LEAD SCORING CASE STUDY

SUMMARY

Lead Scoring Case Study is about the X Education company who sells online courses to industry professionals and identifies the most potential leads or ‘Hot Leads’ who has good probability to join the courses. In this case study, we have build a logistic regression model to analyze the quality of leads, assign a lead score between 0 and 100 to each of the leads such that the customer with higher lead score have a higher conversion chance and lower lead score have a lower conversion chance and on this basis how we can increase conversion rate.

The dataset has been provided to us that consists of various attributes such as Lead Source,Lead origin, Total Time spent on website, specialization, current occupation, last activity etc which were useful in deciding whether a lead will be converted or not.

We have summarized the steps to be taken for analysis by X education.

1).Imported all necessary libraries like numpy, pandas, seaborn, matplot, scikit learn, scipy.

2).Data Loading and Inspecting:

Imported datasets, checking dimensions, columns, statistical summary,columns wise information, data types.

3). Data Cleaning:

Checked duplicates,dropped columns with unique values, checked null values and dropped columns with more than 35% null values, performed missing value treatment by pltting the count plot to check the data distribution of each variable column and imputing the null values with most frequent values or median.

4). EDA:

Univariate Analysis , Bivariate Analysis, calculated imbalance percentage, conversion rate,outlier treatment and analysis, checked correlation

5). Data Preparation:

Converted binary valriable to 0 and 1
Created dummy varibles for some categorical columns

6). Train-Test Split:

Split the data into 70% for training dataset and 30% testing data set
Applied standard scaling to numerical variables for normalization.

7). Model Building:

Employed RFE-Recursive Feature Elimination for feature selection of top 15 significant variables.

Assesed the model using stats model – used manual feature selection based on p-values and VIF.
Made predictions on train dataset.

8).Model Evaluation :

Plotted confusion matrix to find TP, TN, FP, FN
Plotted ROC curve , found optimal cut off = 0.357,
Accuracy = Specificity = Sensitivity = Recall = 80%
Precision = 72%
Precision- recall trade off cut off = 0.43

Variables contributing in Lead Conversion

- Specialization_Others
- Lead Source_Olark Chat
- Lead Origin_Landing Page Submission
- What is your current occupation_Other
- Last Activity_SMS Sent
- Lead Origin_Lead Add Form
- Last Activity_Olark Chat Conversation
- Lead Source_Welingak Website
- Total Time Spent on Website
- Do Not Email
- What is your current occupation_Working Professional
- Last Activity_Unsubscribed
- Last Activity_Converted to Lead

Focusing on these variables will help in targeting potential buyers willing to join the course and generating leads from these fields will increase conversion rate.