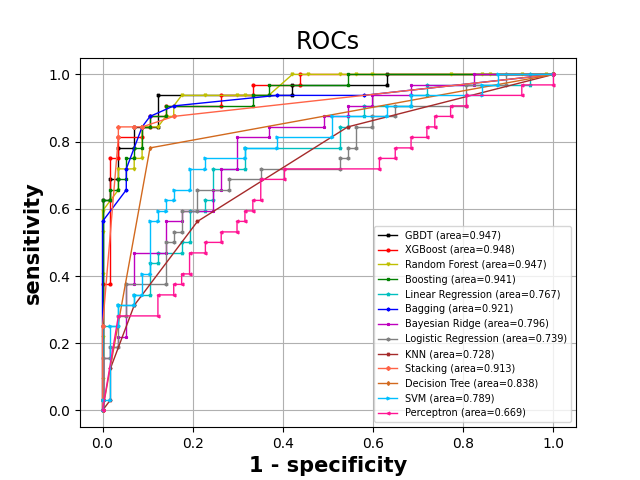
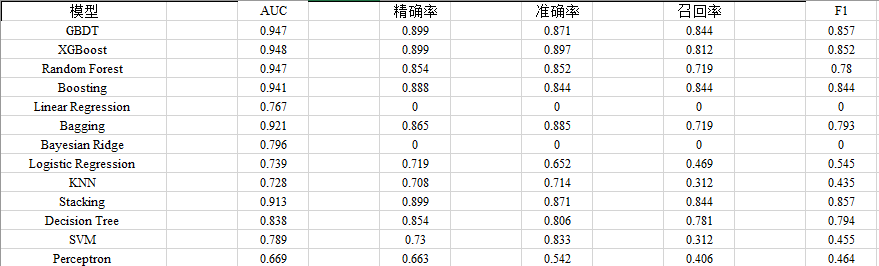
**数据**：去除数据无用列（入院时间、id），将原始数据打乱后按照80%：20%的比例划分训练集和测试集，使用白化操作实现对测试集和数据集中的数据的规格化处理。

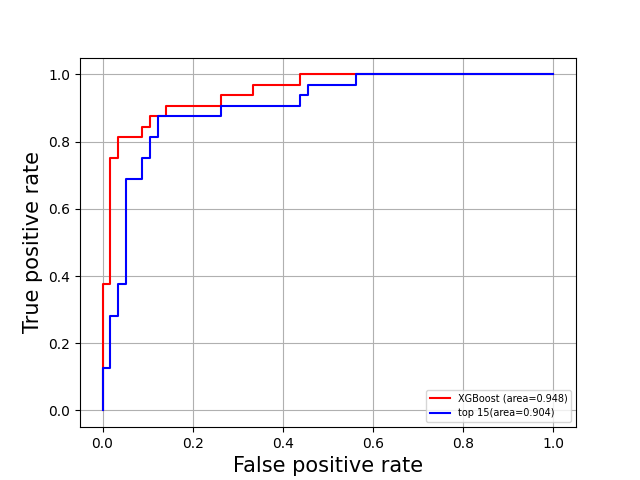
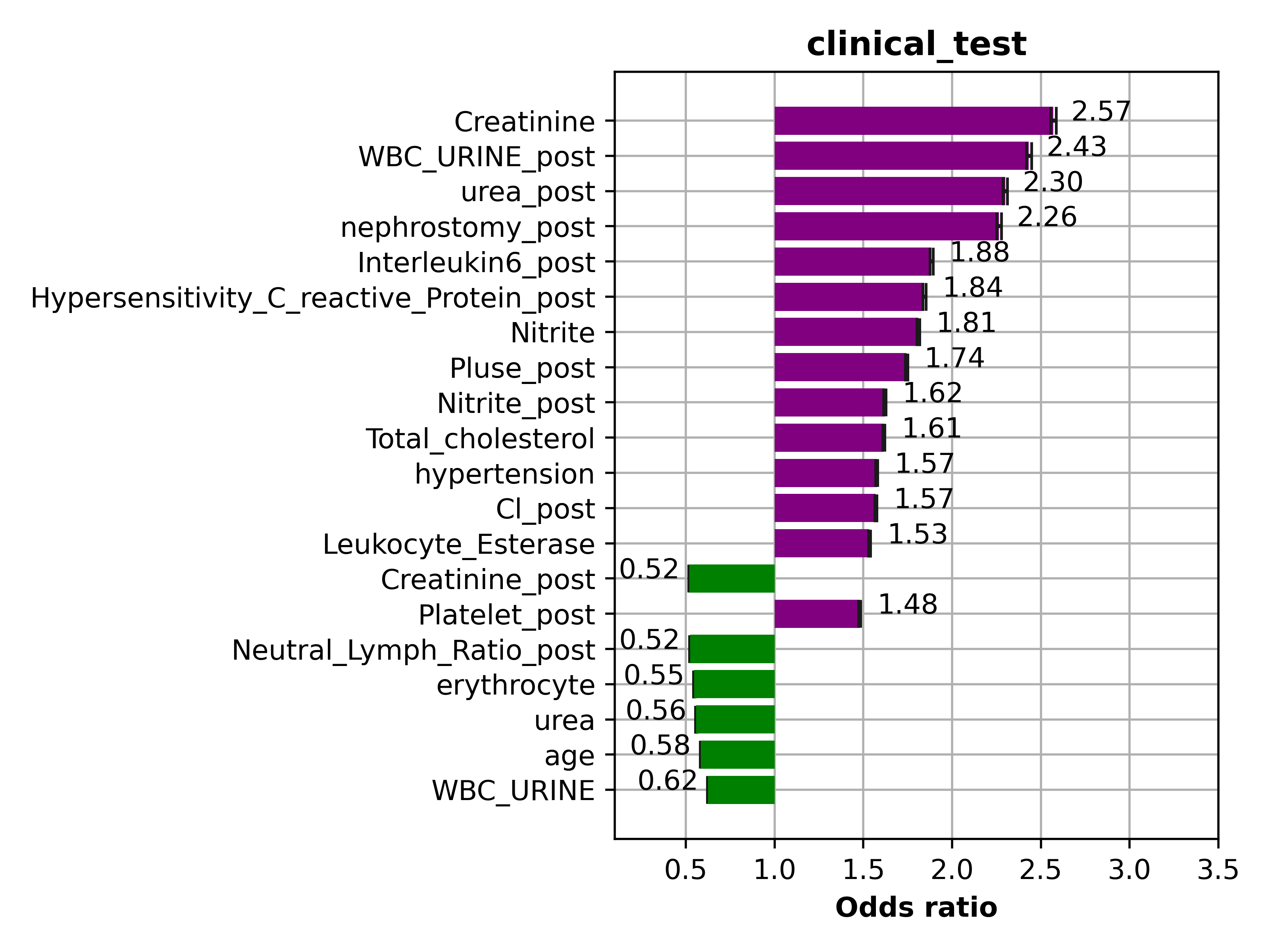
模型训练和分析：

1. 使用sklearn工具库中的GBDT, XGBoost, RF, Adaboost, Linear Regression, Bagging, Bayes Ridge, Logistic Regression, KNN, Stacking, DT, SVM, Perceptron等13种模型在训练集中训练，分别绘制模型在测试集上的性能表现，计算相应的AUC，选择在测试集上性能最佳（AUC最大）的模型（GBDT）作为最终机器学习模型。





1. 由于线性回归和贝叶斯邻回归，只能计算样本的回归概率（即某个样本为正样本的概率，而无法对样本进行分类，因此无法获得其准确率、精准率等数据信息）。
2. 关于为何性能是XGBoost最佳，但是使用GBDT的原因是最初在训练过程中，确有出现GBDT最佳的情况，不过再重新跑了多次实验确实是XGBoost模型性能最佳，后续模型图形均以XGBoost作为基础模型。
3. 补充来说，XGBoost模型和GBDT模型本质是一种模型，XGBoost在GBDT模型的基础上做了略微改善。从两种模型两次训练的结果图来看，其实对于性能各方面图形改变并不显著。
4. sklearn工具包，使用逻辑回归模型进行各特征重要性分析，使用GBDT模型训练特征选择前后的数据，并使用matplotlib绘制相应ROC曲线，计算相应AUC。



1. 为验证模型可信度，我们绘制了测试集上的decision curve analysis分析图表。

