LEAD SCORE CASE STUDY

- GROUP MEMBER
- Supankaj
- Rohit

Problem Statement

- 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.

SOLUTION METHODOLOGY

- DATA CLEANING AND MANIPULATION.

- 1. CHECK AND HANDLE DUPLICATE DATA.
- 2. CHECK AND REMOVE THE COLUMNS WITH MORE THAN 35 % NULL VALUES.
- 3. DROP COLUMNS, IF IT CONTAINS LARGE NUMBER OF MISSING VALUES AND NOT USEFUL FOR ANALYSIS.

SOLUTION METHODOLOGY

- EDA
- FEATURE SCALING AND DUMMY VARIABLE AND ENCODINF OF DATA.
- LOGISTIC REGRESSION USED FOR MODEL MAKING AND PREDICTION.
- VALIDATION OF MODEL.
- MODEL PRESENTATION .
- CONCLUSIONS AND RECOMMENDATIONS.

DATA MANIPULATION

TOTAL NUMBER OF ROWS=6373, TOTAL NUMBER OF COLUMNS=12 (after data cleaning)

FINDING THE NUMBER OF MISSING VALUES AND DROPPING COLUMNS HAVING MORE THAN 35% AS MISSING VALUES SUCH AS “LEAD QUALITY”

LET'S CHECK FOR THE COLUMN WHICH ARE OF NO USE LIKE COUNTRY AND CITY.

DATA CONVERSION

NUMERICAL VALUES ARE NORMALISED

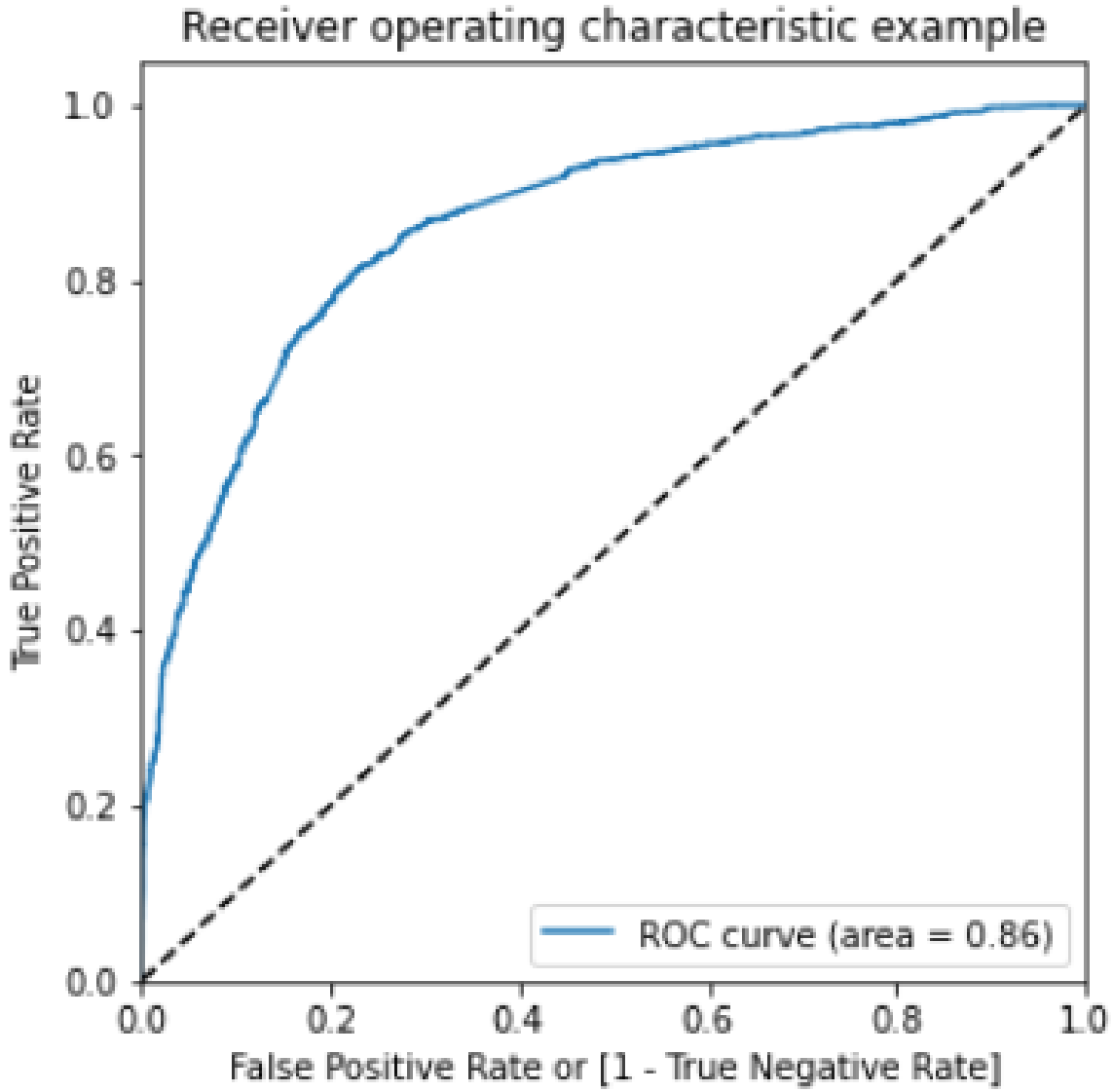
DUMMY VARIABLES ARE CREATED FOR OBJECT TYPE VARIABLES.

CONCATENATING DUMMY AND ORIGINAL VARIABLES.

MODEL BUILDING

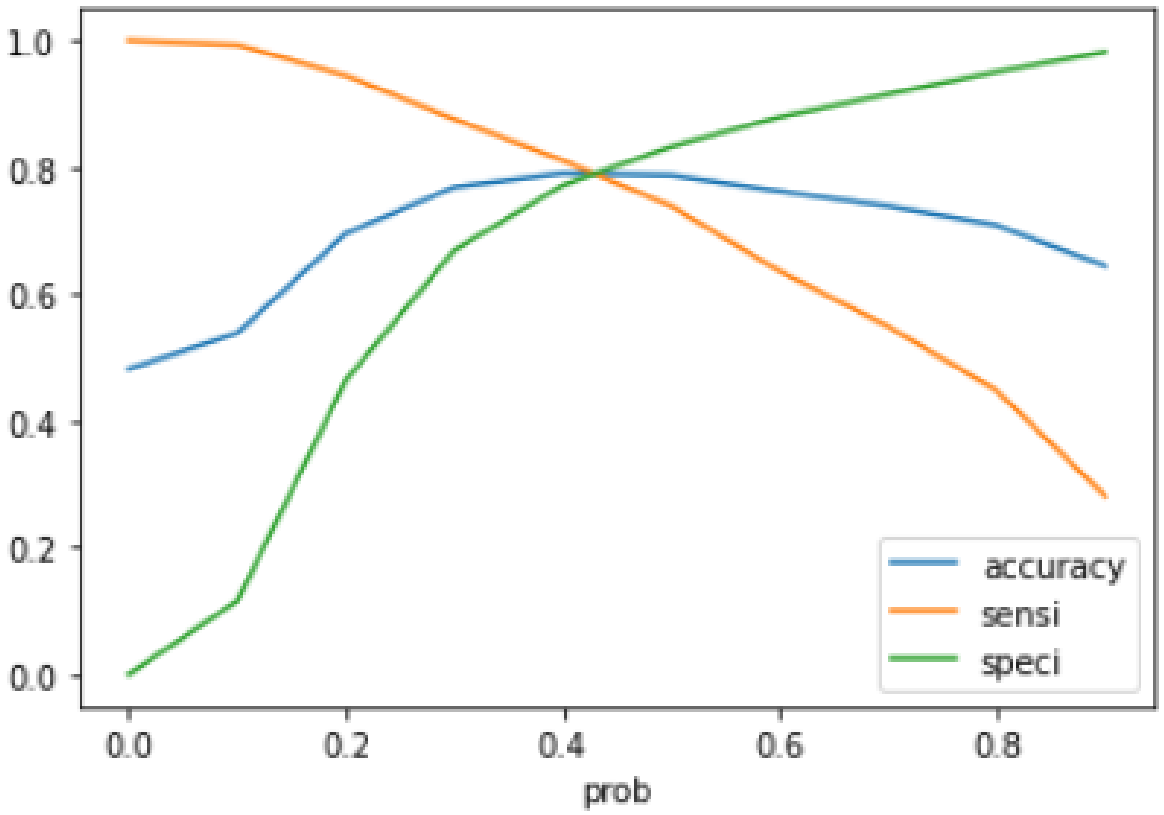
- UNIVARIATE DATA ANALYSIS: VALUE COUNT ETC.
- BIVARIATE DATA ANALYSIS: CORRELATION COEFFICIENTS.
- SPLITTING THE DATA INTO TRAINING AND TESTING SET
- THE FIRST STEP FOR REGRESSION IS PERFORMING A TRAIN-TEST SPLIT, WE HAVE CHOSEN 70:30 RATIO.
- USE OF RFE FOR FEATURE SELECTION
- RUNNING RFE WITH 15 VARIABLES.
- BUILDING MODEL BY REMOVING THE VARIBALES WHOSE P-VALUE IS GREATER THAN 0.05 AND VIF VALUE IS GREATER THAN 5.
- PREDICTION ON TEST DATA SET
- OVERALL ACCURACY 79%

ROC CURVE


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

FINDING THE OPTIMAL CUT-OFF POINT

- 1. PROBABILITY WHERE WE GET BALANCED SENSITIVITY AND SPECIFICITY.
- FROM THE SECOND GRAPH IT IS VISIBLE THAT THE OPTIMAL CUT OFF IS AT 0.41



CONCLUSION

- It was found that the variables that mattered the most in the potential buyers are :

- Total Visits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Last Notable Activity_Unreachable
- Last Activity_Had a Phone Conversation
- What is your current occupation_Unemployed

- 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.