Linear Regression Subjective Question and Answers

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

Answer: In our bike sharing dataset, we have season weathersit mnth and weekday

Are categorical variables. The dependent variable cnt i.e. count of bikes rented are correlated with these.

Season: we have 4 seasons available in the dataset but by analyzing the correlation on heat map we identified that dependent variable is highly correlated with season value of Spring i.e 55%

Weathersit : this is another categorical variable where we have 4 weather situation defined,but out of these we identified that weatherSit value=1i.e Partly cloudy, the rentals are high compare to other two weather and in heacy weather there is almost no rentals.

Month(mnth): Observing the boxplot for month, we have seen like bikes are rented more during month value=5 to 10 compare to other months.

Weekday: Working day are highly correlated to cnt dependent variable i.e almost 63% as observed on the heatmap. Also from boxplot Wednesday is showing high numer of rentals.

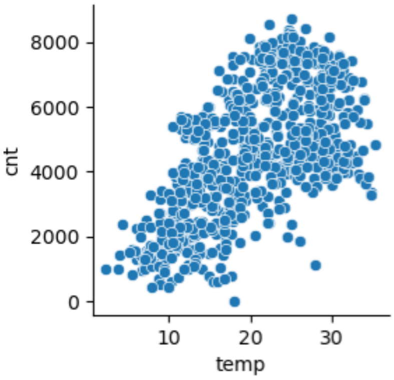
2. Why is it important to use drop_first=True during dummy variable creation?

Answer: When we create the dummies variable for categorical variable, it will create m levels instead of m-1 levels and create un-necessary collinearity between the dummy variables.

Also by using m-1 dummy variable we can infer all levels. That is why we should always use drop_first=True

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

Answer: temp variable is highly correlated with target variable



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

Answer: After building the model, we performed

1. Residual Analysis: In this activity we first calculated the target predicted value from the model. Then we calculated the error term and plotted in histogram. From histogram we identified that mean of the error is almost 0 and error terms are normally distributed. So our assumptions are good.
2. Then we ran the model with test data and predicted the target variable using test data.
   In that we calculated the r-square value of the test data. Then compared the r-square value obtained from training set and test set. We identified that both r-square values are nearly same.

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: Three main feature are:

1.Temperature (temp)

2.Season

3. workingday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a supervised machine learning model that tries to find the linear relationship between independents variable and dependent variable. The Linear regression model tries to basically find the best fir linear line between the dependent(Y) and independent variable(X).

Linear regression is of two type:

a) Simple linear regression: In this kind of model we have one dependent and one independent variable.
Equation: b0+b1X
b) Multiple Linear regression: In this Model we have one continuous dependent variable and multiple independent variable
Equation: b0+b1x1 +b2X2+b3x3+…

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartlet comprises of a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. It basically emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone. The four datasets of Anscombe's quartet include 11 x-y pairs of each data.

3. What is Pearson's R? (

Answer: Pearson R is called Pearson correlation coefficient. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is the ratio between the covariance of two variables and the product of their standard deviations.

The formula to calculate Pearson R is

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

r= correlation coefficient

x_i= values of the x-variable in a sample x= mean of the values of the x-variable

y_i= values of the y-variable in a sample

y= mean of the values of t

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

Answer: Scaling is an activity which we do on numerical data to bring all features to a comparable scale.

The different data in the data set vary in degree of magnitude,ranges and units.So for our Machine learning models to interpret these data on the same unit or scale, we perform scaling.

Normalization: This scaling mechanism is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] Normalization is useful when there are no outliers.

Standardization: This scaling mechanism is used to transform the features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Standard deviation

Standardization can be helpful in cases where the data follows a Gaussian distribution. This takes care of outliers also.

Also it is not bounded to a certain range as compared to normalization.

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

Answer: VIF i.e variance inflation factor determines the collinearity between predictor variables.

Generally vif<5 is good and whatever high is assumed that predictor variables are highly correlated.

So if vif is infinity then there is a perfect correlation between the predictor variable.

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 a probability plot used to compare two probability distributions by plotting their quantiles against each other.

Here we plot x from the first distribution and y taken from same quartile of the second distribution. The curve usually plotted is a paramtric curve.

It 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.

In linear regression, it helps us to compare the sample distribution of the variable against quantiles of a theoretical distribution.