Project : Healthcare - Persistency of a drug

Week 10 Deliverable

Batch: LISUM38

Team: Salus AI Data Avengers				
Name	**Email**	**Country**	**Company**	**Specialisation**
Amrin Shaikh	amrin02@gmail.com	United Kingdom	Freelance	Data Science
Keilor Fallas Prado	kfallasprado@gmail.com	Costa Rica	Immune Technology Institute Madrid	Data Science
Hyejoon Lee	candy0543@gmail.com	United Kingdom	--	Data Science
Raina Singh	rainasinghh1@gmail.com	United States of America	The Ohio State University	Data Science

Problem description:

The objective of this analysis is to explore and analyse pharmaceutical data related to the persistency of a drug. The goal is to develop a classification model that helps identify medication adherence patterns. The target variable for this analysis is the Persistency Flag, which indicates whether a patient is persistent or non-persistent in taking their medication. The dataset contains various demographic and clinical features such as age, gender, race, ethnicity, risk segments, adherence flags, and medical data. The problem focuses on identifying trends and relationships that influence persistency to aid in improving medication adherence.

Exploratory Data Analysis (EDA):

A thorough EDA was conducted to understand the dataset's structure, identify patterns, and prepare the data for modelling. Below are the key findings:

❖ Data Loading and Overview

The dataset was loaded, and initial inspection was performed using basic descriptive statistics and visualizations.

- **Target Variables:** Persistency_Flag (Persistent/Non-Persistent)
- **Key Demographic Variables:** Gender, Race, Ethnicity, Region, Age
- **Risk Factors:** Prior NTM Risk Segment, During Rx Risk Segment
- **Numerical Variables:** Dexa_Freq_During_Rx, Count_Of_Risks.

❖ Data Visualization

- **Persistency Flag Distribution:** A bar chart showing a significant imbalance between Persistent (1289) and Non-Persistent (2135) cases, suggesting that non-adherence is more prevalent.

- **Correlation Heatmap:**
  - Minimal correlation (0.01) between Dexa_Freq_During_Rx and Count_Of_Risks, indicating no significant relationship between DEXA scan frequency and risk count.
  - Strong positive correlation (0.65) between Prior and During Rx Risk Segments, suggesting a close relationship between historical and current risk levels.

- **Demographic Distributions:** Gender, Race, Ethnicity, and Region distributions were analysed. Key observations included:
  - A majority of the dataset is female, with 3230 females compared to 194 males.
  - Racial distribution is skewed towards Caucasians (3148), with small counts of other races (Asian, African American, and Other/Unknown).
  - Most individuals are from the Midwest and South regions, with a smaller count from other regions.

❖ Outlier Detection and Treatment

- Isolation Forest and IQR Method were used to detect outliers. The Isolation Forest method flagged 172 records as outliers across all numerical columns, with the IQR method flagging 460 for the "Dexa_Freq_During_Rx" variable.
- Outliers were addressed using log and square root transformations to make the data more normal and reduce the influence of extreme values on model performance.

❖ Skewness and Transformation

- Dexa_Freq_During_Rx was highly skewed before transformation (skewness of 6.81), but after applying a log transformation, the skewness reduced to 1.33.
- Count_Of_Risks had minimal skewness, and the square root transformation did not flag significant outliers.
- These transformations helped in preparing the data for modelling by improving normality and minimizing outlier influence.

❖ Categorical Variable Encoding

- One-Hot Encoding was used for nominal variables (e.g., Gender, Race, Ethnicity).
- Label Encoding was applied to ordinal variables (e.g., Age Bucket, Risk Segments).
- Binary variables were encoded with 0 and 1.

Final Recommendations:

Based on the EDA and the patterns observed in the dataset, the following recommendations are made:

- **Targeted Interventions:** Focus on Non-Persistent individuals, as they represent a significant portion of the dataset. Special attention should be given to improving medication adherence, especially among female patients, Caucasians, and individuals from the Midwest and South regions.
- **Address Disparities:** Investigate the factors contributing to lower persistence in Caucasians, Non-Hispanic individuals, and older adults. Identifying systemic barriers or personal challenges could guide better support strategies.

- **Improving Risk Management:** Since risk segments during treatment are closely related to prior risk segments, integrating this information into personalized treatment plans could help reduce non-adherence. Ensuring that patients with higher risk levels receive consistent follow-ups may improve persistence.
- **Data Quality Improvement:** Some features have "Unknown" entries (especially for ethnicity and region), which could be contributing to biases. Future data collection should aim to reduce missing or unclear information to improve predictive accuracy.
- **Further Feature Engineering:** Applying more advanced transformations like Box-Cox or Winsorization for skewed data can improve model performance. Additionally, ensuring that features like Age and Risk Segments are thoroughly understood and cleaned may help in more accurate predictions.
- **Modelling Considerations:** Given that tree-based models like Random Forest and Gradient Boosting are robust to skewed data, they would be appropriate for modelling persistency behaviour. However, testing linear models may still be valuable after additional transformations to reduce skewness.

Conclusion:

The analysis of the drug persistency dataset has revealed significant trends and correlations, pointing toward areas where healthcare systems can intervene to improve patient adherence. The data preparation process involved addressing skewness, outliers, and categorical encoding, ensuring the dataset is ready for modelling. Further efforts should focus on improving data quality and developing predictive models to optimize patient care.