LEAD SCORING CASE STUDY

Submitted By: Anand Bhosale

INDEX

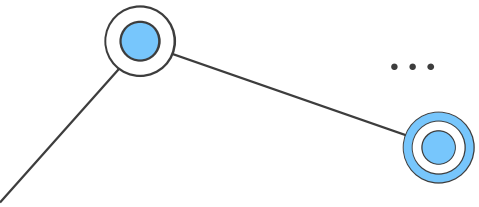
PROBLEM STATEMENT

- X Education sells online courses to industry professionals.

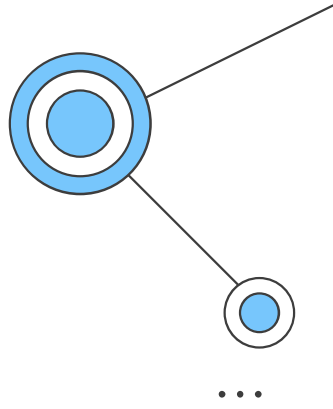
- X Education gets a lot of leads, but its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

- 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.

SOLUTION METHODOLOGY

Data cleaning and data manipulation

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains a large number of missing values and are not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

Exploratory Data Analysis (EDA)

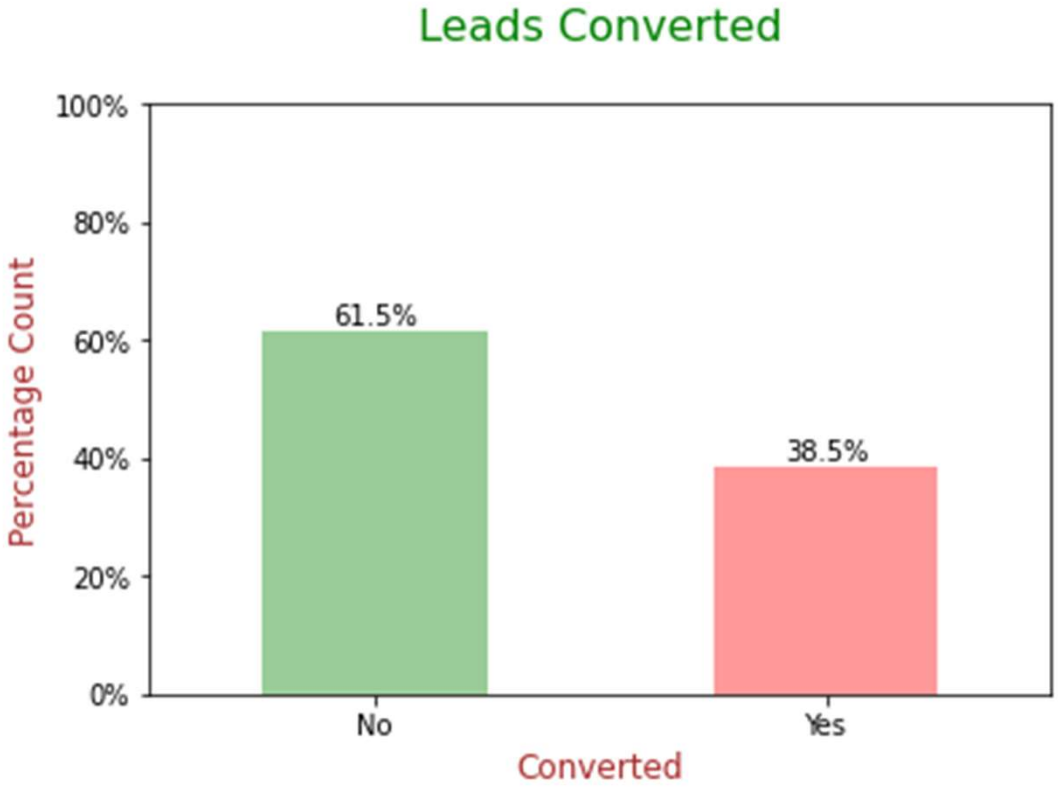
- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy variables and encoding of the data.
- Classification technique: logistic regression is used for model making and 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", "ReceiveMoreUpdates About Our Courses", "Update my supply"
- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "ProspectlD" 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 enough variance, which has dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", 'Newspaper, Article", "XEducation Forums", "Newspaper", 'DigitalAdvertisement" etc.
- Dropping the column shaving more than 35% as missing values such as 'How did you hear about X Education' and 'Lead Profile'.

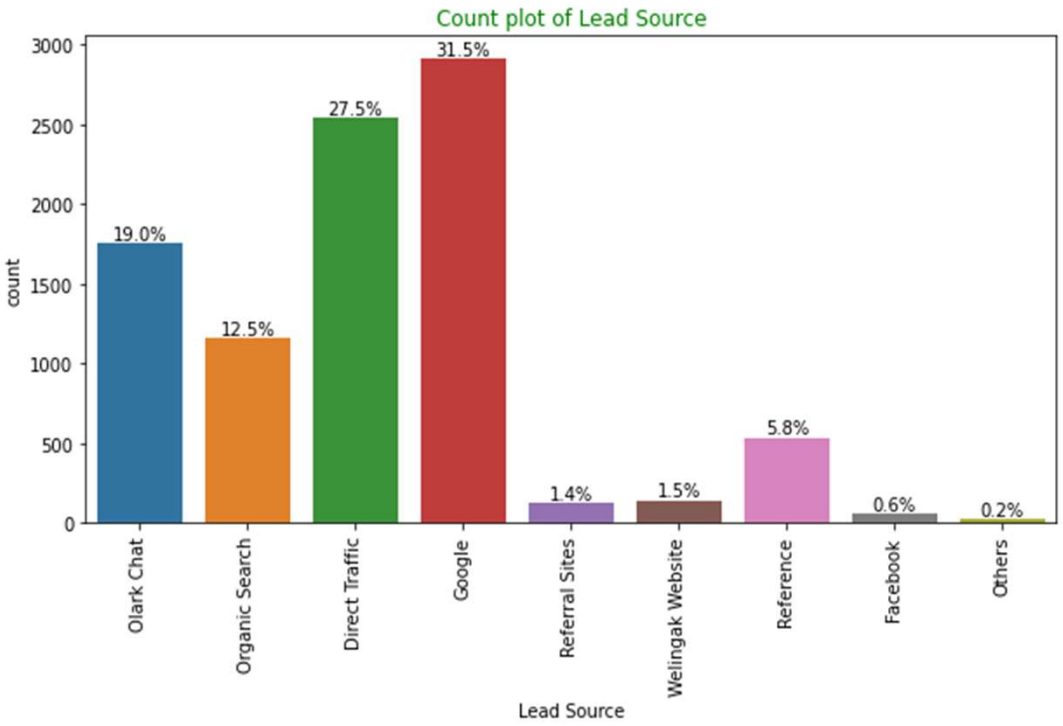
EXPLORATORY DATA ANALYSIS (EDA)

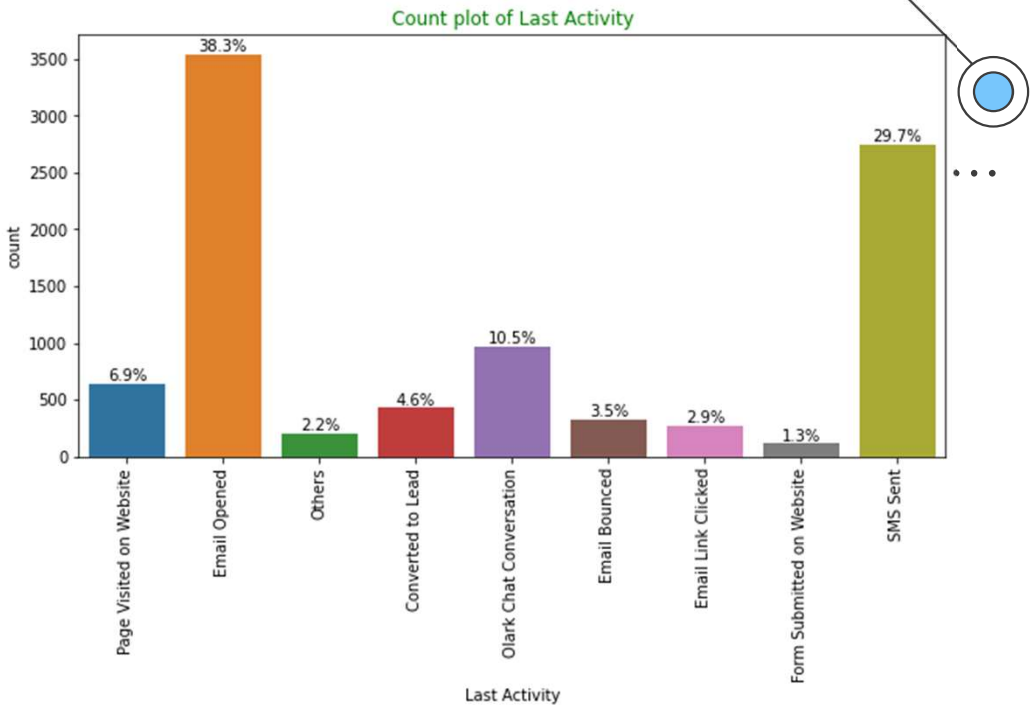
Data is imbalanced while analyzing the target variable.



- The conversion rate is 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

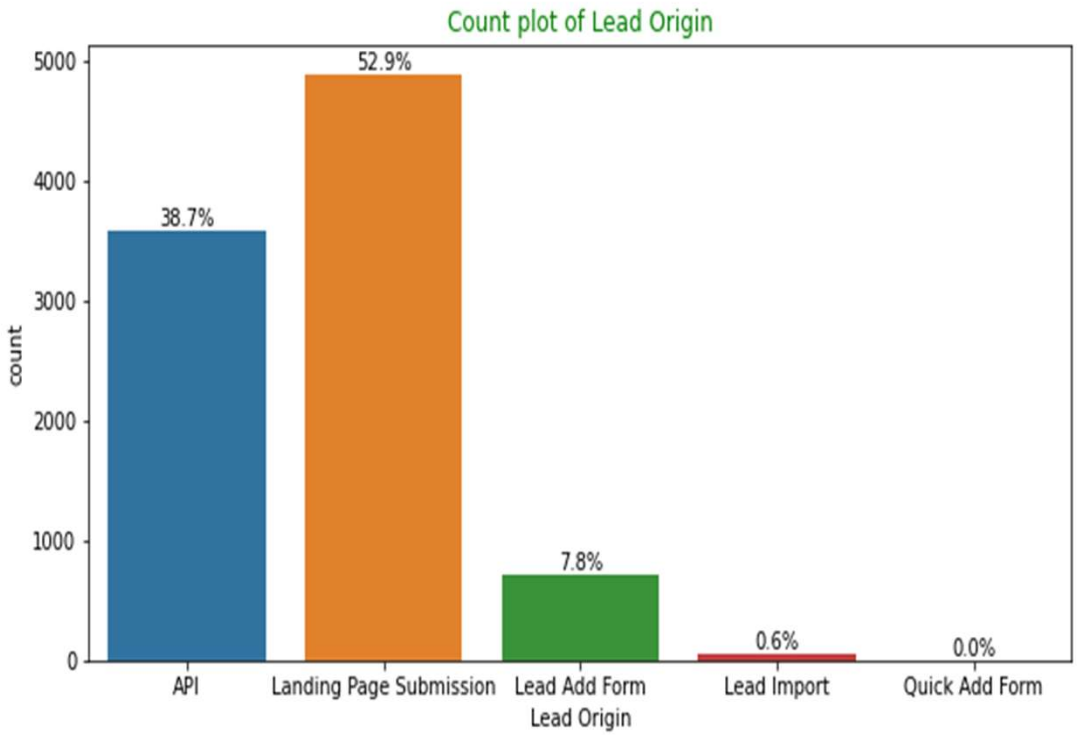


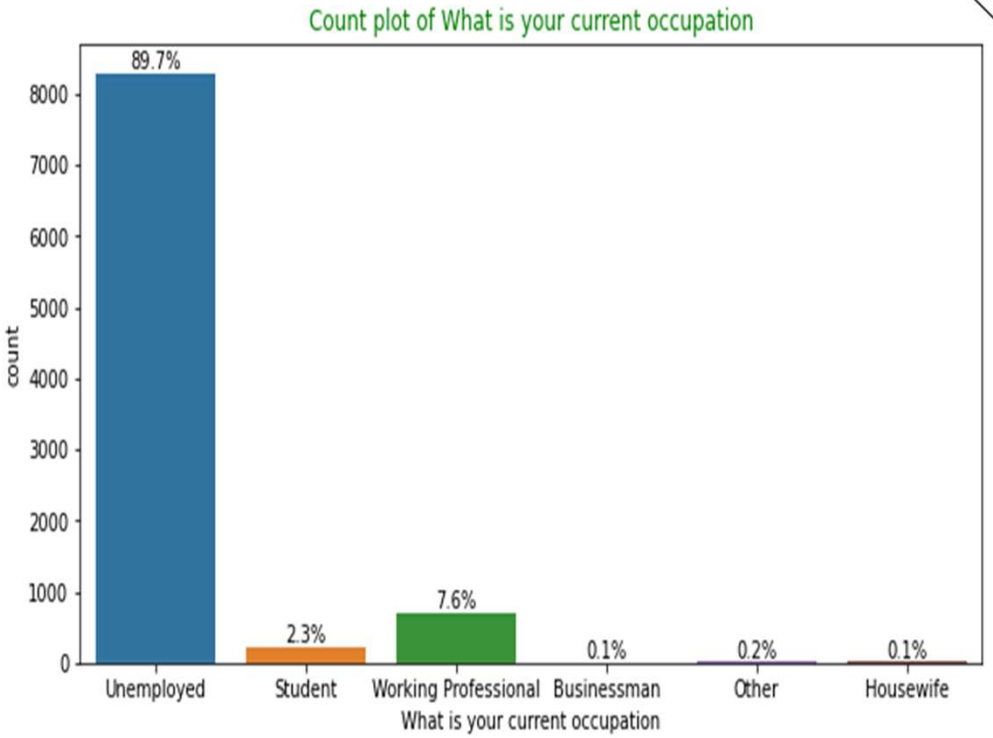


Lead Source: 58% Lead source is from Google & Direct Traffic combined.

Last Activity: 68% of customer's contribution in SMS Sent & Email Opened activities.

Univariate Analysis — Categorical Variables

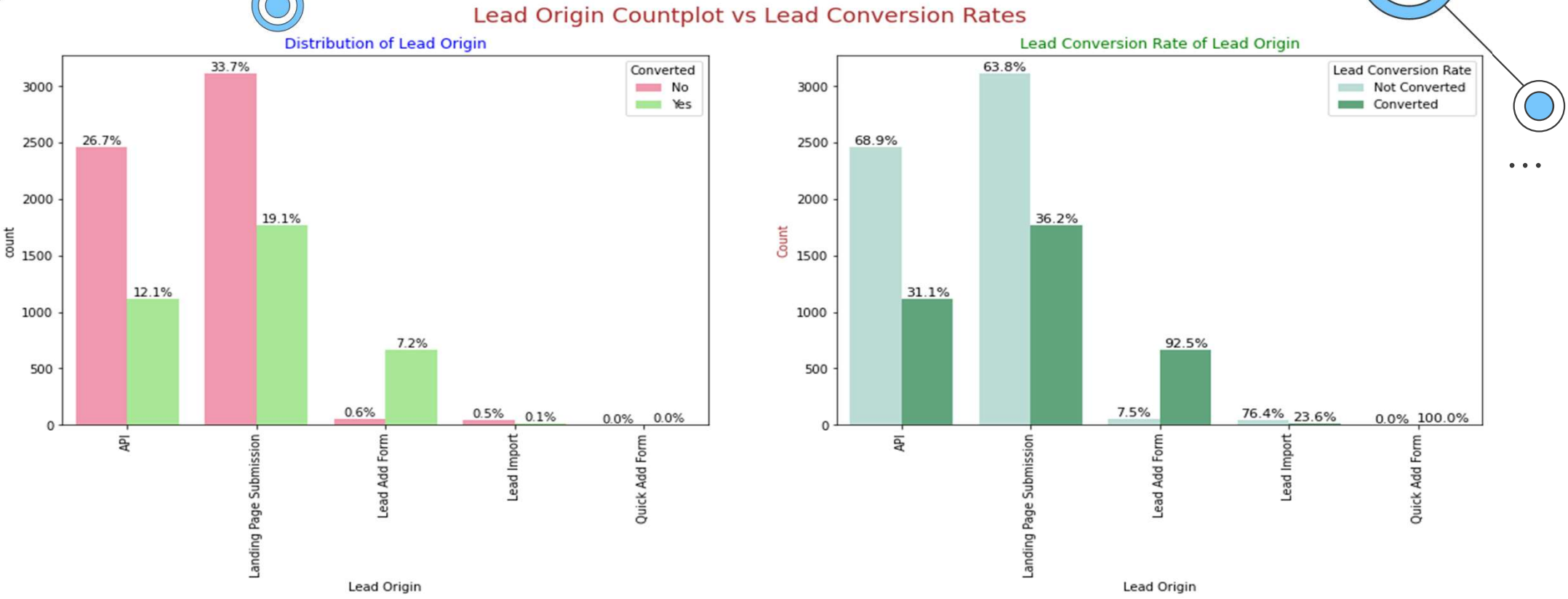




Lead Origin: "Landing Page Submission“ identified 53% of customers, "API“ identified 39%.

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

EDA - Bivariate Analysis for Categorical Variables

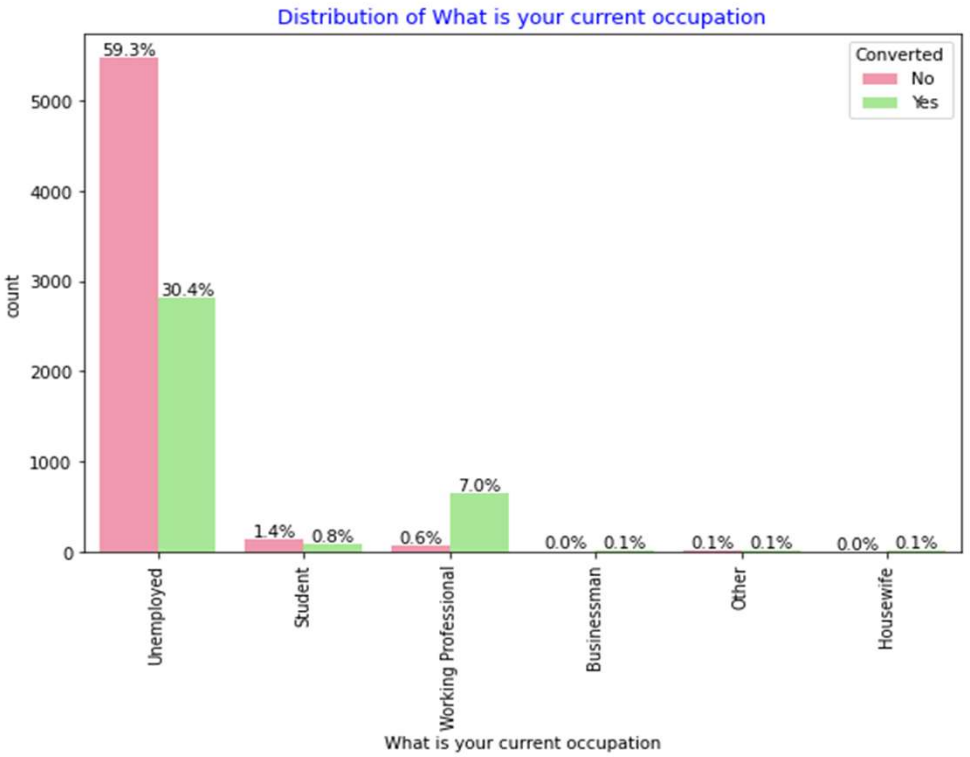


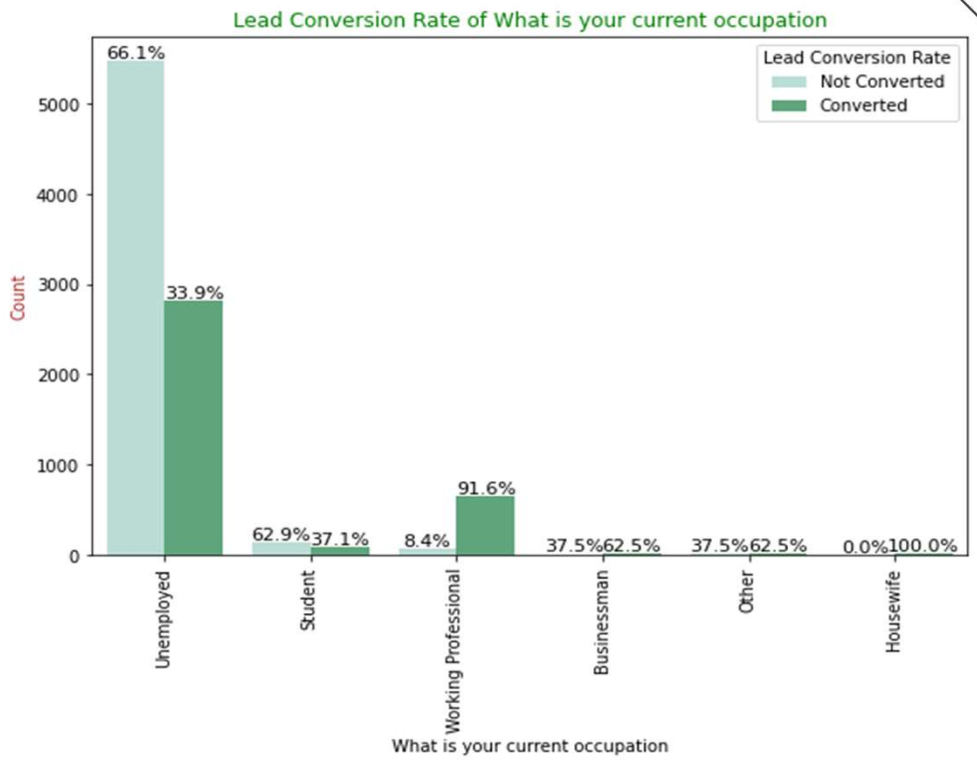
- Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.

- The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

EDA - Bivariate Analysis for Categorical Variables

What is your current occupation Countplot vs Lead Conversion Rates



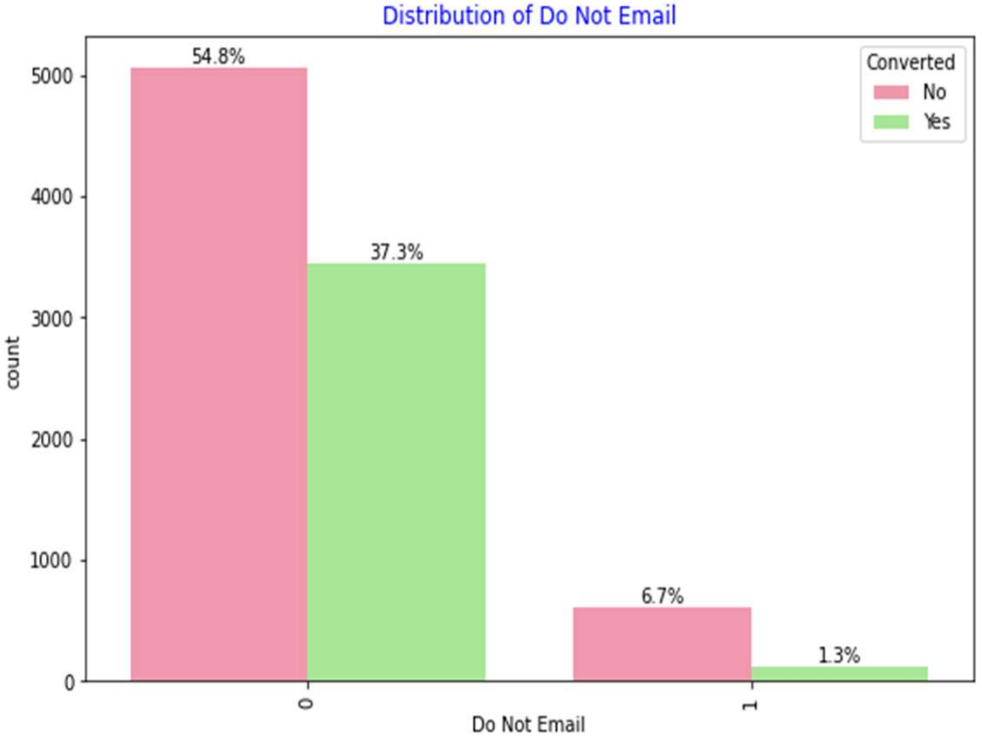


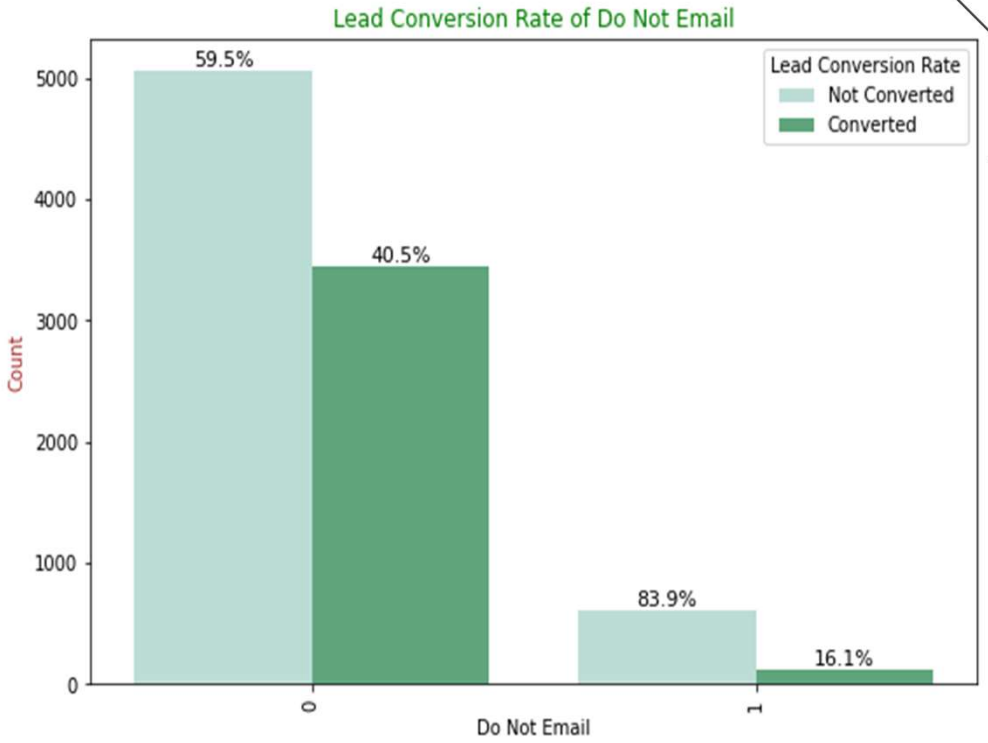
- • Around 90% of the customers are Unemployed, with a lead conversion rate (LCR) of 34%.

- • While Working Professionals contribute only 7.6% of total customers with almost 92% Lead conversion rate (LCR).

EDA - Bivariate Analysis for Categorical Variables

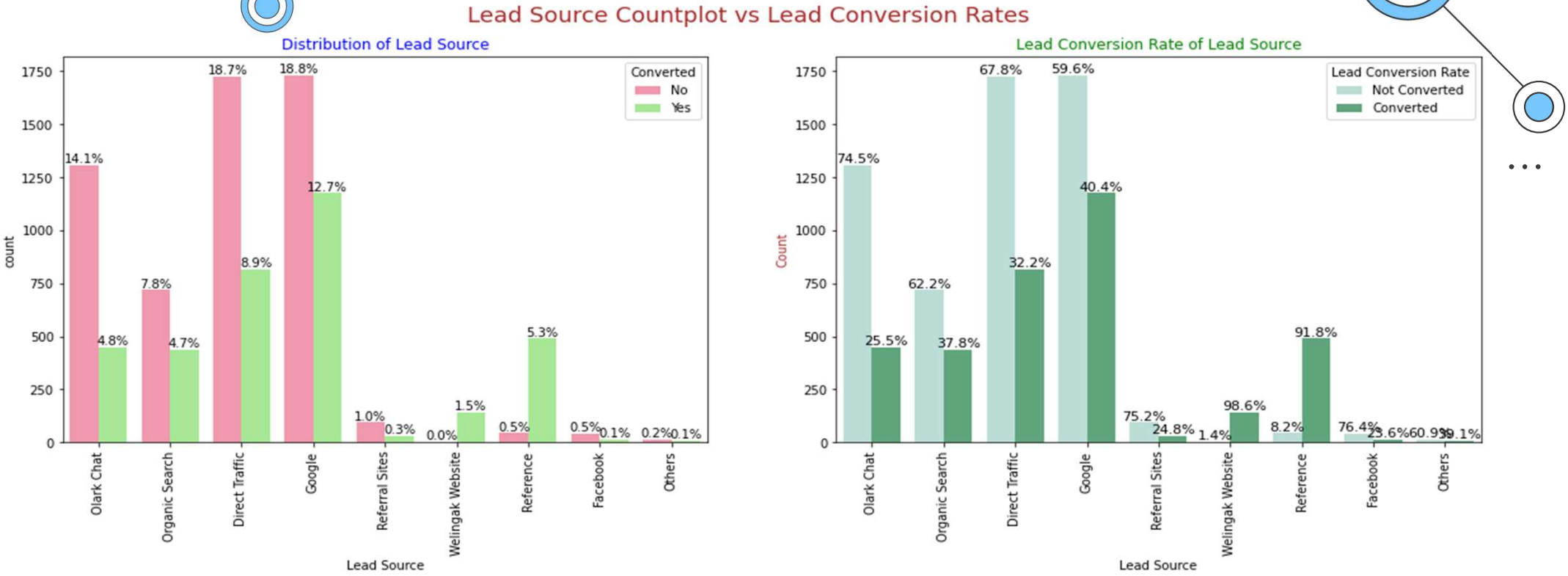
Do Not Email Countplot vs Lead Conversion Rates





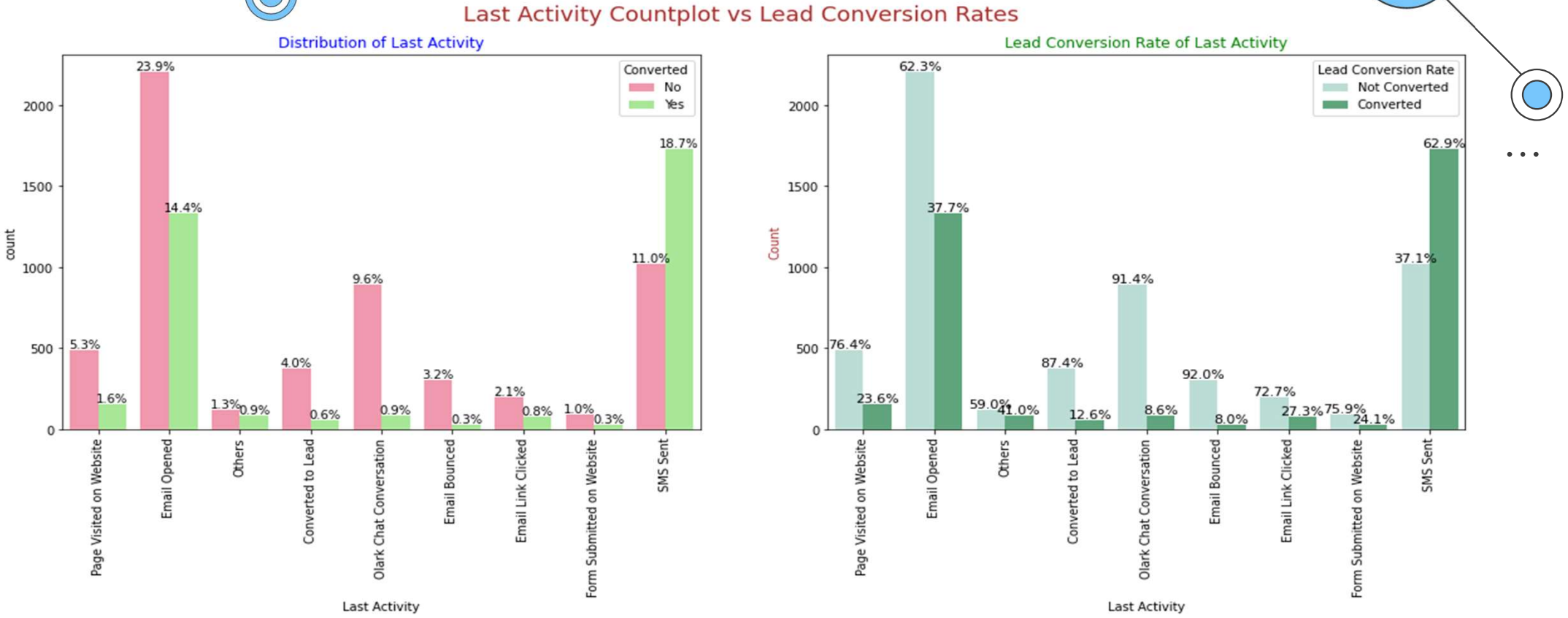
- • 92% of the people have opted that they don't want to be emailed about the course & 40% of them are converted to leads.

EDA - Bivariate Analysis for Categorical Variables



- Google has an LCR of 40% out of 31% of customers,
- Direct Traffic contributes 32% LCR with 27% customers, which is lower than Google,
- Organic Search also gives 37.8% of LCR, but the contribution is by only 12.5% of customers,
- Reference has an LCR of 91%, but there are only around 6% of customers through this Lead Source.

EDA - Bivariate Analysis for Categorical Variables



- 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities,
- 'Email Opened' activity contributed 38% of last activities performed by the customers, with 37% lead conversion rate.

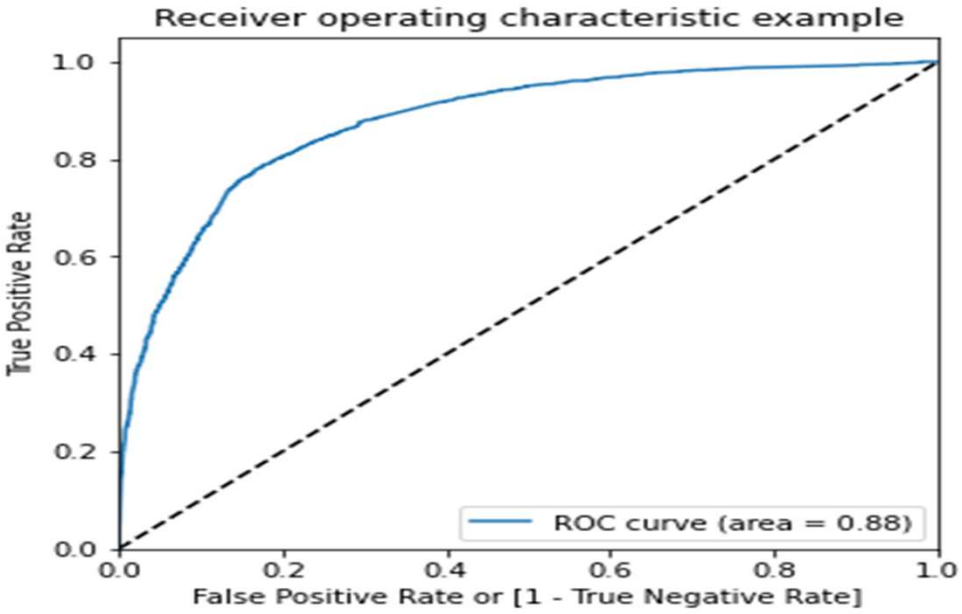
DATA CONVERSION

- Numerical Variables are normalized
- Dummy Variables are created for object-type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 37

MODEL BUILDING

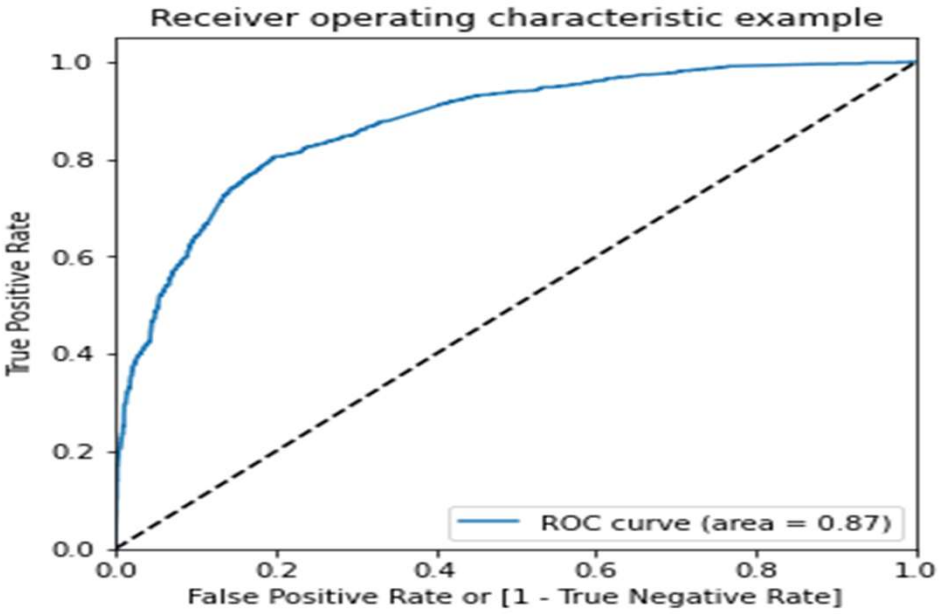
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen a 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5
- Predictions on a test data set
- Overall accuracy 81%

Model Evaluation



ROC Curve — Train Data Set

- The area under the ROC curve is 0.88 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.



ROC Curve — Test Data Set

- The area under the 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

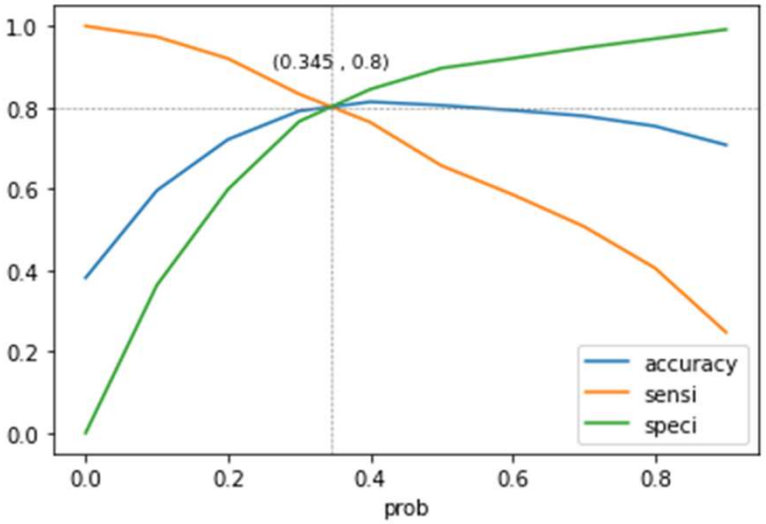
Train Data Set

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

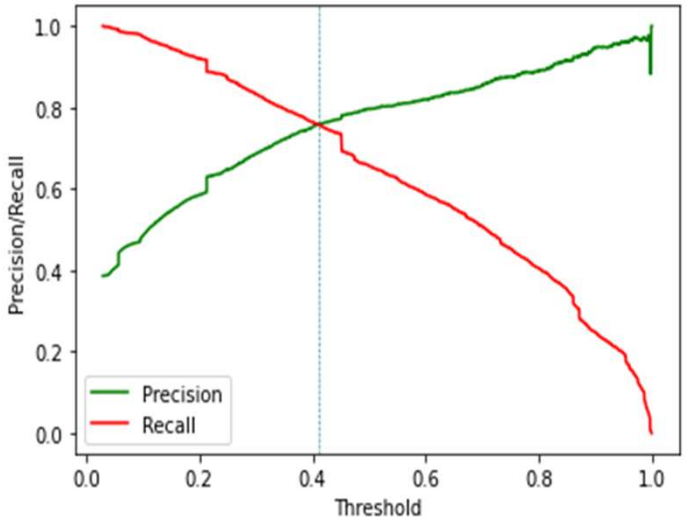
```
**********************************************************

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
[[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
```

****



****



PREDICTION ON TEST SET

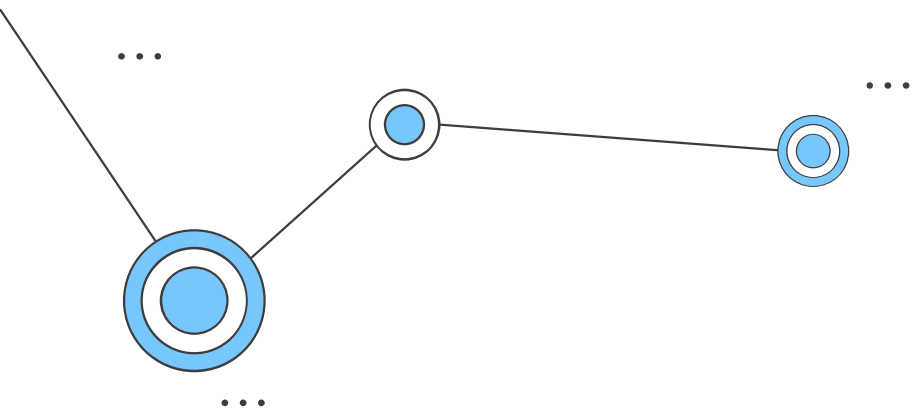
- Before predicting the test set, we need to standardize the test set and need to have the exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this, we did model evaluation i.e. finding the accuracy, precision, and recall.
- The accuracy score we found was 0.82, precision 0.75, and recall 0.75 approximately.
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and reca 11/ sensitivity.
- Lead score is created on test dataset to identify hot leads - high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted

CONCLUSION

- The model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set, using a cut-off value of 0.345.
- Sensitivity in this case indicates how many leads the model identifies correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.

RECOMMENDATIONS

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing references that convert to leads, encourage providing more references.
- Working professionals to be aggressively targeted as they have a high conversion rate and will have a better financial situation to pay higher fees too.



THANK YOU