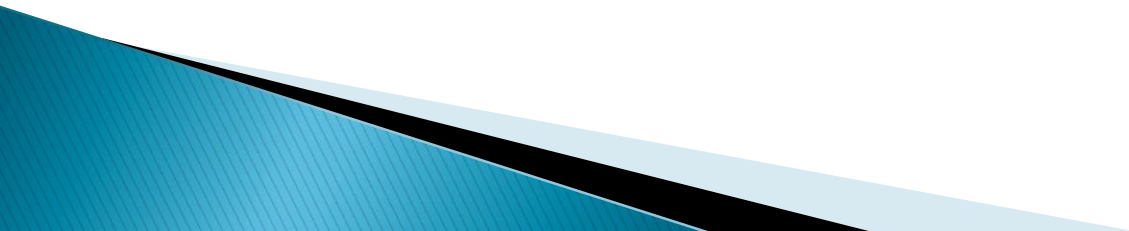


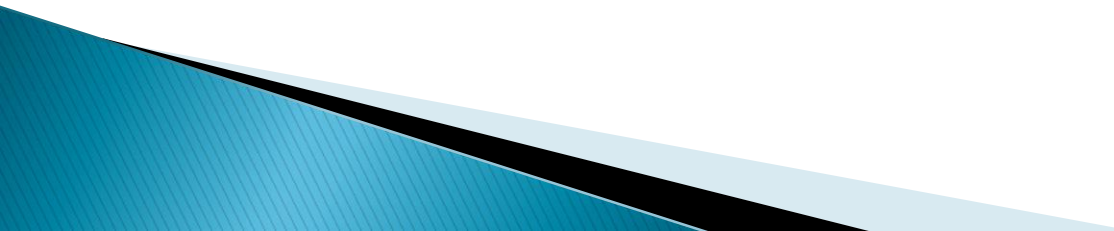
# **Lead Scoring Case Study**



# Problem statement

X Education company sells online courses to industry professionals. The company markets its courses on several websites. The company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, the typical lead conversion rate at X education is around 30%, which is very poor.

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.



# Aim

Building a a logistic regression model wherein we need to assign a lead score between 0 and 100 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%.

# Approach

**Step1:** Reading, Understanding Data and inspecting the data.

**Step2:** Data Cleaning:

**Step3:** Data Transformation: Changed the binary variables into '0' and '1'

**Step4:** Dummy Variables Creation: **Step5:** Test Train Split: The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

**Step6:** Feature Rescaling

**Step 7:** Model Building:

- Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.

- Using the statistics generated, we recursively tried looking at the P-values along with VIF in order to select the most significant values that should be present and dropped the insignificant values.

- Finally, we arrived at the 14 most significant variables.

- The VIF's for these variables were also found to be good.

- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 97% which further solidified the of the model.
- - We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.29.
- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics

Step 8: Conclusions



# Handling "SELECT" values

```
# List of columns having 'Select' as value  
columns_with_select = [col for col in leads_df.columns if len(leads_df[col].isin(["Select"]).unique())>1]  
print(columns_with_select)
```

```
['Specialization', 'How did you hear about X Education', 'Lead Profile', 'City']
```

Above three are the columns having "SELECT" values, which are being replaced nan

```
: # Converting 'Select' values to NaN.  
leads_df = leads_df.replace("Select", np.nan)
```

```
# Calculating Percentage of values more than 40 % missing  
(leads_df.isnull().mean()*100)[leads_df.isnull().mean()*100>40]
```

```
How did you hear about X Education    78.463203  
Lead Quality                          51.590909  
Lead Profile                          74.188312  
Asymmetrique Activity Index           45.649351  
Asymmetrique Profile Index            45.649351  
Asymmetrique Activity Score           45.649351  
Asymmetrique Profile Score            45.649351  
dtype: float64
```

Identified and dropped the columns with 40% missing values

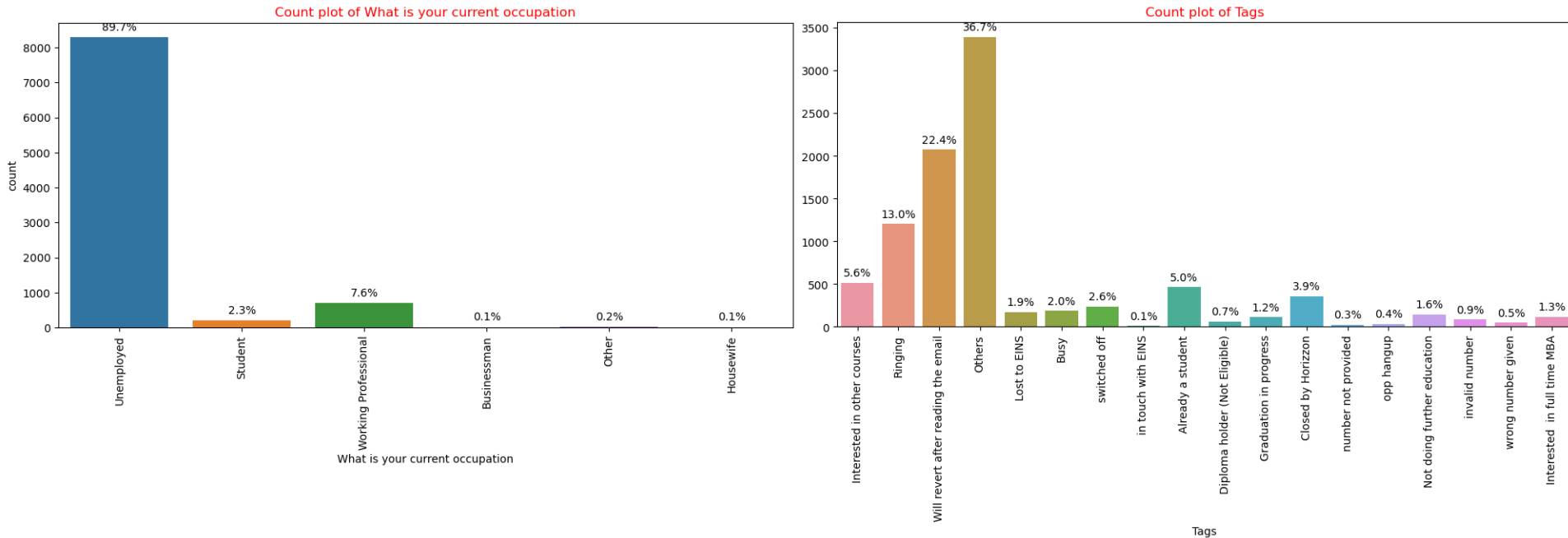
# Handling Missing values of categorical variables

- City:** This column has 39.71 % missing values. One solution is we can impute the missing values with mode value i.e "Mumbai", but doing this will make the data more skewed, which will create problem in model, so, best solution is dropping this column.
- Specialization:** This column has 36.58 % missing values, however the values are evenly distributed. Hence imputation or dropping is not a good choice. Instead we are creating additional category called 'Others'.
- Tags:** Tags has 36.29 % missing values. Tags are assigned to customers indicating the current status of the lead. This column is important in modeling, so instead of imputing the values, we are creating additional category for this column as well called 'Others'.
- What matters most to you in choosing a course:** This variable has 29.32 % missing values. 99.95% customers have selected 'better career prospects'. This is massively skewed and will not provide any insight, so dropping it will be the better choice.
- What is your current occupation:** This variable has 29.11 % missing data. Dropping this variable would not be a good choice. We can impute the missing values with 'Unemployed' as it has the most values. This seems to be an important variable from business context, since X Education sells online courses and unemployed people might take this course to increase their chances of getting employed.
- Country:** This column has 26.63 % missing values and 95.77 % of the customers are from India. Since X Education sells online courses. We either can impute the missing values with "India" or we can drop this column. As, imputing this with India does not make business sense. Hence we will drop this variable.
- Last Activity:** This variable has only 1.11% missing values and "Email Opened" is having highest number of values, hence we will impute the missing values with label 'Email Opened'.
- Lead Source:** Lead Source is having 0.39% missing values and "Google" is having highest number of occurrences, hence we will impute the missing values with label 'Google'.

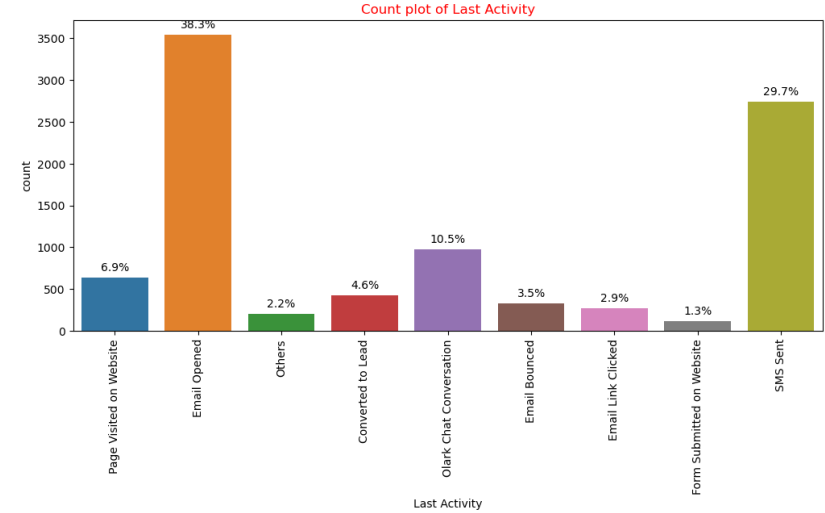
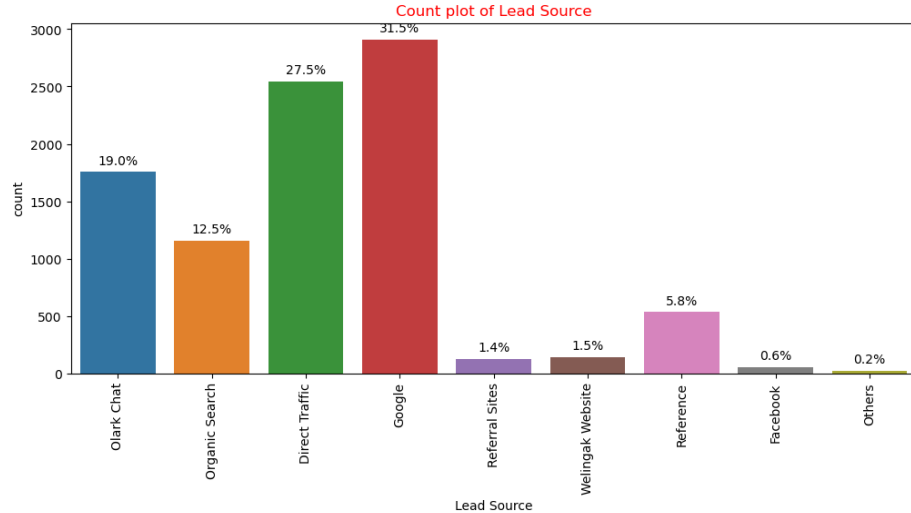


# EDA

# Univariate analysis



- Current\_occupation: It has 89.7% of the customers as Unemployed
- Tags: Apart from others, will revert after reading the email contributes 22.4 %

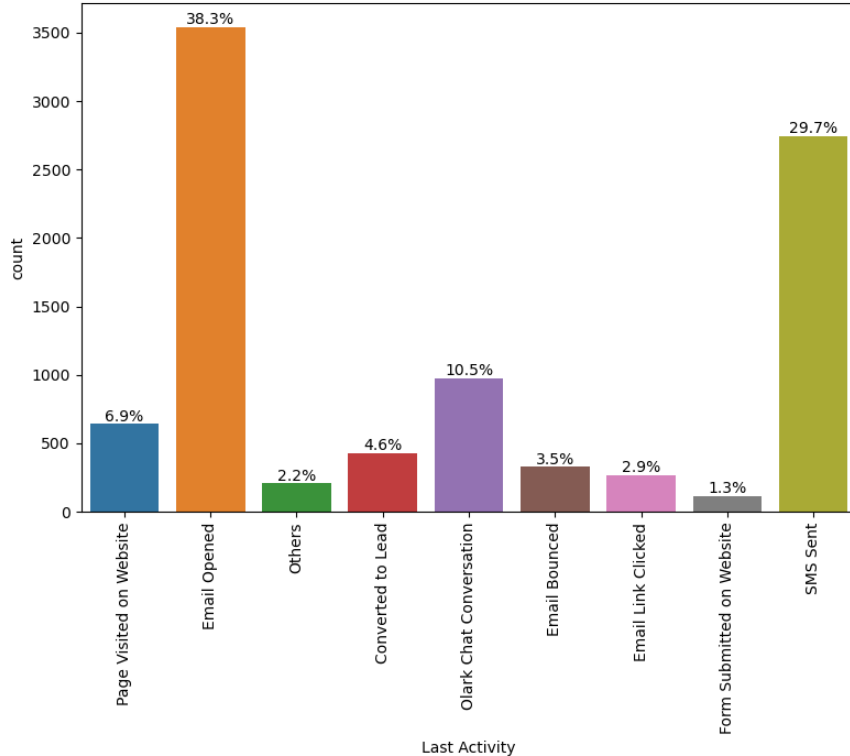


- Lead Source: 31.5% Lead source is from Google, followed by Direct Traffic as 27.5%
- Last Activity: 38.3% of customers opened email as last activity followed by SMS Sent as 29.7%

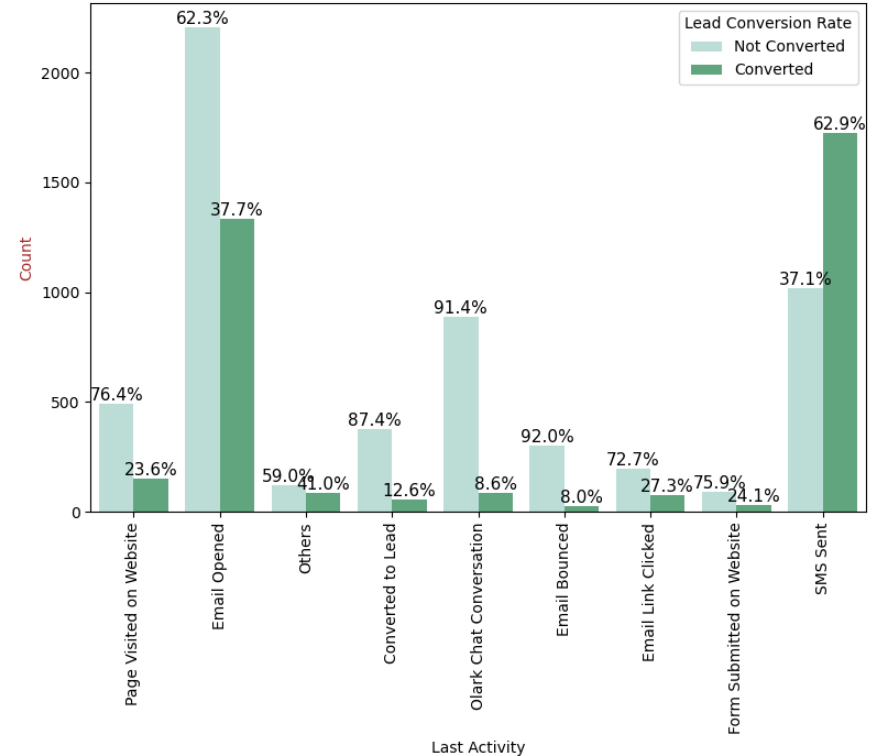
# Bivariate analysis

## Last Activity Countplot vs Lead Conversion Rates

Distribution of Last Activity

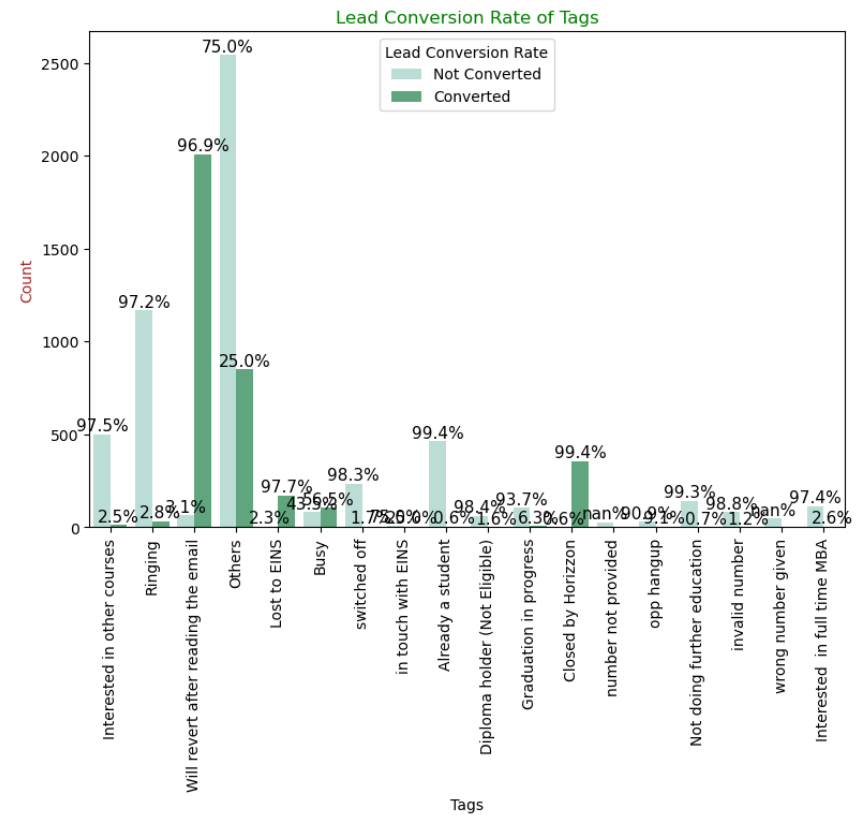
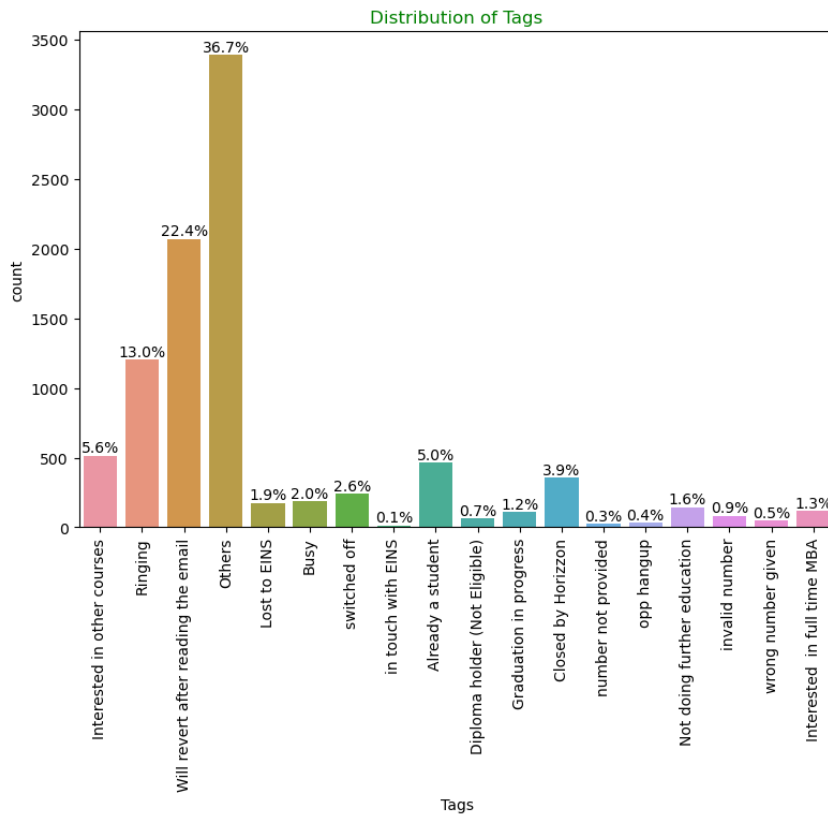


Lead Conversion Rate of Last Activity



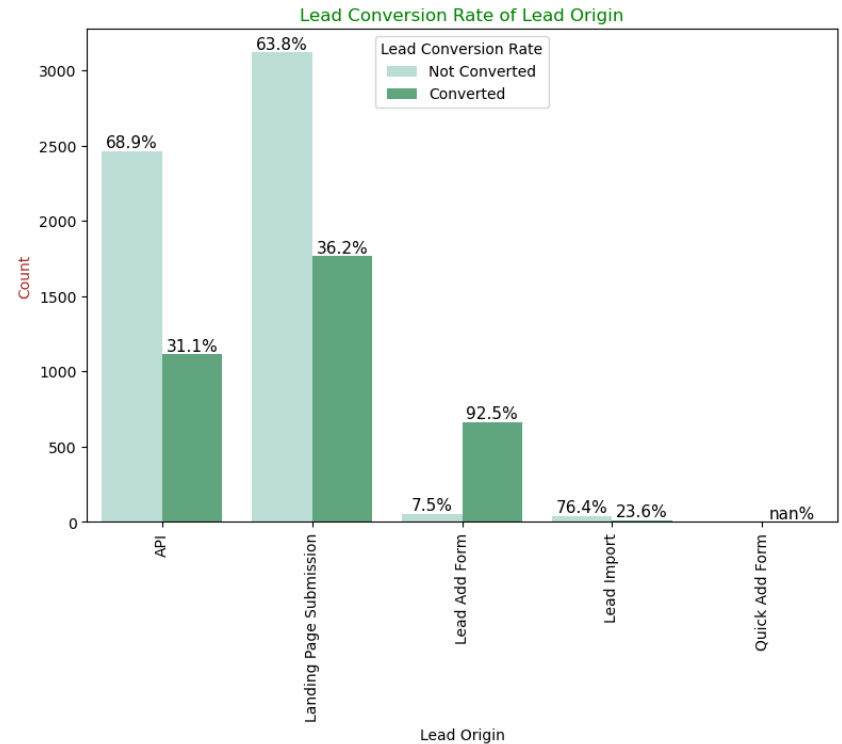
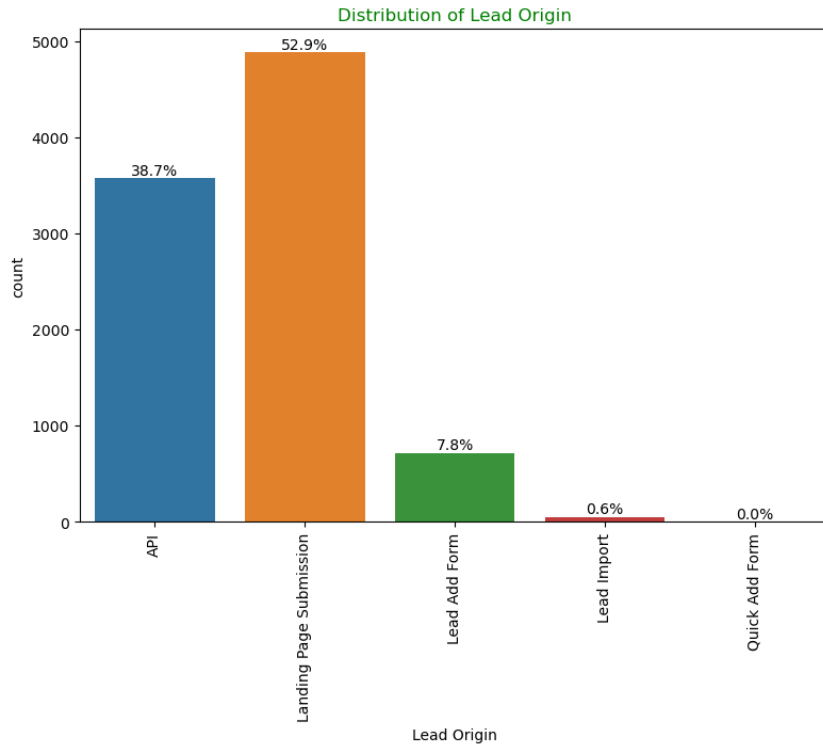
SMS Sent has 62.9% lead conversion rate with 29.7% contribution from last activities and Email Opened activity contributed 38.3% of last activities performed by the customers with 37.7% lead conversion rate.

## Tags Countplot vs Lead Conversion Rates



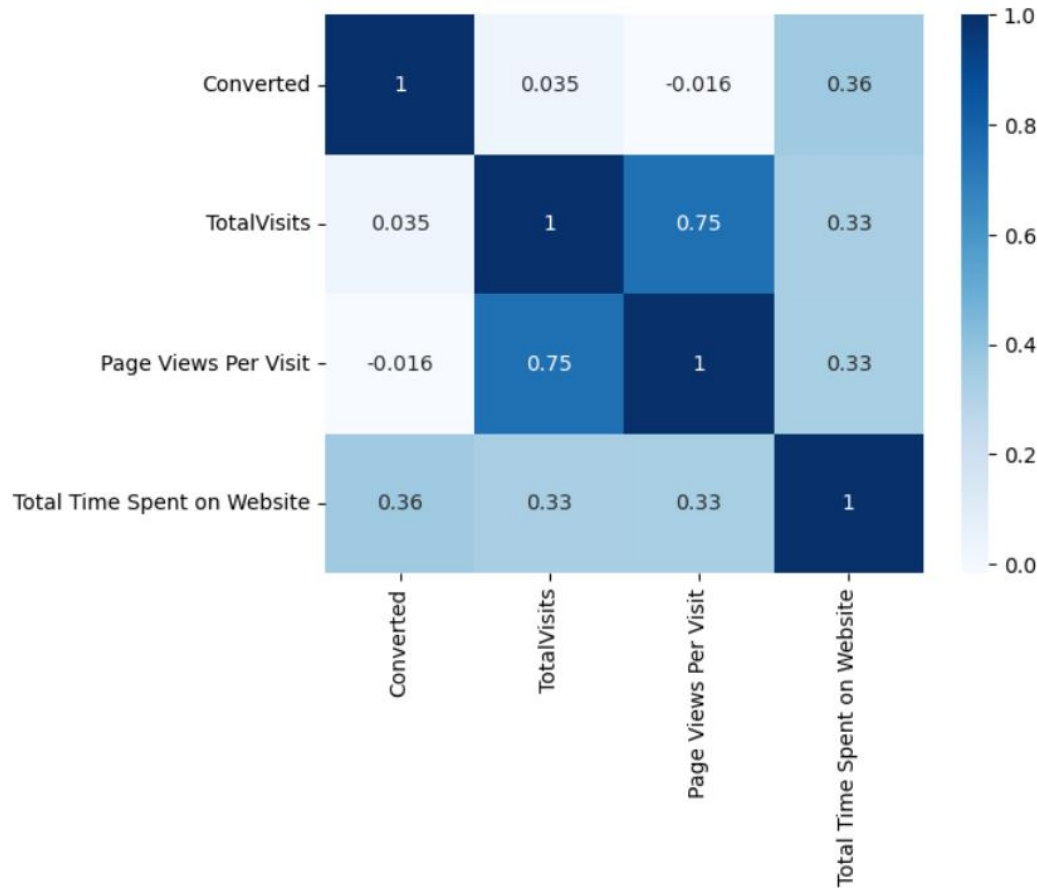
22.4% customers mentioned that they will revert after reading the email and 96.9 % of them converted into leads.

## Lead Origin Countplot vs Lead Conversion Rates



52.9% of leads originated from **Landing Page Submission** with a lead conversion rate of 36.2% .The **API** identified 38.7% of customers with a lead conversion rate of 31.1%.

# Bivariate Analysis for Numerical Variables



Total Time spent on Website has strong positive relation with lead conversion

# Top 20 feature selected using RFE

```
: # columns which are selected by RFE
rfe_col = X_train.columns[rfe.support_]
rfe_col

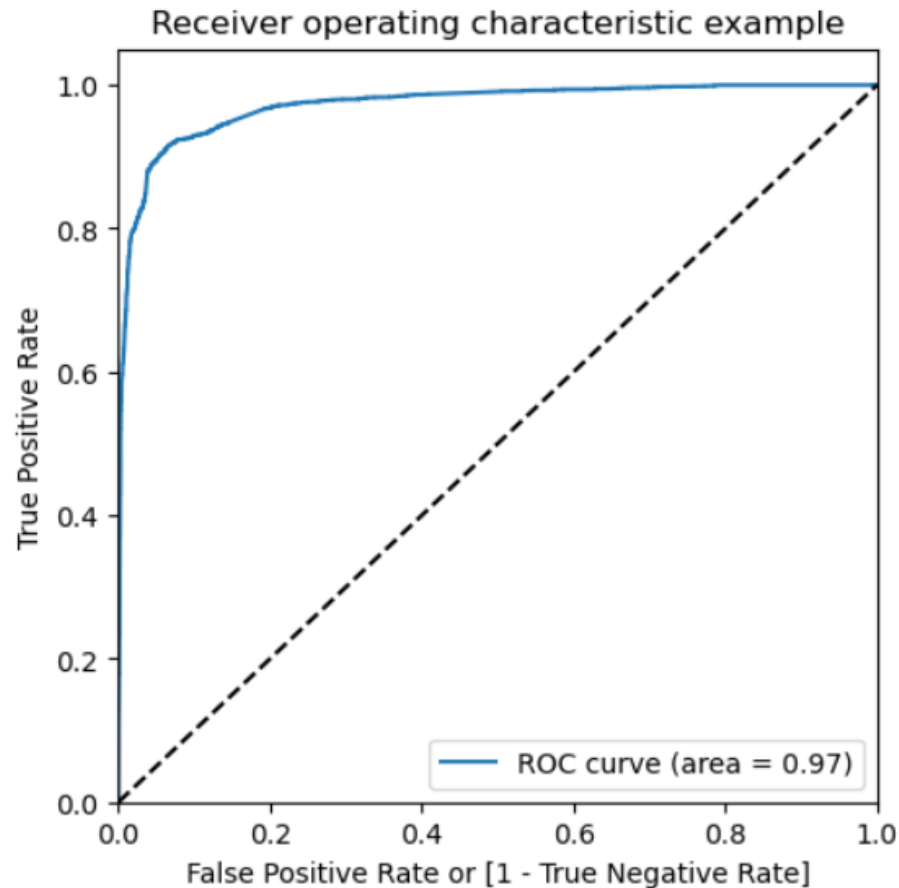
: Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
      'Lead Origin_Lead Add Form', 'Lead Source_Welingak Website',
      'Last Activity_SMS Sent', 'Tags_Busy', 'Tags_Closed by Horizzon',
      'Tags_Lost to EINS', 'Tags_Others', 'Tags_Ringing',
      'Tags_Will revert after reading the email', 'Tags_in touch with EINS',
      'Tags_invalid number', 'Tags_number not provided', 'Tags_switched off',
      'Tags_wrong number given',
      'Last Notable Activity_Had a Phone Conversation',
      'Last Notable Activity_Modified',
      'Last Notable Activity_Olark Chat Conversation',
      'Last Notable Activity_SMS Sent'],
      dtype='object')
```



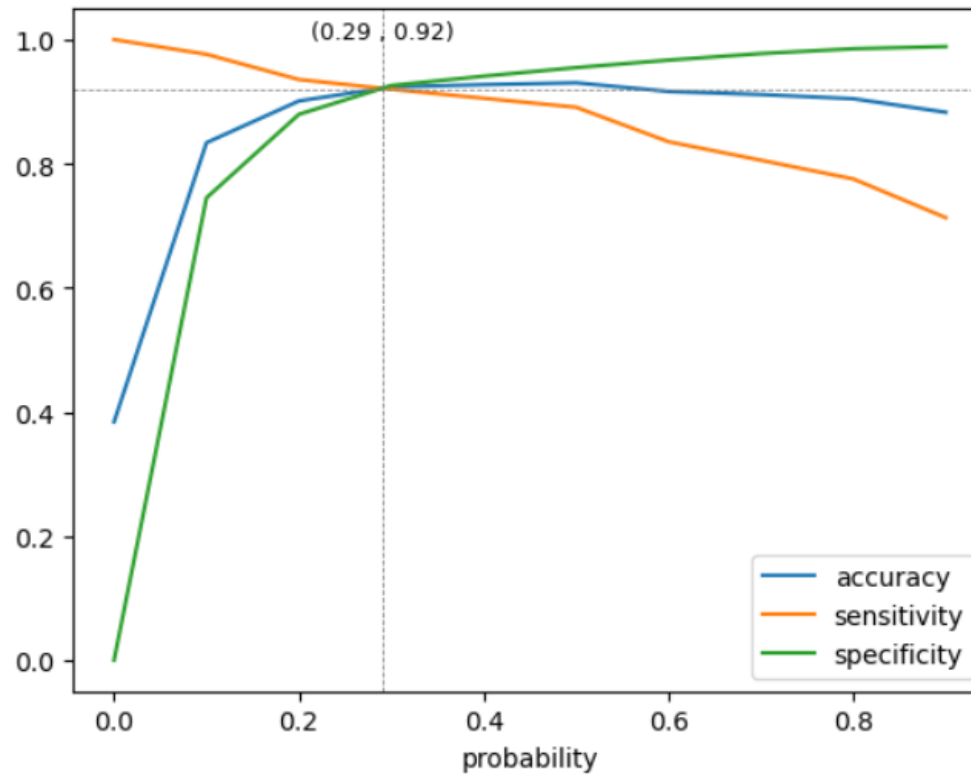
# Final 14 features after dropping based on p-value and VIF

	Features	VIF
0	Lead Origin_Landing Page Submission	2.07
1	Tags_Will revert after reading the email	1.77
2	Last Activity_SMS Sent	1.63
3	Tags_Others	1.61
4	Lead Origin_Lead Add Form	1.49
5	Last Notable Activity_Modified	1.47
6	Tags_Closed by Horizzon	1.31
7	Tags_Ringing	1.31
8	Total Time Spent on Website	1.21
9	Tags_Lost to EINS	1.08
10	Tags_Busy	1.07
11	Tags_switched off	1.07
12	Last Notable Activity_Olark Chat Conversation	1.04
13	Tags_in touch with EINS	1.00

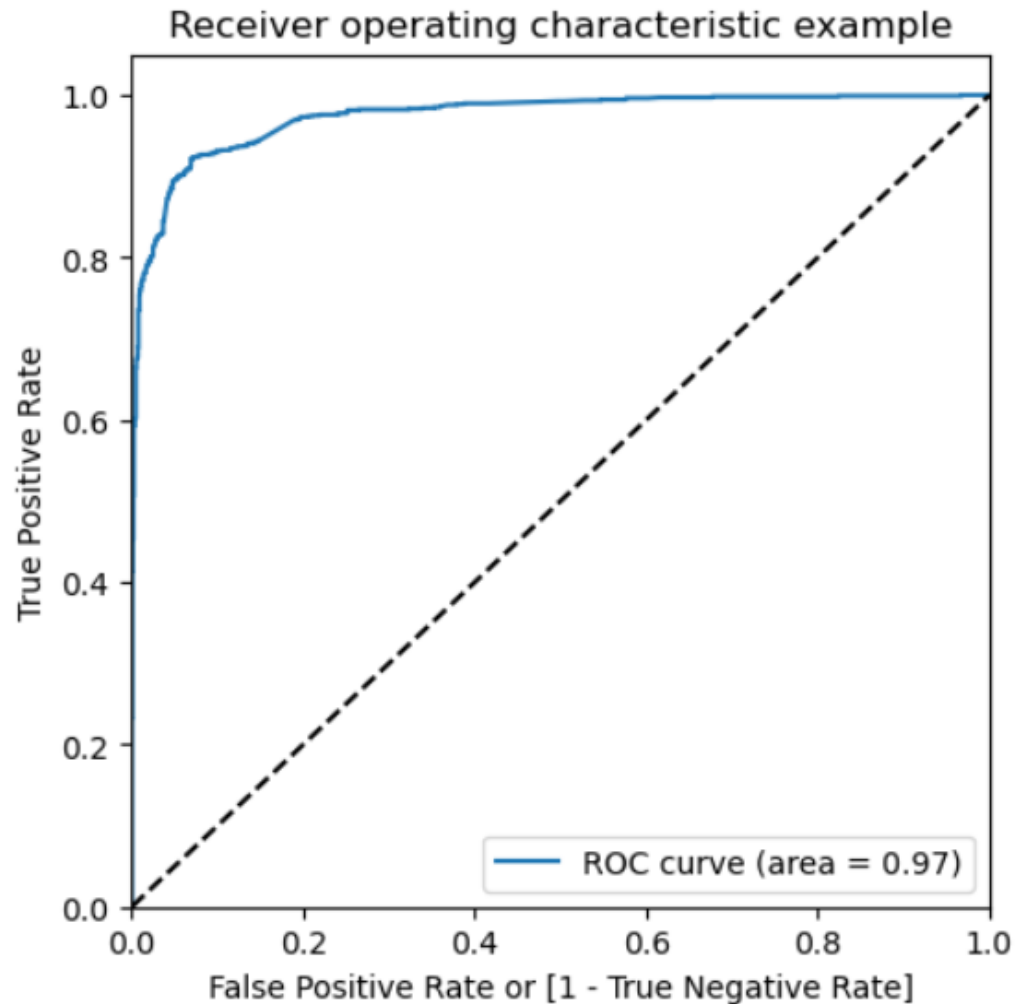
# ROC- curve for train set, with area under curve 0.97



**Optimal Cutoff Point/ Probability** where we get balanced sensitivity and specificity is 0.29



ROC– curve for test set, with area under curve 0.97



# Conclusions

## Train Data Set:

- Accuracy: 92.32%
- Sensitivity: 92.11%
- Specificity: 92.44%

## Test Data Set:

- Accuracy: 92.42%
- Sensitivity: 92.48%
- Specificity: 92.39%

The evaluation metrics are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

- The model achieved a `sensitivity of 92.11%` in the train set and 92.48% in the test set, using a cut-off value of 0.29.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- `The CEO of X Education had set a target sensitivity of around 80%, but we have achieved it as 92.11`
- The model also achieved an accuracy of 92.42%, which is in line with the study's objectives.