Lead Score Case Study using Logistic Regression

Submitted By

Pritkumar Parmar

Lead Score Case Study for X-Education

- Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Strategy

- Source the data for analysis
- Clean and Prepare the data
- Exploratery Data Analysis
- Feature Scaling
- Splitting the data into Train and Test dataset
- Building a Logistic Regression Model
- Evaluting the model by different criteria- Accuracy Score, Sensitivity, Specificity, Precesion-Recall
- Applying the test dataset for the best model by Sensitivity and Specificity Matrics

Problem Solving Methodology

Data Sorting, Cleaning and Preparation

- Read the Data From Source
- Convert data into clean format for analysis
- Remove Duplicate data and outlier Treatment also

Feature Scalinf and splitting

- Feature Scaling of Numerical Data
- Splitting data into train and test split

Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various matrics and criterias

Exploratery Data Analysis

- For the Exploratery Data Analysis, used a **sweetviz** Library, so that we can do all the data visualization detailed analysis and this file will be in html format.
- Syntax of the library is as follows :
- !pip install sweetviz
- import sweetviz as sv
- sweet_report = sv.analyze(df)
- sweet_report.show_html('sweet_report.html')

Analysis from EDA

- We have around 39% Conversion rate in Total.
- The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visi.
- In Lead Origin, maximum conversion happened from Landing Page Submission.
- Major conversion has happened from Emails sent and Phone Calls.
- Major conversion in the lead source is from Google

Variables impacting on Conversion Rate

TotalVisits
Total Time Spent on Websit
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Housewife

What is your current occupation_Student
What is your current occupation_Unemployed
What is your current occupation_Working Professional
Last Notable Activity_Had a Phone Conversation
Last Notable Activity_Unreachable

Observations

Train Data

Accuracy : 78.88%
Sensitivity : 73.84%
Specificity : 83.56%

Test Data

Accuracy : 78.71%
Sensitivity : 76.74%
Specificity : 80.52%

Conclusion

- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission.
- Leads who are spent more time on website is to be converted on successfully entolled for the course.
- The top 3 variables are in the dataset which converts the leads to be converted are Total time spent on website, Lead add from lead origin, Had a phone conversation on phone call.
- Hence the model seems to be good.