Assignment-based Subjective Questions

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

Answer1: We can infer that season (If it's a summer season then more customer will come) ,Weather situation (If the weather worsen then less customer will come) and holiday(If there is holiday then more customer will come) can have impact on target variable.

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

Answer2: 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.

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

Answer3: temp and atemp

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

Answer 4: Plotted the scatter plot to check liner relationship, R- square value on test data is checked, plotting distribution plot of residual of test data, plotting scatter plot of y_test,y_test_pred to check homoscedastic.

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

Answer 5: Year, temperature and Weather situation.

General Subjective Questions

Ques 1. Explain the linear regression algorithm in detail.

Answer1 : Linear regression is simple statistical regression method by which we can be used to show the relationship between Target or dependent variable and Independent variable. Which can be further used for predictive analysis.

If there is only one independent variable, then we called it as a Linear Regression. If more than one independent variable then it is called as Multi Linear Regression.

After analyzing the data and plotting a best fit line, we try to get equation of line which is in the form. Best fit line can be found by minimizing the RSS (Residual sum of square) value. Residual are the difference between value of dependent variable at particular point and predictive value for the same point (ei=yi-ypredicted), the equation of best fit line is mentioned below

Y=B0+ B1x

Y is dependent variable; B0 is intercept/constant of line; x is independent variable and; B1 is coefficient of x.

In order to get the best fit line, we should get the best possible value of B1 and B1 which can be calculated by Gradient Descent.

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.

Ques 2. Explain the Anscombe's quartet in detail.

Answer2: 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.

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

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

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

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

Ques 3. What is Pearson's R?

Answer3: 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.

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

Answer 4: Feature scaling is a method used to normalize the range of independent variable. It helps to bring all feature value in same range i.e between 0 to 1.

It became easy to analyze feature after scaling. In few algorithms like Gradient descent converge much faster after scaling the variables. If the feature is not scale then the higher value range will start dominating in calculation.

There are two scaling features:

1) Standardizing:- In this the variable are scaled such that the mean is 0 and a unit variance.
   x=(x-mean(x))/sd(x)

2) MinMax Scaling:- Also known as normalization. In this variables are scaled between 0 and 1 by using the max and min value of the variable.
   X=(x-min(x))/(max(x)-min(x))

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

Answer 5: VIF can be infinite if there is perfect correlation between two variables.

According to the formula, VIF =1/(1-Rsqaured)

In perfect correlation, Rsquared will be 1, thus the denominator becomes 0 and the VIF will be infinite.

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

Answer 6: 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.