Lead Scoring Case Study

Summary Report

X Education, an online educational institution catering to industry professionals, faces a challenge in terms of lead conversion rates despite a substantial influx of potential leads daily. The company's objective is to elevate its lead conversion rate from the current 30% to an ambitious 80%. This can be achieved by identifying and prioritizing 'Hot Leads,' individuals who are more likely to convert. This report outlines the development of a logistic regression model for lead scoring and provides a set of strategic recommendations for various scenarios.

Data Overview:

The dataset provided comprises approximately 9000 data points and includes a variety of attributes such as Lead Source, Total Time Spent on Website, Total Visits, and Last Activity, among others. The target variable 'Converted' indicates whether a lead was successfully converted (1) or not (0). Additionally, certain categorical variables contain a 'Select' level, which is treated as equivalent to a null value.

Steps Used In Model Development:

1. Data understanding
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis
4. Data Preparation
5. Model Building
6. Model Evaluation
7. Finding the Optimal Cutoff
8. Precision and recall tradeoff
9. Making Predictions on Test Data
10. Conclusion

Disclaimer:

- The model exhibits the following performance metrics on the test set, demonstrating consistency between the train and test sets:

  Accuracy: **80.34%**

  Sensitivity: **79.78% (approximately 80%)**

  Specificity: **80.68%**

- The area under the ROC curve is **0.88**, indicating the model's predictive strength.
- Variables with over **3,000** missing values and numerical data outliers were not considered in the model.
- A Logistic Regression model was constructed using SKLearn, with RFE selecting **15** variables as output.

Key Findings:

Top Three Variables Contributing to Conversion Probability (Hot Leads):

I. Total Time Spent on Website.

II. What is your current occupation - Working Professional.

III. Lead Origin - Lead Add Form.

Top Three Categorical/Dummy Variables for Focused Efforts:

I. What is your current occupation - Working Professional.
II. Lead Origin - Lead Add Form.
III. Lead Source - Welingak Website.

Variables with Negative Correlation

A key highlight after performing logistic regression on data was that some features were identified that had a negative correlation with the target variable which in our case is lead conversion for the education company. So, approaching those prospects with the following features is a no.

I. Lead Source_referral sites
II. Lead Source_direct traffic
III. Last Activity_olark chat conversation
IV. Last Activity_converted to lead
V. Do Not Email_yes
VI. const

Strategy Recommendations:

1. During Intern Hiring Period:

It is advisable to prioritize leads based on characteristics that positively contribute to the conversion rate. In descending order, focus on:

I. Total Time Spent on the website.
II. Current occupation - Working Professional.
III. Lead Origin - Lead Add Form.
IV. Lead Source - Welingak Website.
V. Current occupation - Other.
VI. Last Activity - Had a Phone Conversation.

2. After Achieving Quarterly Targets:

The primary goal is to minimize unnecessary phone calls by shifting focus to the following criteria during this time, so once the above listed variables have been exhausted the variables listed below could be approached for cold calls

I. Current occupation - Student.
II. Last Activity - SMS Sent.
III. Current occupation - Unemployed.