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 :

The categorical dataset that we used were 1) season 2) year 3) month 4) holiday 5) weekday 6) working day 7) weathersit

a) Season : Bike demand was maximum during the fall and was very less during the spring
b) year : There was yearly increase in the demand of the bikes from 2018 to 2019
c) month : Monthy demand is higher months from May to October
d) Weather : Bike demand is more during the clear weather and less during the cloudy weather
e) weekday : There is no much change in demand during the week days
f) working day : Bike demand does not change based on the working and non working day

---

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

Answer:

drop_first=True helps reduce the extra column created during the dummy variable creation and hence avoids redundancy.
If not used it might create corelation among the dummy variables creating a dummy variable trap.

---

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

Answer :

atemp and temp seem to be having linear relation ship and has highest correlation with the target variable

---

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

Answer:
based on the below details
1) Normal distributions of the error
2) Less Multi-collinearity between features (Low VIF)
3) No visible patterns in residual values
4) Linear relation ship among the variables

---

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

Temperature - A unit increase in temperature leads to an increase in bike numbers.
Weather - A unit increase in weather decreases the bike numbers.
Year - Year on year there is increase in bike numbers

General Subjective Questions

---

1. Explain the linear regression algorithm in detail:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task of predicting 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
the regression line is the best-fit line for our model.

y = a1 +a2.x

here, a1 is intercept

a2 is the coefficient of x

x: input training data

y: labels to data

Once we find the best θ1 and θ2 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.

We use CostFunction to update the value of a1 and a2 to get the best fit line Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted
y value (pred) and true y value (y).

---

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that
fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots. 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 The four datasets can be described as

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

---

3. What is Pearson's R?

The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient

is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

---

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

Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling:

Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.
Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

---

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

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity.

Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
VIF = 1/1-R^2

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would meanthat that standard error of this coefficient is inflated by a factor of 2 . The standard error of the coefficient determines the confidence

interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

---

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

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 below that point and 50% lie above it.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small . for example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).