Summary: Leveraging Logistic Regression for Lead Conversion Enhancement in X Education

The education company, X Education, is grappling with the challenge of lead conversion, seeking to identify the most promising leads for its online courses. To address this, a logistic regression model is developed to assign lead scores, enabling the company to focus resources on high-conversion potential leads.

Business Problem:

X Education aims to improve its lead conversion rate, targeting a conversion rate of 80%. The task is to construct a logistic regression model that assigns lead scores, categorizing leads as "hot" or "cold" based on their likelihood of conversion.

Steps to Solve the Business Problem:

➢ **Problem Formulation:** Clarify the business problem and translate it into a data science challenge.

➢ **Data Import and Understanding:** Gather data and gain insights into its characteristics.

➢ **Preprocessing:** Clean the data by handling missing values, outliers, and redundant variables.

➢ **Train-Test Split:** Divide the dataset into training and testing subsets.

➢ **Feature Scaling:** Normalize numeric variables using StandardScaler to mitigate outlier influence.

➢ **Model Building:** Construct a logistic regression model using significant predictors identified through Recursive Feature Elimination (RFE).

➢ **Model Evaluation:** Analyse p-values and Variance Inflation Factor (VIF), refining the model iteratively.

➢ **Insights from Model:** Uncover predictors positively or negatively affecting lead conversion.

Insights from EDA:

➢ Landing page submission is the primary lead origin, while maximum conversions arise from referrals.

➢ Leads from the banking, investment, and insurance domain exhibit the highest conversion rates.

➢ Working professionals constitute a significant portion of converted leads.

Assumptions:

➢ Columns more than 40% missing values were dropped

➢ Columns having very low or no variance were dropped

➢ 'Select' value throughout the data is treated as 'null' value

➢ The missing values in the categorical columns were replaced by the column's respective mode (i.e., most occurring value in the column)

➢ The missing values in the continuous columns were replaced by the column's respective median (i.e., The value below which 50% of a particular column's datapoints reside)

➢ Highly correlated dummy variables were dropped

Feature Scaling:

StandardScaler is employed due to outlier presence in numeric variables.

Model Insights:

Predictors with positive weightage: Last Notable Activity, Current Occupation, Total Time Spent on Website, Last Activity, Lead Source.

Predictors with negative weightage: Lead Origin, Ringing, Interest in Other Courses, already a student.

Model Evaluation and Performance:

The optimal cut-off for classification is identified as 0.36, yielding a model with the following performance parameters on test data:

➢ Accuracy: **87.01%**

➢ Sensitivity: **87.09%**

➢ Specificity: **86.96%**

➢ Positive Predictive Value: **80.93%**

Conclusions and Suggestions:

➢ **Phone Call Priority:** Focus on leads predicted as "hot" by the model for phone calls.

➢ **Segmentation:** Segment leads based on Conversion Probability score, prioritizing those with higher scores.

➢ **Personalized Communication:** Train interns for personalized communication, emphasizing X Education's benefits.

➢ **Follow-Up Strategy:** Implement a structured follow-up plan, increasing contact frequency for positive responses.

➢ **Feedback Loop:** Regular intern feedback to refine calling strategy and make necessary adjustments.

➢ **Offer Discounts:** Provide special discounts to low-scoring leads to boost conversion potential.

➢ **Selective Calling:** Prioritize leads with positive attributes and avoid those with negative indicators.

➢ **Marketing Focus:** Emphasize content marketing, targeted emails, and lead nurturing for engagement.

This approach will empower X Education to enhance lead conversion rates, optimize resource allocation, and ultimately drive revenue growth.