Lead Score

Executive Summary

Contributors: Abhishek G, Ritesh K, Sanket
02/01/2023

Introduction

This analysis is done for X Education which sells online courses to working professionals. The X EducationX gets a lot of leads, however its lead conversion rate is very poor. So, we have built a machine learning model based on logistic regression which helps to determine whether a lead is actually a hot lead or not based on top features variables and where the lead score is greater than 80%.

This will actually improve the online sales of the X Education by converting the leads into the paying customers.

Data Provided

- Leads.csv
- Leads_Data_Dictionary.xlsx

Steps Involved

Data Cleaning:

The dataset "Leads.csv" actually contains 9270 rows and 37 columns. Analyzing the dataset we came to know that there were a lot of missing values in the dataset. So, we decided to drop the columns where there are missing values greater than 3000 count as they don't help in analysis. Similarly, we have also dropped multiple categorical columns where the data was highly imbalanced

So, after all the data clean-up we had 6373 rows i.e., we could retain ~69% of the actual data and that is good enough to build the model

Exploratory Data Analysis

Performing a univariate and bi-variate analysis, we could infer the below from the existing data

- The "Landing Page Submission" in the Lead Origin feature has high non-converted count followed by "API"
- The lead source from "Direct Traffic / Google" has high non-converted rate

- There is no big difference w.r.t Converted feature hued against “Do Not Email”
- There is a good converted count for the variable “SMS Sent” in the feature Last Activity
- The unemployment category has highest count of non-converted count

Dummification:

We have taken all the categorical variables/feature/column and transformed it into several columns and performed a concat to the original dataframe

Train-Test Split:

The split was done at 70% and 30% for train and test data respectively and the randomstate is set to 100

Scaling:

There were a couple of numerical variables/features/columns excluding the target column “Converted” and we have normalized the values by using the MinMaxScaler which scales the data from 0 to 1.

Correlation Matrix:

We tried to look at the correlations by plotting a correlation matrix however there were many features that have been created after the dummification. Hence, it turned out little difficult to understand from the matrix

Model Building:

So, RFE was used to attain the top 15 relevant features. Later, we have analyzed the P-Values of each of the features and before dropping the features with P-Value > 0.05. We calculated the VIF for each of the above features and removed the feature which has both P-Value (> 0..05) and VIF (5) higher. We repeated the same process to build a better model by eliminating the features where it has higher P-Values and VIF score.

Finally, we have arrived at a good model where all the features are less than the range indicated above

Model Evaluation:

The prediction of the Converted feature is made with the cut-off point as 0.5. We have created the confusion matrix from which we could calculate the accuracy, sensitivity and specificity of the model. The accuracy was close to 79% which is a good model.

However, we proceeded further to find the optimal cut off value (using ROC curve) and it turned out to be ~0.42 and we have predicted the converted feature based on the updated cut-off value. We reevaluated the confusion matrix and the accuracy seems to be very close to the initial evaluation.

Test Data Prediction:

Transformed the test data set and prediction was done on the test data frame and the accuracy of the dataset is close to the train set which actually tells that the model is good.

Precision and Recall:

This method was also used to recheck and a cut off of 0.44 was found with Precision around 77% and recall around 76% on the test data frame

Observations:

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spent on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. When the last activity was:
    a. SMS
    b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.