Summary Report
Lead Scoring Case Study

Problem Statement:

An X Education needs assistance in choosing the leads that have the best chance of becoming paying clients. The business wants us to develop a model in which each lead is given a lead score, with higher lead scores indicating a higher likelihood of conversion and lower lead scores indicating a lower likelihood of conversion. The desired lead conversion rate has been estimated by the CEO to be in the range of 80%.

Goals and Objectives:

Create a logistic regression model to provide each lead a lead score between 0 and 100 that the business may use to target potential prospects. In contrast, a lower number would indicate that the lead is chilly and unlikely to convert, while a higher score would indicate that the lead is hot and most likely to convert.

STEP 1: Reading the Dataset

Importing and reading the dataset.

STEP 2: inspecting the Dataset

Checking dimensions. Their variables types, null values percentage in each of the variables. Along with some checks on the continuous data regarding their distribution.

STEP 3: Cleaning the Dataset

Except for a few null values, the data was mostly clean. However, the option choose had to be changed to a null value because it provided little useful information. To avoid losing too much data, a few of the null values were changed to "not provided." Nevertheless, they were later taken out while manufacturing dummies. The elements were altered to "India," "Outside India," and "not provided" because there were a lot of people from India and a small number from elsewhere.

STEP 4: Exploratory Data Analysis

To quickly assess the state of our data, an EDA was performed. It was discovered that several of the categorical variables' components were unnecessary. The numerical figures are accurate, and no anomalies were discovered.

Viewing the distribution of all categorical and continuous variables against the target variable. Searching for underlying patterns. Also, keep an eye out for outliers in the continuous variables that are present and remove them. Several categorical columns were also removed due to data imbalance.

STEP 5: Preparing data for model building:

Converting a categorical variable from Yes/No to binary values 0/1. Creating dummy variables for multiple-category categorical variables.

Creating dummy variables and dividing the data into train and test data at a 70:30 ratio.

STEP 6: Feature Scaling

Every continuous variable was scaled. dividing the data of the variables with a mean of 0 and a variance of 1 Examining the target variable for any imbalances and examining the correlation of all variables.

STEP 7: Building the Model:

Initially, RFE was used to identify the top 15 relevant variables. The remaining variables were then manually removed based on VIF and P values. (VIF5 and P value0.05 were retained)

STEP 8: Model testing and evaluation:

A confusion matrix was created. Later, using the ROC curve, the optimal cut-off value of 0.35 was discovered (0.88 ROC area). Furthermore, the model's accuracy, sensitivity, and calculated values were found to be approximately 81%, 69%, and 88%, respectively.

STEP 9: Making Prediction:

The data frame was used for the prediction, which had an accuracy, sensitivity, and specificity of 80% and an optimal cut off of 0.35.

STEP 10: Precision and Recall Trade off:

On the test data frame, this procedure was also utilised to recheck, and a cut of 0.41 was discovered with accuracy between 73% and 76%.

Conclusion

According to research, the following factors affected potential purchasers the most (in descending order):

1. The total amount of time spent on the Website.

2. The total number of visitors

3. When the source of the lead was: a. a lead add form b. Olark Chat

4. When was the most recent activity: a. SMS b. Olark chat conversation

5. When they are currently employed as a working professional.

Keeping these in mind, X Education can thrive because they have a very high chance of convincing almost all potential buyers to change their minds and purchase their courses.