Final Presentation

Saif Amer

Project Overview

Goal: Predict whether a patient will remain Persistent or become Non-Persistent with a prescribed drug regimen.

Dataset: pharma_data.csv with 3,424 patients and 69 features

Data Understanding

- **Target Variable: Persistency_Flag**
- **Feature Types:**
  - **Categorical: 67 columns (e.g. Gender, Race, Region, Risk Factors)**
  - **Numeric: 2 columns (Dexa_Freq_During_Rx, Count_Of_Risks)**

No missing values in this dataset. Good data quality.

Data Preparation

Steps taken:

1. Label Encoding the target
2. Categorical Imputation: Most Frequent
3. Numeric Imputation: Median
4. Encoding: OneHotEncoder for categorical features

Split:

- 80% Training
- 20% Testing

Model Building

Model Used: Random Forest Classifier

Pipeline Includes:

- Data preprocessing
- Model training

Hyperparameter Tuning:

- Performed using GridSearchCV (5-fold cross-validation)
- Parameters searched:
  - n_estimators: [100, 200]
  - max_depth: [5, 10, 20]
  - min_samples_split: [2, 5]

Best Parameters

Best Parameters:

- n_estimators = 200
- max_depth = 10
- min_samples_split = 2

These parameters gave the best ROC AUC on validation folds

Final Model Evaluation

On Test Data:

- Accuracy: **81.6%**
- Precision: **80.6%**
- Recall: **67.4%**
- ROC AUC: **87.8%**

Confusion Matrix:

- True Negatives: 385
- False Positives: 42
- False Negatives: 84
- True Positives: 174

ROC Curve

- ROC Curve shows a high AUC (0.878)
- Indicates good separation between Persistent and Non-Persistent classes

Graph Highlights:

- X-axis: False Positive Rate
- Y-axis: True Positive Rate
- Diagonal line: Random guess
- Our curve: Above the diagonal (better than random)

Key Takeaways

- Data quality was high (no missing values)
- Random Forest was effective for classification
- Strong model performance with AUC ~88%
- Model may improve with advanced feature engineering or ensemble methods

Next Steps

- Explore SHAP/LIME for interpretability
- Try other models: XGBoost, LightGBM
- Deploy with Flask/Streamlit for end-user access
- Monitor model performance on new patient data