Lead Scoring Case Study

Submitted by:
Jyoti

Problem Statement

- X education data and they sell online courses to industry professionals. People interested in the course land on the website and browse courses.
- 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.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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%.

Steps Followed

- First will take the dataset, understand it and checks the columns description to understand all the parameters involved in the problem.
- After importing the dataset and the libraries will find out the shape, info, description, which will give us the no of rows, columns, null values, mean, median, standard deviation.
- The next step is Data Cleaning, which is the important step, to check the % of null values and how we can impute those values, dropping the columns which are not relevant to the problem at the time.

- So, after checking the missing values, dropped the columns which were having missing values more than 40% for further positive analysis.
- At the start, dataset had 37 columns, after dropping 40%, was left with 30 columns.
- Dropped Lead Number and Prospect ID as well, because these are just generated for the reference of the lead.
- Then, ‘Select’ was present in most of the columns so we have replaced it with NaN.
- Dropped 4 more columns which didn’t seem relevant for the analysis, Now left with 26 columns.
- Some of the columns(‘’Country, Specialization, How did you hear about X Education ,What is your current occupation, What matters most to you in choosing a course’’), still have null values present, we looked into them individually to see what can be done.
- We did value counts and imputed the columns with the mode as there are categorical columns.

- Dropped Country column because in country column- India and Nan combine will give you 97% of the distribution, so its safe to drop it.
- Dropped How did you hear about X Education as well because of the null values which were 78%.
- I have imputed categorical columns with mode and combined the columns for easy and better analysis.

- Now, we are left with 23 columns and all the columns have 0 null values.
- We checked the conversion rate at the start, which is 45.5%.

Exploratory Data Analysis

- cat_cols=['Lead Origin','Lead Source','Do Not Email','Do Not Call','Last Activity','Specialization','What is your current occupation','What matters most to you in choosing a course','Search','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations','Receive More Updates About Our Courses','City','A free copy of Mastering The Interview','Last Notable Activity']
- con_cols=['Converted','TotalVisits','Total Time Spent on Website','Page Views Per Visit']

Univariate analysis of categorical columns:-

- Lead origin is mostly **landed on the page submission** than API.
- Majority of the leads are from **GOOGLE,** then direct traffic , Olark Chat and organic engines.
- Most people said **NO** to **email and calls.**
- **Email opened** was the **highest activity performed** by a customer then SMS sent comes after it.
- **Management courses are more in demand** and after that customer haven't specified the **specialization**.
- **Unemployed** tops in occupation working professional comes afterwards.
- **Better Career Prospects matters** most to the customers who wants to choose a course.
- Big **NO** to **search , Education forums, Newspaper, Digital advertisement.**
- **No one** has come through **recommendations.**
- **No one** wants to receive updates about the courses.
- Most people have **not mentioned their City** after that Mumbai comes in the list.
- **No** to A **free copy** of Mastering The Interview.
- **Modified, email opened and sms sent** was the **Last Notable Activity** for most of the customers.

- <u>Box plot of continuous columns:</u>- As we can see from box plot of continuous columns there are no such ouliers present in the data.Though 'total visits' and 'Page Views Per Visit' has some outliers.
- Treatment of outliers by removing top and bottom 1% of the outlier values.

Bivariate analysis

- Bivariate analysis with target column'converted'.
-
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
No
Yes
Do Not Call

Converted
0
1
count
2500
2000
1500
1000
500
0
No
Yes
Do Not Email

Converted
0
1
count
1200
1000
800
600
400
200
0
Page Visited on Website
Email Opened
Others
Converted to Lead
Email Link Clicked
Form Submitted on Website
Olark Chat Conversation
Email Bounced
SMS Sent
Last Activity

Converted
0
1
count
1600
1400
1200
1000
800
600
400
200
0
Not given
Business_courses
Media and Advertising
Management_courses
Industry_courses
Specialization

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Other
What is your current occupation

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course

Converted
0
1
count
1200
1000
800
600
400
200
0
Not given
Mumbai
Other Metro Cities
Maharastra_cities
Other Cities
Tier II Cities
City

Converted
0
1
count
1200
1000
800
600
400
200
0
Modified
Email Opened
Page Visited on Website
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Email Bounced
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Email Received
Email Marked Spam
Last Notable Activity

Inference:-

- landed on the page submission and API will mostly get converted.
- leads are from GOOGLE, direct traffic ,Olark Chat, organic engines and through reference gets converted.
- People who have Opened email and sent sms mostly gets converted.
- Management courses specialization and not mentioned** anything mostly gets converted..
- Unemployed & working professional will get converted.
- People who wants a Better Career Prospects will choose a course.
- Most people have not mentioned their City, Mumbai and other mahatrastra_cities will take a course.
- People who has done Last Notable Activity- Modified, email opened and sms sent will get converted.
- who said no to Search, newspaper, digital advertisement will get converted.
- who said no to Do not email or calls will get converted.

Numerical columns with Target column

Inference:

- Median of TotalVisits - converted and non converted leads is almost same. Can't say anything on the basis of median.
- Leads **who spends more time on website** will likely to get **converted** so, company should engage their customers on website nby making it more interesting.
- Median of Pages views per visit - converted and non converted leads is almost same. Can't say anything on the basis of median.

Data Preparation

- Converted into binary columns.
- Created dummy columns.
- Then, train-test split the data into 70-30% split.
- Then, Recursive feature elimination and went with 15 varaibles.

Model Building

- Started with model building, checked the p -values and VIf.
- Then if the p values look higher than the range 0.05 dropped that column and did this again and again until we have got p -values and VIF under control.
- So, we build 7 models in total and then we freeze the process and went ahead with the analysis on the model 7.

- Then we predicted the values and made confusion matrix, checked the accuracy , sensitivity, specificity, positive and negative predictive values.
- Plotted ROC curve and we have got the area of 0.87.
- Find the optimal cut-off point - created columns with different probability cutoffs, calculated the accuracy, sensitivity, specificity on various cut-offs.
- Plotted the graph of various cut-offs.
- And, chose 0.4 the optimal cut-off point.

- Then, taking 0.4 the prob, created a lead score column.
- Again calculated the accuracy, sensitivity, specificity.
- And, we got , Accuracy=79.8%, Sensitivity=73%, Specificity=85.9% on the train set.
- Precision and recall values are 81%, 73 % resp.
- Then we plotted precision_recall_curve and again checked the precision, recall values.

Predictions on Test-set

- Transform the test set, predicted the values on the basis of optimal prob, created lead-score column.
- Calculated the accuracy, sensitivity, specificity, precision and recall.

Comparison of Train -Test set

<u>Observations on Test Set:</u>

- Accuracy = 80%
- Sensitivity= 81.5%
- Specificity = 79.1%

<u>Observations on Train Set:</u>

- Accuracy=78.9%
- Sensitivity=71.9%
- Specificity=85.6%

So, we have got a good model of accuracy around 80% and we can present this the CEO.

Summary

There are a lot of leads generated in the initial stage but only a few of them come out as paying customers. In the process to convert them into hot leads, you need to polish the potential leads by educating the leads about the course, what are their requirements, what is their aim and constantly communicating them. First, churn out the best prospects from the leads you have received. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit', 'Last Activity', 'Last Notable activity' which contribute most towards the probability of a lead getting converted. Then, you should keep a list of leads in reach to inform them about new courses, services, job offers and future higher studies. Check each lead carefully so that you can tailor the information you send to them. A fine plan to check the requirements of each and every lead will go to convert them into as prospects. Focus should be converted leads. Have a video session with them , let them ask questions as much as possible to clear out all their doubts. Give them time they need to compare and check .Then at last, make further inquiries with the leads to determine their intention and mentality to join online courses.

“ThankYou.. this is all from my end! ”