Real Estate Valuation Project

This document summarizes the insights I got from the EDA on the UCI house price dataset. I also included the final results from the Machine learning model I built, based on several features such as house_age, distance to the nearest MRT (Mass Rapid Transit) station e.t.c. I completed the project in about 2 hours.

Insights

1. The house_age variable has values between 0 to ~44 years with three different peaks. The highest peak is for houses that are between 13 to 19 years old.
2. The majority of the houses (approximately 67%) are located 0 to 1000 meters (i.e., 1km) away from the nearest MRT (Mass Rapid Transit) station.
3. I noticed that approximately 69% of transactions happened in 2013 - the majority occurred in January, May, and June 2013.
4. The most occurring number of convenience stores is 5 and 0. Approximately 16% (i.e., 67/413) of houses don't have convenience stores around in the UCI real estate valuation dataset. Most likely, these houses aren't located in highly residential areas.
5. I discovered that more expensive houses have a higher number of convenience stores than less costly houses (The average house price for real estate with at least a single convenience store is about 66% higher than houses without any convenience store around).
6. I also noticed that the average house prices tend to decrease as the distance between a house and the nearest MRT station increases. In essence, houses closer to MRT stations are relatively more expensive than houses farther from an MRT station.
7. More than 50% of the houses have anything between 1 and 6 convenience stores.
8. Over 69% of the houses are clustered around latitude [24.96^{o},24.98499^{o}] and longitude [121.53^{o},121.55^{o}].
9. The mean and median house price is 37.98 and 38.34 Ping respectively. Also, over 80% of the houses cost around most houses cost around 20 to 54.59 Ping.
10. I saw a single house that cost 117.5 Ping. This is an outlier, and I had to remove the data points from the dataset where house price >= 117.5
11. The target_variable is a bit skewed, and on average, house prices were relatively higher in 2013 than in 2012. Perhaps, could be linked to a situation that occurred around 2013
12. As the age of a house increases, the average house price is likely to decrease. However, this doesn't hold for houses above 35 years old because the average price increased from 33 to 42 from the 9th to 10th decile.
13. I observed the following correlation between between target variable (house price) and independent variables: distance_to_nearest_MRT_station(-0.69), number_of_convenience_stores (0.61), latitude (0.56), longitude (0.55), and house_age (-0.21).

Evaluation Metrics

Because the target variable is continuous and the task is a regression problem, I used a regressor rather than a classifier. I also choose evaluation metrics appropriate for a regression problem, such as Mean Absolute Error, Mean Squared Error, and R-Squared. After evaluating the model performance, I had the following results on train and test set respectively:

S/N	Evaluation Metric	Train set(1dp)	Test Set(1dp)
1	Mean Absolute Error	-4.7	4.9
2	Mean Squared Error	-43.4	47.7
3	R-squared	0.73	0.72

Next Steps

The following are some other things I'd consider If I were to dedicate more time to this data science project.

1. Hyper-parameter tuning using Gridsearch CV - to get the best set of parameters that increase the performance of the Random Forest Regressor
2. Train other regressors: Xgboost, SVM, Linear e.t.c and see if they perform better than Random Forest Regressor
3. Carry out some geo-location analysis to see if I could do more with the latitude and longitude