Linear Regression Assignment

Subjective Questions

Name- Swaraj Parida

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

- The demand of bike is significantly less in the month of spring in compare with other seasons.
- There is a significant increase in the demand of bike in the year 2019 than year 2018 as it can assume that bike demand will also increase in this way if there will be no lockdown or any other unforeseen situation.
- Demand of bike remains high between the month of June to September.
- Bike demand is less in holidays compare to other days.
- There is no significant difference in the demand throughout weekdays.
- There is no significant difference in the bike demand in working and non-working day.
- The bike demand lowest in case of Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans.

It actually makes the model less complex as it reduces the extra column created during dummy variable creation. Therefore, it reduces the correlations created among dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans.

Both ‘temp’ and ‘atemp’ have the highest correlation with the target variable ‘cnt’ at the pair-plot.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

After building the model on training set, we the validation of assumptions on the following methods.

a. **Linear Relationship**
   For this case, we checked the linear relationship between the predictors and the dependant variable by plotting scatterplots. Here, we confirmed the same.
b. **Absence of Multicollinearity**
   For this, we checked the Multicollinearity by calculating the VIF of our selected features and we got VIF scores less than 5 in our final model.

c. **Normality of Errors**
   Here, we did residual analysis by taking the y train and y-predicted. After that we plotted a distplot were we found that errors are normally distributed with mean value 0.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans.

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

a. Temperature having coefficient 0.549892
b. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds having coefficient -0.287090
c. Year having coefficient 0.233139

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans.

Linear regression is a machine learning algorithm based on supervised learning. Here, we train a model to predict the behaviour of your data based on some variables. Regression models can perform prediction value of target variable based on independent variables. In the case of linear regression two variables which are on the x-axis and y-axis, are linearly correlated.

Mathematically, we can write a linear regression equation as:

$y = mx+ b$

where, m is the slope and b is the intercept.

Q2. Explain the Anscombe's quartet in detail.

Ans.

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.

Q3. What is Pearson's R?

Ans.

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of

their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.

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

Many times, data set that we collect contains features highly varying in magnitudes, units and range. If scaling is not performed, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we need to perform scaling to bring all the variables to the same level of magnitude.

Normalisation Scaling	**Standardized scaling**
Scales values between 0 and 1 or -1 and 1.	It's not limit to a certain range.
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is affected by outliers.	It is not that affected by outliers.
Scikit-Learn provides a transformer called *MinMaxScaler* for Normalization.	Scikit-Learn provides a transformer called *StandardScaler* for standardization.
It is called as Scaling Normalization.	It is called as Z-Score Normalization.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans.

In case of perfect correlation, then VIF is 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. In order to solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans.

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. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

If the two distributions are similar, the points in the Q–Q plot will approximately lie on the line y = x. In the other hand, 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.