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?

In the case study, categorical variables are weather kind that will affect dependent variable like temp, hum etc. that will affect overall model as temperatures like dependent variables have about 65% correlation with our target variable.

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

Its required when we are dealing with binary kind of categorical variable and we need dummy variable in form for 1 and 0. If we have n dummy variable we can use (n-1) variable for model building. That will remove redundancy.

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

Temp

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

I have used p_value significance and VIF method to extract the columns first that were having high P-value and VIF. I have removed temp, hum using this validation as their p_value >0.5and VIF >5 value was high.

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

Spring, Light Snow, yr

General Subjective Questions

Explain the linear regression algorithm in detail?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

LR analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

LR fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best fit line for a set of paired data. You then estimate the value of X dependent variable from Y independent variable.

Explain the Anscombe's quartet in detail?

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph

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, r = .816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values

What is Pearson's R?

is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

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.

If scaling is not done, then algorithm only takes magnitude in 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 brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler

Standardization replaces the values by their Z scores. It brings all the data into a standard normal

distribution which has mean zero and standard deviation.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity

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

Q-Q plots are also known as Quantile-Quantile plots. they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential