Summary Report – Khanh - Ilan - Anubha

Objective:

• Develop a logistic regression model to help X Education effectively identify and prioritize high-potential leads, increasing conversion rates from the current ~30% to a target of ~80%.

Data Preparation & Exploratory Data Analysis:

• Dropped variables with extensive missing values (Asymmetrique Activity Score, Asymmetrique Profile Score).

• Imputed missing numeric values (TotalVisits, Page Views Per Visit) with median.

• Handled outliers by capping at the 99th percentile and validated visually.

• Converted binary variables (Yes/No) to numeric (0/1) and created dummy variables for categorical features.

• Transformed skewed numeric variables to improve normality (log-transformed Total Time Spent on Website, sqrt-transformed Total Visits and Page Views Per Visit).

• Scaled numeric variables using StandardScaler.

Model Building & Evaluation:

• Applied Recursive Feature Elimination (RFE) selecting 10 significant features.

• Important positive predictors for conversion:

• Lead Source: Welingak Website

• Tags: Closed by Horizzon, Lost to EINS, Will revert after reading email

• Last Activity: SMS Sent

• Key negative predictors included:

• Lead Quality: Worst

• Tags: Switched off, Interested in other courses, Ringing

• Asymmetrique Activity Index: Low

• Logistic regression model achieved strong performance:

• Accuracy: ~89%, Sensitivity: ~73%, Specificity: ~98%, AUC-ROC: 0.93

• Selected a probability cutoff of 0.25, effectively balancing precision and recall for aggressive lead targeting.

Business Recommendations:

• During intern-intensive phases, retain a low cutoff (0.25) to aggressively maximize lead conversions.

• To reduce unnecessary calls, increase the cutoff, focusing only on highly probable conversions.

This refined logistic regression model effectively identifies leads with the highest potential, enabling strategic allocation of resources and substantially improving conversion efficiency at X Education.