Problem Statement

-X Education an online education company, faces a challenge with low lead conversion rates despite acquiring many leads daily. Leads are obtained through website interactions and referrals, and only 30% of leads typically convert. The company seeks to enhance efficiency by identifying high-potential "Hot Leads" for better focus. A model is needed to assign lead scored, indicating conversion likelihood. The goal is to increase the lead conversion rate from 30% to the CEO's target around 80%. This involves optimizing lead nurturing efforts to guide potential leads through conversion process and improve overall sales funnel performance.

Steps involved for solving problem

➔ Exploratory Data Analysis

- Columns >35% null values are dropped and outliers are managed.
- Univariate, bivariate and multivariate analyses done for insights.
- High conversion: Last activity - phone talk, lead source - referral/Welingak, origin - lead add form, occupation - housewife/working professional, last notable activity - phone talk/SMS
- Low conversion: No email preference/bounced emails, missing lead profile/specialization values

➔ Pre-processing data for model building

- It is done in 3 parts:
  1) Dummy variables creation for categorical features
  2) Splitting data in train and test
  3) Scaling using MinMaxScaler

➔ Building the Model

- LogisticRegression() with automates RFE for building the model
- Functions created for fitting and summarizing the model
- 11 features in final model: all p-values < 0.05, VIFs < 5 - acceptable
- Top 3 contributing variables:

  1) Total time spent on website

  2) Lead add from (From lead origin)

  3) Working professional (From occupation)

- Other variables in model:

  i) Do Not Email

  ii) Total Visits

  iii) Last Activity having a phone conversation

  iv) Last activity of sending SMS

  v) Olark Chat as source of lead

  vi) Welingak website as lead source

vii) Potential lead as lead profile

viii) Student of some school as lead profile

➔ Evaluating the model on training set

- Training data: accuracy ~82%, precision & recall ~77%, high specificity ~86%

➔ Making the predictions on the test set

- Test set: MinMaxScaler() applied using transform() not fit_transform()
- Model performs similarly on test set as on training set

➔ Handling business scenarios

- For aggressive sales, use low threshold to contact most leads via calls. Choose higher sensitivity (cut-off 0.2) for fewer false negatives, 89% sensitivity reduces chance of missed conversions to 11%.

- When company reaches its target and wants to focus only on high promising leads which minimizes the rate of useless phones, then company should increase the cut-off value. It will leads to very less chances of selecting lead which is not much promising. For that higher value of specificity is looked, which reduces the number of false positive. For this reason, cut-off value is selected as 0.8 for which specificity is 96.5% which means there are only 3.5% chances of selecting lead is not converted and company makes useless phone calls.