Summary Report: Analysis and Model Creation

Introduction

The objective of this assignment was to develop a predictive model for customer conversion using a dataset containing various features related to customer interactions and characteristics. The approach involved thorough data exploration, preprocessing, feature selection, model building, and evaluation. This summary outlines the steps taken and key insights gained throughout the process.

Data Preparation and Exploration

1. Data Exploration: The initial step involved loading and exploring the dataset to understand its structure and content. Key variables included customer interactions such as total time spent on the website, total visits, and page views per visit, along with categorical variables like lead source and city. This exploration provided a foundational understanding of the dataset's distribution and quality.

2. Data Cleaning: The next phase focused on cleaning the data by addressing missing values and detecting outliers. Outlier analysis was performed using box-plots, particularly for numerical variables like total time spent on the website, total visits, and page views per visit. Outliers were handled by removing extreme values based on quantile thresholds, which helped in stabilising model performance.

Data Transformation

3. Encoding Categorical Variables: To prepare the data for modelling, categorical variables were encoded. Binary encoding was used for binary attributes, while one-hot encoding was applied to other categorical variables. This transformation converted categorical features into numerical format suitable for regression analysis.

4. Scaling Features: Standard scaling was applied to numerical features to ensure that all variables contributed equally to the model, thus improving convergence and performance.

Feature Selection

5. Recursive Feature Elimination (RFE): Recursive Feature Elimination was employed to identify the most significant features for the model. An initial model using logistic regression was fitted, and features were ranked based on their importance. Features with high p-values and those contributing to high Variance Inflation Factors (VIF) were removed to mitigate multicollinearity.

6. Model Refinement: Through iterative feature removal based on statistical significance and VIF values, the model was refined to include only the most relevant features, improving interpretability and performance.

Model Building and Evaluation

7. Logistic Regression Model: The refined dataset was used to build a logistic regression model. The model was evaluated using performance metrics such as accuracy, sensitivity, specificity, and the ROC-AUC score. The optimal cutoff for classification was determined to balance sensitivity and specificity, which was crucial for practical application.

8. Model Performance: The model's performance was assessed on both training and test datasets. Training accuracy, confusion matrix, and ROC curve analysis provided insights into the model's effectiveness. Test accuracy and confusion matrix results validated the model's performance on unseen data, ensuring its robustness and generalisability.

Learnings and Insights

1. **Data Preparation**: Proper handling of outliers and encoding of categorical variables are crucial steps in preparing data for modelling. Outlier removal can significantly impact model stability and accuracy.

2. **Feature Selection**: Feature selection is an iterative process that involves evaluating statistical significance and multicollinearity. RFE proved effective in identifying key features, enhancing model interpretability and performance.

3. **Model Evaluation**: Logistic regression is a useful model for binary classification, but its performance can be further enhanced by exploring additional models like Random Forests or Gradient Boosting. Evaluating the model using various metrics (accuracy, sensitivity, specificity, ROC-AUC) provides a comprehensive understanding of its effectiveness.

4. **Optimisation**: The choice of an optimal cutoff for classification is essential to balance between sensitivity and specificity, depending on the business context and objectives.

In conclusion, the assignment underscored the importance of thorough data preprocessing, feature selection, and careful model evaluation. These steps collectively contribute to building a robust and reliable predictive model. Future improvements could involve experimenting with advanced algorithms and hyper-parameter tuning to further enhance model performance.