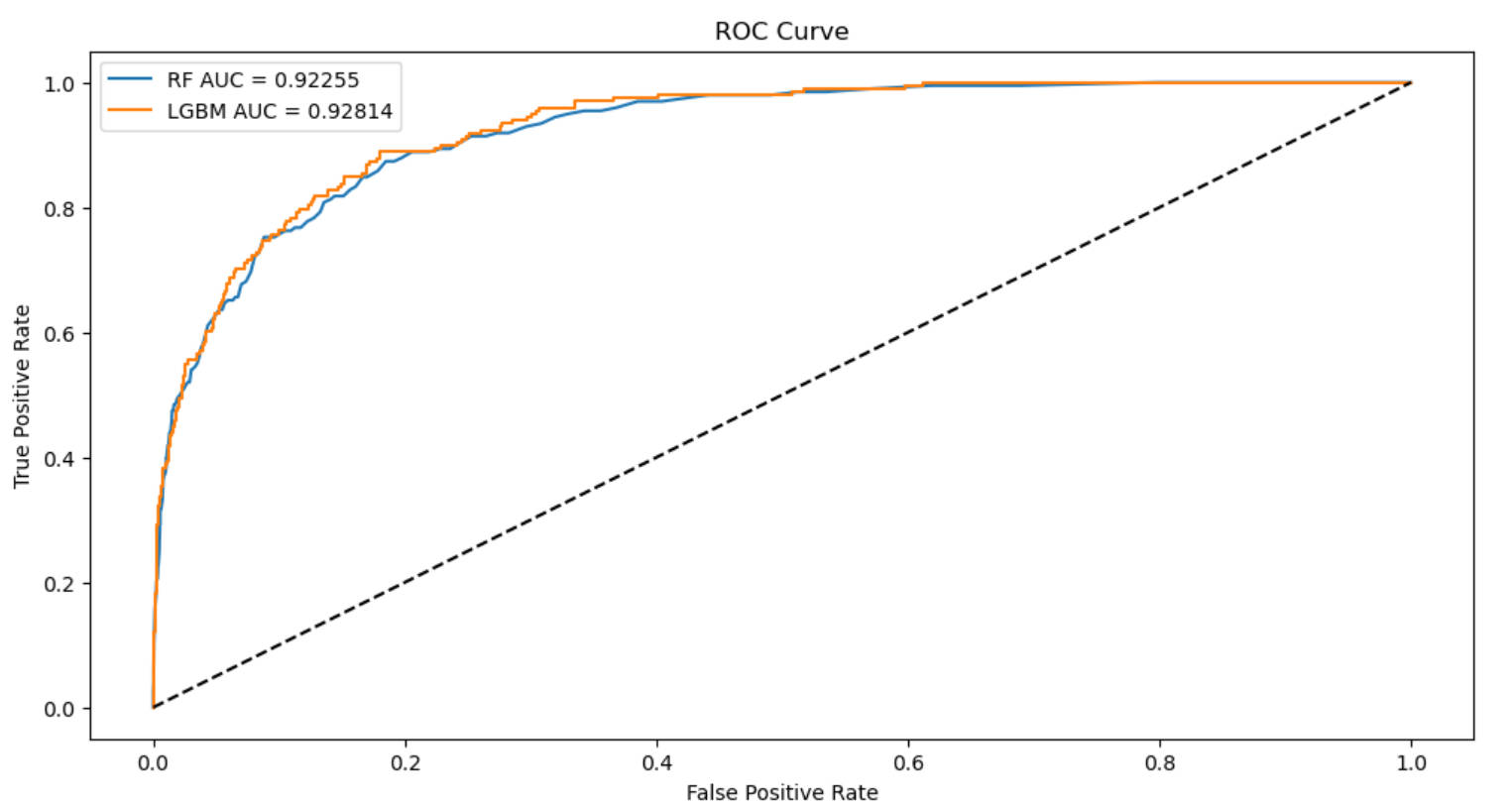
测试报告与案例

# 1. RF与LGBM模型性能

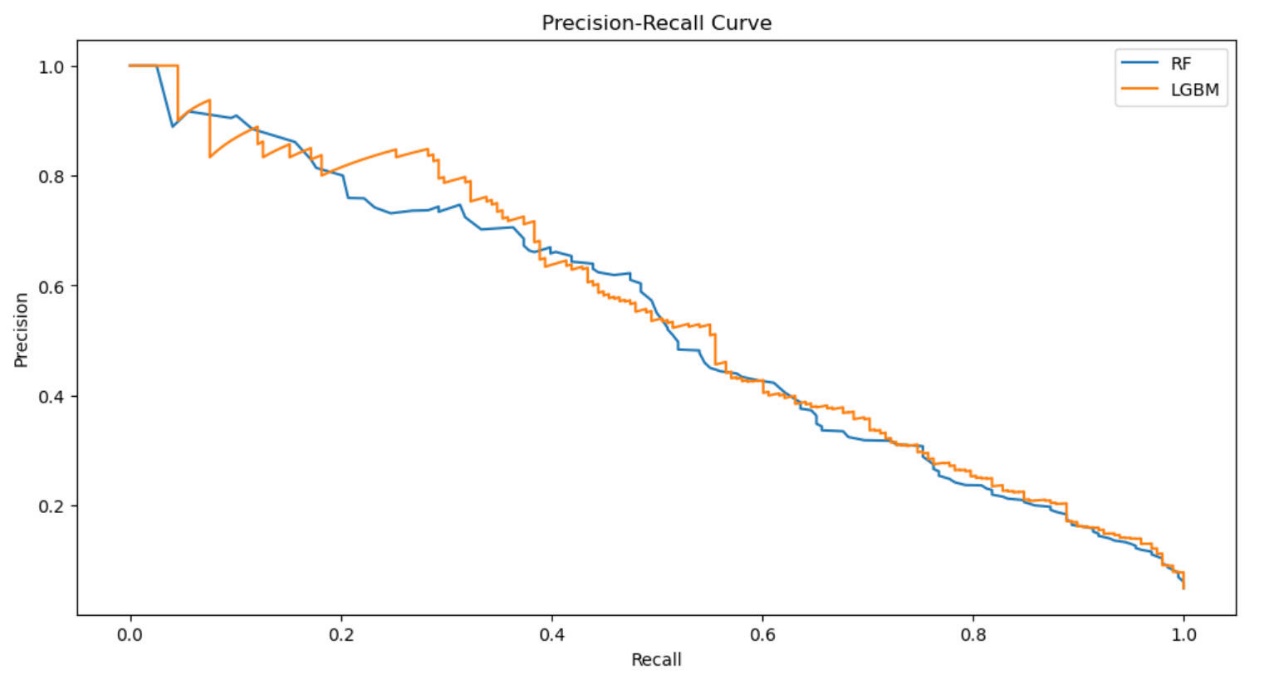
RF特征数-19

LGBM特征数-12

**AUC-ROC变化曲线**



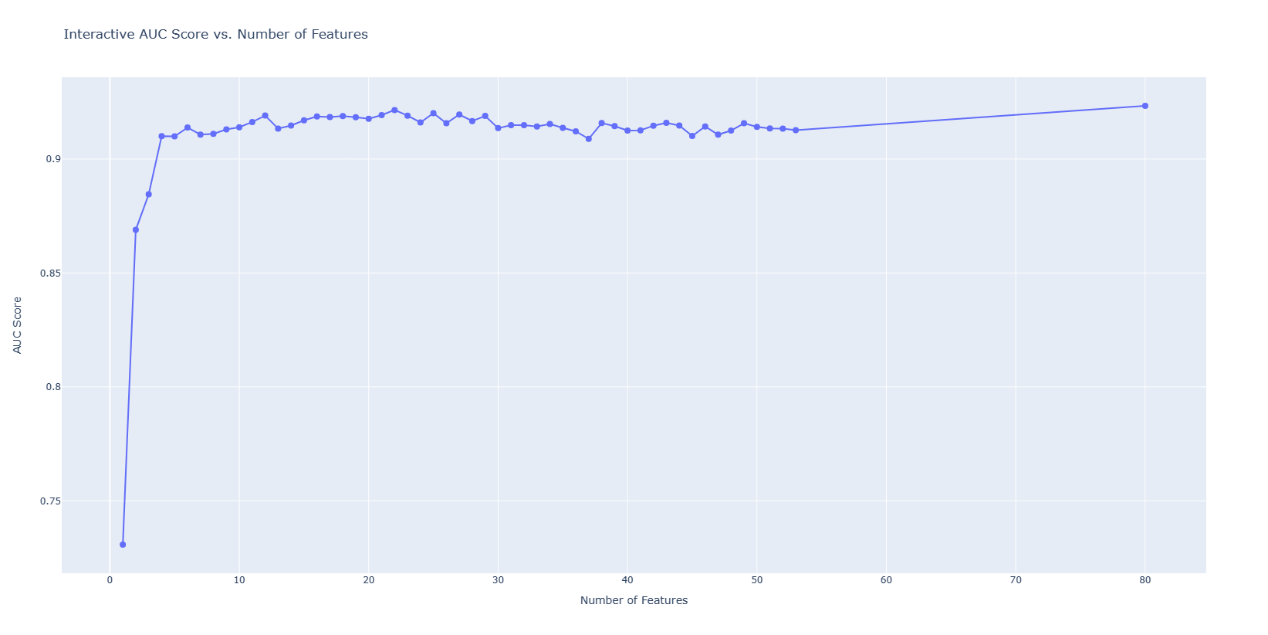
**精确率与召回率变化曲线**



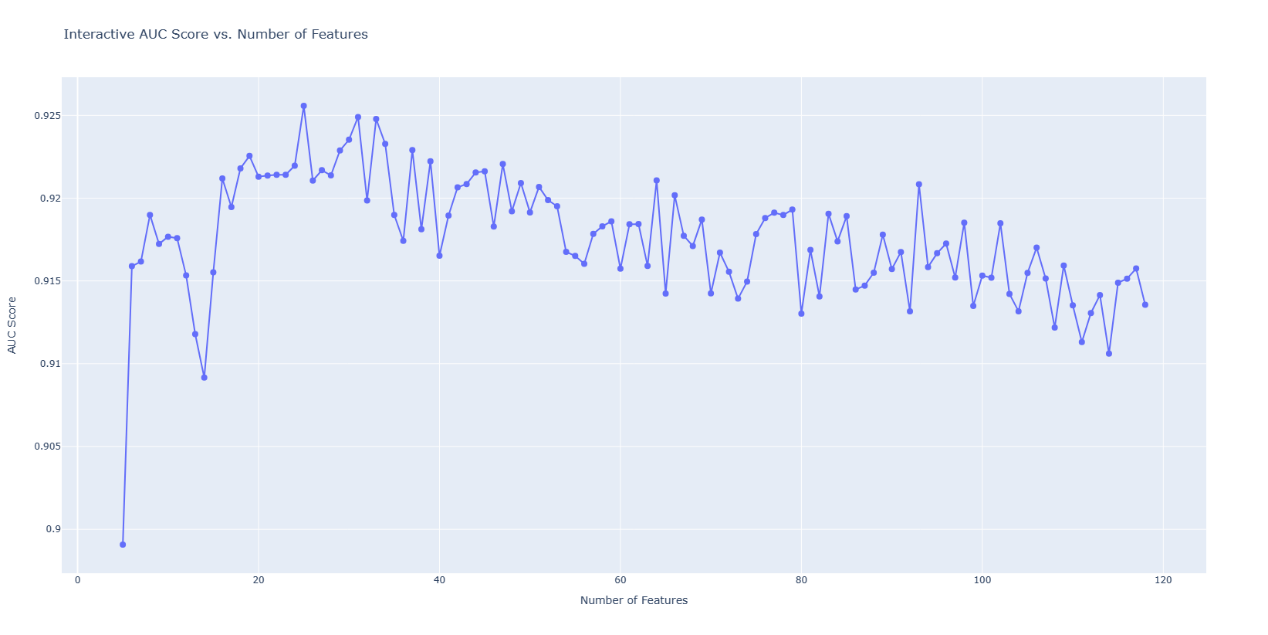
# 2. 特征选择

不同特征数量模型性能变化

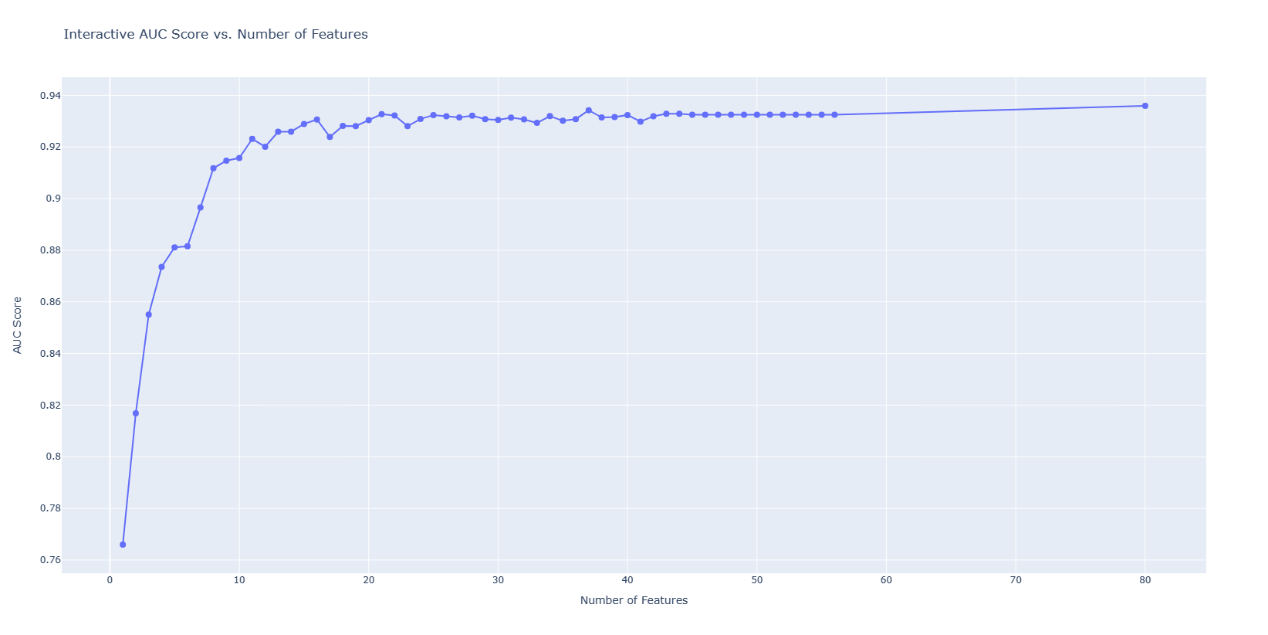
**RF-原数据集**



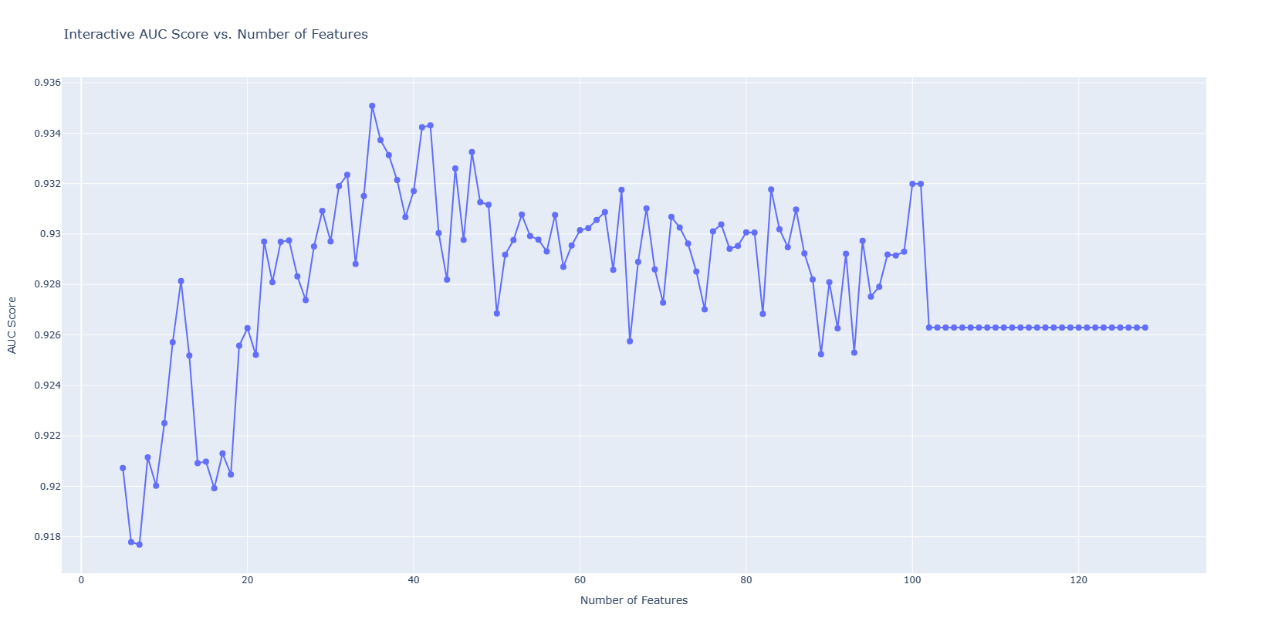
**RF-新数据集**

****

**LGBM-原始数据集**

****

**LGBM-新数据集**

****

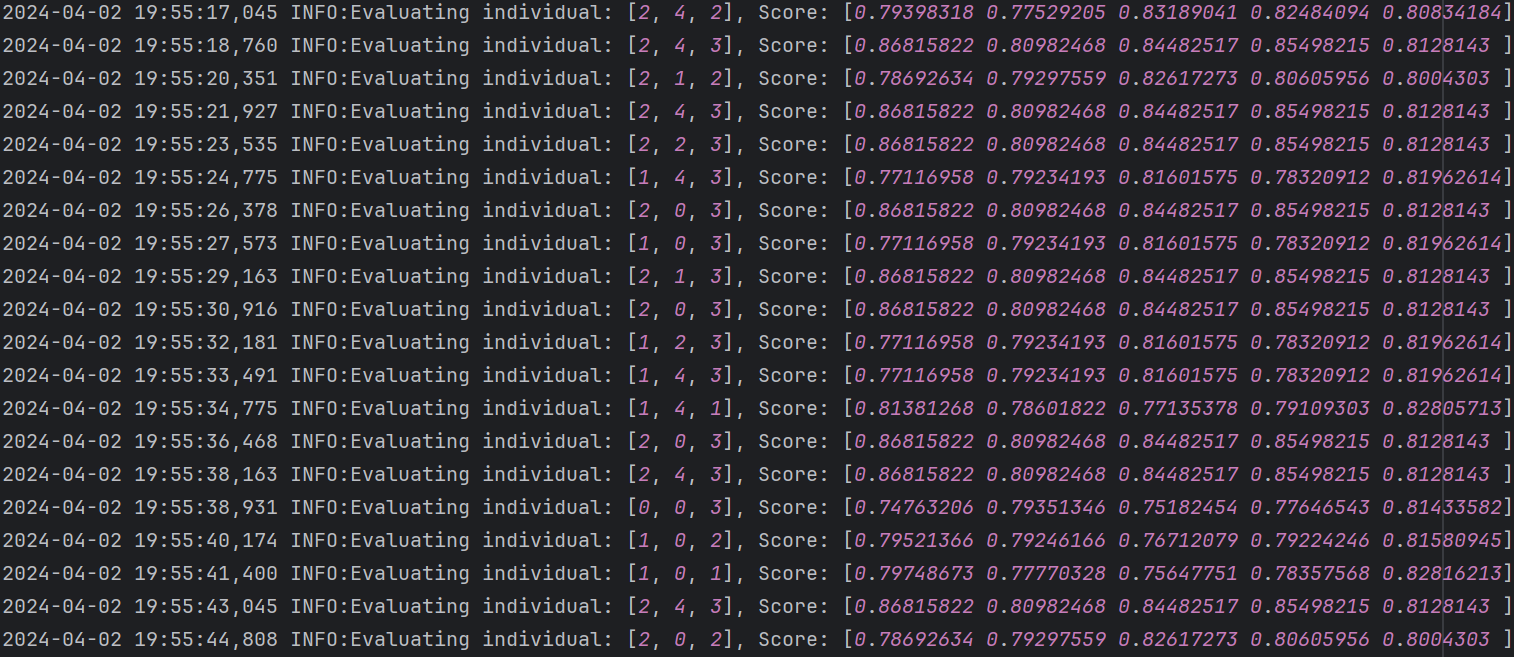
# 3. 超参数调优

决策树超参数调优

|  |  |
| --- | --- |
| 超参数 | 取值/范围 |
| max\_depth | 3, 5, 7, 9, 12, 15, 17, 25 |
| min\_samples\_leaf | 2, 5, 10, 15, 20 |
| min\_samples\_split | 1, 2, 5, 10 |

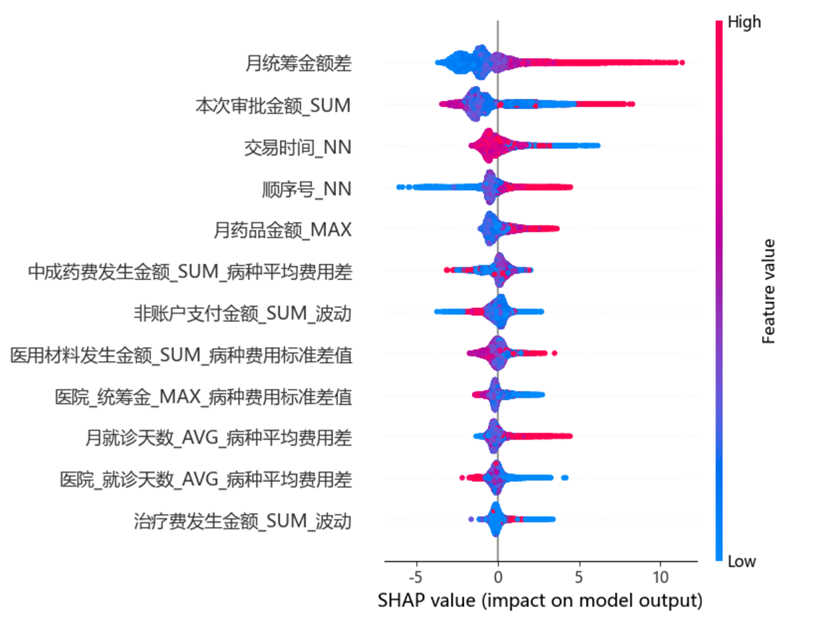
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 调整/评估 | 准确率 | | 宏平均F1 | | 加权平均F1 | | AUC分数 | |
| 训练 | 测试 | 训练 | 测试 | 训练 | 测试 | 训练 | 测试 |
| 调整前 | 1.00 | 0.94 | 1.00 | **0.69** | 1.00 | 0.94 | 1.00 | **0.710** |
| 调整后 | 0.97 | 0.96 | 0.76 | **0.74** | 0.96 | 0.96 | 0.914 | **0.838** |

日志记录（部分）



# 4. 模型案例解释

总结图



1）每个点代表一个数据点的SHAP值，它反映了该数据点的该特征对模型预测的影响。这些点的颜色代表特征的实际数值，从低（蓝色）到高（红色）。特征的实际数值是沿着颜色条的右侧垂直线显示的。

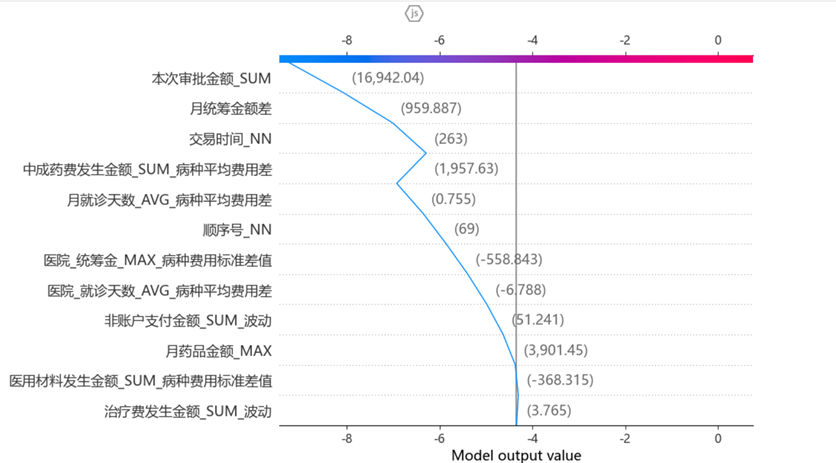
2）解释SHAP总结图：

* 横轴：表示SHAP值，即该特征如何影响模型输出。值可以是正的（推动模型预测向正类），也可以是负的（推动模型预测向负类）。
* 纵轴：列出了数据集中的各个特征，通常按照影响力大小排序。
* 例如，最上面的特征“月药品金额\_MAX”有许多点在SHAP值正值区域，这意味着对于这个特征的高数值，它们推动了模型的预测向正类方向。相反，如果许多点在负值区域，就表明对于这个特征的低数值，它们将模型的预测推向了负类方向。

3）分析总结图的关键点：

* 如果一个特征的点主要集中在SHAP值的正（或负）区域，表明这个特征通常增加（或减少）模型预测正类的概率。
* 如果一个特征的点在SHAP值0附近分布比较广，意味着它对模型预测的影响较小或有正有负，可能依赖于与其他特征的交互。
* 特征点在横轴上分布的宽度显示了这个特征对模型输出的影响的变异性。分布越宽，表明不同数据点上该特征的影响更加多变。
* 从图中可以看出，一些特征（如“顺序号\_NN”和“月统筹金额差”）对模型输出的影响，它们的SHAP值分布宽度较大，并且多数集中在正值区域。

决策图



1）决策图展示了一个单一预测的详细特征贡献。这种图展示如何从一个基线值（通常是数据集的平均预测值）通过各个特征的影响来达到最终的预测值。

2）解释shap决策图：

* 横轴：代表了模型的输出值。在这个例子中，向左的蓝色线表示负面贡献（即这些特征的值导致预测值降低），而向右的红色线表示正面贡献（即这些特征的值导致预测值升高）。
* 纵轴：列出了模型的特征。通常，这些特征会按照它们对预测结果的影响程度进行排序，最顶部的特征对预测结果有最大的影响。
* 线的长度：表示每个特征对预测值的实际贡献大小。线越长，该特征的影响越大。
* 点的位置：表示最终的预测输出值。

3）在图像中，可以看出“本次审批金额\_SUM”特征对模型输出有很大的正向影响，其余特征的贡献量较小，有的正向有的负向，最终这些贡献累积起来形成了最终的模型输出。

更多测试比对结果放在了详细设计报告中