Lead Scoring Case Study

Submitted by:

Rafeek Ponnandy

Priyanka Chatterjee

Radhika Bansal

Contents

Problem Statement

- **X Education**, which is an education company is facing a challenges lead conversion, with only 30% of leads converting into paying customers. The company aims to improve this rate by identifying and prioritizing 'Hot Leads'—those most likely to convert.

- Currently, X Education generates leads through website visits, form submissions, and referrals. Despite acquiring many leads, the conversion rate remains low. To address this, X Education seeks a model to help them with a solution to help them identify the most promising leads, which are highly likely to be converted.

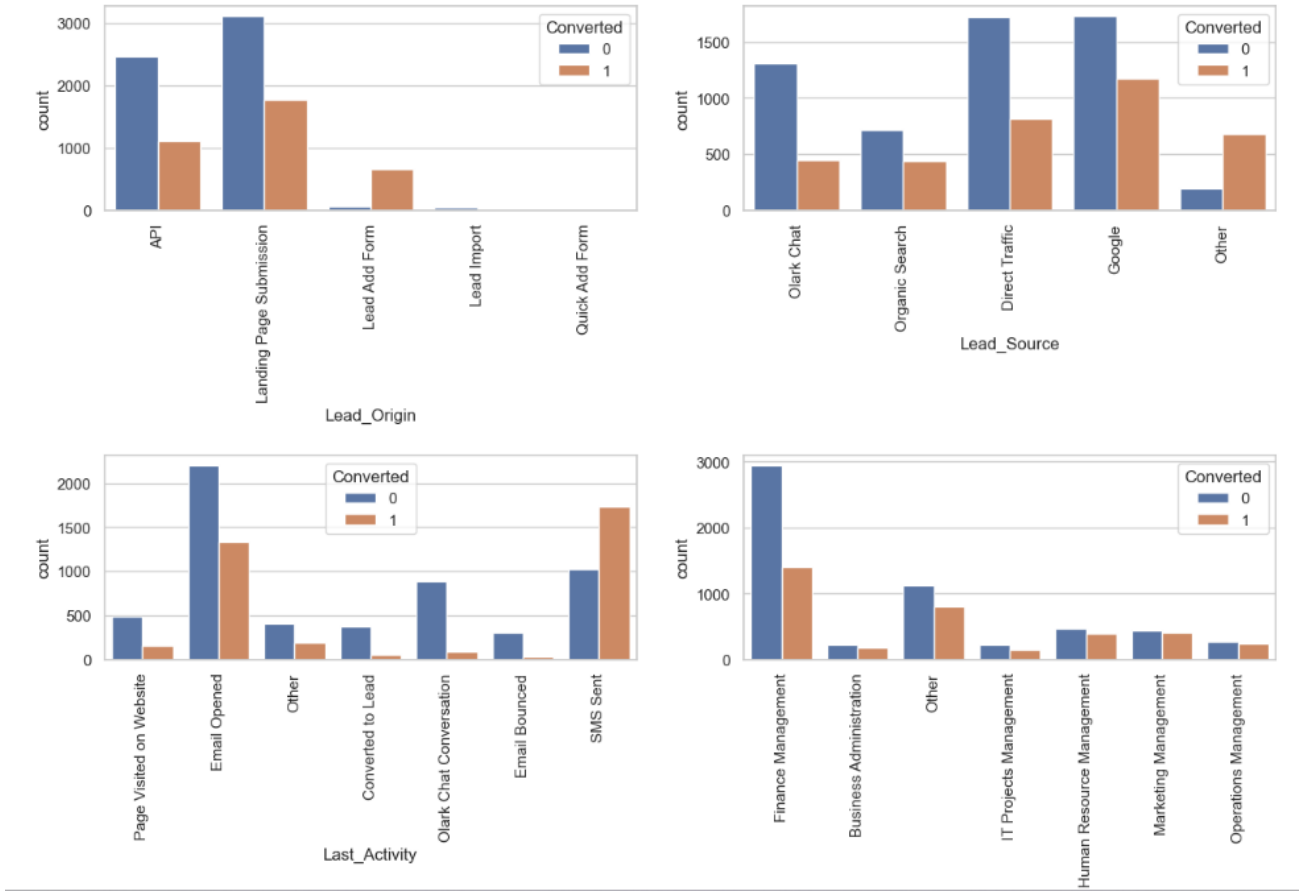
- The objective is to assign a lead score to each lead, indicating their likelihood of conversion. The leads with higher scores should be prioritized as they are more likely to convert into paying customers. The target lead conversion rate is set around 80%.

Solution approach

Approach to Building a Logistic Regression Model to Calculate the Probability of Lead Conversion

- Import and inspect the dataset
- EDA, Data Cleaning and preparation
- Dummy variable creation
- Create Test-Train Split
- Feature Scaling
- Model Building
- Model Evaluation
- Making prediction on the test set
- Conclusion

Exploratory data analysis



Inferences

- Most leads are originated from 'Landing Page
- Most leads came from 'Google' and 'Direct sources'.
- Most leads came from 'Mumbai'
- Last Notable Activity of leads - 'Modified', 'Email Opened','SMS send' in sequence.
- Majority of leads have Last Activity as 'Email Opened'
- Many leads have specialization in 'finance'

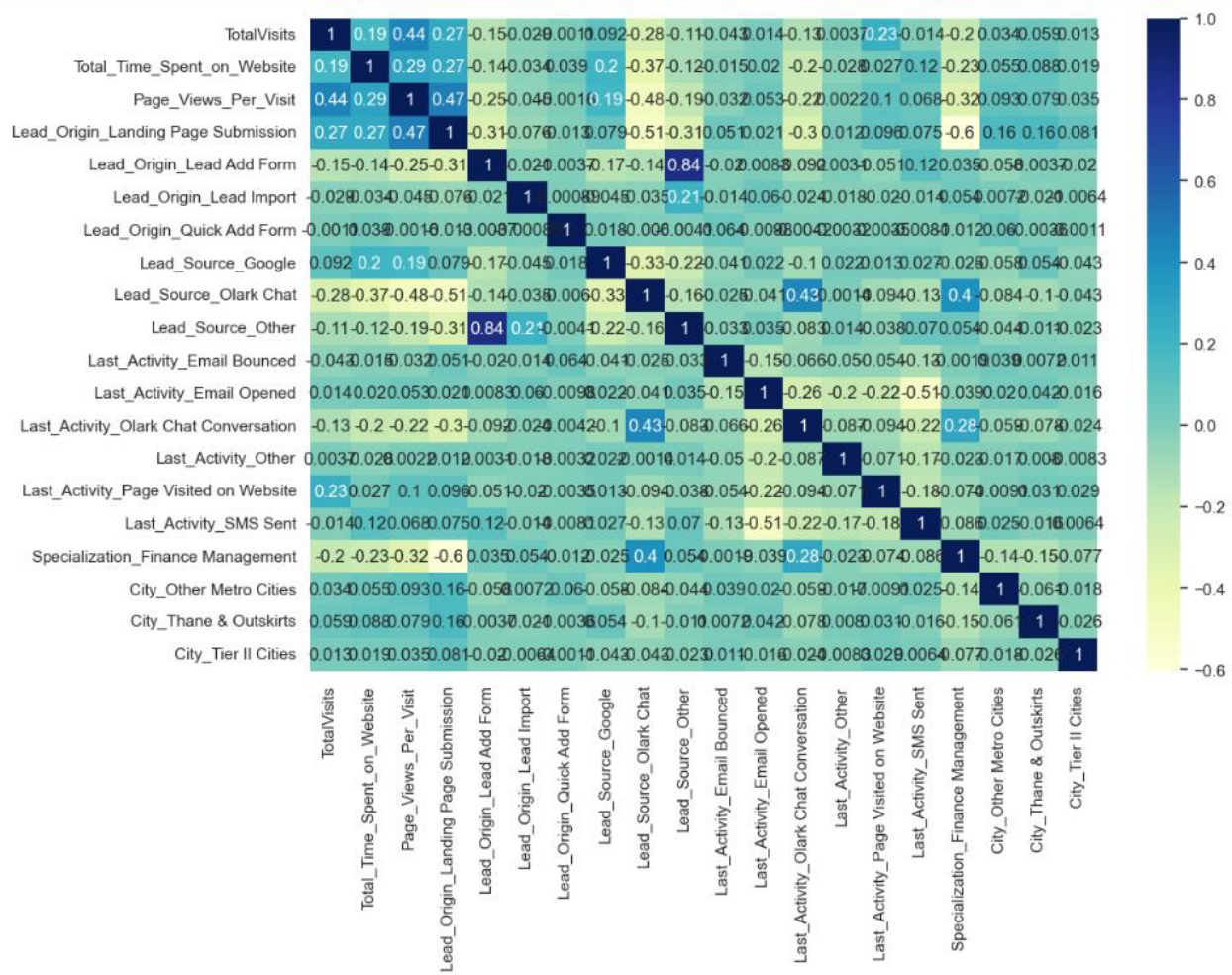
Data Cleaning and Data Preparation

- There are several columns containing missing data
- Columns which have higher frequency of missing values are removed from the model. The cut-off has been set to more than 30% missing
- Missing records from all numerical columns are deleted
- Remaining columns (Categorical) with missing values are labelled with missing category and included in the model

	Column	Null Percentage
0	How did you hear about X Education	78.463203
1	Lead Profile	74.188312
2	Lead Quality	51.590909
3	Asymmetrique Profile Score	45.649351
4	Asymmetrique Activity Score	45.649351
5	Asymmetrique Activity Index	45.649351
6	Asymmetrique Profile Index	45.649351
7	City	39.707792
8	Specialization	36.580087
9	Tags	36.287879
10	What matters most to you in choosing a course	29.318182
11	What is your current occupation	29.112554
12	Country	26.634199
13	Page Views Per Visit	1.482684
14	TotalVisits	1.482684
15	Last Activity	1.114719
16	Lead Source	0.389610
17	Receive More Updates About Our Courses	0.000000

Model Building

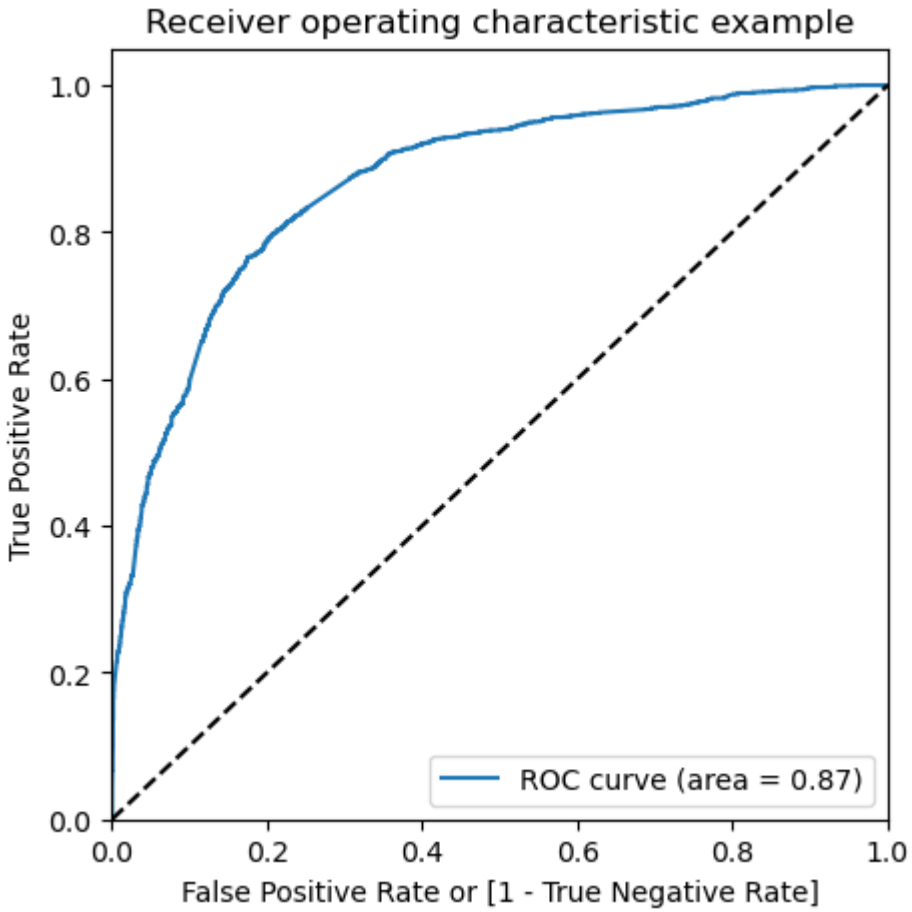
- Dummy variables are created for all categorical features
- Data has been split to Test train data
- Feature scaling done using Min-Max scaler
- Checked the correlation and variable selected further checked for VIF score
- Feature selection has bee done using RFE, to 15 features selected
- Ran a logistic regression model using stats model



Model Evaluation

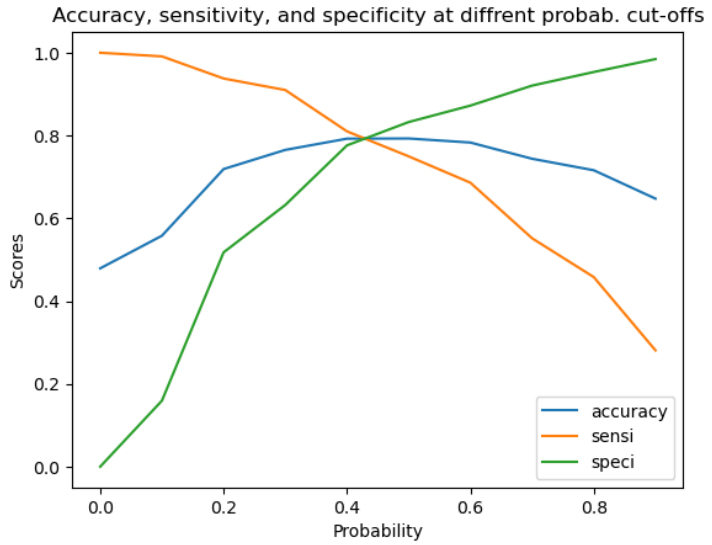
- The Area Under the ROC Curve (AUC-ROC) is 0.87 which shows a good model, this quantifies the ability of the model to discriminate between positive and negative classes

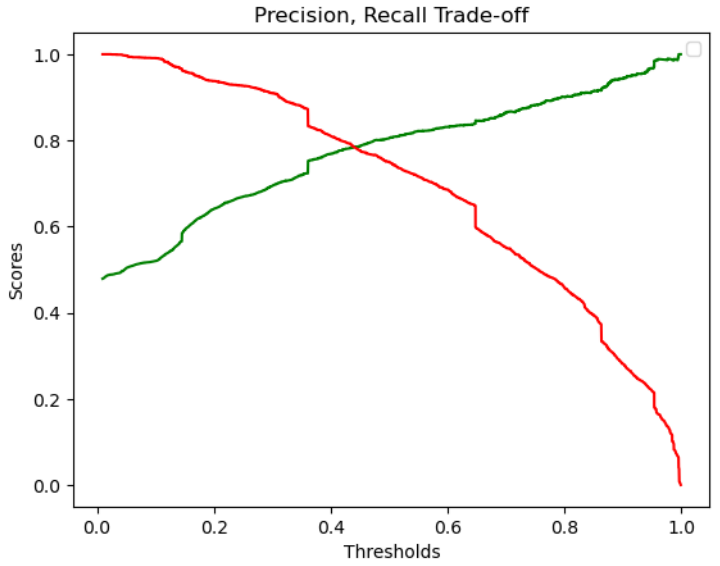
- Models performance with 42% cut-off of probability on training data,
  - Sensitivity (Recall) **79.7**%
  - Specificity is **79**%
  - Precision is **77.7**%
- On Test data
  - Sensitivity (Recall) **79**%
  - Specificity is **76**%
  - Precision is **76**%



Conclusion

- There in an optimal performing model at 42% Probability (precision, recall trade off), at this cut-off Model shows 77% precision and 78% recall rate.
- At a higher probability threshold the model would give increased precision as shown the precision- recall trade off chart.





Observations

Top features

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0578	0.115	-17.933	0	-2.283	-1.833
Total Time Spent on Website	**4.562**	0.187	24.41	0	4.196	4.928
Lead Origin_Lead Add Form	**3.8964**	0.258	15.08	0	3.39	4.403
TotalVisits	**3.0001**	0.599	5.009	0	1.826	4.174
What is your current occupation_Working Professional	**2.4917**	0.191	13.029	0	2.117	2.866
Source_Welingak Website	2.2903	1.047	2.187	0.029	0.238	4.342
Last Activity_Had a Phone Conversation	2.1501	0.703	3.059	0.002	0.773	3.528

Model Performance

Train data

Performance at 42% probability cut-off

Sensitivity (Recall) 79.7%

Specificity is 79%

Precision is 77.7%

Test data

Performance at 42% probability cut-off

Sensitivity (Recall) 79%

Specificity is 76%

Precision is 76%

Thank you

Rafeek Ponnandy

Priyanka Chatterjee

Radhika Bansal