Data Explorer Team Details

Group Name: Data Explorer

Name: Mohammad Tohin Bapari

Email: tohin@gmx.de

Country: Germany

University: Bergische Universität Wuppertal

Specialization: Data Science

Problem Description

Pharmaceutical companies need to understand the persistency of drug usage as per physician prescriptions to improve patient adherence. ABC Pharma aims to automate the process of identifying factors impacting persistency. The goal is to build a classification model to predict whether a patient will be persistent or non-persistent based on various demographic, clinical, and treatment-related factors.

Data Understanding

We have a dataset provided by ABC Pharma, which consists of 69 columns with various features related to patient demographics, provider attributes, clinical factors, and adherence details. The target variable is Persistency_Flag which indicates whether a patient is persistent or not in their medication adherence.

Type of Data for Analysis

The dataset includes:

- **Demographic Information**: Age, Gender, Race, Ethnicity, Region
- **Provider Attributes**: Physician Specialty
- **Clinical Factors**: T-Score, Risk Segment, DEXA Scan Frequency, Glucocorticoid Usage
- **Adherence Information**: Details on therapy adherence

The data is provided in an Excel file format with two sheets:

1. **Feature Description**: Contains descriptions of each feature.
2. **Dataset:** Contains the actual data for analysis.

Dataset Overview

1. Number of rows and columns.
2. Data types.
3. Statistical summary for numerical and categorical columns.
4. Unique values and frequency counts.

Expletory data Analysis of Dataset

Dataset Overview

Description	Value
Number of Rows	3424
Number of Columns	69
Data Types	object

Numerical Summary

Statistic	Count	Mean	Std	Min	25%	50%	75%	Max
Count_Of_Risks	3424	1.23	1.41	0	0	1	2	10

Categorical Summary (Top 5 Examples)

Column Name	Unique Values	Top	Frequency	Null Values
Persistency_Flag	2	Non-Persistent	2435	0
Gender	2	Female	1854	0
Race	4	Caucasian	2320	0
Ethnicity	2	Not Hispanic	3190	0
Region	5	South	1392	0
Age_Bucket	4	>75	1424	0
Ntm_Speciality	22	GENERAL PRACTITIONER	2810	0

Column Name	Skewness
Count_Of_Risks	1.550

Handling Missing Values (NA values)

1. **Remove Rows with Missing Values:**
   - **Approach:** Drop rows where missing values are present.
   - **Why:** This approach is simple and effective when the dataset is large and the number of missing values is small. It prevents introducing bias or errors from imputed values.
2. **Remove Columns with Missing Values:**
   - **Approach:** Drop columns that have a significant number of missing values.
   - **Why:** If a column has a large proportion of missing values, it may be less informative or problematic to impute. Removing such columns can help simplify the model without much loss of information.
3. **Impute Missing Values:**
   - **Approach:** Fill in missing values using techniques like mean, median, mode, or more sophisticated methods such as K-nearest neighbors (KNN) or regression.
   - **Why:** Imputation allows for the retention of all data points and can reduce bias introduced by simply removing rows or columns. It is particularly useful when the amount of missing data is not too high.

4. **Use Algorithms that Support Missing Values:**
    - **Approach:** Use machine learning algorithms that handle missing values natively, such as XGBoost.
    - **Why:** These algorithms can manage missing data internally, reducing the need for preprocessing.

Handling Outliers

1. **Remove Outliers:**
    - **Approach:** Identify and remove outliers using statistical methods (e.g., Z-score, IQR).
    - **Why:** Outliers can skew the results of an analysis and lead to misleading conclusions. Removing them can help improve the performance and accuracy of the model.
2. **Transform Data:**
    - **Approach:** Apply transformations such as log, square root, or Box-Cox to reduce the impact of outliers.
    - **Why:** Transformations can help normalize the distribution and reduce the influence of extreme values.
3. **Impute Outliers:**
    - **Approach:** Replace outlier values with a measure such as the mean, median, or a value based on neighboring data points.
    - **Why:** This approach can mitigate the impact of outliers without losing any data points.
4. **Use Robust Algorithms:**
    - **Approach:** Use algorithms that are less sensitive to outliers, such as tree-based methods (e.g., Random Forest, Gradient Boosting).
    - **Why:** These algorithms can handle outliers more effectively and can still produce accurate results even in their presence.

Why These Approaches

- **Preserve Data Integrity:** Removing or imputing values carefully ensures that the dataset remains as informative and accurate as possible.
- **Reduce Bias:** Proper handling of missing values and outliers minimizes the introduction of bias, which could affect model performance.
- **Improve Model Performance:** Clean and well-prepared data can significantly enhance the performance of machine learning models.
- **Maintain Simplicity:** By removing problematic data points (either rows or columns), the dataset becomes simpler to work with, making the analysis and modeling process more efficient.

Implementing These Approaches

For our dataset, since there are no missing values identified, we don't need to handle NAs. However, we identified skewness in the "Count_Of_Risks" column. Here are potential steps to address it:

1. **Transform the Skewed Data:**
    - Apply a log or square root transformation to reduce skewness.
2. **Check for Outliers:**
    - Identify any extreme values in the "Count_Of_Risks" column using statistical methods and decide whether to remove or impute them.