Linear Regression:Subjective Questions

By : Animesh Anand

==========

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?

–Below are the observations:

a) Overall,shared bike demand had significantly risen in **2019. Demand is more in 2019 as compared to 2018.**
b) Monthwise : **May, June, July, August, September** have more demands than other months.Its true for both year with increase in number in 2019
c) Season Wise :There is more demand in **fall** as compared to any other season.
d) Weather wise : Demand is high in **Clear** weather.
e) Day wise: Demand is high on **Wednesday, Thursday, Friday** and **Saturday.**

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

– because it helps in reducing the extra column created during dummy variable creation. Which inturn reduces the correlations created among dummy variables.

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

–'temp' variable has the highest correlation. It's the same for 'atemp' variable.

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

– Validation of assumptions of linear regression after building model:

a) Histogram to ensure error terms are normally distributed.
b) Heat Map and VIF for multicollinearity check.
c) CCPR plot for linear relationship validation.

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

– Top three feature on which demand on shared bikes depends:

a)Yr
b)temp
c)windspeed

==========

General Subjective Question:

1.Explain the linear regression algorithm in detail.
– Linear Regression is a model based on the linear relationship between dependent variable and independent variable.This means the value of independent variable changes based on the change in dependent variable.

Let's understand this with the help of mathematical equations and examples:
A simple linear regression expression **y=mx +c** where **y** is the dependent variable which depends on the actions of **x** i.e independent variable .

*Linear regression can be of:

a) Simple linear regression: if the number of independent variables is one.
b) Multiple linear regression: if the number of independent variables is more than one.

*Linear regression can be of:

a) Positive type if dependent variable is directly proportional to independent variables.
b) Negative type if dependent variable is inversely proportional to the independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet consists of 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. But when they are plotted on scatter plots they appear differently and they have very different distributions.

This explains the importance of data visualization before applying different algorithms to build models.The data feature needs to be plotted to see the distribution and identify the different details like outliers.

3. What is Pearson's R?
Pearson's R is used to measure the linear correlation.It measures the strength and direction of the relationship between two variables and its value lies between -1 to +1. It can be of

a) Positive type if the variable change is in the same direction and the it's values lie between 0 and 1.
b) Negative type if the variable change is in the opposite direction and then its value lies between -1 and 0.

Also, there is no correction if there is no relationship between variables and then its value is 0.

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

– It is a method performed during data pre-processing to normalize the range of an independent variable. Without scaling, there will be a huge difference in the value regardless of the unit as the algorithm will try to mark greater value higher and smaller value lower .

Normalized scaling Vs Standardized scaling

a) Normalized scaling values are in the range of -1 and 1 or 0 and 1 whereas Standard scaling is not bound to any range.
b) Minimum and maximum values are used for scaling in normalized scaling whereas mean and standard deviation is used in standard scaling.
c) Normalized scaling is more affected by outliers than standard scaling.

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

– Value of VIF is infinity when there is a perfect correlation that means there is a perfect correlation between two independent variables.In the case of perfect correlation, we have

R^2 =1

=> 1/(1-R^2) = infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data possibly came from a common distribution. It is a scatterplot created by plotting two sets of quantiles against one another.

The use or purpose of the QQ plot is to show if two data sets come from the same distribution.

This helps in a scenario of linear regression

a) when we have training and test data set received separately
b) when we have two data sample

and we want to confirm that both the data samples or sets are from populations with same distributions

References used in this assignment:

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2
https://www.kaggle.com/code/irfanakgul/multiple-liner-regression/notebook
https://www.heap.io/blog/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story