Subject

We use the existing housing price and related indicators data to generate a machine learning model and use relevant important parameters to predict the housing price;

File

File Name	Function
boston_house_prices.csv	Dataset file that provide data for training and testing
house_prices_predict_model.ipynb	Script files that process datasets, obtain data for the analysis process, and generate ML models
README.md	Script operation description file
dataset_scatter.png	The script generates a chart that describes the relationship between each indicator and house prices
dataset_heatmap.png	The script generates a chart that describes the correlation between each indicator and house prices
prediction_scatter.png	The script generates a chart that describes the dispersion of predicted and actual result
BostonHousePriceLinearModel.skln	Script generated file that saves the trained model

Library

python 3.9.12

Packages	version	Function
pandas	1.4.2	Read CSV files, view and process data
numpy	1.21.5	Provides array processing methods
matplotlib	3.5.1	Make a visual chart
seaborn	0.11.2	Make a visual chart
sklearn	1.0.2	ML tool that provides methods for training models
warnings	N/A	Filtering Warning Messages

Dataset

boston_house_prices is a publicly available dataset. It consists of 506 cases of information collected by the U.S. Census Bureau on home prices in Boston, Massachusetts, USA.

Feature	Explanation
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDIUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX-	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk−0.63)2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000s

Analyze the correlation between each field and MEDV, chose three factors with the max correlation with MEDV: 'LSTAT','RM','PTRATIO', training model.

Summary

Using data visualization to observe data sets is more intuitive than numbers, and it is easier to find potential relationships between data.

For a newcomer, this ML training is very challenging. It has little to do with the previous Python homework, and I have spent a lot of time and effort to learn it better.

I think it's interesting, machine learning can analyze data much more efficiently than humans, and visualization makes it easier for analysts to do their job.

In the process of learning, I saw many more complex cases, and people made many interesting models, including speech recognition and image recognition.

The project that was so challenging to me was just a very basic part of machine learning, and it still had a lot to learn.