Final Report: Model Performance Evaluation

Introduction

The objective of this analysis is to evaluate different machine learning models to predict the Persistency_Flag in a healthcare dataset. Four models were considered for this evaluation:

1. Random Forest Classifier
2. Logistic Regression
3. Gradient Boosting Classifier
4. Support Vector Machine (SVM)

Dataset

The dataset contains various features related to healthcare and the target variable Persistency_Flag, which indicates whether a patient is persistent or non-persistent.

Data Preprocessing

1. **Numerical Features:** StandardScaler was used to scale numerical features.
2. **Categorical Features:** OneHotEncoder was used to handle categorical features.
3. A **ColumnTransformer** was used to bundle the preprocessing steps.

Model Training and Evaluation

Each model was trained on the preprocessed dataset using a pipeline to ensure consistent preprocessing. The performance metrics considered were accuracy, precision, recall, and ROC AUC.

Performance Metrics

Metric	Random Forest	Logistic Regression	Gradient Boosting	SVM
Accuracy	0.8161	0.8102	0.8058	0.8044
Precision	0.7936	0.7672	0.7788	0.7632
Recall	0.6811	0.7008	0.6654	0.6850
ROC AUC	0.7883	0.7877	0.7770	0.7799

Analysis

- **Random Forest:** Achieved the highest accuracy (81.61%) and precision (79.36%). It also had a competitive ROC AUC score (78.83%), indicating a good balance between sensitivity and specificity.
- **Logistic Regression:** Showed good performance with an accuracy of 81.02% and the highest recall (70.08%). This indicates it was slightly better at identifying persistent cases correctly.
- **Gradient Boosting:** Had a balanced performance with accuracy (80.58%) and precision (77.88%). However, its recall (66.54%) was slightly lower compared to Logistic Regression and Random Forest.
- **SVM:** Achieved the lowest accuracy (80.44%) and precision (76.32%) among the four models. However, it had a relatively good ROC AUC score (77.99%).

Conclusion

- The **Random Forest Classifier** emerged as the best performing model with the highest accuracy and precision, making it a robust choice for this dataset.
- **Logistic Regression** also performed well, especially in terms of recall, which is important for identifying persistent patients correctly.
- **Gradient Boosting** provided a balanced performance and could be considered if further tuning and optimization are applied.
- **SVM** had the lowest overall performance but still provided valuable insights and could be useful in specific scenarios.

Next Steps

- **Hyperparameter Tuning:** Further tuning of model parameters using grid search or random search to potentially improve performance.
- **Cross-Validation:** Implement cross-validation to ensure the models' robustness and to avoid overfitting.
- **Feature Engineering:** Explore additional feature engineering techniques to enhance the predictive power of the models.
- **Ensemble Methods:** Consider combining the strengths of multiple models through ensemble methods like stacking or voting classifiers to achieve better performance.

This concludes the performance evaluation report for the given healthcare dataset. The Random Forest Classifier is recommended as the primary model for predicting the Persistency_Flag.