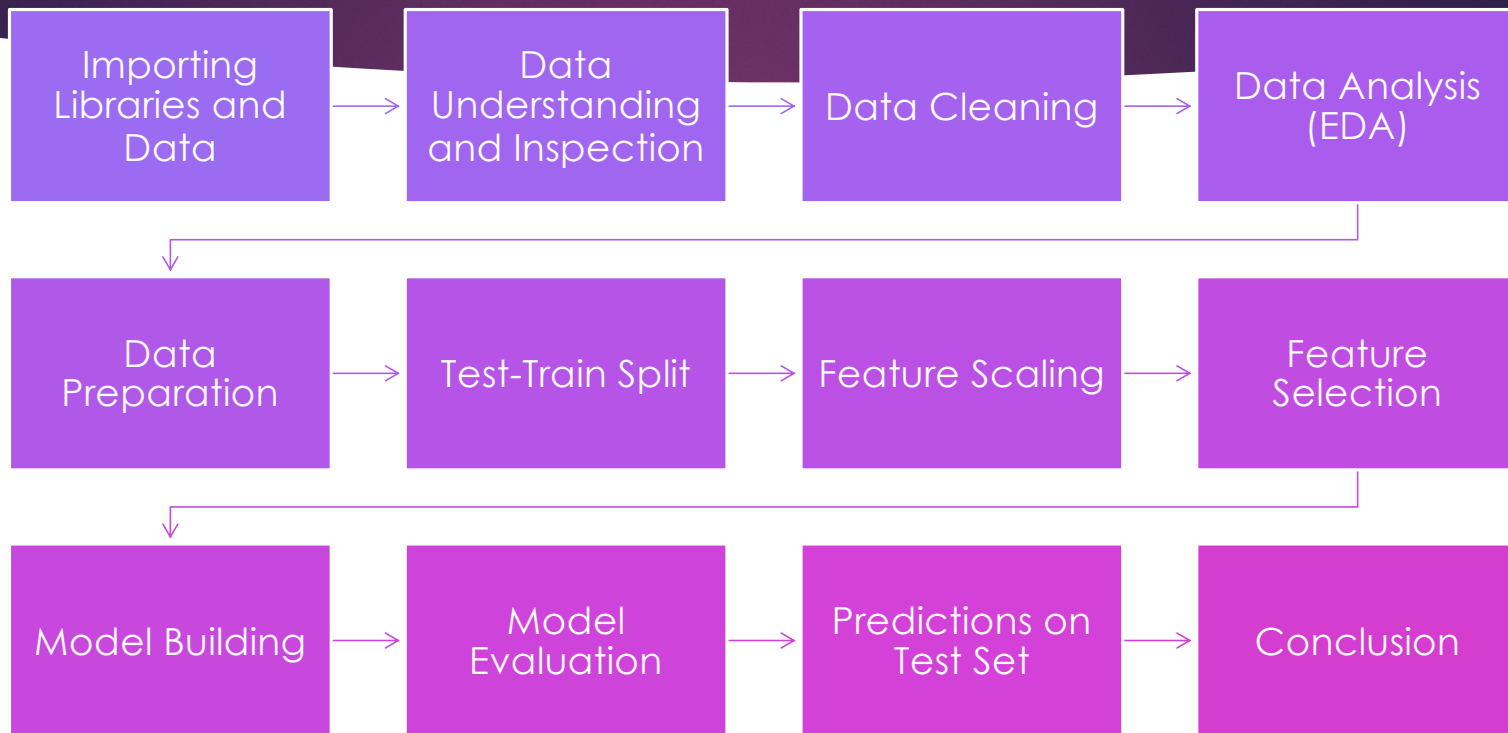# Lead Scoring Case Study

Problem Statement

▶ X Education aims to improve lead conversion rates in its online courses business by developing a lead scoring model. The goal is to identify promising leads with a higher likelihood of conversion.

▶ This involves utilizing historical data, model training, setting conversion thresholds, and ongoing refinement.

▶ Current Lead conversion rate is 38.53%

Solution Methodology

Importing Libraries and Data	→	Data Understanding and Inspection	→	Data Cleaning	→	Data Analysis (EDA)

Data Preparation	→	Test-Train Split	→	Feature Scaling	→	Feature Selection

Model Building	→	Model Evaluation	→	Predictions on Test Set	→	Conclusion

Data Cleaning

Actions Performed:

▶ Removed columns with >= 40% missing values

▶ Treated data skewness

▶ Imputed missing values

▶ Outliers treatment

Outcome: No. of columns reduced from 37 to 11

Data Analysis (EDA)

Univariate Analysis

Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

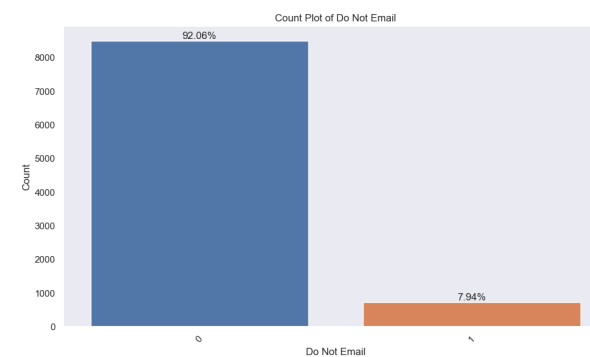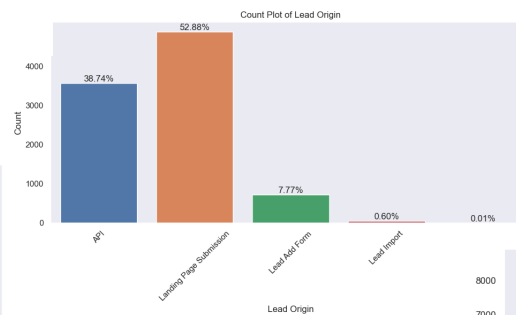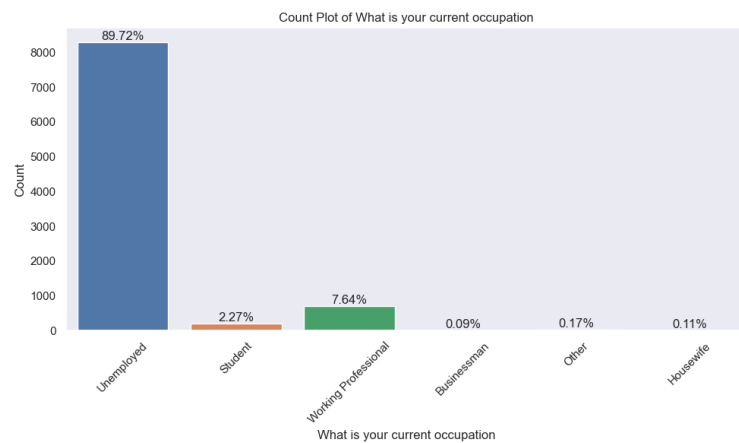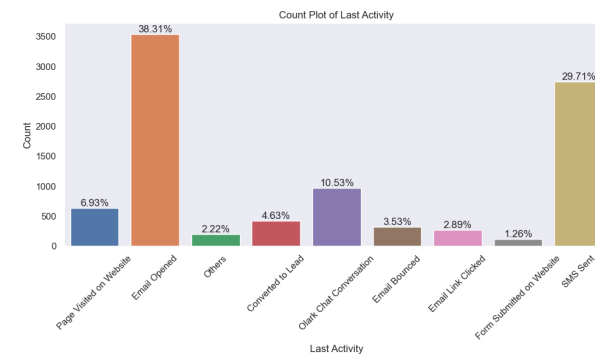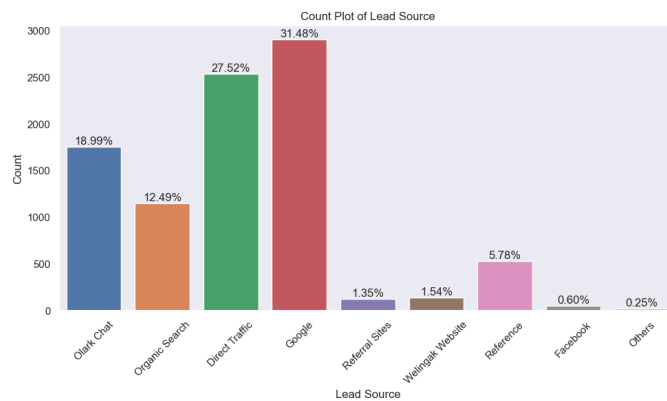**Lead Source:** 58% Lead source is from Google & Direct Traffic combined

Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

Current_occupation: It has 90% of the customers as Unemployed

Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

Count Plot of Lead Source

Lead Source values:
- Olark Chat: 18.99%
- Organic Search: 12.49%
- Direct Traffic: 27.52%
- Google: 31.48%
- Referral Sites: 1.35%
- Welingak Website: 1.54%
- Reference: 5.78%
- Facebook: 0.60%
- Others: 0.25%

Count Plot of Last Activity

Last Activity values:
- Page Visited on Website: 6.93%
- Email Opened: 38.31%
- Others: 2.22%
- Converted to Lead: 4.63%
- Olark Chat Conversation: 10.53%
- Email Bounced: 3.53%
- Email Link Clicked: 2.89%
- Form Submitted on Website: 1.26%
- SMS Sent: 29.71%

Count Plot of Lead Origin

Lead Origin values:
- API: 38.74%
- Landing Page Submission: 52.88%
- Lead Add Form: 7.77%
- Lead Import: 0.60%
- (last): 0.01%

Count Plot of What is your current occupation

What is your current occupation values:
- Unemployed: 89.72%
- Student: 2.27%
- Working Professional: 7.64%
- Businessman: 0.09%
- Other: 0.17%
- Housewife: 0.11%

Count Plot of Do Not Email

Do Not Email values:
- 0: 92.06%
- 1: 7.94%

Bivariate Analysis

▶ **Lead Origin:**

• About 53% of leads stem from "Landing Page Submission," boasting a conversion rate of 36%.

• The "API" accounts for approximately 39% of customers, showing a conversion rate of 31%.

▶ **Current Occupation:**

• Approximately 90% of customers fall under the "Unemployed" category, with a conversion rate of 34%.

• Despite constituting only 7.6% of the total customer base, "Working Professionals" exhibit an impressive 92% conversion rate.

▶ **Do Not Email:**

• A significant 92% of individuals have chosen not to receive email communications regarding the course.
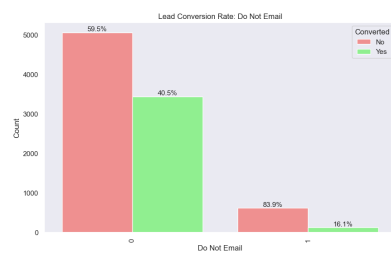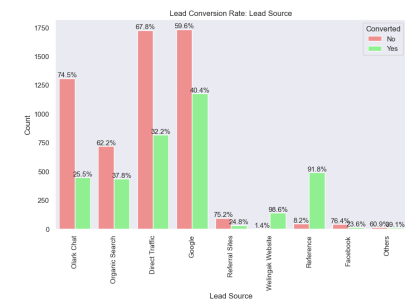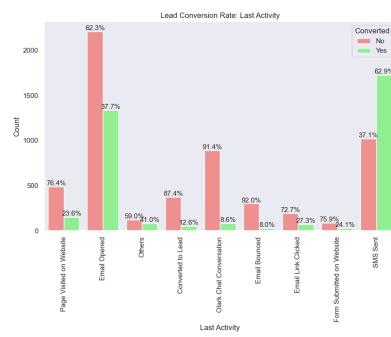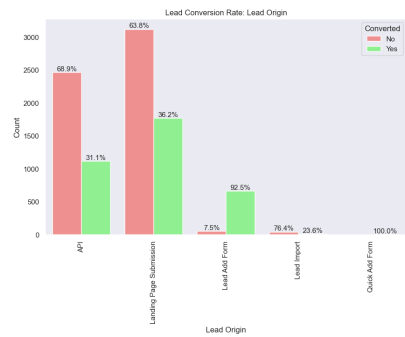
▶ **Lead Source:**

• "Google" yields a conversion rate of 40% among the 31% of customers from this source.

• "Direct Traffic" contributes a lower conversion rate of 32% with a customer percentage of 27%.

• Although "Organic Search" contributes to 37.8% of the conversion rate, it represents only 12.5% of the customer base.

• "Reference" showcases a remarkable conversion rate of 91%, yet it comprises merely around 6% of the customer acquisition.
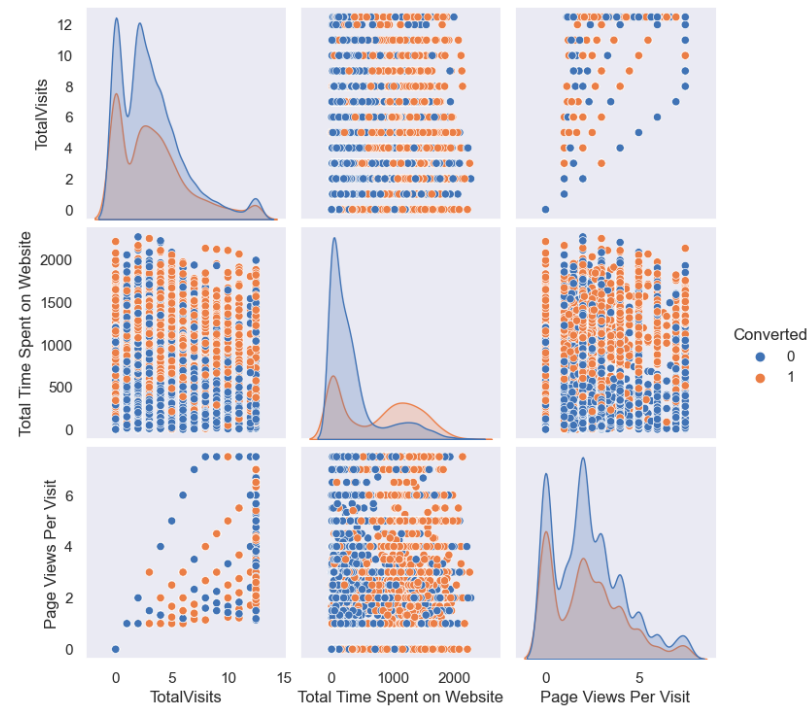
▶ **Last Activity:**

• The act of sending an "SMS" displays a notably high conversion rate of 63%, driven by 30% of last activities.

• "Email Opened" encompasses 38% of the customer's recent interactions, accompanied by a conversion rate of 37%.

▶ **Specialization:**

• "Marketing Management," "HR Management," and "Finance Management" emerge as prominent contributors.

Lead Conversion Rate: Lead Origin

Lead Conversion Rate: Last Activity

Lead Conversion Rate: What is your current occupation

Lead Conversion Rate: Lead Source

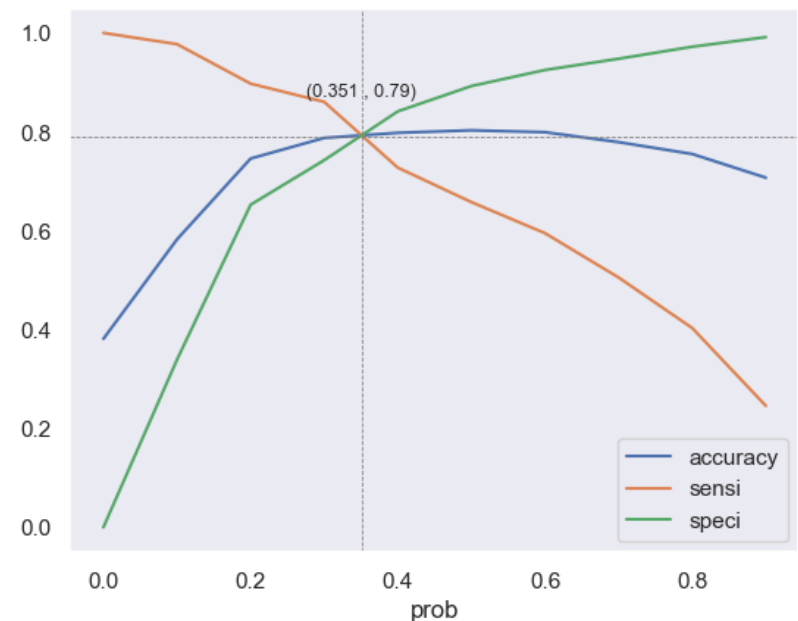Lead Conversion Rate: Do Not Email
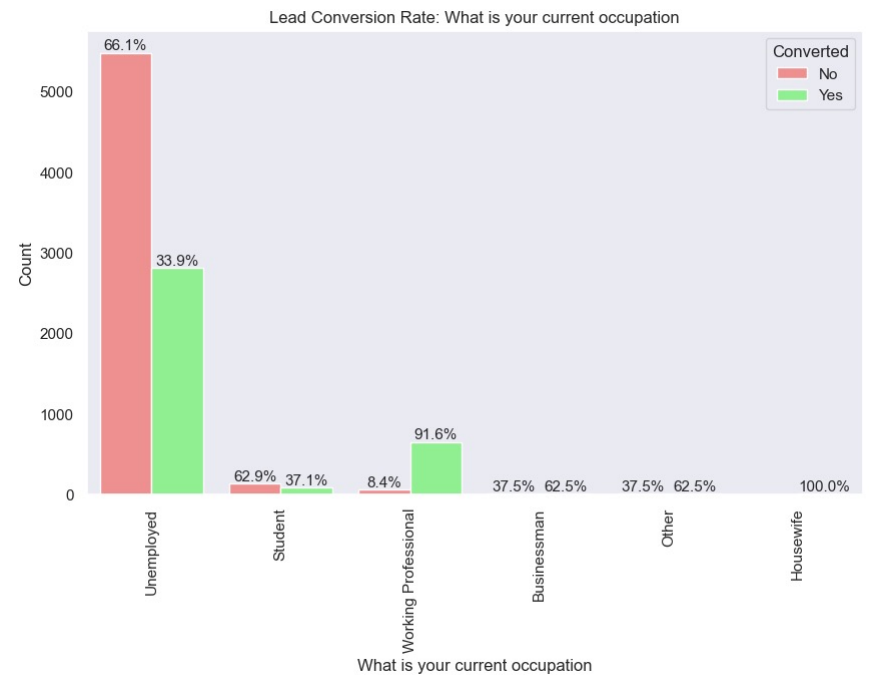
Model Evaluation

Sensitivity and Specificity on Train Set

0.351 is the approx. point where all the curves meet, so 0.351 seems to be our Optimal cutoff point for probability threshold

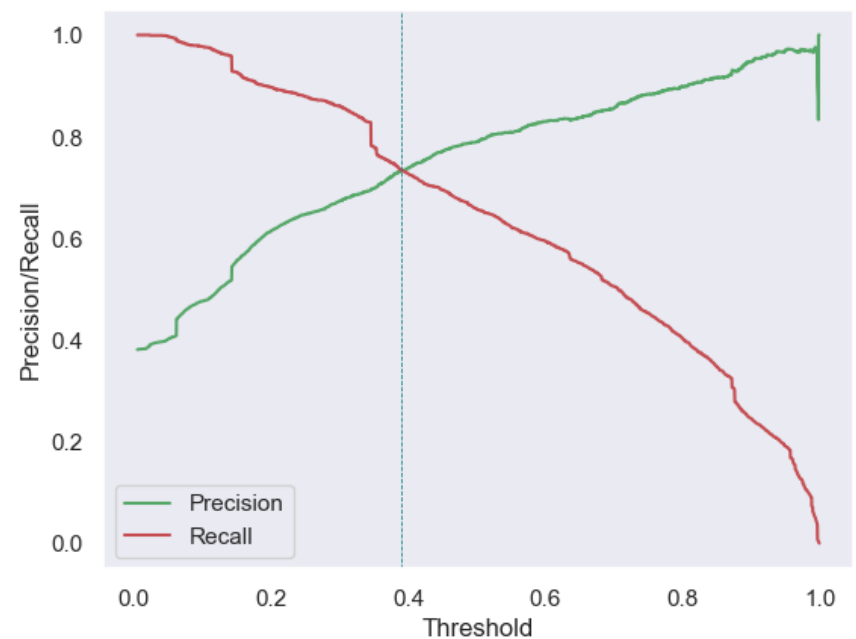| Actual\Predicted	Not Converted	Converted
Not Converted	3174	828
Converted	544	1922

Model Evaluation

Precision and Recall on Train Set

The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimise the performance of the model based on business requirement.
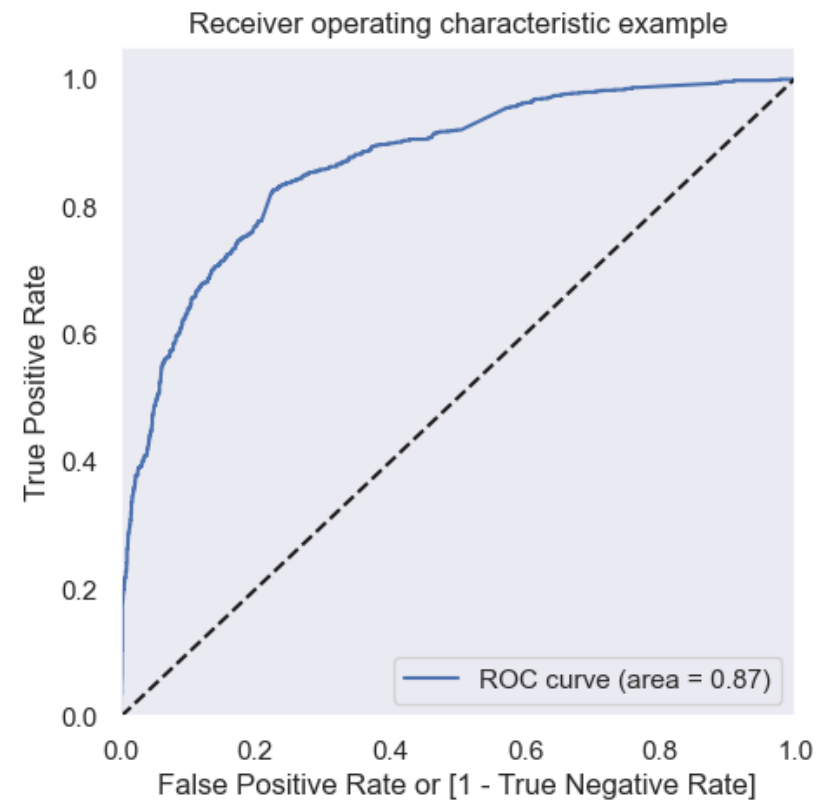
Actual\Predicted	Not Converted	Converted
Not Converted	3341	661
Converted	656	1810

Prediction on Test Set

Area under ROC curve is 0.87 out of 1 which indicates a good predictive model

Actual\Predicted	Not Converted	Converted
Not Converted	1340	337
Converted	253	842

Receiver operating characteristic example

Conclusion

▶ Model performed consistently across different metrics in both train and test datasets.

▶ Models sensitivity is close to the target of 80% set by the CEO.

	Train Set	Test Set
Accuracy	78.79%	78.72%
Sensitivity	77.9%	76.89%
Specificity	79.31%	79.99%

Top 3 Features Contributing to Predicting Hot Leads:

Lead Source_Welingak Website
This source significantly impacts lead conversion.

Total Time Spent on Website
Longer engagement positively correlates with conversion.

Occupation_Working Professional
Leads with this occupation show a higher likelihood of conversion.

Business Recommendations

To Increase Lead Conversion Rates:

1. Prioritize features with positive coefficients to enhance targeted marketing precision.

2. Focus on top-performing lead sources to attract high-quality leads.

3. Tailor communication to connect effectively with employed professionals.

4. Optimize communication channels based on lead engagement impact.

5. Allocate budget for activities on the Welingak Website to leverage its influence.

6. Introduce incentives for successful references to boost lead submission.

7. Target working professionals due to their high conversion rates and better financial capacity.

To Identify Areas of Improvement:

1. Evaluate specialization offerings with negative coefficients.

2. Review the landing page submission process for potential enhancements.

The developed logistic regression model with 11 features proved effective in identifying promising leads for conversion. By focusing on the recommended strategies and refining the model based on ongoing data, X Education can expect to achieve its goal of improving lead conversion rates.