Lead Scoring...

a Classification Case Study

by

Abhinav Kumar & Subas Chandra Giri

Problem Statement ...

Problem Statement

1. X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

2. 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 score have a lower conversion chance.

Business Objective...

Business Objective

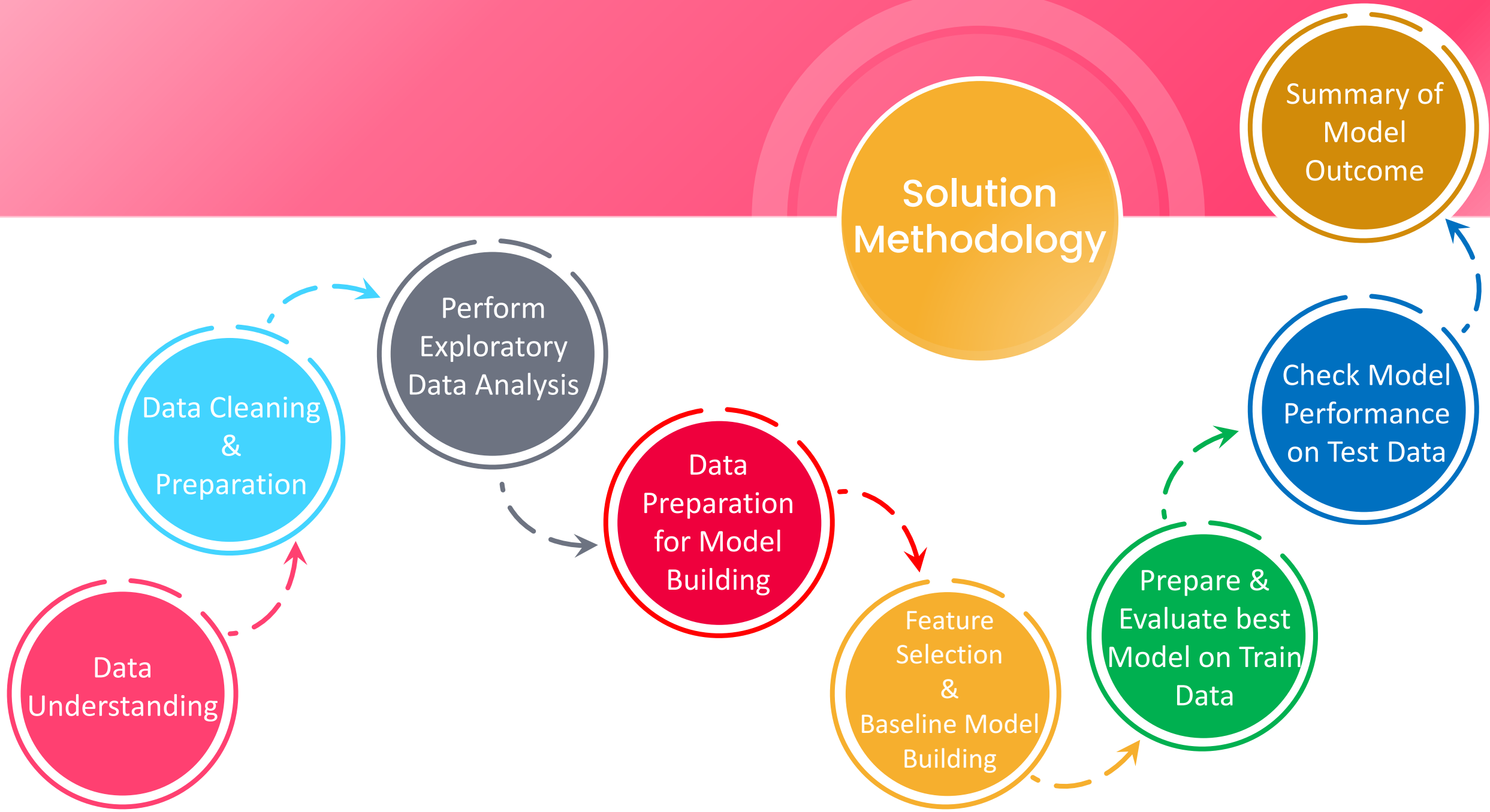
Business Objective

1. Client wants to built a model which can predict the hot leads which will help them to reduce the efforts and cost

2. Also client is looking for such a model which can handle the changes required in future

Solution Methodology...

Solution Methodology



1

Data Understanding

Data Understanding

- In '**Leads**' dataset we observe **9240** rows and **37** columns, out of which 30 are categorical and 7 are numerical variables.
- There are few columns which have more than 40% missing values.
- There are 1943 rows without any missing data while 3719 rows have missing values more than 5 per row and 619 rows have more than 10 per row.
- The class distribution of target variable 'Converted' is 61:39 for No and Yes. The data is imbalanced.
- Lead profile and how did you hear about X-education have missing values more than 70% after we clean the data.

2

Data Cleaning & Preparation

Data Cleaning & Preparation

- We have replaced the 'select' values with NaN. With that, the number of null values increased and so its percentage.
- We have dropped the static columns i.e. columns having same values for all rows & the columns with no variance i.e. columns having more than 90% data for one class
- We dropped columns having over 40% missing values. This also resulted in dropping the rows with higher missing values.
- The maximum number of missing values row wise now dropped form 13 to 3. This step has cleaned our data significantly.
- We have created a new country variable after mapping the city and country columns

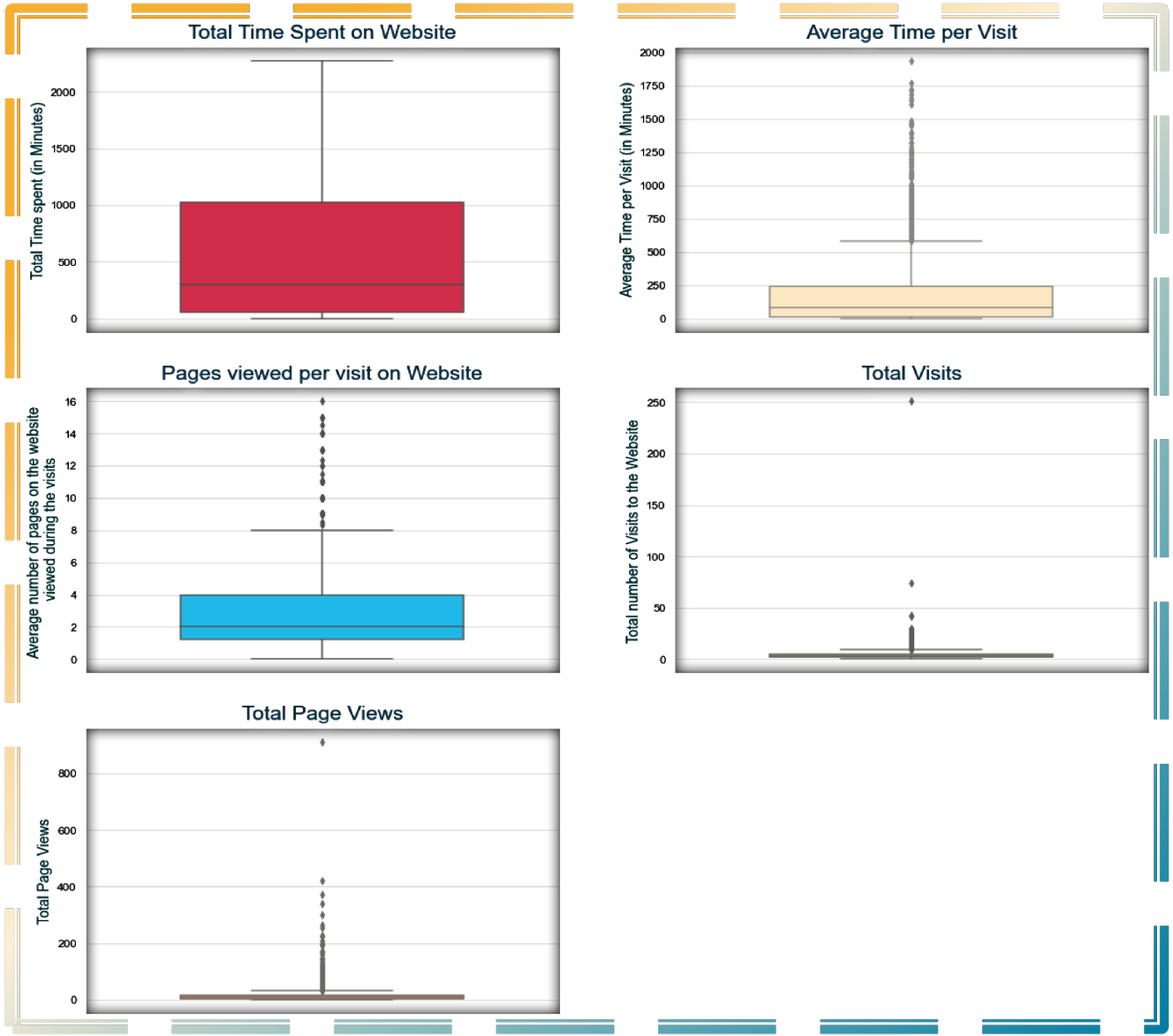
Data Cleaning & Preparation

- We did some feature engineering to specify average time per visit by dividing the total time spent on website by total visits to website.
- Another variable that we defined using feature engineering is total page views obtained dividing page views per visit by total visits to website.
- We subset data based on some important columns like occupation and total time visit
- We mapped some variable like 'lead_source', 'last_activity', 'what_is_your_current_occupation', 'tags' and 'last_notable_activity' to make data more lucid.
- Lastly, we dropped other unnecessary variables like 'prospect_id','country' and 'lead_number' for better analysis of data in this step.

Outlier Detection

- To determine the outliers in numerical variables, we plotted a subplot of boxplots.
- After carefully analysing, we did not feel to do outlier treatment for numerical columns.

Data Cleaning & Preparation



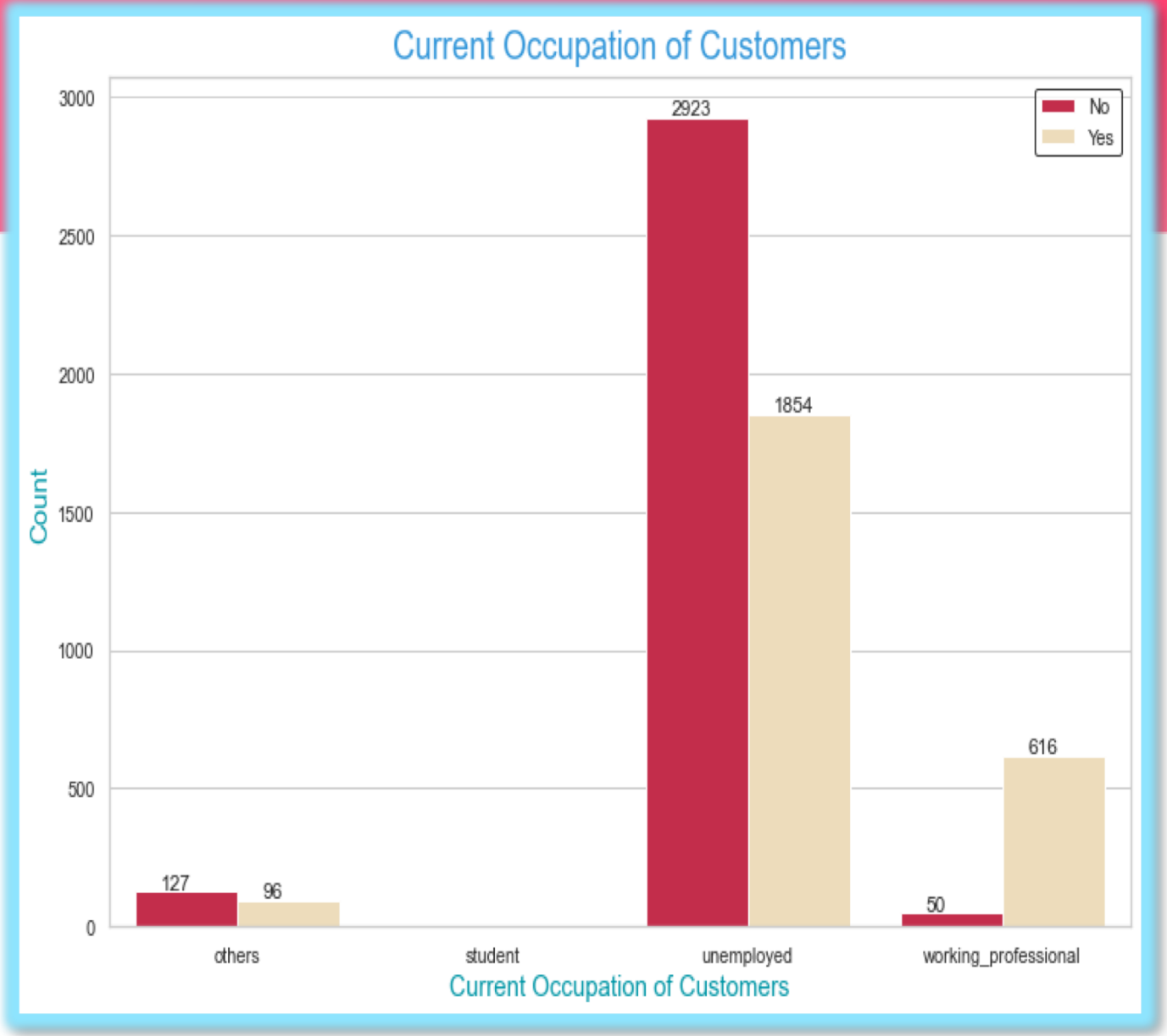
3

Perform Exploratory Data Analysis

Perform Exploratory Data Analysis

Most customers who comes to X-education are unemployed and many are possible leads.

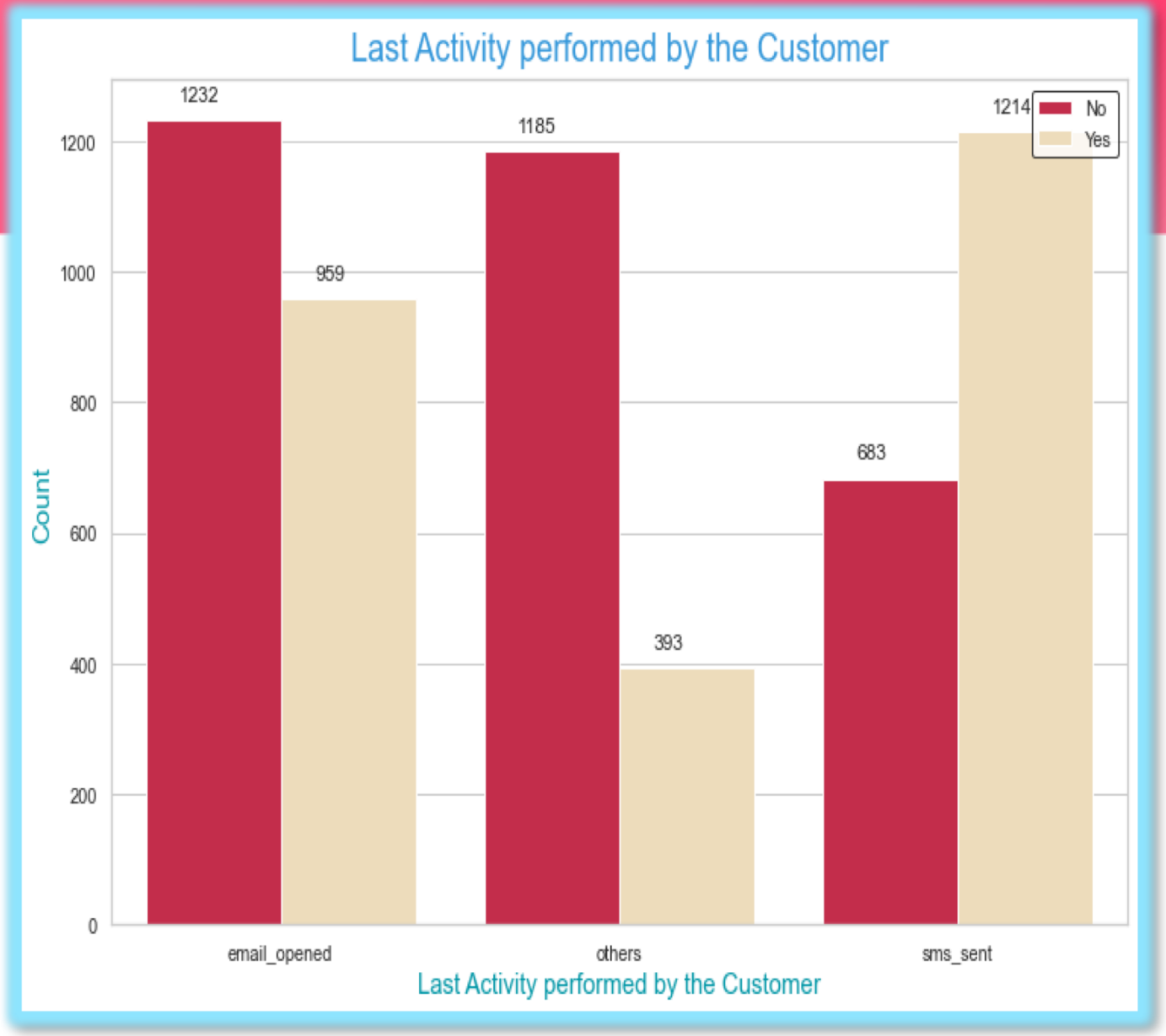
Customers of X-educations are not students.



Perform Exploratory Data Analysis

The last activity performed by the customers tends to churn most students through the SMS sent to student.

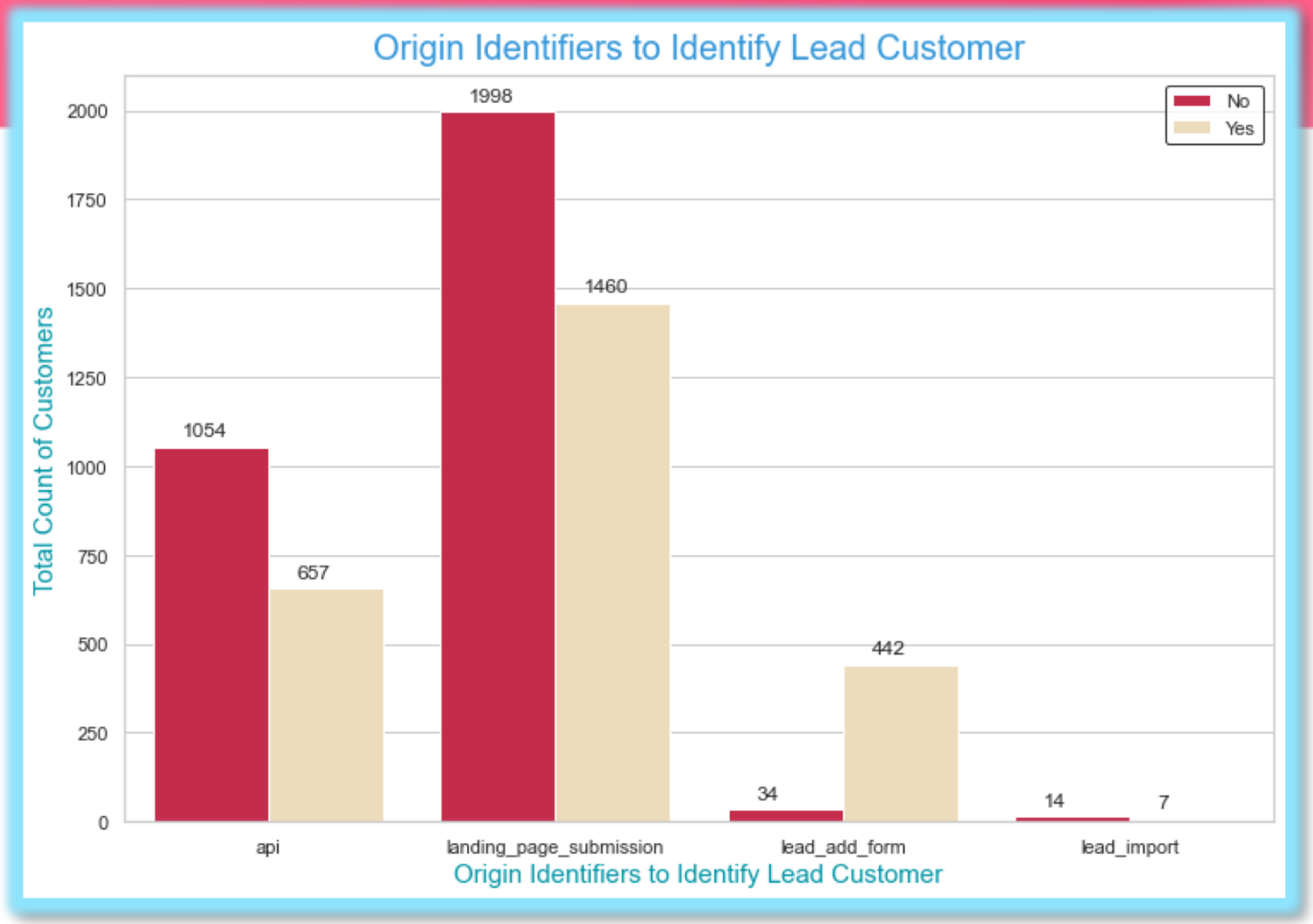
Emails opened by the customers also show the possibility whether the customer will convert into a lead or not.



Perform Exploratory Data Analysis

Mostly the lead customers are identified by the landing page submission whether they will convert into a positive lead or not.

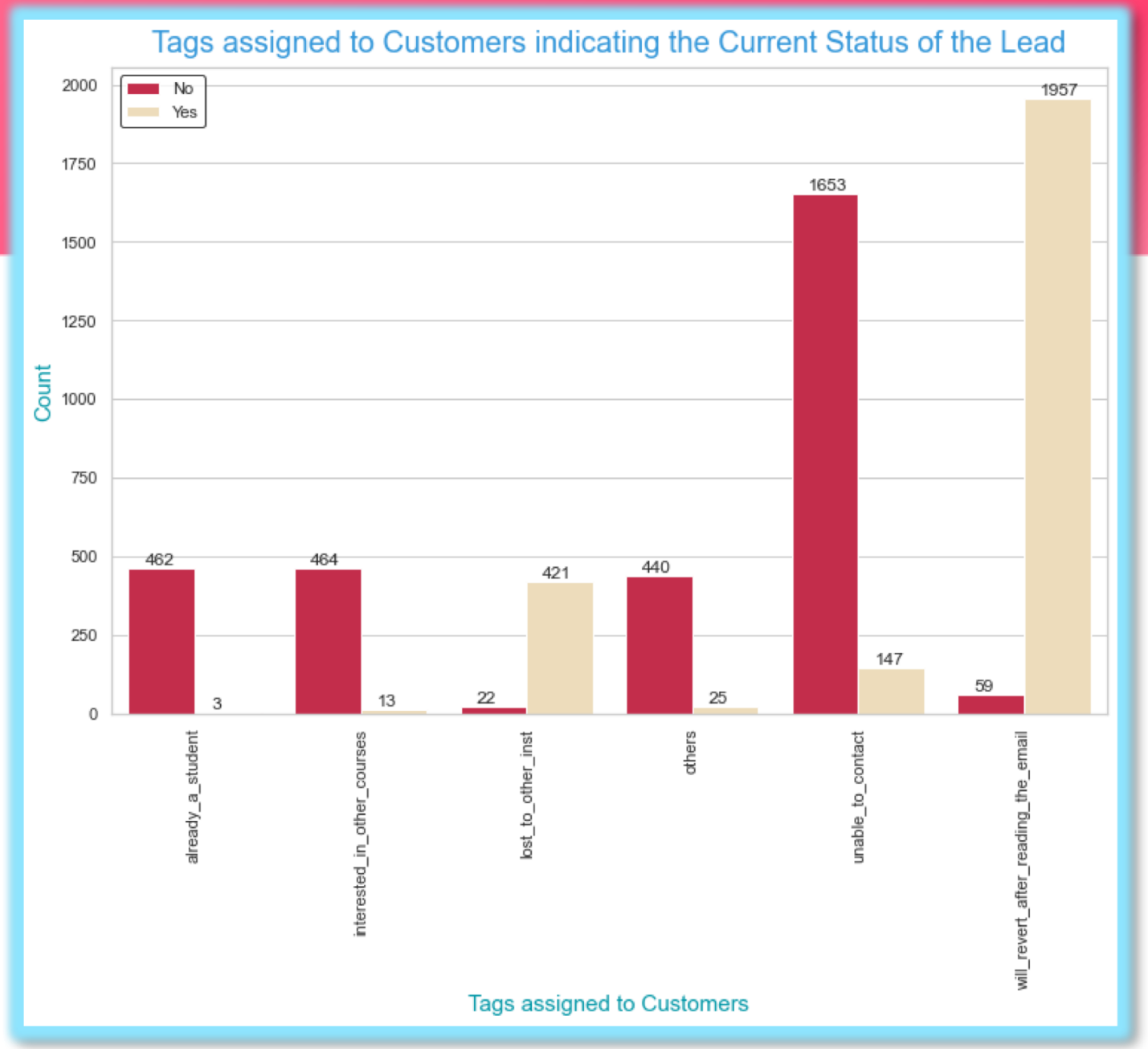
Secondly, ‘api’ is the another way to identify the possible lead.



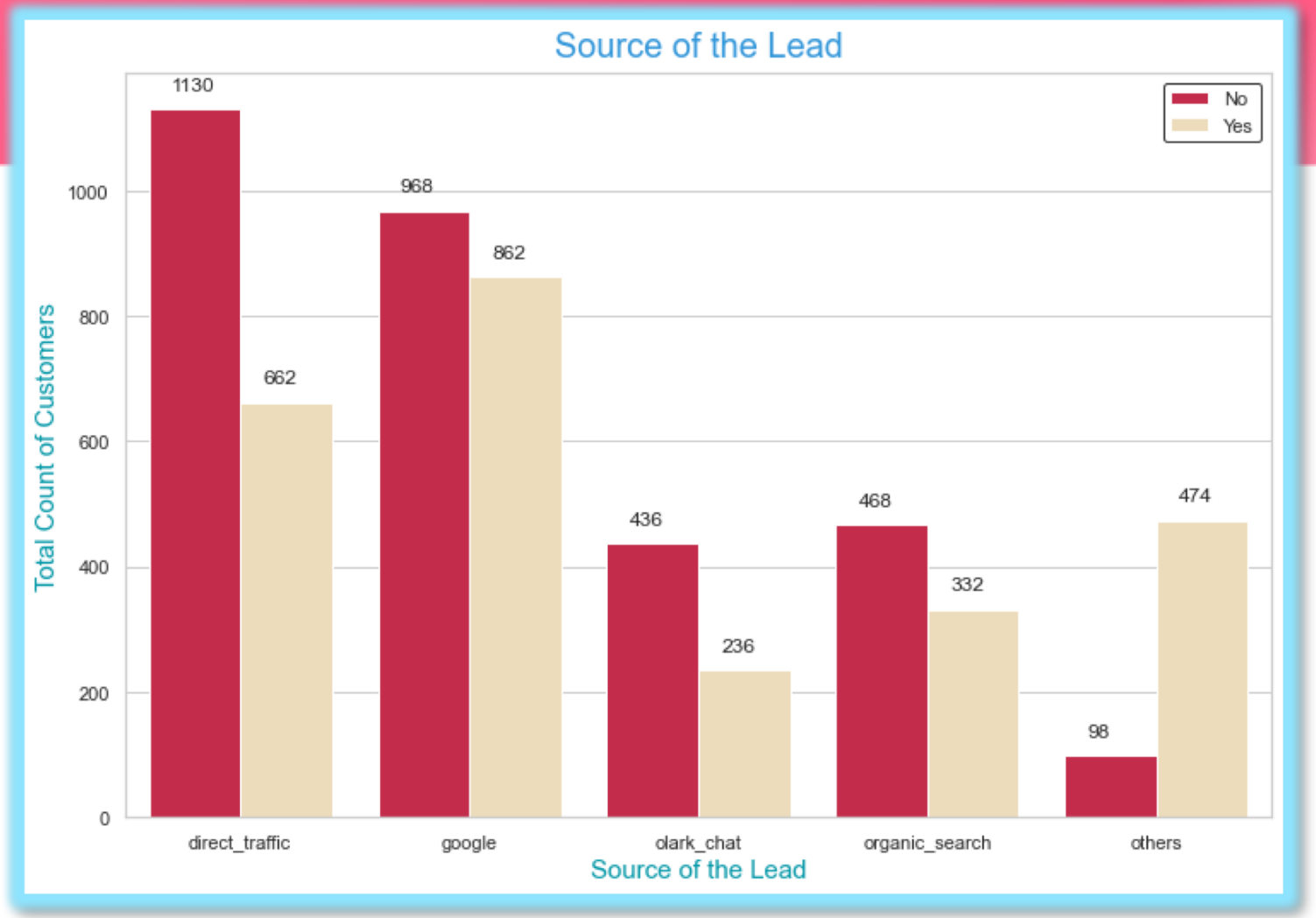
Perform Exploratory Data Analysis

Maximum number of customers who are tagged as 'will revert after reading the email' gets converted into a lead.

Maximum number of customers who are tagged as 'unable to contact' do not get converted into a lead.



Perform Exploratory Data Analysis



Most leads come via direct traffic and Google respectively and both sources have potential to bring traffic to the X-education website.

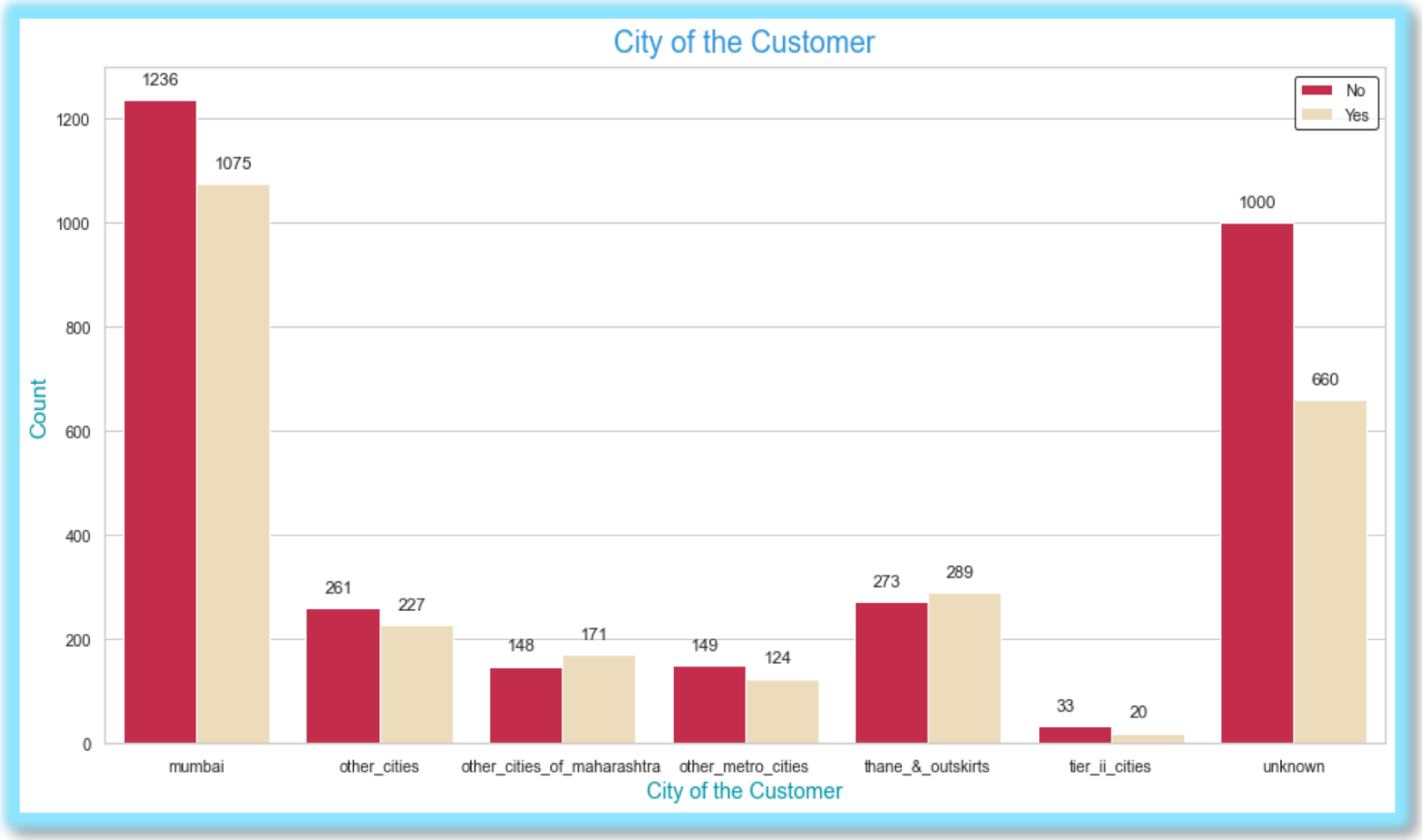
Other sources like dark chat and organic search also contribute to X-education.

Perform Exploratory Data Analysis

Most customers of X-education belongs to Mumbai and only approximately half of the customers those who belong to Mumbai convert into a lead.

Many customers of X-education come from other cities of Maharashtra, metro cities, etc.

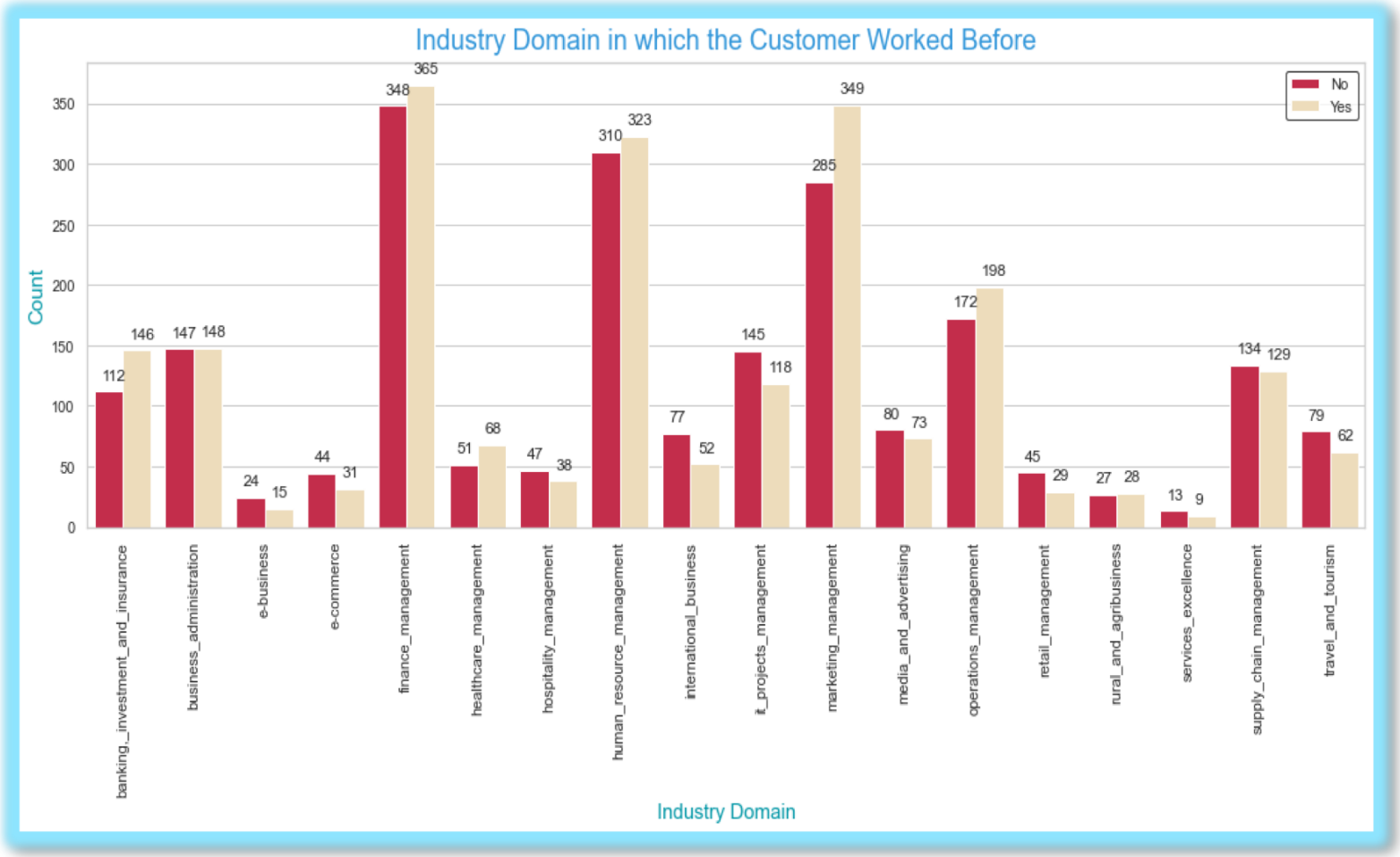
There are good number of customers whose city is unknown for both leads and non-leads.



Perform Exploratory Data Analysis

Out of the most leads who got converted, majority are from finance management, human resource management and marketing management specialisation respectively.

Customers from other domains such as banking investment and insurance, business administration and operation management also contribute to X-education business.



4

Data Preparation for Model Building

Data Preparation for Model Building

After data cleaning we can see the class is balanced with 55:45 ratio of No and Yes

We did one hot encoding for categorical and normalization (min-max scaling) for numeric variables.

We split the train and test data with 70:30 ratio.

5

Feature Selection & Baseline Model Building

Feature Selection & Baseline Model Building

We had total 51 features and we have finalized 25 features using Recursive Feature Elimination (RFE).

After feature selection, we built the baseline model (Generalized Linear Model, GLM) with all 25 features.

Then based on the p-value and VIF we have removed few variables and finally we are left with 10 features with p-value less than 0.05 and VIF less than 5.

Also, we got the cross validation mean accuracy of the model 0.952 which is quite good.

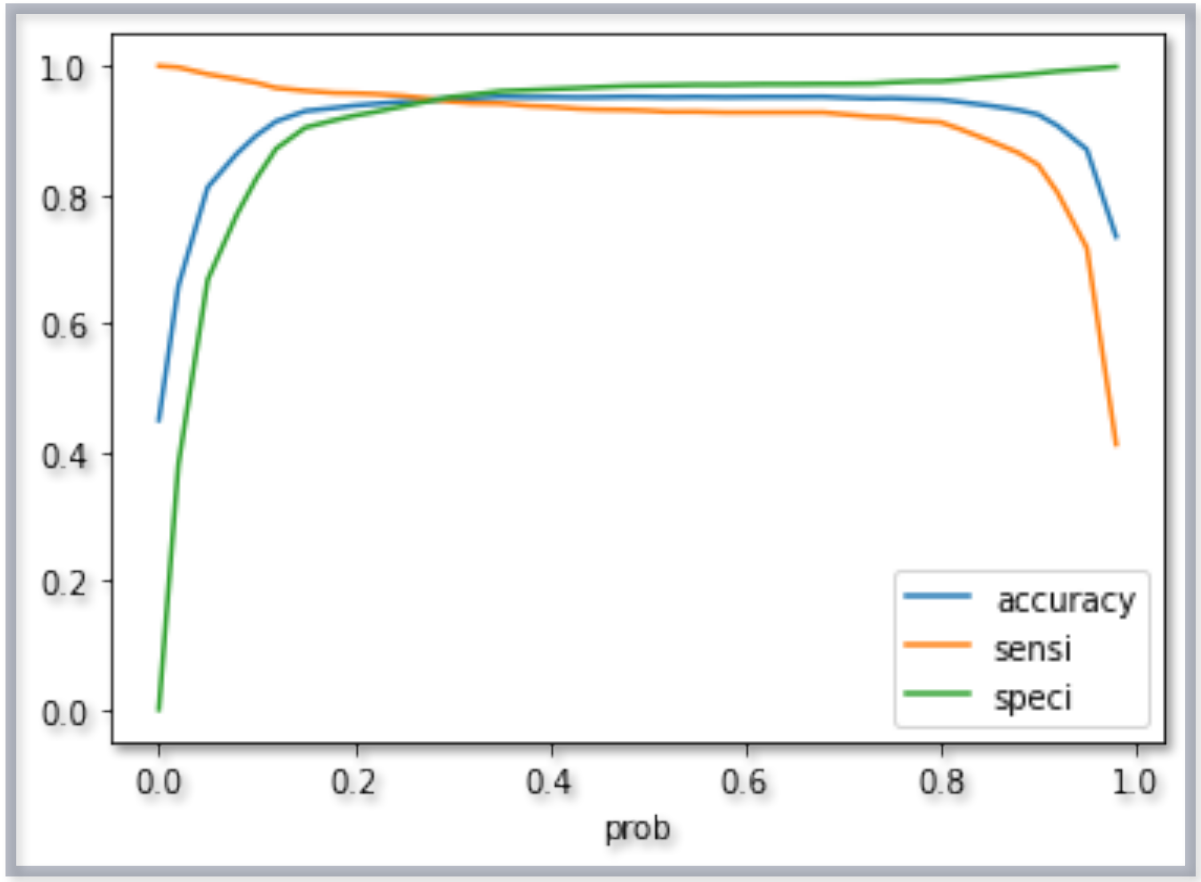
6

Prepare & Evaluate the best Model on Train Data

Prepare
&
Evaluate the best Model on Train Data

Accuracy & Cut-Off

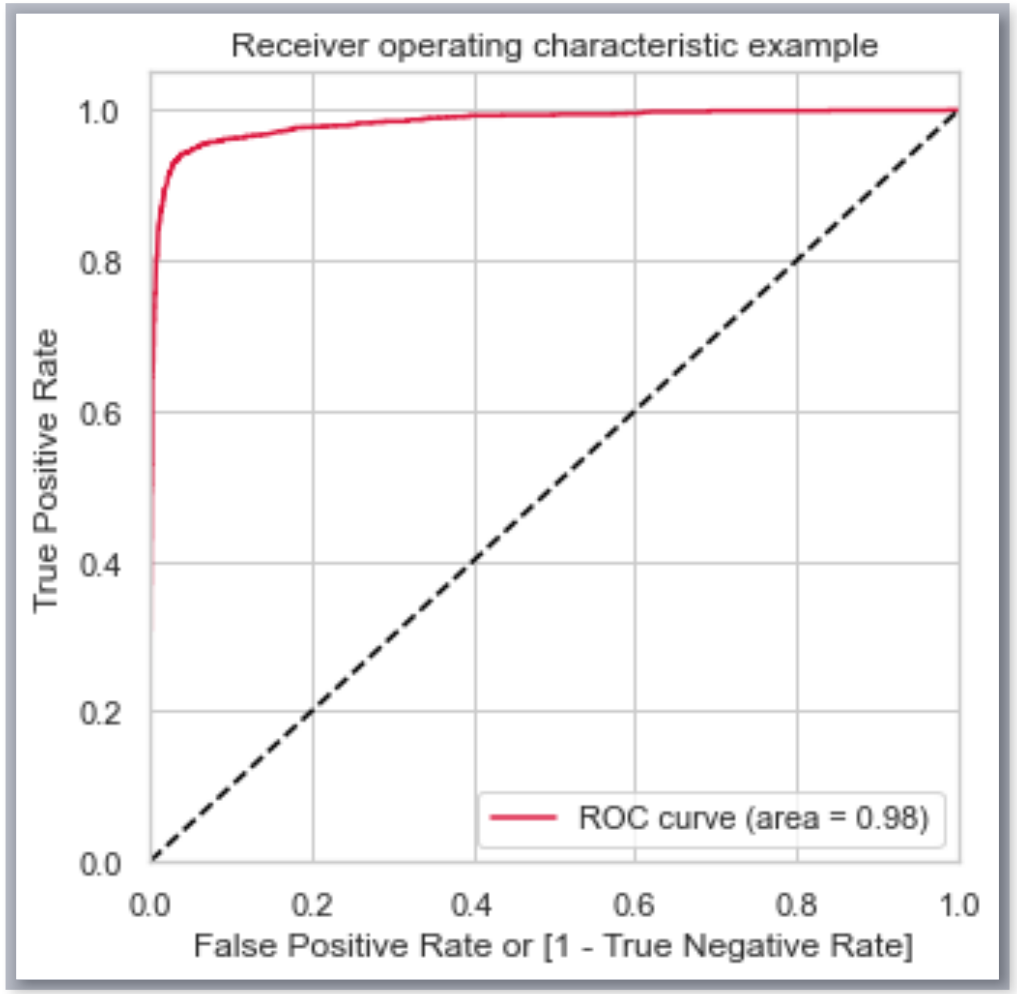
- **Cross Validated Mean Accuracy of Model:** 0.952
- **Optimum Cut-Off value:** 0.28



Prepare
&
Evaluate the best Model on Train Data

Evaluation Matrix Results

- **Accuracy:** 0.94730
- **Sensitivity (Recall):** 0.94843
- **Specificity:** 0.94637
- **Precision:** 0.93532
- **F1-Score:** 0.94183



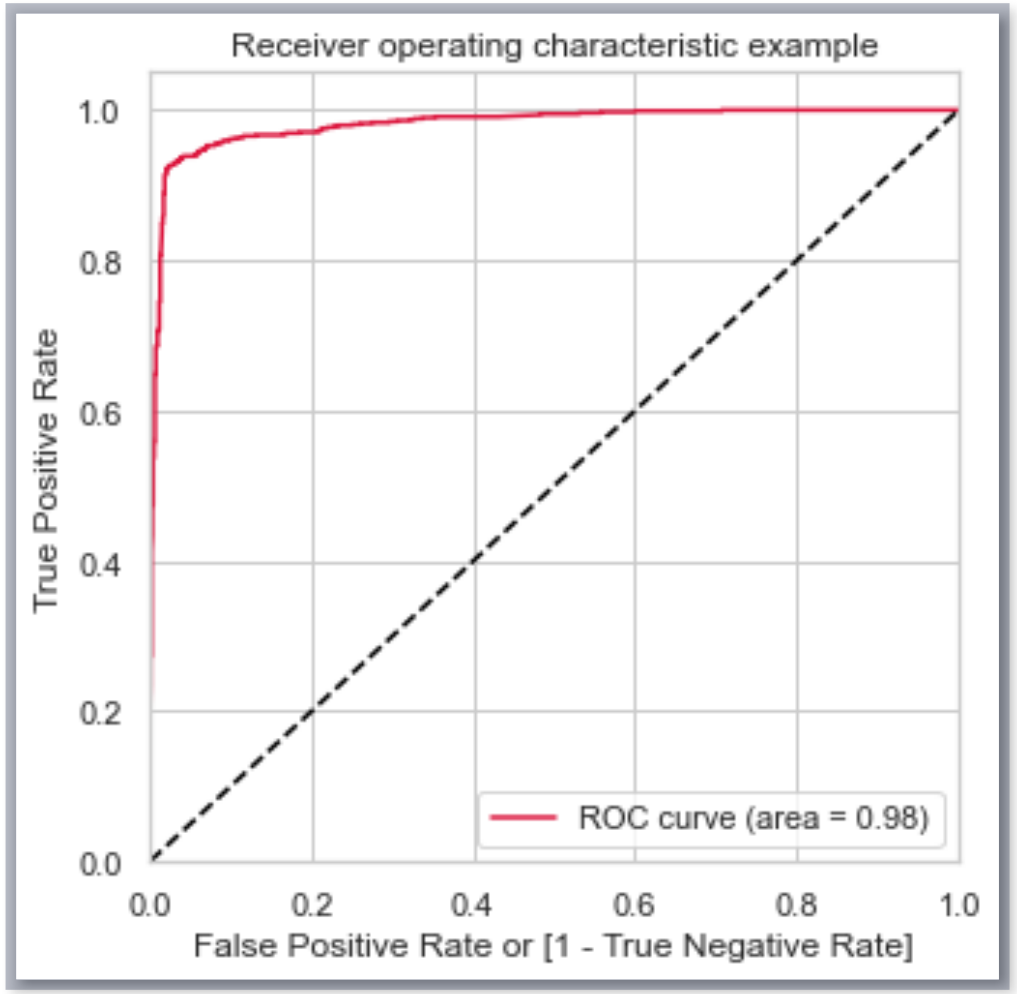
7

Check Model Performance on Test Data

Check Model Performance on Test Data

Evaluation Matrix Results

- **Accuracy:** 0.94470
- **Sensitivity (Recall):** 0.93861
- **Specificity:** 0.94989
- **Precision:** 0.94102
- **F1-Score:** 0.93982



8

Summary of Model Outcome

Summary Of the Model Outcome

According to the final model, the top five predictor factors that influence the target variable ‘converted’ are:

1. tags_lost_to_other_inst
2. tags_will_revert_after_reading_the_email
3. total_time_spent_on_website
4. lead_origin_lead_add_form
5. specialization_rural_and_agribusiness

Thank You