Service Science and Management 服务科学和管理, 2023, 12(1), 37-47
Published Online January 2023 in Hans. https://www.hanspub.org/journal/ssem
https://doi.org/10.12677/ssem.2023.121005

Hans汉斯

结合LightGBM与SHAP的家政服务员离职预测方法研究

刘峰涛，刘雅琦，王扶东

东华大学旭日工商管理学院，上海

收稿日期：2022年12月10日；录用日期：2023年1月10日；发布日期：2023年1月18日

摘　要

针对互联网环境下家政服务员人力资源管理场景的变化，本文将LightGBM算法与SHAP模型结合，形成解决互联网背景下家政服务员离职问题的集成方法。以企业真实数据为研究对象，经数据预处理后建立LightGBM模型进行预测，并与KNN、逻辑回归、决策树、随机森林和GBDT算法对比，结果表明，LightGBM模型的准确率、F1值与AUC值分别为81.23%、84.41%和86.5%，优于其他算法。最终使用SHAP模型分析影响员工离职的重要因素，以此增强模型的可解释性，为企业管理者进行决策提供依据。

关键词

机器学习，LightGBM，SHAP，员工离职

A Study on the Combined LightGBM and SHAP Approach for Predicting Domestic Helper Turnover

Fengtao Liu, Yaqi Liu, Fudong Wang

Glorious Sun School of Business and Management, Donghua University, Shanghai

Received: Dec. 10th, 2022; accepted: Jan. 10th, 2023; published: Jan. 18th, 2023

Abstract

In response to the changes in the human resource management scenario of domestic helpers in the Internet environment, this paper combines the LightGBM algorithm with the SHAP model to form an integrated approach to solve the problem of domestic helpers' leaving in the Internet environment. Using real data from enterprises as the research object, the LightGBM model was es-

文章引用: 刘峰涛, 刘雅琦, 王扶东. 结合 LightGBM 与 SHAP 的家政服务员离职预测方法研究[J]. 服务科学和管理, 2023, 12(1): 37-47. DOI: 10.12677/ssem.2023.121005

tablished for prediction after data pre-processing, and compared with KNN, Logistic Regression, Decision Tree, Random Forest and GBDT algorithm, and the results showed that the accuracy, F1 value and AUC value of LightGBM model were 81.23%, 84.41% and 86.5% respectively, which were better than other algorithms. Finally, the SHAP model is used to analyze the important factors influencing helper turnover, thus enhancing the interpretability of the model and providing a basis for corporate managers to make decisions.

Keywords

Machine Learning, LightGBM, SHAP, Helper Turnover

1. 引言

近年来，我国家政服务业的需求呈现出刚性增长的态势。中国家政服务业市场规模仅用 5 年时间就从 2776 亿元提升至 8782 亿元[1]，其市场总规模保持 20%左右的增速[2]。在信息时代和知识社会的背景下，家政服务行业也逐步与互联网相融合，进行优化升级转型，形成新的发展生态。面对急速扩大的市场规模和新态势下的商业模式，家政服务行业内员工流失率较高、市场供需失衡等问题逐步显现，在一定程度上制约了行业的进一步发展。经走访调研得知，行业内家政服务员流失率在 35%左右。身处一线直面客户的员工稳定性差，导致公司运力不足、客户流失、业绩下滑，成为了制约家政服务行业继续发展的难点与痛点。

互联网时代的到来，致使家政服务行业的人力资源管理场景发生了重大变化。从契约关系角度来看，传统背景下，企业员工与企业签订劳动合同，构成强有力的契约关系，但互联网背景下，家政服务员无需签订劳动合同，且部分员工是以兼职工作者的身份加入企业，导致契约关系变弱；从管理范围角度来看，通常而言，员工与管理者同处一个工作空间，管理者能够密切关注员工动态，但互联网背景下，家政服务员的工作地点脱离了管理者管理范围，管理效力下降，管理者无法第一时间发现员工的异常；从管理模式角度来看，以往员工相关的数据获取较为困难，企业在没有大量数据支撑的情况下，处于“靠人监督、靠人管理”的管理模式，然而互联网背景下，企业内部数字信息剧增，为提升管理效率与管理精确度，企业正逐步向数字化管理模式转变。以上特点使得互联网家政企业内的管理者与家政服务员的关系从以往的强管理关系转变为弱管理关系，管理模式从“靠人管理”转变为数字化管理，这也意味着传统的离职管理方法的失效。

员工离职问题作为人力资源管理领域的核心问题，受到众多学者的关注，其研究使用的科学方法，也经历了定性分析[3] - 调查问卷与实证研究 - 机器学习路径的演变。目前，传统的研究离职问题的方法通常选择填写调查问卷的方式收集数据，进而开展实证研究[4] [5] [6]，一些学者在实证研究的部分使用了多元回归分析[7]、层次多元回归分析[8]来深入探讨影响员工离职的因素。由于该方法的数据来源为人工填写，故所获取到的样本量与数据量较小，同时，特征维度大多也由学者预设而来，无法适用于新场景下的离职问题研究。近年来，部分学者将机器学习的方法引入员工离职领域，使用了支持向量机算法(SVM) [9]、决策树算法(Decision Tree) [10]、随机森林算法(Random Forest) [11]、XGBoost 算法[12]等算法建立了员工离职预测模型，并取得良好的效果。但家政服务员的数据集中不仅包含性别、年龄等静态

数据，还包括每月请假情况、接单情况、获得评价情况等动态数据，导致数据维度较高，数据量大，传统算法处理此类数据效率较低；另外，机器学习算法作为“黑箱模型”，无法有效解释具体特征对员工离职的影响程度及特征间的交互关系，因此模型可解释性较差。

为更好地解决以上问题，本文选取 LightGBM 算法构建员工离职预测模型，该算法在处理样本多、维度高的预测任务时，展现出准确率高、速度快的优势，解决了传统算法在处理海量数据时时间复杂度高、占用内存大、耗时长的问题[13]。以企业真实数据为基准点，将 LightGBM 算法模型与其他机器学习主流算法进行对比，从而验证模型的预测效果。同时，为提高模型的可解释性，使用 SHAP 模型对影响家政服务员离职的因素进行分析，全面展示特征因素的影响力大小与特征之间的交互关系。本文将 LightGBM 算法与 SHAP 模型结合，形成解决互联网背景下家政服务员离职问题的集成方法，为企业管理者提前预判家政服务员的离职状态、前置管理干预行为提供依据，以稳定员工队伍、满足市场需求。

2. 模型方法

本文提出的结合 LightGBM 算法与 SHAP 模型的集成模型方法流程图如图 1 所示，其中，模型主要包括了数据预处理、模型超参数寻优、模型对比分析、模型解释分析等核心步骤。

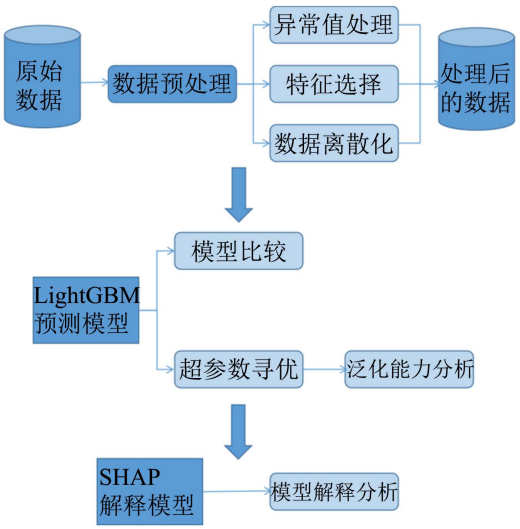


Figure 1. Flow chart of the integrated model approach
图 1. 集成模型方法流程图

2.1. LightGBM 算法

LightGBM (Light Gradient Boosting Machine Method)算法是对 GBDT (Gradient Boosting Decision Tree)算法的改进与升级，它是以决策树为基学习器的分布式梯度提升框架。LightGBM 算法的内存占用低、运行效率高、准确率高、支持多维度并行等特征十分契合互联网环境下的家政服务员的数据特性。自 2017 年算法发布后，LightGBM 算法已被运用于医药[14]、交通[15]、商业预测[16]等领域，并表现出良好的性能。

2.1.1. GBDT 算法

GBDT 算法是一种采用了 Boosting 思想的集成学习算法。该算法是以决策树为弱学习器的加法模型，它将多个弱学习器线性组合，最终得到一个强学习器，其公式如下所示：

$$T(x)=T_0(x)+\sum_{m=1}^{M}\sum_{j=1}^{J}a_{m,j}I\left(x\in R_{m,j}\right) \tag{1}$$

其中，a 为使得损失函数最小的值，m 为决策树个数，j 为第 m 颗决策树叶子结点的个数，$R_{m,j}$ 为决策树对应的叶子结点区域。

GBDT 算法使用前向分布算法与梯度提升算法进行训练，其算法核心是将上一步学习器的预测值与实际值的残差作为目前学习器的特征值进行迭代拟合，使得残差逐步减小，逼近真实值。在 GBDT 算法中，残差即为损失函数的负梯度，如公式(2)所示：

$$r_{m,i} = -\frac{\partial L\left[y_i, T_{m-1}\left(x_i\right)\right]}{\partial T_{m-1}\left(x_i\right)} \tag{2}$$

2.1.2. 直方图算法

传统的基于 GBDT 算法的模型在分裂决策树时，通常使用穷举法寻找最佳分裂点，这种遍历每个数据样本来计算分裂收益的方法虽然很精确，但是代价非常大。直方图算法则是将特征维度中精确的浮点数离散化映射到 k 个区间内，分为 k 个整数；同时构造具有 k 个箱的直方图，每个箱的值为数据离散化后落入该区间的数量，最终依据直方图遍历数据，计算每个箱的样本数、梯度等来寻找最佳分裂点。

在 LightGBM 算法中引入直方图算法能够有效降低计算内存，离散后的整数型数据占用空间相对较小，且不需要对结果进行排序与储存，从而可以降低内存的消耗。此外，相对于计算每个变量所有特征值的分裂收益的预排序算法而言，直方图算法仅需计算所有箱的分裂收益即可，大大降低了计算复杂度。

2.1.3. 有深度限制的 Leaf-Wise 叶子生长策略

一般而言，决策树通常使用按层生长(Level-wise)的生长策略，如图 2 所示。按层生长需要遍历所有数据，对决策树同一层的叶子不加区分地进行分裂。该方法的缺点在于忽视了有些叶子结点分裂增益较小、无需再次进行分裂的情况，冗余计算较多，运行效率低。

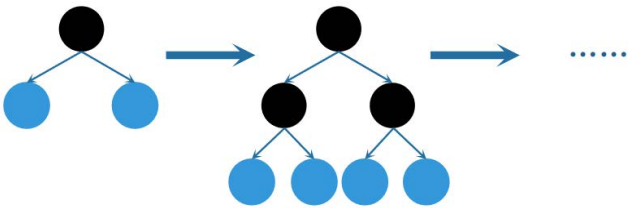


Figure 2. Level-wise growth strategy
图 2. Level-wise 生长策略

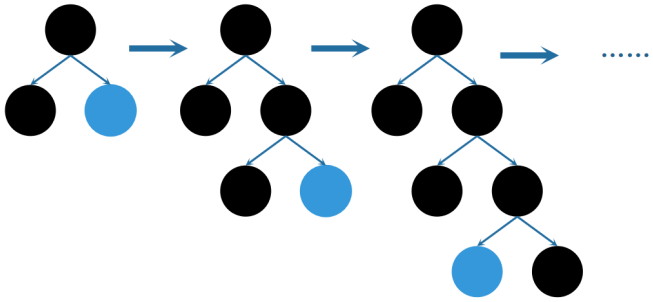


Figure 3. Leaf-wise growth strategy
图 3. Leaf-wise 生长策略

而按叶生长(Leaf-wise)的生长策略针对上述缺点进行了改进，如图 3 所示。该方法是在决策树的同层叶子中寻找分裂增益最大的叶子结点进行分裂，其余分裂增益较小的叶子结点停止分裂，按此规则不断

迭代。最终生成的决策树在同样的分裂次数下能够达到更高的精度，但由于树的深度较深，容易造成过拟合。因此，LightGBM 算法在引用 Leaf-wise 叶子生长策略的基础上，通过 Max-depth 增加了深度限制，降低模型在训练数据集上过拟合的风险。

2.1.4. 互斥特征捆绑

互斥特征捆绑算法(EFB, Exclusive Feature Bundling)能够有效减少特征维度，应用在大规模数据集上，可以显著提升运算效率。通常而言，高维度的数据是十分稀疏的，在这些稀疏的特征中，有许多特征不会同时取非 0 值，这类特征即为互斥特征。EFB 算法将互斥特征进行合并，形成更为紧密的特征，进而降低特征维度。使用 EFB 算法构造直方图的复杂度大大降低，可以实现在不损失精度的情况下提升模型训练速度。合并互斥特征的关键是确保合并后的特征仍然可以识别出原始特征值。由于 LightGBM 的直方图算法储存的是离散型数据，因此可以通过对原始特征值添加偏移量来确保各个箱中的独立特征完成捆绑。

2.2. SHAP 模型

本文使用 LightGBM 算法构建了精度较高的家政服务员离职预测模型，但该模型的高复杂度导致了模型的可解释性降低，使其几乎成为一个黑箱模型。Lundberg 等人于 2017 年提出了 SHAP (SHapley Additive exPlanations)模型[17]，其主要思想来自于组合博弈论中的 Shapley 值[18]，该模型通过改进 Shapley 值使其适用于机器学习模型，并用以解释各类黑箱模型。SHAP 方法以 SHAP 值来计算各样本的指标组合贡献，在得到每个样本的各个指标贡献度后，若某指标在大部分样本上呈现出一致的趋势，则认为这一指标具有重要的正向或负向作用。LightGBM 算法可以通过 Feature Importance 给出各特征的重要程度，但 SHAP 模型能够在此基础上进一步解释各特征是如何影响预测结果的，如特征的影响力、特征影响结果的正负性、特征的依赖关系等。

模型对每个预测样本都会产出一个预测值，而 SHAP 值则是该预测样本中每个特征所获得的值，设第 i 个样本为 x_i，则第 i 个样本的第 j 个特征为 $x_{i,j}$，第 i 个样本的模型预测值为 y_i，设模型的基线(即全部样本的目标变量均值)为 y_{base}，则 SHAP 值服从公式(3)。

$$y_i = y_{base} + f\left(x_{i,1}\right) + f\left(x_{i,2}\right) + \ldots + f\left(x_{i,j}\right) \tag{3}$$

公式中，$f(x_{i,j})$为特征 $x_{i,j}$ 的 SHAP 值，即为第 i 个样本的第 j 个特征对预测值 y_i 的贡献程度，同时该公式表明了所有样本预测值均值与实际预测值之间的偏差。若 $f(x_{i,j}) > 0$，则表明该特征对预测值有正向影响，能够提升预测值；反之，则表明该特征具有负向影响。

3. 模型构建与对比实验

3.1. 数据描述及预处理

本文所使用的数据为某互联网家政服务公司 IT 部门数据库内的真实数据。该公司将连续三个月无接单行为的家政服务人员定义为离职员工，为确保数据的时效性，本文选择在 2021 年 9 月、10 月和 11 月内有接单行为的家政服务人员数据进行建模，共选择了 927 个样本。初始特征包括服务者 ID、姓名、性别、离职状态、是否兼职、门店 ID、学历、评分、出生年、注册时间、9 月请假通过次数、9 月获得 1 星评价次数等 41 个特征。为更好地适配模型方法，将数据样本进行数据预处理与特征选择，步骤如下：

确认离职状态，修正在离职状态、结束合作时间以及 2021 年 9 月、10 月和 11 月接单所获得的评价等特征中有逻辑冲突的样本。将结束合作时间在 2021 年 9 月之前但仍在 2021 年 9 月~11 月有接单行为的员工离职状态修改为 1 (在职)，将其合作结束时间修改为 null 值，共 46 人。

进行数据清洗，删除与算法和模型无关的特征，如服务者 ID、姓名等 5 个特征。

对部分连续型数据进行离散化处理，如注册时间、出生日期、入职时间等特征，将其等距离离散化划分为 n 组。

将通过以上步骤预处理后的数据整合合并，形成新的实验数据集，该数据集共 927 个样本，36 个特征变量。其中，离职员工数量为 284，占全体样本的 30.6%；未离职员工数量为 643，占全体样本的 69.3%。使用 Pyhton 中的 Pandas 库对实验数据集进行描述性统计，部分结果如表 1 所示。

Table 1. Descriptive statistics of the experimental data set
表 1. 实验数据集描述性统计

特征	平均值	方差	最小值	25%分位数	50%分位数	75%分位数	最大值
星级	1.4962	0.6370	1	1	1	2	3
学历	1.8900	1.0175	0	2	2	2	8
总评分	4.9740	0.0556	4	4.9689	4.9897	5	5
注册时间	2019.6	1.7025	2015	2018	2020	2021	2021
出生年	1983.5	161.17	1957	1972	1978	1983	2000
9 月请假通过次数	1.4736	2.5874	0	0	0	2	23
9 月请假未通过次数	0.0302	0.2208	0	0	0	0	4
9 月 1 星评价次数	0.0151	0.1220	0	0	0	0	1
9 月 3 星评价次数	0.0356	0.1967	0	0	0	0	2
9 月 5 星评价次数	36.6224	32.835	0	7	27	60	130

3.2. 模型评价指标

本文选择准确率(Accuracy)、F1 值和 AUC 值(Area Under Curve)这三个常见的模型评价指标来衡量模型的优劣。表 2 为关于预测离职结果的混淆矩阵，依据混淆矩阵可得出相应的模型评价指标值。准确率(Accuracy)是基于所有预测样本，预测正确的样本数与全体样本数之比，是最基础的评价指标。为避免随机划分训练集与测试集所带来的准确率的偶然性，本文使用十折交叉验证的准确率作为评价指标。F1 值为综合评价指标，由于精确率(Precision)和查全率(Recall)在某种情况下无法进行比较(如精确率低、查全率高的模型)，为了使其具有可比性，本文选择 F1 值来同时测量精确率和查全率，它通过惩罚极值的方式用谐波均值代替算数均值。准确率和 F1 值计算公式如公式(4)和(5)所示。

Table 2. Confusion matrix for predicting turnover outcomes
表 2. 预测离职结果的混淆矩阵

离职状态	预测离职	预测未离职
实际离职	TP	FN
实际未离职	FP	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{4}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100\% \tag{5}$$

ROC 曲线为受试者工作特征曲线，其横纵坐标分别是假阳性率和真阳性率。AUC 值可以通过 ROC 曲线下面积求得，其取值范围在 0~1 之间，AUC 值越大表示模型精度越高、泛化能力越好。

3.3. 对比实验及结果分析

本文的实验环境为一台拥有 4G 内存、搭载 2.30 GHz 的 Intel i5-6200U CPU、硬盘空间为 256 G、安装 Windows 10 系统的电脑，编程语言为 Python 3.8.3。

导入实验数据，划分训练数据集与测试数据集，其中训练数据集占总样本的 80%，测试数据集所占比例为 20%，将训练数据集中的特征变量与目标变量输入模型，进行模型训练。

在参数选择方面，本文使用网格搜索与五折交叉验证的方法进行超参数寻优。寻优后的参数设定如表 3 所示。

Table 3. Parameter setting of LightGBM algorithm
表 3. LightGBM 算法的参数设定

参数名称	参数含义	设定值
N-estimators	用于训练的提升树的数量	35
Max-depth	基学习器的最大树深度	7
Max-bin	直方图中最大分桶的桶个数	140
Num-leaves	基学习器的最大叶子数	12
Min-child-samples	叶子结点所需要的最小样本量	20

将本文建立的 LightGBM 家政服务员离职预测模型与 KNN 算法、逻辑回归算法、决策树算法、随机森林算法与 GBDT 算法进行交叉验证实验对比，并使用十折交叉验证的准确率、F1 值与 AUC 值作为评价指标评估。对比结果如表 4 所示。

Table 4. Comparison results of performance evaluation metrics for each model
表 4. 各模型性能评估指标对比结果

模型名称	Accuracy	F1-score	AUC
LightGBM	0.8123	0.8441	0.8650
KNN	0.6073	0.7204	0.6838
Logistic Regression	0.7860	0.8172	0.7867
Decision Tree	0.7210	0.8172	0.7806
Random Forest	0.8018	0.8333	0.8593
GBDT	0.7968	0.8226	0.8637

从表 4 结果分析可知，本文使用的 LightGBM 模型在这三项性能评估指标上的表现均要优于其他模型，其中十折交叉验证的准确率为 81.23%，F1 值为 84.41%，AUC 值为 86.50%，表明该模型的预测精度较高，泛化能力与稳健性更好。整体来看，集成学习模型(如随机森林、GBDT 与 LightGBM)的各项性能表现较单一学习器模型(如 KNN、逻辑回归与决策树)更加突出。其中，KNN 算法的准确率最低，仅有 60.73%；随机森林与 GBDT 算法的准确率均在 80%左右，表现较好，但仍比 LightGBM 模型低 1 个百分点以上。

各模型的 AUC-ROC 曲线如图 4 所示，AUC 值越大，模型的准确性越高，泛化能力越好。可以看出，

KNN、逻辑回归与决策树算法的 AUC-ROC 曲线均被 LightGBM 模型的曲线包裹，此类单一学习器模型的 AUC 值显著低于 LightGBM 模型。而 LightGBM、随机森林与 GBDT 这三种集成学习算法的 AUC-ROC 曲线位置区分并不十分明显，但从具体数值上看，LightGBM 模型的 AUC 值仍高于随机森林与 GBDT。

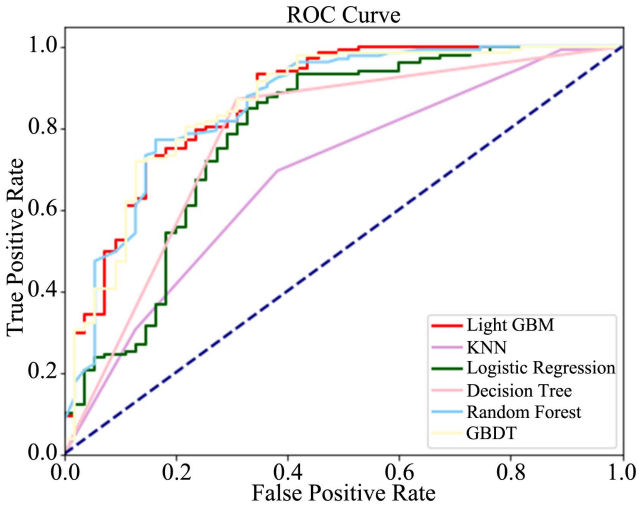


Figure 4. AUC-ROC curves for each model
图 4. 各模型 AUC-ROC 曲线图

4. 基于 SHAP 的模型解释分析

本章通过 SHAP 模型对 LightGBM 家政服务员离职预测模型进行解释分析，主要对特征的影响力、特征影响结果的正负性、特征的依赖关系等进行描述说明。

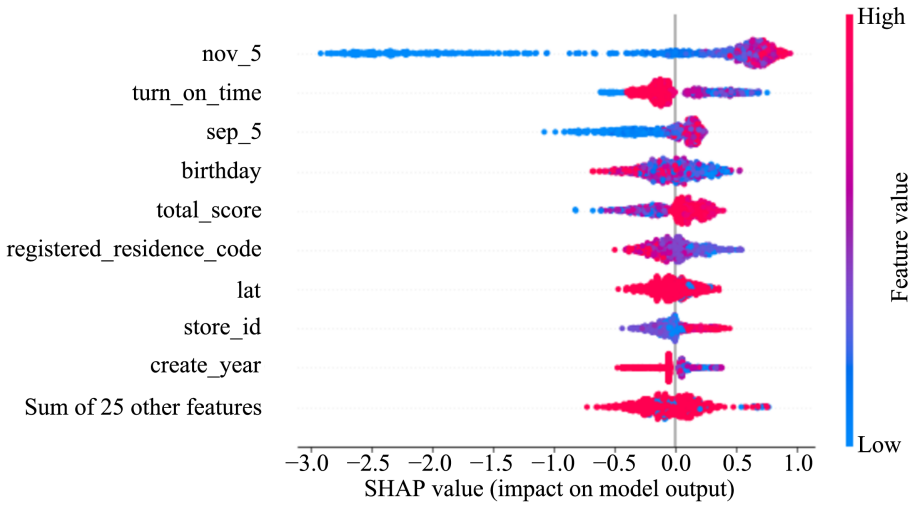
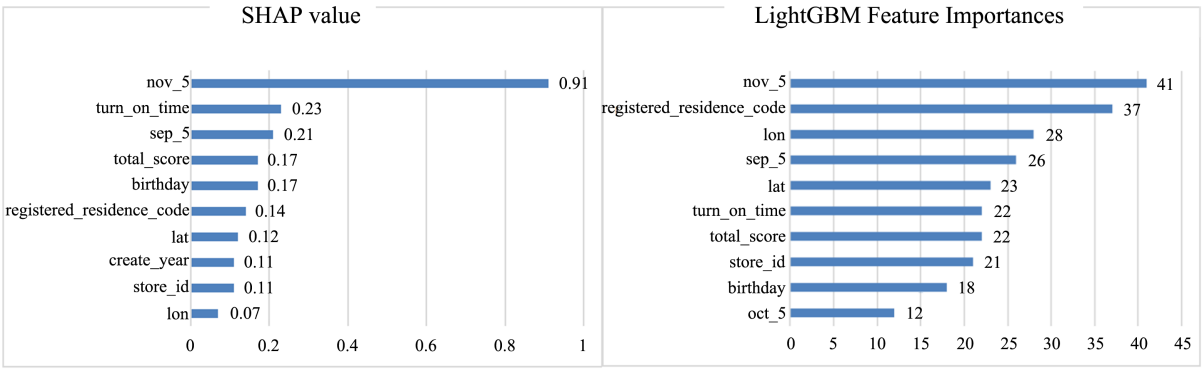


Figure 5. Global interpretation diagram of LightGBM model
图 5. LightGBM 模型的全局解释图

图 5 为通过 SHAP 模型建立的 LightGBM 算法的全局解释图。该图反映了 LightGBM 算法中特征、特征值与 SHAP 值的关系。SHAP 值的绝对值大小表明该特征对离职影响的强弱，绝对值越大，影响力越强。从正负性方面来看，SHAP 值为正，代表该特征对离职具有正面影响；SHAP 值为负，代表该特征对离职具有负面影响。从特征值角度来看，特征值高，颜色呈现红色，反之则呈现蓝色。由图 5 可看出，

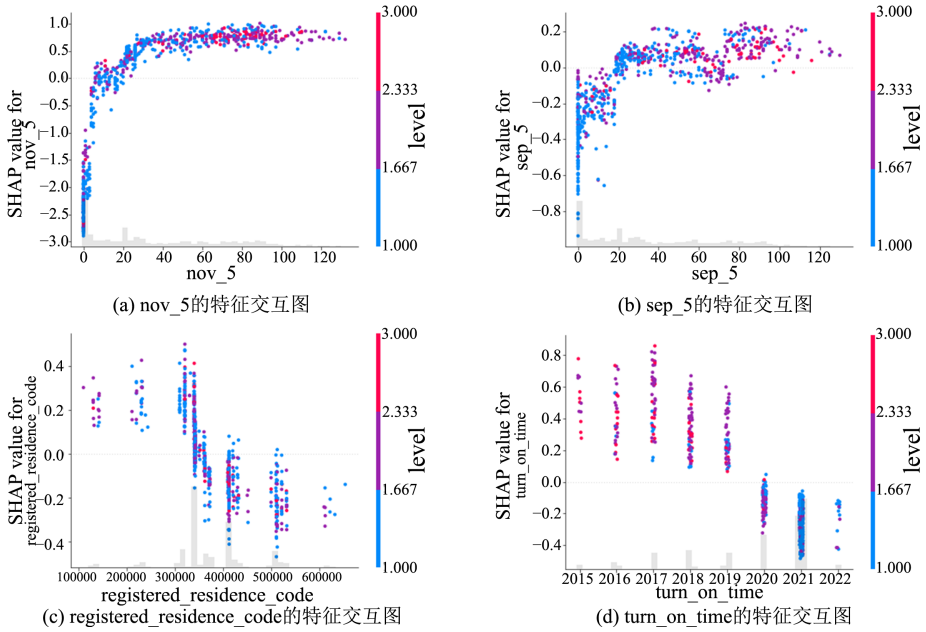
Nov-5 (11 月获得 5 星评价的次数)、turn-on-time (开始合作时间)、Sep-5 (9 月获得 5 星评价的次数)、birthday (出生年)、total-score (总评分)等特征对该模型的影响较为突出。其中，Nov-5 (11 月获得 5 星评价的次数)和 Sep-5 (9 月获得 5 星评价的次数)对模型的影响较为一致，低特征值所带来的负面影响较大，表明若 11 月获得 5 星评价的次数较少，则会给离职带来负向影响，导致离职行为出现的可能性提高；而高特征值会带来正面影响，但影响力较小。turn-on-time (开始合作时间)则呈现出较为复杂的影响能力，高特征值大多会带来负面影响，但影响力较小，表明越晚开始合作的家政服务员，离职的可能性越高；但低特征值对离职的正负向影响均有，相较高特征值而言，低特征值的影响力更大。

图 6 为 SHAP 模型与 LightGBM 模型的特征重要程度排名对比图。可以看出，两个模型的排名顺序以及特征重要性的相对差异并不完全相同。综合来看，Nov-5 (11 月获得 5 星评价的次数)、Sep-5 (9 月获得 5 星评价的次数)和 registered-residence-code (户籍编码)、turn-on-time (开始合作时间)等特征的影响力较为显著。



(a) SHAP　　(b) LightGBM

Figure 6. Comparison of SHAP model and LightGBM model feature importance ranking
图 6. SHAP 模型与 LightGBM 模型特征重要程度排名对比图



(a) nov_5的特征交互图　　(b) sep_5的特征交互图

(c) registered_residence_code的特征交互图　　(d) turn_on_time的特征交互图

Figure 7. Feature interaction diagram of SHAP model
图 7. SHAP 模型的特征交互图

图 7 为 SHAP 模型的特征交互图。图片选取了 Nov-5 (11 月获得 5 星评价的次数)、Sep-5 (9 月获得 5 星评价的次数)、registered-residence-code (户籍编码)和 turn on time (开始合作时间)这四个影响力显著的特征分别与 level (星级)绘制特征交互图。可以明显看出，随着 Nov-5 (11 月获得 5 星评价的次数)和 Sep-5 (9 月获得 5 星评价的次数)的增大，SHAP 值在增大；高星级人群(level 特征的特征值为 2 或 3)在这两个特征上的取值范围大多分布在 60~120，由此可知，高星级人群获得 5 星评价的次数较多，该人群不易离职。分析 registered-residence-code (户籍编码)与 level (星级)的特征交互图可知，华北地区(registered-residence-code 值为 1 开头)与东北地区(registered-residence-code 值为 2 开头)的家政服务员 SHAP 值为正，对离职具有正向影响，这表明该部分人群更加不易离职；河南、湖北等地(registered-residence-code 值为 4 开头)、西南地区(registered-residence-code 值为 5 开头)和西北地区(registered-residence-code 值为 6 开头)的家政服务员更易离职。同时，华北地区与西南地区的高星级人群占比更高。除此之外，从 turn-on-time (开始合作时间)来看，合作时间越早，高星级人群占比越高，越不易离职。

5. 结语

互联网背景下，家政服务行业的人力资源管理场景发生的重大变化，使得企业与员工之间的契约关系及企业对员工的管理范围和管理模式产生了革命性的改变，传统的离职管理方式已不再适用于该场景，因此，企业必须在离职管理方式上寻求新的出路。通过机器学习的方法提高员工离职预测的准确率及可解释性，探寻对员工离职影响力较大的特征因素，对企业管理具有重要的现实意义。

本文基于企业的真实数据集，建立 LightGBM 家政服务员离职预测模型，并使用 SHAP 模型增强其解释性，将二者相结合，形成解决互联网背景下家政服务员离职问题的集成方法。首先，进行数据清洗等数据预处理工作，将处理后的数据用于模型训练；之后，通过网格搜索与交叉验证的方式进行模型参数调优，并与 KNN 算法、逻辑回归算法、决策树算法、随机森林算法与 GBDT 算法进行实验对比，结果表明，LightGBM 算法在准确率、F1 值和 AUC 值上均优于其他主流算法，以此证明 LightGBM 算法的有效性；最终，使用 SHAP 模型进行特征解释分析，发现获得 5 星评价的次数、户籍、开始合作时间、出生年等特征是主要影响员工离职的因素。基于以上研究，企业管理者能够提前预判员工的离职行为，并依据特征因素制定相应政策对人才进行挽留，最大限度降低企业损失。

未来的研究可以着眼于模型的迭代优化，通过扩充数据集与增加特征维度的方式来尽可能全面地评估影响离职的因素，并对特征赋予不同权重以提高模型精度。

基金项目

上海市哲学社会科学规划一般课题“考虑数据伦理的入户服务人员行为风险状态分类研究”(项目编号：2020BGL007)。

参考文献

[1] 李婕. 做好家政兴农大文章[N]. 人民日报海外版, 2021-10-21(005).

[2] 尹双红. 推动家政服务业高质量发展[N]. 人民日报, 2021-07-08(005).

[3] Corredoira, R.A. and Rosenkopf, L. (2010) Should Auld Acquaintance Be Forgot? The Reverse Transfer of Knowledge through Mobility Ties. *Strategic Management Journal*, **31**, 159-181. https://doi.org/10.1002/smj.803

[4] Mendis, M.V.S. (2017) The Impact of Reward System on Employee Turnover Intention: A Study on Logistics Industry of Sri Lanka. *International Journal of Scientific & Technology Research*, **6**, 67-72.

[5] Mittal, S., Gupta, V. and Motiani, M. (2022) Examining the Linkages between Employee Brand Love, Affective Commitment, Positive Word-of-Mouth, and Turnover Intentions: A Social Identity Theory Perspective. *IIMB Management Review*, **34**, 7-17. https://doi.org/10.1016/j.iimb.2022.04.002

[6] Zhang, X., Ma, L., Xu, B., *et al*. (2019) How Social Media Usage Affects Employees' Job Satisfaction and Turnover Intention: An Empirical Study in China. *Information & Management*, **56**, Article ID: 103136. https://doi.org/10.1016/j.im.2018.12.004

[7] Khalid, S.A., Nor, M., Ismail, M. and Razali, M.F.M. (2013) Organizational Citizenship and Generation Y Turnover Intention. *International Journal of Academic Research in Economics and Management Sciences*, **3**, 132-141. https://doi.org/10.6007/IJAREMS/v2-i4/104

[8] Ohunakin, F. and Olugbade, O.A. (2022) Do Employees' Perceived Compensation System Influence Turnover Intentions and Job Performance? The Role of Communication Satisfaction as a Moderator. *Tourism Management Perspectives*, **42**, Article ID: 100970. https://doi.org/10.1016/j.tmp.2022.100970

[9] Khera, S.N. and Divya (2019) Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. *Vision*: *The Journal of Business Perspective*, **23**, 12-21. https://doi.org/10.1177/0972262918821221

[10] Kao, H.W., Lin, S.W. and Wan, S.Y. (2012) Applying Decision Tree to Predict Nursing Turnover—A Case Study in a Public Hospital. *The Journal of Taiwan Association for Medical Informatics*, **21**, 15-29.

[11] Gao, X., Wen, J. and Zhang, C. (2019) An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, **2019**, Article ID: 4140707. https://doi.org/10.1155/2019/4140707

[12] Jain, R. and Nayyar, A. (2018) Predicting Employee Attrition Using Xgboost Machine Learning Approach. 2018 *International Conference on System Modeling & Advancement in Research Trends* (*SMART*), Moradabad, 23-24 November 2018, 113-120. https://doi.org/10.1109/SYSMART.2018.8746940

[13] Ke, G.L., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3146-3154.

[14] Wang, D.H., Yang, Z. and Yi, Z. (2017) LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. *Proceedings of the* 2017 *International Conference on Computational Biology and Bioinformatics*, Newark, 18-20 October 2017, 7-11. https://doi.org/10.1145/3155077.3155079

[15] Zhang, M., Fei, X. and Liu, Z.H. (2018) Short-Term Traffic Flow Prediction Based on Combination Model of Xgboost-lightgbm. 2018 *International Conference on Sensor Networks and Signal Processing* (*SNSP*), Xi'an, 28-31 October 2018, 322-327.

[16] Weng, T.Y., Liu, W.Y. and Xiao, J. (2019) Supply Chain Sales Forecasting Based on LightGBM and LSTM Combination Model. *Industrial Management & Data Systems*, **120**, 265-279. https://doi.org/10.1108/IMDS-03-2019-0170

[17] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[18] Shapley, L.S. (1953) A Value for N-Person Games. *Contributions to the Theory of Games*, **2**, 307-317. https://doi.org/10.1515/9781400881970-018