Broadband Analysis

Ruben Brionez Jr

Bellevue University

DSC680 – Applied Data Science

Matthew Metzger

07/20/2025

Introduction

Business Problem

The business problem for this project is similar to the previous project, but in this project, we are focusing on underserved areas. Underserved for this project means internet speeds across all mediums that are not equal to 1000mbps download (our current base offering).

Background and History

The government has given various ISPs funding to build fiber to underserved areas across the country. This has increased the presence of fiber and made it more important to find areas that lack fiber for a growing fiber communications company.

Data Explanation

To forecast likely areas to remain underserved, I used two primary datasets: *FCC Broadband Deployment Data (2025 release)*: Provides provider counts and technology availability by geography - *ACS S1903 5-Year Estimates*: Supplies median income data for cities, towns and counties. The data was cleaned and ultimately joined to create a single data set that contained all the features needed. These features are focused on speed, median income, and total units.

Methods

The datasets were loaded and explored in Python using pandas in a Jupyter Notebook. Filtering and cleaning were performed to FCC records to county rows using the FIPS ID. Similar steps were taken on the ACS data for cleaning. The FIPS ID was also contained in the data set and made merging easy.

The merged dataset included average speeds, average Gbs availability, and median income values across demographics. This joined data set was used to train a Random Forest model to predict counties that would remain underserved in the next 6-12 months (standard FCC lag time for reporting new data). The model was then sued to predict the same but looking at the state level by grouping counts of underserved counties.

Analysis

The final modeling approach focused on predicting whether a county currently has gigabit broadband availability, using this as a proxy for identifying areas likely to remain underserved in the near future. This approach assumes a six- to twelve-month lag in FCC reporting, which from experience is standard.

Prior to modeling, exploratory data analysis (EDA) was conducted to understand the structure and distribution of key variables. A histogram and boxplot of median household income (see Figures A1 and A2 in the Appendix) revealed a right-skewed distribution, with most counties falling within the $30,000 to $60,000 range. Similarly, a histogram and boxplot of gigabit broadband availability (speed_1000_100) showed that many counties

report either full availability or none at all, indicating wide disparities in high-speed internet access (Figures A3 and A4).

These patterns informed the feature selection process for the model. The cleaned and merged dataset included over 3,000 counties and combined features from the FCC broadband deployment records and ACS income estimates. Predictor features included total units, average availability at multiple speed tiers (including 25/3 Mbps, 100/20 Mbps, 250/25 Mbps, and 1000/100 Mbps), and the median household income.

A Random Forest classification model was trained using stratified sampling to try and address class imbalance, feature importance after the model was run can be seen in Figure A5. The model achieved perfect performance on the test set, though this outcome is partially attributed to many counties that already report some level of gigabit availability. Despite this, the model was used to classify all counties in the dataset to estimate which would likely remain underserved in the future. Predicted underserved counties—those for which the model predicted no gigabit availability—were grouped by state to analyze geographic disparities in broadband deployment. The resulting state-level summary (Figure A6) highlights regions with significant clusters of underserved counties, which may represent high-priority areas for future fiber construction and investment.

Conclusion

Assumptions

This project assumes that broadband availability reported by the FCC provides a useful—though imperfect—signal for understanding infrastructure deployment across U.S. counties. Additionally, it is assumed that median household income remains stable enough at the county level to act as a reliable socioeconomic feature. The model also assumes that counties without gigabit coverage as of the most recent data are unlikely to see immediate service changes, consistent with the FCC's known data lag of approximately 6 to 12 months.

Limitations

A key limitation of the FCC data is its reliance on provider self-reporting, which often overstates coverage—particularly in rural or low-income areas. This makes it difficult to precisely identify unserved or underserved regions using the FCC data alone. Additionally, the extreme imbalance in the dataset (with most counties already reporting some gigabit availability) posed modeling challenges and limited the ability to validate predictions against ground truth. While the ACS dataset provided full national coverage for income data, the merging of datasets required assumptions about geographic consistency that may not hold in every case.

Challenges

One of the primary challenges in this project was dealing with inconsistencies in the FCC data, including inflated provider claims and duplicated entries across service types and business/residential categories. This required multiple rounds of filtering and aggregation to produce a reliable county-level dataset. Another challenge was the need to replace earlier modeling targets (such as fiber presence by city) after discovering insufficient variation and imbalance.

Future Uses

The resulting model could be used quarterly or annually to evaluate changing market conditions. However, additional data is needed to fully verify the number of counties that have access to Gbps speeds. The FCC data may have some discrepancies or varied methods of self-reporting.

Recommendations

The model gives us a good starting point for identifying underserved counties, but it's only as reliable as the data behind it. The FCC data has major flaws—it's self-reported and often overstates coverage—which limits how much we can trust the results. For future work, we'll need better broadband availability data, ideally verified or pulled from more detailed sources. With cleaner data, this model could be a powerful tool for targeting real fiber expansion opportunities.

Implementation

The model can be used to highlight counties that might still be underserved and passed to the business development team for review. These areas can then be looked at more closely for things like permitting, right-of-way access, and construction feasibility. It's not meant to be a standalone decision tool, but rather one piece of a broader market evaluation process. It also pairs well with the model from the previous project to build a more complete view of new market opportunities.

Ethical Assessment

The model doesn't use race or any sensitive demographic features. Income is included only to help estimate economic viability—not to deprioritize lower-income areas. In fact, one goal is to make sure rural and historically underserved communities aren't overlooked just because they haven't been built out yet. Care should be taken to ensure these areas stay in focus during future deployment planning.

Appendix

Figure A1

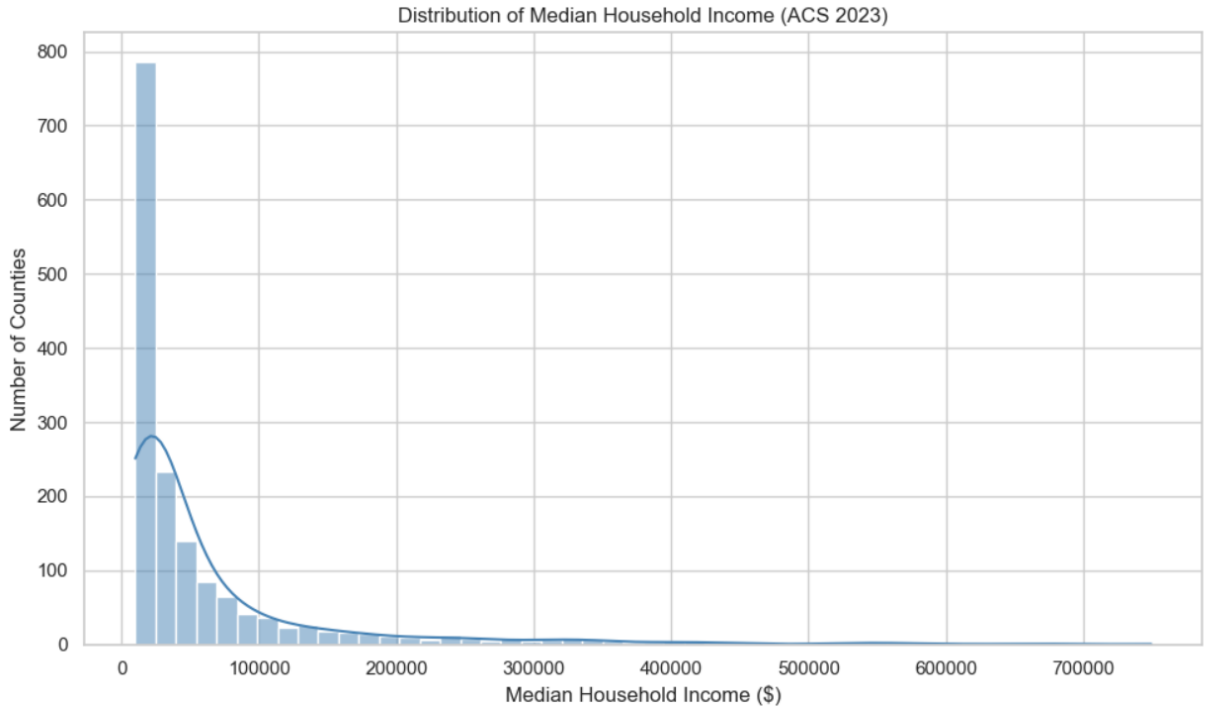

Figure A1 Illustrates the distribution of median income by number of counties. We can see that most counties in the US have a median income less than 100k.

Figure A2

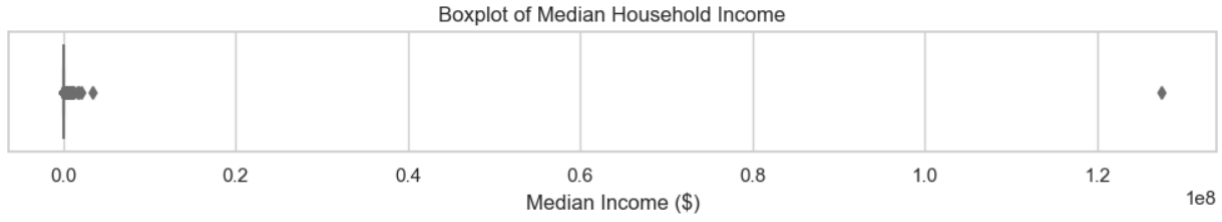

Figure A2 Illustrates the distribution without limiting the maximum income to 750k. We can see there is an outlier here and the reason for the capped histogram.

Figure A3

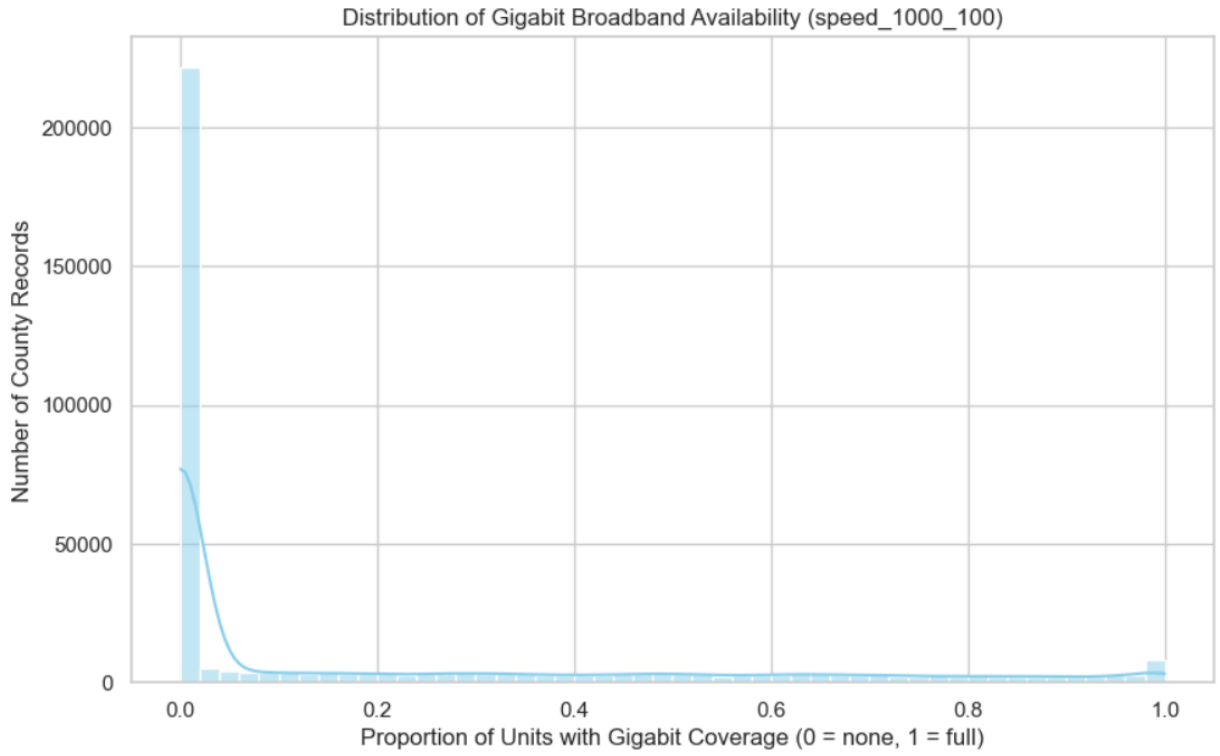

Figure A3 Illustrates the distribution of Units (residential) that have access to Gbs speeds.

Figure A4

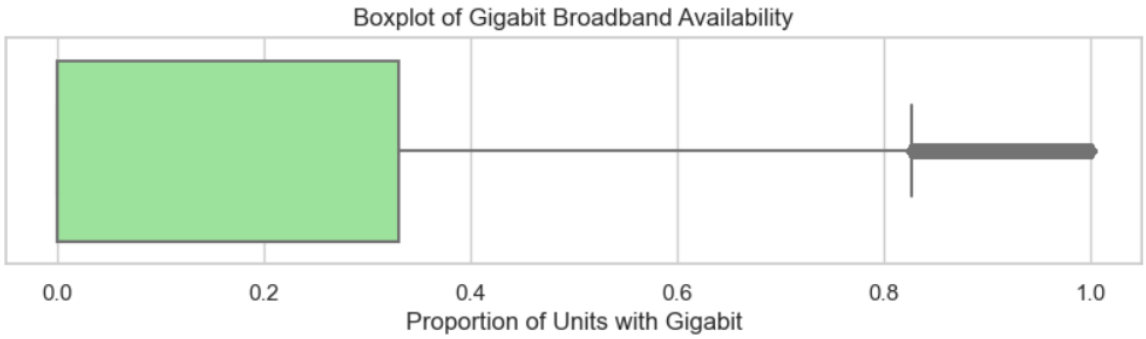

Figure A4 is another illustration of the distribution of access to Gbs speeds.

Figure A5

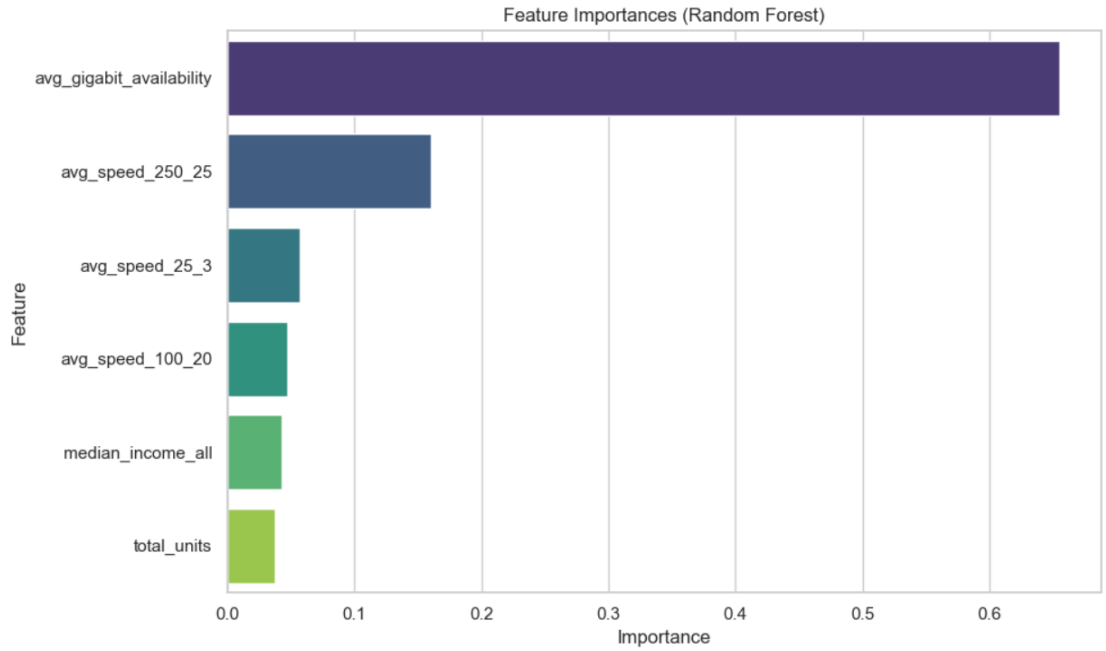

Figure A5 Illustrates the importance of predicting features for the Random Forest Model.

Figure A6

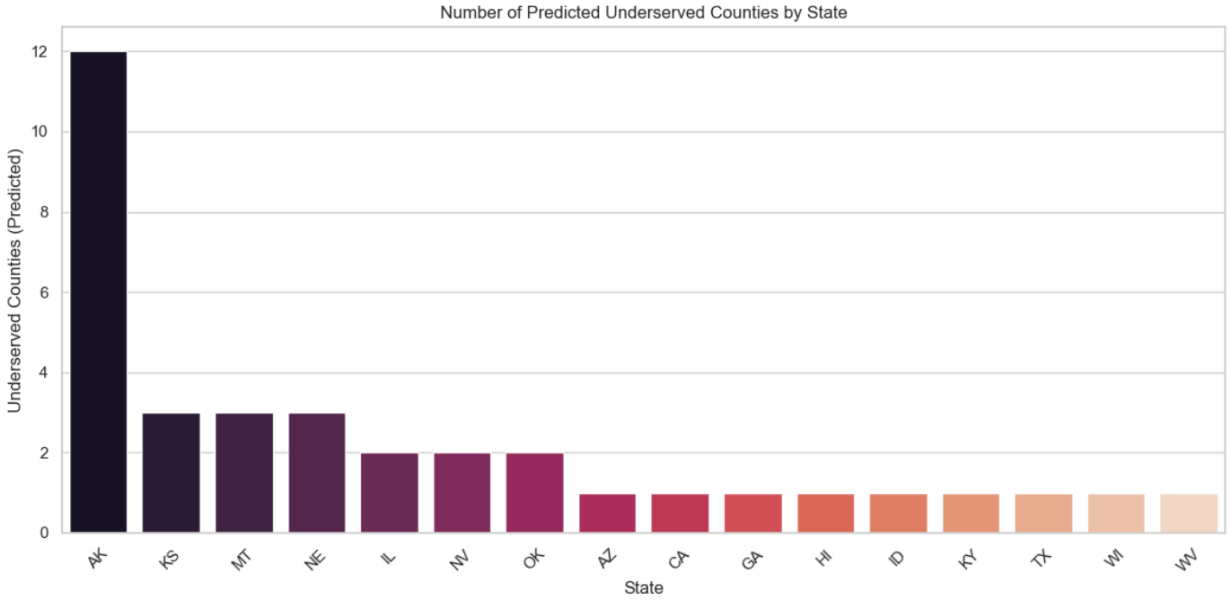

Figure A6 Illustrates the predictions at the state level. The sate data comes from the grouped values of counties by state.

References

Federal Communications Commission. (2025). *Broadband Data Collection (BDC): Fixed Broadband Availability Data*. https://www.fcc.gov/BroadbandData

U.S. Census Bureau. (2024). *American Community Survey 5-Year Estimates, Table S1903: Median Income in the Past 12 Months (in 2023 Inflation-Adjusted Dollars)* [Data set]. https://data.census.gov/

Questions

Q1. *Why did you choose to model gigabit availability instead of fiber specifically?*

A1. Fiber presence in the FCC data wasn't consistently reported or meaningful across counties. Gigabit availability (speed_1000_100) is a clearer indicator of high-performance broadband, and aligns more directly with end-user experience anyway.

Q2. *How did you define an "underserved" county in your model?*

A2. Counties predicted to have no gigabit broadband availability (has_gigabit = 0) were considered underserved. This assumption was based on FCC lag time and the lack of gigabit-level infrastructure.

Q3. *What were the key features used in the model?*

A3. Median household income, total broadband serviceable units, and average availability of four speed tiers: 25/3 Mbps, 100/20 Mbps, 250/25 Mbps, and 1000/100 Mbps.

Q4. *Why did you use Random Forest for this classification task?*

A4. Random Forest handles non-linear relationships well, is robust to outliers, and performs reliably with mixed data types—making it a good fit for infrastructure and demographic data.

Q5. *How did you address class imbalance in your dataset?*

A5. We used stratified sampling in the train-test split to ensure both classes were represented fairly. However, we also acknowledged that most counties already report gigabit availability, which limits class variation.

Q6. *What limitations did you encounter with the FCC dataset?*

A6. The FCC data is self-reported by providers and tends to overstate coverage. It also includes duplicate records per county and doesn't reflect recently completed builds due to a 6–12 month reporting lag.

Q7. *How did you ensure consistency between the ACS and FCC datasets?*

A7. We used standardized FIPS codes (geography_id) as the merge key, filtered both datasets to county level, and dropped rows with missing or invalid values to ensure alignment.

Q8. *What does the histogram of median income tell you about the dataset?*

A8. The distribution is right-skewed, with most counties earning between $30,000 and $70,000. This supports the idea that broadband investments should not rely solely on income thresholds.

Q9. *How could this model be improved in future work?*

A9. Integrating verified deployment data from ISPs or local governments would make the target more accurate. Also, using more granular data—like census block or address-level availability—would improve precision.

Q10. *How do you ensure the model doesn't reinforce historical bias against rural areas?*

A10. We avoided using demographic features like race and treated income as a neutral planning variable. The intent is to highlight underserved areas—especially rural counties—not exclude them.