SUMMARY REPORT: LEAD SCORING CASE STUDY

Problem Statement:

X Education seeks to assign a lead score to industry professionals for their online courses, aiming to identify leads with the highest likelihood of becoming enrolled customers. The target lead conversion rate is around 80%.

Steps involved:

1. Data Cleaning:

The data was mostly clean, except for a few null values. "Option select" was replaced with a null value as it provided little information. Some null values in categorical columns were updated with the mode to preserve data integrity.

2. Exploratory Data Analysis:

Leads primarily originate from Google, with most coming from "Landing Page Submission" and "API." Most customers are unemployed, specializing in "Others" or "Management," and mainly from Mumbai. Higher conversion rates are seen from the Lead Add Form and Reference Source, with SMS Sent converting better than Email Opened. Working professionals show higher conversion rates than unemployed individuals.

Numerical analysis shows 'Total Visits' and 'Page Views Per Visit' are highly correlated. Significant differences are noted in numerical variables across categories, with outliers present that need treatment before modelling.

3. Data Preparation:

The dummy variables were created for the categorical columns. The data was split, with 70% allocated for training and 30% for testing. Numerical features were standardized using Standard Scaler.

4. Model Building:

We used Recursive Feature Elimination (RFE) to identify key variables and removed those with a Variance Inflation Factor (VIF) over 5 or p-values above 0.05, ensuring all remaining variables had VIFs below 3. The ROC curve showed an AUC of 0.97, highlighting excellent model performance. The optimal cutoff point, determined at the intersection of accuracy, sensitivity, and specificity curves, was found to be 0.27.

5. **Model Evaluation:**

The model shows excellent performance with high accuracy on both the training data (91.98%) and test data (92.65%), meaning it generalizes well. A recall of 92.87% on the test set ensures the model identifies most potential leads, while a precision of 88.52% indicates a high percentage of correct predictions among those marked as conversions. The balance between specificity and recall further highlights the model's effectiveness in predicting lead conversions.

Conclusion:

Based on the final model, several key variables are crucial for identifying Hot Leads:

- **Tags Closed by Horizzon:** Customers with this tag show a significant likelihood of conversion.
- **Tags Lost to EINS:** This tag helps identify potential leads that may require more attention.
- **Tags Will Revert After Reading the Email:** Indicates leads that are likely to convert after engaging with email content.
- **Lead Source Welingak Website:** This source has a notable impact on the number of conversions.
- **Lead Origin Lead Add Form:** This origin is critical for identifying high-potential leads.
- **Last Activity SMS Sent:** Indicates higher engagement and conversion likelihood.
- **Lead Source Olark Chat:** Shows leads from this source have a higher chance of conversion.
- **Total Time Spent on Website:** Directly proportional to conversions, emphasizing the importance of an engaging website.

These variables help prioritize leads effectively, enhancing the overall conversion rate by focusing on high-potential leads.