- **机器学习——2023 秋季学期**

UnbalancedDatasetImpact: A Comparative Analysis of Sampling Techniques in Credit Card Fraud Detection

——针对信用卡欺诈检测数据集的系列分析

课　　程: 机器学习-2023 年秋季课程

数据集选择: **Credit Card Fraud Detection**

课 程 教 师: **曹　　鹏**

班　　级: 人工智能 2101 班

学　　院: 计算机科学与工程学院

作　　者: **谢　　山**

学　　号: **20216404**

1. 摘要

本研究旨在通过应用 t-SNE 降维方法对信用卡欺诈检测数据进行预处理，深入探讨该处理对决策树、K 最近邻、逻辑回归和支持向量机等四种传统分类器的影响。在研究过程中，我们发现不同分类器对于不同降维维数的响应存在差异，其中决策树和 KNN 在处理极不均衡的数据集方面表现出色，优于其他两种分类器。此外，我们还对比了随机升采样、随机降采样、SMOTE、ClusterCentor、TomekLinks 和 SMOTETomekLinks 等六种采样算法对决策树和 KNN 分类器的影响。研究结果显示，在该数据集中，TomekLinks 采样的效果明显优于其他五种算法。这一研究提供了在应对极不均衡的信用卡欺诈检测数据时选择降维和采样方法的有益见解。

关键词：传统分类器、采样算法对比、不均衡数据集分析、降维

2. 引言

在当今信息时代，机器学习在处理各种应用中发挥着关键作用，尤其是在分类问题中。然而，面对现实中的不均衡数据集，传统分类器可能面临着严重的性能挑战。我们的研究旨在深入探究不均衡数据对传统分类器的影响，并比较不同采样算法的效果，以提高分类器在这一背景下的性能。

2.1. 研究背景

数据不均衡是指在数据集中各个类别的样本分布不平衡，这在实际问题中是非常常见的。例如，在信用卡欺诈检测中，正常交易相对于欺诈交易的比例可能极其不均衡。这种不均衡性可能导致传统分类器对多数类别过度拟合，而对少数类别的识别能力较弱。

2.2. 研究问题和目标

在这一背景下，我们的研究关注于以下几个问题：不均衡数据对经典分类器的性能产生何种影响？在不同采样算法的支持下，分类器的性能是否得以改善？通过实验，我们将选取信用卡欺诈数据集，并使用决策树、K-最近邻、支持向量机和逻辑回归这四个经典分类器，从而深入分析这些问题。

2.3. 数据集和分类器的选择

我们选择信用卡欺诈数据集作为我们实验的基准，因为它代表了一个典型的不均衡数据场景。同时，我们选取了决策树、K-最近邻、支持向量机和逻辑回归这四个分类器，以确保我们的研究具有一定的广泛性和可比性。

2.4. 采样算法的选取

采样算法主要分为三大类，包括过采样、降采样和混合采样。 在过采样方面，我们采用了随机过采样作为基准，并引入了 SMOTE（Synthetic Minority Over-sampling Technique）采样方法，以增强数据集的样本。 在降采样方面，同样以随机降采样为基准，进一步引入了 TomekLinks 和 Cluster Center 方法，以优化数据集的样本分布。在混合采样方面，我们采用了 SMOTE+TomekLinks 的组合方法，以综合利用两种采样技术的优势进行采样操作。

通过这个研究，我们期望能够为处理不均衡数据集的机器学习任务提供实用的指导，并为选择合适的采样算法提供有力的支持。这对于提高分类器在真实应用中的可靠性和效果具有重要意义。

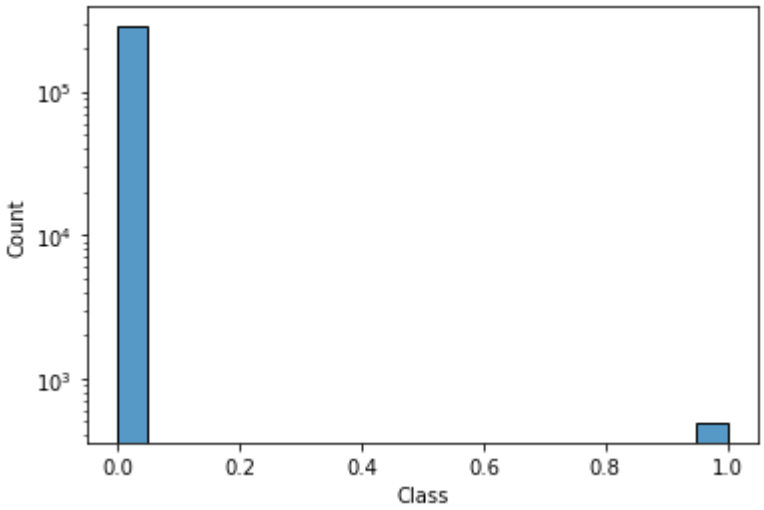
3. 实验和方法

3.1. 数据集

这一数据集涵盖了欧洲信用卡持卡人在 2013 年 9 月进行的交易记录，其中详细呈现了为期两天的交易情况。数据集中包含 492 笔欺诈交易，总共记录了 284,807 笔交易。值得注意的是，该数据集的正类别（欺诈交易）在所有交易中仅占 0.172%。

类别	数目	占比
正常交易	284315	99.827%
欺诈交易	492	0.172%

该数据集主要包含数值型输入变量，这些变量是通过主成分分析（PCA）进行转换得到的结果。为了保护用户隐私，数据集中的特征 V1、V2、……、V28 表示通过 PCA 获得的主成分，而未经 PCA 转换的特征则包括"Time"和"Amount"。其中，"Time"特征表示每笔交易与数据集中第一笔交易之间经过的秒数，而"Amount"特征则表示交易金额。最后，特征"Class"作为因变量，其取值为 1 表示欺诈交易，取值为 0 表示非欺诈交易。



3.2. 评价指标

鉴于我们的数据集呈极不均衡状态，其中类别分布存在显著偏斜，传统的准确率和混淆矩阵等指标在此情境下显得不够敏感。因此，我们决定采用 AUPRC（Precision-Recall Curve 下的面积）作为主要评价指标。AUPRC 在考虑精确率和召回率之间的平衡时表现出色。Precision-Recall 曲线能够全面展示模型在不同阈值下的性能表现，而 AUPRC 则通过综合整个曲线下的面积，更为细致地反映了模型在处理不平衡类别时的性能敏感性。这一选择有助于更准确地评估模型对于罕见事件（如欺诈）的有效性。

3.3. 实验方法

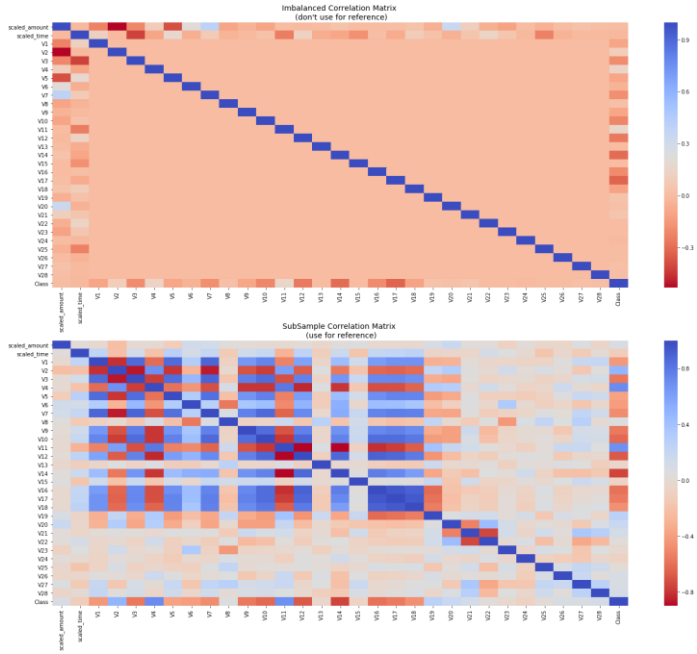
a) 数据预处理

i. Scale

在原始数据中，‘Time’ 和 ‘Amount’ 两个特征的尺度和其余 28 个特征不一致。因为这两个特征是真实的数据，而其余 28 个是真实数据经过主成分析得到的。因此我们对 Time 和 Amount 也进行缩放。

ii. 相关矩阵分析

为了更好的探究该数据集对于传统分类器的影响，我们进一步研究 30 个特征与类别的相关性。对以 50%正类别和 50%负类别降采样后的数据进行相关性分析。



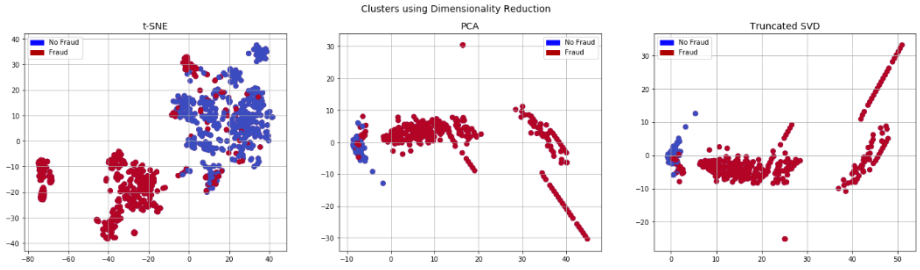
从相关性矩阵热力图可以得出以下结论：

- 负相关关系：V17、V14、V12 和 V10 呈负相关。请注意，这些数值越低，最终结果更有可能是欺诈交易。
- 正相关关系：V2、V4、V11 和 V19 呈正相关。请注意，这些数值越高，最终结果更有可能是欺诈交易。

iii. 降维和聚类

由于原始数据的特征维度高达 30 维，为避免维度灾难，即高维数据所带来的问题，我们采用了降维和聚类的方法。在这一步骤中，我们选择了三种经典的降维方法，分别为 t-SNE、PCA 和 TruncatedSVD。

- t-SNE（t-distributed Stochastic Neighbor Embedding）： t-SNE 是一种非线性降维技术，能够在降低维度的同时保留数据的局部相似性结构。我们对降采样后的数据应用了 t-SNE，并通过可视化手段展示了结果。
- PCA（Principal Component Analysis）： PCA 是一种线性降维方法，通过找到数据中的主成分来减小维度。我们同样采用 PCA 对降采样后的数据进行降维，并进行可视化。
- TruncatedSVD（Truncated Singular Value Decomposition）：TruncatedSVD 是一种主要用于处理稀疏矩阵的降维方法，通常应用于文本挖掘等领域。我们选择了 TruncatedSVD 对数据进行降维，并通过可视化呈现了降维后的结果。



显而易见，在将数据降维到两维的情况下，t-SNE 方法呈现更高的样本分离度。因此，在后续的工作中，我们决定采用 t-SNE 作为主要的降维方法。这一选择基于 t-SNE 在保持数据局部相似性和提高样本可分性方面的优越性能，为进一步的分析和建模提供了更有前景的基础。

b) 超参数实验以及交叉验证

i. 最优参数选择

由于每种分类器有不同的超参数，为了达到发挥出分类器最好的性能，我们采用 GridSearchCV 的方法搜索每个分类器最适合该数据集的参数：

Logistic Regression: *{"penalty": ['l1', 'l2'], 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000]}*

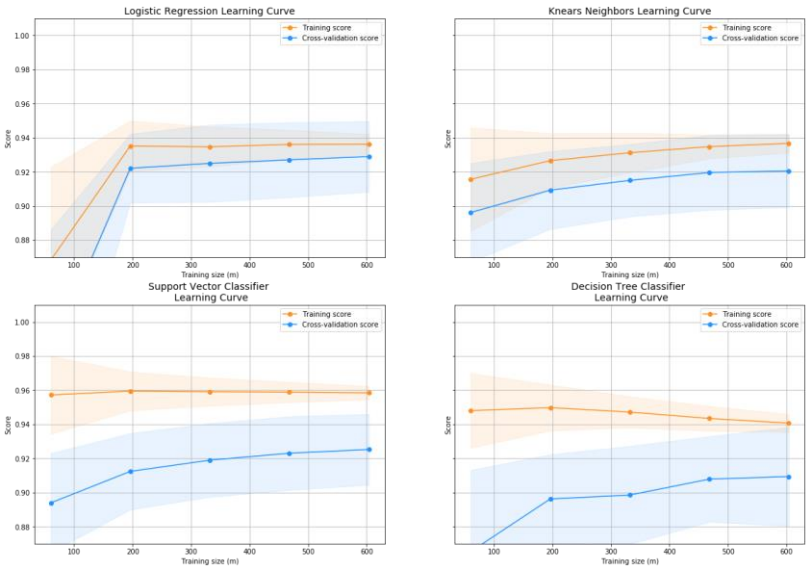
KNearestNeighbor: *{"n_neighbors": list(range(2,5,1)), 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']}*

SuportVectorMechine: *{'C': [0.5, 0.7, 0.9, 1], 'kernel': ['rbf', 'poly', 'sigmoid', 'linear']}*

DecisionTree: *{"criterion": ["gini", "entropy"], "max_depth": list(range(2,4,1)), "min_samples_leaf": list(range(5,7,1))}*

ii. 五折交叉验证

为了减小过拟合的风险，得到更稳健的性能估计，我们增加五折交叉验证获得最稳定的评估参数。



c) 原始数据训练传统分类器

找出四个传统分类器适应该数据集的最佳参数后，我们将其分别在原始数据集上训练，探究不均衡数据集对传统分类器的影响。结果见第四部分。

d) 不同采样方法

在对先前实验数据的详细分析后，我们决定选择在该数据集上表现最为代表性的 KNN 以及 Decision Tree 分类器作为后续采用各种采样方法的基准模型进行实验。

i. 过采样

- **随机过采样：** 通过增加少数类样本的复制来平衡数据集。
- **SMOTE（Synthetic Minority Over-sampling Technique）：** 利用插值方法生成合成的少数类样本，以增加数据集的平衡性。

ii. 欠采样

- **随机欠采样：** 随机减少多数类样本，以平衡数据集。
- **Tomek links：** 通过删除相邻不同类别的样本对，减少多数类样本。
- **Cluster Centroids：** 使用聚类方法对多数类样本进行聚类，并用聚类中心替代多数类样本。

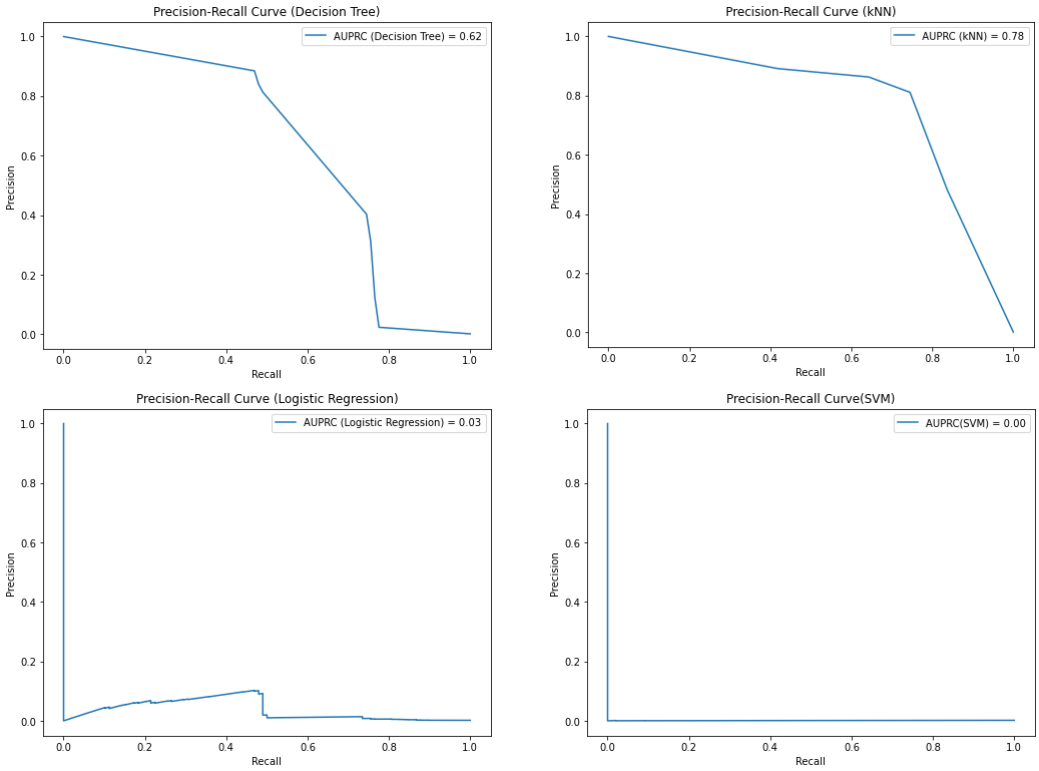
iii. 混合采样

- **SMOTE+Tomek links：** 将 SMOTE 过采样和 Tomek links 欠采样两者结合，以综合考虑过采样和欠采样的优势。

通过对这些采样方法的综合实验，我们旨在找到对于 KNN 以及 Decision Tree 分类器性能优化效果最显著的采样策略，以提高模型对不平衡数据的泛化能力。

4. 实验结果与分析

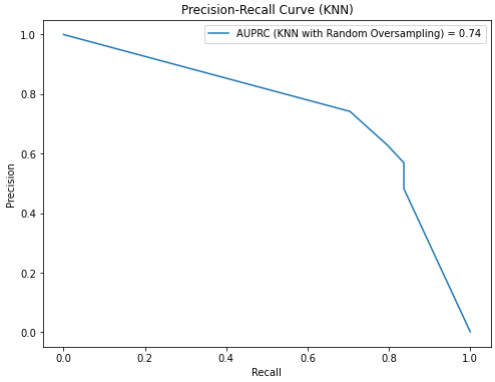
a) 直接处理原始数据集

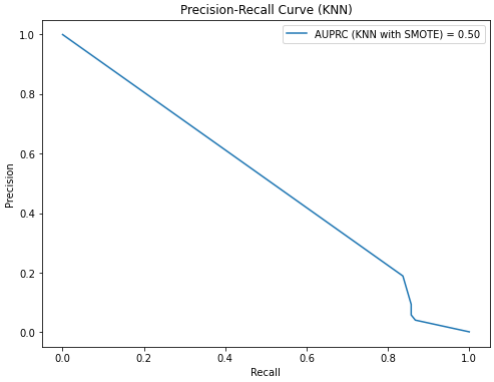


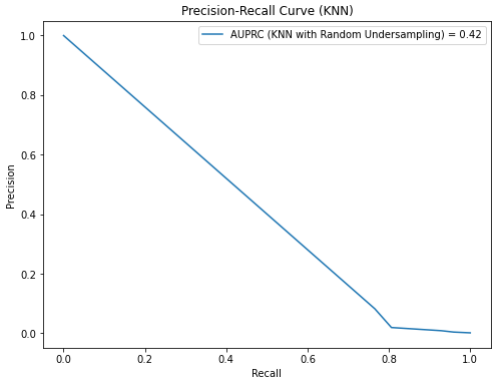
在实验结果中，决策树 (DT)、K 最近邻方法 (KNN)、逻辑回归 (LR)、支持向量机 (SVM) 分别展现了不同的 AUPRC 分数，分别为 0.62、0.78、0.03 和 0.00。通过对原始数据进行 t-SNE 降维，使得数据集的维度从 284807x30 减少到 284807x2。这样的数据集在逻辑回归和支持向量机中难以提取有效特征，因此这两种方法几乎无法准确判别欺诈交易。

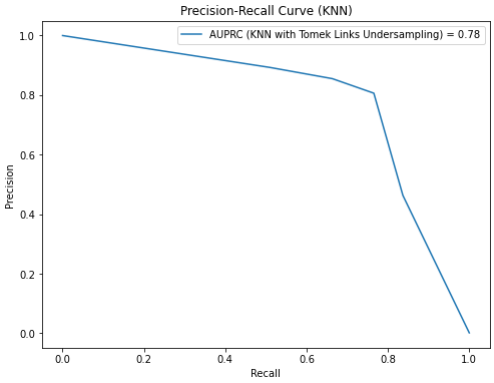
然而，决策树和 K 最近邻方法对于降维操作没有表现出负面效果。相反，降维后的数据对于 K 最近邻方法而言更加有效。这突显了在降维后，某些分类器的性能受到较小影响，而其他分类器可能因为失去了原始数据的重要信息而表现不佳。

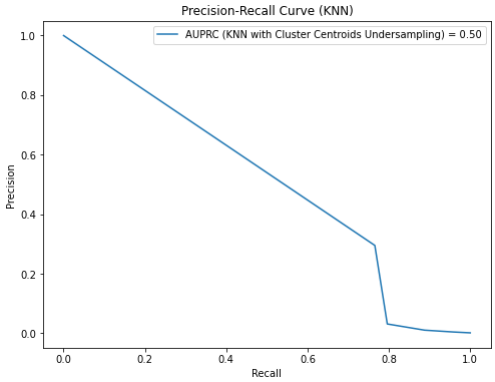
b) KNN 在六种采样策略下的数据

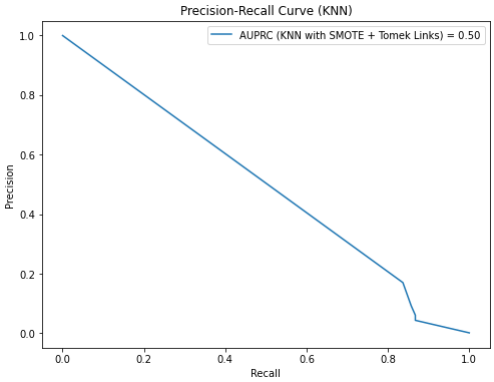




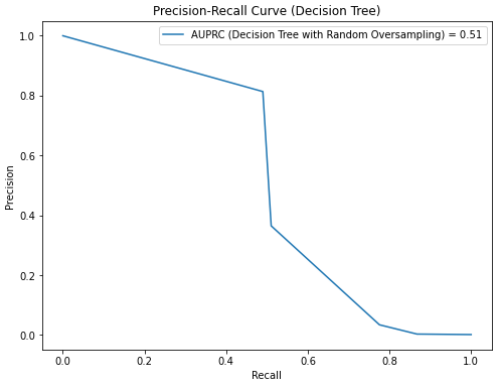


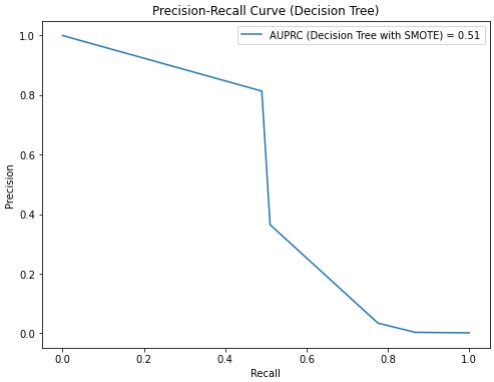


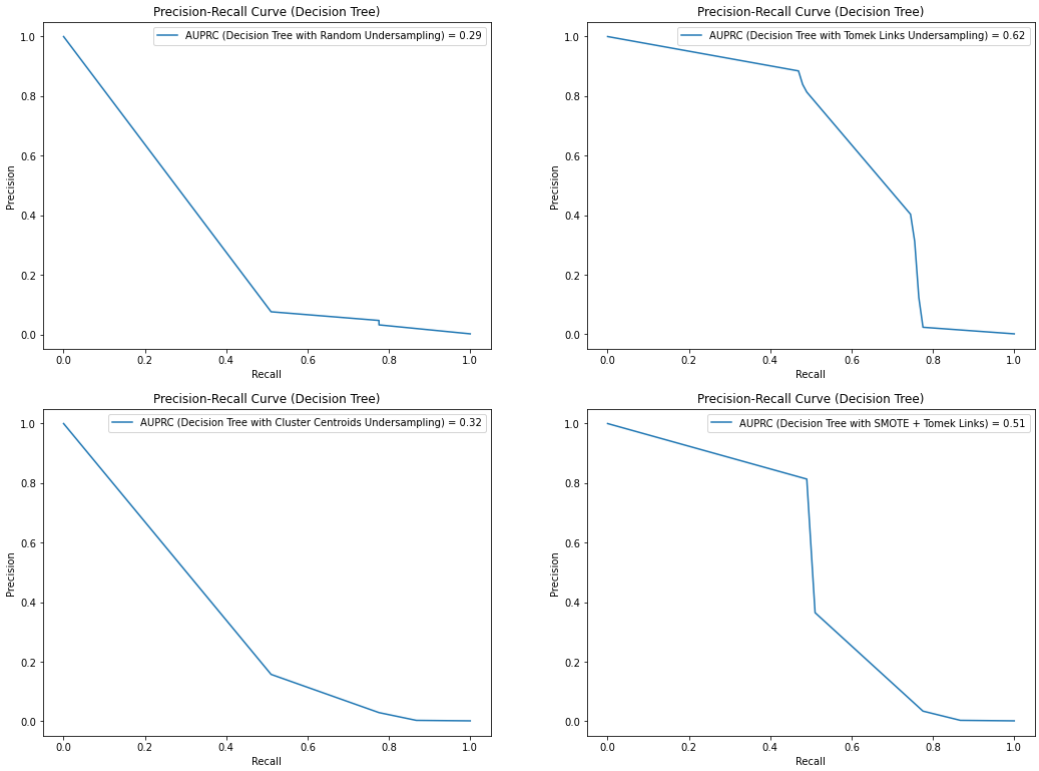




c) Decision Tree 在六种采样策略下的数据







d) KNN 与 Decision Tree 在不同采样算法下的对比

KNN

Origin	*0.78*
Random Oversampling	0.74
SMOTE	0.50
Random Undersampling	0.41
Tomek Links Undersampling	0.78
Cluster Centroids Undersampling	0.50
SMOTE + Tomek Links	0.49

Decision Tree

Origin	*0.62*
Random Oversampling	0.51
SMOTE	0.51
Random Undersampling	0.29
Tomek Links Undersampling	0.62

Cluster Centroids Undersampling	0.32
SMOTE + Tomek Links	0.51

从上表的结果可以看出，六种不同的采样算法对于分类情况的都没有很好的优化效果，只有使用 Tomek Links 降采样能达到原始数据的精度。出现这样的情况有以下几个原因：

- 数据集过于不均衡，在两种升采样的算法处理过后的数据量高达 50 多万条，而这其中 50%的正常交易类别，剩下 50%的欺诈交易类别只能从 492 条通过升采样得到。显然，里面出现了很多重复数据，即使是 SMOTE 方法在这样的数据面前也会出现重复数据，分类器对于这些重复数据学习不到有效内容。
- 三种降采样算法导致大量信息的缺失。
- 在三种降采样算法的对比中，我们可以看到 TomekLinks 方法的效果最优。这可能是因为在降维之后的数据集中，TomekLinks 可以准确的找到删除大部分重复近似的数据，减少了大量冗余。

5. 结论

在经过了以上实验之后，我们得出了以下结论：

- 在处理极度不均衡的数据集时，选择适当的传统分类器至关重要。在这一数据集中，部分传统分类器，如 SVM 和 LR，基本上无法通过训练过程学习到有效的信息。
- 在处理及其不均衡的数据集时，简单的使用 imbalance 库中的采样算法无法达到有效的采样效果。
- 相对于升采样的各种算法，当处理某一类别具有大量相似数据的情况时，通过降维后增强数据可分性，使用 TomekLinks 能够有效减少数据冗余。

6. 实验细节

参照 github 仓库: https://github.com/ShAn3003/Imblanced-Dataset

7. 参考文献

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud