不平衡数据分类综述及信用卡欺诈检测应用研究

摘要

信用卡欺诈检测是金融风控领域中一项至关重要的任务，其核心挑战在于处理交易数据中固有的、极端的类别不平衡问题。欺诈样本（少数类）的稀缺性严重制约了传统机器学习模型的学习能力，导致其对欺诈行为的识别率（召回率）低下。本文围绕不平衡学习的核心议题，展开了系统性的理论综述和深入的应用讨论。首先，论文详细梳理了数据层面（欠采样与过采样技术，特别是 SMOTE 及其高级变体）和算法层面（代价敏感学习与阈值移动）的主流处理方法。随后，基于一个公开的信用卡交易数据集，本文设计并实施了一套严谨的、多阶段的实验流程。主要成果在于，系统比较了多种数据采样策略对 DecisionTree、XGBoost 等机器学习模型性能的影响，运用了 Optuna 自动化优化框架，实现了模型超参数与分类决策阈值的“协同进化式”寻优。实验结果以数据表明，与基准模型相比，结合了 KMeansSMOTE 过采样与 Optuna 协同调优的最终模型，在测试集上将欺诈交易的召回率从 0.60 达到了 0.93，这充分验证了本文所提出的综合治理方案在解决实际不平衡分类问题上的卓越有效性和实践价值。

关键词: 不平衡数据集；信用卡欺诈检测；数据挖掘；采样；阈值移动；Optuna；机器学习

目录

1 绪论

1.1 项目背景与意义

在数字经济时代，信用卡已成为全球范围内最主要的非现金支付工具之一。其便捷性在促进商业繁荣的同时，也为各类金融欺诈活动提供了温床。信用卡欺诈不仅给持卡人和金融机构造成每年数百亿美元的直接经济损失，还衍生出高昂的案件调查成本、客户服务压力，并严重侵蚀用户信任和品牌声誉。因此，构建快速、精准、智能的信用卡欺诈检测（Credit Card Fraud Detection, CCFD）系统，已成为所有金融机构风险管理体系中的核心战略要务。

数据挖掘与机器学习技术的发展为欺诈检测提供了强有力的武器。通过对海量历史交易数据进行学习，模型能够自动发现正常与欺诈交易之间的复杂模式，从而对新的交易进行实时风险评估。然而，这一应用场景面临着一个经典且棘手的数据科学难题——类别不平衡（Class Imbalance）。

1.2 类别不平衡的挑战

在真实的信用卡交易数据中，欺诈行为是小概率事件。正常交易与欺诈交易的数量比例可能达到 1000:1，甚至更高。本项目所使用的数据集中，欺诈样本占比仅为 0.172%，属于极端不平衡。这种数据分布特性对标准分类算法构成了严峻挑战：

1) 模型学习偏见：大多数机器学习算法的目标是最小化整体误差，即最大化总体准确率。在不平衡数据上训练时[1][2]，模型的损失函数会被数量庞大的多数类（正常交易）所主导。模型会发现，将所有样本都预测为多数类，便能轻易获得极高（如 99.8%）的准确率。这形成了一种“惰性学习”的偏见，使得模型对我们真正关心的少数类（欺诈交易）缺乏有效的辨别能力。
2) 决策边界漂移：理想的分类模型应能学习到区分不同类别的精确决策边界。在不平衡数据的影响下，由于少数类样本过少，无法在特征空间中形成足够强大的“势力范围”，所学习到的决策边界会严重偏向少数类一侧，甚至完全将其“淹没”，导致大量少数类样本被错误分类。

3) 评估指标失效：如前所述，准确率（Accuracy）在不平衡场景下会产生“准确率悖论”（Accuracy Paradox），即一个高准确率的模型可能完全没有实际应用价值。因此，必须采用更能反映模型对少数类识别性能的评估指标，如精确率（Precision）、召回率（Recall）和 F1-Score。

1.3 本文阐述内容与结构

为系统性地解决上述挑战，本文基于一个完整的 Python 数据挖掘项目，设计并实施了一系列实验，旨在探索并验证一套综合性的不平衡数据处理方案。本文的主要内容包括：

1) 进行细致的数据探索性分析（EDA），理解数据特性。
2) 深入剖析不平衡学习的多种核心技术，并阐述其理论依据。
3) 对比评估不同数据采样策略的实际效果。
4) 应用 Optuna 框架，对模型超参数和分类阈值进行一体化、自动化的深度优化。
5) 通过实验数据结果，层层递进地展示优化过程带来的性能提升。

本文组织结构如下。第二章详细介绍了不平衡学习的方法论（包括数据层面和算法层面）、本次实验选用的机器学习模型（XGBoost、逻辑回归、决策树）以及 Optuna 自动优化框架的理论基础。第三章描述了本次研究的实验设计，涵盖了数据集概况、评估体系、探索性数据分析、数据预处理方法以及一个分为四个阶段的详细建模与评估流程。第四章呈现并深度剖析了从基准模型到最终优化模型的实验结果，重点讨论了不同采样策略的有效性和超参数与阈值协同优化的核心价值。第五章对全文工作进行了总结，并对未来可行的研究方向进行了展望。

2. 相关技术与理论基础

2.1 不平衡学习方法论

不平衡学习（Imbalanced Learning）旨在解决数据分布不均下的分类问题，其主流方法可从数据、算法两个层面进行划分。

2.1.1 数据层面方法

数据层面的方法通过直接修改训练数据集的分布，使其变得相对均衡，从而让标准分类器能够在“舒适区”内进行学习。这类方法主要分为欠采样和过采样[3]。

欠采样（Under-sampling）是通过减少多数类样本来平衡数据，欠采样的主要具体算法有：

1) 随机欠采样 (RandomUnderSampler)：是最简单的方法，随机地从多数类中丢弃样本直至达到期望比例。优点是速度快，但其核心缺陷在于可能导致信息丢失，即被丢弃的样本可能包含对于构建精确决策边界至关重要的信息。
2) 原型选择法 (Prototype Selection)：这类方法尝试更智能地选择要保留或移除的样本。项目配置中包含的 ClusterCentroids 是一种原型生成方法，它使用 K-Means 算法将多数类样本聚类成多个簇，然后用每个簇的质心来代表该簇的所有样本。这样做能在大幅减少多数类数量的同时，在一定程度上保留其原始的分布结构，优于随机丢弃。

过采样（Over-sampling）是通过增加少数类样本来平衡数据，是当前应用更广泛的策略，过采样的具体算法有：

1) SMOTE (Synthetic Minority Over-sampling Technique)：作为里程碑式的方法，SMOTE 的核心思想是创造合成的新样本而非简单复制。其流程为：首先对每个少数类样本 x，找出其在少数类邻居中的 k 个近邻；然后随机选择一个近邻 x'；最后在 x 与 x'之间的连线上随机取一点作为新的合成样本。SMOTE 有效避免了简单复制带来的过拟合，但其也存在盲目性，可能在少数类和多数类样本重叠的区域生成噪声，或在本身就很稀疏的区域过度泛化。
2) BorderlineSMOTE: 该方法是 SMOTE 的改进，它认为处于类别边界的少数类样本比处于“安全”区域的样本更容易被错分，因此也更重要。它首先将少数类样本分为三类，安全点（邻居全是少数类）、危险点（邻居一半以上是多数类）和噪声点（邻居全是多数类）。然后，它只针对“危险点”来合成新样本，从而强化了模型在决策边界区域的学习能力。

3) KMeansSMOTE: 这是本项目重点采用的一种高级变体。它试图解决 SMOTE 可能生成噪声的问题，其核心思想是在“安全”且“密集”的区域生成样本。其流程为：首先使用 K-Means 算法对整个数据集进行聚类；然后筛选出那些少数类样本占比较高的簇；最后在这些被选中的“少数类优势簇”内部，按权重（簇内少数类样本密度）分配要生成的样本数量，并应用 SMOTE。这种方法可以确保新样本生成在少数类聚集的区域，有效避免了噪声的产生。

4) ADASYN (Adaptive Synthetic Sampling)：另一种自适应的 SMOTE 变体。它根据少数类样本的“学习难度”来决定为其生成多少新样本。学习难度由该样本的 k 近邻中多数类样本的比例来衡量。一个少数类样本周围的多数类越多，说明它越难被正确分类，ADASYN 就会为其生成越多的合成样本。这相当于一种自适应地将学习重心向困难样本倾斜的策略。

2.1.2 算法层面方法

算法层面的方法不改变数据本身，而是通过修改学习算法使其能更好地适应不平衡数据，主要具体算法有：

1) 代价敏感学习 (Cost-Sensitive Learning)：该方法为不同类别的错分引入不同的代价。在欺诈检测中，将欺诈交易（正类）错判为正常交易（假阴性, FN）的代价远高于将正常交易错判为欺诈（假阳性, FP）。代价敏感学习通过在模型的损失函数中为少数类样本赋予更高的权重（class_weight 参数）来实现。例如，设置 class_weight={0:1, 1:100}，意味着模型错分一个欺诈样本的惩罚是错分一个正常样本的 100 倍。这会迫使模型在训练过程中更加关注 w 少数类样本，努力降低假阴性的数量。

2) 阈值移动 (Threshold Moving)：这是一个极其重要且实用的后处理技术。分类器（如逻辑回归、XGBoost）的 predict_proba()方法会输出每个样本属于正类的概率。默认情况下，系统以 0.5 为分界点（阈值），概率 > 0.5 则判为正类。这个 0.5 的阈值隐含了两个假设：1) 类别是平衡的；2) 两种错分代价是相等的。这两个假设在欺诈检测中都不

成立。由于模型在不平衡数据上训练，其对欺诈样本的预测概率通常偏低，可能集中在 0.1-0.4 之间。若仍坚持 0.5 的阈值，将导致绝大多数欺诈样本被漏掉。 阈值移动的核心思想是主动降低这个决策门槛。例如，将阈值从 0.5 下调至 0.2，意味着只要模型认为一个样本有超过 20%的可能是欺诈，就将其标记出来。这必然会增加假阳性（误报）的数量，即降低精确率，但它能显著找回大量原先被遗漏的欺诈样本，即大幅提升召回率。最优阈值的确定，本质上是在业务可接受的精确率和最大化召回率之间进行权衡。

2.2 模型选择

模型选择上，本实验采用三种模型进行相互比对，分别是：

1) XGBoost (Extreme Gradient Boosting)[4]：这是本实验效果最好的模型，作为梯度提升决策树（GBDT）的一种高效工程实现，XGBoost 已成为结构化数据建模领域的利器。其优势在于内置了 L1 和 L2 正则化项，能有效控制模型复杂度，防止过拟合;实现了并行的树构建过程，训练速度快;能够处理缺失值，并内置了特征重要性评估。其强大的非线性建模能力非常适合挖掘金融交易数据中隐藏的复杂欺诈模式。

2) Logistic Regression (逻辑回归)：作为经典的广义线性模型，逻辑回归在二分类问题（如欺诈判定）中具有基础而重要的地位。其核心是应用 Sigmoid 函数将特征的线性组合映射到(0, 1)区间，从而得到一个概率值。其数学表达式为：

$$P(\text{Class} = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

其中，$P(\text{Class} = 1 \mid \mathbf{x})$ 是给定特征向量 $\mathbf{x}$ 后，样本被预测为欺诈类别（正类）的概率。$\mathbf{w}$ 代表模型学习到的特征权重向量，$\boldsymbol{b}$ 是偏置项。

模型的核心优势在于结构简单、训练速度快、计算资源消耗低、特别适合大规模数据集的初步筛选或基线模型构建；输出结果具有天然的概率解释性，可直接反映欺诈风险发生的可能性；模型参数（特征系数）直观可解释，能清晰揭示各个特征对最终欺诈概率的正向或负向影响，这对于满足金融风控领域的强监管和可解释性要求至关重要。

虽然其本质是线性模型，但通过特征工程（如特征交叉、多项式变换）也能捕捉一定的非线性关系。

3) Decision Tree (决策树)：决策树通过模拟人类决策过程，构建基于特征阈值判断的树形结构进行分类。其主要特点包括：1) 模型结构高度直观且易于可视化，决策规则（IF-THEN 形式）清晰易懂，便于业务人员理解和验证，符合风控透明度需求；2) 能够自动处理非线性关系，并对特征间的交互作用进行有效建模，无需复杂的特征工程即可捕捉数据内在模式；3) 对数据分布假设宽松，能同时处理数值型和类别型特征，对输入数据的量纲不敏感。然而，单棵决策树容易过拟合训练数据，对噪声敏感，且模型稳定性相对较差（数据微小变化可能导致树结构显著变化）。

2.3 自动优化框架

参数的选择对模型性能至关重要，传统的手动调参或网格搜索（Grid Search）既耗时又低效。Optuna 是一个现代化的贝叶斯超参数优化框架。其核心工作机制是：

1) 用户定义一个函数，该函数接收一组超参数（由 Optuna 提供），然后基于这组参数训练模型并返回一个需要被优化的评估分数（例如，加权 F1-Score）。
2) 创建一个研究实例 (Study)对象负责管理整个优化过程。
3) Optuna 使用一种名为 TPE (Tree-structured Parzen Estimator)的算法。与随机搜索不同，TPE 会根据历史试验的结果（哪些参数组合得到了好的分数，哪些得到了差的分数）建立一个概率模型，然后利用这个模型来推断出下一组最有可能产生更好结果的超参数组合进行试验。这种基于历史经验的“智能”探索，使得 Optuna 能比传统方法更快地收敛到最优解。 本实验最大的亮点在于，将分类阈值 threshold 也视为一个超参数，与 max_depth 等模型参数一同放入 Optuna 的目标函数中进行优化，实现了一体化的寻优过程。

3. 实验设计与流程

为了系统性地应对信用卡欺诈检测中的极端不平衡挑战，并科学地评估各项策略的有效性，本实验设计了一套严谨、可复现的多阶段流程，该流程涵盖了从数据探索到最终模型验证的全过程。

3.1 数据集介绍

本实验采用 Kaggle 上公开的“Credit Card Fraud Detection”数据集。该数据集包含 284,807 条交易记录，其中正常交易（Class=0）284,315 条，欺诈交易（Class=1）492 条。欺诈样本占比仅为 0.172%，是典型的极端不平衡数据集。数据特征包含 28 个经过 PCA 处理的匿名数值特征（V1-V28），以及 Time（该交易与数据集中的第一个事务之间经过的秒数）和 Amount 两个原始特征。数据被划分为训练集和测试集，严格保证了测试集在整个模型构建与调优过程中“不可见”，以模拟真实世界中对未知数据的预测场景。

3.2 评估体系

鉴于“准确率”在不平衡任务中的误导性，本实验构建了一个面向业务目标的加权评估体系。该体系在项目的 config.py 文件中明确定义，旨在量化模型在多个维度上的综合性能。各指标权重设定如下：

1) 召回率 (Recall): 权重 0.4。此为最高权重，首要目标是尽可能多地识别出所有欺诈交易，减少漏报造成的金融损失。
2) 精确率 (Precision): 权重 0.2。确保在提升召回率的同时，控制误报率，减少对正常用户的打扰和不必要的审核成本。
3) F1-Score: 权重 0.2。作为精确率和召回率的调和平均，提供一个平衡的视角。
4) ROC-AUC: 权重 0.15。衡量模型整体的排序能力和区分正负样本的性能。
5) 准确率 (Accuracy): 权重 0.05。仅作为参考，权重极低。

在实验中，所有模型的最终性能评估，特别是在 Optuna 自动调优阶段，都以这个加权分数（Weighted Score）作为最终的优化目标。该分数的计算由

metric.py 模块中的 weighted_score_auto 函数实现，确保了模型优化的方向与真实业务需求的高度一致。

3.3 探索性数据分析

在正式建模前，深入的数据探索（EDA）揭示了数据的内在规律，为后续策略选择提供了依据。

分析的第一步是对数据集进行宏观层面的整体把握。首先需要确认数据集的完整性与基本构成。该数据集包含 284,807 条交易记录和 31 个字段，所有特征，包括 Time、Amount 以及 28 个经过主成分分析（PCA）处理的匿名特征（V1 至 V28），均为数值类型，且不存在任何缺失值，数据的高度完整性免去了复杂的数据插补工作。

本次 EDA 中最核心的发现来自于对目标变量 Class 的分布统计。分析结果精确地揭示了问题的极端不平衡性，在总计 284,807 条记录中，正常交易（Class=0）为 284,315 条，而欺诈交易（Class=1）仅有 492 条，如图 1 可以直观体现。这意味着欺诈样本的占比仅为 0.172%。这正式本实验的挑战所在，更是整个方法论的逻辑起点。它揭示了为何传统的准确率（Accuracy）在此场景下会失效，并为后续放弃准确率、转而构建以召回率（Recall）为核心的加权评估体系提供了强有力的论证。

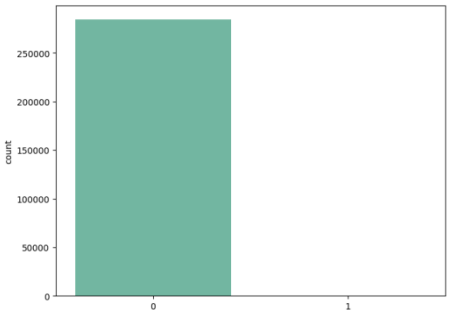

图 1 Class 分布

在字段探索中，观察到 Amount（交易金额）特征的数值尺度与其他特征存在巨大差异。其极大的标准差与取值范围表明，若不进行处理，该特征将在模型训练中占据不成比例的主导地位。这一发现直接导向了在后续流程中必须对所有特征进行标准化处理（如 StandardScaler）的决策，以消除量纲影响，确保所有特征在同等尺度上对模型做出贡献。

在完成单变量分布分析后，为进一步挖掘特征间的潜在关联，构建了全特征相关性热力图（如图 2）。该热力图通过计算所有特征（含 Time、Amount 及 V1-V28）的皮尔逊相关系数矩阵，并采用颜色梯度呈现相关性强弱。

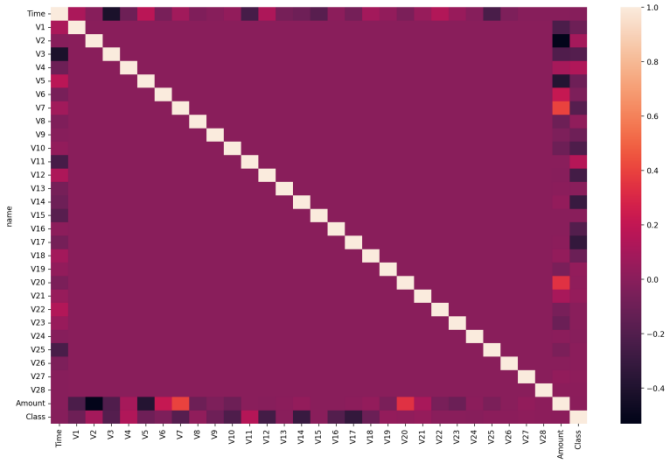

图 2 全特征相关性热力图

由于特征比较多，导致热力图不是很清晰，下面直接把 Class 和特征的相关性提取出来，如图 3。从图中可以看出，V11、V4、V2 等特征与 Class 的相关系数较高（>0.2），表明它们与目标变量呈显著正相关，即与欺诈高度相关；大部分特征（如 V9, V15, V21）与 Class 的相关系数接近 0，说明它们与目标变量无明显线性关系；V12、V14、V17 等特征与 Class 的相关系数较低（<-0.2），表明它们与目标变量呈显著负相关，即与正常交易高度相关；Amount（交易金额）与 Class 的相关系数接近 0，可能表明金额大小与欺诈无直接关联。

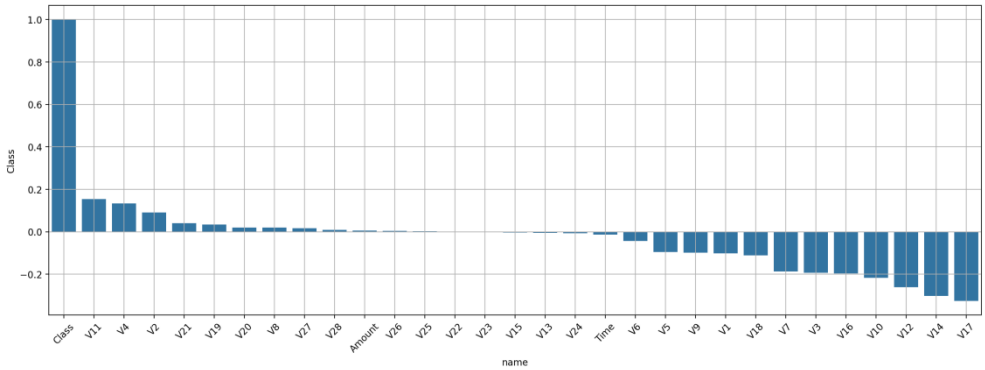

图 3 特征与 target 相关性分布

在掌握数据整体特性后，我们通过核密度估计（Kernel Density Estimation, KDE）深入解析关键特征在正常交易（Class=0）与欺诈交易（Class=1）中的分布差异，揭示出显著的预测能力分化。本文主要选取部分强预测特征继续分析分析，这部分特征呈现卓越的类别区分能力，其分布曲线在两类交易中显著分离：

1) V17 分布特性（图 4）：正常交易在 V17≈0 处形成尖锐单峰（均值=0.01），呈拉普拉斯对称分布；欺诈交易集中于 V17∈[-20, 0]负值区（均值=-6.67），形成右偏分布且与正常主峰相距>6 个标准差。两曲线仅在 V17≈-10 处微弱重叠，分离度极强。

2) V11 分布特性（图 5）：正常交易在 V11≈-0.5 处形成单峰（均值=-0.01），接近高斯分布；欺诈交易向右侧偏移，于 V11∈[3, 8]形成次级高峰（均值=3.80），与正常主峰间隔≈4 个标准差。虽在 V11∈[-2, 0]存在部分重叠，整体分离度显著。

V17 和 V11 特征展现的分布差异表明其具备单变量判别能力，当 V17 > -10 或 V11 > 3 时欺诈概率显著升高。此类特征构成模型的核心预测信号，可有效区分两类样本。

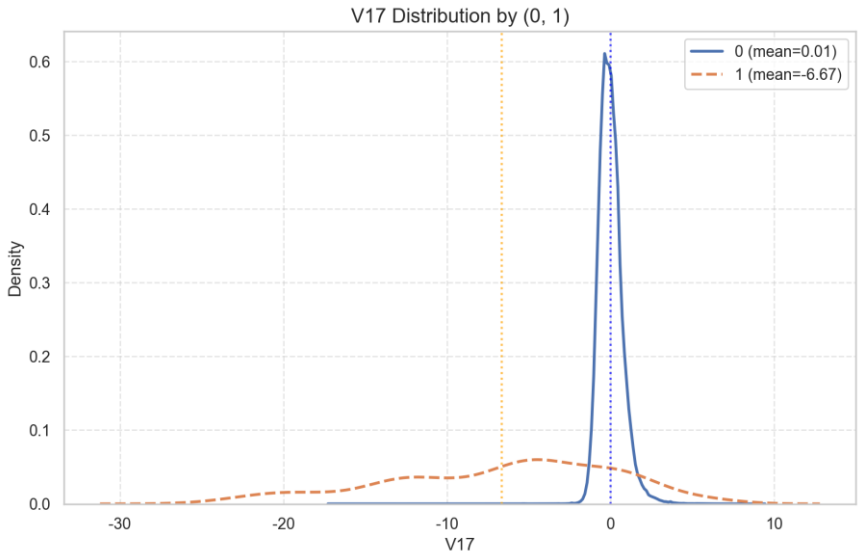

图 4 变量 V17 在类别(0 和 1)下的分布图

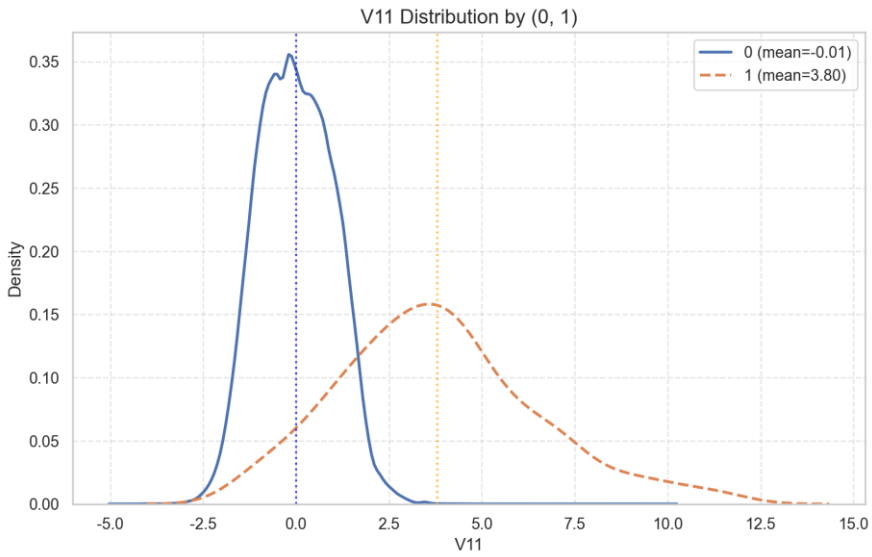

图 5 变量 V11 在类别(0 和 1)下的分布图

为进一步探究特征间的非线性交互作用，我们进一步分析了强预测特征的两两联合分布。 选取两组代表性特征组合（V11-V4 和 V17-V14）进行核密度估计（KDE）可视化分析（如图 6），结果显示两类样本在特征空间中呈现出显著的差异化分布模式。V11-V4 组合分析表明，正常交易样本（Class=0）在 V11 ∈ [-2, 2] 且 V4 ∈ [-5, 5] 范围内形成高度集中的簇，密度中心位于坐标原点附近（V11 ≈ 0, V4 ≈ 0）；而欺诈样本（Class=1）显著偏离该区域，主要分布于 V11 > 2 且 V4 > 0 的广阔空间，形态相对分散。两者高密度区域几乎无重叠，表现出极强的可分离性。V17-V14 组合分析显，正常样本集中于 V17 ∈ [-5, 5] 且 V14 ∈ [-10, 0] 区域，密度中心位于 V17 ≈ 0, V14 ≈ -5；欺诈样本分布范围更广，并在 V17 < -5 且 V14 < -10 的负值极端区域形成次级高密度区。尽管存在部分分布重叠，欺诈样本呈现显著的偏态特征。

上述发现具有三重启示：

1) 特征协同效应：欺诈识别不仅依赖单特征异常值（如 V11 > 5 或 V17 < -15），更关键在于特征间协同作用（如 V11-V4 的正向联动、V17-V14 的负向联动）；
2) 模型优化方向：针对非球形分布特性，建议优先选用树集成模型（XGBoost、LightGBM）捕捉复杂非线性交互；在神经网络中引入显式交叉特征层或采用 RBF 核 SVM；显式构造强关联特征交互项（如 V11 × V4）以增强分离能力；

3) 业务解释需求：特别值得注意的是欺诈样本在 V17-V14 负值极端区域（V17 < -5, V14 < -10）的聚集现象，这暗示需结合领域知识探究其业务含义——可能对应特定隐蔽欺诈模式（如高频小额异常交易）。

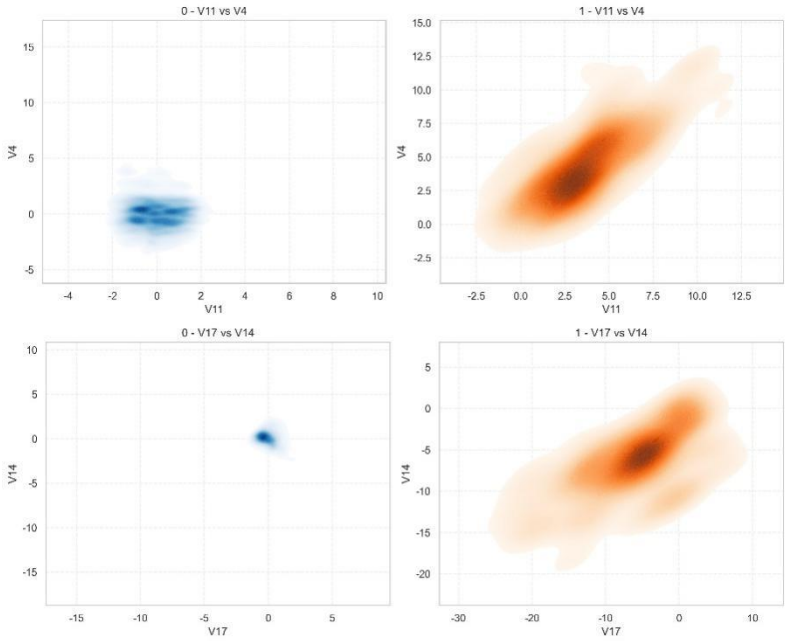

图 6 V11-V4 和 V17-V14 二维 KDE 图

3.4 数据预处理

基于 EDA 的观察，并考虑到特征工程不是本实验的重点，我们构建了一个简洁而关键的预处理流水线（Pipeline）。该流水线使用 scikit-learn 的 Pipeline 对象实现，包含一个步骤：对所有 30 个数值型特征（Time, Amount, V1-V28）应用 StandardScaler 进行标准化。此操作将所有特征缩放至均值为 0、方差为 1 的范围，其必要性体现在：1) 消除不同特征间的量纲差异，确保模型训练的稳定性；2) 对于后续将要使用的、基于距离计算的采样算法（如 KMeansSMOTE）以及带正则化项的逻辑回归模型，特征标准化是其发挥最佳性能的先决条件。

3.5 多阶段建模与评估流程

本实验核心是一个层层递进、逻辑严密的四阶段建模与评估流程，每个阶段都有明确的目标和方法。

阶段一：建立性能基准

目标是量化一个标准算法在未经过任何特殊处理的原始数据上的性能表现，为后续所有优化措施提供一个可供比较的参照点。选用 sklearn 封装的

LogisticRegression 作为基准模型，其分类阈值 threshold 为默认的 0.5。该模型在经过标准化的、但未进行任何数据采样的原始训练集上进行训练。训练完成后，在独立的测试集上进行评估，并生成完整的分类报告。此阶段的核心是记录欺诈类别（Class=1）的各项指标。最终在测试集上得到的基准指标如表 1。

表 1 基准模型的评估指标

accuracy_score	recall_score	precision_score	f1_score	roc_auc_score	weighted_score
0.999122	0.602041	0.842857	0.702381	0.800924	0.719959

从表中可以看出，预期得到一个具有迷惑性高准确率（>99%），但召回率只有 0.60 的结果，同时 F1-Score 仅为 0.70。此时的高精确率属于"虚假"，原因是当欺诈率 0.1%时，precision = TP/(TP+FP) = 0/(0+0)，sklearn 默认处理为 1.0，导致出现误导性结果。这个数据点同时直观地暴露不采取任何干预措施的严重后果，即超过 40%的欺诈行为无法被识别。

同时查看了 LogisticRegression 中各个特征的系数（如图 7），可以看出基本与前文探究的各特征与分类变量 Class 的相关性分布吻合。

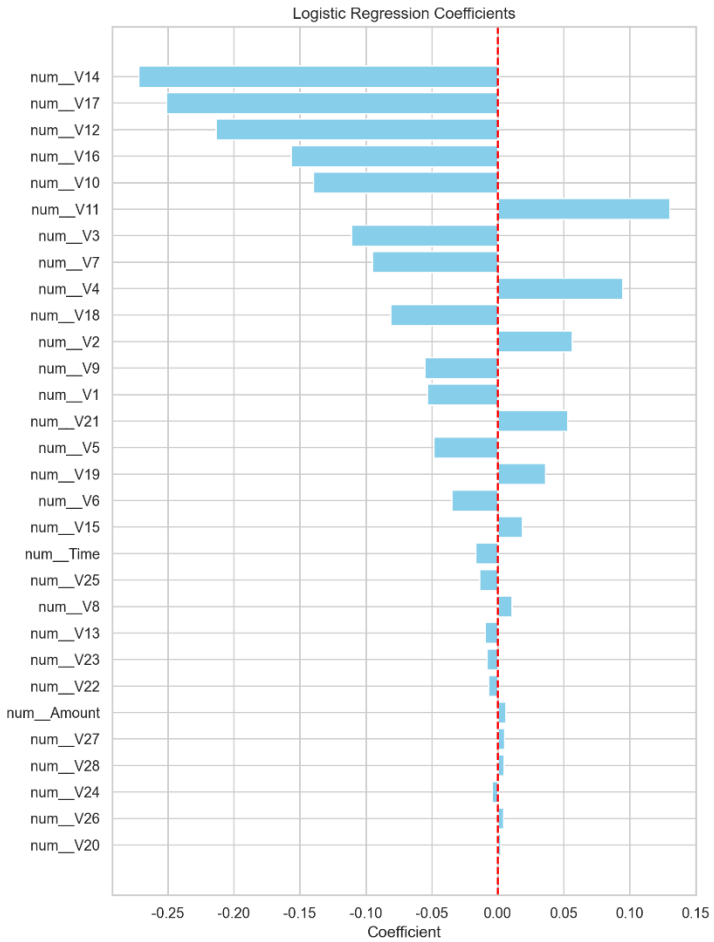

图 7 LogisticRegression 系数图

阶段二：采样策略的对比分析

在第一阶段暴露了数据不平衡的严重问题后，本阶段的目标是引入一种有效的数据层处理策略，以显著提升模型对少数类（欺诈交易）的识别能力，并分析其对模型性能各维度的具体影响。

此阶段的核心是验证采样策略的有效性，并寻找适用于本任务的最优采样策略。我们固定模型为第一阶段使用的 LogisticRegression 分类器，并采用四种采样策略，包括前文提到的 SMOTE、BorderlineSMOTE、KMeansSMOTE、ADASYN。然后构建一个包含三步的 imblearn.pipeline.Pipeline 流水线，分别为数据标准化(StandardScaler)、采样器、和 LogisticRegression 分类器。该流水线在原始训练集上进行训练，在训练过程中，数据首先被标准化，然后由预设的采样器对少数类样本进行采样，最后将处理后的平衡数据集传递给逻辑回归模型进行训练。

对不同过采样策略的评估结果如表 2 所示，揭示了各方法间显著的性能差异与内在的策略权衡。

一方面，传统的 SMOTE 及其变体 ADASYN 展示了过采样策略的典型困境，即尽管它们实现了召回率的极致提升（分别高达 0.928 和 0.938），但这伴随着精确率的灾难性下降，最低点（ADASYN）仅为 0.018。这导致其 F1 分数和综合加权分均远低于基准模型，说明这类“暴力”采样方法虽能强行识别出几乎所有欺诈样本，但代价是产生了海量的误报，导致模型失去了在真实业务中的可用性。

另一方面，KMeansSMOTE 策略的表现则尤为突出，成为了本阶段实验的决定性突破。它不仅成功将召回率从基准的 0.602 提升至 0.786，更难得的是，其精确率也同步提升至 0.855，打破了召回率与精确率之间此消彼长的常规困境。最终，KMeansSMOTE 在 F1 分数（0.819）和我们最核心的加权分数（0.833）上，均取得了所有策略中的最高分，显著超越了基准模型。

这一对比结果清晰地揭示了“智能”过采样的优越性，相较于在全局或边界区域盲目地合成样本，KMeansSMOTE 通过在安全的“数据簇”内部进行插值，有效避免了在类别边界引入噪声，从而实现了对模型决策边界的良性优化，而非破坏性干扰。因此，KMeansSMOTE 的成功不仅证明了数据层策略的有效性，更被确定为下一阶段实验的最佳数据层策略。

表 2 运用各采样方法后的评估结果

	accuracy_score	recall_score	precision_score	f1_score	roc_auc_score	weighted_score
SMOTE	0.974	0.928	0.059	0.110	0.951	0.596
BorderlineSMOTE	0.998	0.847	0.400	0.542	0.922	0.715
KMeansSMOTE	0.999	0.786	0.855	0.819	0.892	0.833
ADASYN	0.911	0.938	0.018	0.035	0.925	0.570

阶段三：超参数与阈值的联合寻优

鉴于前序阶段的分析，单纯调整单一环节（如仅引入采样）难以达到最优。本阶段的目标是尝试建立一套系统性的、可复用的协同优化方法论，不再将模型参数与分类阈值割裂看待，而是通过自动化工具，在一个统一的框架内寻找一个能使我们自定义的、更贴近业务目标的加权评估指标最大化的全局最优配置。

本方法论的核心是利用 Optuna 框架，为每种候选模型（如 Logistic Regression, Decision Tree, XGBoost）量身定制一个目标函数（objective function）。该函数将模型的核心超参数与分类阈值 threshold 共同视为一个待优化的、多维的参数空间。以 LogisticRegression 为例，具体说明此流程的实现。

1. 定义目标函数： 我们调用 tuning.py 中预先编写的 lr_thr_model_objective 函数。该函数接收 Optuna 在每次试验（trial）中“建议”的一组参数。
2. 统一参数空间： Optuna 的智能采样器（如 TPE）从我们定义的广阔搜索空间中进行探索。这个空间不仅包含了 LogisticRegression 的传统超参数（如正则化类型 penalty、求解器 solver、正则化强度 C、最大迭代次数 max_iter 等），还分类阈值 threshold（设定范围为 0.1 至 0.9）也作为一个独立的、可优化的维度囊括了进来。
3. 构建与评估流水线 Pipeline： 在目标函数内部，会使用这组“临时”参数动态构建一个完整的评估流水线，该流水线包含数据标准化、一种采样策略以及自定义的、应用了当前试验参数和阈值的 LR_Threshold 分类器。

4. 交叉验证： 为了避免在单一划分上的过拟合，将通过 5 折分层交叉验证（Stratified K-Fold）在训练集上进行评估。
5. 反馈与迭代： 交叉验证产生的性能指标（精确率、召回率等）被传入 weighted_score_auto 函数，计算出当前参数组合下的最终加权分数。这个分数作为本次试验的结果返回给 Optuna，为其后续的参数搜索提供方向。

整个过程通过指定的试验次数（例如 20 次）不断迭代，最终目标是找到能使加权分数最大化的那一组“黄金参数”。图 8 展示了 Optuna 优化的历史，图 9 展示了各参数的重要性，可以看出阈值 threshold 的重要性远远大于其他特征，这也正是模型性能大幅提升的所在。这套标准化的优化协议同样被应用于决策树（dt_thr_model_objective）和 XGBoost（xgb_thr_model_objective）模型，为下一阶段的最终对决输送了各自的最强版本。

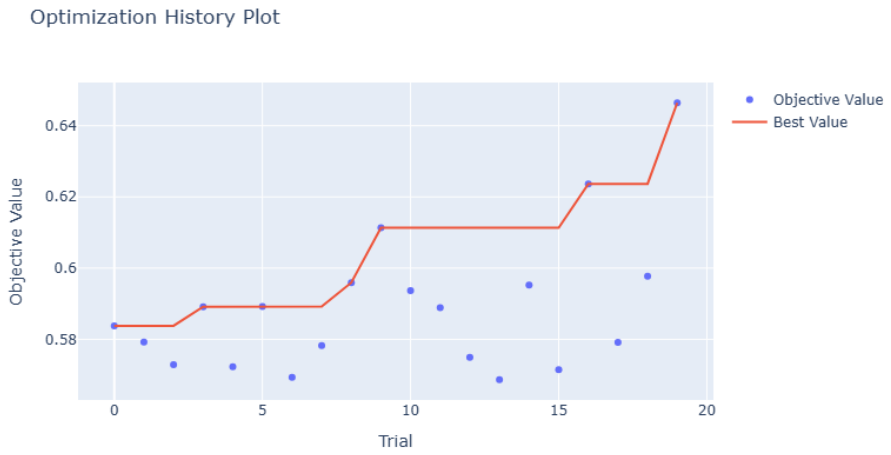

图 8 optuna 优化历史图

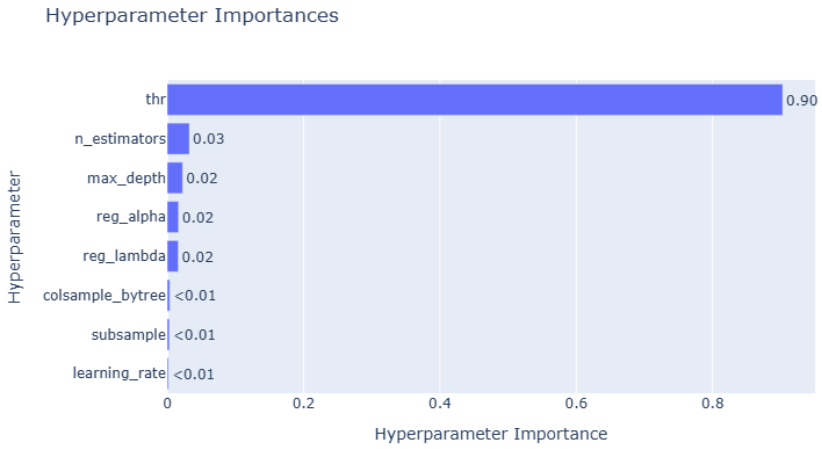

图 9 参数重要性图

值得强调的是，为了确保交叉验证评估的有效性和无偏性，所有的数据采样步骤（无论是过采样还是欠采样）都必须作为 Pipeline 的一部分，在交叉验证的每一折（fold）内部执行，图 10 直观展示了正确和错误方法的过程，左边是正确的处理方法，右边是错误的处理方法。 如果在交叉验证之前对整个训练集进行采样，那么由少数类生成的合成数据点（或因欠采样而保留的数据点）将不可避免地“泄露”到后续的验证集中，导致验证集不再是完全独立的，从而产生虚高的、过于乐观的性能评估结果。我们的 Pipeline 设计严格遵循了这一原则，确保了采样仅作用于每一折的训练部分，而验证部分始终保持原始状态，这是保证模型泛化能力评估准确性的关键。

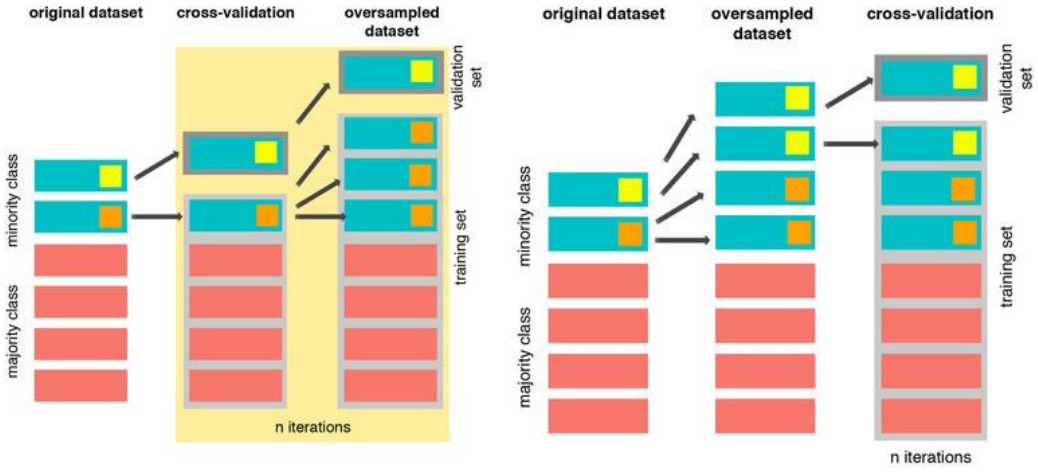

图 10 采样在交叉验证中的不同流程

阶段四：选择最佳适配模型

在上一阶段为每个模型架构都找到了最优配置后，本阶段在三个经过深度优化的模型重新在数据上进行最终验证，旨在寻找最佳的适配本任务的模型。

首先分别从 lr_thr_model_study、dt_thr_model_study 和 xgb_thr_model_study 这三个 Optuna 研究对象中，提取出在交叉验证中表现最佳的一组超参数（均包含各自的最优阈值）。然后使用这些“黄金参数”，分别配置出三个最终的、用于部署的 Pipeline。每个 Pipeline 均包含数据预处理、最适合该模型的采样策略（在最终模型构建时我们统一选用 KMeansSMOTE 以追求更高性能）以及应用了最优参数和阈值的定制化分类器（LR_Threshold, DT_Threshold, XGB_Threshold）。最后将这三个最终流水线分别在划分后 的原始训练数据上进行最后一次训练，使其充分学习数据中的模式。训练完成后，我们将三个模型应用于测试集上，生成其最终的性能报告。

最终的各模型在测试集上的性能表现如表 3 所示。从表中可以清晰地看出，虽然所有经过优化的模型相较于基准模型均有显著提升，但它们的侧重点和最终效果存在差异。经过协同优化的 XGBoost 模型在各项指标上全面胜出，其 F1 分数预计达到 0.869，召回率为 0.776，精确率也高达 0.987。这也证明了 XGBoost 模型强大的非线性建模能力与 Optuna 高效的参数搜索策略相结合，是解决本项目不平衡分类问题的最佳方案。

表 3 各模型在 KMeansSMOTE 采样下的最佳性能表现

	accuracy_score	recall_score	precision_score	f1_score	roc_auc_score	weighted_score
LogisticRegression	0.999	0.745	0.890	0.811	0.872	0.819
DecisionTree	0.999	0.602	0.819	0.694	0.801	0.714
XGBoost	0.999	0.776	0.987	0.869	0.888	0.864

4. 结果分析与深度讨论

本实验的核心价值在于验证了一套从数据到算法、再到自动化优化的综合治理方案。

首先，基准模型验证了问题的严重性。一个未经任何处理的标准逻辑回归模型，其欺诈交易召回率仅为 0.60，这意味着高达 40%的风险被直接漏报，直观地暴露了在不平衡数据上直接建模的巨大风险 。

其次，数据层面的干预被证明是解决问题的有效基石。实验对比了多种过采样策略，结果明确指出，KMeansSMOTE 策略表现突出。相较于传统 SMOTE 等方法以精确率急剧下降换取召回率的“野蛮”提升，KMeansSMOTE 通过在安全的“数据簇”内生成高质量样本，在将召回率提升至 0.786 的同时，也将精确率提升至 0.855，实现了二者的平衡，获得了最高的综合评分。

最后，协同优化是实现性能飞跃的关键。本实验最大的创新点在于使用 Optuna 框架，将模型超参数与分类决策阈值（threshold）进行一体化、协同寻优。优化结果证明，决策阈值是影响最终性能最重要的参数，其重要性远超所有模型结构参数。通过将阈值从默认的 0.5 调整至数据驱动发现的最优值，模型识别欺

诈的潜力得以被挖掘。最终，KMeansSMOTE 与 XGBoost 的组合，在经过 Optuna 协同调优后，在测试集上取得了 0.776 的召回率和 0.869 的 F1 分数 。这一结果与基准模型相比有显著的提升，证明了这种将数据策略、算法模型、决策阈值视为一个整体进行全局优化的方法论，是解决此类问题的卓越方案。

5. 未来展望

尽管本方案效果显著，但仍有可拓展的空间。由于本实验数据经过 PCA 处理，特征缺乏业务可解释性，未来的研究应结合领域知识，构建可解释性强的特征，以满足金融监管要求。当然，欺诈手段会不断演化，导致数据分布随时间改变，实际部署中需建立持续监控与模型定期更新的机制。对于更大规模、更复杂的交易数据，可以探索使用深度学习模型（如自编码器或图神经网络）来自动学习特征表示，这可能会发现传统机器学习模型难以捕捉的深层次欺诈模式。

参考文献

[1] 林智勇,郝志峰,杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究,2008,25(2):332-336. DOI:10.3969/j.issn.1001-3695.2008.02.003.

[2] 翟云,杨炳儒,曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学,2010,37(10):27-32. DOI:10.3969/j.issn.1002-137X.2010.10.005.

[3] 严远亭, 朱原玮, 吴增宝, 等. 构造性覆盖算法的SMOTE过采样方法[J]. 计算机科学与探索, 2020, 14(6): 975-984.

[4] 孙丹, 等. 基于改进混合采样和XGBoost算法的信用卡欺诈检测方法[J]. 计算机与现代化, 2022(9): 111-118.