Valuation on Real Estates Using Machine Learning Models

-Yongle Lu, Yuwen Qiu, Anqi Hu, Ge Wu

Abstract

This project is mainly focused on finding out the properties are selling at below market value by using machine learning models for real estate investors or companies. We apply two models including Simple Linear Regression and Logistic Regression. According to these models, we finally target five properties which have the largest price spread between their actual price and the predicted price.

Index Terms: Machine Learning, Real Estate, Simple Linear Regression, Logistic Regression.

Topic Background

The real estate market around the world is a widespread trade. House valuation is a significant financial decision for individuals who are going to invest or sell the house. House is not only the basic needs for most people but also could be a great investment or asset for some people. Property valuation used to be an imprecise science. Most people always value the house based on their own experience. There are lots of properties that have been undervalued for many reasons. However, now people have more techniques to achieve more accurate results. So, in this paper we aim to apply machine learning techniques to do the house price prediction for finding out the undervalued properties for potential buyers and owners.

Data Description

Our data consists of 9064 examples of houses, and it contains 22 different features. However, we drop 9 features which are not related to our project and only use 13 variables to deploy our models. House price is our dependent variable while attributes including year built, year renovated, numbers of units, numbers of stories,lot size, parking ratio, apartment style, building class, capital ratio, gross rent multiplier, opportunity zone will be our independent variables. Moreover, our dataset also contains features in various formats including numerical factors such as the price of house, the number of units and categorical factors such as the building class and opportunity zone.

This dataset has rich samples and various features which enable us to apply different techniques to figure out the most fitted model and the top undervalued properties with more accurate and comprehensive results.

Data Processing

Firstly, we delete 9 variables including property type, sale type, id, price per unit, address, zip code, ain, crawled id, property subtype which do not contribute to the predict of response variable. Then we preprocess one variable year renovated to indicate whether the property is renovated. We also apply scale and create level factors for variables such as building class. After that, we finalize our data cleaning process by detecting the existence of missing values and avoiding multicollinearity problems.

Lastly, we split our dataset into a training dataset with the size of 8245 and 13 variables, and a testing dataset with the size of 906 and 13 variables.

Model analysis

a. Linear regression model

The first model we used is the linear regression model. Linear regression model is a useful model to research the relationship between one or more explanatory variables and one dependent variable. Obviously, the linear model is effective for the dataset to predict the house price with the 13 variables mentioned above. To avoid multicollinearity and choose the best model, the training set and testing set are divided into 10 groups through rolling windows and then train the model with those different parts of datasets. After comparing the R-square of trained models, we find that the eighth group is the best model which trained by 0-70% and 80%-100% of the training dataset, and the R-square of the model is 90.24%. The next step is to apply this best model to the whole dataset and then get the expected housing price. With the expected housing price, we can divide the dataset into two categories, which are undervalued houses with positive differences from the market price and overvalued houses with negative differences from the market price. However, there is a problem that if one of attributes used in the predicted process is an outlier, the result may be affected and become an error. To avoid this mistake, we use the logistic regression model to classify the dataset in the next process.

b. Logistic regression model

Logistic regression model is a widely used machine learning model in solving binary classification class problems, so it is helpful to classify the dataset correctly. The target variable in this project is that if the house is undervalued, we set a dummy variable as 1 for this house and otherwise is 0. After finding the best logistic model, we apply the model to the dataset, and then find the 5 undervalued houses. The indexes of those 5 houses are 8579, 8363,8549,8348 and 8537.

Conclusion

The goal for this project is to find the undervalued houses by the machine learning model and gain profit from the price spread. Through this study, investors and companies can find the best real estate investment more efficiently with houses' conditions. To accomplish it, we apply the linear regression model to get the expected housing price. Since the attributes of a couple of samples are outliers, we use the logistic regression model to verify the classification. Based on all model results, there are 5 undervalued houses in the dataset.

However, we still have some improvement in the future study. Firstly, we do not consider the region variable in this project. For example, there are some houses that are easily undervalued in some special areas, so we can take zip codes into consideration through rating the geographic areas according to the wealth the residents hold in that area. Moreover, we can try more models to predict the housing price such as time-series models and gradient boosted models and then compare the accuracy to choose the best model in the future study.

Data Resource

The data is obtained from the ECON446 course file "loopnet_data_ca" in the dataset provided by our professor Mendler.