Lead Score Case study

Summary Report

Problem Statement:

Create a machine learning model for an education company, having online platform for their education courses which predicts and assigns a lead score to each lead based on different variables available from historical leads. This is a logistic regression problem due to predicting the classification of the leads as converted or not, with probability of conversion to be predicted.

Methodology:

1. Data Cleaning & EDA

- Initially Data contains 9240 rows and 37 columns
- Checked for nulls in all the columns and removed the the columns where null percentage is greater than 40%
- There are 37% missing values present in the Specialization column. It may be possible that the lead may leave this column blank if he may be a student or not having any specialization or his specialization is not there in the options given. So, created another category 'Others' for this.
- Imputed Missing values in columns like “Tags column”, “What is your current occupation”, 'Country’, 'City’,.
- There were a few columns in which only one value was majorly present for all the data points. These include Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque, What matters most to you in choosing a course. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- After Data cleaning, the dataset contains 9074 rows and 13 columns.
- Analysing Categorical Variables w.r.t Count variable.
- Analysing Numerical Variables w.r.t Count variable.

2. Data Preparation for Model Building & Model Evaluation

- Dummy variable created for the categorical variables and dropping the first one.
- Concatenated the dummy data to the lead data data-frame.
- The dataset contains 9074 rows and 13 columns
- Split the data into train and test data sets in 70:30 ratio.
- Scaled the numerical features using standard scaler.
- RFE to select the columns and build Logistic regression Model.
- Ensuring all p values tending to 0 and VIF values are low, arriving at the final model with n number of variables.
- Making predictions on Train data set based on assumed probability 0.5.
- Calculated the sensitivity, specificity, accuracy of model on train data set with predicted prob of 0.5
- Found optimal cutoff point using ROC curve, 0.36 is the optimum point to take it as cut-off probability
- Recalculating the sensitivity, specificity, accuracy of model on train data set with predicted prob of 0.36
- Evaluating model using Test data set and comparing against Train data set results.

3. Conclusions and Recommendations for the Company Strategy

- The model Evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values. The Recall score is a bit greater than precision score as well. This fits the business needs for the future.
- Top features for good conversion rate:
  1. Lead Origin_Lead Add Form
  2. What is your current occupation_Working Professional
  3.Lead Source_Welingak Website
  4. Total Time Spent on Website
- In order to increase the time that a user spends on webpage, the company to employ more web developers and UI/UX designers to improve the experience for the user and thereby luring the users to spend more time on the webpage, exploring the contents.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- It is also important not to waste a lot of time on some factors that do not contribute much or negatively affect the lead scores