Bike Sharing

Subjective and Assignment Based Questions/Answers

Anviksha Singh
anviksha.singh0110s@gmail.com

Subjective Questions

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? (3 marks)
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Assignment Based Subjective Question

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

There were 6 categorical variables in the dataset. We have used Box plot to understand their effect on the dependent variable ('**totalRentalCount**')

The inference that we could derive were:

- **Season**: ~30% of the bike booking were happening in season3 with a median of over 5000 booking for the period of 2 years.
- **Month:** ~10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month.
- **Weathers:** ~65% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking for the period of 2 years.
- **Holiday:** ~98% of the bike booking were happening when it is not a holiday which means this data is clearly biased.
- **Weekday:** weekday variable shows very close trend (between ~13%- ~14% of total booking on all days of the week) having their independent medians between ~4000 to ~5000 bookings.
- **Workingday:** ~70% of the bike booking were happening in 'workingday' with a median of close to 5000 booking for the period of 2 years.

The above concludes to :

- Season, Month, Weathers, Workingday can be a good predictor for the dependent variable
- Holiday CANNOT be a good predictor for the dependent variable.
- weekday can have some or no influence towards the predictor., thus we have allowed the model to decide if this needs to be added or not

Assignment Based Subjective Question

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

The syntax **drop_first = True** is important as it helps in reducing extra columns created during dummy variable creation. Hence, it reduces the irrelevant correlations that would be created among dummy variables during the further analysis

Syntax for drop_first
drop_first : bool and the default value is False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

For example, if we have 3 types of values in a categorical column and we want to create dummy variables for this column. If one variable is not A and not B, then it is obvious to be C. So we do not need 3rd variable to identify the C

Assignment Based Subjective Question

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

'temp' variable which has been renamed to tempC for better understanding - **has the highest correlation with the target variable** which is names cnt in original dataset and has been renamed to totalRentalCount in Jupiter notebook calculations

Assignment Based Subjective Question

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

I have validated the assumption of Linear Regression Model based on below 5 assumptions:

- **Assumption of Normally Distributed Error Terms :** Error terms should be normally distributed. After building model, we cannot finalise until we prove the residual analysis wherein we check whether the distribution of Error is around 0 or not.
- **Assumption of Error Terms Being Independent :** Pearson Value for Predicted Value Against Residual ==> 9.645062526431047e-16 .From the graph plotted between residual and predicted values, we see that there is almost no relation between Residual & Predicted Value. This is what we had expected from our model to have no specific pattern
- **Multicollinearity check :** There should be insignificant multicollinearity among variables. This assumption is already taken care of while building model by calculating VIF of every predictor.
- **Linear relationship validation** : Linearity should be visible among variables
- **Homoscedasticity** : There should be no visible pattern in residual values. From the graph plotted between predicted and actual values we can say that residuals are equally distributed across predicted value. This means we see equal variance and we do NOT observe high concentration of data points in certain region & low concentration in certain regions. This proves Homoscedasticity of Error Terms

Assignment Based Subjective Question

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

As per our final Model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- **Temperature (tempC)** - A coefficient value of '0.4377' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4377 units.
- **Weather Situation (weather_light)-** A coefficient value of '-0.2929' indicated that, w.r.t weather_light, a unit increase in weather_light variable decreases the bike hire numbers by 0.2929 units.
- **Year (year)** - A coefficient value of '0.2343' indicated that a unit increase in year variable increases the bike hire numbers by 0.2343 units. So, it's suggested to consider these variables utmost importance while planning, to achieve maximum Booking

The next best features that can also be considered are:

- **windspeed**: - A coefficient value of '-0.1586' indicated that, a unit increase in windspeed variable decreases the bike hire numbers by -0.1586 units.
- **season_winter**: - A coefficient value of '0.0887' indicated that w.r.t season_winter, a unit increase in season_winter variable increases the bike hire numbers by 0.0887 units.

General Subjective Question

Explain the linear regression algorithm in detail.

Linear **Regression is a machine learning algorithm based on supervised learning** which performs a regression task. Regression models a target prediction value based on independent variables.

Linear Regression is mostly **used for finding out the relationship between variables and forecasting**. The major differentiation factor between multiple regression model is consideration of **kind of relationship between dependent and independent variables** and the number of independent variables being used.

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.
Consider, 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.
y = B1 +B2.x
where,
B1 is intercept
B2 is the coefficient of x
x: input training data
y: labels to data

We use CostFunction to update the value of B1 and B2 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).

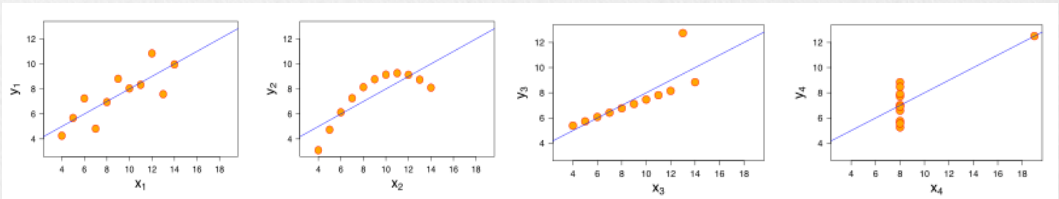
General Subjective Question

Explain the Anscombe's quarter in detail

Anscombe's Quartet can be defined as a group of four datasets 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.

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.
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



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

General Subjective Question

What is Pearson's R.

Pearson's R is a measure of linear correlation between 2 datasets. It is the ratio between the covariance of 2 variables and the product of their standard deviations; thus it is a normalized measurement of the covariance, such that the result always has a value between −1 and 1

The Pearson's correlation coefficient varies between -1 and +1 where:

- **r = 1** means the data is perfectly linear with a positive slope
  (i.e. both variables tend to change in the same direction)
- **r = -1** means the data is perfectly linear with a negative slope
  (i.e. both variables tend to change in different directions)
- **r = 0** means there is no linear association
- **0 < r < 5** means there is a weak association
- **5 < r < 8** means there is a moderate association
- **r > 8** means there is a strong association

General Subjective Question

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

Scaling 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, the **dataset contains features highly varying in magnitudes, units and range**. If scaling is not done then algorithm only takes magnitude in account and not units which leads to incorrect modelling.

To solve the above 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.

A. **Normalization/Min-Max Scaling:** It brings all of the data in the range of 0 and 1.
   - Implemented normalization in python using sklearn.preprocessing.MinMaxScaler

B. **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).
   - Implemented standardization in python using sklearn.preprocessing.scale

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

General Subjective Question

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 accurately access 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 mean that the **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**.

General Subjective Question

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 and above that quantile. For example, the median is a quantile where **50% of the data fall below that point and 50% lie above it**.

The quantile-quantile (q-q) plot **is a graphical technique** for determining if two data sets come from **populations with a common distribution**.

The purpose of Q-Q plots is to find out if two datasets come from the same distribution. A 75 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 **slope tells us whether the steps in our data are too big or too small**.
For example, if we have N+1 observations, then each step traverses 1/N of the data.
So, we are seeing how the step sizes (i.e. quantiles) compare between our data and the normal distribution.

A **steeply slope section** of the Q-Q plot **means** that in this part of the 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**