Linear Regression

Subjective Answers

Name:

Questions

- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

ANS: There is a relationship between the categorical value and other independent variables like if it was Mist & Humid then the cnt were less

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

ANS: As we need only (n-1) values to represent a categorical data with n values.

- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

ANS: Temperature

Questions

- How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANS: Residual Analysis and ensured that the error is a normal distribution. And also verified the model against TEST data and ensured that R squared is in permissible value.

- Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANS: Temerature, Year, WindSpeed, Snow

General Subjective Questions

- Explain the linear regression algorithm in detail. (4 marks)

Ans : Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables. Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. Multiple linear regression is a type of regression analysis where the number of independent variables are many and there is a linear relationship between the independent variables (x1, x2….) and dependent(y) variable.

- Explain the Anscombe's quartet in detail. (3 marks)

ANS: At times there are data sets on which when we apply statistics we get same mean, r, sigma etc but then they are actually not similar in anyways. So its very important to plt the data using scatter plots to understand if the data sets are similar, 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.

- What is Pearson's R? (3 marks)

ANS : Measures the strength of a relationship. Value lies between -1 to 1. + 1 suggests +ve correlation in the same manner and –1 suggests –ve correlation in the same fashion. 0 indicates randomness and no relationship.

General Subjective Questions

- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

ANS So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

Min-Max scaling – Compressed to 0-1, all values are in between 0 and 1

Standardization (mean-0, sigma-1) - Standardized based on a formula, however the values might not range between 0-1 and can be scattered again.

- You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANS : 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)

- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANS : A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. .