AP Outcomes to University Metrics

Erdös Data Science Bootcamp Cohort of Fall 2024

Group Members: Prabhat Devkota, Shrabana Hazra, Jung-Tsung Li, Shannon J. McElhenney, Raymond Tana

GitHub: https://github.com/mcelhens/AP-Outcomes-to-University-Metrics

Streamlit app: https://ap-outcomes.streamlit.app/

Project Overview

This project was designed to investigate the potential relationship between **AP exam performance** and the **presence of nearby universities**. It was initially hypothesized that local (especially R1/R2 or public) universities would contribute to better pass rates for AP exams in their vicinities as a result of their various outreach, dual-enrollment, tutoring, and similar programs for high schoolers. We produced a predictive model that uses a few features related to university presence, personal income, and population to predict AP exam performance and report it in a streamlit app.

Background

AP exams are standardized tests widely available at high schools across the United States. During the 2022-2023 school year, 79% of all public high school students attended schools offering at least five AP courses. These exams are popular for their potential to earn college credits during high school by achieving high scores. In fact, high scores in most AP exams are eligible to receive college credits at roughly 2000 higher-education institutions.

AP exams are scored on a whole number scale between 1 (lowest) and 5 (highest). A student is said to *pass* their AP exam if they score a 3 or higher on the exam. The *pass rate* of a locality would be the proportion of AP exams passed out of all exams taken by its students during a single year. AP outcomes are often correlated to measures of socioeconomic factors: a recent study confirmed that negative socioeconomic factors have a strong negative influence on exam scores; as well as being a non-native English language speaker.

Beyond these socioeconomic factors, we wished to measure the strength of the effect of universities on AP outcomes. Without a clear source of data on all high school outreach programs offered by US universities, we made use of the various classifications offered by the Carnegie Classifications of Institutions of Higher Education. Of particular interest included R1 and R2 (i.e., doctoral with very high or high research activity, respectively), public, or private institutions. Other minority-serving aspects were also considered, such as historically Black, Hispanic-serving, and tribal colleges.

Key Questions

1. How do universities influence local high schoolers' standardized test performance?

2. What benefits can universities offer beyond socio-economic limitations?

Goals

1. Determine whether proximity to universities has the potential to overcome socio-economic obstacles.
2. Uncover possible opportunities for educational equity through university outreach.
3. Provide a tool to predict AP performance in an area.

Stakeholders

- **Universities**: looking for educational equity opportunities.
- **State officials**: strategic planning for improving testing results.
- **Parents**: deciding where their children should live and learn.

Data Search, Cleaning, and Processing

Our university metrics data was obtained from the Carnegie Classifications of Institutions of Higher Education. Our population and income datasets were obtained from the United States Census Bureau, Federal Reserve Bank of St. Louis, and the United States Department of Commerce easily.

We initially contacted the College Board for details on what public datasets were available from them regarding AP performance. Unfortunately, these datasets were highly limited. They make only about 4 years of AP performance, availability, and participation available. Subsequently, we split the states across the team to hunt for county/district-level data and found 4 states with passing score outcomes (3+/5 scores) to develop state- and combined-models. We used the 5 most recent years of each state's data for our investigation.

To resolve issues of counties/districts without universities, we used a "5 nearest" method. We used geopy to calculate the 5 nearest universities to a county/district then averaged those 5 into each university feature.

Modelling Approach

We followed a consistent approach for our models substituting the datasets for each state and the combined model. The only exceptions were to validate the combined model on an unanalyzed year of AP performance from Wisconsin (2017-18), to maintain a categorical feature of state, and to robustly apply XGBoost with PCA and hyper parameterization.

Features

We had 17 predictive features for our output feature "AP Exam Pass Rate". 2 features were university independent.

1. Population

2. Per capita Income

The other 15 features are 3 different "nearest 5" averages applied to 5 categories of universities. The 3 "nearest 5" average metrics were

1. Distance
2. Enrollment
3. Dorm rooms

The 5 categories of universities were

1. R1/R2 research
    a. These institutions represent the highest research spending and output high numbers of PhDs
2. Public
3. Private non-profit
4. Land-grant
    a. Land-grant universities are a subcategory of public universities
5. STEM-specialized
    a. These institutions issue high rates of degrees in STEM

Fitted Models

We fitted several models to our datasets and selected the optimal models individually. We used a test/train split of 20%. We also use Shapley (SHAP) to analyze the key features of these models. We fit a "take the average and call it a day" baseline, a full feature set OLS regression, a ridge regression, an adaboost model, a random forest model, and an XGBoost model. To assess performance, we used RMSE and R^2-scores from 5-fold cross validation and R^2 on the selected model with test data.

Individual State Model Results

On the individual state level, XGBoost and Random Forest models swap back and forth as the optimal model.

Wisconsin

Wisconsin performed best with Random Forest. Outside of income and population, distance to public and private non-profit institutions, and average number of STEM-institution dorms had the largest SHAP impact.

Massachusetts

XGBoost performed best for Massachusetts. Outside of income and population, distance to public and private non-profit institutions, and average enrollment of land grand institutions had the largest SHAP impact.

Georgia

Georgia performed best with XGBoost. Georgia is notable for having a less severe dependence on per capita income and population though they remain the largest influencing features. Land-grant universities also had a large impact from both dorms and distance.

North Carolina

North Carolina performed best with Random Forest.

4-State Model Results

On our 4-state model hyperparameter-tuned XGBoost performs best on the training data (RMSE 10.54, R^2 0.70). On the testing data, the R^2 remains similarly high (0.77). This suggests that our model trained well. To test generalizability, we further tested our model on a previous year of Wisconsin data where the model predicts with an average accuracy of ±8.4% and an R^2 of 0.56. On the 4-state model, the most extreme dependence on income can be seen in the SHAP calculation followed by population, then distance to land-grant, public, or private non-profit universities. Overall, our model performs well on the states where it was trained and the variance within state models does not seem to limit the function of the 4-state model.

Future Work

1. Several improvements for this work can be achieved by improving the AP performance dataset.
    a. More states would improve generalizability across the United States.
    b. Utilizing additional years of scores
    c. Finer dataset resolution to district (or school)
    d. AP performance by exam
2. The project would also be improved by non-performance features and feature optimization.
    a. Including other university metrics like minority-serving or undergraduate-focused institutions
    b. Advanced feature analysis tools (ex. UMAP) would improve the understanding of individual feature impacts versus in combination.