Lead Scoring Case Study

Team Members:
Surabhi Dadhich
Vittal Eswargoud
Vimal Kant

Background of X Education Company

- 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.
- 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%.

Problem Statement & Objective of the Study

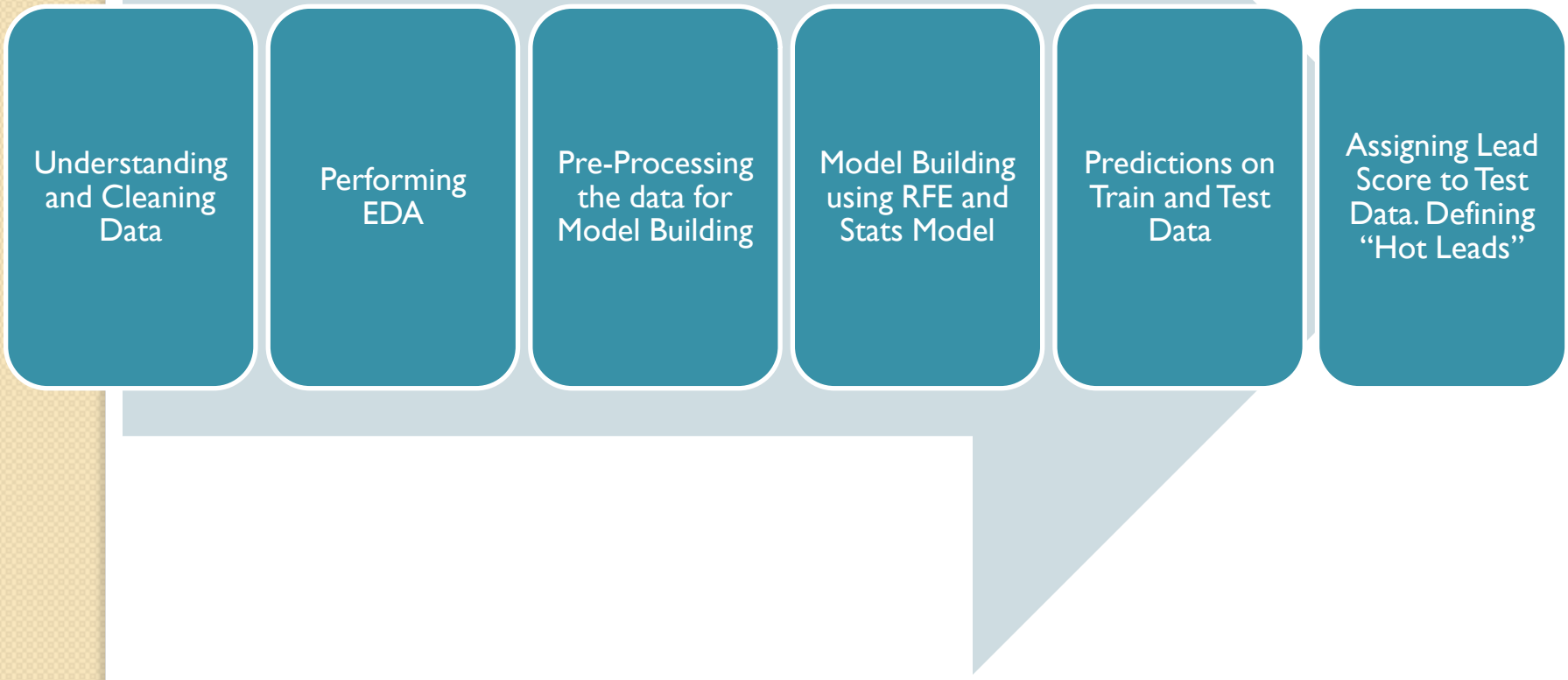
Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone

Objective of the Study:

- To help X Education select 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 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 has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach



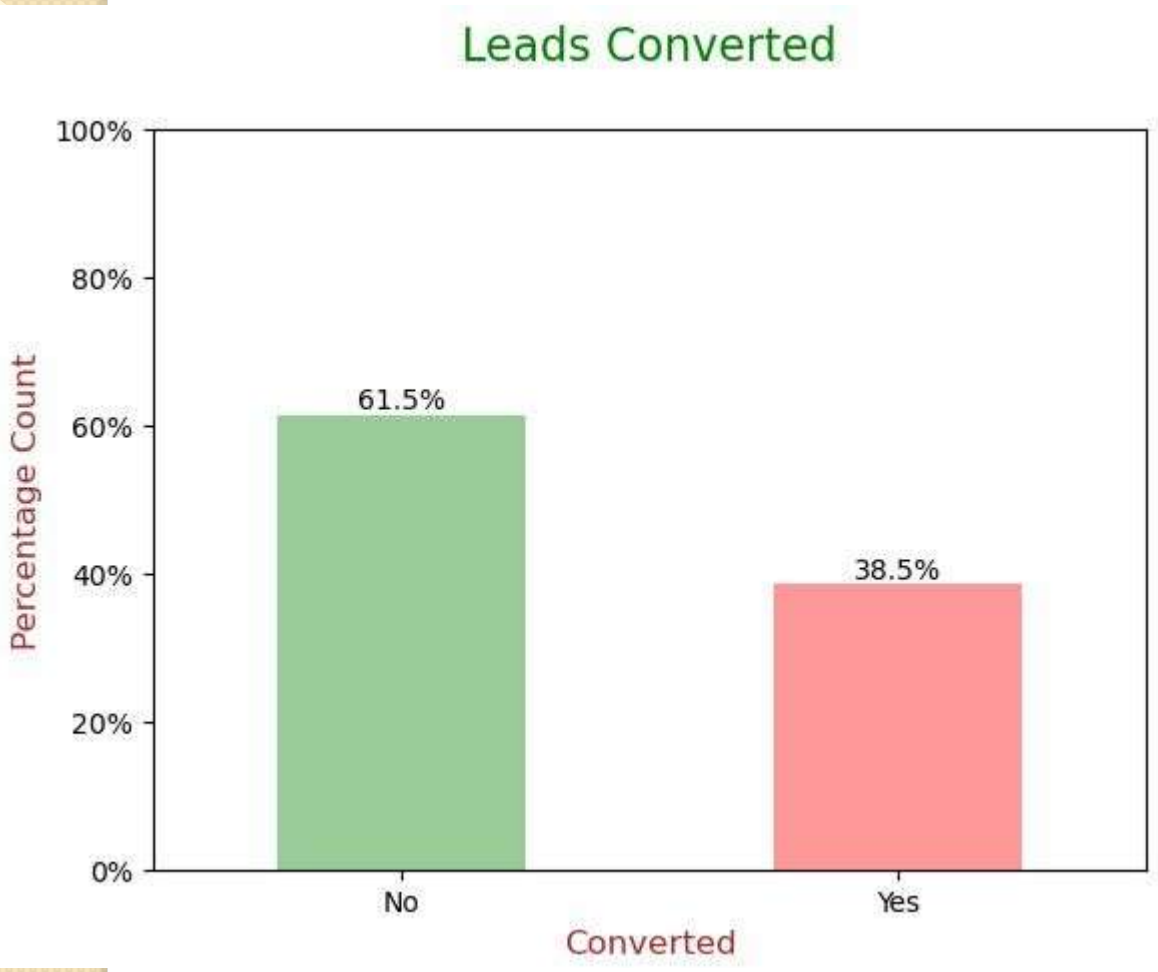
Understanding and Cleaning Data

- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (Tags, country)
- Imputation with mode was done for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modelling (Prospect ID, Lead Number) or only one category of response were dropped.

Understanding and Cleaning Data (Contd)

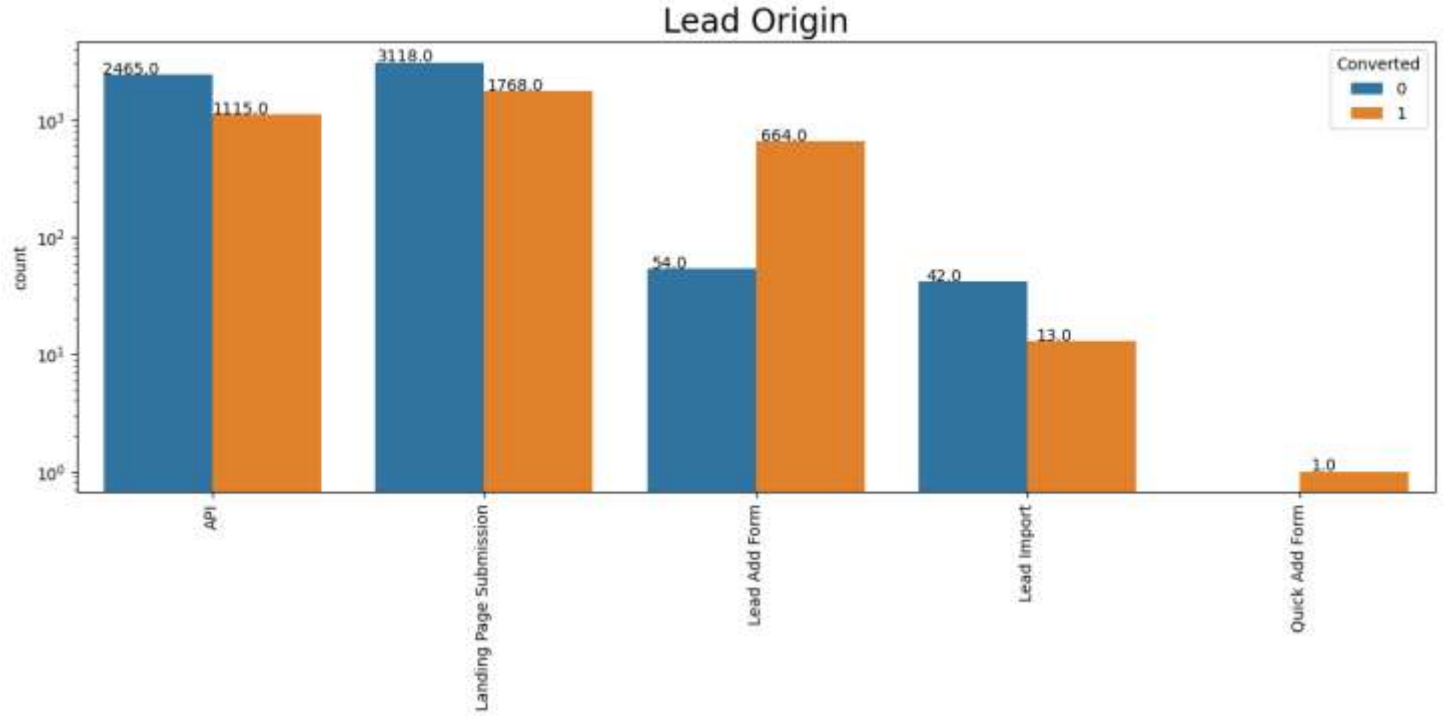
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **TotalVisits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
    - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)

EDA



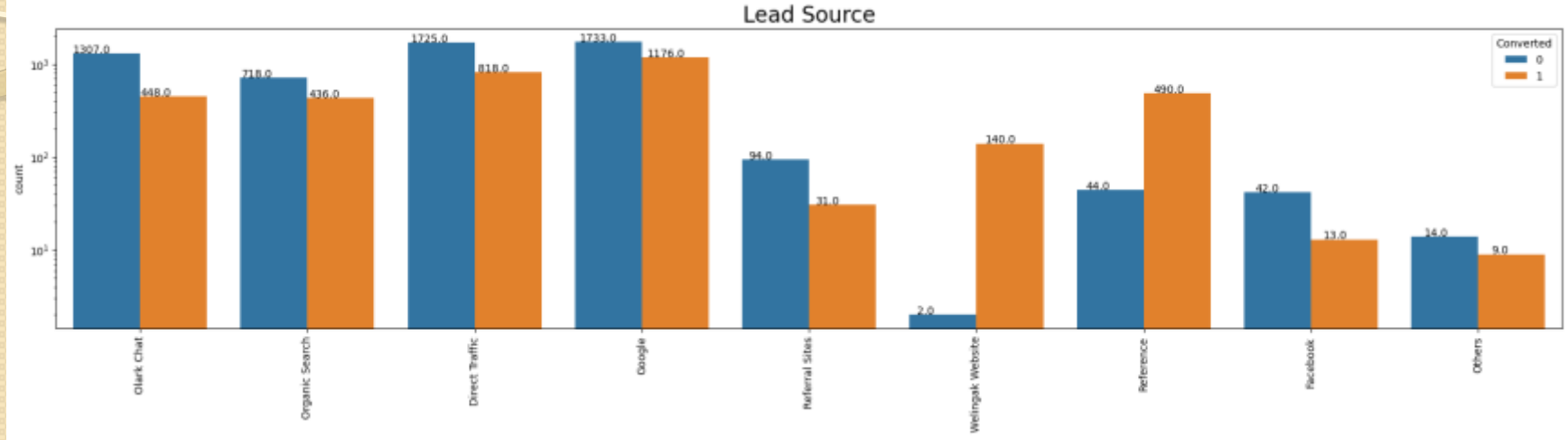
- Only 38.5% of the people have converted to leads.
- While 61.5% of the people didn't convert to leads.

EDA



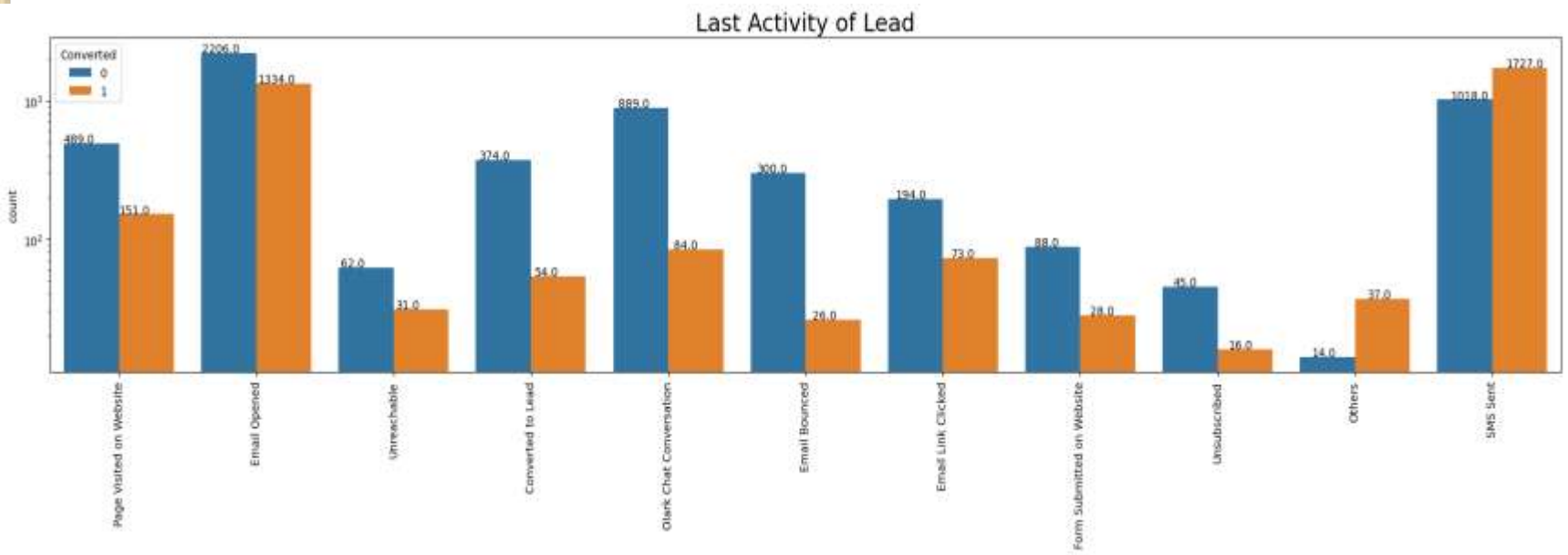
Landing page submission had the highest converted followed by API

EDA



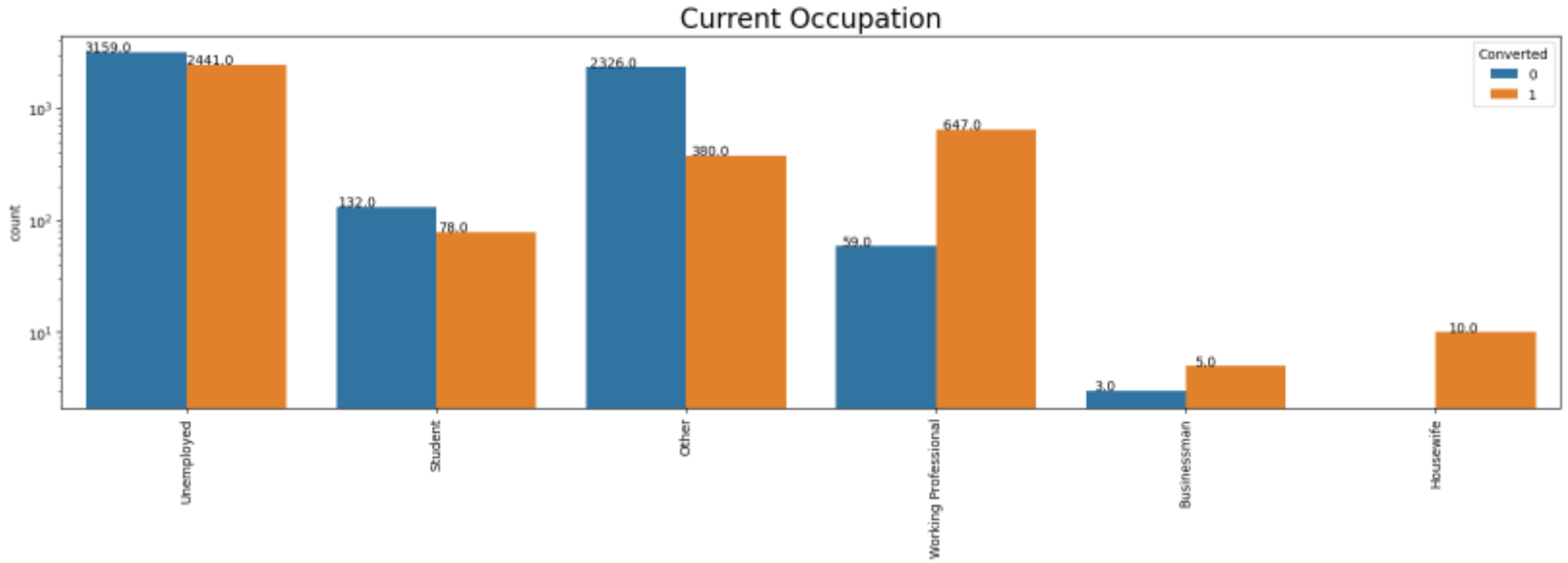
- Higher no of leads are converted from Google as lead source
- Welingkar website and Reference have a higher conversion rate

EDA

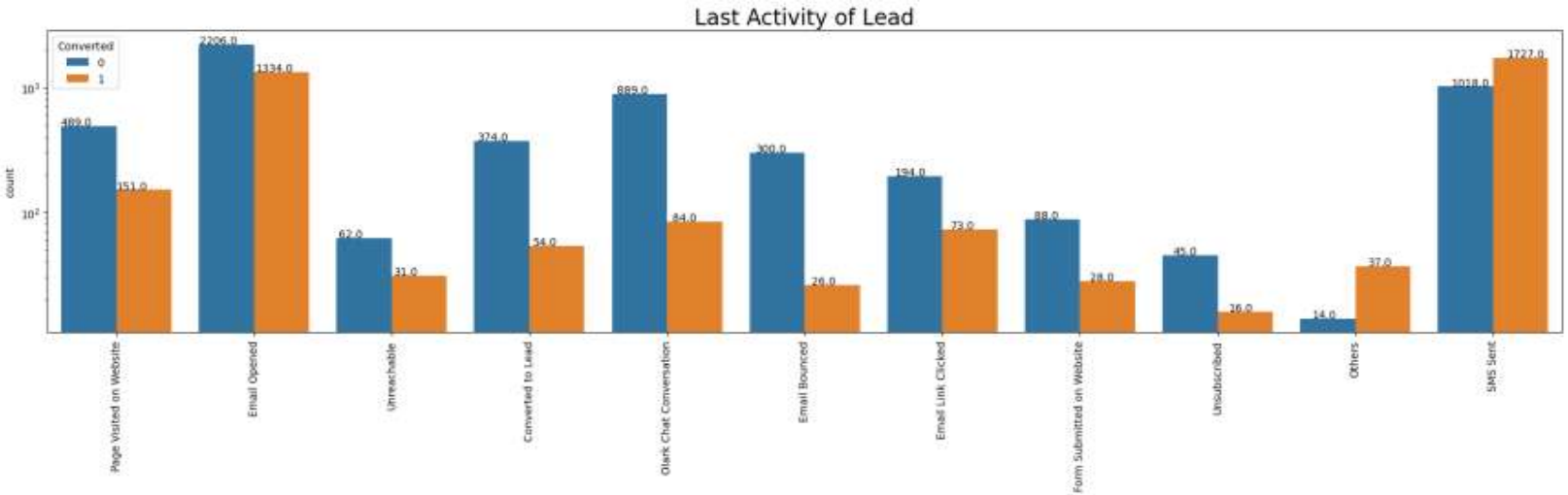


- Email Opened and SMS sent had higher conversion
- Olark chat conversation had the highest non conversion

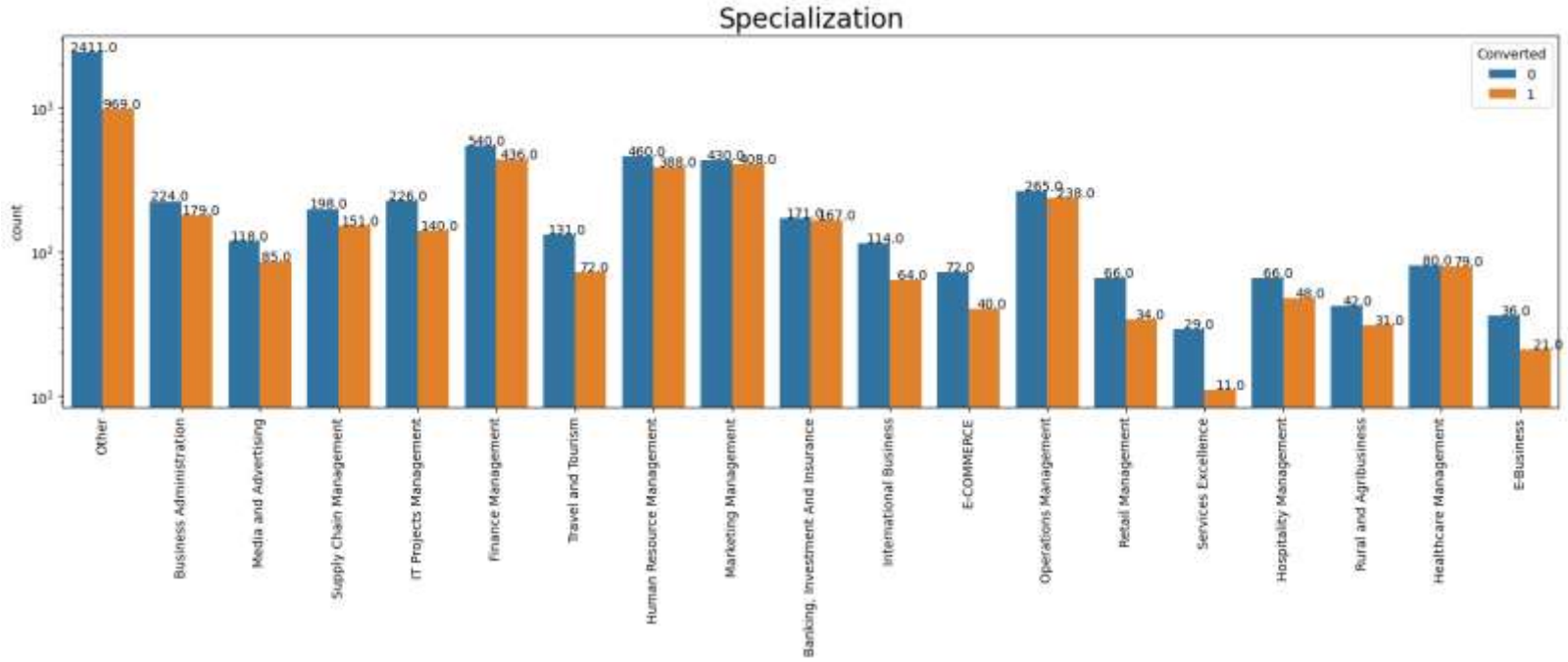
EDA



- Those who selected "Do Not Email" as "No" have higher conversion
- Working professions and Unemployed had a higher conversion
- Others have a lower conversion rate

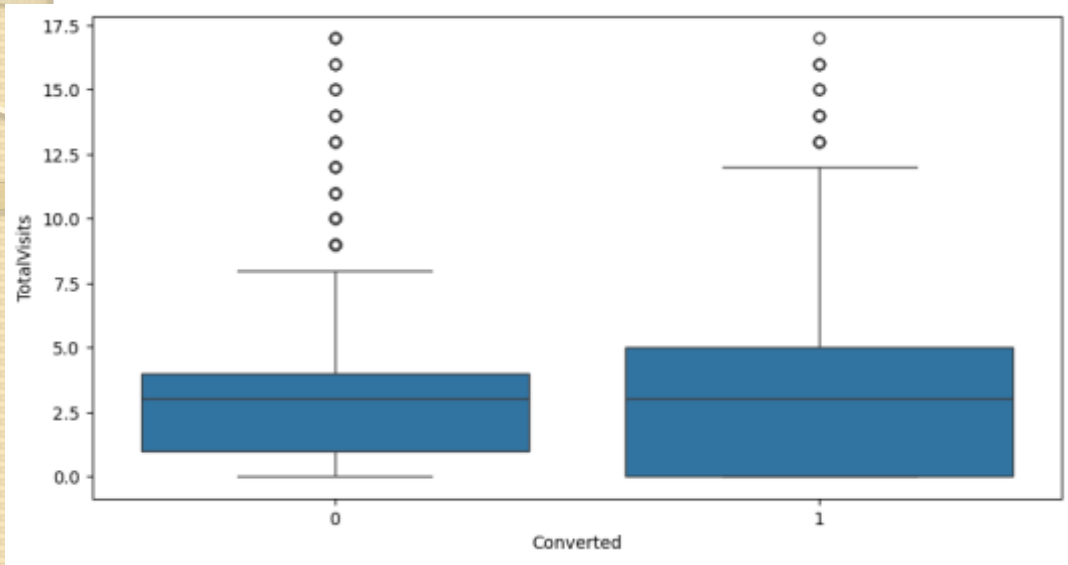


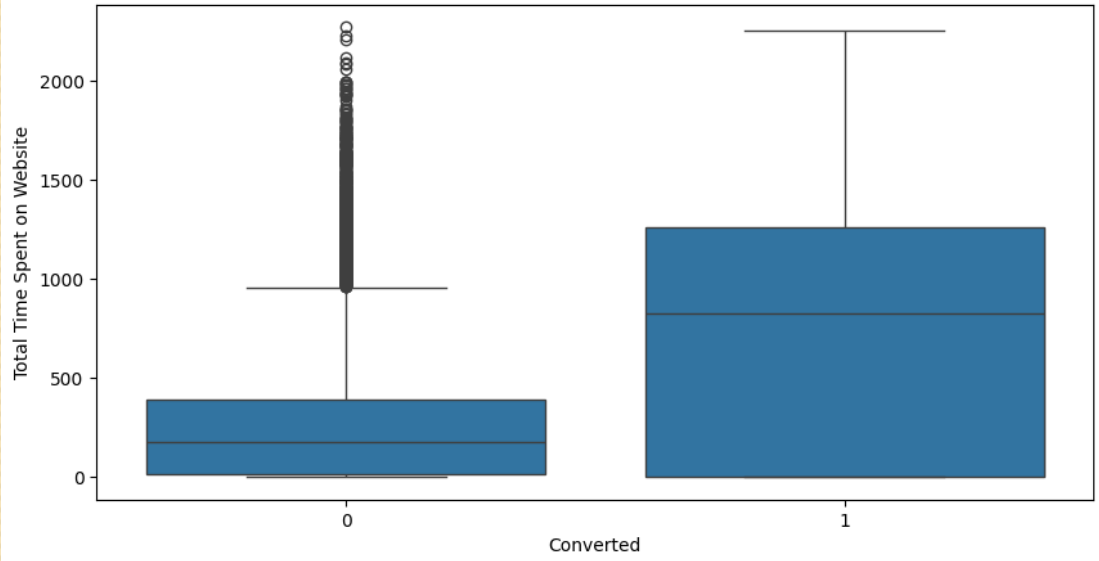
- If Last activity of lead was SMS sent then it was converted
- If Email was opened and also it was converted
- Email bounced had a lower conversion rate



- Those in Human Resource management and those with Marketing management have almost equal conversion and non conversion rate
- Thos in Finance Management have a good conversion rate
- Others have higher non conversion rate

EDA





- Converted value in Total Visits has a higher spread compared to non converted
- Total time spent on website has a higher spread

Data Preparation before Model building

- Created dummy features (one-hot encoded) for categorical variables – Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation'
- Splitting Train & Test Sets
  - 70:30 % ratio was chosen for the split
- Feature scaling
  - Standardization method was used to scale the features
- Checking the correlations
  - 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98.
  - 'Do Not Email' and 'Last Activity_Email Bounced' having higher correlation. 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.
  - 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72.
  - 'Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current Occupation_Working Professionals' having positive correlation with our target variable 'Converted'.
  - These were kept to be handled by the RFE

Model Building

Feature Selection

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
  - Pre RFE – 48 columns & Post RFE – 13 columns

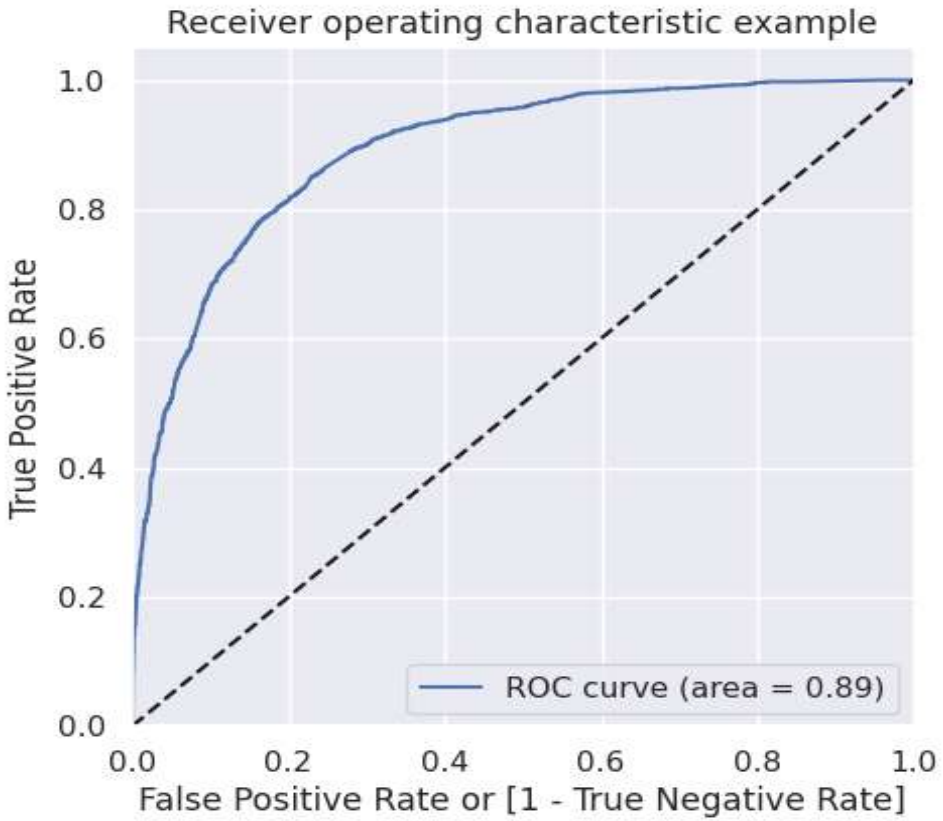
Model Building

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.

- Model 3 looks stable after four iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5

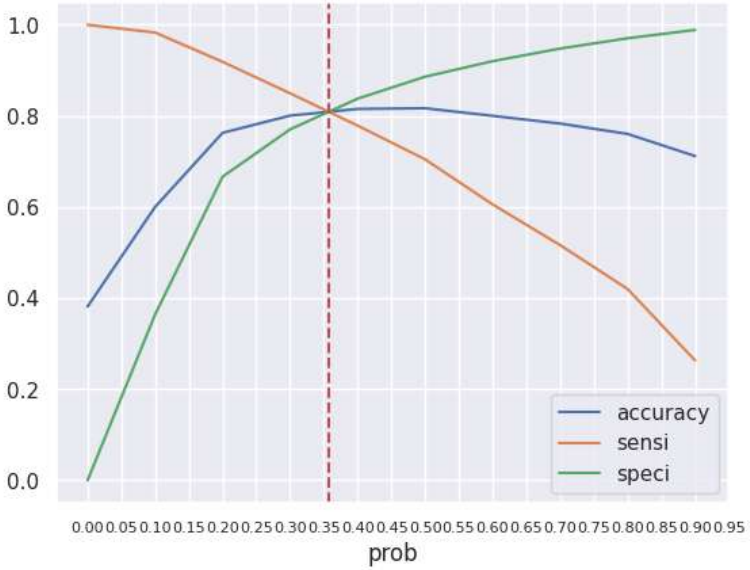
- Hence, **logm3** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

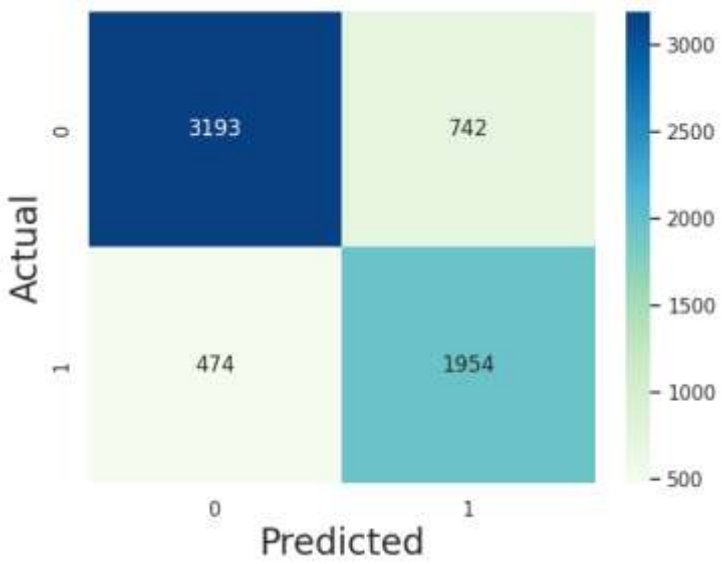
Model Evaluation



ROC curve with an area of 89% which tells us that the model is good

Model Evaluation





Observations:

After running the model on the Test Data , we obtain:

- Accuracy : 80.38 %
- Sensitivity : 81.44 %
- Specificity : 79.69 %

This tells us that the ball park figure of 80% is acheived

Model Evaluation

Upon Comparing the values obtained for Train & Test:

- Train Data:
  - Accuracy : 80.88 %
  - Sensitivity : 80.47 %
  - Specificity : 81.14 %
- Test Data:
  - Accuracy : 80.38 %
  - Sensitivity : 81.44 %
  - Specificity : 79.69 %
- This leads to conclusion that the not only the model achieves the desired figure of 80% but is also consistent

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

- Below is the list of highest possible conversion rate
  - Lead Source_Welingak Website: 3.153724
  - Lead Origin_Lead Add Form : 2.978775
  - What is your current occupation_Working Professional: 2.392955
  - Last Activity_Unsubscribed: 1.445675
  - Last Activity_SMS Sent:1.382380
  - Lead Source_Olark Chat:1.170840
  - Total Time Spent on Website: 1.072358

Recommendation based on Final Model

- Most likely to be converted:

  - Whose Lead Source is Welingak Website and Olark Chat
  - Where the Lead Origin is Lead Add Form
  - Those whose are Working Professional
  - Last Activity is Unsubscribed and SMS Sent
  - Who spent good amount of time on the Website

- Most Unlikely to be converted

  - Whose Lead Origin is Landing Page Submission
  - Who have a selected their Specialization as Other or Hospitality Management
  - Whose current occupation is Other
  - Whose Last Activity was in Olark Chat Conversation
  - Who have explicitly mentioned "Do Not Email"