Summary Report on Lead Conversion Analysis for X Education

The objective of this assignment was to analyse the lead conversion process for X Education, an online course provider facing a conversion rate of approximately 30%. The goal was to identify high-potential leads, referred to as "Hot Leads," and increase the conversion rate to around 80% through data-driven strategies.

Data Preparation

The initial step involved examining a dataset comprising around 9,000 leads, which included variables such as Lead Source, Total Time Spent on Website, Total Visits, and Last Activity. The target variable was 'Converted,' indicating whether a lead converted into a paying customer.

Data cleaning was the first crucial step. I created dummy variables for categorical features and identified columns with excessive missing values. Specifically, I removed columns with more than 3,000 missing values and treated instances of 'Select' as missing data. This approach was necessary to maintain a robust dataset while ensuring that only relevant features were included for analysis. Additionally, I dropped columns with predominantly one value and handled missing entries in allowing us to retain 69% of the data.

Feature Engineering and Model Building

After cleaning the data, I focused on feature engineering. I categorized the variables and created dummy variables to facilitate model training. Subsequently, I split the dataset into training and testing sets to ensure the model's performance could be effectively evaluated.

To build the model, I employed Recursive Feature Elimination (RFE) to select the most impactful features. The logistic regression model was created using statsmodels, with careful consideration of p-values and Variance Inflation Factors (VIF) to identify multicollinearity issues. All variables included in the model had p-values below 0.05, indicating their statistical significance.

Model Evaluation and Insights

Once the model was developed, I created a DataFrame containing actual conversion flags and predicted probabilities. I initially set a cutoff of 0.5 to classify leads as converted or not, but later optimized this threshold by analysing the ROC curve. The area under the curve (AUC) was found to be 0.76, which indicated good model performance. Ultimately, I determined an optimal cutoff point of 0.42 based on the sensitivity and specificity trade-offs.

The model evaluation revealed an accuracy of approximately 69.6%, with precision at 71% and recall at 61%. These metrics highlighted the model's strengths in minimizing false positives while indicating room for improvement in capturing true positive conversions.

Learnings and Recommendations

Throughout this assignment, I learned the importance of data cleaning and feature selection in building a predictive model. Addressing missing values and eliminating irrelevant features were critical to improving model performance. Additionally, the significance of interpreting precision, recall, and accuracy in understanding model effectiveness became clear.

In terms of business implications, the analysis yielded actionable strategies. For aggressive outreach periods, focusing on high-impact variables and leveraging specific categorical segments (like "welingnak" leads and those who are unemployed or students) can optimize the sales team's efforts.

Conversely, during quieter periods, prioritizing outreach based on lead scoring can reduce unnecessary contacts and enhance efficiency.

In conclusion, this assignment reinforced the value of data-driven decision-making in improving sales strategies, ultimately aiding X Education in increasing its lead conversion rate.