HOUSE PRICE PREDICTION REPORT

PROJECT TEAM-ID - PTID-CDS-DEC-24-2145
PROJECT-ID - PRCP-1020

INTRODUCTION

House price prediction is essential in real estate, helping buyers, sellers, and investors make informed decisions. Prices depend on factors like location, economic conditions, interest rates, and property attributes. Accurate predictions support pricing strategies, investment planning, and market analysis.

This project develops a predictive model using machine learning to estimate house prices based on key features such as size, location, and amenities. The approach includes data preprocessing, feature selection, and training **Random Forest** and **Gradient Boosting** models to determine the best-performing one. The final model offers reliable price estimates, benefiting real estate agents, homebuyers, and investors with data-driven insights.

DATASET DESCRIPTION

The dataset used in this project consists of house-related features and their corresponding prices. The key attributes include:

- **Size** (Square footage of the house)
- **Location** (Geographical position)
- **Number of Bedrooms and Bathrooms**
- **Year Built** (Age of the property)
- **Additional Features** (Garage, swimming pool, garden, etc.)

The dataset is divided into training and testing sets for model evaluation.

CHALLENGES FACED

- **Missing Data:** Some features required imputation.
- **Outliers:** Extreme price variations influenced predictions.
- **Feature Correlation:** Some variables were highly correlated.
- **Non-linearity:** Complex relationships among features affected model accuracy.

TECHNIQUES USED

- **Data Preprocessing:** Handling missing values, encoding categorical variables, and feature scaling.
- **Feature Selection:** Identifying key predictors using SelectKBest with f_regression.
- **Machine Learning Models:** Training **Random Forest Regressor** and **Gradient Boosting Regressor**.
- **Hyperparameter Tuning:** Optimizing models using **RandomizedSearchCV**.

SYSTEMATIC METHODS USED

- **Importing Libraries**
- **Loading and Exploring the Data**
- **Handling Missing Values**
- **Encoding Categorical Variables**
- **Feature Scaling and Selection**
- **Model Selection and Training**
- **Hyperparameter Tuning**
- **Model Evaluation and Performance Comparison**

DATA PREPARATION

- Dropped columns with excessive missing values.
- Imputed numerical missing values with the median.
- Imputed categorical missing values with the mode.
- Encoded categorical variables using **OneHotEncoder**.
- Scaled numerical features using **StandardScaler**.
- Selected the most significant features using **SelectKBest**.

MODEL CREATION

Random Forest Regressor

An ensemble learning model that combines multiple decision trees to enhance accuracy and reduce overfitting. Random Forest improves generalization by averaging predictions from multiple trees. Hyperparameter tuning was performed using **RandomizedSearchCV** to optimize depth, number of trees, and feature selection.

Gradient Boosting Regressor

A sequential ensemble model that corrects the errors of weak learners to improve prediction accuracy. Gradient Boosting optimizes model performance by training on residual errors. This model is powerful but requires careful tuning to avoid overfitting. **RandomizedSearchCV** was used to optimize learning rate, number of estimators, and tree depth.

MODEL EVALUATION

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **R-squared (R^2) Score**

Performance Comparison:

Model	**MAE**	**MSE**	**RMSE**	**R^2 Score**
Random Forest	15,200	6.3e8	25,100	88%
Gradient Boosting	13,900	5.7e8	23,800	91%

The **Gradient Boosting Regressor** demonstrated the best overall performance.

COMPARISON

Model	**Accuracy**	**Training Time**	**Scalability**
Random Forest	Very Good	Moderate	High
Gradient Boosting	Excellent	Slow	Medium

RESULT

The **Gradient Boosting Regressor** is the most suitable model for house price prediction due to:

- **3% improvement in accuracy** over Random Forest.
- **Lower error rates across all evaluation metrics**.
- **Better handling of complex relationships in the data**.

CONCLUSION

This project successfully developed a predictive model for house prices using **Random Forest** and **Gradient Boosting**. The **Gradient Boosting Regressor** provided the most accurate predictions, making it the recommended model for deployment. Future improvements may include:

- Incorporating additional real estate market factors.
- Exploring deep learning approaches for further accuracy improvements.
- Enhancing interpretability using SHAP values or feature importance analysis.

With a **3% accuracy improvement** over Random Forest, this model can be deployed as a web application for real-time house price estimation.