TRƯỜNG ĐẠI HỌC KINH TẾ LUẬT

KHOA TÀI CHÍNH - NGÂN HÀNG

MACHINE LEARNING

Predict price house in HoChiMinh City

Lecture	***Nguyen Anh Phong, Phan Huy Tam***
Student	***Huynh Thi Ha Thanh K194141746***
Class	***212CN1701***

Ho Chi Minh City, 23th, June 2022

Contents

PREDICT PRICE HOUSE BY MACHINE LEARNING

Abstract: House price prediction is a significant financial decision for individuals working in the housing market as well as for potential buyers in Vietnam. This study applies a machine learning model to help home buyers and sellers find a reasonable price for their home. I applied Linear Regression, Random Forest Regression, Decision Tree Regression models to find the best forecasting model based on data taken from ChoTot site. The best model to predict house prices in Ho Chi Minh City is Decision Tree Regression.

Index term: House price prediction, Random Forest Regression, Linear Regression, Decision Tree Regression.

1.Introduction

House price prediction is a hot topic in the real estate industry when housing prices skyrocket rapidly in the world and including Vietnam. However, due to many limitations, some research models give not high accuracy results and have very large errors. The reason for this error is that house prices are influenced by many factors such as location, direction, number of years of existence, etc.

The objective of the study is to find the best model to predict house prices in Ho Chi Minh City by some machine learning models. Based on the area, number of bedrooms, bathrooms, location, ... the model can predict the price with the smallest error taken from ChoTot page (Cho Tot is a Vietnamese online marketplace for homes, cars, recruitment, used electronics, pets, and home services).

The difference of this study: Currently, in Vietnam, there are not many house price predictions and data collected at ChoTot (although there are many in the world). Besides, the research papers show that their accuracy is not high.

Cons: With the data taken from the ChoTot site, their accuracy has not been verified. The data has a lot of NaN and there is a lot of noise/outlier, leading to difficult prediction and

large error. Besides, the analysis of factors affecting house prices makes it difficult for the author.

2. Relate work

Previous studies on the real estate market using machine learning approaches can be categorized into two groups: the trend forecasting of house price index, and house price valuation.

According to Changchun Wang and Hui Wu (2018), they found that the Random Forests can capture the nonlinear hidden relationship between house price and house location and give an overall better estimation than benchmark linear regression. This simple model can be scaled up for larger data with more features and captures the nonlinear information traditional models used to neglect based on attributes like: Area, bedrooms, bathrooms, …

According to CH.Raga Madhuri, Anuradha G, M.Vani Pujitha (2019), the applied machine learning models are Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and Ada Boost Regression, the results show that Gradient Boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions.

Phan, T. D. (2018), Regression tree delivers a prediction result as good as linear regression, while Polynomial regression results in lower errors which is acceptable. Furthermore, Neural Network doesn't seem to work effectively with this dataset. This may not represent the effectiveness of modern deep learning methods based on several important attributes such as: House Type, number of bedrooms, number of bathrooms, bumber of Car slots, and Land size3. Describe data.

3 Description Data

The dataset contains the house price that occurred in Ho Chi Minh City, posted Cho Tot online page. The data is crawl by and posted Github (by HungTrinhIT).

The data set including 16 variables and 24949 observations. Out of 16 variables, 3 of them are continuous, 3 of them are discrete, 10 of them categorical.

Variable	Description	Data type
Dependent variable		
Gia	price of the house in HCM City	Numeric
Independent variable		
DiaChi	full name of the location	Nominal
TinhTrangBDS	Đã bàn giao/Chưa bàn giao	Nominal
DienTich	living area in Square feet	Numeric
Gia/m2	VND/m2	Numeric
Phongngu	number of bedrooms	Numeric
TenPhanKhu	object	Category
SoTang	number of floor	Numeric
PhongTam	number of bathrooms	Numeric
Loai	object	Category
GiayTo	object	Category
MaCanHo	object	Category
TinhTrangNoiThat	object	Category
HuongCuaChinh	object	Category
HuongBanCong	object	Category
DacDiem	object	Category

Table 1: Describing the collected variables.

Most of these variables indicate the type of information a typical house buyer would want to know before buying a house such as area, number of bedrooms, number of bathrooms, etc.

The data appeared a lot of NaN and noise/outlier, so I decided to go process them. Besides, the data columns are not standard data form to be able to perform analysis, so we need a data preprocessing step.

4. Data pre-processing and exploratory data analysis

I will represent the steps of making data for the analysis and give some statistical information which helps to understand to data better.

4.1 Data Pre-processing

First, I will check NaN in data set by function info() in Python. It is seen that 14 variables out of 16 variables have missing observations. Among 14 variables, SoTang variable has highest missing value (6726 non-null / 24949 observations). The variable with lowest missing value is DiaChi and Price (24949 non-null/ 24949 obsetvations).

Most of these missing observations indicate that ‘NaN’. So, to exactly predict house, I filter data where DienTich, Phongngu, PhongTam variables we have no basis to fill them, so I

choose to drop these rows that contain the value NaN. Besides, if the value of Gia is 0 I will drop them it because it is noise.

In this house price prediction article, I will study the Quan (District) of the county to compare the difference between them. The Quan column will be extracted from the DiaChi column.

To predict target variable, I must clean Gia column because it has object value. To exact value, I multiplied 1 billion for row which contains “tỷ”, multiplied 1 million for row which contains “triệu”.

Then, we will check outlier in the data. Almost columns have outlier and extreme observations. I can get the rid of these observations by removing them. Therefore, I just process noise value and drop DienTich variable is higher 500 m2 and SoTang variable is higher 81 (Because The fact that in Vietnam, the highest SoTang is 81 and the highest DienTich is 500m2).

4.2 Exploratory Data Analysis

House price is the target we need to pay attention to, the max value is 980 billion and the min value is 10 million, the average price is 2.2 billion. Inferring that the data is skewed to the side, we can see that the data after using the log has tended to be normal.

I will check association of the sale price with category by drawing scatter.

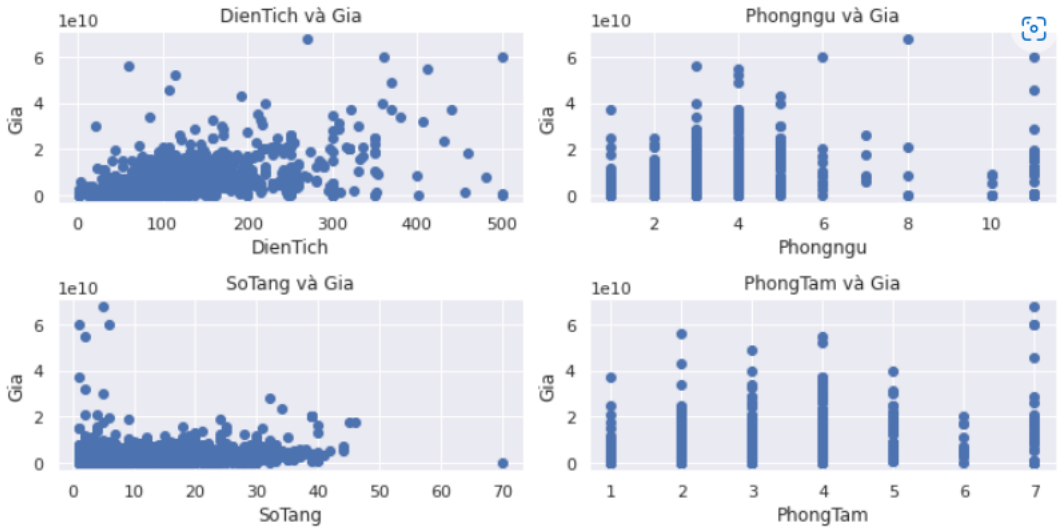

Figure 1: The relationship between numeric variables and target variable.

With DienTich and SoTang column, there are differences when the price goes up.

With Phongngu and PhongTam column, there are no difference when the price goes up.

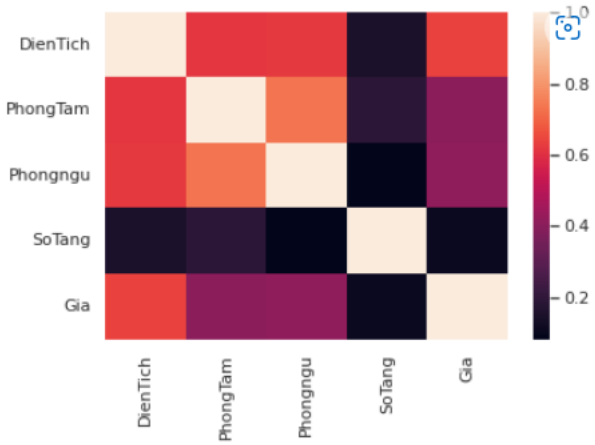

Figure 2: Correlation between the numeric variables and target variable

DienTich have the highest correlation with Gia, PhongTam & Phongngu maybe have correlation with Gia but not as high as DienTich. SoTang has a low correlation with Gia.

The association between the category variables and the house price variable

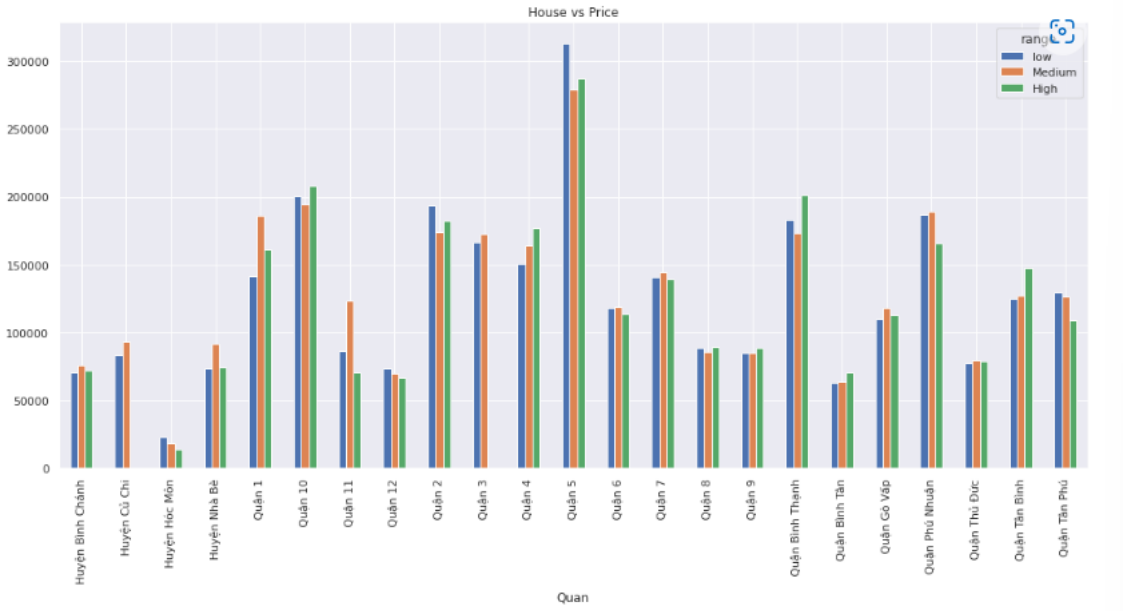

Figure 3: Quan & Gia

It can be seen that of all districts, district 5 has the highest house prices. Meanwhile, Hoc Mon district has the lowest house prices. This shows that there is a difference in house prices across counties.

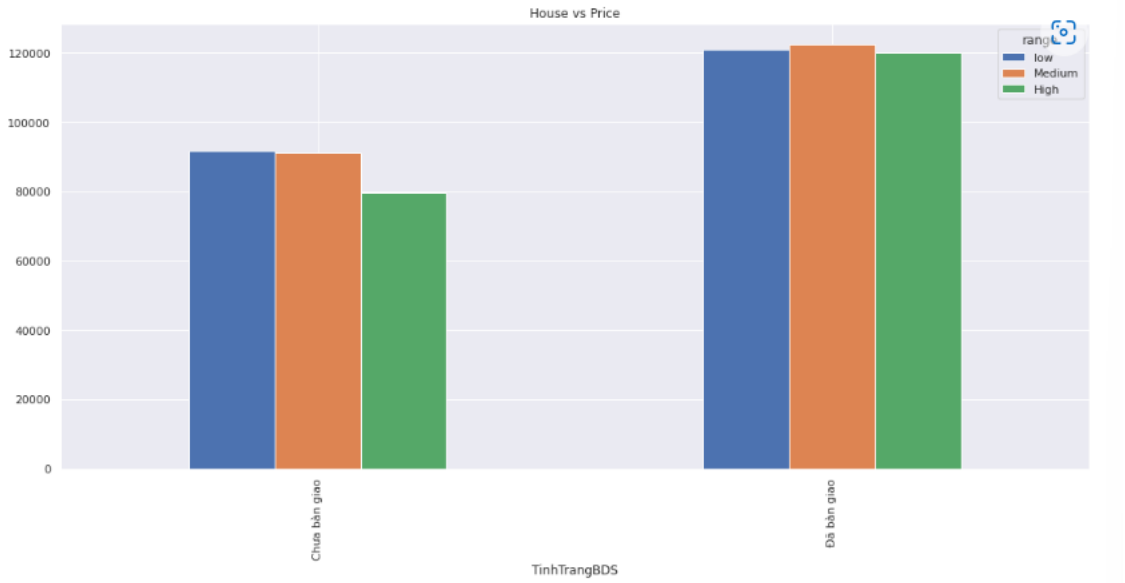

Figure 4: TinhTrangBDS & Gia

In general, the price of the house that has been handed over is higher than the price of the house that has not been handed over. This proves that TinhTrangBDS has an influence on house prices

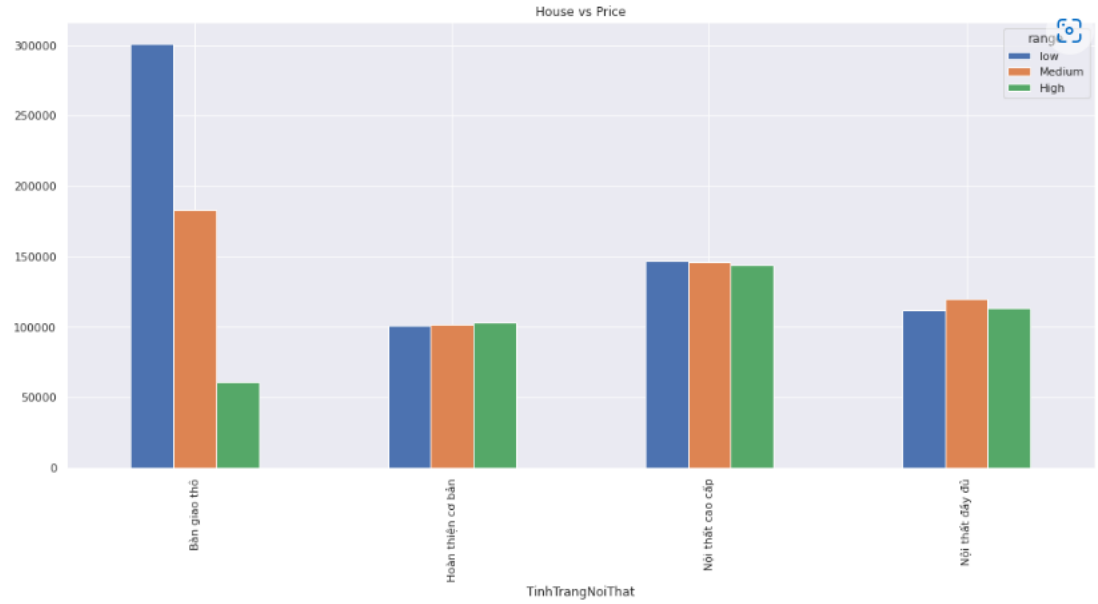

Figure 5: TinhTrangNoiThat & Gia

Đã Bàn Giao Thô has the lowest price in the category of TinhTrangNoiThat.

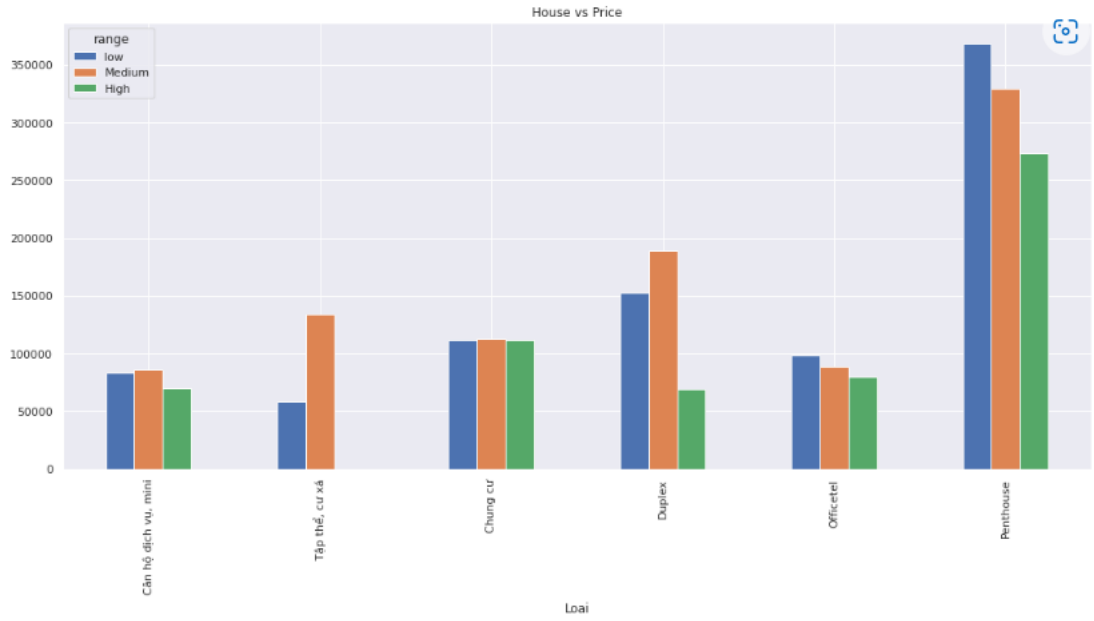

Figure 6: Loai & Gia

It is easy to see that Penthouse is the most expensive type of Loai, Tập thể, cư xá is the type of housing that seems to be the cheapest. This also shows that house type also affects house prices.

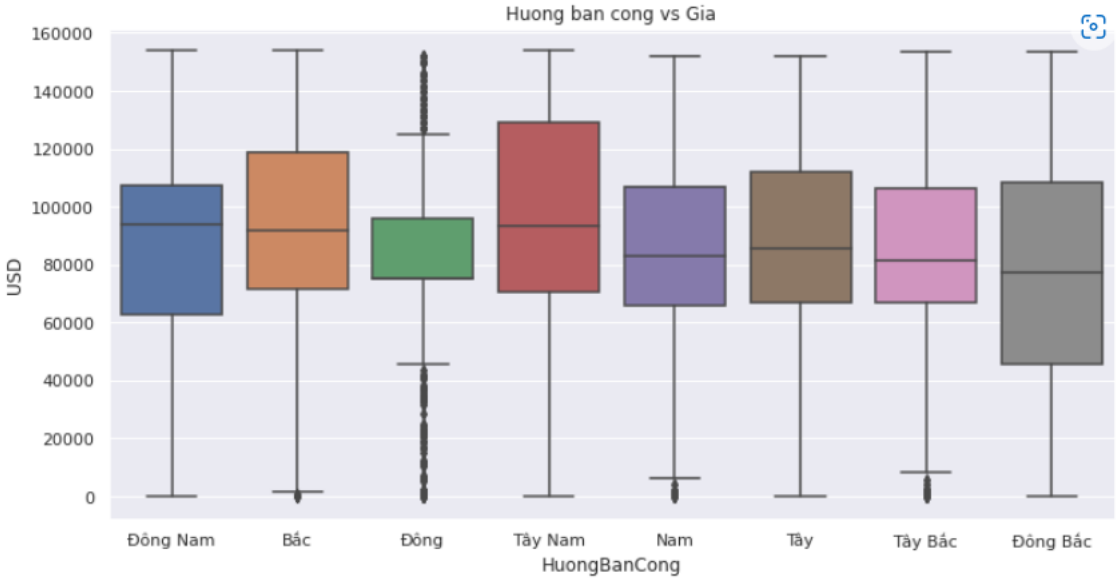

Figure 7: HuongBanCong & Gia

In general, the average price of HuongBanCong is not too big of a difference.

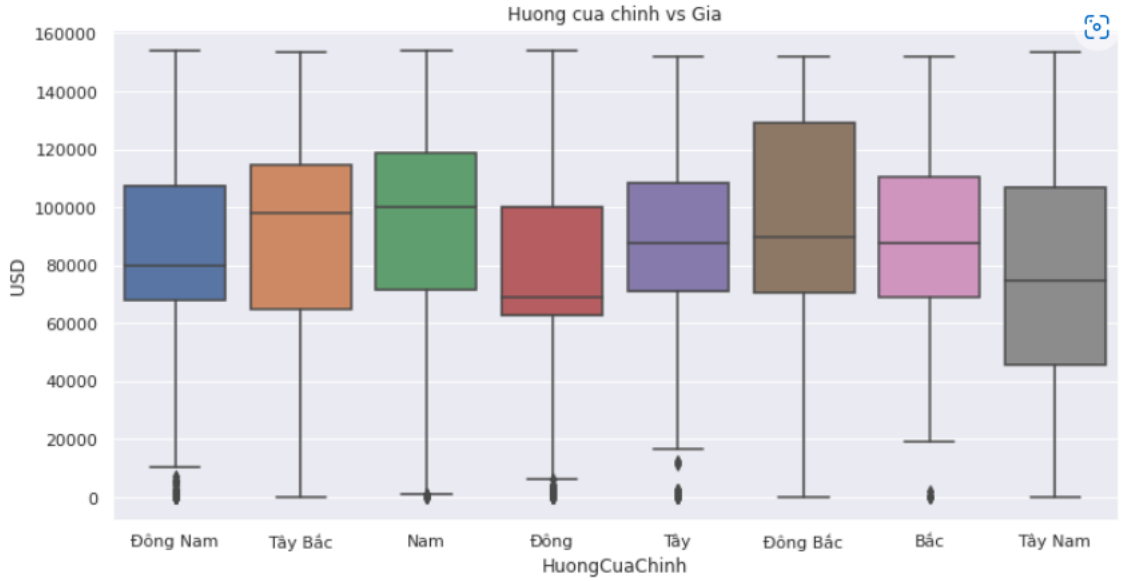

Figure 8: HuongCuaChinh & Gia

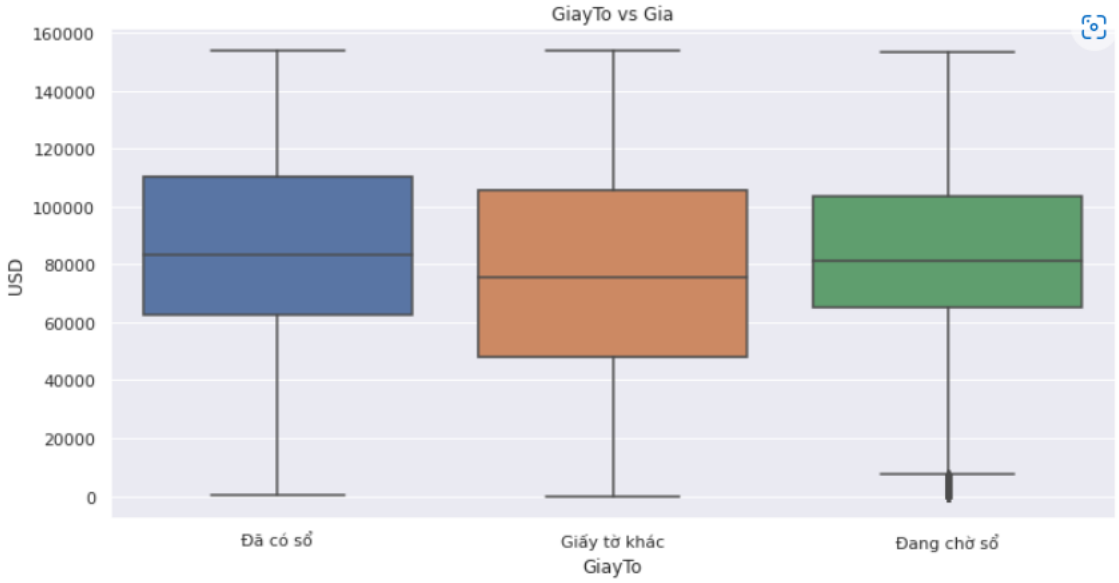

Figure 8: GiaTo & Gia

Similarly, for HuongCuaChinh, so is GiayTo.

5. Modeling

5.1 Data reduction and transformation

Because our data still has many NaN value, I decided to delete all columns with more than 70% NaN and delete all rows that appear NaN. Because the variables used when included in the analysis are not completely appropriate. So, for the category variables, I will use the Python dummies function to separate them. Finally, the variables used in the model are described as follows:

Variable	Description	Data type
Dependent variable		
Gia	price of the house in HCM City	Numeric
Independent variable		
DienTich	living area in Square feet	Numeric
Phongngu	number of bedrooms	Numeric
PhongTam	number of bathrooms	Numeric
TinhTrangBDS_Đã bàn giao	Yes/No	Binary
Loai_ Tập thể, cư xá	number of bedrooms	Binary
Loai_Chung cư	Yes/No	Binary
Loai_Duplex	Yes/No	Binary
Loai_Officetel	Yes/No	Binary
Loai_Penthouse	Yes/No	Binary
GiayTo_Đang chờ sổ	Yes/No	Binary
GiayTo_Đã có sổ	Yes/No	Binary
Quan_ Huyện Củ Chi	Yes/No	Binary
Quan_ Huyện Hóc Môn	Yes/No	Binary
Quan_ Huyện Nhà Bè	Yes/No	Binary
Quan_ Quận 1	Yes/No	Binary
Quan_ Quận 10	Yes/No	Binary
Quan_ Quận 11	Yes/No	Binary
Quan_ Quận 12	Yes/No	Binary
Quan_ Quận 2	Yes/No	Binary
Quan_ Quận 3	Yes/No	Binary
Quan_ Quận 4	Yes/No	Binary
Quan_ Quận 5	Yes/No	Binary
Quan_ Quận 6	Yes/No	Binary
Quan_ Quận 7	Yes/No	Binary
Quan_ Quận 8	Yes/No	Binary
Quan_ Quận 9	Yes/No	Binary
Quan_ Quận Bình Thạnh	Yes/No	Binary
Quan_ Quận Bình Tân	Yes/No	Binary
Quan_ Quận Gò Vấp	Yes/No	Binary
Quan_ Quận Phú Nhuận	Yes/No	Binary
Quan_ Quận Thủ Đức	Yes/No	Binary
Quan_ Quận Tân Bình	Yes/No	Binary
Quan_ Quận Tân Phú	Yes/No	Binary

Table 2: Describing the variables included in the model.

I use StandardScaler to transform data such that its distribution will have a mean value 0 and standard deviation of 1.

5.2 Model selection

The models selected for forecasting are: Linear Regression, Decision Tree Regression, Random Forest Regression.

5.2.1 Linear Regression

It is an algorithm that is used for estimating the real values (cost of houses, number of calls, complete deals and so forth) in view of continuous variable(s). Here, we try to find a best fit line which can get us the relationship between independent and dependent variables.

5.2.2 Decision Tree Regression

It is a tree-based model and is a supervised learning algorithm which can be used regression models here the nodes are decision points having conditions the results of which then extends the tree into more nodes

5.2.3 Random Forest Regression

Forest is a kind of democratic collection of many decision trees, where to tackle the problem of overfitting of a single Decision tree we now do voting, and the most voted class wins and is the result for your target observation.

5.3 Evaluation

5.3.1 MAE

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size. (Wikipedia)

$$MAE = \frac{\sum_{t=1}^{n} |\varepsilon_t|}{n} = \frac{\sum_{t=1}^{n} |Y_t - \hat{Y}_t|}{n}$$

In our case these continuous variables are listing price value and predicted price value of the house property.

5.3.2 MSE

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. (Wikipedia)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

5.3.3 RMSE

MSE sometimes increases the actual error, making it difficult to realize and understand the actual error amount. This problem is resolved by the RMSE measure, which is obtained by simply taking the square root of MSE.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

The goal to choose the best model is RMSE and MSE, the smaller the MAE, the smaller the MAE because they said errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable.

Model	MAE	MSE	RMSE
Linear	863201531.6	3.24E+18	1798987603
Decision Tree	210020923.3	4.83E+17	694692763.5
Random Forest	323022776.2	8.09E+17	899677466.4

Table 3: Model comparison with RMSE, MSE, MAE.

As in the previous discussion the evaluation ratio of each model is equal to its evaluation MSE, MAE, RMSE. The smaller evaluation ratio, the higher accuracy of the model's prediction.

It can be seen from table 3 that Decision Tree Regression delivers a prediction result better than linear regression, Random Forest. With MEA of 210020923.3 (corresponding to an error of nearly 200 million VND), MSE is 4.83E+17, RMSE is 694692763.5 (corresponding to nearly 700 million VND). Besides, the Linear Regression model is no longer suitable for this dataset.

So it can be said that with this dataset, the most suitable model to predict house prices is Decision Tree Regression.

6. Conclusion and discussion

The main goal of this research paper is to help home buyers and sellers not be overpriced or undervalued based on machine learning models. Through the study, Decision Tree Regression is the best model to predict house prices in HCMC when it gives the smallest RSME, MAE, and MSE. Of the variables included in the forecast, most are significant for predicting house prices. It is completely consistent with some previous studies such as Changchun Wang and Hui Wu (2018). Besides, depending on the data set, the traditional Linear model is no longer highly accurate when we forecast.

Besides, the price of a house posted on Cho Tot sometimes does not reflect its true value and inadvertently causes this assessment to be overvalued or undervalued. The fact that when buying/selling a house depends on many other situations such as: land price fluctuations, how is the real estate market, etc. The results of this study are for reference only. However, this will be the most basic thing when you want to buy / sell a house.

Reference

[1] Madhuri, CH.R., Anuradha, G. and Pujitha, M.V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. [online] IEEE Xplore. doi:10.1109/ICSSS.2019.8882834.

[2] www.proquest.com. (n.d.). A new machine learning approach to house price estimation - ProQuest. [online] Available at: https://www.proquest.com/openview/c09e4fd88a8f972d517322da1ef9c769/1?pq-origsite=gscholar&cbl=2041201 [Accessed 23 Jun. 2022].

[3] Rakhra, M., Soniya, P., Tanwar, D., Singh, P., Bordoloi, D., Agarwal, P., Takkar, S., Jairath, K. and Verma, N. (2021). Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review. Materials Today: Proceedings. doi:10.1016/j.matpr.2021.03.261.

[4] Phan, T.D. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (iCMLDE). doi:10.1109/icmlde.2018.00017.

[5] Fan, C., Cui, Z. and Zhong, X. (2018). House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018. doi:10.1145/3195106.3195133.

[6] Thamarai, M. and Malarvizhi, S.P. (2020). House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering and Electronic Business, 12(2), pp.15–20. doi:10.5815/ijieeb.2020.02.03.

[7] Zaman, U., Waqar, M. and Zaman, A. (n.d.). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. Soft Computing and Machine Intelligence Journal, [online] (1), p.2021. Available at: https://www.koreascience.or.kr/article/JAKO202122260767075.pdf.

[8] Wikipedia Contributors (2019). Mean absolute error. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Mean_absolute_error.

[9] Hùng, M. (2022). BÁO CÁO ĐỒ ÁN CUỐI KÌ - KHOA HỌC DỮ LIỆU ỨNG DỤNG. [online] GitHub. Available at: https://github.com/HungTrinhIT/FinalProject-Datascience [Accessed 23 Jun. 2022].

[10] Madhuri, CH.R., Anuradha, G. and Pujitha, M.V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. [online] IEEE Xplore. doi:10.1109/ICSSS.2019.8882834.

[11] kaggle.com. (n.d.). House price prediction. [online] Available at: https://www.kaggle.com/code/narayanyadav/end-to-end-house-price-prediction?scriptVersionId=99062084 [Accessed 23 Jun. 2022.