Bike Sharing Assignment

Prepared & Submitted by:
Diya Biswas

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?
   Ans.

   - The demand for bikes increases in the year of 2019 compared to 2018.
   - In January bike;s demand is low. In June to September the demand is very high.
   - The demand of bike is almost similar throughout the weekdays,
   - The weekday plot indicates that more bikes are rented during Saturday.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans.
   It is important to use because it helps to reduce the extra columns that have been created during dummy variable creation.

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

The numerical variable ‘registered’ has the highest correlation with the ‘cnt’ target variable.

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

By Quantile-Quantile plot we can check the assumptions of LR after building the model on the training set. If the data points on the graph from a straight diagonal line, the assumption will meet. Also error terms can be useful in this case.

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

- Weathersit
- Windspeed
- temp(Positive Correlation
- yr_2019(Positive Correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans.

Linear Regression is one of the very basic forms of machine learning where we build a model and after that we have to train it and test it to predict the behavior of the variables which are mentioned in the dataset.

And also by train and test set(X_train, y_train, X_test, y_test) we can create boxplot, heatmap, etc.

2. Explain the Anscombe's quartet in detail.

Ans.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties but little different after graphed. Each dataset consists of 11 points.
All these were created in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphical before analyzing it and also to check the effect of outliers on statistical properties.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

3. What is Pearson's R?

Ans.

- Pearsons's R measures the strength of the linear relationship between two variables. Pearson's R is always between -1 and +1
- The correlation coefficient lies between -1 and +1. i.e. $-1 \leq r \leq 1$
- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation
- If r = +1, then the correlation is perfect positive • If r = –1, then the correlation is perfectly negative.
- If r = 0, then the variables are uncorrelated.

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

Ans.

Mainly we have multiple variables in our data, they may be on different scales. Eg. From the housing dataset: price, area, no of bedroom, no of bathrooms etc all the variables are in different scales. It is necessary to bring everything between 0 to 1 so that we can effectively put them in perspective and build a model. When every predictive variable is in the same scale, it is easy to interpret them and easy to use in the model.

We do use the fit_transform for the train data.
We can only transform the test data.

Standardizing Scaling is done using standard deviation Min max using the maximum and minimum of the data, values are set between zero and one.

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

Ans.

By any chance if there will be perfect correlation, VIF will be infinite Mainly, VIF shows a perfect correlation between two independent variables.
In case of perfect correlation: R2=1, which leads to 1/(1-R2), it means infinity.
To solve this kind of problem, we need to drop one variable which can cause this multicollinearity

.

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

Ans.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall down below that point and 50% lie above it.

The purpose of Q-Q plots is to find out if two sets of data are coming from the same distribution or not. A 45 degree angle is plotted on the Q-Q plot; if the two data sets are coming from a common distribution, the points will fall on that reference line.

A Q-Q plot is used to compare the shapes of the distributions, providing a graphical view of how all the properties like scale, skewness,etc. Are similar or different in the two distributions.