Evaluating alternative modeling approaches to predict Homelessness Rates using Market Factors

Cassidy Denault – Data 3320

Investigating homelessness in the US

- Investigation into homelessness in the US, specifically how community level market factors affect homelessness rates.
- Are there are alternative modeling approaches that outperform the models described in the U.S. Department of Housing and Urban Development (HUD) report.

Data derived from HUD report

- U.S. Department of Housing and Urban Development (HUD)'s 2019 report "Market Predictors of Homelessness" describe a model-based approach to understanding of the relationship between local housing market factors and homelessness.
- "To continue progressing toward the goals of ending and preventing homelessness, we must further our knowledge of the basic community-level determinants of homelessness. The primary objectives of this study are to (1) identify market factors that have established effects on homelessness, (2) construct and evaluate empirical models of community-level homelessness.."
- Can we improve on their model?

HUD's Report

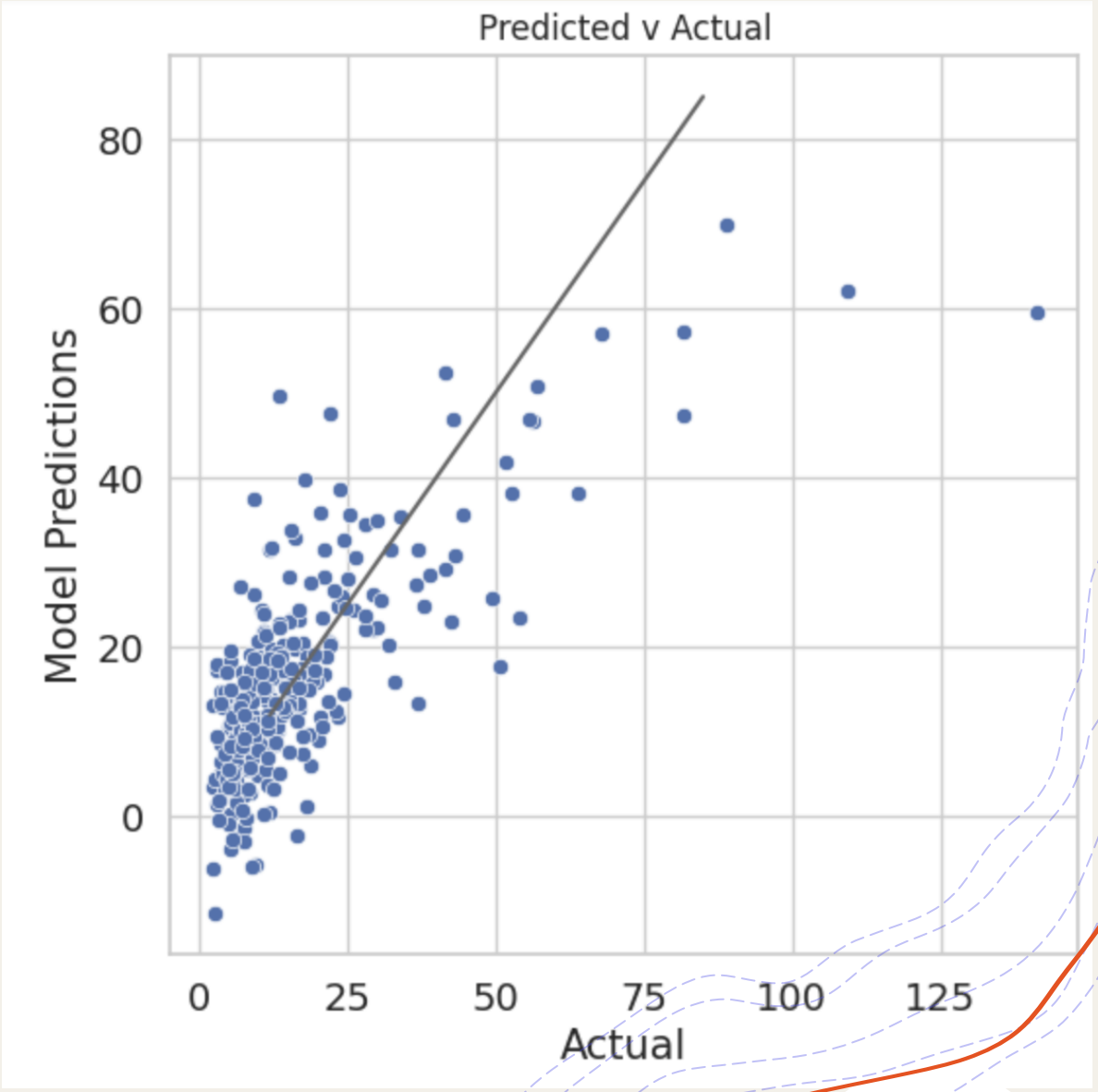
- HUD conducted a literature review to select:
- Independent variables from multiple sources in the areas of housing, economic, safety net, demographic, and climate they believed are associated with homelessness. The report indicates data was carefully chosen for its completeness and usefulness.
- Dependent variables
  - Total number of homeless people per 10,000 population.
  - HUD has estimated the number of people experiencing homelessness—in both sheltered and unsheltered situations—on a single night in the last week of January in approximately 400 Continuums of Care (CoCs).
  - CoCs are local planning bodies responsible for coordinating a full range of homelessness services in a geographic area, which may cover a city, county, group of counties, or an entire state.

Analysis Methods

- To conduct the analysis, the following steps were taken:
  - created test and train splits
  - scaled predictor data
  - created a OLS model
  - used K fold cross validation testing on Ridge, Lasso, and XGBoost models
  - use root mean square error to evaluate effectiveness of a model
- Additional Question: Adding region identifying variable
  - Creating interactions between all other variables and the region identifying variable

OLS model underpredicts higher rates

- RMSE of 8.926
- Actual values of the test data vs what the model predicts when given the test data
- model does an okay job predicting cities on the lower end of rates of homelessness.
- outlier cities with higher rates, our model underpredicts. However, even at the lower end of the homeless rates, the data points are not that close to the identity line. This model is okay, but not super accurate.



K Fold Cross Validation

- K Fold Cross Validation
  - Tests models across multiple iterations with different subsets of the predictive data
  - Access performance of model

- Root Mean Square Error (RMSE) is a commonly used metric to measure the accuracy of a predictive model
  - provides a single value that represents the magnitude of the model's errors, with lower values indicating better accuracy

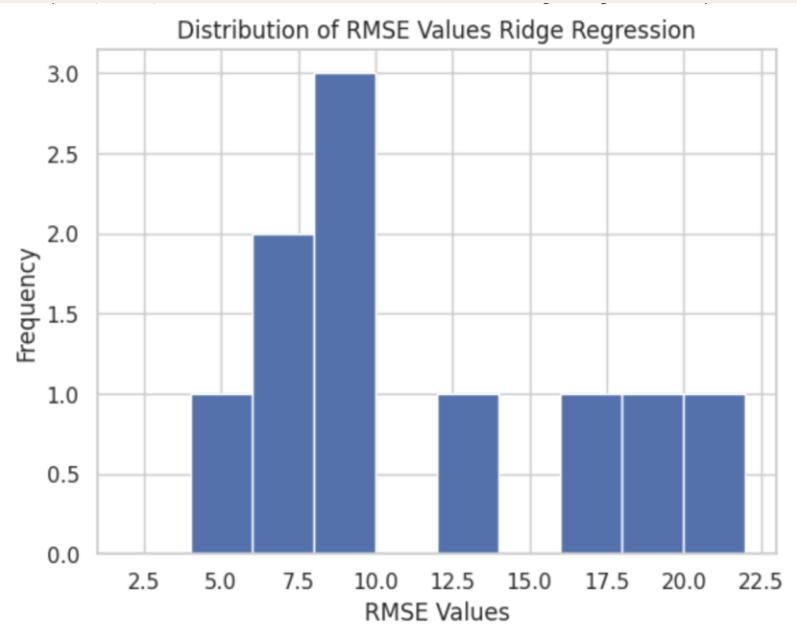
Mean RMSE reveal models' similarity

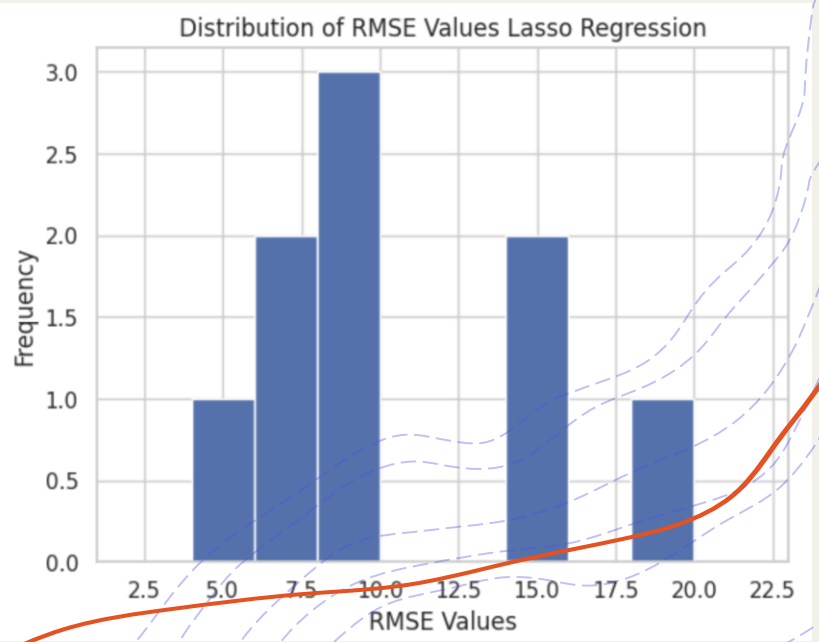
- K Fold Cross Validation
  - Tests models across multiple iterations with different subsets of data
  - Access performance of model

- RMSE of the Lasso, XGBoost, and Ridge values are pretty similar.
- Range from around 5 to 22
- Mean RMSE is ~11 for all of them. The ridge model is a bit lower for min, max, and mean- but not by a significant amount.

Distribution of RMSE reinforce similarities

- Histograms show the number of folds at a certain interval of RMSE
- Again, pretty similar results between lasso and ridge models. The ridge model has slightly more folds at the higher range. The xg boost model also has about the same amounts of folds on the lower end of rmse values as the other models.





Similar effectiveness throughout

- Lasso, Ridge, and XG Boost Models preform similarly and have little variability in their effectiveness.
- On average, models made using the predictor variables and interactions between region and the other variables were off by an average of 11 percent points when predicting. This is a substantial amount.
  - Consider outlier CoCs with uniquely high rates of homelessness.
- At their best these models would only be around 5 percent points off.
- In choosing between the three models, performance is similar across them.
- XG Boost is slightly better.