Lead Scoring Model

X Education Case Study Using Logistic Regression

Presented by : Manisha Singh

(*Data Analyst*
Machine Learning Enthusiast)

Problem Statement & Business Objective

- **Problem:**

X Education is an online education company that offers industry-relevant courses to working professionals. Despite acquiring numerous leads daily through multiple digital channels, the **lead-to-conversion rate remains low (~30%)**. This results in inefficient sales efforts and high operational costs.

- **Business Objective**

To improve the conversion rate by identifying **'Hot Leads'**, i.e., leads most likely to convert into paying customers, so that the sales team can **prioritize outreach and resources** more effectively.

- **Goal**

Develop a **Logistic Regression-based Lead Scoring Model** to:

- Assign a **Lead Score (0–100)** to each lead.
- Predict the **probability of conversion**. Help the sales team focus only on **high-potential leads**, increasing efficiency and revenue.

Approach & Methodology

- **Data Understanding & Cleaning**
- Explored ~9,000 historical lead records.
- Identified and handled missing values & inconsistencies.
- Dropped redundant features and imputed nulls.
- Removed non-informative values like "Select" in categorical fields.

- **Feature Engineering**
- Converted categorical variables into dummy/indicator variables.
- Scaled numeric features using **StandardScaler**.
- Ensured all variables were in a suitable format for modeling.

Model Development

- Chose **Logistic Regression** for binary classification (Converted vs Not Converted).
- Addressed class imbalance using **class_weight='balanced'**.
- Split data into 70% training and 30% testing sets with stratification.

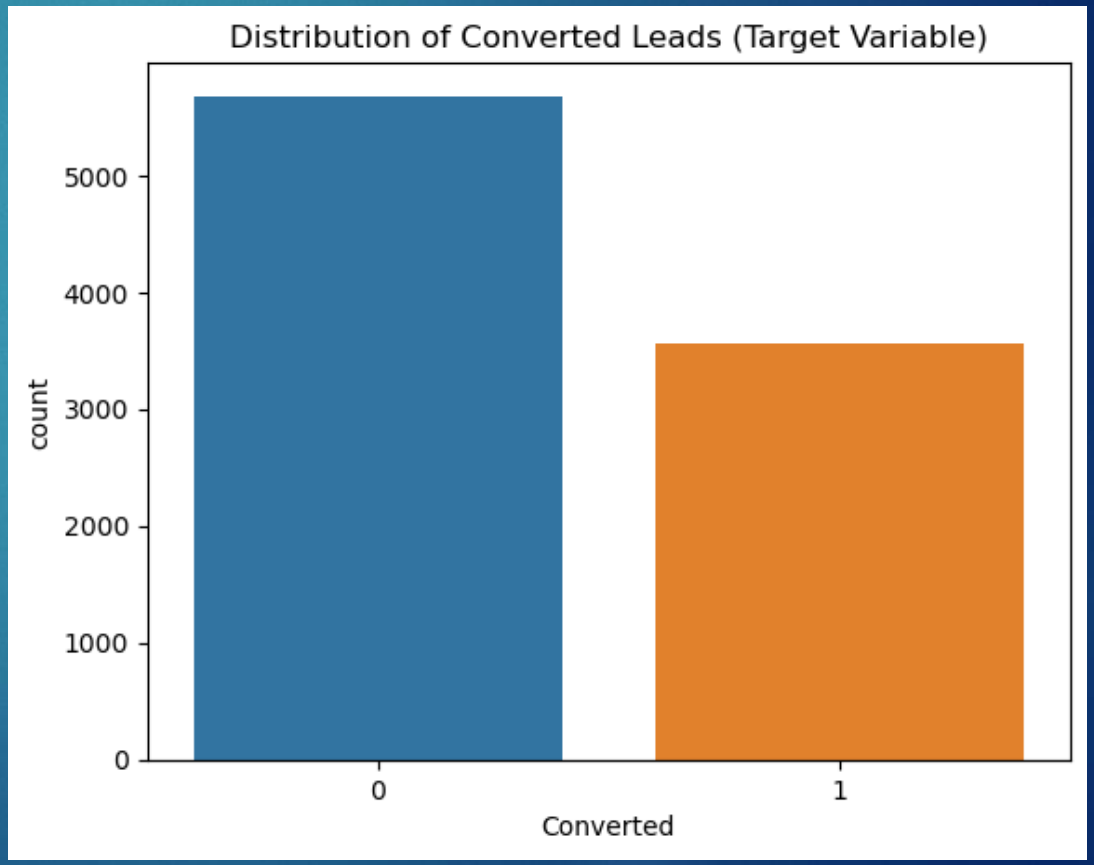
Model Evaluation & Interpretation

- Evaluated model using **Accuracy, Precision, Recall, F1 Score**, and **ROC AUC**.
- Generated **Confusion Matrix**, **ROC Curve**, and **Lead Scores**.
- Identified top influencing variables from model coefficients.

Model Performance & Evaluation

- **Lead Conversion Distribution**

- Class 0 (Not Converted): 61%
- Class 1 (Converted): 39%



- **Confusion Matrix**

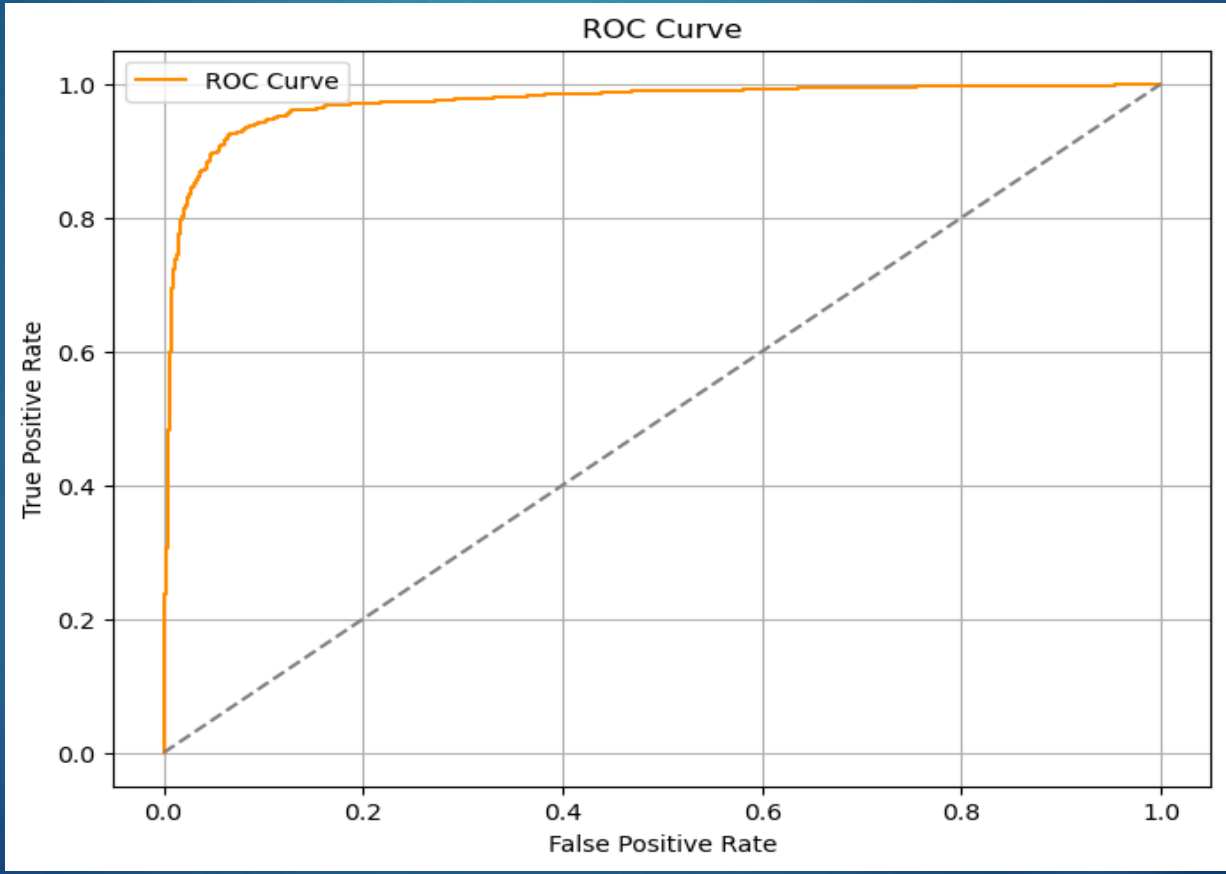
	Predicted: 0	**Predicted: 1**
Actual: 0	**914**	64
Actual: 1	61	**702**

- **Interpretation:**

  - **True Negatives (TN)** = 914
  - **False Positives (FP)** = 64
  - **False Negatives (FN)** = 61
  - **True Positives (TP)** = 702

Logistic Regression Performance

- Accuracy: 92.82%
- Precision: 91.64%
- Recall: 92.01%
- F1 Score: 91.82%
- ROC AUC Score: 97.18%



Classification Report Summary

Metric	Class 0 (Not Converted)	Class 1 (Converted)
Precision	94%	92%
Recall	93%	92%
F1-Score	94%	92%
Support	978	763

Interpretation

- High precision and recall reflect effective identification of hot leads.
- Balanced classification helps reduce false positives and negatives.
- Lead scoring allows prioritizing high-conversion potential leads.

Top Influencing Features

- **Top 3 Contributing Features**

Feature Name	Coefficient
Tags_Will revert after reading the email	**1.89**
Tags_Closed by Horizzon	**1.47**
Tags_Ringing *(Negative Influence)*	**-1.32**

- **Analytical Interpretation**:

  - **Positive Coefficients** imply higher likelihood of lead conversion.
  - Leads tagged "Will revert after reading the email" and "Closed by Horizzon" demonstrate strong conversion potential.
  - **Negative Coefficient** (e.g., "Ringing") suggests lower engagement and lower conversion probability.

Business Recommendations

- The sales team operate during the **intern hiring period** (2 months)

Lower the classification threshold from **0.5 to ~0.3** to increase recall. This ensures almost all potential leads (including borderline cases

Assumption: Interns are available to handle higher outreach volume.) are captured and contacted.

Outcome: More potential leads contacted → Higher chances of conversion → Better intern utilization

- The team act after the **quarterly target is achieved**

Increase the threshold to **0.7 or higher** to prioritize only the most likely conversions, thus reducing unnecessary outreach.

Assumption: The goal is to conserve resources and focus on only high-certainty leads.

Outcome: Minimized effort → Reduced unnecessary calls → Focus on leads with high ROI.

- **lead characteristics (features)** should be prioritized in communication

Focus on leads tagged as:

"Will revert after reading the email"
"Closed by Horizzon"
These had the **highest positive coefficients**, indicating strong conversion potential.
Avoid low-impact or negative tags like **"Ringing"** unless necessary.
Outcome: Smarter segmentation → Targeted messaging → Better use of sales bandwidth.

Summary Report:

This project aimed to build a predictive model to help X Education convert leads more effectively. We performed data cleaning, exploratory data analysis, and logistic regression modeling using balanced class weights.

The model achieved strong results with 92.8% accuracy and a ROC AUC of 97.18%. Key features contributing to lead conversion include specific user engagement tags. We offered two business strategies: lowering the prediction threshold during intern-heavy periods to cast a wider net, and raising it when targets are already met to reduce effort.

Lessons learned included the importance of balancing classes, interpreting model coefficients, and aligning data science with practical sales decisions.

Thank you!