Assignment 1: Linear Regression Analysis on USA Housing Dataset

DSCI 6601: Practical Machine Learning

Introduction

In this assignment, we will analyze the USA Housing dataset using linear regression. The objective is to build a linear regression model, evaluate its performance, and explore how the features influence the predictions. You will perform the tasks listed below and include your Python code, comments explaining each step, and the corresponding outputs for each question.
The dataset can be downloaded from the following link: `https://www.kaggle.com/datasets/vedavyasv/usa-housing`.

Assignment Breakdown

1 Model Fitting and Evaluation

- **Task 1: Data Splitting and Model Fitting**
  Start by splitting the dataset into training (80%) and testing (20%) sets. Fit a linear regression model using the training data.

- **Task 2: Reporting Coefficients and Model Evaluation**
  After fitting the model, report the coefficients of each feature and the intercept. Then, compute the R^2 score using the test data to evaluate the model's performance. Explain what the R^2 score represents in the context of linear regression.

2 Prediction and Comparison

- **Task 3: Predictions on Sample Data**
  Randomly select 20 samples (rows) in random from the test set. Using the fitted model, make predictions for these inputs. Compare the predicted values to the actual values and present your results in a table.

3 Feature Importance and Model Adjustment

- **Task 4: Feature Ranking and Model Refinement**
  Based on the model's coefficients, rank the features by importance. Drop the least important feature and re-fit the model using the remaining features. Recalculate the R^2 score for the new model and compare it to the original R^2. Discuss whether dropping the feature had a significant impact on the model's performance.

Report Submission

Your submitted report should include the following:

- A brief explanation of the linear regression model.
- Detailed answers to each of the tasks mentioned above, along with the Python code and output.
- Clear explanations for all the steps taken and an analysis of the results.