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:
   Final equation is calculated as

   *count = 0.2164 + (-0.0838) * Dec + (-0.0601) * Feb + (-0.0753) * Jan + (-0.0851) * Nov + (0.0839) * Sep + (-0.0260) * Monday + (0.0643) * Saturday + (-0.2002) * Light_Rain_snow + (-0.0530) * Misty + (0.0675) * Summer + (0.1630) * Winter + (0.2420) * year + (0.0515) * workingday + (0.5544) * atemp + (-0.2020) * humidity + (-0.1755) * windspeed*

   ## Temperature

   It can be observed that temperature plays a vital role as most rides are booked during higher temperature. A thorough knowledge and weekly forecast can yield better results.

   ## Season

   As per data count is impacted more during Winters and Summers. Business should focus more during these seasons to drive maximum value.

   ## Month

   September is highly preferred month by riders to share rides. Therefore, pushing marketing efforts during Aug to Oct can better business.

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

   Answer:
   Dummy variable creation must be N – 1, where N represents number of unique values in a column. Therefore, it not only reduces an extra column but also correlation effects.

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

   Answer:
   Temperature had the maximum correlation with count.

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

   Answer:
   We've tested derived model against following LR assumptions:
   A. Linearity
   B. Homoscedansity
   C. Multicollinearity
   D. Residual Independence

E. Error Normality

All tests were positive and as per expectations.

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:

1. Temperature
2. Season
3. Month

General Subjective Questions

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

Answer:

Linear Regression (LR) Model helps to drive relationship between a Dependent variable on one or more Independent variables. Independent variables can be Categorical or Continuous however, Dependent variable will always be Continuous. The multiple regression model is based on the following assumptions:

1. There is a linear relationship between the dependent variables and the independent variables
2. The independent variables are not too highly correlated with each other
3. Observations are selected independently and randomly from the population
4. Residuals should be normally distributed with a mean of 0

LR consists of following steps:

A. EDA and data cleaning.
B. Analysis and conversion of categorical variables.
C. Dividing dataset into Train and Test
D. Rescaling train data using either normalization or Standardization
E. Building the model using train data by either of following:
    a. Top-Down
    b. Bottom-Up
    c. RFE
F. Testing LR principal against derived model.
G. Predicting values against test data
H. Undergoing Residual Analysis

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

Answer:

Anscombe's quartet is a set of four different datasets, which carry identical statistical description. However, these were created to enunciate value of graphical representation of data. This is owing to the reason that statistical inferences – Mean, Median, Variance – might not help to point out anomalies and hence data visualization plays a pivotal role, to achieve that.

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

Answer:

Pearson's correlation coefficient helps to measure association between two continuous variables. Not only it provides the magnitude but also direction of the relationship.

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

Answer:

A. Scaling is a phenomenon to normalize data within an expected range such as 0 to 1.
B. Scaling is performed to bring all variables in same level of magnitude. If not done, algorithm will take magnitude into consideration and produce incorrect modeling.
C. There are two ways of Scaling viz.
    a. Normalization: Scales the variables between 0 to 1
    b. Standardization: Scales the variables to have mean 0 and standard deviation 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

Infinite VIF is possibility when a variable can be expressed exactly by a linear relationship of other variables.

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

Answer:

A quantile-quantile plot is a visualization of the quantiles between two datasets to ascertain whether both originate from population that has common distributions.

A quantile signifies portion of points below a given value. That is, the 0.4 (or 40%) quantile is the point at which 40% percent of the data falls below and 60% fall above that value.

A 45-degree reference line is also plotted and in case two sets are developed from a similar population, points will fall along the line. Higher the distance from reference line, there is a greater chances that respective populations represent different distributions.