X Education - Lead Scoring Case Study

INCREASING LEAD CONVERSION RATE FOR X EDUCATION

Background

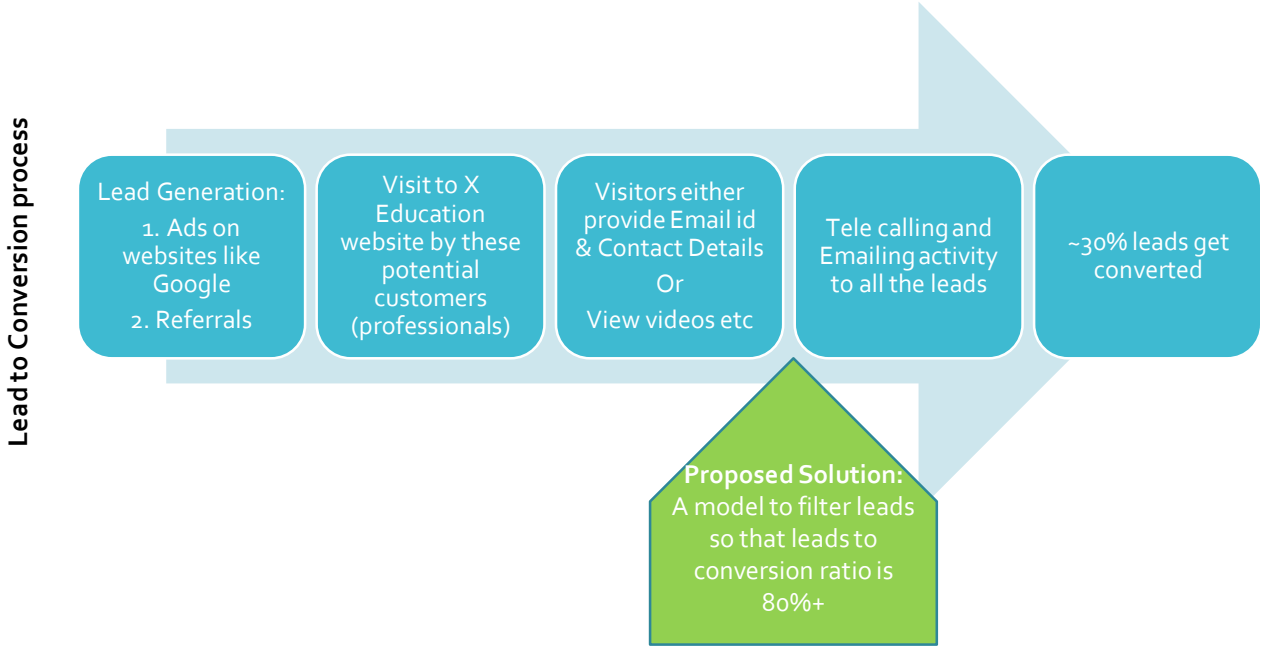
- An education company named X-education sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos
- When these people fill up a form providing their email address or phone number, they are classified to be a lead
- 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%

Problem Statement

X Education Company's Problem

- X Education gets a lot of leads but its lead conversion rate is very poor
- 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
- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required 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
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Lead – Conversion Process



Solution

Selection of Hot Leads

A. 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.

B. EDA

1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

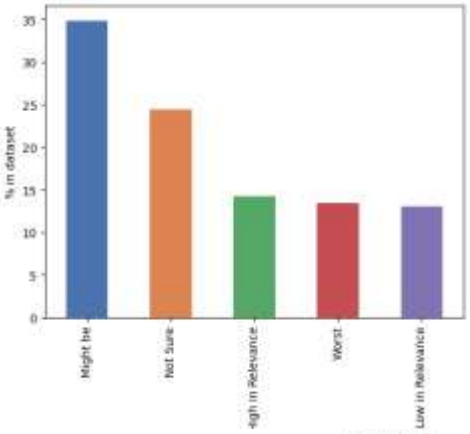
C. Feature Scaling & Dummy Variables and encoding of the data.

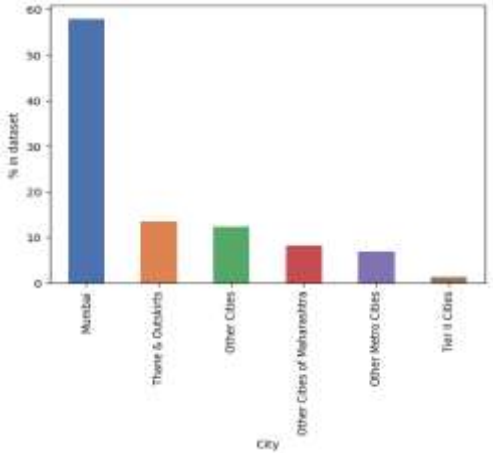
D. Classification technique: logistic regression used for the model making and prediction.

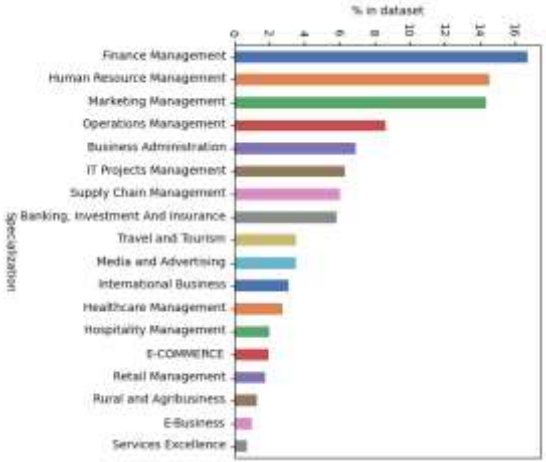
E. Model presentation.

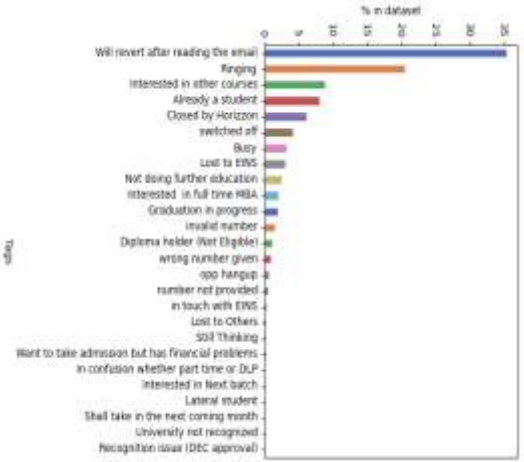
F. Conclusions and recommendations.

EDA

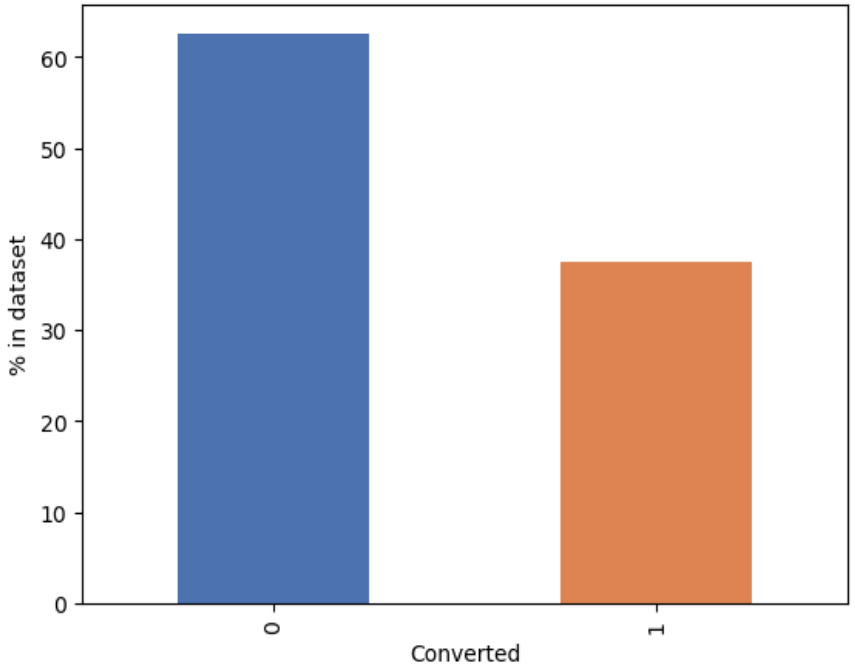

% in dataset
Might be
Not Sure
High in Relevance
Worst
Low in Relevance


% in dataset
Mumbai
Thane & Outskirts
Other Cities
Other Cities of Maharashtra
Other Metro Cities
Tier II Cities
City


% in dataset
Specialization
Finance Management
Human Resource Management
Marketing Management
Operations Management
Business Administration
IT Projects Management
Supply Chain Management
Banking, Investment And Insurance
Travel and Tourism
Media and Advertising
International Business
Healthcare Management
Hospitality Management
E-COMMERCE
Retail Management
Rural and Agribusiness
E-Business
Services Excellence

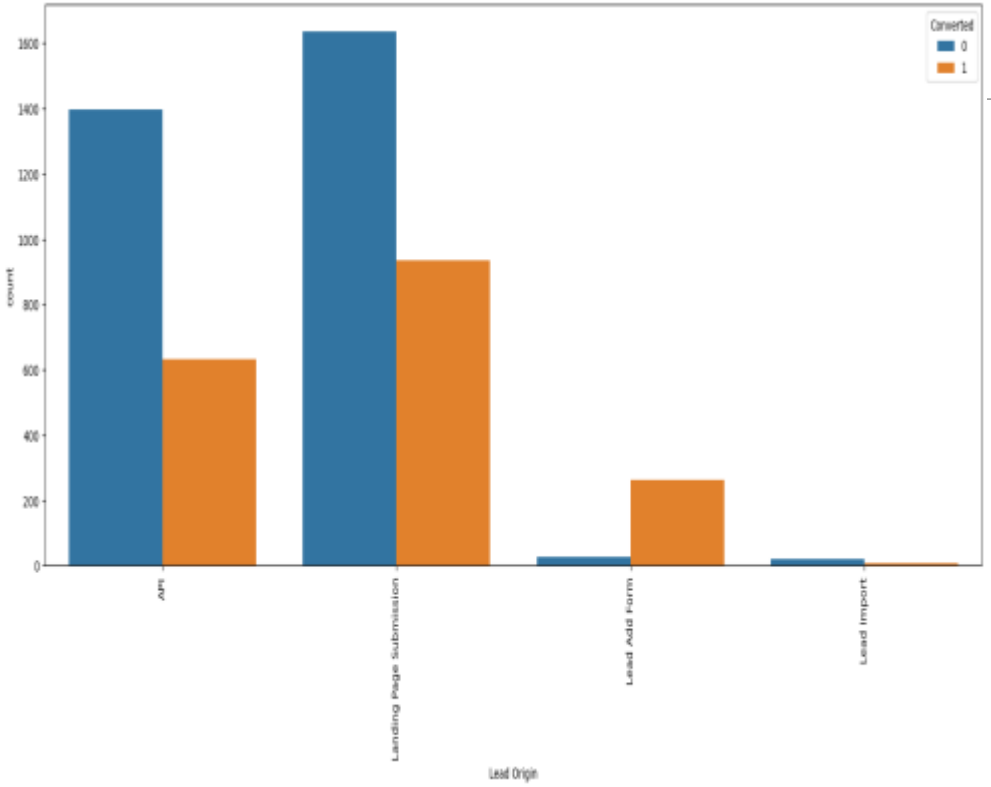

% in dataset
Tags
Will revert after reading the email
Ringing
Interested in other courses
Already a student
Closed by Horizzon
switched off
Busy
Lost to EINS
Not doing further education
Interested in full time MBA
Graduation in progress
invalid number
Diploma holder (Not Eligible)
wrong number given
opp hangup
number not provided
in touch with EINS
Lost to Others
Still Thinking
Want to take admission but has financial problems
In confusion whether part time or DLP
Interested in Next batch
Lateral student
Shall take in the next coming month
University not recognized
Recognition issue (DEC approval)

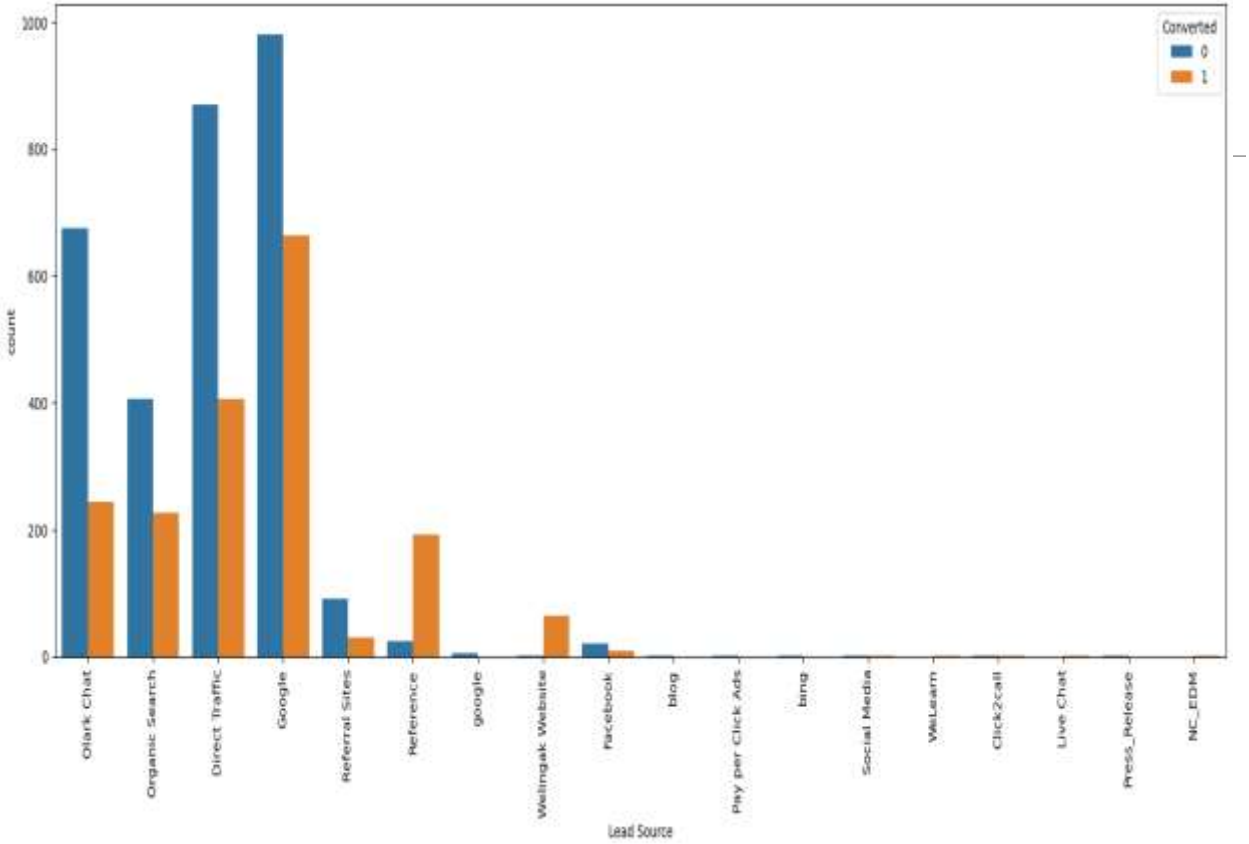
DATA VISUALIZATION - Converted

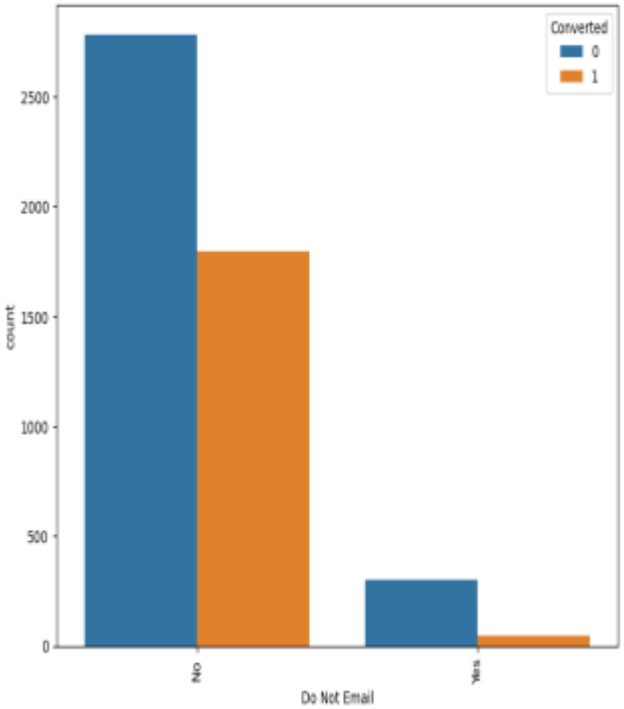


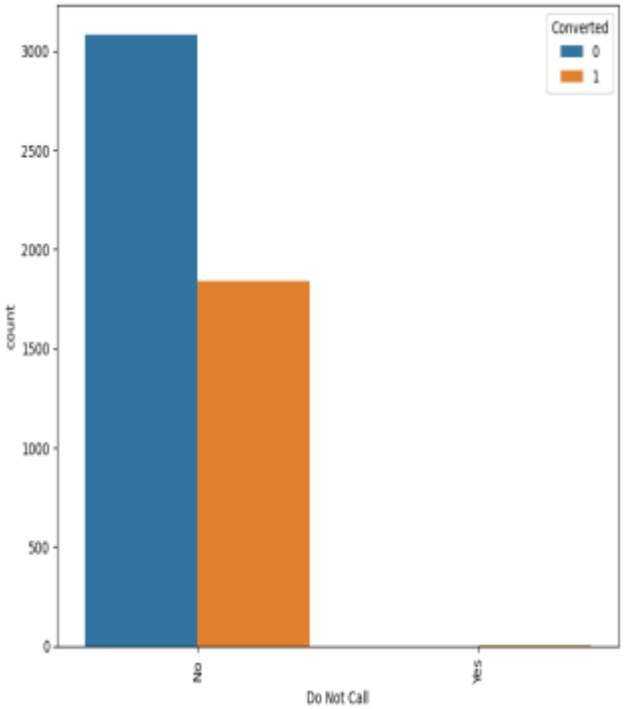
Note: 37.8% of the 'Converted' data is 1 ie. 37.8% of the leads are converted. This means we have enough data of converted leads for modelling.

CATEGORICAL VARIABLES

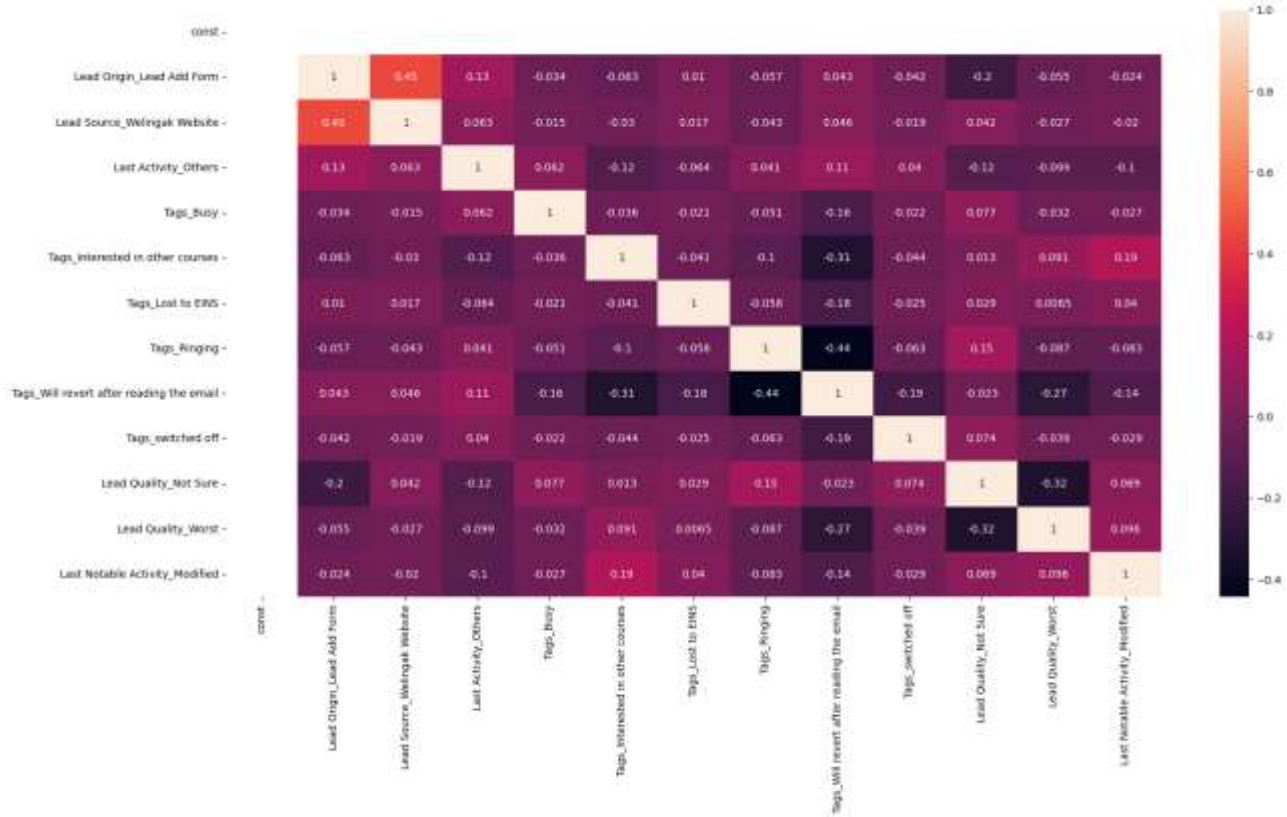



Converted
0
1
count
1000
800
600
400
200
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
Press_Release
NC_EDM
Lead Source


Converted
0
1
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Email


Converted
0
1
3000
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Call

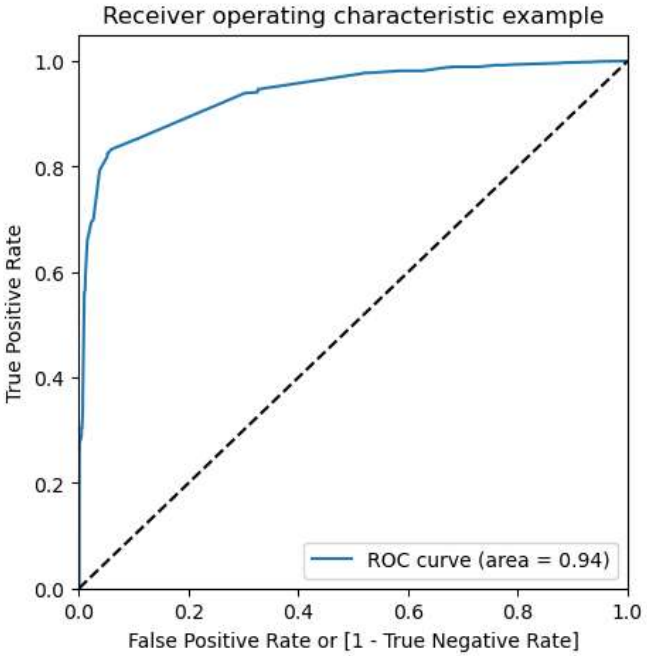
Heatmap

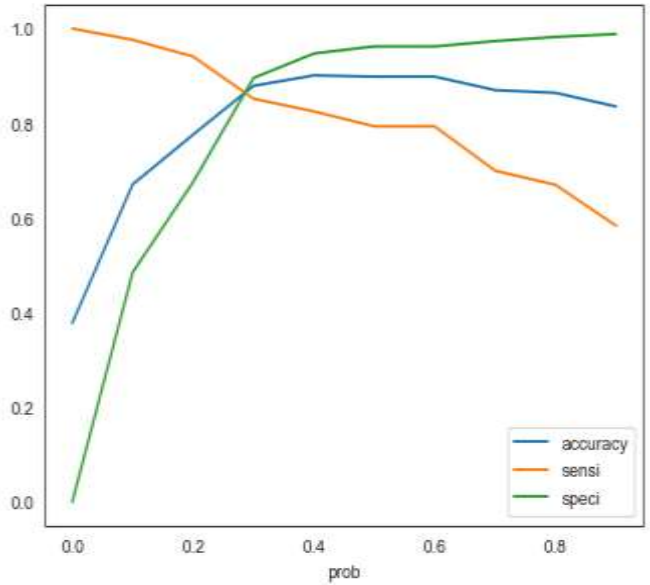

const
Lead Origin_Lead Add Form
Lead Source_Welingak Website
Last Activity_Others
Tags_Busy
Tags_Interested in other courses
Tags_Lost to EINS
Tags_Ringing
Tags_Will revert after reading the email
Tags_switched off
Lead Quality_Not Sure
Lead Quality_Worst
Last Notable Activity_Modified
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Predictions on test data set.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.94)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

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

a. The total time spend on the Website.
b. Total number of visits.
c. When the lead source was: a. Google b. Direct traffic c. Organic search
d. When the lead origin is Lead add format.
e. When their current occupation is as a working professional.

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.

Thank-you