Summary Report

The assignment involved developing a logistic regression model to enhance lead conversion rates for X Education, aiming to improve the current conversion rate of 30% to approximately 80%. The goal was to identify 'hot leads' and optimize resource allocation to boost conversions.

Approach

1. Data Understanding and Exploration

The process began by exploring the dataset, which included around 9000 leads with attributes such as Lead Source, Total Time Spent on Website, and Last Activity. This involved loading the data, inspecting its structure, and identifying patterns and anomalies to understand its characteristics and relationships.

2. Data Cleaning

I addressed data quality issues by handling missing values, particularly converting 'Select' in categorical variables to NaN. Data types were corrected, and inconsistencies were resolved to prepare the dataset for analysis.

3. Exploratory Data Analysis (EDA)

EDA was performed to visualize trends and relationships using boxplots, histograms, scatter plots & heatmap. This step helped in understanding feature distributions and correlations, guiding further analysis and model development.

4. Data Preparation

Data was standardized and categorical variables were encoded into dummy variables. The dataset was split into training and testing sets using a 70/30 ratio to ensure a fair evaluation of the model’s performance.

5. Building the Logistic Regression Model

The logistic regression model was developed by performing feature selection. Automated feature selection using Recursive Feature Elimination (RFE) was combined with manual adjustments based on statistical significance (p-value < 0.05) and multicollinearity checks (Variance Inflation Factor < 5). The model was built using statistical packages to obtain detailed insights.

6. Model Evaluation

The model was evaluated using:

- **Accuracy:** 85%, showing reliable predictions.
- **Sensitivity (Recall):** 80%, indicating effective capture of leads likely to convert.
- **Specificity:** 88%, minimizing false positives and efficient resource use.
- **Precision:** 81%, reflecting accurate identification of potential converters.
- **AUC (Area Under the ROC Curve):** 0.92, demonstrating strong discrimination between converting and non-converting leads.

7. Results and Business Implications

Key findings include:

- Highly Predictive Variables: Variables such as total visits, total time spent on the website, Lead Source, lead origin, and lead quality are strong predictors of conversion. Focusing on these areas can help in targeting and nurturing leads more effectively.

- Negative Indicators: Negative coefficients associated with page views per visit, bounced emails, and certain specializations should be closely monitored. These factors may indicate areas where the company's approach needs adjustment or where additional support could be beneficial.

Learnings

1. **Data Quality:** Ensuring clean and accurate data is crucial for building a robust model.
2. **Feature Selection:** Proper feature selection and handling multicollinearity are essential for effective modeling.
3. **Model Metrics:** Using multiple evaluation metrics provides a comprehensive view of model performance.
4. **Business Impact:** Insights from the model can guide marketing strategies and resource allocation to enhance lead conversion.

This assignment demonstrated the importance of a structured approach to data analysis and modeling, highlighting how data-driven insights can significantly improve business outcomes.

Group Case study by:

1. Vaddi Pradeep Satya Chandra

2. Rishabh Salekar

3. Utkarsh Pandey