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:

I could infer the following details from the analysis of the categorical variables on the dependent variable “cnt” the bike sharing demand:

1. Season fall has highest demand for the bike booking followed by summer and winter season.
2. The demand for bike booking is increasing year over year.
3. There is less demand of bike during Sunday and Monday compared to other days of the week.
4. Demand for bike is more on working day compared to holiday.
5. Weather determines the bike demand. So when the day is clear, then the demand is more. Demand of bike is less during Rainy or Snowy day.
6. Depending upon the weather condition of the day, the demand for the bike increases or decreases. Also the demand increases steadily from April and peaks between Jul-Sep and decreases slowly during winter.

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

Answer:

When we create dummy variable for the categorical variables, we create multicollinearity in the data. To overcome this multicollinearity, we need to have (n-1) levels when we have n levels for the categorical variable. By dropping one column, it reduces the correlations created among the dummy variables.

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

Answer:

‘Temp’ is the numerical 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:

I followed the steps below to validate the assumptions of Linear Regression after building the model on the training set:

- I used scatter plot to visualize the correlation effect among the residual and fitted values.
- I made sure that VIF factor for the variables was less than 5.
- I used q-q plot to check whether the data has normal distribution or not.
- Also I used autocorrelation plot to check correlated pattern in residual values.
- For auto-correlation, I checked Durbin – Watson(DW) statistic. My model value is less than 2(1.945) which means positive correlation.
- I plotted histogram for the error terms to see whether it is normally distributed and has the zero mean.

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

Answer:

The top 3 features contributing significantly towards the demand of the shared bikes dependent upon my final model are:

- Temperature (temp)
- Windspeed (windspeed)
- Weather situation (weathersit)

General Subjective Questions:

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical linear approach for modelling the relationship between a predictor (independent variable) and output variable (target variable). If we use one predictor variable to determine the relationship, then it is known as simple linear regression and if more than one predictor variable is used to determine the relationship with the target variable it is called multiple linear regression.

Linear regression is used for the predictive analysis. In machine learning, linear regression is used to train a model to predict the behavior of the data based on some variables. The relationship between X-axis (independent variable) and Y-axis (depdent variable) should be linear. This model gives a sloped straight line describing the relationship within the variables. We need to make sure that the line that is formed should be best fit straight line.

Linear regression is represented as follows:

$y = B_0 + B_1x$ (For simple linear regression)

Here B_0 is intercept and B_1 is slope

$y = B_0 + B_1x1 + B_2x2 + \ldots + B_nxn$ (For multiple linear regression)

Steps to be followed in linear regression algorithm are:

1. Reading and understanding the data
    - Reading the data and clean the data and make into assessable data for analysis. Data cleaning and manipulation is performed.
2. Visualizing the data
    - Use various plot to visualize the numeric and categorical variables to intercept about the business/domain inferences.
3. Data preparation
    - Convert categorical variables into dummy variables or convert into numeric variables, so that these variables can be represented during model building.
4. Splitting the data into train and test set
    - Split the data into train and test data set. Also rescale the trained model using MixMax scaling or standard coding.
5. Building a linear model
    - Build the model using any one of the method like Recursive Feature Elimination, Forward/ Backward/ stepwise selection based on AIC and Regularization method.
6. Residual analysis of the train data
    - Check whether assumptions are met in this step.
7. Making predictions using the final model and evaluation
    - Predict the test dataset by transforming into trained dataset.
    - Evaluate the model using RMSE. Check for the R^2 and Adjusted R^2 score between trained and test data.

2. Explain the Anscombe's quartet in detail.

Answer:

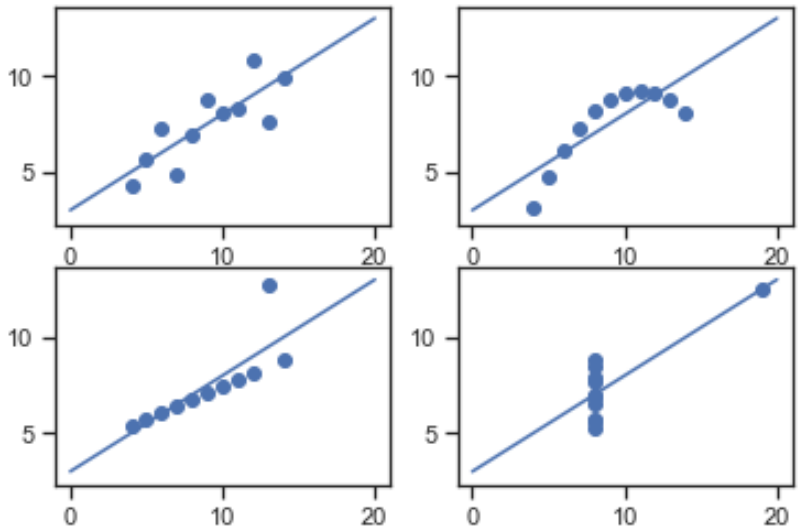
Anscombe's quartet was designed by statistician Francis Anscombe, which tells us about the importance of visualizing data before applying various algorithms to build model. It can be defined as four data sets which are identical in statistical observations but when the data is plotted, they look different from one and other. When these data sets are plotted, they have different distributions and appear differently using scatter plots.

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.

When these four data sets are plotted using scatter plots, we obtain the below figure:



First Data Set depicts linear regression model that fits the best

Second data set depicts the non-linearity among the data and hence it could not fit linear regression model.

Third data set depicts that there is some linear relationship but have different regression line. It shows the outliers involved which influence the data set.

Finally fourth data set shows there is no relationship between variables due to outliers in the data set. But it can produce high correlation co-efficient.

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.

6. 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.

7. 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