Bike Sharing System

Chaitanya Prabhune

Assignment-based Subjective Questions

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

- In spring season users are less.
- Users in 2019 increased than 2018.
- Mean count when there are no holidays is more than with holiday.
- Mean count over weekdays is same.
- Very few users prefer ride in rainy season.

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

- It helps to avoid redundancy and multicollinearity
- By excluding first dummy variable column, we prevent scenarios where the presence of One category can be inferred from absence of other.
- For example, if there is column having values 2018,2019,2020, then when we create dummy variable, if 2019 is 0 and 2020 is 0, then we can infer its from 2018. We dont need to have one more column 2018.

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

- Registered and count variable have highest correlation (0.95)

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

- We perform residual analysis for confirming our assumption
- Residual is nothing but difference between actual and predicted value and we plot this error
- If error looks to be normally distributed around 0, then the assumption is correct.

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

- Temperature
- Wethersit(Light rain)
- Year(2019)

General Subjective Questions

Explain the linear regression algorithm in detail.

- Concept : It model the relationship between dependent variable(target) and independent variables(features) using linear equation. It try to find best-fit line that minimizes the diff between predicted value and actual value
- The general equation

  $y= B0+ B1x1 + B2x2 \ldots\ldots+Bpxp + E$

  Where y is target variable

  B0 is intercept

  X1,x2,x3 are independent variables

  B1,B2…. Are coefficient for each predictor var

  E Error term

Explain the linear regression algorithm in detail.

- Objective : The primary objective in linear regression is to find best fit line that minimise sum of the squared diff between predicted values and actual values(RSS).
  RSS= Summation (i=1 to n)(yi-y^i)2
- Estimation: Coefficients are typically estimated using Ordinary Least Squares (OLS)
  β=(XT X)−1XTy
  Where
  X:Matrix of predictor values(Including column of ones for the intercept)
  Y: Vector of observed values
- Assumption: Relationship between predictors and target is linear. Observations are independent. Residuals are normally distributed. There is constant in variance of residuals

Explain the linear regression algorithm in detail.

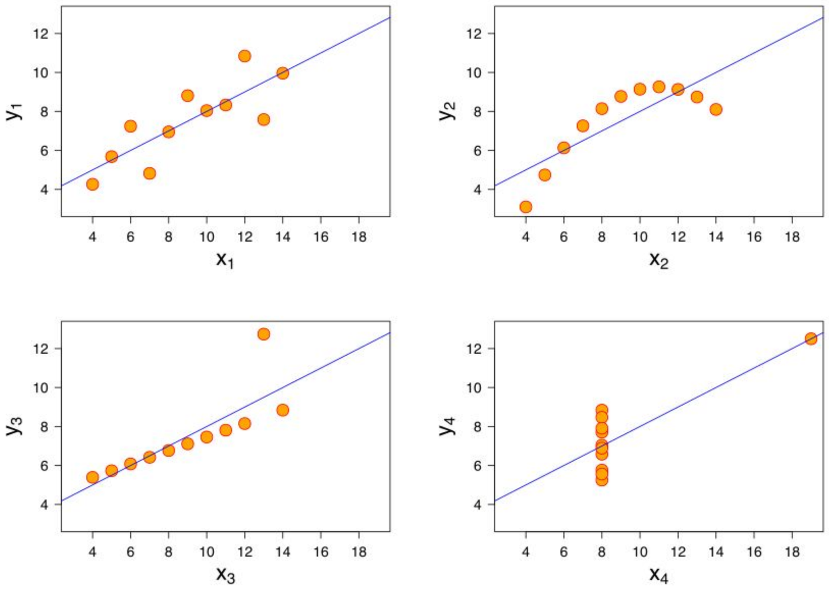
- Evaluation Technique:
    - R-squared: Proportion of variance
    - Mean Absolute Error: Avg absolute diff between predicted and actual value
    - Mean Squared Error: Average squared diff
    - Root Mean Squared Error: Square root of MSE

Explain the Anscombe’s quartet in detail.

- It is set of 4 datasets which illustrate importance of graphical representation of data and limitations of statistical measures. The quartet consist 4 diff dataset with nearly same statistical properties but very diff distribution and patterns
- Key points
    - All 4 have same mean x and y val
    - Same variance for x and y
    - Same correlation coefficient between x and y
    - y=mx +c (Linear regression line) is nearly same

Explain the Anscombe's quartet in detail.

- Visual Diff
  - DS1 : Linear distribution
  - DS2 : More Curvilinear
  - DS3 : Contains outlier
  - DS4 : All x values are same except
    One creating vertical line



Explain the Anscombe’s quartet in detail.

- It illustrates the crucial role of visualizing data
- If only relied on statistics, then it might draw incorrect conclusions

What is Pearson’s R?

- Pearsons R in Linear Regression is a measure of linear correlation between 2 vars, independent variable and dependent variable. It shows strength and direction of linear relationship between these 2 variables.
- Pearsons R or correlation coefficient is a value that ranges from -1 to +1 indicating perfect positive or negative relationship and 0 indicate no relationship.
- Formula:
  r=cov(X,Y)/(sigma x * sigma y)
  cov(X,Y) is the covariance of the variables X & Y
  Sigma x and y are standard deviation of X and Y resp

What is Pearson’s R?

- Pearsons R helps to understand how well the independent variable predicts the dependent variable. Value indicate how well independent variable interpret the dependent variable.
- However correlation doesn't imply causation. Ie. strong correlation doesn't necessarily mean one variable cause change in other

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

- Scaling refers to the process of adjusting the range of features in dataset so that they can be compared on common scale. It is important when distance between data points or magnitude of feature matters
- Scaling is performed to
  - Improve Model Performance: It make optimization process more efficient when features are in same range
  - Ensure Equal Contribution: It insures consistent contribution of each feature to model
  - Enhance Interpretability: It makes easy to interpret result

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

- Normalized and Standardized scaling
    - Normalized: It scaled data between fix range.[0,1] It adjust values to be within the range without affecting the relative diff between data points. It is also called min max scaling
      Xnorm = x - min(x)/max(x)-min(x)
      Rescale feature to specific range. It is useful when you need bounded range of features, in algos that are sensitive to scale of input data
    - Standardized: Transform features to have a mean of 0 and standard deviation of 1
      xstandardized = x - meu/ sigma where meu is the mean and sigma is standard deviation of feature
      Center feature around 0 with unit variance. Useful when data is normally distributed or when dealing with algo that assume Gaussian distribution

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

- The variance inflation factor is a measure used to detect multicollinearity in regression model. It shows how much variance of regression coefficient is inflated due to presence of multicollinearity among independent variables.
- Infinite VIF occurs when the correlation between one independent variable and combination of the other independent variables is perfect (Correlation coefficient is 1 or -1). In this model cant differentiate between the perfectly correlated variables. It happens when determinant of matrix (X’X) used to compute VIF is zero
- It indicate severe multicollinearity which can destabilize the regression coefficients. It makes them sensitive to changes to model which can lead to wrong interpretations.

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

- To avoid
  - We can remove one of the perfectly correlated variables
  - Combine the correlated variables (Principal Component Analysis)

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

- Q-Q(Quantile Quantile Plot) It is graphical tool used to analyse whether a data set follow a specific theoretical distribution like normal distribution. It compares the quantiles of data against quantiles of theoretical distribution to check if data conforms to expected distribution
- It is scatterplot created by plotting 2 sets of quantiles against one other.
- Plot construction:
    - X- Quantiles of theoretical distribution
    - Y- Quantiles from observed data
    - Point represent pair of quantiles from observed and theoretical dist.
    - If data follow theoretical distribution, the points line on straight line

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

- Use:
  Assessing normality of residuals: We assume in linear regression that residuals are normally distributed. QQ plot helps to check this assumption by plot residuals against theoretical normal distribution. If residuals are following normal distribution plot will show straight line
- Importance :
  It provides visual method for checking distributional assumption. It is essential for reliability and validity of model
  It hepls to detect deviations from normality and helps to make adjustments

Thank You!