HOUSE PRICE PREDICTION MODELING USING MACHINE LEARNING

P.DEVENDAR[1], G.SUPRAJA[2], D.HARI PRIYA[3], E.RAVALIKA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF INFORMATION TECHNOLOGY MALLA REDDY ENGINEERING COLLEGE FOR WOMEN (UGC-AUTONOMOUS) MAISAMMAGUDA, HYDERABAD-500100

ABSTRACT: Machine Learning is seeing its growth more rapidly in this decade. Many applications and algorithms evolve in Machine Learning day to day. One such application found in journals is house price prediction. House prices are increasing every year which has necessitated the modeling of house price prediction. These models constructed, help the customers to purchase a house suitable for their need. Proposed work makes use of the attributes or features of the houses such as number of bedrooms available in the house, age of the house, travelling facility from the location, school facility available nearby the houses and Shopping malls available nearby the house location. House availability based on desired features of the house and house price prediction are modeled in the proposed work and the model is constructed for a small town in West Godavari district of Andhrapradesh. The work involves decision tree classification, decision tree regression and multiple linear regression and is implemented using Scikit-Learn Machine Learning Tool.

INDEXTERMS—Decision tree, house price prediction, decision tree regression, multiple linear regression.

I. INTRODUCTION Data Mining is extracting knowledge or useful pattern from large databases[12]. Classification is one of the data mining functionalities, employed for finding a model for class attribute which is a function of other attribute values [13]. Decision Tree is a tool, which can be employed for Classification and Prediction[6][7]. 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[14]. Decision Tree formation as shown in fig. 1 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 [2]. 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. Once the Decision Tree is formed, new instances can be classified easily by tracing the tree from root to leaf node. Classification through Decision Tree does not require much computation. Decision Trees are capable of handling both continuous and Categorical type of attributes. To avoid generation of meaningless and unwanted rules in Decision Trees, tree should not be deeper which results in over fitting. Such a tree with over fitting works more accurate with training data and less accurate with test data. Pre pruning and Post pruning are the techniques used in Decision Tree to reduce the size of the trees and avoid over fitting. In Post Pruning the Decision Tree branches and hence the level (depth) of the tree are reduced after completely building the tree. In Pre Pruning, care is taken to avoid over fitting while building the tree itself. Decision Trees find its major applications in areas such as medicine, weather, finance, entertainment, sports, etc. Decision Tress can also be used for prediction[6], data manipulation and handling of missing values[7]. As an example in digital mammography it is used for classifying tumor cells and normal cells [3]. This paper discusses about an application of Decision Tree, for purchasing a house in a city based on attribute values such as transport facilities, number of bed rooms, and availability of schools, shopping facilities andmedical facilities

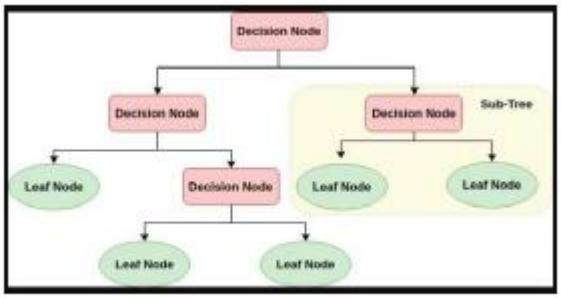

Fig. 1. Decision Tree Structure

II. RELATED WORK

Patel and Upadhyay [4] have discussed various pruning methods and their features and hence pruning effectiveness is evaluated. They have also measured the accuracy for glass and diabetes dataset, employing WEKA tool, considering various pruning factors. 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 [5,6]. Weka interface [7] is used for testing of data sets by means of a variety of open source machine learning algorithms. Fan et al [8] 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. [9] and Berry et al. [10] have also used hedonic based regression for house prediction based on significant characteristics. Shinde and Gawande [11], 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. [12] 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. Timothy C. Au [13] 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[14][15].

EXISTING SYSTEM:

In The Existing system used xgboost for house price prediction. This study aims to explore the important explanatory features and determine an accurate mechanism to implement spatial prediction of housing prices in Beijing..,, based on the housing price and features data in Beijing, China. Our result shows that compared to traditional hedonic methods, machine learning methods demonstrate significant improvements on the accuracy of estimation despite that they are more time-costly. Moreover, it is found that XGBoost is the Less accurate model in explaining and predicting the spatial dynamics of housing prices in Beijing.

DISADVANTAGES OF EXISTING SYSTEM:

- INXgboost, you have to manually create dummy variable/ label encoding for categorical features before feeding them into the models. Catboost/Lightgbm can do it on their own, you just need to define categorical features names or indexes.

- Training time is pretty high for larger datasets.
- Moreover, it is found that XGBoost is the Less accurate model in explaining and predicting the spatial dynamics of housing prices in Beijing.

Algorithm: XGBOOST.

PROPOSED SYSTEM:

The proposed method is based on the linear regression. This project is proposed to predict house prices and to get better and accurate results.The data for the house prediction is collected from the publicly available sources. In validation, training is performed on 50% of the dataset and the rest 50% is used for testing purposes.

This technique splits the dataset into a number of subsets. At that point, it has been attempted for preparing on the entirety of the subsets; however, leave one (k-1) subset for the assessment of the prepared model. This strategy emphasizes k times with an alternate subset turned around for the preparation reason each time.

ADVANTAGES OF PROPOSED SYSTEM:

- The error free prediction provides better planning in the prediction of house price and other industries.
- This would be of great help for the people.
- This would be of great help to the people because the house pricing is a topic that concerns a lot of citizens whether rich or middle class as one can never judge or estimate the pricing of a house on the basis of locality or facilities available.
- Linear Regression is simple to implement and easier to interpret the output coefficients
- The ability to determine the relative influence of one or more predictor variables to the criterion value

Algorithm: Linear Regression (LR)

IMPLEMENTATION:

MODULES:

- User.
- Admin
- Machine learning

MODULES DESCRIPTION:

User:

The User can register the first. While registering he required a valid User email and mobile for further communications. Once the User registers, then the admin can activate the User. Once the admin activates the User then the User can login into our system. After login User will add the data to predict house values.

Admin:

Admin can login with his credentials. Once he logs in he can activate the users. The activated users only login in our applications. The admin will store csv data into our database. we can implement logistic algorithm to predict house and also we can perform cross validation

Machine learning:

Machine learning refers to the computer's acquisition of a kind of ability to make predictive judgments and make the best decisions by analyzing and learning a large number of existing data. The representation algorithms include deep learning, artificial neural networks, decision trees, enhancement algorithms and so on. The key way for computers to acquire artificial intelligence is machine learning. Nowadays, machine learning plays an important role in various fields of artificial intelligence. Whether in aspects of internet search, biometric identification, auto driving, Mars robot, or in American presidential election, military decision assistants and so on, basically, as long as there is a need for data analysis, machine learning can be used to play a role**.**

CONCLUSION AND FUTURE SCOPE This article uses the most fundamental machine learning algorithms like decision tree classifier, decision tree regression and multiple linear regression. Work is implemented using Scikit-Learn machine learning tool. This work helps

the users to predict the availability of houses in the city and also to predict the prices of the houses. Two algorithms like decision tree regression and multiple linear regression were used in predicting the prices of the houses. Comparatively the performance of multiple linear regression is found to be better than the decision tree regression in predicting the house prices. In future the dataset can be prepared with more features and advanced machine learning techniques can be for constructing the house price prediction model.

REFERENCES

[1] Jiawei Han, MichelineKamber, "Data Mining Concepts and Techniques", pp. 279-328, 2001.

[2] J. R.Quinlan," Simplifying decision trees", Int. J. HumanComputer Studies.

[3] Maria-Luiza Antonie, et. al., "Application of Data Mining Techniques for Medical Image Classification", Proceedings of the Second International Workshop on multimedia Data Mining(MDM/KDD'2001) in conjunction with ACM SIGKDD conference. San Francisco,USA, August 26,2001.

[4] Nikita Patel and Saurabh Upadhyay, "Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA", International Journal of Computer Applications, Volume 60– No.12, December 2012, pp 20-25.

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

[6]Pravin Kshirsagar, Nagaraj Balakrishnan & Arpit Deepak Yadav "Modelling of optimised neural network for classification and prediction of benchmark datasets" , Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 8:4, 426-435, DOI: 10.1080/21681163.2019.1711457,2020

[7]Dr. Sudhir Akojwar, Pravin Kshirsagar, Vijetalaxmi Pai "Feature Extraction of EEG Signals using Wavelet and Principal Component analysis", National Conference on Research Trends In Electronics, Computer Science & Information Technology and Doctoral Research Meet, Feb 21st & 22nd ,2014.

[8] 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.

[9] 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.

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

[11] 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

[12]H. Manoharan, et al., “Examining the effect of aquaculture using sensor-based technology with machine learning algorithm”, Aquaculture Research, 51 (11) (2020), pp. 4748-4758.

[13]S. Sundaramurthy, S. C and P. Kshirsagar, "Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 2020, pp. 17-21, doi: 10.1109/DASA51403.2020.9317253.

[14]P. R. Kshirsagar, H. Manoharan, F. Al-Turjman and K. Kumar, "DESIGN AND TESTING OF AUTOMATED SMOKE MONITORING SENSORS IN VEHICLES," in *IEEE Sensors Journal*, doi: 10.1109/JSEN.2020.3044604.

[15] D. Jose, P. N. Kumar and A. A. Hussain, "Fault tolerant adaptive neuro-fuzzy based automated cruise controller on FPGA," *2013 Annual IEEE India Conference (INDICON)*, 2013, pp. 1-5, doi: 10.1109/INDCON.2013.6725977.

