SUMMARY

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to 80%.

Process Followed:

- Reading and Understanding the Data:
    - Imported all libraries required and read the data.
    - Performed basic sanity checks
- Data Cleaning:
    - Checked columns with null value and removed columns based on high null value percentage
    - Fill na in some columns like ‘Specialization’ , ‘What matters most to you in choosing a course’, ‘Country’, ‘What is your current occupation’ with ‘not provided’
    - Rechecked the NA rows for all the columns and removed them. Which were less than 1.5% of data.
- Exploratory Data Analysis:
    - Exploratory Data Analysis: Here we have performed both Univariate as well as bivariate analysis to understand the columns better and to check how they are related with each other.
    - We have also checked for outliers and removed values >0.995 on numerical columns

Below are the observation from the analysis:

- The origin identifier with which the customer was identified to be a lead was mainly “Landing Page Submission” and “API”. “Lead Add form” has a high conversion rate.
- The source of the lead is mainly Google, Direct Traffic, Olark Chart and Organic Search. People coming from Reference have a very high conversion rate.
- Last Activity performed by users mainly includes “Email Opened” and “SMS Sent” followed by “Page visited on website” and “Olark Chat Conversation”. “SMS Sent” have highly converted %.
- Most of the customers are from India.
- Most of the customers focused are “unemployed”, but “working professional” conversion rate is much higher than “unemployed”
- Since most of the customers are “unemployed” they have not selected the “Specialization”. Other Specializations include “Finance”, “HR”, “Marketing” and “Operations” followed by other specializations.
- “Total Time spent on website” has a good correlation (0.36) with “Converted”, which means people who spend higher time on websites have more chances of conversion. “TimeVisits” and “Page Views Per Visit” have high correlation (0.51), which is understandable.

    - Top 3 features which contributes to decision are as following -
        - 'Total Time Spent on Website'
        - 'Lead Origin_Lead Add Form'
        - 'What is your current occupation_Working Professional'
- Model Building:
    - Changed all the binary values as 0 and 1.
    - Created Dummy Variables for all the categorical variables.
    - Removed all the redundant variables
    - Split data into train and test with 70% and 30% respectively.
    - Used MINMAX scaler for numerical variables
    - Used RFE approach to reduce variables to 15
    - Build the first model with Logistic Regression
- Model Optimization:
    - Removed columns with P value > 0.005 and checking the VIF
    - Reduced the number of columns until all columns had P<0.005 and low VIF
    - Got accuracy = 0.80, specificity = 0.88 and sensitivity = 0.66
    - Plotted the ROC curve and got value of 0.88
- Finding the optimal cutoff
    - Plotted the graph for various cutoff for 'accuracy', 'sensitivity', 'specificity'.
    - Got optimal cut off value of 0.35
    - Got accuracy = 0.80, specificity = 0.80 and sensitivity = 0.80 on training dataset
- Running model on test dataset
    - Got accuracy = 0.80, specificity = 0.81 and sensitivity = 0.79 on test dataset
    -