**Housing Price Prediction Capstone**

***Abstract***—In this machine learning project, I apply machine leaning concepts with real estate data. The data is from Kaggle, aim to use different feature of the houses to estimate the market price. There are a total of 16 features used in this task for the prediction.

I.Introduction

To estimate the real price, this report will recognize the most relevant factors at first and then use the selected factors to build different models: Linear regression, Random Forest and Decision Tree. Then I will use the accuracy of different model to evaluate different models and make analysis. Finally, I will print out the best predicted value as the final result.

II.Data Description

The task is to estimate the real estate price based on the following 21 features:

1)Area: area of the building.

2)Garage: how many garages are there in the houses.

3)FirePlace: how many fireplaces are there in the houses.

4)Baths: how many bathrooms are there in the houses.

5)White Marble: how many rooms are with white marble.

6)Black Marble: how many rooms are with black marble.

7)Indian Marble: how many rooms are with Indian marble.

8)Floors: how many floors are there in the houses.

9)City: which city is the house in.

10)Solar: whether the house uses solar energy or not.

11)Electric: whether the house uses electricity or not.

12)Fiber: whether the house uses fiber or not.

13)Glass door: whether the house uses glass door or not.

14)Swimming pool: whether the house has pool or not.

15)Garden: whether the house has garden or not.

16)Prices: house price.

Before the analysis of data, I drop all the null value and check the duplicate values, in this way there are 11 columns left.

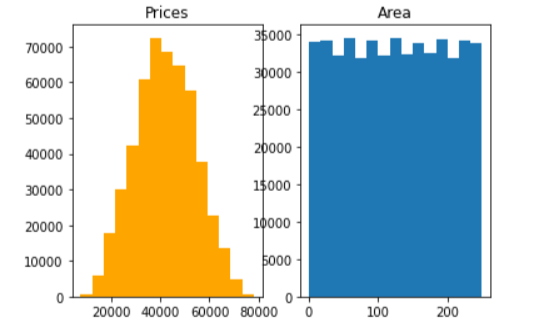


Fig.1. Distribution of prices(left) and area(right)

It’s obvious that the area distribution is quite average and the prices distribution is nearly normal distribution.

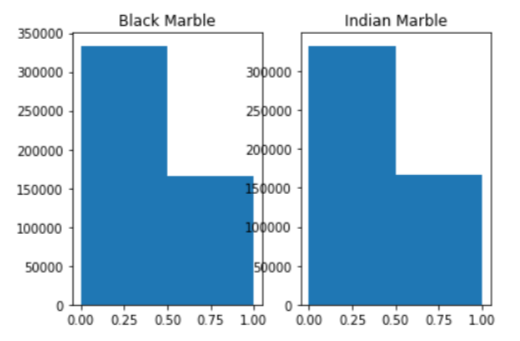
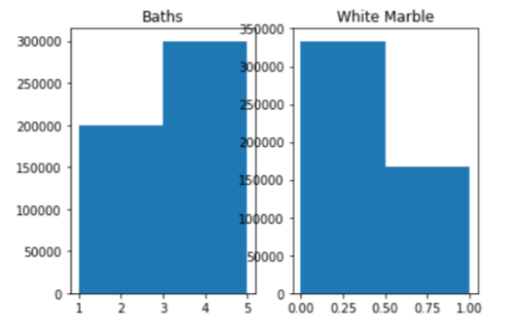
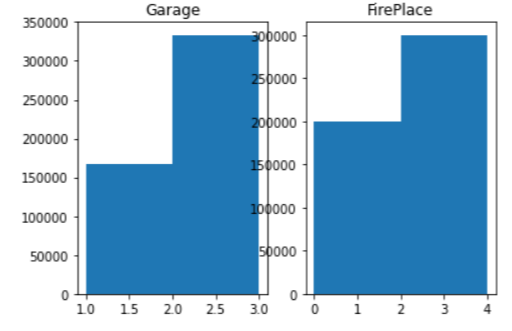


Fig.2. Distribution of Garages, FirePlaces, Baths, White Marble, Black Marble and Indian Marble

The distribution of garage, FirePlace, bath, Black Marble and Indian Marble is left skewness, which means most of the data is smaller than the average rate, while white marble is opposite.

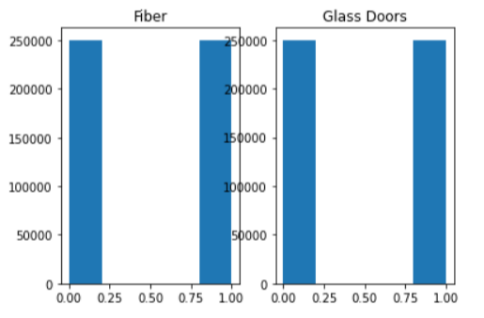
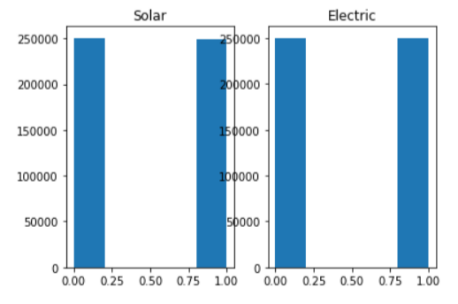
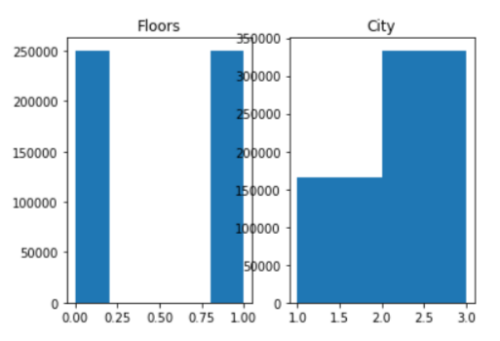
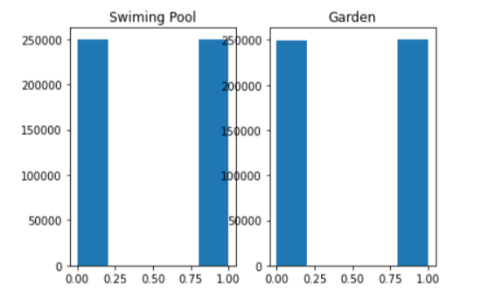
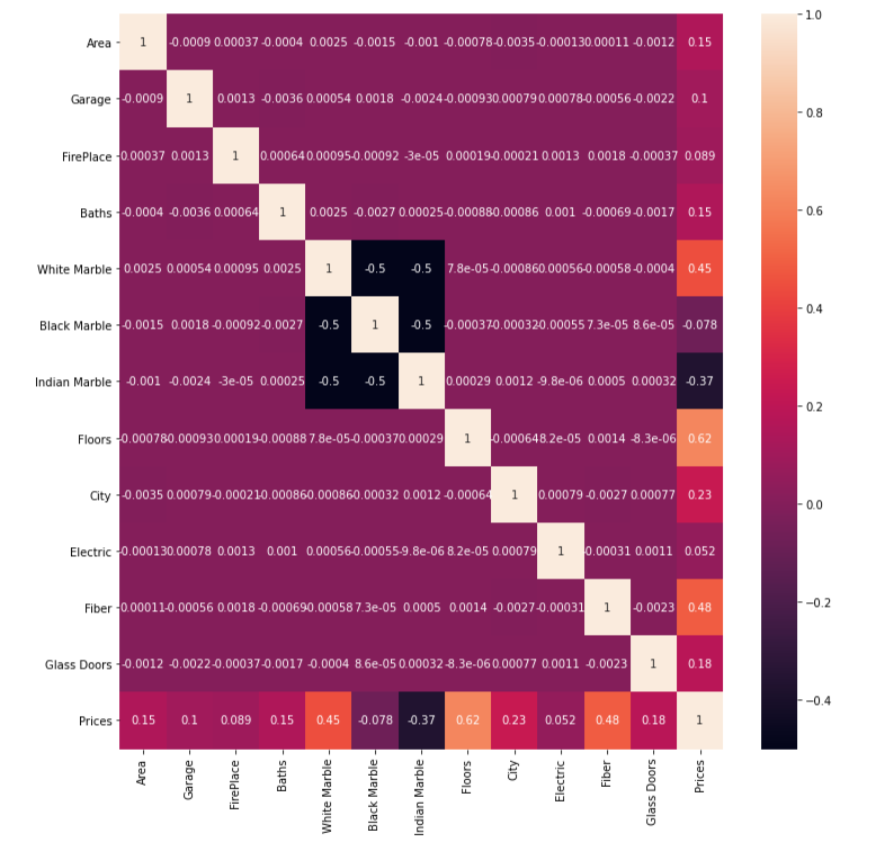
 

Fig.2. Distribution of Citys, floors, solar, electric, fiber, glass door, swimming pool and garden.

The distribution of citys are right skewness while others are 50 to 50 which is quite average.

III.Correlation Analysis

From the pairs plots and the heat map between variables, we can clearly examine the correlation of variables. Classification variables are not suitable to investigate in correlation analysis, so they will be further analyzed according to significance in the following analysis. According to the correlation coefficient, we can eliminate variables that are not closely related to the dependent variable Prices, that is, the absolute value of the correlation coefficient is less than 0.1. Thus, I drop FirePlace, Black Marble, Electric, Solar, Swimming pool and Garden.



IV.Model Building

First, I need to split our dataset into training and testing sets. I will be using the training data to train our model for predicting the real price. Then the testing data will be used to verify the predicted price by the model. In this project, we randomly select 80% of the raw data as the training set, and the remaining 20% of the data is the testing set. That is, the training set contains 400,000 observations, while the testing set contains 100,000 observations.

Secondly, I use Linear regression, Random Forest and Decision Tree to train the model. And also using the scale to normalize the training data, which find out the following R square of those models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | lasso | ridge | ENet | Random Forest | Decision Tree |
| 0.98952 | 0.989516 | 0.989516 | -0.0169 | 0.98516 | 0.98263 |

According to our evaluation results, the R square of Ridge Method is 0.99 which is the largest one compared with other models, referring to the right model to be the best one. And the best models refers to the Ridge Regression with alpha at 0.0001 fit with the scaled training set.