Introduction to Data Science

CA4 - Report

Mohammad Yaser Azad | 810800003

Parham Javan | 810800008

Summary:

This project revolves around the application of regression analysis to predict the progression of diabetes using the Diabetes dataset from scikit-learn. Its main purpose is to provide a practical exploration of regression modeling techniques, loss functions, and model evaluation metrics in the context of medical data analysis.

Purpose:

The project is structured to cover essential steps in regression analysis:

1. Data Preparation: The dataset is loaded, preprocessed, and split into training and testing sets.
2. Function Implementation: Key loss functions (MSE, MAE, RMSE, R^2 Score) are implemented from scratch to understand their mathematical underpinnings.
3. Model Building and Training: A linear regression model is constructed and trained using the diabetes dataset to predict disease progression.
4. Model Evaluation: The performance of the trained model is evaluated using various metrics such as MSE, MAE, RMSE, and R^2 score to assess its accuracy and effectiveness.
5. OLS Analysis: Ordinary Least Squares (OLS) regression is performed to further analyze the model and extract statistical insights.
6. Discussion and Analysis: Questions are posed to encourage critical thinking and interpretation of the results, including the significance of metrics like R^2 and p-values, and the importance of individual features in predicting diabetic condition.

Question1

1. Analyzing and evaluating the values in Table (1):

- To analyze the values in Table (1), you would look at the performance metrics (MSE, MAE, RMSE, R^2 score) calculated for both the training and testing datasets.
- MSE, MAE, and RMSE measure the error between the predicted values and the actual values. Lower values of these metrics indicate better model performance.
- R^2 score, also known as the coefficient of determination, represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 score indicates a better fit of the model to the data.

Question2

2. Reviewing R^2 and Adjusted R^2 values:

- R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher R^2 values (closer to 1) indicate that a larger proportion of the variance is explained by the model, suggesting a better fit.
- Adjusted R^2 adjusts the R^2 value based on the number of predictors in the model. It penalizes complexity, so it generally decreases if irrelevant predictors are added to the model. Adjusted R^2 can be more informative when comparing models with different numbers of predictors.
- Differences between R^2 and Adjusted R^2 arise because Adjusted R^2 penalizes overly complex models. Adjusted R^2 will be lower than R^2 if the model includes irrelevant predictors.

Question3

3. Reviewing p-values:

- P-values obtained from statistical tests (such as those in OLS regression) indicate the probability of observing the data, given that the null hypothesis is true (i.e., the coefficient for the predictor variable is zero).
- Lower p-values suggest stronger evidence against the null hypothesis, indicating that the predictor variable is likely to be related to the response variable.
- A commonly used threshold for significance is 0.05. If the p-value is less than 0.05, the result is considered statistically significant.
- Columns with p-values less than 0.05 are typically considered to have significant effects on the response variable, while those above 0.05 may not be statistically significant.

Question4

4. Assessing the importance of each feature:

- The importance of each feature can be assessed based on various factors such as its coefficient value, p-value, and practical significance.
- Features with higher absolute coefficient values and lower p-values are generally considered more important in predicting the outcome variable.
- Additionally, domain knowledge and medical relevance can help prioritize features. For example, in the context of diabetes prediction, features related to blood sugar levels or body mass index may carry higher importance than others.
- By analyzing the results obtained in part 4, you can identify which features have the most significant impact on predicting an individual's diabetic condition.