# Leads Score Case Study


*Group Members:*

1. *Dev*
2. *Manohar Kumar*
3. *Medha Chauhan*

# PROBLEM STATEMENT


- X Education sells online courses to industry professionals and many professionals browse through their website for courses.
- When these people fill up a form and provide their email address and their phone number they are classified as a lead.
- The Lead conversion rate at X education is around 30% ( if they acquire 100 leads in a day only 30 would get converted) which clearly is a very poor rate.
- The company wishes to identify the most potential leads which are " HOT LEADS".
- If they identify this set of leads, the lead conversion rate would go up as the sales team would focus on communicating with the potential leads rather than making calls to everyone.

# BUSINESS OBJECTIVE

- X Education wants to know the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we will 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%.
- The model should be such that it can be used in the future as well.

# METHODOLOGY USED

**Importing the necessary libraries ( Pandas, Numpy, etc.) and the dataset given to us.**

**Data Cleaning And Data Manipulation**

1. Checking the number of missing values.
2. Finding null % across columns round.
3. Dropping columns if it contains large number of missing values not useful for our analysis.
4. Imputation of values if necessary.
5. Check and handle the outliers in the data.

# METHODOLOGY USED

**Exploratory Data Analysis**

1. Univariate Analysis: value count, distribution of variables etc
2. Bivariate Analysis: Correlation coefficients and patterns between the variables etc.

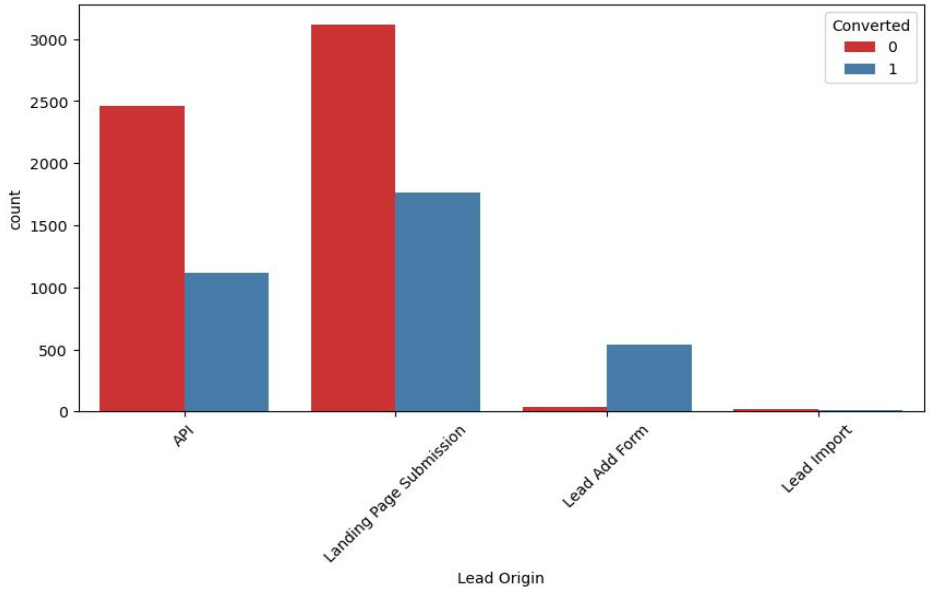
**Data Preparation:**

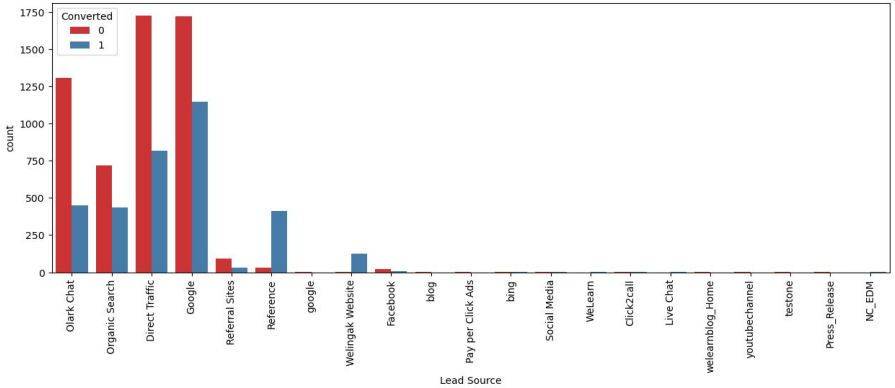
1. Creating a dummy variable for the categorical variables and dropping the first one.
2. Test train split
3. Feature scaling

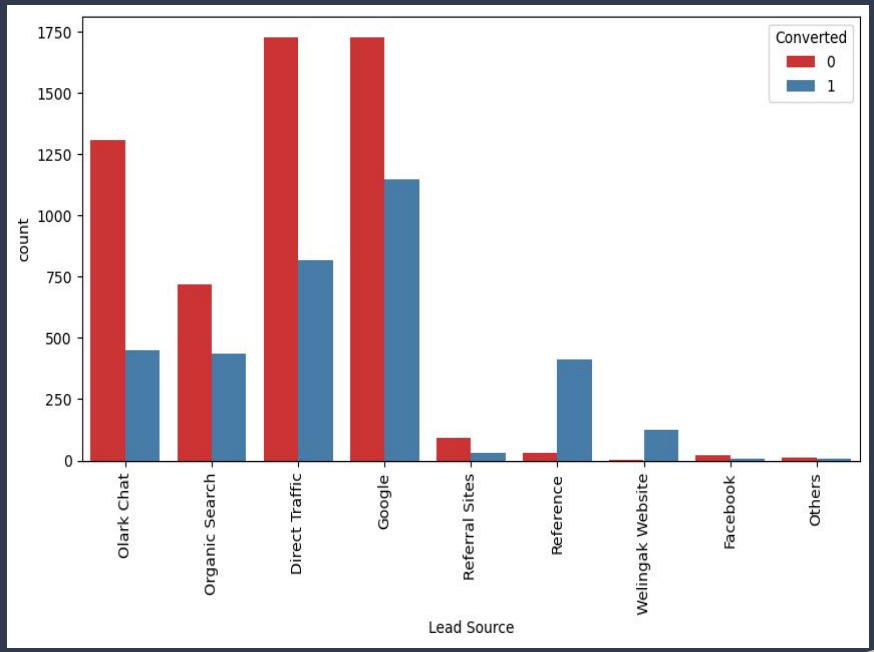
# METHODOLOGY USED

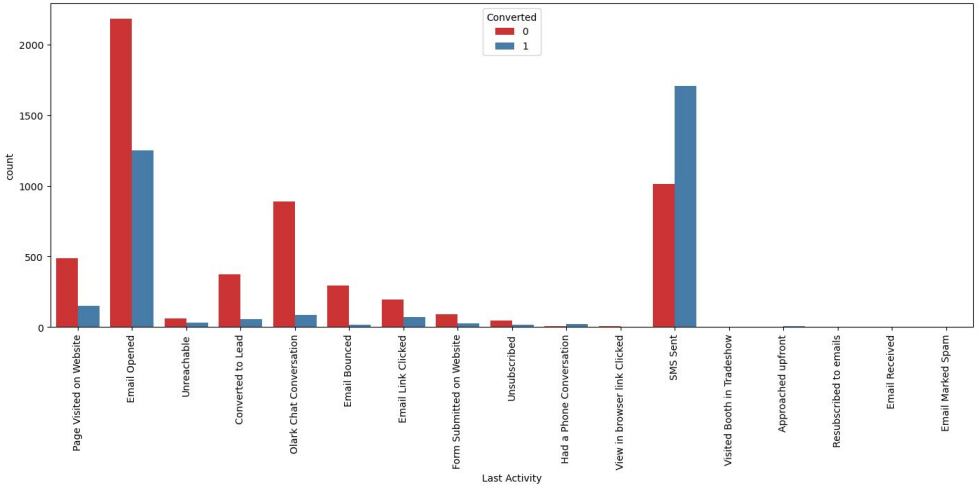
- **Model Building:**
  Logistic Regression is used for model building and prediction.
- **Validation of the model**
- **Model Presentation**
- **Conclusions and Recommendations**

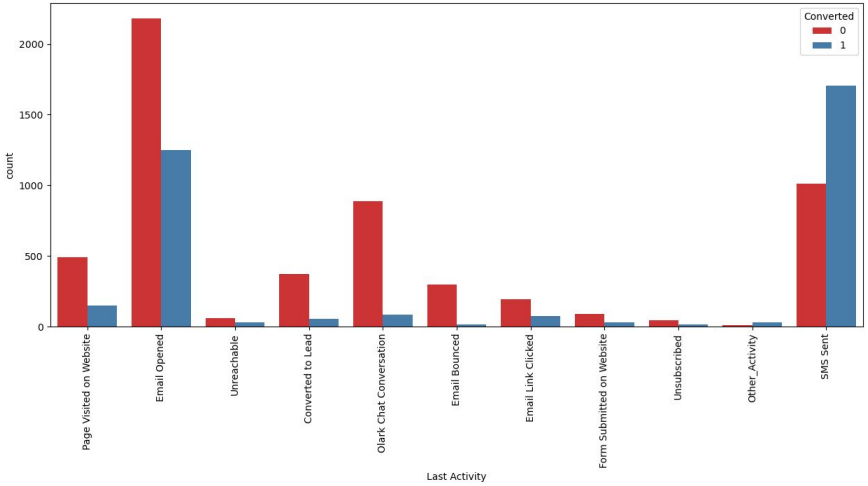
# EXPLORATORY DATA ANALYSIS (EDA)

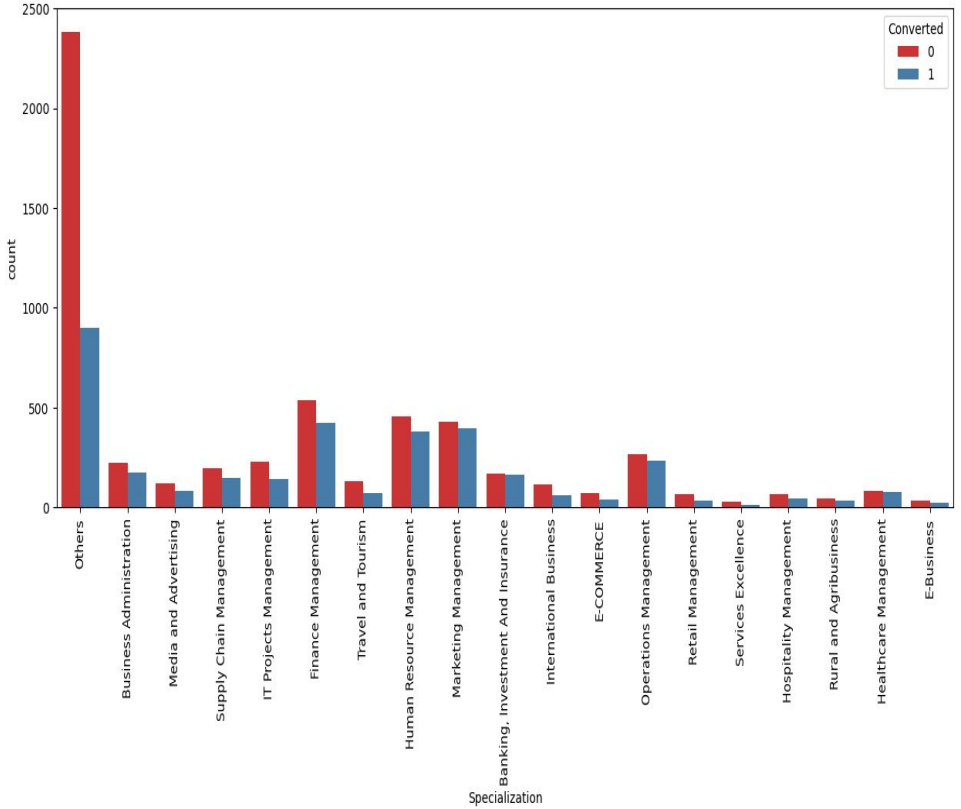


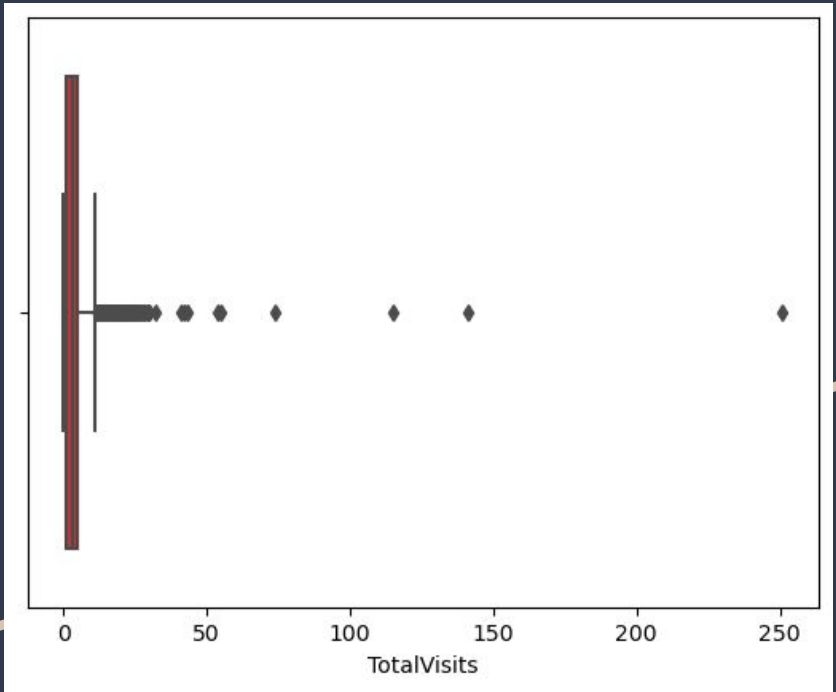

Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source

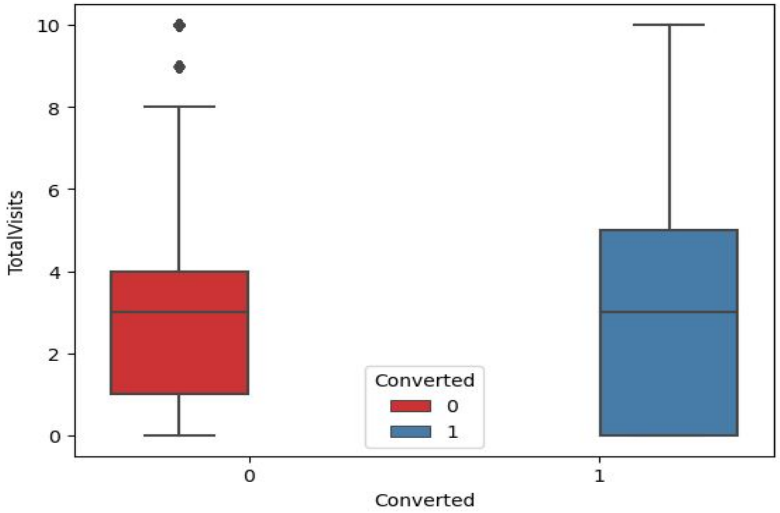

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached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

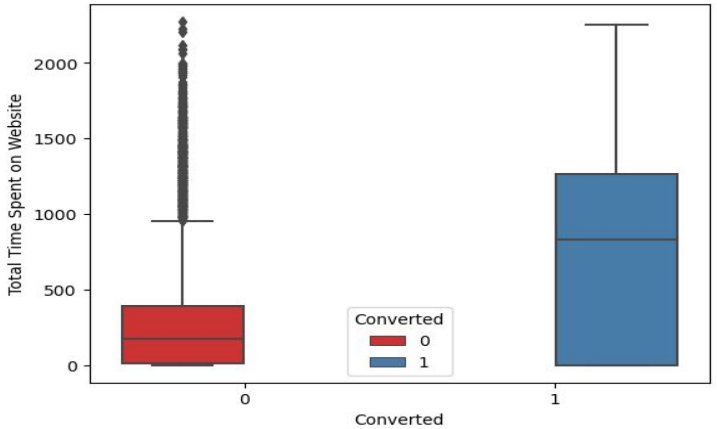

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity


Converted
0
1
count
2500
2000
1500
1000
500
0
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

# BOX PLOT


0
50
100
150
200
250
TotalVisits


10
8
6
4
2
0
TotalVisits
Converted
0
1
0
1
Converted

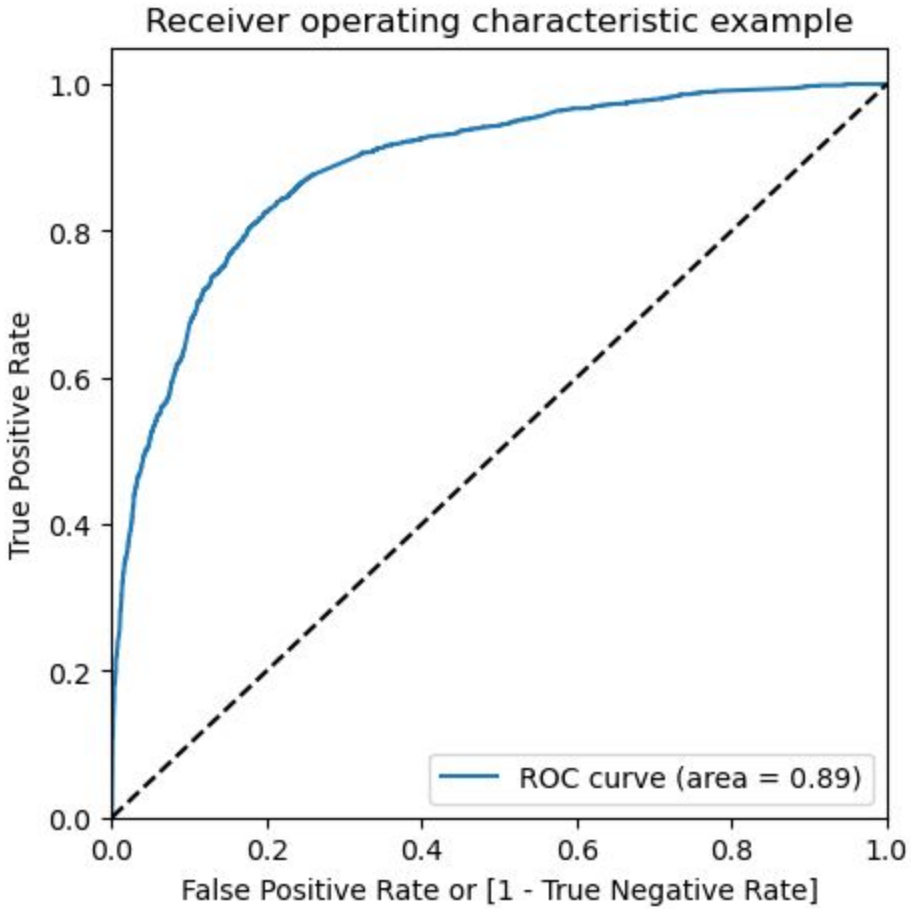

2000
1500
1000
500
0
Total Time Spent on Website
Converted
0
1
0
1
Converted

# MODEL BUILDING

- Splitting the data into training and testing sets. First basic step is the train-test split for this we have taken the 70:30 ratio.
- Use Rfe ( Recursive Feature Elimination)for feature selection.
- Running RFE with 20 variables as output.
- Building model by removing variables whose VIF values are high and the final model has variables having p-value as 0 The final 12 variables.
- Prediction on Test data set.
- **Overall accuracy: 81%**

# Roc curve ( Receiver Operator Characteristic curve)

- Finding optimal cutoff point.



# CONCLUSION

**The model has the ability to adjust to the company's requirement in the future.**

**The variables that mattered the most in the potential buyers are:**

1. **Total time spent on the website.**
2. **When the lead source was :**
   - **Olark chat**
   - **Reference**
   - **Wellingak website**
3. **When their current occupation is working professional.**
4. **When their lead origin is landing page submission.**