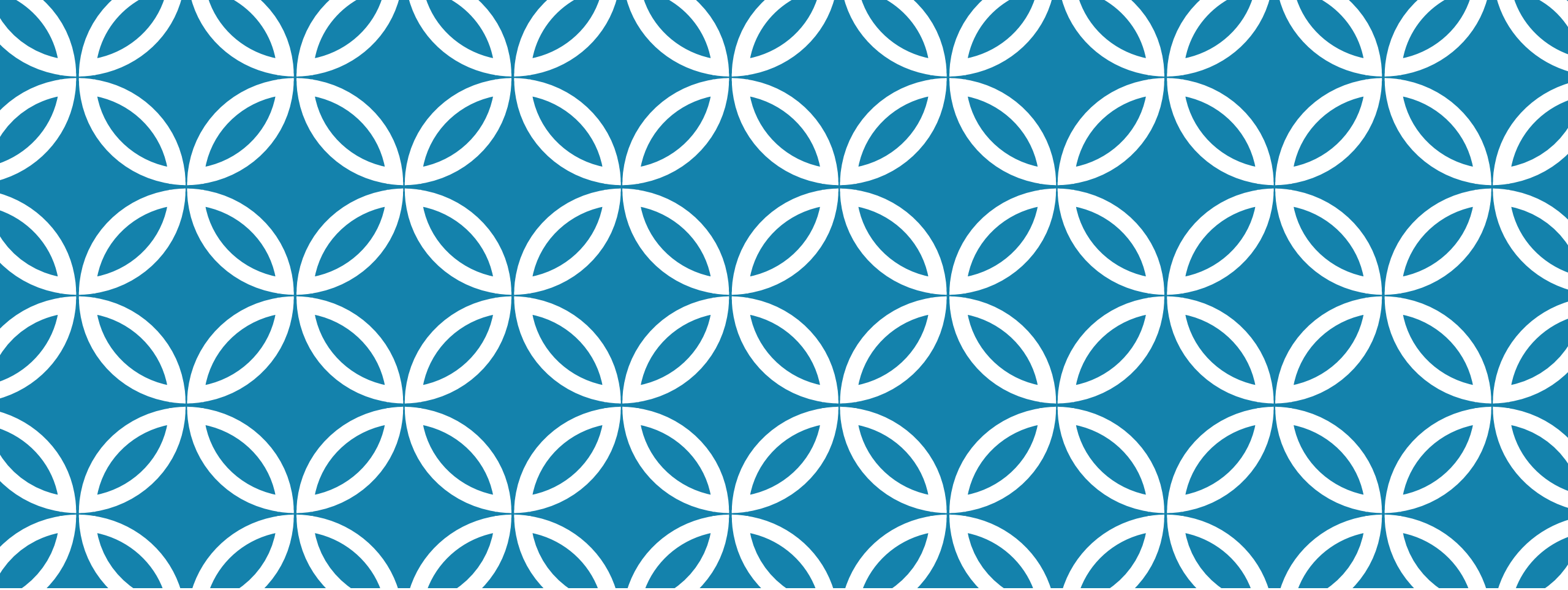


# LEAD SCORE CASE STUDY

Arpith R K
Dheeraj Sapate
Shiva Krishna

# PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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.

- **Business Objective:**
- X education wants to know the most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for future use.

# OBJECTIVES

- To help the company in selecting the most potential leads, also known as ‘Hot Leads’ whose lead conversion rate is around 80%.
- To build 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 scores have a lower conversion chance.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.

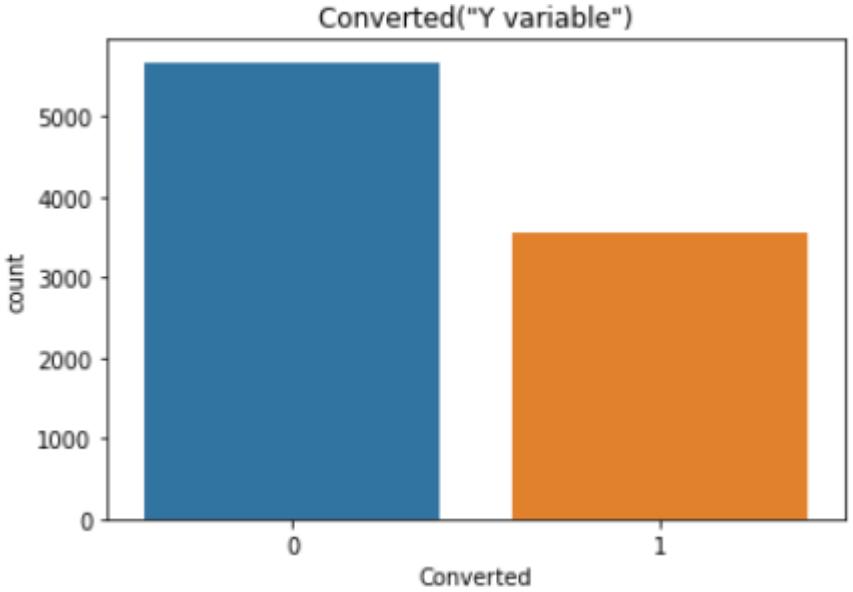
# SOLUTION METHODOLOGY

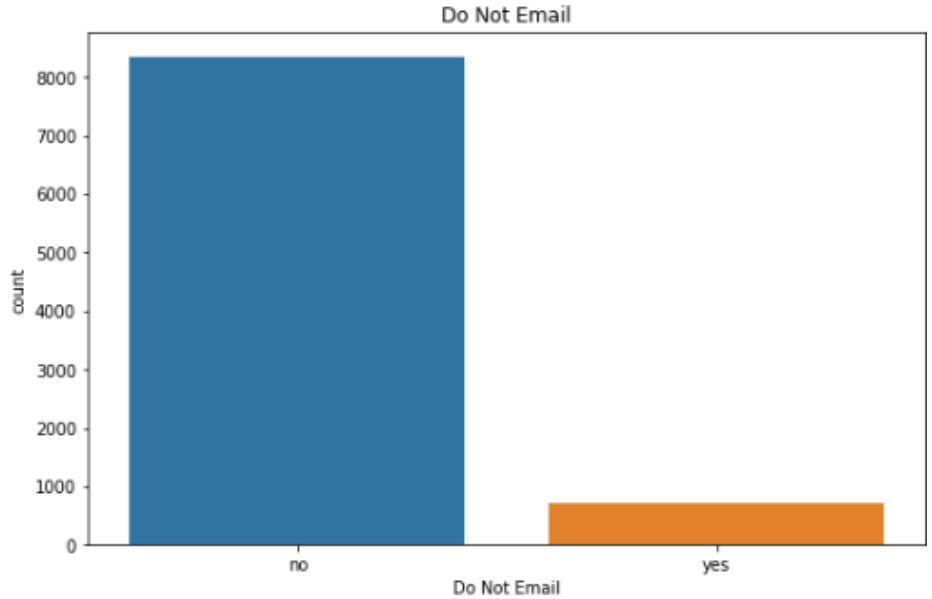
- X Education sells online courses to industry professionals. Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains a large number of missing values and is not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- EDA
    - Univariate data analysis: value count, distribution of variables etc.
    - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression is used for model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

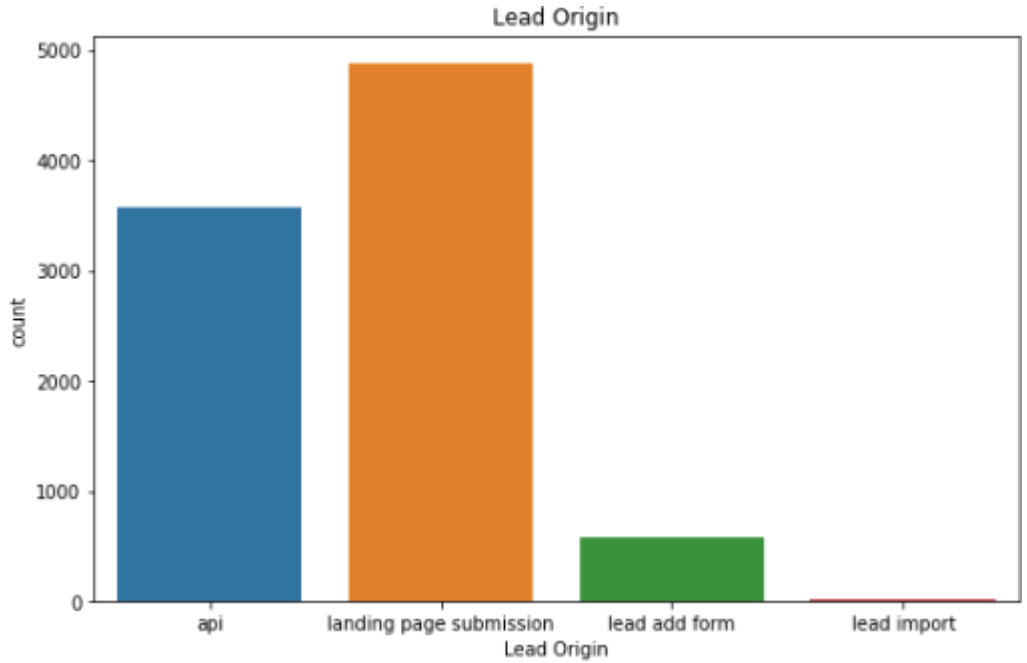
# DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which are not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which have no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper.

  Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as a missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.


Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

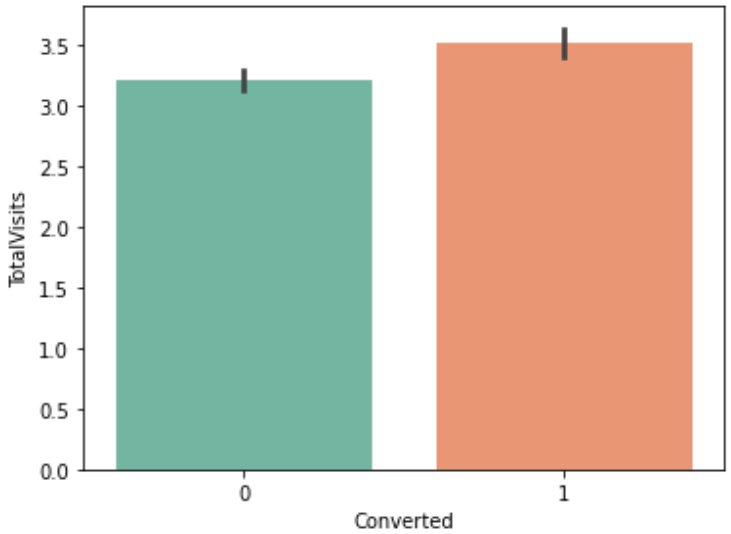

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

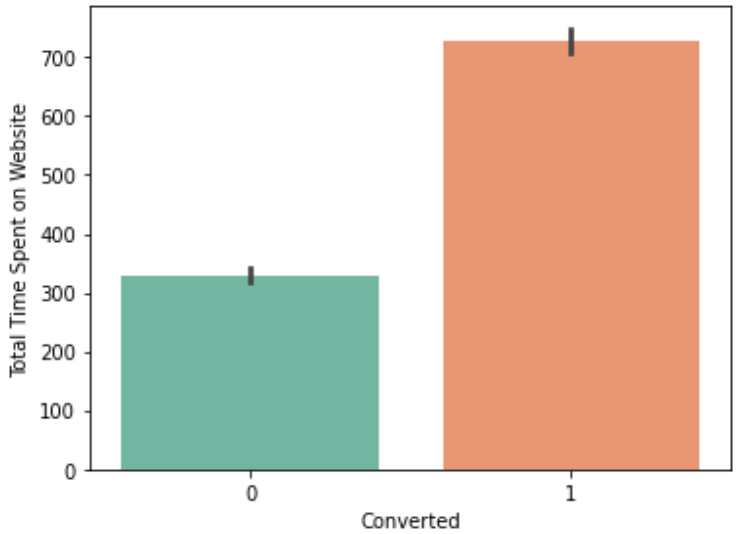

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

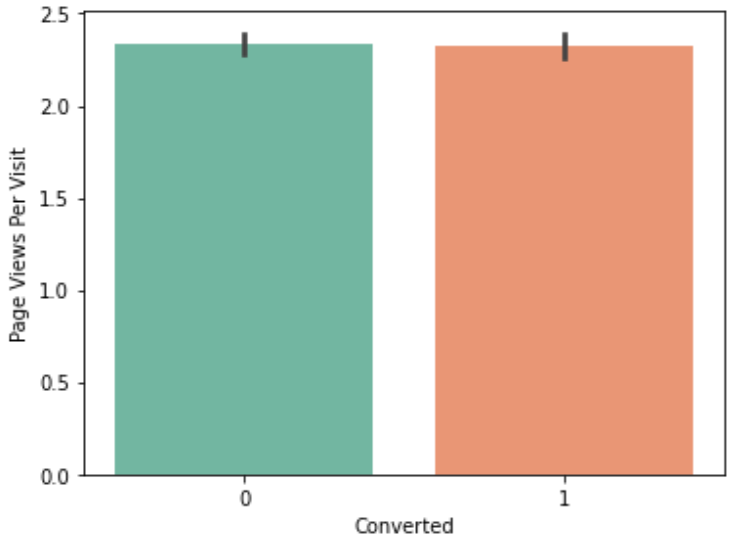
# DATA CLEANING

- Handling the ‘Select’ variable: “Select” variable indicates that the user has not selected any option, We impute the same with null values.
- Dropping column with high null values: Columns having null values greater than 40% does not have meaning to the data, hence we drop these
- Treating Categorical data: Columns having high data imbalance must be removed. For e.g. : Category A has 98%, and Category B has 2% -This data is irrelevant to our analysis as one category is overpowering the other.
- Columns which do not make any insights for our analysis are removed.

# EDA ANALYSIS

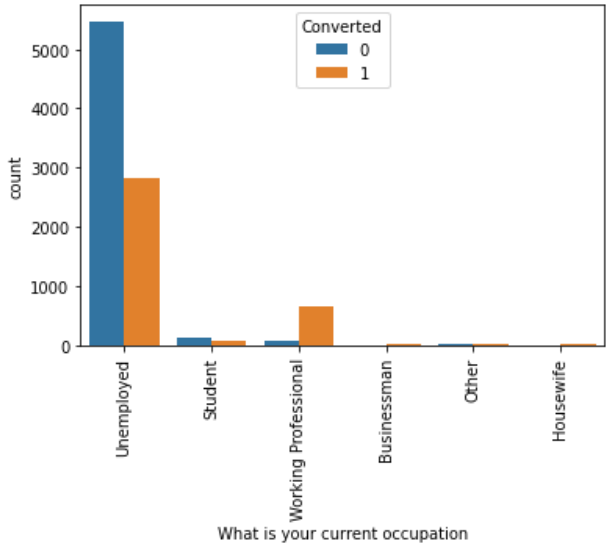


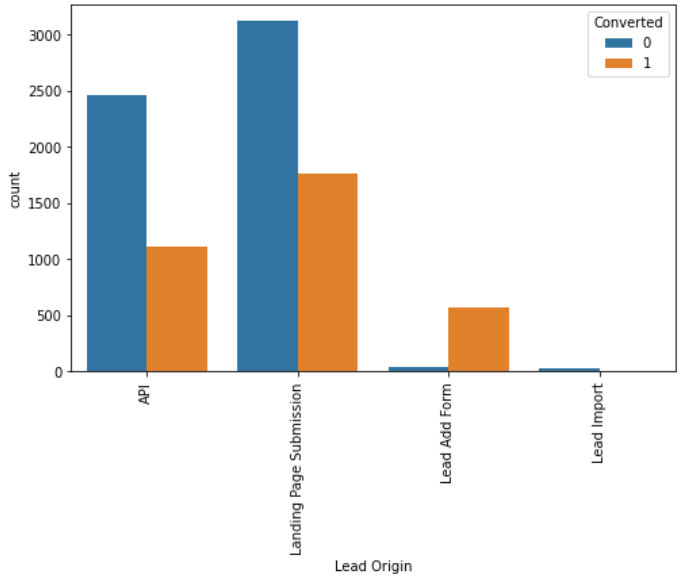



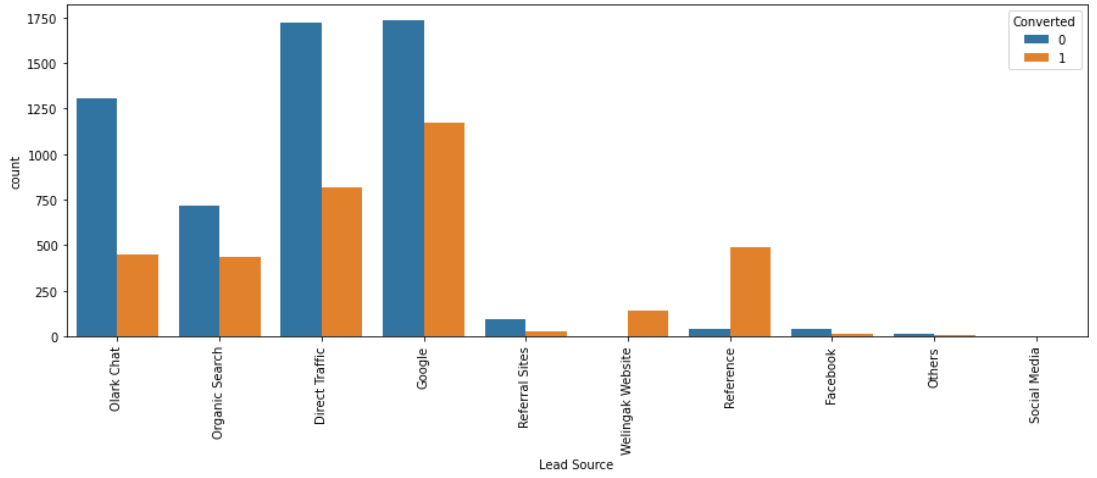




- As the median for both converted and non-converted leads are the same, nothing conclusive can be said on the basis of variable TotalVisits
- Leads spend more time on the website therefore more engagement should be shown on the website.
- Median for converted and unconverted leads is the same, Nothing can be said specifically for lead conversion from Page Views Per Visit

# CATEGORICAL VARIABLE RELATION



# CATEGORICAL VARIABLE RELATION

- Working professionals have a high conversion rate.
- Unemployed personnel are most in terms of numbers.
- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.
- API and Landing Page Submission bring a higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but the count of leads is not very high.
- Lead Import and Quick Add Form get very few leads.


- We can observe that the variables are not highly correlated with each other. But still, there is multicollinearity among some features

# FACTORS RESPONSIBLE IN DRIVING LEADS

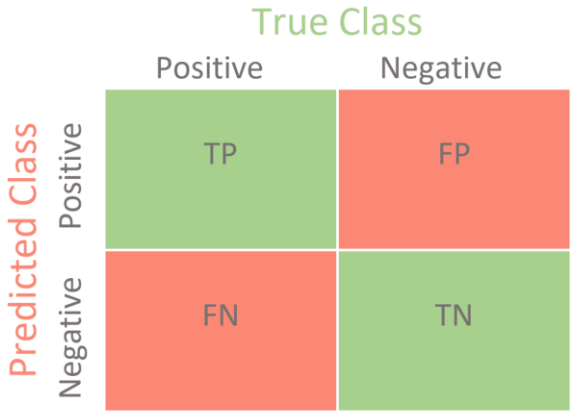
The factors mentioned below are the ones responsible in lead conversion:

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- What is your current occupation_Working Professional
- Lead Source_Olark Chat
- Last Activity_Converted to Lead
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent

# TERMINOLOGIES REQUIRED

Before proceeding ahead, we need to understand a few terminologies

- **Conversion of categorical columns to numerical.** This step is done as our algorithm runs only on numerical data.
- **Feature Scaling.** This is done to bring our data to the same scale.
- **Data Splitting:** We have split the data into 80:20 and named it to train data and test data. We run the model on train data and validate our model on test data.
- **Confusion Matrix:** Where,




True positive (TP): correct positive prediction
False positive (FP): incorrect positive prediction
True negative (TN): correct negative prediction
False negative (FN): incorrect negative prediction

The above metrics are  known as Confusion Metrics, using the above we  have derived the following conclusions:

1. **Accuracy** = (True Negative + True Positive)/Total

    This metric provides the accuracy of the model, where the total is TP + FN + FP +FN

2. **Sensitivity** = True Positive / (True Positive + False Positive)

    **Sensitivity** (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best **sensitivity** is 1.0, whereas the worst is 0.0.

3. **Specificity =** True Negative/ (True Negative + False Negative)

    **Specificity** (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best **specificity** is 1.0, whereas the worst is 0.0.

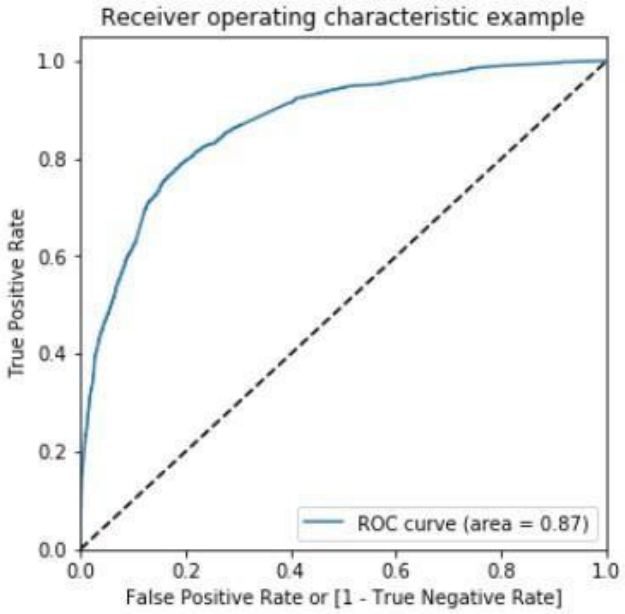
4. **Precision =** True Positive/ (True Positives +False Positives)

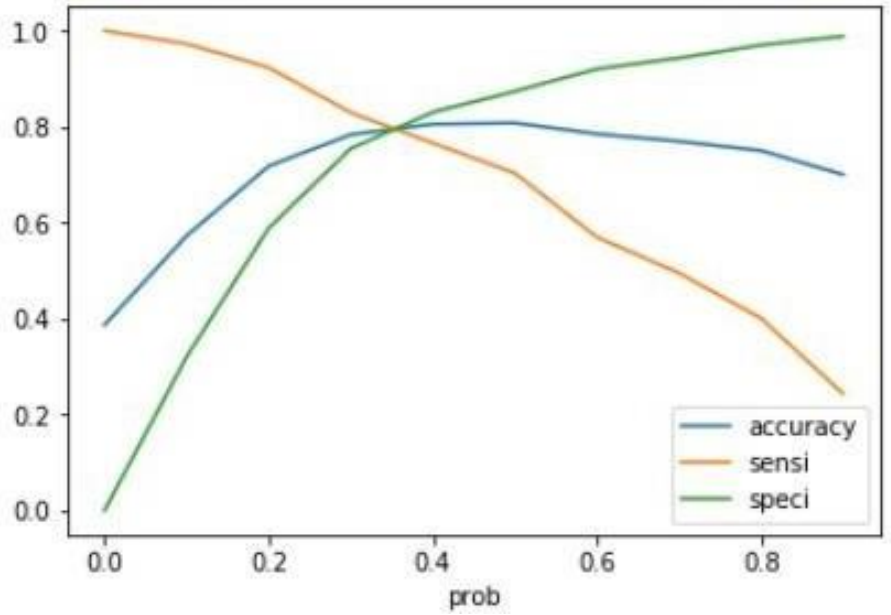
    **Precision** is defined as the number of true positives divided by the number of true positives plus the number of false positives.

5. **Recall** = True Positives/(True Positives +False Negatives)

    The precise definition of **recall** is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect).

# ROC CURVE




- **Finding Optimal Cut off Point**
- Optimal cut-off probability is that
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut-off is at 0.35.

# FINAL OBSERVATIONS

**1. Train Data:**

- Accuracy: 80.5%

- Sensitivity: 79.4%

- Specificity: 81.1%

- Precision: 72.05%

- Recall: 79.41%

**2. Test Data:**

- Accuracy: 80.88%

- Sensitivity: 76.1%

- Specificity: 83.77%

- Precision: 74.34%

- Recall: 76.19%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

# CONCLUSION

- The model seems to predict the conversion rate very well and we should be able to give the CEO confidence in making good calls based on this model.
- We must majorly focus on working professionals.
- It's always good to focus on customers, who have spent significant time on our website.
- More time spent on the website was when the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
- We must majorly focus on leads whose last activity is an SMS sent or Email opened.

**Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their minds and buy their courses.**

# RECOMMENDATIONS

**The company should make calls to the leads :**

- coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- who are the "working professionals" as they are more likely to get converted?
- who spent "more time on the websites" as these are more likely to get converted.
- coming from the lead source "Olark Chat" as these are more likely to get converted.
- whose last activity was SMS Sent as they are more likely to get converted.
- whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- whose Specialization was "Others" as they are not likely to get converted.
- who chose the option of "Do not Email" as "yes" as they are not likely to get converted.