Details:

Name: Bhushan Date
Course: Executive PG Program in Machine Learning, ML & AI
Batch : ML-C38 Feb- 2022

Linear Regression Case study Subjective Question & Answers

Linear Regression Assignment

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

Ans 1: As seen in the dataset provided, there are 2 categorical variables present namely ‘season’ & ‘weathersit’.

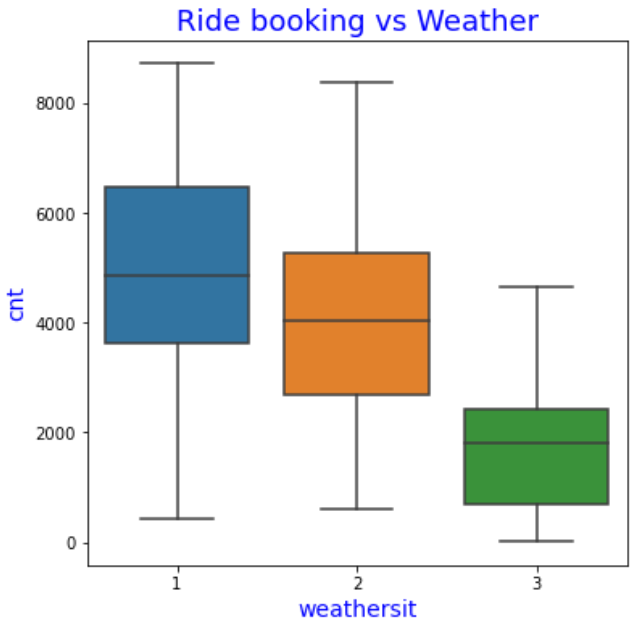


Fig. 1

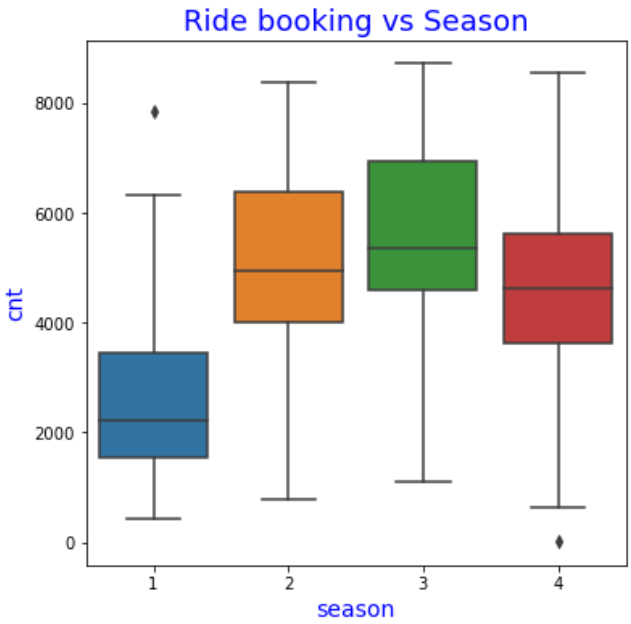


Fig. 2

- ❑ “Wethersit” has following interpretations:
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

As seen in Fig. 1 dependent variable increases in the order of Light rain to better in Cloudy & best in clear as per data.

- ❑ “Season” has following interpretations:
  - 1:spring
  - 2:summer
  - 3:fall
  - 4:winter

As seen in Fig. 2 dependent variable increases in order from spring, winter, summer & fall. Where as fall having max no of rides booked.

Assignment-based Subjective Questions

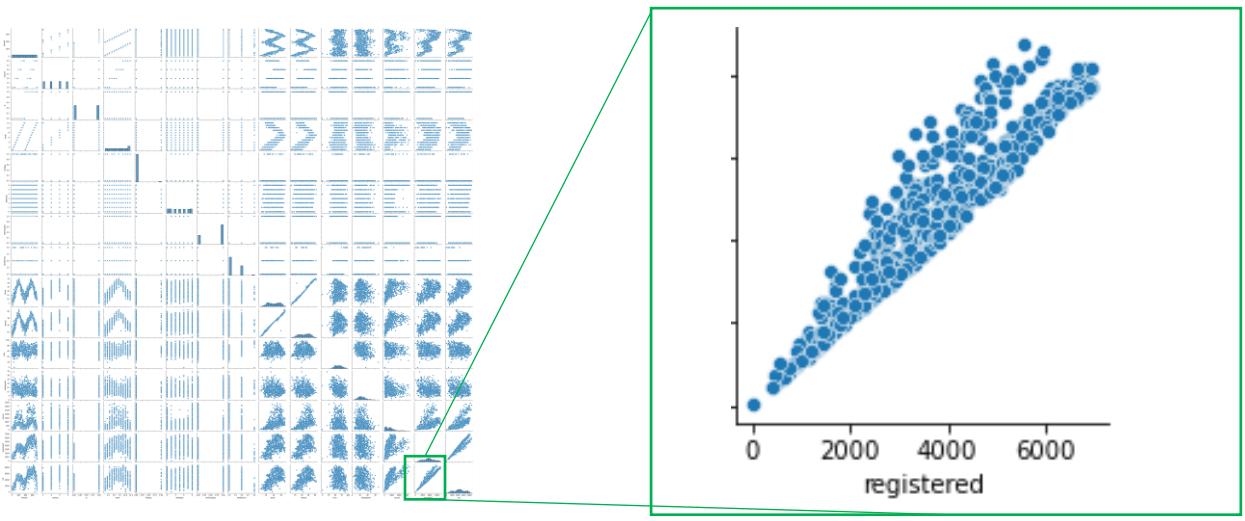
Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans 2: The attribute drop_first is used to drop the extra column created while converting categorical variable to integer format for performing regression. This drops the left most (first) column from the created dataframe. The logic behind this is to reduce the correlation between dummy variables.
For example, a categorical column having n-levels of categories then the whole data can be explained by n-1 columns. As the null value in all the (n-1) columns will explain the data of (n)th column.

Example: Coin Flip
A categorical column having data of coin flipping i.e. Heads / Tails when we convert this data to integer by performing 'pd.get_dummies' the output dataFrame will have 2 columns having data of heads & tails each. The Heads column will contain the output 1 in case of heads & 0 in case of tails & for Tails column the data will be vice versa. This same data can be explained efficiently by any one column in format of 0's & 1's where 1 will represent the presence of that particular column & 0 will represent the significance of 2nd column.

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

Ans3 : Looking at the pair-plot among the numerical variables, ‘registered’ variable has the maximum correlation with target variable i.e. ‘Cnt’. As seen in the scatter plot all the points show linear scattering at 45deg to both axis. Means when 1 variable varies the other varies in the same manner.



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

Ans4 : Linear Regression has following assumptions,

1. First assumption of Linear regression is that the relationship between the independent and dependent variables needs to be linear.

As per pair plots shown while performing EDA all the variables have linear dependencies with target variable.

2. Second Assumption of linear regression is that the mean of residuals need to be zero.

This assumption is verified by plotting distribution of errors. As seen in fig. 3 the mean of error is near to zero, the assumption is satisfied.

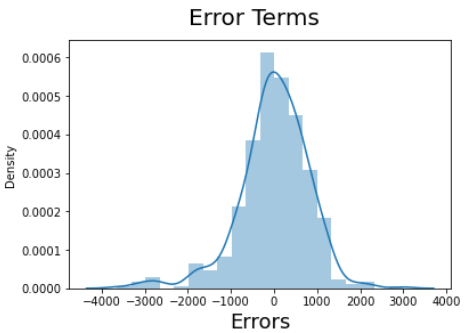


fig. 3

3. Third Assumption is the Distribution of error is normal in nature.

As seen in fig. 3 the distribution of error is normal in nature.

4. Fourth assumption is Error term is independent of each other

As seen in fig. 4 the distribution of error random & no visible pattern is seen

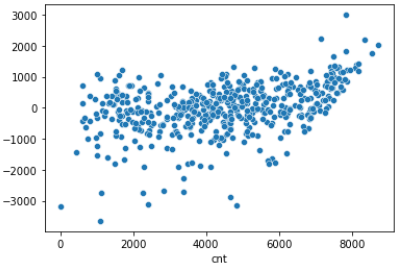


fig. 4

5. Fifth assumption is independent variables are not correlated

As seen in VIF is less than 5 it shows that the independent variables do not have correlation between each other.

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

Ans 5: The Top 3 features contributing significantly towards the demands of the target variable are as follows:

1. Yr

As the coefficient of Yr id '2061.60' & value of T-statistics is '27.92' it is **positively correlated**

2. Atemp

As the coefficient of atemp is '3801.10' & value of T-statistics is '14.195' it is **positively correlated**

3. Weathersit

As per the coefficient of is '-2437.75' & value of T-statistics is '-11.004' it is **negatively correlated**

Linear Regression Assignment

Q1: Explain the linear regression algorithm in detail.

Ans1: Linear regression is basic form of machine learning where a model is trained to predict the variation of target variable depending upon some independent variables.

Mathematically equation of linear regression is written as:

$$y = mX + c$$

Where:
y = dependent variable of dataFrame
m = slope of line
X= independent variables from dataFrame
c = intercept of line

When there is one independent variable then the model is called Simple linear regression model, in case of multiple independent variables the model is called Multiple linear regression model.

Modelling can be done using two methods:

1. Top down method: In this all variables are considered while building model & then individual variables are eliminated considering P-value & VIF score.
2. Bottom up method: in this method individual variables are added to model.

RFE – Recursive feature elimination method is mostly used for eliminating features from a training dataset efficiently.

General Subjective Questions

Q2: Explain the Anscombe's quartet in detail.

Ans2: Anscombe's quartet is a group of 4 data sets, which are identical in simple descriptive statistics with some peculiarities.
The datasets have a peculiarity that fools the regression model if built.

They tends to follow different distribution.

When these models are plotted on a scatter plot, each data sets generates a different kind of plot that isn't interpretable by any regression algorithm

Q3: What is Pearson's R?

Ans3: Pearson's R will be used for identifying the linear correlation between different datasets having numeric values.
It assigns a value between -1 and 1.

The value of 1 signifies that there is positive correlation between both, means increase in one will result in the increase in other

The value of -1 signifies that there is negative correlation between both, means increase in one will result in decrease in the other

The value of 0 represents no correlation.

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

Scaling is a technique to standardize the values of independent variables convert to a predetermined range to same magnitude. Purpose of scaling is to get simple model & for fast execution of code. There are two major methods to scale the variables, i.e. standardization and Minmax scaling namely normalization scaling.

Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

NO.	Normalization	Standardization
1	Minimum and maximum value of variables are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	MinMaxScaler from Sklearn library is used for Normalization.	StandardScaler form Sklearn library is used for standardization.
6	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

Ans 5: When there is a perfectly correlated between independent variable present in the dataFrame, then the value of VIF will be infinite. In the case of perfect correlation, we get R^2 =1, which lead to 1/(1-R^2) infinity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Example:
Considering the data provide for assignment, the independent variable 'temp' was having correlation factor 0.95 with 'atemp'. Hence we eliminated 1 variable from dataset. If the same was not done the vale of VIF would reflect infinite.

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

Ans6: Q-Q plot also known as the Quantile – Quantile plot, plot the quantiles of a sample distribution against quantiles of theoretical distribution. This helps in determining the type of distribution followed by the data such as, normal distribution, exponential distribution, etc

This helps in modelling based on the type of distribution