**Real Estate Sales Price Prediction – Using Regression**

Abstract—

In this machine learning paper, I have tried to analyze the real estate property prices in New York city area. The information on real estate listings is extracted from my company’s dataset (HIGHLINE Residential LLC., I am working as an Intern here.). We are predicting the sales price of real estate properties based on features such as geographical locations, living area, number of bed rooms, baths, etc. Additional geographic features such as nearest subway station and a few others might also be considered. We are using and comparing regression methods such as linear regression, gradient boosting algorithm, Random Forest algorithms and might consider comparing KNN. We will present the details of the prediction questions, the analysis of the real estate listings, and testing and validation results for the different algorithms. In this paper, only an over view of the first two methodologies is mentioned as the work has just started.

Machine Learning Model—

The dataset we will work upon consists information about the location of the house, price and other aspects such as square feet, number of beds, baths etc. When we work on these sort of data, we need to see which column is important for us and which is not. Our main aim is to make a model which can give us a good prediction on the price of the house based on other variables. We are going to use Linear Regression for this dataset and see if it gives us a good accuracy or not.

We start by importing libraries and dataset and then we see how the data looks like and use describe function to see the percentile’s and other key statistics.

Starting, by importing libraries and reading dataset using pandas library. We might need to clean the dataset to remove missing values, outliers from the dataset and etc.

Knowing more about the dataset by using describe() function.Then, we will see some visualization and also going to see how and what can we infer from visualization. Like..

Which is the most common house (Bedroom wise)?

For example, we can see which is most common bedroom number. Looking at this problem from a builder’s perspective, sometimes it’s important for a builder to see which is the highest selling house type which enables the builder to make house based on that. In India, for a good locality a builder opts to make houses which are more than 3 bedrooms which attracts the higher middle class and upper class section of the society.

By this we can find out that for example 3 and 4 bedrooms are highest selling. That arises further questions like, at which locality?

We can find that out by visualizing the location of the houses based on latitude and longitude. We are able to see the common location and how the houses are placed.

Using seaborn, and we get the visualization. Joinplot function can help us see the concentration of data and placement of data and can be really useful. We can infer important things from this visualization such as, for certain latitude there are many houses, which would mean that maybe it’s an ideal location. But longitude wise it differs. Which would mean that most of the buy’s has been for that particular location.

Similarly, we can find out other things such as, how common factors are affecting the price of the houses? We can create a model to which would predict the price of the house based upon the other factors such as square feet, water front etc.

Linear Regression—

In simple words a model in statistics which helps us predicts the future based upon past relationship of variables.

We use train data and test data, train data to train our machine and test data to see if it has learnt the data well or not.

For linear regression we use sklearn (built in python library) and import linear regression from it. we know that prices are to be predicted, hence we set labels (output) as price columns. We again import another dependency to split our data into train and test. Once we have train data, test data and labels for both we can fit our train and test data into linear regression model.

After fitting our data to the model we can check the score of our data i.e., prediction.

For building a better prediction model, we can use gradient boosting regression. It is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

We again fit our training data into the gradient boosting model and check for accuracy. And we might get a better accuracy. We can try to put random seeds and check if it changes the accuracy of the data or not.

References—

[1] <http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf>

[2] <https://towardsdatascience.com/regression-analysis-model-used-in-machine-learning-318f7656108a>

[3] <https://yalantis.com/blog/predictive-algorithm-for-house-price/>

[4] http://arno.uvt.nl/show.cgi?fid=134740

<https://smartech.gatech.edu/bitstream/handle/1853/31763/Corsini_Kenneth_R_200912_mast.pdf>

**-Rucha Sonawane (r34464n@pace.edu)**