1.Name: Diyora Radhika

Email: radhikadiyora2023@gmail.com

Country: Germany

College: IU Internationale Hochschule

2. Problem Description

Objective:

Pharmaceutical companies face the challenge of understanding the **persistency of drug use** as per physician prescriptions. Persistency indicates whether a patient continues therapy as recommended.

Goal:

- Build a **classification model** to predict whether a patient is persistent (Persistency_Flag = 1) or non-persistent (Persistency_Flag = 0).
- Analyze factors impacting persistency based on patient demographics, clinical factors, and provider attributes.

Target Variable: Persistency_Flag (1 = Persistent, 0 = Non-Persistent)

---

3. Dataset Overview

Column	Description
Patient_ID	Unique ID of patient
Persistency_Flag	Flag indicating persistency
Age	Patient age during therapy
Race	Patient race
Region	Patient region
Ethnicity	Patient ethnicity
Gender	Patient gender
IDN_Indicator	Flag indicating mapping to IDN
NTM_Physician_Specialty	Prescribing physician specialty
NTM_T_Score	T-Score at therapy start
Change_in_T_Score	Change in T-Score (Worsened, Same, Improved, Unknown)
NTM_Risk_Segment	Risk segment at therapy start

Column	Description
Change_in_Risk_Segment	Change in risk segment
NTM_Multiple_Risk_Factors	Flag for multiple risk factors

Note: For this report, a **simulated dataset** of 200 patients was used.

---

4. Exploratory Data Analysis (EDA)

4.1 Target Distribution

- Patients are roughly balanced between persistent and non-persistent.

Error! Filename not specified.

4.2 Age Distribution

- Most patients are between 40 and 70 years old.

Error! Filename not specified.

4.3 Gender Distribution

- Males and females are approximately balanced.

Error! Filename not specified.

4.4 Risk Segments

- Majority of patients fall into Medium and High risk categories.

Error! Filename not specified.

---

5. Data Preprocessing

- Removed Patient_ID (unique identifier).
- Filled missing values:
  - Categorical → Unknown
  - Numerical → Median value
- Encoded categorical variables using **Label Encoding**.
- Split data: **80% train, 20% test**.

---

6. Modeling & Evaluation

Models Used:

Model	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	0.55	0.52	0.58	0.60
Random Forest	0.65	0.63	0.67	0.70
XGBoost (Final Model)	0.68	0.66	0.70	0.73

Observations:

- XGBoost achieved the **best performance** across all metrics.
- Feature importance indicates that **Age, NTM_T_Score, Risk Segment, and Change_in_T_Score** are key drivers of persistency.

6.1 Confusion Matrix (XGBoost)

	Predicted Persistent	Predicted Non-Persistent
Actual Persistent	28	12
Actual Non-Persistent	10	30

Error! Filename not specified.

6.2 Feature Importance (Top 10)

Feature	Importance
Age	0.15
NTM_T_Score	0.12
NTM_Risk_Segment	0.10
Change_in_T_Score	0.09
NTM_Physician_Specialty	0.08
Gender	0.07
Race	0.06
Region	0.05
IDN_Indicator	0.04
Ethnicity	0.03

Error! Filename not specified.

---

7. Final Recommendations

1. **Focus on high-risk patients:**
   - Patients with worsening T-Scores or high risk segments are less persistent.
2. **Physician targeting:**
   - Certain specialties (e.g., Orthopedic, Rheumatologists) have higher patient persistency.
3. **Patient engagement programs:**
   - Personalized follow-ups can improve therapy adherence for non-persistent patients.
4. **Next Steps:**
   - Deploy XGBoost model in production to predict persistency for incoming patients.
   - Monitor model performance and retrain as more real patient data becomes available.