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 bike sharing count is increasing from Spring -> Summer -> Fall. And it is dropping in summer. Its explainable that in winter due to heay snow and cold , people would not prefer bike. So weather has a significant role in bike sharing.
- There is a decrease in number of people who share bike in holiday. Its becasue in holiday most of the people prefer to stay at home.
- Bike sharing showing a significant progress from 2018 to 2019. More people are now prefering this system.
- People would not prefer bike in Light snow or raining time. And also when there clear sky or few cloudy people are more likely to use shared bikes. So rain and snow affect will affect the business.

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

Answer:

If a categorical variable has n number of variables, then it can represent by n-1 numer of dummy variables.

For example if an age category there are three values ie young, middle and old. It can represent by dummy variable table

young | middle | old
0 | 0 | 1
0 | 1 | 0
1 | 0 | 0

It has three level. But it will be still explainable if a level dropped ie young.

middle | old
0 | 1
1 | 0
0 | 0

Here if middle and old is zero , then it should be young. so we can still represent the variables with 2 levels.

So drop_first=True will drop the first dummy variable and it can explained by n-1 levels.

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

Answer:

temp : temperature in Celsius

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

Answer:

I validated the assumptions of Linear Regression models by following methods

1. Linear Relationship

   Features should show a visible linearity with target variable

2. Normal distribution of error terms
   The residual errors should be normally distributed
3. No Multicollinearity
   Every feature should be independent each other
4. Homoscedasticity
   There should not be any visible patters in residual erros
5. No auto-correlation or independence
   No auto-correlation

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 three features are :

- Temp
- Light_Snow_Rain
- windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is the machine learing algorithm which is based of supervised learning. This perform regression task and models using indepnedent variables. Main puprpose linear regression is to find the linear relationship berween two or more variable . Based on this linear regression relationship, it predicts the future value of one of the variable using other variables. Lets say it as dependent variable. The other variables are independent variables.

The linear formula is

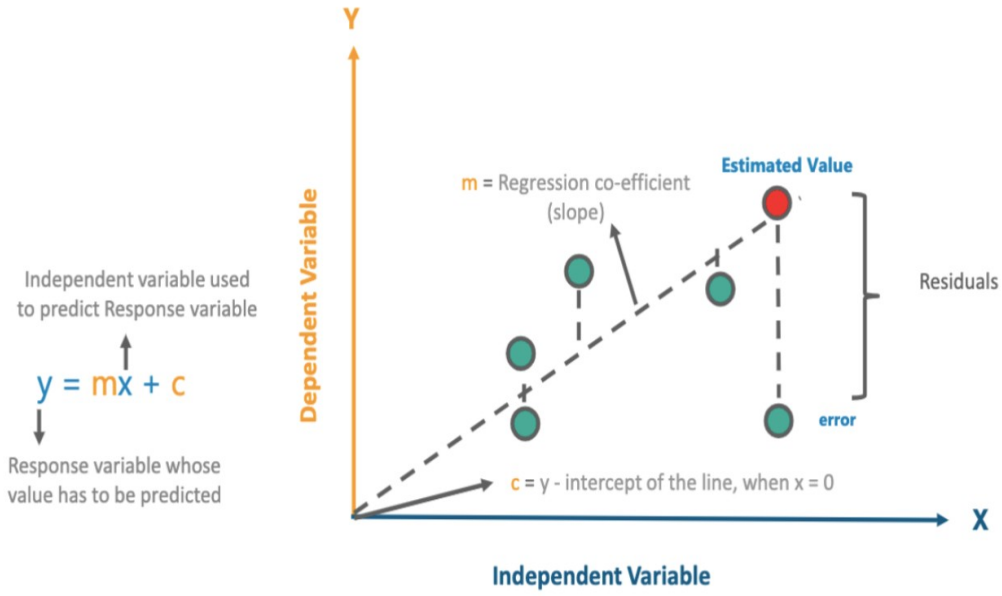
y = mx+c

where m = slope of the line

c = intercept

y = dependent variable

x = independent variable



Main two types of linear regressions are :

- Simple linear regression
  It describes a relationship between one dependent and one independent varaible using a straight line.
- Multiple linear regression
  It predict the outcome of one dependent variable from multiple independent variable using a straight line

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

Answer:

Feature scaling is used to bring all the independent variables under a same standardize range. So that it will be easy to infer on same range of data. It done before the model building. Otherwise the some variables may tend to have high value data while other varaibles have small value of data. Then it will be difficult to compare these variables.

There are two methods for scaling

1. Min-Max scaling
   Min-Max rescale in a range between [0, 1] or [−1, 1]. It is really affected by outliers. It is used when features are in different scale. Scikit learn provide library for this method.
2. Standardisation (mean-0, sigma-1)
   It is rescaled by mean and standard deviation of data. It is bound all data into certain range.

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

Answer:

VIF is the Variance Inflamation Factor that used to caluclate the correlation between variables. If the VIF is infinity, it shows the perfect correlation between variables. That means one of the independent variable can perfectly described by other independent variables. For linear regression model this column should be dropped because it can explain by other variable and in order to avoid multicollenearity.

$$VIF_i = \frac{1}{1 - R_i^2}$$

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this 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.

Answer:

Q-Q plot is the quantile-quantile plot.It is probability plot and it is a graphical method for comparing two probabilities by plotting their quantiles against each other.

It is graphical technique to determine two data set come from a populcation with same distribution.

Use :

- To find that two data set comes from a population of common distribution
- To check these two dataset have common location and scale
- To confirm these two dataset have similiar distribution shape

Its role linear regression is when we get test and train dataset received seperately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions . So it has significant role in linear regression