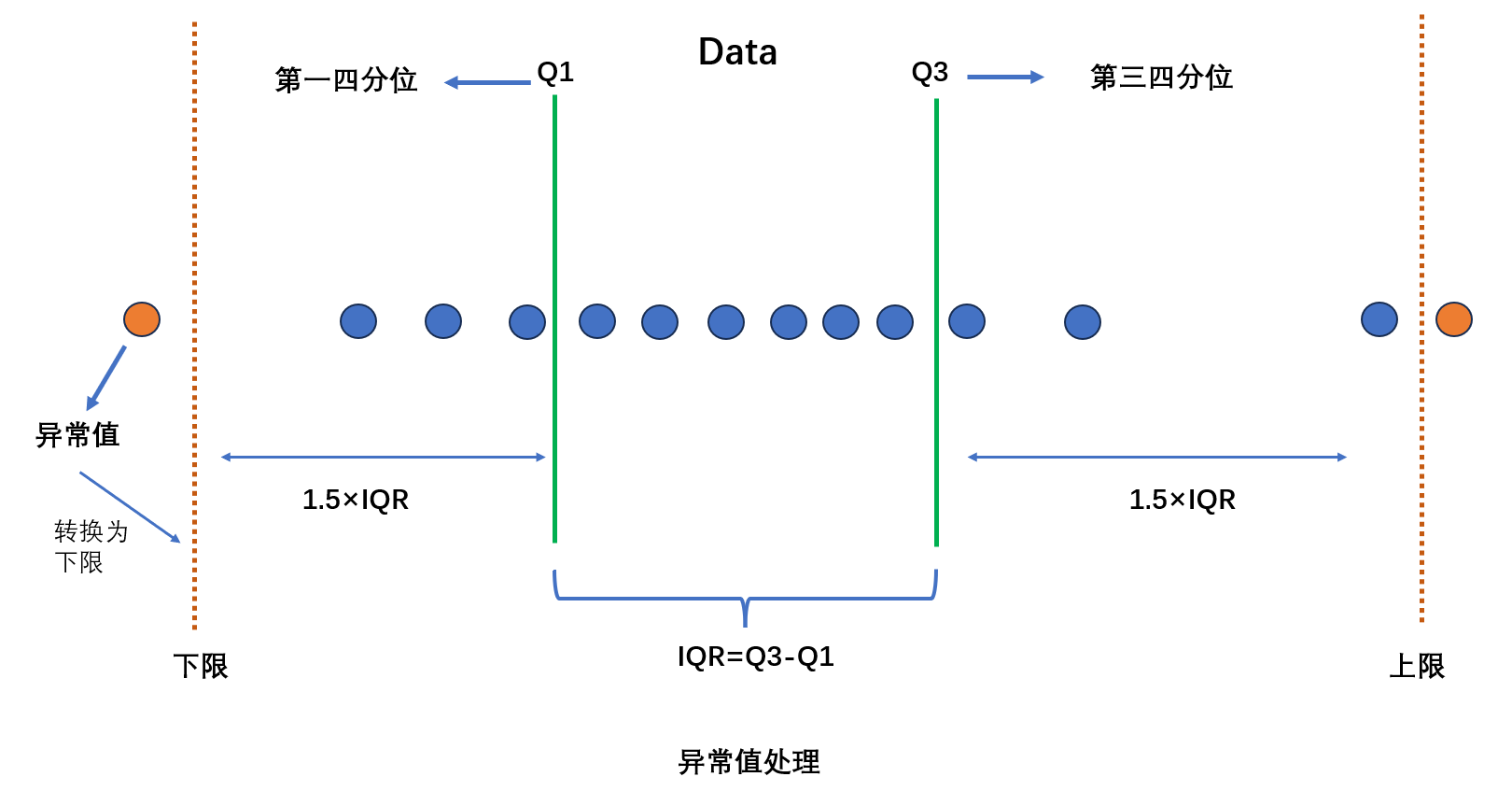
1. **检查并替换异常值**

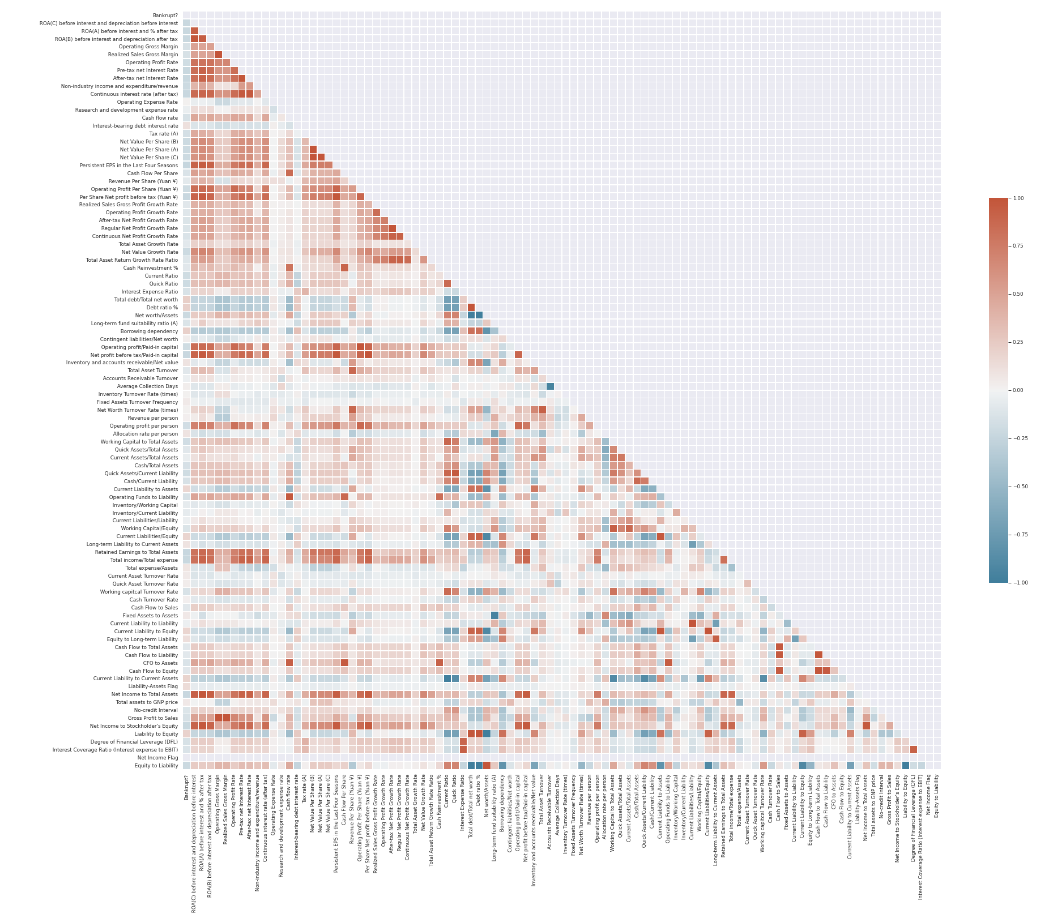
**IQR 方法的依据**

1. **四分位数（Quantiles）**：
   * **第一个分位数（Q1）**：也称为下四分位数，表示数据集从小到大排序后位于 25% 位置的值。
   * **第三个分位数（Q3）**：也称为上四分位数，表示数据集从小到大排序后位于 75% 位置的值。
2. **四分位距（IQR）**：
   * IQR = Q3 - Q1，表示数据中间 50% 的值的范围。
3. **异常值检测**：
   * 数据在 [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] 范围之外的值被认为是异常值。
   * 1.5 是一个经验值，可人为规定
4. **异常值替换**：将数据中的异常值替换为上下限值
5. **删除冗余特征**

1.删除常量特征，即所有值都相同的特征。

* Net Income Flag

1. 删除高度相关的特征，在本方法中删除的是相关系数大于 0.8 的变量，共39个。



1. 删除重复的特征，即特征之间完全相同的列。
2. **BorderlineSMOTE平衡数据集**

**SMOTE取样方法**

SMOTE（Synthetic Minority Over-sampling Technique）是一种用于处理不平衡数据集的过采样方法，主要通过生成合成实例来平衡类别分布。在不平衡数据集中，少数类样本的数量显著少于多数类样本，这可能导致模型在训练过程中偏向于多数类样本。SMOTE 通过在少数类样本之间插值生成新的合成样本，从而增加少数类样本的数量，平衡数据集。

**SMOTE 方法的基本步骤**

1. 选择少数类样本：

从少数类样本中随机选择一个样本。

2. 找到 k 个最近邻样本：

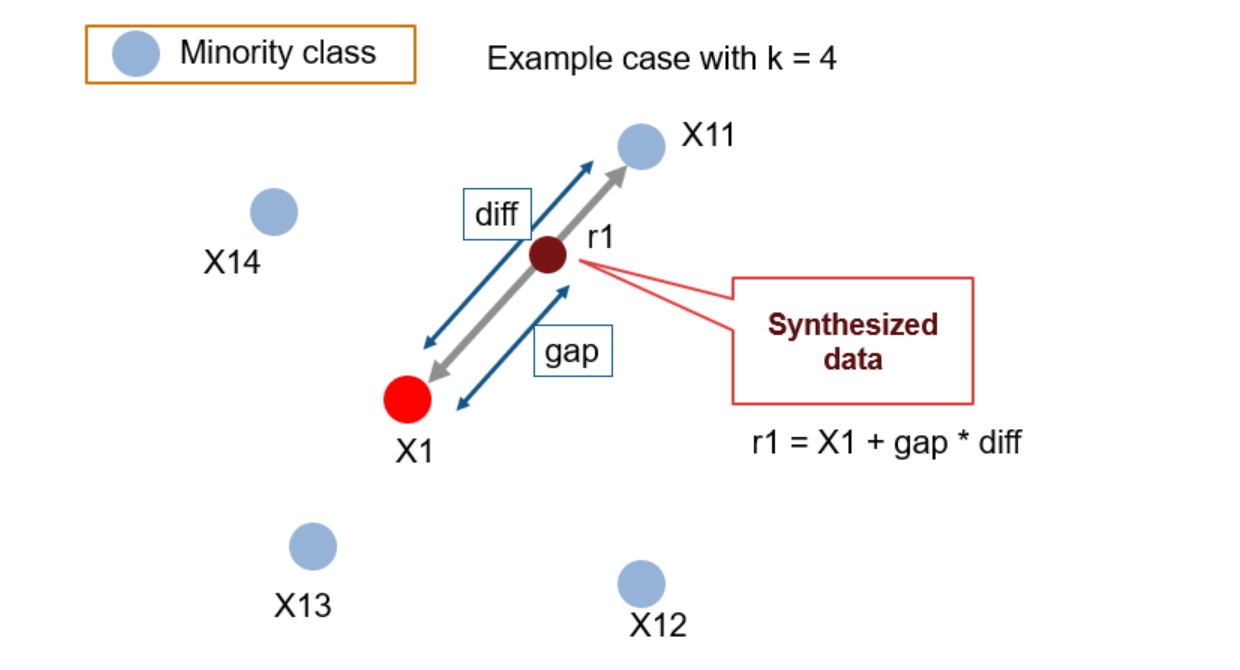
使用欧氏距离或其他距离度量方法，找到该样本的 k 个最近邻样本（通常 k = 5）。

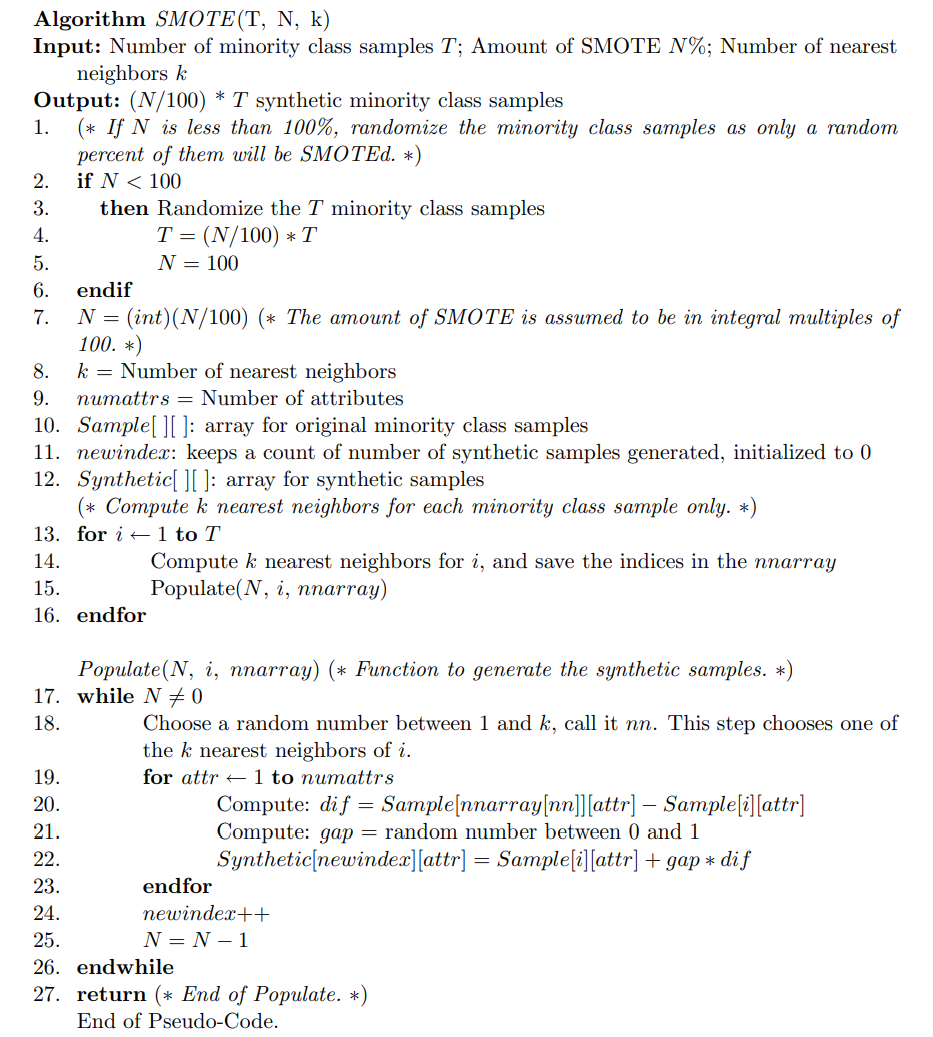
3.生成合成样本：

从这 k 个最近邻样本中随机选择一个样本。

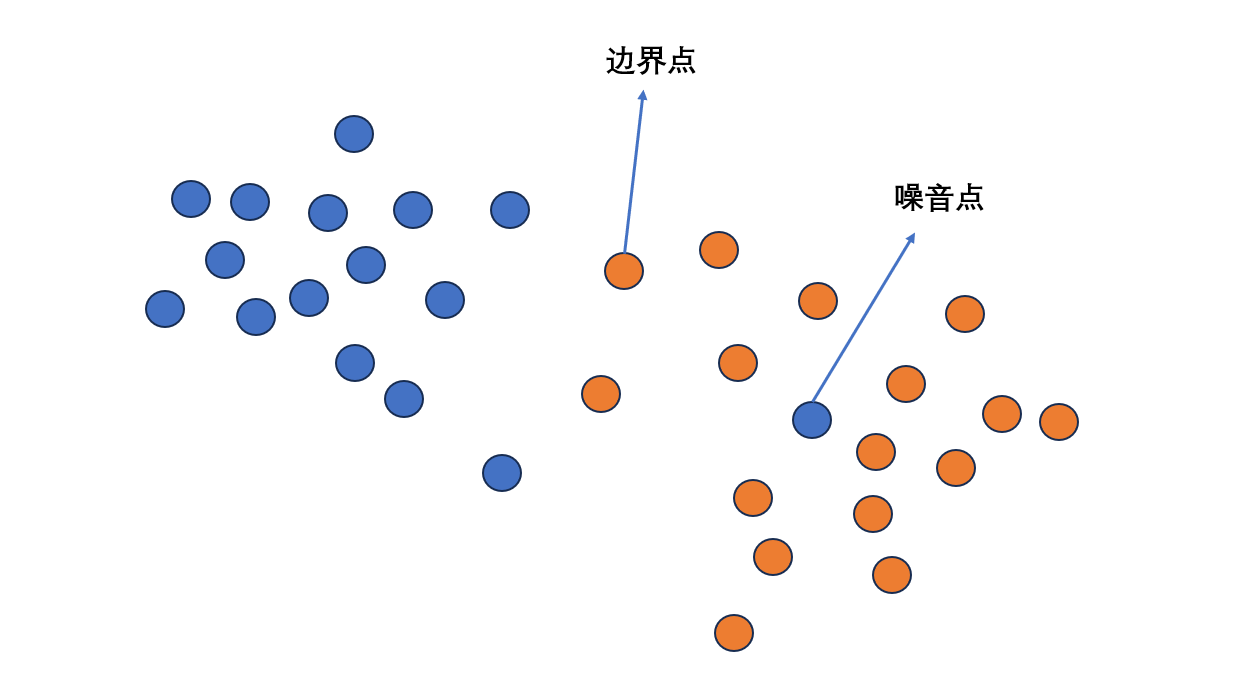
在选择的样本和原始样本之间随机插值生成一个新的合成样本。

4.重复上述步骤，直到生成所需数量的合成样本。



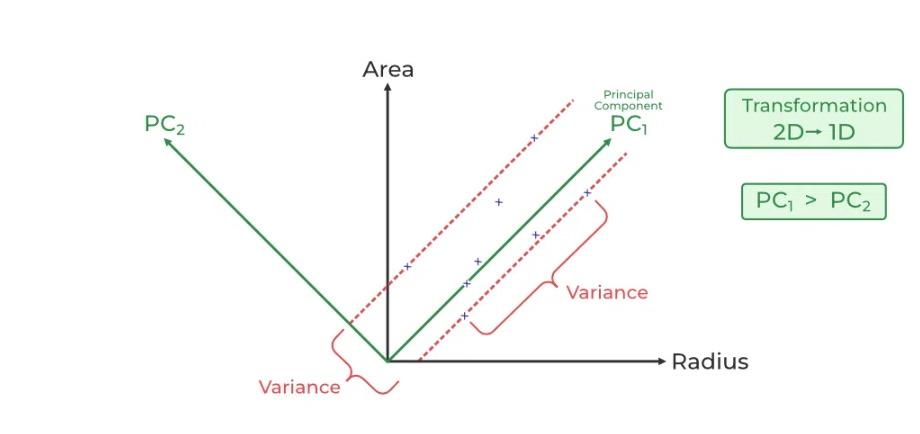


**BorderlineSMOTE**



BorderlineSMOTE算法首先对少数类观察值进行分类。如果某个少数类观察值的所有邻居都是多数类，则将其分类为噪音点，并在创建合成数据时忽略该观察值（类似于DBSCAN）。此外，它将一些具有多数类和少数类邻居的点分类为边界点，然后从除边界点和噪音点外的其他点钟继续采样（通常支持向量机会关注的极端观察值）。

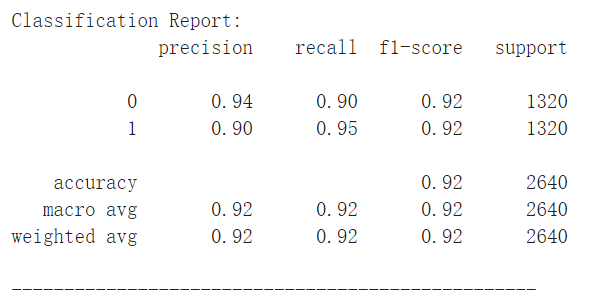
1. **使用PCA方法对特征降维**

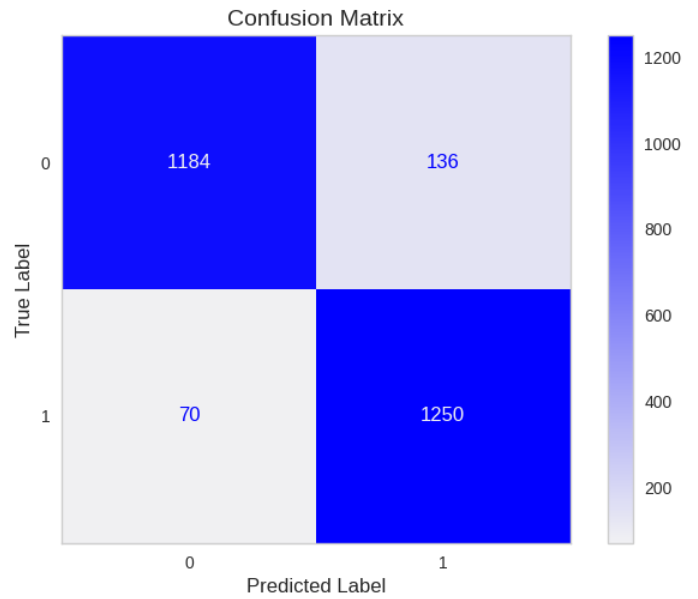


将特征数降为27维

1. **模型选择及结果**
2. **LogisticRegression:**

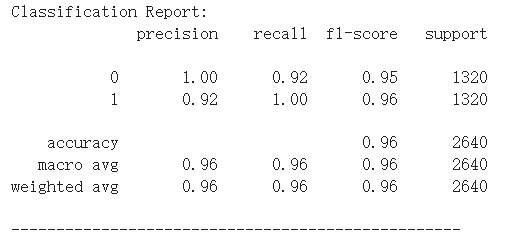
逻辑回归（LogisticRegression）是一种线性模型，常用于二分类问题。使用逻辑函数将输入特征映射到概率。

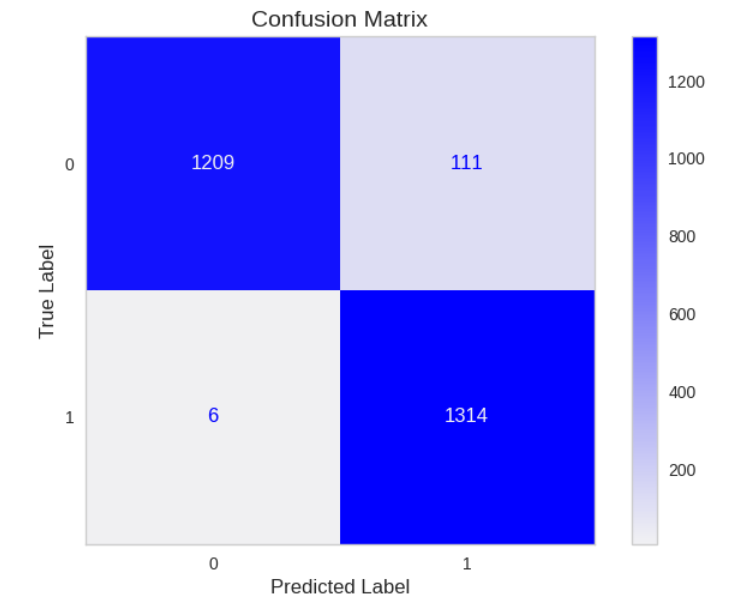




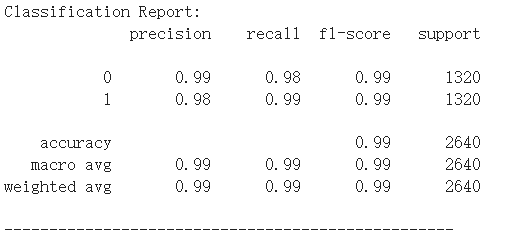
1. **KNeighborsClassifier**

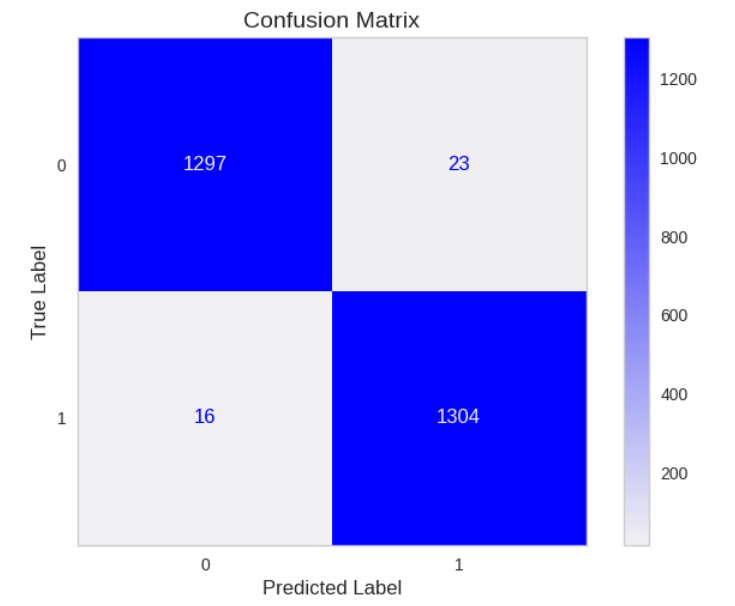
K 近邻分类器（KNeighborsClassifier）是一种基于实例的学习方法，通过计算新样本与训练样本之间的距离来进行分类。



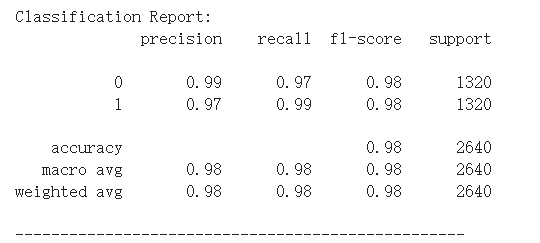


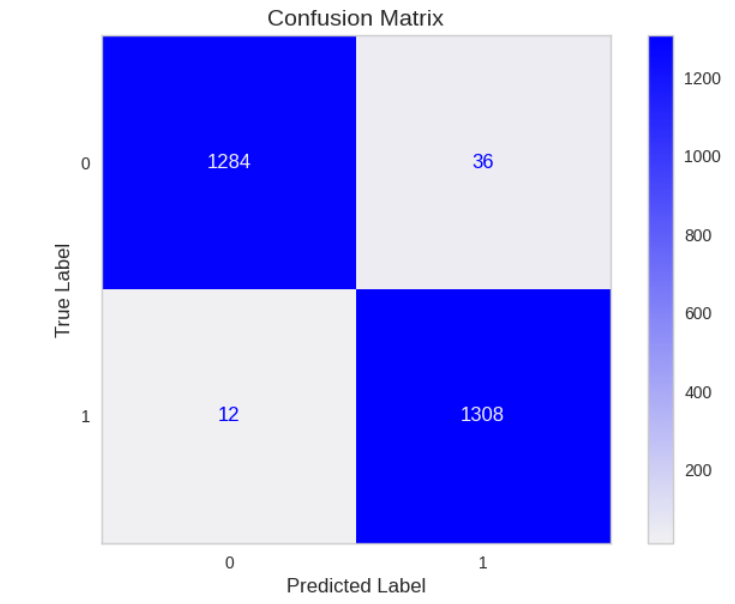
1. **RandomForestClassifier**



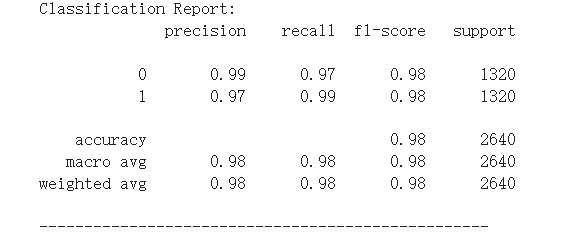


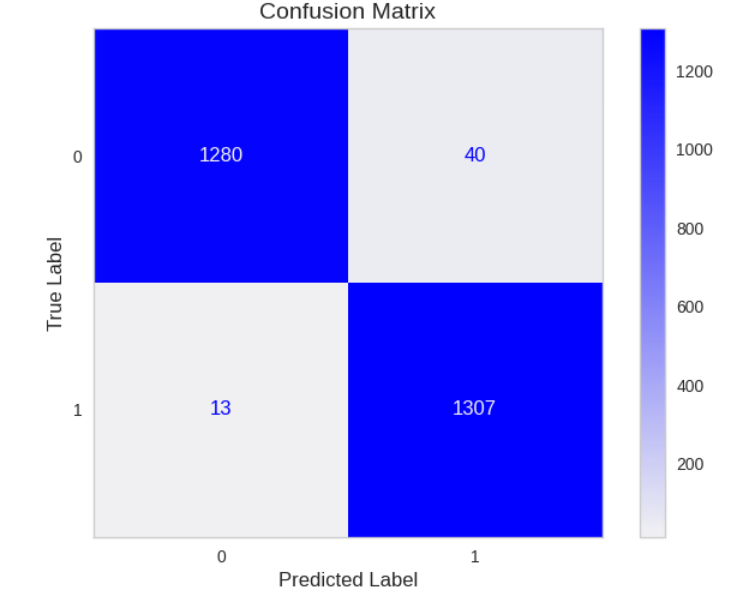
1. **XGBClassifier**



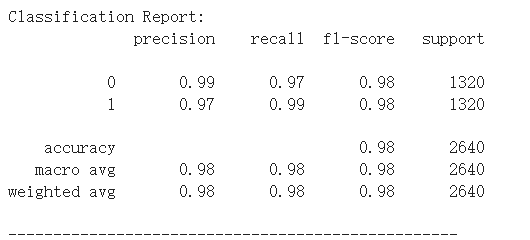


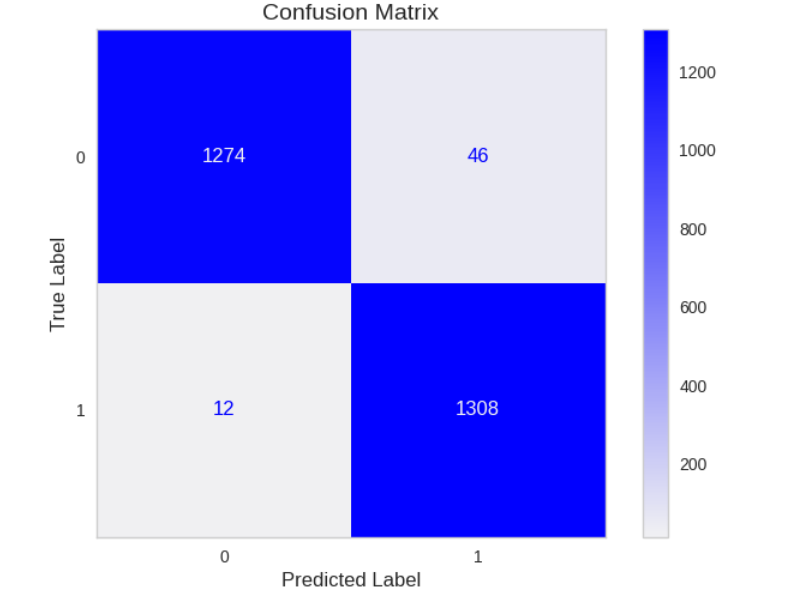
1. **LGBMClassifier**

****

****

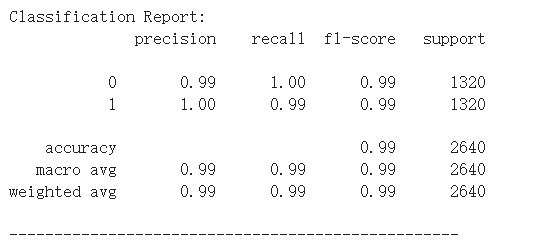
1. **CatBoostClassifier**

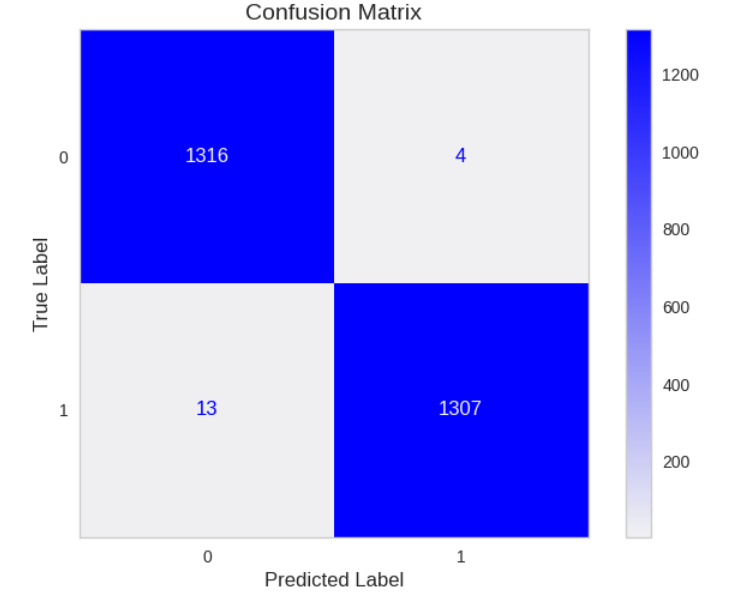
****

****

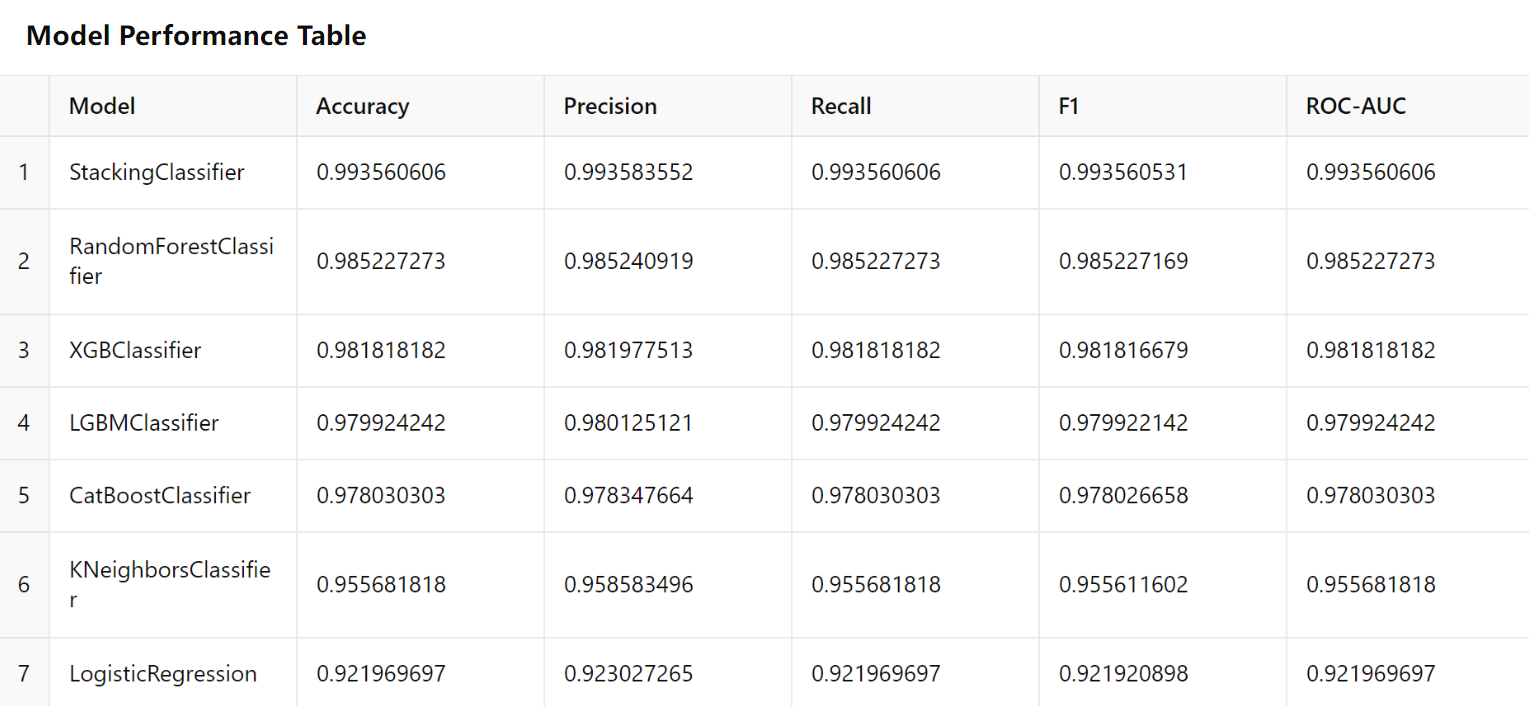
1. **集成学习**

使用 ExtraTreesClassifier、XGBClassifier 和 CatBoostClassifier 作为基础分类器，RandomForestClassifier 作为最终分类器。





1. **结果对比**

****