Leads Scoring Case Study

Ramachandran Ramamoorthy, Sarath Chandran

Business Understanding

Business Understanding:

An EdTech company markets its courses on several websites and search engines like Google. Visitors browse through the courses in various websites. However, few of them convert into paying customers. The CEO of the company wants to predict based on the various attributes which customers are most likely to "Convert" into a paying customer. The CEO of the company has given a target of 80% as per the case study

Business Objective of the X education is 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

Approach:

There are quite a few goals for this case study:

- Understand the data and the characteristics/attributes of visitors and leads
- Build a logistic regression model to assign a score in scale 0 to 100 to each of the leads
- Identify potential leads
- Higher score would mean that the lead is hot, i.e. is most likely to convert
- Lower score would mean that the lead is cold and will mostly not get converted

Understanding of the data

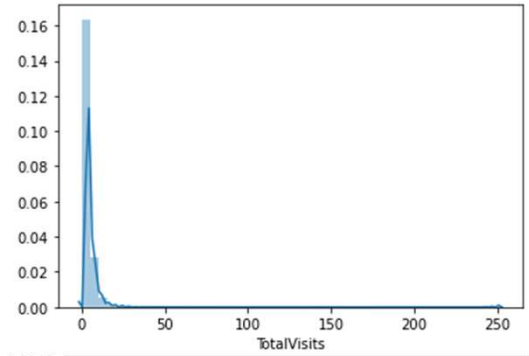
Leads Data

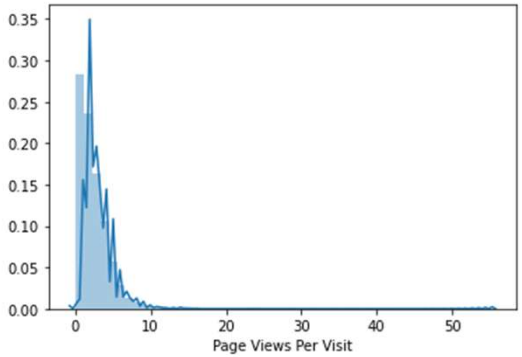
- This data comprises of the details of the leads visiting the website and their attributes
- It has 9240 rows and 37 Columns

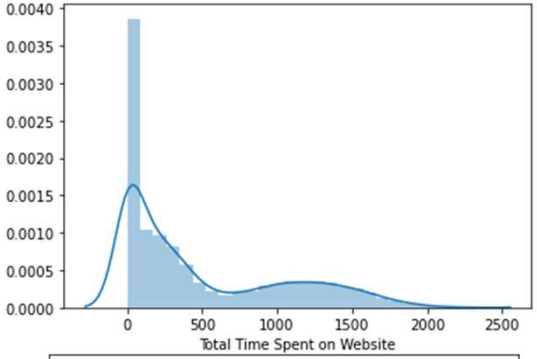
The various columns in the data are:

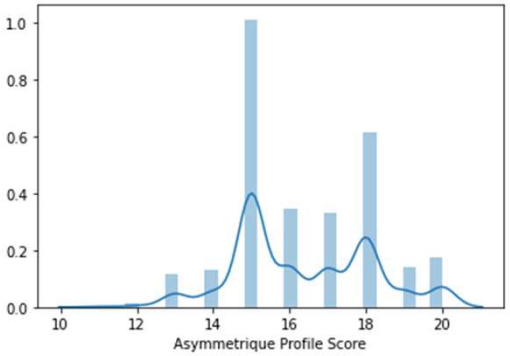
- Prospect ID
- Lead Number
- Lead Origin
- Lead Source
- Do Not Email
- Do Not Call
- Converted
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Last Activity
- Country
- Specialization
- How did you hear about X Education
- What is your current occupation
- What matters most to you in choosing this course
- Search
- Magazine
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations
- Receive More Updates About Our Courses
- Tags
- Lead Quality
- Update me on Supply Chain Content
- Get updates on DM Content
- Lead Profile
- City
- Asymmetrique Activity Index
- Asymmetrique Profile Index
- Asymmetrique Activity Score
- Asymmetrique Profile Score
- I agree to pay the amount through cheque
- a free copy of Mastering The Interview
- Last Notable Activity

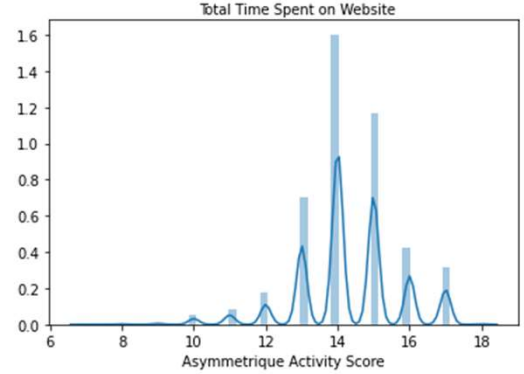
Data Cleaning and EDA









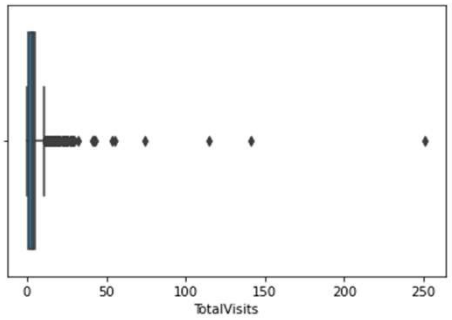


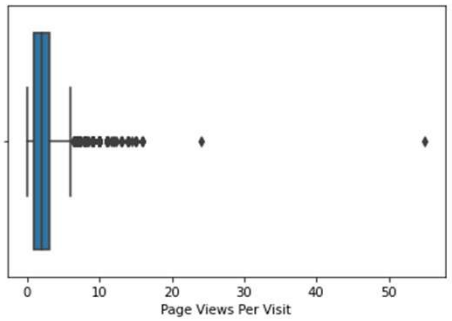
- Numeric and continuous variables are checked for “Normality”

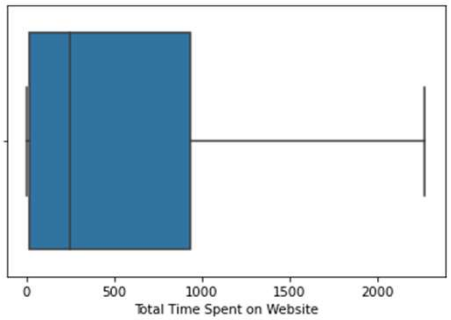
- Total Visits
- Page Views Per Visit
- Total Time Spent on Website
- Asymmetrique per Score
- Asymmetrique Activity Score

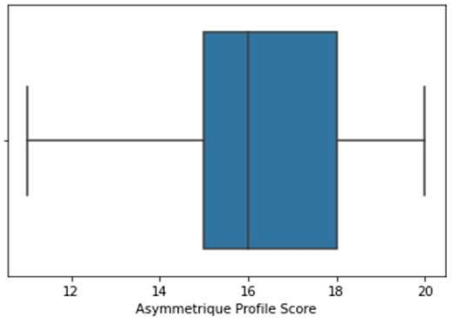
The above are NOT Normally Distributed

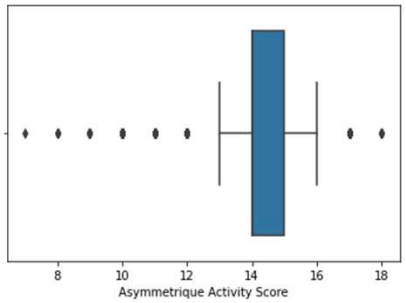
Data Analysis | Categorical Univariate Analysis











- Numeric and continuous variables are checked for "Outliers
- **Total Visits**
- **Page Views Per Visit**
- Asymmetrique Activity Score
- **Total Time Spent on Website**
- Asymmetrique Profile Score

The above are NOT Normally Distributed

Outlier Treatment is done for numeric variables with Median

Data Analysis | Missing Value Analysis

Variable Type	Total	Percent Missing	Unique Data Points	Datatype
Lead Quality	4767	51.590909	5	object
Asymmetrique Activity Index	4218	45.649351	3	object
Asymmetrique Profile Score	4218	45.649351	10	float64
Asymmetrique Activity Score	4218	45.649351	12	float64
Asymmetrique Profile Index	4218	45.649351	3	object
Tags	3353	36.287879	26	object
Lead Profile	2709	29.318182	6	object
What matters most to you in choosing a course	2709	29.318182	3	object
What is your current occupation	2690	29.112554	6	object
Country	2461	26.634199	38	object
How did you hear about X Education	2207	23.885281	10	object
Specialization	1438	15.562771	19	object
City	1420	15.367965	7	object
Page Views Per Visit	137	1.482684	114	float64
TotalVisits	137	1.482684	41	float64
Last Activity	103	1.114719	17	object
Lead Source	36	0.389610	21	object

- The Missing Value Analysis is shown aside
- During the next steps of the analysis all variables with 45% or above missing values are dropped
- The next slide displays values which have either 0 Missing Values or have 1 Unique value

Data Analysis | Missing Value Analysis

Variable Type	Total	Percent Missing	Unique Data Points	Datatype
Receive More Updates About Our Courses	0	0.000000	1	object
I agree to pay the amount through cheque	0	0.000000	1	object
Get updates on DM Content	0	0.000000	1	object
Update me on Supply Chain Content	0	0.000000	1	object
A free copy of Mastering The Interview	0	0.000000	2	object
Prospect ID	0	0.000000	9240	object
Newspaper Article	0	0.000000	2	object
Through Recommendations	0	0.000000	2	object
Digital Advertisement	0	0.000000	2	object
Newspaper	0	0.000000	2	object
X Education Forums	0	0.000000	2	object
Lead Number	0	0.000000	9240	int64
Magazine	0	0.000000	1	object
Search	0	0.000000	2	object
Total Time Spent on Website	0	0.000000	1731	int64
Converted	0	0.000000	2	int64
Do Not Call	0	0.000000	2	object
Do Not Email	0	0.000000	2	object
Lead Origin	0	0.000000	5	object
Last Notable Activity	0	0.000000	16	objec

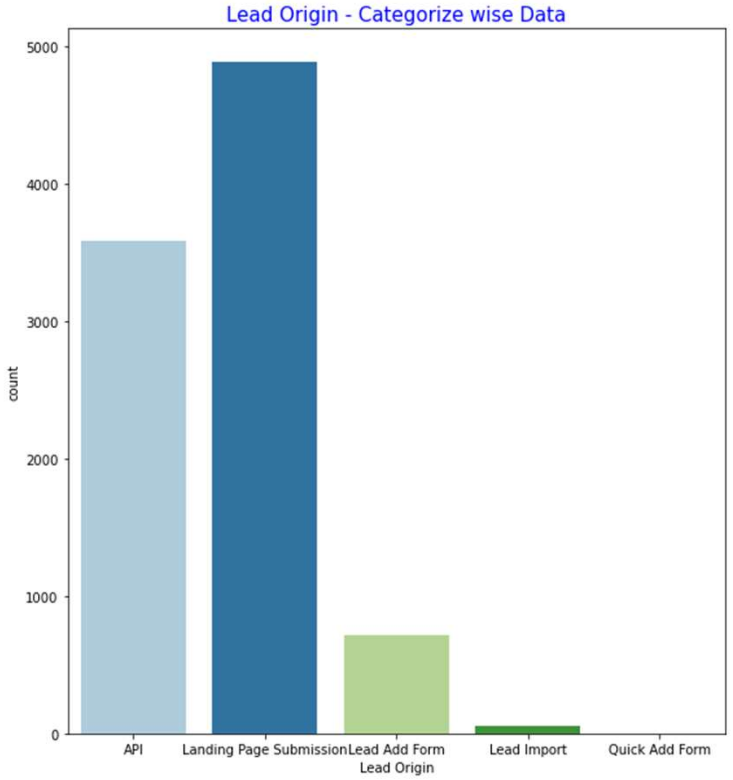
- The Values which have just one Unique Value may not add a lot of value to the model
- Also values depicting unique IDs will not add value to the model being developed

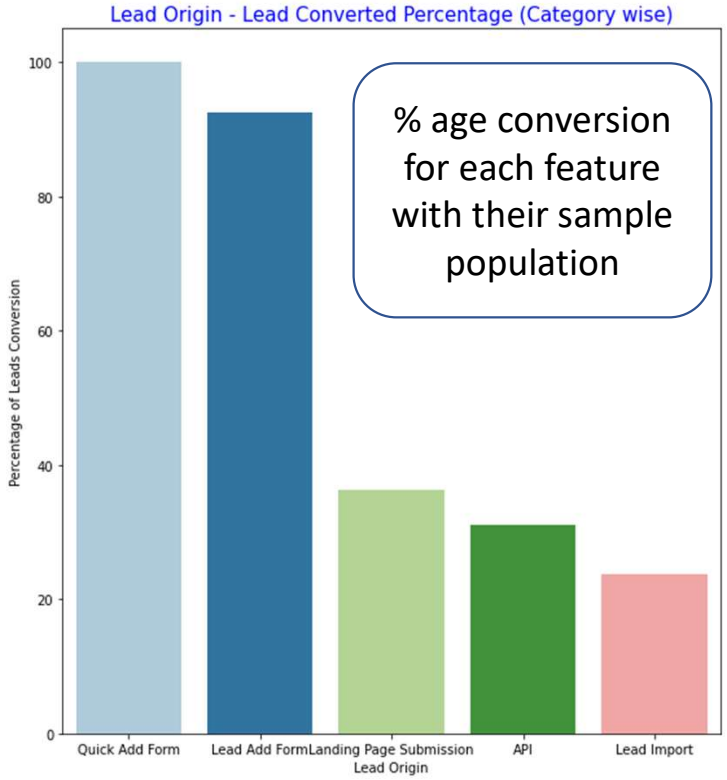
- Let's drop the columns having >45% null values as imputing will lead to skewness

- Columns like the Prospect ID, Lead Number, Tags, Lead Profile etc to be dropped as they would not add value in model building

- *Dropping the Variables:*
  - *'Prospect ID'*
  - *'Lead Number'*
  - *'Lead Profile'*
  - *'Lead Quality'*
  - *'Asymmetrique Profile Score'*
  - *'Asymmetrique Activity Score'*
  - *'Asymmetrique Activity Index'*
  - *'Asymmetrique Profile Index'*
  - *'Tags'*

Data Analysis | Numerical Univariate Analysis



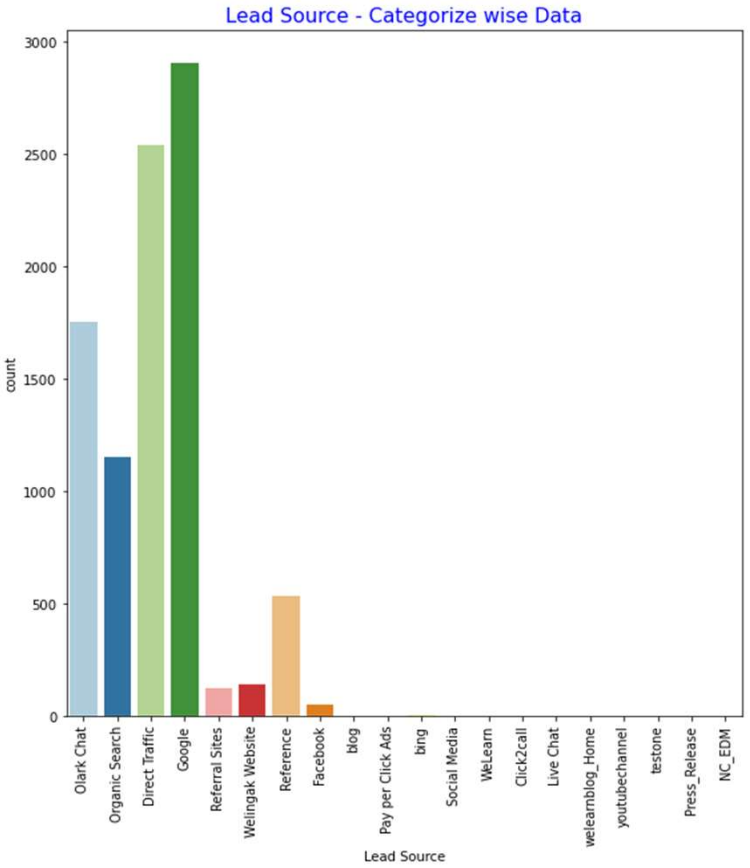


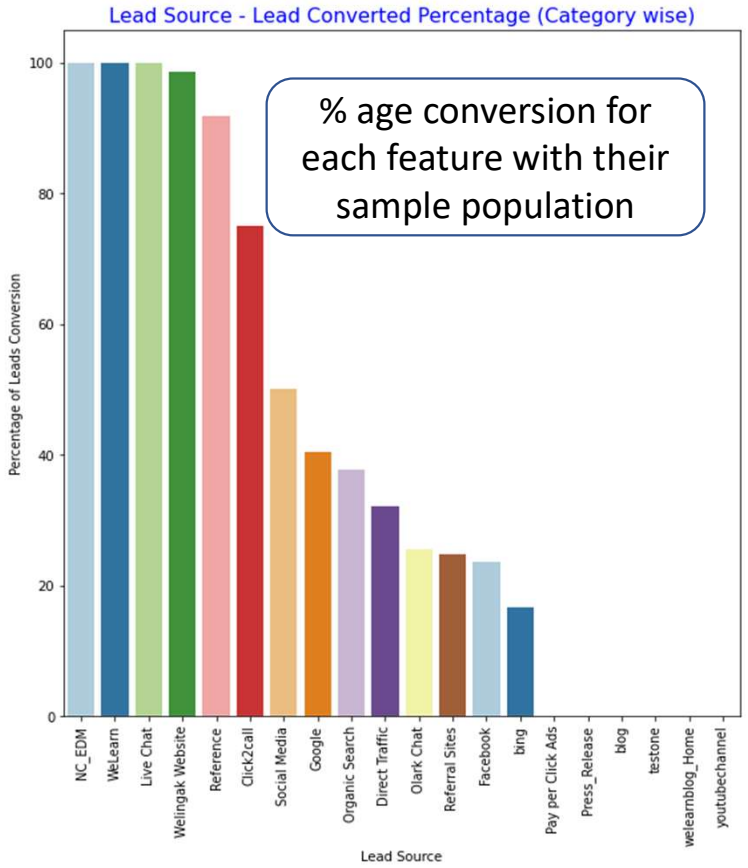
Insights

1. Lead Generation is maximum at the Landing Page Submission, followed by API and Lead Add Form
2. Sources where the conversion rate seems to be good is in the Order of Quick Add Form, Lead Add Form and Landing Page Submission
3. Lead Import has the lowest lead conversion deeming it as least effective

Data Analysis | Numerical Univariate Analysis

Lead Source

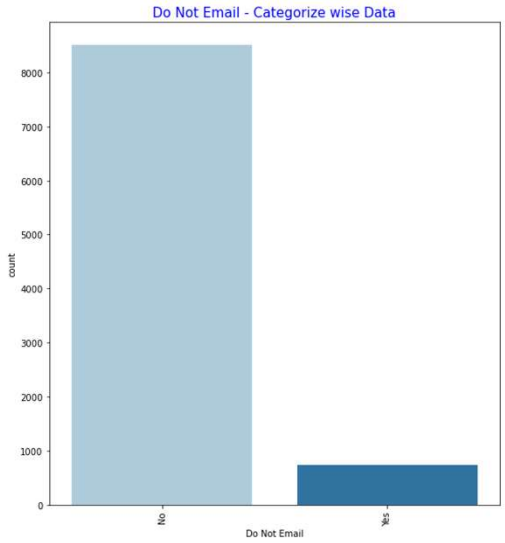


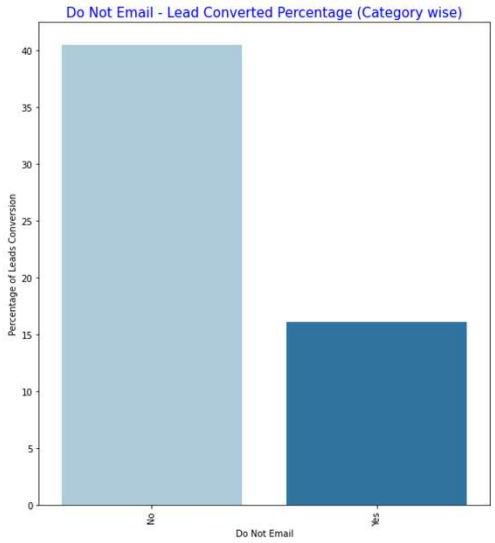


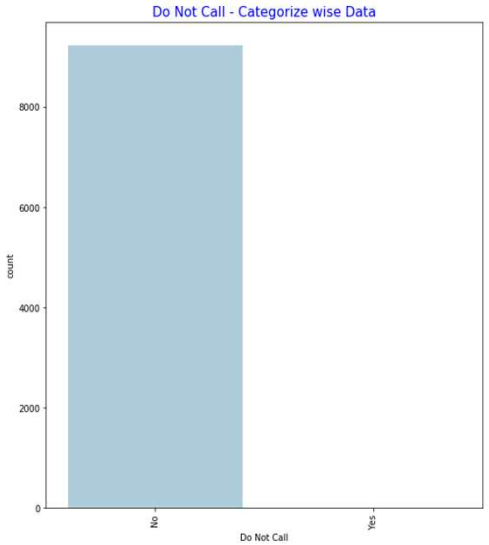
Insights

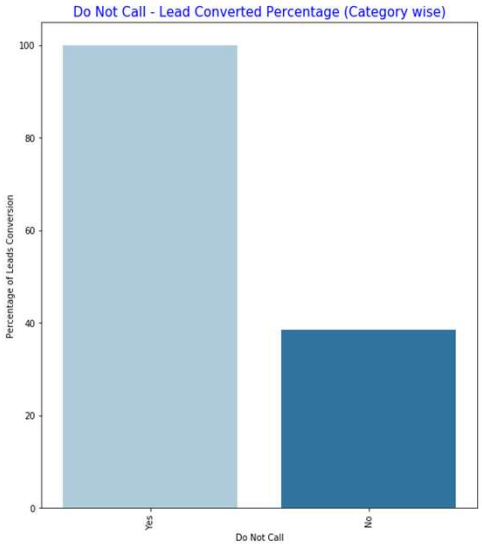
1. Google is the best Source of Lead Generation, followed by Direct Traffic, Olark Chat
2. For Lead Conversion , Google has a good conversion rate in comparison to the volume.
3. Other lead sources in the order of conversion effectiveness are Direct Traffic, Organic Search and Olark Chat

Data Analysis | Numerical Univariate Analysis







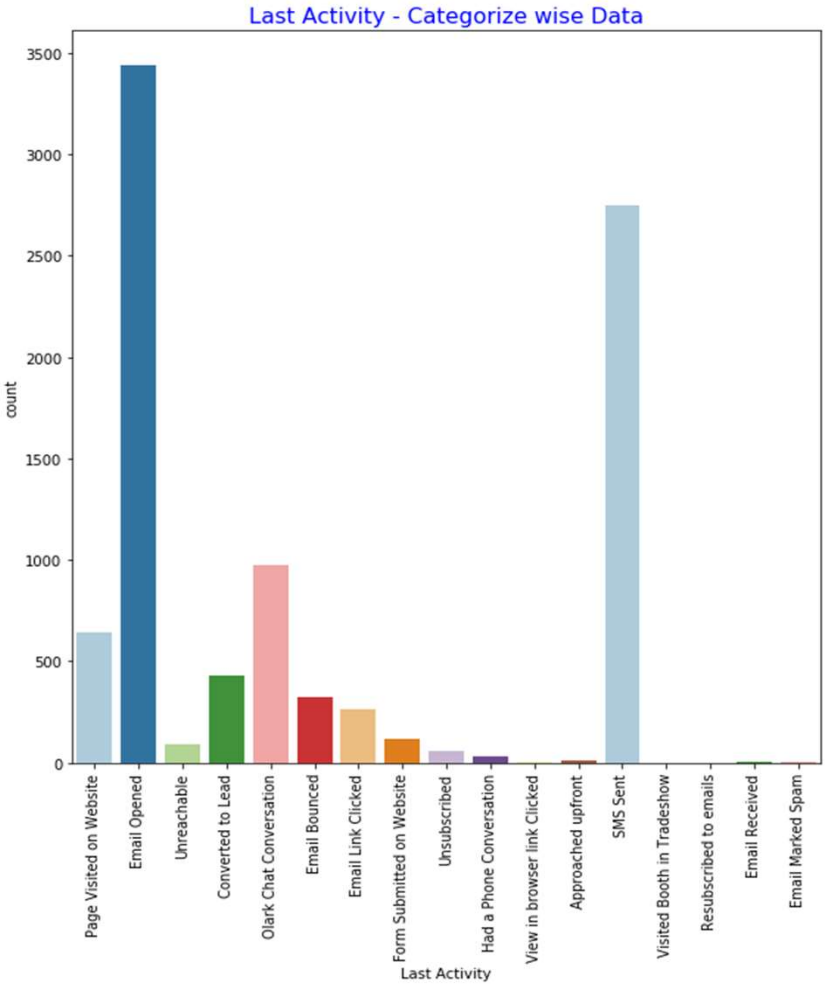


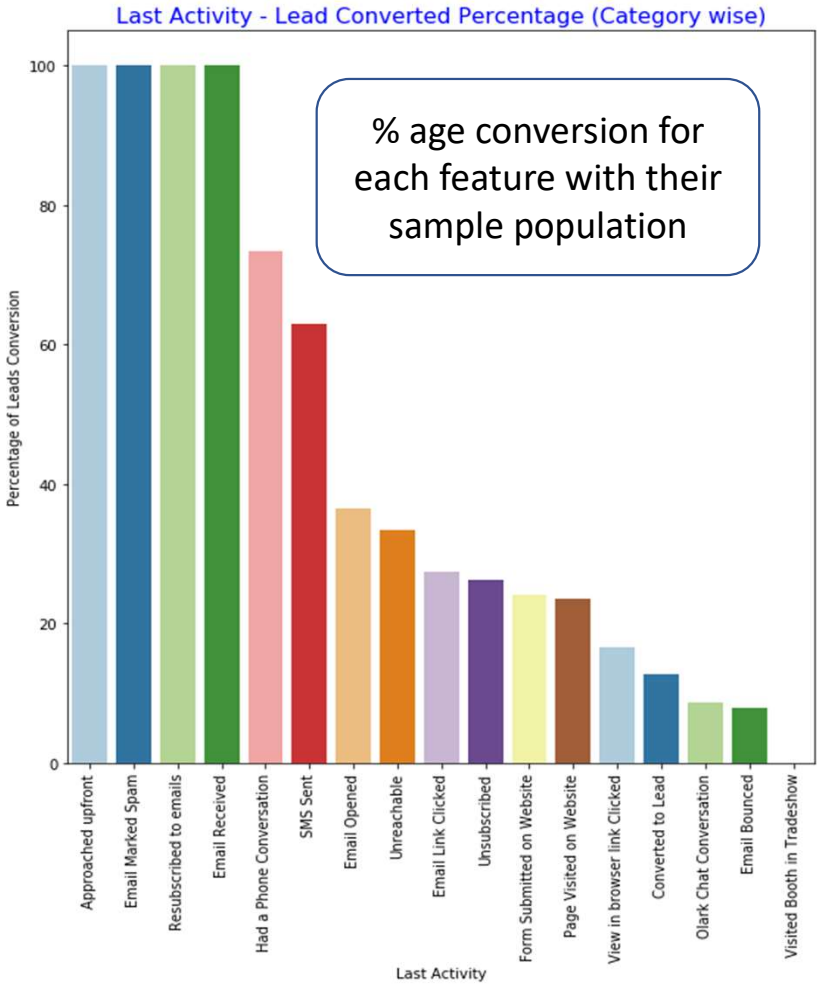
Insights

Leads who want to receive emails are high with a lead conversion %as close to 40%
Leads who do not want to receive emails, are low in number but within their population the conversion % is close to 15%

No of leads have shown interest in getting call updates is high. However, such leads conversion rate is at 40% and leads who are not ok to receive calls is very low; but the conversion rate is 100%. As the data is not balanced, this feature shall be treated insignificant for Model building.

Last Activity





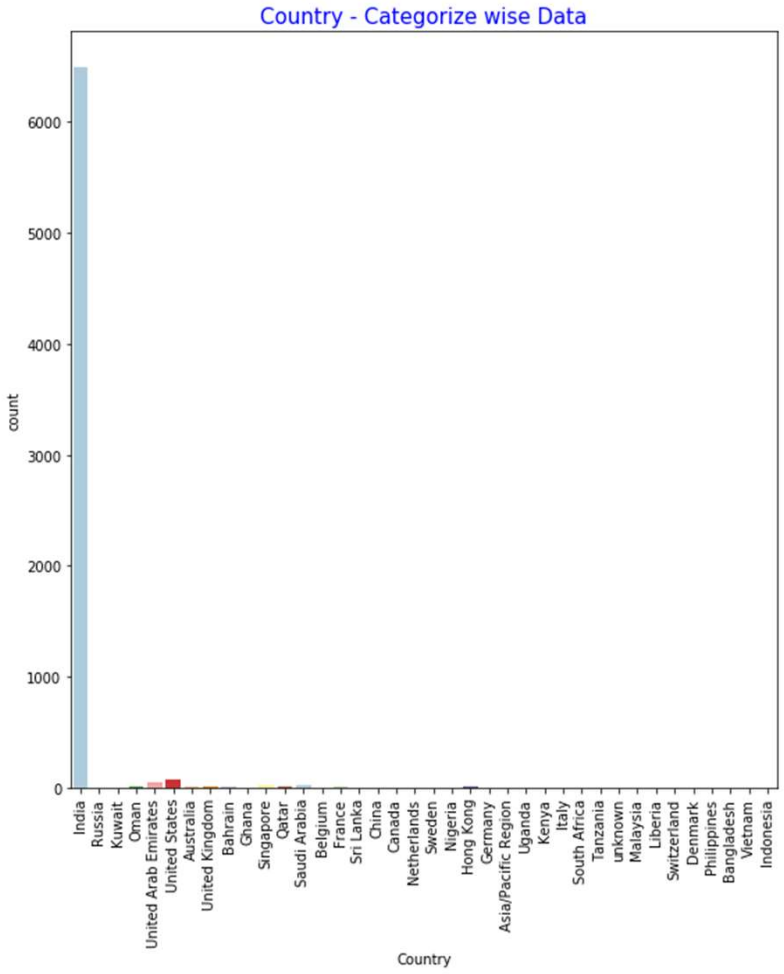
% age conversion for each feature with their sample population

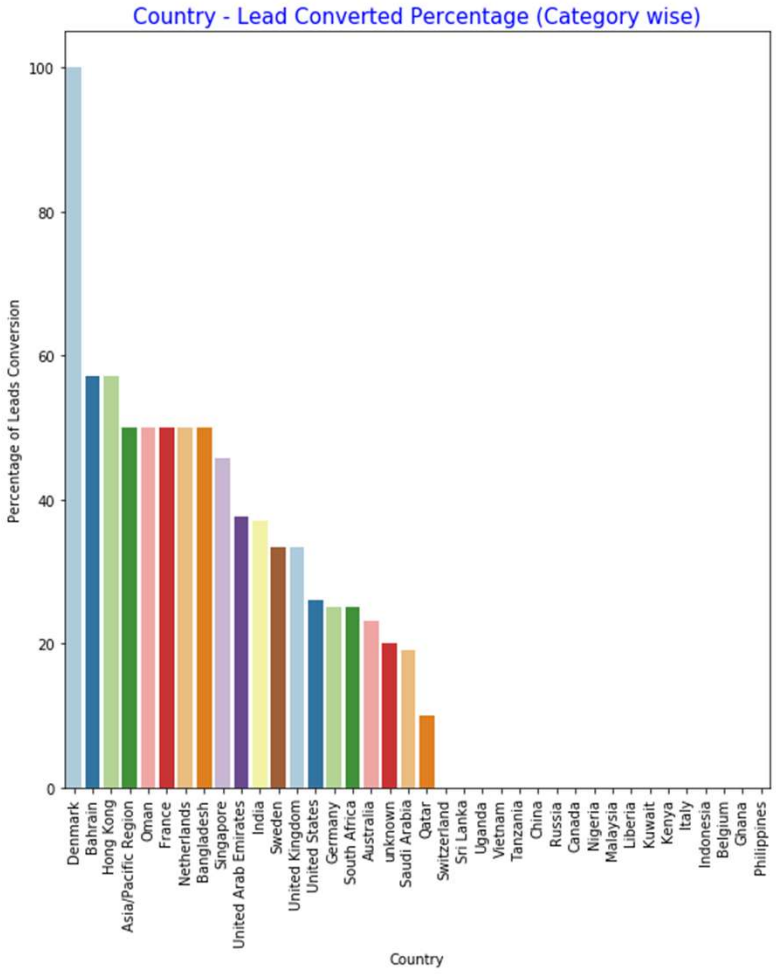
Insight

High number of leads who have their Email Opened, followed by SMS Sent

Leads conversion with SMS sent seem to be higher (60%) compared to the leads with last activity as Email Opened

Data Analysis | Numerical Univariate Analysis



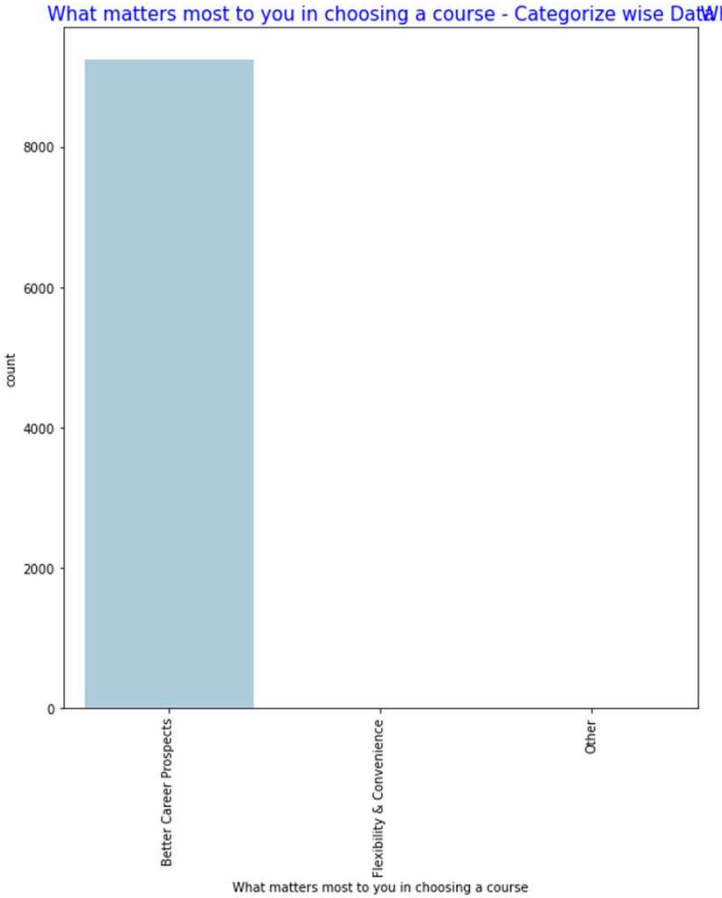


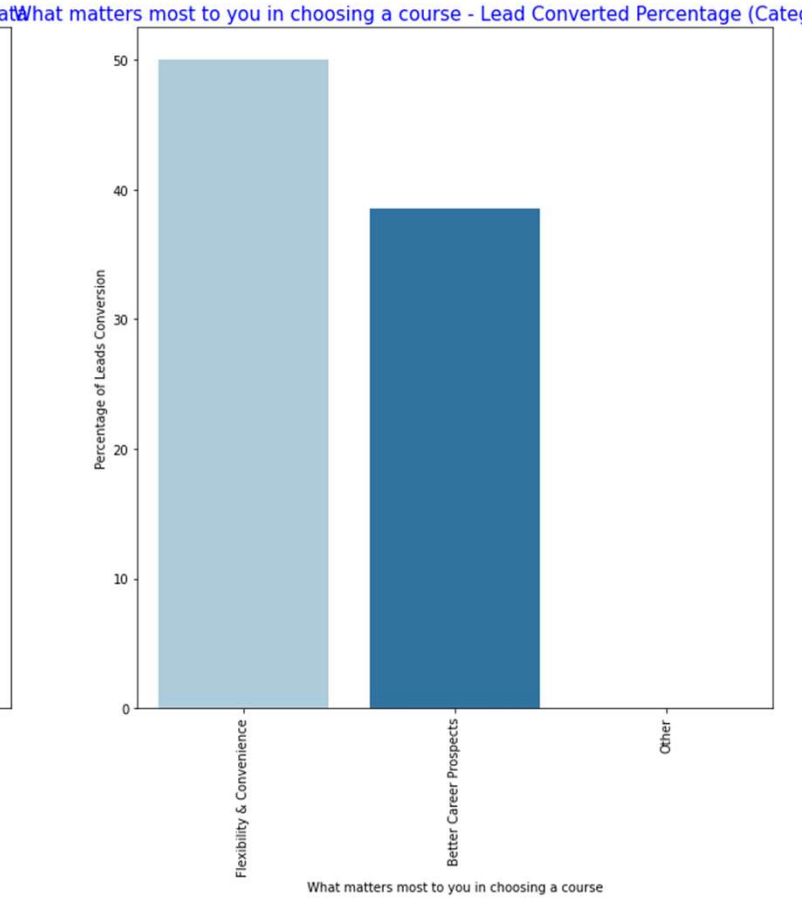
Insight

Scarce data points for Country and Count except India

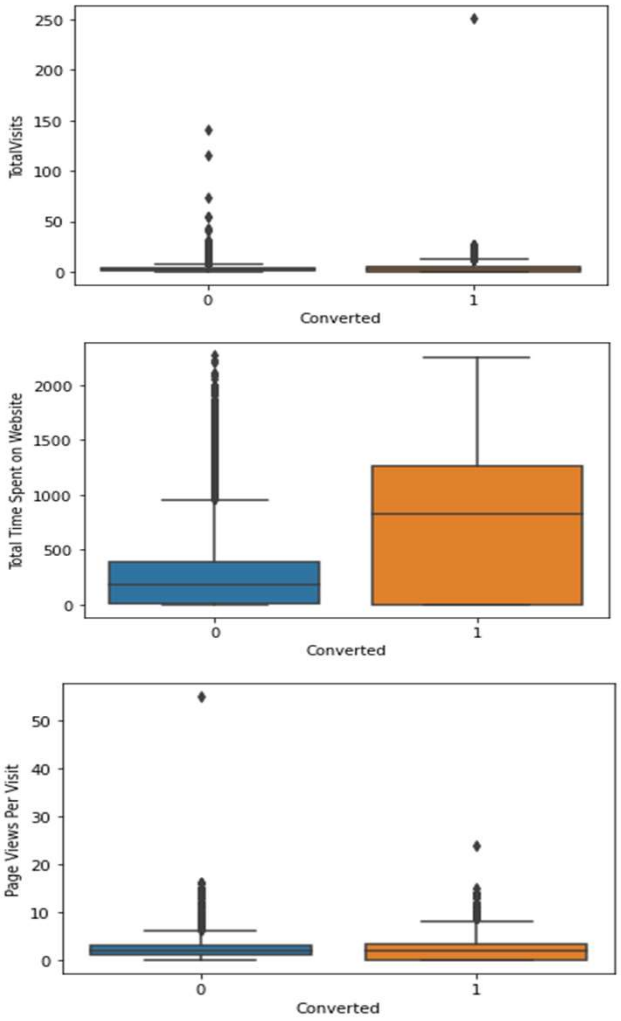
And within India, lead conversion = 40%

Data Analysis | Numerical Univariate Analysis

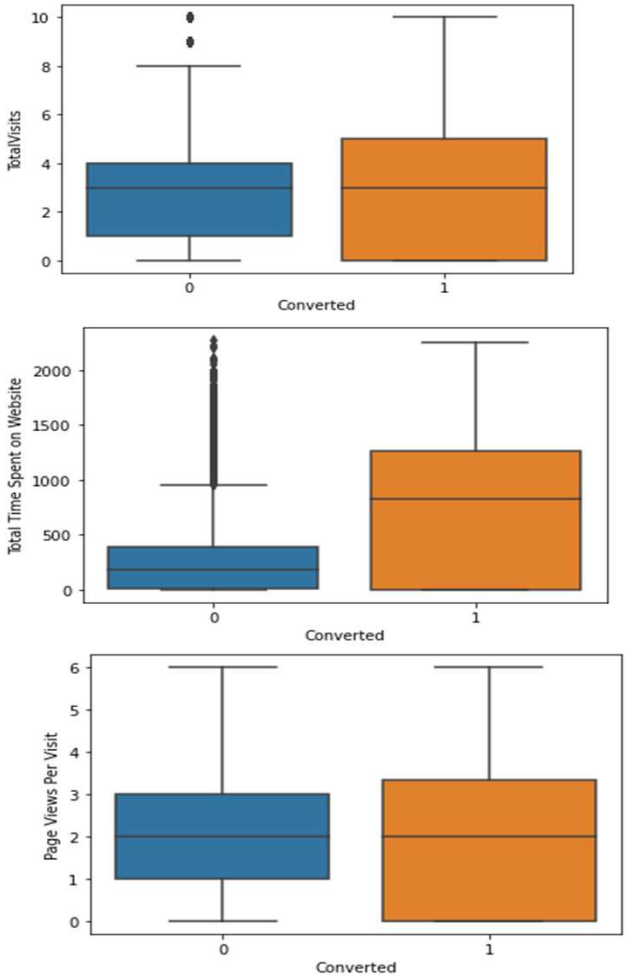




Data Analysis | Outlier Treatment



Post Capping The >95% Percentile



Outlier Capping

To have the least impact in Model Building, Capping of outlier > 95 percentile

Dummy Creation for Categorical Variables | Correlation Analysis

Variables considered for dummy variable creation and later dropped during model building

1. Lead Origin
2. Lead Source
3. Last Activity
4. Specialization
5. What is your current occupation
6. What matters most to you in choosing a course

VAR1	VAR2	Positive Correlation
Lead Source_Facebook	Lead Origin_Lead Import	0.981709
Lead Source_Reference	Lead Origin_Lead Add Form	0.853237
Page Views Per Visit	TotalVisits	0.767585
Last Activity_Email Bounced	Do Not Email	0.618470
Lead Origin_Landing Page Submission	Page Views Per Visit	0.553423
Lead Source_Direct Traffic	Lead Origin_Landing Page Submission	0.528303
Lead Origin_Landing Page Submission	TotalVisits	0.453501
Lead Source_Welingak Website	Lead Origin_Lead Add Form	0.430407
Specialization_Others	Lead Source_Olark Chat	0.429177
Last Activity_Olark Chat Conversation	Lead Source_Olark Chat	0.426248

VAR2	VAR2	Negative Correlation
What is your current occupation_Working Profes...	What is your current occupation_Unemployed	-0.849653
Lead Source_Olark Chat	Page Views Per Visit	-0.573334
Lead Source_Olark Chat	Lead Origin_Landing Page Submission	-0.512950
Specialization_Select	Lead Origin_Landing Page Submission	-0.507078
Last Activity_SMS Sent	Last Activity_Email Opened	-0.500317
Lead Source_Olark Chat	TotalVisits	-0.500094
What is your current occupation_Unemployed	What is your current occupation_Student	-0.450486
Specialization_Others	Lead Origin_Landing Page Submission	-0.424285
Lead Source_Google	Lead Source_Direct Traffic	-0.417704
Lead Source_Olark Chat	Total Time Spent on Website	-0.376768

This will help tune the model in the next steps

Test Train Split

```
Shape of X_train is : (6468, 70)
Shape of y_train is : (6468,)
Shape of X_test is : (2772, 70)
Shape of y_test is : (2772,)
```

Model Building Process| Feature Scaling | RFE

1. Splitting the Data into Training and Testing Sets
2. The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
3. Use RFE for Feature Selection
    - *Total Time Spent on Website*
    - *Page Views Per Visit*
    - *Total Visits*
4. Running RFE with 15 variables as output
5. Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
    - *High P and High VIF – Remove*
    - *High P and Low VIF – Remove*
    - *High VIF and Low P – Remove*
    - *Low VIF and Low P - Keep*
6. Predictions on test data set
7. Overall accuracy 79.84%

Model Building and VIF

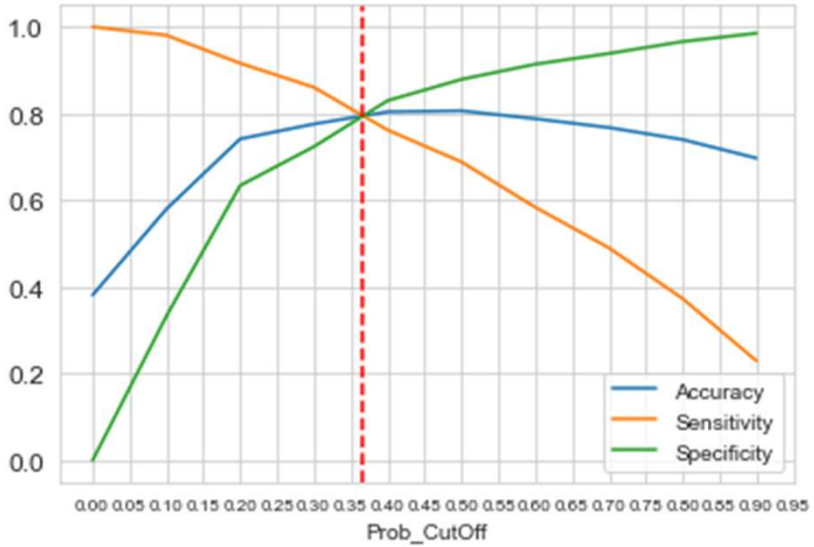
1. Splitting the Data into Training and Testing Sets
2. The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
3. Use RFE for Feature Selection
   - *Total Time Spent on Website*
   - *Page Views Per Visit*
   - *Total Visits*
4. Running RFE with 15 variables as output
5. Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
   - *High P and High VIF – Remove*
   - *High P and Low VIF – Remove*
   - *High VIF and Low P – Remove*
   - *Low VIF and Low P - Keep*
6. Predictions on test data set
7. Overall accuracy 79.84%

After 12th Iteration of model building, VIF Values

Features	VIF
Lead Origin_Landing Page Submission	2.63
Lead Source_Olark Chat	2.07
Lead Source_Direct Traffic	1.92
Do Not Email	1.89
Last Activity_Email Bounced	1.85
Specialization_Others	1.74
Last Activity_SMS Sent	1.74
Specialization_Select	1.57
Last Activity_Olark Chat Conversation	1.54
Lead Origin_Lead Add Form	1.54
Total Time Spent on Website	1.26
Lead Source_Welingak Website	1.26
Last Activity_Page Visited on Website	1.20
Last Activity_Others	1.17
Last Activity_Converted to Lead	1.12
Last Activity_Email Link Clicked	1.07
Last Activity_Form Submitted on Website	1.03
Specialization_Hospitality Management	1.02
Last Activity_Had a Phone Conversation	1.01

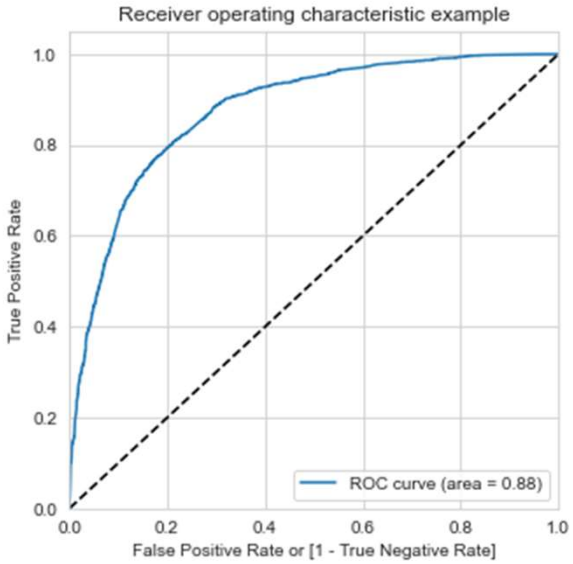
Accuracy , Specificity and Sensitivity | ROC Curve | Top Predictors

Ideal cut-off points = 0.365



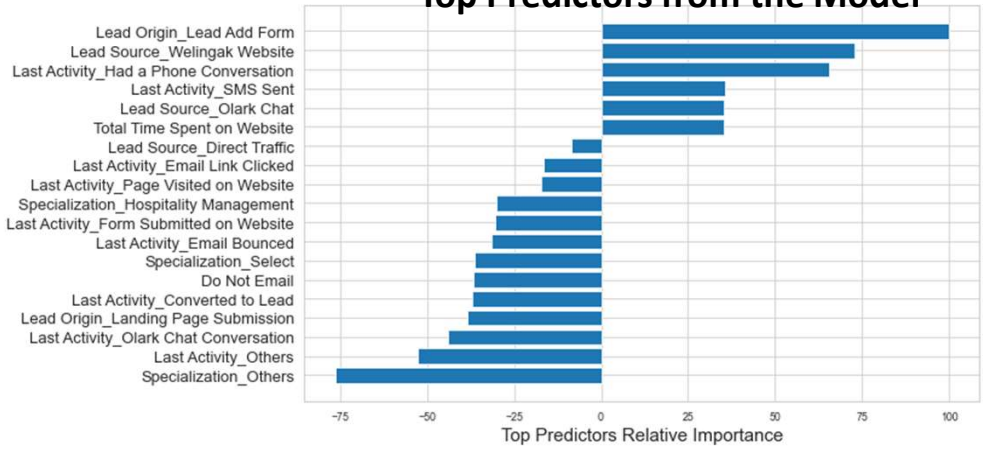
Confusion Matrix

Model Accuracy value is	: 79.84%
Model Specificity value is	: 80.31 %
Model Sensitivity value is	: 79.08 %
Model Precision value is	: 71.22 %
Model Recall value is	: 79.08 %
Model True Positive Rate (TPR)	: 79.08 %
Model False Positive Rate (FPR)	: 19.69 %
Model Poitive Prediction Value is	: 71.22 %
Model Negative Prediction value is	: 86.17 %



Receiver Operating Curve

Top Predictors from the Model



End