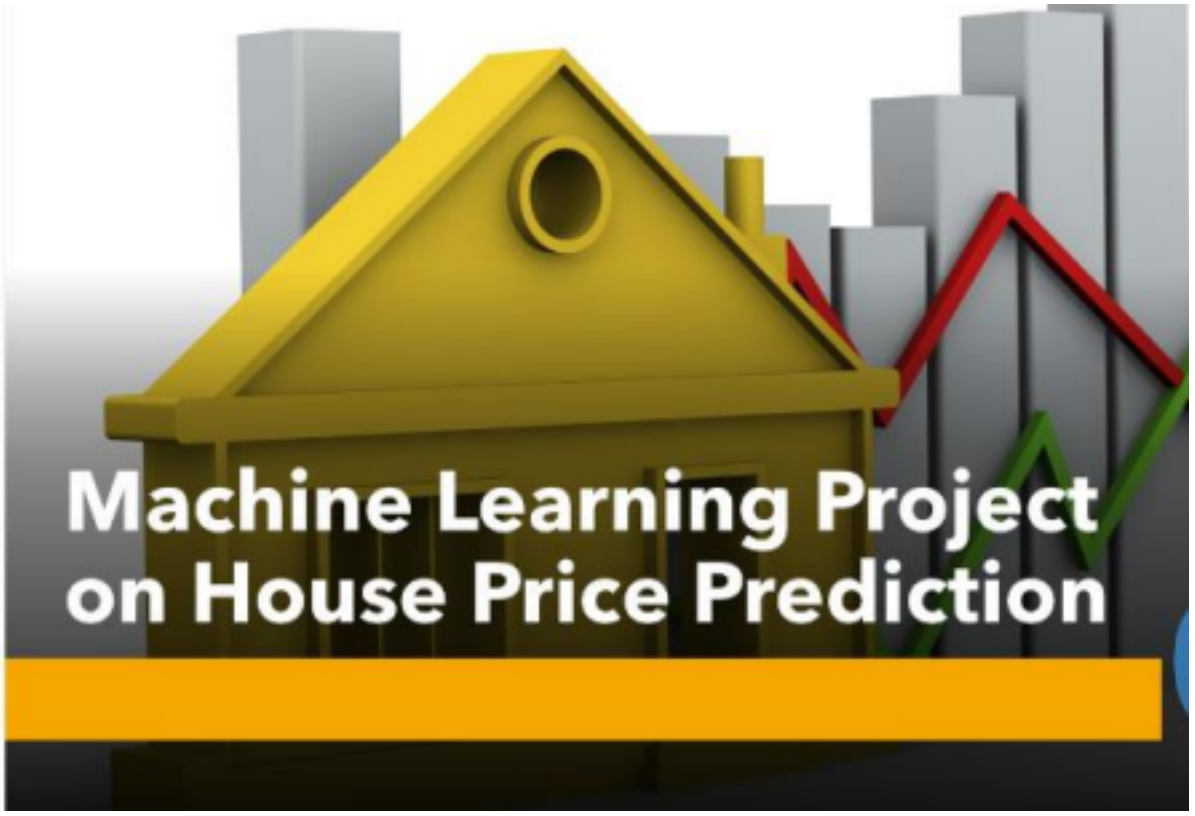
Predicting House Prices using Machine Learning

Team member: D.V.AKASH

Reg no:211521243006

Artificial Intelligence Phase-1 Document



Problem Statement:

The problem is to predict house prices using machine learning techniques. The objective is to develop a model that accurately predicts the prices of houses based on a set of features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

Predicting House Prices using

Machine Learning

Phase-1 : Problem Definition and Design Thinking

Project definition:

- Predict the selling price of residential properties based on various features such as square footage, number of bedrooms, number of bathrooms, location, and other relevant factors.
- Forecast future housing market prices for a specific location or region over a certain time period, taking into account historical price data and potentially economic indicators.
- Predict house prices using a combination of structured data (e.g., property features) and unstructured data (e.g., images of the houses, textual descriptions) for a more comprehensive model.
- Instead of predicting exact prices, classify properties into price brackets (e.g., low, medium, high) based on their characteristics.
- Predict house prices while considering spatial dependencies and geographical factors, such as proximity to schools, parks, transportation, and crime rates.

Project Goals:

Main goal for this model follows:

- The primary goal is to develop models that accurately estimate the selling or listing prices of houses. This ensures that buyers and sellers have a reliable reference point for making informed decisions.
- Customize price predictions for individual buyers or sellers based on their specific preferences, needs, and budget constraints.
- Assist sellers in setting competitive prices for their properties to maximize their chances of selling quickly and at a desirable price. • For real estate investors, the goal may be to manage a portfolio of properties effectively by predicting their future values and optimizing the allocation of resources.

Predicting House Prices using Machine Learning

Design Thinking:

1). Data Source:

- **: Websites like Zillow, Redfin,**

Real Estate Listings Websites

Realtor.com, and MLS (Multiple Listing Service) provide extensive data on property listings, including property details, prices, and location information.

- **: Historical data on property sales in the target**

Historical Sales Data

area can be a valuable source for training machine learning models. This data can include sale prices, transaction dates, and property characteristics.

- **: Economic data, such as inflation rates,**

Economic Indicators

unemployment rates, and interest rates, can be used to analyze the overall economic health of an area and its impact on housing prices.

- **: Detailed information about the**

Property Characteristics Data

physical characteristics of properties, including square footage, number of bedrooms, number of bathrooms, lot size, and building age, is crucial for modeling house prices.

- **: Data on neighborhood quality,**

Neighborhood and School Ratings

crime rates, school ratings, and amenities (e.g., grocery stores, restaurants) can influence property prices and are valuable for analysis

Dataset link :https://www.kaggle.com/datasets/vedavyasv/usa housing

Predicting House Prices using Machine Learning

2). Data preprocessing:

Data Cleaning	:

- Handle Missing Values: Identify and address missing values in the dataset. Common strategies include imputation (replacing missing values with a suitable estimate) or removal of rows or columns with excessive missing data.
- Outlier Detection and Treatment: Identify outliers in the data that may skew the model's predictions. Depending on the situation, you can either remove outliers or transform them to mitigate their impact.
- Data Validation: Check for data consistency and correctness, such as unrealistic values or data entry errors, and resolve any issues.

Feature Engineering	:

- Feature Selection: Choose the most relevant features (attributes) that are likely to have a significant impact on house prices. Eliminate irrelevant or redundant features to reduce model complexity.
- Feature Scaling: Standardize or normalize numerical features to bring them to a common scale, which helps gradient-based algorithms converge faster and avoids certain biases.
- Encoding Categorical Variables: Convert categorical variables (e.g., property type, neighborhood) into numerical representations using techniques like one-hot encoding or label encoding.

Data Transformation:

		• Log Transformations: Apply logarithmic transformations to skewed numerical features to make their distributions more normal, which can improve model performance. • Box-Cox Transformations: Use the Box-Cox transformation to stabilize variance and make the data conform more closely to a normal distribution. • Handling Date and Time Data: Extract relevant information from date and time features, such as year, month, day of the week, or time since a specific event.
	Data Splitting	:
		• Split the dataset into training, validation, and test sets to evaluate the model's performance accurately. Common splits include 70-30 or 80-20 for training and testing, with a separate validation set for hyperparameter tuning.
	Handling Skewed Target Variable	:
		• If the target variable (house prices) is significantly skewed, you may apply a transformation to make it more symmetric (e.g., log transformation) before modeling.

Predicting House Prices using Machine Learning

3).FEATURE SELECTION:

Correlation analysis:

Calculate the correlation between each feature and the target variable (house prices). Features with higher absolute correlation coefficients are typically more relevant. You can use metrics like Pearson's correlation coefficient for numerical features and point-biserial correlation for binary features.

Recurssive feature elimination:

Use RFE with a machine learning model (e.g., linear regression) to recursively

remove the least important features based on model performance. This method iteratively prunes features until the desired number is reached.

:

L1 Regularization (Lasso Regression)

L1 regularization encourages sparse feature selection by penalizing the absolute values of feature coefficients. Features with non-zero coefficients after applying L1 regularization are considered important.

:

Mutual Information

Mutual information measures the dependency between two variables. Calculate the mutual information between each feature and the target variable and select the features with the highest scores.

:

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that can help reduce the number of features while preserving as much variance as possible. It may be useful when dealing with a large number of features.

Predicting House Prices using Machine Learning

4)**.Model Selection:**

There are several machine learning models that can be used for house price

prediction in regression tasks. The choice of model depends on the characteristics of your dataset and the performance you aim to achieve. Here are some commonly used models for house price prediction:

❖ :

Linear Regression

Linear regression is a simple and interpretable model that assumes a linear relationship between the independent variables (features) and the target variable (house price).

❖ **and :**

Ridge Regression

Lasso Regression

These are regularized linear regression models that can help prevent overfitting by adding penalty terms to the linear regression equation.

❖ :

Random Forest

Random Forest is an ensemble method that combines multiple decision trees to reduce overfitting and improve predictive accuracy. ❖ :

Support Vector Machines (SVM)

SVM regression aims to find a hyperplane that best fits the data, and it can be effective for house price prediction when you have a relatively small dataset.

Neural Networks

❖ :

Deep learning models, such as feedforward neural networks and

convolutional neural networks (CNNs), can be applied to house price prediction tasks for capturing complex patterns in the data. However, they may require large amounts of data and computational resources. ❖ :

Time Series Models

If your dataset includes temporal information, you may consider time series models like ARIMA (AutoRegressive Integrated Moving Average) or LSTM (Long Short-Term Memory) networks for capturing time dependent patterns in house prices.

Predicting House Prices using Machine Learning

5).Model Training:

Data Splitting

➢ :

Split your preprocessed dataset into two or three subsets: a training set, a validation set (optional), and a test set. The training set is used to train the model, the validation set is used for hyperparameter tuning (if necessary), and the test set is used for evaluating the final model's performance.

➢ :

Model Selection

Choose the machine learning model you want to use for house price prediction based on your analysis and requirements. For example, you might choose Linear Regression, Random Forest, or a Gradient Boosting algorithm.

➢ :

Model

Training

Train the selected model using the training dataset. This involves fitting the model to the preprocessed feature data (independent variables) and the corresponding target variable (house prices). The specific steps for training depend on the chosen model. Typically, you'll use a function or method provided by your machine learning library (e.g., scikit-learn in Python) to train the model.

➢ :

Model Evaluation

After training the model, evaluate its performance using the test dataset. Common regression metrics for house price prediction include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score. The choice of metrics depends on your specific goals and the nature of the problem. ➢ :

Documentation

Document the entire process, including data preprocessing steps, model selection, training, hyperparameter tuning, and evaluation. This documentation is essential for reproducibility and future reference.

Predicting House Prices using Machine Learning 6).Evaluation:

Mean Absolute Error (MAE)

:

- **MAE measures the average absolute difference between the predicted house prices and the actual prices. It provides a straightforward measure of prediction accuracy without considering the direction of errors.**
- **Formula: MAE = (1 / n) ∑ |actual - predicted|.**
- **Lower MAE values indicate better model performance.**

Mean Squared Error (MSE)

:

- **MSE measures the average squared difference between the predicted house prices and the actual prices. It penalizes larger errors more heavily than MAE, making it more sensitive to outliers.**
- **Formula: MSE = (1 / n) ∑ (actual - predicted)^2.**
- **Lower MSE values indicate better model performance.**

Root Mean Squared Error (RMSE)

:

- **RMSE is the square root of MSE and provides an interpretation of the average prediction error in the same unit as the target variable.**
- **Formula: RMSE = √MSE**
- **Lower RMSE values indicate better model performance.**

R-squared (R2) Score

:

- **R2 measures the proportion of the variance in the target variable (house prices) that is explained by the model. It ranges from 0 to 1, where 1 indicates a perfect fit, and 0 indicates that the model does not explain any variance.**
- **Formula: R2 = 1 - (SSE / SST), where SSE is the sum of squared errors, and SST is the total sum of squares.**
- **Higher R2 values indicate better model performance, with values close to 1 suggesting a good fit.**

Predicting House Prices using Machine Learning

Conclusion:

House price prediction in machine learning is a valuable

application with significant real-world implications. It involves using various predictive models and regression techniques to estimate the prices of residential properties based on relevant features and historical data.