LITERATURE REVIEW

Real Price Predictor
Predict The Market Value of Properties

Submitted by

ALDRIN SIMENTHY

GREESHMA PRASAD P

JYOTHIS K S

VISMAYA DINESH

On

20 - MARCH - 2021

Contents

1 Introduction

Every single organization in today's real estate business is operating fruitfully to achieve a competitive edge over alternative competitors. There is a need to simplify the process for a normal human being while providing the best results. This paper proposes a system that predicts house prices using a regression machine learning algorithm. In case you're going to sell a house, you have to recognize what sticker price to put on it. This regression model is built not only for predicting the price of the house which is ready for sale but also for houses that are under construction.

Decision Tree is a tool, which can be employed for Classification and Prediction. It has a tree shape structure, where each and every internal node represents test on an attribute and the branches out of the node denotes the test outcomes. 80% of the known dataset can be used as training set and 20% can be 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. Using the training set Decision Tree Classifier model is constructed and tested with test data to identify the accuracy level of the classifier.

Decision Tree formation as shown in fig. 3 employs divide and conquer strategy for splitting the training data into subsets by testing an attribute value. This involves attribute selection measures; the attribute which is to be tested first is the one which is having high information gain. Same splitting process is recursively performed on the subsets derived. The splitting process of a subset ends when all the tuples belong to the same attribute value or when no remaining attributes or instances are left with. Decision Tree formation does not need any basic domain knowledge. It can handle data of high dimensions as well. Decision Tree Classifiers have good accuracy in classification.

This paper discusses about an application of Decision Tree, for purchasing a house in a city based on attribute values such as no. of bedrooms, no. of bathrooms, carpet area, built-up area, the floor, age of the property, zip code, latitude and longitude of the property and availability of schools, transport facilities, shopping facilities, medical facilities, Other than those of the mentioned features, which are generally required for predicting the house prices, we have included one more feature is crime rate. These features provide a valuable contribution towards predicting property prices since the higher values of these features will lead to a reduction in house prices.

2 Literature Review

Trends in housing prices indicate the current economic situation and also are a concern to the buyers and sellers. There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms. House price depends upon its location as well. A house with great accessibility to highways, schools, malls, employment opportunities, would have a greater price as compared to a house with no such accessibility. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh [1] had proposed an advanced house prediction system using linear regression. This system's aim was to make a model that can give us a good house price prediction based on other variables. They used the Linear Regression for Ames dataset and hence it gave good accuracy.

ID3 algorithm splits attribute based on their entropy. TDIDT algorithm is one which constructs a set of classification rules through the intermediate representation of a decision tree [2,3]. Weka interface [4] is used for testing of data sets by means of a variety of open source machine learning algorithms.

Timothy C. Au [5] addressed about the absent level problems in Random Forests, Decision Trees, and Categorical Predictors. Using three real data sets, the authors have illustrated how the absent levels affect the performance of the predictors.

Fan et al [6] has utilized decision tree approach for finding the resale prices of houses based on their significant characteristics. In this paper, hedonic based regression method is employed for identifying the relationship between the prices of the houses and their significant characteristics. Ong et al. [7] and Berry et al. [8] have also used hedonic based regression for house prediction based on significant characteristics.

Shinde and Gawande [9], predicted the sale price of the houses using various machine learning algorithms like, lasso, SVR, Logistic regression and decision tree and compared the accuracy. Alfiyatin et al. [10] has modeled a system for house price prediction using Regression and Particle Swarm Optimization (PSO). In this paper, it has been proved that the house price prediction accuracy is improved by combining PSO with regression.

Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) [11] studied linear regression algorithms for prediction of the houses. The goal of the paper is to predict the efficient price of real estate for customers with respect to their budgets and priorities. Analysis of past market trends and price ranges will predict future house pricing.

3 Existing System

In the present situation, the customer visits a real estate agent so that he/she can suggest suitable show places for his investments. But the above method is risky as the agent may forecast wrong prices to the customer and that will lead to loss of customer's investment. This manual technique which is currently used in the market is outdated and has a high risk. So as to overcome the drawback, there is a need for an updated and automated system.

4 Proposed System

The world is shifting from manual to automated systems. The objective of our project is to reduce the problems faced by the customer. Proposed work aims at predicting the availability of houses based on different features of the houses and also the facilities available nearby the location of the houses. Work also includes the price prediction of the houses based on the features of the house and facilities nearby its location.

This work includes two parts namely:

i Decision Tree Classifier is used to predict the availability of houses as per the users' requirement constraints and it produces responses like yes or no respectively to tell whether a house is available or not.

ii Decision tree regression and Multiple Linear Regression methods are used to predict the prices of the houses.

A real time dataset is prepared by analysing the location named Ernakulam District in kerala of India. The dataset contains the following features of the houses such as Number of bedrooms, age of the house, transport facility, schools available in the nearby location and shopping facilities.

The proposed method helps to search houses in big cities based on the following attributes.

1. Number of bedrooms (1BHK, 2BHK and 3BHK etc..).
2. Number of bathrooms, carpet area, built-up area, the floor etc..
3. Age of the property.
4. Education facilities.
5. Medical facilities.
6. Transport facility such as availability of bus, trains and flights.
7. Shopping facility such as small markets, general stores, shopping malls.
8. Analyse the crime rate.

The proposed work is implemented using Scikit Learn, a machine learning tool.

5 System Design and Architecture

Phase 1: Collection of data
Data processing techniques and processes are numerous. We collected data for Kochi's real estate properties from various real estate websites. The data would be having attributes such as Location, carpet area, built-up area, age of the property, zip code, etc. We must collect the quantitative data which is structured and categorized. Data collection is needed before any kind of machine learning research is carried out. Dataset validity is a must otherwise there is no point in analysing the data.

Phase 2: Data preprocessing
Data preprocessing is the process of cleaning our data set. There might be missing values or outliers in the dataset. These can be handled by data cleaning. If there are many missing values in a variable we will drop those values or substitute it with the average value.

Phase 3: Training the model
Since the data is broken down into two modules: a Training set and Test set, we must initially train the model. The training set includes the target variable. The decision tree regressor algorithm is applied to the training data set. The Decision tree builds a regression model in the form of a tree structure.

Phase 4: Testing and Integrating with UI
The trained model is applied to test dataset and house prices are predicted. The trained model is then integrated with the front end using Kivy in python.

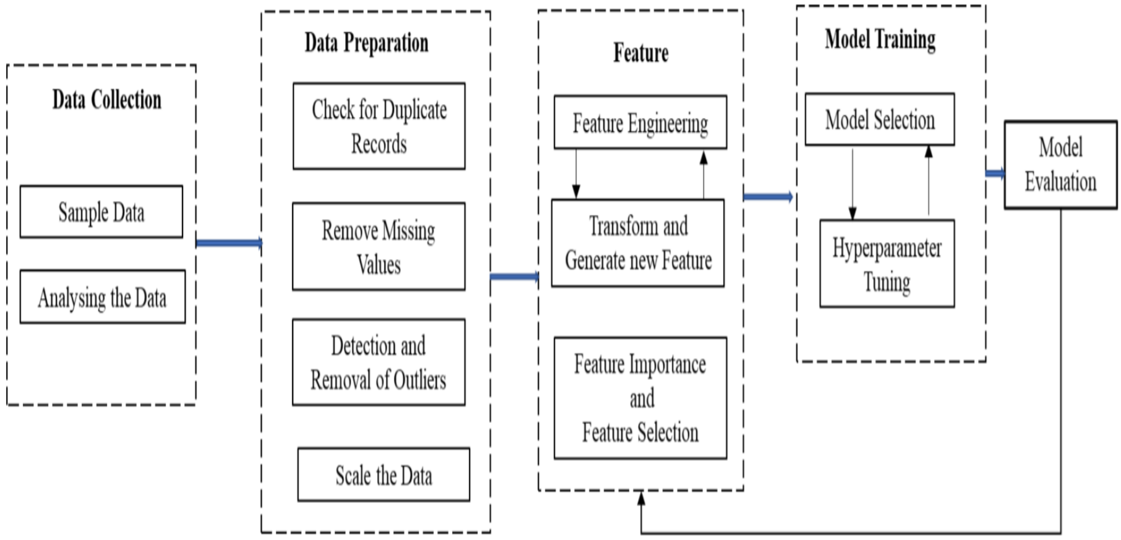

Figure 1: Research framework for the housing price problem.

6 Methodology

6.1 Flow of Development

The below figure describe about the methodology used in the real estate house price predictions and the generic flow of development is given.

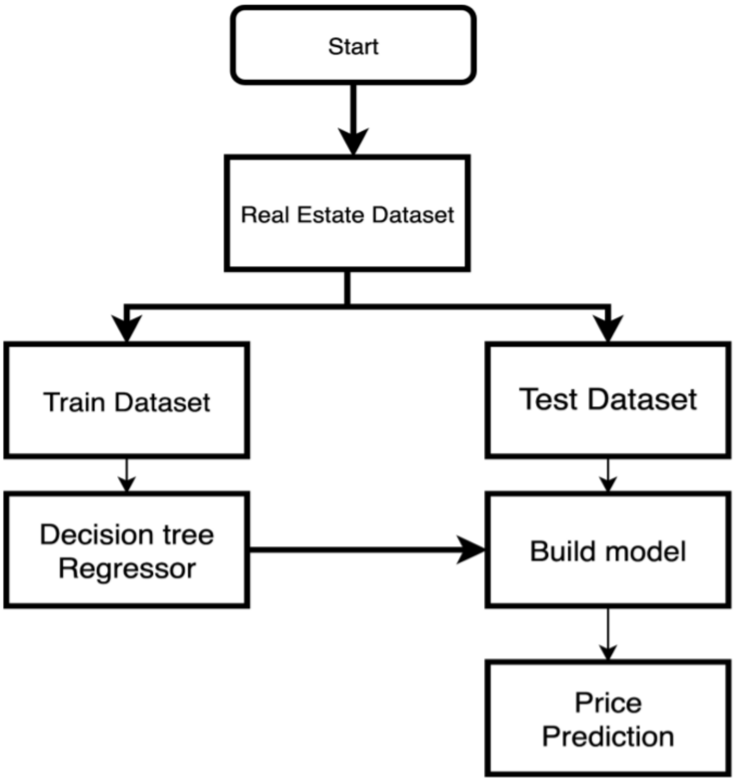

Figure 2: The generic flow of development.

6.2 Studied Algorithms

In the process of developing this model, various regression algorithms were studied. SVM, Random Forest, Linear regression, Multiple linear regression, Decision Tree Regressor, KNN, all were tested upon the training dataset. However, the decision tree regressor provided the highest accuracy in terms of predicting the house prices. The decision to choose the algorithm highly depends upon the dimensions and type of data in the data used. The decision tree algorithm suited best for our dataset.

6.3 Tools

6.3.1 Decision Tree Regressor

The decision tree regressor observes features of an attribute and trains a model in the form of a tree to predict data in the future to produce meaningful output. Decision tree regressor learns from the max depth, min depth of a graph and according to system analyzes the data.

Grid Search CV is a way to deal with parameter tuning that will efficiently manufacture and assess a model for every mix of calculation parameters indicated in a grid. Grid Search CV in this algorithm is used to assess the best value for max-depth, using which the decision tree is constructed.

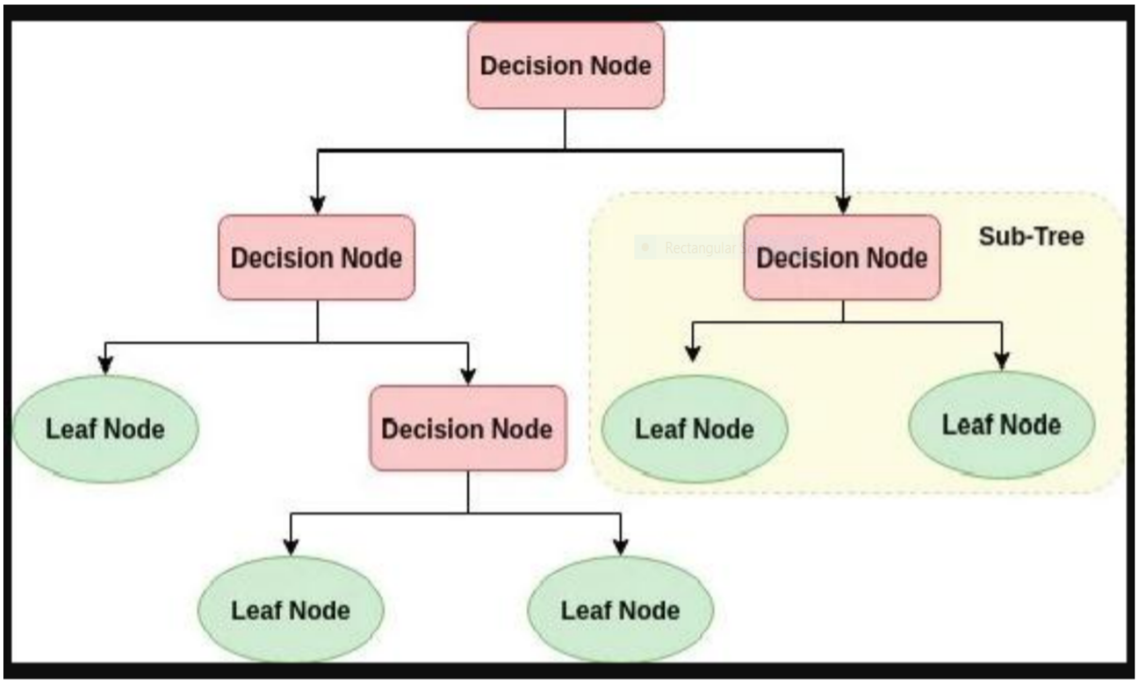

Figure 3: Decision Tree Structure.

6.3.2 Scikit Learn

The Scikit-Learn (SK Learn) is a Python Scientific toolbox for machine learning and is based on SciPy, which is a well-established Python ecosystem for science, engineering and mathematics. Scikit-learn provides an ironic environment with state of the art implementations of many wellknown machine learning algorithms, while sustaining an easy to use interface tightly integrated with the Python language [12],[13]. Scikit-learn features various functionalities like Clustering algorithms, Regression, Classification including random forests, gradient boosting, support vector machines, k-means and DBSCAN, and it has been designed to interoperate in conjunction with the Python scientific and numerical libraries SciPy and NumPy.

The step by step implementation using Scikit Learn is as follows.

Step 1: Import the required libraries.
Step 2: Load the dataset.
Step 3: Assign the values of columns 1 to 6 in the Dataset to “X”.
Step 4: Assign the values of column 7 which is the class label to “Y”.
Step 5: Fit decision tree classifier to the dataset.
Step 6: Predict the class label for the test data.

6.3.3 Kivy

After building the model and successfully giving the result, the next step is to do the integration with the UI, for this purpose Kivy is used.

Kivy is a free and open source Python framework for developing mobile apps and other multitouch application software with a natural user interface. This means Kivy provides you with tools, libraries, and technologies that allow you to build an android application. Kivy is easy to put away routes together and this framework is mainly used for integrating python models.

7 Conclusion

In this paper, the Decision tree machine learning algorithm is used to construct a prediction model to predict potential selling prices for any real estate property. we will add Additional features like crime rate were included in the dataset to help predict the prices even better. These features are not mostly included in the datasets of other prediction systems, which makes this system is different. These features influence people’s decision while purchasing a property, so why not include it in predicting house prices. The trained model is integrated with the User Interface using the Kivy Framework. we expecting the system gives 90% and more accuracy while predicting the prices for the real estate prices.

8 References

[1] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, “A Hybrid Regression Technique for House Price Prediction”, December 2017.

[2] J.R. Quinlan, “C4.5: programs for Machine Learning”, Morgan Kaufmann, New York, 1993.

[3] J.R. Quinlan, “Induction of Decision Trees”, Machine Learning 1, 1986, pp.81-106.

[4] SamDrazin and Matt Montag", Decision Tree Analysis using Weka", Machine Learning-Project II, University of Miami.

[5] Timothy C. Au, "Random Forests, Decision Trees, and Categorical Predictors: The Absent Levels Problem", Journal of Machine Learning Research 19, 2018, pp. no.1- 30.

[6] Gang-Zhi Fan, Seow Eng Ong and Hian Chye Koh, "Determinants of House Price: A Decision Tree Approach", Urban Studies, Vol. 43, No. 12, November 2006, PP.NO.2301- 2315. [7] Ong, S. E., Ho, K. H. D. and Lim, C. H., "A constantquality price index for resale public housing flats in Singapore", Urban Studies, 40(13), 2003, pp. 2705 –2729.

[8] Berry, J., McGreal, S., Stevenson, S., "Estimation of apartment submarkets in Dublin, Ireland", Journal of Real Estate Research, 25(2), 2003, pp. 159–170.

[9] Neelam Shinde, Kiran Gawande, "Valuation of house prices using Predictive Techniques", International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835, Volume-5, Issue-6, Jun.-2018 pp. 34 to 40.

[10] Adyan Nur Alfiyatin , Hilman Taufiq, Ruth Ema Febrita and Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, pp. 323 to 326.

[11] Nihar Bhagat, Ankit Mohokar, Shreyash Mane "House Price Forecasting using Data Mining" International Journal of Computer Applications,2016.

[12] `http://scikit-learn.org/stable/index.html`

[13] `http://scikit-learn.org/stable/auto_examples/index.html`

[14] `https://kivy.org/doc/stable/guide/android.html`