UpGradLogistic Regression Case Study on ‘X Education’

Problem Statement

Current Situation:

1. Many professionals visit the website.
2. Leads are acquired, but conversion rate is low (~30%).

Challenge:

Identify high-potential leads to improve conversion rates.

Business Aspect Objective

Goal:
Develop a model to assign a lead score to each lead.

Target Conversion Rate:
Increase from 30% to 80%.

Outcome:

Focus sales efforts on 'Hot Leads' to improve efficiency and conversion rates.

Data Description

Dataset Size:

9,000 data points

Features:

Various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity.

Target Variable:

'Converted' (1 = Converted, 0 = Not Converted)

Data Exploration and Cleaning

Initial Steps:

1. Identify missing values.
2. Analyse key features for data quality issues.

Key Insights:

1. Imbalances in 'Lead Quality', 'Lead Profile'.
2. Potential issues with 'Tags', 'Current Occupation'.

Data Pre-processing

Handling Missing Values:
Imputation or dropping columns.

Converting Categorical Variables:
Numerical formats for model input.

Feature Engineering:
Creating new features and scaling numerical ones.

Model Training

Data Split:
Training and testing sets.

Model: Logistic Regression.

Training Results:
1. Significant intercept (const) with a coefficient of -2.5983.
2. VIF values within a stable range, indicating no multicollinearity issues.

Model Evaluation

Some of the most important Metrics Used:

Accuracy, Precision, Recall, ROC-AUC score.

Results:

1. ROC-AUC: 0.90 (indicating a strong model).
2. Optimum cutoff probability: 0.36.
3. Precision: 0.696
4. Recall: 0.821
5. Sensitivity: 0.821
6. Specificity: 0.796

Conclusion

Summary:

1. Systematic approach from data exploration to model evaluation.
2. Effective lead management with a reliable predictive model.

Business Impact:

Improved focus on potential leads to enhance conversion rates.