Modern Management 现代管理, 2018, 8(6), 621-630
Published Online December 2018 in Hans. http://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2018.86080

Hans汉斯

Empirical Analysis and Forecast of Turnover Factors of Excellent Employees Based on Random Forest

Qiaoqiao Tang, Haomin Zhang*

School of Science, Guilin University of Technology, Guilin Guangxi
Email: 584940435@qq.com, *Zhanghm@glut.edu.cn

Received: Nov. 20th, 2018; accepted: Dec. 3rd, 2018; published: Dec. 10th, 2018

Abstract

Talents play an important role in constructing the core competence of enterprises and play a decisive role in the sustainable development of enterprises. With the opening and progress of the society, the flow of talents is becoming more and more frequent, especially the outstanding talents. Therefore, more and more attention has been paid to the analysis of factors affecting the turnover of outstanding employees. In this paper, we use the relevant data set on Kaggle website to establish a stochastic forest model to find out the key factors that affect the turnover of outstanding employees in an enterprise. Combining these variables into a single indicator to help the company understand which outstanding employees need to be focused on and forecast the outstanding employees on the job, and judge the probability of their leaving. Thus, we can make effective measures to retain talents for company managers.

Keywords

Excellent Employees, Turnover Factor, Random Forest

基于随机森林的优秀员工离职因素实证分析及预测

唐巧巧，张浩敏*

桂林理工大学理学院，广西 桂林
Email: 584940435@qq.com, *Zhanghm@glut.edu.cn

收稿日期：2018年11月20日；录用日期：2018年12月3日；发布日期：2018年12月10日

*通讯作者。

文章引用：唐巧巧，张浩敏. 基于随机森林的优秀员工离职因素实证分析及预测[J]. 现代管理, 2018, 8(6): 621-630.
DOI: 10.12677/mm.2018.86080

摘　要

人才在构建企业核心竞争力中扮演着重要角色，对企业的可持续发展起着决定性作用。随着社会的开放与进步，人才的流动也越来越频繁，特别是优秀人才。因此对于公司优秀员工离职影响因素的分析越来越受到关注。本文利用kaggle网站上的相关数据集，通过建立随机森林模型来找出影响企业的优秀员工离职的关键因素，将这几个变量综合成一个指标帮助公司了解哪些优秀员工需要被重点关注以及对在职的优秀员工进行预测，判断其离职的概率，从而可以为公司管理人员制订有效的挽留人才措施。

关键词

优秀员工，离职因素，随机森林

1. 引言

21 世纪是一个信息经济和知识经济高度发展的时代，随着国际市场的不断开放以及国内市场自由化程度的提高，市场对人才资源的竞争愈发激烈，因此导致优秀人才的流动意愿也越来越强烈，离职的行为也变得越来越频繁。优秀人才的流失将会严重抑制公司的发展，也将使公司付出巨大的人力物力去重新培养可用人才，因此通过对公司优秀员工的离职因素的研究，有效的降低离职率、减少离职行为变得必不可少。国内外对员工离职的影响因素的研究已有很多，著名学者 Viteles [1]得出在组织中智力测验成绩高于平均分以上的员工留职的时间更长。Abelson [2]认为离职员工的离职大多是因为工作压力大，对工作存在意见。我国学者刘智强、廖建桥和李震[3]在对国企员工离职倾向影响因素的研究中发现，升职制度对于国企员工离职倾向的影响要高于薪酬。颜西平[4]通过对生产一线的生产员工离职情况研究发现，影响其离职的主要因素是个人因素、组织因素及薪酬待遇。车雯[5]从中国银行咸阳分行的员工离职现状着手，发现影响员工离职的最主要原因是薪酬体系设计的不合理、晋升渠道太窄以及相关培训的缺乏。马跃如，余航海，夏冰[6]在研究中证明，组织中的破坏性领导会加大和强化员工的负面情绪，导致更强烈的离职意愿。

综上所述，国内外大部分关于员工离职影响因素的分析大多建立在定性分析的基础上，定量分析涉及的比较少。本文尝试使用随机森林机器学习[7] [8]方法对某公司优秀员工离职因素进行实证分析，找出影响其离职的最关键因素，把多余的无用的特征去掉，降低了维度计算，避免了过拟合，同时用挑选出来的关键变量来形成一个新的指标帮助公司改善管理模式，并对在职的优秀员工进行预测，判断其离职的概率以供公司管理人员采取相应的挽留人才措施。最后，提出有效可行的建议。

2. 样本数据的获取及模型构建

2.1. 样本数据获取

本文采用的数据集来源 kaggle 竞赛项目：HR-Analytics。这个数据集是关于一家大公司的人力资源数据集，共有 14,999 条数据，包含 9 个自变量以及一个因变量。详细说明见表 1。熊梦鸿[9]谈到，在现

代的人力资源管理中，薪酬管理和绩效管理是导致优秀人才严重流失、抑制企业发展的最为关键和重要的内容。在本文中，将基于熊梦鸿[9]的论述，对于优秀员工的选择指标分别为：工作年限大于等于 4 年；绩效评估大于等于 0.72；薪资水平为 high。

Table 1. Explanation of variables
表 1. 变量的解释说明

变量名	解释说明
left	是否已经离职。0 代表未离职，1 代表已经离职
Satisfaction_level	对公司的满意程度。取值范围：0~1
last_evaluation	绩效评估。取值范围：0~1
number_project	参加过的项目数。单位：个，取值范围：2~7
average_montly_hours	平均每月工作时长。单位：小时，取值范围：96~310
time_spend_company	工作年限。单位：年，取值范围：2~10
Work_accident	是否发生过工作差错。0 代表未发生，1 代表已经发生
promotion_last_5 years	五年内是否升职。0 代表未升职，1 代表已经升职
sales	职业。共 10 个水平。accounting、hr 等
salary	薪资水平。共 3 个水平。medium、high、low

2.2. 基本的描述性统计分析

从表 2 可以看到，1) 该公司的离职率将近 20.8%。2) 该公司的优秀员工对公司的满意度在 62%左右。3) 该公司的优秀员工的绩效评估大概在 79.7%。4) 该公司的优秀员工平均每人参加过 4 个项目左右。5) 该公司的优秀员工每月平均工作时长大约在 210 个小时。

Table 2. Description statistics for each variable
表 2. 各个变量的描述统计量

变量名	取值	计数	最小值	中位数	均值	最大值
left	0	7774				
	1	2044				
Satisfaction_level	0~1	9818	0.09	0.68	0.6176	1
last_evaluation	0~1	9818	0.36	0.83	0.7973	1
number_project	2	618				
	3	2636				
	4	3091				
	5	2147				
	6	1078				
	7	248				
average_montly_hours	96~310	9818	96	217	211.4	310
time_spend_company	2~10	9818	2	4	3.911	10
Work_accident	0	8323				
	1	1495				
promotion_last_5years	0	9575				
	1	243				
salary	high	1237				
	low	4478				
	medium	4103				

从图 1(a)中可以看到，该公司的优秀员工工作未发生差错的占比为 84.77%，远高于发生差错的；图 1(b)中可以看出，该公司优秀员工的薪资大多分布于低等－中等水平，高等水平占比比较少；图 1(c)中可以看出该公司优秀员工没有得到晋升的占比为 97.52%，远大于得到晋升的。图 1(d)中可以看出该公司优秀员工参加 4 个项目的人数居多，参加 7 个项目的人数较少，占比为 2.53%。

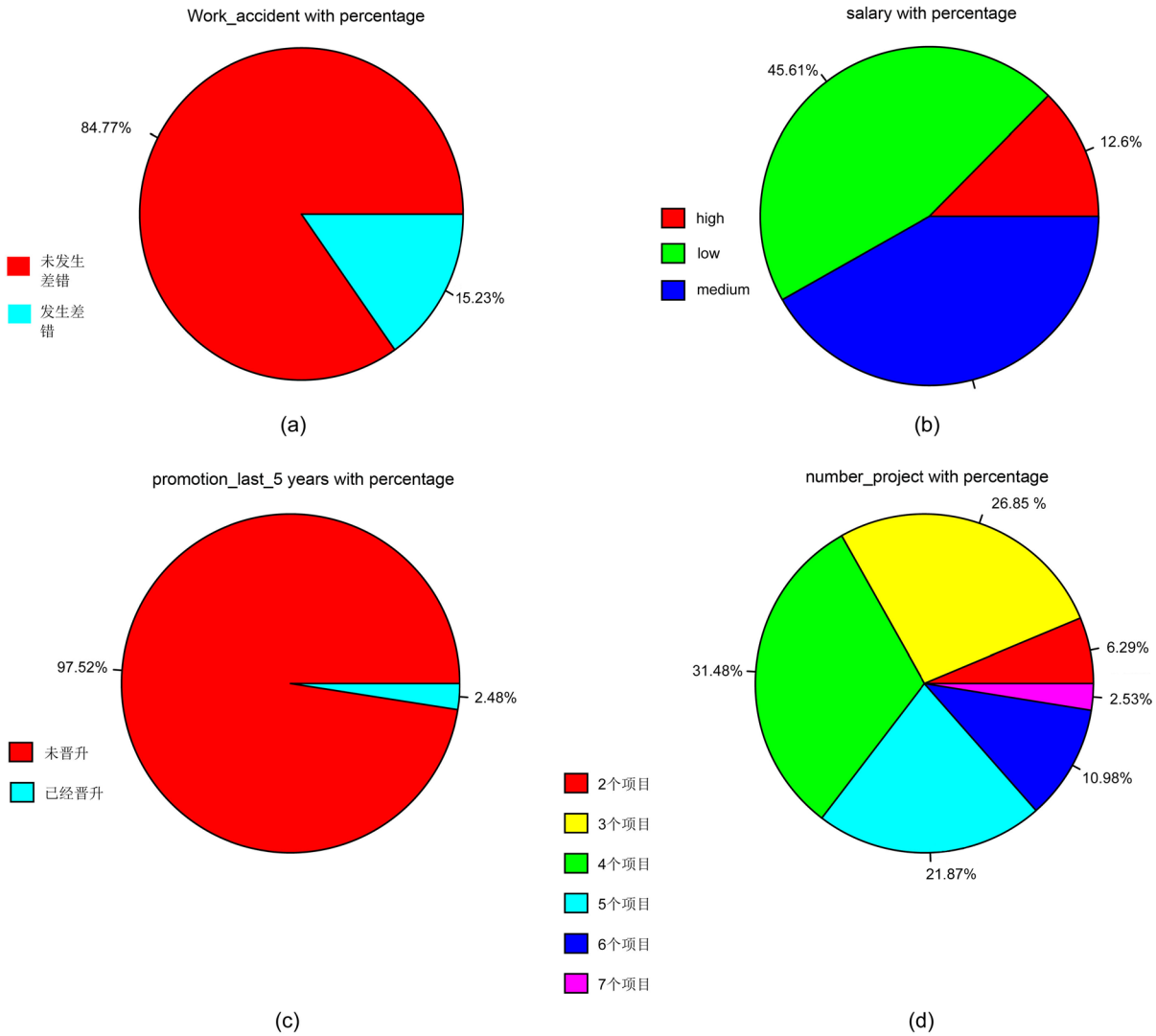

Figure 1. Percentage pie chart for each variable
图 1. 各变量的百分比饼图

接下来通过 R 软件对选择出的优秀员工数据进行分析，将进一步探索各个变量分别与因变量(是否已经离职)之间的关系，结果如下图 2。

从图 2 中可以看出，优秀员工离职的特征：

1) 对公司满意度较低的员工离职率高，对公司满意度较高的员工离职率也不低；

2) 绩效评估较高的优秀员工离职率也高；

3) 平均每月工作时间多的(高于 230 个小时)优秀员工相对于每月工作时间短的员工离职率较高；

相对应的可能原因进一步分析如下：

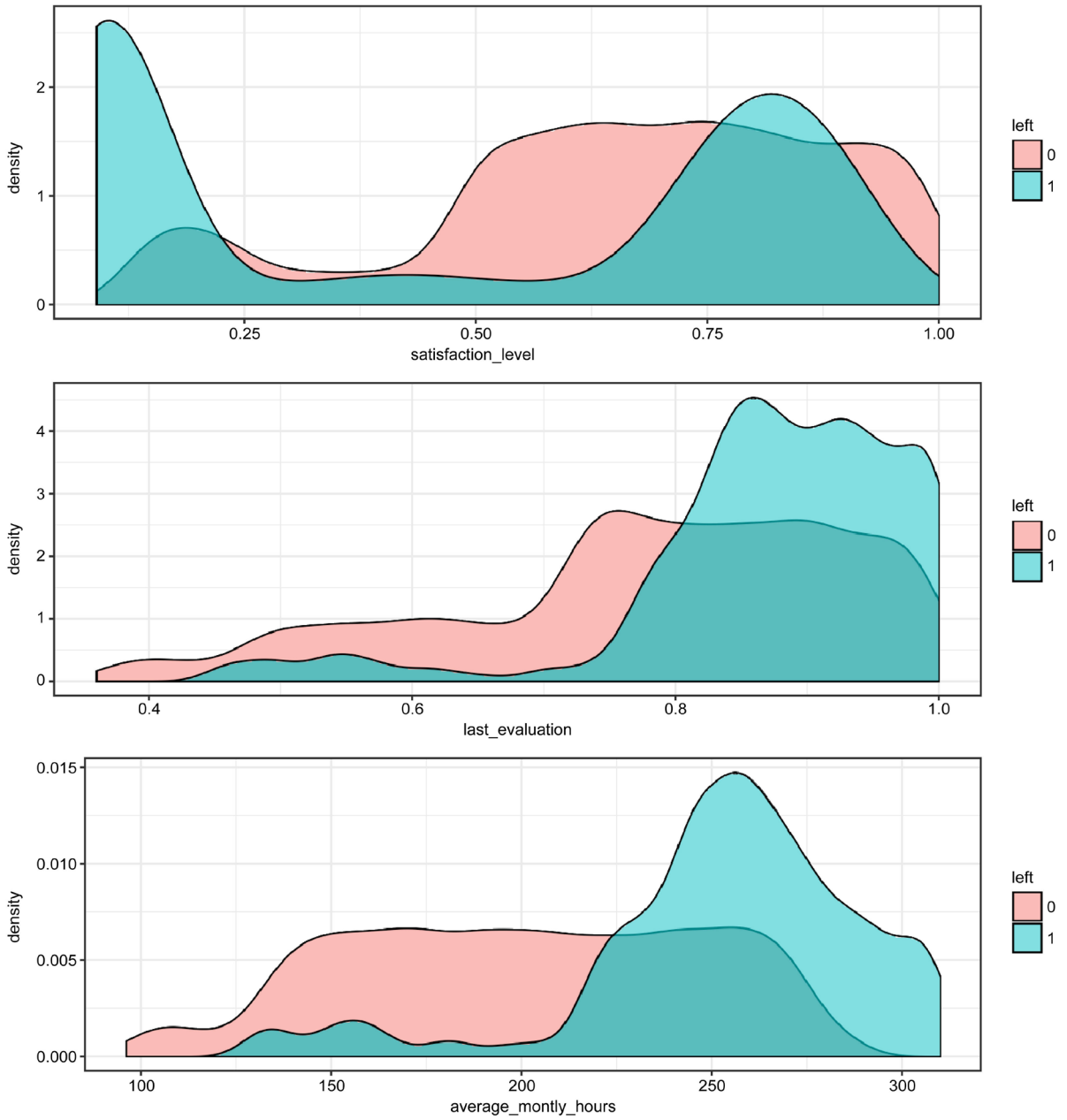

Figure 2. Whether to quit or not to be satisfied with the company (top), performance evaluation (middle), average working hours per month (bottom)
图 2. 是否离职与对公司满意程度(上)、绩效评估(中)、平均每月工作时长(下)

从图 2(上)可以看出，对公司满意程度低于 0.1 的优秀员工基本上离职了，满意度在 0.75~0.9 之间，优秀员工的离职率又达到了一个峰值，这些都是对公司满意程度比较高的员工，说明其离职并不是对公司不满意，可能是他们寻找到了一个更好地工作机会。图 2(中)可以看出，绩效评估在 0.6~0.8 之间，优秀员工有一个比较好的留职情况，但是绩效评估比较优秀的离职密度较大，说明对于绩效评估比较高的优秀员工，公司没有相应的转化到升职和加薪上面。图 2(下)中可以很明显的看出，平均每月工作时间多的(大约高于 230 小时)优秀员工离职率最高，说明一般离开公司的优秀员工大部分属于过度工作的。

从图 3 中可以看出：优秀员工离职的特征：

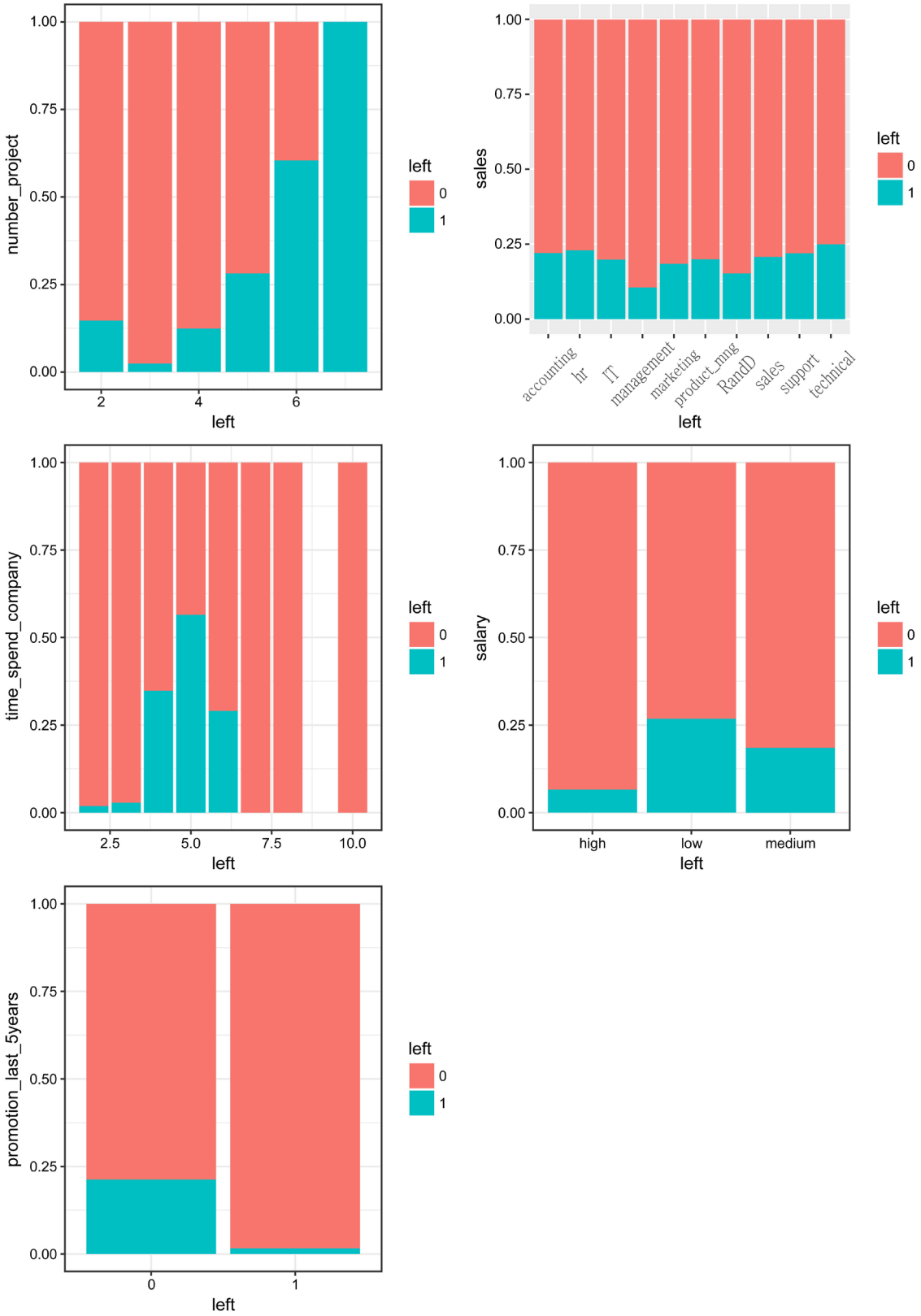

Figure 3. The percentage stacking bar chart of whether a good employee leaves or not and other independent variables
图 3. 优秀员工是否离职分别与其他自变量的百分比堆积条形图

1) 参与项目个数较少的员工选择离职，但参与项目数在 4~7 个之间的员工离职率越来越高；
2) 在公司工作时间较短的优秀员工离职率更高，年数在 4~6 年比较集中；
3) 各个部门的离职率都差不多；
4) 薪资在低等－中等水平的优秀员工离职率高；
5) 5 年内没有得到晋升的员工离职率高；

进一步对其产生的可能的原因进行分析如下：

(1)中有的优秀员工在只有两个项目的时候就选择了离开公司，同样地，有 4~7 个项目的优秀员工离职率愈来愈高，我们可以猜测到：项目数目比较少的优秀员工可能会因为没有受到公司的重视或者自己的才华得不到施展从而离开公司，6 个项目数以上的优秀员工可能是因为工作太劳累从而选择离开公司；(2)中大部分离职的优秀员工都是在公司已经工作了 4~6 年，而在公司待了 7~10 年的优秀员工反而没有人离职。其可能的原因在于年轻的优秀员工更倾向于多尝试挑选适合的公司或岗位，高离职率也意味着员工在短期内难以形成对企业价值观的长期认同。(3)中可以看出，各个部门的离职率情况差不多，但是人力资源(hr)和技术(technical)部门的离职率稍微比较高，可以针对这两个部门的优秀人员深入了解一下情况，多关注。(4)中薪资较低、没有得到晋升的离职率高，很明显，这符合人之常情。说明不定期的加薪和晋升也是必不可少的。

3. 模型构建

随机森林是通过组合多棵决策树分类器进行预测的，因此形成了“森林”，这也就是其名称的由来。从直观上讲，每棵决策树都是一个分类器(针对于分类问题)，那么对于输入一个样本，N 棵树就会有 N 个分类结果。而随机森林将对多个决策树产生的预测结果采取投票的方式，将投票次数最多的类别指定为最终的输出。在本文中，我们将根据因变量进行 7:3 的分层抽样，其中 70%作为训练集，剩余 30%作为测试集进行预测。

对于分类模型，通常采用混淆矩阵来评价其预测能力。混淆矩阵的核心在于预测值与真实值的互联表。显然，在混淆矩阵中，预测值和实际值相符的观测个数是评价模型好坏的一个重要指标，如下表 3 所示。

Table 3. Confusion matrix
表 3. 混淆矩阵

		预测分类	
		0	1
实际分类	0	真正类(*TP*)	假负类(*FN*)
	1	假正类(*FP*)	真负类(*TN*)

其中，强调预测精准程度和查准率的指标为：

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}, \quad \text{Precision} = \frac{TP}{TP+FP}$$

模型的精度，即模型预测正确的个数/样本的总个数。一般情况下，模型的精度越高，说明模型的效果越好；查准率，在模型预测为正类的样本中，真正为正类的样本所占的比例。一般情况下，查准率越高，说明模型的效果越好。

建立模型的实验结果

如表 4。

Table 4. Confusion matrix of test set
表 4. 测试集的混淆矩阵

预测	实际	
	0	1
0	2316	3
1	33	594

由表 4 可以看出，Accuraccy = (2316 + 594)/(2316 + 3 + 33 + 594) = 98.8%

Precision = 2316/(2316 + 33) = 98.6%，模型的精度和查准率都很高，说明该模型的预测效果非常好。

4. 优秀员工离职影响因素的重要性分析

绘制出影响优秀员工离职的各因素重要性分析如下图 4 所示。由于其他文献多使用 Mean Decrease Accuracy 指标，本文我们也重点在于对其的阐述。图 4 即为 Mean Decrease Accuracy 指标，是指平均精确度的降低。如果变量重要，则预测的误差增大，即误差的增大相当于精确度的减少，精确度越小也就反映这个变量越重要。从图中可以看出这些变量的重要性排序，其中比较重要的几个变量依次为对公司的满意程度、工作年限、绩效评估、平均每月工作时长。通过前面的分析也得到了类似的发现。

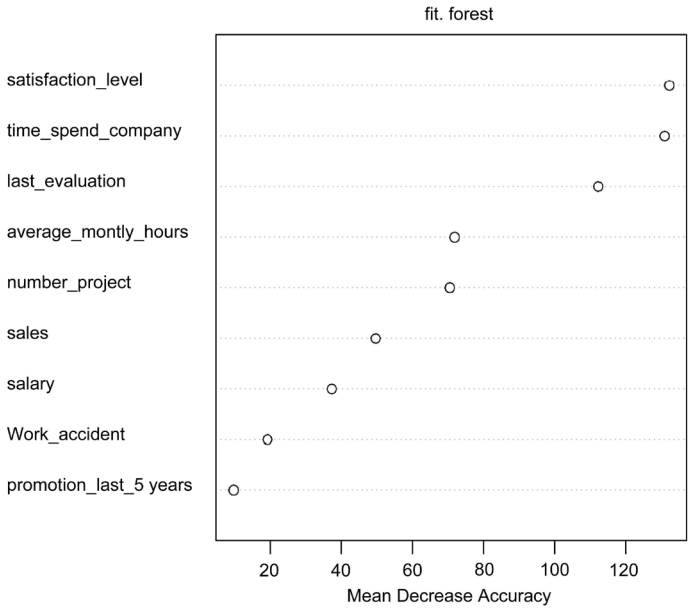

Figure 4. Ranking chart of the importance of factors influencing the turnover of excellent employees
图 4. 各因素对优秀员工离职影响重要性排序图

对在职优秀员工进行预测

目前为止，没有任何数据能够很精准的预测员工的去留，在这里，我们可以利用模型来预测优秀员工在不久的将来是否会离职。接下来我们就用剩余的 30%数据，采用上述方法进行预测，判断优秀员工是否会离职。

由表 5 可以得出，在这剩余 30%的样本数据中，有 597 个优秀员工可能会选择离职。该公司可以选择对这些优秀员工进行重点关注，尝试在优秀员工要离职的时候，采取适当措施以留住优秀人才或者提前招聘、培训新员工，减少公司造成的损失。

Table 5. Forecast excellent employee turnover probability table
表 5. 预测优秀员工离职概率表

编号	未离职的概率	离职的概率	是否会离职
2	0	1	1
4	0	1	1
7	0	1	
9	0	1	1
10	0	1	1
14	0.002	0.998	1
23	0.128	0.872	1
.....	.....	.....	.....
613	0.768	0.232	0
614	0.55	0.45	0
618	0.93	0.07	0
.....	.....	.....	.....
9813	0.01	0.99	1
9815	0.042	0.958	1
9816	0	1	1

5. 结论及建议

5.1. 结论

1) 随机森林算法可获得影响因素的重要性排序，可快速有效地从大量影响因素中辨别出对模型预测影响较大的、较关键的影响因素，降低了模型的维度，并减小了模型过拟合的可能性，在管理领域中可以进行广泛应用。

2) 选择离开该公司的优秀员工平均每月工作时长大多在 230 个小时、4~7 个项目以上且离职的优秀员工的薪资大多在低等－中等水平，说明选择离职的优秀员工存在着过度劳累、薪酬低，付出和回报不对等的情况；

3) 从优秀员工的不同因素的数据比较中可以看出，该公司的优秀的离职员工大多对公司满意程度比较高、绩效评估比较优秀，说明可能该公司没有这些相应的转化到升职和加薪上面，导致该公司优秀员工选择寻求另一个更好的机会；这也说明了结论 1，即使对公司有着很高的满意度，但是还是有其他的因素影响优秀员工的离职。

4) 各个部门的离职率大致都差不多，但是人力资源(hr)和技术(technical)部门的离职率比较高，那么该公司应该对其部门的工作方式和工作量进行关注；

5.2. 建议

1) 该公司应该更加注重优秀员工的工作和生活的平衡，采取人性化的管理方式，提高工作效率，避免加班情况的出现；

2) 该公司应该要合理进行薪酬设计，并且要确保能够及时的根据员工的工作表现和业绩以及市场变化作出相应的客观的变化，这样能够更好地激发优秀员工的工作热情，调动其积极性、创造性。优秀员

工看中的是良好的待遇，和更好的职业发展，这些因素都直接影响员工的主观感受，公司给予了员工高的评价，但没有相应转化到薪资和升职，即使一部分离职的优秀员工对公司有很高的满意度，但依然不能阻挡他们会追寻更好的工作机会。

3) 完善绩效评估的方式，绩效评估的真正意义是为了提高优秀员工的工作业绩水平，但是现在大部分公司的绩效评估缺乏科学的方式和依据，导致评估结果失去公正性以至于造成了人员的流动。该公司可以根据相关管理人员和专业人员工作岗位的性质、职责以及所要承担风险的大小程度来指定考核标准，脱离原本的死板的条条框框。

基金项目

国家自然科学基金项目(61763008，71762008)；广西自然科学基金项目(2016GXNSFAA380194)。

参考文献

[1] Viteles, M.S. (1924) Selecting Cashiers and Predicting Length of Service. *Journal of Personnel Research*, **2**, 467-473.

[2] Abelson, M.A. and Baysinger, B.D. (1984) Optimal and Dysfunctional Turnover: Toward an Organizational Level Model. *Academy of Management Review*, **9**, 331-341. https://doi.org/10.5465/amr.1984.4277675

[3] 刘智强, 廖建桥, 李震. 员工自愿离职倾向关键性影响因素分析[J]. 管理工程学报, 2006(4): 142-145.

[4] 颜西平. 武汉制造型企业一线员工满意度调查——以 F 公司为例[J]. 中国水运, 2010, 10(12): 77-78.

[5] 车雯. 中国银行咸阳分行员工流失问题研究[D]: [硕士学位论文]. 西安: 西北大学, 2017.

[6] 马跃如, 余航海, 夏冰. 破坏性领导对员工离职意愿的影响研究[J]. 贵州财经大学学报, 2018(2): 46-53.

[7] 徐昆, 赵东亮. 餐饮连锁店员工离职倾向预测研究[J]. 合作经济与科技, 2018(8): 170-173.

[8] 常国珍, 曾珂, 朱江. 用商业案例学 R 语言数据挖掘[M]. 北京: 电子工业出版社, 2017.

[9] 熊梦鸿. 基于优秀员工流失的中小民营企业人力资源管理问题[J]. 山海经, 2016(7): 157-158.