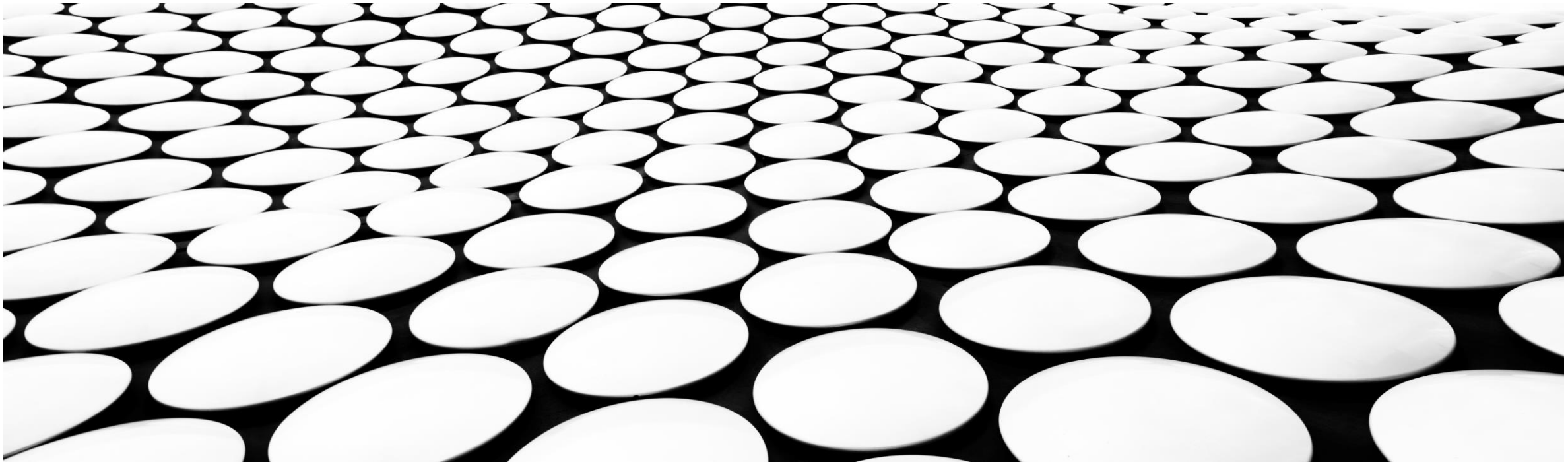
LEAD SCORING CASE STUDY

18TH JULY 2023

SANDEEP S, SANJAY SH, SANJAY K



PROBLEM STATEMENT

X Education, which sells online courses to industry professionals, gets a lot of leads through multiple sources. However, the lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

Business Objective:

- Assist X Education in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Build a model which assigns a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance (Hot Leads) and the customers with a lower lead score have a lower conversion chance. (Cold Leads)
- Target Lead conversion rate to be around 80%.

SOLUTION MODELING METHODOLOGY

- Data Cleaning and Data Preparation
  - Handling Columns with Missing Data
  - Mapping Binary categorical variables to 0 or 1
  - Creating Dummy Variables
- EDA
  - Checking for outliers and Removing outliers
- Test – Train Split
- Feature Scaling
- Identifying Correlations
- Model Building (RFE based, VIF/ p-value based)
- Model Evaluation using Key Metrics
- Model Prediction on Test Set
- Summary: Identifying the Key Important Predictor Variables

DATA CLEANING AND MANIPULATION

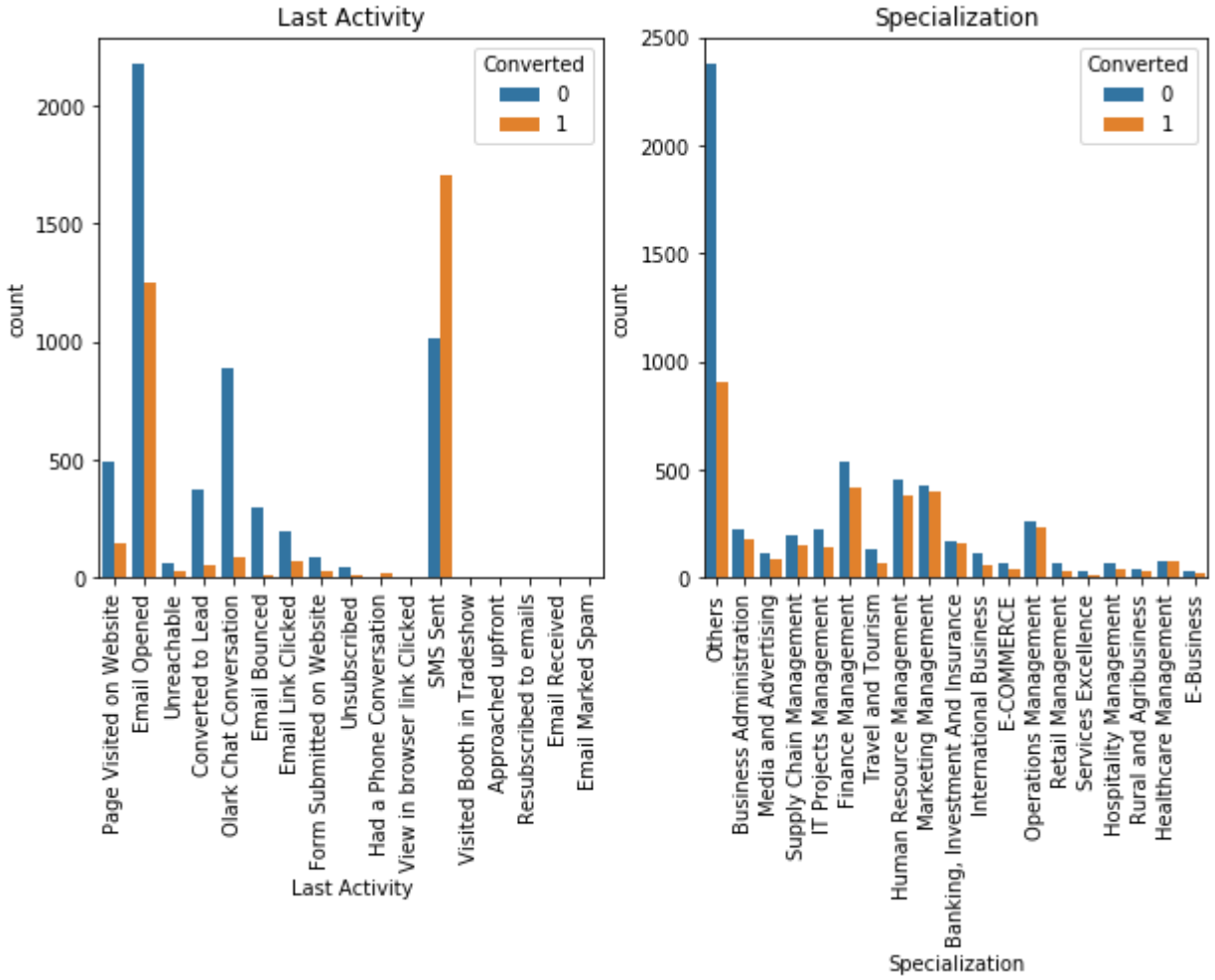
- Replaced ‘Select’ in columns with Null values
- Columns Dropped Due to High % of Missing Values:
  - How did you hear about X education?
  - Lead Profile
  - City
  - Country
  - Asymmetric Activity & Profile Variables
  - Lead Quality
  - Tags & What matters most to you in choosing a course – Data is not distributed properly
- Imputed Null Values with
  - ‘Others’ for Specialization
  - ‘Unemployed’ for Occupation
- Outliers: Picked only the 99% percentile data
  - ‘Total Visits”
  - ‘Page Views per visit’

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000

Outliers

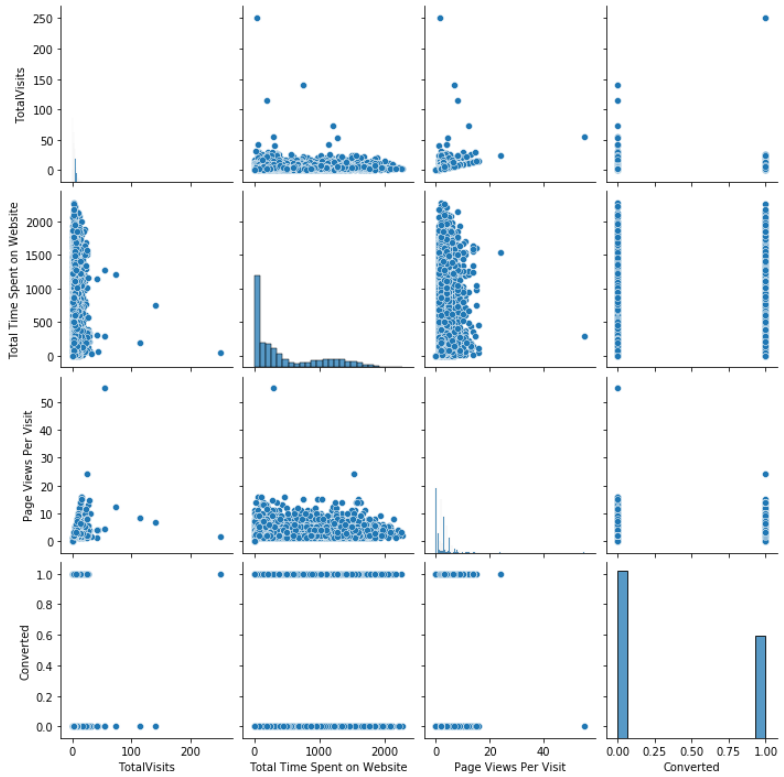
EDA – CATEGORICAL VARIABLES

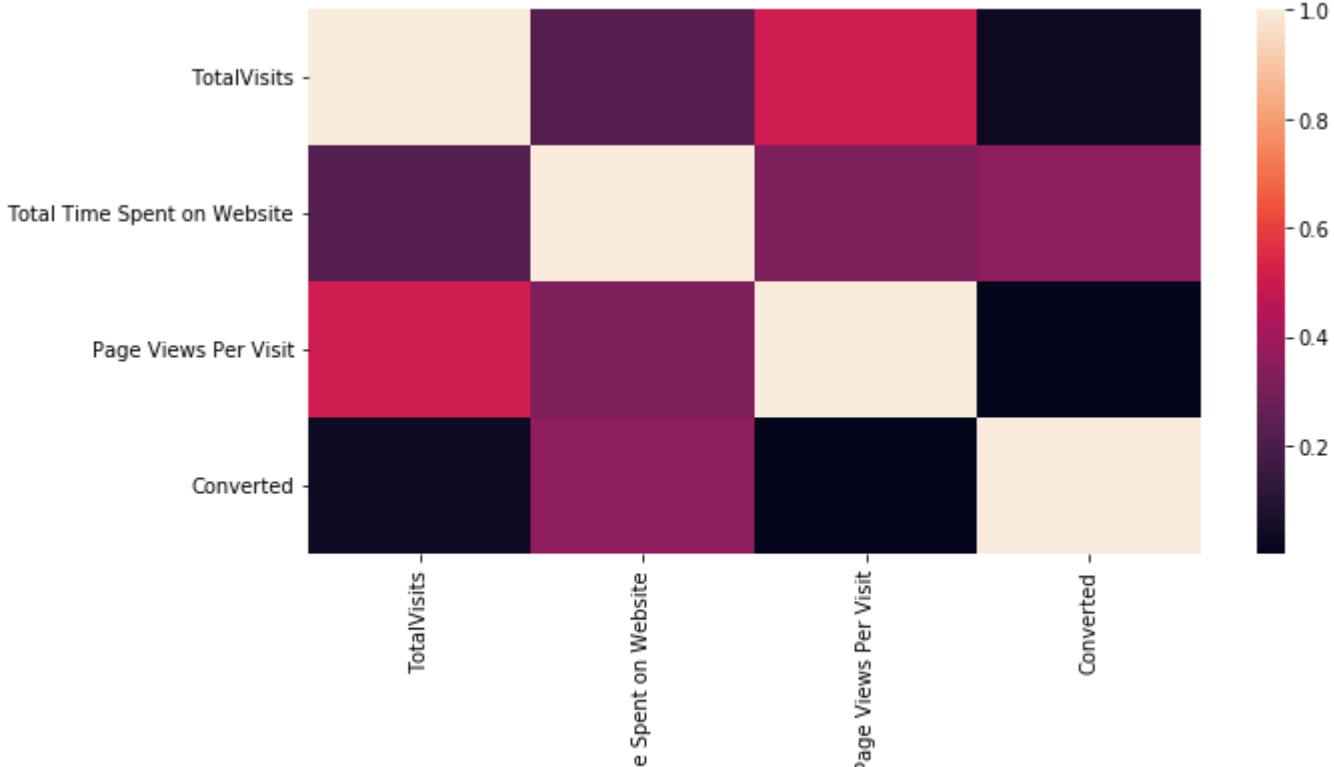
- Higher Conversion Chances When:
  - Last Activity is:
    - SMS is sent
    - Email Opened
  - Specialization is:
    - HR Management
    - Marketing Management
    - Operations Management
    - Finance Management



EDA – NUMERICAL VARIABLES

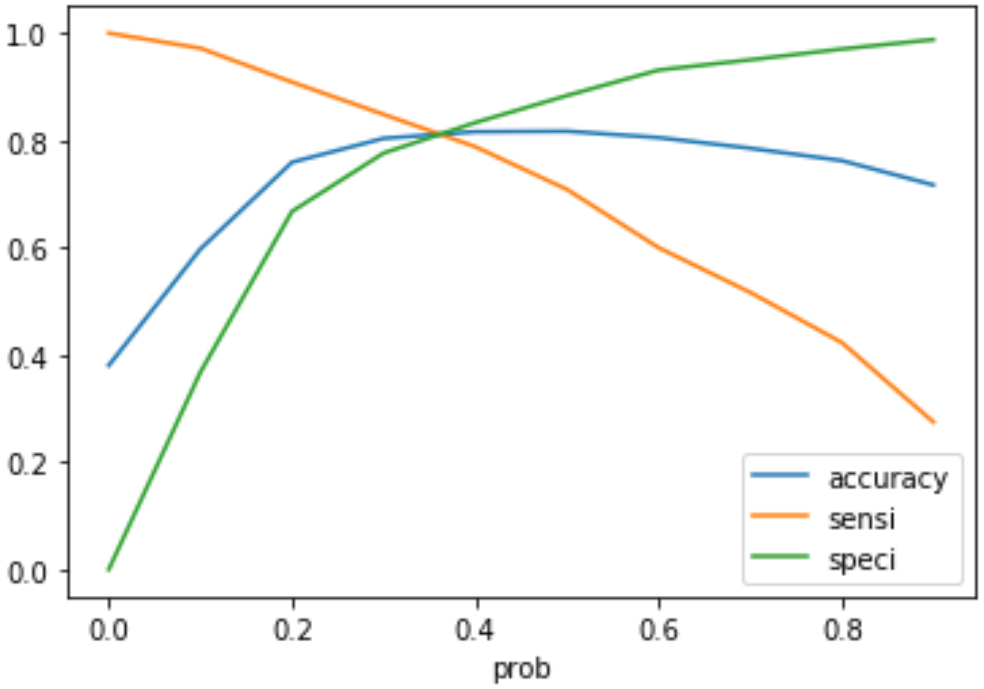
- Doesn’t seem to be any correlation between the numerical variables

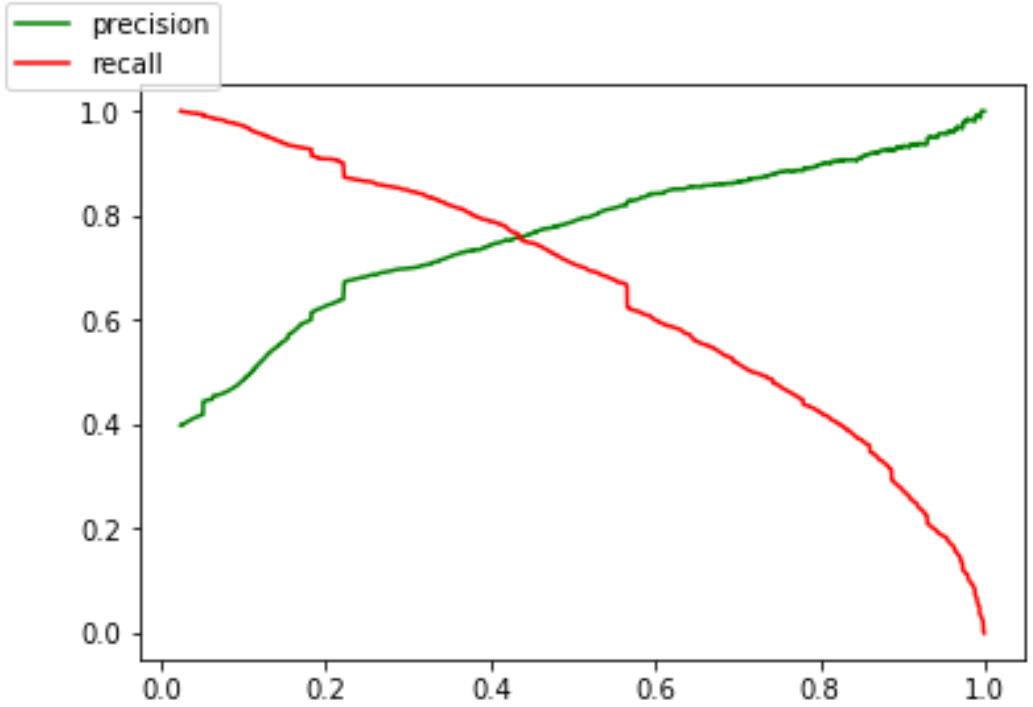




MODEL EVALUATION
ROC CURVE & PRECISION – RECALL TRADEOFF

- 0.4 Seems to be the optimal cut off – based on the plots of accuracy, sensitivity and specificity
- If Conversion Probability >0.4; Categorise them as Hot Leads (1); Else map them to Cold Leads (0)





MODEL EVALUATION METRICS & FINAL FEATURE LIST TRAIN & TEST DATA

- Train Data:
    i. Accuracy: 82%
    ii. Precision: 74%
    iii. Recall: 78%
- Test Data:
    i. Accuracy: 80%
    ii. Precision: 72%
    iii. Recall: 76%

Most Important Attributes or Features with Positive Impact:

- Total Time Spent on Website
- TotalVisits
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Last Activity_Had a Phone Conversation
- What is your current occupation_Working Professional

SUMMARY

Higher Chances of conversion when the lead is :

- Current Occupation as Working professionals
- Spend more time on the website
- Visit the website repeatedly
- Leads who have come from (Lead source or lead origin)
    - Welingak website
    - Lead Add Form
    - Olark Chat
- Where last activity was a phone conversation with the lead