**Lead Scoring Case Study Summary**

**By: Kanika Khattar Ahuja and Divya Naman**

**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:**

There were 9000 data points of leads which consisted of various attributes such as Lead Source, Total Time Spent on Website,Total Visits,Last Activity, etc.

There were a total of 37 attributes out of which we had to determine what attributes will ultimately be useful in deciding whether a lead will be converted or not.

The target column ‘Converted’ was the one which depicted whether a past lead was converted or not wherein 1 meant it was converted and 0 meant it was not converted.

A logistic model was built to assign a lead score between 0 and 100 to each of the leads. This score can be further used by the company to target potential lead.

A higher lead score would mean that the lead is hot and is most likely to convert.

A lower lead score would mean that the lead is cold and will mostly not get converted.

The following steps were used to build a logistic model and further assign a lead score to every lead.

**Step 1. Reading and Understanding the Data**

The very first step was to understand the data and its attributes.

The data consisted of total 9240 rows and 37 attributes out of which 1 was a target variable.

Out of 37 attributes there were total 6 numeric attributes and rest were categorical.

Of all the leads, there were **38.5% of leads that got converted**.

**Step 2. Data Cleaning and EDA**

The next step was to clean the data and visualize that attribute based on the target variable.

**2.1 Missing value check**

Out of all 37 attributes there were many attributes that had missing values.

At the time of filling a form, there are fields where a user does not select any option, and that field is captured with a value of ‘select’ as a default option. Therefore, these ‘select’ values were also treated as a missing value for the attribute.

The very first step in cleaning the data was to delete all the columns where more than 45% of the data was missing. This involved deleting the following attributes:

* Asymmetrique Activity Index
* Asymmetrique Profile Index
* Asymmetrique Activity Score
* Asymmetrique Profile Score
* Lead Quality
* How did you hear about X Education
* Lead Profile

The imputations of the missing values in categorical and numeric variables were handled in the next steps.

**2.2 Cleaning and Visualizing categorical variables**

After visualizing the categorical variables we learnt that there were attributes that were skewed and had only one value more that 90% of the time. These included the following features and were removed as they did not provide any insights.

* Do Not Email
* Do Not Call
* Reason
* Search
* Magazine
* Newspaper Article
* X Education Forums
* Newspaper
* Digital Advertisement
* Through Recommendations
* Receive More Updates About Our Courses
* Update me on Supply Chain Content
* Get updates on DM Content
* I agree to pay the amount through cheque

The following categorical variables were visualized and their missing values were imputed

1. **Lead Source:**
   * It had a total of 20 different values.
   * ‘Welingak Website’ value has the maximum conversion rate of 98% followed by ‘Reference’ with 91% conversion rate.
   * Around 30% of the leads have the value ‘Google’ and ‘Direct Traffic’ and have a good conversion rate of close to 35%.
   * We noticed that Lead Source is highly affected by another attribute Lead Origin. Therefore, we replaced the missing values of Lead Source by the mode of Lead Source depending on itsLead Origin

**Lead Origin Mode of Lead Source**

- API Olark Chat

- Landing Page Submission Direct Traffic

- Lead Add Form Reference

- Lead Import Facebook

1. **Last Activity:**
   * 37% of rows have ‘Email Opened’ as the value of Last Activity
   * Replaced 103 missing values of Last Activity with Email Opened
2. **Country:**
   * There were many missing values for this attribute. Based on the city, the missing values were imputes,
   * Where the city information was not available, the values were set to ‘unknown’.
3. **Specialization:**
   * 36% values were missing which were replaced by ‘Others’ as opposed to deleting these rows which would have resulted in information loss.
4. **Occupation:**
   * 60% of leads were Unemployed and they have good conversion rate of 43%
   * Though number of working professional is less, but their conversion rate was higher at 91%
   * The missing percentage is 29% and replacing this with mode might have skewed the data.Therefore, Replaced the null values in Occupation with value ‘Other’
5. **City:**
   * Close to 40% of the values were missing
   * Each value had equal probability of conversion(~40%), which might not haved helped in the prediction
   * As the city column did not add any value, therefore, deleted the column.

Based on the business knowledge the following columns were deleted:

* **Last Notable Activity:** As it represented an intermediate step recorded by the sales represented. Instead of an intermediate step, the final conclusion by the sales representative was captured in Last Activity was retained.
* **Tags:**These tags were entered by the sales representative after talking to leads. As we need attributes by looking at which sales team should call leads, this attribute will not provide any insight to the sales team.

**2.3 Cleaning and Visualizing numerical variables**

1. **TotalVisits:**
   * Since the Total visits cannot be fractional, after considering both mean and median values the missing values were replaced with 3.
2. **Page Views per Visit:**
   * Since the Page Views per Visit cannot be fractional, after considering both mean and median values the missing values were replaced with 3.

The leads who got converted usually spent more time on the website.

The total visits on the website ranged widely for the leads who got converted.

**2.4 Outlier Treatment**

The outliers were treated by soft capping the 99% value.

**The following variables were finalized**

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

**Step 3. Preprocessing and Data Preparation**

**3.1 Categorizing variables**

Once the data was cleaned and visualized, we noticed that the attributes country and lead source had a large number of distinct values.

* For Lead Source we clubbed all the values whose count was greater than 7 into Others.
* There were 37 different countries and many have very few data. Clubbed them together based on the continents.

**3.2 Creating dummy variables**

* The dummies were created for all the categorical variables. As for n values, n-1 dummy variables are good, for the following variables we deleted the dummy manually based on the business knowledge.
  + Lead Source\_Others
  + Specialization\_Others
  + Country\_unknown
  + Occupation\_Other

**3.3 Train test split**

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

**3.4 Scaling data**

Standard scaling was used to scale all the numerical variables.

**Step 4. Model Building**

* For feature reduction, RFE was used.
* A logistic regression model was built and was finalized with 13 variables.
* Optimal cutoff of 3.5 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:

Features Coeff

* + 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.

Top 3 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.

Top variables that need to be worked on by the team are

* 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)

If company gets are chance to work on some new things, teams should working on the following:

* 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, the 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.