X Education Lead Prediction and Rating

1. Problem Statement and Objective:

- X Education is an online courses company that sells online courses to industry professionals. The primary business of the organization depends on the number of leads being generated and further converted into buying one of the courses.
- We worked on a model that will assign a lead score to each of these leads, such that the customers with higher lead score have higher chances of conversion. The Leads with higher score, whose chances of conversion are higher, is termed as **Hot Leads**.

2. Process Flow Chart

I developed this model **to understand the relationship between the dependent variable and the independent variables** by estimating probabilities using a logistic regression equation. This analysis will help us predict the likelihood of an event happening or a choice being made.

Here, our target variable is **Converted**, and we identified its association with other variables.

A brief stepwise flowchart of the approach is shown as below:

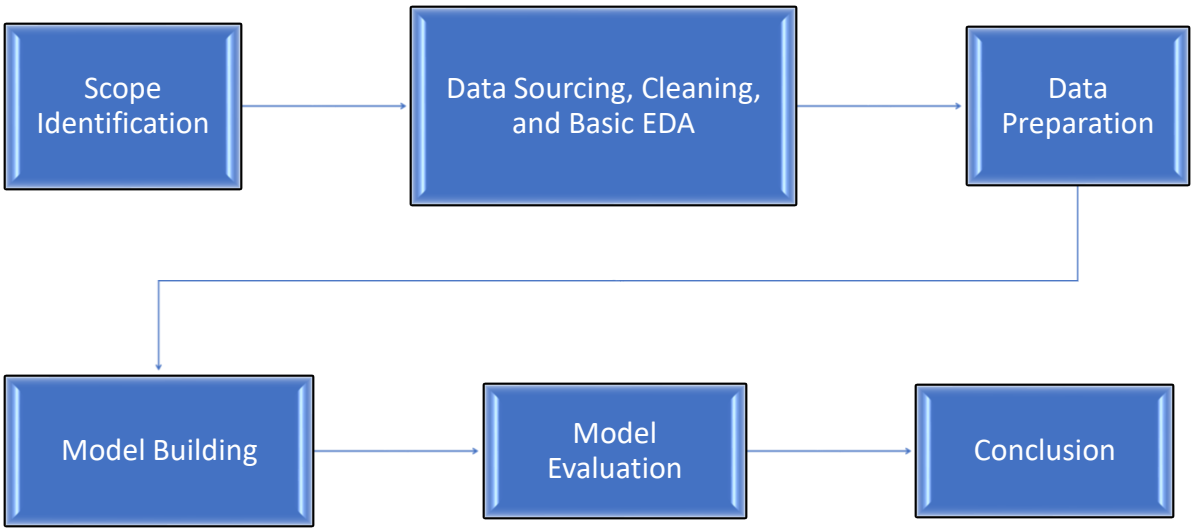

Figure 1: Process Chart of Model Creation

3. Brief Summary Report

Though the above six are the major steps involved in this project, we have broken it down for easy to follow up:

a. Scope Identification of the Project

- In this step, I Identified the scope, defined a set of Objectives considering the business goal and instruction from the X-Education Team.

b. Read and Understand the Data

- Imported the data, read it and did the basic checks like shape of data, null value, unique value, numerical description, and so on.
- Next, jumped into the basic data quality check.

c. Data Cleaning

- Started with cleaning up the data be exploring more into it.
- Identified few places where selections were not made and converted them to Null values.
- Dropped the columns where null values are more than 30%

d. Data Analysis

- Started with Skewness Check using plots and unique counts.
- Did the data imbalance check, and considered 70% as cutoff value.
- Dropped the variables with more than 70% data for a single value to eradicate the chances of error in model.

e. Data Preparation

- Dummyfied the categorical variables.
- Didn't rescale the numerical variables at this point, because we wanted our model to work equally good for non-scaled and scaled data in future.

f. Correlation Check

- Checked the correlation matrix using numerical data as well as Heatmap.
- Identified a good number of correlated variables, made a note of them.
- Didn't drop the variables at this stage, as RFE drops any such highly correlated variables in feature selection step.

g. Test-Train Split

- Split the data in 70-30 as train-test using sklearn model selection after dropping the target variable "Converted".

h. Rescaling the Numerical Features

- Rescaled the numerical features at this stage using MinMax scaling and verified the final dataframe before model creation.
-

i. Model Building Self-Evaluation

- Started with first model creation using Logistic regression Model Building. This model includes all the features we had in the latest train dataframe.

- We used RFE at this point to select 20 best features for the model.
- Recursively re-built the models to check the model parameters, with concentration over revised p-value and VIF after dropping one variable at a time.
- The Fifth Model built seems good to fit on all the parameters with all the **p-value ≤ 0.05** and **VIF ≤ 5**.
- The final model includes 17 variables.
- Created the confusion matrix and tested the model for all the evaluation parameters.

j. ROC Curve

- Validated the model with ROC Curve, an area of 0.89 is indicative of good model here.

k. Model Evaluation on Test Data

- Created a new table with 10 division of different probability cutoff (.1, .2, .3, .. .9, 1)
- Calculated the optimum cutoff probability of 0.38 based on accuracy, sensitivity, and specificity versus probability plot.
- Further, assigned a lead score to each lead based on available probability data.
- Created the confusion matrix and tested the model for all the evaluation parameters.

l. Lead Prediction using Model

- Tested the lead prediction score against the test data to validate and confirm that the total prediction conversion rate is 79.5% on test data against the 80.38% on train data. This looks good against a given target of 80

m. Conclusion

- The final model fits well on all the criteria and good to refer for sales team to predict on who should be the potential customer for a course.