Lead Scoring Analysis

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. 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%. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Analysis Approach:

- Analyse the dataset with the given input variables data quality
- Identifying the missing values, Outliers and treat them respectively
- Standardising the variables to keep all variables on the same scale
- Creating Dummy variables and Label Encoding respectively
- Implementing Logistic Regression from Stats Model
- Implement Logistic Regression from sklearn to fit using RFE to select most 20 important features
- Validating VIF and removing if necessary and re-iterate the entire process.
- Evaluate the results using Accuracy, AUC and Roc, Precision and Recall.
- Interpret the coefficients to identify the most potential lead and factors contributing towards it.

Lead Scoring Analysis

Model Architecture:

Evaluate model with accuracy metrics (ROC, Accuracy, Precision and Recall)

Finding Probability and set threshold based on the cutoff point from metrics.

Identify the potential leads from the probability score from 0-100 and categorise them

Lead Score EDA Analysis:

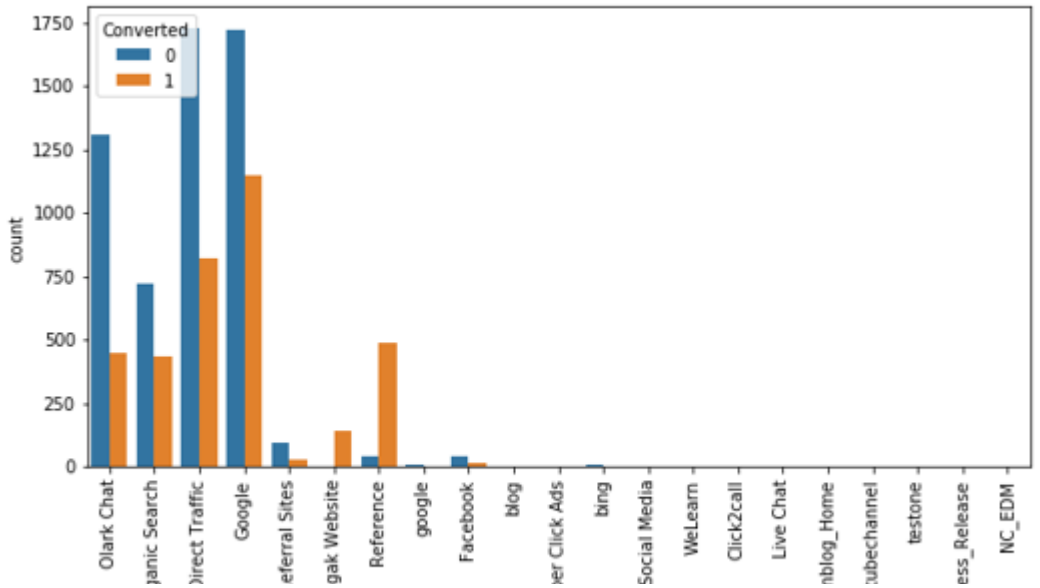
From the given data we have columns like below:

Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	What matters most to you in choosing a course	Search	Maga
API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemployed	Better Career Prospects	No	
API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	Better Career Prospects	No	
Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Student	Better Career Prospects	No	
Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemployed	Better Career Prospects	No	
Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other	Unemployed	Better Career Prospects	No	

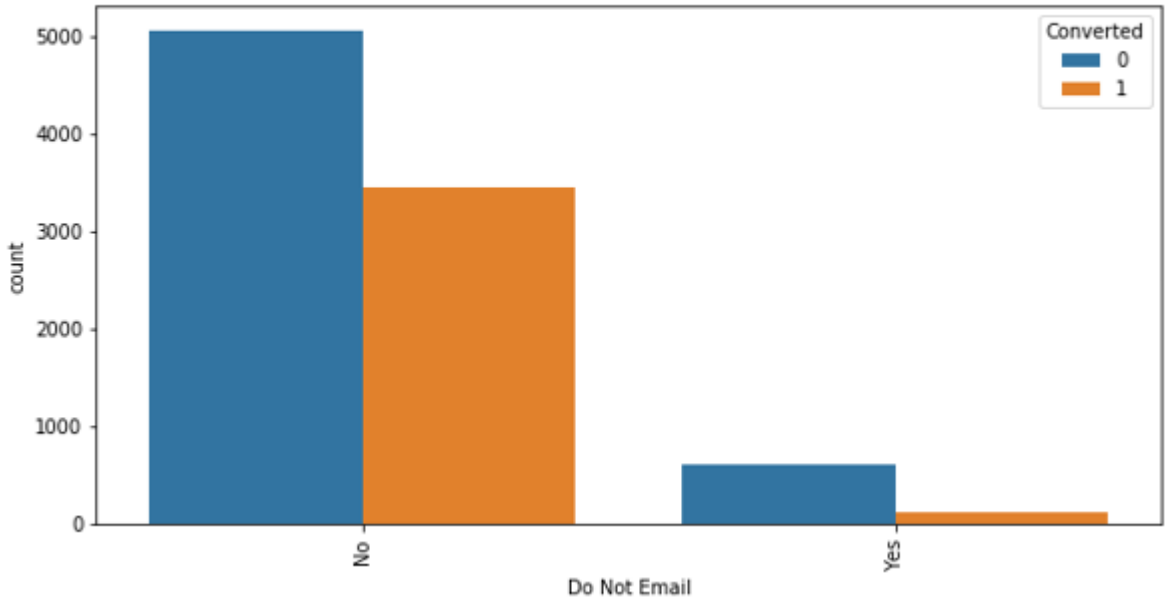
The dataset consists of 9240 rows and 37 columns

Insights From the EDA

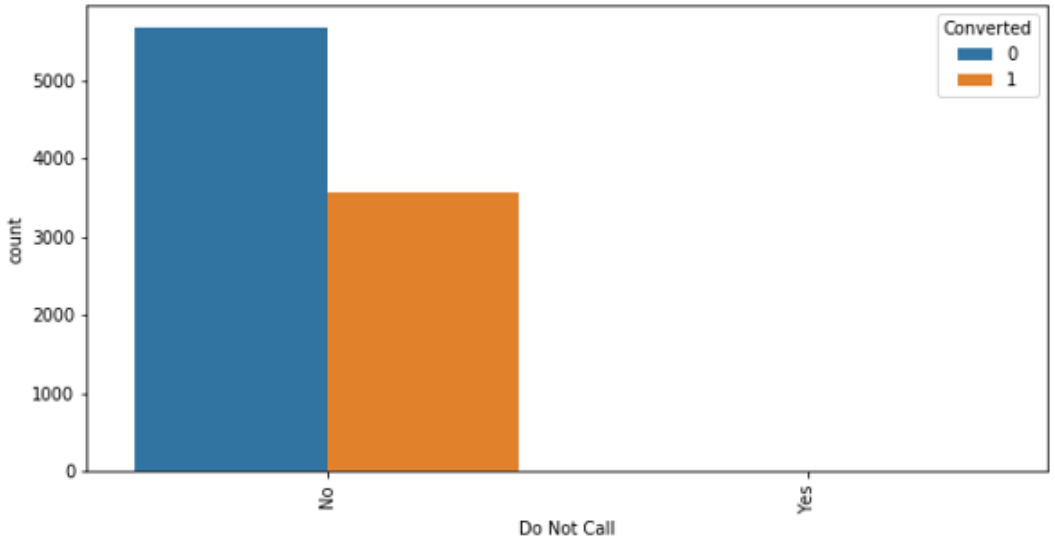
- Most of the leads are came through Google Search and Direct Traffic only



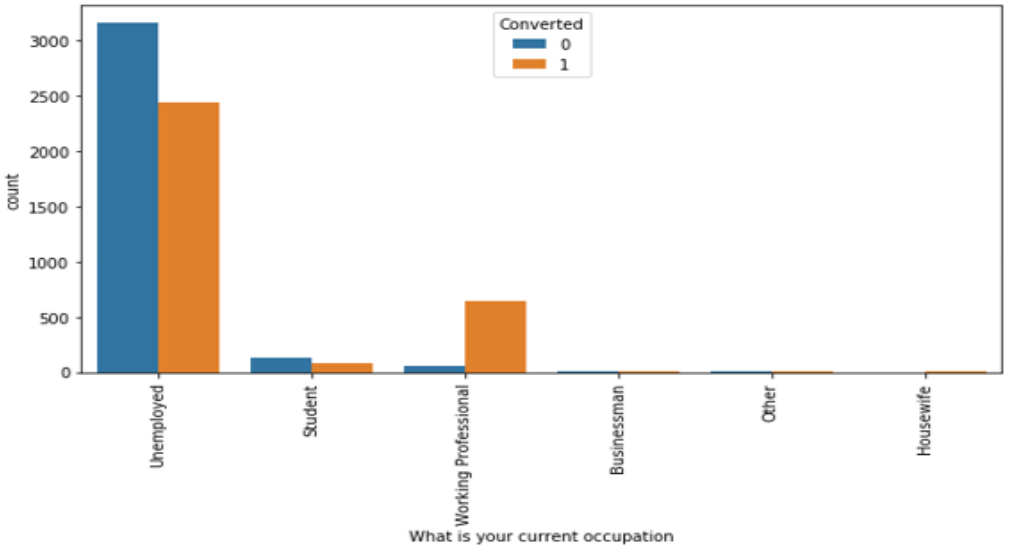
- People who opted "Do not Email" to No, they are mostly converted leads.

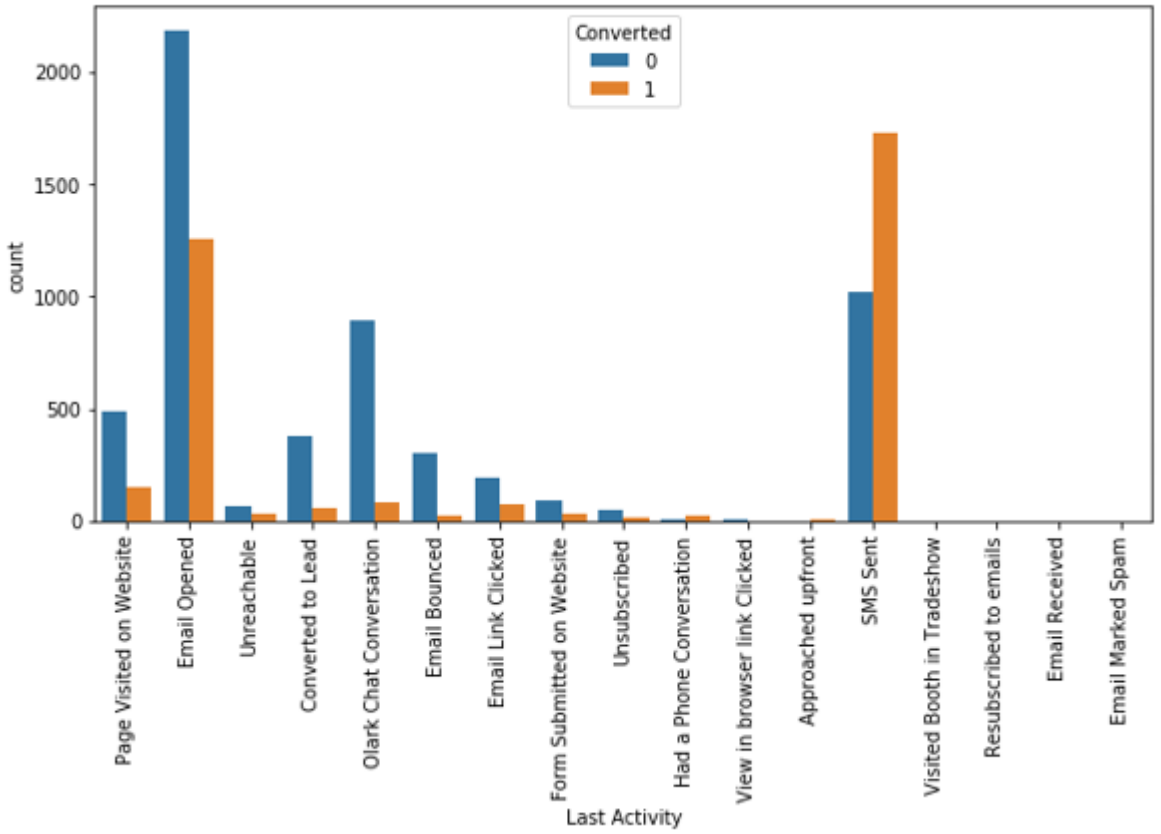


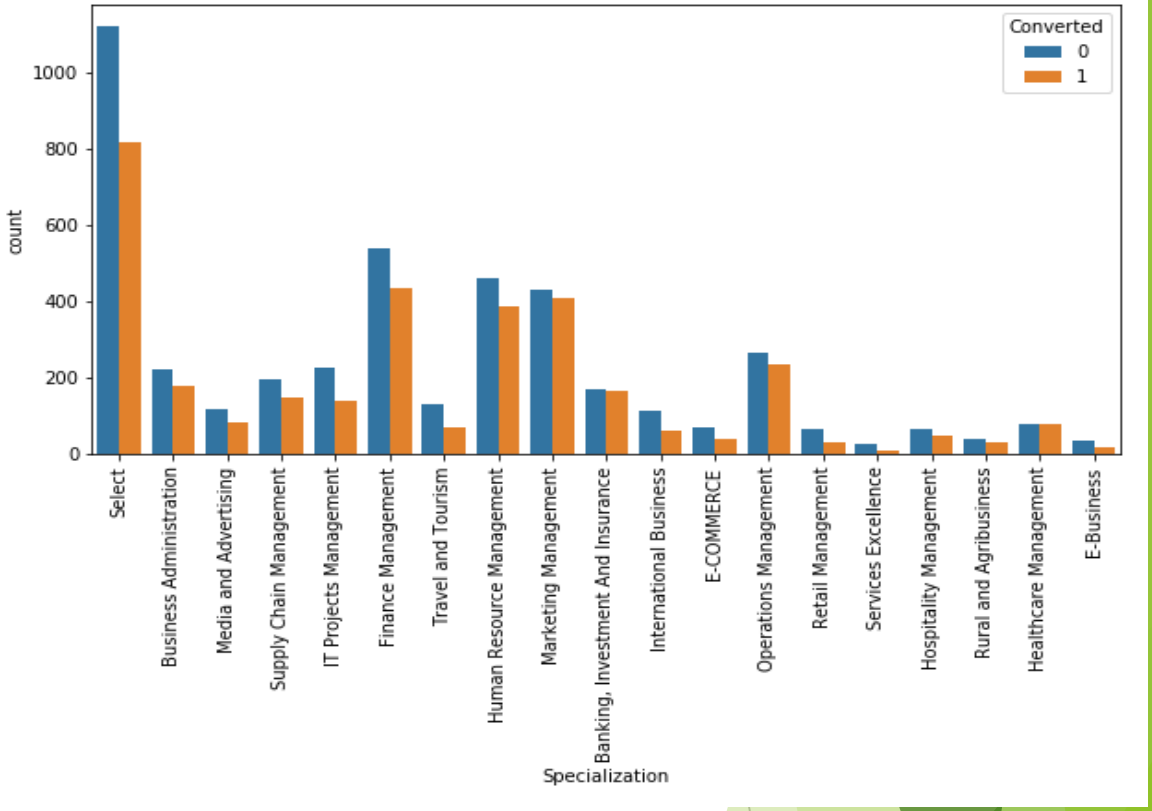
- Leads who don't opted for Do not call option has high conversion rate.



- Compare to all, users who are "Un employed" are most likely to be potential leads.

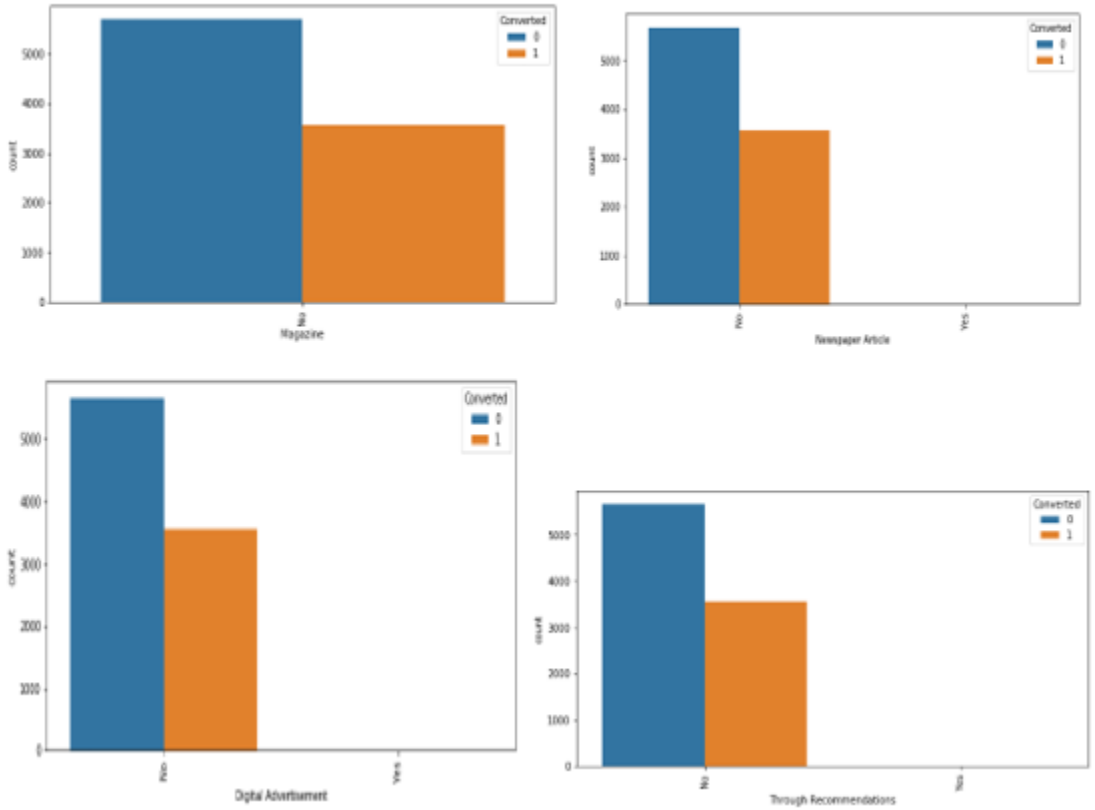




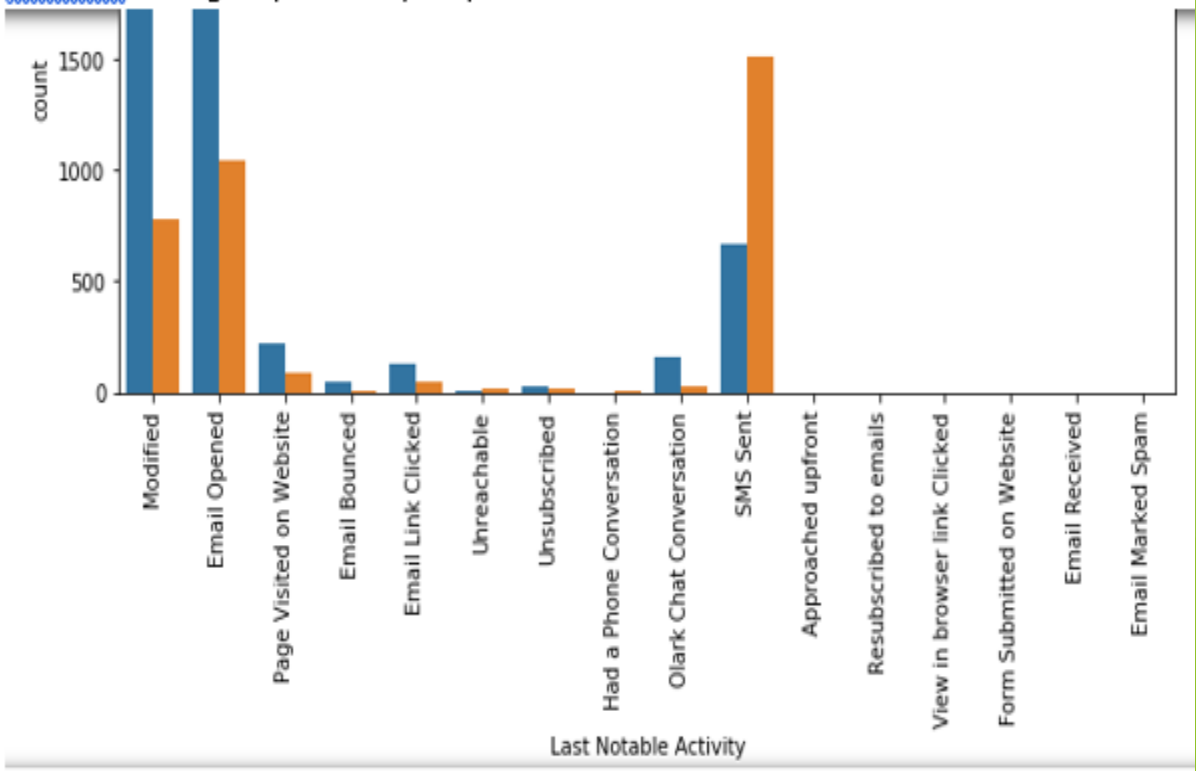


- Users who opened the email and who sent SMS has the high probability of the leads conversion rate.
- There are multiple Specializations, in which many users are not selected the specialization option and left that select column. So many entries are coming under select category. Which needs to be treated as null value only.

- Coming to the Advertisements from Magazine, Newspaper, Digital Advertisement, Through Recommendation has no impact on potential leads.



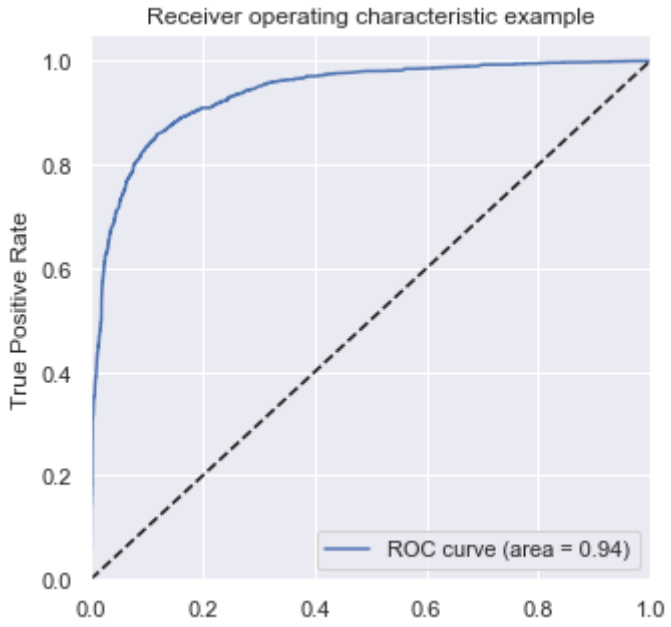
- From the Last Notable Activity Column, the users who SMS sent and Email Opened, Modified has higher probability for potential leads.

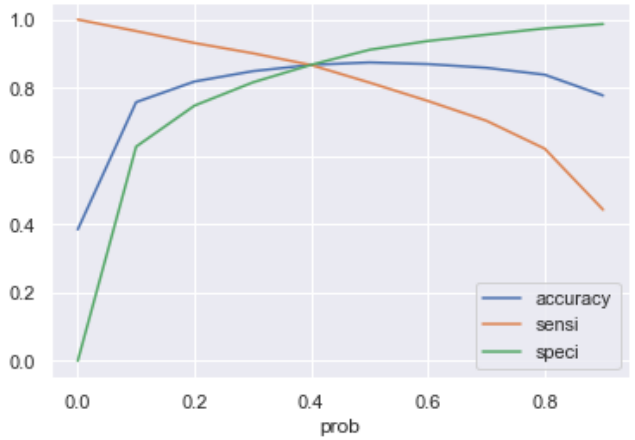


So advertisements are not creating useful impact on the users to increase the potential leads. So company needs to be focus on that.

Important Results and Visualizations from the Analysis

Obtained accuracy of 87.9 %






```
print(metrics.accuracy_score(y_test_pred_final.Converted, y_test_pred_final.Precison_Recall_Cutoff_predicted))
```

```
0.8791486291486291
```

```
confusion4 = metrics.confusion_matrix(y_test_pred_final.Converted, y_test_pred_final.Precison_Recall_Cutoff_pre
confusion4
```

```
array([[1538,  163],
       [ 172,  899]], dtype=int64)
```

```
# Classification report to evaluate the results
from sklearn.metrics import classification_report
print(classification_report(y_test_pred_final.Converted, y_test_pred_final.Precison_Recall_Cutoff_predicted))
```

```
             precision    recall  f1-score   support

          0       0.90      0.90      0.90      1701
          1       0.85      0.84      0.84      1071

avg / total       0.88      0.88      0.88      2772
```

- So for both cutoff points using Sensitivity & specificity and Precision & Recall the overall accuracy will be in the same level

From the above plot, 0.4 is the opimum point to take as cutoff for the probability

Logg Odds and its Intrepretation

- **Factors that are mainly contributing to form Potential Leads:**
  - other_Lead_Origin means 'Lead Add form', 'Lead Import' and 'Quick Add Form' together from 'Lead Origin' is likely to contribute
  - 'Lost to EINS' from Tags is 24.35 times more likely to contribute
  - 'Closed by Horizzon' from Tags is 15.67 times more likely to contribute
  - 'SMS Sent' from last_notable_activity is 14.04 times more likely to contribute
  - 'Will revert after reading the email' from Tags is 11.7 times more likely to contribute
  - Busy from Tags is 3.12 time more like to contribute
  - Total Time Spent on Website is 2.8 times more like to contribute
- **Factors not creating any positive affect on Leads:**
  - 'Olark Chat Conversation' from Last notable activity is 5.97 times not likely creating any impact
  - Do not Email column is 3.82 times not likely
  - 'switched off' from Tags is 2.72 times not likely
  - Ringing from Tags is 2.60 times not likely
  - 'Interested in other courses' from Tags is 1.74 times not likely
  - 'Already a student' from Tags is 1.46 times not likely to create any impact to Potential Leads

Log Odds(coefficients)	Odds Ratio	Column
-1.7873	5.973302757	main_Last Activity_Olark Chat Conversation
-1.3427	3.829368856	Do Not Email
-1.0025	2.725086035	main_Tags_switched off
-0.9571	2.604133525	main_Tags_Ringing
-0.5587	1.748398105	main_Tags_Interested in other courses
-0.3851	1.46976129	main_Tags_Already a student
0.5085	1.662795131	main_Lead_Origin_API
0.5505	1.734119861	main_Last Activity_Email Opened
0.5841	1.793376221	main_Last Activity_other_Last Activity
0.588	1.800384044	Asymmetrique Activity Score
1.0416	2.833747385	Total Time Spent on Website
1.1386	3.122393952	main_Tags_Busy
2.4679	11.79764576	main_Tags_Will revert after reading the email
2.6426	14.04968534	main_last_notable_activity_SMS Sent
2.7519	15.67238114	main_Tags_Closed by Horizzon
3.1929	24.35896611	main_Tags_Lost to EINS
3.215	24.90329191	main_Lead_Origin_other_Lead_Origin

Hot Leads(Who has high probability to become potential lead):

```
# Finding the Hot Leads
output.sort_values(by='Converted_Prob',ascending=False).head()
```

gs	Lead Quality	Update me on Supply Chain Content	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetrique Profile Index	Asymmetrique Activity Score	Asymmetrique Profile Score	I agree to pay the amount through cheque	A free copy of Mastering The Interview	Last Notable Activity	Converted_Prob
ill ert er ng ne ail	Might be	No	No	Select	Mumbai	01.High	01.High	16.0	19.0	No	Yes	SMS Sent	99.98
ill ert er ng ne ail	High in Relevance	No	No	Potential Lead	Mumbai	NaN	NaN	NaN	NaN	No	No	SMS Sent	99.92
ill ert er ng ne ail	High in Relevance	No	No	Potential Lead	Thane & Outskirts	02.Medium	01.High	14.0	19.0	No	No	SMS Sent	99.88
ed by	Might be	No	No	Select	Select	01.High	01.High	17.0	17.0	No	No	Page Visited on	99.88

Cold Leads(Who has low probability to not become potential lead):

```
# Finding the Cold Leads
output.sort_values(by='Converted_Prob',ascending=True).head()
```

s	Lead Quality	Update me on Supply Chain Content	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetrique Profile Index	Asymmetrique Activity Score	Asymmetrique Profile Score	I agree to pay the amount through cheque	A free copy of Mastering The Interview	Last Notable Activity	Converted_Prob
g	Not Sure	No	No	Select	Other Metro Cities	03.Low	02.Medium	8.0	13.0	No	Yes	Modified	0.01
d er s	Not Sure	No	No	Select	Select	03.Low	02.Medium	11.0	15.0	No	No	Modified	0.02
N	NaN	No	No	Select	Select	03.Low	02.Medium	12.0	15.0	No	No	Modified	0.03
n s	High in Relevance	No	No	Potential Lead	Thane & Outskirts	03.Low	02.Medium	8.0	16.0	No	Yes	Modified	0.03

Final Summary of Lead Scoring in Business Terms:

- So mainly Lead Origin, Tags and Last notable activity are the top 3 variables that contribute most towards the probability of lead getting converted.
- None of the Converted leads are came through Advertisements from Newspaper, Digital Advertisement, Magazines etc… So there should be more focus on advertisement part to make it use effectively.
- People who spent high time on website are likely to be potential leads. So offering discounts to them will leads to increase the conversion rate.
- Students are not likely to be part of the converted leads. So students should get hands on and career opportunities to make them attractive to be part of the leads.
- Many un employed users are likely to be part of the conversion. So targeting un-employed people will be the good advantage to the company.