Lead Conversion Analysis Project

Project Overview

This project aims to analyze lead data to identify patterns and factors influencing lead conversion. Utilizing a dataset containing information on various leads, the project follows a structured workflow encompassing data cleaning, preprocessing, exploratory data analysis (EDA), model training, and evaluation. The goal is to build a predictive model that can accurately identify potential conversions and uncover the most influential factors affecting conversion rates.

Data Preparation and Initial Inspection

- **Libraries Used**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.
- **Initial Data Inspection**: Functions were implemented to display the dataset's head, tail, info, description, and percentage of missing values in each column.

Data Cleaning and Preprocessing

- **Missing Values**: Columns with more than 30% missing values were dropped. Missing values in other significant columns were imputed based on insights gained from data visualization, such as replacing missing 'Specialization' values with 'Others'.
- **Data Imputation**: Visual analysis guided the imputation strategy for columns with significant missing values, ensuring a more informed approach to maintaining data integrity.

Exploratory Data Analysis (EDA)

- **Target Variable Distribution**: The target variable 'Converted' showed a moderate imbalance, indicating the need for careful model evaluation metrics.
- **Feature Analysis**: Univariate and bivariate analyses provided insights into the distribution and impact of various features on lead conversion. Notable findings include the influence of 'Total Time Spent on Website', 'Lead Origin', and 'Lead Source' on conversion rates.
- **Feature Engineering**: Based on EDA findings, new features were engineered (e.g., combining 'TotalVisits' and 'Page Views Per Visit') to enhance the predictive model.

Data Engineering for Model Training

- **Preprocessing Pipeline**: A comprehensive preprocessing pipeline was set up for numerical and categorical data, including median imputation, standard scaling, and one-hot encoding.
- **Feature Transformation**: The pipeline ensured that the data was appropriately transformed for model training, with categorical variables converted through one-hot encoding and numerical variables standardized.

Model Training and Evaluation

- **Logistic Regression Models**: Multiple logistic regression models were trained using Recursive Feature Elimination (RFE) to select an optimal number of features.
- **Model Evaluation**: Models were evaluated based on accuracy, sensitivity, and specificity. Cross-validation and classification reports provided insights into model performance across different feature sets.

Dataset Synthesis and Model Creation

- **Synthetic Data Generation**: The project incorporated synthetic data generation to evaluate model performance across datasets of varying sizes and complexities.

Insights from Model Coefficients

- **Feature Importance**: Analysis of model coefficients revealed the most influential features affecting lead conversion. 'Total Time Spent on Website', 'Do Not Email_Yes', and geographical factors were among the top predictors.
- **Predictive Factors**: The findings underscored the importance of user engagement metrics, communication preferences, and lead origin in determining lead conversion likelihood.

Conclusions and Recommendations

- **Key Predictors of Lead Conversion**: The project identified critical factors influencing lead conversion, such as website engagement and specific lead sources, providing actionable insights for marketing strategies.
- **Strategic Recommendations**: Businesses can enhance lead conversion rates by focusing on optimizing the user experience on the website, tailoring communication strategies based on lead preferences, and targeting leads from high-converting sources.