REPUcs Assigment 2024

Overview

This repository contains a Python script that builds, trains, and evaluates a machine learning model using the XGBoost classifier. The model is designed to predict the target column based on the other features in a dataset. It performs hyperparameter tuning, model evaluation, and saves the trained model for future use.

Prerequisites

To run this script, you need Python version 3.11.9 installed. Additionally, the following Python libraries are required:

- scikit-learn
- polars
- yellowbrick
- xgboost
- hyperopt
- pandas
- pyarrow
- shap
- numpy==2.0.2

Installation

Install the required libraries using pip:

```
pip install scikit-learn polars yellowbrick xgboost hyperopt pandas pyarrow
shap numpy==2.0.2
```

Model Evaluation Report

1. Evaluating Overfitting with Learning Curve

To assess the model for overfitting, we use a learning curve, which involves creating subsets of the data for training and observing how the model's evaluation metric changes. In this case, we use the Area Under the Curve (AUC) metric, which is suitable for our label imbalance problem.

As more training data is added, the model's cross-validation score improves, and the training score decreases slightly. This behavior indicates that the model is generalizing well and is not overfitting. The training score is not consistently at 1, and the cross-validation score is improving with more data, showing that the model is learning effectively from the increased data.

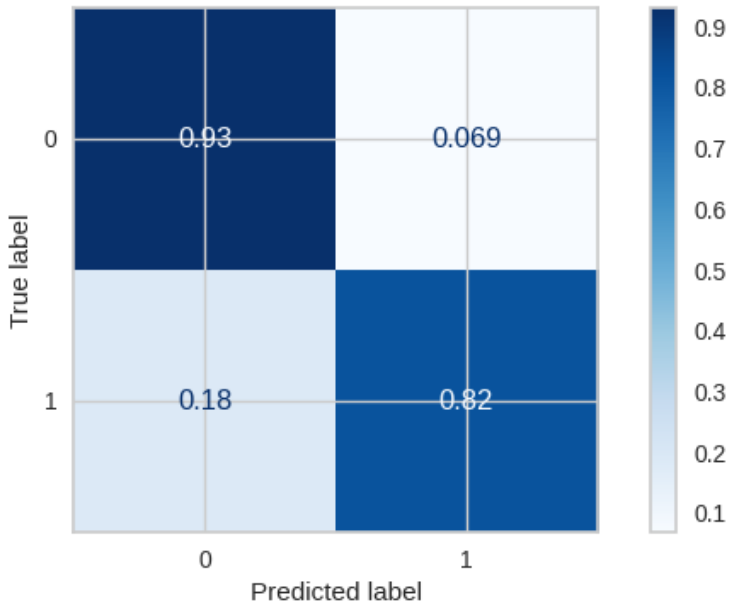
The model is neither underfitting nor overfitting. With more data, we could expect further improvement, indicating the model is robust and has the capacity to learn more.

2. Confusion Matrix for Accuracy Metrics

The confusion matrix provides insights into the performance of the model by comparing true labels with predicted labels.

- True Positive Rate (No heart disease): 0.93, indicating that the model correctly predicted "no heart disease" 93% of the time.
- True Negative Rate (Heart disease): 0.82, indicating that the model correctly predicted "heart disease" 82% of the time.

These metrics suggest that the model is more accurate at identifying cases without heart disease but still performs well in detecting heart disease cases.

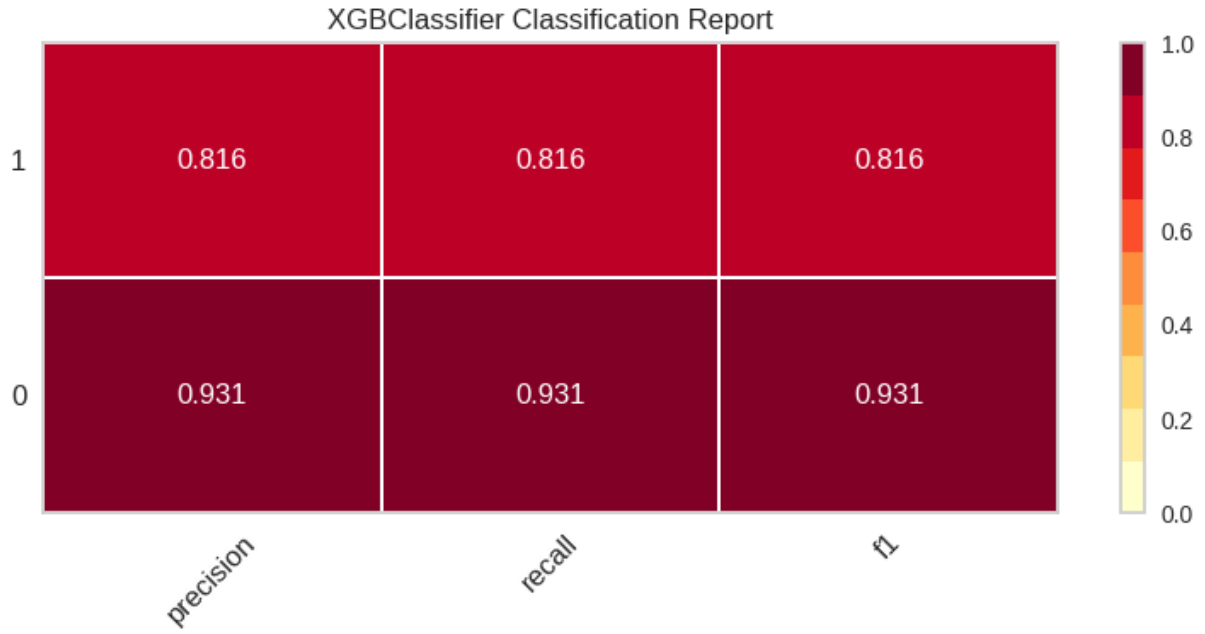


3. F1 Score and Classification Report

The classification report provides detailed metrics, including precision, recall, and F1 score.

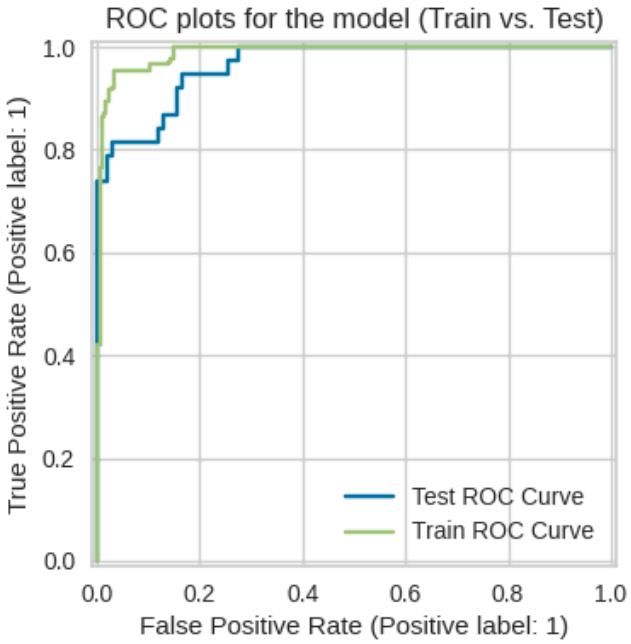
- Precision for "Heart disease": 0.816
- Recall for "Heart disease": 0.816
- F1 Score for "Heart disease": 0.816

These high scores indicate that the model has a good balance of precision and recall, minimizing both false positives and false negatives.



4. ROC Curve

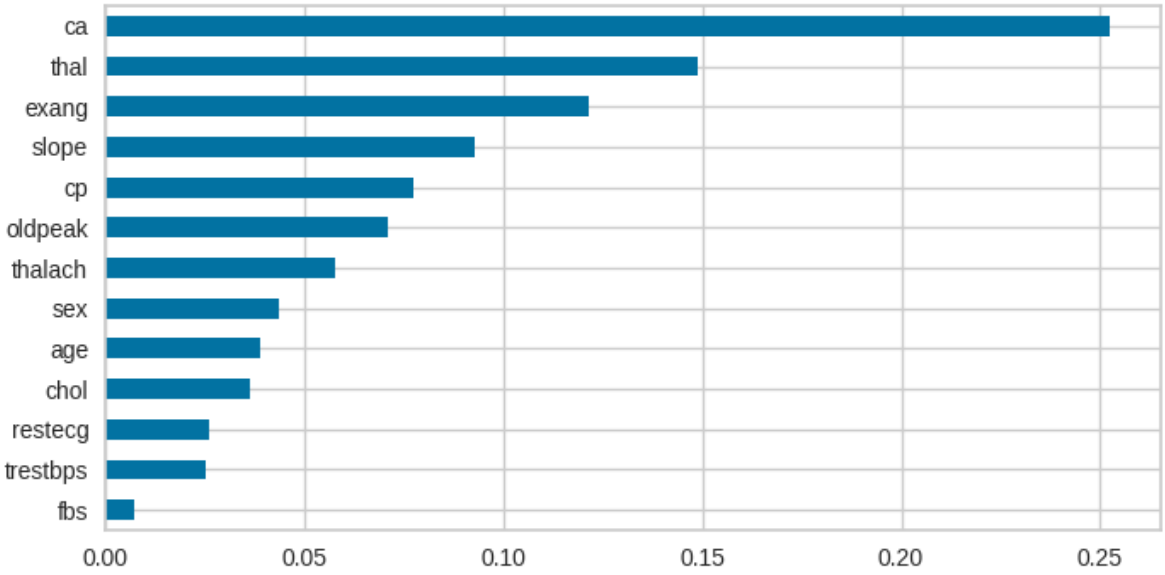
The ROC curve demonstrates the performance of the model in distinguishing between classes. Both the training and testing ROC curves are close to 0.9, indicating that the model has high discriminative power and maintains similar performance on both training and test sets.



5. Feature Importance

Using XGBoost's built-in feature importance, we identify the features that contribute the most to the model's predictions. Features such as `ca` (number of major vessels colored by fluoroscopy), `thalach` (maximum heart rate achieved), `exang` (exercise-induced angina), and `slope` (slope of the peak exercise ST segment) significantly impact the model's output.

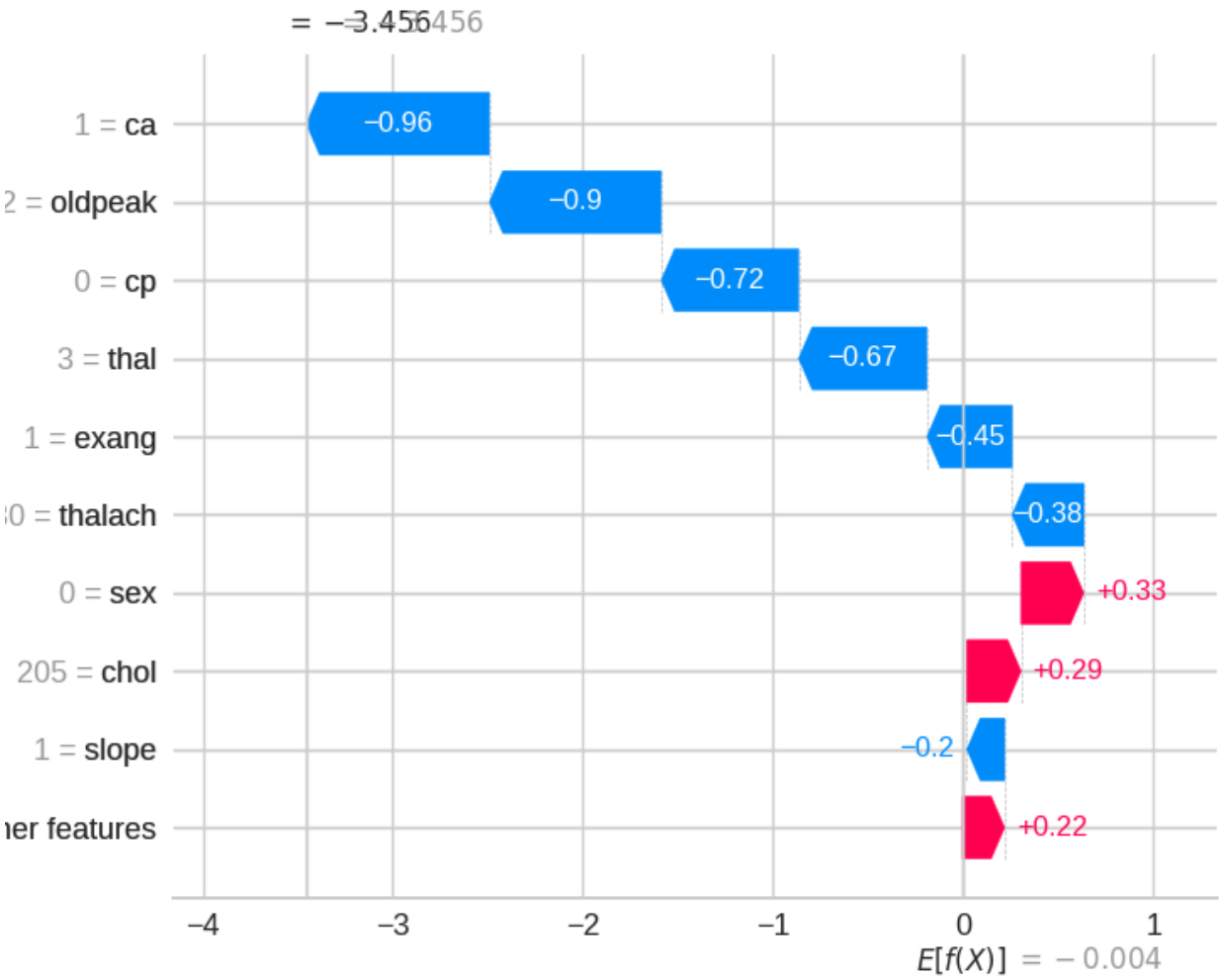
These features have high normalized gain values, indicating their importance in the model's decision-making process.



6. SHAP Values for Model Interpretation

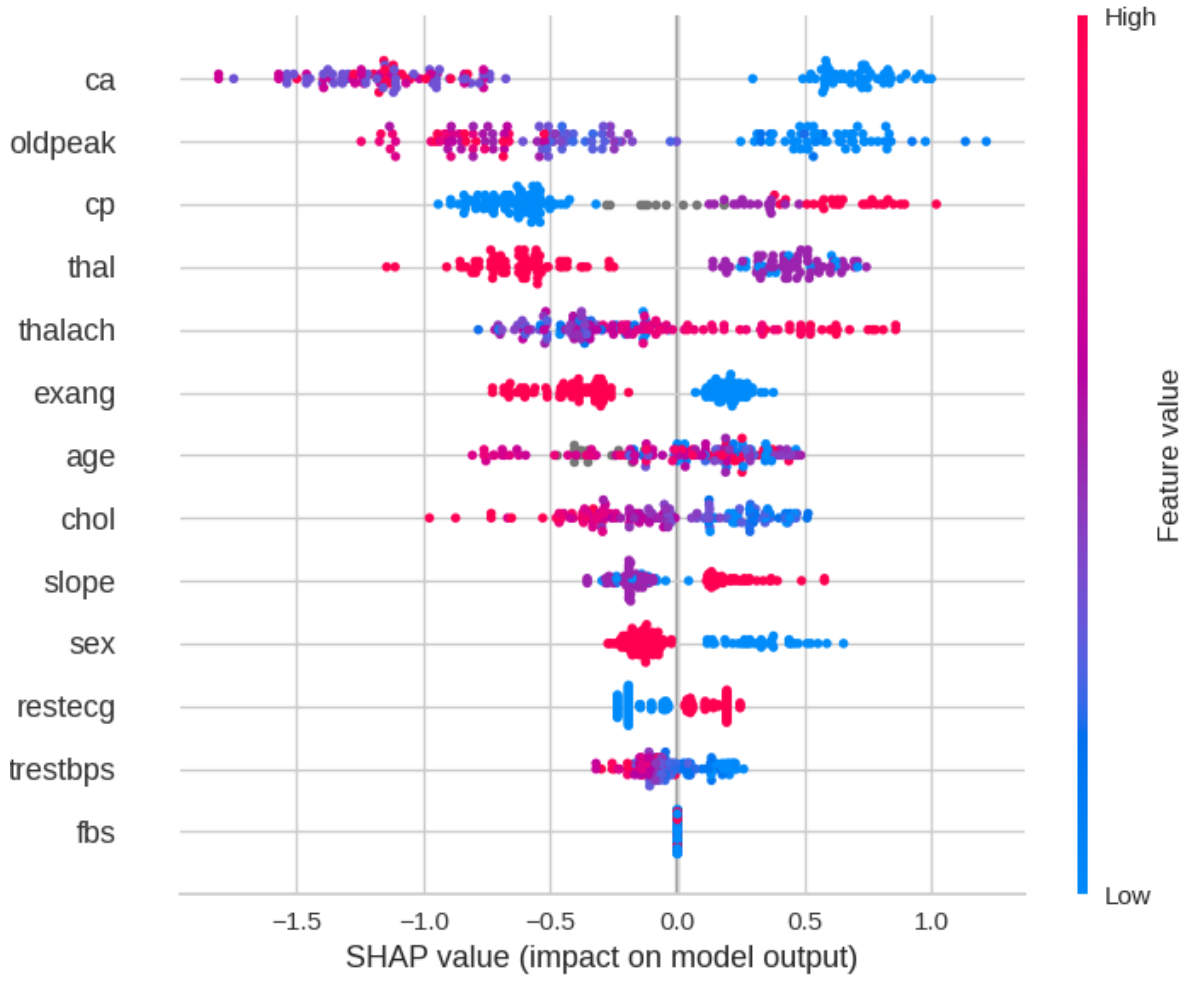
Waterfall Plot (Single Prediction Interpretation)

The SHAP waterfall plot explains how individual features contribute to a specific prediction. For example, the feature `ca` had a significant positive impact, suggesting a higher likelihood of heart disease. Other features, such as `cp` (chest pain type), also influenced the prediction, showing the model's reasoning behind its decision.



Beeswarm Plot (Global Feature Impact)

The SHAP beeswarm plot shows the impact of each feature across all predictions. The feature ca consistently showed a strong influence, where higher values of ca increased the probability of having heart disease. Conversely, lower values of ca were associated with a lower risk of heart disease, demonstrating how the model uses these features to make predictions.



Feature Descriptions

1. **age**: Age of the individual
2. **sex**: Sex of the individual (Male/Female)
3. **cp**: Chest pain type (0-3, where 0 indicates typical angina, 1 indicates atypical angina, 2 indicates non-anginal pain, and 3 indicates asymptomatic)
4. **trestbps**: Resting blood pressure (in mm Hg on admission to the hospital)
5. **chol**: Serum cholesterol in mg/dl
6. **fbs**: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. **restecg**: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. **thalach**: Maximum heart rate achieved
9. **exang**: Exercise induced angina (1 = yes; 0 = no)
10. **oldpeak**: ST depression induced by exercise relative to rest
11. **slope**: The slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)
12. **ca**: Number of major vessels (0-3) colored by fluoroscopy
13. **thal**: Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
14. **target**: Presence of heart disease (1 = presence; 0 = absence)

Evaluation Summary

Evaluation Metrics

- **Accuracy**: 0.9000
- **Precision**: 0.8158
- **Recall**: 0.8158
- **F1 Score**: 0.8158
- **AUC**: 0.8736

Confusion Matrix

	Predicted: 0	**Predicted: 1**
Actual: 0	95	7
Actual: 1	7	31

> *Analysis by:* **Diego Taquiri Díaz**