X Education – Lead Scoring Case Study

Improving Potential lead conversion rate, to boost the revenue of X education company

Case Study Conducted By: Gursewak Singh, Kartikey Mishra and Shivam Sikka

Table of Content

X Education Company Background

- X education company sell online courses to industry professionals
- Company markets its courses on several websites and search engines
- People visiting website fills the form with email and phone number. These people treated as leads
- These leads are then contacted by sales team through email or phone.
- With this process only 30% of the total lead get converted.

Problem Statement and Objective

Problem Statement

- Even though X education get lots of lead but poor conversion rate of 30%
- X education wants to make the process of lead conversion efficient, by focusing only on potential lead (also called Hot Leads)
- Sales team of X education wants to focus on hot lead, instead of making useless phone calls.

Objective of the Study

- In order to assist X education to identify promising leads, we need to build a model wherein leads are assigned lead score. Higher lead score higher the chances of its conversion.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%

Analysis Approach

- **Data Preparation:** encoding categorical variables and handling null values
- **EDA**: Univariate Analysis, Bivariate Analysis, Outlier Detection, Checking Data imbalance
- **Dummy Variable Creation**
- **Test Train Split**
- **Feature Selection**
- **Correlation Analysis**
- **Model Building:** Feature selection manually, Model Improvisation
- **Model Finalization**
- **Model Evaluation**
- **Recommendation**

Since, we have target of 80% conversion rate or better, we would want obtain a high sensitivity of obtaining hot leads. Its important for our model to predict the potential lead that can be converted.

Data Preparation

- Encoding variable, such as "Do Not Email", "Do Not Call" etc with values yes as 1 and no as 0, so that its easier to deal with will performing EDA and during model Building phase.
- Replace all the SELECTS for NaNs, as users didn't select any option from list, therefore Select is good as NaN
- Standardizing String Value in terms of casing and spellings, for example "Google" and "google" are same. But treated different while performing EDA. Required to bring them on same scale, i.e same casing.
- Dropping all the columns with missing value above 70% and then selectively drops columns with missing value above 40%.
- Impute the categorical variable's missing value (less than 40%) with most frequent value, for example City column had above 90% value as Mumbai. Therefore imputed missing value with Mumbai
- Columns with less than 1% of missing values simply drops the rows.

EDA: Exploratory Data Analysis

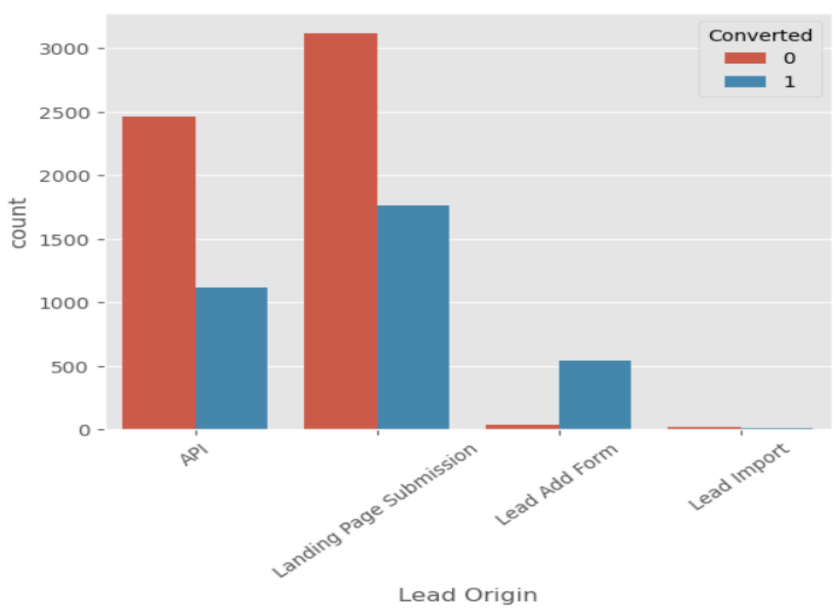
- Data imbalance check for target variable

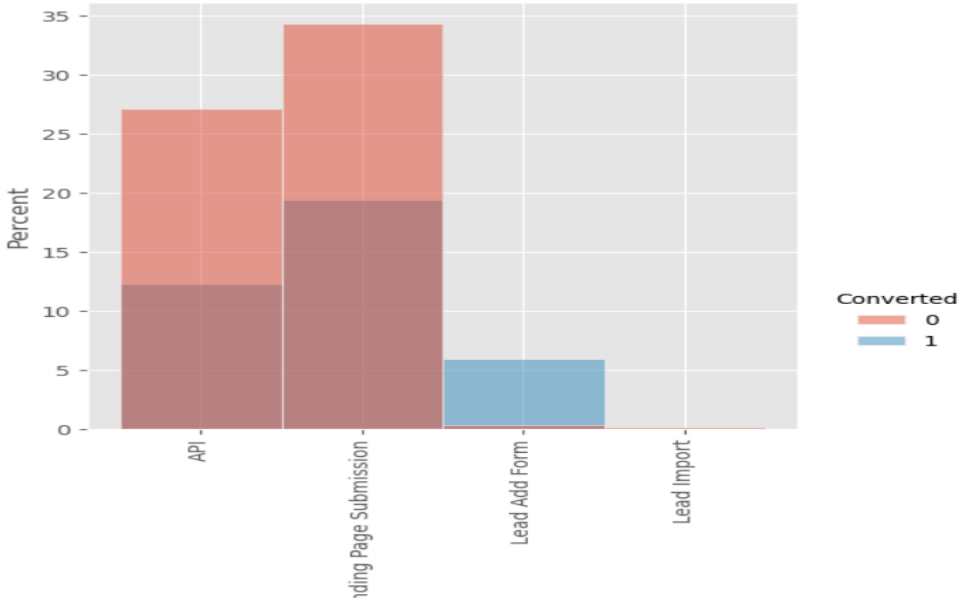
```
]: lead_df["Converted"].value_counts(normalize=True)

]: 0    0.621446
   1    0.378554
   Name: Converted, dtype: float64
```

- Lead that are not converted (0) are 37%, whereas Leads that are converted are 62%
- Since data is not highly imbalance we safely move further with analysis

EDA: Exploratory Data Analysis

- **Univariate and Bivariate analysis:** Lead Origin

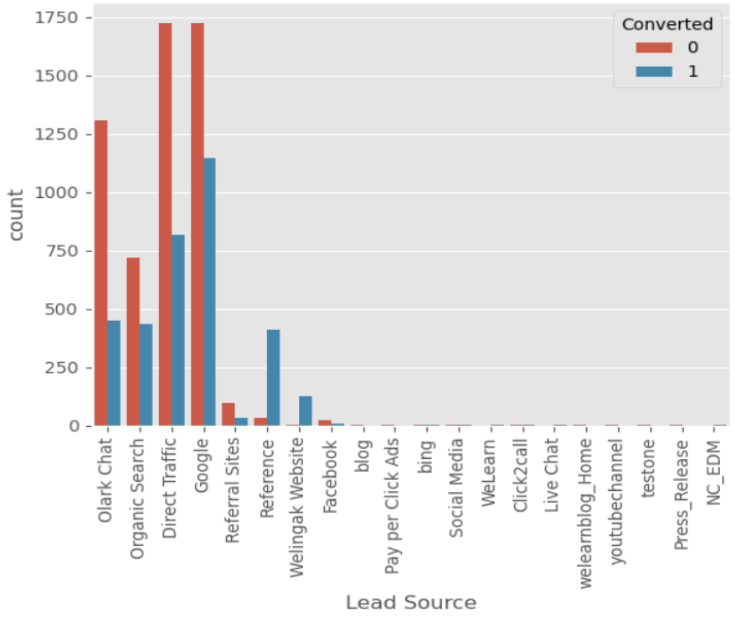


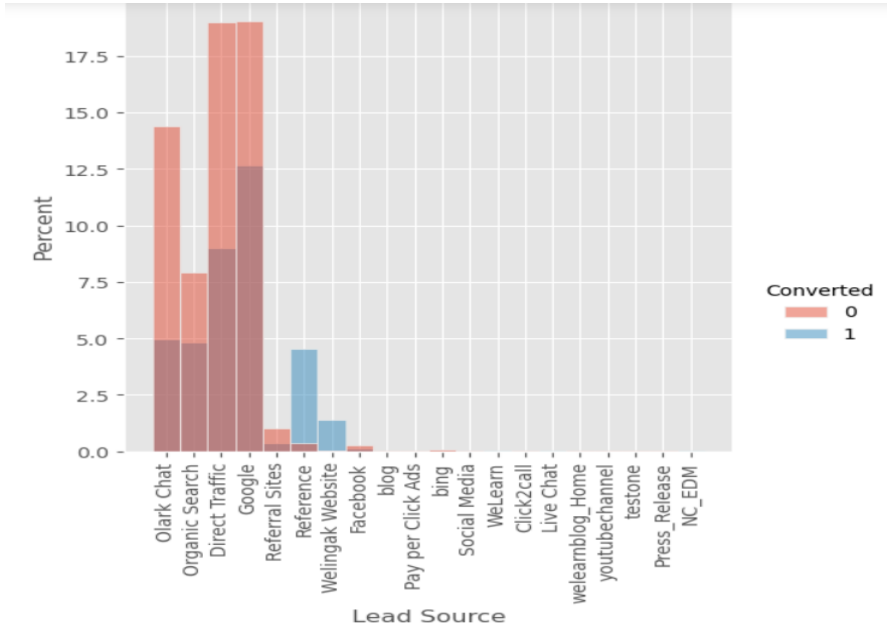


- API and Landing Page Submission has less conversion rate (~13% and ~20%) but count of lead is quite noticeable
- For Lead Add Form, although the conversion is high but have poor lead count
- For Lead Import count as well as conversion rate is ignorable

EDA: Exploratory Data Analysis

- **Univariate and Bivariate analysis:** Lead Source

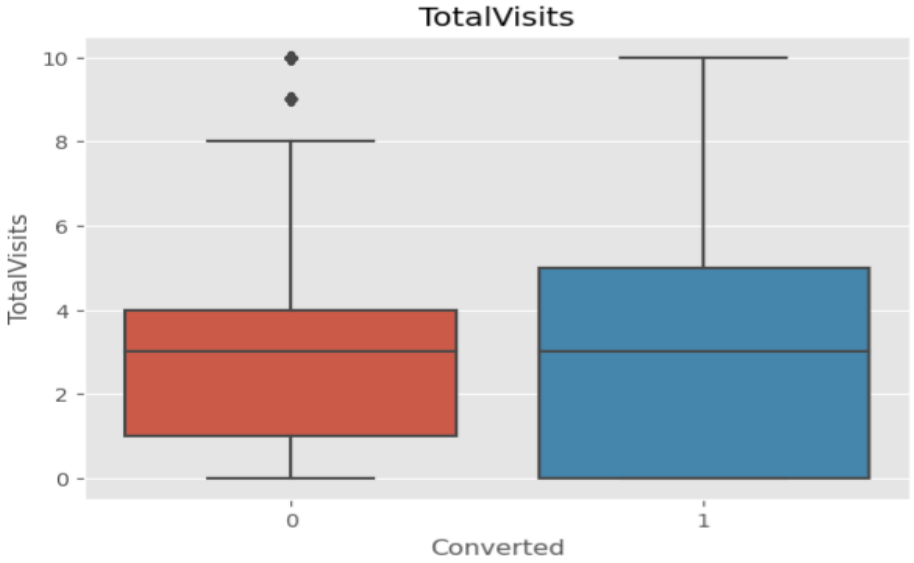


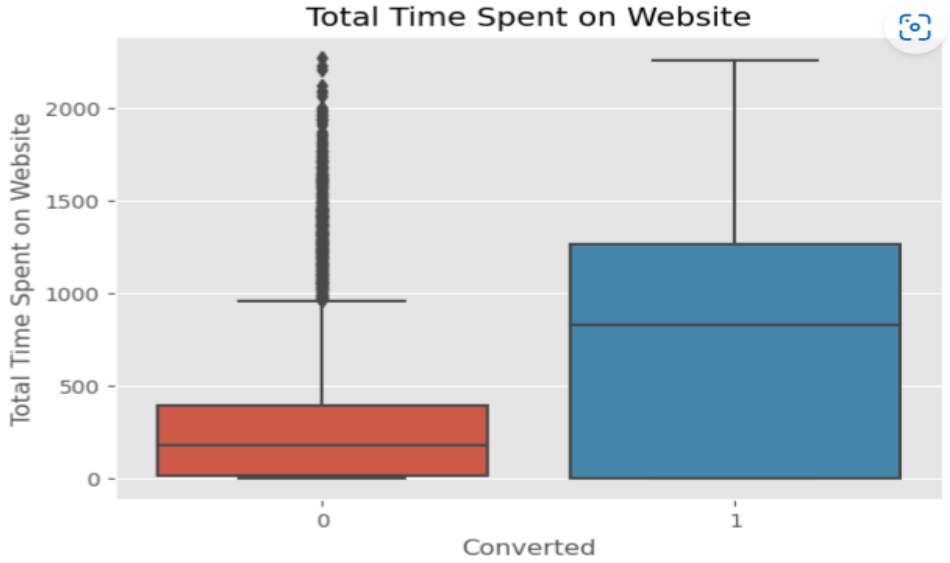


- As noticeable, count from Google, Direct Traffic is high
- Conversion rate of Reference and Welingak Website is maximum even though having poor count lead

EDA: Exploratory Data Analysis

- **Univariate and Bivariate analysis:** Total Visits and Total Time Spent on Website

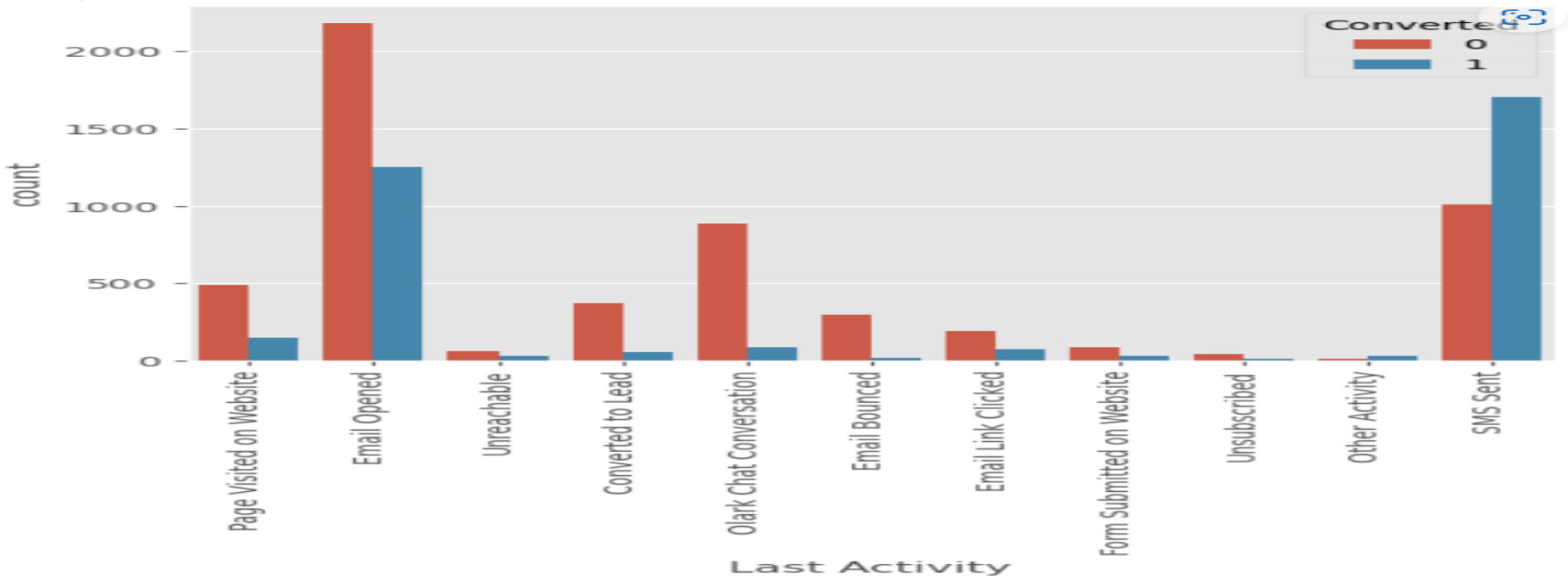




- For TotalVisits, the conversion and non-conversion are same (median), therefore nothing important to conclude over here
- For Total Time Spent on Website, User spending more time on the website are more likely to get converted

EDA: Exploratory Data Analysis

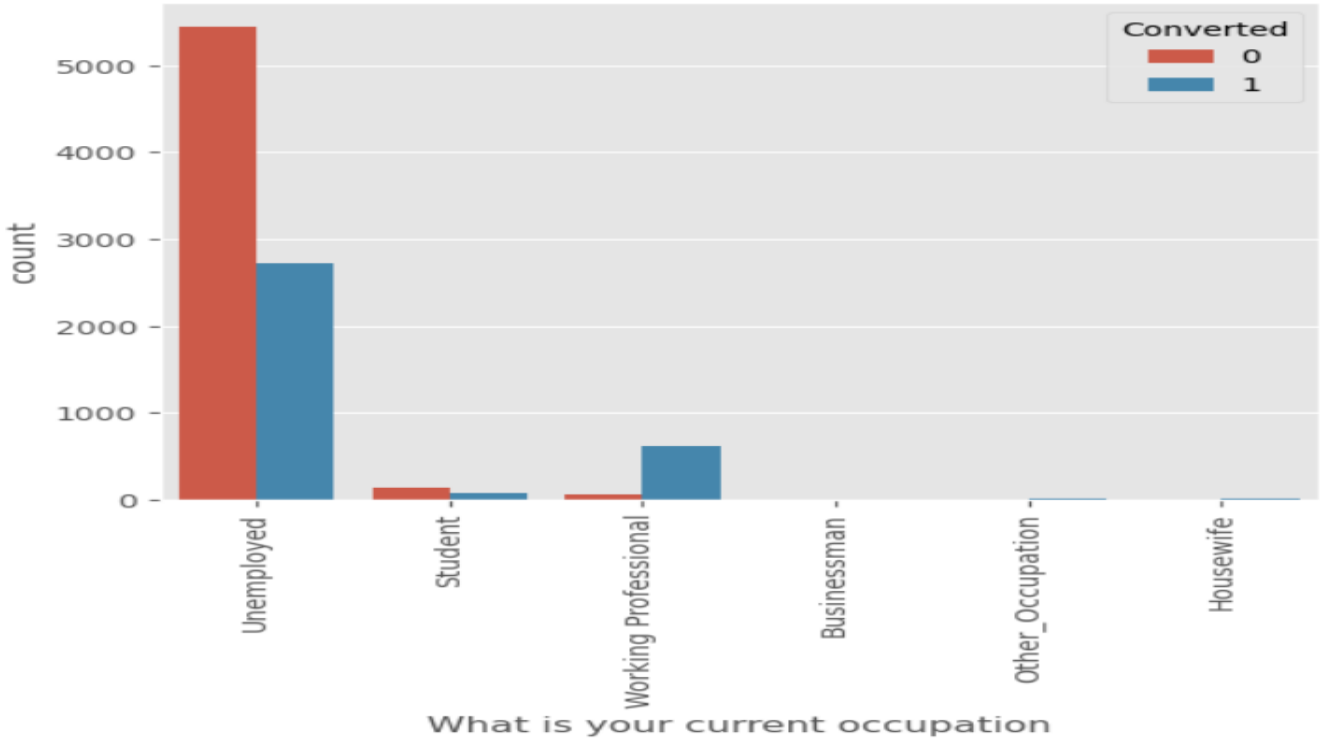
- **Univariate and Bivariate analysis:** Last Activity



- The count for category "Email Opened" is maximum
- the conversionrate of category "SMS Sent" is maximum

EDA: Exploratory Data Analysis

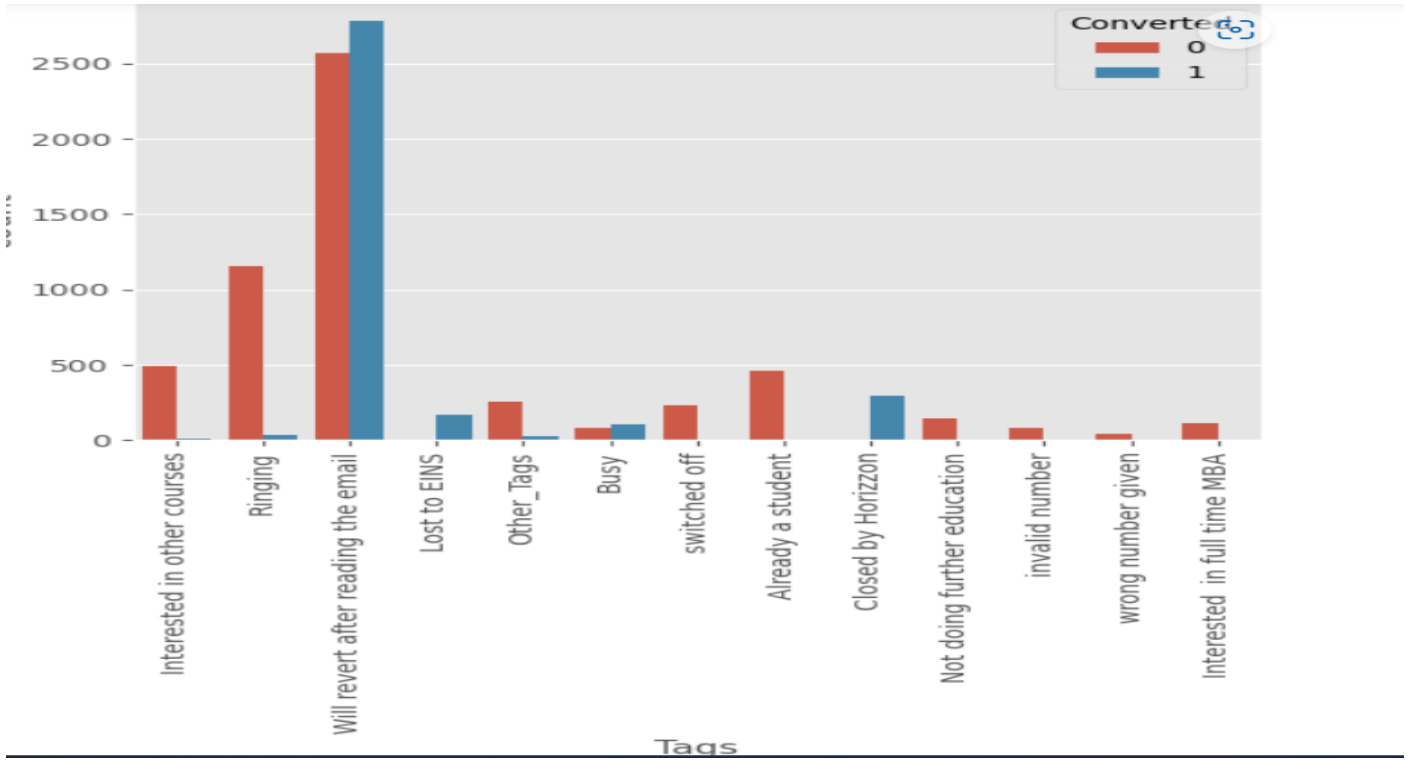
- **Univariate and Bivariate analysis:** What is your current occupation



- Working professionals have higher conversion rate how overall count is low
- Unemployed people have high count but conversion rate is poor

EDA: Exploratory Data Analysis

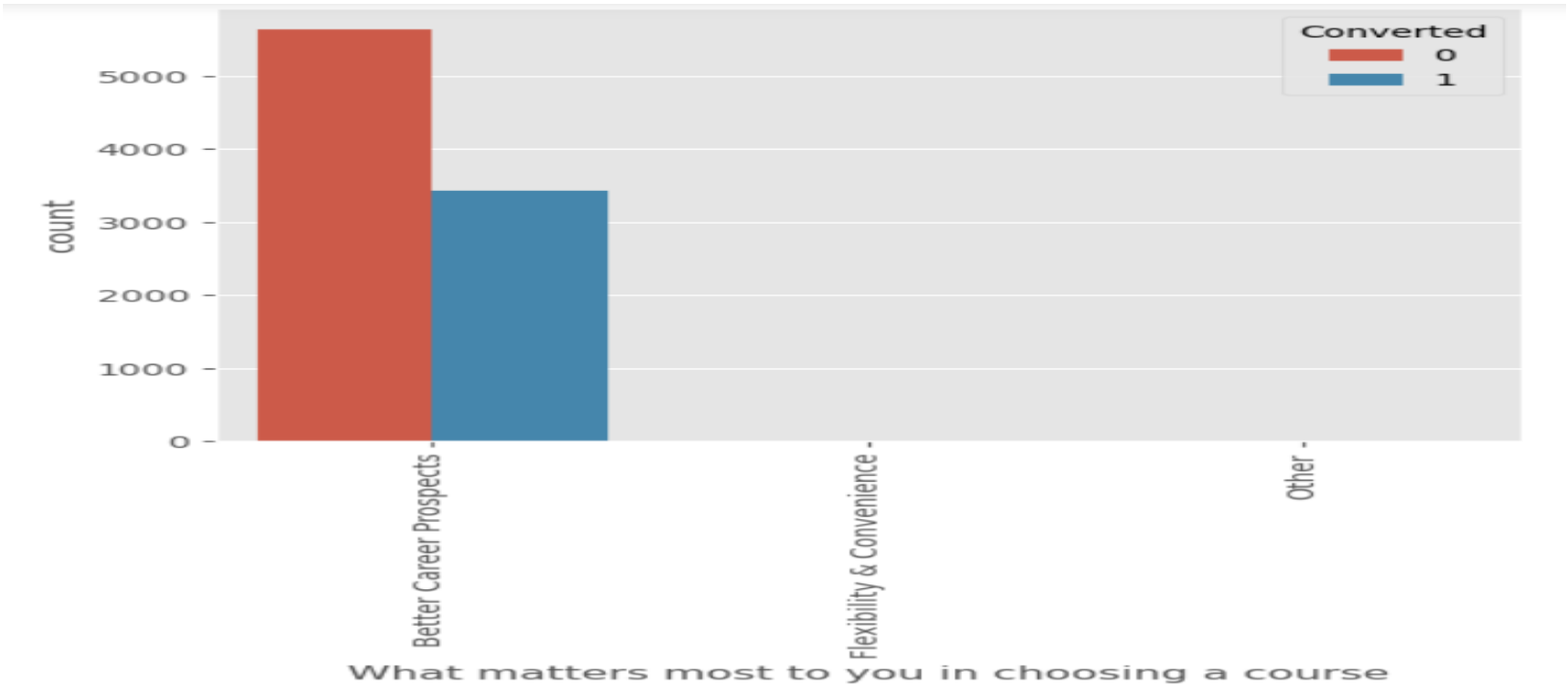
- **Univariate and Bivariate analysis:** Tags



- leads who are tagged as "Will revert after reading the email" have high count and appreciable conversion rate

EDA: Exploratory Data Analysis

- **Univariate and Bivariate analysis:** What matters most to you in choosing a course



- Almost all leads respond with "Better Career Prospects", as a result this column has data which almost constant (or no variation in the response value) and hence no insight can be made.

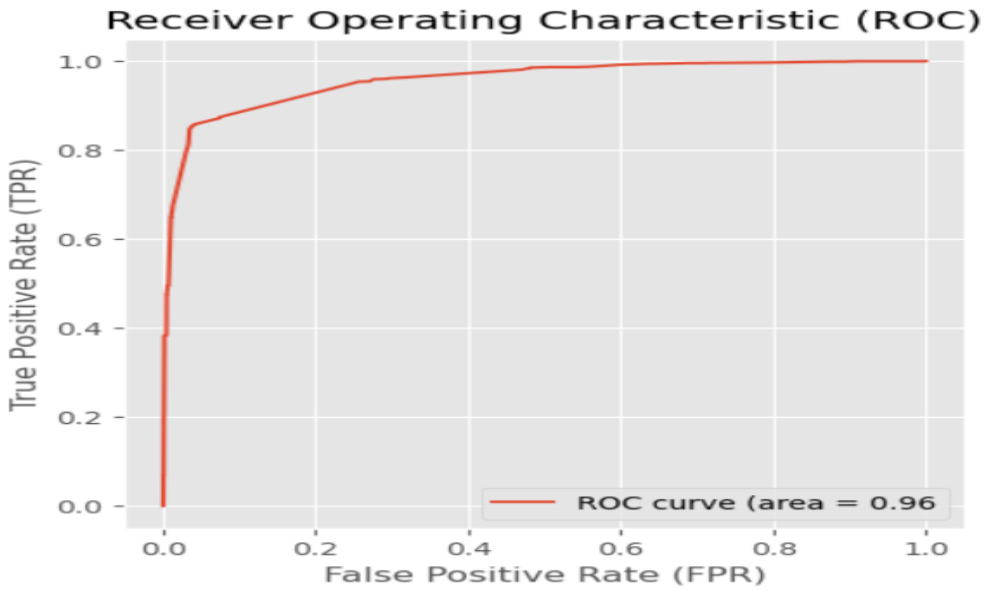
Model Building

- **Feature Selection Using RFA**
  - There are so many feature to deal with
  - This will reduce the model performance and might take high computation time
  - Its important to perform Recurive Feature Elimination (RFE) and to select only the important columns
  - Then manually selecting feature to fine tune the model
  - RFE outcome
    - Top 15 feature selected
  - Manual Feature reduction was employed to build the model and drop the feature with high p-value
  - Arrived at a model (which is out final model) with all p-value less than 0.5 the highest noted p-values is 0.288.

Last Activity_SMS Sent	1.9909	0.102	19.580	0.000	1.792	2.190
What is your current occupation_Working Professional	1.4258	0.296	4.818	0.000	0.846	2.006
Tags_Already a student	-0.8385	0.789	-1.062	0.288	-2.386	0.709
Tags_Busy	3.5651	0.327	10.916	0.000	2.925	4.205
Tags_Closed by Horizzon	8.5613	0.774	11.065	0.000	7.045	10.078
Tags_Lost to EINS	9.4727	0.770	12.310	0.000	7.964	10.981

Model Evaluation

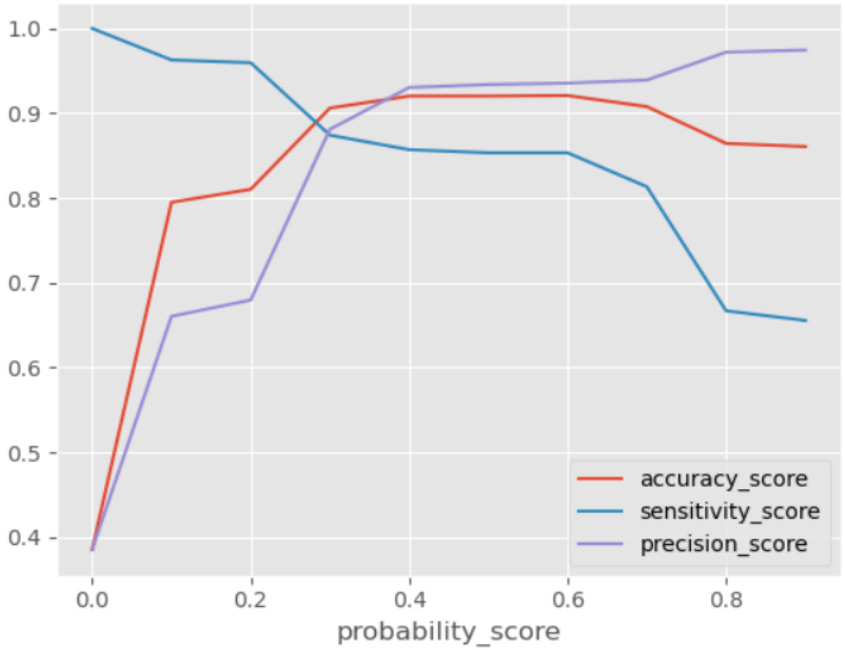
ROC Curve Analysis

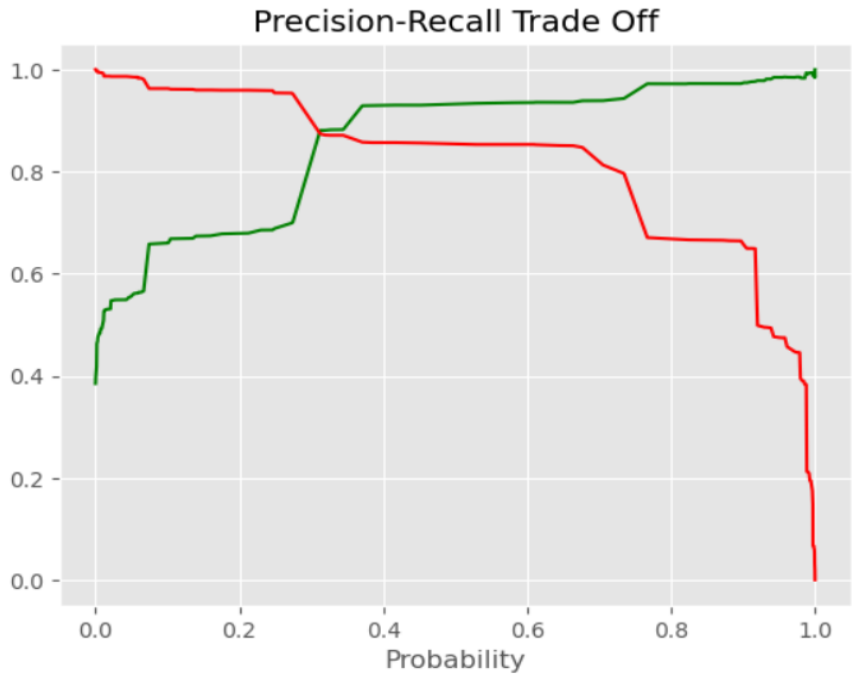


Now if you notice we have area of around 0.96, which tells we have really good model.

Model Evaluation

Optimal Value of Cutoff





- In the plot that shows sensitivity and specificity accuracy, the best value appears to be 0.28. In the precision-recall curve, the optimal value seems to be 0.3.

We are going to take the cutoff of 0.27 as our optimal cutoff

Model Evaluation

Train Dataset Confusion Matrix and Metrics

```
Confusion Matrix:
 [[2902 1003]
 [ 113 2333]]
Accuracy:  0.8242796410014172
Sensitivity:  0.9538021259198691
Specificity:  0.7431498079385404
Precision:  0.6993405275779376
```

Train Dataset Confusion Matrix and Metrics

```
Confusion Matrix:
 [[1260  474]
 [  67  922]]
Accuracy:  0.8013220712449505
Sensitivity:  0.9322548028311426
Specificity:  0.726643598615917
Precision:  0.660458452722063
```

- there is drop in accuracy which not drastic (around 0.02)
- drop is sensitivity is around 0.02 approximately

Sensitivity of 93% means that out of all the converted lead, 93% of them are correctly predicted. Whereas accuracy of 80% means out all the prediction (either converted or not) 80% of them are correct.
We can say that our model doing fine.

Recommendation Based on Final Model

- Focus on the lead having high potential based on the lead score.
- Leads who had last activity as “SMS sent” can be focused as they have higher conversion rate
- In order to enhance the overall rate at which leads are converted, efforts should be directed toward improving the conversion rates of 'API' and 'Landing Page Submission' Lead Origins, as well as increasing the quantity of leads originating from the 'Lead Add Form' channel.
- Focus is required on leads tagged with "Will revert after reading the email" as they have high count and conversion rate is also good. There are chances that customer forgets to read the email, therefore can be approached or give a reminder to increase the conversion rate.
- Website could be made more attractive, with fun-quality content and interactive component might to increase the time spent by users on website.
- Since, lead count of Google, Direct Traffic, Olark Chat, and Organic search is high. Hence more focus is required in regards of conversion rate. Also Reference and Welingak Website have pretty good conversion, more attention is required to increase its count.

THANK YOU