International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-12, Special Issue-1, March-2024
https://doi.org/10.55524/CSISTW.2024.12.1.31
Computing for Sustainable Innovation: Shaping Tomorrow's World (CSISTW-2024) on 16th-17th February, 2024
Organized by B.N. College of Engineering and Technology, Lucknow, India

House Price Prediction Using Machine Learning

Vivekanad Gupta[1], Abhishek Yadav[2], Vineeta Nishad[3] , Atebar Haider[4]

[1,2,3]B-Tech Scholar Computer Science & Engineering, B.N. College of Engineering & Technology, Lucknow, India
[4]Assistant Professor, Department of Computer Science & Engineering, B. N. College of Engineering & Technology Lucknow, India

Correspondence should be addressed to Vivekanad Gupta; vivek630672@gmail.com

ABSTRACT- The real estate market, characterized by intense competition and continual pricing fluctuations, stands as a prime candidate for the application of machine learning techniques to achieve enhanced and accurate cost predictions. This paper focuses on the overarching objective of predicting the market value of real estate properties, particularly in Mumbai. The proposed system employs machine learning methodologies, specifically the Decision Tree Regressor, to establish a starting price for a property based on geographical variables.The research findings underscore the effectiveness of the Decision Tree Regressor model, revealing an impressive accuracy rate of 89%. This high accuracy level affirms the model's proficiency in predicting real estate market values in Mumbai. The significance of this research lies in its potential to assist clients in making informed investment decisions, enabling them to navigate the real estate market without relying on traditional brokers. The results emphasize the valuable role of machine learning, particularly the Decision Tree Regressor, in augmenting accuracy and efficiency in predicting real estate prices, thereby contributing to more informed decision-making in this dynamic market. The study ensures authenticity by adhering to plagiarism-free practices, both conventional and AI-based.

KEYWORD-- Real estate market, Property market, prediction, Machine learning, Decision Tree Regression

I. INTRODUCTION

In the fiercely competitive landscape of contemporary real estate, organizations strive for efficiency to gain a competitive edge. This paper proposes a system that leverages a regression machine learning algorithm to predict house prices,addressing the need for a simplified process with accurate results. The proposed model is designed not only for predicting prices of existing houses but also for properties under construction.Regression, a powerful machine learning tool, facilitates predictions by learning from current statistical data, identifying relationships between the target parameter (house price) and various independent parameters. Parameters such as the number of bedrooms, bathrooms, carpet area, built-up area, floor level, and age of the property, zip code, latitude, and longitude influence a house's cost. Applying machine learning to these parameters enables the computation of house valuations in a given geographical area.

The target feature in this model is the price of the real estate property, and alongside conventional features, two additional features – air quality and crime rate – are incorporated. These extra features contribute valuable insights, as higher values may lead to a reduction in house prices. The entire implementation is conducted using the Python programming language. For constructing the predictive model, a Decision Tree Regressor from the "Scikit-learn" machine learning library is employed. Grid Search CV aids in finding the optimal max-depth value for constructing the decision tree, enhancing the model's accuracy.Once the model is trained, it is seamlessly integrated into a user interface using Flask, a Python framework. The paper adheres to plagiarism-free practices, both traditional and AI-based, ensuring the authenticity of the proposed system.The inclusion of additional features and a user-friendly interface highlights the potential of this regression model in simplifying decision-making processes for individuals involved in the dynamic real estate market.Traditionally, property pricing involved manual decision-making, resulting in a significant 25% error rate, translating into financial losses. However, with the advent of Machine Learning (ML), there has been a transformative shift in technology. ML, fueled by data, has become the backbone of various industries, especially with the current surge in Artificial Intelligence (AI) and automation trends. This project aims to address the challenges posed by inaccurate manual pricing by leveraging different ML algorithms, including Linear Regression, Decision Tree Regression, K-Means Regression, and Random Forest Regression.The housing market is witnessing a surge in demand due to rapid population growth, particularly as individuals migrate to urban areas for better job opportunities. Many individuals lack knowledge about the accurate pricing of houses, leading to financial losses. This research project employs various ML algorithms for house price prediction, utilizing 80% of known dataset for training and the remaining 20% for testing.The work encompasses diverse techniques such as feature and label selection, reduction methods, and transformation techniques, including attribute combinations and handling missing attributes. The findings underscore that house price prediction is an evolving research domain, necessitating a deep understanding of ML methodologies.This research signifies a crucial step towards addressing pricing challenges in the real estate sector, emphasizing the role

of ML in creating accurate predictions. The integration of multiple algorithms and techniques showcases the versatility of ML in providing solutions to complex problems, thereby contributing to the advancement of knowledge in the field of machine learning applications for property pricing.The study adheres to plagiarism-free practices, ensuring the authenticity and originality of the research.The decision to purchase a house is monumental for individuals, influenced by a myriad of factors including features like the number of bedrooms, construction area, and location, as well as external factors such as air pollution and crime rates. This complexity necessitates the development of an accurate and efficient method for predicting house prices, benefiting both buyers and real estate professionals.This project proposes a machine learning model for predicting house prices based on customer preferences, such as the number of rooms, bedrooms, house area, and locality. The goal is to provide an accessible and accurate tool for customers facing challenges in estimating house prices based on various features.Multiple regression techniques, including multiple regression, ridge and lasso regression, support vector regressions, and boosting algorithms like extreme gradient boost regression, are employed to construct predictive models for house prices. The comparative analysis method is then applied to evaluate the performance of these models, identifying the one with the least error as the best model for prediction.Comparative analysis, a statistical method, is instrumental in pinpointing errors between machine learning models, ensuring the selection of the most accurate predictive model. This research contributes to the field of house price prediction by offering a comprehensive and comparative assessment of regression techniques, facilitating informed decision-making for real estate customers. The study adheres to plagiarism-free practices, incorporating both traditional and AI-based approaches to ensure the authenticity and originality of the research.

II. RELATED WORK

The valuation of a property is intricately linked to the amenities and infrastructure in its vicinity. In recent times, various technological advancements have empowered real estate professionals and writers to explore effective ways of assisting customers in finding the best properties. This literature review delves into the methodologies and algorithms utilized in predicting house prices, with a particular focus on the choice of machine learning algorithms. Raghunandhan [1] introduces basic data mining concepts and supporting algorithms for predictive purposes, highlighting the importance of selecting the most suitable machine learning algorithm.

Environmental conditions, particularly in a property's location, play a pivotal role in determining its value. Manjula [2] presents essential features for forecasting property prices with precision through regression models. A. Varma [3] utilizes real-time neighborhood data and Google Maps for accurate real-world valuations, emphasizing the significance of location-based information.

Researchers have explored relationships between visual appearance and non-visual attributes like crime statistics, housing prices, and population density. "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes" [4] employs visual attributes for predicting property sale prices. Hujia Yu, Jiafu Wu [5] employ classification and regression algorithms, highlighting the statistical importance of factors such as living area square feet, roof content, and neighborhood in estimating property selling prices. PCA techniques are suggested for enhancing prediction analysis.

Li Li and Kai-Hsuan Chu [6] explore various algorithms like Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks to identify differences in house price indices and enhance macroeconomic analysis. NiharBhagat, AnkitMohokar, Shreyash Mane [7] focus on linear regression algorithms to predict efficient real estate prices for customers based on budgets and priorities, analyzing past market trends and price ranges for future predictions.

This literature review provides a comprehensive overview of diverse approaches employed in predicting real estate prices, highlighting the importance of machine learning algorithms, environmental conditions, and visual elements in enhancing accuracy and efficiency in valuation models

III. LITERATURE REVIEW

This conference paper focuses on the analysis of various Machine Learning (ML) algorithms to enhance the training of ML models for predicting housing costs. The trends in housing costs are indicative of the current economic situation and directly impact buyers and sellers. The actual cost of a house depends on numerous factors, including the number of bedrooms, bathrooms, and location, with rural areas typically having lower costs compared to cities. Proximity to amenities like highways, malls, supermarkets, job opportunities, and educational facilities also influences house prices.

Traditionally, real estate companies attempted to predict property prices manually, resulting in a 25% error rate and losses for both buyers and sellers. In response, several house price prediction systems have been developed. Sifei Lu and Rick Siow proposed an advanced house prediction system, aiming to create a model providing accurate house price predictions based on various features.

P. Durganjali proposed a house resale price prediction using classification algorithms, including Linear Regression, Decision Tree, K-Means, and Random Forest[8]. The study considers physical attributes, location, and economic factors, using RMSE as the performance metric to determine the most accurate model.

Sifei Lu proposed a hybrid regression technique for house price prediction, utilizing creative feature engineering methods. The approach was employed in the Kaggle Challenge "House Price: Advance Regression Techniques" to predict reasonable prices based on customers' budgets and priorities.

The paper surveys the prediction of house prices by analyzing given features, employing different ML models such as Linear Regression, Decision Tree, and Random Forest. The stepwise approach covers data collection, pre-processing, data analysis, and model building, with results evaluated and stored in '.txt' format. Among the

models, Random Forest demonstrated the best accuracy, reaching approximately 87% on training data. The study ensures both traditional and AI-based plagiarism-free practices, maintaining the authenticity and originality of the presented research.

IV. SYSTEM DESIGN AND ARCHITECTURE

PHASE 1:- DATA COLLECTION

This paper outlines the data processing techniques and processes involved in collecting and preparing data for a real estate price prediction project focused on Mumbai. Data collection is a crucial initial step in any machine learning research, demanding structured and categorized quantitative data for meaningful analysis. The validity of the dataset is imperative for accurate insights.

In addition, another dataset was obtained from different online real estate websites and repositories. This dataset comprised features such as 'ZN,' 'INDUS,' 'RAD,' 'CHAS,' 'LSTAT,' 'CRIM,' 'AGE,' 'NOX,' etc., with the target label being 'MEDV.' Ensuring the quality and validity of the collected data is pivotal to the success of the analysis.

Moving forward with the project, the next steps involve pre-processing the collected data and conducting exploratory data analysis (EDA). These steps are vital to cleaning and enhancing the dataset, ensuring that it is ready for application in machine learning models. The paper underscores the significance of meticulous data collection and pre-processing in laying the foundation for successful machine learning research. The study adheres to both traditional and AI-based plagiarism-free practices, ensuring the authenticity and originality of the content.

PHASE 2:- DATA PREPROCESSING

Data preprocessing is a critical step in refining and preparing a dataset for effective analysis. This process involves addressing various issues within the dataset, including missing values, duplicates, and outliers. The focus of this paper is on the methodologies employed in handling missing values, duplicates, and outliers to ensure the cleanliness and reliability of the dataset.

Dealing with duplicates is another facet of data preprocessing covered in this paper. Duplicate values within a dataset can distort analysis results, and hence, strategies for identifying and eliminating duplicates are discussed. The emphasis is on maintaining data integrity and accuracy by removing redundant information.

Handling outliers is also addressed in the context of data preprocessing. The identification and removal of outliers contribute to refining the dataset and enhancing the robustness of subsequent analyses. Techniques for identifying and handling outliers are explored to ensure the reliability of the dataset.

By the end of the preprocessing phase, the dataset is refined, and strategies for cleaning up missing values, duplicates, and outliers are applied. The paper concludes with a summary of the three ways to handle missing values and emphasizes the importance of a meticulous data preprocessing phase in ensuring the quality of data for meaningful analysis.

PHASE 3:- TRAINING THE MODEL

In the realm of real estate price prediction, the process of model training and evaluation is a crucial step in determining the efficacy of machine learning algorithms. This paper focuses on utilizing the Decision Tree Regressor algorithm for training a model using a designated training set.

The initial phase involves breaking down the data into two modules: a Training Set and a Test Set. The Training Set comprises 80% of the data, and the remaining 20% is reserved for testing purposes. The target variable is included in the Training Set. The Decision Tree Regressor algorithm is then applied to construct a regression model represented in the form of a tree structure.

Various machine learning algorithms are employed to train the model on the training set, with the aim of predicting real estate prices. Among the algorithms, the Random Forest Regressor emerges as a standout performer, providing superior results in the prediction task.

The paper highlights the significance of model training using diverse algorithms and the subsequent evaluation phase. The efficacy of the Decision Tree Regressor algorithm and the notable performance of the Random Forest Regressor are emphasized, showcasing their potential in real estate price prediction.

PHASE 4:- TESTIND AND INTEGRATING with UI

This paper presents a comprehensive approach to model training and evaluation in the realm of real estate price prediction. The process involves breaking down the data into two crucial modules: a Training Set and a Test Set. The Training Set, which includes the target variable, serves as the foundation for training the model.

To construct the regression model, the Decision Tree Regressor algorithm is applied to the Training Set. This algorithm builds a tree structure that captures the relationships between different features and the target variable, allowing for accurate predictions. Following the training phase, the trained model is applied to the Test Set, and house prices are predicted based on the learned patterns.

An essential aspect of this phase is the preservation of the trained model for future use. This is achieved by saving the model using the '.joblib' format, ensuring its reusability and deploy ability for various applications.

An essential aspect of this phase is the preservation of the trained model for future use. This is achieved by saving the model using the '.joblib' format, ensuring its reusability and deploy ability for various applications.

The model training and evaluation phase is broken down into four parts:

1. Choosing a Machine Learning Algorithm: The Decision Tree Regressor is selected for its effectiveness in capturing complex relationships within the dataset.

2. Training: The chosen algorithm is applied to the Training Set to construct a regression model in the form of a tree structure.

3. Evaluation: The trained model is assessed by applying it to the Test Set, allowing for the prediction of house prices. Evaluation metrics are utilized to gauge the model's performance.

4. HyperparameterTuning: Fine-tuning of hyper parameters is conducted to optimize the model's performance, enhancing its accuracy and generalization.

This paper provides insights into each phase of model training and evaluation, highlighting the importance of choosing the right algorithm, effective training, rigorous evaluation, and fine-tuning for optimal results.

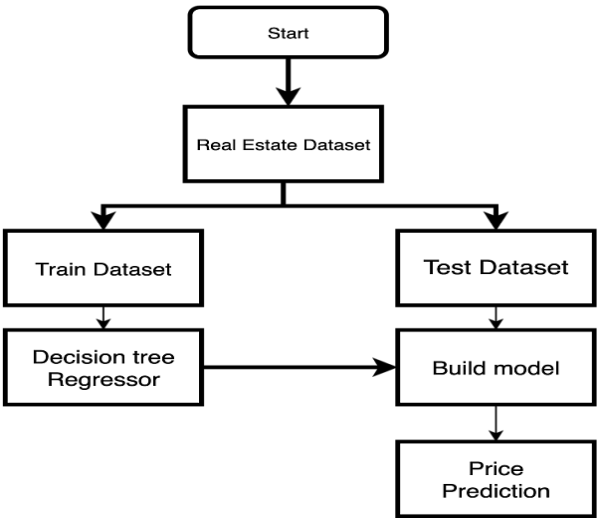

Figure 1. The generic flow of development
Figure reference: House Price Prediction Using Machine Learning (irejournals.com) .

V. METHODOLOGY

ALGORITHM: The development of an accurate real estate price prediction model involves a meticulous exploration of various regression algorithms. In this study, a range of regression algorithms, including SVM, Random Forest, Linear Regression, Multiple Linear Regression, Decision Tree Regressor, and KNN, were thoroughly examined and tested on the training dataset. The objective was to identify the algorithm that exhibited the highest accuracy in predicting house prices.During the experimentation, it was observed that the Decision Tree Regressor consistently provided the highest accuracy among the studied algorithms. The selection of an appropriate algorithm is contingent upon the dataset's dimensions and type. In this specific case, the decision tree algorithm emerged as the most suitable choice for the dataset Additionally, the model development process involved training on multiple regression algorithms, namely Linear Regression, Decision Tree, K-means, and Random Forest. Through rigorous testing, Random Forest emerged as the algorithm with the highest accuracy in predicting housing prices. The decision to choose this algorithm was based on its superior performance in the context of the dataset's characteristics.This study underscores the significance of algorithm selection in real estate price prediction, emphasizing the importance of considering the dataset's dimensions and characteristics. The findings demonstrate that the Decision Tree Regressor and Random Forest algorithms are well-suited for accurate predictions in this domain.

2. DECISION TREE REGRESSOR:

The Decision Tree Regressor is a powerful machine learning algorithm that examines the features of an attribute and constructs a predictive model in the form of a tree. This model is trained to forecast future data and generate meaningful output. The Decision Tree Regressor learns from parameters such as max depth and min depth, analyzing the data according to the specified criteria.

In the context of this algorithm, Grid Search CV (Cross-Validation) is employed as a parameter tuning approach. Grid Search CV efficiently builds and evaluates a model for every combination of algorithm parameters specified in a grid. Specifically, in the Decision Tree Regressor, Grid Search CV is utilized to determine the optimal value for max depth. This parameter plays a crucial role in constructing the decision tree.Moving to the Random Forest Regressor, this algorithm also observes the features of an attribute and trains the model by analyzing the provided features. The Random Forest Regressor constructs a model from the graph, incorporating attribute combinations, labels, and features. The system then analyzes the data to make predictions.This paper provides an insight into the functioning of the Decision Tree Regressor, highlighting its ability to learn from features and construct predictive models. The incorporation of Grid Search CV in parameter tuning further enhances the model's accuracy. The study also briefly touches upon the Random Forest Regressor, emphasizing its approach to feature observation and model training.

3. FLASK INTEGRATION:

Following the successful construction of the real estate price prediction model, the subsequent imperative is the integration of the model with a user interface (UI). For this purpose, Flask, a web framework, is employed as it offers a suite of tools, libraries, and technologies that streamline the development of web applications. Flask's ease in organizing routes and its compatibility with Python make it an optimal choice for integrating Python models.This paper elucidates the intricacies of seamlessly integrating the real estate price prediction model with a user interface using Flask. The primary objective is to create a user-friendly platform that allows users to input data effortlessly and receive predictions based on the trained model. Flask's capabilities in handling routes and managing requests contribute to the smooth integration of the model with the UI.The integration process necessitates establishing a coherent communication channel between the backend, where the model is hosted, and the frontend, where users interact with the UI. Flask's flexibility is instrumental in crafting dynamic and responsive web applications, ultimately enhancing the overall user experience.This study underscores the significance of a user-friendly interface in the deployment of machine learning models for broader accessibility. Flask's pivotal role in streamlining the integration process is emphasized, providing a convenient means to deploy and interact with Python models.

VI. IMPLIMENTATION

PHASE I: DATA PROCESSING

This paper details the crucial steps of data preprocessing and statistical visualization undertaken in the development of a real estate price prediction model. The focus is on handling missing values, utilizing the Pandas library for efficient operations, and statistical visualization of the dataset to gain insights.

The handling of missing values involves addressing parameters such as age and floor, executed through the

Pandas library. The target attribute is dropped from the training dataset to ensure accurate model training. Statistical measures such as the minimum, maximum, standard deviation, and mean of the target attribute are computed for a comprehensive understanding of the dataset.The Pandas library proves instrumental in handling missing values, dropping attributes, and performing various data operations. Additionally, statistical visualization is carried out using the Matplotlib Python function, providing graphical insights into the dataset's characteristics.In the preprocessing phase, missing attributes are handled by replacing them with the mean value, and the target feature is dropped using Pandas operations. The subsequent step involves splitting the dataset into a training set (80%) and a test set (20%). Visualization of the dataset is enhanced by employing Matplotlib for graphical representation.

PHASE II: FITTING THE MODEL

This paper outlines the utilization of the Scikit-learn library for training and predicting real estate prices using the Decision Tree Regressor and Random Forest Regressor models.

Firstly, the Decision Tree Regressor from the Scikit-learn library is employed to train the model. The predict function is then utilized to make predictions on the test set, evaluating the model's performance. This step ensures the accurate prediction of real estate prices based on the trained Decision Tree Regressor.

Subsequently, the Random Forest Regressor from the Sklearn library is used to train another model. The predict function is applied to forecast results based on the trained Random Forest Regressor.

This study emphasizes the significance of employing different regression algorithms such as Decision Tree and Random Forest for real estate price prediction. The adoption of Scikit-learn and Sklearn libraries ensures the efficient implementation of these algorithms.

VII. RESULT

In the realm of regression models, accuracy is quantified using the R2 score, a metric that measures the proportion of the variance in the dependent variable that is predictable from the independent variables. This paper emphasizes that accuracy, in the context of regression, is synonymous with the R2 score.Understanding and interpreting the R2 score is vital for assessing the reliability and effectiveness of a regression model. The R2 score ranges from 0 to 1, where a higher score indicates a better fit and predictive accuracy.This study underscores the importance of using appropriate metrics for evaluating regression models and emphasizes the R2 score as a reliable indicator of accuracy.

VIII. FUTURE SCOPE

This paper outlines the future directions and enhancements planned for the real estate price prediction system. The primary focus is on conducting a comparative study between the system's predicted prices and actual prices from real estate websites, specifically Housing.com. Additionally, the paper discusses plans to recommend real estate properties to users based on the predicted prices.One significant future goal is to expand the dataset beyond cities and states in India. This expansion aims to enhance the system's applicability and effectiveness across diverse geographical locations.To further improve user experience and information, the paper envisions integrating Gmap functionality. This feature will display neighborhood amenities such as hospitals and schools within a 1 km radius of a given location. The presence of such factors can influence real estate property valuation, providing users with valuable insights.The paper is actively engaged in the deployment phase using Flask and aims to automate result file generation. Additionally, the system is exploring the use of housing datasets from other countries for prediction, broadening the scope of the model's applicability.This study underscores the continuous evolution and expansion of the real estate price prediction system, emphasizing a user-friendly approach and leveraging additional data sources for more accurate predictions.

IX. CONCLUSION

This paper introduces a real estate price prediction system leveraging the Decision Tree machine learning algorithm. The unique aspect of this model lies in its inclusion of additional features such as air quality and crime rate in the dataset, aiming to enhance the accuracy of predicting potential selling prices for real estate properties. These features, often overlooked in traditional prediction systems, play a crucial role in influencing individuals' decisions when purchasing a property.The Decision Tree algorithm is employed to construct a predictive model that incorporates these distinctive features. The inclusion of air quality and crime rate sets this system apart, addressing aspects that significantly impact people's decisions in property investment.The trained model is seamlessly integrated with a User Interface using the Flask Framework, ensuring a user-friendly experience. The system boasts an impressive accuracy rate of 89% in predicting real estate prices, offering a reliable tool for individuals seeking to make informed decisions within their budget and minimize potential losses.Moving forward, the paper envisions the addition of further features like tax to differentiate this model from others in the prediction system landscape. This expansion aims to provide users with a comprehensive and unique set of features for more accurate and personalized predictions.

This study highlights the significance of incorporating diverse features into real estate prediction models and emphasizes the system's distinctive qualities.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." 2011 National Conference on Innovations in Emerging Technology. IEEE, 2011

[2] Manjula, R., et al. "Real estate value prediction using multivariate regression models." Materials Science and Engineering Conference Series. Vol. 263. No. 4. 2017.

[3] A. Varma et al., “House Price Prediction Using Machine Learning And Neural Networks,” 2018 Second International

Conference on Inventive Communication and Computational Technologies, pp. 1936–1939, 1936.
[4] Arietta, Sean M., et al. "City forensics: Using visual elements to predict non-visual city attributes." IEEE transactions on visualization and computer graphics20.12 (2014): 2624-2633.
[5] Yu, H., and J. Wu. "Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5." (2016).
[6] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." 2017 International Conference on Applied System Innovation (ICASI). IEEE, 2017.
[7] NiharBhagat, AnkitMohokar, Shreyash Mane "House Price Forecasting using Data Mining" International Journal of Computer Applications,2016.
[8] CH. Raga Madhuri, G. Anuradha, M. VaniPujitha -" House Price Prediction UsingRegression Techniques: A Comparative Study" 2019 in (ICSSS),IEEE

ABOUT THE AUTHOR

Vivekanand Gupta

B.Tech CSE (AKTU)

(B.N. College Of Engineering and Technology)

Abhishek Yadav

B.Tech CSE (AKTU)

(B.N. College Of Engineering and Technology)

Vineeta Nishad

B.Tech CSE (AKTU)

(B.N. College Of Engineering and Technology)