Homelessness Rates in the U.S

By Bayan Farag

Introduction & Data Used

Introduction

Our project is motivated by the goals of the HUB study which includes

- Identifying market factors that have established effects on homelessness
- Construct and evaluate empirical models of community-level homelessness
- Additional step:investigating whether the density of different areas (such as major cities, suburban regions, and rural areas) has an impact on their respective rates of homelessness

Data Used

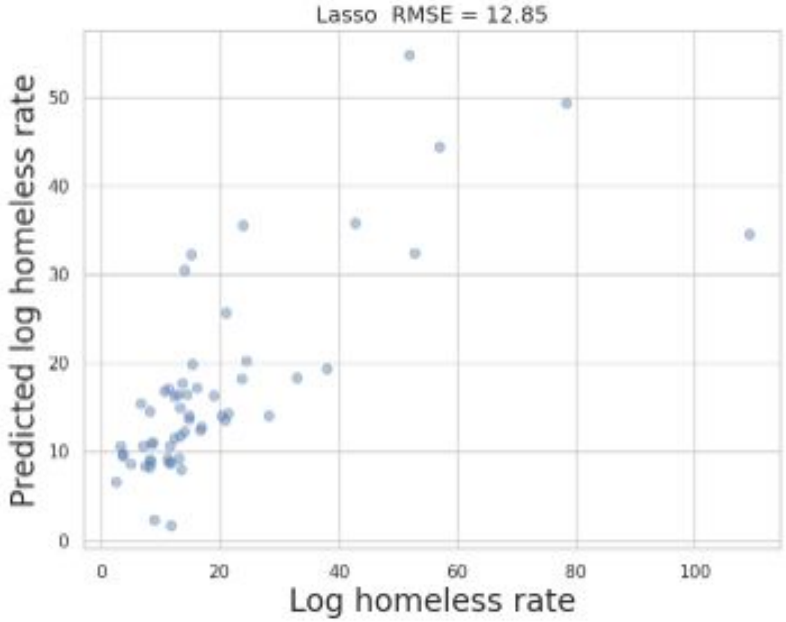
Two sets of data were used; the firsting being factors that influenced homelessness across communities from 2010-2017. The second being a data dictionary which is a centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format.Both which were from the HUB report of 2019

Analysis

STEPS INCLUDE:

- Train & Test Split
- Scaled
- OLS Model
- Ridge
- Lasso
- XGBoost
- RMSE

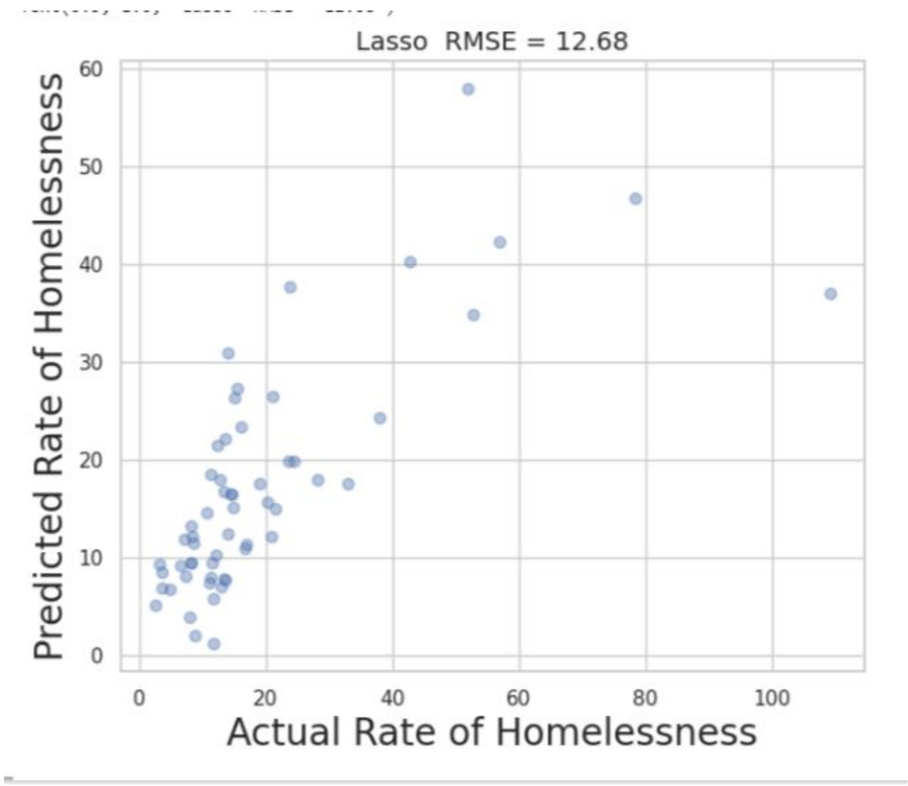
Lasso



	Estimated Coefficient
intercept	12.40
HUD_unit_occupancy_rate	-0.00
average_Jan_temperature	-0.00
average_summer_temperature	-0.00
city_or_urban	0.00
gini_coefficient_2016	0.00
high_housing_density	-0.00
house_price_index_2009	0.00
log_median_rent	0.00
medicare_reimbursements_per_enrollee	-0.00
migration_4_year_change	0.94
net_migration	-0.00
number_eviction	0.00
percent_asian	-0.00
percent_black	-0.64
percent_female_population	-0.00
percent_latino_hispanic	0.00
percent_pacific_islander	0.18
percent_population_0_19	-1.84
percent_population_65_plus	0.00
percentage_excessive_drinking	1.36
percentage_owners_cost_burden_2016	0.43
percentage_renters_severe_cost_burden_2016	0.26
poverty_rate	-0.00
proportion_one_person_households	2.54
rate_unemployment	0.00
rental_vacancy_rate	-0.00

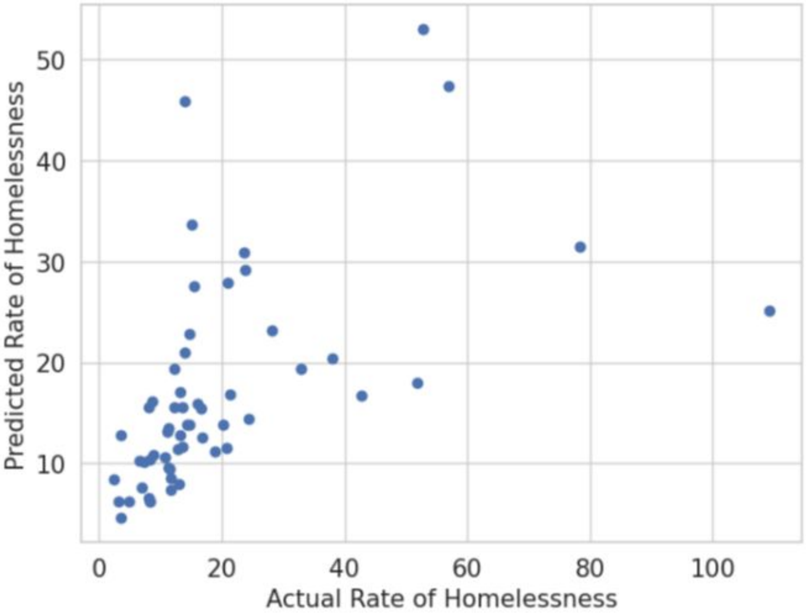
RMSE measures the average squared difference between the predicted values and the actual values

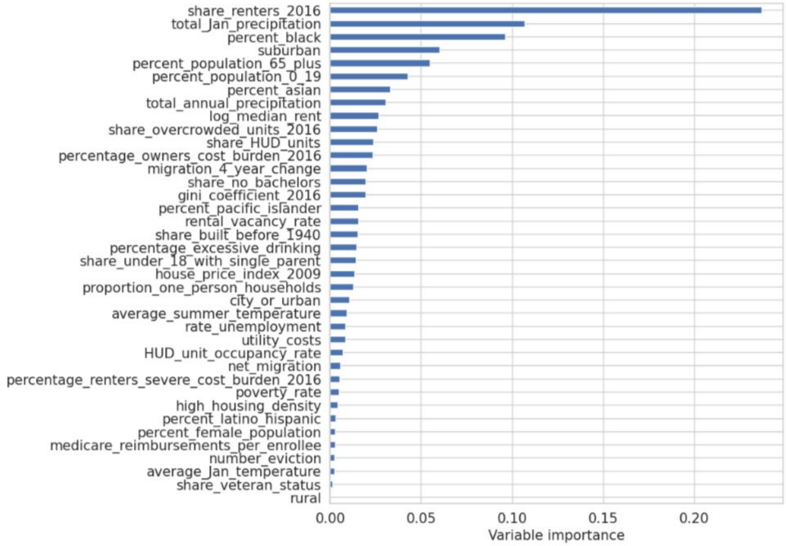
Ridge



	Estimated Coefficient
intercept	12.256
HUD_unit_occupancy_rate	-0.580
average_Jan_temperature	0.144
average_summer_temperature	-0.921
city_or_urban	0.410
gini_coefficient_2016	0.465
high_housing_density	-0.528
house_price_index_2009	1.094
log_median_rent	1.675
medicare_reimbursements_per_enrollee	-0.211
migration_4_year_change	1.393
net_migration	-0.361
number_eviction	0.413
percent_asian	-0.724
percent_black	-1.049
percent_female_population	-1.011
percent_latino_hispanic	0.978
percent_pacific_islander	0.278
percent_population_0_19	-1.253
percent_population_65_plus	0.448
percentage_excessive_drinking	1.001
percentage_owners_cost_burden_2016	1.117
percentage_renters_severe_cost_burden_2016	0.853
poverty_rate	-0.292
proportion_one_person_households	2.115
rate_unemployment	0.640
rental_vacancy_rate	-0.083
rural	0.199

GKBoost

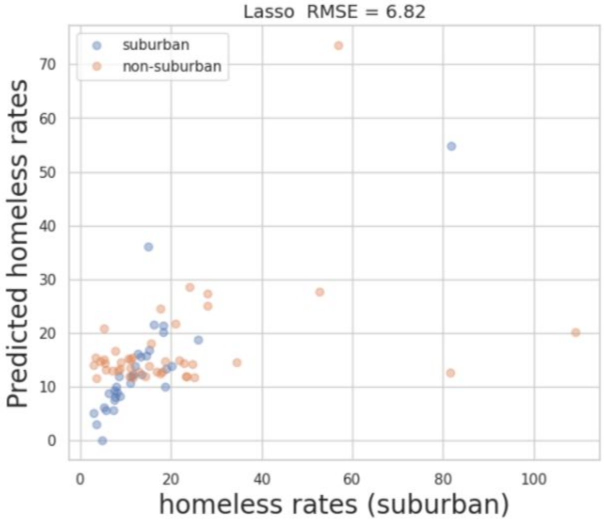


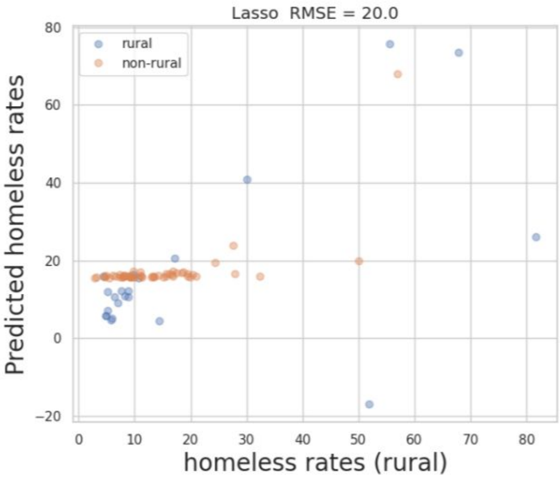


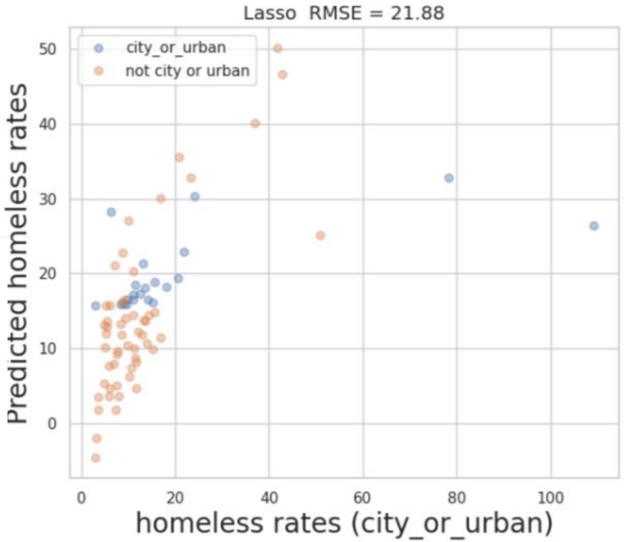
[89]
```
# Computing the RMSE
mean_squared_error(y_test,
```

15.794

Additional Step







Additional step: Investigating whether the density of different areas (such as major cities, suburban regions, and rural areas) has an impact on their respective rates of homelessness and plotted it using the Lasso model.

Conclusion

- All three models Lasso, Ridge, and XGBoost had very similar ranges around 12 percent.
- The additional step had a lot more variations & the ranges for the RMSE were drastic
- Lasso was used on my additional step because it effectively excluding irrelevant variables thought it would be more helpful
- Things to do in the future -> plot ridge & XGboost and compare
- There is not one best regression it depends what you are working on

THANK YOU