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?

A) From our analysis the month,weathersit, weekday are the categorical variables .

In could be infer it :

In weekdays: Sat,sunday,Monday having the more count of booking .

In month :January,March,May,July,August,October and December having the more booking.

In weathersit : Good and moderate categories have more counts.

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column.

If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished

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

A) By observing the pair-plot in the analysis the temp with the cnt variables are the highest correlation with the target variable.

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

A) By checking the dependent and independent variables and between the correlation of the categorical variables.By doing the train split test knowing the R-Square and Adjust R-square.

Finally By conducting the test of the between predected variables and test predicted variable.Assuming the variable for the best fit for the linear model.

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

A) From the analysis of the bike sharing data. The top 3 features are temp, windspeed and hum .

General Subjective Questions

1. Explain the linear regression algorithm in detail

A) Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.The values for x and y variables are training datasets for Linear Regression model representation.

Graph

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$y=mx+c \Rightarrow y= a0+a1x+ \varepsilon$

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

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

A) Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were

constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

A) Correlation means to find out the association between the two variables and Correlation coefficients are used to find out how strong the is relationship between the two variables.The most popular correlation coefficient is Pearson's Correlation Coefficient. It is very commonly used in linear regression.The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

A) Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Techniques to perform Feature Scaling

Consider the two most important ones:

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text {new }}=\frac{X_{i}-\min (X)}{\max (x)-\min (X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text {new }}=\frac{X_{i}-X_{\text {mean }}}{\text { Standard Deviation }}$$

It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between normalized scaling and standardized scaling

Noramlized scaling :

Minimum and maximum value of features are used for scaling.It is used when features are of different scales.Scales values between [0, 1] or [-1, 1].It is really affected by outliers.Scikit-Learn provides a transformer called MinMaxScaler for Normalization

standardized scaling:

Mean and standard deviation is used for scaling.It is used when we want to ensure zero mean and unit standard deviation. It is not bounded to a certain range.It is much less affected by outliers.Scikit-Learn provides a transformer called StandardScaler for standardization.

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

A) If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.Also, it helps to determine if two data sets come from populations with a common distribution. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.

If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.