Assignment-based Subjective Questions - Solutions

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

(1). Solution :

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with workign day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

(2). Solution :

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. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the 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. Hence, 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?

(3). Solution :

Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'

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

(4) Solutions.

Check the assumption using a Q-Q (Quantile-Quantile) plot. If the data points on the graph form a straight diagonal line, the assumption is met. You can also check for the error terms' normality using statistical tests like the Kolmogorov-Smironov or Shapiro-Wilk test.

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

(5). Solution :

Three main contributing and significant features towards the demands of share bikes are:

yr_2019 (Positive correlation).

temp (Positive correlation).

weathersit_Light_Snow (negative correlation).

General Subjective Questions – Solutions

Q1. Explain the linear regression algorithm in detail.

Sol.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

An example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe’s quartet in detail.

Sol.

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.

3. What is Pearson’s R?

Solution:

The Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

4. Sol.

Scaling.: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Scaling performed for following reason:

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.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

1. Minimum and maximum value of features are used for scaling
2. It is used when features are of different scales.
3. Scales values between [0, 1] or [-1, 1].
4. It is really affected by utliers.
5. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
6. This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
7. It is useful when we don’t know about the distribution
8. It is a often called as Scaling Normalization

Standardization Scaling:

1. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

2. sklearn.preprocessing.scale helps to implement standardization in python.
3. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.
4. Mean and standard deviation is used for scaling.
5. It is used when we want to ensure zero mean and unit standard deviation.

6. It is not bounded to a certain range.

7. It is much less affected by outliers.

8. Scikit-Learn provides a transformer called StandardScaler for standardization.

9. It translates the data to the mean vector of original data to the origin and squishes or expands.

10. It is useful when the feature distribution is Normal or Gaussian.

11. It is a often called as Z-Score Normalization.

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

(5) Solution:

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).

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

(6) Solution:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. If the two data sets come from a common distribution, the points will fall on that reference line.

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.