Lead Scoring Case Study

AMRIN MOHAMMED – RUDWAAN VANKAR - IBRAHIM KHAN

Problem Statement

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.

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%.

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:

Lead Conversion Process - Demonstrated as a funnel

Objectives

- To help the company in selecting the most potential leads, also known as **'Hot Leads'** whose lead **conversion rate is around 80%.**
- To build a **model** wherein 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.
- Help the **sales team** to divert their **focus** on **potential leads** & avoid them from making useless phone calls.

Business Goals

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Approach

- **Analyzing Patterns**:
  - Using Exploratory Data Analysis, we have analyzed the patterns present in the Dataset.
    which will provide us intuition that the which features will help in driving the lead conversion.

- **Driving Factors**:
  - Looking at the below data we get an intuition that how the variables are distributed.

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

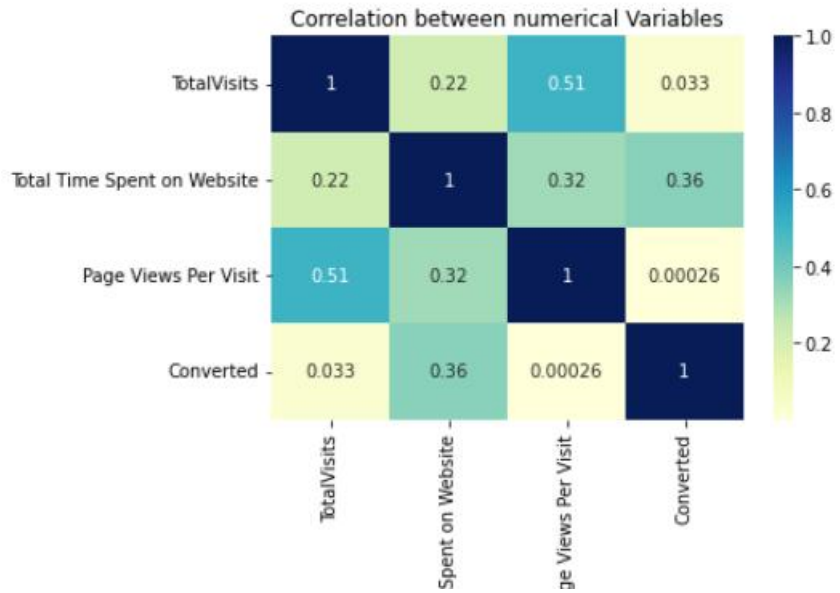
Data Insights

We have total 9240 entries of unique customers and we needs to identify out of these which have the highest probability of getting converted.

Decision Criteria:

- Potential Leads can be bifurcated on the basis of Leads Score (which is probability of getting converted).
- Out of 9240 entries we see that around 37% of leads are converted and 73% of leads are not converted.

Let us observe the correlation among the numerical columns.



We can observe that the variables are not highly correlated with each other. But still there is multicollinearity among some features.

Inferences

- **Lead Source**

  1- Majority source of the Lead is Google & Direct Traffic.

  2- Lead Source from Google have highest probability of conversion.

  3- Leads with source Reference have maximum probability of conversion.

  Lead Origin

  1- Customers identified as Lead from Landing Page submission, constitute the majority of the Leads.

  2- Customers originating from Lead Add Form have high probability of conversion, these customers are very few in number.

  3- Lead origin - API & Lead Import have the least convertion rate, customers from Lead Import are few in number.
- **Do Not Email**

  1- Customers who opt for Do Not Mail have lower conversion rate.

  2- Customers who do not opt for Do Not Mail have higher conversion rate around 40%.
- **Do Not Call**

  1- Customers who do not opt for Do Not Call have higher rate of conversion around 38%.
- **Last Activity**

  1- Customers whose Last activity was SMS Sent have higher conversion rate around 63%.

  2- Customers whose Last activity was Email Opened constitute majority of the customers with around 36% of conversion rate.
- **Specialization**

  1- Maximum Leads have specialisation as Management & Others.

  2- Leads with specialisation as Rural & Agribusiness have least probability of conversion.
- **What is your current occupation**

  1- Maximum Leads have occupation as Unemployed. 2- Very few leads are Housewives

- **What matters most to you in choosing a career**

  1- Number of Leads to whom better career aspects matters most in choosing a career are m ore & have higher probability of conversion.

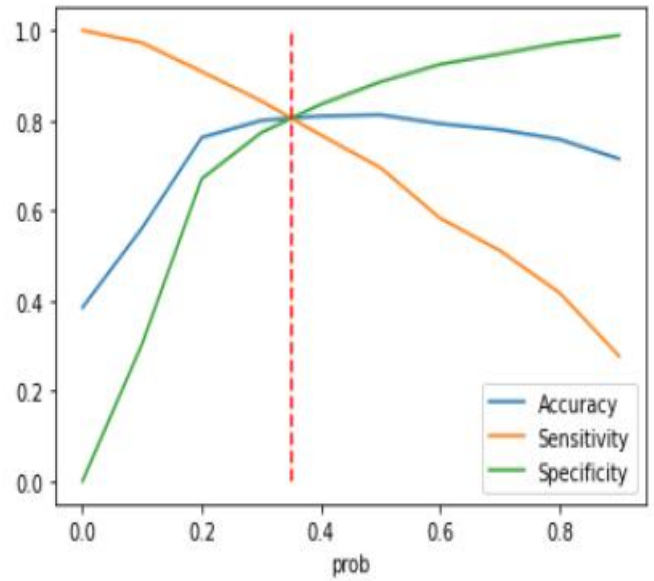
Tags

- More focus shall be given on the leads as will revert after reading the mail & others as these are potential leads and have higher rate of conversion.

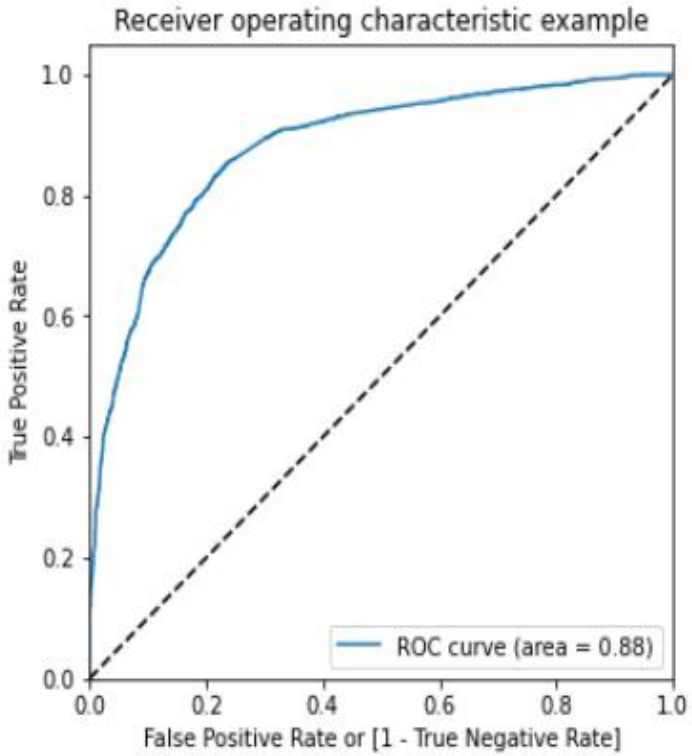
City

- Majority of the leads are from Mumbai city. Customers from Mumbai city should be targeted more as these are the potential leads.
- A Free Copy Of Mastering the Interview Customers who ask for free copy of mastering the interview are less in number but the conversion probability for both type of customers is similar.

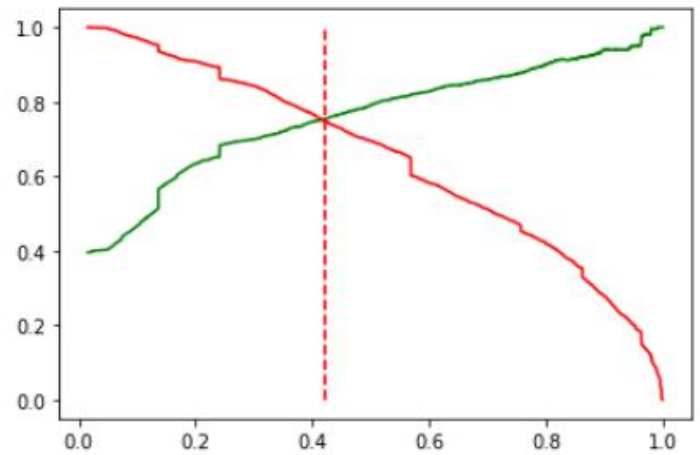
Last Notable Activity 1- Customers whose last notable activity was Modified are more in number. 2- Customers whose last notable activity was SMS Sent have higher probability of conversion.



- From the above graph it can be seen that the cutoff point is 0.35%



Area under roc curve is 88%



- Here the above Green line is indicating Precision and the Red line is indicating Recall
- Both the above lines are joining on 4.2

Below features are most important ones which are responsible for leads conversion :

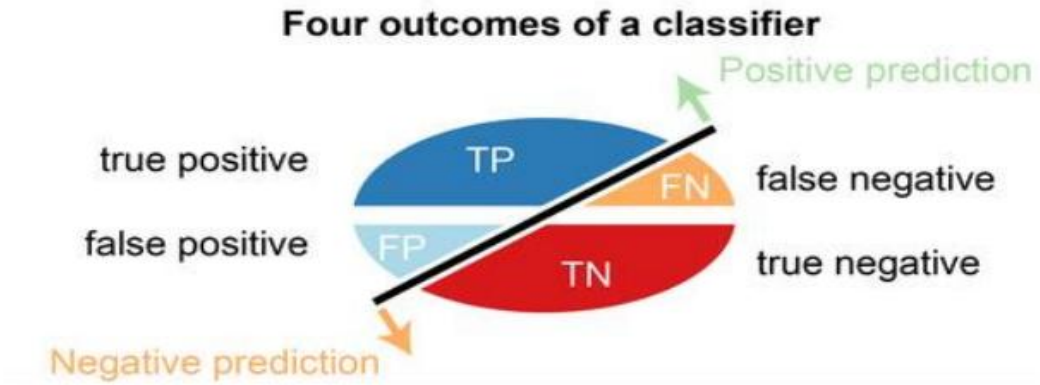
1. 'Total Time Spent on Website'
2. 'Lead Origin_Lead Add Form'
3. 'Lead Source_Olark Chat'
4. 'Lead Source_Welingak Website'
5. 'Last Activity_Email Bounced'
6. 'Last Activity_SMS Sent'
7. 'Tags_Closed by Horizzon'
8. 'Tags_Lost'
9. 'Tags_No phone number'
10. 'Tags_Others'
11. 'Tags_Will revert after reading the email'
12. 'Last Notable Activity_Modified'

'Last Notable Activity_Olark Chat Conversation'.

Terminologies Required

Before proceeding ahead, we need to understand few terminologies

- **Conversion of categorical columns to numerical**- This step is done as our algorithm runs only on numerical data.
- **Feature Scaling**- This is done to bring our data into same scale.
- **Data Splitting**- We have split the data into 70:30 and named it as train data and test data. We run model on train data and validate our model on test data.
- **Confusion Matrix:**



Where,
True positive (TP): correct positive prediction
False positive (FP): incorrect positive prediction
True negative (TN): correct negative prediction
False negative (FN): incorrect negative prediction

Above Metrics is Known as Confusion Metrics, using above metrics we derived following things:

1. **Accuracy** = (True Negative + True Positive)/Total
   This metrics provides the accuracy of the model, where total is TP + FN + FP +FN

2. **Sensitivity** = True Positive / (True Positive + False Positive)
   Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.

3. **Specificity** = True Negative/ (True Negative + False Negative)
   Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.

4. **Precision** = True Positive/ (True Positives +False Positives)
   Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

5. **Recall** = True Positives/(True Positives +False Negatives)
   The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect).

Observations

- Let us compare the values obtained for Train & Test:

Training DataSet

- Accuracy : 80.42%
- Sensitivity : 80.30%
- Specificity : 80.51%

Test Dataset

- Accuracy : 79.72%
- Sensitivity : 78.86
- Specificity : 80.21

Precision and Recall results

We used a cut off value of .42 and got the following results

- Precision : 75.29
- Recall : 74.89

---

we can conclude the model has decent sensitivity, specificity and Accuracy. This will help the X education company find Hot Leads who will be more likely to buy their courses and save their time and resources.

Conclusion Focus:

Company should focus on following features to increase the leads

- **Tags_Closed by Horizzon**: the Leads that have been assigned Tags as ‘closed by horizon’ have highest probability of conversion.
- **Tags_Lost**: Leads that have been tagged as ‘Lost ‘also contribute to the conversion to a considerable extent.
- **Tags_Will revert after reading the email**: Leads that have been tagged as ‘will revert after reading the mail’ also have significant correlation with the conversion.

ThankYou BY- AMRIN MOHAMMED – RUDWAAN VANKAR - IBRAHIM KHAN.