# Lead Scoring Case Study

Group Assignment

Done individually by Tushar Sharma

# Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

# Analysis Approach

**Inspecting Data**

- We started the analysis by first going through the meaning of each variable using dictionary and understanding the impact of each variable on the target variable. Here we focus more on the business aspects of the problem instead of the data.
- Then we inspected the data by going through the values available in each column.

**Handling null values**

- We observed that, in some columns like 'Specialization', 'How did you hear about X Education', 'Lead Profile', etc., there are values named select which is equivalent to a null value. So, we convert them to null value.
- Then, we checked the number of null values for each variable.

- Based on our understanding of the business problem and amount of null values, we discarded the following columns (reason for discarding are mentioned in the python coding notebook):

- Prospect ID
- Lead Number
- Last Activity
- How did you hear about X Education
- Search
- Megazine
- Newspaper Article
- X Education Forum
- Newspaper
- Digital Advertisement
- Through Recommendation
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Get updates on DM Content
- Lead Profile
- Asymmetrique Activity Score
- Asymmetrique Profile Score
- I agree to pay the amount through cheque

- Out of the remaining variables, since the null values for the columns `Lead Source`, `TotalVisits`, and `Page Views Per Visit` are in miniscule amount compared to the data size, we can comfortably remove them.
- For the remaining columns having null values, though the number of null values are large but these variables seems to be important for our analysis. Also, we can not fill these values with mean, median or mode etc., as it may lead to biased results. So, we kept these variables as such without disturbing their null values.

**Standardising data**

- Some of the categorical variables including `Country`, `What is your current occupation`, `What matters most to you in choosing a course`, `Tags`, `City`, `Last Notable Activity`, and `Lead Origin` had large number of subcategories and many of these are less than 1%. We put these subcategories under 'Others' subcategory to make analysis easier.

## Handling Outliers

- We found a huge jump between 99th percentile and maximum value specifically for `TotalVisits` and `Page Views Per Visit` columns. With help of boxplots we removed these outliers, less than 100 rows were removed in this step, hence not affecting our analysis.

## Handling Correlations

- With the help of heat map we found that some of the variables had high positive or negative correlations. So we found the need to do feature selection while building model to avoid multicollinearity.

## Model Building and evaluation

- We build our logistic regression model using statsmodel and Scikit learn libraries and we used metrices like accuracy, sensitivity, specificity, etc., along with ROC curve to evaluate the model.

# Results

- As per our model, the following features have come out as the major contributors for the problem at hand:

- What is your current occupation_Unemployed
- Tags_Will revert after reading the email
- Lead Quality_Might be
- What is your current occupation_Working Professional
- Tags_Ringing
- Last Notable Activity_SMS Sent
- Lead Quality_Not Sure
- Total Time Spent on Website
- Lead Quality_Worst
- Lead Source_Olark Chat
- Lead Source_Reference
- What is your current occupation_Student
- Tags_Interested in other courses
- Lead Source_Welingak Website
- Do Not Email_Yes
- Asymmetrique Activity Index_03.Low

- For the required condition of conversion rate of 80% (converted_prob > 0.8 or Lead score > 80), the following are the evaluation metrices:

Accuracy = 87.3%

Sensitivity = 69.3%

Specificity = 98.2%

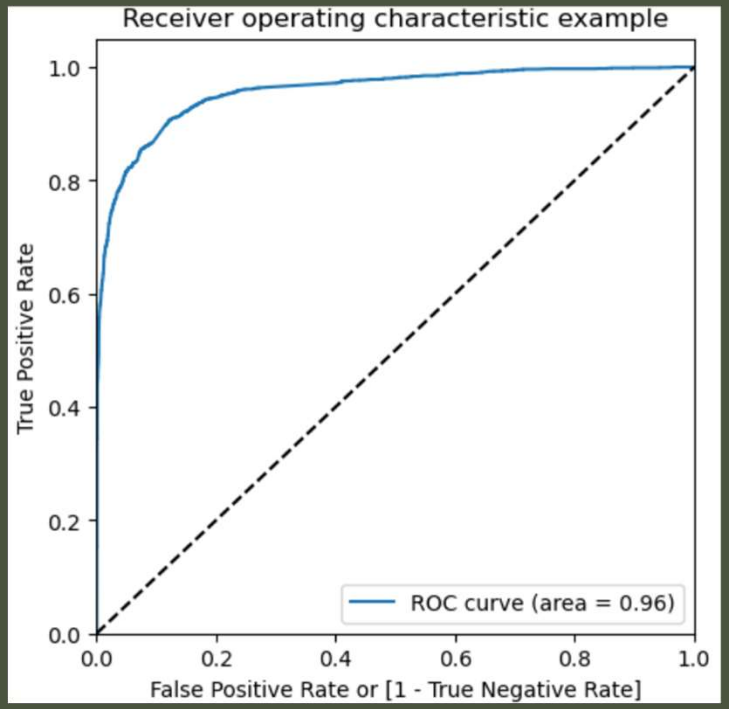
False positive rate = 1.8%

Positive predictive value = 96%

Negative predictive value = 84%

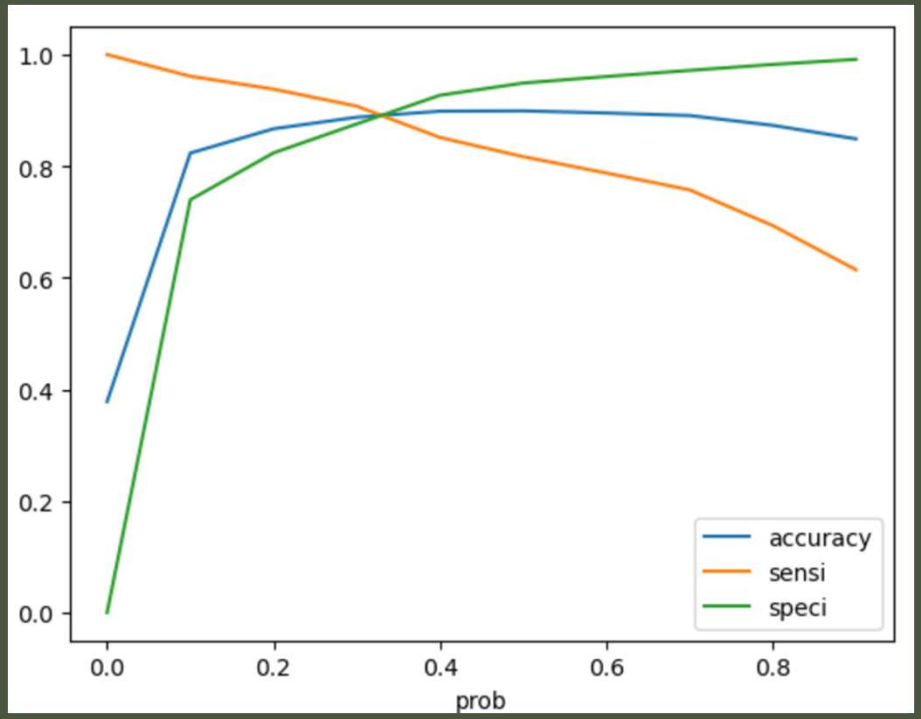
So, all the metrices are showing considerably good values, that means the model is of good quality.

- **ROC curve** for the given data is as follows:




As the curve is closer to the left-hand border and then the top border of the ROC space, the model has good accuracy.

- We also found the **optimal cutoff point** for the given data where we get balanced sensitivity and specificity.




Using the curve above, we take 0.35 as the optimum point to find cutoff probability.

- The metrices at optimal cutoff point are as follows:

Accuracy = 89.3%

Sensitivity = 86.3%

Specificity = 91%

False positive rate = 9%

Note that, here both sensitivity and specificity have optimum value, also the false positive rate is low as required.

- The predictions for the test set are also aligned with those of training set for 80% conversion rate.

Accuracy = 88%

Sensitivity = 70.8%

Specificity = 98.3%

This shows that our model is working well.

# Recommendations

- The company should give special focus to the contributors mentioned in results for increasing the number of hot leads. In particular, the following are the ones having highest positive or negative coefficients:
  - Lead Source_Welingak Website (positive)
  - Lead Source_Reference (positive)
  - What is your current occupation_Working Professional (positive)
  - What is your current occupation_Unemployed (positive)
  - Tags_Interested in other courses (negative)
  - Tags_Ringing (negative)
  - Tags_Will revert after reading the email (positive)
  - Lead Quality_Worst (negative)
  - Asymmetrique Activity Index_03.Low (negative)
  - Last Notable Activity_SMS Sent (negative)

- For the scenario mentioned in the word file, where interns are hired for 2 months by the company and they want to chase as many leads as possible, we should consider all the leads coming under the optimal cutoff range of probability > 0.35. The company will have a good number of leads and overall accuracy will also be good as shown in the results as well.
- Similarly, in the other scenario mentioned in word file, where the company just want to focus on extremely necessary leads and to avoid unnecessary phone, they should take leads with probability above 0.9 percent. Here the false positive rate will be very low as well as sensitivity and specificity would also be good, thus saving time with high productivity.

# Thank You