upGrad | iiit-b International Institute of Information Technology Bangalore

upGrad

Post Graduate Diploma – Data Science

Machine Learning

Lead Score Case Study

Shobhit Sinha

Goutham Dasari

Problem Statement

An education company named 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, although X Education gets a lot of leads, its lead conversion rate is very poor. The task is to build a model wherein you need to 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 chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

1. Data Cleaning : This includes handling Nan values , other values like ‘Select’ , also Exploratory data analysis , to identify variables which are relevant for modelling and removing outliers and dropping non relevant fields
2. Data preparation- This involves creating dummy variables, splitting the data into train and test datasets and scaling of numerical Variable with standardardized scaling technique.
3. Data Modeling- This includes Feature Selection using RFE, Building model on train set ,checking VIF , measuring optimum probability , model accuracy, and other metrics
4. Prediction-This involves making predictions on the test set and measuring the accuracy and other metrics

Data Cleaning

We can see the percentage of missing value in the figure .

1. Most of the variables are user entries , seems to be from website so the fields with no entry have "Select" as value. For modelling we have replaced these values with NAN .

2. Handling missing or Nan values with following steps

   a) we will identify the % of missing values for each field
   b) drop those fields with more than 45% missing values
   c) replacing the Nan values with most occurring values
   d) If there is no obvious most occurring value, then replace Nar with "Unknown"
   e) Dropping the records if the missing % value is lesser than 2%

```
In [23]: # Checking the percentage of missing values
         round(100*(lead_data.isnull().sum()/len(lead_data.index)), 2).sort_values()

Out[23]: Prospect ID                                    0.00
         I agree to pay the amount through cheque       0.00
         Get updates on DM Content                      0.00
         Update me on Supply Chain Content              0.00
         Receive More Updates About Our Courses         0.00
         Through Recommendations                        0.00
         Digital Advertisement                          0.00
         Newspaper                                      0.00
         X Education Forums                             0.00
         A free copy of Mastering The Interview         0.00
         Magazine                                       0.00
         Search                                         0.00
         Newspaper Article                              0.00
         Last Notable Activity                          0.00
         Total Time Spent on Website                    0.00
         Converted                                      0.00
         Do Not Call                                    0.00
         Do Not Email                                   0.00
         Lead Number                                    0.00
         Lead Origin                                    0.00
         Lead Source                                    0.39
         Last Activity                                  1.11
         Page Views Per Visit                           1.48
         TotalVisits                                    1.48
         Country                                       26.63
         What is your current occupation               29.11
         What matters most to you in choosing a course 29.32
         Tags                                          36.29
         Specialization                                36.58
         City                                          39.71
         Asymmetrique Activity Index                   45.65
         Asymmetrique Profile Index                    45.65
         Asymmetrique Activity Score                   45.65
         Asymmetrique Profile Score                    45.65
         Lead Quality                                  51.59
         Lead Profile                                  74.19
         How did you hear about X Education            78.46
         dtype: float64
```

Exploratory Data Analysis

we would need to analyze the distribution of various values of all the above considered fields against the "Converted" value .
Intent is to check if the convertion is dependent on the variable or not

Based on the value distribution following Fields are found to non – relevant fields for Conversion.

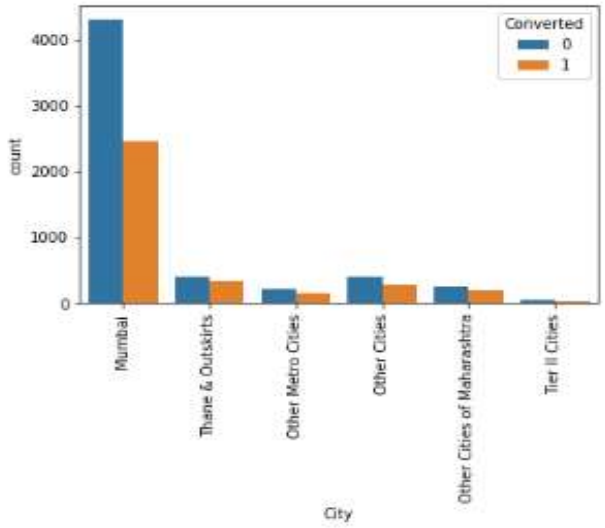
These main observation made which resulted in the decision of dropping these fields :

Most of the values of the field more than 90% of them have same value

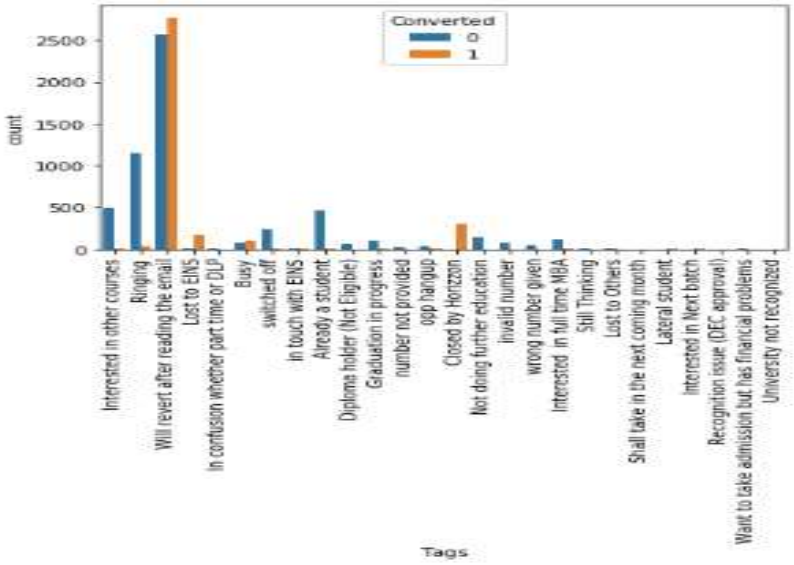
Eg : almost all of the leads are from ‘India’ so the field country doesn’t affect the conversion

'I agree to pay the amount through cheque’,
'Get updates on DM Content’,
'Update me on Supply Chain Content’,
'Receive More Updates About Our Courses'
,'Through Recommendations’,
'Digital Advertisement’
,'Newspaper’,
'X Education Forums’,
'Newspaper Article',
'Magazine’,
'Search’,
'What matters most to you in choosing a course’,
'Country’,
'Do Not Call’,
'Do Not Email',
'A free copy of Mastering The Interview’,

Exploratory Data Analysis

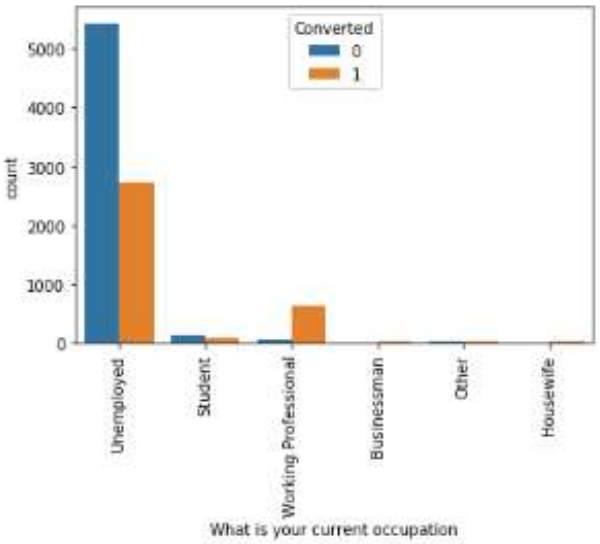


Most of the leads converted from Mumbai. But we cannot infer from this as most of the leads are also from Mumbai

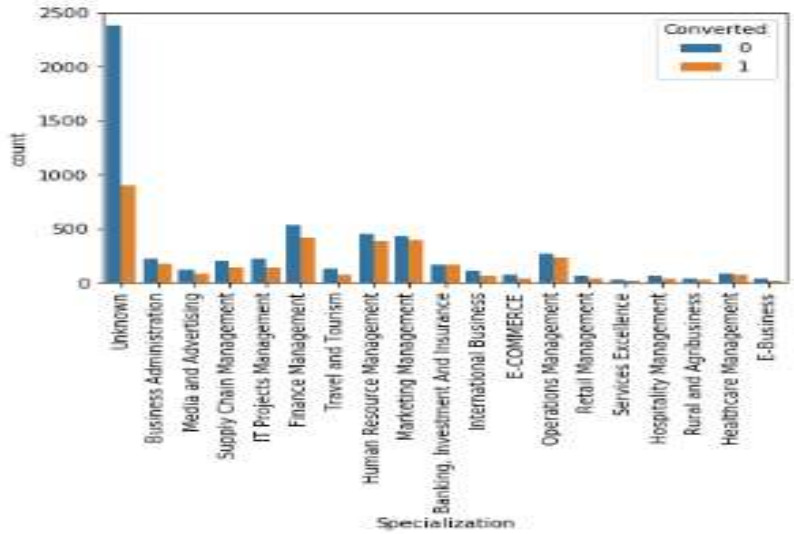


As most of the leads have "revert after reading the email" , more leads are mapped to this. but to be observed that more than 50% of the leads are converted with this status as well.

Exploratory Data Analysis

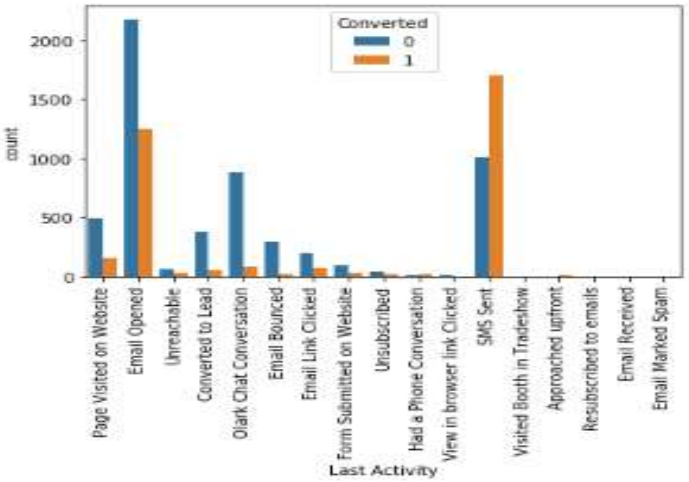


Unemployed leads though are high in total count the conversion rate is low.
Student are very low in count
Working Professionals have high convertion rate



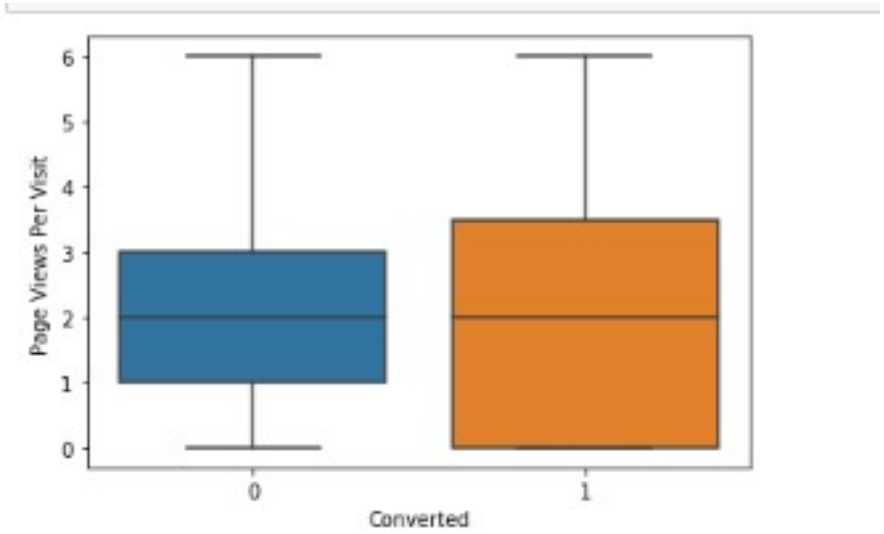
Seems many courses like 'business adminitration','banking investment and insurance','finance management' and few others have high convertion

Exploratory Data Analysis



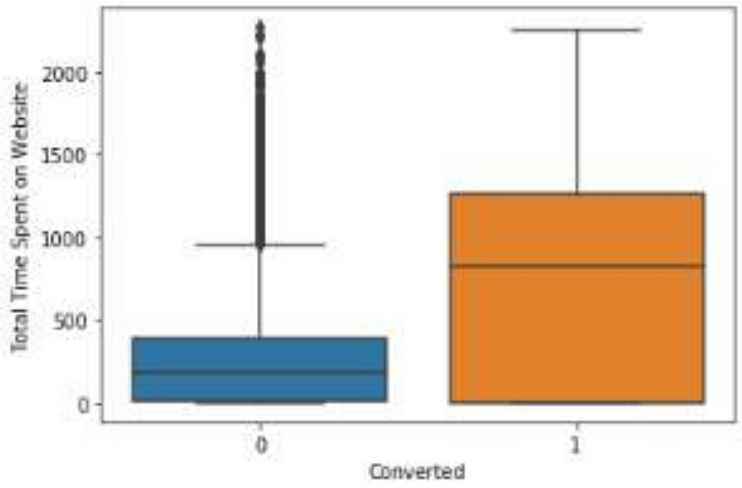
Last Activity SMS sent has very high convertion rate

last Activity Email Opened though has high count but has relatively less convertion

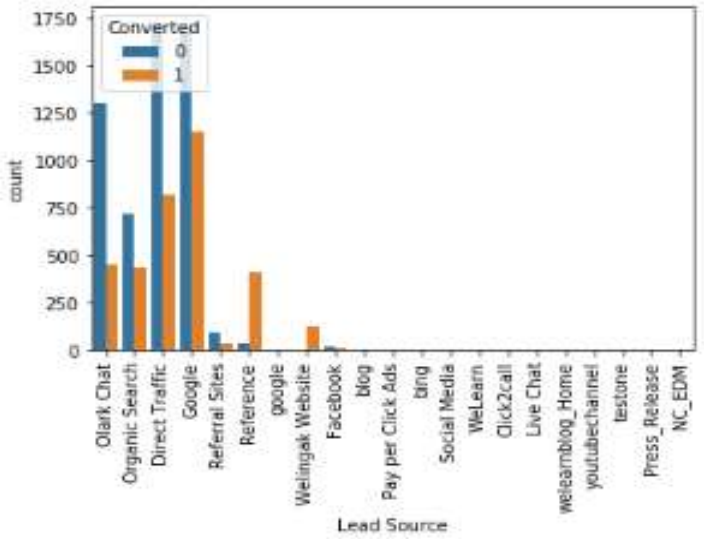


As the median for both Converted and not- converted are almost the same , we cannot colclude if high page views affect the convertion.

Exploratory Data Analysis

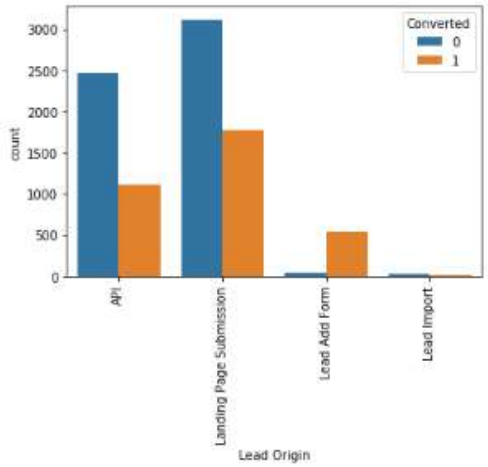


Median for converted is higher , this indicates if leads spend more time on the website there is a higher chance of conversion.

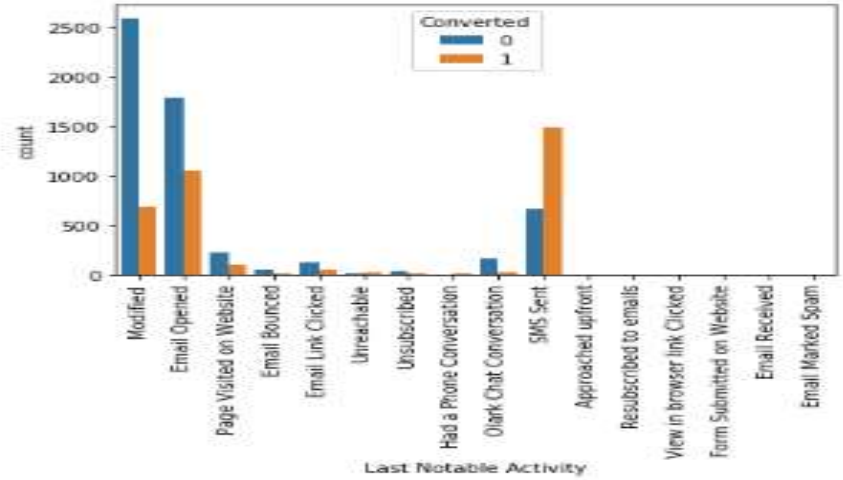


The lead sources like Reference , welingak Website is very high , should implement strategies increasing leads from this Google and , Direct chat,organic search and Olark chat have higher count of leads

Exploratory Data Analysis



The count of leads is higher for Landing page submission and API but the convertion is less the count of leads is low for Lead Add form but the convertion is very high



no particular conlcusion is made out of this

Data Preparation

1. Creating Dummy Variables for all the categorical variables.
2. Scale the necessary variables with standard technique
3. Data split into Train and test set in the ratio 70/30

Model Building

Feature selection using RFE method:

RFE is used to select the attributes automatically. Thus after this process we end up selecting below 15 variables.

1. Lead Origin_Landing Page Submission
2. Lead Origin_Lead Add Form
3. Lead Source_Welingak Website
4. Last Activity_Email Bounced
5. Last Activity_Had a Phone Conversation
6. Specialization_Unknown
7. What is your current occupation_Unemployed
8. Tags_Busy
9. Tags_Closed by Horizzon
10. Tags_Lost to EINS
11. Tags_Ringing
12. Tags_Will revert after reading the email
13. Tags_invalid number
14. Tags_switched off
15. Last Notable Activity_SMS Sent

Model Building

We have run the training model ; from the output of the model we can observe pvalue>0.05

As ***Tags_invalid number*** has high pvalue we drop this variable

Covariance Type: nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5824	0.245	-2.379	0.017	-1.062	-0.103
Lead Origin_Landing Page Submission	-1.4546	0.147	-9.862	0.000	-1.744	-1.166
Lead Origin_Lead Add Form	1.6761	0.317	5.294	0.000	1.056	2.297
Lead Source_Welingak Website	2.5187	0.804	3.133	0.002	0.943	4.094
Last Activity_Email Bounced	-2.0760	0.389	-5.338	0.000	-2.838	-1.314
Last Activity_Had a Phone Conversation	2.7778	0.972	2.857	0.004	0.872	4.683
Specialization_Unknown	-2.1169	0.150	-14.093	0.000	-2.411	-1.822
What is your current occupation_Unemployed	-2.3111	0.187	-12.342	0.000	-2.678	-1.944
Tags_Busy	3.0114	0.307	9.800	0.000	2.409	3.614
Tags_Closed by Horizzon	8.4197	0.743	11.335	0.000	6.964	9.876
Tags_Lost to EINS	8.3874	0.746	11.236	0.000	6.924	9.850
Tags_Ringing	-1.1915	0.317	-3.760	0.000	-1.812	-0.570
Tags_Will revert after reading the email	3.7549	0.205	18.352	0.000	3.354	4.156
Tags_invalid number	-21.6538	1.47e+04	-0.001	0.999	-2.88e+04	2.88e+04
Tags_switched off	-1.2049	0.560	-2.153	0.031	-2.302	-0.108
Last Notable Activity_SMS Sent	2.8209	0.114	24.766	0.000	2.598	3.044

Model Building

Model is built again and from the second model we can observe p value>0.05 for **Tags_switched off** in next Batch.

We can drop this attribute and build model again.

No. Iterations:	23					
Covariance Type:	nonrobust					

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5824	0.245	-2.379	0.017	-1.062	-0.103
Lead Origin_Landing Page Submission	-1.4546	0.147	-9.862	0.000	-1.744	-1.166
Lead Origin_Lead Add Form	1.6761	0.317	5.294	0.000	1.056	2.297
Lead Source_Welingak Website	2.5187	0.804	3.133	0.002	0.943	4.094
Last Activity_Email Bounced	-2.0760	0.389	-5.338	0.000	-2.838	-1.314
Last Activity_Had a Phone Conversation	2.7778	0.972	2.857	0.004	0.872	4.683
Specialization_Unknown	-2.1169	0.150	-14.093	0.000	-2.411	-1.822
What is your current occupation_Unemployed	-2.3111	0.187	-12.342	0.000	-2.678	-1.944
Tags_Busy	3.0114	0.307	9.800	0.000	2.409	3.614
Tags_Closed by Horizzon	8.4197	0.743	11.335	0.000	6.964	9.876
Tags_Lost to EINS	8.3874	0.746	11.236	0.000	6.924	9.850
Tags_Ringing	-1.1915	0.317	-3.760	0.000	-1.812	-0.570
Tags_Will revert after reading the email	3.7549	0.205	18.352	0.000	3.354	4.156
Tags_invalid number	-21.6538	1.47e+04	-0.001	0.999	-2.88e+04	2.88e+04
Tags_switched off	-1.2049	0.560	-2.153	0.031	-2.302	-0.108
Last Notable Activity_SMS Sent	2.8209	0.114	24.766	0.000	2.598	3.044

```
In [100]: col2 = col1.drop('Tags_switched off',1)
```

Model Building

Model is built again and from the second model we can observe all the pvalues are less the 0.05

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1990.5
Date:	Mon, 20 Apr 2020	**Deviance:**	3981.1
Time:	16:24:15	**Pearson chi2:**	1.09e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7700	0.249	-3.089	0.002	-1.259	-0.282
Lead Origin_Landing Page Submission	-1.4894	0.149	-10.012	0.000	-1.781	-1.198
Lead Origin_Lead Add Form	1.7037	0.319	5.334	0.000	1.078	2.330
Lead Source_Welingak Website	2.4855	0.805	3.088	0.002	0.908	4.063
Last Activity_Email Bounced	-2.0792	0.390	-5.331	0.000	-2.844	-1.315
Last Activity_Had a Phone Conversation	2.7710	0.970	2.857	0.004	0.870	4.672
Specialization_Unknown	-2.1375	0.152	-14.100	0.000	-2.435	-1.840
What is your current occupation_Unemployed	-2.4096	0.192	-12.574	0.000	-2.785	-2.034
Tags_Busy	3.3570	0.293	11.471	0.000	2.783	3.931
Tags_Closed by Horizzon	8.7308	0.739	11.807	0.000	7.282	10.180
Tags_Lost to EINS	8.7022	0.743	11.712	0.000	7.246	10.158
Tags_Ringing	-0.8313	0.301	-2.761	0.006	-1.421	-0.241
Tags_Will revert after reading the email	4.0717	0.191	21.358	0.000	3.698	4.445
Last Notable Activity_SMS Sent	2.7642	0.111	24.968	0.000	2.547	2.981

Prediction

1. Post model run , we are trying to calculate the probability of customer being converted
2. Assuming customer with probability greater than 0.5 gets converted we create a new variable if customer is converted or not

Thus the data looks something like this.

	Converted	Converted_prob	Prospect ID	predicted
0	0	0.354966	3009	0
1	0	0.082831	1012	0
2	0	0.002132	9226	0
3	1	0.897234	4750	1
4	1	0.982591	7987	1

Model Evaluation

Confusion matrix is used

```
from sklearn import metrics

# Confusion matrix
confusion_mat = metrics.confusion_matrix(pred_final.Converted, pred_final.predicted )
print(confusion_mat)
```

```
[[3739  166]
 [ 640 1806]]
```

3739 customer were not converted, and model also predicted them as non potential leads, these are called True Negatives(TN)

166 customers were wrongly predicted as potential customers while they were not actually these are called False Positives(FP)

640 customers were converted but the model predicted them as non potential leads these are called False Negative(FN).

1806 customers who were converted was correctly predicted as potential leads these are called True Positives(TP)

Model Evaluation

Further checking various score metrics.

Accuracy : 87.3%

Sensitivity : 73.8%

Specificity : 95.7%

False positive rate : 4.2%

Positive Predictive value : 91.6%

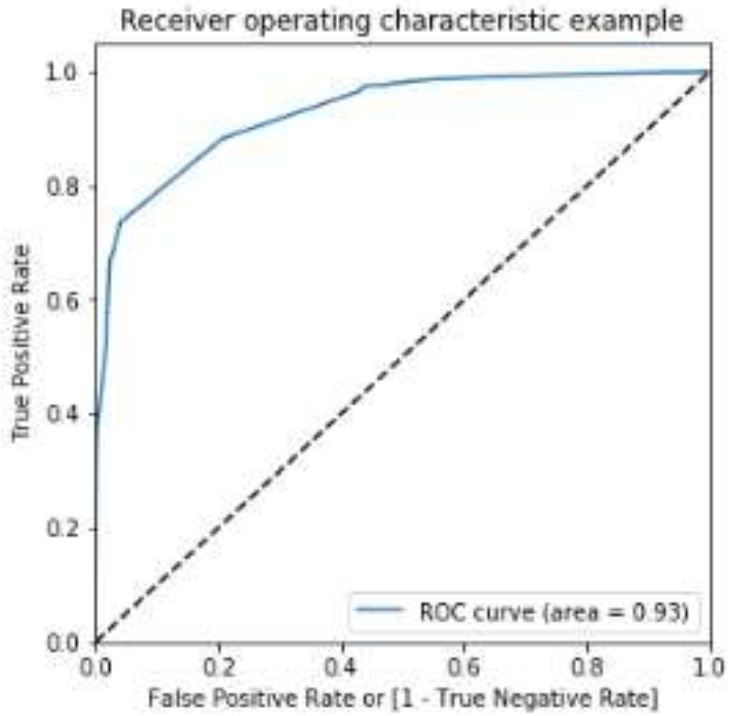
Negative Predictive Value : 85.4%

Model Evaluation

Plotting Roc Curve

The ROC curve between sensitivity and specificity is closed to the left-hand border and then the top border of the space,

We can conclude from the curve that the model accuracy is better



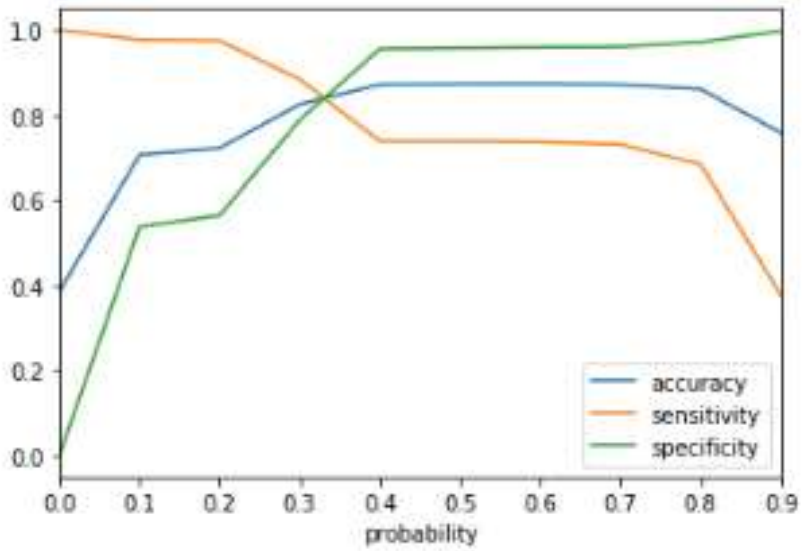
Model Evaluation

Finding the optimal cutoff point for probability

So far, we evaluated the model assuming the customer can be a potential customer if probability is greater than 0.5.

We optimal cut off point is the point where accuracy, sensitivity and specificity curve cross each other.

From the figure we can notice the cutoff is little less than 0.4 and we can take it to be 0.4.



Model Evaluation

Using the optimal cut off point if the prediction is done on the model,

We get the confusion Matrix like this

```
confusion_mat1 = metrics.confusion_matrix(pred_final.Converted, pred_final.final_predicted )
confusion_mat1
```

```
array([[3732,  173],
       [ 640, 1806]], dtype=int64)
```

3732 customer were not converted, and model also predicted them as non potential leads

173 customers were wrongly predicted as potential customers while they were not actually

640 customers were converted but the model predicted them as non potential leads

1806 customers who were converted was correctly predicted as potential

Model Evaluation

Accuracy : 87.2%

Sensitivity : 73.8%

Specificity : 95.6%

False positive rate : 4.2%

Positive Predictive value : 91.2%

Negative Predictive Value : 85.4%

Since the Score is good, we can consider model to be a good performing model and can be used to make predictions on the data.

Prediction on Test data

1. predicting the probability for the customers getting converted
2. Since we know optimal cut off is 0.4, we can use this to introduce predict variable to predict the potential lead.

Confusion matrix :

```
: confusion_mat2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.Final_Predicted )
  confusion_mat2
```

```
: array([[1645,   89],
         [ 291,  698]], dtype=int64)
```

1645 customer were not converted, and model also predicted them as non potential leads

89 customers were wrongly predicted as potential customers while they were not actually

291 customers were converted but the model predicted them as non potential leads

698 customers who were converted was correctly predicted as potential

Prediction on Test data

Accuracy : 86%

Sensitivity : 70.1%

Specificity : 94.8%

Since this score is almost like the sensitivity and specificity score got for trained model the prediction looks correct.

Conclusion

The model has resulted high accuracy results in predicting the leads who can be converted.

So the marketing team can leverage this to make their operations more efficient by reducing the number customer interactions there by improving the conversions as well.