Data MJB

Overview

In Shanghai the prices of real estate are skyrocketing with no correct estimation of price of land for a particular area. Hence we did a project where we solve this problem, estimating the correct price, helping people in making their decisions.

What Problem you have solved

Prices for apartments in Shanghai vary according to districts and sub districts alot. The prices depend on a number of features like the neighbourhood, orientation of the apartment, number of bedrooms and bathrooms, the floor area, whether the society has a garden, a parking lot and many such features. But there is no proper compilation of data for real estate in Shanghai. Large amounts of data is present online with a random collection of some of these features. Therefore we decided to estimate the price per square meter and help people know the real price for a particular neighbourhood.

To solve this problem, we began by scrapping large amounts of data. We wrote different scrappers to collect the vasts amount of data from a number of websites. After collecting the data, the next biggest challenge was to compile it in one location and clean the data. Many of the rows had empty cells, making it difficult for us to work with the data. Once we were done with data cleaning, we had around 18000 rows of non repeated data. With this in hand, we trained our data model using different algorithms. We trained our model on Dense Neural Network, Random Forest Regressor and Random Forest Classifier. The best accuracy was given by Random Forest Classifier when the error rate was kept to $\pm$ 15%.

How Solution Works

We take features like the number of rooms and neighbourhood, and then use a model to get a prediction on price. The accuracy used in regression is defined in the following way.

(1)A prediction is accurate if

$$|prediction - actual\ price| < \beta * actual\ price$$

(2)The accuracy of a regression is:

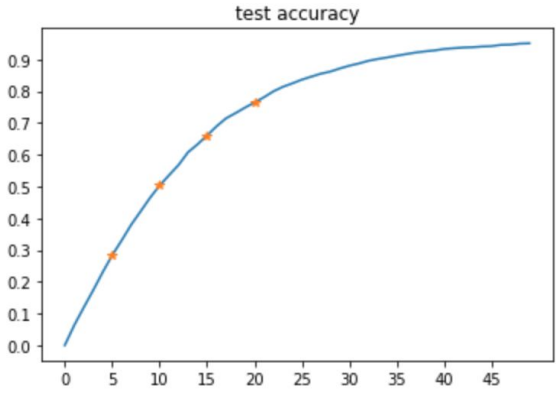
$$accuracy = accurate\ predictions / total\ predictions$$

1.DNN of 55% accuracy

The DNN is constructed on 3 hidden layers, with 500, 300,100 neutron units respectively. The activation function is “tanh” instead of relu because it is a regressor. In this way, we get a accuracy of 55% with ß=15%.

2.Random forest Regressor of 66% accuracy

In random forest regressor we just use the default parameters except that n_estimater is set to be 100. From this model we achieve a 66% accuracy with ß=15%.



ß(%)

3.Random forest Classifier of 75% accuracy(5 labels)

In random forest classifier, we label the price with 5 categories,from “very high” to “very low”. In this case, we can achieve a 75% accuracy.

Future work

In the future, we are going to separate the price into 2 parts-Price index and Price Premium in order to predict both the price of the house and the future price of the house. Also, we are going to complete our User Interface.

Submit a 2- paragraph **"news-like"** story of your project-idea, method and experience. This teaser summary is useful for communicating about your project on the data-x web pages .

Our Project aims at helping people with a better understanding of the house price via technical prediction, since in Shanghai the prices of real estate are skyrocketing with no correct estimation of price of land for a particular area.