LEAD SCORING CASE STUDY

SONJOY ROY CHOUDHURY

Problem Statement

The Case Study aims to identify the indicating factors that results in leads getting converted i.e., the leads that are most likely to convert into paying customers. The company requires a model wherein a lead score needs to be added 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%.

Identification of such leads with high lead score(i.e. higher rate of conversion) using the EDA and ML algorithm is the aim of the provided Case Study.

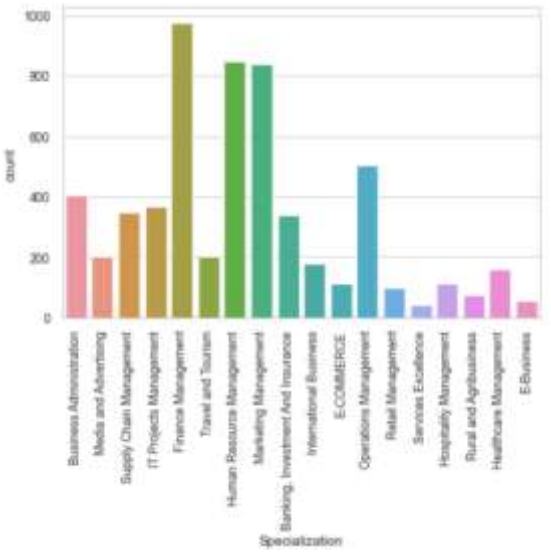
Steps Involved in Analysis.

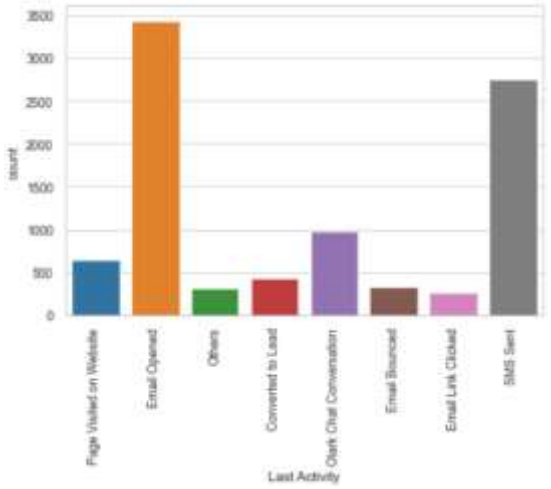
Lead Data

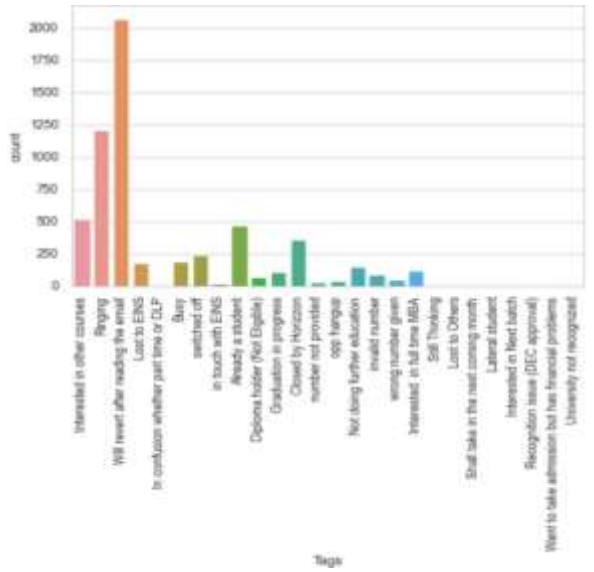
- Importing Libraries
- Inspecting the dataframe.
- Cleaning the Data.
- Analysing the Outliers.
- Univariate Analysis.
- Bivariate and Multivariate Analysis.
- Train and Test Set.
- Feature Scaling.
- Feature Scaling using RFE.
- Checking VIF's.
- Calculating metrics beyond accuracy.
- Plotting ROC curve.
- Finding optimal cut-off point.
- Precision and Recall.
- Predictions on Test Set.
- Calculating lead score for entire dataset.
- Conclusion.

Lead Data

- Importing the necessary libraries.
- Then we inspect the dataframe and understand all the attributes.
- Finding the missing data from the dataframe and then find the percentage of null values,
- Then taking the data drop percentage(40%) and removing columns accordingly from the dataset.
- We then clean the data left and replace the left-over null values by mean value of the respective column as implemented on ’Specialization’, ‘Tags’, ‘Last Activity’ and other columns.

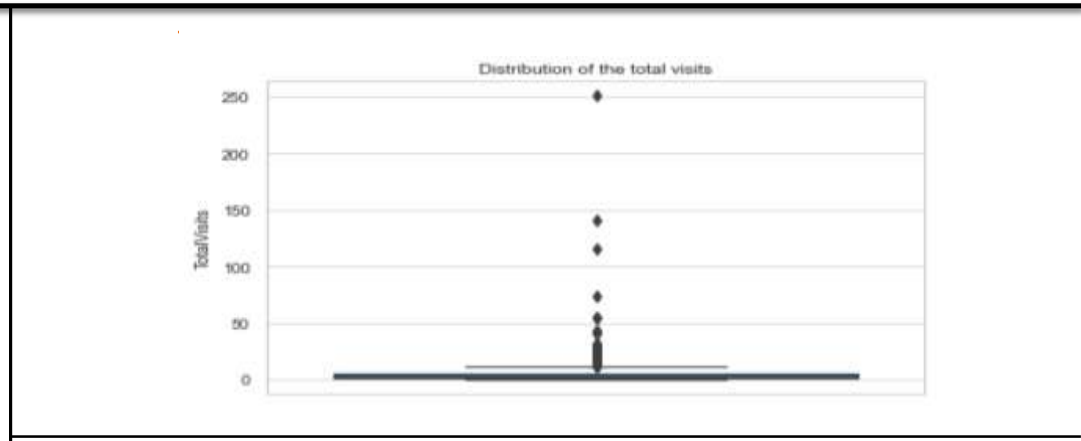


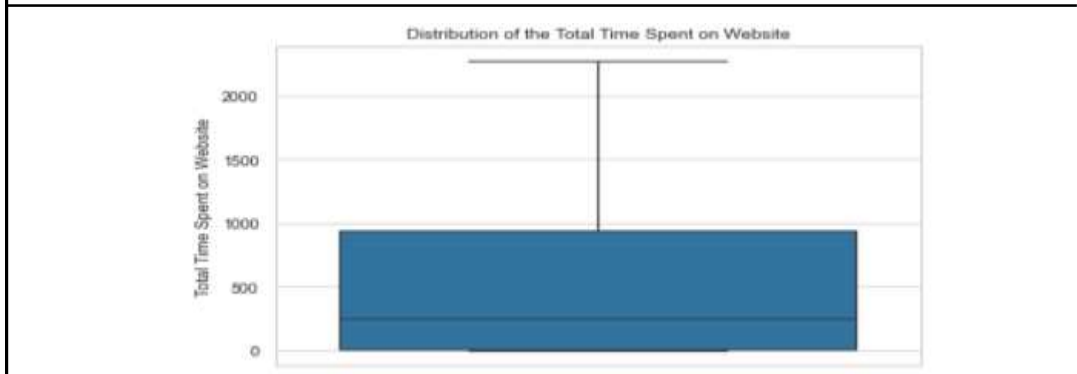


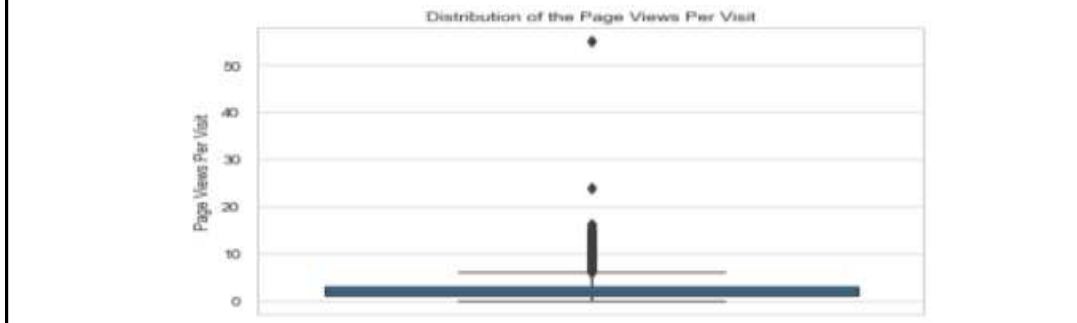


Analyzing the Outliers

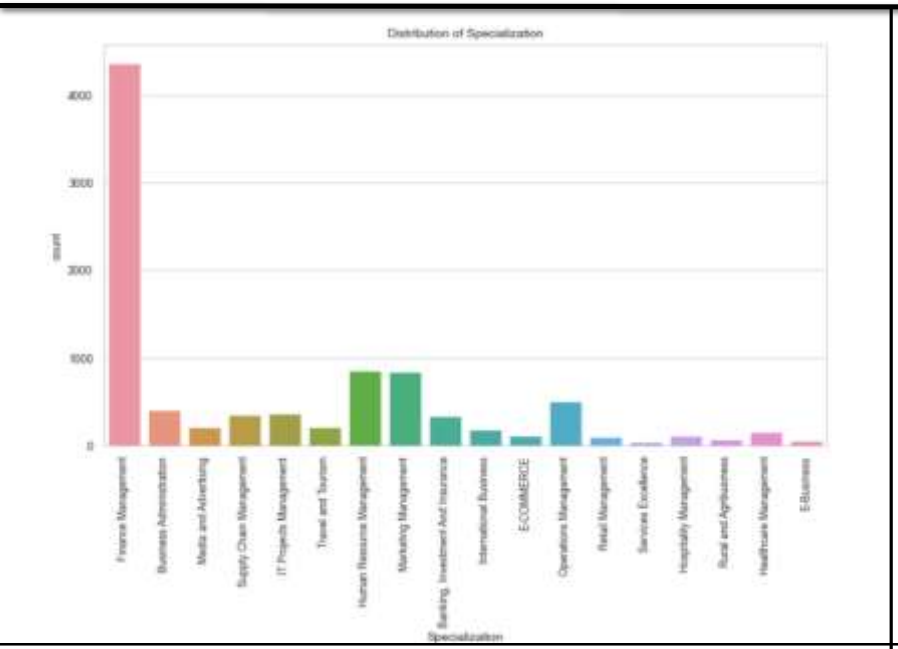
- Analyzing ‘TotalVisits’ we can observe it has quite a few outliers.
- Analyzing ‘Total Time Spent on Website’ we can observe that it has no outliners.
- Analyzing ‘Page Views Per Visit’ we can observe that it has huge number of outliners.

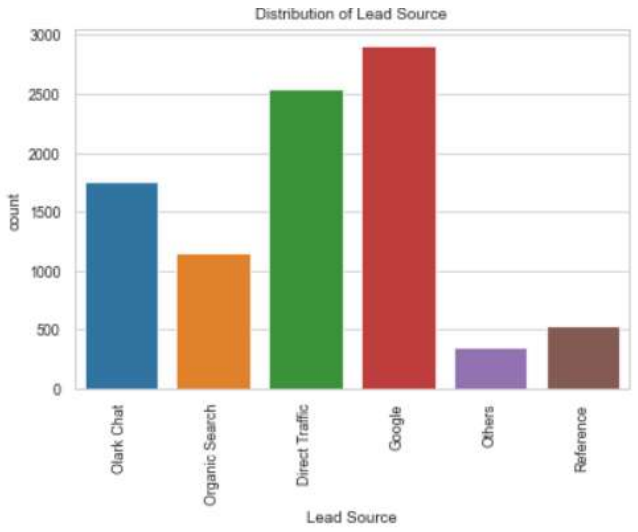


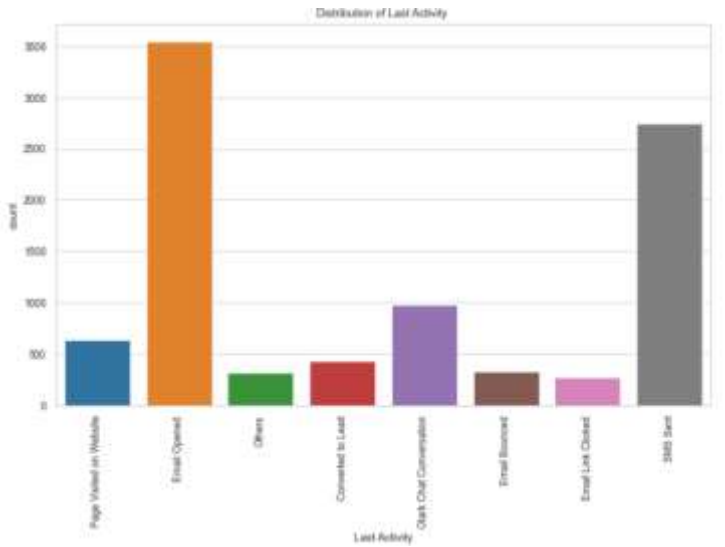




Univariate Analysis







- In the above analysis of Specialization, we can see that Finance Management has the highest count of Lead generation.

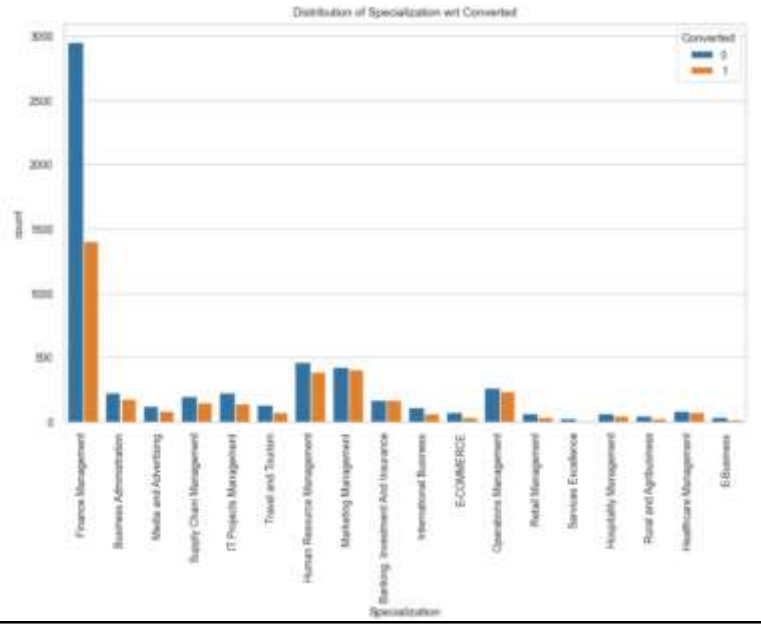
- In the above analysis of Lead Score, we can see that Google has the highest count of Lead generation.

- In the above analysis of Lead Activity ,we can see that Email Opened has the highest count of Lead geration.

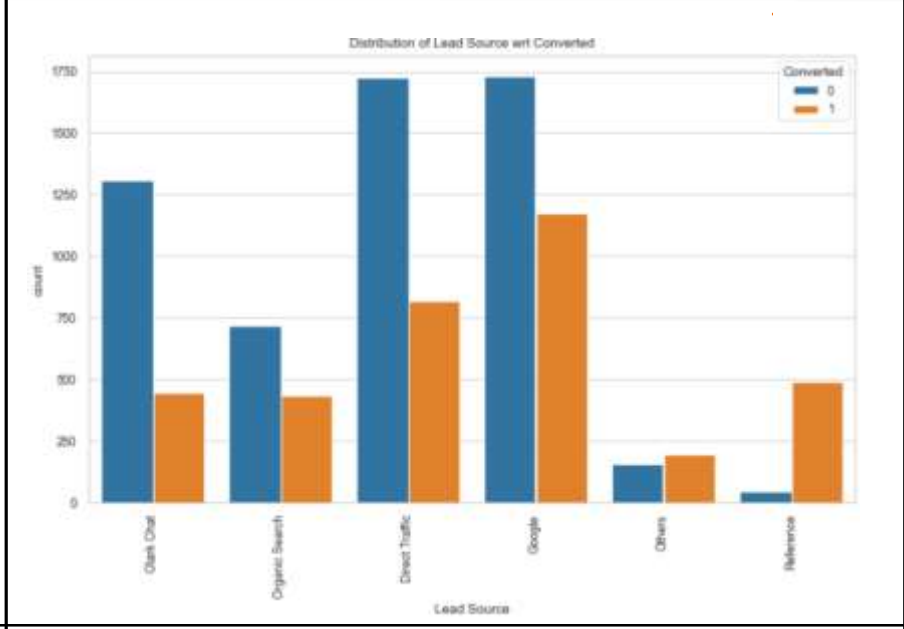
Note:-

* Similar to the above observation's, univariate analysis on some other attributes like 'Lead Origin' and 'Tags' is also done which provides interesting view on the leads for the enrollment.

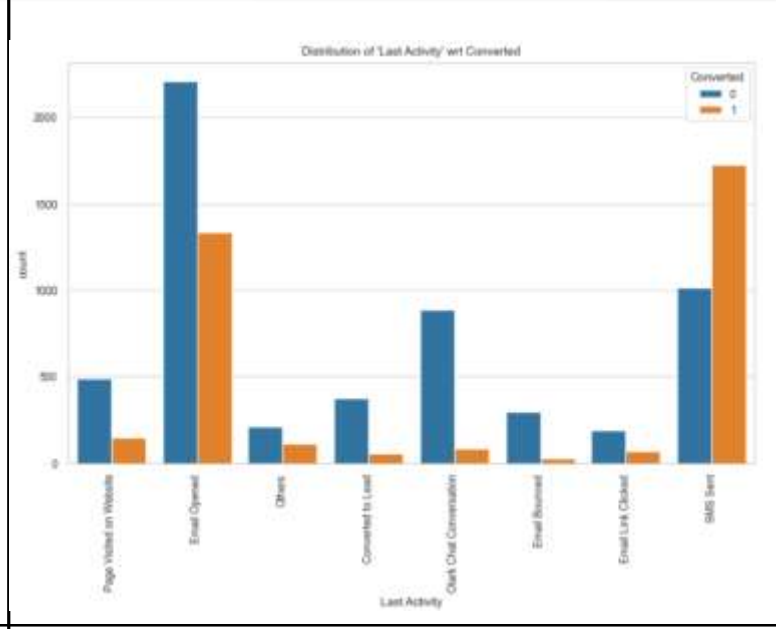
Bivariate Analysis



- In the above analysis we can say that the maximum focus for lead conversion should be at specialization with high conversion rates like Financial Management.



- In the above analysis, we can see that Google and Direct Traffic generate the highest number of leads.



- In the above analysis, we can observe that most of the leads have Email Opened as their last activity. Also SMS Sent has highest conversion rate of leads.

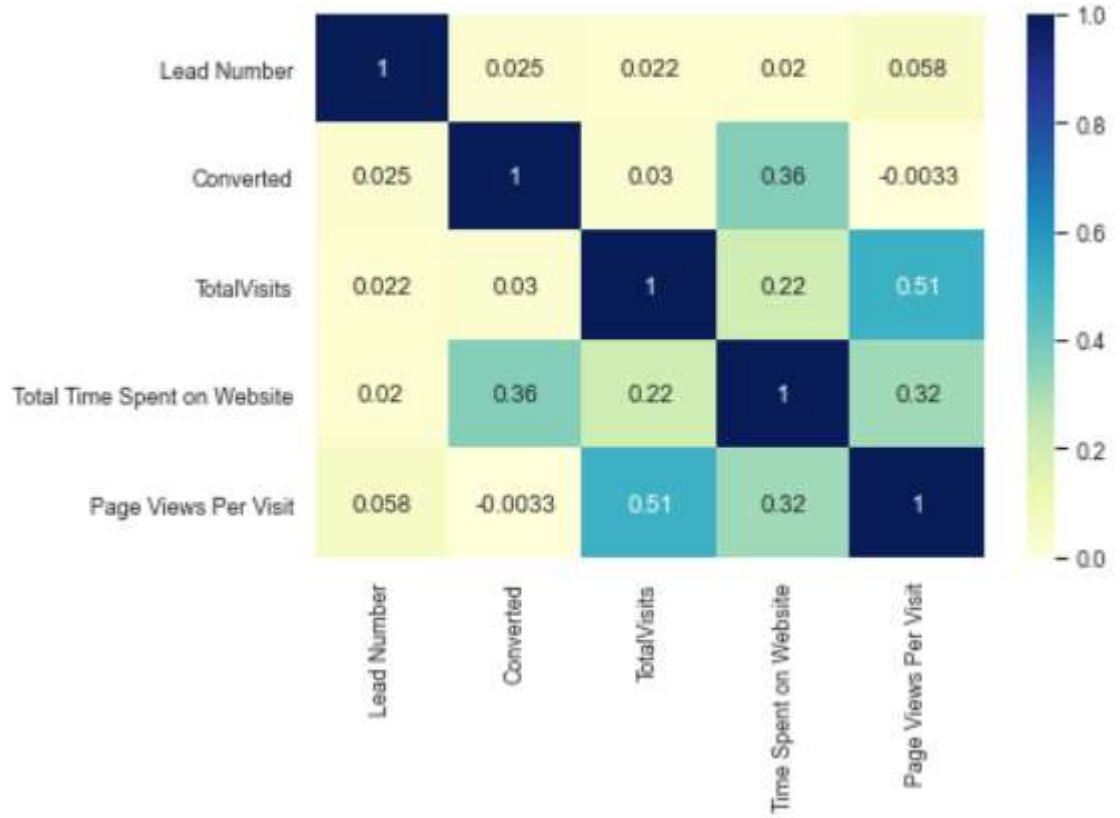
Note:-

* Similar to the above observation's, bivariate analysis on some other attributes like 'Lead Origin', 'Tags', 'TotalVisits', etc is also done which provides interesting view on the leads for the enrollment.

Multivariate Analysis

- There is low variation in Page Views Per Visit and TotalVisits but higher variation in Total Time Spent on Website.
- There are outliers in Page Views Per Visit and TotalVisits which won't be treated further since they look legimate.
- There is positive correlation between Total Time Spent on Website and Conversion
- There is some correlation between Conversion and some categorical columns like Lead Origin and Lead Source.
- There is almost no correlation in Page Views Per Visit and TotalVisits with Conversion.
- Median of time spent on website was more for positive conversion.

.



* To check the Top 10 Correlation for Repayer and Defaulter client, kindly go through the program.

Test-Train and Split along with Feature Scaling

- We first divide the data frame into Training set and Test Set.
- The split ratio for the Train and Test set is 70:30
- The Feature Scaling is also applied on the Training set it is applied on attributes like ‘TotalVisits’, ‘Total Time Spent on Website’ and 'Page Views Per Visit’.
- We then import the statsmodels and apply GLM.
- We then also use Feature Selection using RFE, we start with 20 attributes.
- We then check the VIF’s for the X_train set and check the multicollinearity in the attributes if any.
- Then we drop the attributes multiple times to get p-value <0.5 and the VIF as less as possible.

	Features	VIF
9	Tags_Will revert after reading the email	2.60
19	Last Activity_SMS Sent	2.27
17	Last Activity_Email Opened	2.12
7	Tags_Ringing	1.46
15	Lead Origin_Lead Add Form	1.36
2	Tags_Closed by Horizzon	1.29
18	Last Activity_Others	1.10
13	Tags_switched off	1.10
0	Total Time Spent on Website	1.09
1	Tags_Busy	1.08
16	Last Activity_Email Bounced	1.08
6	Tags_Lost to EINS	1.05
11	Tags_invalid number	1.03
14	Tags_wrong number given	1.02
12	Tags_number not provided	1.01
4	Tags_Interested in Next batch	1.01
3	Tags_Diploma holder (Not Eligible)	1.01
10	Tags_in touch with EINS	1.01
8	Tags_Shall take in the next coming month	1.00
5	Tags_Lateral student	1.00

The VIF check for the first 20 attributes.

* To check the final model and the remaining attributes their p-value and VIF, refer to the ipynb python file.

Matrix beyond Accuracy

- The sensitivity of our logistic regression model is 0.8325.
- The specificity of the model is0.90154.
- The False positive rate is 0.0984.
- The Positive predictive value is 0.8389.
- The Negative predictive value is 0.8972.

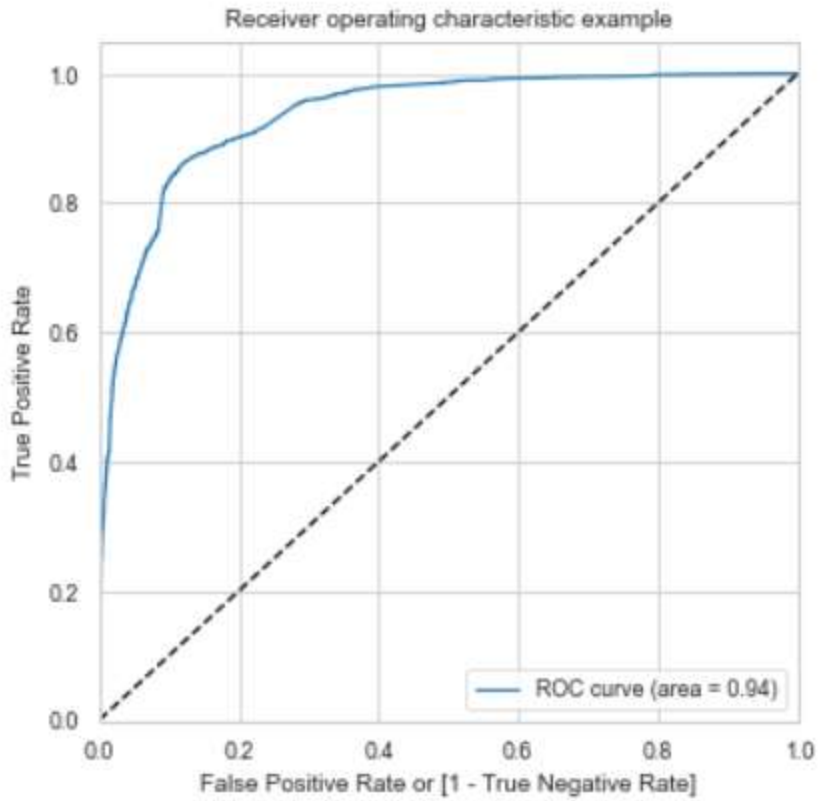
```
[[3608  394]
 [ 413 2053]]
```

Confusion Matrix

TN= 3608
FP= 394
FN= 413
FP= 2053

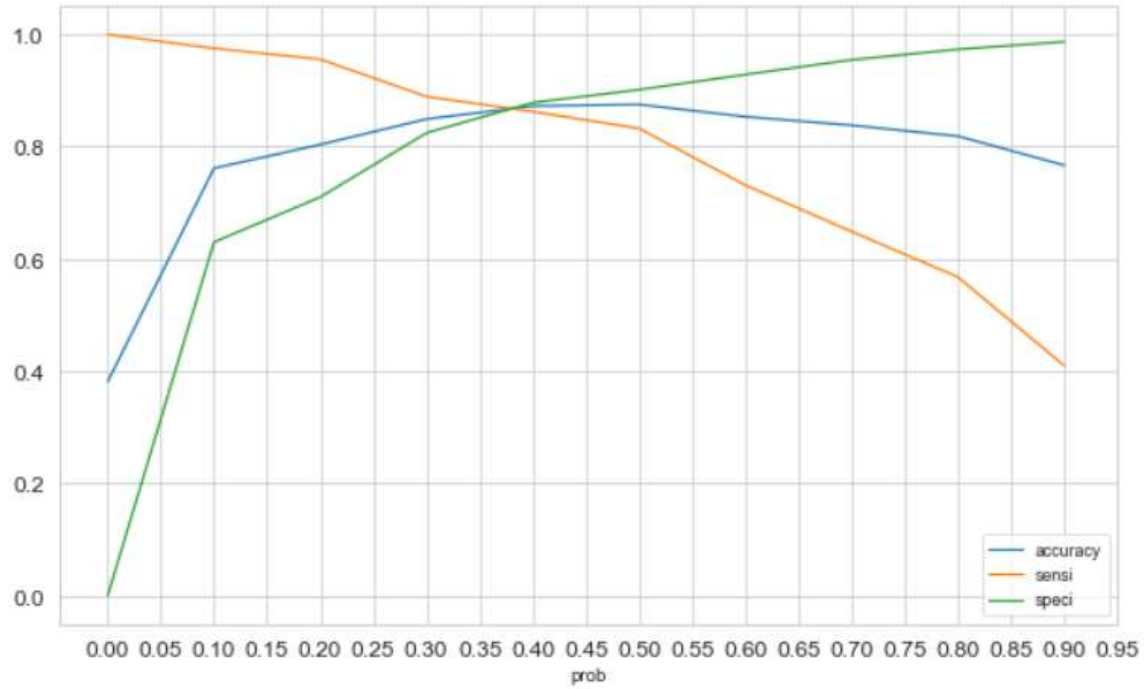
ROC Curve

- The ROC curve shows the tradeoff between the False Positive Rate and True Positive Rate.
- As we can observe that the curve follows the left-hand border and then the top border of the ROC space, it shows that the curve is accurate.
- The area under the curve is 0.936.



Finding the Optimal Cut-off point.

	prob	accuracy	sensi	speci
0.00	0.00	0.38	1.00	0.00
0.10	0.10	0.76	0.98	0.63
0.20	0.20	0.80	0.96	0.71
0.30	0.30	0.85	0.89	0.82
0.40	0.40	0.87	0.86	0.88
0.50	0.50	0.88	0.83	0.90
0.60	0.60	0.85	0.73	0.93
0.70	0.70	0.84	0.65	0.95
0.80	0.80	0.82	0.57	0.97
0.90	0.90	0.77	0.41	0.99



- The above table depicts the various cut-off and calculates the accuracy, sensitivity and specificity.

- From the curve above, 0.37 is the optimum point to take it as a cutoff probability..

Predictions on Test Set

- The Feature Scaling is also applied on the Training set it is applied on attributes like ‘TotalVisits’, ‘Total Time Spent on Website’ and 'Page Views Per Visit’.
- The Overall accuracy is 86%.
- The Sensitivity is 0.8776.
- The specificity of the model is0.855.
- The False positive rate is 0.1449.
- The Positive predictive value is 0.7981.
- The Negative predictive value is 0.9145.
- The Precision is 0.7981.
- The Recall rate is 0.877.
- The F1 score is 0.8360.

```
[[1434  243]
 [ 134  961]]
```

Confusion Matrix

TN= 1434
FP= 243
FN= 134
FP= 961

Conclusion

- The company can increase the engagement of the clients on the Website.
- Increase the amount of SMS sent to the students or potential customers to generate active or hot leads.
- Focus on Tags_Closed by Horizzon, Tags_lost to EINS and Tags_Will revert after reading the email.

Features with Positive Coefficient Values
Tags_Closed by Horizzon
Tags_Lost to EINS
Tags_Will revert after reading the email
Lead Origin_Lead Add Form
Tags_Busy
Tags_in touch with EINS
Last Activity_SMS Sent
Last Activity_Email Opened
Last Activity_Others
Total Time Spent on Website

- The company should work on improving the Tags_Ringing to improve the active lead generation.
- The company's lead generation gets effected due to Email getting bounced.
- Also Tags like Invalid number also results in failure to gather active leads as the customer is not reached.

Features with Negative Coefficient Values
Tags_Ringing
Last Activity_Email Bounced
Tags_invalid number