- # Summary Report

A. Workflow:

1. Importing Libraries
2. Importing Data
3. Cleaning Data
4. EDA
5. Data Modelling
6. Model Building
7. Predicting a Train model
8. Prediction the test dataset
9. Model Evaluation
10. Lead Score assigning.

B. <u>IMPORTING LIABRARIES:</u>

- ➢ The first steps is to import the all necessary libraries required for analysing the data and modelling for eg. Pandas , Numpy etc.

C. <u>Importing Data & Cleaning Data :</u>

- ➢ The second steps is to import the data from the csv. File and analyse the data for null values. While analysing the data we have observed that the form where in candidate/customer doesn't fill anything display as 'select' which we replaced by np.nan
- ➢ Further we check for null values and drop the column having more than 35% of data missing.
- ➢ Using value count we analyse data in each column and impute the missing value with appropriate variable e.g in lead source the missing data is replaced with the 'Goggle' which appears frequently.
- ➢ After imputing missing value we again check the data for null values. After this data is ready for EDA

D. <u>EDA:</u>

- ➢ In EDA we compare the categorical variable to converted (i.e customers who opt for the different courses from X education) using sns.countplot.
- ➢ we understand the most important factor which can be the indicator of customer enrolling for the course.

E. Data Modelling:

- In data modelling we again analyse the data , and drop the column which are are constant across all rows, they don't add any value to the model as well as Drop the column with heavily single-valued data.
- Change the Data type of total visit and page per visit to float for scaling purpose.
- After that we create the dummies column using get dummies and concatenate it with the data frame and result of which the main column for which dummies are created get reinstate and hence we again drop the column for which dummies are created e.g lead origin and specialization.
- After creating the dummies we split the data into train and test and using RFE model we select the top 15 features and proceed with modelling and VIF
- we try to get the model where the p value is less than 0.5 and VIF is less than 5 which represent no sign of multicollinearity.

F. Predicting a Train Model:

- In this we predict the train data set with our final dataset so we create the new data set and save the predicted values in it
- After that we do ROC curve plotting which give us the false positive rate , true positive rate and threshold than calculate auc score and map the x axis and y axis limit
- Using lead_roc we call the function and analyse the curve which will give accuracy percentage which is 89% in our case.

G. Model Evaluation:

- We find the optimal probability cut off point after creating series of points we check the possibilities of choosing any one points from 0 to 0.9. We will do this by finding 'Accuracy', 'Sensitivity' and 'Specificity' for each points. These three methods will tell us how our model is - whether it is having low accuracy or high or number of relevance data points is high or low etc.
- Afterward we observed from the data we have created points for accuracy , sensitivity and specificity for all probability points from 0 to 0.9. Out of this we have to choose one as a cut off point and it is probability cut off = 0.3 because all the accuracy , sensitivity and specificity are having nearly same value which is an ideal point to consider for as we can't ignore any one from three.
- Than we plot this data and see the convergent point or meeting point for all three point 'accuracy' , 'sensitivity' and 'specificity'.

- We observed after analysing the curve that 0.4 is the optimum point for taking probability cut-off as the meeting point is slightly before from 0.4 hence final cut-off we choose is 0.40. Also we can see that there is a trade off between sensitivity and specificity.
- We proceed with precision and recall and create confusion matrix and observed that Our precision percentage is 73% approximately and recall percentage is 77%. This means we have very good model which explains relevancy of 73% and true relevant results about 77%.
- As per our business objective, the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.
- We than again do RFE Test and draws the conclusion