DATA 3320

Homelessness Rates in the U.S.

Tina Chau

Introduction

This project focuses on utilizing data science methodology to investigate the issue of homelessness in the United States.

The primary inspiration for this research stems from the goals outlined in the HUD study, which aims to comprehend the fundamental factors that contribute to homelessness at the community level.

Specifically, HUD seeks to achieve two objectives:

- identifying market factors that significantly influence homelessness
- develop models that effectively represent and predict homelessness within communities.

Source of Data

The U.S. Department of Housing and Urban Development (HUD) produced a report in 2019 Market Predictors of Homelessness that describes a model-based approach to understanding of the relationship between local housing market factors and homelessness.

Data Science Question(s):

- How accurately can we predict homelessness rates?
- Is there a difference in homelessness rates based on the type of district like in urban or suburban areas?

Analysis Methods

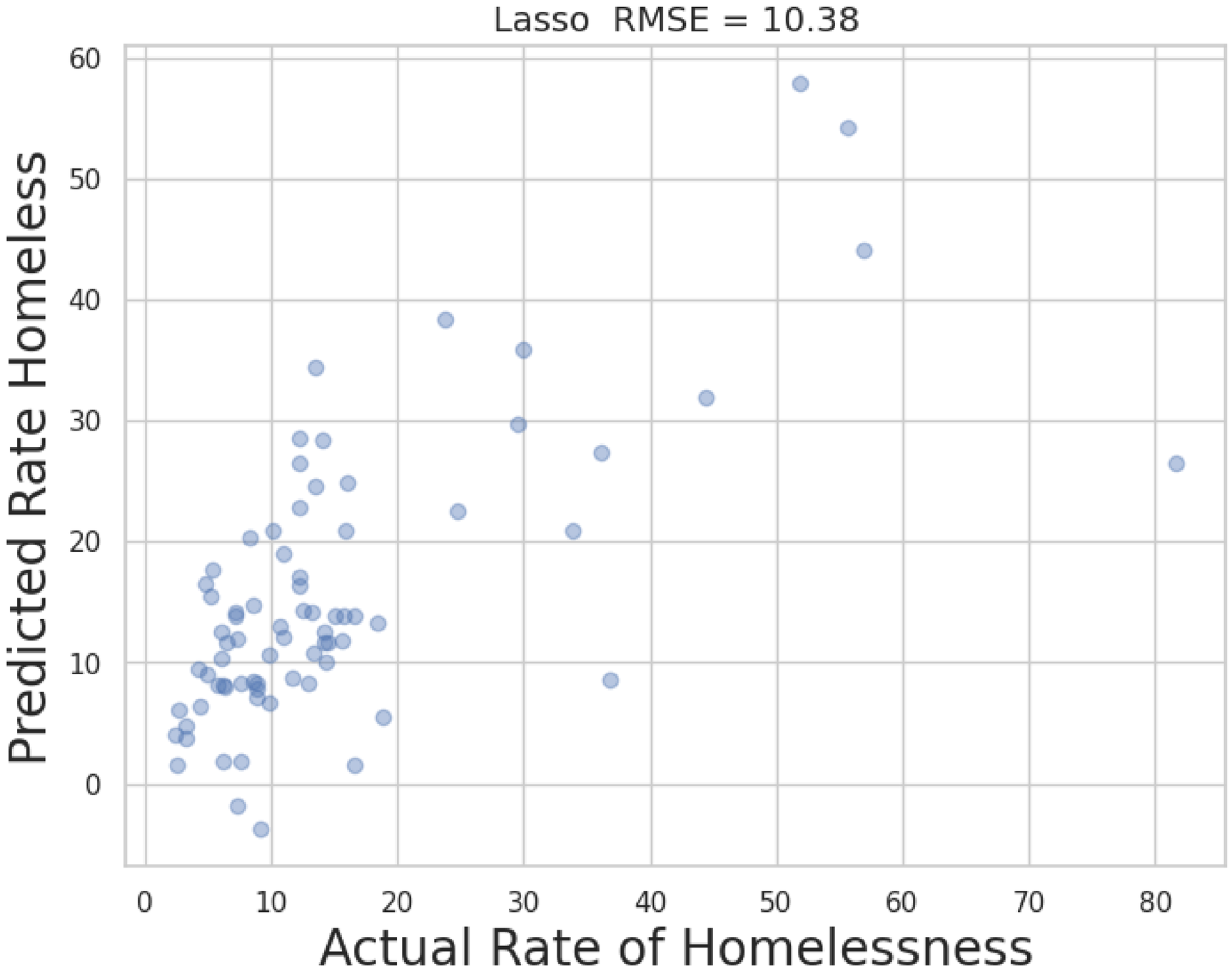
The methods used in this project are:

- Train test splits
- Scaled predictor data
- OLS regressional model
- K fold cross-validation using Lasso, Ridged, and XGBoost models
- Root mean squared

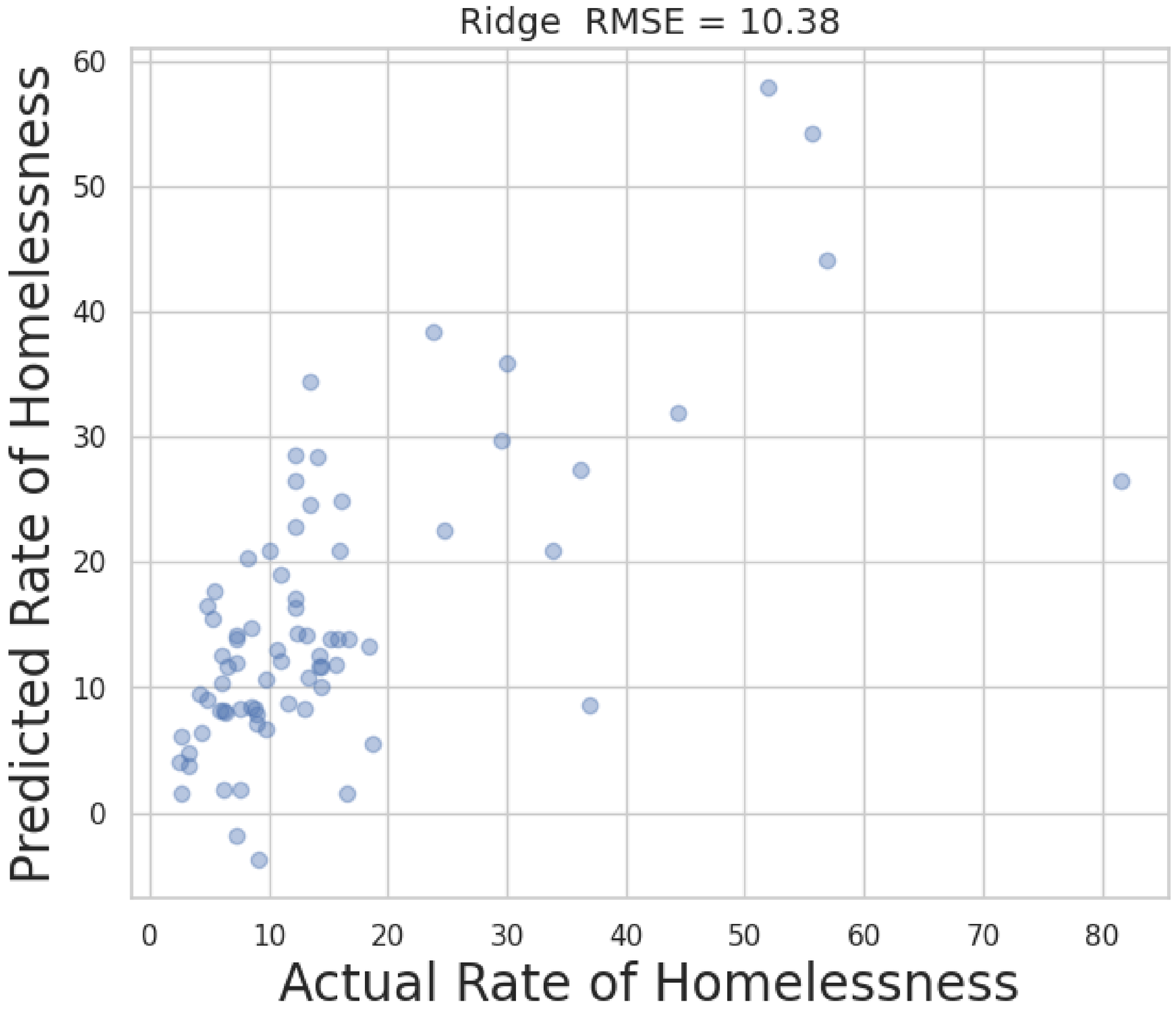
Additional step method:

- K fold cross-validation using Lasso and Ridged on city_or_urban

Lasso

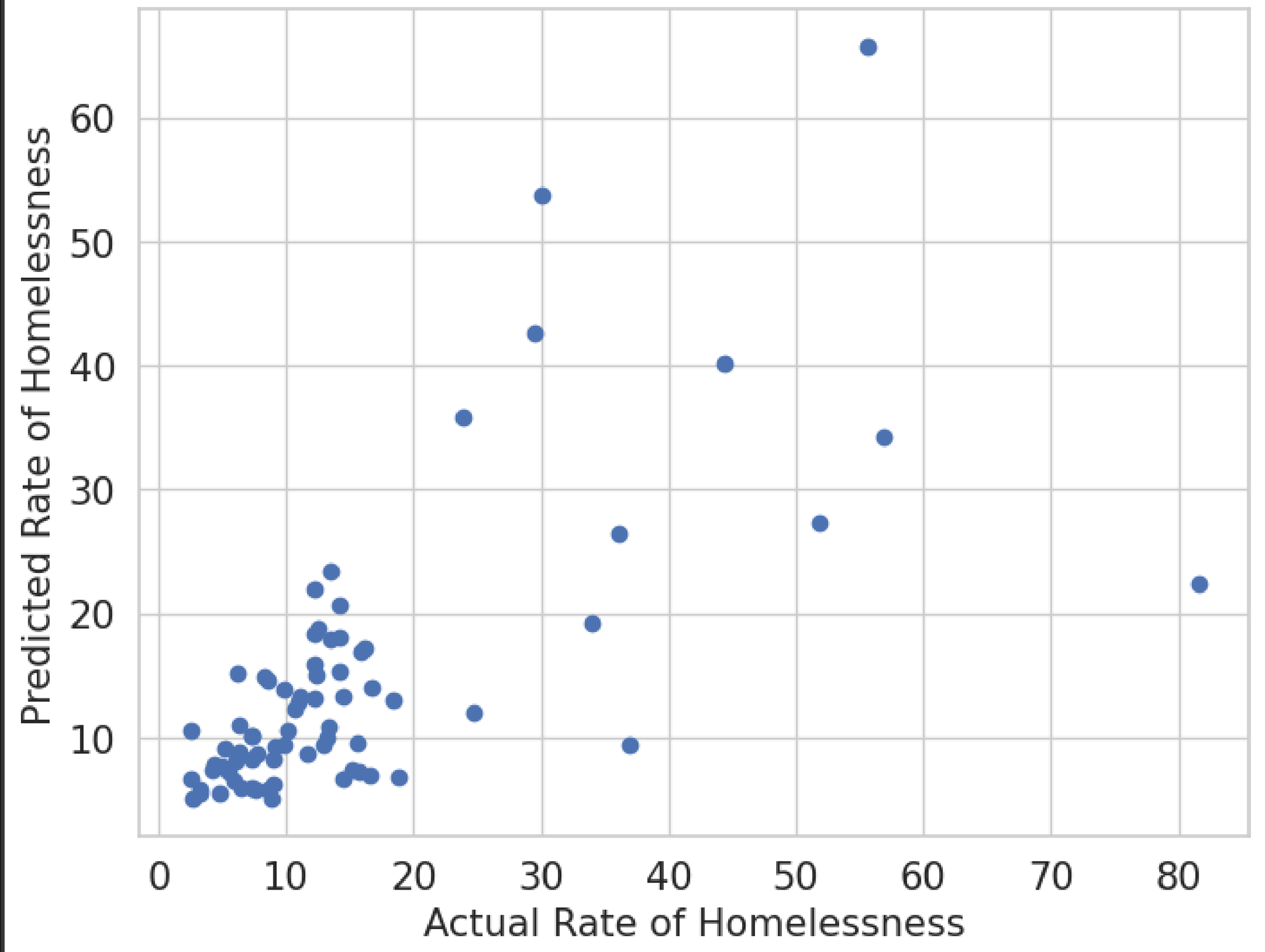

Lasso RMSE = 10.38
Predicted Rate Homeless
Actual Rate of Homelessness
0
10
20
30
40
50
60
70
80

Ridge

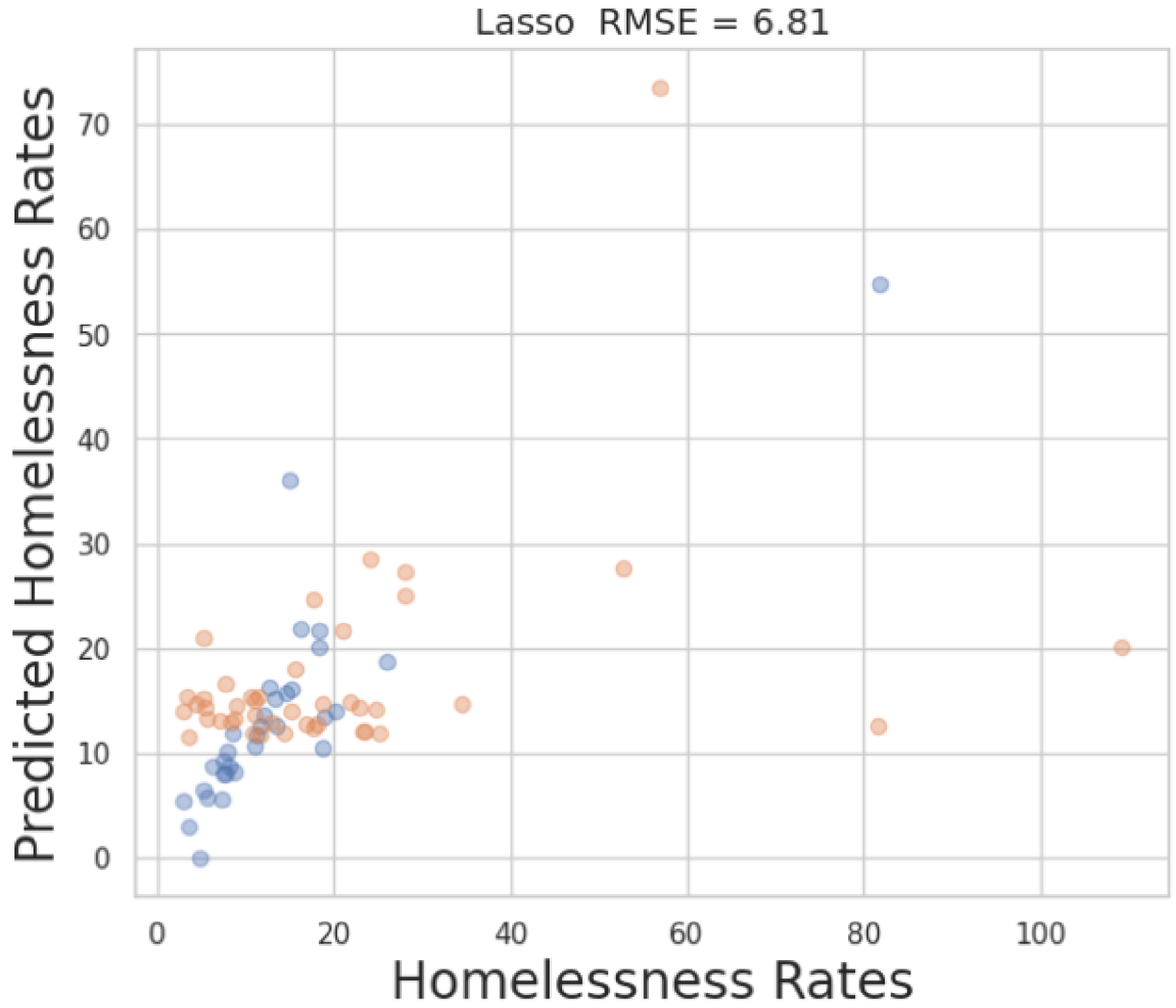

Ridge RMSE = 10.38
Predicted Rate of Homelessness
Actual Rate of Homelessness
0
10
20
30
40
50
60
70
80

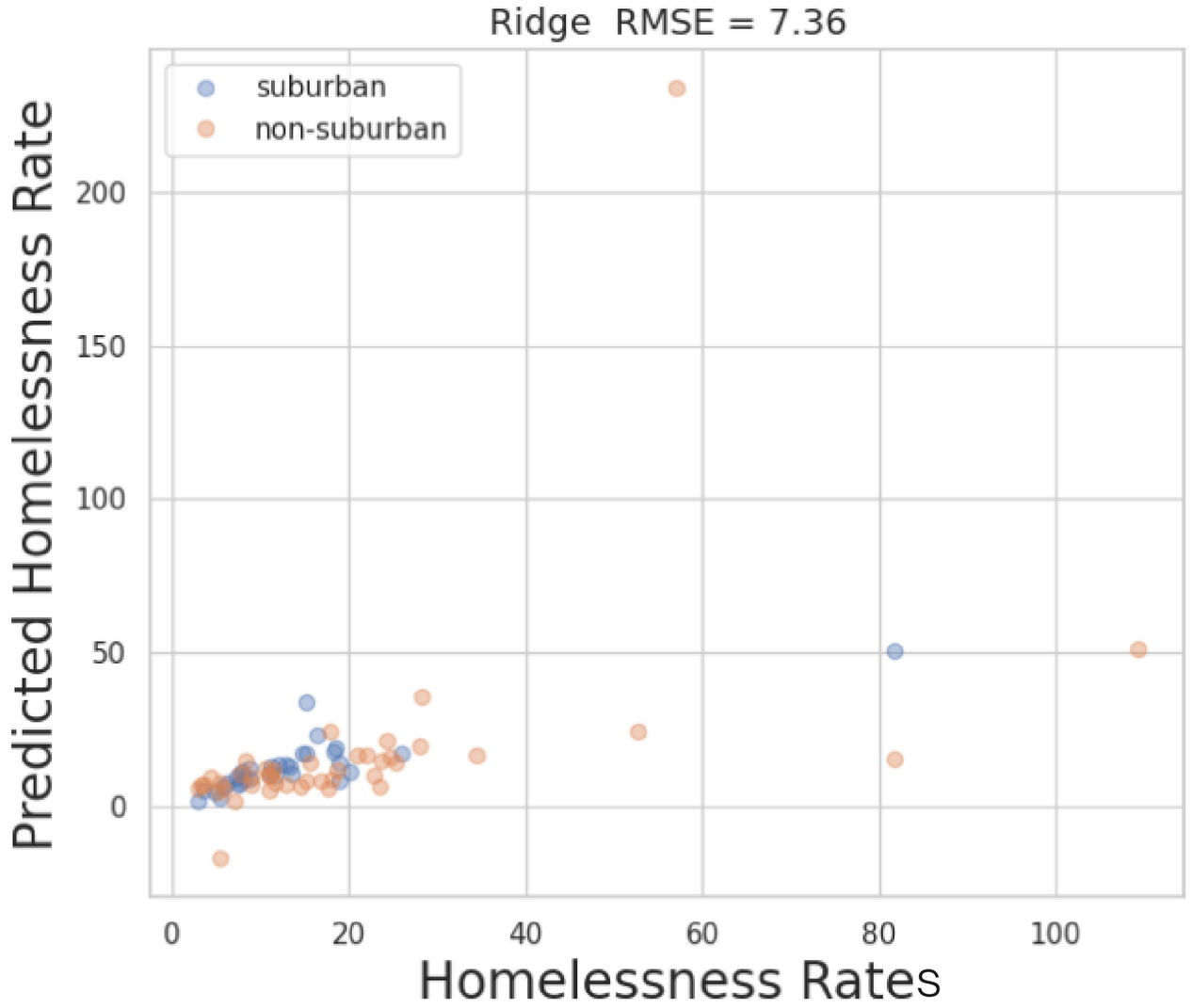
XG Boost

Mean squared error: 10.46



Homelessness on Suburban vs. Non-Suburban Areas





In conclusion, the Lasso, Ridge, and XG Boost models demonstrated similar performance with some variability in their effectiveness. This is evident from our k-fold cross-validation, where the mean RMSE across the three models was approximately 10.5. This indicates that, on average, the models predicted the homelessness rates with an error of around 10.5 percentage points.

Although there is room for improvement, the best-performing models were able to achieve an accuracy of approximately 5 percentage points. However, it is important to note that the majority of non-suburban plot points appeared as outliers, exhibiting high homelessness rates with low predicted rates. This suggests that the models may face challenges in accurately capturing the unique dynamics and factors influencing homelessness in non-suburban areas.

Considering the overall similarity in performance among the models, the choice between them would depend on additional factors such as the data set maybe needing to be larger for better accuracy.