Linear Regression - Subjective questions

Kishor Kunal | DS27

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

-> season,month,weather_situation,weekday and workingday are categorical variables from the dataset. Weekday or working day doesn't affect the sales that much. Fall season has the most bookings, accordingly Clear weather_situation and Sep month have the most sales. Similarly Jan and Spring has the least bookings.

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

-> it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. For example, if we have 3 types of values in a Categorical column and we want to create a dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obviously unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

-> Temperature

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
-> By doing Residual analysis, VIF and R2-square

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
-> Working, Sunday and Feeling Temperature,

General Subjective Questions

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

-> Regression is a data mining function to predict a number. Regression algorithms fall under the family of Supervised Machine Learning algorithms which is a subset of machine learning algorithms. One of the main features of supervised learning algorithms is that they model dependencies and relationships between the target output and input features to predict the value for new data. Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and uses the model to predict the value for new data. Some of the most popular applications of Linear regression algorithms are in financial portfolio prediction, salary forecasting, real estate predictions.

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

-> Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
2. The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
3. In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R? (3 marks)

-> Pearson's r or Pearson correlation coefficient, 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

->

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm
We have to do scaling to bring all the variables to the same level of magnitude.
It just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1.

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

If there is perfect correlation, then VIF = infinity.
Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity.
if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:
1 = not correlated.
Between 1 and 5 = moderately correlated.
Greater than 5 = highly correlated.

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 is Quantile-Quantile (Q-Q) plot. It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with same distributions.