silhouette scoring is well suited for clusters, we thought that this would be an interesting metric to use for generating the model. As it turned out, this caused issues that we could not resolve. Every other run would result in grid search returning parameters that were scorable, in between these runs would be the grid search returning NaN. As a future direction, we would like to do Kernel PCA with the clustered data, as well as using the results to engineer additional features. Also, in our project we only examined AP enrollment data. If possible, we would hope to examine students' AP course performance data as well.

Ethical Considerations

For both supervised and unsupervised learning, if one did not examine and evaluate the data quality carefully before building the model, it could lead to a very biased and poor performing model. When the model output is used to make high-stake educational decisions such as policies related to educational resources allocated, it can impact students' opportunities to receive high quality education. Therefore, as data scientists, we need to always do a data quality check as the first step. Also, we need to be transparent and explicit when presenting modeling results to any stakeholders, making sure that they know how to interpret and act on the data with caution.

For unsupervised learning, in addition to the aforementioned ethical considerations, we must keep in mind that as data scientists, we never should make any assumptions based on our own personal and prior educational experiences (e.g., making assumptions about rural, town, and city schools). As we learned in the data science courses, data scientists need to keep in mind that their personal assumptions might introduce unintentional bias into the model development process.

Statement of Work

Below are the parts that each of us **led**. For parts that we did not lead, we reviewed each other's work and provided **feedback and input**.

- Sophia: Data cleaning (AP data), EDA (AP data), unsupervised modeling, project report (unsupervised part), github repository management, project coordination.
- Meixun: Data cleaning (school features data), data merging (school features, math, AP data), EDA/feature engineering (on merged datasets), project proposal (drafting), project report (initial draft, editing, and finalizing the report), domain expertise in "education".
- Chih-pin: Data cleaning (math data), EDA (data), supervised modeling, project report (supervised part).

References

Antoniou, F., Alghamdi, M. H., & Kawai, K. (2024). The effect of school size and class size on school preparedness. *Frontiers in psychology*, *15*, 1354072.

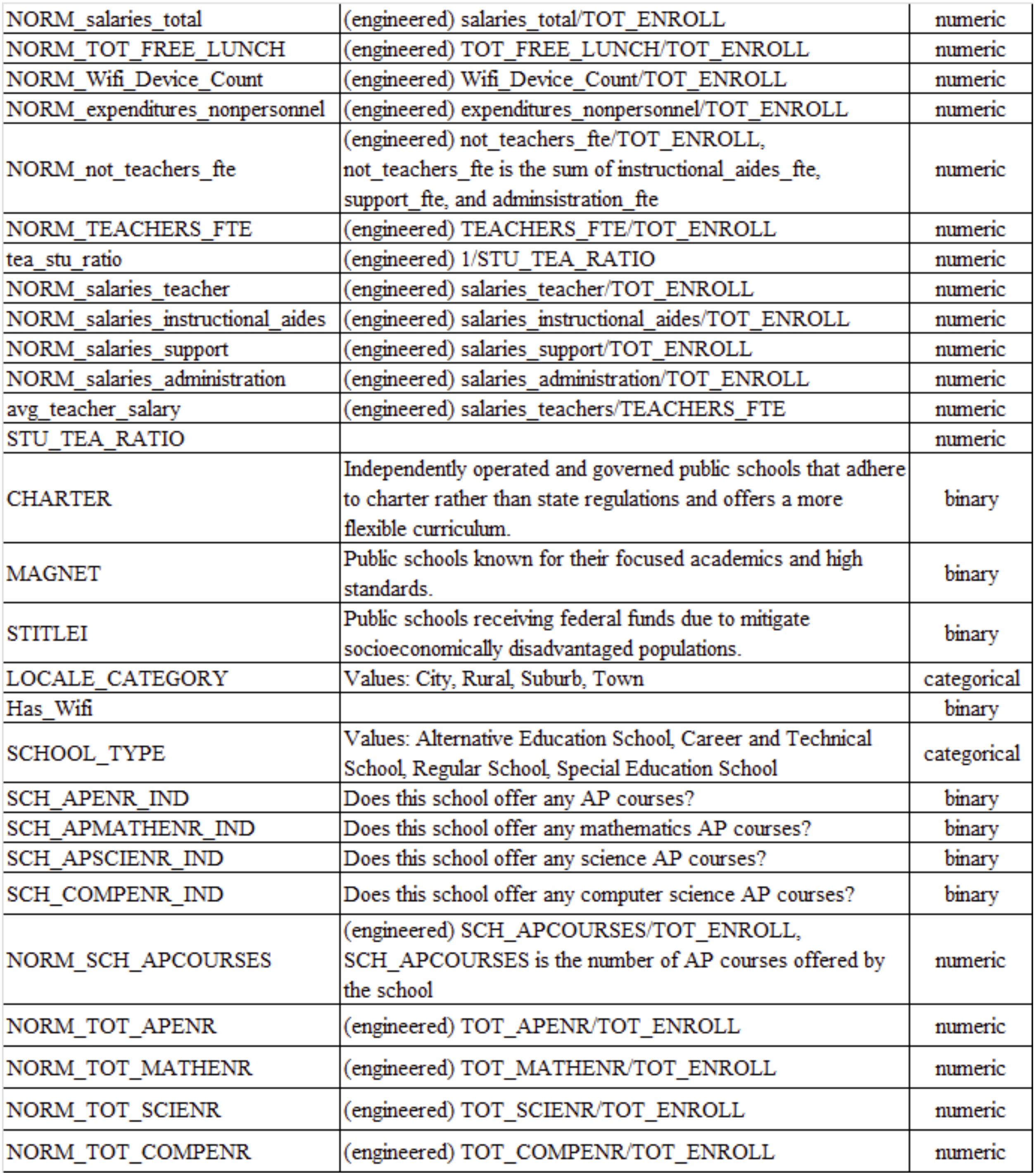
Loeb, S., & Page, M. E. (2000). Examining the link between teacher wages and student outcomes: The importance of alternative labor market opportunities and non-pecuniary variation. Review of Economics and Statistics, 82(3), 393-408.

Warne, R. T. (2017). Research on the Academic Benefits of the Advanced Placement Program: Taking Stock and Looking Forward. *Sage Open*, *7*(1). https://doi.org/10.1177/2158244016682996

Appendix A: Supervised Learning Features

Variables	Meaning	Data type
NCES_SCH_ID	Unique school ID	string
STATE_NAME_ABR	Abbreviated state name	categorical
CHARTER	Whether the school is a charter school or not	categorical
MAGNET	Whether the school is a magnet school or not	categorical
STITLEI	Whether the school is a title 1 school or not	categorical
STU_TEA_RATIO	Student teacher ratio	numeric
LOCALE_CATEGORY	Whether school is a rural, suburban, town, or city school	categorical
avg_teacher_salary	Average teacher salary	numeric
FREE_LUNCH_percent	Percent of students receiving free lunch	numeric
female_enroll_percent	Percent of female students	numeric
male_enroll_percent	Percent of male students	numeric
AM_enroll_percent	Percent of American Indian or Alaska Native students	numeric
AS_enroll_percent	Percent of Asian students	numeric
BL_enroll_percent	Percent of black students	numeric
HP_enroll_percent	Percent of Hispanic students	numeric
HI_enroll_percent	Percent of Hawaiian students	numeric
TR_enroll_percent	Percent of students who are two more races (mixed race)	numeric
WH_enroll_percent	Percent of white students	numeric
wifi_device_per_student	Average number of WIFI device per student	numeric
math_test_pct_prof_midpt	Percent of students meeting grade-level proficiency in math	numeric
disability_Yes_IDEA	percent of students who has disability	numeric
econ_disadvantaged_Yes	percent of students from economically disadvantaged families	numeric
Immigrants_Yes	Percent of students who are immigrants	numeric
military_connected_Yes	Percent of students who relate to military	numeric
homeless_Yes	percent of students who are homeless	numeric
foster_care_Yes	percent of students in foster care	numeric
lep_Yes	percent of students with limited English proficiency	numeric

Appendix B: Unsupervised Learning Features



NORM_salaries_total	(engineered) salaries_total/TOT_ENROLL	numeric
NORM_TOT_FREE_LUNCH	(engineered) TOT_FREE_LUNCH/TOT_ENROLL	numeric
NORM_Wifi_Device_Count	(engineered) Wifi_Device_Count/TOT_ENROLL	numeric
NORM_expenditures_nonpersonnel	(engineered) expenditures_nonpersonnel/TOT_ENROLL	numeric
NORM_not_teachers_fte	(engineered) not_teachers_fte/TOT_ENROLL, not_teachers_fte is the sum of instructional_aides_fte, support_fte, and adminsistration_fte	numeric
NORM_TEACHERS_FTE	(engineered) TEACHERS_FTE/TOT_ENROLL	numeric
tea_stu_ratio	(engineered) 1/STU_TEA_RATIO	numeric
NORM_salaries_teacher	(engineered) salaries_teacher/TOT_ENROLL	numeric
NORM_salaries_instructional_aides	(engineered) salaries_instructional_aides/TOT_ENROLL	numeric
NORM_salaries_support	(engineered) salaries_support/TOT_ENROLL	numeric
NORM_salaries_administration	(engineered) salaries_administration/TOT_ENROLL	numeric
avg_teacher_salary	(engineered) salaries_teachers/TEACHERS_FTE	numeric
STU_TEA_RATIO		numeric
CHARTER	Independently operated and governed public schools that adhere to charter rather than state regulations and offers a more flexible curriculum.	binary
MAGNET	Public schools known for their focused academics and high standards.	binary
STITLEI	Public schools receiving federal funds due to mitigate socioeconomically disadvantaged populations.	binary
LOCALE_CATEGORY	Values: City, Rural, Suburb, Town	categorical
Has_Wifi		binary
SCHOOL_TYPE	Values: Alternative Education School, Career and Technical School, Regular School, Special Education School	categorical
SCH_APENR_IND	Does this school offer any AP courses?	binary
SCH_APMATHENR_IND	Does this school offer any mathematics AP courses?	binary
SCH_APSCIENR_IND	Does this school offer any science AP courses?	binary
SCH_COMPENR_IND	Does this school offer any computer science AP courses?	binary
NORM_SCH_APCOURSES	(engineered) SCH_APCOURSES/TOT_ENROLL, SCH_APCOURSES is the number of AP courses offered by the school	numeric
NORM_TOT_APENR	(engineered) TOT_APENR/TOT_ENROLL	numeric
NORM_TOT_MATHENR	(engineered) TOT_MATHENR/TOT_ENROLL	numeric
NORM_TOT_SCIENR	(engineered) TOT_SCIENR/TOT_ENROLL	numeric
NORM_TOT_COMPENR	(engineered) TOT_COMPENR/TOT_ENROLL	numeric