基于集成学习的电信用户流失预测模型研究

刘 梅 马艮娟
（南京传媒学院，江苏 南京 21000）

摘要：电信市场竞争日益激烈，提前预测用户流失并制定有效的营销策略是确保用户长期稳定的重要环节。本文建立了基于集成学习的电信用户流失预测模型，并对其性能进行了评估。我们选用了 IBM 公司提供的电信用户数据集，首先进行了数据探索与分析，并根据数据特征进行了相应的数据预处理。接着，训练了 2 个深度学习模型和 12 个机器学习模型，并对其性能进行了分析评估，最终选取了表现较为优异的 LR、CatBoost、XGBoost、RF、AdaBoost、ET 和 LGM 模型作为基模型。最后，通过硬投票、软投票和 Boost 集成方法实现了多模型的融合。实验结果表明，该基于集成学习的预测模型的准确率达到了 87.02%。利用该模型预测用户流失，不仅具有显著的实用价值，还展现了广泛的应用前景。

关键词：用户流失预测；集成学习；机器学习

随着移动通信技术的迅猛发展，电信运营商用户数量迅速增长，而电信市场却逐渐趋于饱和。在此竞争激烈的环境中，电信行业面临着诸多挑战。运营商不断加大研发力度，为用户提供更多服务的同时，也为用户提供了更多的选择空间，从而导致许多运营商都面临着用户流失的问题。研究表明，提高用户保留率 5%，将使利润率提高 25%。相比之下，开发一个新用户的成本通常是维护一个老用户成本的 4~5 倍。因此，如何准确预测流失用户，制定相应的营销措施，以保持客户的长期稳定至关重要。

近年来，机器学习和深度学习方法已经被成功应用于识别客户流失问题。Xu Jiabing 等人借助反向传播神经网络（BPNN）算法建立了电信客户流失预测模型，并在 Hadoop 平台上部署了 MapReduce 编程框架。实验表明，BPNN 在某电信运行上数据集预测用户流失准确率为 82.12%，且大大缩小模型训练时间。李宏明等人提出了一种用户流失预测模型 BO-XGBoost，即 XGBoost 模型进行用户预测，同时采用贝叶斯方法优化 XGBoost 模型的调参过程，实验表明，该模型具有的泛化能力。

本文以 IBM 公司提供电信公司的用户数据集为例，首先完成数据预处理、特征工程；再对多个分类器进行训练和评估，最后采用集成学习技术的投票机制进行多模型融合，进步提高了模型预测的准确率。

一、机器学习技术

（一）梯度提升算法

该算法主要包括两个阶段：训练阶段和预测阶段。在训练阶段中，通过迭代优化损失函数，使模型逐步逼近最优解。在每轮迭代中，新的弱学习器预测前一模型的残差（即目标值与当前模型输出的差异），逐步减小整体误差。在预测阶段，将所有弱学习器的预测结果加权求和，得到最终预测结果。由于弱学习器的融合受权重影响，通常使用交叉验证等方法选择最佳的弱学习器。常见的梯度提升算法包括梯度提升决策树（Gradient Boosting Decision Tree，GBDT）、XGBoost（eXtreme Gradient Boosting）、LightGBM（Light Gradient Boosting Machine）和 CatBoost，通常用于解决回归和分类问题。

1. 梯度提升决策树。作为一种强大的集成学习方法，GBDT 以其卓越的预测精度、良好的泛化能力以及对异常值的稳健性，在机器学习领域尤其是回归和分类任务中占据着重要地位。已被广泛应用于信用评分、广告点击预测、疾病诊断等多个实际场景，成为数据科学工作者的重要工具。

2.CatBoost。CatBoost 利用哈希技术对类别型特征进行编码，避免了人工处理特征编码的烦琐过程。此外，CatBoost 还采用了动态增强正则化技术，在训练过程中能够自动调整正则化参数，以防止模型过拟合。CatBoost 适用于处理包含大量类别型特征的数据集，在实践中能够取得较好的性能和准确性。

（二）集成学习

集成学习（Ensemble Learning）是指将多个机器学习模型组合在一起，以提高预测的准确性和稳定性的技术。按照集成模型的实现方式可分为：

1. 投票（Voting）：采用多数投票或加权投票的方式，将多个基学习器的预测结果集成得到最终的预测结果。

2. 平均（Averaging）：将多个基学习器的预测结果进行平均，得到最终的预测结果。适用于回归问题。

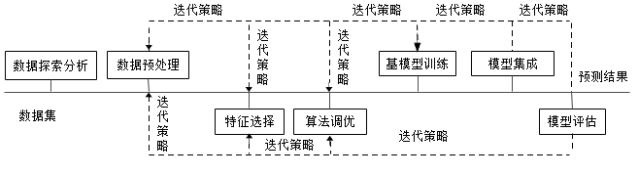
3. 堆叠（Stacking）：通过将多个基学习器的预测结果作为输入，再训练一个元学习器进行预测得到结果。

4. 装袋（Bagging）：通过随机采样生成多个子数据集，再针对不同的子数据集训练不同的基学习器，最后将所有的输出结果进行平均或投票来得到最终预测结果。随机森林就是一种基于 Bagging 思想的集成学习算法。

5. 提升（Boosting）：通过一系列迭代训练基学习器，每一轮训练的目标是修正上一轮学习器的预测错误，最终通过加权求和或级联的方式将所有的学习器组合起来，构建出更强大的整体学习器。典型的 Boosting 的算法有 GBDT 和 Catboost。

二、建模

电信用户流失预测模型的建模过程通常包括以下主要步骤：数据探索分析、数据预处理、特征选择、基模型训练、算法调优和模型评估。1）数据探索分析与预处理：对数据的类型、缺失值、分布情况等进行探索性分析，根据分析结果进行数据预处理操作，如填充缺失值、处理异常值等。2）特征工程：包括处理特征异常值、进行特征选择等操作，以提高模型性能和泛化能力。3）算法调优：通过网格搜索等技术，结合性能指标进行超参数调优，以获得最佳模型参数设置。5）基模型训练：训练多个模型，评估它们的预测准确性，并挑选出表现最佳的基模型用于集成。6）模型集成：根据模型评估结果，选择合适的集成方法，如投票、平均、堆叠等，将多个基模型集成为一个更强大的模型。在整个建模过程中，每个步骤的运作都与前后步骤相互关联，共同致力于构建一个准确并高效的用户流失预测模型。具体的流程如图 1 所示。



预测模型建模过程

三、 实验与结果

（一）数据简介与探索

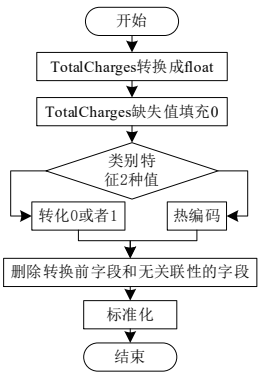
本文利用了 IBM 公司提供的某电信运营商的数据集，其中包含了用户订购的服务、账户信息以及人口学特征。该数据集共有 7043 条数据，每条数据包括 20 个特征和一个标签“Churn”（用于表示是否在最后一个月流失的用户，“Yes”表示流失，“No”表示未流失）。

进行探索性分析时，我们发现该数据集占据 7.8MB，其中只有两个特征的数据类型为 float，其余 19 个均为 object 类型。查看每个特征的唯一值数量发现，共有 18 个特征为类别数据，其中 TotalCharges 特征（入网至今的总费用）包含 6531 个值，且其范围跨度较大，另外还包含 11 个空值。进一步分析 TotalCharges 与入网时间的关系发现，这 11 个空值均对应于当月入网的用户。分析 Churn 标签后发现，流失用户占比为 26.54%。

综上所述，本次研究对象为一个数据不平衡的二分类问题。

（二）数据预处理

由数据探索分析可知，需要完成以下的操作：1）数据类型转换：需要将 TotalCharges 列数据转成 float 类型；2）缺失值填充：该数据集中 TotalCharges 的 11 个空值，用 0 值填充；3）对类别特征编码，特征的值只有 2 值分别设置为 0 和 1，当特征类别数大于 2 时，采用热编码；4）删除无效字段编号 customerID。5）数据标准化，特征 MonthlyCharges 和 TotalCharges 的值跨度较大，需要调用 StandardScaler 对其进行标准化处理。6）不平衡处理：过采样；整体的预处理的流程如图 2 所示。



预处理流程图

（三）基模型训练

基模型训练步骤包括：1）将数据集划分为训练集和测试集，比例为 8：2；2）采用网格搜索方法对超参数进行调优，以寻找到最优参数组合；3）在训练集上使用选定的超参数对模型进行训练；4）采用准确率、精准率、召回率和 F1 得分等指标对模型进行评估；5）从评估结果中选择准确率超过 85% 的模型作为基模型。本文选用了训练模型为随机森林（Random Forest，RF）、支持向量机（Support Vector Machine，SVM）、逻辑回归（Logistic Regression，LR）、K 最近邻（K-Nearest Neighbors，KNN）、朴素贝叶斯（Naive Bayes，NB）、决策树（Decision Tree，DT）、AdaBoost（Adaptive Boosting）、梯度提升树（Gradient Boosting Decision Tree，GBDT）、XGBoost（Extreme Gradient Boosting、XGB）、极端随机树（Extremely Randomized Trees，ET）、LightGBM（Light Gradient Boosting Machine）、CatBoost（Categorical Boosting）、深度神经网络（Deep Neural Network，DNN）和多层感知器（Multi-Layer Perceptron，MLP）等共 14 种模型进行训练。其中 DNN 和 MLP 为深度学习模型，其余 12 种均为机器学习模型。训练的性能指标如表 1 所示。

表 1 14 种分类器训练结果

	recall	precision	F1	accuracy
RF	85.75%	85.75%	85.75%	85.70%
SVM	59.82%	67.29%	63.34%	65.25%
LR	86.65%	85.34%	85.99%	85.83%
KNN	82.86%	74.15%	78.27%	76.91%
NB	87.68%	71.03%	78.48%	75.88%
DT	81.58%	80.19%	80.88%	80.64%
AdaBoost	87.68%	85.27%	86.46%	86.22%
GBDT	87.42%	85.77%	86.59%	86.41%
XGB	85.37%	85.97%	85.67%	85.67%
CatBoost	86.46%	87.47%	86.96%	86.99%
LGM	86.59%	86.92%	86.75%	86.73%
ET	82.67%	84.90%	83.77%	83.93%
DNN	83.91%	83.96%	83.88%	83.91%
MLP	77.58%	78.68%	77.21%	77.58%

分析表 1 可知，CatBoost、GBDT、RF、LR、AdaBoost、XGBoost 和 LightGBM 模型的平均准确率为 86.22%。其中，CatBoost 表现最佳，其准确率高达 86.99%，而 XGBoost 的准确率最低为 85.67%。七个模型的准确率均在 85% 以上，因此可选作集成学习的基模型。

（四）模型集成

本文采用软投票、硬投票和 Boosting 方式进行 7 个基模型的集成。集成模型的性能指标如表 2 所示。

表 2 集成学习模型性能指标

	recall	precision	F1	accuracy
硬投票	86.34%	86.34%	86.34%	86.34%
软投票	87.02%	87.02%	87.02%	87.02%
Boost	86.51%	86.51%	86.51%	86.51%

通过对表 2 的分析可知，采用三种模型集成方式后，准确率均优于 7 个基模型的平均准确率，表明多模型融合的效果明显优于单一模型。其中，硬投票方法的表现最佳，准确率达到 87.02%，比 7 个基模型的平均值高出 0.8%。进一步验证了模型集成在提升预测性能方面的重要性。

四、总结

本文旨在解决电信用户流失问题，这对运营商的后续发展具有重要影响。传统的用户流失分析方法依赖人工手段，存在效率低下的局限性。为此，本文引入机器学习技术，构建了一个基于集成学习的电信用户流失预测模型。本文详细描述了从数据分析、预处理、基模型训练到多模型集成的完整建模过程，并以 IBM 公司提供的电信用户数据为实验基础。最终结果显示，集成流失模型的准确率高达 87.02%，进一步验证了基于集成学习的模型在综合性能上的优势。该模型为电信运营商提供了更智能、高效的解决方案，助力其有针对性地制定营销措施，从而稳定客户群并实现持续盈利。

参考文献：

[1] 叶成，郑红，程云辉 . 基于多模型融合的流失用户预测方法 [J]. 计算机工程与科学，2019，41（11）：2027-2032

[2] 祝元丽，冯向阳，闫庆武，吴子豪 . 基于梯度提升决策树的东北黑土区农田土壤有机碳空间分异及主控因子研究 [J/OL]. 中国环境科学 .https：//doi.org/10.19674/j.cnki.issn1000-6923.20240016.004

[3] 阎馨，卓志远，屠乃威 . 基于 PCA-LDA-CatBoost 的煤与瓦斯突出预测研究 [J/OL]. 控制工程 .https：//doi.org/10.14107/j.cnki.kzgc.20230597