Lead Scoring Case Study

Presented By:

- Anand Krishnan
- Indri Lonka
- Navneet Kumar

BUSINESS PROBLEM

- An education company named X Education sells online courses to industry professionals.
- 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.

BUSINESS OBJECTIVE

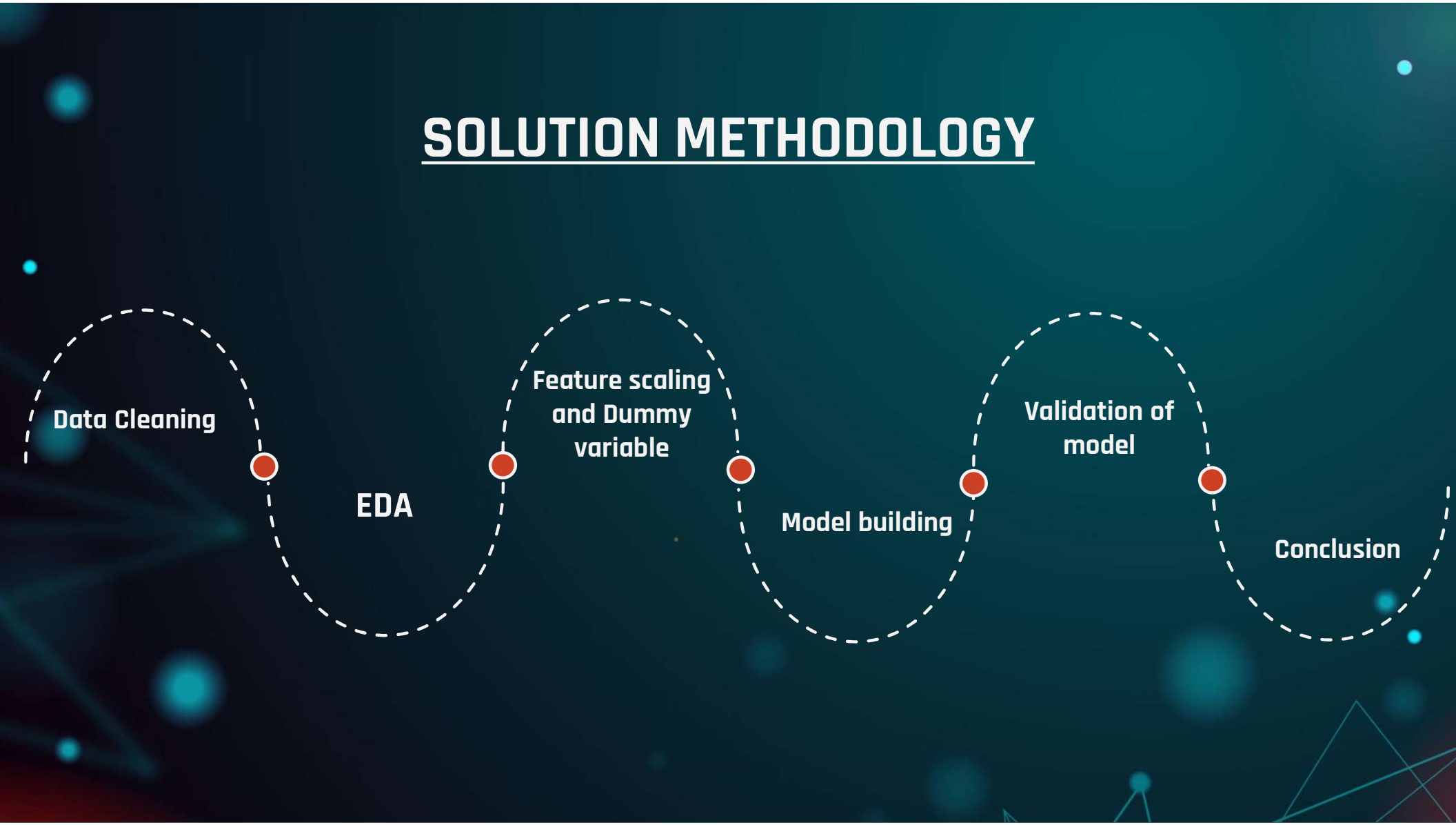
- X Education wants to know the most promising leads

- The business objective is to build a logistic regression model to identify the potential leads and achieve the target lead conversion rate to be around 80%

DATA KNOWLEDGE

- Data Set Used: **“Leads.csv”**
- Total number of customer: **9240**
- Total number of features: **37**
- Model Used: **Logistic Regression**
- After initial study, we can see that there are many variables that affect conversion rate.
- The target columns in our dataset: **“Converted”**
- We need to reduce the features to maximize the conversion rate
- Current conversion rate = **38.54%**

SOLUTION METHODOLOGY



Solution Methodology

- Data Cleaning:
  - Check and handling duplicate data
  - Checking and handling NA and missing values.
  - Drop column, if missing values are more.
  - Check and handle outliers in data.

- EDA:
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Scaling and Dummy variables and encoding of the data.

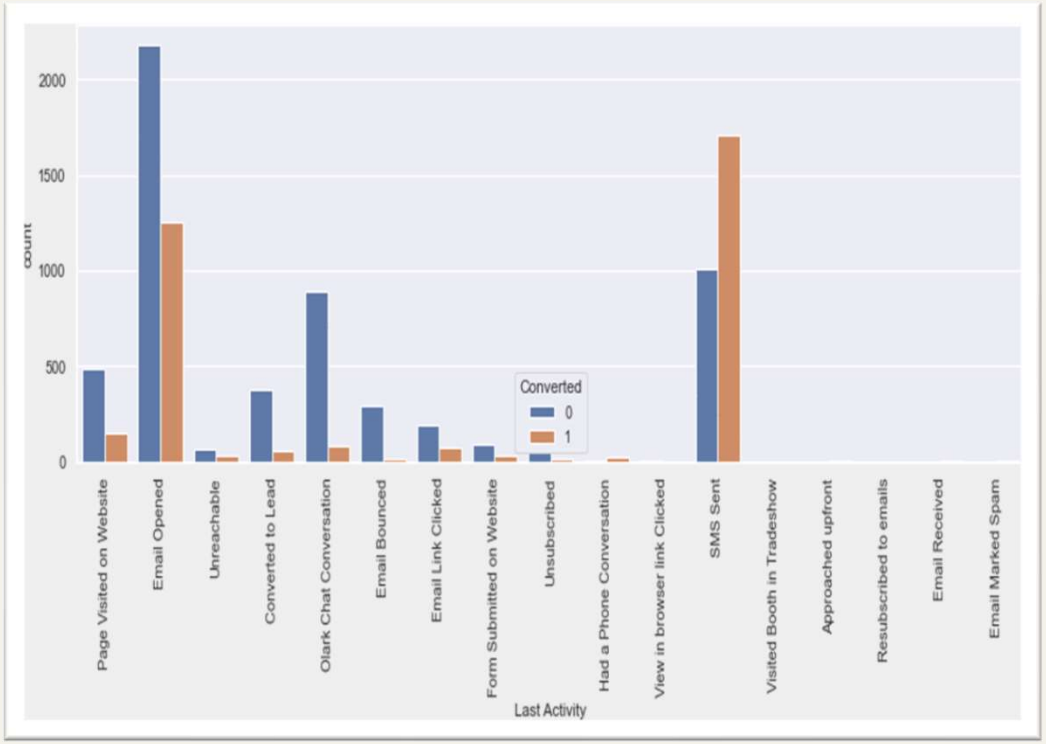
- Model Building: Use of Logistic Regression for the Model Buildings

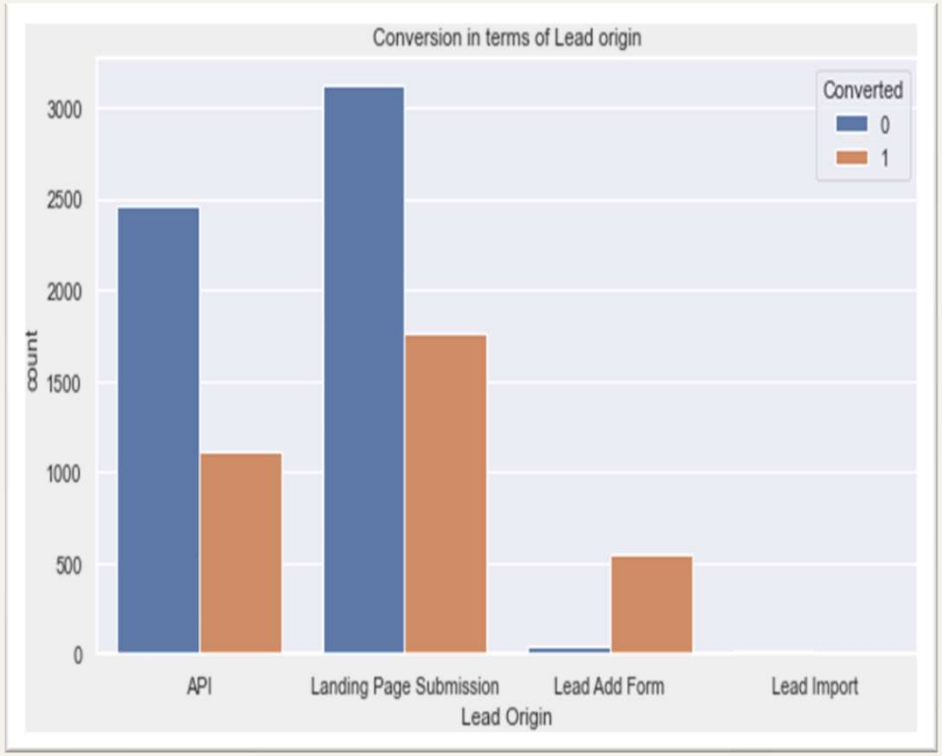
- Validation of Model : Validate the model

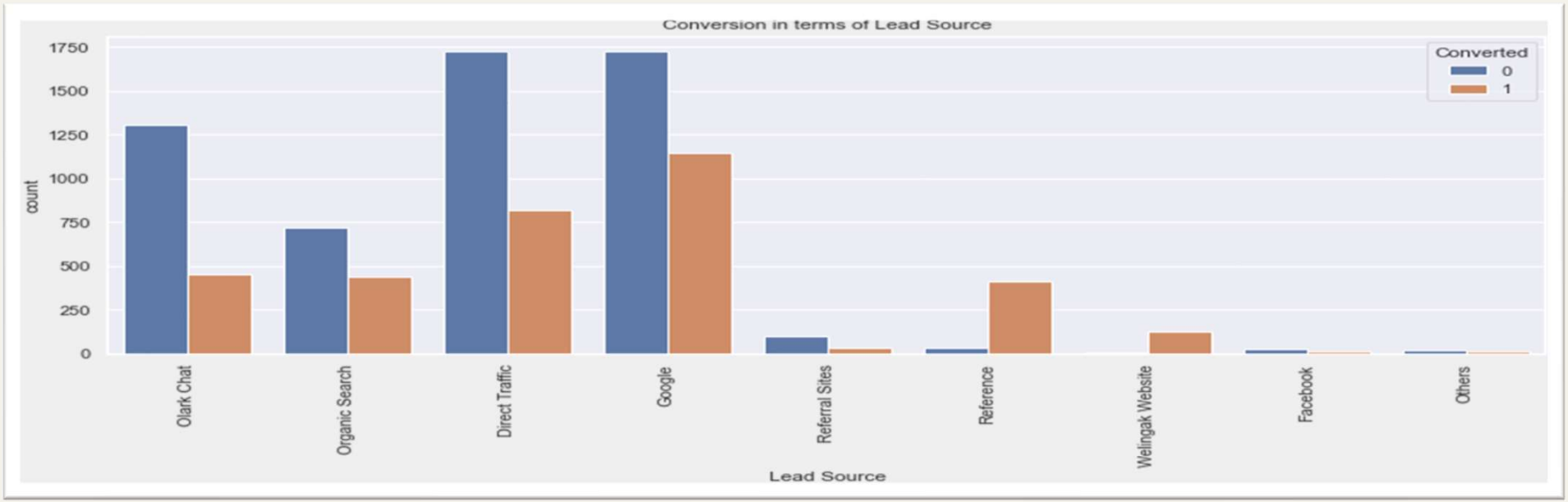
- Conclusion and recommendation

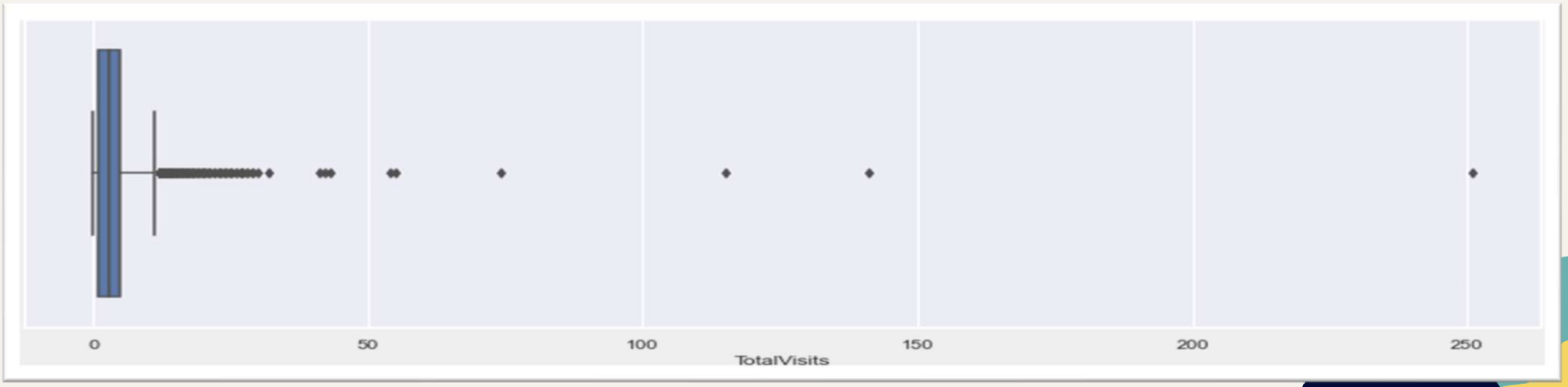
EDA: Categorical Data

Lead Source	Last Activity	Lead Origin
• Google and Direct Traffic both are the important source of lead conversion	• Since the probability of a response is greater when leads are converted, we must reach people via SMS and emails.	• When a landing page is submitted, a higher conversion rate is observed. • We can also see that there is a very high lead convert rate when the lead source is an add form.

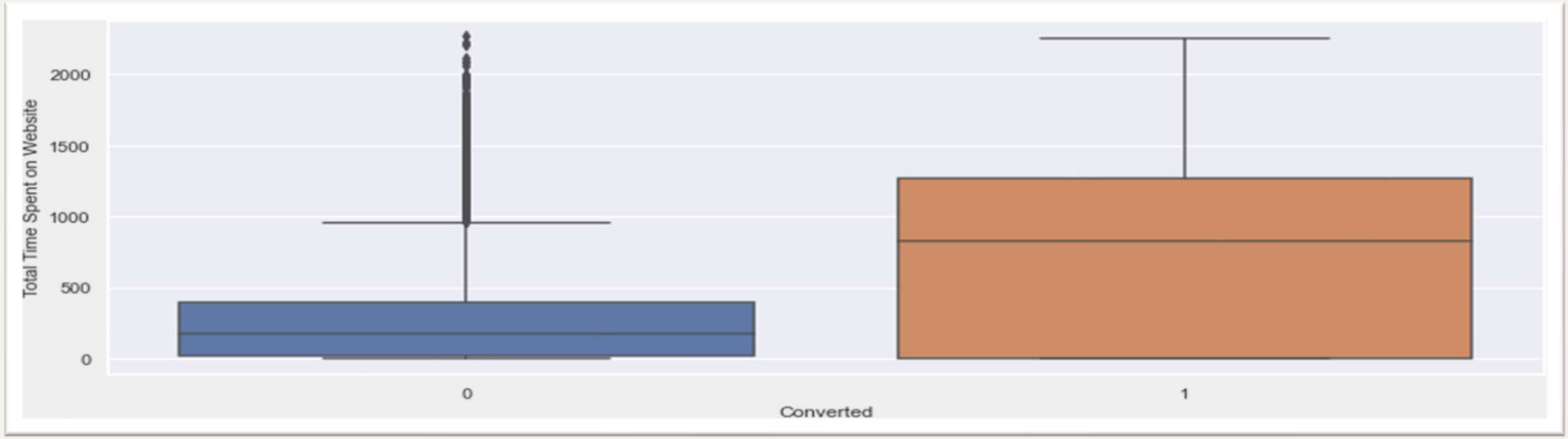

2000
1500
1000
500
0
count
Converted
0
1
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

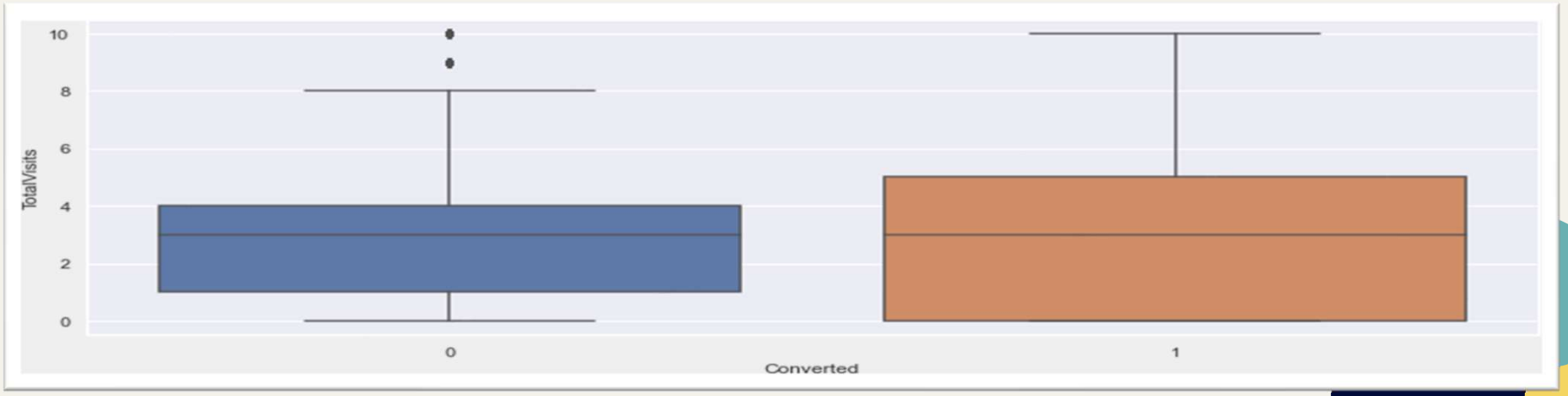

Conversion in terms of Lead origin
3000
2500
2000
1500
1000
500
0
count
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

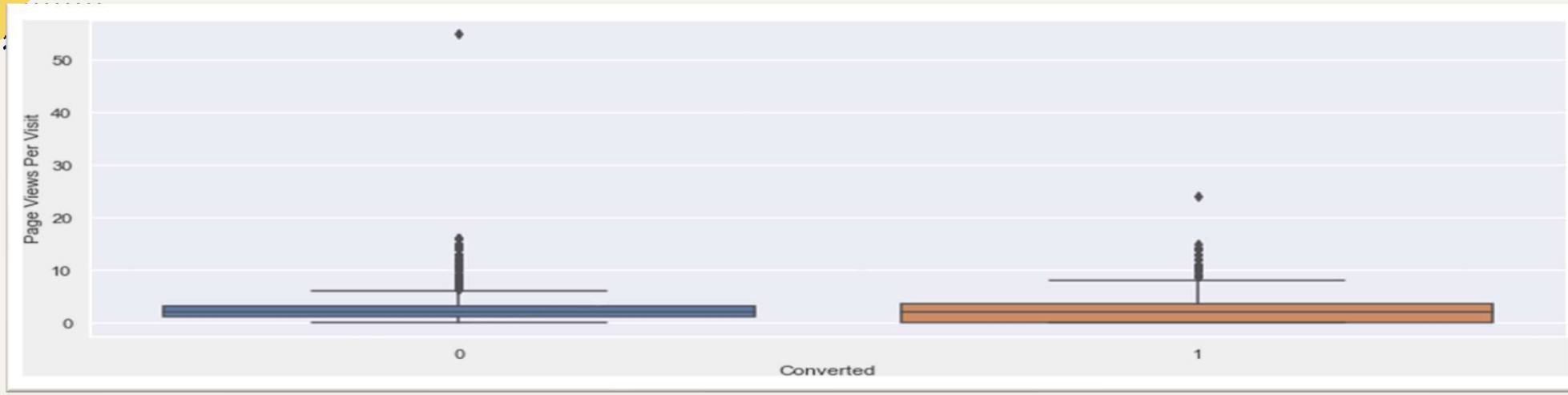

Conversion in terms of Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source

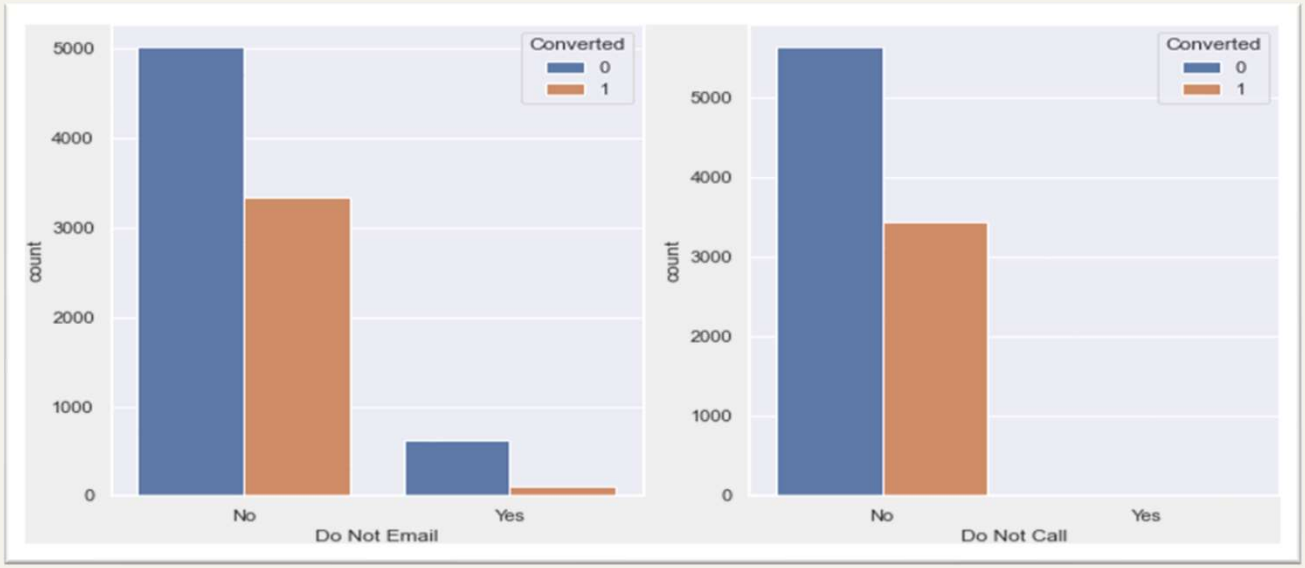

0
50
100
150
200
250
TotalVisits

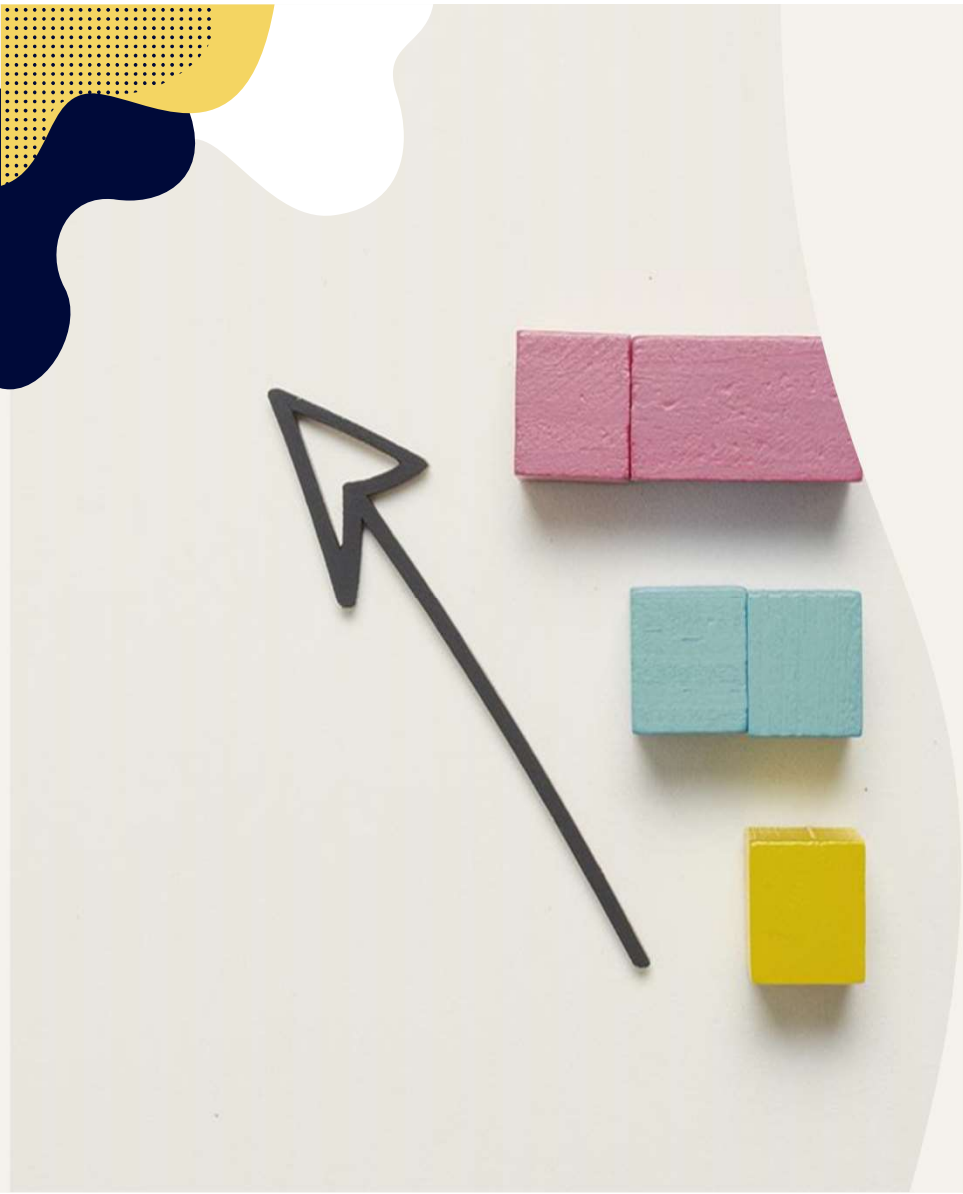
EDA Numerical Data






Page Views Per Visit
50
40
30
20
10
0
0
1
Converted

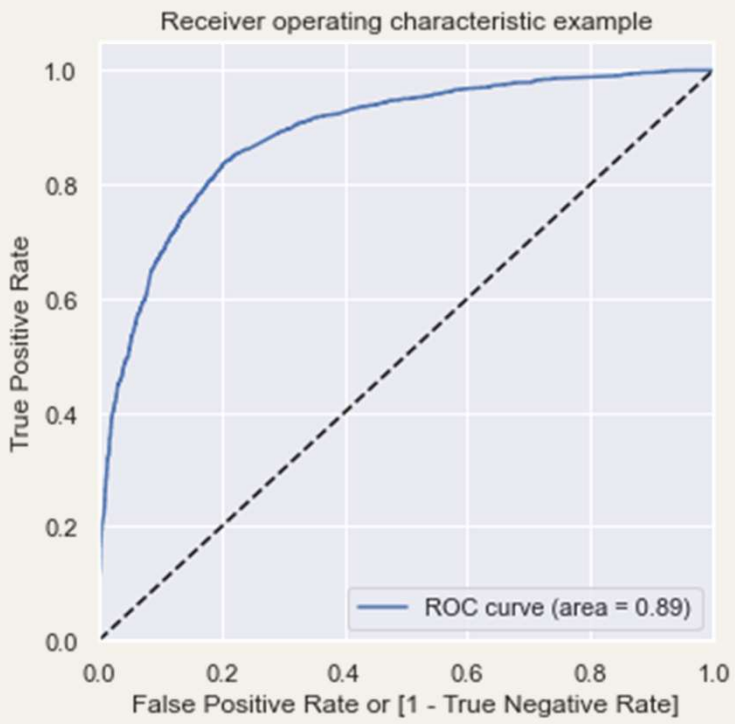

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

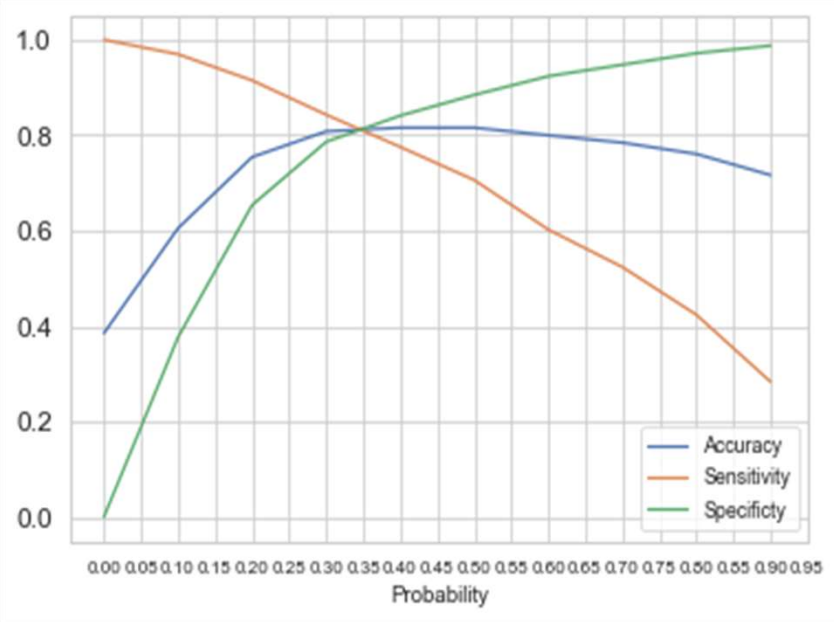


MODEL BUILDING

- Splitting the data into training and testing sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio
- Use RFE for feature selection
- Running RFE with 18 variables as output
- Building model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- Prediction on test data set
- Overall accuarcy is 81%

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy
Sensitivity
Specificty
Probability
0.0
0.2
0.4
0.6
0.8
1.0

Continue…..

ROC Curve

- **Finding optimal cut off point**
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is a 0.35

Conclusion

- It was found that the following factors affected potential buyers the most:

a) What is your current Occupation_unemployed
b) Lead origin_lead Add Form
c) Lead source_reference

- There are many important variables like city, specialization , occupation which can potentially explain Conversion better.It is important for the management to make few of these information mandatory to fill , so that we can use in our model and build important decisions for the business.

- The 'Cold Leads'(Customer having lead score <= 35) should be focused after the Sales Team is done with the 'Hot Leads'.

Continue.....

- High specificity will ensure that leads who are on the edge of being converted or not are not selected, whereas high sensitivity will ensure that almost all leads who are likely to convert are accurately forecasted.
- Customers who do not want to be contacted about the course should receive the relatively little attention.
- If the Last Notable Activity is Modified, he/she may not be the potential lead.


Thank You