REGRESSION Project Write-up

Predicting house price in different cities and states of the USA

Marjan Rezvani

Abstract

House Price prediction is important to drive Real Estate efficiency, also customers, companies such as Zillow, and Banks can use this analysis to get better informed and decide marketing strategy.
The aim of this project is to predict the house prices using Machine learning algorithm Linear Regression considering factors such as lot square feet, number of bedrooms, number of bathrooms, address, city, state, zipcode and region.
I focused on real estate sales price in different cities of The United States, as listed on the Century21 website, and scraped the website using BeautifulSoup library. Then leveraged numerical and categorical feature engineering along with a linear regression to achieve promising results for this problem.

Design

First, I scraped the site of Century21 and extract useful features and then I did regularization and feature engineering on data and analyzed house price predications on different parameters and evaluated my model to get the best result that can best predict home prices given reasonable test/train splits in the data. I used linear regression with K-fold cross validation to predict the Target using the features listed below.

Data

The dataset I made after web scraping, contains 1753 observations and 10 attributes (9 predictors and 1 response). There were several variables that might help determine the sale price of houses such as lot square feet, number of bedrooms, number of bathrooms, address, city, state, and zipcode of the area.

Algorithms

Web Scraping:

Data is collected for various states and their cities from different regions by scraping the Century21 website for houses price.

Feature Engineering

1. Converting categorical features to binary dummy variables
2. Selecting subsets of the total unique values for categorical features that were converted to dummies
3. Converting data into a format required for a number of information processing needs by encoding
4. Doing particular calculations on some features to use them for modeling during EDA
5. Making new variables like region

Analysis:

Seaborn and matplotlib were used to do visualization of the data. The visualization gave some very good information about the trends we see with respect to the type of property.

Models

Machine Learning Linear Regression Model and Polynomial regression were used to find the model with strongest cross-validation performance.
Also tried Lasso and Ridge models, to a get a model that minimizes errors

Model Evaluation and Selection

The entire training dataset was split into train, val, and test, and all scores reported below were calculated with 5-fold cross validation on the training portion only.

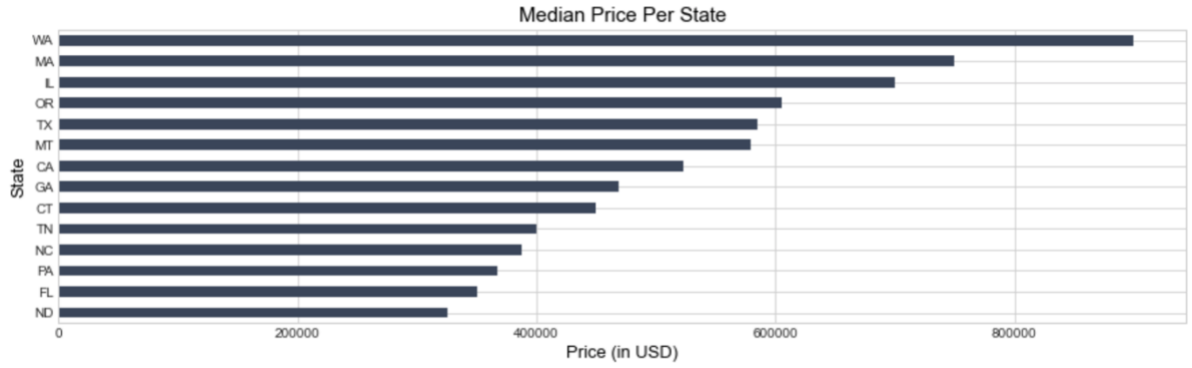
Linear Regression: 0.8437278269210082
Polynomial Regression: 0.9023538924173838

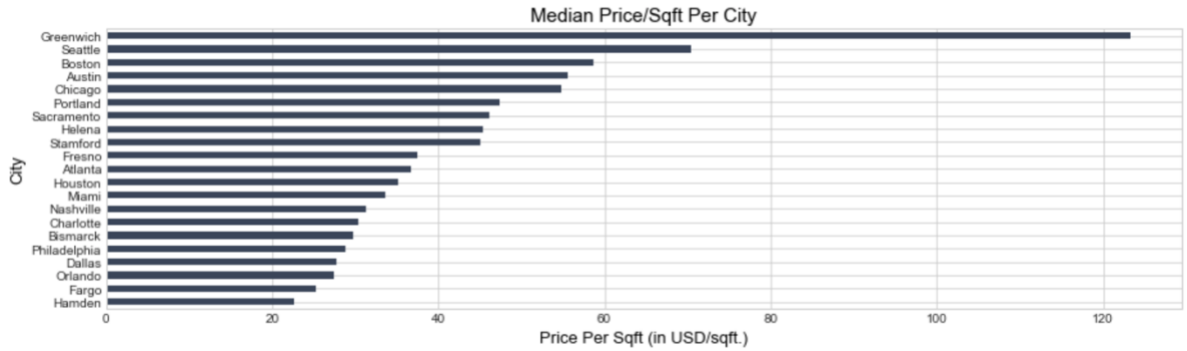
Lasso Regression: 0.7847769127129868
Ridge Regression: 0.712747428001326

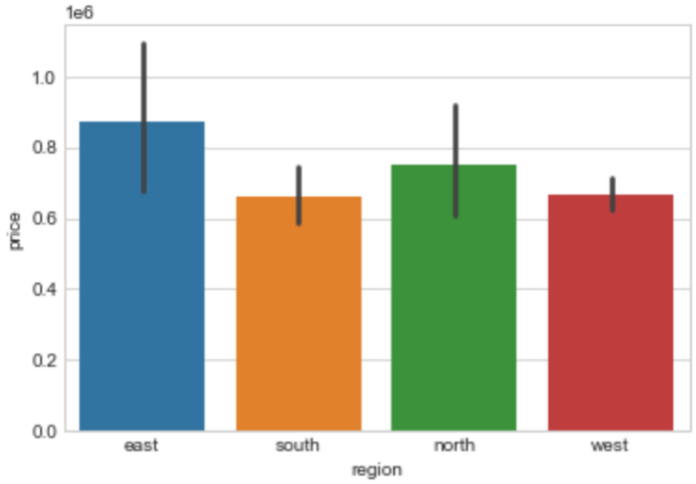
Tools

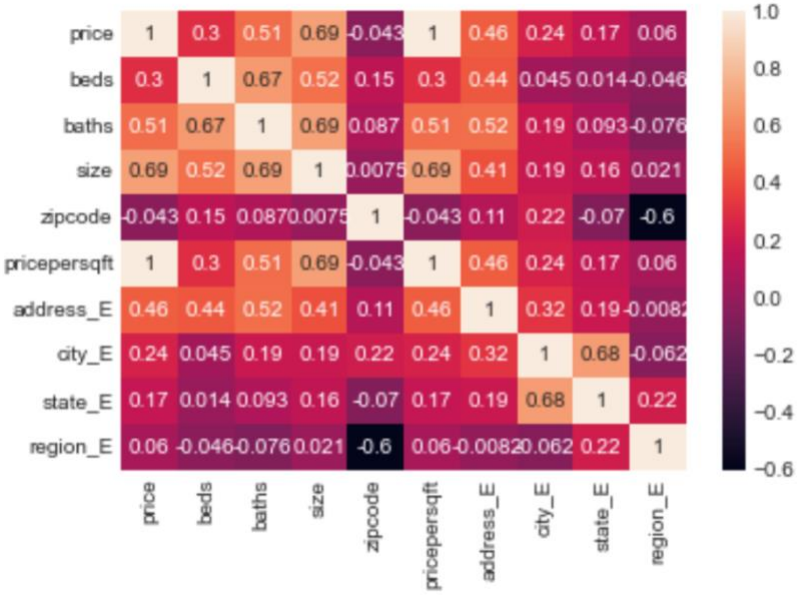
- BeautifulSoup and Requests for web scraping and parsing the data
- NumPy and Pandas for data manipulation
- Scikit-learn for modeling and evaluation
- Matplotlib and Seaborn for plotting

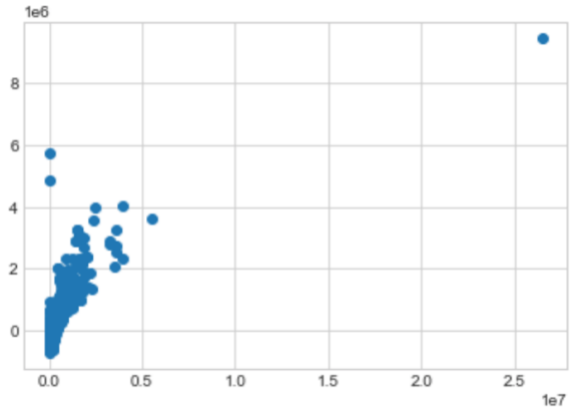
Communication



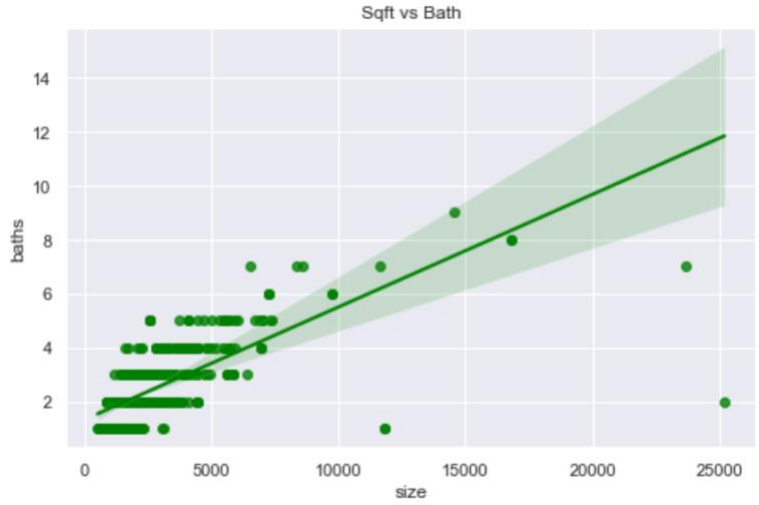

Median Price/Sqft Per City
Greenwich
Seattle
Boston
Austin
Chicago
Portland
Sacramento
Helena
Stamford
Fresno
Atlanta
Houston
Miami
Nashville
Charlotte
Bismarck
Philadelphia
Dallas
Orlando
Fargo
Hamden
City
0
20
40
60
80
100
120
Price Per Sqft (in USD/sqft.)

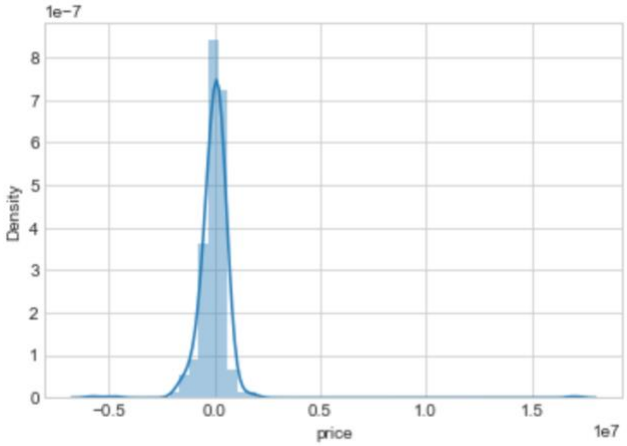

1e6
1.0
0.8
0.6
0.4
0.2
0.0
price
east
south
north
west
region

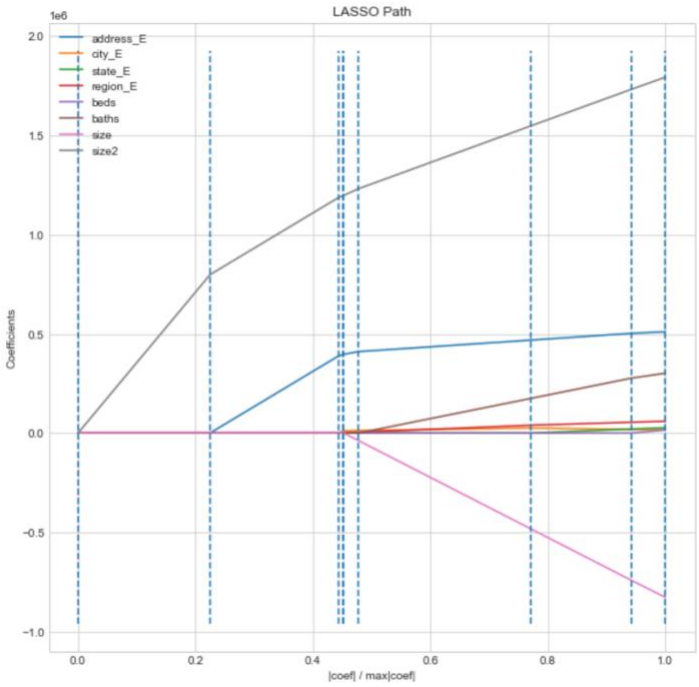

price 1 0.3 0.51 0.69 -0.043 1 0.46 0.24 0.17 0.06
beds 0.3 1 0.67 0.52 0.15 0.3 0.44 0.045 0.014 -0.046
baths 0.51 0.67 1 0.69 0.087 0.51 0.52 0.19 0.093 -0.076
size 0.69 0.52 0.69 1 0.0075 0.69 0.41 0.19 0.16 0.021
zipcode -0.043 0.15 0.087 0.0075 1 -0.043 0.11 0.22 -0.07 -0.6
pricepersqft 1 0.3 0.51 0.69 -0.043 1 0.46 0.24 0.17 0.06
address_E 0.46 0.44 0.52 0.41 0.11 0.46 1 0.32 0.19 -0.008
city_E 0.24 0.045 0.19 0.19 0.22 0.24 0.32 1 0.68 -0.062
state_E 0.17 0.014 0.093 0.16 -0.07 0.17 0.19 0.68 1 0.22
region_E 0.06 -0.046 -0.076 0.021 -0.6 0.06 -0.008 -0.062 0.22 1
price
beds
baths
size
zipcode
pricepersqft
address_E
city_E
state_E
region_E
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


1e6
8
6
4
2
0
0.0
0.5
1.0
1.5
2.0
2.5
1e7


Size vs Price
25000000
20000000
15000000
10000000
5000000
0
price
0
5000
10000
15000
20000
25000
size


Sqft vs Bath
14
12
10
8
6
4
2
baths
0
5000
10000
15000
20000
25000
size


1e−7
8
7
6
5
4
3
2
1
0
Density
−0.5
0.0
0.5
1.0
1.5
price
1e7


LASSO Path
1e6
address_E
city_E
state_E
region_E
beds
baths
size
size2
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
|coef| / max|coef|