Lead Score Group Case Study

Submitted By:

Ms. Shalu Lulla Nair

Mr. Vijayendra Kumar N Joshi

Ms. Nanda Pattanashetti

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.

Business Goal

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you 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.

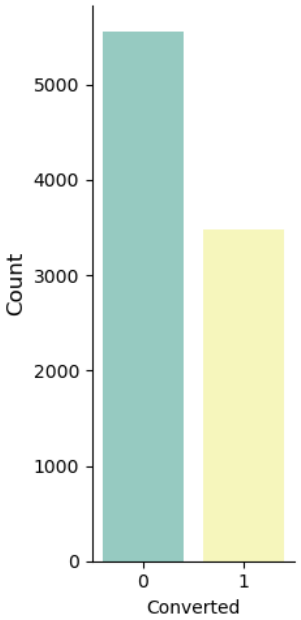
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

1. Importing and Cleaning of the Data
2. Formatting / Grouping for EDA analysis.
3. Performing Univariate & Bivariate analysis on Categorical and Numerical Columns
4. Min-Max Feature Scaling
5. Splitting Dataset in to Train and Test
6. Building Logistic Regression Model
7. Calculate Lead Score
8. Evaluating the model by using different metrics (Sensitivity , Specificity Or Precision and Recall)
9. Applying the best model on the test based on Sensitivity and Specificity Metrics

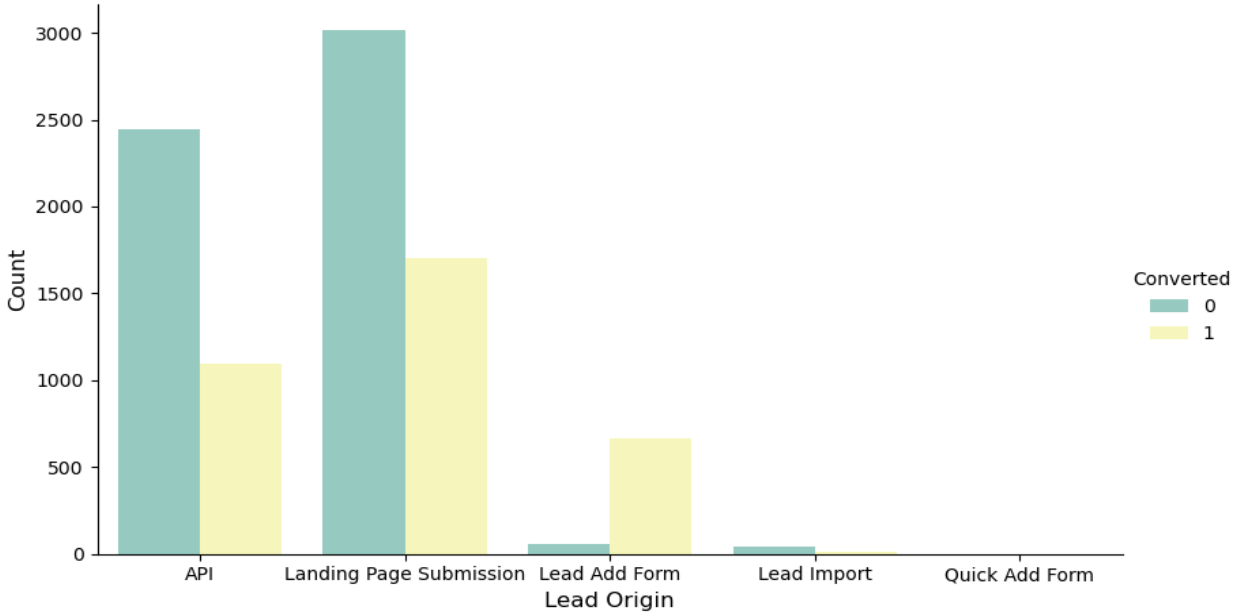
EDA - Univariate Analysis and Bivariate Analysis

Conversion Rate



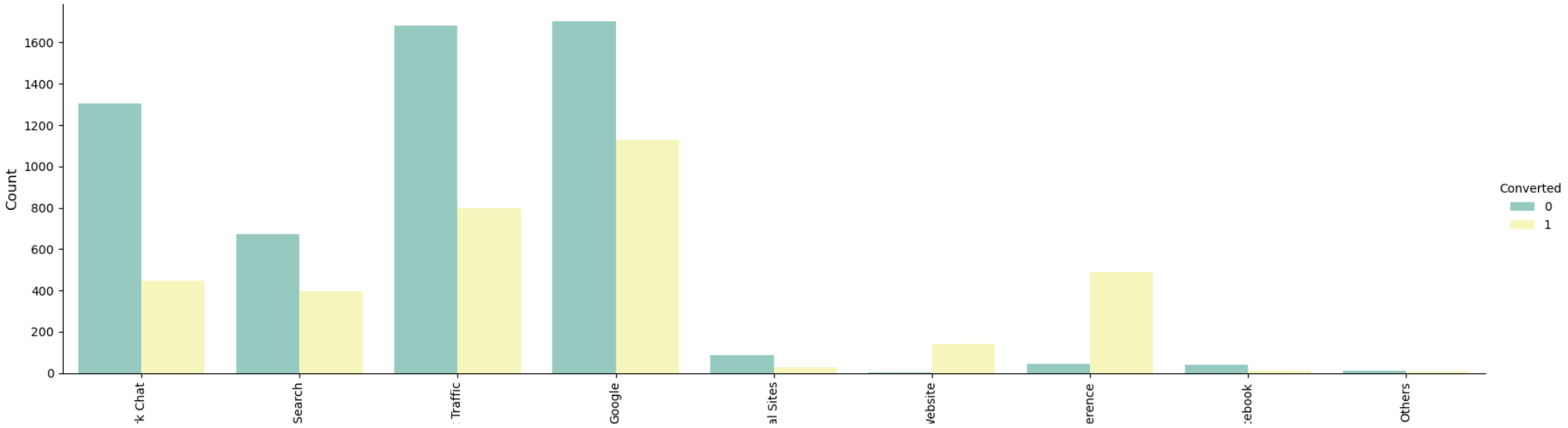
Overall conversion rate of around 38.51%

Lead Origin Vs Converted



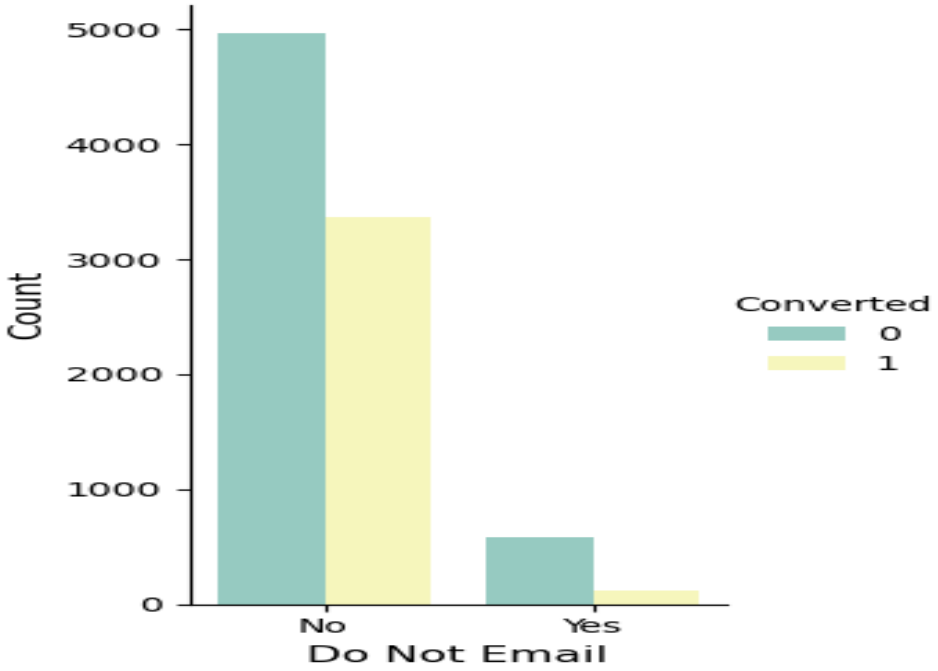
The maximum conversion happened from Landing Page Submission Also there was only one request from quick add form which got converted.

Lead Source Vs Converted

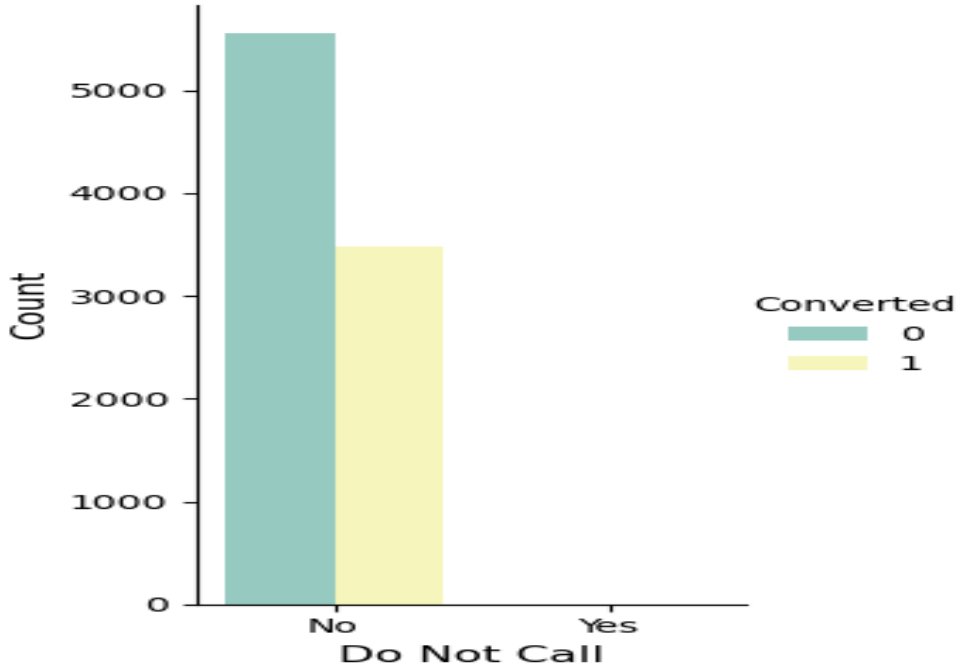


Major conversion in the lead source is from 'Google'

Do Not Email , Do Not Call Vs Converted

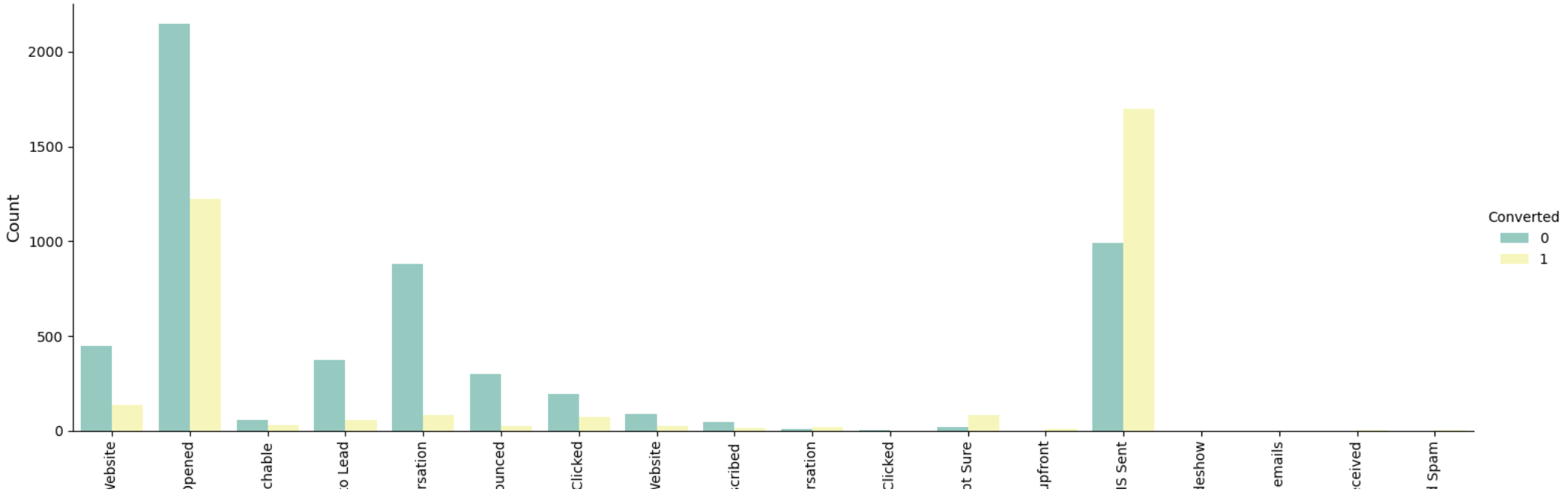


Major conversions happened when calls were made.



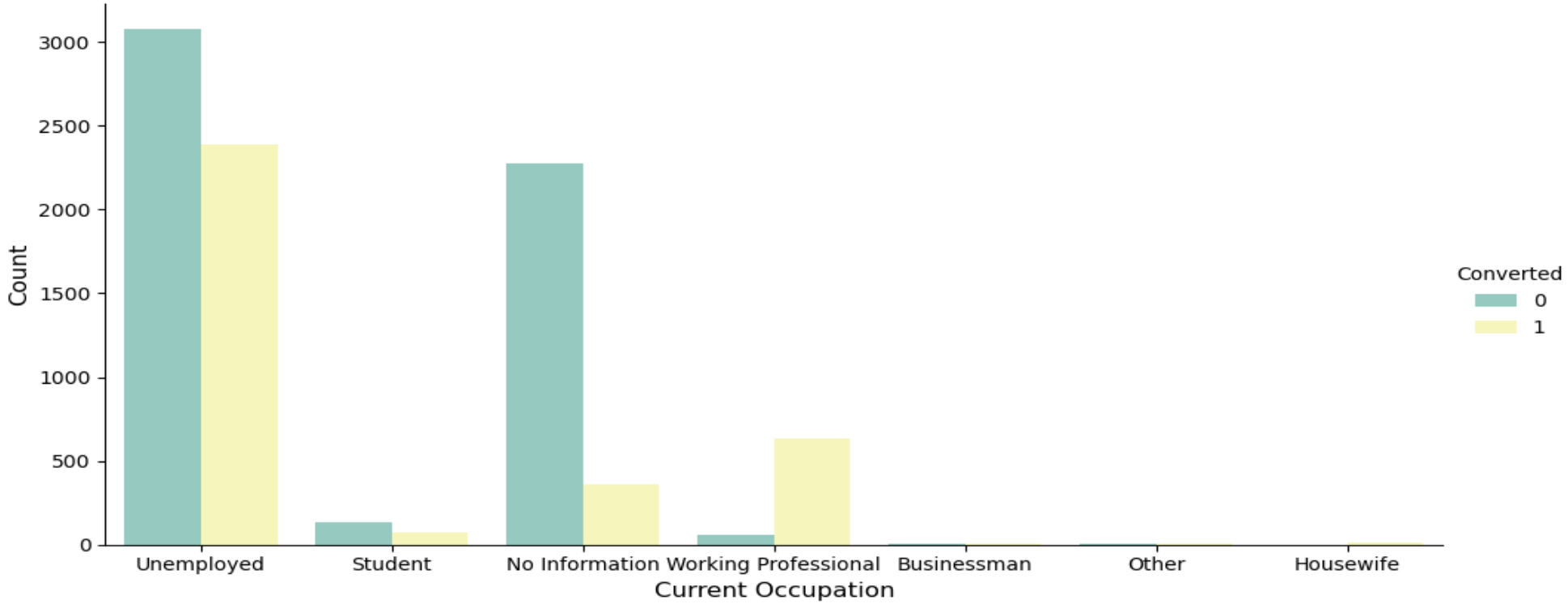
Major conversions happened when calls were made.¶

Last Activity Vs Converted



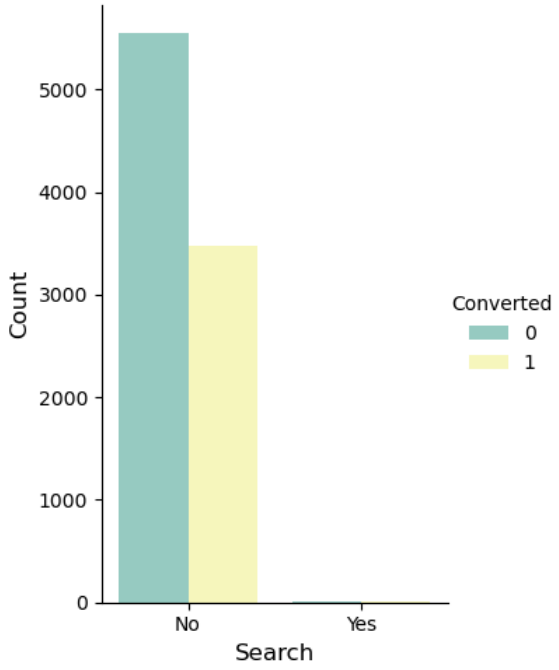
As per the above graph, last activity value of 'SMS Sent' had more conversion

Current Occupation Vs Converted

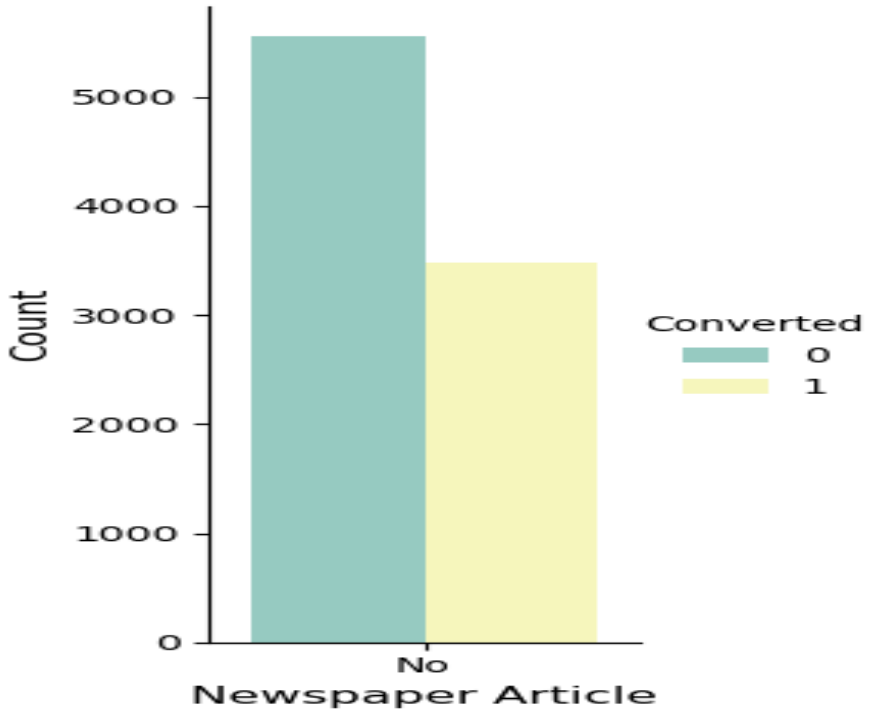


More conversion happened with people who are unemployed.

Search , Newspaper Article Vs Converted

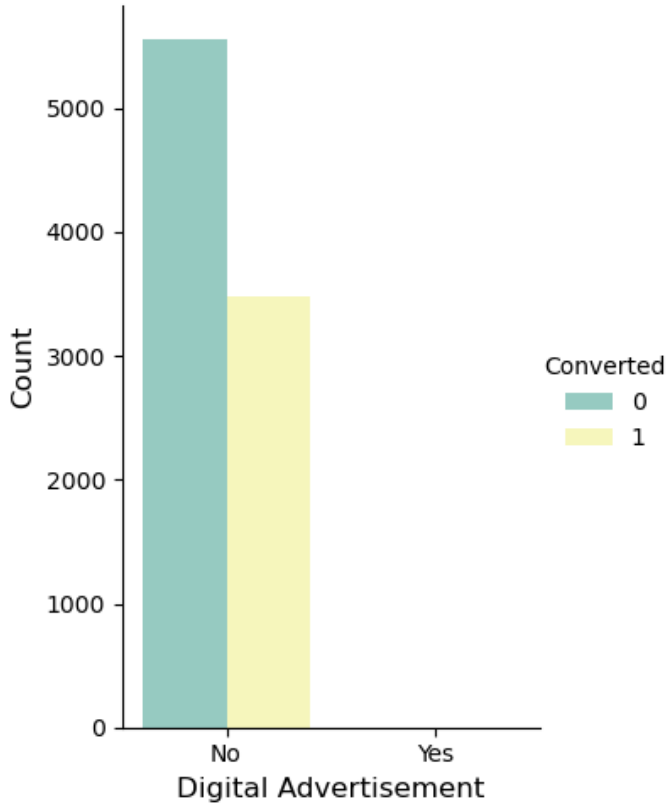


Conversion rate is high on leads who are not through search



Since "Newspaper Article" column now has only one value for all rows - "No" , it is safe to drop this column

Digital Advertisement



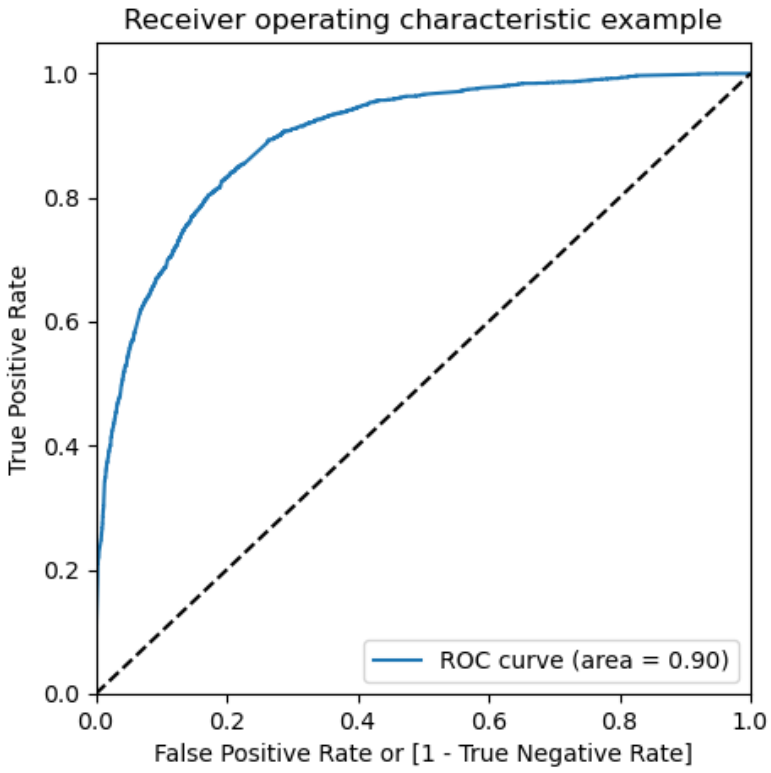
No Lead Generation from Digital Advertisement

Columns Dropped

Based on the univariate analysis dropped below mentioned columns for further analysis

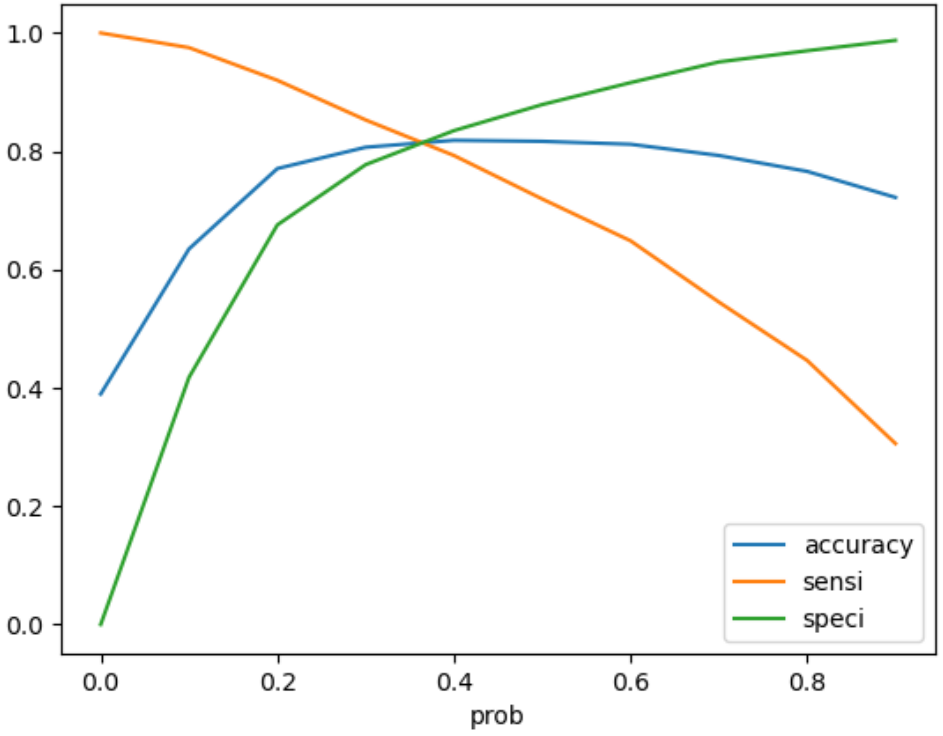
1. Country
2. Search
3. Digital Advertisement
4. Through Recommendations
5. A free copy of Mastering The Interview

ROC



The curve follows the left-hand border, higher (0.89) area under the ROC curve

Model Evaluation : Sensitivity and Specificity



Confusion Metrix

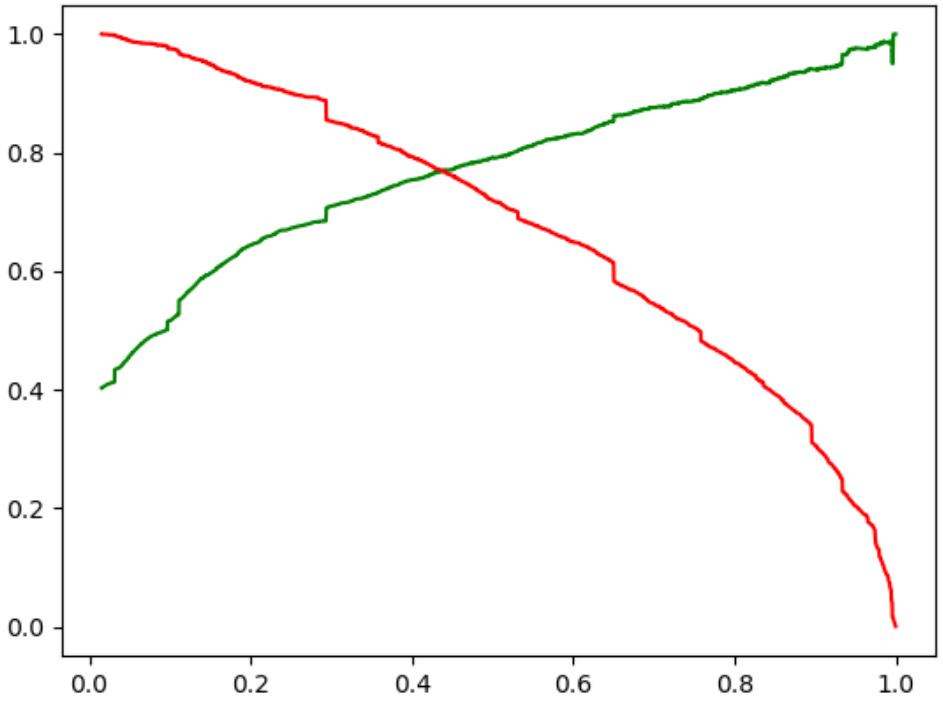
[3221, 637]
[510, 1952]

Train Data:

Accuracy : 81.8 %
Sensitivity : 79.28 %
Specificity : 83.48 %

Sensitivity : 0.7928513403736799
Specificity : 0.834888543286677
False Positive rate : 0.16511145671332297
Negative Predictive Value : 0.8633074242830341

Model Evaluation : Precision and Recall



Precision : 0.7913315460232351

Recall : 0.7193338748984566

Comparing Train & Test Values

Train Data:

Accuracy : 81.8 %

Sensitivity : 79.28 %

Specificity : 83.48 %

Test Data:

Accuracy : 81.8 %

Sensitivity : 79.01 %

Specificity : 83.47 %

Summary

- 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 81%, 79% and 82% 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%.
- The company should make calls to the leads who are the "working professionals" , who spent "more time on the websites", leads coming from the lead sources "Olark Chat" ,leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation", "Landing Page Submission", Specialization was "Others", who chose the option of "Do not Email" as "yes" as they are not likely to get converted.


thank you!