California Housing Price Prediction

Akbanu Tleubayeva

TOC

Overview

Background of 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.

Problem Objective :

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.

Dataset description

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Data Sources:

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

Download dataset from here https://www.kaggle.com/camnugent/california-housing-prices

Content

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

median_house_value **is our target feature, we will use other features to predict it.**

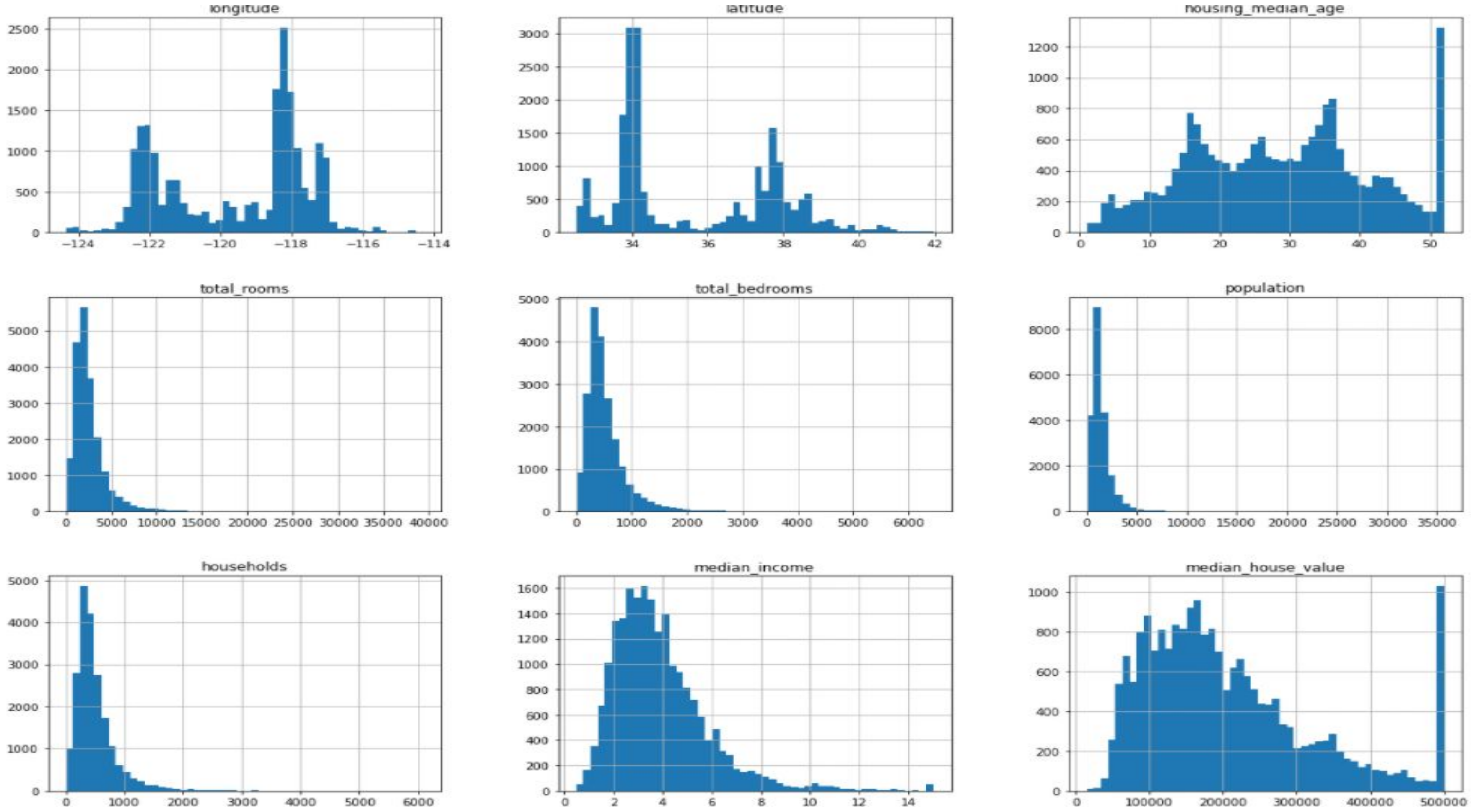
The task is to predict how much the houses in particular block cost (the median) based on information of blocks location and basic socio demographic data

Key Findings

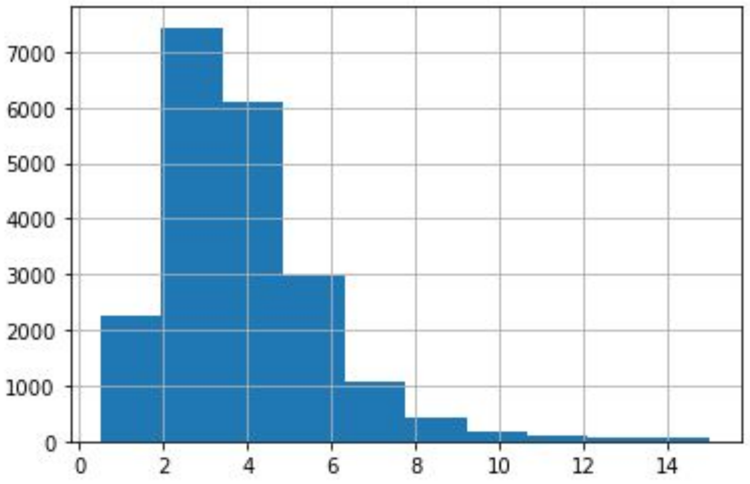
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

Column `total_bedrooms` has about 200 missing values; `ocean_proximity` is not numerical data.

Visualisation and Missing Values Treatment

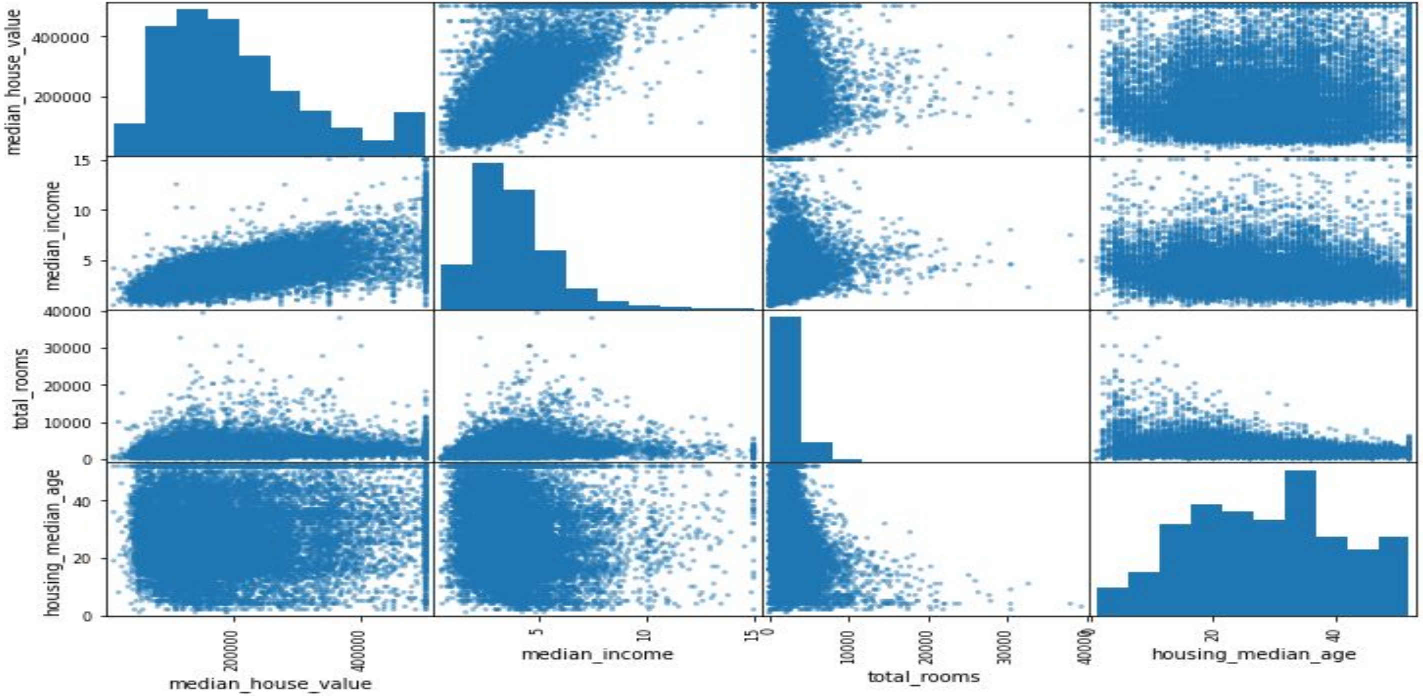

longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value

- According to the pictures, these attributes have very different scales.
- The housing_median_age and the median_house_value were capped. The median_house_value may be a serious problem since it is the label to predict. The Machine Learning algorithms may learn that prices never go beyond that limit. We need to check to see if this is a problem or not. If precise predictions even beyond 500,000 is needed, then we have two options:
    - Option 1: Collect proper labels for the districts whose labels were capped.
    - Option 2: Remove those districts from the dataset.
- Many attributes are right skewed. This may make it a bit harder for some Machine Learning algorithms to detect patterns. We will try transforming these attributes to have more bell-shaped distributions.



From this graph, we can observe median_income is important feature.

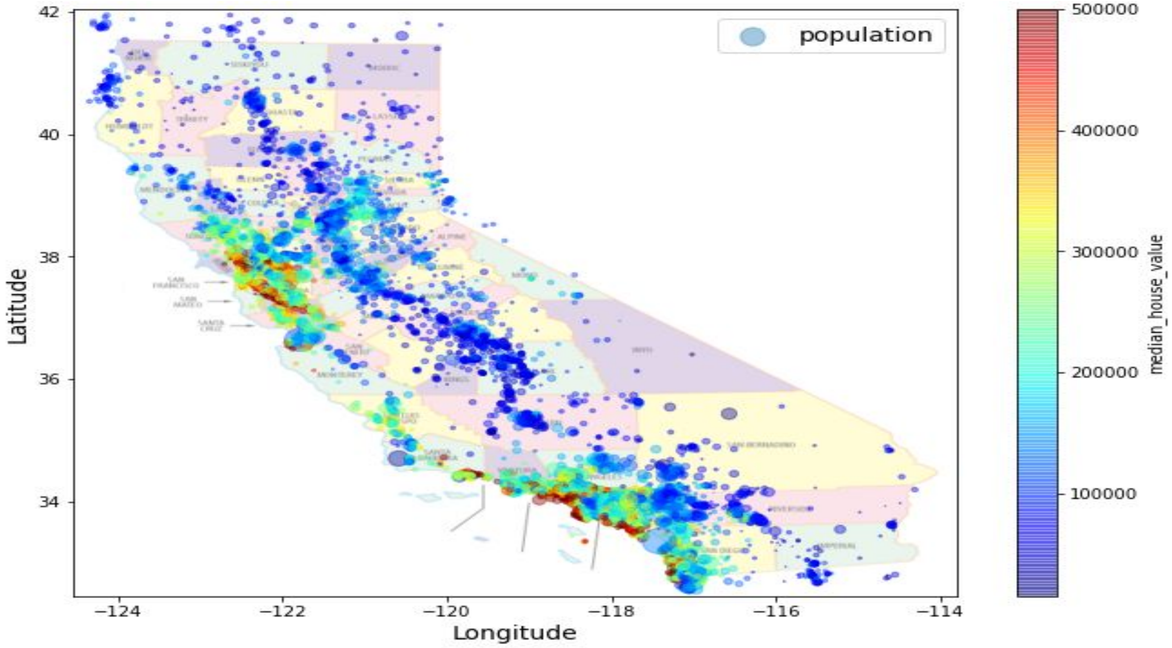
Data Processing



Dependencies between some numerical features.

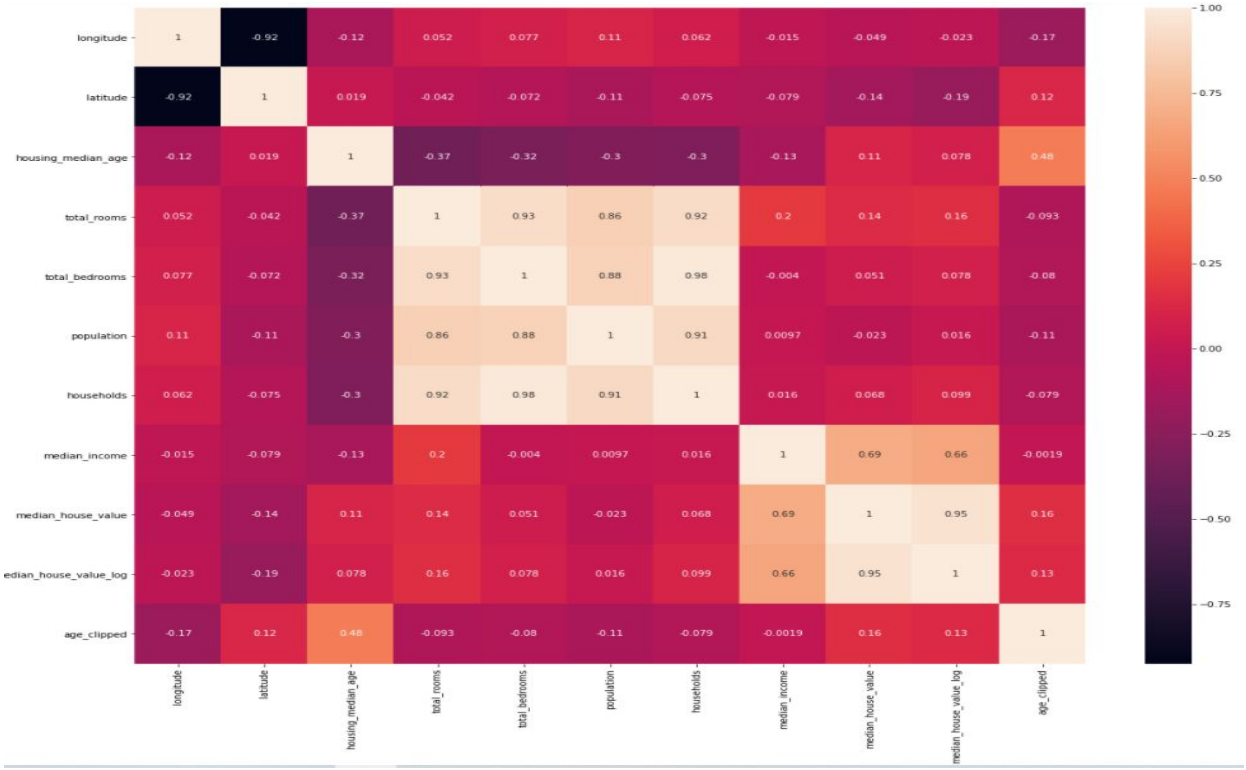
We can see that on any local territory (you can play with local_coord and euc_dist_th) the linear dependencies between variables became stronger, especially median_income_log / median_house_value_log. So the coordinates are a very important factor for our task.

Gain Insight



This image tells that the housing price is very much related to the location and to the population density.

Correlation Coefficient



We can see some patterns here:

- House values are significantly correlated with median income.
- Number of households is not 100% correlated with population, we can try to add average_size_of_household as a feature
- Longitude and Latitude should be analyzed separately (just a correlation with target variable is not very useful)
- There is a set of highly correlated features: number of rooms, bedrooms, population and households. It can be useful to reduce dimensionality of this subset, especially if we use linear models.
- total_bedrooms is one of these highly correlated features, it means we can fill NaN values with high precision using simplest linear regression

Modeling and Analysis

Model Selection

We studied the performance of 3 classification models:

- Linear Regression
- Decision Tree Regression
- Random Forest Regression

The pros and cons of each model are summarized.

Final Model Selection and Hyperparameters Tuning

	Models	Accurasy score
1	Linear Regression	0.605 - 0.627
2	Desision Tree Regression	0.593 - 0.642
3	Random Forest Reg	0.426 - 0.815