Linear Regression 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? (3 marks)

Answer:

- The CNT median value increases significantly when it's not a Holiday. This is on expected lines as the consumers would be using the bikes more for reaching office / official travel. So we should see more bookings when it's not a holiday.
- The CNT median value increases in the Summer and Fall season. This is also on the expected lines as driving bikes in the Winter season wouldn't be preferred. The amount of booking s in the Summer / Fall season would be higher in the summer season.
- The CNT median value increases when there is clear weather, mist weather situation as compared to the light snow / heavy rain scenario. This is also on the expected lines.
- The CNT median value doesn't show too much deviation based on the weekday category variable.

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

Answer:

Dummy variables are created from categorical variables. For example, if we create the dummy variables from a categorical variable of 5 values, we can completely define the categorical variable with just 4 dummy variables making the 5th dummy variable a redundant variable. If we have all the 5 dummy variables, the model can be hit by multicollinearity of the dummy variables. By dropping the first dummy variable, the multicollinearity is avoided.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Temp (temperature) is the numerical variable which 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? (3 marks)

Answer:

The assumptions can be validated in the following ways

1) The important assumption after training the model on the training data is that the residual values are normally distributed. Residual values are the differences between the observed V/s Predicted values. The residual values can then be plotted on the dist plot to see that the values are normally distributed
2) We can also check for the multicollinearity by checking the VIF (variance inflation factor) values. The VIF values should be less than 5. If we see a VIF of more than 5 or more than 10 value, then some multicollinearity exists.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:
The most important categorical variables from the model generated and their effect on the dependent variable are as follows

Temperature: With a positive coefficient of 0.5928. Temperature is the highest feature variable which affects the Count of Bikes getting booked. With every increase of the 0.5928 units of temperature, the bike bookings increases by the same value

Humidity: With a negative coefficient of -0.2784, the Humidity variable inversely affects the Count of bikes getting booked. With the increase in Humidity, the number of bikes getting booked are lesser.

Year: With a positive coefficient of 0.2268, the year variable proportionately increases the Count of bikes getting booked.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression in a type of supervised machine learning algorithm that shows the relationship between the dependent variable and one or more independent variables. Linear regression models are relatively simple to understand and the linear equation gives a simple mathematical formula for doing predictions. From a high level, the linear algorithm uses a large amount of data to calculate the linear equation. The data is decided into Train and Test data. The user first trains the algorithm on train data to come up with the equation. This is then used on the test data for predictions. Various values like R squared or VIF are used to see if the predictions are falling under acceptable conditions.

There are two types of linear regression:

1) Simple linear regression
2) Multiple linear regression.

Simple linear regression:

Simple linear regression studies the linear correlation between 2 variables. Simple linear regression has the following equation

$$Y= \beta0*X + \beta1 + \varepsilon$$

β0 and β1 are two unknown constants representing the regression slope, whereas ε (epsilon) is the error term.

As we can see the simple linear regression can be used for just 2 variables. It can be used to predict the expansion of the metal based on the temperature of the metal, the value of a flat based just on the number of rooms in the flat.

Multiple linear regression:

Multiple linear regression studies the correlation between the predicted and 2 or more independent variables. Multiple linear regression has the following equation

$$Y = \beta0 + \beta1x1 + \beta2x2 + \cdots + \beta kxk + \varepsilon$$

Where Y is the dependent variable
B0, B1 are the coefficients, the describe the change in Y for per unit change of the variable
ε term describes the random error (residual) in the model.

As we can see from the equation, in the Multilinear residual regression, we have multiple variables which are affecting the Y value. Example: The price of a company share can be predicted by using the revenue, sales and other factors making this prediction as a multiple linear regression.

Steps for creation of the Multiple linear regression model

1) First step is to do clean the data and then do the EDA for that data
2) After first level EDA is done, substituting the categorical value with the dummy variables is done.

3) After this, we split the data into Train data set and Test data set is done usually in the ratio of 70:30
4) First we try to create the model on the Train data set.
5) Then by using RFE, we can reduce the number of variables to the required number of variables
6) After this, we will keep checking the value of P for the variables which are present.
7) Any parameter’s P value which is higher than 0.05 will be removed from the train data set
8) Once all the parameters have values less than P < 0.05, we can check for multicollinearity with the help of VIF.
9) Once all the VIF values of the parameters are less than 5, check for the R squared value
10) This completes the training of the model
11) Now we can check for the Residual analysis to see whether the errors are normally distributed.
12) Now we will start the prediction of the data with the help of trained model
13) Use the trained model to predict the predicted value based on the trained model
14) At the end use the Rsqured code to check if the Rsquare value from the test data is near the R squared value of the train data set.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s quartet is a set of 4 datasets which are having identical values of mean, R squared, correlation and linear regression lines but having different representations when the scatter plot is used.

Anscombe’s quartet is used to show the importance of the Exploratory data analysis and to avoid over emphasizing on the summary statistics. It promotes the usage of the data visualization graphs and tools with which we can spot the outliers, trends which cannot be deciphered from the statistics alone.

3. What is Pearson’s R? (3 marks)

Pearson's R correlation coefficient is a number between -1 to 1 and it measures the strength and direction of the relationship between two variables.

If the value is between 0 and 1, its said to be a positive correlation and in this scenario when one variable changes, the other variable also changes in the same direction. Example of positive correlation are like when the phone usage increases the phone heating also increases.

If the value is 0, then it suggests that there is no correlation between the variables. Example of no correlation is like price of the railway ticket v/s price of bus fare.

If the value is between 0 and -1, then it suggests that there is a negative correlation between the variables. Example of the negative correlation is that the price of metal goes down depending upon how abundant the metal is.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling of variables in linear regression refers to the process of transforming the numerical values of the independent or dependent variables on similar scale. This ensures that the variables contribute equally to the regression analysis.

Some of the reasons as to why scaling is done are as follows

1) Comparability of the coefficients: When the variables are on different scales, the coefficients of the variables are not directly comparable. Scaling of variable ensures that the coefficients represent the change in the dependent variable per unit change in the scale variable.

2) Numerical Stability: Scaling of the variables can improve the numerical stability and the convergence of optimization algorithms used to estimate the coefficients

Common methods of scaling include

1) Normalized Scaling: This method scales the values to a specific range , from 0 to 1.
2) Standardized Scaling: This method scales the values to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

In linear regression analysis, VIF is a measure of the degree of multicollinearity of one variable with the other variables. Multicollinearity happens when one or more variable combined can define the other variable completely. In this scenario, having multicollinearity variables will increase the variance of regression coefficient estimates. To detect the multicollinearity we use the VIF. The greater the VIF, the higher the degree of multicollinearity. If the variable is equal to a linear combination of other variables, the VIF tends to be infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plot(Quantile-quantile) plot is a probability plot. Its a graphical method for comparing the two probability distributions by plotting their quantities against each other. If the two distributions are similar, then the Q-Q plot will approximately lie on the identity line (x - y). Q -Q plots are used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale are similar or different in two distributions.