Assignment-based Subjective Questions

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

Answer: Below the three variables can infer and impact on the target variables

1) season, if the season is summer then more customer will come
2) weathersit, if weathersit is clean then more customer will come
3) holiday, for holiday more customer will come

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

Answer: There is redundant column created while creating dummies variable. If there are n level in any categorical variable then we need only n-1 dummies variable for analysis.

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

Answer: temp and atemp

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

Answer: Plotted the scatter plot to check liner relationship, R-squared value on test data is checked plotting distribution plot of residual of test data, plotting scatter plot of y_test,y_pred to check homoscedastic.

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

Answer: year, temp and weathersit.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer: Linear regression algorithm is the simplest form of regression by which we can show the relationship between dependent variable and independent variable. Which can be further used for predictive analysis.

If we have only one independent variable then it is called Linear Regression but if we have more than one independent variable then it is called as Multi Linear Regression.

To found the best fit line, minimize the RSS (Residual Sum if Squares) value. Where residuals are difference between dependent value actual data and dependent value predicted data that is:

$$e_i = y_i - y_{pred}$$

The below equitation used for best fit line:

$$y = \beta_0 + \beta_1 X$$

where,

y — is dependent variables

β_0 — intercept/ constant of line

β_1 — is coefficient of X

X — independent variables

There are two types of linear relationship:

1) Positive Linear Relationship- Dependent variable increases as independent variable increases.

2) Negative Linear Relationship- Dependent variable increases as independent variable decrease and vice-versa.

Question 2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet are set of 4 dataset which have very similar statistical summary but when it graphed it had very different distribution. Each of these datasets have eleven variables.

It shows the importance of plotting data before analyzing it and the effect of outliers on statistical summary.

1 st plot- In this plot, datapoint appear to be following a linear regression with some variance.

2 nd plot- In this, the dataset fit a neat curve but it is not following Linear regression.

3 rd plot- In this, perfect linear regression can be observed with minor variance.

4 th plot- In this, value on x axis, remain constant with some outlier.

Question 3: What is Pearson’s R?

Answer: Pearson’s R is also referred as the Pearson’s Correlation Coefficient, the Pearson product moment correlation coefficient (PPMCC), or bivariate correlation. It is used to measure the correlation between two variables. Its values lies between -1 and 1.

It can’t capture non linear relationship between two variable and it can’t differentiate between dependent and independent variable.

Below mentioned are the requirement for Pearson’s R:

1) Measurement scale should be interval or ratio

2) There should be linear relationship

3) Data should be outlier free.

4) Data should be approximately normally distributed.

(with the reference of Wikipedia)

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

Answer: Scaling is a method used to normalize the range of independent variables in multi linear regression in between 0 to 1. It became easy to analyze feature after scaling. If the feature is not scale then the higher value range will start dominating in calculation.

There are two scaling features:

1) Standardisation: - In this the variable are scaled such that the mean is 0 and sigma is 1

Formula: $x = x - \text{mean}(x) / \text{sd}(x)$

2) MinMax Scaling:- Also known as normalization. In this variables are scaled between 0 and 1

Formula: $x = x - \min(x) / \max(x) - \min(x)$

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

Answer: VIF can be infinite if there is perfect correlation between two variables.

Formula, VIF = 1 / (1 - Rsqaured)

In the perfect correlation, Rsquared value will be 1, thus the denominator becomes 0, in this case VIF will be infinite.

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

Answer: QQ (Quantile- Quantile) plots is the way to analyze whether the data is normally distributed or not. It is a plot of two quantile against each other. It is also used to find whether two data set of data comes from same distribution or not.

If the two compared distribution as similar then the in the Q-Q plot, the points will lie on the line y=x.

If there are linear relationship then the points in Q-Q plot will lie near to y=x but not necessarily on the line