Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € € $
Converted Leads

LEAD SCORING CASE STUDY

(LOGISTIC REGRESSION)

Contents

1. Business Problem Statement
2. Technical Approach
3. Exploratory Data Analysis - Numerical Variables
4. Exploratory Data Analysis - Categorical variables
5. Model Summary
6. Final Evaluation Metrics and Confusion Matrix
7. Top 5 features contributing to Lead Conversion

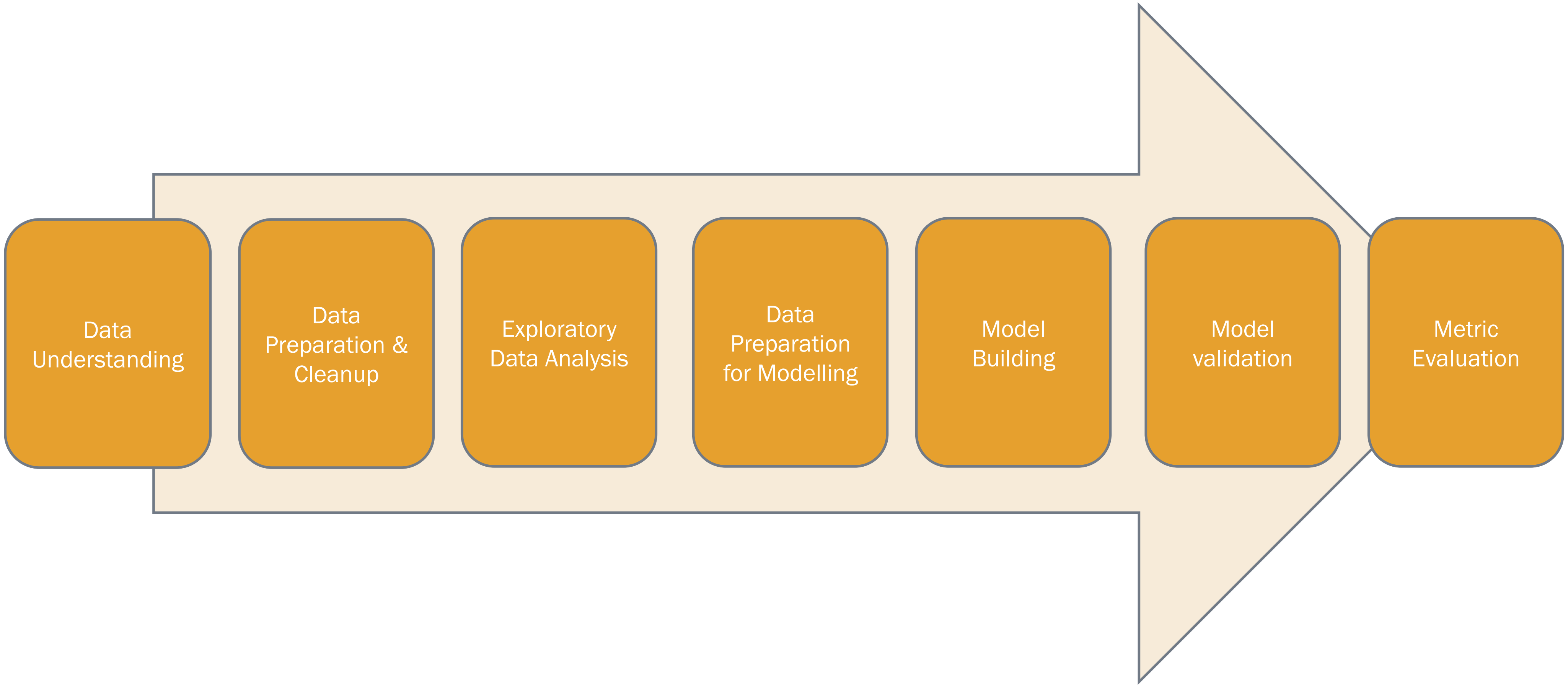
Business Problem Statement

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.

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%.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

Technical Approach



Data Understanding

Data Preparation & Cleanup

Exploratory Data Analysis

Data Preparation for Modelling

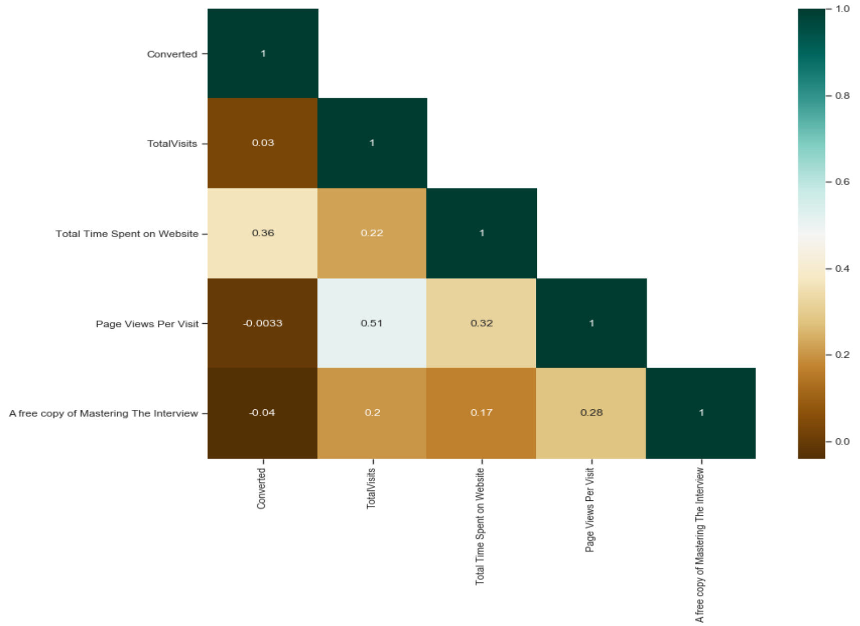
Model Building

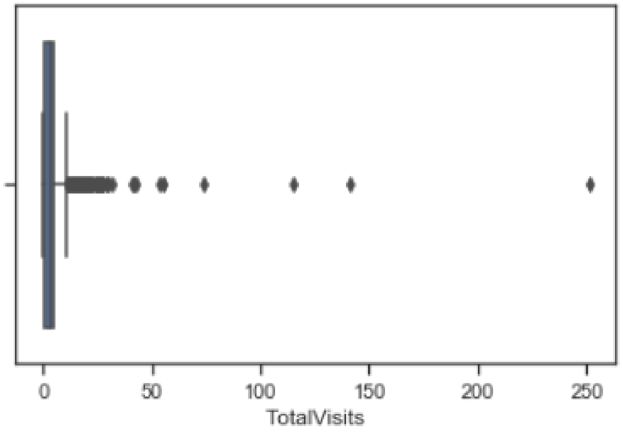
Model validation

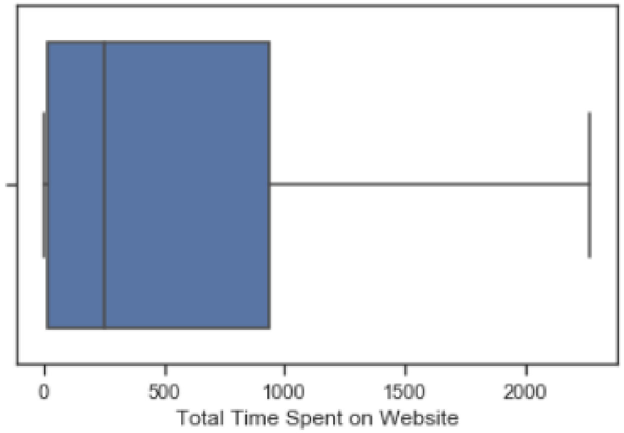
Metric Evaluation

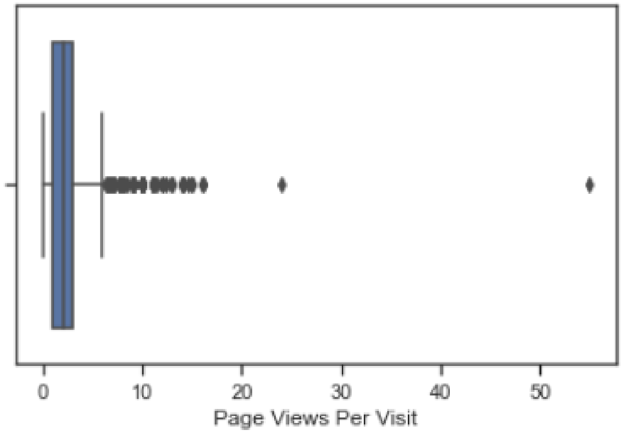
Exploratory Data Analysis - Numerical Variables

- Missing values **imputed with mean** for the columns.
- A strong **correlation** between 'TotalVisits' & 'Page View Per Visit' as indicated by Heat Map was handled later.
- Outliers were checked as shown below and handled by **Capping at 25th and 75th percentile** for lower and upper bounds respectively.



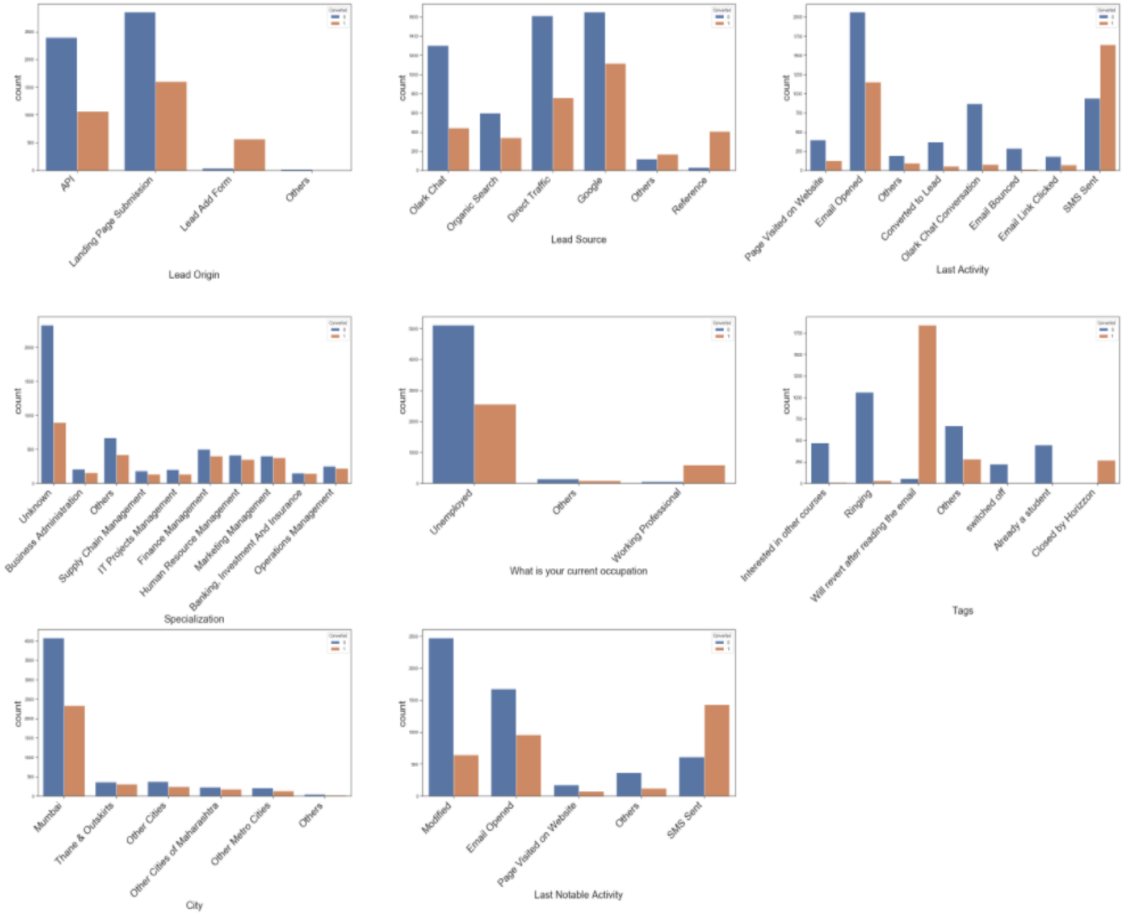






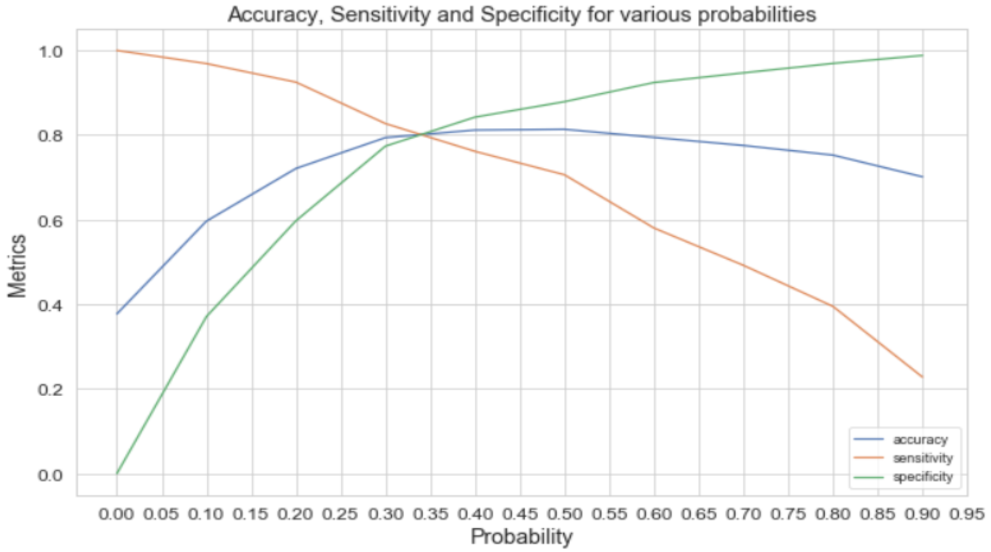
Exploratory Data Analysis – Categorical variables

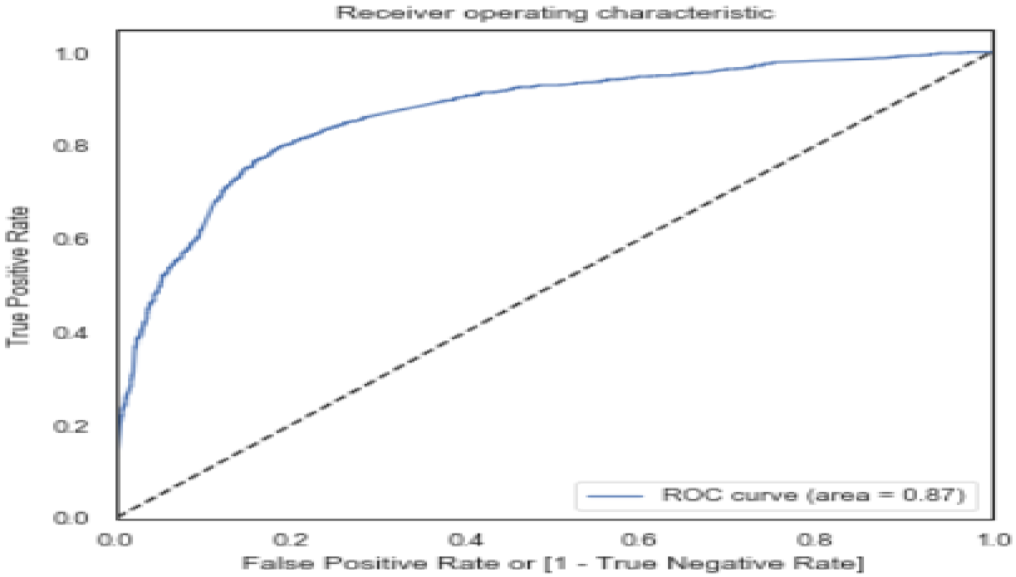
- Select values replaced by ‘**Nan**’ values
- Missing value check performed. If less than 40% missing values, then **imputation by mode** of that categorical variable was performed.
- For some categorical variables having low frequency in the column**, bucketing was performed** and valued were clubbed and categorized as **‘Others’ category**.
- Univariate and Bivariate analysis performed to understand the impact on Lead Conversion.



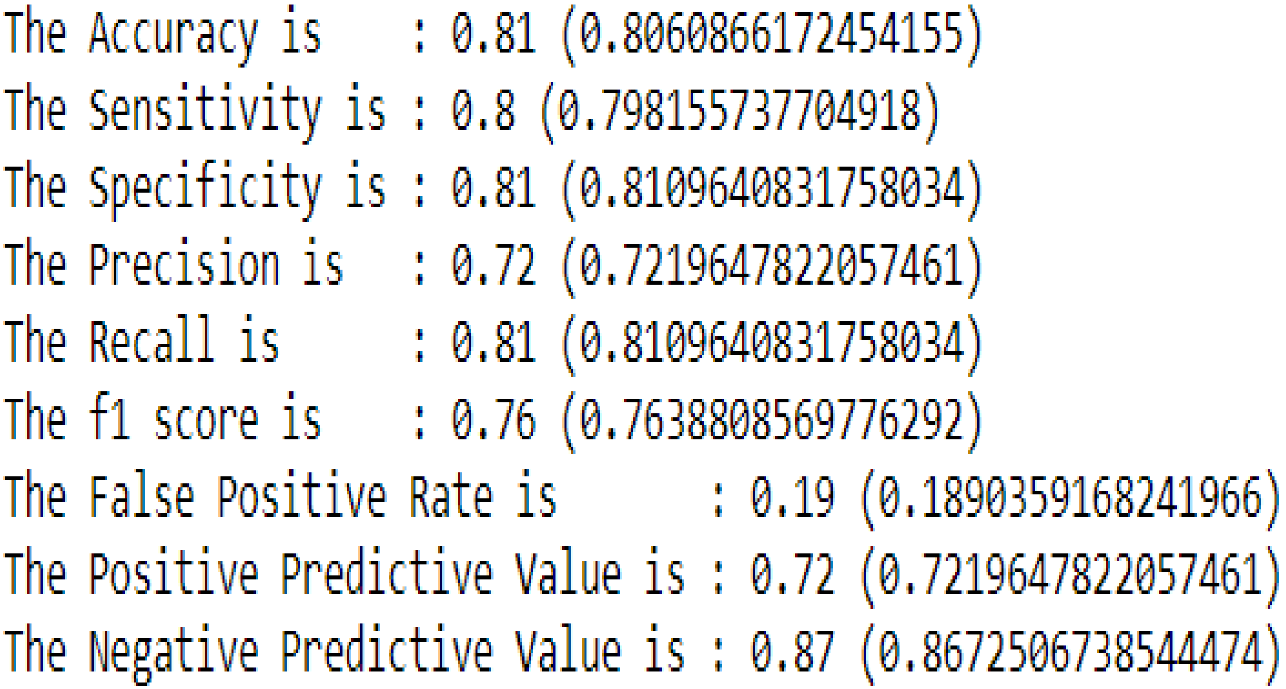
Model Summary

- We first applied **Recursive feature elimination** over the train data to extract **top 20** features.
- We then used a combination of **p-value and VIF** to eliminate unwanted features.
- Sensitivity / Specificity curve for the different cut-offs was plotted to identify the optimal cut-off value of **0.35.**

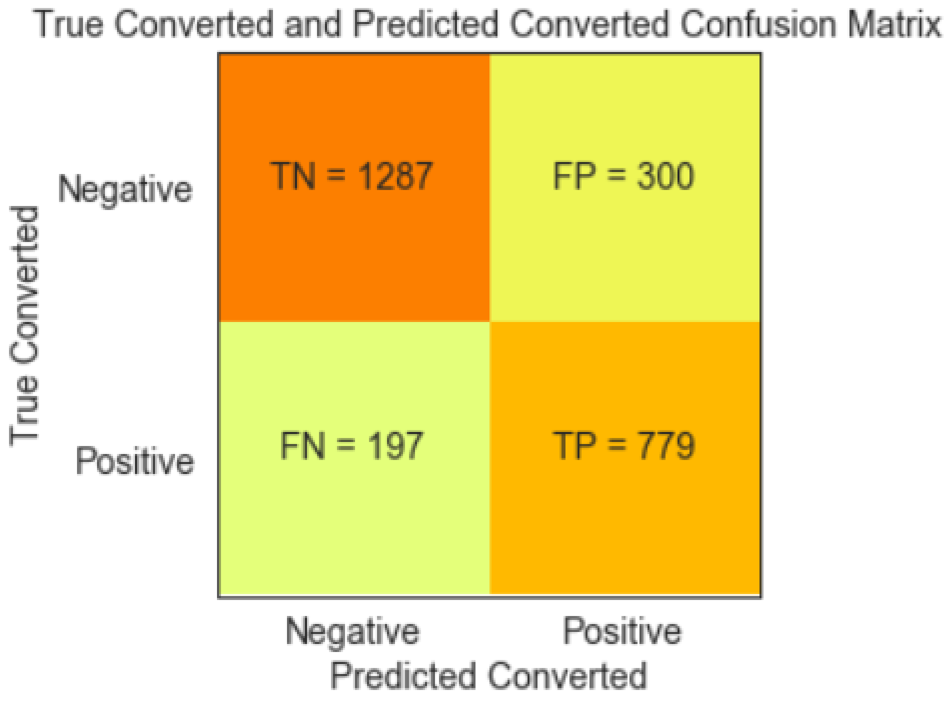




Final Evaluation Metrics and Confusion Matrix



```
The Accuracy is    : 0.81 (0.8060866172454155)
The Sensitivity is : 0.8 (0.798155737704918)
The Specificity is : 0.81 (0.8109640831758034)
The Precision is   : 0.72 (0.7219647822057461)
The Recall is      : 0.81 (0.8109640831758034)
The f1 score is    : 0.76 (0.7638808569776292)
The False Positive Rate is       : 0.19 (0.1890359168241966)
The Positive Predictive Value is : 0.72 (0.7219647822057461)
The Negative Predictive Value is : 0.87 (0.867250673854474)
```



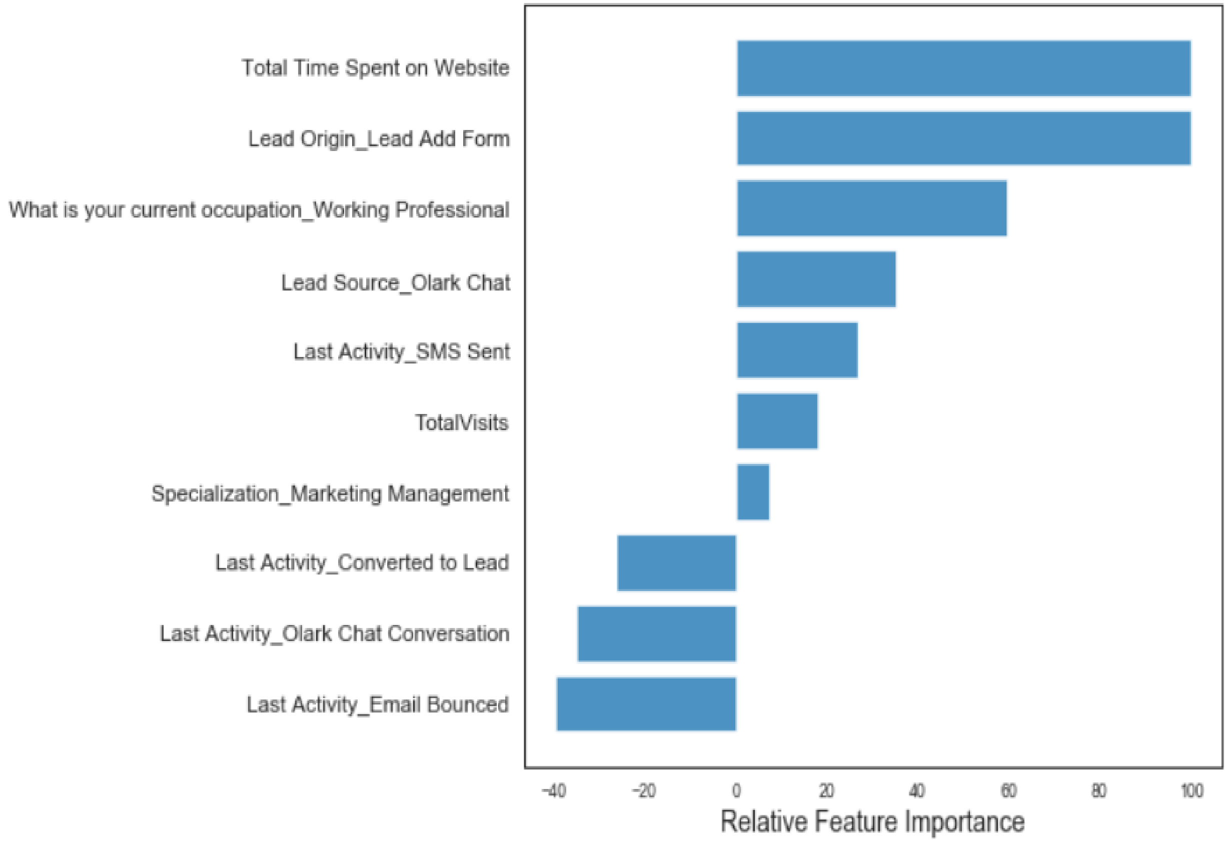
True Converted and Predicted Converted Confusion Matrix

True Converted	Negative	Positive
Negative	TN = 1287	FP = 300
Positive	FN = 197	TP = 779

Top 5 features contributing to Lead Conversion

X-Education has a better chance of converting a potential lead when:

- **The total time spent on the Website is high**
- **Current Occupation is working professional**
- **When the Lead origin was Lead Add form**
- **Number of Total Visits were high**
- **Lead Source Olark Chat**


Total Time Spent on Website
Lead Origin_Lead Add Form
What is your current occupation_Working Professional
Lead Source_Olark Chat
Last Activity_SMS Sent
TotalVisits
Specialization_Marketing Management
Last Activity_Converted to Lead
Last Activity_Olark Chat Conversation
Last Activity_Email Bounced
Relative Feature Importance