Lead Scoring - Presentation

Created by : Harsh Chavda and
Anuj Kumar

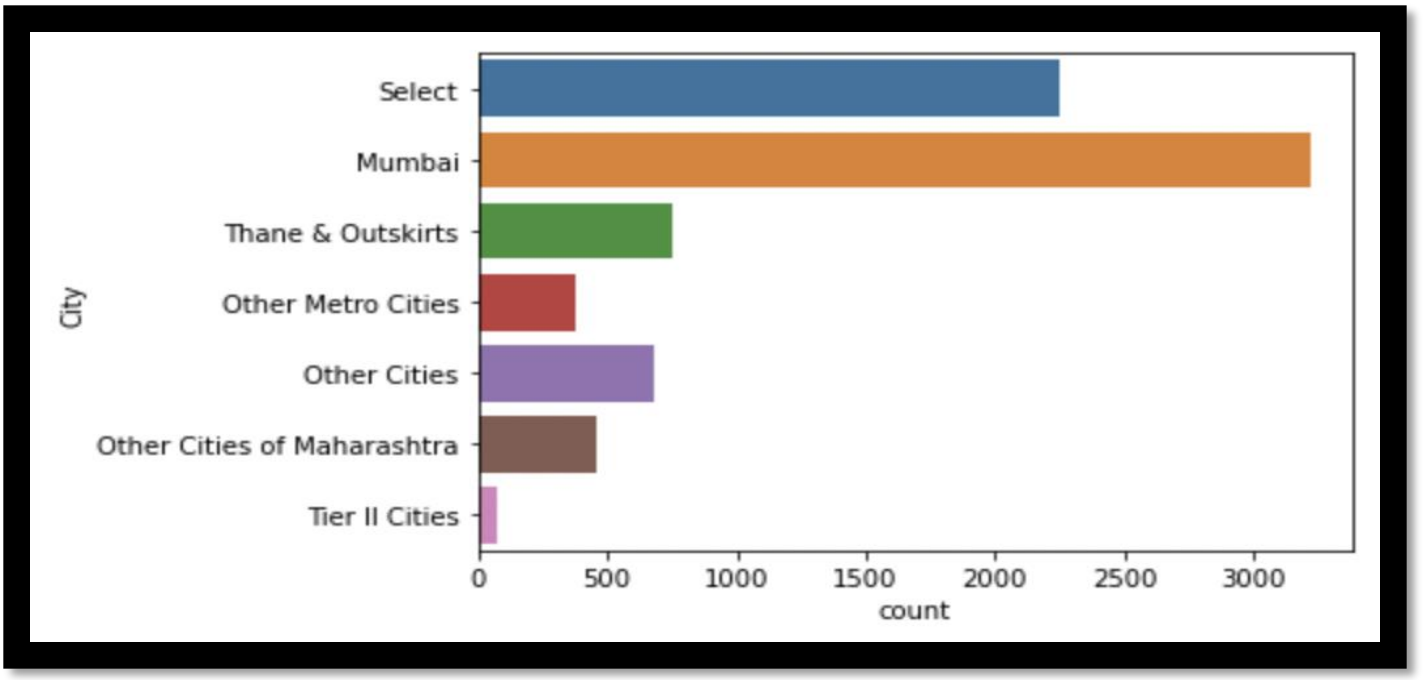
Problem Statements & Assumptions

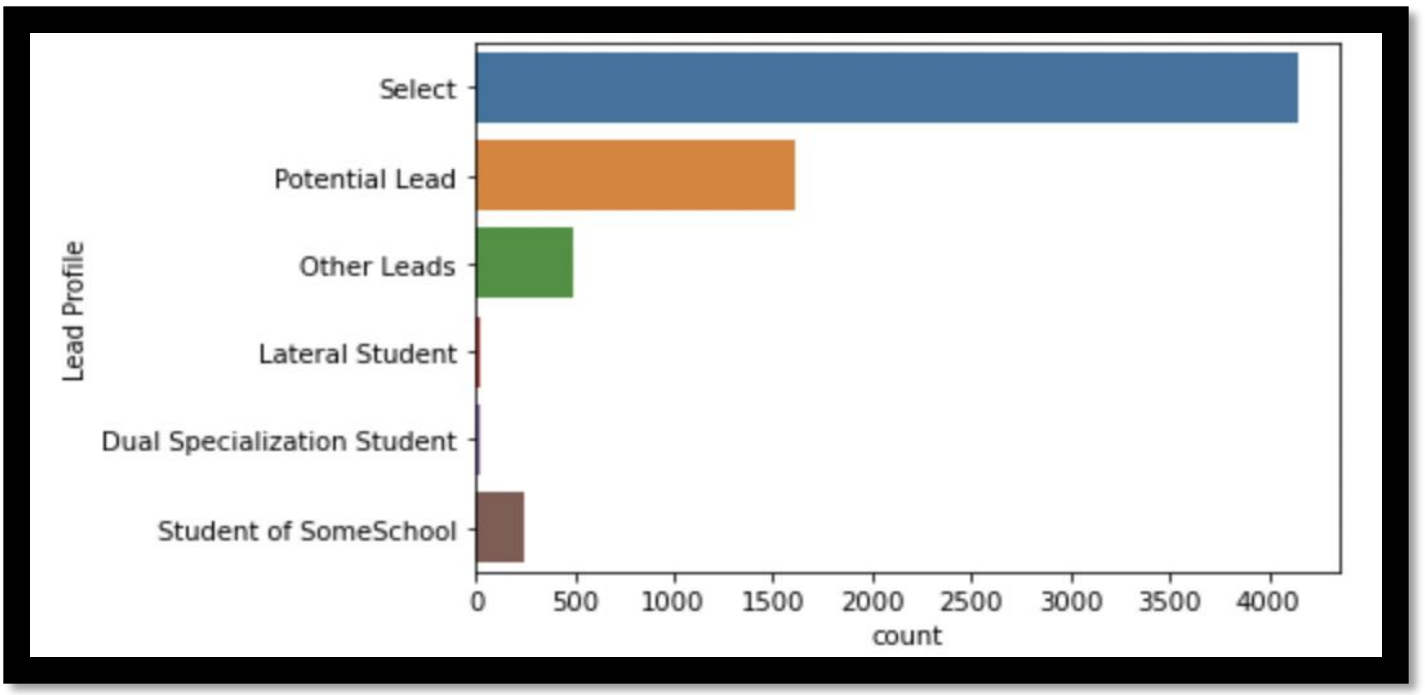
- 1. We are hired by X Education company, which is an online education company. They have their websites on multiple platforms where people can visit for info regarding offered courses.
- 2. Once a visitor provides email/phone no, they are leads which are further contacted by telecommunications team of the company.
- 3. At the current rate only 30% of leads actually buys the course (I.e. becomes a Hot-lead).
- 4. We need to make a logistic regression model which evaluates the leads and scores them on their chances of converting to hot-leads.
- 5. Our target is to raise the lead conversion rate to 80%.

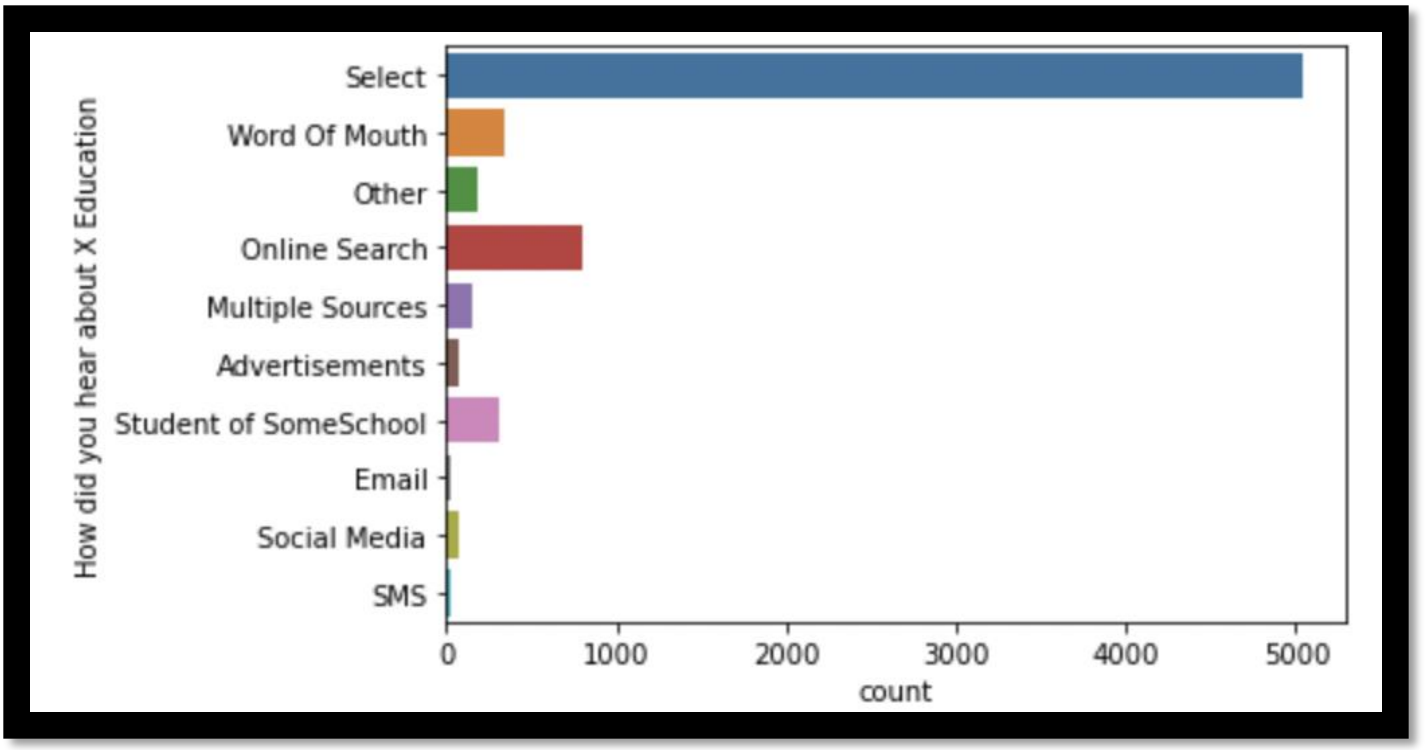
Approach

- 1. We began with Meta-Data Check, I.e. Basic descriptive stats and rows and cols of data frame.
- 2. We performed Data Cleaning and EDA simultaneously, ultimately removing cols and rows with missing values while plotting count plots of required variables.
- 3. Followed by data preparation, I.e. Fit_transform for Categorical variables and MinMaxScaler for Numerical variables.
- 4. Splitting data into train-test set, RFE and VIF were used to eliminate non-relevant variable via Automatic-Manual approach based on there p-values and VIF values.
- 5.With variables selected in the final model, it was evaluated via confusion matrix, I.e. Accuracy, Specificity and Sensitivity.
- 6. Using ROC curve the optimal cutoff point was approximated via graph I.e. 0.42, to calculate the matrices, same cutoff was used to make prediction on Test data set.
- 7. For both train and test data set, required values of accuracy were met accordingly.

Exploratory Data Analyses

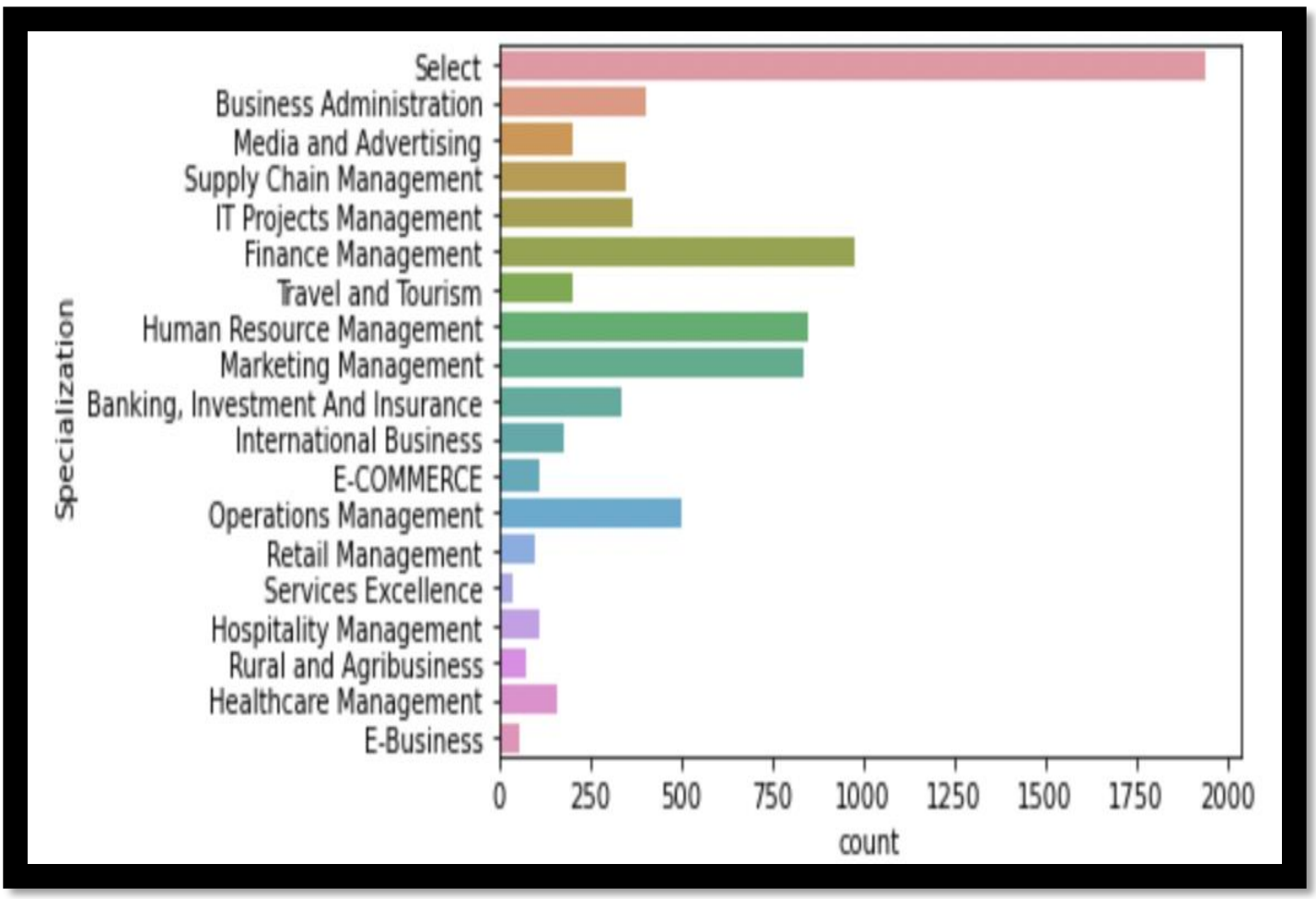






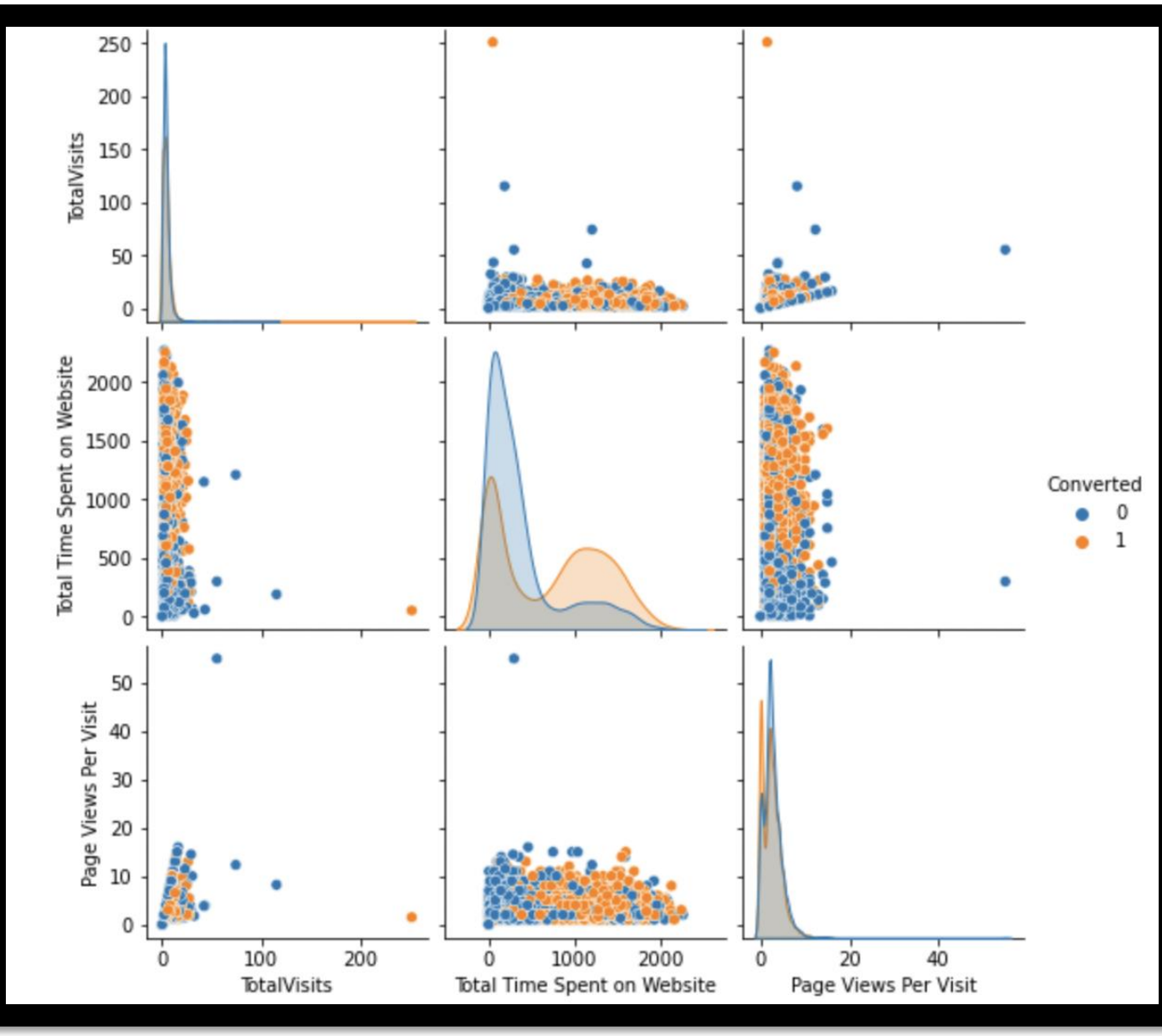
- The above mentioned graphs contains count plots of few Categorical variables among many, which were dropped due to either high missing values or low relevant variance of data contained within it.

Exploratory Data Analyses



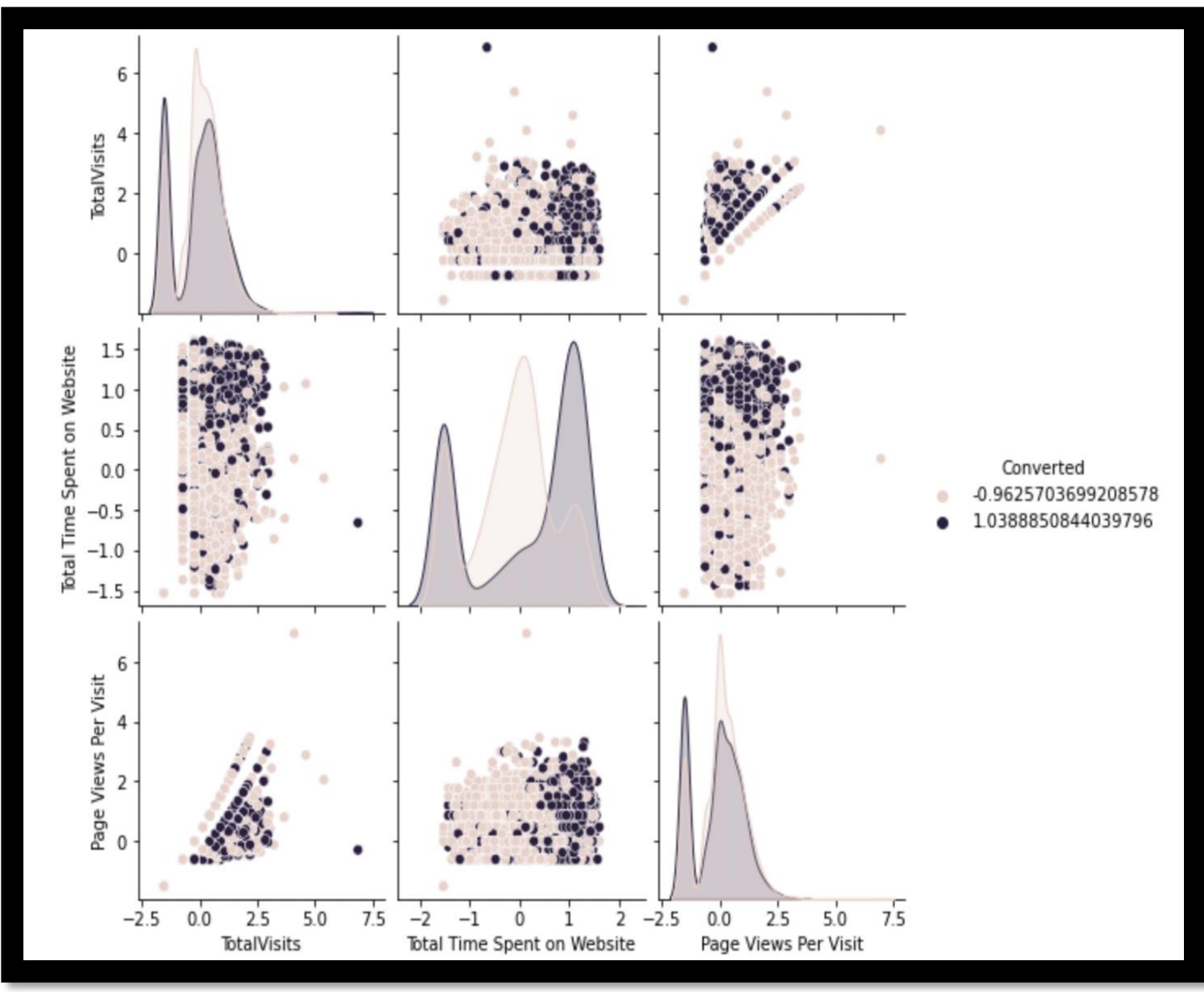
- Categorical columns with low missing value percentage were cleaned by dropping rows containing miss data, one exception to this was 'Specialization' column, this col was still kept due to it's business relevance

Exploratory Data Analyses



- Pair-plot showing trends of correlation among Numerical variables.
- Unique trends were observed each pair.

Exploratory Data Analyses



- Change in correlation observed after Fit_transforming the variables

Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Tue, 12 Apr 2022	**Deviance:**	4158.1
Time:	21:40:41	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

- All the p-values are now in the appropriate range.

- Final Model#5 with all the selected variables with p-values below 0.5.

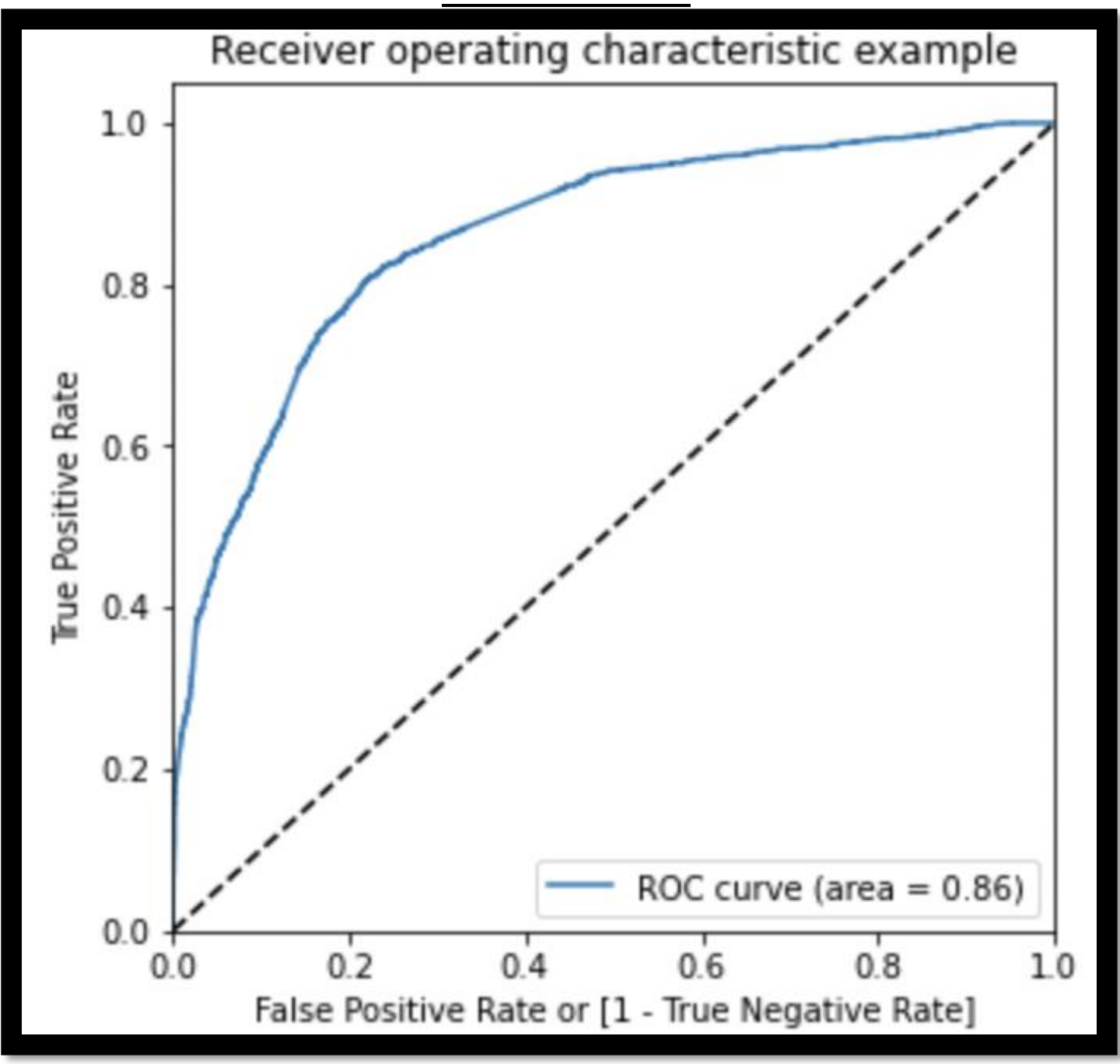
Results

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

- Final Model#5 with VIF values below 5 for selected variables.

Results

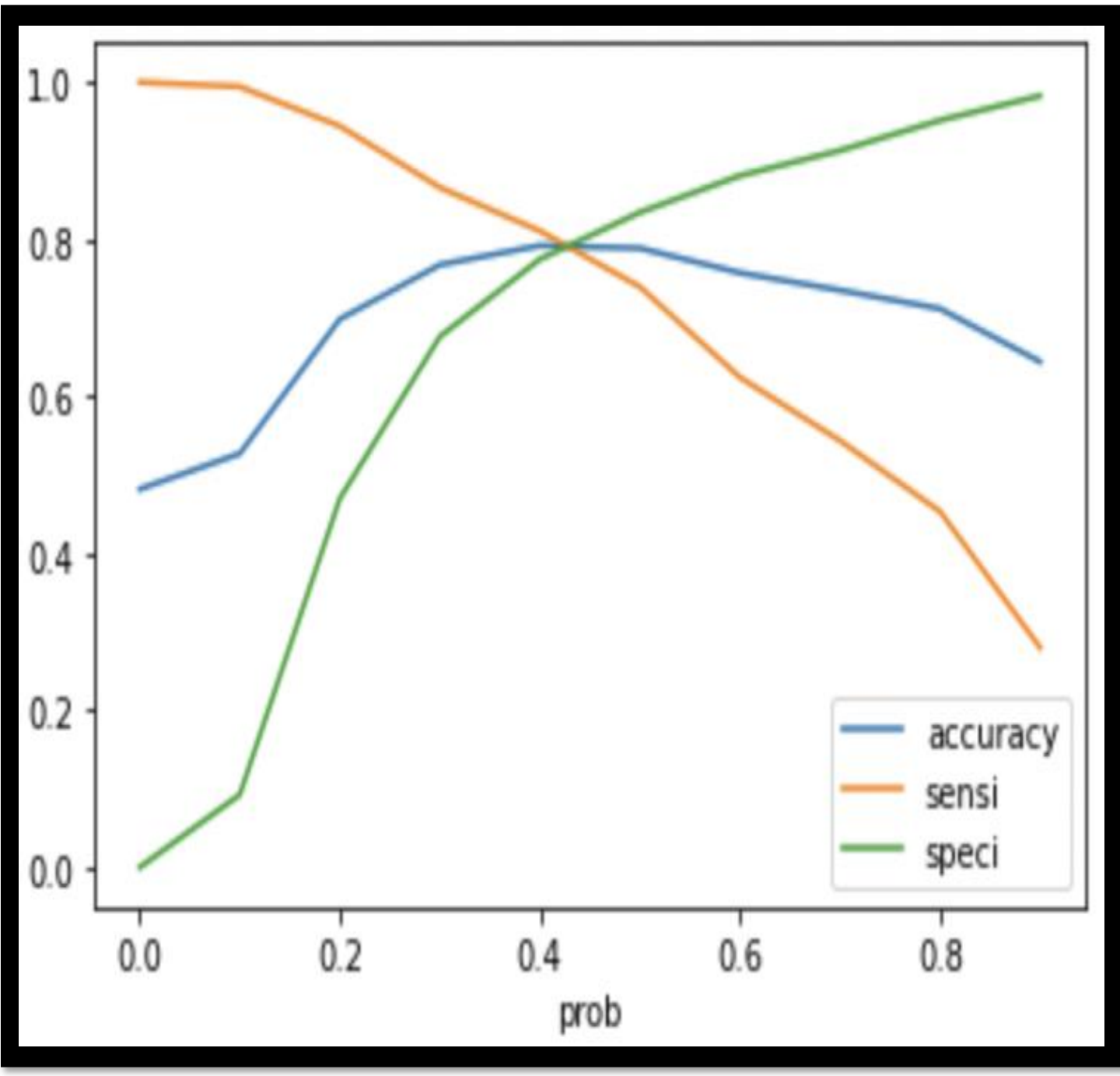
- ROC Curve



- The ROC's area under the curve is 0.86, which is excellent. As a result, we appear to have a good model. To discover the best cutoff point, we'll look at the sensitivity and specificity tradeoff.

Results

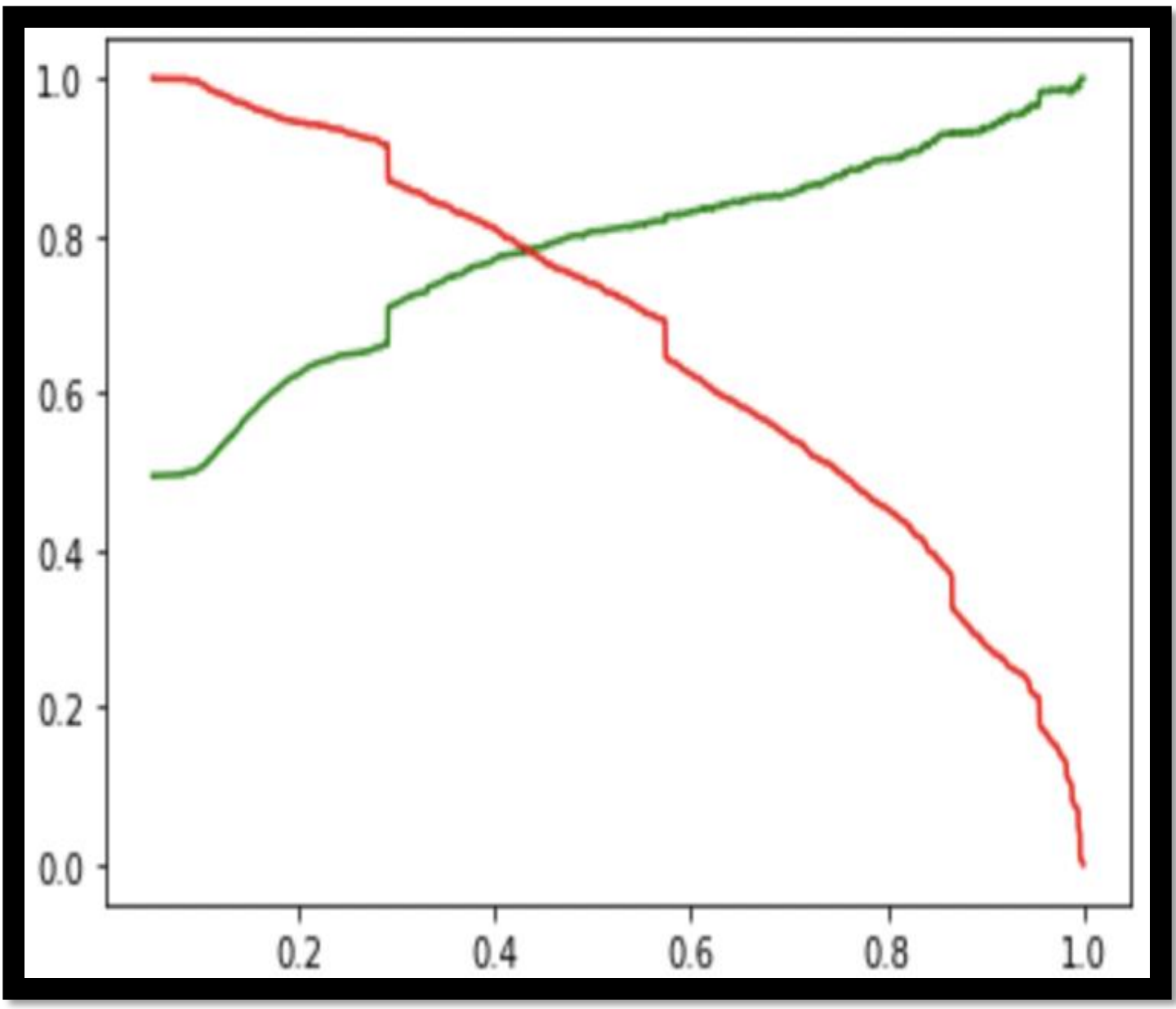
- Cutoff plot



- We chose 0.42 as our cutoff as three metrics meet there.

Results

- Precision Recall curve



- We revised our cutoff to 0.44 accordingly as per the cutoff value of the curve.

Results

- Train Data
- Accuracy:- 79.08%
- Specificity:- 79.33%
- Sensitivity:- 78.84%
- Test Data
- Accuracy:- 78.45%
- Specificity:- 77.94%
- Sensitivity:- 78.91%

- Confusion matrix parameters for both Train and Test data set were calculated with appropriate cutoff value, results meet the requirement of the education company; I.e. ~80% accuracy.