LEAD SCORING CASE STUDY

BY AMIT, RASHMI & MANPRIT

PROBLEM STATEMENT :

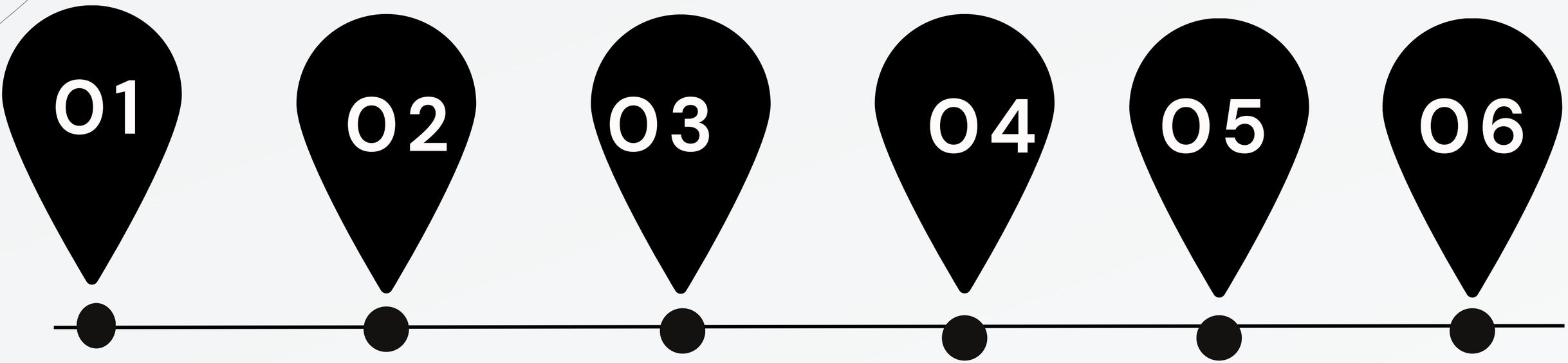
INTRODUCTION:

AN EDUCATION COMPANY, X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS. THE COMPANY MARKETS ITS COURSES ON VARIOUS WEBSITES AND SEARCH ENGINES SUCH AS GOOGLE
ONCE PEOPLE LAND ON THE WEBSITE, THEY MIGHT BROWSE THE COURSES OR FILL UP A FORM FOR THE COURSE OR WATCH SOME VIDEOS. WHEN THESEPEOPLE 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. THE TYPICAL LEAD CONVERSION RATE AT X EDUCATION IS AROUND 30%

BUSINESS GOALS:

Company wishes to identify the most potential leads, also known as “Hot Leads”
The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead scorehave a higher conversion chance and customer with lower lead score have a lower conversion chance
The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%

OVERALL APPROACH



DATA CLEANING AND IMPUTING MISSING VALUES

EXPLORATORY DATA ANALYSIS : UNIVARIATE , BIVARIATE AND MULTIVARIATE ANALYSIS

FEATURE SCALING AND DUMMY VARIABLE CREATION

LOGISTIC REGRESSION MODEL BUILDING

MODEL EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION AND RECALL

CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY

DATA CLEANING AND PREPARATION

- Read data from source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier treatment
- Exploratory data analysis

SPLITTING THE DATA AND FEATURE SCALING

- Splitting the data into train and test dataset
- Feature scaling of numerical variables

MODEL BUILDING

- Feature selection using RFE, VIF and p-value
- Determine optimal model using Logistic Regression
- Calculate various evaluation metrics

RESULT

- Determine Lead score and check if target final prediction is greater than 80% conversion rate
- Evaluate final prediction on test set

DATA CONVERSION

01 CONVERTING THE VARIABLE WITH VALUESYES/NO TO 1/0S

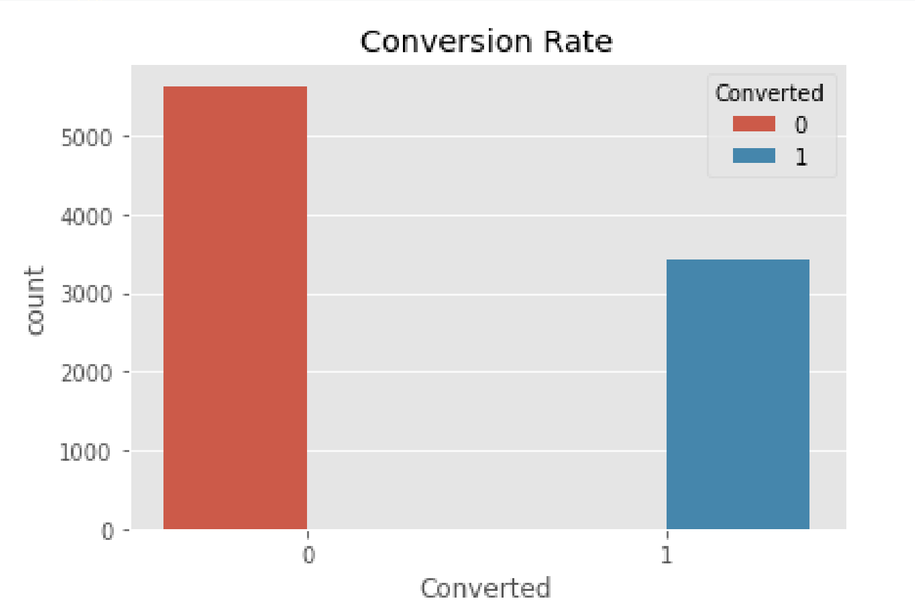
02 CONVERTING THE ‘SELECT’ VALUES WITH NANS

03 DROPIING THE COLUMNS HAVING >70% OF NULL VALUES

04 DROPPING UNNECESSARY COLUMNS

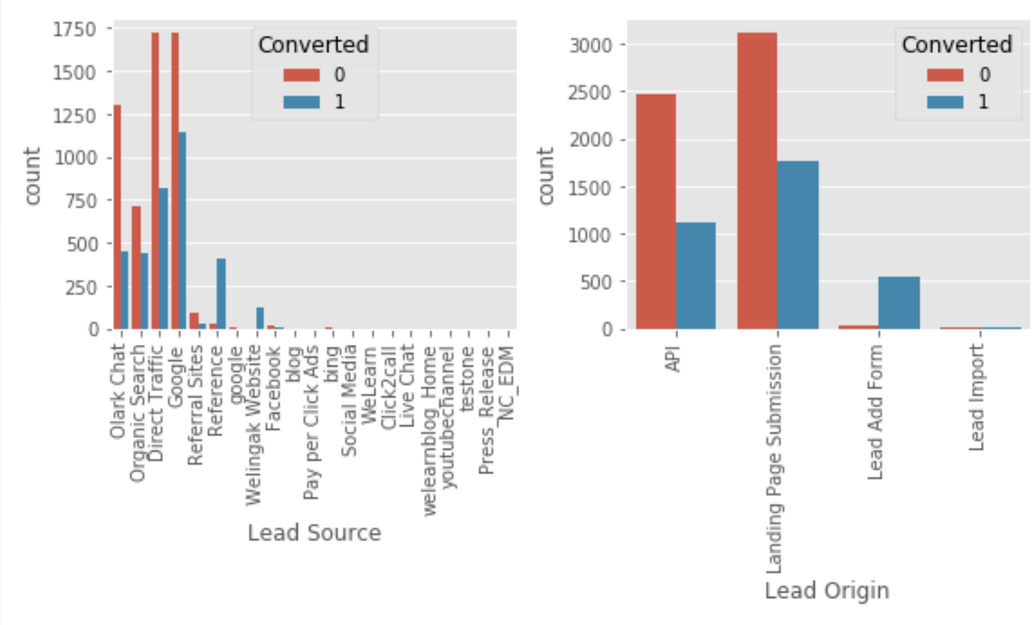
05 DROPPING THE ROWS AS THE NULL VALUES WERE <2%

EXPLORATORY DATA ANALYSIS

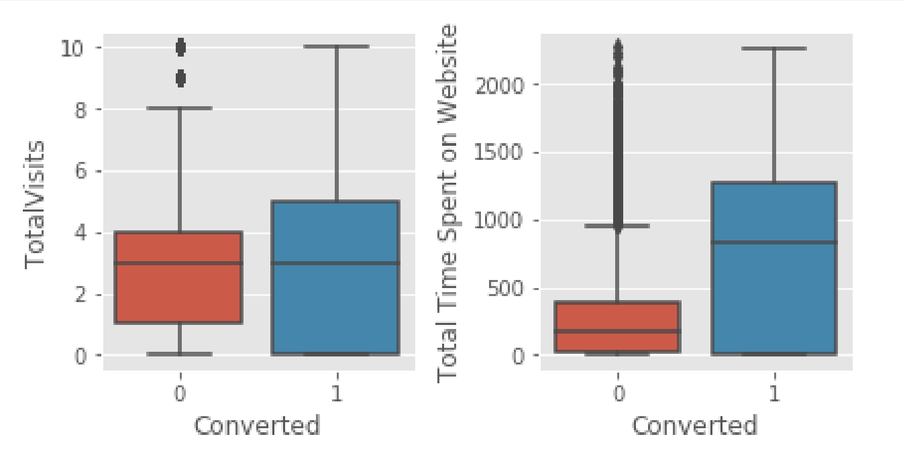


- We have around 30%of Conversion Rate

- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and WelingakWebsite is maximum
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

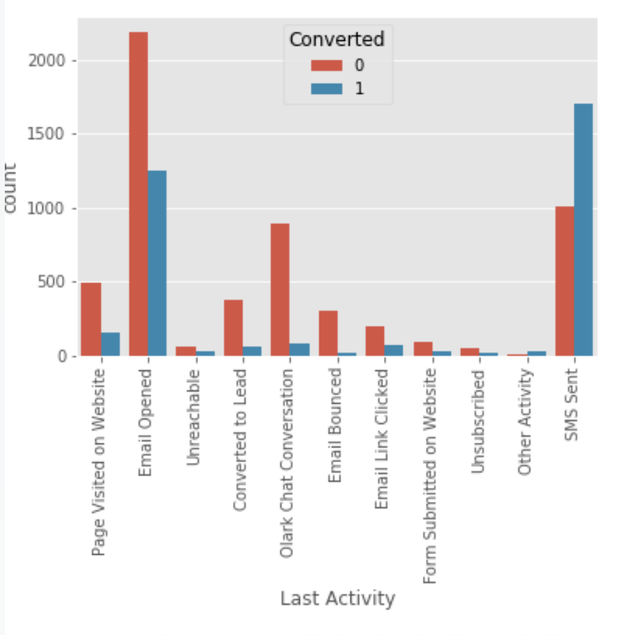


EXPLORATORY DATA ANALYSIS

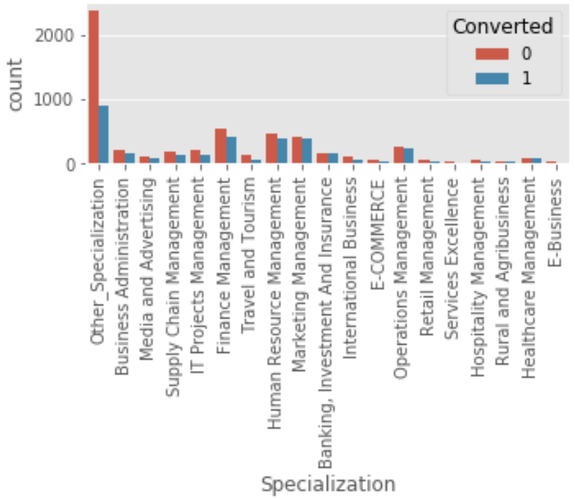


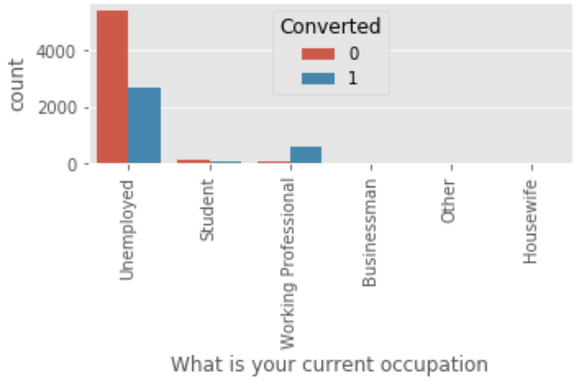
- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information
- Users spending more time on the website are more likely to get converted

- The count of lead's last activity as "Email Opened" is maximum
- The conversion rate of SMS sent as last activity is maximum



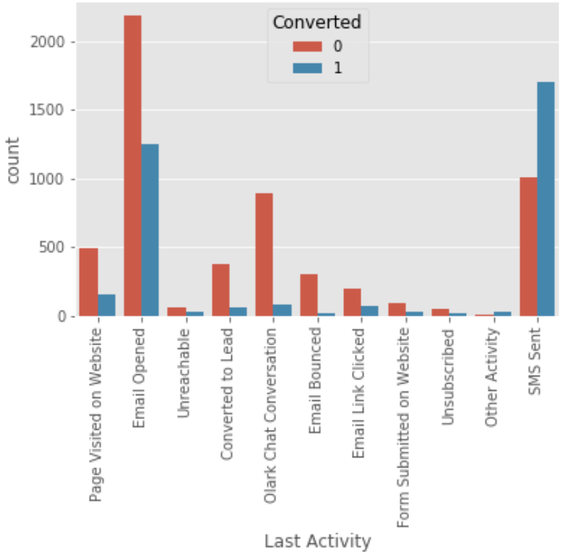
EXPLORATORY DATA ANALYSIS





- Looking at above plot, no particular inference can be made for Specialization
- Looking at above plot, we can say that working professionals have high conversion rate
- Number of Unemployed leads are more than any other category

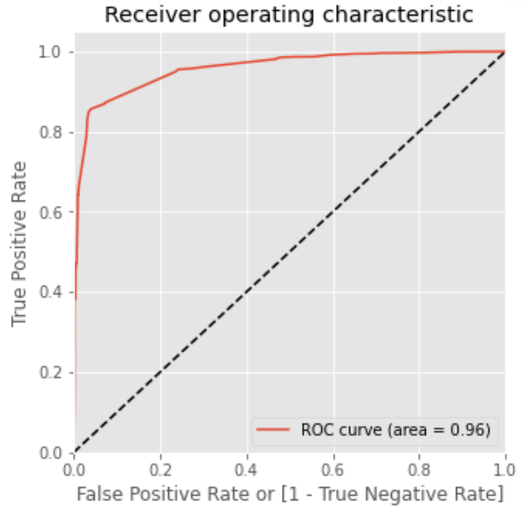
- 'Will revert after reading the email' and 'Closed by Horizzon' has high conversion rate



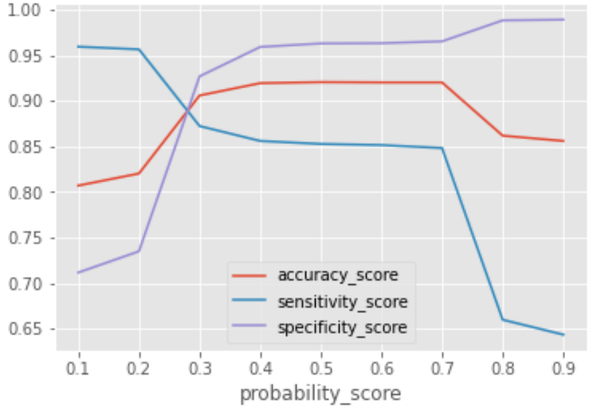
MODEL BUILDING

- SPLITTING THE DATA INTO TEST AND TRAINING SETS
- WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30
- USING RFE TO CHOOSE TOP 15 VARIABLES
- BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5
- PREDICTIONS ON TEST DATASET
- OVERALL ACCURACY IS 92.0 %

ROC CURVE



OPTIMAL CUT-OFF



MODEL EVALUATION

- CALCULATED ACCURACY, SENSITIVITY AND SPECIFICITYFOR VARIOUS PROBABILITY CUTOFFS FROM 0.1 TO 0.9
- AS PER THE GRAPH AND LOOKING AT THE OTHER SCORES, IT CAN BE SEEN THAT THE OPTIMAL POINT IS 0.27

	probability_score	accuracy_score	sensitivity_score	specificity_score
0.1	0.1	0.807117	0.959526	0.711652
0.2	0.2	0.820343	0.956664	0.734955
0.3	0.3	0.905999	0.872445	0.927017
0.4	0.4	0.919540	0.856092	0.959283
0.5	0.5	0.920642	0.852821	0.963124
0.6	0.6	0.920328	0.851594	0.963380
0.7	0.7	0.920328	0.848324	0.965429
0.8	0.8	0.861912	0.659853	0.988476
0.9	0.9	0.856086	0.643500	0.989245

ACCURACY	**83.59%**
PRECISION	71.6%
SENSITIVITY	94.9%
SPECIFICITY	76.5%

TRAIN DATA -CONFUSION MATRIX

PREDICTED ACTUAL	**NOT CONVERTED**	**CONVERTED**
NOT CONVERTED	2987	918
CONVERTED	124	2322

MODEL PREDICTION

```
----------------------Feature Importance--------------------
const                                              -1.248649
Do Not Email                                       -1.180501
Lead Origin_Lead Add Form                           0.908052
Lead Source_Welingak Website                        3.218160
Last Activity_SMS Sent                              1.927033
Tags_Busy                                           3.649486
Tags_Closed by Horizzon                             8.555901
Tags_Lost to EINS                                   9.578632
Tags_Ringing                                       -1.771378
Tags_Will revert after reading the email            3.831727
Tags_switched off                                  -2.336683
Lead Quality_Not Sure                              -3.479228
Lead Quality_Worst                                 -3.943680
Last Notable Activity_Modified                     -1.682075
Last Notable Activity_Olark Chat Conversation      -1.304940
```

ACCURACY	**81.5%**
PRECISION	68.0%
SENSITIVITY	92.8%
SPECIFICITY	75.1%

TRAIN DATA -CONFUSION MATRIX

PREDICTED ACTUAL	**NOT CONVERTED**	**CONVERTED**
NOT CONVERTED	1303	431
CONVERTED	71	918

CONCLUSION

The logistic regression model is used to predict the probabillty of conversion of a customer.

While we have calculated both sensitivity-specificityas well as Precision-Recallmetrics, we have considered optimal cut off on the basis of sensitivity-specificity for final prediction

Lead Score calculated shows the conversion rate of final predicted model is around 92% in test dataas compared to 95% in train data

In Business terms, this model has capability to adjust with the company’s requirements in coming future

TOP variables that contributes for lead getting converted in the model are:

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Lead Quality_Worst