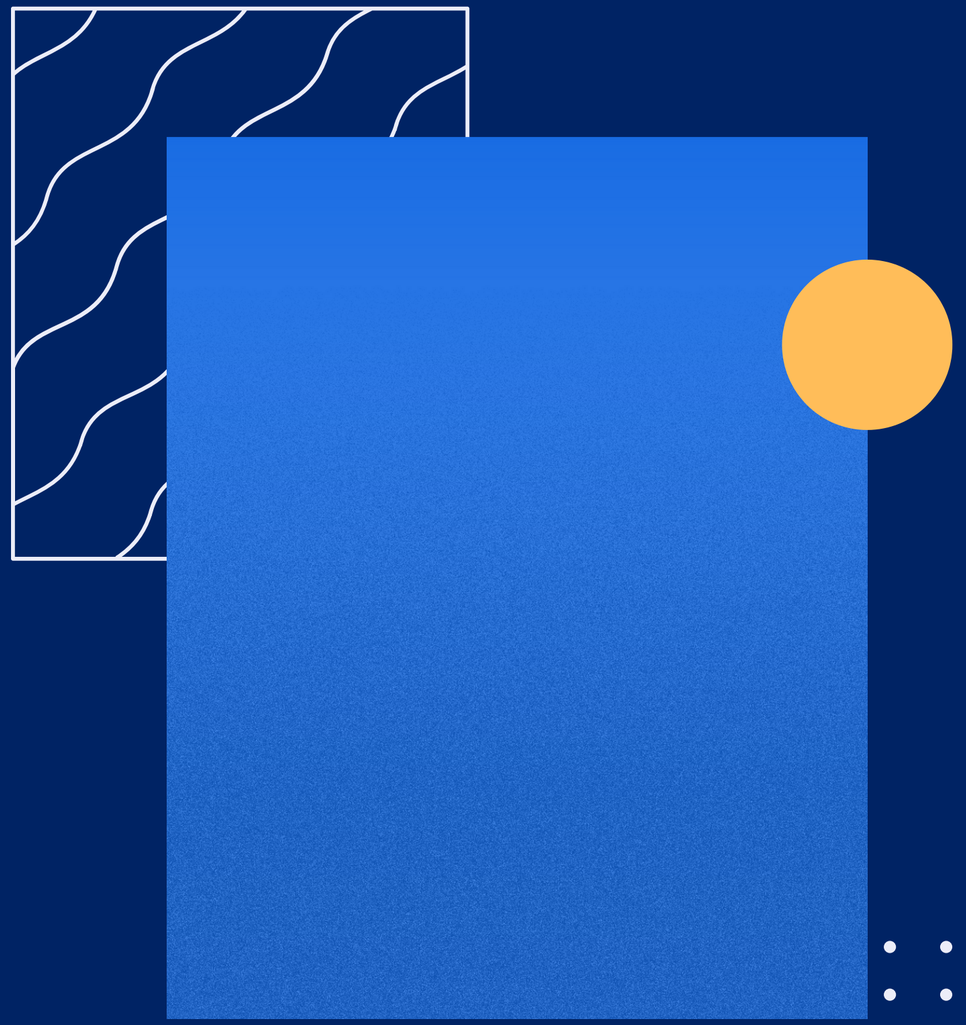
LEAD SCORING CASE STUDY



BY:
SHIVANG NAIK
AISHWARYA GYANJYOTI

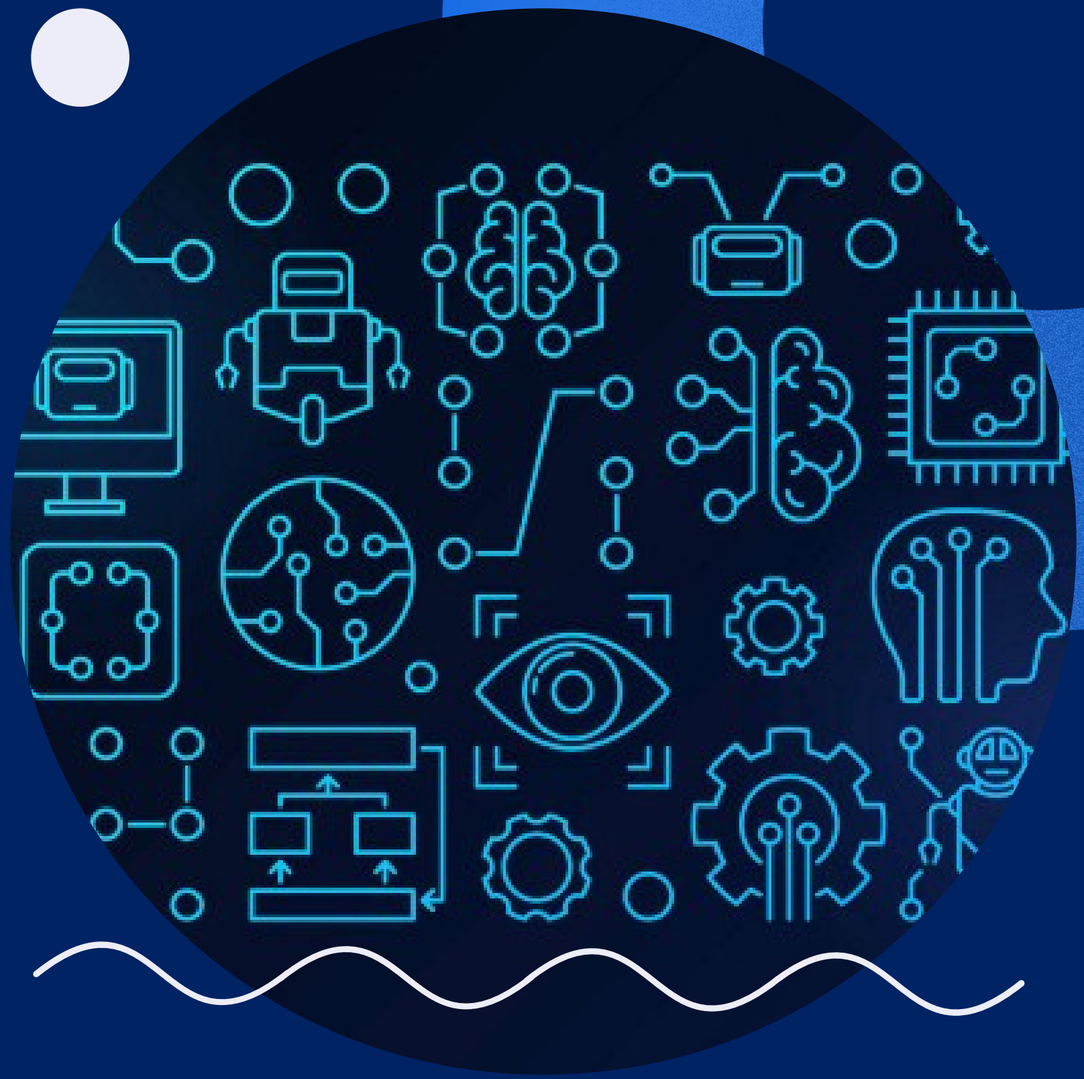
PROBLEM STATEMENT

- Building a logistic regression model for X education by assigning a lead score between 0 and 100 for targeting particular leads to be converted or not. Leads as in finally enrolling for a course by giving contact details like phone number and email address.

- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

- Adjusting the company's requirement for further changes in the future.

APPROACH



Exploratory data analysis and Data Preparation

By creating dummy variables, Train-test split, Scaling of data and RFE

Matrix Score test

Finding confusion matrix, Accuracy along with sensitivity and specificity. Plotting ROC curve and finding optimal cut-off along with precision recall curve.

Data Inspection and Cleaning

Checking the information, numerical description, null values of the dataset, and removing unwanted columns.

Building a logistic regression model

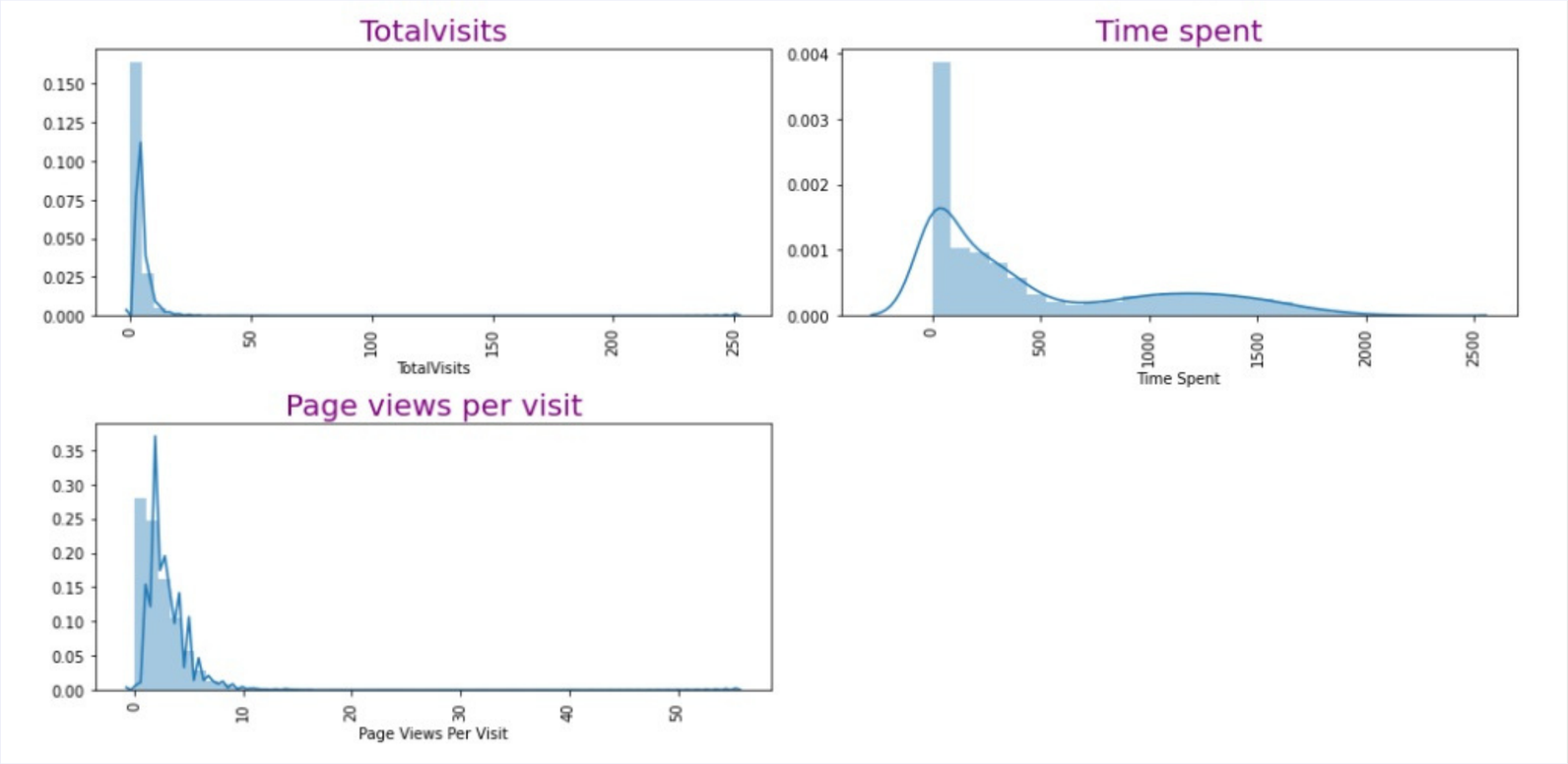
Features to be eliminated one by one having high P-values and VIF.

Prediction on Test Data

The final prediction of test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

DATA VISUALIZATION

UNIVARIATE ANALYSIS

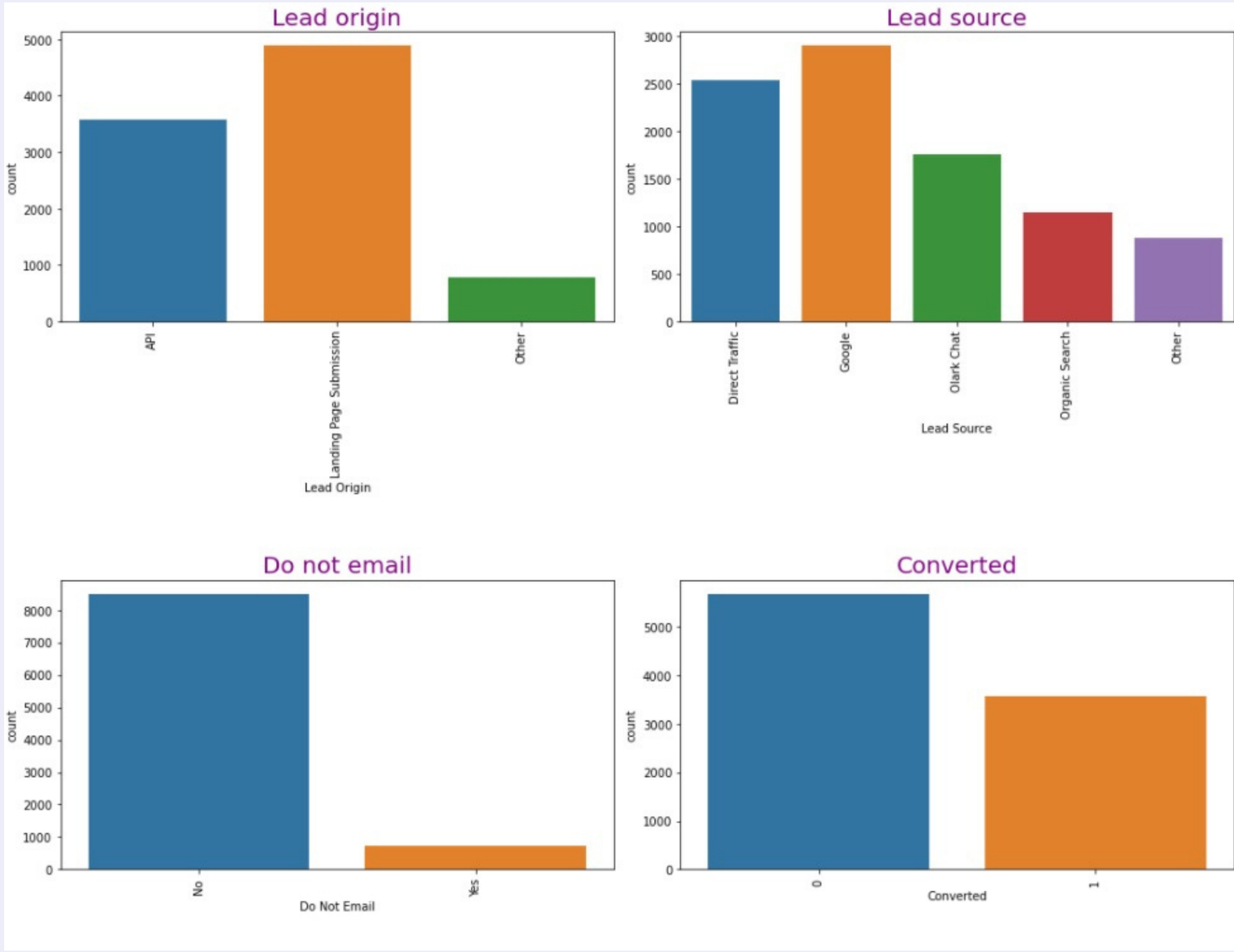


Normal distribution of the above variables.

- Landing page submission seems to have the majority among other origins of the lead.

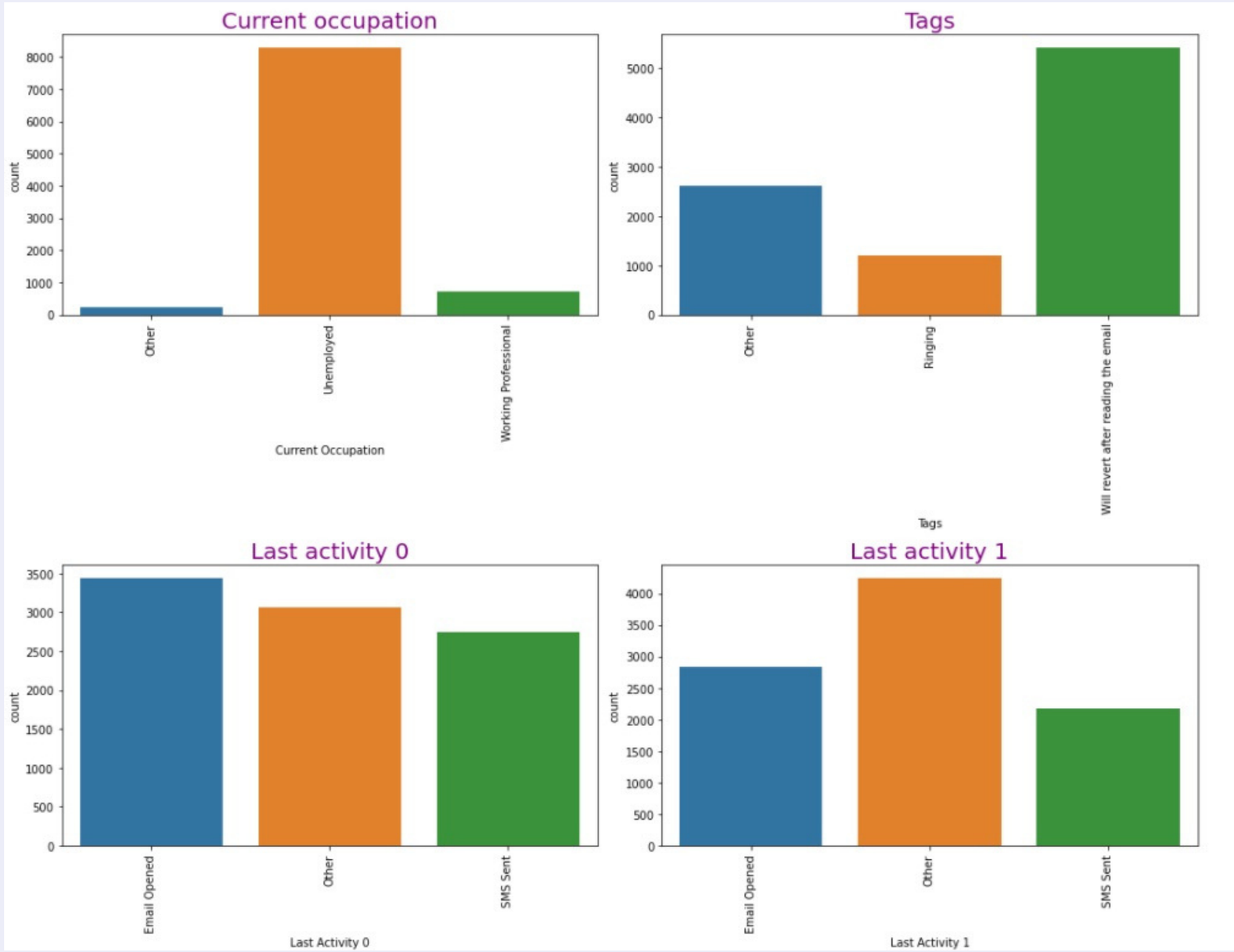
- 5000+ leads aren't converted yet i.e., they won't enroll in the course.

- Seems like the customer doesn't opt for receiving emails related to the course after looking at "Don Not Email".

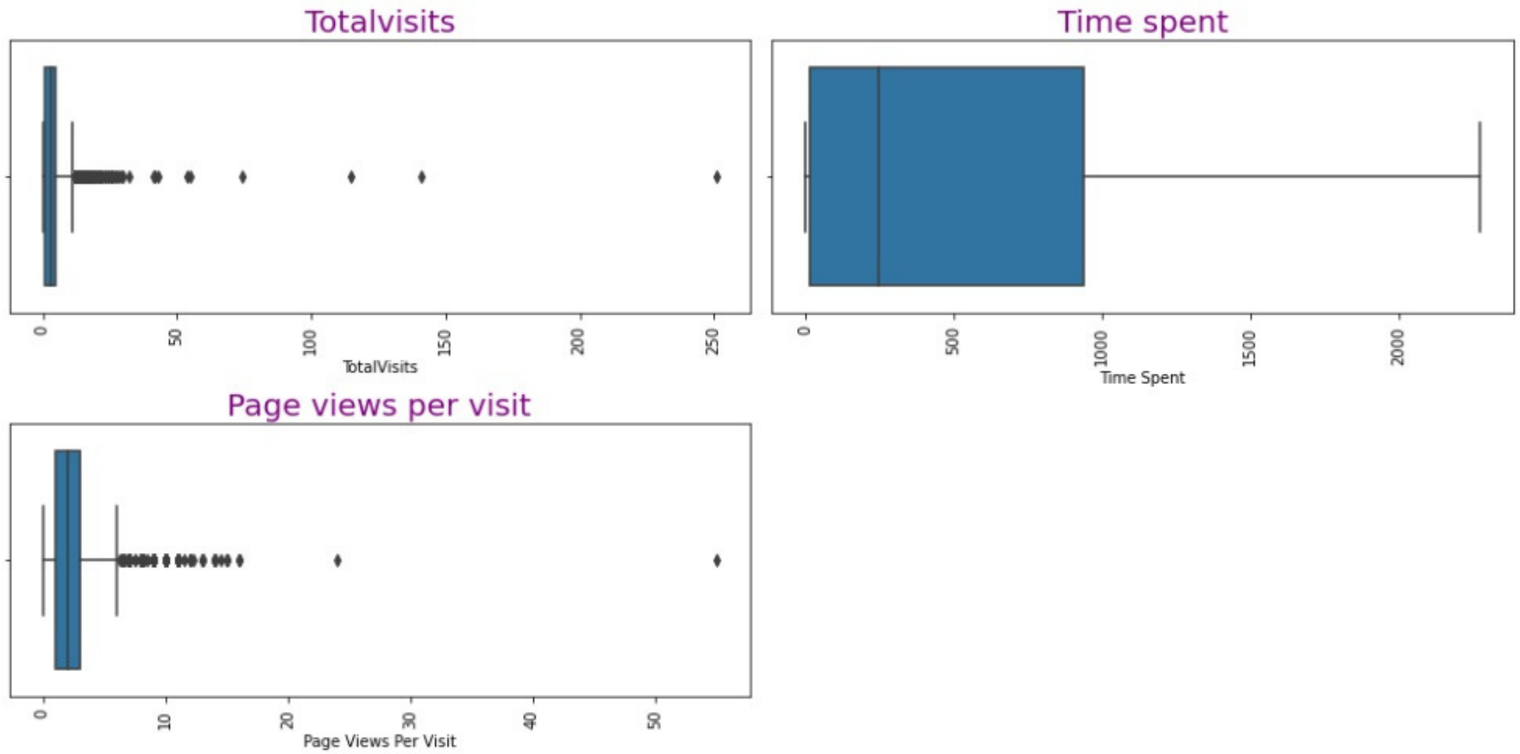


- 8000+ customers with current occupation as "unemployed" are interested in this course.

- As per the current status in"Tags", most of them will revert after receiving the mail

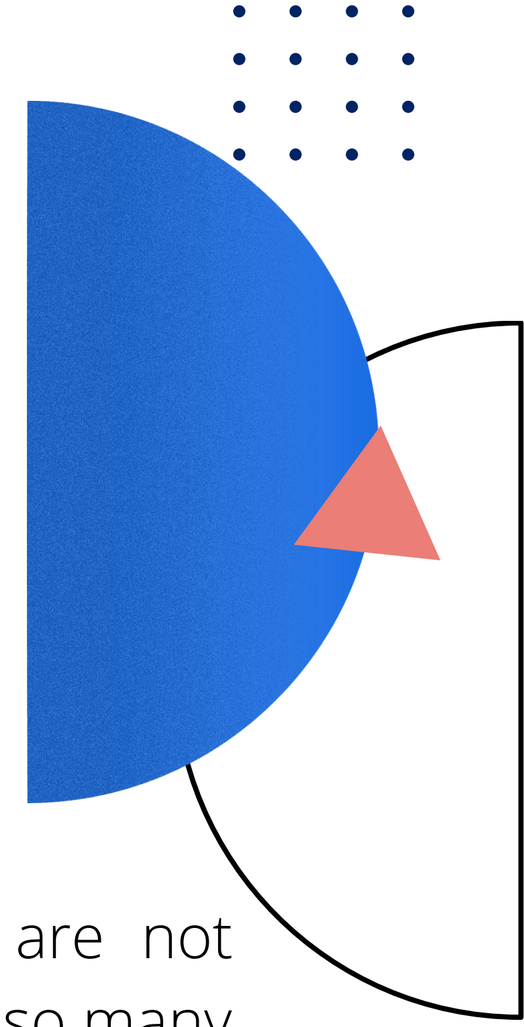


CHECKING OUTLIERS



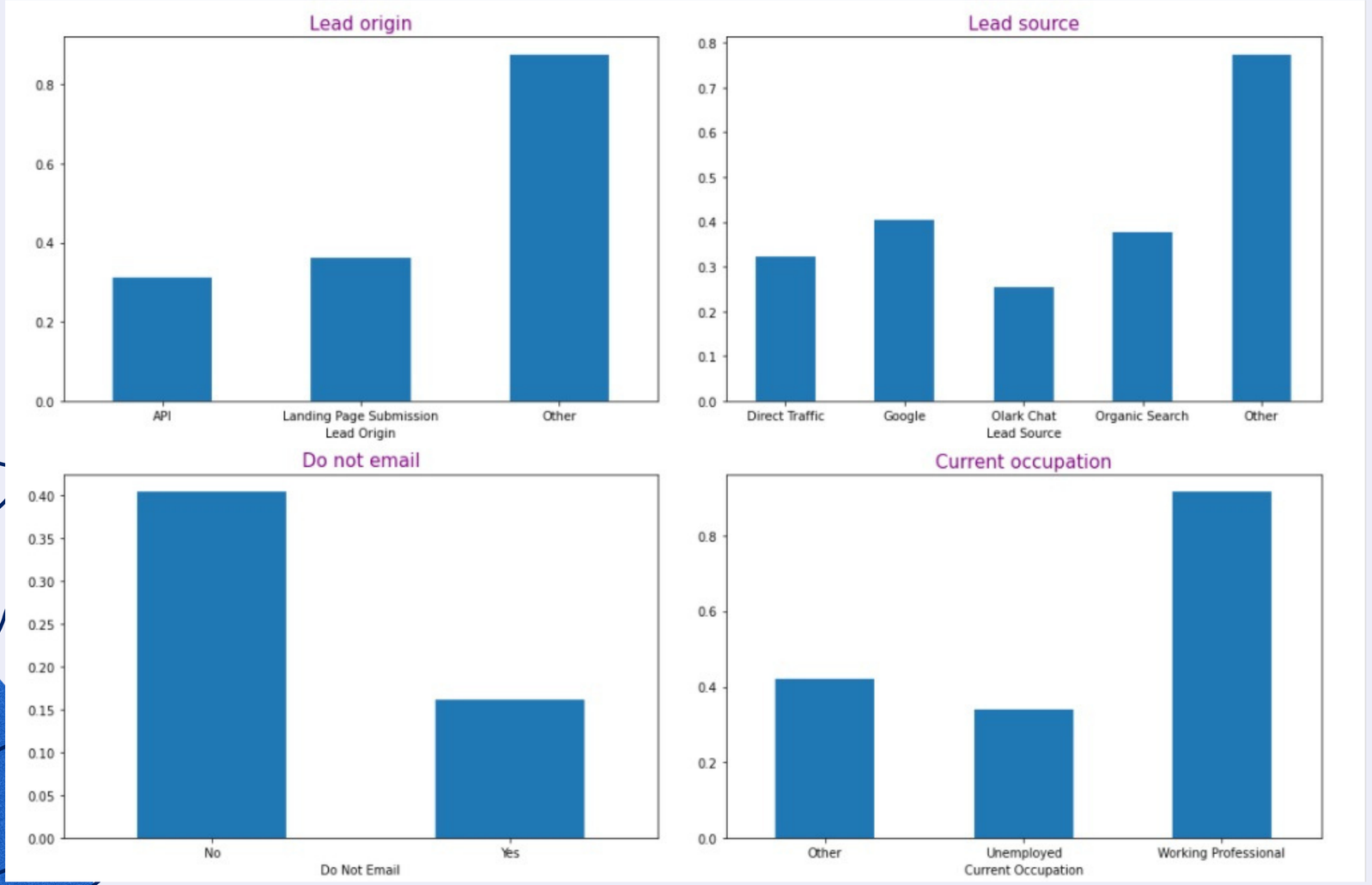
Visibility of outliers in Total Visits and Page Views per visit.

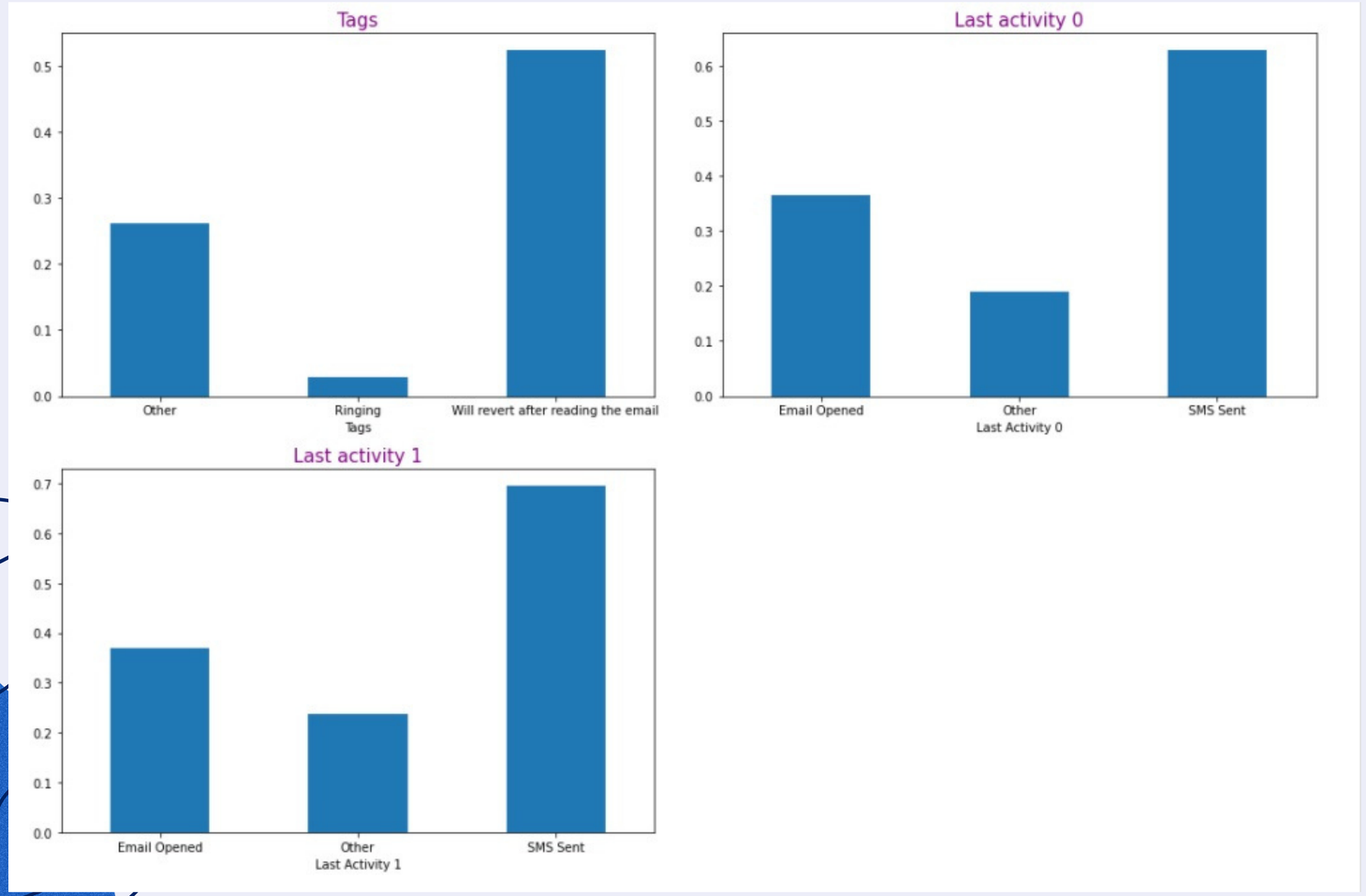
	Lead Origin	Lead Source	Do Not Email	Converted	TotalVisits	Time Spent	Page Views Per Visit
1160	Landing Page Submission	Direct Traffic	No	0	43.0	57	3.91
2190	Landing Page Submission	Direct Traffic	Yes	0	55.0	297	55.00
2322	Landing Page Submission	Direct Traffic	Yes	0	141.0	755	6.71
5283	Landing Page Submission	Direct Traffic	Yes	0	74.0	1209	12.33
5530	Landing Page Submission	Direct Traffic	Yes	0	41.0	311	1.14
5538	Landing Page Submission	Other	No	0	54.0	1280	4.50
5608	Landing Page Submission	Other	Yes	0	115.0	187	8.21
6102	Landing Page Submission	Direct Traffic	No	1	251.0	49	1.48
6580	API	Google	No	0	32.0	26	1.78
8230	Landing Page Submission	Direct Traffic	Yes	0	42.0	1148	3.82



- If Total Visits are more than 30 then most of times they are not converted.It doesn't look convenient that person visiting site for so many times.
- For that , outlier treatment is done by removing it.

BIVARIATE ANALYSIS

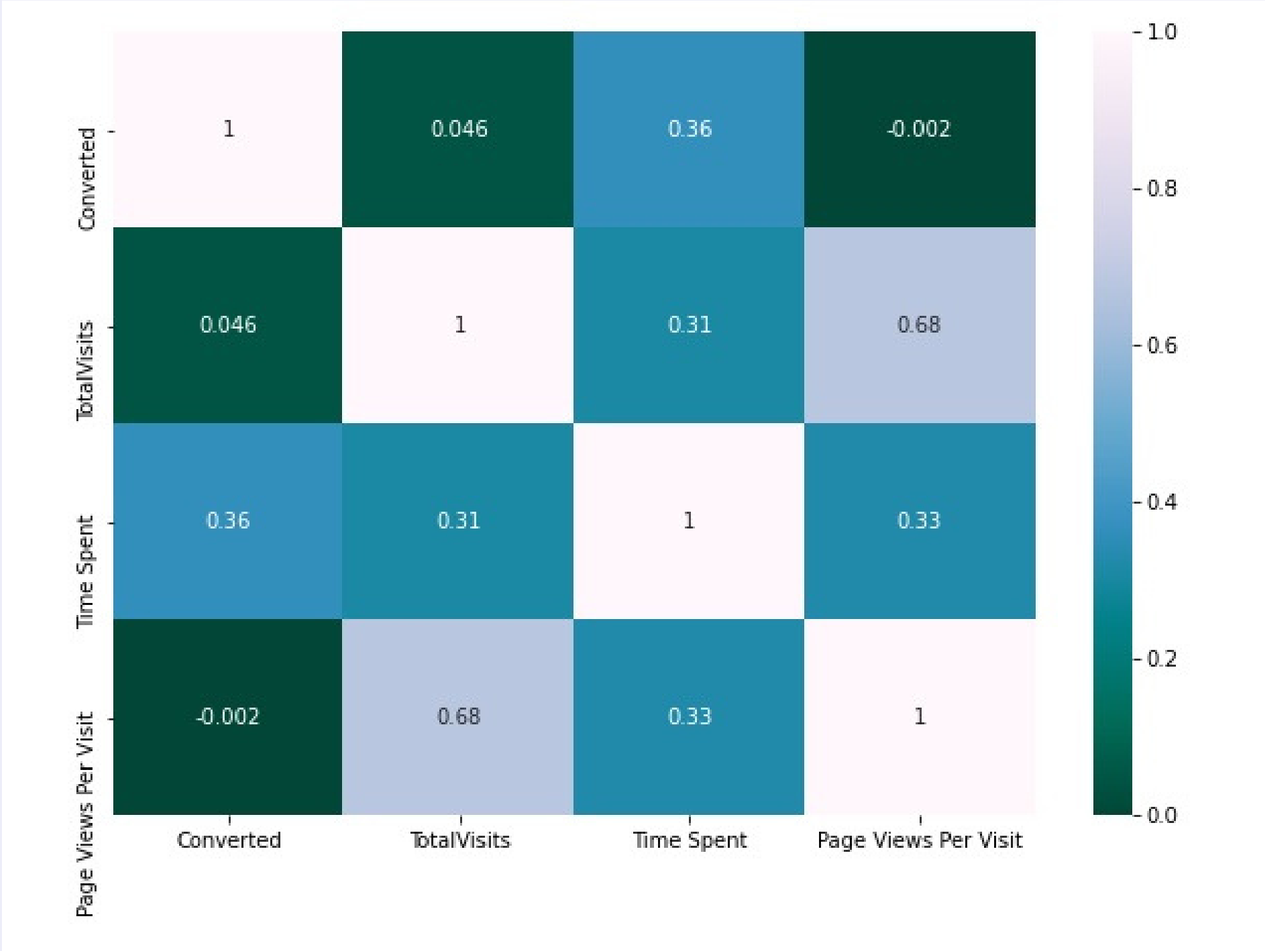


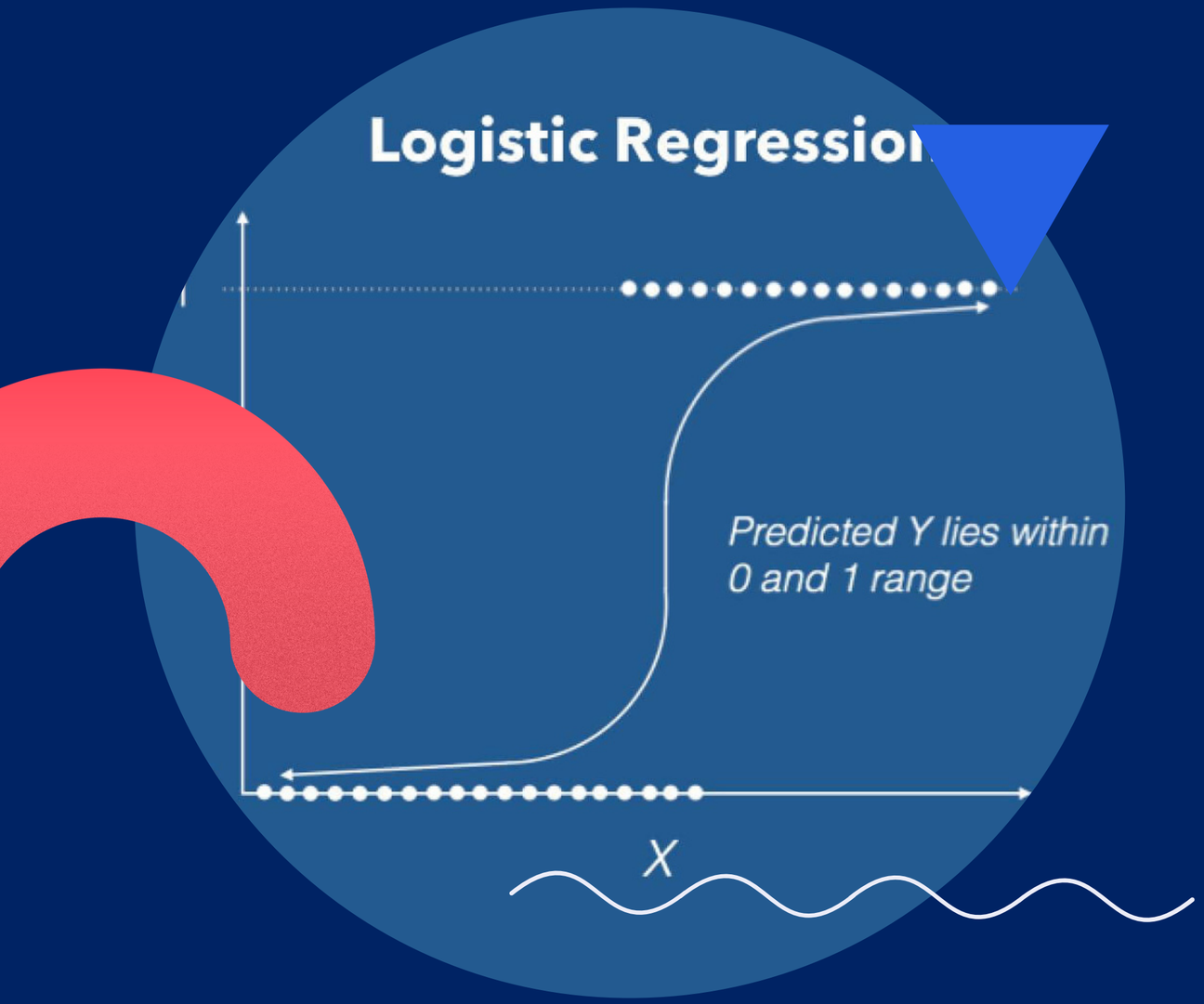

Tags
0.0
0.1
0.2
0.3
0.4
0.5
Other
Ringing
Will revert after reading the email
Tags
Last activity 0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Email Opened
Other
SMS Sent
Last Activity 0
Last activity 1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Email Opened
Other
SMS Sent
Last Activity 1

- "Converted" has been used as the target variable. It tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- In Last Activity 0 & 1, `SMS Sent` has a good Converted ratio.
- `Working Proffesional` has more interest towards course.

CORRELATION BETWEEN THE VARIABLES

Total Visits and Page Views Per Visit are positive correlates with each other.





BUILDING A MODEL

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6461
Model:	GLM	Df Residuals:	6450
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2317.2
Date:	Sun, 07 Mar 2021	Deviance:	4634.4
Time:	14:47:40	Pearson chi2:	9.20e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8758	0.126	-22.882	0.000	-3.122	-2.629
TotalVisits	2.9698	0.466	6.369	0.000	2.056	3.884
Time Spent	4.7681	0.183	26.014	0.000	4.409	5.127
Page Views Per Visit	-2.5564	0.657	-3.889	0.000	-3.845	-1.268
Lead Origin_Other	3.8656	0.186	20.809	0.000	3.501	4.230
Lead Source_Olark Chat	0.7773	0.126	6.179	0.000	0.531	1.024
Current Occupation_Working Professional	2.4780	0.188	13.206	0.000	2.110	2.846
Tags_Ringing	-3.3492	0.240	-13.928	0.000	-3.821	-2.878
Tags_Will revert after reading the email	1.1170	0.085	13.069	0.000	0.949	1.285
Last Activity 0_Other	-1.1074	0.090	-12.339	0.000	-1.283	-0.931
Last Activity 1_SMS Sent	1.8386	0.100	18.433	0.000	1.643	2.034

Final Model

	Features	VIF
2	Page Views Per Visit	4.52
0	TotalVisits	3.92
7	Tags_Will revert after reading the email	2.66
1	Time Spent	2.10
4	Lead Source_Olark Chat	1.61
8	Last Activity 0_Other	1.55
9	Last Activity 1_SMS Sent	1.55
6	Tags_Ringing	1.32
3	Lead Origin_Other	1.20
5	Current Occupation_Working Professional	1.16

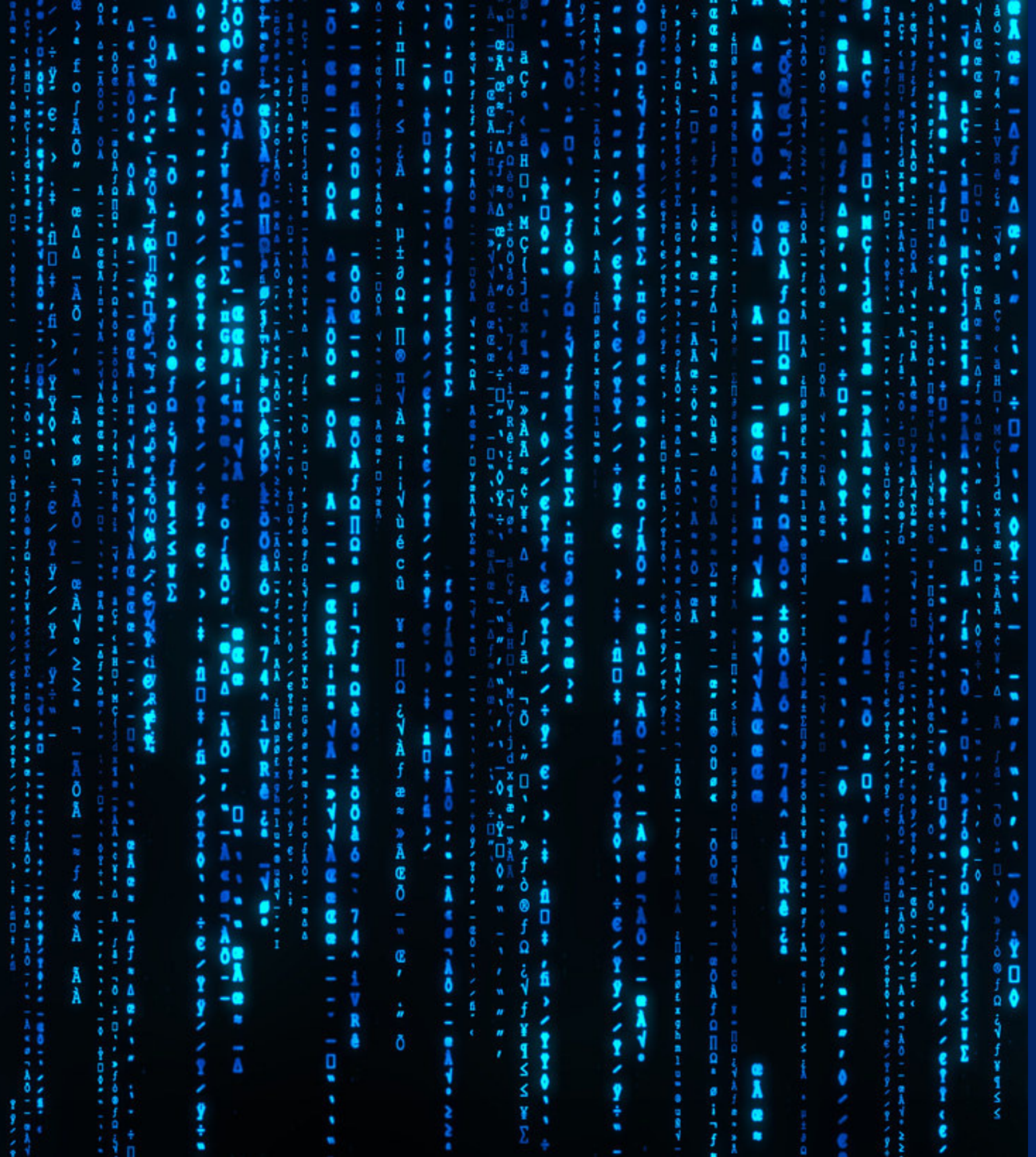
Final VIF check

- Model building along with feature selection (RFE) and VIF check for multicollinearity.

- After the removal of unnecessary variables, further modelling is conducted by taking 15 variables as an output. Irrelevant features may affect the model without getting the actual output in the end.

- Features with high P- Values (0.05+) and VIF (more than 5) were removed eventually by building the model again.

- Lastly, the model with 10 features and low P and VIF values are good to go for further predictions.

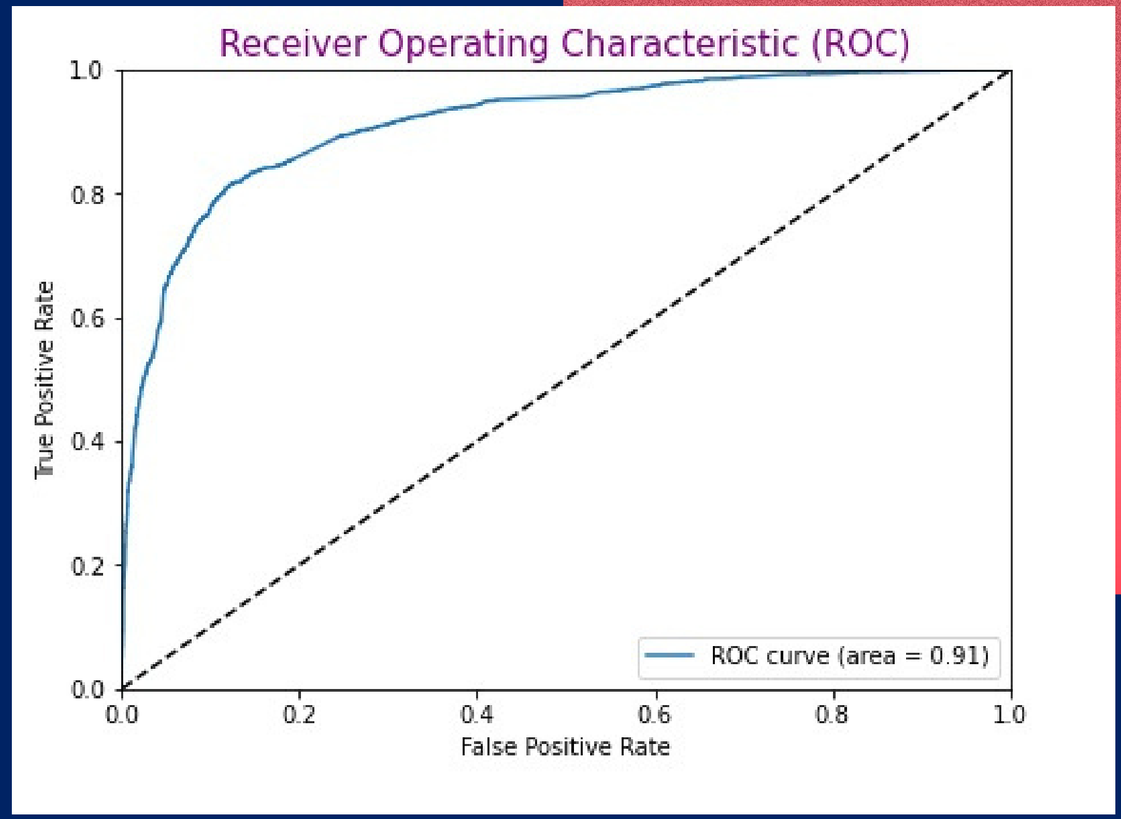


MATRIX SCORE TEST

ROC CURVE

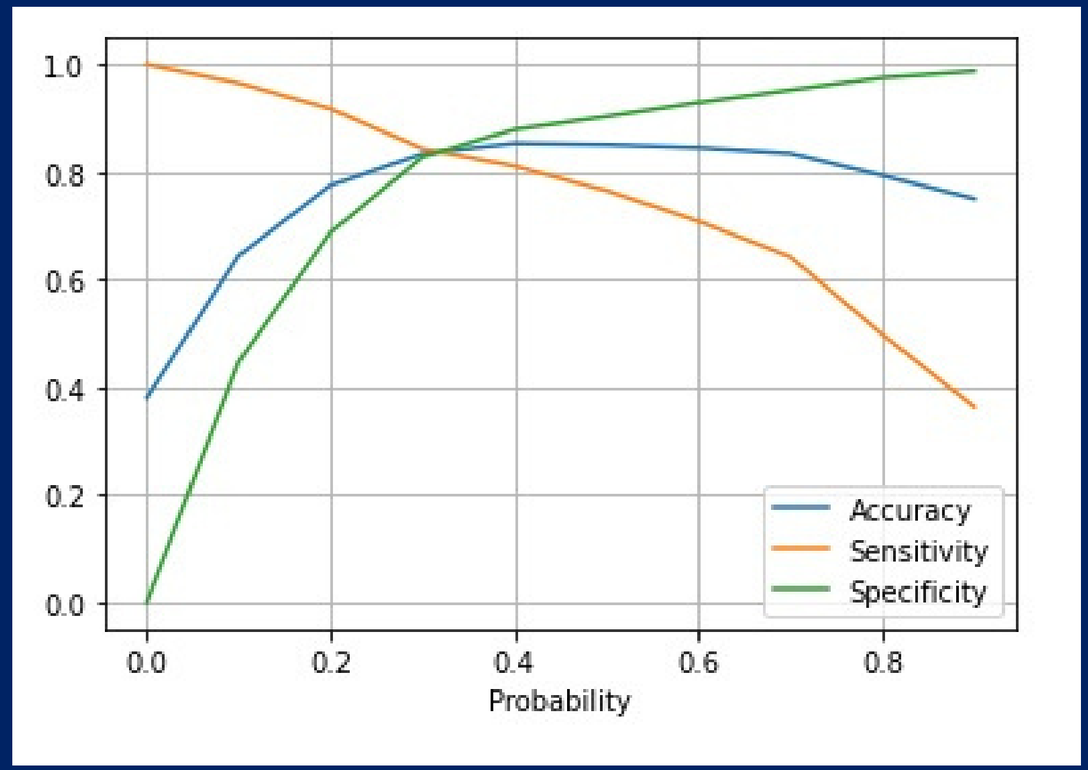
A ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- As per the graph, the ROC curve covered almost 91% of the total area.



OPTIMAL THRESHOLD

- The values of accuracy, sensitivity, and specificity were used for finding out the optimal threshold.
- From the curve, 0.3 is the optimum point to take as a cutoff probability.

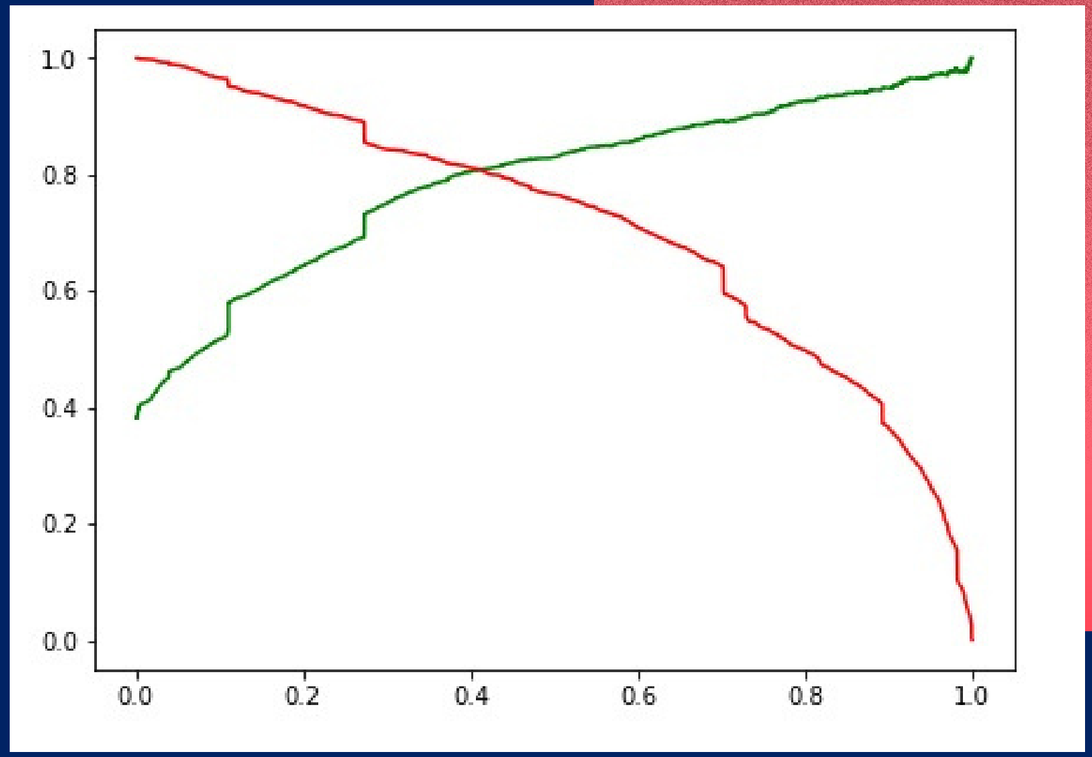


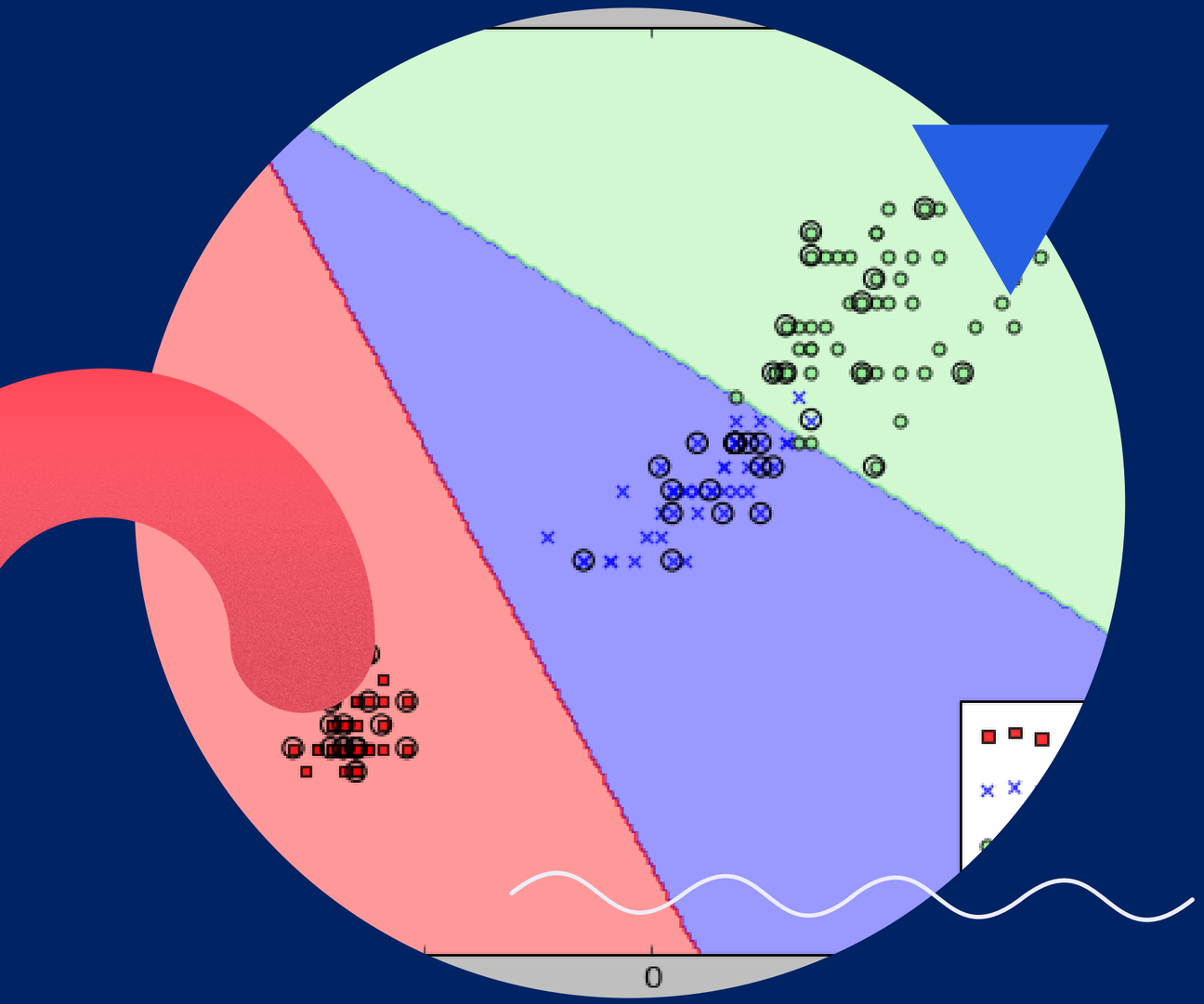
PRECISION - RECALL CURVE

- Precision is a metric that quantifies the number of correct positive predictions and recall is as same as sensitivity used for quantifying the positive predictions by finding true positive and false negative.

- Precision and recall are used for knowing the exact terminologies can be helpful as both of these pairs of metrics are often used in the industry.

- The right side graph shows a glimpse of the jumpy curve for precision. Both curves meet at 0.4.





PREDICTION ON TEST DATA

THESE ARE THE FOLLOWING VALUES AFTER DOING THE MODEL EVALUATION FOR TEST DATA

- FINAL PREDICTION :

	Lead Number	Converted	Converted_prob	Final_Predicted
0	600788	0	0.032204	0
1	600639	1	0.706023	1
2	644182	0	0.079234	0
3	599247	0	0.020375	0
4	596748	0	0.110185	0

- CONFUSION MATRIX:

PREDICTED → / ACTUAL ↓	NOT CONVERTED	CONVERTED
NOT CONVERTED	**1472 (TN)**	**195 (FP)**
CONVERTED	**218 (FN)**	**884 (TP)**

- Model Accuracy (Correctly predicted labels / Total no. of labels): 0.850
- Sensitivity (TP / TP + FN) : 0.802
- Specificity (TN / TN + FP) : 0.883

RECOMMENDATIONS

- Though the page per visit is high, the customer will be in a dilemma related to the course like comparing this course with others because of which conversion ratio is low.

- As per the business aspect, working professionals have probable chances of getting converted.

- Important features that contributed most towards the probability of a lead getting converted are :
    - Time Spent
    - Lead Origin_Other
    - TotalVisits
- Features that need to be focused more :
    - Last Activity 0_Other
    - Page Views Per Visit
    - Tags_Ringing

THANK YOU