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?

Answer:

In the Linear regression model

CNT= 0.26 + .24 * yr + 0.04 * weekday -0.15 * windspeed + 0.27 * season_2 + 0.31 * season_3 + 0.23 * season_4 -0.07 * weathersit_2 + -0.28 * weathersit_3

Following affects seen for each feature mentioned below.

season_2 0.2719

season_3 0.3168

season_4 0.2254

weathersit_2 -0.0743

weathersit_3 -0.2834

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

Answer :

We remove one of the fake variables in order to escape the dummy variable trap. This implies that your model has only n−1 dummy variables.

This is possible since the other dummy variables implicitly contain the information about the discarded category.

Drop_first=True lowers the dimensionality of the feature space, improves interpretability, and helps avoid multicollinearity.

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

Answer : temp/atemp

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

Answer :

Test data:

```
from sklearn.metrics import r2_score
r2_score(list(y2_test),y2_test_pred)
0.7374797033768311
```

Train data:

Dep. Variable:	cnt	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.756

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

Answer :

Form the model summaery following top 3 features contributing significantly towards explaining the demand of the shared bikes

yr 0.2396

season_3 0.3168

season_4 0.2254

General Subjective Questions

==========

1. Explain the linear regression algorithm in detail.

Linear Regression is an algorithm used to model the relationship between the sets of numerical data of one or more independent variables (also called predictors or features and a dependent or target variable. It finds the best fit line which minimizes sum of squared errors between the predicted variables and the target variable. This data helps in calculate the coefficients (slopes) of the linear regression model. The model performance is evaluated using various metrics like R-squared, mean squared error, etc. to check how well it fits the observed data.

We can measure the strength of the linear relationship, by using a correlation coefficient. These coefficients are determined by taking the partial derivatives of the sum of squared residuals with respect to each coefficient and setting them equal to zero.Coefficient Correlation r ranges from -1 to +1, If r=0 → It means there is no linear relationship. It doesn't mean that there is no relationship

linear regression finds the coefficients that best fit a straight line to the observed data by minimizing the residual sum of squares (RSS)using the method of least squares. The goal of the linear regression algorithm is to get the best values for B0 and B1 to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

where ypredicted = B0 + B1 Xi

MSE stands for Mean Squared Error. It is a measure of how well a model fits the data.

The MSE is calculated as follows:

MSE = ∑(residuals)2 / n

Where:

∑(residuals)2 = The sum of the squared residuals. This is calculated by squaring each residual and adding them all up.

n = The number of observations or data points.

2.Anscombe's quartet refers to a set of four datasets that have very different distributions, yet have nearly identical statistical properties when analyzed through basic descriptive statistics. The datasets have the exact same means, variances, correlations, and regressions. However, their distributions and plots appear quite different.

However, when plotted as x-y graphs, they appear very differently:

1: Linear trend

2: Curvilinear trend

3: No trend, outlier point

4: No trend, symmetrical

This implies only the statistics are not enough for the analysis but graphical analysis is required to truly gain insights. Anscombe's quartet is a very important lesson in statistics and data analysis.Visualization and modeling play a key role in properly analyzing, interpreting, and applying conclusions based on data.

Dataset 1:

Linear Trend

*

Dataset 2:

Curvilinear Trend

*

*

*

*

*

*

*

Dataset 3:

No Trend, Outlier Point

*

*

Dataset 4:

No Trend, Symmetrical

*

*

*

*

*

*

*

The spacing and alignment of the asterisks attempts to represent the differences in distribution shape, trends, outliers between the four datasets. While limited without true data plots, hopefully this gives a sense of how varied the data can look despite the same summary statistics!

3. What is Pearson's R

The value of the Pearson correlation coefficient product is between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

The Pearson r can be calculated as:

$$r = \sum(x - \bar{x})(y - \bar{y}) / \sqrt{\sum(x - \bar{x})2} \sqrt{\sum(y - \bar{y})2}$$

where x and y are the scores, $\bar{x}$ and $\bar{y}$ are the means, and $\sum(x - \bar{x})2$ and $\sum(y - \bar{y})2$ are the variances.

Pearson's r is widely used due to its simple interpretation, connection to linear regression models. But the limitations highlight the need to visualize data and make prudent assumptions when applying it. Used in various fileds like Correlation analysis, regression analysis, Compare risk-return profile of assets in finance

It has limitations as well the outlier can distort the results, it only measures the linear relationship.

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

Scaling refers to transforming the values of a feature in a dataset to a common scale, without distorting differences in the ranges of values. Scaling is commonly performed that normalize the data to a common scale so that features with larger ranges do not dominate features with smaller ranges. It standardize the data so that it has a mean of 0 and standard deviation of 1. This is useful for many machine learning algorithms as it can speed up convergence.

Both normalized and standardized scaling are useful, but have some key differences.

Normalized scaling preserves the shape of the distribution, standardized scaling results in a normal distribution.

Normalized scaling rescales to a 0 to 1 range, standardized scaling rescales to a mean of 0 and SD of 1.

Normalized scaling depends on the minimum and maximum values, standardized scaling depends on the mean and standard deviation.

Standardized scaling allows you to compare observations from different normal distributions where as normalized scaling only allows comparing observations within the same distribution.

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

The value of Variance Inflation Factor (VIF) may be infinite, this occurs when one independent variable can be written as a linear combination of other independent variables. In other words, one variable is perfectly predicted by the others. This results in a singular covariance matrix and the VIF becomes infinite.

If an independent variable is a constant (does not vary at all), then its variance is zero. Dividing by zero results in an infinite VIF.

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

A Quantile-Quantile (Q-Q) plot is a graphical technique to determine if two data sets come from populations with a common distribution, helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

In linear regression generally the error terms (residuals) are normally preferred distributed. A Q-Q plot of the residuals allows you to visually check whether they follow a normal distribution. If the residuals lie close to the diagonal line, they are close to normal. Any major deviations indicate non-normal errors, which conveys how well the model fits the data.

A Q-Q plot allows easy identification of potential outliers as points that detach from the diagonal at the extremes. Outliers show up as detached extreme points.

Q-Q plots of residuals validate regression assumptions, evaluate goodness of fit, spot outliers and help compare alternative models which are important aspects of a regression analysis.