Lead Scoring Case Study Summary

Problem Statement

X Education, an online course provider for industry professionals, aims to improve its lead conversion rate. Currently, only 30% of leads convert. The goal is to develop a lead scoring model using logistic regression to assign scores based on the likelihood of conversion, with a target conversion rate of 80%.

1. Data Understanding & Preprocessing

- **Handling Missing Values:**
  - 9,240 records and 37 features.
  - Columns with over 45% missing values were dropped, except 'Lead Quality' (51.6% missing but deemed important).
  - Categorical variables were imputed with appropriate values.
  - Numerical variables like 'Total Visits' and 'Page Views Per Visit' had missing values under 2%, so these rows were dropped.

---

3. Outlier Detection & Treatment

- **Outliers checked using boxplots:**
  - 'Total Visits' & 'Page Views Per Visit'→ capped at 95th percentile.
  - 'Total Time Spent on Website': No significant outliers found.

---

4. Exploratory Data Analysis (EDA)

- **Numerical Variables Analysis:**
  - 'Total Visits' and 'Page Views Per Visit' had similar median values for converted and non-converted leads, making them inconclusive.
  - 'Total Time Spent on Website' positively correlated with conversion – more time spent increases conversion probability.
- **Categorical Variables Analysis:**
  - **Lead Origin & Occupation:**
    - 'API' and 'Landing Page Submission' generate the most leads but have a low 30% conversion rate.
    - 'Lead Add Form' generates fewer leads but has a high conversion rate.
    - 'Working Professional' has the highest conversion rate.
    - 'Unemployed' leads are numerous but convert poorly.

    - **Lead Last Activity:**
        - Most leads are generated when the last activity is 'Email Opened.'
        - Highest conversion rate is observed when the last activity is 'SMS Sent.'
    - **Tags & Lead Quality:**
        - Most leads and highest conversion rates are for the tag 'Will revert after reading the email.'
        - 'Lead Quality' confirms that 'Might be' has the highest conversion rate, while 'Worst' has the lowest.

---

5. Data Preparation for Modeling

- **Feature Engineering:**
    - Dummy variables were created for categorical features, increasing the dataset from 37 to 88 columns.
    - Standard Scaling was applied to all numerical variables.
- **Train-Test Split:**
    - Data was split into 70% training and 30% test datasets.

---

6. Model Building & Feature Selection

- **Recursive Feature Elimination (RFE) used to select top 15 features.**
- **High p-value features were removed:**
    - 'Tags_invalid_number' & 'Tags_number_not_provided' → Dropped due to high p-values.
- **Multicollinearity Check using Variance Inflation Factors (VIF):**
    - VIF values showed no significant collinearity, so no further features were dropped.

---

7. Model Performance & Evaluation

- **Training Accuracy:**
    - 88.67% at a probability threshold of 0.05.
- **ROC Curve Analysis:**
    - AUC (Area Under the Curve) = **0.96204**, indicating a highly accurate model.
- **Final Model Predictions on Test Data:**
    - Achieved 80% conversion rate, aligning with CEO’s target.

8. Key Findings & Business Insights

- Prioritize leads with high lead scores.
- Focus on leads engaging through 'SMS Sent' as their conversion rate is high.
- Improve conversion efforts on 'API' and 'Landing Page Submission' leads.
- Increase efforts to attract 'Working Professionals,' as they convert the most.