Summary Report

Introduction

X Education, a company offering online courses to industry professionals, has been facing challenges in converting a significant portion of its leads into paying customers. Despite generating a large number of leads daily through its website and various marketing channels, the lead conversion rate remains low, hovering around 30%. The company seeks to enhance this conversion rate by identifying the most promising leads, or "Hot Leads," allowing the sales team to focus their efforts more effectively.

To address this challenge, a lead scoring model was developed. The model aims to assign a score between 10 to 100 to each lead, indicating the likelihood of conversion. The ultimate goal set by the CEO is to achieve an 80% lead conversion rate.

Data Understanding and Preparation

The initial dataset consisted of 9,240 records with 37 columns, including identifiers like "Prospect ID" and "Lead Number." These columns were dropped as they did not add value to the predictive model.

Handling Missing Data

- Columns with more than 3,400 missing values were removed to streamline the dataset.

- Missing values in critical fields were either imputed or those records were dropped to ensure data quality.

Outlier Detection

- Boxplots were utilized to identify outliers in numerical variables such as "TotalVisits," "Total Time Spent on Website," and "Page Views Per Visit." Extreme outliers were removed to stabilize the data distribution.

Feature Engineering

- Dummy variables were created for categorical columns such as "Lead Origin," "Lead Source," and "Specialization," among others. This expanded the dataset to include binary indicators for each category, making it suitable for logistic regression modeling.

Exploratory Data Analysis (EDA)

The EDA phase involved both univariate and bivariate analysis to understand the distribution of data and the relationships between variables. Correlation analysis was performed to identify potential multicollinearity issues, which could affect the model's reliability.

Model Building

A logistic regression model was developed, using Recursive Feature Elimination (RFE) to select the most important predictors. The selected features were those that contributed significantly to predicting lead conversion.

Key Predictors

- Total Time Spent on Website: Leads that spend more time on the website are more likely to convert.

- TotalVisits: The number of visits also plays a crucial role in determining the likelihood of conversion.

- Lead Origin_Landing Page Submission: Leads originating from landing page submissions show a higher conversion probability.

Model Evaluation

The model was evaluated using several metrics on both the training and test datasets:

- Accuracy: The model achieved an accuracy of approximately 79% on both the train and test sets.

- Sensitivity: Around 77%, indicating the model's ability to correctly identify actual converters.

- Specificity: Approximately 81%, showing the model's ability to correctly identify non-converters.

- Precision: The precision of the model was about 80%, reflecting the accuracy of the positive predictions.

An ROC curve was plotted, and the area under the curve (AUC) was found to be 0.87, indicating a strong model performance.

Threshold Optimization

The model's default threshold of 0.5 was optimized to a trade-off value of 0.47, where precision and recall were balanced. Leads with a conversion probability higher than 47% were identified as "Hot Leads."

Strategies for Implementation

1. During Peak Periods:

  - Interns and additional manpower can focus on leads with the highest conversion probabilities, maximizing outreach efforts during peak times.

2. After Reaching Quarterly Targets:

  - Once the conversion target is achieved, efforts should be concentrated on maintaining high-quality customer interactions. This can include nurturing long-term leads, refining sales strategies, or focusing on professional development for the sales team.

Conclusion

The lead scoring model developed for X Education effectively predicts the likelihood of lead conversion, with a well-balanced performance between accuracy, sensitivity, and specificity. By focusing on high-probability leads, the company can optimize its sales efforts, improve conversion rates, and achieve the CEO's target of an 80% conversion rate. The model is also designed to adapt to future constraints, making it a valuable tool for sustained business growth.

This approach not only enhances immediate sales performance but also prepares the company for long-term success by focusing on quality over quantity in lead management.