

Problem Statement 1: Predicting Housing Prices Using Linear Regression

1. Introduction:

In real estate markets, predicting housing prices accurately is crucial for buyers, sellers, and real estate agents. Housing prices depend on various factors such as the size of the house, the number of bedrooms, location, age of the property, and other amenities. In this problem, we aim to develop a machine learning model using linear regression to predict housing prices based on these features.

2. Objective:

The primary objective of this project is to build a linear regression model that can predict the price of a house given its features. The model will be trained on a dataset of housing prices and evaluated based on its accuracy and performance.

3. Dataset:

For this problem, we will use the California Housing Prices dataset. This dataset contains information about various houses in California and includes features such as the number of rooms, population, median income, house age, and location coordinates.

4. Dataset Details:

Link to Dataset: California Housing Prices

(<https://www.kaggle.com/datasets/camnugent/california-housing-prices>)

Features:

longitude: The longitudinal coordinate of the house.

latitude: The latitudinal coordinate of the house.

housing_median_age: The median age of the house.

total_rooms: The total number of rooms in the house.

total_bedrooms: The total number of bedrooms in the house.

population: The population in the neighborhood.

households: The number of households in the neighborhood.

median_income: The median income of the households in the neighborhood.

median_house_value: The median value of the house (target variable).

5. Methodology:

Data Preprocessing:

- Handle missing values.
- Feature scaling and normalization.
- Encoding categorical variables if necessary.

Exploratory Data Analysis (EDA):

- Visualize the distribution of the target variable.
- Analyze the relationship between features and the target variable.
- Identify and handle outliers.

Model Development:

- Split the data into training and testing sets.
- Train a linear regression model on the training data.
- Evaluate the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) on the testing data.

Model Evaluation and Tuning:

- Perform cross-validation to ensure the model's robustness.
- Tune hyperparameters to improve model performance.

Model Deployment:

- Save the trained model for future predictions.
- Develop a simple user interface (UI) for users to input features and get housing price predictions.

6. Tools and Technologies:

- Programming Language: Python
- Libraries:
 - Data Analysis: pandas, numpy
 - Visualization: matplotlib, seaborn
 - Machine Learning: scikit-learn
- Development Environment: Jupyter Notebook or any other preferred IDE

7. Expected Outcome:

A linear regression model capable of predicting housing prices with reasonable accuracy. The model's performance will be evaluated based on the chosen metrics, and it will be fine-tuned to achieve the best possible results. Additionally, a simple UI will be developed to demonstrate the model's predictions in a user-friendly manner.

8. Conclusion:

This project will provide a practical application of linear regression in predicting housing prices, demonstrating the process from data preprocessing to model deployment. The insights gained from this project can be extended to more complex models and other real-world applications in the field of machine learning.