现代统计软件论文

吴宇翀

高思琴

陈蔚

Table of Contents

[摘要 2](#_Toc42961390)

[背景 2](#_Toc42961391)

[数据集说明 2](#_Toc42961392)

[数据预处理 3](#_Toc42961393)

[描述分析 4](#_Toc42961394)

[年龄 4](#_Toc42961395)

[债务数量 4](#_Toc42961396)

[月收入 5](#_Toc42961397)

[Logit 回归 6](#_Toc42961398)

[拟合 6](#_Toc42961399)

[预测 8](#_Toc42961400)

[混淆矩阵与验证结果 8](#_Toc42961401)

[接受者操作特征（ROC）曲线 9](#_Toc42961402)

[模型选择 11](#_Toc42961403)

[抽样、训练与评价指标 11](#_Toc42961404)

[Logit 回归 11](#_Toc42961405)

[线性判别分析（LDA） 12](#_Toc42961406)

[偏最小二乘判别分析（PLSDA） 13](#_Toc42961407)

[SVM 16](#_Toc42961408)

[随机梯度助推法（GBM） 18](#_Toc42961409)

[模型间的比较 22](#_Toc42961410)

[总结 24](#_Toc42961411)

[参考文献 24](#_Toc42961412)

[附录 24](#_Toc42961413)

[数据 24](#_Toc42961414)

[模型间的比较 25](#_Toc42961415)

[模型间准确率和 Kappa 的比较 25](#_Toc42961416)

[Logit 回归结果 26](#_Toc42961417)

# 摘要

识别与预测信用卡是否将会逾期

待完善

# 背景

识别与预测信用卡是否将会逾期（信用卡风控部门）

陈蔚：待完善，数据集的背景什么的可以到英文网站上翻译[[1]](#footnote-1)

# 数据集说明

变量描述解释

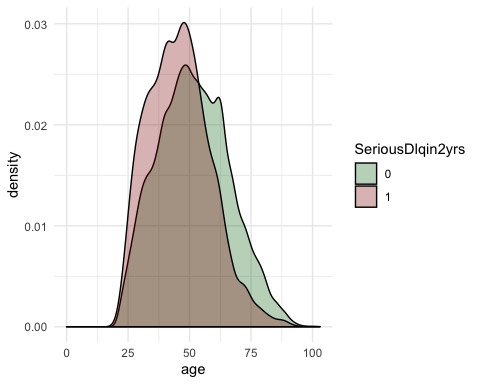
|  |  |  |
| --- | --- | --- |
| 变量名 | 描述 | 变量类型 |
| 是否逾期 | 是否有超过90天的逾期 | Y/N |
| 无担保放款的循环利用 | 无分期付款债务的信用卡和个人信用额度总额 | 百分比 |
| 年龄 | 借款人年龄 | 整数 |
| 过去2年间逾期30-59天的次数 | 有逾期30-59天，但在过去2年没有更糟的情况出现的次数 | 整数 |
| 负债比率 | 每月债务支付，赡养费，生活费用除以月总收入 | 百分比 |
| 月收入 | 每月的收入 | 实数 |
| 未偿还贷款数量 | 开放式贷款的数量和信用额度（如信用卡） | 整数 |
| 90天逾期次数 | 借款人逾期90天或以上的次数 | 整数 |
| 不动产贷款或额度数量 | 按揭及房地产贷款数目，包括房屋净值信贷额度。 | 整数 |
| 过去2年逾期60-89天的次数 | 借款人逾期60-89天的次数，但过去两年更糟的情况出现 | 整数 |
| 家属人数 | 不包括自己在内的家属（配偶，子女等）数量。 | 整数 |

# 数据预处理

1. 由于样本量已经足够大，我们删除所有包含缺失值的观测。
2. 由于**信用卡和个人信贷额度的总余额**和**负债比率**两个指标为百分比，我们将这两个指标中小于0的数据调整为0，将大于1的数据调整为1。

# 描述分析

## 年龄



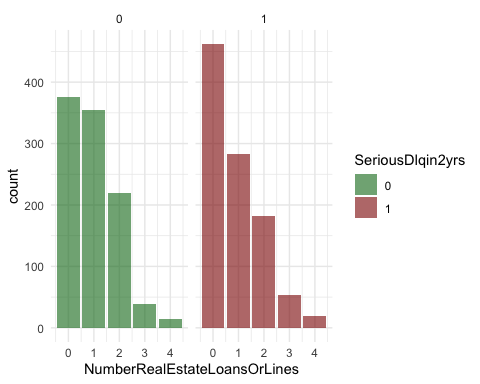
信用卡逾期与否两类人群的年龄分布（红色代表逾期）

从上图中我们可以看到，信用卡逾期与否的两类人群年龄上有着较为明显的差别。信用卡逾期者普遍年龄较小，这可能与信用卡使用者…….有关。

待完善（陈蔚）

## 债务数量

我们在信用好和差的持卡人中各抽取1000人，且由于数量多于5的持卡人非常少，为了方便画图，我们删去这些样本。

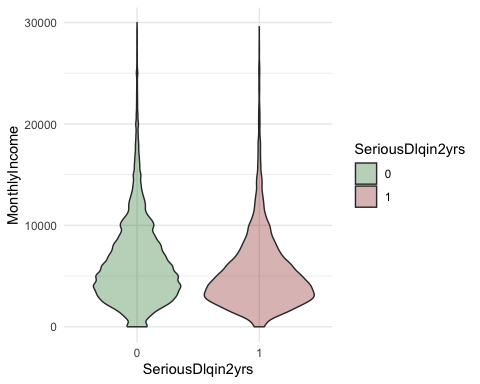


信用卡逾期与否两类人群的债务数量（红色代表逾期）

陈蔚：与上面的分析类似

## 月收入

且由于月收入高于30000的持卡人非常少，为了方便画图，我们删去这些样本。



信用卡逾期与否两类人群的月收入（红色代表逾期）

陈蔚：与上面的分析类似

# Logit 回归

## 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）。

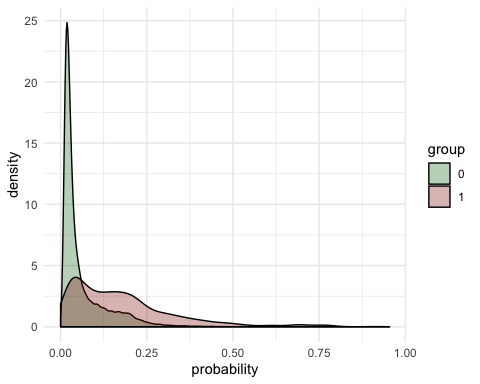
Logit回归系数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| （截距） | -3.56 | 0.07 | -51.33 | 0 |
| 无担保放款的循环利用 | 2.47 | 0.04 | 57.73 | 0 |
| 年龄 | -0.01 | 0.00 | -11.96 | 0 |
| 过去2年间逾期30-59天的次数 | 0.32 | 0.01 | 22.80 | 0 |
| 负债比率 | 0.25 | 0.06 | 3.96 | 0 |
| 月收入 | 0.00 | 0.00 | -7.31 | 0 |
| 未偿还贷款数量 | 0.03 | 0.00 | 8.73 | 0 |
| 90天逾期次数 | 0.28 | 0.02 | 15.66 | 0 |
| 不动产贷款或额度数量 | 0.06 | 0.01 | 4.18 | 0 |
| 过去2年逾期60-89天的次数 | -0.57 | 0.02 | -26.53 | 0 |
| 家属人数 | 0.07 | 0.01 | 6.45 | 0 |

可以看到，所有系数的 p 值在四舍五入后都为0，变量全部显著。

陈蔚：结合我们的数据集背景，分析自变量对因变量（是否逾期）的正负向作用。（Estimate 那一列为正，代表该变量的增加会引起逾期的可能性增大）

## 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

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 | 27910 |
| 1 | 0 | 2003 |
| 0 | 1 | 68 |
| 1 | 1 | 86 |

验证结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | Kappa | AccuracyLower | AccuracyUpper | AccuracyNull | AccuracyPValue | McnemarPValue |
| 0.931 | 0.068 | 0.928 | 0.934 | 0.995 | 1 | 0 |

可以看到：尽管准确率达到了 0.931, 但是还低于0.995的无信息率准确度（No Information Rate）。

灵敏度和特异度等指标表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sensitivity | Specificity | Pos Pred Value | Neg Pred Value | Precision |
| 指标值 | 0.558 | 0.933 | 0.041 | 0.998 | 0.041 |

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**

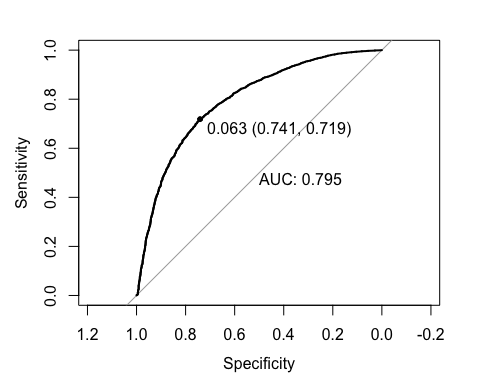
如果模型的准确度稳定在一个水平,通常会在灵敏度和特异度之间做一个权衡。直觉上，增加灵敏度会使特异度下降，因为更多的样本被预测为“发生”。当不同类型的错误对应惩罚不同时,在灵敏度和特异度间做出潜在权衡或许是合理的。在过滤垃圾邮件时我们通常关注特异度,如果家人和同事的邮件能不被删除,大多数人愿意接受看一些垃圾邮件。

陈蔚：这段话要进行改写，垃圾邮件要改为我们的数据案例，注意规避查重。

## 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] were designed as a general method that, given a collection of continuous data points, determine an effective threshold such that values above the threshold are indicative of a specific event. ROC curve can be used for determining alternate cutoffs for class probabilities.（陈蔚：待翻译）[[2]](#footnote-2)



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。[[3]](#footnote-3)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。

在上式中，O是观测的准确性，E是基于混淆矩阵边缘计数得到的期望准确性。该统计量取值在-1和1之间;0值表示观测类与预测类之间没有一致性，1值表示模型的预测与观测类完全一致。负值表示预测与事实相反，但在建立预测模型过程中绝对值大的负值很少出现。当各类分布相同时，总精确度与 Kappa是成比例的。取决于具体情况，Kappa值在0.30到0.50之间代表合理的一致性。（Agresti 2002）

陈蔚：这段话要进行语序修改，规避查重。

## Logit 回归

在重抽样下 Logit 模型的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| parameter | Accuracy | Kappa | AccuracySD | KappaSD |
| none | 0.931 | 0.202 | 0.014 | 0.173 |

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：

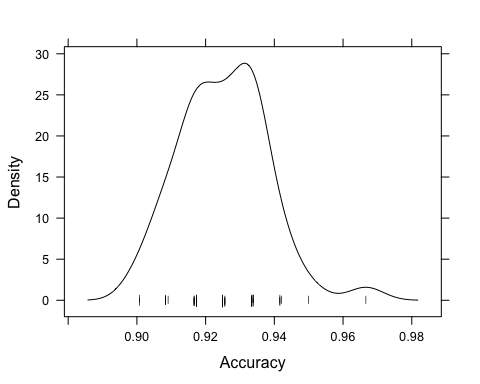
由于我们的分类只有两类，所以只有一个判别向量，不需要优化判别向量的数目，即不需要模型调优，计算速度较快。

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[[4]](#footnote-4)，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 | 0.925 | 0.122 | 0.013 | 0.161 |



在重抽样下 LDA 模型的准确率分布

## 偏最小二乘判别分析（PLSDA）

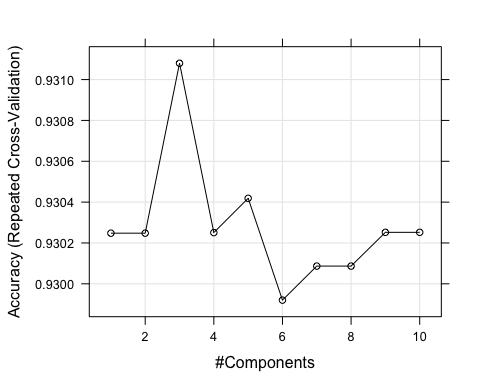
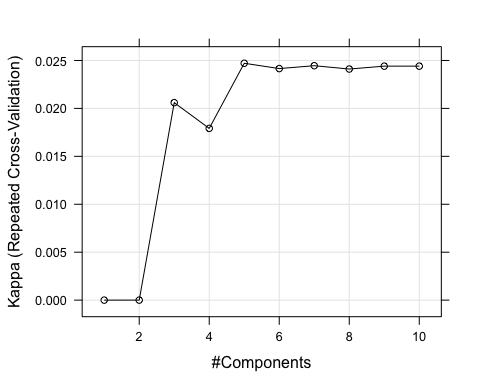
由于 LDA 不太适合多重共线性的变量，我们可以试着使用主成分分析压缩变量空间的维度，但 PCA 可能无法识别能将样本分类的较好变量组合，且由于没有涉及被解释变量的分类信息（无监督），很难通过 PCA 找到一个最优化的分类预测。

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。

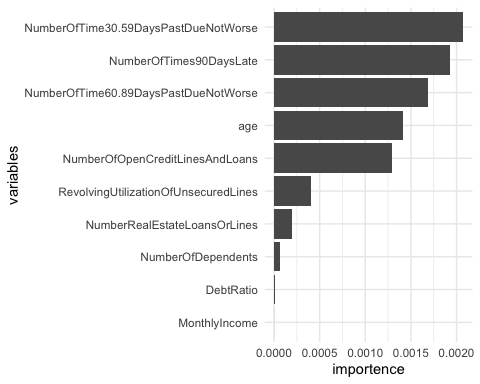
我们只使用前十个 PLS 成分

在重抽样下 PLSDA 模型的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| parameter | Accuracy | Kappa | AccuracySD | KappaSD |
| none | 0.925 | 0.122 | 0.013 | 0.161 |



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

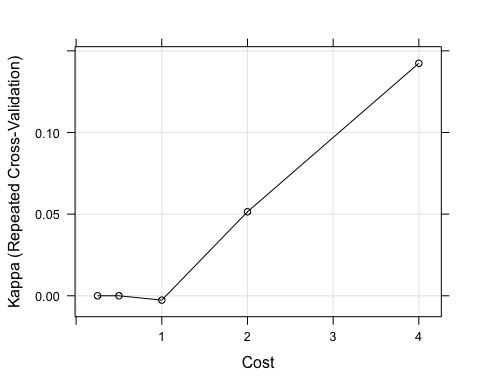
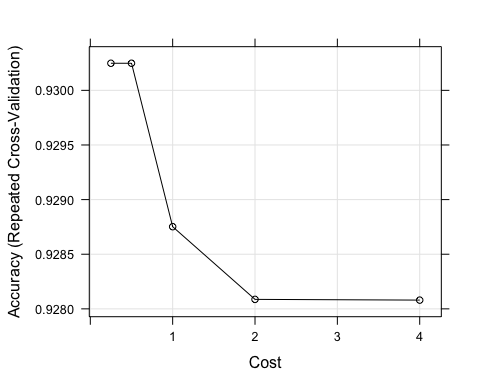
陈蔚：变量重要程度待分析

## 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 |
| 0.149 | 0.25 | 0.930 | 0.000 | 0.004 | 0.000 |
| 0.149 | 0.50 | 0.930 | 0.000 | 0.004 | 0.000 |
| 0.149 | 1.00 | 0.929 | -0.003 | 0.005 | 0.006 |
| 0.149 | 2.00 | 0.928 | 0.051 | 0.008 | 0.095 |
| 0.149 | 4.00 | 0.928 | 0.142 | 0.012 | 0.147 |



## 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。

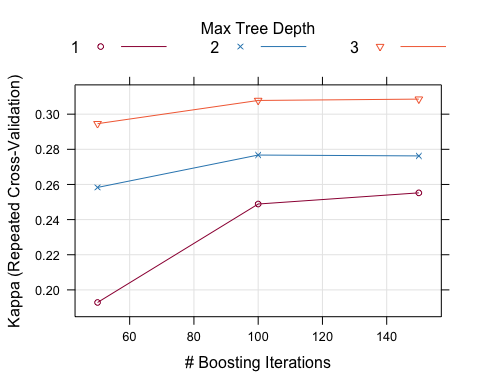
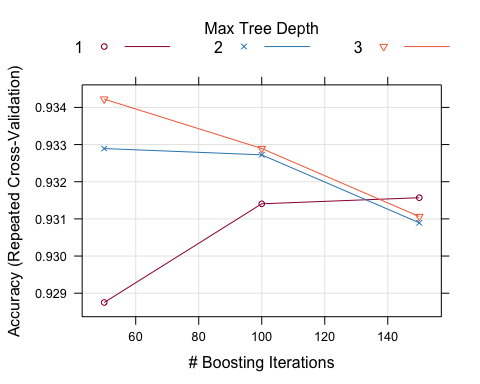
Friedman等（2000） [9] 发现分类问题可以当作是正向分布可加模型，通过最小化指数损失函数实现分类。

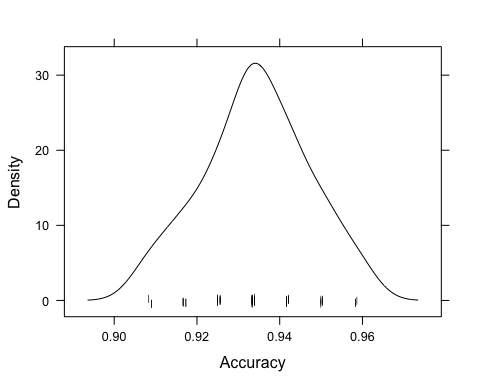
首先我们设定样本预测初始值为对数发生：

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

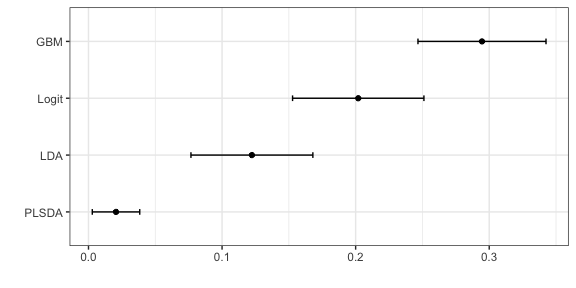




在重抽样下 GBM 模型的准确率分布

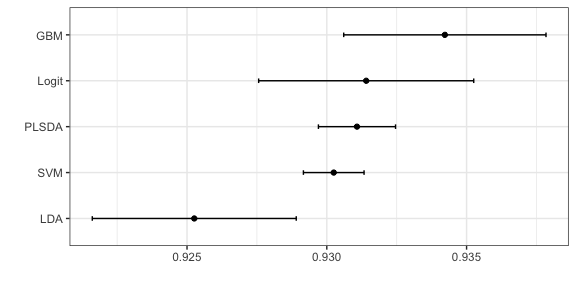
## 模型间的比较

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。[[5]](#footnote-5)



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。

# 总结

待完善

# 参考文献

[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.

# 附录

## 数据

## 'data.frame': 150000 obs. of 11 variables:  
## $ SeriousDlqin2yrs : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...  
## $ RevolvingUtilizationOfUnsecuredLines: num 0.766 0.957 0.658 0.234 0.907 ...  
## $ age : int 45 40 38 30 49 74 57 39 27 57 ...  
## $ NumberOfTime30.59DaysPastDueNotWorse: int 2 0 1 0 1 0 0 0 0 0 ...  
## $ DebtRatio : num 0.803 0.1219 0.0851 0.036 0.0249 ...  
## $ MonthlyIncome : int 9120 2600 3042 3300 63588 3500 NA 3500 NA 23684 ...  
## $ NumberOfOpenCreditLinesAndLoans : int 13 4 2 5 7 3 8 8 2 9 ...  
## $ NumberOfTimes90DaysLate : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ NumberRealEstateLoansOrLines : int 6 0 0 0 1 1 3 0 0 4 ...  
## $ NumberOfTime60.89DaysPastDueNotWorse: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ NumberOfDependents : int 2 1 0 0 0 1 0 0 NA 2 ...

## 模型间的比较

### 模型间准确率和 Kappa 的比较

模型间准确率的比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| LDA | 0.9008264 | 0.9168388 | 0.9250000 | 0.9252590 | 0.9333333 | 0.9666667 | 0 |
| PLSDA | 0.9256198 | 0.9256198 | 0.9333333 | 0.9310799 | 0.9333333 | 0.9416667 | 0 |
| SVM | 0.9256198 | 0.9256198 | 0.9333333 | 0.9302479 | 0.9333333 | 0.9333333 | 0 |
| GBM | 0.9083333 | 0.9256198 | 0.9333333 | 0.9342231 | 0.9416667 | 0.9586777 | 0 |
| Logit | 0.9008264 | 0.9250000 | 0.9333333 | 0.9314091 | 0.9416667 | 0.9750000 | 0 |

模型间准确率差异矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LDA | PLSDA | SVM | GBM | Logit |
| LDA |  | -0.0058209 | -0.0049890 | -0.0089642 | -0.0061501 |
| PLSDA | 0.027190 |  | 0.0008320 | -0.0031433 | -0.0003292 |
| SVM | 0.076479 | 0.237793 |  | -0.0039752 | -0.0011612 |
| GBM | 0.001116 | 0.929403 | 0.356758 |  | 0.0028140 |
| Logit | 0.003293 | 1.000000 | 1.000000 | 1.000000 |  |

模型间 Kappa 的比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| LDA | -0.0386266 | -0.0150376 | 0.1180075 | 0.1223394 | 0.1879195 | 0.7321429 | 0 |
| PLSDA | 0.0000000 | 0.0000000 | 0.0000000 | 0.0206005 | 0.0000000 | 0.2105263 | 0 |
| SVM | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 0 |
| GBM | -0.0377358 | 0.1805415 | 0.3023256 | 0.2945850 | 0.4221800 | 0.5969354 | 0 |
| Logit | -0.0377358 | 0.1046918 | 0.1830008 | 0.2019465 | 0.3222845 | 0.7567568 | 0 |

模型间Kappa差异矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LDA | PLSDA | SVM | GBM | Logit |
| LDA |  | -0.0058209 | -0.0049890 | -0.0089642 | -0.0061501 |
| PLSDA | 0.027190 |  | 0.0008320 | -0.0031433 | -0.0003292 |
| SVM | 0.076479 | 0.237793 |  | -0.0039752 | -0.0011612 |
| GBM | 0.001116 | 0.929403 | 0.356758 |  | 0.0028140 |
| Logit | 0.003293 | 1.000000 | 1.000000 | 1.000000 |  |

## Logit 回归结果

##   
## Call:  
## glm(formula = SeriousDlqin2yrs ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5488 -0.3724 -0.2387 -0.1852 4.5549   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## （截距） -3.559e+00 6.933e-02 -51.331 < 2e-16 \*\*\*  
## 无担保放款的循环利用 2.471e+00 4.280e-02 57.727 < 2e-16 \*\*\*  
## 年龄 -1.368e-02 1.143e-03 -11.962 < 2e-16 \*\*\*  
## 过去2年间逾期30-59天的次数 3.177e-01 1.393e-02 22.801 < 2e-16 \*\*\*  
## 负债比率 2.466e-01 6.234e-02 3.956 7.63e-05 \*\*\*  
## 月收入 -2.978e-05 4.071e-06 -7.314 2.59e-13 \*\*\*  
## 未偿还贷款数量 2.842e-02 3.255e-03 8.730 < 2e-16 \*\*\*  
## 90天逾期次数 2.818e-01 1.800e-02 15.660 < 2e-16 \*\*\*  
## 不动产贷款或额度数量 5.884e-02 1.407e-02 4.182 2.89e-05 \*\*\*  
## 过去2年逾期60-89天的次数 -5.721e-01 2.157e-02 -26.526 < 2e-16 \*\*\*  
## 家属人数 7.441e-02 1.154e-02 6.451 1.11e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 45518 on 90201 degrees of freedom  
## Residual deviance: 38190 on 90191 degrees of freedom  
## AIC: 38212  
##   
## Number of Fisher Scoring iterations: 6

1. 数据来源: <https://www.kaggle.com/c/GiveMeSomeCredit/overview> [↑](#footnote-ref-1)
2. ROC曲线是一个较为常用的方法，它给出了一系列连续数据点，便于确定一个有效的阈值，将超过某个阈值的值表示一个特定的事件。 [↑](#footnote-ref-2)
3. 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-3)
4. 一般要求数据集含有至少预测变量5——10倍的样本 [↑](#footnote-ref-4)
5. 重抽样 50 次：10 折交叉验证重复 5 次 [↑](#footnote-ref-5)