Where Should I Live?

Eli Curme

Problem Statement

The "Where Should I Live?" app is an online interactive tool designed to help users find the best deals on homes. There are many online resources already available that can display the cheapest homes in a given area. This app uses machine learning to improve upon this idea. Using data gathering techniques like web scraping, a dataset of over 5,000 U.S. municipalities was assembled, with many features including crime rate per capita, Gini coefficient, and population density. A random forest regression model was trained to predict the average home price per town. The app operates by first filtering the data based on the user's constraints, then sorting by the residuals between predicted average home price and actual average home price. In this way, the app displays the towns that have the best "deals" on home price.

Data Gathering

Web scraping operations:

- Crime data > cityrating.com
- School data > greatschools.org

Downloaded data was found at 3 levels:

- Zip code
- Town
- County

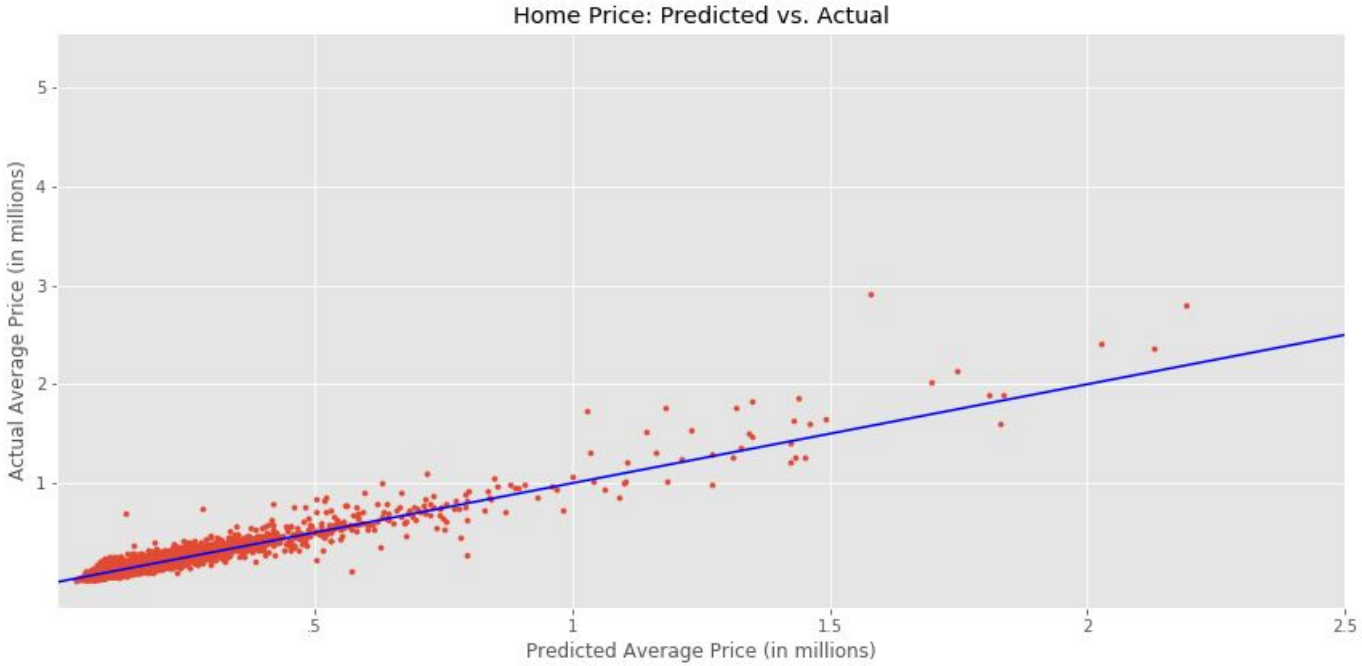
Aggregation done by mean or sum depending on variable

2016 Crime (Actual Data)*	Incidents
Aggravated Assault	17
Arson	0
Burglary	20
Larceny and Theft	81
Motor Vehicle Theft	4
Murder and Manslaughter	0
Rape	3
Robbery	0
Crime Rate (Total Incidents)	126
Property Crime	105
Violent Crime	20

Features

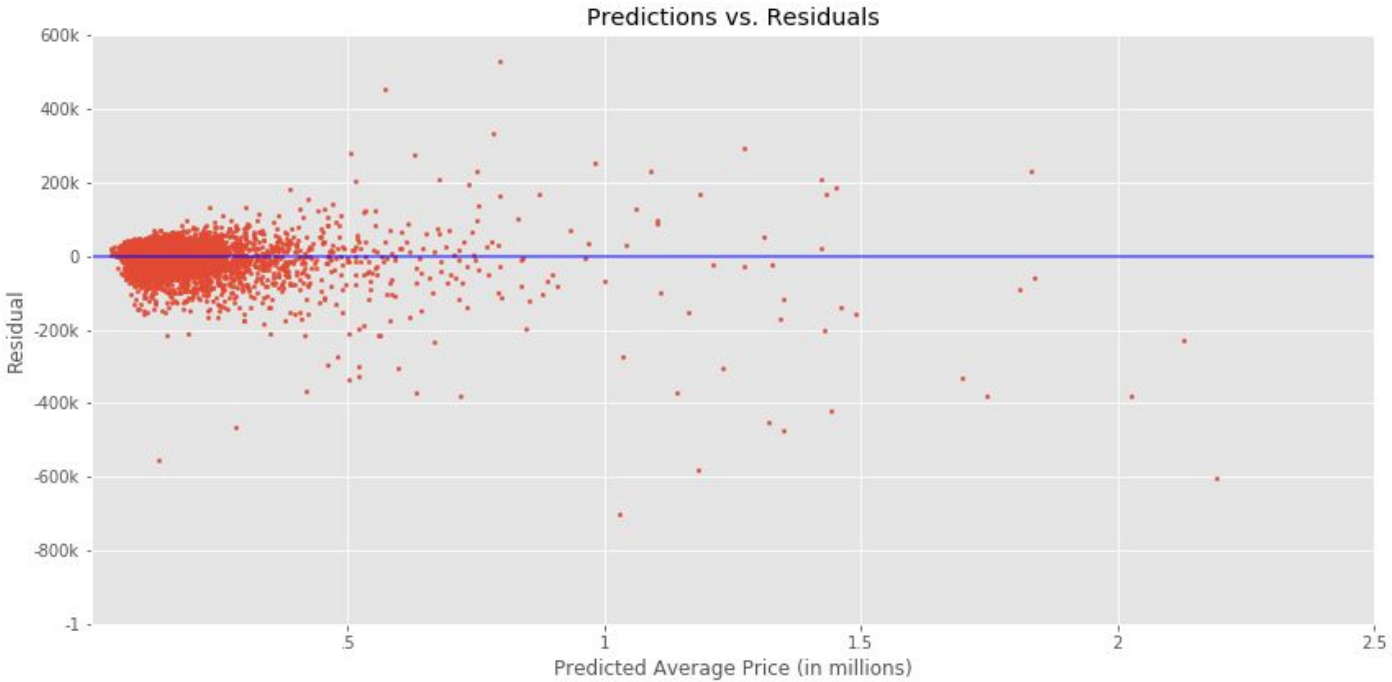
- Median income
- Poverty rate
- High school completion rate
- Crime rate
- Crime rate per capita
- Property crime rate
- Violent crime rate
- Average student to teacher ratio
- Number of households
- Gini Index
- Population
- Population Density
- Unemployment rate
- Latitude & Longitude

Predictions



Random Forest: R2 of 0.85

Residuals


Predictions vs. Residuals
600k
400k
200k
0
-200k
-400k
-600k
-800k
-1
Residual
.5
1
1.5
2
2.5
Predicted Average Price (in millions)

Model Comparison

Most Impactful features:

- Median income
- Longitude
- High school completion rate
- Crime rate

Random Forest R2: 0.85

Linear Regression R2: 0.63

Random Forest allows for more complex feature relationships to be captured

Ex. A certain variable might be a strong predictor in a small town but weak in a large city

Clustering

- Optimal # of clusters is 2 based on silhouette score and inertia.
- Clusters represent urban and non urban areas - Density is defining characteristic

cluster	0	1
med_income	50131.416991	61773.124949
poverty	18.930562	16.456916
hs_completion	84.440810	87.289388
population	8913.701921	35149.411765
density	76.738927	538.410105
lat	37.991461	37.620970
lng	-92.408172	-94.007546
students_per_teacher	15.088422	17.504096
gini	0.457146	0.456076
crime_rate	94.644055	613.724960
crime_rate_pc	0.010762	0.023653
property_crime	84.453683	545.242448
violent_crime	10.190372	68.482512
unemployment_rate	4.552334	4.551322
home_price	151580.640267	287721.327909

Flask Demo

Where Should I Live?

Conclusion / Next Steps

Predictive power was excellent considering the challenges:

- Data combined from many sources in different formats
- Multiple web scraping operations
- Different methods of aggregation

More complete product would have more specific input constraints, a more detailed output, as well as a cleaner, more interactive display.

Sources

- https://data.census.gov/cedsci/
- https://simplemaps.com/data/us-zips
- https://www.cityrating.com/crime-statistics
- https://www.greatschools.org/
- https://www.zillow.com/research/data/
- https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml