



# Lead Scoring Case Study



# Goal

- The goal is to improve how we rank potential leads by taking into account different attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc., using different methods to assign scores to leads and focussing on the Hot leads for higher Lead conversion.



# Problem Statement



- X Education, an online education company, caters to industry professionals by offering courses. Professionals visiting the website daily after finding the company's courses on various platforms may explore different courses and enrol in them by submitting forms. Leads are generated when visitors provide their contact details, either through forms or referrals. The sales team then contacts these leads through calls and emails, resulting in a typical conversion rate of 30%.
- Despite acquiring numerous leads, X Education struggles with a low conversion rate. To improve efficiency, the company aims to identify the most promising leads, referred to as 'Hot Leads.' By focusing efforts on these potential leads, the conversion rate is expected to increase, enhancing the effectiveness of the sales team's interactions.





# Goals of the Case Study

1. Build 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.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.



# Steps involved

- Reading and understanding data
- Data Cleaning
- Univariate and BI-variate analysis
- Data Preparation
- Splitting the data
- Scaling the cloumns

- 
- 
- Feature selection using RFE
  - Building the model
  - Model Evaluation
  - Finding the optimal cutoff of the probability, using ROC curve
  - Precision & Recall
  - Making predictions on test
  - Feature Importance



# Data Sourcing Cleaning and Preparation

- Reading the Data from the CSV file.
- Outlier Treatment
- Data Cleaning, Handling Null Values and Removing Higher Null Values Data
- Redundant Data removal from Columns
- EDA

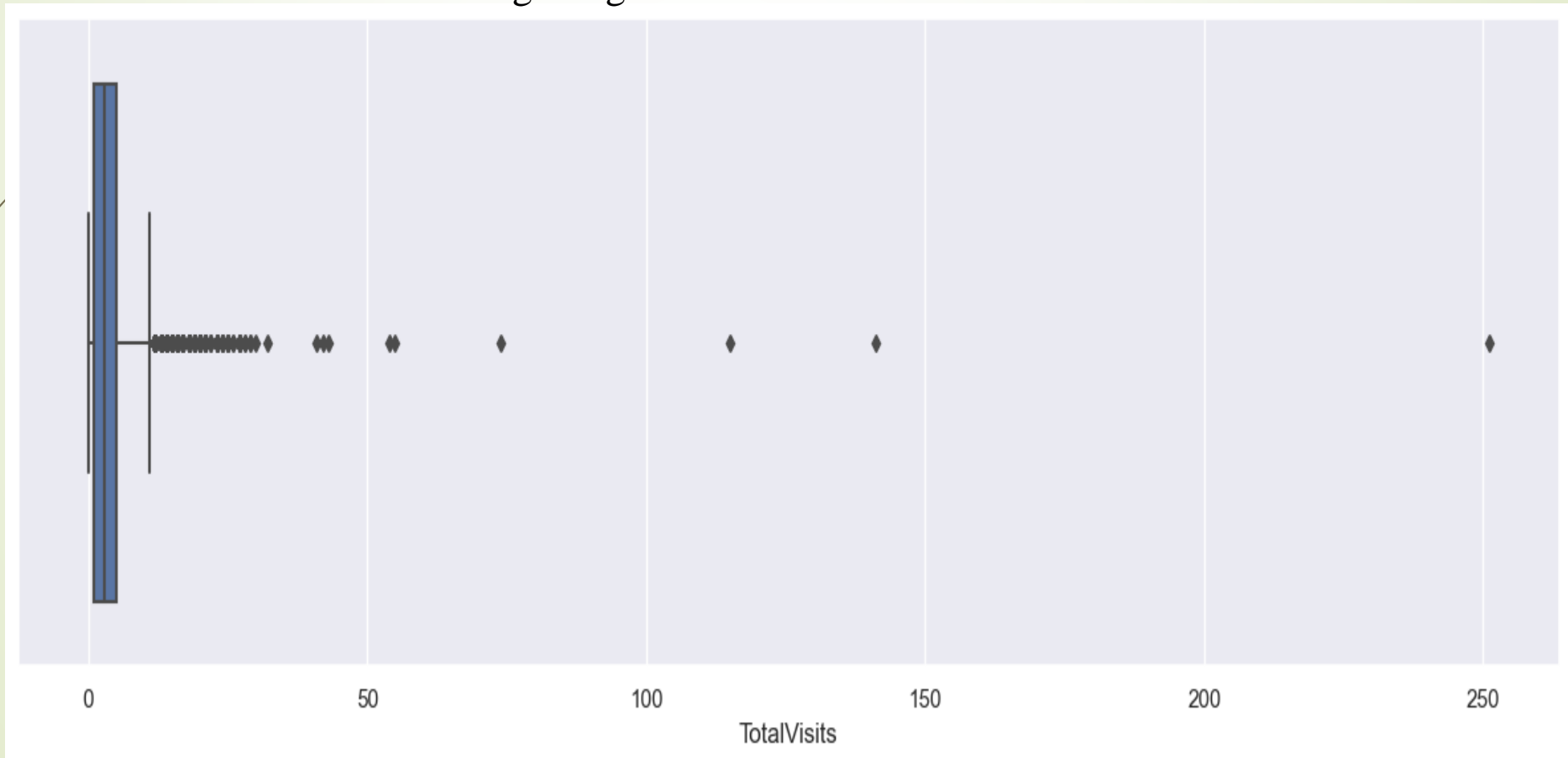


# Outliers

Restricting Data to 95 Percentile to handle Outliers.

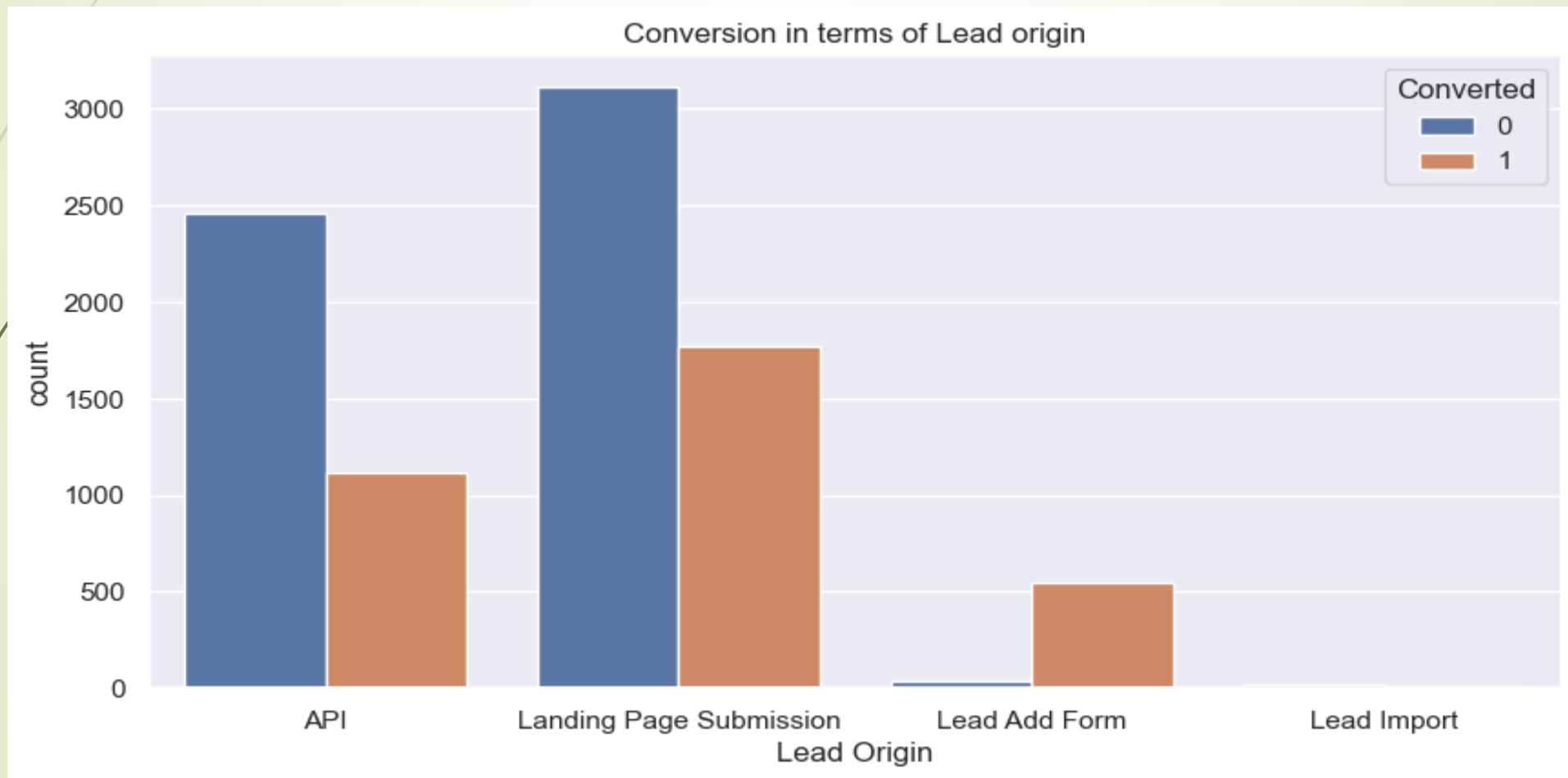
## Insights

- Median for both are same.
- Increase in total visit has a slight higher chance to be converted.






# Univariate Analysis of Categorical Variables





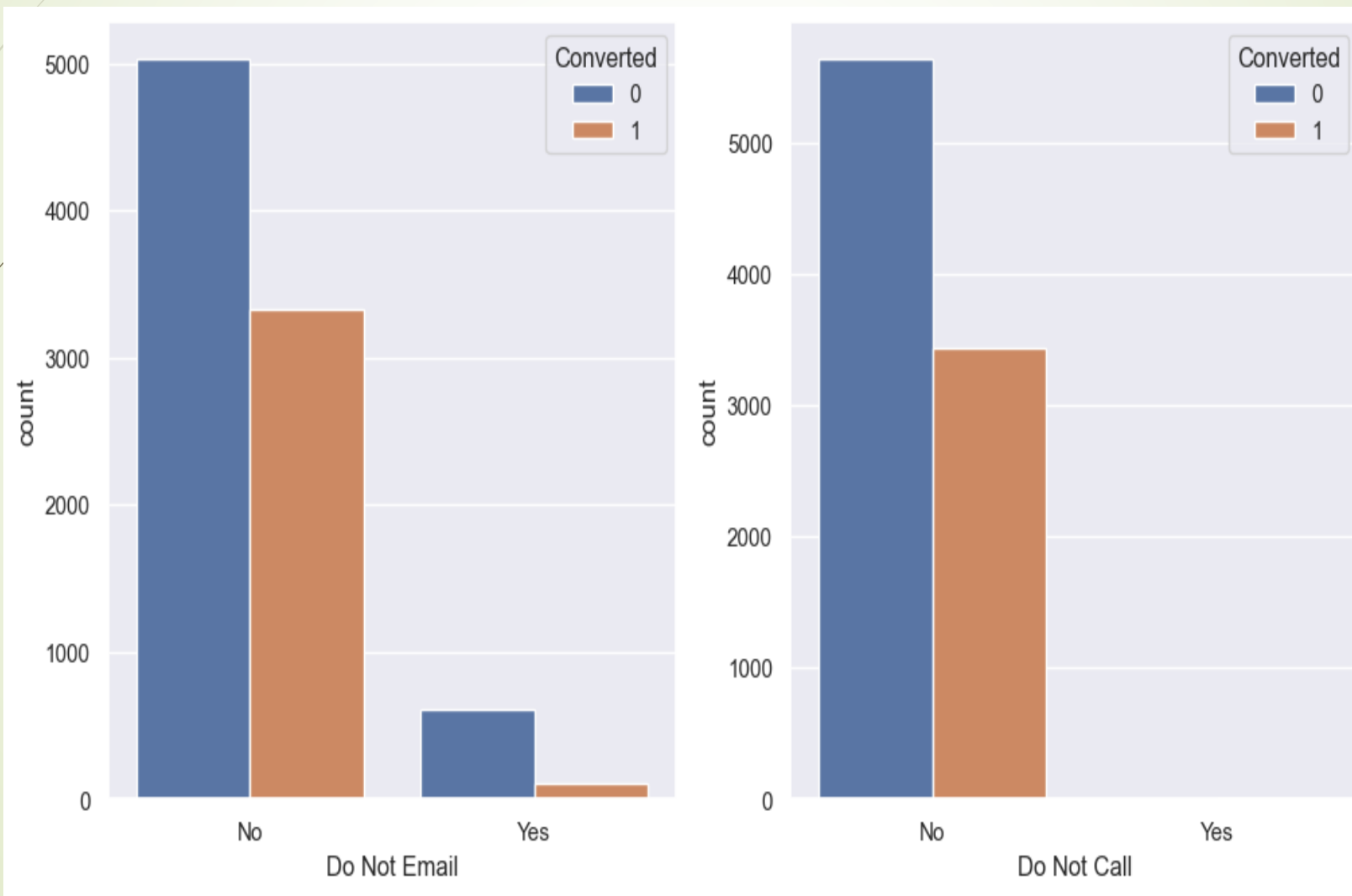
## ■ Insights

- Lead Add Form & Landing Page Submission though has avg of 33% conversion, it generates the most no. of leads.
  - To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also, generate more leads from Lead Add form since they have a very good conversion rate
- 

# Univariate Analysis of Communication Method

## Insights

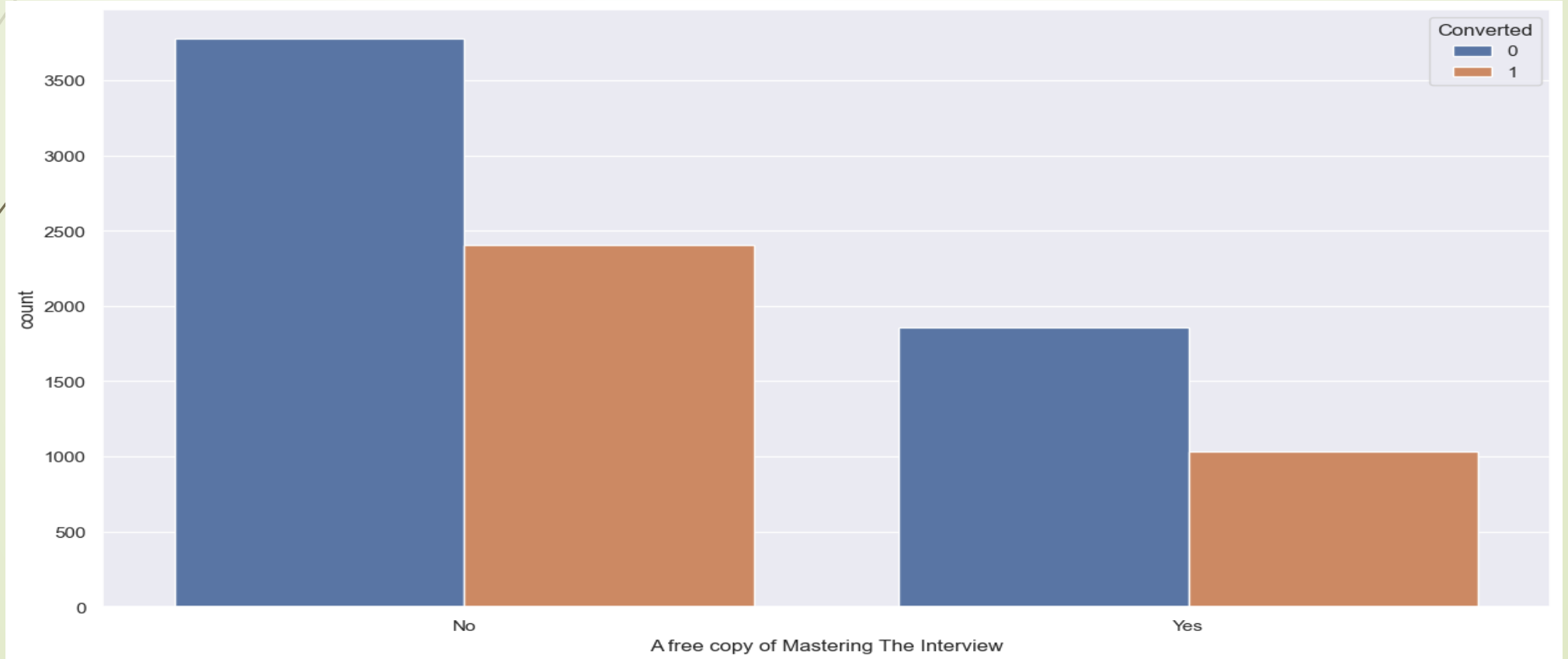
- More than 95% of the leads do not prefer to be called or emailed



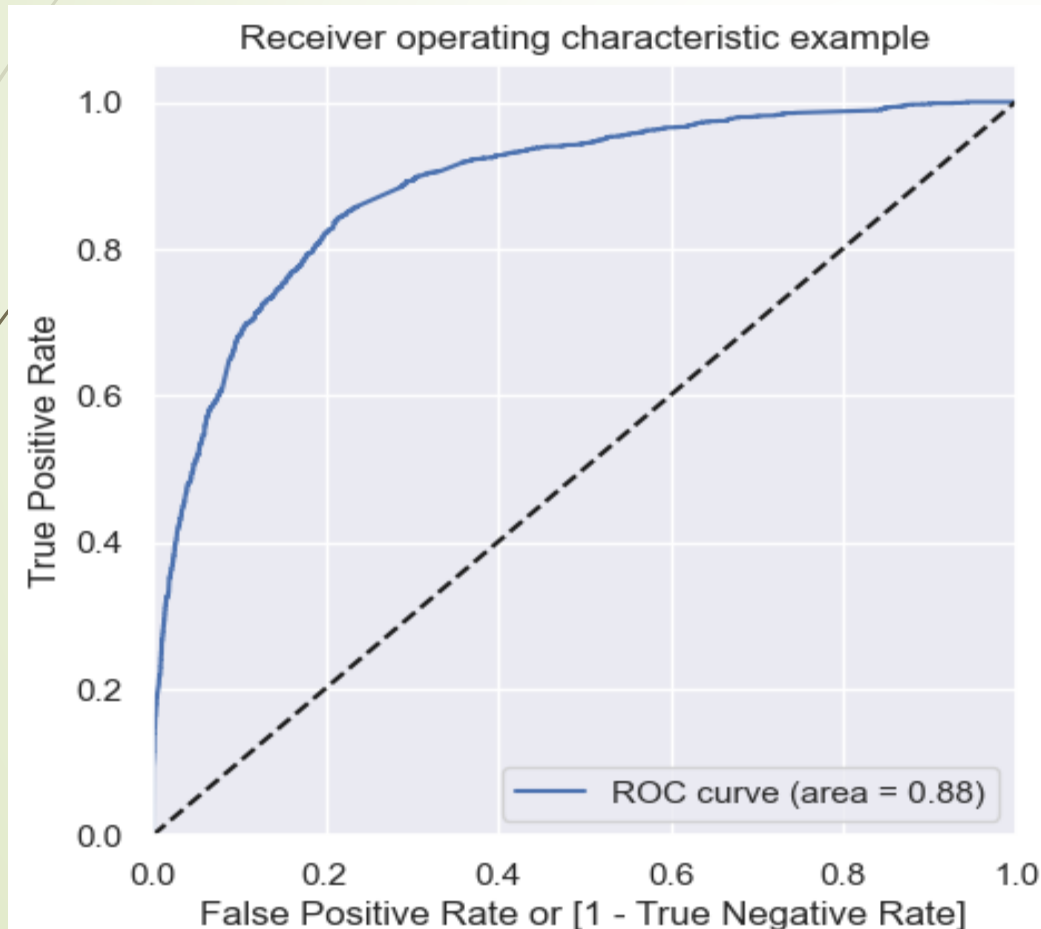
# Bivariate Analysis of Newsletter Subscription

## Insights

- Most of the lead opted out for the free copy and those opted for it their conversion rate is lower than those opted in.



# Finding the optimal cutoff of the probability, using ROC curve



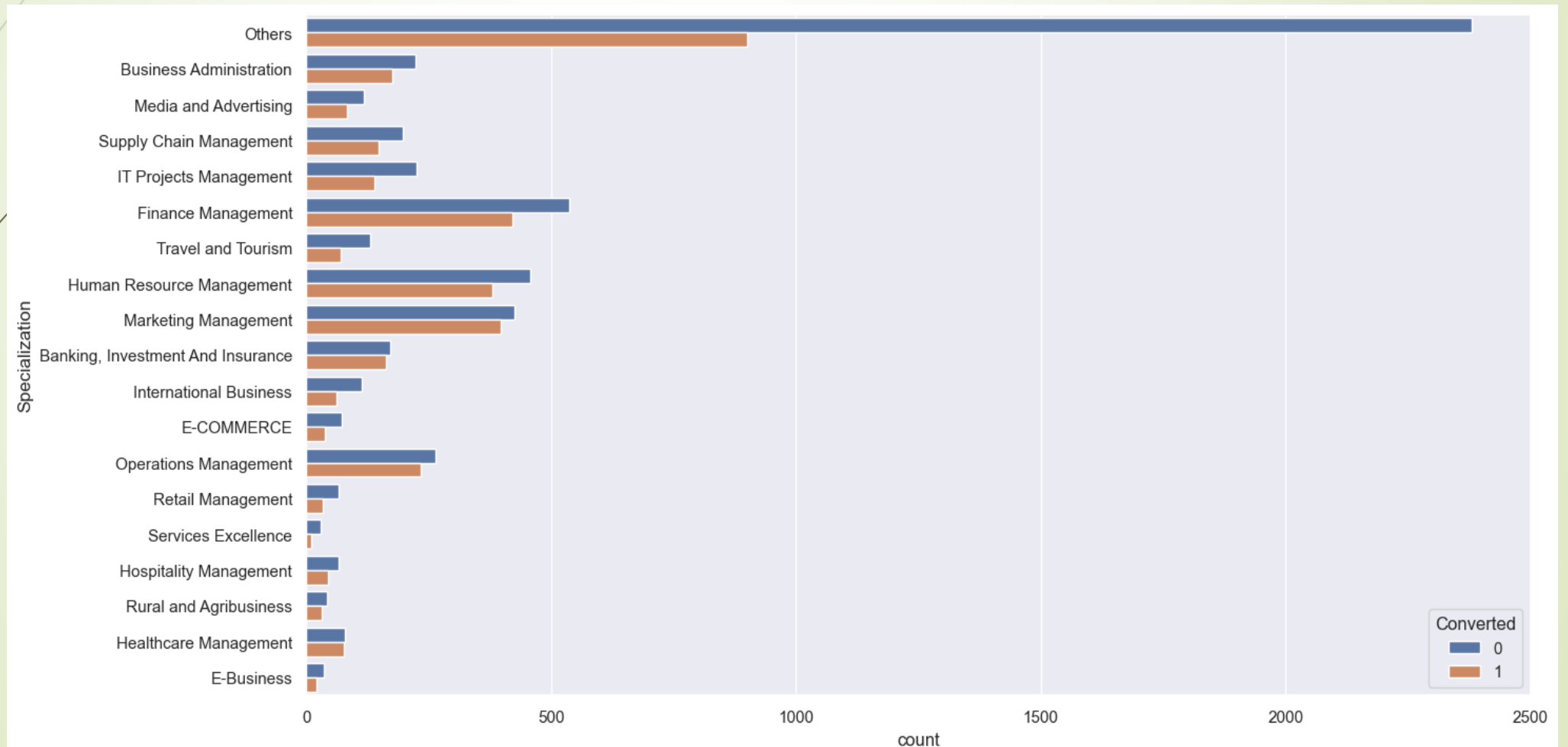
## Insights:

The area under the curve should be a value closer to 1, since it is 0.88 our model is good

# Bivariate Analysis of Specializations

## Insights

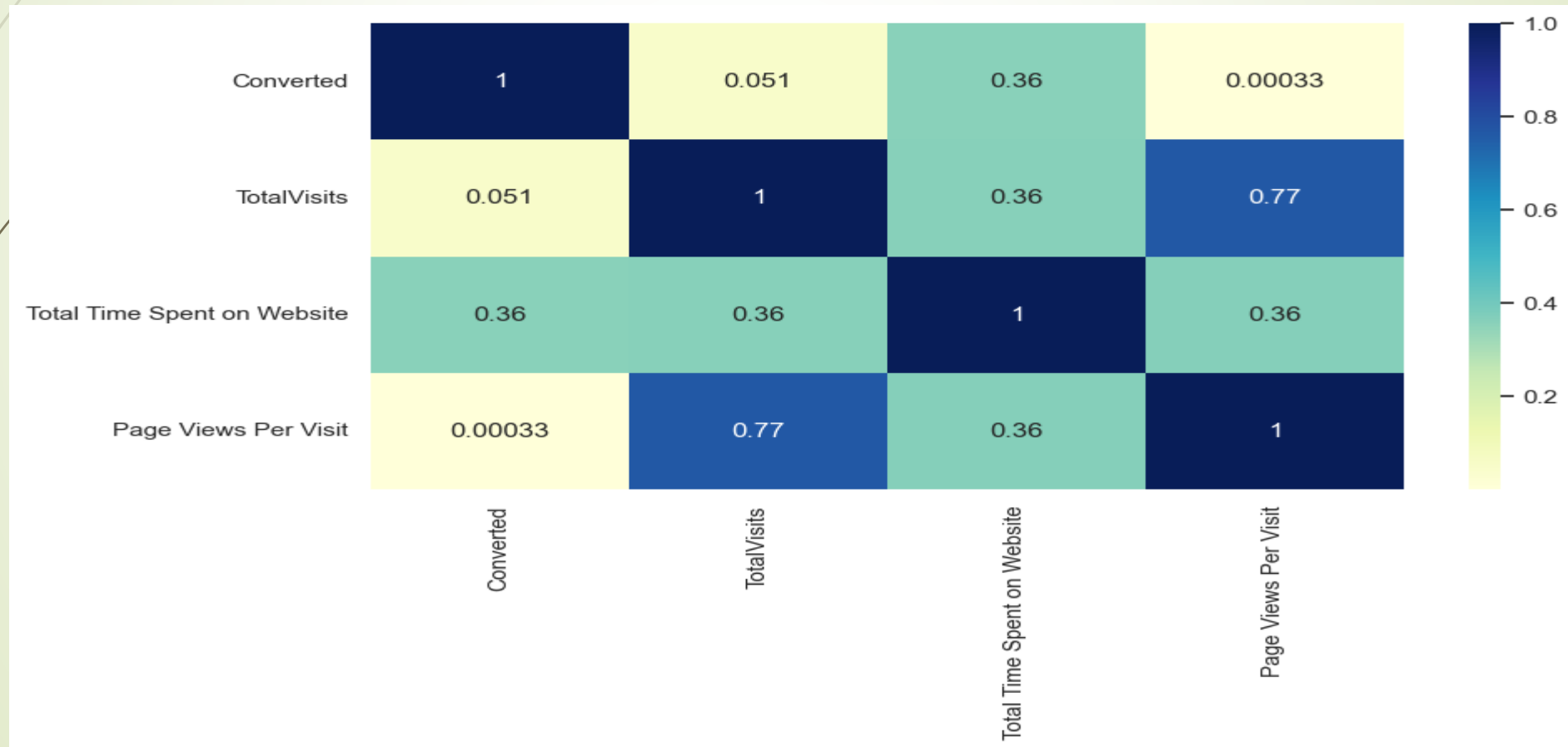
- Most of the Specialization has a more than 40 % conversion rate, with Finance Management and Human Resource Management having higher leads and conversion rates.



# Multivariate Analysis Using Heatmaps

## Insights:


We can see the Total Visits and Page Views Per Visit has hi co-lonearity, hence either of the two has to be there.



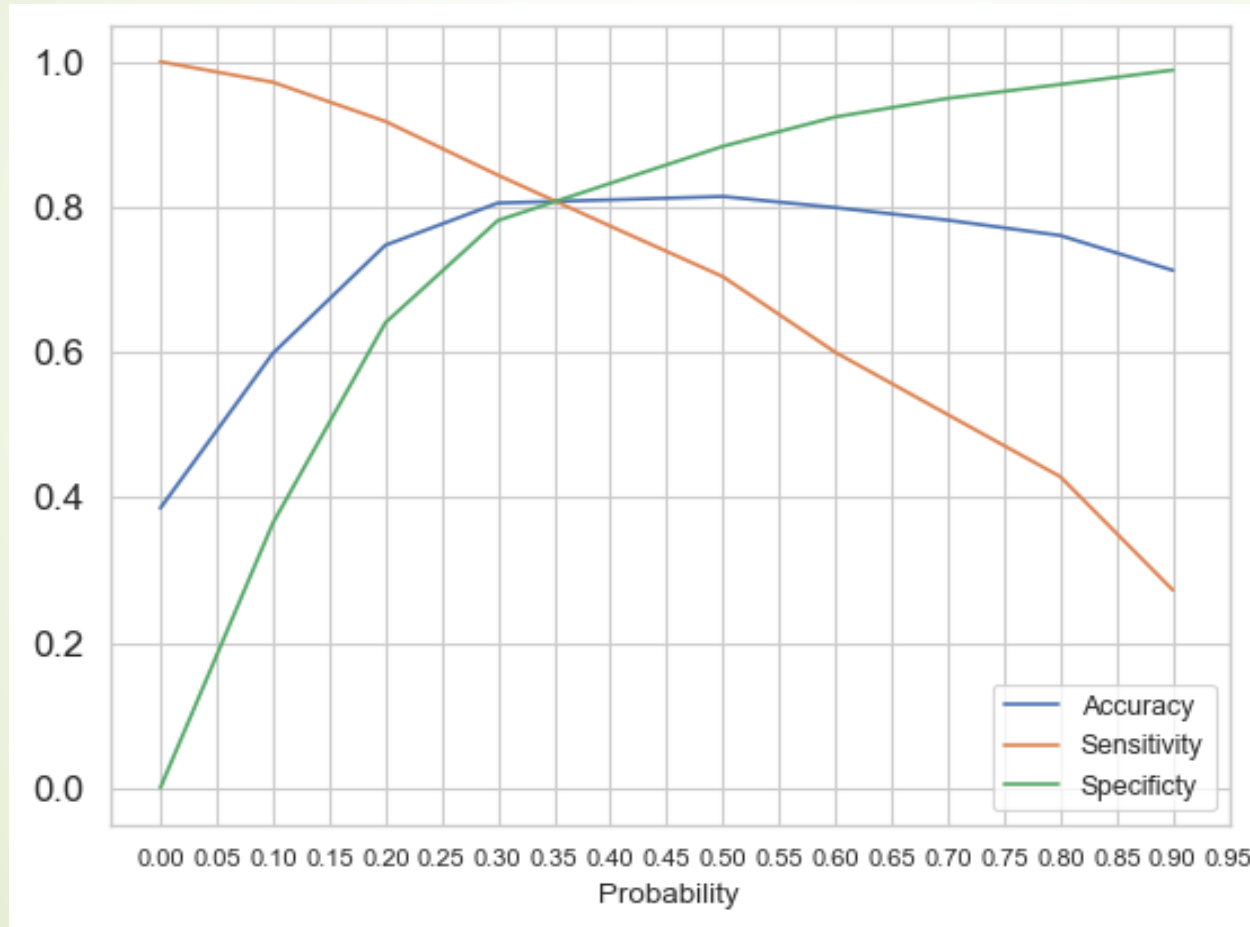




# Model Building

1. Feature Selection using RFE
  2. Determined Optimal Model using Logistic Regression
  3. Accuracy, Sensitivity and Specificity Calculation
- 

# Model Evaluation-Sensitivity and Specificity on Training Data Set



Selecting Cutoff as 0.35 from graph based on:

1. Accuracy = 81%
2. Specificity = 88% and
3. Sensitivity = 70%



# Impact Variables

1. Total Time Spent on Website
2. Lead Source\_Welingak Website
3. Last Activity\_Email Opened

The above three variables are the most impactful variables for predicting conversions.

Other Impactful variables are:

1. Total Visits and Total Time Spent on Website
2. Last Activity:
3. Specialization:
4. Current Occupation
5. Do Not Email and Do Not Call
6. Last Notable Activity.





# Result

- The model achieves a reasonably good accuracy, sensitivity, and specificity, suggesting that it is capable of identifying hot leads to a certain extent.
- - The accuracy score is around 81%, indicating that the model's predictions are accurate for a significant portion of the dataset.
- - The sensitivity score is also around 70%, which means that the model is able to correctly identify a substantial proportion of actual converting leads.
- - The specificity score is around 88%, implying that the model can effectively distinguish between converting and non-converting leads.



# Recommendations

- Lead Source and Generation: Focus on generating more leads from sources like 'Welingak Website', 'Reference', and 'Google' as these sources have higher conversion rates.
- Focus on 'Working Professionals' and 'Unemployed' individuals highlighting course efficacy and career benefits.
- Prioritize Leads engagement through emails and SMS as they show a higher tendency to get converted into customers.
- Segment the Leads into “Hot” and “Cold” leads on the basis of the lead scores. Prioritize communication and follow-ups with 'Hot Leads' for maximum conversion.

- 
- 
- Continuous Monitoring and Adaptation of the Lead Scoring Model by adjusting threshold and updating the model periodically will help in improving the accuracy of prediction.
  - The sales team should focus on leads with a higher lead score since they have a higher probability of conversion. They should also consider leads with slightly lower scores, as they might still have a decent chance of converting.
  - Paying adequate attention using proper communication and relevant information, to leads with the 'Last Notable Activity' as 'Modified', can boost their chances of conversion.