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 aligned with the original proposal. I am developing a data-driven system to identify market opportunities for E-Rate consulting services by analyzing USAC funding patterns, school district demographics, and library operational data.

No major changes to scope. The project continues to focus on: 1. Building a machine learning classification model to predict E-Rate application approval probability 2. Performing geographic clustering (K-means) to identify underserved regions 3. Creating a market opportunity scoring system for business development prioritization

The project structure has been organized to support both academic requirements and future business productization, with clear separation between data collection, preprocessing, modeling, and visualization modules.

Data Sources

1. USAC E-Rate Recipient Details and Commitments API ✓

API Endpoint: `https://opendata.usac.org/resource/avi8-svp9.json`

Status: Successfully accessed via Python

Description: The USAC Open Data API provides comprehensive E-Rate funding data through a Socrata Open Data API (SODA) interface. The API supports: - Filtering by funding year, state, applicant type, and other parameters - Pagination for large datasets (50,000 record limit per request) - Optional app token for higher rate limits - JSON response format

Key Fields Retrieved: - `entity_name`: Name of the applicant (school/library) - `ben`: Billed Entity Number (unique identifier) - `funding_year`: E-Rate program year - `state`: State abbreviation - `zip_code`: Location - `total_commitment`: Approved funding amount - `discount_rate`: Percentage discount (based on NSLP) - `applicant_type`: School or Library - `service_type`: Category 1 (connectivity) or Category 2 (internal connections) - `frn_status`: Application status

(approved/denied/pending) - `nslp_percentage`: National School Lunch Program eligibility (poverty indicator) - `locale_code`: Urban/rural designation

Implementation Details: - Created `ERateDataCollector` class with methods for: - Single API calls with flexible filtering - Batch fetching across multiple years - Automatic pagination handling - Summary statistics calculation - Data persistence (CSV export) - Built-in error handling and logging - Rate limiting courtesy (0.5s delay between requests)

Sample Query Results: - Successfully fetched test data for California FY2024 - Confirmed data structure matches proposal specifications - Verified all key fields are present and accessible

2. NCES Common Core of Data (via Urban Institute) - Planned

API Endpoint:
`https://educationdata.urban.org/api/v1/schools/ccd/directory/2024/`

Status: Implementation pending (next phase)

Plan: Will collect school and district demographic data including enrollment, NSLP percentages, urbanicity classification, and geographic coordinates for merging with E-Rate application data.

3. IMLS Public Libraries Survey - Planned

Source: `https://www.imls.gov/sites/default/files/2024-06/fy2022_pls_data_file_csv.zip`

Status: Implementation pending (next phase)

Plan: Will download and parse CSV file containing public library operational data including service populations, revenue, expenses, and urban/rural designation for library-focused prospect identification.

Issues / Difficulties

Resolved Issues

1. **API Rate Limiting Considerations**
   - **Issue:** USAC API has rate limits that could impact large data collection
   - **Solution:** Implemented batch fetching with pagination and polite rate limiting (0.5s delays). Also added support for optional app tokens to increase limits if needed.
2. **Project Structure Organization**
   - **Issue:** Needed to balance academic requirements with future business use
   - **Solution:** Created modular structure with separate directories for data collection, preprocessing, modeling, and visualization. This allows code reuse when transitioning to production business tool.

Current Challenges

1. **Data Volume Management**
   - **Challenge:** With 100k-150k applications per year across 9 years (2016-2024), full dataset will be ~1 million records
   - **Approach:** Implementing incremental data collection with state-by-state or year-by-year fetching. Will use pandas for initial processing and may consider Dask for larger-scale operations if memory becomes an issue.
2. **Data Quality and Consistency**
   - **Challenge:** Historical E-Rate data may have missing values, inconsistent formatting, or schema changes across funding years
   - **Approach:** Building robust data validation and cleaning pipelines in the preprocessing module. Will document any data quality issues discovered during EDA.

Potential Issues Expected

1. **Feature Engineering Complexity**
   - The model will require calculating derived features such as:
     - Historical application success rates per entity
     - Comparison of entity's discount rate vs. actual NSLP percentage
     - Time-based features (application trends, funding cycles)
   - Will need careful planning to avoid data leakage in train/test splits
2. **Data Integration Challenges**
   - Merging three data sources (USAC, NCES, IMLS) will require careful handling of:
     - Different entity identifiers across datasets
     - Geographic matching (zip codes, FIPS codes, coordinates)
     - Temporal alignment (different data collection periods)
3. **Model Selection and Validation**
   - Need to balance model interpretability (important for business insights) with predictive accuracy
   - Will need to properly handle class imbalance if approval rates are very high or very low

---

Next Steps

1. **Complete remaining data collection modules** (NCES and IMLS)
2. **Exploratory Data Analysis** to understand distributions, correlations, and data quality
3. **Feature engineering** and dataset preparation for modeling
4. **Model development** starting with baseline logistic regression
5. **Validation and testing** using 2024 data as holdout set

GitHub Repository

URL: https://github.com/princessmanifest/erate-prospector

Repository includes: - Complete data collection module for USAC E-Rate API - Comprehensive test suite demonstrating API connectivity - Project documentation (README, requirements.txt, .gitignore) - Modular structure ready for expansion