Linear Regression Subjective Q&A

Assignment-based Subjective Questions

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

The categorical variables had barely any impact on the dependent variables. The ones that could be counted as having any influence on the dependent variables were the *months* and *year*.

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

If not used then the first column would be considered redundant data as the remaining columns already hold the binary information available in the first column.

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

Merely looking at the pair-plot, it is hard to distinguish the right answer to this question. It could either be the *temp* or *atemp* variable that has higher correlation. But using the .corr() method lets us confirm that the *atemp* variable has the highest correlation among the numerical variables.

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

Using the statsmodel functions, we eliminate those variables that have a low coefficient or a high p-value.

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

The top 3 features contributing to the demand of bike sales would be *year* and the months of *June* and *September.* This could be point to the seasonal shifts that occur in these months as well as the increase in demand for these bikes that occur every year, especially in this dataset that shows the increase in demand that has occurred from 2018 to 2019.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A linear regression algorithm finds the most appropriate constants from a dataset that can help align to the line of an equation. From this equation, predictions can be made to come up with values for a target variable.

These values should not follow the line so strictly as to not account for unforeseen circumstances (or variables) called overfitting and they should not be so laxed that they do not obey the equation called underfitting.

Q3. What is Pearson's R?

The correlation between two variables, or simply put the amount by which one variable has an influence over the value of another variable, is measured using several ways one of which is the Pearson Correlation Coefficient (r)

The value of r can range from –1 and 1 which indicates how strong the relationship is between the two variables measured and whether it positively influenced or negatively (which means 1 value increases, the other value decreases)

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

Scaling is the act of bringing values of numerical variables all down to a proportional scale so that generally higher values are given the appropriate weight when building the model.

For example, a 1,000,000$ for 4000 sqft. flat in Dubai might actually be a budget apartment when considering the rest of the flat area and pricing in the dataset even though it might sound like a lot.

So scaling is performed to give values of variables the appropriate weight when considering the entire dataset.

The main feature that I have understood between normalized and standardized scaling is that normalization is used when variables are of different scales while standardization is used when the variables are similar.

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

It is understood from the term that a high VIF or Variance Inflation Factor means that there is a correlation between the independent variable and the target. Therefore an infinite VIF value means that there is a high multicollinearity present and we need to drop the variable(s) influencing this behavior.