Capstone Project

Yusuke Yamamoto

Introduction

In this report, we are going to analyse the housing price in Tokyo in Japan. In the course work, we looked at the Boston housing data set where we built a model using the number of rooms in home, neighbourhood poverty level and student-teacher ratio of nearby schools. The goal of this project is to understand whether the housing market in a different city has different dynamics where other factors determine the housing price. Predicting the housing price is important in many respects. One is to understand the economy of a country. It is imperative for policy makers to understand if the housing price is so high that the county is facing the bubble economy.
Another aspect is personal. Buying a house is an important decision for anyone because the price is high and most people need to apply for a mortgage. When it comes to selling a house, you also want to know if you are selling it at a fair price. Understanding the fair market value of a house you are trying to buy or sell is of significant interest to anyone.

The price of houses in Japan is said to be correlated mainly with the age of the house. Some areas are popular so the price tends to remain the same. People in Tokyo normally commute by train so which train lines are close to the house should also determined the house price. The location is also relevant, and it is often reported that the housing price in the west part of Tokyo tends to be higher than that in the east part. Some areas are known to be popular for well-off people. Another factor to consider is whether it is a house or apartment. Because the land is limited in the centre of Tokyo, people normally buy apartments. People who want to buy a house go to a suburb of Tokyo. The price of a house includes the price of its building and the land, so the pricing mechanism is considered to be different from that for apartments.

The problem we are trying to solve in this report is to build a regression model that can accurately predict the price of apartments in Tokyo. The model is to predict the price from different characteristics of the apartments such as size, age and location. The accuracy of the model is quantifiable by comparing the prediction and the actual price. Mean Squared Error (MSE) is used for the evaluation metric. This is a regression problem where we predict the price of apartments, and the predicted values take non-negative real numbers. R2 can be an alternative metric. In theory, it takes values between 0 and 1, and the close the value is to 1, the better the model is. The problem is that this holds only when we calculate R2 for the same data set that we use to train the model. Another problem is that R2 changes when you change the number of variables in the model, so you cannot compare two models that have different number of variables. For these reasons, we prefer MSE to R2.

Datasets and Inputs

We use data from Ministry of Land, Infrastructure, Transport and Tourism (MLIT) which cover the price of houses and apartments in Japan (http://www.land.mlit.go.jp/webland/servlet/MainServlet). The price is rounded to the nearest one million yen (approximately 10,000 USD). Other variables such as age, month of the transaction and the size in square meters are obfuscated. This is due to the privacy reason so that the users of the system cannot tell the exact price or location of each house and apartment.
The data have 28 features. MLIT web page has a data set for each quarter, and there are about 3500 records for the price of apartment in Tokyo per quarter. The price of apartment in Tokyo is used in the project.

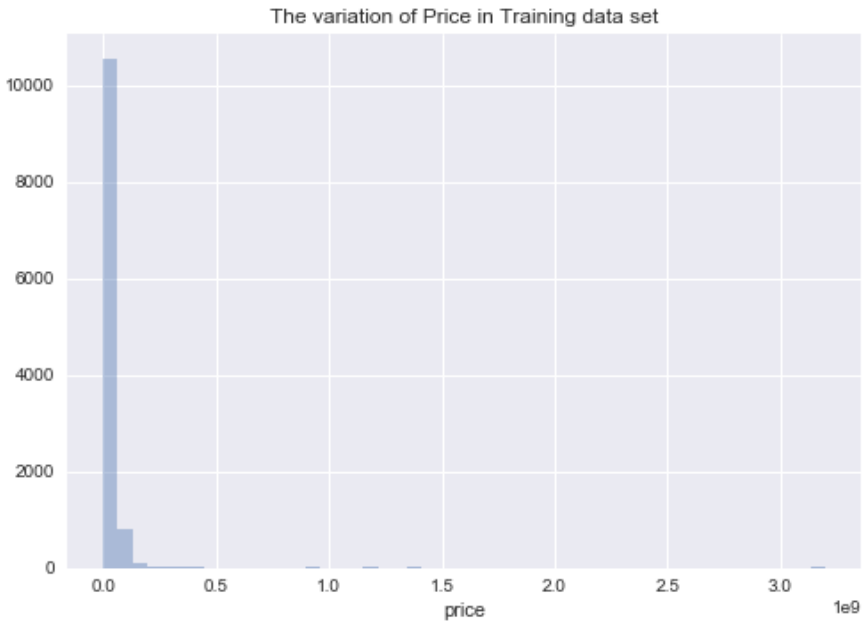
We build a supervised model that can predict the price of apartments using the characteristics of the apartments such as age and location. To build a model, we use data in Q2 2015 through Q1 2016 as the training data set, Q2 2016 as the validation data set and Q3 2016 as the test data set.

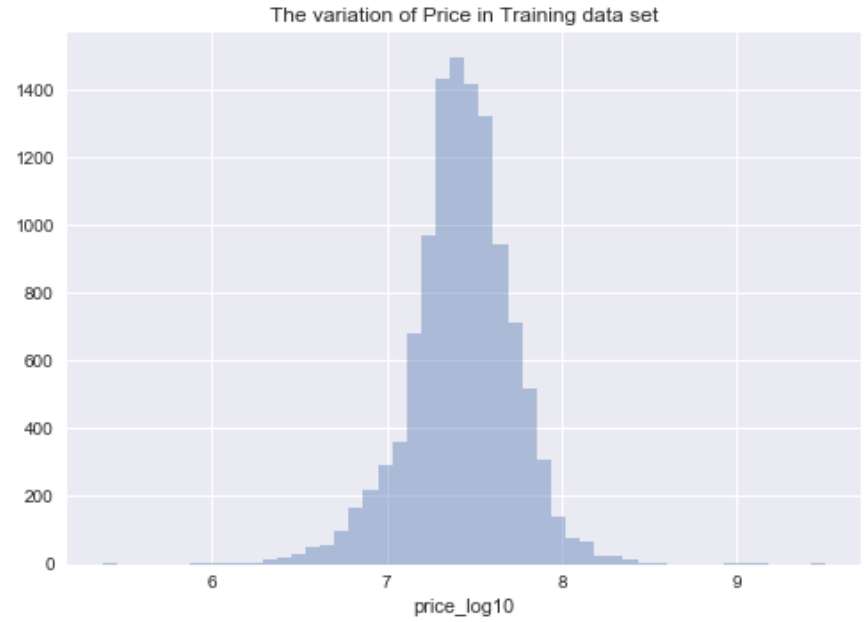
Analysis

We first look at the distribution of the data set. This is to understand how price is distributed, what other variables seem to be correlated in what way, and what variables appear to be useful to predict the price. In the training data set, the price distribution is summarized in a table below. The average price is 33.6 million JPY (310K USD). 50% of the records take values between 18 million JPY and 40 million JPY.

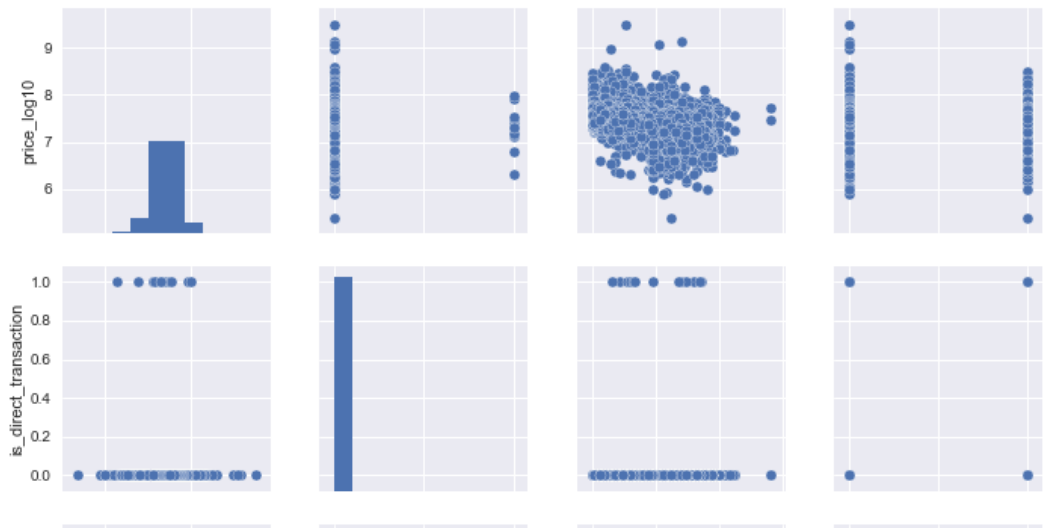
Min	25%	Median	Mean	75%	Max
0.24 M	18 M	27 M	33.6 M	40 M	3,200 M

The histogram of price shows very skewed distribution. Almost all apartments are priced below 50 million JPY (460,000 USD), and you can observe some expensive apartments that cost well over 100 million JPY (920,000 USD).
The skewed distribution of price can be difficult to model. Taking the log of price in base 10 makes the distribution almost symmetric.

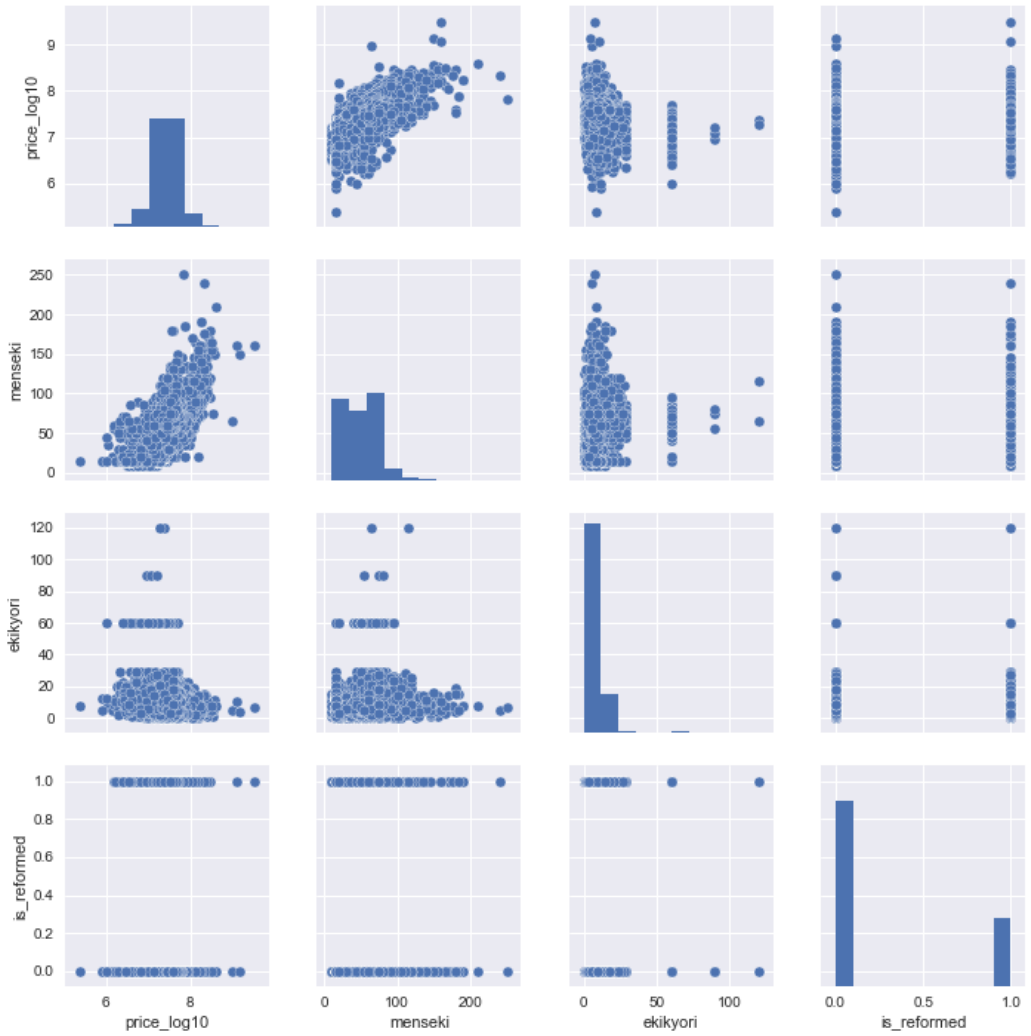




We also look at the relationships of log price with other variables. Graphs below show the histogram of each variable, and the scatter plot of every combination of two variables. Most variables do not show an apparent correlation. The size of the room



(menseki) appears to be positively correlated as the scatter plot shows so that the



larger room costs higher, which is reasonable.

A positive correlation can also be observed between price and age. The older the room is, the lower the price is. The visual inspection does not show other correlations.

Data Preprocessing

From the original data set, we included records for apartments only that were sold between Q2 in 2015 and Q3 in 2016. The data between Q2 in 2015 and Q1 in 2016 were used as the training data, the data that were sold in Q2 in 2016 were used as the validation data, and the data in Q3 2016 were used as the test data set.
The distance between an apartment to the nearest station in minutes is available in the data set. Some records show only approximate distance such as 30 minutes - 60 minutes, 1 hour - 1.5 hours, 1.5 hours - 2 hours and more than 2 hours. These were replaced with 60 minutes, 90 minutes, 120 minutes and 120 minutes respectively.
Building-to-land ratio and floor-area ratios were denoted in percentages. These ratios were conveyed into decimal points.
In order to calculate the age of each apartment as of the transaction date, we calculated it from the different between the year it was build and the transaction year. Some records have the building year before 1945, and these were all treated as being built in 1945.
The data included character variables, such as town's name, area's name, the name of the nearest station, material of the apartment, zoning. We converted these into dummy variables so that each column has either 0 or zero entries. Regarding the names of the station and area's name, there are just too many types in these variables, so we did not use these variables. In order to avoid the model identifiability problem, we dropped one level from each dummy variable. In the end, we have the following dummy variables.

- town: indicates where the apartment is located in Tokyo.
- material: what the apartment building is made of; reinforced concrete (RC), steel reinforced concrete (SRC) and steel.
- purpose: how the room is used for; housing, office or other.
- zoning: indicates how certain land uses are permitted or prohibited such as residential and commercial. https://en.wikipedia.org/wiki/Zoning#Japan

Finally, any rows that has missing values for at least one variable were excluded. The number of records for the training, validation and testing reduced to 11479, 2895 and 1880 respectively from 14733, 3664 and 2464. There are 77 predicting variables.

Benchmark Model

We use a simple linear regression model using all the features in the data set as the benchmark model. The target variable is the log transformed price. The performance is measured against the validation data set. The table below shows the performance of the linear model.

	Training MSE	Validation MSE

Linear Model	0.0150	0.0196

The benchmark of the prediction model is 0.0196 for the validation data set.

Model Building

We build a model using Random Forest. The method makes a collection of regression trees and takes the average of the estimate.

	Training MSE	Validation MSE
Random Forest Model	0.003	0.016

As the table above shows, the training MSE dropped significantly from 0.015 of the benchmark to 0.003. The problem is that the validation MSE did not drop very much and it decreased to 0.016. To improve the validation MSE further, we also tried to tune the hyper-parameters using randomized parameter optimization. The tuned model gave rise to a slightly better result for the validation data set as below.

	Training MSE	Validation MSE	Testing MSE
Tune Random Forest Model	0.0018	0.0137	0.0155

This is about 30% (=1 - 0.0137/0.0196) improvement from the benchmark model, and we choose this as the final model. One thing to notice is that the validation MSE is nearly 10 times as high as the training MSE. The difference between the training and the validation MSE in the benchmark linear model was small (0.015 vs 0.0196), so the gap between the training and the validation performance became wider.
Finally, we fit the model with the testing data set, and the performance on the test data set was 0.0155.

Conclusion

We built a Random Forest model to predict the apartment price in Tokyo area that were sold between Q2 2015 through Q1 2016 and tested it on the data that were sold in Q3 2016. The model predicts the log-transformed apartment price, and we were able to achieve the better performance than the benchmark model, although we were not able to get the same level of improvement in the validation data set as we get in the training data set.
One possible reason is that the structure of the housing market is completely different from quarter to quarter. For example, the average market price in Q2 2016 may be

higher than that in Q1 2016. The model does not have variables that are related the such market trend.
Another possibility is that the RF model may not be suitable to predict the apartment price. We did not test other models such as XGBoost or SVM. Correctly tuned models using those methods may have given rise to better performance.