Lead Score Case Study Summary:

Objective:

The aim was to build a logistic regression model to assign a lead score between 0 and 100, helping the company prioritize potential leads based on their likelihood of conversion.

Data Overview:

The dataset contained 9,240 rows and 37 columns. Initial exploration revealed significant missing values, prompting data cleaning steps. Columns with over 40% missing data were dropped, while others were imputed logically (e.g., filling "Specialization" with "Others" and "What matters most" with "Not provided"). Unemployed and Indian leads, being common, were used to fill relevant columns. After data cleaning, 98% of the original dataset was retained.

Exploratory Data Analysis (EDA):

Key insights included:

- **Lead Sources:** “API” and “Landing Page Submission” had low conversion rates but significant lead counts, requiring focused improvement. “Lead Add Form” showed high conversion but low lead count.
- **Lead Origins:** Most leads came from Google, Direct Traffic, Olark Chat, and Organic Search. Efforts should target boosting conversion from these sources while increasing leads from "Reference" and "Welingkar Website."
- **User Engagement:** Leads spending more time on the website were more likely to convert, suggesting the need for an engaging website.
- **Last Activity:** Most leads had "Email Opened" as the last activity, while leads with "SMS Sent" showed a 75% conversion rate. A follow-up call strategy was recommended.
- **Specialization:** Working professionals converted the most, suggesting targeted outreach on LinkedIn.

Data Preparation:

After cleaning, dummy variables were created for categorical features. Continuous variables like "TotalVisits," "Total Time Spent on Website," and "Page Views Per Visit" were scaled. Highly correlated variables were removed.

Model Development:

Recursive Feature Elimination (RFE) selected key features, resulting in a model with 14 significant variables after evaluating p-values and VIF scores. The final logistic regression model achieved:

- **Training Accuracy:** 85%
- **Test Accuracy:** 83%
- **ROC-AUC:** 0.93

The optimal cutoff probability of 0.42 was selected based on the Sensitivity-Specificity-Accuracy plot, ensuring a balanced prediction performance.

Key Features and Coefficients:

Top variables included encoded tags like:

- Tags_Closed by Horizzon (Coefficient: 9.33)
- Tags_Lost to EINS (Coefficient: 9.21)
- Tags_Will revert after reading the email (Coefficient: 4.16)

Model Evaluation Metrics:

- **Sensitivity:** 83.6% (correctly predicted 83.6% of actual conversions)
- **Precision:** 74% (74% of predicted hot leads were true conversions)

Conclusion:

The model effectively prioritizes leads by assigning accurate lead scores. Recommendations include enhancing website engagement, improving follow-up processes, and focusing on professional platforms to target high-potential leads.