SUMMARY

PROBLEM STATEMENT:

X Education has a high number of leads but a low lead conversion rate. They want to improve their lead conversion rate by identifying the most potential leads, referred to as 'Hot Leads,' to prioritize their communication efforts.

PROPOSED SOLUTION:

X Education wants to build a model that assigns a lead score to each lead, indicating the likelihood of conversion. The goal is to focus on leads with higher scores, increasing the chances of conversion. The company aims to achieve a target lead conversion rate of around 80%.

STEPS INVOLVED:

- **Importing:** In this step, we import all the required libraries that are required for building the Logistic regression model and also the data provided by the company.
- **Inspecting the Data:** We checked the shape of the data, the columns present, and the number of null values in the columns and the datatypes of these columns.
- **Data Cleaning:** Dropped the columns that had very high amount of null values (more than 40%), dropped the columns that were not really adding any value to building a model. Then dropped few category columns that were skewed.
- **Outlier Analysis:** Checked for outliers in the numerical columns and then treated the outliers via capping and flooring.
- **EDA:** Did univariate analysis to all the categorical columns and bivariate analysis to the all the columns with respect to Converted column.
- **Data Preparation:** Created dummy variables to all the categorical columns, spit the data for test and train datasets and applied standard scaling to the numerical columns.
- **Model Building:** Reduced the number of features to 15 by using RFE and later further features were reduced based on their p-values and VIF values (requirement was the p-values < 0.05 and VIF < 5).

- **Model Evaluation:** Confusion matrix was made. Later optimum cut-off value was found using ROC curve and this was used to find the accuracy, sensitivity, specificity, precision and recall.
- **Prediction on test dataset:** Used the created logistic regression module to predict on the test dataet.

Conclusion

Train Data Set:

- **Accuracy:** 80.46%
- **Sensitivity:** 80.05%
- **Specificity:** 80.71%

Test Data Set:

- **Accuracy:** 80.34%
- **Sensitivity:** 79.82% ~80%
- **Specificity:** 80.68%

Model Parameters:

Top 3 features that contributing positively to predicting hot leads in the model are:

- **Lead Source_Welingak Website**
- **Lead Source_Reference**
- **Current_occupation_Working Professional**

Recommendations:

To increase our Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement:

- Analyse negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.