LEAD SCORING CASE STUDY

Group members:

Thimmaiah K

Pramod Lohani

Ankur Katiyar

Problem statement

Problems for leads conversion faced by X Education

X Education sells online courses to industry professionals.

The company markets its courses on various platforms also X Education gets leads through past referrals.

The lead conversion rate for X Education is very poor, although they gets lots of leads by only able to convert 30% leads.

The company wishes to identify the most potential lead 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

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

Business objective

Business Objective for X Education

X Education needs a solution to identify the Hot Leads i.e., the leads that are most likely to convert into paying customers.

The company wants to build a model that will assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Deployment of the model for future use by the company.

Solution Approach and steps

Data Cleaning and manipulation: →

Understanding the data using data dictionary.

Check and handle missing values in the data set.

Dropping columns with more than 40% missing/Null values

Handling ‘NaN’ and ‘Select’ values in the data set.

Imputation of values and Handling outliers in the data set.

EDA: →

Creating categorial and continuous features from the data set.

Performing Univariate data analysis: value count, distribution of variables etc.

Performing Bivariate data analysis.

Feature Scaling and Creating Dummy variables for categorial columns. →

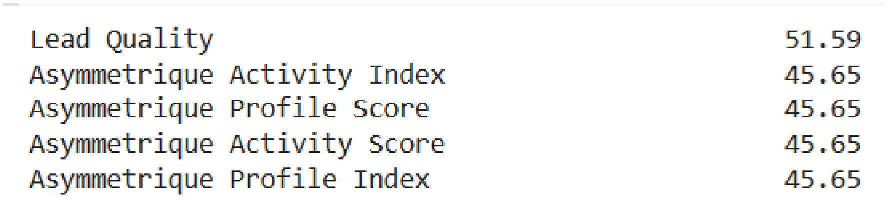
Building the model using Logistic Regression for prediction. →

Model Evaluation and presentation →

Conclusion and Recommendations.

Data manipulation

- Data set has 37 columns and 9240 rows.
- Dropped columns Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Score, Asymmetrique Profile Score & Asymmetrique Profile Score as more than 40% values were missing

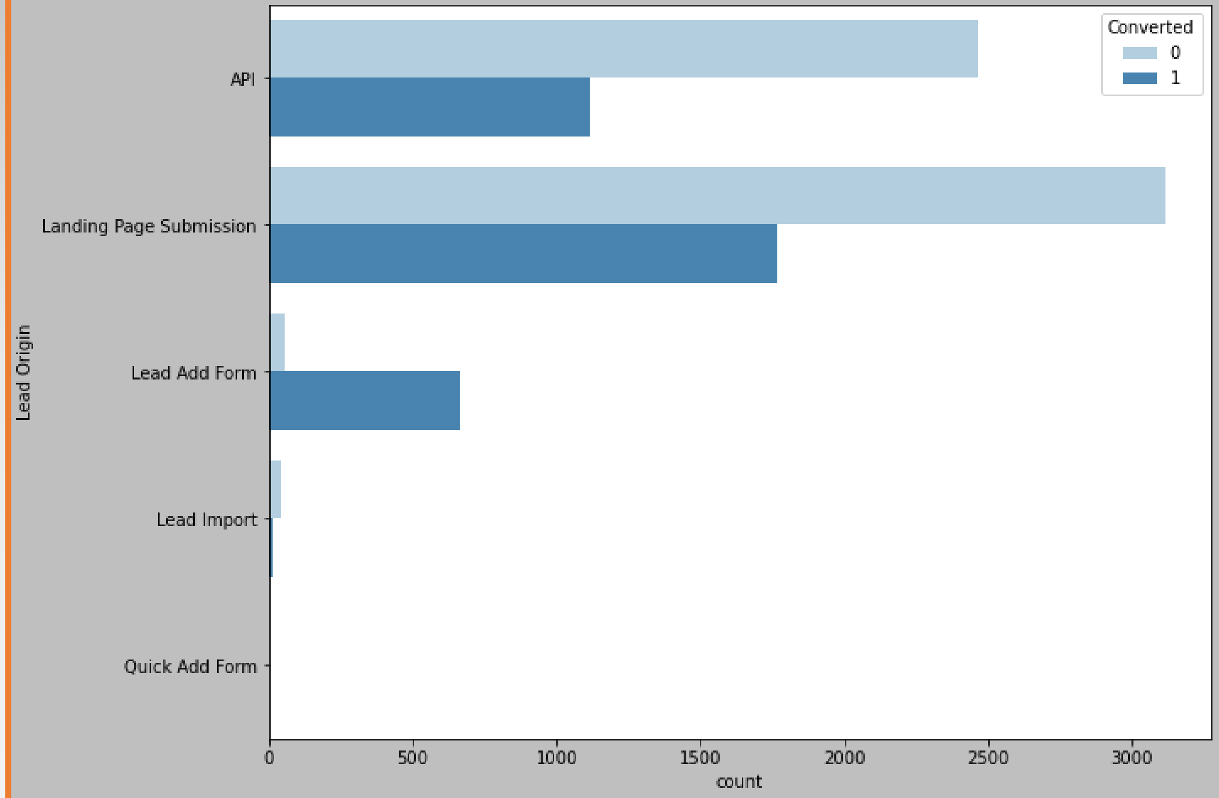


```
Lead Quality                                  51.59
Asymmetrique Activity Index                   45.65
Asymmetrique Profile Score                    45.65
Asymmetrique Activity Score                   45.65
Asymmetrique Profile Index                    45.65
```

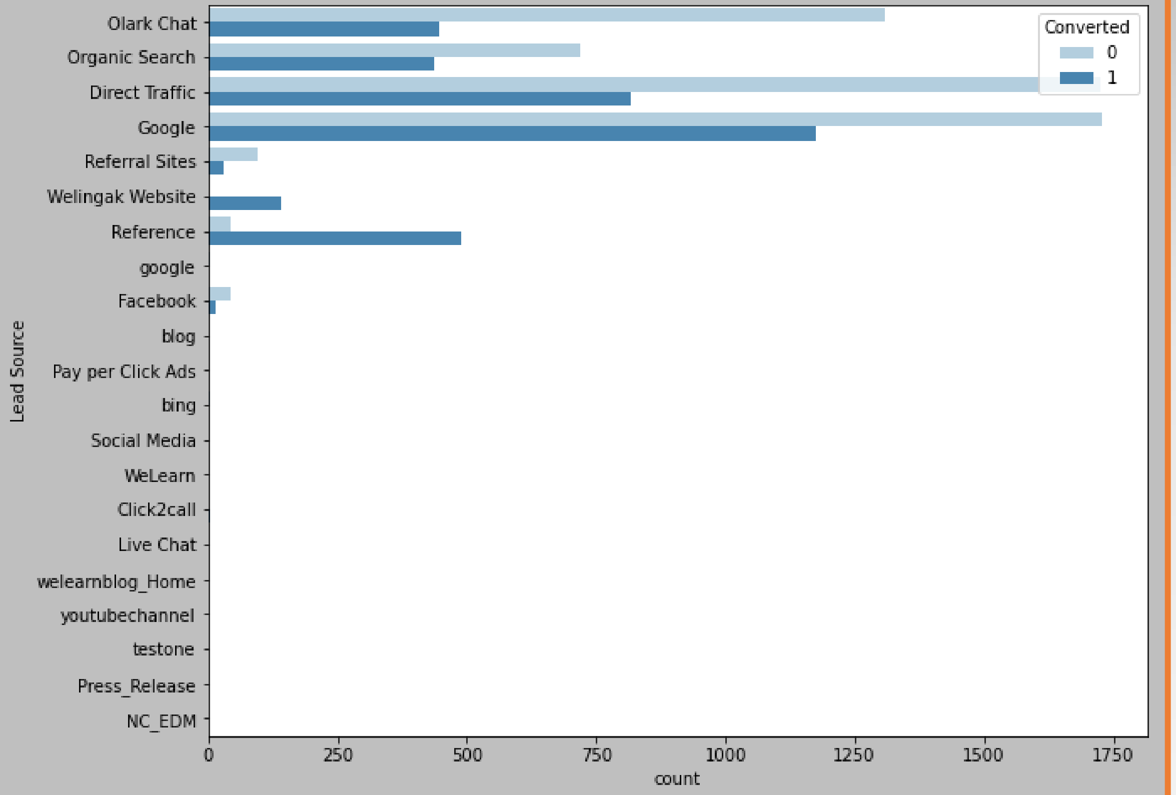
- Handling columns with values as ‘Select’. This value is as good as ‘Null’ so updated the value to ‘Not Selected’ for both ‘Select’ and ‘NaN’ values.
- Imputing Null values for column ‘Tags’, ‘What matters most to you in choosing a course’, ‘What is your current occupation’, ‘Country’, ‘Page Views Per Visit’, ‘TotalVisits’, ‘Last Activity’ & ‘Lead Source’.
- Performed EDA
- Dropped column Prospect ID and setting ‘Lead Number’ as Index.
- Dropped columns that only has one Value ‘No’or has single value that has more the 99% occurrence- columns dropped are Do Not Call, Search, Magazine, Newspaper Article, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Receive More Updates About Our Courses, Receive More Updates About Our Courses & Receive More Updates About Our Courses.

EDA – Exploratory DATA Analysis

EDA: Categorical Variable Relation

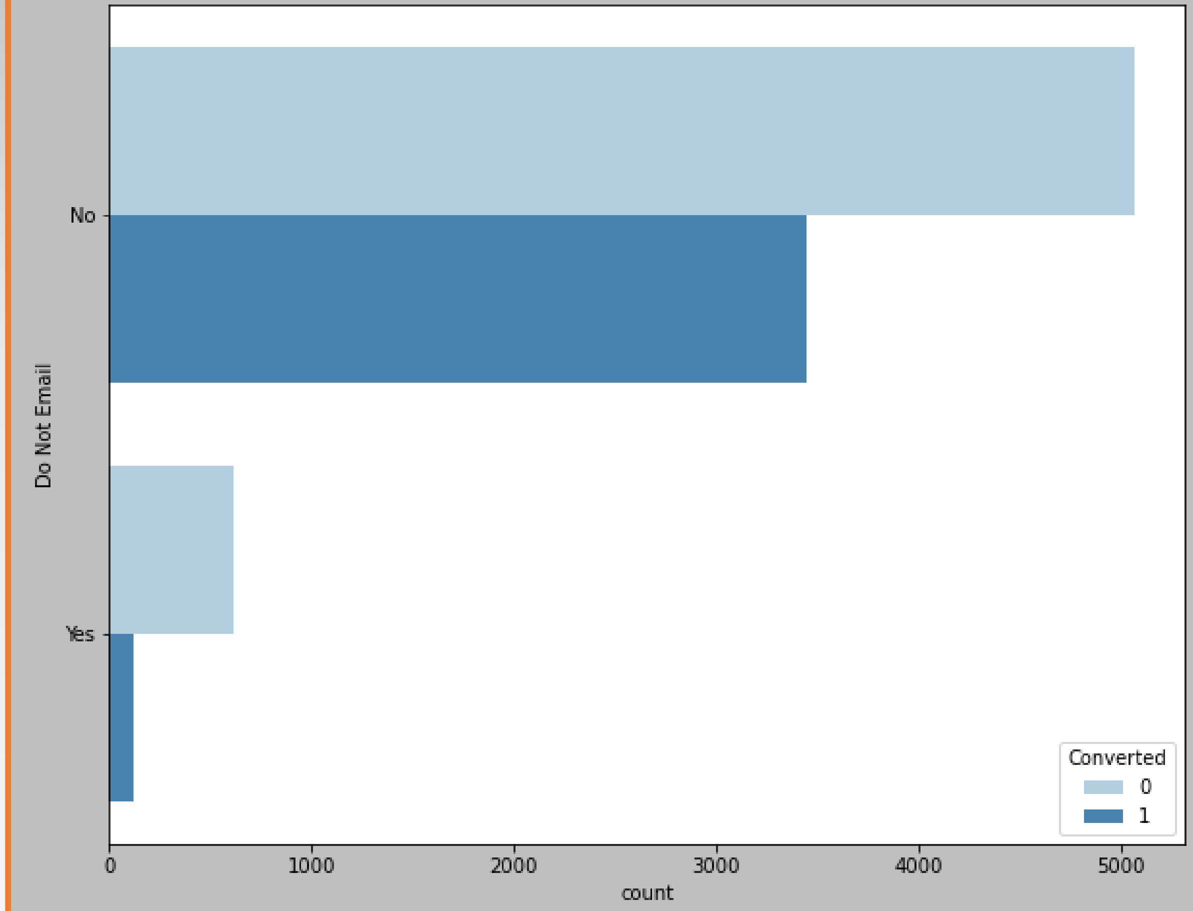


Lead Origin

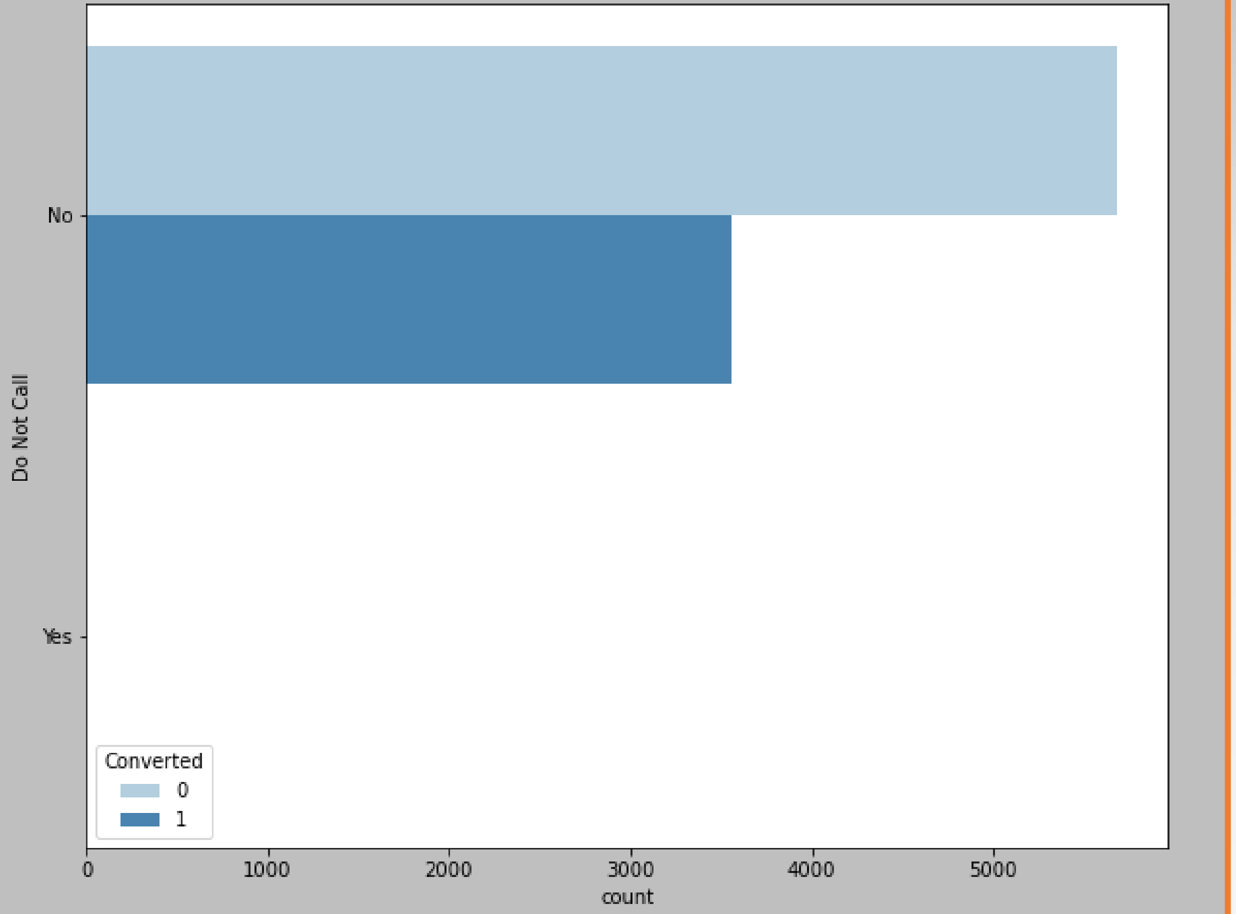


Lead Source

Categorical Variable Relation

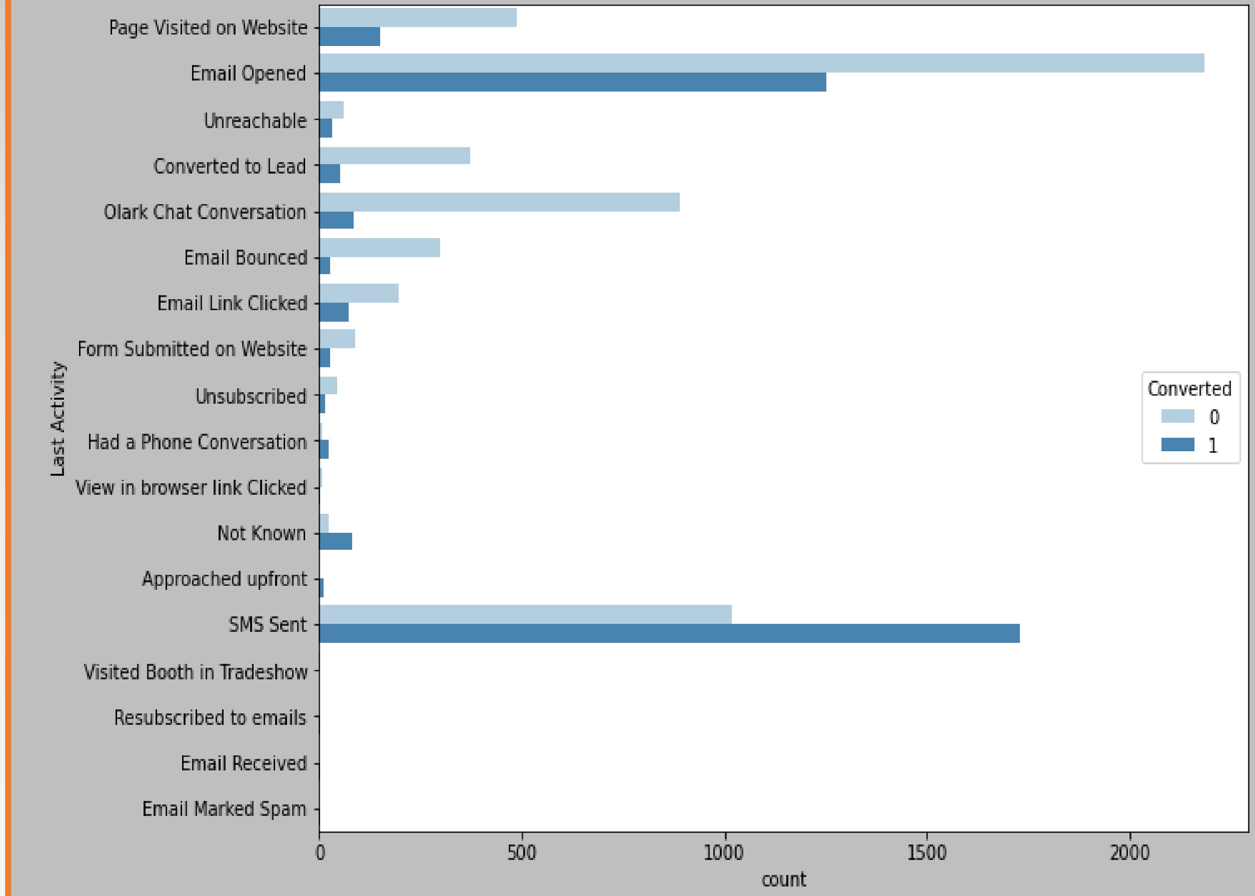


Do not Email

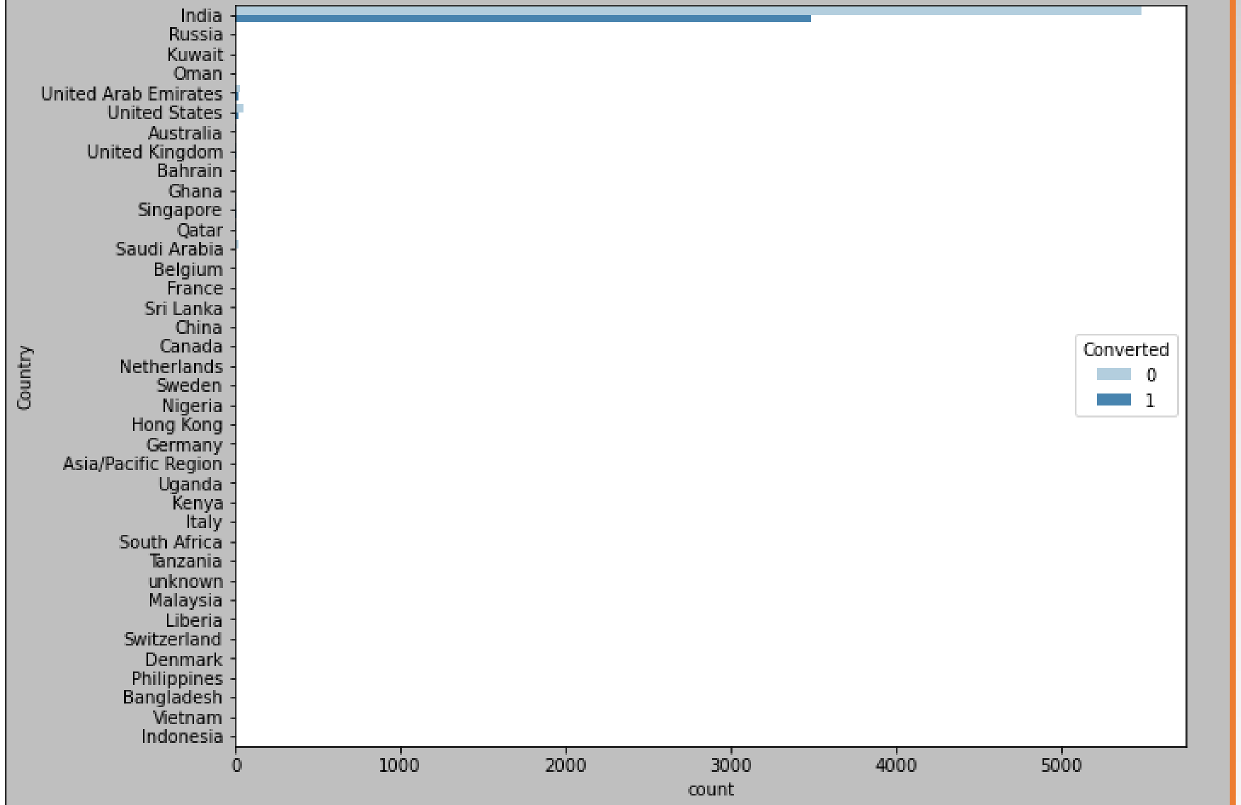


Do not Call

Categorical Variable Relation

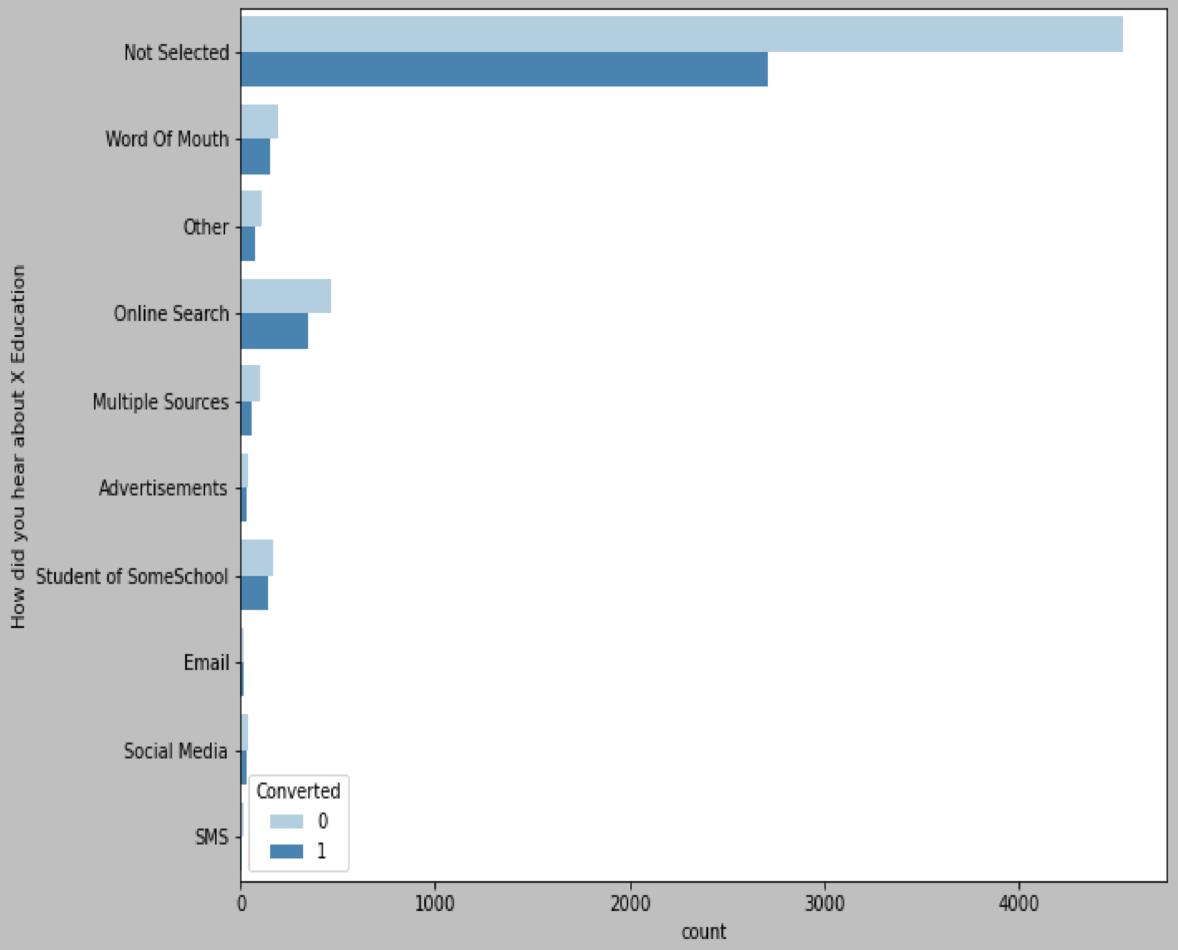


Last Activity

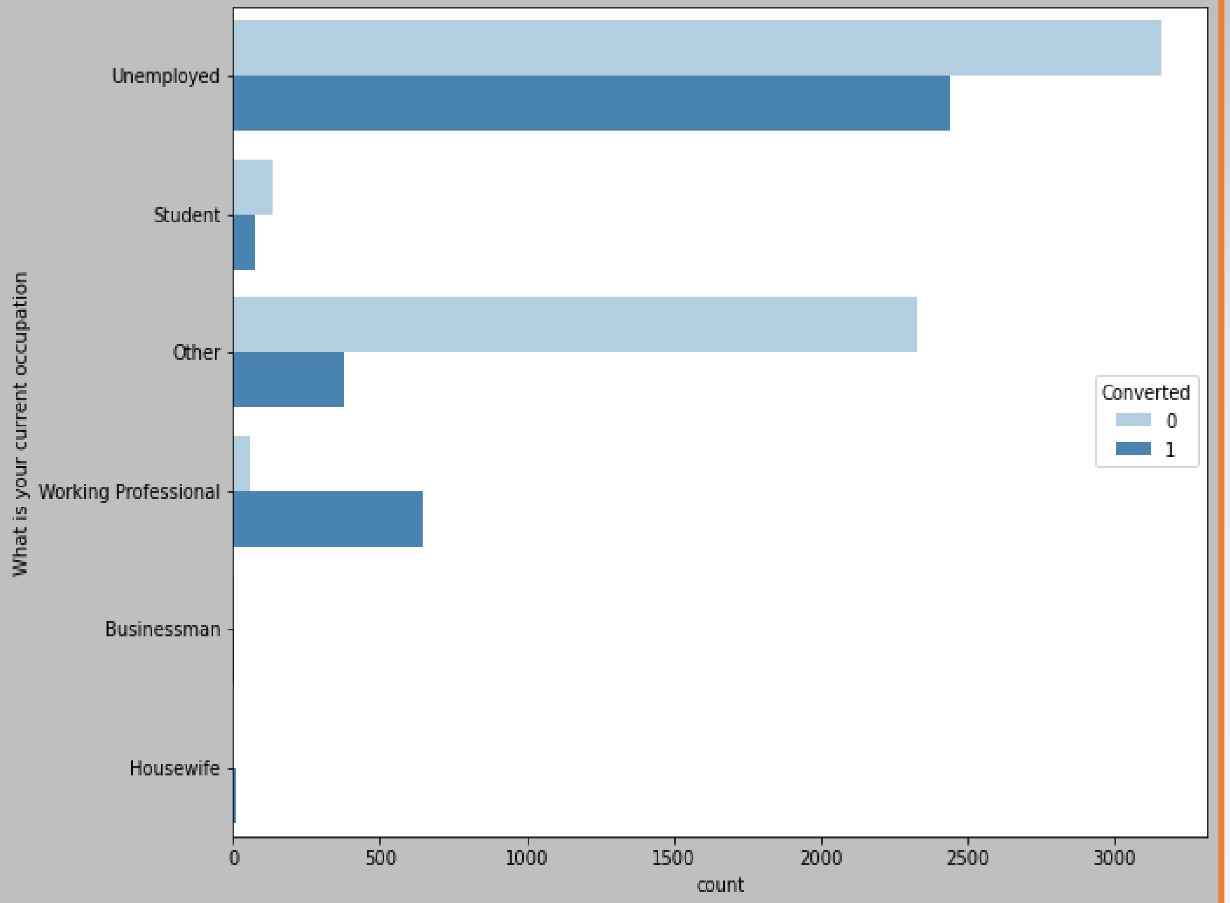


Country

Categorical Variable Relation

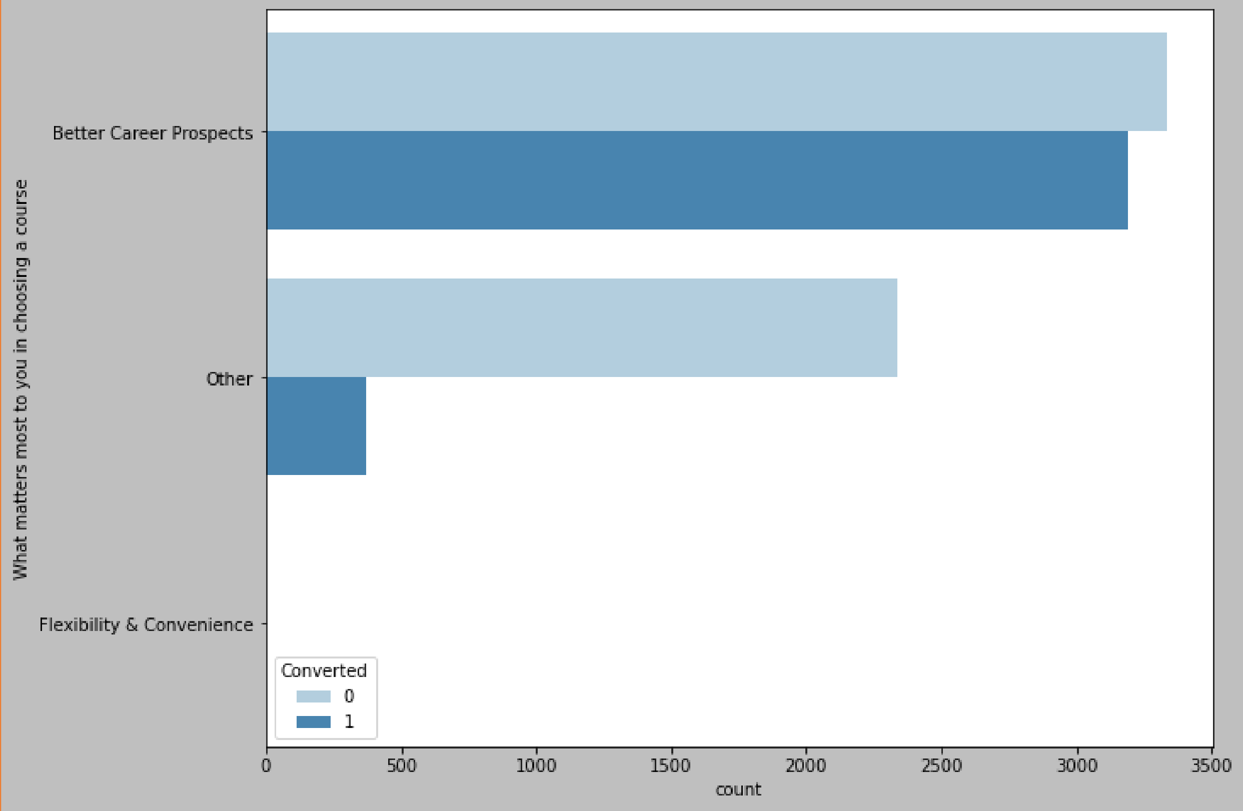


How did you hear about X Education

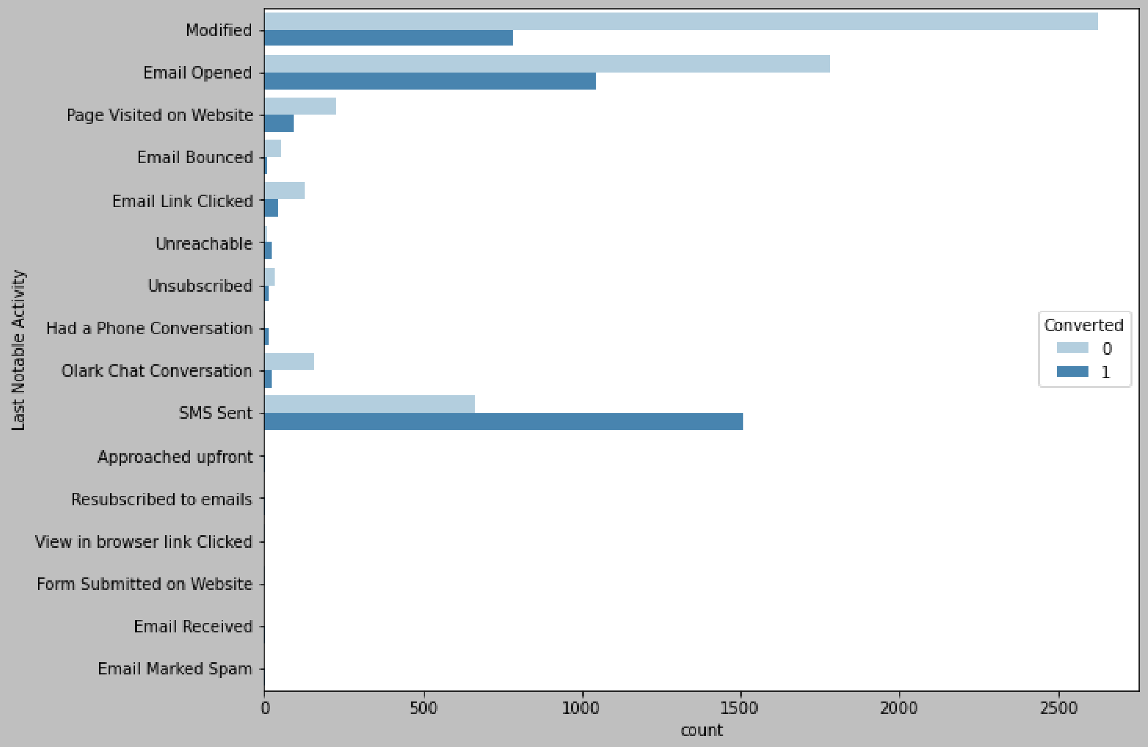


What is your current occupation

Categorical Variable Relation



What matters most to you in choose a course



Last Notable Activity

Insights from EDA

Insights for lead conversion

Lead Origin & Lead Source

Highest conversion in ‘Landing Page Submission’

High conversion from ‘Google’ followed by ‘Direct Traffic’.

Specialization

Highest conversion for 'Not Selected', this could mea
n these customers were not employed previously

Last Activity, Country & City

Conversion is highest among students who sent SMS.

Highest conversion of leads from India.

Mumbai has highest conversion rate.

Data Conversion

Data conversion for Model Building

- Numerical Variables are Normalized.
- Dummy variables are created for categorical variables.
- Dropped irrelevant columns after EDA
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 72

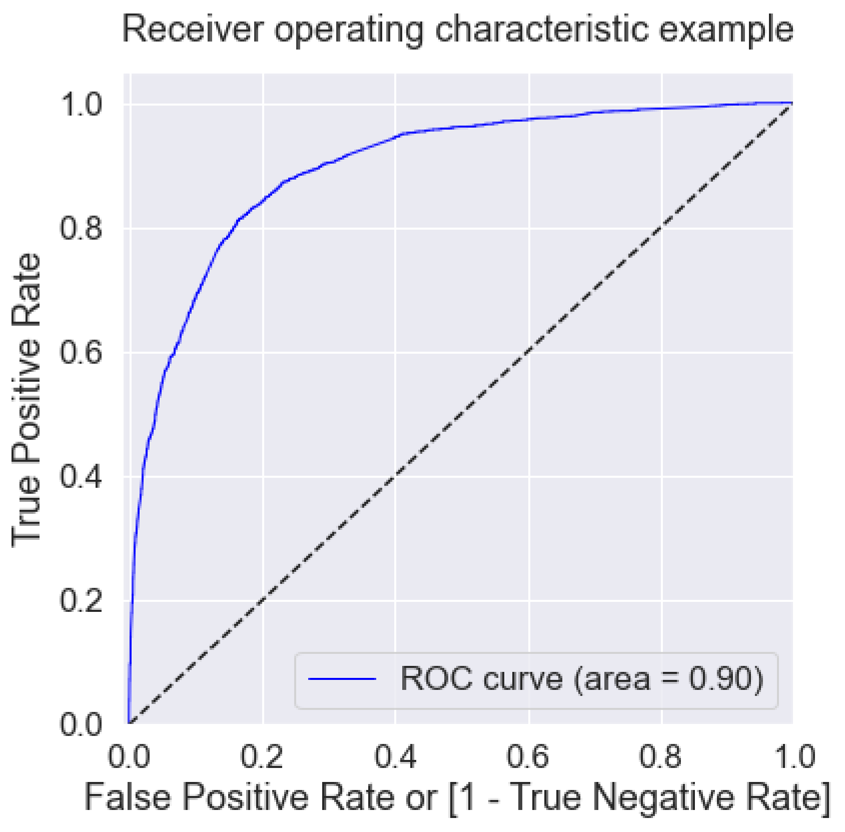
“BUILDING THE MODEL”

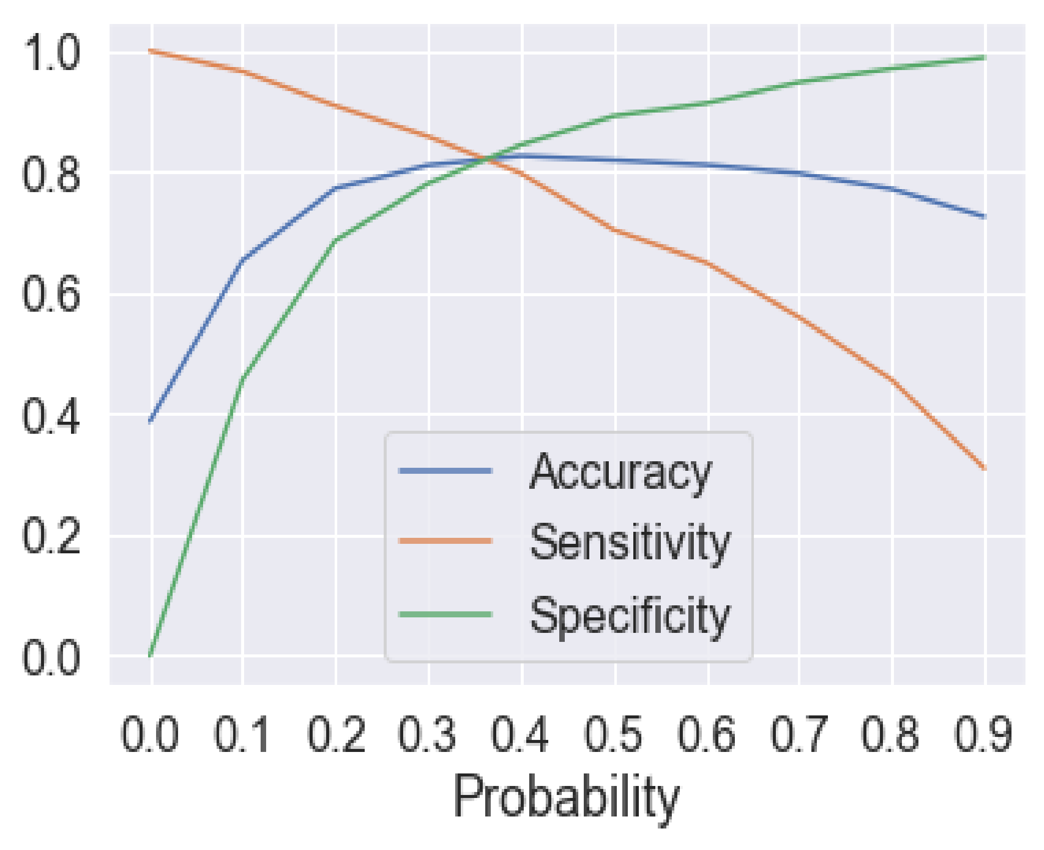
- Splitting the data into training and test set. Choosing train-test split using 70:30 ratio.
- Feature Scaling for TotalVisits, Total Time Spent on Website & Page Views Per Visit.
- Selecting Logistic Regression model since the target variable is binary.
- Using RFE and selecting 15 variables for model.
- Dropping column ‘What is your current occupation_Housewife’ and ‘Lead Profile_Lateral Student’ as the p-value is greater than 0.05.

“Model Evaluation”

- Model 3 is the most suitable model with p value as 0 for all the columns.
- VIF score is also less then 5%.
- Overall Accuracy score is 82%

ROC Curve





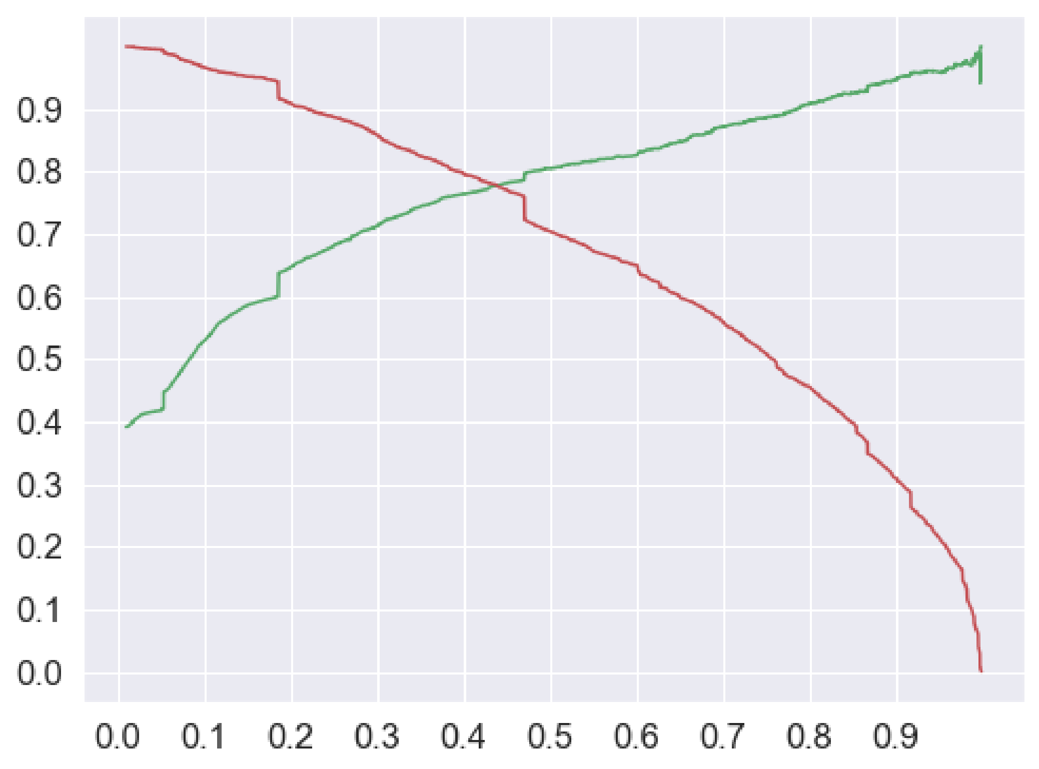
Finding Optimal Cut off point

- From the curve above, 0.36 is the optimum point to take it as a cutoff probability.

PRECISION AND RECALL TRADEOFF

Final Model Prediction

- Precision value is evaluated as 0.75 and Recall value is 0.82 in the final model.
  - Final accuracy score of the prediction is 0.82.
  - Sensitivity % = 80.0
  - Specificity % = 82.0
  - Precision Score % = 73.0
  - Recall Score % = 80.0
- F1 = 2 * (precision * recall) / (precision + recall) = 0.77



Summary Report

- *The below features contributed most towards the probability of a lead getting converted*
  1. *The Total Time Spent on Website*
  2. *Total Number of Visits*
  3. *Most numbers of leads were converted from Lead Source_Welingak Website*

- *Maximum traffic for Leads came from:*
  - *Google*
  - *Direct traffic*
  - *Organic Search*
  - *Olark Chat*
  - *Welingak website*

- *X Education should focus the most on the below points in order to increase the probability of lead conversion*
  - *Focus more on the traffic coming from Welingak Website.*
  - *Lead probability is highest where the lead origin Is Lead Add Form*
  - *Should target more where the current occupation is Working Professional*