upGrad

#LifeKoKaroLift

Assignment and General Subjective Questions

Submitted by: Vishal Arora

iit-b

Agenda

upGrad

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)
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
3. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Assignment based Subjective Questions

Assignment based subjective questions

upGrad

Question 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: Following are inference made by analyzing categorical variable from the dataset on dependent variable

1. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.

2. Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.

3. Overall spread in the month plot is reflection of season plot as fall months have higher median.

4. People rent more on non-holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals

5. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.

6. Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that

7. Clear weather is most optimal for bike renting, as temperate is optimal, humidity is less, and temperature is less

Assignment based subjective questions

upGrad

Question 2. Why is it important to use drop first=True during dummy variable creation? [2 Marks]

Answer: A variable within levels can be represented by n-1 dummy variables. So, if we remove the first column then also, we can represent the data. If the value of variable from 2 to n is 0, it means that the value of 1st variable is 1. Example: 'Relationship with three levels, namely, 'Single', 'In a Relationship", and "Married.) would create a dummy table like the following:

Relationship Status	Single	In a Relationship	Married
Single	1	0	0
In a Relationship	0	1	0
Married	0	0	1

But I can clearly see that there is no need to define three different levels. If I drop a level. say "Single, I would still be able to explain the three levels. Let us drop the dummy variable Single from the columns and see what the table looks like:

Relationship Status	In a Relationship	Married
Single	0	0
In a Relationship	1	0
Married	0	1

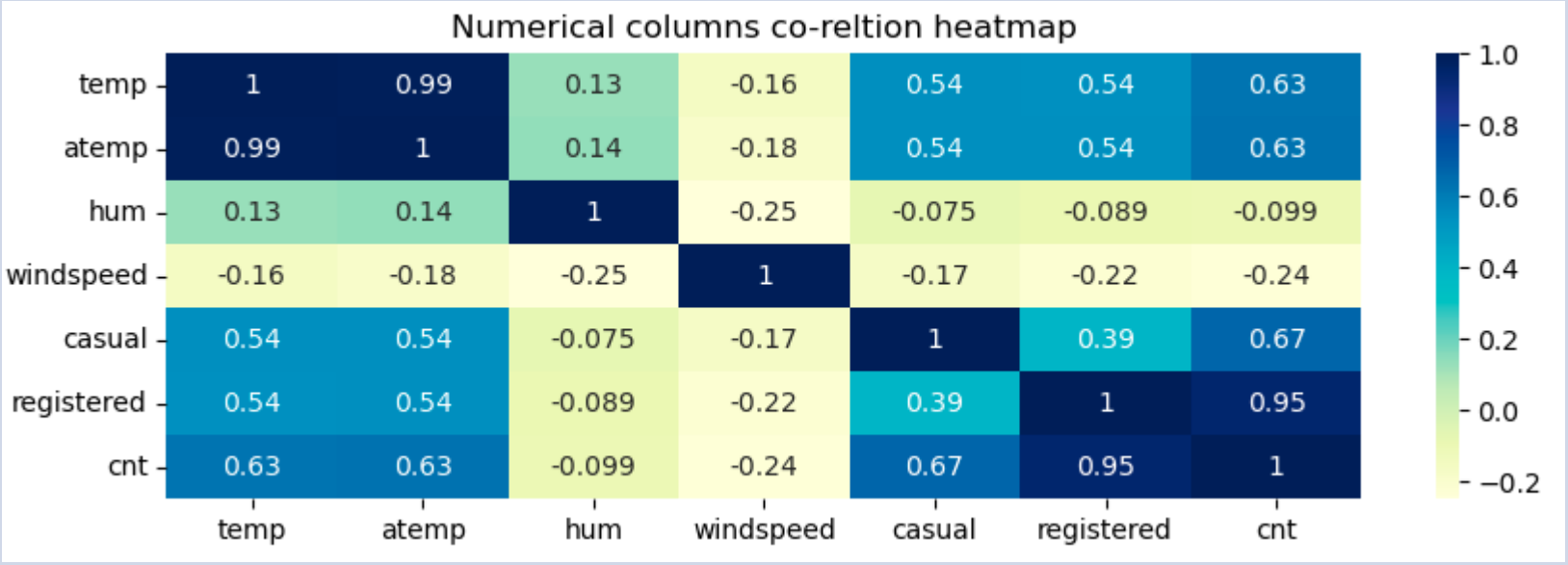
If both the dummy variables, namely. “In a Relationship” and "Married”, are equal to zero, that means that the person is single. If “In a relationship” is one and “Married” is zero, that means that the person is in a relationship, and finally, if "In a relationship” is zero and "Married" is 1, that means that the person is married.

Assignment based subjective questions

upGrad

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? [1 Marks]

Answer: Temp has highest correlation coefficient of 0.63

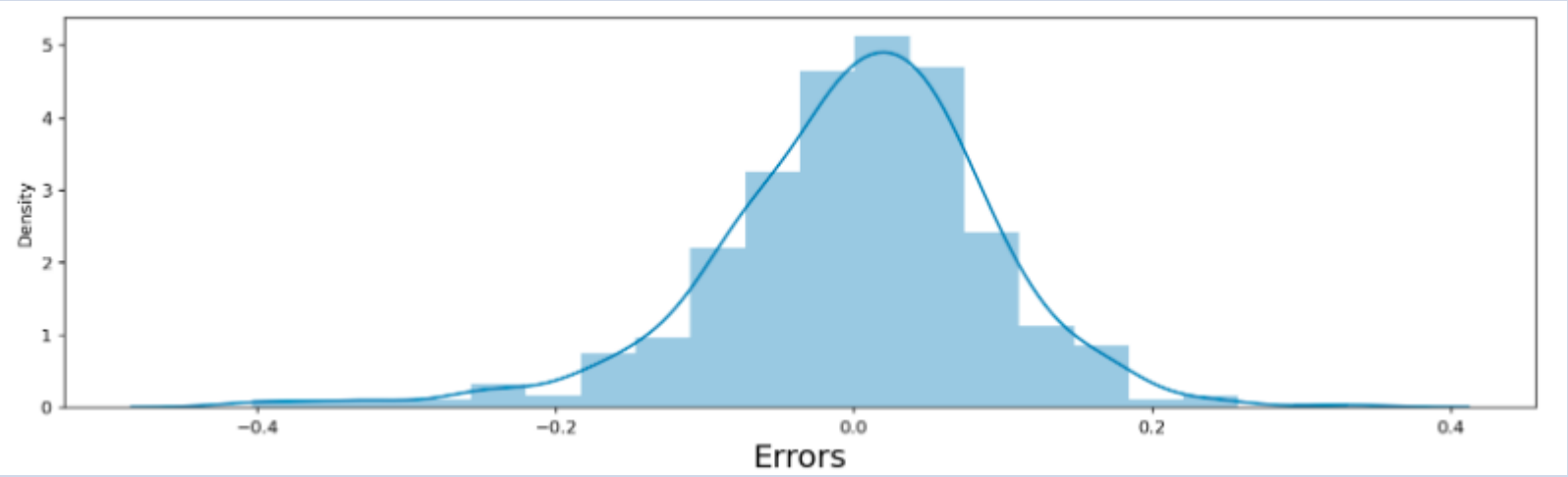


Assignment based subjective questions

upGrad

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set? [3 Marks]

Answer: By plotting the residual distribution. It came out to be a normal distribution with a mean value of 0



The distribution of residual should be normal. We test this residual by producing a distplot of residual to see if they follow a normal distribution or not.
The residual are scattered around mean = 0 as seen in the diagram above

Assignment based subjective questions

upGrad

Question 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 following are the top 3 features contributing significantly towards explaining the demand of the shared bikes:
(1) Temp – 0.5527
(2) Year 2019 - 0.2332
(3) Cloudy – 1.49

Based on equation derived from analysis

$$y=0.1209+0.5527\times temp-0.1552\times windspeed+0.2332\times Year2019+0.0894\times summer+1.48\times cloudy+0.1281\times winter+0.0978\times sept-0.2785\times light-rain$$

General Subjective Questions

Assignment based subjective questions

upGrad

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

Answer: A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables only. Following steps are performed while doing linear regression:

- The dataset is divided into test and training data
- Train data is divided into features independent) and target (dependent) datasets A linear model is fitted using the training dataset. Internally the api's from python uses gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares
- In case of multiple features, the predicted variable is a hyperplane instead of line. The predicted variable takes the following form:

$$y(x) = p0 + p1x1 + p2x2 + \ldots + p(n)x(n)$$

- The predicted variable is than compared with test data and assumptions are checked.

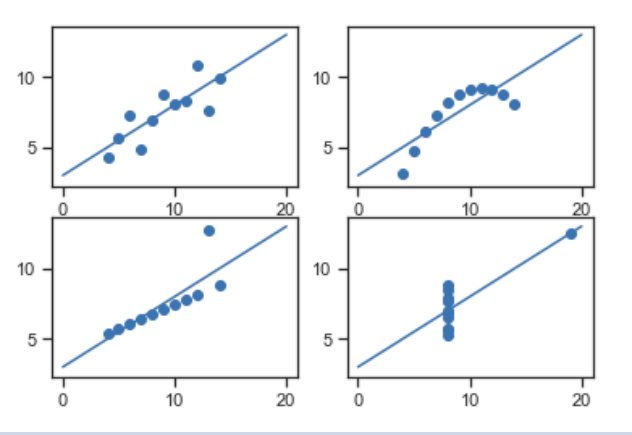
Assignment based subjective questions

upGrad

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

Answer: Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically. The simple statistics consist of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value, Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. The graphs are shown below

(image source: https://medium.com/analytics-vidhya/anscombes-quartet-an-importance-of-data-visualization-856b3d1bd403#:~:text=It%20comprises%20of%20four%20data,etc)%20but%20different%20graphical%20representation.)



1. The first plot is simple linear regression relationship
2. The second plot (side right) is not distributed normally and correlation coefficient is irrelevant as it shows a nonlinear relationship
3. The third plot (bottom left) is linear but has different regression line. This is happening because of the outliers present in the data
4. The fourth plot (bottom right) does not show linear relationship however due to outliers the statistics got adjusted.

In a nutshell, it is a better practice to visualize data and remove outliers before analysing it.

Assignment based subjective questions

upGrad

Question 3: What is Pearson's R? (3 marks)

Answer: Pearson's R measures the strength of association of two variables. It is the covariance of two variables divided by the product of their standard deviation. It has a value from 1 to -1.

A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase

A value of 0 means no correlation

A value of -1 means a total negative correlation. It means that if one variable Increase then other will decrease

Assignment based subjective questions

upGrad

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

Answer: Scaling of a variable is performed to keep a variable in certain range. Scaling is a pre-processing stop in linear regression analysis. The reason we scale a variable is to make the computation of gradient descent faster. The step size of gradient descent are generally low for accuracy, if the data has some small variables (values in the range of 0-1) and some big variables (values in the range of 0-1000) than the time taken by the gradient descent algorithm will be huge.

Normalised Scaling	Standardized Scaling
Called as min max scaling, scales the variable such as that the range is 0-1	Values are centered around mean with a unit standard deviation
Good for non-Gaussian distribution	Good for Gaussian distribution
Value id bound between 0 and 1	Value is not bounded
Outlier are also scaled	Does not affect outliers

Assignment based subjective questions

upGrad

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

Answer: The formula of VIF is

$$VIF = \frac{1}{1 - R_i^2}$$

So, if R square is 1 than VIF will become infinite. This means that there is perfect correlation between the features

Assignment based subjective questions

upGrad

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

Answer: A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two sets of data came from the same distribution. It is a visual check of data. If the data is from same source than the plot will appear as a line.

The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot.Because both sets of quantiles came from the same distribution, the points should form a tine. That's a fairly straight line,

The q- plot is used to answer the following questions

- ❑ Do two data sets come from populations with a common distribution
- ❑ Do two data sets have common location and scale?
- ❑ Do two data sets have similar distributional shapes?
- ❑ Do two data sets have similar tail behaviour?

upGrad

#LifeKoKaroLift

Thank You!