Lead Scoring Assignment "X Education"

By
- Lavkesh Sharma
- Aniket Koltharkar

Problem Statement

X Education gets a lot of leads and 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'.

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

Business Goal

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.The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

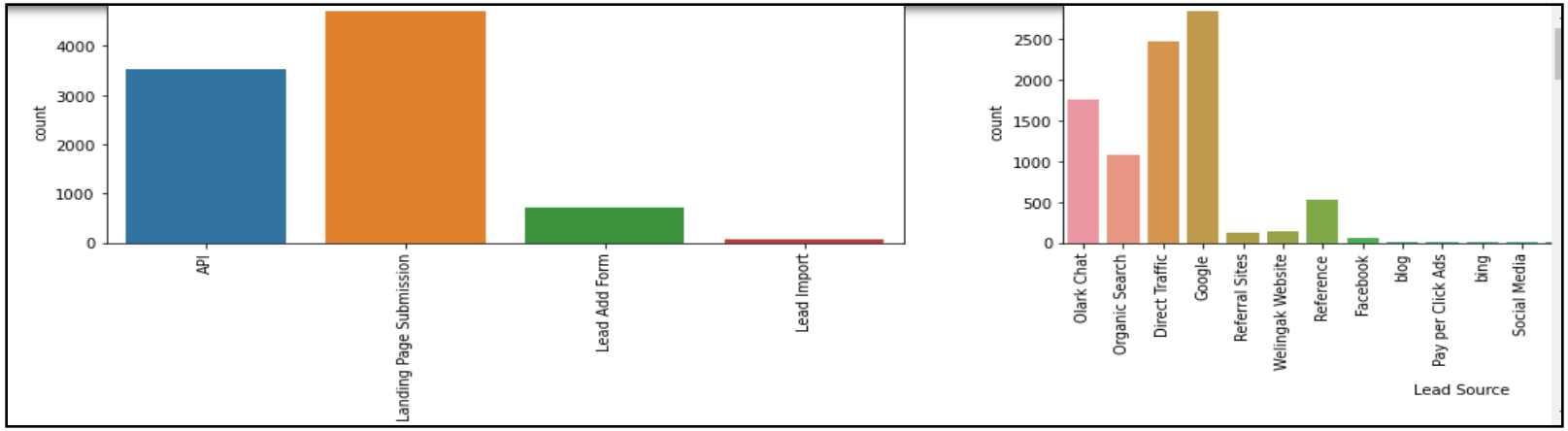
Case Study Steps performed

- Data Reading and Understanding the data.
- Data Cleaning
- EDA
- Data Preparation
- Split the Data into train and test Data set
- Feature Scaling
- Building the Model
- Feature selection through RFE
- Evaluation of the model
- Performing the prediction of test dataset.

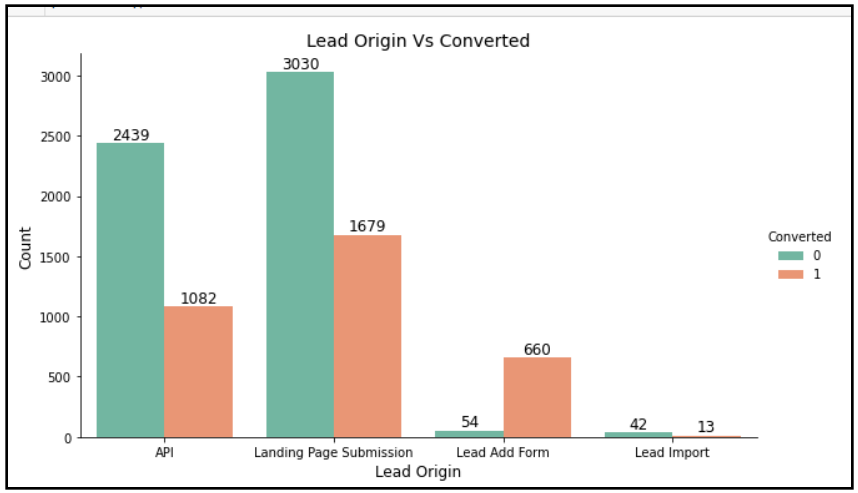
Data Cleaning(Manipulation)

- Total Numbers of row and columns : (9240, 37).
- There were null values in few columns:
- Some of the columns had values as “Select” which were imputed to NaN.
- Dropped the columns having more then 35% null values.
- “What is your current occupation” nan values imputed with “info not available.
- “Page Views Per Visit” column nan values imputed with median.
- Google and google were two category which merged into one In “Lead Source” column.
- Outlier were treated (replaced with 99% percentile).

EDA

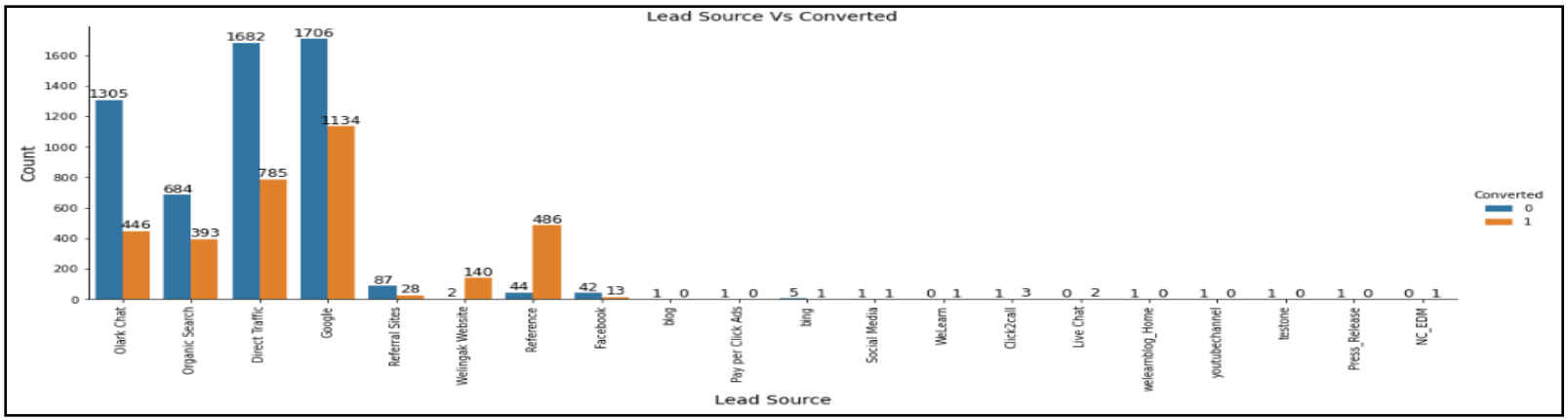


Above plot show count for the categories present in Lead source(Right) and Lead Origin(left).

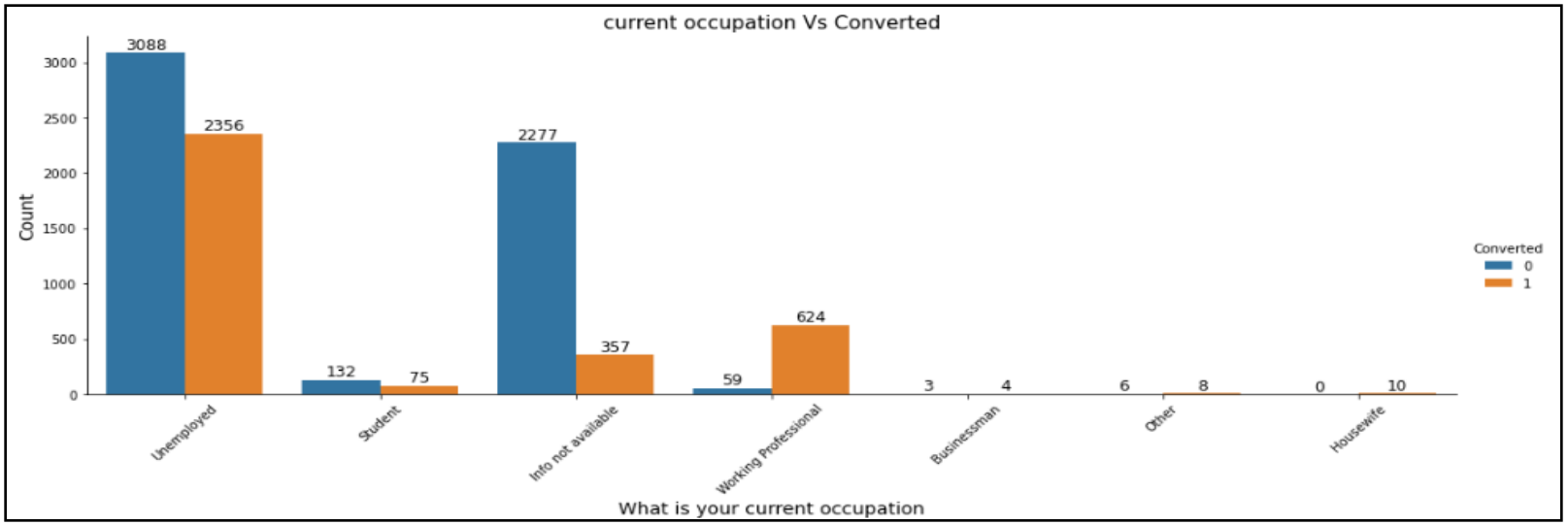


From above plot we can see that "Lead Page Submission" were converted the most in Lead Origin.

EDA

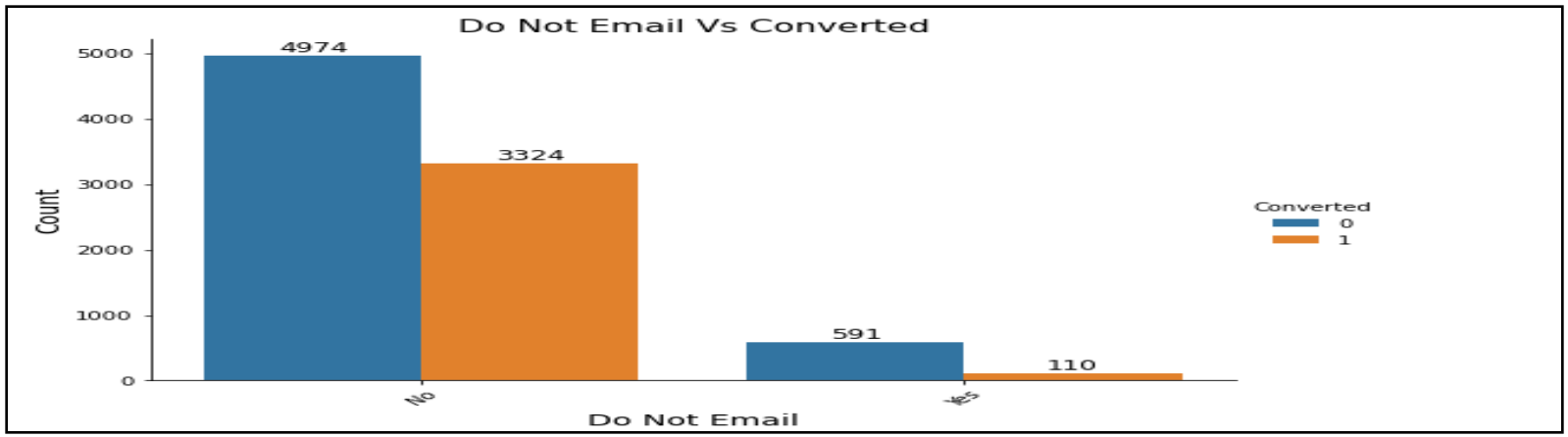


From above plot we can observe that "Direct Traffic" and "Google" were the two most converted category in Lead Source.

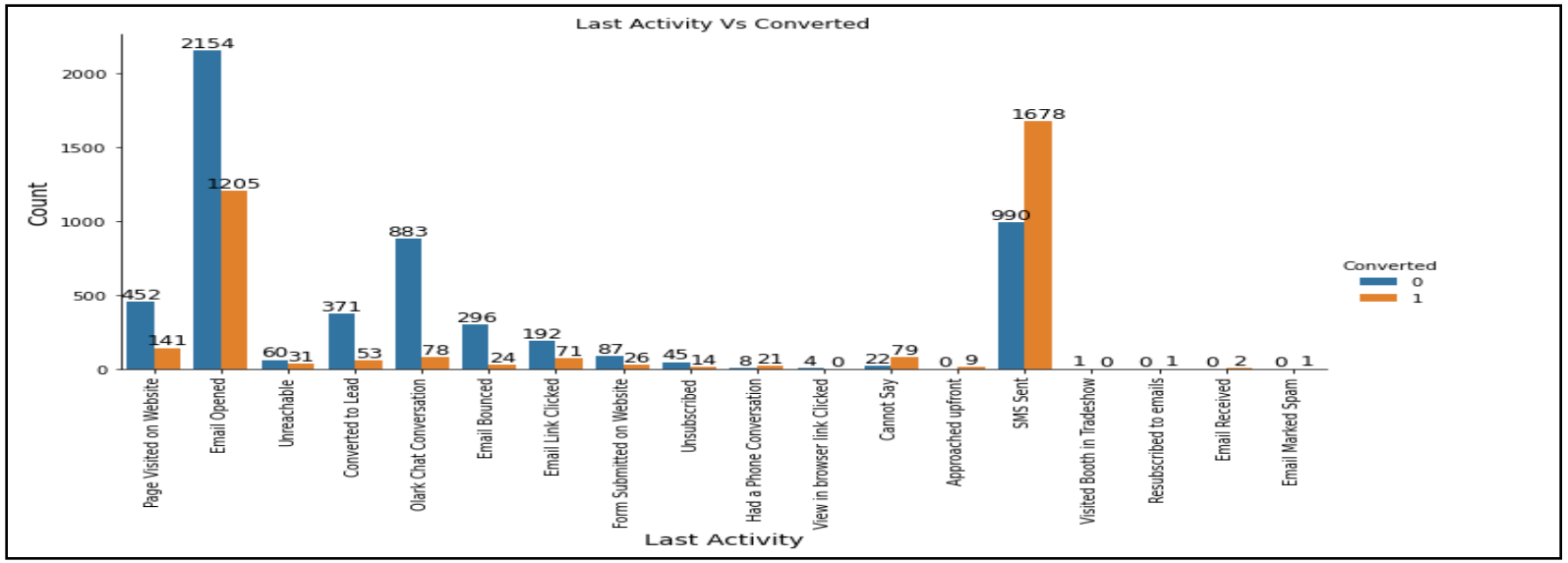


From above plot we can observe that "Unemployed" was the most converted category in Lead Source.

EDA



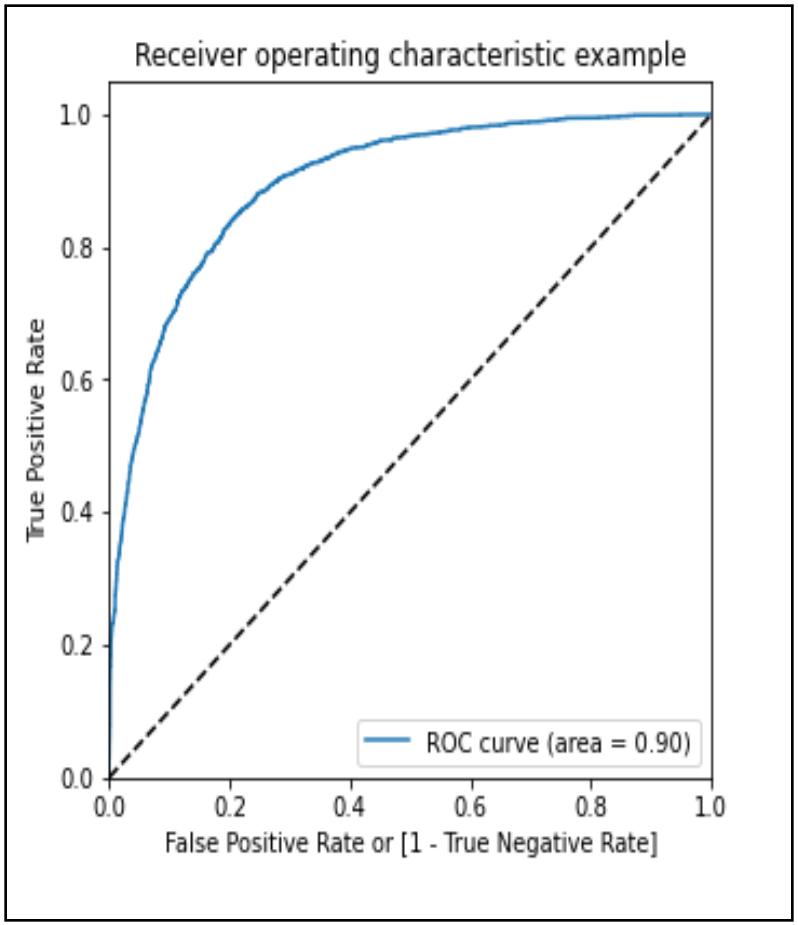
From above plot we can observe that email :No is most converted category in Lead Source.



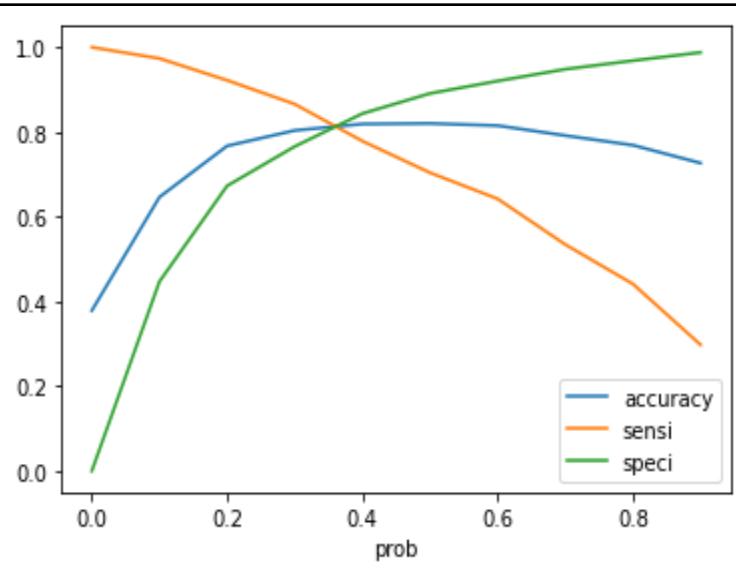
From above plot we can observe that “SMS sent" was the most converted category in Last Activity.

Features and ROC Curve of Model

	Features	VIF
const		786.85
Lead Origin_Lead Add Form		4.16
Lead Source_Reference		3.82
Lead Origin_Landing Page Submission		2.07
Lead Source_Olark Chat		1.99
Last Activity_Email Bounced		1.71
Do Not Email_Yes		1.66
Last Activity_Olark Chat Conversation		1.48
Lead Source_Direct Traffic		1.42
Last Notable Activity_Modified		1.36
Total Time Spent on Website		1.29
Last Activity_Converted to Lead		1.21
Last Activity_SMS Sent		1.18
What is your current occupation_Info not available		1.15
What is your current occupation_Working Professional		1.12
Lead Number		1.08



Model Evaluation: Train data set



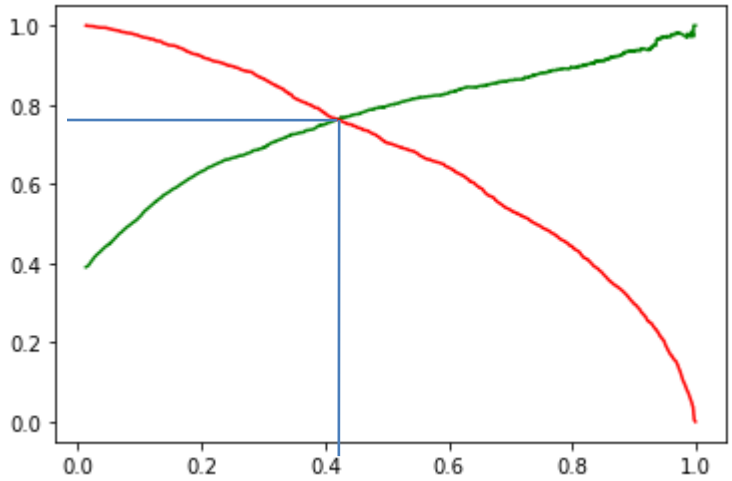
- From above plot we can observe that the optimum cut can be 0.38

Confusion Matrix :

3240	675
485	1899

- ❖ Accuracy: 81.58%
- ❖ Sensitivity : 79.6%
- ❖ Specificity : 82.75%
- ❖ False positive rate : 17.24%
- ❖ Positive predictive rate : 73.77%

Model Evaluation: Train data set



Confusion Matrix :

3240	675
485	1899

- Precision : 79.71%
- Recall : 70.38%

The graph shows the cut off 0.41(based on recall and precision)

Model Evaluation: Test data set

- Accuracy: 80.85%
- Sensitivity :79.65%
- Specificity : 82.75%
- Precision : 74.83%
- Recall : 76.47%

Conclusion

➢Accuracy- 80.85%, Sensitivity- 79.65% and Specificity- 82% values of test set which are approximately closer to the respective values calculated using trained set.

➢The Following are the top three variables in our model which contribute most towards the probability of a lead getting converted:

- **Lead Origin Lead Add Form**
- **Total Time Spent on Website**
- **What is your current occupation Working Professional**

➢ To avoid unnecessary phone they can use this model.

➢They can use bots to chat and get a more serious candidate

➢They search for more features of past converted leads, so that it can be helpful to build more sensitive and accurate model to convert future leads