Lead Scoring Case Study

Case study group: Thilagavathy Ragunath, Anil.S

Problem Statement

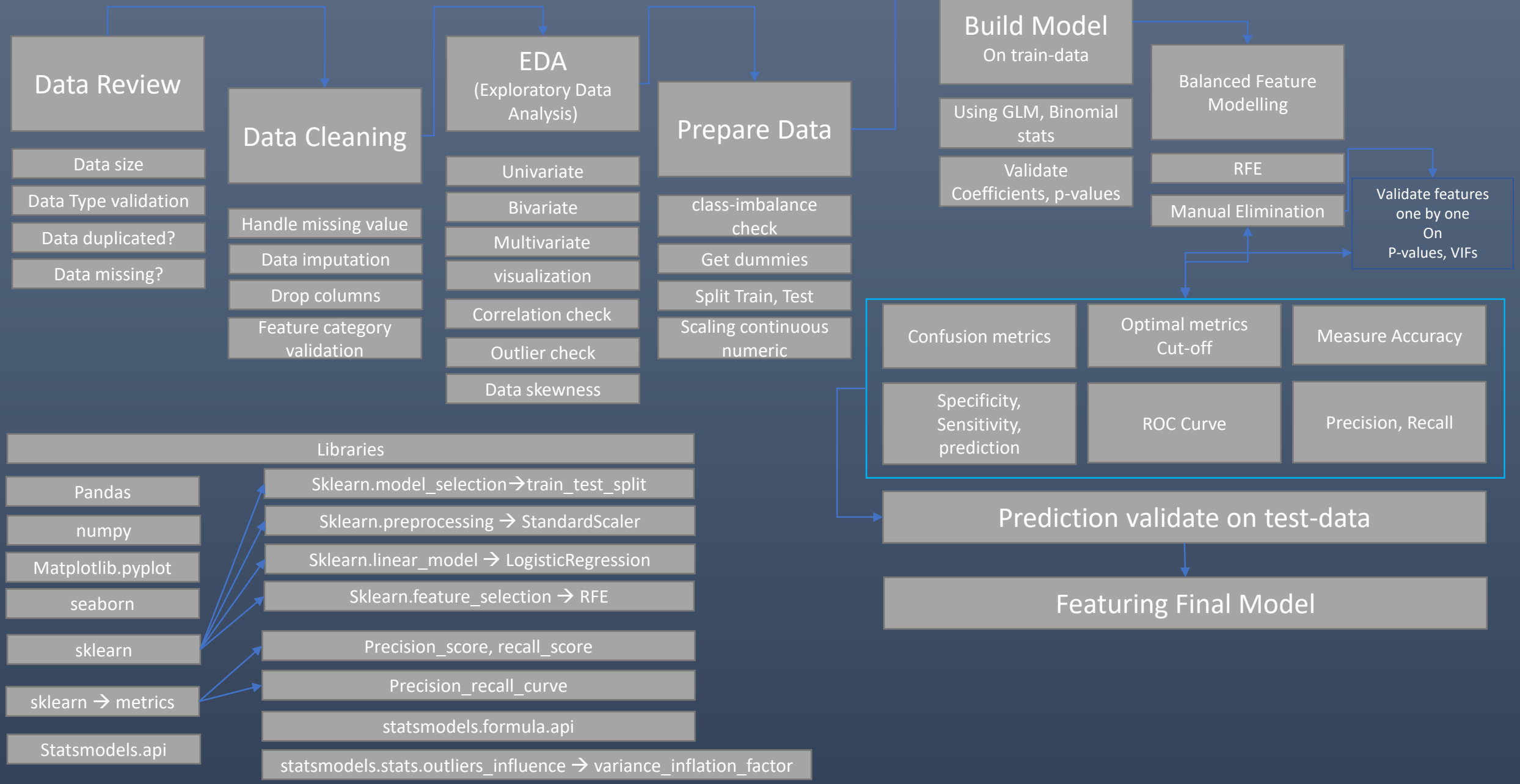
“X Education” sells online courses to industry professionals.
This company markets the courses on several platforms. Customers are identified as leads when they fill-up X Education forms providing their email and phone numbers. Once the customer details are acquired, sales team start their action in the process of converting the leads. However the conversion rate is achieved only 30%

X Education wants to identify the potential customers addressed as “Hot Leads” in order to achieve the better conversion rate. The company requires a model which assign a lead score to each leads, identifying high-lead score will have higher conversion chances(“Hot Leads”), the lower-lead score will have lower conversion chances(“Cold Leads”).

Create a Logistic Regression model for the given 9000+ data records and predict the target variable “Converted” which describes if the past lead was converted to Hot Leads or not.

This data has provided with many features such as Lead Profile, occupation, time spent on website & pages viewed, last activities, Ads sources, source of contact

Model Process Flow



Data Handling - missing values

1. Dropping columns(high % of missing)

Lead Quality	51.59%
Asymmetrique Profile Score	45.65%
Asymmetrique Activity Score	45.65%
Asymmetrique Profile Index	45.65%
Asymmetrique Activity Index	45.65%
Tags	36.28%

How did you hear about X Education- 71% of data not selected

2. **Data Imputation: What matters most to you in choosing a course**
Null values are imputed with mode value, “Better Career Prospects”

3. **Data Imputation: Lead Profile**
“Select” is imputed as null values. Dropped the column as 74% of null data

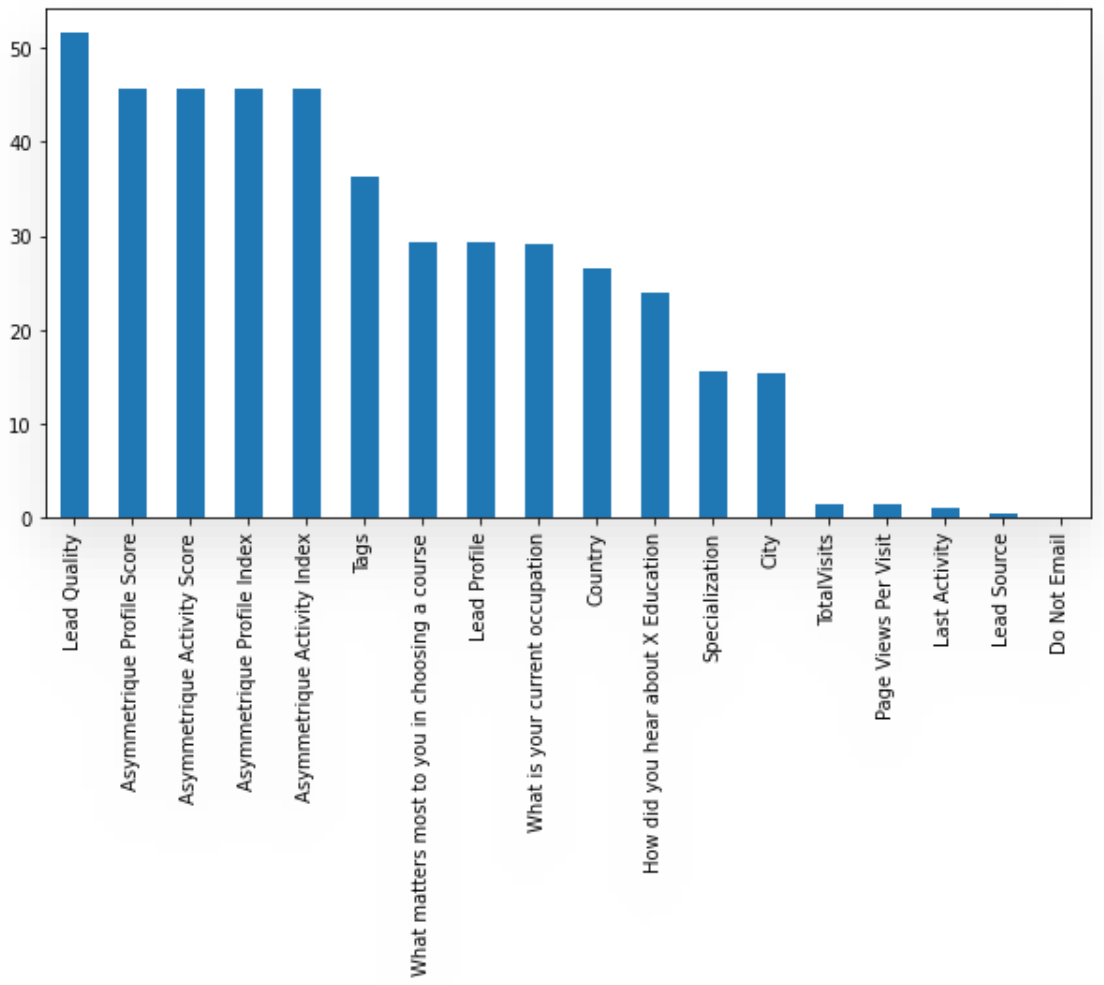
4. **Data Imputation: What is your current occupation**
null is imputed as “Unemployed” as most of the data are in same

5. **Data Imputation: Specialization, Null is imputed as “Unknown”**

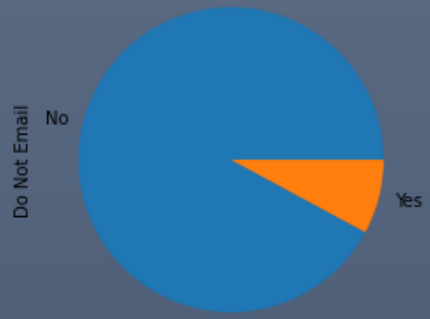
6. **City/Country:**
Cities are in "Thane & Outskirts", "Other Cities of Maharashtra: "City_In_Maharashtra“.
Cities are in "Select", "Other Cities", "Other Metro Cities", "Tier II Cities",
"Other Cities of Maharastra" , updated to "City_In_India"
If City is null but the Country is "India“, updated to "City_In_India"
If City is null or "Select" but the Country is NOT "India", updated to "Others"
If City is null and Country is NOT "India“, This is highly confused situation to either
update City or If Country is null, but the City is "Mumbai" --> update to "India"
If Country is null but the City is "Others“, update to "Other_Countries"

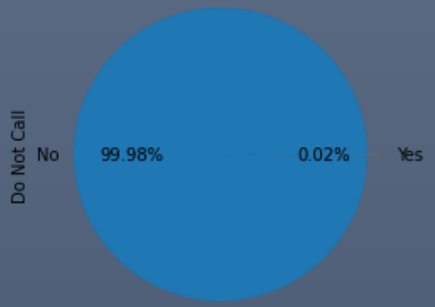
7.**Data Correction: Lead Source**, “google” is corrected as Google

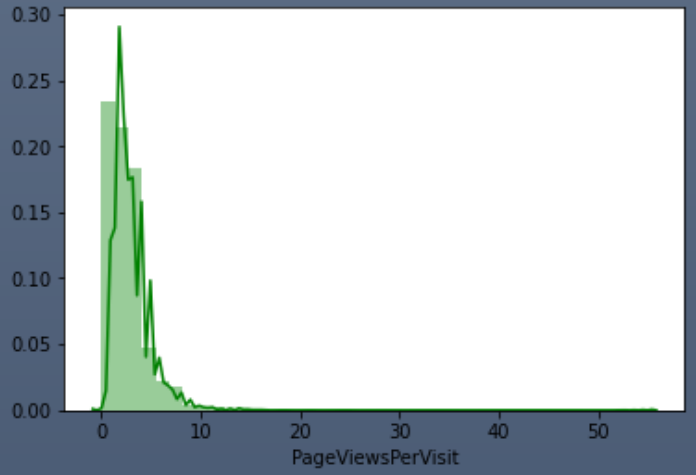
8. **Dropping Rows of missing values in lesser than 2%,**
- TotalVisits
- Page Views Per Visit
- Last Activity

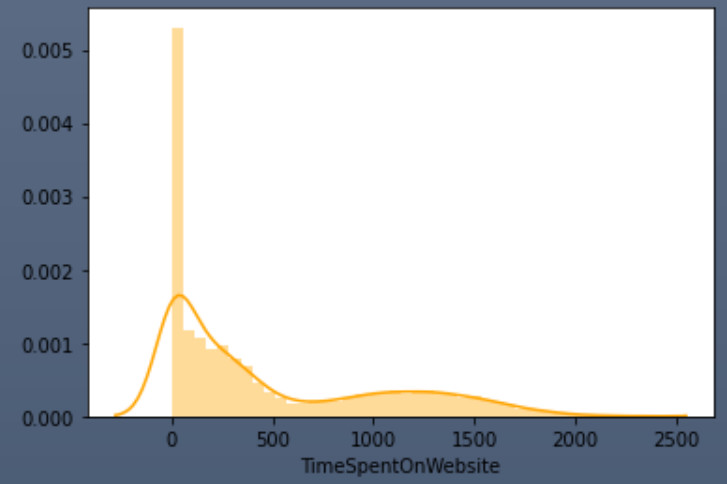


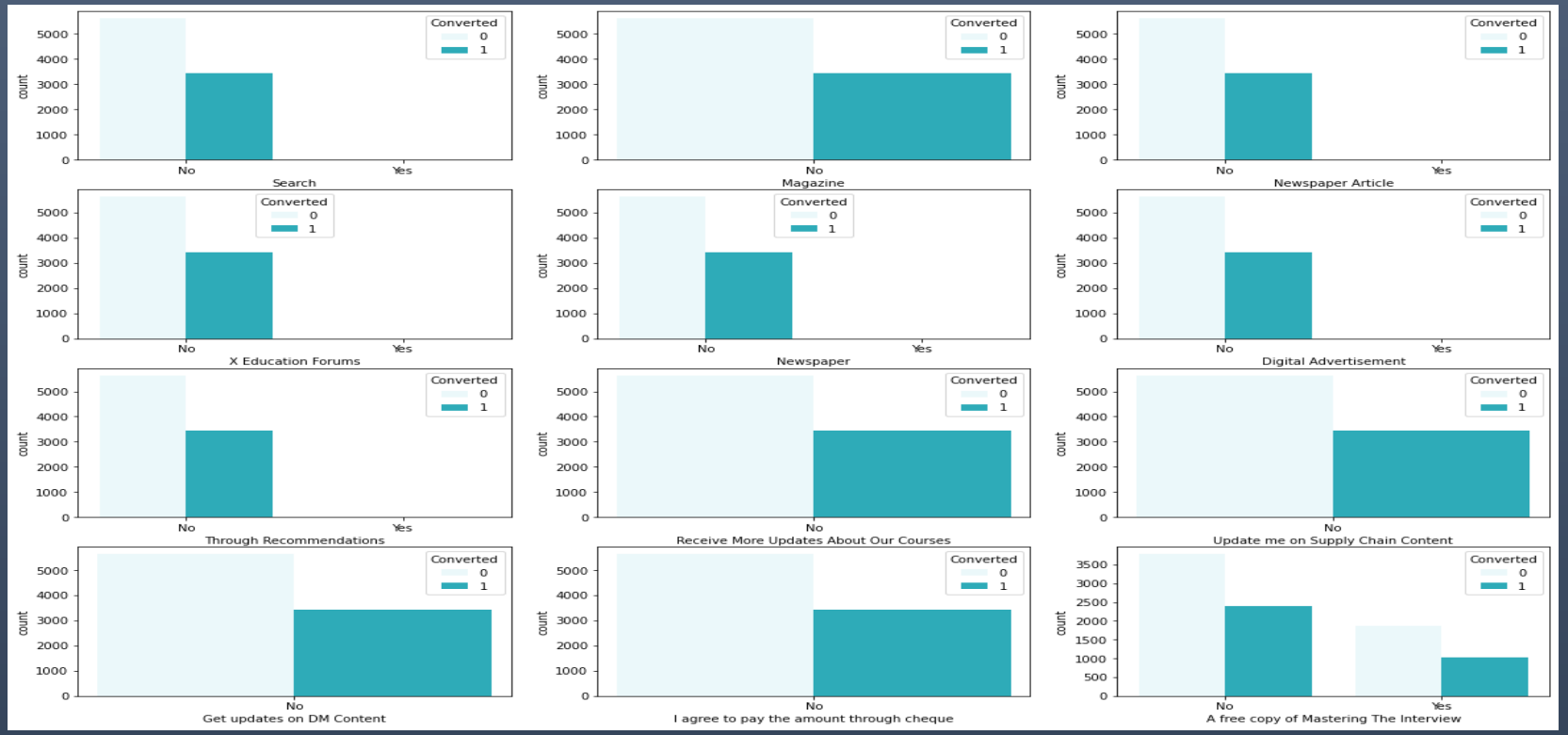
Exploratory Data Analysis (EDA)

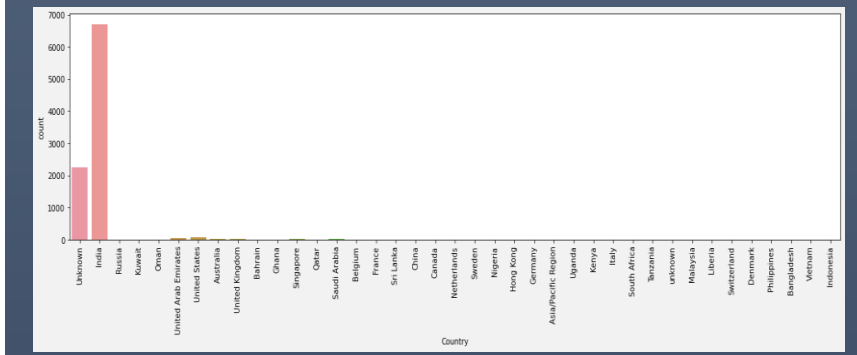


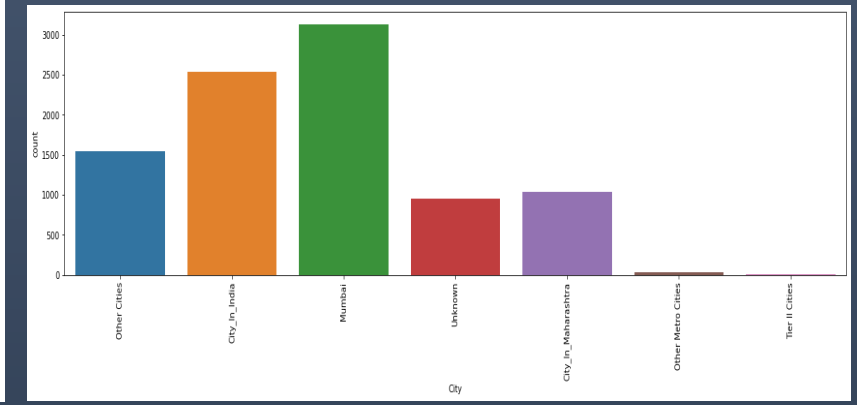


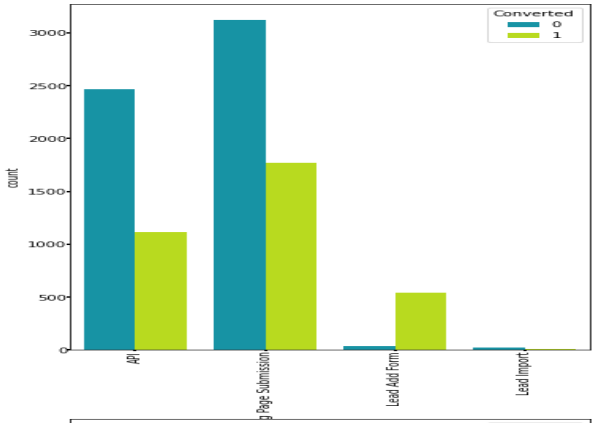


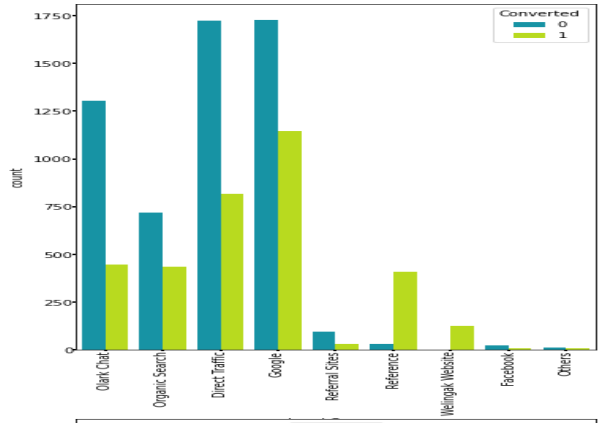


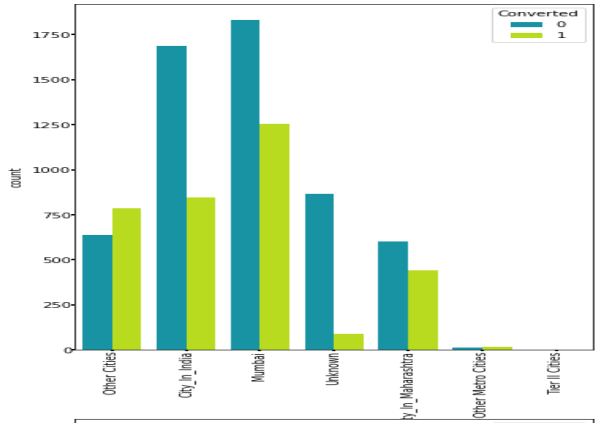


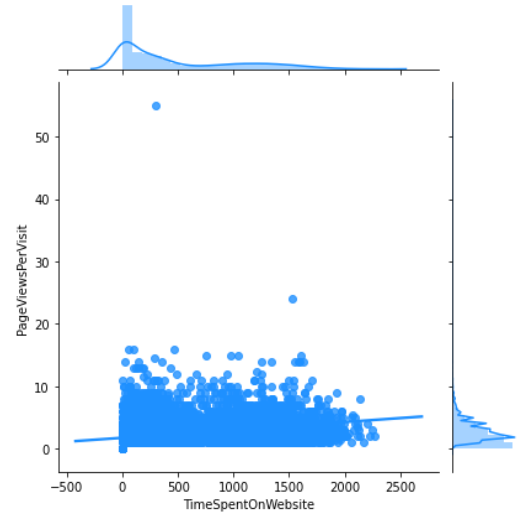


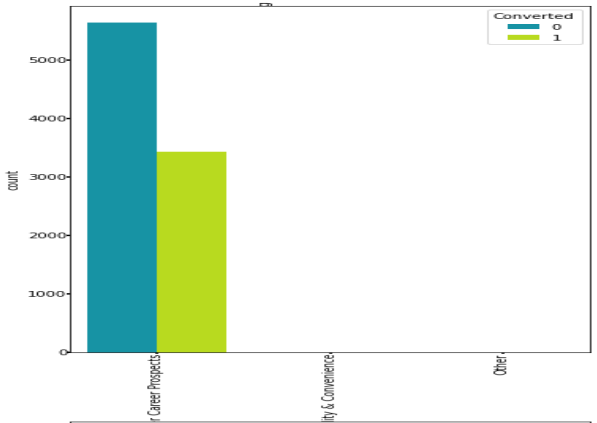


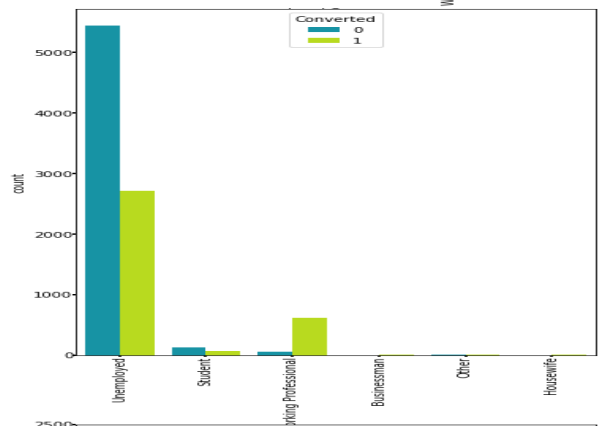

Converted
0
1
count

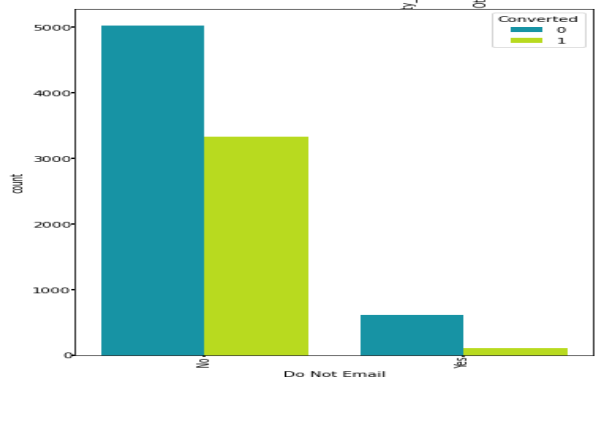

Converted
0
1
count

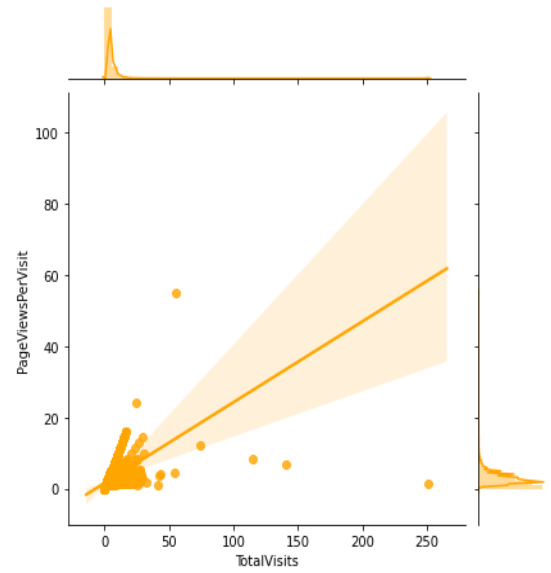

Converted
0
1
count

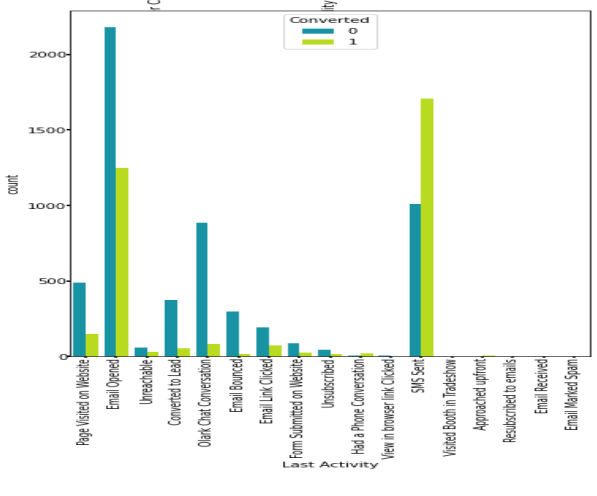

PageViewsPerVisit
TimeSpentOnWebsite

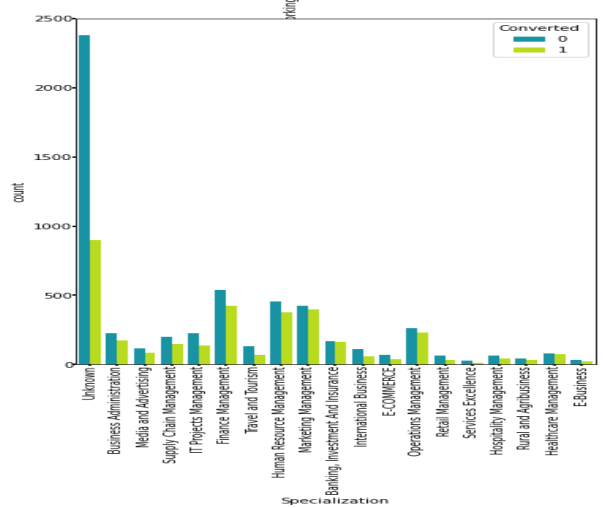

Converted
0
1
count

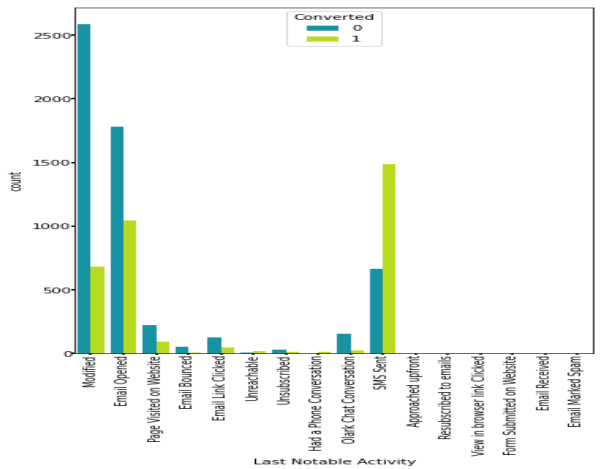

Converted
0
1
count


Converted
0
1
count
Do Not Email


PageViewsPerVisit
TotalVisits


Converted
0
1
count
Last Activity


Converted
0
1
count
Specialization

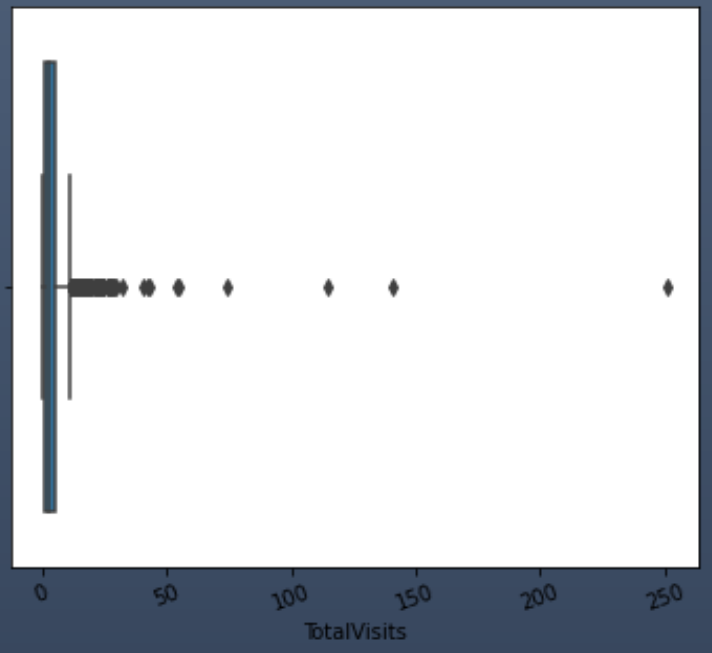

Converted
0
1
count
Last Notable Activity

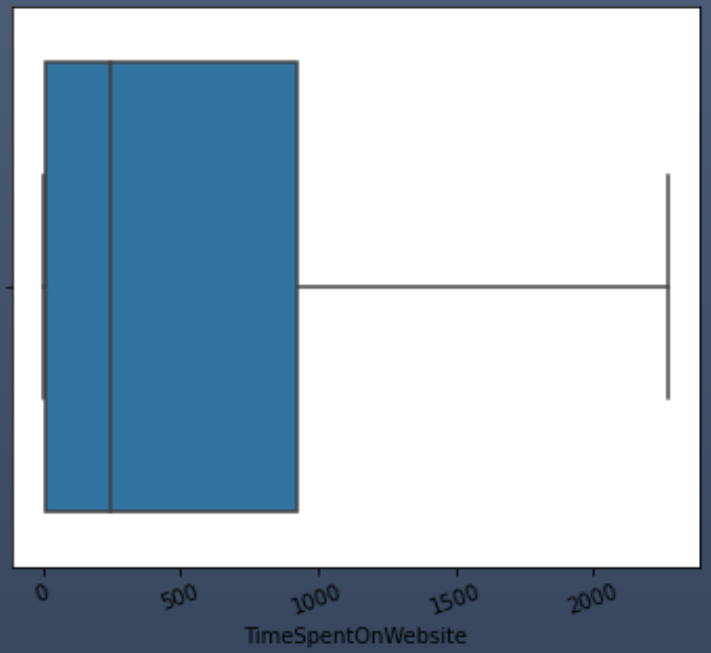
EDA - summarized

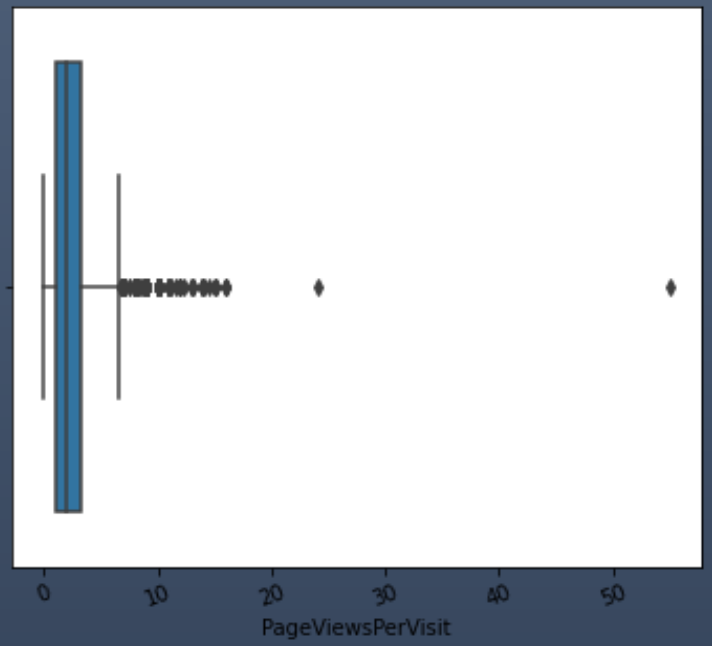
- *Most of the customers not showing interest in calls than emails(email acceptance rate: ~8%*
- *Time spent on website has all the ranges, but maximum records are in 0*
- *As the total number visits increases, pages viewed count also increases*
- *99% of the customers not interested in following up Ads, Articles, Forums, Newspapers, Receiving frequent updates, reactive to DMs also not willing to pay through cheque. However, they show interest in materials on "Mastering The Interview".*
- *Lead Origin: Maximum of the customer is landing on the submission page to take the decision as per the record. Comparatively Lead Add Form status customers are having high probability of conversion*
- *Lead Source: Google, Direct Traffic zones are having lot of customers in record. Even though the reference-category is in smaller amount, the lead conversion rate is high*
- *City: Most of the customers are from India and particularly Mumbai. Also, they are having greater chances of converting-leads*
- *Course purpose: everyone is looking for better career prospects*
- *Occupation: Maximum of the records are in unemployed but still leads are appearing to be the working professionals*
- *Last Activity: SMS sent is the most popular category with high conversion rate.*
- *Specialization: various specializations are popular among customers like, Finance, HR, Markgeting and operation management details. The good picture can be measured when all the customers are insisted filling this data as maximum records are unknown*
- *Last Notable Activity: SMS Sent, email opened are popular ones. Though "Modified" has lots of customers, the conversion rate is very low*

Data Preparation

- Categorical to Numeric conversion
  - Do Not Email
  - A Free Copy of Mastering The Interview
- Dummy creation
  - Lead Origin
  - Lead Source
  - Specialization
  - Last Notable Activity
  - Occupation
- Outlier Check and data scale check
  - Total Visits
  - Time Spent On Website
  - Page Views Per Visit



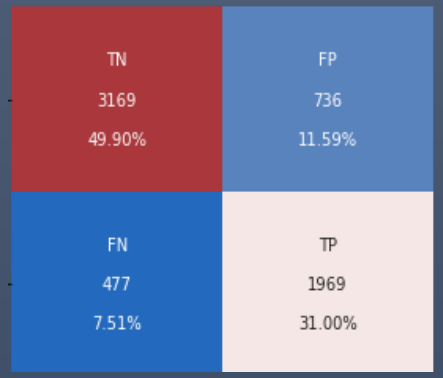


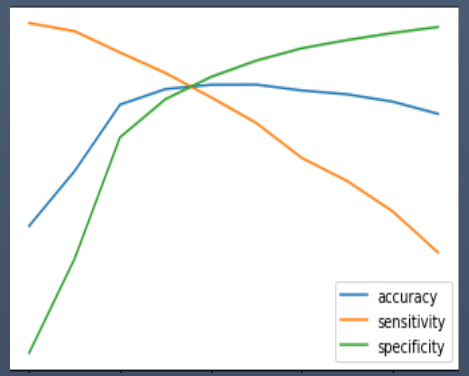


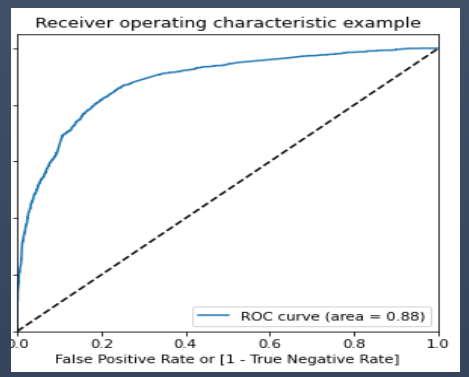
Data Modelling – Train Data

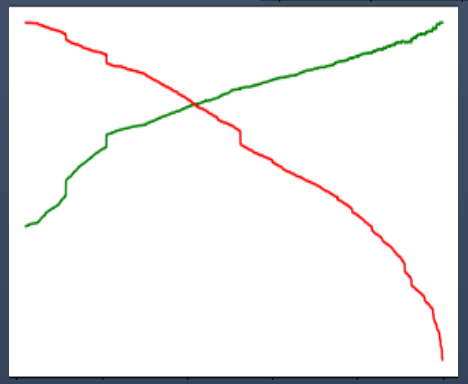
Model prediction on Train-data

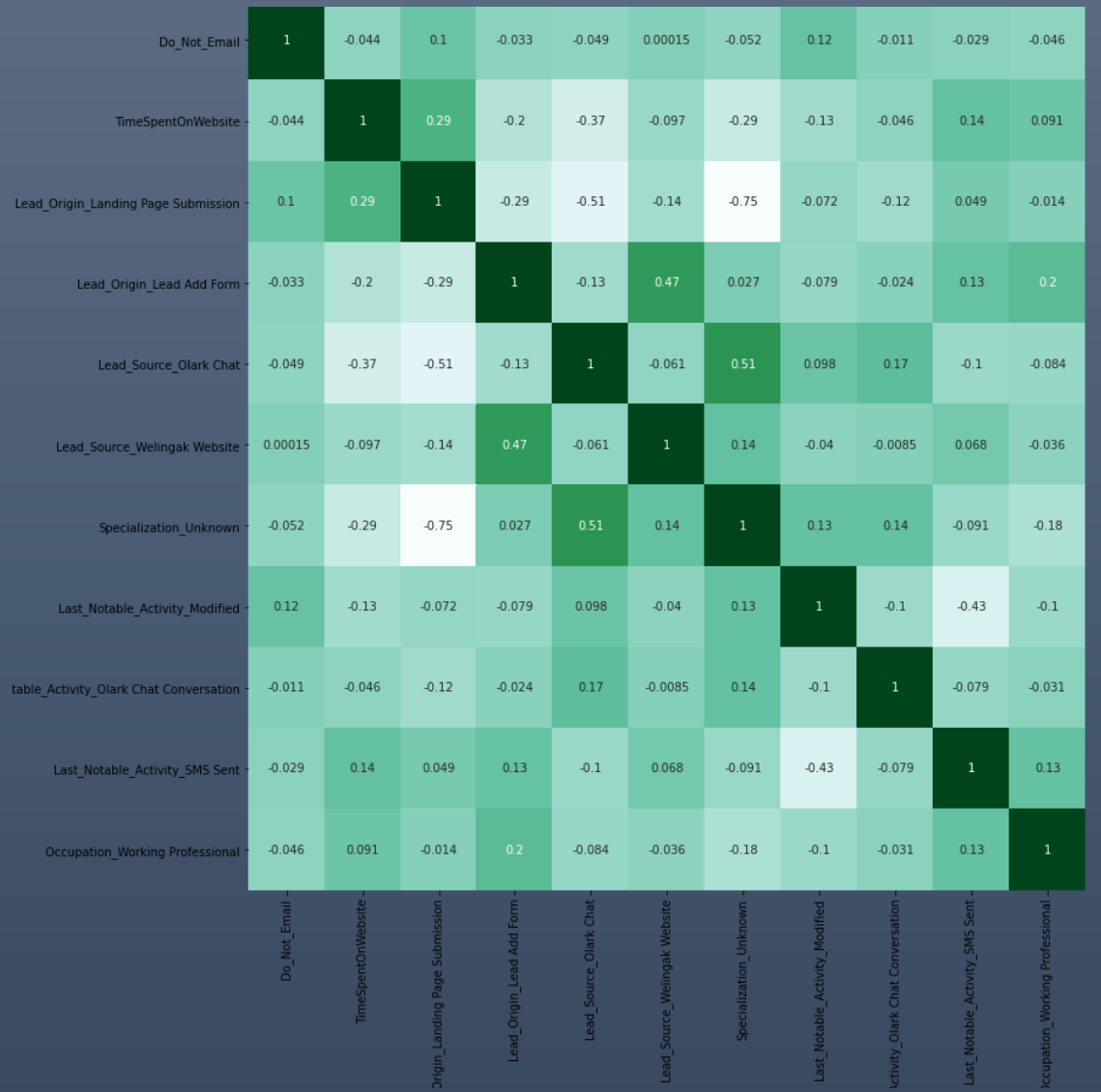
Accuracy= 0.810, Sensitivity= 0.806, Specificity= 0.813, False Positive Value= 0.18, Negative Prediction Value= 0.87, Positive Prediction Value= 0.73, Precision= 0.73, Recall= 0.80, F1-Score= 0.767, Threshold cut-off=0.4, ROC Curve Area=0.88









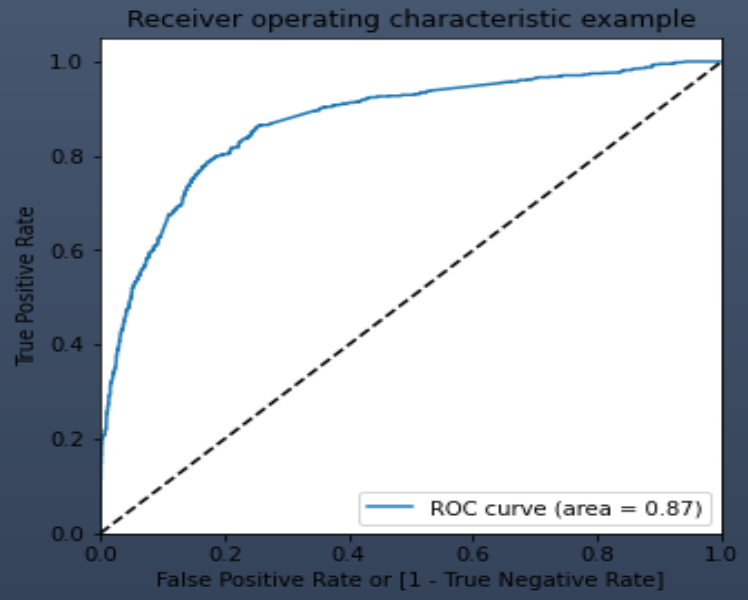


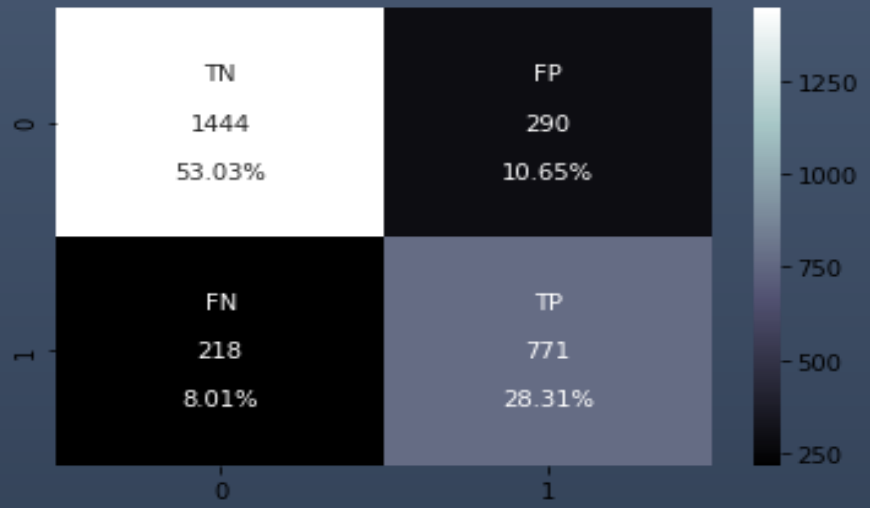
Data Modelling On Test Data

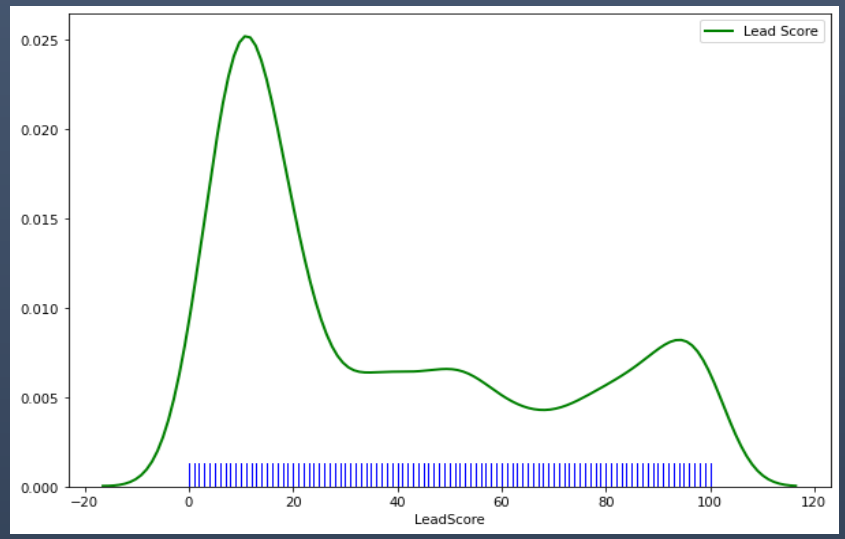
Prediction On Test-data Accuracy= 0.813, Sensitivity= 0.779, Specificity= 0.832, False Positive Value= 0.16, Negative Prediction Value= 0.86, Positive Prediction Value= 0.72, Precision= 0.72, Recall= 0.77, F1-Score= 0.75, Roc Curve Area = 0.87.

Top Features with coefficients:

Lead Origin: Lead Add Form	3.39	Last Notable Activity: SMS Sent	1.42
Lead Source: Welingak Website	2.70	Time Spent On Website	1.10
Occupation: Working Professional	2.61	Lead Source: Olark Chat	0.95







Conclusion

Customers who has,

- *Lead origin which has "Lead Add Form" are having highest conversion rate from the past records*
- *With Leas source: Welingak Website are high chances of conversion*
- *Who had answered "Working professional" for "what is your current occupation" can be reached*
- *Their last notable activity as "SMS sent" can be reached as record shows the high conversion rate for this category.*
- *The time spent on website are more likely to be converted as leads. The maximum the time the conversation rate increases.*

- *In the features having negative correlation, Lead Origin: Landing Page Submission users are more likely to be converted in the past records*

Identified Features with coefficients

Lead_Origin_Lead Add Form	3.39
Lead_Source_Welingak Website	2.70
Occupation_Working Professional	2.61
Last_Notable_Activity_SMS Sent	1.42
TimeSpentOnWebsite	1.10
Lead_Source_Olark Chat	0.95
Last_Notable_Activity_Modified	-0.71
Lead_Origin_Landing Page Submission	-1.15
Specialization_Unknown	-1.21
Last_Notable_Activity_Olark Chat Conversation	-1.30
Do_Not_Email	-1.48

	Features	VIF
6	Specialization_Unknown	2.21
4	Lead_Source_Olark Chat	1.92
8	Last_Notable_Activity_Modified	1.89
2	Lead_Origin_Landing Page Submission	1.79
10	Last_Notable_Activity_SMS Sent	1.62
3	Lead_Origin_Lead Add Form	1.61
5	Lead_Source_Welingak Website	1.37
1	TimeSpentOnWebsite	1.30
11	Occupation_Working Professional	1.19
0	Do_Not_Email	1.12
9	Last_Notable_Activity_Olark Chat Conversation	1.09
7	Last_Notable_Activity_Had a Phone Conversation	1.00