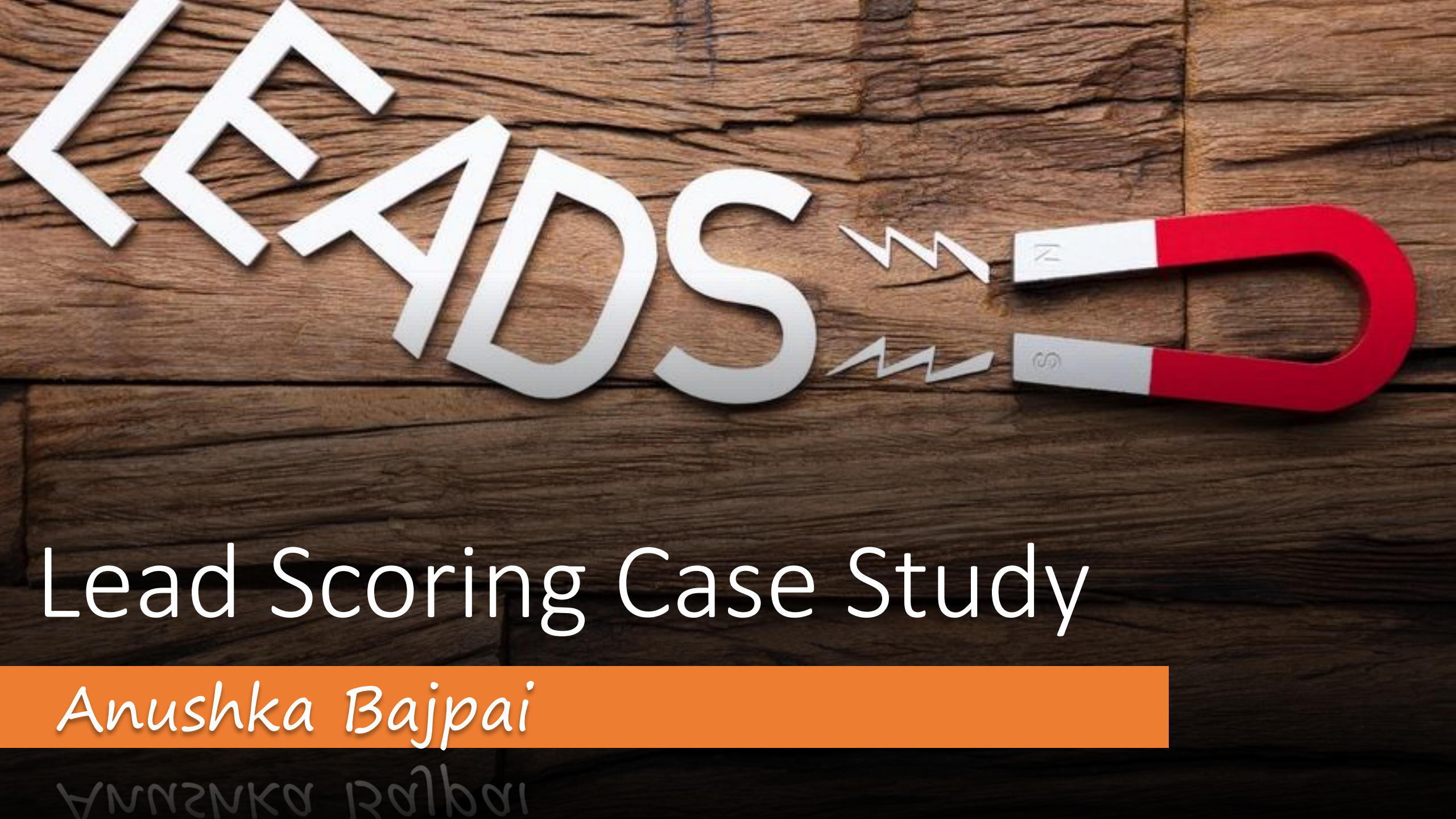



# Lead Scoring Case Study

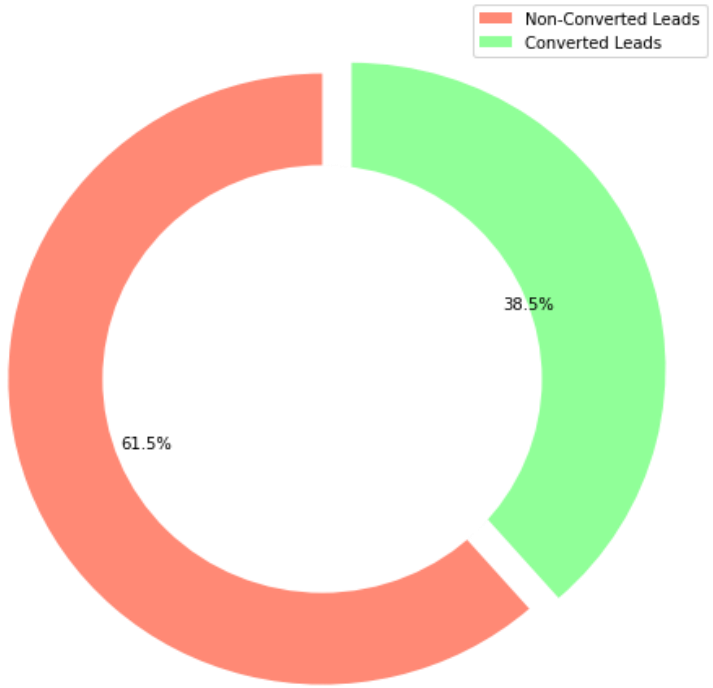
Anushka Bajpai

# PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- Although X Education gets a lot of leads, its lead conversion rate is very poor (around 30%). To improve this, the company wishes to identify the most potential leads ("hot leads")


## Business Objective :

- X Education wants to successfully identify the right set of leads
- In order to achieve this, they want to know most promising hot leads
- They need to implement the right set of practices on communicating with these leads




## Goals of the Case Study

1. *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 (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 mostly not get converted).*
2. *There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.*

# Methodology and Approach

## 1. DATA INSPECTION

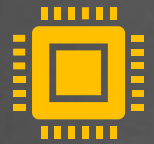
- Data Loading
- Data parameters check
- Duplicate value check
- Null value check
- Imbalance check

## 2. DATA CLEANING

- Drop variables with nulls > 40%
- Drop rows with nulls > 70%
- Drop unique valued cols
- Drop heavily skewed cols
- Replace 'Select' values with NaN
- Club low occurrence values

## 3. EDA

- Univariate
- Bivariate
- Multivariate
- Correlation plot (heat-map) for numeric variables



## 4. FEATURE SELECTION & DATA SCALING

- Dummy Variable creation
- Data encoding
- Train-Test Data Split
- RFE

## 5. MODEL BUILDING

- Logistic Regression used
- Mixed approach used : RFE Statsmodel

## 6. MODEL VALIDATION

- Model metrics
- Confusion matrix evaluation
- Metrics evaluation
- Probability cutoff
- ROC curve
- Lead scores assigned

## 7. FINAL MODEL PRESENTATION

- Train-Test metrics comparison
- Trade-off visualization

# Data Inspection Findings

- Total Number of records: 9240 | Total number of Variables : 37 (original + Sales team provided)
- **Presence of Duplicate values :** None | **Presence of Null values : 5 columns ( > 40% nulls)**

# Data Cleaning Steps :

- Presence of Null/redundant/unique values (in some columns) :

  **Fix –** a) Dropped columns with more than **40% null values** (total : 5 cols)

  b) Columns with '**Select**' values replaced with **NaN** since they are as good as nulls.

  c) Imputed the missing values with Mode in categorical and Median for Numeric columns.

  d) For columns with multiple "**Select**" values, a separate category was added as "**Not_specified**" rather than imputing with mode/median

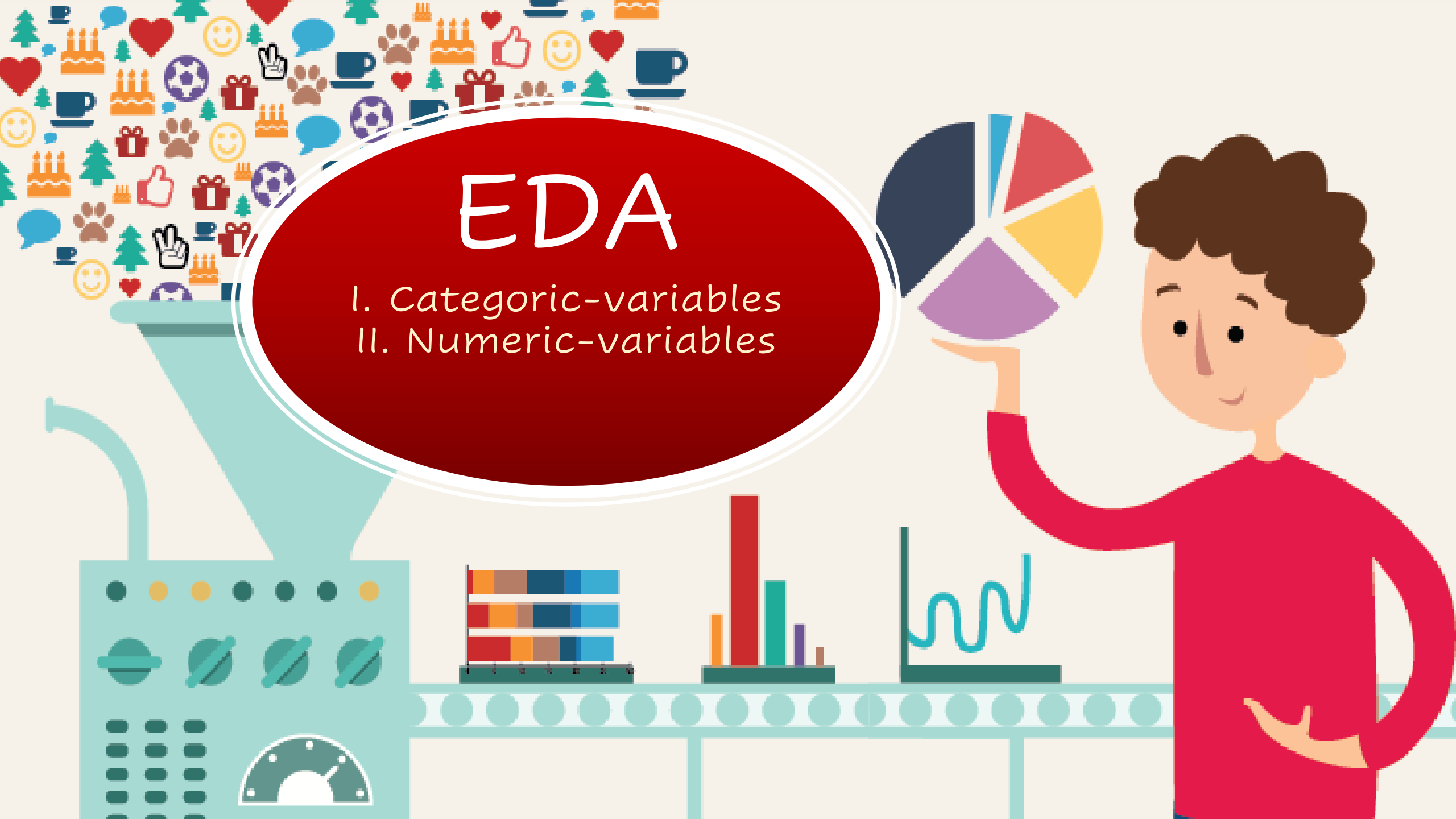
  e) **Dropped unique valued** columns (Lead Number, Prospect ID)

  f) Dropped rows with null values greater than 70% **(noisy data)**

- Clubbed low occurrence values together for a better analysis
- Dropped highly skewed columns which had a very low variability and could make the model biased
- Dropped variables provided by Sales Team (post-EDA) to avoid overfitting of our final model.

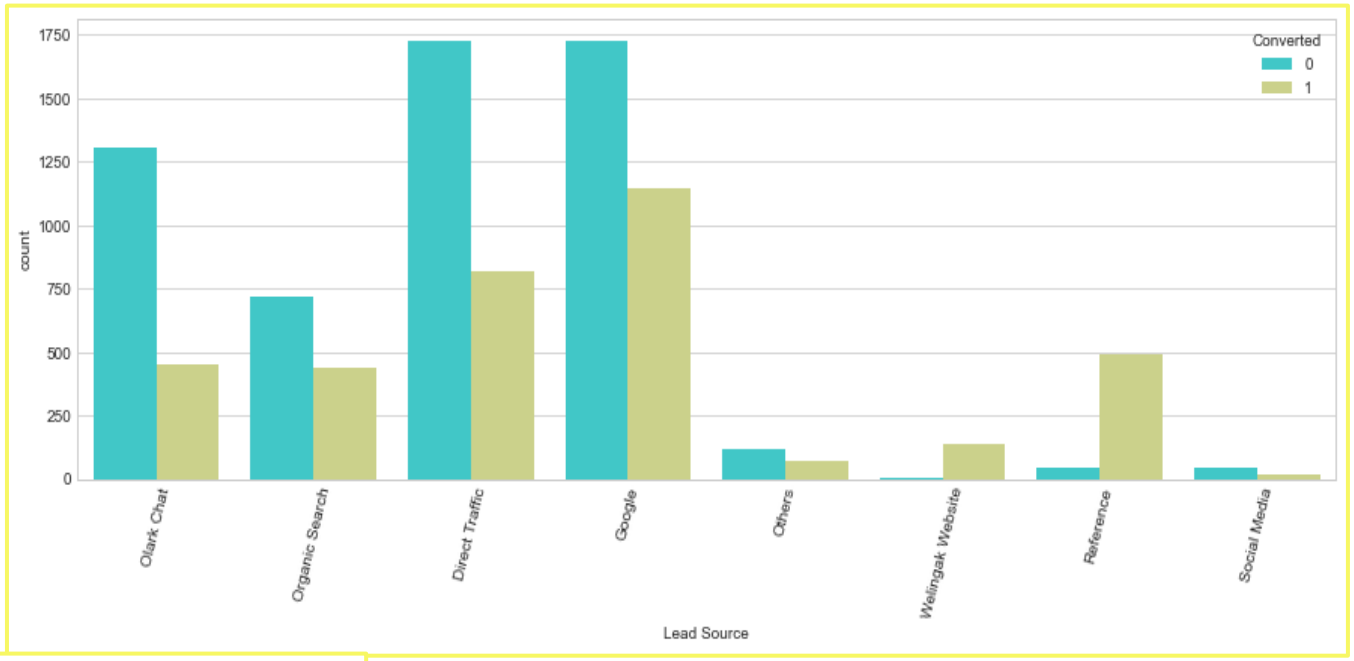
# EDA

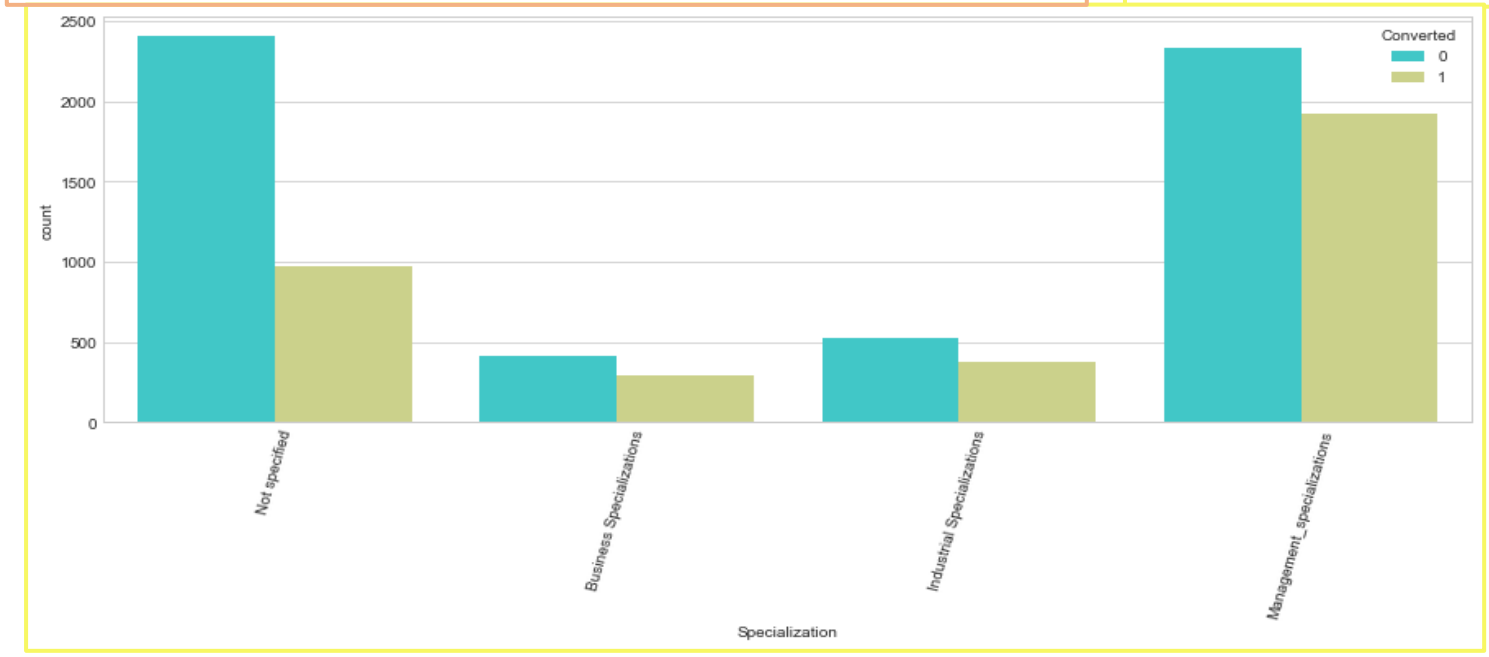
I. Categoric-variables
II. Numeric-variables



# Categoric Variables

- **Majority of the converted leads come from Google and Direct Traffic**. These should be used to generate further leads.
- Conversion rate is highest for leads coming from **Welingak Website followed by Reference**.
- Leads from other sources show a medium or poor conversion rate and hence company needs to focus and act upon these sources.
- We will consider this variable for our final analysis

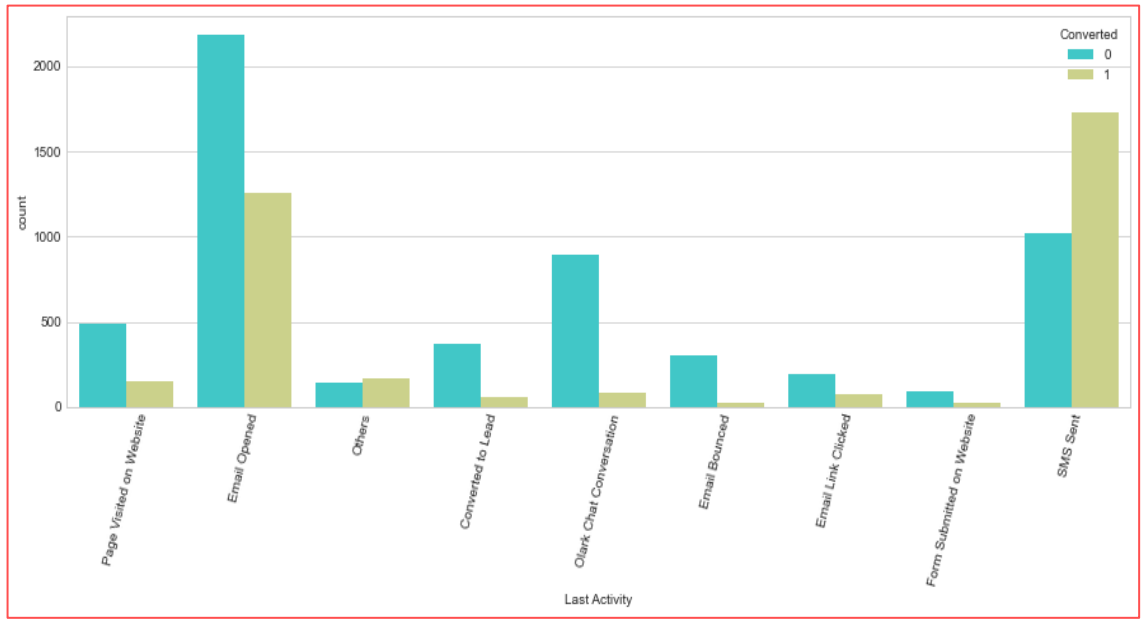



## Lead Source and Specialization

- There is a sense of high skewness here despite clubbing the valued under specific Specialization.
- Management has the highest number of Leads as well as Conversion Rate as compared to the other two
- This is likely to result in bias (even after we drop Not_specified column while creating dummies)
- Considering this in mind, we do not consider it in our final analysis (Model Preparation)

# Tags, Last Activity and Last Notable Activity



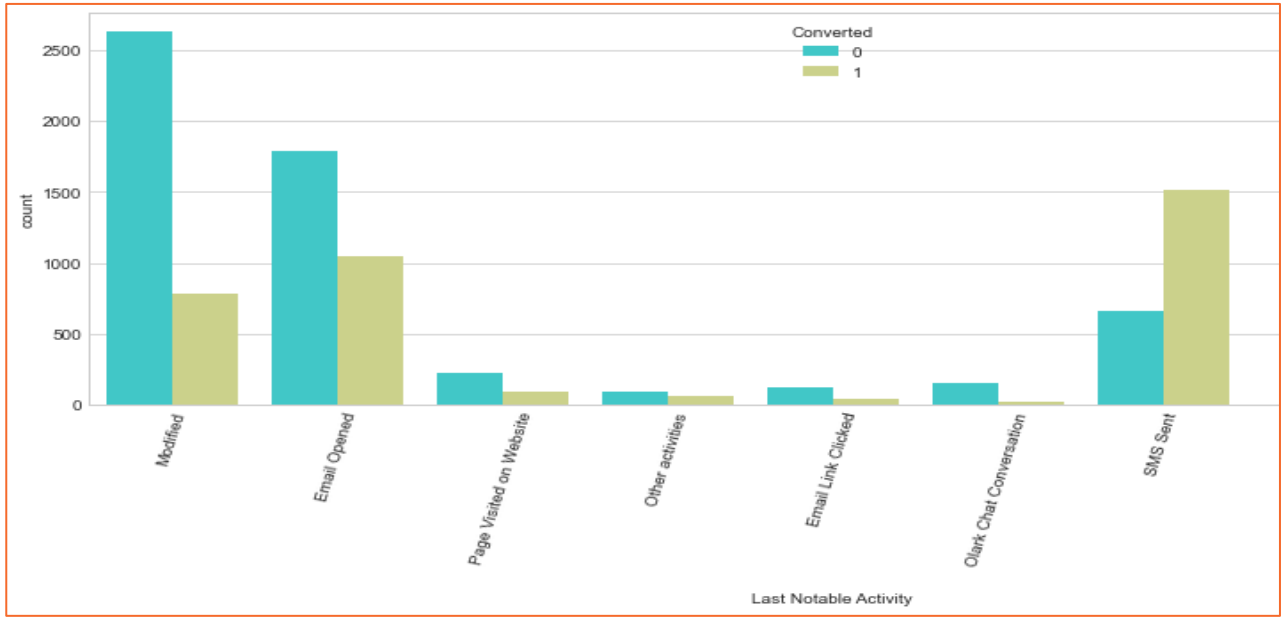

1. Leads with SMS Sent as last activity have the highest conversion rate.

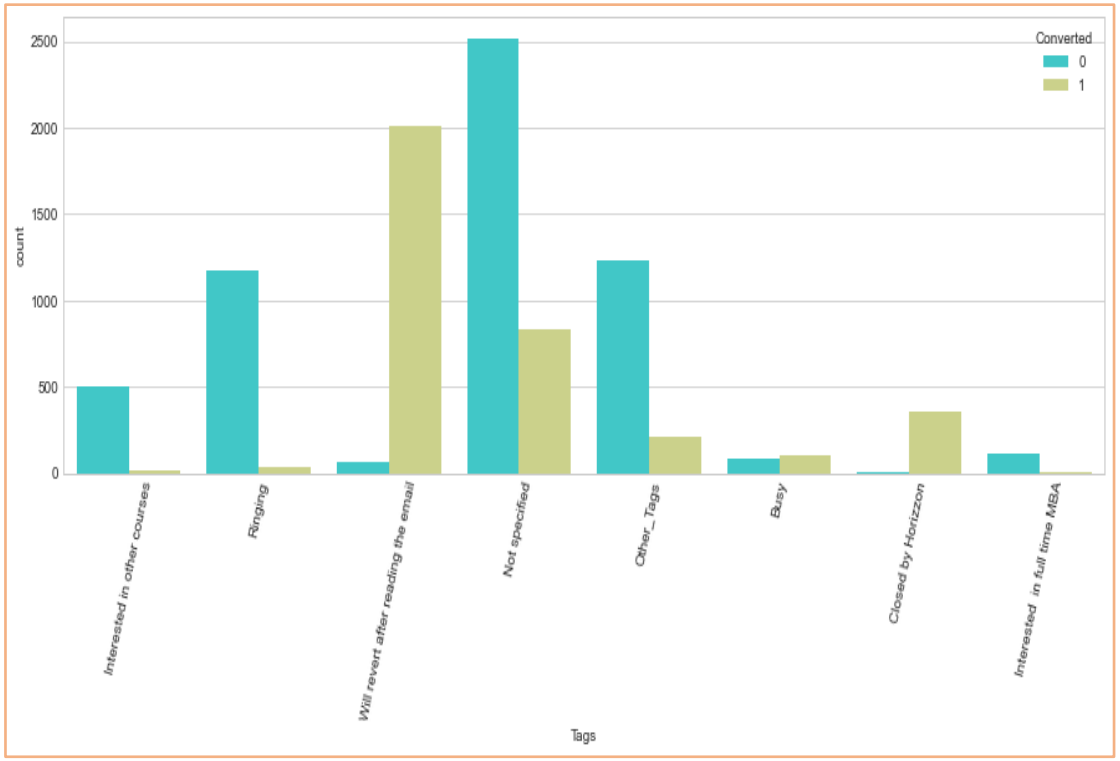
2. Leads with Email Opened have a moderate chance of getting converted (around 50%). It indeed needs attention

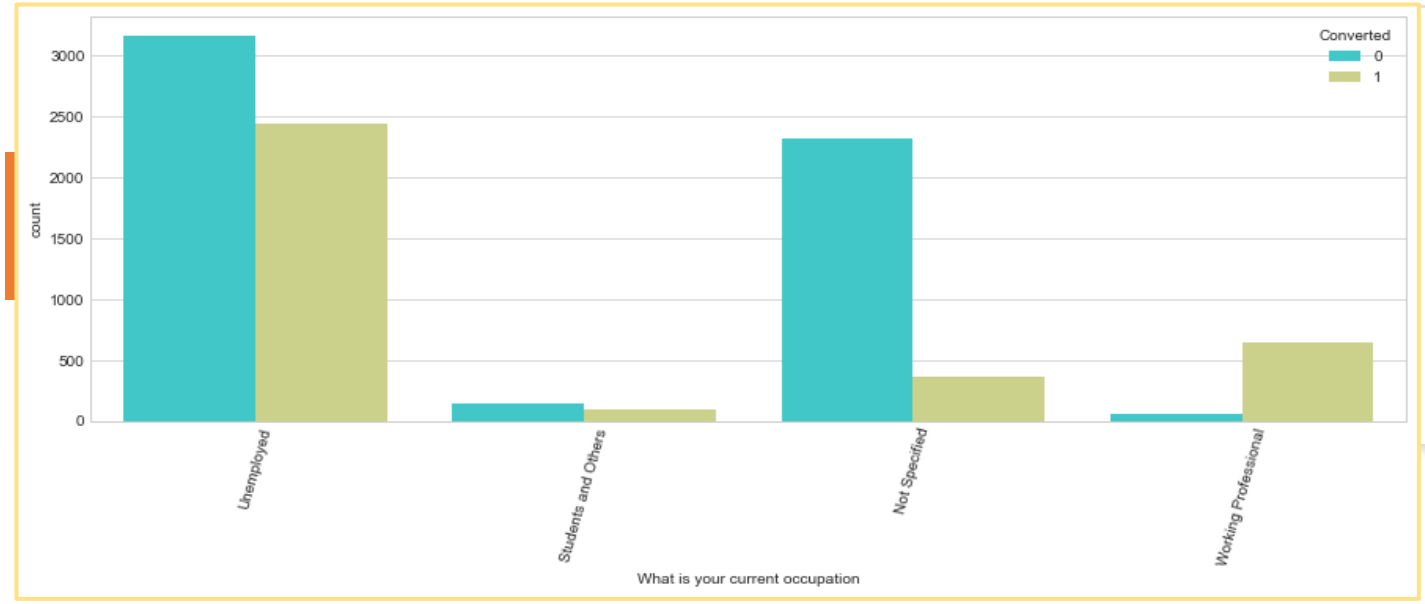
3. The maximum leads who converted belong to 'Will revert after reading the email' Tag category

4. Tags Closed by Horizzon also have a very high conversion rate (almost 100%)

As all these 3 variables have been provided by the **SALES TEAM** they are likely to result in **overfitting** of our model, hence we **will not consider them in our Model Preparation**






- **Unemployed leads** are in majority, however their conversion rate is nearly around 50%
- **Working professionsals** are second highest leads and there conversion rate is high
- **We will consider this column for our final analysis**

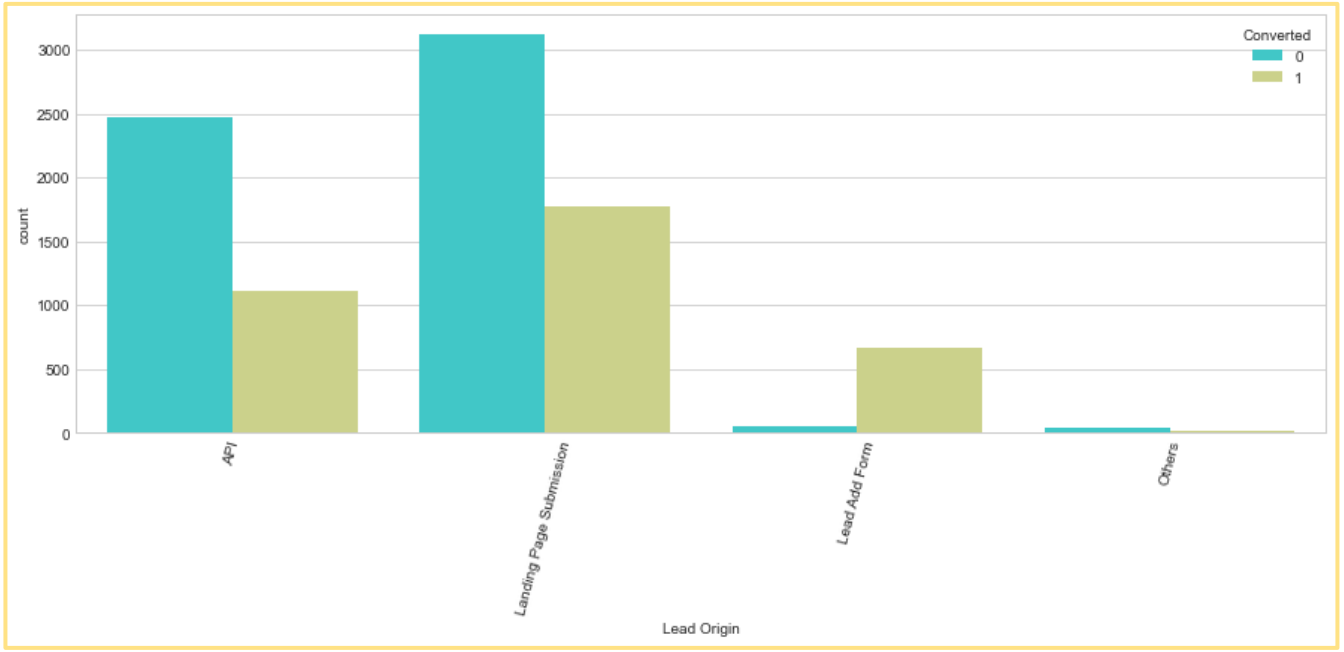
# Current Occupation and Lead Origin

- **Landing Page Submission followed by API bring in higher number of leads** and have moderate to fair conversion rate. Company needs to improve lead conversion here

- **Lead Add Form has a very high conversion rate,** however not many leads come in through this. Company needs to focus here in order to bring more leads.

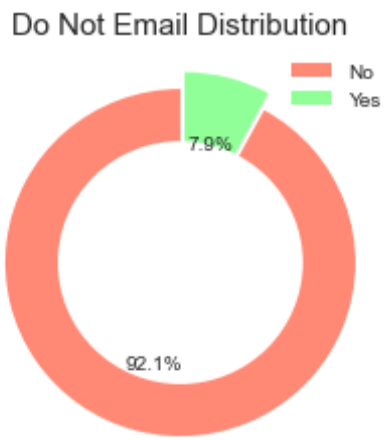
- Lead Import and Quick Add Form has has a poor conversion rate and bring in very minimal leads

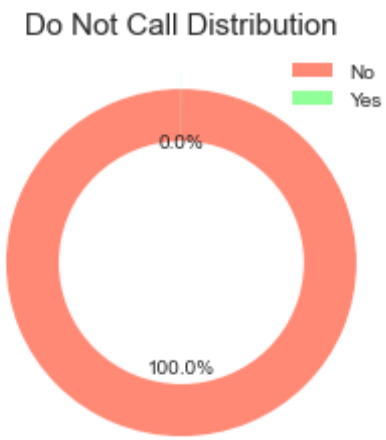
- **We will consider this column for our final analysis**

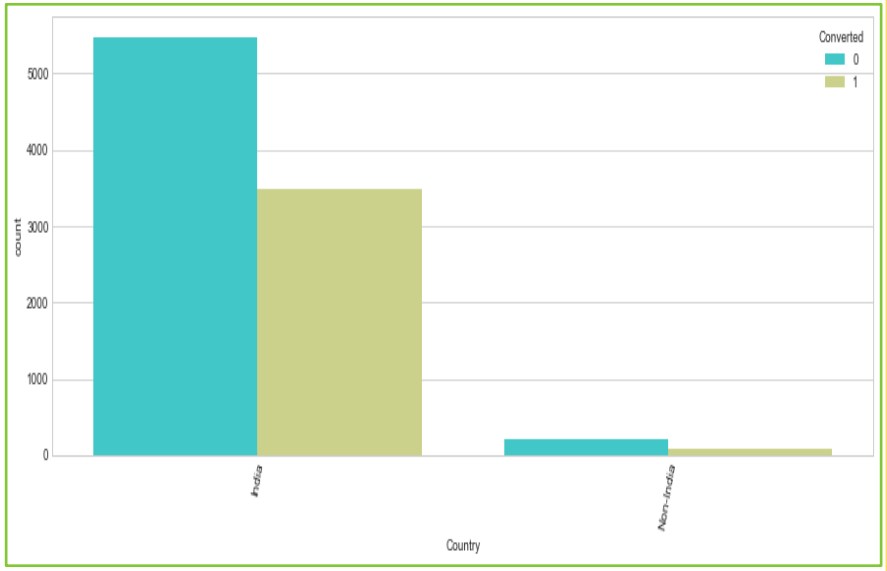


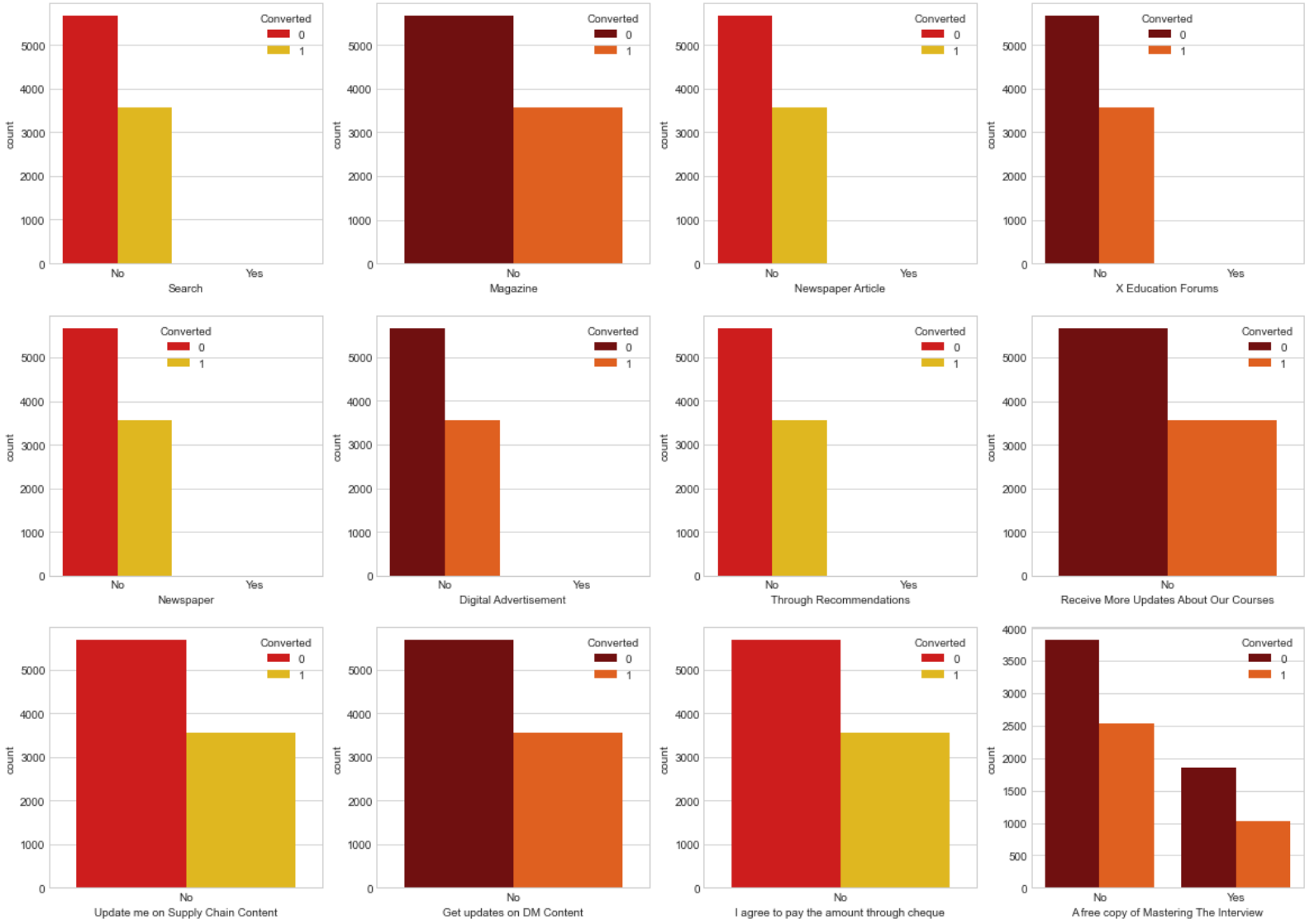
# EDA – Skewed columns

[These columns had to be dropped due to high bias (except for : ‘Do Not Email and ‘A free copy of Mastering the Interview)]



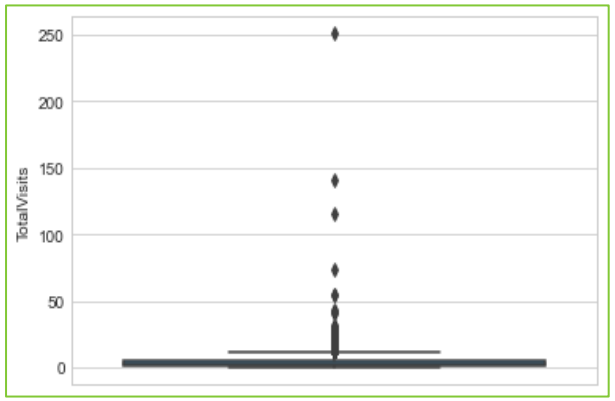



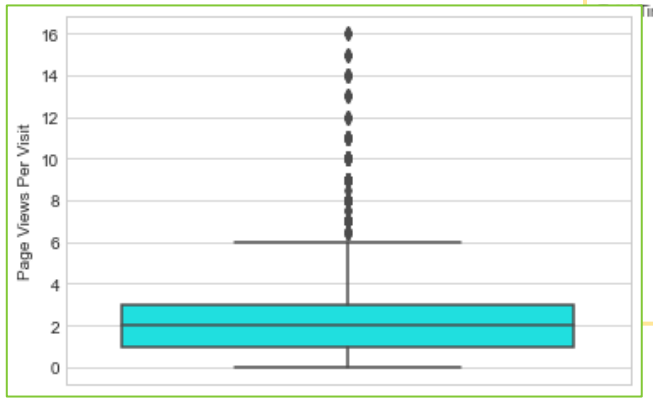



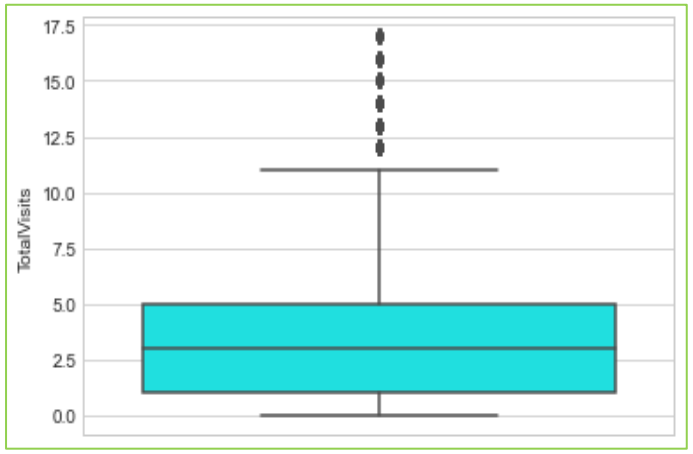



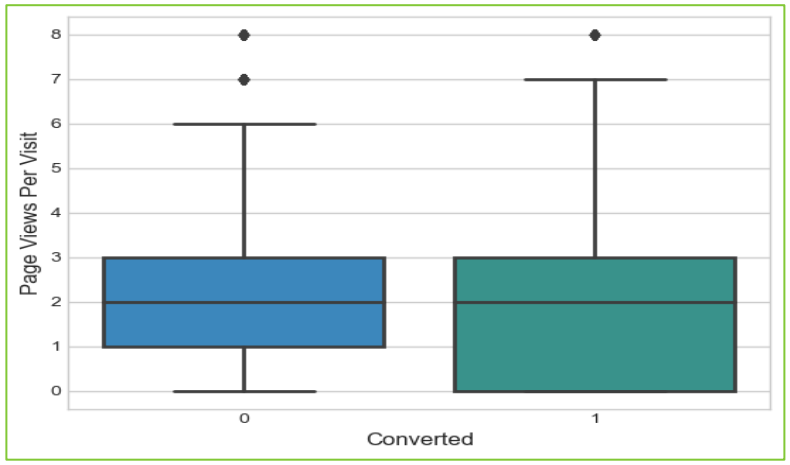
# EDA –Numeric Columns

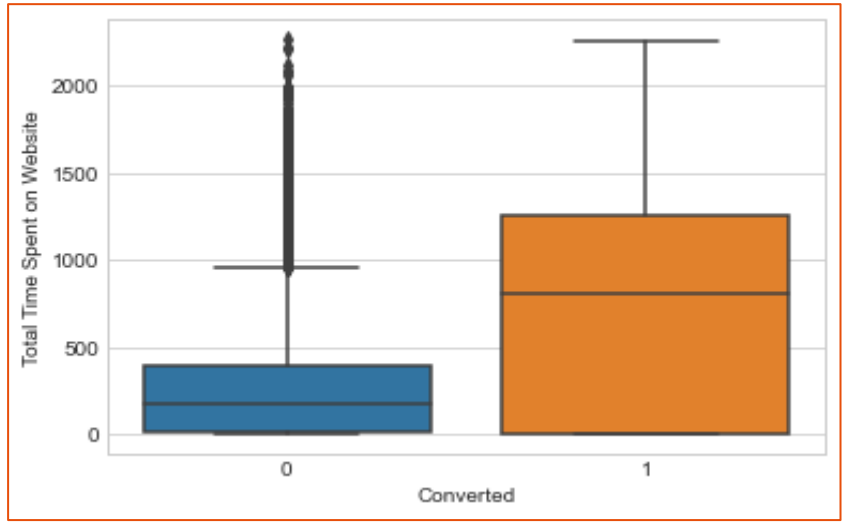
- No significant correlation observed between variables
- Due to presence to outliers, outlier treatment had to be performed for : **Total Visits and Page Views Per Visit**

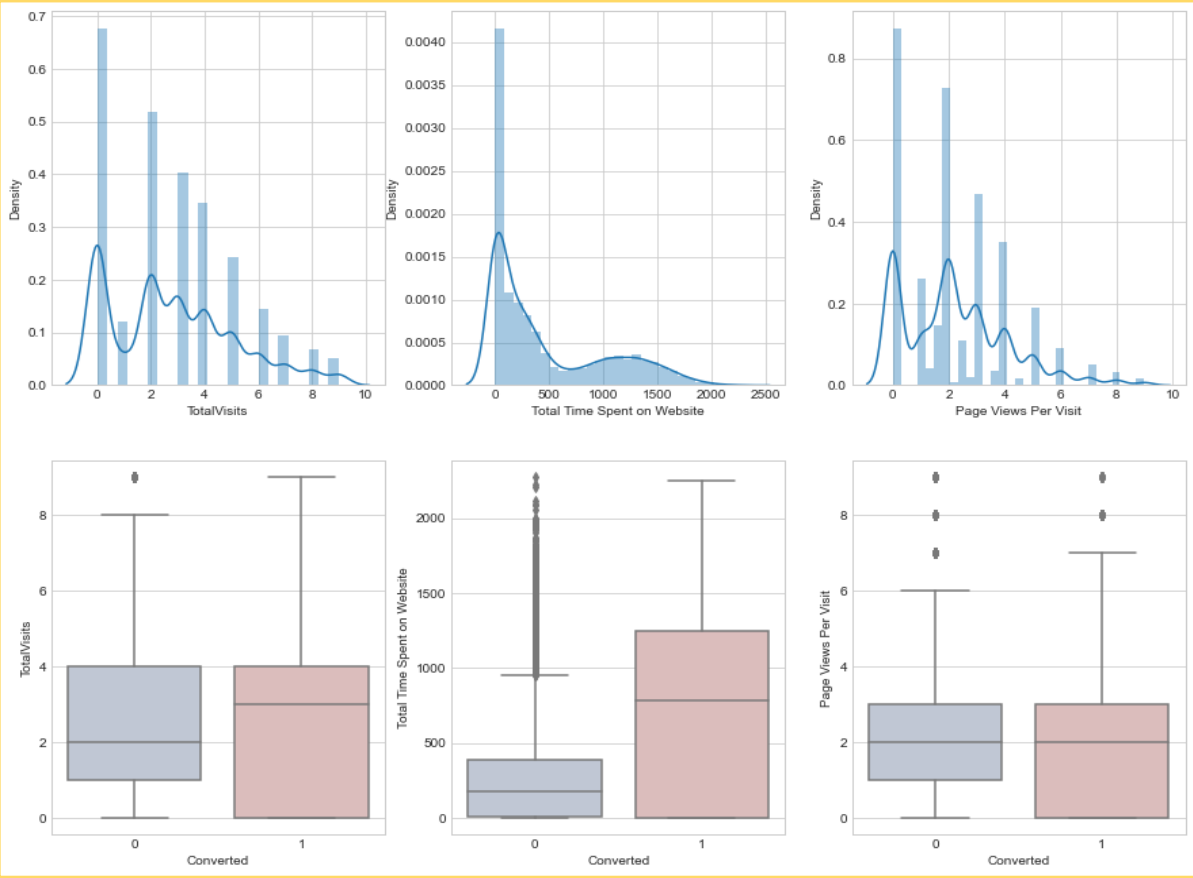



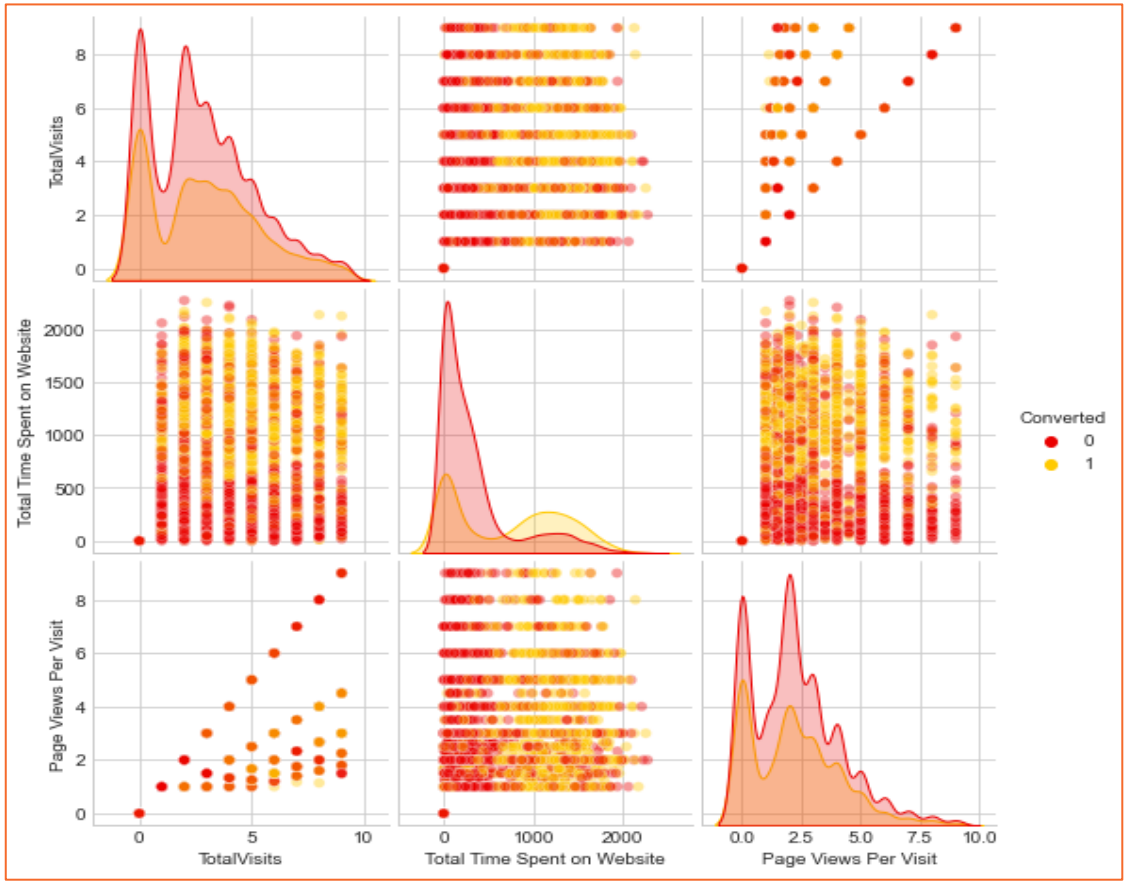



# Univariate and Bivariate analysis (after outlier treatment)




After outlier treatment, the distributions look better and moderate trends are visible :

- Converted leads have a higher total visits to the website than non-converted (as per median)
- Converted Leads have are likely to spend more time on website than those who didn't
- All leads have a similar trend in terms of the pages they viewed during each visit.

The data is not normally distributed (no major info or correlation can be drawnfrom the above pair plot)

# Feature Selection

## Dummy variable creation

- For all variables with 'Select' option, created dummies for each option/level and then dropped the first column by explicitly specifying the "not-specified" / "others" level.

- Encoded variables with suitable labels. Also mapped binary-valued columns (Yes/No) to 0 and 1 respectively.

**Steps followed :**

1. Get dummies
2. Drop first column
3. Add the results to the main dataframe (leads)

**Note :** For all columns having the "Select_Not specified" label, we need to create dummies first and then drop that level explicitly by specifying

```
dummy = pd.get_dummies(leads['Lead Origin'], prefix ='Lead Origin')
dummy = dummy.drop(['Lead Origin_Others'], 1)

leads = pd.concat([leads,dummy], 1)
```

```
dummy = pd.get_dummies(leads['What is your current occupation'], prefix ='What is your current occupation')
dummy = dummy.drop(['What is your current occupation_Not Specified'], 1)

leads = pd.concat([leads,dummy], 1)
```

```
dummy = pd.get_dummies(leads['Lead Source'], prefix ='Lead Source')
dummy = dummy.drop(['Lead Source_Others'], 1)

leads = pd.concat([leads,dummy], 1)
```

**We can finally drop the original categorical columns after dummy creation**

## 18 variables selected after running RFE on dataset are :

```
cols = x_train.columns[rfe.support_]
cols
```

```
Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'A free copy of Mastering The Interview',
       'Lead Origin_API', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form',
       'What is your current occupation_Students and Others',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Lead Source_Direct Traffic', 'Lead Source_Google',
       'Lead Source_Olark Chat', 'Lead Source_Organic Search',
       'Lead Source_Reference', 'Lead Source_Social Media',
       'Lead Source_Welingak Website'],
      dtype='object')
```

Splitting the entire data set in the two parts using train_test_split method from SKLEARN library in the ratio of ***70:30 :***

- **TRAIN : 70 %**
- **TEST : 30 %**

# Model Building

- Type : Logistic Regression (Classification)
- **RFE :** Ran RFE with **18** variables as output
- **Stats-Model** : Ensured p-value (< 0.05) followed by VIF values (< 5) while dropping variables (priority given to p-value).

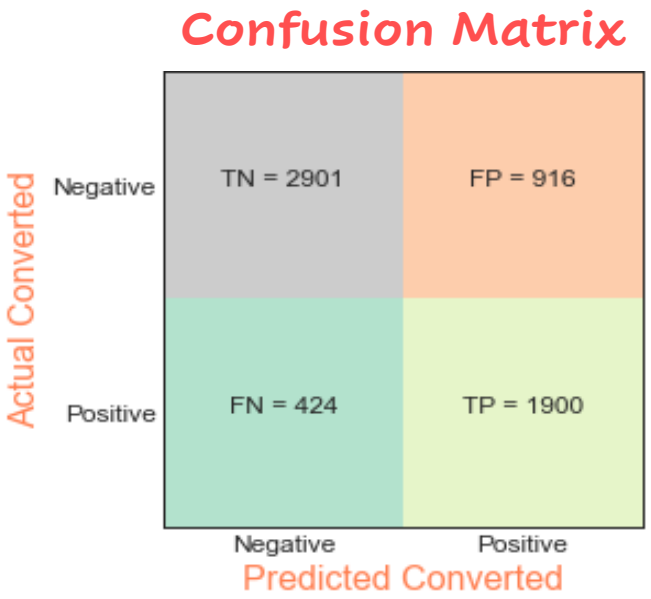

## Final Model :

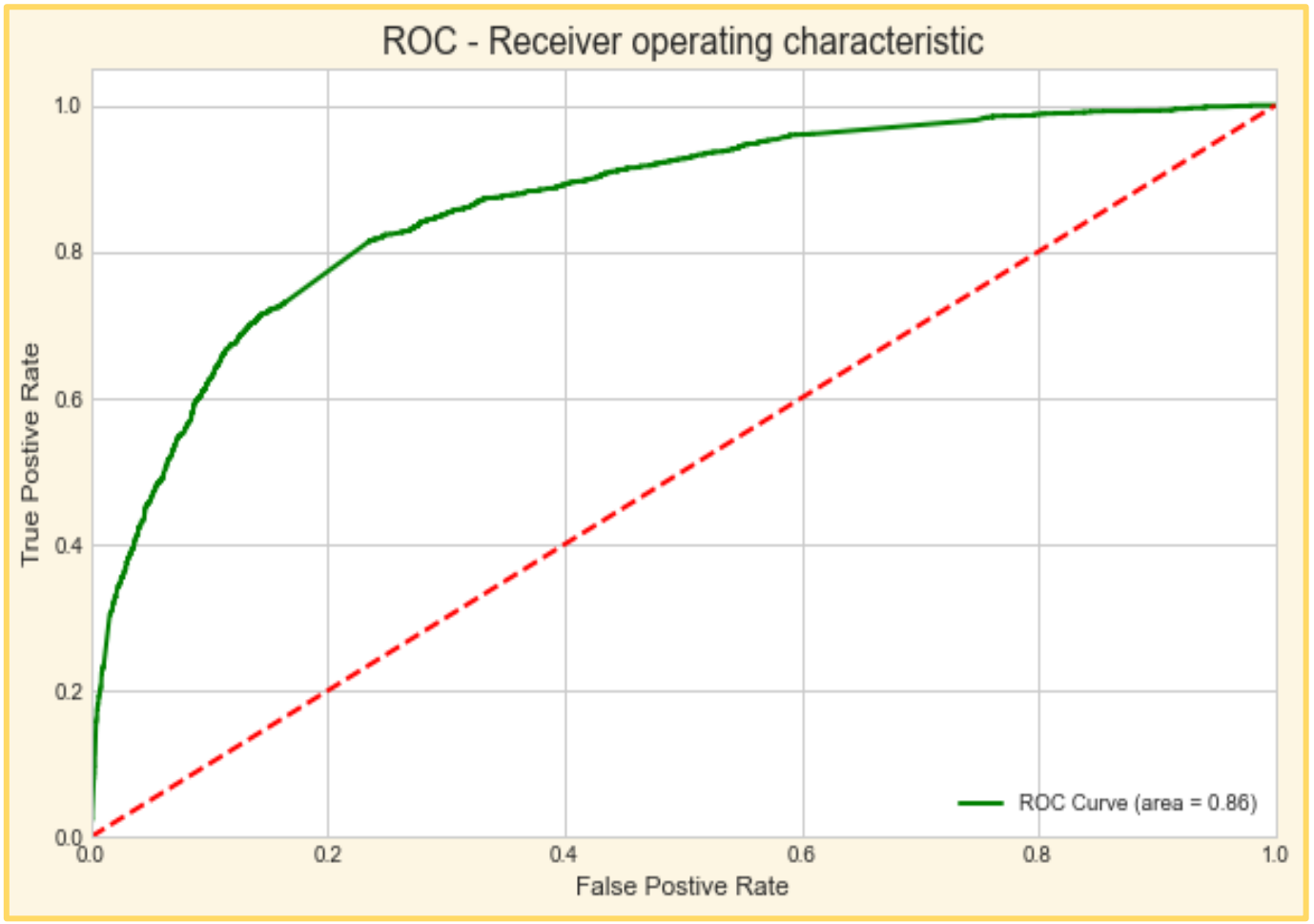
| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6141 |
| **Model:** | GLM | **Df Residuals:** | 6130 |
| **Model Family:** | Binomial | **Df Model:** | 10 |
| **Link Function:** | logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -2748.3 |
| **Date:** | Wed, 11 Aug 2021 | **Deviance:** | 5496.6 |
| **Time:** | 15:15:44 | **Pearson chi2:** | 7.54e+03 |
| **No. Iterations:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -2.9933 | 0.114 | -26.218 | 0.000 | -3.217 | -2.770 |
| **Do Not Email** | -1.1078 | 0.156 | -7.096 | 0.000 | -1.414 | -0.802 |
| **TotalVisits** | 0.2228 | 0.045 | 4.917 | 0.000 | 0.134 | 0.312 |
| **Total Time Spent on Website** | 0.0020 | 7.28e-05 | 27.329 | 0.000 | 0.002 | 0.002 |
| **Lead Origin_Landing Page Submission** | -0.2348 | 0.087 | -2.712 | 0.007 | -0.405 | -0.065 |
| **Lead Origin_Lead Add Form** | 3.8183 | 0.214 | 17.862 | 0.000 | 3.399 | 4.237 |
| **What is your current occupation_Students and Others** | 1.2399 | 0.202 | 6.132 | 0.000 | 0.844 | 1.636 |
| **What is your current occupation_Unemployed** | 1.3493 | 0.086 | 15.722 | 0.000 | 1.181 | 1.518 |
| **What is your current occupation_Working Professional** | 3.7009 | 0.194 | 19.067 | 0.000 | 3.320 | 4.081 |
| **Lead Source_Olark Chat** | 1.1788 | 0.130 | 9.074 | 0.000 | 0.924 | 1.433 |
| **Lead Source_Welingak Website** | 2.0232 | 0.740 | 2.733 | 0.006 | 0.572 | 3.474 |

| | Features | VIF_Value |
|---|---|---|
| **0** | What is your current occupation_Unemployed | 2.51 |
| **1** | Lead Origin_Landing Page Submission | 2.39 |
| **2** | Total Time Spent on Website | 1.94 |
| **3** | TotalVisits | 1.87 |
| **4** | Lead Origin_Lead Add Form | 1.78 |
| **5** | Lead Source_Olark Chat | 1.67 |
| **6** | What is your current occupation_Working Profes... | 1.34 |
| **7** | Lead Source_Welingak Website | 1.26 |
| **8** | Do Not Email | 1.10 |
| **9** | What is your current occupation_Students and O... | 1.08 |

# Model Evaluation




TP = confusion[1,1] ==> actual = predicted = 1 (converted)
TN = confusion[0,0] ==> actual = predicted = 0 (not-converted)

FP = confusion[0,1] ==> actual (0) != predicted (1)
FN = confusion[1,0] ==> actua (1)l != predicted (0)

**ROC Curve** has been drawn to find the Model stability with the AUC Score (area under the curve).

- From the graph on the right, it is évident that **AUC = 0.86** which a great score altogether.
- Moreover, the plot is leaned towards the **left** border and **uppe**r edge which indicates a **very good accuracy**

# Metrics with Optimal Cutoff Point

- Plot on the right shows the Accuracy, Sensitivity and Specificity for each of the points between 0 to 1.
- The point where all three are at the closest range (near trade-off) is at **0.32 and this is considered as our Optimal probability cutoff.**

## Values at optimal probability :

Accuracy : 78.17

Sensitivity : 81.75

Specificity : 76

## Other Metrics :

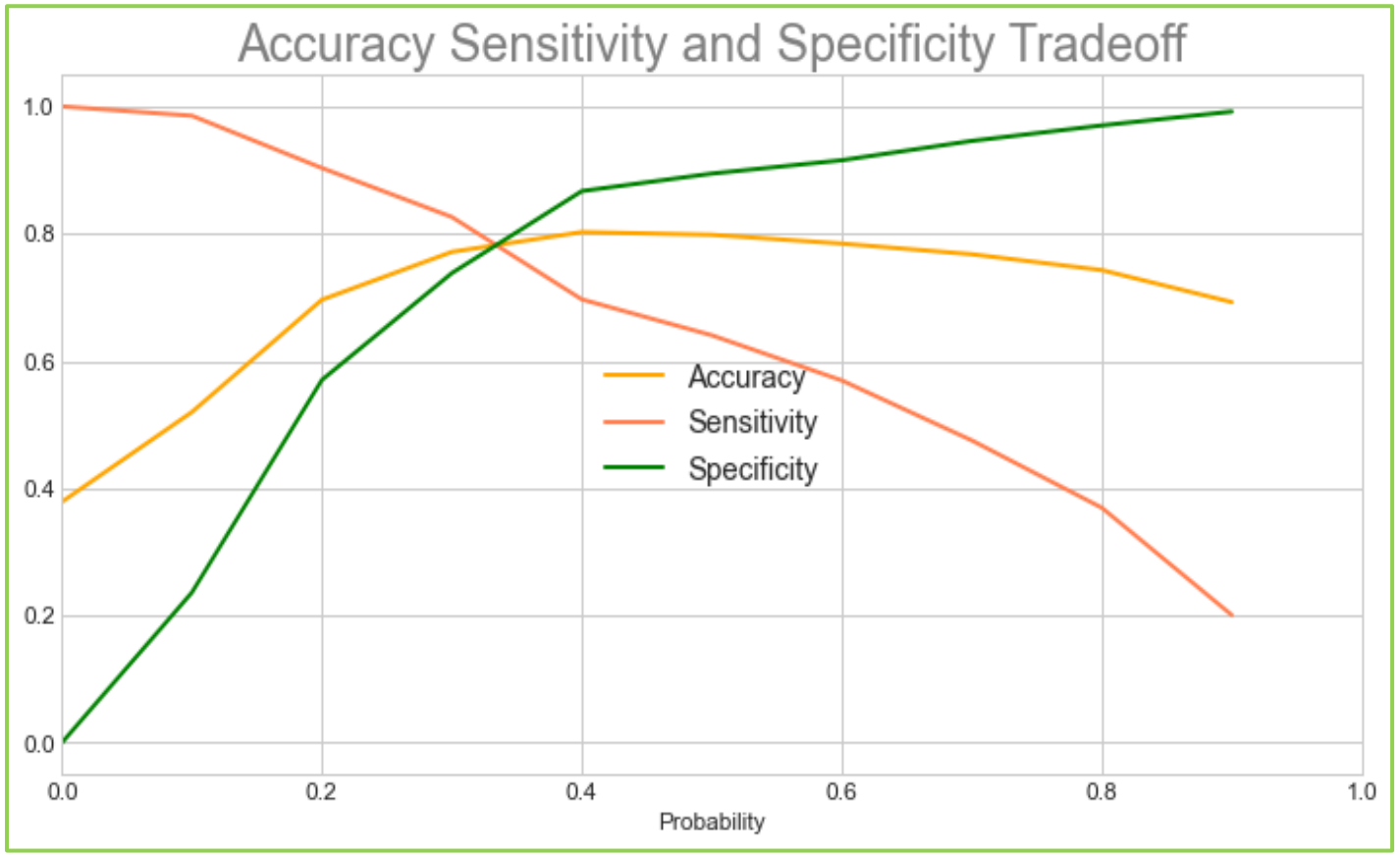
Positive Predictive Value : 67.5

Negative Predictive Value : 87.24

Precision : 67.5

Recall : 81.75

**FI - SCORE** (Harmonic mean of Precision and Recall score) = 74



# Precision and Recall

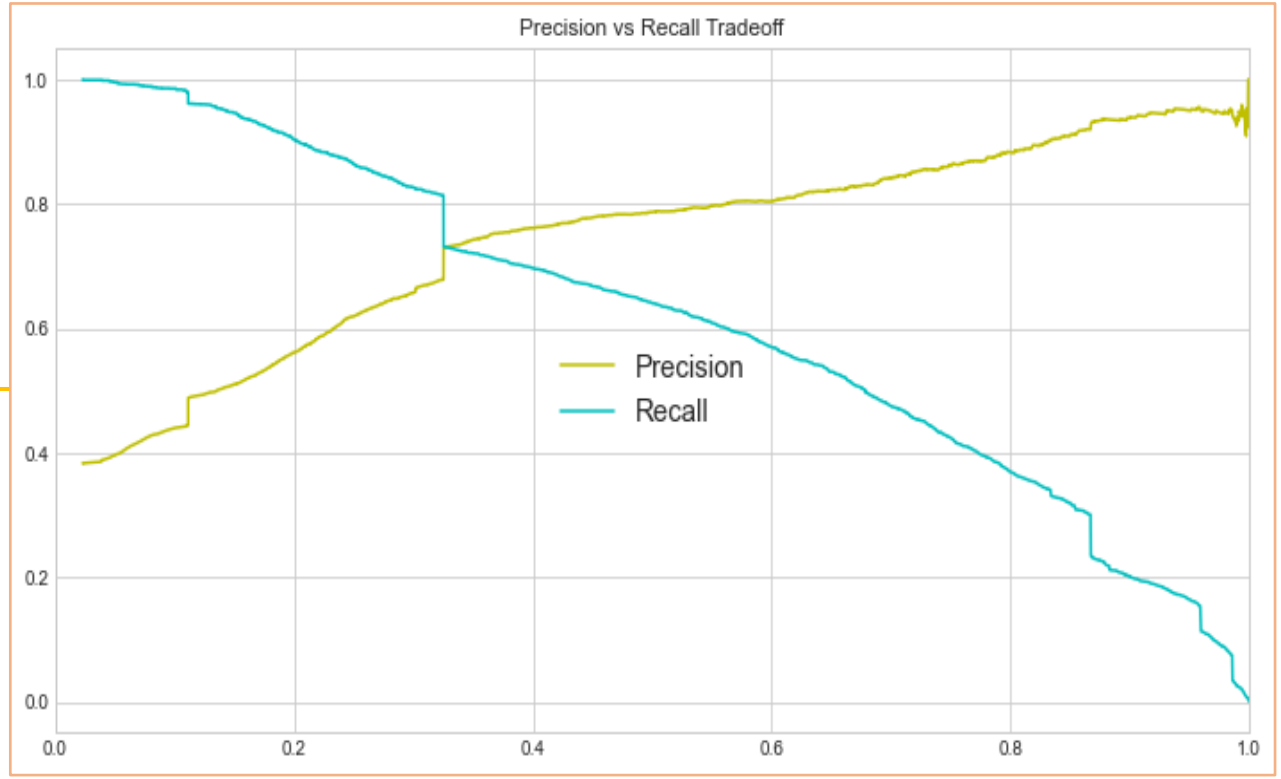
Both the above metrics play a crucial role in determining **performance** **of the model** and understand how **Business oriented** it is.

Precision score : 67.5
Recall score : 81.75

Recalling our Business Objective, we know that Recall score is more important as compared to Precision as we want to ensure we don't miss out on any Hot Leads who are actually willing to get Converted.

We can still manage with a little low precision, which mainly refers to chances of Cold leads being predicted as Hot Leads.

From the plot on right side, it is evident that the trade-off between the two is somewhere around 0.3 (approx)




We know that Precision has a bouncy, wavy nature at the end, whereas Recall remains comparatively smooth.

Precision is bound to change as denominator changes, but denominator remains fixed in case of Recall and hence a smoother landing.

# Predictions on Test Set

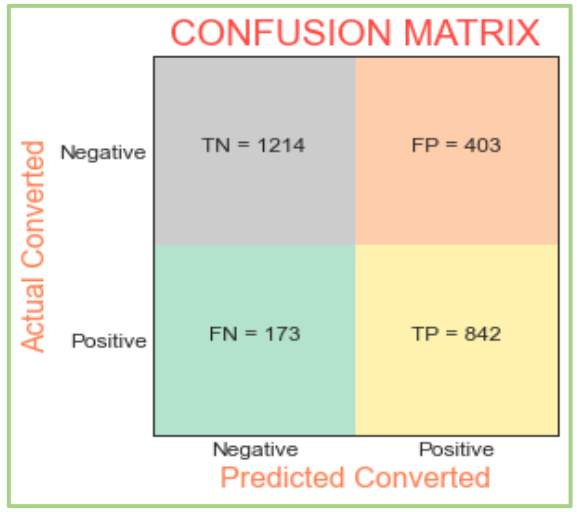
- Before we proceed with Test Set, we need to apply **Standard scalar** on it to ensure it has the same columns present as that in our final train dataset.

  **x_test[numeric_cols] = scaler.transform(x_test[numeric_cols])**
- Next, we perform same predictions on the Test set and save them to a new dataframe.
- The **AUC score for test data = 0.87** indicating strong stability of the model.

## Test and Train Data Metrics Comparison

**Train Data :**

1. Accuracy : 78.17
2. Sensitivity : 81.75
3. Specificity : 76
4. Precision : 67.5
5. Recall : 81.75




**Test Data :**

1. Accuracy : 78.11
2. Sensitivity : 82.95
3. Specificity : 75
4. Precision : 67.63
5. Recall : 82.95

The similarity between Train-Test metrics obtained, indicate that our Model is accurately able to predict the results by a very good margin.

```
y_pred_final= pd.concat([y_test_df, y_pred_df], axis=1)
y_pred_final.head()
```

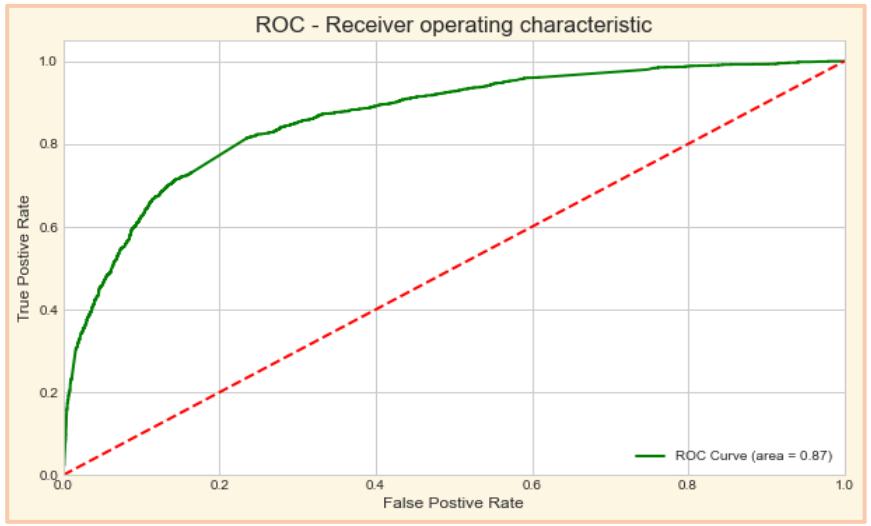
| | Converted | Prospect ID | 0 |
|---|---|---|---|
| 0 | 0 | 4595 | 0.76 |
| 1 | 0 | 898 | 0.19 |
| 2 | 0 | 6597 | 0.19 |
| 3 | 0 | 8303 | 0.01 |
| 4 | 0 | 4300 | 0.36 |

```
# Let's rename column 0 as Converted_Probability

y_pred_final = y_pred_final.rename(columns={0: 'Converted_Probability'})
```

```
y_pred_final.head()
```

| | Converted | Prospect ID | Converted_Probability |
|---|---|---|---|
| 0 | 0 | 4595 | 0.76 |
| 1 | 0 | 898 | 0.19 |
| 2 | 0 | 6597 | 0.19 |
| 3 | 0 | 8303 | 0.01 |
| 4 | 0 | 4300 | 0.36 |



# Lead Score Assignment

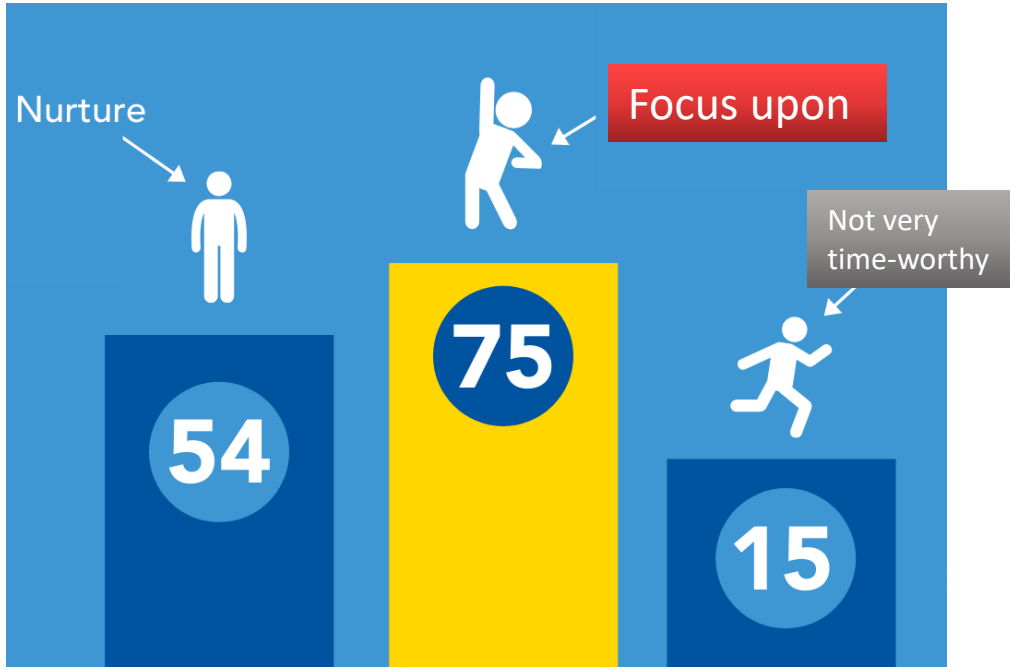
Created **Lead Score** on the Test Dataset in order to identify **HOT LEADS** (with higher scores)**,** so as to be able to focus on the right set of Leads.

Higher the Lead score, higher the chances of them getting converted and higher the need to focus on them**.**

This was done by calculating the converted probability and then converting it to percentage :
y_pred_final.Converted_Probability.map(lambda x: round(x*100))

Results obtained were the Lead Scores : between 0 -100 based on the correct probability of a lead getting converted as predicted by our model

This will give a lot of clarity to the company to focus on the right set of customers and improve further on their Conversion Rate.




Lead Score Table (with first 5 records)

| Prospect ID | Converted | Converted_Probability | Lead_Score | Final_predicted |
|---|---|---|---|---|
| 4595 | 0 | 0.74 | 74 | 1 |
| 898 | 0 | 0.16 | 16 | 0 |
| 6597 | 0 | 0.17 | 14 | 0 |
| 8303 | 0 | 0.01 | 1 | 0 |
| 4300 | 0 | 0.37 | 37 | 1 |

# Insights and Conclusion

- Crucial features which contribute more towards the conversion of Leads are (descending order of importance):

Top 3 Features post StatsModel applied (final model) :

1. Lead Origin_Lead Add Form
2. What is your current occupation_Working Professional
3. Lead Source_Welingak Website

- As per Business term & conditions, the model has an ability to adjust and perform well with the company's changing requirements in future : **Recall score** : 82.95 and **Precision** : 67.63 [Recall > Precision ensures no potential leads will be missed]

- F1-SCORE : 74.51

- Accuracy : 78.11 , Sensitivity : 82.95 and Specificity : 75

All the above 3 show promising and high scores on Test set and match well with the training data set metrics indicating that our model is able to correctly predict and classify the Hot leads.


- **AUC also shows a very high value (0.86)** on both train and test data sets indicating a good predictive model.

# WHAT NEEDS TO BE DONE ?

**As per our analysis and final model predictions following are some of the crucial measures that needs to be taken by X Education to improve the Conversion Rate :-**

- **Relying on the data provided by Sales team,** taking the following actions can help :

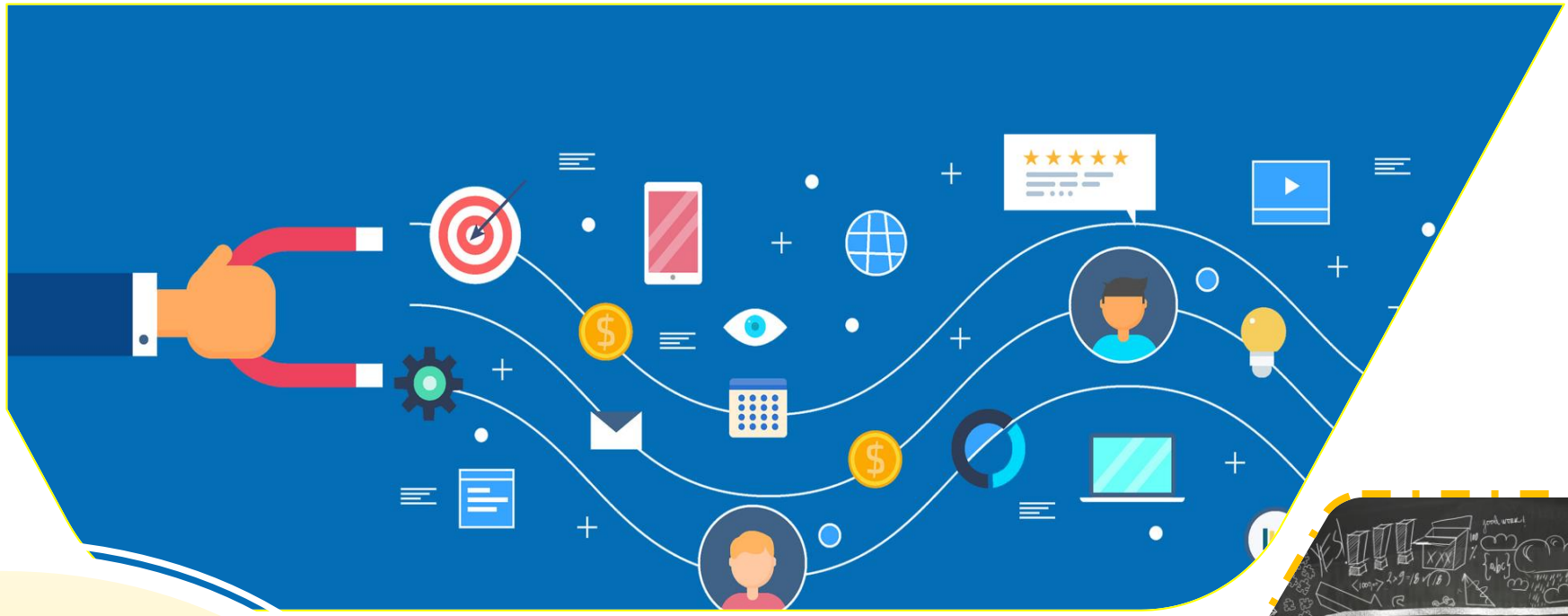
1. Focus more on Leads with "**Will revert after reading email**" and "**Closed by Horizzon**" tags
2. Focus more on leads with : **Last activity marked as SMS Sent and Email Opened**
3. Do not focus upon leads with **Last notable activity : Modified** as they are high in number but very less likely to get converted and hence not worth the time.

- **As per our final model :**

1. Leads generated through **Lead add form and Welingak Website** show very high conversion rates and should be focused upon
2. Leads who are **working professionals** also show great conversion rates and should be focused upon.
3. Leads who are currently **unemployed, also show good conversion rate**. Focussing more on them and **providing benefits like additional discount** can improve it

- **As per our EDA analysis :**

1. Leads with **Management specialization** are more likely to get converted & must be focused upon.
2. Leads through source : **Reference and Google** have a fair conversion rate, but can be improved further through better advertising and offers.
3. Most Leads look out for this course in search for **Better career prospects** and hence the company can ensure they **include pay hike / job switch/ promotion examples and real stories from their alumini on their website in order to gain more leads.**



# Thank You