Class project

COM SCI-X 450.1 Introduction to Data Science

California housing price prediction

by Supervised Machine Learning

Boonsita Wetakarn

UCLA Extension, Fall 2022

California Housing price prediction with Machine Learning

Outline

1. Project Overview
2. Dataset and variable description
3. EDA and Data Visualization
4. Data Transformation
5. Statistical model
6. Evaluate Model Performance
7. Conclusion

1. Project Overview

This project aims to develop the model to predict the median house value using variables such as location, housing age, number of rooms, population and households, income, and proximity to the ocean. The data was published in 1997 in a paper titled Sparse Spatial Autoregressions by Pace, R. Kelly, and Ronald Barry. This led to a business question: Which aspects of housing significantly influence California housing prices?

2. Dataset and variable description

Dataset and variables description

The dataset has 20,640 observations of 10 attributed as described below

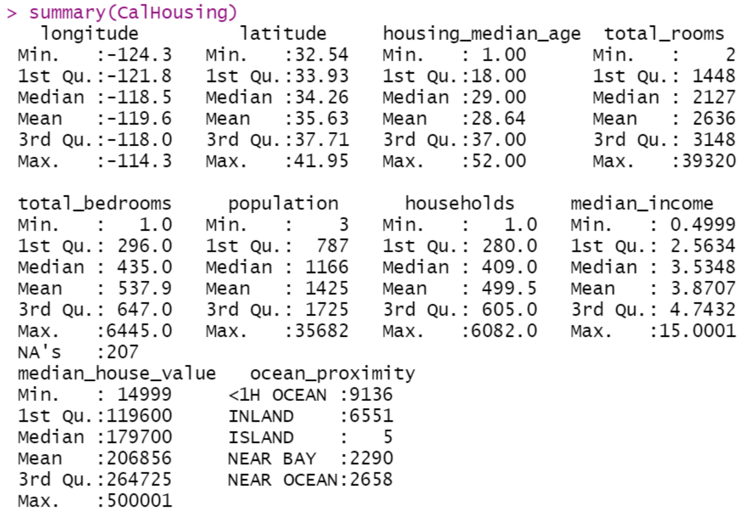
Variables	Class	Description
longitude	Numeric	A measure of how far west a house is; a higher value is farther west.
latitude	Numeric	A measure of how far north a house is; a higher value is farther north
housing_median_age	Numeric	The median age of a house within a block; a lower number is a newer building
total_rooms	Numeric	Total number of rooms within a block
total_bedrooms	Numeric	Total number of bedrooms within a block
population	Numeric	Total number of people residing within a block
households	Numeric	Total number of households, a group of people residing within a home unit, for a block
median_income	Numeric	The median income for households within a block of houses (measured in tens of thousands of US Dollars)
median_house_value	Numeric	Median house value for households within a block (measured in US Dollars)
ocean_proximity	Character	Location of the house w.r.t ocean/sea

Source: https://www.kaggle.com/datasets/camnugent/california-housing-prices

3. EDA and data visualization

3.1 EDA (Statistical summary and correlation analysis)

First, I explore the summary of each variable in the dataset using the *summary ()* function.

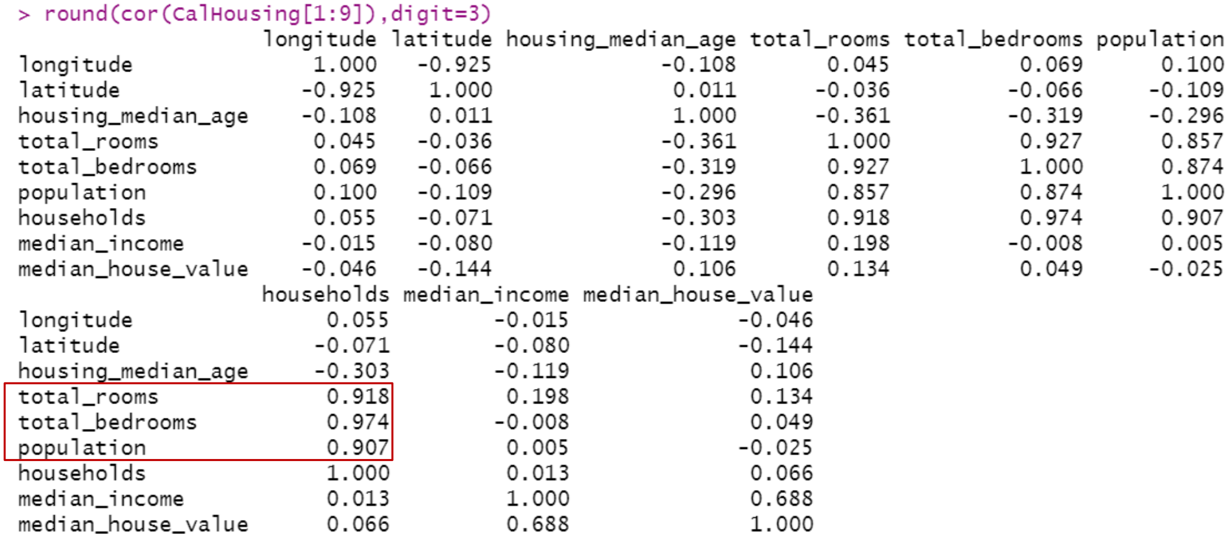


```
> summary(CalHousing)
   longitude        latitude     housing_median_age  total_rooms   
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2  
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448  
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127  
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636  
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148  
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320  
                                                                    
 total_bedrooms     population      households     median_income    
 Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999  
 1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634  
 Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348  
 Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707  
 3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432  
 Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001  
 NA's   :207                                                        
 median_house_value   ocean_proximity
 Min.   : 14999     <1H OCEAN :9136  
 1st Qu.:119600     INLAND    :6551  
 Median :179700     ISLAND    :   5  
 Mean   :206856     NEAR BAY  :2290  
 3rd Qu.:264725     NEAR OCEAN:2658  
 Max.   :500001                      
```

From the observation, all variables are numeric except *ocean_proximity,* the only category variable with five unique values. Also, 207 missing values of total_bedrooms need to be handled in the data transformation process.

In terms of statistical value, I found that medians of all numeric variables are less than means implying the right skewness of the data.

Next step, I perform a correlation analysis on the numeric variable of the data frame.



```
> round(cor(CalHousing[1:9]),digit=3)
                   longitude latitude housing_median_age total_rooms total_bedrooms population
longitude              1.000   -0.925             -0.108       0.045          0.069      0.100
latitude              -0.925    1.000              0.011      -0.036         -0.066     -0.109
housing_median_age    -0.108    0.011              1.000      -0.361         -0.319     -0.296
total_rooms            0.045   -0.036             -0.361       1.000          0.927      0.857
total_bedrooms         0.069   -0.066             -0.319       0.927          1.000      0.874
population             0.100   -0.109             -0.296       0.857          0.874      1.000
households             0.055   -0.071             -0.303       0.918          0.974      0.907
median_income         -0.015   -0.080             -0.119       0.198         -0.008      0.005
median_house_value    -0.046   -0.144              0.106       0.134          0.049     -0.025
                   households median_income median_house_value
longitude               0.055        -0.015             -0.046
latitude               -0.071        -0.080             -0.144
housing_median_age     -0.303        -0.119              0.106
total_rooms             0.918         0.198              0.134
total_bedrooms          0.974        -0.008              0.049
population              0.907         0.005             -0.025
households              1.000         0.013              0.066
median_income           0.013         1.000              0.688
median_house_value      0.066         0.688              1.000
```

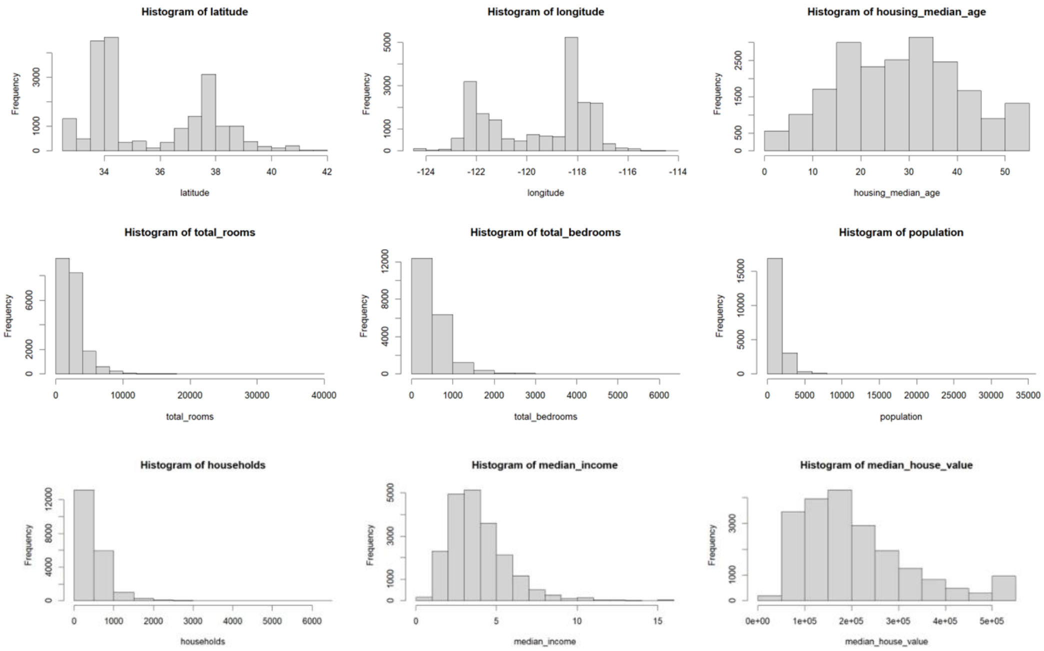
*** Result below after filling NA in total_bedrooms*

From the correlation matrix, it is evident that there is a high correlation between the number of households and other variables, namely total rooms, total bedrooms, and population. This could be an implication of the multicollinearity issue.

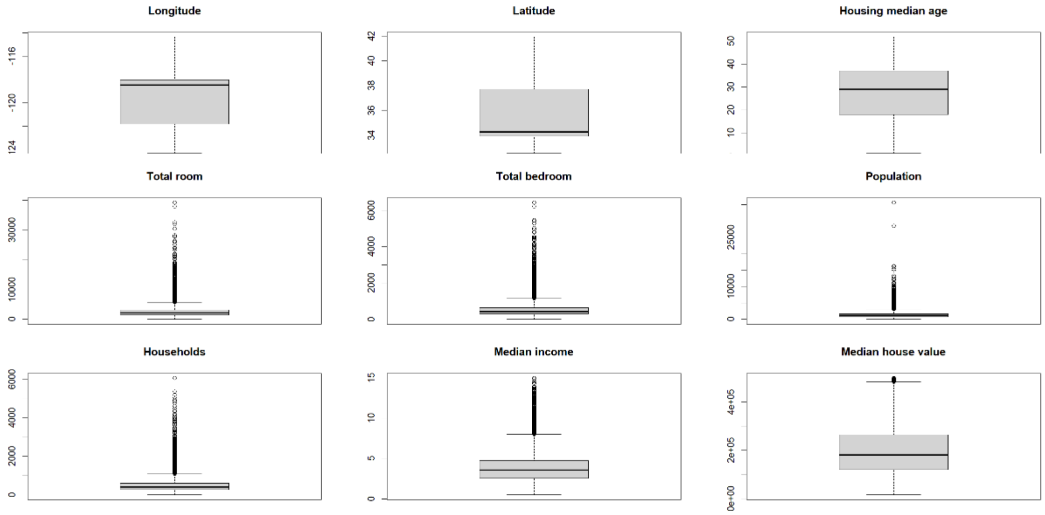
3.2 Data Visualization

The following histograms and boxplots present the distribution and skewness of the numeric variables. Most of them are skewed right to some extent.

3.2.1 **Histograms**

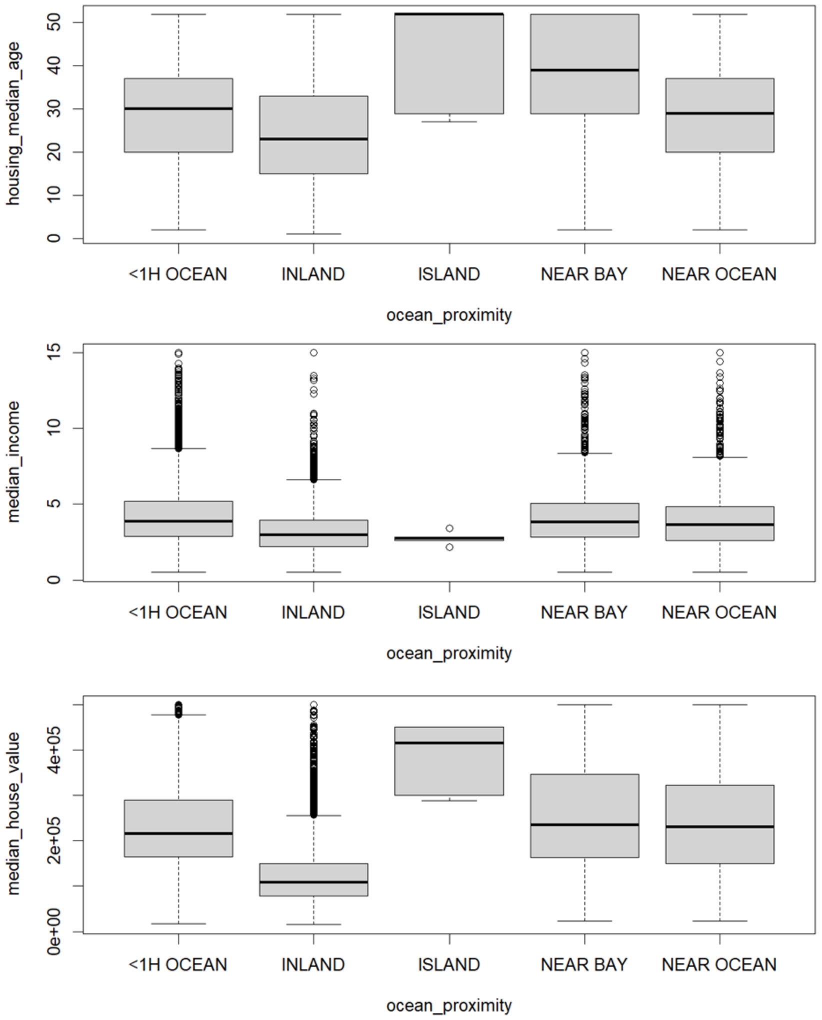


3.2.2 **Box plots**



3.2.3 **Box plot by ocean proximity**

We can see some observations below by examining the box plot by ocean proximity level.



There is a similar pattern in terms of median age and median house value. ISLAND has the highest values in both terms, followed by NEAR BAY, <1H OCEAN, NEAR OCEAN. INLAND has the least values. Conversely, the median income of ISLAND shows the lowest value compared to other proximity groups.

4. Data munging process (Data Pipeline process)

In this project, I have performed data pipeline process to create a cleaned_housing data frame as below:

1. Revise variable type by casting ocean_proximity character variable to the factor class
2. Deal with 207 missing values in the total_bedrooms variable by filling in missing values with median imputation
3. Create binary categorical variables by splitting the ocean_proximity variable into 1 and 0
4. Create new variables mean_number_bedrooms and mean_number_rooms by computing the ratio between the number of bedrooms/rooms and households
5. Perform feature scaling for all numeric variable (except the response variable) to normalize the value of variables that has different magnitudes
6. Remove unused variables in a data frame, namely total_bedrooms, total_rooms, and ocean_proximity
7. Rearrange a new data frame as follow:

“NEAR BAY",
"<1H OCEAN",
"INLAND",
"NEAR OCEAN",
"ISLAND",
"longitude",
"latitude",
"housing_median_age,”
"population",
"households",
"median_income",
"mean_bedrooms",
"mean_rooms",
"median_house_value"

5. Statistical model

In this project, I applied the random forest algorithms, one of the supervised Machine Learning. The algorithm randomly selects several feature variables and learns a decision tree. Ultimately, each tree in the forest votes for the most popular class.

For the first random forest model, “median_house_value” is the response variable, and other numeric variables are feature variables.

Variables	**Variable names**
Feature variables	NEAR BAY, <1H OCEAN, INLAND, NEAR OCEAN, ISLAND longitude, latitude, housing_median_age, population, households, median_income, mean_bedrooms, mean_rooms
Response variables	median_house_value

Dataset is split into two sets for the model fit process. 70% for the training dataset and 30% for the test. I applied the randomForest() function in R for the training dataset as below.

$$rf = randomForest(x = train_x, y = train_y,\\ ntree = 500, importance = TRUE)$$

while “ntree=500” specifies the number of trees to grow and
“importance=TRUE” for the algorithm to assess the importance of the predictors

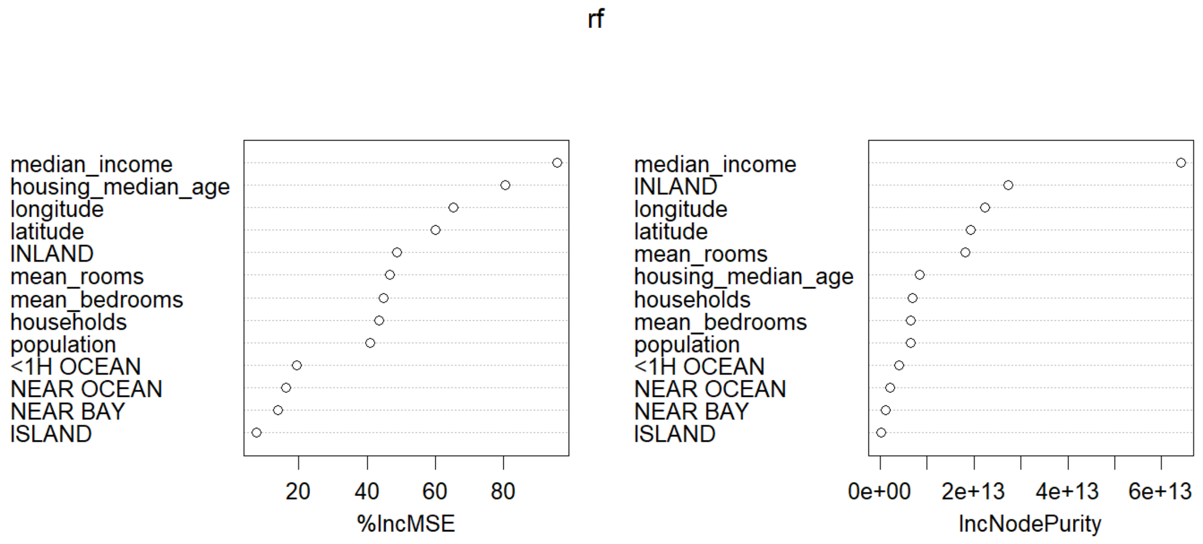
6. Result of performance metric

To evaluate the model performance, I calculate the root mean squared error (RMSE) for the trained model to compare with RMSE for the test set.

Random Forest	Training set	Test set	%Var
RMSE	49,165.43	49,786.27	1.3%

When the model is evaluated on test data, it has an RMSE of 49,786.27. This means that our model's predicted housing price will be ± 49,786.27 in comparison to the actual price on average. Moreover, The RMSE of the test set is slightly greater than the train set. This could be concluded that the model is not overfitting and make a good prediction.

Moreover, I run the **varImpPlot(rf)** function, which provides a dot chart displaying the importance of each variable measured by a random forest algorithm.



From %IncMES and IncNodePurity plots suggest, the following feature variables tend to be a good predictor on the median_house_value.

1. median_income
2. housing_median_age
3. longitude
4. latitude

Then, I re-train the random forest model based on four feature variables as abovementioned. The result of the re-train model compared to the first model presents below.

Random Forest	Training set	Test set	%Var
RMSE – 1st	49,165.43	49,786.27	1.247%
RMSE – 2nd	50,932.61	51,083.02	-0.29%

As seen from the table, the RSME from the re-train model indicates a slightly lower accuracy of the model when compared to the first model. However, the RSME obtained through the training set and the test dataset is closer to the re-train model. Thus, it could be concluded that there is an improvement in the overfitting in the latest model.

Conclusion:

The objective of this project is to develop models to predict the housing price and find the important housing aspects that influence the price.

In order to achieve the objective, I perform data exploration and munging processes by filling the missing value of total_bedroom with a median value, creating new binary categorical variables, and performing feature scaling for all numeric variables.

To fit the model with the dataset, I apply the Random Forest model, which provides a low RMSE in both the training and test set, indicating a good predictor. I also re-train the model by using the variable suggested by varImpPlot(rf) function, which concludes that the median housing price could be predicted by median_income, housing_median_age, and location (longitude and latitude).