DSCI 510 Final Project Progress Report
Student: Jasmine Adams
Project: *E-Rate Market Opportunity Analysis: Identifying Underserved Schools and Libraries Using Multi-Source Data Integration*
Date: November 13, 2025

Project Scope Update

The project scope remains consistent with the proposal. I am developing a data-driven system to identify underserved schools and libraries for E-Rate consulting opportunities by integrating multiple public datasets. Key objectives include:

1. Predicting E-Rate application approval probability using a classification model
2. Performing geographic clustering (K-means) to locate underserved regions
3. Building a market opportunity scoring system

The project structure is modular, separating data collection, preprocessing, modeling, and visualization, to support both academic goals and future business use.

Data Sources

1. USAC E-Rate Open Data API

- **Endpoint:** https://opendata.usac.org/resource/avi8-svp9.json
- **Status:** Fully implemented via Python
- **Details:** Provides E-Rate funding data (entity name, funding year, discount rate, service type, etc.).
- **Implementation:** Created an ERateDataCollector class to automate pagination, error handling, and data export.
- **Progress:** Successfully fetched and validated California FY2024 data.

2. NCES Common Core of Data (Urban Institute API)

- **Status:** Next phase
- **Goal:** Collect school demographic and geographic data for model enrichment.

3. IMLS Public Libraries Survey

- **Status:** Next phase
- **Goal:** Add library operational data for library-focused opportunity analysis.

Issues / Difficulties

Resolved:

- **API rate limits**: Added batch fetching, pagination, and delay controls.
- **Project structure**: Organized modules for scalability and reuse.

Current Challenges:

- **Large data volume:** Managing ~1M records via incremental collection and potential Dask integration.
- **Data consistency:** Addressing missing or inconsistent fields across years.

Expected Challenges:

- Complex feature engineering (historical success rates, time trends)
- Merging datasets with differing identifiers and formats
- Balancing interpretability and accuracy during model selection

Next Steps

1. Complete NCES and IMLS data integration
2. Conduct EDA and feature engineering
3. Train baseline logistic regression model
4. Validate using 2024 holdout data

GitHub: github.com/princessmanifest/erate-prospector