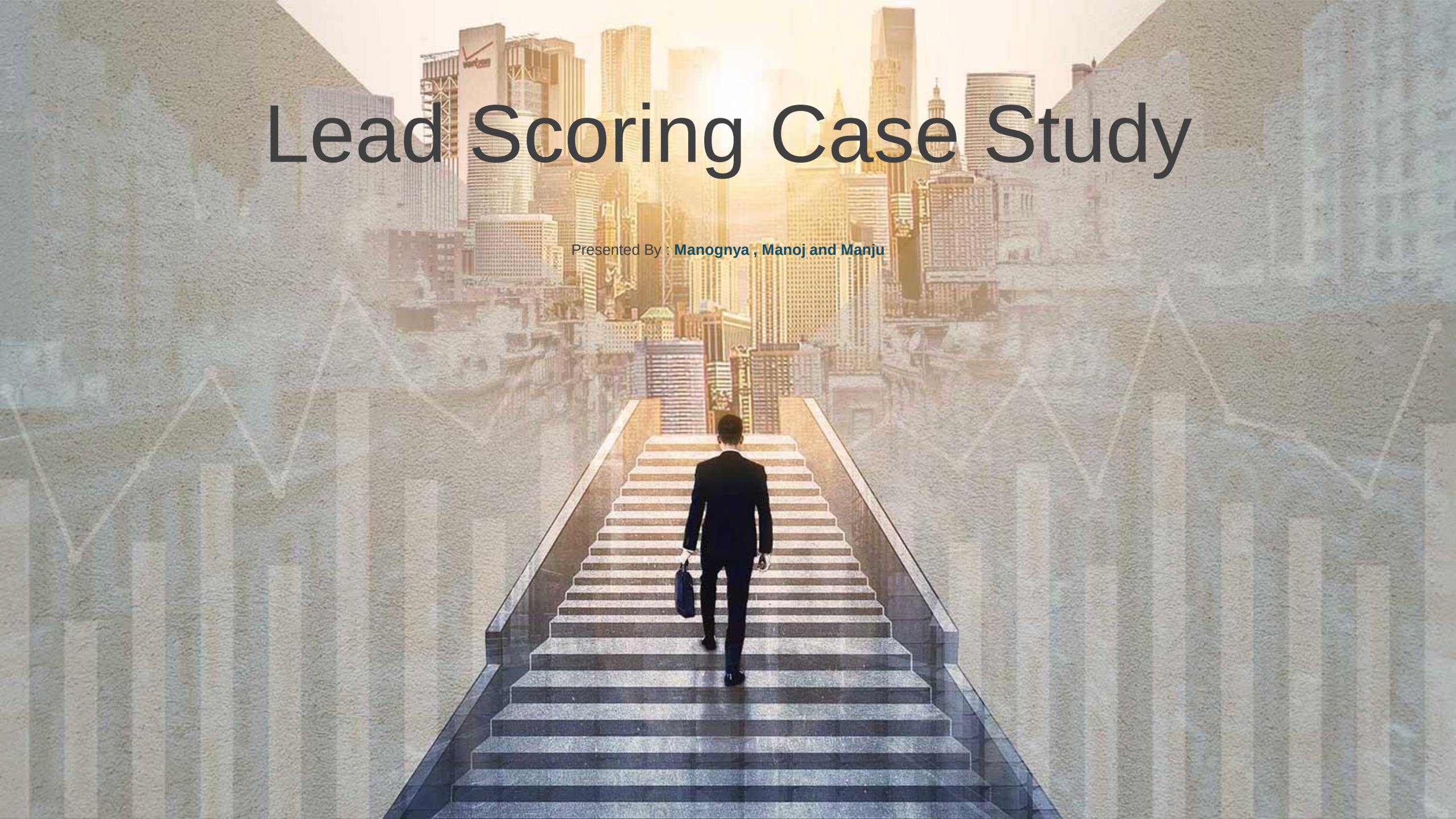
Lead Scoring Case Study

Presented By : **Manognya , Manoj and Manju**



Agenda

01 Problem Statement

02 Problem Approach

03 EDA

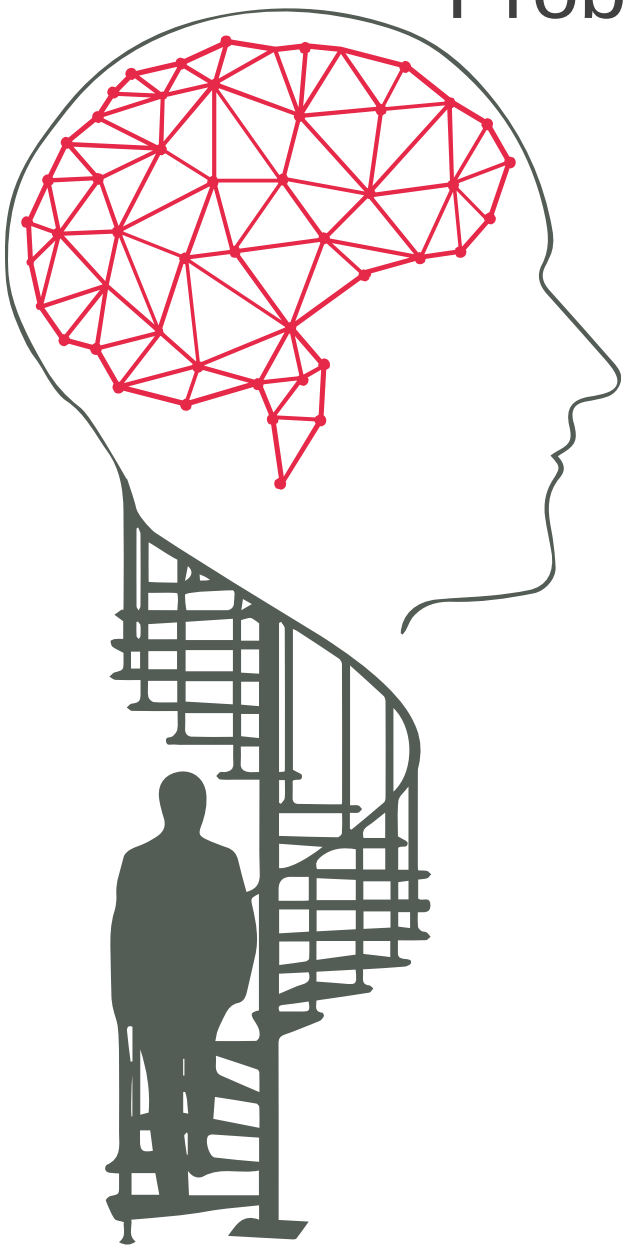
04 Correlation

05 Model Evaluation

06 Observation

07 Conclusion

Problem Statement & Business Goal



Problem Statement:

X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.

Strategy

Feature Scaling

Exploratory Data Analysis

Splitting Data into test and Train

Clean and prepare the data

Building a logistic Regression model and calculating Lead Score.

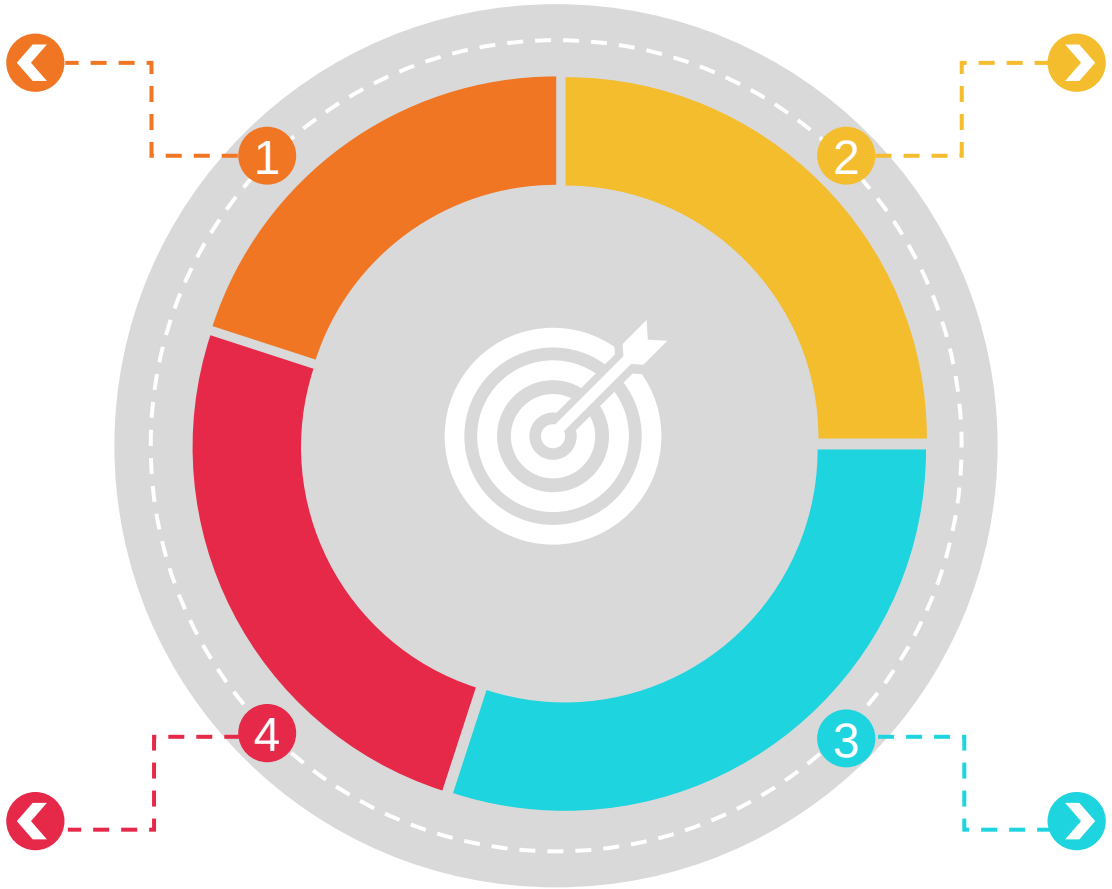
Source the data for analysis

Evaluating the model by using different metrics and applying the best model in Test data based on the Sensitivity and Specificity Metrics

Problem solving methodology

Data Sourcing, Cleaning and Preparation

- ✓ Read the Data from the Source and convert it into a clean format
- ✓ Remove duplicate data
- ✓ Exploratory Data Analysis
- ✓ Feature Standardization



Feature Scaling and Splitting Train and Test Sets

- ✓ Feature Scaling of Numeric data
- ✓ Splitting data into train and test set

Result

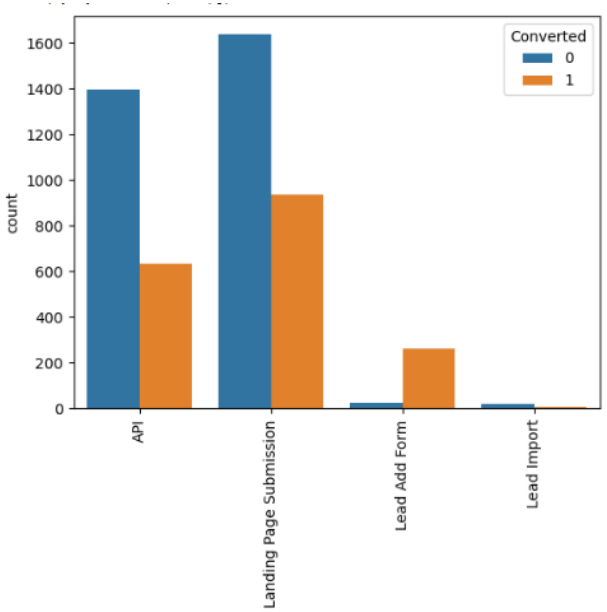
- ✓ Determine the lead score and check if the target final predictions amount to an 80% conversion rate.
- ✓ Evaluate the final prediction on the test set using the cut-off threshold from sensitivity and specificity metrics

Model Building

- ✓ Feature Selection using RFE
- ✓ Determine the optimal model using Logistic Regression
- ✓ Calculate various metrics like accuracy, sensitivity, specificity, precision and recall.
- ✓ Evaluate the model

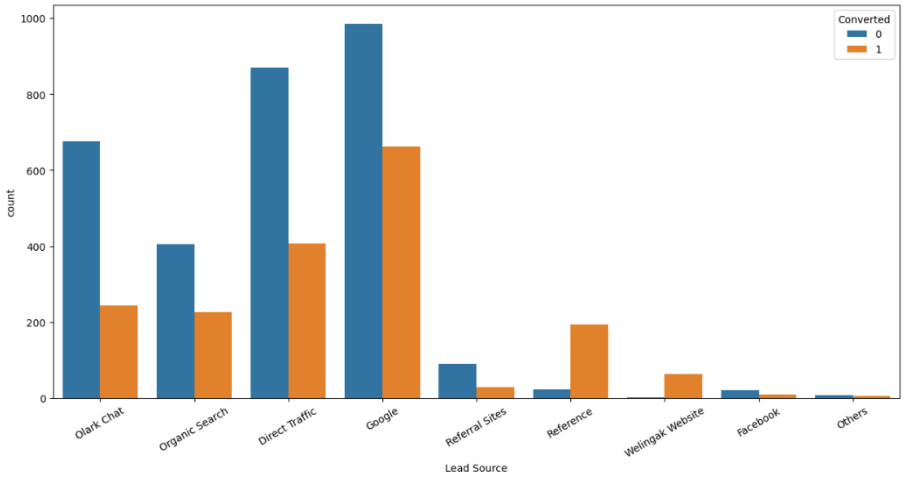
Exploratory Data Analysis

Univariate Analysis



Observations

- API and Landing Page Submission have a 30-35% conversion rate but the count of leads originating from them is considerable.
- Lead Add Form has a more than 90% conversion rate, but the count of leads is not very high.
- Lead imports are very low in the count.
- *To improve the overall lead conversion rate, the focus will be on improving lead conversion of API and Landing Page Submission origin and generating more leads from the Lead Add Form.*

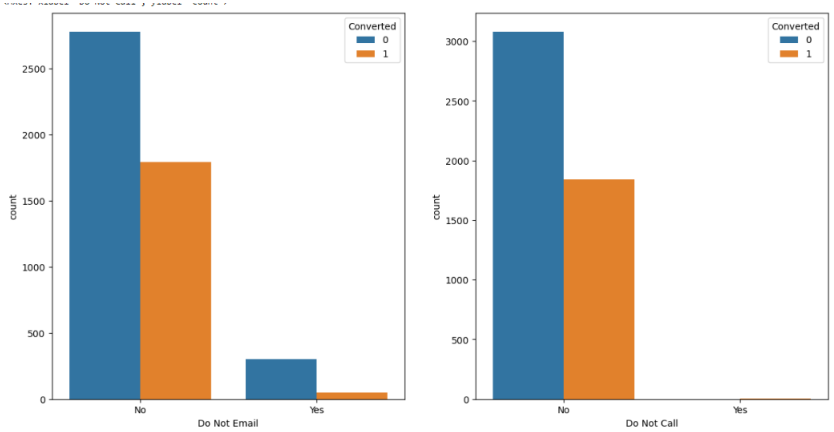


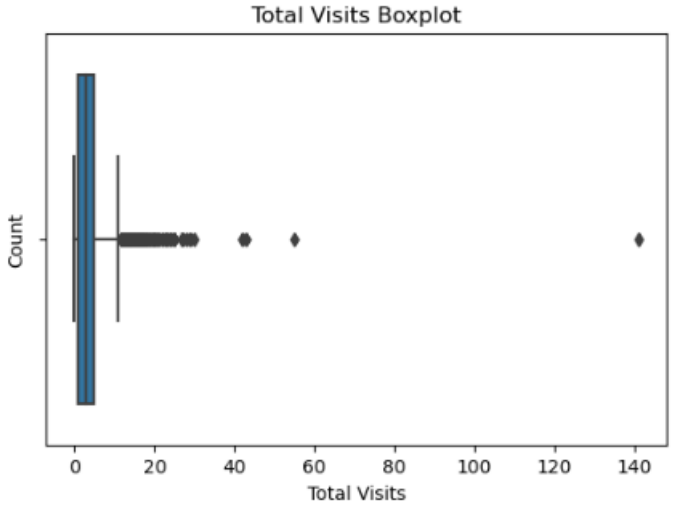
Observations

Google and Direct traffic generate a maximum number of leads.
Conversion Rate of reference leads and leads through welingak website is high.
To improve the overall lead conversion rate, the focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and Google leads and generating more leads from reference and the welingak website

Exploratory Data Analysis

Univariate Analysis

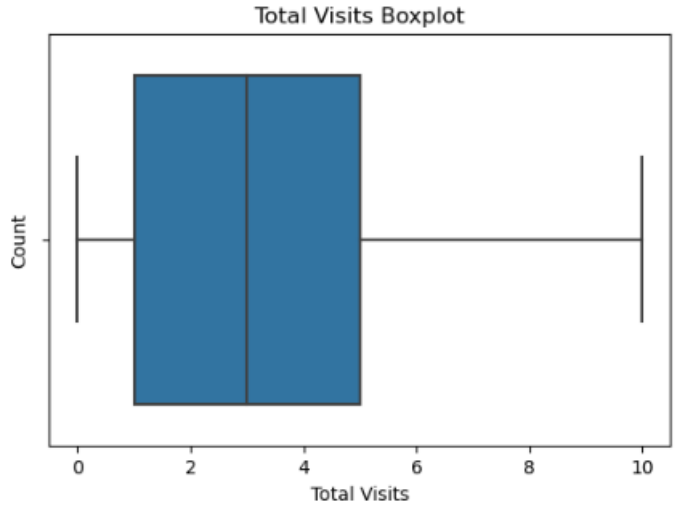


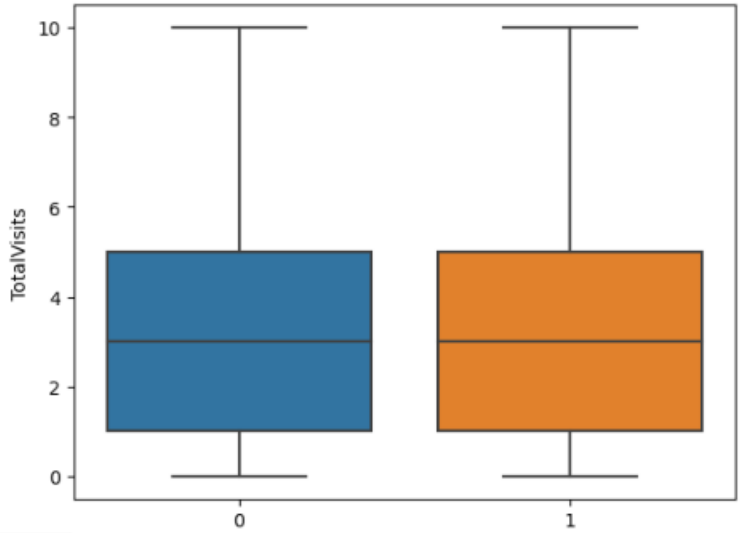


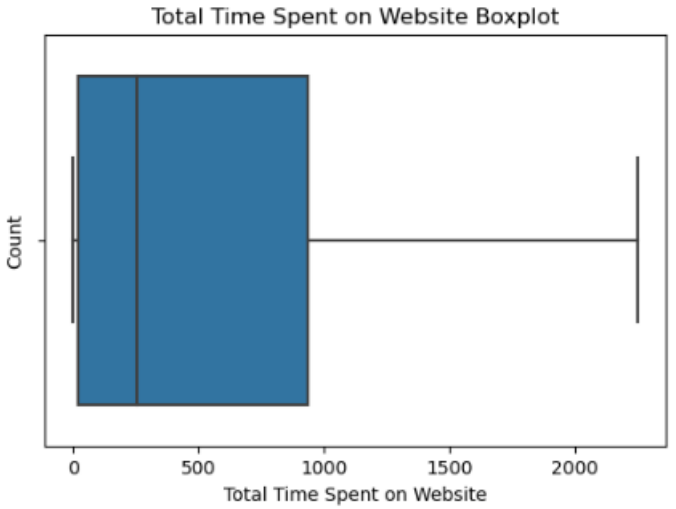
Observations

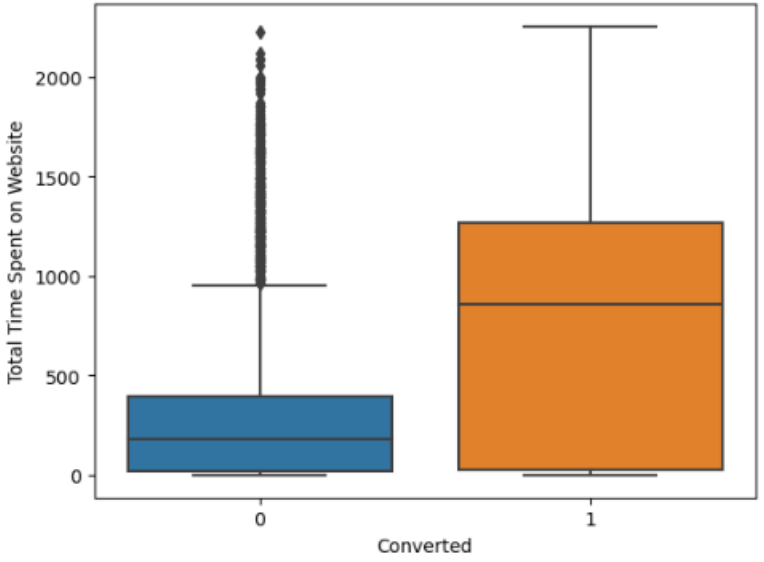
As we can see there are several outliers in the data. We will cap the outliers to 95% value for analysis.

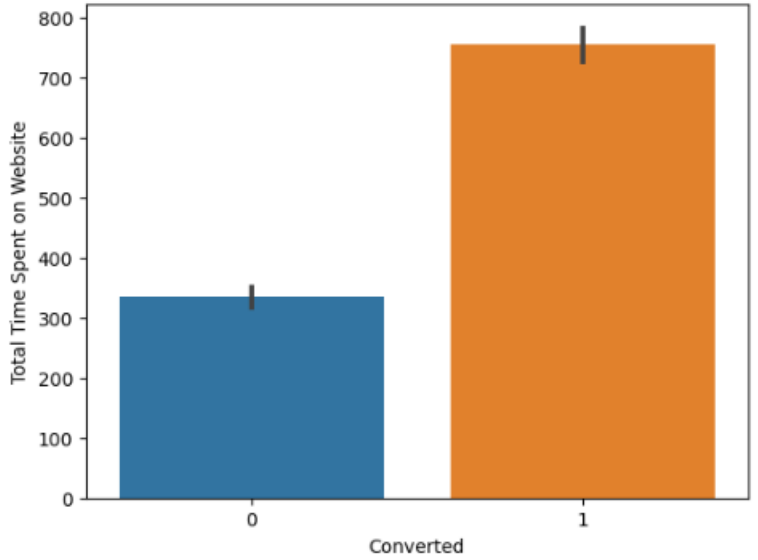
Exploratory Data Analysis











Observations

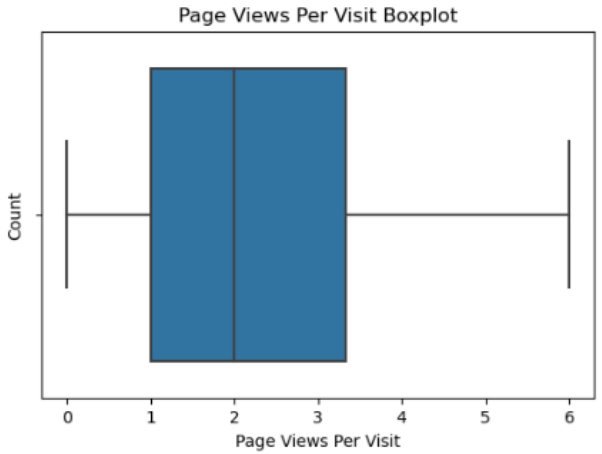
- Leads that are spending more time on the weblise will be more likely to be converted.
- The website should be more engaging to make leads spend additional time.

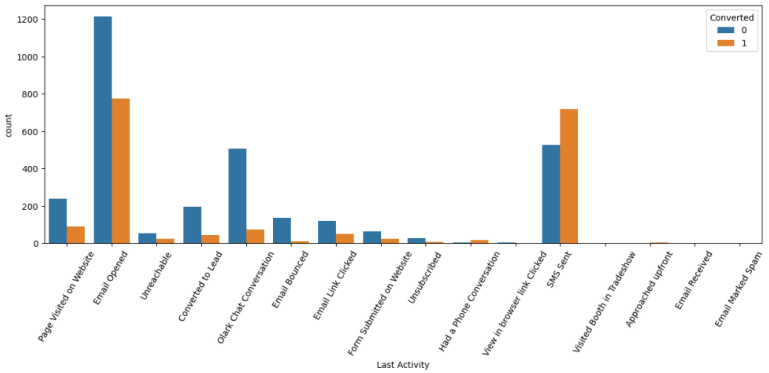
Exploratory Data Analysis

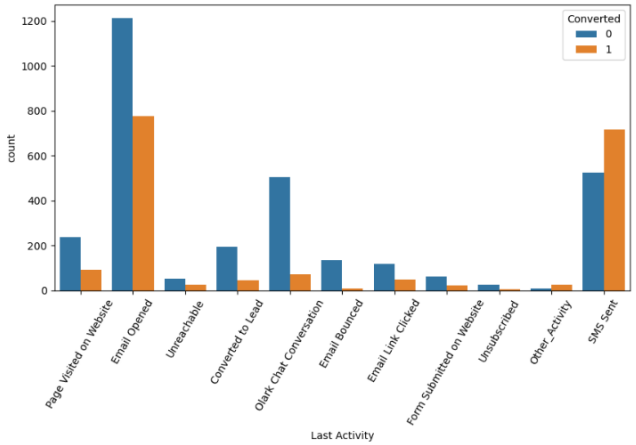
- **Observations**

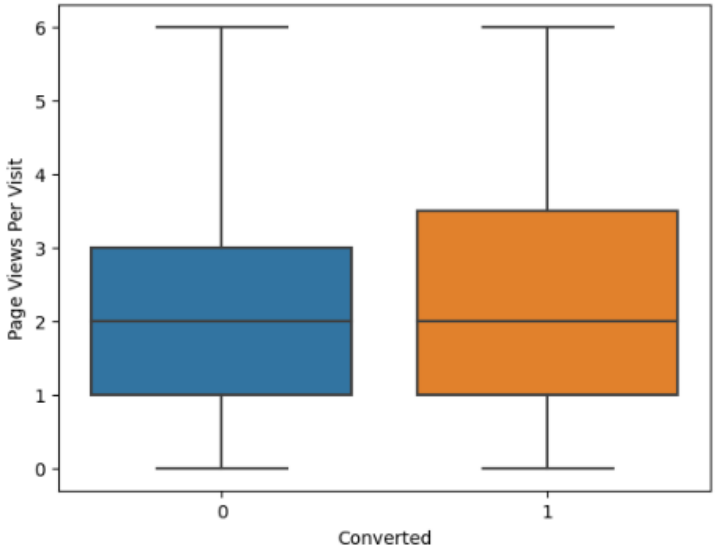
- Most of the leads have their Email opened as their last activity.

- Conversion rate for leads with last activity as SMS Sent is almost 60%.









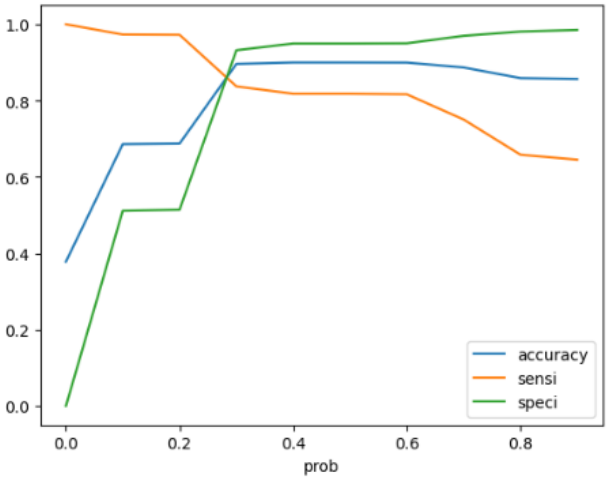
Model Evaluation

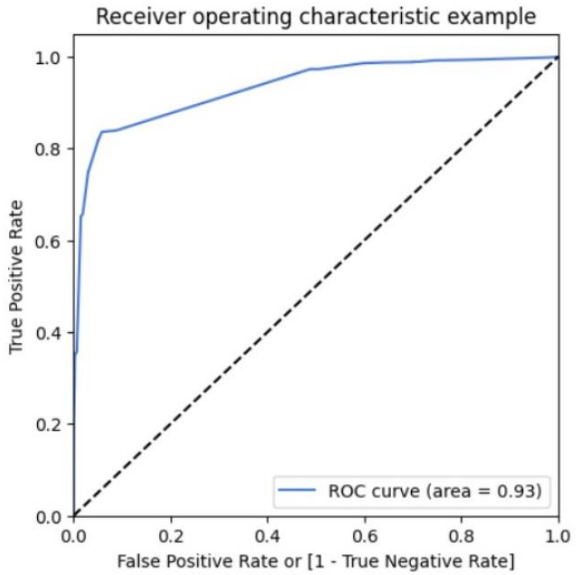
	Features	VIF
12	Last Notable Activity_SMS Sent	3.03
9	Tags_switched off	2.66
8	Tags_Will revert after reading the email	1.50
2	What is your current occupation_Working Profes...	1.25
5	Tags_Interested in other courses	1.18
10	Lead Quality_Not Sure	1.12
3	Tags_Busy	1.09
7	Tags_Ringing	1.09
1	Last Activity_Email Bounced	1.07
6	Tags_Lost to EINS	1.07
4	Tags_Closed by Horizzon	1.05
0	Lead Source_Welingak Website	1.04
11	Lead Quality_Worst	1.03

The VIF of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables.

Model Evaluation

Sensitivity and Specificity on Train Data Set





1	2
Accuracy: 89.98 %	Accuracy: 89.37%
Sensitivity: 81.79 %	Sensitivity: 85.52%
Specificity: 94.95 %	Specificity: 91.57%
Train Data	Test Data

Observations :

From the curve above, 0.25 is the optimum point to take as a cutoff probability.

Conclusion

Suggestions for increasing conversion rates:

Very likely to convert into leads:

1. More than 68 for the lead score.
2. Over twelve hours were spent on the website overall.
3. Welingak Website and Reference as the primary source.
4. Conversion rate for leads with last activity as SMS Sent is almost 60%.

Extremely unlikely to become qualified leads

1. Clients who chose the "Do not email" option.
2. Lead with a score below fifteen.
3. Less than five hours were spent on the website overall.
4. Google, Organic Search, Referral Sites, and Direct Traffic as Lead Sources.
5. The consumers' most recent actions include completing an online form, clicking an email link, visiting a page on the website, having an Olark chat session, and having an email bounce.