LEAD SCORING

Aniket Prakash Manwatkar

Meethu P G

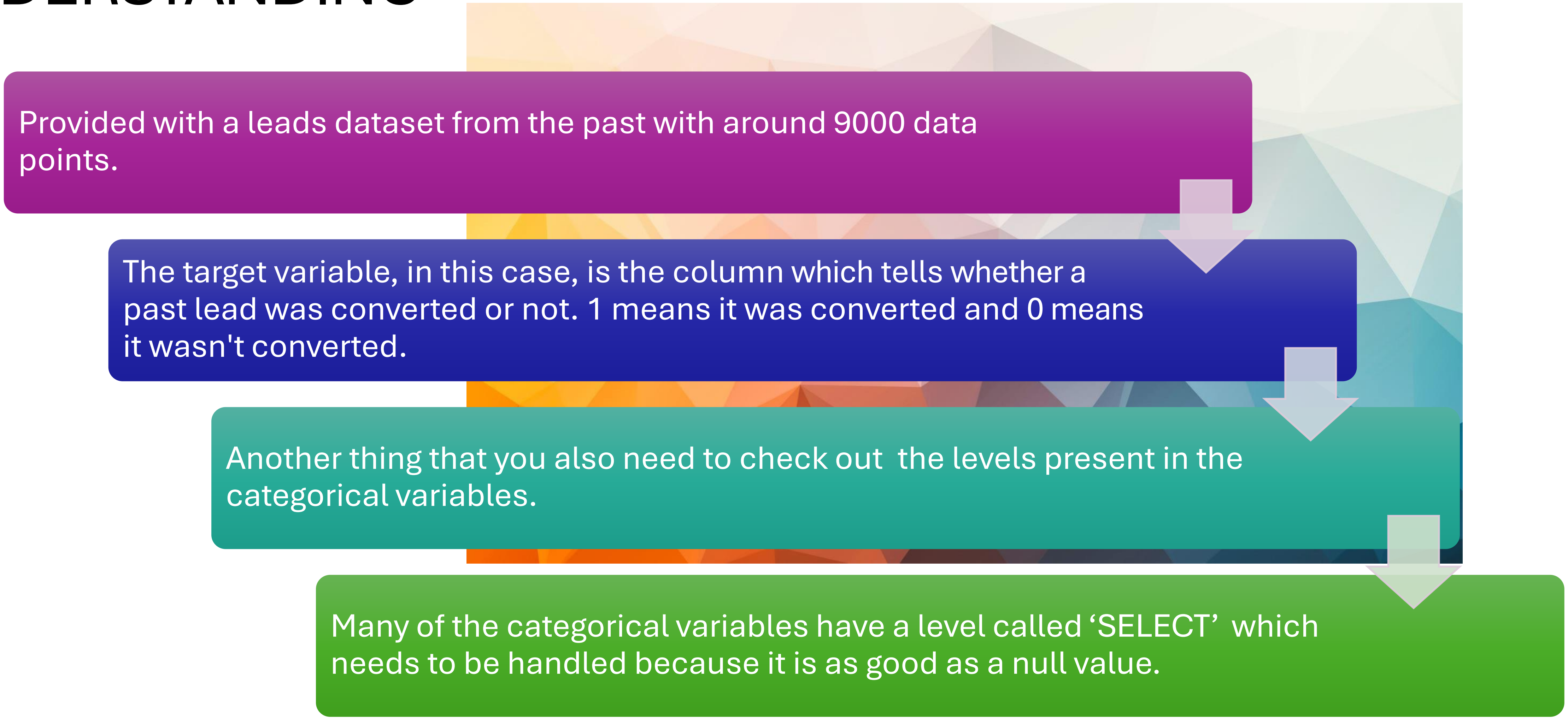
Manikandan Natarajan

Problem Statement

- X Education sells online courses to industry professionals, who are identified as leads through various websites and search engines. The company's lead conversion rate is around 30%, but it is currently low. To improve efficiency, the company aims to identify potential leads, or 'Hot Leads', and focus on communicating with these leads instead of making calls to everyone. This will allow the sales team to focus on converting more leads.

- X Education has hired you to assist in selecting promising leads for conversion into paying customers. The company requires a lead score model, with a target conversion rate of 80%, based on the CEO's expectations.

UNDERSTANDING



Provided with a leads dataset from the past with around 9000 data points.

The target variable, in this case, is the column which tells whether a past lead was converted or not. 1 means it was converted and 0 means it wasn't converted.

Another thing that you also need to check out the levels present in the categorical variables.

Many of the categorical variables have a level called ‘SELECT’ which needs to be handled because it is as good as a null value.

OBJECTIVE

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

These problems are provided in a separate doc file.

Fill it based on the logistic regression model you got in the first step.

DATA CLEANING

1. Importing Necessary libraries
2. Exploring data sources, types, quality, and structure.
3. Identifying potential challenges or biases.
4. Preparing data for analysis and decision making.
5. Removing the unnecessary characters, replacing null values
6. Dropping the columns which as more than 40% of null values to get accurate analysis

DATA VISUALIZATION

Data analysis is the process of inspecting, cleansing, transforming, and modelling data to discover useful information. Plotting is a graphical technique for representing a data set, usually as a graph showing the relationship between two or more variables

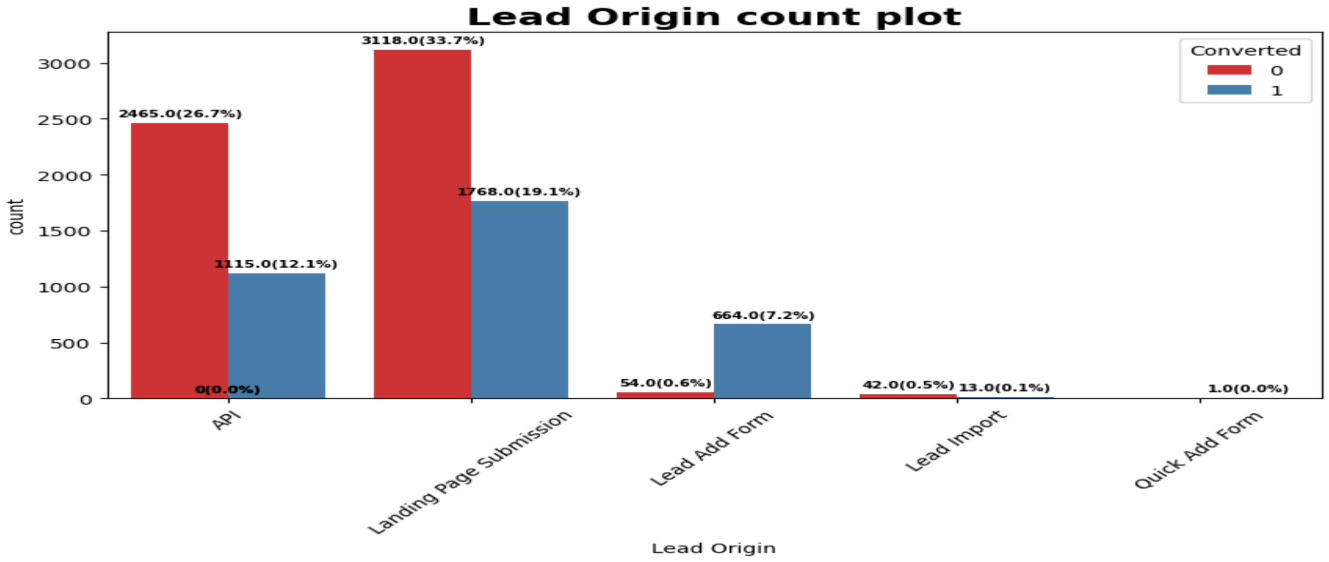
Visualizing the Tag:

- The highest call status is "Not Defined"
- The 40% of respondents will revert after reading the email.
- The 20% of respondents call status is ringing

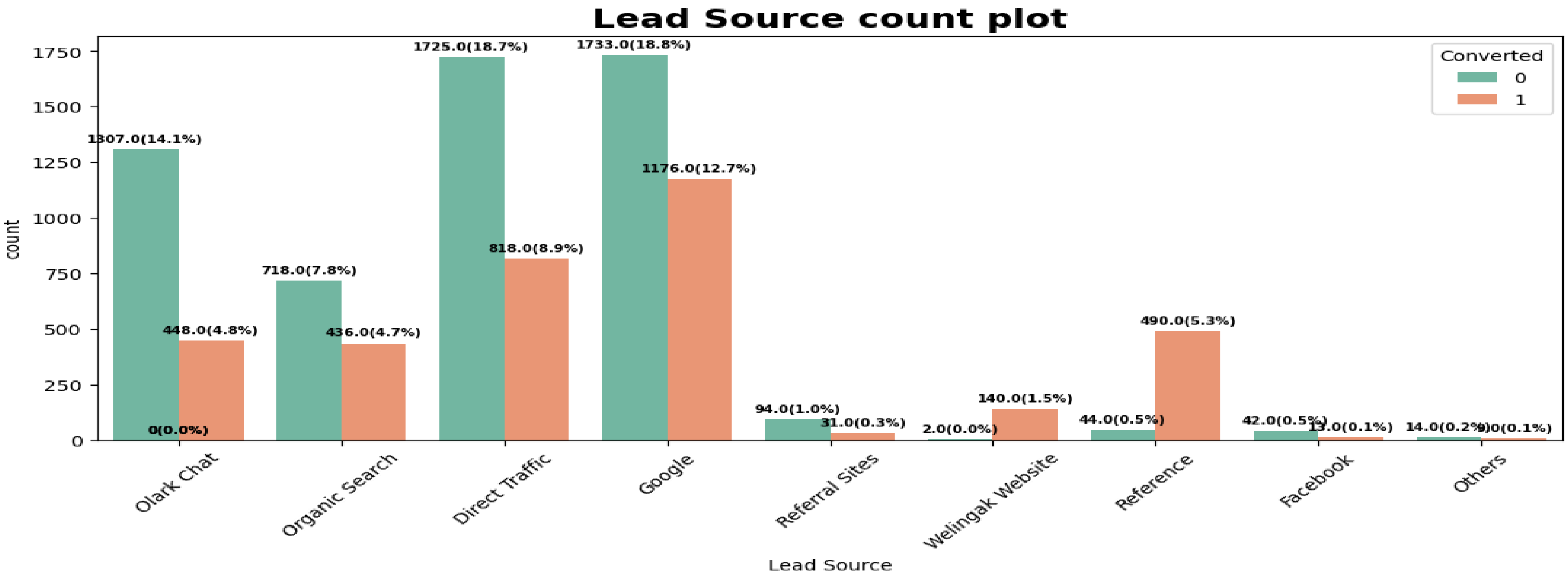
Analyzing the column “What matters most to you in choosing a course”

The Highest Votes for choosing the course is Better Career Prospects

LEAD CONVERSION RATE is 38 %

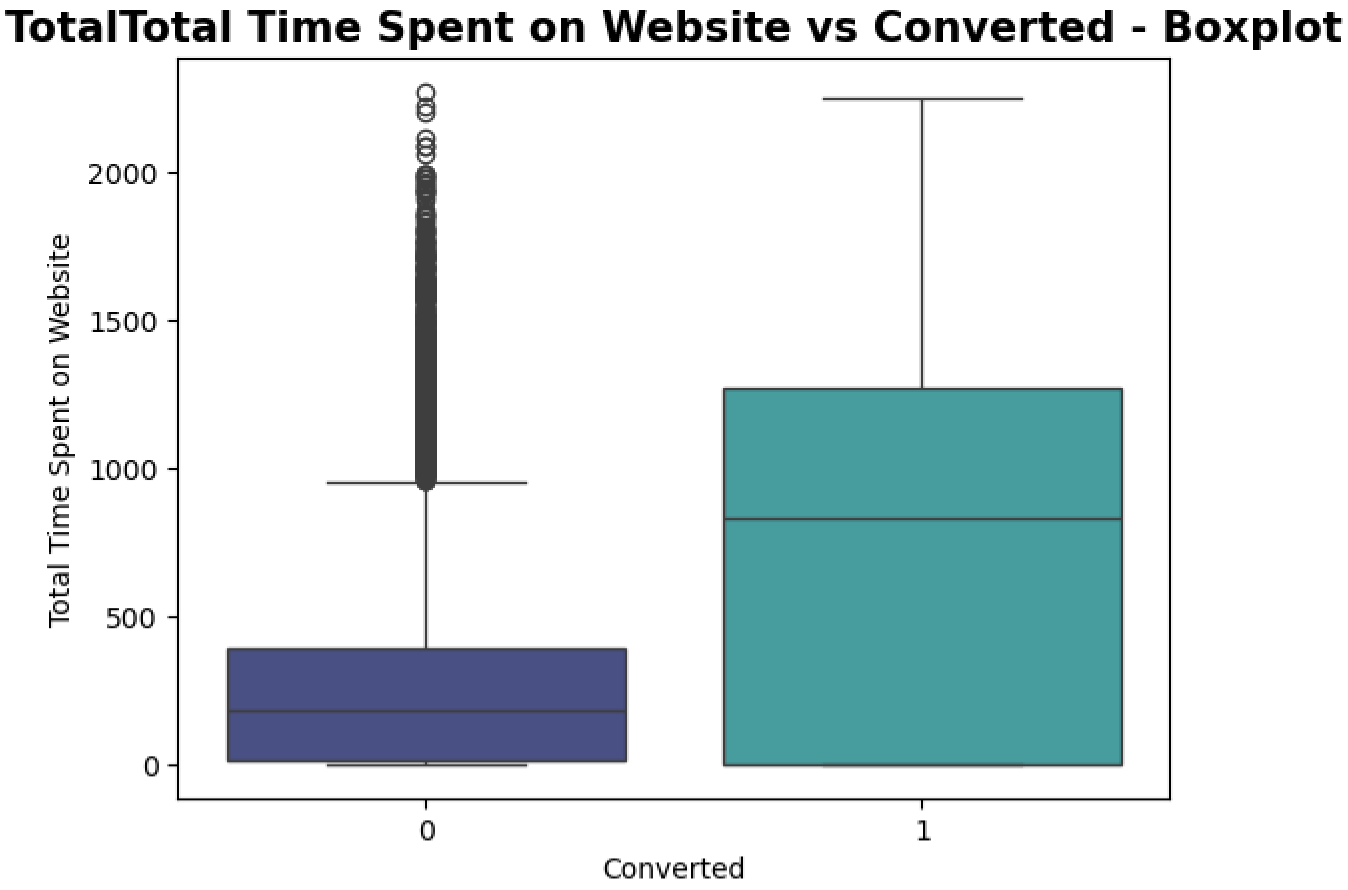


To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



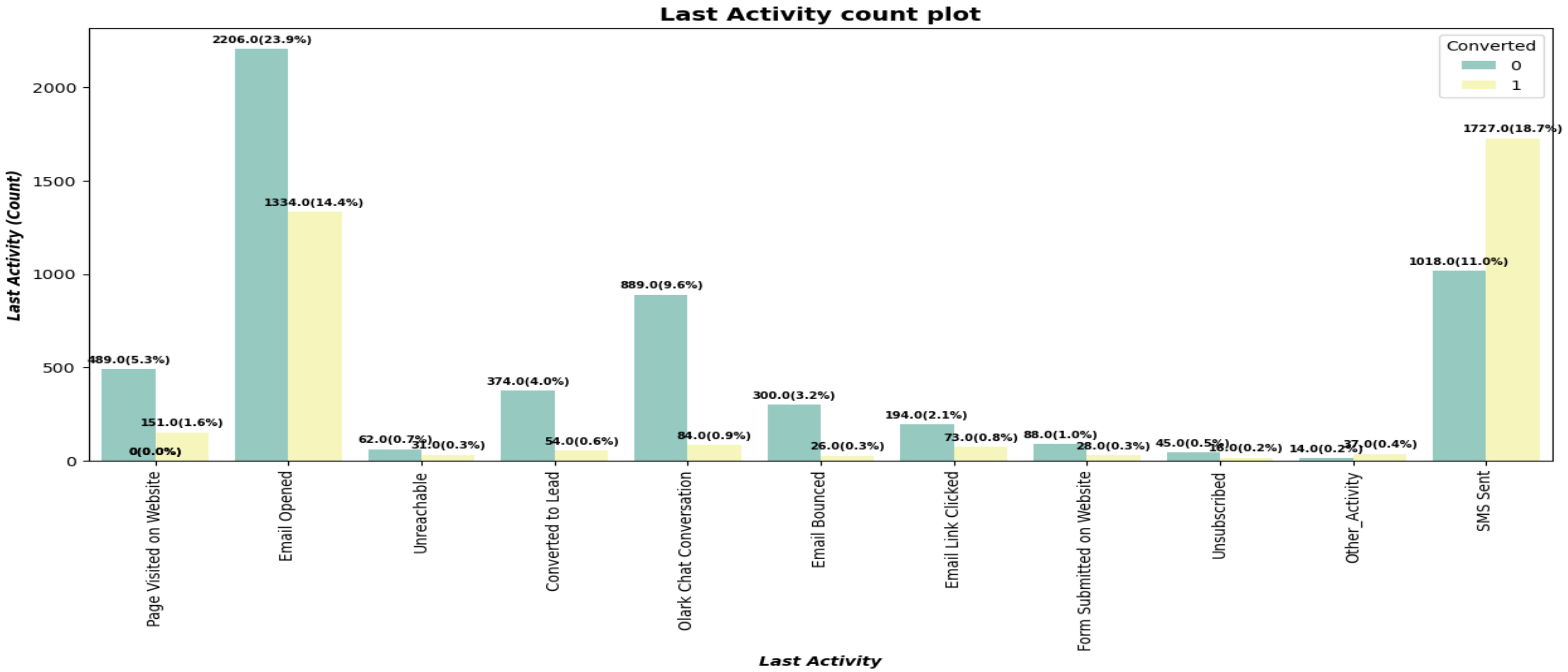
1. The most substantial number of leads is generated through Google and Direct traffic sources.
2. The conversion rate for reference leads and leads originating from the Welingak website is notably high.

To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website



- ✓ Leads spending more time on the website are more likely to be converted.

- ✓ Website should be made more engaging to make leads spend more time.



Inference:

- ✓ Most of the lead have their Email opened as their last activity.
- ✓ Conversion rate for leads with last activity as SMS Sent is almost 60%.

Creating Dummy variables and converting them into numerical values to get a better model evaluation for prediction

FEATURE SELECTION AND SCALING

SCALING THE NUMERICAL COLUMNS IN THE DATASET WHICH HAVE DIFFERENT SCALES. SCALING THE DATA USING STANDARD SCALER.

LEAD SCORING INVOLVES ASSIGNING SCORES TO LEADS BASED ON VARIOUS ATTRIBUTES TO PRIORITIZE AND IDENTIFY POTENTIAL CUSTOMERS.

THE CHOICE OF FEATURE SELECTION AND SCALING METHODS MAY DEPEND ON THE SPECIFIC CHARACTERISTICS OF YOUR DATASET AND THE ALGORITHM YOU PLAN TO USE FOR LEAD SCORING

SPLITTING AND MODEL EVALUATION

- A training and testing data split is the division of a dataset into two subsets:

Training set: Used to train the AI model

Testing set: Used to evaluate the performance of the model

- The accuracy of the model is 0.82

PREDICTIONS ON THE TEST DATA

- The accuracy for the prediction 0.82 and precision is 0.75

CONCLUSION:

The prediction testing proves that the model can effectively detect hot leads. By using the optimal cutoff value, the model showed it could accurately identify hot leads. The classification report at the end gives a quick overview confirming that the model is pretty good at determining if a lead is hot or not.