A Model for House Price Prediction Using Machine Learning Algorithms

Harshada S. Belsare[1], Prof. K. V. Warkar[2]
[1,2]*Department of Computer Engg, Bapurao Deshmukh College of Engineering, Sevagram. Wardha*

Abstract:-In this, we will anticipate housing prices using machine learning techniques. Every year, there is a significant growth in the price of homes, thus we felt the need for a system that can forecast future home prices. People cannot accurately estimate the price of homes due to a lack of understanding about real estate assets. As a result, we realised the need for a model that can accurately forecast home prices. Therefore, the primary goal of our research is to accurately and profitably forecast the price of a property. The outcomes of the algorithms utilised are compared in this survey, and the model with the best accuracy and lowest error rate will be adopted. When selecting a prediction method, We frequently evaluate and compare a variety of prediction methods while choosing a prediction method. Due to its reliable and probabilistic model selection process, we frequently use the linear, and random forest regression models. Our findings demonstrate that the problem-solving strategy should be successful and flexible enough to produce forecasts that can be compared to other models for predicting housing prices. We frequently suggest a house price prediction model to help a consumer determine the accurate value of a home.

Index-Terms-Data standardisation, home price forecasting, machine learning, and machine learning algorithms are some of the keywords.

I. INTRODUCTION

Housing price prediction using machine learning techniques is anexciting area for doing research and application in the area of datascience. House price prediction is an essential task in the real estate industry, and machine learning techniques have showngreat potential in solving this problem. Machine Learning (ML) is a vital aspect of day to day business and research. It successively improves the performance of computer systems by using so amny techniques and neural network models [6]. With the increasing availability of real estate data, along with the advancements inmachine learning algorithms, it has become possible to accurately predict the prices of houses in various locations using various features such as location, size, number of rooms, amenities, and more. Machine learning models can analyze large amounts of data, identify patterns, and make accurate predictions. The size of the latitude and longitude, the population, number of households, the number of bedrooms and bathrooms, and other details that may define the interior elements of the home can all have an impact on a home's price. [3]

The objective the goal of home price prediction is to create a model that can correctly forecast a house's price based on its features. Many people buy houses to stay their and real estate brokers buy it to sell them for making profit and consider it as their income source. the main thing is that each and everyone should get the house they deserve for which they are paying [7].that is no one should get cheated while purchasing a house. According to their expectations and what they are paying for the home should worth it. This model can be used by real estate agents, property buyers, and sellers to make informed decisions. Machine learning algorithms such as linear regression, random forests, and support vector machines are commonly used in house price prediction. the goal of this project is to build a machine-learning model that can predict the prices of houses based on various features such as total rooms, population, number of rooms, latitude etc.

After addressing missing values, encoding category variables, and scaling numerical features, we will pre-process the data. Next, we will select an appropriate machine learning algorithm,such as linear regression, support vector machine, or random forests, and train our model on the pre-processed data. We will use techniques such as cross-validation to ensure that our model generalizes well and does not overfit the predicted house prices using a dataset The goal is to create a model that can correctly forecast housing prices and pinpoint the essential factors that affect those values. Finally, we will evaluate the performance of our model. If our model performs well, we may employ it to provide forecasts based on fresh, unforeseen data and offer insightful information to buyers, homeowners, and real estate brokers.

Supervised learning:

A sort of machine learning technique known as supervised learning uses labelled data to train the computer. Labeled data refers to data that has already been categorized or classified by humans, such as images labeled with the objects they contain, or text labeled with their corresponding categories The fundamental idea behind supervised learning is to utilise this labelled data to educate the computer how to see patterns or correlations in the data so that it can subsequently accurately predict or classify fresh, unlabeled data. In supervised learning, the computer is given a set of input features and a corresponding output label, and its goal is to learn a function that maps the inputs to the correct outputs. The process of training the computer involves adjusting the parameters of the function based on how well it performs on the labeled data, and then testing its performance on new, unseen data.

Algorithms for supervised learning that are often used include support vector machines, decision trees, logistic regression, and linear regression. Applications for these techniques include image classification, natural language processing, and predictive modelling.

Unsupervised learning:

Unsupervised learning is a type of machine learning algorithm in which the computer is trained using unlabeled data. Unlike supervised learning, there are no predefined labels or categories for the computer to learn from. Instead, the computer is tasked with finding patterns or relationships in the data on its own.

The goal of unsupervised learning is to identify hidden structure in the data, such as clusters or groups of similar data points, without any prior knowledge of what those clusters might represent. Unsupervised learning algorithms do this by analyzing the data and identifying patterns that can be used to group the data into clusters.

One common type of unsupervised learning algorithm is clustering, in which the computer groups similar data points together based on their proximity to each other. Dimensionality reduction is a different kind of unsupervised learning technique in which the computer minimises the amount of characteristics in the input while still keeping the most crucial details. Applications for unsupervised learning algorithms include customer segmentation, anomaly detection, and picture and text analysis. Principal component analysis (PCA), hierarchical clustering, k-means clustering, and t-SNE are examples of popular unsupervised learning techniques.

II.PROBLEM DEFINITION

To create a predictive model that can calculate house prices with accuracy using a variety of factors, including location, size, the number of rooms, amenities, and more. This forecast aims to assist real estate brokers, purchasers, and vendors in making knowledgeable choices regarding the acquisition, sale, and valuation of their properties.

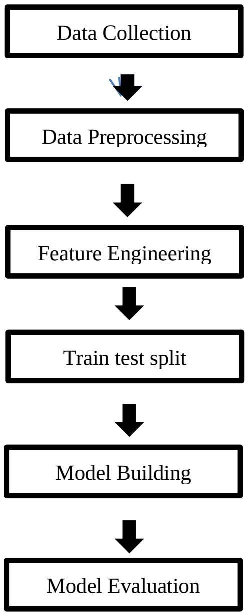
The main challenge in this problem is to develop a model that can accurately capture the complex relationships between the features and the target variable (house price) in a given real estate dataset. This requires careful selection and preprocessing of the features, choosing an appropriate machine learning algorithm, and tuning the model's hyperparameters to optimize its performance.

In addition to accuracy, the model's interpretability is also important as to which features played a significant role in the predictions. This is important for real estate agents, buyers, and sellers to understand and make informed decisions based on the predictions. It involves developing a model that is accurate, interpretable, and can provide insights into the factors that influence house price prediction techniques.

III.PROPOSED METHODOLOGY

The job of determining a residential property's worth based on several elements, such as its latitude, longitude, age, and attributes, is known as house price prediction. Both buyers and sellers need accurate home price predictions in order to make educated judgements about purchasing or selling a property.

Machine learning algorithms have gained popularity in recent years for home price prediction because they can analyse vast volumes of data and spot intricate patterns that may escape the notice of human specialists. The system's flow is seen below.



From the above diagram, we can see that how the system will be built. first, we have to collect data then it will go through cleaning and pre-processing. Then feature engineering will take place to remove the outlier, data scaling, etc. the data will be split into training and testing and then finally a model will be built and evaluated.

Collection: Gathering information on factors like location, size, number of bedrooms and baths, amenities, etc. that affect home pricing is the first step. Several sources, including internet real estate directories, public documents, and real estate brokers, are available for this data.

The quality and quantity of data that you collect can significantly affect the accuracy and reliability of your machine-learning model. Here are some tips for collecting data for house price prediction:

Determine the variables: The variables that are most relevant to house price prediction include longitude, population, number of bedrooms. Determine which variables you want to include in your analysis and then collect data for those variables.

Data sources: There are several sources you can use to collect data for house price prediction. These include real estate websites. You can also collect data from local real estate agents and property management companies.

Data cleaning: Once you have collected your data, you will need to clean it to remove any missing or inconsistent values. You may also need to transform or normalize the data to make it more suitable for analysis.

Data labeling: To train a supervised machine learning model, you will need to label your data. In the case of house price prediction, you will need to label each data point with the actual price of the property.

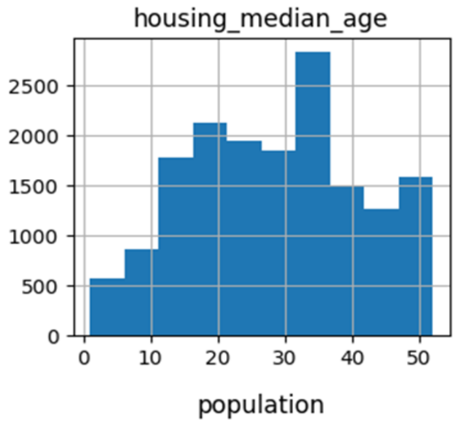
Data augmentation: You can enhance the quality and quantity of your data by augmenting it.

Data splitting: After collecting and cleaning your data, you will need to split it into training, validation, and testing sets. This will enable you to train your machine learning model on a subset of the data and then test its accuracy on a separate subset of data.

Data cleaning and preprocessing: Once the data is collected, it needs to be cleaned and preprocessed to remove any missing or irrelevant data, perform feature engineering, and transform the data into a format suitable for machine learning algorithms. The collected data will likely contain missing values, outliers, and other errors that need to be dealt with before modeling. Therefore preprocessing of the data to remove any errors or inconsistencies.

Data visualization: Data visualization is an important aspect of machine learning (ML) because it helps to better understand and interpret data, which is a critical step in building accurate and effective ML models. By creating visual representations of data, it becomes easier to identify patterns, trends, outliers, and relationships within the data, which can then be used to inform the development of ML models.

There are various types of data visualizations that can be used in ML, including histograms, heat maps, and box plots, among others. Each type of visualization can be used to highlight different aspects of the data, depending on the research question or problem being addressed. Data visualization can also be used to evaluate the performance of ML models.



Normal distribution

The normal distribution is a probability distribution that describes the probability of a continuous random variable taking on a range of values. The normal distribution is defined by two parameters: the mean (μ) and the standard deviation (σ).

The shape of the normal distribution is symmetric around the mean, with the highest probability density at the mean. The standard deviation determines the spread of the distribution, with a larger standard deviation resulting in a wider and flatter distribution. The normal distribution is widely used in statistics, scientific research, and engineering to model and analyze various phenomena, such as the heights of individuals in a population, the test scores of students, and the measurement errors in scientific experiments.

In the above graph we can see a normal distribution in the feature named housing age from our dataset.

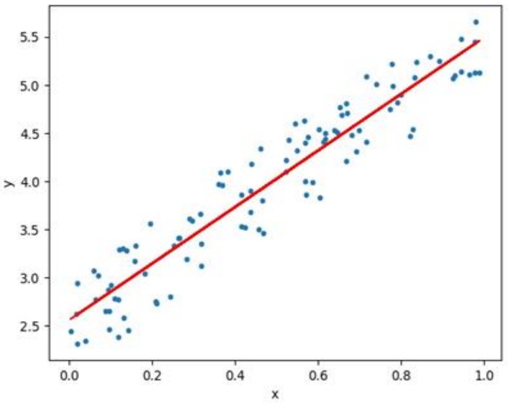
Feature engineering: This step involves transforming the data into a format that can be used by the machine learning algorithm. This includes selecting the relevant features, scaling and normalizing the data, and encoding categorical variables.it also includes finding the correlation between the. variables that are strongly and weakly related.

Split data into training and testing sets: The data is split into two sets: a training set and a testing set. The training set is used to train the machine learning algorithm, and the testing set is used to evaluate the performance of the algorithm.

Choose a machine learning algorithm: There are several machine learning algorithms that can be used for house price prediction, including linear regression, decision trees, and random forests. The choice of algorithm depends on the specific problem and the size of the dataset. we will use Linear regression in this.

Linear Regression:

The connection between a dependent variable (commonly referred to as the target or response variable) and one or more independent variables (often referred to as predictors or features) is modelled using the popular machine learning approach known as linear regression. The price of a home would be the dependent variable in the context of predicting house prices, and the independent factors may include things like the number of bedrooms, the size of the property, the neighbourhood, and so on. Both classification and regression tasks in machine learning require linear regression.



Linear regression

The objective is to find the best-fit line that minimizes the sum of the squared differences between the predicted and actual values of the dependent variable.

The equation for simple linear regression, where there is only one independent variable, is:

$$Y = \beta0 + \beta1X + \varepsilon$$

where Y is the dependent variable, X is the independent variable, $\beta0$ is the intercept, $\beta1$ is the slope, and ε is the error term. The objective of the linear regression is to estimate the values of $\beta0$ and $\beta1$ that minimize the sum of squared errors.

Random Forest:

A common ensemble learning approach for

classification, regression, and other problems in machine learning is called Random Forest. It is a member of the family of decision tree-based models, which combine the predictions of numerous decision trees that have been trained on different subsets of training data.

The technique builds a number of decision trees, each of which is generated using a subset of the input characteristics and the training data that is randomly chosen. Combining all of the decision trees' projections yields the ultimate conclusion.

The main advantages of using Random Forest are:

It can handle a large number of input features and noisy data.

It can perform well on both classification and regression tasks.

It is less prone to overfitting compared to other decision tree-based algorithms.

It provides feature importance measures, which can be useful in feature selection.

To build a Random Forest model, the following steps are typically taken:

Data preprocessing: The input data is preprocessed by handling missing values, encoding categorical variables, and scaling the data if necessary.

Splitting the data: The input data is split into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance.

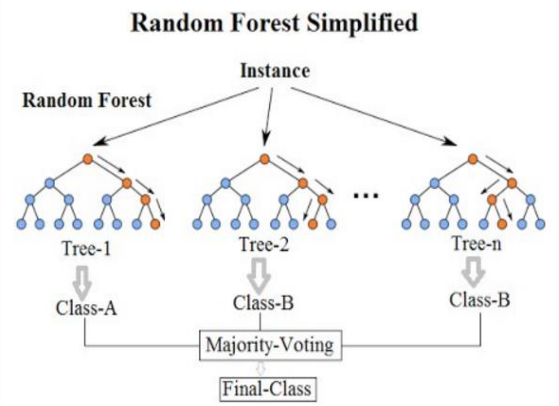
Creating decision trees: A fixed number of decision trees are created, each trained on a randomly selected subset of the input data.

Growing decision trees: Each decision tree is grown by recursively splitting the data into smaller subsets, based on the most informative features, until a stopping criterion is met.

Making predictions: The final prediction is made by aggregating the predictions of all the decision trees. For classification tasks, the most common prediction is selected, and for regression tasks, the average prediction is used.

Evaluating the model: The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1 score.

Tuning the model: The model hyperparameters can be tuned to improve its performance on the validation set. Common hyperparameters to tune include the number of decision trees, the depth of each tree, and the size of the subset of features to consider at each split.



Random Forest models are commonly used for a variety of machine learning tasks such as classification, regression, and feature importance analysis. They are known for their ability to handle noisy data and prevent overfitting. However, they can be computationally expensive, and their performance can be affected by the choice of hyperparameters.

Evaluation: After the model is trained, we need to evaluate its performance using various metrics such as mean squared error, R-squared, and accuracy. The model can be further fine-tuned using hyper parameter tuning techniques to improve its performance.

IV. SYSTEM REQUIREMENTS

The software requirements include the programming languages and libraries needed to implement machinelearning algorithms. Python is a popular language used for machine learning, and libraries such as scikit-learn, pandas, and numpy are commonly used for data preprocessing, feature selection, and model training

Hardware Requirements:

Computer:

The hardware requirements depend on the size of the dataset and the complexity of the machine learning algorithm used. A powerful computer with a multi-core processor and a dedicated GPU is recommended for training large datasets and complex models. Software Requirements:

Python:
Python is a high-level, interpreted programming language. It is widely used in web development, scientific computing, data analysis, artificial intelligence, and many other areas. One of the main advantages of Python is its large and active community of developers, who contribute to the development of libraries and tools that make it easier to work with Python. Most popular libraries include NumPy, Pandas, SciPy, Matplotlib, and Tensor Flow.

Google colaboratory:
A Jupyter notebook environment hosted in the cloud is offered by Google under the name Colab. It is a free platform that enables users to create and execute Python code in a web browser without having to install any software on their computer.

Numpy :
"Numerical Python" is the name of a Python library. It is an essential Python module for scientific computing, supporting big, multi-dimensional arrays and matrices as well as a sizable number of high-level mathematical operations on these arrays. Data science, machine learning, scientific research, and other domains that need for effective numerical computations on huge datasets frequently employ NumPy. Mathematical operations: NumPy comes with a sizable library of mathematical operations that may be used with arrays. These operations range from simple ones like addition and subtraction to more intricate ones like matrix multiplication and trigonometric operations.Linear algebra: NumPy provides a suite of functions for linear algebra operations, such as matrix multiplication, matrix inversion, and solving linear equations. Overall, NumPy is a powerful library that is essential for many scientific computing and data analysis tasks in Python. Its efficient array processing capabilities and extensive mathematical functions.

Pandas:
A well-liked Python package for data analysis and manipulation is called Pandas. For processing structured data, especially data frames, which resemble database tables, it offers strong data structures. Its ability to handle structured data and perform complex data transformations makes it an ideal tool for data preprocessing and feature engineering. Its ability to integrate with other data analysis and visualization libraries makes it a popular choice for machine learning projects.it is usedin machine learning for data preprocessing, feature engineering, data visualization, data analysis etc.0

Seaborn :
Built on top of Matplotlib, Seaborn is a well-known Python module for data visualisation. Heatmaps, scatterplots, line charts, bar charts, and m-plots are just a few examples of the statistical visuals that can be produced using its high-level interface. Seaborn is particularly useful for creating complex visualizations with minimal code. Oveerall, Seaborn is a powerful library for creating complex visualizations with minimal code. Its integration with Pandas and built-in statistical visualizations make it a popular choice for data visualization in Python.

Matplotlib:
It is a well-known Python data visualisation package that offers several capabilities for producing static, animated, and interactive visualisations. Multiple plot types, interactive plots, publication-quality plots, and interaction with pandas are some of Matplotlib's standout features.
Overall, it is a powerful library for creating static, animated, and interactive visualizations in Python. Its customization options and support for multiple plot types make it a versatile tool for data visualization and exploration. Its integration with Pandas and support for multiple backends make it a popular choice for data analysis and machine learning projects.

CONCLUSION

House price prediction using machine learning techniques is a challenging task, but it can be achieved with good accuracy by employing appropriate models and data preprocessing machine learning models can be used for house price prediction. These models can be trained on various features related to the house, such as longitude, latitude, number of rooms, population, and so on.

Feature selection and feature engineering can also play a critical role in improving the accuracy of the model. In conclusion, house price prediction using machine learning is a promising field that can benefit real estate agents, buyers, and sellers. However, the accuracy of the model heavily depends on the quality and quantity of data, the selection of appropriate features, and the choice of a suitable machine-learning algorithm. Therefore, it is essential to have a thorough understanding of the problem, data, and models to achieve accurate predictions.

REFERENCE

[1] Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, DR. Vinayk A. haradi (2021)," House Price Prediction Using Machine Learning" Finolex Academy of Management and Technology, Mumbai University.

[2] Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla, Manav Rachna, (2020) "Prediction of House Pricing Using Machine Learning with Python" International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)

[3] Data Imran, Umar Zaman, Muhammad Waqar and Atif Zaman ‘‘Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing" Soft computing and machine intelligence journal (2021)

[4] Yong Piao, Ansheng Chen, Zhendong Shang, ‘‘Housing Price Prediction Based on CNN" 2019 9th International Conference on Information Science and Technology (ICIST).

[5] Pei-Ying Wang, Chiao-Ting Chen, Jain-Wun Su, Ting- Yun Wang, and Szu-Hao Huang, (2021) ‘‘Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism".

[6] Smith Dabreo, Shaleel Rodrigues, Valiant Rodrigues, Parshvi Shah, "Real Estate Price Prediction" International Journal of Engineering Research and Technology (2021)..

[7] Abigail Bola Adetunjia , Oluwatobi Noah Akande , Funmilola Alaba Ajala , Ololade Oyewo , Yetunde Faith Akande , Gbenle Oluwadara, ‘‘House Price Prediction using Random Forest Machine Learning Technique "The 8th International Conference on Information Technology and Quantitative Management Procedia Computer Science 2022.

[8] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei "Housing Price Prediction via Improved Machine Learning Techniques" 2019 International Conference on Identification Information and Knowledge in the Internet of Things (IIKI2019)

[9] Imran, Umar Zaman, Muhammad Waqar and Atif Zawan "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data" (Soft Computing and Machine Intelligence Journal, Vol (1), Issue (1), 2021

[10] Keren He, Cuiwei He, "Housing price analysis using linear regression and logistic regression: A comprehensive explanation using Melbourne real estate data" 2021 IEEE International conference on Computing (ICOCO)

[11] Bandar Almaslukh, "A gradient boosting method for effective prediction of housing price in complex real estate systems" 2020 International conference on technologies and applications of artificial intelligence (TAAI)

[12] Maryam Heidari, Samira Zad, Setareh Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision"2021 IEEE

[13] Karshiev Asanjar, Olimov Bekhzod, Jaesoo Kim, Anand Paul and Jeonghang Kim "Missing data imputation of geolocation-based price prediction using KNN-MCF Method", international journal of geo-information (2020).

[14] Kassahun Abebe, Pallavi V Patil, "Housing price forecasting using machine learning algorithm" IJARET (2021)

[15] Xiangqin Cheri, "Optimizations of training dataset on house price estimation"2021 2nd International Conference on big data economy and international management (BDEIM).

[16] Ping-Feng Pai and Wen-Chang Wang "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices" (2020) Department of Information Management, National Chi Nan University, 1 University Rd., Puli, Nantou 54561, Taiwan

[17] Debanjan Banerjee, Suchibrota Dutta "Predicting the Housing Price Direction using Machine Learning Techniques" IEEE

International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017)

[18] Atharva chouthai Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari, Vijay Kukre “House Price Prediction using Machine Learning” International Research Journal of Engineering and Technology (IRJET).