Housing Affordability and Telecommunication Expansion

Ruben Brionez Jr

Bellevue University

DSC680 – Applied Data Science

Matthew Metzger

06/29/2025

Introduction

Business Problem

As part of a startup telecommunications company and 12 years' experience in the industry, I have first-hand experience in the struggles and challenges of growing a telecommunications company from the start. The business problem that needs to be solved is potential new market growth, meaning as a company, we need to find the markets/cities in Ohio, Pennsylvania, and Michigan that are growing cities in order to construct our fiber internet infrastructure in areas that are predicted to be growing in population.

Background and History

With over a decade of experience in the telecommunications industry — which inluded both mid-sized and national providers — I've observed that traditional approaches to market expansion are often slow and reactive. Now working with a fast-paced startup, I believe that data-driven analysis is essential for identifying underserved or high-potential markets more efficiently than legacy methods. This project stems from that belief: to leverage housing affordability and income data to proactively guide strategic infrastructure expansion.

Data, Process, and Analysis

Data Explanation

To begin the process of finding an answer to the business problem analyzing data of income and housing affordability metrics will be required. I will be using data sets from the

US Census, Zillow and Redfin in an effort to provide analysis on affordable housing and income for the selected states and cities within.

The data from the US census provides information on median and mean income across different demographics, while the Zillow data provides some city and state level information on housing prices and estimated income needed for a standard 20% down.

Methods

Various methods were used to clean and process the data while maintaining the core of the data. CSVs were downloaded from the public facing sites of Zillow and the government census site. Those CSVs were then loaded using pandas and Jupyter Notebook to perform the exploratory data analysis as well as the visualizations.

A variety of Python libraries were used to clean, prep and build a model from the data. More details will be provided in the Analysis section, but after the EDA was completed and some visualizations created a logistic regression model was created to predict if some cities would move in or out of the top 20 affordability list.

Analysis

The analysis incorporated four publicly available datasets — two from Zillow and two from the U.S. Census Bureau — to evaluate affordability trends in Ohio, Michigan, and Pennsylvania (U.S. Census Bureau 2023). Zillow data provided monthly city housing prices and estimated income needed to purchase a home with the standard 20% down (Zillow Research, 2025). The U.S. Census data contributed state-level median household income across multiple demographic groups. The racial demographics were combined to get an overall median income estimate (Figure A1). After cleaning and filtering the data, an affordability ratio was calculated for each city, the affordability ratio was calculated as the income needed divided by the median home price. This allowed for a direct comparison of affordability across cities and time (Figure A2).

To identify which cities might become attractive markets for expansion, a logistic regression model was initially developed to classify cities that would rank in the top 20 most affordable. However, due to limited accuracy (Figure A3) and the lack of a strong linear relationship, a Random Forest classifier was implemented as an alternative. The Random Forest model achieved over 88% accuracy and 100% recall on test data, successfully predicting cities like Philadelphia, Columbus, and Pittsburgh as candidates likely to remain or enter the top 20 (Figure A4). Features included affordability ratios from three prior months, making it a time-sensitive yet practical tool.

Conclusion

Assumptions

One of the assumptions made during this analysis was the assumption that combining the median income across races would yield a more encompassing median. Since race is not a factor for telecommunication expansion plan (the physical construction portion) it seemed that excluding race would not have any significant impact on the analysis.

A second assumption made was that the slightly dated income data was still relevant. The data for income comes from the US Census bureau and is not gathered annually. It is assumed that the income has not significantly changed since 2023 for the analysis.

Limitations

One limitation to note is the date for income data used for analysis. The US Census does not appear to collect income data every year; this means the income data used for analysis is slightly dated. This is not expected to have a significant impact on analysis but is worth noting.

Challenges

Model selection for predictions was a challenge. The initial plan was to include a regression model to predict the future affordability of housing in the three selected states. The challenge found was that there did not appear to be any strong linear relationship between the predictors used. The regression model had to be removed, and a new model had to be selected and trained.

Future Uses

The future use of the analysis and model could be implemented if found useful for the business problem. The groundwork has been laid out to train and use the model for predictions, and the analysis provides a template of the overall scope of work for what the business should consider when looking for new market/cites for expansion.

Recommendations

The analysis being complete and the model trained, a recommendation that can now be advised is to incorporate updated data into future versions of this model and analysis. This will keep the model and analysis relevant and as up to date as necessary for a growing start-up company.

Implementation

The implementation of this model should proceed based on the data and analysis. In the process of implementation, the model should only be used as an initial guide for potential new markets and cities. The model should be implemented in this way; the predicted markets are provided from the model, and those markets are then checked against the existing markets of the company, if not an existing market, then those predicted can be moved to the business development team to perform a deeper level of analysis which would include different metrics for expansion.

Ethical Assessment

Ethical considerations were kept in mind while completing the analysis. This was partially the reason I combined the race data for income; I did not want to create a potential ethical issue regarding the race and location of expansion markets. Looking at the overall median income, while excluding race, allows for an unbiased analysis of potential new markets.

Appendix A

Figures, Illustrations and Supporting Documents

Appendix A contains figures and illustrations referenced in the white paper.

Figure A1

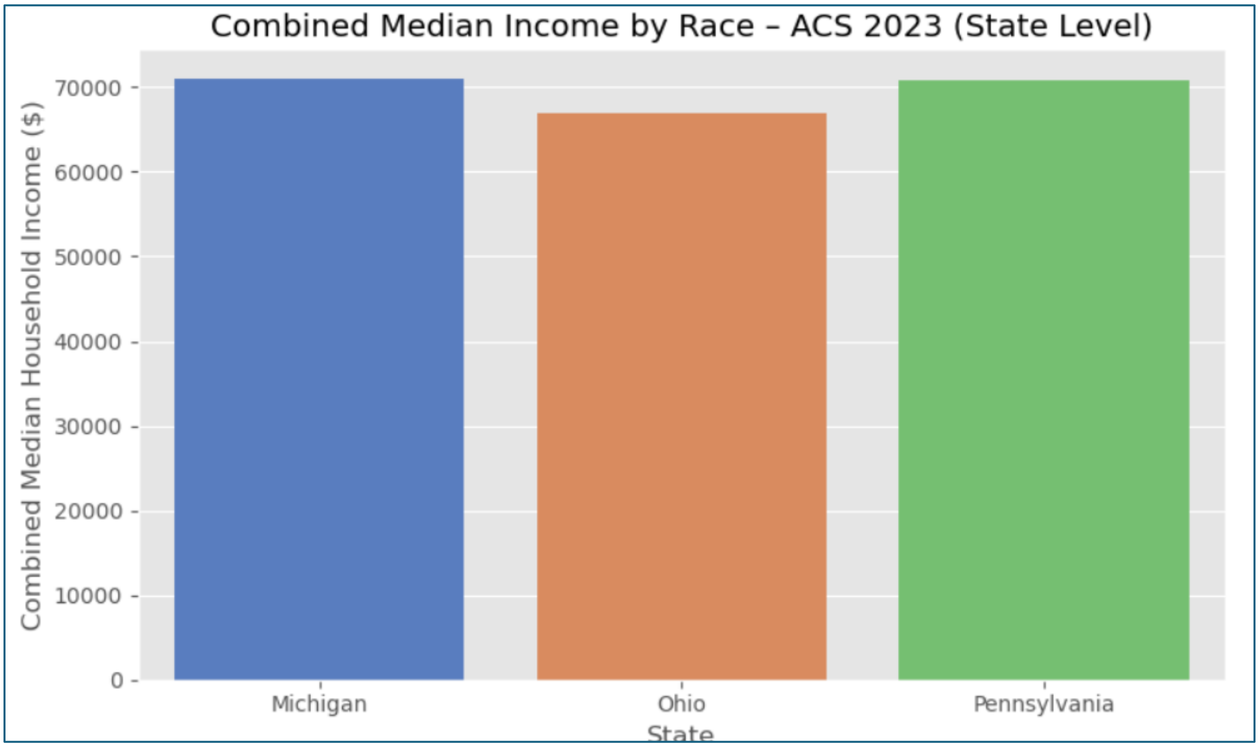
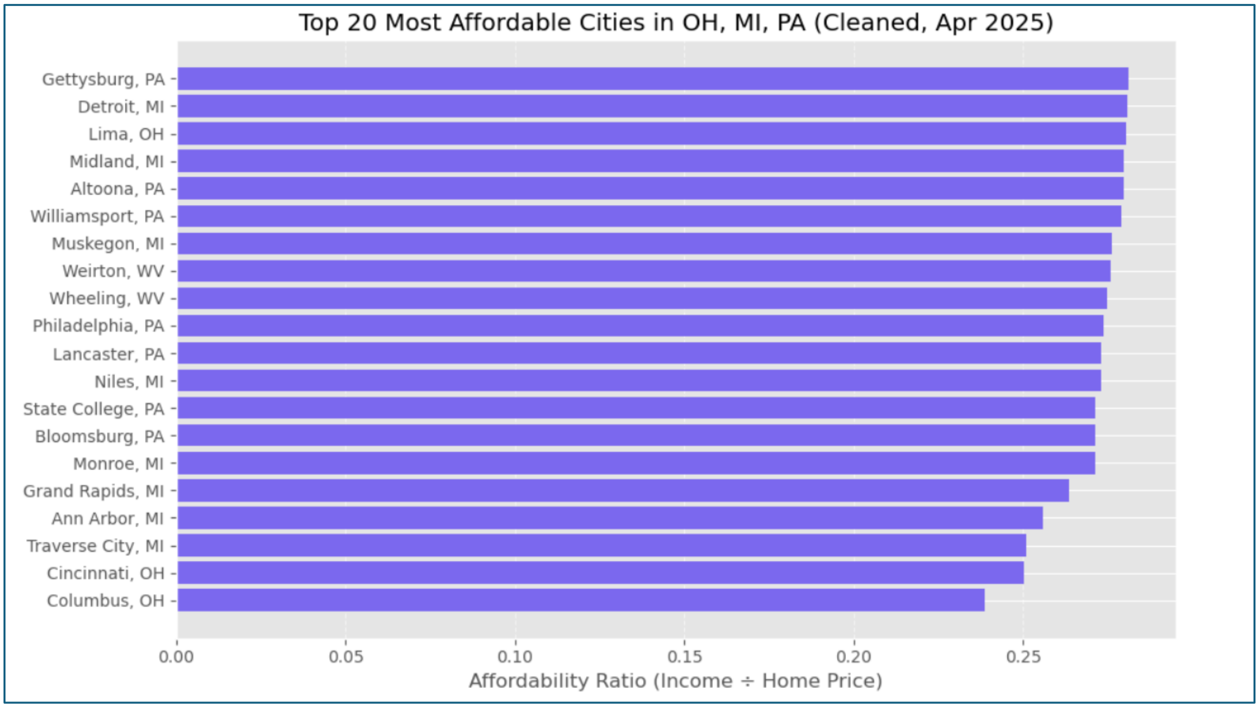


Figure A1 Illustrates the median income for all races combined

Figure A2



Figures A2 Illustrates the top 10 cities based on the calculated Affordability Ratio.

Figure A3

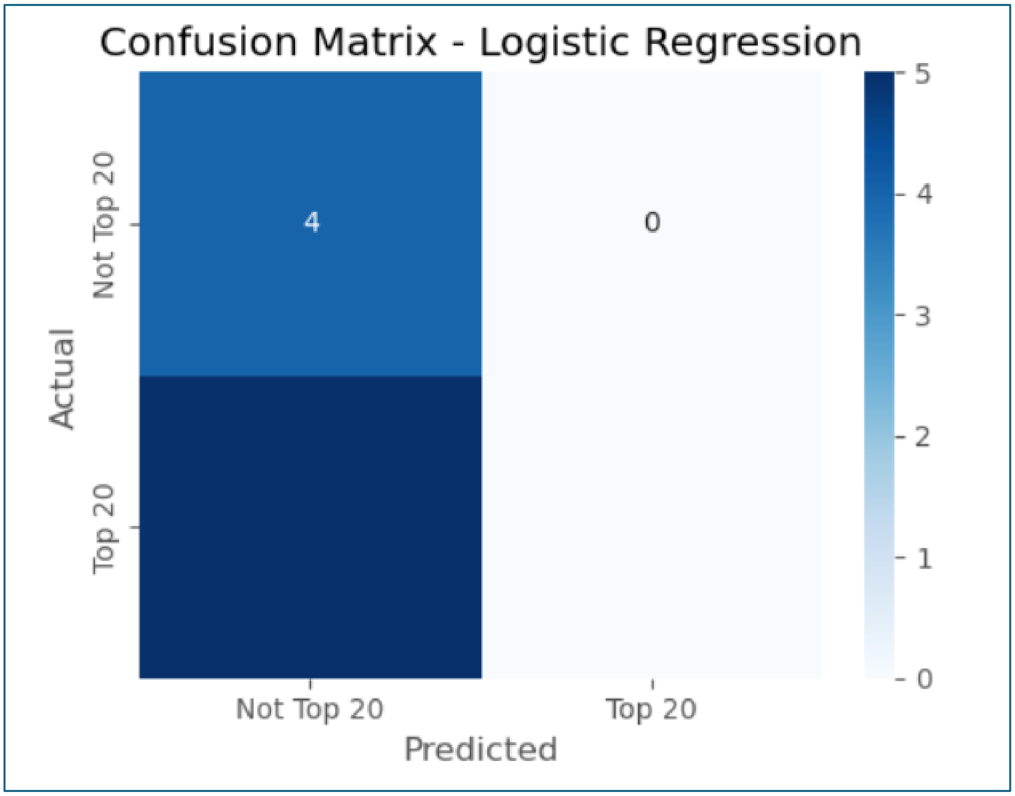
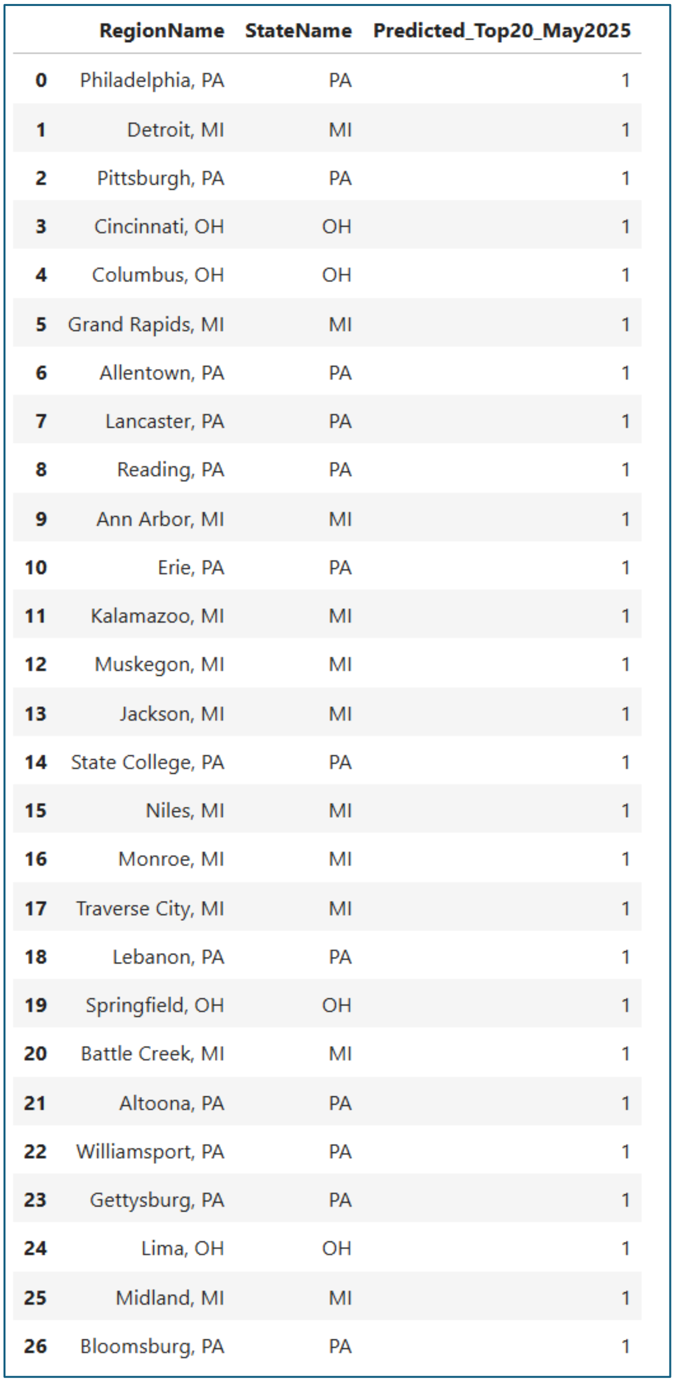


Figure A3 Illustrates the failed predictions for the Logistic Regression Model

Figure A4



	RegionName	StateName	Predicted_Top20_May2025
0	Philadelphia, PA	PA	1
1	Detroit, MI	MI	1
2	Pittsburgh, PA	PA	1
3	Cincinnati, OH	OH	1
4	Columbus, OH	OH	1
5	Grand Rapids, MI	MI	1
6	Allentown, PA	PA	1
7	Lancaster, PA	PA	1
8	Reading, PA	PA	1
9	Ann Arbor, MI	MI	1
10	Erie, PA	PA	1
11	Kalamazoo, MI	MI	1
12	Muskegon, MI	MI	1
13	Jackson, MI	MI	1
14	State College, PA	PA	1
15	Niles, MI	MI	1
16	Monroe, MI	MI	1
17	Traverse City, MI	MI	1
18	Lebanon, PA	PA	1
19	Springfield, OH	OH	1
20	Battle Creek, MI	MI	1
21	Altoona, PA	PA	1
22	Williamsport, PA	PA	1
23	Gettysburg, PA	PA	1
24	Lima, OH	OH	1
25	Midland, MI	MI	1
26	Bloomsburg, PA	PA	1

Figure A4 Illustrates the predictions made by the Random Forest model

References

U.S. Census Bureau. (2023). *Median income in the past 12 months (in 2023 inflation-adjusted dollars) – S1902*. American Community Survey 1-Year Estimates. https://data.census.gov/

U.S. Census Bureau. (2023). *Household income by race and Hispanic or Latino origin of householder – S1903*. American Community Survey 1-Year Estimates. https://data.census.gov/

Zillow Research. (2025). *Homeowner income needed for a 20% down payment on a typical home*. Zillow Economic Research. https://www.zillow.com/research/data/

Zillow Research. (2025). *Zillow Home Value Index (ZHVI) – Middle price tier*. Zillow Economic Research. https://www.zillow.com/research/data/

Audience Questions

1. How did you ensure the affordability ratio was a reliable indicator of market opportunity for telecom expansion?
    a. The affordability ratio took into account income and house price,
2. Why were Ohio, Pennsylvania, and Michigan specifically chosen for this analysis?
    a. The states selected are areas that our fiber network area has already been built; by looking to grow in the existing states we can utilize our existing network to help mitigate the new construction costs.
3. What factors led to the decision to switch from a logistic regression model to a Random Forest classifier?
    a. The confusion matrix for the logistic regression model showed that there were not any strong linear relationships between variables. The random forest was better suited for nonlinear relationships.
4. Could using metro-level income data (instead of state-level) improve the accuracy of the affordability analysis?
    a. Using metro-level income data could absolutely improve the accuracy of the affordability analysis. However, during the initial analysis this data was not readily available.
5. How would this model adapt to rapid economic shifts, such as inflation or a recession?

    a. The model would not update well to rapid changes since a Random Forest model is not inherently dynamic. The model would need to be re-trained on a new data set.

6. What ethical implications did you consider when using race-specific income data, and how were those addressed?
    a. I did not want to exclude any specific race from potential markets, in order to mitigate the chances of bias due to race the median income was aggregated across races.
7. Is there a risk of excluding high growth but less affordable cities that may still be valuable markets?
    a. Yes, there is potential to exclude high growth but less affordable market, the model would need to be re-trained to incorporate population growth data.
8. How might this model be scaled or adapted for use in other states or regions?
    a. The model should adapt well to other regions and states but would require re-training with the new data sets.
9. What additional data sources (e.g., broadband adoption rates, population growth) could improve future versions of the model?
    a. Population growth data and existing broadband infrastructure would be great additional data to have that would improve the model and yield more actionable results.
10. If multiple cities are predicted as viable, what criteria should be used to prioritize construction or investment?

a. As mentioned above, the existing broadband infrastructure in the predicted market would be great data to have. This would allow for analysis of which predicted markets have no fiber internet making them even more desirable for construction.