Business Problem Statement

- Identify the most potential leads, also known as ‘Hot Leads’
- 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 chanThe CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

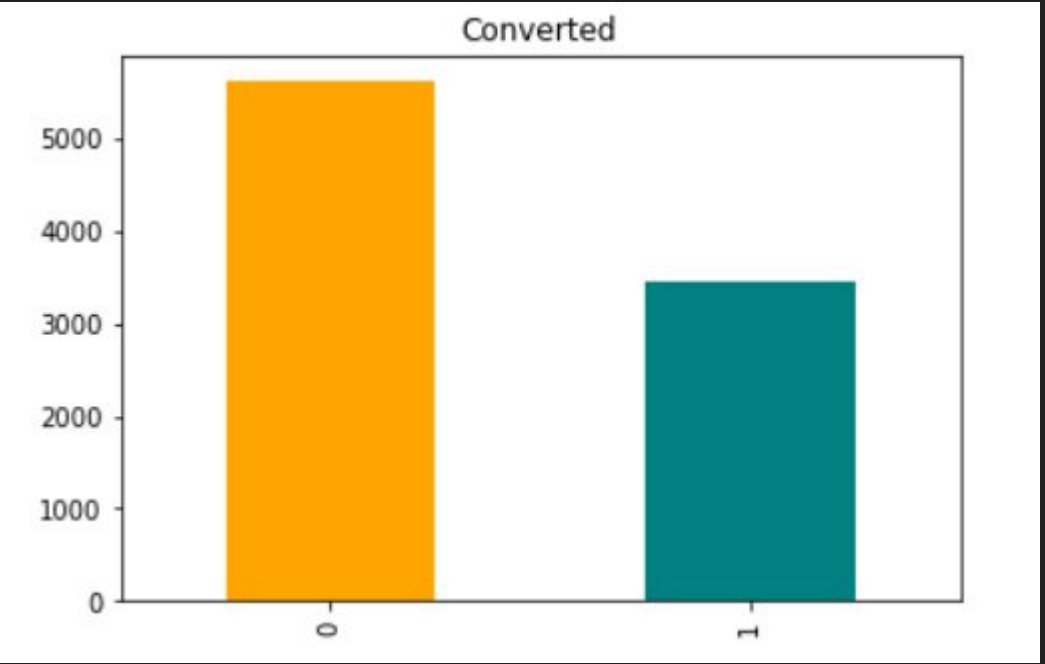
- Basic Data Exploration
- EDA
- Data Preprocessing
- Feature Selection
- Model Training
- Evaluation

Basic Data Exploration

- 9240 leads
- 36 Independent Variables
- Target Variable Converted is Categorical Variable, So it needs a Classification Predictive Model
- Datasheet has initially collected data and sales team updated data
- In some variables missing data represented as ‘Select’. So replaced it with NaN
- Redundant data is corrected. Ex. Google, google
- Independent variables with no variance were dropped
- Independent variables with missing values percentage 40 or above were dropped
- Leads with missing values percentage 70 or above were dropped
- Missing values in categorical variables imputed with NA - Not Available
- Missing values in continuous variables imputed with median
- Extreme outliers leads were dropped

EDA

- Not balanced
- Not converted rate is 2x to conversion rate

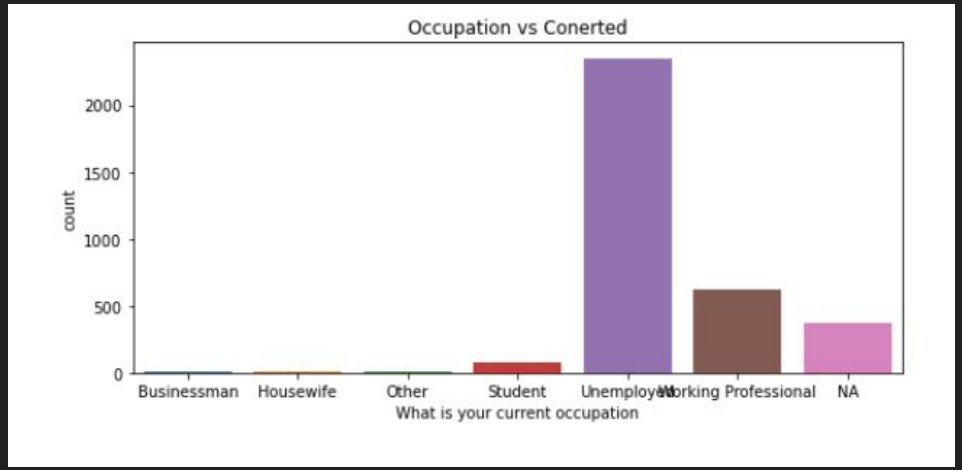


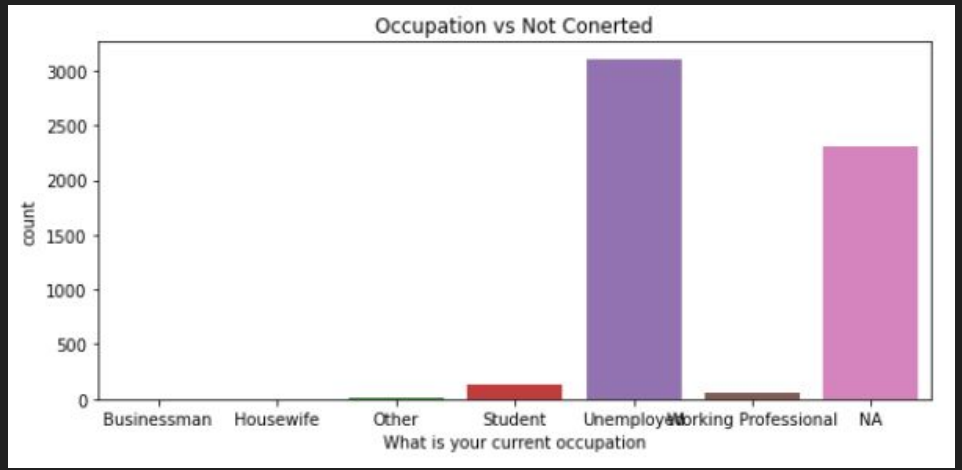
Occupation

1. Working Professional Conversion rate is high

2. Leads which doesn't have occupation details are less likely to convert

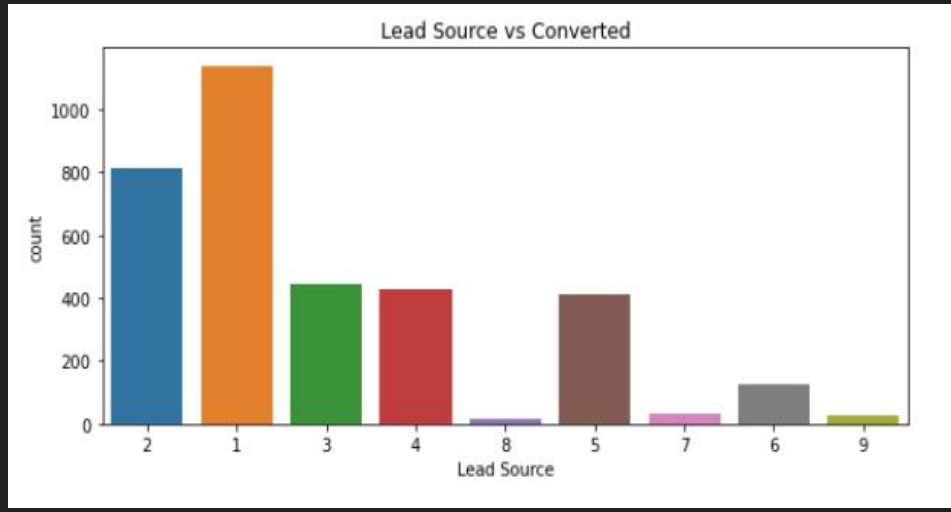
3. unemployed and Student interest is bit unpredictable

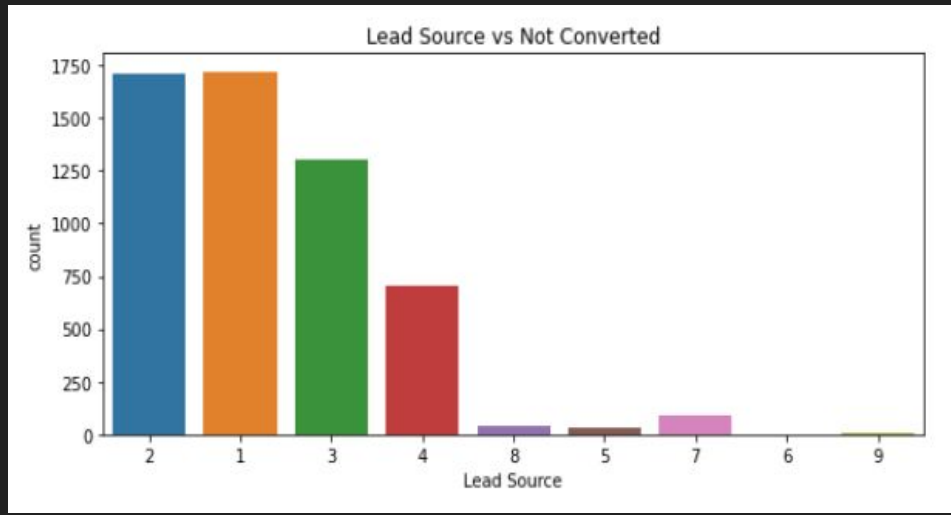




Lead Source

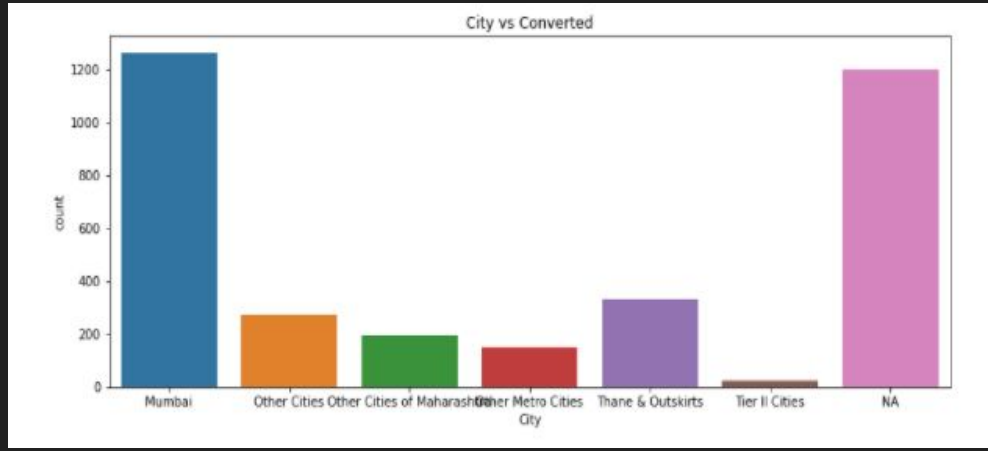
- Reference and Welingak Website conversion rate is high
- Google 50% conversion rate
- Other 'Sources' converting chances are lesser

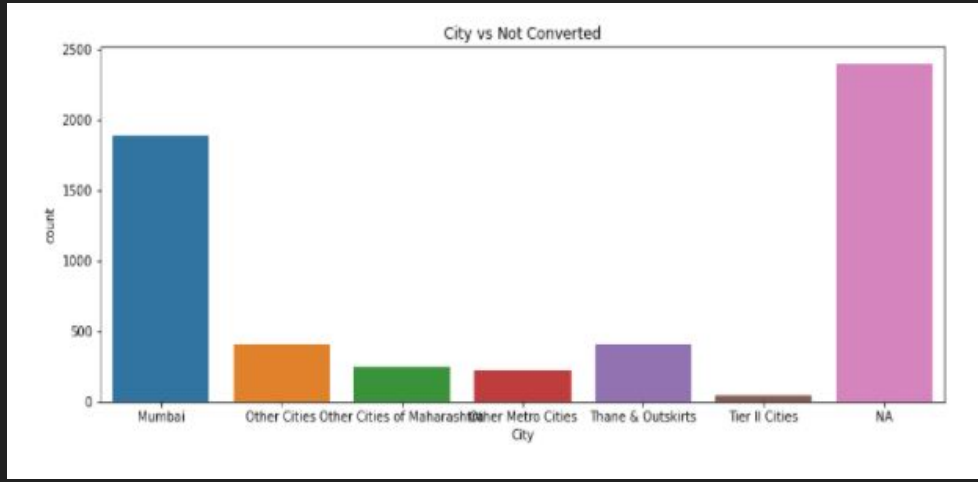




City

Mumbai, Thane & outskirts conversion rate is higher

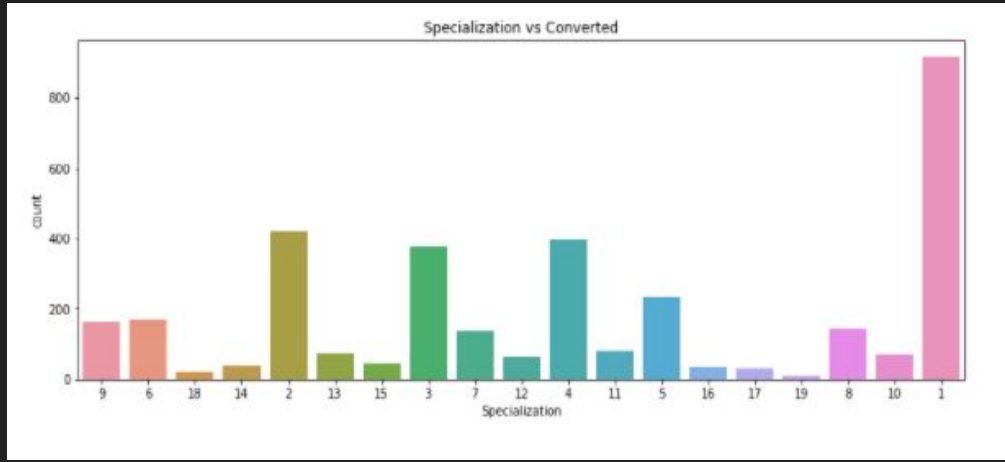


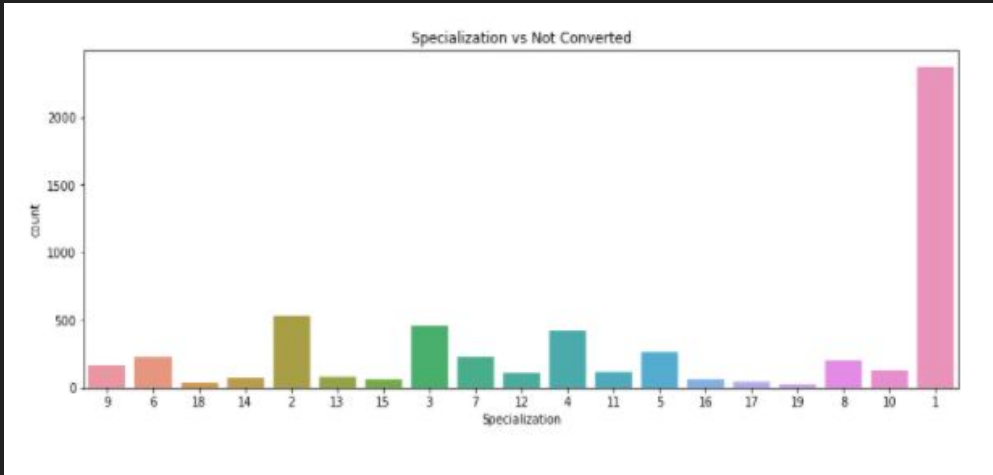


Specialization

NA: 1, Finance Management: 2, Human Resource Management: 3, Marketing Management: 4, Operations Management: 5, Business Administration: 6, IT Projects Management: 7, Supply Chain Management: 8, Banking, Investment And Insurance: 9, Travel and Tourism: 10, Media and Advertising: 11, International Business: 12, Healthcare Management: 13, E-COMMERCE: 14, Hospitality Management: 15, Retail Management: 16, Rural and Agribusiness: 17, E-Business: 18, Services Excellence: 19

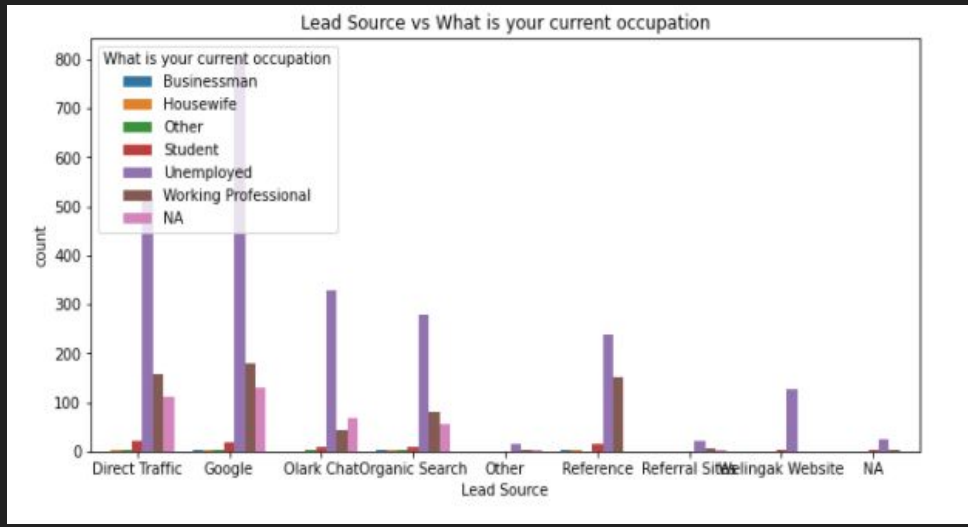
If the leads has specialization chances of conversion is bit higher

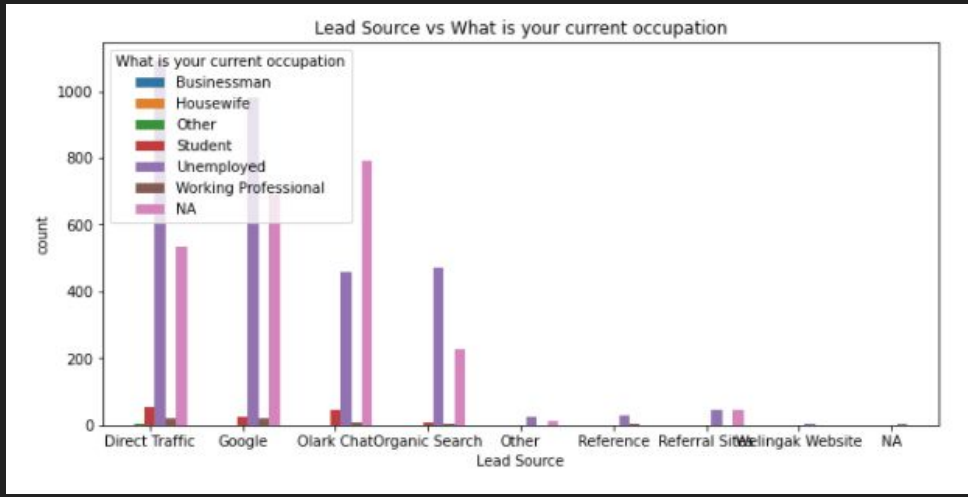




Leads Source vs Occupation

- Working Professional leads are hot leads
-
- Welingak Website leads are hot leads





Correlation

TotalVisits and Per Page Views Per Visit ha high correlation

Data Preprocessing

- Dummies Creation For Categorical Data
- Created new category ‘Other’ If the number of categories are more with less percentage of data
- Create Train(70% train data) Test(30%) Data
- MinMax Scaling for continuous data

Model Training

- RFE used to select to 20 predictors
- Then manually dropped predictors as per p value and VIF
- Logistic Regression with balanced class since given data is imbalance
- Cut Off Identified 0.43

Model Training Contd.

	Features	VIF
0	TotalVisits	2.723435
1	Total Time Spent on Website	2.107680
2	Lead Source_Other	1.047926
3	Lead Source_Welingak Website	1.294429
4	Country_NA	2.673369
5	What matters most to you in choosing a course_NA	1.589661
6	Lead Origin_Landing Page Submission	2.934648
7	Lead Origin_Other	1.860712
8	Do Not Email_Yes	1.101066
9	Specialization_Hospitality Management	1.017684
10	Specialization_NA	2.574784
11	What is your current occupation_Working Profes...	1.187890

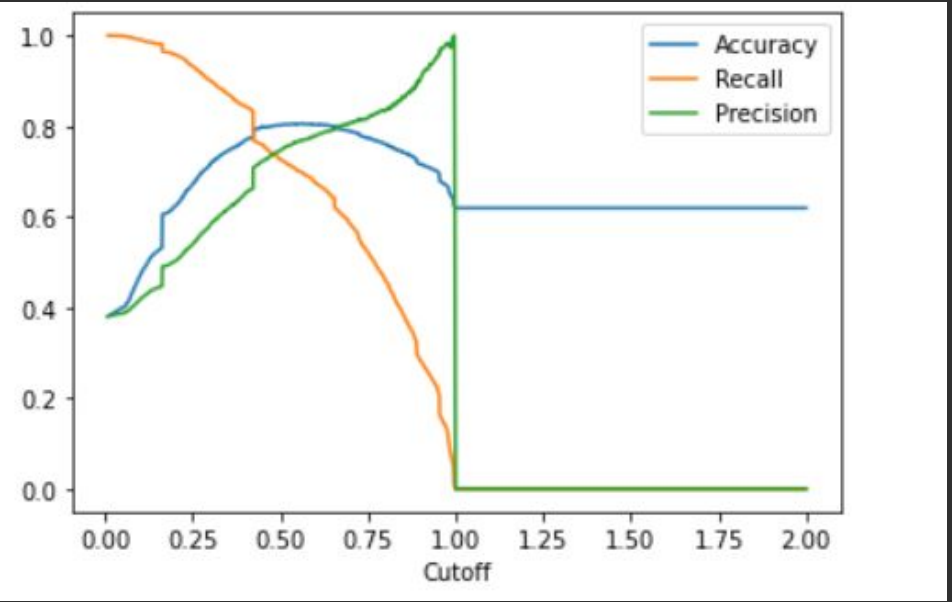
Logit Regression Results

Dep. Variable:	Converted	No. Observations:	6330
Model:	Logit	Df Residuals:	6317
Method:	MLE	Df Model:	12
Date:	Wed, 11 Aug 2021	Pseudo R-squ.:	0.3337
Time:	17:28:26	Log-Likelihood:	-2801.6
converged:	False	LL-Null:	-4204.8
Covariance Type:	nonrobust	LLR p-value:	0.000

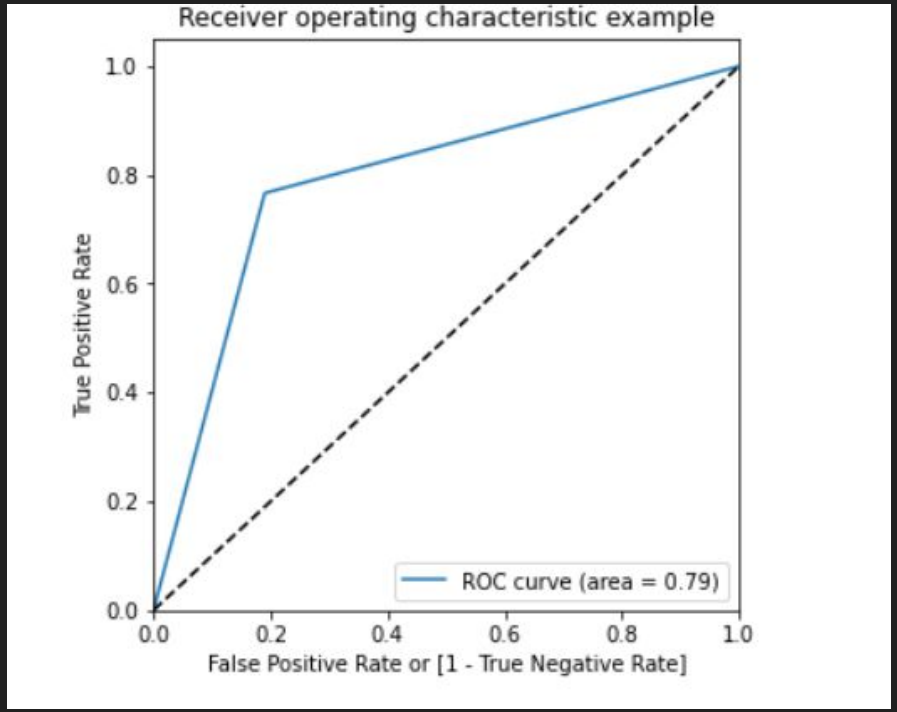
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1258	0.137	-8.233	0.000	-1.394	-0.858
TotalVisits	1.1168	0.246	4.546	0.000	0.635	1.598
Total Time Spent on Website	4.5777	0.163	28.110	0.000	4.259	4.897
Lead Source_Other	-2.6994	0.503	-5.365	0.000	-3.686	-1.713
Lead Source_Welingak Website	2.4720	0.909	2.721	0.007	0.691	4.253
Country_NA	1.3114	0.124	10.577	0.000	1.068	1.554
What matters most to you in choosing a course_NA	-1.3356	0.084	-15.873	0.000	-1.501	-1.171
Lead Origin_Landing Page Submission	-0.8555	0.120	-7.143	0.000	-1.090	-0.621
Lead Origin_Other	2.5672	0.232	11.073	0.000	2.113	3.022
Do Not Email_Yes	-1.1667	0.159	-7.324	0.000	-1.479	-0.854
Specialization_Hospitality Management	-1.0263	0.338	-3.034	0.002	-1.689	-0.363
Specialization_NA	-0.9924	0.122	-8.146	0.000	-1.231	-0.754
What is your current occupation_Working Professional	2.3995	0.192	12.481	0.000	2.023	2.776

Model Training Contd.

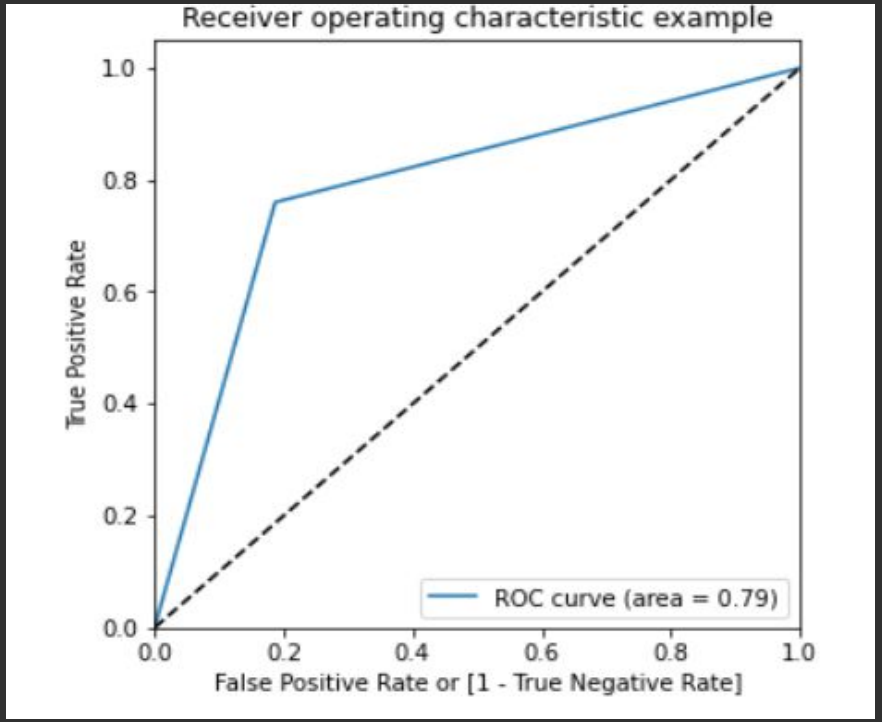
Cut off 0.43



Train ROC



Test ROC



Conclusion

- Concentrate on working professionals
- Reference and Welingak Website conversion rate is high
- Google 50% conversion rate
- Mumbai, Thane & outskirts conversion rate is higher
- the leads which has specialization chances of conversion is bit higher

Model Achieved testing and training sensitivity score 79 and added lead scores to test leads