Assignment-based Subjective Questions

Que_1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: we build the model based on top seven highly significant feature variables as named 'yr', 'temp', 'windspeed', 'season_spring', 'mnth_july', 'weathersit_cloudy', 'weathersit_thunderstorm'. And based on those variables we derived the following equation: $y = 0.2349X1 + 0.4126X2 - 0.1591X3 - 0.1462X4 - 0.0837X5 - 0.0774X6 - 0.2696X7 + 0.3001.$

Where X1= Yr, X2= temp, X3: windspeed, X4: season_spring, X5: mnth_july, X6: weathersit_cloudy, X7: weathersit_thunderstorm and y: cnt(dependent variable)

Que_2: Why is it important to use drop_first = True during dummy variable creation?

Ans: This is something related to multi-collinearity in case of multiple-linear regression. keeping n dummy variable for n levels of categorical variable is good idea but there is catch and we can see it in the form of correlation. And to reduce this factor it's always preferable to drop the first columns, which anyway does not impact our overall analysis.

Que_3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: 'registered' column shows highest correlation with target variable 'cnt', but since 'cnt' itself derived from 'registered' it is obvious that they have high correlation among them. But besides the 'registered' variable there is a 'temp' feature which shows a high correlation of 0.63 with target variable.

Que_4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: we will validate the assumption by plotting scatter plot between the feature variables and target variable.

Que_5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: the top three feature variable which is significantly contributing towards explaining the model are:

1. Temp (temperature of the city)
2. Yr (year)
3. Weathersit_thunderstorm

General Subjective Questions

Que_1: Explain the linear regression algorithm in detail.

Ans: We categories machine learning algorithms broadly into three types supervised learning, unsupervised learning and reinforcement learning. linear regression comes under supervised learning. Which is used to predict the value of dependent variable based on the value of at least one independent variable. And target variable always follows the linear equation.

Y (dependent variable) = m*x (independent variable) + c

Que_2: Explain the Anscombe's quartet in detail.

Ans: The issue with real dataset, we often look at summaries and aggregation of the data and that's perfectly valid but it's also important to be able to dig a little deeper and visualize the data graphically there might be more insights in the data than those numbers might tell you. To support this statement, Anscombe has performed an experiment with 4 different dataset which then later known as Anscombe's quartet. The experiment shows that these four datasets are almost identical in terms of their basic descriptive statistics including mean, variance and correlation. And one must conclude that after looking at statistics the four datasets are identical in nature. but looked totally different when plotted out. So, we can conclude that don't trust summary statistics. Always visualize your data first.

Que_3: What is Pearson's R?

Ans: It is also known as correlation coefficient which helps to measure the strength of the linear relationship between two variables and as well as tell us whether they are moving in same direction or not. And it is always in between –1 to 1.

Que_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 time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude into account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalized scaling: it will bring all the data in the range of 0 to 1.

$$X = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardized scaling: Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$X = \frac{x - x_\mu}{x_\sigma}$$

Que_5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: This will happen when there is perfect correlation between two independents variable, it simply means R2 Score = 1 and its eventually leads VIF equal to infinity as VIF = 1/ 1- R2 (i.e. 1/1-1 = 1/0). And to solve this problem we need to drop any one of the feature variable.

Que_6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. 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 are similar or different in the two distributions.