Linear Regression Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:
Categorical variables such as **Year, Humidity, windspeed, certain weekdays, seasons and months** have strong influence on the dependent variable "count". Some of these variables have different significance for "Casual" and "Registered" rental counts of bikes. Overall these variables are significant in predicting the demand for shared bikes

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:
When a categorical variable is converted into dummy variables to assign numerical values, we can safely drop first variable, which implies that if remaining dummy variables are assigned value 0 then it's belongs to first variable which is dropped. This helps in creating unnecessary redundant variables and helps in avoiding collinearity.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:
Variables "atemp" and "temp" have similar high correlation with target variable among all numeric variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:
Calculated the error terms (y_train - y_train_pred) on the training set to verify if the plotted curve for error terms is normally distributed. If the curve peaks near 0, it indicates that model is accurate.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:
'hum', 'windspeed', 'season' and 'weekday' are the top features contributed significantly.

General Subject Questions

1. Explain the linear regression algorithm in detail.

Answer:
Linear regression algorithm attempts to calculate the coefficients for independent variables in such a way that resulting straight line is able to explain the variance in dependent variable values to the best degree possible.

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots\ldots + \beta nXn$

It calculates the intercept and slope based on training data, to fit the line with minimal error terms possible.

2. Explain the Anscombe's quartet in detail

Answer:

Anscombe's quartet is group of 4 datasets which are different in nature but produce same mean, standard deviation and regression line.

This theory emphasizes the importance of not trusting the summary statistics of regression model. Instead highlights the need for analysing the dataset in different ways such as line plotting, box plotting, dist plotting, scatter plotting etc to understand the data patterns. This helps in validating the relevance of regression model for the given dataset.

3. What is Pearson's R?

Answer:

Pearson's R represents the correlation coefficient between 2 quantitative variables. R value indicates the strength and direction of linear relationship between variables. Value ranges from -1 to 1.

Between -1 and 0 : Negative correlation
0 : No correlation
Between 0 and 1 : Positive correlation

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

Answer:
Scaling is the data pre-processing step to scale the data for numerical variables within a finite definite range.

When quantitative variables contain values which vary in magnitudes, and range of values are different among independent variables, it can lead to incorrect modelling resulting in error in regression model. To not confuse the regression algorithm, scaling is performed on numerical variables to get the data in same range across.

Normalized scaling : variables are scaled in the range of 0 and 1
Min max scaling $x = x - min(x) / (max(x) – min(x))$

Standardized scaling: replaces the values with their Z scores.
Standardization : $x = x - mean(x) / StandardDeviation(x)$

With normalization we will lose some information about dataset such as **outliers** as all the data is brought in the range of 0 to 1.

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

Answer:
When correlation between 2 independent variables is perfect, then VIF value is infinity.
This indicates strong and high multicollinearity.

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

Answer:
Quantile-Quantile plot is used to determine if dataset follows certain probability distribution and if 2 samples of data belong to population with same distribution or not.

This helps to compare the quantiles from 2 datasets, to check if training sample and test sample belong to same distribution/population.