LEAD SCORING CASE STUDY

ANKUR NAPA
AMANDEEP KAUR

upGrad

International Institute of Informat Technology Bangal

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 rate should go up to 80% 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 most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

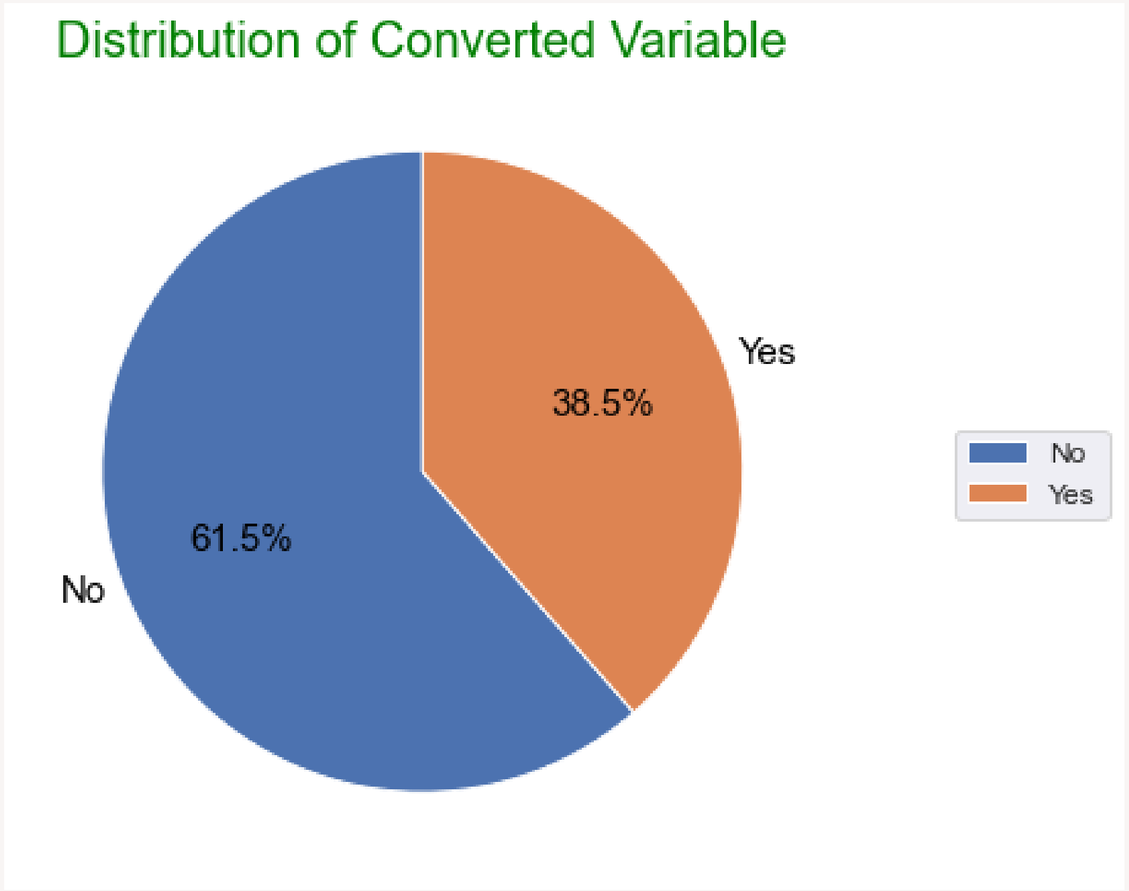
Analysis Approach

OVERVIEW PROCESS

Data cleaning and data manipulation.	EDA	Feature Scaling & Dummy Variables	Logistic regression	Validation of the model.	Model Presentation	Conclusions and recommendations
Duplicates, dropping & imputation	Univariate and Bivariate	Data encoding	Classification and modeling	Quantifying the ability of model	Constructing model	Analysing features

Data Cleaning

- Some columns were having label as ‘Select’ which means the customer has chosen not to answer this question. Hence, we changed those labels from ‘Select’ to null values.
- Removed columns having more than 45% null values.
- Imputed values mode () for Numerical columns having missing values.
- Analyzed missing values in the Categorical features and decided to drop columns for which data was skewed e.g. City, Country, Search, Do not call etc..Dropped “Tags” because it had values like incorrect number, switched off etc. which it seemed to be created by the sales team based on the current status of the lead.
- Removed columns irrelevant for our model building e.g. Prospect_ID, Lead_Number and Last Notable_Activity (updated by sales team).

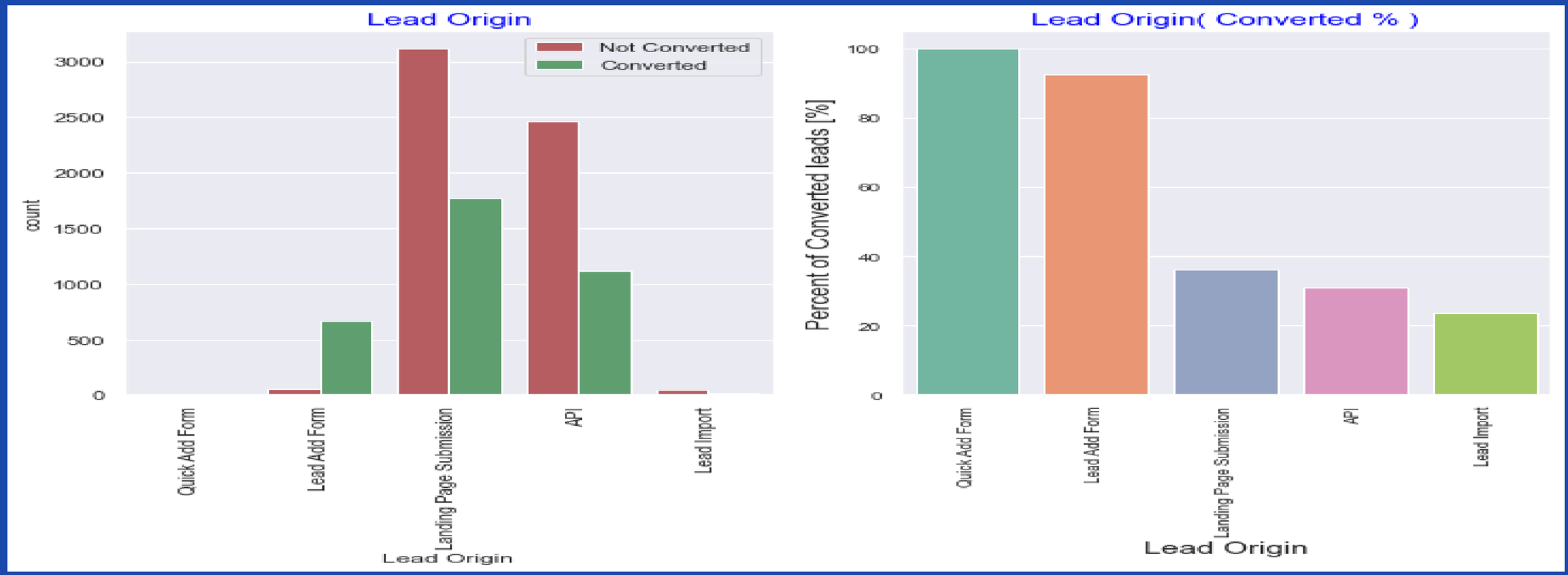


Data Imbalance

38.5% of the customers have converted to leads whereas 61.5% did not convert to a lead. It is not a well-balanced dataset

.

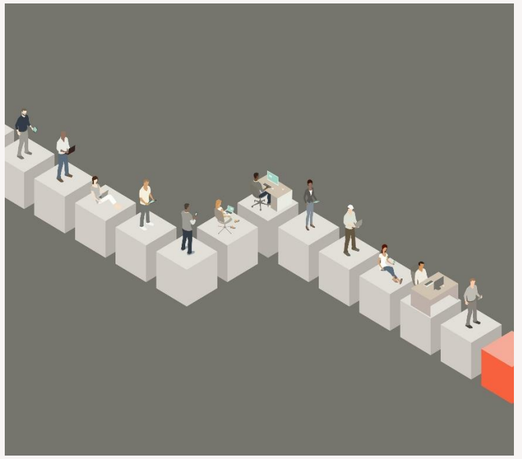
Outlier Analysis



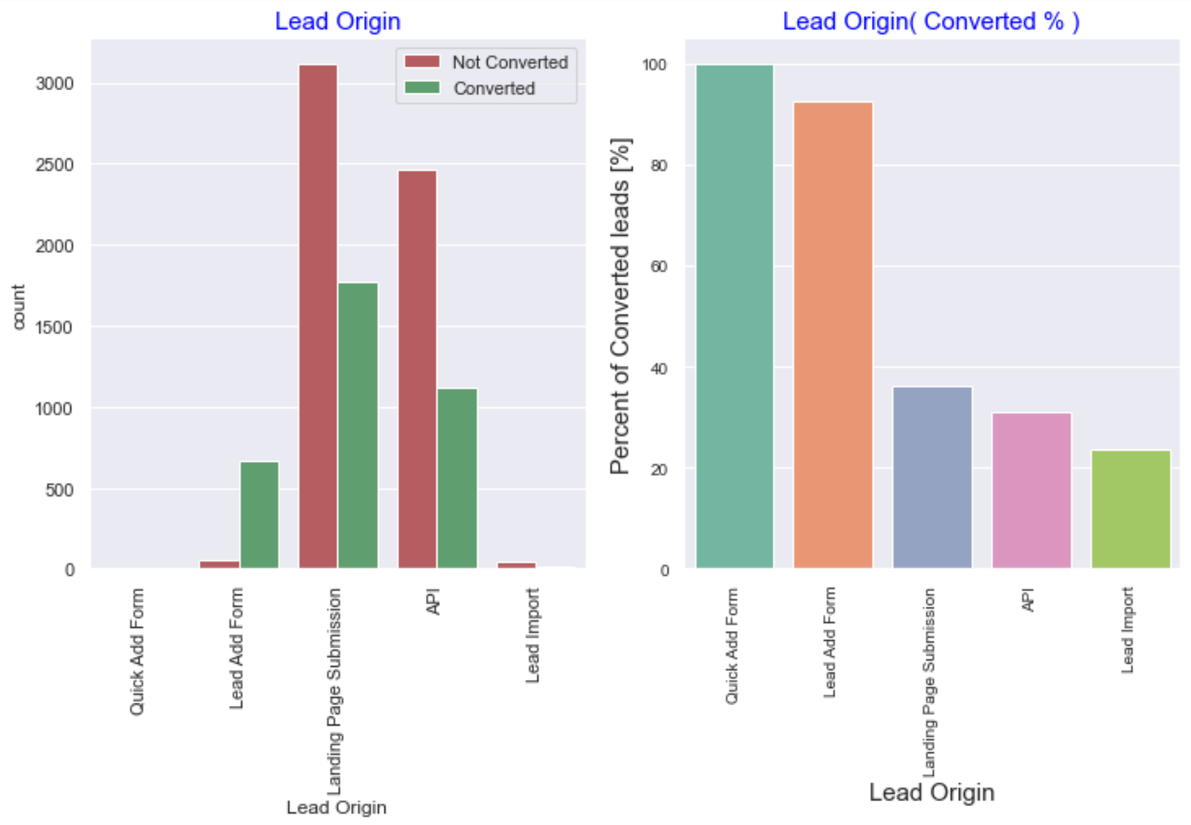
Insights:-

- Total Visits: It has some outliers which needs to be treated.
- Total Time Spent on Website: People whose spend more time has higher chance of getting converted.
- Page Views Per Visit: It has some outliers which needs to be treated.

Exploratory Data Analysis



Univariate Analysis - Categorical



Lead Origin

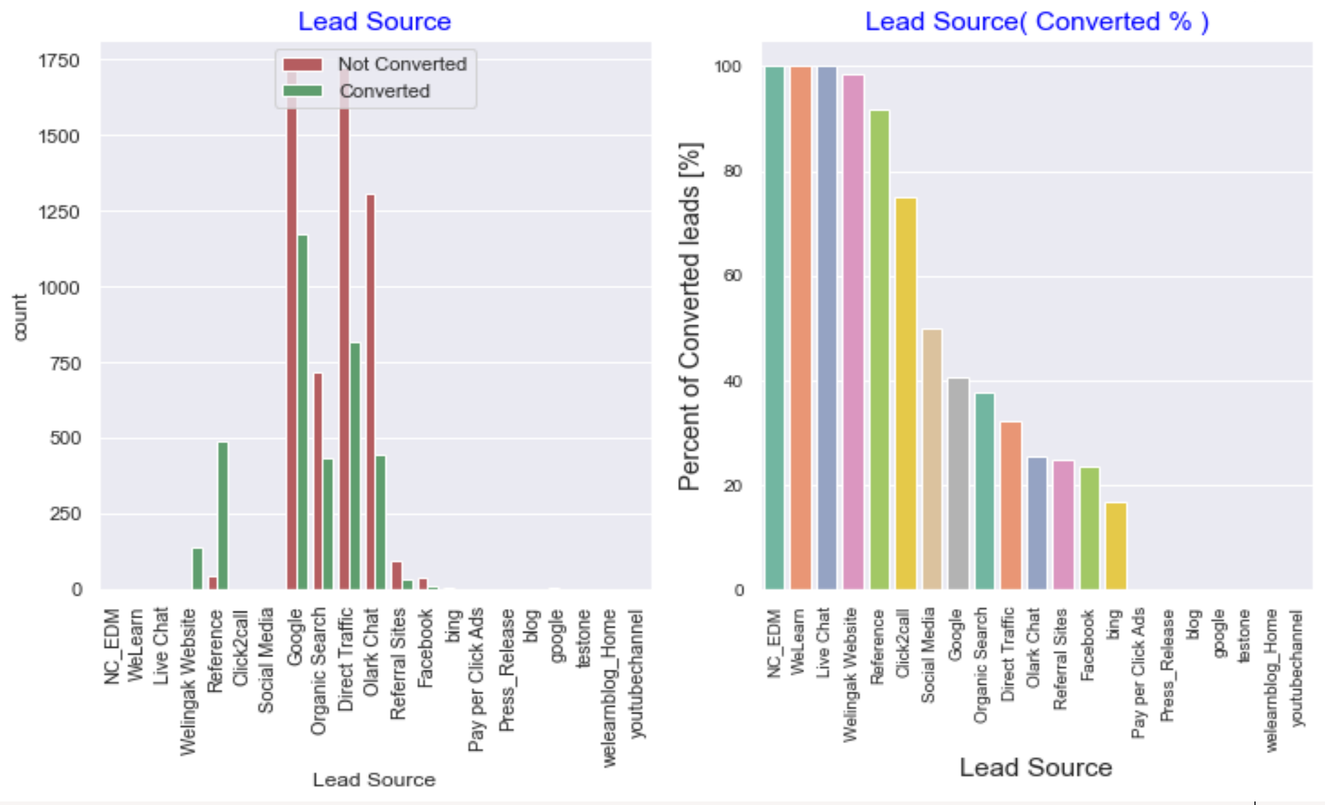
Insights:

- Most Leads are from "Landing Page submissions" out of which around 38% got converted, followed by "API", where around 32% are converted.
- Leads from the "Lead Add Form" have third highest conversions with conversion rate around 90%.
- "Lead Import" has only 55 records with the lowest conversion rate if around 22%
- "Quick Add Form" are 100% Converted with just 1 lead from this category.

To improve overall lead conversion rate, we should work on

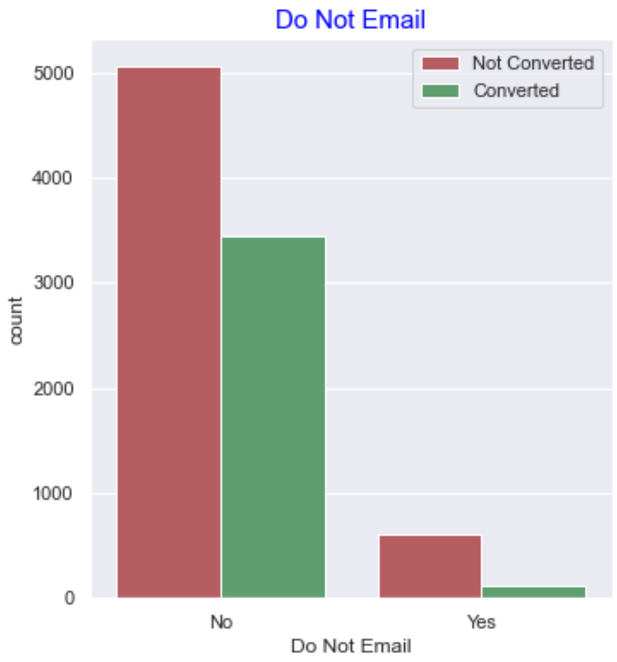
- Improving lead conversion of "API" and "Landing Page Submission origin"
- Generating more leads from "Lead Add Form" and "Quick Add Form"

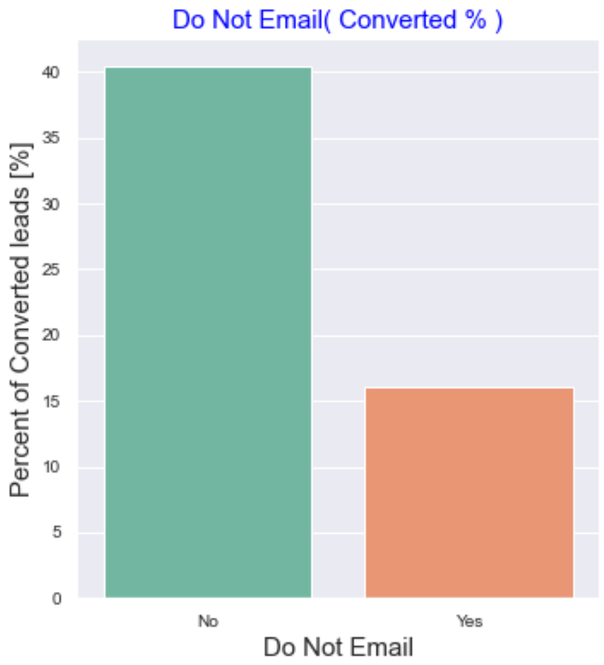
Lead Source



Insights:-

- The source of most of the leads was "Google" with conversion rate of 40%.
- "Direct Traffic","Organic search" and "Olark chat" comes next with around 32%, 38% and 28% conversion rate respectively.
- A lead that came from a "reference" has over 90% conversion even though number of such cases are just 534.This option should be explored more to increase lead conversion
- Welingak Website has around 98% lead conversion rate. This option should be explored more to increase lead conversion
- To increase lead count, we should encourage and incentivize existing members to bring more of their referrals as it has 90% conversion rate.

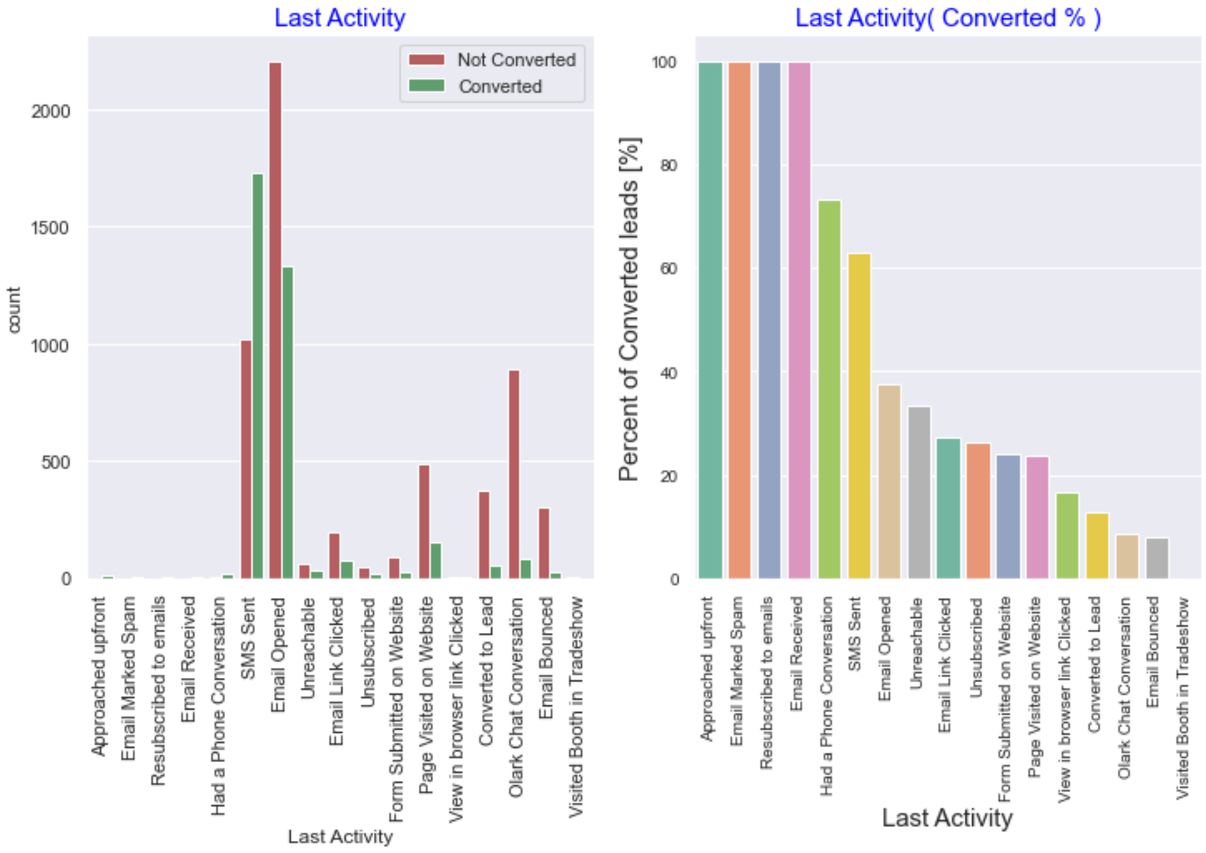




Do Not Email

Insights:-

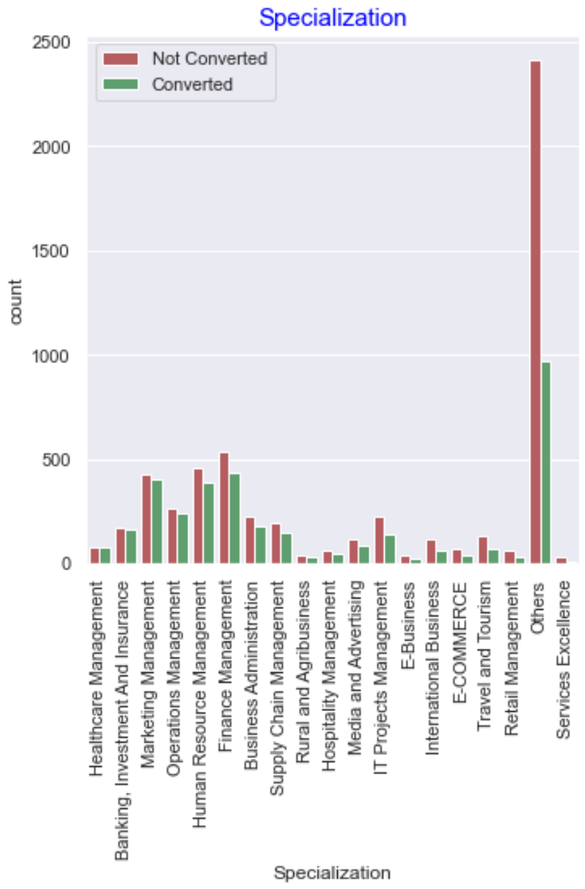
- Majority of the people(approx. 92%) are fine with receiving email.
- People who are ok with email has conversion rate of 40%
- People who have opted out of receive email has lower number of records and also have rate of conversion (only 15%)

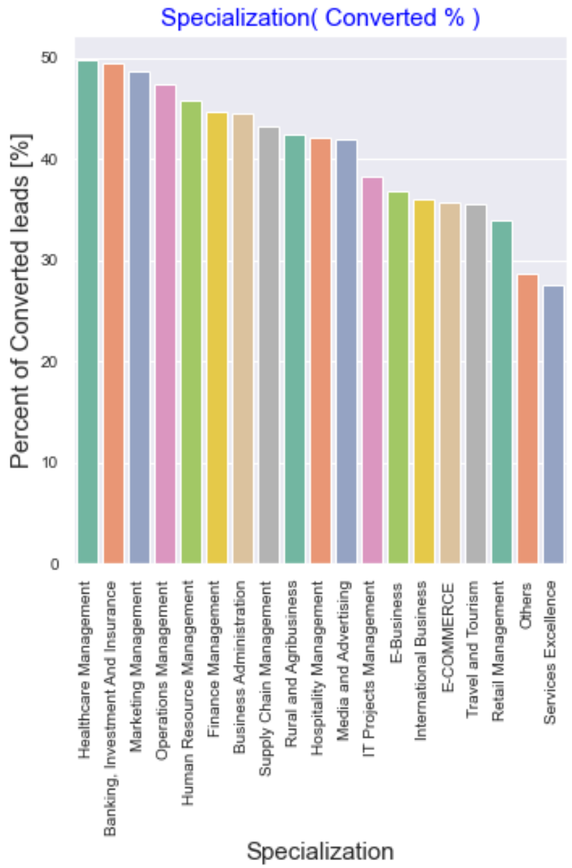


Last Activity

Insights:-

- "Email opened" is the last activity for most of the leads with conversion rate 38%.
- "SMS Sent" is the second highest last activity with Conversion rate of around 62%
- We can considering all other smaller Last Activity types as Other Activity.

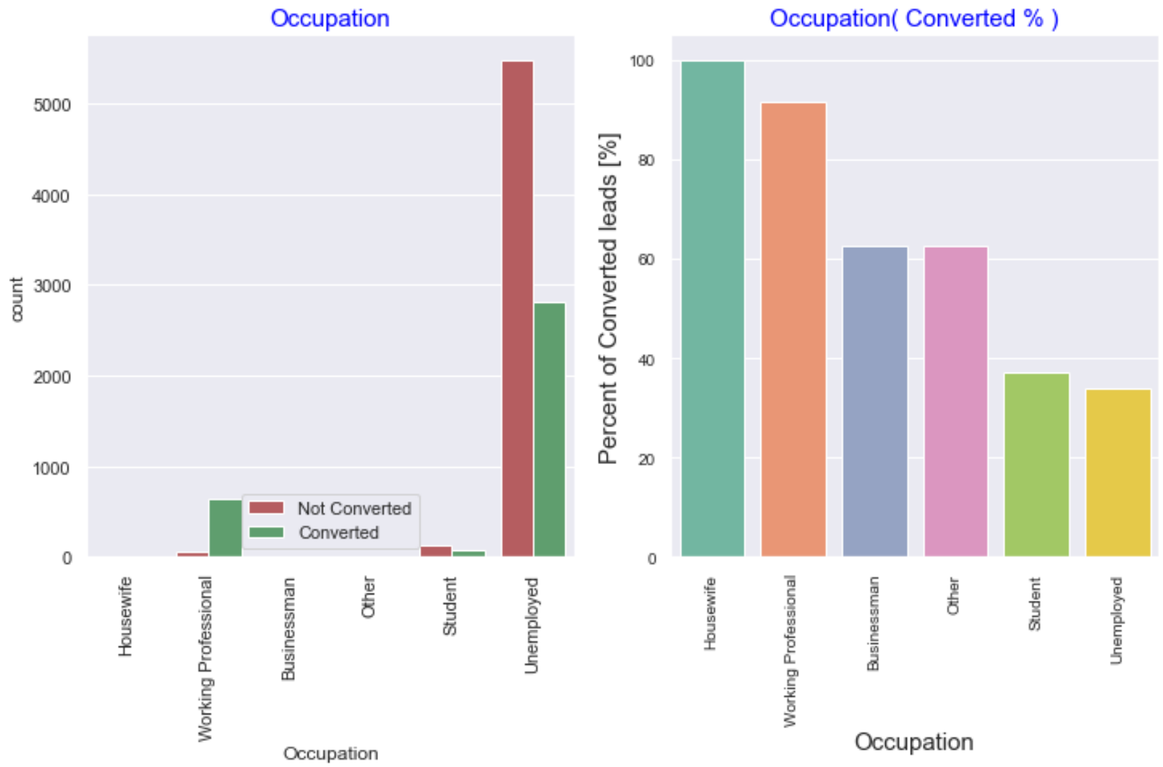




Specialization

Insights:-

- Most of the leads are in "Others" specialization category and around 28% of those conversion rate
- Leads with "Finance management" and "Marketing Management" and "Human Resource Management" are the next best specialization categories even though their numbers are in 3 digit figures but their conversion rate varies between 40-45%.



Occupation

Insights:-

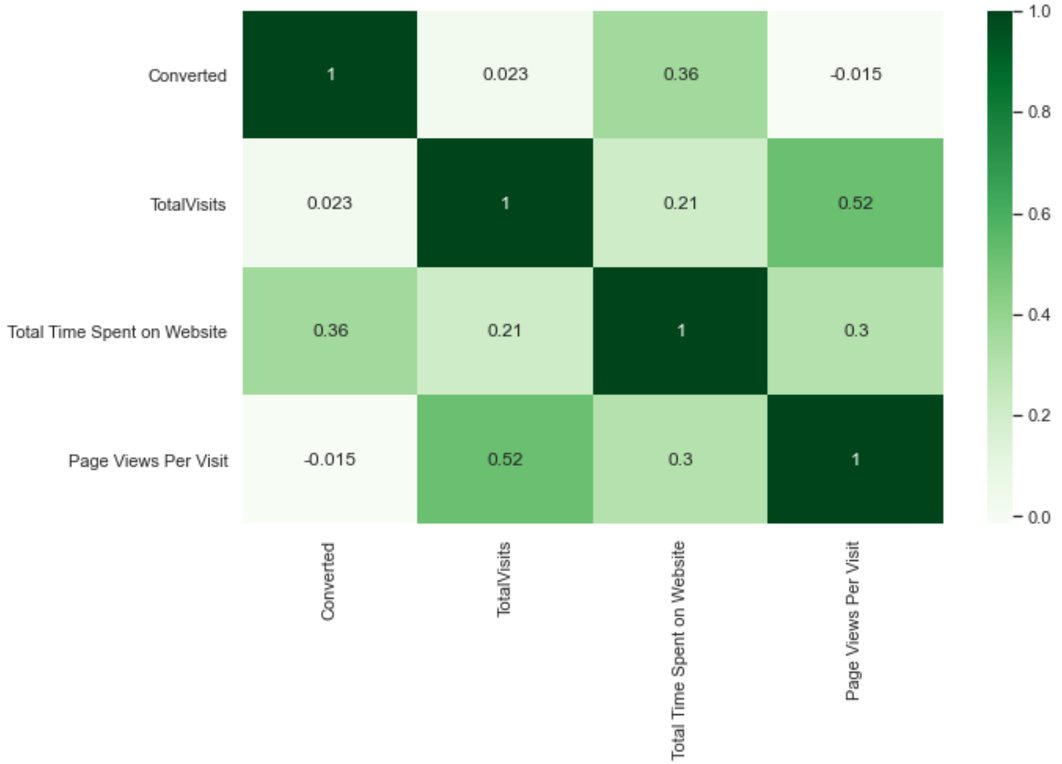
- Though Housewives are very few in numbers, they have 100% conversion rate, so we should try to increase their numbers
- Working professionals, Businessmen and Other category have high conversion rate
- Though Unemployed people have been contacted in the highest number, the conversion rate is low (~40%)
- We cannot combine smaller value categories as their conversion rate is very different. Combing them may lead to wrong predictions.

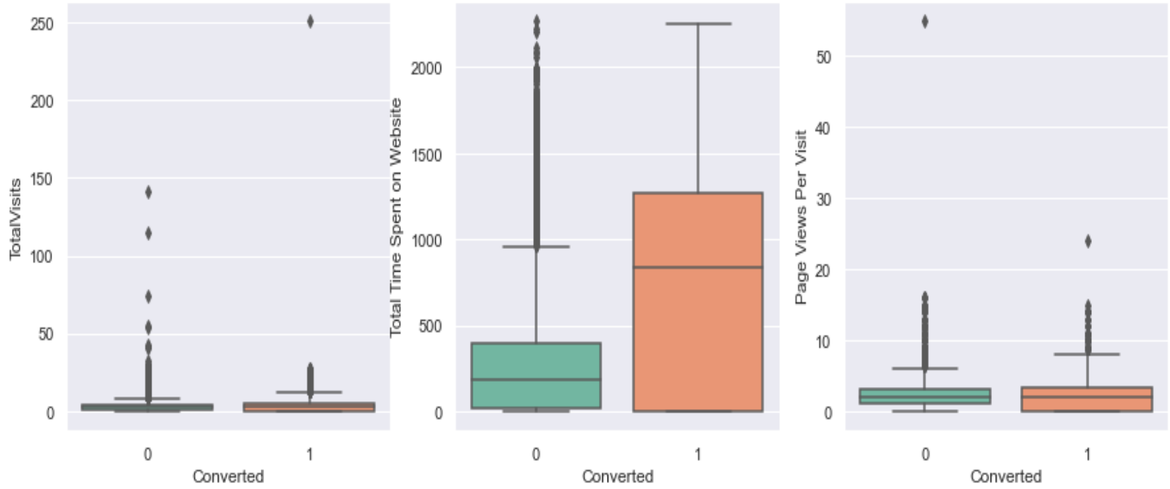
Free Copy

Insights:-

- Free-Copy of Mastering Interview doesn't seem to have much variance if free copy is given or not, so this column is not adding any value to our analysis, hence we will drop it.

Univariate Analysis - Numerical

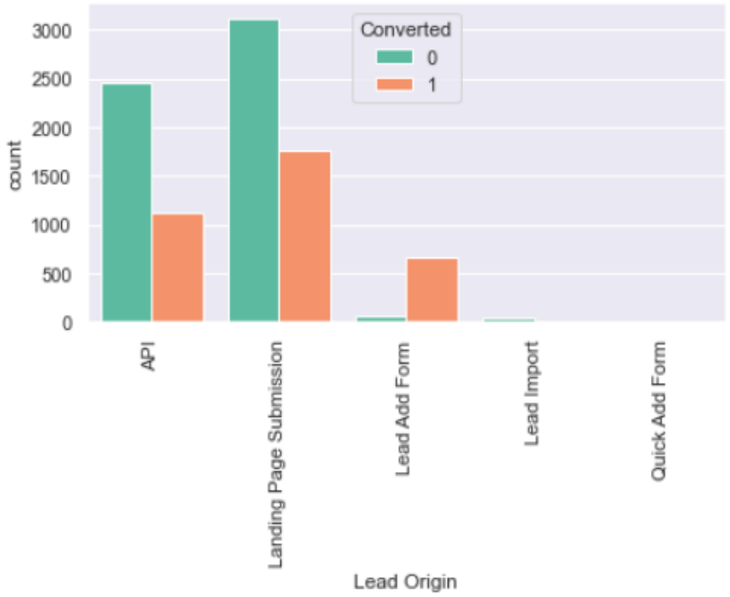


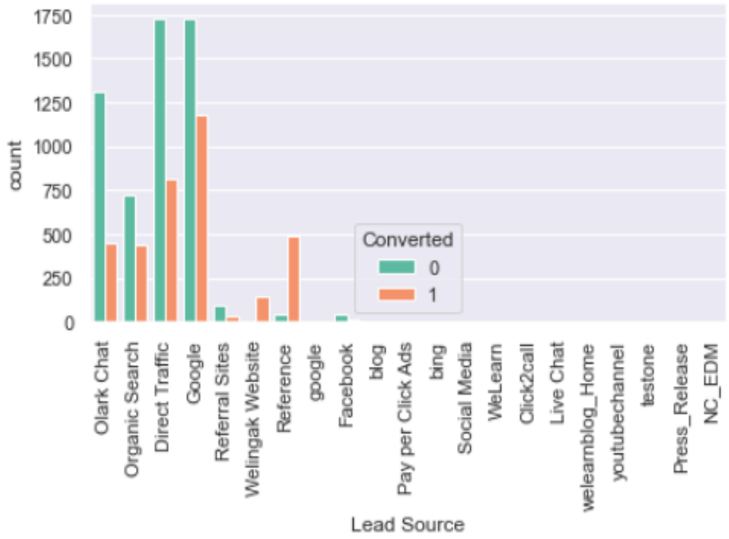


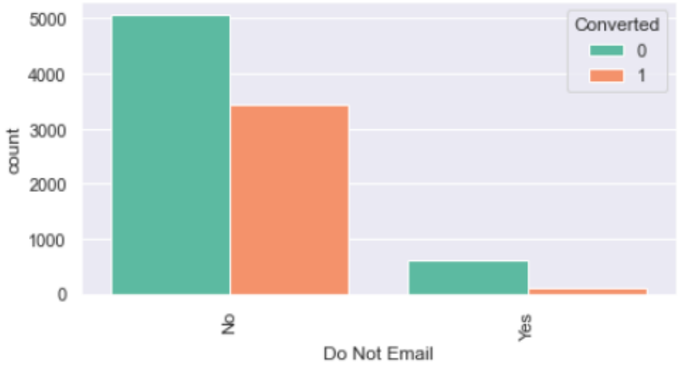
Insights:-

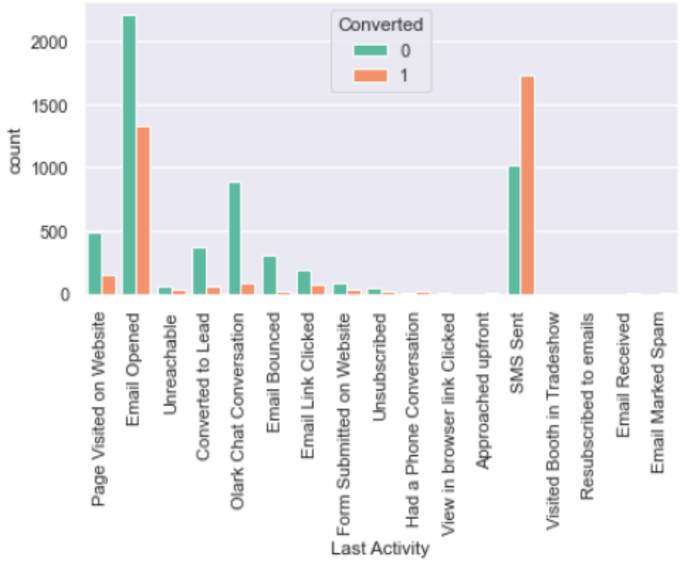
- TotalVisits: It has some outliers which needs to be treated.
- Total Time Spent on Website: People whose spend more time has higher chance of getting converted.
- Page Views Per Visit: It has some outliers which needs to be treated.

Bivariate Analysis





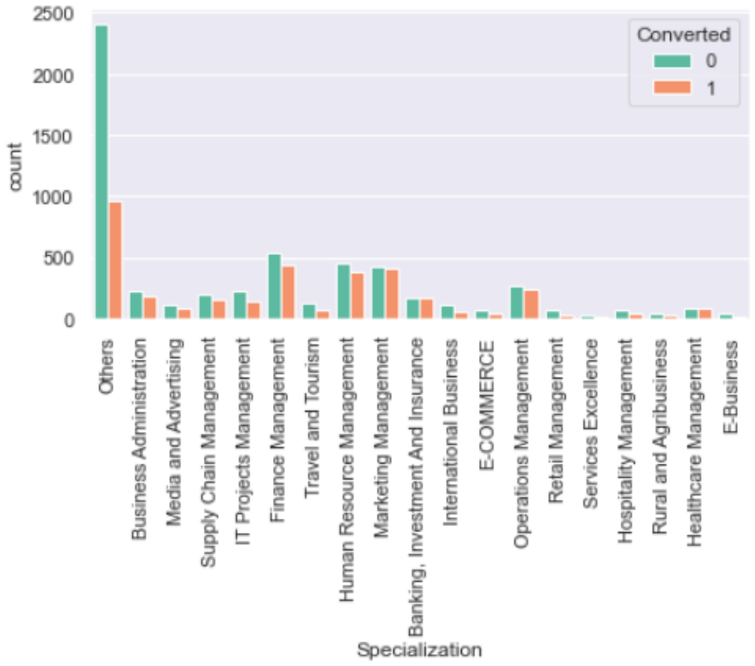


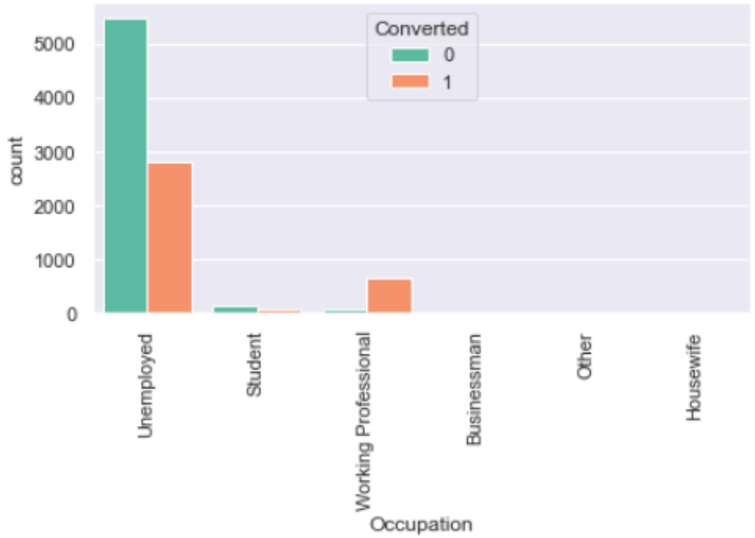


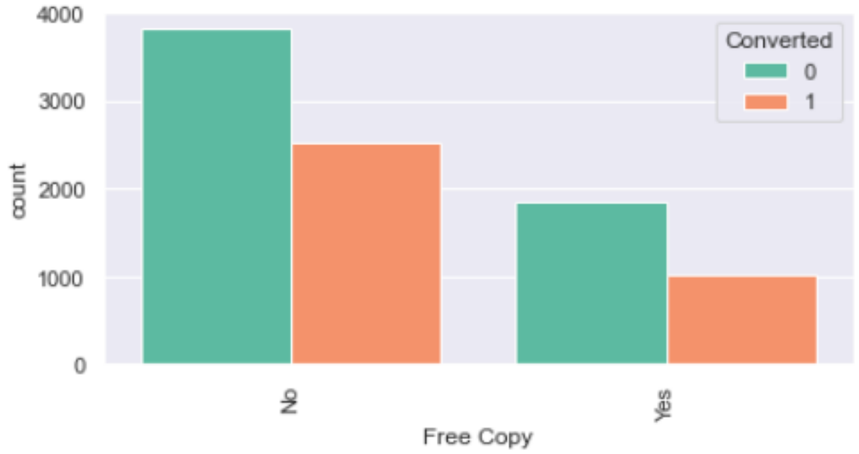
Bivariate Analysis (Categorical)

Insights:-

- Lead Original: Higher leads in "Landing Page Submission" and "API" category
- Lead Source: leads are higher in "Direct Traffic" and "Google" Category
- Do not email: No has higher converted as well as non-converted population
- Last Activity: The number of Hot leads is higher in SMS and in EMAIL category.





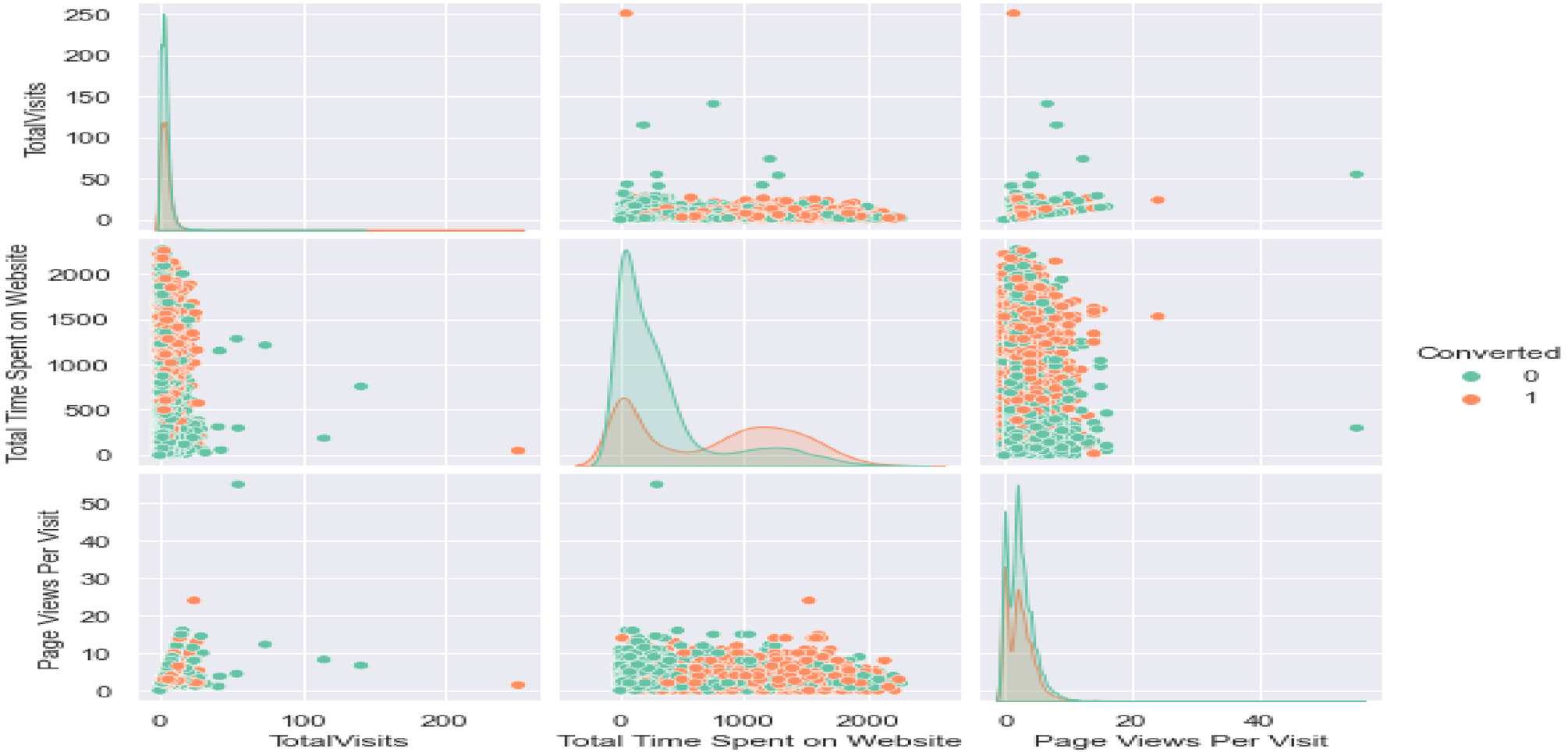


Bivariate Analysis (Categorical)

Insights:-

- Specialization the most of the leads are comes from Finance management.
- Occupation: Mostly unemployed population is catterd to the site, so they have higher converted as well as non-converted population than others.
- Free Copy: Seems to have no significant impact as converted as well as non-converted population are distributed across both categories. We can keep it for now and see its behaviour during modelling step

Bivariate Analysis - Numerical



Insights:-

- Data is not normally distributed.
- Total Visits and Page Views Per Visit has positive correlation among each other.

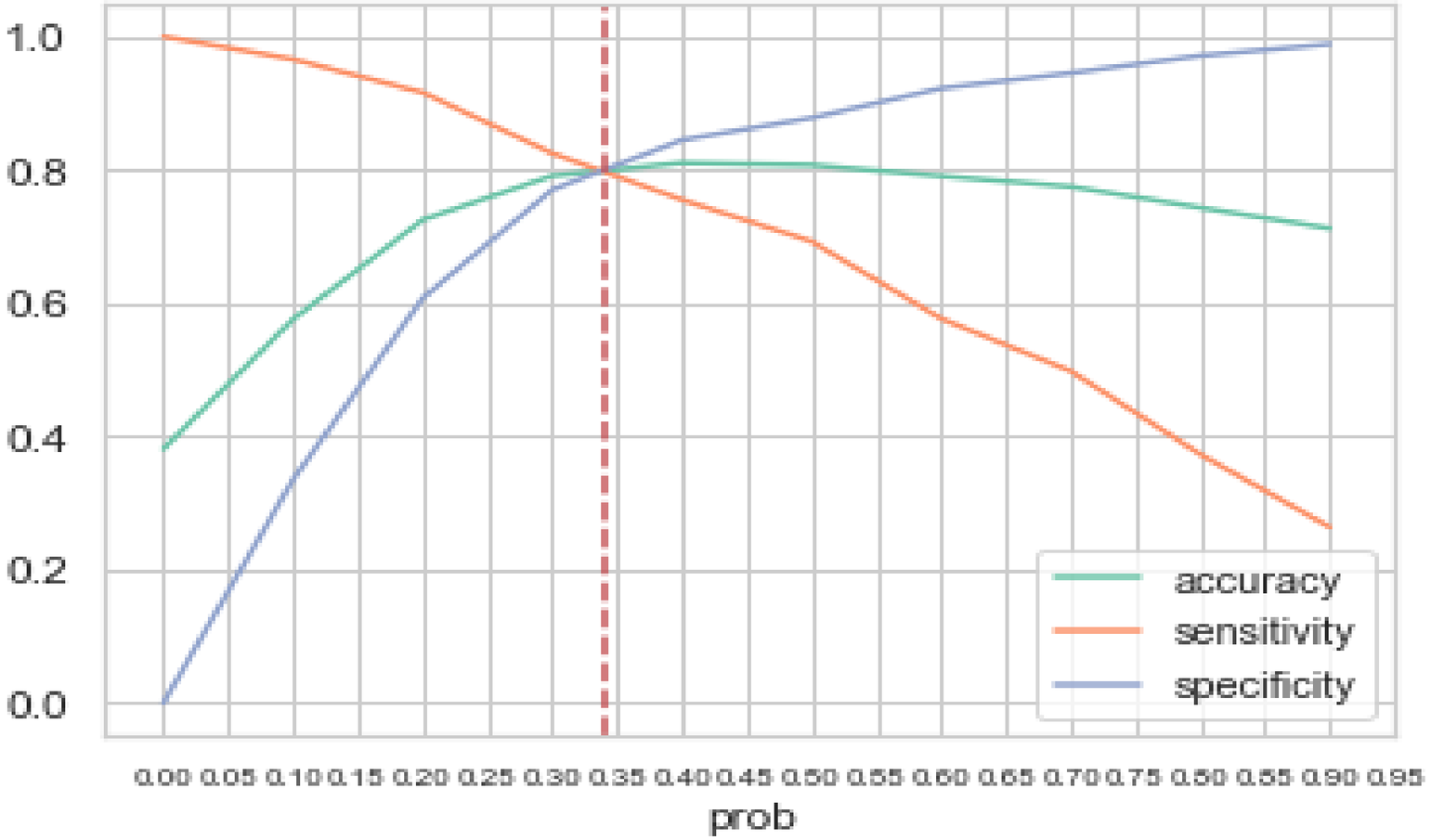
Model Building

Steps Involved:-

- Split the Data into Training and Testing Sets
- The first basic step for regression is to perform a train-test split, we have chosen 70:30 ratio.
- Used RFE for Feature Selection
- Ran RFE with 15 variables as output
- Performed Manual Feature Reduction to build model by removing the variable whose p-value is greater than 0.05 and/or VIF value is greater than 5.
- Performed Predictions on the train data and checked for accuracy, sensitivity, specificity and recall. It is in 80% range
- Performed Predictions on the test data set.
- Overall accuracy, sensitivity, specificity and recall of the test model is also in 80% range.

Model Evaluation

Optimal Cutoff Point



Insights:-

Optimal cut-off probability is that probability where sensitivity and specificity and accuracy meet. We are getting cut-off of 0.34

Specificity = TN/(TN+FP) Sensitivity = TP/(TP+FN) Accuracy = (TP+TN)/(TP+TN+FP+FN)

Confusion Matrix and Logistic Regression Metrics

Train

```
##############################
  Confusion Matrix
[[3225  777]
 [ 492 1974]]
##############################
True Negative                  :  3225
False Positve                  :  777
False Negative                 :  492
True Positive                  :  1974
Model Accuracy value is        :  80.38 %
Model Sensitivity value is     :  80.05 %
Model Specificity value is     :  80.58 %
Model Precision value is       :  71.76 %
Model Recall value is          :  80.05 %
Model True Positive Rate (TPR) :  80.05 %
Model False Positive Rate (FPR) :  19.42 %
##############################################
```

Test

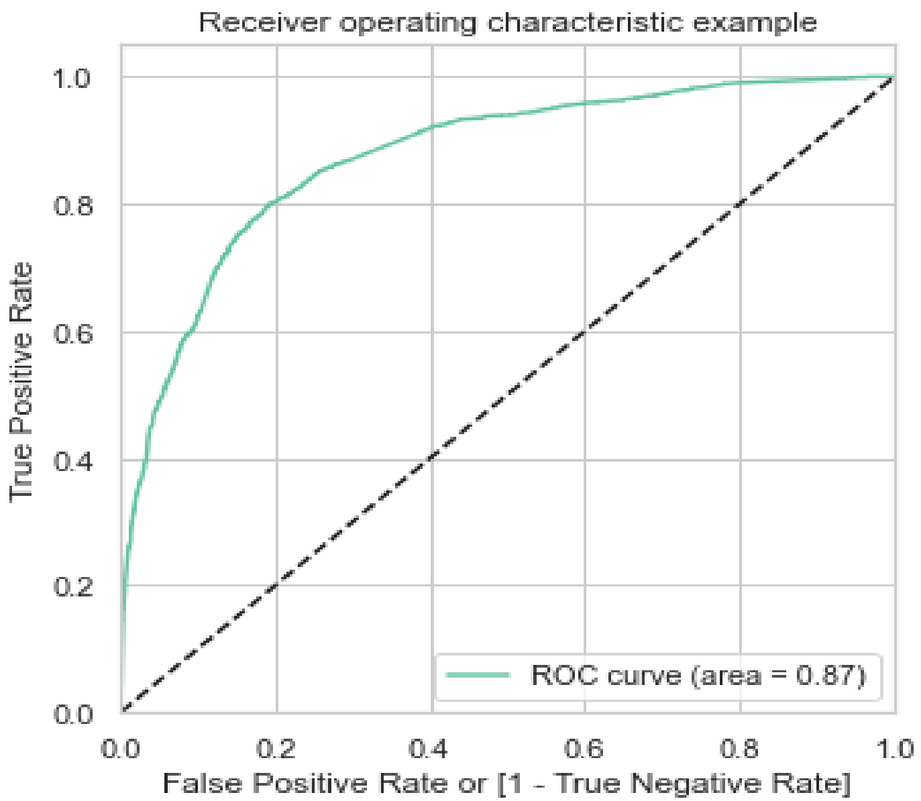
```
##############################
  Confusion Matrix
[[1358  319]
 [ 217  878]]
##############################
True Negative                  :  1358
False Positve                  :  319
False Negative                 :  217
True Positive                  :  878
Model Accuracy value is        :  80.66 %
Model Sensitivity value is     :  80.18 %
Model Specificity value is     :  80.98 %
Model Precision value is       :  73.35 %
Model Recall value is          :  80.18 %
Model True Positive Rate (TPR) :  80.18 %
Model False Positive Rate (FPR) :  19.02 %
##############################################
```

Insights:-

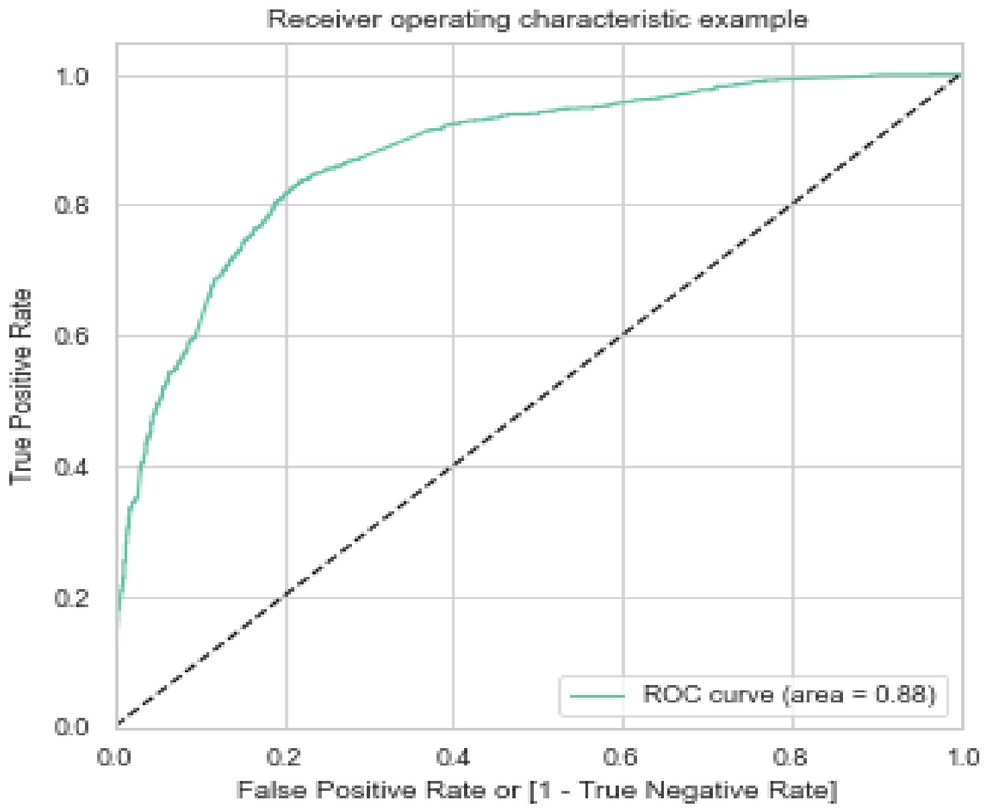
Using cut off value at 0.34 Sensitivity of 80.05% in Train and 80.18% in Test. Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting. More than 80% is what the CEO has requested in this case study. Accuracy is also 80.38% which is also good.

ROC Curve

Train



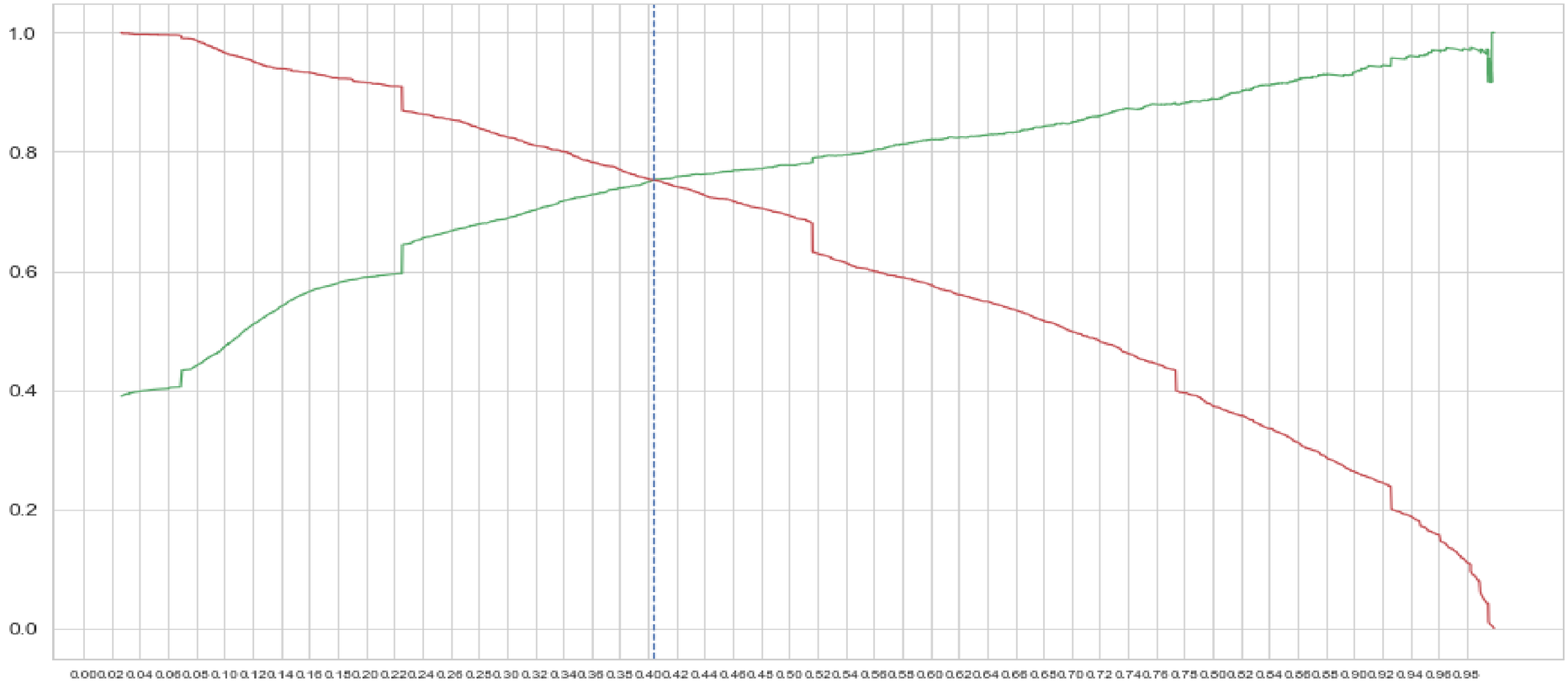
Test



Insights:-

ROC Curve area is 0.87 for Train and 0.88 for Test model, which indicates that the model is good because the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate is it.

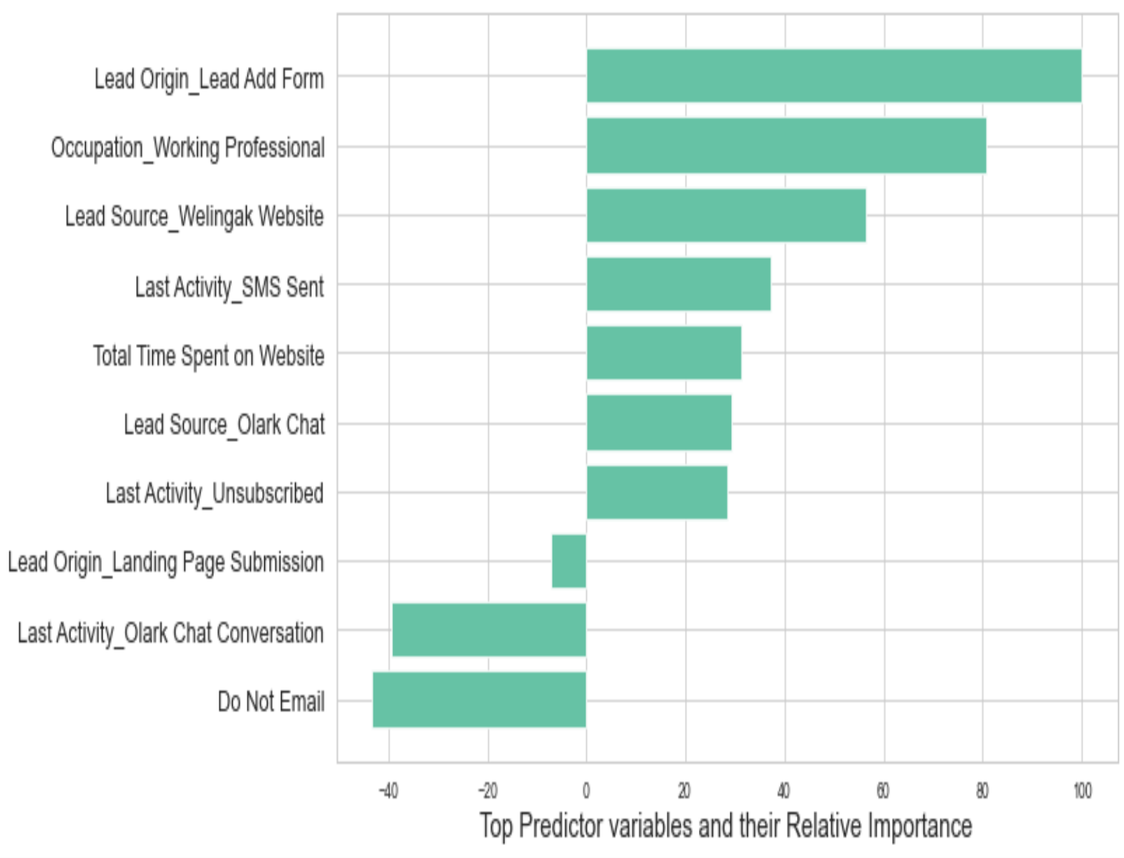
Precision – Recall Trade-off



Insights:-

Based on Precision-Recall Trade off curve, the cut-off point is0.404. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. He wants people to be correctly identified as leads with 80% success i.e. True Positive , True Positive Rate ,Sensitivity, Recall should be close to 80%, which we are getting using the previous cut off of 0.34. These number decreased by using 0.404 cut-off . Hence we will go for 0.34 cut-off.

Summary



After running the model on the Train and Test Data evaluation metrics meet the goals of X-Education CEO, which is to achieve 80% target lead conversion rate to be around 80%. It give the CEO confidence in making good calls based on this model.

Evaluation Metrics are:-

Train Data:

Model Accuracy value is : 80.38 %
Model Recall value is : 80.05 %
Model Sensitivity value is : 80.05 %
Model Specificity value is : 80.58 %

Test Data:

Model Accuracy value is : 80.66 %
Model Recall value is : 80.18 %
Model Sensitivity value is : 80.18 %
Model Specificity value is : 80.98 %

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable. Here, the logistic regression model is used to predict the probability of conversion of a customer.
- The Optimum cut off we chose after building the mode is 0.34 i.e. any lead with greater than 0.34 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.34 or less probability of converting is predicted as Cold Lead (customer will not convert)
- Our final Logistic Regression Model is built with 10 features.
- Features used in final model are ['Do Not Email', 'Total Time Spent on Website', 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'Last Activity_Unsubscribed', 'Occupation_Working Professional']
- The top three variables which impact the lead conversion positively are:-
  - Lead Origin_Lead Add Form 3.48
  - Lead Source_Welingak Website 1.97
  - Occupation_Working Professional 2.81
  - Prospects having Lead Origin as "Lead add form" and source as "Welingak Website" and those who are working professionals have high probability of conversion and should be catered to maintain the Sensitivity of the model
- The three variables which impact the lead conversion negatively are:-
  - Do Not Email -1.51
  - Lead Origin_Landing Page Submission -0.25
  - Last Activity_Olark Chat Conversation -1.37
- The final model has Sensitivity of 80.05 % , this means the model is able to predict 80.05 % customers out of all the converted customers, (Positive conversion) correctly.

Recommendation

Recommendations to the X Education for Potential buyers/Hot Leads:-

- During EDA we found that, Leads from the "Lead Add Form" have third highest conversions with conversion rate around 90%. It also emerged as a most significant attribute for Hot Leads hence we should try to put Lead Add Forms on the social media websites specially on the Welingak Website and we should give more importance to customers you came through this channel.
- Working professionals have higher chance(around 90%) to convert as they dont have any monetaries restrictions, So more focus should be given in engaging with the Working professionals.
- Welingak Website has around 98% lead conversion rate. More adds should be given on this website to cater the leads from their, as it has higher chance to conversion.
- During EDA, we discovered that a lead that came through a "reference" has over 90% conversion. To increase lead count, we should encourage and incentivize existing members to bring more of their referrals.

Thank you

Ankur Napa
Amandeep Kaur