Summary Report

Introduction

The goal of the case study was to develop a logistic regression-based lead scoring model for X Education to improve its lead conversion rate from ~30% to 80%. The sales team was spending time on low-quality leads, giving low efficiencies. With the aid of machine learning, we identified hot leads i.e., most possible conversion, allowing better prioritization and resource allocation.

Stepwise approach

1. Data Understanding & Pre-processing

AT first we explored and cleaned the data by focusing on missing values, duplicates, & inconsistencies:

- Handling Missing Data: Columns with >30% missing values were dropped; others were imputed using mode (categorical) or median (numerical) values.
- Identifying 'Select' as NULL: Several categorical variables (e.g., ‘Specialization’, ‘Lead Profile’, ‘City’, ‘How did you hear about X Education’) contained the value ‘Select’, which was actually missing data may be due to non-selection by the leads. We treated these as missing values.
- Fixing Categorical Issues: Standardized inconsistent values such as 'google' and 'Google'.
- Outlier Detection & Removal: The 1.5x IQR method (due to skewness of data) was used to treat outliers in numerical features like 'Total Visits' and 'Page Views Per Visit'.

2. Exploratory Data Analysis (EDA)

Conducted EDA to understand data distributions, detect anomalies, and find relationships between features and lead conversion:

- Univariate Analysis (using histograms and count plots):
  - Identified feature distributions and missing value patterns.
  - Found skewness in 'Total Visits' and 'Page Views Per Visit', confirming the need for outlier treatment.
- Bivariate Analysis (using box plots and bar plots):
  - Explored relationships between independent features & target variable ('Converted').
  - Found that leads from 'Google' and 'Olark Chat' had high conversion rates.
  - 'Total Time Spent on Website' showed a strong correlation with conversions.
- Correlation Matrix (Heatmap):
  - Helped identify highly correlated features and reduce multicollinearity.
  - Variables like 'Last Activity_SMS Sent' and 'Last Notable Activity_SMS Sent' were highly correlated due to very high correlation (~0.9-1.0).

3. Feature Selection & Model Development

Using Recursive Feature Elimination (RFE), we selected the top 20 features, further refining them using p-values and VIF analysis to remove multicollinearity and achieve model with features <=15. Starting RFE with 15 features and no fine-tuning may lead to unnecessary or redundant features by not allowing comprehensive set of relevant predictors.

- Train-Test Split: Data split into 70% training, 30% testing.
- Feature Scaling: StandardScaler applied to numerical features.

- Model: A logistic regression model (GLM - binomial family) was trained.
- Model refinement iteratively: Final Model with reduced to 14 features for better stability.

4. Model Evaluation: ROC-AUC & Optimal Cutoff Selection

To assess model performance, we plotted the Receiver Operating Characteristic (ROC) curve.

- AUC-ROC Score: 0.87, confirming strong classification ability.
- Sensitivity-Specificity Tradeoff: A cutoff of 0.35 was optimal, balancing:
    - Sensitivity (Recall) = 81% (high conversion prediction)
    - Specificity = 81% (avoiding unnecessary sales efforts)

5. Precision-Recall Tradeoff & Business Decision on Cutoff

Since X Education's goal was to maximize lead conversions, we analyzed the Precision-Recall Tradeoff curve:

- Precision-Recall suggested an optimal cutoff at 0.42, which increased precision but lowered sensitivity to 76%.
- Since business objective was ~80% conversion rate, we chose 0.35 as final cutoff, ensuring:
    - More potential leads were correctly identified (higher recall).
    - The model achieved with the CEO's goal of 80% lead conversion.

Key Learnings

1. Handling missing values, standardizing data, and treating outliers significantly improved model performance.
2. Understanding strange Missing Data: Recognizing 'Select' as missing values prevented data distortions and improved feature quality.
3. EDA Helps in Feature Engineering & Selection: Univariate and bivariate analysis helped identify important predictors and detect outliers.
4. Feature Selection Enhances Model Stability: RFE + p-value + VIF analysis ensured a robust, interpretable model.
5. ROC-AUC for Performance Evaluation: A 0.87 AUC score confirmed strong classification ability.
6. Sensitivity-Specificity vs. Precision-Recall Tradeoff:
    - Precision-Recall (0.42) improved precision but lowered sensitivity.
    - Sensitivity-Specificity (0.35) balanced ensures an 80% conversion rate.
7. Choosing the Right Cutoff is Business-Driven: A 0.35 threshold aligned best with the company's conversion goal, ensuring higher lead prioritization without excessive false positives.

Conclusion & Business Impact

- The final model achieved desired 80% conversion target.
- The lead scoring system can help sales team prioritize high-quality leads, improving efficiency.
- The 3 most important features are:
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional
    - Lead Source_Olark Chat