Lead Scoring — Summary Report

This summary outlines the end-to-end approach used to build a robust, business-ready lead scoring model for X Education. The objective was to assign a calibrated score (0–100) to every lead, enabling the Sales team to focus on high-propensity “hot” leads and operate flexibly under aggressive or conservative outreach goals.

We began with data understanding and quality checks. Identifier columns and potential leakage fields were excluded; categorical “Select” placeholders were treated as missing and imputed systematically. Numeric fields were median-imputed and capped for extreme outliers. Categorical variables were one-hot encoded with drop-first to limit multicollinearity, and numeric features were standardized for modeling stability.

We split the data into train and test sets using stratification to preserve conversion ratios. A logistic regression model with class-weight balancing was chosen for its interpretability and strong baseline performance. Model quality was assessed using ROC-AUC, precision, recall, F1, and confusion matrices. To reflect business objectives, we tuned the decision threshold using the precision–recall curve rather than relying on the default 0.5, allowing us to trade off precision and recall depending on staffing and quarterly targets.

Interpretability was addressed by fitting a StatsModels logistic regression on the engineered design matrix to obtain coefficients and odds ratios. The most influential features typically included engagement behaviors (e.g., Total Time Spent on Website, TotalVisits), channel/source signals (e.g., specific Lead Source and Lead Origin categories), and certain last-activity indicators (e.g., Email Opened, SMS Sent). These factors align intuitively with sales behavior: more engaged and well-sourced leads tend to convert at higher rates.

For deployment, we generate a lead score as predicted probability × 100 and export it for the Sales team, along with the predicted label at the current operating threshold. Two playbooks are recommended. In Aggressive Mode (e.g., during the two-month intern period), set a lower threshold to maximize recall, prioritize top deciles by score and recent engagement for fast multi-channel touchpoints, and use interns to expand calling capacity. In Conservative Mode (e.g., post-target attainment), raise the threshold to maximize precision, limit calls to the highest-intent leads, and move others to automated nurture sequences until their engagement increases.

This process yields an interpretable, auditable model, strong baseline performance, and a pragmatic operating framework. The deliverables include a clean, commented notebook; a polished presentation focusing on business implications; a concise subjective-answers brief; and this summary report. Together, they demonstrate the technical rigor and the business clarity required to improve conversion rates and help the team work smarter, not just harder.