Operations Research and Fuzziology 运筹与模糊学, 2024, 14(3), 843-851
Published Online June 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.143320

Hans汉斯

基于改进的Tomek Link方法预测电信客户流失

郑诗滢

福建师范大学数学与统计学院，福建 福州

收稿日期：2024年1月8日；录用日期：2024年6月20日；发布日期：2024年6月28日

摘　要

为提高电信客户流失检测的准确性和效率，提出了一综合方法。首先应用XGB-RFE方法对特征进行筛选，以选择最相关的特征，其次采用基于分层交叉验证框架下的数据平衡技术，来处理不平衡数据。通过这两种方法的综合应用，旨在提高模型性能和可解释性。结果表明，基于分层交叉验证框架下的Tomek Link欠采样技术，显著提高了各个模型的性能。此外，将该方法应用于TabNet模型中，同样取得了良好的效果。这一综合方法对于电信客户流失预测具有实际应用价值，有望提高流失检测率，改善业务决策。

关键词

Tomek Link，Logistic Regression，XGB-RFE，TabNet，客户流失

Predicting Telecom Customer Churn Based on Improved Tomek Link Method

Shiying Zheng

School of Mathematics and Statistics, Fujian Normal University, Fuzhou Fujian

Received: Jan. 8th, 2024; accepted: Jun. 20th, 2024; published: Jun. 28th, 2024

Abstract

In order to address the issue of low detection rates in telecom customer churn, a comprehensive approach is proposed. Firstly, the XGB-RFE method is applied for feature selection to choose the most relevant features. Secondly, a data balancing technique is employed under a stratified cross-validation framework to address the issue of class imbalance. Through the combined application of these two methods, the aim is to enhance model performance and interpretability. The results demonstrate that the Tomek Link under sampling technique under the stratified cross-validation framework significantly improves the performance of various models. Additionally, successfully applying this

文章引用: 郑诗滢. 基于改进的 Tomek Link 方法预测电信客户流失[J]. 运筹与模糊学, 2024, 14(3): 843-851.
DOI: 10.12677/orf.2024.143320

method to the TabNet model also yields favorable results. Therefore, this comprehensive approach holds practical value for predicting telecom customer churn, with the potential to improve churn detection rates and enhance business decision-making.

Keywords

Tomek Link, Logistic Regression, XGB-RFE, TabNet, Customer Churn

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着社会竞争日益激烈，尤其是移动技术行业的迅猛发展，客户拥有越来越广泛的选择，导致客户流失的潜在风险增加，同时留住现有客户的成本也在不断上升[1]。然而，与获取新客户相比，维护老客户的成本相对较低，因此近年来各行业都开始重视客户流失问题的分析[2] [3]。

近年来，学者们对电信客户流失问题进行了多方面的研究。例如，周颖等[4]采用 K-Means 聚类方法对客户进行细分，并建立了相应的聚类模型。张等[5]将深度学习和生存分析理论相融合，利用深度学习模型对电信客户流失数据进行模型构建。他们根据客户的生存状态与观察时间，进行客户流失风险分析。邱一卉[6]提供了剪枝随机森林的新方法，该模型具备高度准确的预测性能、较小数量的组合分类器以及更高的计算效率。

并且，鉴于这类数据通常呈现不平衡分布，少数类样本的研究价值较高，需要关注少数类样本的流失原因并采取针对性的补救措施。然而，由于数据的不平衡性质，传统的分类方法更注重对多数类样本的预测，这可能导致准确性不足。为了解决这个问题，国内外研究者提出了多种方法[7]。

一种常见的方法是对不平衡数据本身进行处理，其中 Tomek Link 方法是通过移除位于类别边界附近的重叠样本，以提升模型的分类性能。这种方法能够有效地减小数据集的规模，从而改善类别不平衡问题。本文对传统的 Tomek Link 方法进行了改进。提出了一种基于分层交叉验证框架的 Tomek Link 欠采样方法，以解决数据泄露问题。这种训练模型方式更为可靠，能够提高客户流失预测的效果。

2. 相关技术原理

2.1. XGBoost-RFE 方法

XGBoost 模型是陈天奇博士[8]提出的一种集成学习算法，该模型将多棵分类决策树组合在一起，具有在训练完成后提供特征重要性排序信息的能力。本文将 XGBoost 模型与递归特征消除(Recursive Feature Elimination, RFE)的思想相结合，采用 XGB-RFE 方法对特征进行筛选。这种方法的目标是通过多次训练 XGBoost 模型，并利用 XGBoost 特征重要性度量中的 F-score 方法对特征进行排名，来剔除最不重要的特征来选择最相关的特征，以提高模型性能和减少模型的复杂性。

步骤一：将数据集分割成训练集、验证集和测试集。

步骤二：使用包含所有特征的训练集训练 XGBoost 模型，采用五折交叉验证来寻找最优参数配置。

步骤三：使用训练好的模型计算特征的 F-score，用于对特征进行排序，识别哪些特征对目标变量的影响较大。

步骤四：经过多次试验，我们将阈值 h 的取值定为 0.2，然后根据这一阈值来判断是否保留刚刚删除的特征。如果在验证集上的 AUC 值低于 0.2，我们将删除相应的特征。

步骤五：通过反复进行步骤四，权衡特征数量和模型效果，筛选出对客户流失预测有重要影响的特征，并降低特征之间的冗余。

2.2. Logistic Regression 模型

逻辑回归一般用来预测类别，它模型简单且可解释性强，主要用于二分类问题。假设输入特征为 x，待预测的二元分类结果为 y。首先，对于每个样本 i，计算线性组合：

$$z = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

其中，b_0 是截距(bias)，$b_1, b_2, \cdots, b_k$ 是特征权重(Coefficients)，$x_1, x_2, \cdots, x_k$ 是输入特征。

接下来，将线性组合 z 应用于 Logistic 函数(也称为 Sigmoid 函数)，获得一个在 0 到 1 之间的概率值，表示样本属于类别 1 的概率。Logistic 函数的数学表达式如下：

$$P\left(y=1|x\right) = \frac{1}{1+\mathrm{e}^{-x}}$$

最后，我们可以根据得到的概率值来进行分类决策。示性函数是一种数学函数，通常用来表示某个条件是否为真，在表达式中，通常使用符号“I”来代表示性函数。因此我们将分类结果用 $y = I\left\{P\left(y|x\right) \geq 0.5\right\}$ 来表示，它在给定输入 x 的情况下，根据条件概率 $P\left(y|x\right)$ 的值来进行分类决策。具体而言，如果条件概率 $P\left(y|x\right)$ 大于或等于 0.5，那么示性函数的结果为 1。如果条件概率小于 0.5，示性函数的结果为 0。

2.3. 分层 k 折交叉验证

k 折交叉验证是一种常见的模型评估方法，它将数据集分成 k 个子集，然后依次选择其中一个子集作为验证集，将剩余的 $k-1$ 个子集合并作为训练集来进行模型评估。然而，标准的交叉验证可能存在一个问题，即随机划分可能导致某些折内只包含某一类别的数据，从而影响模型的训练。为了解决这个问题，分层 k 折交叉验证采用分层抽样的方式，考虑了各个类别的样本比例，以确保每个折中都包含各类别的样本，从而更好地保留了数据的分布特点，提高了模型评估的可靠性。这种方法有助于防止模型在训练过程中学到错误的分布，同时减少了不平衡数据导致的过拟合问题，进一步提高了测试精度。

由此本文通过分层 k 折交叉验证，将训练集(x_train 和 y_train)划分为 k 个折叠，确保每个折叠中的类别分布与整个训练集相似。并提出一种基于分层 k 折交叉验证的框架下使用 Tomek Link 技术的方法，旨在应对类别不平衡问题。该方法的主要目标是确保在处理不平衡的数据集时，模型能够公平评估并提高性能，从而防止过拟合。

2.4. SMOTE 过采样

SMOTE 过采样是一种基于特征空间的合成少数类样本的过采样方法，旨在应对不平衡数据集问题。它通过使用线性插值的方法来合成新的少数类样本，从而增加数据集中少数类的样本数量，以降低不平衡性。样本合成方式如下：

步骤一：对于每个少数类样本，计算它与其 k 个最近邻样本之间的差值。这将生成一个向量，表示在特征空间中少数类样本与其邻居之间的距离和方向。

步骤二：随机选择一个介于 0 到 1 之间的随机数。

步骤三：将随机数乘以上述生成的向量，并将结果添加到所考虑的特征向量(少数类样本)上。这相当

于在两个特征向量之间的线段上随机选择一个点，然后将该点作为新的合成样本。

2.5. Tomek Link 欠采样

最简单的过采样的算法思想是有放回的对正例样本随机抽取采样复制，使正反例样本数目达到相对平衡的目的。但是该算法的缺点是容易造成数据冗余，进而产生过拟合现象。在经过采样算法得到的平衡数据集，虽然可以使正反例样本分布平衡但同时也会造成正例样本数据空间扩展到反例样本数据空间，可能出现本来属于反例样本的空间被正例样本入侵的现象。可以用删除 Tomek Link 对的途径清除这些边界点。

Tomek Link 是一项关键的数据清理技术，用于处理由于过采样而引入的噪声数据和重叠样本[9]。对于数据集中的每个样本 x，计算它与其他样本 y 之间的欧氏距离 dist (x, y)。如果不存在第三个样本点 z，满足 dist (x, z) < dist (x, y)并且 dist (y, z) < dist (x, y) (即样本点 x, y 是最近邻)，并且 x 和 y 属于两个不同的类别，那么这两个样本点被称为“Tomek 链接对”。传统的 Tomek Link 欠采样方法通常会排除大类别中的样本，以清除这些链接对，从而减少噪声数据的影响，或者处理位于类别边界上的样本。

2.6. 分层 k 折交叉验证——Tomek Link 欠采样

分层交叉验证可以分为以下几个步骤：

步骤一：将原始训练集划分为 5 个折叠，其中一个作为验证集，其余作为训练集，以评估模型性能，并减少数据不平衡引起的偏差。能够考虑数据中不同类别的分布情况。它确保每个折叠中的类别分布与整个数据集中的类别分布大致相同。

步骤二：在每个交叉验证迭代中，并在每个折叠上使用 Tomek Link 欠采样技术进行合成样本，以解决类别不平衡的问题。

步骤三：通过循环遍历每个交叉验证折叠来进行模型训练。

步骤四：使用机器学习模型对验证折叠进行预测，并计算其各项指标，这些指标用于衡量模型在不同数据子集上的性能。

步骤五：分析性能指标的平均值，以了解模型的整体性能和稳定性。

步骤六：在完成所有交叉验证折叠的训练、验证、测试后，选择性能最佳的模型作为最终模型，以用于后续的实际应用。

3. 数据处理

3.1. 数据来源

Table 1. Description of data attribute features
表 1. 数据的属性特征描述

	属性名称	属性说明
客户基本信息	CustomerID	客户 ID
	Count	数量
	Country	国家
	State	州
	City	城市

续表

客户基本信息	Zip Code	邮政编码
	Lat Long	经纬度
	Latitude	纬度
	Longitude	经度
	Gender	性别
	Senior Citizen	是否为老年人
	Partner	有无伴侣
	Dependents	有无家属
客户开通的各项服务	Phone Service	有无电话服务
	Multiple Lines	是否多线路
	Internet Service	上网方式
	Online Security	上网安全
	Online Backup	在线备份
	Device Protection	设备保护
	Tech Support	技术支持
	Streaming TV	流媒体电视
	Streaming Movies	流媒体电影
客户账单信息	Contract	合同期限
	Paperless Billing	是否无纸化支付
	Payment Method	付款方式
	Monthly Charges	月租
	Total Charges	总费用
	Tenure Months	留存期
流失相关	Churn Label	流失标签
	Churn Value	流失值
	Churn Score	流失分数
	CLTV	客户生命周期价值
	Churn Reason	流失原因

文中使用来源于 Kaggle 网站的预测电信客户流失情况数据集[10]，该数据集共有 7043 条数据、33 个属性特征，其中属性特征可以分为 4 类，分别为客户基本信息、客户开通的各项服务、客户的账单信息和流失相关。数据集的属性特征描述，如表 1 所示。

3.2. 数据预处理

经过仔细观察，发现“Total Charges”列中存在 11 个空白字符串，而这 11 条数据对应的“Tenure Months”列的数值为 0，这意味着这 11 条数据属于新用户，他们还没有产生费用。为了更正这个问题，可以将这 11 条数据的“Total Charges”和“Monthly Charges”设置为 0，表示这些新用户的费用为 0。这一步骤可以确保数据的一致性和准确性，以便后续分析和建模。

在机器学习领域，对于分类变量通常需要采取一定的方法将其转换成数值型数据。本文选择使用标签编码对分类变量进行编码，并对连续型变量采用标准化处理方式。这种方法的优势在于它将分类变量映射为数值，减少了数据维度。对于连续型变量，标准化处理可以确保它们具有相似的尺度，有助于提高模型的性能和可解释性。

3.3. 特征选择

数据预处理后，将数据集以 8:2 的比例随机分成训练数据集和测试数据集。训练数据集将用于 XGBoost 分类器的学习和训练，通过不断调整模型参数来提高预测精度，最终确定最佳模型参数。接着，使用 XGBoost 分类器和递归特征消除(RFECV)的方法，筛选出最重要的 23 个特征，以提高模型的性能和可解释性。

4. 实验分析

4.1. 评估标准

分类常用指标为准确率，但对于不平衡样本机器学习大多数算法为提升准确率，分类结果偏向多数类，本文所选客户流失数据中多数类为未流失客户，这样会出现有更多流失客户被划分为未流失客户，对于电信公司来说有更大损失。因此本文引入精确率、召回率、F1、AUC 值。

4.2. 模型评估

4.2.1. 模型效果比较

采用 SMOTE 过采样方法，平衡训练集中的类别分布，然后使用机器学习方法进行模型训练，采取五折交叉验证进行效果评估。实验结果如下表 2：

Table 2. Effect of traditional SMOTE method
表 2. 传统 SMOTE 方法效果

Model	Precision	Accuracy	Recall	F1	AUC
XGBoost	0.7383	0.7962	0.7387	0.7385	0.7387
Random Forest	0.7505	0.8054	0.7605	0.7551	0.7605
Logistic Regression	0.7365	0.7763	0.7895	0.7464	0.7895
AdaBoost	0.7409	0.7876	0.7852	0.7530	0.7852
LightGBM	0.7439	0.8004	0.7476	0.7457	0.7476
CatBoost	0.7449	0.8011	0.7498	0.7473	0.7498
Gradient Boosting	0.7439	0.7983	0.7668	0.7532	0.7668

基于分层交叉验证的框架下使用 SMOTE 技术对模型进行训练，实验结果如下表 3：

Table 3. Effect of improved SMOTE method
表 3. 改进 SMOTE 方法效果

Model	Precision	Accuracy	Recall	F1	AUC
XGBoost	0.7296	0.7898	0.7207	0.7248	0.8494
Random Forest	0.7494	0.8047	0.7557	0.7524	0.8568

续表

Logistic Regression	0.7319	0.7713	0.7844	0.7412	0.8686
AdaBoost	0.7385	0.7898	0.7738	0.7503	0.8619
LightGBM	0.7483	0.8040	0.7466	0.7475	0.8564
CatBoost	0.7548	0.8089	0.7526	0.7537	0.8640
Gradient Boosting	0.7500	0.8040	0.7698	0.7583	0.8668

基于分层交叉验证的框架下使用 Tomek Link 技术对模型进行训练，实验结果如下表 4：

Table 4. Effect of improved Tomek Link method
表 4. 改进 Tomek Link 方法效果

Model	Precision	Accuracy	Recall	F1	AUC
XGBoost	0.7391	0.7969	0.7332	0.7360	0.8523
Random Forest	0.7621	0.8139	0.7500	0.7556	0.8604
Logistic Regression	0.7630	0.8153	0.7612	0.7621	0.8690
AdaBoost	0.7485	0.8040	0.7543	0.7513	0.8626
LightGBM	0.7318	0.7912	0.7294	0.7305	0.8583
CatBoost	0.7449	0.8011	0.7498	0.7473	0.7498
Gradient Boosting	0.7575	0.8111	0.7566	0.7570	0.8680

基于分层交叉验证的框架下使用 Smote Tomek 技术对模型进行训练，实验结果如下表 5：

Table 5. Effect of improved Smote Tomek method
表 5. 改进 Smote Tomek 方法效果

Model	Precision	Accuracy	Recall	F1	AUC
XGBoost	0.7353	0.7940	0.7270	0.7309	0.8477
Random Forest	0.7484	0.8040	0.7518	0.7501	0.8600
Logistic Regression	0.7326	0.7727	0.7845	0.7423	0.8687
AdaBoost	0.7398	0.7926	0.7706	0.7509	0.8661
LightGBM	0.7430	0.7997	0.7463	0.7446	0.8578
CatBoost	0.7521	0.8068	0.7468	0.7494	0.8630
Gradient Boosting	0.7472	0.8018	0.7658	0.7551	0.8640

从上表可以看出采用数据平衡方法后进行训练的模型性能没有基于分层交叉验证的框架下使用数据平衡方法进行训练的模型性能好。其中基于分层交叉验证的框架下使用 Tomek Link 欠采样技术对模型进行训练的效果最佳。这表明，使用分层交叉验证的框架可以更好地处理不平衡数据集，因为它确保了每个折叠中都包含了各个类别的样本，从而更全面地评估模型性能。此外，Tomek Link 在处理不平衡数据集时表现出色，有助于提高模型的性能。综合使用分层交叉验证和 Tomek Link 等技术，可以显著提升模型性能，使其更适应不平衡数据集，提高分类准确性和鲁棒性。

4.2.2. 三阶段 TabNet 模型

TabNet 是一种深度学习模型，专门设计用于处理表格数据，比如 CSV 文件或数据库表。该模型由 Google Cloud AI 团队[11]于 2019 年首次提出，并在多个表格数据任务中表现出竞争性的性能。TabNet 以其在特征选择、注意力机制和逐步决策方面的独特设计而广受赞誉，为处理复杂的结构化数据提供了一种强大而灵活的解决方案。本文应用了三阶段 TabNet 模型，充分发挥其在处理表格数据时的优越性能。

首先，对客户流失数据集进行预处理，然后应用 SMOTE 过采样算法。这一步骤的目的是通过生成流失客户的合成样本来平衡数据集，增加了流失客户的样本数量。这样做可以使训练集中不同类别之间的分布更均衡，使模型更专注于少数类样本。然而，需要注意的是，SMOTE 算法可能引入一些噪声。接下来，对预处理后的客户流失数据集应用 SMOTE Tomek 混合采样算法。这一步骤结合了 SMOTE 和 Tomek Link 算法，通过同时增加流失客户样本和删除留存客户样本来更好地处理类别不平衡问题。这有助于提高模型的性能，同时减少噪声的影响。最后，对客户流失数据集应用 Tomek Link 欠采样算法，以获得分类边界更为明确的新数据集。这一步骤有助于进一步改善模型的性能，并提供更准确的分类结果。

整个过程分为三个阶段，每个阶段都使用 TabNet 模型进行训练。最终得到一个经过多次改进的模型，以更好地处理客户流失问题和不平衡的数据集。这种方法旨在提高模型的泛化能力，以更准确地预测客户流失情况。选取结合 Tomek Link 欠采样和 TabNet 的 TL-TabNet 算法、结合 SMOTE 过采样和 TabNet 的 SMOTE-TabNet 算法、结合 SMOTE Tomek 采样和 TabNet 的 ST-TabNet 算法作为对比算法，实验结果见表 6，可以看出 MultiS-TabNet 算法能有效地预测电信行业流失客户，MultiS-TabNet 算法的 AUC 值在此基础上提高了 0.5%。

Table 6. TabNet model performance
表 6. TabNet 模型效果

Model	Precision	Accuracy	Recall	F1	AUC
Smote-TabNet	0.7162	0.7315	0.7762	0.7104	0.8572
ST-TabNet	0.7183	0.7774	0.7326	0.7237	0.8368
TL-TabNet	0.7627	0.8118	0.7305	0.7434	0.8584
MultiS-TabNet	0.7320	0.7791	0.7752	0.7435	0.8634

5. 结语

通过对电信客户流失情况的数据集进行分析，研究了 33 个变量对电信客户流失(Churn)的预测，本文使用了多个不同的机器学习模型，包括 XGBoost、Random Forest、Logistic Regression、AdaBoost、LightGBM、CatBoost、Gradient Boosting 和 TabNet 模型，对预测结果进行了比较，并评估了这些模型在 Precision、Accuracy、Recall、F1 和 AUC 性能指标上的表现。通过对建模过程中平衡数据方法进行比较可以发现，本文提出的基于分层交叉验证的框架下使用数据平衡技术的方法，对各个模型的性能都具有一定的提升，尤其是在数据平衡方面使用了 Tomek Link 欠采样技术，取得了最佳的效果。此外，本文还将这一数据平衡方法成功应用于 TabNet 模型中，并取得了良好的效果。这表明该方法在不同类型的模型中都具有一定的实用价值，有助于提高预测电信客户流失的准确性和可靠性。

参考文献

[1] 周婉婷, 赵志杰, 刘阳, 等. 电子商务客户流失的 DBN 预测模型研究[J]. 计算机工程与应用, 2022, 58(11): 84-92.

[2] 李波, 谢玖祚. 生成对抗网络的银行不平衡客户流失预测研究[J]. 重庆理工大学学报: 自然科学, 2021, 35(8): 136-143.

[3] Wei, C.P. and Chiu, I.T. (2002) Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Systems with Applications*, **23**, 103-112. https://doi.org/10.1016/S0957-4174(02)00030-1

[4] 周颖, 吕巍, 井淼. 基于数据挖掘技术的移动通信行业客户细分[J]. 上海交通大学学报, 2007(7): 1142-1145.

[5] 张俊春, 王庶民, 徐峰. 深度生存分析在电信客户流失预测中的应用[J]. 价值工程, 2021(20): 165-167.

[6] 邱一卉. 基于剪枝随机森林的电信行业客户流失预测[J]. 厦门大学学报: 自然科学版, 2014, 53(6): 817-823.

[7] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019(4): 16.

[8] Chen, T., He, T. and Benesty, M. (2016) XGboost: Extreme Gradient Boosting. https://api.semanticscholar.org/CorpusID:225317050

[9] TANKY (2020) Telco Customer Churn: IBM Dataset. https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset/data

[10] Devi, D., Biswas, S.K. and Purkayastha, B. (2016) Redundancy-Driven Modified Tomek-Link Based Undersampling: A Solution to Class Imbalance. *Pattern Recognition Letters*, **93**, 3-12. https://doi.org/10.1016/j.patrec.2016.10.006

[11] Arik, S.O. and Pfister, T. (2019) TabNet: Attentive Interpretable Tabular Learning. arXiv preprint arXiv:1908.07442.