House Price Prediction Using Regression Algorithms

1st Md. Jusef
Dept. of CSE
East West University
2017-2-60-160@std.ewubd.edu

2nd Sk. Amir Hamza
Dept. of CSE
East West University
2017-1-60-091@std.ewubd.edu

3rd Tanim Hasan Mahmud
Dept. of CSE
East West University
2017-1-60-130@std.ewubd.edu

***Abstract*—people are cautious when they are attempting to purchase a new house with their budgets and advertise methodologies. The objective of the paper is to figure the coherent house costs for non-house holders based on their financial conditions and their goals. The paper includes expectations utilizing distinctive Regression techniques like Linear Regression, Random forrest Regression, Decision tree Regression. House cost expectation on a data set has been done by utilizing all the over mentioned techniques to discover out the finest among them. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house. A few of the related variables that affect the cost were moreover taken into thought such as physical conditions, concept and area etc.**

***Index Terms*—house price prediction,linear regression,random forest,decession tree**

I. Introduction

The most inspiration of the project forecasting varieties on house price was to make the most excellent thinkable forecast of house costs by using appropriate calculations and finding out which among them is best appropriate for anticipating the cost with low error rate. This is an curiously issue since most of the individuals will eventually buy/sell a house. This problem allows us to learn more about the housing market and helps with making more informed decisions. The analysis that were done in this paper is basically based on the Boston housing data. In this paper we try to use suitable Regression techniques for solving the issue. We used Random Forrest Regression, Linear Regression and Decision Tree Regression and calculate the error rate.

II. Dataset

A. Features Description

In this study, I used Boston Housing dataset. The dataset comprises 13 input features and 1 target feature. The Boston Data frame has 506 rows and 14 columns. Attribute information are given below:

1. CRIM per capita crime rate by town
2. ZN proportion of residential land zoned for lots over 25,000
sq.ft.
3. INDUS proportion of non-retail business acres per town
4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX nitric oxides concentration (parts per 10 million)
6. RM average number of rooms per dwelling
7. AGE proportion of owner-occupied units built prior to 1940
8. DIS weighted distances to five Boston employment centres
9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per 10,000 doller

11. PTRATIO pupil-teacher ratio by town
12. B 1000(Bk - 0.63) power 2 where Bk is the proportion of blacks
by town
13. LSTAT percent lower status of the population
14. MEDV Median value of owner-occupied homes in 1000's doller

III. Methodology

I was using three algorithms for evaluate the model. For Figuring out best model I take some process like as linear regression, Decision tree and random forest algorithm.

A. Plotting Histogram

This study I used plotting Histogram for data analyze easily. All features histogram are here. From "CHAS" features I see that counts of 1 is greater than 0. From "age" features I see that more than 60 building age is 100 years. So we can easily read dataset through histogram.

B. Missing Attributes

To take care of missing attributes we have three options
1. Get rid of the missing data points
2. Get rid of the whole attribute
3. Set the value to some value (0, mean or median)

Identify applicable funding agency here. If none, delete this.

Following this option, we can easily handle with missing data. Picking the right evaluation metric Picking the right evaluation metric will help us to evaluate whether our model's performance is good. Through root mean square error, mean and standard deviation I evaluate my model.

C. Train test splitting

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

• Train Dataset: Used to fit the machine learning model.
• Test Dataset: Used to evaluate the fit machine learning model
We used 80 percent of the known dataset as training set and 20 percent used as test data set. Each record in the dataset denotes X and Y values, where X is a set of attribute values and Y is the class of the record which is the last attribute in the dataset.

D. Linear Regression

Multiple Linear Regression is the most common form of linear regression. The steps to perform multiple linear regression are almost similar to that of simple linear regression. In linear regression it predicts a independent value by using dependent values. In our model average number of rooms, full-value property-tax, proportion of owner-occupied is considered as dependent value. The house price Is independent value which will predict our model.

E. Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numeric

F. Random Forest

Random Forest Regression is also supervised learning algorithm. A Random forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

IV. Discussion and Result

After evaluating Linear regression, Decision tree and Random forest regression algorithms, I observe that Random forest regression give me best result both of them. So I used Random forest regression algorithm in my model. I separated my model as regression.joblib file.

I input all features value and my model predict me price is 23.578.

A. Figures and Tables

All performance matrics value is given below:

TABLE I
Performance matrics value

Algorithm	**Table Column Head**		
	RMSE	***MEAN***	***STANDARD DEVIATION***
Linear Regression	4.143	5.037	1.059
Decision Tree Reg.	3.972	4.271	0.777
Random Forest Reg.	2.913	3.301	0.704

V. Conclusion

In this study our models are trained with Boston housing data utilizing model based linear regression, decision tree and random forest regression algorithm. Given my dataset used in this report, my main conclusion is that Random Forest is able to generate comparably accurate price. My model has shown that machine learning algorithms like Linear Regression, Decision Tree and Random Forest are promising tools for housing price prediction.

References

[1] thamarai2020house, title=House Price Prediction Modeling Using Machine Learning., author=Thamarai, M and Malarvizhi, SP, journal=International Journal of Information Engineering & Electronic Business,year=2020

[2] madhuri2019house, title=House price prediction using regression techniques: a comparative study,author=Madhuri, CH Raga and Anuradha, G and Pujitha, M Vani,booktitle=2019 International Conference on Smart Structures and Systems (ICSSS),pages=1–5,year=2019, organization=IEEE

[3] fan2006determinants, title=Determinants of house price: A decision tree approach, author=Fan, Gang-Zhi and Ong, Seow Eng and Koh, Hian Chye,journal=Urban Studies,volume=43,number=12,pages=2301–2315,year=2006, publisher=Sage Publications Sage UK: London, England

[4] thamarai2020house, title=House Price Prediction Modeling Using Machine Learning., author=Thamarai, M and Malarvizhi, SP, journal=International Journal of Information Engineering & Electronic Business,volume=12,number=2,year=2020