Detailed Report on Claimants Based on Predictive Analysis

Introduction

This report presents a detailed analysis of claimants based on a dataset, focusing on predicting whether a claimant is represented by an attorney. The analysis includes data diagnosis, handling missing values, outlier detection and treatment, and applying various classification algorithms to predict the target variable 'ATTORNEY'.

Data Overview

The dataset consists of several columns, each representing different attributes of claimants. Below are the key attributes considered in the analysis:

- **ATTORNEY**: Whether the claimant is represented by an attorney (1) or not (0).
- **CLMSEX**: Gender of the claimant (1: Male, 0: Female).
- **CLMINSUR**: Whether the claimant is insured (1) or not (0).
- **SEATBELT**: Whether the claimant was wearing a seatbelt (1) or not (0).
- **CLMAGE**: Age of the claimant.
- **LOSS**: Financial loss incurred.

Data Diagnosis

Missing Values

The dataset contains missing values in several columns. A detailed check reveals the following:

- **CLMSEX**: 12 missing values (0.9%).
- **CLMINSUR**: 41 missing values (3.1%).
- **SEATBELT**: 48 missing values (3.6%).
- **CLMAGE**: Missing values detected.
- **LOSS**: Missing values detected.

Handling Missing Values

A custom imputation strategy was applied:

- **Numerical Columns**: Imputed using the mean.
- **Categorical Columns**: Imputed using the mode.
- **Datetime Columns**: Imputed using interpolation.

Outlier Detection

Outliers were detected and treated in numerical columns using the Interquartile Range (IQR) method. Outliers were replaced with the median values to maintain the robustness of the models.

Predictive Analysis

Imbalance in Data

The dataset showed significant imbalance, particularly in the 'ATTORNEY' column. The target variable has the following distribution:

- **0 (Not represented by an attorney)**: 51.1%
- **1 (Represented by an attorney)**: 48.9%

Resampling

To address the imbalance, RandomUnderSampler was applied to balance the data, ensuring that the predictive models are not biased towards the majority class.

Classification Results

Several classification algorithms were applied to predict whether a claimant is represented by an attorney. The performance metrics for each algorithm are summarized below:

Algorithm	Accuracy	F1-Score	Precision	Recall
Decision Tree Classifier	0.5746	0.5747	0.5749	0.5746
Gradient Boosting Classifier	0.7015	0.7016	0.703	0.7015
K-Nearest Neighbors Classifier	0.6828	0.6826	0.6826	0.6828
Logistic Regression	0.7164	0.7119	0.7237	0.7164
Random Forest Classifier	0.6418	0.6418	0.6418	0.6418
XGBoost Classifier	0.653	0.6521	0.6526	0.653

Best Performing Model

- **Logistic Regression** emerged as the best-performing model with an accuracy of 0.7164, an F1-Score of 0.7119, precision of 0.7237, and recall of 0.7164. This indicates Logistic Regression is well-suited for this binary classification problem.

Conclusions

1. **Data Imbalance**:
    - Significant imbalance in the dataset required resampling to ensure balanced training data for predictive models.
2. **Handling Missing Values**:
    - Custom imputation strategies effectively handled missing values, ensuring data completeness.

3. **Outlier Detection and Treatment**:

    - Outliers were detected and treated using median imputation to maintain data robustness.

4. **Model Performance**:

    - Logistic Regression was identified as the most effective model for predicting attorney representation, outperforming other tested algorithms.

5. **Recommendations for Future Work**:

    - Further improvements could include advanced resampling techniques, feature engineering, hyperparameter tuning, and validating models on separate validation sets for better real-world applicability.

This report provides a comprehensive overview of the data analysis and predictive modelling efforts to understand and predict whether claimants are represented by attorneys, helping inform decision-making processes and improve data-driven strategies.