Team Member Details

Group Name: *SoloVision Analytics*

Name: Ankita Roy
Email: roy.anki39@gmail.com
Country: India
College/Company: Institute: Great Learning, Gurgaon/ Internship: Data Glacier
Specialization: Data Science

Problem Description

Drug persistency is a key metric in determining the long-term success of a treatment regimen. It reflects whether patients continue using a prescribed therapy as advised. Pharmaceutical companies like ABC Pharma often struggle to identify the factors that cause patients to discontinue treatment.

To address this, ABC Pharma has sought the help of data science to automate the detection of persistency patterns among patients using their drug. The aim is to develop a **classification model** that predicts whether a patient is likely to persist with the treatment based on demographic, clinical, and behavioural data.

The binary target variable for this task is **Persistency_Flag**, where the goal is to classify patients as either **persistent** or **non-persistent**.

EDA Performed on the Data

- **Univariate Analysis:**
  - **Target:** Pie chart confirmed a moderate class imbalance
  - **Numerical columns:** Dexa_Freq_During_Rx distribution and outliers analysed with displot and boxplot
  - **Categorical columns:** Countplots for Gender, Race, Ethnicity, Region and other categorical columns to understand distributions and imbalance
- **Bivariate Analysis:**
  - **Numeric vs Target:** Displots and boxplots showed higher Dexa scan frequency linked to higher persistency
  - **Category vs Category:** Grouped bar plots showed relationships between persistency and categorical features
- **Statistical Testing:**
  - Shapiro – Wilk test indicated non-normal distribution
  - Levene's test showed unequal variances
  - Mann – Whiteney U test confirmed significant difference in Dexa scan frequency between groups ($p < 0.05$)

Key Insights: Higher Dexa scan frequency strongly correlates with higher persistency – to be retained as an important predictor.

Modelling Performed

The dataset was used to train multiple families of classification models to find the best fit. The following models were evaluated:

Model Family	Model	Accuracy	Precision	Recall	F1 Score	Cohen Kappa
Linear	Logistic Regression	~80%	~78%	~66%	~72%	~0.56
Linear Weighted	Logistic Regression (class weight)	~80%	~74%	~74%	~74%	~0.58
KNN	K-Nearest Neighbours	~78%	~81%	~57%	~67%	~0.51
Tree	Decision Tree	~78%	~72%	~71%	~71%	~0.53
Ensemble	Random Forest	~80%	~81%	~65%	~72%	~0.57
Boosting	Gradient Boost	~81%	~80%	~69%	~74%	~0.59
Boosting	XG Boost	~81%	~79%	~70%	~74%	~0.59
Ensemble	Voting Classifier	~81%	~78%	~71%	~74%	~0.59
Boosting Final	XG Boost	~82%	~80%	~70%	~74%	~0.60

Boosting algorithms i.e., Gradient Boost and XG Boost as well as an Ensemble method i.e., Voting Classifier provided the best balance of all scores i.e., precision, recall and F1 score as well as accuracy. XG Boost was selected as final model after fine tuning due to its faster training time and robust performance.

Final Recommendation

- **Selected Model:** XG Boost Classifier
- **Rationale:** High F1 score, reasonable interpretability through feature importance plots, fast execution
- **Next Steps for Improvement:** To further enhance model performance, we recommend exploring additional hyperparameter tuning and possibly collecting more data to handle class imbalance or rare cases
- **Implementation Readiness:** The prepared models are ready to be integrated into further analysis pipelines
- **Monitoring & Review:** It is advised to monitor the model's performance periodically to ensure it stays aligned with changing business scenarios and data patterns

These recommendations, if implemented, will help the business team make data – driven decisions to improve patient persistency outcomes and optimize resource allocation.

Deliverables

- **Codebase:** Complete pipeline, feature engineering, EDA, model training, hyperparameter tuning code uploaded
- **PowerPoint:** Separate slide deck summarizes business problem, EDA visuals, key statistical findings, modelling results and final recommendation

GitHub Repository Link

https://github.com/Ankita-ar/PERSISTENCY-OF-A-DRUG

NOTE: This is a solo project, all tasks – from business understanding to importing the data to EDA till model building and fine tuning were completed single handed.