Assignment-based Subjective Questions

****

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** (3 marks)

Answer:

From the analysis of the categorical variables from the dataset, we can infer the following insights-

1. Fall has highest demand for rental bikes

2. Demand is growing each month till June. September month has highest demand. After September, demand is decreasing.

3. Weekday is not giving clear picture about demand but at Saturdays demand is slightly high.

4. The clear weather shit has highest demand

5. When there is a holiday, demand has decreased.

6. Demand for year 2019 year has grown

****

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

Answer:
drop_first = True, helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. It helps to avoid multicollinearity regression models. This improves the overall stability and interpretability of the model.

****

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:
'temp' variable has the highest correlation with the target variable.

****

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
(3marks)

Answer:
I have validated the assumption of Linear Regression Model based on following points–

1. Error terms should be normally distributed
2. There should be insignificant multicollinearity among variables.
3. Linearity should be visible among variables

4. There should be no visible pattern in residual values.

****

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Answer:
The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows-

- temp
- Winter
- Sep

****

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the difference between the predicted and actual values of the dependent variable. Here's a detailed explanation of the linear regression algorithm:

Mathematically the relationship can be represented with the help of following equation

$$Y = b0 + b1 \cdot X1 + b2 \cdot X2 + \cdots + bn \cdot Xn$$

Here, Y is the dependent variable we are trying to predict.

X1, X2, X3, …, X4 are the independent variable we are using to make predictions.

b0 is the y-intercept

b1, b2, …, bn are the coefficients (slope) representing the relationship between each independent variable and the dependent variable.

The goal of linear regression is to find the values of the coefficients (β) that minimize the sum of squared differences between the predicted and actual values of the output variable. This process is known as "fitting" the model to the data.

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model −

- Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- Relationship between variables – Linear regression model assumes that the relationship

between response and feature variables must be linear.

- Normality of error terms – Error terms should be normally distributed
- Homoscedasticity – There should be no visible pattern in residual values.

Types of Linear Regression:

- Simple Linear Regression: One independent variable.
- Multiple Linear Regressions: Multiple independent variables.

Regularization Techniques:

Lasso Regression (L1 regularization): Adds a penalty term to the absolute values of the coefficients.

Ridge Regression (L2 regularization): Adds a penalty term to the square of the coefficients.

Linear regression is widely used due to its simplicity and interpretability. However, it is important to ensure that the assumptions hold true for reliable results. If assumptions are violated, adjustments or alternative models might be necessary.

****

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. This collection of datasets was created to emphasize the importance of graphical exploration and visualization in understanding the underlying patterns and characteristics of data. The quartet is designed to highlight the limitations of relying solely on summary statistics without visualizing the data.

The four datasets in Anscombe's quartet share the following common characteristics:

1. Each dataset consists of 11 data points.
2. The mean and variance of both the x and y variables are almost identical.
3. The correlation coefficient between x and y is approximately the same for all datasets.

In practice, Anscombe's quartet is often used to illustrate the concept that exploring data graphically is essential for a comprehensive understanding of its behavior and to avoid making decisions based solely on summary statistics.

****

3. What is Pearson's R? (3 marks)

Answer:

Pearson's correlation coefficient, denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative It ranges from -1 to 1, where:

r=1: Perfect positive linear correlation

r=−1: Perfect negative linear correlation

r=0: No linear correlation

****

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. The key difference between normalized scalingand standardized scaling are:

- Range: Normalized scaling adjusts the range to a specific interval, while standardized scaling centers the data around a mean of 0 and scales by the standard deviation.
- Sensitivity to Outliers: Standardized scaling is generally more robust to outliers compared to normalize scaling, which can be influenced by extreme values.
- Use Cases: Normalized scaling is often preferred when the algorithm expects input features in a specific range. Standardized scaling is suitable when dealing with data that follows an approximately normal distribution.

The choice between normalized and standardized scaling depends on the characteristics of the data and the requirements of the machine learning algorithm being used.

****

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

Answer:
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. So we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

****

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:
A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset adheres to a specific theoretical distribution, such as the normal distribution. It compares observed quantiles to expected quantiles, with points ideally forming a straight line for a good fit. In linear regression, Q-Q plots are crucial for checking the normality assumption of residuals, identifying outliers, and assessing model adequacy. Systematic deviations in the plot may suggest model improvements. Overall, Q-Q plots play a key role in validating assumptions and enhancing the robustness of linear regression models.

****