LEAD SCORING CASE STUDY

Soumya Ranjan Bhanja
Mithilesh Hubli
Jagadeswar Reddy CH
Batch: DS57

Problem Statement

- X Education is an Organization which provides online courses for industry professional. The company marks its courses on several popular website like Google.

- X Education wants to select most promising leads that can be converted to paying customers.

- Although the company generates a lot of leads only a few are converted into paying customers, where in the company wants a higher lead conversion. Leads come through numerous modes like email, advertisement, on websites, Google searches etc.

- The Company has 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

Business Goal

The Company requires a model to be built for selecting most promising leads.

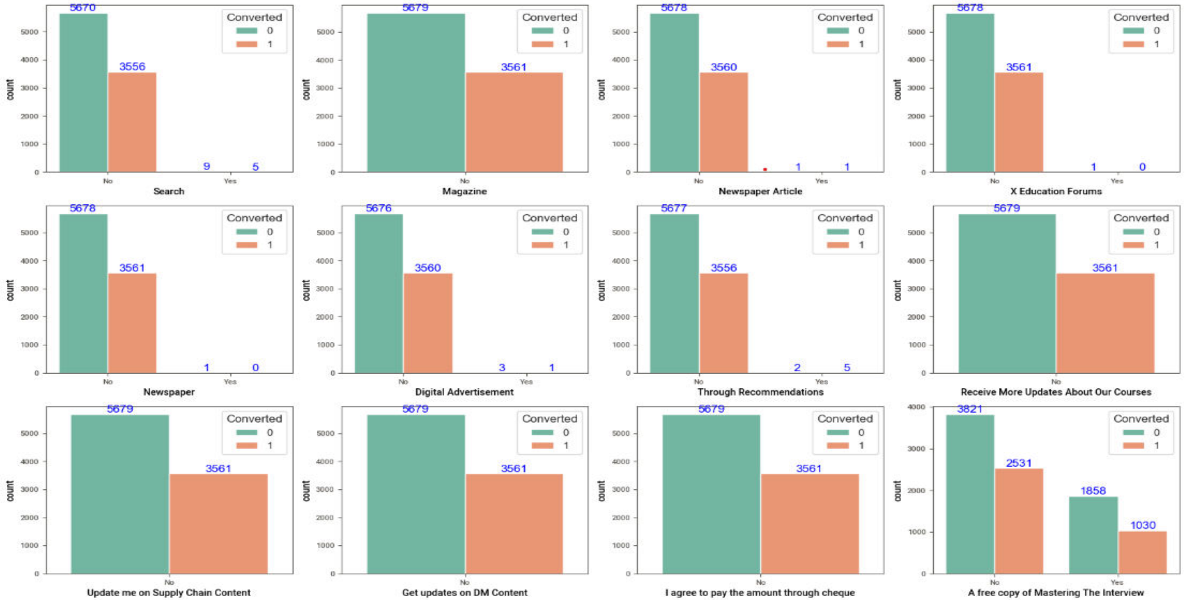
Lead score to be given to each leads such that it indicates how promising the lead could be. The Higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversation.

The model to be built in lead conversion rate around 80% or more.

Strategy

- Understanding the dataset
- Clean and preparing the acquired data for further analysis
- Exploratory data analysis
- Prepare the data for model building
- Outlier Treatment
- Data Preparation
- Dummy Variable Creation
- Build Logistic Regression Model
- Feature Scaling
- Model Building using Stats Model & RFE
- Measure the accuracy of the model and other metrics for evaluation

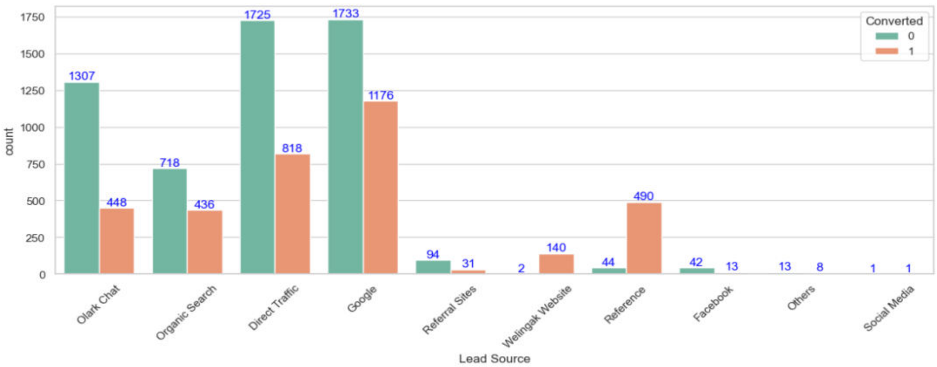
Categorical Attributes Analysis
Imbalanced Variables



- In this Graph columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus we will drop them

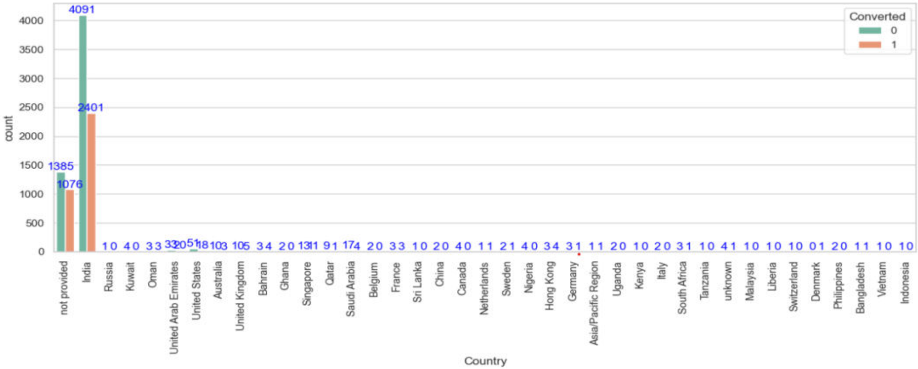
- "A free copy of Mastering The Interview" is a redundant variable so we will include this also in list of dropping columns.

Exploratory Data Analysis



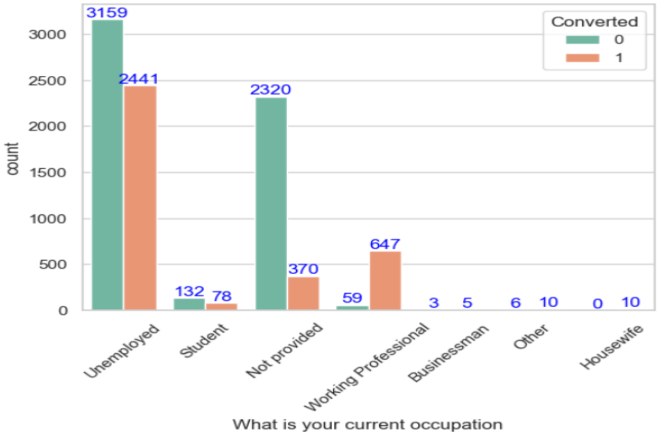
Lead Source vs. Converted

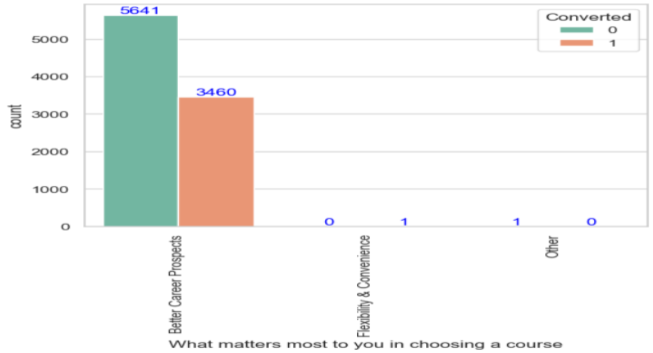
- In this graph Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.



Country vs. Converted

- As we can see that most of the data consists of value 'India', no inference can be drawn from this parameter. Hence, we can drop this column



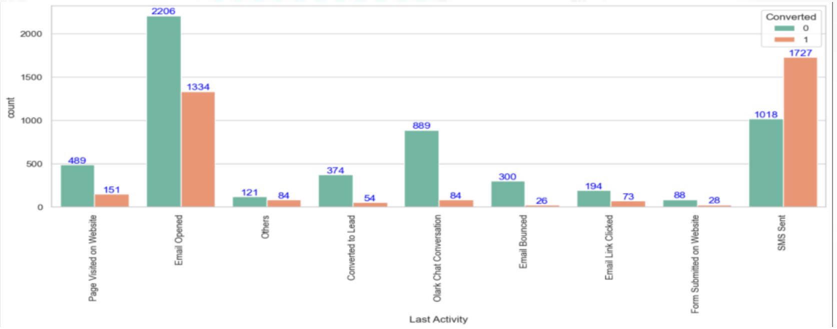


Current occupation vs. Converted

- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.

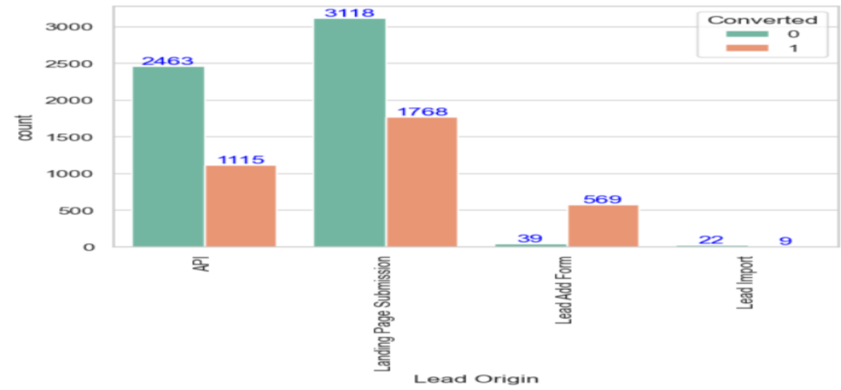
What matters most to you in choosing a course vs. Converted

This column spread of variance is very low , hence it can be dropped.



Last Activity vs. Converted

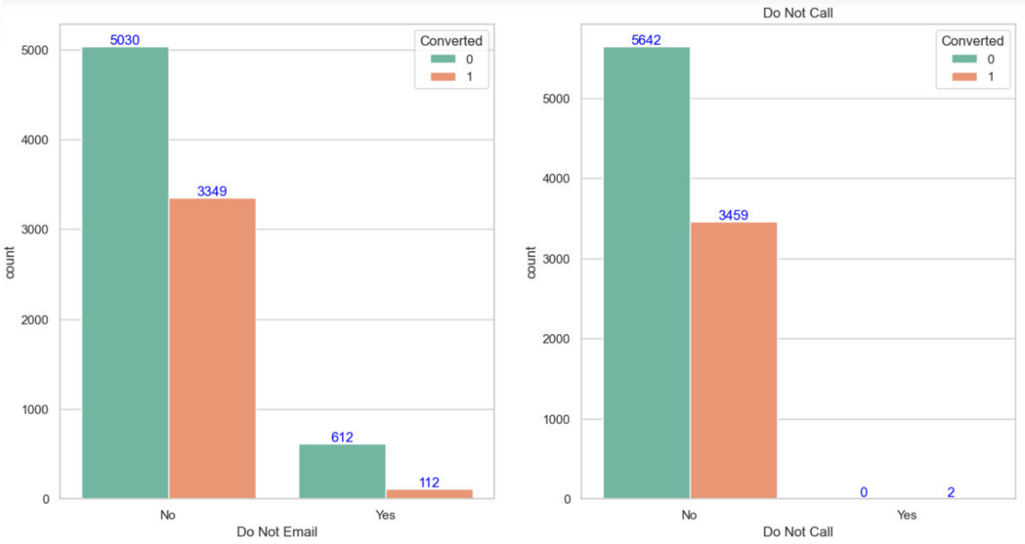
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last activity has high conversion rate.



Lead Origin vs. Converted

- Landing Page submission has had high lead conversion

Do Not Email & Do Not Call vs. Converted

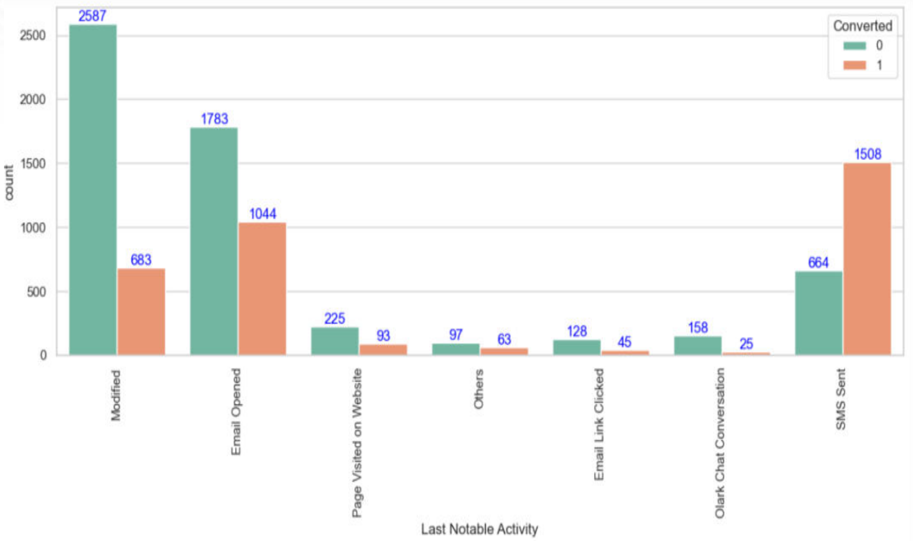


Do Not Email vs. Converted

Google searches has had high conversation compared to other modes, what references has had high conversion rate.

Do Not Call vs. Converted

Most leads prefer not to informed through phone.



Last Notable Activity vs. Converted

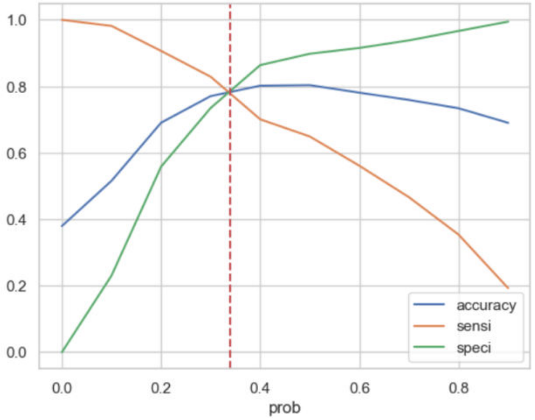
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.

- SMS sent as last activity has high conversion rate.

Model Building

- Model Building using Stats Model & RFE
- Calculating VIF
- Predicting a Train model
- Evaluate Accuracy & Other Matrices
- Plotting ROC Curve
- Finding Optimal Cutoff Point
- Precision & Recall
- Prediction on the test set
- Precision and Recall metrics for the test set

Model Evaluation Train



Accuracy Sensitivity & Specificity

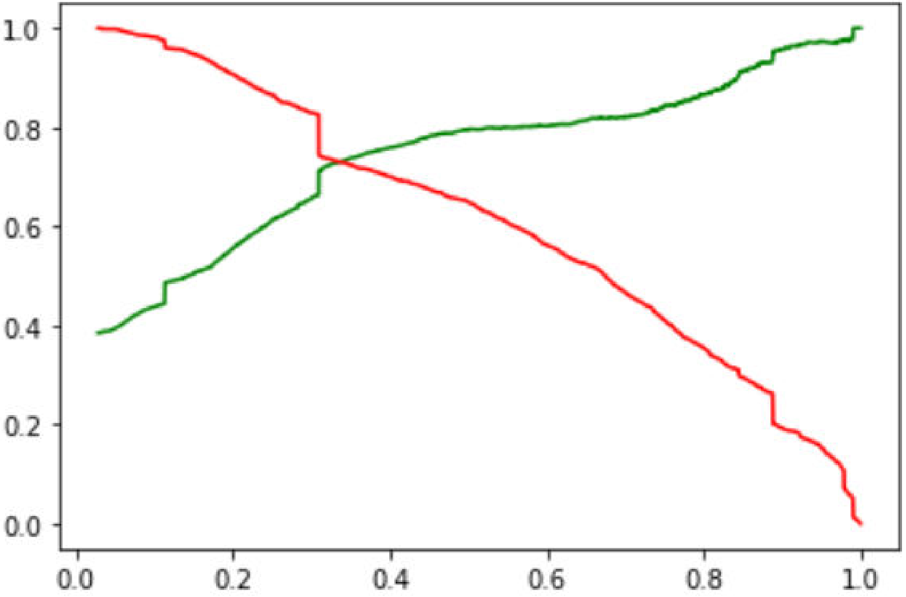
2905	1048
414	2005

Accuracy: 77.05%
Sensitivity: 82.89%
Specificity: 73.49%

Precision & Recall

65% Precision
82% Recall

Model Evaluation Test



Accuracy Sensitivity & Specificity

1252	437
177	865

Accuracy : 77.52%
Sensitivity :83.01%
Specificity : 74.13%

Precision & Recall

66% Precision
83% Recall

Hence We found the final parameter

Do Not Email	-0.360034
Total Time Spent on Website	1.102320
Lead Origin_Lead Add Form	4.611875
Lead Source_Direct Traffic	-1.049608
Lead Source_Google	-0.780419
Lead Source_Organic Search	-0.863852
Lead Source_Reference	-1.742494
Lead Source_Referral Sites	-1.374889
What is your current occupation_Student	1.134167
What is your current occupation_Unemployed	1.261263
What is your current occupation_Working Professional	3.757549

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.
- Hence overall this model seems to be good.

Thank You