

Home Price Prediction

Aaryan*,Himanshu Singh**

*,**(Department of Computer Science Engineering, RKGIT Ghaziabad)

Under Guidance of Ms niketa

(Department of Computer Science,RKGIT Ghaziabad)

Abstract

Home price prediction is an important topic for the real estate .It is useful because it help various dealer, builder, contractor and buyer to assist the know the value of the home at the actual area on the basis of the price predict by this project any person can make the future plan and make the budget according to it. This project also help for the people which are know the price of the country which is completely new to it. In this project there are the technology used for making the model and the model using the regression technology used in it. The idea behind this project is come from when the person especially student come in the new city for education and jobs they don't know about the lifestyle cultural and price of the various home this can be able to the price of the Bangalore city. In our project the dataset used of the Bangalore city and this dataset take from the online site.

Keywords- regression, dataset, learning algorithm, training set, unlabeled data.

Introduction

This project has been done as part of my course for the B.tech Engineering RKGIT ,Ghaziabad .Supervised by Mr , I had five months to fulfill the requirements in order to succeed the module. Every weeks, a gathering was organized to point out and report my progress and fix the following objectives.

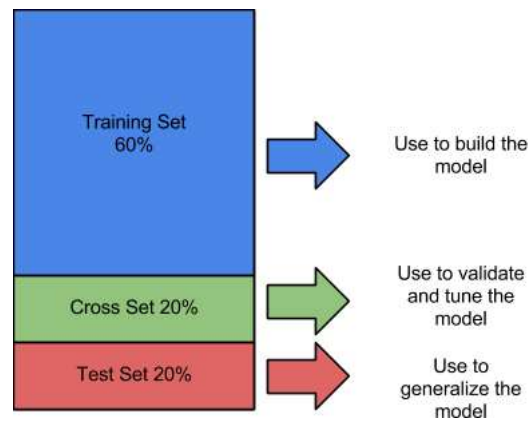
The fact that we are going to make estimations, predictions and provides the power for machines to seek out by themselves is both powerful and limitless in term of application possibilities. We will use Machine Learning in Finance, Medicine, almost everywhere.[1] That's why I made a decision to conduct my project round the Machine Learning. Technology and tools wise this project covers,

- 1) Python
- 2) Numpy and Pandas for data cleaning
- 3) Matplotlib for data visualization
- 4) Sklearn for model building
- 5) Jupyter notebook in anaconda, visual studio code and pycharm
- 6) Python flask for http server
- 7) HTML/CSS/Javascript for UI

The Machine Learning part is about trying to seek out the most effective learning algorithm for a given problem whether or not it is highly conditioned by how well the information has been processed and tune some parameters to enhance it. Counting on the matter, if it is supervised (meaning we build a model from labeled training set, the worth of the dependent variable is known) or if it unsupervised (the model is made on unstructured and unlabeled data), if it is a regression or classification problem, many learning algorithm exist each with their benefits and downsides.

Sampling

The training set as its name indicates it is wont to train the learning algorithm. The cross validation set is employed to validate the model and make optimizations. Indeed, because the model is built from the training set, we can't check it on the identical set because the result would be overly optimistic. It are often also useful for optimization as I will show within the following sections. The test set is used only to figure out how well the learning algorithm is generalized, meaning how it performs with unknown data.



Problem Statement And Solution

To design a website that predict price of home on basis of loacality, bhk, no. of bathrooms, gross area. The first problem was where am i able to get the info to make an oversized enough dataset since I would like to be able to predict the worth for a given apartment according to the agency chosen.

The idea is to simulate the human behavior on different websites and parse the information to avoid wasting them

Cleaning, is that the primary module called to clean the item and verify that every one the data in it correspond to the pattern accustomed extract it. As an example, suppose that we've several estates on a given web content. Each estate is presented using images and much of basic information like, the gross area, the saleable area, the price, etc. When the spider extract them, it's not unlikely that some noise (like extra characters) was present with the worth, or just the worth doesn't exist. The cleaning module removes the noise, and check that all the values are not empty, otherwise the item is dropped. This is often done for simplicity, indeed, it can be better to try to inference them later. After the cleaning part done, the item is distributed to the formatting module.[2]

- **Formatting**, the second module is used to format the item's values as we would like. A basic example may be for the price, initially got being string type, is converted as float. This is done for each numeric values. The item formatted is then sent to the last module called Integrating.[3]

- **Integrating**, this module, the last one, is essentially the one in charges of saving the things that format that we wish. It also checks that there is no redundancies between the tuples. In my case, I made a decision to save them in an excel sheet for each website.[4]

Indeed, two attributes highly correlated (not using price) could be useless because they will not have a great impact on the regression result and should be reduced to one (cf: Principal Component Analysis). [5]

Conclusion

In this paper we have discussed basic needs of home price prediction in this project there are various type of problem solve for buying selling and estimate the price of any home in the specified city according to the input in the webpage like bhk,yard, bathroom location etc.and the efficiency of project depend upon the dataset collection.

References

To conduct this project the following tools have been used :

1. Python 3.1
2. Pandas (Library) : <http://pandas.pydata.org/>
3. Numpy (Library) : <http://www.numpy.org/>
4. Scikit-learn (Library) : <http://scikit-learn.org/>
- 5.The techniques used to visualize and preprocess the data has been inspired from the book "Data Mining Concepts and Technique".