Team member's details:
Group Name: **Data Dominators**

chau.devin031602@gmail.com	Devin Chau	United States	San Jose State University	Data Science

rohankhatri0507@gmail.com	Rohan Khatri	United States	San Jose State University	Data Science

ethan05dy@gmail.com	Ethan Dy	United States	San Jose State University	Data Science

Project: Data Science:: Healthcare - Persistency of a drug

Problem description

We are building a predictive model that classifies patients into "persistent" or "non-persistent" categories based on factors like their demographics, medical history, physician characteristics, and treatment details. Factors like the patient level such as their age, risk factors, previous test results, or provider type allows for insights into why some patients continue therapy while others drop off. Thus understanding "persistence" levels. By analyzing these data points and finding patterns, the predictive model helps explain patient behavior and supports the creation of targeted interventions to improve adherence.

Exploratory Data Analysis (EDA)

During the exploratory data analysis (EDA), we performed the following steps:

- **Data Visualization**: Visualized distributions of numerical features such as Dexa_Freq_During_Rx and Count_Of_Risks.
    - Histograms showed high skewness in Dexa_Freq_During_Rx, leading to the decision to apply log transformation for normalization

- **Correlation Analysis**: Calculated correlations between numerical variables. Count_Of_Risks showed a moderate correlation with the target variable Persistency_Flag

- **Categorical Feature Inspection**: Explored distributions for categorical variables like Ntm_Speciality and Gender. Identified the imbalanced classes in Ntm_Speciality which could maybe lead to overfitting (?)

- **Outlier Detection**: Used the Interquartile Range (IQR) method to detect outliers. Dexa_Freq_During_Rx had 460 extreme values,
    - while Count_Of_Risks had 8 outliers (addressed using capping methods)

- **Handling Missing Values**: Discovered multiple columns with the value "Unknown" instead of NaNs, indicating hidden missing data. Handled as separate categories

Final Recommendations

Based on the analysis and identified data issues, we recommend:

- **Handling Missing Values**: Maintain “Unknown” entries as a separate category to preserve data integrity and avoid loss of potentially valuable patterns

- **Normalization**: Continue using MinMaxScaler for numerical features like Dexa_Freq_During_Rx and Count_Of_Risks to ensure consistent scaling across variables

- **Outlier Handling**: Use the IQR method to cap extreme values for Dexa_Freq_During_Rx and Count_Of_Risks

- **Categorical Encoding**: Apply one-hot encoding for categorical variables with multiple categories. Using binary encoding for columns with simple Y/N values

- **Feature Engineering**: Consider grouping rare categories in columns like Ntm_Speciality to avoid overfitting due to high cardinality

- **Data Consistency**: Ensure proper standardization of data types and consistency across the entire dataset before model training

Github Repo link

https://github.com/rohankhatri7/DataGlacier-Internship/tree/main/Week10