California Housing Prices Prediction

Problem Statement

The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

Context and Content

The project aims at building a model of housing prices to predict median house values in California using the provided dataset. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

Districts or block groups are the smallest geographical units for which the US Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset.

Data Source

We will use California Housing data as an example. It contains data drawn from the 1990 U.S. Census: related literature: Pace, R. Kelly, and Ronald Barry, "Sparse Spatial Autoregressions," Statistics and Probability Letters, Volume 33, Number 3, May 5 1997, p. 291-297.* *The following is the description from the book author:*

We collected information on the variables using all the block groups in California from the 1990 Census. In this sample a block group on average includes 1425.5 individuals living in a geographically compact area. Naturally, the geographical area included varies inversely with the population density. We computed distances among the centroids of each block group as measured in latitude and longitude. We excluded all the block groups reporting zero entries for the independent and dependent variables. The final data contained 20,640 observations on 9 characteristics.

The dataset sourced from https:
//www.kaggle.com/harrywang/housing
And consist of two datasets: anscombe.csv, housing.csv

Constraints and Scope

This kernel is divided into 7 parts. The most important part of "house price prediction" is knowing the data we are working with. Almost 75% of the prediction effort goes into getting familiar with data, data cleaning and making the data ready for machine learning algorithms. So, broadly you can divide the seven substeps into 2 major categories.

Working with data

- Get to know your data. The house.csv and anscombe.csv will be imported and cleaned via Python. Missing values will be handled appropriately based on specific factors.
- Data Cleaning. Categorical variables will all be encoded to numerical variables using Ordinal Encoding, One Hot Encoding, or dummy variable encoding techniques, based on if the features have natural rank ordering or not.
- Scaling the data into machine learning readable format. Dividing the data into train/test. The cleaned dataset will be explored visually in order to find interesting trends/correlations in the data. Pairs of columns with correlation coefficient higher than a threshold will be reduced to only one in order to avoid multicollinearity.
- Applying machine learning algorithms. Testing the effectiveness of machine learning Algorithms. Trying different Algorithms. Multiple models including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors will be used to train the aug_train.csv dataset after reducing multicollinearity from both the house.csv and anscombe.csv.

Apply the trained models on the house.csv dataset, compare the accuracy of each model and pick up the one with the highest accuracy.

Note:

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."

Features

Data consists of 20640 rows and 10 features:

1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

Criteria for Success (Inspiration)

- *median_house_value* is our target feature, we will use other features to predict it.
- The task is to predict how much the houses in particular block cost (the median) based on information of blocks location and basic socio demographic data

Approach

Multiple steps will be taken to build a predictive model for this project as well as to analyze the resulting predictions.

1. Build a model of housing prices to predict median house values in California using the provided dataset.
2. Train the model to learn from the data to predict the median housing price in any district, given all the other metrics.
3. Predict housing prices based on median_income and plot the regression chart for it.

Deliverables

The final draft of the project will be presented in the form of a slide deck and formal project report. Jupyter Notebooks will be delivered detailing each step taken and code written for the analysis of the project. A Github repository for the project will be created as well.