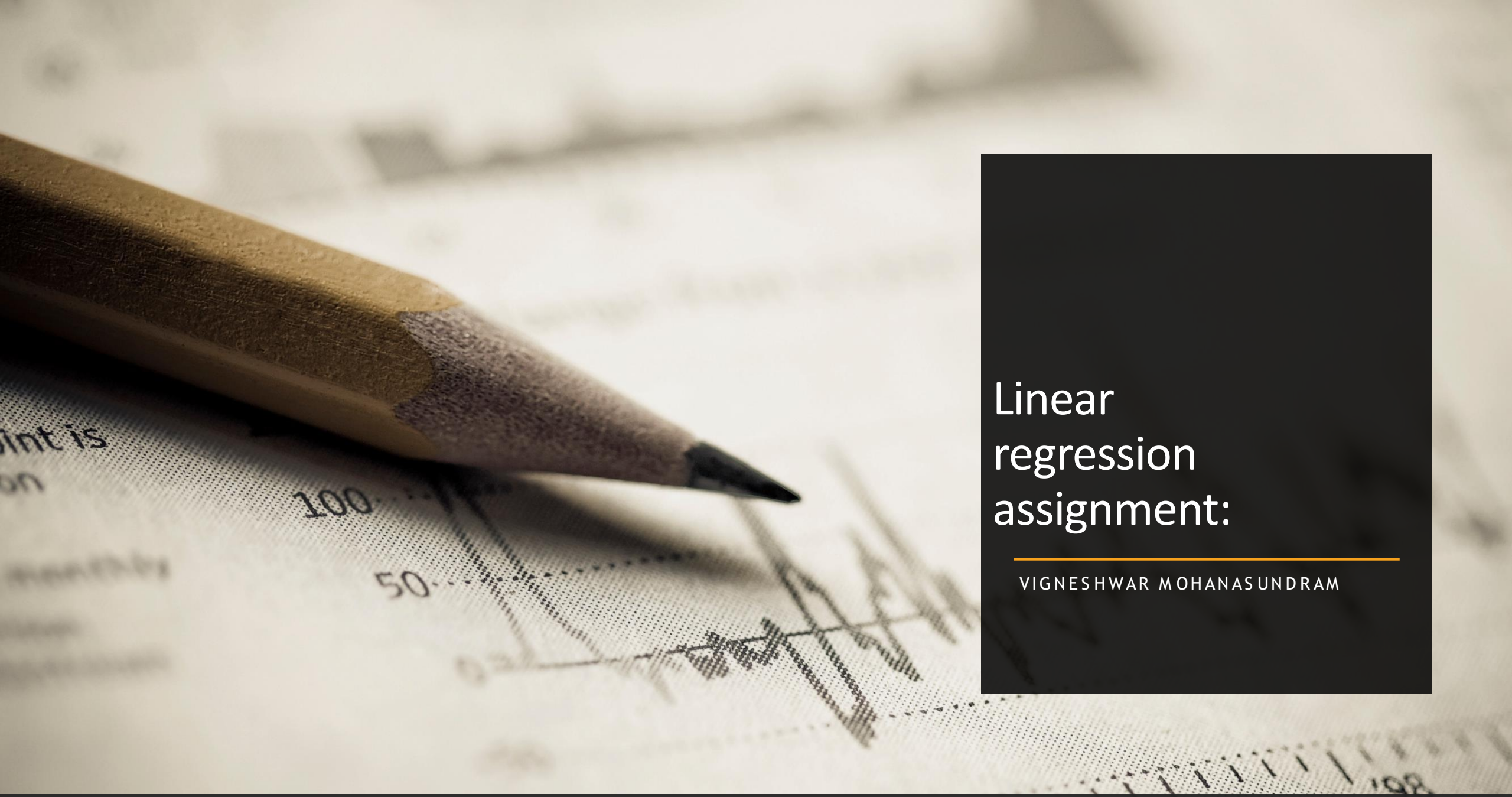

Linear regression assignment:

VIGNESHWAR MOHANASUNDRAM

Problem statement

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

A US bike-sharing provider **BoomBikes** has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.
They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

Business Goal:

We are required to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features.

They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations.

Further, the model will be a good way for management to understand the demand dynamics of a new market.

Analysis approach

- Performed EDA and analysis the basic variables
- Creating dummies and drop the unneeded rows
- Create the linear model and check for P values
- Drop column if P value is higher than limit
- Check for VIF and drop the column if VIF is higher.
- After multiple model building we found that model 11 is best fit.

Insights

- As per summary our model's R-squared model is 79.7 and adjusted R-squared model is 79.1
- Test R-squared is 77.5 and adjusted R-squared model is 77.4
- Best equation fits with respect to above summary for cnt= 0.246 X yr -0.0836 X holiday - 0.198 X Spring - 0.321 X Light snow - 0.089 X Mist+Cloudy +0.063 X 3 + 0.123 X 5 +0.151 X 6 +0.153 X 8 + 0.193 X 9 - 0.049 X Sun + 0.126 X 7 + 0.115 X 10
- Demand of bike can be say using below column values
- -yr ,holiday,Spring,Mist+Cloudy,Light Snow,3,5,6,7,8,9,10,sunday -Demands increases in the month of 3,5,6,7,8,9,10 -Demand decreases if it is holiday, Spring,Light snow, Mist+cloudy, Sunday.
- -Demand is higher in month of 3, 5 , 6, 8, 9 ,7 and 10

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?

The demand for bikes is less in the period of spring when distinguished accompanying different seasons. The demand for bikes raised in the old age 2019 when distinguished accompanying the old age 2018.

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

Dummy variables are valuable cause they allow us to use a alone reversion equating to show diversified groups drop_first=True is main to use, as it helps in lowering the extra procession founded all along mannequin changeable production.

Assignment-based Subjective Questions

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

From duplicate pair plot, we commit celebrate that temporary has the topmost certain equating accompanying mark changeable cnt .

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

We can plot the pred and test value in scatter plot and confirm also p value and VIF should be less value.

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

The top 3 countenance providing considerably toward the demands of share bikes are

Weathers Light Snow (Negative Correlation),The old age 2019 (Positive Correlation),Temp (Positive Correlation)Temp (Positive Correlation)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is an ML treasure secondhand for directed learning. Linear reversion acts the task to call a helpless variable(goal) established the likely liberated variable(s). So, this reversion method learns a uninterrupted friendship between a helpless changeable and the different likely independent variables.

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

General Subjective Questions

3. **What is Pearson’s R?**

Pearson equating coefficient (PCC) — as known or named at another time or place Pearson's r, the Pearson produce-importance equivalence coefficient (PPMCC), the bivariate equivalence,[1] or colloquially utterly as the equating cooperative[2] — is a measure of linear equivalence middle from two points two sets of dossier. It is the percentage betwixt the covariance of two variables and the product of their standard departures; accordingly it is basically a normalized calculation of the covariance, such that the result forever has a profit betwixt −1 and 1.

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

Independent variables are pre-processed to into steps by normalizing the data into particular range and its speeds us the calculation. 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.

General Subjective Questions

- Normalization uses minimum and maximum value of features are used for scaling whereas standardization uses mean and standard deviation is used for scaling.
- Normalization scales values between [0, 1] or [-1, 1] and standardization not bounded to a certain range.
- Normalization is useful when we don’t know about the distribution whereas Standardization is useful when the feature distribution is Normal or Gaussian.

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

If skilled is perfect equating, before VIF = infinity. This shows a perfect equating 'tween two liberated variables. In the case of perfect equivalence, we take R2 =1, which bring about 1/(1-R2) endlessness. To answer this question we need to drop one of the variables from the dataset that is inducing this perfect multicollinearity.

An limitless VIF advantage displays that the corresponding changing can be meant accurately by a undeviating combination of different variables (that show an limitless VIF also).

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

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

Thank you