Project 9: Predicting House Prices using Machine Learning

Project Title: House Price Predictor

Problem Statement:

The housing market is an important and complex sector that impacts people's lives in many ways. For many individuals and families, buying a house is one of the biggest investments they will make in their lifetime. Therefore, it is essential to accurately predict the prices of houses so that buyers and sellers can make informed decisions. This project aims to use machine learning techniques to predict house prices based on various features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors.

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

Project Steps

Phase 1: Problem Definition and Design Thinking

Problem Definition:

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.

Design Thinking:

1.Data Source:

Choose a dataset containing information about houses, including features like Income, House Age, Number of Rooms, Number of Bedrooms, Population, Price and Address.

Features/Attributes: These are variables or characteristics of a property that are believed to have an impact on its price. Common features includes

- Number of Rooms and Bedrooms: The count of rooms and bedrooms in the house, which can affect its price due to the increased area and size.
- Address: The geographical location of the property, which can include factors like neighbourhood, city, and proximity to amenities or landmarks.
- House Age: The age of the house or building can also influence its price.
- Income: The income earned by the people living in those areas can also influence its price.
- Population: The number of people who are currently living in those areas can also influence the price of the houses.
- Price: The total cost of the house can be displayed here based on certain factors like area, number of rooms, age etc..

Target Variable: This is the variable you're trying to predict, which is the price of the house or property. It is typically represented as the dependent variable in a regression analysis.

Dataset Size: The number of records or data points in the dataset, which can range from a few hundred to thousands or more.

Data Sources: Information about where the data was collected from, such as real estate listings, government records, or surveys.

2.Data preprocessing :

It is a crucial step in preparing a house price prediction dataset for analysis or machine learning. Here's a short overview of the key steps involved:

Data Cleaning:

Handle missing values by filling them with appropriate values (e.g., mean, median, or mode).Remove duplicates if they exist in the dataset.Correct any inconsistent or erroneous data entries.

Feature Selection:

Choose relevant features that are likely to impact house prices.Remove irrelevant or redundant features to simplify the dataset. Consider using domain knowledge and feature importance techniques.

Feature Encoding:

Convert categorical variables into numerical representations through techniques like one-hot encoding or label encoding. Standardize or normalize numerical features to have a consistent scale (e.g., using Min-Max scaling or z-score normalization).

Outlier Detection and Handling:

Identify and handle outliers in the data, either by removing them or transforming them. Use visualization and statistical methods (e.g., Z-scores or IQR) to detect outliers.

Data Splitting:

Split the dataset into training, validation, and test sets to assess model performance effectively.A common split ratio is 70-80% training, 10-15% validation, and 10-15% testing.

Handling Skewed Data:

If the target variable (house prices) or some features are highly skewed, consider applying transformations like log transformations to make the data more symmetric.

Scaling and Normalization:

Ensure that numerical features are scaled or normalized to have similar scales, preventing some features from dominating others during modeling.

Data Transformation (if needed):

For some modeling algorithms, you might need to transform the data to meet their assumptions (e.g., transforming the target variable for linear regression).

Data preprocessing ensures that the dataset is clean, well-structured, and ready for analysis or modeling. The specific steps may vary depending on the dataset and the machine learning algorithm you plan to use, but these general steps provide a solid foundation for preparing data for house price prediction tasks.

3.Feature selection:

Feature selection for a house price prediction dataset involves choosing the most relevant and informative features while excluding irrelevant ones to improve the accuracy of your prediction model. Here's a concise guide to feature selection:

Correlation Analysis: Identify features that have a strong correlation with the target variable (house price) using techniques like Pearson correlation. Keep features with high correlations and eliminate those with low or negative correlations.

Feature Importance: If you're using tree-based models (e.g., Random Forest, XGBoost), use their feature importance scores to rank and select the most important features.

Cross-Validation: Use cross-validation to evaluate different feature subsets' performance and select the one that yields the best model performance metrics.

Regularization Hyperparameters: When using models like Ridge or Elastic Net, tune their regularization hyperparameters to encourage feature selection during training.

4.Model Selection:

Choose a suitable regression algorithm (e.g., Linear Regression, Random Forest Regressor) for predicting house prices.

Linear Regression: Start with a basic linear regression model, which is simple and interpretable. It's a good baseline.

Decision Trees: Try decision tree-based models like Random Forest and Gradient Boosting. They often perform well and handle non-linear relationships.

Neural Networks: Experiment with deep learning models, such as neural networks, if you have a large dataset and complex features.

5. Model Training:

- Train the selected model using the preprocessed data.
- Train the selected model on the training data using the features to predict house prices.
- The model will learn the relationships between the features and target variable.

6. Evaluation:

Evaluate the model's performance using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

- **MAE:** It represents the average absolute error in house price predictions. Lower MAE values indicate better model accuracy. It's easy to understand and suitable for comparing models.

- **RMSE:** RMSE penalizes larger errors more than MAE because of the squaring. It's also sensitive to outliers. Lower RMSE values indicate better model accuracy.

- **R-squared**: R-squared measures the goodness of fit. It tells you the proportion of variance in the target variable that is explained by the model. Higher R-squared values (closer to 1) indicate a better fit.

These steps provide a simplified overview of the model training process for house price prediction. The specific implementation details and choice of algorithms may vary depending on the dataset and the goals of the project.