LEAD SCORING CASE STUDY

UPGRAD

By :
Swathi Setti
Sai Priyesh Chittoori

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. 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. 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% which is very poor rate.

Business Goal

- Our job 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%.

Data Understanding

This assignment has 2 files as explained below:

- ‘Leads.csv’ provides leads dataset from the past with around 9000 data points.
- ‘Leads Data Dictionary.csv' is data dictionary which describes the meaning of the features.

Steps followed for Analysis

1. **Data Cleaning :**
   - ❑ Handling 'Select' & Missing values
   - ❑ Handling Unique columns
   - ❑ EDA to identify variables
     - ❑ Univariate Analysis
     - ❑ Bivariate Analysis
   - ❑ Outlier Treatment

2. **Data Preparation :**
   - ❑ Creating dummy variables
   - ❑ Train and Test Spilt (70-30 ratio)
   - ❑ Scaling of Numerical Variables
     - ❑ Standardized Scaling techniques

3. **Data Modeling :**
   - ❑ Feature Selection using RFE
   - ❑ Apply Logistic Regression using GLM model on train data with RFE
   - ❑ Remove the features which are having P-value is greater than 0.05 and VIF 5
   - ❑ Measuring optimum probability
   - ❑ Model accuracy & other metrics

4. **Predictions :**
   - ❑ This involves making predictions on the test set
   - ❑ Measuring the accuracy and other metrics
   - ❑ Based on the model, identifying the variables which can influence the objective
   - ❑ Draw the recommendations based on the model

Data Cleaning

Inference:
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 modeling we have replaced these values with NAN .

2. Handling missing or Nan values with following steps :
   - Drop those fields with more than 45% missing values
   - Drop the records if the missing % value is lesser than 2%
   - Replacing the NaN values with most occuring values
   - If there is no obvious most occuring value then simply replace NaN with "Others"

```
# Checking the percentage of missing values again

round(100*(df.isnull().sum())/len(df), 2).sort_values(ascending=False)
```

```
How did you hear about X Education              78.46
Lead Profile                                    74.19
Lead Quality                                    51.59
Asymmetrique Profile Score                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Index                     45.65
City                                            39.71
Specialization                                  36.58
Tags                                            36.29
What matters most to you in choosing a course   29.32
What is your current occupation                 29.11
Country                                         26.63
TotalVisits                                      1.48
Page Views Per Visit                             1.48
Last Activity                                    1.11
Lead Source                                      0.39
```

```
# Rechecking % of null value columns

round(100*(df.isnull().sum())/len(df), 2)
```

```
Lead Number                              0.0
Lead Origin                              0.0
Lead Source                              0.0
Converted                                0.0
TotalVisits                              0.0
Total Time Spent on Website              0.0
Page Views Per Visit                     0.0
Last Activity                            0.0
Specialization                           0.0
What is your current occupation          0.0
Tags                                     0.0
City                                     0.0
A free copy of Mastering The Interview   0.0
Last Notable Activity                    0.0
dtype: float64
```

```
# Finally after data cleaning, the percentage of data present when compare with old data.

round((len(df)/9240)*100 , 2)
```

```
99.61
```

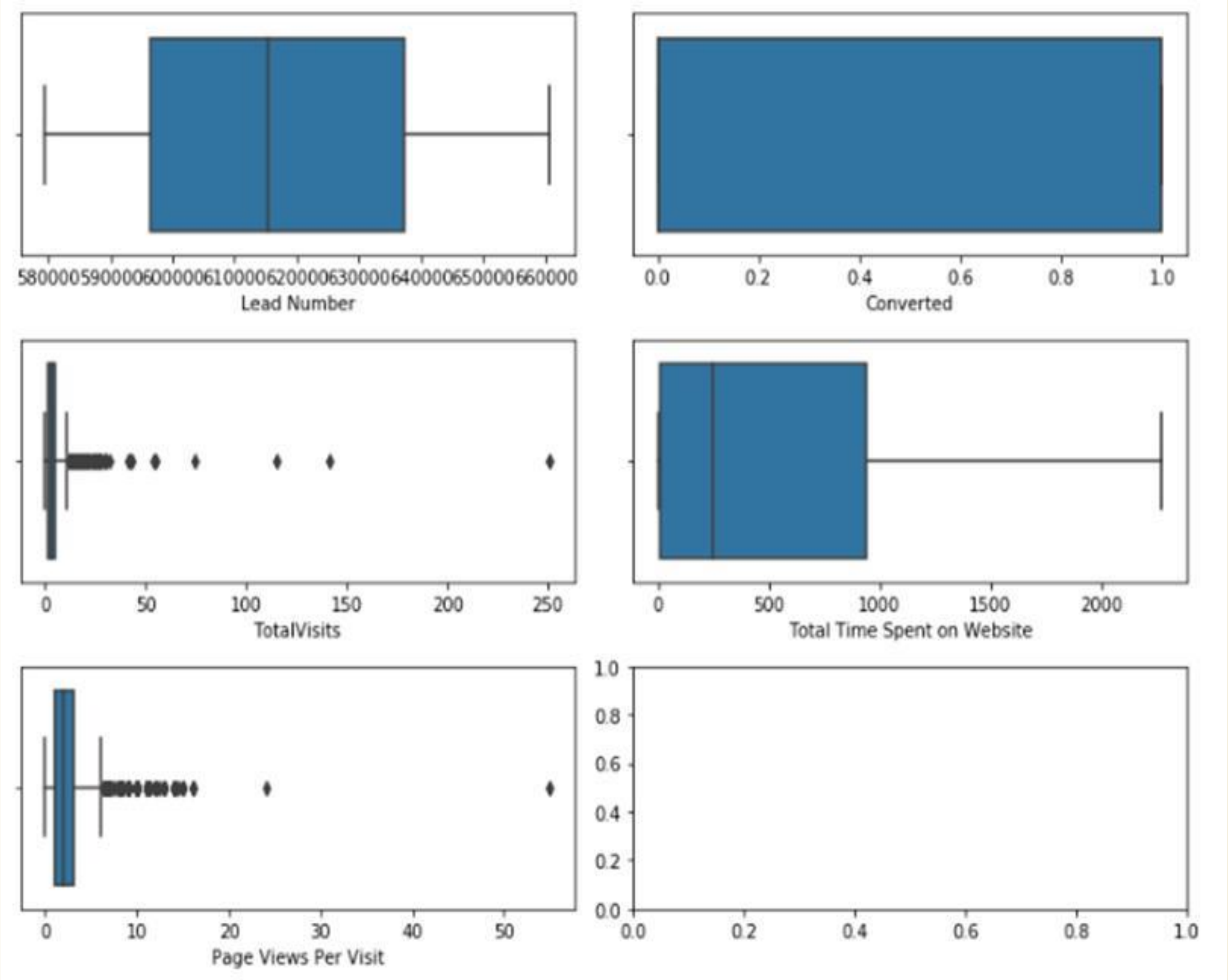
Outlier Analysis & Treatment

- We have observed that there are outliers for few variables (TotalVisits & Page Views Per Visit)

- we will use capping technique to treat the outliers

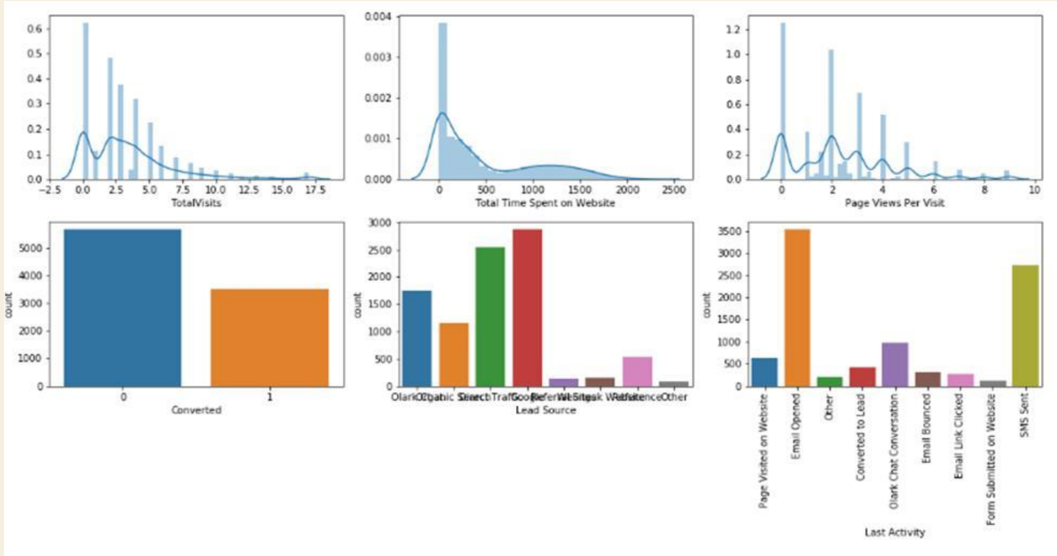
- We used the data <= **0.01** and data >= **0.99** percentile into one group

- As you can observe, most of the outliers have been treated and we can go ahead with this data



EDA (Exploratory Data Analytics)

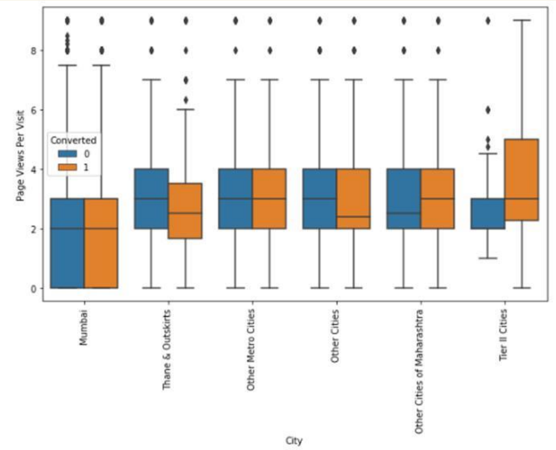
Univariate

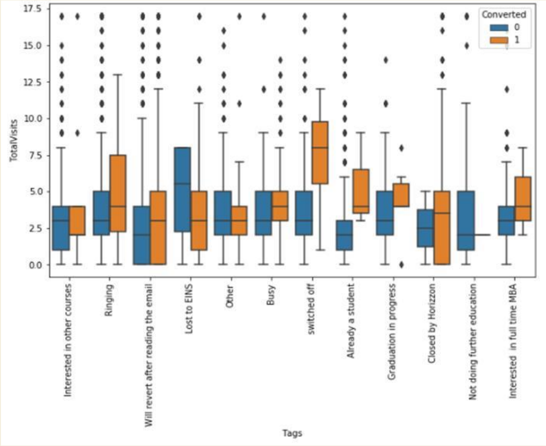


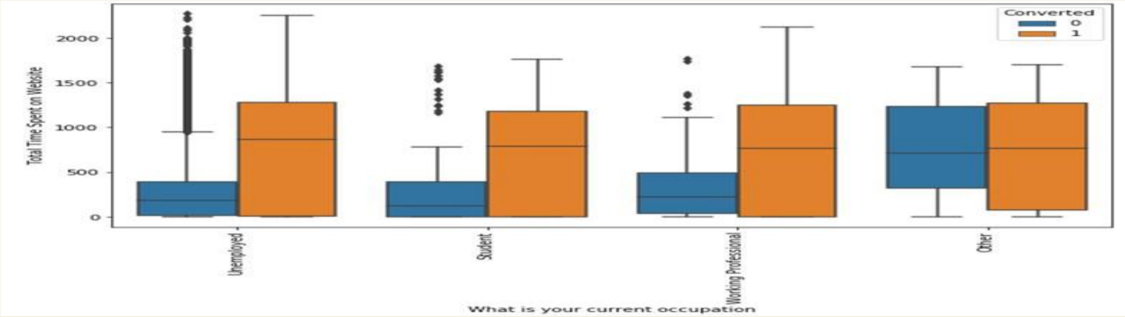
Inference :

•From above graghs we can see lot of variation in 'TotalVisits', 'Total Time Spent on Website'& 'Page Views Per Visit'
•Converted rate is low in comparision
•Lead Source - 'Google' and 'Direct Traffic' have higher count of leads
•Last Activity - 'Email Opened' ; 'SMS Sent' has high number of activities

Bivariate







Inference:

•Total time spent on website is high in Unemployed cases
•Total Visits are high in Switched off and ringing cases
•Page Views per Visit is high in Tier II Cities

Data Preparation

```
df_final.head()
```

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Other
0	0	0.0	0	0.0	0	0	0	0	0	1	0	0
1	0	5.0	674	2.5	0	0	0	0	0	0	1	0
2	1	2.0	1532	2.0	1	1	0	0	0	0	0	0
3	0	1.0	305	1.0	0	1	0	0	0	0	0	0
4	1	2.0	1428	1.0	0	1	0	0	1	0	0	0

```
df_train.describe()
```

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Google	Lead Source_Olark Chat	L Source_Org Se
count	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000000	6442.000
mean	0.394598	0.199376	0.219115	0.262456	0.311549	0.533685	0.076063	0.005123	0.314654	0.186122	0.127
std	0.488802	0.191336	0.241963	0.220388	0.463163	0.498903	0.265120	0.071395	0.464414	0.389236	0.333
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	0.000000	0.058824	0.007923	0.111111	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
50%	0.000000	0.176471	0.114437	0.222222	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000
75%	1.000000	0.294118	0.418904	0.364444	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000

- Creating Dummy Variables for all the categorical variables.
- Scale the necessary variables with standard technique
- 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 variables.
  1. Total Time Spent on Website
  2. Lead Origin_Landing Page Submission
  3. Lead Origin_Lead Add Form
  4. Lead Source_Welingak Website
  5. Last Activity_Email Bounced
  6. Last Activity_Olark Chat Conversation
  7. Specialization_Others
  8. What is your current occupation_Unemployed
  9. What is your current occupation_Working Professional
  10. Tags_Busy
  11. Tags_Closed by Horizzon
  12. Tags_Interested in other courses
  13. Tags_Lost to EINS
  14. Tags_Not doing further education
  15. Tags_Ringing
  16. Tags_Will revert after reading the email
  17. Tags_switched off
  18. Last Notable Activity_Modified
  19. Last Notable Activity_Page Visited on Website
  20. Last Notable Activity_SMS Sent

Model Building

Recursive Feature Elimination (RFE) to perform Variable Selection

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6857	0.389	-4.336	0.000	-2.448	-0.924
Total Time Spent on Website	4.3824	0.201	21.797	0.000	3.988	4.776
Lead Origin_Landing Page Submission	-1.6574	0.161	-10.286	0.000	-1.973	-1.342
Lead Origin_Lead Add Form	1.8534	0.285	6.502	0.000	1.295	2.412
Lead Source_Welingak Website	3.4239	1.055	3.246	0.001	1.357	5.491
Last Activity_Email Bounced	-1.6532	0.357	-4.629	0.000	-2.353	-0.953
Last Activity_Olark Chat Conversation	-1.1006	0.199	-5.523	0.000	-1.491	-0.710
Specialization_Others	-1.4786	0.161	-9.163	0.000	-1.795	-1.162
What is your current occupation_Unemployed	-1.5506	0.324	-4.792	0.000	-2.185	-0.916
What is your current occupation_Working Professional	0.9632	0.407	2.368	0.018	0.166	1.760
Tags_Busy	3.1821	0.317	10.039	0.000	2.561	3.803
Tags_Closed by Horizzon	8.4608	0.766	11.050	0.000	6.960	9.962
Tags_Interested in other courses	-0.6070	0.517	-1.175	0.240	-1.620	0.406
Tags_Lost to EINS	8.8269	0.773	11.424	0.000	7.312	10.341
Tags_Not doing further education	-1.2619	1.168	-1.081	0.280	-3.550	1.026
Tags_Ringing	-1.1363	0.324	-3.511	0.000	-1.771	-0.502
Tags_Will revert after reading the email	3.5136	0.234	15.017	0.000	3.055	3.972
Tags_switched off	-1.4401	0.585	-2.460	0.014	-2.587	-0.293
Last Notable Activity_Modified	-1.0525	0.109	-9.700	0.000	-1.265	-0.840
Last Notable Activity_Page Visited on Website	-0.8796	0.248	-3.550	0.000	-1.365	-0.394
Last Notable Activity_SMS Sent	2.0652	0.121	17.075	0.000	1.828	2.302

- Removed below features which are having P-value is greater than 0.05 and VIF 5
  - Tags_Not doing further education
  - Tags_Interested in other courses
  - What is your current occupation_Unemployed
  - Tags_switched off
  - Tags_Ringing
- All the VIFs & p-values are now in the appropriate range. We are good to go with further predictions

Variance Inflation Factor (VIF) to measure co-linearity

	vars	VIF
7	What is your current occupation_Unemployed	14.555034
15	Tags_Will revert after reading the email	6.613062
1	Lead Origin_Landing Page Submission	5.460936
6	Specialization_Others	3.924065
17	Last Notable Activity_Modified	2.288804
8	What is your current occupation_Working Profes...	2.248977
14	Tags_Ringing	2.223206
0	Total Time Spent on Website	2.162898
2	Lead Origin_Lead Add Form	1.959888
19	Last Notable Activity_SMS Sent	1.757316
10	Tags_Closed by Horizzon	1.610642
11	Tags_Interested in other courses	1.495424
5	Last Activity_Olark Chat Conversation	1.461445
3	Lead Source_Welingak Website	1.350524
16	Tags_switched off	1.256804
9	Tags_Busy	1.224686
12	Tags_Lost to EINS	1.209555
13	Tags_Not doing further education	1.174179
4	Last Activity_Email Bounced	1.139064
18	Last Notable Activity_Page Visited on Website	1.097202

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	Predicted
5310	0	0.033663	0
2181	0	0.006498	0
8215	0	0.034536	0
8887	0	0.778241	1
7920	0	0.278894	0

```
# Confusion matrix

confusion = metrics.confusion_matrix(y_train_pred_final['Converted'], y_train_pred_final['Predicted'] )
print(confusion)
```

```
[[3624  276]
 [ 364 2178]]
```

```
# Predicted     not_lead     lead
# Actual
# not_lead        3624        276
# lead            364         2178
```

```
# Let's check the overall accuracy.

print(metrics.accuracy_score(y_train_pred_final['Converted'], y_train_pred_final['Predicted']))
```

```
0.9006519714374418
```

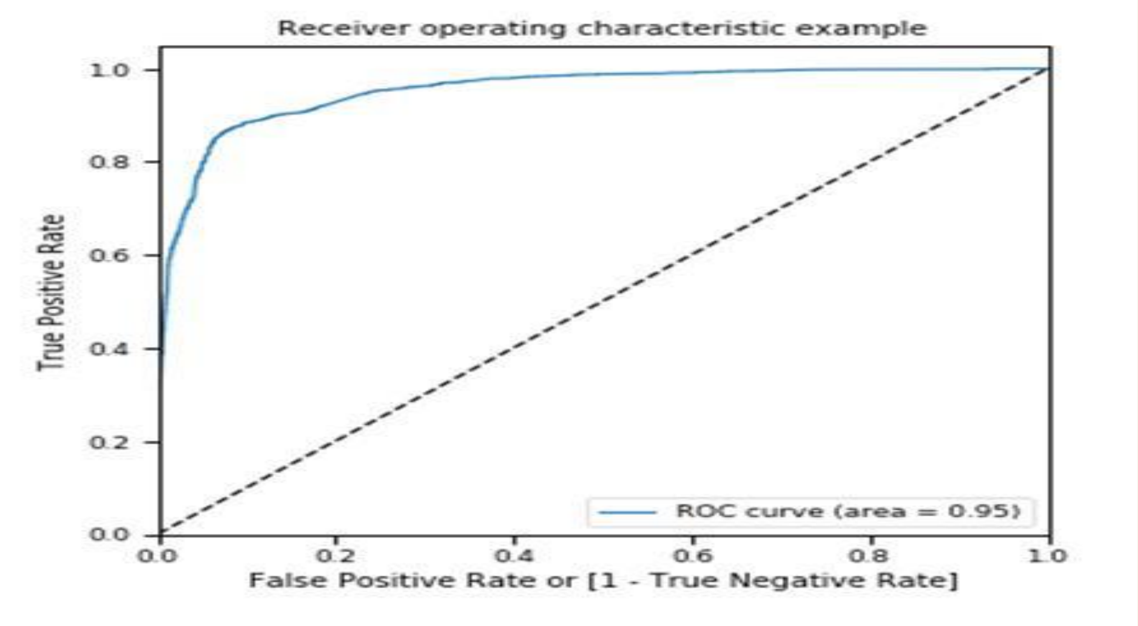
Model Evaluation

- **Confusion matrix is** used :

- 3624 customer were not converted, and model also predicted them as non potential leads, these are called True Negatives(TN)
- 276 customers were wrongly predicted as potential customers while they were not actually these are called False Positives(FP)
- 364 customers were converted but the model predicted them as non potential leads these are called False Negative(FN).
- 2178 customers who were converted was correctly predicted as potential leads these are called True Positives(TP)

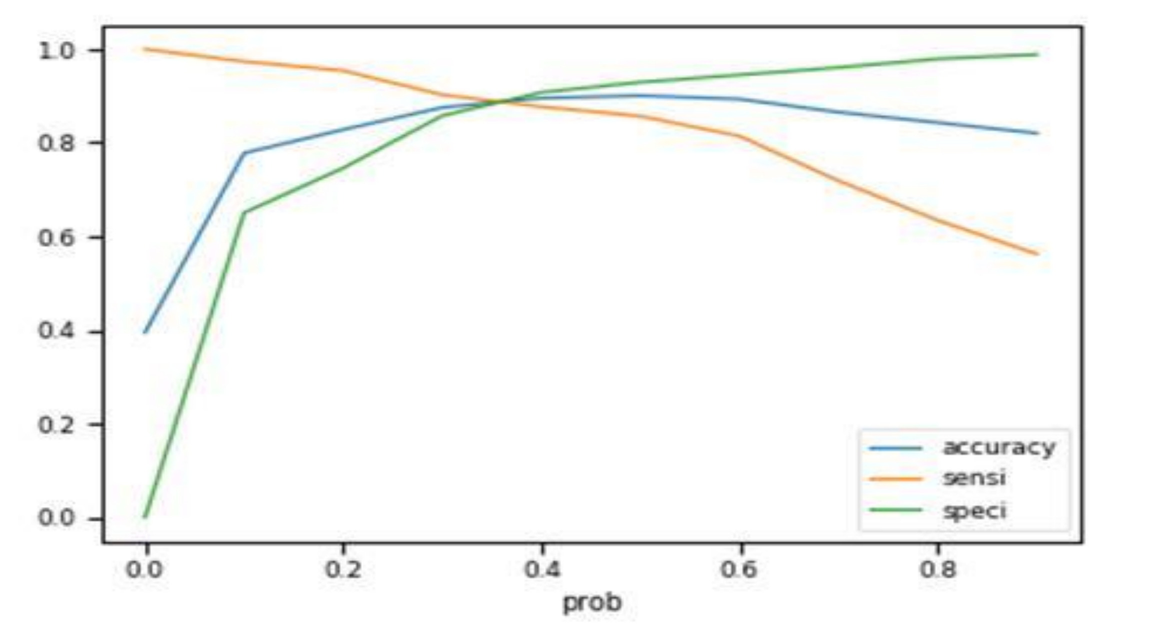
- ***That's around 90%accuracy which is a very good value***

ROC Curve



- The area under the curve of the ROC is 0.89 which is quite good.
- Model seems to be accurate.
- Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.

Optimal Cutoff Point



- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
- From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Train – Accuracy, Precision and Recall

- **Accuracy** - 0.888699161751009
- Sensitivity (Recall) is 0.8886703383162864
- Specificity (Precision) is 0.8887179487179487
- Positive predictive value is 0.8388414407723728
- Negative predictive value is 0.9245132035209389

Test – Accuracy, Precision and Recall

- **Accuracy** - 0.8924692251991311
- Sensitivity (Recall) is 0.8888888888888888
- Specificity (Precision) is 0.8944695259593679
- Positive predictive value is 0.8247422680412371
- Negative predictive value is 0.9351032448377581

```
# metric

# [ TN   FP ]
# [ FN   TP ]

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

Recommendations

Since 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.
The top three variables that contribute towards the probability of a lead getting converted are:

- Leads who tag with Lost to EINS
- Leads tag with Closed by Horizon
- Total Time Spent on Website

Phone calls should be done for the following people:

- They spend a lot of time in the website and this can be done by making the website interesting and thus bringing them back to the site
- They are seen coming back to the website repeatedly
- Their last activity is through SMS or through Olark chat conversation
- They are working professionals

THANK YOU