
Team :

Ayesha Taranum
Aishwarya Behera
Sumit Bansal

# LEAD SCORING CASE STUDY

# PROBLEM STATEMENT

- X Education acquires leads from website visits, form submissions, and referrals.
- Typical lead conversion rate is around 30%, with 100 leads resulting in approximately 30 conversions.
- Improve lead conversion rate to target 80% by identifying and focusing on "Hot Leads."
- Develop a lead scoring model to prioritize leads based on their likelihood to convert.
- Higher lead scores indicate greater conversion potential; lower scores suggest lower potential.
- Utilize the lead scoring model to enable the sales team to focus on high-potential leads, increasing overall conversion rates.
- Streamlined sales efforts, improved efficiency, and a significant rise in lead conversion rates towards the target of 80%.

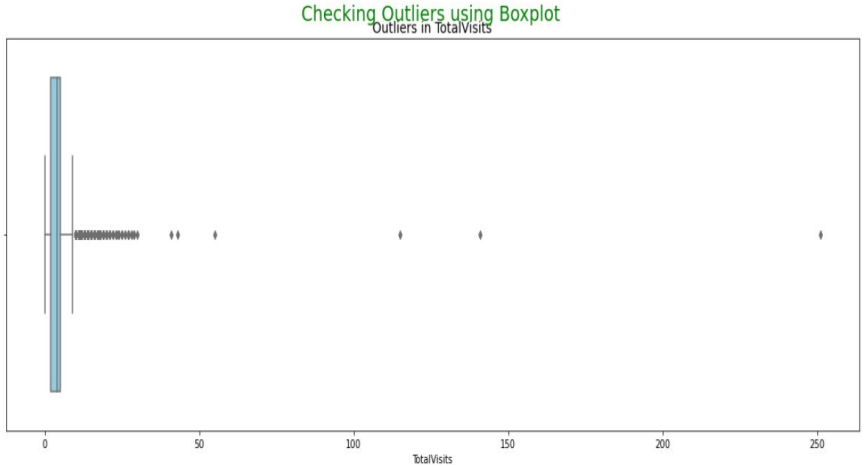
# STRATEGY

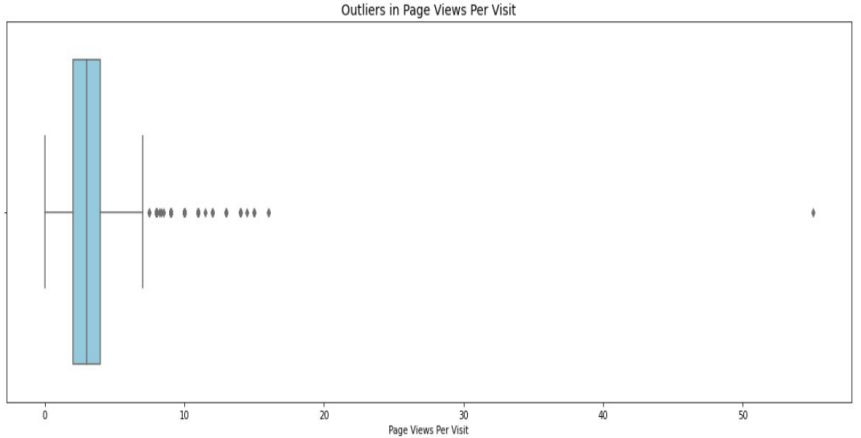
- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out the most helpful attributes for conversion
- Dummy variable creation
- Test - Train split
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation.

# DATA MANIPULATION

- Total number of Columns 37. Total number of rows 9240
- Dropping the columns having more than 3000 missing values like Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Score, Asymmetrique Activity, Asymmetrique Profile Index, Tags
- The variable "What matters most to you in choosing a course" is overwhelmingly dominated by the single value. Hence removed this column from the dataset.
- Single value features like ‘I agree to pay the amount through cheque’, 'Get updates on DM Content’, 'Update me on Supply Chain Content’, 'Receive More Updates About Our Courses’ and 'Magazine’ also got dropped.
- Highly skewed variables can introduce bias and inaccuracies into logistic regression models. Hence, 'Do Not Call’, 'Search', 'Newspaper Article’, 'X Education Forums’, 'Newspaper’, 'Digital Advertisement’ and ‘Through Recommendations’ also got dropped.
- Removing the variables 'Prospect ID' and 'Lead Number’ as they do not hold any significant relevance or utility for our analysis.
- There are few columns, who have “Select” as level. 'Specialization', 'How did you hear about X Education', 'City’, ‘Lead Profile’ columns have to be taken care. The reason is that the customer did not select any option from the list and hence for such columns the data remained as default.

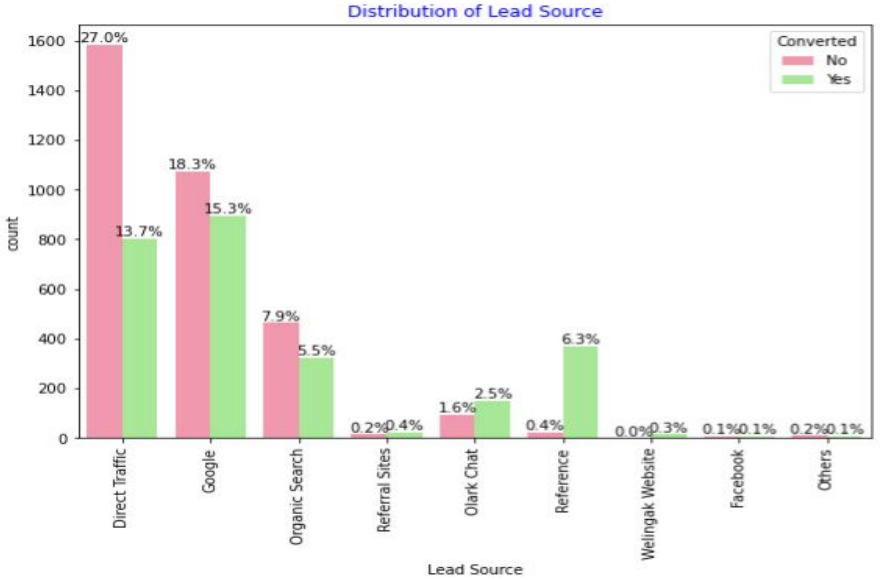
# OUTLIER ANALYSIS

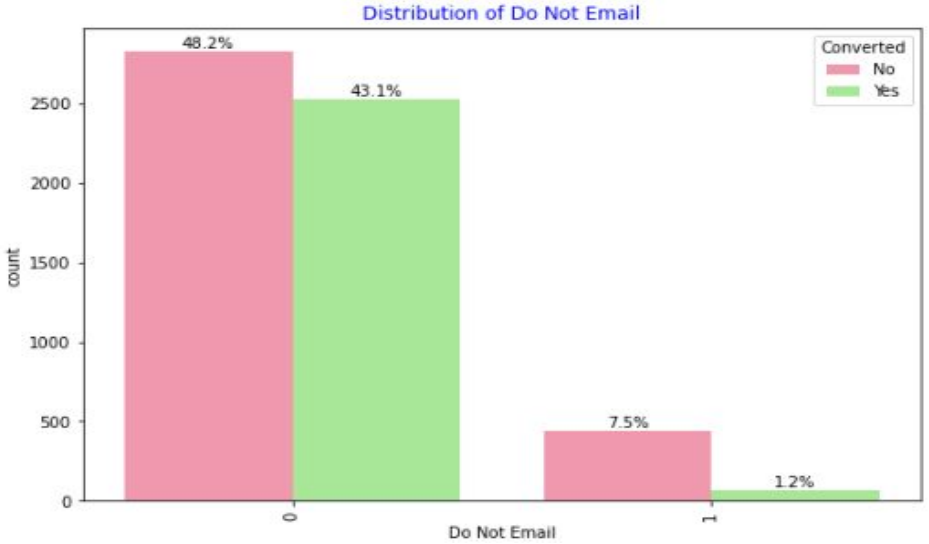



"TotalVisits" and "Page Views Per Visit" both exhibit outliers, as evident from the boxplots. These were taken care.

# EXPLORATORY DATA ANALYSIS



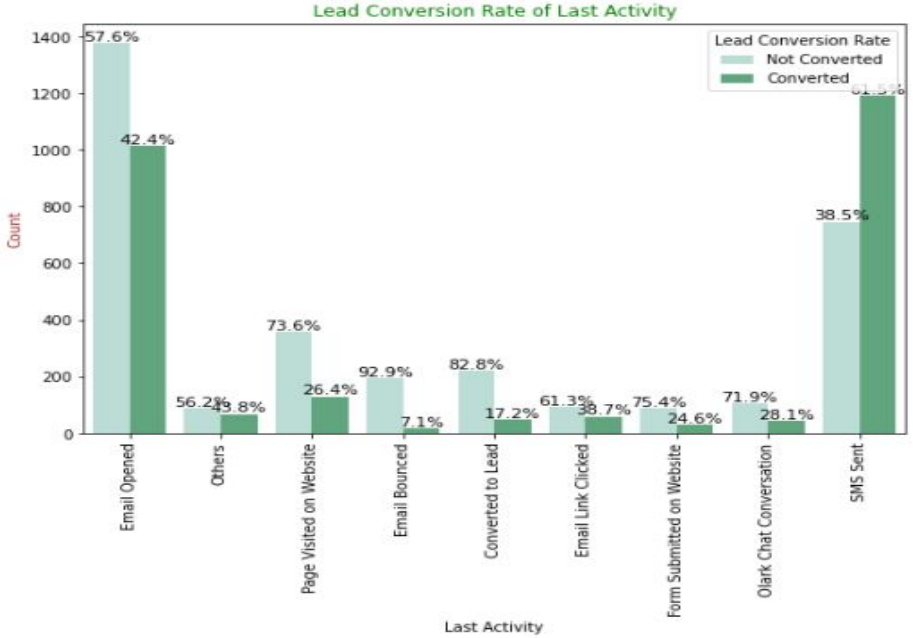

## LEAD SOURCE VS CONVERTED

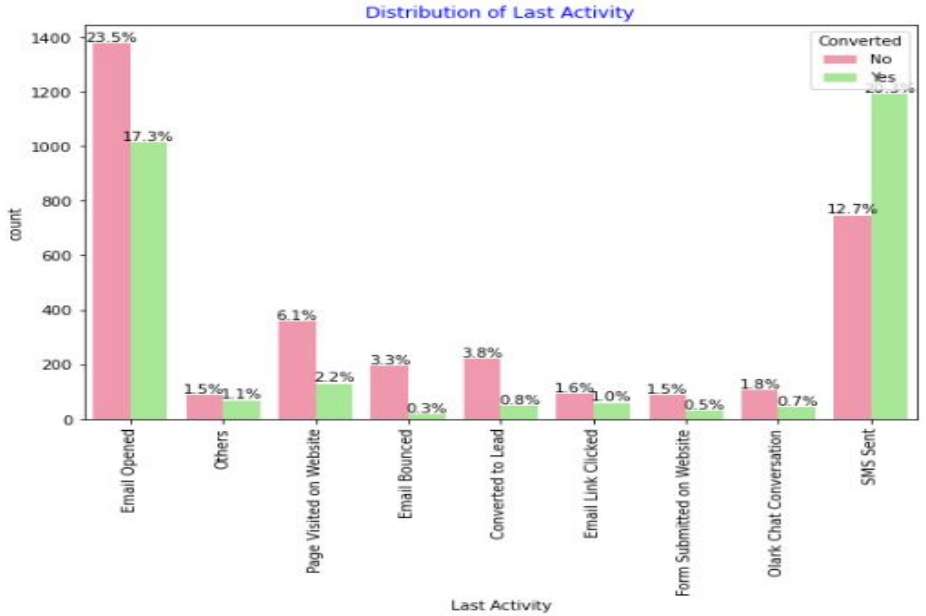
Direct Traffic has had high conversions as compared to the other modes.

## DO NOT EMAIL VS CONVERTED

Customers preferred not to be informed through emails

# EXPLORATORY DATA ANALYSIS



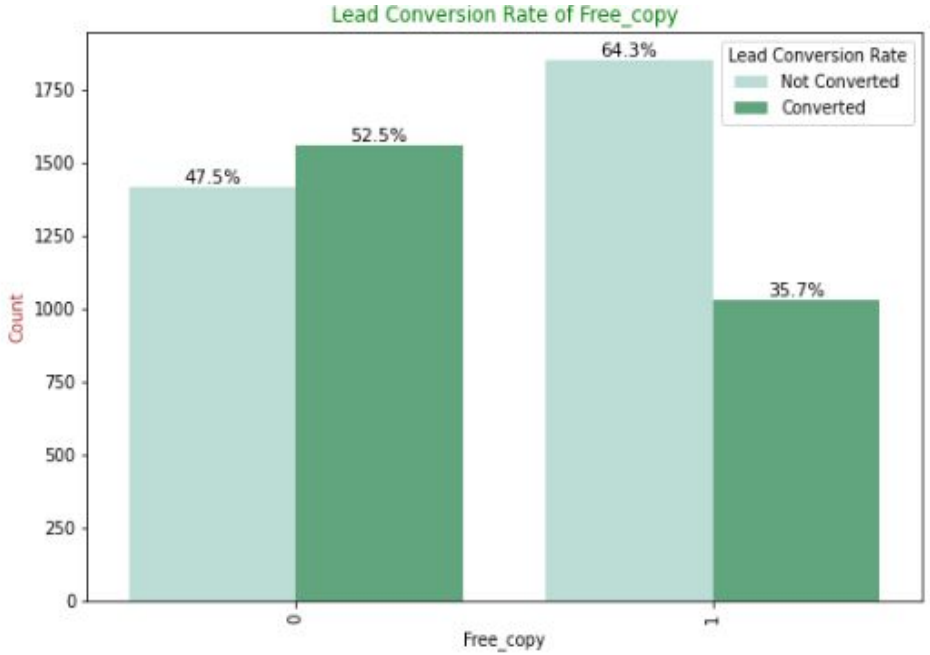

## LAST ACTIVITY VS CONVERTED

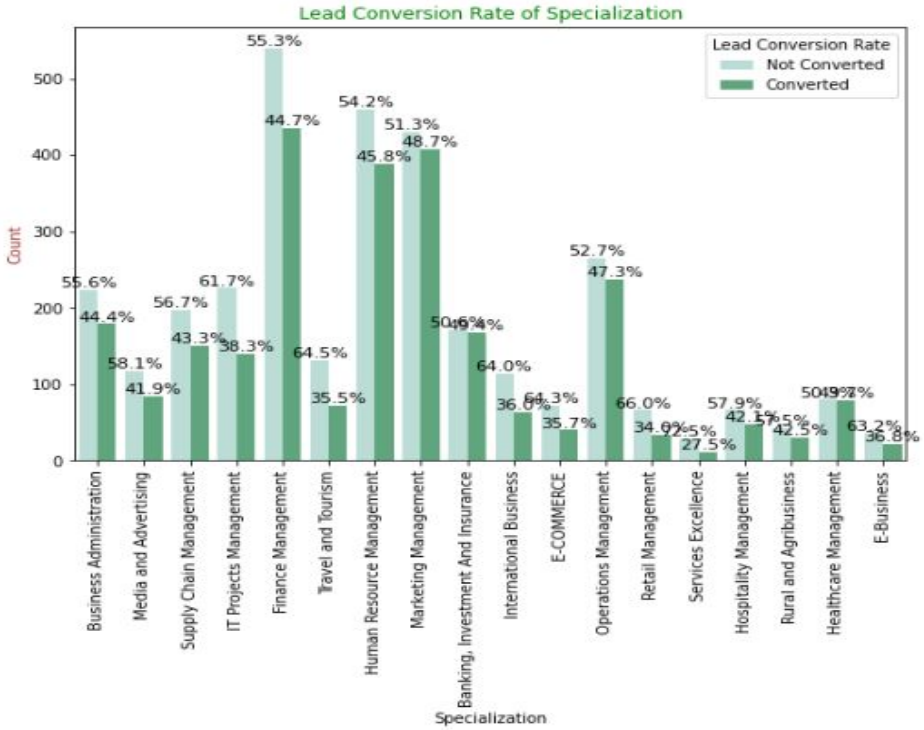
SMS has shown to be promising method for getting higher confirmed leads, emails also has high conversions

## LAST NOTABLE ACTIVITY VS CONVERTED

Most leads are converted with messages. Emails also induce leads.

# EXPLORATORY DATA ANALYSIS




## A FREE COPY OF MASTERING THE INTERVIEW VS CONVERTED

Leads prefer Less copies of interviews

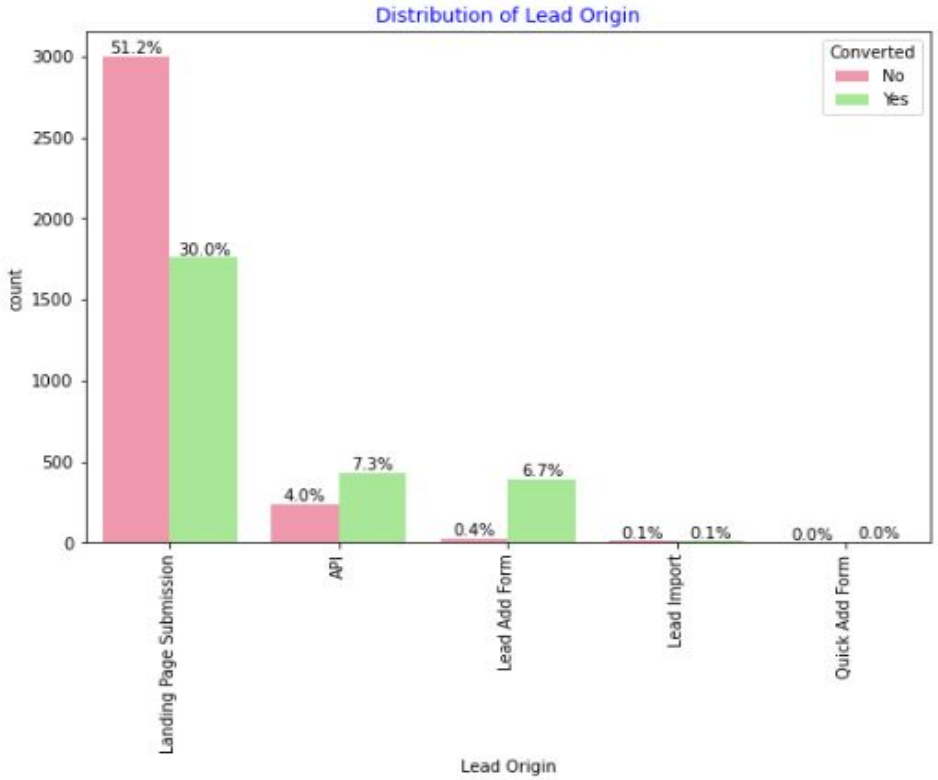
## SPECIALIZATION VS CONVERTED

Most of the leads have no information about specialization. On the other hand, marketing management, human resources management and finance management has high conversion rates. People from these specializations can be promising leads

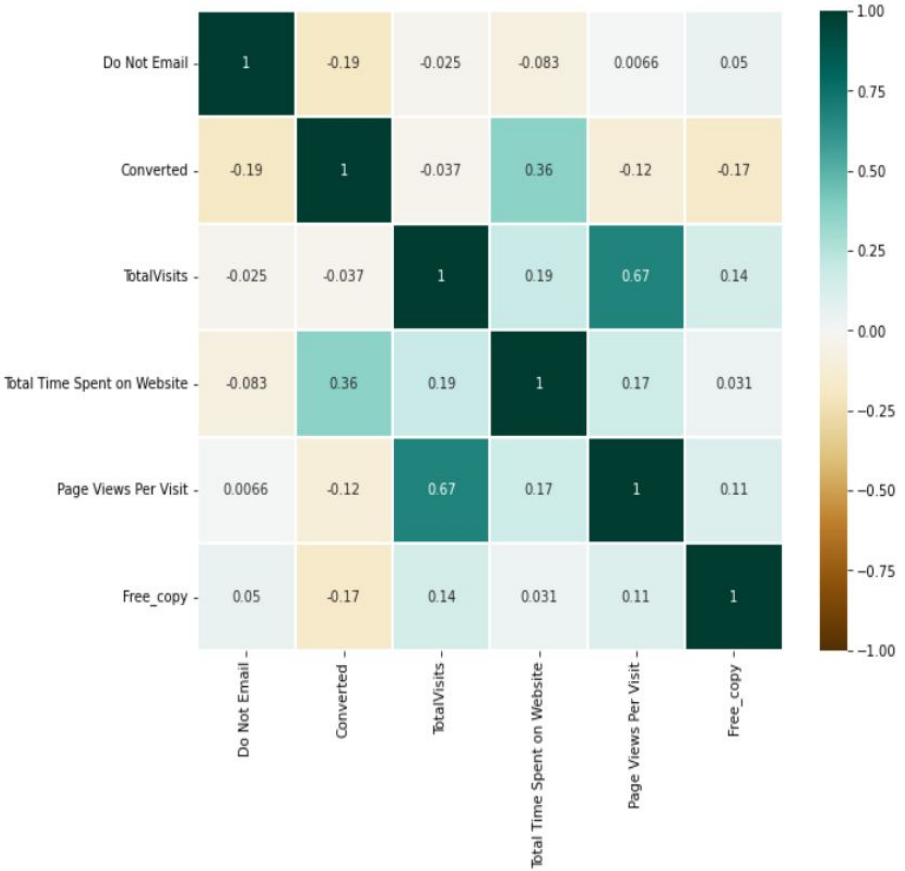
# EXPLORATORY DATA ANALYSIS

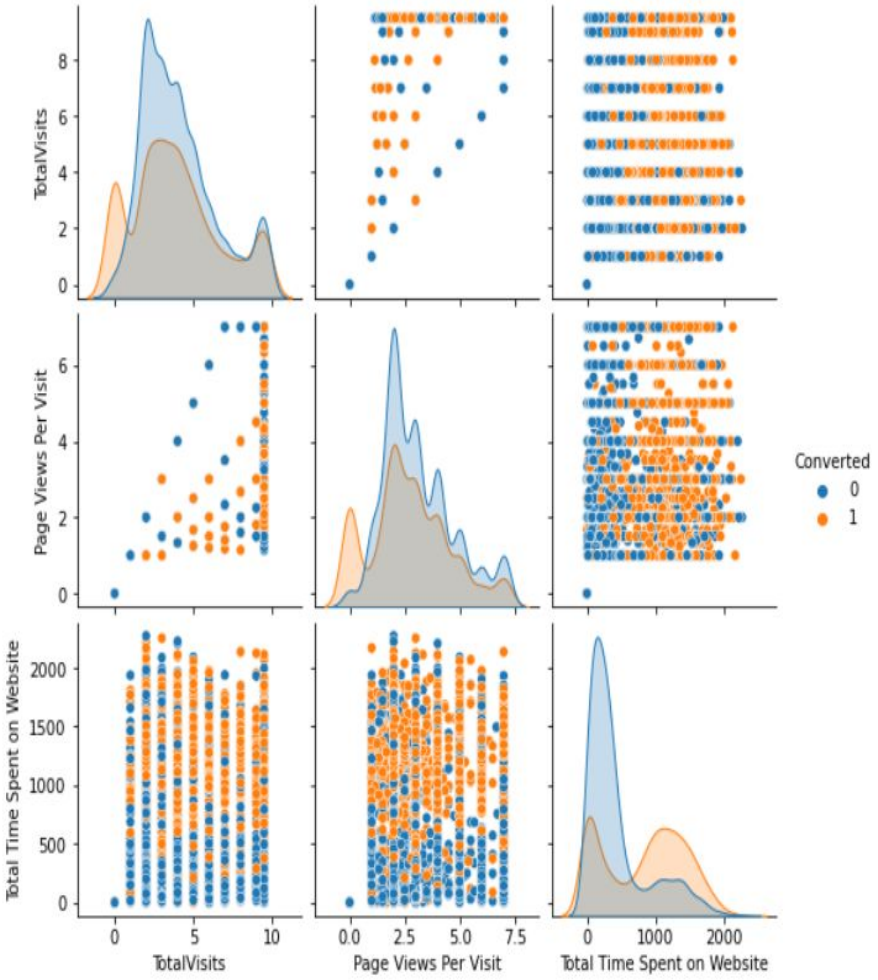
## LEAD ORIGIN VS CONVERTED

Landing page submissions have consistently resulted in high lead conversion rates.

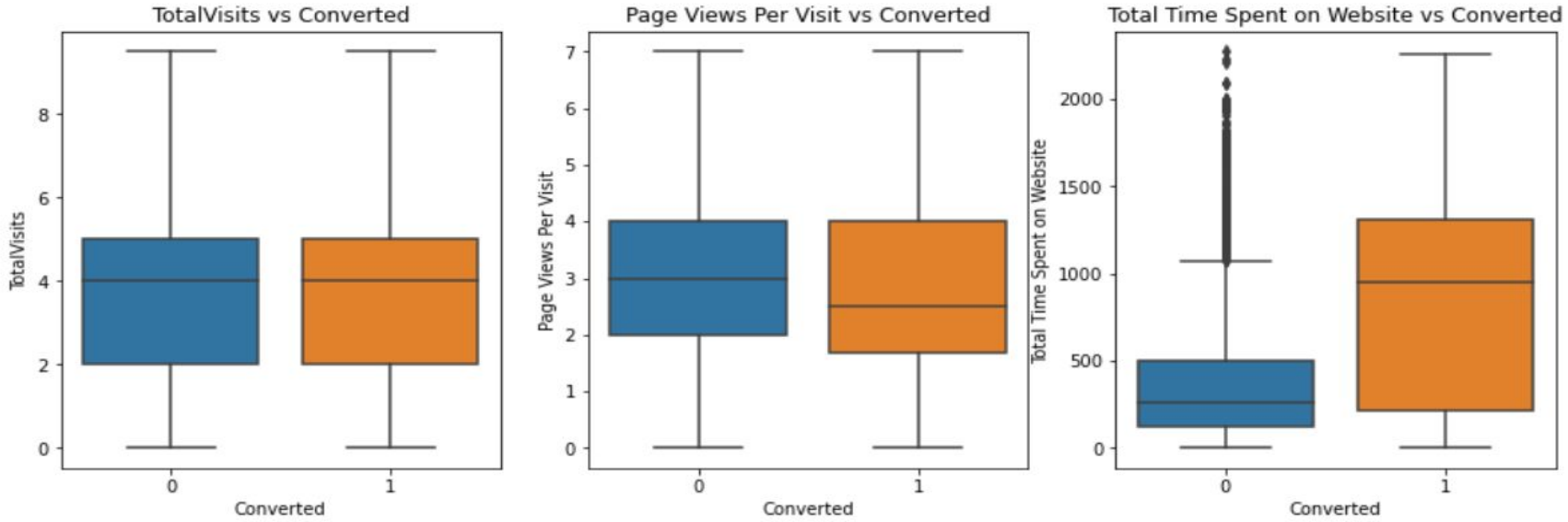


# BIVARIATE ANALYSIS



# BIVARIATE ANALYSIS

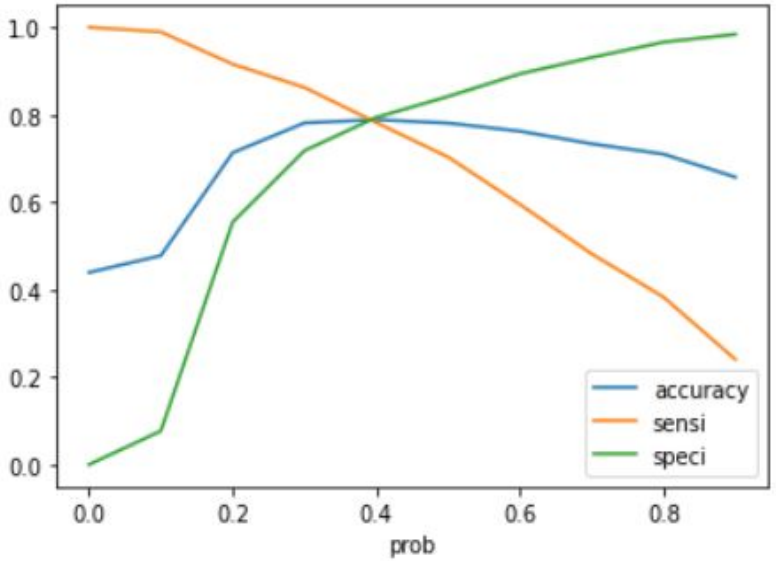


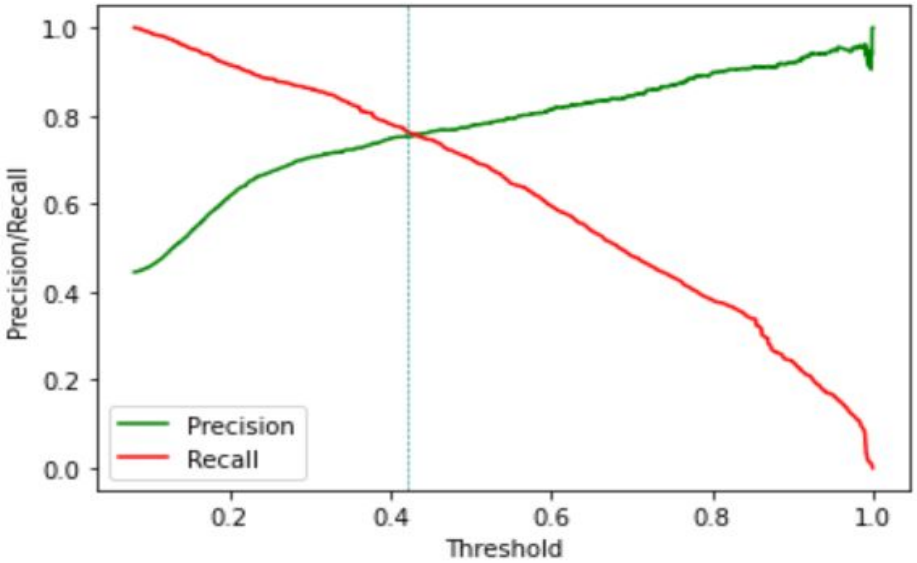

The boxplot for "Total Time Spent on Website" suggests that past leads who spend more time on the website tend to be more successfully converted compared to those who spend less time. This could indicate that spending more time on the website might be associated with higher engagement and interest in the products or services offered, which in turn leads to a higher conversion rate. This insight can be valuable for designing strategies to engage and convert leads effectively.

# DATA MODELLING/MODEL BUILDING

- Dummy Variable Creation
- Splitting the data into training and test set.
- Split the dataset into 70% train and 30% test
- Scale the three numeric features present in the dataset
- RFE for feature selection
- Running RFE with 25 variables
- Building first model by dropping columns whose p-value is greater than 0.05 and vi value is greater than 10.
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using Train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions

# ACCURACY, SENSITIVITY AND SPECIFICITY




78.9% Accuracy

78.2% Sensitivity

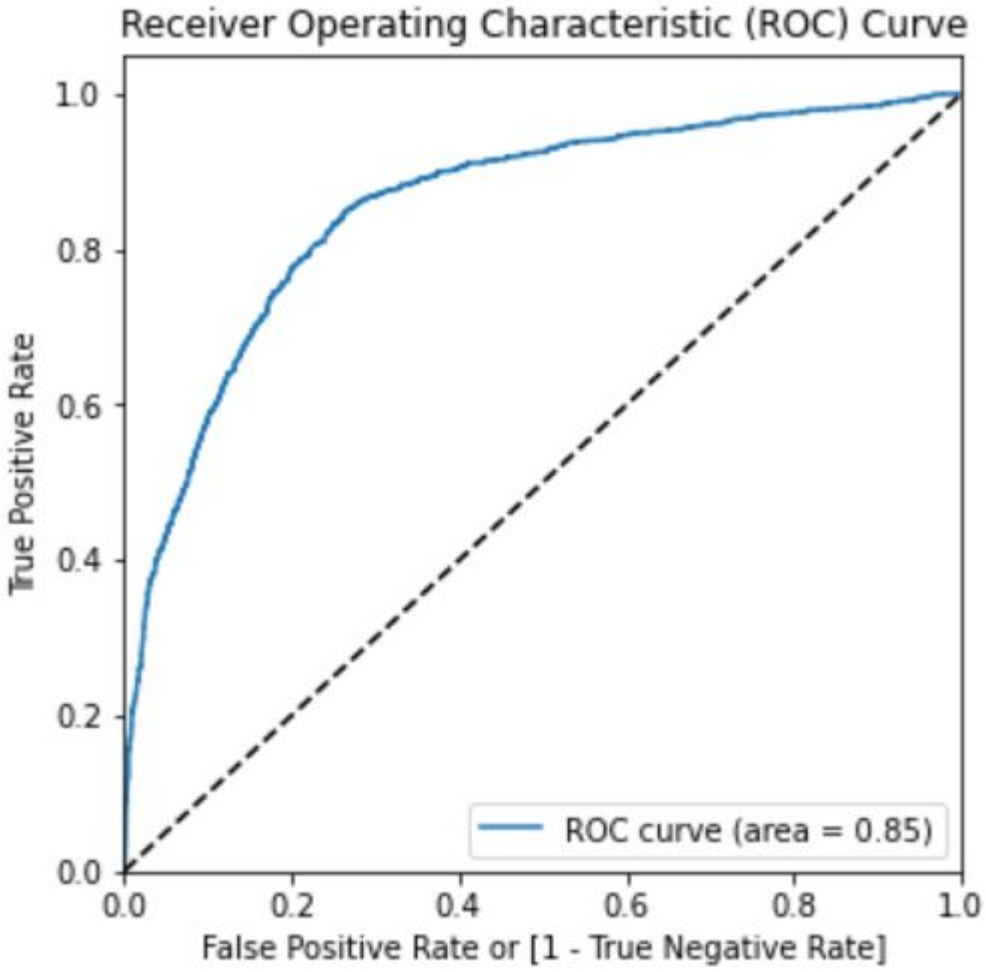
79.4% Specificity

74.82% Precision

78.19% Recall

# ROC CURVE

- An area under the ROC curve of 0.85 is indeed indicative of a good predictive model.
- The ROC curve measures the model's ability to distinguish between positive and negative cases, and an AUC of 0.85 suggests that the model is performing well in this regard.
- Generally, a curve above 0.7 is considered good, and an ROC of 0.85 is a strong indication of the model's predictive power.
- Test set threshold has been set as 0.45



# Logistic Regression Model

## Metrics for Logistic Regression Model:

**Accuracy:** 0.4050056882821388

**Precision:** 0.4231926770708285

**Recall:** 0.8924050632911392

**Specificity:** 0.0072314049586776866

**F1 Score:** 0.5741042345276874

# CONCLUSION

**Exploratory Data Analysis -**

- People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversions
- SMS messages can have a high impact on lead conversion
- Landing page submissions can help find out more leads
- Marketing management, human resources management has high conversion rates. People from these specializations can be promising leads
- References and offers for referring a lead can be good source for higher conversions.
- An alert messages or information has seen to have high lead conversion rate

# CONCLUSION

## Logistic Regression Mode -

- The model shows high close to 79% accuracy
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and precision, recall curves.
- The model shows 78% sensitivity and 79% Specificity
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate