

Upgrad Assignment-based Subjective Questions

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

A. From the Analysis, below inferences were drawn:

- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more bookings.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

A. `drop_first = True` is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax -

`drop_first`: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one

variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

A. 'temp' Variable

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

A. The assumption of Linear Regression Model were validated based on below 5 assumptions –

- Normality of error terms
- Multicollinearity check
- Linear relationship validation
- Homoscedasticity
- Independence of residuals

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

- temp
- winter
- windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. Mathematically the relationship can be represented with the help of following equation –
$$Y = mX + c$$

Here, Y is the dependent variable and X is the independent variable we are using to make predictions.

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

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

Furthermore, the linear relationship can be positive or negative in nature as explained below–
Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variable increases.

Negative Linear relationship:

A linear relationship will be called negative if independent increases and dependent variable decreases.

Linear regression is of the following two types –

Simple Linear Regression

Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-collinearity –Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation –Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables –Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms –Error terms should be normally distributed

Homoscedasticity –There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

A. Anscombe's quartet comprises a set of four datasets, 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. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

A. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r , can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

A. 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. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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

A. If there is perfect correlation, then $VIF = \infty$. 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, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

A. 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.