Lead Scoring Case Study Summary

Problem Statement

X Education, an online course provider, faces a low lead conversion rate of 30%. Despite many professionals visiting the website and filling out forms, only a small portion convert into customers. The sales team engages with all leads equally, leading to inefficient resource use. The goal is to identify and prioritize "Hot Leads" to improve conversion rates, reduce wasted outreach, and increase sales efficiency.

Business Objective

Develop a model that assigns a lead score between 0 and 100 to each lead, helping identify Hot Leads and enhance conversion rates. The CEO aims for an 80% lead conversion rate. The model should also address future challenges like peak time management and maximizing resource utilization.

Problem Approach

1. **Data Import and Inspection:** Initial data review and cleaning.
2. **Data Preparation:** Handling missing values and irrelevant features.
3. **Exploratory Data Analysis (EDA):** Univariate, bivariate, and correlation analysis.
4. **Dummy Variable Creation:** For categorical variables.
5. **Test-Train Split:** Dividing data into training and testing sets.
6. **Feature Scaling:** Normalizing numerical data.
7. **Model Building:** Using Recursive Feature Elimination (RFE) and optimizing based on p-values and Variance Inflation Factor (VIF).
8. **Model Evaluation:** Assessing model performance using metrics like accuracy, sensitivity, specificity, precision, and recall.
9. **Predictions:** Making predictions on the test set.

Data Pre-Processing

- **Dataset Overview:** 37 rows and 9,240 columns.
- **Dropped Features:** Single-value features and irrelevant identifiers.
- **Low-Variance Features:** Removed features with insufficient variance.
- **Missing Data Handling:** Excluded columns with over 35% missing values.

Exploratory Data Analysis (EDA)

- **Conversion Rate:** 38.5% leads converted.

- **Key Predictors:** Total Time Spent on Website, Lead Source, Last Notable Activity.
- **Class Imbalance:** Majority leads are not converted (~61.5%).

Model Building

- **Data Normalization:** Numerical data normalized.
- **Dummy Variables:** Created for categorical variables.
- **Data Shape:** (9240, 111) after preparation.
- **Data Split:** 70:30 ratio for training and testing.
- **RFE:** Selected 15 significant variables.
- **Model Optimization:** Iterative removal of variables based on p-values and VIF.
- **Final Model:** Achieved 92.9% training accuracy.

Model Evaluation

- **ROC Curve:** ROC-AUC score of 0.97, indicating excellent model performance.
- **Optimal Cut-Off:** 0.37 based on accuracy, sensitivity, specificity, and precision-recall trade-off.
- **Metrics:**
    - **Training Data:** Accuracy 92.86%, Sensitivity 90.31%, Specificity 94.49%, Precision 91.04%, Recall 90.31%.
    - **Test Data:** Accuracy 92.17%, Sensitivity 89.57%, Specificity 93.82%, Precision 90.16%, Recall 89.57%.

Key Features

- **Total Time Spent on Website**
- **Lead Source (Welingak Website)**
- **Last Activity (SMS Sent)**
- **Tags (e.g., Busy, Closed by Horizzon)**
- **Occupation (Student, Unemployed, Working Professional)**

Conclusions

- **Key Predictors:** Total Time Spent on Website, Lead Source, Last Activity, and specific Tags.

- **Occupation Impact:** Higher conversion rates for Students, Unemployed, and Working Professionals.
- **Model Performance:** High accuracy, sensitivity, specificity, precision, and recall, indicating a reliable model.

Recommendations

- **Focus on High-Engagement Leads:** Prioritize leads with higher website engagement and specific tags.
- **Leverage Lead Source & Tags:** Target leads from high-conversion sources and key tags.
- **Optimize Sales Efforts:** Direct resources towards high-probability leads and use automated channels for lower-scoring leads.
- **Monitor and Adjust:** Continuously update the model based on evolving lead behaviours for better predictions.