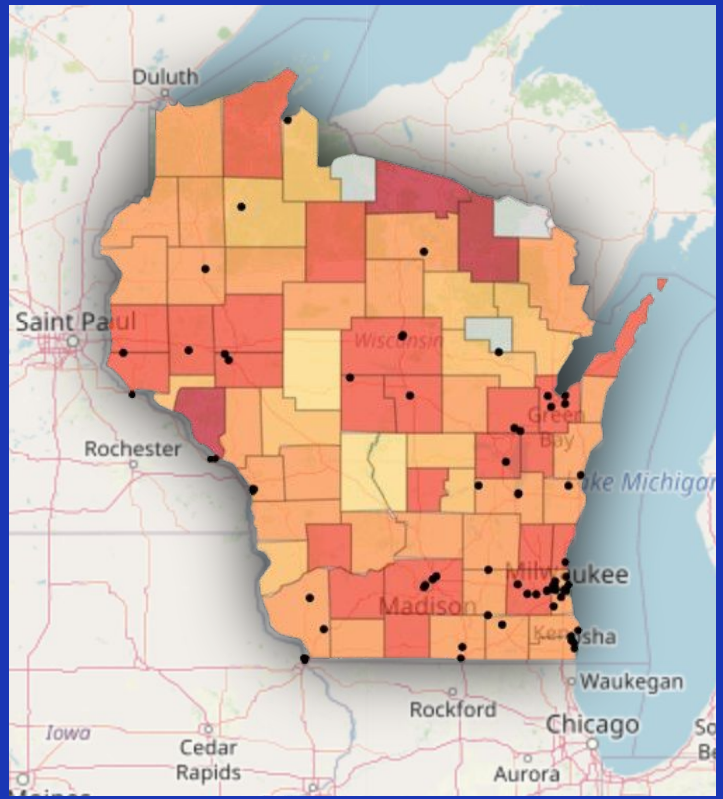
ERDŐS PROGRAM 2024

Universities and AP Scores

Universities for Educational Equity
2 December 2024

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



Motivation

Key Questions:

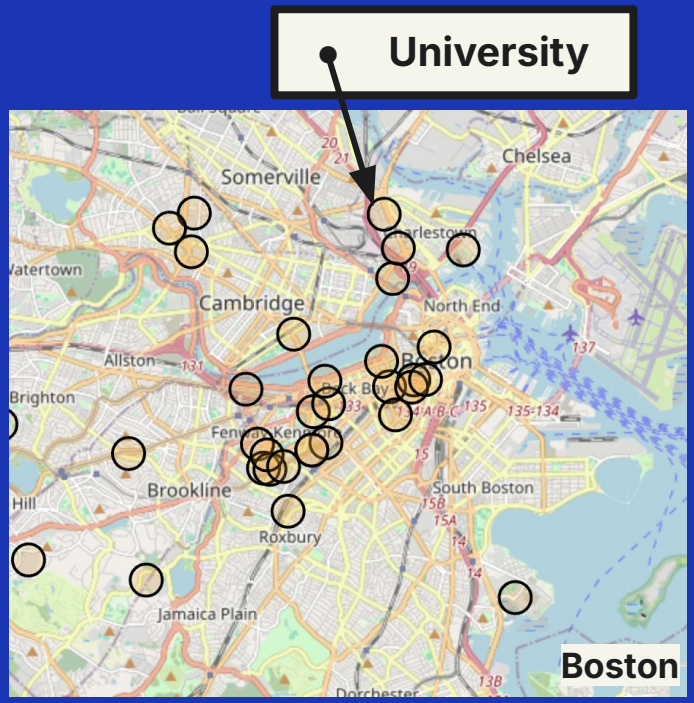
- How do universities influence local high schoolers' standardized test performance?
- What benefits can universities offer beyond socio-economic limitations?

Goal:

- Determine whether proximity to universities has the potential to overcome socio-economic obstacles.
- Uncover possible opportunities for educational equity through university outreach.
- 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.



Datasets

Carnegie classifies all US universities:

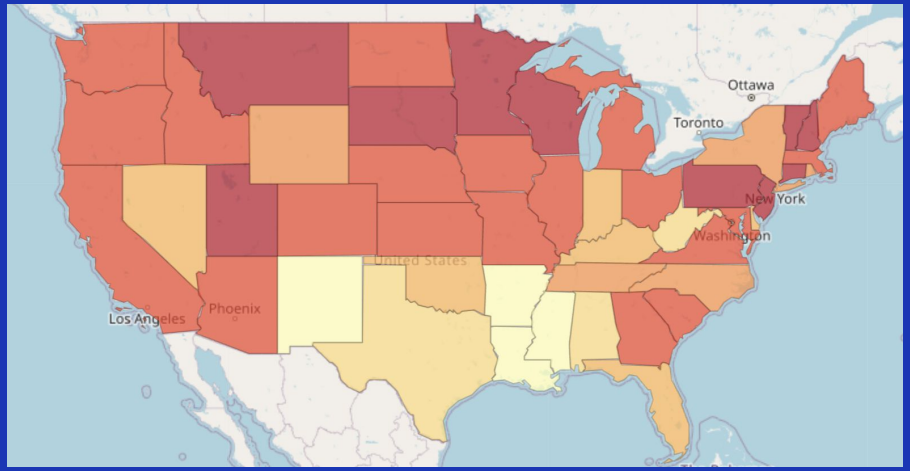
- Research level
- Public vs. Private
- Minority-serving
- Land grant status

CollegeBoard provides in-depth AP exam data *only at the state level*

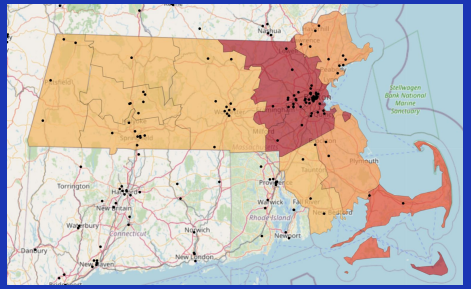
A few states self-report AP performance by county or school-district:

- Massachusetts
- Wisconsin
- Georgia
- North Carolina

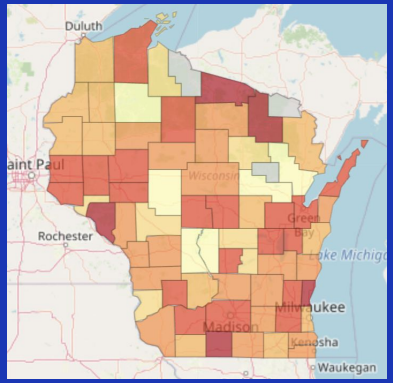
US Census Bureau, **Federal Reserve Bank of St. Louis**, and **US Department of Commerce** provide other features for localities



CollegeBoard: National AP Data



Carnegie & FRBSL in Massachusetts

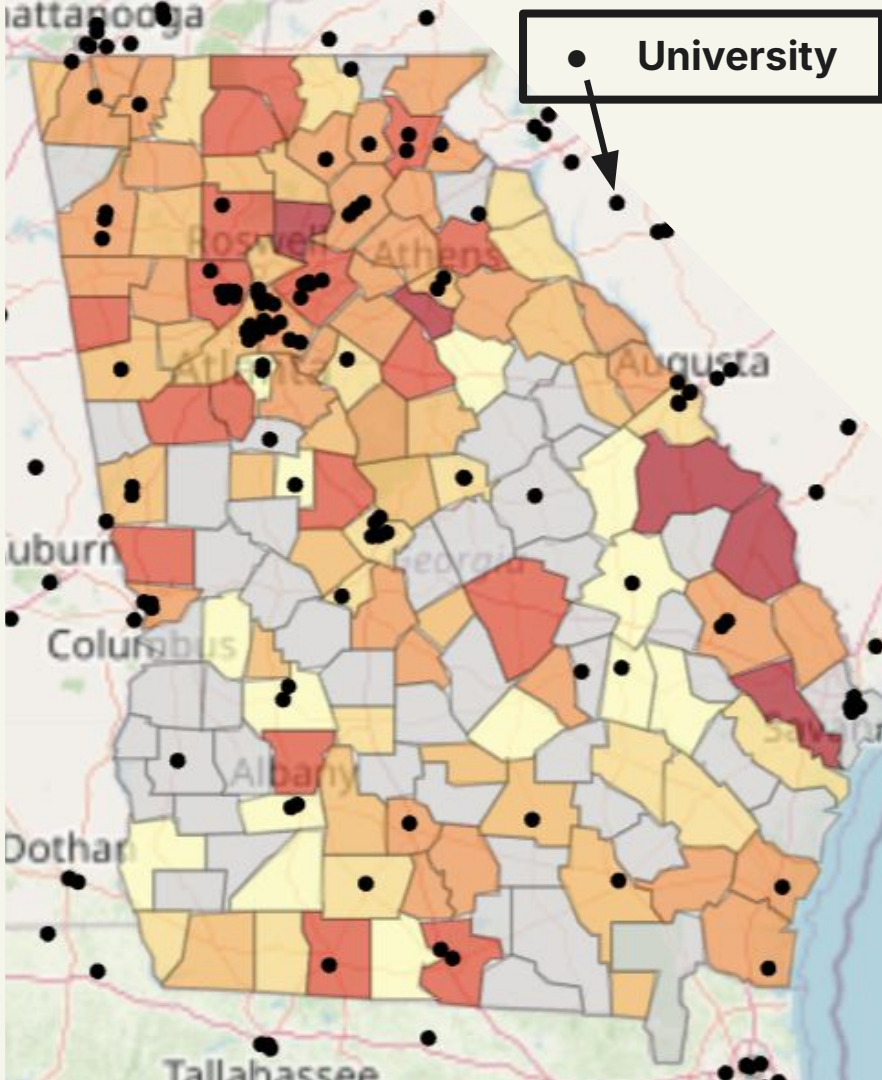


Wisconsin AP Data

Data Processing

We collect various features about counties or school districts:

- Population
- Per-capita income
- Categorized universities
  1. R1/R2 research
  2. Public
  3. Private non-profit
  4. Land-grant
  5. STEM-specialized
- “Closest 5”
  - Many counties have no universities
  - Closest 5 universities eliminates 0-regions
  1. Distance
  2. Enrollment
  3. Dorm-Rooms
- Selection of key features
  - 17 features needs interpretability (**SHAP**)



Models

Multiple scopes to model:

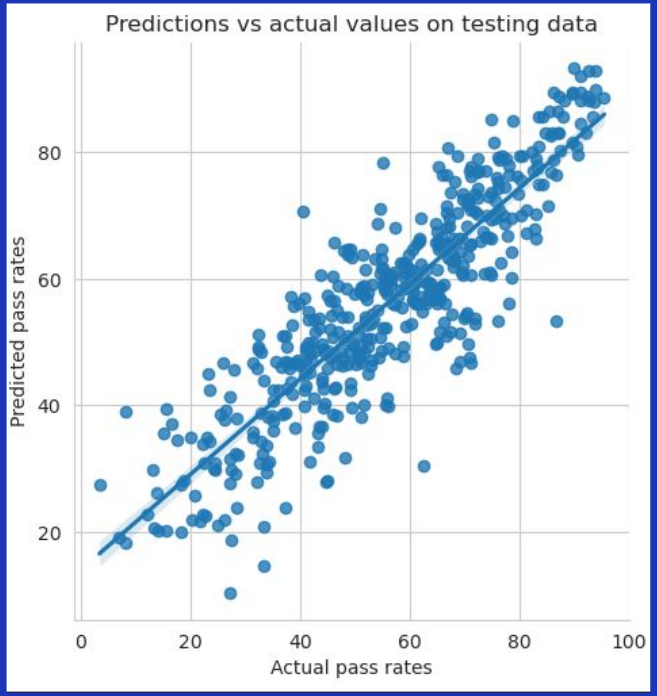
- Separate models on **each state**'s counties/school districts.
- Combined model on **four states**' data.

Create and compare various machine learning models from `sklearn` and `xgboost`.

XGBoost performs the best:

- Use PCA to reduce the number of features.
- Hyperparameter tuning.

Final model: XGBoost model after PCA(0.95).



Predictions on Test Set

The final model had **RMSE = 9.23** and **R^2-score = 0.77** on the testing data.

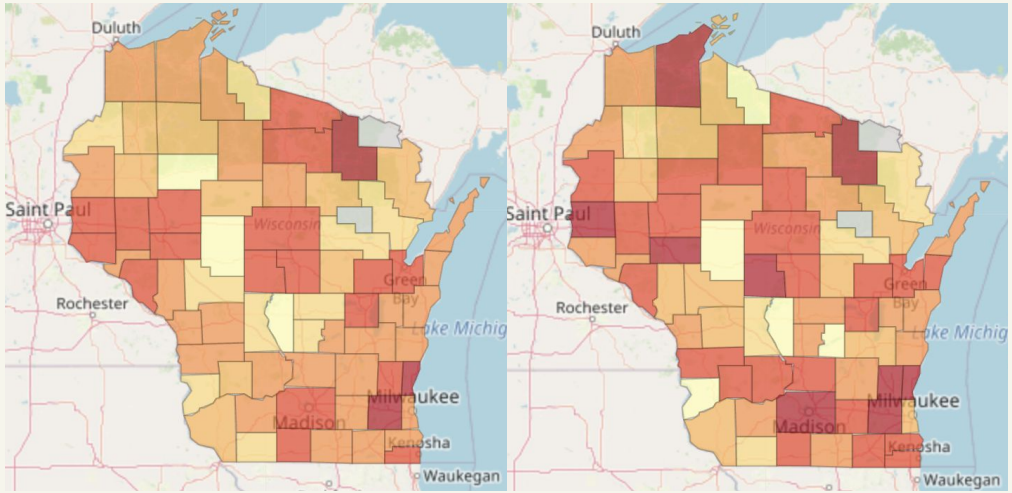
Predicting a new Year

Running our *combined model* (trained on four states' AP performance data over five years), **we predict AP performance in Wisconsin during 2017-2018**. The model was not trained on any data from this year.

Predictions were fairly accurate:

- **RMSE**: 8.392
- **R^2**: 0.562

The combined model can at least give decent predictions for other years in the states on which it was trained.



Predicted vs. **Reality**

Wisconsin 2017-2018 AP Pass Rate

Results

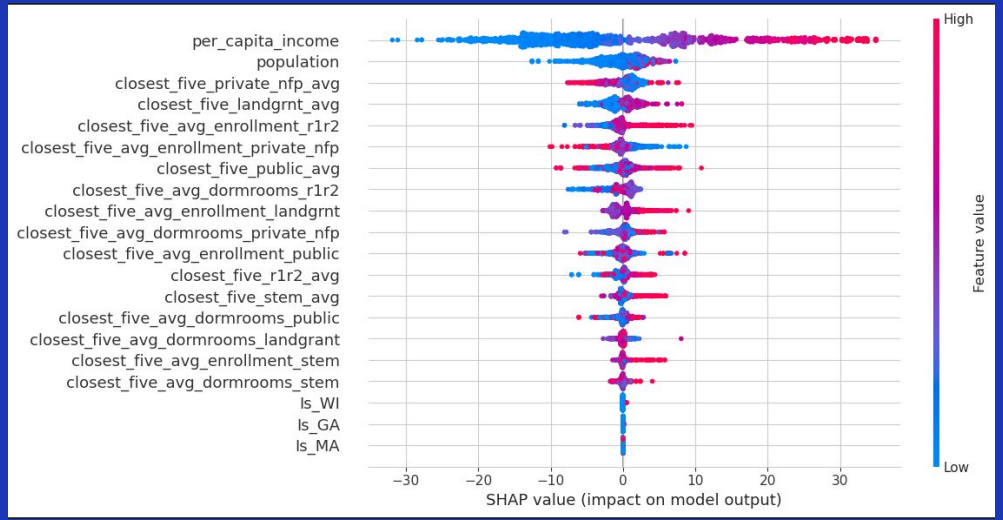
1. *Family income* is the most important feature for obtaining higher AP pass rate.

2. SHAP values identify important features for boosting the AP pass rate:
   a. Living in *high-population county*
   b. Living *close to private universities*
   c. Living *close to R1/R2 universities with high enrollment*

3. SHAP values indicate that distances to STEM universities, their enrollment, and the number of dorm beds are less important features.

4. We conclude that living closer to universities can overcome social-economical barriers.

MA, WI, GA, NC combined result



Future Directions

1	**More data**	Having data for more states will be helpful to improving our models.
2	**Finer resolution**	Ideally, we would want to work with district level data (or even school by school).
3	**Advanced tools**	Using tools like UMAP to identify key feature correlations and improve the model's predictive accuracy.

Streamlit App

We have made our model interactive and offered more analysis in a Streamlit application.

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

Acknowledgements

Thank you to Steven Gubkin, Roman Holowinsky, and Alec Clott at the Erdős Institute for their support throughout the Fall 2024 Data Science Bootcamp.

Thank you to Gleb Zhelezov for his insight and mentorship throughout the project.

ERDŐS PROGRAM 2024

DATA SCIENCE BOOTCAMP