Central University of Finance and Economics

中央财经大学

经济统计学课程

企业员工离职预测及模型比较

吴宇翀

2017310836

WUYUCHONG.COM

指导老师：王文静

2020 年 6 月 21 日

摘要

我们通过案例研究分析一个企业的员工离职情况，并从中找出员工离职问题的原因，通过建立多种算法模型，预测每个**员工离职的概率**，并进行模型效果的比较。同时，预测模型并不局限于该企业，而具有很强的普适性，可以为其它企业乃至社会研究提供商业服务支持、政策制定支撑。

首先，我们建立简单的 Logit 回归以初步解释各个变量的效应。在使用混淆矩阵得出**灵敏度**和**特异度**之后，我们使用 **ROC 曲线**结合业务情形在两者之间进行权衡。

在使用相同的重抽样方法进行重复 5 次的 10 折**交叉验证**的前提下，我们将准确率和 Kappa 作为衡量指标，比较了 **Logit、线性判别、偏最小二乘判别、支持向量机、随机梯度助推模型**的优劣。

综合来看，**PLSDA** 模型在两项评判指标下都具有最好的效果。然而，在模型的应用方面，由于 Logit 模型计算速度较快、可解释性强的，在对准确率要求不高而更加重视变量的可解释性的场景下，Logit 模型也不失为一个较好的选择。

在变量选择上，最重要的变量**婚姻状况**和**绩效**。这两个变量对于不同部门、不同工作内容、不同工作地位的员工具有较强的普适性，属于对员工个人的刻画，对于预测员工是否离职较为重要。重要程度最低是薪资、特殊完成项目的数量和是否在 IT/IS 部门，这三个变量与员工个人的性格、工作能力、家庭关系较小，属于对工作分类的刻画，对于预测员工是否离职的重要性较低。

关键词：**员工离职**，**分类预测模型**，**机器学习**，**变量选择**，**模型比较**

目录

1 背景

随着市场化和国际化的不断推行，行业的更迭越来越快，企业间的竞争变得越来越激烈，人力资源的流动也变得越来越快。可以说，人力资源的流动是一家公司、整个社会不可或缺也不可避免的一部分，一方面它使得人力资源的分配更高效；然而，另一方面，它也带来了一些摩擦性失业。

1. 从社会科学的角度，对人力资源流动的合理预判有利于调整就业市场，尽可能地减小摩擦性失业。
2. 从人才就业服务中心的角度，将合适的人放置于合适的企业、匹配的岗位是其工作的核心。
3. 从企业的角度，由于企业的核心技术和运营业务被优秀的人才掌握，公司希望尽量避免自己所器重的员工为了更好的就业机会而主动离职，避免公司竞争力降低的同时竞争对手掌握主动权；同时，对于非核心的员工，公司则希望能够预判员工的离职，以降低离职率，减轻员工离职对公司正常经营活动的影响，节省公司新招聘员工的成本。

大部分企业都会设立人事部专门管理人力资源问题，企业有着招聘员工需要付出一定的成本，而员工入职后的培训和磨合也需要不小的费用。能够胜任工作的优秀员工的离职对于企业来说是不小的损失。所以，人员的频繁离职是人力资源部极为重视的问题之一。

引起员工离职的原因有很多，员工的薪资、满意度、工龄等等都是重要因素，通过描述性统计，我们能大致刻画出主动离职和被解雇员工的群体画像，得到一些结论。但是相较于得出群体结论，企业或是人才服务中心更加关注精确到每位员工上，从每位员工自身的角度，离职的原因又有很强的自身独特性，准确地预测出每位员工是否处于离职的边缘较难。

2 文献综述

在企业人力资源流动上，不同学者对员工离职的不同方面进行了研究。

国内学者张梓嫣和杨喆麟都使用案例分析的方法研究企业员工离职问题。张梓嫣用问卷调查的方法对 BJM 公司进行案例分析，深入剖析了 BJM 公司留才策略的详细计划和执行现状，同时对其他的

影视公司乃至整个影视行业如何缓解新员工频繁离职具有很高的参考价值。[1] 杨喆麟作为管理者和参与人力资源规划的研究人员以案例分析的形式，以星巴克公司驻中国部作为案例研究对象，研究招聘培训以及薪酬福利的特点、提高员工工作积极性的方法；并进一步对星巴克员工离职问题提出优化措施找到切实可行的优化策略，以降低离职率、提高企业竞争优势。[2]

在筛选重要变量，找出离职原因方面，学者赵西萍、刘玲和张长征在则问卷调查的基础上，采用因子分析法和多元相关分析法提取主要因子，以对员工工作态度进行测度，提取引起员工离职倾向的关键因素。[3]

在预测员工离职概率上，学者张紫君使用梯度提升分类树（GBDT）算法构建模型预测企业员工离职，采用 smote 算法对样本进行倾斜处理、采取网格搜索法模型的调优、使用混淆矩阵和 ROC 曲线进行评价模型、运用梯度提升分类树决定重要特征。[4]

3 研究方法

在我们的研究中，我们吸收了几位学者优秀的研究成果，并在他们的基础上继续加以研究创新。

1. 通过案例研究的方式，对数据进行分析和处理，建立模型对企业员工离职倾向进行预测，选用多种机器学习模型进行模型比较，为离职概率的预测提供更多的评判思路。但与此同时，研究并不局限于一个案例，而是具有普适性，在实际应用过程中能够有效地推广到不同企业，为商业服务和社会政策决策提供参考依据。
2. 我们应用机器学习算法探究导致离职的决定性因素，将重要的变量筛选出来，理清楚其影响关系，使用数据集中的这些变量预测有离职的倾向的员工。在数据中向更深的层次进行挖掘，通过探究内在的问题，提前采取措施，从而避免造成更多的损失。
3. 结合心理学中的需求层次理论、双因素理论、公平理论和职业生涯理论探究导致员工的离职的深层次原因。从员工个人家庭婚姻因素，到职业发展因素，再到工作内容和工作岗位，又结合员工的更深层次的感受：参与感、满足感、价值感，多方面地探究导致员工离职最重要的因素。
4. 在实际应用中，并不局限于此数据集，此研究为离职的预测提供思路，在应用时可以加以调整，使用更大维度的数据集，在分布计算环境下进行学习预测。

4 数据集与描述分析

4.1 数据集说明

我们使用一个公开的数据集[1] ，它有 35 个变量，310 个观测。[2]

[1]数据来源: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

[2]模型的变量取值和分布见附录

表 1: 变量解释和类型

变量名	变量描述	数据格式
Employee Name	Employee's full name	Text
EmpID	Employee ID is unique to each employee	Text
MarriedID	Is the person married (1 or 0 for yes or no)	Binary
MaritalStatusID	Marital status code that matches the text field MaritalDesc	Integer
EmpStatusID	Employment status code that matches text field EmploymentStatus	Integer
DeptID	Department ID code that matches the department the employee works in	Integer
PerfScoreID	Performance Score code that matches the employee's most recent performance score	Integer
FromDiversityJobFairID	Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no	Binary
PayRate	The person's hourly pay rate. All salaries are converted to hourly pay rate	Float
Termd	Has this employee been terminated - 1 or 0	Binary
PositionID	An integer indicating the person's position	Integer
Position	The text name/title of the position the person has	Text
State	The state that the person lives in	Text
Zip	The zip code for the employee	Text
DOB	Date of Birth for the employee	Date
Sex	Sex - M or F	Text
MaritalDesc	The marital status of the person (divorced, single, widowed, separated, etc)	Text
CitizenDesc	Label for whether the person is a Citizen or Eligible NonCitizen	Text
HispanicLatino	Yes or No field for whether the employee is Hispanic/Latino	Text
RaceDesc	Description/text of the race the person identifies with	Text
DateofHire	Date the person was hired	Date
DateofTermination	Date the person was terminated, only populated if, in fact, Termd = 1	Date
TermReason	A text reason / description for why the person was terminated	Text
EmploymentStatus	A description/category of the person's employment status. Anyone currently working full time = Active	Text
Department	Name of the department that the person works in	Text
ManagerName	The name of the person's immediate manager	Text
ManagerID	A unique identifier for each manager.	Integer
RecruitmentSource	The name of the recruitment source where the employee was recruited from	Text
PerformanceScore	Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)	Text
EngagementSurvey	Results from the last engagement survey, managed by our external partner	Float
EmpSatisfaction	A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey	Integer
SpecialProjectsCount	The number of special projects that the employee worked on during the last 6 months	Integer
LastPerformanceReviewDate	The most recent date of the person's last performance review.	Date
DaysLateLast30	The number of times that the employee was late to work during the last 30 days	Integer

Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these

features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The “Bounce Rate”, “Exit Rate” and “Page Value” features represent the metrics measured by “Google Analytics” for each page in the e-commerce site. The value of “Bounce Rate” feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session. The value of “Exit Rate” feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The “Page Value” feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The “Special Day” feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

4.2 数据预处理

1. 为了保护员工的个人隐私，我们对数据进行了脱敏处理。
2. 建模前，我们对原始数据进行了相关预处理，包括将数据中的缺失值、重复值、异常值的处理，对每个变量的数据分别进行标准化。对分类变量，我们采用因子化编码的处理方法，选定一个因子水平作为基准水平，将其余的因子水平拆分为各个虚拟变量 (Dummy Variables) 。
3. 我们将**主动辞职**和**被迫离职**的标记为离职，与**在职**相对应，生成一个虚拟变量。

4.3 描述分析

4.3.1 工作日/周末

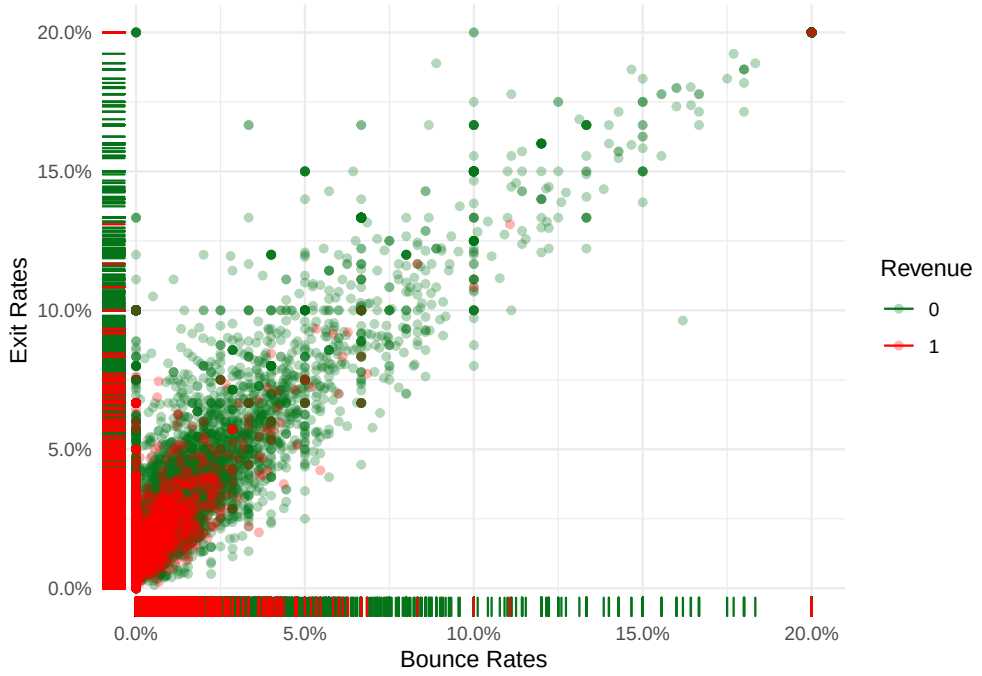

图 1: 离职与在职两类员工日薪分布密度图（红色代表离职）

离职的员工与在职的员工的有着相似的薪资分布情况，员工的时薪在 20 - 25 美元和 45 - 60 美元之间较为集中。但有所不同的是，离职的员工集中于 20 - 25 美元的比例更大，而集中于 45 - 60 美元的比例更小。这说明薪资过低的确是离职的一个较为重要的理由之一。

4.3.2 页面价值/周末

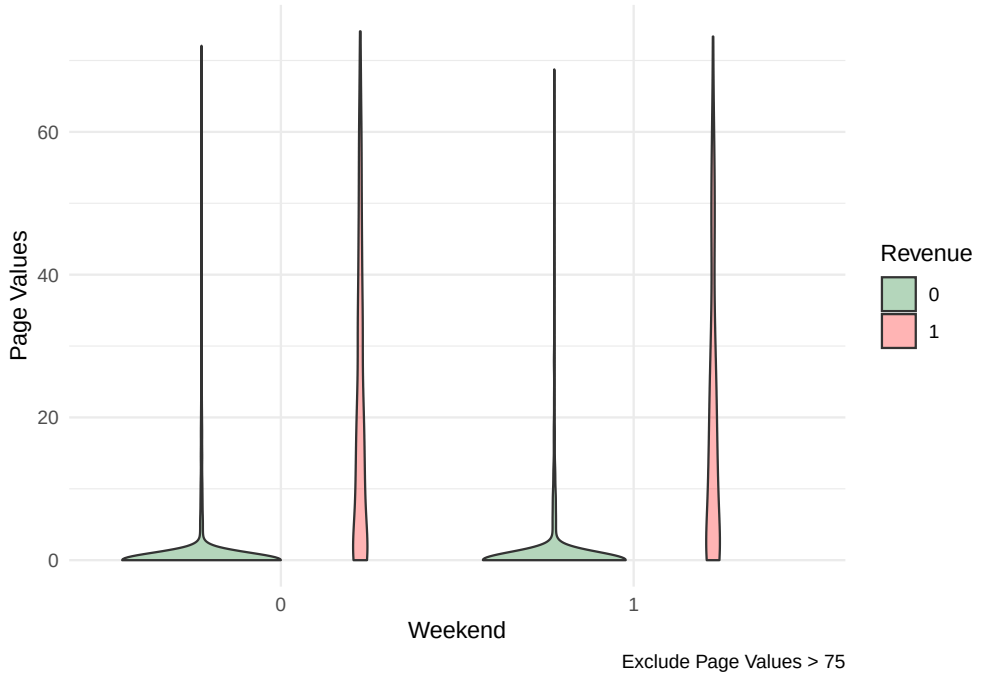

图 2: 不同性别员工日薪分布密度图（蓝色代表已婚）

女性相比男性，时薪在 20 - 25 美元的低水平处聚集更加明显。

已婚员工群体相对于未婚员工群体，最低薪资相对更高一些，低收入范围的平均工资更高一些，且中等收入范围的员工比例更多。考虑到婚姻状况与年龄强相关，我们认为这很有可能是由员工的工作经验所带来的影响。

4.3.3 产品相关/特殊日/收入

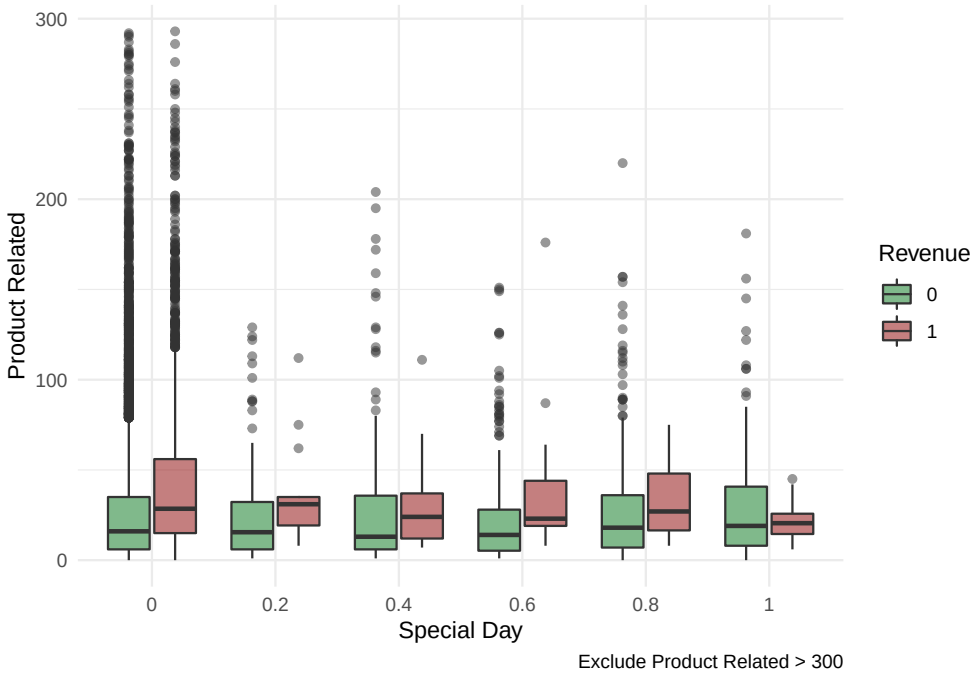

图 3: 离职与在职两类不同部门员工参与感箱线图（红色代表离职）

技术和销售部门员工的参与感不如生产和行政部门。从需求层次理论出发，员工有着实现自身价值的需要。对于 IT 和软件开发部门，离职者的参与感比在职者弱，他们离职的原因之一是没有获得足够的归属感和价值感。而对于销售和生产部门，离职着的参与感比在职者更强，工作太过繁忙可能是促成他们离职的一大原因。而对于行政岗位，没有足够的参与感往往意味着地位不足，可能收到部门的排挤和边缘化，促成了人员的离职。

4.3.4 绩效

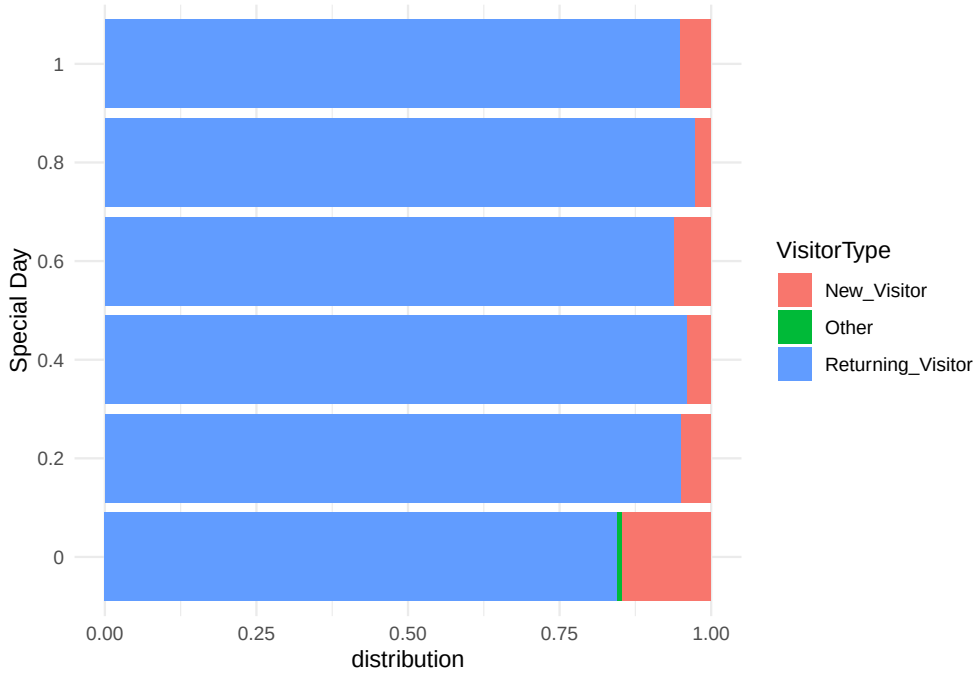

图 4: 不同任职状况的员工绩效（红色代表离职）

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.06143 0.00000 1.00000
```

我们发现:

1. 在职员工的绩效普遍都在优秀和良好两档
2. 自愿离职的员工绩效优秀的比例小一些
3. 被解雇的员工绩效为“需要提高”的比例非常高，同时解雇前已经被列入“解雇缓冲”的比例远高于在职和自愿离职的员工

可见绩效表现不好、不适应工作是员工自愿离职或被解雇的重要原因之一。

5 建立解释模型

5.1 拟合

因为 logit 模型相对简单，求解速度快，且具有较强的可解释性，故我们使用 logit 模型对样本进行拟合。[3]

我们将**离职**作为响应变量，选取的自变量有：

1. 性别
2. 婚姻状况：包括离婚、已婚、分居、未婚、配偶去世
3. 所在部门：包括行政部、总裁办公室、IT 部门、产品部门、销售部门、软件工程部门
4. 绩效：超过、符合要求、需要提高、进入淘汰流程
5. 员工参与感：1-5 员工自行打分
6. 员工满意度：1-5 员工自行打分
7. 在过去的 6 个月内员工进行的特殊项目个数
8. 时薪（美元）

对于连续型变量，我们直接将它们加入模型之中；对于因子型变量，我们将它们转换成为隐变量。

[3]模型详细见附录

表 2: Logit 回归系数表

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.88	0.17	-11.25	0.00
Administrative	0.00	0.01	0.35	0.73
Administrative_Duration	0.00	0.00	-0.56	0.58
Informational	0.03	0.03	1.10	0.27
Informational_Duration	0.00	0.00	0.37	0.71
ProductRelated	0.00	0.00	1.13	0.26
ProductRelated_Duration	0.00	0.00	2.23	0.03
BounceRates	-4.53	3.34	-1.36	0.17
ExitRates	-16.86	2.38	-7.08	0.00
PageValues	0.08	0.00	34.13	0.00
SpecialDay	-0.13	0.24	-0.56	0.58
MonthDec	-0.60	0.18	-3.32	0.00
MonthFeb	-1.82	0.64	-2.85	0.00
MonthJul	0.08	0.22	0.36	0.72
MonthJune	-0.32	0.27	-1.17	0.24
MonthMar	-0.52	0.18	-2.89	0.00
MonthMay	-0.57	0.17	-3.30	0.00
MonthNov	0.54	0.16	3.32	0.00
MonthOct	0.01	0.20	0.04	0.97
MonthSep	0.01	0.21	0.03	0.98

在 Z 检验的 p 值中，在众多的因素之中，只有**婚姻状况**中的 “分居” 和 “单身”，和**绩效表现**中的 “有待提高” 是统计上显著的。[4]

- 分居者和单身者的离职概率都显著较低，这可能与他们在经济上的独立性有关。分居者和单身者相比结婚合居者，在经济上不太依赖他人，有稳定的收入对他们来说更为重要，离职率自然较低一些。
- 绩效表现较差的员工离职概率也较高，这一方面可能是由于员工自身品质不佳或能力不足造成的不胜任岗位；另一方面也可能是员工与企业的文化不契合，对于工作内容或是上级不适应不喜欢；还有可能是企业处于末位淘汰制度或是效益不好，而对员工进行主动辞退的操作。

同时，**部门**中的销售部门、**绩效表现**中的 “良好” 和**特殊项目数量**这三个变量也有一定的显著性。[5]

[4]在 95% 置信区间下

[5]在 90% 置信区间下

- 相比技术部门，销售部门人员离职概率较低，这可能与销售人员在行业内专一产品方向所积累的经验和人脉有关。相比 IT 技术人员，销售人员的人脉可能更加局限于某一细分行业，跳槽的机会较少；而且，随着经验和人脉的积累，销售部门人员在企业内逐渐拿到更多的销售提成，对于企业的价值越来越大，企业对资深销售人员的待遇逐渐抬升；反过来，销售人员也一定程度上依赖着企业的平台，跳槽对于销售人员的不确定性较高。
- 特殊项目的数量与离职率负相关，从心理学的角度，这与员工的成就感和价值感有关，由于他们的工作不仅局限于日常工作，其它的项目推进让他们有更多的参与感和成就感，进而增强了对企业的归属感；同时，反过来说，参加特殊项目多的员工很可能本身就是为企业器重的核心人员，他们本身待遇和地位都较高，离职倾向不明显。

5.2 预测

我们对样本进行随机抽样，划分为 75% 的训练集和 25% 的测试集（验证集）。

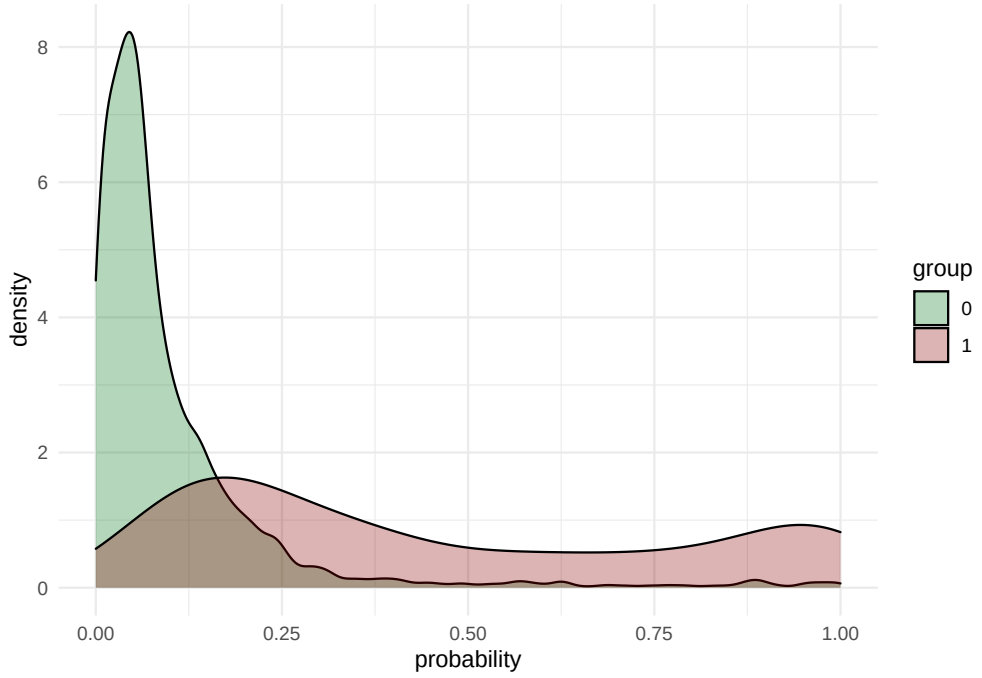

图 5: 预测的离职概率值（红色代表已知为离职）

从预测概率分布图，对于真实情况为在职的员工，我们预测出的离职概率值的分布是有偏的；比较之下，对于真实情况为在职的员工，我们预测出的离职概率值的分布则显得较为均匀。

对于真实情况为离职的员工，大部分得到的预测离职概率值的确都比较高。但对于真实情况为在职的员工，虽然大部分得到的预测离职概率较低，但仍然有相当一部分的预测离职概率值超过 50%。

为此，我们猜想：**我们的模型将没有离职倾向的员工错预测为离职的概率较低，但是较难识别出可能离职的员工。**

为了验证我们的猜想，我们使用混淆矩阵来计算预测模型的灵敏度和特异度。

5.3 混淆矩阵与验证结果

我们将预测概率大于 50% 的判定为离职。

灵敏度（Sensitivity）

$$\text{灵敏度} = \frac{\text{正确判定为“离职”的样本数量}}{\text{观测到的“离职”的样本数量}}$$

特异度（Specificity）

$$\text{特异度} = \frac{\text{正确判定为“在职”的样本数量}}{\text{观测到的“在职”的样本数量}}$$

假离职率

$$\text{假离职率} = 1 - \text{观测到的“在职”的样本数量}$$

表 3: 混淆矩阵表

Prediction	Reference	Freq
0	0	2534
1	0	71
0	1	286
1	1	191

表 4: 验证结果表

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.884	0.457	0.872	0.895	0.845	0	0

可以看到: 使用简单的 Logit 回归模型进行预测的准确率大致为: 59.2% , 95% 置信区间为 (0.5131, 0.7394) ，并不算高，甚至低于无信息准确率（即不经预测直接将所有员工归为在职）。但这并不代表模型是无用的。[6]

[6]事实上，在预测一些有偏分布的小概率事件时，模型准确率通常会低于无信息准确率。

表 5: 灵敏度和特异度等指标表

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision
指标值	0.4	0.973	0.729	0.899	0.729

从灵敏度和特异度来看：4% 的有离职倾向的员工会被模型成功捕捉到；对于模型捕捉到的员工，只有 13.7% 的误判率。

这验证了我们的猜测：**对于真正离职的员工，模型不一定能准确预测到；不过模型预测认为有离职倾向的员工在绝大部分情况下的确会发生离职。**

在模型准确度稳定的前提下，需要我们在灵敏度和特异度之间有所取舍。实际上，由于样本会更多的被认为是 “发生”，所以灵敏度上升会使特异度下降。这二者之间的潜在利弊的权衡是合理的，因为不同类型的错误对应不同的惩罚。在对员工是否会离职做识别和预测的时候我们通常关注特异度，只要模型能够捕捉到部分可能离职的员工，模型对于企业人力资源部门或是劳动力服务中心还是有很强的实用性的。

5.4 接受者操作特征（ROC）曲线

为了在灵敏度和特异度二者间权衡，我们使用接受者操作特征（ROC）曲线。

ROC 曲线 (Altman 和 Bland 1994; Brown 和 Davis 2006; Fawcett 2006) [5] [6] [7] 是一种常用方法, 在给定连续数据点集合的情况下，确定有效阈值，使阈值以上的值表示特定事件。ROC 曲线可以用来决定分类概率的阈值。

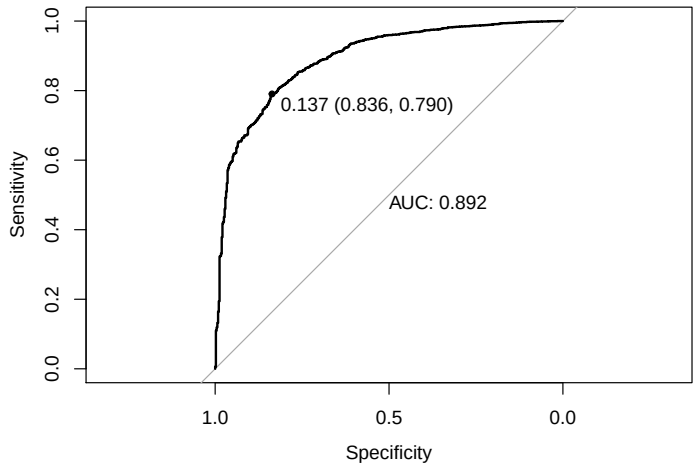

图 6: Logit 模型的 ROC 曲线

前文计算灵敏度和特异度时，我们默认 50% 概率阈值。为了捕获更多真阳性样本的方式提高灵敏度，我们可以通过降低阈值的方法，将灵敏度从 4% 提高到了 96% ，特异度从 86.3% 降低到了 35.3%。

也就是说，降低阈值有利于我们识别出更多有离职倾向的员工，但同时也会使误判的几率上升。

在实际操作中，我们可以通过**确定不同的阈值来达到不同的效果**，例如：

1. 当业务要求尽量减少筛选出的离职员工并减少错判时，可以通过提高阈值的方式增加特异度。
2. 当业务要求尽量识别覆盖范围更广时，可以通过降低阈值的方式提高灵敏度，以检测出更多潜在离职者以扩大服务范围。
3. 在进行人数评估时，通过平衡错判的成本与查漏的损失，确定适中的阈值以达到估计的准确性。

6 预测模型的选择

6.1 抽样、训练与评价指标

由于数据量较大，我们随机抽取部分数据用于模型的训练和验证，使用 10 折交叉验证，重复 2 次的方法进行重抽样，使用 Kappa 和准确率作为模型的评价指标。[7]

Kappa 统计量（Cohen 1960）[8] 最初是一个用来评估两个估价者评估结果的一致性，同时也考虑到了由偶然情况引起的准确性误差。

$$\text{Kappa} = \frac{O - E}{1 - E}$$

在上面的公式里，O 代表的是准确性，E 则代表着根据混淆矩阵边缘计数得出的期望准确性。0 值意味着观测类和预测类是不同的，1 值表示模型的预测与观测类是相同的，这个统计的量取值在-1 和 1 之间。虽然绝对值大的负数值在模型预测中出现的很少，但负数代表实际和预测值是相反的。总精确度在各类分布相同的时候与 Kappa 是成比例的。Kappa 值在 0.30 到 0.50 间代表着合理的一致性，这要依具体情况而定。（Agresti 2002）

6.2 Logit 回归

表 6: 在重抽样下 Logit 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.889	0.472	0.022	0.121

[7]由于数据集样本量过大，难以完成较为复杂的模型求解，且没有分布式计算的环境，我们从总样本中随机抽取 10% 的数据用于各种模型的训练和验证。

Logit 是一个受到非常广泛应用的模型，它十分简单、计算速度非常快，而且具有很强的可解释性。虽然 Logit 模型已经有很好的预测分类能力，但如果我们仅仅关注这一预测准确性这一指标，可能还有其它模型有更佳的表现。

6.3 线性判别分析（LDA）

Fisher（1936）[9] 和 Welch（1939）[10] 分析了获得最优判别准则的方式。

由贝叶斯法则：

$$\Pr[Y = C_\ell | X] = \frac{\Pr[Y = C_\ell] \Pr[X | Y = C_\ell]}{\sum_{\ell=1}^{C} \Pr[Y = C_\ell] \Pr[X | Y = C_\ell]}$$

对于二分类问题，如果：

$$\Pr[Y = C_1] \Pr[X | Y = C_1] > \Pr[Y = C_2] \Pr[X | Y = C_2]$$

我们就将 X 分入类别 1，否则分入类别 2。

为了计算 $\Pr[X|Y = C_\ell]$，我们假设预测变量服从多元正态分布，分布的两个参数为：多维均值向量 μ_ℓ 和协方差矩阵 Σ_ℓ，假设不同组的均值向量不同且协方差相同，用每一类观测样本均值 $\bar{x}_\ell$ 估计 μ_ℓ，用样本协方差 S 估计理论协方差矩阵 Σ，将样本观测 μ 代入 X，第 ℓ 组的线性判别函数为：

$$X'\Sigma^{-1}\mu_\ell - 0.5\mu_\ell'\Sigma^{-1}\mu_\ell + \log(\Pr[Y = C_\ell])$$

由于我们的分类只有两类，所以只有一个判别向量，不需要优化判别向量的数目，即不需要模型调优，计算速度较快。

当我们仔细观察线性判别函数时，我们会发现 Fisher 的线性判别方法有两点缺陷：

1. 而且，由于线性判别分析的数学构造，随着预测变量数目的增加，预测的类别概率越来越接近 0 和 1。这意味这，在我们的数据集下，由于变量较多，如前文所述的调整概率阈值的方法可能有效性会降低。这在单纯分类**在职倾向**和**离职倾向**的员工时可能并不是问题，但在需要进一步平衡灵敏度和特异度以达到更好效果时将很难进行。

2. 由于线性判别分析的结果取决于协方差矩阵的逆，且只有当这个矩阵可逆时才存在唯一解。这意味着样本量要大于变量个数[8]，且变量必须尽量相互独立。而在我们的数据集中，变量之间有很强的多重共线性，这在一定程度上会降低预测的准确性。

[8] 一般要求数据集含有至少预测变量 5——10 倍的样本

表 7: 在重抽样下 LDA 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.887	0.44	0.019	0.123

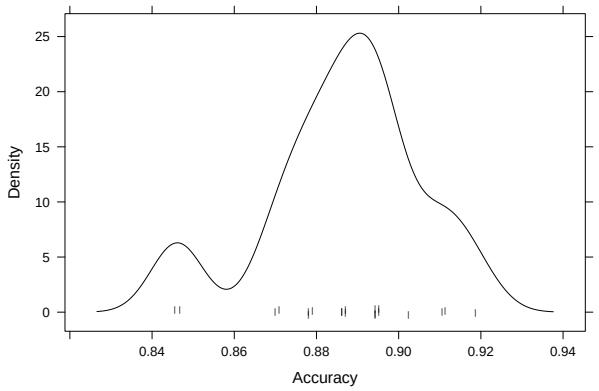

图 7: 在重抽样下 LDA 模型的准确率分布

6.4 偏最小二乘判别分析（PLSDA）

由于 LDA 不太适合多重共线性的变量，我们可以试着使用主成分分析压缩变量空间的维度，但 PCA 可能无法识别能将样本分类的较好变量组合，且由于没有涉及被解释变量的分类信息（无监督），很难通过 PCA 找到一个最优化的分类预测。

所以，我们使用偏最小二乘判别分析来进行分类。Berntsson 和 Wold（1986）[11] 将偏最小二乘应用在了问题中，起名为偏最小二乘判别分析（PLSDA）。尽管 Liu 和 Rayens（2007）[12] 指出，在降维非必须且建模目的时分类的时候，LDA 一定优于 PLS，但我们希望在降维之后，PLS 的表现能超过 LDA。

我们只使用前十个 PLS 成分

表 8: 在重抽样下 PLSDA 模型的表现

ncomp	Accuracy	Kappa	AccuracySD	KappaSD
1	0.844	0.016	0.005	0.036
2	0.844	0.039	0.008	0.056
3	0.847	0.120	0.019	0.118
4	0.879	0.355	0.019	0.130
5	0.880	0.358	0.019	0.132
6	0.878	0.349	0.017	0.120
7	0.878	0.349	0.017	0.121
8	0.878	0.343	0.017	0.123
9	0.878	0.349	0.017	0.121
10	0.877	0.343	0.016	0.112

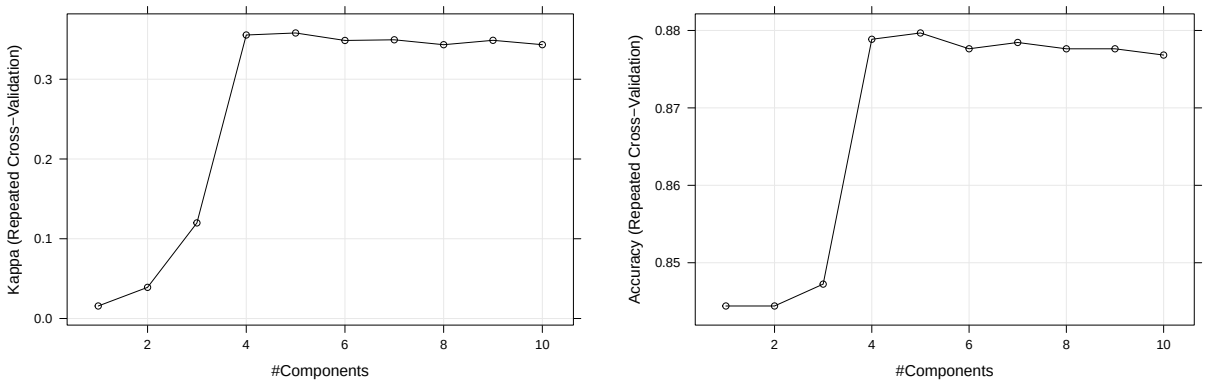

图 8: Kappa 指标和准确率随主成分个数的变化

我们可以看到 Kappa 指标随主成分个数的增多而先上升，后有所下降；准确率指标随着主成分个数的增加而先下降、后上升到顶峰、再下降。可见，在此模型中，选取前 8 个主成分不管是 Kappa 还是准确率指标都是最佳状态。

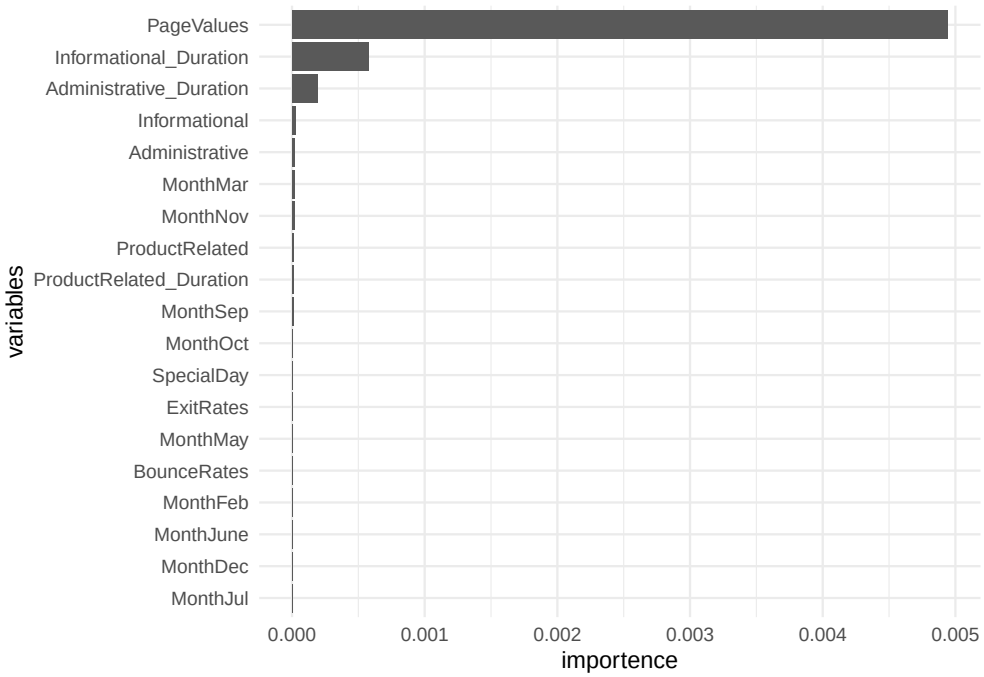

图 9: 变量重要程度

我们将变量按照其在 PLSDA 模型中的重要性进行排序：排在前三名的是 “婚姻状况中的独居”、“绩效表现中的较差一类” 和 “婚姻状况中的已婚”。这三个变量对于不同部门、不同工作内容、不同工作地位的员工具有较强的普适性。属于对员工个人的刻画，对于预测员工是否离职较为重要。

而重要程度最低的三个变量分别是 “薪资水平”、“完成项目的数量” 和 “是否在 IT/IS 部门”。这三个变量与员工个人的性格、工作能力、家庭关系较小，属于对工作分类的刻画，对于预测员工是否离职的重要性较低。

6.5 SVM

Logit、LDA、PLSDA 本质上都是线性模型，即模型结构产生线性类边界，这一类模型的优点是不太会受到无信息变量的干扰。然而，在我们的数据中，并没有存在大量无信息变量的情况，所以我们考虑使用非线性模型进行训练。

表 9: 在重抽样下 SVM 模型的表现

sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
0.061	0.25	0.881	0.370	0.017	0.125
0.061	0.50	0.886	0.420	0.016	0.103
0.061	1.00	0.893	0.478	0.018	0.117
0.061	2.00	0.895	0.508	0.018	0.113
0.061	4.00	0.897	0.524	0.017	0.101

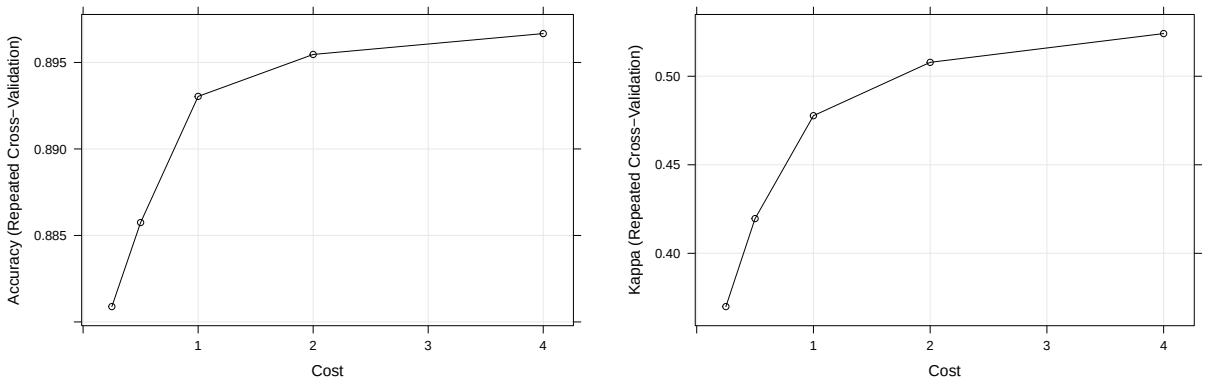

图 10: 调优参数不同取值下的准确率和 Kappa 指标变化

在损失参数增大的同时，准确率指标与 Kappa 指标的变化趋势相反，准确率有所降低而 Kappa 有所上升。

6.6 随机梯度助推法（GBM）

第三类被广泛应用的模型是分类树与基于规则的模型，在此，我们使用助推法这种树结构与规则的融合方法。

Friedman 等（2000）[13] 发现分类问题可以当作是正向分布可加模型，通过最小化指数损失函数实现分类。

首先我们设定样本预测初始值为对数发生：

$$f_i^{(0)} = \log \frac{\widehat{p}}{1-\widehat{p}}$$

其中，$f(x)$ 是模型的预测值，$\widehat{p}_i = \frac{1}{1+\exp[-f(x)]}$

接着从 $j = 1$ 开始进行迭代：

1. 计算梯度 $z_i = y_i - \hat{p}_i$
2. 对训练集随机抽样
3. 基于子样本，用之前得到的残差作为结果变量训练树模型
4. 计算终结点 Pearson 残差的估计 $r_i = \frac{1/n\sum_i^n (y_i - \hat{p}_i)}{1/n\sum_i^n \hat{p}_i(1-\hat{p}_i)}$
5. 更新当前模型 $f_1 = f_i + \lambda f_i^{(j)}$

表 10: 在重抽样下 GBM 模型的表现

	shrinkage	interaction.depth	n.minobsinnode	n.trees	Accuracy	Kappa	AccuracySD	KappaSD
1	0.1	1	10	50	0.889	0.535	0.024	0.104
4	0.1	2	10	50	0.898	0.569	0.019	0.102
7	0.1	3	10	50	0.904	0.591	0.018	0.087
2	0.1	1	10	100	0.892	0.552	0.023	0.104
5	0.1	2	10	100	0.899	0.579	0.019	0.095
8	0.1	3	10	100	0.899	0.571	0.023	0.110
3	0.1	1	10	150	0.890	0.541	0.024	0.111
6	0.1	2	10	150	0.897	0.573	0.019	0.087
9	0.1	3	10	150	0.894	0.551	0.021	0.094

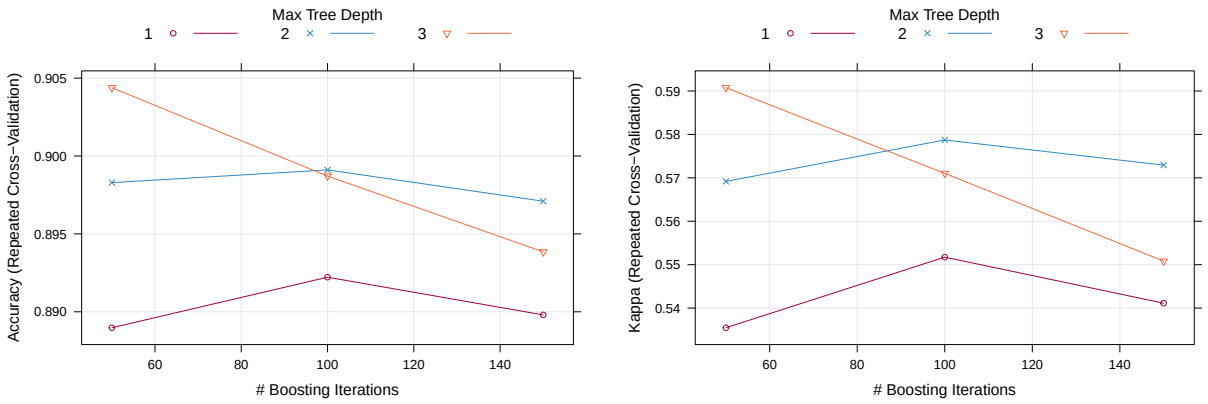

图 11: 调优参数和迭代次数不同取值下的准确率和 Kappa 指标变化

助推树的加深和迭代次数的增多一般引起 Kappa 指标的上升，随着迭代次数的增加，准确率变动先下降后上升。

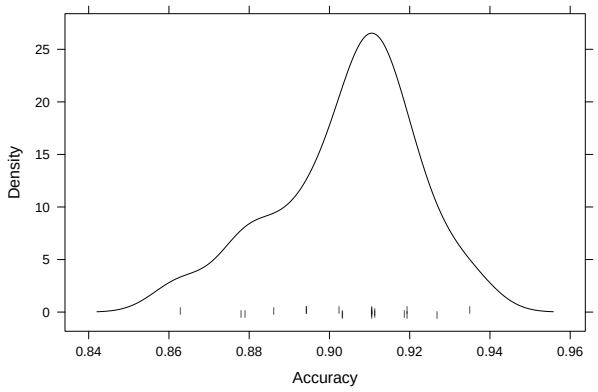

图 12: 在重抽样下 GBM 模型的准确率分布

6.7 模型间的比较

我们对训练的 4 个不同的模型进行比较，所有模型都使用相同的重抽样方法估计各自的模型表现。且由于设置的随机数种子相同，故不同模型使用的重抽样样本完全一致。[9]

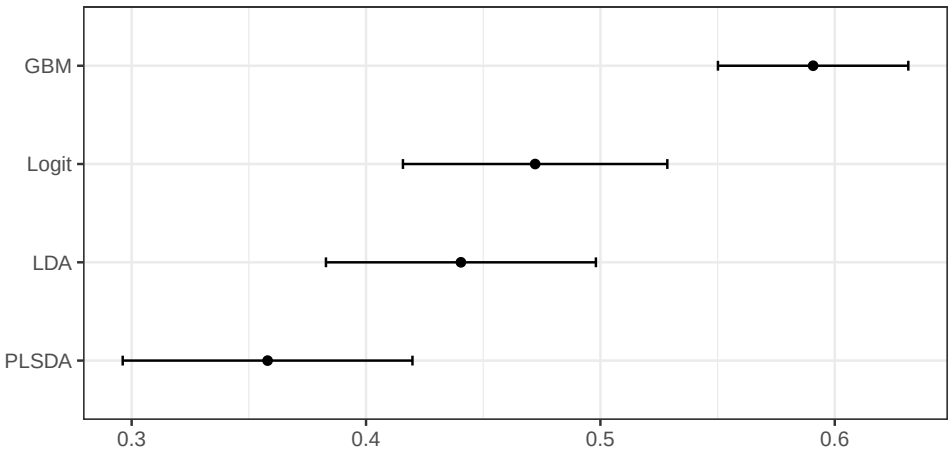

图 13: 模型间 Kappa 的比较（0.95 置信区间）

[9] 重抽样 50 次：10 折交叉验证重复 5 次

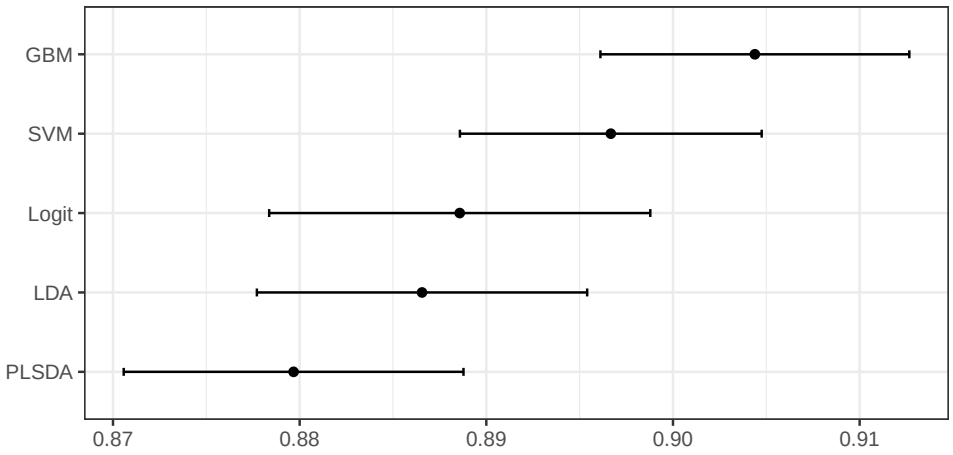

图 14: 模型间准确率的比较（0.95 置信区间）

在 **Kappa** 这一效果衡量指标下，PLSDA 有着最好的效果，LDA 和 Logit 模型次之，GBM 模型远差于前面 3 个模型。

在**准确率**这一效果衡量指标下，从偏差的角度来看，PLSDA 有着最好的效果，SVM 模型次之；从方差的角度来看，SVM 模型具有明显较小的方差。

7 结论

在此研究中，我们主要研究了企业**员工的离职预测问题**。我们通过这个研究一个案例，所得出的该企业结论员工离职情况，对整个人力市场有一定的启发性。

于此同时，我们研究此案例的方法具有较好的普适性。对于此研究建立的多种预测离职预测模型，完全可以在其它企业中适当地调整后加以应用。

同时该研究的受益者不仅仅是企业，就业市场服务中心、社会科学研究在涉及到人力资源流动时，均可参考这些模型和方法，对人力资源的流动进行方向上和比例上的评估和监测，为商业服务、政策制定提供解决方案。

7.1 变量解释

在我们的案例研究中，众多的变量中有一些变量是在统计学上显著的。[10]

- 分居者和单身者的离职概率都显著较低，由职业生涯理论，这可能与他们在经济上的独立性有关。分居者和单身者相比结婚合居者，在经济上不太依赖他人，有稳定的收入对他们来说更为重要，离职率自然较低一些。
- 绩效表现较差的员工离职概率也较高。由双因素理论[11] 员工对不满意因素的心理感受强于激励因素，容易引发主动离职。而绩效评定较差的原因，这一方面可能是由于员工自身品质不佳或能力不足造成的不胜任岗位；另一方面也可能是员工与企业的文化不契合，对于工作内容或是上级不适应不喜欢；还有可能是企业处于末位淘汰制度或是效益不好，而对员工进行主动辞退的操作。
- 相比技术部门，销售部门人员离职概率较低，这可能与销售人员在行业内专一产品方向所积累的经验和人脉有关。相比 IT 技术人员，销售人员的人脉可能更加局限于某一细分行业，跳槽的机会较少；而且，随着经验和人脉的积累，销售部门人员在企业内逐渐拿到更多的销售提成，对于企业的价值越来越大，企业对资深销售人员的待遇逐渐抬升；反过来，销售人员也一定程度上依赖着企业的平台，跳槽对于销售人员的不确定性较高。
- 特殊项目的数量与离职率负相关，由需求层次理论，这与员工的成就感和价值感有关，由于他们的工作不仅局限于日常工作，其它的项目推进让他们有更多的参与感和成就感，进而增强了对企业的归属感；同时，反过来说，参加特殊项目多的员工很可能本身就是为企业器重的核心人员，他们本身待遇和地位都较高，离职倾向不明显。

7.2 阈值选择

结合具体的业务，为了达到最高的效率，我们可以通过**确定不同的预测阈值来达到不同的效果**，例如：

1. 在企业进行潜在离职者的一对一谈心和了解情况时，可以通过提高阈值的方法提高特异度，以尽量避免错判。

[10] 在 90% 置信区间下

[11] （two factor theory）亦称“激励一保健理论”

2. 在就业服务中心进行潜在离职者的筛选时，通过降低阈值的方式提高灵敏度，以检测出更多潜在离职者以扩大服务范围。
3. 在政策制定需要估计离职率时，通过平衡错判的成本与查漏的损失，确定适中的阈值以达到估计的准确性。

7.3 模型选择

在 **Kappa** 这一效果衡量指标下，PLSDA 有着最好的效果，LDA 和 Logit 模型次之，GBM 模型远差于前面 3 个模型。

在**准确率**这一效果衡量指标下，从偏差的角度来看，PLSDA 有着最好的效果，SVM 模型次之；从方差的角度来看，SVM 模型具有明显较小的方差。

综合来看，**PLSDA** 模型具有最好的效果。然而，在模型的应用方面，由于 Logit 模型计算速度较快、可解释性强的，在对准确率要求不高而更加重视变量的可解释性的场景下，Logit 模型也不失为一个较好的选择。

7.4 变量选择

在 PLSDA 模型中的各变量重要性排序：排在前三名的是 “婚姻状况中的独居”、“绩效表现中的较差一类” 和 “婚姻状况中的已婚”。这三个变量对于不同部门、不同工作内容、不同工作地位的员工具有较强的普适性。属于对员工个人的刻画，对于预测员工是否离职较为重要。

而重要程度最低的三个变量分别是 “薪资水平”、“完成项目的数量” 和 “是否在 IT/IS 部门”。这三个变量与员工个人的性格、工作能力、家庭关系较小，属于对工作分类的刻画，对于预测员工是否离职的重要性较低。

7.5 模型应用

1. 从企业角度：通过对员工是否将离职进行预测，可以为企业提前找到潜在的离职员工，并提前作出应对策略，通过改进用人制度和政策等措施留住企业并不想解雇的员工，以减小离职率。在减小离职率的同时，公司可以通过对有离职倾向的员工数量进行评估，提前准备后备人才以便随时顶岗，减小因个别人才的流失带来的损失。
2. 从社会角度：利用大数据，通过对社区人员的信息统计，对数据进行脱敏处理后，可以预测人力资源流动、监测摩擦性失业指标，为政策决策提供依据，达到减小社会失业率的目的。

8 参考文献

[1] 张梓嫣. BJM 公司新员工离职问题分析及对策研究 [D]. 江苏大学, 2019.

[2] 杨喆麟. 星巴克 (中国) 公司员工离职问题分析与优化策略 [D]. 兰州大学, 2017.

[3] 赵西萍, 刘玲, 张长征. A Multi- variable Analysis on Factors Influencing Employee's Turnover Intention[J]. 中国软科学, 2003, 000(3): 71–74.

[4] 张紫君. 企业员工的离职预测模型 [D]. 重庆大学, 2018.

[5] ALTMAN, DOUGLAS, G., 等. Diagnostic tests 3: receiver operating characteristic plots.[J]. Bmj British Medical Journal, 1994.

[6] BROWN C D, DAVIS H T. Receiver operating characteristics curves and related decision measures: A tutorial[J]. Chemometrics & Intelligent Laboratory Systems, 2006, 80(1): 24–38.

[7] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861–874.

[8] COHEN J A. A Coefficient of Agreement for Nominal Scales[J]. Educational & Psychological Measurement, 1960, 20(1): 37–46.

[9] FISHER R A. The Use of Multiple Measurements in Taxonomic Problems[J]. Annals of Eugenics, 1936, 7(7): 179–188.

[10] L. W B. (ii) Note on Discriminant Functions[J]. Biometrika, 1939(1-2): 1–2.

[11] BERNTSSON P, WOLD S. Comparison Between X-Ray Crystallographic Data and Physicochemical Parameters with Respect to Their Information about the Calcium Channel Antagonist Activity of 4-Phenyl-1,4-dihydropyridines[J]. Quantitative Structure Activity Relationships, 1986, 5(2): 45–50.

[12] LIU Y, RAYENS W. PLS and dimension reduction for classification[J]. Computational Statistics, 2007, 22(2): 189–208.

[13] BEN-DOR, AMIR, BRUHN, 等. Tissue Classification with Gene Expression Profiles[J]. Journal of Computational Biology, 2000.

9 附录

9.1 模型间准确率和 Kappa 的比较

表 11: 模型间准确率的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.846	0.878	0.887	0.887	0.895	0.919	0
PLSDA	0.839	0.868	0.879	0.880	0.886	0.919	0
SVM	0.862	0.893	0.899	0.897	0.903	0.927	0
GBM	0.863	0.894	0.911	0.904	0.913	0.935	0
Logit	0.839	0.884	0.891	0.889	0.903	0.919	0

表 12: 模型间准确率差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		0.006881	-0.010113	-0.017830	-0.002019
PLSDA	0.406844		-0.016994	-0.024712	-0.008900
SVM	0.367854	0.026244		-0.007717	0.008094
GBM	0.001186	2.984e-05	0.638866		0.015811
Logit	1.000000	0.435876	0.860707	0.003634	

表 13: 模型间 Kappa 的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.179	0.352	0.442	0.440	0.530	0.635	0
PLSDA	0.162	0.244	0.364	0.358	0.416	0.603	0
SVM	0.266	0.491	0.535	0.524	0.576	0.676	0
GBM	0.388	0.545	0.610	0.591	0.645	0.751	0
Logit	0.245	0.407	0.498	0.472	0.571	0.635	0

表 14: 模型间 Kappa 差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		0.006881	-0.010113	-0.017830	-0.002019
PLSDA	0.406844		-0.016994	-0.024712	-0.008900
SVM	0.367854	0.026244		-0.007717	0.008094
GBM	0.001186	2.984e-05	0.638866		0.015811
Logit	1.000000	0.435876	0.860707	0.003634	

9.2 Logit 回归结果

```
##
## Call:
## glm(formula = Revenue ~ Administrative + Administrative_Duration +
##     Informational + Informational_Duration + ProductRelated +
##     ProductRelated_Duration + BounceRates + ExitRates + PageValues +
##     SpecialDay + Month, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.1120  -0.4669  -0.3408  -0.1629   3.5081
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.882e+00  1.673e-01 -11.251  < 2e-16 ***
## Administrative            3.835e-03  1.095e-02   0.350 0.726141
## Administrative_Duration -1.084e-04  1.936e-04  -0.560 0.575700
## Informational             2.968e-02  2.690e-02   1.103 0.269967
## Informational_Duration    8.065e-05  2.208e-04   0.365 0.714905
## ProductRelated            1.285e-03  1.141e-03   1.127 0.259846
## ProductRelated_Duration   5.977e-05  2.680e-05   2.230 0.025726 *
## BounceRates              -4.533e+00  3.341e+00  -1.357 0.174871
## ExitRates                -1.686e+01  2.380e+00  -7.082 1.42e-12 ***
## PageValues                8.198e-02  2.402e-03  34.134  < 2e-16 ***
## SpecialDay               -1.323e-01  2.369e-01  -0.558 0.576707
## MonthDec                 -6.016e-01  1.814e-01  -3.316 0.000914 ***
## MonthFeb                 -1.819e+00  6.385e-01  -2.849 0.004393 **
## MonthJul                  7.833e-02  2.180e-01   0.359 0.719360
```

```
## MonthJune               -3.214e-01  2.741e-01  -1.173 0.240938
## MonthMar                -5.196e-01  1.795e-01  -2.894 0.003802 **
## MonthMay                -5.725e-01  1.733e-01  -3.303 0.000957 ***
## MonthNov                 5.372e-01  1.620e-01   3.316 0.000914 ***
## MonthOct                 7.194e-03  2.013e-01   0.036 0.971489
## MonthSep                 6.113e-03  2.123e-01   0.029 0.977024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10624.8  on 12329  degrees of freedom
## Residual deviance:  7179.6  on 12310  degrees of freedom
## AIC: 7219.6
##
## Number of Fisher Scoring iterations: 7
```

9.3 数据指标明细

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : Factor w/ 8 levels "1","2","3","4",..: 1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : Factor w/ 13 levels "1","2","3","4",..: 1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : Factor w/ 9 levels "1","2","3","4",..: 1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ Revenue                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

##  Administrative   Administrative_Duration Informational
```

```
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser         Region      TrafficType
##  May    :3364   2      :6601     2      :7961   1      :4780   2      :3913
##  Nov    :2998   1      :2585     1      :2462   3      :2403   1      :2451
##  Mar    :1907   3      :2555     4      : 736   4      :1182   3      :2052
##  Dec    :1727   4      : 478     5      : 467   2      :1136   4      :1069
##  Oct    : 549   8      :  79     6      : 174   6      : 805   13     : 738
##  Sep    : 448   6      :  19     10     : 163   7      : 761   10     : 450
##  (Other):1337   (Other):  13     (Other): 367   (Other):1263   (Other):1657
##            VisitorType    Weekend  Revenue
##  New_Visitor      : 1694   0:9462   0:10422
##  Other            :   85   1:2868   1: 1908
##  Returning_Visitor:10551
##
##
##
##
```