# **Lead Scoring Model for X Education – Summary Report**

#### I. Problem Statement

X Education, an online course provider, faces a low lead conversion rate (~30%) despite generating a large number of leads daily. To improve sales efficiency, we developed a Lead Scoring Model using Logistic Regression, assigning scores from 0 to 100 to help prioritize leads. The goal is to increase conversion rates to around 80% by focusing on high-potential leads.

### II. Data Preprocessing & Exploration

The dataset included categorical and numerical features related to user engagement and lead sources. The key preprocessing steps were:

## 1. Handling Missing Data:

- Columns with >45% missing values were dropped if they lacked significance (e.g., Asymmetrique Activity/Profile Index & Score).
- Some missing values were replaced with mode or categorized as 'unknown' (e.g., Lead Quality was retained despite 51% missing values).
- 'Select' responses were treated as missing values and replace by 'unknown' value.

#### 2. Feature Selection:

- Columns with only one value (e.g., Magazine, Update me on Supply Chain Content) were dropped.
- Low-frequency categories were grouped under 'Other' to reduce dimensionality.

## 3. Categorical Encoding & Scaling:

- One-hot encoding was applied to categorical variables.
- Min-Max scaling was used for numerical variables.

#### 4. Outlier Treatment:

Boxplots revealed outliers in Total Visits and Page Views Per Visit, but they were retained as they were business-relevant.

## 5. Exploratory Analysis:

Univariate and bivariate analysis helped identify key patterns and relationships with conversion rates.

#### III. Model Building

A Logistic Regression Model was chosen for its interpretability and efficiency in binary classification.

## 1. Key Steps:

- Recursive Feature Elimination (RFE): Selected the most relevant predictors.
- Variance Inflation Factor (VIF) Analysis: Removed highly correlated variables to reduce multicollinearity.
- Optimal Cut-off Selection: The Sensitivity-Specificity Intersection was chosen as the cut-off to balance lead conversion (sensitivity) and avoiding unnecessary sales calls (specificity).

#### 2. Model Performance:

Metric	Train Set	Test Set
Accuracy	91.30%	90.40%
Sensitivity (Recall)	91.80%	91.70%
Specificity	91%	89.70%
<b>False Positive Rate</b>	9%	10%
<b>Positive Predictive Value</b>	86.50%	83%
Negative Predictive Value	94.60%	95.10%

## IV. Lead Scoring & Implementation

Once finalized, the model was applied to assign lead scores across the dataset:

#### 1. Lead Score Calculation:

The model's probabilities were scaled between 0 to 100.

A higher score = "hot lead" (likely to convert).

A lower score = "cold lead" (unlikely to convert).

#### 2. Final Model Validation:

The model performed consistently on both training and test sets, making it deployment-ready.

## V. Key Learnings & Takeaways

- Data Preprocessing is Critical: Handling missing data, encoding, and feature selection significantly impacted model quality.
- Feature Engineering Improves Performance: Grouping rare categories and removing redundant variables enhanced interpretability.
- Threshold Selection Matters: Choosing the right cut-off impacts precision and recall, affecting business decisions.
- Model Validation is Essential: Ensuring consistent performance on training and test sets prevents overfitting.

#### VI. Conclusion

This Lead Scoring Model provides data-driven insights to help X Education prioritize high-conversion leads, optimize sales efforts, and improve overall business growth. The final model is ready for deployment, enabling the sales team to maximize efficiency and increase revenue effectively.