



Lead Scoring Case Study

**Submitted By: Yamika Jupally, Manish Kumar,
Jaikirat Singh Seth**

Problem Statement

- X Education sells **online courses** to industry professionals. 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.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is **around 30%** only which is very poor.

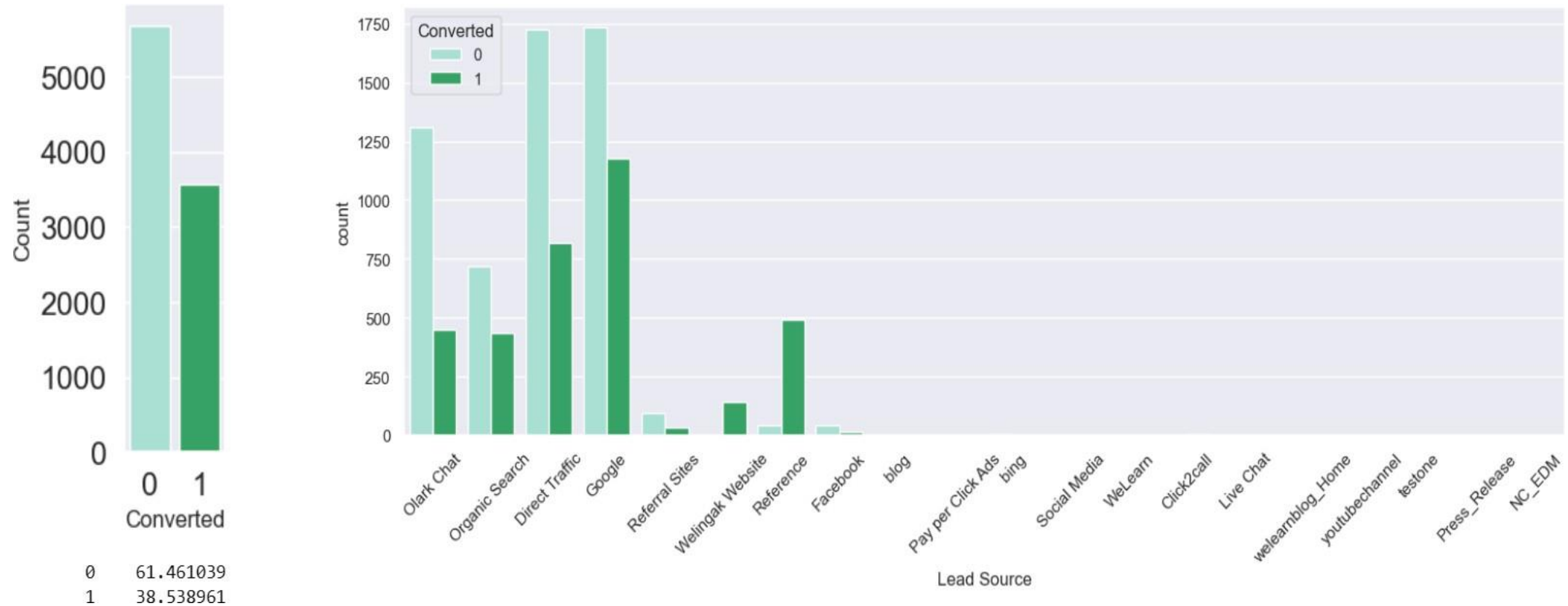
Business Objective

- Company wants make this process more efficient and wishes to identify the most potential leads, also known as 'Hot Leads'.
- Company want us to identify this set of leads, so that 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.
- 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.

Solution Strategy

- Collecting data and inspecting the Dataset
- Data Cleaning and Manipulation:
 - Checks for duplicate data.
 - Convert Select value to Null value.
 - Checks and handle Outliers.
 - Drop columns with large number of missing data.
 - Impute missing value, if necessary.
- Exploratory Data Analysis:
 - Univariate and Bivariate Analysis.
 - Creating Dummy Variable \ Binary mapping and Feature Scaling.
 - Logistic Regression is used for making model and predicting the result.
 - Model Evaluation
 - Conclusion and Summary of model.

Exploratory Data Analysis

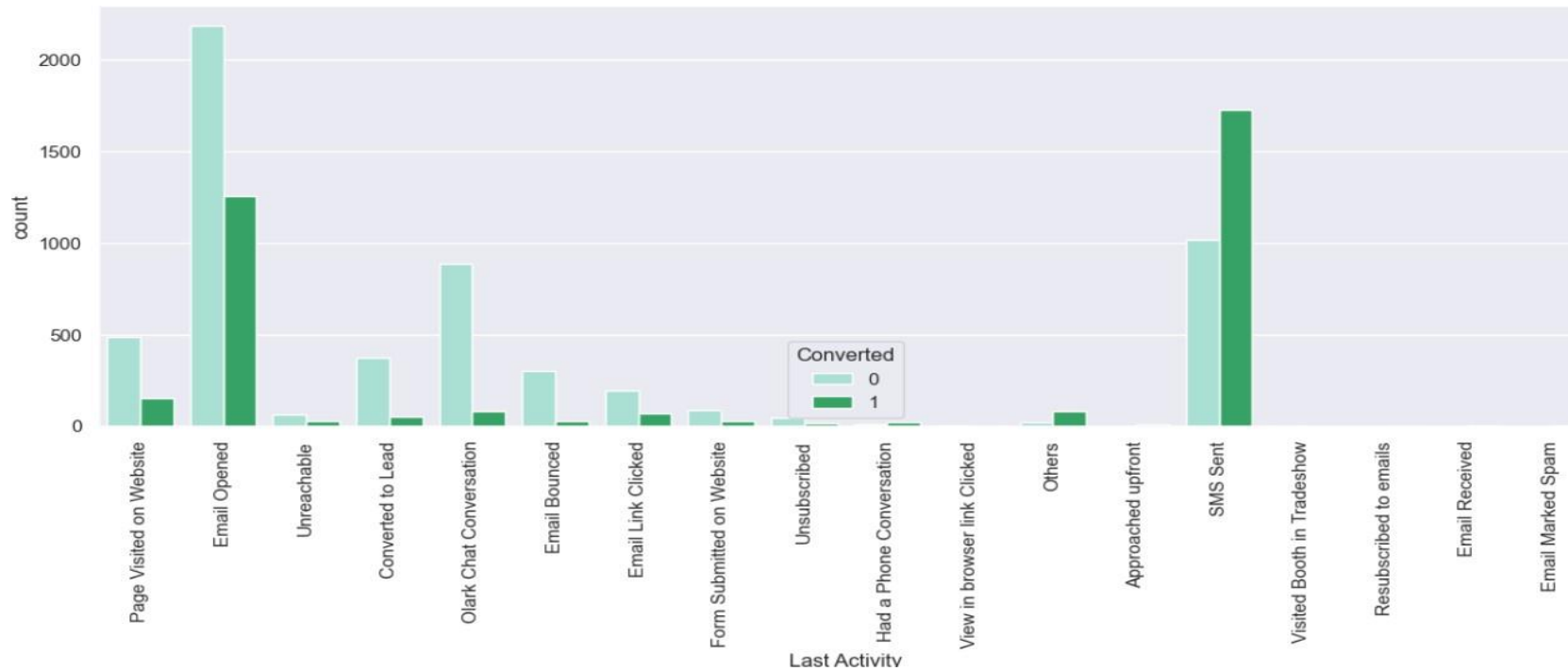


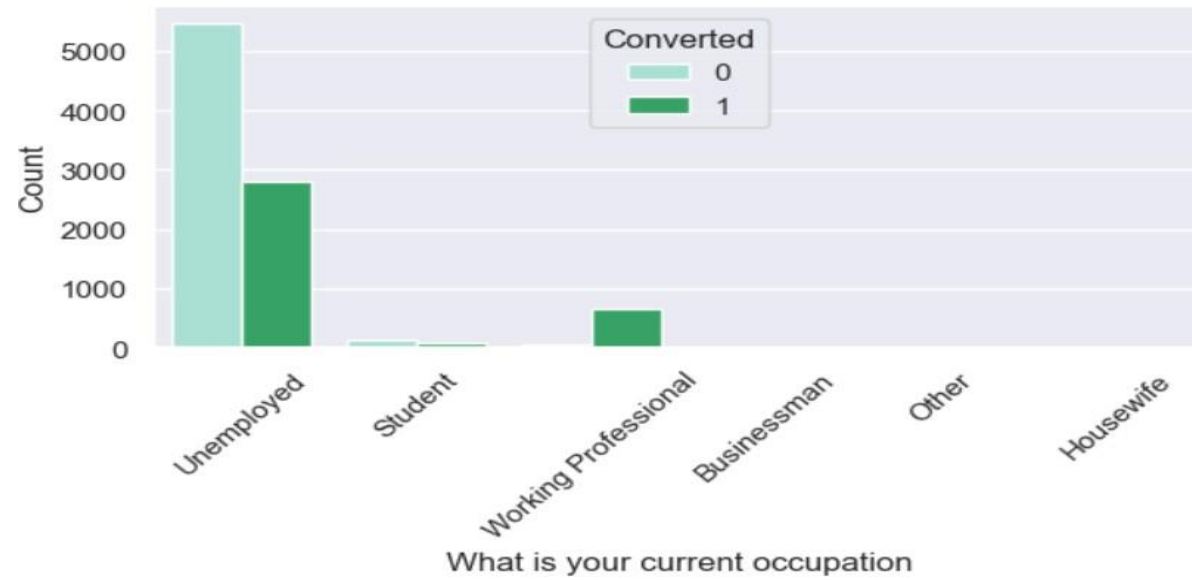
Total Converted rate in data is around 39% and most of the conversion happened through Google.



Most of the lead are converted which lands on a submission page followed by API. Conversion rate is high on Lead Add Form.

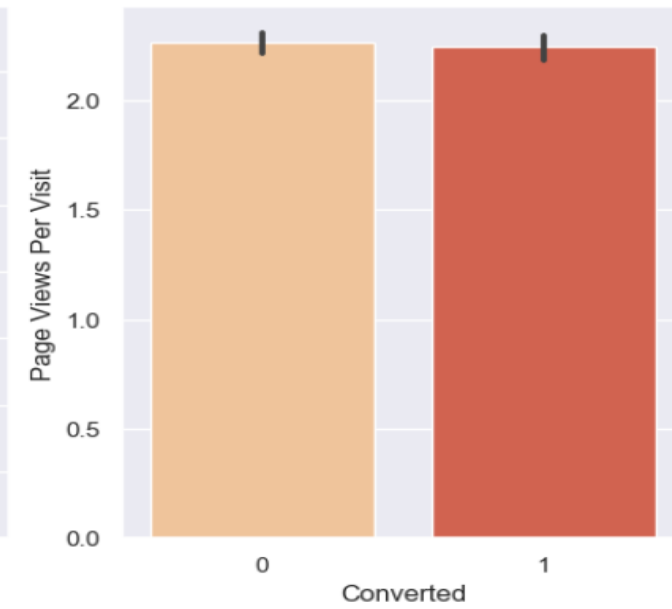
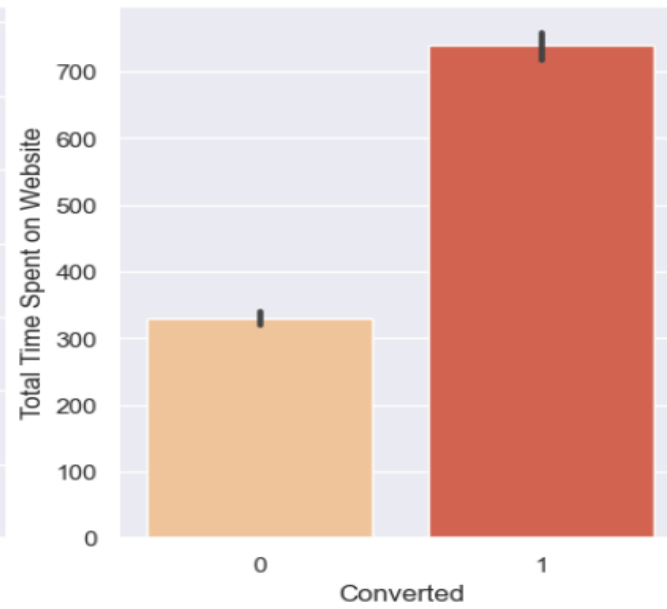
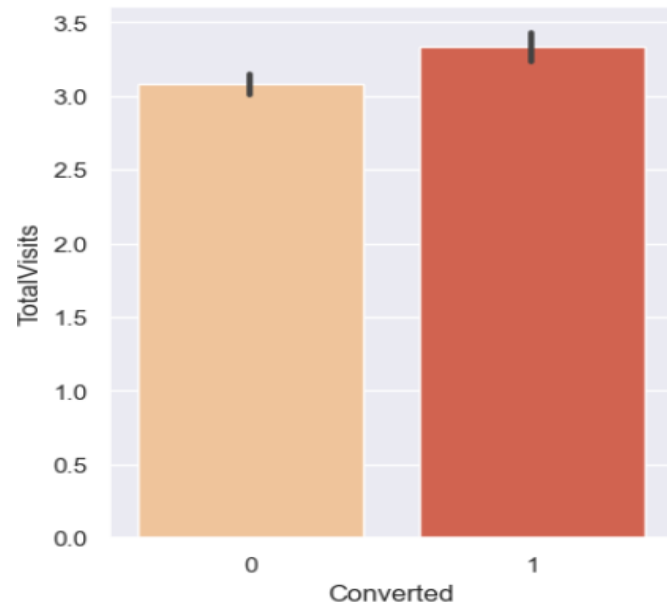
Conversion rate is very high for SMS Sent value, and we have more lead for Email Opened but the conversion rate is 50%.

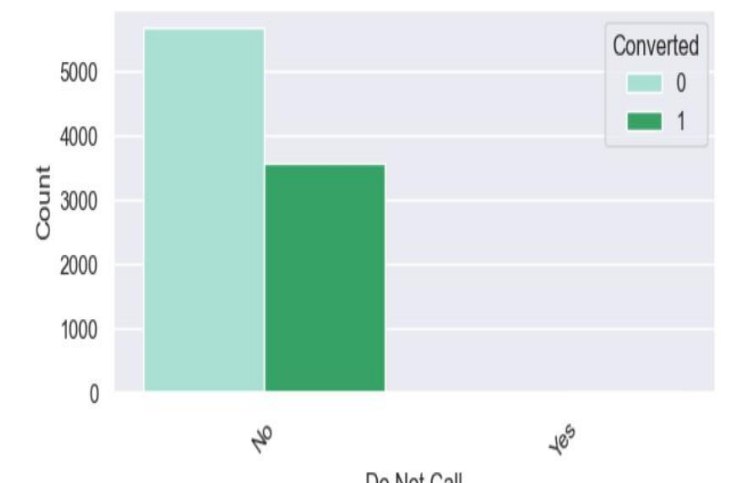
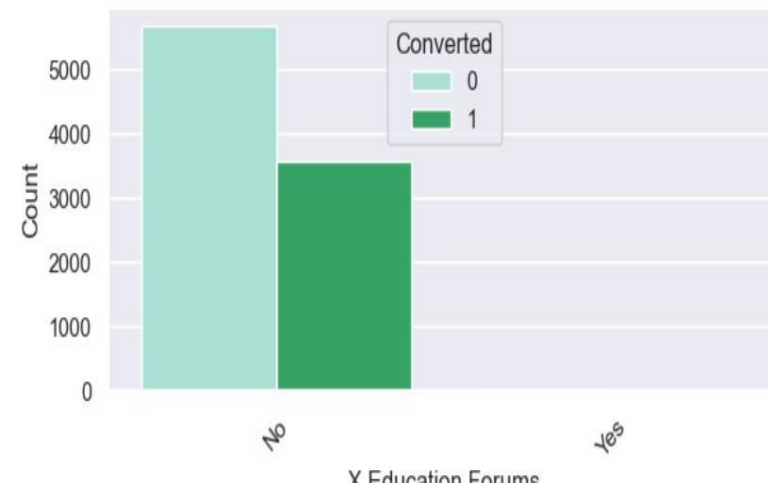
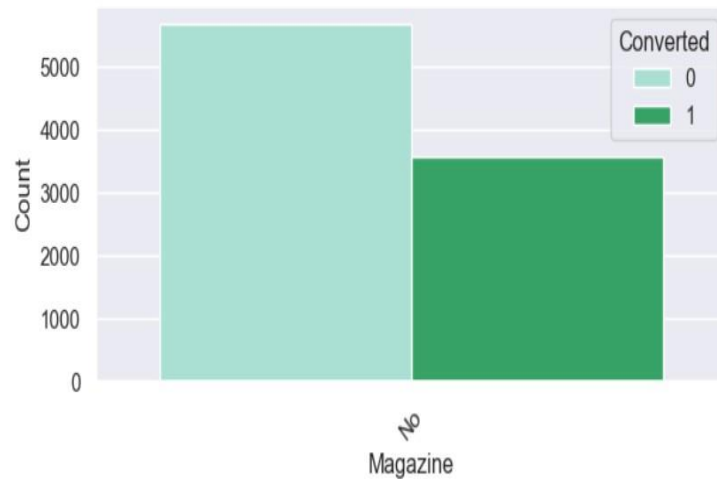
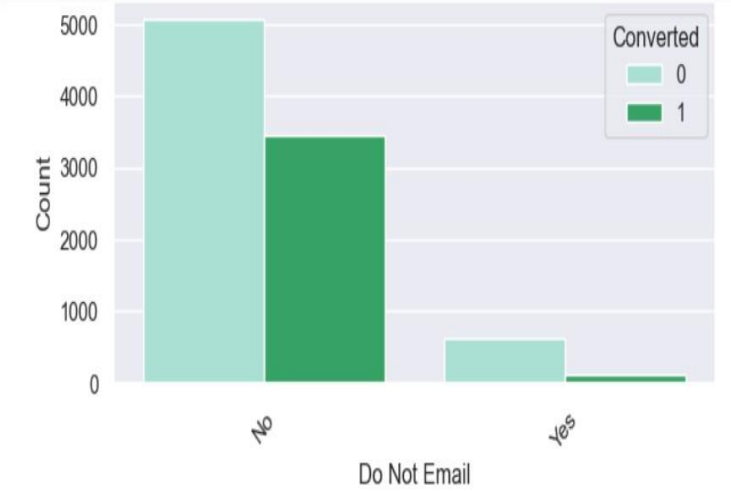
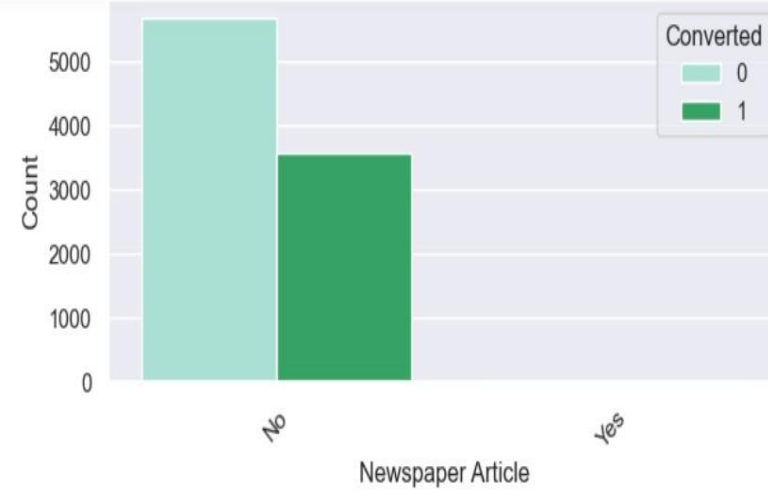
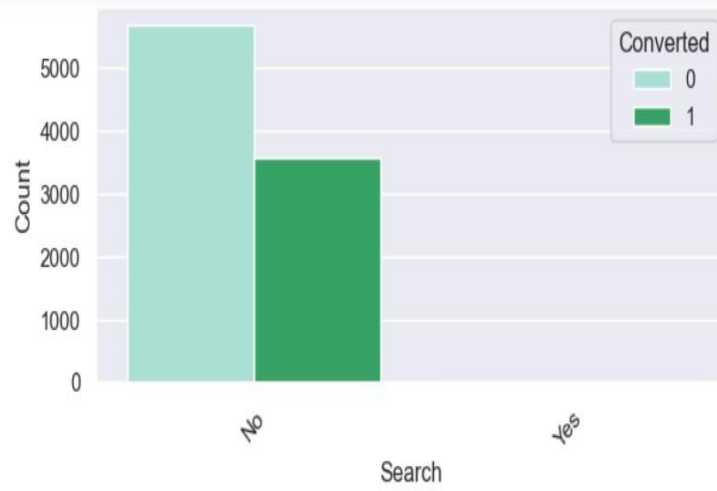




It is clear that high number of conversion are of Working professional, also we can see that the unemployed people are in huge numbers out which almost 50% got converted.

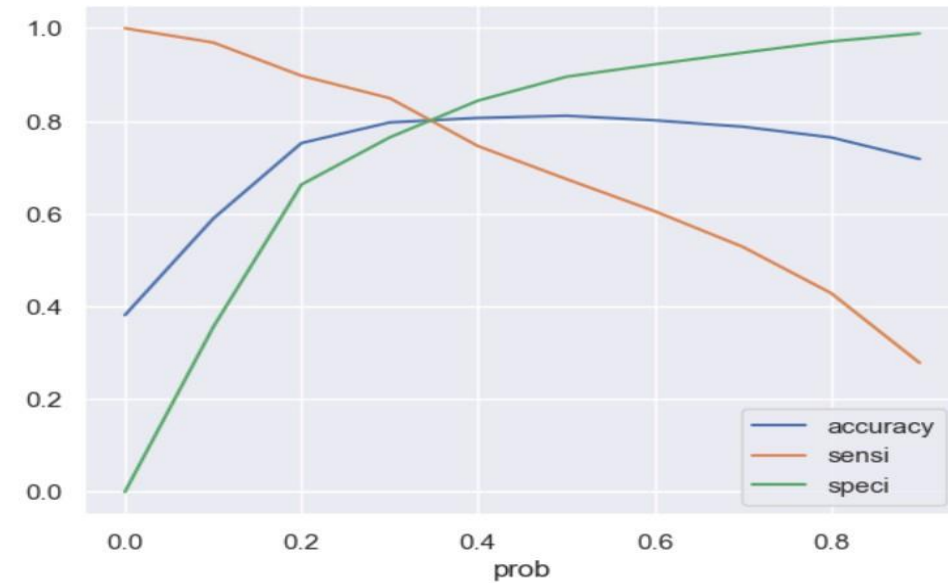
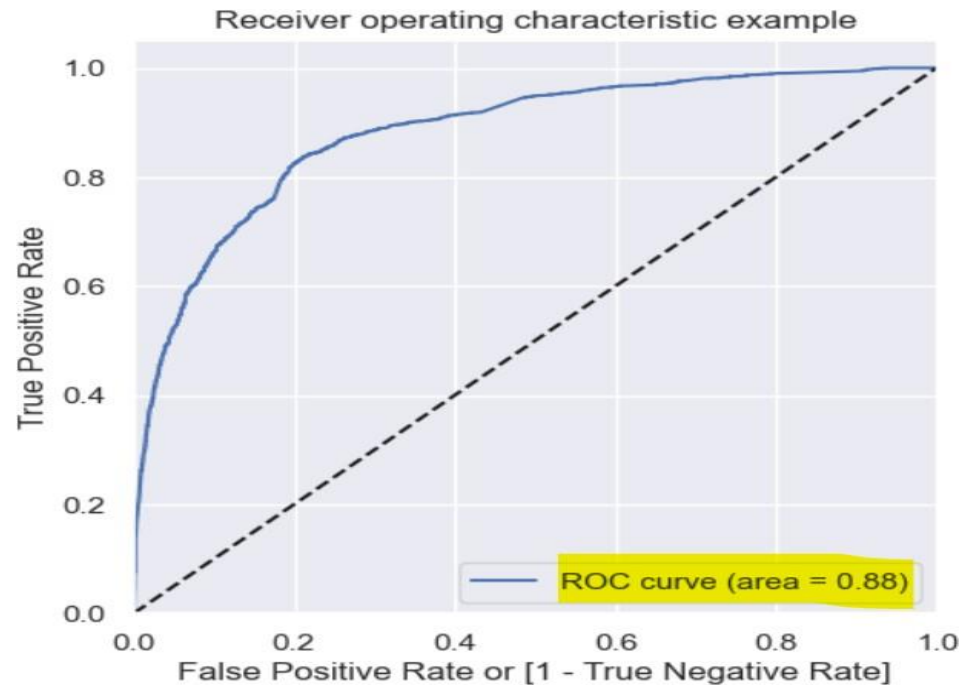
Conversion rate is much higher in Total Time Spent on Website feature.





From Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendation not able to draw any inference as almost 99% of the data has No in response.
For Do not Email and Do not call, almost all the people opted for No, but still some of the lead got converted.

Model Evaluation on Training Dataset

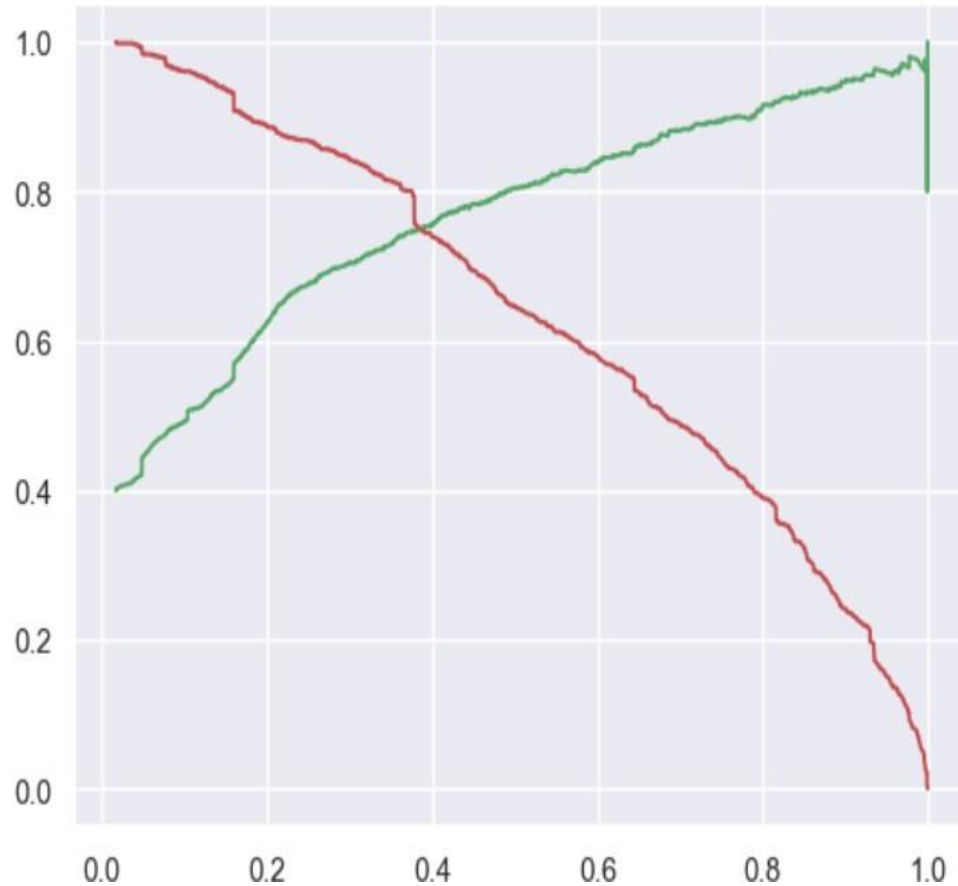


From the curve above, 0.37 is the optimum point to take it as a cutoff probability.

With the final Model, we get ROC curve area=0.88 which is very good and with cut-off value at 0.37 we get below data:

- Accuracy : 80.8 %
- Sensitivity : 80.2 %
- Specificity : 81.2 %

Model Prediction on Test Dataset



Optimal cut-off based on Precision and Recall is close to .38 which is close to ROC cut-off.

With the cut-off value at 0.37 we get below data for test dataset which is very similar to Train data hence the model is good and reliable:

- *Accuracy : 81.1 %*
- *Sensitivity : 80.09 %*
- *Specificity : 82.11 %*

Feature that are responsible to create model with positive coefficient:

- CurrentOccupation_Working Professional
- LastActivity_Had a Phone Conversation
- LastActivity_SMS Sent
- Total Time Spent on Website
- LastActivity_Email Opened
- LastNotableActivity_Unreachable

Feature that are responsible to create model with Negative coefficient:

- Do Not Email
- LeadOrigin_Landing Page Submission
- LeadOrigin_Lead Import
- LeadOrigin_API
- Specialization_Others
- Specialization_Hospitality Management
- LastNotableActivity_Modified

Conclusion:

- The overall model is able to predict the conversion rate up to 80% which was expected from it on both Train and Test data with cut-off 0.37
- Company should focus more on Working Professionals, People who spent more time on Website also where Lead last activity found as Phone Conversation an SMS sent, as they most likely to convert.
- Leads which are most likely to convert are came from:
 - Reference
 - Google
 - Weblinks
- Lead Add Form has most conversion rate in Lead Origin.
- Company should also look at negative coeff features to improv the conversion rate. And to offer something that will engage their attention.