**House Price Prediction**

**Problem Statement:**

The problem statement is to predict house prices using machine learning techniques. Our goal is to create a model that can accurately estimate the prices of houses based on a set of key features, including location, square footage, number of bedrooms, number of bathrooms, and other relevant factors. This project encompasses data preprocessing, feature engineering, model selection, model training, and model evaluation.

**Design Thinking:**

**1. Data Source Selection:**

To tackle this problem, our first step is to source a dataset containing comprehensive information about houses. This dataset should include essential features such as location, square footage, bedrooms, bathrooms, and their corresponding prices. We will choose a dataset that is reliable, up-to-date, and representative of the problem we aim to solve. Potential data sources include real estate databases, public datasets, or web scraping if necessary.

**2. Data Preprocessing:**

Data preprocessing is a vital stage in ensuring the quality and reliability of our dataset. We should perform the following tasks:

**Data Cleaning**: It is the process which handles any issues related to missing or erroneous data. This may involve handling missing values, removing outliers, or correcting inaccuracies.

**Data Transformation**: Many machine learning algorithms require numerical data as input. Therefore, we will convert categorical features into numerical representations. Techniques like one-hot encoding or label encoding will be employed based on the nature of the categorical data.

**3. Feature Selection:**

Selecting the most relevant features which impact the price of the house for predicting house prices is a crucial step. To achieve this, we will use various techniques:

**Feature Correlation Analysis**: We will measure the correlation between each feature and the target variable i.e house price. Features with strong correlations will be prioritized.

**Feature Importance Scores**: For machine learning models that provide feature importance scores, we will use these scores to identify the most informative features.

By focusing on the most relevant features, we can improve model efficiency and interpretability.

**4. Model Selection:**

The choice of a regression model is important for accurate price prediction. We have several options to consider:

**Linear Regression**: This is a straightforward model that is easy to interpret and can serve as a baseline.

**Random Forest Regressor**: A more complex ensemble model that can capture non-linear relationships in the data.

The selection of the model will depend on factors such as dataset size, complexity, and the level of interpretability required.

**5. Model Training:**

Once we have selected the most suitable regression model, we will move on to the model training phase. This involves the following steps:

**Data Split**: We will divide the dataset into training and testing sets to measure the model performance.

**Model Fitting**: We will train the chosen model on the training data. This process includes parameter tuning to optimize model performance.

**6. Evaluation:**

To determine the effectiveness of our predictive model, we will employ relevant regression metrics such as:

**Mean Absolute Error (MAE)**: This measures the average absolute errors in our predictions.

**Root Mean Squared Error (RMSE)**: This assesses the square root of the average squared errors, giving more weight to larger errors.

**R-squared (R^2)**: This metric evaluates how well the model explains the variance in the data.

**Summary:**

By following the design thinking steps detailed above, we intend to construct a robust machine learning model capable of accurately predicting house prices based on relevant features. This model will provide valuable insights for real estate market participants, enabling informed decision-making.