文章编号: 1000-5641(2020)05-0167-12

基于聚合支付平台交易数据的商户流失预测

徐一文, 黎潇阳, 董启文, 钱卫宁, 周 昉

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 在聚合支付领域, 为了减少聚合支付平台的运营成本、提高平台利润率, 要解决的一个关键问题是确保平台中达到较低的商户流失率. 本文所关注的是聚合支付平台的商户流失预测问题, 目标是帮助平台及时挽回可能流失的客户. 基于交易流水数据和商户基本信息, 本文提出了与商户流失密切相关的特征, 采用多种传统机器学习模型进行流失预测. 考虑到商户的交易流水数据具有时序性, 增加了基于 LSTM 的多种时间序列模型来建模. 在真实数据集上的实验结果表明手动提取的特征具有一定的预测能力, 结果具有可解释性; 采用时间序列模型能够较好地学习到数据的时序特征, 从而进一步提升预测结果.

关键词: 流失预测; 特征工程; 时间序列模型

中图分类号: TP399 **文献标志码**: A **DOI**: 10.3969/j.issn.1000-5641.202091016

Merchant churn prediction based on transaction data of aggregate payment platform

XU Yiwen, LI Xiaoyang, DONG Qiwen, QIAN Weining, ZHOU Fang

(*School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*)

Abstract: In the field of aggregate payments, ensuring a low dropout rate of merchants on the platform is a key issue to reduce the overall platform operating cost and increase profit. This study focuses on the prediction of merchant churn for aggregate payment platforms and aims to help the platform reactivate potential churn merchants. The paper proposes a series of features that are highly relevant to merchant churn and applies a variety of traditional machine learning models for prediction. Given that the data analyzed contains sequential information, the study, moreover, applies LSTM-based techniques to address the prediction problem. Experimental results on a real dataset show that the proposed features have a certain predictive ability and the results are interpretable. And, the LSTM-based approaches are capable of capturing the timing characteristics in the data and further improve prediction results.

Keywords: churn prediction; feature engineering; time series models

0 引 言

随着移动通信设备在我国的飞速普及, 移动支付正逐渐取代现金支付成为支付的主流方式. 常见的移动支付平台有支付宝、微信、银联等, 然而不同的用户对支付平台有不同的偏好, 在线下消费场景中, 商户需要同时安装多种支付软件来满足不同顾客的需求, 这给商户和顾客都带来了一定程度的不便, 因此聚合支付技术随之出现. 聚合支付技术支持收款码一码通用以及自动识别多种平台的付款

收稿日期: 2020-08-16

基金项目: 国家自然科学基金 (61902127); 上海市自然科学基金 (19ZR1415700)

通信作者: 周 昉, 女, 副研究员, 研究方向为数据挖掘与机器学习. E-mail: fzhou@dase.ecnu.edu.cn

码, 提供聚合支付服务的公司会从每一笔交易里收取一定比例的费率. 考虑到中国人口数量和移动支付的普及, 这是一个巨大的市场且竞争十分激烈, 目前常见的提供聚合支付服务的公司有 ping++、快钱、收钱吧、美团等.

在竞争如此激烈的市场中, 用户的留存率是一个至关重要的问题. Bhattacharya 的研究表明, 增加一个新用户的运营成本是保留一个老用户的运营成本 5 至 6 倍[1]; 根据 Reichheld 和 Detrick 的研究, 企业的客户流失率每减少 5%, 平均利润就会增加约 25%[2]. 本项研究是针对聚合支付领域的商户流失预测问题进行的. 聚合支付平台的业务部门每天需要对可能流失的用户进行回访, 越早进行回访, 成功挽回的概率就越大. 如果某商户连续两天没有任何交易, 则认为该商户有可能会流失. 在这个条件下, 聚合支付平台每天都会出现大量的疑似流失商户, 但是真正有流失倾向的只占一小部分. 如果对每个疑似流失商户都进行电话回访会带来两个问题: 一是工作量太大, 需要耗费大量的人工成本和时间成本; 二是如果频繁地对本身并无流失倾向的商户进行回访, 反而会极大地降低客户的满意度. 因此, 如何准确地预测出具有流失倾向的商户成为一个至关重要的问题.

本文关注的是具有挽回价值的疑似流失商户, 该商户需要达到一定的历史交易金额和历史交易笔数. 实验使用真实数据对疑似流失商户进行预测. 针对数据的特点, 预测任务面临如下挑战.

(1) 如何提取有效的特征来进行预测? 实验数据包含商户的基本信息和一个月的历史交易流水. 针对商户的基本信息, 需要筛选出对预测有帮助的字段作为特征. 例如, 入网时长能反映商户的忠诚度, 且加入模型后能够提高预测结果的准确性, 所以将其作为特征; 又如, 对商户的行业字段进行 one-hot 编码后, 因为太过稀疏, 入模后并不能提高模型的表现, 因此不能将其作为特征. 针对流水数据, 最直接的做法是使用交易金额和交易笔数的统计信息, 如均值、最值、中位数、标准差等作为特征, 但这样得到的预测结果并不理想. 因此我们需要设计出更加有效的特征, 使其既能体现流水数据的动态信息, 同时又具有较强的预测能力.

(2) 如何解决交易记录间的时间不规律问题? 商户不一定每天都有交易记录, 两笔交易间有可能间隔几天或几个星期. 本次实验取商户一个月时间范围内的流水记录, 有交易的天数无规律地分布在 [1, 30] 区间内. 由于商户近期的数据更能反映其流失倾向, 如果交易记录之间的时间间隔较大, 应给予近期数据更多的关注, 降低早期数据的影响. 因此在建模时需要将记录之间的时间间隔考虑在内.

针对第一个挑战, 本文不仅提出了商户信息方面的特征, 如入网时长、连锁门店数量, 以及交易流水方面的特征, 如平均交易金额、无交易天数占比, 还提出了一系列风控特征用来反映商户存在的刷单套现等异常行为, 如信用卡支付笔数占比、整 10 金额交易笔数. 本次实验将这 3 类特征输入传统机器学习模型, 验证了所提出特征的有效性.

针对第二个挑战, 常见的序列模型如 RNN、LSTM (Long Short-Term Memory)[3] 没有考虑到数据间的时间间隔. 本文首次将 T-LSTM (Time-aware LSTM) 模型[4] 用于商户流失预测任务, T-LSTM 模型能够接收时间信息, 根据输入的时间间隔对长短期记忆进行分解, 使得时间间隔越长, 前序时间节点的输入对模型的影响越小. 实验部分通过与 LSTM 和 Bi-LSTM (Bidirectional LSTM) 的对比验证了其预测结果的有效性.

本文的主要贡献如下.

(1) 提出了有效的特征用于商户流失预测, 具有可解释性, 为业务提供指导.

(2) 使用考虑了序列数据中时间间隔属性的 T-LSTM 模型, 使预测结果得到较大提升, 验证了商户流失问题中数据的时间敏感性, 为商户流失预测的研究提供了新的思路.

(3) 使用自编码器来学习商户的表征, 再根据商户的表征进行分类, 进一步提升了预测结果的准确度. 同时自编码器输出的商户表征不仅可用于分类, 也可用于聚类分析等各种任务, 为机器学习系统的构造提供了新的思路, 即可将自编码器输出的表征作为各种任务的共享数据, 提高了机器学习项

目的开发效率.

本文后续结构: 第 1 章介绍相关工作; 第 2 章介绍实验数据及问题描述; 第 3 章介绍根据商户基本信息和交易流水数据提取的 3 类特征; 第 4 章简单介绍本文所使用的各种时间序列模型; 第 5 章为实验部分, 对各种模型的实验结果进行对比分析; 第 6 章总结所做的工作并展望未来的研究方向.

1　相关工作

本章从两个方面进行介绍. 先介绍国内外在用户流失预测方向上的相关工作, 包括基于特征工程和传统机器学习模型的方法, 以及使用深度学习模型的方法. 再介绍考虑了序列数据中时间属性的相关工作.

1.1　用户流失预测

近年来用户流失预测问题在学术界引起了广泛关注, 范围涉及 MOOC 平台[5-6]、社交平台[7-9]、电信[10-12] 等多个领域. 在早期电信领域的研究中, 研究者通过领域的业务知识设计各种特征, 再使用随机森林、梯度提升树等集成模型就能获得准确度较好且具有可解释性的结果. 例如, 使用随机森林模型可获得特征的重要性系数, 系数值越高表示该特征在模型中所起的作用越大, 在业务上代表这个特征跟流失的关系越密切, 从而对业务流程起到指导作用. 同时重要性系数可用于特征选择, 对于重要性系数值太低的特征, 可认为与用户流失相关性不大, 甚至对模型结果起到反作用, 所以可将其剔除. 这种方法具有较好的可解释性, 但是, 特征选取的好坏将直接影响最终的预测结果. 随着深度学习的兴起, 神经网络模型在众多任务上的表现超过了大多数传统模型. 在 MOOC 平台用户流失预测的研究中[5], 作者通过大量的数据分析提取了有效的用户行为特征与全局环境特征, 使用 Embedding、CNN 和 Attention 机制将两种特征进行结合, 得到了较好的预测结果. 另外, 作者提到一种可用的模型集成方法, 将神经网络第 (L−1) 层的输出提取出来, 与原始特征结合放入 XGBoost[13] 等集成模型中, 可以得到更好的预测结果. 在社交应用的用户流失预测研究中[7], 作者将预测任务分为了两步, 第一步对用户进行聚类; 第二步同时训练多个 LSTM 网络, 设计损失函数将分类结果与聚类结果相关联, 极大提高了预测的准确性. 这种方法通过聚类分析使结果拥有较好的可解释性, 同时使用原始的用户行为数据训练 LSTM 模型, 避免了繁杂的特征工程.

1.2　T-LSTM

传统的 RNN、LSTM 模型并没有考虑到序列数据中隐含的时间属性, 会导致某些场景下模型不能得到理想的预测结果. 例如, 在医疗诊断预测的研究中, 病人的病历记录时间往往是不规律的, 两条记录之间的时间间隔短则几天, 长则数年, 建模时应该考虑时间的影响. 为此有研究者提出 T-LSTM 模型, 该模型修改了 LSTM 的内部结构, 将序列数据中两个相邻输入之间的时间间隔加入模型. 同时作者将其扩展为 T-LSTM 自编码器结构, 用于学习病人的表征, 可用于分类、聚类等任务. 但是各种疾病对于时间的敏感度是不同的, 流感、肠胃炎等急性病症在短时间内会造成很大影响, 但在较长的时间跨度下其影响可忽略不计, 而糖尿病等慢性疾病会持续造成影响, 为此有研究者提出 Timeline 模型[14]. Timeline 模型的主体部分采用 Bi-LSTM[15] 结构, 在序列数据输入 Bi-LSTM 之前对其进行处理, 将时间跨度和不同疾病种类对时间跨度的敏感度同时考虑在内, 并引入注意力机制为每个诊断结果分配权重, 从而得到了表现良好并具有可解释性的结果.

2　数据及问题描述

本章正式介绍实验数据和商户流失预测问题. 如图 1 所示, 本文所关注的疑似流失商户在 t_1 和 t_2 两天内无任何交易记录, 且在 t_{-2} 到 t_0 三天内有一定的交易量, 该商户在 t_2 时间点被系统检测到并标记

为疑似流失. 若在未来 28 天内, 即 t_3 至 t_{30} 时间段内, 该商户仍无交易发生, 则认为该商户已流失, 否则为未流失. 本文将根据 t_1 前 30 天, 即 t_{-29} 至 t_0 区间内的历史记录数据, 预测 t_2 后 28 天, 即 t_3 至 t_{30} 区间内商户是否会发生交易. 下文将对重要概念和数据集进行详细介绍, 并给出商户流失预测问题的正式定义.

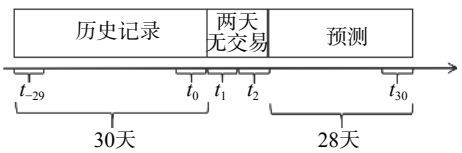

图 1 商户流失预测问题定义

Fig. 1 The definition of merchant churn prediction

2.1 概念定义

定义 1 疑似流失商户 本文考虑的是疑似流失但具有挽回价值的商户, 商户在 t_1 至 t_2 两天内没有交易, 但是在 t_{-2} 至 t_0 三天内有 10 笔以上大于 2 元的交易.

定义 2 系统预警时间点 系统预警时间点是指疑似流失商户被系统检测到的日期, 对应于图 1 中的 t_2 时间点.

定义 3 商户流失 商户流失是指疑似流失商户在 t_2 时刻被系统预警, 且后续 m 天仍无任何交易发生. 本文实验设置 $m=28$, 即 t_3 至 t_{30} 时间段内仍无交易. 因此一个商户连续 30 天无交易记录, 则认为该商户已流失.

2.2 数据介绍

本文采用上海市 2019 年 8 月至 11 月商户的数百万条支付数据作为实验数据, 数据均以二维表结构的形式存储在关系型数据库中, 商户基本信息包括商户标识码、门店标识码、省份、城市、区、入网时间、所属行业等字段, 其中同一个商户标识码可以对应多个门店标识码; 交易流水数据包括门店标识码、交易类型、支付方式、交易金额、信用卡支付金额、创建时间等字段. 对于一些入网时间非常短的商户来说, 其交易流水数据的数量较少, 无法准确描述商户行为. 因此, 本文剔除了入网时长少于 30 天以及数据项存在缺失值的商户, 并最终选取出 83198 家疑似流失商户作为实验数据. 在模型构建阶段, 将 2019 年 8 月至 10 月的疑似流失商户作为训练集 (64325 家疑似流失商户), 11 月的疑似流失商户作为测试集 (18873 家疑似流失商户), 占比接近 3∶1.

2.3 问题定义

用 x_i 表示一个疑似流失商户, $i\in\{1,2,\cdots,n\}$. 商户 x_i 根据交易流水数据和商户基本信息提取的特征向量使用 $\boldsymbol{u}_i=\{u_{i,1},u_{i,2},\cdots,u_{i,d}\}\in\mathbb{R}^d$ 表示, 其中 d 表示特征维度. 在根据商户 x_i 的交易流水数据提取的时序特征中, 用 a,c 分别表示一天的总交易金额和总交易笔数, 商户 x_i 在第 j 天的输入特征用二维向量 $\boldsymbol{v}_{i,j}=\{a_{i,j},c_{i,j}\}$ 表示, 则商户 x_i 的时序特征用 $\boldsymbol{v}_i=\{\boldsymbol{v}_{i,1},\boldsymbol{v}_{i,2},\cdots,\boldsymbol{v}_{i,l}\}\in\mathbb{R}^{l\times 2}$ 表示, 其中 l 表示序列长度. 给定 n 个疑似流失商户, $\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n\}\in\mathbb{R}^{n\times d}$ 表示用于传统机器学习模型的特征矩阵, $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}\in\mathbb{R}^{n\times l\times 2}$ 表示用于时间序列模型的特征张量. 目标向量 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_n\}\in\mathbb{R}^n$, 其中 $y_i\in\{0,1\}$, $y_i=0$ 代表商户未流失, $y_i=1$ 代表商户流失. 本文的商户流失预测问题定义为给定疑似流失商户的特征 $\boldsymbol{X}\in\{\boldsymbol{U},\boldsymbol{V}\}$, 预测其在未来时间段 ($t_3$ 至 t_{30}) 内是否会发生交易, 即学习函数:

$$f(\boldsymbol{X})\to\boldsymbol{y}.$$

3 特征描述

本章详细介绍根据交易流水和商户信息提出的特征, 这些特征将用于传统机器学习模型. 通过对

原始数据集进行数据处理和相关性分析, 本文选取能够反映商户经营状况的重要特征用于商户流失预测. 特征被划分为三组: 商户信息特征、交易流水特征、风控特征.

3.1　商户信息特征

商户的基本信息中拥有大量描述字段, 本文考虑其中能够反映商户规模、商户忠诚度和经营模式的特征, 如入网时长反映了商户的忠诚度, 而连锁门店数量反映了商户的规模, 详细描述如下.

入网时长: 入网日期是指商户开始注册使用目标产品的日期. 入网时长是指商户从入网日期到系统预警时间点的时间长度 (单位: 月), 反映了商户对于产品的忠诚度. 给定入网日期 t_{s}, 则入网时长表示为 $(t_2 - t_{\mathrm{s}})/30$.

连锁门店数量: 指该商户旗下所有门店的总数量, 反映了商户规模.

系统预警次数: 指商户在 t_{-29} 至 t_0 历史时段内被系统预警的总次数. 对于不同类型的商户, 存在不同的经营模式, 如大多餐饮门店全周经营, 每天均存在交易, 而 4S 店可能一周只有几笔交易, 会出现多次预警.

3.2　交易流水特征

商户交易流水数据中包含了大量信息, 能够反映出商户的经营状况、经营模式等, 为商户流失预测提供支持, 如商户历史时段的交易金额标准差反映了商户经营状况的稳定性, 系统预警时间点是否为节假日和双休日反映了商户的经营模式. 我们需要考虑以下交易流水信息特征, 详细描述如下.

无交易占比: 指 t_{-29} 至 t_0 历史时段内商户不存在交易的天数与总天数的比值, 可以反映商户的经营状况. 给定历史时段商户存在交易的天数为 T, 则无交易占比表示为 $(30-T)/30$.

休息日: 指系统预警时间点 t_2 是否为节假日或双休日, 反映了商户的经营模式, 如一些写字楼旁的餐铺只在工作日经营, 而大多数知名餐饮门店则全周经营. 商户如果在假期被系统预警, 可能是假期停业休息. 若系统预警时间点为节假日和双休日, 该特征取值为 1, 否则取值为 0.

平均交易笔数: 指 t_{-29} 至 t_0 历史时段内商户每天交易笔数的平均值, 该指标越大, 则商户经营状况越好. 给定交易笔数集合 $C = \{c_1, c_2, \cdots, c_T\}$, 其中 c_i 为第 i 天的交易笔数, $\forall c_i \neq 0$, 则平均交易笔数表示为 $\bar{c} = \sum_{i=1}^{T} c_i/T$.

平均交易金额: 指 t_{-29} 至 t_0 历史时段内商户每天交易金额的平均值. 给定交易金额集合 $A = \{a_1, a_2, \cdots, a_T\}$, 其中 a_i 为第 i 天的交易金额, $\forall a_i \neq 0$, 则平均交易金额表示为 $\bar{a} = \sum_{i=1}^{T} a_i/T$.

笔均交易金额: 指 t_{-29} 至 t_0 历史时段内商户每笔交易金额的平均值. 给定历史时段内总交易笔数为 $\sum_{i=1}^{T} c_i$, 交易金额为 $\sum_{i=1}^{T} a_i$, 则笔均交易金额表示为 $\sum_{i=1}^{T} a_i / \sum_{i=1}^{T} c_i$.

交易失败占比: 指 t_{-29} 至 t_0 历史时段内商户交易失败笔数与总交易笔数的比值. 当出现网络波动、机器故障时, 会出现交易失败的情况, 进而对商户的产品忠诚度造成影响. 给定历史时段交易失败总笔数为 c_{tf}, 总交易笔数为 $\sum_{i=1}^{T} c_i$, 则交易失败占比表示为 $c_{\mathrm{tf}} / \sum_{i=1}^{T} c_i$.

三天交易笔数: 指商户在 t_{-2} 至 t_0 时段内的总交易笔数, 可以反映商户近期的经营状况.

三天交易金额: 指商户在 t_{-2} 至 t_0 时段内的总交易金额.

平均交易笔数比较: 指三天平均交易笔数与历史平均交易笔数的比值取对数. 设商户在 t_{-2} 至 t_0 时段内的三天平均交易笔数为 $\bar{c}_{\mathrm{t}}$, 则平均交易笔数比较表示为 $\log(\bar{c}_{\mathrm{t}}/\bar{c})$.

平均交易金额比较: 指三天平均交易金额与历史平均交易金额的比值取对数. 设商户在 t_{-2} 至 t_0 时段内的三天平均交易金额为 $\bar{a}_{\mathrm{t}}$, 则平均交易金额比较表示为 $\log(\bar{a}_{\mathrm{t}}/\bar{a})$.

夜间交易占比: 指 t_{-29} 至 t_0 历史时段内商户夜间 (00: 00 ~ 06: 00) 总交易笔数与总交易笔数的比值, 设夜间总交易笔数为 $\tilde{c}$, 则夜间交易占比表示为 $\tilde{c} / \sum_{i=1}^{T} c_i$.

金额标准差: 指 t_{-29} 至 t_0 历史时段内商户每天交易金额的标准差, 反映商户经营状况的稳定性. 金额标准差表示为 $\sqrt{\frac{1}{T}\sum_{i=1}^{T}(a_i-\bar{a})^2}$.

3.3 风控特征

商户交易流水数据中隐藏了大量与风控有关的信息. 这些信息反映了商户的异常行为, 如商户信用卡支付笔数及金额占比, 反映了商户的信用卡支付情况, 通过对该指标的分析可判断商户是否存在刷单套现等不良行为. 关于风控信息特征的详细描述如下.

信用卡支付笔数占比: 指 t_{-29} 至 t_0 历史时段内, 商户的所有收款记录中使用信用卡支付的笔数与总交易笔数的比值, 如果占比较高, 则可能存在刷单套现等行为. 给定信用卡支付笔数 c_{r}, 信用卡支付笔数占比表示为 $c_{\mathrm{r}}/\sum_{i=1}^{T}c_i$.

信用卡支付金额占比: 指 t_{-29} 至 t_0 历史时段内, 商户的所有收款记录中使用信用卡支付的金额与总交易金额的比值. 给定信用卡支付金额 a_{r}, 信用卡支付金额占比表示为 $a_{\mathrm{r}}/\sum_{i=1}^{T}a_i$.

整 10 金额交易笔数: 指 t_{-29} 至 t_0 历史时段内满足交易金额是整十数的总交易笔数, 整 10 金额数量越多, 则越可能存在严重的刷单行为.

高额信用天数: 指 t_{-29} 至 t_0 历史时段内满足信用卡支付大于 2000 元的总天数. 该指标越大, 商户越可能存在刷单套现行为.

高额信用笔数: 指 t_{-29} 至 t_0 历史时段内满足信用卡支付大于 2000 元的总笔数.

4 模型介绍

根据上文对商户流失问题和数据特征的描述, 本文实验部分将选取两类模型: 第一类为传统机器学习模型, 包括逻辑回归、随机森林、XGBoost; 第二类为序列模型, 包括 LSTM、Bi-LSTM, 以及考虑时间间隔的 T-LSTM. 下面将对第二类模型进行简单介绍.

4.1 LSTM

LSTM 是一种特殊的 RNN, 主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题而提出的. LSTM 能够在长序列中有更好的表现, 其内部各个单元的计算方法如下. (公式中, 符号“·”表示点乘, “*”表示数乘.)

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f),$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i),$$

$$\tilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c),$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t,$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o),$$

$$h_t=o_t*\tanh(C_t).$$

f_t 是遗忘门控, 用于对上一个节点的输入 C_{t-1} 进行选择性遗忘; i_t 是输入门控, 用于对当前阶段的候选状态 $\tilde{C}_t$ 进行选择性记忆; f_t 和 i_t 共同作用, 得到当前阶段的长期记忆 C_t; o_t 是输出门控, 决定 C_t 中哪些部分将会作为当前状态的输出. 与普通 RNN 类似, 输出 y_t 往往也是由 h_t 通过 sigmoid 或者 softmax 等变换得到的.

4.2　Bi-LSTM

传统的 LSTM 只能依据之前时刻的时序信息来预测下一时刻的输出, 但在有些问题中, 当前时刻的输出不仅和之前的状态有关, 还可能和未来的状态有关. Bi-LSTM 由两个 LSTM 叠加在一起, 第一层从左边作为系列的起始输入, 第二层从右边作为系列的起始输入, 反向做与第一层一样的处理, 最后的输出由这两个 LSTM 的状态共同决定, 真正做到了结合上下文信息进行预测.

4.3　T-LSTM

上节中介绍的 LSTM 以及 Bi-LSTM 都没有考虑到序列之间的时间间隔, 为此有研究者提出了 T-LSTM 模型用以解决该问题. T-LSTM 的主要思想为将记忆状态分解为短期记忆和长期记忆, 根据输入之间的时间间隔调整短期记忆的影响, 时间间隔越长, 短期记忆的影响越小. 接着将调整后的短期记忆与长期记忆重组为新的记忆状态. T-LSTM 的模型结构如图 2 所示, 其各个内部单元的计算方法如下.

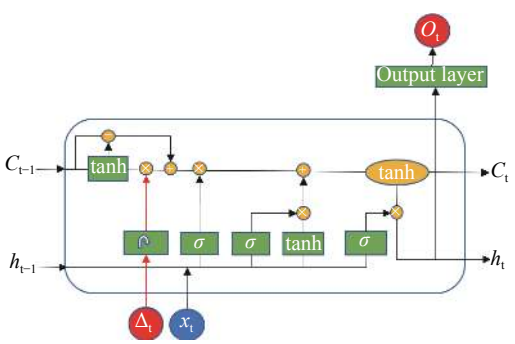

图 2　T-LSTM 模型结构

Fig. 2　The structure of T-LSTM

$$C_{t-1}^{S}=\tanh(W_d\cdot C_{t-1}+b_d),$$

$$\widehat{C}_{t-1}^{S}=C_{t-1}^{S}*g\left(\Delta_t\right),$$

$$C_{t-1}^{T}=C_{t-1}-C_{t-1}^{S},$$

$$C_{t-1}^{*}=C_{t-1}^{T}+\widehat{C}_{t-1}^{S},$$

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f),$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i),$$

$$\tilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c),$$

$$C_t=f_t*C_{t-1}^{*}+i_t*\tilde{C}_t,$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o),$$

$$h_t=o_t*\tanh\left(C_t\right).$$

C_{t-1}^{S}是由上一时刻的记忆状态C_{t-1}分解得到的短期记忆, 其参数在训练过程中根据数据自动习得. C_{t-1}^{T}表示长期记忆, 由记忆状态C_{t-1}减去短期记忆得到. Δ_t表示时间间隔, $g\left(\Delta_t\right)$表示对时间间

隔的映射函数, 一般可取 $g(\Delta_t)=1/\Delta_t$ 或 $g(\Delta_t)=1/\log(\mathrm{e}+\Delta_t)$. $\widehat{C}^S_{t-1}$ 表示经时间权重更新后的短期记忆, C^*_{t-1} 为长期记忆与更新后的短期记忆之和, 表示更新后的记忆状态, 用于替换原始的 C_{t-1}, 其余门控单元的机制和算法与传统的 LSTM 一致. 由此便得到了具有时间间隔敏感性的 LSTM 网络.

4.4 自编码器

自编码器的目标是学习函数 $h(x)\approx x$, 也就是要学习一个近似的恒等函数, 使得输出 $\hat{x}$ 近似等于输入 x, 训练的损失函数一般可用均方误差函数 (MSE). 自编码器通常分为两个部分: 编码器和解码器. 在编码器部分输入原始特征, 经过学习后得到一个 Embedding 表示, 再将 Embedding 输入解码器, 层层解码还原出原特征. 模型训练完毕后, 可用单独的编码器部分编码出数据的表征进行各种机器学习任务. 常见的自编码器有 RNN 自编码器[16]、LSTM 自编码器[17] 等. 本文的实验部分将加入 LSTM、Bi-LSTM、T-LSTM 对应的自编码器, 学习商户表征再进行预测. 实验证明这种方法能够提升预测结果.

5 实 验

5.1 实验环境及数据处理

本文根据真实的交易流水数据提取特征, 特征入模前需要进行进一步处理. 与交易金额和交易笔数相关的特征, 由于数值较大, 需要取对数, 然后对所有特征进行归一化操作, 可加快模型的收敛速度. 实验使用 Python 语言编写代码, 操作系统为 CentOS 7.0, 机器配置为 32 核 2.1 GHZ Intel Xeon E5 处理器, 256 G 物理内存, Tesla P4 显卡.

5.2 度量指标

预测问题中对实验结果有两个要求: 一是预测出尽可能多的流失商户, 即召回率要高; 二是预测为流失的商户中, 真流失的商户占比要尽可能地多, 即精确率要高. 为了同时兼顾这两个要求, 我们采用 AUC 作为模型的度量指标. AUC 被定义为 ROC 曲线下与坐标轴围成的面积, 坐标轴横轴为假阳率, 表示所有真实类别为 0 的样本中预测类别为 1 的概率; 纵轴为真阳率, 表示所有真实类别为 1 的样本中预测类别为 1 的比例. AUC 值越接近 1, 代表模型的整体性能越好.

5.3 基于特征工程的实验

为了验证所提取的特征 $\boldsymbol{U}$ 的有效性, 本文对特征的流失相关性进行了分析, 采用传统机器学习模型逻辑回归、随机森林和 XGBoost 进行预测, 并进行结果对比.

5.3.1 特征分析

为了验证特征的有效性, 本文对特征进行了数据分布和流失相关性两方面的分析. 首先使用箱图观察流失样本与未流失样本的数据分布 (见图 3—图 5), 若两者间差异较为显著, 说明该特征具有较好的区分性; 然后计算特征与样本类别之间的皮尔逊相关系数 (见表 1), 得到特征与流失之间的相关性, 由相关性进一步分析业务上的解释. 由于篇幅限制, 本文将仅展示对以下 3 个特征的分析.

系统预警次数 在流失样本和未流失样本中, 系统预警次数的分布如图 3 所示. 由图 3 可知未流失样本的上四分位数和上边缘都显著大于流失样本, 因此该特征具有一定的区分性. 表 1 中系统预警次数与流失的相关性系数为–0.19, 呈现负相关, 即系统预警次数越大, 该商户越倾向于未流失. 经过分析, 这是由于某些商户具有特殊的经营模式, 会经常性地暂停营业, 因此虽然被系统频繁预警, 但并不会流失. 例如, 学校周边的快餐店往往周末两天停止营业, 从而每周都会被系统预警.

无交易天数占比 由图 4 可知, 未流失样本中位数显著大于流失样本, 因此该特征具有一定的区分性. 表 1 中无交易天数占比与流失的相关性系数为–0.11, 该特征与流失呈负相关, 即无交易天数占比

越高, 越倾向于未流失. 原因与系统预警次数中的情况相同, 由于部分商户的经营模式不同, 会经常性地暂停营业, 造成一个月内无交易天数较多.

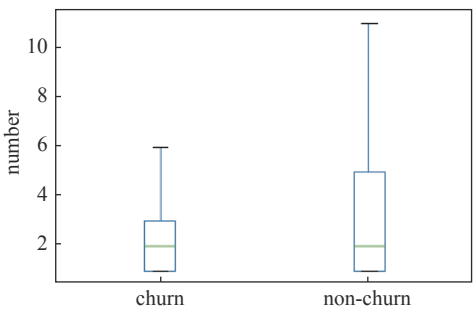

图 3　特征分析: 系统预警次数

Fig. 3　Feature analysis: number of system warnings

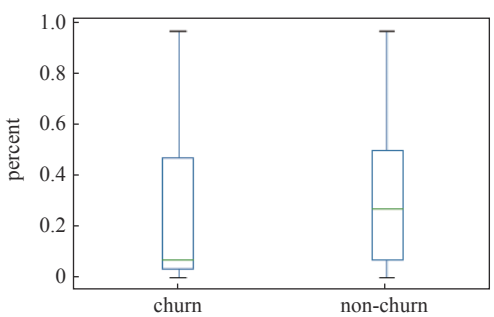

图 4　特征分析: 无交易天数占比

Fig. 4　Feature analysis: percentage of no trading days

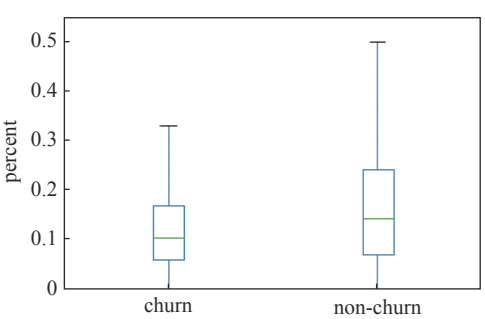

图 5　特征分析: 信用卡支付金额占比

Fig. 5　Feature analysis: percentage of amount paid by credit card

表 1　特征与流失的相关系数

Tab. 1　The correlation coefficient between features and churn

特征	相关性系数
系统预警次数	–0.19
无交易天数占比	–0.11
信用卡支付金额占比	–0.14

信用卡支付金额占比 由图 5 可知, 未流失样本的中位数、上四分位数、上边缘均大于流失样本且差异较为显著, 因此该特征具有一定的区分性. 表 1 中信用卡支付金额占比与流失的相关性系数为

–0.14, 该特征与流失呈现负相关, 信用卡支付金额占比越高的商户越不倾向于流失. 业务上认为, 这是由于这类商户具有比较稳定的客源, 一般经营状况良好, 所以不容易流失.

5.3.2 传统机器学习模型实验效果

为了验证所提取的特征能够有效地进行预测, 本文选取逻辑回归、随机森林和 XGBoost 这 3 种传统机器学习模型进行实验. 实验在训练集上使用 5 折交叉验证对各模型进行超参数调优. 逻辑回归模型的超参数设置为: 正则化方法 penalty = 'L2', 正则项系数 C = 0.1, 最大迭代次数 max_iter = 1 000, 类别权重 class_weight = 'balanced', 其余均使用默认参数. 随机森林模型的超参数设置为: 决策树数量 n_estimators = 200, 最大树深度 max_depth = 3, 类别权重 class_weight='balanced', 其余均使用默认参数. XGBoost 模型的超参数设置为: 决策树数量 n_estimators = 200, 最大树深度 max_depth = 3, 类别权重 scale_pos_weight = 2.5, 其余均使用默认参数.

各模型所得结果如图 6 所示. 由于逻辑回归是简单的线性模型, 因此得到 3 个模型中最低的 AUC 值 0.747, 随机森林和 XGBoost 的 AUC 值分别为 0.779 和 0.789. 与随机森林相比, XGBoost 具有以下 3 个优点: ①在损失函数里引入了正则项控制模型复杂度; ②优化时对损失函数进行了 2 阶泰勒展开; ③使用 shrinkage 机制防止过拟合, 因此取得了 3 个模型中最好的结果. XGBoost 支持查看特征的重要性系数, 表 2 列举了重要性系数最高的 4 个特征和对应的相关系数. 无交易天数占比获得了最高的重要性系数, 其相关性和业务解释在上文中已做了详细分析; 休息日获得了第二高的重要性系数, 与类别的相关系数为–0.14, 呈现负相关, 这表示若商户的系统预警时间点为休息日, 更可能是由于假期或周末暂停营业, 倾向于未流失; 连锁门店数量的重要性系数排第三位, 呈现负相关性, 其特征反映的是商户的经营规模, 值越高表示商户规模越大, 经营状况更加平稳, 更不倾向于流失; 平均交易笔数与流失呈现正相关性, 表示该值越大, 则越倾向于流失. 这是由于数据集中餐饮行业的占比较大, 经过数据分析, 餐饮行业的流失率远大于其他行业, 而餐饮行业表现为高频次的交易, 所以造成该特征与流失的正相关性.

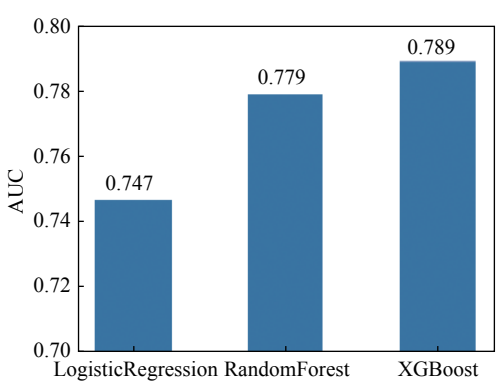

图 6 传统机器学习模型实验结果

Fig. 6 Results of traditional machine learning models

表 2 特征的重要性系数和相关系数

Tab. 2 The importance coefficient and correlation coefficient of features

特征	重要性系数	相关系数
无交易天数占比	0.204	–0.11
休息日	0.178	–0.14
连锁门店数量	0.082	–0.06
平均交易笔数	0.074	0.02

5.4　基于时序数据的实验

为了验证交易流水数据的时序性以及对时间间隔的敏感性, 本文将时序特征 $\boldsymbol{V}$ 输入一系列时间序列模型进行了实验. 模型选择 LSTM、Bi-LSTM、T-LSTM 及对应的自编码器, 为公平起见, 各模型使用了相同的参数设置. 在 LSTM、Bi-LSTM、T-LSTM 中, 递归网络中隐藏层单元数 units = 64, 激活函数 activation = 'relu', 返回序列 return_sequence = False, 再通过一层全连接网络输出预测结果, 激活函数 activation = 'sigmoid', 模型的损失函数 loss = 'binary_crossenrtopy', 优化函数 optimizer = 'adam'. 在自编码器模型中, 编码单元参数与上述网络一致, 解码单元参数为隐藏层单元数 units = 64, 激活函数 activation = 'relu', 返回序列 return_sequence=True, 通过层封装器 TimeDistributed 结合全连接层 Dense 输出 30×2 维的序列数据. 模型的损失函数 loss='mse'. 使用自编码器得到的表征进行分类时, 模型选用一层的全连接网络, 激活函数 activation = 'sigmoid', 损失函数 loss = 'binary_crossenrtopy', 优化函数 optimizer = 'adam'. 对所有模型, 训练的迭代次数 epochs = 25, 块大小 batch_size = 128.

模型所得结果如图 7 所示. 根据结果可以看出, 在仅使用交易金额和交易笔数这两个特征的情况下, 所有模型的表现均优于使用了包含流水信息和商户信息的逻辑回归模型, 其中 LSTM 自编码器和 Bi-LSTM 自编码器的预测结果均优于 XGBoost 模型. 结果表明仅使用时序数据就能够产生较好的预测结果. 在 LSTM、Bi-LSTM、T-LSTM 模型中, T-LSTM 模型取得最高的 AUC 值, 验证了实验数据中时间间隔敏感性的假设, 将时间间隔特征加入模型能够有效地提升预测结果的准确性. LSTM 自编码器、Bi-LSTM 自编码器和 T-LSTM 自编码器的预测结果均优于原始的 LSTM、Bi-LSTM、T-LSTM 模型, 这表明使用自编码器结构能够更好地学习到序列数据的表征. 与图 6 相比, 本实验不仅提高了预测结果的准确性, 同时还避免了繁杂的特征工程, 为商户流失预测的研究提供了可行的方法.

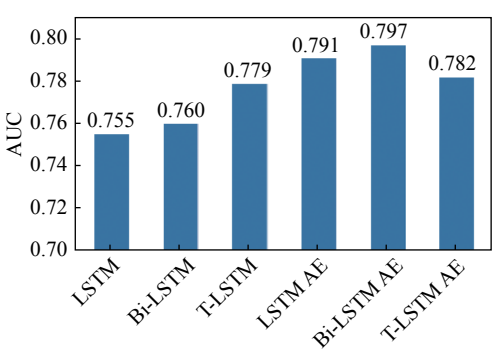

图 7　时间序列模型实验结果 (AE 表示自编码器)

Fig. 7　Results of sequential models (AE denotes AutoEncoder)

5.5　基于组合特征的实验

为了能进一步提高预测结果的准确性, 本文尝试将特征工程方法和序列模型方法相结合. 本实验将特征 $\boldsymbol{V}$ 与自编码器学习的表征 (记为 $\boldsymbol{R}$) 拼接在一起形成新的特征矩阵 $[\boldsymbol{V}, \boldsymbol{R}]$, 再将这个特征矩阵放入 XGBoost 模型中进行最终的预测, 实验结果见图 8. 由图 8 可知模型效果略有提升. 这表示可以采用模型集成的方法, 将自编码器学习出的表征与其他特征相结合, 从而进一步提升预测结果.

6　总　　结

为了准确预测流失商户, 本文根据商户信息和交易流水数据提出了与商户流失相关的 3 类特征, 实验证明所提特征既能获得较好的预测结果, 又具有可解释性. 同时本文根据交易流水数据提取了时序特征, 使用时间序列模型获得了更好的预测结果. 最后使用了模型集成的方法将两种方案结合, 进

一步提升了预测结果的准确性. 在未来的研究中, 考虑将行业信息引入时间序列模型, 使模型同时拥有时间敏感性和行业敏感性, 并将考虑时间序列模型在商户流失预测中的可解释性.

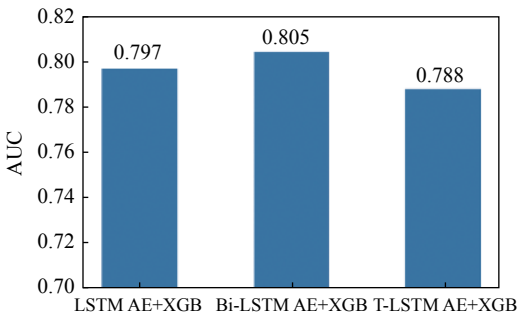

图 8 模型集成实验结果 (AE 表示自编码器)

Fig. 8 Results of model ensembles (AE denotes AutoEncoder)

[参 考 文 献]

[1] BHATTACHARYA C B. When customers are members: Customer retention in paid membership contexts [J]. Journal of the Academy of Marketing Science, 1998, 26(1): 31-44.

[2] REICHHELD F, DETRICK C. Loyalty: A prescription for cutting costs [J]. Marketing Management, 2003, 12(5): 24-24.

[3] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] BAYTAS I M, XIAO C, ZHANG X, et al. Patient subtyping via time-aware LSTM networks [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 65-74.

[5] FENG W, TANG J, LIU T X. Understanding dropouts in MOOCs [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 517-524.

[6] FEI M, YEUNG D Y. Temporal models for predicting student dropout in massive open online courses [C]// 2015 IEEE International Conference on Data Mining Workshop. IEEE, 2015: 256-263.

[7] YANG C, SHI X, JIE L, et al. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 914-922.

[8] LU Y, YU L, CUI P, et al. Uncovering the co-driven mechanism of social and content links in user churn phenomena [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 3093-3101.

[9] XIE Y, LI X, NGAI E W T, et al. Customer churn prediction using improved balanced random forests [J]. Expert Systems with Applications, 2009, 36(3): 5445-5449.

[10] WEI C P, CHIU I T. Turning telecommunications call details to churn prediction: A data mining approach [J]. Expert Systems with Applications, 2002, 23(2): 103-112.

[11] DASGUPTA K, SINGH R, VISWANATHAN B, et al. Social ties and their relevance to churn in mobile telecom networks [C]// Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology. 2008: 668-677.

[12] HUANG Y, ZHU F, YUAN M, et al. Telco churn prediction with big data [C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015: 607-618.

[13] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[14] BAI T, ZHANG S, EGLESTON B L, et al. Interpretable representation learning for healthcare via capturing disease progression through time [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 43-51.

[15] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5/6): 602-610.

[16] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. (2014-09-03) [2020-07-05]. https://arxiv.org/pdf/1406.1078v3.pdf.

[17] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using LSTMs [C]// International Conference on Machine Learning. 2015: 843-852.

(责任编辑: 林 磊)