# 摘要

待完善

# 背景

# 目的

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，通过一定措施留住企业并不想解雇的员工，以减小离职率。
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。

# 数据

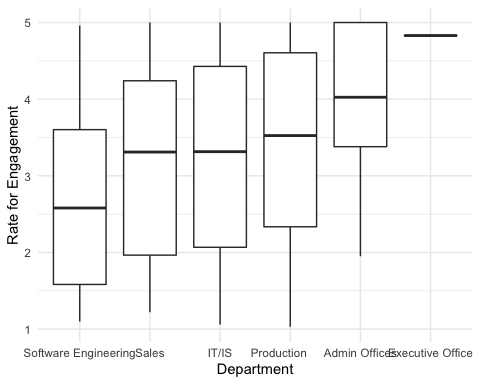
数据进行脱敏处理

## [1] "EmpID" "MarriedID"   
## [3] "MaritalStatusID" "GenderID"   
## [5] "EmpStatusID" "DeptID"   
## [7] "PerfScoreID" "FromDiversityJobFairID"   
## [9] "PayRate" "Termd"   
## [11] "PositionID" "Position"   
## [13] "State" "Zip"   
## [15] "DOB" "Sex"   
## [17] "MaritalDesc" "CitizenDesc"   
## [19] "HispanicLatino" "RaceDesc"   
## [21] "DateofHire" "DateofTermination"   
## [23] "TermReason" "EmploymentStatus"   
## [25] "Department" "ManagerName"   
## [27] "ManagerID" "RecruitmentSource"   
## [29] "PerformanceScore" "EngagementSurvey"   
## [31] "EmpSatisfaction" "SpecialProjectsCount"   
## [33] "LastPerformanceReview\_Date" "DaysLateLast30"

共有 310 个观测

# 描述分析

## 参与感



不同部门员工参与感箱线图（评分1-5）

# Logit 回归

## 数据处理

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。

## 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。

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，

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）。

Logit回归系数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 0.70 | 1.52 | 0.46 | 0.65 |
| SexM | -0.04 | 0.27 | -0.16 | 0.88 |
| MaritalDescMarried | -0.57 | 0.43 | -1.33 | 0.18 |
| MaritalDescSeparated | -2.60 | 1.13 | -2.30 | 0.02 |
| MaritalDescSingle | -1.15 | 0.43 | -2.67 | 0.01 |
| MaritalDescWidowed | 0.08 | 0.85 | 0.09 | 0.93 |
| DepartmentExecutive Office | -14.66 | 882.74 | -0.02 | 0.99 |
| DepartmentIT/IS | 1.32 | 1.21 | 1.09 | 0.28 |
| DepartmentProduction | -1.05 | 1.10 | -0.96 | 0.34 |
| DepartmentSales | -2.10 | 1.22 | -1.72 | 0.09 |
| DepartmentSoftware Engineering | 1.51 | 1.22 | 1.24 | 0.22 |
| PerformanceScoreFully Meets | 0.77 | 0.45 | 1.73 | 0.08 |
| PerformanceScoreNeeds Improvement | 1.61 | 0.65 | 2.47 | 0.01 |
| PerformanceScorePIP | 1.14 | 0.84 | 1.36 | 0.17 |
| EngagementSurvey | 0.02 | 0.10 | 0.21 | 0.83 |
| EmpSatisfaction | 0.06 | 0.16 | 0.38 | 0.71 |
| SpecialProjectsCount | -0.52 | 0.28 | -1.91 | 0.06 |
| PayRate | -0.01 | 0.02 | -0.90 | 0.37 |

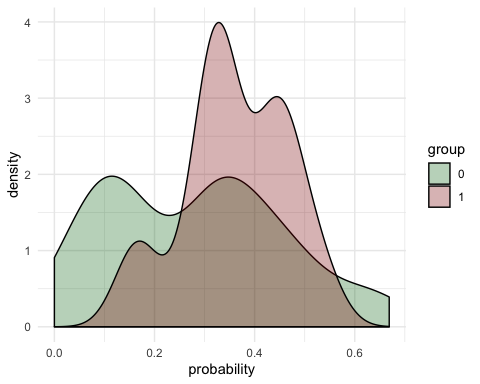
无担保放款的循环利用次数越多，逾期可能性越高。由于个人信用总额度大，借款的数量多，从而导致还款不及时或者到指定日期还款能力不足。

年龄越大逾期可能性越小，但影响程度不高。年龄越大工作生活的状态越稳定，心智更加成熟，收入会比年龄小的人群更多，需要使用信用卡借款的可能性越低。

过去两年间逾期30-59天的次数越多，逾期的可能性越高。有过短时间逾期经历的人群对截止日期的敏感程度会降低，重视程度也会下降，从而导致超过还款日期的可能性增高。

负债比率越高，逾期的可能性越高。日常开销和债务占收入的比例越大，可支配金额越少，此时可能个人基本生活开销都会出现问题，还款的可能性降低。 月收入的增加会使得逾期的可能性降低。月收入越高，可支配金额越多，还款能力越强，但是否及时还款还是需要根据个人还款习惯。即使收入增加，借贷人忘记还款日期也依然会造成逾期。 未偿还贷款数量越多会导致逾期的可能性增加。有大量负债说明资金链已经出现问题，需要偿还的债务很多，及时归还信用卡贷款的可能性低。 90天逾期次数越多，逾期的可能性越高。在这种情况下，持卡人出于各种自身原因长期拖欠，陷入资金不正常循环。 不动产贷款或额度数量越多，逾期可能性越大。不动产贷款或者额度越多，会使持卡人背负较大的偿还压力，一旦收入来源出现问题，将导致信用卡无法如期偿还，逾期概率增加。 家属人数越多，日常开支越大，从而导致用于还款的收入剩余越少，逾期的可能性越大。

## 预测



预测的逾期概率值（红色代表已知为逾期）

可以看出，对于真实情况为信用好的持卡人，我们预测出的逾期概率值的分布是有偏的，大多数预测概率的非常低。然而，比较之下，对于真实情况为逾期的持卡人，我们预测出的逾期概率值的分布则显得较为均匀。

为此，我们猜想：**我们的模型将信用好的持卡人错认为逾期的概率较低，但是较难识别出逾期的客户。**

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。

## 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）

特异度（Specificity）

假阳性率为 1 - 特异度

混淆矩阵表

|  |  |  |
| --- | --- | --- |
| Prediction | Reference | Freq |
| 0 | 0 | 46 |
| 1 | 0 | 23 |
| 0 | 1 | 5 |
| 1 | 1 | 2 |

验证结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | Kappa | AccuracyLower | AccuracyUpper | AccuracyNull | AccuracyPValue | McnemarPValue |
| 0.632 | -0.022 | 0.513 | 0.739 | 0.908 | 1 | 0.001 |

## 0 1   
## 207 103

## [1] 0.6677419

可以看到：准确率大致为： 0.632，置信区间 XXX ，并不算高。

灵敏度和特异度等指标表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sensitivity | Specificity | Pos Pred Value | Neg Pred Value | Precision |
| 指标值 | 0.286 | 0.667 | 0.08 | 0.902 | 0.08 |

从灵敏度和特异度来看：55.8% 的将会逾期的客户会被模型成功捕捉到；对于模型捕捉到的客户，只有 6.7% 的误判率。

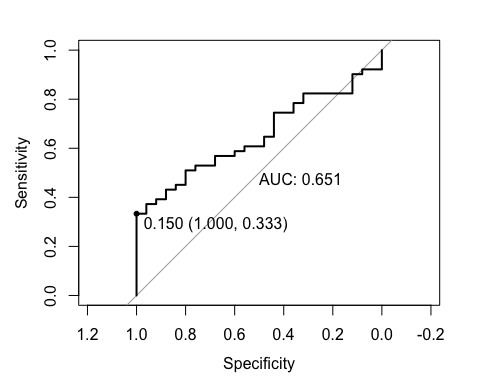
这验证了我们的猜测：**当持卡人逾期时，模型不一定能准确预测到；不过模型预测认为是逾期的客户绝大部分情况下的确会发生逾期**

在模型准确度稳定的前提下，灵敏度和特异度之间需要我们有所取舍。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。或许在二者之间的潜在权衡利弊是合理的，因为不同类型的错误会导致不同的惩罚。在对信用卡是否会逾期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逾期客户，信用卡风控部门还是可以使用模型进行预测的。

## 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) [1] [2] [3] 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。



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。将阈值降低至 6.3% , 此时，灵敏度从 55.8% 提高到了 71.9% ，特异度从 93.3% 降低到了 74.1%。

也就是说，降低阈值有利于我们识别出更多逾期的持卡人，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. 在进行交易风控、信用卡降额的自动化系统构建时，通过确定较高的阈值以提高特异度，避免错判。
2. 在进行逾期自动化预测以便于进一步调查时，通过降低阈值的方式提高灵敏度，以检测出更多潜在逾期持卡人。
3. 通过平衡错判的成本与查漏的损失，确定适中的阈值以谋求商业利益最大化。

# 模型选择

## 抽样、训练与评价指标

由于数据集样本量过大，难以完成较为复杂的模型求解。[[1]](#footnote-37)我们从总样本中随机抽取 1% 的数据用于各种模型的训练和验证。

我们使用10折交叉验证，重复5次的方法进行重抽样。

我们使用 Kappa 和准确率作为模型的评价指标。

Kappa 统计量（Cohen 1960） [4] 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。

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）

## Logit 回归

在重抽样下 Logit 模型的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| parameter | Accuracy | Kappa | AccuracySD | KappaSD |
| none | 0.653 | 0.096 | 0.057 | 0.138 |

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。

## 线性判别分析（LDA）

Fisher（1936）[5] 和 Welch（1939）[6] 分析了获得最优判别准则的方式。

由贝叶斯法则：

对于二分类问题，如果：

我们就将 X 分入类别1，否则分入类别2。

为了计算 ，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 和协方差矩阵 ，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 估计 ，用样本协方差 估计理论协方差矩阵 ，将样本观测 代入 ，第 组的线性判别函数为：

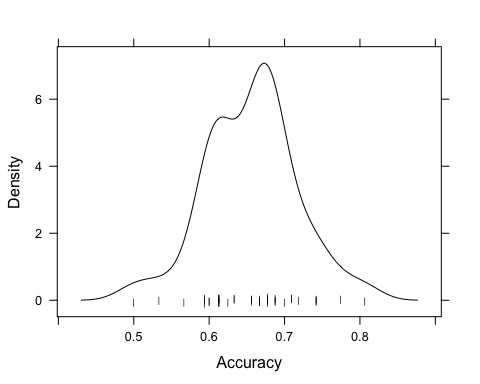
由于我们的分类只有两类，所以只有一个判别向量，不需要优化判别向量的数目，即不需要模型调优，计算速度较快。

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**逾期**和**信用良好**的持卡人时可能并不是问题，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[[2]](#footnote-40)，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。

在重抽样下 LDA 模型的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| parameter | Accuracy | Kappa | AccuracySD | KappaSD |
| none | 0.656 | 0.104 | 0.06 | 0.142 |



在重抽样下 LDA 模型的准确率分布

## 偏最小二乘判别分析（PLSDA）

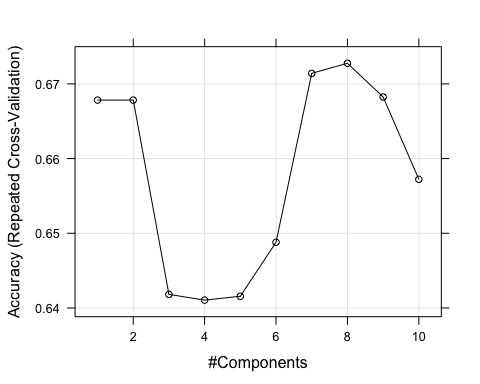
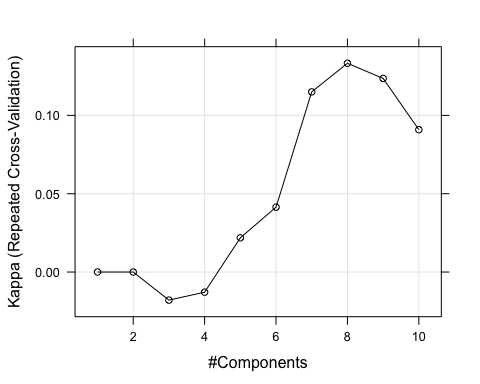
由于 LDA 不太适合多重共线性的变量，我们可以试着使用主成分分析压缩变量空间的维度，但 PCA 可能无法识别能将样本分类的较好变量组合，且由于没有涉及被解释变量的分类信息（无监督），很难通过 PCA 找到一个最优化的分类预测。

所以，我们使用偏最小二乘判别分析来进行分类。Berntsson 和 Wold（1986） [7] 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） [8] 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。

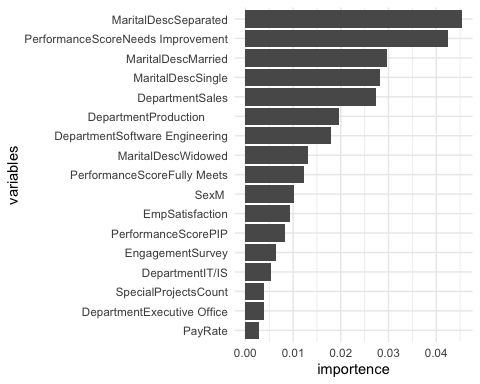
我们只使用前十个 PLS 成分

在重抽样下 PLSDA 模型的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| parameter | Accuracy | Kappa | AccuracySD | KappaSD |
| none | 0.656 | 0.104 | 0.06 | 0.142 |



我们可以看到 Kappa 指标随主成分个数的增多而先上升，后基本保持不变。可见，在此模型中，选取前 5 个主成分效率最高。



变量重要程度

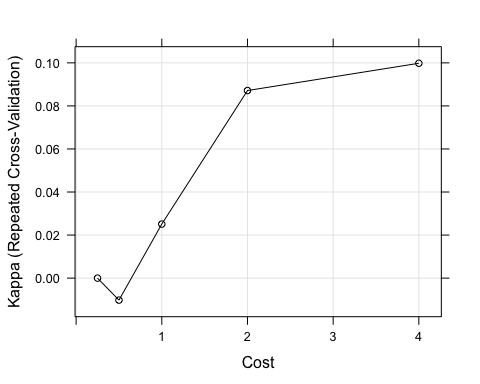
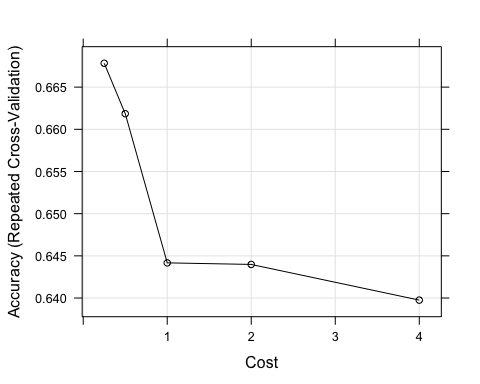
由上图所示，各个变量的重要程度有明显不同。排在前三名的是“过去逾期30-59天的次数”,“逾期超过90天的次数”和“过去两年内逾期60-89天的次数”。这三个变量都属于同一类型的变量，它们描述的都是借贷者过去是否有信用卡逾期的先例，它们代表的数据能很好地反映出借贷者的信誉情况。过去出现过逾期的先例，那么未来信用卡逾期的可能性就大大增加。而重要程度最低的三个变量分别是“家属人数”“负债比率”和“月收入”。这三个变量与个人信誉的关系不大，属于外界因素。即使家属人数多，负债比率高，月收入低，借贷人依旧可以通过合理规划及时还款。即借贷人如果有良好的还贷习惯，这些变量的影响较低。

## 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。

在重抽样下 SVM 模型的表现

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sigma | C | Accuracy | Kappa | AccuracySD | KappaSD |
| 0.059 | 0.25 | 0.668 | 0.000 | 0.010 | 0.000 |
| 0.059 | 0.50 | 0.662 | -0.010 | 0.021 | 0.038 |
| 0.059 | 1.00 | 0.644 | 0.025 | 0.053 | 0.123 |
| 0.059 | 2.00 | 0.644 | 0.087 | 0.069 | 0.142 |
| 0.059 | 4.00 | 0.640 | 0.100 | 0.071 | 0.162 |



## 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。

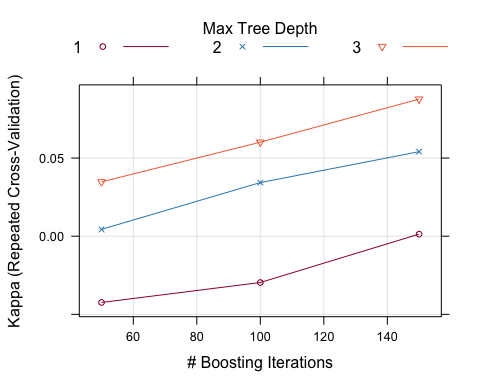
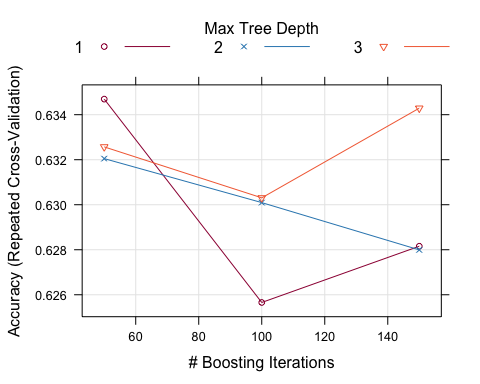
Friedman等（2000） [9] 发现分类问题可以当作是正向分布可加模型，通过最小化指数损失函数实现分类。

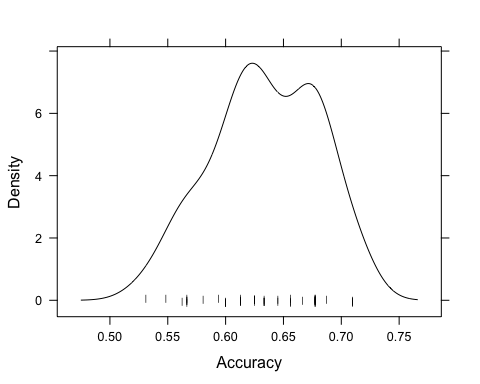
首先我们设定样本预测初始值为对数发生：

其中， 是模型的预测值，

接着从 开始进行迭代：

1. 计算梯度
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
5. 更新当前模型

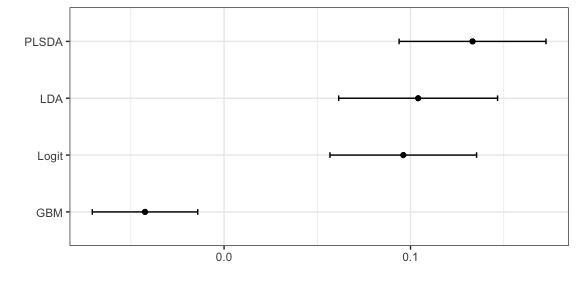




在重抽样下 GBM 模型的准确率分布

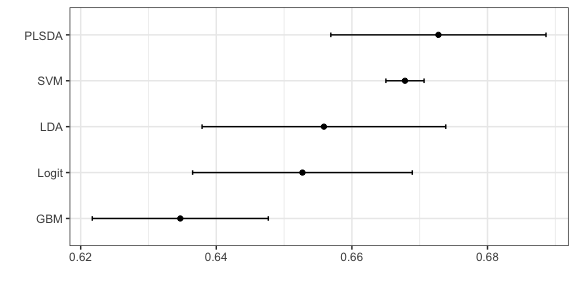
## 模型间的比较

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。[[3]](#footnote-54)



模型间 Kappa 的比较（0.95 置信区间）

在**Kappa**这一效果衡量指标下，GBM 有着最好的效果，Logit 模型次之，PLSDA 模型表现最差。



模型间准确率的比较（0.95 置信区间）

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

综合来看，**GBM**模型具有最好的效果，**Logit**模型次之。然而，在模型的应用方面，我们更加倾向于使用计算速度较快、可解释性强的 Logit 模型。

# 附录

## 数据

## 'data.frame': 310 obs. of 34 variables:  
## $ EmpID : int 1103024456 1106026572 1302053333 1211050782 1307059817 711007713 1504073368 1403065721 1408069481 1306059197 ...  
## $ MarriedID : int 1 0 0 1 0 1 1 0 0 1 ...  
## $ MaritalStatusID : int 1 2 0 1 0 1 1 0 0 1 ...  
## $ GenderID : int 0 1 1 0 0 0 0 0 0 1 ...  
## $ EmpStatusID : int 1 1 1 1 1 5 5 1 1 1 ...  
## $ DeptID : int 1 1 1 1 1 1 6 6 6 6 ...  
## $ PerfScoreID : int 3 3 3 3 3 3 3 3 1 3 ...  
## $ FromDiversityJobFairID : int 1 0 0 0 0 1 0 0 0 0 ...  
## $ PayRate : num 28.5 23 29 21.5 16.6 ...  
## $ Termd : int 0 0 0 1 0 1 1 0 0 0 ...  
## $ PositionID : int 1 1 1 2 2 2 3 3 3 3 ...  
## $ Position : Factor w/ 32 levels "Accountant I",..: 1 1 1 2 2 2 3 3 3 3 ...  
## $ State : Factor w/ 28 levels "AL","AZ","CA",..: 11 11 11 11 11 11 26 27 28 16 ...  
## $ Zip : int 1450 1460 2703 2170 2330 1844 21851 5664 98052 3062 ...  
## $ DOB : Factor w/ 306 levels "01/02/51","01/04/64",..: 283 87 204 223 115 139 129 110 109 219 ...  
## $ Sex : Factor w/ 2 levels "F","M ": 1 2 2 1 1 1 1 1 1 2 ...  
## $ MaritalDesc : Factor w/ 5 levels "Divorced","Married",..: 2 1 4 2 4 2 2 4 4 2 ...  
## $ CitizenDesc : Factor w/ 3 levels "Eligible NonCitizen",..: 3 3 3 3 3 3 1 3 3 3 ...  
## $ HispanicLatino : Factor w/ 4 levels "no","No","yes",..: 2 2 2 2 2 2 2 2 4 2 ...  
## $ RaceDesc : Factor w/ 6 levels "American Indian or Alaska Native",..: 3 3 6 6 6 2 3 6 6 1 ...  
## $ DateofHire : Factor w/ 99 levels "1/10/2011","1/20/2013",..: 20 9 94 32 52 92 84 86 35 86 ...  
## $ DateofTermination : Factor w/ 94 levels "","01/15/16",..: 1 1 1 13 1 41 81 1 1 1 ...  
## $ TermReason : Factor w/ 18 levels "","Another position",..: 12 12 12 1 12 4 2 12 12 12 ...  
## $ EmploymentStatus : Factor w/ 5 levels "Active","Future Start",..: 1 1 1 4 1 5 5 1 1 1 ...  
## $ Department : Factor w/ 6 levels "Admin Offices",..: 1 1 1 1 1 1 5 5 5 5 ...  
## $ ManagerName : Factor w/ 21 levels "Alex Sweetwater",..: 4 4 4 4 4 4 13 13 13 13 ...  
## $ ManagerID : int 1 1 1 1 1 1 17 17 17 17 ...  
## $ RecruitmentSource : Factor w/ 23 levels "Billboard","Careerbuilder",..: 4 22 9 17 22 4 20 1 22 17 ...  
## $ PerformanceScore : Factor w/ 4 levels "Exceeds","Fully Meets",..: 2 2 2 2 2 2 2 2 4 2 ...  
## $ EngagementSurvey : num 2.04 5 3.9 3.24 5 3.8 3.14 5 2.3 3.6 ...  
## $ EmpSatisfaction : int 2 4 5 3 3 4 5 5 1 5 ...  
## $ SpecialProjectsCount : int 6 4 5 4 5 4 0 0 0 0 ...  
## $ LastPerformanceReview\_Date: Factor w/ 43 levels "","1/10/2019",..: 5 7 8 1 5 1 1 11 18 20 ...  
## $ DaysLateLast30 : int 0 0 0 NA 0 NA NA 0 0 0 ...

## EmpID MarriedID MaritalStatusID GenderID   
## Min. :6.020e+08 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:1.101e+09 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :1.203e+09 Median :0.0000 Median :1.0000 Median :0.000   
## Mean :1.200e+09 Mean :0.3968 Mean :0.8097 Mean :0.429   
## 3rd Qu.:1.379e+09 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.988e+09 Max. :1.0000 Max. :4.0000 Max. :1.000   
##   
## EmpStatusID DeptID PerfScoreID FromDiversityJobFairID  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :0.00000   
## 1st Qu.:1.000 1st Qu.:5.000 1st Qu.:3.000 1st Qu.:0.00000   
## Median :1.000 Median :5.000 Median :3.000 Median :0.00000   
## Mean :2.397 Mean :4.606 Mean :2.984 Mean :0.09355   
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:3.000 3rd Qu.:0.00000   
## Max. :5.000 Max. :6.000 Max. :4.000 Max. :1.00000   
##   
## PayRate Termd PositionID   
## Min. :14.00 Min. :0.0000 Min. : 1.00   
## 1st Qu.:20.00 1st Qu.:0.0000 1st Qu.:18.00   
## Median :24.00 Median :0.0000 Median :19.00   
## Mean :31.28 Mean :0.3323 Mean :16.84   
## 3rd Qu.:45.31 3rd Qu.:1.0000 3rd Qu.:20.00   
## Max. :80.00 Max. :1.0000 Max. :30.00   
##   
## Position State Zip DOB   
## Production Technician I :136 MA :275 Min. : 1013 06/14/87: 2   
## Production Technician II: 57 CT : 6 1st Qu.: 1901 07/07/84: 2   
## Area Sales Manager : 27 TX : 3 Median : 2132 09/09/65: 2   
## Production Manager : 14 VT : 2 Mean : 6570 09/22/76: 2   
## Software Engineer : 9 AL : 1 3rd Qu.: 2357 01/02/51: 1   
## IT Support : 8 AZ : 1 Max. :98052 01/04/64: 1   
## (Other) : 59 (Other): 22 (Other) :300   
## Sex MaritalDesc CitizenDesc HispanicLatino  
## F :177 Divorced : 30 Eligible NonCitizen: 12 no : 1   
## M :133 Married :123 Non-Citizen : 4 No :281   
## Separated: 12 US Citizen :294 yes: 1   
## Single :137 Yes: 27   
## Widowed : 8   
##   
##   
## RaceDesc DateofHire DateofTermination  
## American Indian or Alaska Native: 4 1/10/2011: 14 :207   
## Asian : 34 3/30/2015: 12 08/19/13 : 2   
## Black or African American : 57 1/5/2015 : 11 09/24/12 : 2   
## Hispanic : 4 9/29/2014: 11 09/26/11 : 2   
## Two or more races : 18 5/16/2011: 10 2001/9/12: 2   
## White :193 7/5/2011 : 10 2004/1/13: 2   
## (Other) :242 (Other) : 93   
## TermReason EmploymentStatus  
## N/A - still employed :196 Active :182   
## Another position : 20 Future Start : 11   
## unhappy : 14 Leave of Absence : 14   
## more money : 11 Terminated for Cause : 15   
## N/A - Has not started yet: 11 Voluntarily Terminated: 88   
## career change : 9   
## (Other) : 49   
## Department ManagerName ManagerID   
## Admin Offices : 10 Elijiah Gray : 22 Min. : 1.00   
## Executive Office : 1 Kelley Spirea : 22 1st Qu.:10.00   
## IT/IS : 50 Kissy Sullivan: 22 Median :15.00   
## Production :208 Michael Albert: 22 Mean :14.58   
## Sales : 31 Amy Dunn : 21 3rd Qu.:19.00   
## Software Engineering: 10 Brannon Miller: 21 Max. :39.00   
## (Other) :180 NA's :8   
## RecruitmentSource PerformanceScore  
## Employee Referral : 31 Exceeds : 37   
## Diversity Job Fair : 29 Fully Meets :243   
## Search Engine - Google Bing Yahoo: 25 Needs Improvement: 18   
## Monster.com : 24 PIP : 12   
## Pay Per Click - Google : 21   
## Professional Society : 20   
## (Other) :160   
## EngagementSurvey EmpSatisfaction SpecialProjectsCount  
## Min. :1.030 Min. :1.00 Min. :0.00   
## 1st Qu.:2.083 1st Qu.:3.00 1st Qu.:0.00   
## Median :3.470 Median :4.00 Median :0.00   
## Mean :3.332 Mean :3.89 Mean :1.21   
## 3rd Qu.:4.520 3rd Qu.:5.00 3rd Qu.:0.00   
## Max. :5.000 Max. :5.00 Max. :8.00   
##   
## LastPerformanceReview\_Date DaysLateLast30  
## :103 Min. :0   
## 1/14/2019: 18 1st Qu.:0   
## 2/18/2019: 12 Median :0   
## 1/21/2019: 10 Mean :0   
## 1/28/2019: 9 3rd Qu.:0   
## 2/25/2019: 9 Max. :0   
## (Other) :149 NA's :103

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

[2] 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.

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

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

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

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

[7] 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.

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

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

1. 由于条件所限，本研究小组只有单台计算机的算力。在有分布式计算的环境下，可能不需要此步操作。 [↑](#footnote-ref-37)
2. 一般要求数据集含有至少预测变量5——10倍的样本 [↑](#footnote-ref-40)
3. 重抽样 50 次：10 折交叉验证重复 5 次 [↑](#footnote-ref-54)