• Summary Report

Our Business Problem was: An education company named 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%.

Our Business Objective is: Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

We proceed with following steps :

1. Data Importing: Sourcing the required libraries

2. Data Reading, Interpreting & Understanding :

Reading the dataset "Leads.csv" and Understanding it :-

a. Performing basic data checks like shape (number of rows and columns), info (data type of each column and null values with percentage), describe (mean & median for all numerical columns), converting data types wherever required, etc.

b. Missing value checks and filling them with mean, median, mode or percentiles.

c. Outliers Analysis.

d. Check for duplicates.

3. Data Cleaning: In case studies with machine learning models, Data cleaning plays an important role. The quality and efficiency of the model depends on the data cleaning. Hence it must be done preci

a. “Select” values, which are considered as null is replaced with NAN.

b. Calculation of missing values for each column and dropping columns named Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, and Asymmetrique Profile Score.

c. Dropping the columns with high percentage (60%) of missing values.

d. Mapping binary variables ‘Yes’ and ‘No’ to corresponding 0’s and 1’s.

e. If the columns are highly skewed with one category, such columns will be dropped like Country(India) . Combining different categorical levels in the columns with less percentage values into "Others" category.

f. Dropping the rows with least missing values percentage for that column.

g. Analysing each column separately to check for missing values imputation manually like lead quality.

h. Checking the final dataset for number of rows and columns kept after performing all of the above steps.

4. EDA : As part of EDA**, 1**. **Univariate and Bi-Variate analysis** was done on both numerical and categorical variables and checked for outliers. **2**. **Outlier Treatment**: We performed capping of lower and upper range outlier values for TotalVisits and Page View Per Visit columns. **3**. We also replaced columns with low percentage of levels with one single level ‘Others’ and **4**. dropped columns with like 95% values with No’s since they were highly imbalanced, thus adding no value to the model.

5. Data Preparation : In this step, we imported few essential libraries. We performed Data Preprocessing, dummy variables creation, train test data splits are performed and scaled the numerical columns using standard scaler.

6. Model Building & Model Evaluation:

a. Before model building, we had 31 columns and we created generalized linear model taking into account all variables. Since there were lots of variables to deal with, we moved to automatic feature selection.

b. We used both RFE and manual elimination methods to get the final list of columns. During which the most insignificant, highly correlated columns were dropped & lastly we had 13 columns in our final model.

c. We know that the relationship between ln(odds) of 'y' and feature variable "X" is much more intuitive and easier to understand. The equation is:

y = -1.9863 * const + 2.0380 * Lead Source_Welingak Website + Last Activity_Email Bounced * -1.8657 + Last Activity_SMS Sent * 2.1375 + What is your current occupation_Unemployed * -3.1890 + Tags_Busy * 2.5655 + Tags_Closed by Horizzon * 8.0682 + Tags_Lost to EINS * 10.6212 +Tags_Ringing * -1.7877 + Tags_Will revert after reading the email * 4.8001 + Tags_switched off * -2.1994 + Lead Quality_Not Sure * -1.7327 + Lead Quality_Worst * -3.5567 + Last Notable Activity_Modified * -1.4413

d. Created the confusion matrix and evaluated accuracy, sensitivity and specificity scores for training data. We then went on to create Roc curve and in order to find optimal cut off point, plotted accuracy sensitivity and specificity for various probabilities.

e. We chose the cutoff probability as 0.40 from Accuracy, Sensitivity, Specificity curve and calculated lead score for all the leads. The sensitivity of model (on test data) was around 90% and the conversion rate increased from 38% to 73%.

7. Conclusion:

From our logistic regression model, we coclude:

- Major focus should be on Working professionals.
- Major focus on leads whose last activity is SMS sent or E-mail opened.
- Good to focus on customers who have spent significant time on our website.
- If the leads are referral, they may not be potential leads.
- If the people didn't filled specialization or chose others, they may not know what to study and are not right people to target. So it is better to care less for such cases.
- The customer who fills the form are potential leads.

8. Recommendations

- Since the company hires some extra helping hands, they should focus more and reach out to people who have low probability of conversion to help improve the overall conversion rate.
- Focusing more on leads showing lower conversion rate will help in improving overall conversion rate.
- It's good to collect data more often and get in touch with potential leads. It is believed that the best time to connect with potential leads is just after few hours the leads show interest in your product or service like once the link is clicked sent via email or sms and showed interest in the content by spending time on your page.

- While mailing it is good to send personalized messages to particular set of leads as it will have positive impact on leads.

- Try reducing the no. of attempts such as phone calls to get the leads converted, for ex: if it usually takes 5-6 phone-calls to get them converted try reducing it to 3 to 4 by taking appointments and calling, sending emails and providing right information and keep the leads in touch, hence there will be more time available which can be used to convert more leads.

[Created by : Saurabh Ghavghave, Santosh K and Rohit Rokade]