

# 特征选择与模型训练流程说明（基于 Minimal Depth 策略）


本项目旨在建立一个鲁棒的分类器模型，在“高维低样本”的背景下，**有效挑选出具有判别力的特征**，并通过逐步训练过程获得性能最优的模型结构。整个流程包括三个核心部分：

## 一、特征选择机制设计

### 1. Lasso 初筛（稀疏建模）

**\*\*目的：\*\***在原始的高维特征空间中（可能是成百上千个），快速删除与分类目标无关的特征，缩小搜索空间。

- Lasso（L1正则化逻辑回归）会压缩某些特征系数为0；
- 非零系数对应的是可能对分类任务有意义的特征；
- Lasso 具有稳定性强、可解释性好的特点，适合作为第一步降维手段。

 **\*\*好处：\*\***大幅压缩特征维度，减少后续模型复杂度。

### 2. Minimal Depth（最小深度）排序

**\*\*目的：\*\***使用随机森林分析“每个特征在树中首次分裂的位置”，也就是它被模型首次使用的深度。

- 如果一个特征经常在靠近根节点分裂 → 表示它能早期将样本划分开，重要性较高；
- 多棵树平均后，最小深度越小，表示特征越重要；
- 更稳健于特征之间的相关性。

 **\*\*作用：\*\***在 Lasso 初筛后进一步对特征重新排序，获取精细的“实际重要性”排序。

### 3. 逐步递减特征数（逐轮训练）

**目的：**从最小深度排名中，逐步减少特征数量，观察模型性能的变化。

每一轮：

- 重新选择 top-K 个特征；
- 重新训练一个 RandomForest 模型；
- 使用调参工具（RandomizedSearchCV）找到最优超参数；
- 用最优模型在测试集上预测，记录 AUC 和 Recall。

⚠️ **注意：**每一轮都会重新训练随机森林 → 重新排序 → 再决定下一轮的特征。

✅ **优势：**这是一种模拟“特征重要性稳定性”的方式，寻找最优的特征组合规模。

## 🎯 二、模型评估指标说明

### 1. ROC AUC (Area Under Curve)

**定义：**ROC 曲线下的面积，用于衡量模型区分正负样本的能力。

- AUC 越接近 1，模型越优秀；
- AUC = 0.5 表示模型跟随机分类差不多；
- 对于**类别不平衡问题**非常适合，衡量模型整体判别力。

✅ **本项目核心优化指标。**

### 2. Recall (召回率)

**定义：**真正例 / (真正例 + 假负例)，也就是预测出的正例中，有多少是真正的正例。

- 强调模型“不要漏掉”正例；
- 对于某些任务如疾病检测、欺诈识别尤为重要；

- 和 AUC 搭配使用可以监控模型稳定性。

✓ 作为辅助指标，判断模型在不同特征数量下的实用性。

### 🔄 三、训练过程概述

步骤	内容	说明
①	加载数据并标准化	分 cohort 独立归一化，防止信息泄露
②	Lasso 特征初筛	稀疏建模，去除冗余无关特征
③	初始化候选特征集	Lasso 输出结果作为初始 pool
④	进入循环，从最多特征到最少特征（不少于5个）	每轮减一维，训练 + 排序 + 筛选
⑤	每轮使用 RandomForest 得到当前特征集的最小深度排序	每轮排序都更新，不是静态排序
⑥	基于 Top-K 特征调参训练模型	交叉验证 + AUC评分
⑦	记录 AUC、Recall、重要性、SHAP图等	为后续分析保留完整路径
⑧	输出性能变化曲线	找出最佳特征数量对应的模型
⑨	输出最终模型参数与特征组合	作为推荐最终方案

### 📝 附：核心参数说明

参数名称	建议默认值	含义与备注
<code>max_iter</code> (Lasso)	10000	防止迭代未收敛
<code>cv</code> (Lasso/RF)	5 或 10	数据量小建议10，大可调为5加速
<code>n_iter</code> (RandomSearch)	30	每轮超参搜索次数，越大越精细
<code>k</code> 起点	<code>len(features)</code>	每轮减少1个特征直到5个
<code>shap_values</code>	维度大时可选取 sample	否则时间开销很大

### ✓ 总结

本代码实现的是一种“结构化、多轮、自动化的特征评估与模型选择流程”，通过：

- Lasso 控制规模、
- Random Forest + Minimal Depth 评估强弱、
- 多轮比较寻找最优模型结构，

最终得到一个在测试集上**AUC 最优、可解释性强、特征维度控制合理**的模型，适用于后续部署、报告撰写、以及科研成果产出。