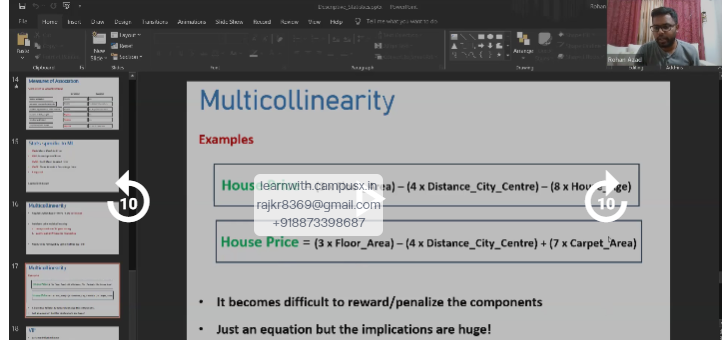
***Credit Risk Modelling Using Machine Learning***

* Problem statement:- we have customers data, and we need to predict whether we need to provide the lone or not.
* This is real world Banking Data that we are going to work.
* In this project these are the few things that is explained by the instructor:-
  + Asset = Loan product
    - housing loan
    - personal loan
    - vehicle loan
    - group loan
    - educational loan
    - credit card
  + Liability
    - current account
    - saving account
    - fixed deposit (FD)
  + Disbursed amount:- this is loan amount that is given to the customer
  + OSP:- It stands for outstanding Principle
    - 1 lakh loan and our EMI is 8000 / month and the money which is left after each payment is called outstanding principle.
    - At the end of last EMI payment, the OSP is 0 and then the loan is closed.
  + DPD:- It stands for Date past due , it means how much day is due to do payment the EMI .
    - let suppose your every month 1 you need to pay your emi but you didn’t do payment in that day then after all the dates is DPD
    - make sure your DPD is 0
  + PAR:- It stands for Portfolio at risk
    - OSP when DPD >0
  + NPA:- This is a loan account when the DPD is Greater than 90 days.
  + Credit Risk Types in Banking:-
    - DPD (zero) : NDA (Non delinquent account) = no default account = Timely payment EMI
    - DPD (0 TO 30) : SMA1 (Standard Monitoring account)
    - DPD (61 to 90): SMA2(Standard Monitoring account)
    - DPD(90 to 100): NPA
    - DPD (>180) : Written-off :- it is something loan is not present.
      * NPA improve :- loan portfolio quality of the bank will be better.= Market sentiment will be good = stock price will improve.
* NPA:-
  + GNPA: it is gross NPA. It is always in the range of(3-5%) = OSP default
  + NNPA :- It is Net LPA = (0.01 to 0.06 %) = Provisioning Amount Subtracted.
  + Bank quality assess, we always need to check the GNPA for evaluating how bank is good.

There are Two Dataset is available for this project:-

* 1. Company internal data:- In this dataset there is information for each customer for this bank ,in our case this is bank of Baroda dataset.
* 2. civil dataset :- In which the customer civil score and other information is present and based on these information’s we need to predict where we need to provide the load to this person or not.
  + The civil dataset is collected by every bank to the civil organisation in which all banks give the information of each customer civil data.
  + The range of civil score is from ( 350 to 900)
* In our dataset there are so much -9999 values , it represent null values

***EDA( Exploratory Data analysis)***

* here we have two columns, and we need to see if there is relationship between the columns or associated with each other or not
  + columns name :- MARITALSSTATUS and APPROVED\_FLAG
  + for this purpose, we are going to use chi-square test
  + the chi-square test created a contingency table where it contain all possible combination in the form of number
  + where these tests used:-
    - chi-square:- categorical vs categorical
    - T-test:- Categorical vs Numerical (2 categories)
    - Anova:- Categorical vs Numerical (>= categories)
* Multicollinearity vs Correlation
  + In Multicollinearity we calculate Predictability of each feature by other features.
  + Correlation is specific to linear relationships between columns.
  + In convex function always gives misleading values
  + 
  + There are two ways for VIF Calculation
    - sequential and Parallel