California Median House Value Prediction

Author: Yosif Yosif

Date: 07/10/2021

Data Description

The ‘California Housing Prices Data’ data set is a data set by fedesoriano on Kaggle. The data set contains useful information to predict the median housing value in a certain area. It contains 20640 data points.

Features:

Median\_House\_Value: The median house value in each area.

Median\_Income: Median income in each household in that area.

Median\_Age: Median age of the people in a household in each area.

Tot\_Rooms: Total number of rooms in each area.

Tot\_Bedrooms: Total number of bedrooms in each area.

Population: Population in each area.

Households: Number of households in each area.

Latitude: Coordinates.

Longitude: Coordinates.

Distance\_to\_coast: Distance from area to coast.

Distance\_to\_LA: Distance from area to LA.

Distance\_to\_SanDiego: Distance from area to San Diego.

Distance\_to\_SanJose: Distance from area to San Jose.

Distance\_to\_SanFrancisco: Distance from area to San Francisco.

Objective:

The goal for this analysis is to create a model that predicts Median\_House\_Value using all of the other features.

Steps:

1. Importing the required packages into jupyter notebook
2. Importing the California housing pricing data and perform EDA
3. Data Visualization
4. Modeling the data using linear regression, ridge regression, lasso regression, and elastic net.
5. Evaluating the built model using the R^2 score

Explanation:

First thing to do in order to predict the median\_housing\_value is to first clean the data. The data cleaning process would be to first remove any outliers in the features then check if there are any null values in any of the rows. I removed any outliers and then found out that there is not a single missing value. Next step is data visualization, I used the matplotlib library to plot the features in a heatmap and a histogram to see how the data is distributed. Most of the features are either skewed or form a bimodal distribution. All the features except the distances are skewed and the distances are bimodal distributions. Latitude and longitude have some negative values. So, I had to perform the absolute value function on the negative value. After that, I log transformed every feature so that they can be more normal.

Model Testing:

KFold:

Used the following Kfold to split the data into 3 parts:

kf = KFold(shuffle=True, random\_state=72018, n\_splits=3)

Linear Regression:

In linear regression, I first used the MinMaxScaler() to further normalize the data and make the features at an equivalent scale. Next, I performed the linear regression using the pipeline GridSearchCV.

Ridge Regression:

In ridge regression, I first used the MinMaxScaler() to further normalize the data and make the features at an equivalent scale. After that, I added PolynomialFeatures() with degrees of [1,2,3]. Next, I performed the ridge regression using the pipeline GridSearchCV. I used an alpha of np.geomspace(0.001, 10, 200).

Lasso Regression:

In lasso regression, I first used the MinMaxScaler() to further normalize the data and make the features at an equivalent scale. After that, I added PolynomialFeatures() with degrees of [1,2,3]. Next, I performed the lasso regression using the pipeline GridSearchCV. I used an alpha of np.geomspace(0.001, 10, 200). The lasso took a very long time.

Elastic Net:

In elastic net, I first used the MinMaxScaler() to further normalize the data and make the features at an equivalent scale. After that, I added PolynomialFeatures() with degrees of [1,2,3]. Next, I performed the lasso regression using the pipeline GridSearchCV. I used an alpha of np.geomspace(0.001, 10, 200) and an l1 ratio of np.linspace(0.1, 10, 20). The elastic net took the most operation time overall.

Model Scores:

Linear Regression:

Linear regression scored a cross-validation score of 0.66 and an R^2 score of 0.71.

Ridge Regression:

Linear regression scored a cross-validation score of 0.74 and an R^2 score of 0.81.

Lasso Regression:

Linear regression scored a cross-validation score of 0.71 and an R^2 score of 0.78.

Elastic Net:

Linear regression scored a cross-validation score of 0.78 and an R^2 score of 0.85.

Conclusion:

The elastic net performed the best overall with a R^2 score of 0.85. I recommend using Elastic Net if timing is not of importance. On the other hand, if timing is important, I recommend using the Ridge model with a R^2 score of 0.81.

Key Findings:

Overall, features Tot\_Rooms and Tot\_Bedrooms had the most impact on the median\_house\_value and the least coefficient was the Distance\_to\_SanJose feature.

Further work:

Using more types of regression might further enhance the test results. Also, making the model complex by adding more features like Tot\_Bathrooms. Increasing the kfold might also play a part in improving the score.