SUBJECTIVE QUESTIONS

06 February 2022 18:09

Assignment-based Subjective Questions

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

  - There is an increase in bike sales from 2018 to 2019
  - There are a greater number of users in Summer and Fall than in Winter and Spring
  - Fall is completely dependent on both the temperature variables

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

  - It is important to use drop_first=True or dropping the first variable during dummy variable creation because during VIF(Variance Inflation Factor) it will become infinity if we don't drop it.

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

  - Looking at the pair-plot among the numerical variables, Actual Temperature , Temperature and Fall has the highest correlation with the target variable(Total_Count) . We dropped one among actaul temperature and temperature because they both are highly co related.

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

  - The independent variables are not correlated in the final model

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

  - the top 3 features contributing significantly towards explaining the demand of the shared bikes\
    a) Temperature
    b) Year
    c) Windspeed (negative effect)
    d) Clear weather

General Subjective Questions

- Explain the linear regression algorithm in detail.

  - There are 2 types of regression they are

(i) Simple linear regression - contains one independent variable

(ii) Multiple linear regression - contains multiple independent variables
- P value should be less than 0.05
- R^2 value higher the better any values greater than 0.7 or 0.8 are considered good
- Prob F statistic value should be less than 0.05
- Adjusted R^2 value should be always less than R^2 value

➢ Explain the Anscombe's quartet in detail.

- Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit.

➢ What is Pearson's R?

- Pearson's R is a measure to determine the relationship between two quantitative variables and the degree to which the two variables coincide with one another.

➢ What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling?

- Scaling is a method used to normalize the range of independent variables
- It is used to normalize the values which are way out of range because it will be easier to form a model
- We perform scaling on the variables that are required to get them into a uniform way [0,1] in normalization or min max scaling. whereas standardization uses standard deviation.

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

- If we don't drop drop_first=True Then VIF will be infinite . This can be avoided by dropping the first variable during dummy variable creation because during VIF(Variance Inflation Factor) it will become infinity if we don't drop it.

➢ What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- A Q–Q *plot* is *used* to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness