Lead Score Case Study for X Education: Summary

Background:

X Education, a prominent player in online education for professionals, faced challenges with its lead conversion rates. Despite a substantial number of daily leads, the conversion rate lingered around 30%. The objective was clear: improve efficiency by identifying 'Hot Leads' - those with a high likelihood of conversion.

Approach:

The project commenced with an in-depth analysis of the provided dataset, consisting of approximately 9000 data points with diverse attributes. The data, reflecting various aspects of customer interaction and behavior, was a goldmine for understanding lead conversion dynamics.

Data Preparation:

Initial steps involved a thorough examination of the dataset. This revealed several nuances:

- The dataset comprised 37 attributes, including categorical and numerical data.

- A significant amount of missing data and 'Select' entries (effectively null values) in categorical variables posed initial challenges.

- No duplicate entries were found, ensuring data integrity.

The cleaning process involved handling null values and removing columns with high percentages of missing data. 'Select' entries in categorical variables were treated as null values. Columns with more than 35% missing values were dropped, streamlining the dataset to 25 relevant attributes.

Exploratory Data Analysis (EDA) and Feature Engineering:

The EDA phase focused on understanding patterns and correlations in the data:

- Categorical variables with imbalanced data or redundant information were dropped.

- Numerical variables like 'Total Visits' and 'Page Views Per Visit' underwent outlier treatment to ensure statistical robustness.

- The most striking insight was from 'Total Time Spent on Website,' which showed a strong correlation with lead conversion, indicating that engaged leads are more likely to convert.

Model Building and Optimization:

A logistic regression model was chosen for its aptitude in binary classification problems. The process involved:

- Splitting the data into training and test sets.

- Employing Recursive Feature Elimination (RFE) to identify the most impactful features.

- Iteratively refining the model by evaluating p-values and VIF (Variance Inflation Factor) to address multicollinearity.

Model Performance and Metrics:

The final model showcased an impressive balance of accuracy, sensitivity, and specificity:

- An overall accuracy of around 77% was achieved, both on training and test data.

- Sensitivity and specificity metrics indicated the model's robustness in identifying true positives and true negatives, respectively.

Key Learnings and Recommendations:

- Time spent on the website is a critical factor in lead conversion. Strategies to engage leads on the site can boost conversion rates.

- Certain lead sources and origins (e.g., 'Google', 'Direct Traffic') are more likely to convert, suggesting a potential area for marketing focus.

- The occupation of leads plays a significant role, with working professionals showing higher conversion rates.

Conclusion:

This analysis provides X Education with actionable insights to refine its lead targeting strategy, focusing on the most promising leads. The model's high sensitivity ensures fewer potential customers are overlooked, while its specificity minimizes wastage of resources on less likely conversions. This data-driven approach is poised to significantly enhance X Education's lead conversion efficiency.