Lead Scoring Case Study Summary

Problem Statement:

X Education is an online course provider targeting industry professionals. The company seeks assistance in identifying the most promising leads, those with the highest likelihood of converting into paying customers. To achieve this, they require a lead scoring model that assigns a lead score to each lead, with higher scores indicating a higher chance of conversion and lower scores indicating a lower chance of conversion. The CEO has set a target lead conversion rate of approximately 80%.

Solution Summary:

Step 1: Reading and Understanding Data

- The dataset was read and inspected to gain a thorough understanding of its contents.

Step 2: Data Cleaning

- Variables with unique values were dropped.

- Columns containing the value 'Select,' indicating leads who didn't choose any option, were replaced with null values.

- Columns with more than 35% null values were dropped.

- Imbalanced and redundant variables were removed, missing values were imputed, and outliers were identified and addressed. Label inconsistencies were fixed.

- Sales team-generated variables were removed to ensure clarity in the final solution.

Step 3: Data Transformation

- Binary variables were converted to '0' and '1'.

Step 4: Dummy Variables Creation

- Dummy variables were created for categorical variables, eliminating repeated and redundant variables.

Step 5: Test Train Split

- The dataset was divided into training and testing sets, with a 70-30% proportion.

Step 6: Feature Rescaling

- Min-Max Scaling was applied to rescale the original numerical variables.

- A heatmap was generated to analyze variable correlations, and highly correlated dummy variables were dropped.

Step 7: Model Building

- Recursive Feature Elimination was employed to select the top 15 important features.

- Insignificant variables were identified and dropped based on p-values.

- The final model consisted of 11 significant variables with favorable VIF values.

- Optimal probability cutoff was determined, accuracy, sensitivity, and specificity were evaluated.

- The ROC curve was plotted, achieving an area coverage of 86% and validating the model's performance.

- The model's accuracy in predicting the converted column was assessed, and precision, recall, sensitivity, and specificity were analyzed.

- A cutoff value of approximately 0.3 was chosen based on the precision and recall trade-off.

- The learnings from the train set were applied to the test set, resulting in an accuracy of 77.52%, sensitivity of 83.01%, and specificity of 74.13%.

Step 8: Conclusion

- The lead scoring model demonstrated a conversion rate of 83% on the test set, meeting the CEO's target of an 80% conversion rate.

- The model's high sensitivity ensures the selection of the most promising leads.

- Features with significant contributions to the conversion probability include:

i. Lead Origin_Lead Add Form

ii. What is your current occupation_Working Professional

iii. Total Time Spent on Website

This concludes the summary of the lead scoring case study. The detailed implementation and findings can be found in the corresponding documentation and presentation.