House Price Forecasting using Machine Learning

1st Shriya Reshi
Department of Computer Engineering
Maharashtra Institute of Technology
Pune, India

2nd Kunal Lawangare
Department of Computer Engineering
Maharashtra Institute of Technology
Pune, India

3rd Shuhul Nehru
Department of Computer Engineering
Maharashtra Institute of Technology
Pune, India

4th Rajeshwar Singh
Department of Computer Engineering
Maharashtra Institute of Technology
Pune, India

5th Prof.Mrs.Archana Khandekar
Department of Computer Engineering
Maharashtra Institute of Technology
Pune, India

***Abstract*—Buying a house is an emotional desire as well as a popular investment option. The prices of houses, properties etc keep on increasing every year making it difficult for the people to keep the track of the prices. House prices differ from place to place and even from locality to locality, sometimes even in the same building their might be apartments with different prices. It becomes very difficult for the people to know the prices of house in each and every locality, therefore they usually hire a broker who has the information of the houses in his area. But a major disadvantage is that the broker may charge huge amount of fees for this. Therefore we need a system that can predict the prices of the house. This system will help both the buyer and the seller as it can allow the seller to put on the house for sell and the most important for the buyer to be able to search houses in the locality he needs and the house type as per his need. With this system the buyer can buy the house at a reasonable price. This research aims to predict the house prices in the city. Various types of regressions can be used for the prediction of the house prices, but the ones with the best accuracy and minimum error are the Lasso Regression and the Gradient Boosting Regression.**

***Index Terms*—House prediction, Regression analysis, Lasso Regression, Gradient Boosting Regression**

I. Introduction

Investment is an activity in which many people are interested as it gives a good returns most of the times. People often invest their money in various objects such as gold, stocks etc but the most common and well know place for the investors to invest their money is in the property, especially in the apartments and the plots. This is one of the most common investment asset for the people as it guarantees profitable returns. People know the house prices increase every year but their is no guarantee over the stock market or if they invest their money in the gold, therefore the investment in the property is one of the most common practises. In order to get good returns people must choose their property very carefully. With a well chosen asset people can enjoy excellent returns. As the population world wide increases every year it is obvious to assume that the number of houses will keep on increasing as long as the population increases. Buying a house at a reasonable price is a concern for the buyer and to sell it at a profitable price is a concern for the seller.

While looking for the houses the buyer should be very careful while searching, as sometimes their are people with fraud schemes, the buyer may end up loosing the money. House prices differ from one locality to other locality. It is very difficult for an ordinary person to know the house price of every locality in the city. The reason why house prices differ so much is because, house prices depend on various factors such as the living area, number of storeys, locality, the number of bedrooms in the apartment, the amenities provided by the society, whether the house is near to a public transport such as railway/metro station, bus stops, airports etc, the house prices also depend on whether the house is near a school or hospitals etc. We can build a system for the buyers to search for the houses with respect to locality of their choice or with respect to the type of an apartment they are looking for with number of rooms in the apartment. Their is a lot of demand for these kind of models as it can detect the prices of the houses at an early stage and thus it can help the buyers or investors for buying the house. With the help of these systems their can be less number of frauds and also the brokers who cut off lot of fees from both buyers and the sellers can be eliminated and hence saving money to both.

Lasso Regression in machine learning is used to perform predictive analysis for enhancing accuracy and interpretability of the model. We have used the Lasso Regression analysis in our model as it is vast enough to improve those inclination of the model on over-fit. Least ten variables can foundation over fitting and also is enough to cause computational tests. This circumstance could emerge in the event from claiming a large number or billions about characteristics.

Gradient boosting 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. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

A. Problem Statement

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, future prices will be predicted.

The rest of the paper work is organised as follows: section II talks about the related work; section III talks of our approach; section IV shows the experiment results.

II. Related Work

A. Real Estate Grading

Real Estate Grading is a technique by which the investors can see how much risk is involved in the house for investment. House grading gives a view of the true value of real estates and each grade shows a certain level of investment risk.

Grade scale usually consists of 5 levels:-

Grade 0- Weak recommendation with high risks.

Grade 1- Weak recommendation with low risks.

Grade 2- Moderate recommendation.

Grade 3- Strong recommendation with low rewards.

Grade 5- Strong recommendation with high rewards.

This method helps the investors to look into the risks involved in buying the property. It gives the investor a brief information on how much risk is involved in buying the house. This way the investor can avoid buying houses in which their is high risk involvement factor.

This technique may alert the investor by giving the risk factor but it cannot predict the prices of the house that will change in the future.

B. Hedonic Pricing

Hedonic Analysis is one of the most common used method to predict the house prices. Hedonic Analysis is also used for estimating the risk involved in buying the property. All the housing attributes such as number of bedrooms, living area, amenities, number of storeys etc are combined into a multiple regression with the sale price as the dependable variable. Hedonic Analysis can be used to predict the future sale price of the house i.e what will the price of the house after few years be profitable or will the buyer or investor suffer a loss. Hedonic pricing can give a huge data that can be measured within a long time frame. But it is quite uncertain to predict the prices efficiently. One of the problems with assessing and predicting future sales prices using traditional hedonic models, is the chosen functional relation between spatial factors and sales prices.

It is very common that the location plays an very important role in the house prices. There are various ways by which we can include the factors such as neighbourhood characteristics. Though Hedonic Analysis work efficiently on large datasets, there is a need to improve this approach for the smaller datasets.

C. Artificial Neural Networks

It is a well know fact that the Artificial Neural Networks is used for various prediction. Artificial Neural Networks is also used to predict property prices. While housing attributes in hedonic models are typically fitted with linear, log or squared relationships with price, artificial neural networks are used to fit more complex functional relationships.

The performance between Artificial Neural Networks and the multiple regression models can be compared, it is shown that the multiple regression models perform better than artificial neural network models at small sample sizes. Also, Artificial Neural Networks are more complicated than the regression models. Regression models are also more widely used than the Artificial Neural Networks. Hence, we do not consider the Neural Networks any further.

III. Proposed Method

The price prediction system contains following steps:

A. Data Collection

B. SQL Connection

C. Data Preprocessing

D. Regression

E. Prediction Analysis

Let us discuss these steps one by one.

A. Data Collection

Data Collection is first step in almost every predictive model. Data Collection is a process of collecting information on the variables of interest. The data collected is of related to the predictive model which is to be made. Data can be in the form of a database or a datasets consisting of various number of rows and columns. Hypothesis are tested on this data and the results are evaluated. The data collected should be accurate, the data should contain minimum number of redundant values and garbage data. One should avoid collecting data with NULL values in it. This data redundancy, NULL values etc can be eliminated in the Data Pre-processing, but it should be kept at minimum. This will give us wrong result, which will mislead the researchers to pursue fruitless avenues of investigation.

There are many ways to collect data, one can get it from various websites others can make their own datasets, though it is only possible if the dataset is small. We have collected the data from the Kaggle website. The data that we have collected contains the attributes of the factors affecting the house prices, which are:- Living Area, Neighborhood, Condition of the house, Building Type, Number of bedrooms, the year in which house was built, various amenities provided by the society etc. These are just few of many attributes from our dataset. The data collected is divided into parts for training and testing. The training dataset contains 80 columns and 1460 rows whereas the testing dataset contains 79 rows and 1459 column.

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis services randomly samples the data

to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

B. SQL Connection

SQL connection is required to import the data tables from SQL database into our application. SqlConnection object represents a unique session to a SQL Server data source. With a client/server database system, it is equivalent to a network connection to the server.

The train file is used for training the model whereas the test file is used for the prediction. We made two tables in SQL, First table contains test file attributes whereas the Second table contains house ID and the Predicted price. These tables are needed to be shown in our GUI and also used for filtering the data. This is done by connecting our SQL to the python platform. SQL connection is done via ODBC driver. ODBC is Open DataBase Connectivity that allows applications to access data in database management systems (DBMS) using SQL as a standard for accessing the data.

C. Data Preprocessing

The data present in our collected data may have redundancy, missing values, NULL values, garbage values etc. In other words the data is not clean and if that data is used for our predictive model then it will give wrong results and less accuracy. To avoid this the data is cleaned. This cleaning of raw data to remove redundancy, ambiguity etc is known as data pre-processing. Data Preprocessing allows the data to transform into useful and efficient format which can be used for the analysis purpose.

The data which we collected had some missing values, and unnecessary attributes. We have dropped the id because it is an unnecessary attribute which will not affect our prediction. After dropping the id we have calculated the skewness and kurtosis before and after the log transformation. We then have calculated the correlationship between attributes. We then have plotted the scatter plot and looked for the outliers. Outliers are the values which are not in the cluster and hence they need to be deleted. We have replaced the missing values in the dataset. We then have transformed the numerical features into categorical features.

These are few ways by which we have performed the data pre-processing technique to remove the redundant data and to remove ambiguity.

D. Regression Analysis

Regression Analysis is a method by which we can estimate the relationship between the dependent variables and the independent variables. Regression Analysis provides us detailed insight which can be applied to improve the product. Regression Analysis is very important as it helps determine which factors matter most, which it can ignore, and how those factors interact with each other. The importance of regression analysis lies in the fact that it provides a powerful statistical method that allows a business to examine the relationship between two or more variables of interest.

We have used Lasso Regression and Gradient Boosting Algorithm. Linear regression is a type of regression technique which gives us the linear relationship between the dependent variables and the independent variables. Lasso regression uses the shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. Gradient Boosting algorithm is the technique which produces a prediction model. Gradient Boosting Algorithm technique provides us with good predictive accuracy.

E. Predictive Analysis

Predictive Analysis is technique which is used to make prediction about the unknown future events. It uses many techniques from data mining, statistics, modelling, machine learning, and artificial intelligence to analyze current data to make predictions about future. Predictive analysis is used in various sectors such as retail, health, sports etc.

We have used predictive analysis in forecasting the house prices as it analyses the current events affecting on the house price, i.e location, amenities provided, also whether in future will it be near to a public transport, hospital, school etc, and analysing the given current events it is able predict the price of the house in the future.

IV. Experiment Results

In this section, we will be discussing about the results we get from our machine learning model. Earlier we had divided our dataset into training data and testing data. Here, we run our model on the testing data to find the accuracy of the model.

TABLE I
EXPERIMENT RESULTS

Sr No.	Method	Accuracy(%)
1	Lasso Regression	73
2	Gradient Boosting	75

We can observe that the Gradient Boosting gives us the better accuracy result than the Lasso Regression.

V. Conclusion

The aim of this project was to predict house prices using machine learning. Out of the two models that we used, Gradient Boosting gave us the better accuracy compared to the Lasso Regression in the prediction of the house prices.

Thus, we can conclude that the Gradient Boosting Regression is the best we used to build our model as not only it gave us the better accuracy but also because of its adaboost approach which enables a model to build on Weak Learners and neglect the less desirable result by assigning higher weight to it. Gradient Boosting also reduces the higher complexity of our model.

REFERENCES

[1] Yanjie Fu, Hui Xiong, 2015 IEEE 15th International Conference on Data Mining Workshops A Discovery System for Finding High-Value Homes.

[2] Neelam Shinde and Prof.Kiran Gawande, 2018 International Conference on Automation and Computational Engineering (ICACE 2018) Survey on predicting property price.

[3] Jiaying Kou , Xiaoming Fu , Jiahua Du , Hua Wang , Geordie Z. Zhang, 2018 IEEE Understanding Housing Market Behaviour from a Microscopic Perspective.

[4] 2017 IEEE Research of second-hand real estate price forecasting based on data mining.

[5] Swati Singh and Gaurav Dubey International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-X, Issue-X Finding interest of people in purchasing real estate by using data mining techniques.

[6] https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

[7] https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

[8] https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques