

Lead Scoring Case Study

Submitted by,

- Anushree V
- Jeevan Kumar
- Suket J

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- The company currently markets its courses on websites and search engines. The lead conversion rate through this process is very poor.
- The company aims to achieve a target lead conversion rate of 80%



BUSINESS GOAL



Build a logistic regression model for company to target most potential leads.



A lead score is to be given to each lead – higher the lead score, more promising the lead is



The model should be able to adapt to the changes in future requirements as well

APPROACH

Reading and
Cleaning the Data

Reading & Understanding the Data [1](#)

Cleaning the Dataset

EDA and Data
Visualization

Univariate and Bivariate Analysis

Data Preparation for
Model Building

Model Pre-processing - Train Test Split [1](#)

Model Pre-processing - Feature Scaling

Model Building &
Feature Selection

Running Initial GLM Regression Model with all the features using Statsmodel [1](#)

Feature Selection using Recursive Feature Elimination (RFE) [1](#)

Model Building

Model Evaluation -
Confusion Matrix,
Accuracy, Sensitivity
(Recall), Specificity,
Precision, etc.

Plotting the ROC Curve

Finding the Optimal Cut-off Point [1](#)

Precision and Recall

Making Predictions
on the Test Set

ANALYSING VARIABLES THAT IMPACT CONVERSION RATE

Lead origin – Lead add form

Lead source – Welingak Website and Reference, Google and Direct Traffic Sources

Calling and email preference – Email marketing

Total visits

Total time spent on website

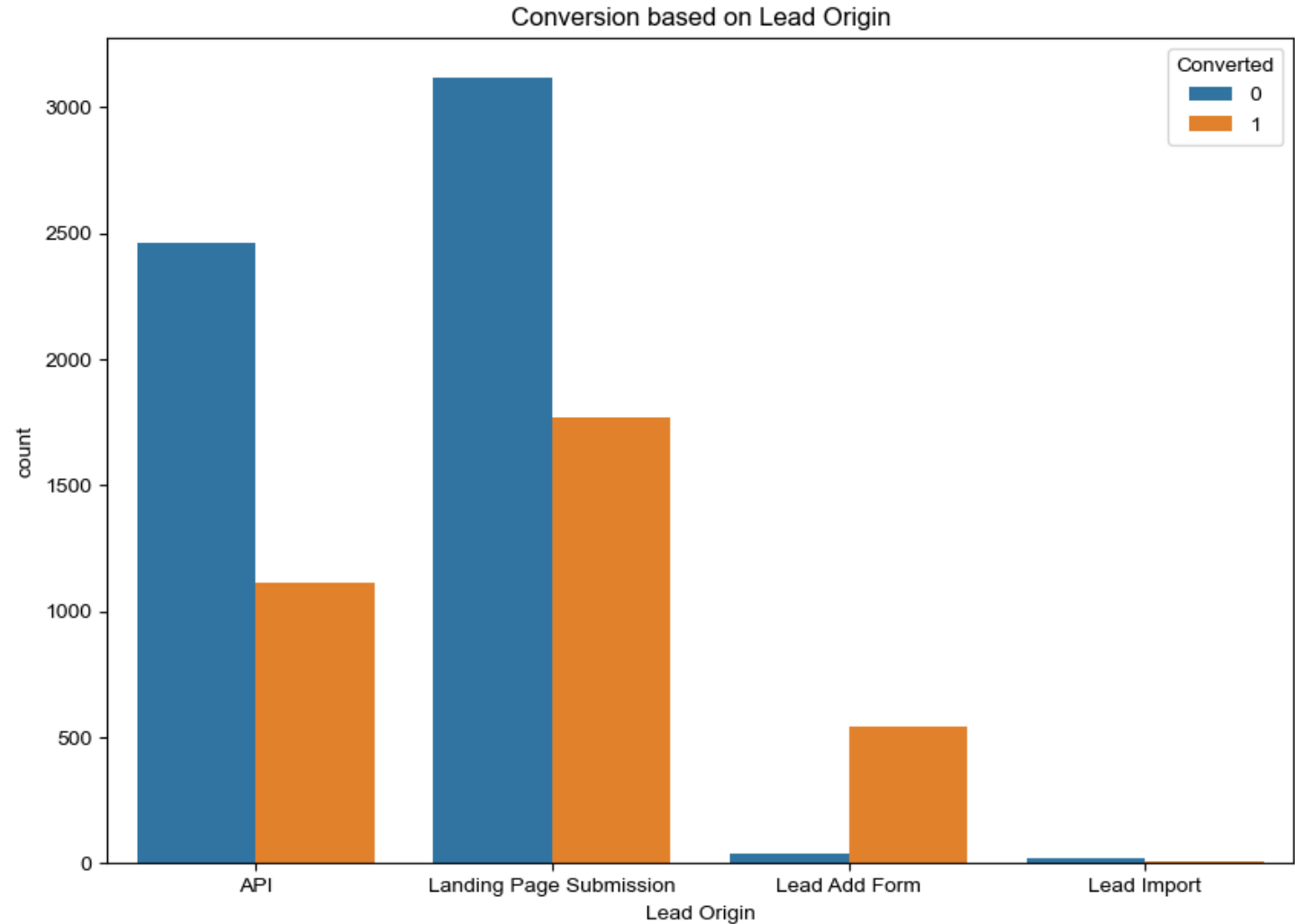
Page views per visit

Occupation

Region

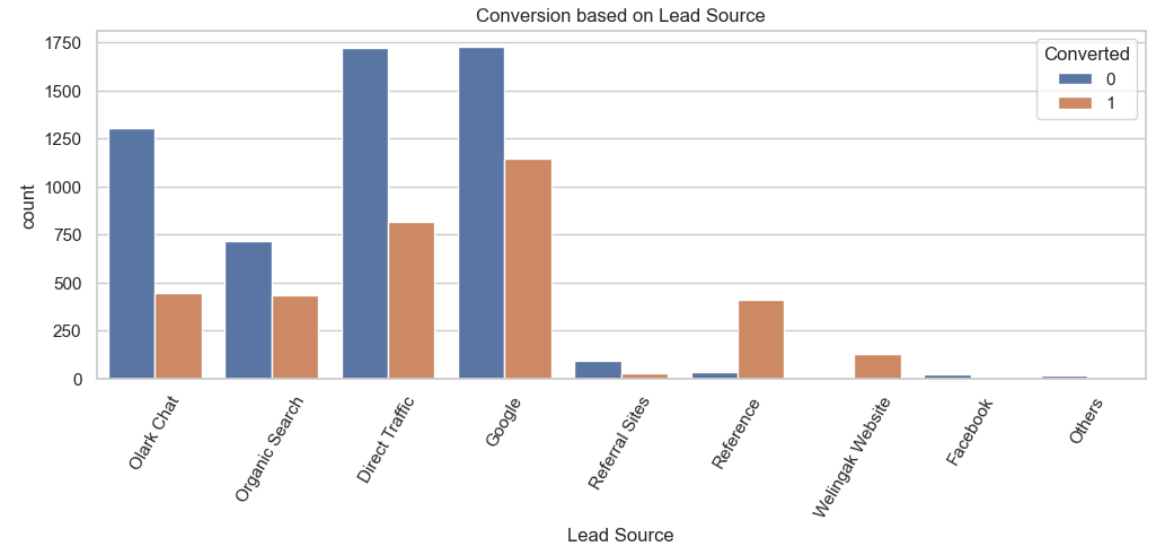
LEAD ORIGIN

- Lead Add Form have a conversion rate of 94. hence the company should consider generating more leads from Lead Add Form
- Landing Page Submission & API have a conversion rate of 36% and 31% respectively but they are also the place where most of the leads come from. Hence the company should focus on improving the conversion rate from these two sources.



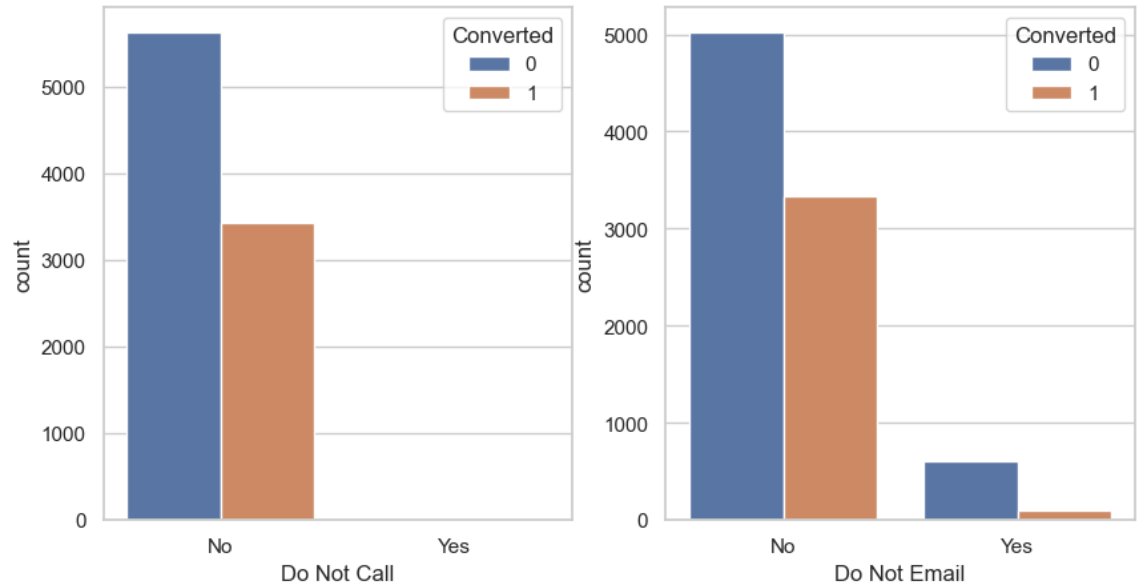
LEAD SOURCE

- Welingak Website and Reference are the Sources which have the best conversion rates however, they generate very few leads. Hence the company should focus on generating more leads from these two sources.
- Google and Direct Traffic Sources generate the most leads out of all sources but their conversion rates are around 40% and 32% respectively. The company should be focusing on improving the lead conversion rate of these two sources.
- Sources such as Olark Chat & Organic Search generate quite a few number of leads but the conversion rate from these two sources is still low i.e., 26% and 38% respectively. The company should focus on improving the conversation rate of these 2 sources as well and see if they can generate more leads from the Organic search source.



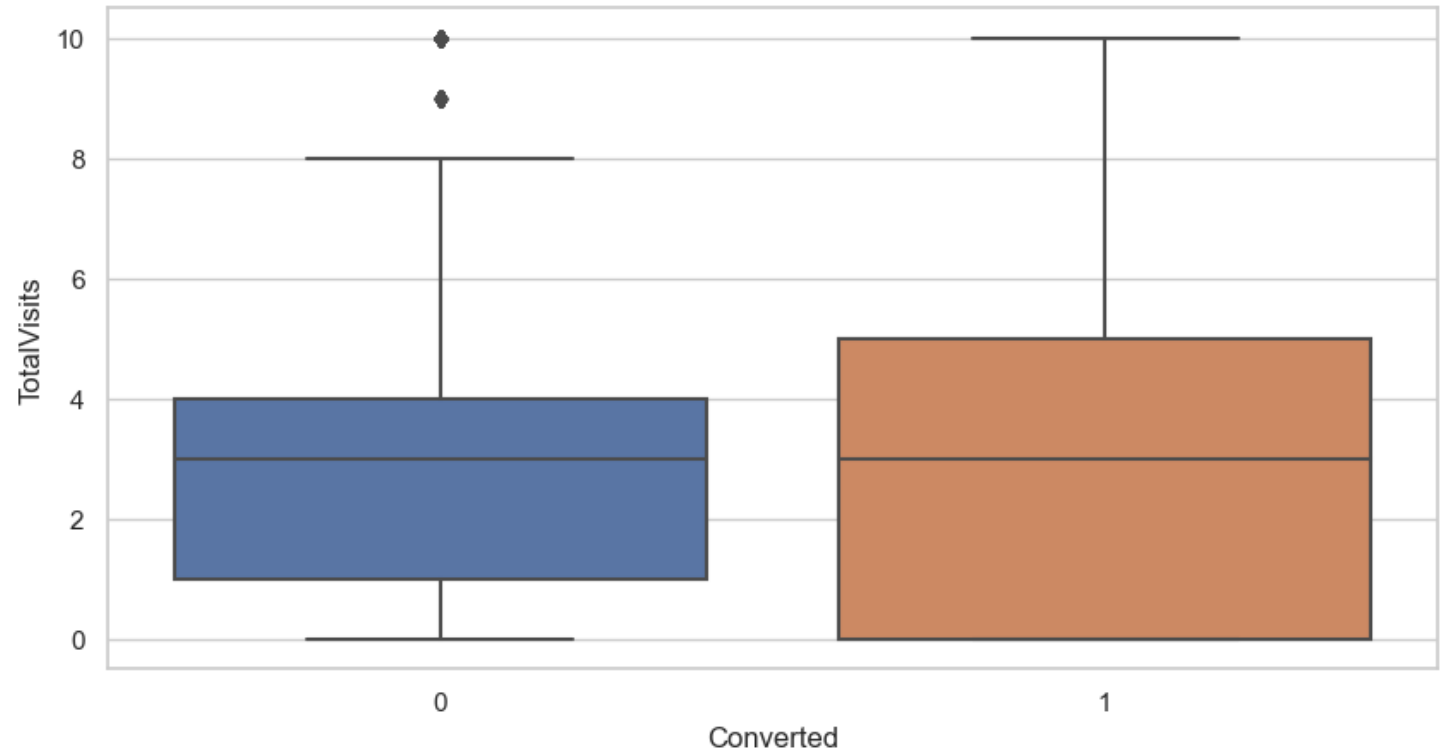
CALLING AND EMAIL PREFERENCE

- Most people do not like to be called or emailed and those who do like to be called are more or less interested in the product and hence the conversation rate can be seen as 100%.
- However, still the conversion rate by email is still greater than conversion rate by call hence it might not be a bad idea to do a segmented email marketing campaign to increase the conversion rate.



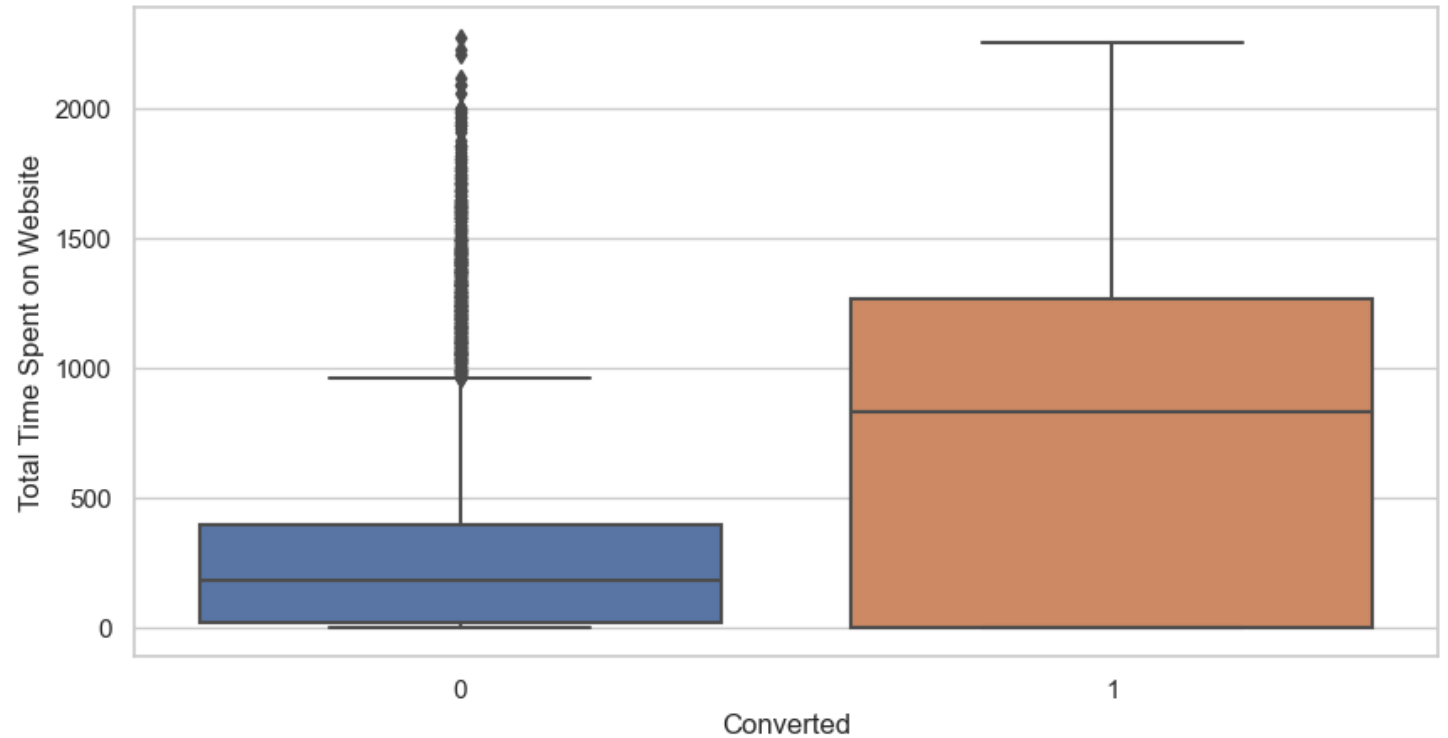
TOTAL VISITS

- The medians of the users who converted and those who not converted after visiting their platform is similar hence there is an equal chance of a user getting converted (applying for the course) and not getting converted (not applying for the course) after visiting the company's platform.



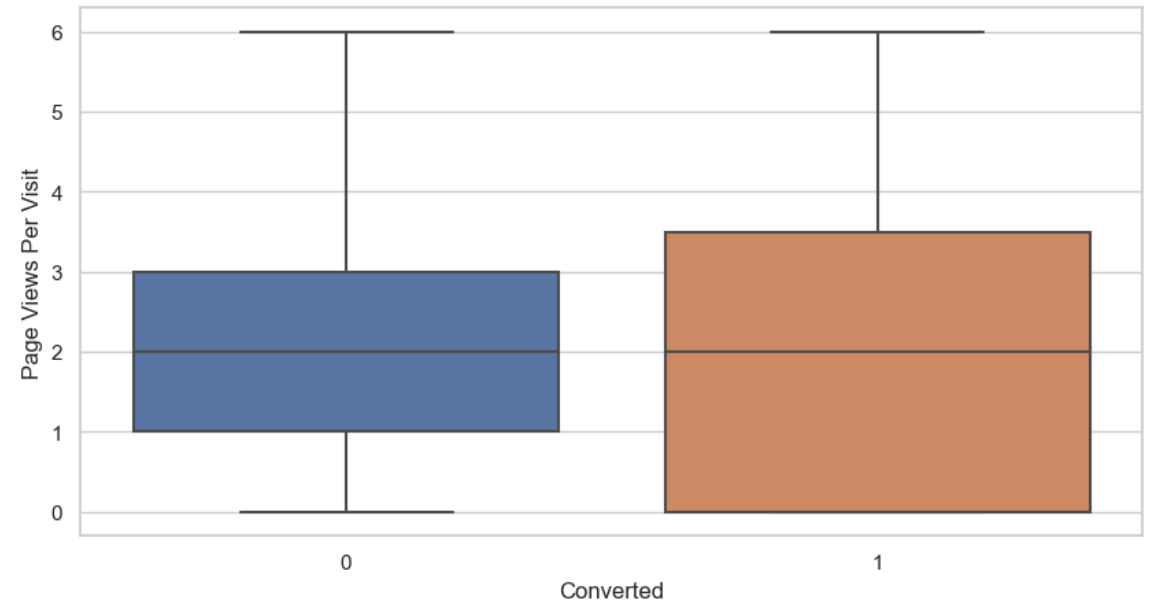
TIME SPENT ON WEBSITE

- Those people spending more time on the website are more likely to convert on average.
- There are some outliers who have spent a lot of time on the website but still haven't converted. It could be inferred from this that they were either idle after visiting their website or even after reading the course details they were not interested or did not like what they saw and did not convert



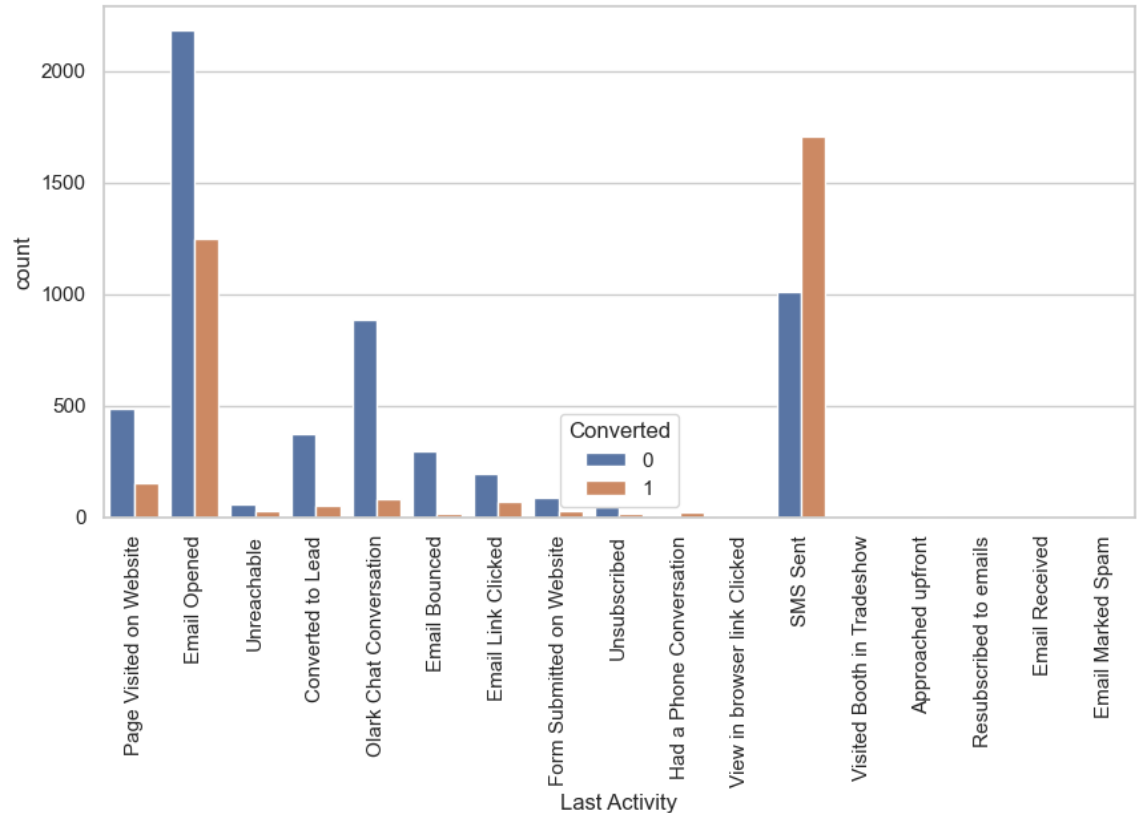
PAGE VIEWS PER VISIT

- People who view 1-3 pages per visit to the platform have a 50-50 chance of converting and we can see that by observing the equal medians for both.
- However, people who have viewed 0-1 pages per visit to the platform have a higher chance of conversion.



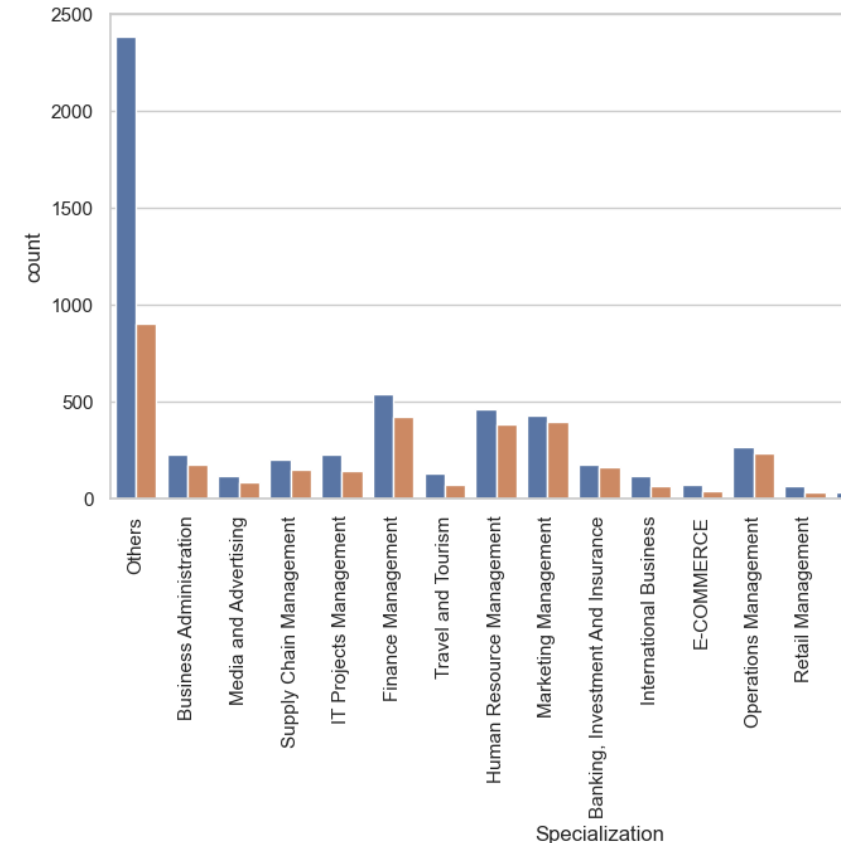
LAST ACTIVITY

- Maximum leads are generated via SMS Sent and Email Opened categories and the conversion rates are 63% and 36% respectively.
- Olark Chat Conversation & Page Visited on Website also generate quite a few leads but the conversion rate for these two categories is still too low.
- People with whom the company has Had a Phone Conversation, the conversion rate for those is 80%.
- If the company wants to improve their overall conversion rate, they can focus on improving the conversion rate for Olark and Page visited categories and have more Phone Conversations with the leads. There is a high chance that people who are okay with receiving a call and are willing to listen will get converted and apply for the course.



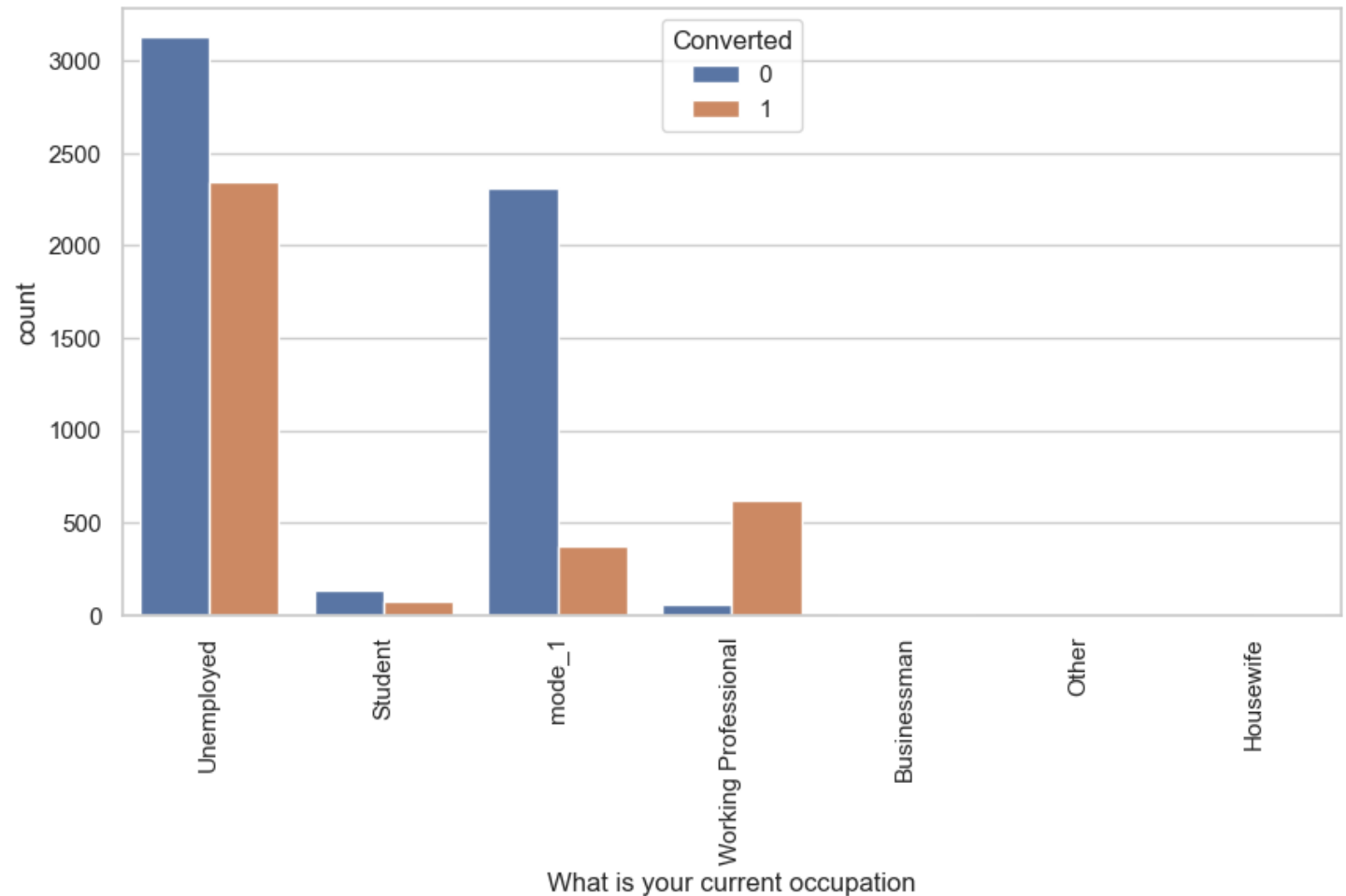
SPECIALIZATION

- Almost all of the specializations that are recognized by the company have a conversion rate between 30% - 50%.
- Others section which dealt with information about the users who either don't have a specialization or didn't find it in the specializations recognized by our company brings in the most leads with a conversion rate of around 27%.



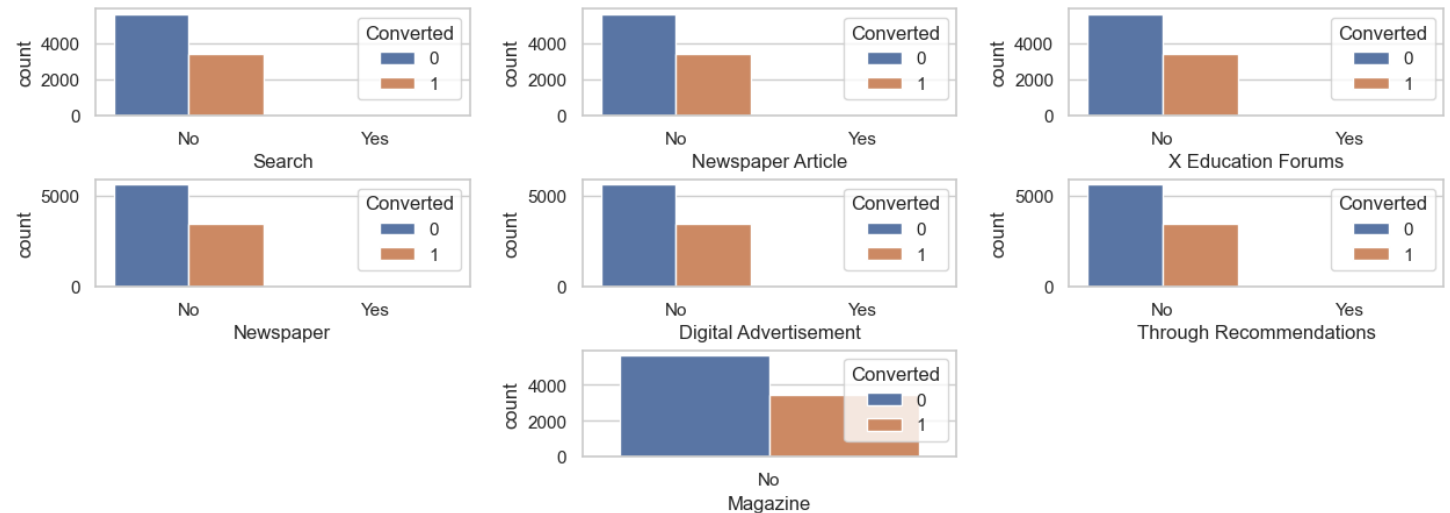
CURRENT OCCUPATION

- Working Professional and Unemployed categories seems to bring in a lot of leads with a conversion rate of 92% and 43% respectively.
- The company should be focusing on generating more leads from Working Professionals and try to increase the conversion rate of Unemployed category users.



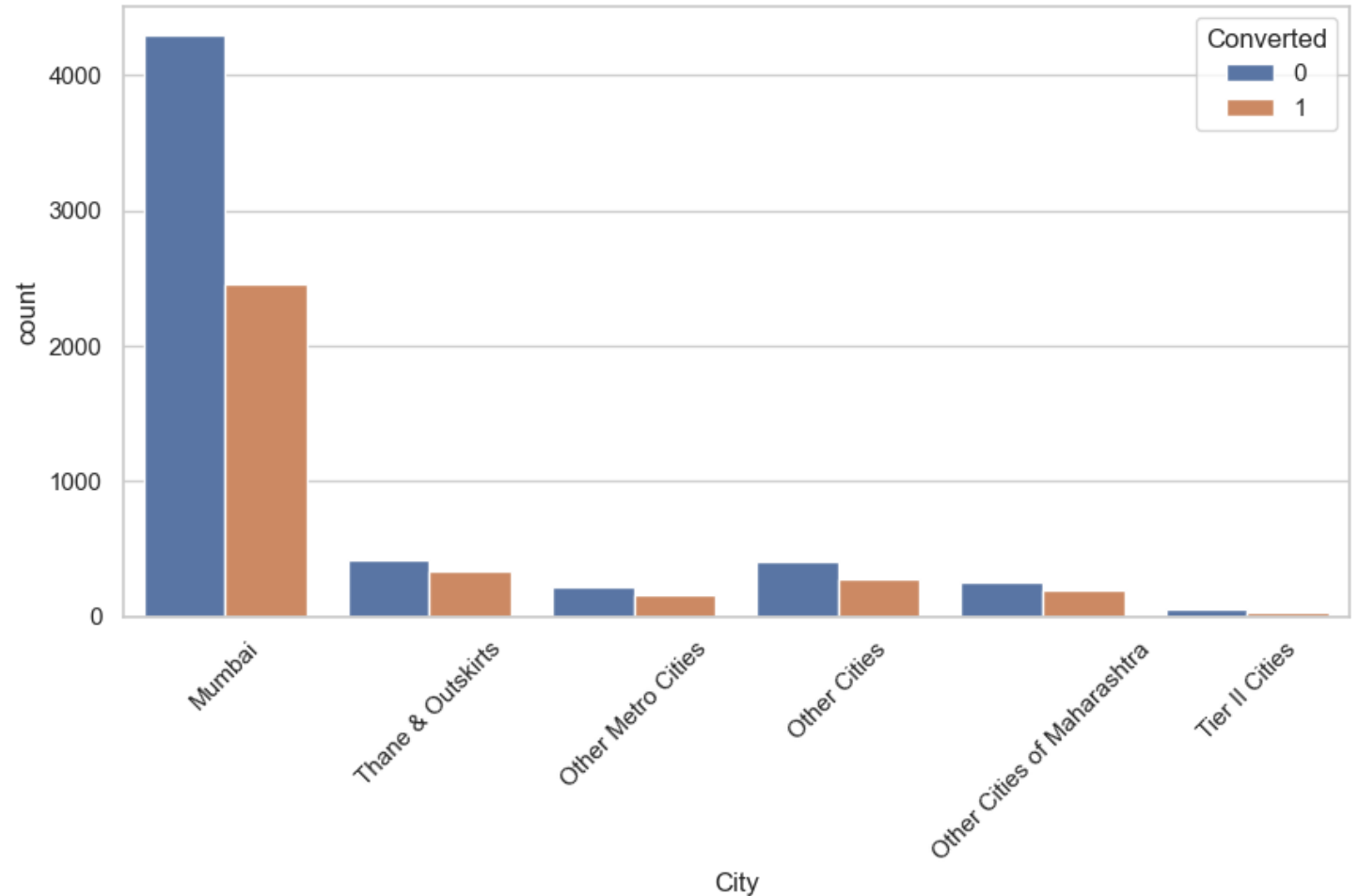
FEATURES

- Features such as `search`, `Newspaper`, `Newspaper Article`, `X Education Forums`, `Digital Advertisement`, `Through Recommendations` have not played an important role in lead generation. Almost 99% of users and 100% of `Magazine` users did not come through these platforms for the course.
- All of the leads have a similar conversion ratio and does not seem to be dependent on these sources.



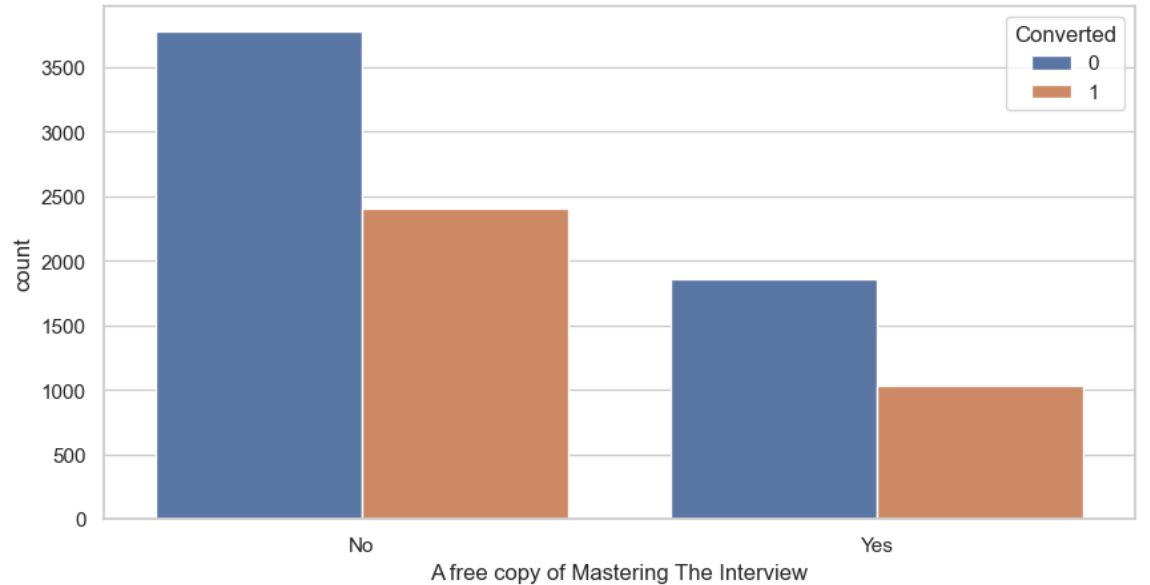
REGIONAL

- The city Mumbai brings in the most leads but does not have a significant conversion rate.
- Cities like Thane and other outskirts cities of Maharashtra also generate some leads and have 40 - 50 % conversion rate as well.
- The company should focus on increasing the conversion rate of Mumbai and also focus on generating more leads from other cities.



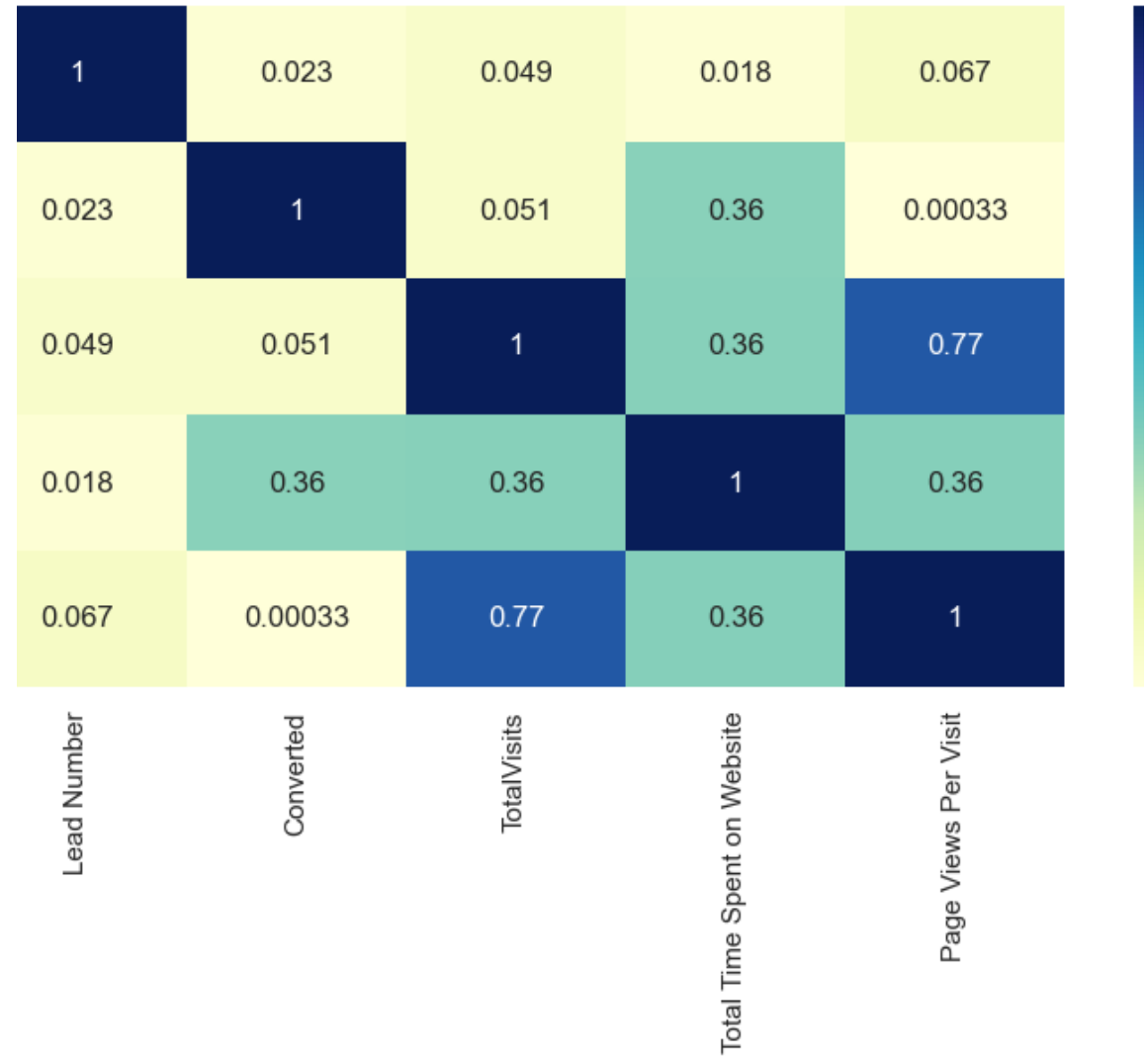
FREEBIE

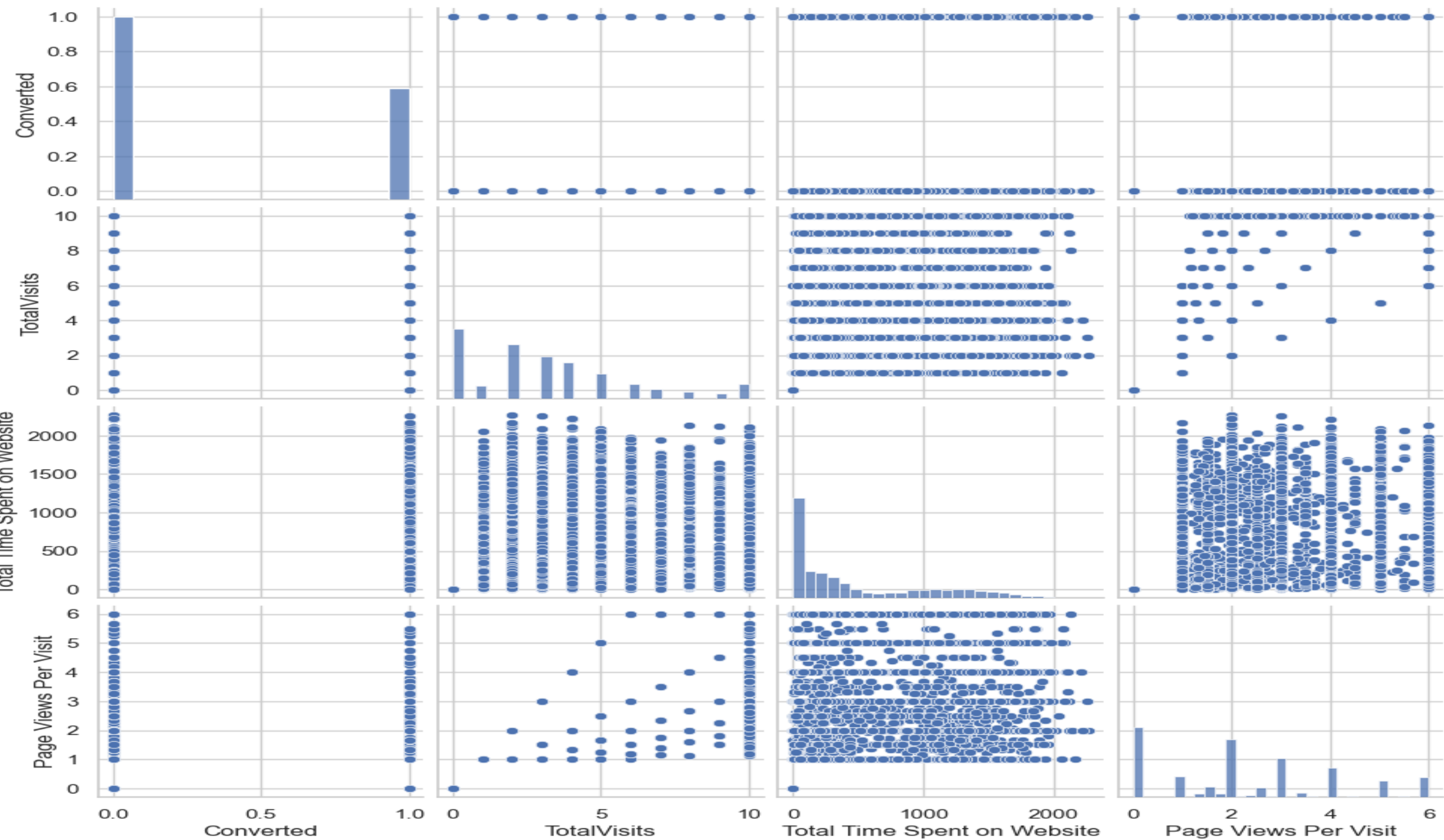
- A lot of customers did not want a free copy of Mastering The Interview however, that did not seem to have a significant impact on the conversion rate of the leads. Which implies this is not a significant feature in deciding the conversion rate.



HEATMAP

- TotalVisits and Page Views Per Visit are highly correlated. This collinearity needs to be treated between multiple variables.



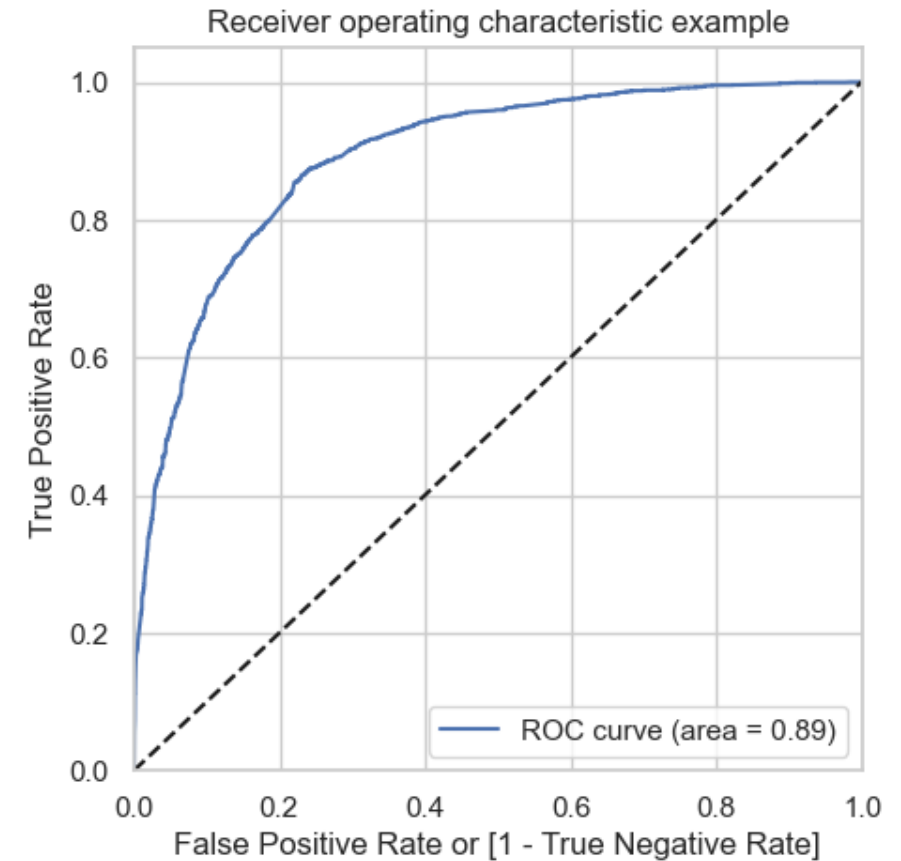


MODEL EVALUATION

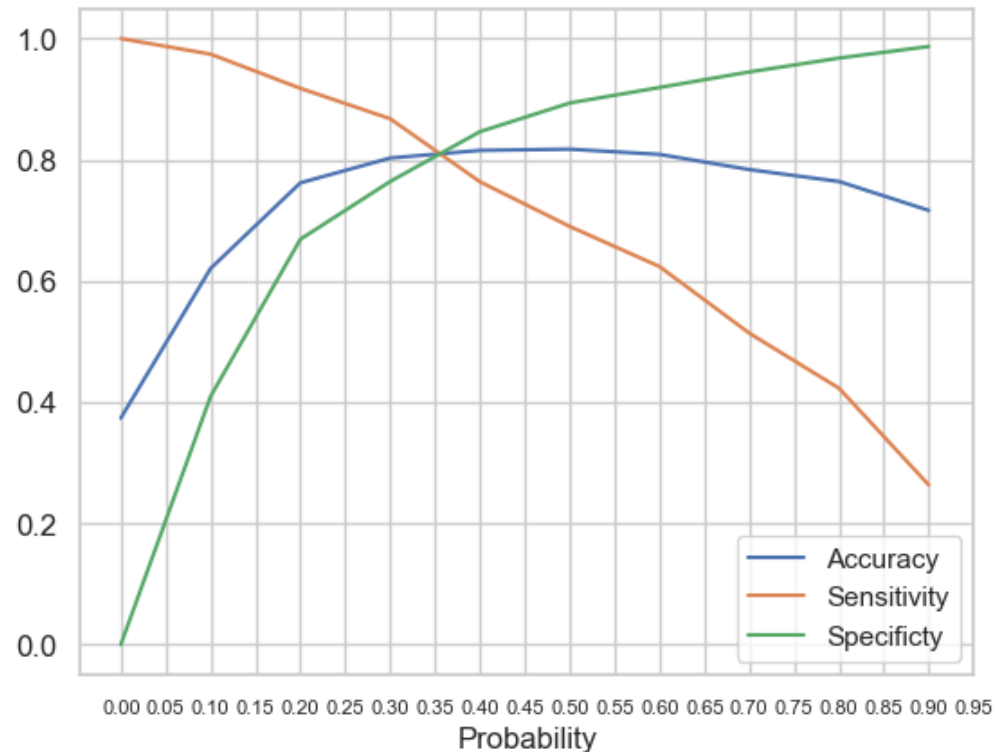
- **Confusion Matrix, Accuracy, Sensitivity (Recall), Specificity, Precision**
- Using different evaluation metrics machine learning model's performance is evaluated.
- To assess the efficacy of a model during initial research phases it is also used for model monitoring.

ROC CURVE

- The area under the curve is 0.89 for this model and this value leans towards 1. The more close it is to 1, the better the model will be. Which means this is a fair predictive model.



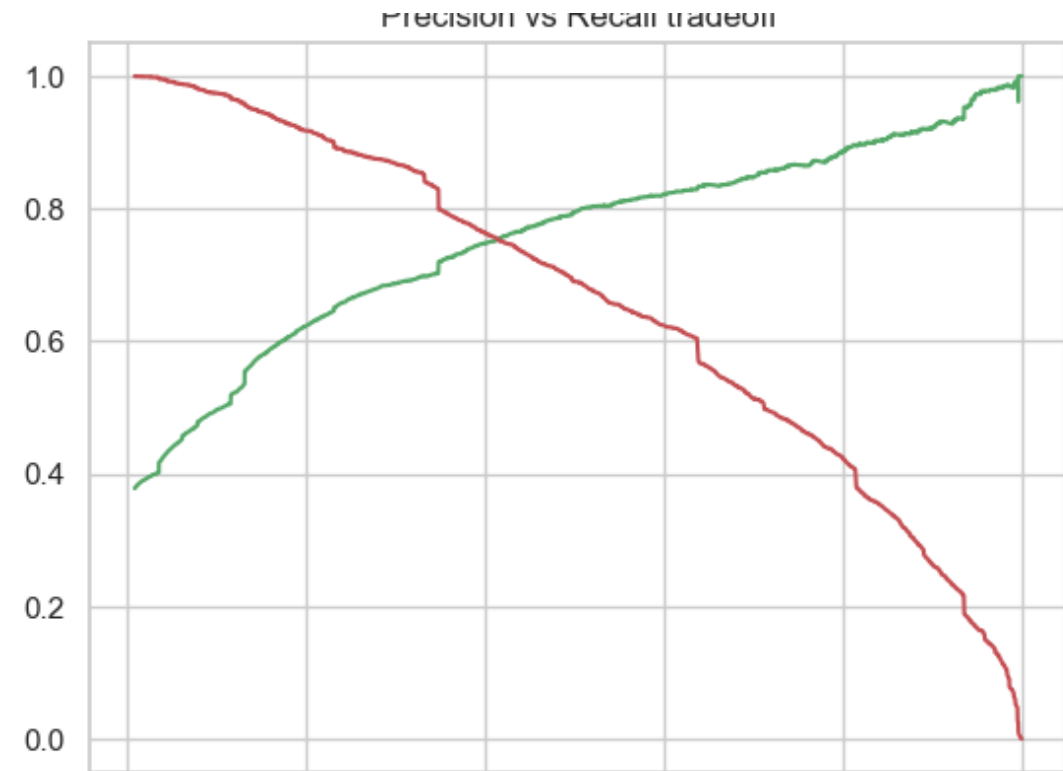
OPTIMAL CUT OFF POINT



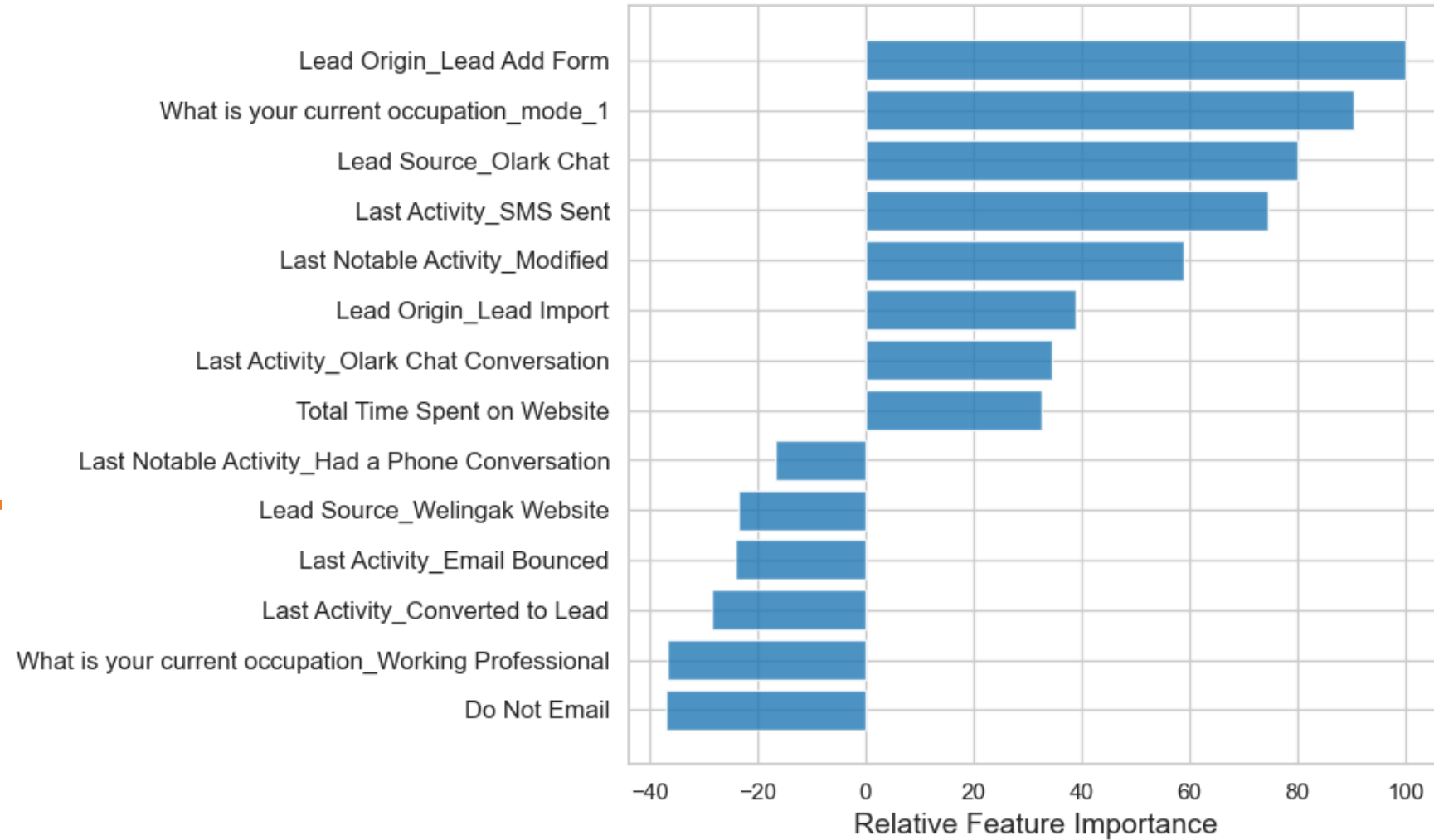
- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity
- The plot has an optimal cutoff point of 0.35. Accuracy, Sensitivity & Specificity are well balanced and are above 80%
 - Accuracy: 80.9 %
 - Sensitivity: 80%
 - Specificity: 82%

PRECISION AND RECALL

- Precision and Recall are used metrics in the industry to decide model performance and build business understanding
- 80% recall rate suggests a good model
- Precision and Recall are inversely proportional which suggests that there is tradeoff between the two.







FEATURE SELECTION





SUGGESTIONS FOR X EDUCATION

- X-Education should focus on leads having Lead Origin as Lead Add Form , Occupation as Working Professional , Lead Source as Wellingak Website and Last Notable Activity as Had a Phone Conversation.
 - We have categorized all the leads as either Hot Leads or Cold Leads. Hot Leads are those having a Lead Score of more than 35. The sales team should give priority to Hot Leads.
 - Location, Specialization, Occupation are a few variables that can probably help the company determine the nature and/or behavior of the leads a little better. This can help in determining the quality of lead and the company should be able to categorize the leads in Hot and Cold. In order to utilize these variables, the company should make these variables mandatory for the users to fill on the platform.
 - The company could also have a section on their platform which asks the users whether they would like to receive a callback and those who opt Yes will have a higher probability of converting, hence the company should focus on those leads more.
 - Our Model has a higher recall score than precision which suggests that this model can adjust with X-Education's requirements in coming future.
 - We could also see that visitors who have spent a higher time on the website were more likely to convert than the users who have spent less time.
 - We can also see that users having Last Activity as SMS Sent are more likely to convert than those who have a Last Activity as Email Bounced or Modified.
 - Users who Don't Want to be Called or Emailed are very less likely to convert, hence the company should not put their primary focus on such users.
- 
- 
- 
- 

CONCLUSION



Similar accuracy on train and test data's performance metrics suggests that the final model didn't overfit on the training data and is performing well.



Higher sensitivity will ensure that the leads who are likely to get Convert are correctly predicted where as high Specificity will ensure that the leads that have a lower probability of getting converted are not selected.



Depending on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.