Higher Education Synthesized

Extending Higher Education Data with Synthetic Data at the Onset of the Demographic Cliff

Problem Statement — Introduction/Motivation

Data Mining + Diminishing Data = No Good

- After decades of enrollment increases, Higher education is now facing decreasing enrollments, either gradually over time or with a sharp drop off a demographic cliff, depending on who you ask
- One corollary to declining enrollments is declining amounts of data (might rearrange to put a visual showing increase and decrease over time to match narrative)
- This project will explore setting up data mining pipelines with IPEDS institution-level data, and, to handle less data going forward, we will explore employing generative modeling to create tabular synthetic data based on the IPEDS subset we are working with (separate slide clarifying synthetic data with example)
- This iteration of the project will focus on institution-level, not student-level, data mining using Integrated Postsecondary Education Data System (IPEDS) data from the National Center for Education Statistics (NCES), a well-known and reliable set of higher education data overseen by the federal Department of Education (once we decide later, put more details on what IPEDS data we end up using)

Synthetic Data Example (visual slide)

Image of real tabular data

Image of synthetic tabular data

Related Work

- Not a lot directly related to the focus of this project
- Did not find much running synthetic data for higher education or IPEDS data
- Did not find much on data mining on higher education data
- Found some about data mining on education data generally, focusing on different predictive approaches in particular
- Did find research speaking to the importance of bringer more advanced data methods to higher education data
- Found research highlighting the lack of big data practices in higher education
- Subsection of research focused on synthetic data generally

Proposed Work

(Turn this into a visual once we have more concrete details)

- Scrape IPEDS data
- Relational database in Docker and leveraging database to mimic real-world scenario where data will live in a database or some sort of data store to start with
- Python scripts for automation with Jupyter Notebooks for experimentation and prototyping
- Can use database for preprocessing through triggers
- The read-eval-print-loop (REPL) functionality of Jupyter Notebooks helps for building up to automated Python scripts
- We do not want notebooks to play much, if any, role in the final pipeline, with the exception of where end users of the pipeline may work for their own analysis and visualization
- Database is already available for warehousing
- Synthetic Data Vault Project for synthetic data libraries to synthesize tabular IPEDS data
- Some amount of likely regression or classification, though clustering seems like it may be a good final step for modeling

Data Pipeline (visual to match text slide)

Series of images with arrows to help visualize the flow of data and key processes

Evaluation

- Two sets of evaluation: synthetic data and modeling components that run against the synthetic data
- Synthetic data evaluation focuses on how similar the synthetic data is to the real data
  - Per-attribute similarity: are measures like the mean and standard deviation for a given feature similar across the real and synthetic data
  - Machine learning score: how similar is the performance when we train classifiers on the real and synthetic data and test them on a held-out test set
  - Possibly look at others in the Synthetic Data Vault Project's SDGym library
- The second group of evaluations are the standard ones we use with modeling
  - Accuracy, precision, recall, F1 score for classification, and mean square error (MSE) or adjusted r-squared for regression
- On a project level, a successful project will look like one where have sensible evaluations for synthetic data and a demo for modeling on the synthetic data

Evaluation Summary (visuals)

Synthetic Data Metrics

Modeling Metrics

Timeline and Future Work

- 3-4 week limit for working on this project, if possible — I have a habit of adding too much scope, so a shorter timeline can help to contain that
- Biggest risk is that the data is much more complex to work with than expected
- Focus on this iteration is synthetic data, so next iterations can include more domains of IPEDS data and wider range of analytical and modeling experimentation in next iteration of project
- Maybe add a table to summarize possible directions in which future work can go