Assignment-based Subjective Questions

Name: Somvir Singh Nain
Batch: September 30, 2022

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

Answer: Based on the data visualizations made in Python, I have made the following observations and inferences.

a. Irrespective of the season or month, the number of users have significantly gone up in 2019 as compared to 2018.
b. If we make a monthwise analysis, we can see that for both the years, the months of may, june, july, aug, sep and oct have seen more bookings of users as compared to the other months.
c. If we look at the seasons, no of users are the highest during the fall season and lowest during spring. Bookings in summer are higher compared to winter. This maybe true because given the harsh winters in the US, people would prefer staying at home during winters.
d. The seasonal data almost corroborates approximately with the monthly data if we compare the months and seasons in the US.
e. Number of bike users has been higher during clear weather during both the years and lower when it has been raining or snowing. This is absolutely normal.
f. We can also see clearly that demand is more during Thursdays, Fridays and Saturdays. Also demand has increased in 2019 as compared to 2018 for all days.
g. There's a good trend as far as bookings are considered as irrespective of any factor, 2019 has seen higher number of users than 2018.

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

Answer: drop_first=True is used so that we can avoid creating an extra column during dummy variable creation. Eg. Suppose we have 3 kinds of values in a categorical column (YES, NO & NA) and we want to create a dummy variable for that column. If one variable is not YES neither NO, then It is obviously NA. So we do not need a 3rd variable to identify the NA entries.

This just helps to reduce the correlations among dummy variables which otherwise would have to be removed later. Our aim should always be to keep the model as simple as possible.

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 with the target variable.

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

Answer: All the below assumptions have been validated by me:

a. It has been ensured that error terms are approximately normally distributed. This is evident from the histogram.
b. Multicollinearity check was conducted. There was no multi collinearity as is evident from the VIF calculations and also from the heatmap.
c. Linear relationship validation was made using CCPR plot and linearity was visible among the variables.
d. Homoscedasticity check was done. No visible pattern was found among residual variables.
e. `R2 score of test data predictions: 0.69441`
f. `R2 score of train data predictions: 0.70236`
g. `Absolute difference between R2 scores of test and train dataset predictions: 0.00795`
h. Difference bw train and test model results is less than 5%. so it is generalized model with app 1% difference.

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

Answer: Based on Demand temp, wind speed and year which are the top 3 features.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$y = ax + b$

y is the dependent or response variable that we are trying to predict

x is the independent or explanatory variable we are using to make predictions

a is the slope of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b

Such a linear relationship can be either positive or negative.

A linear relationship will be called positive if both independent and dependent variable increases. Eg. Increase in inflation leads to an increase in the interest rates

A linear relationship will be called negative if independent variable increases and dependent variable decreases. Eg. A fall in the Share market leads to an increase in gold prices.

Linear regression can either be Simple linear regression or multiple linear regression depending on the number of explanatory variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Since Pearson's correlation coefficient measures linear association it may give a low result when variables have a strong, but non-linear relationship.

Reference: Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American

Statistician 27 (1): 17–21. JSTOR 2682899

3. What is Pearson’s R?

Answer:

Pearson's r (also called Pearson’s product-moment correlation coefficient) is a measure of the strength of the linear association between 2 variables. If the variables tend to move together in the same direction, the correlation coefficient (r) will be positive. If the variables tend to move in the opposite direction, then the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take values ranging from +1 to -1. A value of 0 indicates that there is no association whatsoever between the two variables. This was put forth by Mathematician Karl Pearson.

Pearson’s R is heavily affected by outliers and is less reliable in such cases.

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Some of the differences between normalized scaling and standardized scaling are as under:

a. In Normalized scaling, Minimum and maximum value of features are used for scaling whereas in Standardized scaling, Mean and standard deviation is used for scaling.
b. Normalized scaling is used when features are of different scales. Standardized scaling is used when we want to ensure zero mean and unit standard deviation.
c. Standardized scaling is not bound to a certain range whereas normalized scale values are always between 0 and 1 or -1 and 1.
d. Standardized scaling is not much affected by outliers whereas normalized scaling is really affected by outliers.
e. StandardScaler is the transformer used in Sci-Kit learn library of python for standardized scaling
f. Scikit-Learn provides two transformer but we used MinMaxScaler for Normalization scaling.

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation,

$$R^2 = 1$$

$$\frac{1}{1 - R^2} \rightarrow \infty$$

To solve this, we need to drop one of the variables from the dataset which is causing perfect multi collinearity.

6. 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

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.