Lead Scoring Case Study

Lavanya Ramakrishnan | Madhuri Kamble | Manvendra Ghatode

Contents

- Company Background
- Problem Statement and Objective
- Approach
- Data Cleaning
- Analysis
- Data Preparation
- Model Building
- Model Evaluation
- Recommendations

Company Background

X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Problem Statement

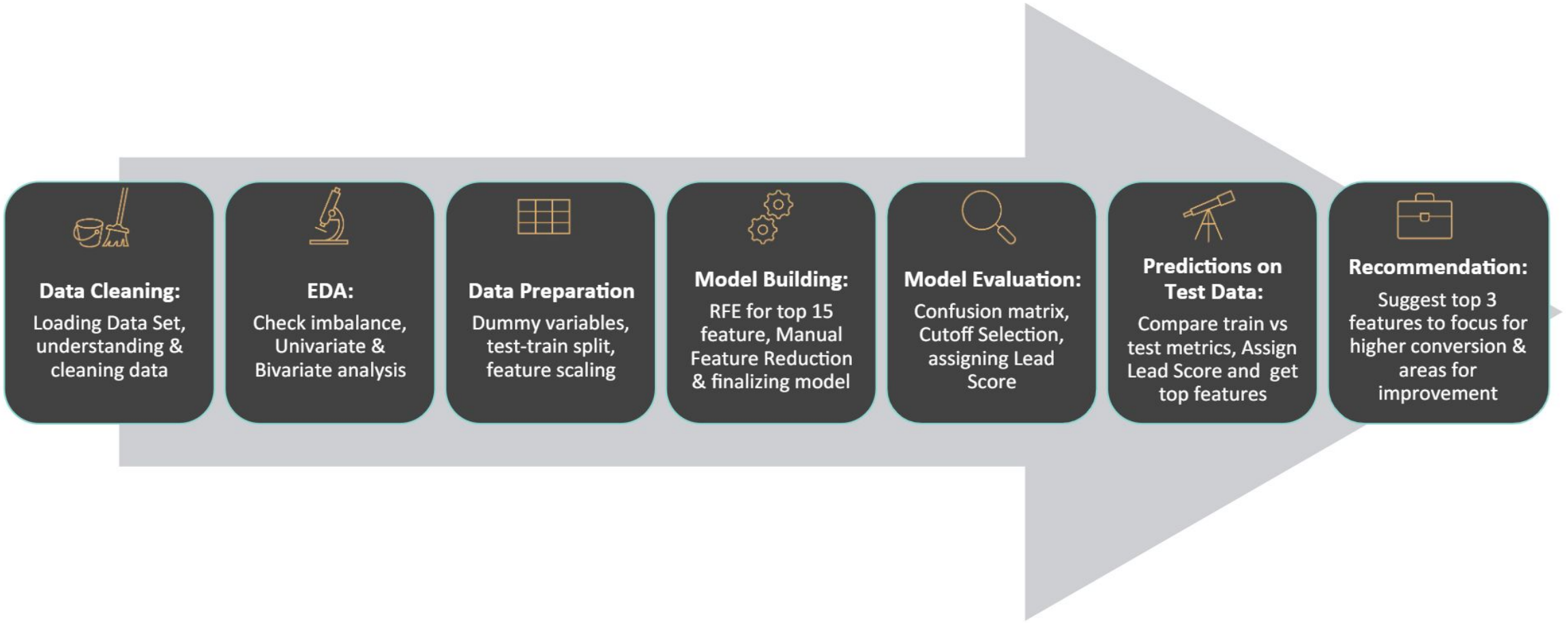
After leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%, which is very poor.

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. So the lead conversion rate should go up.

Objective

To help the company to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach


Data Cleaning:
Loading Data Set, understanding & cleaning data
EDA:
Check imbalance, Univariate & Bivariate analysis
Data Preparation
Dummy variables, test-train split, feature scaling
Model Building:
RFE for top 15 feature, Manual Feature Reduction & finalizing model
Model Evaluation:
Confusion matrix, Cutoff Selection, assigning Lead Score
Predictions on Test Data:
Compare train vs test metrics, Assign Lead Score and get top features
Recommendation:
Suggest top 3 features to focus for higher conversion & areas for improvement

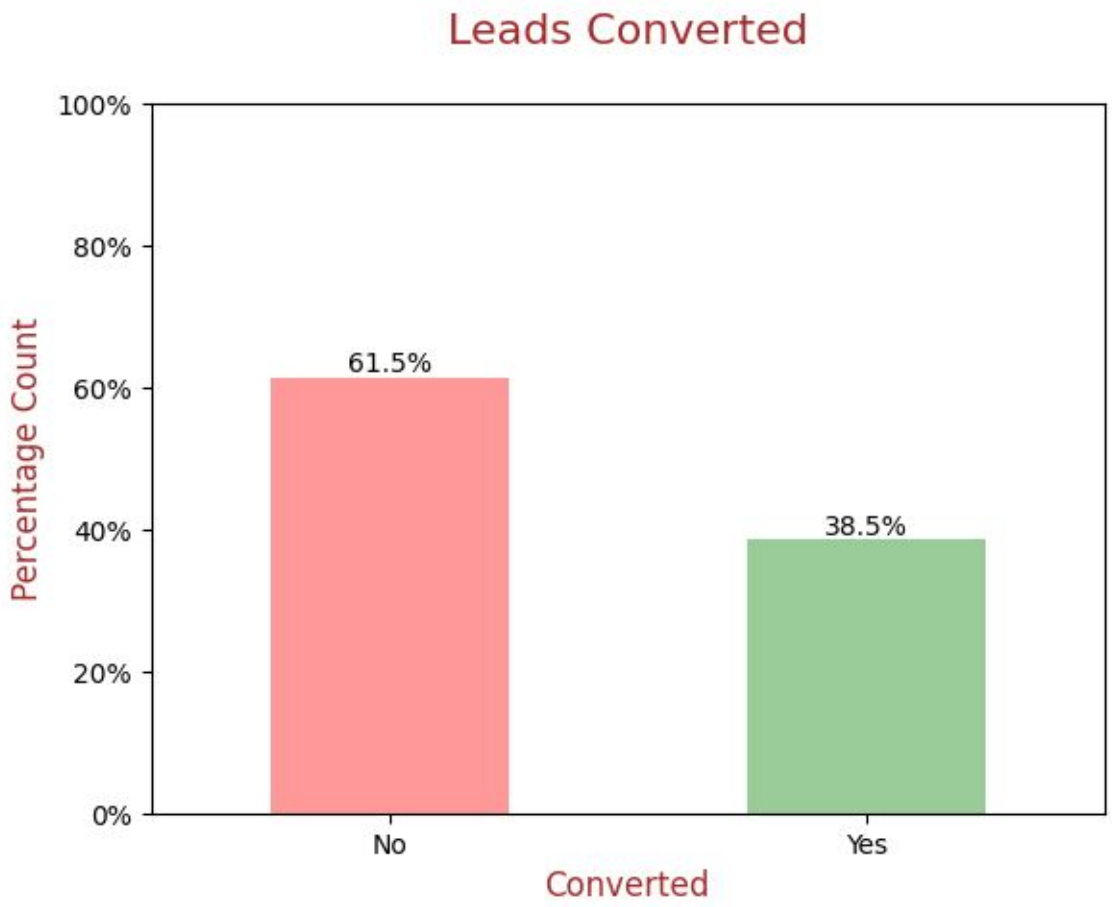
Data Cleaning

- 'Select' values which are same as null values are removed, as the customer left the field blank.
- Columns having null values greater than 40% were dropped.
- Missing values and Imputation was performed based using different techniques as per requirements.
- Some additional columns were dropped as they are do not add value to our study.
- Skewed category columns were also dropped to avoid bias in modelling.
- Outlier were removed using capping and flooring.
- Low frequency values were also grouped to limit dummy variable creation.

Analysis

➢ **Data Imbalance - Target variable (Converted)**

- Conversion rate for is only 38.5% which is not good for business. We have to improve this rate.
- The imbalance ratio is 1.59.



Analysis

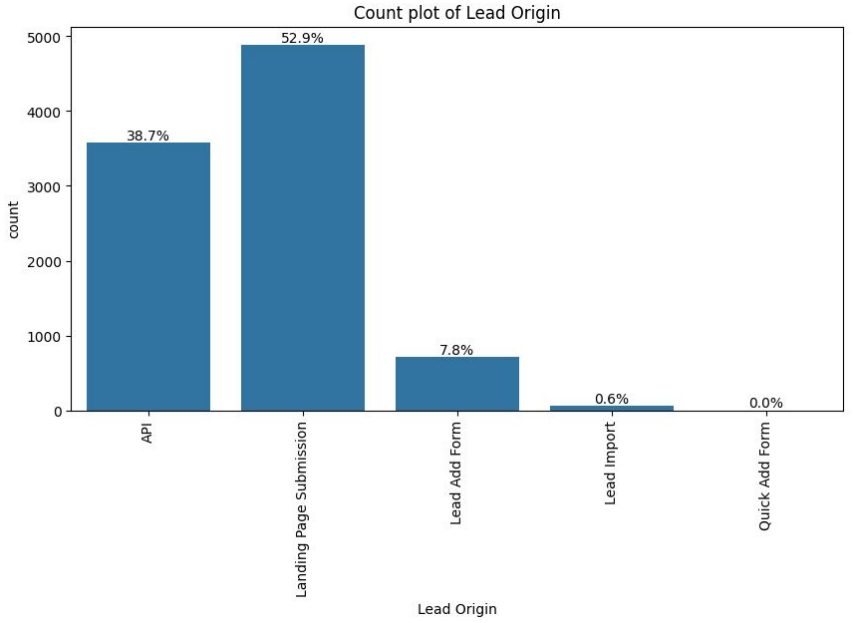
➢ **Univariate Analysis**

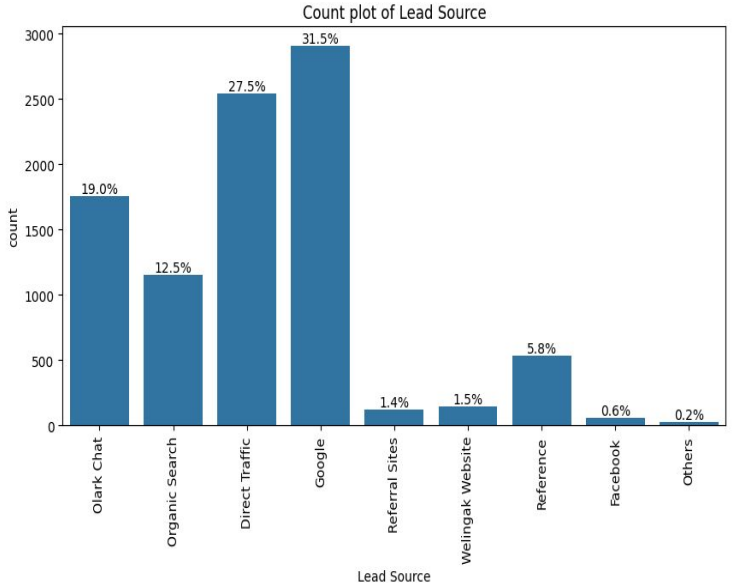
Lead Origin:
"Landing Page Submission" identified 53% customers following "API" with 39%.

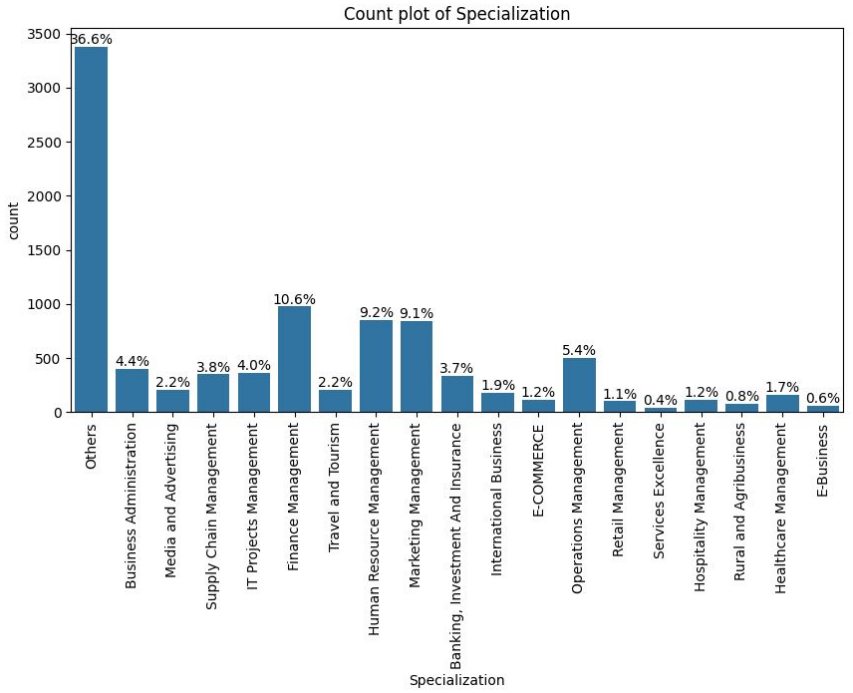
Current_occupation:
90% of the leads are listed as 'Unemployed'.

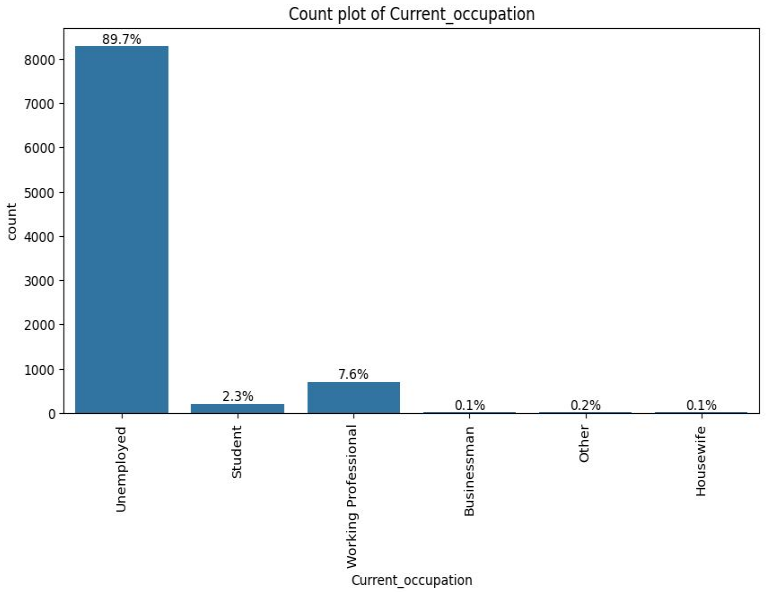
Lead Source:
58% leads are from Google & Direct Traffic combined.

Specialization:
Leads having Finance, HR and marketing management as specialization are the top contributing categories. (We are ignoring "Others" as we imputed null values with it)



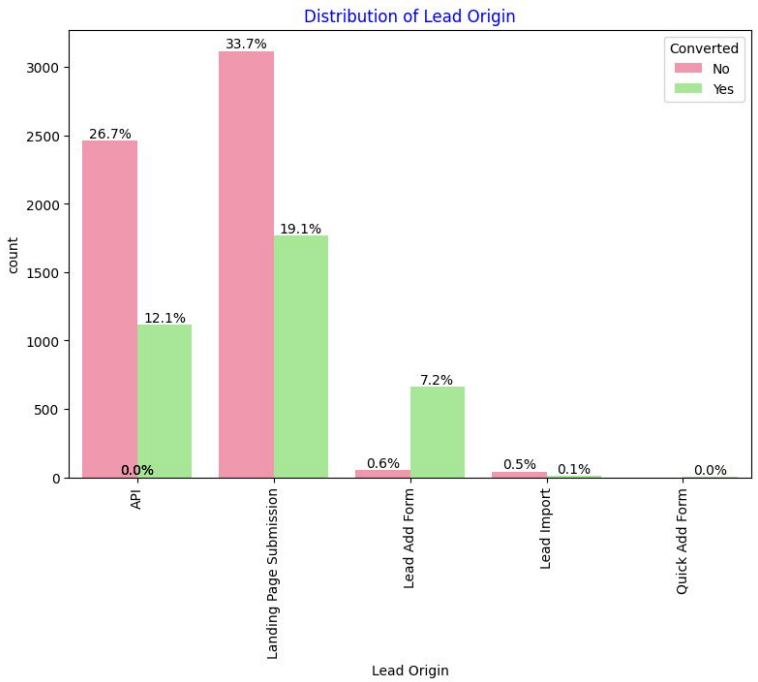


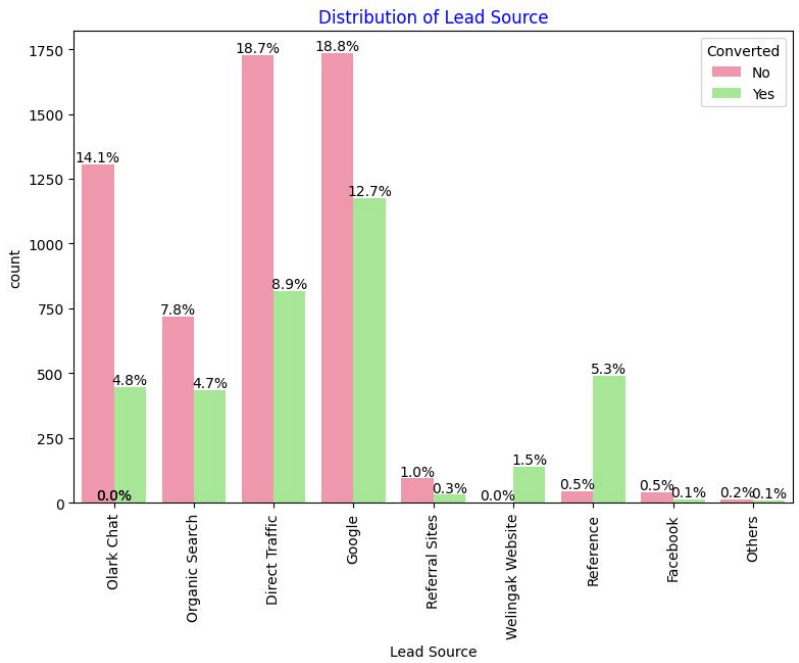


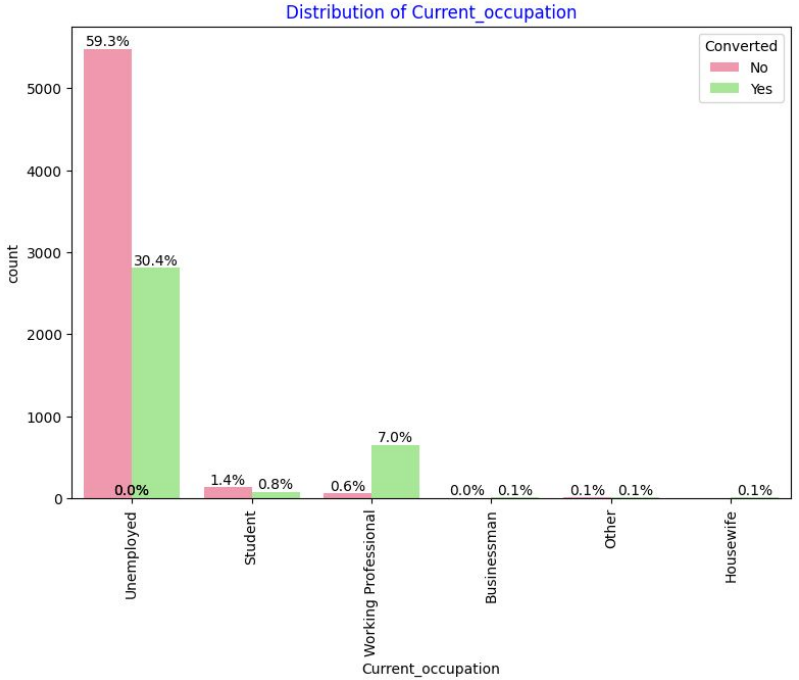


Analysis

➢ **Bivariate Analysis - Categorical**







Lead Origin:

Around 30% of leads converted successfully originate from "Landing page Submission" and "API".

Lead Source:

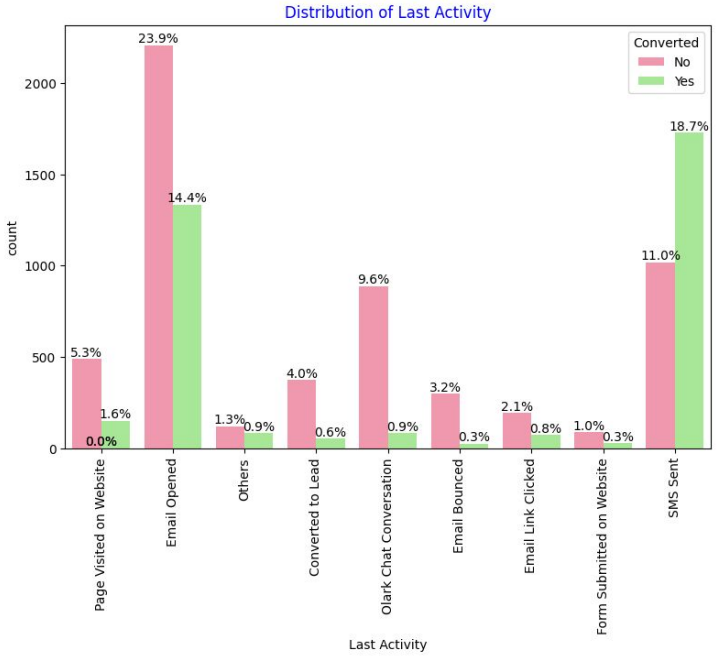
Google and Direct Traffic contribute to 21% leads converted. Notably, Welingak Website and Referance contribute less traffic but have very high conversion rate of above 90%.

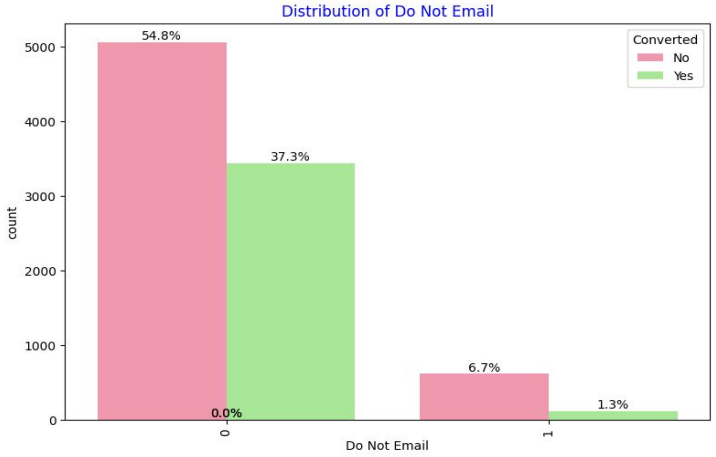
Current_occupation:

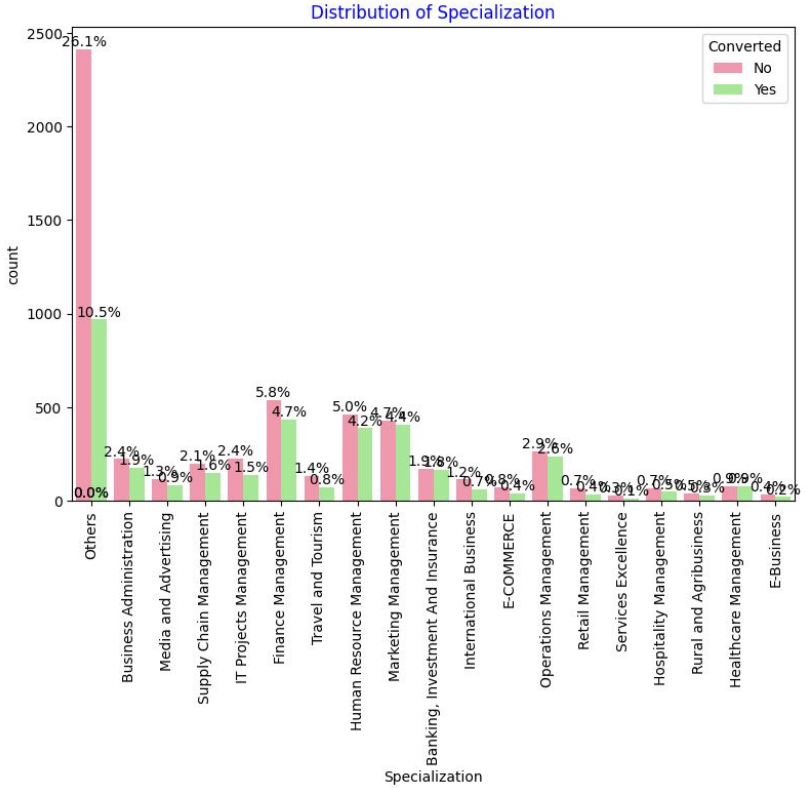
Approximately 90% of the customers are Unemployed with 30% overall conversion. While Working Professional contribute only 7.6% of total customers with almost 7% overall conversion.

Analysis

➢ **Bivariate Analysis - Categorical**







Last Activity:

'SMS Sent' has high lead conversion rate with 18.7% leads converted, 'Email Opened' contribute around 40% of leads but only around 35% are converted from them i.e, 14.4%.

Do Not Email:

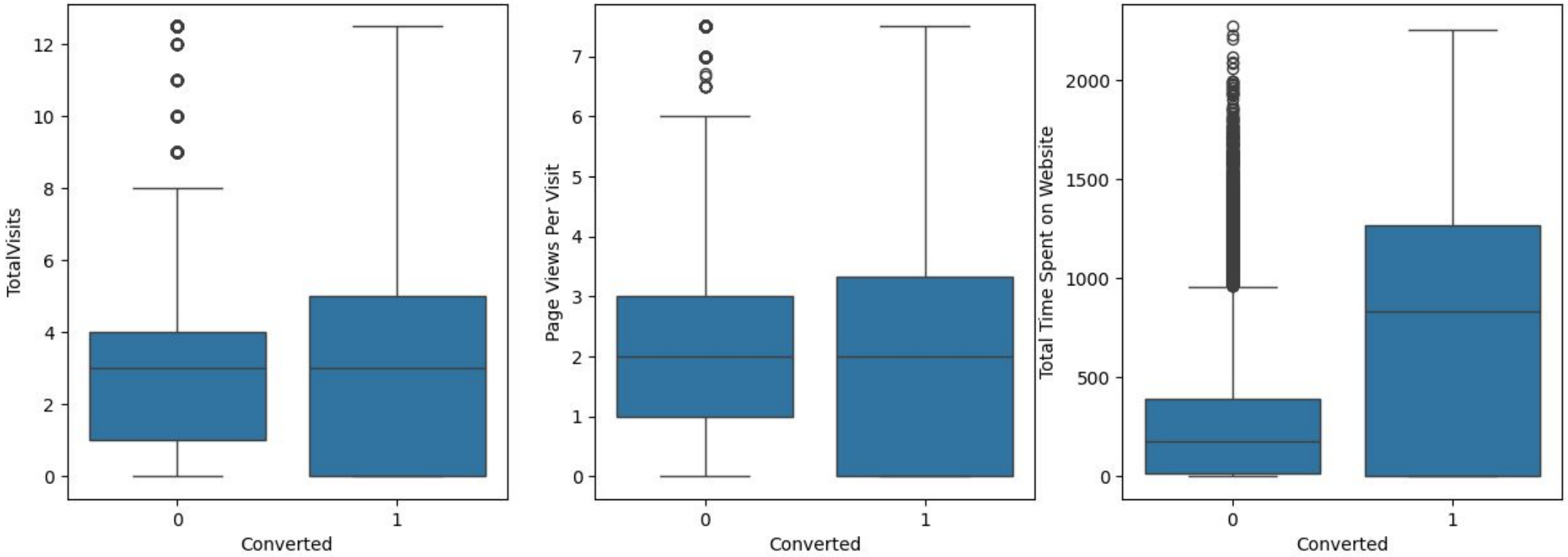
92% of the people has opted that they dont want to be emailed about the course.

Specialization:

Conversion rate is evenly soread across categories but Marketing Managemt,HR Management and Finance Management has the highest contrbution

Analysis

➢ **Bivariate Analysis - Numerical**



- Leads which spent more time on website on an average are more likely to be converted.

Data Preparation

- Binary level categorical columns weremapped to 1 / 0.
- Created dummy features (one-hot encoded) for categorical variables - Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation.
- Splitting Train & Test Sets in 70:30 ratio.
- Feature scaling was done using MinMaxScalar method.
- Checking the correlations
  - Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

- Feature Selection using Recursive Feature Elimination (RFE) to select only the important columns.

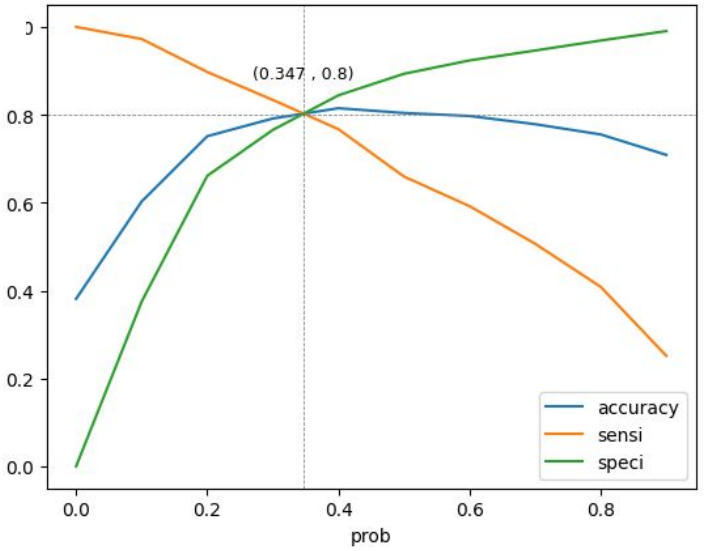
- Further, manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05 amd VIF < 5.

- Model 4 looks stable with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5.

Model Evaluation

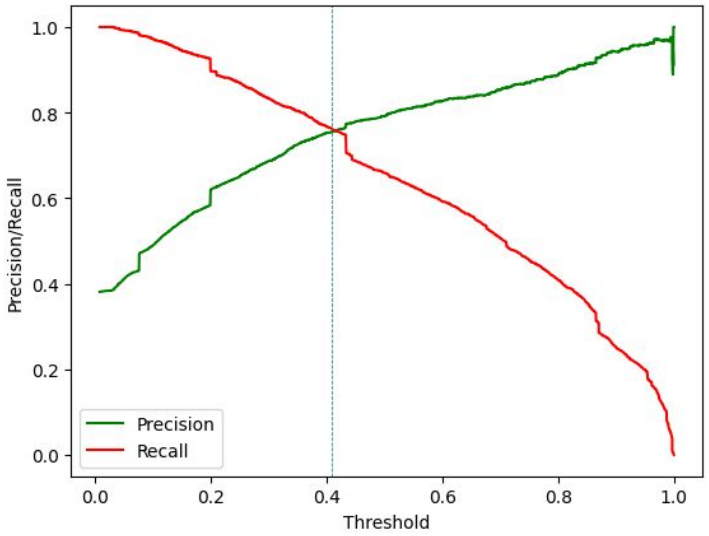
Confusion Matrix and Evaluation Metrics - 0.347

```
Confusion Matrix
[[3240  762]
 [ 476 1990]]

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

True Negative                  :  3240
True Positive                  :  1990
False Negative                 :  476
False Positve                  :  762
Model Accuracy                 :  0.8086
Model Sensitivity              :  0.807
Model Specificity              :  0.8096
Model Precision                :  0.7231
Model Recall                   :  0.807
Model True Positive Rate (TPR) :  0.807
Model False Positive Rate (FPR) :  0.1904


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



Confusion Matrix and Evaluation Metrics - 0.41

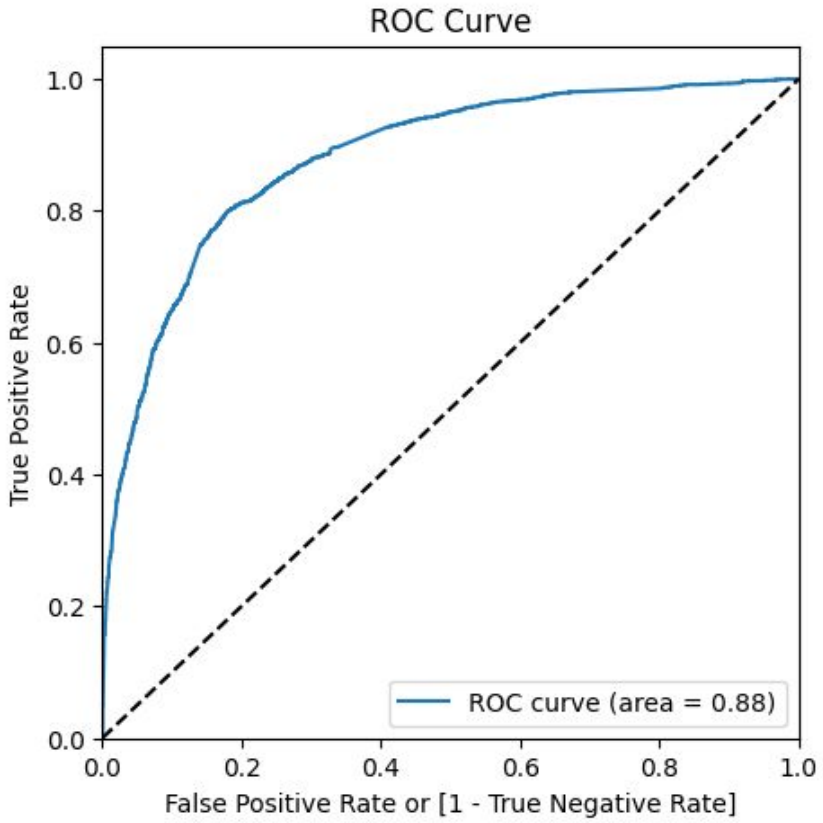
```
Confusion Matrix
[[3393  609]
 [ 590 1876]]

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

True Negative                  :  3393
True Positive                  :  1876
False Negative                 :  590
False Positve                  :  609
Model Accuracy                 :  0.8146
Model Sensitivity              :  0.7607
Model Specificity              :  0.8478
Model Precision                :  0.7549
Model Recall                   :  0.7607
Model True Positive Rate (TPR) :  0.7607
Model False Positive Rate (FPR) :  0.1522


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



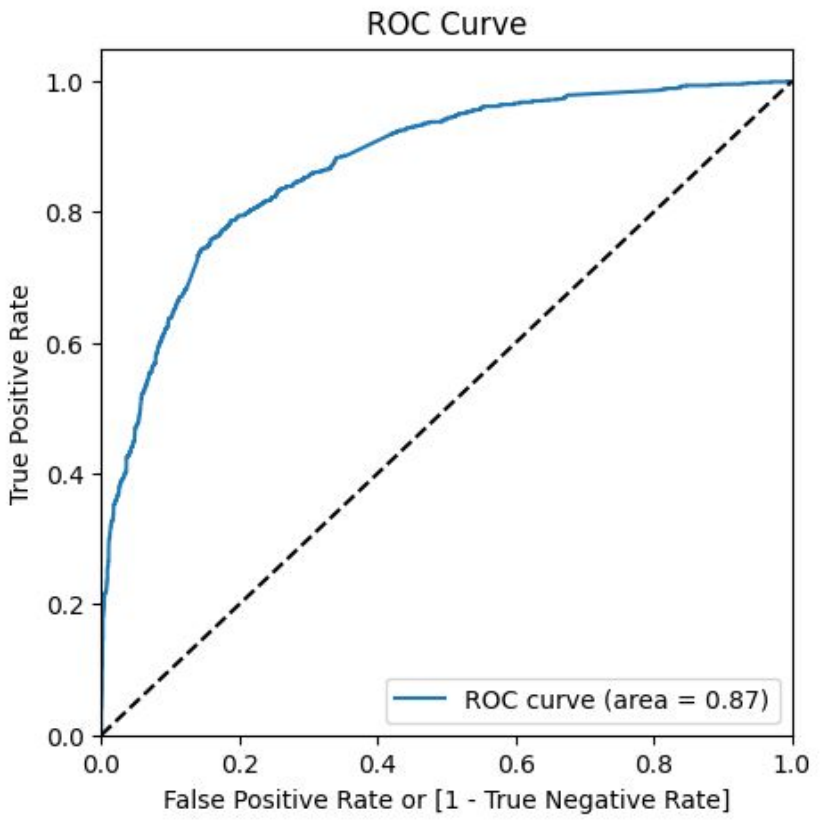
- 0.347 is considered as cutoff after comparing evaluation metrics for both values.

Model Evaluation

ROC Curve - Train



ROC Curve - Test



- Area under ROC Curve in both cases is similar and indicates good predictive model.

Model Evaluation

Confusion Matrix and Evaluation Metrics -Train

```
Confusion Matrix
[[3240  762]
 [ 476 1990]]

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

True Negative                  :  3240
True Positive                  :  1990
False Negative                 :  476
False Positve                  :  762
Model Accuracy                 :  0.8086
Model Sensitivity              :  0.807
Model Specificity              :  0.8096
Model Precision                :  0.7231
Model Recall                   :  0.807
Model True Positive Rate (TPR) :  0.807
Model False Positive Rate (FPR)  :  0.1904


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

Confusion Matrix and Evaluation Metrics - Test

```
Confusion Matrix
[[1340  337]
 [ 225  870]]

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

True Negative                  :  1340
True Positive                  :  870
False Negative                 :  225
False Positve                  :  337
Model Accuracy                 :  0.7973
Model Sensitivity              :  0.7945
Model Specificity              :  0.799
Model Precision                :  0.7208
Model Recall                   :  0.7945
Model True Positive Rate (TPR) :  0.7945
Model False Positive Rate (FPR)  :  0.201


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

- The CEO of X Education had expected the sensitivity to be around 80%, which is achieved by our model.
- The model also has high accuracy of around 80%.

Recommendations

To increase Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- 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 reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement:

- Analyze negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.

```
Lead Source_Welingak Website                  5.656737
Total Time Spent on Website                   4.338266
Lead Source_Reference                         3.182821
Current_occupation_Working Professional       2.712793
Last Activity_SMS Sent                        2.226035
Last Activity_Others                          1.401533
Last Activity_Email Opened                    1.103569
Lead Source_Olark Chat                        1.056458
TotalVisits                                   0.819276
Specialization_Hospitality Management        -1.085516
Specialization_Others                        -1.164978
Lead Origin_Landing Page Submission          -1.266641
const                                        -2.385554
dtype: float64
```

Based on our model, the following features with high positive coefficients should be prioritized and those having negative coefficient require necessary improvements.