BIKE SHARING ASSIGNMENT

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?**

Answer:

The following details could be derived from the analysis of the categorical variables on the dependent variable "cnt":

1. Booking bikes is on the rise year over year
2. On working days, there is a higher demand for bikes compared to holidays.
3. There is a high demand for bikes in the fall, followed by the summer and then the winter
4. demand increases steadily from April and peaks between Jul-Sep and decreases slowly during winter
5. On a clear day, bike demand is higher, on a rainy day, bike demand is lower
6. As well as the outside temperature, if the weather warms up, bike demand also grows
7. The demand for bikes decreases in humid weather
8. The demand for bikes decreases in the Misty day
9. Windspeed decreases the bike demand

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

Answer:

Our data becomes multicollinear when we use dummy variables for categorical variables. When the categorical variable has n levels, we must have (n-1) levels in order to overcome this multicollinearity. As a result of dropping one column, the correlations between dummy variables are reduced.

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

Answer:

‘Temp’ is the variable that has highest correlation with the target variable “cnt”.

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

Answer:

The following steps were followed to validate the assumptions of Linear Regression after building the model on the training data:

1. A scatter plot was used to visualize the correlation between the residual values and the fitted values.
2. In order to determine whether the data had a normal distribution, I used a q-q plot.
3. The VIF factor for the variables was less than 5.
4. To examine correlated patterns in residual values, I used autocorrelation plots.
5. The histogram of the error terms was plotted to see if it was normally distributed and had a zero mean.

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

Answer:

- Temperature
- Windspeed
- Weather situation

General Subjective Questions:

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior 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

2. Explain the Anscombe's quartet in detail.

Answer:

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 has four data set that has information about variance, mean of x and y for all four data set. By plotting the data set using scatter plot, it helps us to identify various anomalies present in the data set like outliers, diversity of data, linear separability of data and distribution of data.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The above table depicts the four data sets of Anscombe's quartet.

3. What is Pearson's R?

Answer:

The most popular Correlation coefficient which is used to measure strong relationship between two variables is Pearson's R. Pearson's correlation (Pearson's R) is a correlation coefficient commonly used in linear regression. It is also called as Product-Moment correlation coefficient or bivariate correlation. It is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. It is represented by two letters, rho (ρ) for a population and the letter "r" for a sample. It has numerical value between -1 and +1. If the value is 1, then we have positive correlation and -1 for negative correlation. If the value is 0, then there is no correlation between the variables.

The formula for Pearson's R is given as below:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Pearson's R cannot capture nonlinear relationships between two variables and clso cannot differentiate between predictor and output variables. It doesn't provide any information about the slope of the line.

Following are some requirements for PMCC:

- Scale of measurement should be interval or ratio
- Variables should be normally distributed
- The association should be linear
- There should be no outliers in the data

R value determines the relationship between the variables.

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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

Answer:

Scaling is a data pre-processing step which is applied to independent variables to normalize the data within a particular range. When there is lot of independent variables with different scales, then we need normalize the data so that we can easily interpret the data and provides faster convergence for gradient descent method. It also helps to speed up the calculation in the algorithm.

For example — if we have multiple independent variables like age, salary, and height; With their range as (18–90 Years), (25,000–75,000 rupees), and (4–6 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.

Two types of scaling method are normalized scaling and standardize scaling. A normalized dataset will always have values that range between 0 and 1. A standardized

dataset will have a mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for the maximum and minimum values. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Standardization may be used when data represent Gaussian Distribution, while Normalization is great with Non-Gaussian Distribution. Normalization is good to use when the distribution of data does not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors. Normalization is often called as Scaling Normalization while standardization is called as Z-score normalization. Normalization is used when features is of different scales while standardization is used when we want to ensure zero mean and unit standard deviation.

In Scikit-Learn, MinMaxScaler is used for normalization while StandardScaler is used for standardization.

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

Answer:

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. 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.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. It also help to assess if a set came from distribution such a normal, exponential or uniform distribution.

This helps in a scenario of linear regression when we have training and test data set to confirm whether using Q-Q plot that both the data sets are from populations with same distributions. It is helpful to determine if residuals follow a normal distribution. We also can verify the error terms assumption in linear regression. It is useful to determine the skewness of distribution. A Q–Q plot requires more skill to interpret.

It is used to check following scenarios:

If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Advantages of this plot is that it can be used for sample sizes and to detect distributional aspects like shifts in location, presence of outliers and shifts in the scale.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x-axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x-axis