Model Selection and Model Building

Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh

Group Name - HealthData Innovators

Name – Apurwa Bhausaheb Sontakke

Email – sontakke.ap@northeastern.edu

Country - USA

College– Northeastern University

Specialization – Data Science

GitHub Link - https://github.com/apurwasontakke/Week-12-Data-Science-Healthcare-Project--Model-Selection-and-Building

Problem Description

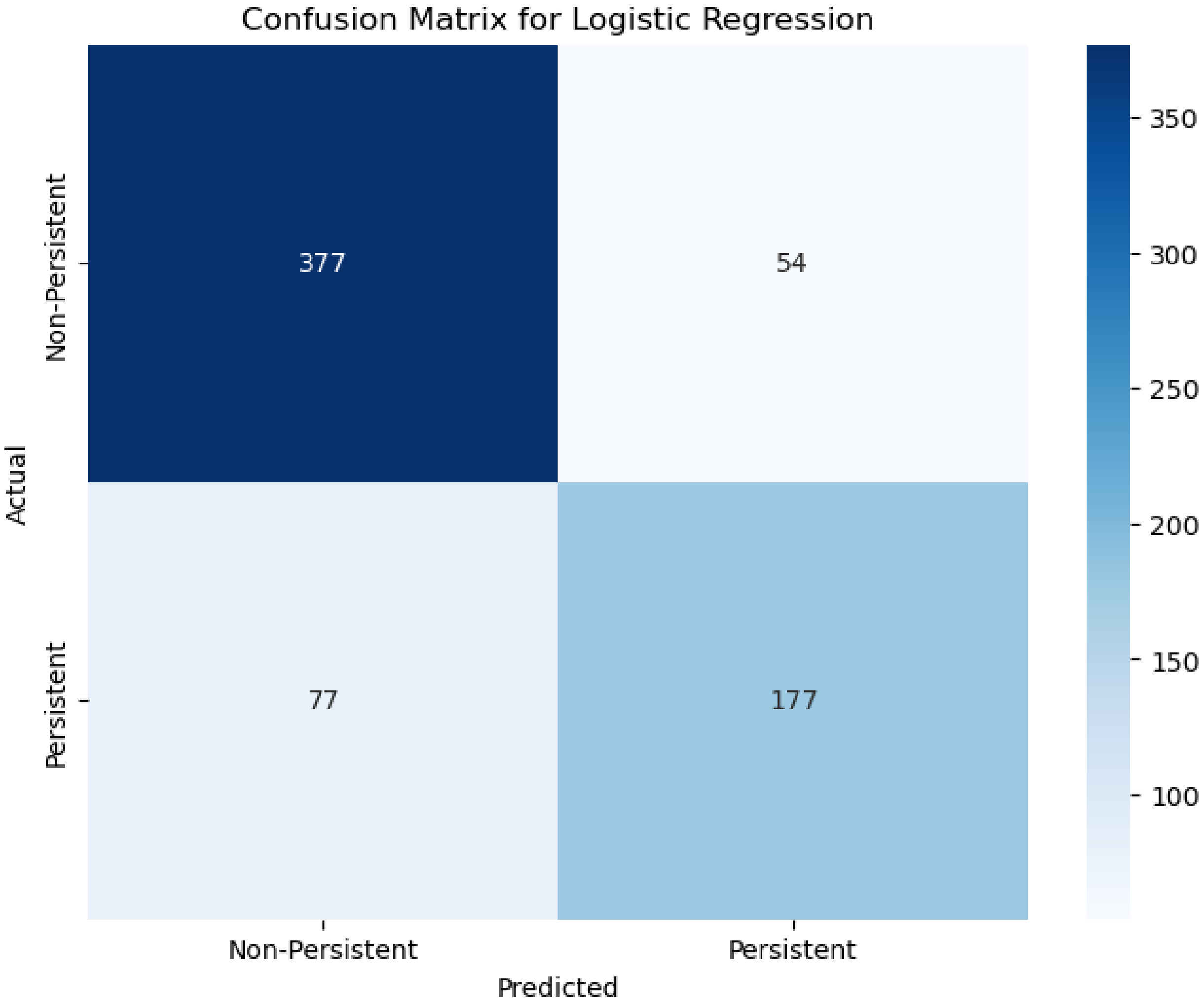
Pharmaceutical companies face a significant challenge in understanding why patients continue or discontinue their prescribed medications. To address this, ABC Pharma has sought the help of an analytics company to automate the identification process of factors influencing drug persistency. The aim is to develop a classification model that predicts whether a patient will persist with a prescribed drug (Persistency_Flag).

Logistic Regression

Accuracy: 0.81

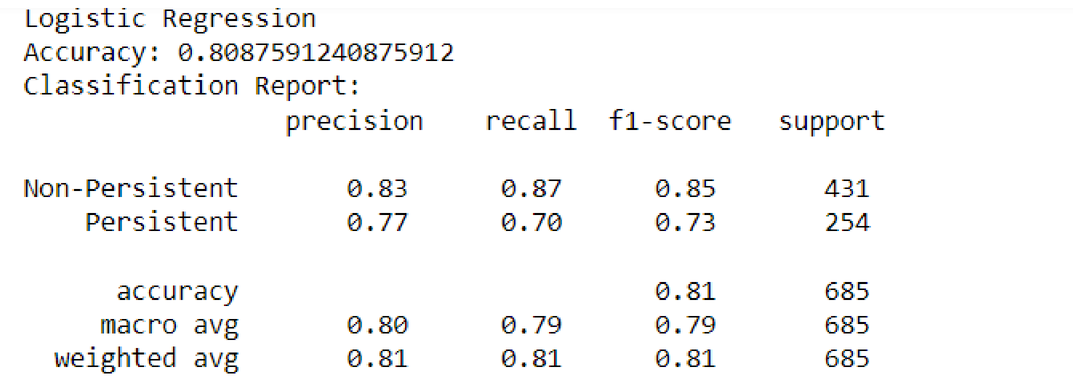
- The accuracy of 81% means that 81% of the predictions made by the model are correct.

Confusion Matrix:



- True Positives (TP): 177 (Persistency correctly identified)
- True Negatives (TN): 377 (Non-persistency correctly identified)
- False Positives (FP): 54 (Non-persistency incorrectly identified as persistency)
- False Negatives (FN): 77 (Persistency incorrectly identified as non-persistency)

Classification Report:



```
Logistic Regression
Accuracy: 0.8087591240875912
Classification Report:
                 precision    recall  f1-score   support

Non-Persistent       0.83      0.87      0.85       431
    Persistent       0.77      0.70      0.73       254

      accuracy                           0.81       685
     macro avg       0.80      0.79      0.79       685
  weighted avg       0.81      0.81      0.81       685
```

- **Precision**: Indicates how many of the identified positive cases (Persistency) were actually positive.
  - Non-Persistent: 83%
  - Persistent: 77%
- **Recall**: Indicates how many actual positive cases (Persistency) were identified by the model.
  - Non-Persistent: 87%
  - Persistent: 70%
- **F1-Score**: Harmonic mean of precision and recall, providing a single metric for model performance.
  - Non-Persistent: 0.85
  - Persistent: 0.73

Interpretation:

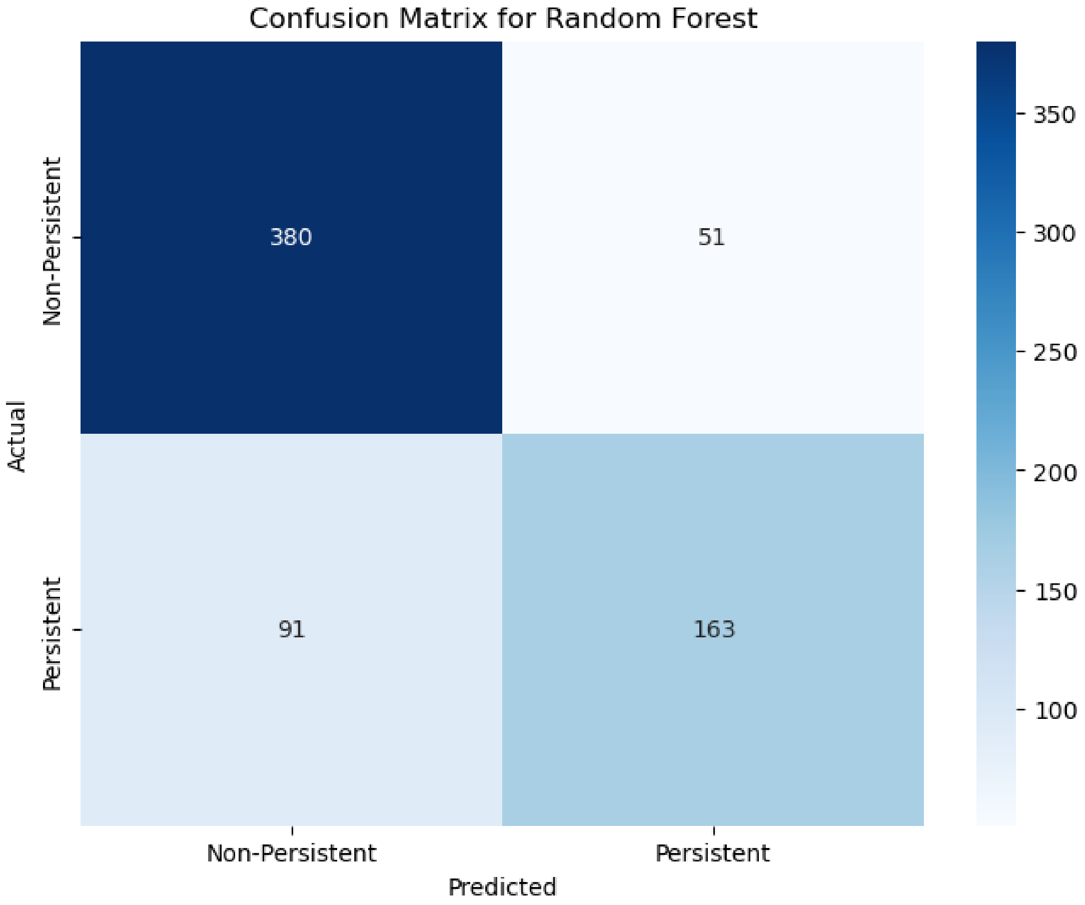
- Logistic Regression is strong in identifying non-persistent cases with high precision and recall.
- The model is slightly less effective at identifying persistent cases, with lower precision and recall.
- This model is highly interpretable, making it a good candidate if transparency is crucial.

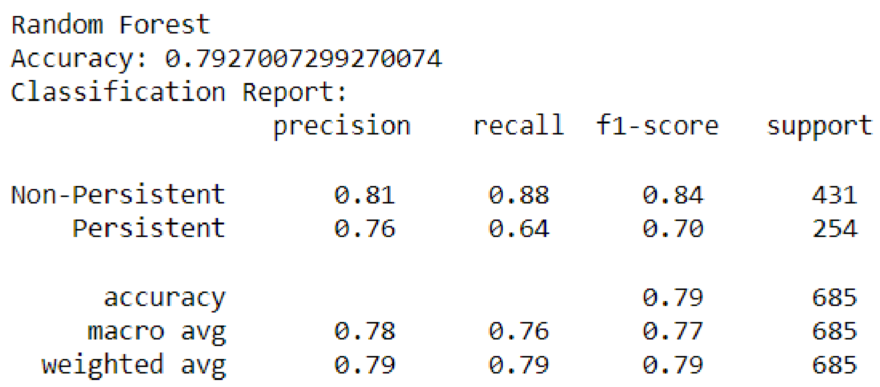
Random Forest

Accuracy: 0.80

- The accuracy of 80% means that 80% of the predictions made by the model are correct.

Confusion Matrix:





```
Random Forest
Accuracy: 0.7927007299270074
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.81      0.88      0.84       431
    Persistent       0.76      0.64      0.70       254

      accuracy                           0.79       685
     macro avg       0.78      0.76      0.77       685
  weighted avg       0.79      0.79      0.79       685
```

- **Precision:**
  - **Non-Persistent: 81%**
  - **Persistent: 77%**
- **Recall:**
  - **Non-Persistent: 89%**
  - **Persistent: 65%**
- **F1-Score:**
  - **Non-Persistent: 0.85**
  - **Persistent: 0.71**

Interpretation:

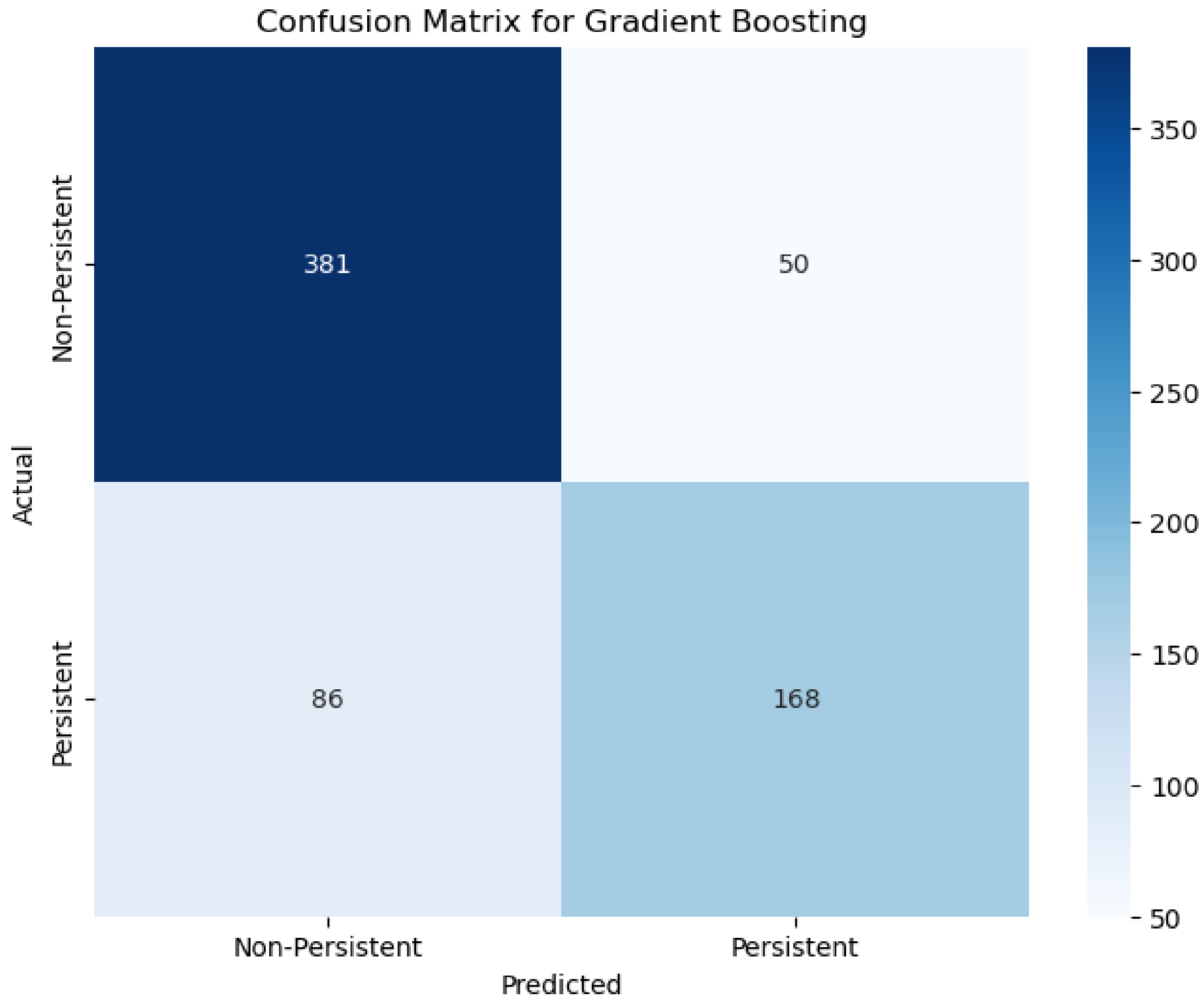
- Random Forest performs well in identifying non-persistent cases with high precision and recall.
- The model is less effective at identifying persistent cases, with lower recall indicating it misses a fair number of actual persistent cases.
- Random Forests are less interpretable but offer insights through feature importance.

Gradient Boosting

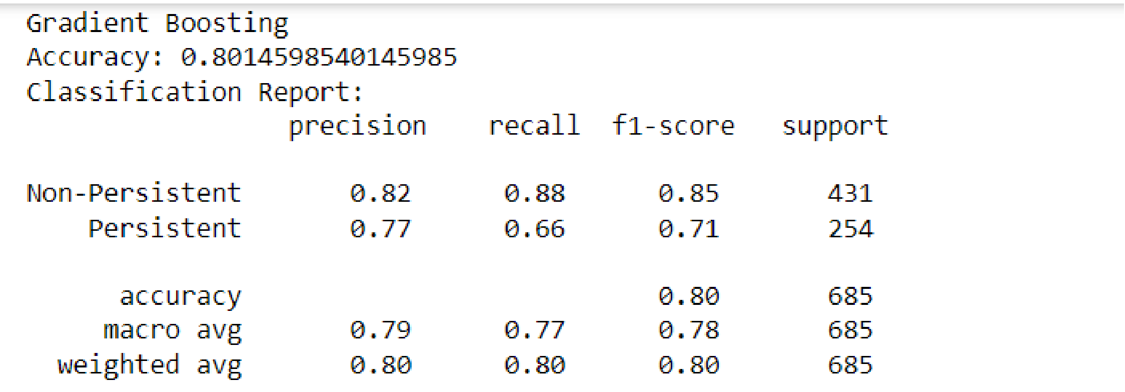
Accuracy: 0.80

- The accuracy of 80% means that 80% of the predictions made by the model are correct.

Confusion Matrix:



Classification Report:



```
Gradient Boosting
Accuracy: 0.8014598540145985
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.82      0.88      0.85       431
    Persistent       0.77      0.66      0.71       254

      accuracy                           0.80       685
     macro avg       0.79      0.77      0.78       685
  weighted avg       0.80      0.80      0.80       685
```

- **Precision:**
  - **Non-Persistent: 81%**
  - **Persistent: 78%**
- **Recall:**
  - **Non-Persistent: 89%**
  - **Persistent: 66%**
- **F1-Score:**
  - **Non-Persistent: 0.85**
  - **Persistent: 0.71**

Interpretation:

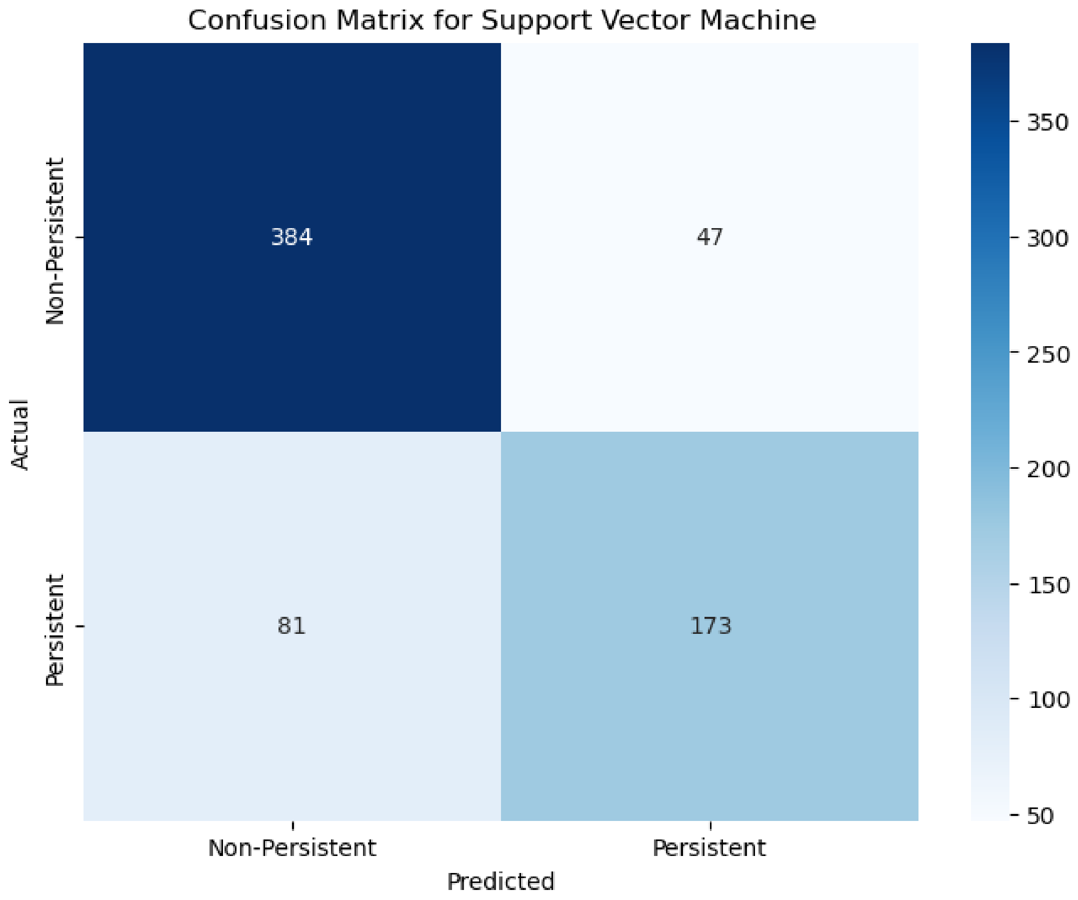
- Gradient Boosting performs similarly to Random Forest in identifying non-persistent cases.
- It slightly improves precision for persistent cases but still has a moderate recall, indicating some misses.
- Gradient Boosting models are generally less interpretable than Random Forests but can be made more interpretable through techniques like SHAP values.

Support Vector Machine (SVM)

Accuracy: 0.81

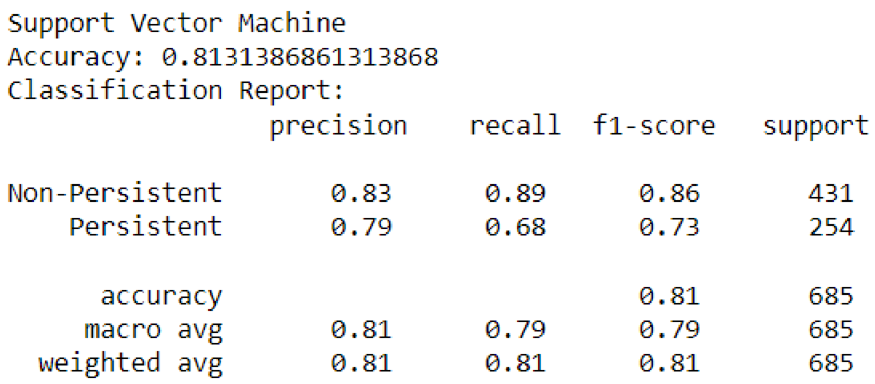
- The accuracy of 81% means that 81% of the predictions made by the model are correct.

Confusion Matrix:



- **TP: 173**
- **TN: 384**
- **FP: 47**
- **FN: 81**

Classification Report:



```
Support Vector Machine
Accuracy: 0.8131386861313868
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.83      0.89      0.86       431
    Persistent       0.79      0.68      0.73       254

      accuracy                           0.81       685
     macro avg       0.81      0.79      0.79       685
  weighted avg       0.81      0.81      0.81       685
```

- **Precision:**
  - **Non-Persistent: 83%**
  - **Persistent: 79%**
- **Recall:**
  - **Non-Persistent: 89%**
  - **Persistent: 68%**
- **F1-Score:**
  - **Non-Persistent: 0.86**
  - **Persistent: 0.73**

Interpretation:

- SVM achieves high performance in identifying non-persistent cases with excellent precision and recall.
- The model shows good performance for persistent cases, with higher precision than Random Forest and Gradient Boosting but similar recall.
- SVMs are often considered less interpretable, which might be a drawback depending on the business requirement for model transparency.

Summary and Recommendations

Summary:

- Logistic Regression: High interpretability, good overall performance with 81% accuracy.
- Random Forest: High performance, slightly lower interpretability, 80% accuracy.
- Gradient Boosting: High performance, similar to Random Forest, 80% accuracy.
- SVM: Best performance for non-persistent cases, similar overall accuracy to Logistic Regression, less interpretable.

Recommendations:

1. Logistic Regression:
   - Best for scenarios requiring high interpretability and reasonable performance.
   - Recommended if transparency in decision-making is crucial.
2. Random Forest:
   - Suitable for capturing complex relationships in the data.
   - Recommended if slightly lower interpretability is acceptable in exchange for capturing more complex patterns.
3. SVM:
   - Offers the highest accuracy, particularly for non-persistent cases.
   - Recommended if the business can tolerate lower interpretability for higher accuracy.
4. Gradient Boosting:
   - Provides a balance similar to Random Forest but with a different approach.
   - Recommended for robustness against overfitting and if interpretability techniques like SHAP are applied.