Computer Science and Application 计算机科学与应用, 2022, 12(9), 2084-2093
Published Online September 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.129211

Hans汉斯

基于代价敏感逻辑回归的电信客户流失预测研究

彭　科*，彭　龑

四川轻化工大学计算机科学与工程学院，四川　宜宾

收稿日期：2022年8月3日；录用日期：2022年9月1日；发布日期：2022年9月9日

摘　要

针对电信客户流失数据集的多维特征和不均衡问题，本文给出了一种基于代价敏感的逻辑回归的电信客户流失预测模型。通过对不平衡样本集分别采用不同权重调整，将代价敏感学习与传统分类算法相结合，建立基于逻辑回归的电信客户流失预测模型，最后对实际电信客户流失进行验证。通过与其他分类器模型的对比显示此方法在各种评估指标上均有更好的表现，更加符合电信业预测客户流失的实际情况。

关键词

代价敏感学习，逻辑回归算法，数据挖掘，客户流失，预测

Research on Telecom Customer Churn Prediction Based on Cost-Sensitive Logistic Regression

Ke Peng*, Yan Peng

School of Computer Science & Engineering, Sichuan University of Science & Engineering, Yibin Sichuan

Received: Aug. 3rd, 2022; accepted: Sep. 1st, 2022; published: Sep. 9th, 2022

Abstract

Aiming at the multi-dimensional characteristics and imbalance of the telecom customer churn dataset, this paper presents a cost-sensitive logistic regression-based prediction model for telecom customer churn. By adjusting the unbalanced sample sets with different weights, combining

*通讯作者。

文章引用: 彭科, 彭龑. 基于代价敏感逻辑回归的电信客户流失预测研究[J]. 计算机科学与应用, 2022, 12(9): 2084-2093. DOI: 10.12677/csa.2022.129211

cost-sensitive learning with traditional classification algorithms, a logistic regression-based telecommunication customer churn prediction model is established, and finally the actual telecommunication customer churn is verified. The comparison with other classifier models shows that this method has better performance in various evaluation indicators, which is more in line with the actual situation of predicting customer churn in the telecom industry.

Keywords

Cost-Sensitive Learning, Logistic Regression Algorithm, Data Mining, Customer Churn, Prediction

1. 引言

随着信息技术与移动网络的巨大进步，电信行业的竞争日趋严峻，与此同时，获得一个新用户所花费的成本是挽留一位老客户所花费成本的 5~6 倍[1]。因此，对客户流失的分析与预测已经成为电信行业提高核心竞争力的重要方式。为了准确预测流失用户，许多学者采用了传统机器学习的分类算法进行识别客户流失。但由于电信行业的数据正负样本的不平衡，这使得电信行业的流失预测不切实际，机器学习往往认定训练样本中各类样本数量都是均衡的，但在具体问题上却往往不能满足这种数据平衡的条件[2] [3]。数据不均衡会使机器学习训练模型关注于数量较多的样本类型，而忽视数量少的样本类型，从而降低了机器学习模型的测试泛化能力。例如，在训练集中，其中正常的样本有 99 个，负例的样本有 1 个。如果没有考虑到样本的不均衡，则该学习方法会导致分类器放弃对负例的预测，因为分类器将所有的样本分成正例，则可以达到 99%的分类准确率。为此，应该考虑不同的实例在分类器中的错误分类成本。本文介绍了一种代价敏感学习方法，用以对电信业的客户流失，使用欠采样对不同实例的错误分类成本进行修改，结合比较不同传统机器学习分类算法，得到一个符合电信业实际情况的预测模型[4] [5] [6]。

2. 电信客户流失预测原理

在传统的分类算法中，一般都会假定不同类型的样本数量趋于均衡，从而导致对大部分类别的样本进行预测，忽略了少数类别的样本。在电信用户的流失中，存在着严重的数据分配失衡现象[7]。若采用传统的数据挖掘方法，对所有的用户进行预测，其准确率也很高。从表面上来看，这是一种非常有效的方法，但是，当一个具有高价值的用户被认为是一种潜在的用户时，它就失去了其研究的意义。由此可得，在不均衡数据中，准确度并非是一个合适的衡量标准。在不平衡数据中，可以考虑使用混淆矩阵来评估分类器的性能。混淆矩阵包括四种样本类别，分别是真正例 TP、假负例 FN、假正例 FP 和真负例 TN。样本总数 $N = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}$ [8]。将预测结果与样本实际类别结合得到混淆矩阵，如表 1 所示。

Table 1. Confusion matrix
表 1. 混淆矩阵

	预测为正	预测为负
实际为正	TP	FN
实际为负	FP	TN

准确率(accuracy)是指分类正确的样本结果占总样本的百分比，定义为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{1}$$

精确率(precision)是指在预测结果中所有被预测为正的样本中实际也为正样本所占的百分比。召回率(recall)是指在原样本中分类器正确检测到的正样本占全部正样本的比例。两者的计算公式分别为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

一个分类器的精确率和召回率往往呈负相关关系，两者在某一分类阈值会达到平衡。F1-score 是召回率和精确率的调和平均，即二者之间的一个平衡点，可以通过评估 F1 分数来找到精确率和召回率的最佳组合。F1-score 公式如下：

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

ROC 曲线是以 FP/(FP + TN)为横轴，即以预测为正但实际为负的样本占所有负例样本的比例为横坐标，以 TP/(TP + FN)为纵轴，即以预测为正但实际为正的样本占所有正例样本的比例为纵坐标。图 1 则是一个标准的 ROC 曲线。AUC 是 ROC 曲线下的面积，反映的是根据分类器计算得到的 score 后对样本的排列顺序的概率。AUC 越大，则分类器的效果越好。

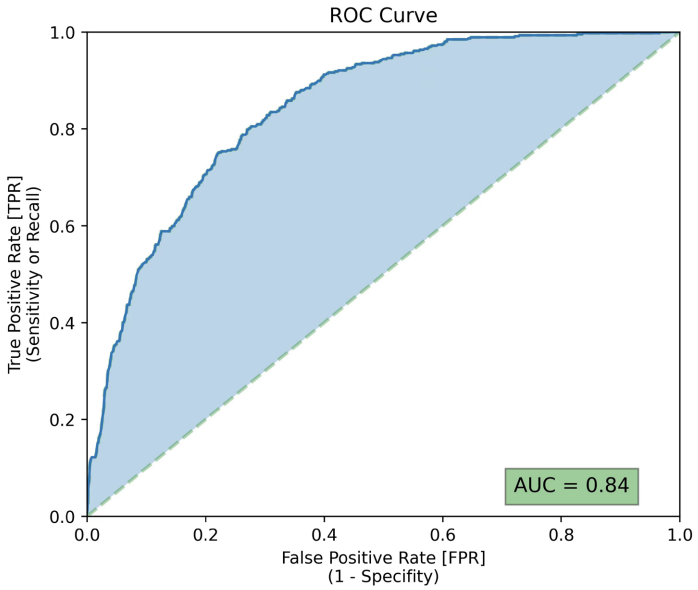

Figure 1. ROC curve
图 1. ROC 曲线

由于数据不均衡，因此评估电信用户流失的学习指标并没有采用 Accuracy 来进行评估，而是选取了聚焦于正例上的评估指标。此外基于代价敏感的客户流失，本文定义了三个额外的评估标准，分别叫做 Revenue Retained Monthly (RRM)，Revenue Retained Total (RRT)，Revenue Retained Monthly per true churn customer (RRM/customer)。其计算公式分别为：

$$\mathrm{RRM}=\frac{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times y_{pred}^{(i)}\times monthlycharges^{(i)}}{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times monthlycharges^{(i)}} \tag{5}$$

$$\mathrm{RRT}=\frac{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times y_{pred}^{(i)}\times totalcharges^{(i)}}{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times totalcharges^{(i)}} \tag{6}$$

这两个指标是针对电信客户样本中加权的月收入特征的召回率分数，可以被看作为等价的召回率。它们表示模型由于其正确的客户流失预测(即真阳性)而保留的收入。

$$\frac{\mathrm{RRM}}{\mathrm{customer}}=\frac{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times y_{pred}^{(i)}\times monthlycharges^{(i)}}{\sum_{i=1}^{n_{test}} y_{test}^{(i)}\times y_{pred}^{(i)}} \tag{7}$$

此指标具有指示分类器识别出最有价值的客户或高支出的客户的能力。值越高表示该分类器能够更擅长识别更多的有价值的客户。

3. 基于代价敏感逻辑回归的电信客户流失预测

3.1. Logistic 回归原理

Logistic 回归是一种减少预测范围，将预测值限定为[0, 1]间的处理二分类的回归模型，它通过一个 sigmoid 激活函数与线性回归得到，这个 sigmoid 函数的性质满足了函数的定义域 x 的输入为全体实数，而值域输出 $h_\theta(x)$ 总是[0, 1]，最终是以一种概率的形式表示，概率大于 0.5 的分类结果为正类，小于 0.5 的分类结果为负类。其中 sigmoid 函数具体表达形式如下：

$$g(z)=\frac{1}{1+\mathrm{e}^{-z}} \tag{8}$$

对于线性回归的情况，方程如下：

$$\theta_0+\theta_1x_1+\cdots+\theta_nx_n=\theta_0+\sum_{i=1}^{n}\theta_ix_i=\theta^{\mathrm{T}}x \tag{9}$$

因此，可以得到 Logistic 回归的具体输出为：

$$h_\theta(x)=g\left(\theta^{\mathrm{T}}x\right)=\frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}} \tag{10}$$

3.2. 代价敏感学习

针对不均衡的数据，采用过采样或欠采样方法，虽然可以在某种程度上解决数据不平衡问题，但其并没有考虑到对不同的分类错误，其代价也不相同的问题。由于误分类代价都是一样的且很容易出现过拟合问题，因此，分类代价不平衡的问题还没有得到解决。由于电信客户数据样本的特殊性，如果将一个有电信流失想法的高价值客户预测为仍在使用电信的用户，则损失了更多的代价，因此需关注模型对少数类样本的识别能力。基于代价敏感学习的方法，对不同类型的错误进行了不同的成本惩罚，从而降低了产生较高的误分率以及降低了总体的分类误差[9]。

代价敏感学习方法的错分代价一般是由代价矩阵确定。代价矩阵如表 2，其中 $cost_{ij}$ 表示把第 i 类数据预测为第 j 类数据的惩罚代价。通常初始惩罚代价 $cost_{ij}$ 为零，假如将第零类误分为第一类所导致的损失程度更大，则 $cost_{01}>cost_{10}$。损失程度差异越大，则 $cost_{01}$ 与 $cost_{10}$ 的值差别也越大。当 $cost_{01}$ 与 $cost_{10}$ 值相同时，则是属于代价不敏感的学习问题。

Table 2. Cost matrix for customer churn predictions
表 2. 客户流失预测的代价矩阵

	预测不流失	预测流失
实际不流失	0	$cost_{01}$
实际流失	$cost_{10}$	0

代价敏感学习的实现共有两种方式：在数据预处理方面，通过调整训练数据的正负两个样本的代价权重，使得分类模型更加符合代价敏感的特征；在分类算法方面，在训练过程中将错分代价与损失函数结合起来。通过为不同类别设置不同的错分代价，得到对应的代价敏感模型，该方法将各种类别的错分代价加权到损失函数上，对损失函数进行近似于误分代价的优化过程，给一个非代价敏感分类算法加入了代价敏感因子，从而得出一种具有倾向性的算法[10]。

本文主要基于第二种代价敏感学习方法，在分类算法上通过调整权重进行处理，对损失函数设置权重，侧重于关注少数类，在 python 的 scikit-learn 使用 class_weight 参数设置权重，可以用来权衡各类错误所带来的不同损失，对错误赋予非均等代价。class_weight 通常是用于表示分类模型中多个类型的权重，也可以是一个字典用于定义每个类别的标签，在默认情况下是不输入，即不考虑权重[11]。如果选择输入的话，可以选择 balanced 让库自己统计类型权重，或者自己输入不同类型的权重。例如对于一个二分类模型，我们可以设置 class_weight = {0:0.8, 1:0.2}，则表示分配的 0 类的权重为 80%，1 类的权重分配为 20%。如果设置 class_weight 为 balanced，则会根据训练样本数量来统计计算权重。针对某种类型样本数量，若数量越多，则权重越低，相反，若数量越少，则权重越高。当 class_weight 选择 balanced 时，权重计算公式如下：

$$\text{class\_weight} = \frac{n_{\text{samples}}}{n_{\text{classes}} * \text{np.bincount}(\text{y})} \tag{11}$$

其中 n_{samples} 为样本数，n_{classes} 为类别数量，np.bincount(y)则是统计每个类的样本数。

3.3. 基于代价敏感的逻辑回归

针对电信客户流失数据的不平衡问题，不但要考虑正负样本比例差异较大的情况，还要考虑由于错分代价所产生的挽留成本情况。传统的逻辑回归虽通过损失函数最小化得到最优解，但并没有考虑将一个有电信流失想法的高价值客户预测为仍在使用电信的用户的代价成本。因此，本文采用了一种代价敏感逻辑回归方法进行电信客户流失预测。它通过优化算法拟合逻辑回归算法的系数，该优化算法可将训练数据集上模型的损失最小化，然后在减少模型损失方向上调整系数，修改给定系数集的损耗计算，为每个类别的重要性权重进行调整，以此考虑类平衡，加权损失函数 $J(\theta)$ 公式如下：

$$J(\theta) = -\left(w_0 * \log(yhat_i) * y_i + w_1 * \log(1 - yhat_i) * (1 - y_i)\right) \tag{12}$$

加权应用于损失函数，使得多数类的损失函数得到较低的加权值和较低的误差计算，从而在电信客户流失预测时减少了代价敏感逻辑回归模型系数的更新，同时少数类的损失函数得到较大的权重值和较大的误差值，进而对于代价敏感逻辑回归模型产生较多系数的更新。

4. 仿真研究

4.1. 数据集

本文实验数据源于 Kaggle 平台的某电信公司的数据，数据集中客户样本总共有 7043 条数据，其中

每条样本都包括了 21 个特征属性，由几个维度的客户信息和用户是否最终流失的标签信息构成，其中已经流失的客户样本数据有 1869 条，尚未丢失的客户样本数据有 5174 条[12]。客户信息分为基本信息、开通业务信息、合同信息三类，其中基本信息：包括性别、是否为老年人、经济独立、入网时长等，如表 3 所示；开通业务信息：包括是否开通电话业务、多线服务、网络电视电影、设备保护服务等，如表 4 所示；签订的合同信息：包括签订合同年限、月费用、总费用等，如表 5 所示。

Table 3. Basic customer profile characteristics
表 3. 客户基本信息特征

特征名	中文含义
CustomerID	用户 ID
Gender	性别
SeniorCitizen	是否为老年人
Partner	是否有配偶
Dependents	是否经济独立
tenure	客户入网时长

Table 4. Customer service information characteristics
表 4. 客户服务信息特征

特征名	中文含义
PhoneService	是否开通电话服务业务
MultipleLines	是否开通多线服务
InternetService	是否开通互联网服务
OnlineBackup	是否开通在线备份业务
OnlineSecurity	是否开通网络安全服务
DeviceProtection	是否开通设备保护业务
TechSupport	是否开通技术支持服务
StreamingTV	是否开通网络电视
StreamingMovies	是否开通网络电影

Table 5. Customer contract information
表 5. 客户合同信息

特征名	中文含义
Contract	签订合同年限
PaperlessBilling	是否开通电子账单
PaymentMethod	付款方式
MonthlyCharges	月费用
TotalCharges	总费用

4.2. 数据预处理

通过对数据的查找，发现 TotalCharges 数据类型存在异常，总费用的数据类型为 object 类型，这与实际情况不符合，于是将其进行强制转换成 float 数值类型。再对数据集进行查找，发现 TotalCharges 存在 11 条记录缺失。对该数据进行分析后发现，该 11 条缺失数据的客户入网时长记录为 0，由此可以判断该 11 名用户由于刚入网并未使用电信网，月费均为 0，所以删除缺失值对数据的影响不大。在删除与该缺失值相对应的整行数据之后，该数据中的连续性特征主要有“tenure”，“MonthlyCharges”和“TotalCharges”，而其余则为离散性特征。对于连续的特征，则用标准化的方法进行处理。对于离散特征，不存在大小关系，则使用 one-hot 编码。在属性间存在大小相关联的情况下，采用数值映射。

4.3. 几种模型仿真结果对比分析

该文采用目前流行的 11 种分类模型进行比较，在 python 的 scikit-learn 使用 class_weight 参数设置权重，并选择 Accuracy，AUC，Recall，Precision，F1-score，RRM，RRT，是否过拟合等作为模型的评价指标。该 11 种分类算法均由代价敏感建立模型，在模型训练完成后各模型均可输出特征重要性排序。其中，基于逻辑回归的特征重要性度量排名、ROC 曲线与 Precision、Recall 曲线结果如图 2 所示。各分类模型对比结果如图 3 所示。图 4 为所提分类算法模型的 ROC 曲线以及 Precision、Recall 曲线比较。

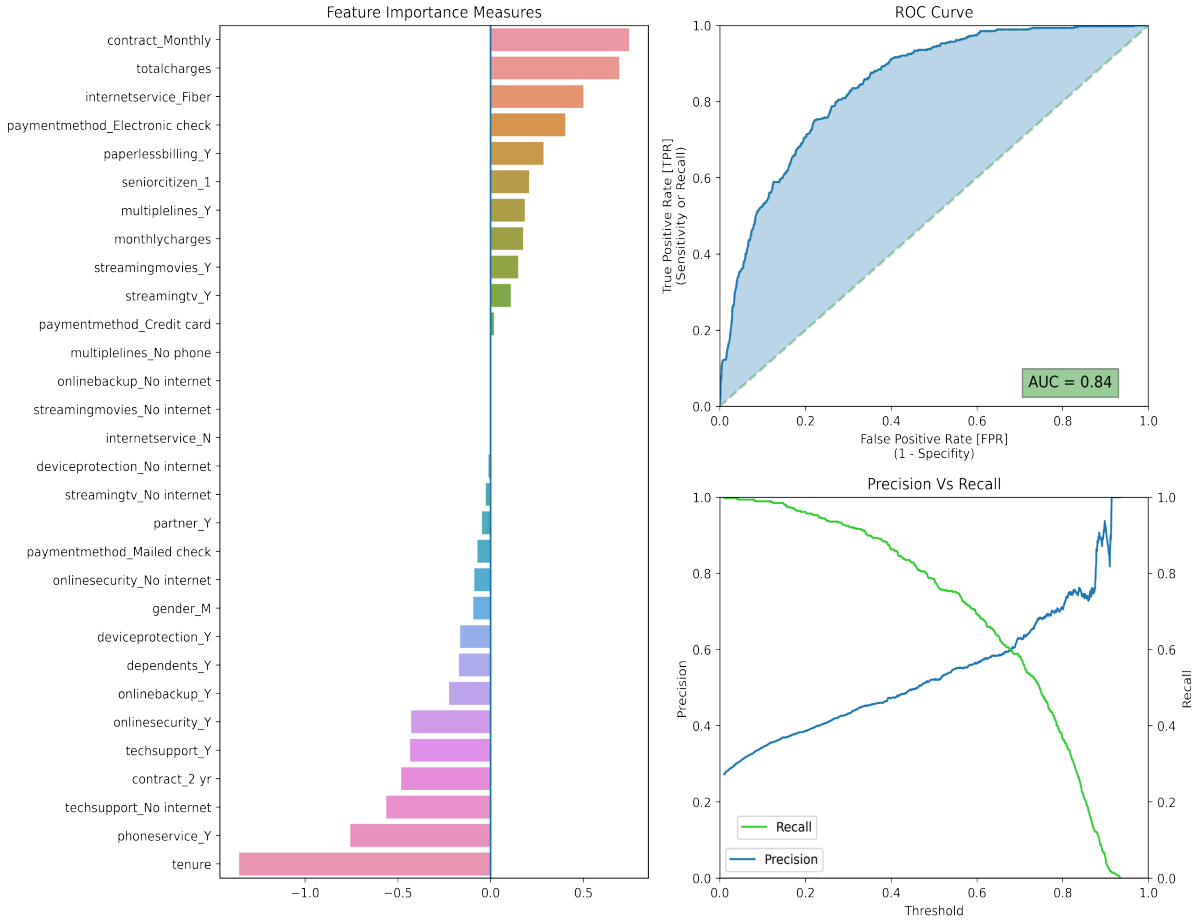

Figure 2. Simulation results based on logistic regression model
图 2. 基于逻辑回归模型仿真结果

	Train Accuracy	Test Accuracy	Overfitting	ROC Area	Precision	Recall	F1-score	Support	RRM/customer	RRM %	RRT %
LogisticRegressionCV	0.751	0.751	False	0.844624	0.521368	0.783726	0.626176	467	77.847268	80.909103	60.717482
RandomForestClassifier	0.7	0.7	False	0.833962	0.463504	0.815846	0.591156	467	76.688583	82.971437	64.536319
Linear SVC	0.7	0.701	False	0.832447	0.46398	0.813704	0.59098	467	76.612368	82.671422	64.477005
CategoricalNB	0.739	0.738	True	0.821244	0.504098	0.79015	0.615513	467	76.813957	80.489536	58.436784
KNeighborsClassifier	0.8	0.797	True	0.842982	0.630952	0.567452	0.59752	467	79.109623	59.531676	28.480971
DecisionTreeClassifier	0.753	0.75	True	0.81769	0.521084	0.740899	0.611848	467	78.011127	76.648838	58.947564
BaggingClassifier	0.748	0.743	True	0.837233	0.511594	0.755889	0.610199	467	76.585836	76.770803	48.239706
AdaBoostClassifier	0.749	0.746	True	0.847724	0.513587	0.809422	0.628429	467	76.257672	81.855576	62.016566
XGBClassifier	0.788	0.754	True	0.834755	0.525849	0.762313	0.622378	467	75.987219	76.818084	53.091100
RBF SVC	0.751	0.742	True	0.829481	0.50922	0.768737	0.612628	467	76.768663	78.262073	53.470557
CatBoostClassifier	0.779	0.758	True	0.847855	0.530973	0.770878	0.628821	467	76.273472	77.973843	56.658716

Figure 3. Several models prediction simulation results
图 3. 几种模型预测仿真结果

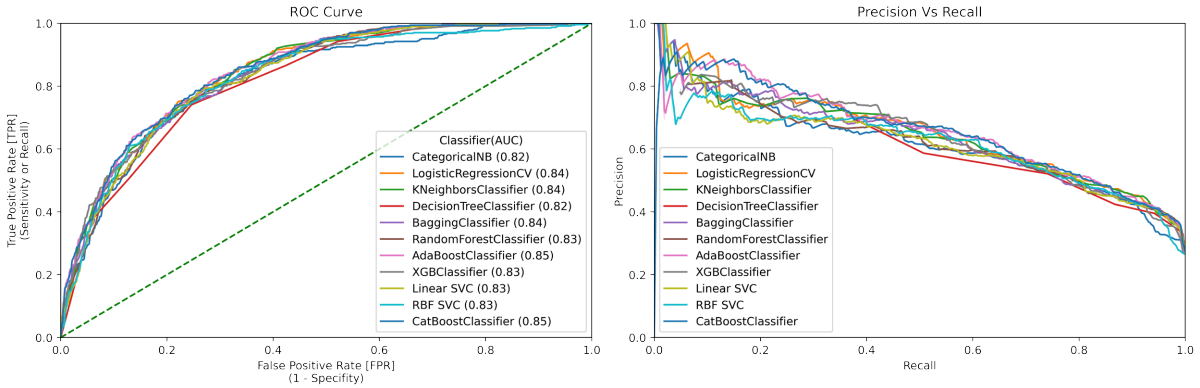

Figure 4. The ROC curves of several models are compared with Precision and Recall
图 4. 几种模型的 ROC 曲线与 Precision、Recall 对比

经过实验发现，由于大多数分类模型具有过拟合现象，我们最终对逻辑回归、随机森林、线性 SVC 三种分类模型进行比较。通过对各评价指标判断，随机森林具有最高的召回率，而逻辑回归具有最佳的 F1 分数，基于在精度和召回率之间做出了正确的权衡，因此逻辑回归具有最佳的整体性能。尽管逻辑回归模型的召回率低于线性 SVC 和随机森林，但逻辑回归保留了很大一部分月收入的特征。即便逻辑回归能够正确识别出流失客户较弱，但它识别出的流失客户的月费更高，因此逻辑回归的 RRM%是接近随机森林和线性 SVC。而通过逻辑回归的高 RRM/customer 值可以判断逻辑回归具有识别高价值客户的能力。对于各分类器没有达到更高准确度分数的原因是由于流失和非流失类重叠，使得分类器很难在不牺牲精确度或召回率的情况下识别较完美的决策边界。

如果该电信公司想要选择在现实世界中部署分类器，公司将不得不在想要保留的收入金额和愿意花在客户保留计划上的收入金额之间进行权衡。对于这种权衡也是精确度和召回率之间权衡的结果。为了保留更多的收入，公司需要选择一个召回率更高、精度更低的分类器，比如随机森林。然而，这将导致更高的误报，使公司在保留计划中包含无流失客户，从而增加其支出。但是，如果公司选择保留相当数量的收入而不在保留计划上花费太多，则必须选择逻辑回归，因为它具有更高的 F1 分数和 RRM/customer 值，而它的 RRM%也接近于随机森林模型。

4.4. 结果分析

基于电信客户不平衡的数据，通过分析 11 种设置代价敏感的分类模型中选择最优的逻辑回归模型进

行客户预测。参数设置方面使用五层交叉验证，最大迭代次数为 50，使用 L1 项用于指定惩罚，损失函数的优化方法则是采用 liblinear 库。对于传统的逻辑回归预测模型和使用代价敏感学习的逻辑回归建立预测模型，结果如表 6 所示。

Table 6. Comparison of results plots for different logistic regression models
表 6. 不同逻辑回归模型对比结果图

Classifier	Precision	Recall	F1	Overfitting	ROC
LR	0.66	0.52	0.58	True	0.83
Cost-LR	0.52	0.78	0.63	False	0.84

通过表 6 可以看出，使用代价敏感学习的逻辑回归预测模型效果较好，相比于未使用代价敏感学习的传统逻辑回归，虽精度下降了 0.14，但在召回率方面提高了 0.26，F1-Score 提高了 0.05。同时可以发现基于代价敏感学习的逻辑回归提升了传统逻辑回归预测模型的泛化性能，从而减少了过拟合现象。通过分析表明此模型在预测电信客户流失时，能够最大可能性地将挽留成本投入到真正的流失客户中。

5. 结束语

本文将代价敏感学习与训练模型的算法结合起来，通过调整对不同算法不同的误分代价设置不同的权重，从而让模型更加专注于对正样本的分类。通过几种分类模型进行多方面比较，最终选择表现突出的代价敏感的 Logistic 回归算法来构建用户流失预测模型。该模型具有最佳的 F1-score 分数且 RRM 也突出，表明该算法可以选择出预测电信客户流失最重要的特征并且对于样本流失具有较好的分类效果。

基金项目

企业信息化与物联网测控技术四川省高校重点实验室(2021WYJ04)。

参考文献

[1] 张力. 基于代价敏感的 XGBoost 算法在电信用户流失预测中的应用[D]: [硕士学位论文]. 广州: 暨南大学, 2020. https://doi.org/10.27167/d.cnki.gjinu.2020.001169

[2] 蒋国瑞, 司学峰. 基于代价敏感 SVM 的电信客户流失预测研究[J]. 计算机应用研究, 2009, 26(2): 521-523.

[3] 王乐, 韩萌, 李小娟, 张妮, 程浩东. 不平衡数据集分类方法综述[J]. 计算机工程与应用, 2021, 57(22): 42-52.

[4] 李赵飞. 基于代价敏感 AdaBoost 的贷款违约风险预测研究[D]: [硕士学位论文]. 南昌: 江西财经大学, 2021. https://doi.org/10.27175/d.cnki.gjxcu.2021.001591

[5] 叶科挺. 基于代价敏感支持向量机与随机森林的个人信用风险评估研究[D]: [硕士学位论文]. 广州: 暨南大学, 2021. https://doi.org/10.27167/d.cnki.gjinu.2021.000413

[6] 冀慧杰, 倪枫, 刘姜, 陆祺灵, 张旭阳, 阙中力. 基于 XGB-BFS 特征选择算法的电信客户流失预测[J]. 计算机技术与发展, 2021, 31(5): 21-25.

[7] 刘定祥, 乔少杰, 张永清, 韩楠, 魏军林, 张榕珂, 黄萍. 不平衡分类的数据采样方法综述[J]. 重庆理工大学学报(自然科学), 2019, 33(7): 102-112.

[8] 向鸿鑫, 杨云. 不平衡数据挖掘方法综述[J]. 计算机工程与应用, 2019, 55(4): 1-16.

[9] 王学玲, 王建林. 基于代价敏感的 AdaBoost 算法改进[J]. 计算机应用与软件, 2013, 30(10): 123-125+138.

[10] 谷琼, 袁磊, 熊启军, 宁彬, 李文新. 基于非均衡数据集的代价敏感学习算法比较研究[J]. 微电子学与计算机, 2011, 28(8): 146-149+153. https://doi.org/10.19304/j.cnki.issn1000-7180.2011.08.041

[11] 万毅斌, 王绍宇, 秦彦霞. 基于代价敏感加权支持向量机的员工离职分类预测[J]. 智能计算机与应用, 2021,

11(12): 43-46+53.

[12] 张三妞, 张智斌. 基于改进粒子群的随机森林优化算法客户流失预测研究[J]. 现代信息科技, 2021, 5(22): 75-78. https://doi.org/10.19850/j.cnki.2096-4706.2021.22.022