Real Estate Price Prediction Project

Econ 445 Group: Haomiao Yu, He Sun, Shengying Li, Qian Feng, Xianyi Deng

March 2022

***Abstract*—**In this machine learning project, we apply machine leaning concepts taught in the class with real estate data. We collect date from loopnet_data_ca, aim to use different feature of the real estate to estimate the market price of different project. There are a total of 21 features used in this task to predict the market price of the real estate project.

I. Introduction

The price of real estate is always a big question in investment field. Except for own using, people invest in the real estate for the future growth value. There are some problems to estimate the real value by manual testing because of the complicated influence factors. Therefore, this report will analyze the market value of the real estate based on different regression methods in machine learning.

To estimate the real price, this report will recognize the most relevant factors at first and then use the selected factors to build 5 different models: Linear regression, SVR, Random Forest, Decision Tree and KNN. The we will use the accuracy of different model to evaluate different models and make analysis. Finally, we will compare the predicted price with the given price to find the most valuable real estate as our good deals.

II. Data Description

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

1) *Id:* id of the building.
2) *Crawled id:* id during the web scraping on loop net.
3) *Zip:* zip code of the building.
4) *Address:* detail location of the building.
5) *Ain:* phone number of the owner of the building.
6) *Size:* the size of the building in square feet.
7) *Sale_type:* the type of the building for sale, has three categories of value: investment, owner use, investment of owner use.
8) *No_story:* the number of the stories of the building.
9) *Property_type:* the type of the property.
10) *Property_subtype:* sub type of the property, which is a blank column.
11) *Year_built:* the year the building built.
12) *Year_renovated:* the year the building rebuilt.
13) *Parking_ratio:* the ratio of parking lot to the size of the building.
14) *Price_per_unit:* the price of the building divided by per unit.
15) *No_units:* the number of units in the building.
16) *Lot_size_ac:* the total horizontal land area within the boundaries of the building.
17) *Apartment_style:* the sort of the apartment, with the value of garden, high rise, low rise, mid-rise, single-family house, townhouse 5 kinds.
18) *Building_class:* the class of the building, from A to C and F.
19) *Cap_rate:* the rate of return on a real estate investment property based on the income that the property is expected to generate.
20) *Gross_rent_multiplier:* the gross multiplier of renting the building.
21) *Opportunity_zone:* whether this building is located in an opportunity zone or not.

Before the analysis of data, we drop column "property_subtype" since it has no values. Second, we drop column "ain", "zip", "address" since tel numbers don't affect house value. Then, we replace year_built with those houses which has year_renovated and build a factor that indicates the house's age. Finally we drop all the NaN.

On the next stage, we need to qualify some label

data. We change variable type of 'building_class', 'sale_type' and 'opportunity_zone'.

As the target of predicting house price, columns of id, address and phone number can be irrelevant. Most of the value in the property type column and property sub type column are null value, while the price per unit column doesn't make any sense to the price prediction. In order to make the features clear during modeling, we concat the built year and renovate year of the building into a column called house age. In this way we have 11 columns left.

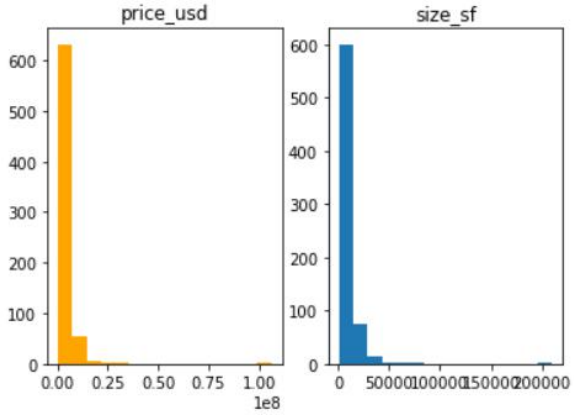

Fig.1. Distribution of price_usd(left) and sixe_sf(right)

It's obviously that the price and size of the building shows the distribution with left skewness, most of the price data drops in the 0-0.25*le8 range, plotted in orange color representing as the target column, and most of the size of the building is under 250 square feet.

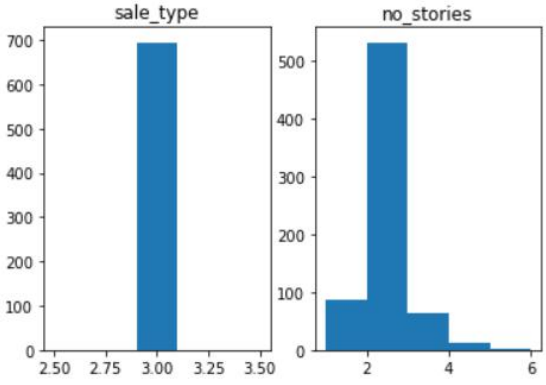

Fig.2. Distribution of sale_type(left) and no_stories(right)

Among the three types sale, most of the data has the sale type of 3, which means investment. Most of the buildings is has two stories.

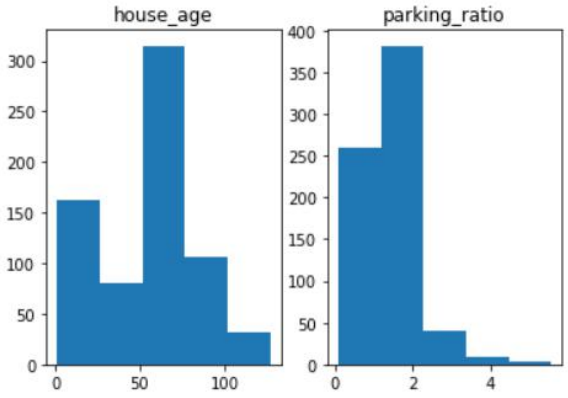

Fig.3. Distribution of house_age(left) and parking_ratio(right)

Most of the buildings has an age of fifty years to 75 years, next is lower than 25 years. Majority of the parking ratio drops in 1-2, next is lower than 1, showing the distribution left-skewed.

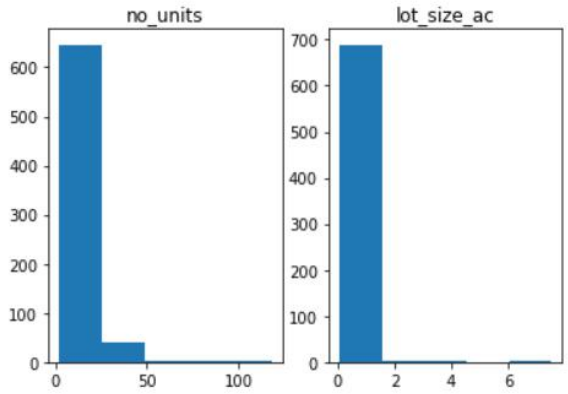

Fig.4. Distribution of no_units(left) and lot_size_ac(right)

Most of the buildings has units less than 25, and lot size range in 0-2, which shows relevant with the feature of numbers of stories, showing most of the buildings has 1 or 2 floors.

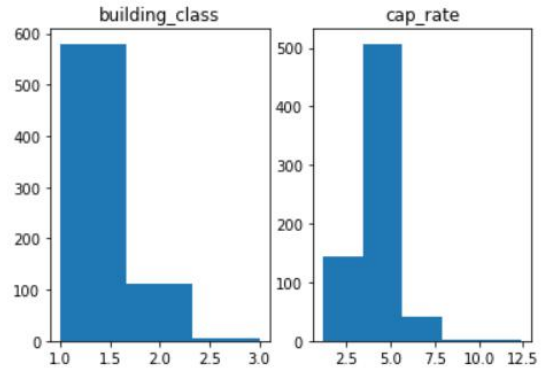

Fig.5. Distribution of building_class(left) and cap_rate (right)

The distribution of building class showing that most of the buildings are at the class of 1-1.5, which represents that most of the buildings are at C rank, with few at A rank while few at F rank. The majority of the cap rate drops around 5.

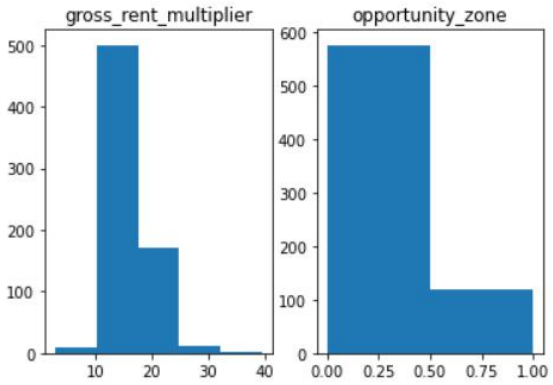

Fig.6. Distribution of gross_rent_multiplier (left) and opportunity_zone(right)

The distribution of gross rent multiplier shows that most of the buildings has a multiplier at 10-20, which is relevant with the low cap rate. For the opportunity zone, as the large quantity represents in the 0-0.5 range, which shows that most of the buildings are not in an opportunity zone.

III. Correlation Analysis

First, we build a correlation matrix to evaluate the correlation between each fators.

Table I

	price_usd	size_sf	no_stories	house_age
price_usd	1.0	0.9	0.5	-0.2
size_sf	0.9	1.0	0.5	-0.1
no_stories	0.5	0.5	1.0	-0.3
house_age	-0.2	-0.1	-0.3	1.0
parking_ratio	-0.1	-0.1	-0.2	0.0
no_units	0.8	0.9	0.4	0.0
lot_size_ac	0.4	0.5	0.1	0.0
cap_rate	0.0	0.0	0.0	-0.2
gross_rent_multiplier	0.0	-0.1	0.0	0.1

	no_units	lot_size_ac	cap_rate	gross_rent_multiplier
parking_ratio	-0.1	0.8	0.4	0.0	0.0
no_units	-0.1	0.9	0.5	0.0	-0.1
lot_size_ac	-0.2	0.4	0.1	0.0	0.0
cap_rate	0.0	0.0	0.0	-0.2	0.1
gross_rent_multiplier	1.0	-0.1	0.0	0.1	-0.1
price_usd	-0.1	1.0	0.5	0.1	-0.2
size_sf	0.0	0.5	1.0	0.0	-0.1
no_stories	0.1	0.1	0.0	1.0	-0.8
house_age	-0.1	-0.2	-0.1	-0.8	1.0

According to the correlation coefficient, we can eliminate variables that are not closely related to the dependent variable 'price_usd', that is, the absolute value of the correlation coefficient is less than 0.1. Thus, we drop 'parking_ratio', 'cap_rate' and 'gross_rent_multiplier'.

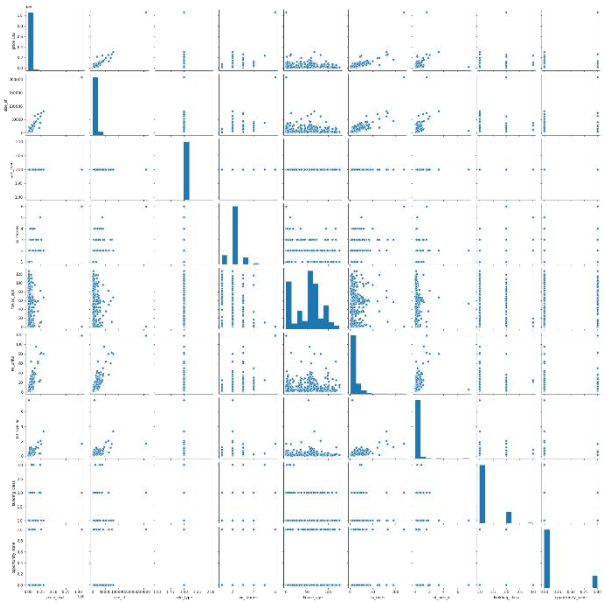

Fig.7. Seaborn Pairs Plot

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. Due to the correlation plot, we drop 3 variables which are sale_type, building_class and opportunity_zone.

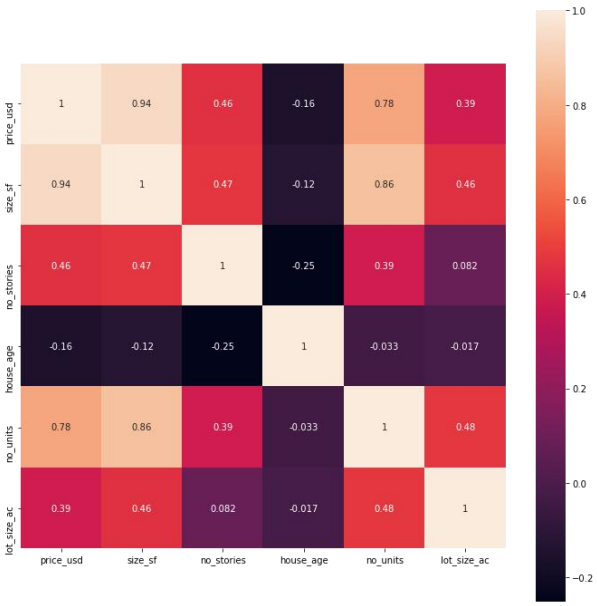

Fig.8. Heat Map

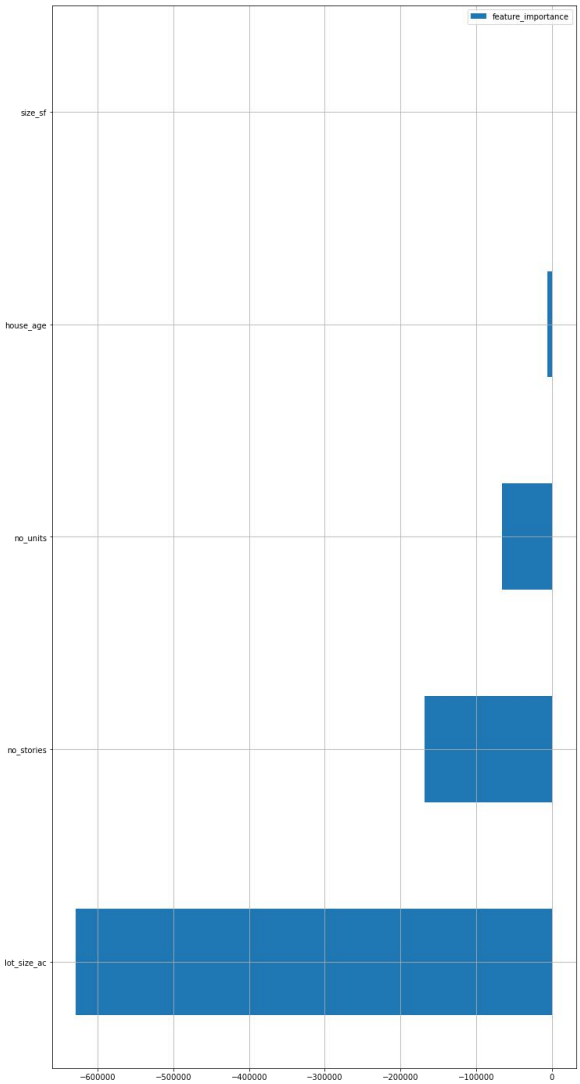

Fig.9. Feature Importance

From the plot above the most important features to the price is number of stories, number of units, size, house age, and lot size. Built train set using these five features.

IV. Model Building

In this part, we will use 5 algorithms: Linear regression, SVR, Random Forest, Decision Tree and KNN to solve this regression problem. The target estimator is 'price_usd' column, and we include 5 features in the model.

- Step 1: Observe the outliers and drop outliers

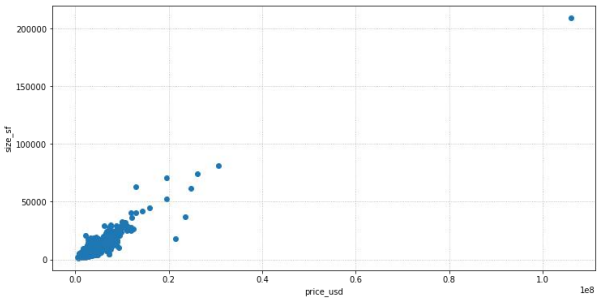

Fig.10. Outliers of the data

From the plot, we can see there are some outliers in our model which may influence the accuracy of our model. Therefore, we drop the data which have the 'size_sf' larger than 50000.

- Step 2: Split data

First, we need to split our dataset into training and testing sets. We 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 550 observations, while the testing set contains 138 observations.

- Step 3: Run the training on 5 algorithms

We will use Linear regression, SVR, Random Forest, Decision Tree and KNN to train the model.

In this project, we use Linear Regression

function from sklearn.linear_model; SVR function from sklearn.svm; Random Forest Regression function from sklearn.ensemble; Decision Tree Regression function from sklearn.tree; KNN Regression function from sklearn.neighbors.

- Step 4: Testing and Evaluation

Since our models are trained, we can make predictions on the testing set (138 samples). And we change the factors in different models to find the optimal factor.

We use R_square to estimate the accuracy in different models. The different R_square for linear regression are as follows:

Table II

	Lasso Scale	lasso	Ridge Scale	Ridge	ENet Scale	ENet
R Square	0.78	0.80	0.78	0.80	-0.01	-0.61

Now we can compare the accuracy in the predicting price among these 5 algorithms.

Table III

	Linear	SVR	RF	DT	KNN
R Square	0.80	-0.08	0.72	0.35	0.78

According to our evaluation results, the accuracy of Linear Regression Method is 0.79 which is the largest one compared with other 4 models.

V. Find the 5 Good Deals

According to the R_square, we choose linear regression as our model. Next, we will use this model to predict the market value of the project ("y_pred") and compare the market value with the real value to choose 5 good deals.

We use linear model to predict the y_pred and calculate the difference between the predicted price and the real price. Then we sort the difference ratio of all the real estate and find the top 5 good deals which have the smallest different ratio

Table IV

	price_usd	predicted	different	difference
8148	410000	1321017	-911016.6	-222.199164
4834	1095000	3253640	-2158640	-197.136098
417	2200000	6227425	-4027425	-183.064761
3188	675000	1817584	-1142584	-169.271762
747	2725000	5867222	-3142222	-115.310895

VI. Conclusion

To sum up, after training our model by Linear Regression method, SVR, Random Forest, Decision Tree and K-Nearest Neighbors method, we obtain five forecasting result of label.

In order to compare different results, we run R_square. The accuracy indicates that the Linear Regression method is the best model we estimate the market value of the real estate in our research. Therefore, we could use Linear Regression method as a great and accurate recognition way to find the good deals in real estate projects.