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?

Answer-:

- From the analysis of the categorical variable various inferences from the data could be looked upon, some of the following are,
- Most of the people prefer a clear weather to rent bikes whereas the cnt of people renting bikes on bad weather day is much less.
- Most of the people rent bikes on holidays.
- Monday is the day where less number of bikes are rented as compared to other days of the week whereas Sunday being the day where most of the bikes are rented.
- People do not prefer to rent more bikes in the season of spring which means that during spring the people usually don’t tend to rent the bikes more.

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

Answer –

- Drop_first = True is very useful during dummy variable creation because the number of rows created by the dummy variable is reduced significantly.
- For eg- If we have 10 columns with 3 categories each it would mean that the create dummies will create 30 columns whereas by using drop_first = True the number of columns created would significantly go down from 30 to 20. Hence we use the drop_first=True to lessen the number of columns in the dataframe making it more simple.

3) Looking at the pair-plot among the numerical variables, which one has the highest

correlation with the target variable?

Answer-

The ‘temp’ column has the highest correlation with the target variable ‘cnt’.

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

Answer-

- After building the model on the training set we carry the residual analysis on the predicted values on the model. After plotting a distribution plot of the difference between the actual and predicted terms , if the error terms difference are concentrated around zero it means that the assumptions of linear regression are validated.
- Also after checking the VIF we can also deduce the assumption of multi collinearity i.e. if the VIF values of the column are less than 5 we can deduce that there is no multi collinearity between two variables.

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

Answer

‘temp’, ‘year’ and ‘windspeed’ are the top 3 features that are contributing significantly towards the explaination of the demand of the shared bikes.

General Subjective Questions

1) Explain the linear regression algorithm in detail ?

Answer –

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Let’s say we have a dataset which contains information about the relationship between ‘number of hours studied’ and ‘marks obtained’. A number of students have been observed and their hours of study along with their grades are recorded. This will be our training data. Our goal is to design a model that can predict the marks if number of hours studied is provided. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used to apply for a new data. That is, if we give the number of hours studied by a student as an input, our model should be able to predict their mark with minimum error.

The assumptions made while doing a simple linear regression are as follows-

- The relation between the dependent and independent variables should be almost linear.
- The data is homoscedastic, meaning the variance between the results should not be too much.
- The results obtained from an observation should not be influenced by the results obtained from the previous observation.
- The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

2) Explain the Anscombe's quartet in detail.

Answer –

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3) What is Pearson's R?

Answer-

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. A correlation of -1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables.

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

Answer-

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling used?

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.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and

  1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer –

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a 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

Answer –

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis