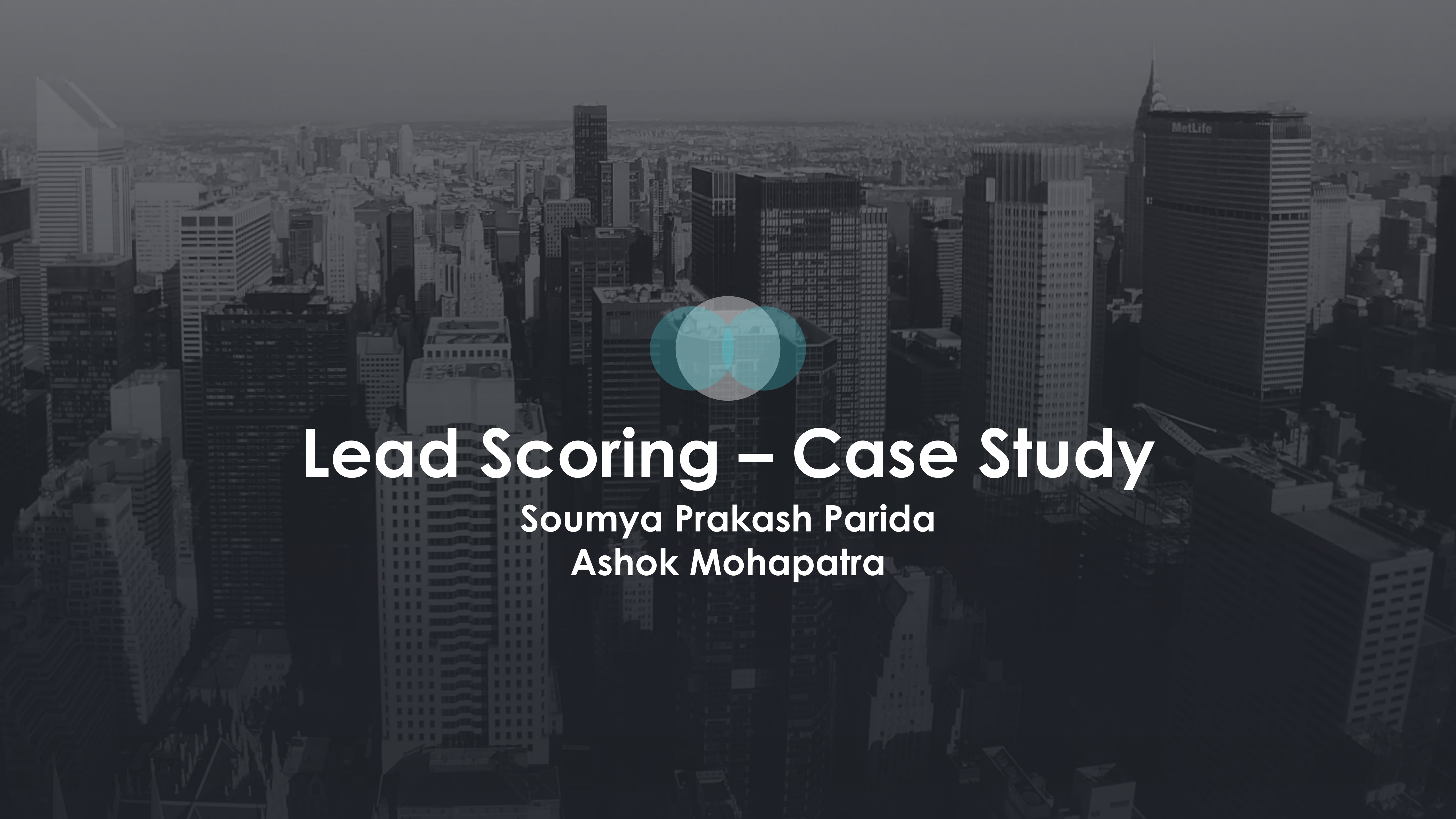


Lead Scoring – Case Study

Soumya Prakash Parida
Ashok Mohapatra

X EDUCATION

Objective

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objective

- 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.

Approach

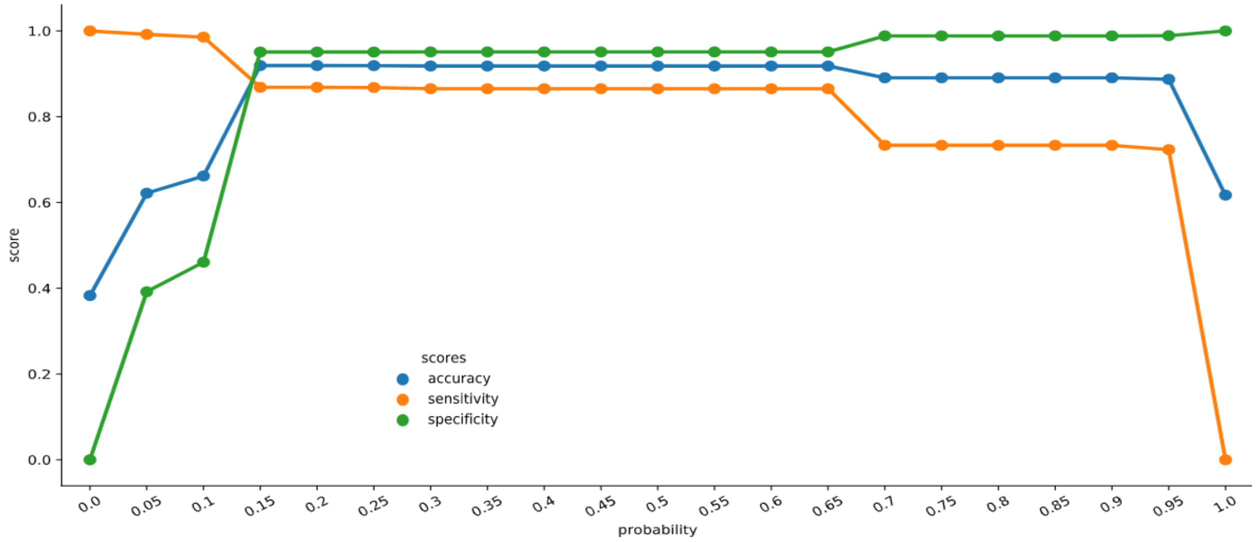
We have used *Logistic Regression* to determine and assign the probability score against each customer based on the available data and choose the candidates which have better probability of joining the course.
The lead dataset is split into Train and Test datasets and once the model is built on the Train dataset, it is used on the Test dataset to verify the accuracy, sensitivity and specificity of the model, which in turn indicates how good the prediction model is.

Logistic Regression

Identified Features

+			−
Tags - Closed by Horizzon	7.2988	- 3.8509	Tags- switched off
Tags - Lost to EINS	5.4513	- 3.3128	Tags - Ringing
Tags - Will revert after reading the email	4.8733	- 3.2165	Tags - Already a student
Lead Source – Wellingak Website	4.5341	- 2.9317	Tags - Not doing further education
Last Notable Activity - SMS Sent	2.6536	- 2.7201	Tags - opp hangup
		- 2.0225	Tags- Diploma holder
		- 1.7019	Tags- Interested in other courses

Model Performance



- The following model performance has been calculated on the "Complete Dataset".
- The cut-off curve and ROC Curve are based on the training dataset to produce an optimal cutoff point.

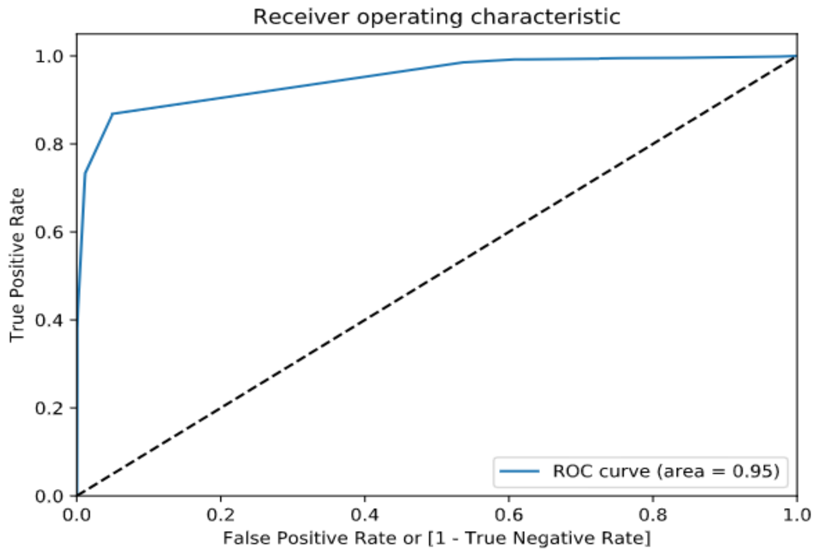
92% Accuracy

87% Specificity

95% Sensitivity

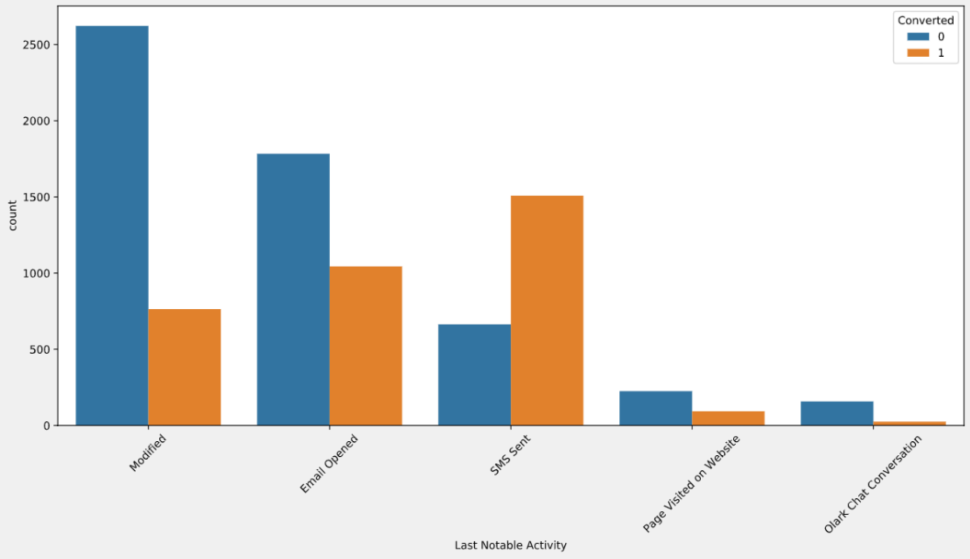
ROC

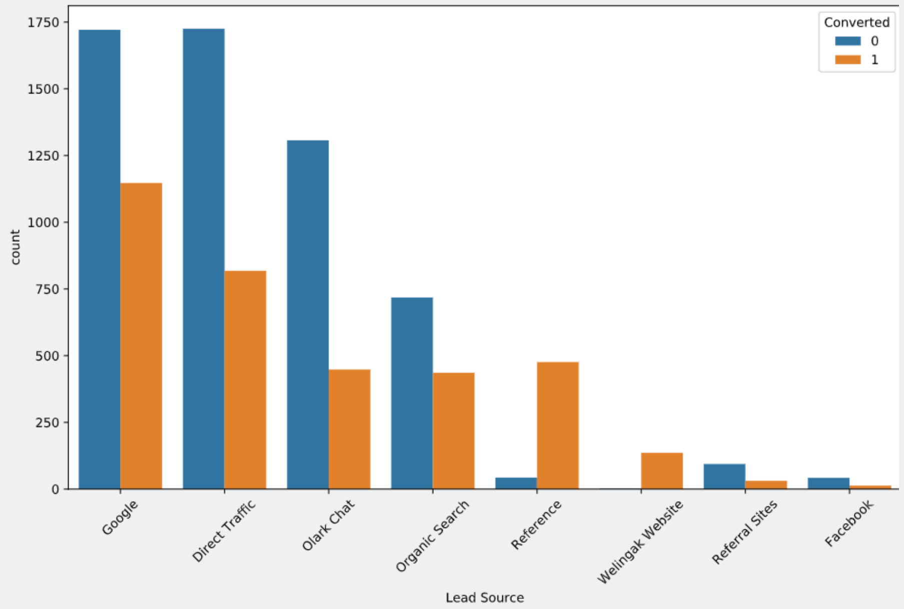
AUC : 0.95



Confusion Matrix		
Actual/Predicted	Converted	Not
Converted	3082	460
Not	274	5403

Data Analysis





Out of the Last Notable Activity categories, "SMS Sent" was found to be significant during EDA and RFE process.

"SMS Sent" stands out here, because the trend is opposite to all other categories in this column.

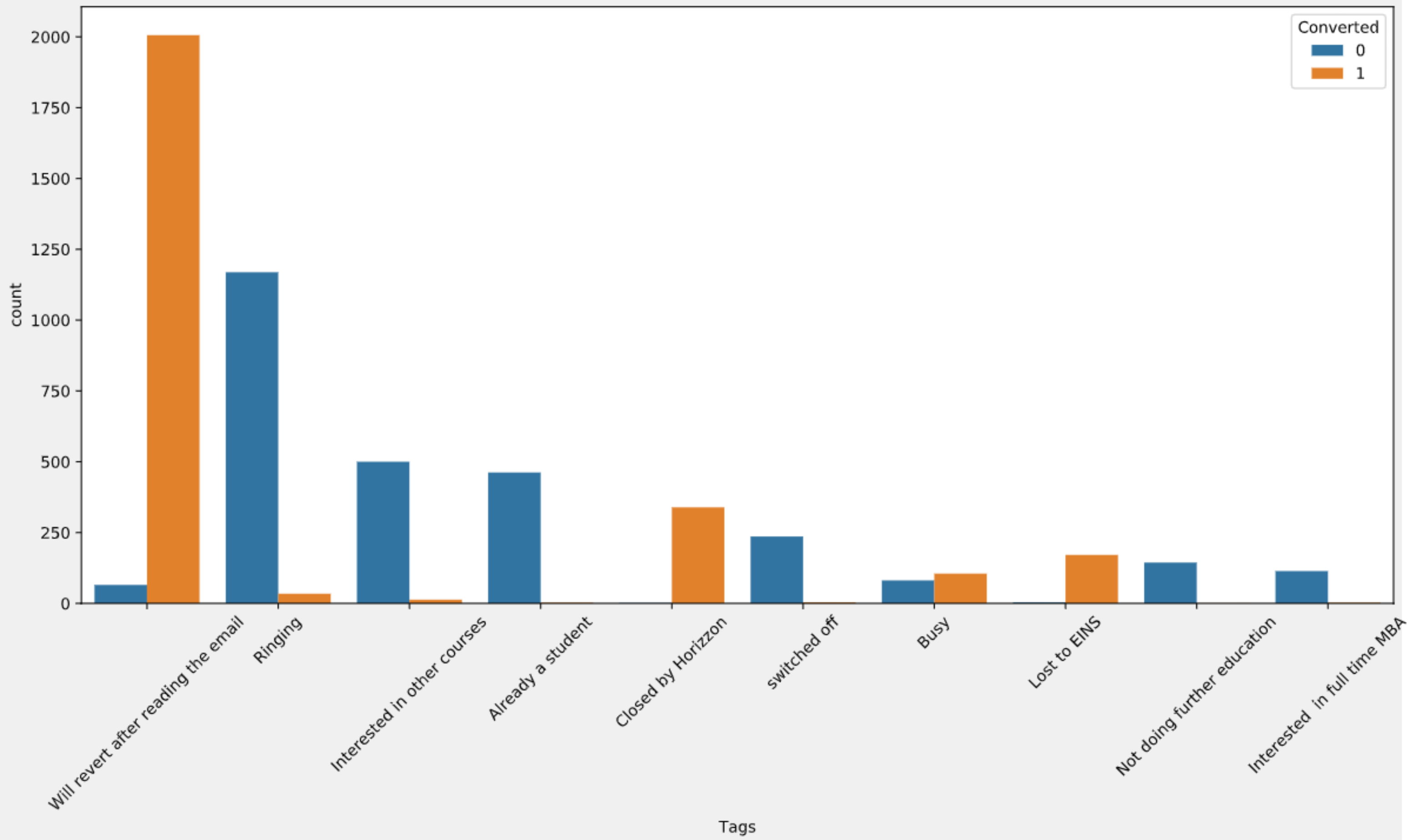
The Lead Source also follows a similar ratio for conversion for all categories, however, Referral Sites and Wellingak Website Reference have exceptional conversion rates.

However, Reference was eliminated by RFE as it was correlated and had a small coefficient if used.

Data Analysis

TAGS:

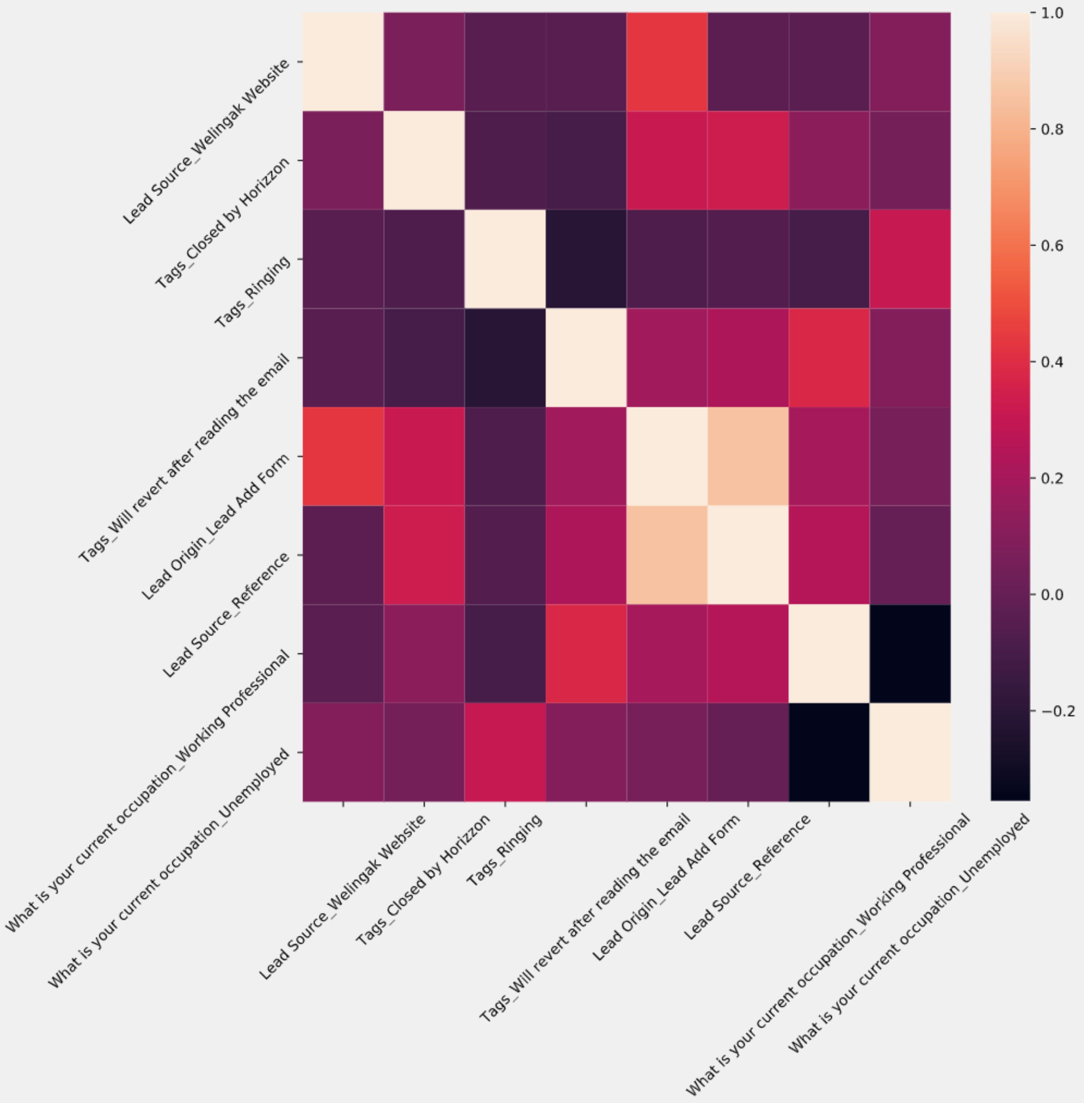
- This column and the categories within was found to be most significant during EDA and RFE method.
- The distinctive trend in these categories is that each category has a much definite outcome prediction.
- For instance “Will Revert…”, “Closed by Horizzon”, “Lost to EINS” indicate to a definite positive result.
- Whereas “Ringing”, “Interested in other…”, “Already a student” etc are definite negative indicators.
- Hence these tags are selected during RFE.
- There is strong indication that these Tags are also correlated to other categories thereby reducing the number of selected features.



Data Analysis

Other notable categories:

- There were some other notable categories which looked significant during EDA.
- However these were removed with RFE.
- The plot on the right side shows the correlation of these categories with other fields which were selected.
  - Lead Source Reference : This category has high correlation with Lead Add Form and 2 other highly significant categories "Wellingak…", "Horizzon…"
  - Lead Origin –Lead Add Form: This is highly corelated to Reference above.
  - Working Professional – This category is correlated with "Will revert.." which has very high significance.
  - Unemployed – The significance of this column was low, and further it is found to be correlated with "Ringing.."



Observations/Recommendations

Final Observations:

- X Education has very high turnover from Wellingak, Horizzon which seem to be generating good leads. However other advertising areas do not have a good conversion rate. So the advertising needs to be targeted more to the below demography.
- Working Professionals have a high conversion rate, so all advertisements can focus on that demography a bit more.
- Direct References and Lead Add Form have high chance of joining course. A focus on increasing referrals would help in generating good leads.
- The tags with negative impacts, denote that leads with no number, incorrect number and unanswered numbers are cold leads and should not be pursued. These have the lowest probability for conversion.
- Further, lot of cold leads also include graduates and leads already in a course. These are in tags, which mean these information were collected during cold calls.
- The above information can be gathered during website registration and or referrals to eliminate them as leads. As these do not get converted.

- With the given dataset, the probability of a lead getting converted is found to 38.4 % from the actual conversions, whereas with the predictive value it is found to be 36.4 %

THANK YOU

Soumya Prakash Parida
Ashok Mohapatra