Lead Scoring Model for X Education

Introduction:

X Education, an online course company, has a low 30% lead conversion rate. The firm acquires leads from the website, search engines, and word of mouth but few turn into paying clients. The key issue is to categorize the high-potential leads that have a high likelihood of conversion so that they can be targeted by the sales team for follow-ups. To solve this problem, a logistic regression-based lead scoring model was developed. This model's aim is to deliver a probability score for each lead so that the sales team focuses on leads that have a high likelihood of being converted. A target of the CEO is to increase the rate of lead conversion to 80%, which this model should accomplish by optimizing the prioritization of leads.

Data Preparation & Analysis:

There exist 9,240 records and 37 features-in both categorical and numerical-before preprocessing. Extensive data cleaning and preprocessing steps were carried out to render the dataset analysis-ready before model development.

Some Steps on How Data Cleaning was Done:

Handling Missing Values:

1. Features with more than 45% of their values missing were removed for their condition in data quality.
2. Select values in the categorical variables were treated as missing and removed.
3. Other features in categorical variables had their missing values filled in by mode in order to maintain data consistency.

Feature Engineering:

1. Categorical columns needed to be dummy-coded to be feasible for logistic regression.
2. Continuous features such as total visits, page views per visit, and total time spent on the site were standardized for optimal model performance.
3. Categorical variables were grouped into broader categories in the case of Lead Source, Last Activity, Tags, and Country, in order to reduce noise.

Exploratory Data Analysis (EDA):

Univariate Analysis:

1. The distribution of numerical features was studied through histograms and box plots.
2. The conversion rate was analyzed against individual lead sources and activities.

Bivariate Analysis:

1. Total time spent on the site and lead conversion had a strong positive correlation.
2. Leads that were engaged through SMS or email scored higher on conversion.

Multivariate Analysis:

1. A correlation heatmap was developed to identify relationships among the numerical variables.
2. TotalVisits and PageViewsPerVisit were more weakly, but positively, correlated with conversion while Total time spent seems to have a much stronger effect.

Model Development & Feature Selection:

Among several approaches considered for predicting mergers and acquisitions, the best approach is logistic regression because:

It produces a probabilistic score and is therefore suitable for ranking leads.

It helps create interpretability so that the company understands where to direct efforts for better conversions.

It will be computationally efficient, enabling easy deployment and scaling.

Feature Selection Using Recursive Feature Elimination (RFE):

Selected 15 most important features to improve the model performance.

Model was trained with LogisticRegression(class_weight='balanced') to address the class imbalance.

Model Performance & Threshold Optimization:

Metrics for evaluating model performance:

Performance on Train Set:

Train Set Accuracy: 89.87%

Train Set Recall: 85.4%

Performance on Test Set:

Test Set Accuracy: 87.12%

Test Set Recall: 83.2%

Specificity: 92.67%

ROC Curve Analysis: The advantages of precision were well balance against recall.

Threshold optimization:

The initial threshold for classifying leads was 0.5, but to achieve the target of 80% recall, various thresholds were evaluated. The threshold that was chosen corresponds to the closest recall value of 80% without missing out on potential conversions.

Business Recommendations:

Top Variables for Lead Conversion:

Total Time Spent on Website – Strongest predictor; higher engagement correlates with conversion.
Last Notable Activity - SMS Sent – Leads receiving SMS updates are more likely to convert.
Lead Quality - Not Sure (Negative Impact) – Uncertain leads have a lower probability of conversion.

Top Categorical Variables to Focus On:

Lead Source - Welingak Website – Leads from this source have a high conversion rate.
Lead Origin - Lead Add Form – Manually added leads show better conversion potential.
Tags - Will Revert After Reading Email – Indicates potential customers who may need follow-ups.

Business situations will call for different approaches:

When it comes time to hire an intern:

1. The threshold probability can be relaxed to above, say, 0.4.
2. Target the very engaged users (e.g., those who spent a long time on the website or engaged through SMS/email).
3. Increase follow-up and reminders.

When meeting quarterly targets:

1. Should raise threshold probability above, say, 0.7, to ensure only high-value leads are contacted.
2. Automated emails/SMS to lower-priority leads.

Conclusion & Business Impact:

This lead scoring model gives excellent prediction on the conversion probability and allows high-potential leads to be prioritized by X Education.

Key benefits,

1. The sales efficiency, including justifying stretch expectations on a few potential leads,
2. The operational cost included with unnecessary follow-up.
3. An expected increase in conversion rates consistent with the expectations of the CEO (80%).