Name of dataset: California Housing Prices

(Median house prices for California districts derived from the 1990 census.)

About Dataset

Context

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

Content

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. Be warned the data aren't cleaned so there are some preprocessing steps required! The columns are as follows, their names are pretty self explanitory:

longitude

latitude

housing_median_age

total_rooms

total_bedrooms

population

households

median_income

median_house_value

ocean_proximity

2: The number of columns: 10, 9 feature, 1 predection

3: number of samples :20641 sample

4: the number of samples used in training and testing .

the total number of samples is 20641, and the data is split by 80% for training and 20% for testing, the number of training and testing sets can be calculated as follows:

Training set size:
Training set size=20641×0.8=16512.8Training set size=20641×0.8=16512.8

Testing set size:
Testing set size=20641×0.2=4128.2Testing set size=20641×0.2=4128.2

Since the total number of samples is an integer, we can approximate the numbers to the nearest integer instead of fractions:

Training set size: 16513 Testing set size: 4128

Therefore, the training set size is 16513, and the testing set size is 4128.

Implementation details:

Number of feature : 9

Names of feature :

1. longitude
2. latitude
3. housingMedianAge
4. totalRooms
5. totalBedrooms

6. population
7. households
8. medianIncome
9. oceanProximity