Lead Scoring Case Study

By : Surabhi Das, Sushil Sawai and Swapna Vijay

Goals of the Case Study

1. 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Steps to analyze the data

- Reading and Understanding the Data
- Data Cleaning
- Data Preparation
- Test-Train Split
- Feature Scaling
- Model Building
- Prediction on train model
- Overall Metrics

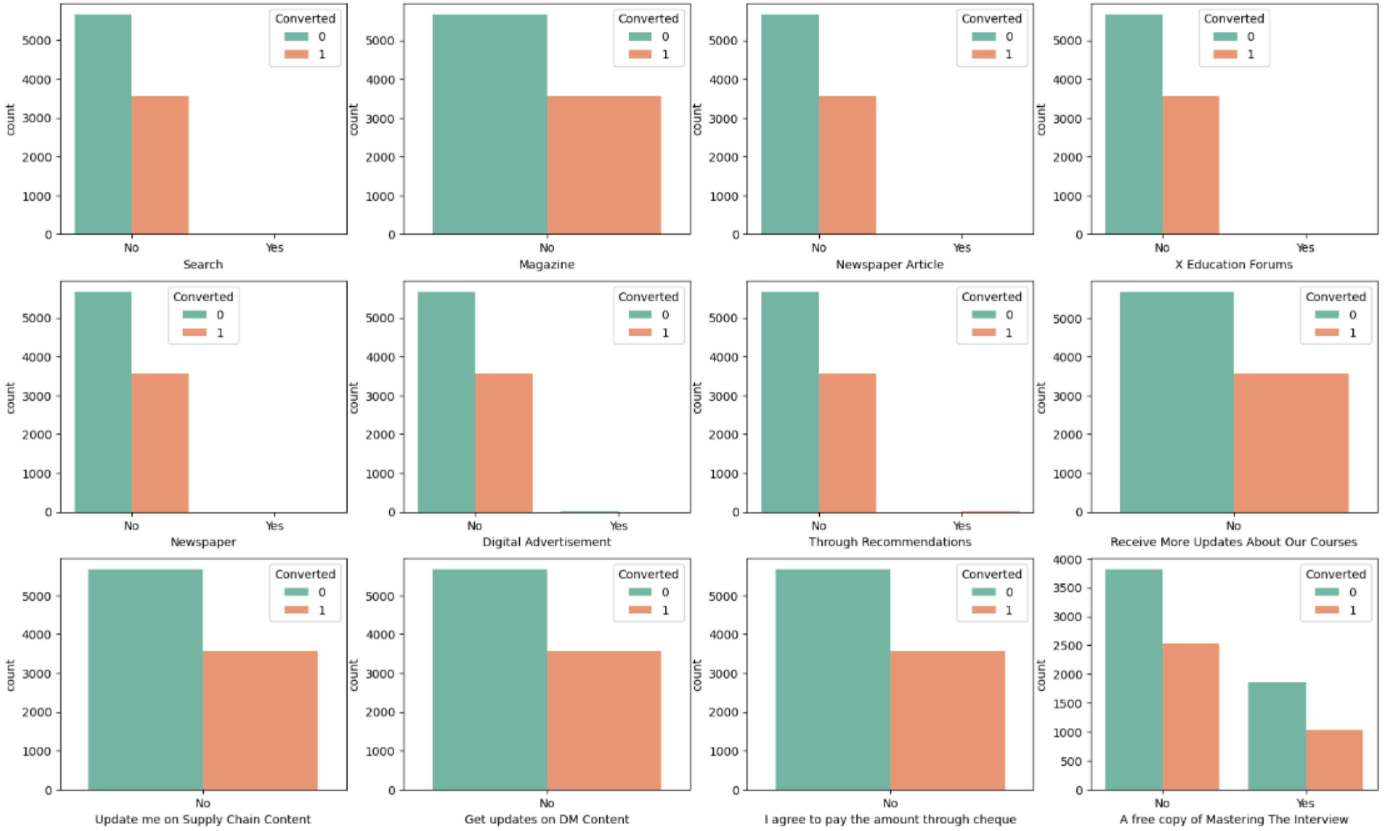
Reading and Understanding the Data

- Checking heads
- Checking shape
- Data description
- Checking info of Columns
- Checking Duplicates (Prospects ID, Lead Number)

Data Cleaning

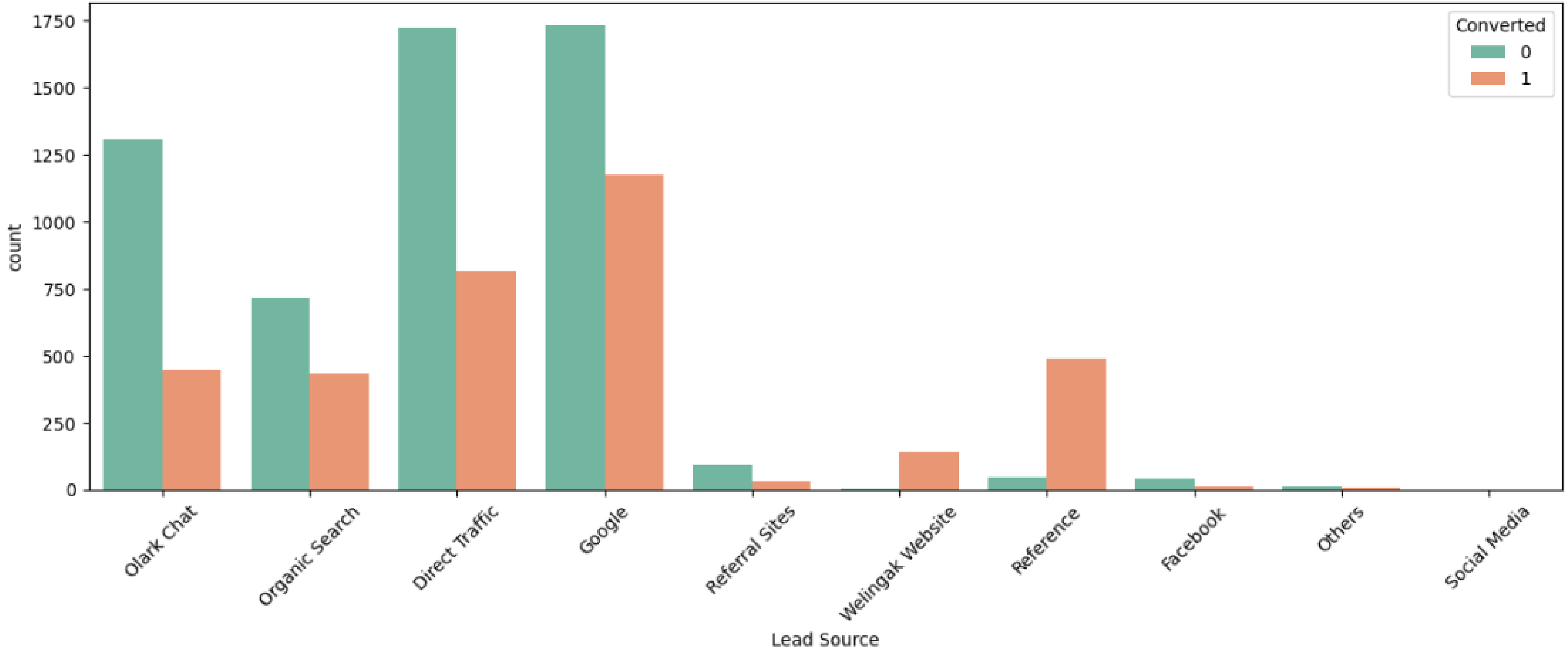
- Checking Missing Values
- Dropping Columns with missing values>=35%
- Categorical Features Analysis
- Numerical Features Analysis

Visualizing variables for Imbalancing



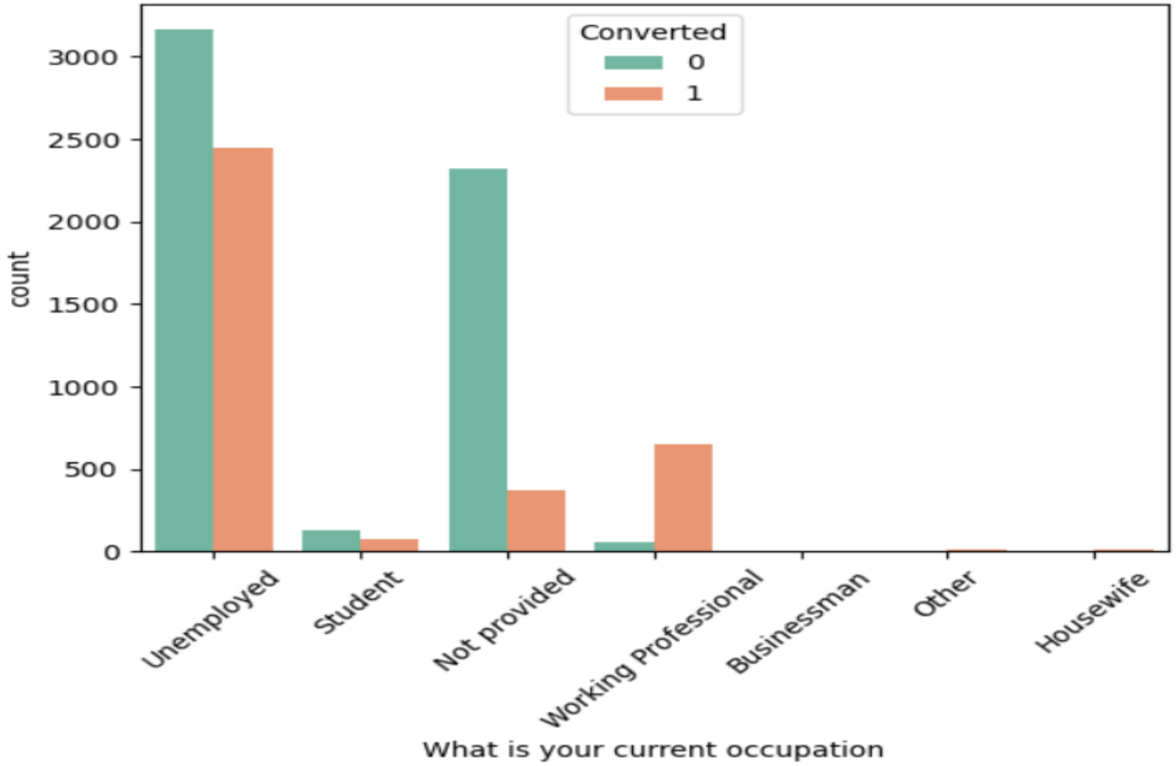
- ❖ As we can see from graph, except 'A free copy of Mastering The Interview' variable all other are highly imbalance and since 'A free copy of Mastering The Interview' is reductant variable so we will drop them.

Plotting count of Lead Source Variable based on Converted value



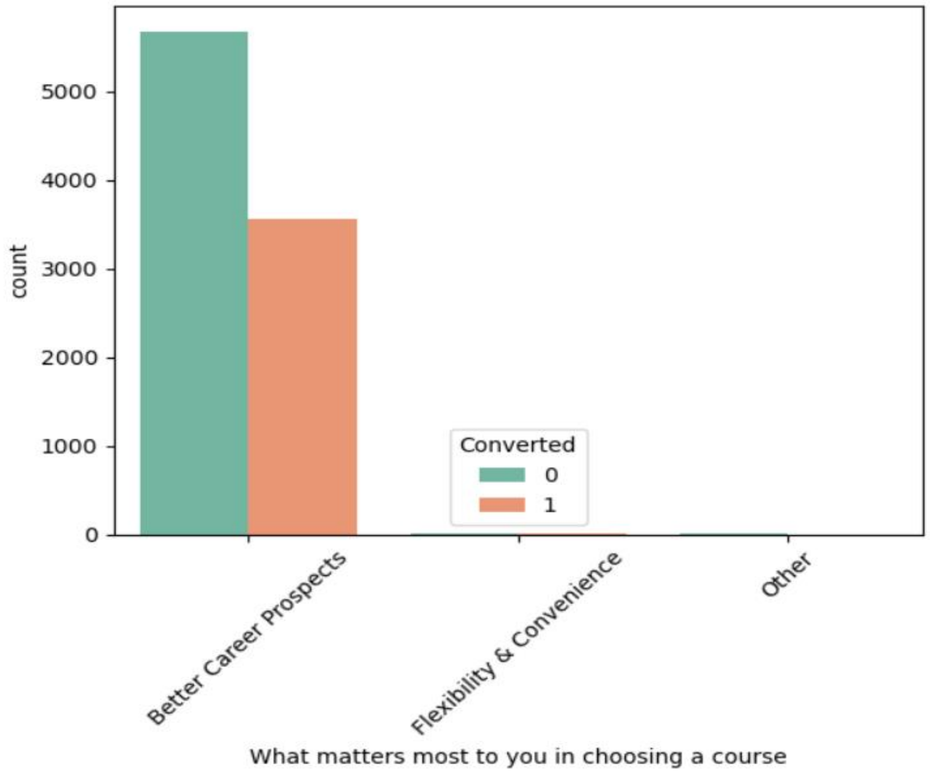
High number of leads are generated by Google and Direct Traffic and Conversion rate of Reference leads and Welingak Website leads is very high.

Plotting count of Variable based on Converted value



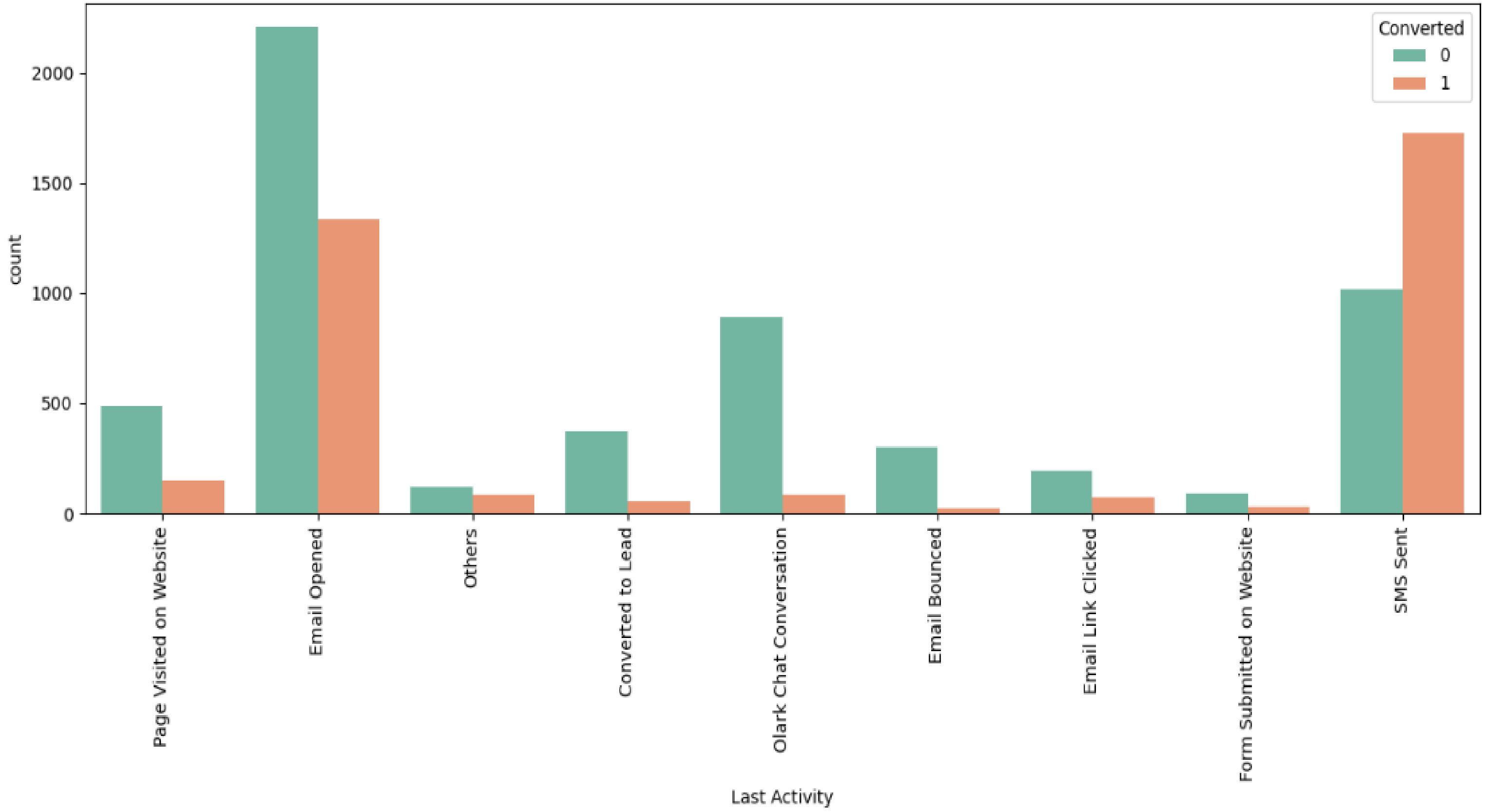
- Maximum leads generated from unemployed whose conversion rate is more than 50% and Conversion rate of working professionals is also very high.

Plotting count of Variable based on Converted value



- As we can observe that this column has low spread of variance which do not provide much insights.

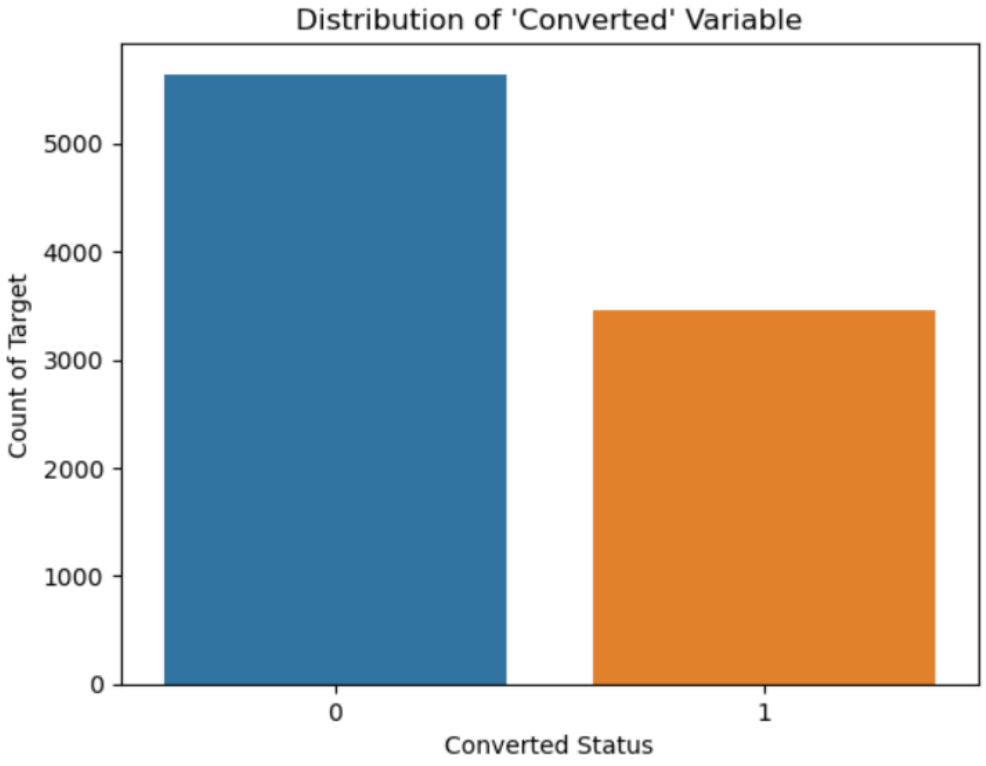
Plotting count of Last Activity Variable



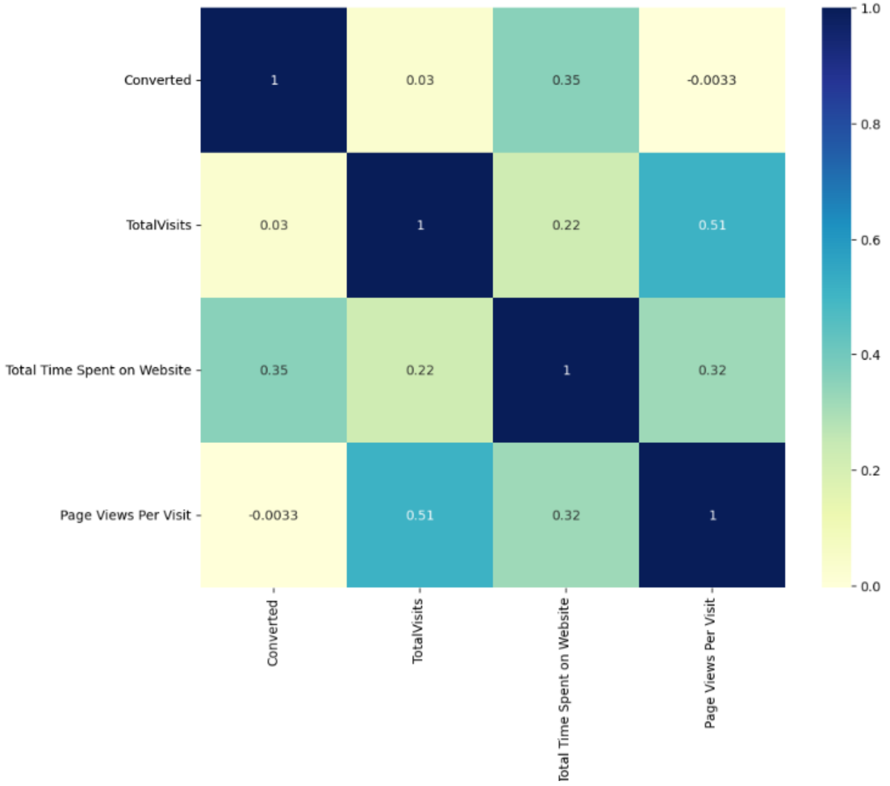
- Maximum leads are generated from last activity as Email opened but conversion rate is not that high and SMS sent as last activity has high conversion rate.

Numerical Features Analysis

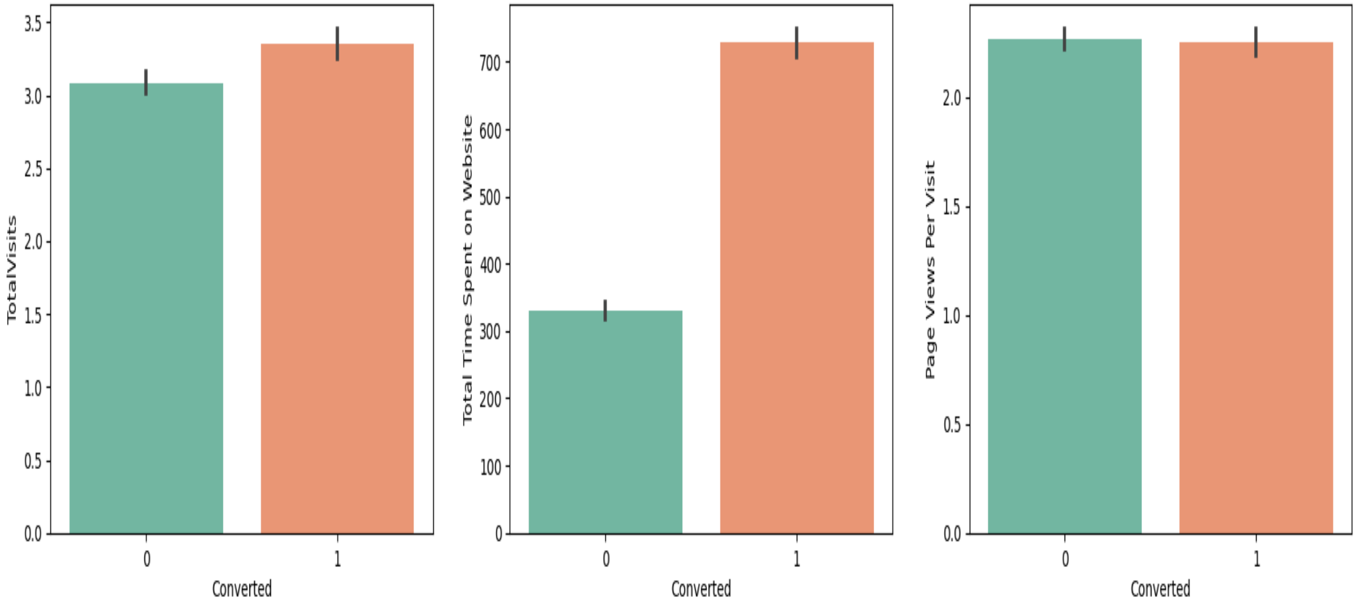
(Plotting distribution of converted variable)



Checking correlations of numeric values using heatmap



Conversion for Numeric Values

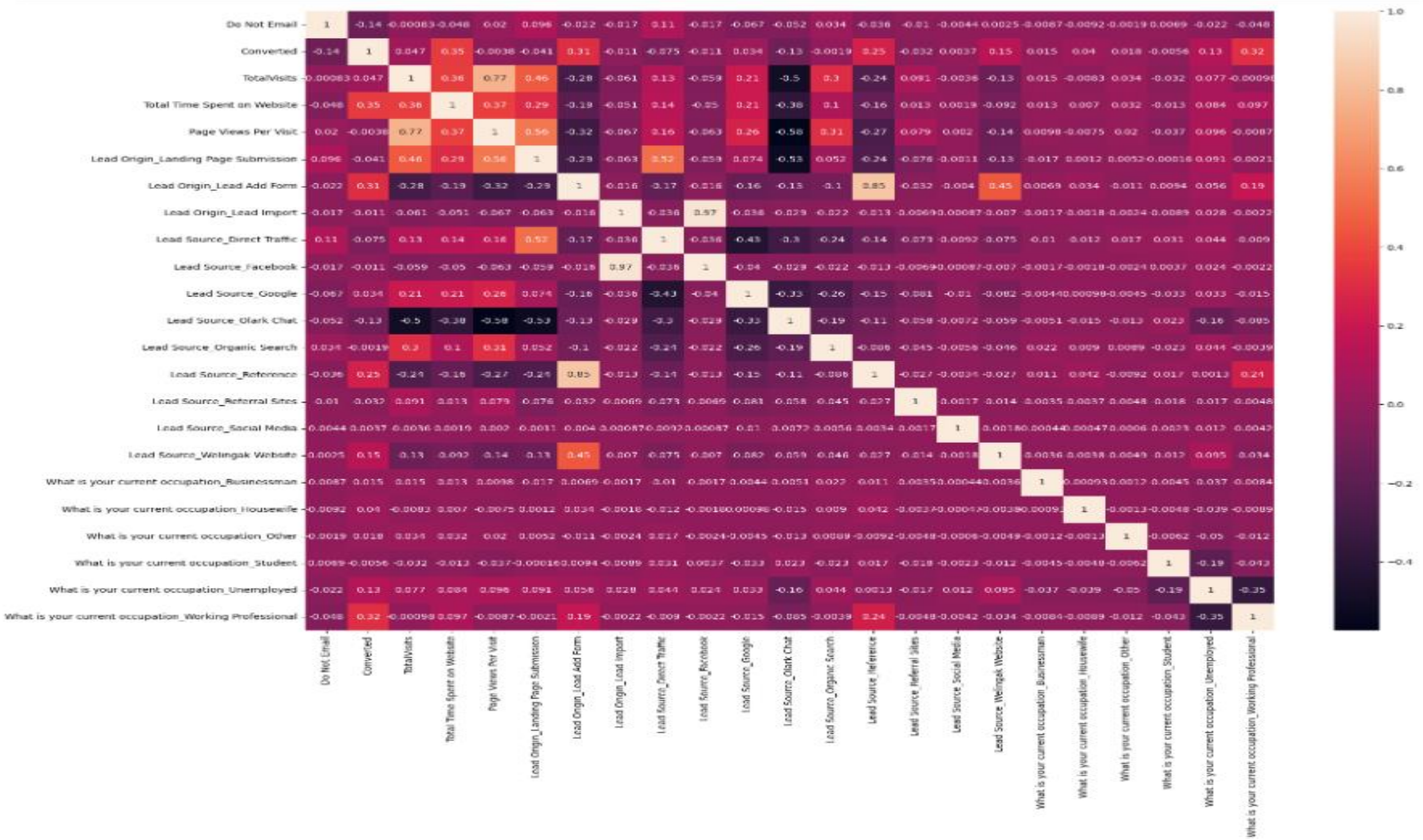


•conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit.

Data Preparation

- Converting binary variables(Yes/No) to (0/1)
- Create Dummy Variable

Correlation Matrix



- As we can see that 'Lead Source_Olark Chat' and 'Lead Origin_Landing Page Submission' are highly correlated dummy variables.

Model Building-1

MODEL 1

```
# MODEL 1

X_train_sm = sm.add_constant(X_train[cols])
logm1 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm1.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6356
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2862.8
Date:	Sun, 15 Oct 2023	**Deviance:**	5725.6
Time:	22:50:11	**Pearson chi2:**	6.38e+03
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.3490
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2420	0.096	-12.979	0.000	-1.430	-1.054
Do Not Email	-0.3583	0.043	-8.295	0.000	-0.443	-0.274
Total Time Spent on Website	1.0998	0.038	28.576	0.000	1.024	1.175
Lead Origin_Lead Add Form	4.1642	0.774	5.379	0.000	2.647	5.682
Lead Source_Direct Traffic	-1.0592	0.108	-9.834	0.000	-1.270	-0.848
Lead Source_Google	-0.7850	0.103	-7.616	0.000	-0.987	-0.583
Lead Source_Organic Search	-0.8803	0.124	-7.094	0.000	-1.123	-0.637
Lead Source_Reference	-1.3303	0.806	-1.650	0.099	-2.911	0.250
Lead Source_Referral Sites	-1.3703	0.336	-4.075	0.000	-2.029	-0.711
Lead Source_Welingak Website	0.7219	1.055	0.684	0.494	-1.347	2.790
What is your current occupation_Businessman	1.5018	0.999	1.503	0.133	-0.456	3.460
What is your current occupation_Housewife	23.8830	1.6e+04	0.001	0.999	-3.14e+04	3.14e+04
What is your current occupation_Other	1.3577	0.641	2.118	0.034	0.101	2.614
What is your current occupation_Student	1.1827	0.225	5.268	0.000	0.743	1.623
What is your current occupation_Unemployed	1.3095	0.083	15.683	0.000	1.146	1.473
What is your current occupation_Working Professional	3.8054	0.189	20.105	0.000	3.434	4.176

- We can observe here that p-value of column 'What is your current occupation_Housewife' is high so we have to drop it.

Model Building-2

MODEL 2

```
# MODEL 2

X_train_sm = sm.add_constant(X_train[cols])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6357
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2872.3
Date:	Sun, 15 Oct 2023	Deviance:	5744.6
Time:	22:50:14	Pearson chi2:	6.40e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3470
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2247	0.095	-12.862	0.000	-1.411	-1.038
Do Not Email	-0.3597	0.043	-8.331	0.000	-0.444	-0.275
Total Time Spent on Website	1.0996	0.038	28.619	0.000	1.024	1.175
Lead Origin_Lead Add Form	4.1662	0.774	5.381	0.000	2.649	5.684
Lead Source_Direct Traffic	-1.0517	0.108	-9.778	0.000	-1.262	-0.841
Lead Source_Google	-0.7756	0.103	-7.540	0.000	-0.977	-0.574
Lead Source_Organic Search	-0.8645	0.124	-6.984	0.000	-1.107	-0.622
Lead Source_Reference	-1.3089	0.806	-1.623	0.105	-2.889	0.272
Lead Source_Referral Sites	-1.3681	0.336	-4.072	0.000	-2.027	-0.710
Lead Source_Welingak Website	0.7294	1.055	0.691	0.490	-1.339	2.798
What is your current occupation_Businessman	1.4744	1.000	1.475	0.140	-0.485	3.434
What is your current occupation_Other	1.3321	0.641	2.079	0.038	0.076	2.588
What is your current occupation_Student	1.1579	0.224	5.160	0.000	0.718	1.598
What is your current occupation_Unemployed	1.2836	0.083	15.498	0.000	1.121	1.446
What is your current occupation_Working Professional	3.7795	0.189	19.999	0.000	3.409	4.150

- We can observe here that p-value of column 'Lead Source_Welingak Website' is high so we have to drop. it.

Model Building-3

MODEL 3

```
#MODEL 3
X_train_sm = sm.add_constant(X_train[cols])
logm3 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm3.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6358
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2872.5
Date:	Sun, 15 Oct 2023	**Deviance:**	5745.1
Time:	22:50:16	**Pearson chi2:**	6.42e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3470
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2215	0.095	-12.847	0.000	-1.408	-1.035
Do Not Email	-0.3606	0.043	-8.350	0.000	-0.445	-0.276
Total Time Spent on Website	1.1006	0.038	28.654	0.000	1.025	1.176
Lead Origin_Lead Add Form	4.6079	0.523	8.807	0.000	3.582	5.633
Lead Source_Direct Traffic	-1.0559	0.107	-9.832	0.000	-1.266	-0.845
Lead Source_Google	-0.7818	0.103	-7.623	0.000	-0.983	-0.581
Lead Source_Organic Search	-0.8687	0.124	-7.026	0.000	-1.111	-0.626
Lead Source_Reference	-1.7536	0.564	-3.109	0.002	-2.859	-0.648
Lead Source_Referral Sites	-1.3724	0.336	-4.085	0.000	-2.031	-0.714
What is your current occupation_Businessman	1.4745	1.000	1.475	0.140	-0.485	3.434
What is your current occupation_Other	1.3324	0.641	2.080	0.038	0.077	2.588
What is your current occupation_Student	1.1571	0.225	5.154	0.000	0.717	1.597
What is your current occupation_Unemployed	1.2843	0.083	15.505	0.000	1.122	1.447
What is your current occupation_Working Professional	3.7806	0.189	20.002	0.000	3.410	4.151

- We can observe here that p-value of column 'What is your current occupation_Businessman' is high so we have to drop it.

Model Building-4

MODEL 4

```
# MODEL 4

X_train_sm = sm.add_constant(X_train[cols])
logm4 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm4.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6359
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2873.5
Date:	Sun, 15 Oct 2023	Deviance:	5747.1
Time:	22:50:19	Pearson chi2:	6.42e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3468
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2155	0.095	-12.809	0.000	-1.401	-1.030
Do Not Email	-0.3610	0.043	-8.360	0.000	-0.446	-0.276
Total Time Spent on Website	1.1004	0.038	28.661	0.000	1.025	1.176
Lead Origin_Lead Add Form	4.6094	0.523	8.810	0.000	3.584	5.635
Lead Source_Direct Traffic	-1.0547	0.107	-9.823	0.000	-1.265	-0.844
Lead Source_Google	-0.7815	0.103	-7.622	0.000	-0.983	-0.581
Lead Source_Organic Search	-0.8655	0.124	-7.003	0.000	-1.108	-0.623
Lead Source_Reference	-1.7436	0.564	-3.091	0.002	-2.849	-0.638
Lead Source_Referral Sites	-1.3729	0.336	-4.087	0.000	-2.031	-0.715
What is your current occupation_Other	1.3254	0.641	2.069	0.039	0.070	2.581
What is your current occupation_Student	1.1497	0.224	5.122	0.000	0.710	1.590
What is your current occupation_Unemployed	1.2770	0.083	15.469	0.000	1.115	1.439
What is your current occupation_Working Professional	3.7733	0.189	19.975	0.000	3.403	4.143

- We can observe here that p-value of column 'What is your current occupation_Other' is high so we have to drop it.

Model Building-5

MODEL 5

```
# MODEL 5

X_train_sm = sm.add_constant(X_train[cols])
logm5 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm5.fit()
res.summary()
```

Generalized Linear Model Regression Results

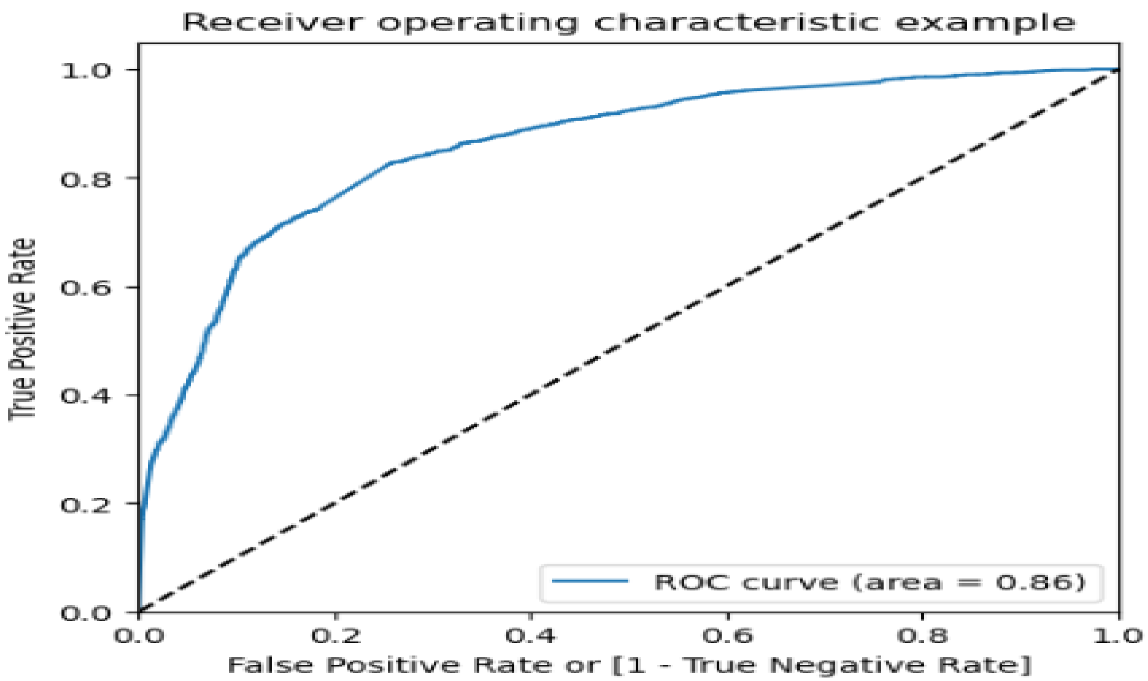
Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6360
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2875.6
Date:	Sun, 15 Oct 2023	Deviance:	5751.2
Time:	22:50:21	Pearson chi2:	6.43e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3464
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2020	0.094	-12.723	0.000	-1.387	-1.017
Do Not Email	-0.3600	0.043	-8.348	0.000	-0.445	-0.276
Total Time Spent on Website	1.1023	0.038	28.710	0.000	1.027	1.178
Lead Origin_Lead Add Form	4.6119	0.523	8.816	0.000	3.587	5.637
Lead Source_Direct Traffic	-1.0496	0.107	-9.783	0.000	-1.260	-0.839
Lead Source_Google	-0.7804	0.102	-7.615	0.000	-0.981	-0.580
Lead Source_Organic Search	-0.8639	0.124	-6.987	0.000	-1.106	-0.622
Lead Source_Reference	-1.7425	0.564	-3.089	0.002	-2.848	-0.637
Lead Source_Referral Sites	-1.3749	0.336	-4.094	0.000	-2.033	-0.717
What is your current occupation_Student	1.1342	0.224	5.057	0.000	0.695	1.574
What is your current occupation_Unemployed	1.2613	0.082	15.384	0.000	1.101	1.422
What is your current occupation_Working Professional	3.7575	0.189	19.919	0.000	3.388	4.127

- As model 5 seems to be stable enough with significant p-value

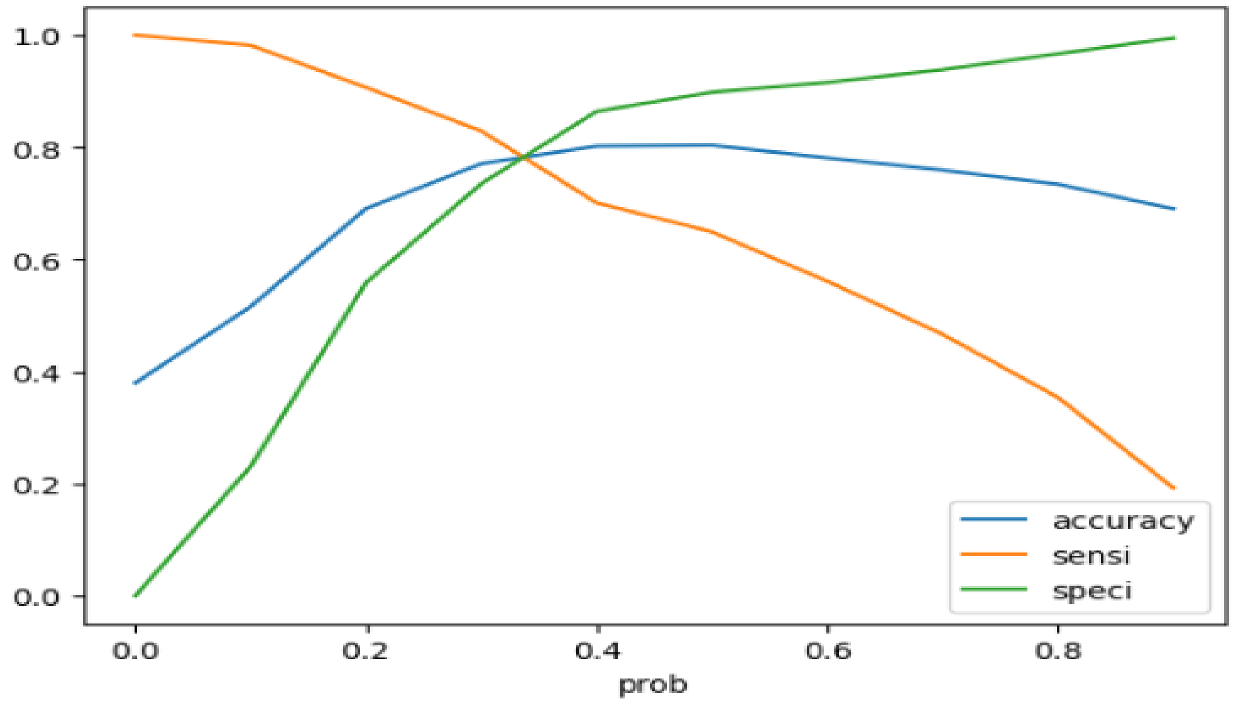
Prediction on train model

(Plotting ROC Curve)



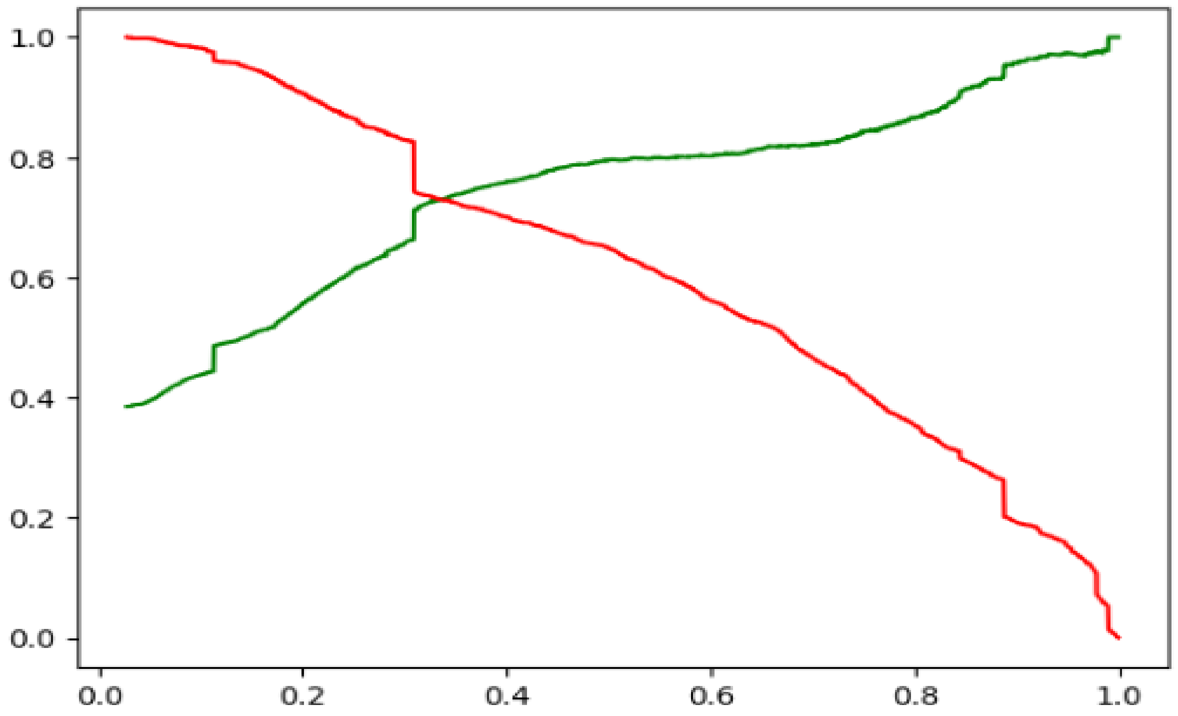
We are getting 0.86 which indicating a good predictive model as ROC should be close to 1

Plotting accuracy sensitivity and specificity



- From the above curve, we can see that the optimum point to take as cut off probability is 0.3

Precision-Recall Curve



we got 0.34 as the Cut-off as Precesion-Recall Thresholds

Overall Metrics-I

Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity, False Postive Rate, Positive Predictive Value, Negative Predicitive Value on final prediction on test set

```
# checking overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_Predicted)
```

```
0.7751739289637496
```

```
confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_Predicted )
confusion2
```

```
array([[1252,  437],
       [ 177,  865]], dtype=int64)
```

```
TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

```
#Checking sensitivity of our  model
TP / float(TP+FN)
```

```
0.8301343570057581
```

```
# Calculating specificity
TN / float(TN+FP)
```

```
0.7412670219064535
```

Overall Metrics-II

Precision and Recall matrics on test set

```
#Importing precision_score
from sklearn.metrics import precision_score
precision_score(y_pred_final.Converted , y_pred_final.final_Predicted)
```

```
0.6643625192012289
```

```
#Importing recall_score
from sklearn.metrics import recall_score
recall_score(y_pred_final.Converted, y_pred_final.final_Predicted)
```

```
0.8301343570057581
```

Inference

After running the model on the Test Data these are the figures we obtain:

Accuracy : 77.52% Sensitivity :83.01% Specificity : 74.13%

Conclusion:

- As we have checked Sensitivity-Specificity and Precision-Recall , we considered optimal cut off based on sensitivity and specificity to calculate final prediction.
- Accuracy, sensitivity and specificity values of test data set are 77.54%, 83.01% and 74.13% which are quite closer to the values we get on train data set.
- Lead score calculated on train data set showing conversion rate on final prediction model is around 80% which means our model is good to go.