Summary Report on the Lead Scoring Case Study Assignment 2024

NOVEMBER 26

Post Graduate Program in Data Science DS 68
Sole contributor: Nguyen Van Thuong

Executive Summary:

This report details the development of a logistic regression model to predict lead conversion for X Education, an online education company. By assigning lead scores, the model aims to help the sales team prioritize "hot leads" and improve conversion rates from the current 30% to a target of 80%.

Data Analysis Process:

The project involved the following key stages:

1. **Data Acquisition and Preparation:**
    - The initial dataset contained 9,240 records and 37 features.
    - Categorical features were identified for conversion into dummy variables.
    - Irrelevant and incomplete data were removed, resulting in a final dataset with 68.97% of the original rows and an increased conversion rate from 38.54% to 48.09%.
2. **Exploratory Data Analysis (EDA):** Visualizations using Matplotlib and Seaborn revealed key insights, such as the influence of lead source and website engagement on conversion likelihood.
3. **Feature Engineering:** Categorical features were converted into dummy variables to optimize model performance, resulting in 74 features.
4. **Preparing Data for Modelling:**
    - The dataset was split into training (70%) and testing (30%) subsets.
    - Robust scaling was applied to mitigate the impact of potential outliers in numerical variables.
5. **Model Buiding:**
    - Recursive Feature Elimination (RFE) and stepwise elimination based on Variance Inflation Factor (VIF) and p-values were used to select the most relevant features.
    - The final model included 12 features at the 4^{th} step.

6. **Initial Model Evaluation:**
    - A conventional 0.5 cutoff was used for classification.
    - The model achieved an accuracy of 78.05% on the training set but showed a high number of false negatives.
    - To optimize the classification threshold, ROC curve analysis and Precision-Recall Tradeoff analysis were conducted.
7. **ROC Curve Optimization:**
    - ROC curve analysis yielded an optimized cutoff of 0.418, achieving higher recall on both training and testing sets.
    - This cutoff prioritizes minimizing false negatives, aligning with X Education's goal of capturing potential leads.
8. **Precision-Recall Tradeoff Optimization:**
    - This analysis resulted in an optimized cutoff of 0.431, providing a balanced tradeoff between precision and recall.
    - This cutoff is suitable for scenarios like marketing campaigns where a balance between identifying potential customers and minimizing wasted resources is crucial.

Key Learnings: This project provided valuable insights into the data science workflow:

- EDA: The importance of thorough EDA in identifying trends and informing feature selection.
- Data Integrity: The necessity of handling missing values effectively to maintain data integrity.
- Feature Engineering: The role of feature engineering in enhancing model performance.
- Evaluation Metrics: Understanding how different evaluation metrics can guide business decisions regarding lead prioritization.
- Optimized Cutoffs: How to determine and select the appropriate optimized cutoffs based on different business objectives.

Conclusion: This case study reinforced technical skills in data analysis and modeling and highlighted the importance of aligning analytical outcomes with business objectives. The logistic

regression model, with optimized cutoffs determined through ROC curve and Precision-Recall Tradeoff analysis, provides X Education with a robust tool to predict lead conversion. The choice of cutoff will depend on the specific business objective, whether it is prioritizing lead capture or balancing precision and recall. Continuous monitoring of model performance is crucial to ensure alignment with evolving business needs.