Lead Conversion Project for X Education

This project aimed to develop a predictive model for **X Education** to optimize their lead conversion process. The organization receives thousands of leads through various digital and offline channels, but only a small portion (~30%) converts into actual enrollments. The objective was to create a **lead scoring model** using logistic regression to identify and prioritize high-potential leads. By assigning a score from **0 to 100**, the sales team could focus their efforts on leads with the highest likelihood of conversion, thereby improving operational efficiency and return on marketing investment.

Approach and Methodology

The analysis started with a comprehensive review of the dataset, which included over 160 variables, both categorical and numerical. Initial preprocessing involved **handling missing values**, treating placeholders such as 'Select' as nulls, and removing low-variance or redundant features. Categorical variables were transformed using **one-hot encoding**, creating a clean dataset ready for modeling.

A **stratified train-test split** (70/30) was performed to maintain class balance. Since the dataset was imbalanced, with far fewer converted leads, we used **class_weight='balanced'** in logistic regression to address this issue. All numerical features were **standardized** using StandardScaler to ensure the model treated them fairly during training.

A logistic regression model was trained with a maximum iteration cap to ensure convergence. Post-training, model performance was evaluated using key metrics: **accuracy (92.8%)**, **precision (91.6%)**, **recall (92.0%)**, **F1 score (91.8%)**, and **ROC AUC (97.2%)**. These high scores indicated that the model was robust and well-balanced, making it suitable for real-world deployment.

Results and Interpretations

The model's output probabilities were converted into **lead scores** by scaling them to a 0–100 range, making them intuitive for the business team to use. These scores enable the sales team to prioritize leads dynamically. For instance, a lead with a score above 80 could be classified as high priority, while those below 20 might be excluded from immediate outreach.

We also analyzed feature coefficients to understand **which factors most influenced lead conversion**. Features like *'Tags_Will revert after reading the email'*, *'Tags_Closed by Horizzon'*, and *'Tags_Ringing'* (negative impact) were the most influential. This allowed us to derive not just predictive power but **actionable business insights** from the model.

Key Learnings and Business Recommendations

This model can be adapted to **different business situations** by modifying the classification threshold. For example:

- During periods when interns are available, **lower the threshold** to increase recall and engage more leads.
- When the team reaches its target, **raise the threshold** to focus only on the most likely converters and reduce unnecessary outreach.

This project taught the importance of **feature engineering, class balancing, and threshold tuning** in real-world classification problems. The logistic regression model proved to be not only accurate but also interpretable, making it highly applicable for business decision-making.

--Manisha Singh