Summary

Objective

The objective of the case study was to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads, which can be used by the company to target potential leads with a high probability of conversion (>80%).

Methodology

The following steps were performed:

Data Understanding, Cleaning, and Preparation

- All variables in the dataset were analysed based on domain knowledge.
- Duplicate entries were checked and found to be absent.
- Exploratory data analysis (EDA) was performed, treating missing values and handling outliers for numerical features.
- The distribution of data was analysed using count plots, and correlation for numerical variables was checked using a heatmap.
- The overall conversion rate was found to be 38%.

Train-Test Split and Scaling

- The dataset was split into train and test sets with a 70:30 ratio and a random_state of 100.
- Numerical features were standardized.

Model Building

- The top 15 most important variables were selected using recursive feature elimination (RFE).
- A logistic regression model was built using these variables, and the model was analyzed with respect to parameters such as variance inflation factor (VIF) and p-value.

- Variables were dropped if the p-value was greater than 0.05 or VIF was greater than 5.
- The optimal model was obtained with VIF < 5 and p-values < 0.05 for all variables.

Model Evaluation

- Various model metrics were evaluated, including accuracy, sensitivity, and specificity.
- A receiver operating characteristic (ROC) curve was plotted to determine the threshold, which was found to be 0.3.
- Other metrics such as precision, recall, and F1_score were also examined.
- The results obtained on the train set for the ideal cutoff of 0.3 were:
    - Accuracy - 92.6%
    - Sensitivity - 91.8%
    - Specificity - 93.1%
    - Precision - 89.3%
    - Recall - 91.8%
    - F1_score - 90.5%

- The model was then evaluated on the test set, and the results were similar to those obtained on the train set:
    - Accuracy - 91.4%
    - Sensitivity - 89.8%
    - Specificity - 92.2%
    - Precision - 86.8%
    - Recall - 89.8%
    - F1_score - 88.3%

Key Factors Impacting Lead Conversion Rates

- The following features were found to have the greatest impact on lead conversion rates:
    - Tags_Closed by Horizzon: Leads assigned the tag "closed by horizon" had the highest probability of conversion.
    - Tags_Lost: Leads assigned the tag "Lost" also contributed to conversion to a considerable extent.
    - Tags_Will revert after reading the email: Leads assigned the tag "will revert after reading the mail" had a significant correlation with conversion.
- Other factors that had a positive effect on conversion rates included overall time spent on the website, lead origin from landing page submission, lead source from the business website, Olark Chat, and Last Activity from SMS Sent.
- The company should prioritize leads whose lead source is from the "Welingak Website" and generate as many leads as possible through this region.
- To enhance the visitor experience, lengthen the time spent on the platform, and increase lead conversion rates, it is also advisable to ask the development team to improve the user interface of the company's app.

In conclusion, the logistic regression model identified key factors that impact lead conversion rates and provide actionable insights for the company to improve its marketing initiatives.