Research Article

Prediction of House Price Using XGBoost Regression Algorithm

J.Avanija[a], Gurram Sunitha[b], K.Reddy Madhavi[c] , Padmavathi Kora[d], and R.Hitesh Sai Vittal[e]

[a]
Associate Professor,Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
[b]Professor,Department of CSE, Sree Vidyanikethan Engineering College, Tirupati,,India.
[c]Associate Professor,Department of CSE, Sree Vidyanikethan Engineering College,
Tirupati, India.
[d]Professor,Department of ECE, GRIET, Hyderabad,,India.
[e]Department of CSE, Sree Vidyanikethan Engineering College, Tirupati,India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: House price fluctuates each and every year due to changes in land value and change in infrastructure in and around the area. Centralised system should be available for prediction of house price in correlation with neighbourhood and infrastructure, will help customer to estimate the price of the house. Also, it assists the customer to come to a conclusion where to buy a house and when to purchase the house. Different factors are taken into consideration while predicting the worth of the house like location, neighbourhood and various amenities like garage space etc. Developing a model starts with Pre-processing data to remove all sort of discrepancies and fill null values or remove data outliers and make data ready to be processed. The categorical attribute can be converted into required attributes using one hot encoding methodology. Later the house price is predicted using XGBoost regression technique.

Keywords: Machine Learning; XGBoost Regression; Gradient Boost; Ensemble Learning; House Price Prediction

1. Introduction

Machine learning has so many real-life applications in health industry, business and is slowly venturing into every sector there is. Applications of machine learning include image processing, speech processing, classification of spam mails, medical diagnosis, different movie and product recommendations and is now a days into various domains. Real world implementations of problems are complicated due to so many constraints related to various attributes of data. For example, taken from this paper Predicting a price of house is very complicated due to many constraints in data, so in order to put a price on house for a realtor, needs a computer algorithm which is trained with all constraints both categorical and non categorical attributes.

Machine learning is used for building models and predict data from learnt models. Supervised learning uses data for learning which is directly mapped to labelled output. Unsupervised learning uses data for learning where target data is not labelled but data is predicted instead of values are distributed into clusters or classes.

Regression is a supervised learning algorithm in machine learning which is used for prediction by learning and forming a relationship between present statistical data and target value i.e., Sale Price in this case. Different factors are taken into consideration while predicting the worth of the house like location, neighborhood and various amenities like garage space etc. if learning is applied to above parameters with target values for a certain geographical region as different areas differ in price like land price, material cost. Regression is simple, after drawing conclusions and relationships in presented data between attributes of data and target value, real world data is then fed into algorithm for target value prediction.

Thousands of houses are getting bought everyday by a customer. There are many questions pops up into customer that puts him/her in a dilemma. Some questions such as Am I paying the right price for the house, what is the actual price of the house etc. By taking many different parameters into consideration for house for predicting sale price of the house prediction, is made into accurate and precise model. XGBoost regression model is used in this paper.

Ensemble learning involves the process training and combining individual models termed as the base learners in order to get single prediction. XGBoost is an ensemble learning approach used to train many models and can be used to produce a single best output.

2. Related Work

In Czech Republic there are only few institutions which focuses on long term real estate changes one is Association of Realtors and the Czech Republic (AKR CR) and other is developed by authors. AKR CR takes care of advertisement from different real estate firm and makes sense of real estate development and price fluctuations and is collected manually therefore making scope of data retrieval miniscule. The issue of database for storing real estate attributes is unstructured data which is solved by team of researchers from university of Columbia [1].

A precise price of the house is to be predicted for different reasons for example for bank to give loans to customers who are in need, to calculate mortgage values and to provide insurance to customers from insurance companies. Hedonic price models are generally used for predicting house prices where house is assumed as a commodity which need to be estimated as a combination of different attributes, where each attribute is taken as an individual component. Neural network is an artificial intelligence model which is generally used for replicating brain learning process which has three layers input data layer, hidden layer and output layer where price of house is predicted [2].

Land property isn't just the important need of a man however today it like wise speaks to the wealth and glory of an individual. Interest in land by large all accounts are taken into productive where property estimates don't fluctuate that easily. Land is the least predictable industry in our biological system because it depends on the land where it is present, is it on terrain, lodging costs change all day every day from now and then[3]. There are three factors that majorly impact the price of the house which include physical condition where it is placed based on terrain; idea refers to style like modern, medieval or eclectic and area in sqft or sqmts.

The topic that is not taken into consideration and dug deep is the Street based Local Area (SLA) and estimate how much of this is taken into what extent and how does it associates with house price. The author takes Metropolitan London into consideration for estimating Street based Local Area (SLA) and the effect of it on house pricing using a hedonic price approach which is previous discussed term from previous research paper from Visit Lumsombunchai et al. Street based Local Area (SLA) is defined as a local area that is ; first street – based, second topological/ configurational, third has membership in discrete form and fourth is larger than a home are about smaller than a city [4].

Scikit – learn is a module written in python for introducing and grouping a vast artillery of machine learning algorithms used commonly for supervised and unsupervised learning. Main focus of using this package is making layman understand machine learning terms using a general – purpose high level language. Main intention behind developing this package is to help user design algorithm with ease with interface consistency provided by Scikit learn. It is designed in such a way that no matter the setting i . e., commercial or academic can be used without any difficulty [5].While cost of house always rise, fact known for centuries, though recent rise noticed in recent years is unique to its own and is common. A dramatic increases in prices in homes has been noticed in since the late 1990s in countries like Australia, Canada, China, France have shown much growth among other countries. These factors are taken into consideration for long time prediction of price of house with all the other attributes [6,7].

House price index shows the summarized changes in price value of houses. While for the prediction of single family house price more accurate methods are needed based on the factors such as type of house, size, year built, amenities and other factors that affects the demand and supply of house. The authors examine composite pre processing and feature engineering methodology by considering only .limited features and dataset. A hybrid Gradient boost regression and Lasso model had been proposed to predict the price of single house [8,10].

The proposed approach has recently been deployed as the key kernel for Kaggle Challenge "House Prices: Advanced Regression Techniques". The common method used for house price prediction is the customer contacts an agent of real estate to manage and suggest relevant estate for their investments. The risk incurred is high in this method since the prediction may go wrong leading to loss for customers in their investment. The manual approach used to predict the market value of estates used currently involves heavy risk for customers [9,11].

3. Methodology

Data pre-processing is a process used for refining data before fed into model. Data pre-processing is vaguely divided into four stages called a) Data cleaning b) Data Editing c) Data reduction d) Data wrangling. Data cleaning is process where inaccurate data or if a data field is empty, then value is filled using mean or median or entire record is deleted from data. If data is recorded manually these problems tend to happen. Calculate the mean value considering the value of attributes and number of records in the data. Data editing is process where outliers are picked from data and eradicated. Outliers are mainly recorded in data mainly due to experimental errors produced by machined due to malfunctioning or due to some other parameters. Data reduction is termed as the process of reducing data using some kind of normalisation for easy process of data. Z score is one of processes used for normalisation. Data wrangling is termed as a process where data is transformed or mapped. Data munging, data visualisation and data aggregation comes under this process. Data visualisation is process where statistics are used for producing graphs. Data aggregation is process where data is filtered before fed into model.

During data pre-processing all machine learning algorithms are able to learn from categorical data and converted into numbers using a process called one hot encoding. By using this method categorical values are converted into numbers for both in input and output. Categorical values are converted into binary vector using one hot encoding. Categorical values are converted into integers by mapping the values from binary vector. Then each data value can be represented as a integer in a binary vector where 1 is used to represent data value and all others are represented as zeroes. For Example, take a attribute called colour with values red, red, yellow, green where binary vector is represented as [1,0,0], [1,0,0], [0,1,0], [0,0,1] then fed into model which is easier to understand for most machine learning algorithms. XGBoost regression also uses one hot encoding for understanding categorical values.

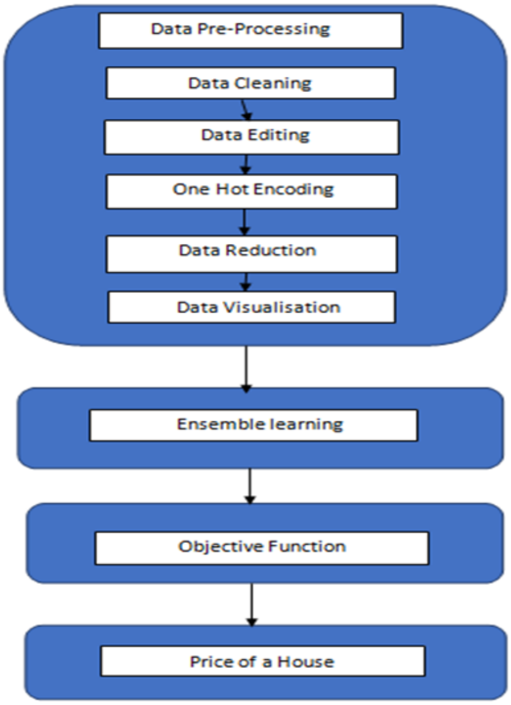

Fig.1.Flow Diagram for House Price Prediction

XGBoost regression is short form for extreme gradient boost regression. It works well compared to others machine learning algorithms.XGBoost is one of the best supervised learning algorithms which can inferred by the way it flows, it consists of objective function and base learners.Loss function is present in objective function which shows the difference between actual values and predicted values whereas regularisation term is used for showing how far is actual value away from predicted value. Ensemble learning used in XGBoost considers many models which are known as base learners for predicting a single value. Not all base learners are expected to have bad prediction so that after summing up all of them bad prediction cancelled out by good prediction. A regressor is the one that fits a model using given features and predicts the unknown output value. Dataset for processing is taken from a Kaggle competition and fed into the model after pre-processing as specified in Figure 1. The algorithm for prediction of price of house using XG Boost Regression is specified below.

3.1 XG Boost Regression Algorithm for House Price Prediction
Input: House attributes dataset.
Output: Price of house.

1. Check input dataset for missing values and calculate d mean is replaced in place of missing value.
2. Divide attributes based on values in data fields as categorical and non-categorical rows.
3. Check Non categorical rows for outliers using outlier detection techniques and remove all outliers.
4. Convert categorical rows into binary vectors using one hot encoding.
5. Divide dataset for cross validation using train test split.
6. Apply Ensemble learning through training and combining individual models termed as base learners in order to derive a single prediction.
   a) Calculate Mean Squared Error (MSE) with true values to predicted values.
   b) Classify independent models as weak-learners and strong-learners using error detection.
   c) Total mean cancels bad prediction with good prediction.
7. Objective function contains the loss function and regularisation term to calculate difference between actual value and predicted value.

4. Results and Discussion

The proposed system is implemented using Anaconda software and the datasets are taken from Kaggle competition on house price prediction. Dataset consists of 80 data attributes and nearly 1500 records. The plot graph for train dataset is specified in Fig.2.

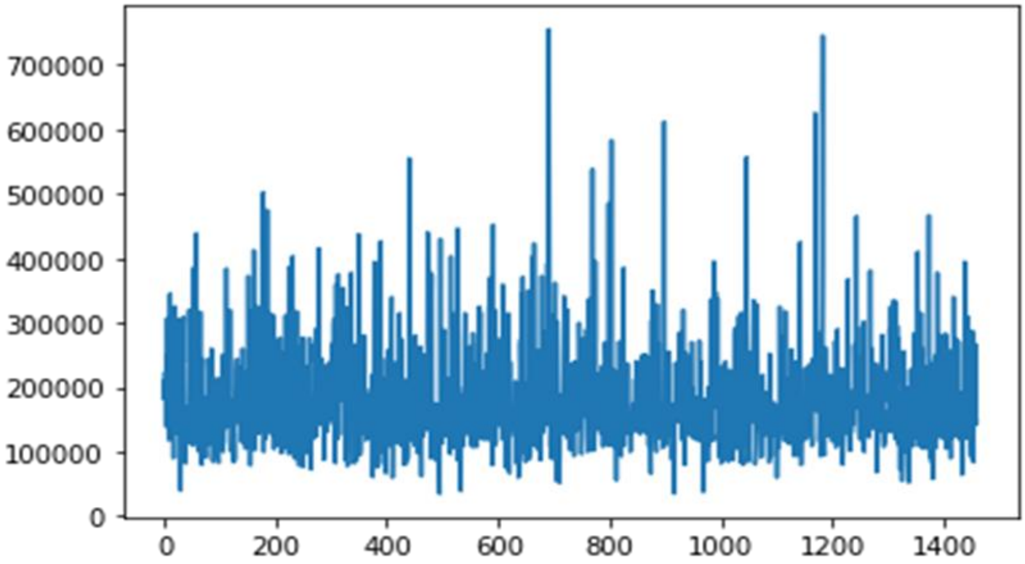

Fig.2. Plot Graph for Train Dataset

Dataset is divided into three subsets for training, validation and testing. Three different models are applied on datasets through different ratios of training, validation and testing data subsets, so that appropriate proportion of three datasets for low test error is calculated as shown in Table.1.

Table.1. Mean Test Errors for 3 Different Models on Different Runs

Training Data (%)	Testing Data (%)	Validation Data (%)	Test Error (%)
60	20	20	5.4
70	15	15	4.6
80	10	10	4.4
90	5	5	4.8

After training each model, the model is fit into the best regularization parameter to avoid overfitting. Table.1. shows the mean test errors for these three different models on different runs. From above table it's evident that for training, test, validation split 80, 10, 10 shows low error rate. The plot graph for test data set is shown in Fig,3.

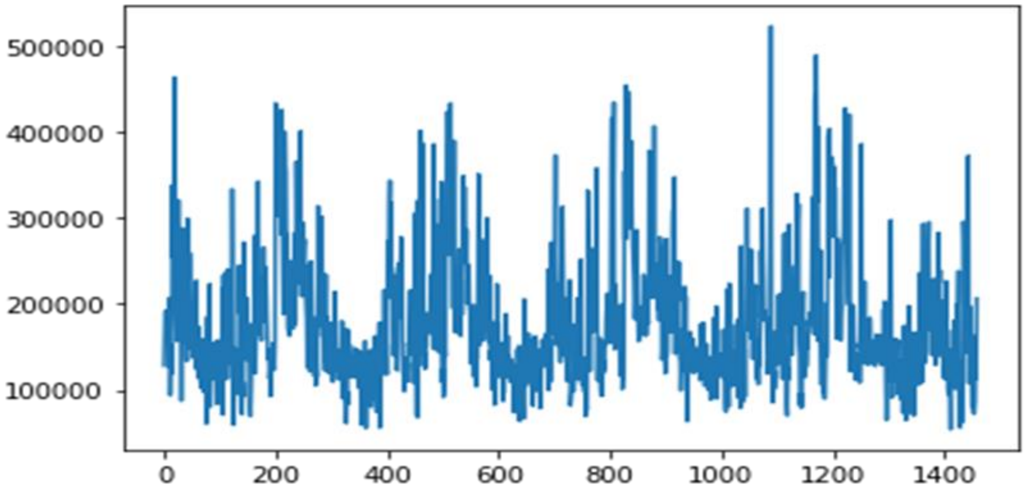

Fig.3. Plot Graph for Test Dataset

5. Conclusion

A model for house price prediction that assists both buyer and seller had been proposed. The proposed house prediction system helps seller to sell house at best price and it also helps buyer to buy house at best price. Price of said land costs, material varies from place to place. Prediction of price of house is difficult as it varies from place to place as all attributes doesn't have same proportion in all places. Deep learning algorithms may enhance the prediction of price of house and decrease test error rate percentage. The XGBoost regression algorithm helps to satisfy the needs of customers by increasing accuracy of the choice of estates and decreasing the risk for customers to invest in real estate. The system can be made widely acceptable by including more number of features. Future scope is to create estate database including more cities in order to help customers explore more number of estates and get accurate decision.

References

1. Shiller R J. Understanding Recent Trends in House Prices and Home Ownership. Proceedings – Economic Policy Symposium - Jackson Hole (2007) ; 89-123.
2. Limsombunchai, Christopher Gan, Minsoo Lee ,House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. American Journal of Applied Sciences (2004) ; 3:193-201 .
3. Eduard, Hromada. Mapping of Real Estate Prices using Data Mining Techniques. Procedia Engineering (2015): 123:233- 240.
4. Stephen Law, Defining Street-based Local Area and Measuring Its Effect on House Price Using a Hedonic Price approach: The Case Study of Metropolitan London. Cities (2007);60 :166–179.
5. Pedregosa, Fabian, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (2011); 12:2825-830.
6. Joep Steegmans, Wolter Hassink. Financial Position and House Price Determination: An Empirical Study of Income and Wealth Effects. Journal of Housing Economics (2017); 36:8-24.
7. Nihar Baghat,Ankit Mohokar, Shreyash Mane. House Price Forecasting Using Data Mining, International Journal of Computer Applications (2016); 152:23-26.
8. Torgo,Luis, Joao Gama. Logic Regression Using Classification Algorithms. Intelligent Data Analysis (1997); 4: 275-292.
9. Bork M, Moller, V.S. House Price Forecast Ability: A Factor Analysis. Real Estate Economics. Heidelberg (2016) ; 46:582-611.
10. Quang Troung, Minh Nguyen, Hy Dang, Bo Mei. House Price Prediction via Improved Machine Learning Techniques. Precedia Engineering (2020); 174:433-442.
11. Atharva Chogle ,Priyankakhaire , Akshata Gaud , Jinal Jain. House Price Forecasting using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering (2017); 6: 24-28.