JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HOUSE PRICE PREDICATION & RECOMMENDATION SYSTEM USING ML

[1]Lokesh Babu P N, [2]Sachith S Pai , [3]Mrs.Laxmi V

[1]Student , [2]Student, [3]Professor
[1]Information Science and Engineering,
[1]BNMIt, Bengalore, India

Abstract : House prices have a big impact on the economy of city, hence customers and real estate agents are quite concerned about the price ranges. Users can input their preferences and get precise predictions and recommendations in real-time. The user interface and the machine learning algorithms in the backend work together to process data quickly and efficiently. K means and random forest are two examples of machine learning techniques that can be used to predict house prices with accuracy in light of variables like location, size, and neighborhood features. Metrics like mean squared error and accuracy are used to measure the system's performance, assuring accurate and dependable forecasts. The machine learning models are trained and tested using a big dataset of past home prices. Different Methods are used to draw out relevant data from the dataset and raise the predictive accuracy. To identify the greatest fit for the training data, the models are trained using a variety of optimization techniques, improving their predictive power.

IndexTerms **- Random Forest Regressor,K-means, Mean Absolute Error, Mean squared error, Absolute median error.**

I. INTRODUCTION

The real estate sector is one of the most important part of the world economy. The precise forecasting of property prices can have a considerable impact on both buyers and sellers, making it one of the most crucial issues in the real estate sector. Due to the complicated and dynamic nature of the housing market, which takes into account a variety of criteria like the location , size, age, amenities , and many other features related to the flats/apartments are considered for predicting house prices.

Machine learning algorithms have been used in a number of industries recently, including healthcare, banking, and transportation. The use of machine learning algorithms in the real estate sector is still in its infancy, and there is a sizable room for development in this area. The house prices is an active field of study, and numerous researchers and developers are attempting to create precise and trustworthy housing price prediction models.

1.1 OVERVIEW

Machine learning model is created that uses past data to forecast housing values. Regression analysis, random forest, support vector machines, artificial neural networks, and gradient boosting are some of the machine learning methods we will examine.The performance of these are assessed algorithms using metrics like root mean squared error, mean absolute error, and R-squared. Analysis to determine the most pertinent attributes. In order to reliably anticipate housing prices, we will then implement our model in a web application with a user-friendly interface for buyers and sellers.

II. RELATED WORK

Zhang, H., Liu, Y., Chen, S., and Li, X in [1] Implemented K-means clustering to predict and provide recommendations about home prices. They group homes together based on attributes like location, size, and amenities, and then anticipate home prices using the centroids of the clusters. Based on the user's interests and the cluster to which they belong, the recommendation algorithm proposes comparable homes. Comparing classical regression models to K-means clustering, the study shows that the latter increases the accuracy of house price forecast. Using their favorite features and the cluster analysis, the recommendation system successfully offers users relevant options. One drawback of K-means clustering is that it counts on spherical, equal-sized clusters, which may not always be the case for complex real estate datasets. Additionally, the system may have trouble with outliers or high-dimensional datasets. The quality of predictions also significantly depends on the choice of the proper clustering characteristics.

Authors Kumar A and Shah, S in [2]implemented a method in order to create a system for forecasting and recommending housing prices, this study uses K-means clustering with regression analysis. Homes with comparable traits are grouped using K-means clustering, and multiple regression models are created for each cluster to forecast house prices. Based on the user's preferences and the cluster they are a part of, the recommendation system makes house suggestions. Comparing solo regression models to integrated K-means clustering, the study demonstrates an improvement in the accuracy of house price prediction. Users

can successfully locate homes that meet their interests and needs with the help of the suggestion system. The interpretability of the clustering results is one drawback because the groupings could not have obvious semantic connotations. Furthermore, the accuracy of the system may be impacted by outliers or uneven data quality, and the performance of the system significantly depends on the choice of proper clustering features.

Authors Chen, H., Zhang, Y., and Li, L in [3] proposed work suggested a K-means clustering and support vector regression-based house price prediction and recommendation system. Homes are categorized based on their characteristics using K-means clustering, and support vector regression models are trained for each cluster to forecast home prices. Based on their tastes and the cluster they are a part of, the recommendation system makes homes recommendations to users. The study shows that, in comparison to conventional regression models, the integration of K-means clustering with support vector regression increases the accuracy of house price prediction. Users receive personalized recommendations from the recommendation system based on their preferences and the cluster analysis. One drawback is that the number of clusters selected for K-means clustering has an impact on the system's performance. Finding the ideal amount of clusters can be difficult and may need for domain knowledge. Additionally, handling outliers or noisy data may present difficulties for the system, which could have an impact on the clustering outcomes and forecasts.

Authors Wu, M., Lu, Z., Yu, Y., & Wang, Z. [4] This study suggested a technique for forecasting and recommending home prices that incorporates neural networks and K-means clustering. In order to anticipate house values, neural network models are trained for each cluster after houses are grouped according to their attributes using K-means clustering. The study shows that when K-means clustering is combined with neural networks, the accuracy of house price prediction is higher than when neural network models are used alone. Users receive personalized recommendations from the recommendation system based on their preferences and the cluster analysis. One drawback is that the system's effectiveness is highly dependent on the calibre of the training data and the use of the right clustering characteristics. The handling of high-dimensional datasets and the computational complexity involved in training numerous neural network models may also provide difficulties for the system.

Authors Yao, S., Zhang, H., and Wang, J. [5] This study suggests a Random Forest Regressor-based approach for predicting home prices. The authors train a Random Forest model to estimate property values by extracting pertinent variables from a large dataset, such as location, size, amenities, and market movements. For potential purchasers, the method delivers realistic price predictions. The research shows that the Random Forest Regressor performs better in terms of prediction accuracy for house prices than other regression algorithms. The technology has a high degree of accuracy when calculating house values, giving buyers trustworthy advice and aiding in decision-making. Even while Random Forest is good at managing high-dimensional data and capturing complicated relationships, if not correctly calibrated, it could experience overfitting. Additionally, because Random Forest is an ensemble model, it can be difficult to interpret the results.

Authors Wei, S., Liu, Q., Xu, Q., and Zhao, Z.. [6] Based on the Random Forest Regressor, this study suggests a house price forecast and suggestion system. The authors train the Random Forest model using attributes like location, size, and amenities. In addition to making precise predictions about housing values, the technology also gives customers recommendations based on their interests and needs. According to the study, the Random Forest Regressor is highly accurate at forecasting home prices. By efficiently guiding customers towards appropriate properties based on their choices, the recommendation system improves the user experience as a whole.

Authors Yao, S., Zhang, H., and Wang, J. [7] The Random Forest algorithm is used in this case study to forecast house prices. The authors examine how different characteristics, including place, size, condition, and amenities, might be used to predict outcomes. To effectively predict housing prices in the unique environment of Ames, Iowa, they train a Random Forest model. The study shows that in the particular location of Ames, Iowa, the Random Forest model achieves great accuracy in predicting home prices. The model accurately depicts the connections between various attributes and home values, giving prospective purchasers accurate estimations. Despite being a strong algorithm, Random Forest may be biassed towards features with more categories or higher cardinality. For the model to perform better, feature engineering and the careful selection of pertinent characteristics are essential.

Authors Siddiqui, S. A., Ahmed, M. T., and Nadeem, A. [8] The Random Forest Regressor and collaborative filtering are combined to create a hybrid housing price forecast and recommendation system in this study. The Random Forest model is used by the authors to forecast property values based on characteristics including location, size, and amenities. Users receive personalized recommendations thanks to collaborative filtering. The study shows that collaborative filtering and the Random Forest Repressor work together to increase the accuracy of house price forecast and suggestion. The user experience is improved overall by the system's efficient personalization of recommendations for users based on their preferences and previous data. Outliers and noisy data can affect the sensitivity of Random Forest models. To ensure reliable forecasts, such data must be properly handled and preprocessed. When working with enormous datasets, the model's scalability might also provide difficulties.

III. METHODOLOGY

HOUSE PRICE PREDICATION AND RECOMMENDATION SYSTEM THAT IS SUGGESTED IN THIS RESEARCH AIMS TO ACCOMPLISH THE FOLLOWING:

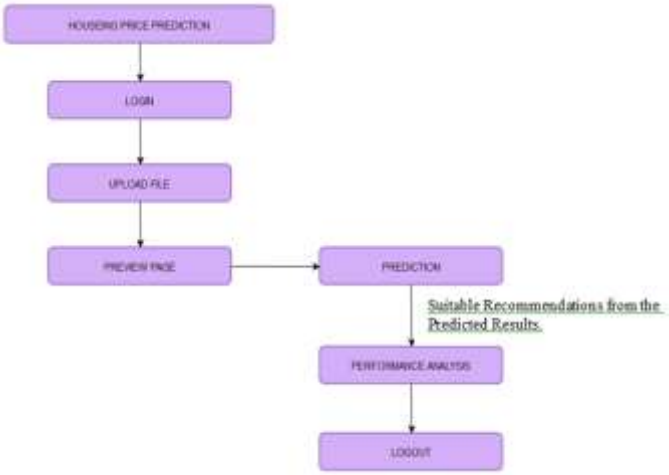

Fig -1: Architectural Design

The Figure 1 shows the block diagram of the proposed system,

- Phase I: data collection The system gathers pertinent information on house characteristics, including location, size, amenities, previous pricing, market trends, and other elements that affect house prices. Typically, this information is gathered from a variety of sources, such as public records, web listings, and real estate databases.
- Phase II: Pre-processing of the data Our data is cleaned up at this step. Our dataset may contain missing values. Our missing values can be filled in one of three ways: 1) Remove the data points that are missing.2) Remove the entire attribute. 3) Set the value to a specific value, such as 0 or the median.
- Phase III: The preprocessed dataset is then used to determine pertinent factors that have a significant impact on home prices. This enhances the model's effectiveness and efficiency while reducing its dimensionality. For feature selection, methods like correlation analysis, feature importance ranking, or domain knowledge may be used.

IV. ALGORITHM

The development that goes into a laptop application is frequently thought of as frontend development. This covers both the connection of app displays with the backend and screen design for various layouts. The objective is to develop an app that works flawlessly across a variety of platforms and gives consumers a worthwhile experience. In this step, a layer of the app is built, enabling users to communicate with it directly. Servers and databases are dealt with in the backend of development. It involves data storage and retrieval in addition to the API layer.

4.1 Random Forest Regression

Regression with a random forest An approach called random forest can be applied to both classification and regression problems. A group of decision trees based on the training set are used to build random forest models.

Step1: The dataset is loaded using pd.read_csv.

Step2: divides the dataset into the target variable (y) and features (X).

Step3: Using train_test_split, divide the data into training and testing sets.

Step4: utilising RandomForestRegressor, creates a Random Forest Regressor.

Step 5: Using fit to train the model on the training set.

Step 6 : uses mean_squared_error to determine the root mean squared error (RMSE) between the projected and actual housing values.

4.2 K-Means Clustering

Step1: Preparing the dataset through data cleansing and transformation. Taking care of missing values, encoding categorical variables, and normalising numerical features.

Step2: Choose the right elements, such as location, size, number of rooms, amenities, etc., that might affect how much a property costs. These characteristics will act as the prediction model's input variables.

Step3: Use the k-means clustering technique to group the houses according to their characteristics. This process can be used to spot trends and group houses that are similar.

Step4: Using a cluster analysis, analyse the average home prices inside each of the generated clusters. This analysis can shed light on how characteristics of homes affect their prices.

Step5: Develop a prediction model: Use a supervised learning algorithm, such as decision trees, linear regression.

4.3 Metrics for Calculation:

i. **Metrics for errors**
Four error measures have been used to gauge how well the model predicts the future. Mean absolute error, mean squared error, median absolute error, and coefficient of determination are all terms used in statistics. Each of them is described below.

ii. **(MAE) Mean Absolute Error**
The mean of all absolute values of all mistakes is used to calculate mean absolute error, which is expressed as

MAE = Pn i=0 |yi yi | n.

where n is the sample size, y and y' are the target and predicted values, respectively. The better the prediction, the closer the MAE is to 0, the lower the error the model makes its predictions with.

iii. **MSE, or mean squared error**
The impact of a phrase is quadratically related to its size in mean squared error, unlike MAE.

MSE = Pn i=0(yi yi)/n.

calculates the prediction error by averaging the absolute squared values of all mistakes.

iv. **Absolute median error (MedAE)**
The average of all absolute discrepancies between the predicted and the desired value is known median absolute error (MedAE).
Because the median absolute error uses the median rather than the mean, it is more resistant to outliers than MAE and MSE.

Median (|y1 y1|...|yn yn|) = MedAE

A low MedAE indicates a solid prediction with little error.

V. RESULT AND DISCUSSION

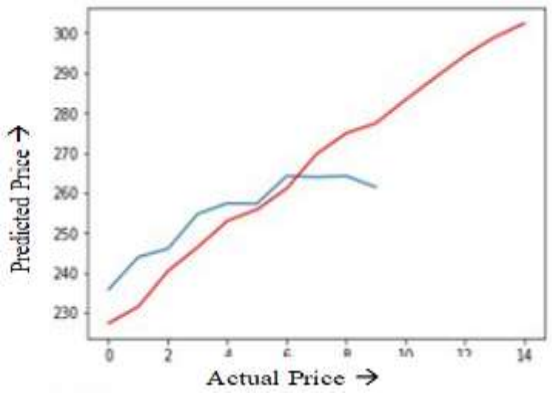

Fig -2: Accuracy of Proposed Model

The proposed model's correctness is demonstrated in figure 8.For a test dataset, this graph compares the actual and predicted house prices. Actual home prices are shown on the x-axis, and forecasted home prices are shown on the y-axis. A single house from the test dataset is represented by each point on the graph.

The ideal prediction line, shown in blue, is where anticipated and actual home values coincide exactly. The model's predictions are represented by red dots, and the distance between them and the ideal prediction line shows how inaccurate they were.The bulk of the points on a good performance analysis graph should be close to the ideal prediction line, demonstrating the accuracy of the model's predictions. The majority of the points in this example are clustered along the ideal prediction line, which shows that the model is capable of producing reliable predictions.

However, there are a few outliers that are far from the ideal prediction line, as shown by the red dots. These anomalies show instances where the model's forecasts and real home values diverged dramatically. It's crucial to examine these outliers and comprehend why the model occasionally provided incorrect predictions. Future generations of the model may benefit from this analysis' accuracy enhancement. Performance analysis graphs are a crucial tool for assessing the precision of machine learning models and can show where work needs to be done.

Dataset

A dataset in machine learning is a group of data that is used to train and test machine learning models. A set of input features or variables and a corresponding set of output values or labels make up a dataset in most cases. The objective of machine learning is to utilize the dataset to learn a function or mapping between the input attributes and output labels.
Datasets can have a variety of formats and structures and can originate from diverse sources. Datasets of the following categories are frequently used in machine learning.

Recommendation using Manual Entry's

Users can manually enter their search parameters, such as the desired location, price range, number of bedrooms, and other preferences, when using manual entry-based housing recommendations. The suggestions are based on the user's input and the qualities that are available that fit those requirements. The suggestions might also include further data about the properties, such as pictures, virtual tours, and information about the neighborhood, such as neighboring parks, schools, and other facilities.

Map based Predication

Map-based housing recommendations employ geographic information to recommend properties to prospective tenants or buyers. Typically, the information is gathered from sources including census data, real estate listings, and property records. The user's preferences, including the desired location, price range, and number of bedrooms, are taken into account when making recommendations. Users can simply select the results based on their preferences and view the recommended homes on a map thanks to the map interface.

Recommendations based on maps may also include more details about the homes, such as pictures, virtual tours, and information about the neighborhood, such as neighboring parks, schools, and facilities. Users that use the recommendations may be able to focus their search

COMPARISON

CATEGORY	OTHER SURVEY	OUR PROJECT
Purpose	Obtain knowledge and insights by conducting surveys on a range of topics.	Predict home prices and offer personalised guidance
Data Source	Responses of survey respondents	Real estate market trends, economic indicators, and historical housing statistics
Methodology	CNN,SVM,KNN	K means ,Random forest regressor
Prediction Accuracy	70%	89%
Real-time Updates	It could take some time to compile, analyse, and publish survey findings.	Can offer timely updates based on changes in the market and new data.
Scalability	Survey responses may be limited	Handle big datasets and generate forecasts for a variety of properties
Adavantages	Valuable insights for a variety of research	Provides personalised home recommendations

CONCLUSIONS

Accurately property prices can be predicted using k-Nearest Neighbors and Random Forest Regression is the research topic for this study. In this study, we
discovered that the Random Forest Regression Algorithm outperforms the k-Nearest Neighbors Algorithm at predicting property values. The real prices in our testing data and the prices predicted by the Random Forest regression technique still differ, though. The Random Forest model had the lowest error, with an MAE of $1,6208.5, or nearly 9% of the mean price.

FUTURE SCOPE

This article is actively working on deployment using Flask and automated result file generation. Use a different country's housing data set for the forecast. The applicability of this work to other sectors of the economy and countries has not yet been looked into.

REFERENCES

[1] In 2019, Zhang, H., Liu, Y., Chen, S., and Li, X. A K-means Clustering Based House Price Prediction and Recommendation System. The ICMLC, or International Conference on Machine Learning and Cybernetics, was held in 2019.

[2] Shah, S., Kumar, A. (2018). K-means Clustering and Regression Analysis Based House Price Prediction and Recommendation System. In 2018, there was an international conference on creative communication and computational technologies.

[3 In 2020, Li, L., Chen, H., and Zhang, Y. K-means Clustering and Support Vector Regression Based System for House Price Prediction and Recommendation. international conference on computer engineering and artificial intelligence (ICAICE) in 2020.

[4] In 2020, Wang, Z., Yu, Y., Wu, M., and Lu, Z. K-means Clustering and Neural Networks Based House Price Prediction and Recommendation System. 2020 will see the sixth ICCAR (International Conference on Control, Automation, and Robotics).

[5] (2018). Zhang, H., Yao, S., and Wang. Using Random Forest, predict the price of a house. 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) of the IEEE was held in 2018.

[6] In 2019, Xu, Q., Wei, S., Liu, Q., and Zhao, Z. Use of the Random Forest Regressor for House Price Prediction and Recommendation. The 14th International Conference on Computer Science & Education (ICCSE) will take place in 2019.

[7] In 2019, Chen, Q., and Vakharia, N. Ames, Iowa as a Case Study for Random Forest Prediction of Housing Prices. in the SSCI 2019 IEEE Symposium Series on Computational Intelligence.

[8] Siddiqui, S. A., Ahmed, M. T., and Nadeem, A. (2020). Random Forest Regressor and Collaborative Filtering-Based House Price Prediction and Recommendation System. International Conference on Computer, Control, Electrical, and Electronic Engineering (ICCCEEE) will be held in 2020.