Lead Scoring Case Study Summary

By: Kanika Khattar Ahuja and Divya Namani

Problem Description:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
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 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.

Solution Approach:

Going by the problem statement,A logistic Regression model has been built to classify whether a student whether a lead is going to be converted.The below solution methodology has been adopted to arrive at the best model that can assign a Lead Score between 0 to 100 based on various features.

Data Analysis:

From the past data,only 38.5 % leads got converted.Hence,a thorough EDA has been conducted to identify key drivers that can help achieve a conversion rate of atleast 80%.

Handling Categorical variables:

1.Variables with high skewness have been removed.

2.Handling Missing values and deriving new categories:

- All the missing values (NaN,Select) have been clubbed as they both imply the same value and columns with >45% missing values have been dropped as imputing may result in biased inferences.
- For columns like Lead Source which is related to Lead Origin,the missing values are imputed based on the mode value of the Lead Source for each Lead Origin Category.

- The missing values in the column – Country have been imputed based on the column-City as there is a logical relationship between both these variables.
- Low frequency Categories in field –Lead Source and Missing values in column-Specialization have been replaced with “Others” to avoid information loss.

3.Columns describing comments from the sales team have been dropped as future data for which the model assigns the predicted lead score may not have that information and thus may result in inappropriate results.

4.All the categorical columns have been visually analysed using boxplots,bar charts as needed.

Handling Numerical variables:

1.Of the 5 numerical variables,the two variables - Asymmetrique Activity Score, Asymmetrique Profile Score have > 45% missing values and hence dropped from analysis.

2.For the remaining 3 columns,the missing values in the 2 columns-Total Visits,PageViews Per Visit,missing values have been imputed using the median(mean was also close to median) of the columns.

3.The outliers in these columns have been visualized using boxplots and have been treated by soft capping the 99% value.

The following variables considered for modelling are:

- Lead Origin
- Lead Source
- Last Activity
- Country
- Specialization
- Occupation
- Free copy required
- TotalVisits
- Website Time
- Page Views per visit

Preprocessing and Data Preparation

Dummy variable creation for categorical variables:

Binary variables with values Yes/No have been converted to 1/0.

Dummy variables for different categories have been created and the dummy columns of category “Others” have been dropped.

Train test split

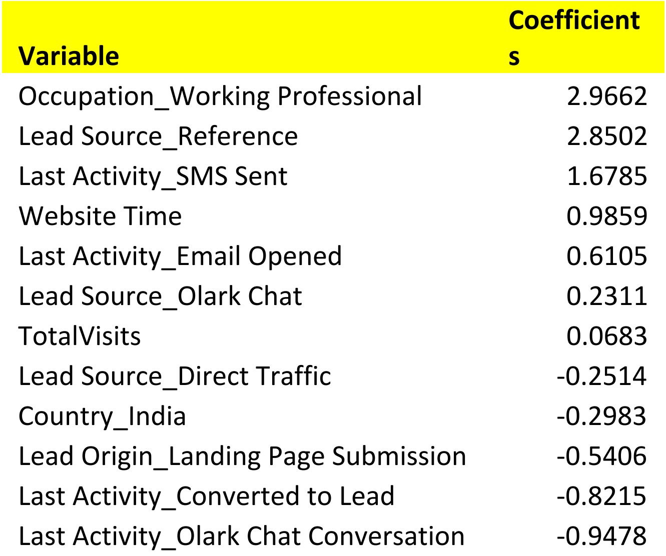
Train size of 70% data and test size of 30% data was used.

Scaling data

Standard scaling was used to scale all the numerical variables.

Model Building

- For features identification, a combination of manual and automated(RFE) approach have been adopted.
- A logistic regression model was built and was finalized with 13 variables with VIF < 5 and p-value < 0.1.
- Optimal cutoff of 0.35 was decided based on accuracy,sensitivity and specificity trade off.
- The model predicted the probabilities which were further used to calculate the Lead Score.
- The following features which ultimately helped in deciding the conversion of a lead are:



Variable	Coefficients
Occupation_Working Professional	2.9662
Lead Source_Reference	2.8502
Last Activity_SMS Sent	1.6785
Website Time	0.9859
Last Activity_Email Opened	0.6105
Lead Source_Olark Chat	0.2311
TotalVisits	0.0683
Lead Source_Direct Traffic	-0.2514
Country_India	-0.2983
Lead Origin_Landing Page Submission	-0.5406
Last Activity_Converted to Lead	-0.8215
Last Activity_Olark Chat Conversation	-0.9478

Step 5. Final Analysis

As sensitivity is 80%, therefore, all the leads with lead score > 35 (a potential lead) have 80% chance of getting converted.

Dummies that contributes most towards increasing the probability are:

- Occupation_Working Professional (currently 92% conversion rate)
- Lead Source_Reference (currently 91% conversion rate)
- Last Activity_SMS Sent (currently 63% conversion rate)

Of all the leads who are predicted as Converted, the sales team should follow the below strategy to increase the conversion rate:

- If all the leads who are a working professional are targeted, 92% of them have a chance of getting converted. These leads should be targeted first.
- Leads who come by reference have a chance of 91% conversion. Therefore, the sales team should next target these leads.

Variables that need to be worked on by the team are

- Country_India(38% conversion rate)
- Last Activity_Olark Chat Conversation(currently 8% conversion rate)
- Last Activity_Converted to Lead(currently 12% conversion rate)
- Lead Origin_Landing Page Submission (currently 36% conversion rate)

To increase the probability of conversion, teams should working on the following:

- A whooping 73% of the leads are from the Country-India ,but only 38% of them are converted.Hence,this is an indicator of area of improvement.
- Even though a major share of the leads originate from Landing page submission – around 52% of them ,only 36% of them are converted. Team can focus more on this area as there are many leads who have not converted.
- Olark Chat Conversion comprises only 10% of the leads and out of which only 8% are converted. Chat Conversation is a good tool to know more about candidates and market the products. Therefore, team should work more on getting candidates from Olark chat and should improve the marketing skills on this platform.
- Converted to lead in Last Activity comprises only 4% of the leads and out of which only 12% are converted. Converted to lead seems to be the first step when a particular person is treated as a lead. Teams should work on reaching to these customers at the earliest through other mediums like email, SMS or chat.