5. 建模与调参

5.1 常用模型

在金融风控领域常用模型包括逻辑回归，树以及集成模型。它们各具优缺点：

5.1.1 逻辑回归

优点：

- 训练速度快，计算量仅和特征数目相关；
- 模型可解释性好，从权重可看出特征对结果的影响；
- 适合 binary classification；
- 占用内存小，只需储存各个维度特征值。

缺点：

- 需要提前进行数据处理；
- 无法解决非线性问题；
- 对多重共线性敏感，难以处理数据不平衡问题；
- 准确率低，难以拟合数据真实分布。

5.1.2 决策树

优点：

- 模型直观，方便可视化；
- 数据无需预处理，不需要归一化，不需要处理缺失值；
- 离散值和连续值都能处理；

缺点：

- 易过拟合，泛化能力不强（可通过剪枝改善）；
- 使用贪心算法，易得到局部最优解。

5.1.3 Ensemble Method

集成学习中，会训练多个模型（弱学习器）解决相同问题，并将它们结合起来以获得更好结果。其中的重要假设为：当弱模型被正确组合时，我们可以得到更精确和/或更具鲁棒性的模型。大多数情况中，这些弱学习器本身性能并不好（high bias / high variance）。集成方法的思想是通过将这些弱学习器的 bias / variance 结合起来，从而创建一个“强学习器”以获得更好的性能。

集成方法主要包括 Bagging 和 Boosting，常见基于 Bagging 思想的集成模型有：随机森林；基于 Boosting 思想的集成模型有：Adaboost、GBDT、XgBoost 以及 LightGBM 等。Baggin 和 Boosting 的区别总结如下：

- 样本选择上：Bagging 方法的训练集是从原数据集有放回的选取，所以从原始集中选出的各轮训练集之间是相互独立的；Boosting 方法需要每轮训练集不变，而是每个样本在分类器中的权重发生变化。而权值是根据上一轮的分类结果进行调整的；
- 样例权重上：Bagging 方法使用均匀取样，每个样本权重相等；Boosting 方法根据错误率不断调整样本权重，错误率越大权重越大；
- 预测函数上：Bagging 方法中所有预测函数的权重相等；Boosting 方法中每个弱分类器都有相应权重，对于分类误差小的分类器有更大的权重；
- 并行计算上：Bagging 方法中各个预测函数可以并行生成；Boosting 方法各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。

5.2 数据集划分

训练集上的误差称之为训练误差或经验误差；测试集上的误差称之为测试误差。对于数据集的划分通常需要满足两个条件：

- Training 和 testing 分布要与样本真实分布一致，从样本真实分布中独立同分布采样而得；
- Training 和 testing 互斥。

数据划分方法有：留出法、交叉验证法以及自助法：

- 留出法：直接将数据集划分为两个互斥的集合。需要注意的是在划分的时候尽可能保证数据分布的一致性，常采用分层采样的方式。
- 交叉验证法：k 折交叉验证通常将数据集分为 k 份，k-1 份作为 training，剩余一份为 testing，最终返回 k 个测试结果的均值（k=1 时为留一法）。
- 自助法：每次从数据集中去一个样本作为训练集中的元素，然后放回，重复该操作 m 次，把没有出现过的样本作为测试集。这样的采样方式会使得数据集中约有 36.8%的数据没有在训练集中出现过。留出法和交叉验证法都是分层采样，自助法是有放回的重复采样。

当数据量充足时，常采用留出法和 k 折交叉验证法；当数据量小且难有效划分时，使用自助法；当数据量小且可以有效划分时，使用留一法。

5.3 LightGBM Model

GBDT（Gradient Boosting Decision Tree）常用于多分类、点击率预测、搜索排序等任务，据统

计 Kaggle 上比赛有一半以上的冠军方案都是基于 GBDT 的。LightGBM 是实现 GBDT 算法的框架，支持高效率的并行训练，且具有更快的训练速度，更低的内存消耗，更好的准确率，支持分布式可以快速处理海量数据等优点。

常见机器学习算法（NN 等）都可以使用 mini-batch 方式训练，不受内存限制。而 GBDT 在每一次迭代的时候都需要遍历整个训练数据多次。LightGBM 提出的主要原因就是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于工业实践。

LightGBM 具有以下优缺点：

优点：

- 采用直方图算法将遍历样本转变为遍历直方图，极大降低了时间复杂度；
- 训练过程中采用单边梯度算法过滤掉梯度小的样本，减少计算量；
- 采用了基于 Leaf-wise 算法的增长策略构建树，减少了不必要的计算量；
- 采用优化后的特征并行、数据并行方法加速计算，当数据量非常大的时候还可以采用投票并行。
- 对缓存也进行了优化，增加了缓存命中率；
- XGBoost 使用与排序后需要记录特征值及其对应样本的统计值的索引，LightGBM 使用了直方图算法将特征值转变为 bin 值，且不需要记录特征到样本的索引，将空间复杂度从 O(2#data)降到 O(bin)；
- 在训练过程中采用互斥特征捆绑算法减少了特征数量，降低了内存消耗。

缺点：

- 可能会训练出过深的决策树，产生过拟合。因此 LightGBM 在 Leaf-wise 上增加了最大深度限制，保证高效的同时防止过拟合；
- Boosting 族是迭代算法，每次迭代都根据上一次迭代的预测结果对样本进行权重调整，随着迭代不断增加，误差会越来越小，bias 降低。LightGBM 是基于偏差的算法，所以会对噪点敏感；
- 寻找最优解时，依据的是最优切分变量，没有将最优解是全部特征的综合考虑进去。

```
[LightGBM] [Warning] Unknown parameter: silent
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[656]   valid_0's auc: 0.730304
```

图 5.1 LightGBM 训练及验证结果

未调参前lightgbm单模型在验证集上的AUC：0.730303735603381

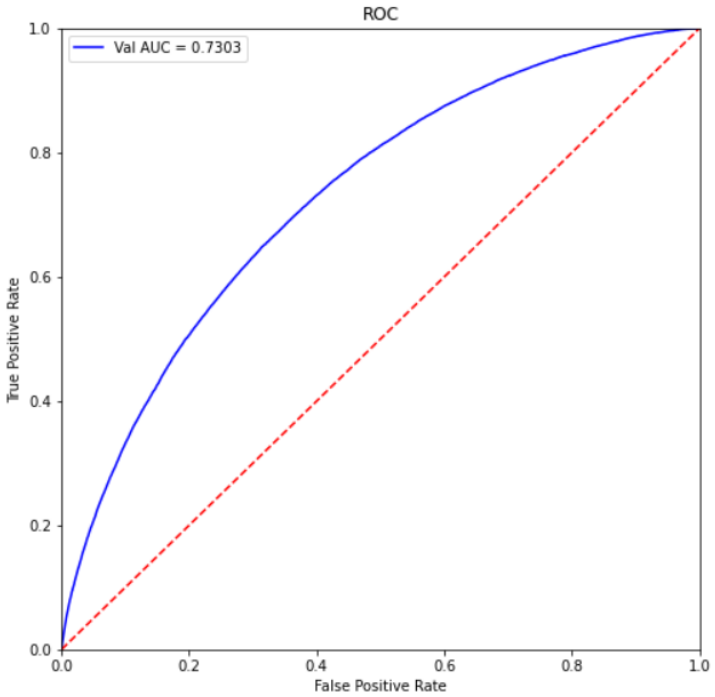

图 5.2 LightGBM 单模型 ROC、AUC 结果

```
********************************** 1 **********************************
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[429]   valid_0's auc: 0.730216
[0.7302164635050385]
********************************** 2 **********************************
[LightGBM] [Warning] Unknown parameter: silent
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[528]   valid_0's auc: 0.726281
[0.7302164635050385, 0.7262811069770285]
********************************** 3 **********************************
[LightGBM] [Warning] Unknown parameter: silent
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[419]   valid_0's auc: 0.731149
[0.7302164635050385, 0.7262811069770285, 0.7311489916644855]
********************************** 4 **********************************
[LightGBM] [Warning] Unknown parameter: silent
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[547]   valid_0's auc: 0.729642
[0.7302164635050385, 0.7262811069770285, 0.7311489916644855, 0.7296419000454826]
********************************** 5 **********************************
[LightGBM] [Warning] Unknown parameter: silent
Training until validation scores don't improve for 200 rounds
Early stopping, best iteration is:
[388]   valid_0's auc: 0.728283
[0.7302164635050385, 0.7262811069770285, 0.7311489916644855, 0.7296419000454826,
0.7282826373391076]
lgb_scotrainre_list:[0.7302164635050385, 0.7262811069770285, 0.7311489916644855,
0.7296419000454826, 0.7282826373391076]
lgb_score_mean:0.7291142199062286
lgb_score_std:0.001694210248122758
```

图 5.3 5 折 LightGBM 模型训练及验证结果

成功构建 LightGBM 后，使用贝叶斯调参得到最优参数：

{'target': 0.7287715528474514,
 'params': {'bagging_fraction': 0.951149142190159,
  'bagging_freq': 24.662212438801678,
  'feature_fraction': 0.8017744633932344,
  'max_depth': 18.771725892792887,
  'min_child_weight': 7.219417186363603,
  'min_data_in_leaf': 75.76252265140555,
  'min_split_gain': 0.4412052082831305,
  'num_leaves': 20.862098032929193,
  'reg_alpha': 8.049345149297617,
  'reg_lambda': 2.923583101548590 2}}

图 5.4 贝叶斯调参最优参数结果

带入最优参数再次训练，得到最优迭代次数为 13176，最终模型 AUC 为 0.732。将模型带入到验证集：

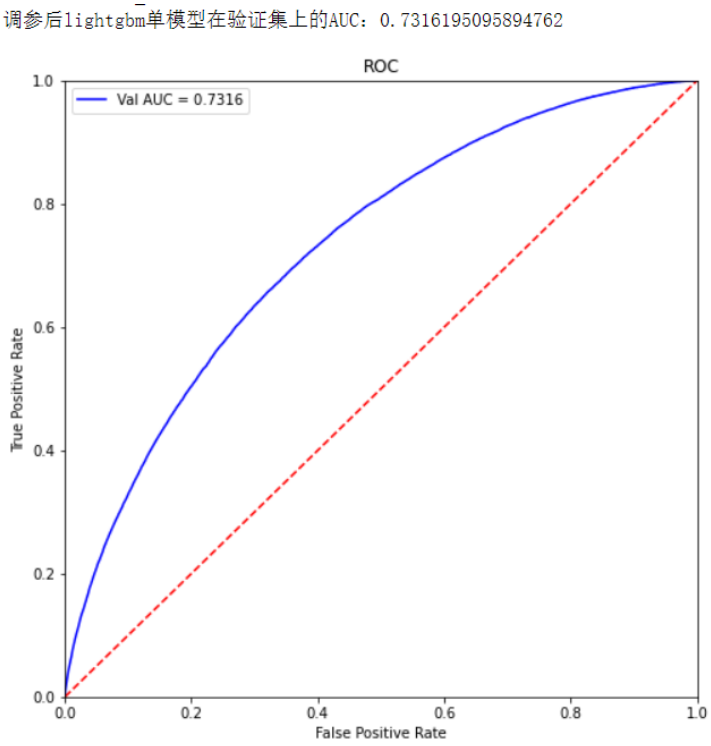

图 5.5 调参后模型于验证集上结果