LEAD SCORING CASE STUDY

BY ANKITA, SRAVANI, DEEPAK

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, 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 rateshould go up as the sales team will now be focusing more oncommunicating with the potential leads rather than making calls toeveryone.

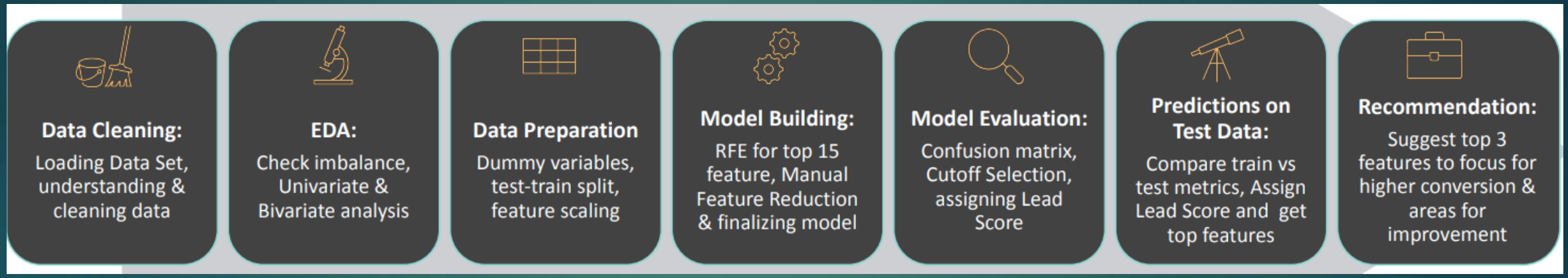
Business Goal

- 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 have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach



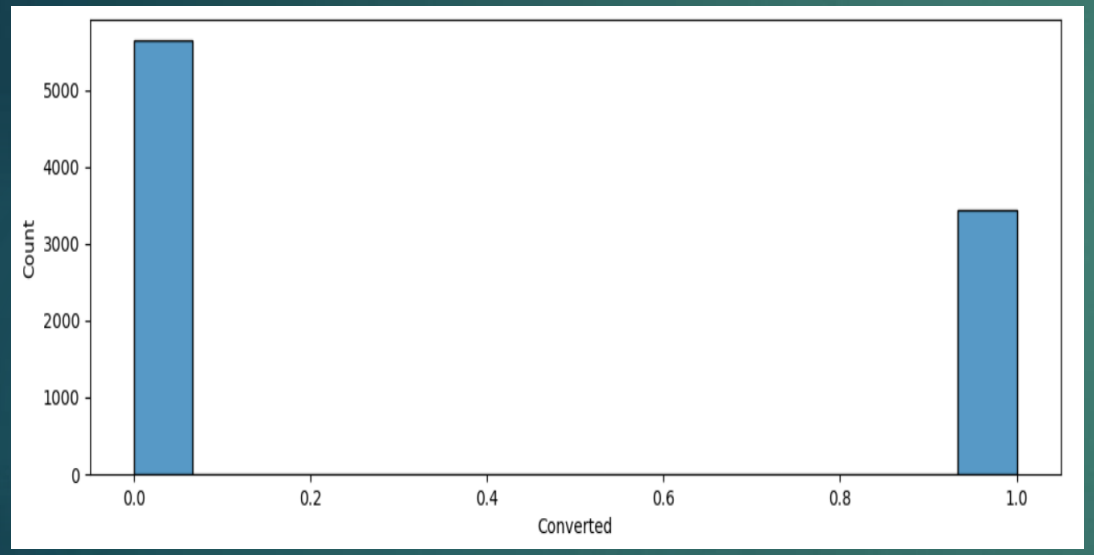
Data Cleaning

- ● "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
  ● Columns with over 40% null values were dropped.
  ● Missing values in categorical columns were handled based on value counts and certain considerations.
  ● Drop columns that don't add any insight or value to the study objective (tags, country)
  ● Imputation was used for some categorical variables using mode.
  ● Additional categories were created for some variables.
  ● Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
  ● Numerical data was imputed with mode after checking distribution.

Data Cleaning

- ● Skewed(unbalanced) category columns were checked and dropped to avoid bias in logistic regression models
- ● Low frequency values were grouped together to “Others”.
- ● Outliers in TotalVisits and Page Views Per Visit were treated and capped.
- ● Binary categorical variables were mapped.

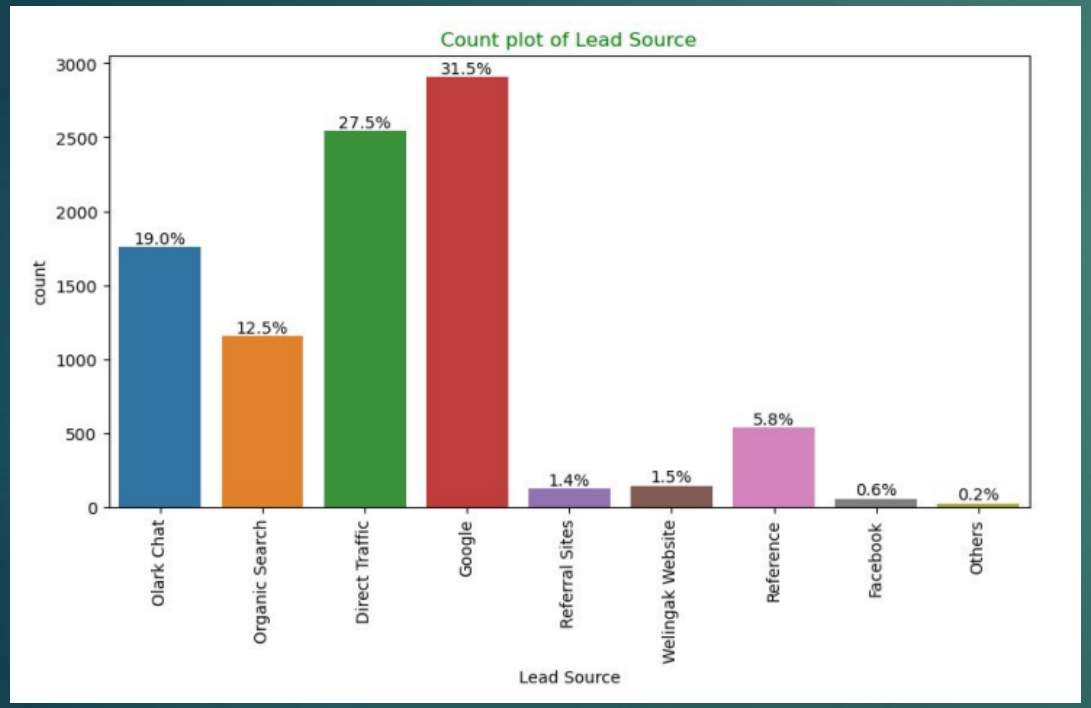
EDA



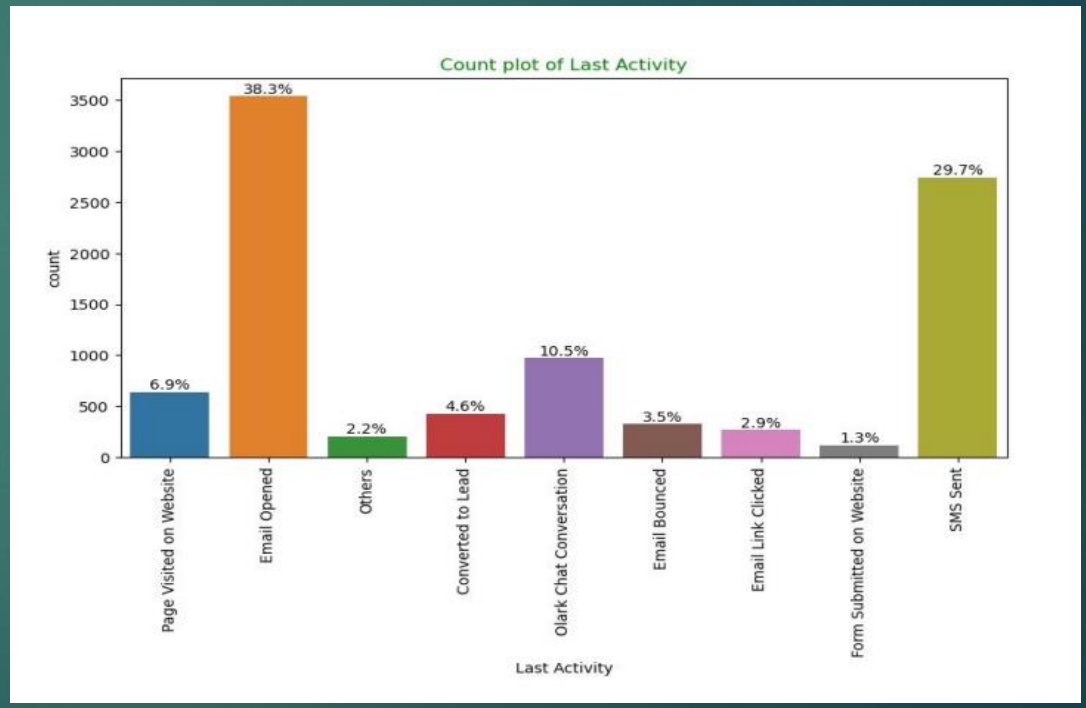
- Data is imbalanced while analyzing target variable.
- • Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- • While 61.5% of the people didn't convert to leads. (Majority)

Univariate Analysis(categorical variables)

- Lead Source: 58% Lead source is from Google & Direct Traffic combined.

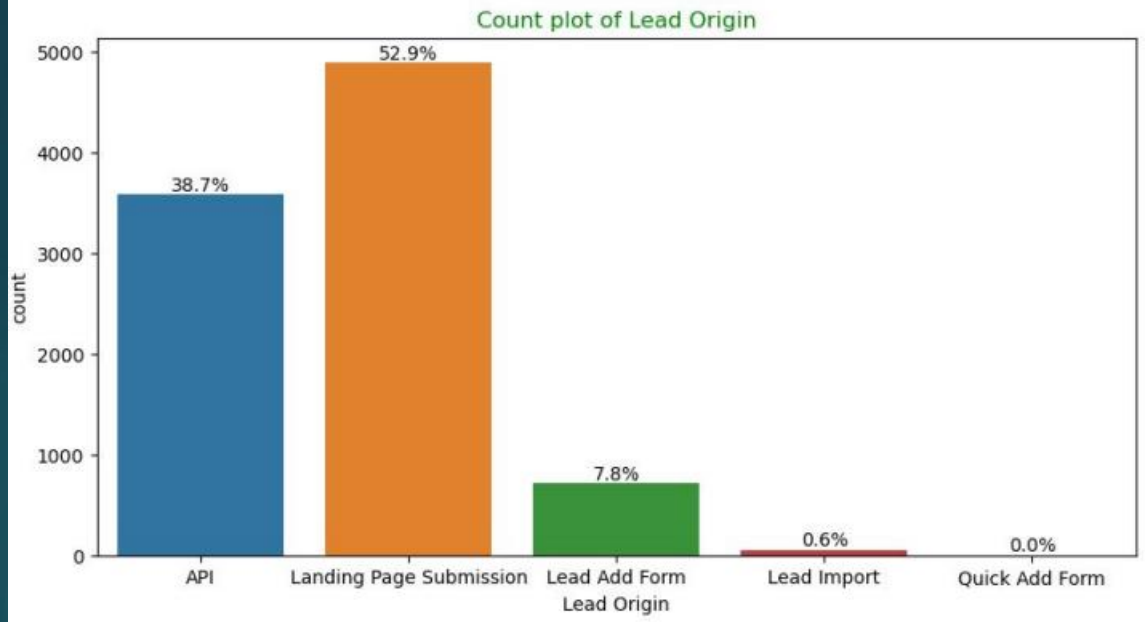


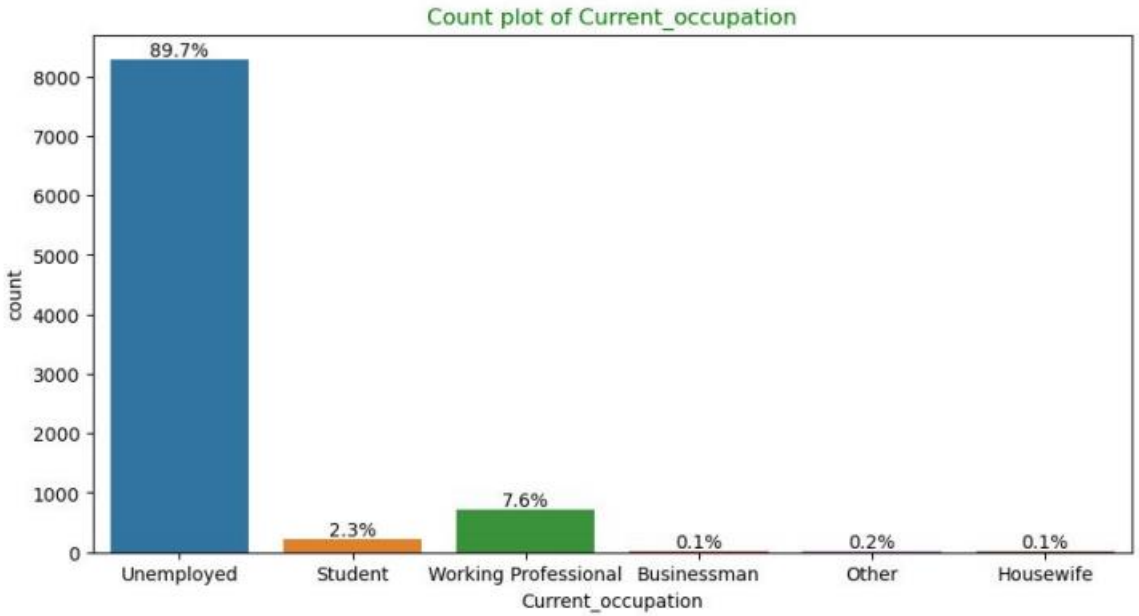
- Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities.



EDA

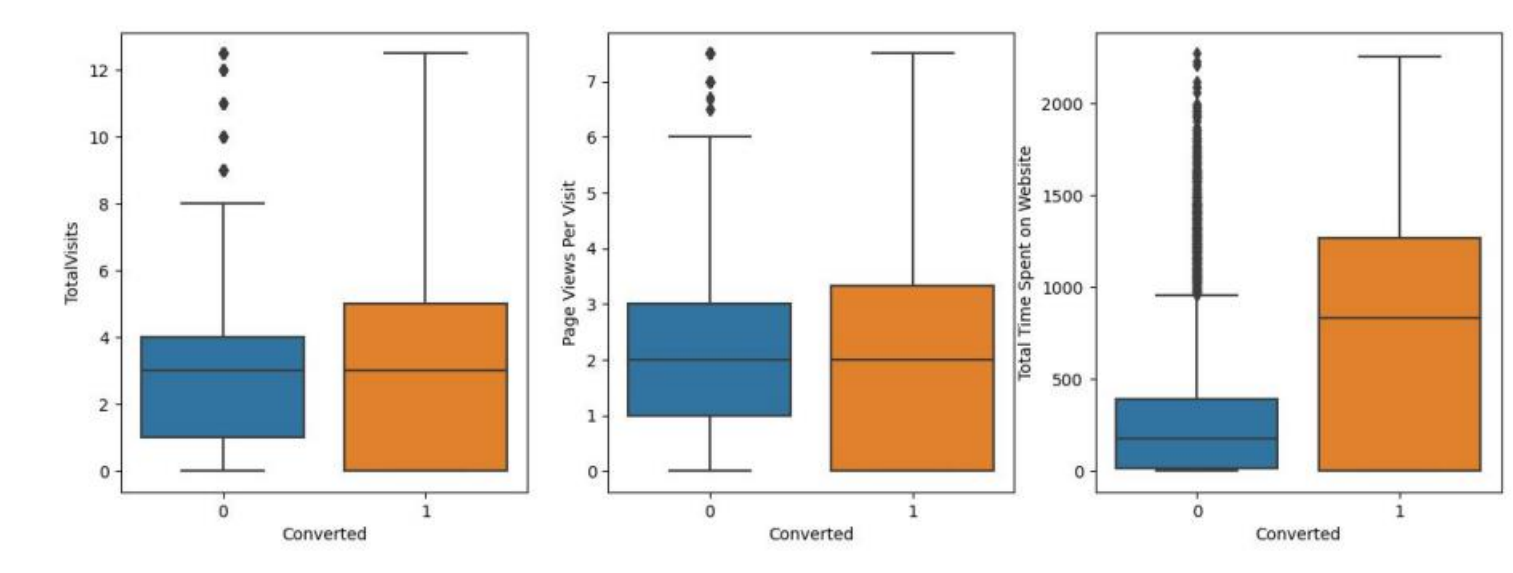
Univariate Analysis – Categorical Variables





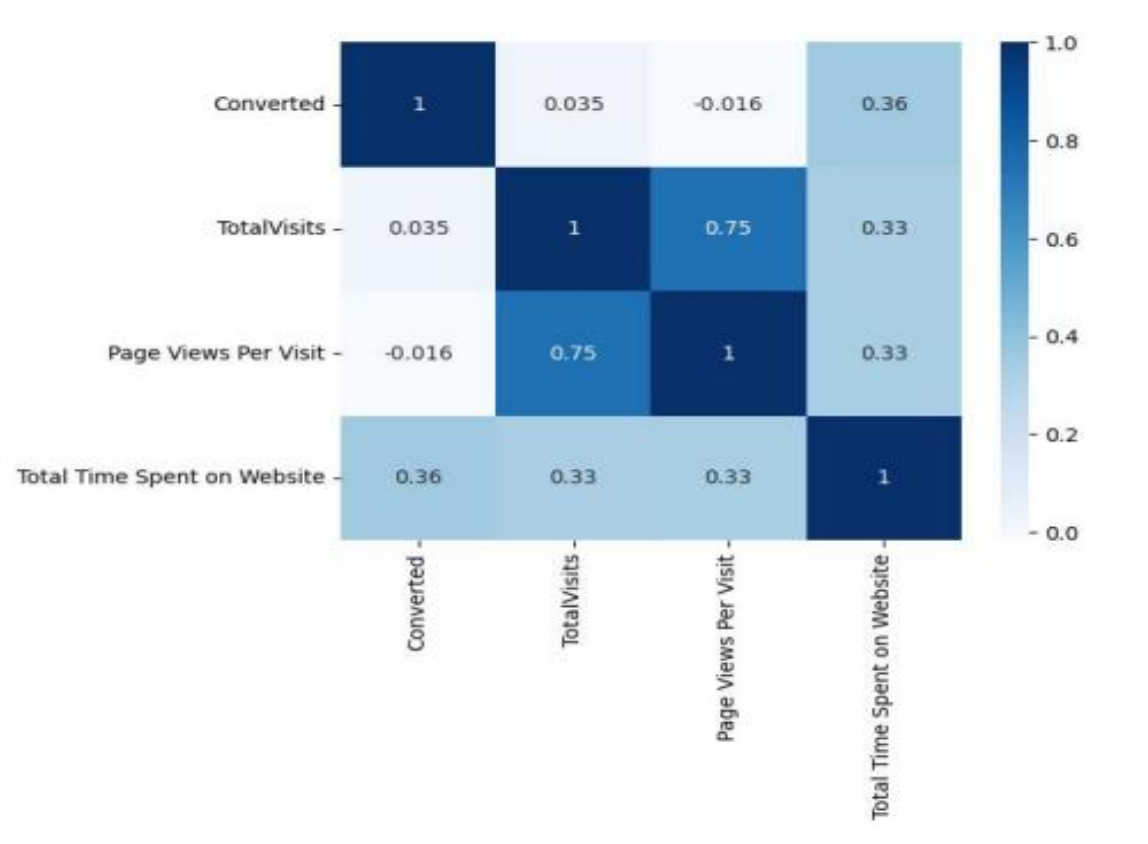
- **Lead Origin:** "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- **Current_occupation:** It has 90% of the customers as Unemployed.

EDA- Bivariate Analysis For Numerical Variables



- 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

Multivariate Analysis



- TotalVisits and Page Views Per Visit are highly correlated variables as visible in the heatmap plot.

Data Preparation before Model building

- ● Binary level categorical columns were already mapped to 1 / 0 in previous steps
- ● Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- ● Splitting Train & Test Sets
- ● 70:30 % ratio was chosen for the split
- ● Feature scaling
- ● Standardization method was used to scale the features
- ● Checking the correlations

Model Building

- Feature Selection
- ● The data set has lots of dimension and large number of features.
- ● This will reduce model performance and might take high computation time.
- ● Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- ● Then we can manually fine tune the model.
- ● RFE outcome
- ● Pre RFE – 47 columns & Post RFE – 15 columns

Model Building

- ● Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- ● Model 3 looks stable after three iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
  - Hence, logm3 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

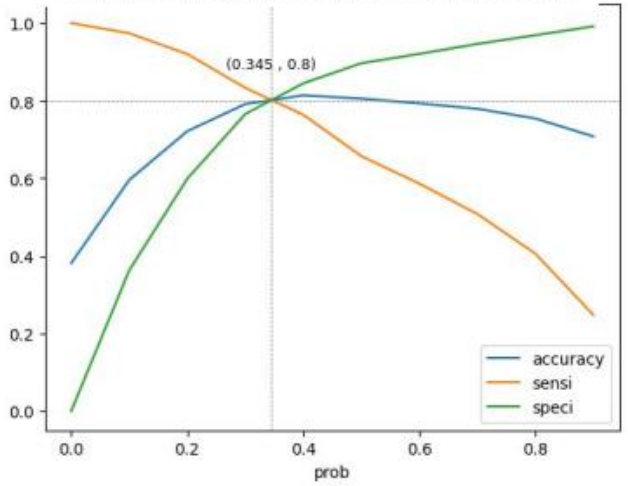
Model Evaluation

Train Data Set

It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

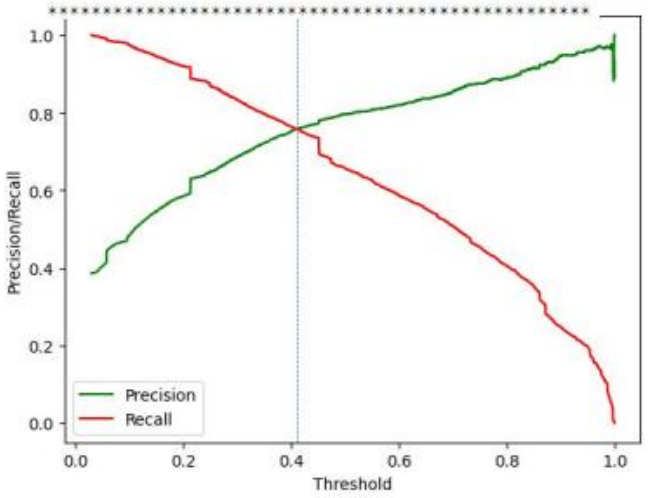
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                  :  3230
True Positive                  :  1974
False Negative                 :  492
False Positve                  :  772
Model Accuracy                 :  0.8046
Model Sensitivity              :  0.8005
Model Specificity              :  0.8071
Model Precision                :  0.7189
Model Recall                   :  0.8005
Model True Positive Rate (TPR) :  0.8005
Model False Positive Rate (FPR) :  0.1929

**************************************************
```



Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

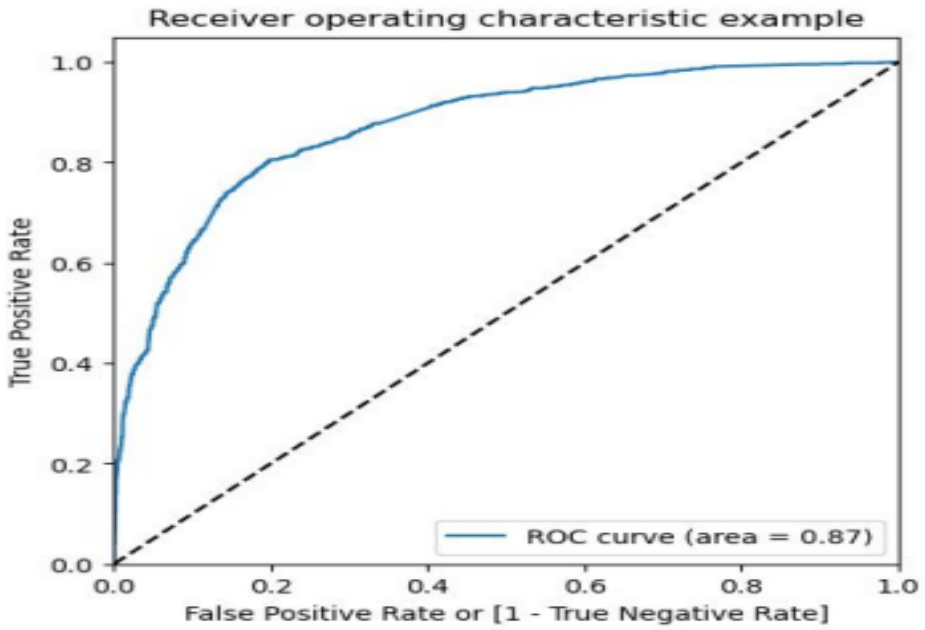
```
**************************************************

Confusion Matrix
[[3406  596]
 [ 596 1870]]

**************************************************

True Negative                  :  3406
True Positive                  :  1870
False Negative                 :  596
False Positve                  :  596
Model Accuracy                 :  0.8157
Model Sensitivity              :  0.7583
Model Specificity              :  0.8511
Model Precision                :  0.7583
Model Recall                   :  0.7583
Model True Positive Rate (TPR) :  0.7583
Model False Positive Rate (FPR) :  0.1489

**************************************************
```



Model Evaluation

ROC Curve – Test Data Set

- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

Train Data Set

```
array([[3145,  760],
       [ 487, 1959]], dtype=int64)
```

Train Data:

- **Accuracy : 81.05 %**
- **Sensitivity : 80.08 %**
- **Specificity : 80.53 %**

Test Data Set

```
array([[1403,  331],
       [ 207,  782]], dtype=int64)
```

Test Data:

- **Accuracy : 80.24 %**
- **Sensitivity : 79.06 %**
- **Specificity : 80.91 %**

Recommendations

- The company should contact the leads who are "working professionals" as they are more likely to get converted.
- The company should contact leads where the lead origin is 'Lead Add Form' as they are more likely to get converted.
- The company should contact leads whose last Activity was 'Phone Conversation' or 'SMS Sent' they are more likely to get converted.
- The company should contact leads coming from the lead sources 'Welingak Website' they are more likely to get converted.
- The company should not contact leads coming from lead_sources 'Google' , 'Organic Search' , 'Direct Traffic','Referral Sites' as they are not likely to get converted.
- The company should not contact leads who have Last_Activity 'Converted to Lead' , 'Olark Chat Conversation','Email Bounced' as they are not likely to get converted.

Thank You