Lead Scoring for X Education

This Case study gives an idea of applying Logistic Regression in real business scenario where an education company X is trying to improve its Lead conversion rate.

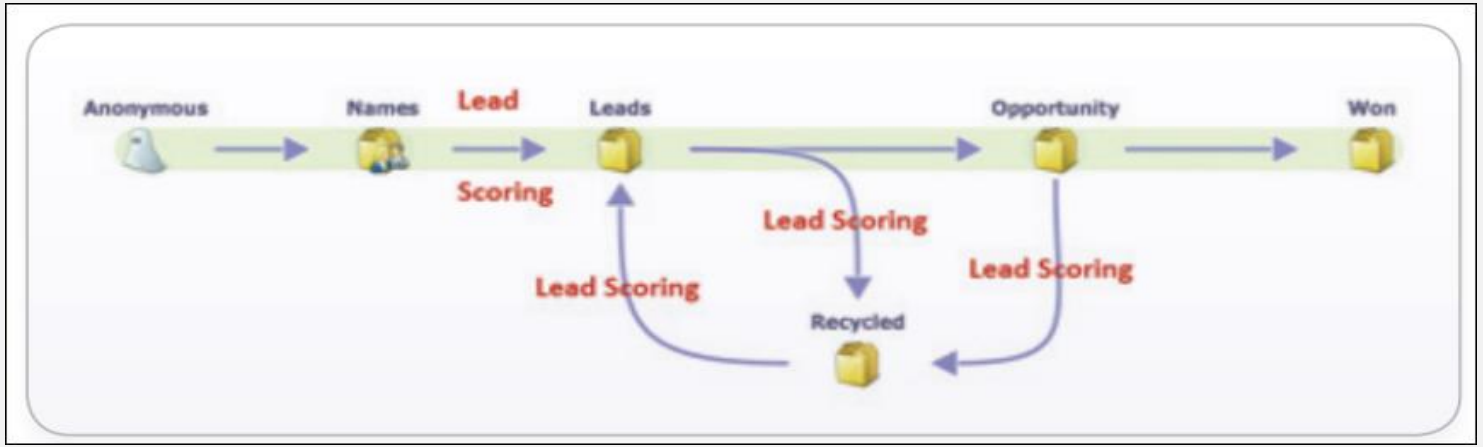
Lead Scoring is a way to categorize prospects by level of commitment. As a marketing segmentation, one can group people based on certain categories and then predict based on those factors that lead is going to convert or not.

The objective is to help X Education select the most promising leads by building a model and assigning a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

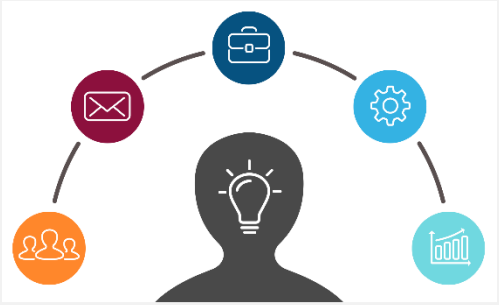
Bhavnish Marwaha
Piyushi Prenam

Business Understanding

- Lead scoring provides a segmentation engine that your marketing automation activities need to thrive, hence majority of effective marketers counted lead scoring as a top contributor to their revenue



Business Objective

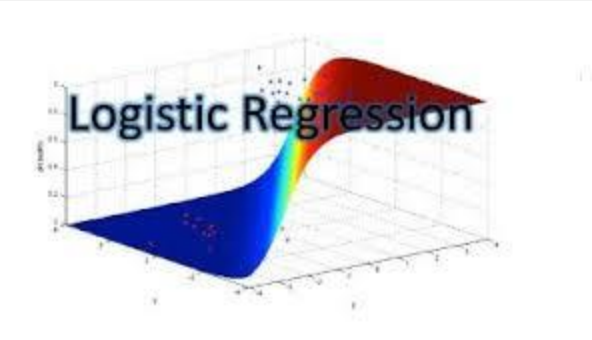


- Business Objective here is to find the most promising leads by building a model and assigning a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

- And to achieve above two things must be taken care. First is the Large scale and second is differentiating between the likelihood percentage.

- Lead Score can introduce more values related to demographics(ex: geography,gender,age etc) and use behavior(actions taken on different pages),resulting in super-targeted and personalized real time campaigns for leads moving through every step of the funnel.

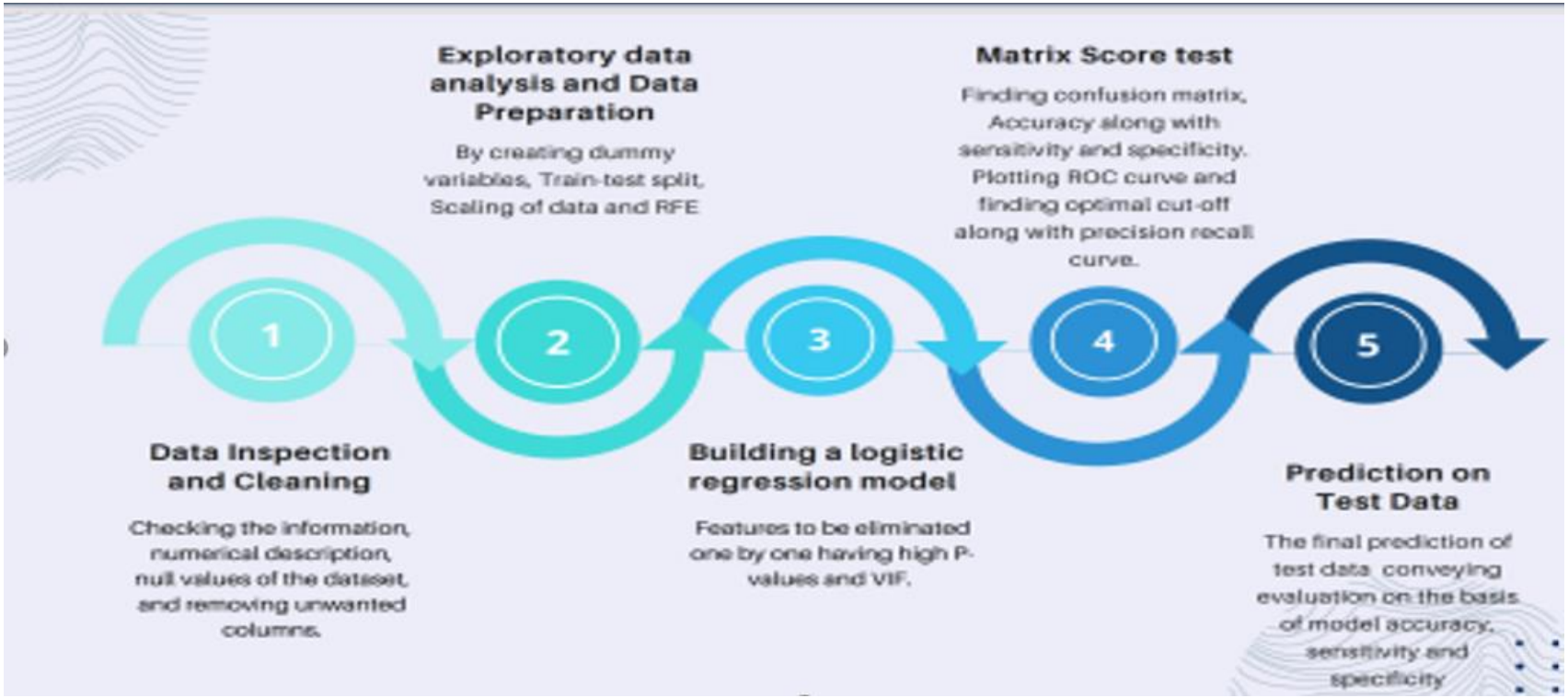
Table of Content



Libraries Used

1. NumPy: Add support for large, **multi-dimensional arrays** and matrices, along with a large collection of **high-level mathematical functions** to operate on these arrays
2. Pandas: Fast, powerful, flexible and easy to use open source data analysis and manipulation tool. We are using pandas libraries to **read the file and do data analysis**
3. Matplotlib: Comprehensive library for creating static, animated, and interactive **visualizations in Python**
4. Seaborn: For making **statistical graphics in Python**. It builds on top of matplotlib and integrates closely with pandas data structures
5. Sklearn: **Scikit-learn** is a free machine learning **library** for **Python**. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports **Python** numerical and scientific **libraries** like NumPy and SciPy
6. Statsmodel: **Statsmodels** is a Python package that allows users to explore data, **estimate statistical models, and perform statistical tests**. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator

Approach


1
Data Inspection and Cleaning
Checking the information, numerical description, null values of the dataset, and removing unwanted columns.
2
Exploratory data analysis and Data Preparation
By creating dummy variables, Train-test split, Scaling of data and RFE
3
Building a logistic regression model
Features to be eliminated one by one having high P-values and VIF.
4
Matrix Score test
Finding confusion matrix, Accuracy along with sensitivity and specificity. Plotting ROC curve and finding optimal cut-off along with precision recall curve.
5
Prediction on Test Data
The final prediction of test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

Approach Taken

- Data Inspection and Cleaning

  Checking the basic characteristics of the data as shape, info etc

  Data Imbalance Analysis is done using Target Variable – "Converted"

  Data Cleaning:

  - 'Select' are mostly the values where nothing is selected so replacing them with Nan.
  - Checking % of missing values and drop or Impute missing values depending on the %
  - Dropping Unwanted columns
  - Drop columns that are going to be used after lead is converted by Sales Team

- Exploratory Data Analysis and Data Preparation

  - Constant features can be removed as these provide no information in classification
  - Categorial features should be identified for creating dummy variables for data preparation
  - Univariate analysis on Boolean features 'Yes','No' can be mapped to 0 and 1 or dropped.
  - Missing value handling for Numeric Features
  - Outlier handling for Numeric Features
  - Categorial feature analysis (replacing missing values by mode and used bucketing)
  - Visualizing the data using various plots
  - Checking correlation between variables

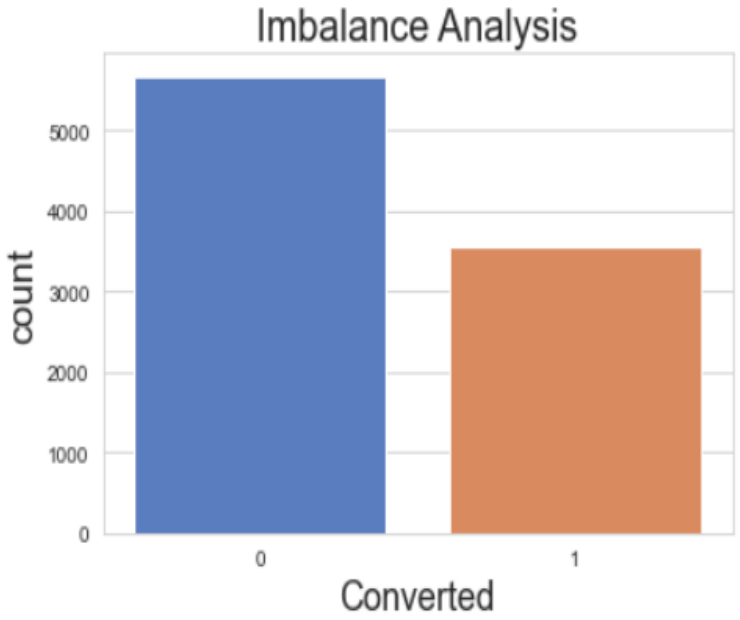
Approach Taken

- Building a Logistic Regression Model:
  - Dummy variable creation
  - Split test-train data at 70%,30% ratio
  - Scaling the numeric features using MinMaxScalar
  - Build the model using combination of automatic (RFE) and manual processing (dropping features one by one)
  - Fit the training data

  Getting the predicted values on train set
  - Used cutoff of 0.5 for the initial predictions
  - Evaluating the metrics by deriving Classification report and classification metrics with the initial cutoff and predictions (Confusion metrics calculation)
  - Deriving AUC (area under ROC curve)
  - Getting the optimal cutoff and final evaluation Metrics for Train Dataset
  - Plotting Confusion metrics for final predictions on train data
  - Assigning Lead score to the training data set
  - Measuring the precision-recall tradeoff
- Matrix Score test
- Prediction on Test Data
- Deriving Classification report and Classification metrics and making final predictions
- Final observation and Recommendations

Analyzing Data Imbalance using Target Variable – ‘Converted’

We can see that the Application data is not too much imbalanced with imbalance ratio of approx. 2%



Handling Null Values

	count_missing	percent_missing
How did you hear about X Education	7250	78.46
Lead Profile	6855	74.19
Lead Quality	4767	51.59
Asymmetrique Profile Score	4218	45.65
Asymmetrique Activity Score	4218	45.65
Asymmetrique Profile Index	4218	45.65
Asymmetrique Activity Index	4218	45.65
City	3669	39.71
Specialization	3380	36.58
Tags	3353	36.29
What matters most to you in choosing a course	2709	29.32
What is your current occupation	2690	29.11
Country	2461	26.63
TotalVisits	137	1.48
Page Views Per Visit	137	1.48
Last Activity	103	1.11

- For this analysis, created Function to **calculate missing values** and **missing percentage for the dataframes**
- **Drop unwanted variables** 'Prospect Id' and 'Lead Number'
- Analysis done on **Score columns** assigned by sales team and then dropped them.

  'Lead Quality'

  'Asymmetrique Activity Index' and 'Asymmetrique Activity Score'

  'Asymmetrique Profile Index' and 'Asymmetrique Profile Score'

- **'Tag' and 'Last Activity'** are the columns added by sales team and do not directly contribute and dropped.

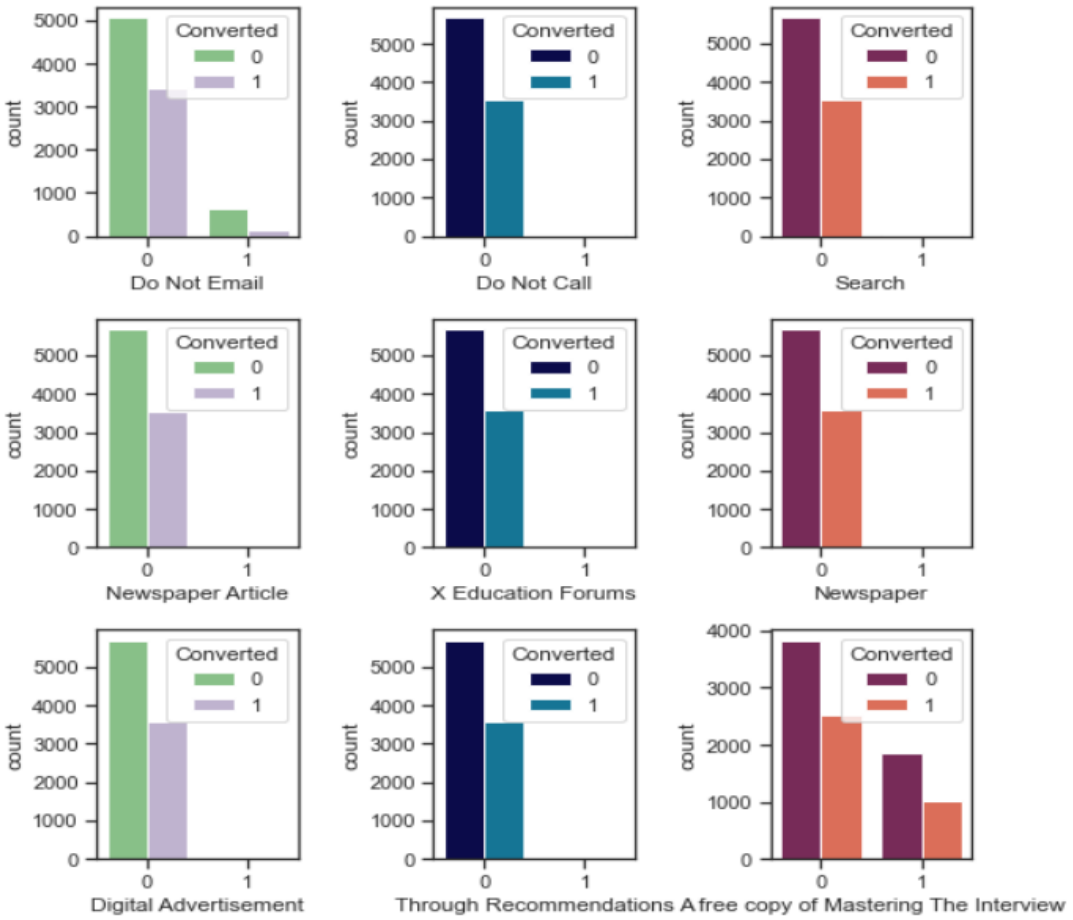
- Dropped value if columns has >70% (almost null values) and replaced with mode for Categorial values if values >45% null values.

Univariate Analysis

- Univariate analysis for Boolean features
  Map ‘Yes’ or ‘No’ to 1 and 0
  Drop the skewed data if present

‘Do Not Email’
‘Do Not Call’
‘Search’
‘Newspaper Article’
‘X Education Forums’
‘Newspaper’
‘Digital Advertisement’
‘Through Recommendations’
‘A free copy of Mastering the Interview’

Recommendations: Only, ‘**A free copy of Mastering the Interview**’ and ‘**Do Not Email**’ has (majority of values) both values as 0 and 1, so dropped all other columns with very less value as 1 as they do not contribute much.



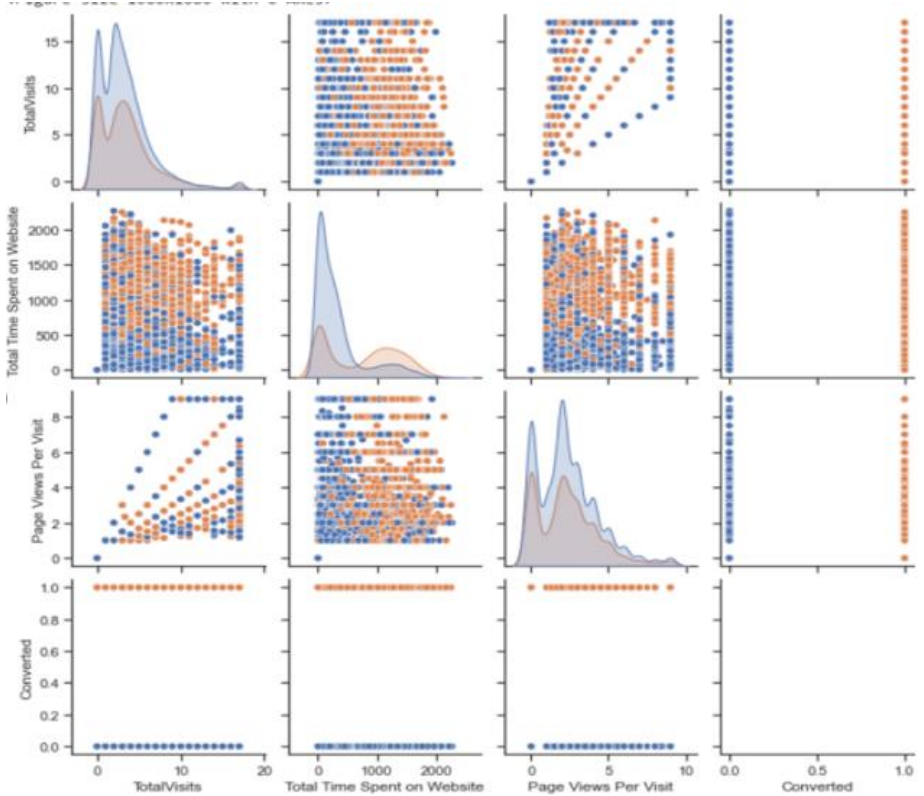
EDA and Data Analysis for Numerical Features

Numerical Features:

- ‘Total Visits’
- ‘Total Time Spent on Website’
- ‘Page Views per Visit’

Recommendation:

- We can use capping for ‘Total Visits’ and ‘Page Views per Visit’ as they have outliers for better analysis and data preparation

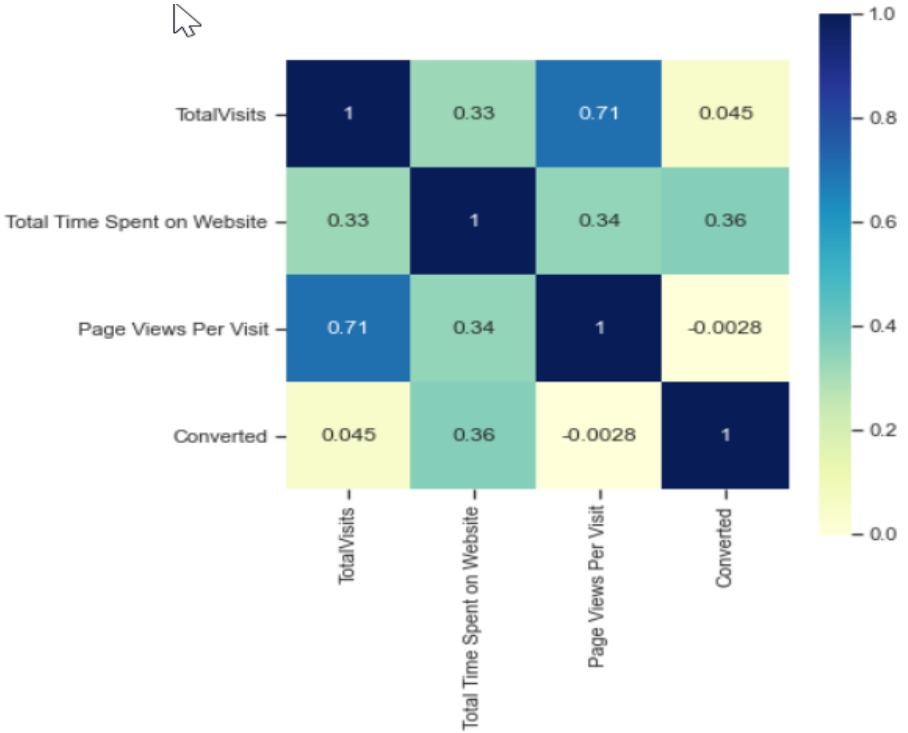


EDA and Data Analysis for Numerical Features

Correlation matrix of Numerical Features:

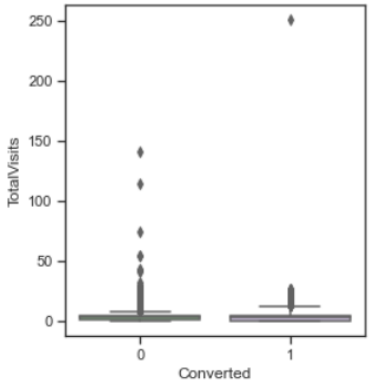
- ‘Total Visits’
- ‘Total Time Spent on Website’
- ‘Page Views per Visit’

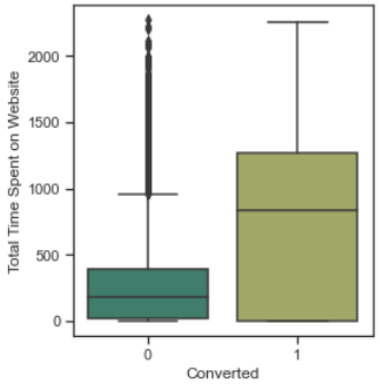
Observation: Total Visits and Page Views per Visit are correlated to each other

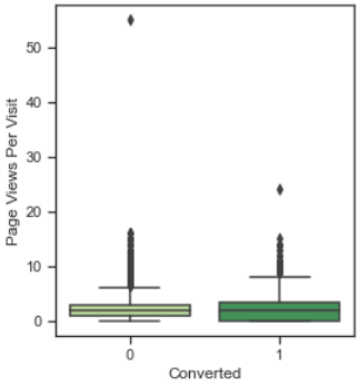


Handling Outliers

- 'Total Visits' and 'Page Views per visit' columns has outliers.
- Created Box Plot and frequency distribution plots to analyze them.
- Used floor(1%) and capping(99%) for removing outliers
- Below plots shows Before capping and after capping views

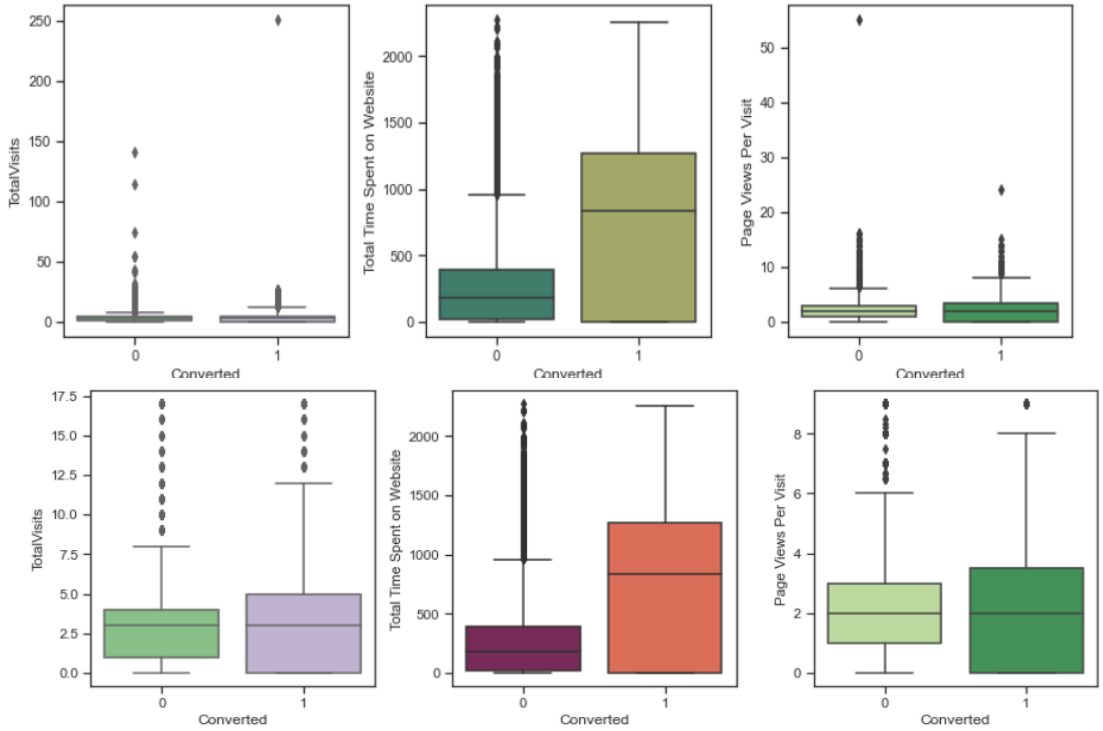






Handling Outliers

- 'Total Visits' and 'Page Views per visit' columns has outliers.
- Created Box Plot and frequency distribution plots to analyze them.
- Used floor(1%) and capping(99%) for removing outliers
- Below plots shows Before capping and after capping views



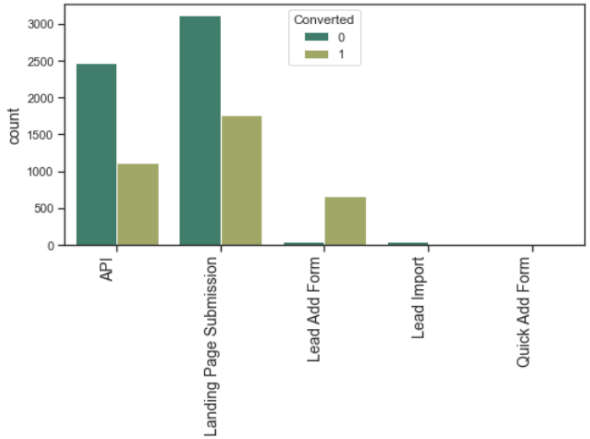
EDA and Data Analysis for Categorial Features

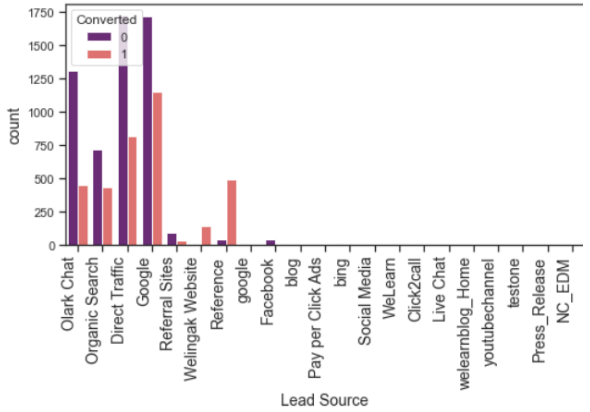
Categorial Features:

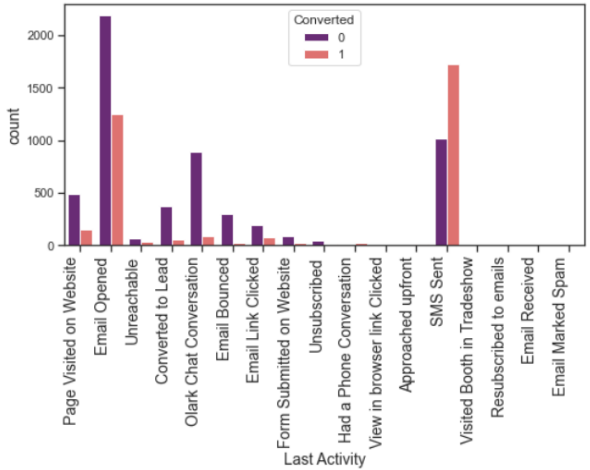
- ‘Lead Origin’
- ‘Lead Source’
- ‘Last Activity’
- ‘Country’
- ‘Specialization’
- ‘What is your current occupation’
- ‘What matters most to you in choosing a course’
- ‘City’

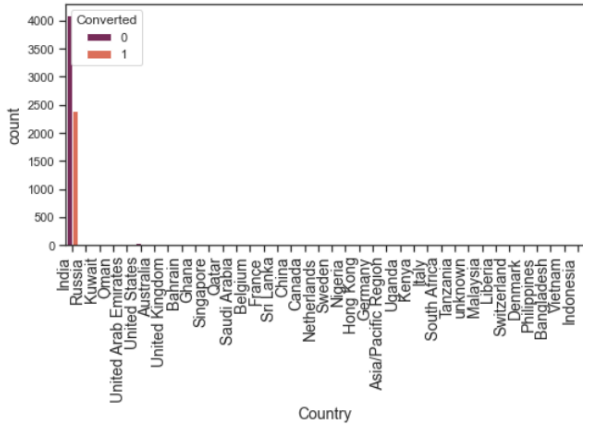
Recommendation:

- We can drop ‘**Country**’ and ‘**What matters most to you in choosing a course**’ as they are highly skewed columns.









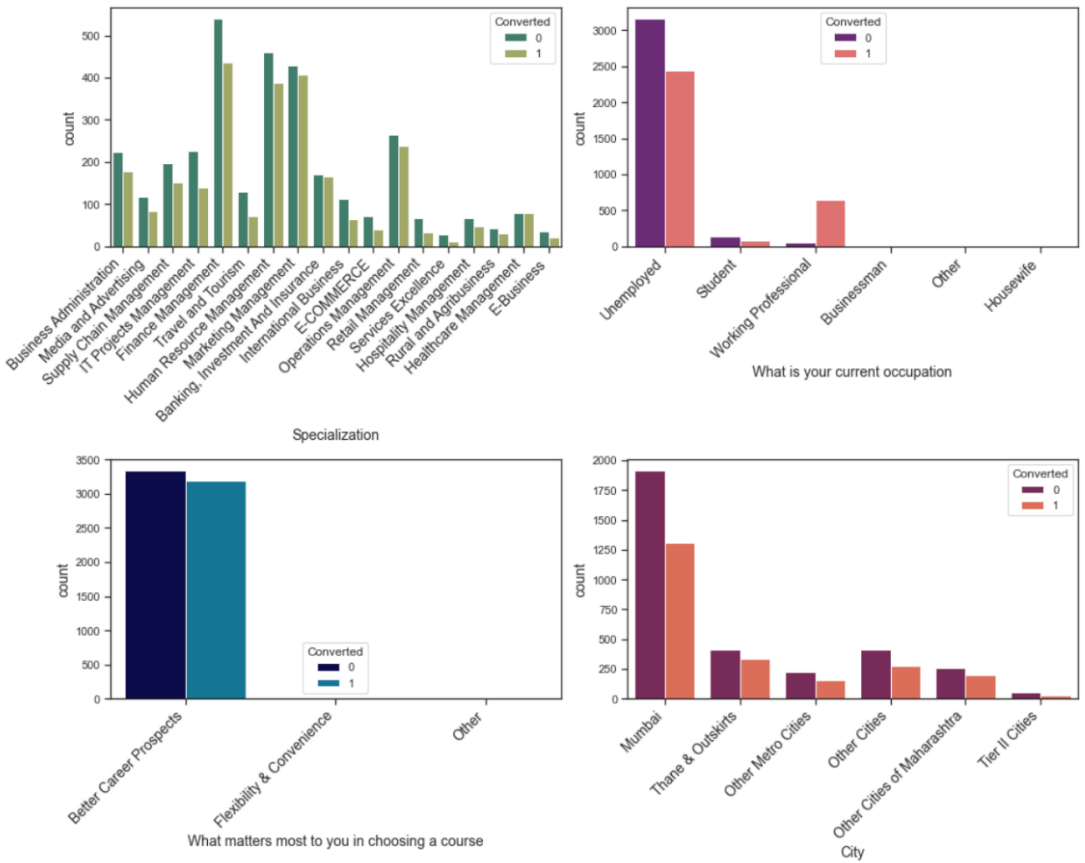
EDA and Data Analysis for Categorial Features

Categorial Features:

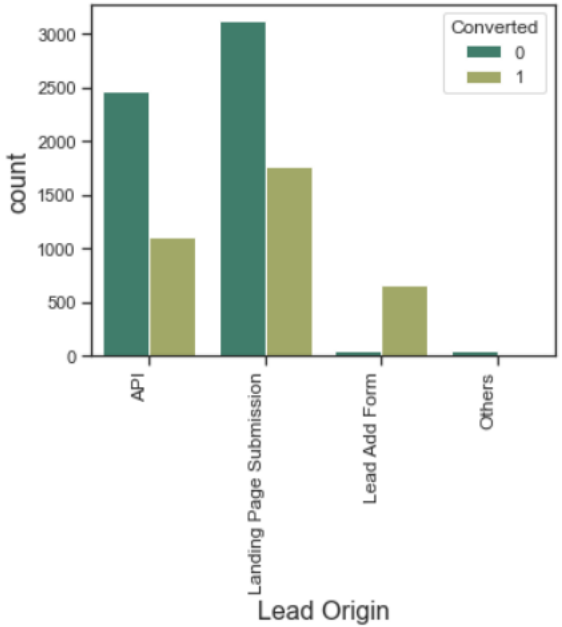
- ‘Lead Origin’
- ‘Lead Source’
- ‘Last Activity’
- ‘Country’
- ‘Specialization’
- ‘What is your current occupation’
- ‘What matters most to you in choosing a course’
- ‘City’

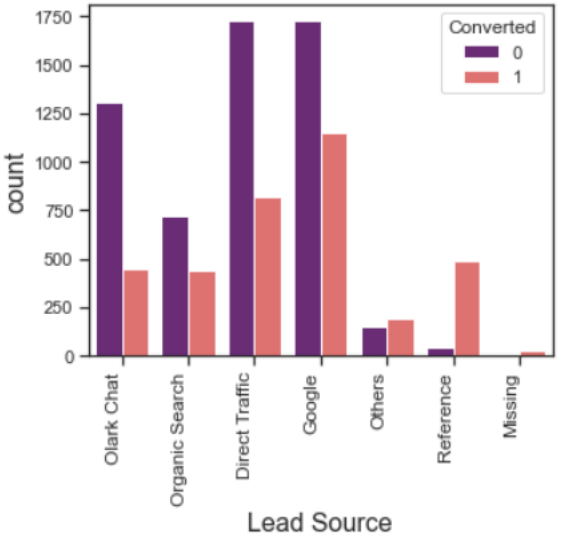
Recommendation:

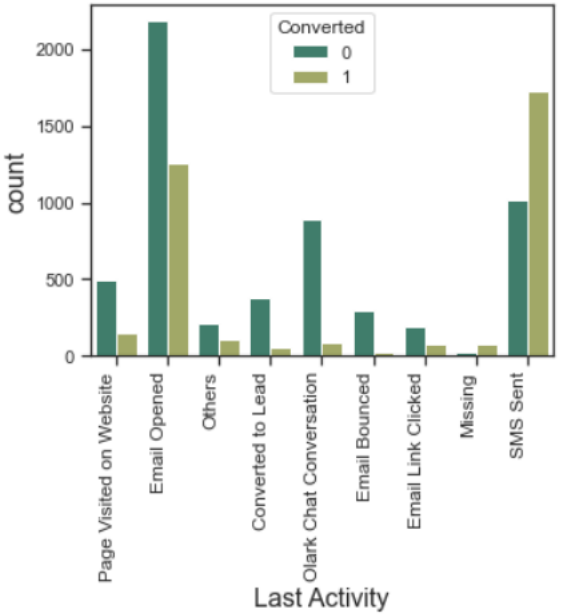
- We can drop ‘**Country**’ and ‘**What matters most to you in choosing a course**’ as they are highly skewed columns.



EDA and Data Analysis for Categorial Features after handling missing values and Bucketing or binning (Bivariate Analysis)

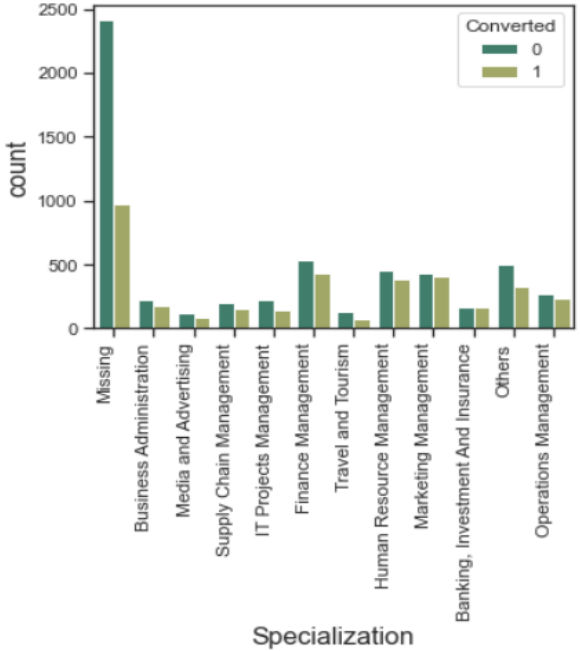


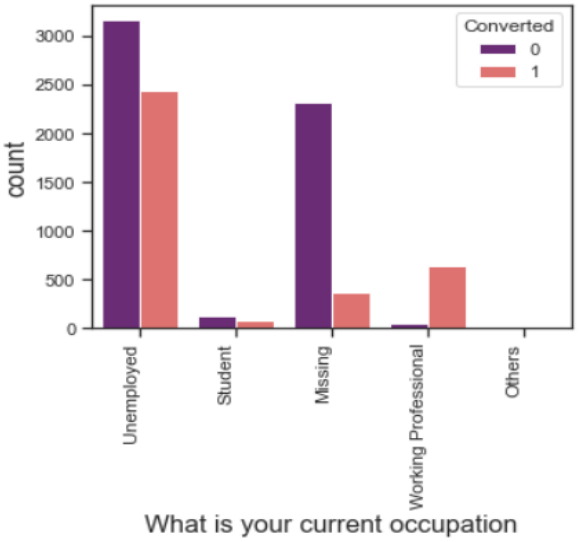


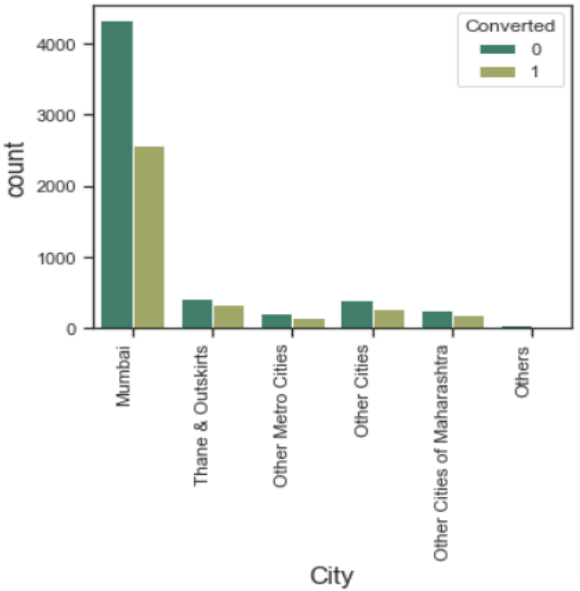


Observation: 'Email Opened' and 'SMS Sent' has high conversion rate. Also those who landed to submission page also have good conversion rate. Lead source having Google or direct traffic has good conversion rate.

EDA and Data Analysis for Categorial Features after handling missing values and Bucketing or binning (Bivariate Analysis)







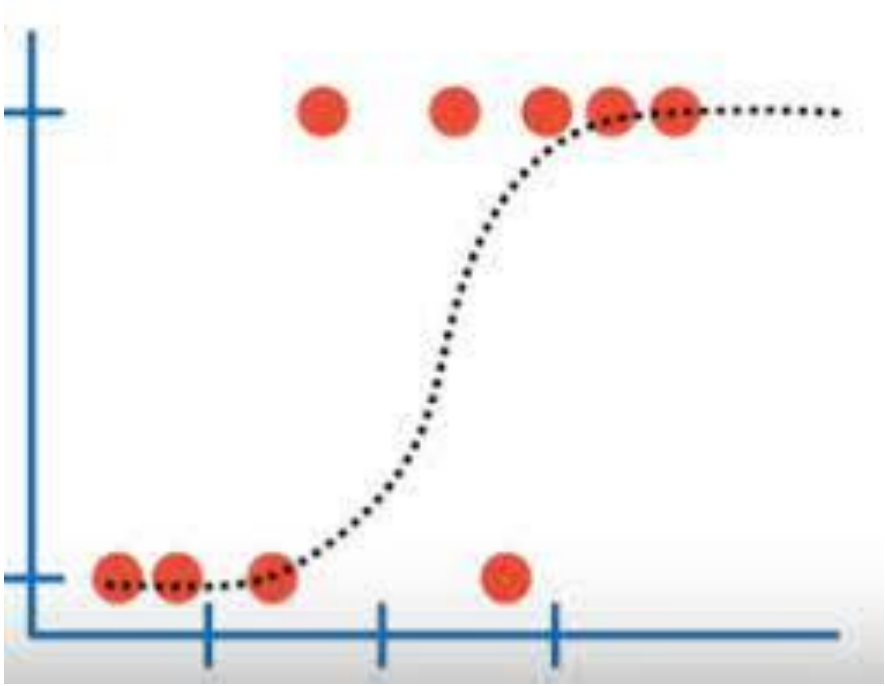
Observation: 'Working Professional' showed higher conversion rate. 'Unemployed' shows both converted as well as non-converted.

Logistic Regression Model Building

Process followed:

- Automatic (using RFE Features)
- Manual(eliminating features one by one)

Defined a generic functions for generate model and calculate vif and used in 6 iterations



Final Model

Observation:

14 variables lead to meaningful results.
Total Visits,Total Time Spent on Website, Last Activity_SMS_Sent, current occupation Working Professional shows positive coefficient and contribute to good lead conversion rate.

Last_Activity_Olark_chat_conversion, Do_not _Email shows negative coefficient

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6453
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2671.2
Date:	Sun, 28 Mar 2021	Deviance:	5342.3
Time:	14:53:38	Pearson chi2:	6.69e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9116	0.155	-25.171	0.000	-4.216	-3.607
Do Not Email	-1.3283	0.174	-7.613	0.000	-1.670	-0.986
TotalVisits	1.0007	0.219	4.561	0.000	0.571	1.431
Total Time Spent on Website	4.4864	0.166	27.060	0.000	4.161	4.811
Lead Origin_Landing Page Submission	-0.3171	0.089	-3.568	0.000	-0.491	-0.143
Lead Origin_Lead Add Form	3.6056	0.199	18.148	0.000	3.216	3.995
Lead Source_Olark Chat	1.3750	0.128	10.775	0.000	1.125	1.625
Last Activity_Email Opened	0.7197	0.108	6.657	0.000	0.508	0.932
Last Activity_Olark Chat Conversation	-0.7617	0.186	-4.103	0.000	-1.125	-0.398
Last Activity_Others	0.9241	0.201	4.598	0.000	0.530	1.318
Last Activity_SMS Sent	1.9050	0.111	17.219	0.000	1.688	2.122
What is your current occupation_Others	2.1324	0.518	4.117	0.000	1.117	3.148
What is your current occupation_Student	1.2275	0.237	5.187	0.000	0.764	1.691
What is your current occupation_Unemployed	1.2189	0.086	14.152	0.000	1.050	1.388
What is your current occupation_Working Professional	3.7400	0.205	18.278	0.000	3.339	4.141

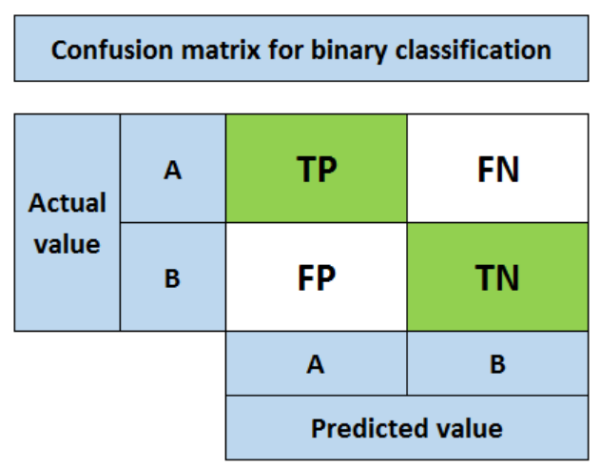
Final VIF's

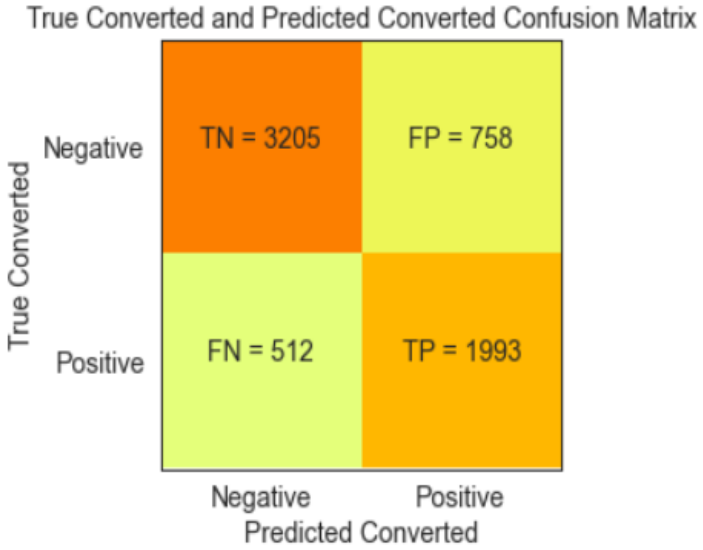
Observation: VIF's are less than 3 in final model from training data set.

Features	VIF
Lead Origin_Landing Page Submission	3.07
What is your current occupation_Unemployed	2.91
TotalVisits	2.69
Last Activity_Email Opened	2.38
Total Time Spent on Website	2.20
Last Activity_SMS Sent	2.18
Lead Source_Olark Chat	1.93
Last Activity_Olark Chat Conversation	1.57
Lead Origin_Lead Add Form	1.41
What is your current occupation_Working Profes...	1.37
Do Not Email	1.15
Last Activity_Others	1.14
What is your current occupation_Student	1.07
What is your current occupation_Others	1.02

Confusion Matrix(Training dataset)

Confusion Matrix gives insight into the predictions. It goes deeper than classification accuracy by showing the correct and incorrect predictions on each class





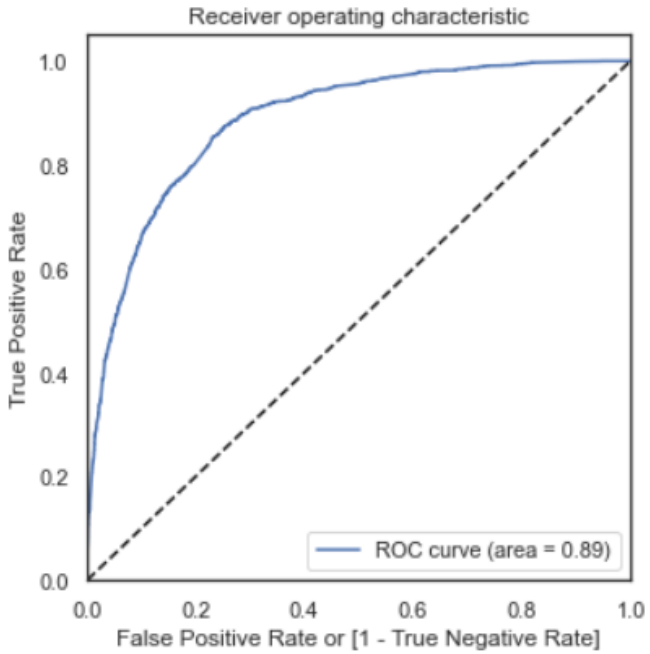
ROC Curve

ROC curve demonstrates several things:

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

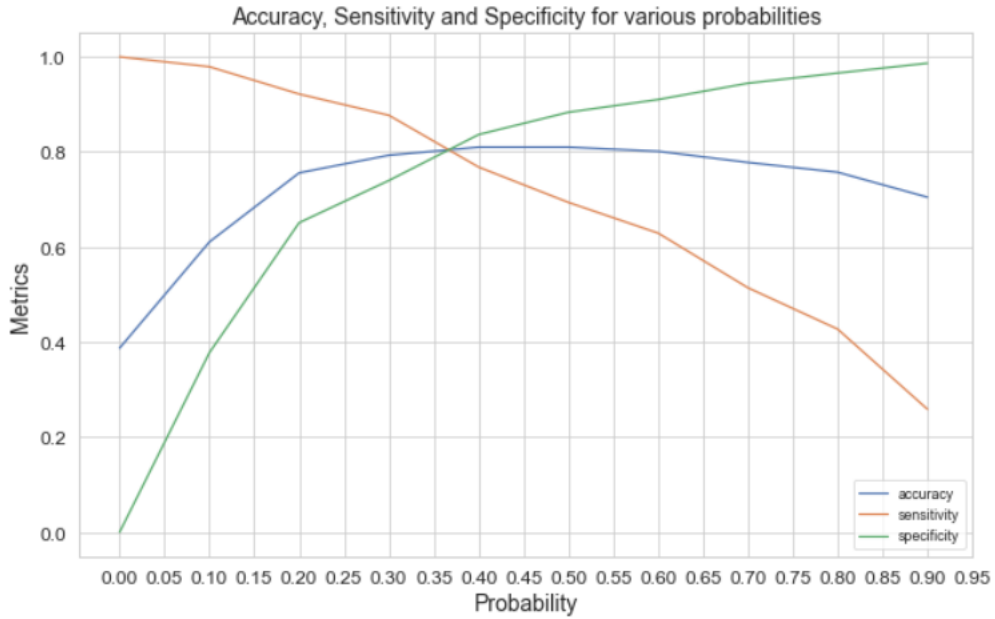
Observation:

ROC Curve area is coming as 0.89 which is a good score

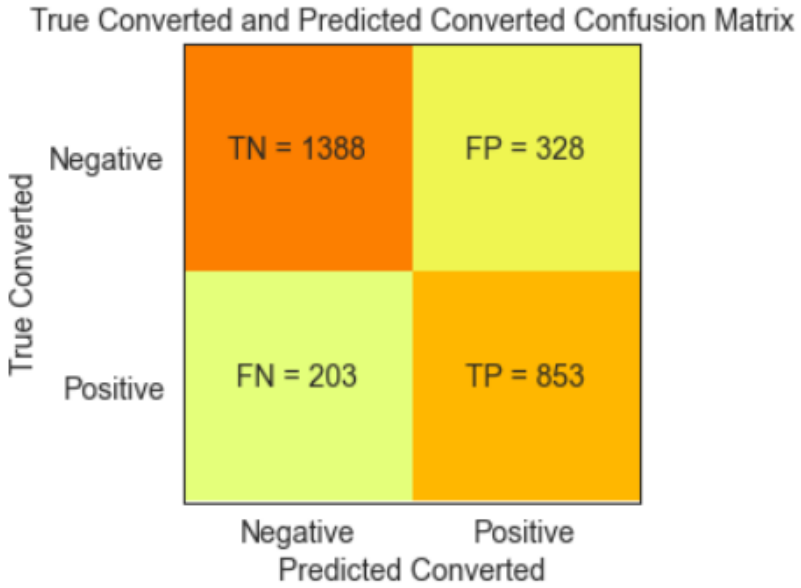


Accuracy, Sensitivity and Specificity for various probabilities

Observation: From the curve above, 0.36 can be taken as the optimum point to take it as a cutoff probability



Confusion Matrix (on Test data at cut off 0.36)

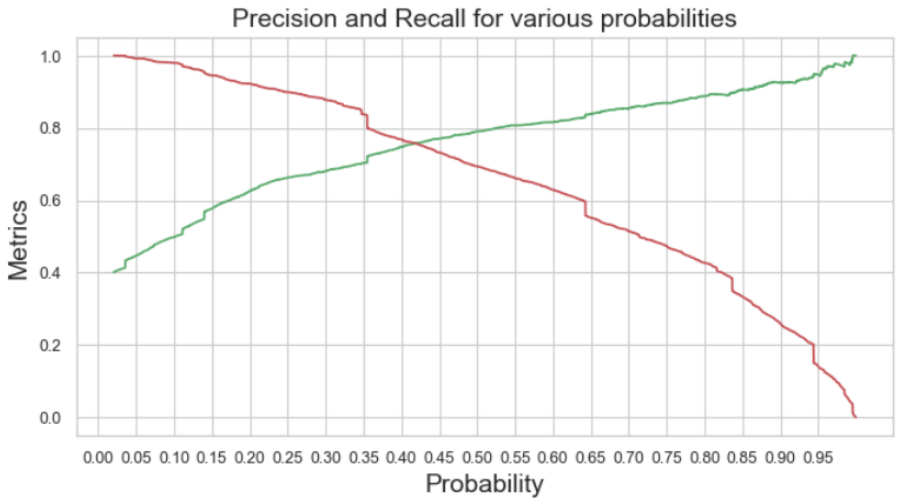


Precision and Recall

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



	precision	recall	f1-score	support
0	0.87	0.81	0.84	1716
1	0.72	0.81	0.76	1056
accuracy			0.81	2772
macro avg	0.80	0.81	0.80	2772
weighted avg	0.82	0.81	0.81	2772

Feature Importance

Observation: 'Total time Spent on Website' has highest Feature Importance and making this variable having high likelihood for converting into a lead and we also see the negative relative importance for 'Do not Email' having least likelihood of converting into a lead

Final Observations and Recommendations

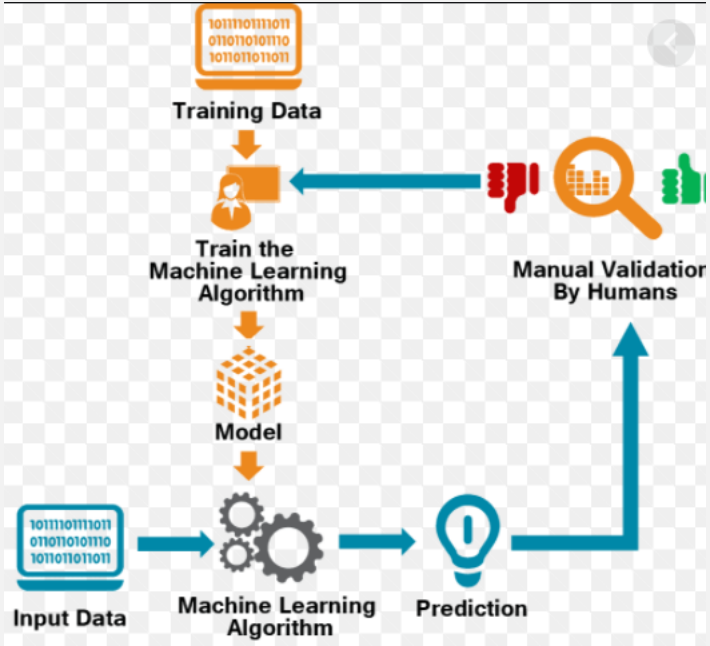
The Final Evaluation Metrics for the Train Dataset:

Accuracy: 0.80
Sensitivity: 0.80
Specificity: 0.81
Precision: 0.73
Recall: 0.81
f1 score : 0.76

The Final Evaluation Metrics for the Test Dataset:

Accuracy : 0.81
Sensitivity: 0.81
Specificity: 0.81
Precision: 0.72
Recall: 0.81
f1 score : 0.76

Final Observations and Recommendations



X-Education has a better chance of converting a potential lead when:

The total time spent on the Website is high:

Leads who have spent more time on the website have converted

Current Occupation is specified:

Leads who are working professionals have high chances of getting converted. People who were looking for better prospects like Unemployed, students, Housewives and Business professionals were also good prospects to focus on.

When the Lead origin was Lead Add form

Leads who have responded/ or engaged through Lead Add Forms have had a higher chances of getting converted

Number of Total Visits were high

Leads who have made a greater number of visits have higher chances of getting converted.

When the last activity was SMS sent or Email opened

Members who have sent an SMS for enquiry or who have opened the email have a higher chance of getting converted.

Thank you

Bhavnish Marwaha
Piyushi Prenam

Post Graduate Diploma - Data Science
International Institute of Information Technology Bangalore - Upgrad