LEAD SCORING CASE STUDY

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 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%.
- THE COMPANY REQUIRES US TO BUILD A MODEL WHEREIN WE NEED TO ASSIGN A LEAD SCORE TO EACH OF THE LEADS SUCH THAT THE CUSTOMERS WITH HIGHER LEAD SCORE HAVE A HIGHER CONVERSION CHANCE AND THE CUSTOMERS WITH LOWER LEAD SCORE HAVE A LOWER CONVERSION CHANCE. THE CEO, IN PARTICULAR, HAS GIVEN A BALLPARK OF THE TARGET LEAD CONVERSION RATE TO BE AROUND 80%.

APPROACH FOLLOWED IN STEPS

- DATA CLEANING
- EXPLORATORY DATA ANALYSIS
- SCALING AND DUMMY VARIABLE CREATION
- TEST-TRAIN SPLIT
- MODEL BUILDING
- MODEL EVALUATION
- PRECISION AND RECALL
- PREDICTION

DATA CLEANING AND PREPARATION

THE INITIAL ANALYSIS OF DATA-SET IS DONE BY TESTING ITS INFO, DTYPE AND SHAPE

- FEW COLUMNS WERE DROPPED AS THEY ARE HAVING HIGH MISSING VALUES AND THOSE WERE SALES TEAM GENERATED COLUMNS.
- COLUMNS WITH SINGLE CATEGORICAL VARIABLES ARE DROPPED.
- SKEWNESS OF THE COLUMNS ARE TESTED AND THEN THE COLUMNS IN WHICH THE SKEWNESS WAS FOUND ARE DROPPED.
- THE COLUMN WITH THE VALUES ‘SELECT’ ARE REPLACED AS NULL VALUES.
- THE COLUMNS WITH MISSING VALUES LESS THAN 30% ARE TREATED BY IMPUTING THEM WITH MODE OR REPLACING THEM AS NOT MENTIONED OR NOT PROVIDED.
- FEW OF THE CATEGORICAL VARIABLES ARE HAVING A LOT OF VARIABLES SO THEY ARE TREATED BY CATEGORIZING THEM INTO A GENERAL CATEGORY.

MODEL BUILDING

- FIRST THE MODEL IS BUILT AND THEN THE VARIABLES ARE SELECTED BASED ON THE RECURSIVE FEATURE ELIMINATION METHOD (RFE)
- THE RFE IS RUN WITH THE OUTPUT VARIABLES AS 20

```
column = X_train.columns[rfe.support_]
column
```

```
Index(['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit',
       'Lead Origin_Lead Add Form', 'Lead Origin_Lead Import',
       'Lead Source_Google', 'Lead Source_Olark Chat',
       'Lead Source_Organic Search', 'Country_Not mentioned',
       'Specialization_Finance Management',
       'Specialization_Hospitality Management',
       'Specialization_International Business',
       'Specialization_Retail Management',
       'Specialization_Rural and Agribusiness', 'occupation_Housewife',
       'occupation_Other', 'occupation_Working Professional',
       'course_Flexibility & Convenience', 'course_Not Provided',
       'course_Other'],
      dtype='object')
```

```
X_train.columns[~rfe.support_]
```

```
Index(['Lead Origin_Landing Page Submission', 'Lead Origin_Quick Add Form',
       'Lead Source_Social Media', 'Country_Other Countries',
       'Specialization_Business Administration', 'Specialization_E-Business',
       'Specialization_E-COMMERCE', 'Specialization_Healthcare Management',
       'Specialization_Human Resource Management',
       'Specialization_IT Projects Management',
       'Specialization_Marketing Management',
       'Specialization_Media and Advertising',
       'Specialization_Operations Management',
       'Specialization_Services Excellence',
       'Specialization_Supply Chain Management',
       'Specialization_Travel and Tourism', 'occupation_Student',
       'occupation_Unemployed', 'Interview_Yes'],
      dtype='object')
```

MODEL BUILDING

- GENERALIZED LINEAR MODELS FROM STATSMODELS IS USED TO BUILD THE LOGISTIC REGRESSION MODEL.
- THE MODEL IS BUILT INITIALLY WITH THE 20 VARIABLES SELECTED BY RFE.
- UNWANTED FEATURES ARE DROPPED SERIALLY AFTER CHECKING P VALUES (< 0.5) AND VIF (< 5) AND MODEL IS BUILT MULTIPLE TIMES.
- THE FINAL MODEL IS BUILT WITH 10 VARIABLES

	Features	VIF
6	Specialization_Finance Management	2.31
4	Lead Source_Olark Chat	1.80
1	Total Time Spent on Website	1.73
3	Lead Source_Google	1.73
9	course_Not Provided	1.48
0	TotalVisits	1.46
5	Lead Source_Organic Search	1.29
2	Lead Origin_Lead Add Form	1.20
8	occupation_Working Professional	1.17
7	Specialization_Hospitality Management	1.02

ANALYSIS BASED ON THE LOGISTICS REGRESSION MODEL

- THE ANALYSIS OF THE MODEL WAS DONE WITH 10 VARIABLES

- THE ANALYSIS WAS DONE BASED ON THE FOLLOWING PERSPECTIVES
    - BUSINESS PERSPECTIVE
    - TECHNICAL PERSPECTIVE.

BUSINESS ASPECT OF THE MODEL

TOTALVISITS AND TOTAL TIME SPENT ON WEBSITE :

IF THE LEAD IS REALLY INTERESTED , THEN THE CANDIDATE WILL VISIT THE WEBSITE AGAIN AND AGAIN AND ALSO SPENDS LOT OF TIME EXPLORING THE KEY SELLING POINTS OF THE COURSE. SO THESE TWO FEATURES SIGNIFICANTLY INFLUENCE IF THE CANDIDATE IS MOST LIKELY TO GET CONVERTED OR NOT. HIGHER THE VISITS AND TIME SPENT ON THE WEBSITES , HIGHER IS CONVERSION CHANCE.

OCCUPATION_WORKING PROFESSIONAL :

MOST OF THE PROFESSIONALS ARE LOOKING FOR UPGRADING THEMSELVES TO FIT INTO THE NEW DIGITAL WAVE.THEY ARE ALSO LOOKING FOR DISTANCE LEARNING AS THIS IS CONVENIENT FOR THEM TO LEARN ANY TIME.

LEAD SOURCE_GOOGLEAND LEAD SOURCE OLARK:

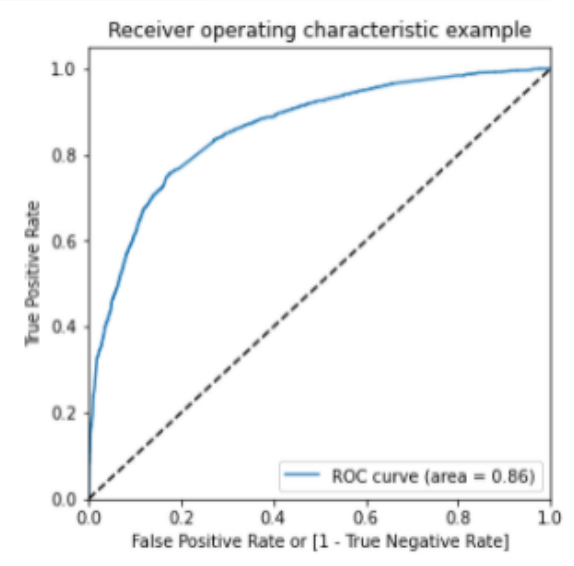
NOWADAYS , PEOPLE DO LOT OF RESEARCH BEFORE BUYING ANYTHING AND THE SAME IS APPLICABLE FOR ONLINE LEARNING. IF THE LEAD IS ACQUIRED BASED ON HIS EXPLORATIONS IN GOOGLE , HE / SHE IS MOST LIKELY HAS ALREADY MADE MIND TO ENROLL FOR THIS COURSE. HENCE, THERE IS MORE CHANCE THAT HE / SHE GETS CONVERTED. SIMILARLY , IF THE CANDIDATE HAS USED THE CHAT HELP , HE IS MORE LIKEY INTERESTED THAN THE OTHERS TO TAKE THE COURSE.

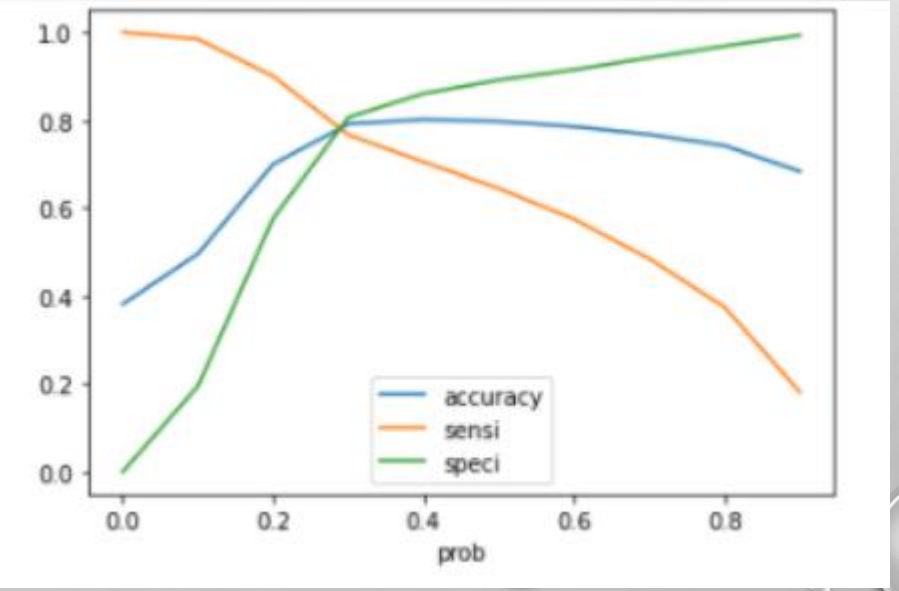
TECHNICAL ASPECT OF THE MODEL

- AFTER SELECTING THE 20 FEATURES FOR THE MODEL ON THE FIRST ITERATION , THE FINAL NUMBER OF FEATURES WAS REDUCED TO 10 , BASED ON THEIR VIF SCORES AMONG OTHER FEATURES. THIS WAS DONE BY ITERATIVELY REDUCING THE FEATURES TO A POINT WHERE THE VIF'S WAS WELL BELOW 2.
- BASED ON THE VARIOUS METRICS SUCH AS ACCURACY , SPECIFICITY AND SENSITIVITY , WE CHOOSE A CUT OFF PROBABILITY OF 0.33. SO ANYTHING ABOVE 33% CHANCE OF CONVERSION IS TREATED AS HOT LEADS AND MORE EFFORT WOULD BE PUT ON THESE LEADS.

MODEL EVALUATION

- AFTER BUILDING THE MODEL THE CONVERSION PROBABILITY IS PREDICTED AND THEN A GENERAL CUT-OFF OF 0.5 IS TAKEN TO FIND THE FINAL PREDICTIONS IF THE LEAD IS CONVERTED OR NOT, AND THEN THE CONFUSION MATRIX IS CREATED AND THEN THE ACCURACY, SENSITIVITY AND SPECIFICITY ARE CHECKED.
- THEN THE OPTIMAL CUT-OFF IS FOUND BY USING THE ROC CURVE
- AFTER THE ROC CURVE IS DRAWN THE OPTIMAL CUT-OFF IS TAKEN AS 0.33





MODEL EVALUATION ON TEST SET

- THE CONVERSION PROBABILITY IS PREDICTED IN THE TEST SET AND THEN THE FINAL PREDICTION IS GENERATED AS 0 AND 1 WITH THE CUT-OFF TAKEN AS 0.33
- THEN THE LEAD SCORE IS GENERATED BY MULTIPLYING THE CONVERSION PROBABILITY BY 100.
- ANY LEAD SCORE GREATER THAN 33 CAN BE CONSIDERED FOR CONVERTING THEM INTO A POTENTIAL LEAD
- AFTER RUNNING THE MODEL ON TEST DATA , A VERY GOOD ACCURACY (80%) AND VERY GOOD PRECISION/RECALL NUMBERS WERE OBTAINED. DETAILS ABOUT THESE ARE PRESENTED SEPARATELY IN ANOTHER DOCUMENT.
- ONE IMPORTANT METRIC WAS RECALL , WHICH WE WANTED TO GET A HIGHER AND WE HAVE IT 75% WHICH IS RELATIVELY GOOD.

RECOMMENDATIONS TO SALES TEAM

FROM THE MODEL IT IS FOUND THAT MOST OF THE POTENTIAL BUYERS ARE THOSE WHO SPEND MOST

1. THE TOTAL TIME SPEND ON THE WEBSITE.
2. TOTAL NUMBER OF VISITS.
3. WHEN THE LEAD SOURCE WAS:
    - A. GOOGLE
    - B. OLARK CHAT
    - C. ORGANIC SEARCH
4. WHEN THE LEAD ORIGIN IS FROM LEAD ADD FORMAT.
5. 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.