Improving Lead Conversion at X Education with Logistic Regression

This report details the development of a logistic regression model to identify high-potential leads for X Education, an online course provider. The model aims to assign a lead score between 0 and 100, indicating the likelihood of conversion into a paying customer.

Background

X Education struggles with a low lead conversion rate (around 30%), despite a high volume of leads. To improve efficiency, the company wants to identify "hot leads" for the sales team to focus on, potentially boosting conversion rates to 80%.

Data and Approach

A historical dataset of 9,000 leads with various attributes (lead source, website activity, etc.) was used. The target variable was "Converted" (1=converted, 0=not converted).

Here's a breakdown of the model building process:

1. **Data Cleaning and Preprocessing:**
   - Missing values were addressed.
   - Unnecessary and duplicate data were removed.
   - Categorical features were converted for model compatibility.
2. **Exploratory Data Analysis (EDA):**
   - Initial data exploration helped understand the distribution of features.
3. **Feature Engineering and Selection:**
   - Dummy variables were created for categorical features.
   - Feature selection techniques like Recursive Feature Elimination (RFE) identified the most relevant features for the model.
4. **Model Building and Evaluation:**
   - A logistic regression model was trained using Statsmodels.
   - Multicollinearity was checked, and redundant features were removed to improve model performance.

- Model performance was evaluated on a separate test set using metrics like accuracy, precision, recall, F1-score, and ROC AUC.
- Different probability cutoffs were tested to determine the optimal threshold for classifying leads as "hot" or "cold."

Key Learnings

- Data cleaning and feature engineering are crucial for building an effective model.
- Feature selection techniques can improve model performance by focusing on the most relevant factors.
- Addressing multicollinearity helps to avoid model instability.
- The chosen probability cutoff significantly impacts the model's ability to identify hot leads.
- Evaluating the model on unseen data helps assess its generalizability.

Model Outcome

The logistic regression model assigns a lead score between 0 and 100, enabling X Education to prioritize leads based on their conversion probability. This data-driven approach can significantly improve sales team efficiency by focusing efforts on the most promising leads, potentially leading to an 80% conversion rate.