Lead Scoring Case Study

By Rupam, Abhinow, Aadil

Introduction

About: This report contains the overall anlaysis on Lead Scoring Case study using the machine learning tools. This is to predict the leads that are most likely to convert into paying customers.

Problem Statement: An education company named X Education sells online courses to industry professionals.

- **Current Scenario:**
  - High volume of leads but low conversion.
  - Only about 30 out of 100 leads convert.
- **Goal:**
  - Assign lead scores to focus on promising leads.
  - Improve efficiency of the sales team.

Objective: Identify high-potential leads ("Hot Leads") to improve conversion rates to around 80%.

Data Overview

Shape = (9074, 37)

Attributes: Prospect ID', 'Lead Number', 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity', 'Country', 'Specialization', 'How did you hear about X Education', 'What is your current occupation', 'What matters most to you in choosing a course', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Tags', 'Lead Quality', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'Lead Profile', 'City', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'Last Notable Activity'

Target Variable: 'Converted' (1 = converted, 0 = not converted).

Challenge: Handle categorical variables with 'Select' as a level (acts as a null value).

Data Handling

Data Preparation

- Steps:
  - Data Sanity Checks: Using head(), shape, columns, describe(), info().
  - Handling Null Values:
    - Removed columns with >40% missing data.
    - Imputed or deleted rows for columns with <15% missing values.
  - Dummy Variables: Created for categorical columns.
  - Data Split: 70% train, 30% test.
  - Scaling: Normalized numerical variables.

Correlations

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google
Converted	1.000000	0.032855	0.359261	0.000260	-0.037481	0.300775	-0.009328	-0.073186	-0.010651	0.029960
TotalVisits	0.032855	1.000000	0.219723	0.511068	0.290347	-0.169742	-0.037808	0.095571	-0.036983	0.106848
Total Time Spent on Website	0.359261	0.219723	1.000000	0.318350	0.292571	-0.188526	-0.050742	0.140793	-0.050248	0.215390
Page Views Per Visit	0.000260	0.511068	0.318350	1.000000	0.484119	-0.268415	-0.056068	0.133118	-0.053735	0.204870
Lead Origin_Landing Page Submission	-0.037481	0.290347	0.292571	0.484119	1.000000	-0.282445	-0.062195	0.523695	-0.059438	0.078455
...	...	...	...	...	...	...	...	...	...	...
Last Notable Activity_Resubscribed to emails	0.013451	-0.007468	-0.009298	-0.011516	-0.011337	-0.002746	-0.000605	-0.006551	-0.000615	-0.007137
Last Notable Activity_SMS Sent	0.360233	-0.001620	0.137169	0.059445	0.052736	0.115585	-0.027600	0.016095	-0.023765	-0.001771
Last Notable Activity_Unreachable	0.037893	0.005513	0.009594	0.019415	-0.000847	0.007222	-0.003426	-0.016425	-0.003483	0.011539
Last Notable Activity_Unsubscribed	-0.016286	0.003061	0.003951	0.021668	0.018171	-0.018465	-0.004066	0.004851	-0.004133	-0.000753
Last Notable Activity_View in browser link Clicked	-0.008194	0.009819	-0.007584	0.001457	-0.011337	-0.002746	-0.000605	-0.006551	-0.000615	0.015443

We can clearly see that Converted and Total Time Spent On Website are highly correlated.

Model Building with Logistic Regression

- **Feature Selection:** RFE identified 15 important features. These features are:

  TotalVisits', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Do Not Email_Yes', 'Last Activity_Converted to Lead', 'Last Activity_Had a Phone Conversation', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Not Known', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Unreachable'

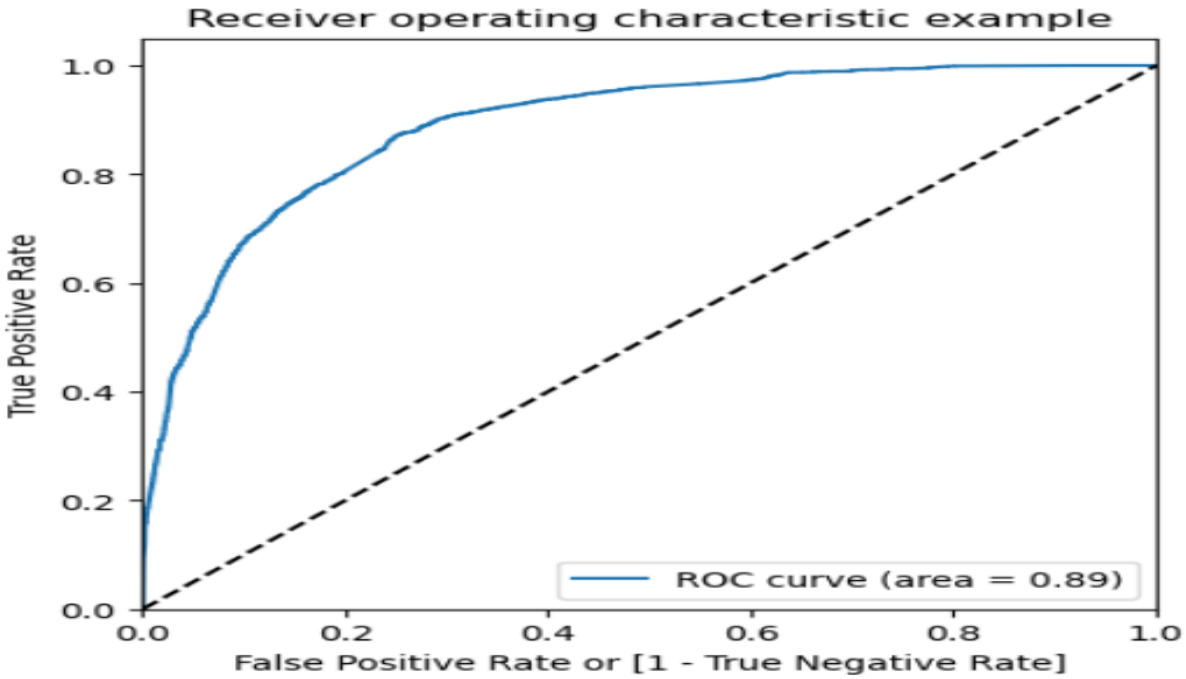
- **Model Refinement:** Removed features with p-value > 0.05 and VIF > 5.

	Features	VIF
1	Total Time Spent on Website	1.65
3	Lead Source_Olark Chat	1.52
2	Lead Origin_Lead Add Form	1.49
8	Last Activity_SMS Sent	1.47
7	Last Activity_Olark Chat Conversation	1.41
9	What is your current occupation_Not Known	1.41
0	TotalVisits	1.39
4	Lead Source_Welingak Website	1.33
10	What is your current occupation_Working Profes...	1.19
5	Do Not Email_Yes	1.07
6	Last Activity_Converted to Lead	1.04
12	Last Notable Activity_Unreachable	1.01
11	Last Notable Activity_Had a Phone Conversation	1.00

By Looking at the VIF Chart we can say that LastNotable Activity_Had a Phone Conversation is not correlated, hence multicollinearity doesn’t exist.

Model Evaluation

- **Training Results:**
  - **Accuracy:** 81.27%
  - **Confusion Matrix Metrics:**
    - Sensitivity: 0.69
    - Specificity: 0.88
  - **ROC AUC:** 0.89

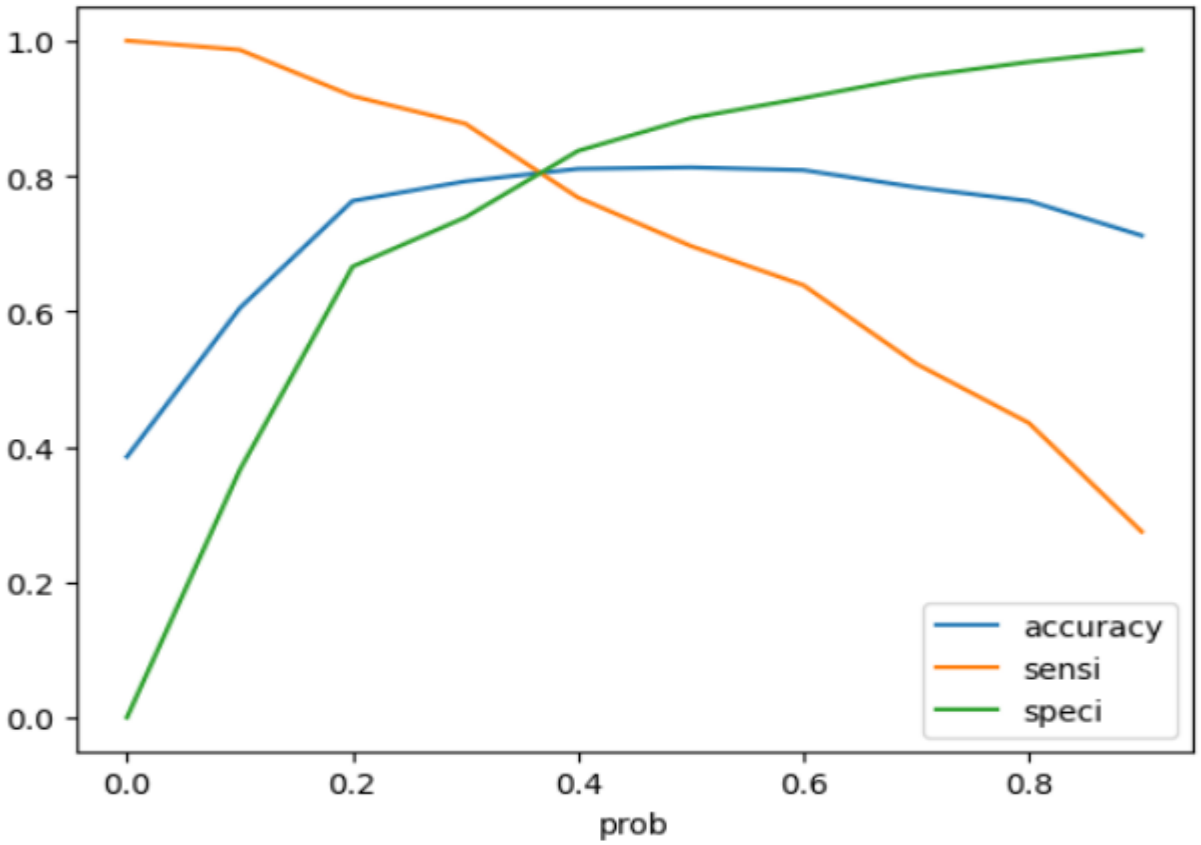


	prob	accuracy	sensi	speci
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.604629	0.986509	0.365429
0.2	0.2	0.763187	0.918234	0.666069
0.3	0.3	0.792159	0.877351	0.738796
0.4	0.4	0.810581	0.768193	0.837132
0.5	0.5	0.812785	0.696648	0.885531
0.6	0.6	0.808692	0.638594	0.915237
0.7	0.7	0.783341	0.522486	0.946735
0.8	0.8	0.763029	0.434996	0.968502
0.9	0.9	0.712014	0.274325	0.986172

This table shows how the performance of a classification model can vary depending on the probability threshold we choose.
By considering the trade-offs between accuracy, sensitivity, and specificity, you can choose the threshold that best suits your needs.

Model Evaluation

- **Optimal Probability Threshold:** 0.38 (balance between accuracy, sensitivity, and specificity).

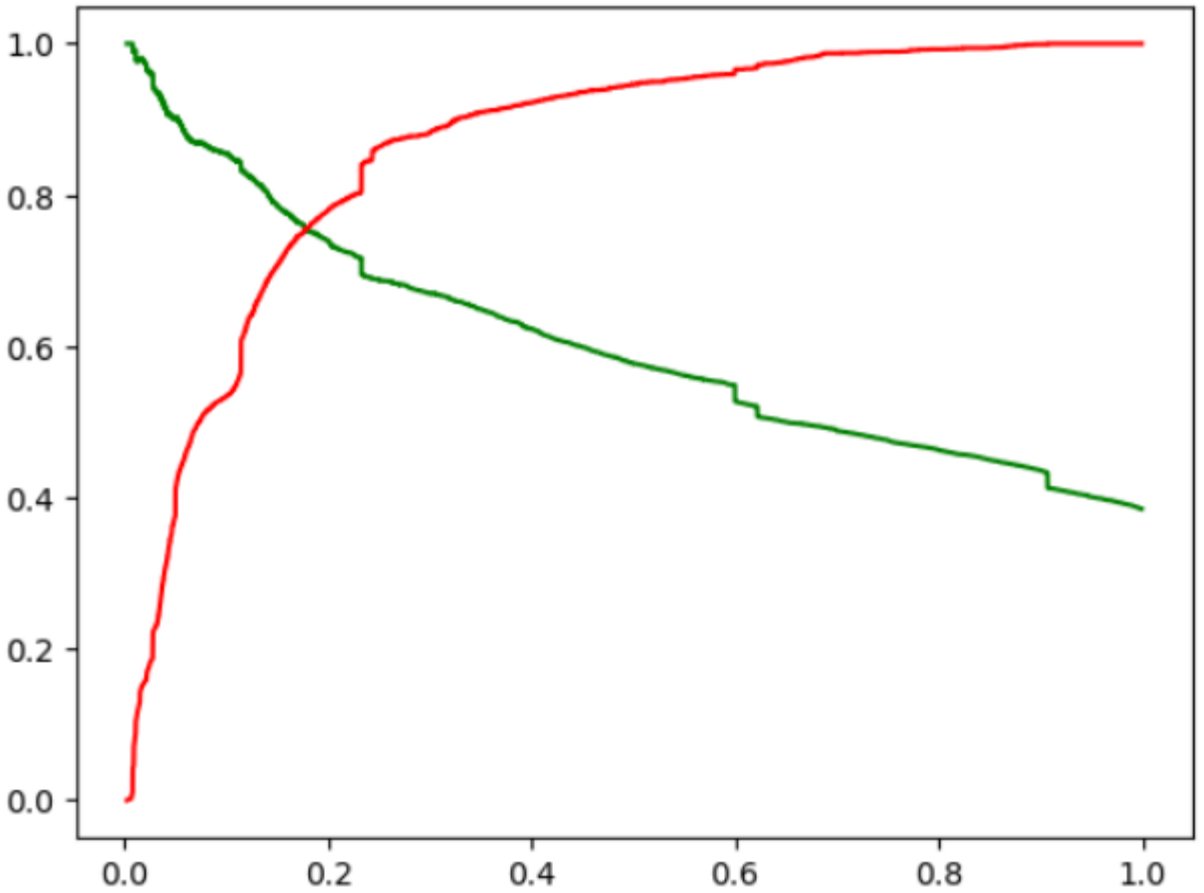


Final Model Evaluation

- **Final Model Metrics:**
  - **Accuracy:** 80.90%
  - **Confusion Matrix:**
    - Sensitivity: 0.78
    - Specificity: 0.82
  - **Precision:** 0.79
  - **Recall:** 0.69
  - **Test Data Accuracy:** 81.12%

Here the threshold between precision and recall is 0.18.

Precision and Recall Tradeoff



Overall overview

Key Insights

- **Top Features:**
  - Total Visits
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
- **Top Categorical Variables:**
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website

Strategic Recommendations

- **During Aggressive Conversion Phase:**
  - Focus on high engagement leads (high Total Visits, Time Spent on Website).
  - Target working professionals and students.
- **During Less Aggressive Phase:**
  - Avoid leads with low engagement (low visits, time on site).
  - Ignore less relevant occupations (housewives, unemployed).

Conclusion

- **Summary:**
  - Improved lead scoring model boosts conversion efficiency.
  - Strategic focus on high-potential leads can meet the desired conversion rate.
- **Next Steps:**
  - Implement the model in the sales process.
  - Continuously monitor and refine based on performance feedback.

Thank you