LEADS SCORING CASE STUDY

ABHAY DESAI RAJAGANESH KONERU TAPU RAJAK

PROBLEM STATEMENT – LOGISTIC REGRESSION TO PREDICT LEADS TO CONVERSION

- X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS.
- 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 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.

BUSINESS UNDERSTANDING

- X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS.
- 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.
- 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.
- 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.
- THE MODEL SHOULD BE ABLE TO ADJUST TO SOME MORE PROBLEMS PRESENTED BY THE COMPANY, IF THE COMPANY'S REQUIREMENT CHANGES IN THE FUTURE.

BUSINESS OBJECTIVES

GOAL:

- IDENTIFY PARAMETERS WHICH HELPS IN IDENTIFYING LEADS, WHICH CAN BE EASILY CONVERTED TO BUSINESS OPPORTUNITY SUCH AS :
  - ➢ LEAD ORIGIN,
  - ➢ TIME SPENT ON WEBSITE,
  - ➢ IS THERE ANY COMPETITION,
  - ➢ IS A REPEAT CUSTOMER, OR A NEW CUSTOMER,
  - ➢ TOTAL NUMBER OF VISITS TO WEBSITE AND PAGE VIEWS PER VISIT,
  - ➢ LAST ACTIVITY DONE, ETC.
- THIS WILL ENSURE THAT THE COMPANY UTILIZES ITS RESOURCES EFFECTIVELY TO TARGET "LOW HANGING FRUIT" AND INCREASE ITS BUSINESS AND REVENUE.
- IDENTIFICATION OF SUCH POTENTIAL LEADS USING LOGISTIC REGRESSION IS THE AIM OF THIS CASE STUDY

PYTHON NOTEBOOK FLOW

- Import required libraries such as pandas, numpy, matplotlib, seaborn, sklearn and statsmodels
- Load leads data having details of all leads got converted
- Verify leads data columns for NULL values in columns
    - If a column is having more than 45% null values, then drop that column
    - Convert all “Select” values to NULL, “Select” are those values which are optional and client hasn’t selected
    - For all other columns, updated NULL to Mode / Median of that column
- If a column is having unique values more than 95%, then drop that column
- Verify categorical parameters and update NULL / Missing values with highest ranking category
- Replace categorical parameters having very low representation and can be clubbed into a common category
- Drop all columns which are having highly imbalanced data
- Plot various graphs for different columns
- Plot co-relation for different columns
- Create dummy variables for categorical variables
- Split data into training and test data
- Scale training data by using StandardScaler
- Build Logistic Regression Model and calculate accuracy, sensitivity and specificity

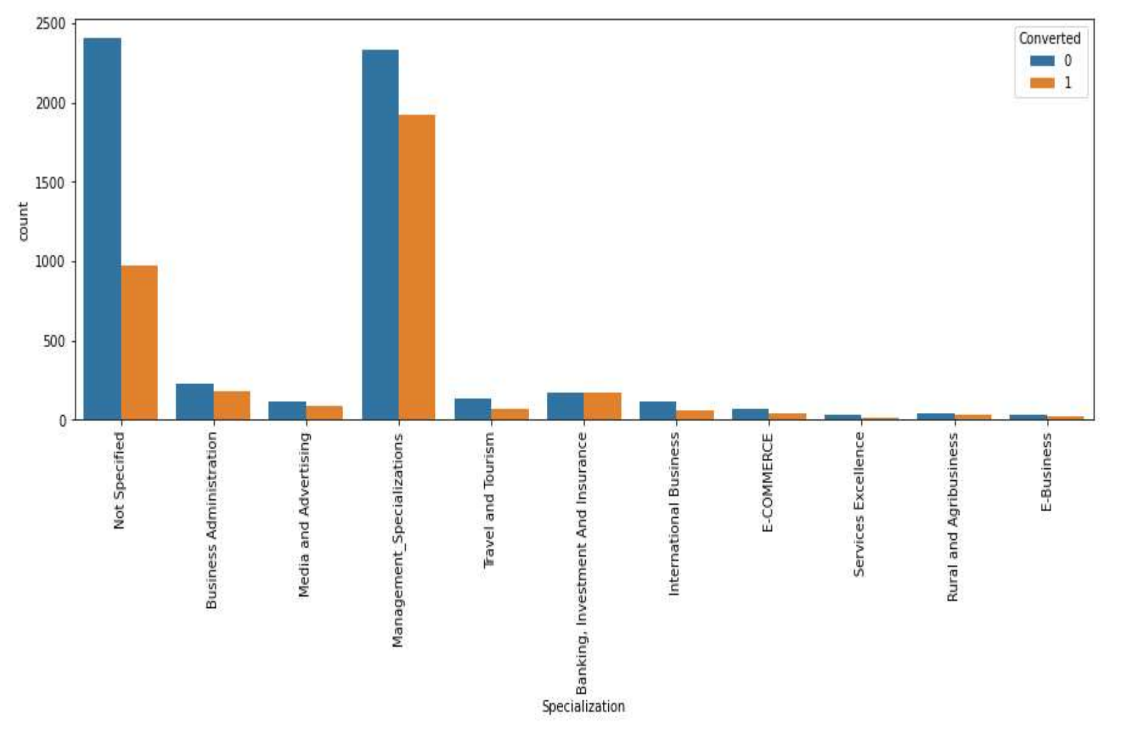
MODEL BUILDING

- SPLITTING THE DATA INTO TRAINING AND TESTING SETS
- THE FIRST BASIC STEP FOR REGRESSION IS PERFORMING A TRAIN-TEST SPLIT, WE HAVE CHOSEN 70:30 RATIO.
- USE RFE FOR FEATURE SELECTION
- RUNNING RFE WITH 15 VARIABLES AS OUTPUT
- BUILDING MODEL BY REMOVING THE VARIABLE WHOSE P- VALUE IS GREATER THAN 0.05 AND VIF VALUE IS GREATER THAN 5
- PREDICTIONS ON TEST DATA SET
- OVERALL ACCURACY 81%

DATA MANIPULATION

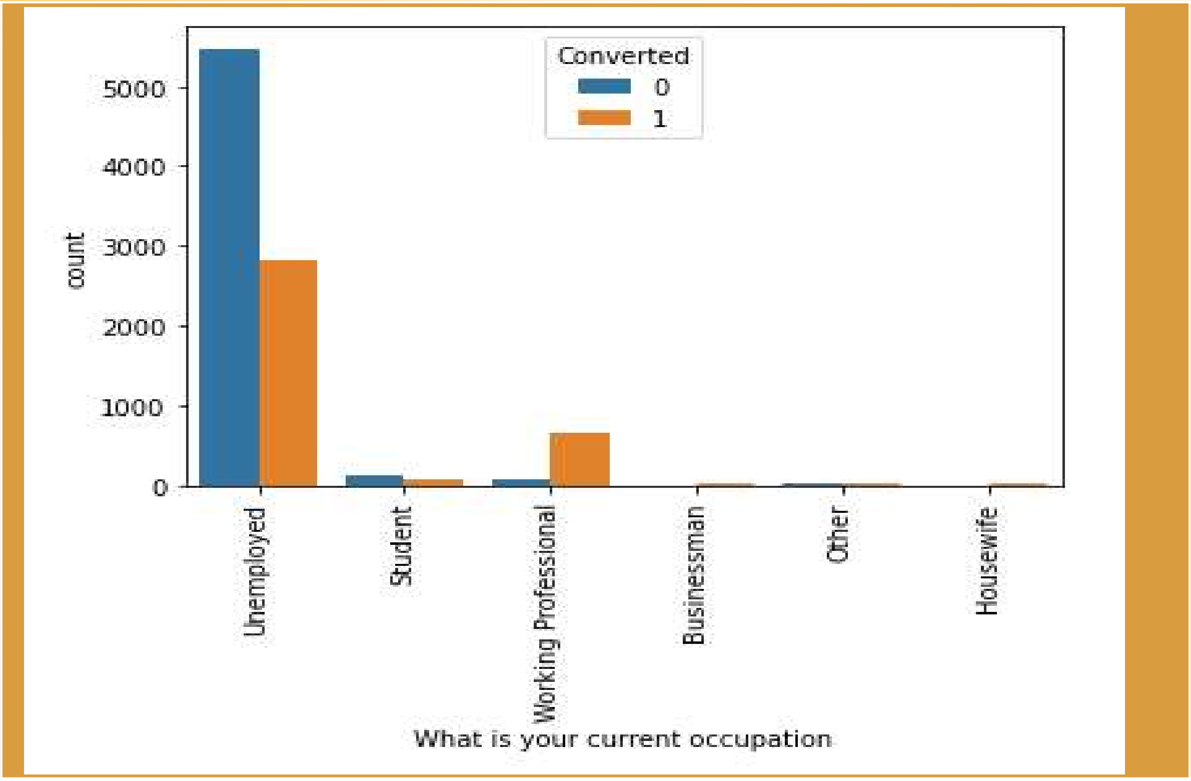
- TOTAL NUMBER OF COLUMNS = 37, TOTAL NUMBER OF ROWS = 9240.
- SINGLE VALUE FEATURES LIKE “MAGAZINE”, “RECEIVE MORE UPDATES ABOUT OUR COURSES”, “UPDATE ME ON SUPPLY CHAIN CONTENT”, “GET UPDATES ON DM CONTENT”, “I AGREE TO PAY THE AMOUNT THROUGH CHEQUE” ETC. HAVE BEEN DROPPED.
- REMOVING THE “PROSPECT ID” AND “LEAD NUMBER” WHICH ARE UNIQUE VALUES AND NOT NECESSARY FOR THE ANALYSIS.
- AFTER CHECKING FOR THE VALUE COUNTS FOR SOME OF THE OBJECT TYPE VARIABLES, WE FIND SOME OF THE FEATURES WHICH HAS NO ENOUGH VARIANCE, WHICH CAN BE DROPPED, THOSE FEATURES ARE: “DO NOT CALL”, “WHAT MATTERS MOST TO YOU IN CHOOSING COURSE”, “SEARCH”, “NEWSPAPER ARTICLE”, “X EDUCATION FORUMS”, “NEWSPAPER”, “DIGITAL ADVERTISEMENT” ETC.
- DROPPING THE COLUMNS HAVING MORE THAN 45% AS MISSING VALUE SUCH AS “HOW DID YOU HEAR ABOUT X EDUCATION” AND “LEAD PROFILE”.

GRAPHS & INFERENCES



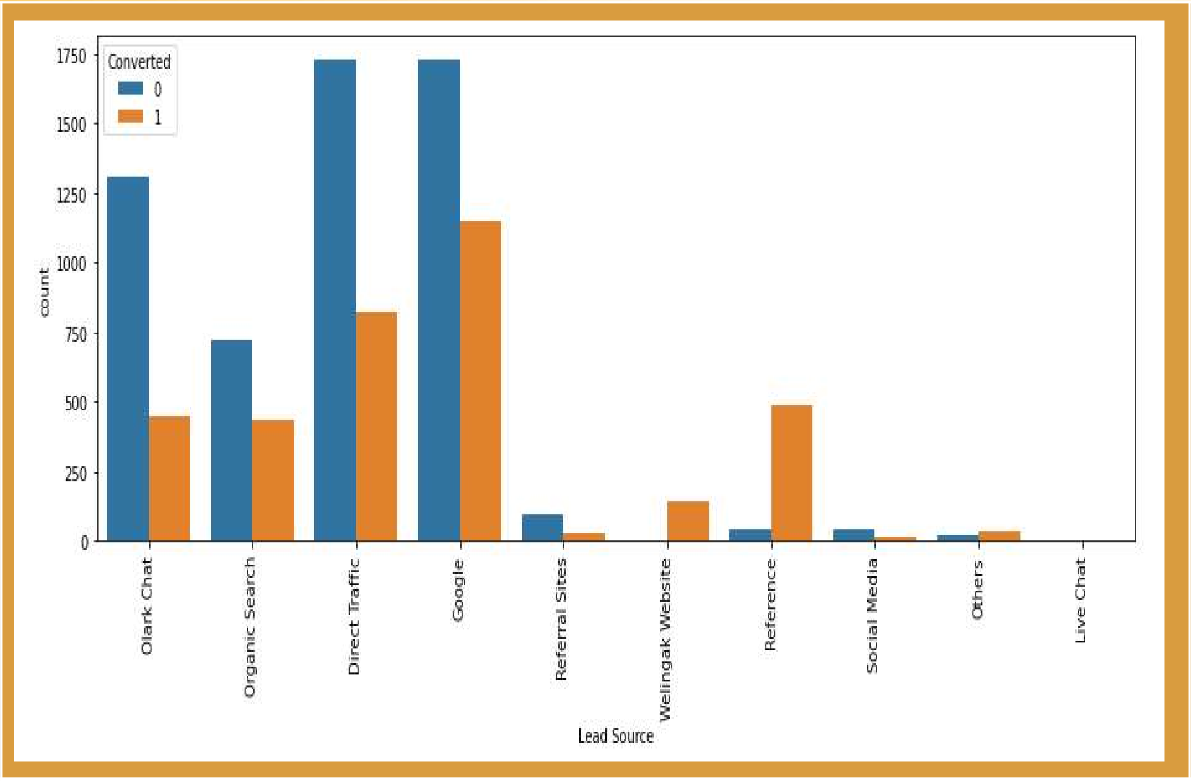
Observations-

The courses having maximum conversion are "Management Specializations".



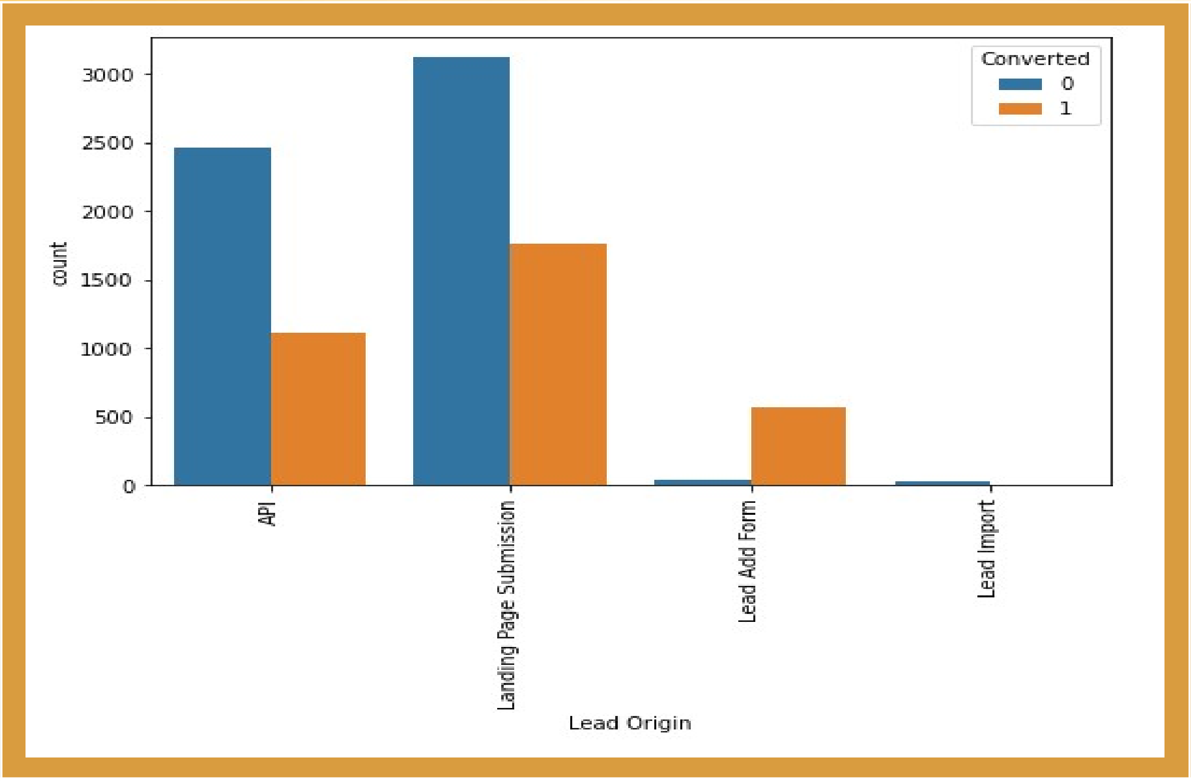
Observations-

The category of people having maximum interest and conversion in pursuing courses in absolute numbers are "Unemployed" followed by "Working Professional".



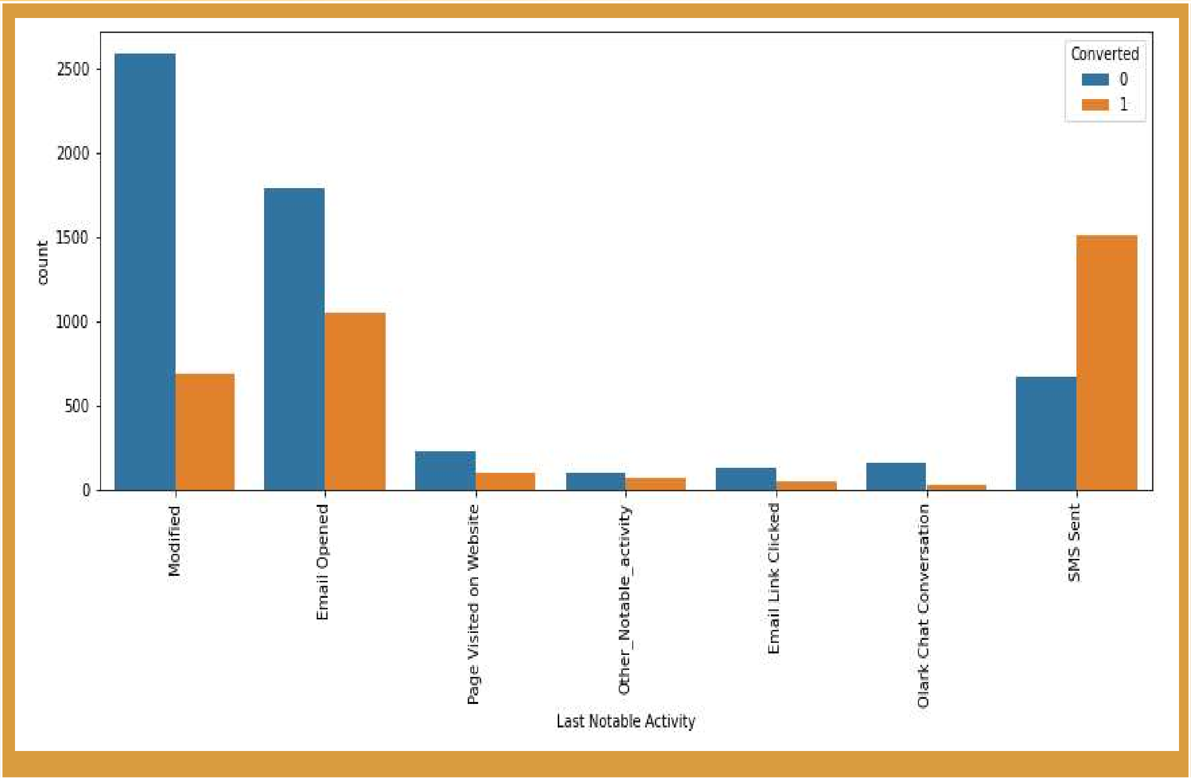
Observations-

Maximum number of leads are generated by Google and Direct traffic. Conversion Rate of reference leads and leads through welingak website is high. To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



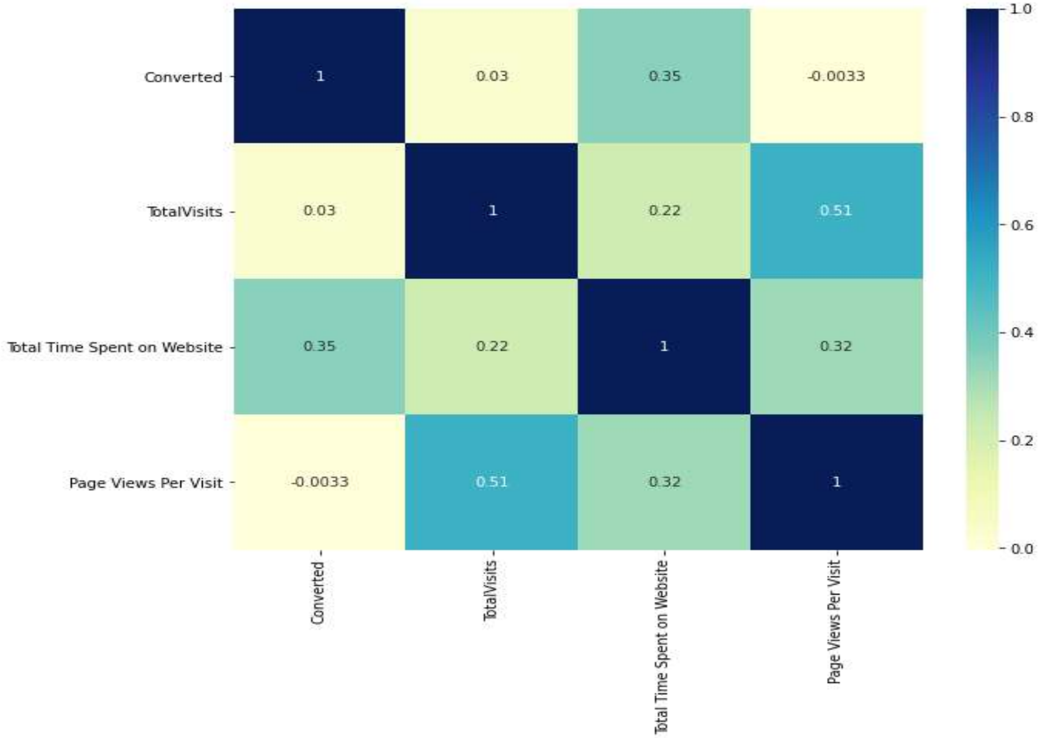
Observations-

API and Landing Page Submission bring higher number of leads as well as conversion. Lead Add Form has a very high conversion rate but count of leads are not very high. Lead Import and Quick Add Form get very few leads. In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



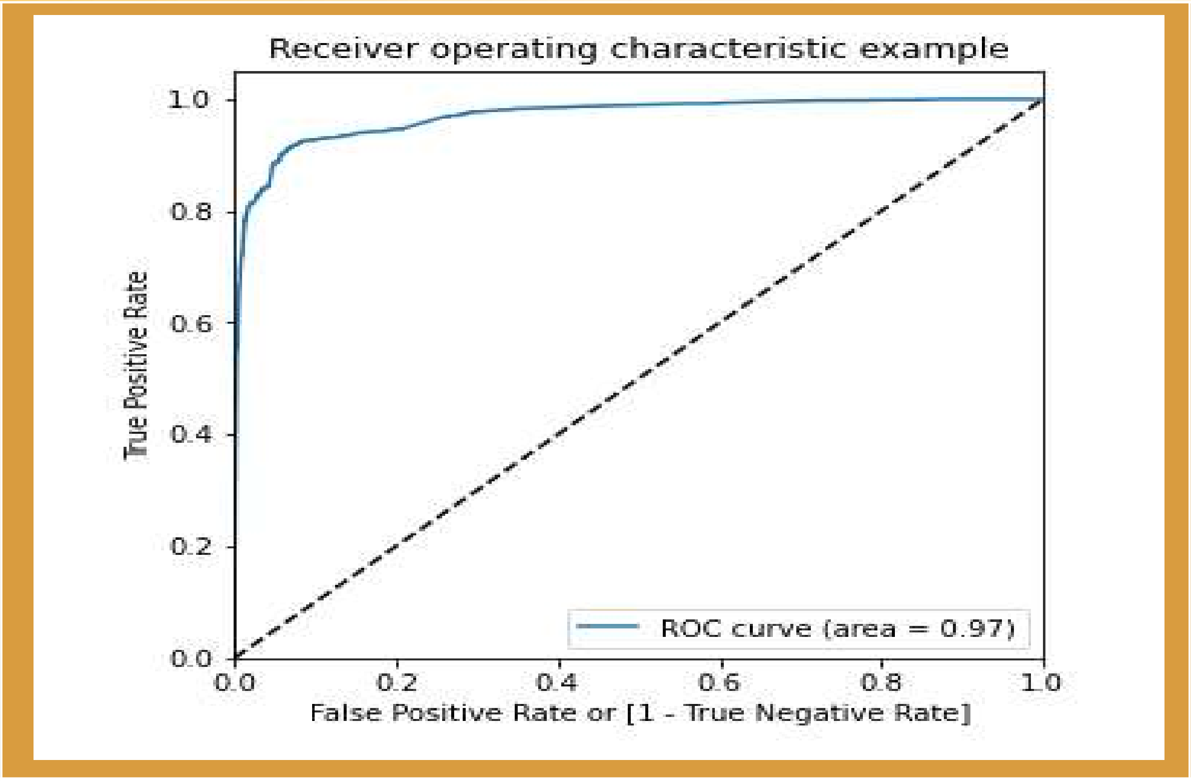
Observations-

The activity resulting in maximum conversion is "SMS Sent" followed by "Email Opened".



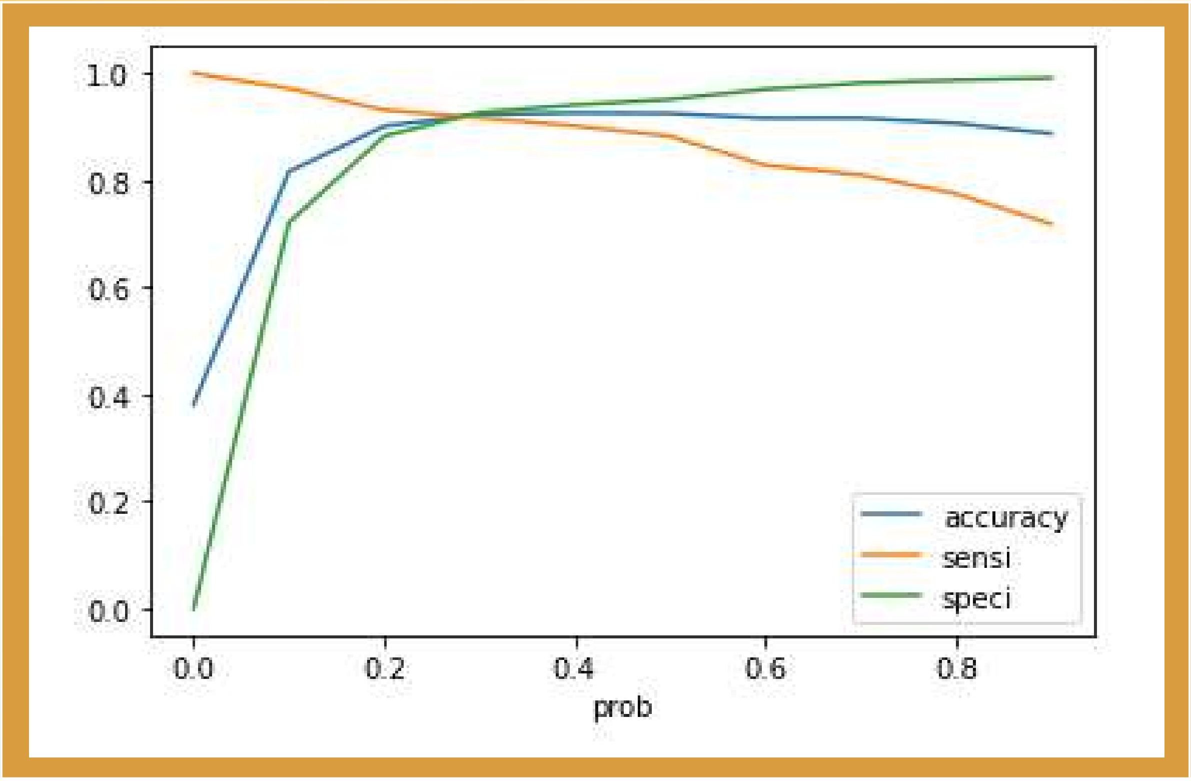
Observations-

The "Total Time Spent on Website" is directly proportional to conversion rate. The more time a prospect spends on website, more are the chances of him/her getting converted as a business opportunity.



Observations-

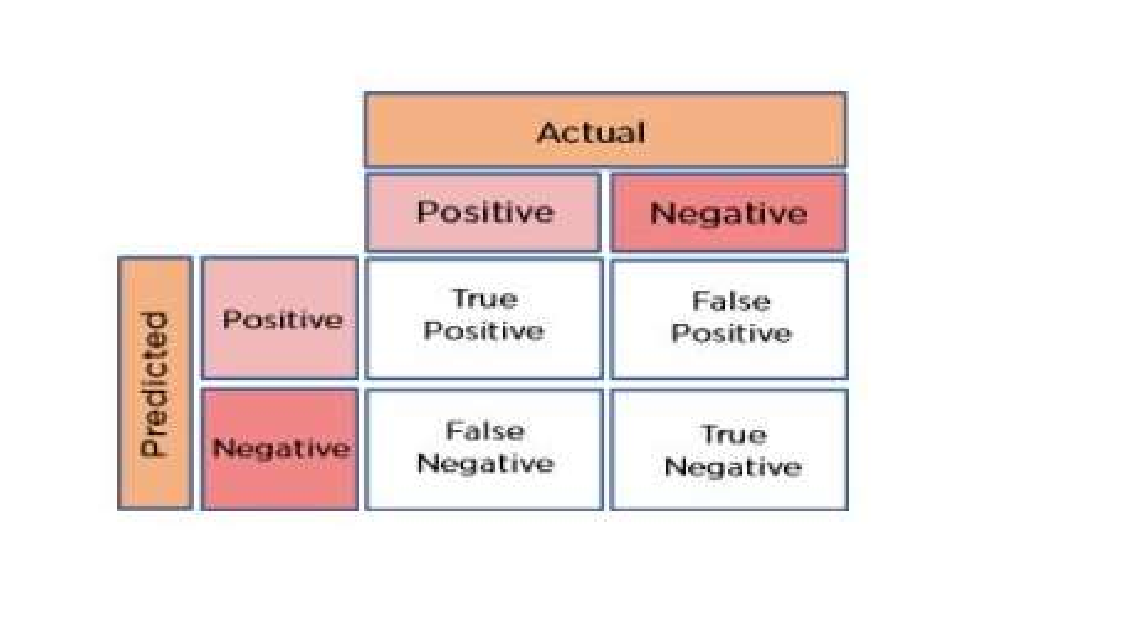
Ideally ROC curve should be a value close to 1. We are getting a good value of 0.97, which is an indication of a good model.



Observations-

Based on values of various accuracy, sensitivity and specificity, optimum cutoff point looks like 0.3.

Confusion Matrix



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Observations-

The confusion matrix for given dataset is as under:

		Actual	
		Positive	Negative
Predicted	Positive	2187	198
	Negative	285	3597

GENERAL OBSERVATIONS:

- THE TRAINING DATA GIVES ROC CURVE VALUE OF 0.97, ACCURACY OF 92.29%, SENSITIVITY OF 91.70% AND SPECIFICITY OF 92.66%.
- THE PRECISION SCORE IS 88.47% AND RECALL SCORE IS 91.70% FOR TRAINING DATA.
- THE TEST DATA IS GIVING ACCURACY OF 92.78%, SENSITIVITY OF 91.98% AND SPECIFICITY OF 93.26%, WHILE PRECISION SCORE IS 89.15% AND RECALL SCORE IS 91.98%.

ASSIGNMENT SUBJECTIVE QUESTIONS:

1. WHICH ARE THE TOP THREE VARIABLES IN YOUR MODEL WHICH CONTRIBUTE MOST TOWARDS THE PROBABILITY OF A LEAD GETTING CONVERTED?

   ANSWER:- THESE ARE THE TOP VARIABLES THAT CONTRIBUTE TOWARDS THE RESULT

   - TOTAL TIME SPEND ON THE WEBSITE.
   - TOTAL VISITS.
   - LEAD SOURCE WITH ELEMENTS GOOGLE.

2. WHAT ARE THE TOP 3 CATEGORICAL/DUMMY VARIABLES IN THE MODEL WHICH SHOULD BE FOCUSED THE MOST ON IN ORDER TO INCREASE THE PROBABILITY OF LEAD CONVERSION?

   ANSWER:- TOP 3 CATEGORICAL/DUMMY VARIABLES TO INCREASE PROBABILITY ARE:

   - LEAD SOURCE WITH ELEMENTS GOOGLE.
   - LEAD SOURCE WITH ELEMENTS DIRECT TRAFFIC.
   - LEAD SOURCE WITH ELEMENTS ORGANIC SEARCH.

ASSIGNMENT SUBJECTIVE QUESTIONS:

3. X EDUCATION HAS A PERIOD OF 2 MONTHS EVERY YEAR DURING WHICH THEY HIRE SOME INTERNS. THE SALES TEAM, IN PARTICULAR, HAS AROUND 10 INTERNS ALLOTTED TO THEM. SO DURING THIS PHASE, THEY WISH TO MAKE THE LEAD CONVERSION MORE AGGRESSIVE. SO THEY WANT ALMOST ALL OF THE POTENTIAL LEADS (I.E. THE CUSTOMERS WHO HAVE BEEN PREDICTED AS 1 BY THE MODEL) TO BE CONVERTED AND HENCE, WANT TO MAKE PHONE CALLS TO AS MUCH OF SUCH PEOPLE AS POSSIBLE. SUGGEST A GOOD STRATEGY THEY SHOULD EMPLOY AT THIS STAGE.

ANSWER:- PHONE CALLS MUST BE DONE TO PEOPLE, IF-

- THEY SPEND A LOT OF TIME IN THE WEBSITE AND THIS CAN BE DONE BY MAKING THE WEBSITE INTERESTING AND THUS BRINGING THEM BACK TO THE SITE.
- THEY ARE SEEN COMING BACK TO THE WEBSITE REPEATEDLY.
- THEIR LAST ACTIVITY IS THROUGH SMS OR THROUGH OLARK CHAT CONVERSATION.
- THEY ARE WORKING PROFESSIONALS.

ASSIGNMENT SUBJECTIVE QUESTIONS:

4. SIMILARLY, AT TIMES, THE COMPANY REACHES ITS TARGET FOR A QUARTER BEFORE THE DEADLINE. DURING THIS TIME, THE COMPANY WANTS THE SALES TEAM TO FOCUS ON SOME NEW WORK AS WELL. SO DURING THIS TIME, THE COMPANY'S AIM IS TO NOT MAKE PHONE CALLS UNLESS IT'S EXTREMELY NECESSARY, I.E. THEY WANT TO MINIMIZE THE RATE OF USELESS PHONE CALLS. SUGGEST A STRATEGY THEY SHOULD EMPLOY AT THIS STAGE.

ANSWER:- IN THIS CONDITION THEY NEED TO FOCUS MORE ON OTHER METHODS, LIKE AUTOMATED EMAILS AND SMS. THIS WAY CALLING WON'T BE REQUIRED UNLESS IT IS AN EMERGENCY. THE ABOVE STRATEGY CAN BE USED BUT WITH THE CUSTOMERS THAT HAVE A VERY HIGH CHANCE OF BUYING THE COURSE.

CONCLUSION

IT WAS FOUND THAT THE VARIABLES THAT MATTERED THE MOST IN THE POTENTIAL BUYERS ARE (IN DESCENDING ORDER) :

- THE TOTAL TIME SPEND ON THE WEBSITE.
- TOTAL NUMBER OF VISITS.
- WHEN THE LEAD SOURCE WAS:
  - GOOGLE
  - DIRECT TRAFFIC
  - ORGANIC SEARCH
  - WELINGAK WEBSITE
- WHEN THE LAST ACTIVITY WAS:
  - SMS
  - OLARK CHAT CONVERSATION
- WHEN THE LEAD ORIGIN IS LEAD ADD FORMAT.
- WHEN THEIR CURRENT OCCUPATION IS AS A WORKING PROFESSIONAL.

KEEPING THESE IN MIND THE X EDUCATION CAN FLOURISH AS THEY HAVE A VERY HIGH CHANCE TO GET ALMOST ALL THE POTENTIAL BUYERS TO CHANGE THEIR MIND AND BUY THEIR COURSES.

THANK
YOU