




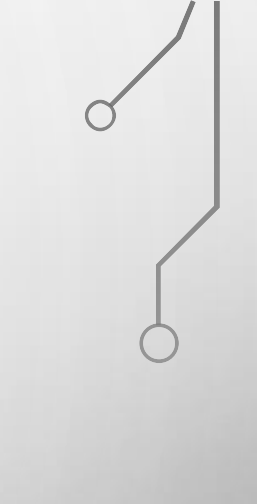
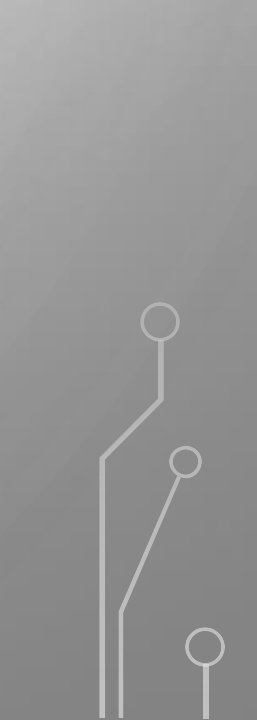
# **LEAD SCORING CASE STUDY**

BY

- ALLAUDDIN
- VISHAL
- SHWETA


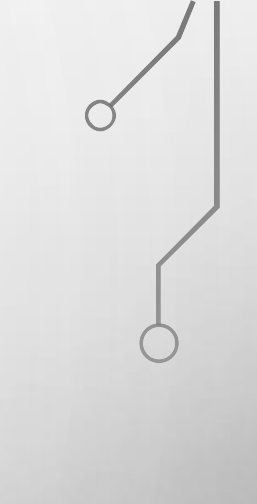
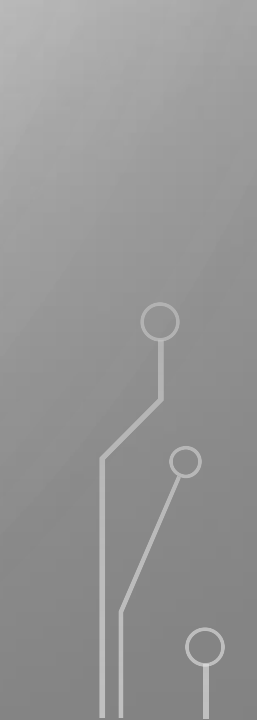


# PROBLEM STATEMENT

- A Company named X-Education sells online industry relevant courses. Many people land on their website through different platforms (e.g. Google Search, Reference etc.) and fill up a form and a lead is generated
  - The sales team then make calls to these leads to convert make a sale of the course they are interested in.
  - Current conversion rate of leads is very low approximately 30%
- 
- 
- 



# BUSINESS GOAL

- X-Education need us to build a model which will generate more promising leads i.e leads which have higher chance of conversion to provide them to their sales team
  - We have to perform exploratory data analysis to identify the features which are important for the model
  - Build a logistic regression model to identify the leads which have more probability of conversion. At the same time assigning lead score to each lead such that higher the lead score higher the chance of conversion
- 
- 
- 

# STEP BY STEP APPROACH

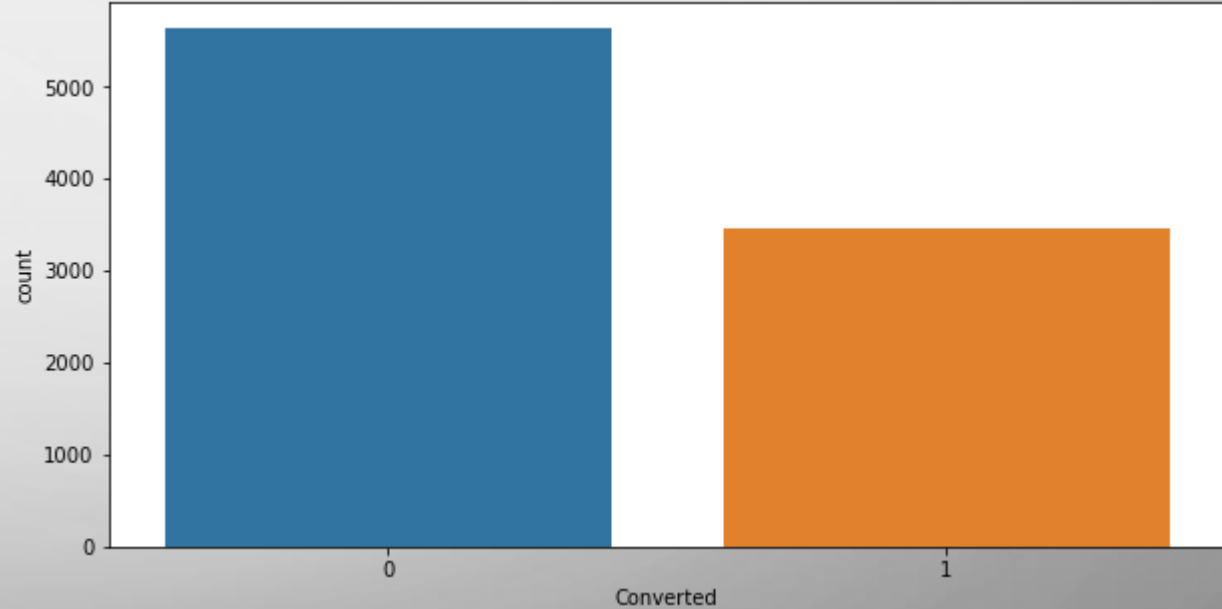
- Reading and Understanding the Data
- Cleaning the Data
  - Missing Value treatment
  - Outlier Value treatment
- Exploratory Data analysis
  - Categorical Variables Analysis
  - Numerical Variables Analysis
- Preparing the Data for Modelling
  - Creating Dummy Variables (Categorical Variables)
  - Feature Scaling (Numerical Variables)
  - Splitting the Data into Train-Test Data sets
- Creating Logistic Regression model
- Evaluating the Model on the Train Data Set
- Evaluating the Model on the Test Data Set



Calculating Accuracy, Sensitivity, Specificity on both train and test data sets to compare and evaluate the model

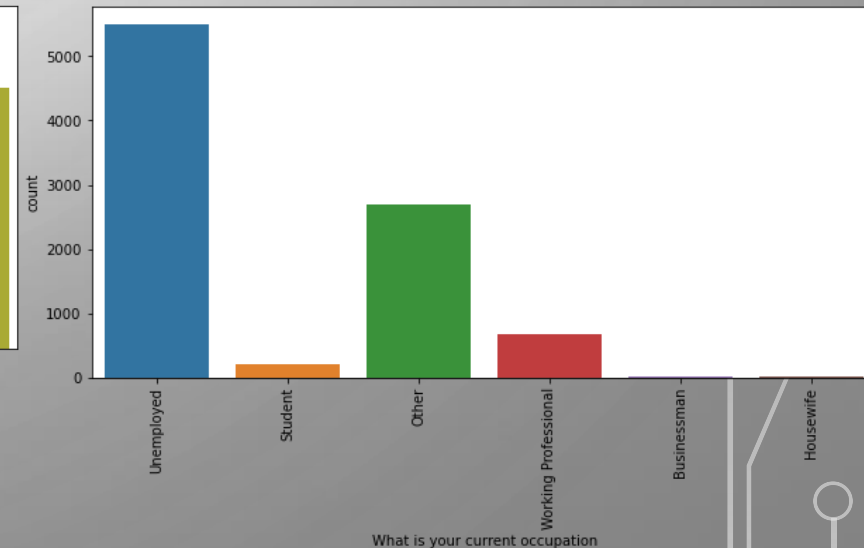
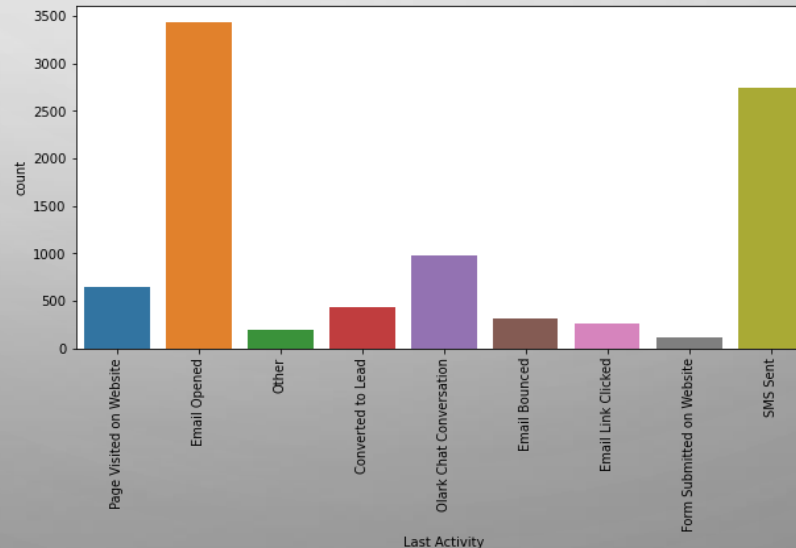
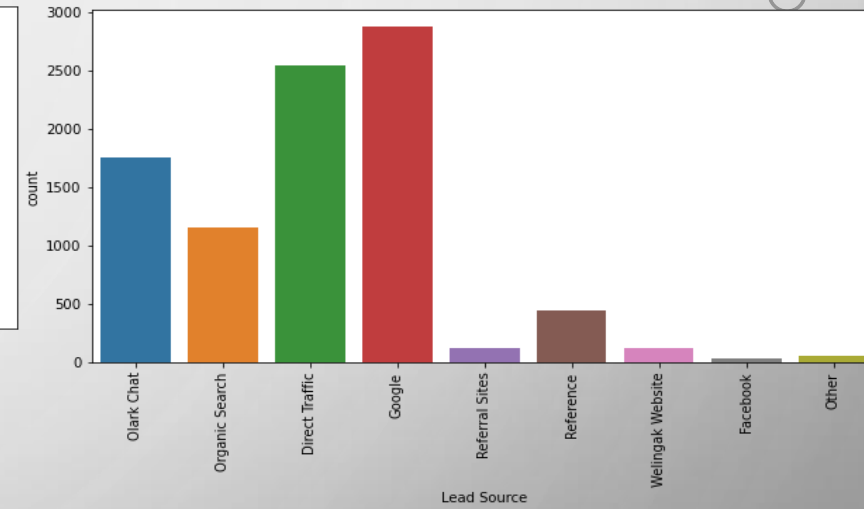
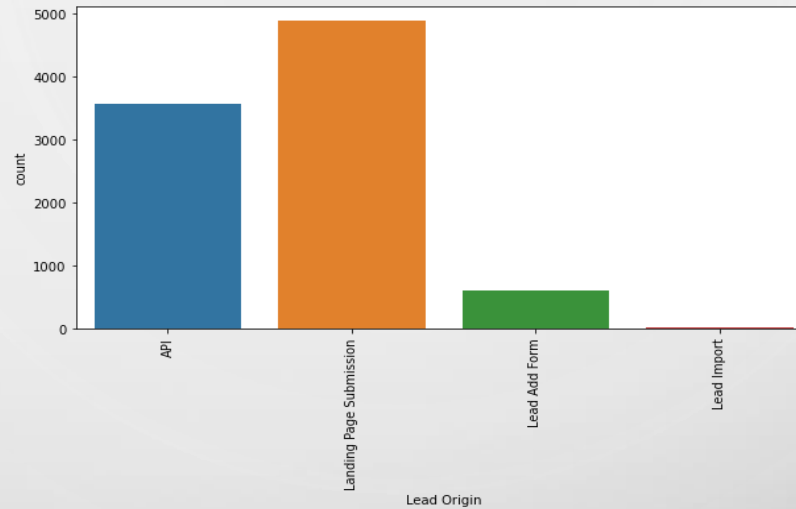
# EXPLORATORY DATA ANALYSIS

- Checking the conversion rate of the leads at the current time we can see there is huge data imbalance the current conversion rate is approximately 38%.



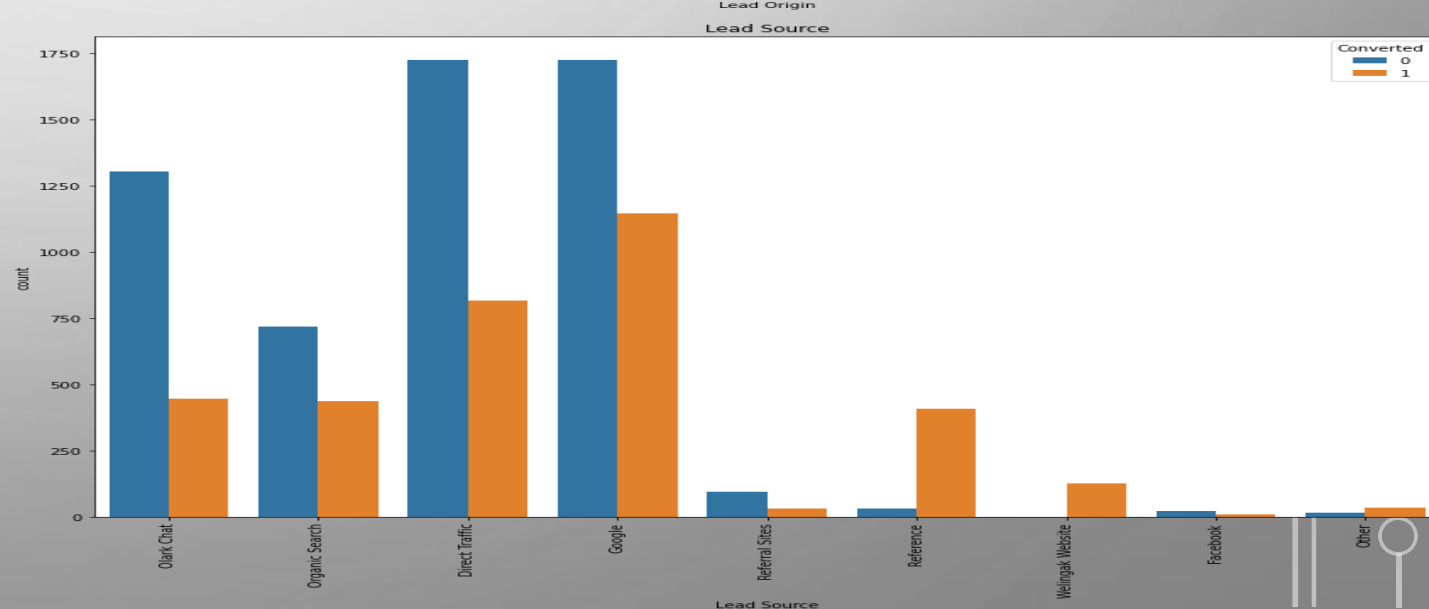
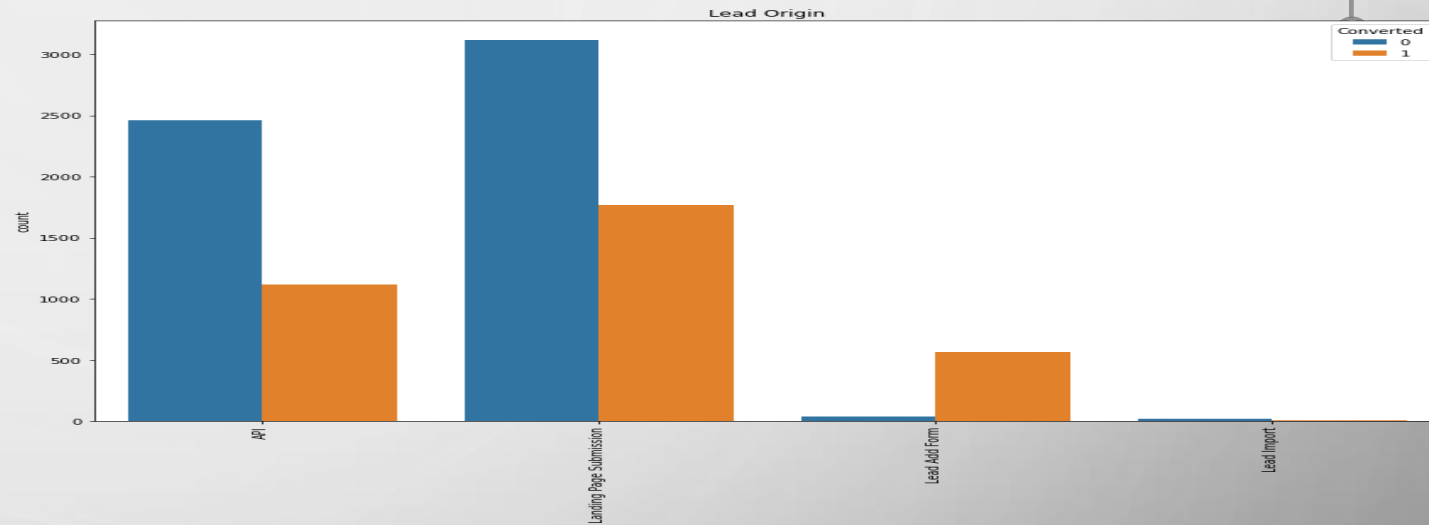
# CATEGORICAL VARIABLE ANALYSIS (UNIVARIATE)

- Having a look at the count of the variables we can see that most of the leads are originated from landing page submission.
- Google, Direct Traffic, Olark Chat are the main sources for our leads.
- The last activity of the most of the leads is seen as Email Opened and SMS sent.
- Most the leads are Unemployed who are looking for industry relevant courses.



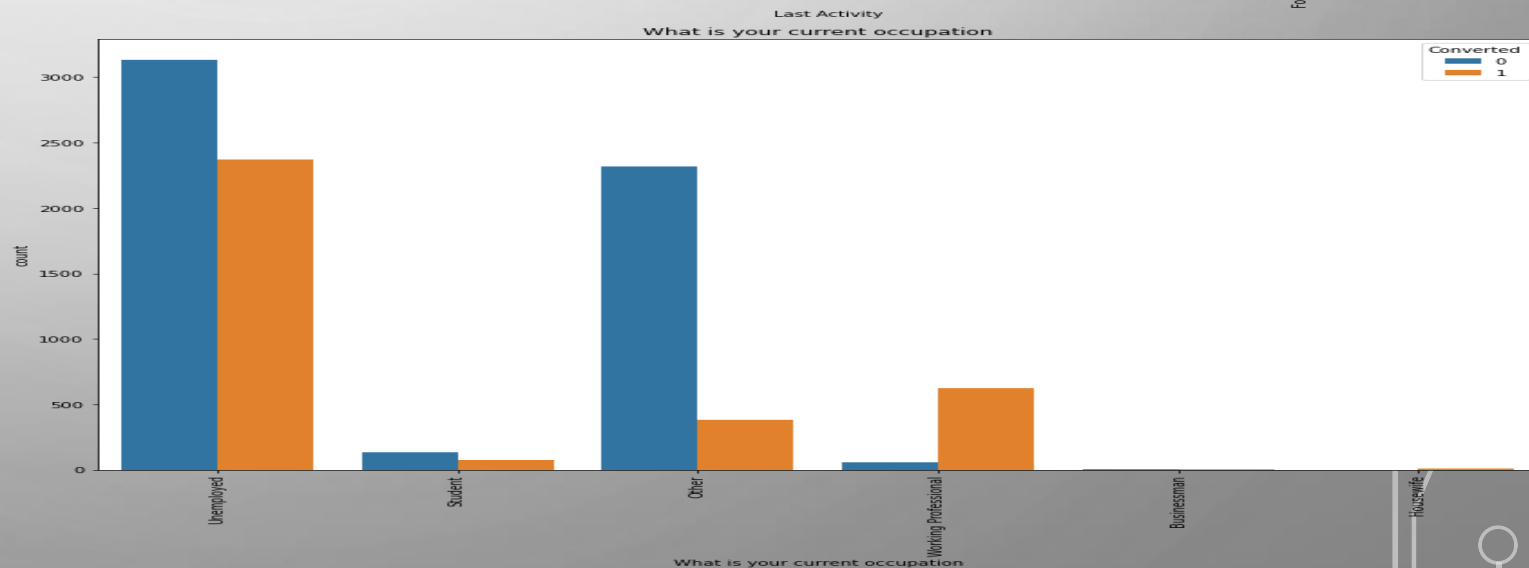
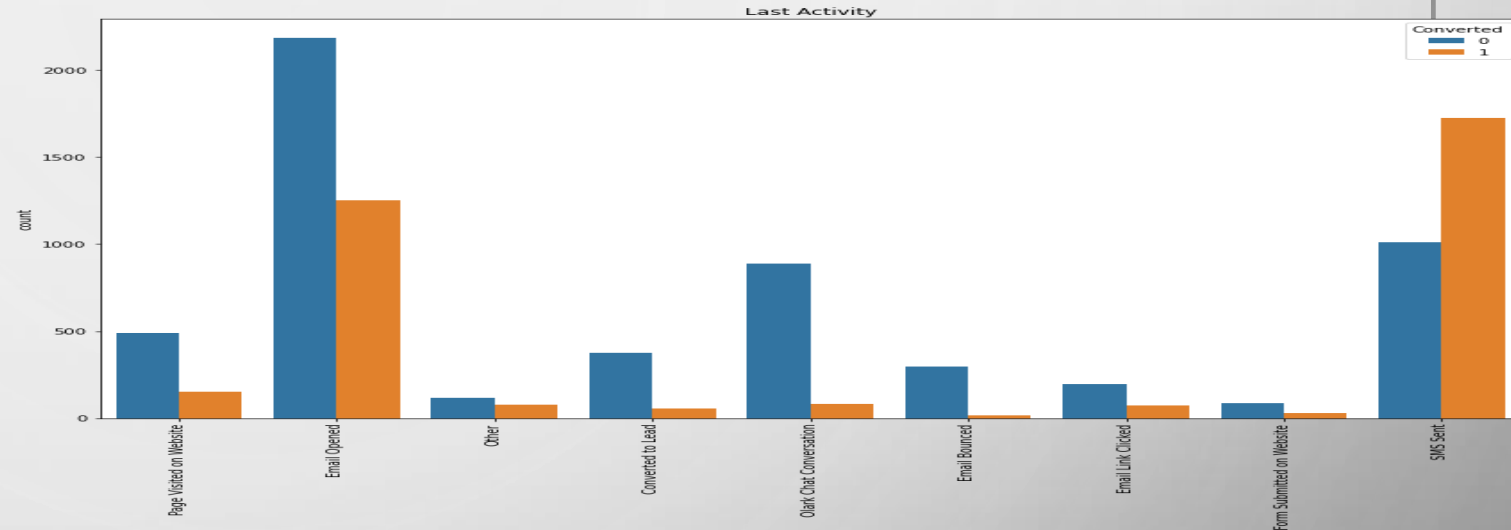
# CATEGORICAL VARIABLE ANALYSIS (BIVARIATE)

- By comparing the lead origin with conversion rate we can see that most number of lead converted are originated from landing submission page. Although lead add form has generated a very few leads we can see that the conversion rate for the seem is comparatively high.
- By Comparing lead source with conversion rate we can see that the highest converted leads are generated from Google, Direct Traffic. Here too we can see that References though have low count of lead sourcing has a high conversion rate.



# CATEGORICAL VARIABLE ANALYSIS (BIVARIATE)

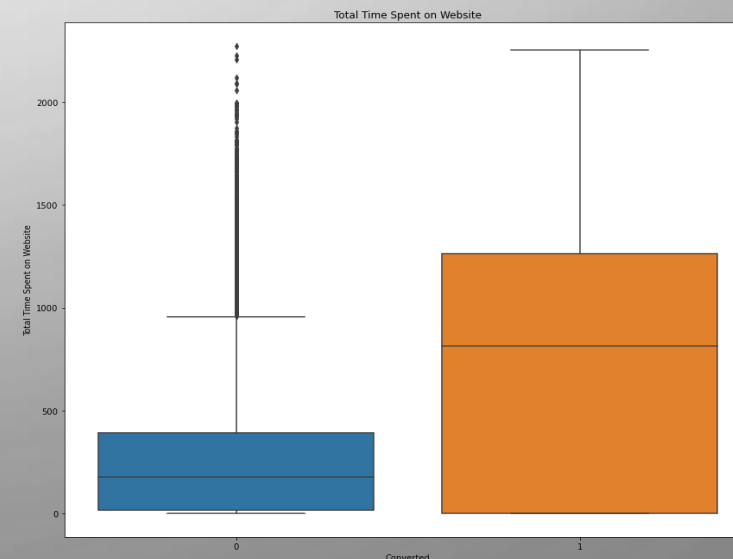
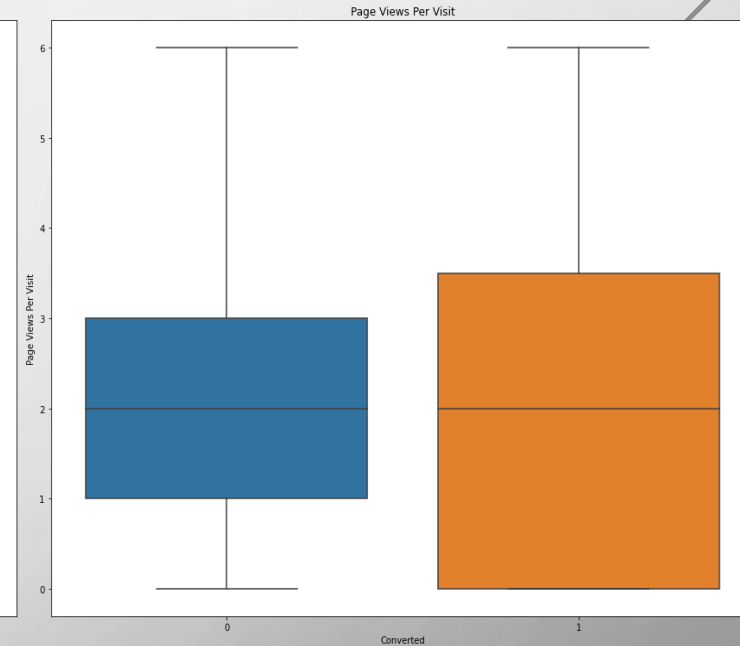
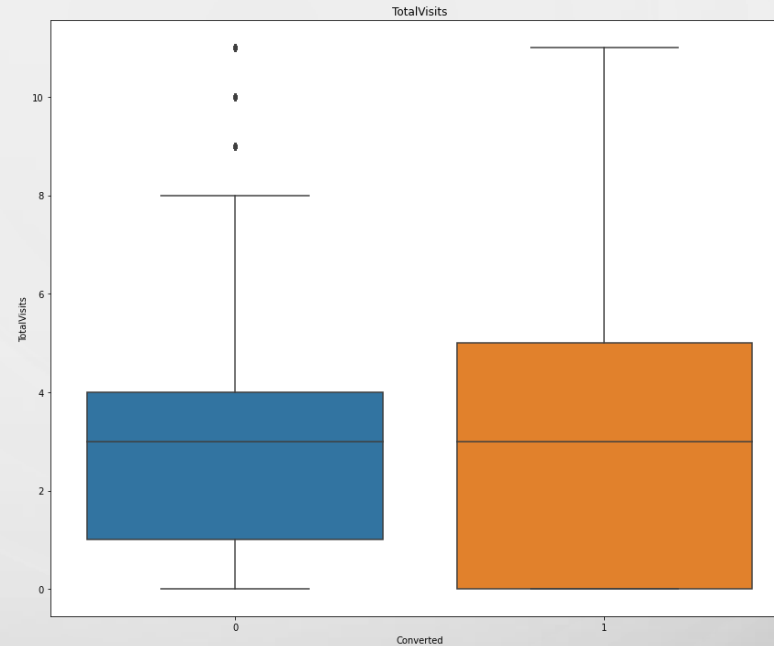
- Comparing Last Activity and conversion rate we can clearly see that leads which have their last activity as Email Opened and SMS sent has the most number of converted leads.
- Finally, upon looking Current Occupation and Conversion rate together we can see majority of the leads converted are unemployed (since this are industry relevant courses people looking for Jobs). Another observation we can see that working professional though are small in count have high conversion rate.





# NUMERICAL VARIABLES ANALYSIS

- Here upon comparing the numerical variables with conversion we can see that total visits and page view per visits have there median approximately same for converted and not converted leads hence we are not able to get much insights.
- However, looking as total time spent on website we can clearly see that leads who have spent more time on the website are the ones who got converted.



# MODEL BUILDING AND EVALUATION

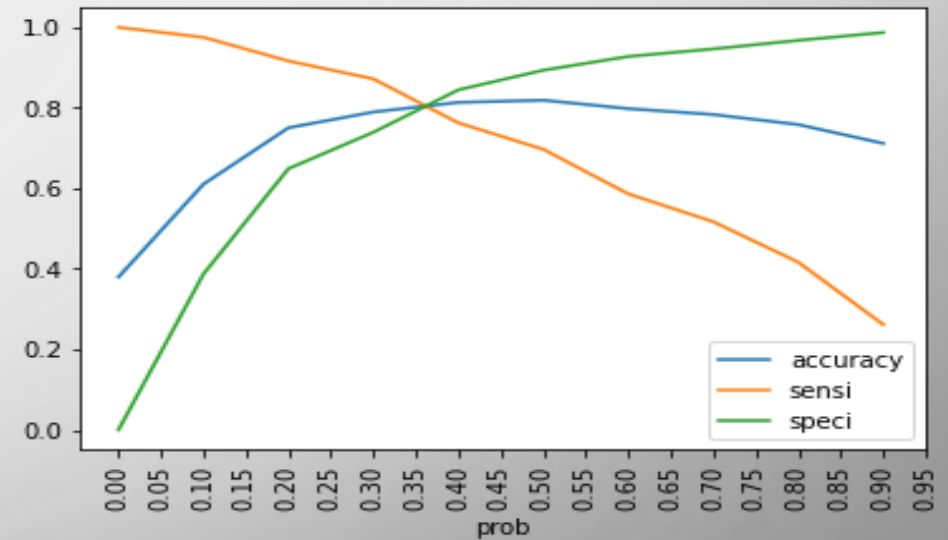
- From the final model we can clearly see that main features impacting the conversion rate are the following
  - Do Not Email
  - Total Time Spent on Website
  - Lead Origin\_Lead Add Form
  - Lead Origin\_Lead Import
  - Lead Source\_Olark Chat
  - Lead Source\_Wellignak Website
  - Last Activity\_Email Link Clicked
  - Last Activity\_Email Opened
  - Last Activity\_Other
  - Last Activity\_Page Visited on Website
  - Last Activity\_SMS Sent
  - What is your current occupation\_Other
  - What is your current occupation\_Working Professional

	coef	std err	z	P> z	[0.025	0.975]
const	-2.2777	0.118	-19.289	0.000	-2.509	-2.046
Do Not Email	-1.2321	0.175	-7.050	0.000	-1.575	-0.890
Total Time Spent on Website	1.1164	0.040	27.570	0.000	1.037	1.196
Lead Origin_Lead Add Form	3.6582	0.223	16.409	0.000	3.221	4.095
Lead Origin_Lead Import	0.9391	0.450	2.085	0.037	0.056	1.822
Lead Source_Olark Chat	1.2609	0.104	12.107	0.000	1.057	1.465
Lead Source_Welingak Website	1.8629	0.760	2.451	0.014	0.373	3.353
Last Activity_Email Link Clicked	0.7517	0.240	3.137	0.002	0.282	1.221
Last Activity_Email Opened	1.3077	0.122	10.739	0.000	1.069	1.546
Last Activity_Other	1.8416	0.237	7.773	0.000	1.377	2.306
Last Activity_Page Visited on Website	0.8408	0.178	4.724	0.000	0.492	1.190
Last Activity_SMS Sent	2.4052	0.124	19.332	0.000	2.161	2.649
What is your current occupation_Other	-1.2082	0.087	-13.889	0.000	-1.379	-1.038
What is your current occupation_Working Professional	2.5505	0.188	13.563	0.000	2.182	2.919

# MODEL EVALUATION ACCURACY, SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET

- Finding the optimal cut-off point as 0.35 by plotting the Accuracy, Sensitivity and Specificity
- Using 0.35 as probability we get the confusion matrix as shown in the table
- The following Scores are calculated for the above probability on train data set

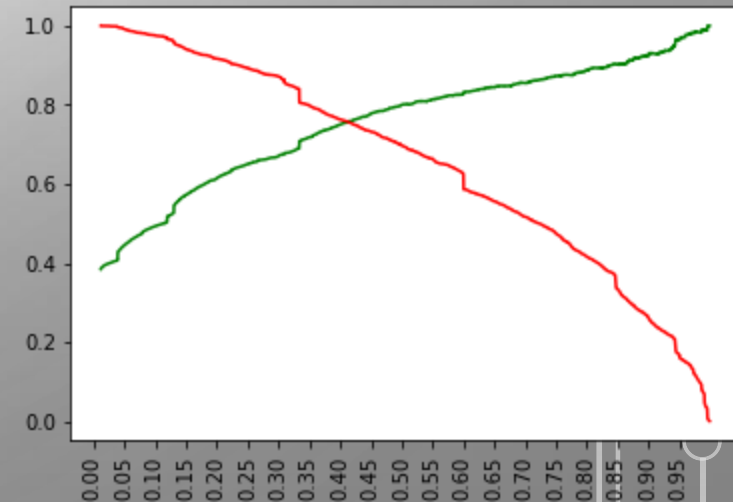
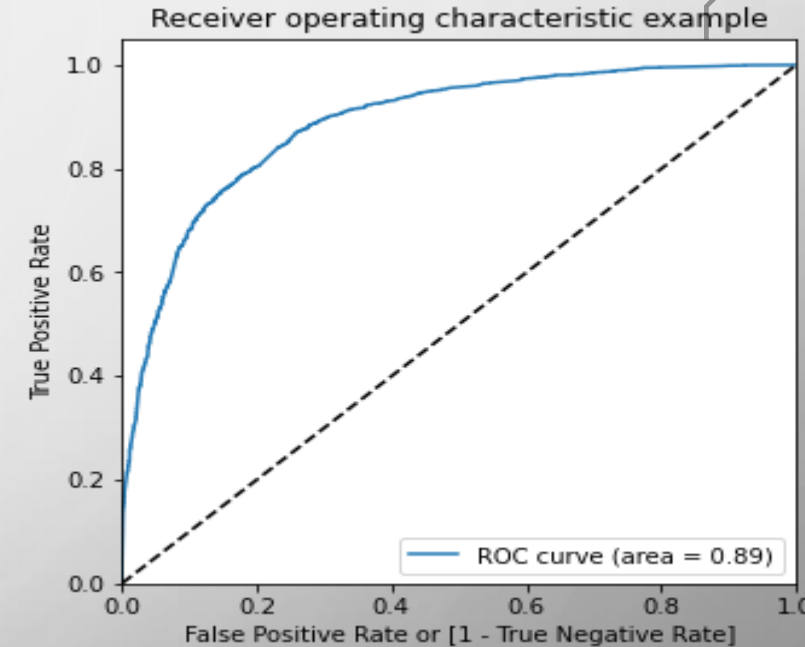
- Accuracy : 80.5%
- Sensitivity : 79.8%
- Specificity : 80.8%
- false positive rate : 19.1%
- positive predictive value : 71.8%
- Negative predictive value : 86.8%



CONFUSION MATRIX	
3196	757
488	1931

# ROC CURVE, PRECISION AND RECALL

- Here we can see that Area under the ROC curve 0.89 which is really good for our logistic regression model meaning our model performance is better in distinguishing positive and negative classes.
- The Precision and Recall Scores calculated at 0.35 Probability are as follows
  - Precision Score : 71.8%
  - Recall Score : 79.8%
- When plotting the precision and recall curve we see the optimum probability is approximately 0.4.



## MODEL EVALUATION ACCURACY, SENSITIVITY AND SPECIFICITY ON TEST DATA SET

- The confusion matrix for the test data set after predicting values at 0.35 probability can be seen in the table.
- Evaluating the model on test data set and predicting values at 0.35 probability the following scores are obtained
  - Accuracy : 80.4%
  - Sensitivity : 79.3%
  - Specificity : 81.2%

CONFUSION MATRIX	
1372	317
216	826

# CONCLUSION

- We calculated optimal cut-off points by plotting accuracy, sensitivity and specificity curve using multiple probabilities at  $\sim 0.35$  and by plotting precision and recall curve at  $\sim 0.4$ . However, we considered 0.35 as optimal cut-off point.
- The Accuracy, Sensitivity and Specificity scores for Train Data set are 80.5%, 79.8% and 80.8% respectively.
- The Accuracy, Sensitivity and Specificity scores for Test Data set are 80.4%, 79.3% and 81.2% respectively.
- The predictions made on the Train and Test Data Sets are approximately same.
- Hence we can say that our model is good at distinguishing positive and negative classes.
- Finally, most important features effecting our model are
  - Lead Origin\_Lead Add Form, What is your current occupation\_Working Professional, Last Activity\_SMS Sent
- Followed by these Features
  - Lead Source\_Welingak Website, Last Activity\_Other, Last Activity\_Email Opened

The background features a series of concentric circles in a light gray tone, centered on the page. In the four corners, there are decorative line art elements resembling circuit boards or neural networks, with thin black lines and small circles.

**THANK YOU..!**