LEAD SCORING CASE STUDY

BY
-NIKHIL KUMAR SINGH
-ADITY DWIVEDI
-KARTEEK RAO

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%.

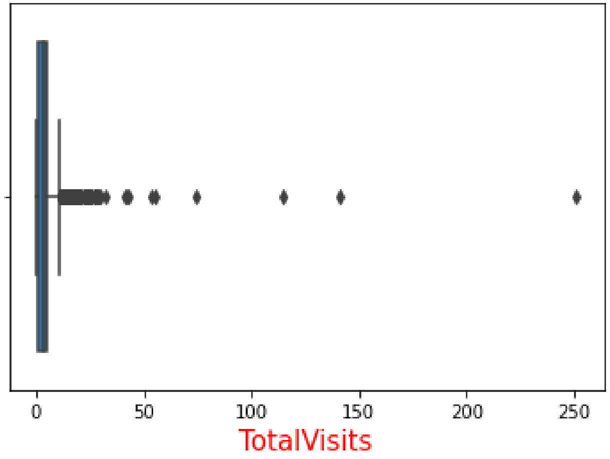
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

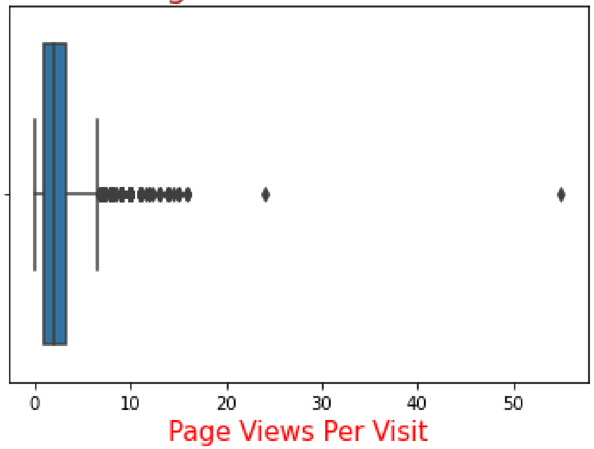
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

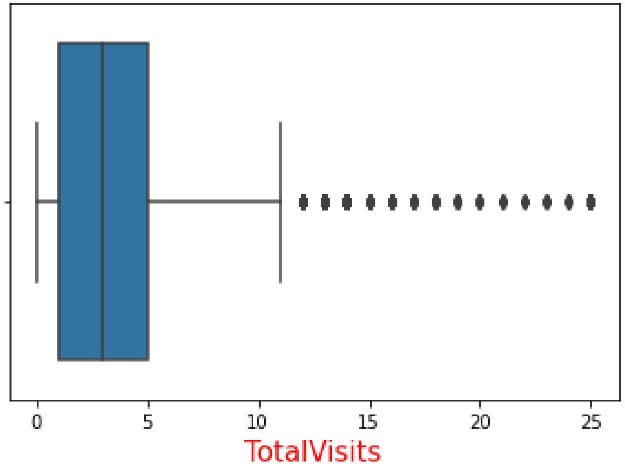
THE STEPS WE FOLLOW IN THIS EXERCISE ARE AS FOLLOWS:

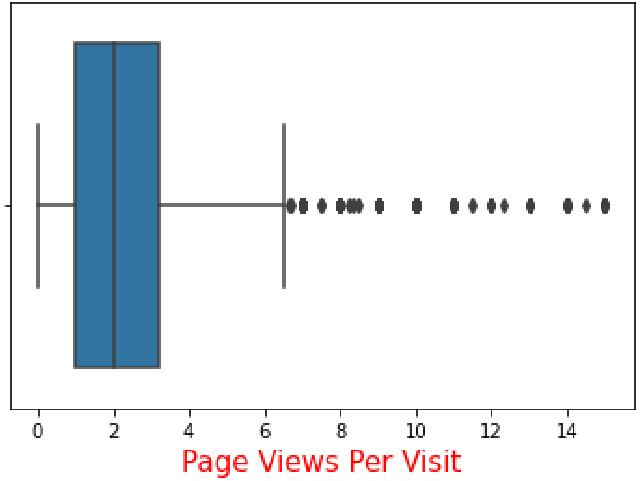
- READING & UNDERSTANDING THE DATA.
- HANDLING MISSING VALUES.
- TREATING OUTLIERS & IMBALANCE DATA.
- VISUALISING THE DATA.
- DATA PREPRATION(mapping & dummy variables).
- SPLITTING INTO TRAIN-TEST SET(dividing into X & y).
- RESCALING.
- BUILDING A LINEAR MODEL (using RFE & statsmodel & checking VIF also).
- PLOTTING ROC CURVE.
- FINDING OPTIMAL CUTOFF POINT.
- PRECISION & RECALL (precision_recall_curve).
- PREDICTION & EVALUATION ON TEST-SET.

AFTER HANDLING OUTLIERS

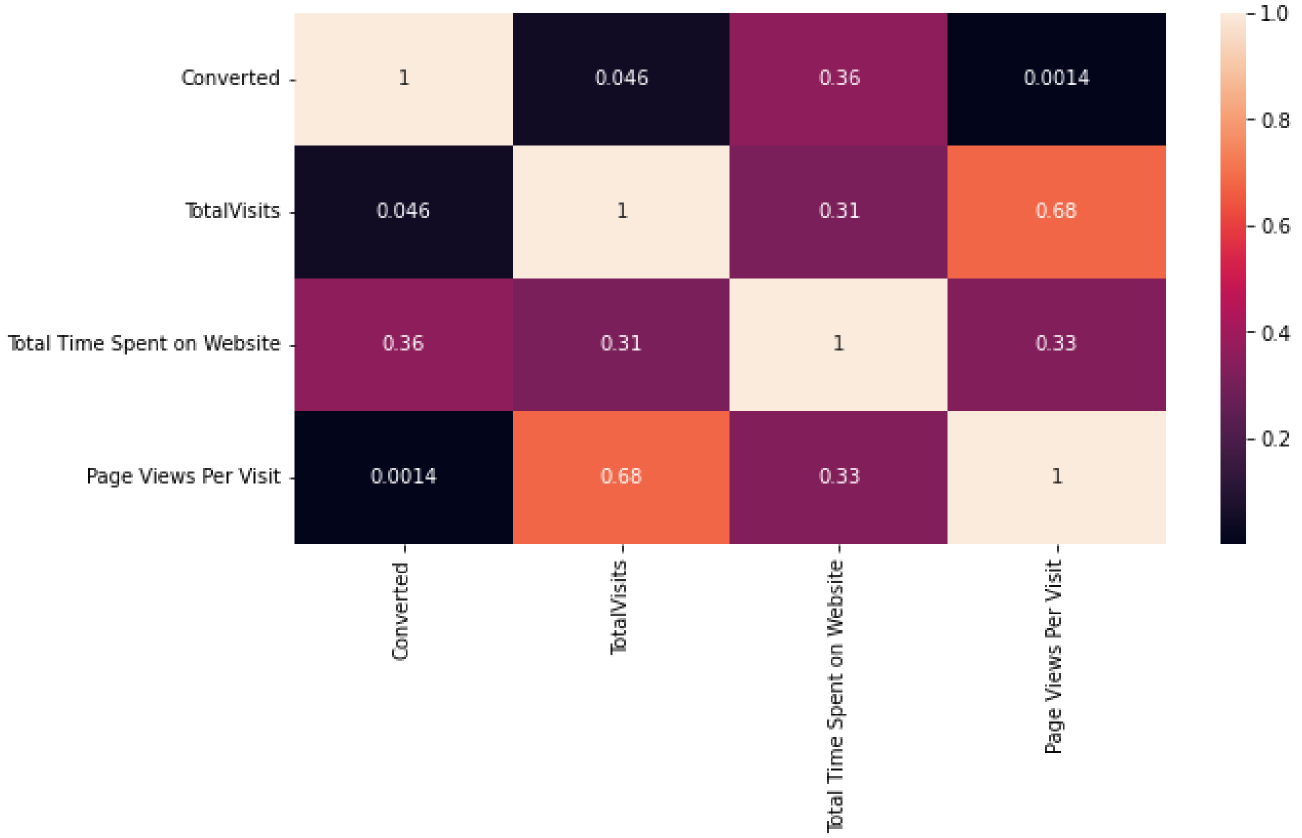

Distribution of TotalVisits


Distribution of Page Views Per Visit


Stastical Distribution of TotalVisits

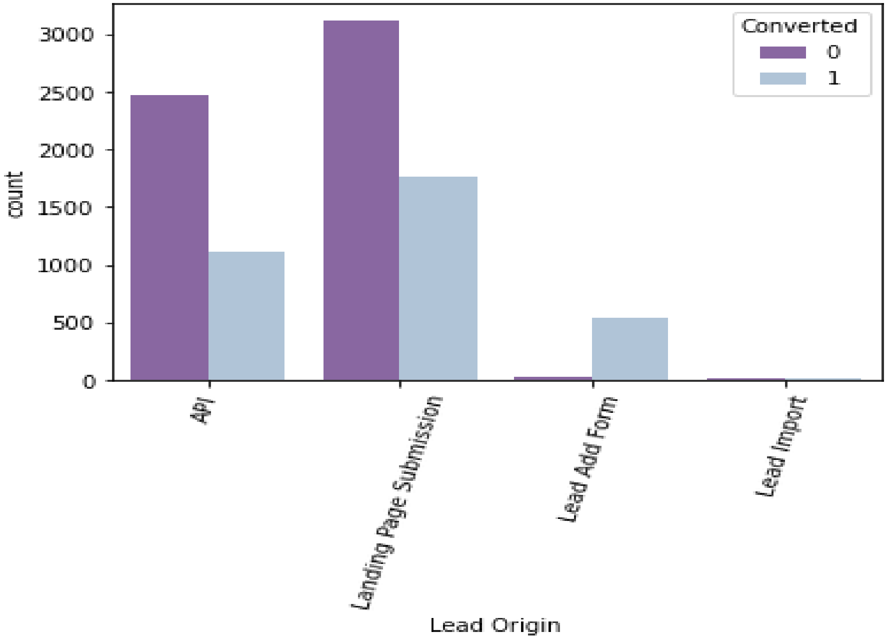

Stastical Distribution of Page Views Per Visit

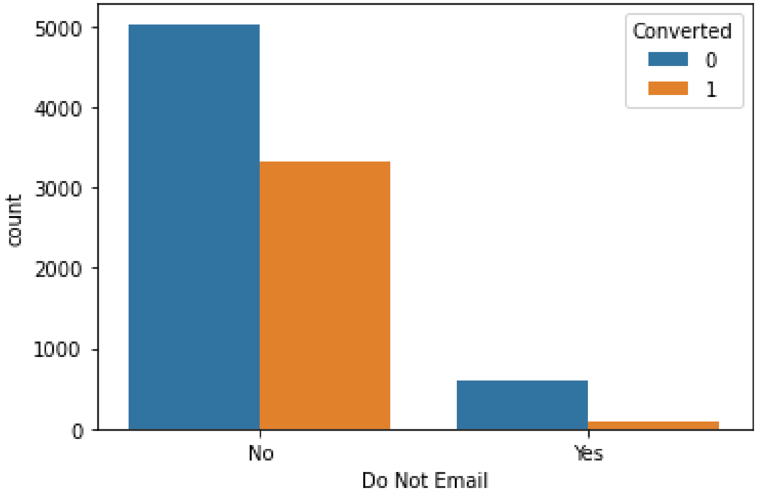
HEAT MAP FOR CORRELATION OF NUMERICAL VARIABLES



- **TOTAL VISITS & PAGE VIEWS PER VISIT ARE POSITIVELY CORRELATED (0.68).**
- **CONVERTED IS POSITIVELY CORRELATED TO TOTAL TIME SPENT ON WEBSITE (0.36).**

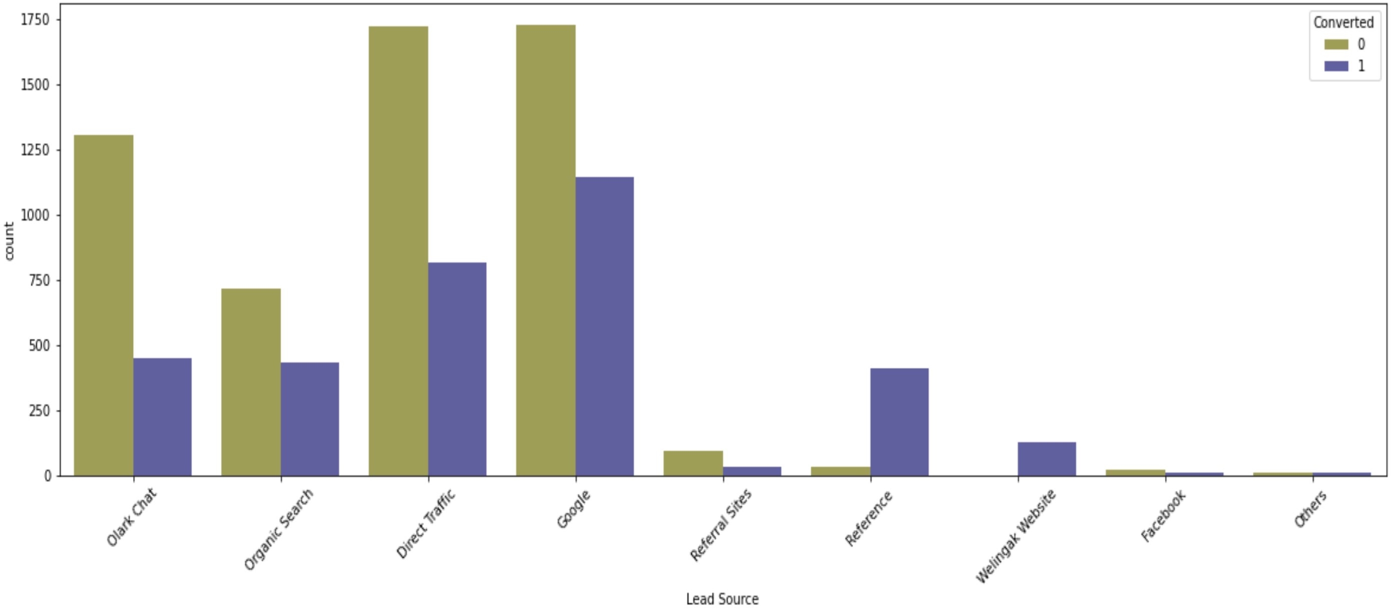
VISUALIZING VARIABLES W.R.T CONVERTED





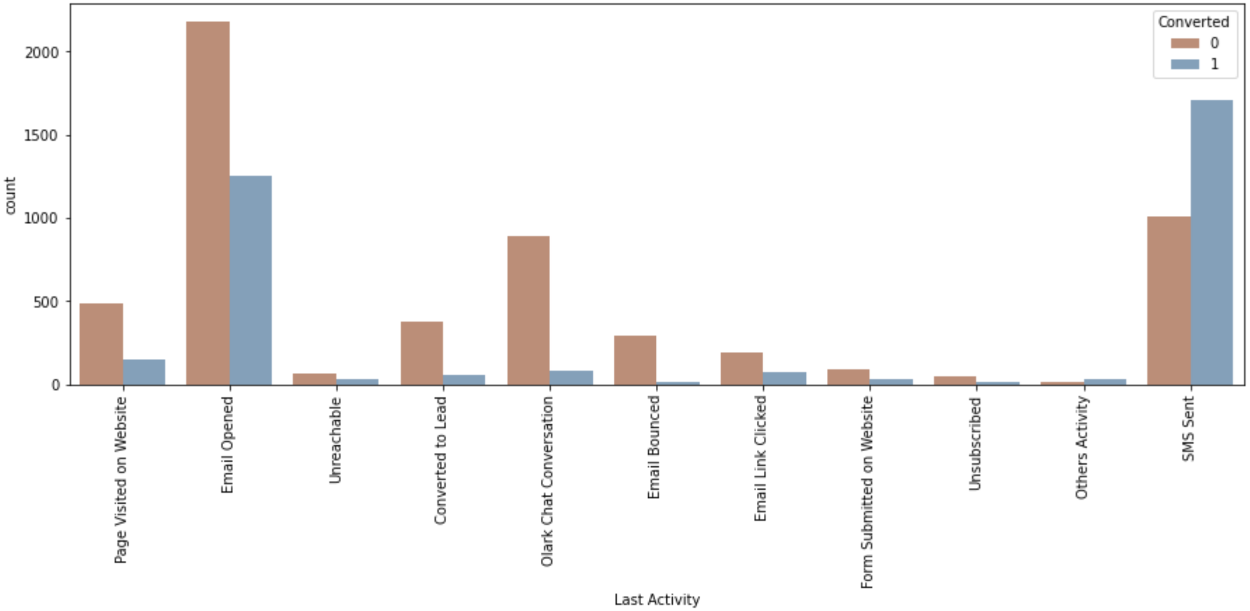
- **MOST OF THE LEADS DON'T WANT TO BE EMAILED ABOUT THE COURSE**

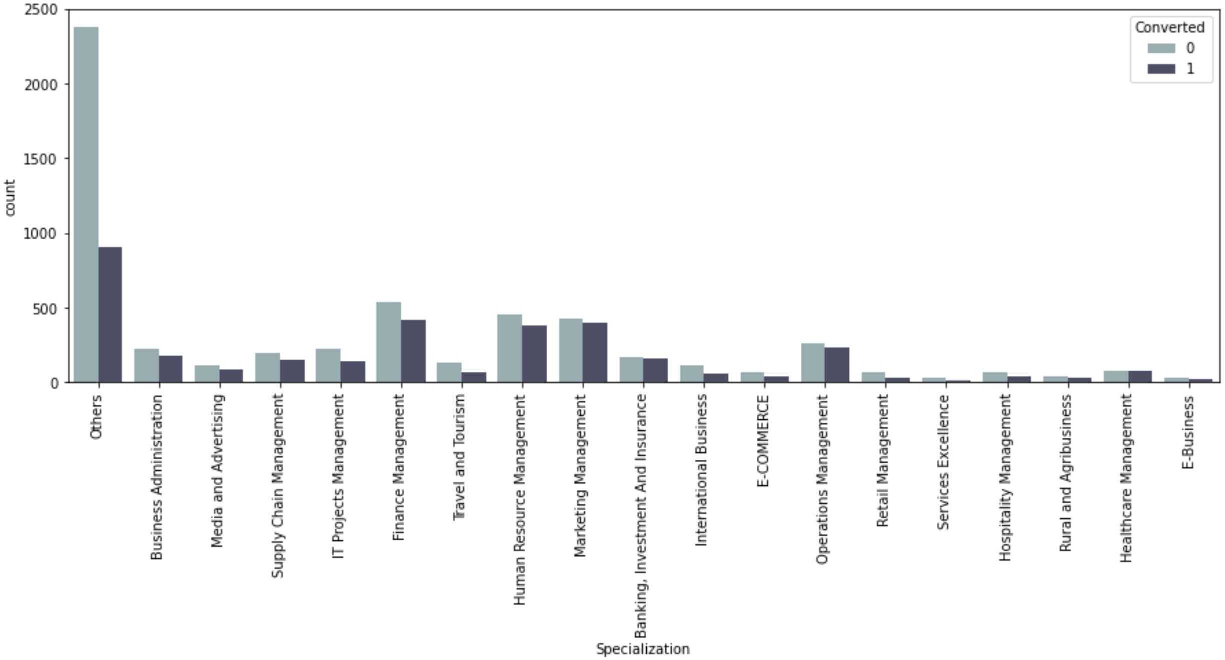
- **API & LANDING PAGE SUBMISSION BRING HIGHER NUMBER OF LEADS COUNT BUT THERE CONVERSION RATE IS LESS.**
- **CONVERSION RATE OF LEAD ADD FORM IS HIGH BUT THE NUMBER OF COUNT IS LESS.**
- **LEAD IMPORT ARE VERY VERY LESS IN COUNT.**



- **GOOGLE & DIRECT TRAFFIC GENERATES MAXIMUM NUMBER OF LEADS.**
- **CONVERSION RATE OF REFERENCE & WELINGAK WEBSITE IS HIGH.**

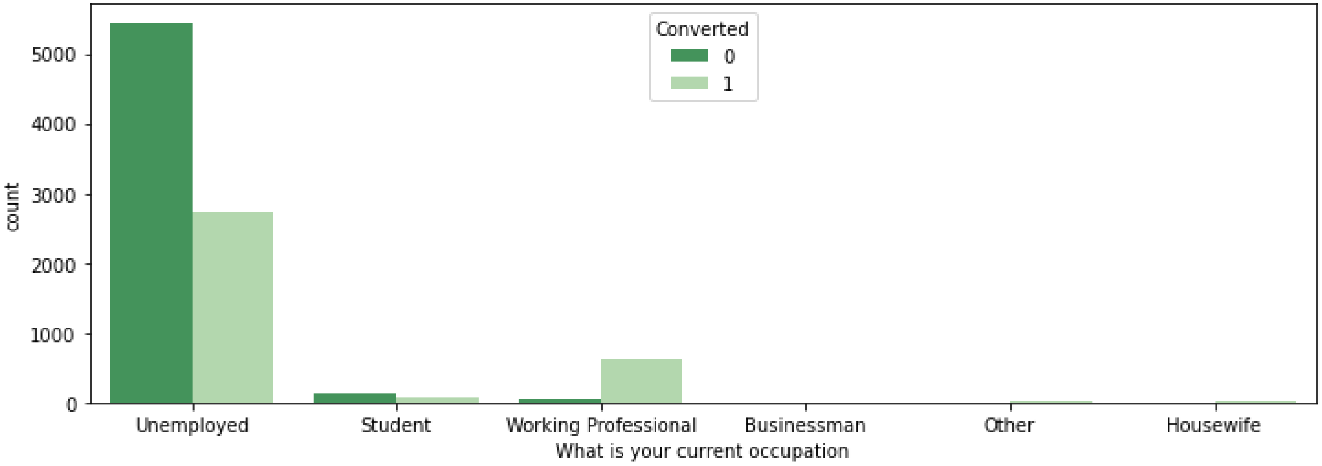
- **CONVERSION RATE FOR LEADS WITH THERE LAST ACTIVITY AS SMS SENT IS HIGH.**
- **MOST OF THE LEADS HAVE THERE EMAIL OPENED IN THERE LAST ACTIVITY.**

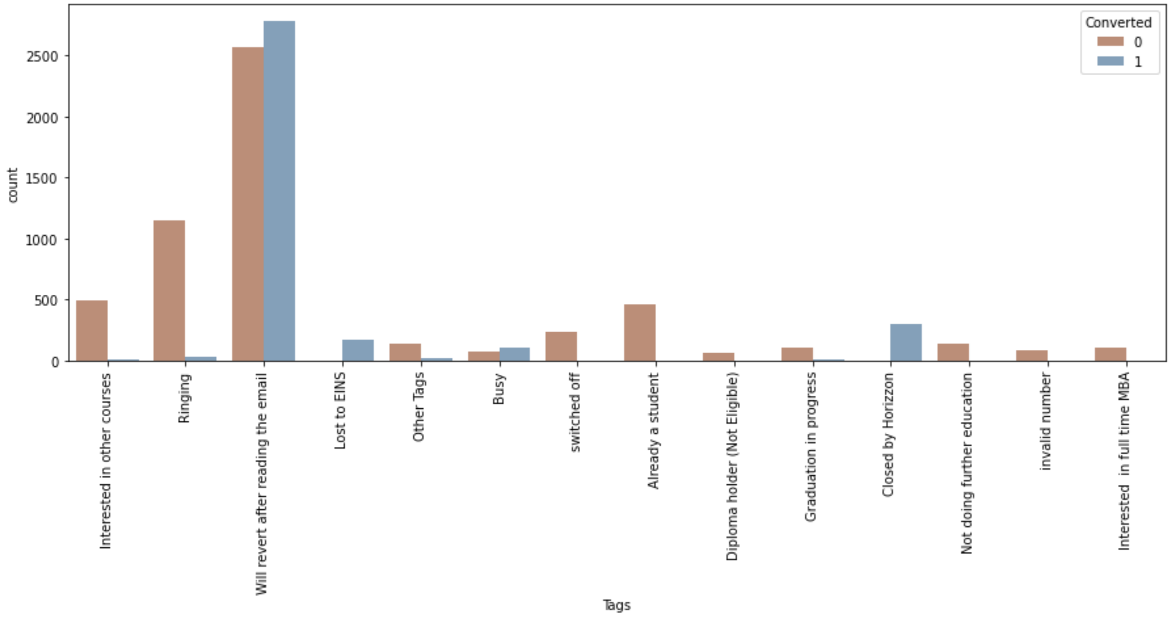




- **NON-CONVERSION RATE IS MORE THAN CONVERSION RATE IN EVERY SPECIALISATION.**
- **OTHER's CATEGORY SPECIALISATION(maybe student's or someone else) HAS MAXIMUM CONVERTED LEADS.**

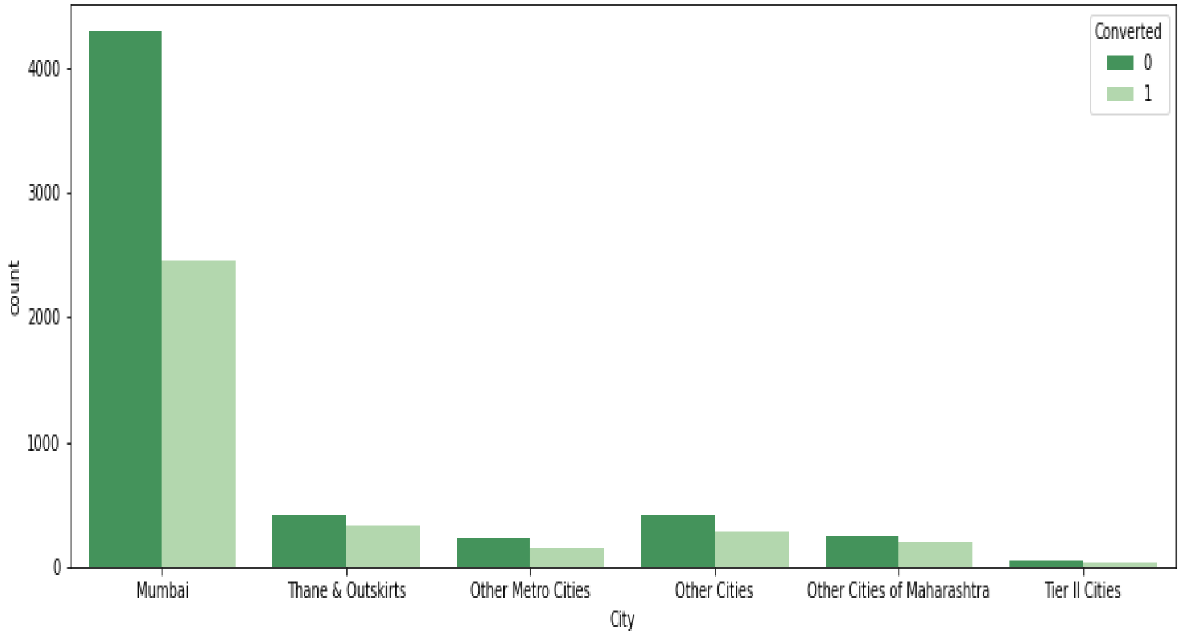
- **UNEMPLOYED LEADS ARE THE MOST IN NUMBERS BUT THERE CONVERSION RATE IS HALF THE NUMBER OF LEADS.**
- **WORKING PROFESSIONAL HAV E HIGH CHANCES OF JOINING THE COURSE AS THERE CONVERSION RATE IS GOOD.**

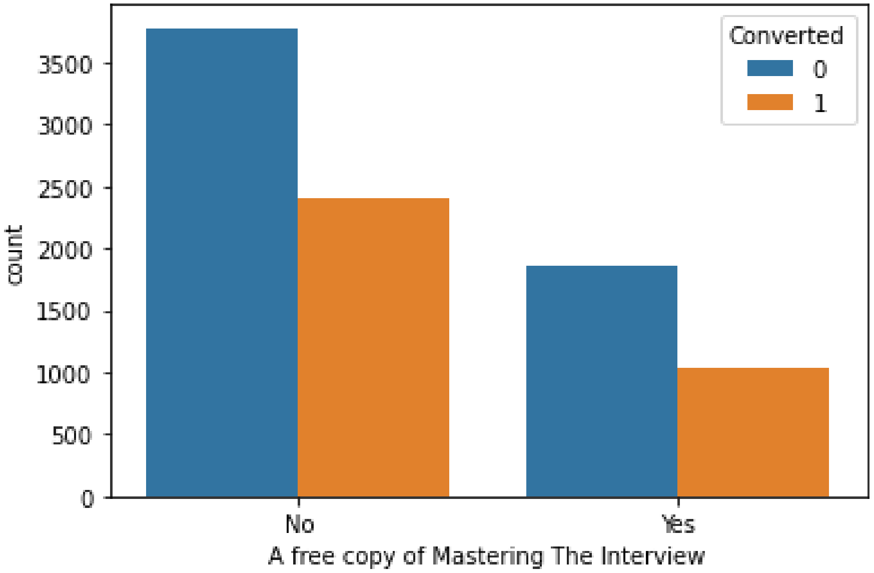




- **WILL REVERT AFTER READING THE EMAIL HAS MAXIMUM NUMBER OF LEADS COUNT & IT's CONVERSION RATE IS ALSO GOOD.**
- **CLOSED BY HORIZON & LOST TO EINS HAS GOOD CONVERSION RATE BUT THERE NUMBER OF COUNT IS NOT GOOD.**

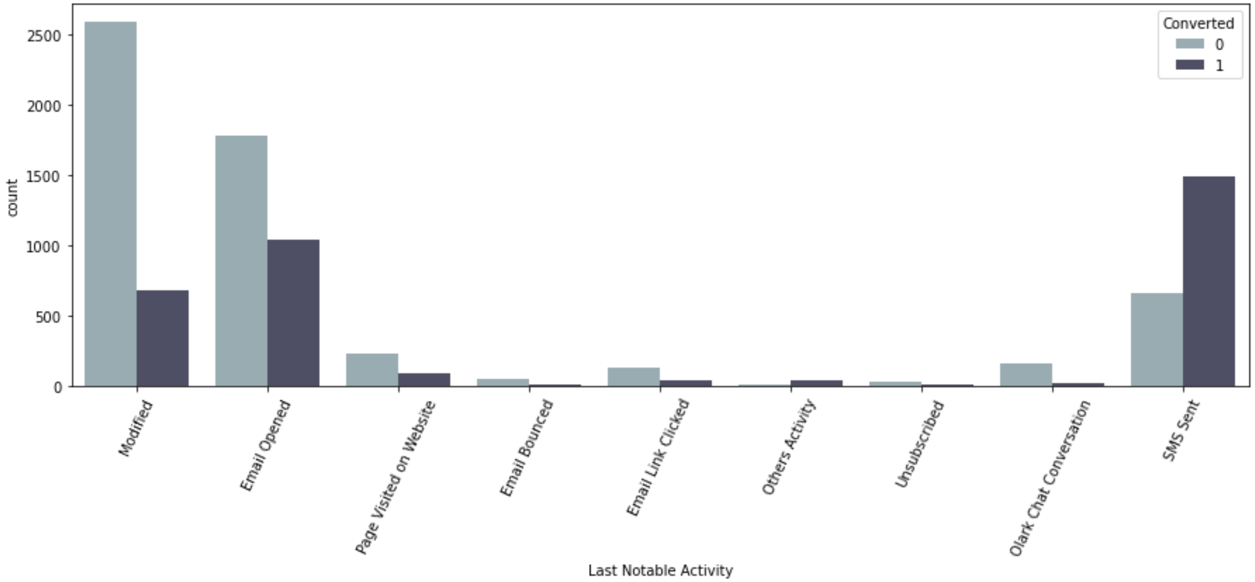
- **MOST LEADS ARE FROM MUMBAI.**
- **CONVERSION RATE OF ALL THE CITY ARE LESS.**



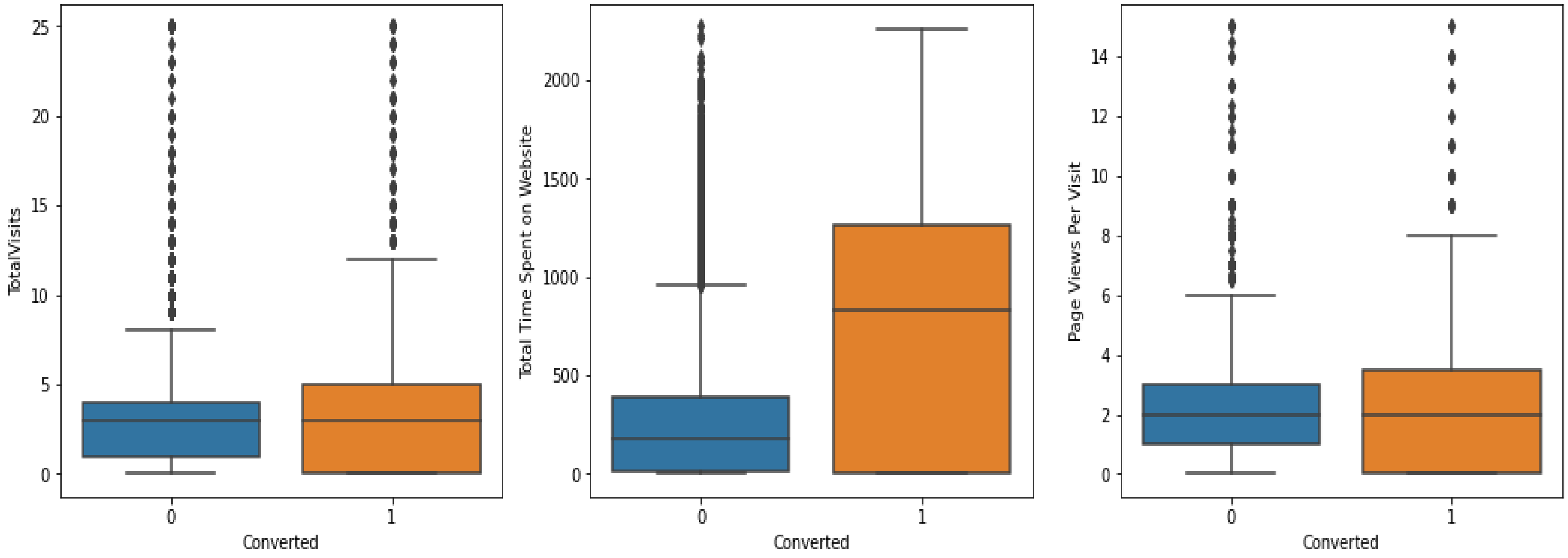


- •THE PATTERNS LOOK EXACTLY THE SAME FOR BOTH YES & NO, BUT THE COUNT OF NO’s ARE MORE.
- •NO INFERENCE CAN BE DRAWN FROM THIS.

- **CONVERSION RATE FOR LEADS WITH THERE LAST NOTABLE ACTIVITY AS SMS SENT IS HIGH.**
- **MOST OF THE LEADS HAVE MODIFIED & HAVE THERE EMAIL OPENED IN THERE LAST NOTABLE ACTIVITY.**



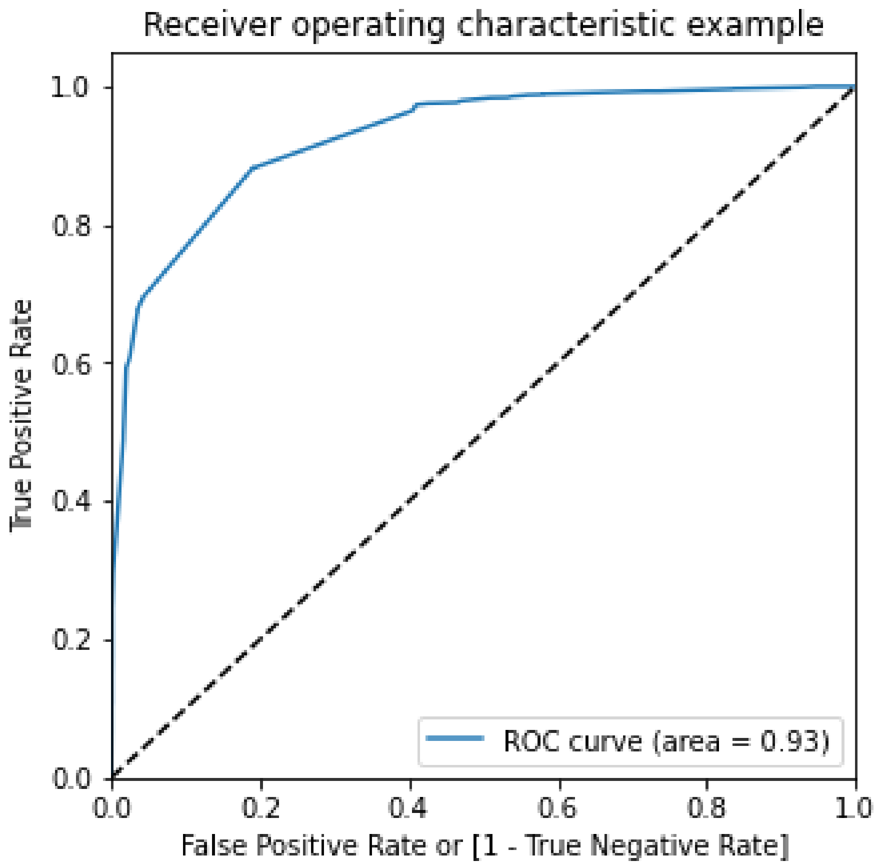
VISUALISING THE NUMERICAL VARIABLES



- **LEADS SPENDING MORE TIME ON THE WEBSITE ARE MORE LIKELY TO BE CONVERTED.**
- **MEDIANS FOR CONVERTED & NOT-CONVERTED OF TOTAL VISITS & PAGE VIEWS PER VISIT ARE SAME.**

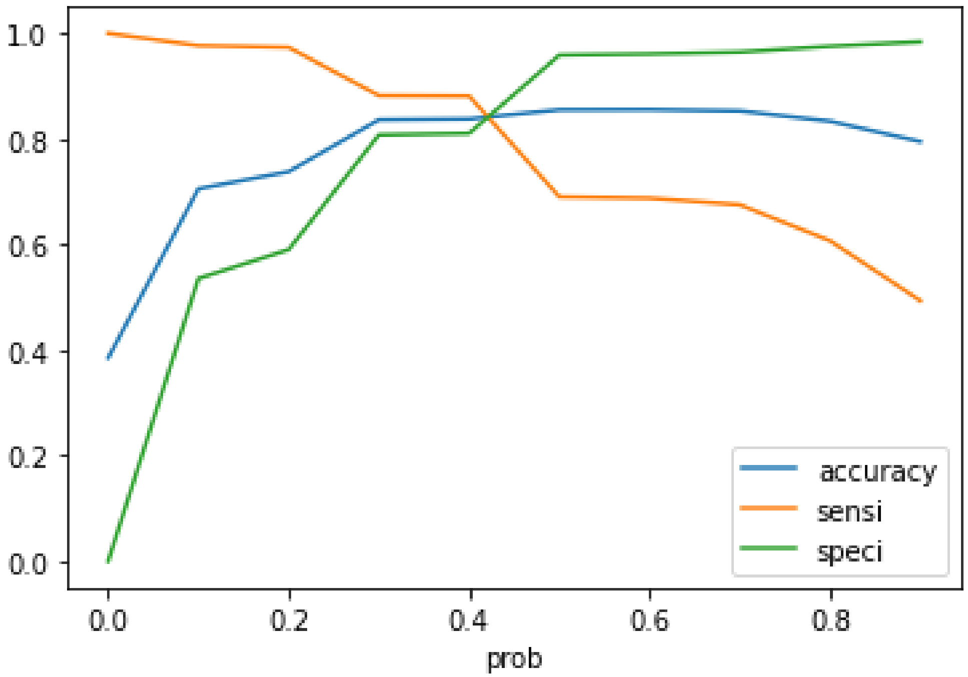
ROC CURVE

- 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.



OPTIMAL CUT OFF POINT

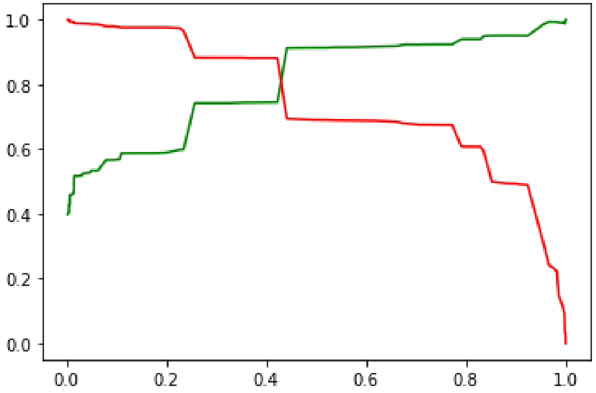
- POINT WHERE WE GET BALANCED SENSITIVITY AND SPECIFICITY.
- IN THIS CASE IT IS 0.42.

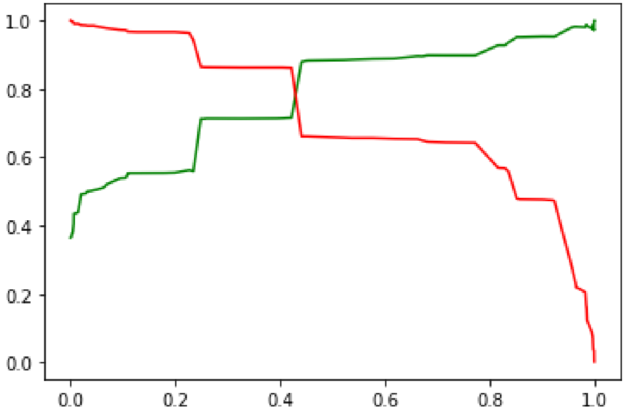


PRECISION RECALL CURVE

TRAINING DATA

TESTING DATA





- **THE PRECISION RECALL CURVE FOR BOTH TRAINING DATA AND TESTING DATA ARE EXACTLY THE SAME.**
- **THE FINAL PREDICTION OF CONVERSION HAVE A TARGET RATE OF 86%, WHICH IS MORE THAN EXPECTED(80%).**

COMPARING ACCURACY, SENSITIVITY, SPECIFICITY, PRECISION & RECALL SCORE OF TRAIN & TEST SET

- Train-set accuracy score : 0.8379782711384034
  Test-set accuracy score : 0.8255600440690415
- Train-set sensitivity : 0.8810302534750614
  Test-set sensitivity : 0.8614762386248737
- Train-set specificity : 0.81101152368758
  Test-set specificity : 0.8050749711649365
- Train-set precision score : 0.7449014863463532
  Test-set precision score : 0.7159663865546219
- Train-set recall score : 0.8810302534750614
  Test-set recall score : 0.8614762386248737

CONCLUSION

- **THE VARIABLES THAT MATTERS MOST ARE :-**
- TOTAL TIME THEY SPENT ON WEBSITE.
- TOTAL NUMBER OF VISITS.
- WHEN THE LEAD SOURCE WAS :-
  - GOOGLE
  - DIRECT TRAFFIC
  - REFERENCE
  - WELINGAK WEBSITE
- WHETHER THERE CURRENT OCCUPATION IS WORKING PROFESSIONAL OR IF THEY ARE UNEMPLOYED.
- WHEN THE LEAD ORIGIN IS LEAD ADD FORM.
- WHEN LAST ACTIVITY WAS :-
  - SMS SENT
  - EMAIL OPENED.
- WHEN TAGS WAS :-
  - CLOSED BY HORIZZON
  - LOST TO EINS.