PANDIT DEENDAYAL ENERGY UNIVERSITY SCHOOL OF TECHNOLOGY

Course: Machine Learning
PROJECT REPORT

B.Tech.

(Computer Science and Engineering) Semester 7

Submitted To:

Dr. Himanshu Gajera

Submitted By:

Shreya Panengaden
20BCP046

G2 batch

Introduction

House price prediction is a critical task in the real estate industry and has garnered considerable attention in the field of machine learning and data science. The prediction of real estate property prices is of paramount importance in the housing market, as it empowers buyers, sellers, and investors with valuable insights for informed decision-making. In this study, we employ machine learning techniques to develop a robust and accurate house price prediction model. The primary objective is to provide a reliable tool for estimating property values, thereby aiding stakeholders in making well-informed choices.

Linear Regression provides a fundamental baseline model for predicting house prices, assuming a linear relationship between the independent variables and the target variable. Decision Tree Regression, on the other hand, leverages a non-linear approach, building a tree-like structure to capture complex interactions among features. Finally, Random Forest Regression combines multiple decision trees to enhance predictive accuracy and reduce overfitting. The project evaluates the performance of each regression model in terms of metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), to assess predictive accuracy and model generalization.

The developed house price prediction model is assessed for its predictive accuracy and robustness through comprehensive testing on a separate dataset. The results demonstrate the model's ability to provide accurate property price estimates, with promising real-world applications in real estate valuation, investment, and decision support.

In conclusion, it showcases the potential of machine learning in house price prediction, enabling stakeholders in the housing market to make informed choices and investments. The model's high accuracy, reliability, and adaptability make it a valuable tool for addressing the dynamic and complex nature of property valuation. Future work may involve the integration of additional data sources and the development of a user-friendly interface for wider adoption in the real estate industry.

Dataset

Boston housing dataset UCI ML repository

Link : https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set

Sources:

- Origin: The dataset was sourced from the StatLib library, which is maintained at Carnegie Mellon University.
- Creator: The dataset was created by Harrison, D. and Rubinfeld, D.L. for their research on "Hedonic prices and the demand for clean air," published in the Journal of Environmental Economics & Management in 1978.
- Date: The dataset was documented on July 7, 1993.

Attributes: There are 13 continuous attributes, along with one binary-valued attribute, for a total of 14 attributes. The "MEDV" attribute serves as the class attribute.

- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS: Proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if the tract bounds the river; 0 otherwise)
- NOX: Nitric oxides concentration (parts per 10 million)
- RM: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distances to five Boston employment centers
- RAD: Index of accessibility to radial highways
- TAX: Full-value property-tax rate per $10,000
- PTRATIO: Pupil-teacher ratio by town
- B: 1000 times (Bk - 0.63)^2, where Bk is the proportion of Black residents by town
- LSTAT: Percentage of lower status population
- MEDV: Median value of owner-occupied homes in $1000s (the target variable)

The dataset contains no missing attribute values.

Methodology

Data Collection

We load the dataset and study some properties of the dataset. We also try to visualize the data. We also look for correlations.

Data Preprocessing

We check for missing values and handle them through imputation techniques, such as median replacement. Standardize numerical features to ensure all attributes have the same scale. We create a pipeline by performing these operations

Data Splitting

Split the dataset into training and testing sets to facilitate model evaluation. Typically, use a 80% training set and 20% testing set. We perform stratified split on CHAS column to maintain the same proportion of target classes or categories as the original dataset.

Model Training

As houseprice prediction involves prediction of continuous values based on a curve we choose three regression models for house price prediction: Linear Regression, Decision Tree Regression, and Random Forest Regression. We train the models and obtain predictions for each model.

Model Evaluation

Evaluate the models on the validation dataset using appropriate metrics. Calculate Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to assess predictive accuracy. Identify the best-performing model based on validation results.

Model Tuning

Fine-tune the hyperparameters of the models to optimize their performance. For example, adjust the learning rate, regularization parameters, and feature subsets for Linear Regression.

Tune the maximum depth, minimum samples per leaf, and other hyperparameters for Decision Tree and Random Forest models.

Model Testing

Assess the selected model on the testing dataset to obtain a final evaluation of its predictive accuracy and generalization. Finally, we evaluate this model using root mean square error and save the model.

Model Use

Develop We use the model to predict for new data by entering all important features

Result

After training each model, the root mean square error, we obtained are shown in the table. The lowest error was obtained in Random Forest Regression and hence we used it for testing.

Model	Root Mean Square Error
Linear Regression	4.945
Decision Tree Regression	2.8162
Random Forest Regression	1.3823

Conclusion and Future Work

Our project on house price prediction using Random Forest Regression has yielded exceptionally promising results. The low error rates, as indicated by the metrics, emphasize the model's accuracy in estimating property values. The success of this project underscores the effectiveness of Random Forest Regression in addressing the complex and dynamic nature of the real estate market. This model can be a valuable tool for buyers, sellers, and investors, providing them with informed insights for decision-making in the housing sector. Additionally, the robustness and adaptability of Random Forest Regression make it well-suited for real-world applications, such as real estate valuation and investment.

As we look to the future, the integration of additional data sources and the development of a user-friendly interface can further enhance the model's utility and accessibility in the real estate industry.