LEAD SCORING CASE STUDY

By

Ayesha Taranum
Aishwarya Behera

Sumit Bansal

INTRODUCTION

X Education Co. 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.

The company markets its courses on several websites and search engines like Google. 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

STRATEGY

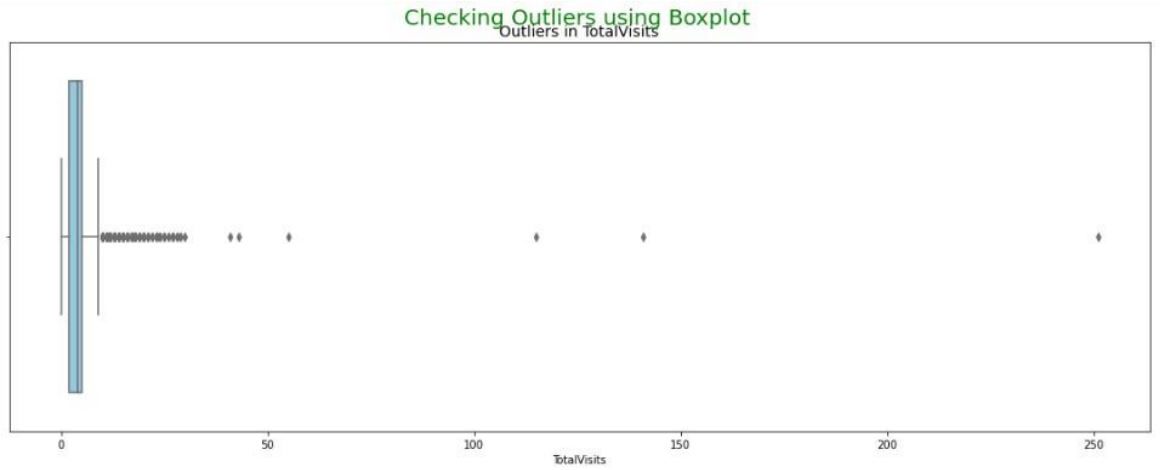
- Import data

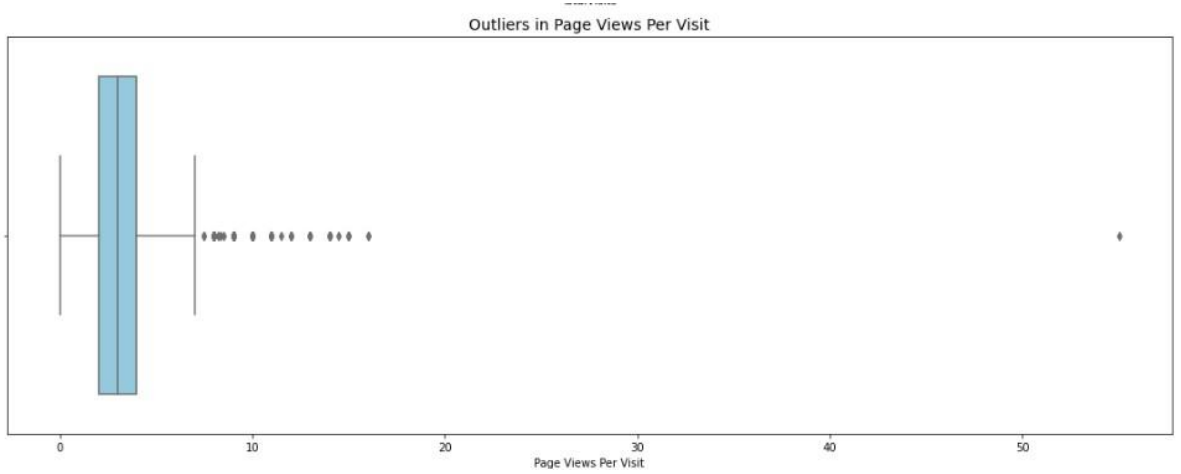
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Dummy variable creation
- Test – Train Split
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score 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.
- "TotalVisits" and "Page Views Per Visit" both exhibit outliers, as evident from the boxplots. These were taken care.

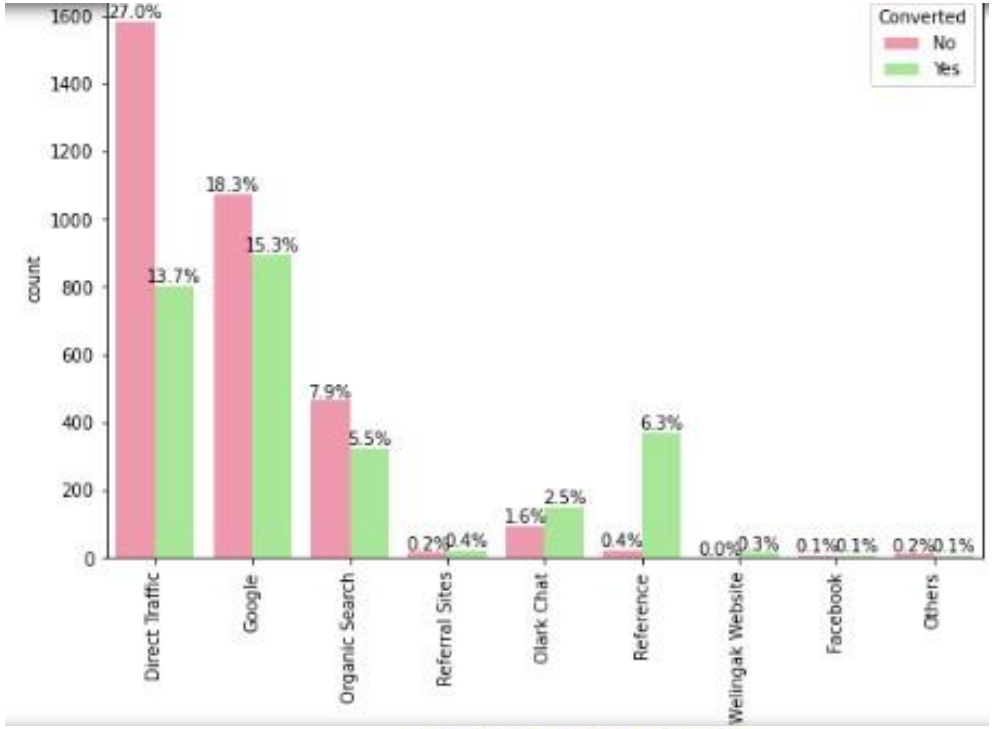




EXPLORATORY DATA ANALYSIS

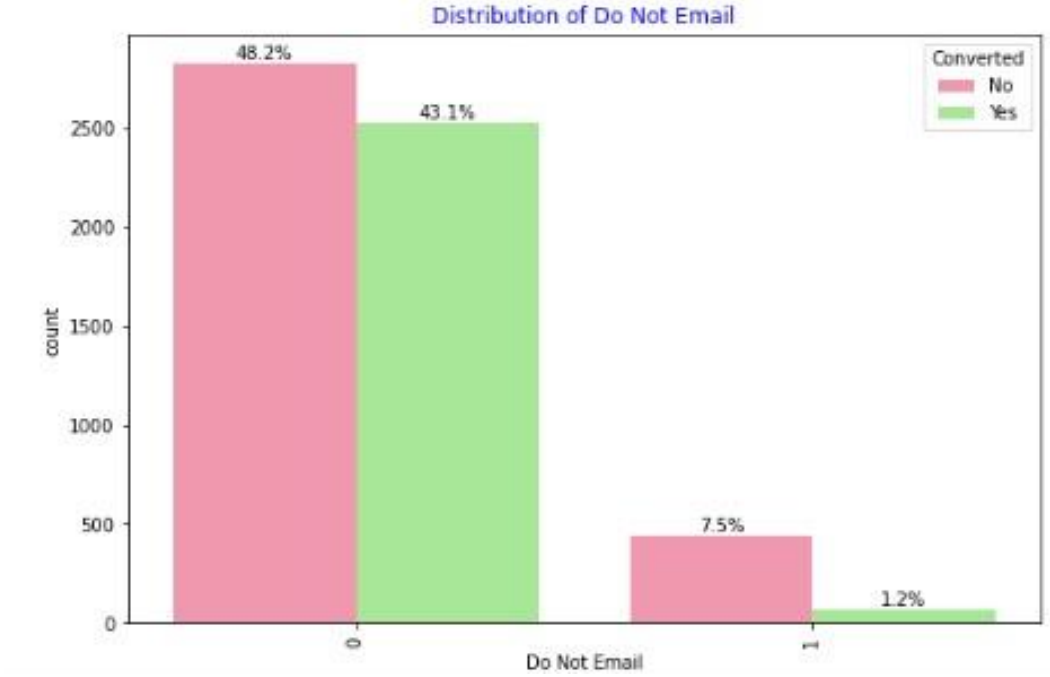
LEAD SOURCE VS CONVERTED

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



DO NOT EMAIL VS CONVERTED

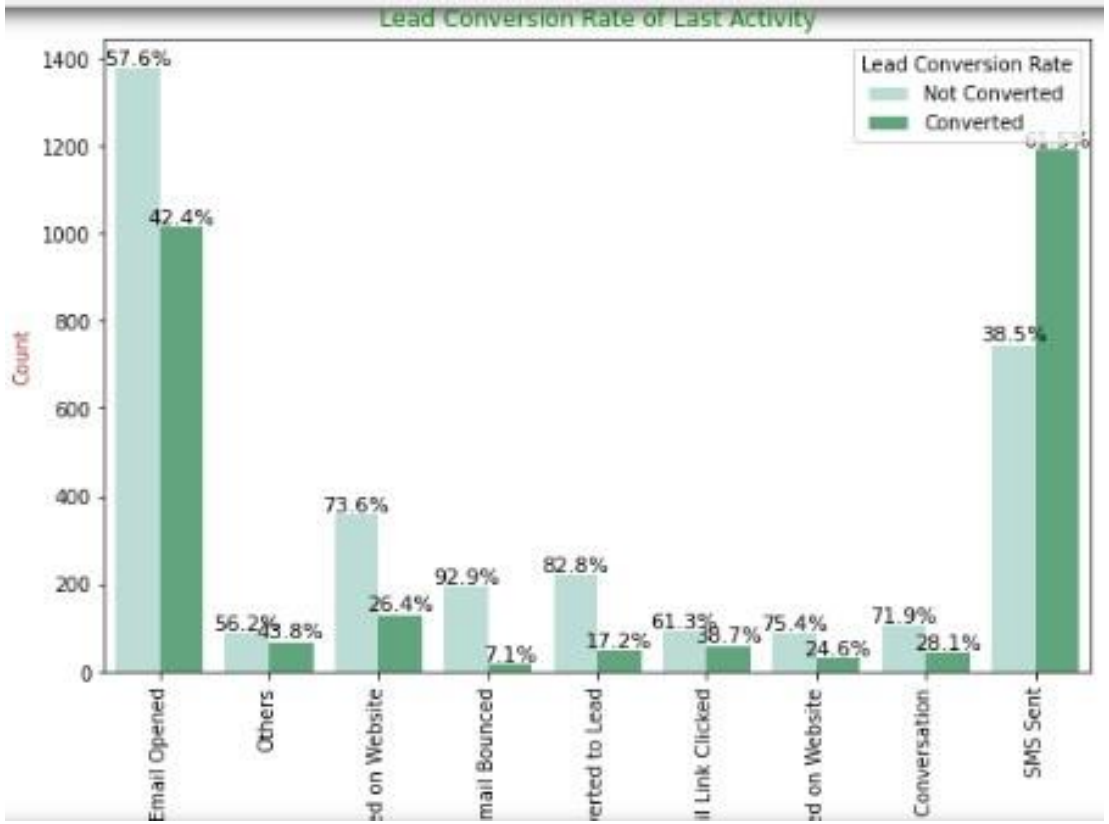
Customers preferred not to be informed through emails.



EXPLORATORY DATA ANALYSIS

LAST ACTIVITY VS CONVERTED

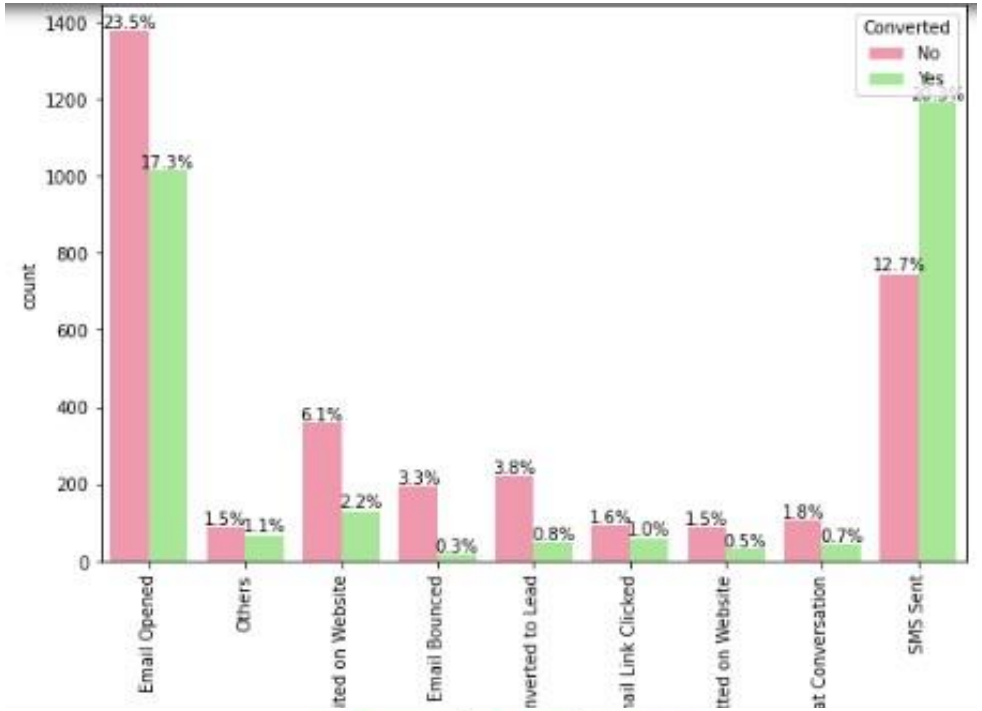
SMS has shown to be promising method for getting higher confirmed leads, emails also has high conversions.



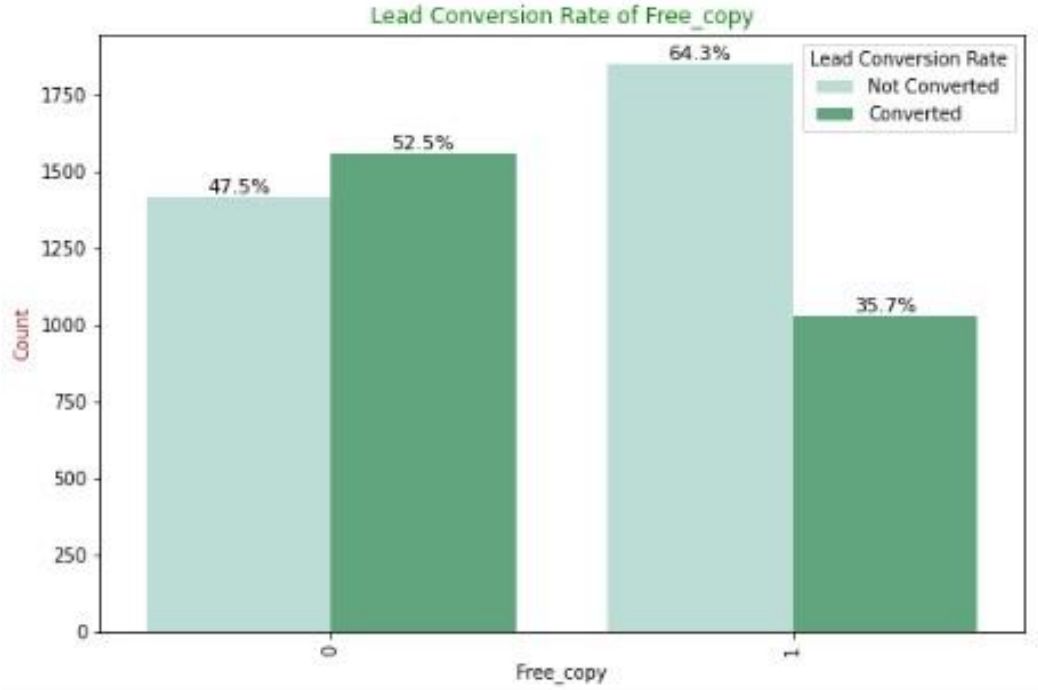
EXPLORATORY DATA ANALYSIS

LAST NOTABLE ACTIVITY VS CONVERTED

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



A FREE COPY OF MASTERING THE INTERVIEW VS CONVERTED



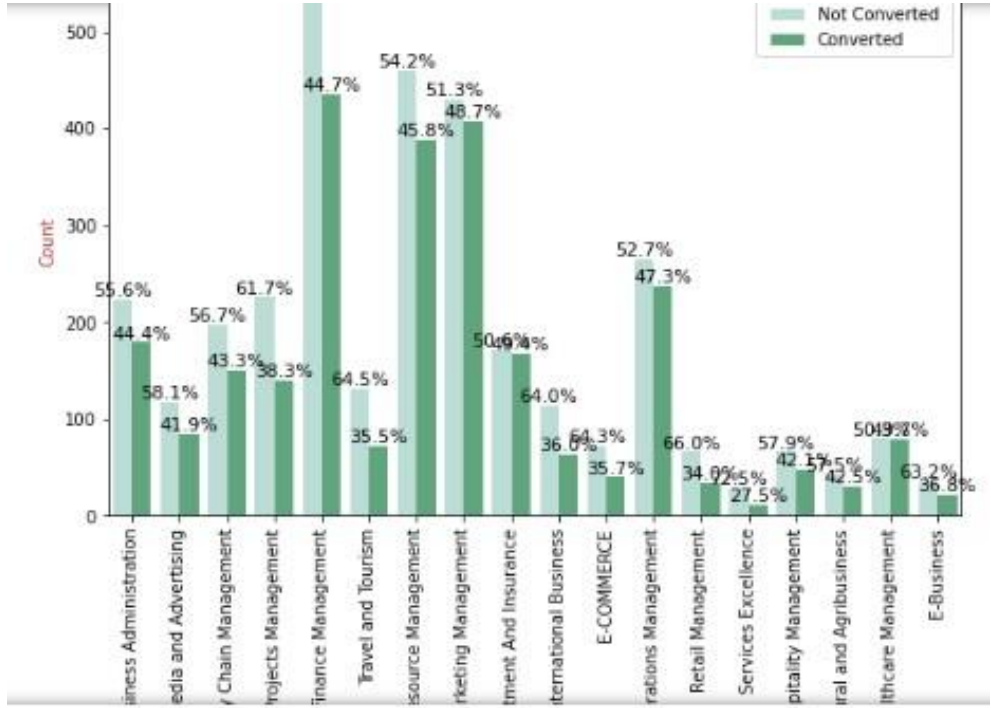
EXPLORATORY DATA ANALYSIS

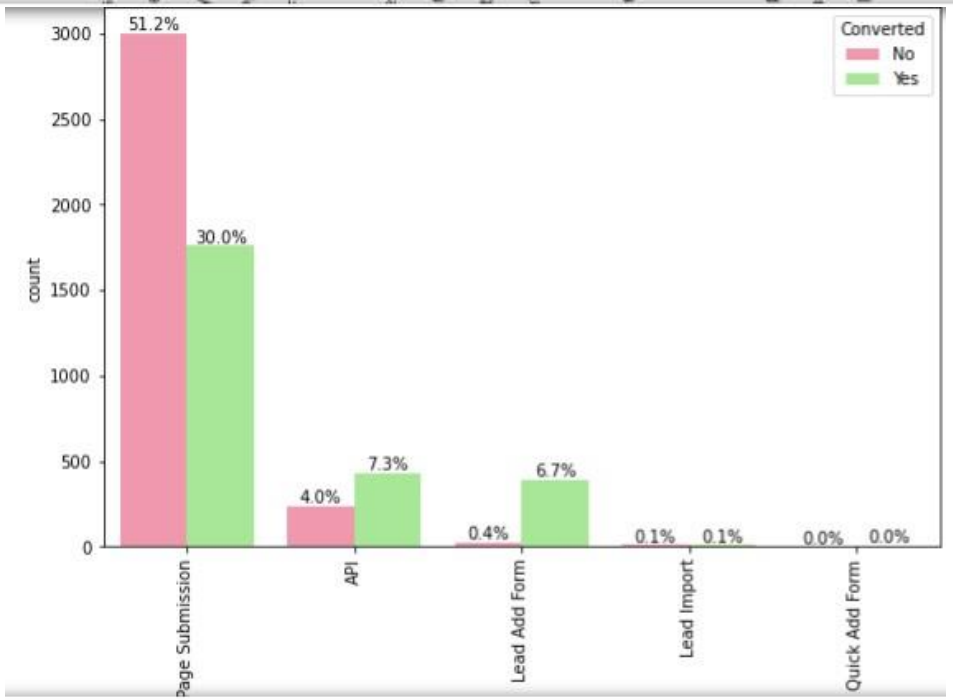
Leads prefer Less copies of interviews.

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.

LEAD ORIGIN VS CONVERTED

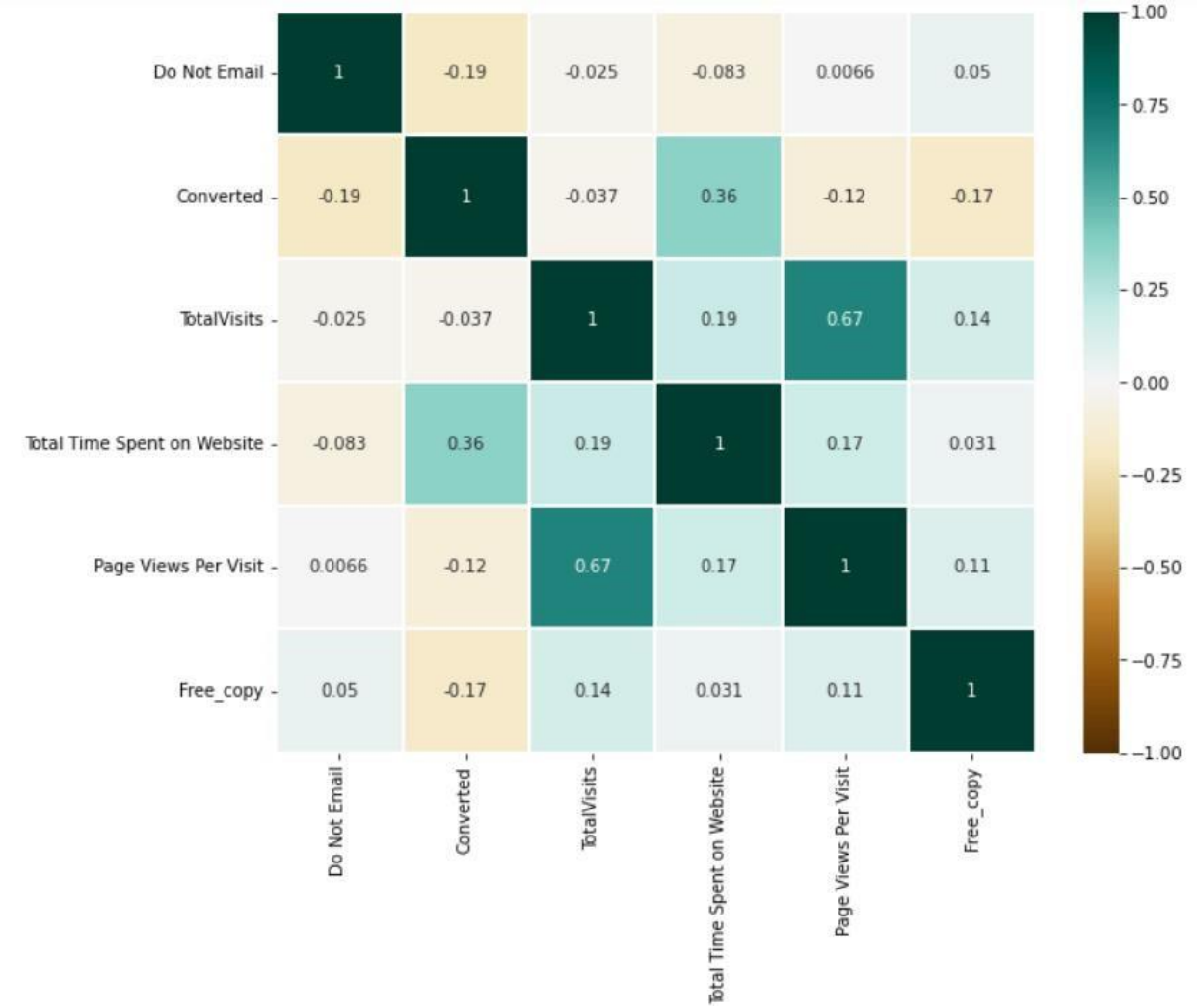


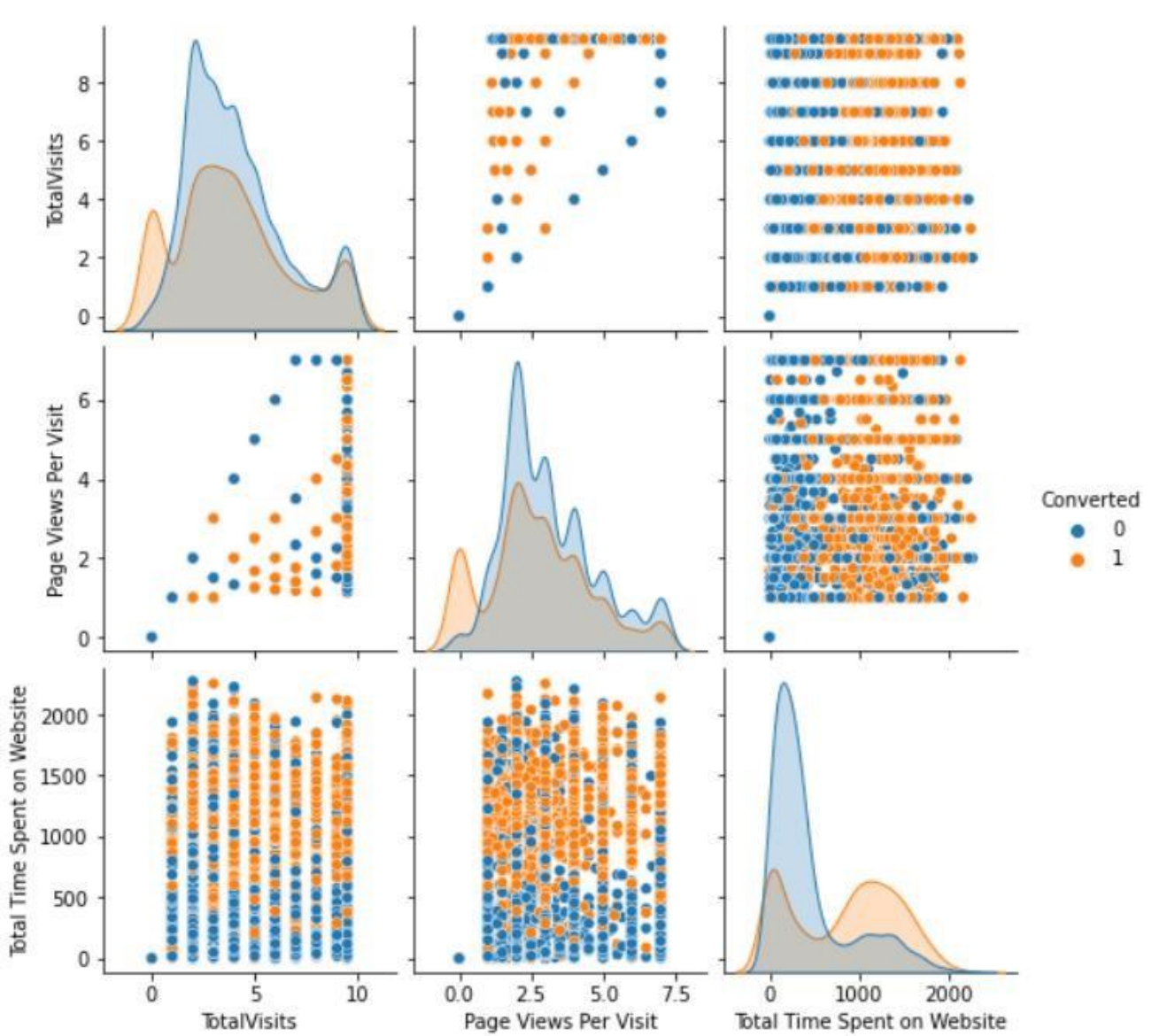


EXPLORATORY DATA ANALYSIS

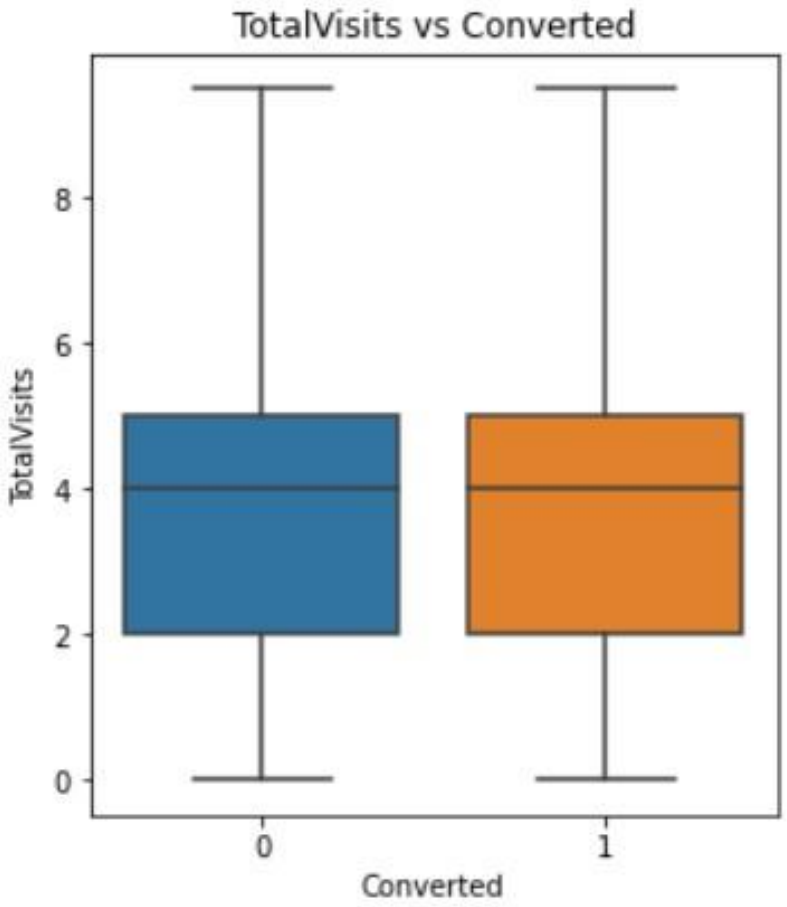
Landing page submissions has had high lead conversions.

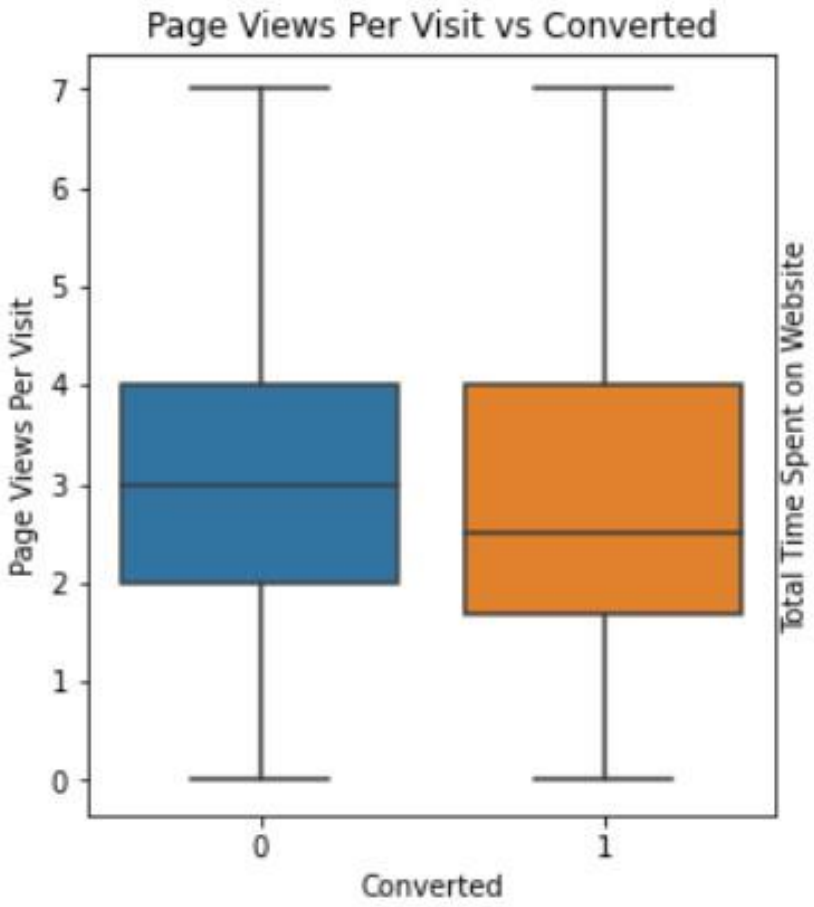
BIVARIATE ANALYSIS

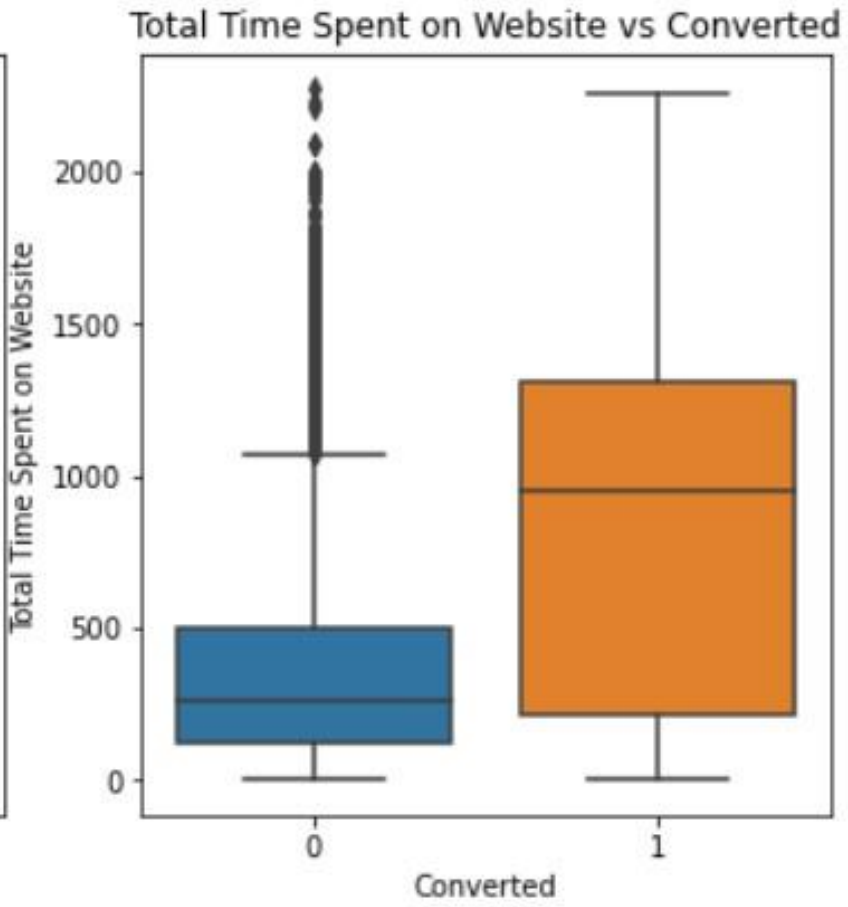


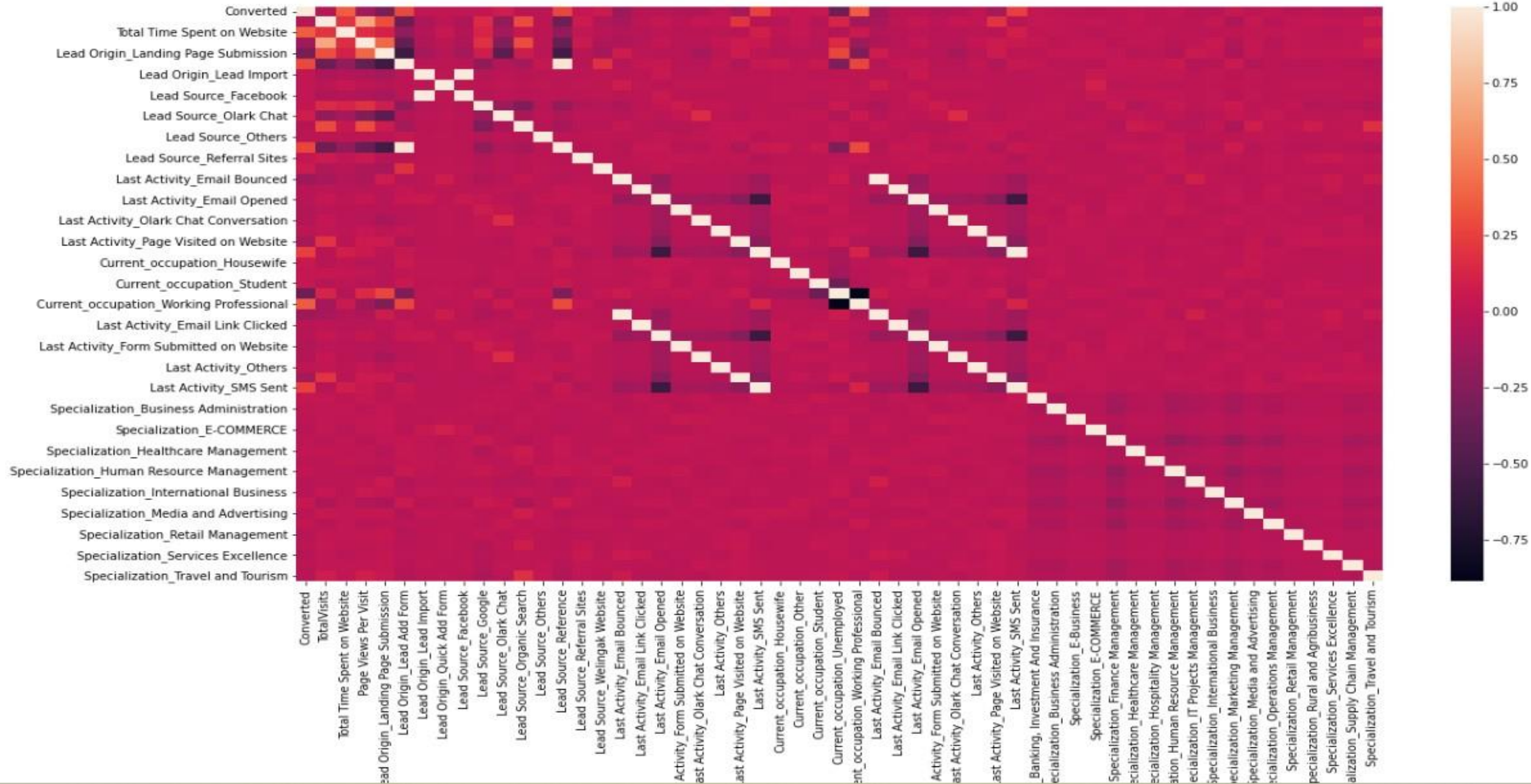


BIVARIATE ANALYSIS








Converted
Total Time Spent on Website
Lead Origin_Landing Page Submission
Lead Origin_Lead Import
Lead Source_Facebook
Lead Source_Olark Chat
Lead Source_Others
Lead Source_Referral Sites
Last Activity_Email Bounced
Last Activity_Email Opened
Last Activity_Olark Chat Conversation
Last Activity_Page Visited on Website
Current_occupation_Housewife
Current_occupation_Student
Current_occupation_Working Professional
Last Activity_Email Link Clicked
Last Activity_Form Submitted on Website
Last Activity_Others
Last Activity_SMS Sent
Specialization_Business Administration
Specialization_E-COMMERCE
Specialization_Healthcare Management
Specialization_Human Resource Management
Specialization_International Business
Specialization_Media and Advertising
Specialization_Retail Management
Specialization_Services Excellence
Specialization_Travel and Tourism
TotalVisits
Page Views Per Visit
Lead Origin_Lead Add Form
Lead Origin_Quick Add Form
Lead Source_Google
Lead Source_Organic Search
Lead Source_Reference
Lead Source_Welingak Website
Current_occupation_Unemployed
Current_occupation_Other
Banking, Investment And Insurance
Specialization_Business Administration
Specialization_E-Business
Specialization_E-COMMERCE
Specialization_Finance Management
Specialization_Healthcare Management
Specialization_Hospitality Management
Specialization_Human Resource Management
Specialization_IT Projects Management
Specialization_International Business
Specialization_Marketing Management
Specialization_Media and Advertising
Specialization_Operations Management
Specialization_Retail Management
Specialization_Rural and Agribusiness
Specialization_Services Excellence
Specialization_Supply Chain Management
Specialization_Travel and Tourism
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

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

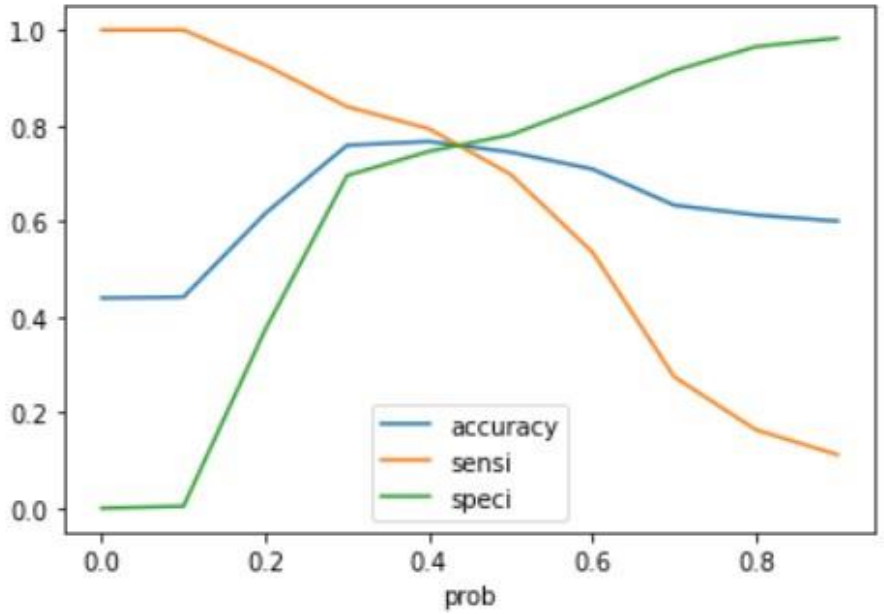
78.2% Sensitivity

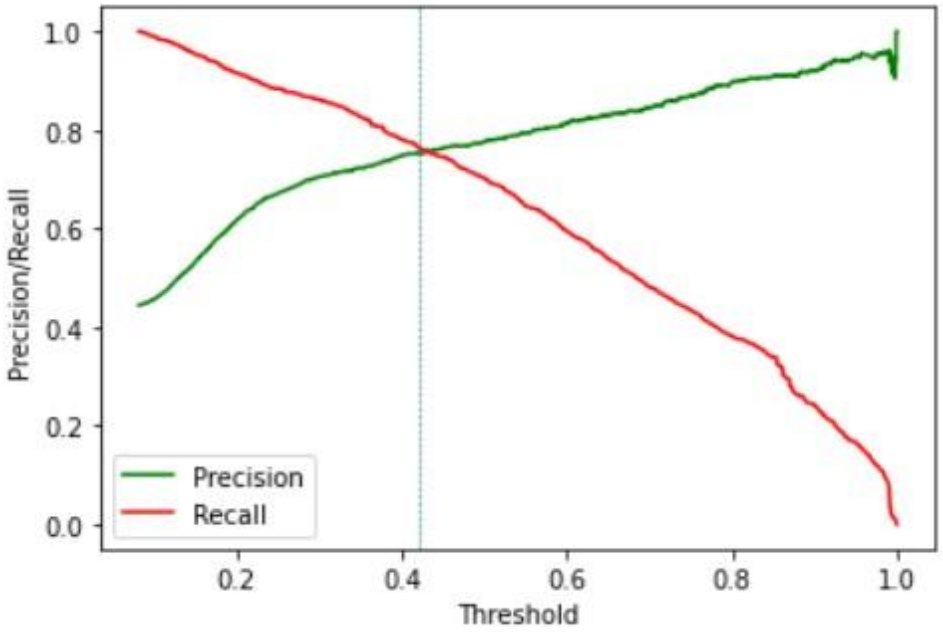
79.4% Specificity

PRECISION AND RECALL

74.82% Precision

78.19% Recall



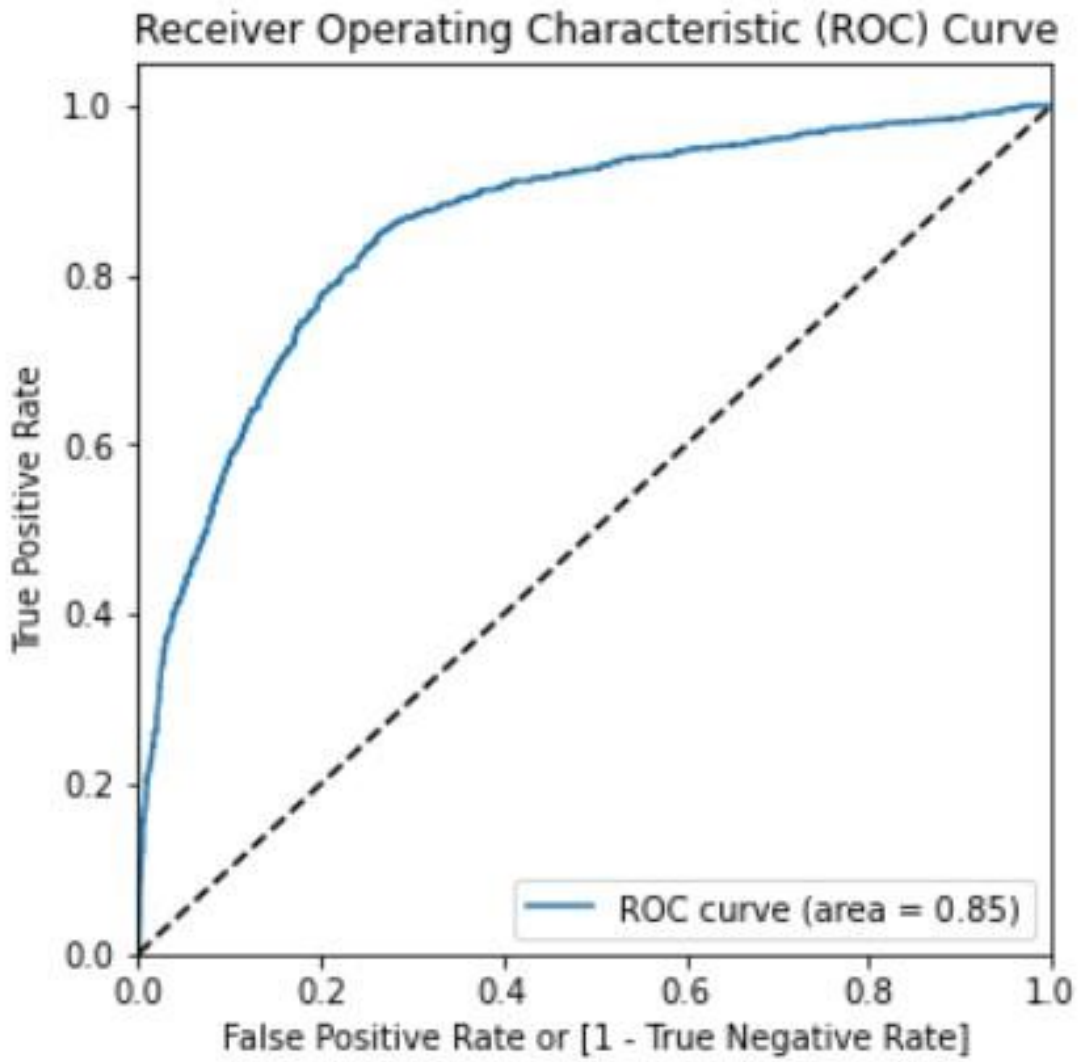


ROC CURVE

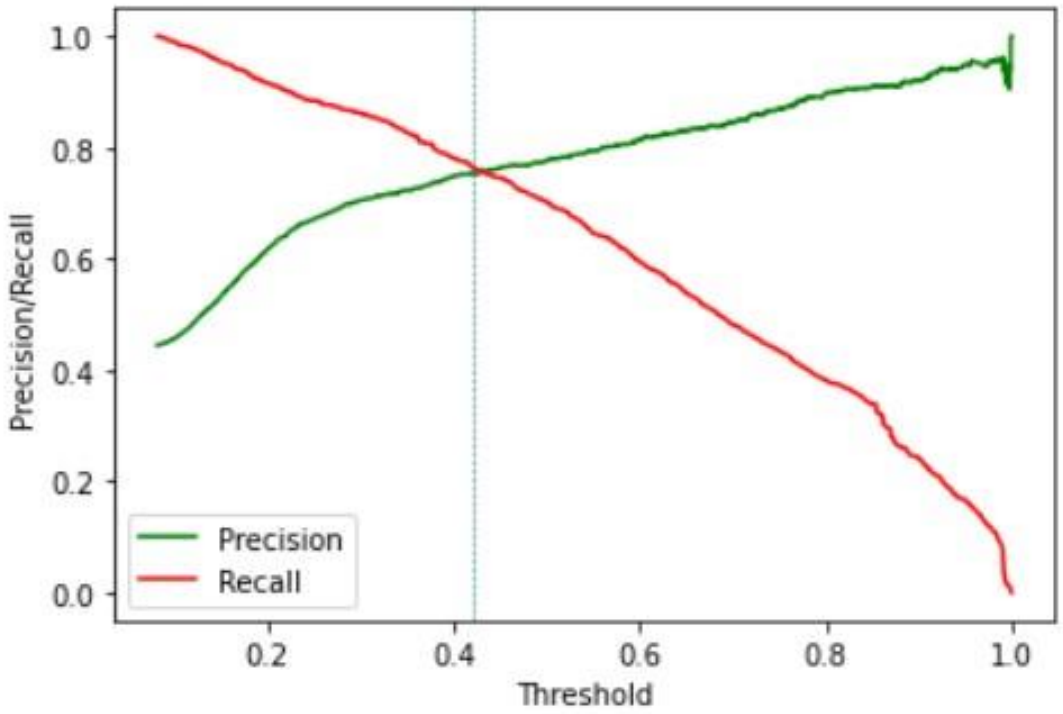
An area under the ROC curve of 0.85 is indeed indicative of a good predictive model.

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



Metrics for Logistic Regression Model:

Accuracy: 0.4050056882821388

Precision: 0.4231692677070828 5

Recall: 0.892405063291392

Specificity: 0.007231404958677686

F1 Score: 0.5741042345276874

CONCLUSION

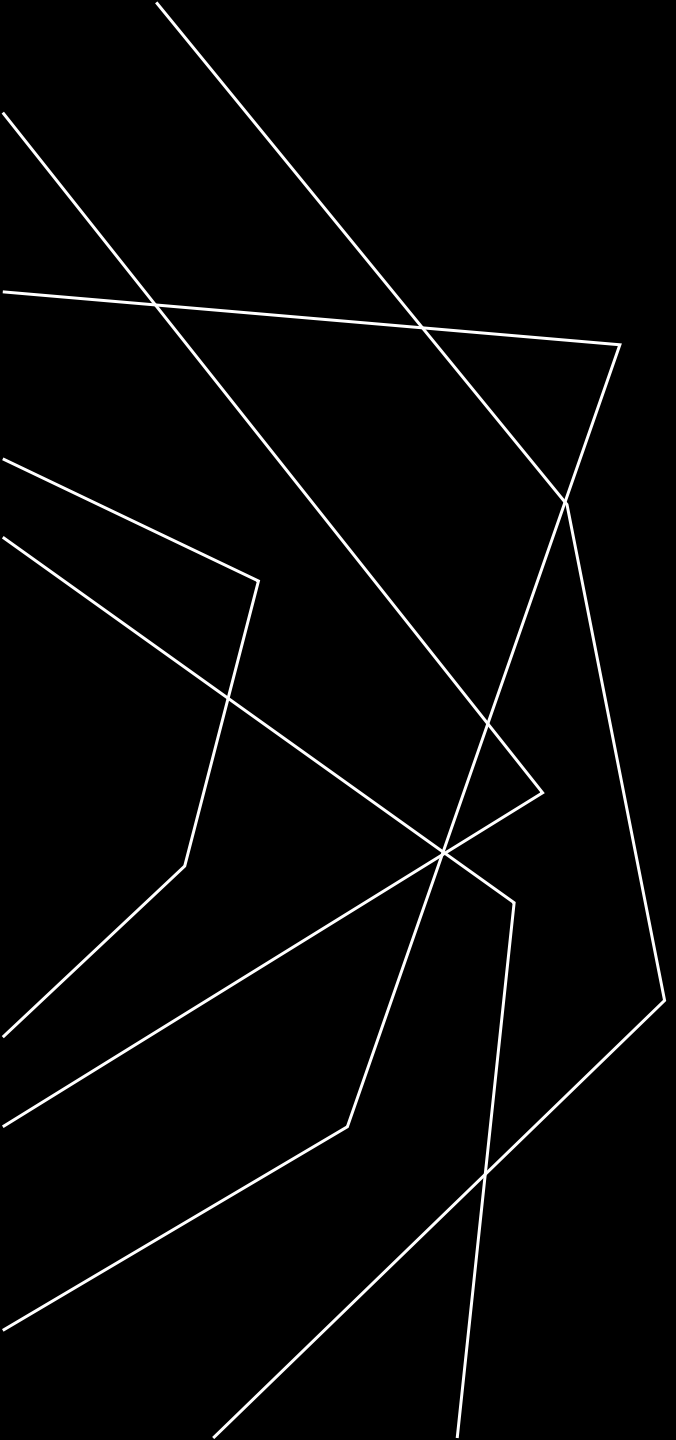
EDA:

- 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

Logistic Regression Model:

- 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



THANK YOU

Ayesha Taranum

Aishwarya Behera

Sumit Bansal