LEAD SCORING CASE STUDY

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.

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% which is very less and to make this process more efficient, the company wishes to identify the most potential leads.

BUSINESS OBJECTIVE

X education wants to find out the most promising leads, i.e. the leads having higher probability of conversion so that the company can save time and money on leads having less probability of conversion . The company wants to create a Logistic Regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. The company CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

DATA PROVIDED

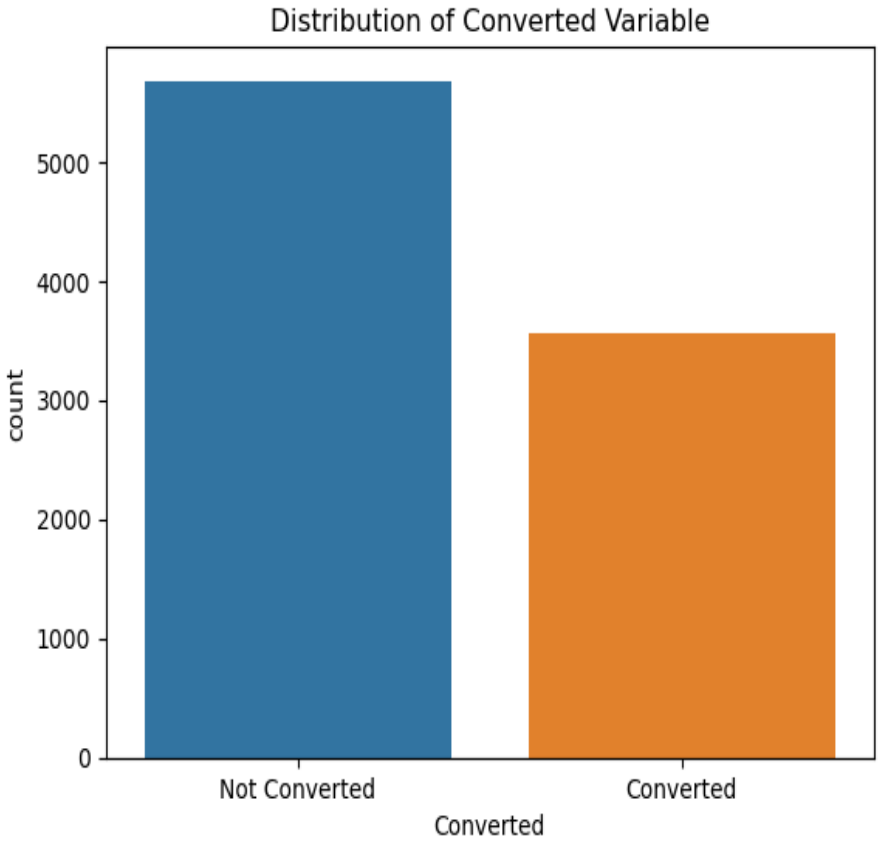
Leads dataset from the past includes data of around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein.

APPROACH

- Import the data
- Study dataset; understand each of the variables and problem statement.
- Clean the data by finding out the missing values, imputing the missing values, or by removing the columns or rows having missing values. Also handle outliers if there are any, Standardising the values of columns ,finding out the Data imbalance percentage
- After cleaning the data and standardising the values of the data identify the categorical and numerical columns to perform analysis.
- Based on the nature of column perform univariate, bivariate or multivariate analysis on variables which seems to be of relevance. Infer the details obtained through the analysis and note down the patterns followed or correlations between variable.
- Data pre-processing like values mapping for categorical data and inserting dummy variables for categorical variables.
- Splitting the data to Train dataset and Test data set.
- Rescaling the numerical variables.
- Create the Logistic regression model using the Train dataset and note down the lead score for each lead.
- Evaluate the model created using evaluation matrix like accuracy, precision, sensitivity, specificity, recall etc.
- Apply the final model on Test datasets and check the same evaluation matrix

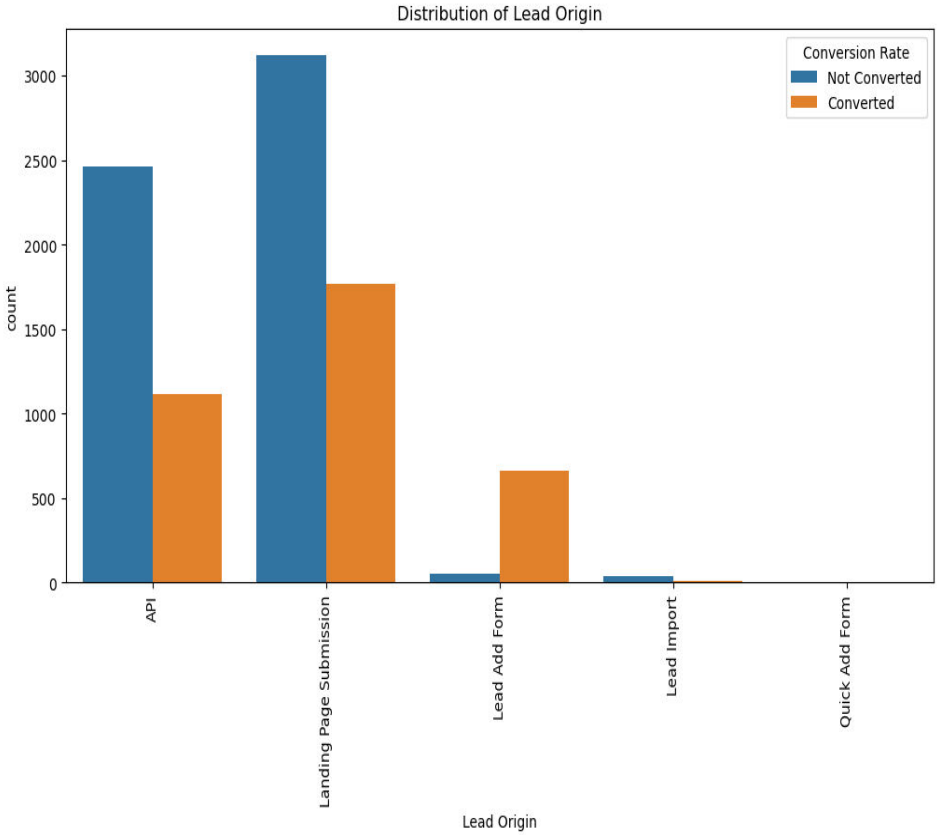
EXPLORATORY DATA ANALYSIS

1. Data imbalance



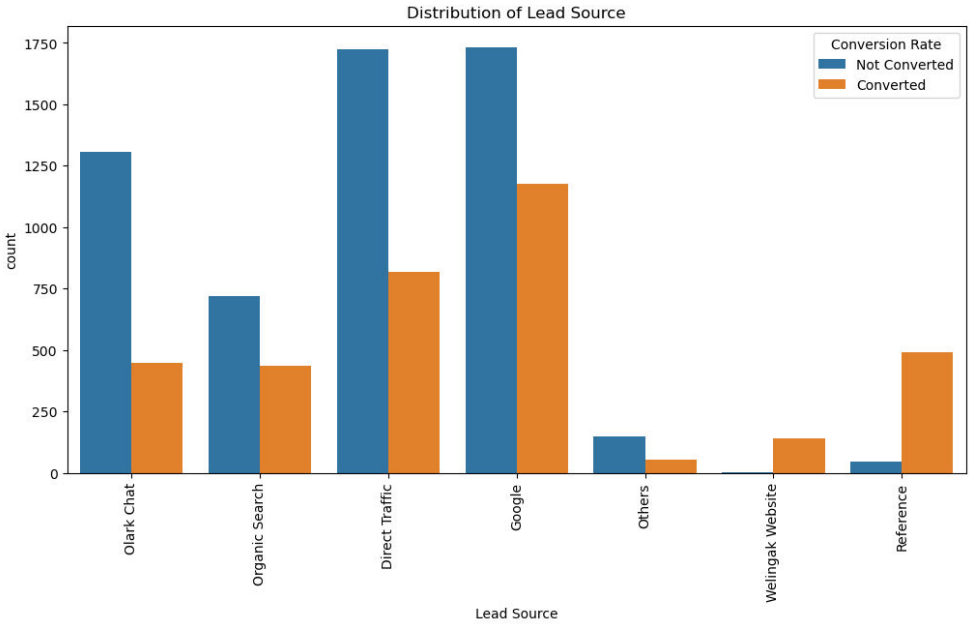
There is a clear imbalance between 'Converted' and 'Not Converted' in the target variable. Around 38% conversion rate is there in the data

2. Lead Origin



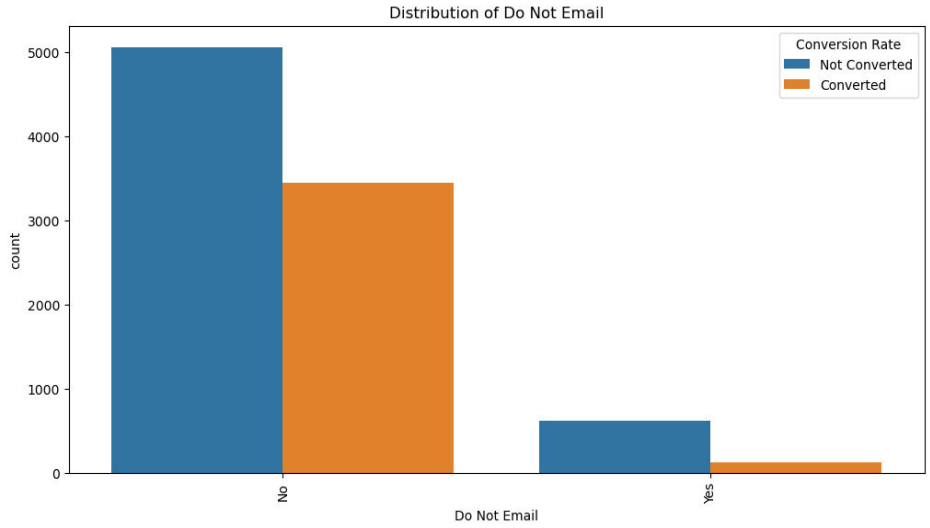
Landing Page Submission has maximum number of lead origin followed by API.Landing Page Submission has high conversion rate followed by API, Lead Add form

3. Lead Source



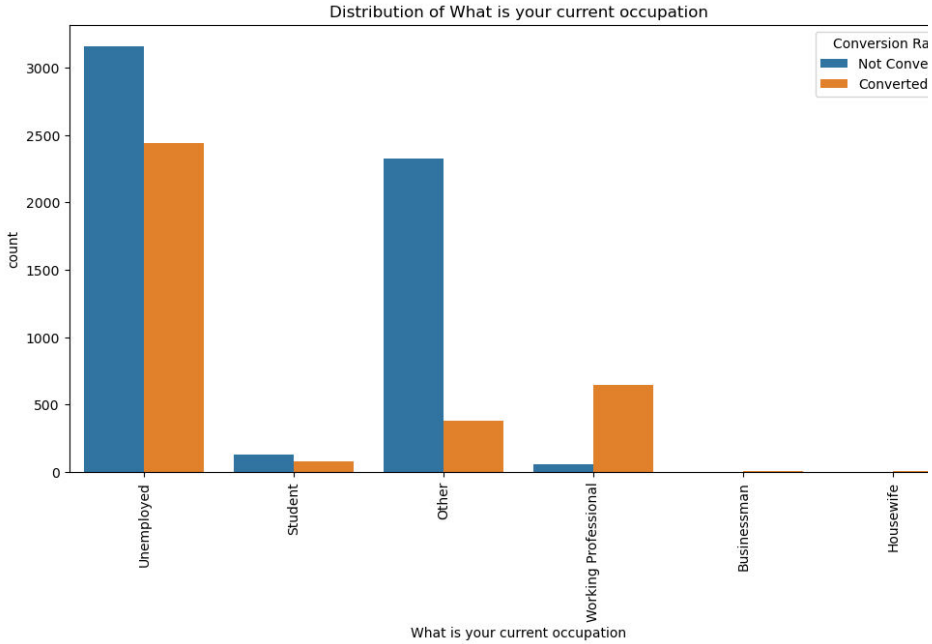
Google and Direct Traffic generates maximum leads. Google has the high conversion rate followed by Direct Traffic, Reference, Olark Chat.

4. Do not email



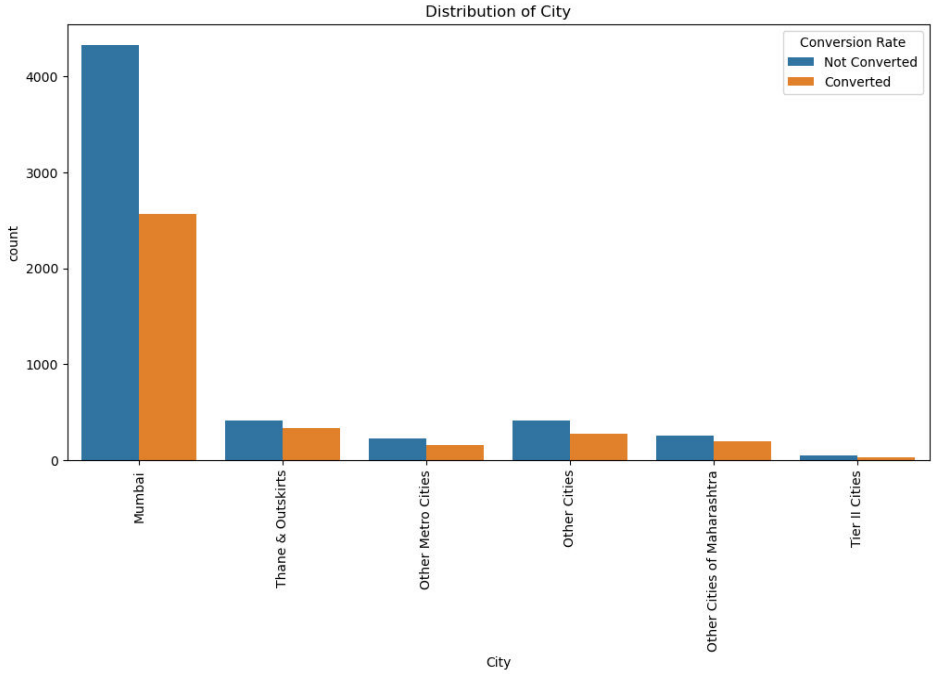
The people who prefer communication via e-mail tends to convert than those who doesn't prefer.

5. Current Occupation



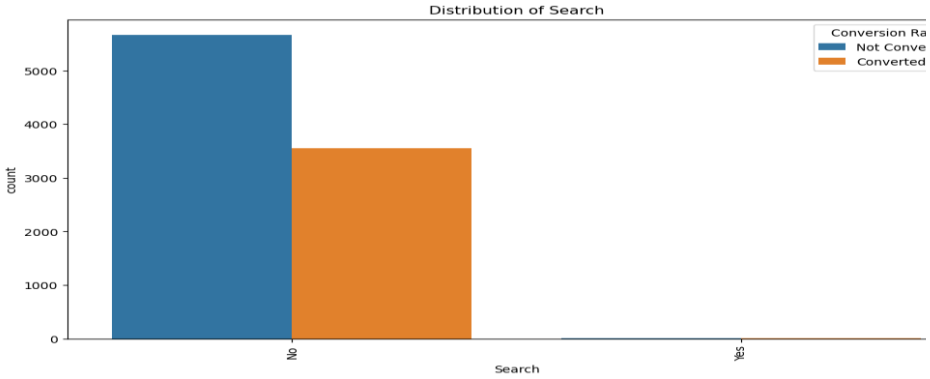
'Unemployed' people prefer to do the courses more, followed by 'Working Professional' and 'Other' category.

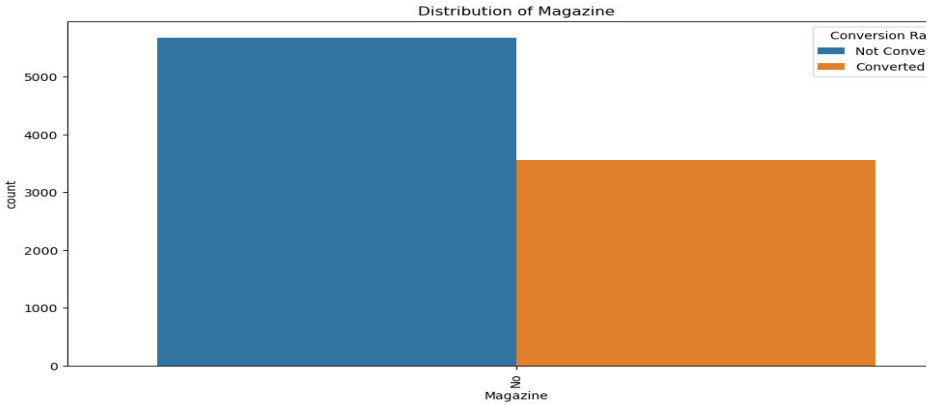
6. City of the applicant

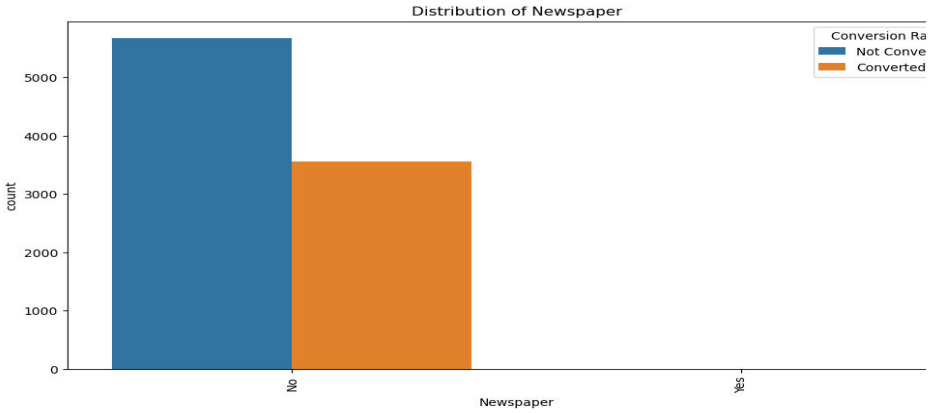


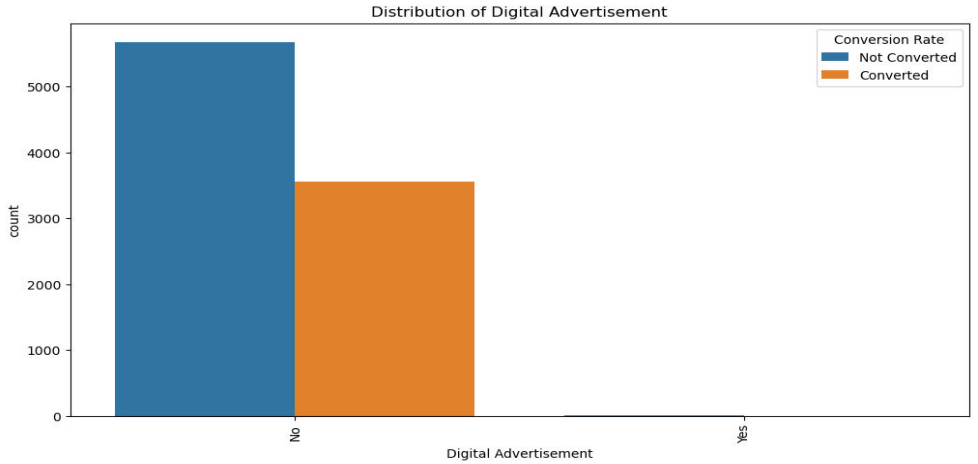
Applicants from Mumbai tends to make the lead conversion higher compared to other cities .

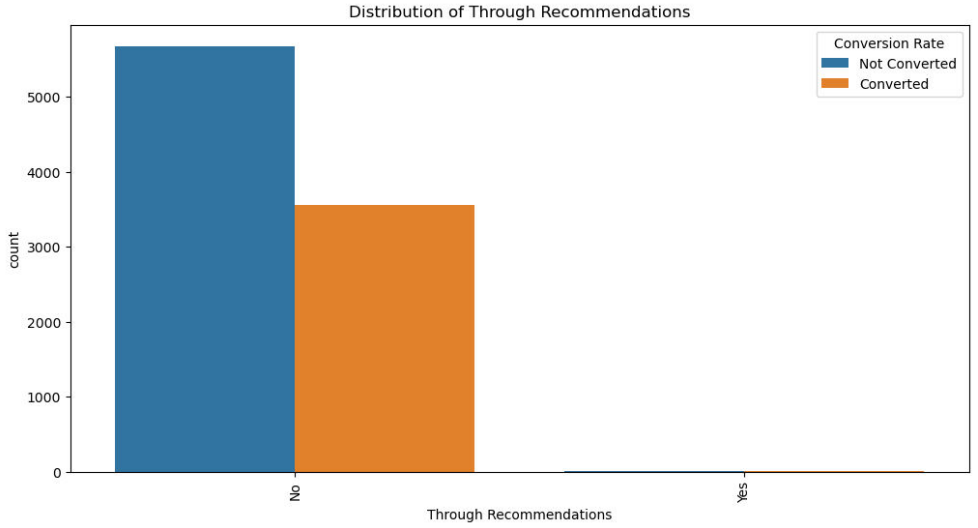
7. No impact





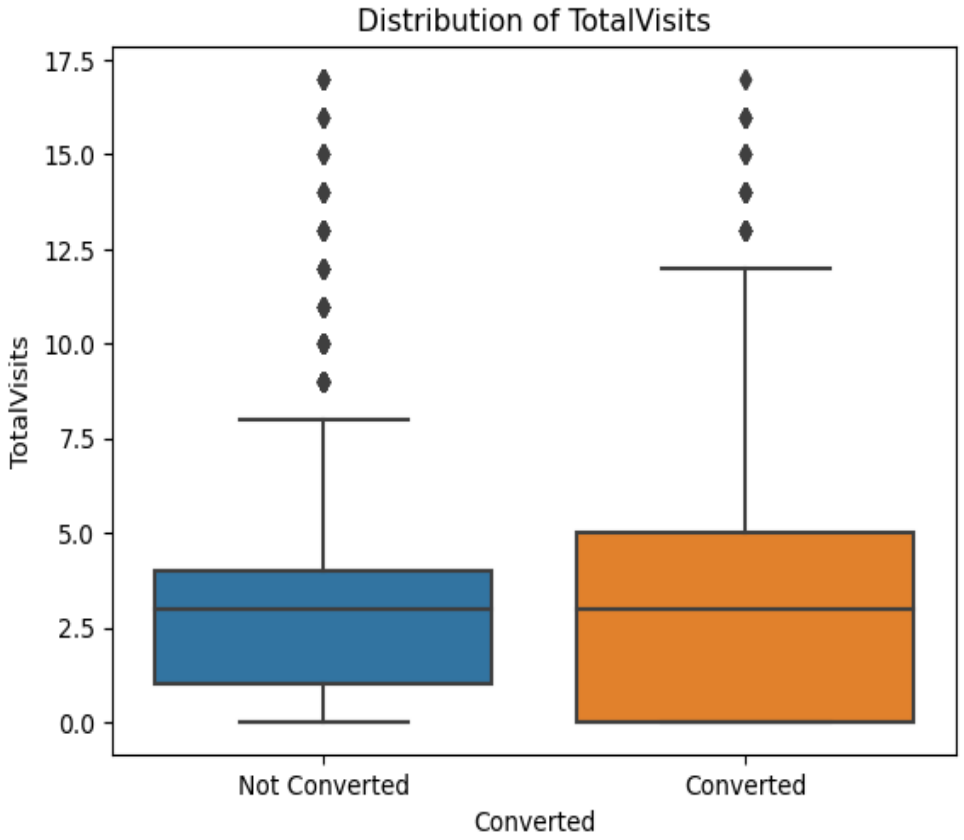






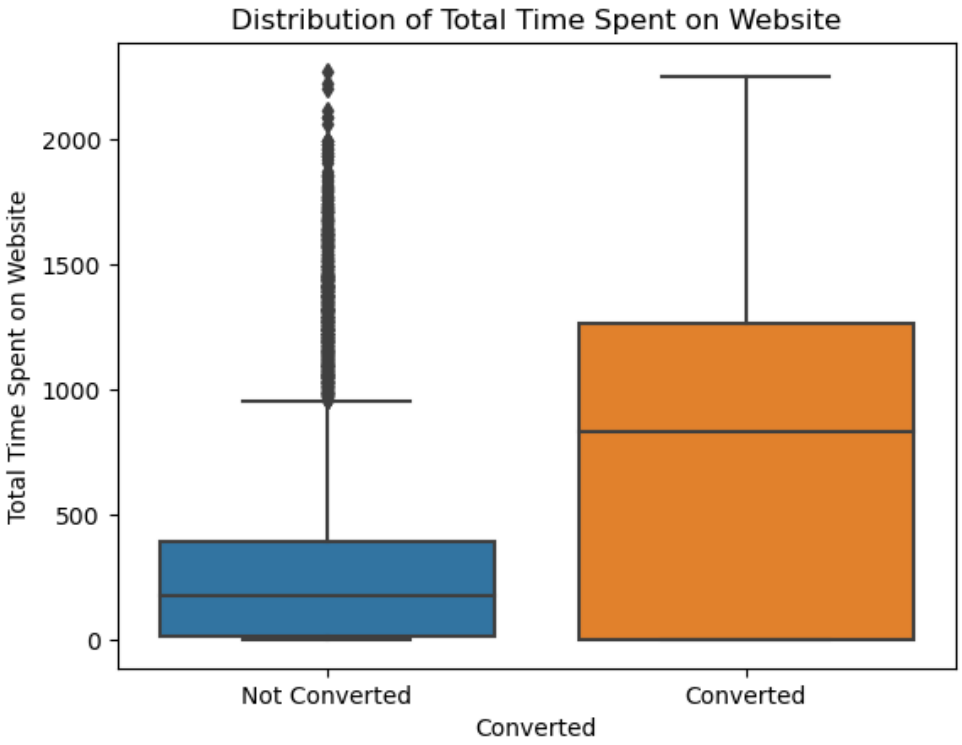
No impact on conversion rate through the variables like ‘search’ , ‘Magazine’ , ‘Newspaper’ , ‘Digital Advertisement’ , ‘Through Recommendation’ .

8. Total Visit



Applicants those who make the most visit to the website seems to get converted than others.

9. Time spent on website



Applicants those who spent more time on website have higher chances of getting converted.

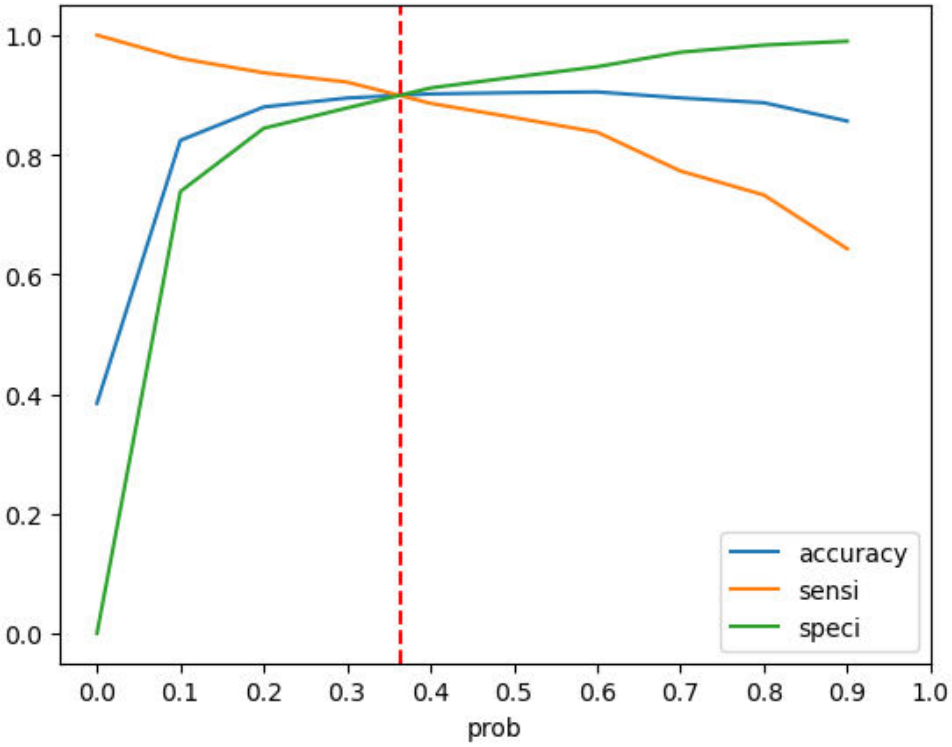
VARIABLES INCLUDED IN THE FINAL MODEL

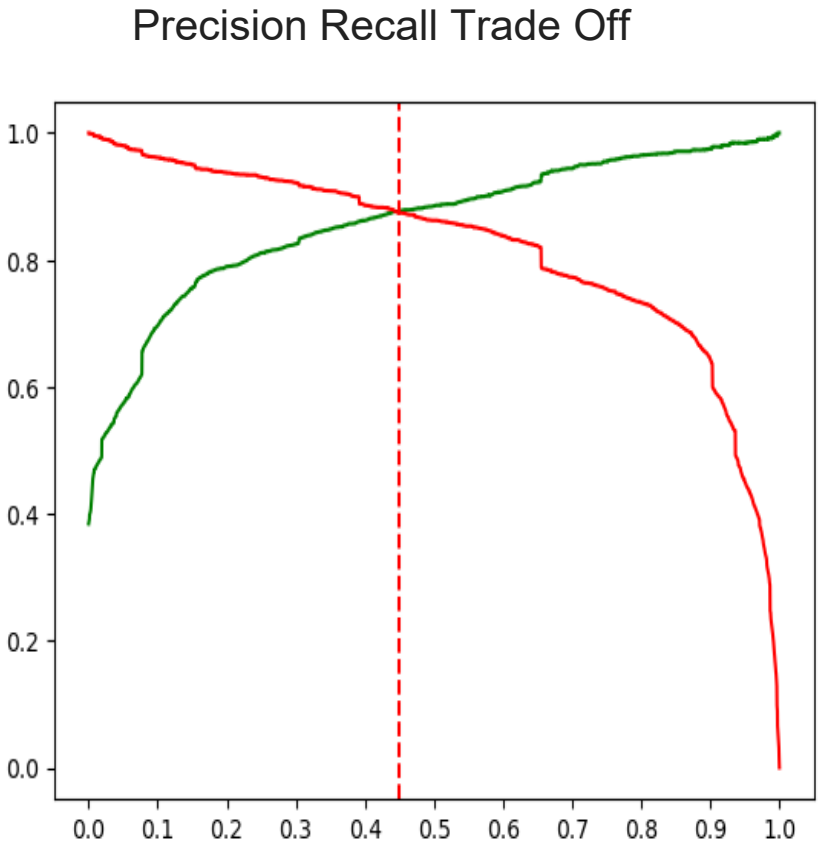
1.Do Not Email
2.Total Time Spent on Website
3.Page Views Per Visit
4.Lead Origin_Lead Add Form
5.Last Activity_Email Bounced
6.Last Activity_Olark Chat Conversation
7.What is your current occupation_Other
8.What is your current occupation_Working Professional
9.Tags_Closed by Horizzon
10.Tags_Others
11.Tags_Will revert after reading the email
12.Last Notable Activity_Others
13.Last Notable Activity_SMS Sent

The final model looks like :

Probability(Converted)= -1.51*Do Not Email + 4.08*Total Time Spent on Website - 1.43*Page Views Per Visit + 1.58*Lead Origin_Lead Add Form -1.45*Last Activity_Email Bounced - 1.46*Last Activity_Olark Chat Conversation - 3.12*What is your current occupation_Other + 1.28*What is your current occupation_Working Professional + 9.53*Tags_Closed by Horizzon + 3.61*Tags_Others + 5.95*Tags_Will revert after reading the email + 1.38*Last Notable Activity_Others + 2.04*Last Notable Activity_SMS Sent

OPTIMUM PROBABILITY CUT OFF





By assigning various value as cut off ranging from 0 to 1 and using accuracy, sensitivity and specificity against the probability the optimum cut off value is obtained as 0.36.Also from Precision Recall Trade off it is obtained 0.46 as the optimum cut off for the probability.

CONFUSION MATRIX AND OTHER EVALUATION MATRIX ON TRAIN DATA

CONFUSION MATRIX: **[[3529, 387]**

[228, 2219]]

EVALUATION MATRIX:

- **Accuracy : 90%**
- **Sensitivity : 90%**
- **Specificity : 90%**
- **Precision : 88%**
- **Recall : 87%**

MODEL EVALUATION ON TEST DATA

CONFUSION MATRIX: **[[1585, 94]**

[112, 936]]

EVALUATION MATRIX:

- **Accuracy : 92%**
- **Sensitivity : 91%**
- **Specificity : 90%**
- **Precision : 88%**
- **Recall : 86%**

CONCLUSION

- From EDA it is clear that people who spend time on website and who make most visits on websites have a higher chance of conversion.
- Unemployed people and applicants from Mumbai have higher chances of conversion .
- The top three variables in the model are:
    1. Total Time Spent on Website
    2. What is your current occupation_Other
    3. Tags_Closed by Horizzon
- The created model has an accuracy ,sensitivity ,specificity ,precision ,recall as 90%,90%,90%,88%,87% on train data .
- The created model has an accuracy ,sensitivity ,specificity ,precision ,recall of test data as 92%,91%,90%,88%,86% .
- Hence the model satisfies the required condition of 80% conversion rate so we can conclude our model as good one.