**Problem Assessment: Lead Scoring for X Education**

**Problem Statement**

An education company that sold online courses to industry professionals, had a typical lead conversion rate of 30%. A lead is defined as the one that fills up a form for the company and provides their email and mobile numbers. The CEO wants the lead scoring rate to be around 80%. To resolve this, a dataset was provided to perform an analysis and create a model with which sales team could target those leads that has a higher chance of conversion.

**Data Analysis and Modelling**

Steps performed were as below:

1. Data reading and Analysis:

The Data was read and then the EDA was carried out on the data. It was found that 5 columns had more than 45% of null values and were dropped, namely –

* Lead Quality
* Asymmetrique Activity Index
* Asymmetrique Profile Index
* Asymmetrique Activity Score
* Asymmetrique Profile Score

Post this there were other columns that had null values for which data imputation was carried out. A new category, **Not Specified**, was created for the columns were the null values could not be imputed as the percentage of the null values was too much too predict. These columns were identified as: **TAGS, Specialization and City**. Also it was found that for a few columns there was certain categories had a category class of **SELECT.** For these, again the values were imputed as the one for the null values. These columns were – **Specialization, Lead Profile, How did you hear about X Education and City**.

While performing EDA, it was also found that there were certain columns wherein for multiple category class had less than two percentage contribution in the whole distribution. Such Columns were identified as – **TAGS, Lead Profile, What is your current occupation, Country, How did you hear about X Education, Specialization, City, Last Activity, Lead Source and Last Notable Activity**

For these category classes, they replaced with the class of the imputed one or if OTHER category already existed then they were merged into that class or a new OTHER class was created. In a few cases the category classes were merged with already existing clases as in the case of city and specialization

There were a number of binary variables in the dataset provided. For a few the distribution was heavily tilted towards one class. The variables were not dropped as due to high number of category class, we had to use the RFE technique to filter out the variables. The variables were: *Do Not Email, Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque and a free copy of Mastering The Interview.*

1. Feature Engineering:

Post the analysis, first all the binary variables were mapped into binary values. One Hot encoding on variables with multiple category classes. Post this the data set was split in the ratio of 70:30 for training and testing the model. Also, Standard Scaler was used on the numerical columns: TotalVisits, Total Time Spent on Website and Page Views Per Visit.

1. Model Building:

RFE was performed to get the features top 15 features and a model was build that has low p-value and a low VIF. The features selected should not have the Not Specified class as these were the imputed categories.

1. Model Evaluation:

The ROC\_AUC\_SCORE was calculated at 0.96. The cut-off probability was taken as 0.4. This gave the recall, precision and F1-score as 0.89 for all. Which looked to be very well balanced.

The Lead score was also calculated.

1. Model Testing:

The model performed very good on the test data too, giving the below values:

**Precision**: 0.8525

**Recall**: 0.9342465753424658

**F1-Score**: 0.8915032679738562

**Subjective questions:**

The main parameters for the model are as below:

* Tags
* Lead Origin
* Last Activity

The three important categorical features created for the model are as below:

* Tags\_Closed by Horizzon
* Tags\_Lost to EINS
* Tags\_Will revert after reading the email

**Strategic Approach:**

* To get higher rate of leads the company can look for the lead score of 10 and above that would give 95% of recall value when the company plans to market very aggressively.
* To minimize the rate of phone call, the company would like to get a higher precision rate. This could be to get back all those customers with lead score of above 90 which would give the precision rate of 97%.