Capstone Proposal

Yusuke Yamamoto

Proposal

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.

Domain Background

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, normally apartments are bought. 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.

Problem Statement

The problem 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.

Datasets and Inputs

We use data from Real Estate Information Network System (REINS) which cover the price of houses and apartments in Japan (http://www.contract.reins.or.jp). The price is rounded to the nearest one million yen (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. REINS 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.

Solution Statement

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.

Benchmark Model

We use a simple linear regression model using all the features in the data set as the benchmark model.

Evaluation Metrics

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.

Project Design

First, we visualise the training data set. This is to understand which factors are correlated with the price in what way, and to understand the distribution of the price and variables, which will help us decide whether we need to transform the variables. Second, we do the pre-processing of the training data set based on the finding of the visualisation part. This involves the normalisation, transformation and the outliers removal. Third, we build a supervised model using Random Forest. We start with a simple model using one or two variables and then refine it to a more complex model. In the refinement process, we are going to look at samples that has a large deviation.