

# Where Should I Live?

Eli Curme

# Project Goals

- Assemble dataset of municipal socioeconomic statistics
- Build a model to predict average home price of US towns
- Create interactive tool to show best deals based on predictions

# Data Gathering

Web scraping operations:

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

Downloaded data was found at 3 levels:

- Zip code
- Town
- County

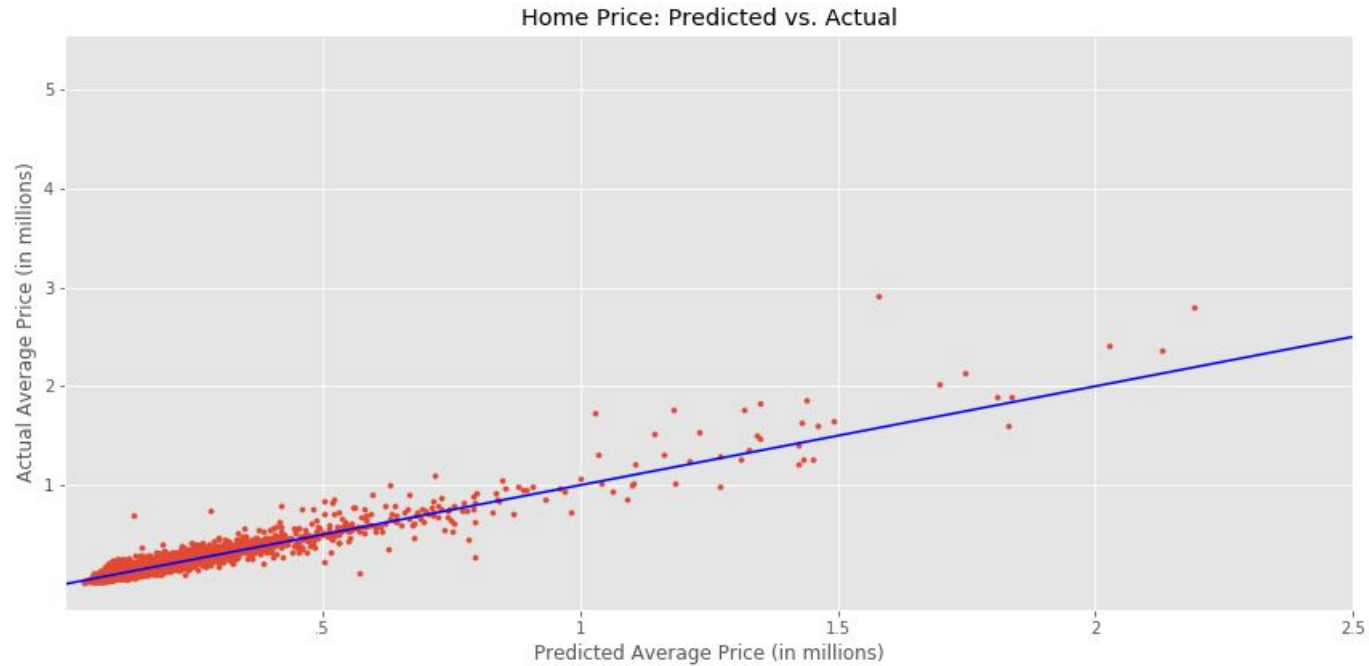
Aggregation done by mean or sum depending on variable

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

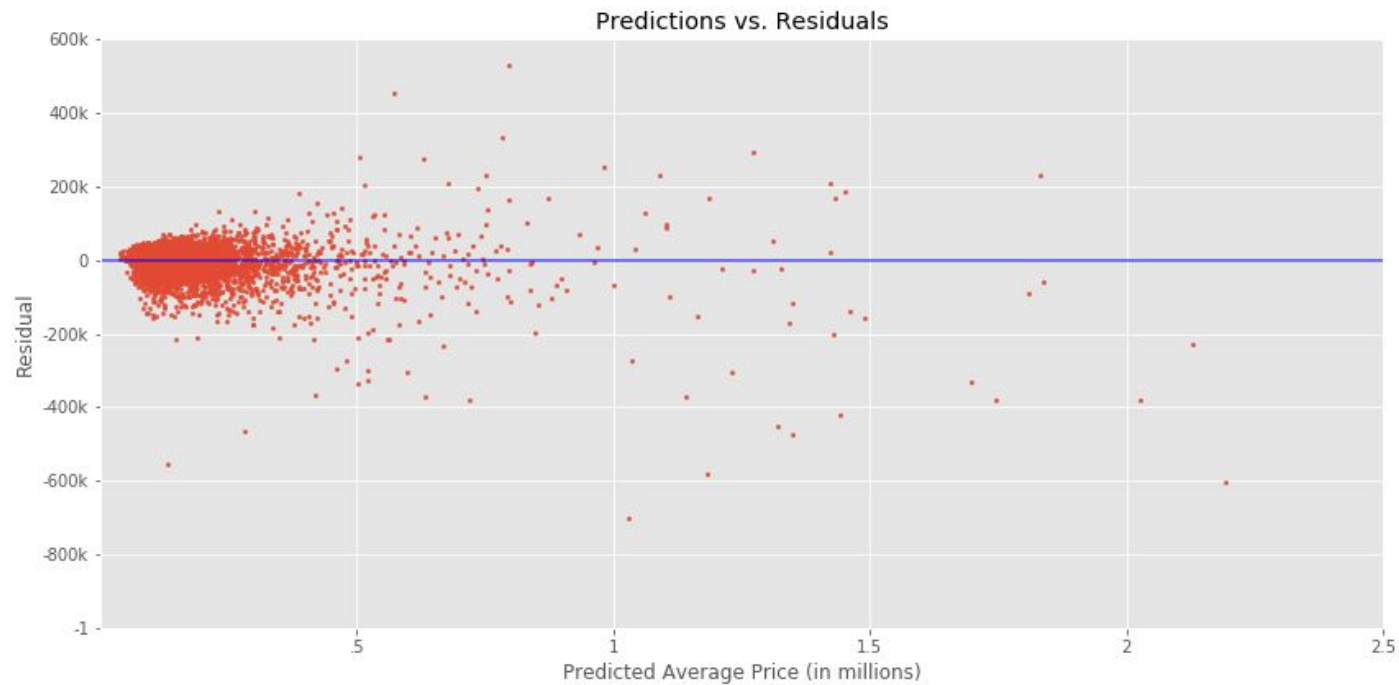
- Median income
- Poverty rate
- High school completion rate
- Crime rate
- Crime rate per capita
- Property crime rate
- Violent crime rate
- Average student to teacher ratio
- Number of households
- Gini Index
- Population
- Population Density
- Unemployment rate
- Latitude & Longitude

# Predictions



Random Forest Regressor:  $R^2$  of 0.85

# Residuals



# Model Comparison

Most Impactful features:

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

Model Scores:

- Random Forest R2: 0.85
- Linear Regression R2: 0.63

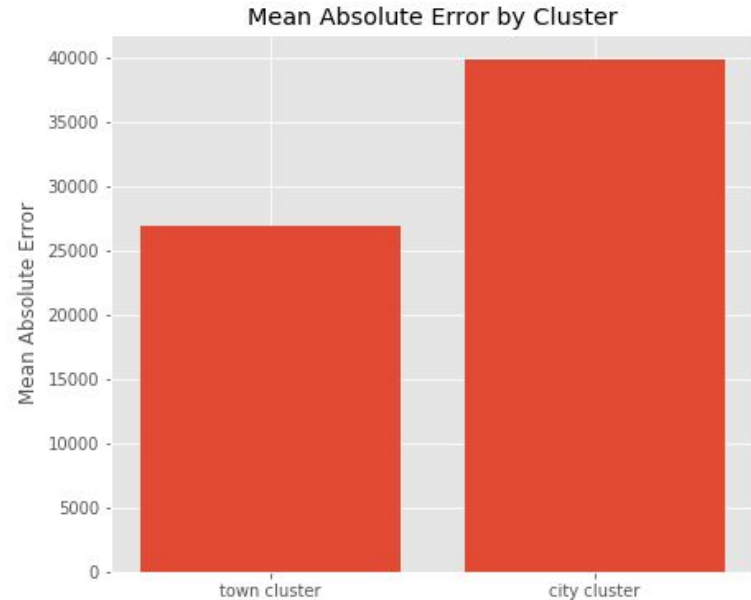
Random Forest allows for more complex feature relationships to be captured

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

# Clustering

cluster	0	1
med_income	50131.42	61773.12
poverty	18.93	16.46
hs_completion	84.44	87.29
population	8913.70	35149.41
density	76.74	538.41
lat	37.99	37.62
lng	-92.41	-94.01
students_per_teacher	15.09	17.50
gini	0.46	0.46
crime_rate	94.64	613.72
crime_rate_pc	0.01	0.02
property_crime	84.45	545.24
violent_crime	10.19	68.48
unemployment_rate	4.55	4.55
home_price	151580.64	287721.33

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





# Best & Worst Deals

## Best

townstate	Los Alamos, New Mexico
preds	696556
home_price	269224
residuals	427332
percent_savings	0.613493
poverty	6.4
med_income	110190
crime_rate_pc	0.00855306
hs_completion	97.6
n_households	7525
population	18356
density	124.2
crime_rate	157
property_crime	139
violent_crime	18
students_per_teacher	13
gini	0.46135
lat	35.8423
lng	-106.291
unemployment_rate	4.58761
state	New Mexico
log_density	4.82189
cluster	0

Name: 2859, dtype: object

## Worst

townstate	Monteagle, Tennessee
preds	79330.3
home_price	205665
residuals	-126335
percent_savings	-1.59252
poverty	30.9
med_income	43094
crime_rate_pc	0.0202522
hs_completion	81.6
n_households	1001
population	2617
density	39.7
crime_rate	53
property_crime	49
violent_crime	4
students_per_teacher	12
gini	0.46135
lat	35.229
lng	-85.8242
unemployment_rate	4.58761
state	Tennessee
log_density	3.68135
cluster	0

Name: 3204, dtype: object

# Flask Demo

Where Should I Live?

# Conclusions

## Challenges:

- Data combined from many sources in different formats
- Multiple web scraping operations
- Different methods of aggregation
- Adjusting target variable and project goals

## Improvements:

- More specific input constraints
- More detailed output
- 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>