Lead Scoring Case Study

Presented by

Akshay Magotra

&

Vishal Kumar

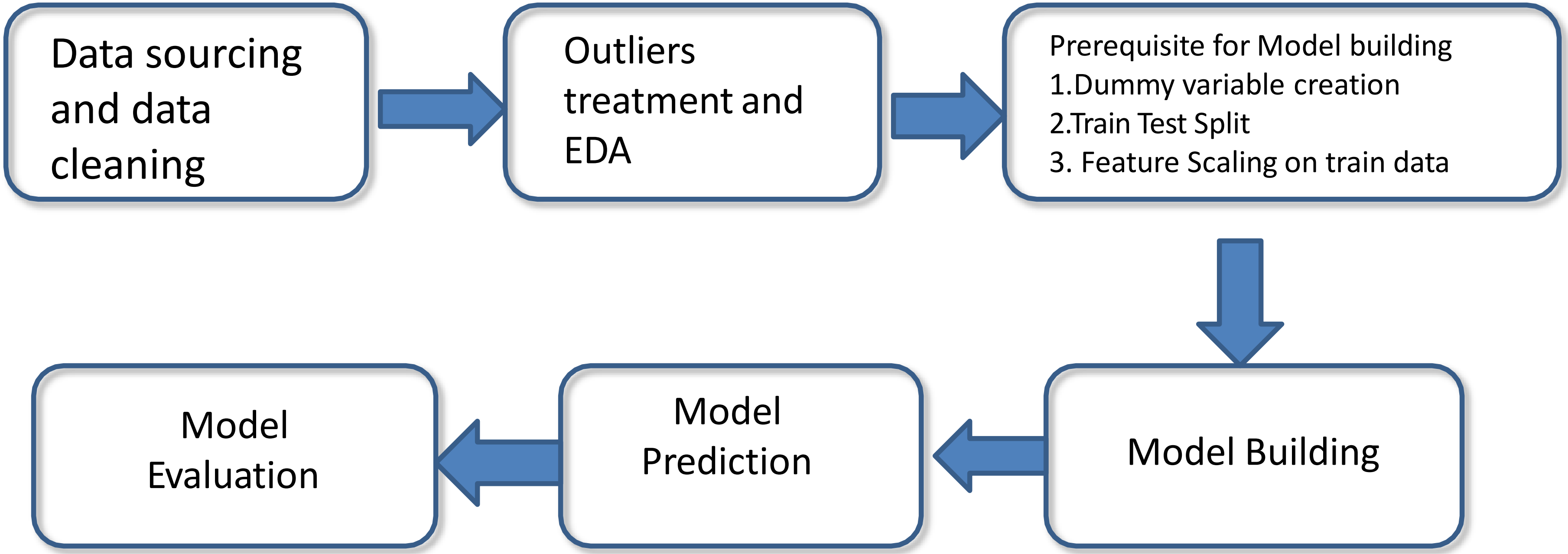
Problem Statement :

- An education company X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- The leads are acquired when people land on the website and fill up forms providing phone number or email address, also from past referrals.
- After acquiring leads from various sources, marketing team starts lead conversion process by making calls and writing emails. The company gets a lot of leads but
- the current lead conversion rate at X education is poor, around 30%.
- X education wishes to make lead conversion process more efficient by finding potential leads or hot leads.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objectives to be achieved :

- ➢ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads
- ➢ Model to be flexible enough to accommodate future requirements.

Analysis Approach



Data sourcing and data cleaning

Outliers treatment and EDA

Prerequisite for Model building
1.Dummy variable creation
2.Train Test Split
3. Feature Scaling on train data

Model Building

Model Prediction

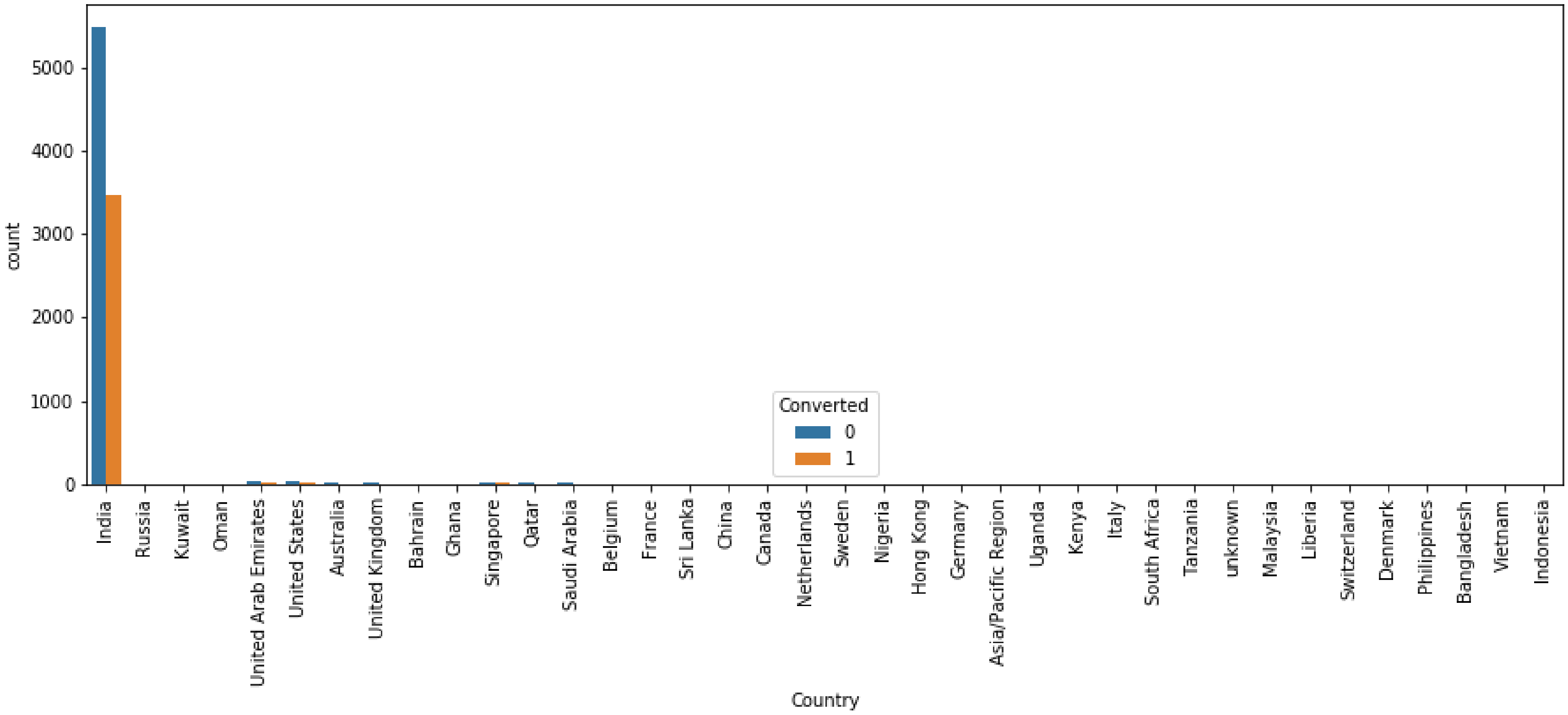
Model Evaluation

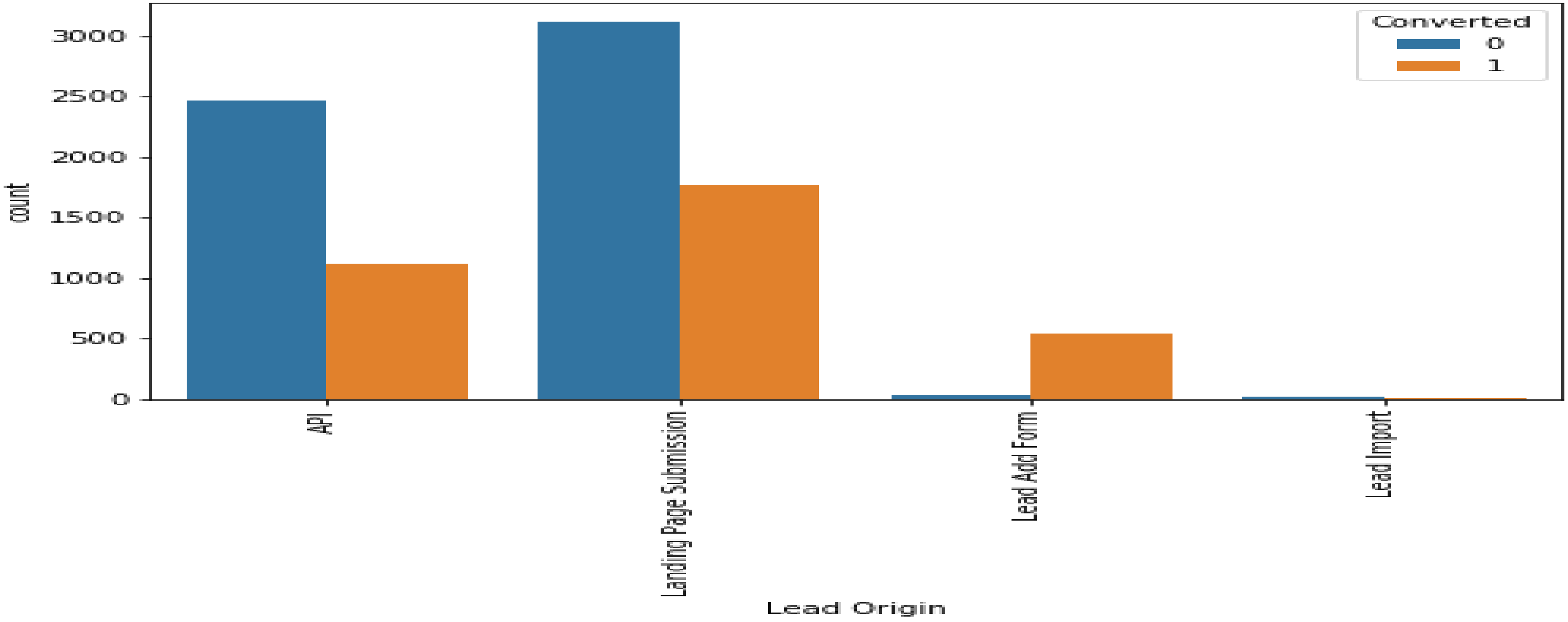
Data sourcing and Exploration :

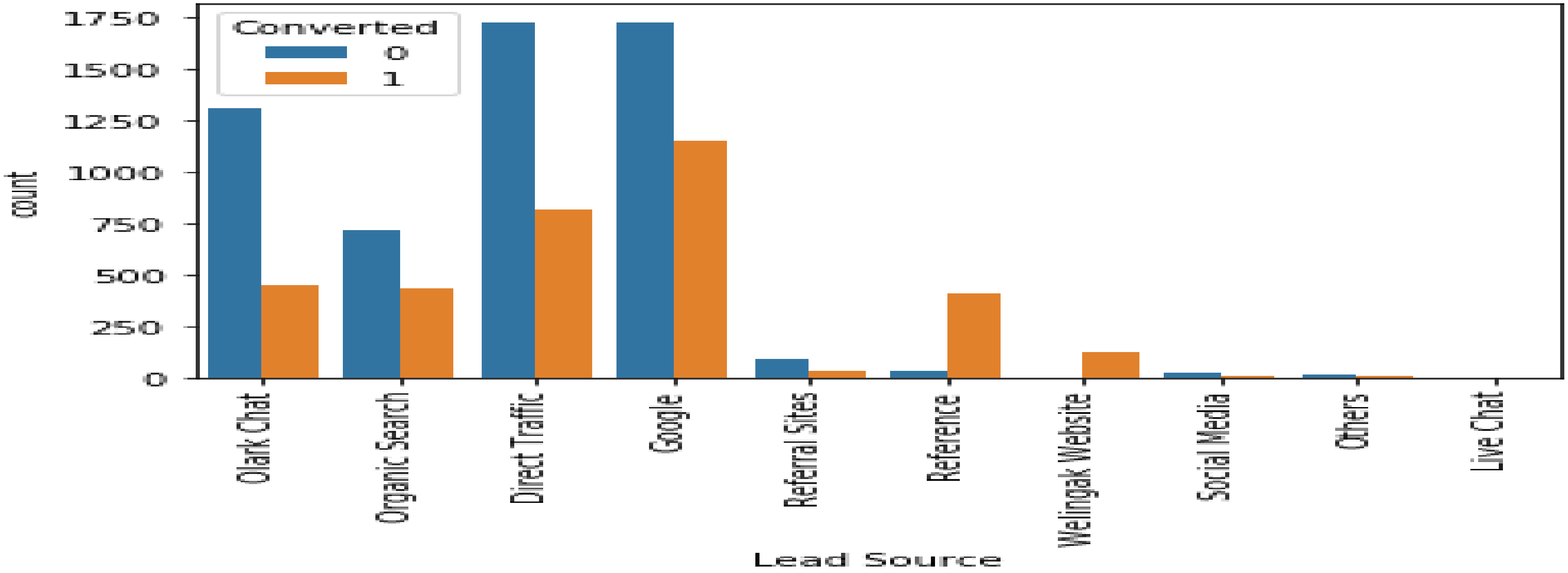
Got some useful insights after data exploration.

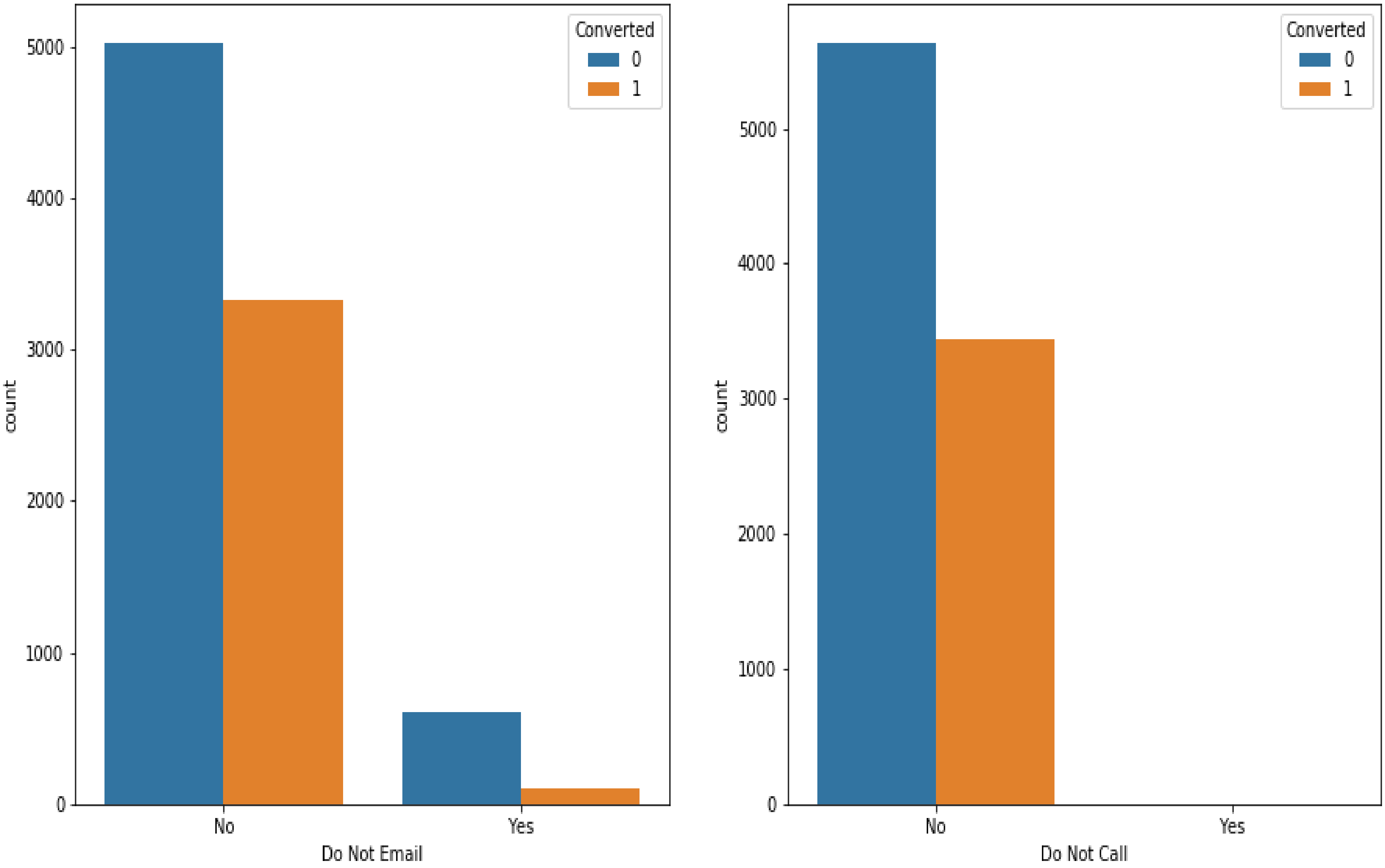
1. Columns with high % missing values, columns with unique values, columns with ‘Select’ values which is nothing but null, rows with missing values, removed columns which were least influential in analysis ex. Country
2. Imputed null values for categorical variables.
3. Few features having outliers will impact the model, so treated outliers using IQR method
4. Data visualisation in univariate and bivariate analysis helped to visualise data spread across the features.

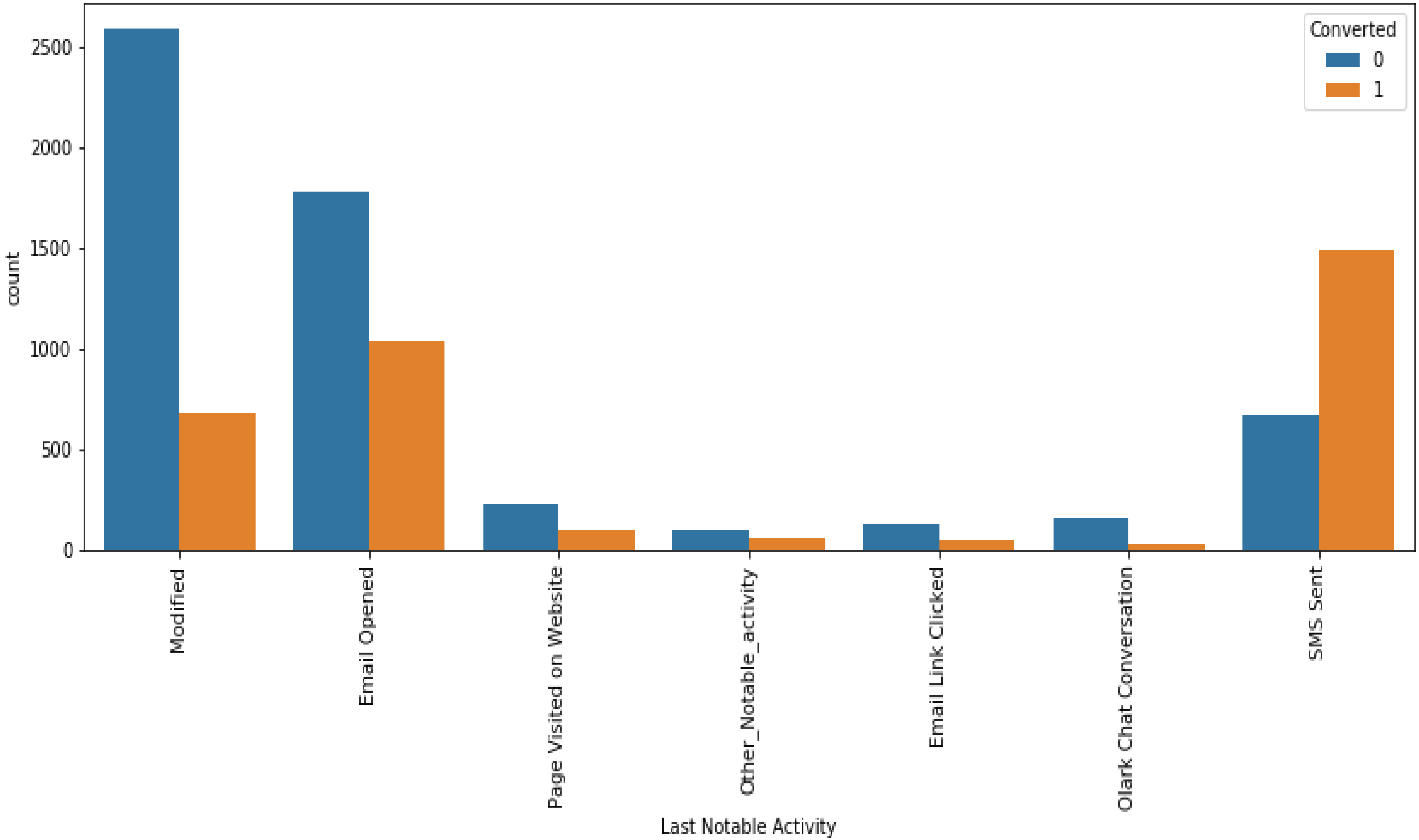
Exploratory Data Analysis




Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Social Media
Others
Live Chat
Lead Source


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call



Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Other_Notable_activity
Email Link Clicked
Olark Chat Conversation
SMS Sent
Last Notable Activity

Prerequisites for Model Building :

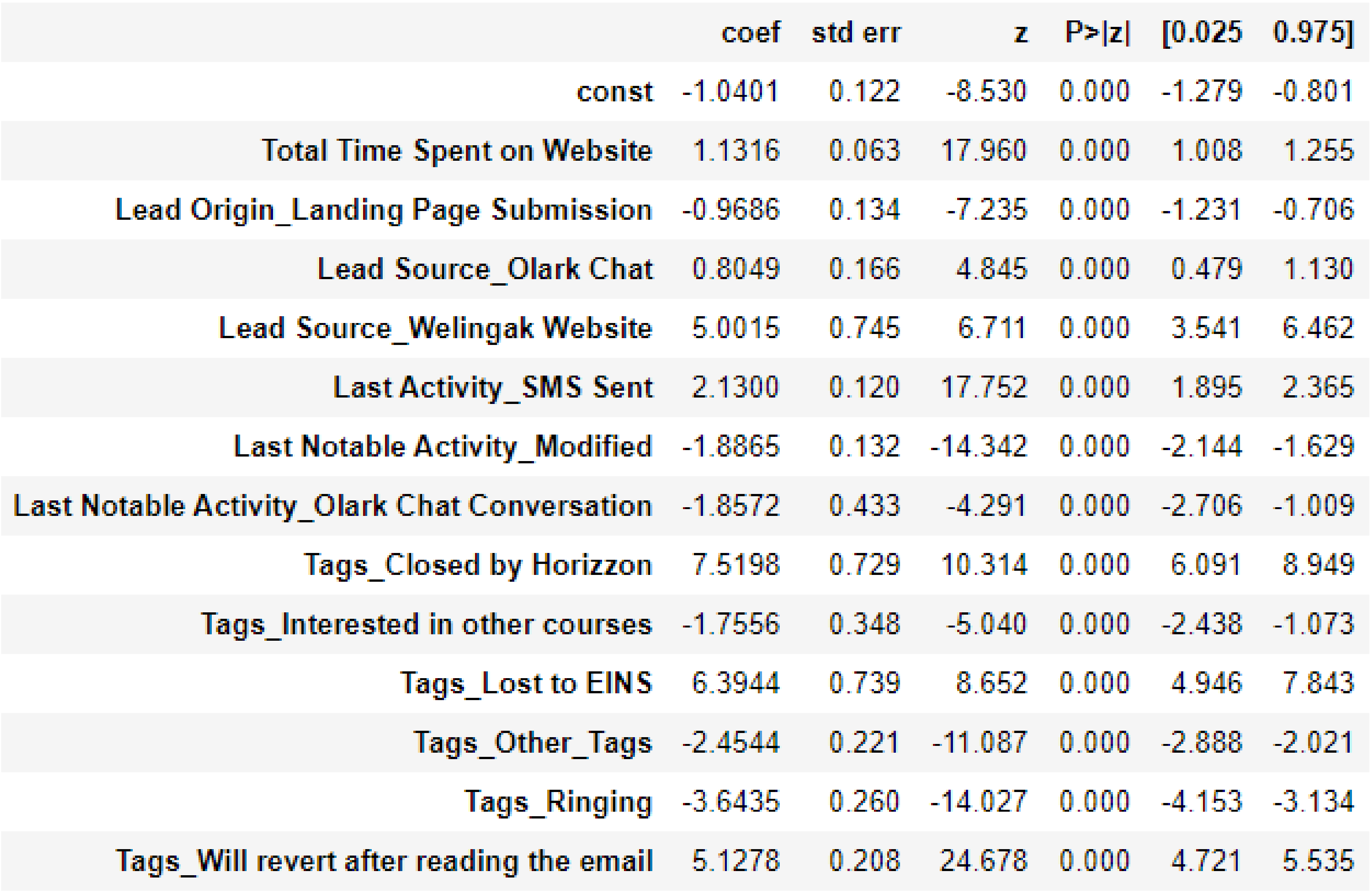
Before we start building logistic regression model, we need to prepare dataset accordingly.

1. Creating dummy variables : For categorical variables , we have to create dummy variables using pd.get_dummies().
2. Train Test Split : We build model on training data set and evaluate on test dataset. Hence , we need to split original dataset in train and test sets with 70-30 ratio.
3. Feature scaling : If numerical feature contains high range values , it will create bias in model, so we have to standardize/ scale feature . We have used min-max scaler to scale data of all features between 0 to 1.

Logistic Regression Model Building :

This is basically process of finding most significant features in the dataset which defines probability of lead conversion into paying customers.

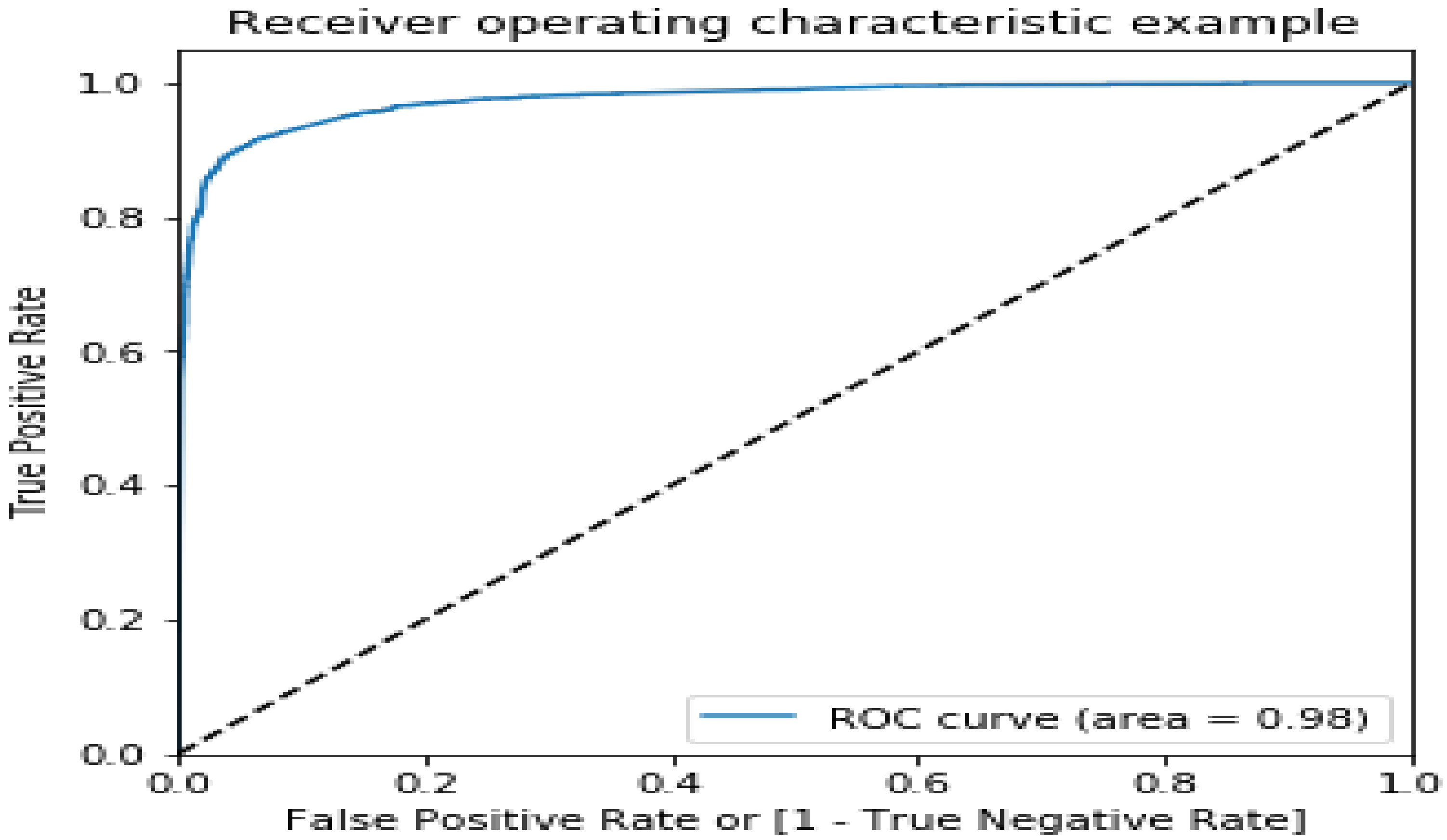
- Once we build initial model using Train data, we get coefficients of each variable present.
- We have used RFE method, to select most significant 15 features.
- After its iterative process to remove feature with high p value and VIF and build model again. We keep removing insignificant feature one by one until we get p value for all feature in model < 0.05 and VIF < 5
- Below is the final regression model we build



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0401	0.122	-8.530	0.000	-1.279	-0.801
Total Time Spent on Website	1.1316	0.063	17.960	0.000	1.008	1.255
Lead Origin_Landing Page Submission	-0.9686	0.134	-7.235	0.000	-1.231	-0.706
Lead Source_Olark Chat	0.8049	0.166	4.845	0.000	0.479	1.130
Lead Source_Welingak Website	5.0015	0.745	6.711	0.000	3.541	6.462
Last Activity_SMS Sent	2.1300	0.120	17.752	0.000	1.895	2.365
Last Notable Activity_Modified	-1.8865	0.132	-14.342	0.000	-2.144	-1.629
Last Notable Activity_Olark Chat Conversation	-1.8572	0.433	-4.291	0.000	-2.706	-1.009
Tags_Closed by Horizzon	7.5198	0.729	10.314	0.000	6.091	8.949
Tags_Interested in other courses	-1.7556	0.348	-5.040	0.000	-2.438	-1.073
Tags_Lost to EINS	6.3944	0.739	8.652	0.000	4.946	7.843
Tags_Other_Tags	-2.4544	0.221	-11.087	0.000	-2.888	-2.021
Tags_Ringing	-3.6435	0.260	-14.027	0.000	-4.153	-3.134
Tags_Will revert after reading the email	5.1278	0.208	24.678	0.000	4.721	5.535

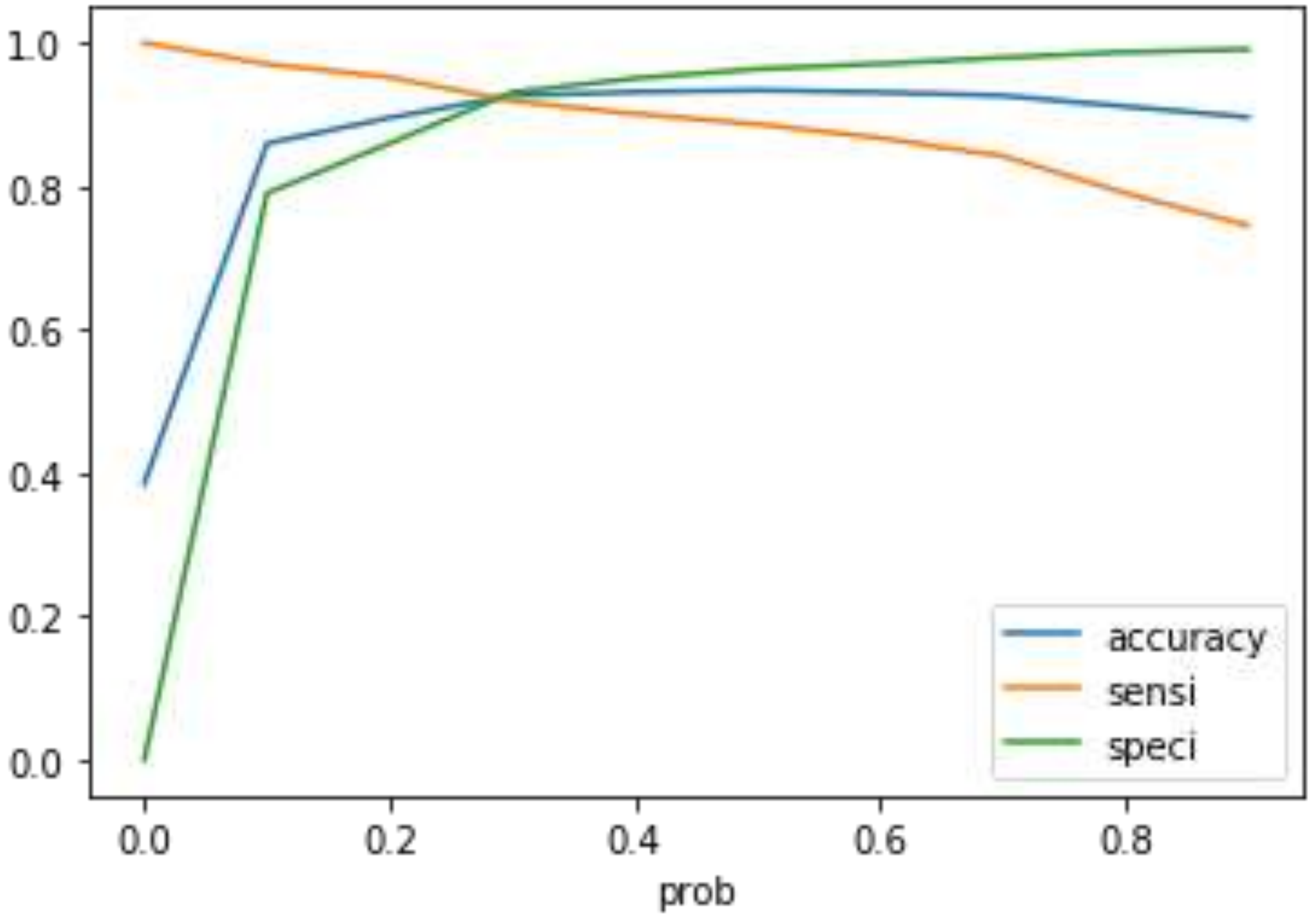
ROC Curve :

Receiver operating characteristic curve is a tool to select optimal model.
It depicts relative trade offs between true positive(benefits) and false positive(costs).
Greater the area under curve, better is the model.



Sensitivity-Specificity View:

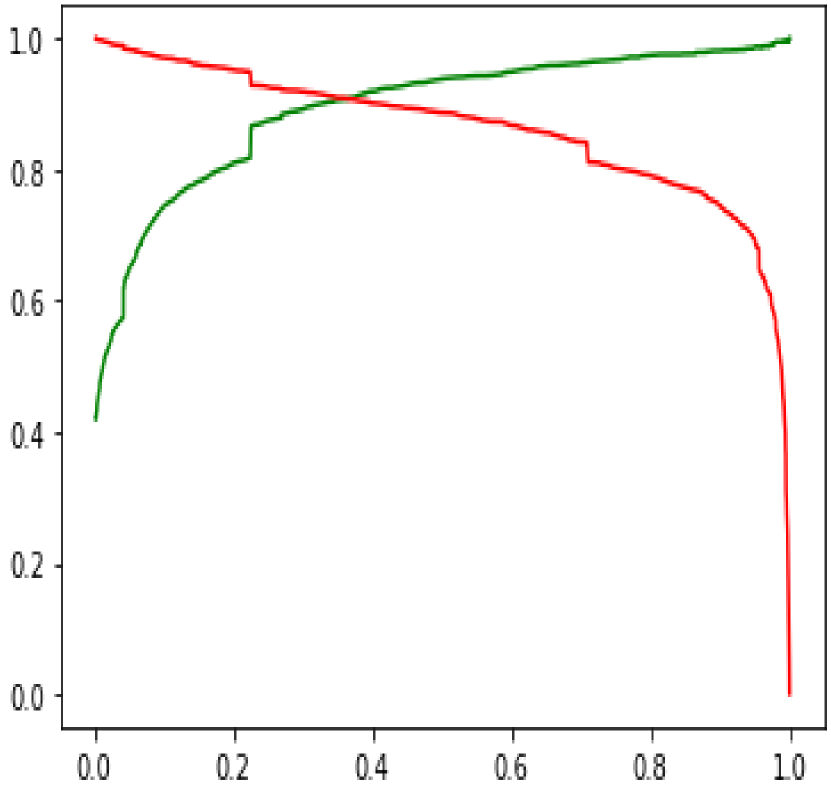
To find balanced statistics we have to find tradeoff between accuracy, sensitivity and specificity



Model Evaluation: Precision and Recall

- As per business requirement, we have chosen 0.39 as a Cut-Off value, which gives better results for both accuracy and precision
- Accuracy: ~93%
- Precision: ~89%
- Recall: ~92%
- The graph shows a trade-off

  between Precision and Recall



Model Evaluation on Test data:

- We predict the probabilities for each lead using build model on train data.
- We predict conversion column using cutoff we decided using Precision- Recall view which is converted_prob> 0.39 results in 1 else 0.
- Using actual converted column and predicted conversion column we build confusion matrix and find statistics

Statistics for test data at cutoff value for Converted_prob > 0.39

- ➢ Overall_Accuaracy :0.91
- ➢ Sensitivity :0.90
- ➢ Specificity : 0.92
- ➢ Precision - Recall : 0.87 , 0.90

Conclusion :

Final Observation:
Train Data:
1) Accuracy : 92.6%
2) Sensitivity : 91.90%
3) Specificity : 93.08%
Test Data:
1) Accuracy : 91.44%
2) Sensitivity : 89.98%
3) Specificity : 92.27%
This Model seems to be quite good and fit to predict the Conversion Rate very well and we are able to give the CEO confidence in making good calls based on this above model

- The model is prepared for prediction of the conversion of the leads. The probability values are generated by the model. The cut- off decided for the model is 0.3. All leads whose probability is generated above this threshold value can be classified as Hot Lead.