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 of a demographic cliff, depending on who you ask
- One corollary to declining enrollments is declining amounts of data
- This project will explore synthesizing institution-level 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
- This iteration of the project will use generative models in the Synthetic Data Vault Project (https://github.com/sdv-dev) to create tabular synthetic IPEDS data and evaluate the quality of that data
- Source code at https://github.com/jhleakakos/msds-data-mining-project

IPEDS Data

Domains of Data

Institutional Characteristics: information related to higher education institutions

12-Month Enrollment: unduplicated counts of students who enroll anytime during a 12-month period

Completions: counts of students who complete degree or non-degree credentials

Difficulties

- IPEDS data dictionaries are not clear about mapping categorical feature numeric values to definitions
- IPEDS data dictionaries are not clear about how to group across categorical features to get the correct summary counts
- Number of features per data file — some files have hundreds of features
- Inconsistencies between encodings (binary fields represented as different pairs of numbers), even with the same files

Synthetic Data

- Synthetic data is “artificially generated [data] that resemble[s] the actual data — more precisely, having similar statistical properties” (*)

- Synthetic data allows us to create more IPEDS data to counter the reduction in real higher education data, either to replace or supplement that real data

Real Data

state_abbr	bea_region	is_hbcu	control_affiliation	enrollment	completions…
AR	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private for-profit	52	20
WA	Far West AK CA HI NV OR WA	n	Private not-for-profit indepe…	1560	355
CA	Far West AK CA HI NV OR WA	n	Private for-profit	2278	829
GA	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	y	Private not-for-profit religi…	342	22
NY	Mid East DE DC MD NJ NY PA	n	Public	19441	1382

Synthetic Data

state_abbr	bea_region	is_hbcu	control_affiliation	enrollment	completions…
IA	Mid East DE DC MD NJ NY PA	n	Private not-for-profit indepe…	487	36
PR	Plains IA KS MN MO NE ND SD	n	Private not-for-profit religi…	740	78
WY	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	y	Private not-for-profit religi…	3115	0
MN	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private for-profit	561	162
PR	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private not-for-profit religi…	1534	136

* Emiliano De Cristofaro. 2024. Synthetic Data: Methods, Use Cases, and Risks. arXiv:2303.01230 [cs.CR] https://arxiv.org/abs/2303.01230

Related Work

We're On Our Own

- Did not find much research directly related to data mining or generative modeling for higher education or IPEDS data
- Found some research about data mining on education data generally, particularly focusing on different predictive approaches
- Found 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
- Found a subsection of research focused on general use of synthetic data
- In summary, no research that covers each of the aspects of this project

Proposed Work

1. Scrape IPEDS data from the IPEDS website using Python scripts, storing flat files on disk

2. Read through IPEDS data dictionaries to determine a set of fields that we want to explore for possible modeling

3. Run a local PostgreSQL database in Docker, using schemas to separate out data lake and data warehouse

4. Design, create, and load data lake with SQL scripts based on findings in data dictionaries

5. Pull data out of data lake and into Jupyter Notebook for an initial pass of EDA

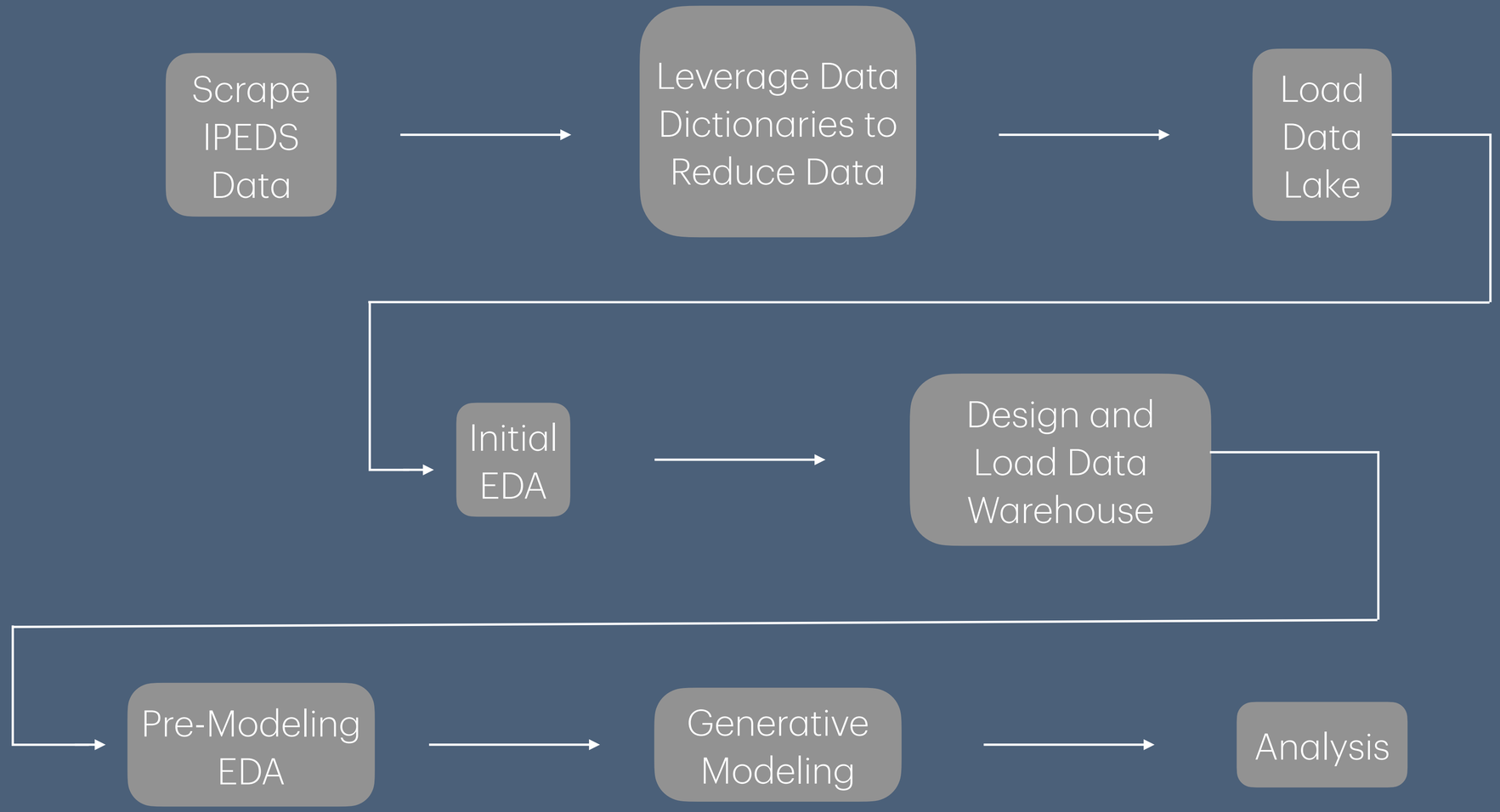
6. Use EDA to inform design of data warehouse, using a very small-scale version of a dimensional model to mimic a larger warehouse concept

7. Pull data out of data warehouse and into Jupyter notebook for generative modeling, using Synthetic Data Vault Project libraries to generate synthetic IPEDS data

8. Evaluate synthetic data quality

9. Load real and synthetic data into a separate Jupyter Notebook in order to perform deeper modeling and run analysis on real data, synthetic data, and a combination of both

Data Pipeline



Evaluation

Is Our Fake Data Any Good?

Two Sets of Evaluation

1. Evaluating the quality of the synthetic data to determine how similar the synthetic data is to the real data
    - Are measures like the mean and standard deviation for a given feature similar across the real and synthetic data
    - Do synthetic columns have the same distribution shapes as their associated real columns
    - Do trends between columns in the real data show the same trends in the synthetic data
    - Diagnostics that check that the synthetic data is valid, meaning it has the right types of data in the right ranges and more
    - How similar is the performance of classifiers or regressors when trained on the real and synthetic data and tested on a held-out test set
2. Evaluating common modeling outputs
    - Accuracy, precision, recall, or F1 score for classification, and mean square error 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 Metrics (visuals)

We will fill this in for the next iteration of the slides

Synthetic Data Metrics

(will fill in later)

Modeling Metrics

(will fill in later)

Timeline and Future Work

- I have a habit of adding too much scope, so I have limited my timeline to 3-4 weeks in order to work outside of my comfort zone and require that I find areas to descope
- I am currently around 3 weeks into the project and plan to finish within about a week, keeping to my agreed-upon timeline
- I have finished working on the pipeline up until the final analysis and modeling on top of the real and synthetic data
- The biggest challenge so far has been that the IPEDS data is much more complex to work with than expected
- The biggest remaining challenge is that the data I have pulled out for analysis may not answer the questions I am hoping it will answer
- The focus in this iteration is on creating synthetic IPEDS data and exploring its quality and suitability for handling reductions in real IPEDS data that are coming
- Next iterations of the work could include more domains of IPEDS data, a more robust data warehouse facilitating analysis on the wider range of IPEDS domains and their associated complexity, and introducing more modeling and analytical and modeling techniques, both for generative modeling and for asking important questions of the data