Lesson 3:- Linear Regression.

▷ The Business Problem.

Qui:- What decision we need to be made?
What information is need to inform that decision?
What type of analysis is need to get the information need to make that decision?

How do we supply enough resources we need?
How much ticket we need per week?
we need ~~predi~~ predictive analysis to help us obtain the data to inform the decision we need to make.

▷ Approach the Busines Problem :-

What decision need to be make?
Do we have enough capacity on the Support team to handle the Support tickets from the new customer?
and if not, how many people do we need to add to the Support team to reach the desired capacity.

What information do we need to inform this decision?
we need to calculate the average number of tickets per customer per week. we can then aggregate the average no. of tickets for each customers to get

A Table Modification

a total average no. of support tickets that we predict will be submitted per week. Once we have this information, we need to compare the predicted average no. of tickets with the current capacity of the support staff, specifically, the average no. of tickets team member can handle.

Data Understanding Quiz :-

Linear Example Data :-

Client ID	Average No. of tickets	No. of employees	value contract	Indus.
TK669	90	561	200000	Retail
:	:	:	:	:

Q. which methodology would best assist in solving the business problem ?

Continuous Numeric Model.

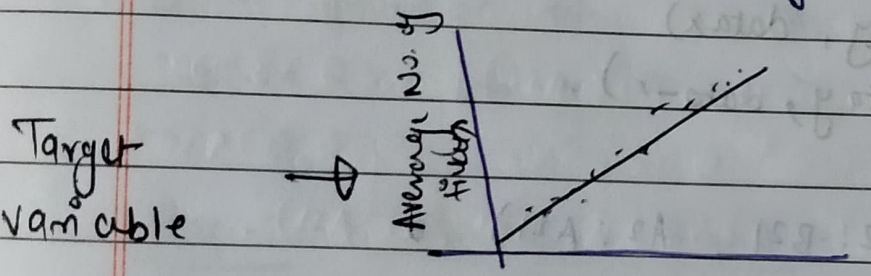
Q. Describe the steps you took to determine the appropriate model.

We need to Calculate the Average no. of tickets per customer per week (which was given) → Data Rich → We need numeric value (i.e. Numeric Analysis → continuous

based Numeric model. (Because we are not trying to predict what will happen over time).

▷ Introduction to Linear Regressions:-

Data displayed in the scatter plot, It appears that we have a linear relationship between the number of employees and the number of tickets.



$$y = mx + b.$$

Target variable:- It is the variable we are trying to understand and predict. It is also referred to as dependent variable. for ex:- we are trying to predict Y (or the average no. of tickets).

Predictor Variable:- These are used to try to predict the target variable and are also known as independent Variable.

▷ Linear Equation in Google Sheets :-

No. of employees	Average No. of tickets.
S-1	S-
68	5
:	:

= slope (data y, data x)
= intercept (data y, data-x).

= slope (B2 : B21, A2 : A21)
= intercept (B2 : B21, A2 : A21)

$$y = 0.1833x - 11.055$$

▷ Linear Regression Validation :-

- **Validation:-**

Now that we've performed the analysis and run the linear regression model, we need to validate the results of the model. In other words, is there a way to measure how good the model is? or in this case, is the linear expression we calculated a

good fit of our data ?

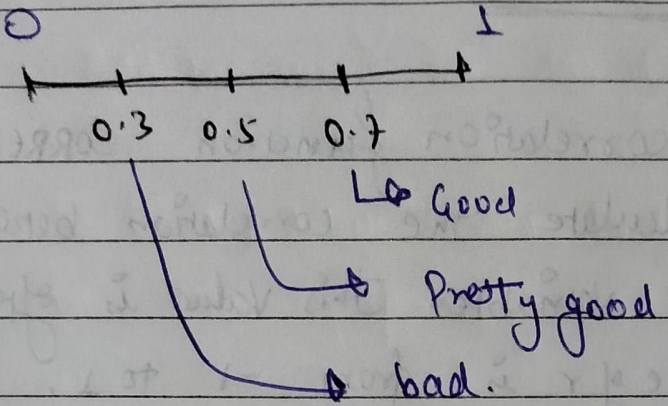
Step 1 :- Correlation :-

Using the correlation function CORREL (datay, datax), we can calculate the correlation between the target and predictor variable. This value is often referred to as r. The range of r is from -1 to 1. The closer r is to plus or minus 1, the higher the correlation between x and y.

Step 2 :- Calculate r-squared :-

While a strong correlation is good, we really want to known how well the data fits our line. Fortunately, we can get a sense of how good the formula is at approximating the data by calculating the coefficient of determination or r-squared : R-Squared is a coefficient between 0 and 1. R-Squared is interpreted as the percentage of variance in observation that is explained by the model, or the explanatory power of the model. An R-Squared value close to 1 would means that nearly all variance in the target variable is explained by the model. An R-Squared value close to zero

could means that nearly none of the variance in the target variable is explained by model.



R^2.

⇒ `=Correl(B2:B21, A2:A21)` = 0.987109

⇒ R^2 `= C5 * C5` = 0.974384.

Or `=RSQ(B2:B21, A2:A4)`.

▷ Simple Linear Regression Quiz:-

Slope = 0.181736

intercept = −7.54648.

Correlation = 0.886683

R^2 = 0.786207.

$y = 0.181736 x - 7.54648$

$y(525) = 87.865 \approx 88$.

Page No. : 24
Date : 14/01/2021

Introduction to Multiple Linear Regression :-

Linear Regression :- $y = b + mx$.

Multiple Linear Regression :- $y = b_0 + b_1 x_1 + b_2 x_2$

y = target variable.
b, b_0 = y-intercept
b_1, b_2, m = coefficient.

Initial Steps :- Preparing and Understanding your Data

Given any data set, we must make sure that the ~~idea~~ data is not clean and not biased. Once we have clean data set, the next step to prepare for your multiple linear Regression is to understand the relationship between each of your predictor variable and your target variable.

Important :- This is important because linear regression models assume that our numerical predictor variable should have a linear relationship with the target variable. It's good practice to analyze the individual variables first before you run your variable through the linear model.

Steps for Multiple Linear Regression in Excel :-

Step1 :- Make sure you have the Analysis ToolPak Add-in active in MS EXCEL.

Step2 :- In Excel, select Data Analysis. Select Regression in the pop-up window and select OK.

Step3 :- Input Y Range should be the range of your target variable

Input X Range should be the range of data of your predictor variable.

Step4 :- Click OK to see the Results.

▷ R Squared Vs Adjusted R-squared :-

The adjusted r-squared value should be used with multiple linear regressions due to phenomenon that occurs when adding additional variables to the model. In a nutshell, the more variables that are included, the higher the r-squared value will be - even if there is no-relationship between the additional variables and the target variable. There, we use the Adjusted R-squared value.