Lead Scoring Case Study – Summary Report

1. Problem Overview

X Education faces a low lead conversion rate (~30%) despite receiving a high number of leads. The goal is to:
- Identify high-potential leads using data.
- Prioritize sales outreach efficiently.
- Ultimately improve the overall conversion rate.

2. Dataset Overview

- Total Rows: 9240
- Total Features: 37
- Target Variable: Converted (1 = lead converted, 0 = not converted)

Key Types of Features:
- Demographic & behavioral data (e.g., TotalVisits, Time Spent)
- Categorical (e.g., Lead Source, Last Activity)
- Binary (e.g., Do Not Email, A free copy of Mastering The Interview)

3. Data Cleaning & Preprocessing

- Replaced 'Yes'/'No' binary values with 1/0.
- Replaced 'Select' and NaN values in key columns appropriately.
- Dropped features with > 40% missing values.
- Imputed or dropped other columns based on data sparsity and business relevance.

4. Feature Engineering

- Converted categorical variables into dummy variables.
- Retained key numeric predictors: TotalVisits, Total Time Spent, Page Views Per Visit.

5. Exploratory Data Analysis

- High time spent on website and more visits correlate with higher conversions.
- Certain Lead Sources and Last Activities show stronger conversion potential.
- Conversion rate: ~38%

6. Model Building

- Model: Logistic Regression
- Feature Selection via RFE, followed by VIF and p-value filtering.
- Final model includes key numeric and categorical variables.

7. Model Evaluation

- Metrics: Accuracy, Precision, Recall, ROC-AUC
- Threshold tuning strategy to maximize recall or precision based on business context.

8. Key Drivers of Conversion

Top Numeric Features:
1. Total Time Spent on Website
2. Page Views Per Visit
3. TotalVisits

Top Categorical/Dummy Features:
1. Last Activity – Email Opened, Olark Chat

2. Lead Source – Google, Direct Traffic
3. Specialization – Finance, IT, HR

9. Strategic Recommendations

During Aggressive Sales Phases:
- Lower threshold to increase recall.
- Nurture medium-potential leads with follow-ups.

During Conservative Phases:
- Raise threshold to maximize precision.
- Focus on top-scoring leads.
- Reallocate sales team to strategic efforts.

Conclusion

The model effectively scores leads based on their likelihood of conversion, helping X Education:
- Streamline outreach
- Save time and resources
- Improve conversion rate

The model is a strong foundation for data-driven sales prioritization.