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?

ANS : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? Answer: The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

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

ANS : drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

ANS: Looking at the pair-plot among the numerical variables, which one hasthe highest correlation with the target variable? The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if weconsider all the features.

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

ANS : One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables. If you try to fit a linear relationship in a non-linear data set, the proposed algorithm won't capture the trend as a linear graph, resulting in an inefficient model. Thus, it would result in inaccurate predictions. The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

ANS : Based on the final model, the top 3 features contributing significant towards explaining the demand of the shared bikes are : holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy, Sunday.

GENERAL SUBJECTIVE QUESTIONS

1) Explain the linear regression algorithm in detail ?

ANS : Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data and linear regression, logistic regression, ridge regression, Lasso regression, Polynomial regression are the other types of regression.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward. Linear regression is used to predict a quantitative response Y from the predictor variable X. Here, x and y are two variables on the regression line.

2) Explain the Anscombe's quartet in detail.

ANS: 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.

The most important insight from Anscombe's quartet :Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

3) What is a Pearson's R?

ANS: Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Pearson’s r is a bivariate statistical model that analyses two variables. Pearson's correlation may ALWAYS be used to test an associative research hypothesis as long as the variables being analyzed are both quantitative.

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

ANS : *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.*

Most of the times, collected data set contains features highly varying in magnitudes, 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.

*It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

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.*

$$\text{MinMax Scaling: } \boldsymbol{x} = \frac{x - min(x)}{max(x) - min(x)}$$

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**).*

$$\text{Standardisation: } \boldsymbol{x} = \frac{x - mean(x)}{sd(x)}$$

***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?

ANS : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

ANS : *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.*

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.