International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 4th Aug 2021. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=ISSUE-08

DOI: 10.48047/IJIEMR/V10/I08/07

Title PREDICTION OF HOUSE PRICE USING MACHINE LEARNING WITH PYTHON

Volume 10, Issue 08, Pages: 41-45

Paper Authors
Mr V. RAHAMATULLA, Mr B. RAVI TEJA

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

PREDICTION OF HOUSE PRICE USING MACHINE LEARNING WITH PYTHON

Mr V. RAHAMATULLA, Assistant Professor, Dept of MCA, SVIM - Sree Vidyanikethan Institute of Management, Tirupati.

Mr B. RAVI TEJA, VIth semester, Dept of MCA, SVIM - Sree Vidyanikethan Institute of Management, Tirupati.
Email.id: ravitejamar9@gmail.com

ABSTRACT
This paper gives an overview of how to forecast housing expenses with the use of python libraries and several regression algorithms. The suggested approach took into account the more sophisticated features of house price computation and provided a more accurate estimate. It also gives a rundown of the many graphical and numerical approaches that will be needed to forecast a house's price. This paper explains what a machine learning-based house pricing model is and how it works, as well as which dataset is used in our proposed approach.

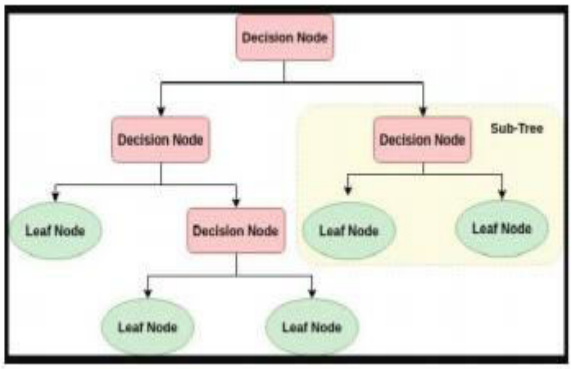
Keywords – Machine learning, Regression Technique, Classification Technique, Cross validation Technique, K-means

I INRODUCTION
Data mining is the process of finding valuable patterns or information from huge databases. One of the data mining functionalities is classification, which is used to build a model for a class attribute that is a function of other attribute values [1]. A decision tree is a tool that may be used to classify and predict data. It has a tree-like structure, with each internal node representing a test on an attribute and the test outcomes denoted by the branches that branch out from the node. The known dataset may be utilised as a training set for 80% of the time, and the test data set for 20% of the time. Each record in the dataset has X and Y values, with X denoting a collection of attribute values and Y denoting the record's class, which is the dataset's final attribute.

The Decision Tree Classifier model is built using the training set and tested using test data to determine the classifier's accuracy level. The divide and conquer technique, is used to break the training data into subsets by evaluating an attribute value. This entails attribute selection criteria; the first attribute to be evaluated is the one with the highest information gain. On the resulting subsets, the same splitting procedure is used recursively [2]. A subset's splitting procedure is complete when all of the tuples have the same attribute value or when there are no more attributes or instances to split. The construction of a decision tree does not need any prior domain expertise. It can also handle data

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

with a lot of dimensions. Decision Tree Classifiers offer a high level of classification accuracy. Once the Decision Tree has been created, new instances may be quickly categorised by tracing the tree from root to leaf node, which does not take a lot of work. Both continuous and categorical characteristics can be handled by Decision Trees.



Decision tree model

Medicine, weather, economics, entertainment, sports, and other fields use Decision Trees extensively. Decision Tress may also be used for data modification, prediction, and missing value management. It is used to distinguish tumour cells and normal cells in digital mammography, for example.

II RELATED WORK

Pow, Nissan, Emil Janulewicz, and L. Liu [11] utilised four regression approaches to estimate the property's pricing value: Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN), and Random Forest Regression, as well as an ensemble approach combining KNN and Random Forest Technique. The prices were predicted with the least error of 0.0985 using the ensemble technique, while PCA did not improve the prediction error. Several research have also looked at how to collect characteristics and how to extract them. Wu and Jiao Yang [12] evaluated several feature selection and feature extraction methods with Support Vector Regression. To anticipate housing values, several academics have built neural network models. To anticipate housing values, Limsombunchai compared hedonic pricing structure with artificial neural network model [13].

When compared to the hedonic model, the R-Squared value produced by the Neural Network model was higher, while the RMSE value of the Neural Network model was lower. As a result, they found that the Artificial Neural Network outperforms the Hedonic model. The hedonic pricing model is used by Cebula to forecast home prices in Savannah, Georgia. The number of bathrooms, bedrooms, fireplaces, parking spaces, storeys, and total square feet of a house have all been proven to be positively and significantly linked with the log price [14]. From 1993 through 2002, Jirong, Mingcang, and Liuguangyan used support vector machine (SVM) regression to forecast China's housing prices. The hyper-parameters of the SVM regression model were tuned using the genetic method. The SVM regression model's error scores were less than 4% [15]. In forecasting apartment prices, Tay and Ho contrasted the pricing predictions of regression analysis with artificial neural networks. With a mean absolute error of 3.9 percent, it was found that the neural network model outperforms the regression analysis model [16].

III IMPLEMENTATION

Data collection

The practise of obtaining information on variables in a systematic manner is known as data collection. This aids in the

discovery of answers to several questions, the formulation of hypotheses, and the evaluation of outcomes. Data collecting is the first step toward planning a social event and estimating data on certain elements in a structured framework, which then allows you to answer important questions and analyse outcomes. In all disciplines of study, including physical and sociologies, humanities, and business, information collection is an important element of research. While techniques change by discipline, the emphasis on ensuring exact and legal selection remains the same. It has been attempted on Kaggle for a variety of datasets that would fit our project's goal. This dataset was discovered after searching through a large number of datasets. It's a house pricing dataset for Ames, Iowa. This is a widely used machine learning dataset that has less mistakes and variances.

Visualising data

The pictorial or graphical depiction of data is known as data visualisation. It aids in the comprehension of complex topics and the recognition of new patterns. Many organisations regard data visualisation as a cutting-edge visual communication. It entails the development and examination of visual representations of data. Information representation use quantifiable drawings, charts, data patterns, and other devices to convey data clearly and effectively. Customers can separate and reason about data and verification with the help of effective visualisation. It gradually makes complicated data more accessible, rational, and usable.

Data pre-processing

It is the transformation of data before it is fed into the algorithm. It is used to convert unclean data into a clean dataset. It's an information-mining method that entails converting raw data into a logical structure. The final dataset used for preparation and testing is the outcome of data preprocessing. Data preparation is an information mining method that transforms raw data into a useful and productive format. Data Preprocessing is the step in any Computer Learning method when the information is modified, or encoded, to get it to a state where the machine can parse it without difficulty. The progressions applied to our data before dealing with it to the estimate are referred to as pre-dealing with.

IV PROPOSED WORK

The proposed study seeks to forecast the availability of houses based on various housing characteristics as well as the services accessible near the houses' location. The work also include estimating the price of properties based on their qualities and the amenities available in the area.

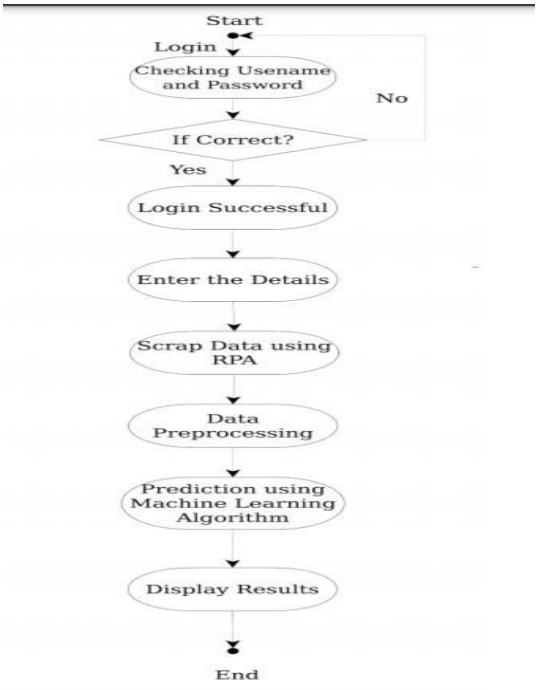
(i)A Decision Tree Classifier is used to forecast the availability of houses based on the users' requirements, and it generates yes or no replies to indicate whether or not a house is available.

(ii)To predict the values of the houses, decision tree regression and Multiple Linear Regression techniques are employed. Analyzing the site Tadepalligudem in the West Godavari District of Andhrapradesh, India, yielded a real-time dataset. The dataset includes information on the number of beds, age of the residence, transportation, neighbouring schools, and retail amenities. The suggested technique aids in the search for

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

properties in large cities based on the criteria listed below.

1. The number of rooms (1BHK, 2BHK and 3BHK).
2. Transportation options, such as bus service, rail service, and flight service.
3. Educational opportunities, such as government schools, matriculation, and CBSE.
4. Retail establishments such as local markets, general shops, and shopping malls
5. Houses range in price from 10 lakhs to 30 lakhs.
6. The house's age ranges from one to five years. Scikit Learn, a machine learning tool, is used to accomplish the suggested work.



Flow Chart Of The System

Yandex's CatBoost is an open-source machine learning algorithm. If necessary, it can readily connect with other deep learning frameworks. CatBoost can handle a wide range of data types and does not need significant data training. Unlike previous algorithms like XGBoost and Light GM, Catboost can automatically cope with categorical variables without displaying the type conversion error. It allows you to concentrate on fine-tuning your model rather than correcting minor flaws. "Min" and "Max" are the two modes used to handle missing data. The missing value for a feature is the feature's minimal value. A value is assigned to the missing value that is smaller than all of the other values. This ensures that when picking splits, a split that separates missing data from all other values is taken into account. The maximum values among all the values are treated as a missing value in "Max."

CONCLUSION

RPA's numerous advantages have made it one of the top contenders in the present industry as a field of interest for many companies across the world. RPA technology is already being used by the majority of businesses because it produces more accurate and consistent processes that are less prone to mistakes. The system extracts data using RPA and makes best use of machine learning algorithms, ensuring that the customer is satisfied by giving correct results and avoiding the danger of investing in the wrong home.

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] J.R. Quinlan, "Induction of Decision Trees", Machine Learning 1, 1986, pp.81-106.

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

[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.

AUTHOR PROFILE

[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

:

Mr B. RAVI TEJA, VIth semester, Dept of MCA, SVIM - Sree Vidyanikethan Institute of Management, Tirupati.
Email.id:ravitejamar9@gmail.com