Lead Scoring Case Study - Enhanced Version

SUMMARY:

Problem Statement:

X Education requires assistance in identifying the most promising leads - those with the highest likelihood of becoming paying customers. The company seeks the development of a **lead scoring model** that assigns a lead score to each lead, enabling prioritization based on **conversion potential**. A higher lead score corresponds to an **increased conversion chance**, while a lower score corresponds to a **lower chance**

Solution Approach:

1. Data Exploration and Pre-processing:

The dataset consisted of **9247 records** with **37 attributes**

Addressed **missing values** by capping null values at **40%**; anything exceeding 40% was **dropped**.

Identified columns with **significant bias and low variance**, leading to their **removal.**

Refined the dataset to **9247 records** and **16 attributes**.

2. Outlier Management:

- Performed **univariate analysis** and applied **99% capping** to potential **outliers**.

3. Data Visualization:

- Conducted **bivariate analysis**, revealing valuable insights:

  a. 'Lead Add Form' origin displayed the **highest conversion ratio** compared to other lead origins.

b. 'Reference' lead source exhibited **strong performance**, followed by 'Google' and 'Direct Traffic.'

c. The 'SMS Sent' category within the 'Last Activity' column showed the **highest conversion ratio,** followed by 'Email Opened.'

d. 'Working Professional' category had a **higher conversion rate** compared to 'Unemployed' individuals.

4. Feature Engineering:

- Converted **categorical variables into dummy variables**.
- Applied **scaling** to both **training and test datasets**.
- **Reduced dimensionality** of categorical variables to enhance model efficiency.

5. Train-Test Split and Model Selection:

- Split the data into **70/30 ratio** for **training and testing**.
- Initiated **logistic regression** as the **baseline model**.

6. Model Building:

- Employed **recursive feature elimination** and **variance inflation factor** to **refine the model**.
- The **7th model iteration** revealed **p-values below 0.05** and **VIF values below 5**, indicating statistical significance and low multicollinearity among variables.

7. Prediction and Optimization:

- Analyzed **ROC curve** to determine the **optimal cut-off**.
- Opted for a **cut-off of 0.37**, yielding improved **accuracy**.
- Identified **top three influential variables**:
  1. Tags: **"Will revert after reading the email"**
  2. Lead Origin: **"Lead Add Form"**
  3. Lead Source: **"Welingak Website"**

Scenario-Driven Tuning:

- In the event of the company achieving its target ahead of schedule and seeking to focus on new opportunities, the model can be tuned for **high specificity**.

- Increased **specificity ensures accurate prediction** of non-conversions.
- Adjusting the **cut-off value higher** can achieve this goal effectively.

This comprehensive lead scoring approach equips X Education with the means to **prioritize leads efficiently** and adapt the model as per evolving business requirements.