# X Education - Lead Scoring Case Study

Detection of Hot Leads to optimize sales efforts, improving lead conversion rates for X Education

Team Members: Sagar Pawar, Sachin Bhatnagar, Saikat Pal

# Table of Contents


- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building Model Evaluation
- Recommendations


# 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 high potential leads, also known as Hot Leads
- Using this info, their sales team would optimize their efforts and focus on quality leads.

## 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 indicate a higher conversion chance and the customers with a lower lead score indicate a lower conversion chance.
- The CEO has given a ballpark target lead conversion rate to be around 80%.

# Suggested Ideas for Lead Conversion

## Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

## Better Communication

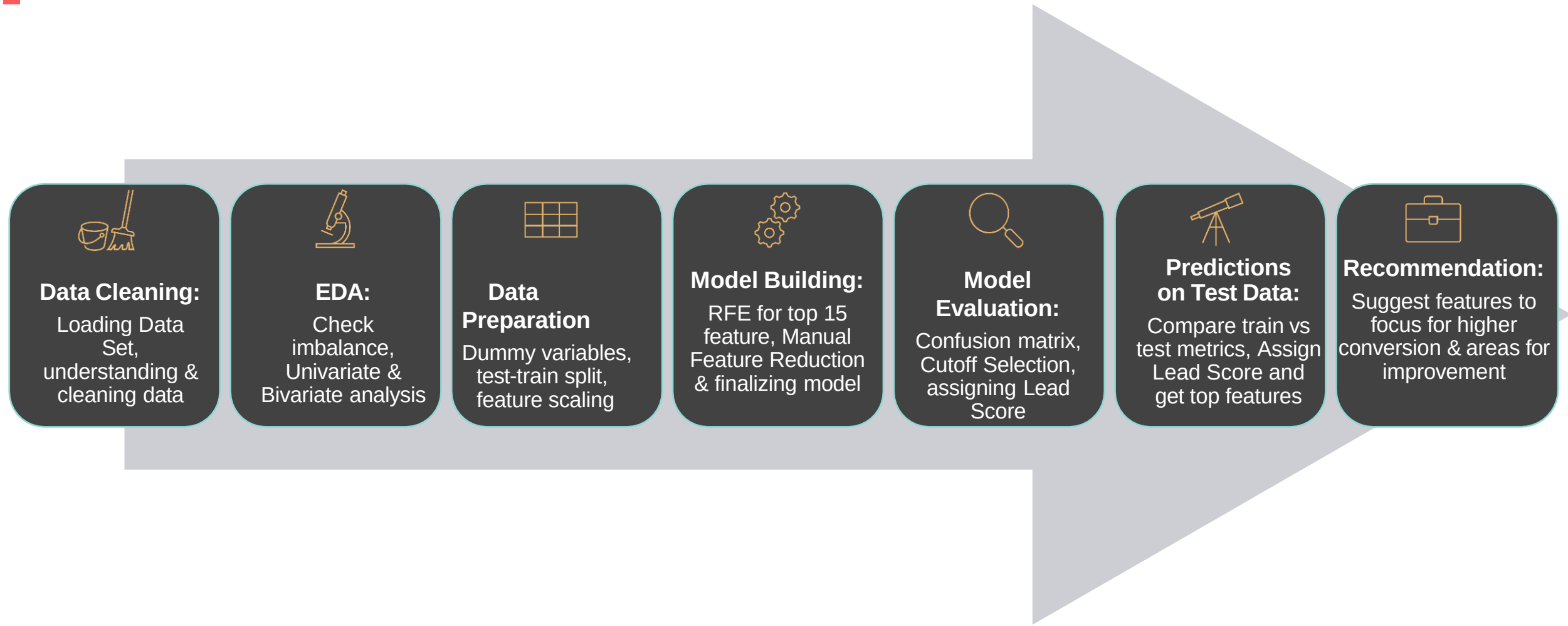
- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

## Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high **Recall** in obtaining hot leads.

# Analysis Approach



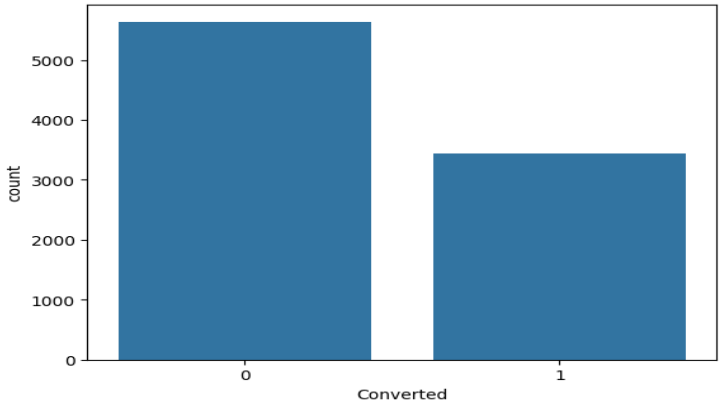
# Data Cleaning

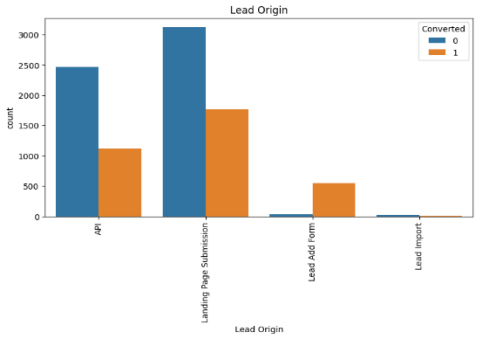
- Dropping the Prospect ID and retaining Lead Number for linking with Lead Score(0-100).
- Replace “Select” value to “NaN”
- Drop columns having only 1 value as they do not have any analytical value(ex: 'Magazine', 'Receive More Updates About Our Courses', 'I agree to pay the amount through cheque‘, etc.).
- Drop column which have more then 45% null value
- We replace “Nan” to "not provided“, because if we drop the NaN column, we lose a lot of information.
- Instead of individual countries, using buckets for the country column
- For country column, replacing missing values in a dataset with "India" because it's the most common non-missing value.

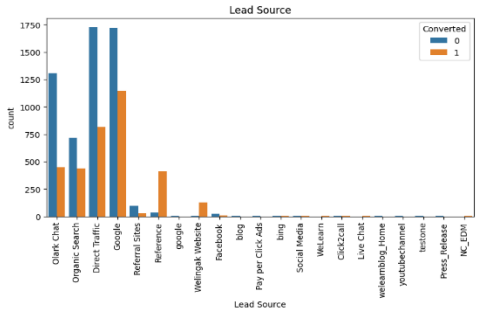
# EDA

## Univariate/Bivariate Analysis – Categorical Variables

Here we seem to have both the classes represented decently and there is no major class imbalance.

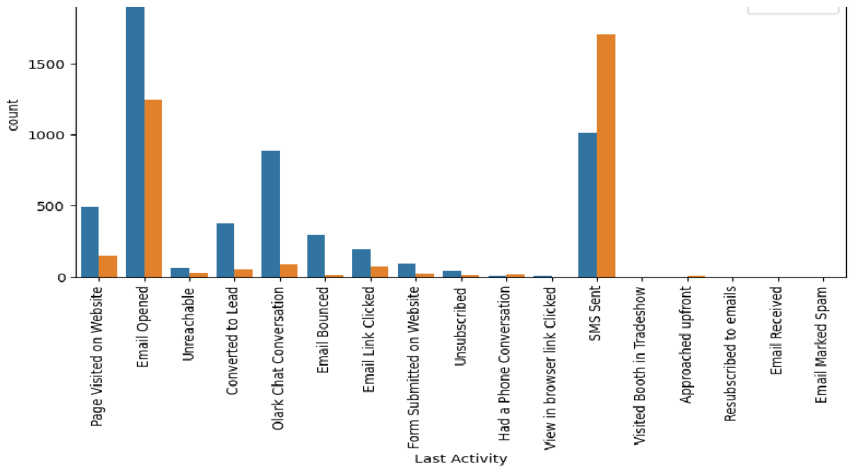


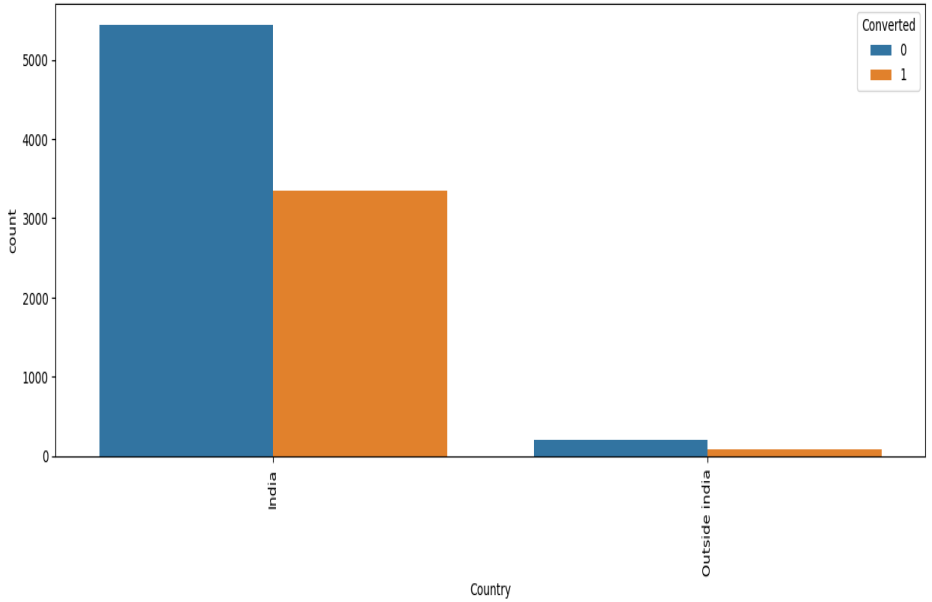

**Lead Source:** Most of the Lead source is from Google & Direct Traffic combined.

**Last Activity:** SMS Sent & Email Opened activities play a key role in Activity.



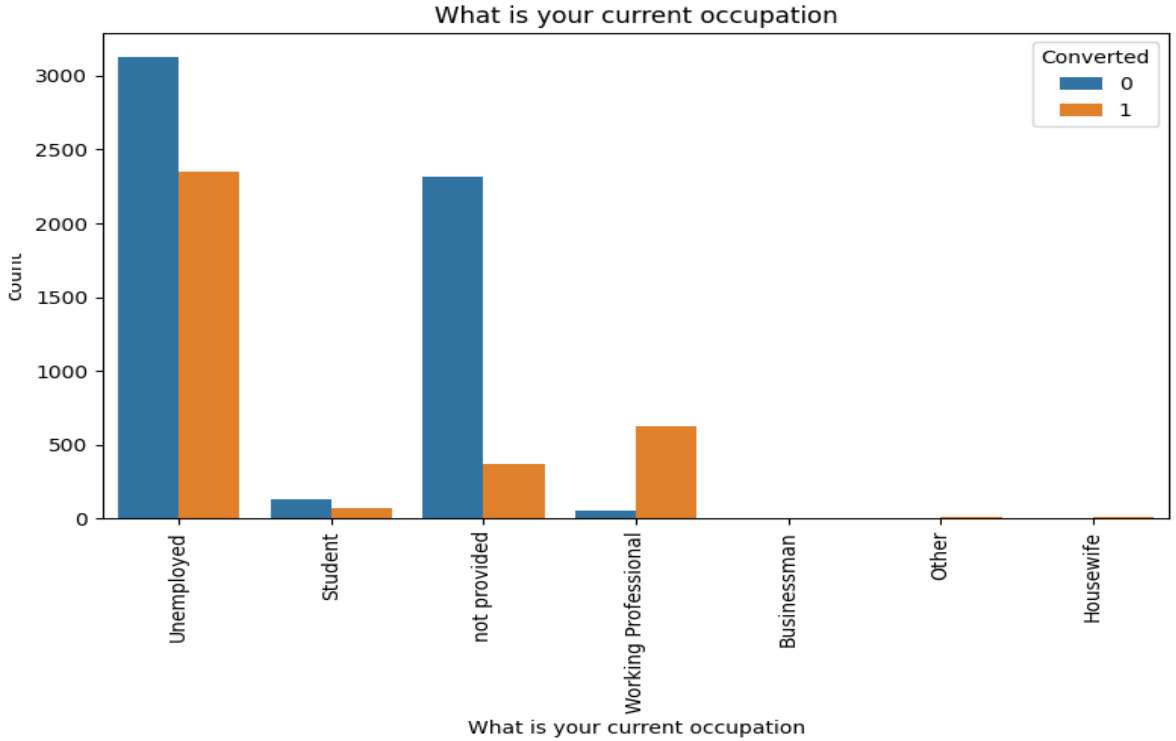
# EDA

## Bivariate Analysis – Categorical Variables



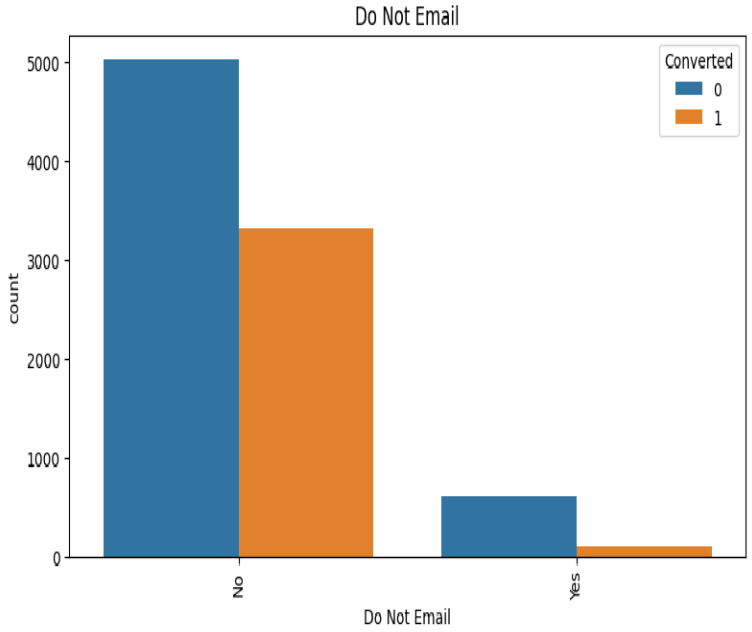

**Lead Origin:** Most of the leads are from India

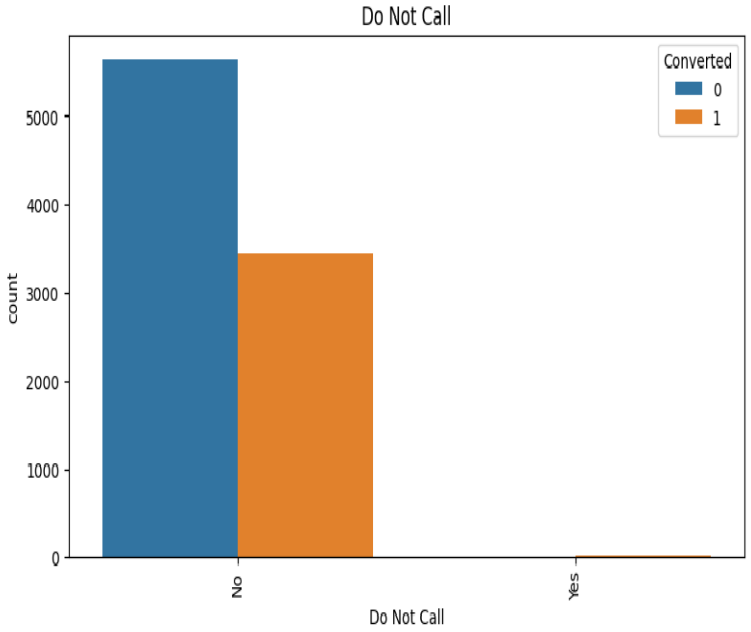
**Current_occupation:** It has 90% of the customers as Unemployed.



# EDA

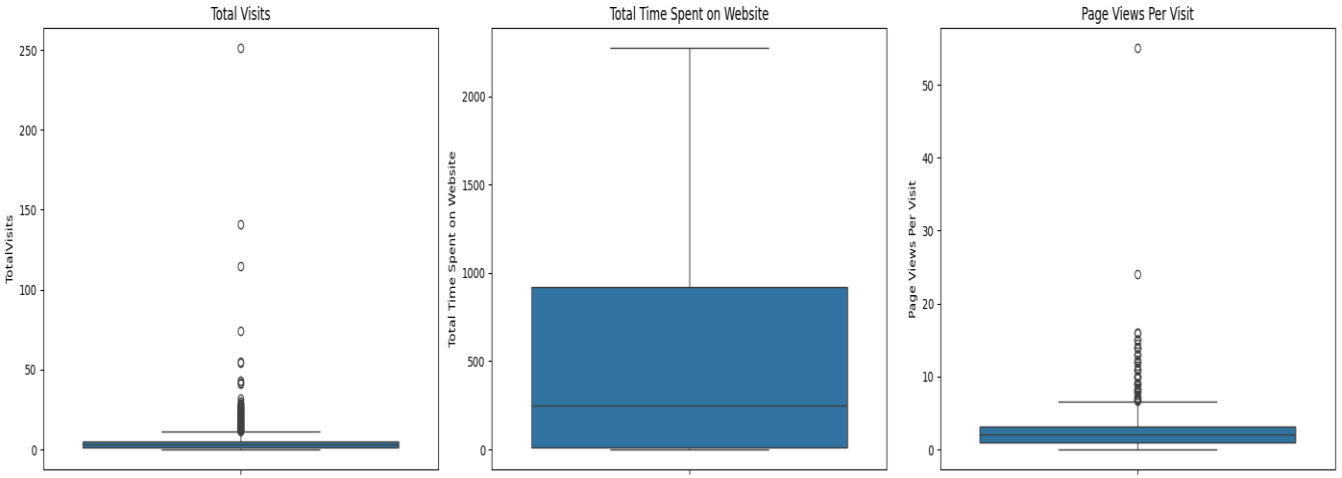
## Bivariate Analysis – Categorical Variables



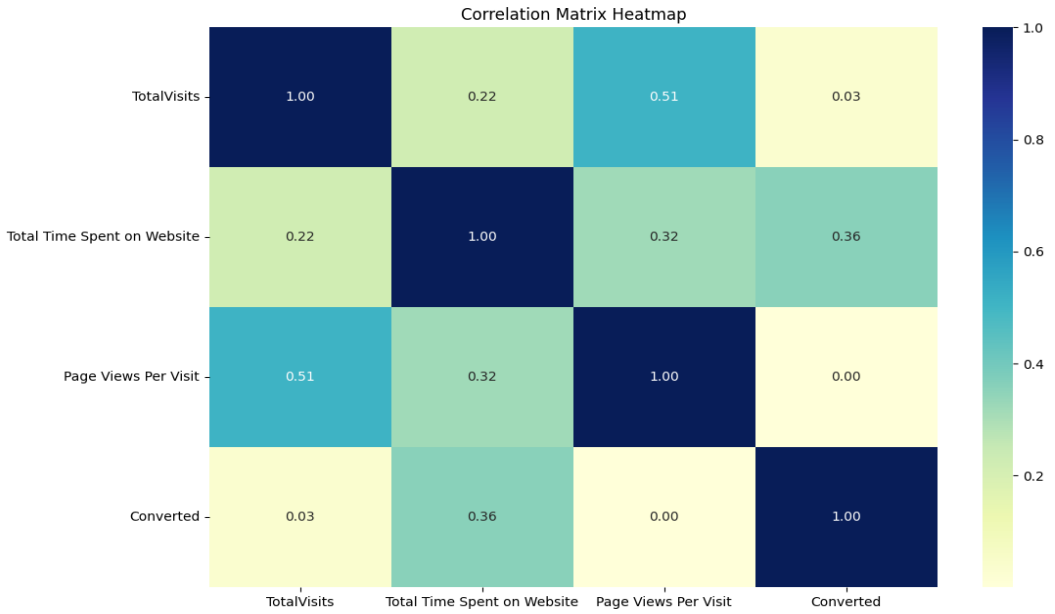

**Engagement through call and email based:** We get some engagement through email. But a greater number of the people have opted that they don't want to be emailed and called about the .

# EDA - Bivariate Analysis for Numerical Variables




Based on the data we can see the users tend to spend a lot less time on the website. Also, the visits and content browsed are towards the lower end of the spectrum..

Past Leads who **spends more time on the Website** have a higher chance of getting successfully converted than those who spends less time as seen in the **box-plot**



# Data Preparation before Model building

- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- Drop the categorical variables from model_df
- Splitting Train & Test Sets
  - 70:30 % ratio was chosen for the split
- Feature scaling
  - Standardization method was used to scale the features

# Model Building

- Used a Hybrid approach for feature selection.

  First used **Recursive Feature Elimination** (RFE) to select 15 out of ~130 columns.
- Then, removed features by dropping variables with p – value greater than 0.05 and VIF value greater than 5.
- Model 4(final_model) 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

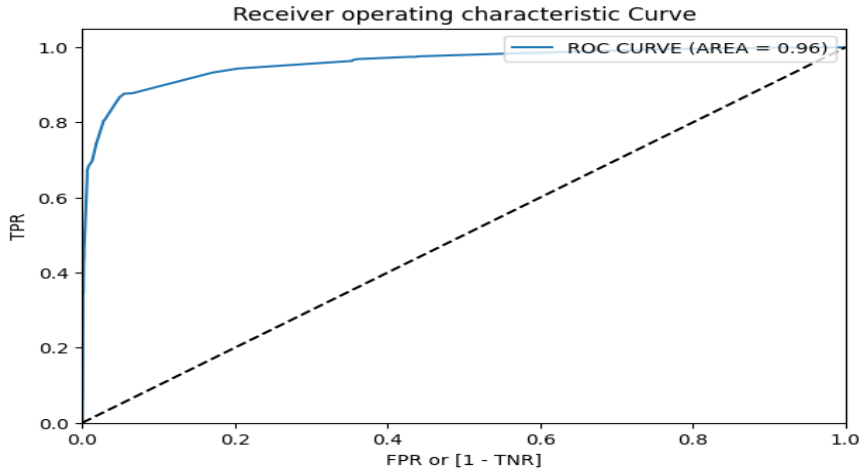
| | variables | VIF |
|---|---|---|
| 2 | Last Activity_SMS Sent | 1.56 |
| 10 | Last Notable Activity_Modified | 1.44 |
| 3 | What matters most to you in choosing a course_... | 1.39 |
| 8 | Tags_Will revert after reading the email | 1.33 |
| 7 | Tags_Ringing | 1.11 |
| 1 | Do Not Email_Yes | 1.10 |
| 5 | Tags_Closed by Horizzon | 1.06 |

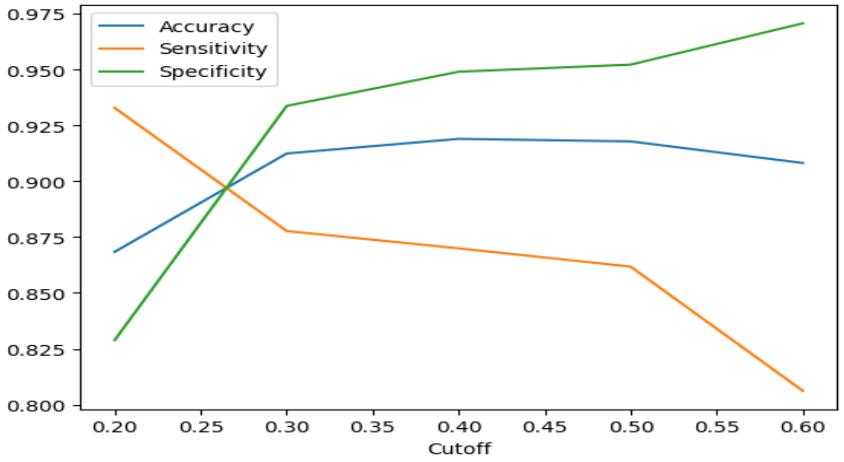
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6109 |
| **Model:** | GLM | **Df Residuals:** | 6096 |
| **Model Family:** | Binomial | **Df Model:** | 12 |
| **Link Function:** | Logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -1423.5 |
| **Date:** | Sat, 15 Feb 2025 | **Deviance:** | 2846.9 |
| **Time:** | 12:55:23 | **Pearson chi2:** | 1.44e+04 |
| **No. Iterations:** | 8 | **Pseudo R-squ. (CS):** | 0.5778 |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -1.2099 | 0.087 | -13.967 | 0.000 | -1.380 | -1.040 |
| **Lead Source_Welingak Website** | 4.2452 | 0.741 | 5.732 | 0.000 | 2.794 | 5.697 |
| **Do Not Email_Yes** | -1.3536 | 0.237 | -5.703 | 0.000 | -1.819 | -0.888 |
| **Last Activity_SMS Sent** | 2.2211 | 0.109 | 20.316 | 0.000 | 2.007 | 2.435 |
| **What matters most to you in choosing a course_not provided** | -0.7779 | 0.113 | -6.911 | 0.000 | -0.998 | -0.557 |
| **Tags_Busy** | 0.7865 | 0.235 | 3.353 | 0.001 | 0.327 | 1.246 |
| **Tags_Closed by Horizzon** | 7.0758 | 0.722 | 9.800 | 0.000 | 5.661 | 8.491 |
| **Tags_Lost to EINS** | 6.4033 | 0.607 | 10.544 | 0.000 | 5.213 | 7.594 |
| **Tags_Ringing** | -3.4824 | 0.247 | -14.122 | 0.000 | -3.966 | -2.999 |
| **Tags_Will revert after reading the email** | 4.5829 | 0.180 | 25.515 | 0.000 | 4.231 | 4.935 |
| **Tags_switched off** | -3.9229 | 0.599 | -6.550 | 0.000 | -5.097 | -2.749 |
| **Last Notable Activity_Modified** | -1.8094 | 0.119 | -15.157 | 0.000 | -2.043 | -1.575 |
| **Last Notable Activity_Olark Chat Conversation** | -1.2099 | 0.422 | -2.869 | 0.004 | -2.037 | -0.383 |

# Model Evaluation & Prediction (Optimum Cut off 0.30)




The model evaluation metrics for test set are almost identical to the model evaluation metrics on the train set. This means that our model generalizes very well over the data and is fit for use for scoring the leads.
**Again here the Recall metric of the model on the test set is ~89% which is in line with the business goal.**

Train Set Metrics

```
Accuracy: 0.9124242920281552
Sensitivity: 0.8776916451335056
Specificity: 0.9337206231845788
Precision: 0.8903451288772389
Recall: 0.8776916451335056
```

Test Set Metrics

```
Accuracy :  0.9175257731958762
Sensitivity :  0.8869653767820774
Specificity :  0.9358582773365913
Precision :  0.8924180327868853
Recall :  0.8869653767820774
```

# Recommendation based on Final Model

Tags seems to be an important feature as it has the top 3 positive coefficients within the model.

- **Tags_Closed by Horizzon (coef: 7.0758)**
- **Tags_Lost to EINS (coef: 6.4033)**
- **Tags_Will revert after reading the email (coef: 4.5829)**

The above three features should be considered as the most influential in predicting lead conversion.

The odds of conversion are also high when **lead source** is "**Welingak Website**" and the **last activity** is "**sms sent**".

People do not like invasive follow-ups, hence people with Do not email set to Yes negatively impacts conversion.
Also, it seems to be a waste of time to follow-up on leads that have ringing or switched off as tags, possible indicating that the contact details are stale.

# Recommendation based on Final Model

## To increase our Lead Conversion Rates

- Focus on leads from Welingak Website, those who have interacted with SMS Sent, and those tagged with Closed by Horizzon , Lost to EINS and Will revert after reading the email, as these variables have high positive coefficients and indicate higher conversion potential. Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Segment leads based on urgency: hot leads (most likely to convert) should be called immediately, warm leads (moderate probability) should be followed up via email first, then called, and low-engagement leads should receive personalized follow-ups before calling.
- For the phone call strategy, make follow-up calls immediately after an email or SMS is sent (considering the positive coefficient for "Last Activity_SMS Sent") and use insights from tags like "Busy" or "Ringing" to tailor the conversation, perhaps scheduling calls at times when leads are less likely to be busy.

## To identify areas of improvement

- Look at alternative channels for leads that have tags as switched off
- Reach out to leads at alternative timeslot where tags is marked as ringing

***Regularly review the conversion rates to adjust strategies. Use feedback from interactions to refine the engagement approach (call script, sms message etc)***

*Thank You!*