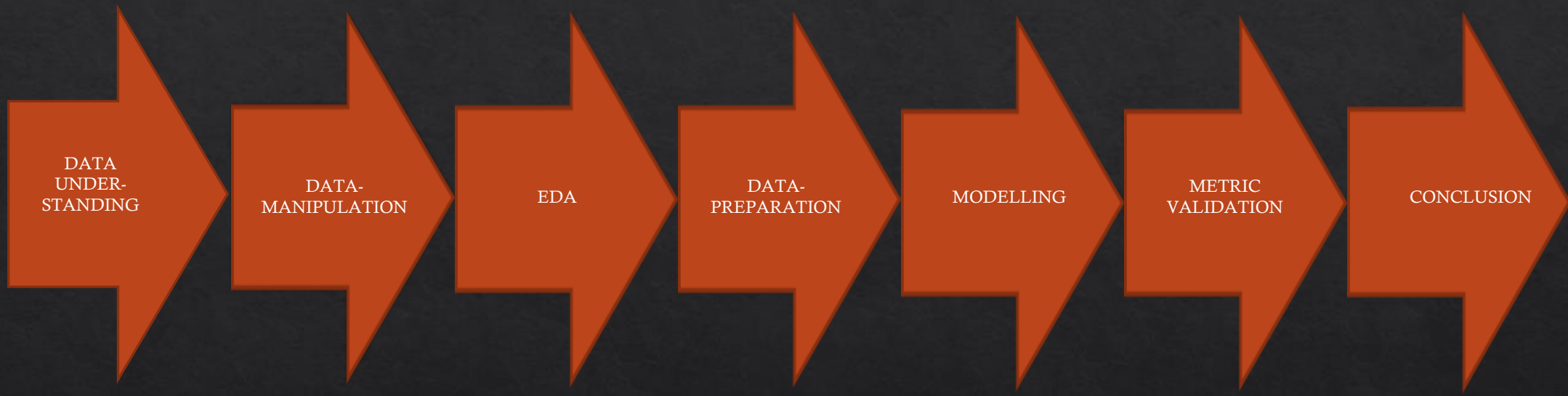
LEAD SCORING CASE STUDY

BY:

BHARATH MD

&

RISHABH SHUKLA

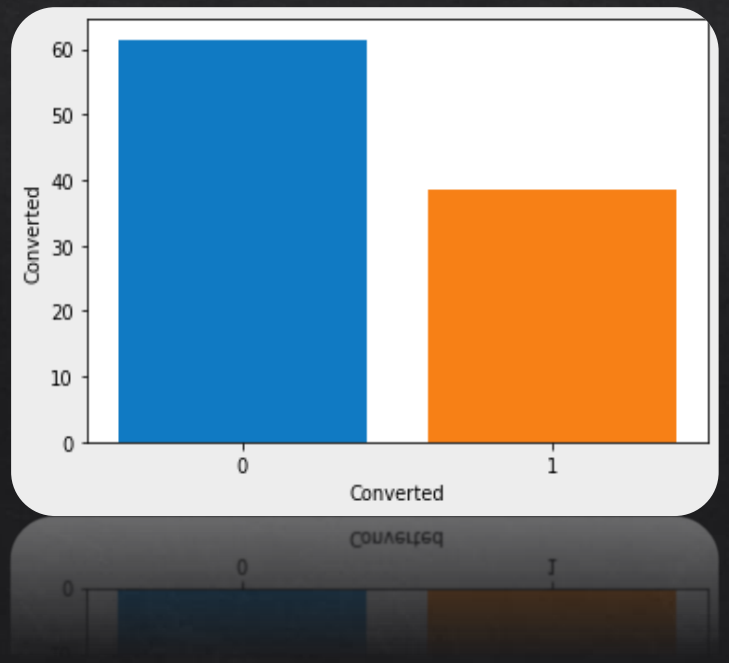

DATA UNDER-STANDING
DATA-MANIPULATION
EDA
DATA-PREPARATION
MODELLING
METRIC VALIDATION
CONCLUSION

PROBLEM STATEMENT

- TO HELP X EDUCATION SELECT THEIR MOST PROMISING LEADS THAT ARE MOST LIKELY TO CONVERT TO PAID CUSTOMERS.
- BUILD A LOGISTIC REGRESSION MODEL TO ASSIGN A LEAD SCORE BETWEEN 0 AND 100 TO EACH OF THE LEADS WHICH CAN BE USED BY THE COMPANY TO TARGET POTENTIAL LEADS. A HIGHER SCORE WOULD MEAN THAT THE LEAD IS HOT, I.E. IS MOST LIKELY TO CONVERT WHEREAS A LOWER SCORE WOULD MEAN THAT THE LEAD IS COLD AND WILL MOSTLY NOT GET CONVERTED.
- MODEL SHOULD BE ABLE TO ADJUST TO IF THE COMPANY'S REQUIREMENT CHANGES IN THE FUTURE SO YOU WILL NEED TO HANDLE THESE AS WELL.
- SUMMARIZE THE PREDICTIONS IN THE END AND HIGHLIGHT FACTORS AFFECTING THE PROCESS.

DATA EXPLORATION

- FILE NAME : LEADS.CSV
  - THE FILE CONTAINS 9240 COLUMNS AND 37 ROWS
  - 7 NUMERIC COLUMNS AND 30 CATEGORICAL COLUMNS DESCRIBE THIS DATA.
  - FROM THE GRAPH PLOTTED WE SEE THAT THE CURRENT CONVERSION RATE IS ~39%



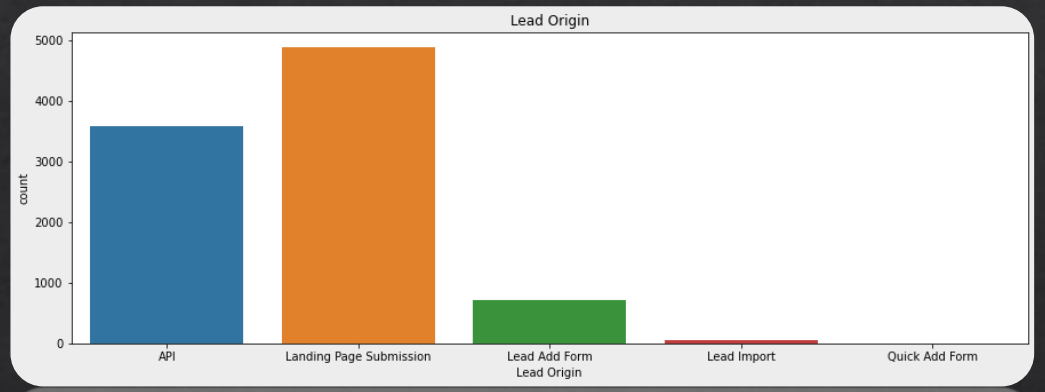
DATA PREPARATION

- SELECTING ONLY IMPORTANT COLUMNS TO ANALYSE THAT CONTRIBUTES TO THE MODEL; since these variables have only 2 categories within themselves with a very high margin difference we can drop them.
- COLUMNS HAVING ‘SELECT A’S A CATEGORY WITHIN IS REPLACED WITH ‘NAN’
- AFTER CHECKING FOR NULL VALUES , THE COLUMNS WERE DROPPED WHOSE NULL VALUES WERE MORE THAN 45% AND STORED IN A SEPARATE DATAFRAME ; now we have 30 columns to work with
- REPLACEMENT OF CERTAIN MISSING VALUES WITH MODE FOR:
  - lead source , last activity.
- REPLACEMENT OF OTHER MISSING VALUES TO CATEGORY ‘OTHERS’ FOR REST OF THE COLUMNS.
- REMOVE PROSPECT ID AND LEAD NUMBER AS THEY HAVE UNIQUE VALUES.

CATEGORICAL ANALYSIS

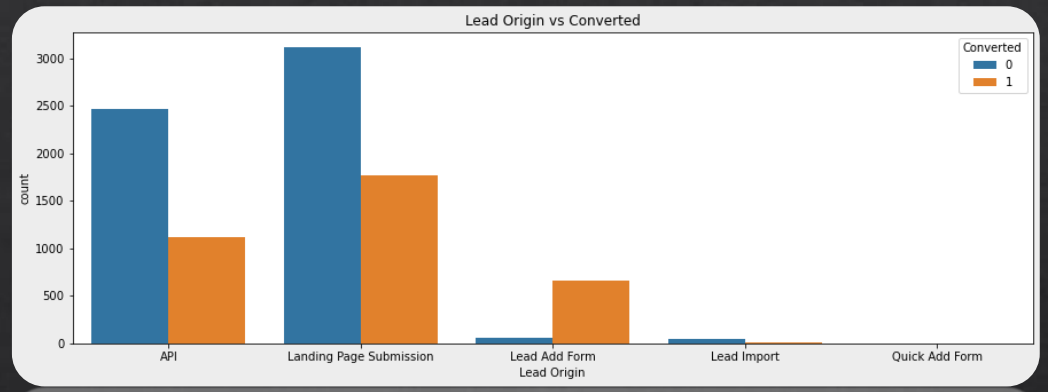
UNIVARIATE AND BIVARIATE

LEAD ORIGIN



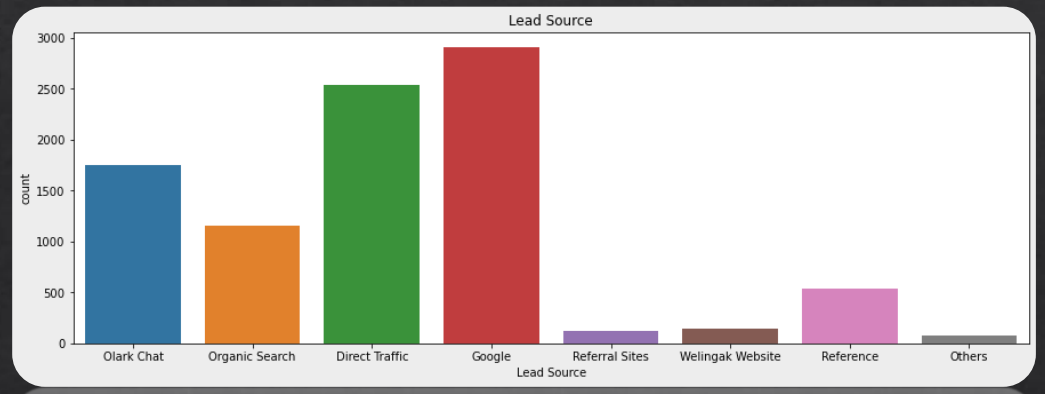
1. ~53% OF LEAD IS FROM LANDING PAGE SUBMISSION FOLLOWED BY
2. API

VS CONVERTED



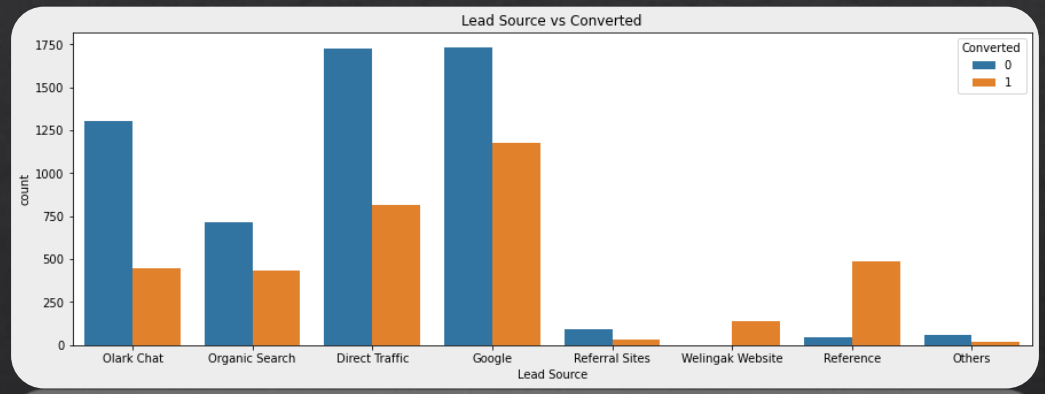
1. ~36% CONVERSION IS SEEN FOR LANDING PAGE SUBMISSION
2. LEAD ADD FORM SHOWS MORE CONVERSION RATE WHICH IS GOOD TO IMPROVE BUSINESS

LEAD SOURCE



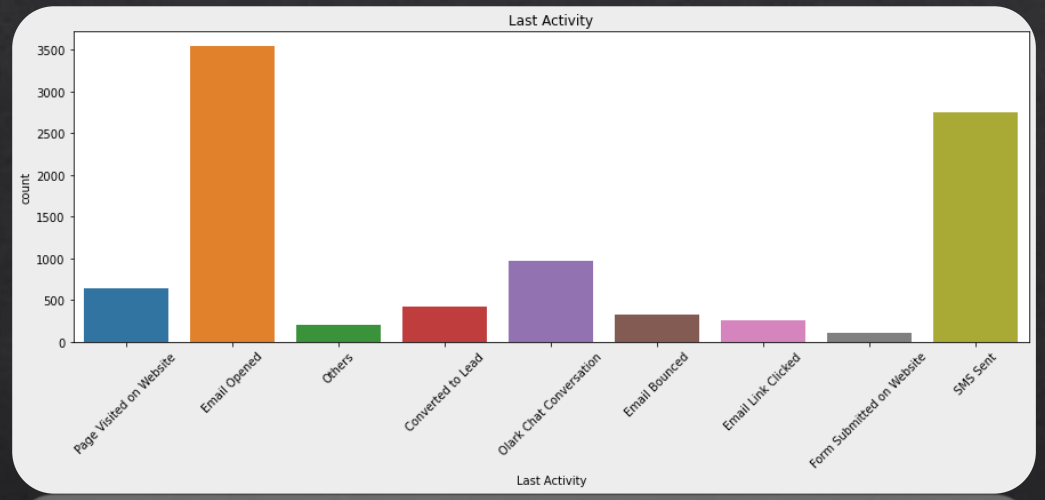
1. ~31% LEAD SOURCE IS FROM GOOGLE FOLLOWED BY
2. DIRECT TRAFFIC

VS CONVERTED



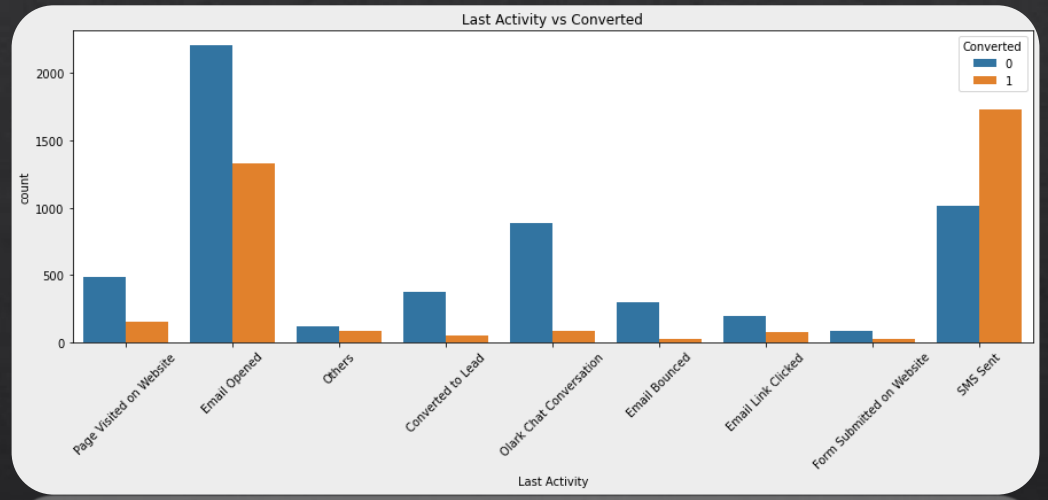
1. ~40% CONVERSION RATE FOR GOOGLE .
2. ~90% CONVERSION RATE FOR REFERENCES WHICH IS GOOD TO IMPROVE BUSINESS.

LAST ACTIVITY



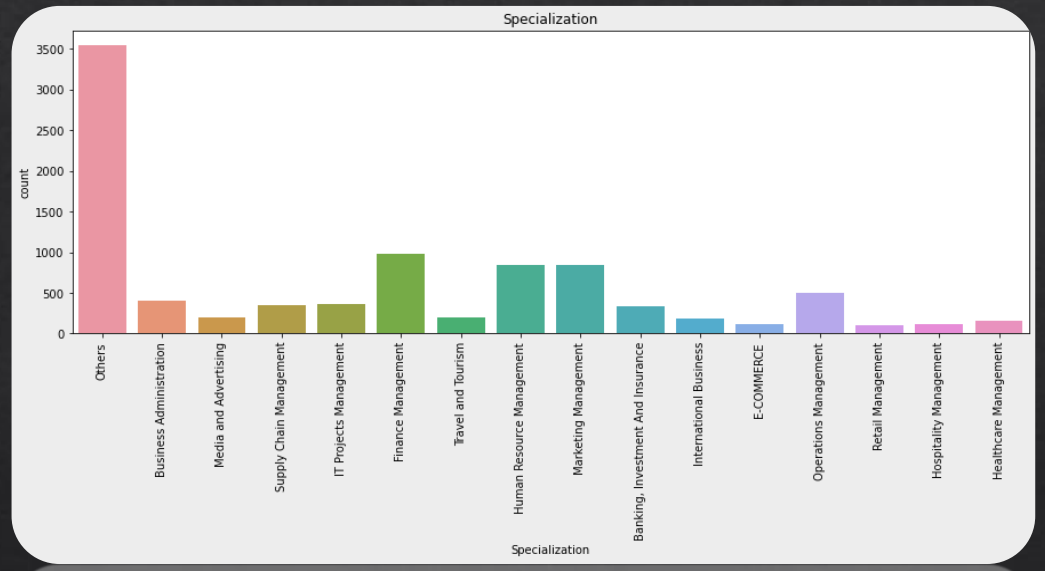
1. ~38% OF THE LAST ACTIVITY SEEN IS EMAIL-OPENED FOLLOWED BY
2. ~30 SMS SENT

VS CONVERTED



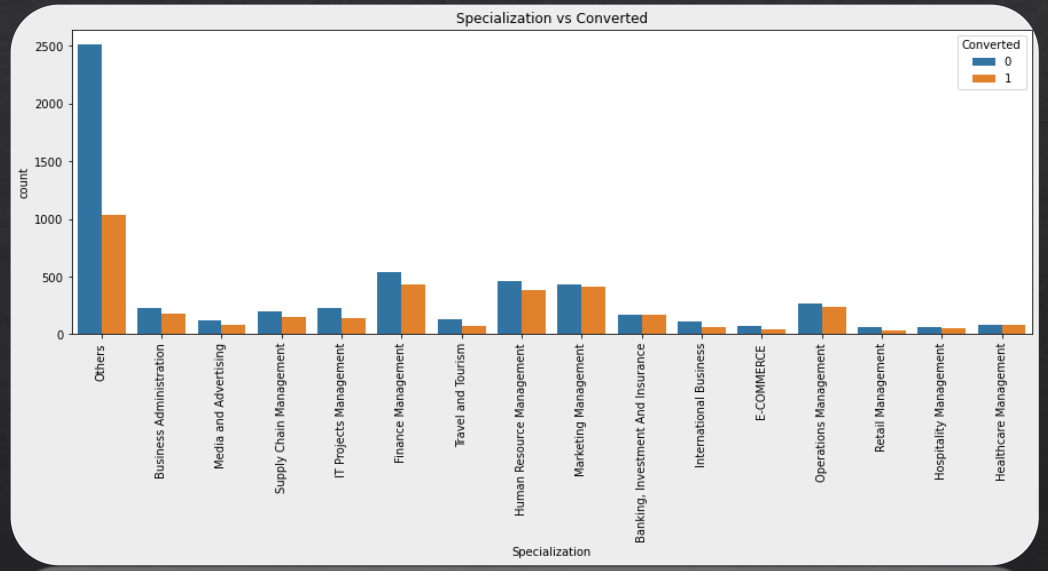
1. ~38% CONVERSION RATE FOR EMAIL OPENED
2. ~63% FOR SMS SENT , WHICH IS A GOOD TURNOVER FOR BUSINESS

SPECIALIZATION



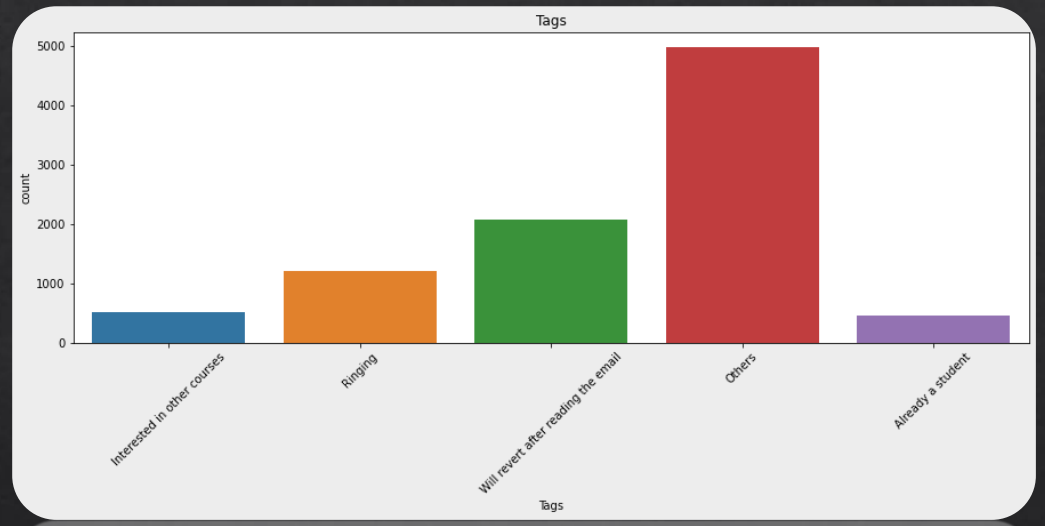
1. ~11% BELONG TO FINANCE MANAGEMENT

VS CONVERTED



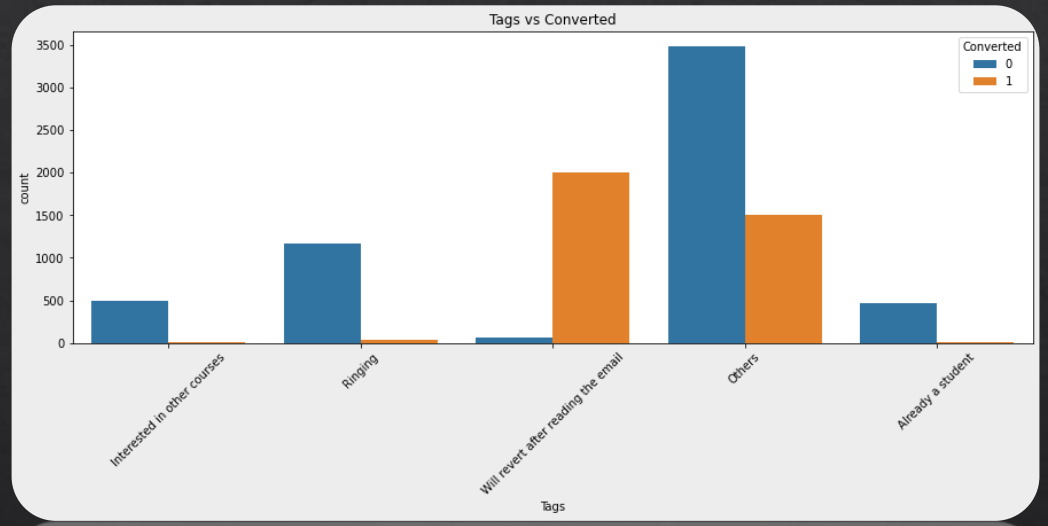
1. ~45% CONVERSION IS SEEN IN FINANCE MANAGEMENT
2. 46% IN HUMAN RESOURCE

TAGS



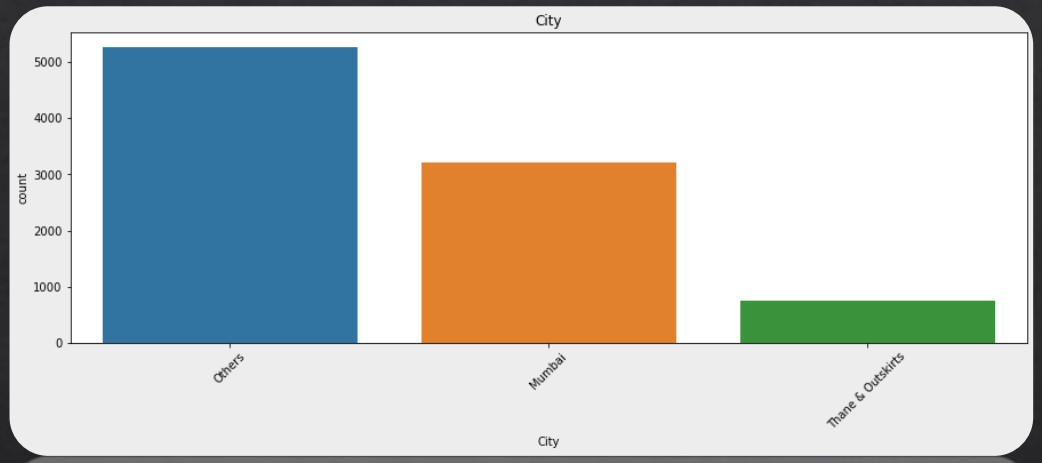
1. ~22% REVERT AFTER READING EMAIL IS SEEN
2. 13% WHERE STATUS HAS BEEN ON RINGING

VS CONVERTED



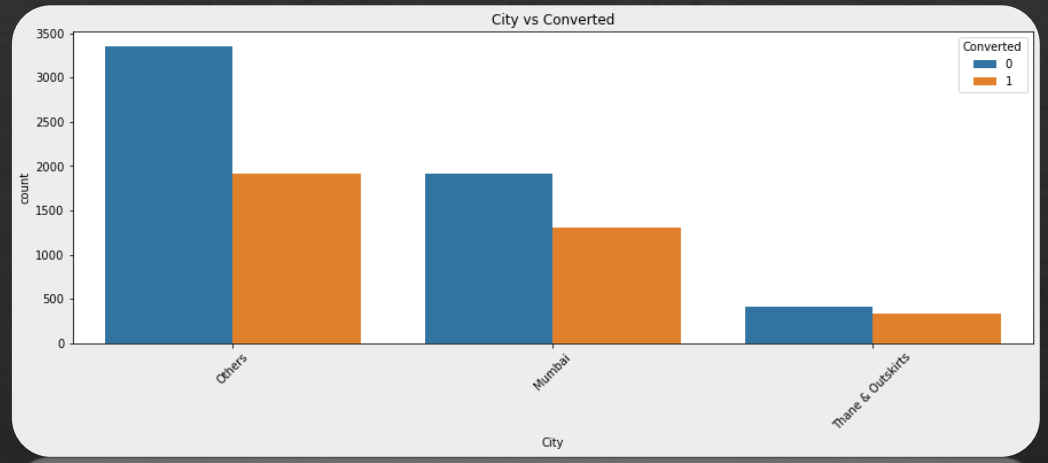
1. ~97% CONVERSION RATE IS SEEN FOR REVERT AFTER READING EMAIL
2. RINGING HAS ONLY A 3% CONVERSION.

CITY



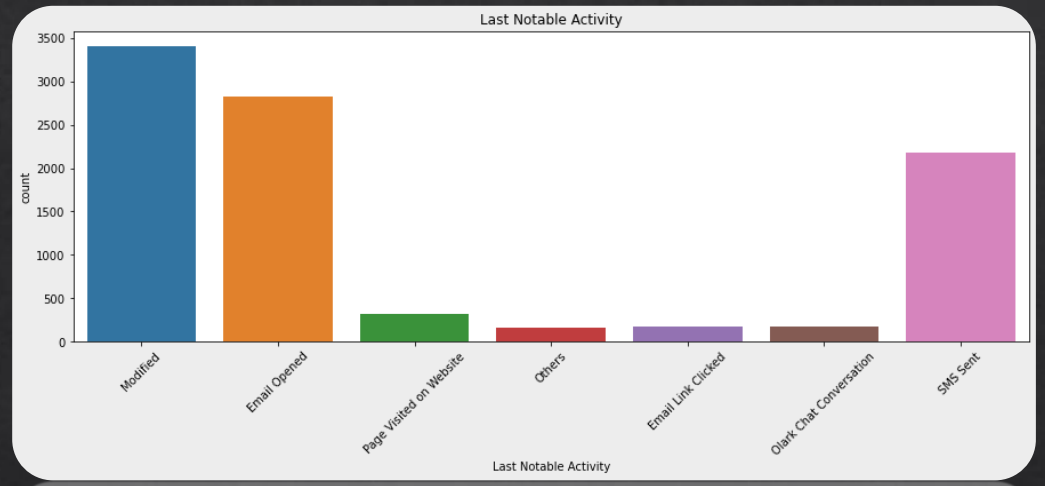
1. ~35% BELONGS ONLY TO MUMBAI

VS CONVERTED



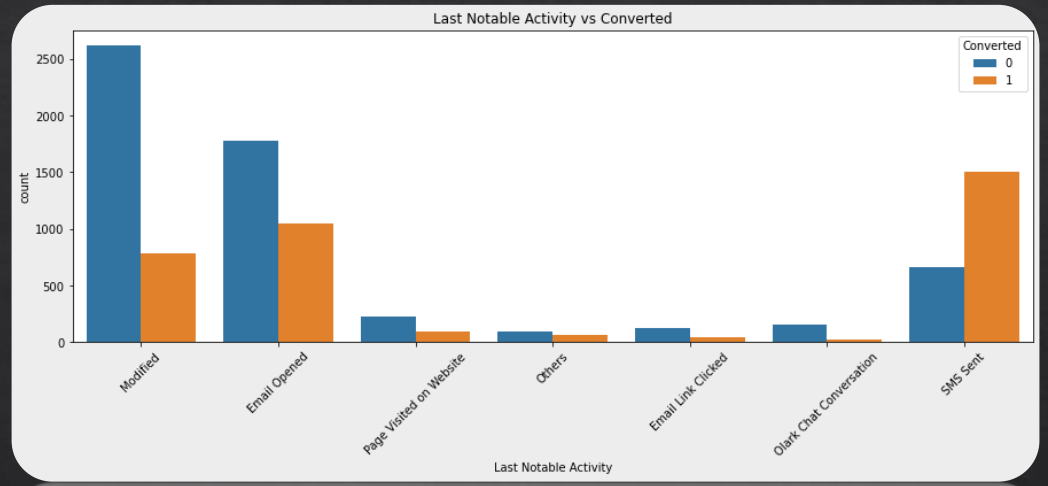
1. ~41% CONVERSION RATE IS OBSERVED FOR MUMBAI

LAST NOTABLE ACTIVITY



1. ~37% BELONGS TO MODIFIED
2. ~31% EMAIL OPENED
3. ~24% SMS SENT

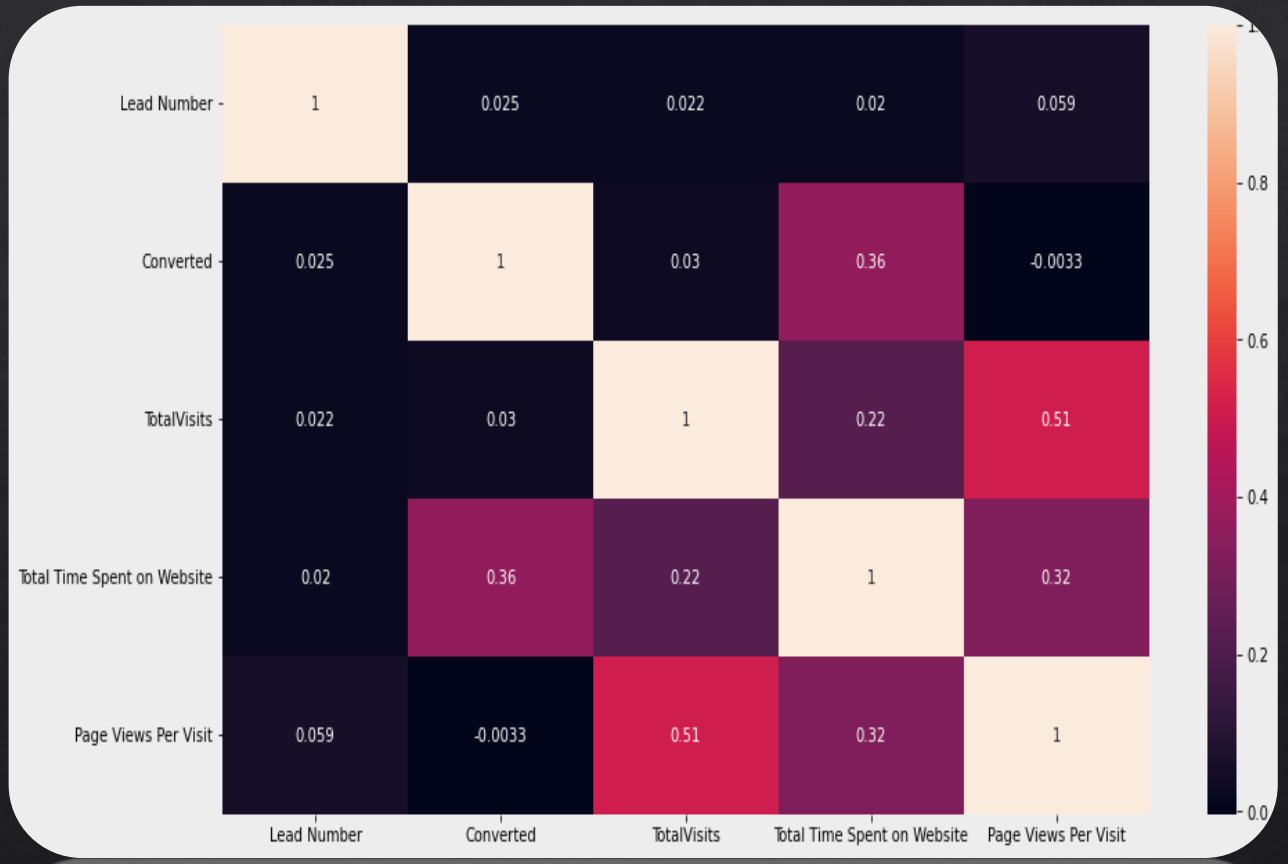
VS CONVERTED



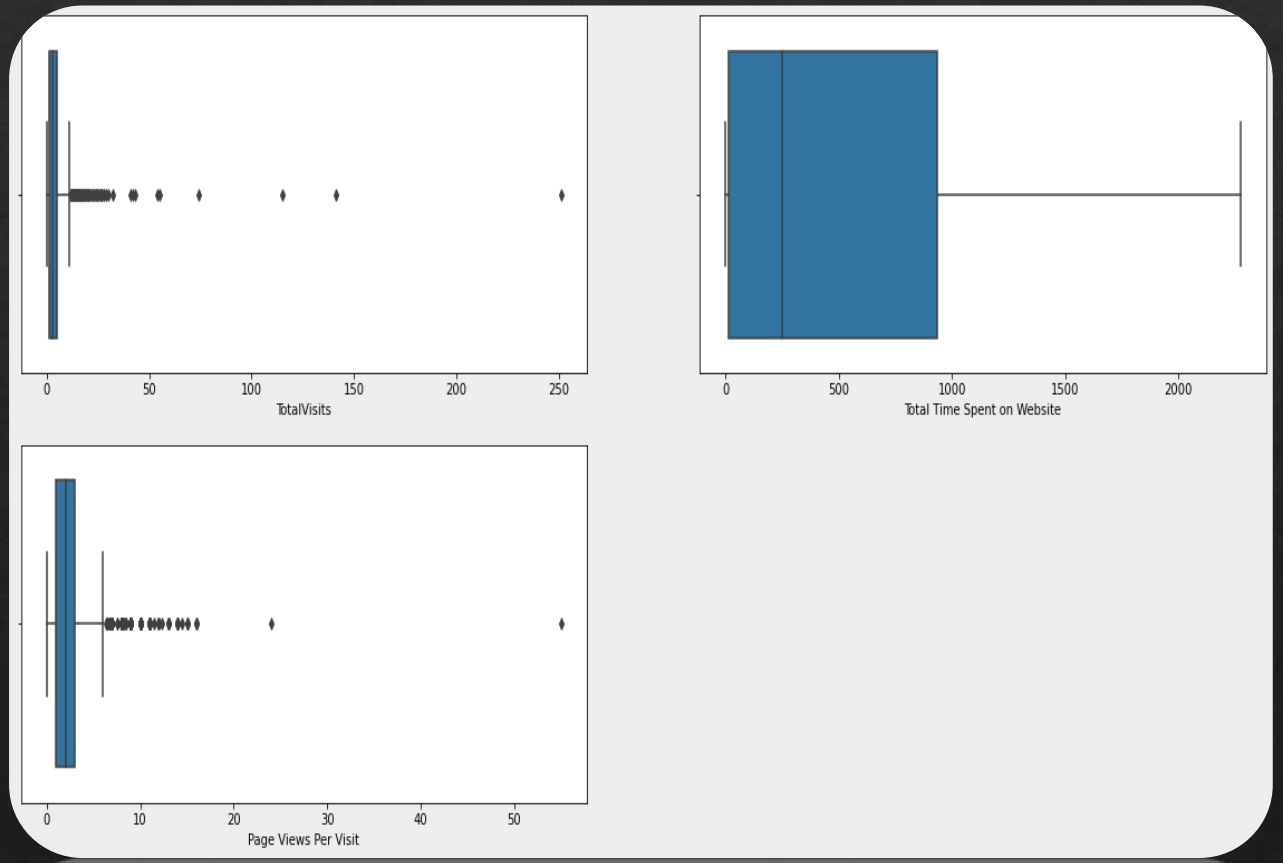
1. ~23% CONVERSION FOR MODIFIED
2. ~37 CONVERSION FOR EMAIL OPENED
3. ~69% FOR SMS SENT , WHICH IS HIGH.

NUMERICAL ANALYSIS

- PAGES PER VISIT HAS A GOOD AMOUNT OF CORRELATION WITH TOTAL VISITS.
- OTHERWISE THE DATA IS GOOD ENOUGH TO BE CONSIDERED FOR MODELLING



- WE CAN SEE MANY OUTLIERS PRESENT IN TOTAL VISITS AND PAGE VIEWS PER VISIT
- THESE OULIERS HAVE BEEN HANDLED BY CAPPING THEM UPTIL 1.5*IQR (Q3-Q1, 0.75 AND 0.25 RESPECTIVELY) VALUES FOR A CLEAN ANALYSIS



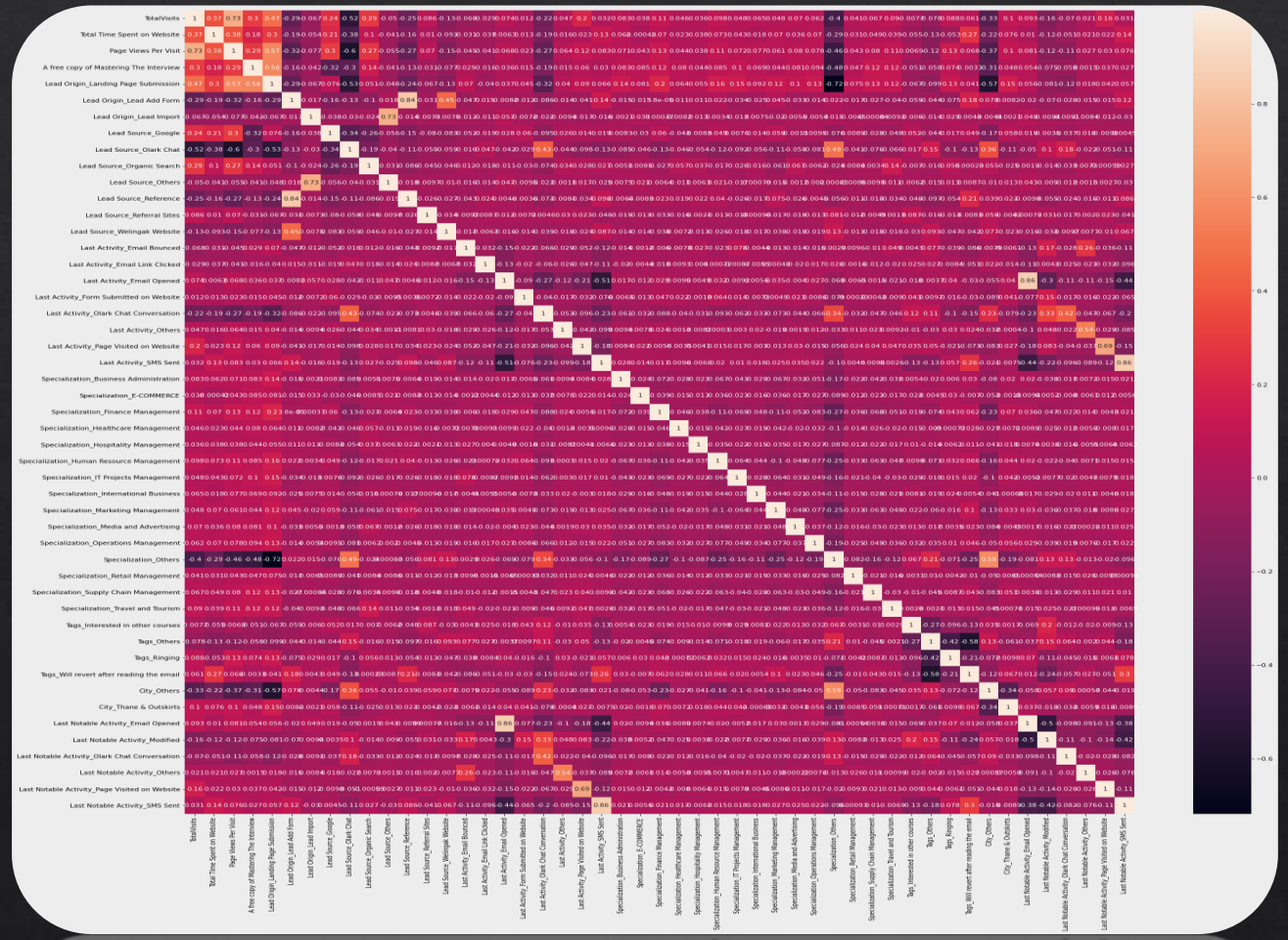
DUMMIES CORRELATION AFTER DATA SPLIT (80:20)

W/

HIGH POSITIVE AND NEGATIVE CORRELATIONS

	VAR1	VAR2	Correlation
2123	Last Notable Activity_Email Opened	Last Activity_Email Opened	0.862896
2373	Last Notable Activity_SMS Sent	Last Activity_SMS Sent	0.857740
544	Lead Source_Reference	Lead Origin_Lead Add Form	0.844642
98	Page Views Per Visit	TotalVisits	0.729415
496	Lead Source_Others	Lead Origin_Lead Import	0.727693
2323	Last Notable Activity_Page Visited on Website	Last Activity_Page Visited on Website	0.692017

	VAR1	VAR2	Correlation
1621	Specialization_Others	Lead Origin_Landing Page Submission	-0.718026
394	Lead Source_Olark Chat	Page Views Per Visit	-0.597803
1998	Tags_Will revert after reading the email	Tags_Others	-0.579584
2013	City_Others	Lead Origin_Landing Page Submission	-0.568112
396	Lead Source_Olark Chat	Lead Origin_Landing Page Submission	-0.526359
392	Lead Source_Olark Chat	TotalVisits	-0.521520



MODEL PREPARATION

- THE VARIABLES WITH A HIGH CORRELATION WERE DROPPED FOR MODELLING LEAVING US WITH 33 COLUMNS
- SCALING (STANDARD SCALER) WAS PERFORMED ON TWO NUMERICAL COLUMNS, i.e., TOTAL VISITS AND TOTAL TIME SPENT ON WEBSITE
- WE FURTHER GO AHEAD WITH FEATURE SELECTION USING RFE AND CONSIDER 15 VARIABLES TO START MODELLING

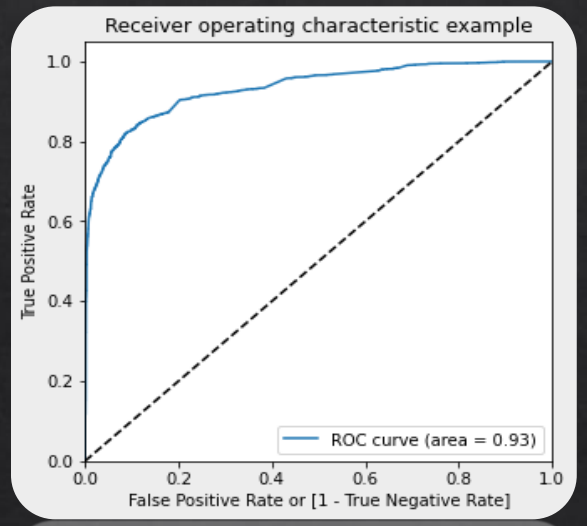
MODEL BUILDING
TRAIN SET

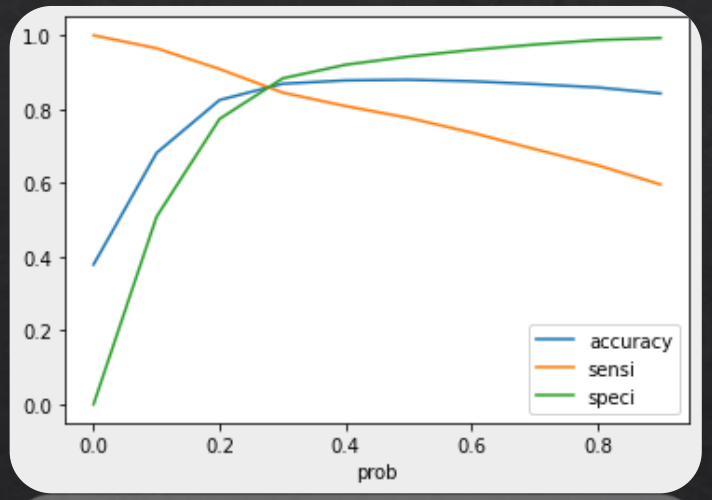
- WE BUILT THE MODEL USING LOGISTIC REGRESSION THROUGH MANUAL BUILDING AND ITERATIONS TILL NO SIGNIFICANT p-VALUES WERE OBSERVED
- THE FINAL MODEL HAS 10 VARIABLES (listed in the end) CONTRIBUTING HEAVILY TOWARDS THE PREDICTION
- CONFUSION MATRIX:

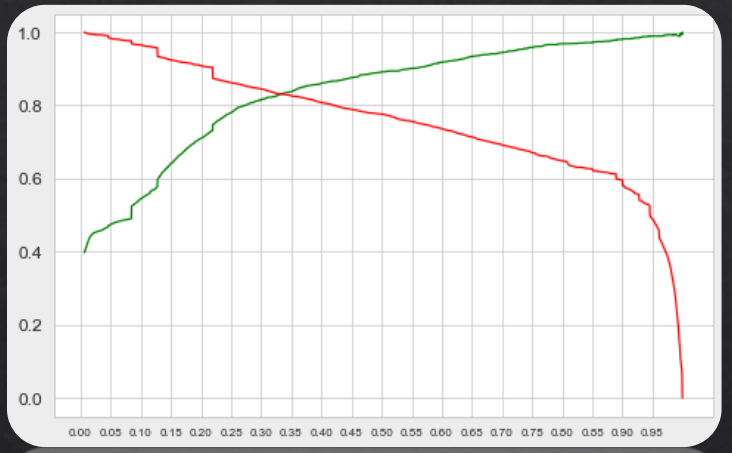
4264	**260**
617	2141

ACCURACY	~88%
SENSITIVITY	~78%
SPECIFICITY	~94%
FALSE-POSITIVE	~6%
POSITIVE PRED	~89%
NEGATIVE PRED	~87%

EVALUATION MATRIX







- THE GOODNESS OF THE MODEL IS DETERMINED USING ROC ; 0.93 IN OUR TRAIN CASE
- 0.28 WAS SEEN TO BE THE OPTIMUM THRESHOLD POINT, BUT AFTER CONSIDERING OUR PRECISION-RECALL TRADE-OFF WE GOT THE THRESHOLD OF 0.34

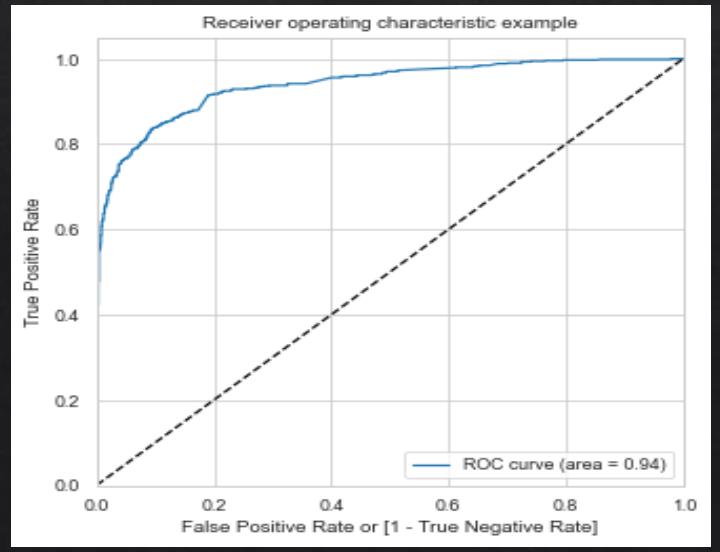
0.34 THRESHOLD

ACCURACY	~87%
SENSITIVITY	~83%
SPECIFICITY	~90%
FALSE POSITIVE RATE	~10%
POSITIVE PREDICTED VALUE	~83%
NEGATIVE PREDICTED VALUE	~90%
PRECISION	~83.4%
RECALL	~82.9%

THIS THRESHOLD SEEMED TO GIVE BETTER RESULTS AND WE STICK TO THIS THRESHOLD FOR OUR TEST SET

TEST SET PREDICTIONS

ACCURACY	~88%
SENSITIVITY	~84%
SPECIFICITY	~90%
PRECISION	~84.6%
RECALL	~83.6%



THE ROC / AREA UNDER THE CURVE WE GOT IS 0.94

~85% of the predicted values are converted
~84% of the conversion case values are predicted correctly

CONCLUSION

Lead Origin_Lead Add Form	4.471787
Tags_Will revert after reading the email	4.123419
Total Time Spent on Website	0.950652
Last Activity_Email Opened	-0.654462
Last Activity_Email Link Clicked	-0.767566
Last Activity_Page Visited on Website	-0.974386
Last Notable Activity_Modified	-1.116981
Last Activity_Email Bounced	-2.064843
Tags_Interested in other courses	-2.438755
Tags_Ringing	-3.353523

PARAMETERS USED IN THE MODEL ALONG WITH ITS CO-EFFICIENTS

LOG ODDS

WE CAN USE LOG ODDS ALONG WITH IT'S COEFFICIENT VALUES TO FIND OUT IF A NEW LEAD WILL BE CONVERTED OR NOT.

$$\text{Ln(P1-P)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + B_n x_n$$

- WHAT WE FOUND OUT FROM THIS ARE THE FACTOR VARIABLES THAT INFLUENCE THE PEOPLE TO BECOME POTENTIAL CLIENTS ARE
  - ORIGIN OF HIGHEST CONVERSION LEAD ADD FORM
  - TAGS , WHERE CONVERSION IS HIGHER IN GETTING BACK AFTER READING THE MAIL
  - TOTAL TIME SPENT ON THE WEBSITE
- NEGATIVE COEFFICIENTS HAVE BEEN OBSERVED FOR : HENCE WHEN THESE DEPENDENT VARIABLE DECREASES , THE INDEPENDENT VARIABLE INCREASES
- LAST ACTIVITIES WHERE , -EMAIL WAS OPENED ,-EMAIL LINKED WAS CLICKED ,-PAGE VISITED ON THE WEBSITE, -IF THE EMAIL HAS BOUNCED ARE THE AFFECTING AFCTORS
- TAGS WHEN ,
  - SELECTED INTEREST IN OTHER COURSES
  - PHONES BEEN ONLY IN RING STATUS
- HENCE THE COMPANY CAN FOCUS ON SENDING OUT MORE EMAILS AND SMSs
- TARGET PREVIOUS LEARNERS AND REFER MORE PEERS
- MAKE CALLS ONLY WHEN PEOPLE ARE SEEN TO VISIT THE WEBSITE FREQUENTLY OR SPENDING MORE TIME ON THE WEBSITE

THANK YOU