BIKE SHARING CASE STUDY

SUBJECTIVE QUESTION PRESENTATION

Assignment-based Subjective Questions

Question 1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

•There is a strong correlation between the season and months in the categorical columns.

•This correlation likely influences the dependent variable, suggesting seasonal trends or patterns in the data.

Question 2 - Why is it important to use drop_first=True during dummy variable creation?

Answer:

- Using drop_first=True during dummy variable creation is important to avoid the issue of multicollinearity in regression models, particularly in linear regression.
- It removes one category from each categorical variable, reducing redundancy and improving model stability.

Question 3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

- **Temperature** has the strongest correlation with the target variable among the numerical variables.
- This indicates that temperature is a key factor in predicting the dependent variable.

Question 4 - How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

•Ensuring these assumptions hold helps improve the accuracy and reliability of your linear regression model, leading to better interpretations and predictions:

- **Linearity:** Checked if the relationship between predictors and the target is linear.
- **Independence:** Ensured observations were independent.
- **Homoscedasticity:** Verified that the residuals had constant variance.
- **Normality:** Confirmed that residuals were normally distributed.

Question 5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

- **Temperature**, **casuals**, and **registered** are the three main contributors significantly towards explaining the demand for shared bikes.
- These features explain most of the variability in bike demand.

General Subjective Questions

Question 1 - Explain the Linear Regression Algorithm in Detail (4 Marks)

1.Definition:

1. Linear regression models the relationship between a dependent variable yyy and one or more independent variables x1,x2,…,xnx_1, x_2, \dots, x_nx1,x2,…,xn.
2. **Objective:** Predict yyy by finding the best-fit line.

2.Model Equation:

1. **Simple Linear Regression:** y=β0+β1x+ϵy = \beta_0 + \beta_1x + \epsilony=β0+β1x+ϵ
2. **Multiple Linear Regression:** y=β0+β1x1+⋯+βnxn+ϵy = \beta_0 + \beta_1x_1 + \dots + \beta_nx_n + \epsilony=β0+β1x1+⋯+βnxn+ϵ

3.Assumptions:

1. Linearity, Independence, Homoscedasticity, Normality of errors.

4.Method:

1. Use **Ordinary Least Squares (OLS)** to minimize the Mean Squared Error (MSE) and determine coefficients β0,β1,…,βn\beta_0, \beta_1, \dots, \beta_nβ0,β1,…,βn.

Question 2 - Explain Anscombe's Quartet in Detail (3 Marks)

1.Definition:

1. A collection of four datasets with identical statistical properties (mean, variance, correlation), but different distributions.

2.Purpose:

1. Demonstrates the importance of visualizing data before analyzing it.

3.Key Insight:

1. Despite having similar statistical summaries, each dataset behaves differently when plotted, highlighting the risk of relying solely on summary statistics.

Question 3 - What is Pearson's R? (3 Marks)

1.Definition:

1. Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables.

2.Range:

1. Values range from -1 to 1:
    1. **+1:** Perfect positive correlation.
    2. **0:** No correlation.
    3. **-1:** Perfect negative correlation.

3.Use:

1. Helps in understanding the strength and direction of the linear relationship between variables.

Question 4 - What is Scaling? Why is Scaling Performed? What is the Difference Between Normalized Scaling and Standardized Scaling? (3 Marks)

1.Definition:

1. Scaling is the process of adjusting the range of data features.

2.Why Performed:

1. Ensures that features contribute equally to the model, especially in algorithms sensitive to feature magnitudes (e.g., SVM, KNN).

3.Types:

1. **Normalized Scaling:** Rescales data to a [0, 1] range.
2. **Standardized Scaling:** Centers data around the mean and scales to unit variance (z-score).

Question 5 - You Might Have Observed That Sometimes the Value of VIF is Infinite. Why Does This Happen? (3 Marks)

1. Definition:

1. VIF (Variance Inflation Factor) quantifies the extent of multicollinearity in a regression model.

2. Reason for Infinity:

1. Occurs when there is perfect multicollinearity, meaning one predictor is an exact linear combination of others.

3. Implication:

1. Indicates redundant predictors, causing issues in estimating regression coefficients.

Question 6 - What is a Q-Q Plot? Explain the Use and Importance of a Q-Q Plot in Linear Regression. (3 Marks)

1.Definition:

1. Q-Q plot (Quantile-Quantile plot) compares the distribution of data to a theoretical distribution (e.g., normal distribution).

2.Use in Linear Regression:

1. Assesses if the residuals (errors) follow a normal distribution, a key assumption in linear regression.

3.Importance:

1. Helps identify deviations from normality, such as skewness or kurtosis, guiding model diagnostics and improvements.