

Lead Scoring Case Study: X Education

1. Business Understanding: X Education aims to improve its lead conversion rate by identifying high-potential leads among the professionals visiting its website daily. The goal is to increase the conversion rate from 30% to 80% by prioritizing "Hot Leads" for the sales team.

2. Problem Statement: X Education needs to develop a model that assigns a lead score to each prospect, predicting the likelihood of conversion. The objective is to identify and focus on leads with higher conversion potential to achieve the target conversion rate.

3. Game Plan Overview:

- **Data Importing:** The data was imported and prepared for analysis.
- **Data Understanding:** Initial examination of the data to understand its structure.
- **Data Preparation:** Cleaning and transforming the dataset, including handling missing values, treating categorical variables, and creating dummy variables.
- **Exploratory Data Analysis:** Outlier detection and binning were performed to handle extreme values.
- **Test-Train Split:** The dataset was split into training and testing sets.
- **Feature Scaling:** Important features were scaled to ensure uniformity.
- **Model Building:** Recursive Feature Elimination (RFE) was used to build the model, selecting the most important predictors.
- **Model Evaluation:** The model was evaluated using metrics such as VIF, ROC curve, accuracy, sensitivity, and specificity.

4. Data Preparation: The dataset underwent several cleaning and transformation steps:

- **Missing Values:** Missing values were imputed using the most frequent values for categorical variables.
- **Categorical Variables:** Categorical variables were converted into numerical variables using dummy variables.
- **Dropping Redundant Columns:** Columns with extreme bias or low relevance, such as 'Country', were dropped.

5. Exploratory Data Analysis: Outliers were identified and treated by binning continuous variables. This approach allowed the preservation of important data while managing extreme values.

6. Model Building: Several logistic regression models were built and refined using Recursive Feature Elimination (RFE). The final model was selected based on statistical significance (p-values) and multicollinearity (VIF).

7. Model Evaluation: The model was evaluated using:

- **ROC Curve:** The area under the curve (AUC) was 88%, indicating high model accuracy.

- **Optimal Cut-off Point:** The best balance between sensitivity and specificity was determined at a probability cut-off of 0.4.
- **Precision and Recall:** The final model exhibited strong precision and recall scores, reflecting its reliability in predicting conversions.

8. Interpretation: The final model demonstrated high accuracy and stability, with an AUC of 88%. This indicates the model's effectiveness in distinguishing between high and low-potential leads. The trade-off between sensitivity and specificity was carefully balanced, leading to a robust and reliable lead scoring system.

9. Conclusion: The lead scoring model developed for X Education effectively predicts the likelihood of conversion, enabling the company to focus on high-potential leads. By prioritizing these leads, X Education can significantly improve its conversion rate, enhancing the efficiency of its sales team and achieving its business goals.