LEAD SCORING CASE STUDY

(TO CORRECTLY IDENTIFY POTENTIAL LEADS)

Dibya Ranjan Mohanty
(DS C56 May 2023 Cohort)
ranjanresponsible@gmail.com
9008111765

PROBLEM STATEMENT

We have an education firm named ‘X Education’ who wants to identify potential leads with significant accuracy so that resources can be focused on only them for their conversion (them buying courses from the company).

A thorough identification of prospective leads will help the company in two ways:

Prospective leads will proper attention to help them understand the courses. Financial benefits – like discounts etc. will be focussed on these leads only

Less prospective leads will get identified. So, time, effort and money can be saved by not contacting them.

ANALYSIS OBJECTIVE

We are given with one datasets. This provides actual data of lead conversion and has other driver features as well.

The objective of this data analysis is to utilize the capabilities of logistic regression and to find out what are the leading or the prominent drivers which influence the lead conversion factor for prospective leads. Also, the objective is to fine tune the data analysis with proper cut-offs so that the analysis could deliver such sensitivity and specificity that, prospective leads are identified with significant accuracy and the conversion rate can be significantly improved.

METHODOLOGY

This logistic regression data analysis comprises of the following steps:

Step 1: Reading and understanding the data

Step 2: Cleaning and preparing the data

Step 3: Train-test split and standardizing numeric variables

Step 4: Building and evaluating the model

Step 5: Making predictions on test set

*All steps are executed and explained in the python notebook.

DATA ANALYSIS

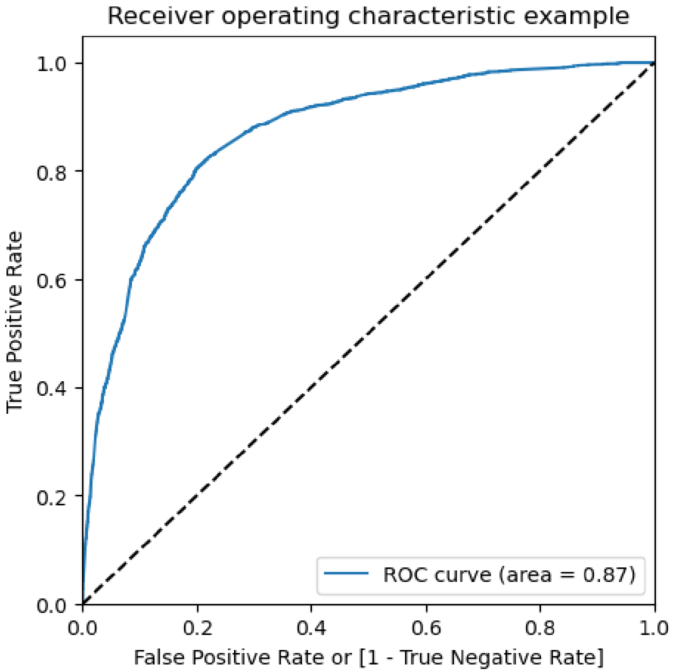
It was found that the class imbalance in the data set was not significant.

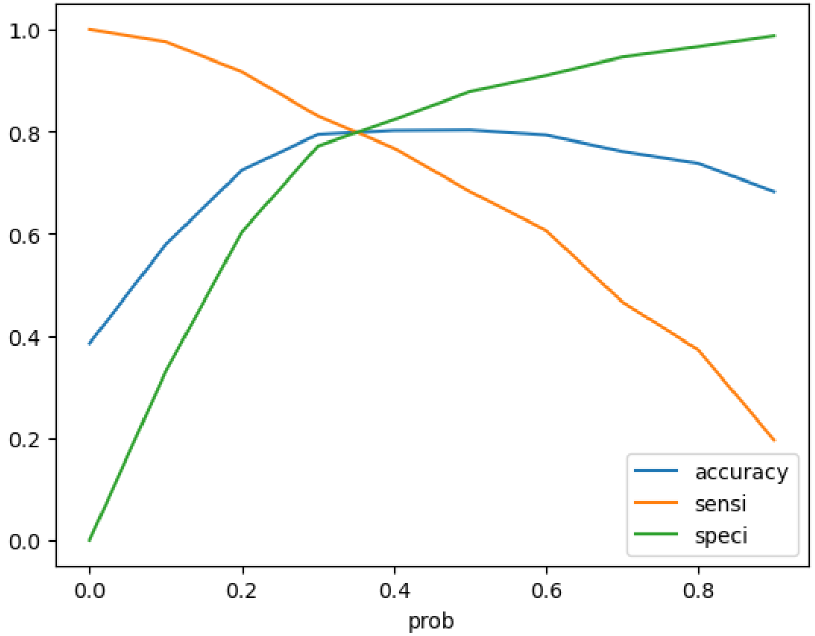
In fact, a 38%-62% target variable split is good enough to make a good model.

The null values were appropriately handled. Dummy variables were created for categorical features and standardization was performed for numeric features.

There were many driver features. To reach to the final model, features with high p-value as well as features with high VIFs were carefully and iteratively removed, ensuring that accuracy, sensitivity and specificity were not compromised.

DATA ANALYSIS





The model was able to deliver 79% accuracy, 78% sensitivity and around 80% specificity

The area under ROC was 0.87 which is indication of a good prediction model

A cut-off probability at 0.35 was arrived keeping in view of high accuracy, high sensitivity and high specificity

SUMMARY

1. Features related to phone calls were insignificant. Making phone calls is not a good driver for lead conversion
2. Welingak website and references are good sources of prospective leads
3. ‘Olark Chat Conversation’ and ‘Email Link Clicked’ features need to be improved