Lead Scoring Case Study

Summary Report

Introduction: -

This document is crated to brief about the procedure followed to build a model to assign the score to the leads of X Education company based on the Leads data set. Based on the model company will be able to concentrate more on paying leads by filtering the non-paying leads.

Steps: -

1. Importing all the necessary libraries

   In this section imported all the required libraries which are required for various steps of model building.

2. Read and understand the data

   In this section we read the data set with Pandas library. Performed basic checks like shape of data set to know the size of dataset, information to know the data type of all the columns, describe to understand statistical details of continuous columns in the data.

3. EDA

   First checked for null values in data set, dropped columns having more than 3000 null values in them. Then deleted the other columns which were not relevant for the model building. The columns which were important and having null values, the rows were dropped.

   Second step was to visualize the data to understand how the data is distributed in individual columns and how the columns are correlated to each other. After this we left 8 columns.

4. Prepare the data for Model Building

   Before building the model all the data should be converted to numeric.
   Categorical columns with binary data were converted to numeric with binary encoding.
   Remaining categorical columns were converted to numeric with one hot encoding with help of get_dummies function of pandas library.

   Then the data was split into test and train with 30 and 70% proportion.

All the continuous columns in the train data set were scaled using the standard scalar. After this step we left with 74 columns

5. Model Building

First, we built logistic regression model with sklearn library. Then with recursive feature elimination (RFE) selected only 15 columns which are important for the model.

With 15 features built model with stats model and started manually removing the features one by one based on the P value and VIF (Variance Inflation Factor) value.

Below 11 features are significant for the model.

TotalVisits
Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Student
What is your current occupation_Unemployed
Last Notable Activity_Unreachable

6. Model Evaluation

In model evaluation, compared the actual data and predicted data with cut off of 10 to 100% and computed accuracy, sensitivity and specificity. Then plotted all the values to find the optimal cut off (In our model the cut off is 42% and accuracy, sensitivity and specificity are above 78% on training data)

7. Making Predictions on the Test Set

Scaled test dataset with the standard scaler.
Then the model was applied to the test data with 42% cut off to check if the model works good for the unknown data. On test data all the three metrics (accuracy, sensitivity and specificity) are above 77%. This concludes the model is good for assigning the score to leads based on the data.

Then with precision and recall found the optimal value as 44%. With 44% cut off accuracy, precision and recall were above 76%

8. Summary

This section contains the recommendation for the X Education to improve the conversion rate.