Project Summary Report: Lead Scoring Model for X Education

Objective:

X Education, a prominent online education firm, experiences a lead conversion rate of about 30%. The project's goal is to identify 'Hot Leads,' or potential customers most likely to convert, thereby optimizing the lead engagement process and elevating the conversion rate towards the CEO's target of 80%.

Data Overview:

The dataset, consisting of 9,240 entries, includes diverse information on leads. Initial inspection revealed a mix of categorical and numerical data and a moderate imbalance in the target variable 'Converted.'

Data Preparation:

A cleaning process was implemented to handle missing values, with a significant culling of data where necessary. Categories with 'Select' levels indicating unselected options were addressed, and dummy variables were created for categorical data. The dataset was then split for training and testing, and numeric features were standardized.

Exploratory Data Analysis (EDA):

The EDA phase focused on the distribution of the 'Converted' target variable and a deep dive into categorical variables. Visualizations such as countplots provided insights into the lead origins, sources, and last activities that correlate with high conversion rates.

Model Building and Evaluation:

Logistic regression was chosen for its appropriateness for binary classification. Recursive Feature Elimination (RFE) selected the top 15 features. The model displayed an impressive ROC AUC of 0.896, indicating a strong ability to differentiate between converted and non-converted leads. Precision and recall scores were well-balanced, and the model achieved an accuracy of 82.27% on the training data.

Feature Importance:

The model identified 'Lead Origin_Lead Add Form,' 'What is your current occupation_Working Professional,' and 'Lead Source_Welingak Website' as the most influential features for lead conversion. These insights suggest a strategic focus on enhancing lead acquisition methods and targeting working professionals for improved conversion rates.

Model Optimization:

Precision-recall tradeoffs were examined to fine-tune the classification threshold. The optimal balance was found at a threshold of 0.4, which was then applied to the test set, yielding consistent performance metrics.

Operationalization:

A function was crafted to process new lead data and identify 'hot leads' with a conversion probability above 80%. This tool allows for dynamic lead scoring, essential for real-time marketing and sales strategies.

Strategic Recommendations:

- **Aggressive Conversion Phase**: During the internship period, prioritize high-scoring leads for phone calls. Provide interns with scripts that emphasize the identified influential features to maximize the efficiency of their outreach.
- **Post-Target Achievement**: After quarterly targets are met, raise the threshold to focus only on the most promising leads. Utilize automated marketing for ongoing engagement, allowing the sales team to focus on new strategic areas.

Conclusion:

The lead scoring model developed for X Education significantly enhances the lead conversion process by identifying high-potential leads. It offers a nuanced understanding of the factors contributing to lead conversion, enabling the company to fine-tune its marketing and sales efforts. The model not only meets the immediate objective of improving conversion rates but also serves as a tool for strategic decision-making, ensuring that the company's customer acquisition efforts are both effective and efficient.