California Housing Data (1990)

**California Housing Price Prediction**

Source

This is the dataset used in this book: <https://github.com/ageron/handson-ml/tree/master/datasets/housing> to illustrate a sample end-to-end ML project workflow (pipeline). This is a great book !

The data is based on California Census in 1990.

About the Data (from the book):

"This dataset is a modified version of the California Housing dataset available from Luís Torgo's page (University of Porto). Luís Torgo obtained it from the StatLib repository (which is closed now). The dataset may also be downloaded from StatLib mirrors.

The following is the description from the book author:

This dataset appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal. They built it using the 1990 California census data. It contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

The dataset in this directory is almost identical to the original, with two differences: 207 values were randomly removed from the total\_bedrooms column, so we can discuss what to do with missing data. An additional categorical attribute called ocean\_proximity was added, indicating (very roughly) whether each block group is near the ocean, near the Bay area, inland or on an island. This allows discussing what to do with categorical data. Note that the block groups are called "districts" in the Jupyter notebooks, simply because in some contexts the name "block group" was confusing."

About the Data (From Luís Torgo page):

[http://www.dcc.fc.up.pt/%7Eltorgo/Regression/cal\_housing.html](http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html)

This is a dataset obtained from the StatLib repository. Here is the included description:

"We collected information on the variables using all the block groups in California from the 1990 Cens us. In this sample a block group on average includes 1425.5 individuals living in a geographically co mpact area. Naturally, the geographical area included varies inversely with the population density. W e computed distances among the centroids of each block group as measured in latitude and longitude. W e excluded all the block groups reporting zero entries for the independent and dependent variables. T he final data contained 20,640 observations on 9 variables. The dependent variable is ln(median house value)."

End-to-End ML Project Steps (Chapter 2 of the book)

1. Look at the big picture
2. Get the data
3. Discover and visualize the data to gain insights
4. Prepare the data for Machine Learning algorithms
5. Select a model and train it
6. Fine-tune your model
7. Present your solution
8. Launch, monitor, and maintain your system

The 10-Step Machine Learning Project Workflow

1. Define business object
2. Make sense of the data from a high level
   * data types (number, text, object, etc.)
   * continuous/discrete
   * basic stats (min, max, std, median, etc.) using boxplot
   * frequency via histogram
   * scales and distributions of different features
3. Create the training and test sets using proper sampling methods, e.g., random vs. stratified
4. Correlation analysis (pair-wise and attribute combinations)
5. Data cleaning (missing data, outliers, data errors)
6. Data transformation via pipelines (categorical text to number using one hot encoding, feature scaling via normalization/standardization, feature combinations)
7. Train and cross validate different models and select the most promising one (Linear Regression, Decision Tree, and Random Forest were tried in this tutorial)
8. Fine tune the model using trying different combinations of hyperparameters
9. Evaluate the model with best estimators in the test set
10. Launch, monitor, and refresh the model and system