Assignment-based Subjective Questions

Q.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

1. The demand for rental bikes in fall season is the highest
2. Demand is more in year 2019 as compared to year 2018
3. Demand is highest in the months of august, September, October
4. demand is less on a holiday
5. Weekday is not giving clear picture about demand.
6. when the weather is clear the demand is more

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

Answer:

drop_first=True. It is important to use at it drops the first column of dummy variable thus the redundancy is reduced, and one extra column is eliminated. If we don't drop the first column then our dummy variables will be highly correlated with each other. Suppose we have a variable gender, we don't need both male and female dummy. Just one will be fine. That is, if male =1 then the person is male and if male =0 then the person is female.

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

Answer:

temp and atemp has the highest correlation with cnt

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

Answer:

We can validate the assumptions of linear regression model by looking at the distplot of residuals or error terms. They follow normal distribution and are centered around 0. So, I validated the model by residual analysis.

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

Answer:

The top three features and there coefficients are as follow:

1. **Temp** - 0.4617
2. **yr_2019** - 1.0322
3. **weathersit_Light_Snow** - -1.1093

General Subjective Questions

Q.1 Explain the linear regression algorithm in detail.

Answer:

Linear regression is a kind of supervised machine learning. In linear regression we try to find out the relationship between a target variable and one or more predictor variables (independent variables). In linear regression problems the target variables in continuous and the independent variables can be of categorical or continuous types both. It is based on the equation **'y=mx+c'**. It assumes that the target variable and the independent variable(s) have a linear relationship with each other. We ought to find the best fit line which describes the relationship between the dependent and independent variables.

Linear regression is broadly classified into two types:

1. Simple Linear regression.
2. Multiple Linear regression.

When the target variable is one and it depends on only one independent variable then it is known as a simple linear regression problem. But when the target variable depends on two or more then two independent variables then it is known as a Multiple linear regression problem.

The equation for simple linear regression is:

$$y= \beta x + \alpha + \epsilon$$

Where:

y= target variable

β = slope

X = independent variable

α = intercept

The equation for multiple linear regression is:

$$y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

where:

y= target variable

$\beta 0$= intercept

X_1-X_p = independent variables

ϵ= error term

Q.2 Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet was developed by a statistician Francis Anscombe. It depicts the importance of fraphing data before analysing it as well as the effect of outliers on the statistical properties.

It consists of four datasets and each dataset consists 11 (x,y) points which have almost identical statistical properties, but they appear very different when plotted on a graph as their distribution is very different.

Q.3 What is Pearson's R?

Answer:

Pearson's R is a numerical summary of the strength of the association between the continuous variables. The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.

Suppose, if two variables A and B have a negative correlation then if A increases then B decreases and B increases as A increases if the correlation is positive between A and B. In simple words, it determines the effect of change in one variable when the other varuable is changing.

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

Answer:

Feature scaling is one of the most important steps during the pre-processing of data while creating a machine learning model. It is a technique used to normalize the range of data. In a dataset, the range of values of different features can be very wide. So we re-scale the data into interpretable format so that different machine learning functions can work. It doesn't affect the parameters like f-statistics, p- value or R-squared.

It feature scaling is not done then the machine learning model will weigh greater values, higher and consider smaller values as the lower ones, irrespective of the units of values.

Majorly there are two methods of scaling:

1. **Standardisation**: It is helpful when the data follows a normal distribution. It converts data into the mean vector of original data. It doesn't get affected by the outliers in the data because there is no predefined range of converted features. It can be used when we want mean to be 0 and a unit standard deviation. It is also called as Z-score normalization.

$$X= \frac{x-mean(x)}{\mathrm{sd(x)}}$$

2. **Normalisation** or **Min-max scaling**: When the features are of different scales it converts all the data on a scale of 0,1 or -1 ,1. It is useful when the data has no outliers. Features like age can be scaled by Min-max scaling but no the features like salary because some of the people can have high incomes and can lead to outliers.

$$X= \frac{\mathrm{x-min(x)}}{\mathrm{max(x)-min(x)}}$$

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

Answer:

The full-form of VIF is variance inflation factor. It determines how the variance of an estimated regression coeff. Increases due the collinearity.

$$VIF= \frac{1}{1-R_i^2}$$

The R_i is the r-squared value of the individual independent variable. If the value of R_i is 1 then the value of VIF will be infinity. If the value of R_i is 1 that means that variable is perfectly explained by the other individual variables and it has a perfect correlation with the other variables.

If R_i =1 then,

$$VIF= \frac{1}{1-1} = \frac{1}{0} = \text{infinity}$$

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

Answer:

Q-Q plot means quantile-quantile plots are plot of the quantiles of the first data-set against the quantiles of the second data-set. A quantile is a fractional number where certain values fall below that quantile. It is basically used to compare the shapes of distribution and it also shows the graphical views of how properties like scale, location and skewness are similar or different from each other. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

Generally, the Q-Q plots are used to answer some specific questions like If the two different data-sets are being taken from same set of population with a common distribution.