Statistics and Application 统计学与应用, 2021, 10(4), 634-641
Published Online August 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.104064

Hans汉斯

基于概率校准的客户流失预测模型研究

徐小燕[1]，夏　燕[2]

[1]中国海洋大学数学科学学院，山东　青岛
[2]武警警官学院，四川　成都
Email: 957779422@qq.com

收稿日期：2021年7月5日；录用日期：2021年7月29日；发布日期：2021年8月4日

摘　要

本文首先利用逻辑回归、线性判别分析、K近邻、支持向量机、贝叶斯判别和决策树模型给出了客户流失预测模型，定量分析了客户是否会流失。然后对用于客户流失预测的分类器进行十折交叉验证，选取了精度最高的线性判别分析模型。其输出结果只有流失或者未流失，不能得到客户流失的概率，我们进一步将线性判别分析模型的输出结果转化成了概率，即概率校准。最后利用Brier分数和概率校准曲线等评价标准对校准前后的模型进行了评估，得到了更好的结果。

关键词

分类器，十折交叉验证，概率校准，Brier分数

Research on Customer Churn Prediction Model Based on Probability Calibration

Xiaoyan Xu[1], Yan Xia[2]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
[2]Officers College of PAP, Chengdu Sichuan
Email: 957779422@qq.com

Received: Jul. 5th, 2021; accepted: Jul. 29th, 2021; published: Aug. 4th, 2021

Abstract

In this paper, we first used logistic regression, linear discriminant analysis, k-nearest neighbor, support vector machine, Bayesian discriminant and decision tree model to give a customer churn prediction model and quantitative analysis of whether the customer will be lost. Then, through the ten-fold cross validation of the six models, we selected the linear discriminant analysis model with

文章引用: 徐小燕, 夏燕. 基于概率校准的客户流失预测模型研究[J]. 统计学与应用, 2021, 10(4): 634-641.
DOI: 10.12677/sa.2021.104064

the highest accuracy. The output result of the linear discriminant analysis model is only loss or no loss, and we cannot get the probability of customer loss, so we further transformed the output result of the linear discriminant analysis model into probability, namely probability calibration. Finally, the Brier score and probability calibration curve were used to evaluate the model before and after calibration, and better results were obtained.

Keywords

Classifier, Ten-Fold Cross Validation, Probability Calibration, Brier Score

Copyright © 2021 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

全球通讯行业迅速发展，通讯需求也日渐扩大。我国电话通信行业形成移动、联通、电信三大运营商并行的现状。当前我国通信行业处于 4G 时代的末期，5G 时代即将全面到来，通信行业的技术革新周期越来越短，运营商也需要即时更新营销策略，服务是否能满足客户对于通信的服务需求成为了运营商盈利的关键。对于电信行业来说，客户的数量和质量决定着企业的竞争力和效益，竞争也日趋激烈。在行业内竞争如此激烈的情况下，准确预测客户流失的风险对企业效益有很大的影响。在三大运营商并行的情况下，客户会有更多的选择，因此，要保证留住客户，研究客户流失概率预测是一个具有现实意义的指导性问题，值得我们深入研究。由此可以进一步研究高流失率和低流失率客户特征，给出运营商有针对性的建议，结合运营商业务提出新优惠套餐，有利于提升用户体验，留住老客户，吸引新客户以提高自身竞争力。

在分类领域中，对于客户流失这一问题，大量学者做了诸多有意义的探索。在当前对客户流失问题的研究中，大多数为定性的分析现状和影响因素，很少有定量分析客户流失的概率大小，现有的定量分析中大多使用单一模型并且在模型的选取上有一定的随机性。客户流失本质上是一个分类问题，它只有两种情况，是或否(分别取值为 1，0)。针对二分类问题，Logistic 回归、线性判别分析、K 近邻、支持向量机、贝叶斯判别和决策树可以将两个不同的类别很好的分开，从而可以识别客户是否流失[1]-[6]。有些分类器只能给出类别而不能给出取某类的概率，为解决此问题，肖瑶等人在概率校准方法的领域做了一些有意义的探索[7] [8] [9]。

在一般分类问题中，都是根据已知类别样本的特征来拟合分类器，寻找分类的规律，然后根据分类器判断未知类别样本属于哪类样本。大多数分类器只能够判断出样本的类别，不能给出判为某一类的概率是多少，在很多情况下不光要判断类别，我们还需要得到样本属于某一类的概率。比如在电信客户流失问题中，电信公司会根据客户的购买能力和个人信用等因素来判断客户的流失风险。但是电信公司并不局限于识别出客户是否会流失，更希望能确定客户流失的风险，以便计算客户违约的概率。在有些可以得到概率的模型中，如果判为不同类别的概率相差不大，那么预测结果可能存在一定的误差，所以也需要进一步做概率校准，以得到更精准的结果。

2. 概率校准

概率校准是利用一个函数将原始分类模型的分类结果转换为准确的概率，得到样本属于某类的概率。

一个校准良好的预测模型可以反映潜在的客户流失概率。概率校准的方法可分为参数和非参数方法两大类。其中 Platt Scaling (普拉特缩放)校准是参数方法的代表性方法，保序回归在非参数方法中得到了较为广泛的应用。

2.1. Platt Scaling 校准

Platt scaling 校准以分类器的输出结果作为输入特征，使用的函数是 sigmoid 函数，用于调整原始模型，拟合概率校准模型，将原始分类器的输出结果转化为概率值。也就是以原始模型的输出结果为自变量，以原数据的目标变量作为因变量拟合逻辑回归模型，是一种参数方法。假设 f 为原始模型的输出值，Platt Scaling 校准后的输出结果如式(1)所示：

$$P(y=1|f)=\frac{1}{1+\exp(Af+B)} \tag{1}$$

其中，参数 A、B 由极大似然估计得出。

Platt Scaling 校准步骤如图 1 左半部分所示。

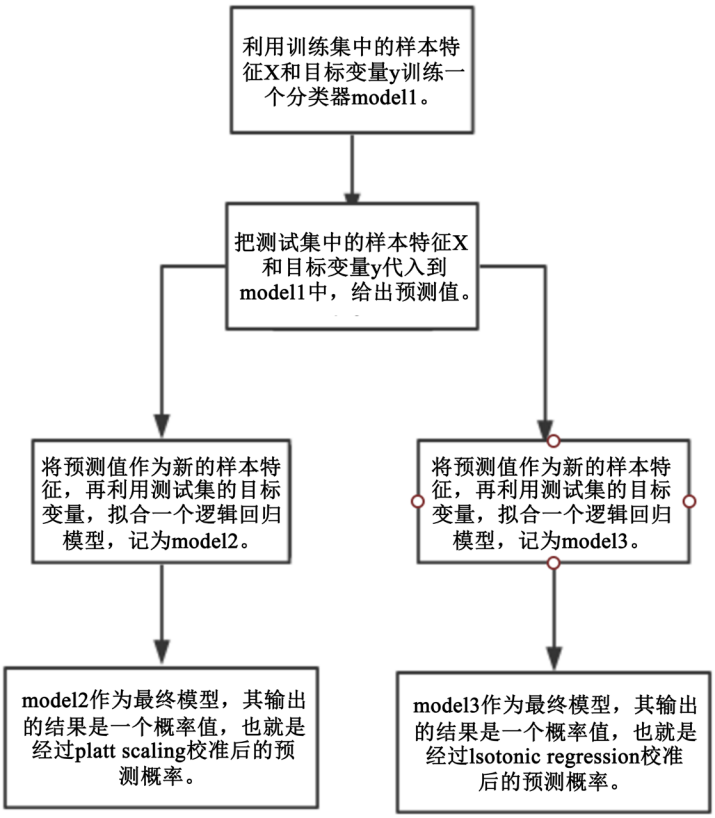

Figure 1. Probability calibration flow chart
图 1. 概率校准流程图

2.2. 保序回归

保序回归(Isotonic regression)是一种非参数回归方法，可用作概率校准方法。它是在给定一组数字序列y的情况下，通过一系列的运算，改变序列中每个元素的值，得到一个非递减序列 y'，同时满足 y 和 y' 的误差(取二者之差的平方)最小化。保序回归校准的步骤如图 1 右半部分所示。

3. 分类器理论

3.1. 逻辑回归

关于逻辑回归模型，适用于输出结果为 0 和 1 的二分类问题。式(2)中普通线性回归得到的估计值为实值，我们必须将其转化成 0/1 值。要把实值 z 转化为 0/1 值，并且使其取值在区间(0, 1)内连续，就运用到 sigmoid 函数对 z 进行运算，如式(3)所示，其中 P 指样本为正例的概率。由式(4)可以看出，实际上对数几率 $\ln\frac{P}{1-P}$ 可以用式(2)中的普通线性回归值 z 逼近，这个模型叫做逻辑回归。

$$z = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{2}$$

$$P = \frac{1}{1+\mathrm{e}^{-z}} \tag{3}$$

$$\text{Logistic}(p) = \ln\frac{P}{1-P} = z \tag{4}$$

3.2. 支持向量机

支持向量机算法通过计算找到一个最佳的分类超平面，它位于两类样本的中间，可以很好的分离不同类别，最接近这个超平面的几个样本被称为“支持向量”，支持向量也分为正类和负类，正类和负类的支持向量与超平面之间的距离相加得到“间隔”，需要找到一个最大间隔，即划分超平面需要满足间隔最大化，且正负类样本分别属于支持向量两侧。

3.3. 贝叶斯判别

贝叶斯判别器通过在样本中统计先验概率，用先验概率表示对数据集的认识，然后随机采样得到训练样本 X，来优化这种认识，计算出样本属于某类 G_t 的后验概率，形成分类模型。后验概率即在已知某个样本的情况下，把它归为某类的条件概率 $P(G_t|X)$，样本属于哪个总体的后验概率大就将它判为哪个总体。当有一个未知类别的样本 x'，就可以根据后验概率分布求出属于各个类别的后验概率，进而判别出 x'属于哪个总体。

3.4. K 近邻

K 近邻的基本思想：基于距离度量，在训练集中找出最近的 k 个样本，然后根据这 k 个样本的特征和标签来进行预测。基于分类任务，使用投票法提供预测结果，采用样本频率最大的类别标签；基于回归任务，应使用 k 个样本的实际输出值的平均值作为预测结果。本文研究的为分类问题，K 近邻算法采用投票法给出类别标签。

3.5. 决策树

决策树是一种划分类别的树状的模型。二分类问题可视为“该样本是否属于正类”的决策过程。CART 决策树采用的划分准则是 GINI 指数。其在选择属性进行树枝的分叉时，随着 GINI 指数的减小，纯度也会相应的提升。因此，最优划分属性是使 GINI 指数最小的属性。决策树由最优划分属性开始生成。

3.6. Fisher 判别

Fisher 判别的规则是先投影，后判别。投影是把样本的点投影到 p 维空间的某一个方向上。通常可以找到一个方向，使得组间平方和与组内平方和的比值在组内平方和为 1 的约束条件下，在这个方向上

的一条直线上达到极大，样本能分开得最好，这样就达到降维的目的。这条投影线就可以作为判别函数，然后计算判别函数值和判别临界值，根据距离判别准则得出判别标准，将未知类别的样本的各个特征代入到判别函数，根据标准，判断其属于哪类总体。

4. 实证分析

4.1. 数据介绍

本文所用数据集来源于某电信公司，共 19 个特征，3463 个样本，其中第一个字段为客户 ID，用于区分客户，不作为建模特征。本数据集特征详细介绍如表 1：

Table 1. Variable interpretation table
表 1. 变量解释表

No	变量名	变量名翻译	数据类型	变量取值范围
1	subscriberID	个人客户的 ID	数值型	不同 ID 代表不同客户
2	churn	是否流失	二分类	0-否，1-是
3	gender	性别	二分类	0-男，1-女
4	AGE	年龄	数值型	9~82
5	edu_class	受教育程度	数值型	0~3
6	incomeCode	居住区域平均收入	数值型	1-62
7	duration	在网时长	数值型	2~73
8	peakMinAv	统计期间内最高单月通话时长	数值型	0~1269.3333
9	peakMinDiff	统计期间结束月份与开始月份相比通话时长增加数量	数值型	−1494~1118
10	posTrend	通话时长是否呈现出上升态势	二分类	0-否，1-是
11	negTrend	通话时长是否呈现出下降态势	二分类	0-否，1-是
12	nrProm	电信公司营销的数量	数值型	0-3
13	prom	最近一个月是否被营销过	二分类	0-否，1-是
14	curPlan	统计开始时套餐最高通话时长	数值型	1 = 200 分钟；2 = 300 分钟；3 = 350 分钟；4 = 500 分钟
15	avgplan	统计期间内平均套餐最高通话时长	数值型	1 = 200 分钟；2 = 300 分钟；3 = 350 分钟；4 = 500 分钟
16	planChange	统计结束时套餐档次的变化	数值型	>0 表示提高，<0 表示下降，=0 表示不变
17	posPlanChange	统计期间是否提高套餐	二分类	0-否，1-是
18	negPlanChange	统计期间是否降低套餐	二分类	0-否，1-是
19	call_10086	统计期间是否拨打 10086	二分类	0-否，1-是

4.2. 建模流程

在建模前，首先随机划分原始数据集，70%的样本用作训练集，用于训练分类模型，30%的样本用作测试集，用来验证概率校准效果。为保证最终模型的精确度和稳定性，本文在验证概率校准效果之前，先用训练集作为原始数据计算十折交叉验证误差，比较不同基分类器的准确度，选出综合效果最好的分类器，对其进行概率校准后得到最终模型，并用一系列指标进行评价，建模流程如图 2。

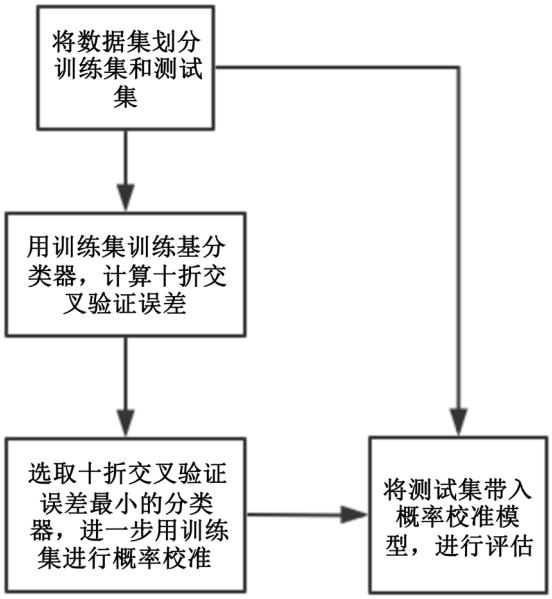

Figure 2. Modeling flow chart
图 2. 建模流程图

4.3. 评价标准及结果分析

在选取模型的过程中，模型评价标准的选取是非常重要的。模型的选择不光要验证结果的准确性，还要验证模型的稳定性。丰富的指标和评价标准的适用性有所不同。由于所用模型都是用于二分类问题的预测，目标变量属于二分类变量，考虑到模型的准确度和稳定性，本文选取了分类问题常用的十折交叉验证误差以及基于混淆矩阵的查准率、查全率和 F1 值。此外，对于概率校准问题还应用了 Brier 分数和概率校准曲线，对比了校准前后结果的准确性和稳定性。

4.3.1. 十折交叉验证误差

为了评估模型的预测精度，十折交叉验证将原始数据集划分成十个大小相似的互斥的子集。每次，十个子集的一个作为测试集，将其余九个的组合用作训练集。这将产生十个预测结果，计算这十次预测结果的预测误差，再取平均值获得十折交叉验证误差，其取值越小说明模型精度越高。十折交叉验证综合考虑了所有样本，对所有样本都进行了检验，估计结果更可靠。

本文分别用 6 种单分类模型拟合模型，计算十折交叉验证误差得到表 2 中结果，发现十折交叉验证误差最小的线性判别模型，所以对线性判别模型进一步做概率校准。

Table 2. Comparison table of classification models
表 2. 分类模型对比表

模型	十折交叉验证准确率	十折交叉验证误差
逻辑回归模型(LR)	0.8096	0.1904
线性判别模型(LDA)	0.8134	0.1866
K 近邻(KNN)	0.6912	0.3088
CART 决策树(CART)	0.7877	0.2123
贝叶斯判别(NB)	0.7704	0.2296
支持向量机(SVM)	0.7657	0.2343

4.3.2. 混淆矩阵

用训练集训练出各个模型，再将测试集代入到模型中，将预测结果和真实标签作对比，得到混淆矩阵如表 3 所示。

Table 3. Confusion matrix
表 3. 混淆矩阵

混淆矩阵		预测结果	
		正例	反例
真实结果	正例	TP (真正例)	FN (假反例)
	反例	FP (假正例)	TN (真反例)

查准率 P 为在预测结果为正例的样本中，预测正确样本所占的比例；
查全率 R 为在真实结果为正例的样本中，预测正确样本所占的比例；
F1 综合考虑了查准率与查全率。

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{5}$$

$$\mathrm{R} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{6}$$

$$\mathrm{F1} = 2*\mathrm{P}*\mathrm{R}/(\mathrm{P}+\mathrm{R}) \tag{7}$$

4.3.3. Brier 分数

Brier 分数(记作 BS)是用于验证概率校准效果的指标，它是对所有预测事件的预测概率与实际概率离差平方和取平均得到的值，值越小代表校准效果越好，预测误差越小。

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{8}$$

通过表 4 中实验结果对比发现，在进行 Isotonic regression 校准之后，LDA 模型的 Brier 分数降低，说明校准效果较好，查准率由 0.741 提高为 0.773，F1 值由 0.789 提高到 0.791，准确度也有所提升。

Table 4. Comparison of models before and after probability calibration
表 4. 概率校准前后模型对比表

模型	Brier 分数	Precision	Recall	F1
LDA	0.134	0.741	0.842	0.789
LDA + Isotonic	0.132	0.773	0.810	0.791
LDA + Sigmoid	0.132	0.768	0.801	0.784

4.3.4. 概率校准曲线

概率校准曲线也是验证概率校准模型性能的一项指标，展示的是实际类别频率(即实际概率)与预测概率之间的关系，它是将[0, 1]按 0.1 的宽度等宽划分成 10 个区间，计算每个区间内预测概率的均值作为概率校准曲线的横坐标，该区间中正类样本所占的比例作为纵坐标，对角线表示实际值与预测值相等，越接近对角线校准效果越好。

由图 3 可见，经过概率校准后实际值与预测值更加接近，而且 Isotonic 校准的效果更佳。

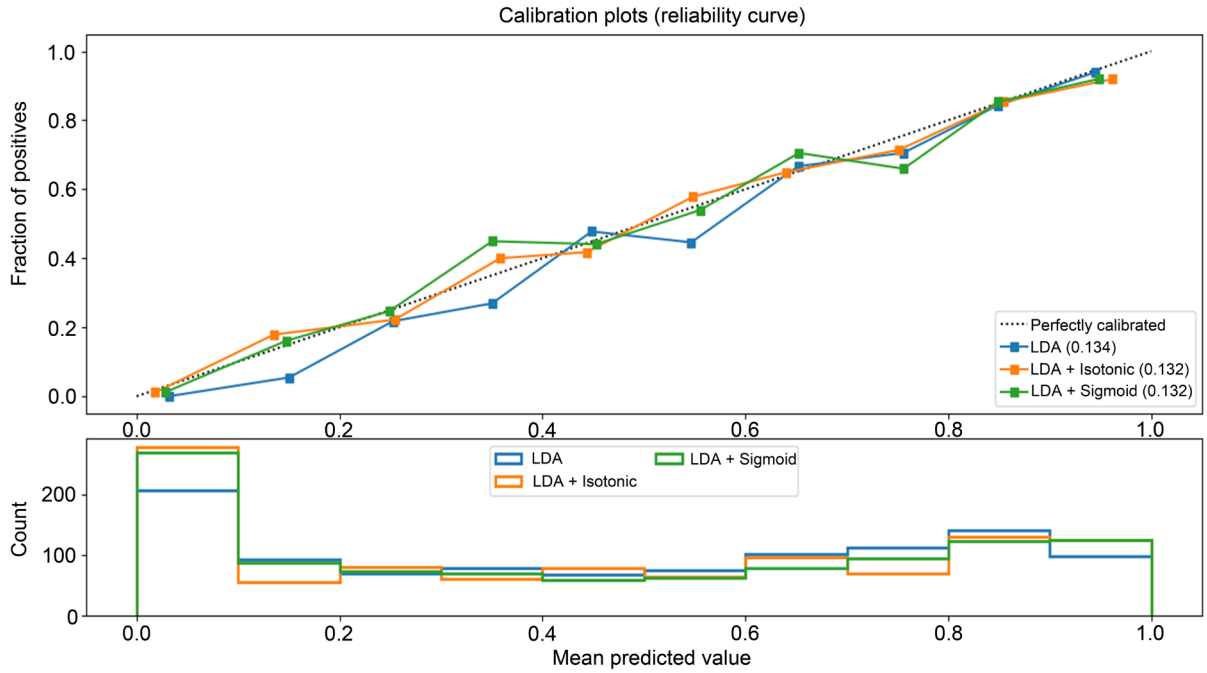

Figure 3. Probability calibration chart
图 3. 概率校准图

5. 结论

本文通过对分类问题以及概率校准方法研究，应用电信客户是否流失的分类数据进行了实证分析，客户流失概率的预测得以实现。在客户流失预测中，本文先用 Logistic 回归、线性判别分析、K 近邻、支持向量机、贝叶斯判别和决策树 6 个分类器实现对客户是否流失的分类，对比十折交叉验证误差发现，线性判别分析模型的预测精度最高，又进一步对其进行概率校准得到客户流失的概率。实验结果表明，在概率校准之后，模型的 Brier 分数降低，概率校准曲线也有所改善，精确度有所提高，充分证明了概率校准可以有效的改善分类器精确度。在对 LDA 模型的校准中，经过 Isotonic regression 校准后的模型显然更好。并且概率校准方法可以得到样本取正类的概率值，为概率预测提供了有效的参考。

参考文献

[1] 顾正云. 信用评分模型有效性比较[D]: [硕士学位论文]. 南京: 南京大学, 2011.

[2] 张国政, 陈维煌, 刘呈辉. 基于 Logistic 模型的商业银行个人消费信贷风险评估研究[J]. 金融理论与实践, 2015(3): 53-57.

[3] 肖铮. 常用的三种分类算法及其比较分析[J]. 重庆科技学院学报(自然科学版), 2020, 22(5): 101-106.

[4] 姜飞, 杨明, 刘雨欣. 基于支持向量机混合采样的不平衡数据分类方法[J]. 数学的实践与认识, 2021, 51(1): 88-96.

[5] 曹玲玲, 潘建寿. 基于 Fisher 判别分析的贝叶斯分类器[J]. 计算机工程, 2011, 37(10): 162-164.

[6] 李衍. 移动互联网背景下客户流失预测研究[D]: [硕士学位论文]. 厦门: 厦门大学, 2018.

[7] 肖瑶, 谢贵才, 朱兵. 浅谈分类问题中的概率校准[J]. 中国统计, 2018(5): 35-37.

[8] 姜正申, 刘宏志. 基于概率校准的集成学习[J]. 计算机应用, 2016, 36(2): 291-294, 407.

[9] 罗艳虹, 李治, 余红梅, 郭虎生, 曹红艳, 王蕾, 宋春英, 郭兴萍, 张岩波. 基于代价敏感性和概率校准的先天性心脏病概率预测模型研究[J]. 中国卫生统计, 2019, 36(1): 36-39.