# LEAD SCORING CASE STUDY – X EDUCATION

DS C50 – TANMAY PRIYADARSHI , AISHWARYA AVINASH AND SAHANA UMESH

# INDEX

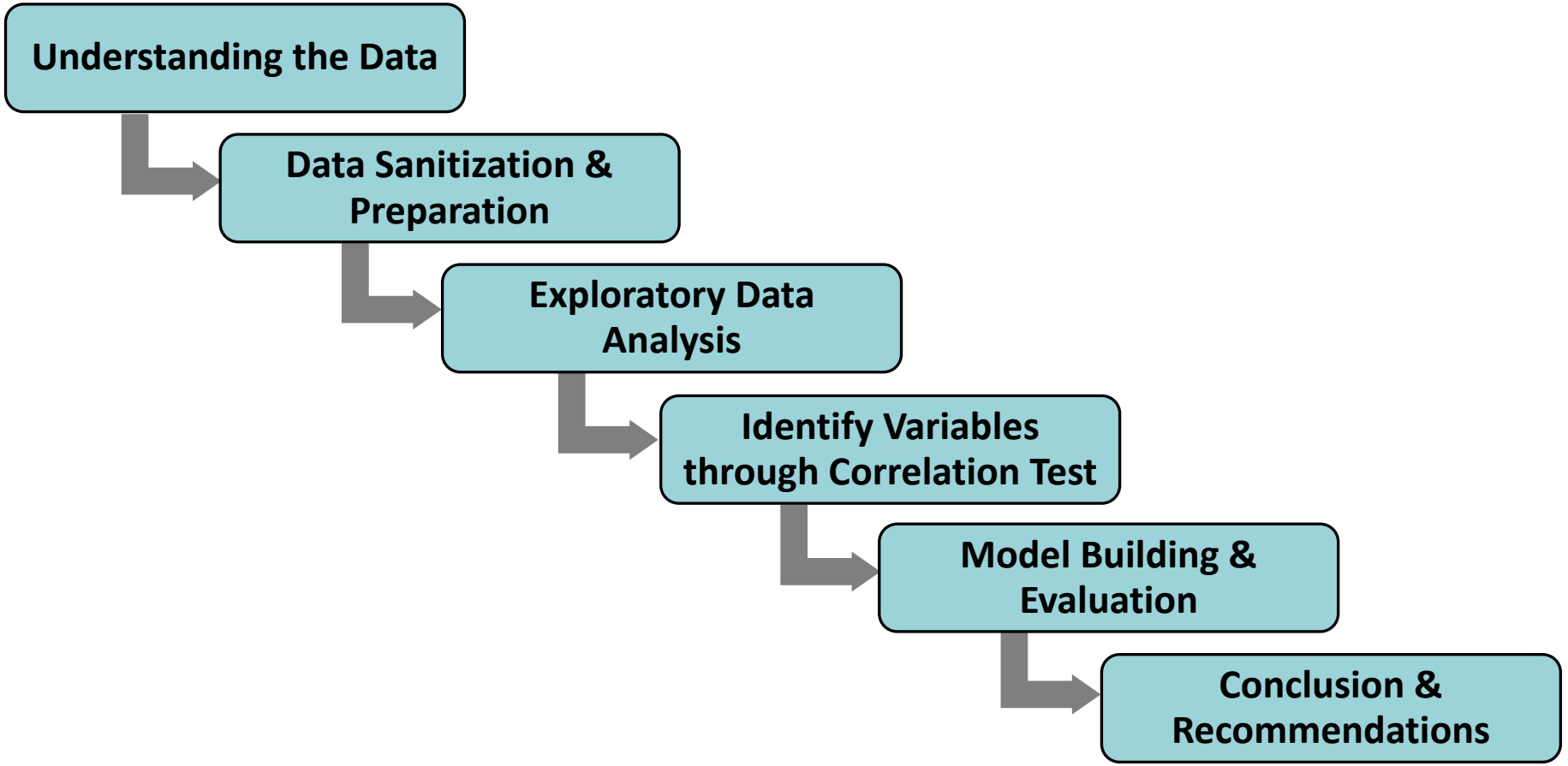
# INTRODUCTION TO PROBLEM STATEMENT

- X Education sells courses online for industry professionals. Everyday, various professionals who are keen on such courses browse for courses on their website.
- Those interested may browse the courses or fill up a form for the course or watch some videos, which is known as a lead. Once the leads are acquired, the sales team will start interacting with the leads to convert them into customers.
- Through the above explained process, so far the company is able to reach 30% lead conversion rate.
- To make the lead scoring process more efficient, the has now decided to identify the most potential leads, known as 'Hot Leads' and focus on communicating more with these potential leads rather than with everyone.

**The goal of the case study -**

1. Logistic Regression Modelling:
   To build a logistic regression model assigning a lead score between 0 and 100 for X Education to target potential leads. A higher score implies that the lead is hot and is most likely to convert into a customer.
2. Recommendations:
   The problems presented should be resolved by the model to fit the company's requirements that may change in the future. The model should be adjustable based on the needs of X Education

# APPROACH TAKEN



# DATA PREPARATION

- Categorical variables having value “Select” pertains to cases where no option was chosen. It is treated as a null value. “Select” has been replaced with NaN for the same.
- Columns with null values greater than 40% in the dataset have been dropped.
- Missing values are imputed with the modal values for each column respectively where suitable.
- Columns that are not useful in building the regression model or have only one unique value have been dropped.
- Highly skewed columns can affect the result of logistic regression models, as they can lead to biased results or inaccurate parameter estimation. Therefore these column can be dropped as they will not add any value to the model.

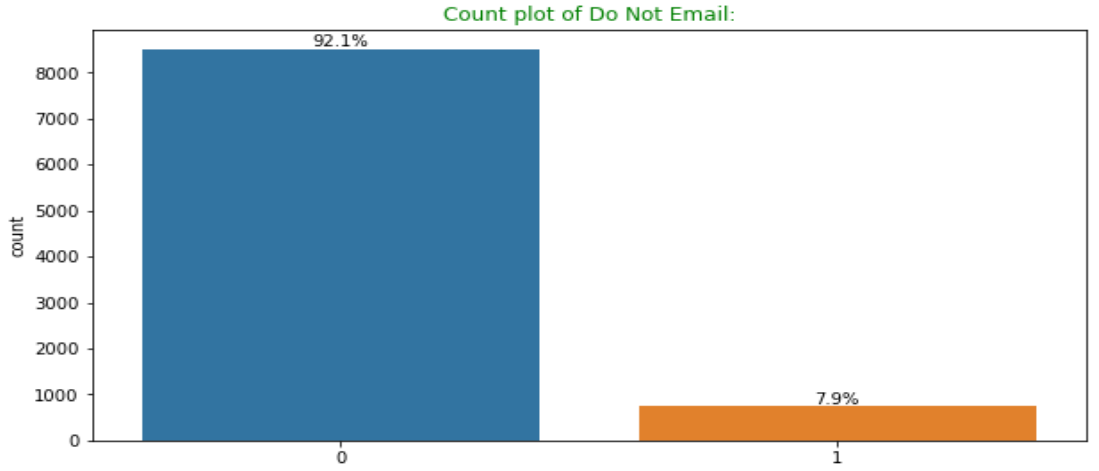
# DATA STANDARDIZATION & OUTLIER CHECK

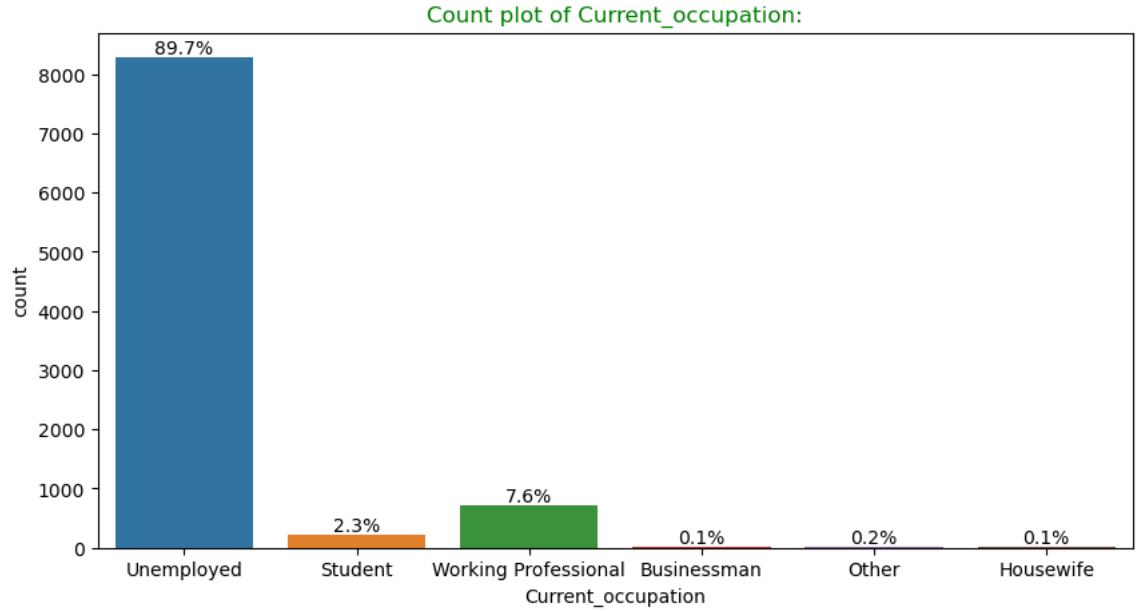
- The columns "Lead Score" and "Last Activity" have a number of labels whose value count is negligible and therefore are grouped together under "Others" to remove columns that are not necessary for regression analysis.
- "Free_copy" & "Do Not Email" are both binary categorical columns. To standardize the same, the values ‘yes’ and ‘no’ to 0 & 1.
- Inconsistencies in case, for example "Google" & "google" are same in "Lead Source", are sanitized by replacing lower case value to upper case, and standardize the data.
- Outliers in the data checked using boxplots. They are treated by defining the upper and lower limits and replacing the values that lie outside the defined range.

# EDA – UNIVARIATE ANALYSIS

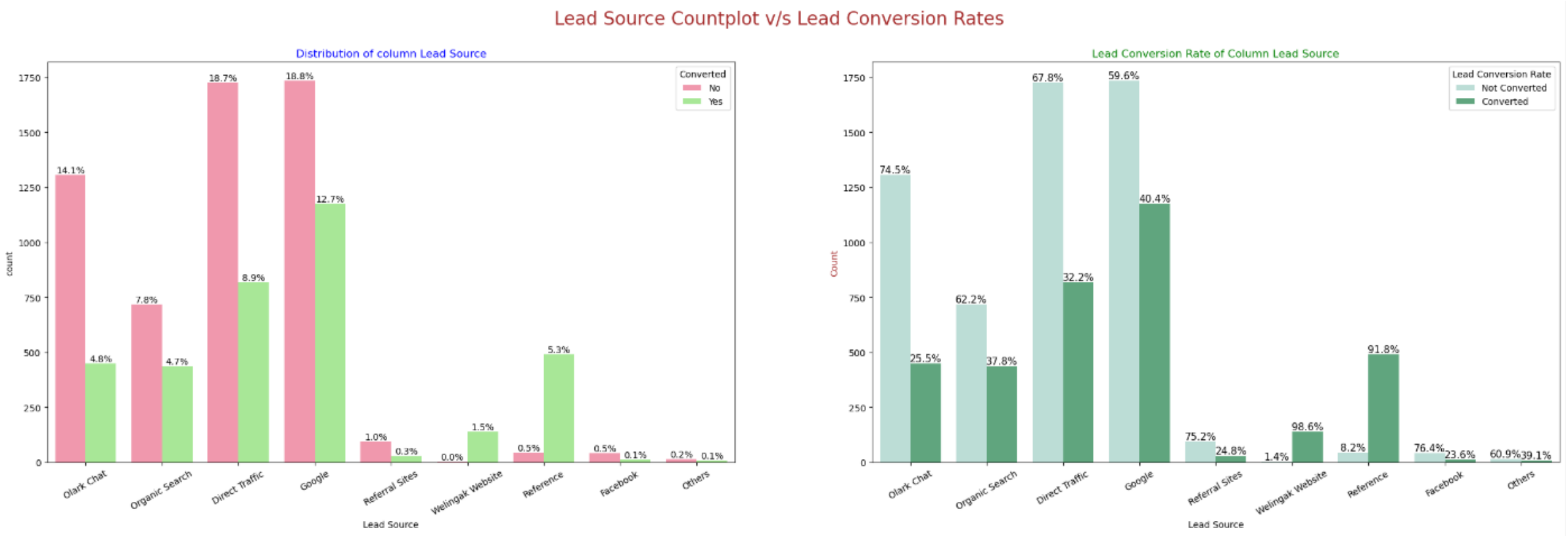
**Insights From Univariate Analysis:**

- Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.
- Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.



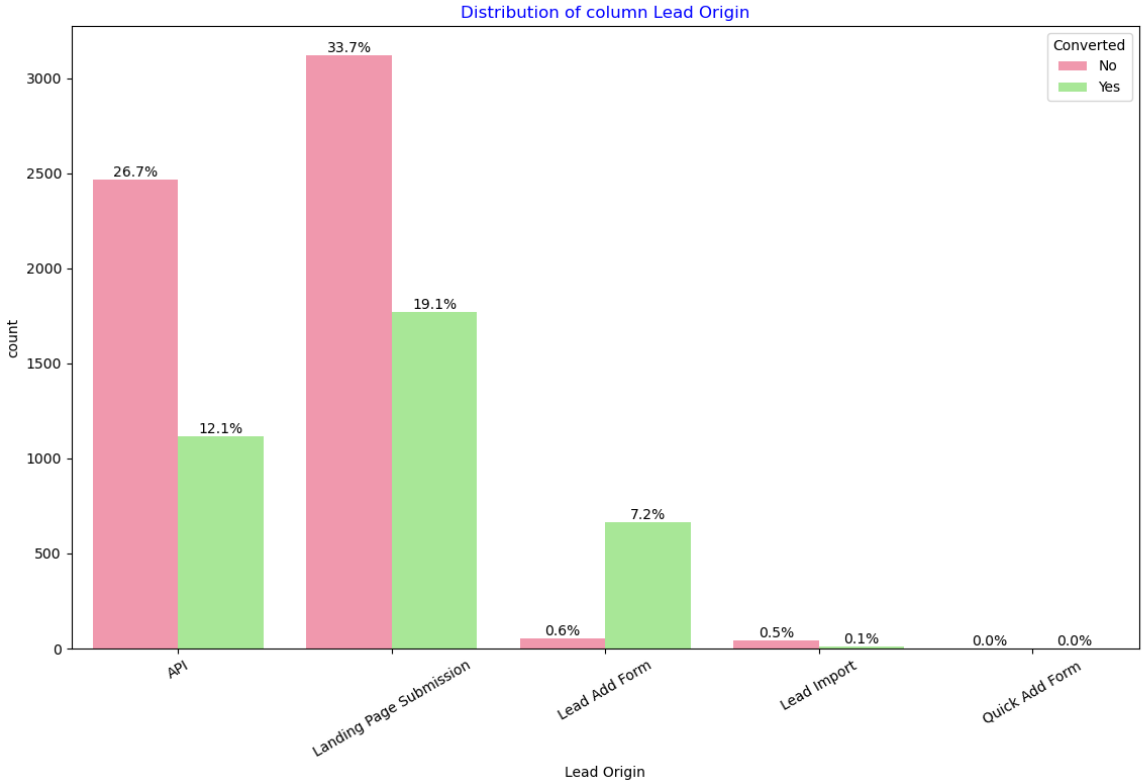
# EDA – BIVARIATE ANALYSIS

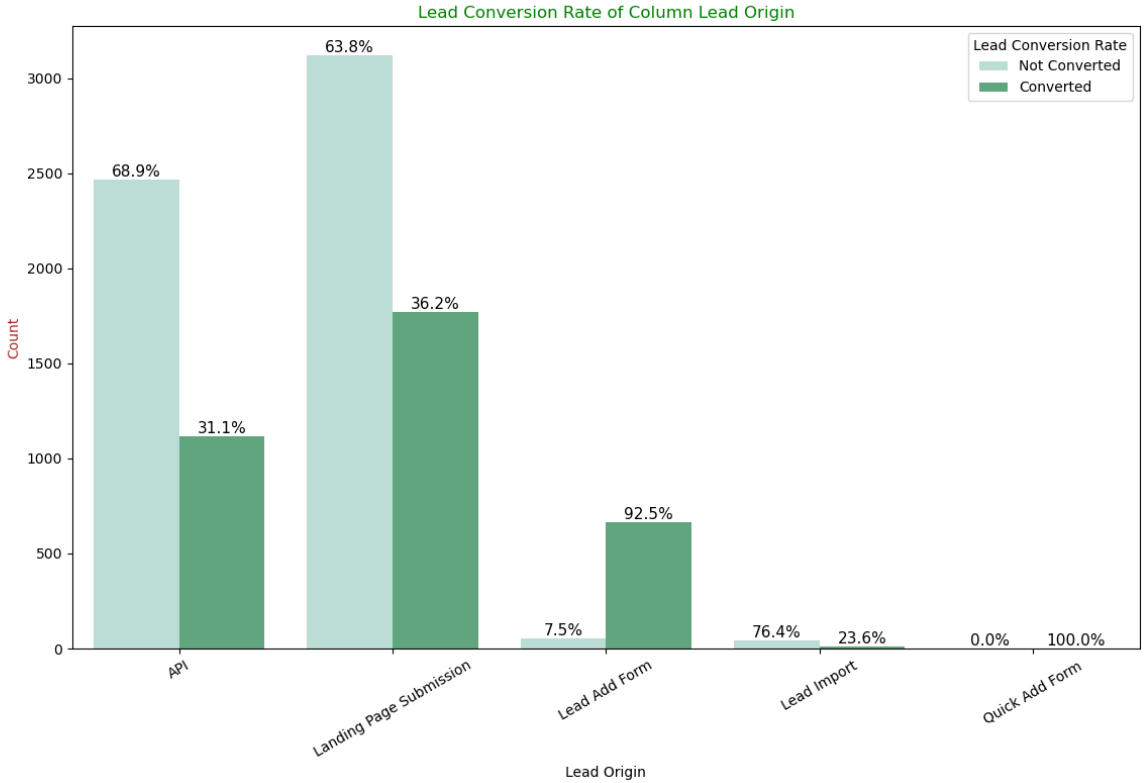



*Google* converts 40% leads out of 31% customers , *Direct Traffic* contributes 32%, *Organic Search* also gives 37.8% conversion rate but contributes to merely 12.5% of customers. *Reference* converts 91% but there are only around 6% of customers through this Lead Source. Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

# EDA – BIVARIATE ANALYSIS

Lead Origin Countplot v/s Lead Conversion Rates



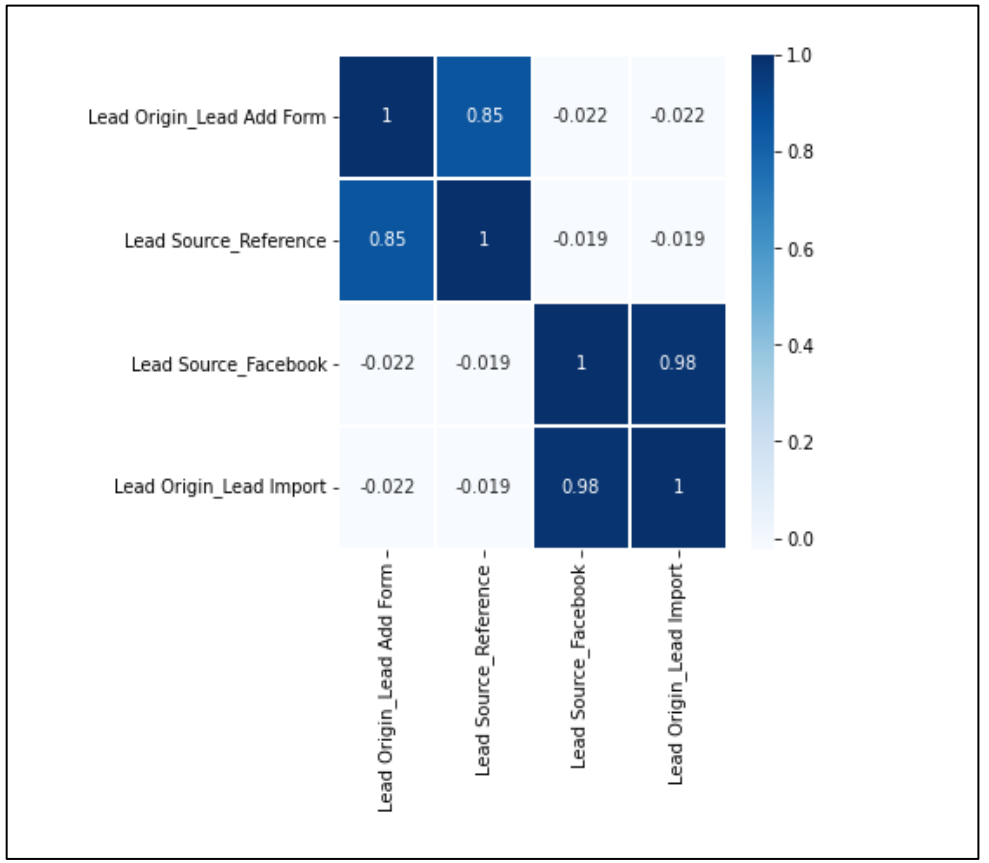

Lead Origin: Around 52% of all leads originated from *"Landing Page Submission"* with a lead conversion rate (LCR) of 36%.The *"API"* identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

# MODEL BUILDING

- A logistic Regression Model has been used for predicting categorical variables in the case study.
- The process involved two stages for selecting appropriate features: RFE (Recursive Feature Elimination) with coarse tuning and manual fine-tuning using p-values and VIFs (Variance Inflation Factors).
- The steps undertaken in building the model are as follows:
  - Creation of Dummy Variables
  - Splitting the Dataset into train and test set
  - Scaling of Features
  - Correlation Checking
  - Feature Elimination based on Correlation Checking
  - Feature Selection Using RFE (Recursive Feature Elimination)
- Using statsmodel, a detailed model is built. The process is repeated 4 times and columns with high p-value above the accepted threshold of 0.05 p-value are dropped.
- Model 4 is ultimately the final model.
- Model 4 is stable and has significant p-values within the threshold (p-values less than 0.05) with VIF <5 for all columns, and thus can be used for further analysis.

# CORRELATION TEST

- The heatmap as shown in the slide is gives an overview of all possible variables with high correlation. It has been derived from a correlation matrix of all variables in the dataset.
- Predictor 'Lead Origin_Lead Import' is highly correlated with 'Lead Source_Facebook' and 'Lead Origin_Lead Add Form' is also highly correlated with 'Lead Source_Reference' with 0.98 and 0.85 as correlation values respectively . Therefore 'Lead Origin_Lead Import' and 'Lead Origin_Lead Add Form’ are dropped.



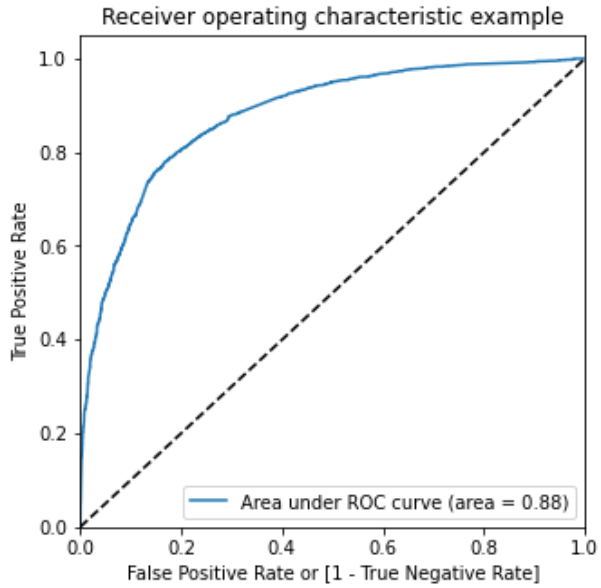
# MODEL EVALUATION

- Tools used in evaluating the model:
  - Confusion Matrix
  - Accuracy
  - Sensitivity and Specificity
  - Threshold determination using ROC & Finding Optimal cutoff point
  - Precision and Recall

- An ROC curve demonstrates:

  1. any increase in sensitivity will be accompanied by a decrease in specificity. It shows the tradeoff between sensitivity and specificity .
  2. The closer the curve follows the top-left hand border and then the top border of the ROC space, the more accurate the test.
  3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

  **(Area under ROC curve is 0.88 out of 1 which indicates a good predictive model)**

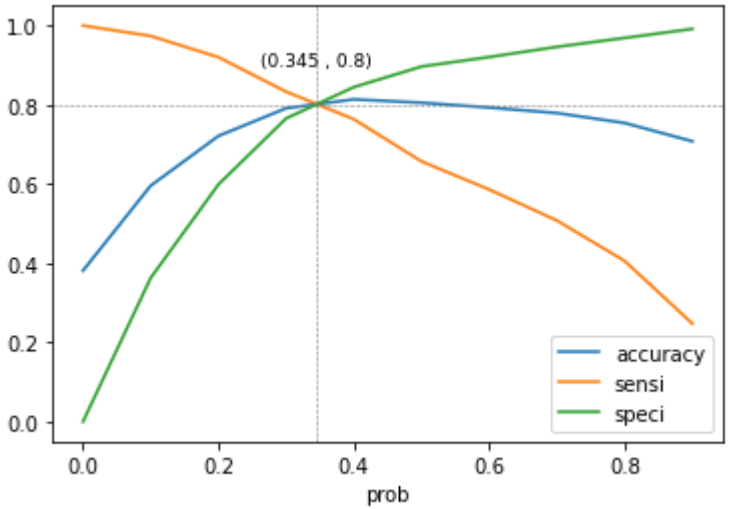



ROC Curve

**Optimal Cut-Off Point:**

To determine the optimal cutoff point or probability, it is necessary to identify the threshold that achieves a balance between sensitivity and specificity.

From the graph shown it is understood that the point 0.345 (approx.) is the optimal cut-off value.

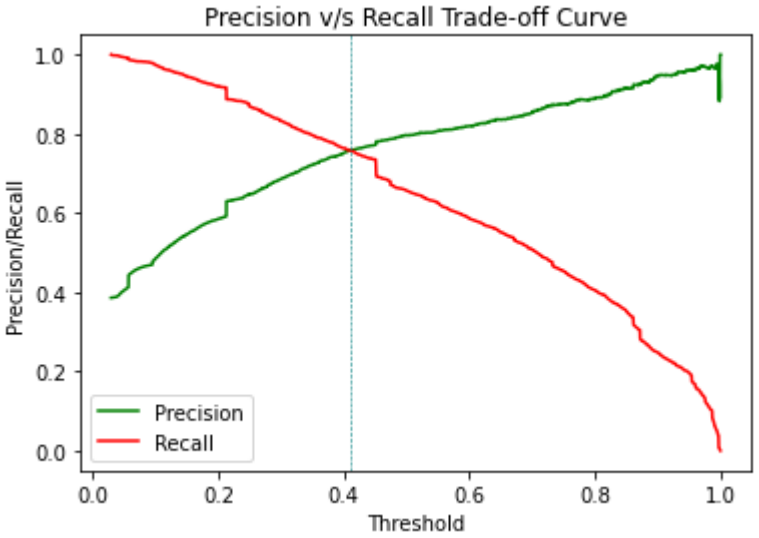



Optimal Cut-off Point

**Precision vs Recall Trade-Off:**

By comparing the Precision-Recall view with the Specificity-Sensitivity view, we can identify the threshold value that provides the best balance between these metrics.

From the above plot we get to know that Precision v/s Recall curve achieves balance at intersection that is the location were find optimal threshold value which in this case is 0.41 approx.



# INSIGHTS

- **Based on the model evaluation:**
  According to Business Requirement the metrices are to be close to 80% and when compared metrices obtained from both Precision-Recall and sensitivity-specificity the values of those metrices drops around 75%.

  We have obtained 80% for the metrics with the sensitivity-specificity cut-off threshold of 0.345. Therefore, we will go ahead with sensitivity-specificity view for our Optimal cut-off for final predictions.

- **Adding Lead Score Feature to Training Dataframe**
  - A higher the score greater is the chance to get converted (i.e HOT) i.e. is most likely to convert
  - Whereas a lower score would mean that the lead is cold and lesser is the chance to get converted.

- The test dataset is scaled to make predictions. Evaluation of the test dataset gives us the following:
  Accuracy : 80.34%
  Sensitivity : 79.82% ≈ 80%(approx)
  Specificity : 80.68%

  **Note**: Since the matrices obtained from test set is very close to train set, it shows that the model lgsm4 is performing consistently.

# CONCLUSION

- The final Logistic Regression Model has 12 features:
  Top 3 features that contribute positively to predicting hot leads in the model are:
  1. Lead Source_Welingak Website
  2. Lead Source_Reference
  3. Current_occupation_Working Professional

- The Optimal threshold/cutoff probability point is 0.345.Converted probability predicted having value greater than 0.345 will be predicted as Converted lead (Hot lead) while those smaller than 0.345 will be predicted as not Converted lead (Cold lead).

- The model also achieved an accuracy of 80.34%, which is in line with the study's objectives.

**Parameters from the Final Model**

```
Lead Source_Welingak Website                5.388662
Lead Source_Reference                       2.925326
Current_occupation_Working Professional     2.669665
Last Activity_SMS Sent                      2.051879
Last Activity_Others                        1.253061
Total Time Spent on Website                 1.049789
Last Activity_Email Opened                  0.942099
Lead Source_Olark Chat                      0.907184
Last Activity_Olark Chat Conversation      -0.555605
const                                      -1.023594
Specialization_Hospitality Management      -1.094445
Specialization_Others                      -1.203333
Lead Origin_Landing Page Submission        -1.258954
dtype: float64
```

# RECOMMENDATIONS

- Focus should be given to the features having positive as well as high coefficients to enhance target market strategies.
- Professionals can be engaged for customized marketing material and tailored infomercials.
- Optimize communication channels to improve lead engagement impact.
- Re-budget to increase spending on Welingak Website as it is the top performer and other such budgeting tactics.
- Commissions to sales force on conversion of hot leads to customers can improve efficiency
- Enhance target market to working professionals as they have high conversion rate as well as the readiness to pay resulting from higher financial position of such leads.