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 –

a. Bike demand in year 2019 was higher than 2018.
b. Demand is high in Summer, winter, Sept, Weather with Mist Broken and Light Snow are also correlated variable
c. Spring is negatively correlated variable

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

Answer –

It is important to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

It is also used to reduce the collinearity between dummy variables

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

Answer –

Temperature is highest correlated variable.

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

Answer –

a. I checked linear relationship between variables and cnt
b. Dropped the Insignificant variables by checking following, here considered P value should be < 0.05 –
    a. High P, High VIF value
    b. High P, Low VIF value
    c. High VIF, Low P value
    d. VIF is > 5
    e. Low P, Low VIF
c. Checked the error distribution
d. Checked and compare the Adjusted R^2 and R^2 of train and test dataset

e. Based on the relation and values done the prediction

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 –

a. A bike-sharing provider BoomBikes can focus more on Temperature
b. Demand is high in Summer, winter, Sept, Weather with Mist Broken and Light Snow are also correlated variable. So company can focus on it
c. Its been observed that the spring season has negative coefficients and negatively correlated. So marketing team can focus on it in order to improve demand of bikes
d. We can see demand for bikes was more in 2019 than 2018

General Subjective Questions

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

Answer –

The algorithm of Linear Regression as mentioned below

1. Reading, understanding and visualizing the data - Reading, understanding and visualizing the data with numpy, pandas, shape, info, describe, check if relation is there or not. Bifurcation between categorical and continuous variables
2. Preparing the data for modeling –
    a. Convert bicategorical variables. e.g. from yes / no to 1 / 0.
    b. User dummy variables for other categorical variables
    c. Splitting dataset into test and train
    d. Rescaling of variables
3. Training the model - Add variables one by one or all at once and remove insignificant variables with the help of R^2, P value and VIF stats.
4. Residual analysis and Predictions - y_train - y_train_pred
5. Predictions and evaluation on the test - Here test the test dataset and check its R^2 it should be near about same with the train R^2

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

Answer –

Anscombe's Quartet is a group of four data sets which are near about identical in simple descriptive statistics. They have very different distributions and appear differently when plotted on scatter plots.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

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

Answer –

The Pearson correlation coefficient (r) is the common way of measuring a linear correlation.

It is a number between –1 and 1 where if r=1 means perfect positive correlation and -r means perfective negative correlation. But correlation does not prove that one of the variable caused the other. e.g. using umbrella does not mean that it will rain.

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

Answer –

1. Scaling makes interpretation easy
2. It is a step of data Pre-Processing which is applied to independent variables to normalize the data.
3. Scaling just affects the coefficients and no change in the other parameters like t-statistic, F-statistic, p-values, R-squared, etc
4. Most of the times the collected data set contains features highly varying in units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
5. Differences between Normalization and Standardization scaling –
    a. Normalization/Min-Max Scaling:
        i. This method scales the model using minimum and maximum values.
        ii. When the feature distribution is unclear, it is helpful.
        iii. Values on the scale fall between [0, 1] and [-1, 1].
        iv. Formula - Normalization Scaling: x=x−min(x)max(x)−min(x)
    b. Standardization Scaling:
        i. Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one
        ii. When the feature distribution is consistent, it is helpful.
        iii. Values on a scale are not constrained to a particular range.
        iv. Formula - x=x−mean(x) / sd(x)

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 can be infinity. In case of perfect correlation R^2=1 and VIF = 1/(1-R^2) is infinity. To solve this issue we may need to drop one variable from dataset.

If VIF = infinite indicates corresponding variables may be expressed exactly by the linear combination 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 –

Quantile-Quantile (Q-Q) plot is graphical tool to show the 2 datasets comes from same distribution. if train and test data set are different, in this case we can confirm using Q-Q plot if both datasets are from populations with same distribution.