Using Regression Models to Predict Home Affordability Ratios

Author: Freda Xin

Problem Statement:

What is home affordability ratio?

Home Affordability Ratio = **Median Home Price / Median Annual Household Income**

Example: Manhattan, 2018

Median Home Price = $ 944,600

Median Annual Household Income = $ 82,459

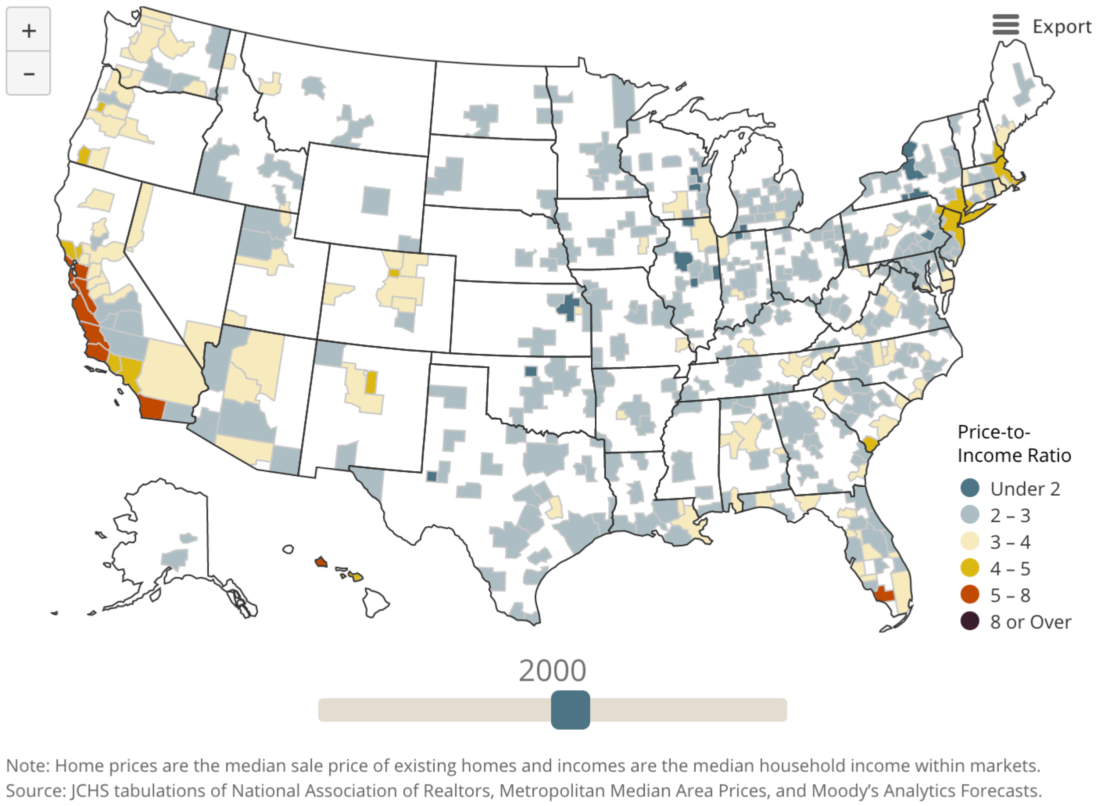
Home Affordability Ratio ≈ 11. 5 years

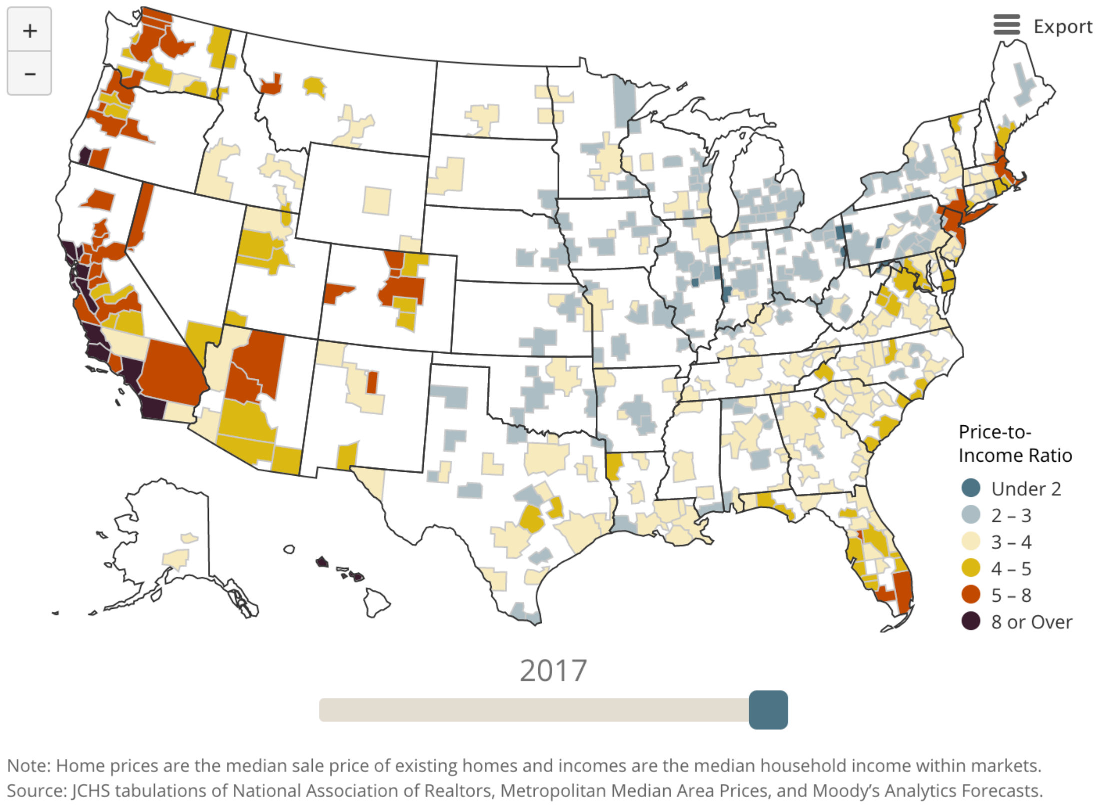
* Data Source: U.S. Census

Problem Statement:

Why is home affordability ratio important?
Who should care about it?

Image Source: [Home Price-to-Income Ratios](#) by Joint Center for Housing Studies of Harvard University



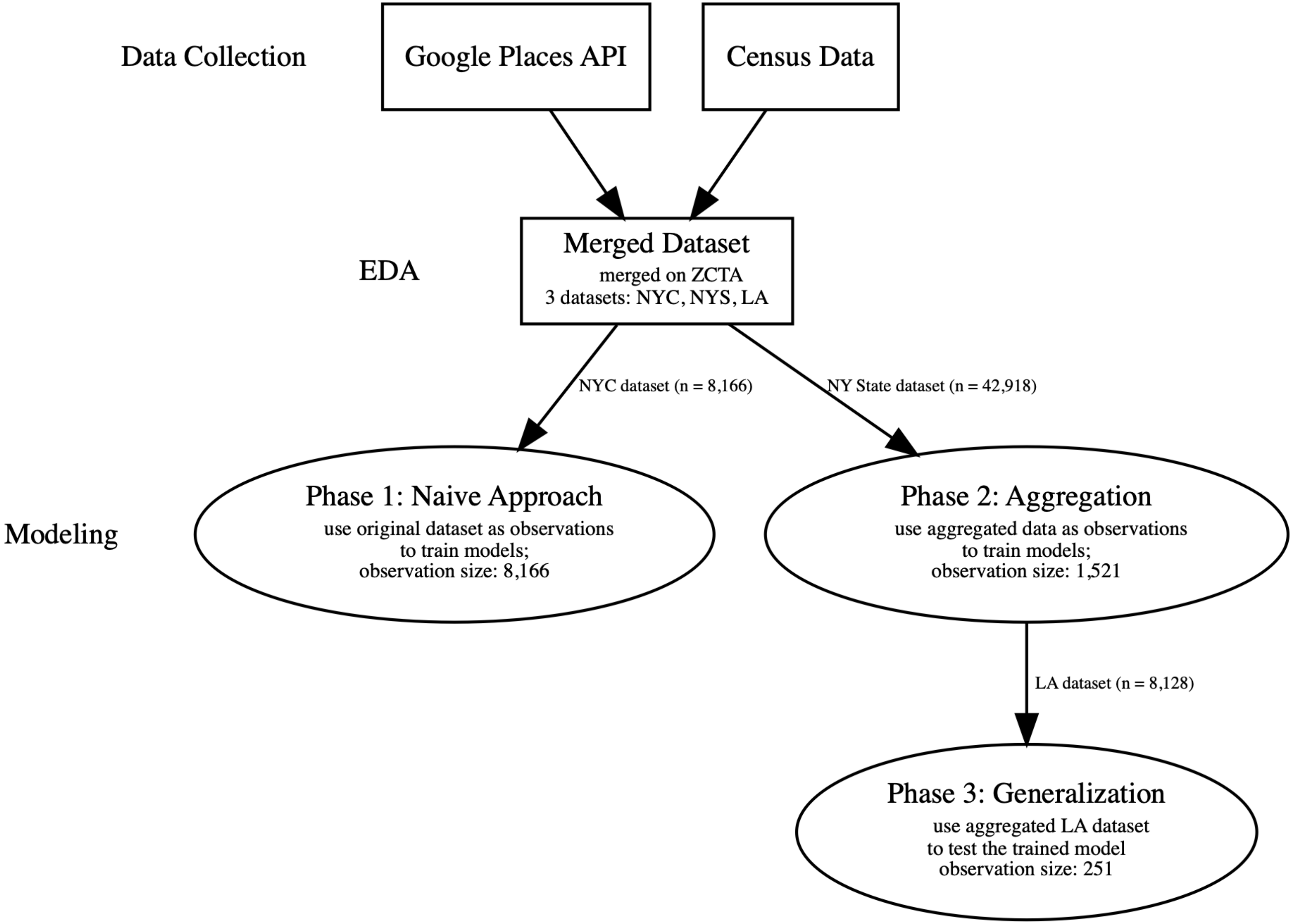


Problem Statement: Project Scope

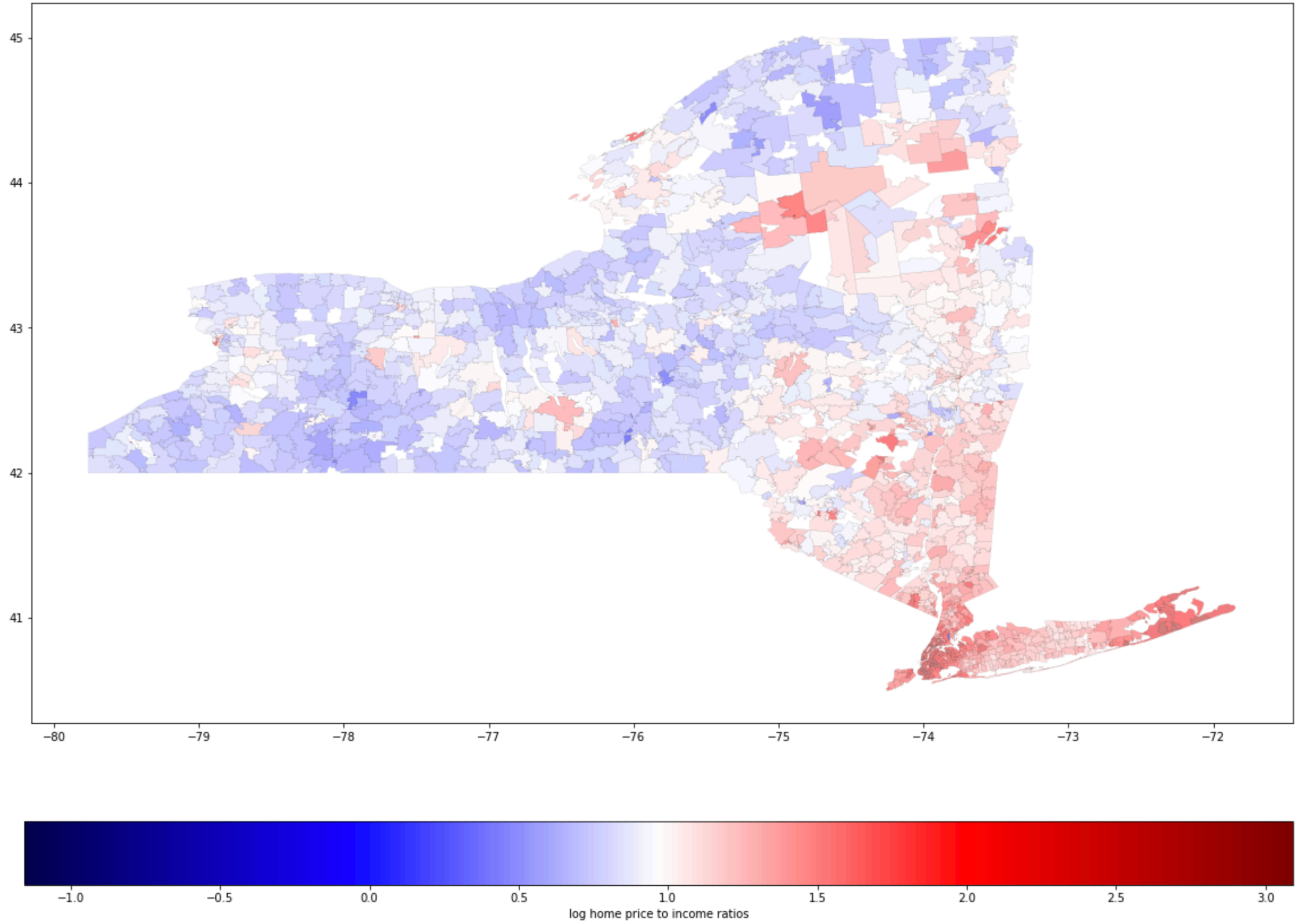
Using Regression Models to Predict Home Affordability Ratios

- Project Scope: explore whether **commercial activities** in a given neighborhood can be predictive for **home affordability ratios.**
- Project Goal: to develop regression models that can make quick predictions given the latest commercial activities in a neighborhood.
- Target Client: municipalities and the general public
- Metric: R2 score

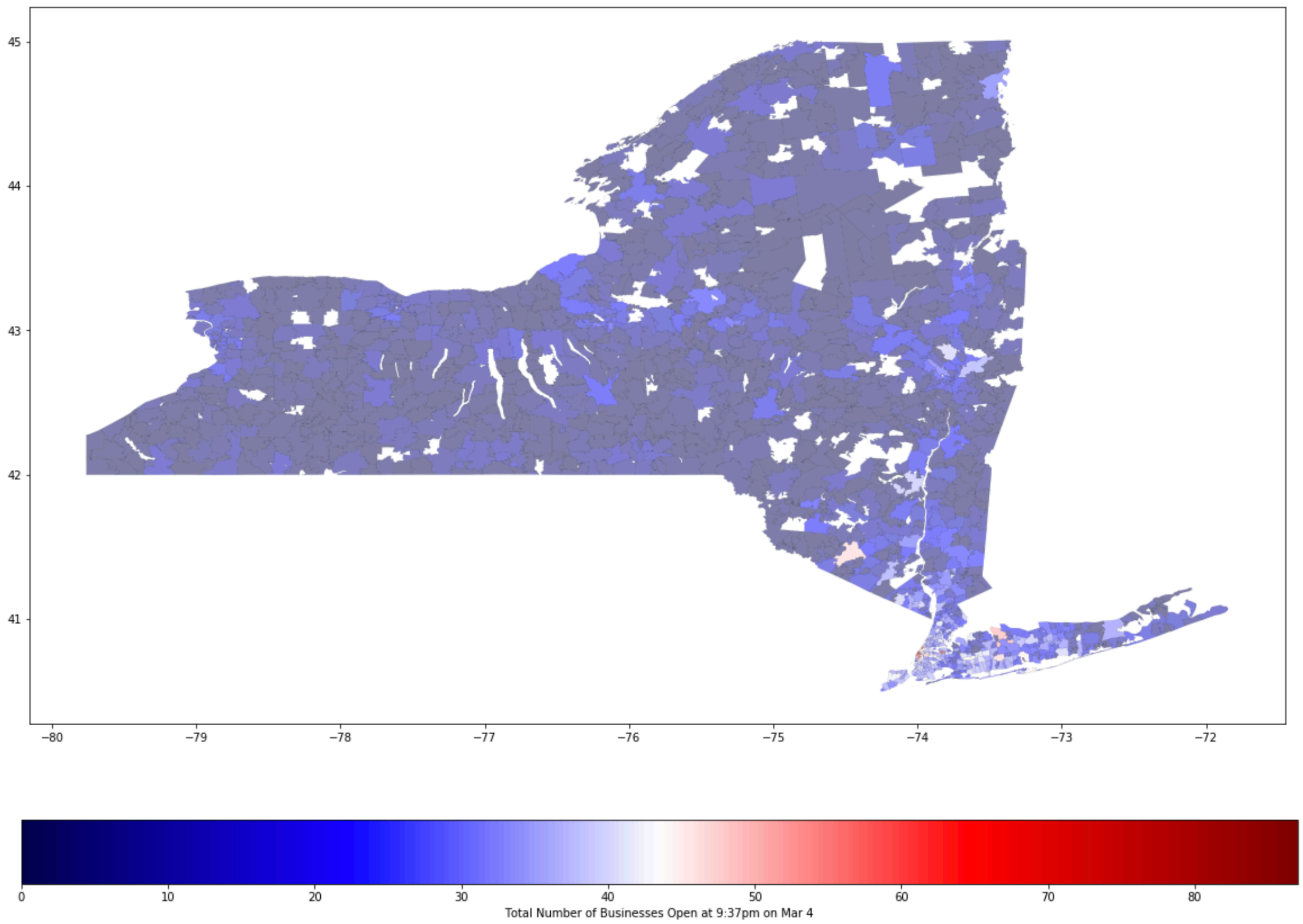
Project Workflow



EDA:
Analysis of Target – Home Affordability Ratios in NYS

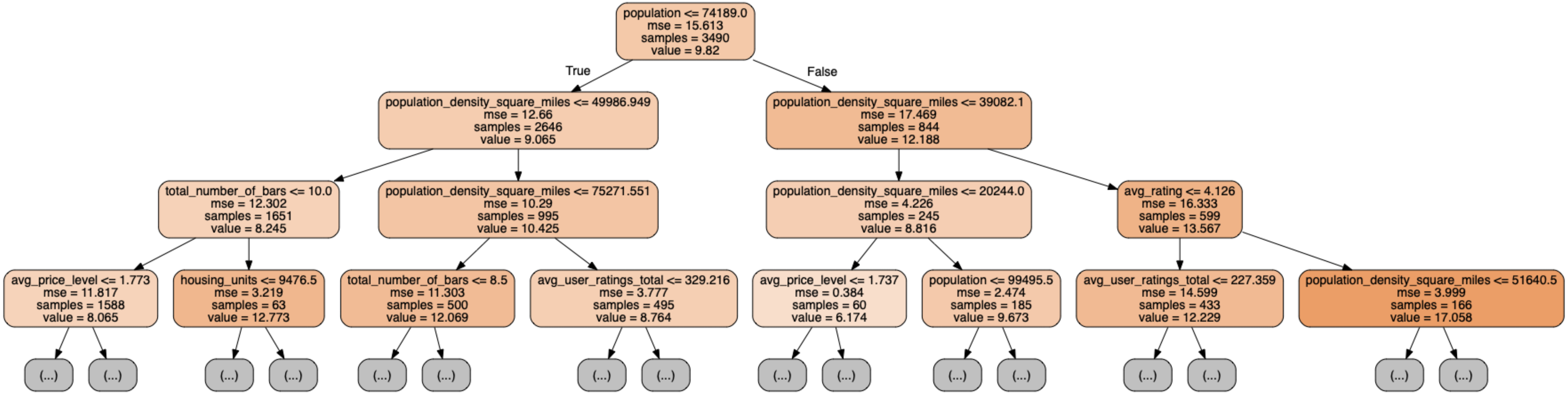


EDA:
Analysis of Feature – "total_open_now_True"

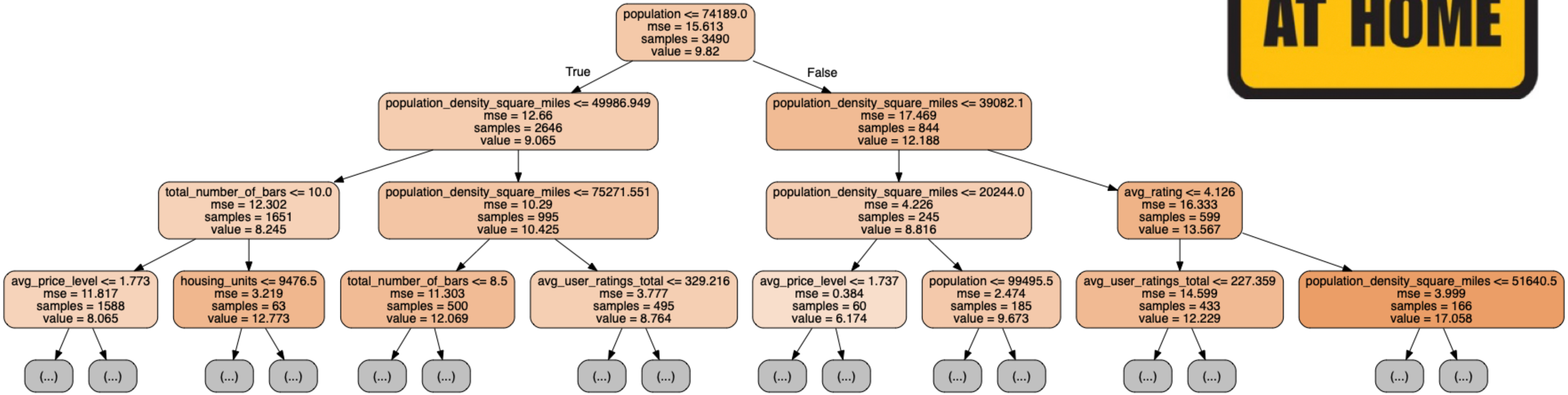


Modeling Phase 1: Naive Approach

Observations: NYC dataset (n=8,166)
Models: Linear Regression, KNN, Decision Tree
Best Model: Decision Tree (Test R2 Score = 0.99)



Modeling Phase 1: Naive Approach
Conclusion: Data Leakage!



Modeling Phase 2: Aggregation

• Original NYS dataset (n = 42,918) -> **Aggregated by ZCTA (n = 1,521)**

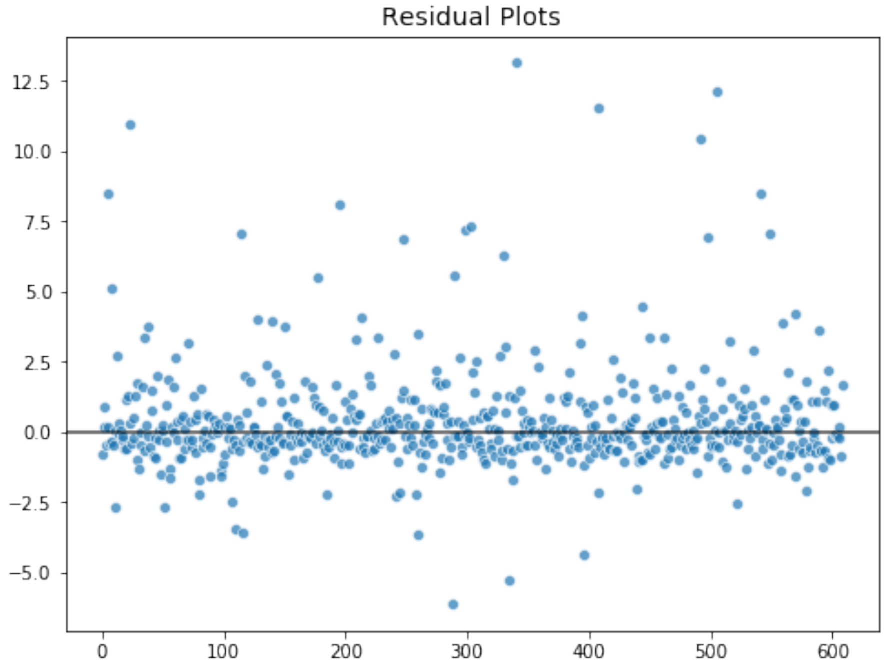
• Method: Pattern submodel to handle missing data: Pattern 0 & Pattern 1

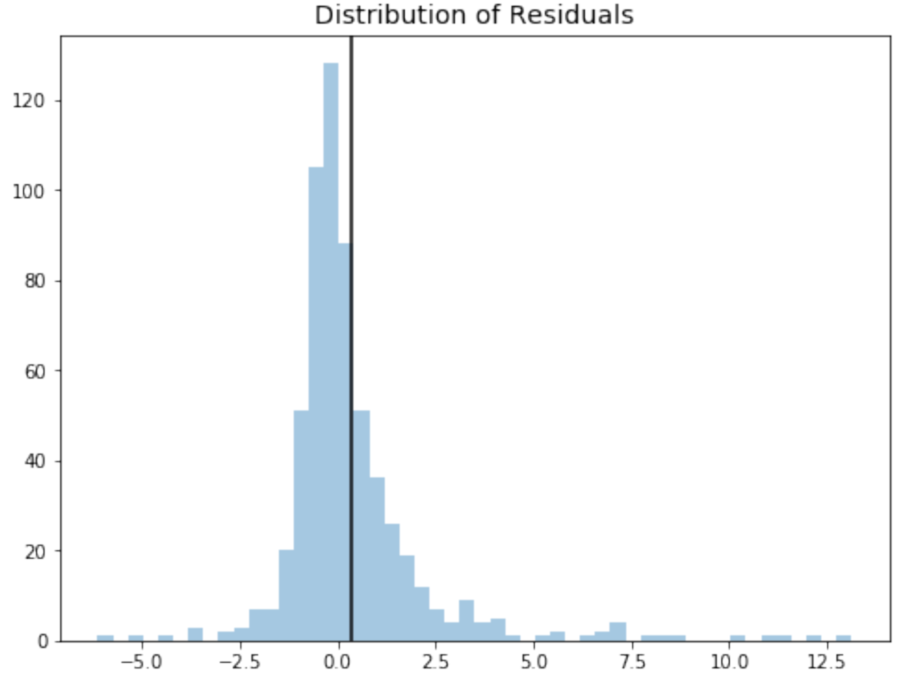
• Model Types:
Linear Regression(+ L1, L2, PCA), Polynomial Regression, KNN, Tree Based, SVR, Stochastic Gradient Decent

• Best Model: BaggingRegressor

• Test R2:
Pattern 0 = 0.6541; Pattern 1 = 0.0046

• Baseline R2:
Pattern 0 = -0.0057; Pattern 1 = -0.0128





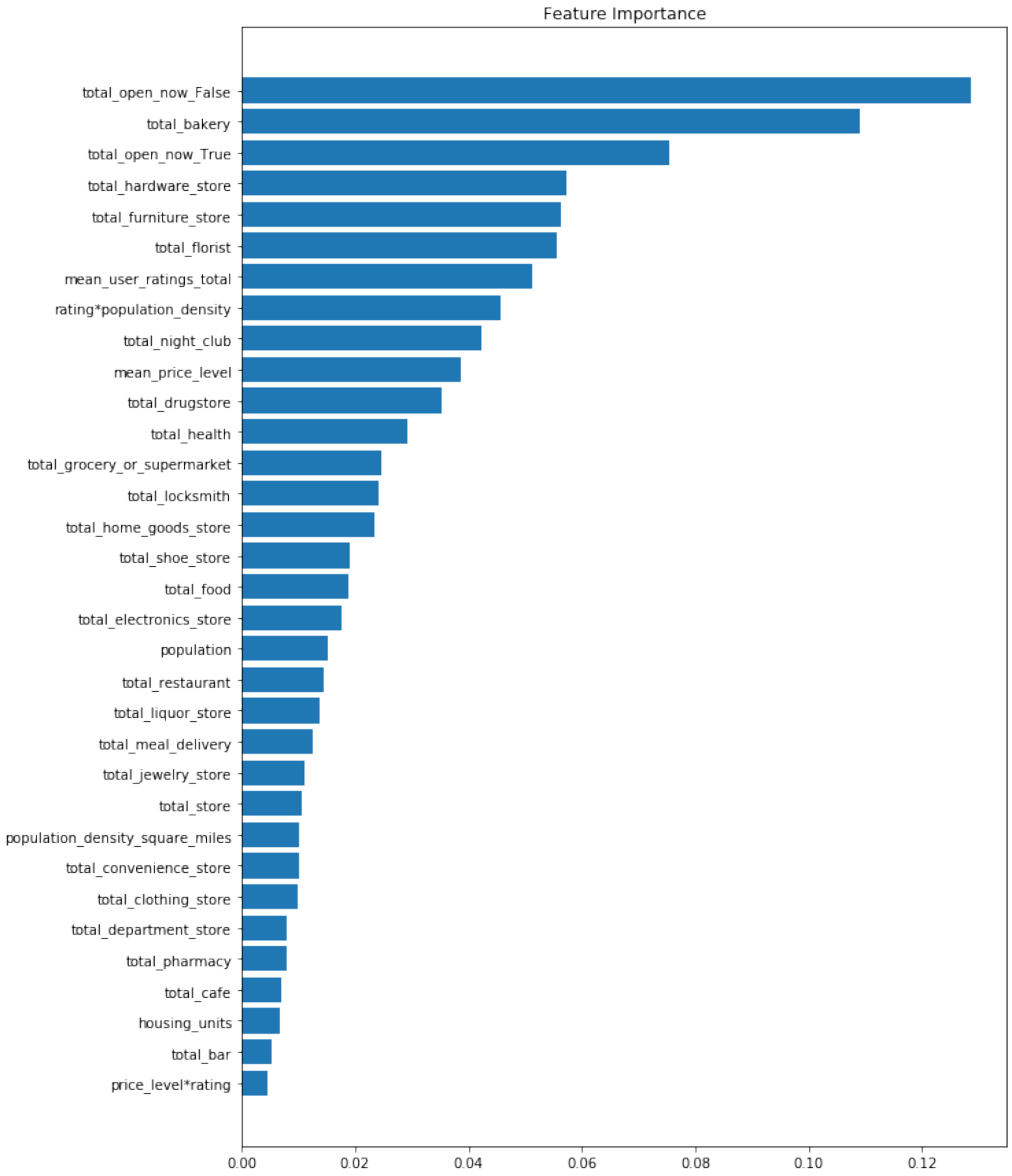
Modeling Phase 2: Aggregation

• Best Model: **BaggingRegressor**

Phase 2 Conclusion:

• Census Data are important for model performance.

• “**open_now**” features are good predictors.



Modeling Phase 3: generalization

- Original LA dataset (n = 8,128) -> **Aggregated by ZCTA (n = 251)**

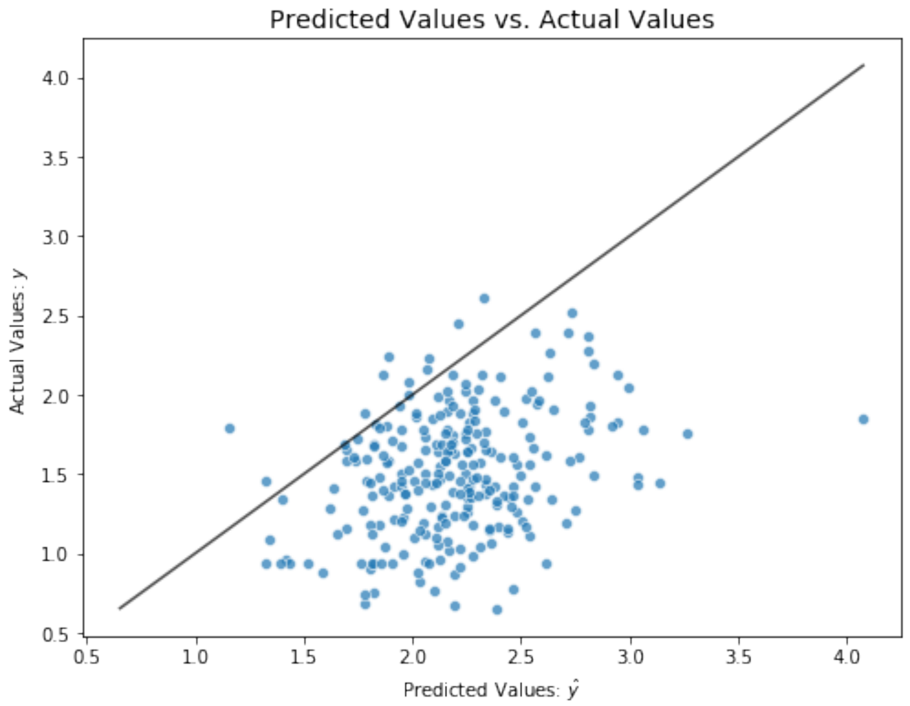
- Goal: Train the model on NYS dataset WITHOUT census features, then test it on the LA Dataset

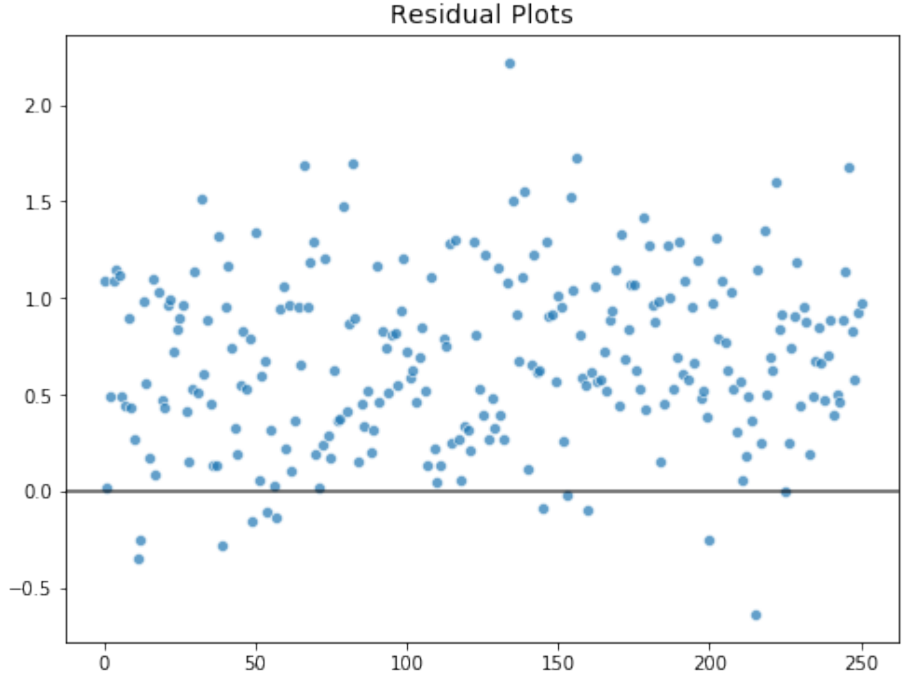
- Model Types:
Linear Regression(+ L1, L2, PCA), Polynomial Regression, KNN, Tree Based, SVR, Stochastic Gradient Decent

- Best Model:
Linear Regression Model with L1 Regularization

- R2: Score
Pattern 0 = -4.0596; Pattern 1 = -5.7796

- Baseline R2:
Pattern 0 = -0.0057; Pattern 1 = -0.0128





Conclusion, Limitation, and Next Steps

Conclusion:

- Commercial activities information collected from Google Places API **alone** are not good predictors for home affordability ratios; Using features combined with Census data improved model performance.
- The model trained on NYS dataset is NOT transferable to LA

Next Steps:

- Improve data quality
- Reevaluate the assumption: are commercial activities in a neighborhood predictive for home affordability ratios?

Limitations:

- Google Place API
- Budget

References

- Home Price-to-Income Ratios by Joint Center for Housing Studies of Harvard University
- Home Price to Income Ratio by longtermtrends.net
- The Impact Of Commercial Development On Surrounding Residential Property Values by Jonathan A. Wiley, Ph.D.
- Predicting Neighborhoods' Socioeconomic Attributes Using Restaurant Data by Lei Dong, Carlo Ratti, and Siqi Zheng
- Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life by Edward L. Glaeser, Scott Duke Kominers, Michael Luca, Nikhil Naik