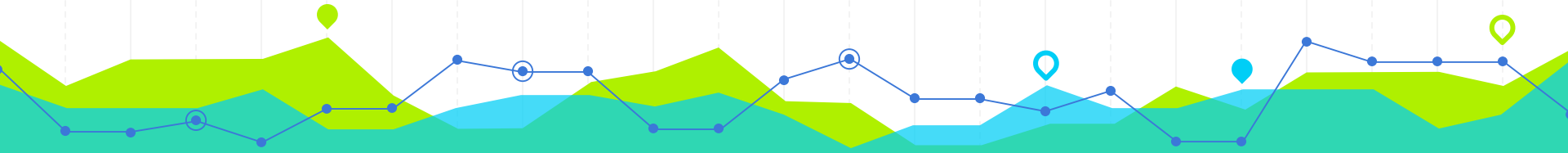
Lead Score Assignment

Submitted by
Sarath Chandrika
Kunal Sadana
DS C40 Batch

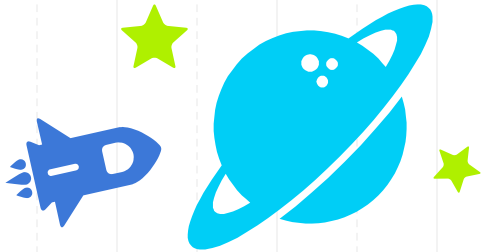


Problem Statement

Lead Conversion Process

Problem Statement

X Education has appointed you to help them 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.



Approach

The steps followed are:

- Importing necessary Libraries
- Loading and Understanding Data
- Data Inspection:
  - Removing unwanted variables
  - Dealing with Null values
  - Dealing with Outliers
  - Binning different categories into one category for categorical columns
- Data Visualization
- Data Preparation:
  - Converting binary variables to 0/1
  - Creating Dummy variables
  - Splitting data into X and y
  - Doing train_test_split to divide into train and test data
  - Scaling
- Building Model
- Model Assessment
- Finding Optimal Threshold Value
- Model Prediction

Data Cleaning

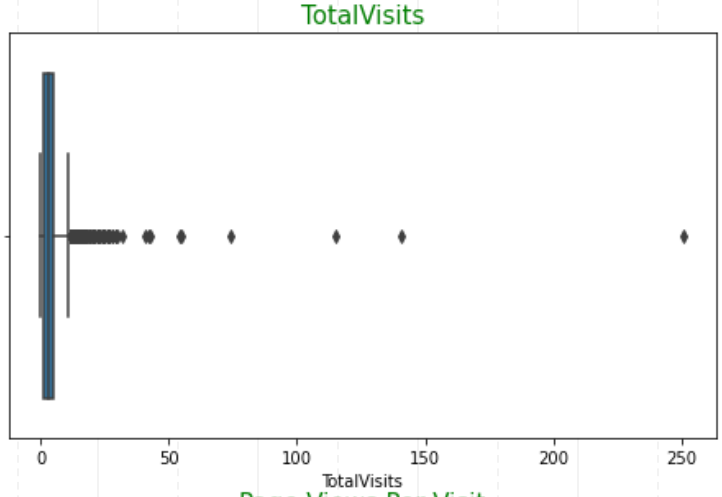
Removing unwanted columns

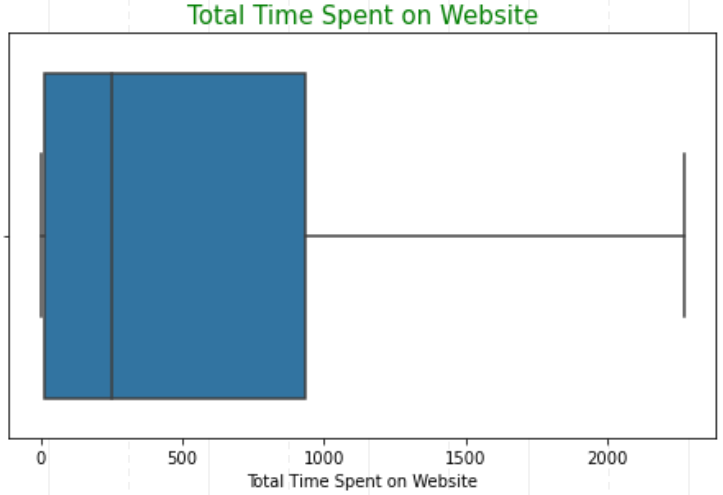
- Few of the columns are removed that are
  - not relevant to business case study
  - biased to a single category.
  - One of the two unique identification columns is dropped i.e, ‘Prospect ID’.

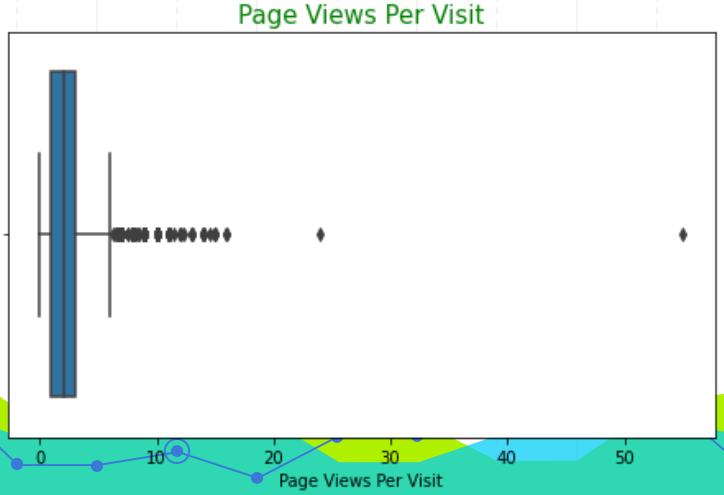
Dealing with Null Values

- Select value across most of the columns is considered as null.
- Drop column with highest null values (‘How did you hear about X Education’ 78%).
- Replace null values of
  - Lead Source with Mode
  - Total Visits, Page Views Per Visit with Median
  - Few categories are grouped as others or unknown.

Dealing with Outliers

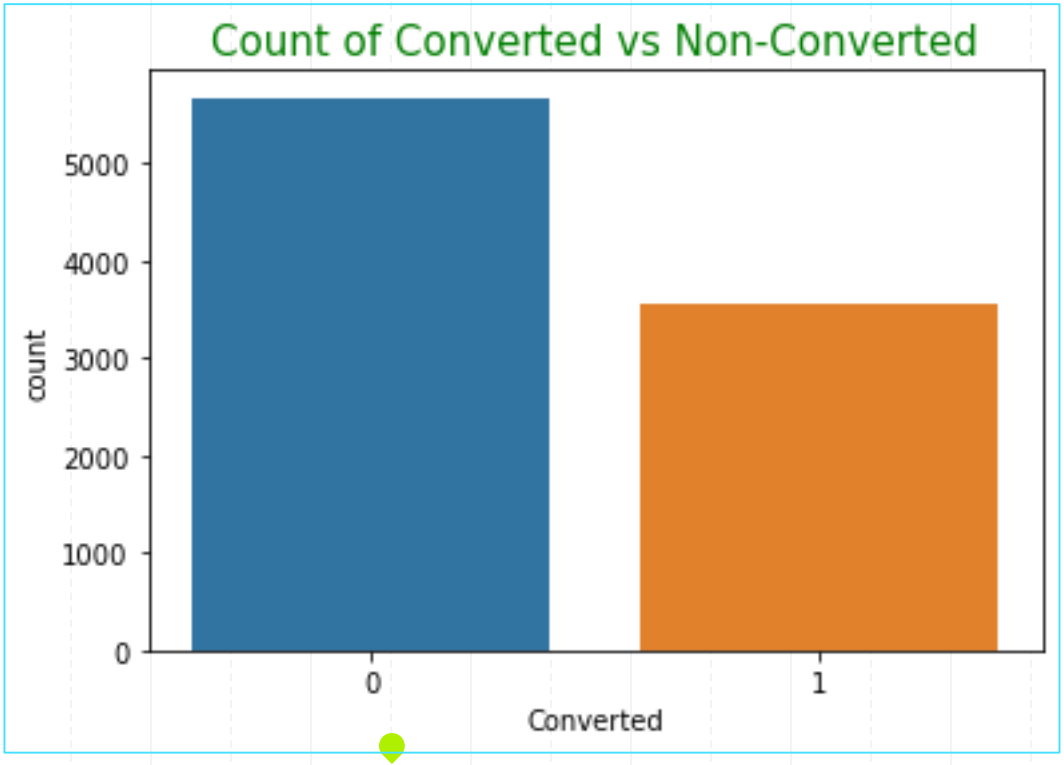




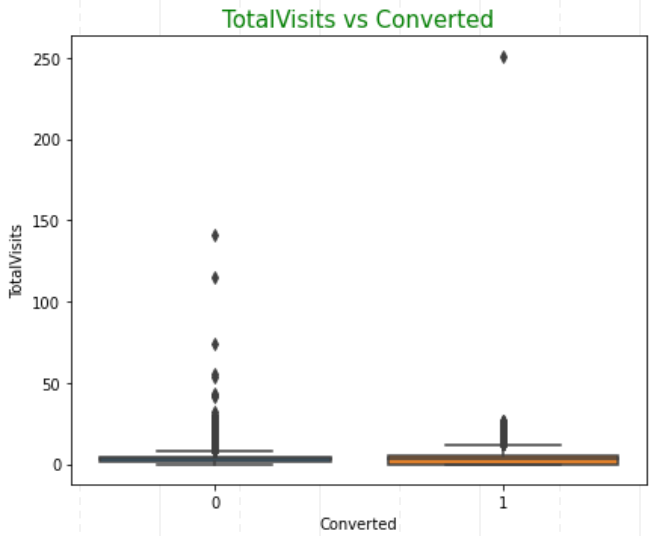


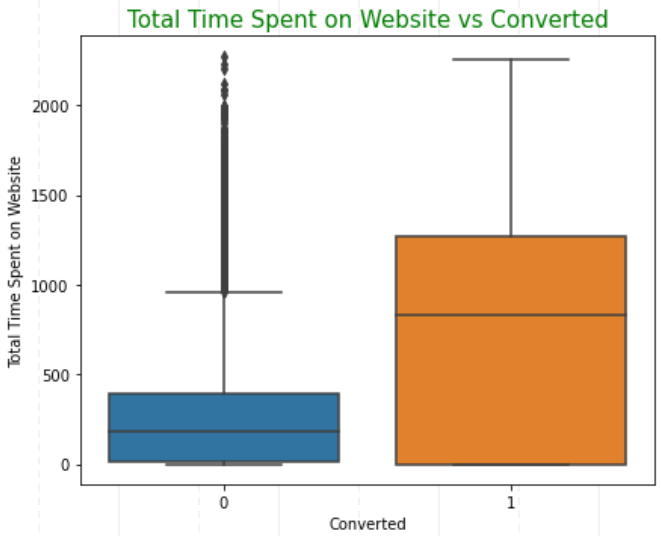
Outliers present in `TotalVisits` and `Page Views Per Visit` but we cannot remove them as they are of siginificance

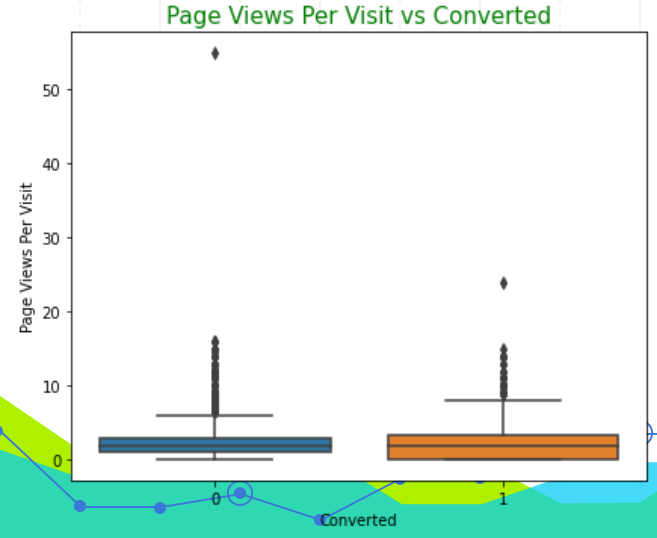
Data Visualisation



Number of converted leads (38.54%) is significantly lower than non-converted (61.46%)

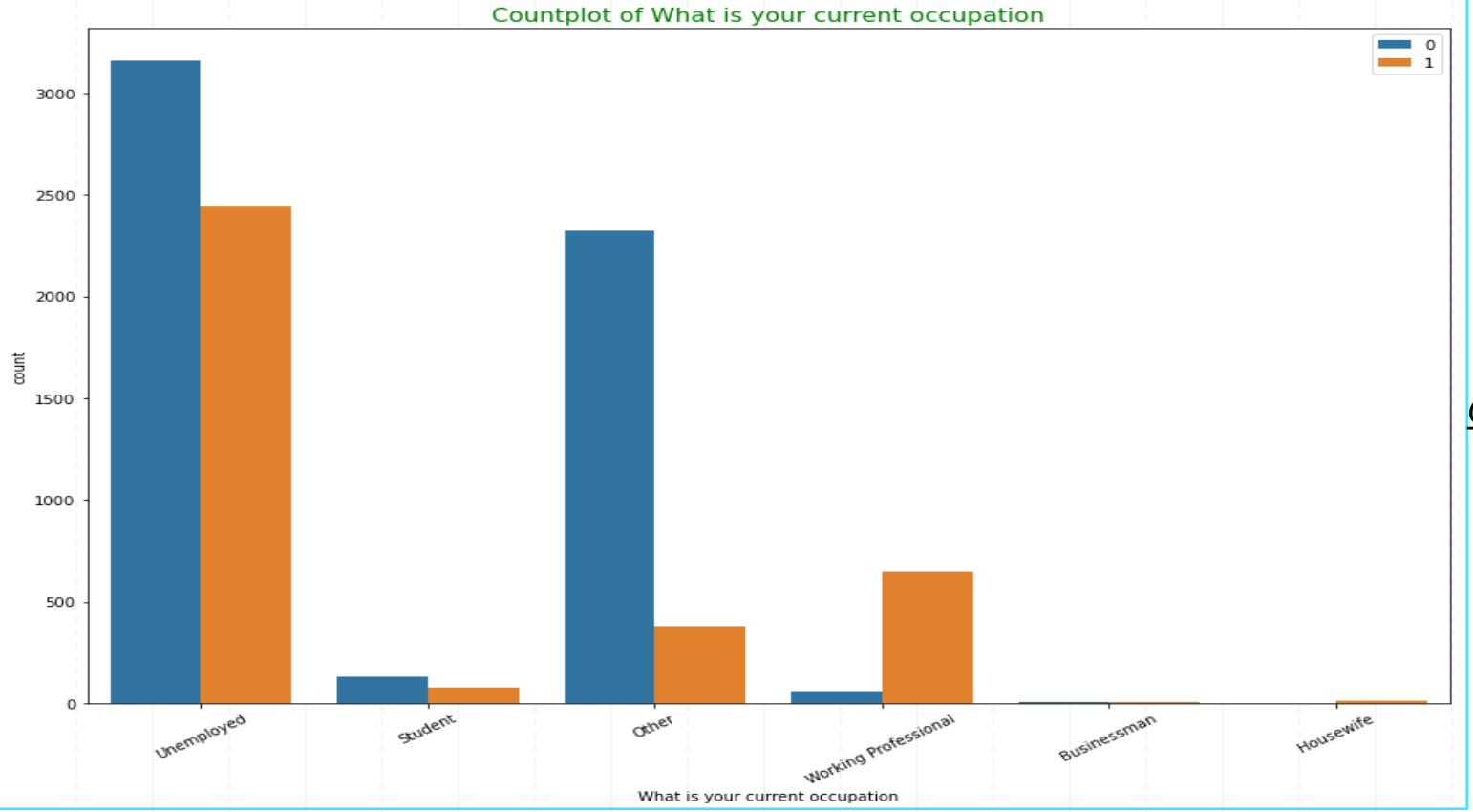




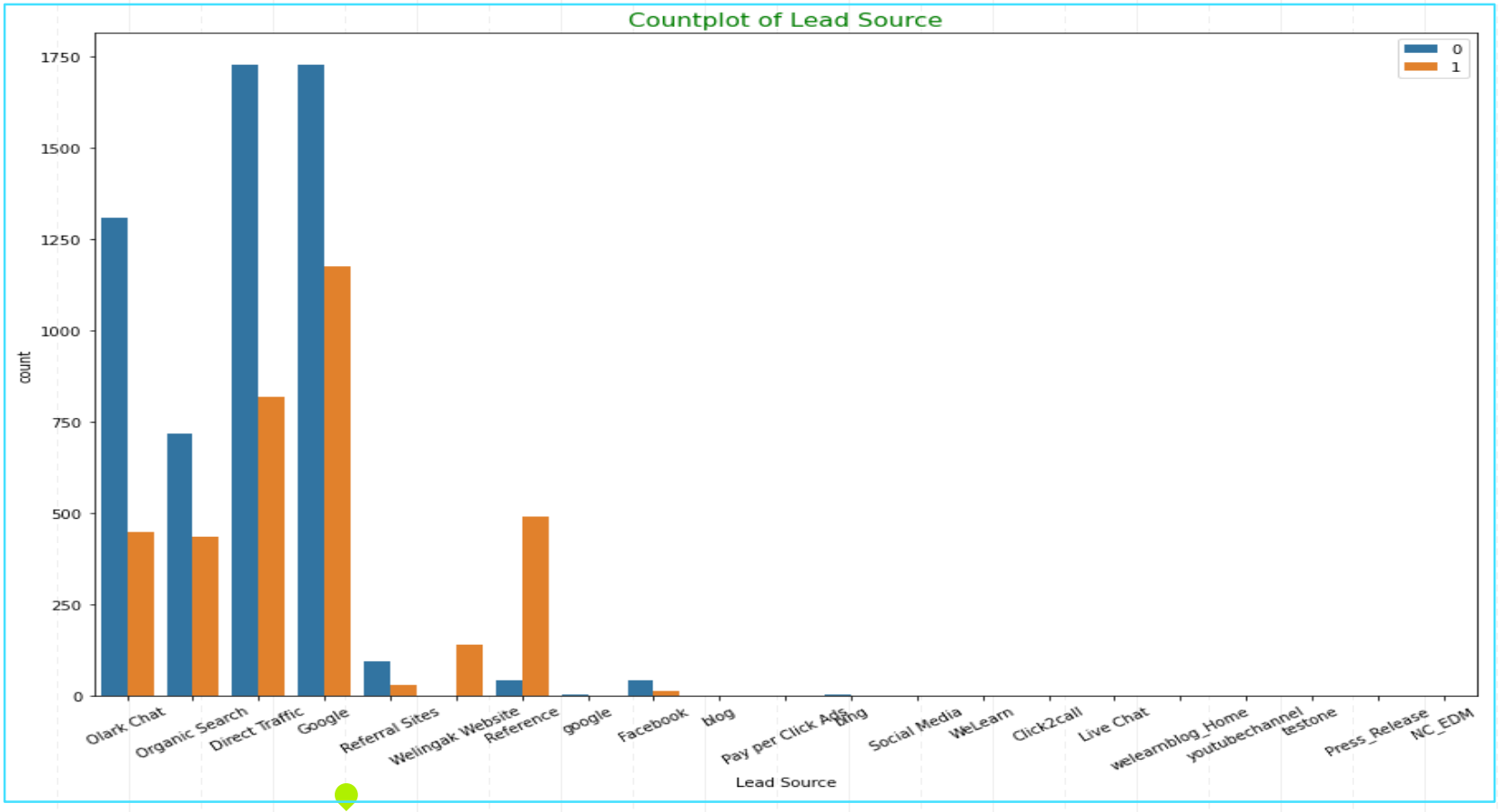


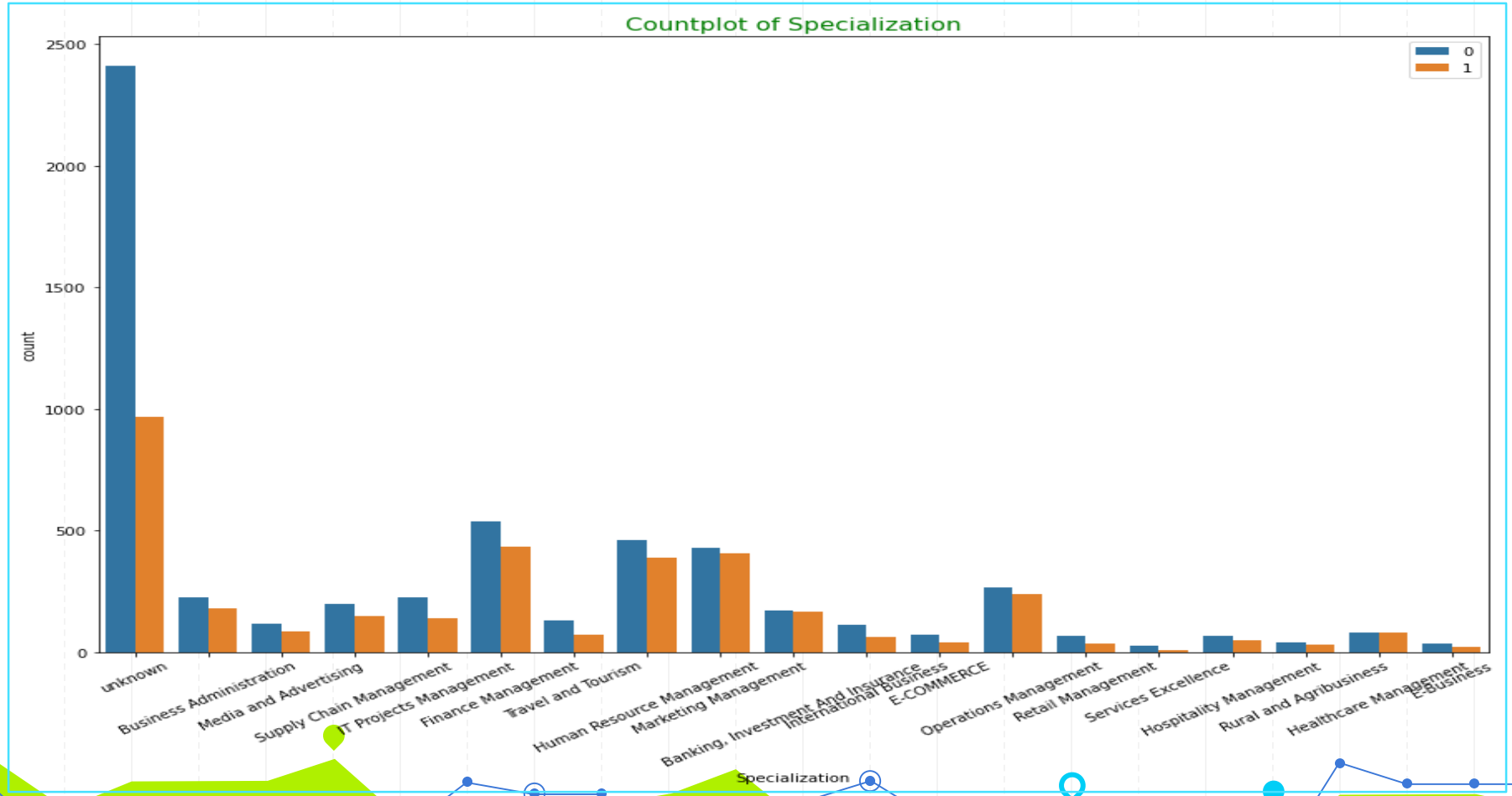
Continuous Variables

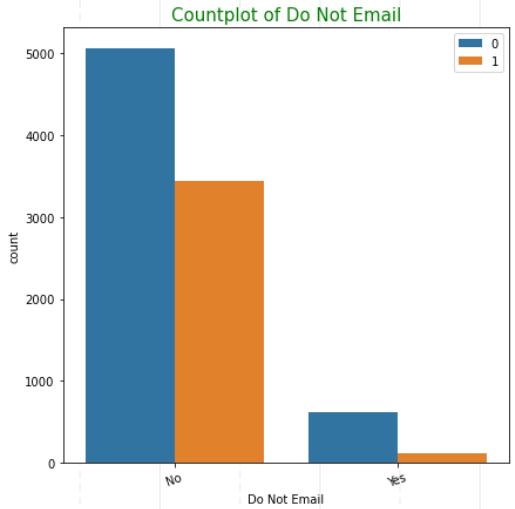
- For `Total Time Spent on Website` converted leads is more than that compared to non-converted

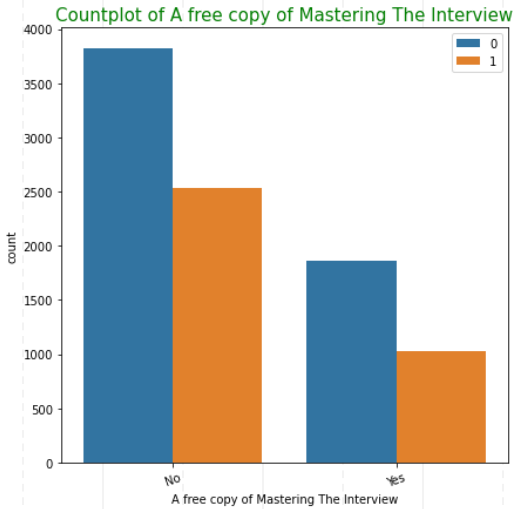


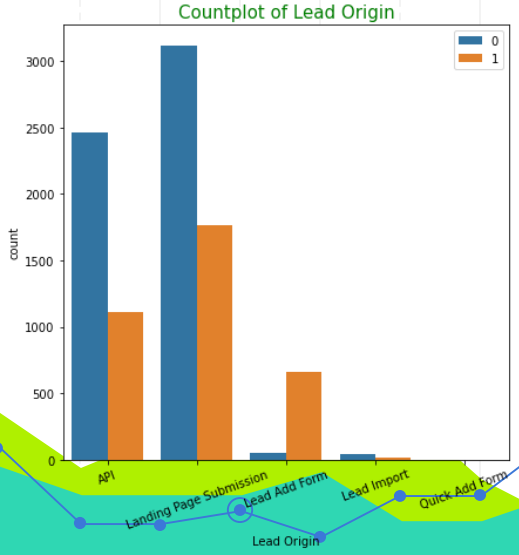
Categorical Variables

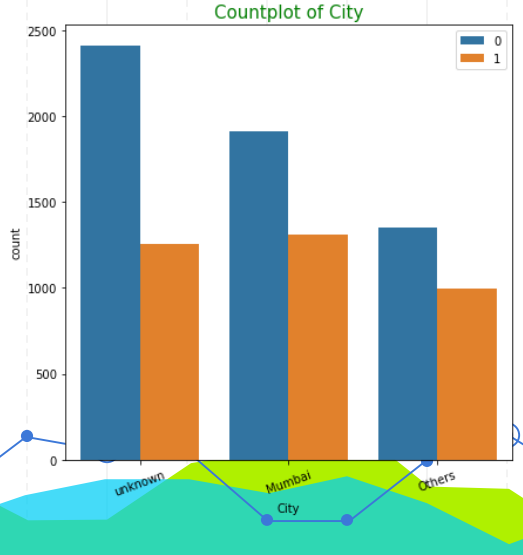

Countplot of Lead Source
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Countplot of Specialization
0
1
count
2500
2000
1500
1000
500
0
unknown
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization



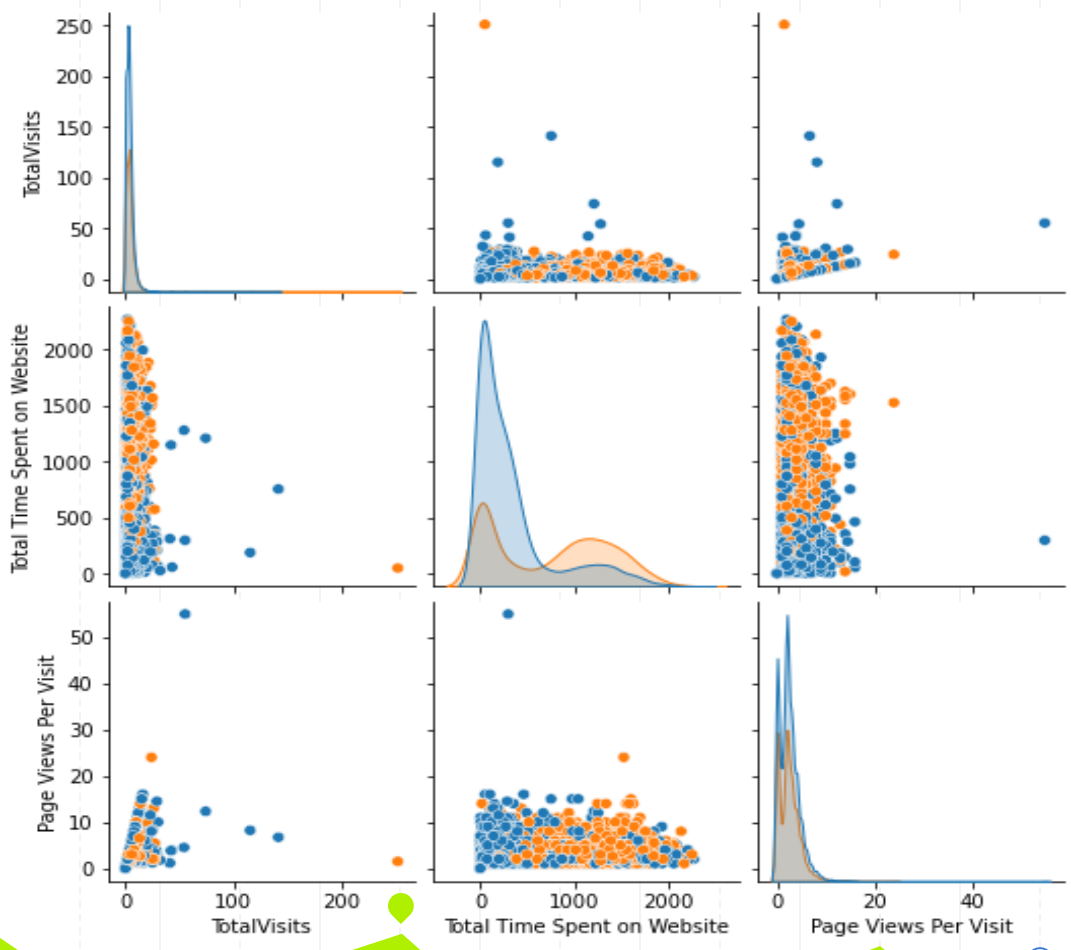


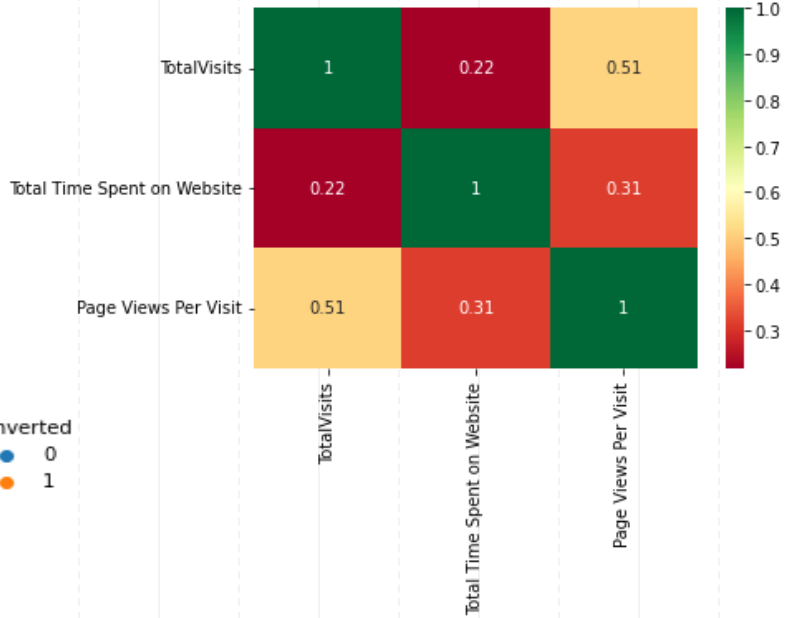




Categorical Variables

- Most cases non-converted is more than converted leads which is to be expected since total percent of non-converted is more
- For `Lead Origin`:'Lead Add Form' converted leads is significantly higher
- For `Lead Source`: 'Welingak Website' and 'Reference' has higher number of Converted
- For `What is Your Current Occupation`: 'Working Professionals' have a higher number a Converted





Multivariate Analysis

- `TotalVisits` and `Page Views Per Visit` have correlation on the higher side

Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2857.6
Date:	Tue, 12 Jul 2022	**Deviance:**	5715.1
Time:	17:24:49	**Pearson chi2:**	8.00e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1485	0.113	1.313	0.189	-0.073	0.370
Do Not Email	-1.3015	0.160	-8.128	0.000	-1.615	-0.988
Total Time Spent on Website	1.0921	0.039	28.339	0.000	1.017	1.168
Lead Origin_Landing Page Submission	-0.9296	0.118	-7.856	0.000	-1.162	-0.698
Lead Origin_Lead Add Form	2.1972	0.198	11.075	0.000	1.808	2.586
Lead Origin_Lead Import	-1.4872	0.537	-2.771	0.006	-2.539	-0.435
Lead Source_Welingak Website	2.5357	0.743	3.412	0.001	1.079	3.992
Country_unknown	0.9646	0.114	8.492	0.000	0.742	1.187
Specialization_Hospitality Management	-0.9872	0.320	-3.082	0.002	-1.615	-0.359
Specialization_unknown	-1.0654	0.119	-8.940	0.000	-1.299	-0.832
What is your current occupation_Other	-1.1739	0.083	-14.160	0.000	-1.336	-1.011
What is your current occupation_Working Professional	2.3873	0.185	12.934	0.000	2.026	2.749

	Features	VIF
6	Country_unknown	2.82
8	Specialization_unknown	2.13
3	Lead Origin_Lead Add Form	1.84
9	What is your current occupation_Other	1.58
2	Lead Origin_Landing Page Submission	1.35
5	Lead Source_Welingak Website	1.26
1	Total Time Spent on Website	1.25
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.11
4	Lead Origin_Lead Import	1.03
7	Specialization_Hospitality Management	1.02

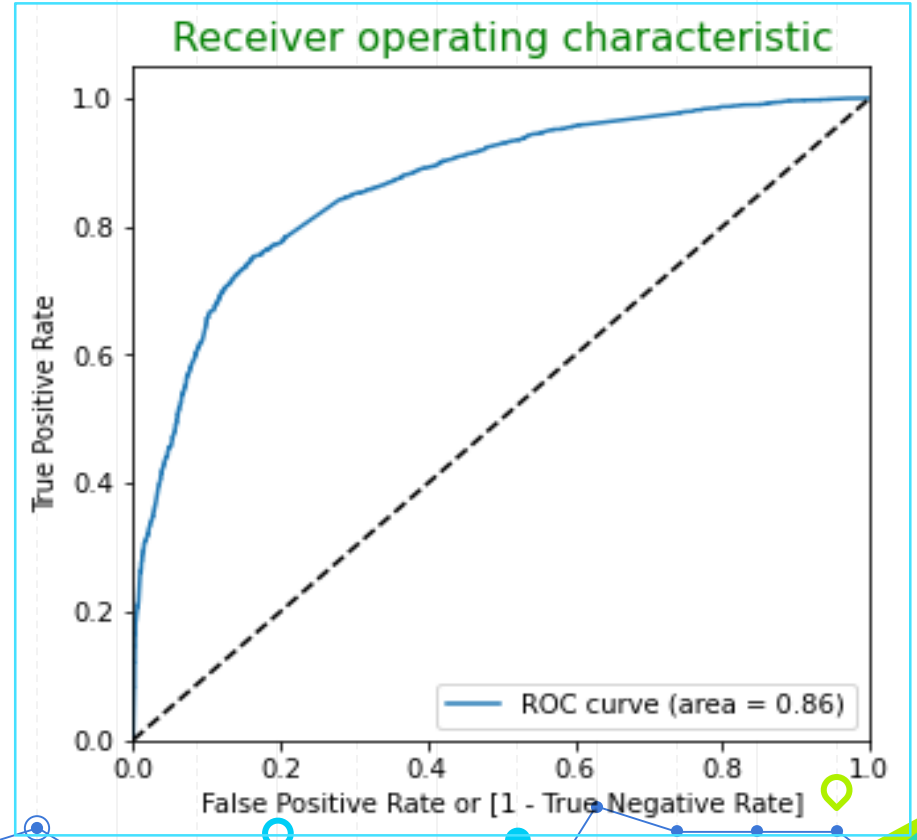
Variables in the final model with their respective coefficient values.

Model Assessment

Evaluation metrics based on 0.5 cutoff
Accuracy: 0.808
Sensitivity/Recall: 0.668
Specificity: 0.894
Positive predicted values/Precision: 0.795
Negative Predicted Values: 0.814
F-score: 0.726

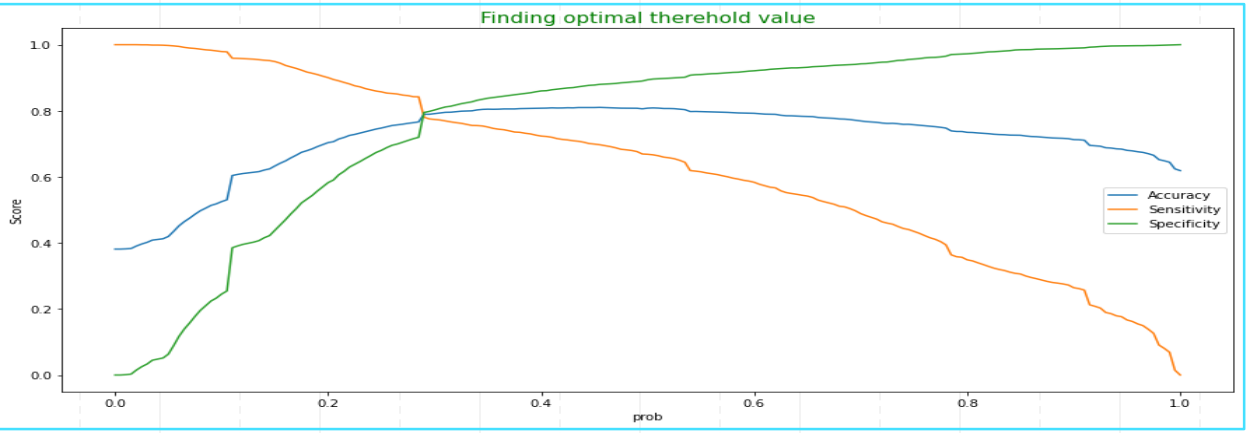
Confusion Matrix

Actual/Predicted	Not Converted	Converted
Not Converted	3577	425
Converted	818	1648



Optimal Threshold Value

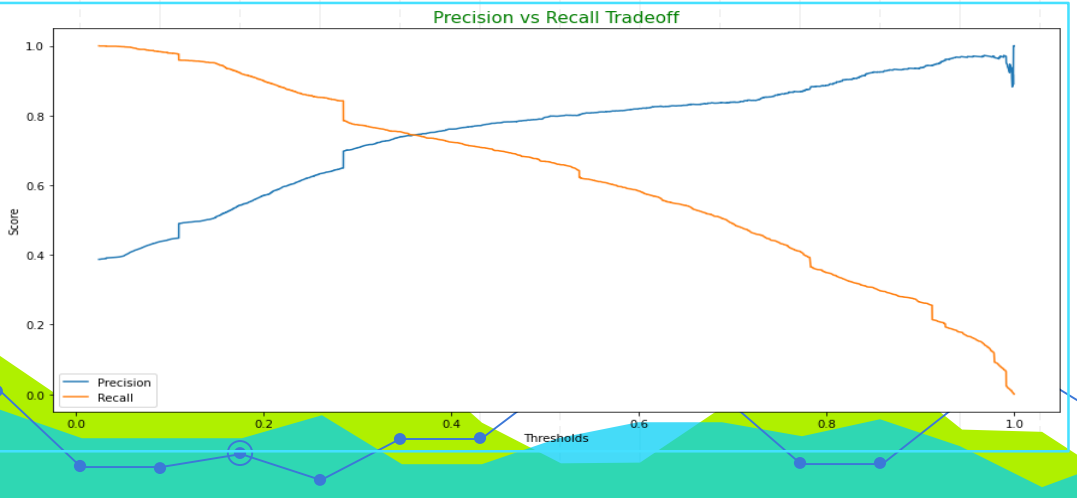
We can clearly see, taking 0.29 as cutoff value gives better Sensitivity or Recall value than at 0.36



Evaluation metrics (cutoff = 0.29 from curve)
Accuracy: 0.789
Sensitivity/Recall: 0.781
Specificity: 0.795
Positive predicted values/Precision: 0.701
Negative Predicted Values: 0.855
F-score: 0.739

Confusion Matrix

Actual/Predicted	Not Converted	Converted
Not Converted	3181	821
Converted	541	1925



Evaluation metrics (cutoff = 0.36 from curve)
Accuracy: 0.805
Sensitivity/Recall: 0.744
Specificity: 0.842
Positive predicted values/Precision: 0.744
Negative Predicted Values: 0.842
F-score: 0.744

Confusion Matrix

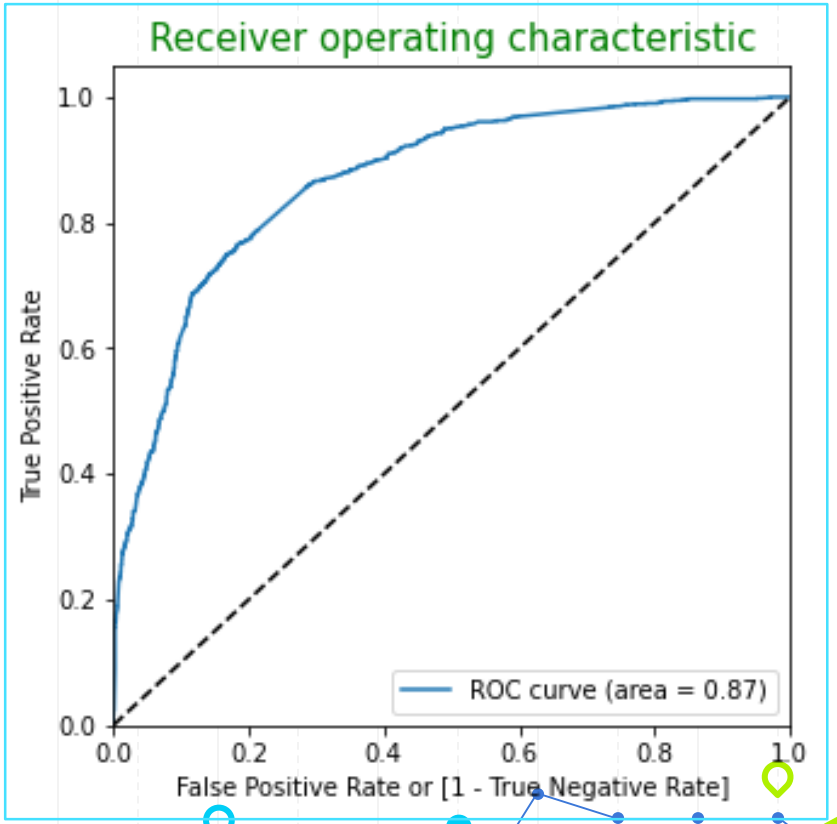
Actual/Predicted	Not Converted	Converted
Not Converted	3370	632
Converted	631	1835

Model Predictions

Evaluation metrics for test data (cutoff = 0.29):
Accuracy: 0.789
Sensitivity/Recall: 0.784
Specificity: 0.792
Positive predicted values/Precision: 0.711
Negative Predicted Values: 0.849
F-score: 0.745

Confusion Matrix

Actual/Predicted	Not Converted	Converted
Not Converted	1328	349
Converted	237	858



Results

- Sensitivity of trained and test models is around 0.784 and 0.781.
- The value is near 80% and difference in values of the test and train data is very less.
- So we can say the model when applied will lead to a good lead conversion rate

	LeadID	Converted_actual	Converted_prob	final_predicted	Lead Score
7014	3478	1	0.999697	1	99.97
4891	8074	1	0.999674	1	99.97
3115	2656	1	0.999717	1	99.97
6312	3428	1	0.999511	1	99.95
8873	5921	1	0.999446	1	99.94

Top 5 rows of Table containing the LEAD SCORES

Summary

Based on the analysis following are top variables as well as categorical variables in model which contribute most towards the probability of a lead getting converted and should be focused the most:

- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Lead Origin_Lead Add Form

In a scenario where the company wish to make the lead conversion more aggressive, and want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible, they should focus more on:

- People who put their source as Welingak Website
- Working Professionals
- People for whom the origin identifier is Add Form
- People spending a lot of time on website
- The company should rigorously do follow ups and nurture such leads by sending them emails or messages to make sure they get converted.

Summary

In a scenario where the company wants the sales team to focus on some new work as well and thus the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls, they should focus less on:

- People who have opted for 'Do Not Email'
- People whose Current Occupation is 'Others'
- People whose specialization is in Hospitality Management or Unknown
- People for whom the origin identifier is Landing Page submission or Import
- People having less Lead Score but predicted as Converted by the model can be sent automated Emails and SMS instead of calling them.

THANKS!