LEAD SCORE CASE STUDY

SUBMISSION BY: SANMITHA S J, ANJALI BHALLA

Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned 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%

Steps followed:

- Importing the necessary libraries for analysis
- Data cleaning
- Data preparation
- Exploratory Data Analysis
- Train and Test data split
- Feature Scaling
- Model Building
- Model Evaluation
- Applying the best results obtained from the Train set on the Test set

Problem Solving Approach

Data Sourcing , Cleaning and Preparation

- Read and inspect the Data
- Treating missing values
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Creating Dummy variables

Splitting Train and Test Sets and Scaling the data

- Splitting data into train and test set.
- Feature Scaling of Numeric data

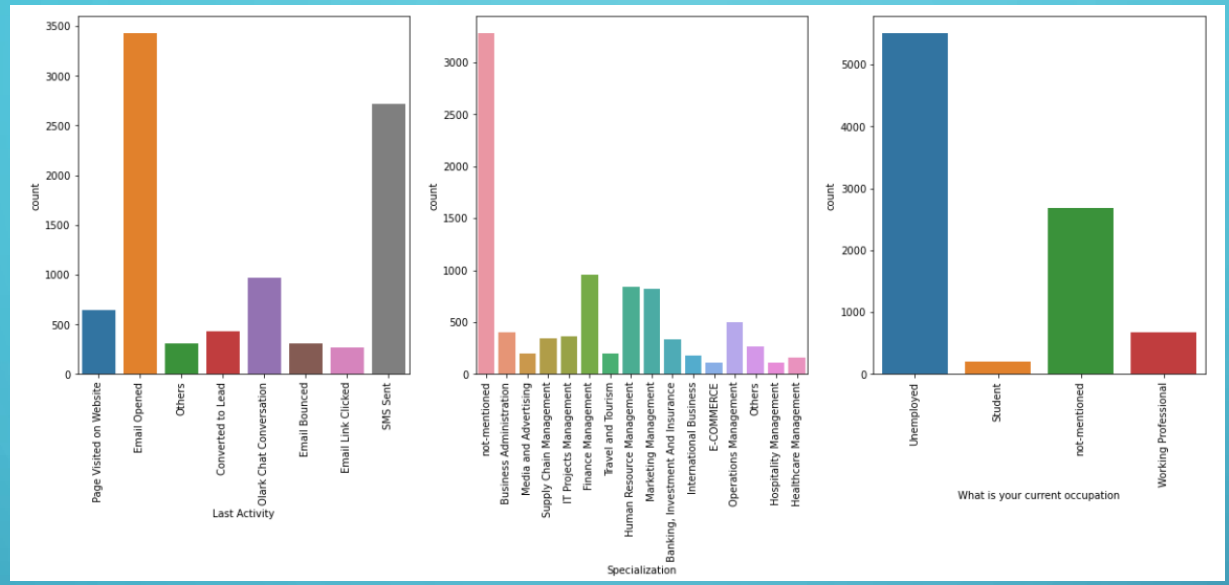
Model Building

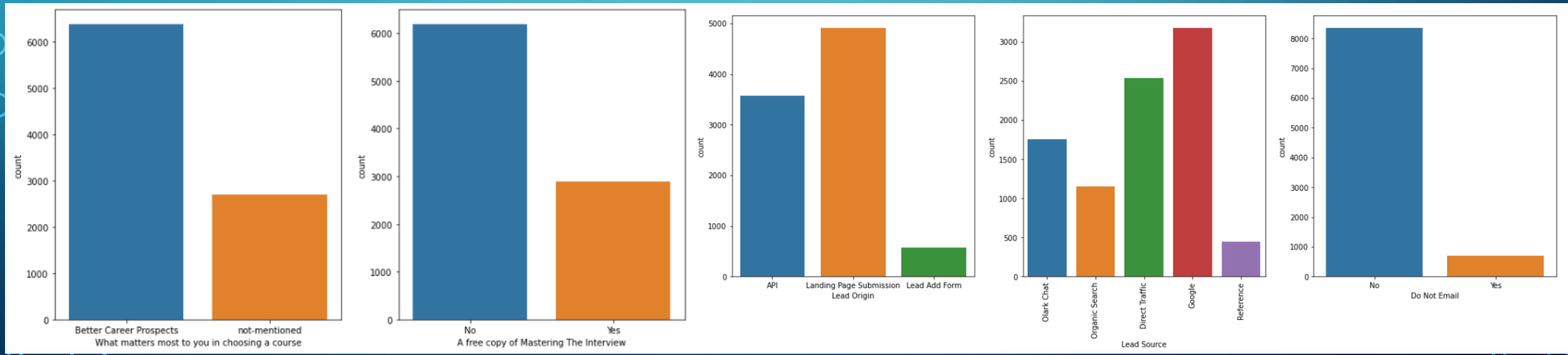
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

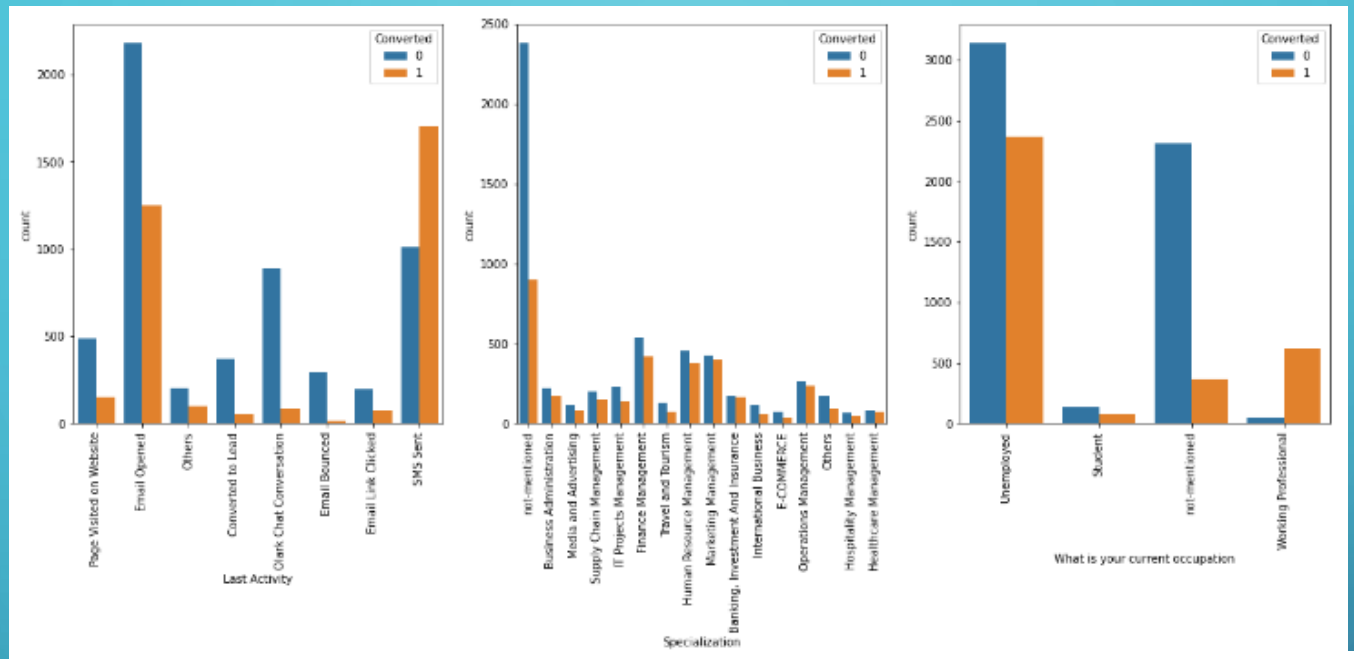
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.

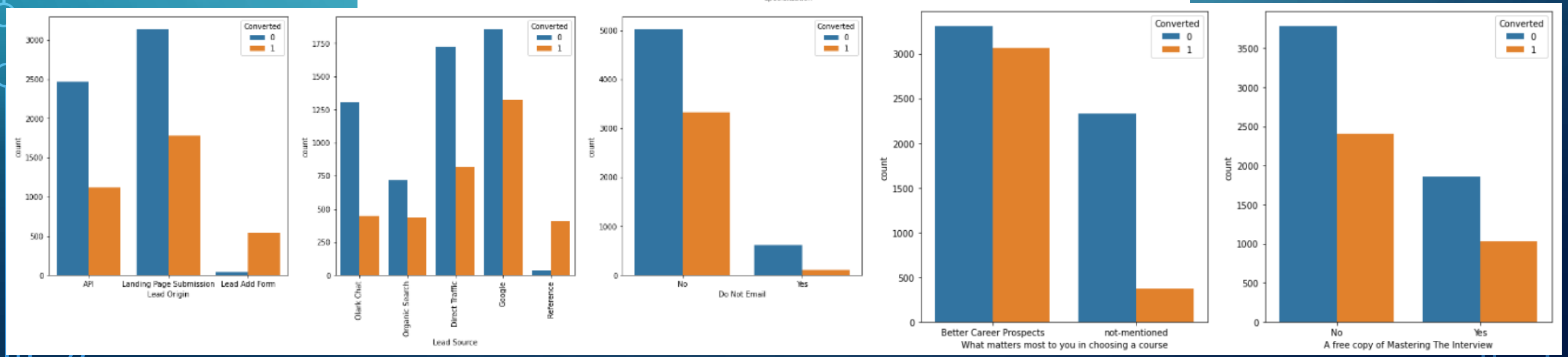
EXPLORATORY DATA ANALYSIS (univariate analysis)





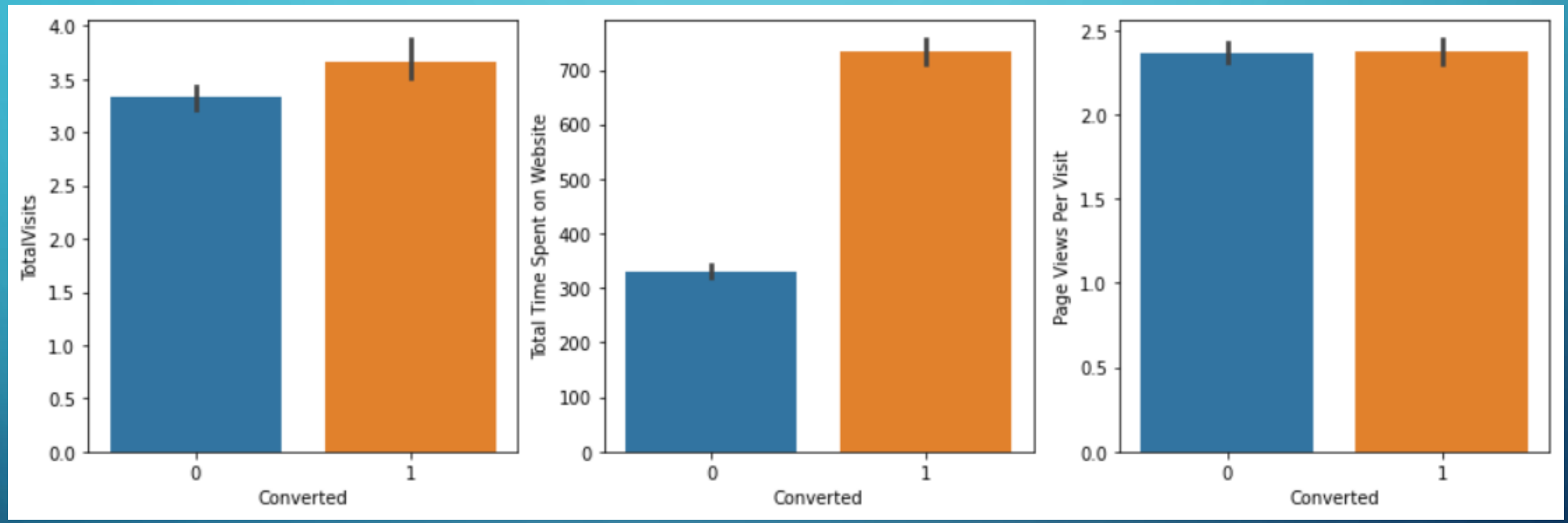
EXPLORATORY DATA ANALYSIS (bivariate analysis)





EXPLORATORY DATA ANALYSIS (heat-map of numerical variables)

Numerical variables effect on Conversion – target variable



Building model with RFE = 20

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6333
Model Family:	Binomial	Df Model:	17
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2541.5
Date:	Tue, 25 Oct 2022	Deviance:	5082.9
Time:	15:40:55	Pearson chi2:	6.38e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6384	0.199	-13.255	0.000	-3.029	-2.248
TotalVisits	9.7931	2.420	4.047	0.000	5.050	14.536
Total Time Spent on Website	4.5208	0.168	26.891	0.000	4.191	4.850
Page Views Per Visit	-1.7707	0.575	-3.081	0.002	-2.897	-0.644
Lead Origin_Landing Page Submission	-0.8011	0.132	-6.062	0.000	-1.060	-0.542
Lead Origin_Lead Add Form	4.6351	0.541	8.564	0.000	3.574	5.696
Lead Source_Google	0.2813	0.082	3.411	0.001	0.120	0.443
Lead Source_Olark Chat	1.3546	0.148	9.162	0.000	1.065	1.644
Lead Source_Reference	-1.4231	0.581	-2.450	0.014	-2.562	-0.285
Do Not Email_Yes	-1.3266	0.186	-7.138	0.000	-1.691	-0.962
Last Activity_Email Link Clicked	0.9295	0.242	3.834	0.000	0.454	1.405
Last Activity_Email Opened	1.3763	0.135	10.200	0.000	1.112	1.641
Last Activity_Others	1.6005	0.218	7.345	0.000	1.173	2.028
Last Activity_Page Visited on Website	0.8175	0.193	4.236	0.000	0.439	1.196
Last Activity_SMS Sent	2.5419	0.138	18.462	0.000	2.272	2.812
Specialization_not-mentioned	-0.8261	0.128	-6.465	0.000	-1.077	-0.576
What is your current occupation_Working Professional	2.4223	0.194	12.466	0.000	2.041	2.803
What matters most to you in choosing a course_not-mentioned	-1.2537	0.089	-14.062	0.000	-1.428	-1.079

	Features	VIF
4	Lead Origin_Lead Add Form	4.68
7	Lead Source_Reference	4.45
3	Lead Origin_Landing Page Submission	4.16
2	Page Views Per Visit	4.09
10	Last Activity_Email Opened	3.01
14	Specialization_not-mentioned	2.92
13	Last Activity_SMS Sent	2.81
6	Lead Source_Olark Chat	2.20
1	Total Time Spent on Website	2.19
0	TotalVisits	1.94
5	Lead Source_Google	1.93
16	What matters most to you in choosing a course_...	1.62
12	Last Activity_Page Visited on Website	1.44
15	What is your current occupation_Working Profes...	1.22
11	Last Activity_Others	1.19
8	Do Not Email_Yes	1.19
9	Last Activity_Email Link Clicked	1.14

Final model will have variables with p-value < 0.05 and VIF < 5

Model Evaluation using a random cut-off as 0.5

Confusion matrix:

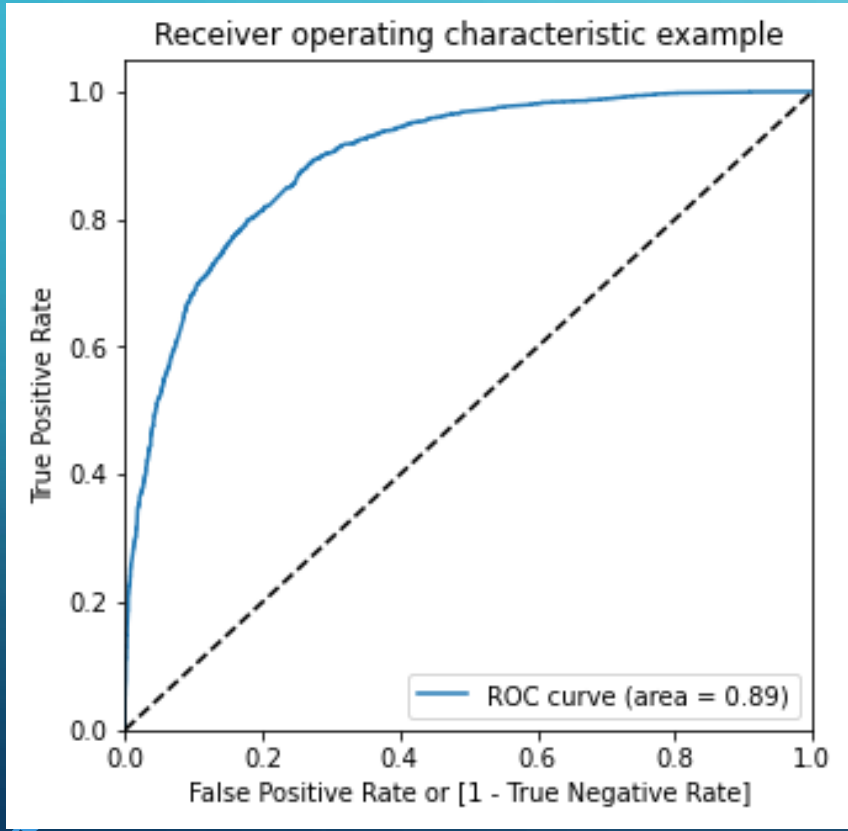
TP – True positive
FP – False positive
TN – True negative
FN – False negative

	TP	FP
	3452	453
	709	1737
	FN	TN

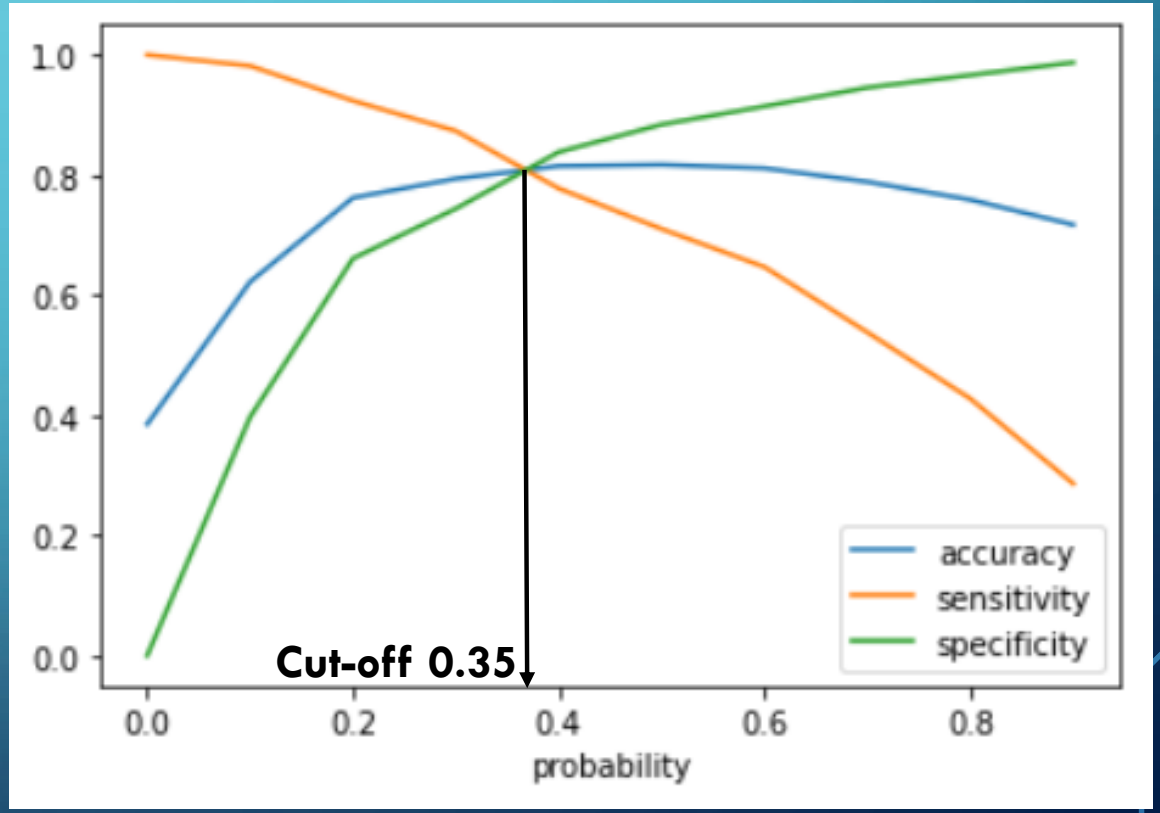
Accuracy = 81.70%
Sensitivity = 71.01%
Specificity = 88.39%

ROC – curve and Probability graph to check the Significance of the model

ROC - curve



Probability graph



Model Evaluation using a cut-off as 0.35 from the graph

Confusion matrix:

TP – True positive FP – False positive
TN – True negative FN – False negative

	TP	FP
	3155	750
	471	1975
	FN	TN

Accuracy = 80.77%
Sensitivity = 80.74%
Specificity = 80.79%

Model Evaluation using a random cut-off as 0.35

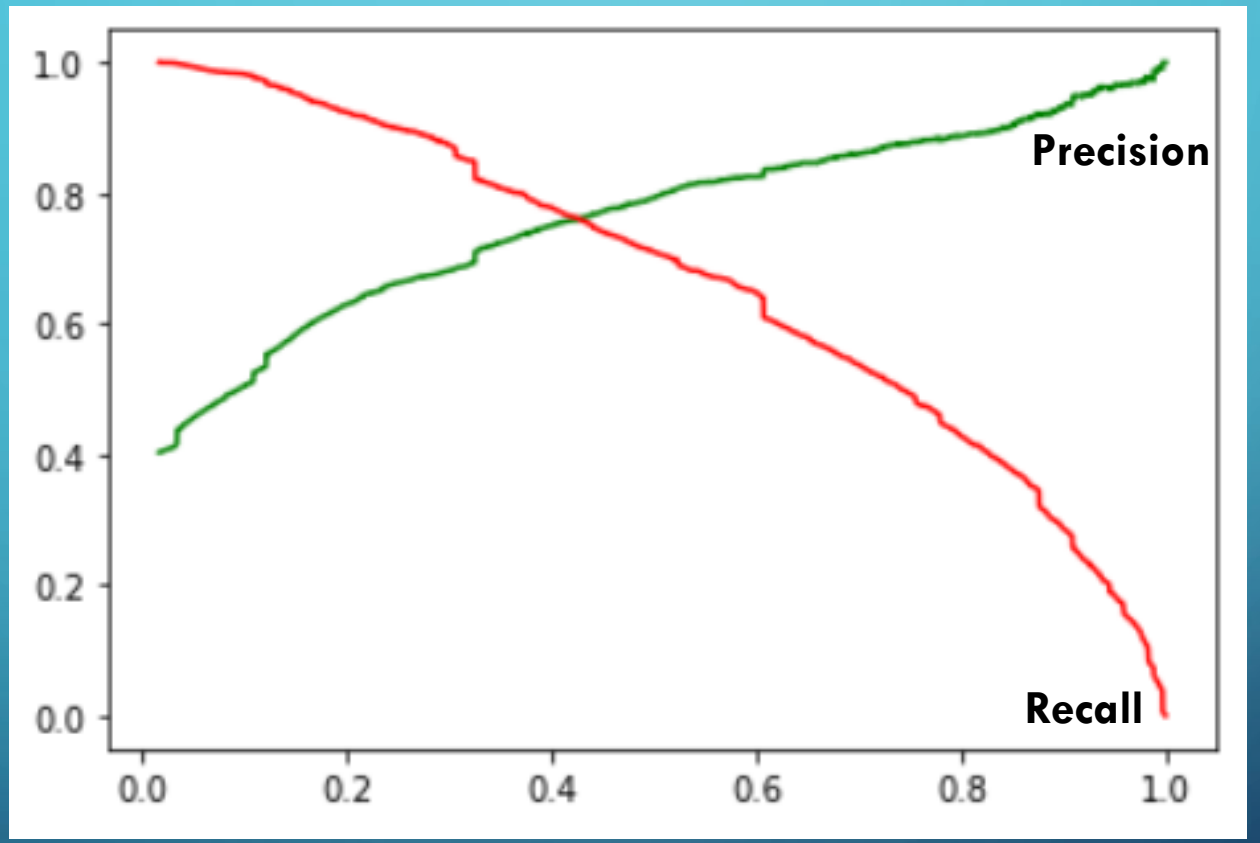
Confusion matrix:

TP – True positive FP – False positive
TN – True negative FN – False negative

```
                TP      FP
array([[1419,   315],
       [ 194,   795]],
        FN      TN
```

Accuracy = 81.31%
Sensitivity = 80.38% = Recall
Specificity = 81.83%
Precision = 71.62%

Precision – Recall graph



Lead-Score generated

	Prospect ID	Converted	Converted_prob	Lead_Score
0	3271	0	0.053108	5
1	1490	1	0.962960	96
2	7936	0	0.048342	5
3	4216	1	0.875446	88
4	3830	0	0.036900	4
5	1800	1	0.691565	69
6	6507	0	0.387317	39
7	4821	0	0.321334	32
8	4223	1	0.908005	91
9	4714	0	0.305396	31

Conclusion :

-- The top 3 variables contribute most towards the probability of a lead getting converted are:

- Total-Visits
- Lead Origin_Lead Add Form
- Total time spent on website

-- From our analysis we can observe that few top variables like SMS-sent, working professionals have higher conversion rates. Thus, taking a look at which specialization, current occupation, choice of course are some of the good choice of variables can result in better lead generation.

-- The sales team should work on getting new customers or leads by looking deep into people who are looking for upskilling their skills, ways to follow-up to name a few.

THANK YOU.