

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? (3 marks)

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

Ans: Any categorical variable having 'n' distinct values can be expressed with n-1 dummy variables having values 0 or 1. Like only 1 dummy column can be used to express 2 values of the categorical column ie 0 and 1, two dummy variables can be used to express the categorical column with 3 distinct values (0,0) (0,1) (1,0). Thus instead of creating n dummy columns to express n values of categorical column we can just create n-1 dummy columns. Which leads to less no. of variables and thus complexity of our model.

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

temp

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

By plotting dist plot of residuals and it should be close to normal distribution.

`residuals = y_test - y_test_predict`

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression equation is defined as:

$$y = C_0 + C_1.x_1 + C_2.x_2 + \dots + C_n.x_n$$

where y is the target variable, $C_0 \dots C_n$ are coefficients and $x_1 \dots x_n$ are the predictor feature variables

The aim of this linear regression algorithm is to find the values of coefficients which will predict the value of y as close as possible to observed value of target variable say Y .

Firstly random coefficient values are picked and RSS(Residual Sum of Squares) is calculated for those set of coefficients.

Then with every iteration those values are tuned by minimizing the RSS(Residual Sum of Squares) using gradient decent.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to show the importance of plotting datasets.

Anscombe's quartet is a collection of 4 datasets, which have same mean, standard deviation, regression line. But their plots are very different.

3. What is Pearson's R?

Pearson's R is a measure of correlation between variables, its value ranges from -1 to 1, a value close to -1, 1 shows more correlation. A value towards zero shows less correlation.

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

Ans:

- a) scaling is a technique used during model training to bring features to same scale
- b) If scaling not done, the coefficients we will get for features after training show vast difference, and will not show the correct significance of features. And our inference will go wrong.

Also the convergence for gradient decent will take more time.

- c) standardized scaling – changes the distribution of feature into standard normal distribution ie with mean = 0 and std dev of 1. Given by $\frac{x - \text{mean}(x)}{\text{stddev}(x)}$.

min max scaler – it normalizes the value of x between 0 and 1. Given by

$$\frac{x - X_{\min}}{X_{\max} - X_{\min}}$$

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

VIF is given by $1/(1-R^2)$, the value of R^2 varies from 0 to 1. If 2 variables are highly correlated, R^2 will tend to 1 and value of VIF will tend to infinity.

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

Q-Q plots are used to see if

- i) Data set follows a certain type of distribution – theoretical quantiles for distribution against which data set is to be compared are calculated, actual quantiles of data set are calculated and theoretical and actual quantiles are plotted. A straight line means the data set follows the distribution being compared
- ii) 2 data sets follow same kind of distribution – quantiles of the two datasets are calculated and plotted against each other, a straight line means that the datasets have similar kind of distribution

In linear regression Q-Q can be used to:

- i) To check if error terms follow the normal distribution which is important assumption for linear regression
- ii) Check if the training and test set follows the same distribution