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.

Recommendation and Proposed Solution
To address the challenge of predicting whether patients will remain persistent on their therapy or become non-persistent, we recommend implementing a classification model using **XGBoost** (Extreme Gradient Boosting).

- **Regularization**:
  - XGBoost incorporates both L1 (Lasso) and L2 (Ridge) regularization techniques, which help reduce overfitting a common issue in healthcare prediction models where numerous features can lead to overly complex models
- **Handling Complexity:**
  - By using a maximum tree depth parameter, XGBoost can effectively navigate complex relationships within the data while mitigating the tendency to fit noise
- **Performance and Speed**:

- XGBoost is known for its speed and scalability, making it efficient for large datasets. It also provides robust tools for parallel computation, which can expedite training

Model Implementation and Findings

Our model was trained on patient demographics, medical history, physician characteristics, and treatment details to classify patients as “persistent” or “non-persistent.” We used a standard train-test split (70:30 ratio) and evaluated performance on the set

- **Label Encoding**: Converted categorical variables (e.g., age group, provider type, and other features) into numeric form
- **Train-Test Split**: Segregated our dataset into training and testing subsets to validate the model’s performance
- **XGBoost Classifie**r: Initialized with use_label_encoder=False and an evaluation metric of logloss, then trained on the prepared dataset

Here are our results:

Accuracy: 79.77%
Precision for class 0 (Non-Persistent): 0.83
Recall for class 0 (Non-Persistent): 0.85
Precision for class 1 (Persistent): 0.73
Recall for class 1 (Persistent): 0.70

	precision	recall	f1-score	support
0	0.83	0.85	0.84	654
1	0.73	0.70	0.72	374
accuracy			0.80	1028
macro avg	0.78	0.78	0.78	1028
weighted avg	0.80	0.80	0.80	1028

XGBoost accuracy: 0.7976653696498055

The overall weighted f1-score stands at 0.80, indicating a moderate yet promising performance for our initial implementation. Despite these decent results, there is room to enhance the accuracy. Especially for identifying persistent patients, since higher recall and precision in the persistent class are crucial for practical healthcare applications.

Future Enhancements

Although XGBoost yielded acceptable performance, we believe further refinements can improve accuracy and reliability. Potential avenues for future exploration include:

- **Hyperparameter Tuning**

    - Fine-tune learning rate, max depth, subsampling ratio, and regularization parameters to optimize model performance

- **Ensemble Learning**
    - *Blending*: Combine the predictions of multiple base models to leverage their strengths
    - *Bagging*: Use techniques like Random Forest or bootstrap aggregating to reduce variance
    - *Stacking*: Layer multiple models, using the outputs of some as inputs to others for improved predictive power

- **Autoencoder-Generated Features**
    - Integrate deep learning approaches to automatically extract high-quality features from the existing data, potentially unveiling hidden patterns missed by traditional feature engineering

Conclusion

Through these steps, we aim to enhance the model's predictive power for patient persistence, providing more accurate insights that can aid in designing targeted interventions. Such improvements in prediction will support better healthcare outcomes by identifying at-risk individuals and enabling interventions to keep patients on critical therapies.