

# Lead Scoring Case Study

Submitted by:

- Indrani Sonar
- Thaneesh Shanand L A
- Aanand Prabhu
- Arun S





# Problem Statement

X Education, an online course provider, aims to improve its lead conversion process by identifying high-potential leads and enhancing sales efficiency.

1. **Low Lead Conversion Rate:** X Education's current lead conversion rate is approximately 30%, meaning only 30 out of 100 leads convert into paying customers. The company wants to improve this efficiency.
2. **Identifying Potential Leads (Hot Leads):** The company seeks to build a logistic regression model that can assign a lead score (0-100) to identify leads most likely to convert. This would help the sales team focus on high-potential leads.
3. **Target Conversion Rate:** The CEO expects the model to help achieve a target lead conversion rate of 80%, by better prioritizing lead engagement and communication.
4. **Lead Nurturing for Higher Conversion:** The lead conversion process requires nurturing potential leads (e.g., educating them about the product and maintaining consistent communication) to increase conversion rates through the sales funnel.



# Business Objective

**Improve Lead Conversion Efficiency:** Increase the lead conversion rate from the current **30%** by identifying and prioritizing high-potential leads (Hot Leads).

**Develop a Scoring System:** Build a logistic regression model to assign a lead score (0-100) that reflects the likelihood of a lead converting into a paying customer.

**Optimize Sales Efforts:** Enable the sales team to focus on high-potential leads by improving the lead nurturing process, leading to better communication and higher conversions.



# Data Overview

- **Dataset:** ~9,000 leads with multiple features:
  - Lead Source, Time Spent on Website, Total Visits, Last Activity, etc.
- **Target Variable:** Converted (1 = Converted, 0 = Not Converted).
- **Challenges:**
  - Missing values in attributes.
  - 'Select' in categorical variables (treated as null).
- To be cleaned and prepared for modeling.



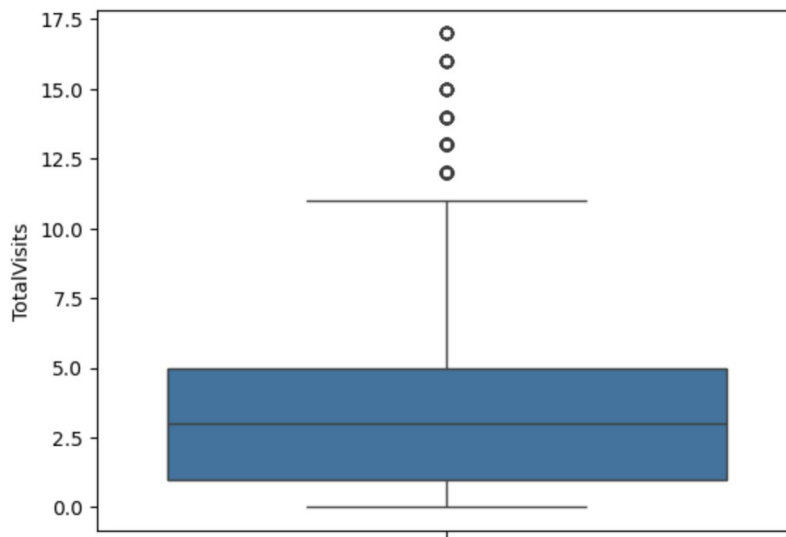
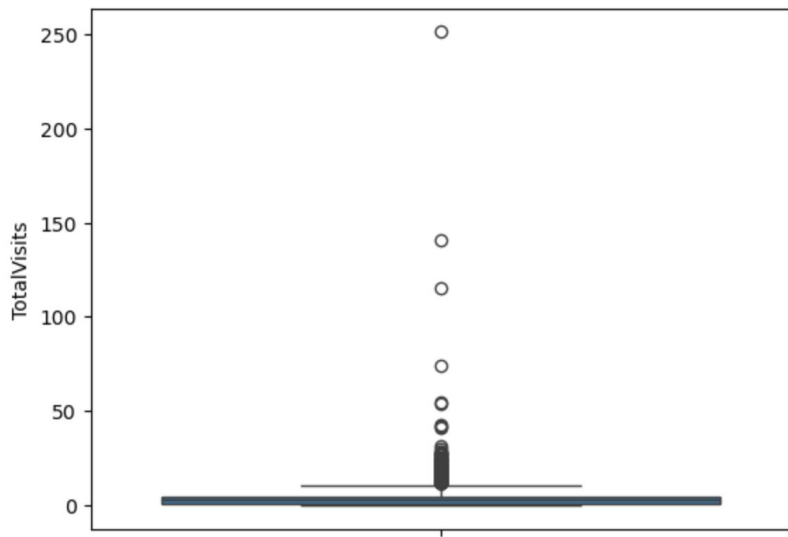
# Data Cleanup

- 'Prospect ID' column is dropped since 'Lead Number' is our unique identifier.
- Dropped columns where missing values are more than 40%.
- Certain columns have 'Select' values which are changed to 'Unknown' and are handled as missing values.
- Dropped columns again where the count of 'Unknown' values are more than 40% since it is considered as missing values.



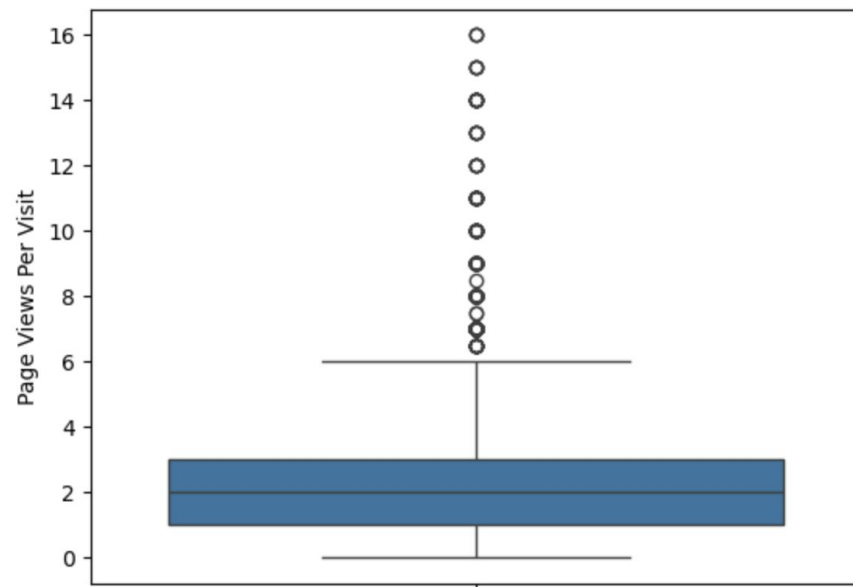
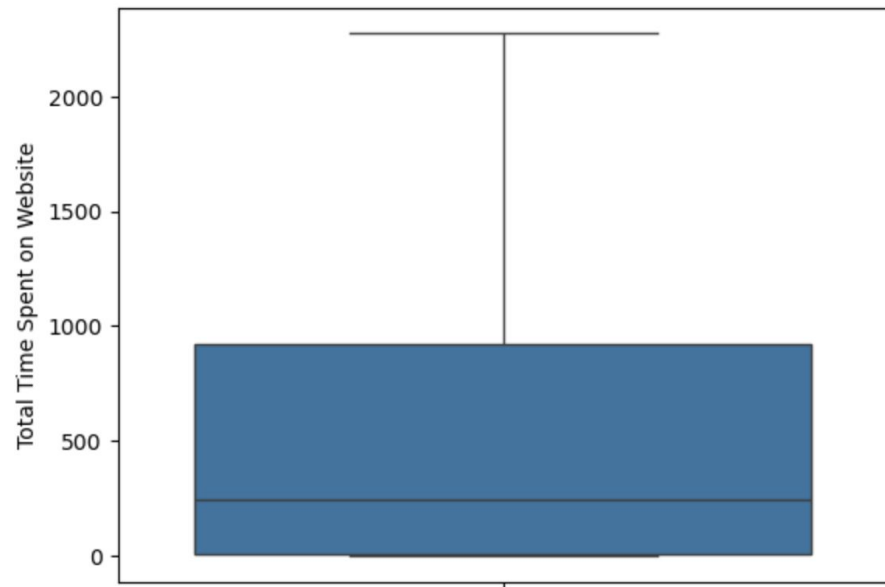
# Data Preparation - Outliers handling

- 'TotalVisits' variable had some outliers. These were removed above the 99 percentile.





# Outliers Handling



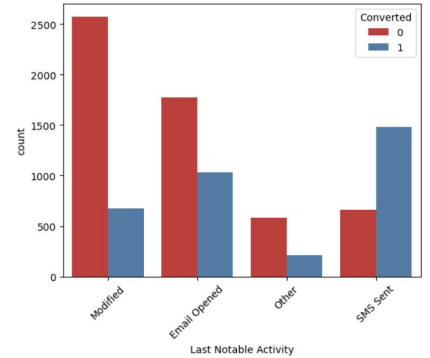
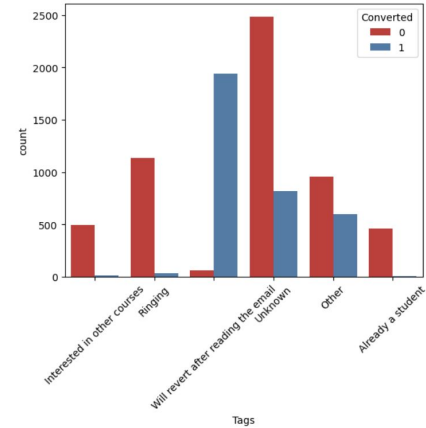
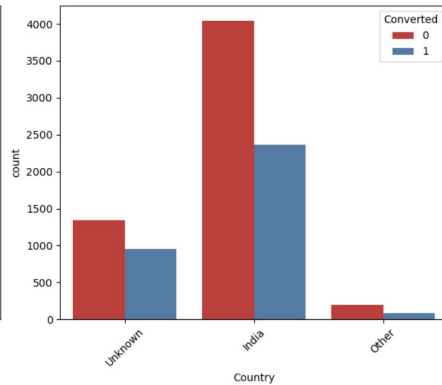
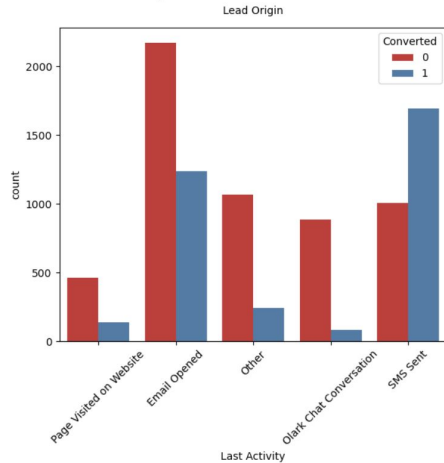
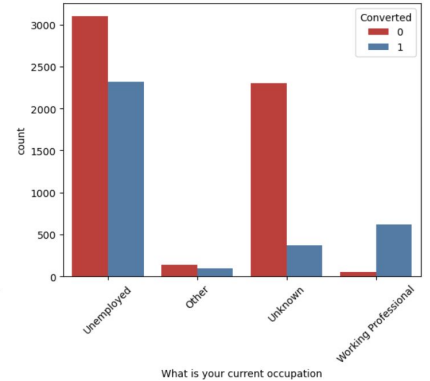
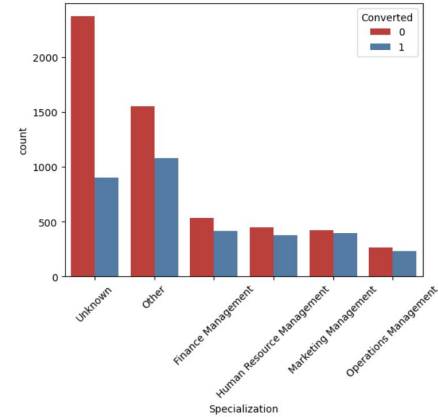
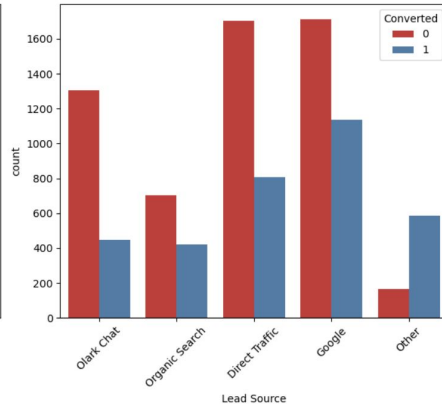
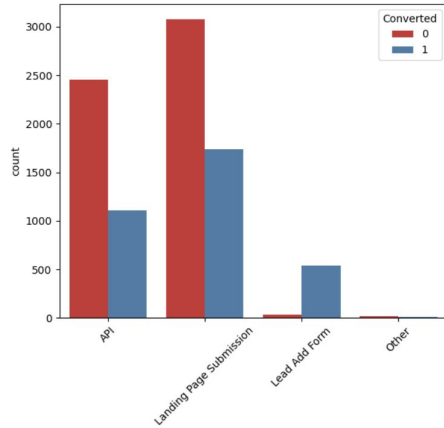


# Categorical variables preparation

- Columns have 'Yes' or 'No' values are replaced with 1 and 0.
- Many of the categorical variables were having categories that are infrequent. If the percentage is less than the threshold of 5%, we are grouping them a 'Other' categories as they wouldn't add much value to the analysis.
- Dummy variables are created all the categorical variables. The categories with the '\_Other' prefix are dropped to make the data relevant. The original columns are also removed to remove duplicates.



# Categorical variables after grouping to Other category





# Model building summary

Using RFE, the following features have been identified, and below is the model summary

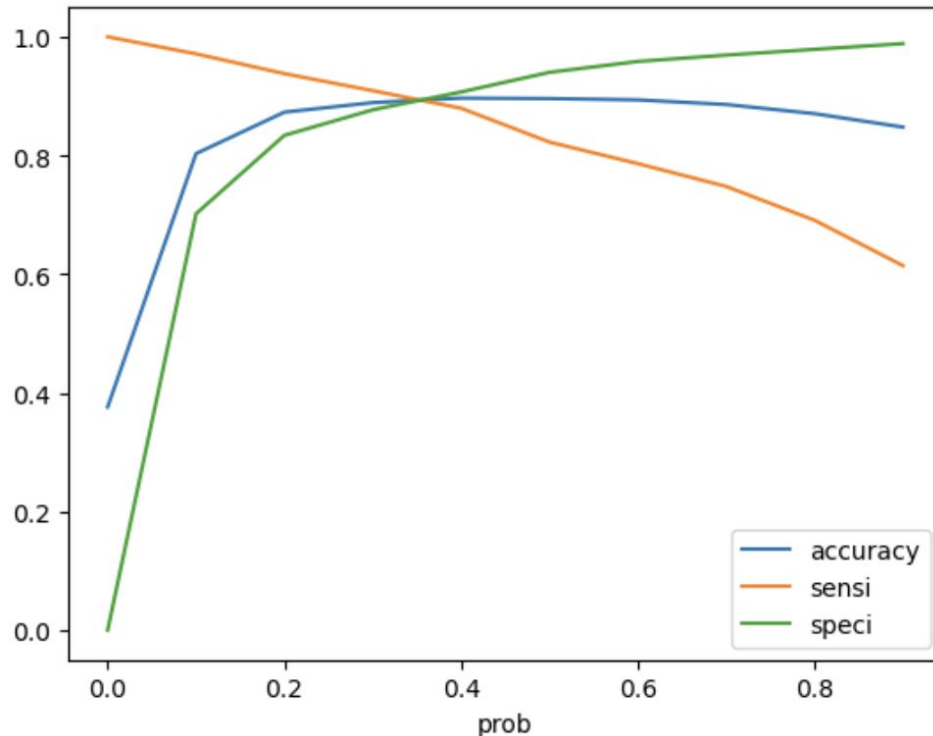
	Features	VIF
5	Country_Unknown	1.96
11	Tags_Will revert after reading the email	1.89
7	CourseGoal_Unknown	1.83
4	LastActivity_SMS Sent	1.72
3	LastActivity_Email Opened	1.71
12	LastNotableActivity_Modified	1.62
2	LeadOrigin_Lead Add Form	1.45
1	Total Time Spent on Website	1.42
10	Tags_Ringing	1.35
6	CurrentOccupation_Working Professional	1.29
9	Tags_Interested in other courses	1.18
8	Tags_Already a student	1.13
0	Do Not Email	1.12

Generalized Linear Model Regression Results							
Dep. Variable:	Converted	No. Observations:	6293				
Model:	GLM	Df Residuals:	6279				
Model Family:	Binomial	Df Model:	13				
Link Function:	Logit	Scale:	1.0000				
Method:	IRLS	Log-Likelihood:	-1546.3				
Date:	Tue, 17 Dec 2024	Deviance:	3092.7				
Time:	20:11:11	Pearson chi2:	7.69e+03				
No. Iterations:	9	Pseudo R-squ. (CS):	0.5652				
Covariance Type:	nonrobust						
		coef	std err	z	P> z	[0.025	0.975]
	const	-1.1708	0.134	-8.762	0.000	-1.433	-0.909
	Do Not Email	-1.4677	0.224	-6.566	0.000	-1.906	-1.030
	Total Time Spent on Website	1.1826	0.055	21.337	0.000	1.074	1.291
	LeadOrigin_Lead Add Form	3.0270	0.302	10.035	0.000	2.436	3.618
	LastActivity_Email Opened	0.4180	0.130	3.227	0.001	0.164	0.672
	LastActivity_SMS Sent	1.7340	0.132	13.139	0.000	1.475	1.993
	Country_Unknown	1.3922	0.135	10.332	0.000	1.128	1.656
	CurrentOccupation_Working Professional	1.3303	0.288	4.623	0.000	0.766	1.894
	CourseGoal_Unknown	-1.4200	0.104	-13.649	0.000	-1.624	-1.216
	Tags_Already a student	-5.2729	1.017	-5.186	0.000	-7.266	-3.280
	Tags_Interested in other courses	-2.9827	0.346	-8.629	0.000	-3.660	-2.305
	Tags_Ringing	-4.0646	0.240	-16.954	0.000	-4.535	-3.595
	Tags_Will revert after reading the email	3.3002	0.182	18.144	0.000	2.944	3.657
	LastNotableActivity_Modified	-0.7037	0.109	-6.464	0.000	-0.917	-0.490



# Plotting the ROC curve

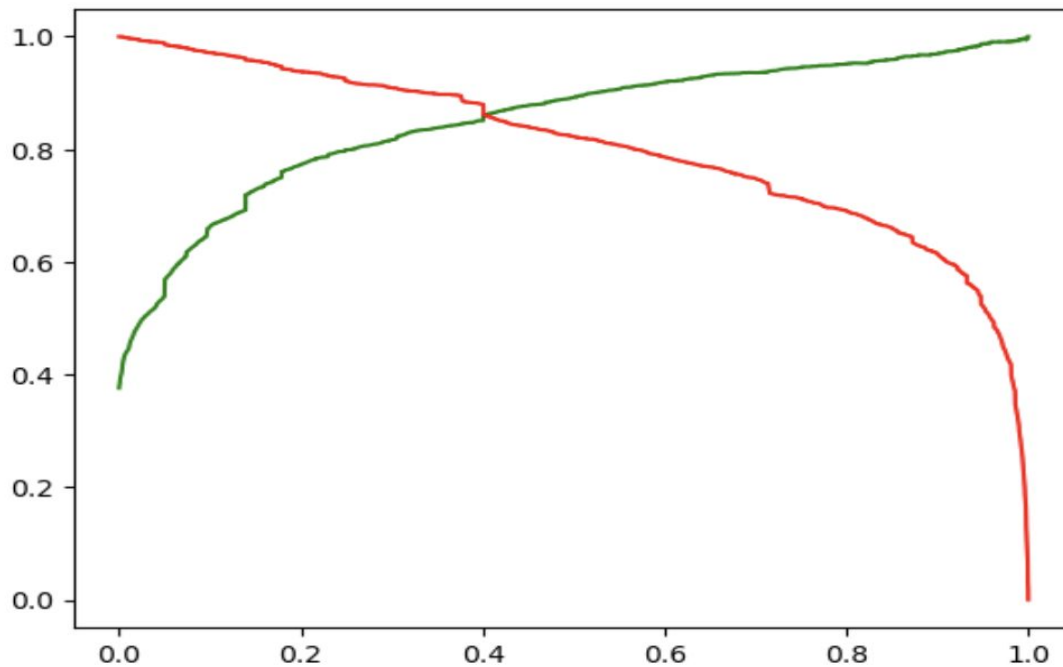
- Using ROC curve, we are finding a cutoff value of 0.35. With this cutoff, we are getting below metrics on the training data.
- Accuracy - 89.62%
- Sensitivity - 89.78%
- Specificity - 89.53%





# Precision and recall trade-off

Precision and recall have a meeting point at around 0.4.





# Final metrics

## Train data metrics

- Accuracy: 89.62%
- Sensitivity: 89.78%
- Specificity: 89.53%
- False positive rate: 10.47%
- Positive predictive value: 83.79%
- Negative predictive value: 93.56%

## Test data metrics

- Accuracy: 88.88%
- Sensitivity: 84.69%
- Specificity: 91.48%
- False positive rate: 8.52%
- Positive predictive value: 86.02%
- Negative predictive value: 90.61%



# Key Insights

- Top Variables Contributing to Lead Conversion (based on coefficients and RFE selection):
  - Total Time Spent on Website
  - LastActivity\_SMS Sent
  - LeadOrigin\_Lead Add Form
- Model Performance
  - The model achieved consistent accuracy, sensitivity, and specificity on both train and test data.
  - Sensitivity on test data (84.69%) meets the business requirement of identifying most potential leads for conversion.
- Lead Scores
  - Higher probabilities (>35% threshold) are mapped to higher lead scores, enabling targeted follow-ups.