HOUSE PRICE PREDICTION

USING MACHINE LEARNING

Harshita

*Assistant Professor, Department of Computer Science and Engineering*

*Inderprastha Engineering College*

Ghaziabad, India

Adarsh Kr. Singh

*Student, Department of Computer Science and Engineering*

*Inderprastha Engineering College*

Ghaziabad, India

Ayush Katiyar

*Student, Department of Computer Science and Engineering*

*Inderprastha Engineering College*

Ghaziabad, India

Chinmay Kansal

*Student, Department of Computer Science and Engineering*

*Inderprastha Engineering College*

Ghaziabad, India

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract---The House Price Prediction System is an innovative solution designed to estimate the market value of residential properties using advanced machine learning techniques. This system integrates multiple data sources, including historical property prices, location demographics, property features, and economic indicators, to generate accurate predictions.**  **At its core, the system employs algorithms like regression analysis, decision trees, or neural networks to analyze large datasets and capture complex relationships between variables. The incorporation of data preprocessing techniques, such as feature scaling and handling missing values, ensures the reliability and robustness of the predictions. Geographic Information System (GIS) integration provides insights into location-based trends, highlighting the impact of neighborhood characteristics on house prices.**  **End-users, including buyers, sellers, and real estate agents, interact with a user-friendly interface to input specific property details and view predictions instantly. The system offers visualization tools to display market trends and comparisons, making it a valuable tool for strategic decision-making in real estate transactions.**  **By leveraging predictive analytics, the platform enhances transparency and reduces uncertainties in property valuation. Furthermore, its adaptability allows periodic updates to models, ensuring that predictions align with changing market dynamics.**  **With scalability and real-time capabilities, the House Price Prediction System is a transformative step towards modernizing property valuation, offering benefits to stakeholders and fostering data-driven real estate practices.**  I. INTRODUCTION  Over long ago, there is manually decide the price of any property. But problem is that in manually there are 25% percent error is occurred and such affect is loss of money. But now there is big change by changing the old technology. Today’s Machine Learning is trending technology. Data is the heart of Machine Learning. Nowadays the booming of AI and Machine Learning in market. All industry are move towards automation. But without data we can’t train model. Basically in Machine Learning involves building these model from previous data and by using them to predict new data. The market demand for housing is increases daily because our population is rising rapidly. In rural area there  prediction of cost of any real estate property. They are decide price manually by analysing previous data. But there 25% of error is occurred on that prediction.so there is loss of buyers as well as sellers. Hence there are many systems are developed for house price prediction. Sifei Lu, Rick Siow had proposed advance house prediction system. The main objective of this system’s was to make a model which give us a good house price prediction based on other features.  House price prediction is a significant area of research due to its implications for real estate markets, housing policies, and economic forecasting. Predicting house prices involves analyzing various factors that influence real estate values, such as location, size, amenities, and market trends. Accurate predictions can help buyers, sellers, and policymakers make informed decisions.Recent studies increasingly focus on machine learning (ML) techniques to enhance the accuracy of house price predictions. Below are five notable research papers that contribute to this field:   1. Zulkifley, N. H., & Nasir, M. H. (2020). *Machine Learning for House Price Prediction: A Review of Techniques and Applications*. This paper reviews various machine learning techniques applied to house price prediction, focusing on the effectiveness and efficiency of each method. 2. Khan, A. A., & Khan, M. N. (2021). *A Comparative Study of Machine Learning Algorithms for House Price Prediction*. In this study, the authors compare several ML algorithms, including Linear Regression, Decision Trees, and Random Forests, assessing their prediction accuracy using various datasets. 3. Cheng, Y., & Huang, X. (2019). *Factors Influencing House Prices: A Statistical Analysis*. This research examines various economic and demographic factors influencing house prices and emphasizes the importance of integrating these variables into predictive models. 4. Agus, M., & Ismail, M. (2022). *Leveraging Natural Language Processing for Real Estate Price Prediction*. This innovative study focuses on the use of NLP techniques to analyze property descriptions, demonstrating the potential for textual data to enhance prediction accuracy. 5. Gao, Y., & Zhang, F. (2023). *Deep Learning Approaches for Housing Price Prediction: A Systematic Review*. This paper systematically reviews the application of deep learning models in predicting house prices, highlighting their advantages over traditional methods and discussing future research directions.   A critical aspect of house price prediction research is the comparative analysis of different algorithms. The study by Khan and Khan emphasizes the importance of evaluating models based on accuracy and error rates. By comparing various machine learning techniques, researchers aim to identify the model that provides the best performance in predicting house prices. The House Price Index (HPI) is commonly referenced in studies as a metric for estimating changes in housing prices. Research indicates that housing prices are strongly correlated with various factors, including economic indicators, demographic trends, and geographical features. Understanding these correlations is essential for developing robust predictive models.An innovative approach discussed in the literature involves using textual descriptions of properties to enhance prediction accuracy. The work by Agus and Ismail leverages natural language processing (NLP) techniques to analyze    V. SPECIFIC RESEARCH  House price prediction is a critical area of research that plays a significant role in the real estate market, influencing economic decisions, housing policies, and appraisal accuracy. The complexity of this challenge arises from various factors that affect housing prices, including location, property features, economic conditions, and market trends. Researchers have employed a range of methodologies to enhance the accuracy of predictions, with a notable shift towards machine learning techniques in recent years.One prominent approach is the use of the House Price Index (HPI), which serves as a key indicator of changes in housing prices. Studies have identified strong correlations between housing prices and various influencing factors, such as interest rates, employment rates, and demographic trends. By utilizing machine learning algorithms, researchers can model these relationships more effectively, providing robust frameworks for forecasting future prices.Ensemble methods, such as Random Forest and Gradient Boosting, have also gained traction in house price prediction research. These methods combine multiple models to improve prediction accuracy and robustness, addressing the limitations of traditional single-model approaches. For example, XGBoost has proven to be particularly effective, outperforming other algorithms in various studies due to its ability to handle large datasets and capture complex relationships.In addition to numerical data, the integration of textual data has emerged as a key innovation in house price prediction. Researchers have begun to leverage natural language processing (NLP) techniques to analyze unstructured text data, such as property descriptions, listings, and reviews. By extracting relevant keywords and features from this textual information, predictive models can achieve greater accuracy and provide a more nuanced understanding of property values.Moreover, regression models, particularly generalized linear regression, remain a foundational approach in this field. These models combine traditional statistical techniques with modern data mining approaches, allowing researchers to analyze various data points and their relationships comprehensively. This integration of methodologies enhances the reliability of housing price predictions and provides valuable insights into market trends.The impact of external factors, such as economic downturns or global events like the COVID-19 pandemic, has prompted researchers to adapt their models continually. Studies have explored how these events influence housing prices and the effectiveness of machine learning algorithms in adjusting to changing market conditions.  Phase II: Looking for correlations  We are trying to find out some new correlation between various attribute. This correlation gives either strong positive correlation with our label or gives strong negative correlation. From pandas library use scatter\_matrix for attribute combination.  Find out some new correlations  Try to find out new attribute from collision of old attribute. For ex. By using ‘TAX’ and ‘RM’ find ‘TAXRM’ is new attribute. Our MEDV= 1.00000 and TAXRM = -0.558626 which shows that ‘TAXRM’ strongly negative correlation with ‘MEDV’  Fig 4. New attribute correlations  Hare, show the data point of new attribute ‘TAXRM’ with respect to ’MEDV’. By analysing such data we can say that it is very good relation for our model. Similarly try to find out some new combinations from old attribute.  Phase III: To fill missing attributes  There are three ways to set a missing vales in data as:  1) get rid of the messing data point. 2) Get rid of the whole attribute.3) set the value to some value (0, mean or median). Hare, can’t use the first option because we cannot drop the data point from the data. Option second is not valid. We have to use option no three for set missing attributes.  Fig 5. Before setting missing attributes  Methods   1. Data Preprocessing    * Data Cleaning: Handle missing values, remove duplicate records, and correct inconsistent formats.    * Feature Engineering: Generate new variables like price per square foot, distance to city center, and neighborhood quality indices.    * Normalization/Standardization: Scale numeric data to improve model convergence and comparability.    * Categorical Encoding: Use one-hot encoding or label encoding for categorical attributes (e.g., house type, region). 2. Exploratory Data Analysis (EDA)    * Visualization: Correlation heatmaps to study attribute relationships with price, scatter plots for distribution analysis, and bar charts for categorical breakdowns.    * Outlier Detection: Identify anomalies in data, such as extremely high/low prices, using boxplots and z-scores.    * Statistical Analysis: Use regression diagnostics to validate assumptions of linearity and independence. 3. Model Selection    * Baseline Models: Start with linear regression, decision trees, and k-nearest neighbors (KNN) for initial testing.    * Advanced Models:      + Gradient Boosting Machines (XGBoost, LightGBM).      + Random Forests for handling non-linear interactions.      + Deep learning models like Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) to capture complex data patterns.    * Ensemble Techniques: Combine multiple models (e.g., bagging or stacking) for better performance. 4. Hyperparameter Tuning    * Use grid search or Bayesian optimization to fine-tune parameters such as learning rates, tree depth, and dropout rates for neural networks. 5. Model Evaluation    * Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for accuracy.    * Cross-Validation: Perform k-fold cross-validation to ensure model generalization. 6. Implementation of Explainable AI (XAI)    * Use techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to interpret model predictions and attribute importance. 7. Deployment and Testing    * Deployment Frameworks: Use Flask or FastAPI to build RESTful APIs for model deployment.    * Testing: Ensure system robustness by testing against unseen data and edge cases.   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 graphics 20.12 (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] Nihar Bhagat, Ankit Mohokar, Shreyash Mane "House Price Forecasting using Data Mining" International Journal of Computer Applications,2016.  [8] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5.  [9] Pow, Nissan, Emil Janulewicz, and Liu Dave Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal." Course project, COMP-598, Fall/2014, McGill University (2014).  [10] Sampathkumar, V., Santhi, M. H., & Vanjinathan, J. (2015). Forecasting the land price using statistical and neural network software. Procedia Computer Science, 57, 112-121.  [11] Banerjee, Debanjan, and Suchibrota Dutta. "Predicting the housing price direction using machine learning techniques." 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, 2017. | is lack of jobs due to this public is migrating for financial purpose.so result is increasing demand of housing in cities. People who don’t know the actual price of that particular house and they suffer loss of money. In this project, the house price prediction of the house is done using different Machine Learning algorithms like Leaner Regression, Decision Tree Regression, K- Means Regression and Random Forest Regression. 80% of data form kwon dataset is used for training purpose and remaining 20% of data used for testing purpose. This work applies various techniques such as features, labels, reduction techniques and transformation techniques such as attribute combinations, set missing attributes as well as looking for new correlations. This all indicates that house price prediction is an emerging research area and it requires the knowledge of machine learning.        II. LITERATURE SURVEY    In this conference paper we have to analyse the different Machine Learning algorithms for better training Machine Learning model. Trends in housing cost show the current economic situation and as well as to directly concern with buyers and sellers. Actual cost of house is depending on so many factors. They include like no of bedrooms, number of bathrooms, and location as well.in rural area cost is low as compare to city. The house price grate with like near to highway, mall, super market, job opportunities, good educational facilities etc. Over few years ago, the real estate companies trying to predict price of property by manually. In company there is special management team is present for  descriptive data, providing a unique perspective on how qualitative factors can influence pricing. In conclusion, the literature on house price prediction systems reveals a dynamic field that integrates traditional statistical methods with advanced machine learning techniques. By continuously refining models and incorporating diverse data types, researchers aim to improve the accuracy and reliability of house price forecasts. This ongoing research is crucial for stakeholders in the real estate market, as it directly impacts investment decisions and policy formulation.  III. PROPOSED SYSTEM  In this proposed system, we focus on predicting house price using machine learning algorithms like Leaner Regression, Decision Tree, k-Means, and Random Forest. We proposed the system “House Price Prediction Using Machine Learning” we have predict the house price using multiple features. In this proposed system, we are able to train model from various features like ZN, INDUS, CHAS, RAD etc. the previous data taken and out of this 80% of data is used for training purpose and remaining 20% of data used for testing purpose. Hare, the raw data is stored in ‘.csv’ file. We are majorly used two machine learning libraries to solve these problems. The first one was ‘pandas’ and another one is ‘numpy’. The pandas used for to load ‘.csv’ file into Jupiter notebook and also used to clean the data as well as manipulate the data. Another was sklearner, which was used for real analysis and it has containing various inbuilt functions which help to solve the problem.one more library was used which is nothing but numpy. For the purpose of train-test splitting numpy was used.  IV. SYSTEM DESIGN AND ARCHITECTURE  Phase I: collection of data  We are collected data for real estate from different online real estate websites and repository. In such data have features like ‘ZN’, ‘INDUS’, ‘RAD’, ‘CHAS’, ‘LSTAT’, ‘CRIM’, ‘AGE’, ‘NOX’ etc. and one label is ‘MEDV’. We must collect the data which is well structured and categorized. When we are start to solve any machine learning problem first data is must require. Dataset validity is must otherwise there is no point in analysing the data.  Phase II: Data pre-processing  In this phase, our data is clean up. There is might be missing values in our dataset. There are three ways to fill our missing values: 1) Get rid of the missing data points.2) Get rid of the whole attribute. 3) Set the value to some value (0, mean or median).  Phase III: Training the model  In this phase, data in broken down into two part: Training and Testing. There are 80% of data is used for training purpose and reaming 20% used for testing purpose. The training set include target variable. The model is trained by using various machine learning algorithms and getting the result. Out of these Random forest regressions predict batter results.  Phase IV: Testing the model  Finally, the trained model is applied to test dataset and house price predicted. The trained model is save by using ‘.joblib’.  This adaptability is crucial for maintaining accurate predictions in a dynamic real estate landscape. In conclusion, the research on house price prediction is evolving rapidly, driven by advancements in machine learning, the integration of textual data, and the application of robust statistical methods. As the real estate market continues to change, ongoing research will be vital in refining predictive models and enhancing their accuracy. The ability to forecast housing prices accurately is essential for stakeholders in the real estate sector, providing them with the information needed to make informed decisions and navigate the complexities of the market effectively.  VI. METHODOLOGY   1. Algorithms: In the process of developing this model, various machine learning algorithms were studied. The model is trained on Leaner regression, Decision tree, K-mean and Random forest algorithms. Out of this Random Forest give a highest accuracy in prediction of housing prices. The decision to choose the algorithm is depends on the dimensions and type of data is used. Random Forest is best fit for our model. 2. Random Forest Regressor: The random forest regressor observes features of an attribute and train the model by analysing given features. Random Forest regressor from the graph, attribute combination, labels including features and according to system analyses the data.   VII. IMPLEMENTATION  Phase I: Data Processing  In this phase, the missing attribute is handle by using mean value. The target is feature is drop out. By using Pandas library the operation is performed. For visualization of dataset graph use Matplotlib python function. After that try to catch some attribute combination and set the missing values. We split the data in the proportion of 80% for Training and remaining 20% use for Testing. Once data processing done, create suitable pipeline for execution of model.    Fig 3.Visualization of data graph      Fig 6. After setting missing attributes  In above data, the ‘RM’ column have total 399 data point out of 404.some data point are missing. To use value of median to set missing points. After setting missing point ‘RM’ column has all total 404 data points are fulfil. After that, creating a pipeline for the execution. For this purpose from sklearn import pipeline.  Phase IV; Fitting the model  From the Sklearn library, a Random forest regressor is used to train a model. The predict function use to predict results and model is save by using ‘.joblib’.    VIII. MATERIALS AND METHODS  Materials   1. Datasets    * Kaggle Housing Prices Dataset: Offers extensive data on housing attributes (e.g., size, location, and condition) along with their corresponding prices.    * Zillow Real Estate Data: Contains up-to-date real estate market data, including pricing trends, transaction records, and property details.    * UCI Machine Learning Repository - Housing Dataset: A smaller dataset suitable for testing baseline models.    * Additional Sources: OpenStreetMap for geographic data, government land records, and Census Bureau statistics for demographic information. 2. Software and Tools    * Programming Languages: Python (NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn), R (for statistical modeling).    * Development Platforms: Jupyter Notebook, Google Colab.    * Libraries for ML and DL: TensorFlow, Keras, PyTorch for neural network models; XGBoost and LightGBM for boosting algorithms.    * Visualization Tools: Tableau, Power BI for exploring and visualizing data trends.    * Version Control: GitHub or Git for collaboration and tracking changes. 3. Hardware Requirements    * Processing Units: High-performance CPUs or GPUs (e.g., NVIDIA RTX series) for training complex models.    * Memory: Minimum of 16GB RAM for efficient model processing and data handling.    * Storage: SSDs for faster read/write speeds during data handling and model training.   IX. RESULTS  To use various machine learning algorithms for solving this problem. Out of that the Random forest is predict better accuracy than other models.   |  |  |  | | --- | --- | --- | | Final RMSE = 2.9131988953 | Mean | Standard Deviation | | Leaner  Regression | 4.221894675 | 0.752030492 | | Decision Tree | 4.189504504 | 0.848096620 | | K-Means | 21.91834139 | 2.115566025 | | Random Forest | 3.494650261 | 0.762041223 |     Table 1. Models Outputs  X. CONCLUSION  In conclusion, we have successfully developed a machine learning web solution to predict house prices based on various features. The solution involves collecting and cleaning data, building and training a linear regression model. Moreover, we have incorporated hyperparameter tuning to optimize the model's performance further. This improves the model's ability to predict house prices accurately, leading to better decision-making for both buyers and sellers in the real estate market. By implementing the model in a web-based solution, users can input data on a house, and the solution will provide an estimated price based on the model's predictions. This makes it easier for buyers and sellers to obtain a rough estimate of a property's value without the need for extensive research. Overall, this machine learning web solution for house price prediction provides a valuable tool for the real estate industry and can aid in making more informed decisions regarding property values.  XI. FUTURE WORK  Further exploration of data with additional features should be conducted through comprehensive feature engineering to enhance the model's predictive capabilities. It's essential to investigate advanced ensemble methods such as stacking or blending to leverage the strengths of multiple models for improved performance. Additionally, the enhancing model interpretability through techniques like feature importance analysis and SHAP values can provide insights into the factors influencing house prices. To address imbalanced data issues, consider employing sampling techniques or alternative evaluation metrics. It's crucial to develop a robust deployment strategy for the model, ensuring scalability and efficient prediction handling. Implement continuous monitoring mechanisms to track model performance over time and detect potential issues promptly. Enhance the user interface of the application to improve user experience and usability. Lastly, incorporate a feedback loop to gather user feedback and iteratively improve the model.  XII. 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 |