Lead Score Model

By - Prerna Prakash,

Murali Krishna Kalla

Problem Statement:

For X Education, there are a lot of leads generated in the initial stage but only a few of them come out as paying customers finally. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

We had to help X Education in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers by building a model wherein we needed 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. We had to maintain the target lead conversion rate to be around 80%, as per the requirement laid by the CEO of X Education.

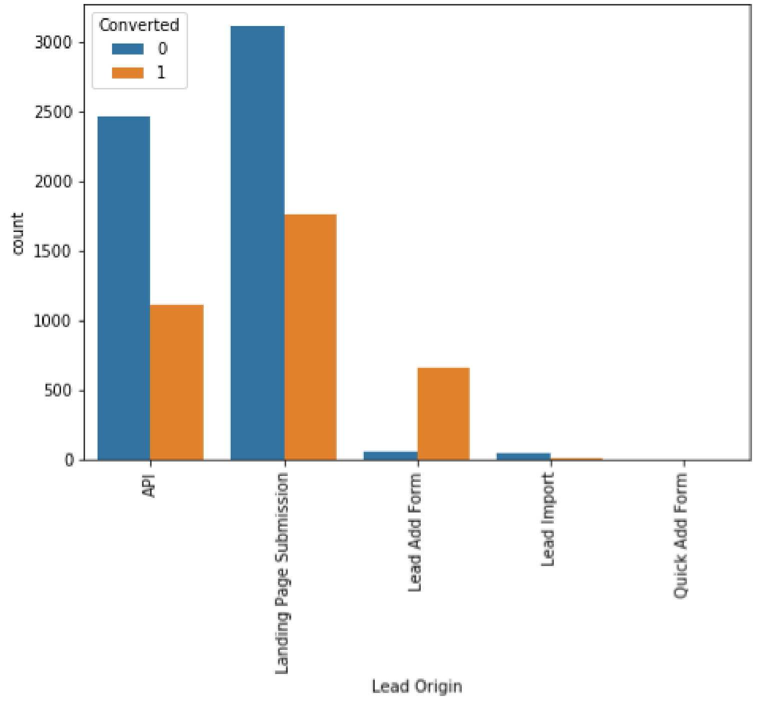
Analysis Approach

- As we have seen from the problem statement the target variable (Converted) is categorical, so we deduced that this a classification type problem that we are dealing with and hence we could proceed with getting the resolution by using Logistic Regression model.
- We have started off with understanding the data and cleaning it. Then we divided 70% of the data in training set and remaining 30% into the test set. The training set will be used to train the model to predict the conversion of a lead. Post this we have performed scaling on the train set and continued with the building a model. Once the apt model was ready whose coefficients p value was less than 0.05 and VIF less than 3, we proceeded with finding the optimal cut off point at which the lead can divided as converted(1) or not converted(0). We achieved a high sensitivity of 83.4% on test data and 83.2% on train data.

Explanation in Business terms

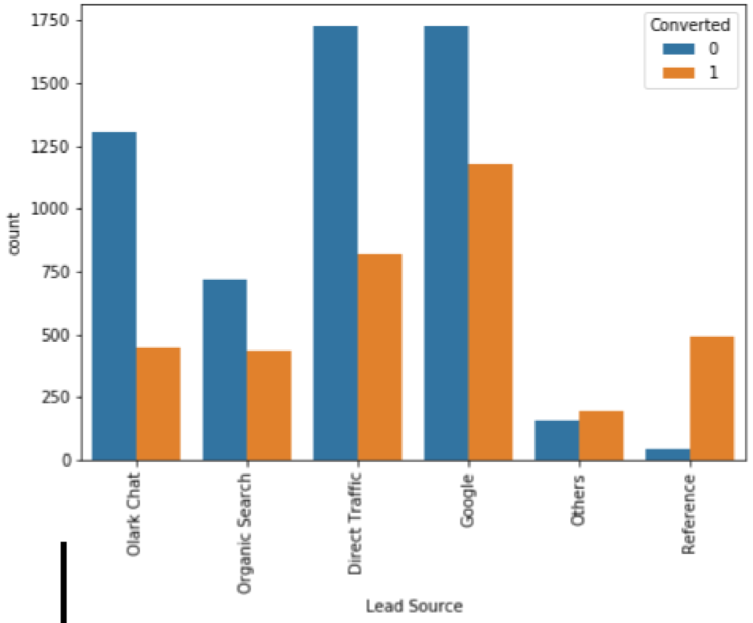
As X Education is aiming for getting to know the most promising leads, i.e. the leads that are most likely to convert into paying customers, we aimed at building a model with high sensitivity and assigned a lead score to each lead. As we know ballpark for target lead conversion rate decided by the CEO is around 80%, we got a model with sensitivity as high as 83%(approx.).

Exploratory Data Analysis: (contd.)



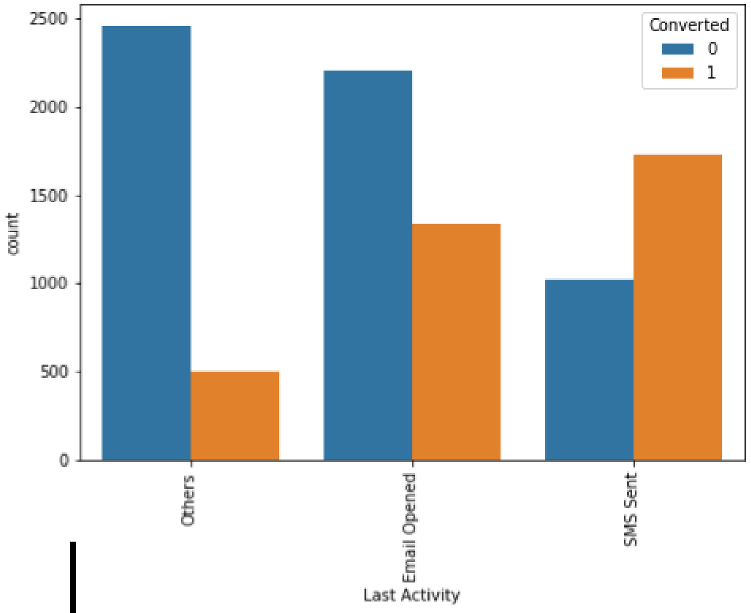
Lead Origin:

Leads originated from API and Landing page submissions have higher conversion rate than other categories. Even though the conversion rate in the same category is lesser in comparison to non conversion rate.



Lead Source:

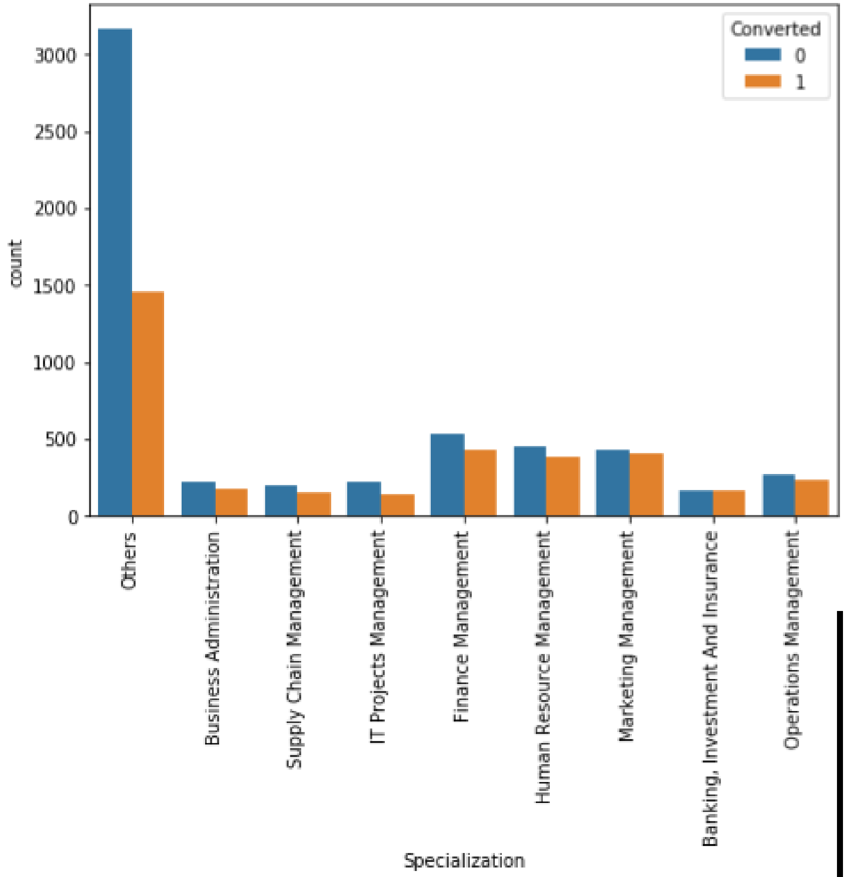
Google and Direct traffic generates maximum number of leads.



Last Activity:

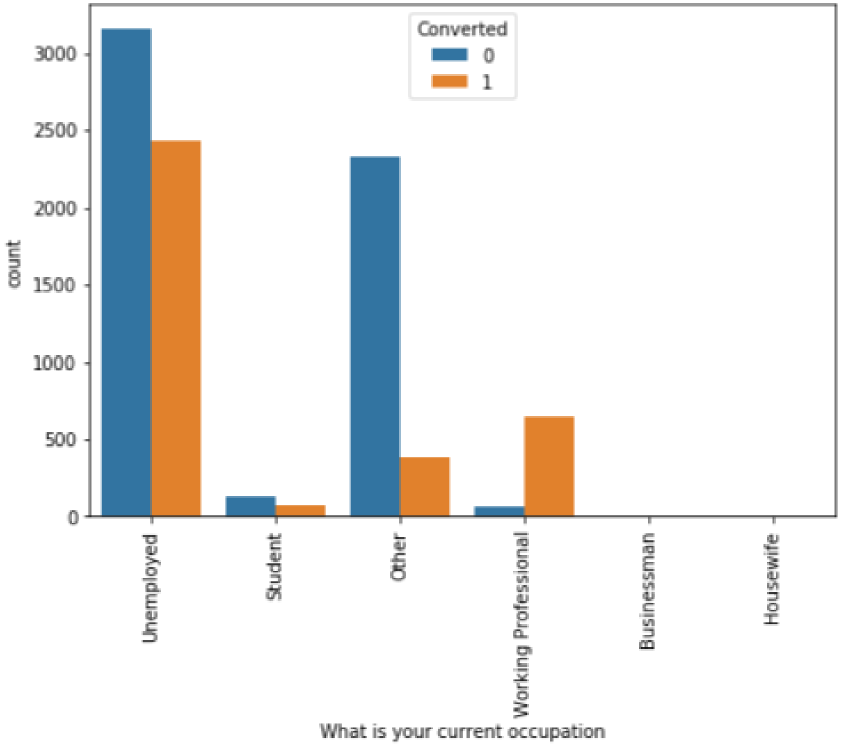
Most of the lead have their Email opened as their last activity.

Conversion rate for leads with last activity as SMS Sent is higher comparatively to all other categories.



Specialization:

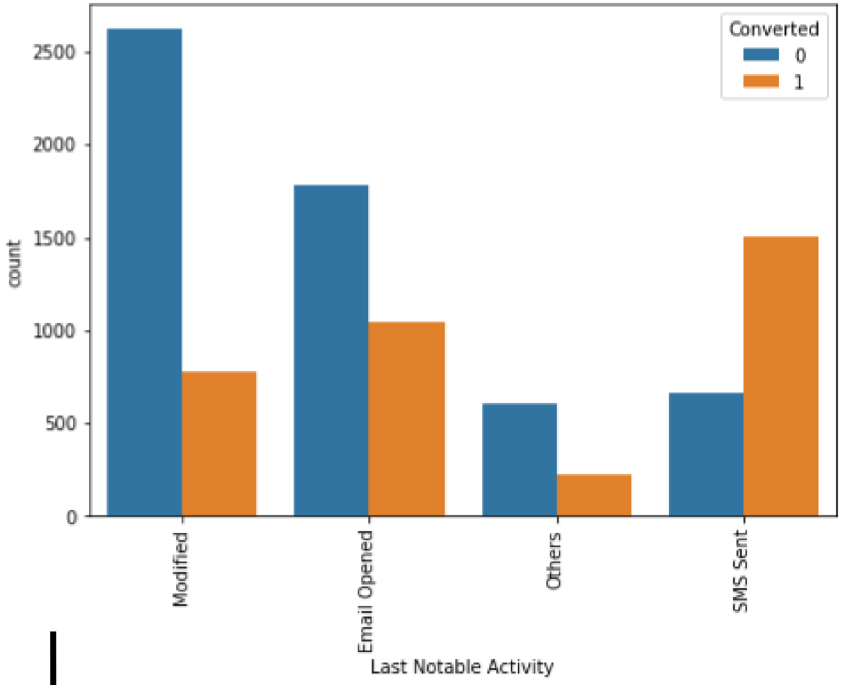
Most conversions and non conversions of leads are noted in the group of Others.



What is your current occupation:

Working Professionals going for the course have high chances of joining it.

Unemployed leads are the most in numbers but has around 30-35% conversion rate.



Last notable activity:

Conversion rate is higher on email opened in comparison to other categories.

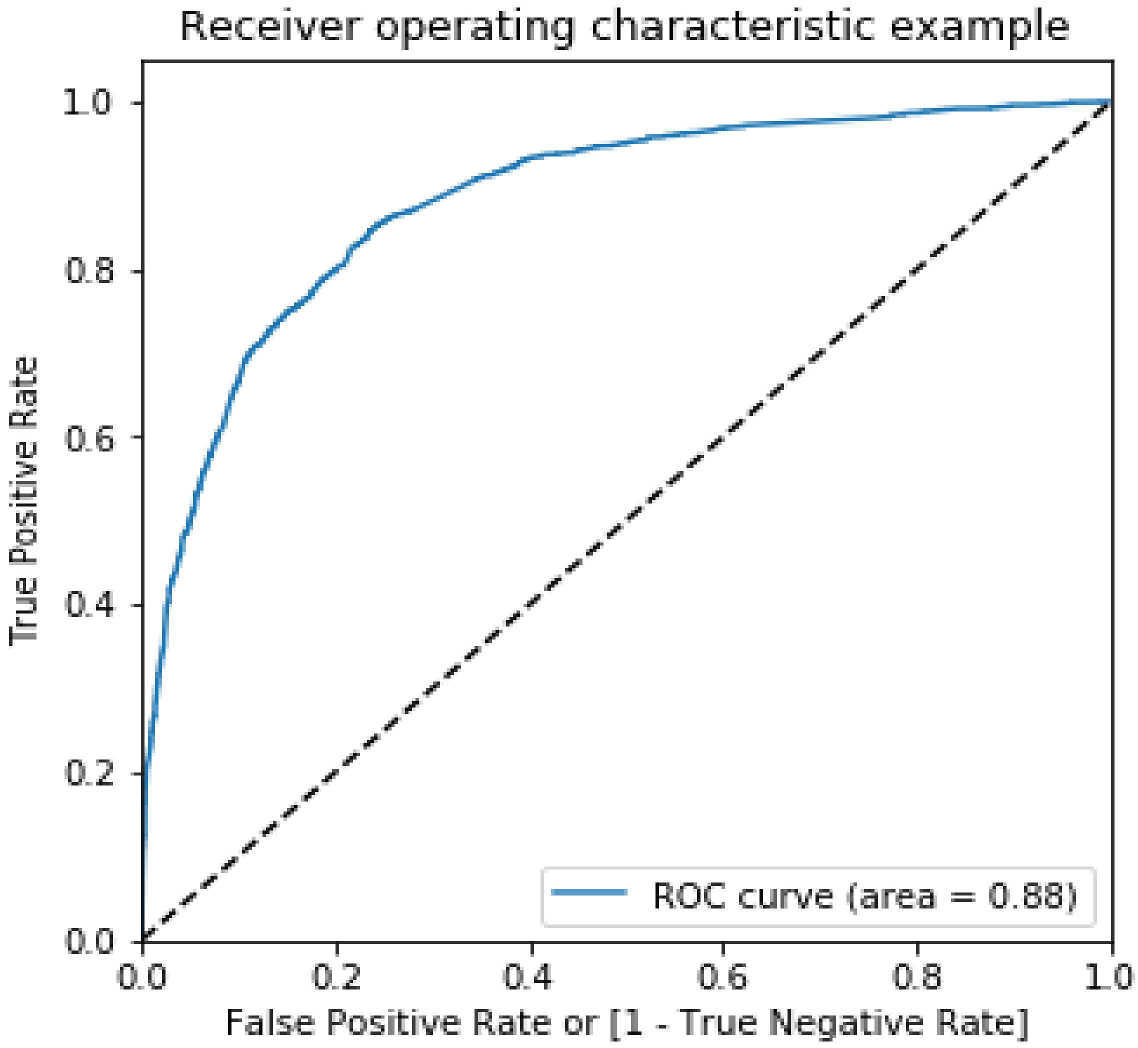
ROC curve:

To get a better sense of the model we plot the ROC curve, where

x-axis: Sensitivity(True Positive Rate)

y-axis: 1-Specificity(False Positive Rate).

ROC curve illustrates the diagnostic ability of the binomial model. A good ROC curve is the one that touches the upper left corner of the graph, so higher the area under the curve the better the model constructed is. In our case, we observe that the ROC curve area = 0.88.

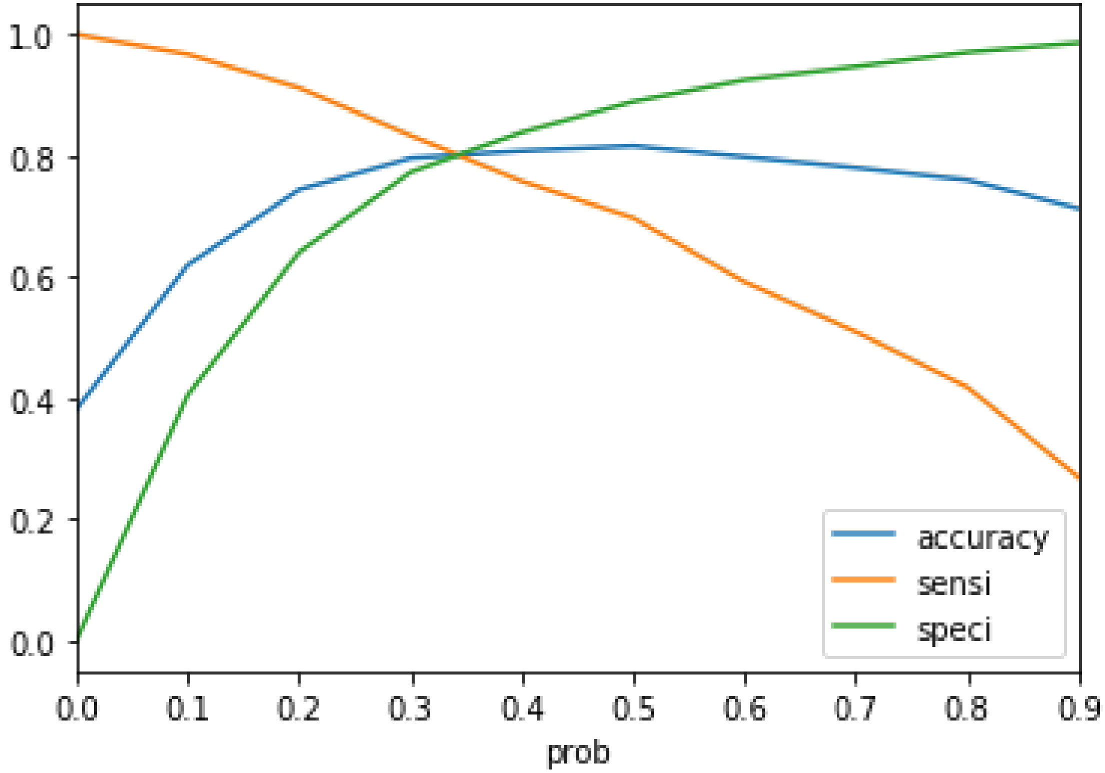


Plotting accuracy sensitivity and specificity for various probabilities:

We will plot a graph with accuracy, sensitivity, and specificity (on y axis) for various probabilities (on x axis).

The optimal cut-off point exists where the values of accuracy, sensitivity, and specificity are decent and almost equal.

At the cut-off of almost 0.3 probability, we observe that there is intersection of accuracy, specificity and sensitivity. This is the optimal value of threshold that we can consider.



Calculated Lead Score:

Using the model built we have calculated the lead score for each of the leads and arranged them in descending order.

We have calculated the lead score by using:

Lead Score = Lead conversion probability * 100

The top 5 hot leads are listed below in the table.

Lead Number	Converted	Converted_Prob	final_predicted	Lead_Score
627106	1	0.999410	1	99.94
600952	1	0.998939	1	99.89
604411	1	0.998932	1	99.89
651812	1	0.998734	1	99.87
606508	1	0.998472	1	99.85