Forecast of California House Values Based on the Random Forest Algorithm

By Hongyi Wang

Class: COM SCI X450.1

Instructor: Daniel D. Gutierrez

1. Introduction

In California, because of the high house price, housing problem is one of the most essential social issues. The house value is affected by various factors. Recently, more and more people started to pay attention to the forecast of median house values. Accurately predicting the values of real estate is extremely important for social and economic development and people's life, and can provide reference for economic decision-making.

This report uses random forest algorithm for training and inference. The goal of this report is to predict the of median house values of various blocks groups based on their demographic data.

2. Data Set

2.1 Overview

This report uses the data set from a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in *the Statistics and Probability Letters* journal. This data set consists of demographics data of different block groups across the California in 1990. Each row contains a census data of a block group including variables as follow:

longitude and ***latitude***: the geographical position of a block

housing_median_age: the median housing age of a block

total_rooms and ***total_bedrooms***: the number of total rooms and bedrooms within a block

population: the population of a block

households: the number of households within a block

median_income: the thousands dollar of median income of a block

median_house_value: the dollars of median house value of a block

ocean_proximity: a feature variable indicating a block's distance from ocean including levels of NEAR BAY, 1H OCEAN, INLAND, NEAR OCEAN and ISLAND.

2.2 Exploratory Data Analysis

The first and last six rows of data set are presented as below:

Table 1: First six rows of data set

longitude	latitude	housing_ median_ age	total_ rooms	total_bed rooms	populatio n	household s	median_ income	median_ house_ value	ocean_ proximity
-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY

-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY

Table 2: Last 6 rows of data set

-121.56	39.27	28	2332	395	1041	344	3.7125	116800	INLAND
-121.09	39.48	25	1665	374	845	330	1.5603	78100	INLAND
-121.21	39.49	18	697	150	356	114	2.5568	77100	INLAND
-121.22	39.43	17	2254	485	1007	433	1.7	92300	INLAND
-121.32	39.43	18	1860	409	741	349	1.8672	84700	INLAND
-121.24	39.37	16	2785	616	1387	530	2.3886	89400	INLAND

In order to get a sense for the data classes, range of values for numeric variables, and levels for factor variable, this report presents the summary statistics for all variables in data set:

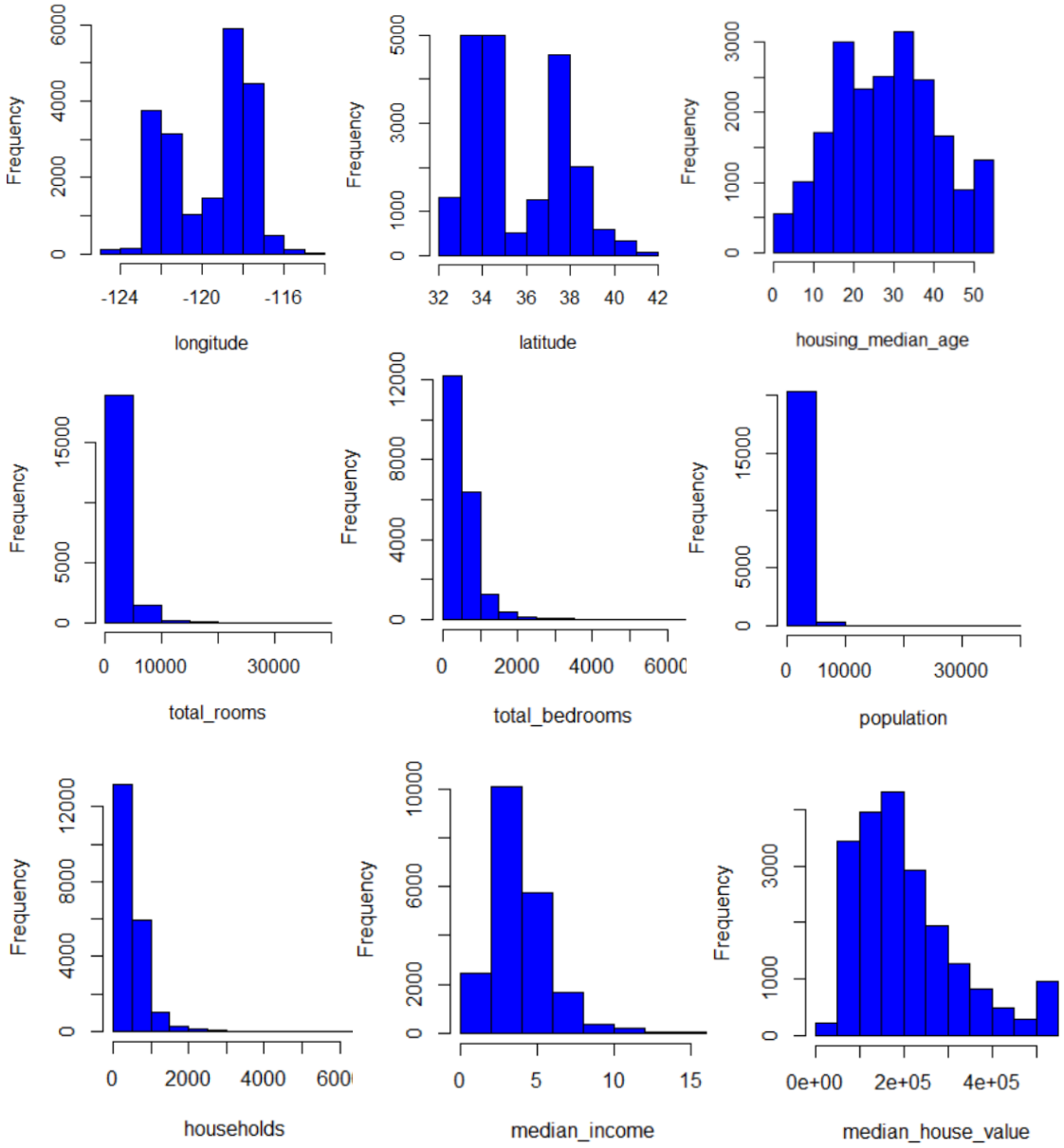
Table 3: Summary statistics

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

2.3 Data Visualization

The histograms show the frequency of the data. The histograms of all numerical variables are as below:

Figure 1: Histograms of all numerical variables



According to the histograms, the distributions of each variables are different from each other. The variables, ***total_rooms, total_bedrooms, population*** and ***households,*** have large range and are positive-skewed, and the ***housing_median_age***, ***median_income*** and ***median_house_value*** are approximately normal distributions. Many blocks concentrate in the geographic position around 122 degrees west longitude, 34 degrees north latitude and 118 west degrees longitude, 37.5 degrees north latitude. Moreover, the scales of the data are various. The unit of ***median_income*** is thousands of dollar, but the unit of ***median_house_value*** is one dollar.

3. Data cleaning

3.1 Missing data

There is missing data in the ***total_bedrooms***, according to the summary statistics above. Missing data will cause some damaging results. Therefore, we replace the missing data with the median value of ***total_bedrooms***. The replacement of the median value of the variable instead of the mean value can decrease the influence of some outliers of the data.

3.2 Feature variable

The variable, ***ocean_proximity*** contains levels of NEAR BAY, 1H OCEAN, INLAND, NEAR OCEAN and ISLAND. This report splits this variable into five binary variables: ***NEAR_BAY, oneH_OCEAN, INLAND, NEAR_OCEAN*** and ***ISLAND.*** For example, if a block group is near the bay, then its ***NEAR_BAY*** will equal to 1, otherwise equal to 0. The same goes for all the other binary variables.

3.3 Data transformation

Variables***, total_rooms*** and ***total_bedrooms*** represent the total number of rooms and bedrooms within a block. This report divides them by their population and get the ***mean_rooms*** and ***mean_bedrooms***. The mean of rooms and bedrooms make more sense in the model. Therefore, we replace the ***total_rooms*** and ***total_bedrooms*** with ***mean_rooms*** and ***mean_bedrooms***.

Moreover, according to the summary statistics, the scales of variables are different. Hence, we preform feature scaling of all numerical variables except ***median_house_value*** in order to give equal weight in the random forest algorithm.

Finally, the cleaned data set is named as **cleaned_housing** data frame.

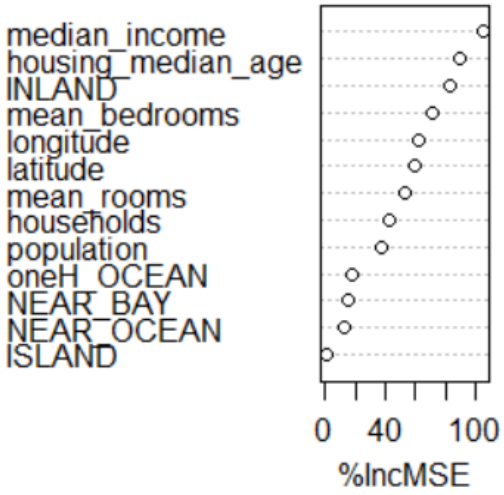
4. Supervised machine learning — random forest

Random forest is a supervised learning algorithm, using multiple decision trees to train and predict samples. Each tree is constructed using a different bootstrap sample from the original data and classifies the feature of the data. Finally, the outputs of each trees can be donated as a set and the prediction is the majority vote on this set. Therefore, combining the learned models of each trees we can increase the overall effect of prediction.

This report randomly split the **cleaned_housing** into a training set named **train** consisting of 80% of the rows of the original data frame and a test set name **test** consisting of 20% of the rows of the original data frame. This report uses the function randomForest() to train the train set and predict the median house value of the test set.

Getting the model of random forest for the train set, we use the rf1$importance to display the importance of predictor in the model.

Figure 2: Importance of predictors



Mean Squared Error (MSE**)** is calculated as blew:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_{i-}\widehat{y_i})^2$$

where m is the sample size; y_i is the actual value; y_i hat is the estimated value.

MSE equals to the mathematical expectation of the square of difference between the estimated value and the actual value, measuring the deviation of an estimator. The lower the MSE is, the more accurate the model is. **%InMSE** is defined as the percentage of the increase in MSE of predictions when the given variable is shuffled. If a variable is important to a model then the shuffled data of this variable will make great deviation, lower the accuracy of the model and increase the MSE. Therefore, the InMSE acts as a metric of that given variable's importance in the performance of the model. A higher number indicates a more important predictor, so we can see that ***median_income*** and ***housing_median_age*** are very important.

Root Mean Squared Error (RMSE) is the root of MSE and it helps evaluate the performance of the model. The resulting RMSE of this model is **47733.42**. It represents the prediction of median price of a house in a given district to within a RMSE delta of the actual median house price.

Next, this report computes the RMSE of prediction of test data to see how well the model predicts using the test data. The resulting RMSE of test data is **45754.81**. We can see that the model score roughly the same on the training and testing data, suggesting that it is not overfit and that it makes good predictions.

5. Model refinement

According to the RMSE reported above, the RMSE of this model is quite high, so it is necessary to refine the model in some ways. This report focuses on the data transformation. One the one hand, this report changes the feature levels of ***ocean_proximity*** into numerical values from one to five, instead of a set of binary variables. From the figure below, we can see that there is an increase in the importance

of all predictors and the ***ocean_proximity*** is a very importance variable in the refined model, suggesting a better fitting of this model.

Figure 3: Importance of predictors of refined model

On the other hand, given the large scale of the response variable ***median_house_value***, this report takes the logarithm value of this variable as the response variable, thereby reducing the model error significantly. The resulting RMSE is **0.2226408**, and the resulting RMSE of test set is **0.2199046**. Taking logarithm of the ***median_house_value*** also has the advantage that we can easily get the predict median house value by taking the e-th power of the predicted response variable. However, the reduce of RMSE does not mean that this model is more accurate, because the reduce may result from a smaller value of the response variable.

6. Conclusions

This report predicts the median house value based on the random forest algorithm. We can conclude that: the median income and median house age are very essential factors when it comes to determining the median house value of a block; the way of dealing with the feature variable and the data scale of response variable influence the performance of the model. By transforming the feature variable into numerical variable and taking the logarithm of response variable, this report gets a RMSE of **0.2226408.** Future researches should improve the model further and move forward, trying other statistical models to get a better result of prediction.