SUMMARY REPORT

Submitted by: Josemon Joy, Mohd Ibney Ali, Juli Kanaujia

Introduction

The assignment focused on developing a robust lead scoring model to predict potential lead conversion using logistic regression. The aim was to assign a lead score between 0 and 100, enabling the company to prioritize high-conversion potential leads while optimizing resources.

Methodology

Data Understanding and Preparation:

- We began by conducting exploratory data analysis (EDA), using univariate and bivariate techniques to assess variable distributions and relationships.
- Outliers in key numerical variables, such as "Total Visits," "Total Time Spent on Website," and "Page Views Per Visit," were identified and handled by setting upper limits.
- For categorical variables, categories were grouped and transformed to enhance interpretability. For example, lead origins were clustered into broader categories like “Indian” and “Non-Indian-Asians and Non- Asians.”

Feature Selection and Model Building:

- Recursive Feature Elimination (RFE) was employed to identify the most influential variables. This process resulted in 15 critical features.
- Variables with high p-values (>0.05) or multicollinearity issues (high VIF scores) were eliminated to ensure model robustness.
- Continuous variables were scaled, and the dataset was split into training and test sets for model validation.

Model Evaluation and Optimization:

- The model achieved an accuracy of 89.8% and a precision of 85.86% on the test set, correctly identifying 87% of converted customers.
- A lead scoring cutoff of 0.27 was determined by optimizing sensitivity and specificity through ROC curve analysis.

Key Insights from the model

- Certain variables, such as “Lead Source - Welingak Website” and “Tags - Lost to EINS,” significantly increased conversion probability.
- Conversely, attributes like “Lead Quality - Not Sure” and “Last Activity - Olark Chat Conversation” negatively influenced conversions.

Business Recommendations

- Prioritize high-impact leads, particularly those tagged with positive predictors.
    - Prioritize leads tagged as **"Lost to EINS"** and **"Closed by Horizon"**

    - Actively engage with leads tagged as **"Busy"** or those who will **"Revert after reading the email"**
    - Leverage the **Welingak Website**
    - Use **SMS as the last activity**
    - Develop tailored strategies for **working professionals**
- Address negative predictors by refining lead qualification processes and improving follow-up strategies.
    - Reduce the proportion of leads categorized under **"Lead Quality - Not Sure"** and **"Lead Quality - Worst"**
    - Avoid **"Last Notable Activity - Modified"** and **"Olark Chat Conversation"**
    - Implement strategies to re-engage leads tagged as **"Switched off"** or **"Ringing"**
- Leverage insights to fine-tune marketing efforts, particularly targeting working professionals and enhancing communication through SMS.

Learnings

Through this assignment, we gained several important insights:

- Practical Application of EDA: Cleaning and preparing data is critical to model performance. Handling outliers and understanding variable relationships helped improve the quality of our model.
- Feature Engineering: Clustering categorical variables into meaningful groups added depth to our analysis and model interpretation.
- Model Optimization: Using statistical techniques like RFE and assessing metrics like VIF helped refine our model for better predictions.
- Performance Evaluation: Understanding and optimizing trade-offs between sensitivity and specificity are crucial in practical applications and also precision & recall optimization
- Business Translation: The process reinforced the importance of translating technical findings into actionable business strategies, ensuring alignment with organizational goals.
- This assignment was a comprehensive exercise in combining statistical modeling with business acumen to deliver data-driven insights and recommendations.