Lead Scoring Case Study - Summary Report

1. Introduction

The objective of this assignment was to build a predictive model for lead conversion based on historical data. The dataset contained multiple variables that influenced the likelihood of conversion. The main goal was to identify key factors affecting conversions and develop a model to improve business decision-making.

2. Steps Taken

Step 1: Data Collection and Cleaning

- The dataset was loaded and explored using `Pandas`. Initial observations included checking missing values, duplicate entries, and data types.
- Necessary data cleaning steps were applied, including handling missing values and categorical encoding.
- Categorical variables were converted into dummy variables for further analysis.
- Outliers were identified and treated where necessary, using techniques such as capping and transformations.

Step 2: Exploratory Data Analysis (EDA)

- Univariate and bivariate analyses were conducted to understand distributions and relationships.
- Heatmaps and pairplots were used to visualize correlations between variables.
- The top influencing features were identified based on initial correlation values and domain understanding.

Step 3: Feature Scaling

- Continuous variables such as `TotalVisits`, `Page Views Per Visit`, and `Total Time Spent on Website` were scaled using `MinMaxScaler` to ensure uniformity.
- The impact of scaling was validated to confirm that the feature distributions remained meaningful.

Step 4: Feature Selection using RFE (Recursive Feature Elimination)

- RFE was applied using logistic regression as the base model to select the most significant predictors.
- A reduced set of features was chosen to optimize the model's performance while reducing multicollinearity.

Step 5: Checking for Multicollinearity using VIF

- Variance Inflation Factor (VIF) was calculated to check multicollinearity among independent variables.
- Features with high VIF were removed iteratively to improve model stability and prevent redundancy.

Step 6: Model Building using Logistic Regression

- A logistic regression model was trained on the refined feature set.
- Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score.

Step 7: Model Evaluation using Confusion Matrix

- The confusion matrix was used to determine True Positives, True Negatives, False Positives, and False Negatives.
- The default probability cutoff of 0.5 was analyzed, and further fine-tuning was performed using ROC curves to select an optimal threshold.

3. Key Learnings

- **Feature Selection Matters**: Using RFE and VIF helped in selecting the most significant predictors while eliminating redundant variables, leading to a more interpretable and efficient model.
- **Scaling is Crucial**: Standardizing numerical features using `MinMaxScaler` ensured that all features contributed appropriately to the model.
- **Multicollinearity Impacts Performance**: High VIF values indicated that some variables were highly correlated, which could lead to overfitting. Eliminating these improved model generalizability.
- **Cutoff Optimization is Key**: The default probability cutoff of 0.5 may not always be ideal. Fine-tuning using ROC analysis can help achieve better classification results.
- **Lead Scoring Insights**: The most influential factors for lead conversion were **Total Time Spent on Website, Lead Origin (Lead Add Form), and Working Professional status**.

4. Conclusion

This assignment provided hands-on experience in data preprocessing, feature selection, and model evaluation. By applying logistic regression effectively, we developed a model that helps businesses prioritize leads with a high likelihood of conversion. The learnings from this exercise can be extended to real-world scenarios where lead scoring is essential for sales and marketing strategies.