LEAD SCORING CASE STUDY

BY

SAGNIK CHAKRAVARTHY & PRASHANTH REDDY

PROBLEM STATEMENT

- AN EDUCATION COMPANY **“X EDUCATIONS”** SELLS ONLINE COURSES TO THE INDUSTRY PROFESSIONALS.
- THE COMPANY GENERATES LOT OF LEADS AND AS PER THE CONVERSIONS DONE IT SHOWS ONLY 30% IS THE CONVERSION RATE
- TO MAKE THE PROCESS MORE EFFICIENT, THE COMPANY WISHES TO IDENTIFY THE MOST POTENTIAL LEADS KNOWN AS **“HOT LEADS”**
- IF THEY COULD 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.

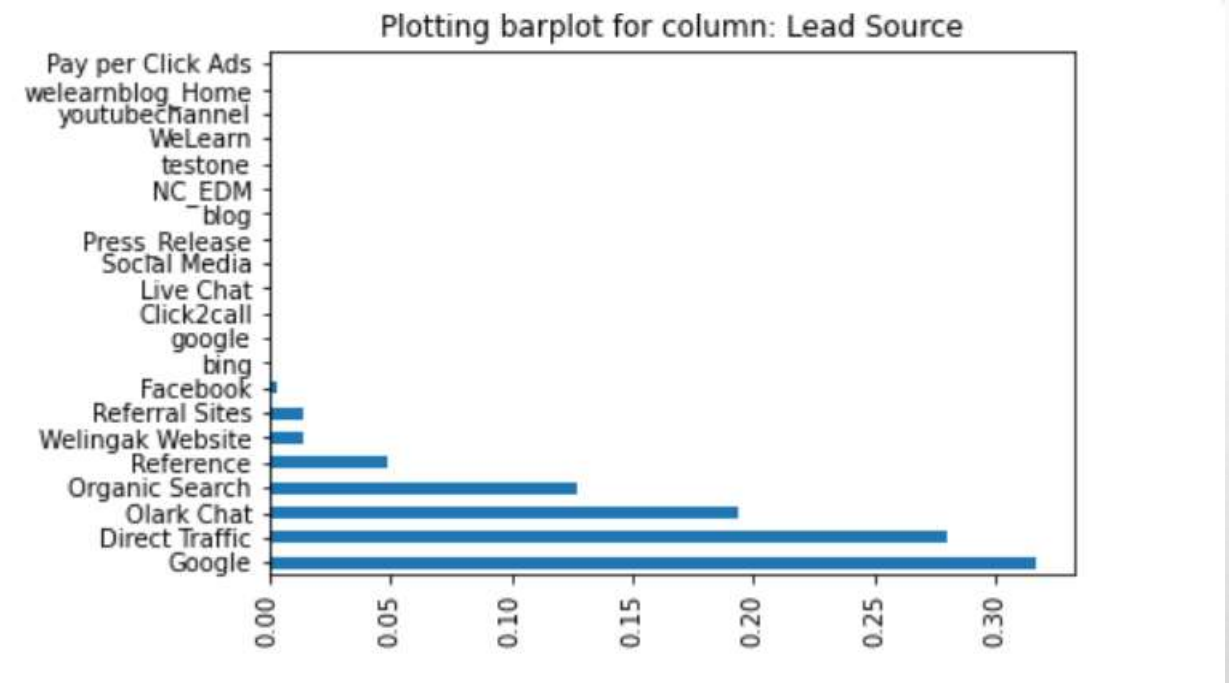
BUSINESS OBJECTIVES

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

TECHNICAL APPROACH

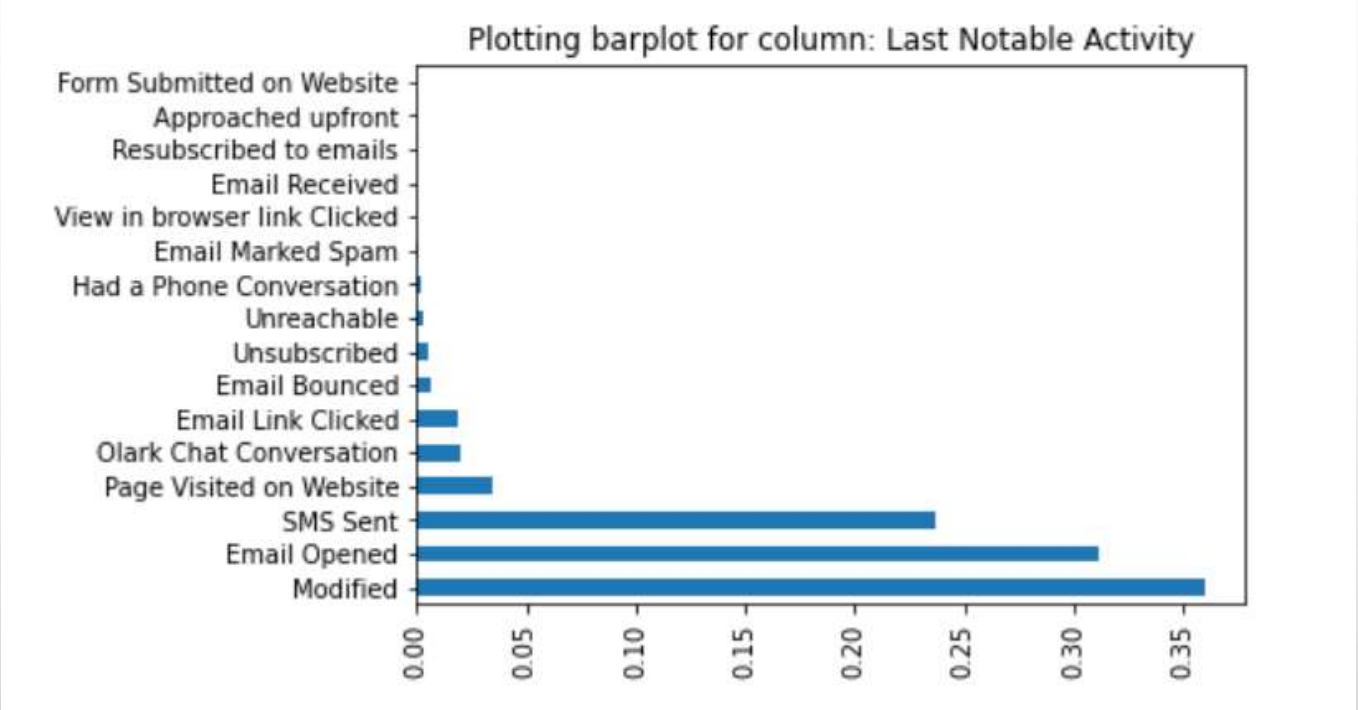
- **DATA CLEANING AND DATA MANIPULATION.**
  1. CHECK AND HANDLE DUPLICATE DATA.
  2. CHECK AND HANDLE NA VALUES AND MISSING VALUES.
  3. DROP COLUMNS, IF IT CONTAINS LARGE AMOUNT OF MISSING VALUES AND NOT USEFUL FOR THE ANALYSIS.
  4. IMPUTATION OF THE VALUES, IF NECESSARY.
  5. CHECK AND HANDLE OUTLIERS IN DATA.
- **EDA**
- **FEATURE SCALING & DUMMY VARIABLES AND ENCODING OF THE DATA.**
- **CLASSIFICATION TECHNIQUE:** LOGISTIC REGRESSION USED FOR THE MODEL MAKING AND PREDICTION.
- **VALIDATION OF THE MODEL.**
- **MODEL PRESENTATION.**
- **CONCLUSIONS AND RECOMMENDATIONS.**

EDA-GRAPH



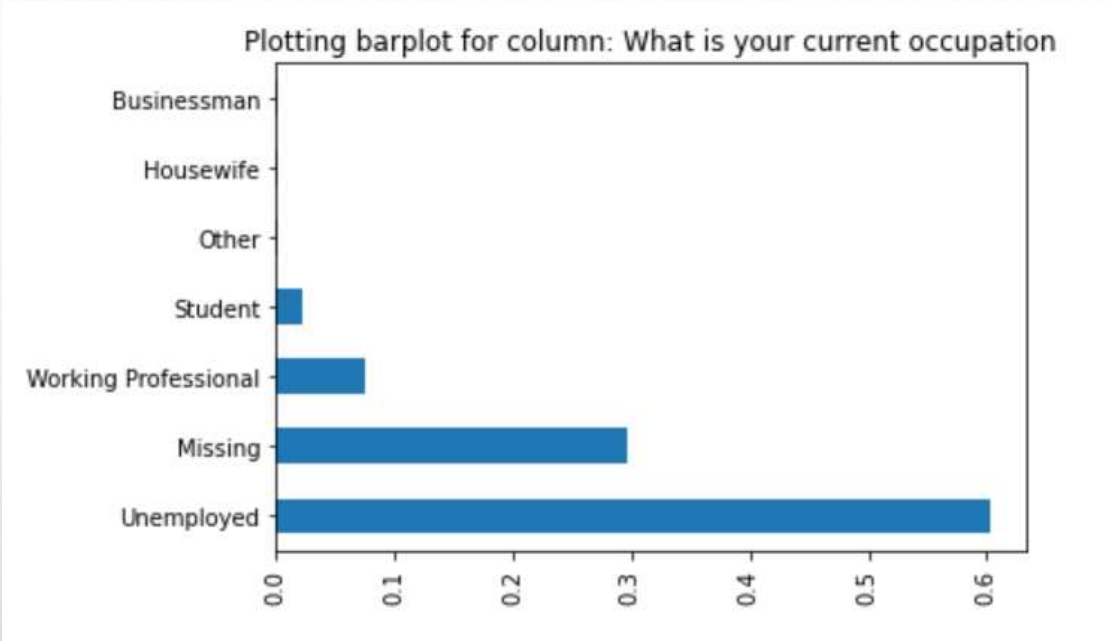
- GOOGLE AND DIRECT TRAFFIC ARE THE BIGGEST LEAD SOURCES

EDA-GRAPH



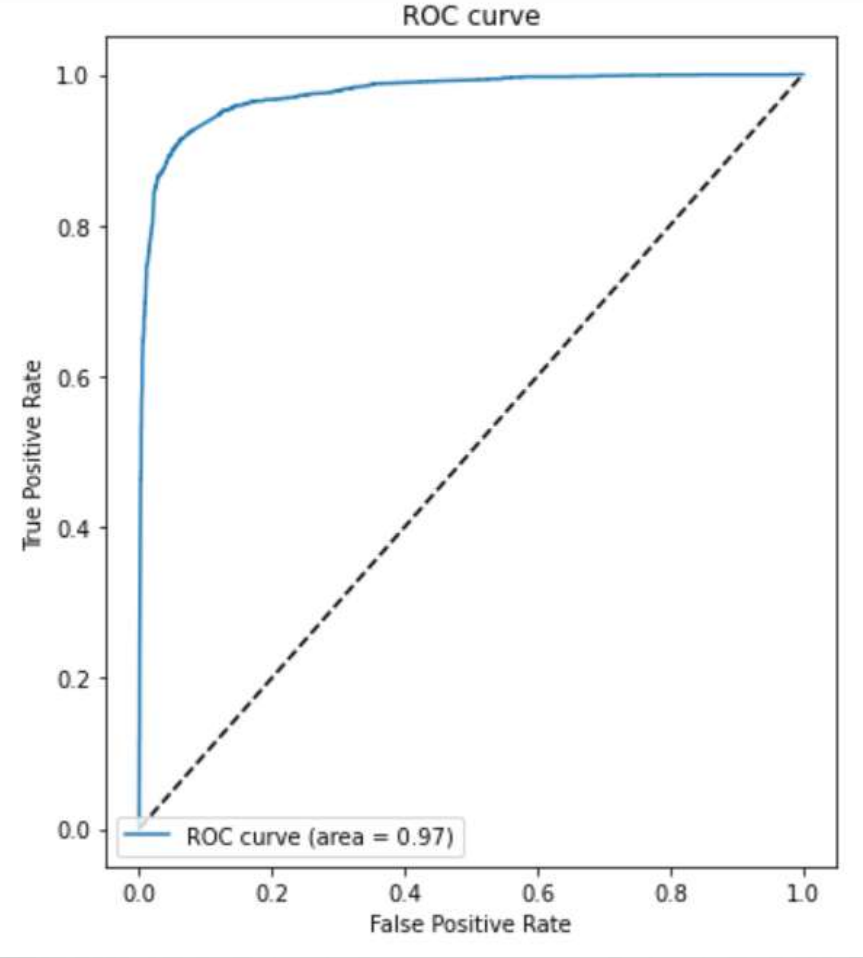
- THE LAST NOTABLE ACTIVITIES OF STUDENTS/LEADS ARE 'MODIFIED', 'EMAIL OPENED' AND 'SMS SENT'

EDA-GRAPH



- MAJORITY OF THE LEADS ARE CURRENTLY UNEMPLOYED

ROC CURVE


ROC curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.97)

CONCLUSION

- MODEL PERFORMANCE
  - TRAIN (ACCURACY, SENSITIVITY AND SPECIFICITY)
  - TEST (ACCURACY, SENSITIVITY AND SPECIFICITY)

- CONFUSION MATRIX
  - TRAIN
  - TEST

MODEL PERFORMANCE

- SPLITTING THE DATA INTO TRAINING AND TESTING SETS
- THE FIRST BASIC STEP FOR REGRESSION IS PERFORMING A TRAIN-TEST SPLIT, WE HAVE CHOSEN 75:25 RATIO.
- USE RFE FOR FEATURE SELECTION AND RUNNING RFE WITH 30 VARIABLES AS OUTPUT. LATER THE REST OF THE VARIABLES WERE REMOVED MANUALLY DEPENDING ON THE VIF VALUES AND P-VALUE WHERE IT ENDED UP WITH 17 VARIABLES.

MODEL EVALUATION

- OVERALL ACCURACY
  - TRAIN : 93.15%
  - TEST : 91.75%
- SENSITIVITY
  - TRAIN : 88.49%
  - TEST : 85.99%
- SPECIFICITY
  - TRAIN : 96.04%
  - TEST : 95.07%
- WE ARE GETTING A PRECISION OF AROUND 90.93% OR 91% ON THE TEST SET.

CONFUSION MATRIX

- 4032 OF THEM ARE TO BE TRUE NEGATIVES FOR TRAIN AND 1370 OF THEM ARE TO BE TRUE NEGATIVES FOR TEST
- 166 OF THEM ARE TO BE FALSE POSITIVES FOR TRAIN AND 71 OF THEM ARE TO BE FALSE POSITIVES FOR TEST
- 300 OF THEM ARE TO BE FALSE NEGATIVES FOR TRAIN AND 116 OF THEM ARE TO BE FALSE NEGATIVES FOR TEST
- 2307 OF THEM ARE TO BE TRUE POSITIVES FOR TRAIN AND 712 OF THEM ARE TO BE TRUE POSITIVES FOR TEST

RECOMMENDATIONS

- TO GET CONNECTED TO HOT LEADS, ONE SHOULD FOCUS ON TAGS CLOSED BY HORIZON, AS THESE SET OF LEADS ARE WITH INTENT AND ARE LOOKING FORWARD TO JOIN THE PARTICULAR PROGRAM.
- FOCUS SHOULD BE ON UNEMPLOYED AS CONVERSION RATE IS HIGHEST THERE.
- GOOGLE AND DIRECT TRAFFIC ARE THE BIGGEST LEAD SOURCE AND ARE VERY GOOD FOR BUSINESS. BUSINESS SHOULD INCREASE ITS ENGAGEMENT WITH THEM. EMAILS SEEMS TO BE MORE EFFECTIVE THAN SMS
- MAJORITY OF THE LEADS ARE LOOKING FOR A COURSE FOR BETTER CAREER PROSPECTS AND HENCE WEBSITE SHOULD BE MADE MORE ENGAGING ON THE SAME FACTOR.

THANK YOU