Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/29/20231181

Machine learning techniques for predicting home rental prices in India

P. Jayadharshini[1,2], S. Santhiya[1], S. Keerthika[1], N. Abinaya[1] and S. Priyanka[1]

[1]Assistant Professor, Department of Artificial Intelligence, Kongu Engineering College, Perundurai, Tamil Nadu, India

[2]jayadharshini.p08@gmail.com

Abstract. Predicting the selling price of houses has become increasingly crucial as land and housing prices rise annually. This task is particularly challenging for metropolitan areas like Chennai and Bangalore. Therefore, there is a growing demand for an easier and more effective approach to forecast house rental prices, ensuring future generations have access to reliable predictions. Several key factors, such as the house's location and area, significantly influence rental prices. In this paper, a dataset comprising ten similar crucial features is utilized. The model is developed using a Python library, where the data is preprocessed and prepared to ensure cleanliness for constructing the model. Various machine learning algorithms, including Random Forest, Linear Regression, Decision Tree Regression, and Gradient Boosting, are employed. Through feature extraction, it is determined that area and property type are the most important features that significantly impact rental prices. Among the techniques used, gradient boosting yields the most satisfactory predictive results for rent based on evaluation metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared Metric (R^2).

Keywords: house rental price, decision tree, random forest, linear regression, gradient boosting.

1. Introduction

Real estate or property on land consists of a particular piece of land and everything permanently built on or below it. From being a basic human need, it has become a symbol of everyone's wealth and luxury. It is considered a safe investment because the value of the property increases over time. Therefore, the real estate sector is an attractive investment option for various investors. Many investors, policymakers and banks are strongly affected by changes in real estate prices. Therefore, forecasting the price of an asset is an important trading indicator. Predicting the selling price of a house is increasingly vital because both the price of land and the price of a house rise annually. It is more difficult to create models that anticipate the rental price of homes, particularly for places like Chennai and Bangalore. Therefore, in the future, our future generations will require a simple and more effective method to anticipate the price of renting a dwelling. The location, size, and other characteristics of a home can all have an impact on how much it costs to rent it out.

In this paper, several machine learning algorithms are implemented on dataset containing attributes in Bangalore and Chennai. This article introduces the most recent regression research topics that can be used to forecast rental prices, such as Linear regression, Random Forest regression, Decision Tree

DOI: 10.54254/2755-2721/29/20231181

regression and Gradient Boosting. For any machine learning model, data quality is a key factor in achieving accurate prediction results. Initial projections of home prices and missing values in the data are difficult. So to achieve higher accuracy, feature engineering plays an important role. In general, the price of a property increases over time and its market value must be calculated. This market price is required when selling the property and when applying for a mortgage. These market prices are usually determined by some professional appraiser. The main downside to this exercise, however, is that these review meetings can be biased because buyers and sellers are interested in each other. Therefore, an unbiased automated model is needed to predict asset values. For first-time buyers and inexperienced customers, this automatic model will be very useful to check if the price is exaggerated or exaggerated.

2. Related works

A significant statistic used to evaluate real estate investment decisions, [1] the speed of comeback, is connived in large part by real estate rent forecast in housing marketing research. Accurate property rent projection investment will help generate capital gains and ensure financial success. Many frequently give out [2] LUCE (Land Use And Circulation Element), the main long-term prognostic model for automated property value. LUCE addresses the lack of recent oversubscribed costs and the consequent lack of housing knowledge, which are two major issues with property assessment. As mentioned [3], a summary of how to forecast housing prices using various regression techniques with the aid of Python modules. The proposed method considered the more sophisticated aspects that were taken into account in the house value calculation and gave an additional accurate estimate House value changes are typically determined using the House Price Indicator (HPI) [4].

Despite exceptions of HPI, it is difficult to evaluate a person's home value due to the high correlation between housing value and a number of factors, including location, area, and population. As a result, this study may examine the differences between a number of complex models and analyse the varied effects of options on prediction strategies using standard and Cutting-Edge Machine Learning models. The typical property worth prediction methods don't have the capacity to analyse a lot of data, which results in poor information usage. [5] As mentioned, it proposes a house worth prediction model supported by deep learning, implemented on the TensorFlow framework, to address these issues. The model is trained using the Adam optimizer, with the Relu function being used as the activation function. The ARIMA (AutoRegressive Integrated Moving Average) model is then predicted to be supported by the trend in house value.

3. Methodology

Real estate is seen as a secure investment since real estate values rise over time. Some qualified appraisers often decide on the market pricing. Due to the conflicts of interest that buyers and sellers have, these evaluating individuals may be biased. Losses occur to inexperienced and first-time clients. The objective is to create an automated model that is objective and can anticipate property values. It will be highly beneficial for first-time shoppers and less seasoned customers to determine whether prices are undervalued or overvalued. Algorithms for machine learning are utilised to create the predictive model. The best prediction model is found by comparative analysis.The work that has been implemented focuses on a regression model that forecasts the price of homes based on key dataset characteristics including layout type, city, seller type, etc. Prior to passing the preprocessed data to the machine learning algorithms to estimate rent prices, the housing data is first preprocessed using Label Encoder.

3.1. Dataset description

The dataset utilized in this study was sourced from Kaggle and focuses on rental rates for 32,823 residences in two major Indian cities, Bangalore and Chennai. The objective of the study is to examine the key factors that influence house rent prices in these metropolitan areas. The dataset is comprehensive, containing no duplicates or missing values, ensuring data integrity. It consists of 9 informative features that provide detailed information about house rent prices in metropolitan cities.The description of each attributes that present in the dataset are, Seller Type approach i.e., through Owner or Agent. Property

DOI: 10.54254/2755-2721/29/20231181

Type includes apartment, house, flat, or villa. Locality describes about the specific area or neighborhood within Chennai & Banaglore where the rental property is located. The Layout features gives 2 values as BHK (Bedroom, Hall, Kitchen) and RK(1 Room, Kitchen). The furnishing status of the property, indicating whether it is fully furnished, semi-furnished, or unfurnished. The total area of the rental property in square feet. The monthly rental price for the property.

3.2. Preprocessing the housing data

The dataset has several labels spread across several columns. For instance, our dataset has attributes like provider type. The process of converting such tags into a digital form that is machine-readable is known as label encoding. Algorithms for machine learning can then choose the best way to mine these labels. This is a crucial supervised learning preparation step for structured data sets.

By locating important data features, feature engineering is a method for reducing the original set of data for machine learning. By using this feature engineering procedure, the number of features needed for processing can be decreased without losing any crucial or pertinent data. Additionally, feature engineering helps an analysis by reducing the amount of redundant data. The five most crucial characteristics out of the ten in the dataset are obtained for the chosen dataset following feature extraction. The accuracy of the model is increased by these five factors: area, property type, bedroom, furniture type, and seller type. Unimportant features have been eliminated because they could make the model less accurate.

3.3. Machine learning techniques

3.3.1. Linear regression. Unquestionably, one of the most used statistical modelling techniques is linear regression. Although the general idea and calculation techniques are the same, simple regression (with only one feature variable) and multiple regression (with numerous feature variables) are typically distinguished from one another. According to the linear regression concept, n quantitative feature variables, such as X1, X2, …, Xn, are combined linearly to model a quantitative dependent variable, Y.

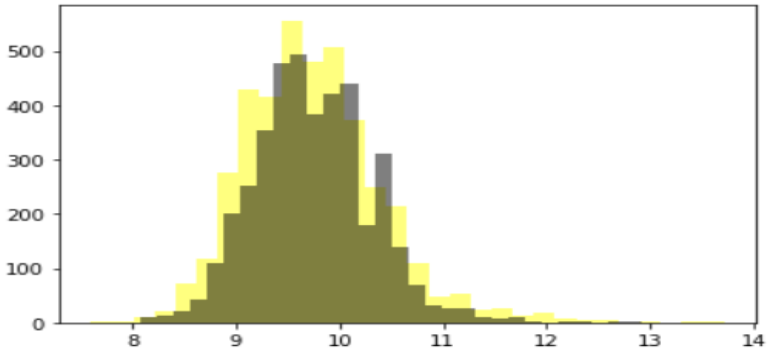

Figure 1. Linear regression.

3.3.2. Random forest regression. Here, we have utilised Random Forest for regression because it is capable of both classification and regression. It is a meta estimator, that is, it makes accurate forecasts, and it's simple to grasp the outcomes. It attempts to average the findings after matching some determinants on subsamples of the data set, which increases accuracy. It delivers results that are more accurate than the decision tree algorithm.

DOI: 10.54254/2755-2721/29/20231181

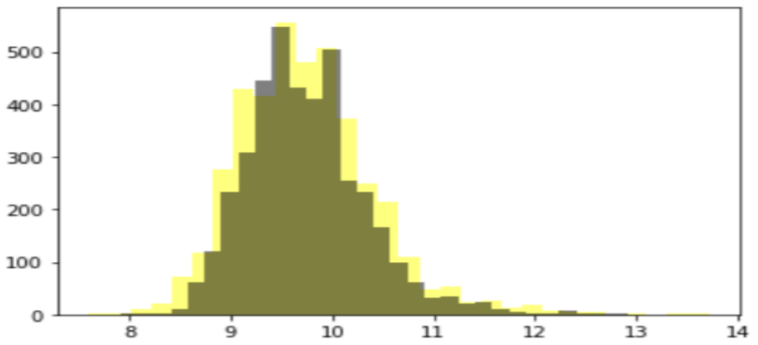

Figure 2. Random forest regression.

3.3.3. Decision tree regression. As a tree structure, decision trees can also conduct classification and regression. One of the most popular methods in supervised learning is this one. Although it separates the data set into smaller sections, a decision tree is gradually constructed in conjunction with it. Decision nodes and leaf nodes make up the tree. Tested characteristics are represented by decision nodes, while decisions are represented by leaf nodes. Decision trees can process category and numerical data.

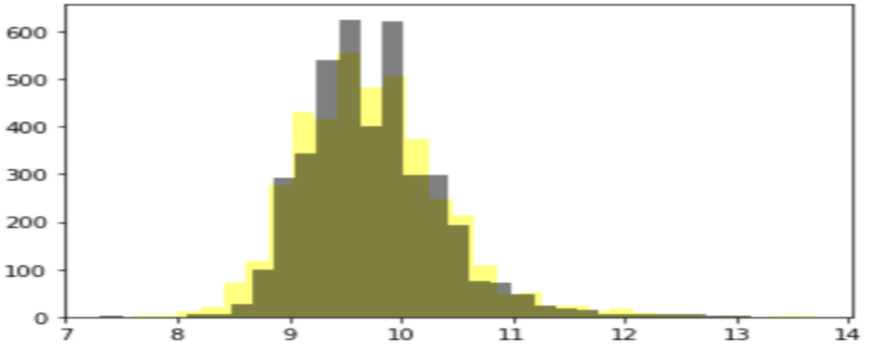

Figure 3. Decision tree regression.

3.3.4. Gradient boosting. Regression and classification can both be done with Gradient Boosting. It is a method for overcoming data matching, a problem with decision trees. This is accomplished by adding decision trees repeatedly so that mistakes in one decision tree are fixed by the next. Results from gradient boosting depend greatly on characteristics. The test results from this regression, however, are superior to those from the random forest regression when the feature settings are correct.

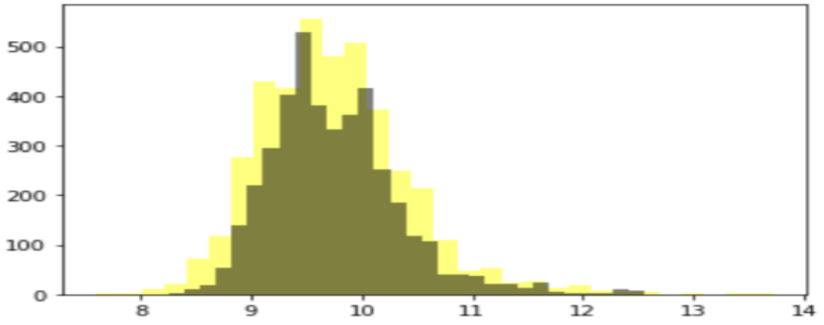

Figure 4. Gradient boosting.

DOI: 10.54254/2755-2721/29/20231181

4. Experimental results and discussion

The proposed machine learning models are implemented in Google Colaboratory. The proposed regression models use python libraries. Library pandas will be required to work with data in tabular representation. It will be necessary to use library numpy to round the data in the correlation matrix. Matplotlib and Seaborn libraries are necessary for data visualisation. Once implemented, we must assess the performance of our models to decide which one is best at making predictions. To make a technique comparison for this reason, we require a few parameters. To compare the algorithms, we used R-Squared Metric (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These three metrics listed are the most frequently used for assessing forecasts on regression machine learning issues:

4.1. R-squared metric(R^2)

It gives an notation of how well a set of predictions fits the data. This measurement is known as the coefficient of determine in statistical literature. This number falls between 0 and 1, with 1 representing a perfect match.

Table 1. Model Accuracy based on R^2 Score.

S.No.	Algorithm	R^2 Score
1	Linear Regression	0.7294196283193385
2	Decision Tree	0.7498128412512541
3	Random Forest	0.7120506578892738
4	Gradient Boosting	0.76703893865587

4.2. Root mean squared error (RMSE)

It is similar to mean absolute error in that it gives a general indication of the error's magnitude. The difference between two continuous variables is measured by this metric.

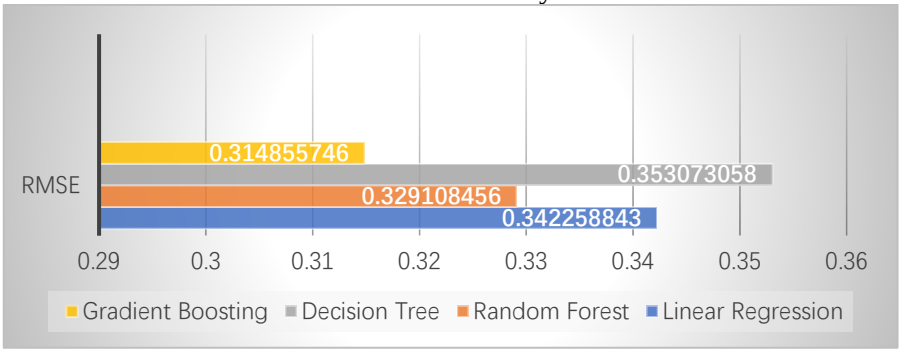

Figure 5. RMSE of the models.

4.3. Mean absolute error (MAE)

It's overall data value of the absolute differences between predictions and actual data is the mean absolute error. It illustrates how inaccurate the projections were. The measurement provides a sense of the error's size but not its direction.

DOI: 10.54254/2755-2721/29/20231181

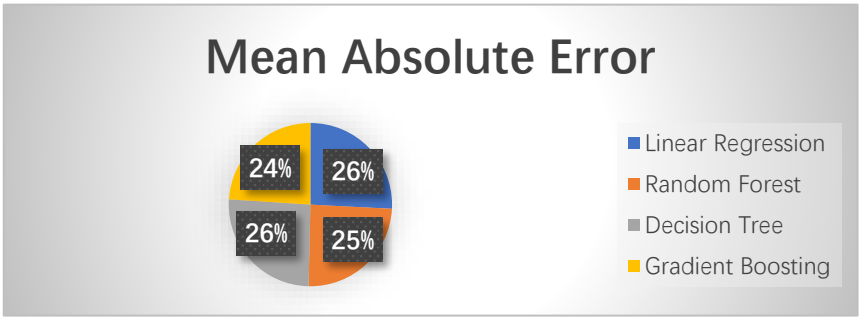

Figure 6. MAE of the models.

We can easily assess the fit of various models using Table 1 to decide which approach would produce the best results. The performance of the various models is shown in the RMSE and MAE-based graphic representations shown in Figures 2 and 3 below. By examining the error in each time point's forecast, RMSE is utilized to determine prediction performance.

Gradient Boosting combines multiple weak learners, typically decision trees, to create a strong ensemble model. This enables it to capture complex interactions and non-linear relationships within the data, resulting in more accurate predictions compared to linear models like Linear Regression. Furthermore, Gradient Boosting iteratively adjusts for errors by focusing on samples with high errors in each iteration. This helps to reduce overfitting and improve the overall performance of the model. In contrast, models like Decision Trees can be prone to overfitting and may not generalize well to unseen data. The combination of these factors allows Gradient Boosting to effectively capture the nuances and patterns present in the rental price data, resulting in higher accuracy and better predictive performance.

5. Conclusion

For many people, owning a home is a dream. Without the assistance of biassed professional appraisers, the methodology we've suggested can assist people in purchasing homes and other assets at the right price. Additionally, certain big businesses can use this model to make precise predictions and fix the market price, saving time and money. The core of real estate is legitimate real estate market values, which this approach guarantees. The model with the highest accuracy is Gradient Boosting, with a value of testing data accuracy equal to 77%, according to the modelling and evaluation results. Among all the models used, gradient boosting has the least RMSE 31% and MAE 24%. Gradient Boosting, which combines multiple weak learners (decision trees), iteratively adjusts for errors, reducing overfitting and improving the overall performance. Gradient Boosting, based on RMSE, MAE, and R^2, provides the most gratifying rental price prediction results of all the approaches. For homebuyers and sellers, accurate rental price predictions provided by Gradient Boosting can assist in making informed decisions and negotiating fair prices. It eliminates the reliance on biased professional appraisers, empowering individuals to determine the right price for homes and assets.

Acknowledgement

We would like to render our acknowledgement to Rakshithaa J, Kannan N and Tharun P, Second Year AIML students for their support in developing the proposed model.

References

[1] Maryam Heidari, Samira Zad, Setareh Rafatirad 2021. Ensemble of Supervised and Unsupervised Learning Models to Predict a Profitable Business Decision . 2021 IEEE International IOT, Electronics and Mechatronics Conference (AIMTRONICS), doi:10.1109/iemtronics52119.2021.9422649.

[2] H Peng, Jianxin Li, Z Wang, Renyu Yang, Mingzhe Liu, Mingming Zhang, Philip S Yu and Lifang He 2021 Lifelong Property Price Prediction: A Case Study for the Toronto Real Estate Market IEEE Transactions on Knowledge and Data Engineering pp. 1–1 doi:

10.1109/TKDE.2021.3112749.

[3] R Jyothsna 2022 House Price Prediction Using Machine Learning International Journal of Research Publication and Reviews Journal homepage www.ijrpr.com vol. 3 no. 11 pp. 371–380 https://ijrpr.com/uploads/V3ISSUE11/IJRPR7732.

[4] Q Truong, M Nguyen, H Dang and B. Mei 2020 Housing Price Prediction via Improved Machine Learning Techniques Procedia Computer Science vol. 174 pp. 433–442 doi: 10.1016/j.procs.2020.06.111.

[5] F Wang, Y Zou, H Zhang, and H. Shi 2022 House Price Prediction Approach based on Deep Learning and ARIMA Model IEEE Explore https://ieeexplore.ieee.org/abstract/document/8962443.