Lead Scoring Case Study Summary

Problem Statement: X Education, an online course provider, seeks to enhance its lead conversion rate, currently at a modest 30%. To optimize efficiency, they aim to identify 'Hot Leads'—those with the highest likelihood of conversion—to focus their efforts effectively.

Goal: Develop a logistic regression model to assign lead scores between 0 and 100, aiding in the identification of potential leads. Higher scores denote hotter leads with a greater probability of conversion, while lower scores signify colder leads less likely to convert. Additionally, address potential future challenges outlined by the company to ensure the model's adaptability.

Solution Summary:

Data Import and Understanding:

Import essential libraries and load the dataset.

Check for duplicates and assess the dataset's structure.

Examine data types and perform initial data exploration.

Data Cleaning:

Handle 'select' values and missing data.

Drop columns with more than 70% null values and replace remaining nulls with mode values.

Data Exploration:

Analyze value counts and outliers.

Plot count plots and box plots for categorical and numerical columns.

Investigate correlations and drop highly correlated columns.

Model Building:

Prepare data for logistic regression.

Create dummy variables for categorical columns.

Split data into training and testing sets.

Perform feature scaling and selection using RFE.

Build logistic regression model and fine-tune features based on p-values and VIF.

Evaluate model performance using various metrics.

Model Evaluation:

Calculate accuracy, precision, recall, and F1-score on both training and testing data.

Determine the optimal cutoff value.

Plot ROC curve and Precision-Recall curve to assess model performance visually.

The ROC curve shows a strong performance, with an area under the curve (AUC) of 96%, indicating the model's reliability. Additionally, by plotting Accuracy, Sensitivity, and Specificity, we determined the optimal cutoff point at 0.3. Furthermore, the Precision-Recall curve highlights a balanced trade-off between precision and recall, reaffirming the effectiveness of the model.

Train Data:- ♣ Accuracy: 90.19%- ♣ Sensitivity: 87.28%- ♣ Specificity: 92.01% ♣ Precision – 92.91% ♣ Recall – 85.81%

Test Data:- ♣ Accuracy: 89.60%- ♣ Sensitivity: 85.23%- ♣ Specificity: 92.09% ♣ Precision – 86.02% ♣ Recall – 85.23% o Optimal cutoff =0.3

Conclusion:

The model achieves a conversion rate of 85% on the test dataset, exceeding the CEO's expectation of an 80% conversion rate.

High sensitivity indicates the model's ability to identify promising leads effectively.

Top three variables influencing lead scoring are 'Current Occupation', 'How did you hear about X Education', and 'Lead Source'.

Key dummy variables in the model include 'Tags_lost to EINS', 'Tags_Ringing', and 'Tags_Closed by Horizzon'.

This comprehensive approach ensures the development of a robust lead scoring model that aligns with the company's objectives and adapts to future challenges.