Real Estate price prediction

Ali Bakhshes

***Abstract*— Real estate price prediction is a critical area of research and application in the field of real estate economics and finance. The accurate prediction of real estate prices is essential for various stakeholders, including home-buyers, sellers, investors, and policymakers. This study explores the use of statistical and machine learning techniques to predict real estate prices based on a variety of relevant factors such as location, age, being convenient, etc. In this project we have used Linear Regression method to predict the price based on the mentioned factors.**

I. INTRODUCTION

In this project we are going to work on a dataset including information of real estates such as longitude, latitude, age, etc. At first we have done some analysis on the data using visualization and doing some statistical tests such as two-sample t_test, ANOVA, etc. At the end we prepared data by doing some preprocessing methods such as removing outliers(z_score method) and scaling(MinMax scaling) then implemented a linear regression model using scikit-learn package to predict the price.

II. DATA ANALYSIS

The dataset include following features:

- Transaction date
- House age
- Distance to the nearest MRT station
- Number of convenience store
- Latitude
- Longitude
- House price of unit area

There was also No. column in the data that it was useless and we dropped it at the beginning. In the following we did some visualization to get more information about the features and investigate the effect of each of them on the target feature.

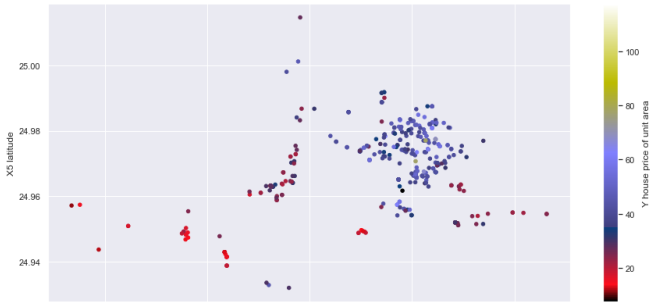

Fig. 1. This plot shows the relation between the location of a house and its price

This plot shows that houses located in the north and east are more valuable. To show this relation more clearly we separated houses to two groups based on their longitude. (we have used median longitude 121.538630). The following plot shows how the price in these two groups differs.

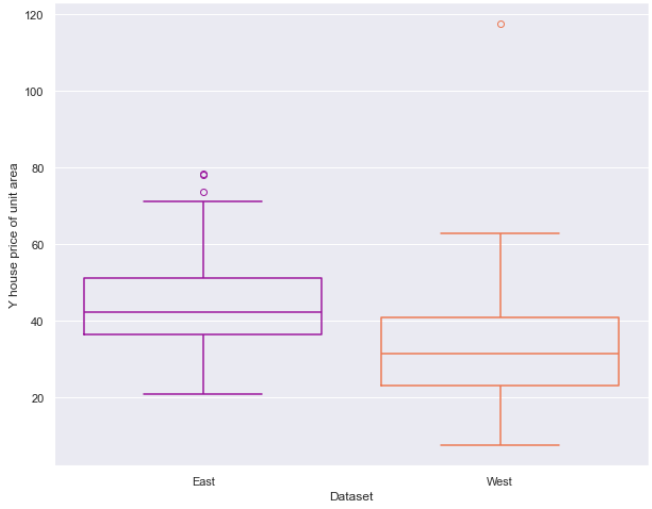

Fig. 2. This plot shows the distribution of the price of houses in different areas

Other investigations will be mentioned in the next section using hypothesis tests.

III. HYPOTHESIS TEST

In this part we were asked three question and also to make an extra question and answer all of

them using appropriate hypothesis test. We will mention the questions and their result respectively.

A. Test if the average price per unit area of houses above the median age is significantly different from those below the median age. Use p-value method.

We used *Two-sample t_test* two answer this question including following hypothesis:

$$H_0 : \mu_{old} = \mu_{new}$$

$$H_1 : \mu_{old} \neq \mu_{new}$$

The result for this test is as follow:

```
P_value: 5.404452714497797e-07
The null hypothesis will be rejected!
```

Fig. 3. Result for the first test

It means that we can say that the average price of two groups are significantly different. It is obvious also by looking at following plot.

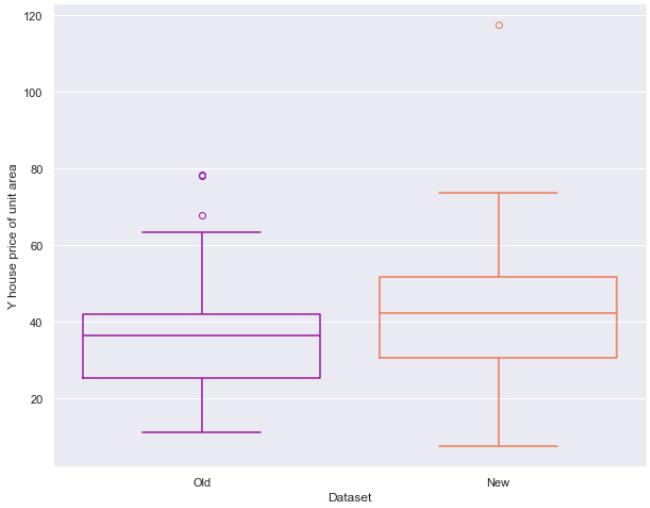

Fig. 4. This plot shows the distribution of the price of houses with different ages

B. Investigate if the average price per unit area significantly differs by the number of convenience stores (categorized by "X4 number of convenience stores").

We have used *ANOVA* method two answer this question including following hypothesis:

$$H_0 : \mu_0 = \mu_1 = \ldots = \mu_{10}$$

$$H_1 : \mu_0 \neq \mu_1 \neq \ldots \neq \mu_{10}$$

The result for this test is as follow:

```
P_value: 1.1781067247237561e-36
The null hypothesis will be rejected!
```

Fig. 5. Result for the second test

The following plot also confirms the result of the test:

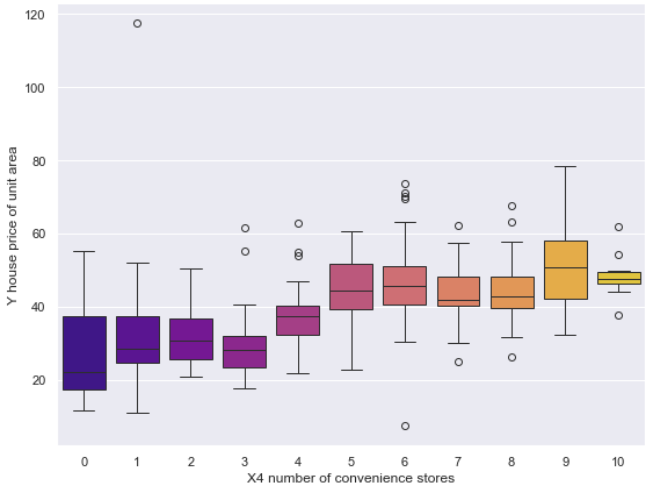

Fig. 6. This plot shows the distribution of the price of houses with different number of convenient stores

C. Investigate if there is a relation between the location of houses and their age.

This is the question that we made ourselves. To answer that we made two new columns in the dataset:

1) descriptive longitude
2) descriptive age

We have used Chi^2 *test* to answer this question. The result for this test is as follow:

```
0.5540364433965241
The null hypothesis will be accepted!
```

Fig. 7. The result for the third test

It means that although we can not say how they are related, there is a semantic relation between these two features.

IV. CALCULATION THE CORRELATION BETWEEN EACH FEATURE AND THE TARGET FEATURE

In this part we were asked to calculate the correlation between each feature and the target feature and report the features with the highest correlation. The correlation matrix is as follow:

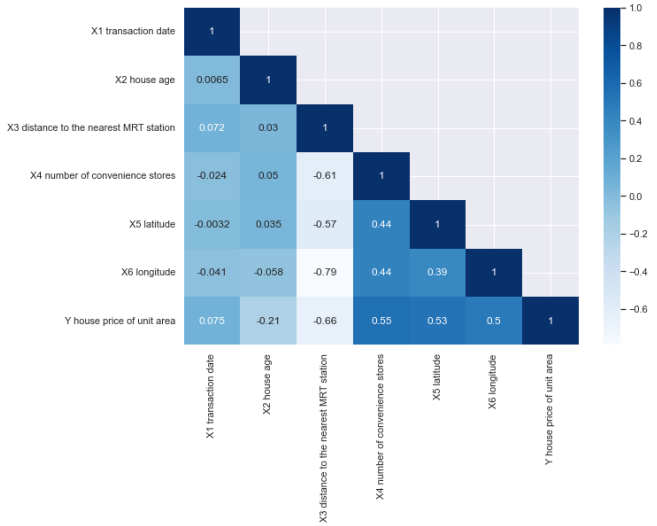

Fig. 8. correlation matrix of all features

And the following features has the highest correlation with the target feature respectively:

1) Distance to the nearest MRT station (-0.66)
2) Number of convenience stores (0.55)
3) Latitude (0.53)
4) Longitude (0.5)

V. PREPROCESSING AND MODEL TRAINING

The last step in our project is training a linear regression model to predict the price of houses based on mentioned features. Before training the model we had to prepare the data so we have done following steps:

1) Check for missing values: Fortunately there was no missing value in the data.
2) Omitting outliers: we have removed outlier using z_score method. In this method the an upper-bound and a lower-bound will be defined and values bigger than upper-bound or smaller than lower-bound will be removed.

$$upper\_bound = \mu + 3\delta$$

$$lower\_bound = \mu - 3\delta$$

3) Scaling data: To scale data we have used two methods:
   a) Min-Max scaling: In this method each value is calculated as follow:
   $$\bar{x} = \frac{x_{max} - x}{x_{max} - x_{min}}$$
   b) Standard scaling: In this method each value is calculated as follow:
   $$\bar{x} = \frac{x - \mu}{\delta}$$
4) Eventually in the last step we trained our linear model on both data (standard scaled and min-max scaled) and the results is shown in fig 8 and 9.

```
train_data report:
MSE: 0.013087957876542307
R_squared: 0.6093244091994814
test_data report:
MSE: 0.010680699663415092
R_squared: 0.6817629246321395
Coeficients:
[ 0.06472329 -0.16526613 -0.27705742  0.17327     0.23258929  0.01314123]
```

Fig. 9. result on min-max scaled data

```
train_data report:
MSE: 0.39067559080050984
R_squared: 0.6093244091994902
test_data report:
MSE: 0.3188189242758957
R_squared: 0.6817629246321502
Coeficients:
[ 0.10857955 -0.23400278 -0.36892303  0.27805151  0.22435427  0.01426122]
```

Fig. 10. result on standard scaled data

VI. CONCLUSION

After doing these analytical investigations we can generally say that although the location and age are important factors to specify the price of a house, being convenient has more effect on the price than the other features. As it is clear the features that we saw in part IV that are highly correlated with the target feature are *Distance to the nearest MRT station* (negative correlation) and *Number of convenience stores* and also the coefficients of these features after training model were bigger than other ones.