Project By:

- **Niharika Trivedi**
- **Nikhil Kishore**
- **Ojas Haldankar**

LEAD SCORING CASE STUDY - SUMMARY

Problem Statement

- Help is needed for an X Education to identify the most promising leads—those that have the highest chance of becoming clients.
- The business wants us to create a model in which every lead must be given a lead score, meaning that consumers who have higher lead scores are more likely to convert than those who have lower lead scores.
- The intended lead conversion rate, as stated by the CEO in particular, is approximately 80%.

Objective

- Build a ***logistic regression*** model that will allow the business to target potential leads by giving each lead a score between 0 and 100.
- A higher score would indicate that the lead is hot and likely to convert.
- A lower number would indicate that the lead is cold and unlikely to convert.

Steps Followed:

1. Data Cleaning

- After loading the dataset into the Python notebook, data cleaning procedures were performed, including duplicate detection, missing value handling, and outlier handling.
- It was discovered that every entry in the dataset is distinct.
- The features that had more than 40% of missing values were eliminated, and the remaining features were handled in accordance with standard procedures.
- For numeric columns, outliers were detected and dealt with.

2. Exploratory Data Analysis

- Three stages of EDA were carried out: target variable analysis, bivariate analysis, and univariate analysis.

- 'Landing Page Submission' was shown to be the primary source of leads, generating the majority of traffic from Google and Direct Traffic.
- The prospects were searching for "Better career prospect" and were okay with follow-ups.
- Heatmaps and pairplots were also used to find correlations between the variables.
- It was determined that the education firm's current conversion rate was 38%.

3. Data Preparation

- One hot encoding was used to encode categorical variables. Moreover, a few unnecessary columns were removed.
- Training data accounted for 70% of the total data.
- The data was scaled using the normalization approach to eliminate the impact of various units.

4. Model Building

- By creating various models, the model building step served as a comparative study.
- The model was trained and unimportant variables were removed using the variance inflation factor method and the GLM utility of statsmodels.
- Plotting the ROC curve yielded the ideal cut off probability value while maintaining a balance between accuracy, sensitivity, and specificity.
- To determine the ideal stage, recall and precision approaches were also taken into account.
- The training dataset was used to evaluate the model, and using the ideal parameter values, the predictions were made on the testing dataset.

Recommendations

The following variables are crucial in determining the conversion rate:

- **Total Time Spent on Website**: A longer visit will result in a higher conversion rate. The team should hence concentrate on optimizing the webpage.
- The conversion rate is typically lower when the customer has **not chosen to receive email follow-ups.**
- Conversion rates are higher when the client is a **working professional**.
- Leads obtained from **references** are more likely to convert.
- Conversion appears to increase when the last noteworthy action was **sending an SMS**.
- Better conversion is seen when the **Welingak website** is used as the lead source.
- **Lead Add Forms** have a higher conversion rate, but they produce less leads overall. To increase lead generation, businesses should concentrate on this source.