基于集成森林元学习网络的客户流失预测

李龙戈[1]，郑铿城[2]
（1. 中国通信建设集团设计院有限公司数字化咨询中心，河南 郑州 450000；
2. 中南财经政法大学，湖北 武汉 430073）

摘　要：为解决树模型在客户流失预测任务中较难捕捉时序特征的问题，提出了基于集成森林元学习网络（ensemble forest meta-learning network，EFML）的流失预测方法。首先通过分组等策略进行数据提质，并结合下采样技术解决样本类别不平衡问题。然后，基于EFML的语义图构造器构建用户时序特征的语义向量，以描绘用户细粒度行为，形成语义图并显式融合。最后，训练多个基础树模型作为元学习器-多层感知机（multilayer perceptron，MLP）的输入，生成综合的流失预测结果。实验证明，EFML能充分挖掘客户历史行为差异，捕获和学习基础树模型间的互补关系，相对于随机森林（random forest，RF），其AUC提升2.7%，AP提升3.7%，预测精度提升显著。该框架结合树模型和微观特征，具备卓越的解释性，为运营商实现更精细的用户化管理提供新视角。

关键词：语义图；客户流失；历史行为差异；树模型
中图分类号：TP183
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024159

Customer churn prediction based on the integration of meta-learning network of the forest

LI Longge[1], ZHENG Kengcheng[2]
1. Digital Consultation Centre, China Communications Construction Group Design Institute Co., Ltd., Zhengzhou 450000, China
2. Zhongnan University of Economics and Law, Wuhan 430073, China

Abstract: To address the challenge of capturing temporal features in customer churn prediction tasks by tree models, a churn prediction method based on ensemble forest meta-learning network (EFML) was proposed. Firstly, data quality was improved through grouping strategies and class imbalance issues were addressed with undersampling techniques. Secondly, semantic vectors of user temporal features were constructed using EFML's semantic graph constructor to depict fine-grained user behavior, forming a semantic graph and explicitly integrating it. Finally, multiple base tree models were trained as meta-learners, with the inputs being multilayer perceptron (MLP) to generate comprehensive churn prediction results. Experimental results demonstrate that EFML can effectively exploit differences in customer historical behaviors, capture and learn complementary relationships between base tree models. Compared to random

收稿日期：2024−01−14；修回日期：2024−04−20

forest (RF), EFML shows a 2.7% increase in AUC, a 3.7% increase in AP, and a significant improvement in prediction accuracy. This framework, combining tree models and micro-level features, possesses excellent interpretability, providing a new perspective for operators to achieve more refined user-centric management.

Key words: semantic graph, customer churn, historical behavior difference, tree model

0 引言

作为电信业务的基石，客户的留存与流失会直接影响运营商的市场占有率。随着“携号转网”“降费提速”等一系列政策的施行，如何减少客户流失对运营商愈发重要[1]。当前，运营商客户的流失预测问题正逐步演进为数据驱动的模式，通过机器学习的方法探索客户数据中潜藏的高价值信息并应用，可有效降低客户的流失率，这对运营商具有重要的理论意义和实践价值。

在近年的研究中，客户流失预测任务的发展主要集中在两个方向[2]。

第一个方向为基于深度神经网络结构的算法优化。黄子璇等[2]提出了基于图算法的自动特征交叉构造的框架来挖掘交叉特征，为训练预测模型做准备，然后基于时序图注意力机制实现用户流失的在线预测。钟保权等[3]根据业务经验，由用户流失原因反推导致用户流失的因子，根据推导的因子、深度学习算法和卷积神经网络（convolutional neural network，CNN）来预测用户流失的概率。Zhang等[4]将客户行为看作一种时序行为，从而利用长短期记忆（long short-term memory，LSTM）网络[5]进行用户流失预测，所得实验结果与其训练随机森林（random forest，RF）模型相当，且优于逻辑回归模型。这些深度学习模型致力于自动提取显式和隐式高阶特征，降低对特征工程的依赖，具备强大的自适应性和鲁棒性，适合大规模数据下的流失预测任务，但其性能略次于基于回归树的集成模型[2]，可解释性差，易发生过拟合现象。

第二个方向为基于机器学习树状结构算法的集成和改进，通过构建决策树等模型，提升分类性能，且获得较好的可解释性。Caigny等[6]采用决策树与逻辑回归结合的方式，在保证模型的分类性能的同时还具备良好的可解释性。UllahI等[7]采用RF进行分类，使用余弦相似性的k均值聚类对流失客户进行分段，分析了可能造成流失的因素，增强了分析深度。HöppnerS等[8]提出了使用期望最大利润集成的决策树模型，并使用进化算法进行学习，取得了分类性能的显著提升。Ahmad等[9]使用社交网络做特征提取，弥补了XGBoost在特征提取上的缺陷，提高了XGBoost在真实流失数据集上的预测准确率。这类模型的构建在特征工程阶段需要耗费一定的成本，大规模数据时代的普适性也相对较差。

树状结构模型在处理时序行为等复杂特征时存在一定的局限性，导致模型表征能力受限；而深度神经网络模型虽然具备自动提取特征能力，但是其解释性不足，难以深入探索用户流失机制。在此背景下，本文提出基于集成森林元学习网络（ensemble forest meta-learning network，EFML）的新型预测框架，旨在综合利用树状结构和深度神经网络的优势，克服两者的局限性。EFML框架通过构建用户语义图的方式，捕捉特征之间的关联依赖，实现用户多维异构特征的融合表达；此外，还引入基于多层感知机（multi-layer perceptron，MLP）的元学习器，集成汇聚不同基础树结构分类器的预测结果，全面优化模型性能。实验结果表明，本文方法相较传统技术在预测效果与模型可解释性两个方面均获得显著提升，该研究对于电信企业实现精细化客户维系与针对性营销决策均具有重要指导意义。

表1　客户流失数据集

年月	用户ID	在网时长/月	VIP等级	…	是否标记流失
201608	U3114031824148707	62	99	…	0
201608	U3115062769878916	5	3	…	0
201608	U3115060569699130	171	4	…	1

1　数据准备

1.1　数据说明

本研究使用了Kaggle数据科学竞赛平台提供的某运营商客户流失数据集。该数据集共记录了用户个人信息、在网时长、VIP等级等34项有效特征，以及因变量客户是否流失的标记。数据的时间跨度为3个月，总计包含900 000条样本记录，其中6月和7月的流失标记值为空。客户流失数据集见表1。

1.2　数据预处理

由于系统宕机、人为误操作和网络流失等因素，电信客户数据中夹杂了噪声，如数据缺失值、重复值和异常值等，噪声会影响预测模型的精度，在模型训练与测试前必须进行预处理，处理过程如下。

（1）数据去重。初始载入数据集后，采用drop_duplicates函数进行数据重复性检查，去除重复记录来确保数据的唯一性。经检测，重复数据为96条，对模型精度有害无益，予以删除。

（2）数据分组与合并。数据集中相同的用户ID视为同一用户，每个用户对应3条样本，3条样本的时间跨度为2016年6月到2016年8月共3个月，数据标注集中在2016年8月的样本中。因此需要通过用户ID（USER_ID）和标注进行数据分组，将数据处理为单个用户单条样本，完成数据合并。合并后共保留299 968条样本，其中标注为0的样本数量为290 289条，标注为1的样本数量为9 679条。合并策略包括：

- 离散数据的处理，采用取平均值、中位数和独热编码；
- 连续数据的处理，采用取平均值或者非空行的值替换；
- 其他属性的数据，比如客户ID、手机型号、用户性别等常量，取2016年8月样本的值。

（3）特征筛选。移除在客户流失预测中不具有显著影响的列，包括手机品牌、手机型号、手机系统、星座，这类特征对目标变量缺乏显式的解释能力或相关性。

（4）缺失值处理。使用近邻值填充的方式处理缺失值，保证数据的完整性和可用性。

预处理完成后，数据形状为299 968行、29列，特征及字段描述见表2。

2　客户流失预测模型简述

2.1　预测流程

EFML在流失预测任务中，包含一系列关键步骤：首先，将用户的时间序列行为转换为语义图，与合并后的分组特征相融合；然后，拟合LightGBM等基础分类器并生成预测概率向量；最后，将概率向量送入元学习器，整合LightGBM等基础分类器的预测结果表示，完成客户状态分类。EFML模型预测流程如图1所示。

2.2　树结构模型算法简述

LightGBM是基于决策树算法的梯度提升框架，最初由微软亚洲研究院在NIPS系列论文[10-11]中提出。它适用于排序、分类、回归等多种机器学习任务。随着不断迭代，LightGBM增加了许多功能，比如支持叶子索引预测、最大深度设置、特征重要性排序、多分类、正则化、支持特征缺失值、RF模式等。它以回归树作为弱学习器，利用每次预测结果与目标值的残差作为

表2 特征及字段描述

名称	字段描述
INNET_MONTH	在网时长/月
IS AGREE	是否合约有效用户 (0或1)
AGREE EXP_DATE	合约计划到期时间(日期)
CREDIT LEVEL	信用等级
VIP_LVL	VIP等级
ACCT_FEE	本月费用 /元
CALL DURA	通话时长/s
NO_ROAM LOCAL CALL DURA	本地通话时长/s
NO ROAM GN LONG CALL DURA	国内长途通话时长/s
GN ROAM CALL DURA	国内漫游通话时长/s
CDR NUM	通话次数/次
NO ROAM LOCAL CDR NUM	本地通话次数/次
NO ROAM GN LONG CDR NUM	国内长途通话次数/次
GN ROAM_CDR NUM	国内漫游通话次数/次
NO_ROAM CDR NUM	非漫游通话次数/次
P2P_SMS _CNT_UP	短信发送数/条
TOTAL FLUX	上网流量/MB
LOCAL FLUX	本地非漫游上网流量/MB
GN ROAM FLUX	国内漫游上网流量/MB
CALL DAYS	有通话天数
CALLING DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL RING	语音呼叫圈
CALLING RING	主叫呼叫圈
CALLED RING	被叫呼叫圈
CUST SEX	性别
CERT AGE	年龄
TERM TYPE	终端便件类型(4=4G、3=3G、2=2G)
Is LOST	用户在3月是否流失标记 (1=是，0=否)，1月和2月值为空

下一轮学习的目标，生成当前的残差回归树。每棵树都学习前面树的结论和残差，最终将多棵决策树的结果相加作为最终预测输出。通过直方图算法对特征进行预排序，并采用节点展开方式构建树，这使得它成为一个高效、高精度、高性能的分类算法。传统的梯度提升决策树（gradient boosting decision tree，GBDT）算法是以分类回归树（classification and regression tree，CART）为基模型的一种集成学习算法，通过基函数的线性组合，迭代减小残差，利用损失函数（loss）作为准确度指标，从而达到分类或回归的效果。相较于GBDT算法，LightGBM能训练更少的样本与特征、占用更少的内存，其基本思想是利用弱分类器迭代训练得到最优模型。LightGBM算法在GBDT算法上进行的优化包括基于直方图（histogram）的决策树算法、支持高效并行、GOSS算法、直接支持类别特征和Cache命中率优化[12-14]等。直方图算法将连续属性值划分为k个离散的分桶（bin），构建宽度为k的直方图。在遍历训练数据时，记录离散值在直方图中的累

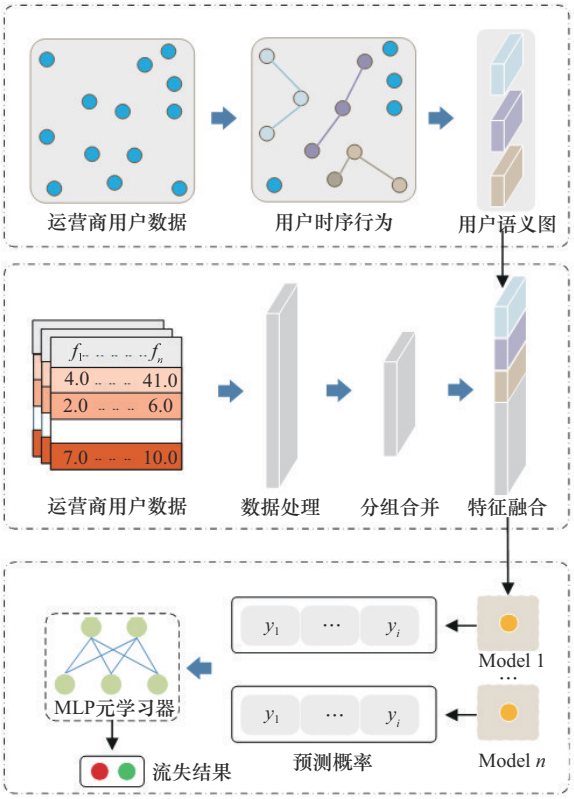


图 1　EFML 模型预测流程

积统计量。然后对直方图的离散值进行排序遍历，找到具有最大增益的节点作为最佳分割点。

CatBoost 算法是由俄罗斯顶尖技术公司 Yandex 在 GitHub 上开源的基于 GBDT 的改进算法。该算法的核心思想借鉴了对称完全二叉树，即每次只划分出两条路径，且划分路径的顺序是随机的。在划分过程中，特征维数不会减小，但用于划分的特征会与其他类别特征通过贪婪算法[15]的方式相结合，形成新特征。此外，在逐个添加样本的过程中，CatBoost 能够自动检测并剔除干扰样本，随着样本数量的增加，预测结果会变得更加准确。

RF 是基于决策树的集成学习算法，相较于其他分类器，其性能表现更为优越。它能有效解决过拟合问题，并具有较高的准确性。此外，RF 还能处理具有高维特征的输入样本，无须进行特征删除。由于在每次划分时仅考虑少数属性，它在大型数据库的应用中表现出良好的效果。

XGBoost 是由 Chen 等[16]提出的一种基于梯度提升算法的集成学习模型框架。它充分发挥了 GBDT 算法的优势，并在其基础上进行了一系列优化，这使得最终训练的 XGBoost 模型准确性高于 GBDT，模型计算代价更低，且运行速度更快。XGBoost 本质上是一种决策树算法的集成，它在基础算法上添加了迭代过程。与 GBDT 相比，XGBoost 有很多创新点，如在代价函数中加入了正则化项、用于控制模型的复杂度、防止过拟合等。

2.3　特征重要度

树结构模型大多以回归树作为基学习器，可以追踪每个特征在树节点中的出现次数。特征出现次数越多，代表其对分类的贡献越大，即树结构模型的特征重要性可视为该特征在所有树中出现次数的累加：

$$\mathrm{IM}_{X_i}=\sum_{j=1}^{N_{\mathrm{trees}}} n_j^{(i)} \tag{1}$$

其中，IM_{X_i} 为特征 X_i 的重要性，N_{trees} 为模型构建的树的棵数，$n_j^{(i)}$ 为特征 X_i 在第 j 棵树中出现的次数。

2.4　用户语义图

语义图特征是被用于挖掘语义信息的关键，参与语义图构建的子特征及描述见表 3。

表 3　参与语义图构建的子特征及描述

特征名称	示意
NO_ROAM_CDR_NUM	非漫游通话次数/次
P2P_SMS_CNT_UP	短信发送数/条
TOTAL_FLUX	上网流量/MB
LOCAL_FLUX	本地非漫游上网流量/MB
GN_ROAM_FLUX	国内漫游上网流量/MB
CALL_DAYS	有通话天数
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLING_RING	主叫呼叫圈
CALLED_RING	被叫呼叫圈

给定的样本集合中的两个样本x_i和x_j，生成用户语义图数学形式的推导如下：

$$\text{feature}_{i,j}=\sum_{i,j}\left(\sum_{k=1}^{n}|x_{i,k}-x_{j,k}|^2\right) \tag{2}$$

其中，$\text{feature}_{i,j}$表示样本x_i和x_j提取的时序行为特征，$x_{i,k}$和$x_{j,k}$分别表示样本x_i和x_j的第k个特征值，n表示特征的数量，$|x_{i,k}-x_{j,k}|^2$表示两个样本在第k个特征上差值的二次方。

样本间的差分乘积再进行求和能够衡量样本间行为特征差异的总和，通过对多条样本差分乘积再求和，生成由不同feature组成的语义向量，以获取样本之间行为的相对差异程度，将语义向量装配成语义图特征进行显示融合。

2.5 MLP元学习器

MLP是一种经典的前馈神经网络，在元学习任务中至关重要，其组成包括输入层、隐藏层和输出层。每一层的节点都与下一层全连接，使用线性整流（rectified linear unit，ReLU）函数激活以引入非线性特性，有效地解决梯度消失和学习率下降的问题。在优化过程中，设置二元交叉熵损失函数指导网络参数的更新。

MLP作为元学习器，能够基于感知器实现非线性判别，可近似大多数非线性函数，并且支持以近似的方式表示连续输入和输出的任意函数，为EFML模型端到端节点的映射提供了基本思路，多层感知机结构如图2所示。

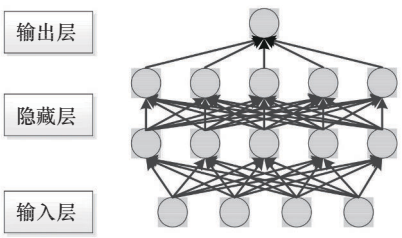


图2 多层感知机结构

3 实验分析

实验的运行环境为：Windows10操作系统、Python3.8解释器和Anaconda3集成开发环境。实验所用到的算法均调用了开源的机器学习框架sklearn和Numpy及Pandas等数据处理库。

3.1 评估指标

本文旨在针对电信领域的客户群体，精准预测其未来的流失或留存状态，该问题是典型的二分类问题。为准确评价分类结果，实验采用了混淆矩阵进行分析，混淆矩阵见表4。

表4 混淆矩阵

真实结果	预测结果	
	流失用户	留存用户
流失用户	TP（真流失用户）	FP（假留存用户）
留存用户	FN（假流失用户）	TN（真留存用户）

注：TP表示预测为流失状态且真实为流失状态的客户样本统计数量；FP表示预测为留存状态但真实为流失状态的客户样本数量；TN表示预测为留存状态且真实为留存状态的客户样本数目；FN表示预测为流失状态但真实为留存状态的客户样本个数。

实验首先设置查准率和查全率作为评价指标。查准率用来统计所有预测为流失时，实际为流失的占比；而查全率为在所有实际流失中，成功预测为流失的占比。指标定义如下：

$$\text{Precision}(\text{查准率/精确率})=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

$$\text{Recall}(\text{查全率/召回率})=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

由于电信领域的客户样本普遍存在分布不均衡的特点，这对模型能否正确预测流失客户提出了更高的要求。因此，对于非平衡的数据集除了使用AUC来评估模型的分类性能外，还应使用AP和F1-score综合衡量模型对于正类的预测性能。F1-score能够同时衡量查准率和查全率，是P和R的调和平均值[17]；ROC曲线下面积即AUC；PR曲线与坐标轴围成的面积即AP，AP不受阈值划分的影响，能直观反映分类器对正类样本进行正确分类的能力。

$$\text{F1-score}=\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{5}$$

$$\mathrm{AUC}=\frac{\sum_{i\in \mathrm{PositiveClass}}\mathrm{rank}_i-\frac{M(1+M)}{2}}{M\times N} \quad (6)$$

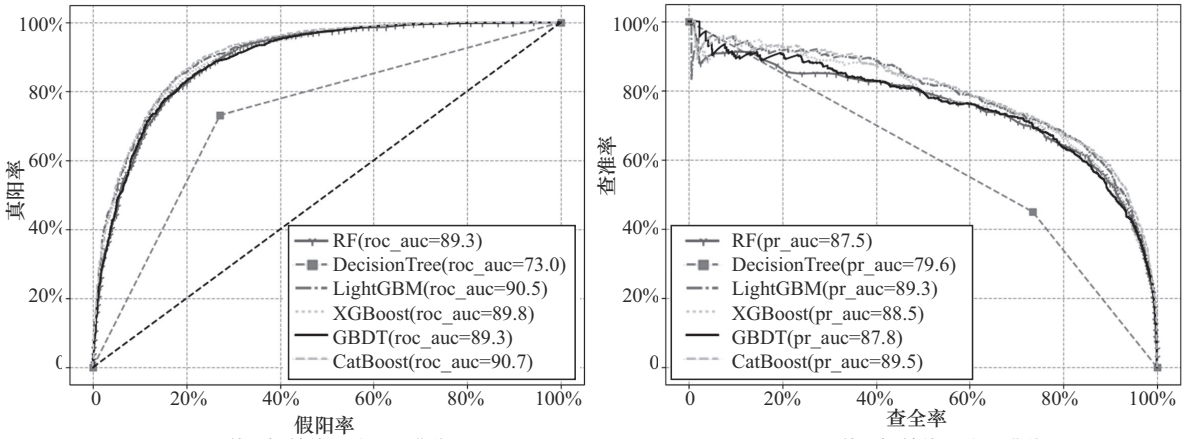
$$\mathrm{AP}=\int_0^1 \mathrm{Precision}\,(\mathrm{Recall})\,\mathrm{dRecall} \quad (7)$$

3.2　模型对比

首先将上述运营商的客户数据集通过Python的sklearn库中的train_test_split函数按8:2比例进行切分，80%作为模型的训练集，20%作为模型的测试集。

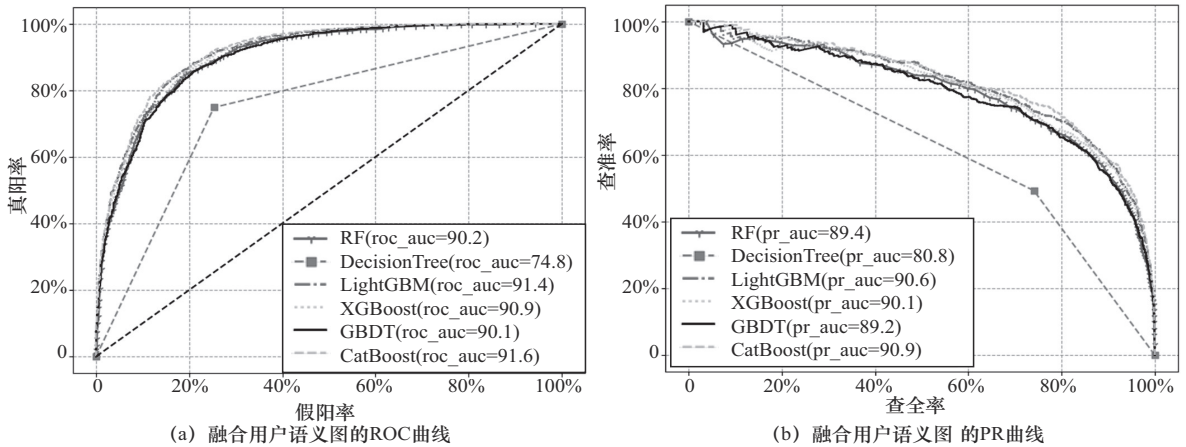
在基学习器的选择问题上，笔者也曾尝试过使用其他算法进行训练，包括多层感知机、BP神经网络[18]和逻辑回归等，但效果不佳，性能远低于其他模型。因此，本文选择sklearn及官方库中性能较好的RF、DecisionTree、LightGBM、XGBoost、GBDT和Catboost共6类常规Classifier用于训练模型，并设置了2组对比实验，一组为基于预处理后的初始特征进行模型训练，一组为融合用户语义图特征后进行模型训练，2组实验均设置相同的超参数。由于该数据集样本分布不平衡，模型的精确率等指标表现较差，因此2组实验均采用下采样策略解决正负样本分布不均匀的问题，最后通过十折交叉验证结果对比分析效果最优的算法。为直观分析预测模型性能，本文绘制了2组实验下6个分类器的ROC曲线和PR曲线。基于初始特征的模型训练曲线如图3所示，融合用户语义图的模型训练曲线如图4所示。

各种常规算法基于初始特征的模型性能见



(a) 基于初始特征的ROC曲线　　(b) 基于初始特征的PR曲线

图3　基于初始特征的模型训练曲线



(a) 融合用户语义图的ROC曲线　　(b) 融合用户语义图 的PR曲线

图4　融合用户语义图的模型训练曲线

表5，融合语义图后的模型性能见表6。

表5　初始特征的模型性能

模型	AUC	查准率	查全率	F1值	AP
RF	0.893	0.791	0.851	0.820	0.875
DecisionTree	0.730	0.720	0.723	0.721	0.796
LightGBM	0.905	0.812	0.855	0.833	0.893
XGBoost	0.898	0.805	0.851	0.828	0.885
GBDT	0.893	0.792	0.855	0.822	0.878
CatBoost	0.907	0.814	0.859	0.836	0.895

表6　融合语义图后的模型性能

融合语义图	AUC	查准率	查全率	F1值	AP
RF	0.902	0.802	0.861	0.831	0.894
DecisionTree	0.745	0.748	0.748	0.748	0.808
LightGBM	0.914	0.822	0.858	0.840	0.906
XGBoost	0.909	0.814	0.856	0.835	0.901
GBDT	0.901	0.802	0.858	0.829	0.892
CatBoost	0.916	0.817	0.858	0.837	0.909
EFML	**0.920**	**0.831**	**0.869**	**0.849**	**0.912**

通过对比表5和表6可见，融合用户语义图的CatBoost和LightGBM等模型在测试集上的查准率、查全率和F1值等5项指标均有较大程度的提升，证明了图特征提升客户流失预测精度的有效性和普适性。

为持续优化模型性能，EFML框架设计了基于MLP的元学习器，在本文实验中，MLP的输入层被设置为16个节点，以支持容纳多种模型预测结果，输出层为1个节点，中间包含2个隐藏层，各含16个节点。EFML模型精度提升对比见表7，通过观察表7的结果可见，将RF、LightGBM、XGBoost和CatBoost的预测数据送入MLP进行学习后，EFML的AUC为0.92，AP为0.912，性能指标远超其他单一模型，进而证明了MLP在整合和优化多个模型预测表示时的关键作用。

EFML模型在测试集上的混淆矩阵如图5所示，1代表预测为流失，0代表预测为留存。

树模型算法可以追踪不同特征在树节点中的

表7　EFML模型精度提升对比

对比模型	EFML模型对比指标				
	AUC	查准率	查全率	F1值	AP
RF	+0.027	+0.04	+0.018	+0.029	+0.037
DecisionTree	+0.19	+0.111	+0.146	+0.128	+0.116
LightGBM	+0.015	+0.019	+0.014	+0.016	+0.019
XGBoost	+0.022	+0.026	+0.018	+0.021	+0.027
GBDT	+0.027	+0.039	+0.014	+0.027	+0.034
CatBoost	+0.013	+0.017	+0.01	+0.013	+0.017

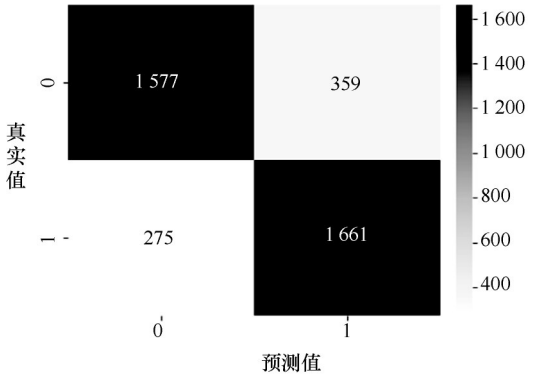


图5　EFML模型在测试集上的混淆矩阵

出现次数，并通过特征出现的频度，衡量特征其对分类的贡献程度。本文以分类性能较好的LightGBM为例，绘制了该模型的特征重要度柱状图，验证由用户语义向量分解的样本差分1（Diff_Product_1）、样本差分2（Diff_Product_2）和样本差分3（Diff_Product_3）这3个特征的重要性。LightGBM的特征重要度如图6所示。

实验发现，3个样本差分类特征的重要性分数分别为163、248和134，排名位于第7、第2和第10位，在所有参与树模型构建的特征中，重要度偏高。这表明，预测目标变量时，这3个特征能提供重要的区分信息，说明语义图与目标变量之间存在着强烈的相关性或条件依赖关系。另一方面，得益于图特征的有效融合，EFML模型的性能也得到有效提升。

3.3　应用建议

本文通过使用Matplotlib绘图工具，以重要度最高的本月费用（元）和样本差分2特征为例

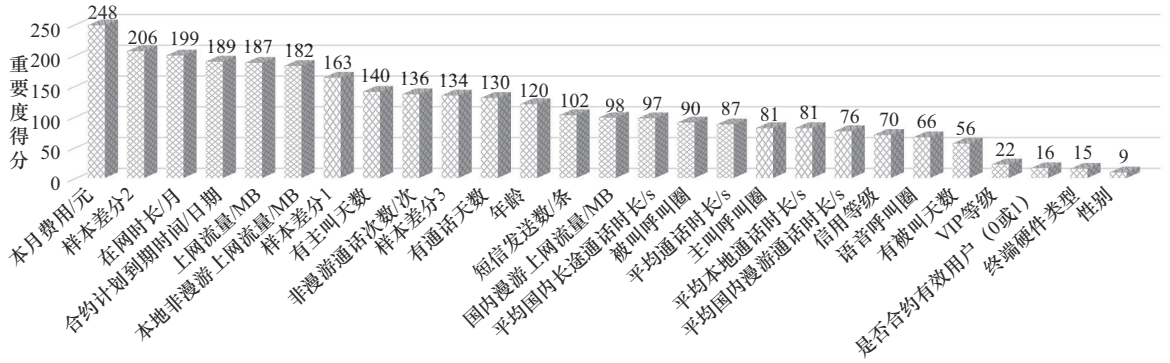


图6　LightGBM 的特征重要度

进行探索，借助散点图分析特征对客户流失的影响。本月费用/元和样本差分 2 散点分布如图 7 所示，由图 7 可知大部分流失客户的月费用较少较小，行为差异较小，被约束在相似的区间内，这说明行为差异波动较大、每月消费较高的用户更不易流失。

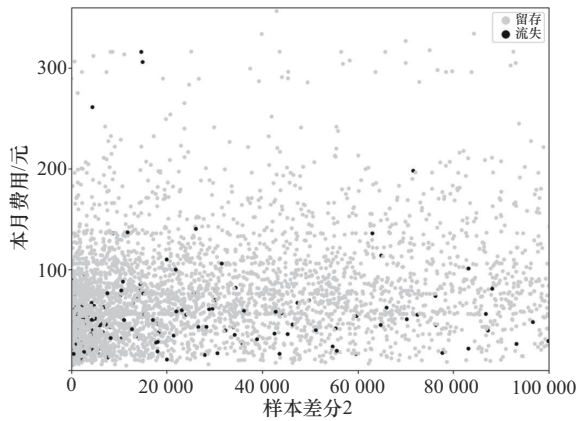


图 7　本月费用和样本差分 2 散点分布

基于 EFML 的模型分析以及运营商营运环境，提出以下几点方案。

（1）关注客户本月费用/元特征，针对经常通话、每月消费较高的用户不易流失的情况，可以通过新颖的营销方式，提升用户的消费兴趣，努力维系存量客户。

（2）针对客户在网时长(月)对用户流失概率的影响，面向不同的入网时长客户，可设计不同的方案。

（3）通过分析用户近数月来的行为动态差异，可以进一步跟踪并提供专属服务、个性化套餐等，增强用户的满意度，从而减少客户流失风险。

（4）策略调整后进行 EFML 的 A/B 测试，模拟推导策略的有效性。

4　结束语

本文针对树状结构算法较难捕获时序特征等问题，提出了一种基于 EFML 的预测流失模型。通过构造用户语义图的方式表征用户时序行为，同时引入基于 MLP 的元学习器来捕捉 LightGBM 等基础模型和预测数据间的潜在非线性关系，以提取更高层次的抽象表示，进而提高 EFML 对各类样本的鉴别能力。实验结果验证，EFML 模型结合了主流树状模型结构及神经网络结构特性，在流失预测任务中表现出优良的性能，具备较强的解释性和工程应用价值。

实验虽取得一定进展，但仍然存在一些待改进的地方：（1）模型算法优化，向技术方向继续延伸，结合数据特征尝试采用更先进、更多样的机器学习算法来挖掘客户行为信息；（2）如何与高阶特征有效结合，后续研究可以进一步融合用户的高阶特征进行分析建模；（3）探索更细粒度的数据结果可视化及分析模式，使本文方法能够在实际的运营商客户维系中得到更广泛的应用与推广。

参考文献：

[1] 阿克弘, 胡晓东. 基于GAN数据重构的电信用户流失预测方法[J]. 电信科学, 2023, 39(3): 135-142.
A K H, HU X D. GAN data reconstruction based prediction method of telecom subscriber loss[J]. Telecommunications Science, 2023, 39(3): 135-142.

[2] 黄子璇, 夏壬焕, 张雄涛. 基于注意力机制和图卷积的电信客户流失预测[J]. 计算机工程与设计, 2023, 44(6): 1685-1691.
HUANG Z X, XIA R H, ZHANG X T. Prediction of telecom customer churn based on attention mechanism and graph convolution[J]. Computer Engineering and Design, 2023, 44(6): 1685-1691.

[3] 钟保权. 基于图学习的电信用户流失预测算法[D]. 杭州: 浙江大学, 2021.
ZHONG B Q. Graph Learning Algorithm for Tele-com User Churn Prediction[D]. Hangzhou: Zhejiang University, 2021.

[4] ZHANG T Y, MORO S, RAMOS R F. A data-driven approach to improve customer churn prediction based on telecom customer segmentation[J]. Future Internet, 2022, 14(3): 94.

[5] MARTINS H. Predicting user churn on streaming services using recurrent neural networks[D].KTH Royal Institute of Technology, 2017.

[6] DE CAIGNY A, COUSSEMENT K, DE BOCK K W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees[J]. European Journal of Operational Research, 2018, 269(2): 760-772.

[7] ULLAH I, RAZA B, MALIK A K, et al. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector[J]. IEEE Access, 2019, 7: 60134-60149.

[8] HÖPPNER S, STRIPLING E, BAESENS B, et al. Profit driven decision trees for churn prediction[J]. European Journal of Operational Research, 2020, 284(3): 920-933.

[9] AHMAD A K, JAFAR A, ALJOUMAA K. Customer churn prediction in telecom using machine learning in big data platform[J]. Journal of Big Data, 2019, 6(1): 28.

[10] MENG Q, KE G L, WANG T F, et al. A communication-efficient parallel algorithm for decision tree[EB]. 2016: 1611.01276.

[11] KE G, MENG Q, FINLEY T, et al. Lightgbm: a highly efficient gradient boosting decision tree [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM Press, 2017. 3149-3157.

[12] 魏志强, 张浩, 陈龙. 一种采用SmoteTomek和LightGBM算法的Web异常检测模型[J]. 小型微型计算机系统, 2020, 41(3): 587-592.
WEI Z Q, ZHANG H, CHEN L. Web anomaly detection model using SmoteTomek and LightGBM algorithm[J]. Journal of Chinese Computer Systems, 2020, 41(3): 587-592.

[13] SHI H. Best-first decision tree learning[D]. Hamilton: The University of Waikato, 2007.

[14] SUN X L, LIU M X, SIMA Z Q. A novel cryptocurrency price trend forecasting model based on LightGBM[J]. Finance Research Letters, 2020, 32: 101084.

[15] MIRONOV S V, SIDOROV S P. Duality gap estimates for weak Chebyshev greedy algorithms in Banach spaces[J]. Computational Mathematics and Mathematical Physics, 2019, 59 (6): 904-914.

[16] CHEN T Q, GUESTRIN C. XGBoost: A scalable treeboosting system[C]//Proceedings of the 22nd ACMSIGKDD International Conference on KnowledgeDiscovery and Data Mining. San Francisco: ACM Press, 2016. 785-794.

[17] ZHENG H, LI H B, LU X J, et al. A multiple kernel learning approach for air quality prediction[J]. Advances in Meteorology, 2018: 3506394.

[18] 周艳聪, 郝园媛. 基于机器学习的运营商客户行为分析[J]. 科学技术与工程, 2022, 22(2): 585-592.
ZHOU Y C, HAO Y Y. Research of operator customer behavior analysis based on machine learning algorithm[J]. Science Technology and Engineering, 2022, 22(2): 585-592.

[作者简介]

李龙戈（1992– ），男，现就职于中国通信建设集团设计院有限公司数字化咨询中心，主要研究方向为大数据分析与处理。

郑铿城（1997– ），男，中南财经政法大学博士生，主要研究方向为大数据分析与处理。