Summary Report on Lead Scoring Case Study Assignment

Introduction

The primary objective of this assignment was to develop a lead scoring model for X Education, an online course provider, to enhance their lead conversion rate, which currently stands at 30%. The goal was to identify 'Hot Leads' that are more likely to convert into paying customers, thereby increasing the conversion rate to approximately 80%. This report summarizes the approach taken, methodologies employed, and key learnings derived from the assignment.

Data Collection and Preparation

The analysis commenced with the acquisition of a dataset containing around 9,240 leads, featuring various attributes such as Lead Source, Total Time Spent on Website, and Last Activity. The first step involved data cleaning, where missing values and irrelevant features was addressed. Columns with excessive missing data (over 3,000 null values) were removed to streamline the dataset.

Categorical variables were transformed into dummy variables for effective modelling, including features like Lead Origin and Last Activity. After handling null values, the dataset was reduced to 6,373 rows, ensuring a robust foundation for analysis.

Exploratory Data Analysis (EDA)

Following data preparation, exploratory data analysis (EDA) was conducted to understand the relationships between different features and the target variable, 'Converted'. Visualizations such as histograms and correlation matrices were employed to identify patterns and correlations, which were crucial in determining the most impactful features on lead conversion.

Model Development

The next phase involved building a logistic regression model to assign lead scores. The dataset was split into training and testing sets, using 70% of the data for training and 30% for testing. Logistic regression was chosen for its interpretability and effectiveness in binary classification tasks.

To enhance model performance, Recursive Feature Elimination (RFE) was utilized to select the most relevant features. This process identified key predictors such as Total Visits, Total Time

Spent on Website, and specific Lead Sources. The model was evaluated using metrics like accuracy, precision, recall, and the ROC AUC score, which indicated strong predictive capability with an AUC of 0.86.

Results and Business Implications

The final model demonstrated an accuracy of approximately 79% on the training set. The insights gained from the analysis revealed that focusing on leads with higher scores could significantly improve the efficiency of the sales team. By prioritizing 'Hot Leads', X Education could allocate resources more effectively, ultimately leading to a higher conversion rate.

Learnings

This assignment provided several key learnings:

1. **Data Cleaning is Crucial:** Thorough data cleaning and preparation are critical for ensuring the quality of analysis and model performance.
2. **Feature Selection Matters:** Identifying and selecting the right features is essential for building an effective predictive model. RFE proved to be a valuable tool in this process.
3. **Model Evaluation is Key:** Understanding various evaluation metrics is crucial for assessing model performance and making informed decisions.
4. **Business Relevance:** Translating analytical results into actionable business strategies is essential for driving improvements in sales and resource allocation.

Conclusion

In conclusion, the lead scoring case study was a comprehensive exercise in data analysis, model building, and deriving business insights. The skills and knowledge gained from this assignment will be invaluable in future data-driven decision-making processes.