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

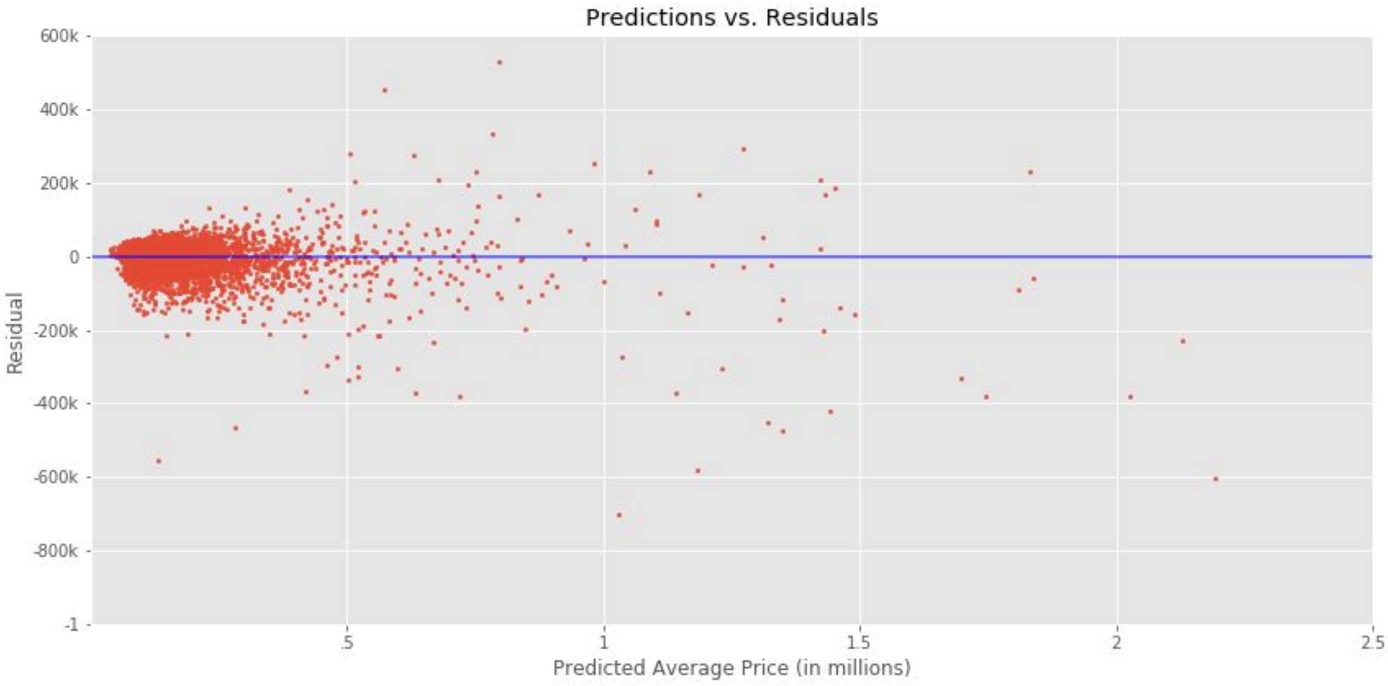
Features

- 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: 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

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

- 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

Best & Worst Deals

townstate	Los Alamos, New Mexico
poverty	6.4
med_income	110190
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
home_price	269224
crime_rate_pc	0.00855306
vcrime_rate_pc	0.000980606
pcrime_rate_pc	0.00757246
people_per_household	2.43934
state	New Mexico
preds	696556
residuals	427332
percent_savings	0.613493

Name: 2859, dtype: object

townstate	Monteagle, Tennessee
poverty	30.9
med_income	43094
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
home_price	205665
crime_rate_pc	0.0202522
vcrime_rate_pc	0.00152847
pcrime_rate_pc	0.0187237
people_per_household	2.61439
state	Tennessee
preds	79330.3
residuals	-126335
percent_savings	-1.59252

Name: 3204, dtype: object

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