Assignment-based Subjective Questions

Q.1: From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?

Answer

Following are some inferences and their effect on dependent variables:

- The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.
- instant column is a record index which does not have any significance in out analysis. So we will drop the column
- The varibles casual and registered are summed up to get cnt which is our target variable. Also during prediction we wll not be having these data, so we will drop these two variables which we are not going to use in the model.
- We are going to use weekday varible which is derived from dteday, so we will not be using dteday and will drop it.
- temp and atemp are directly correlated among each other. We will use temp and drop atemp.

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

Answer

drop_first=**True** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among **dummy variables.**

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.
Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Q.3: Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable?

Answer

Following are the few assumptions that can hold up from the pairplot with respect to Count as Target Variable:

- Ride Count seems to be very much linearly dependent on Temperature

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

Answer

Following are the assumptions to validate Linear Regression:

1. Assumption of Normally Distributed Error Terms.

2. Assumption of Error Terms being Independent.

3. Homoscedasticity

4. Multicorrelation.

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

Answer-

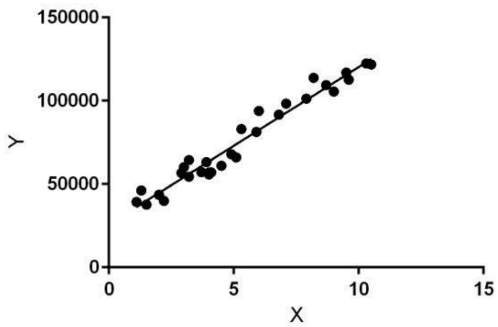
Based on final model top three features contributing significantly towards explaining the demand are:

1. Temperature
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. year .

General Subjective Questions

Q.1: Explain the linear regression algorithm in

detail.



Answer

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Updating θ_1 and θ_2 values to get the best fit line.

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \qquad minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Q.2: Explain the Anscombe's quartet in detail.

Answer

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Application:
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Q.3: What is Pearson'sR?

Answer-

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample. Following is the correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Disadvantages:

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in. In addition, the PPMC will not give you any information about the slope of the line; it only tells you whether there is a relationship.

Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

Q.4: What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?

Answer

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

Answer-

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1 - R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

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

Answer-

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also.

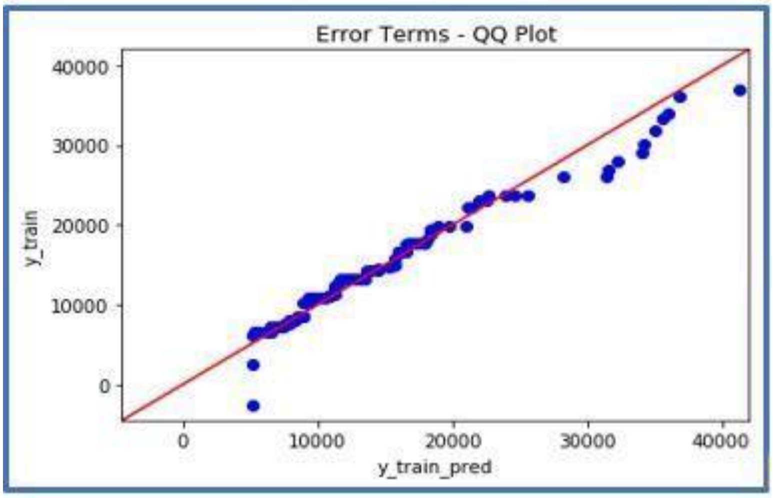
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:

If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour.

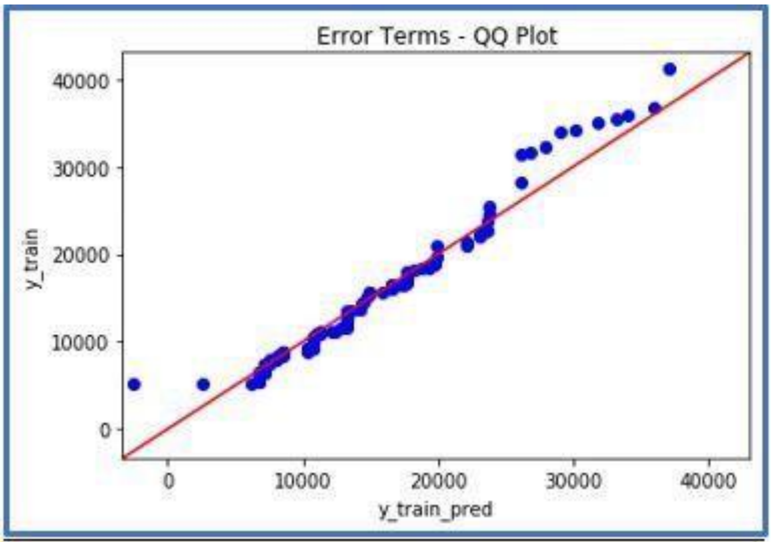
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



*c) **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.*



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis

Note: **statsmodels.api** provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.