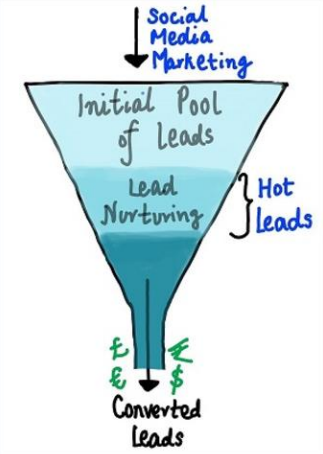
Lead Scoring Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Batch : DS C68

Problem Statement

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance



Lead Conversion Process - Demonstrated as a funnel

Goals

There are quite a few goals for this case study:
1.Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Assumptions

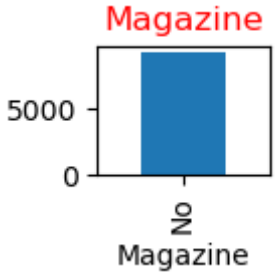
- Columns which have empty cells more than 30 % have been dropped
- Columns which have very less (<2%) percent of empty cells those empty rows also have been dropped (insignificant nos.)
- Remaining numerical columns nan value have been imputed as median
- Remaining categorical columns nan value have been imputed as median

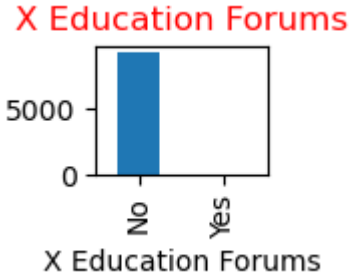
Preprocessing

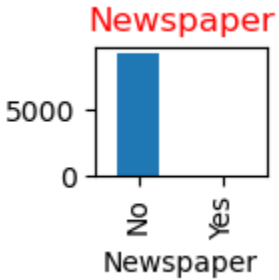
Lead Quality	51.590909
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Score	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Index	45.649351
Tags	36.287879

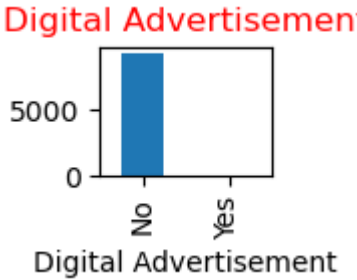
- Columns with more than 30% null values are dropped
- There were a few columns in which only one value was majorly present for all the data points. These include
  - `Do Not Call`, `Search`, `Magazine`, `Newspaper Article`, `X Education Forums`, `Newspaper`, `Digital Advertisement`, `Through Recommendations`, `Receive More Updates About Our Courses`, `Update me on Supply Chain Content`, `Get updates on DM Content`, `I agree to pay the amount through cheque`.
- Since practically all of the values for these variables are same, it's best that we drop these columns as they won't help with our analysis.
- Also, the variable `What matters most to you in choosing a course` has the level `Better Career Prospects` `6528` times while the other two levels appear once twice and once respectively. So we should drop this column as well.
- country and 'Do not email' are majority one type data so we ll drop those
- Dropping rows with null values
- Imputing means in numerical columns where there is null
- Select column is dropped after dummy creation

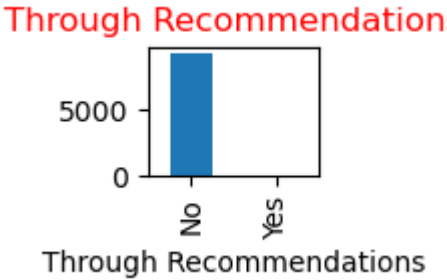
Below Categorical variables did not provide much variation







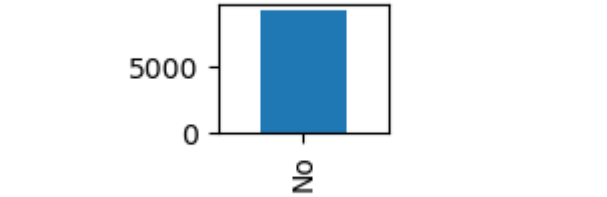




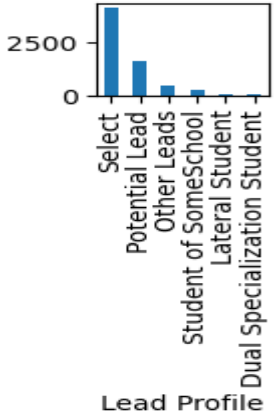
Above Categorical variables Dropped as no variation

Below variables did not provide much variation

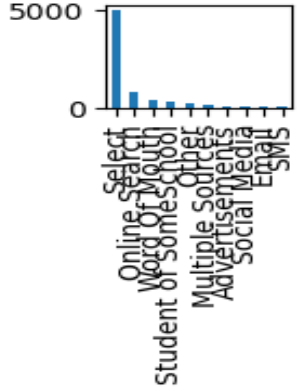
I agree to pay the amount through cheque



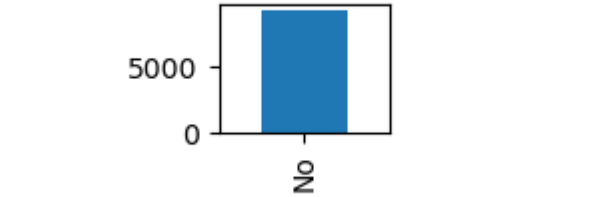
Lead Profile



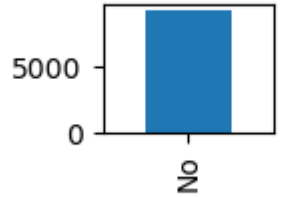
How did you hear about X Education



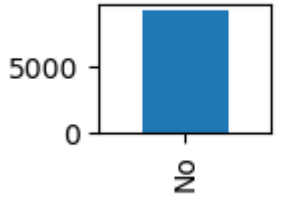
Receive More Updates About Our Courses



Update me on Supply Chain Content

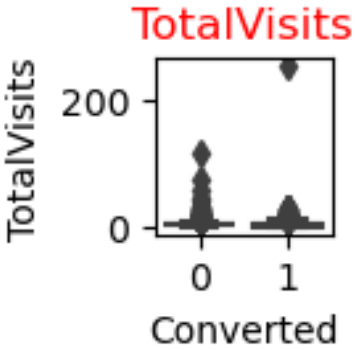


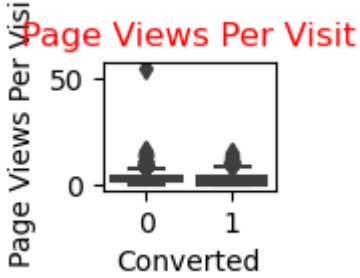
Get updates on DM Content



Above Categorical variables Dropped as no variation

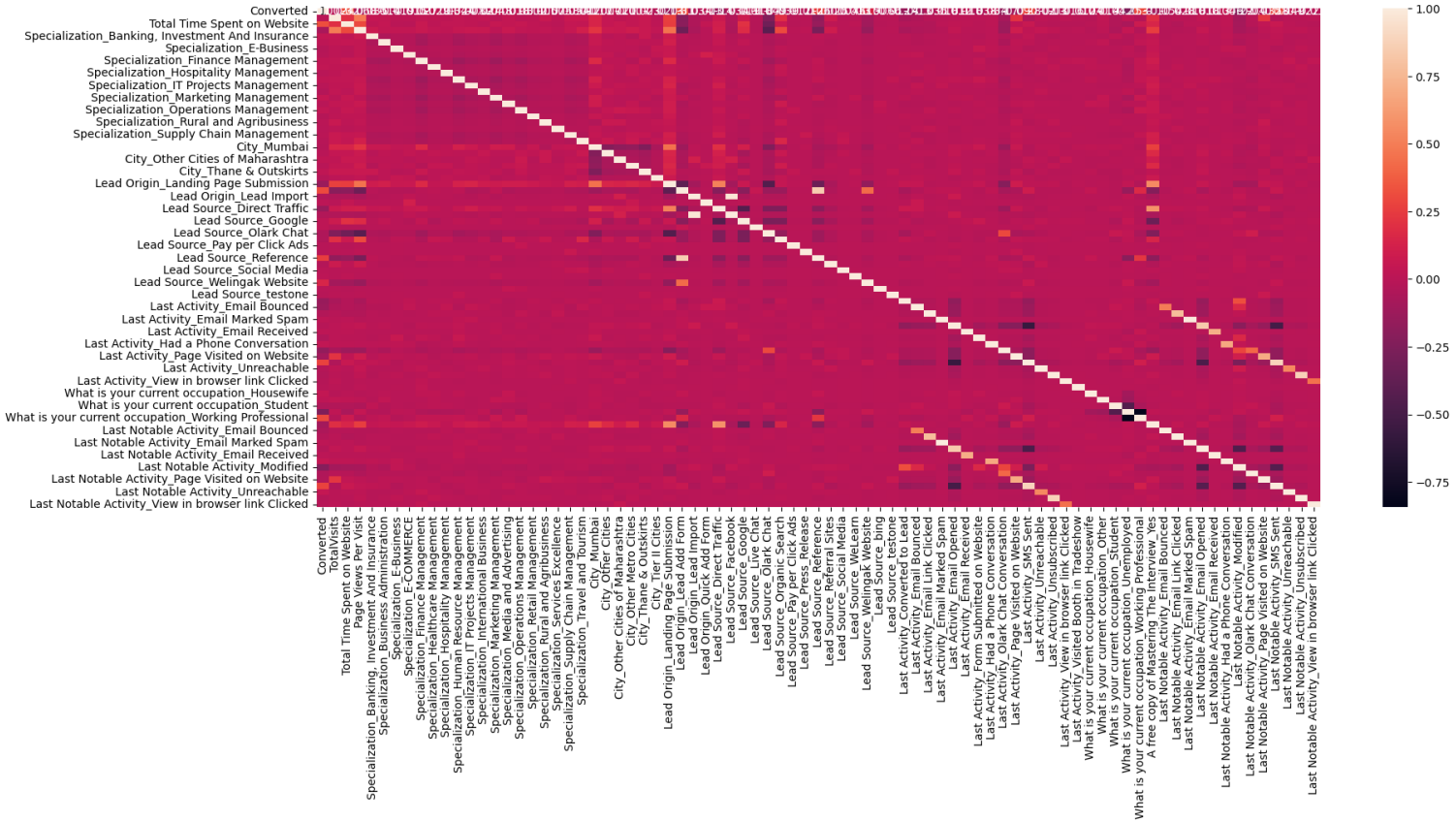
Numerical variable distribution





Numerical ariables scaled around 0

Correlation of variable



Variables are not much correlated, so not dropped on this basis

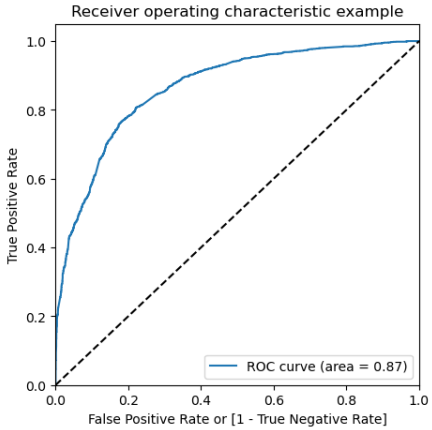
Feature scaling

- Standard scalar to scale numerical columns
  - ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']

Model Building

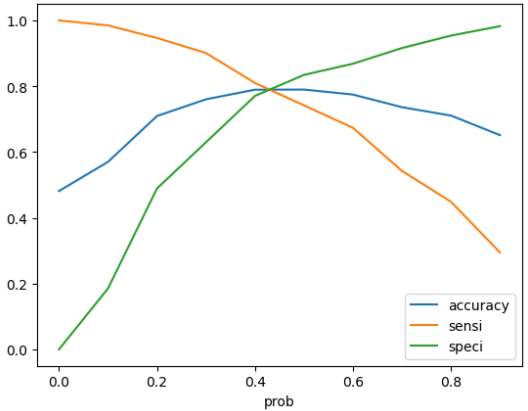
- First top 15 features selected using RFE
  - ['Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Email Bounced', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable', 'Last Notable Activity_Unsubscribed‘]
- Analysing p value and VIF using statsmodel and dropping features which are having p value greater than 0.05
  - What is your current occupation_Housewife, Last Notable Activity_Had a Phone Conversation, Last Notable Activity_Unsubscribed,
- and VIF more than 5 (no such feature)
- Final features:
  - 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Email Bounced', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable

Model Evaluation



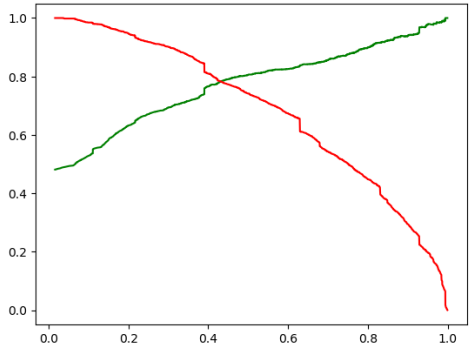
Accuracy-78.97
Sensitivity- 74.17
Specificity-83.43

Optimal cut off comes to be 0.435



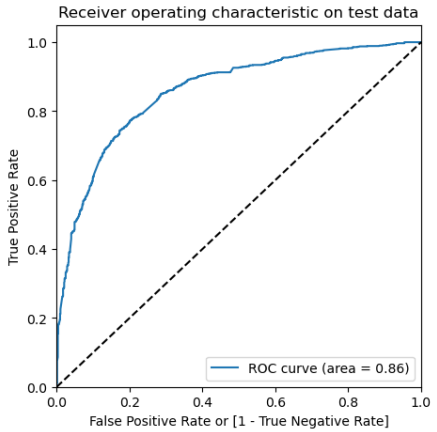
Accuracy-79.14
Sensitivity- 78.07
Specificity-80.15

Precision- 80.60
Recall- 74.17



Precision recall curve also meet at 0.435

ROC on test data



Accuracy-78.52
Sensitivity- 77.03
Specificity-80.02

So final model is acceptable as it has good Accuracy, sensitivity and specificity. Also variation of these parameters on train and test data is very less

Driver Features

So our final list of driver features are:

Feature	Coefficien
Lead Origin_Lead Add Form	3.47
Last Notable Activity_Unreachable	2.5707
Lead Source_Welingak Website	2.5379
Lead Source_Olark Chat	1.4356
Total Time Spent on Website	1.1236
What is your current occupation_Working Professional	1.1093
Last Activity_SMS Sent	0.9757
Last Activity_Olark Chat Conversation	-0.7984
Last Notable Activity_Modified	-0.8383
Last Activity_Email Bounced	-1.3495
What is your current occupation_Unemployed	-1.4383
What is your current occupation_Student	-1.4921

Conclusion

- The logistic regression model offers actionable insights into lead prioritization and conversion. By leveraging high-impact features like **Lead origin**, **Lead Source**, and **Last Notable Activity**, X Education can significantly improve its efficiency and conversion rate.
- Additionally, customized strategies for aggressive conversion and call minimization phases ensure the model adapts well to changing business needs.
- Leads who spent more time on website are more likely to convert

Thank You!