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)

ANS. - Season 3 fall has the highest demand.

- On Holidays demand has decreased.
- Clear weather_sit has great demand
- Demand is contineously growing each month till June.
- September month has highest demand , while in other months like november , December they can do better.

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

ANS. If we do not use drop_first = True then n dummy variables will created. and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

ANS. 'temp' and "atemp" has the highest correlated values among all the others.

4.) How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANS. - The distribution of residuals should be normal and mean is 0.

- We test this residual assumption by producing a distplot of residuals to see If they follow normal distribution or not.

- we can clearly seen in diagram, residuals is around mean = 0

5.) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANS. Top 3 predicted variables For me that influence bike booking accr to our final model is temp, atemp, and yr .

GENERAL QUESTIONS –

1). Explain the linear regression algorithm in detail

Ans. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behaviour of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

2). Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3). What is Pearson's R?

Ans. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Below is a formula for calculating the Pearson correlation coefficient (r):

The formula is easy to use when you follow the step-by-step guide below. You can also use software such as R or Excel to calculate the Pearson correlation coefficient for you.

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

Ans. 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.

NORMALIZING MIN/MAX SCALLLING –

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

STANDARDIZATION SCALLING

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).

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

It seems that a some variables are able to create perfect multiple regressions on other variables (which would explain why all the VIF are infinity).

In order to identify them, I would try to do some actual regressions $Xj=X\backslash j\beta+\epsilon$ and check the coefficients in order to try to identify the problematic variables.

Otherwise, maybe this comes from your dataset. I am not familiar with chemistry data at all. According to your experience, is it common to have very high VIF in such a dataset ? Also, what is the size of your dataset ? How big is the number of observations in comparison to your 2000 regressors ?

EDIT:

Based on your comment, it is likely that the issue comes from the fact that you have way more variables than observations ($k=2000>N=45$).So all the regressions end up having $R2=1$, which corresponds to your VIF equal to 1 for all variables.

As a consequence, I suggest you try to find a way to use a small number of regressors for your regressions. A technique like forward stepwise regression could help, but will prevent you from doing inference with your resulting model. This will be fine if you are only interested in predictions.

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

Ans. 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 behavior

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.