



LEAD SCORING CASE STUDY

PRESENTED BY:

SAURAV ROY


DEVSHRI PATIL

KARTHIK K



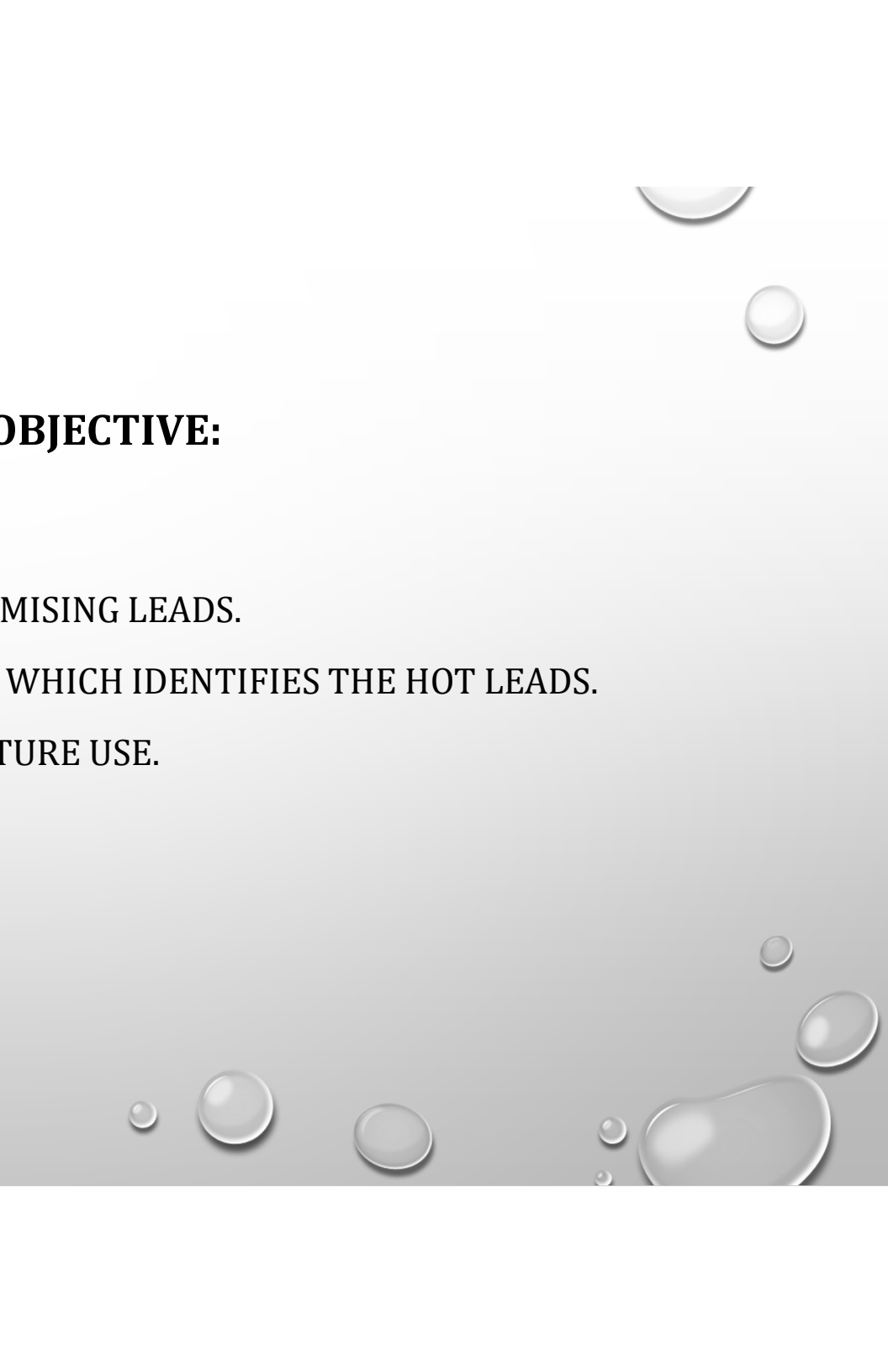


PROBLEM STATEMENT

- 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%.
- 



BUSINESS OBJECTIVE:

- X EDUCATION WANTS TO KNOW MOST PROMISING LEADS.
 - FOR THAT THEY WANT TO BUILD A MODEL WHICH IDENTIFIES THE HOT LEADS.
 - DEPLOYMENT OF THE MODEL FOR THE FUTURE USE.
- 

SOLUTION METHODOLOGY:

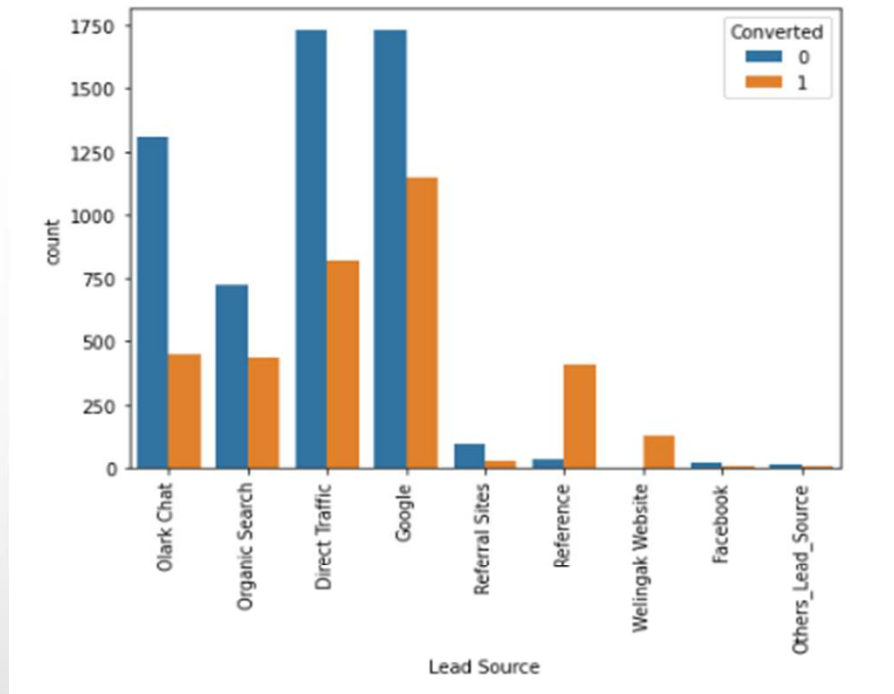
- DATA CLEANING AND DATA MANIPULATION:

1. CHECKING THE TOTAL SHAPE WHICH IS (9240,37).
2. FINDING THE % OF MISSING DATA AND HANDLING THE MISSING DATA ACCORDINGLY. INCASE THE % OF MISSING DATA IS MORE THAN 40%, THE COLUMN IS DROPPED EXCEPT FOR 'LEAD QUALITY' AS IT IS AN IMPORTANT PARAMETER.
3. COLUMNS WHICH HAVE 'SELECT' AS VALUE IS EQUIVALENT TO NULL, SO HANDLING THOSE VALUES. IN MOST CASES IT IS MAPPED WITH THE MOST FREQUENT VALUE AND IN OTHER CASES IF THE % OF SUCH DATA IS HIGH, THE ENTIRE COLUMN IS DROPPED.
4. COLUMNS HAVING DATA AS 'YES/NO' ARE MAPPED UNDER 0/1.

- EDA:

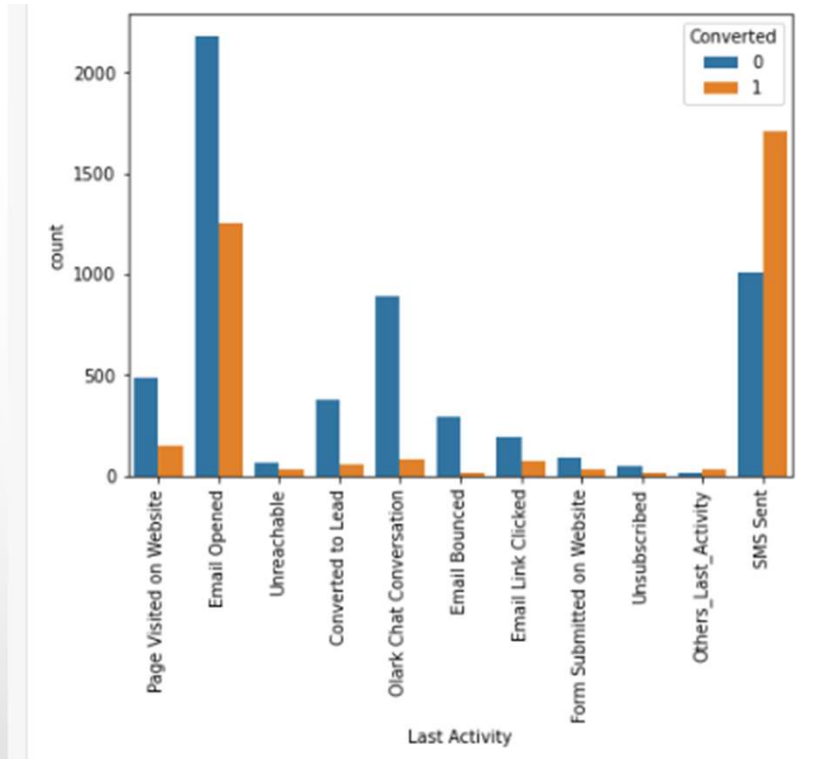
1. IT IS CHECKED AFTER DATA CLEANING WHETHER THE BALANCE IN DATASET IS PROPER OR NOT. AND '0' BEING 62% AND '1' BEING 37% DEPICT THAT THERE IS NO IMBALANCE IN THE DATASET.
2. IN FEW COLUMNS WHERE THE REPRESENTATION OF PARAMETER IS VERY LOW, THE SAME IS MAPPED UNDER NEW PARAMETERS. EG. UNDER COLUMN NAME 'LEAD SCORE' - BLOG, SOCIAL MEDIA ETC ARE MAPPED UNDER 'OTHERS_LEAD_SOURCE'
3. IN SOME CASES SPELL CHECK IS CORRECTED AND HAVE BEEN MAPPED UNDER ONE CATEGORY. EG GOOGLE UNDER COLUMN NAME 'LEAD SOURCE'.

SCREENSHOTS OF THE ANALYSIS IS PROVIDED IN THE NEXT SLIDES.



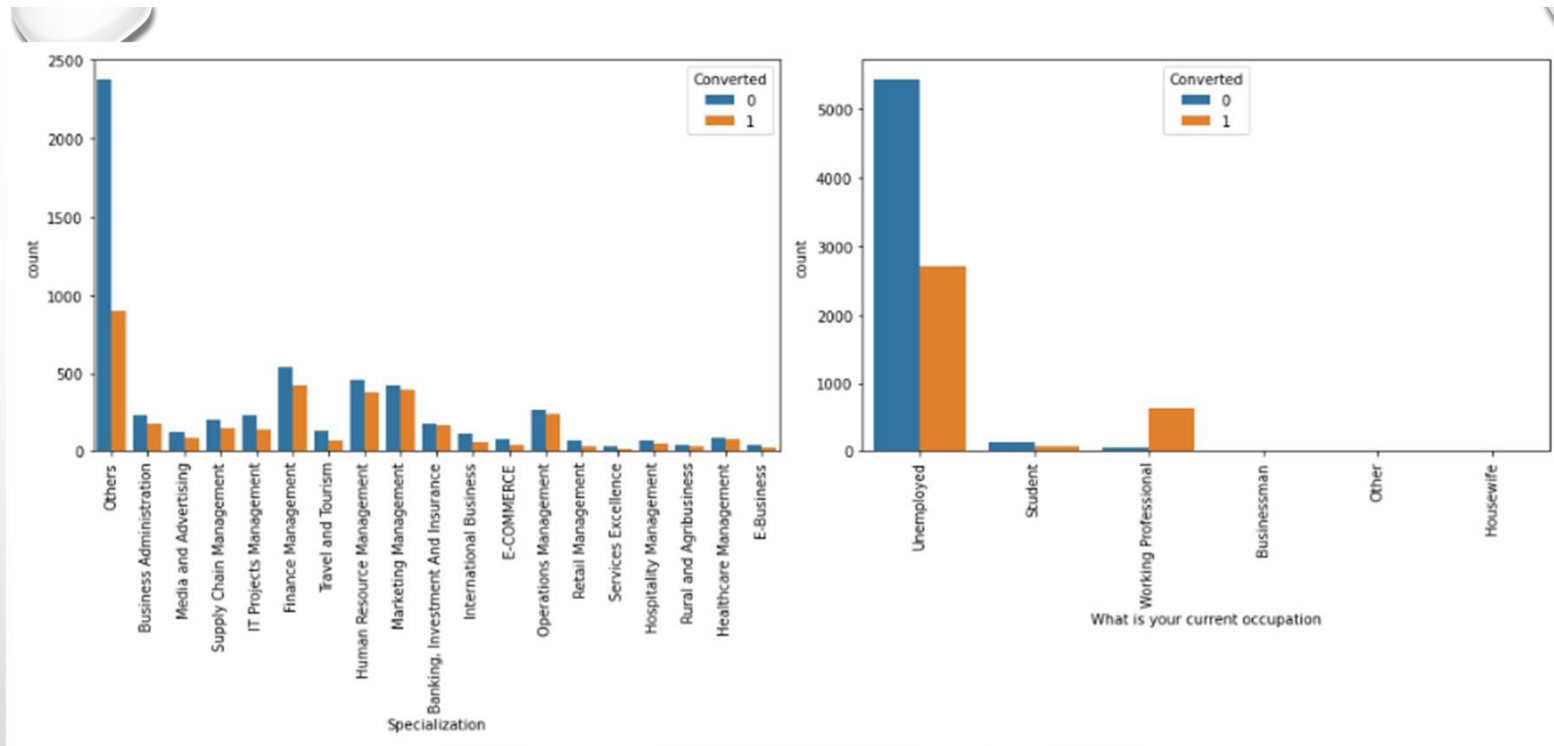
- **FROM THE ABOVE 'LEAD SOURCE' GRAPH FOLLOWING OBSERVATIONS ARE MADE:**

1. THE COUNT OF DIRECT TRAFFIC AND GOOGLE IS HIGH AND SO AS THE CONVERSION RATE IS. BUT THE NUMBER OF 'NO' IS ON A HIGHER SIDE.
 2. IT IS OBSERVED THAT CONVERSION RATE THROUGH REFERENCE AND WELINGAK WEBSITE IS HIGH.
 3. THE CONVERSION RATE OF OLARK CHAT IS LOW COMPARED TO THE COUNT.
- THUS, TO INCREASE THE OVERALL CONVERSION RATE, FOCUS SHOULD BE TO INCREASE THE CONVERSION RATE THROUGH GOOGLE AND DIRECT TRAFFIC. ALSO, THE % OF REFERENCE NEEDS TO BE INCREASED AS IT IS OBSERVED THAT CHANCES OF CONVERSION THROUGH REFERENCE IS HIGH ALONG WITH WELINGAK WEBSITE. ALONG WITH THE ABOVE ORGANIC SEARCH AND OLARK CHAT SHOULD ALSO BE FOCUSED AS THE POTENTIAL OF CONVERSION IS HIGH.



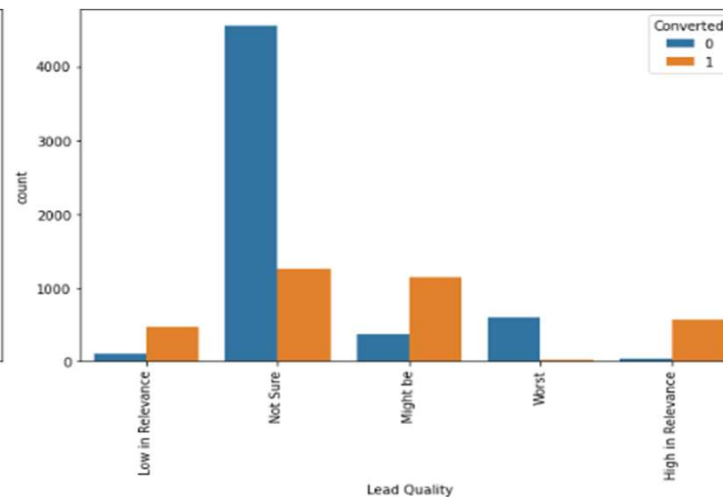
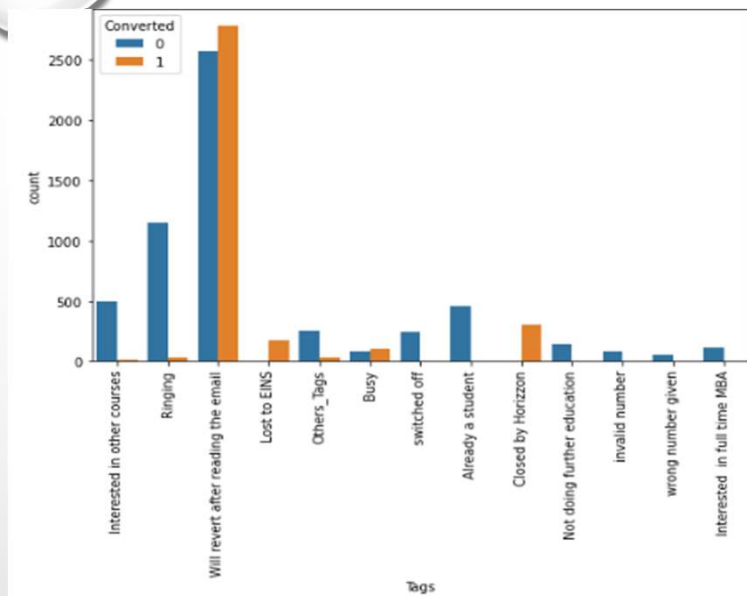
- **FOLLOWING OBSERVATION CAN BE DRAWN FROM THE ABOVE,**

1. THE COUNT OF EMAIL OPENED IS HIGHER THAN ANY OTHER ACTIVITY, HOWEVER THE RATE OF CONVERSION IS LESS THAN THE ONES NOT CONVERTED.
 2. IT IS OBSERVED THAT RATE OF CONVERSION FOR SMS_SENT IS THE HIGHEST AMONG ALL.
- THUS IT CAN BE CONCLUDED THAT MORE SMS NEEDS TO BE SENT AS THE CHANCES OF CONVERSION IS THE HIGHEST. ALSO, THE EMAILS WHICH ARE SENT SHOULD HAVE ALL THE NECESSARY INFORMATION SO THAT THE BASIC DOUBTS GETS CLARIFIED WHICH WOULD RESULT IN INCREASING THE CONVERSION RATE.



- **FROM THE ABOVE FOLLOWING OBSERVATIONS ARE MADE:**

1. FROM THE SPECIALIZATION, SINCE OTHERS HAVE BOTH NULL AND SELECT VALUE SO NO CONCLUSION CAN BE DRAWN.
 2. THE COUNT OF UNEMPLOYED IS MORE THAN ANY OTHER PARAMETER.
 3. IT IS OBSERVED THAT WORKING PROFESSIONAL HAS A HIGHER RATE OF CONVERSION.
- THUS IT CAN BE CONCLUDED THAT, FOCUS SHOULD BE TO INCREASE THE TOTAL COUNT OF APPLICATIONS UNDER WORKING PROFESSIONAL AS THE CONVERSION RATE IS HIGH. ALSO, SALES TEAM SHOULD SPECIFICALLY PITCH THE WORKING PROFESSIONAL FOR A HIGHER CAREER GROWTH, THEN THE CONVERSION RATE WILL INCREASE. ALSO, AMONG UNEMPLOYED, SALES TEAM SHOULD GIVE ATTENTION AND SHOULD TRY TO INCREASE THE CONVERSION RATE.



• **FROM THE ABOVE, FOLLOWING OBSERVATIONS ARE MADE:**

1. 'WILL REVERT AFTER READING THE EMAIL' HAS THE HIGHEST RATE OF CONVERSION.
 2. THE COUNT OF 'NOT SURE' IS HIGH, ALSO THE CONVERSION RATE IS LOW.
 3. 'MIGHT BE' CONVERSION RATE IS HIGH.
 4. 'HIGH IN RELEVANCE' CONVERSION RATE IS HIGH.
 5. 'CLOSED BY HORIZON' HAS A HIGH RATE OF CONVERSION COMPARED TO ITS COUNT.
- THUS IT CAN BE CONCLUDED THAT SALES TEAM SHOULD REACH OUT EVERY INDIVIDUAL THROUGH MAIL AND SHOULD WAIT FOR THEIR RESPONSES. ALSO, THOSE WHO PROPOSES TO COMMUNICATE AFTER READING THE MAIL, SHOULD BE IN TOUCH CONSTANTLY FOR INCREASING THE CONVERSION RATE. ALSO, THE TEAM SHOULD FOCUS ON INCREASING THE CONVERSION AND COUNT OF 'CLOSED BY HORIZON'. AMONG LEAD QUALITY FOCUS SHOULD BE TO INCREASE THE 'MIGHT BE' AND 'HIGH IN RELEVANCE' PARAMETER AS THE CHANCES OF CONVERSION IS HIGH. ALSO, THOSE WHO ARE 'NOT SURE' SHOULD BE PURSUED ON A REGULAR BASIS TO INCREASE THE CONVERSION RATE.

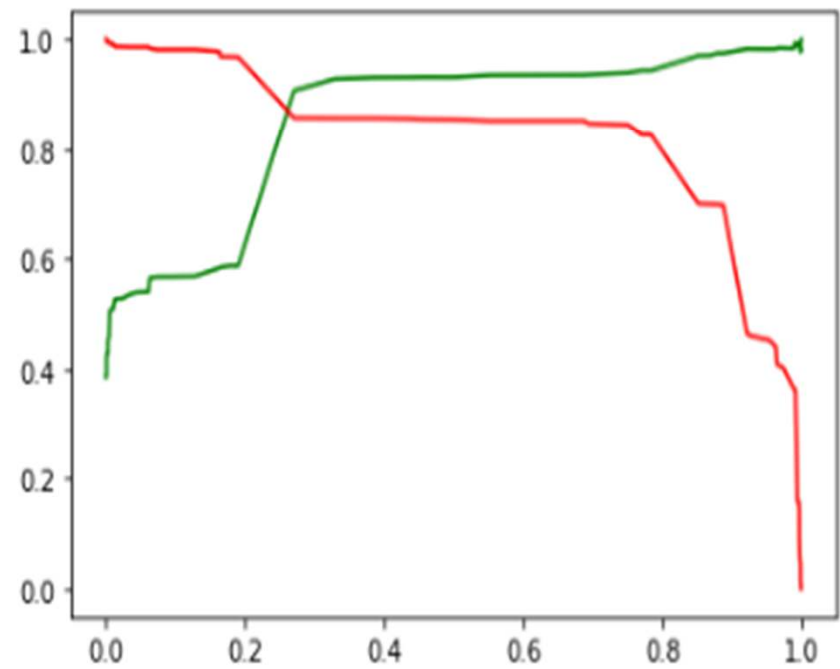
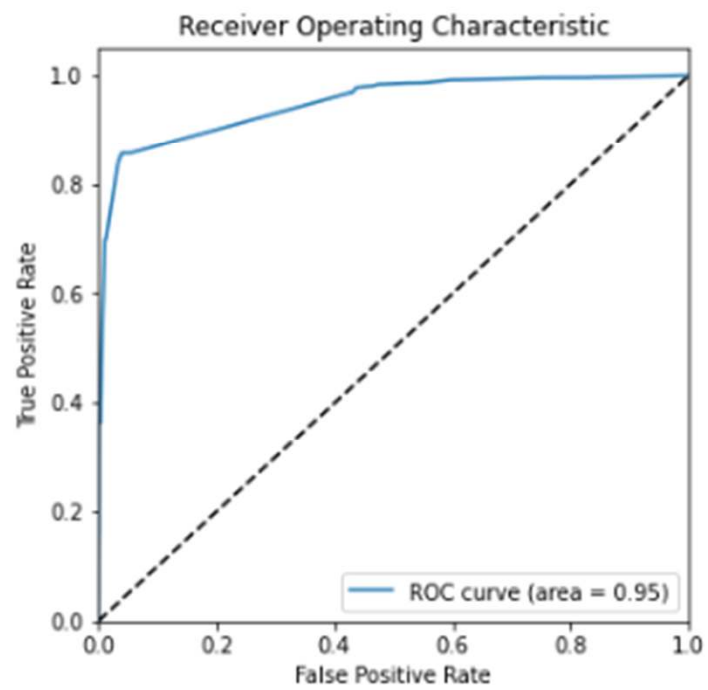
DATA CONVERSION

- NUMERICAL VALUES ARE NORMALISED
- DUMMY VARIABLES ARE CREATED.
- AND THE SAME IS THEN CONCATINATED WITH THE MAIN DATA SET AFTER REMOVING THE ORIGINAL COLUMNS.
- AFTER THAT TRAIN AND TEST SET IS CREATED AND MODEL IS BUILT ON TOP OF IT.

MODEL BUILDING

- RFE IS USED FOR FEATURE SELECTION.
- FINAL MODEL IS SELECTED HAVING P-VALUE LESS THAN 0.05 AND VIF LESS THAN 5.
- CONFUSIUON MATRIX IS CREATED AND ACCURACY, SENSITIVITY AND SPECIFICITY IS CALCULATED.
- ROC CURVE AND PRECISION RECALL TEST IS DONE TO FIND THE OPTIMUM CUT-OFF POINT WHICH IS FOUND OUT TO BE 0.27.

(GRAPHS PROVIDED BELOW)



FINAL OBSERVATION

- **TRAIN DATA:**

- ACCURACY: 91.11%
- SENSITIVITY: 85.73%
- SPECIFICITY: 94.49%

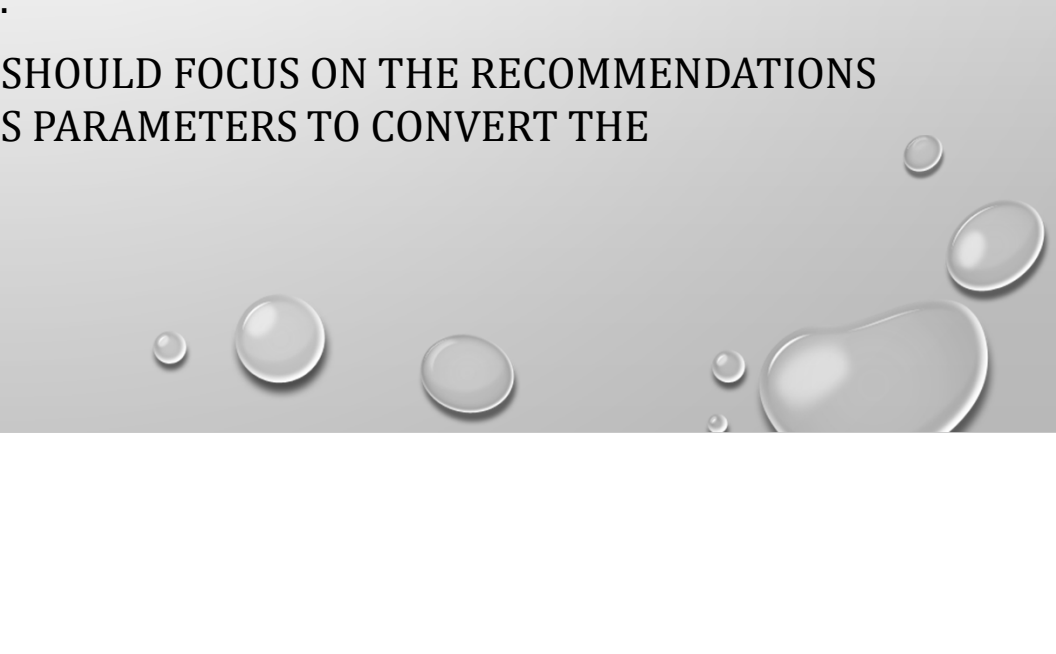
- **TEST DATA:**

- ACCURACY: 90.78%
- SENSITIVITY: 84.12%
- SPECIFICITY: 94.57%

THE MODEL PREDICTS THE CONVERSION RATE VERY WELL WITH A ACCURACY GREATER THAN 90%. SALES TEAM SHOULD FOCUS MORE ON LEAD SCIRE WHICH ARE HIGH TO INCREASE THE LEAD CONVERSION RATE



CONCLUSION

- HERE IN THE ABOVE LOGISTIC REGRESSION MODEL, THE OPTIMUM CUT-OFF IS CONSIDERED AS 0.27.
 - THIS SIGNIFIES THAT ANY LEAD WHICH SIGNIFIES A PROBABILITY GREATER THAN 0.27 CAN BE CONSIDERED AS A HOT LEAD.
 - THE ACCURACY OF THE MODEL IS MORE THAN 90%. THIS PREDICTION WOULD HELP THE CEO MAKE DECISIONS APPROPRIATELY.
 - APART FROM THE ABOVE THE SALES TEAM SHOULD FOCUS ON THE RECOMMENDATIONS MADE ABOVE AS PER BEHAVIOR OF VARIOUS PARAMETERS TO CONVERT THE CONVERSION RATE.
- 

THANK YOU