CENTRAL UNIVERSITY OF FINANCE AND ECONOMICS

中央财经大学

现代统计软件课程

信用卡逾期预测判别及模型比较

吴宇翀

高思琴

陈蔚

指导老师：杨玥含

2020 年 6 月 20 日

摘要

信用卡业务，是商业银行的核心业务；与此同时，信用卡的风险控制，一直以来都是信用卡业务最为密切关注的重要一环。信用卡逾期预测算法，直接为银行的信用卡风控业务提供支持。在此研究中，我们使用公开数据集，通过建立多种算法模型，计算预测的**信用卡逾期概率**。

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。

在模型选择上，**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。

在变量选择上，数据集中最重要的变量描述的是持卡人过去是否有信用卡逾期的先例，它们很好地反映出借贷者的信誉情况。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”，即这些变量并不直接反应持卡人的还贷习惯，影响较小。

关键词：**信用卡逾期**，**分类预测模型**，**机器学习**，**变量选择**，**模型比较**

目录

1 背景

银行在市场经济中扮演着至关重要的角色。他们决定谁可以得到资金，并需要什么样的条件，他们可以做出投资决定，当然也可以取消投资决定。在当今的社会条件下，为了使市场和生活正常运转，个人和公司常常需要获得信贷。作为新兴的消费工具，信用卡已经成为许多人的必备。然而信用卡一旦逾期，会给银行带来很大的风险。所以识别和预测信用卡是否将会逾期成为银行信用卡风控部门的重要工作。

信用评分算法是银行用来判断是否应该发放信用卡的一种重要解决方案，通过它可以对违约概率进行预测。银行利用持卡人的各种指标，通过预测持卡人遭遇财务困难的可能性，来提高信用评分的准确性。我们通过建立多个模型来帮助信用卡风控部门做出最佳的商业决策。

2 文献综述

在信用卡评分体系的分析方面，我国学者对信用卡风控的不同方面进行了的研究：李延东、郑小娟学者对信用卡评分体系的发展和应用做了详细的介绍，作者采用了分类分析法，对于各类信用卡的体系和发展进行了阐述。[1] 而在如何提高国内信用卡风险的可控性方面，国内众多学者提出了不同的思路：学者宋杰、王芳春提出在大数据基础上实现信用卡的自动审批，提高风险测评的自动化能力。学者李冰研究与各类商业银行信用卡所匹配的风险业务措施。[2] 学者叶纯青研究“互联网 + 信用评估”相结合的方式，尝试引入第三方征信机构，如支付宝的“芝麻信用”。[3]

在我们的研究中，我们吸收了几位学者优秀的研究成果，并在其基础上继续加以研究创新，通过对数据进行分析和处理，建立模型对信用卡逾期进行预测，选用多种机器学习模型进行模型比较，为信用卡风评部门提供更多的评判思路。

3 数据集说明

我们使用一个公开的数据集[1]，它有 11 个变量，150000 个观测。[2]

表 1: 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

4 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。

[1]数据来源: https://www.kaggle.com/c/GiveMeSomeCredit/overview

[2]模型的变量取值和分布见附录

5 描述分析

5.1 年龄

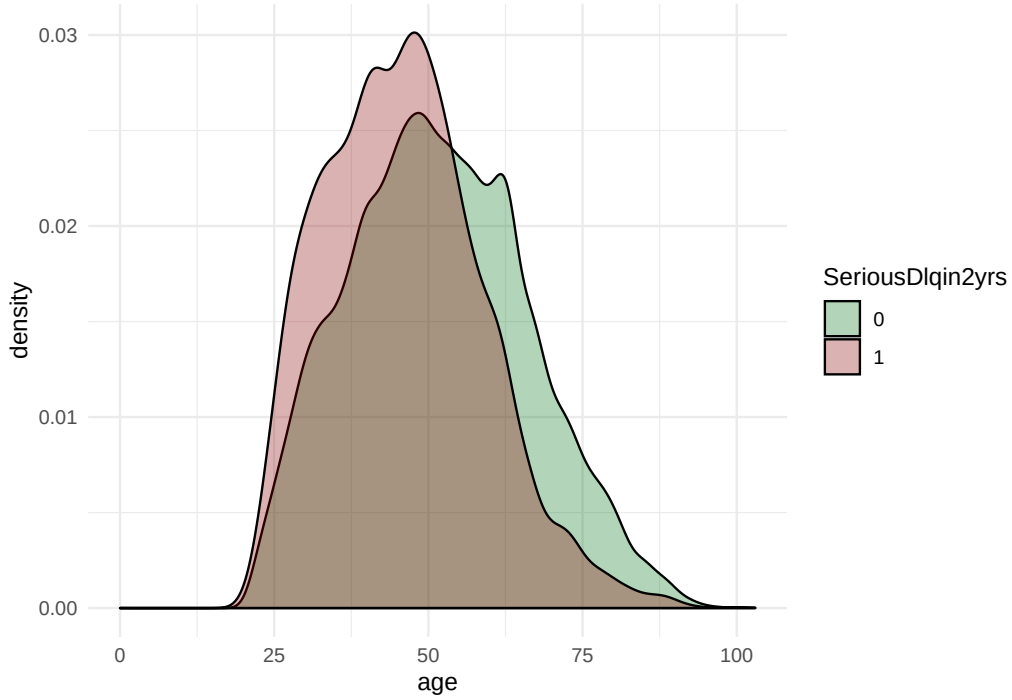

图 1: 信用卡逾期与否两类人群的年龄分布（红色代表逾期）

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的使用习惯有关。

相比年龄较大的人群，年轻人群当中奉行享乐主义者较多，一旦控制不当或者出现突发情况，就容易通过透支信用卡来填补空缺，而一旦信用卡数量过多而导致忘记还款，就会造成信用卡逾期的情况。除此之外，年龄较小人群工作稳定性不足，可能会有创业资金链破裂或者因为初入社会工资无法满足日常生活的问题，这也会导致无法及时还款。

5.2 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。

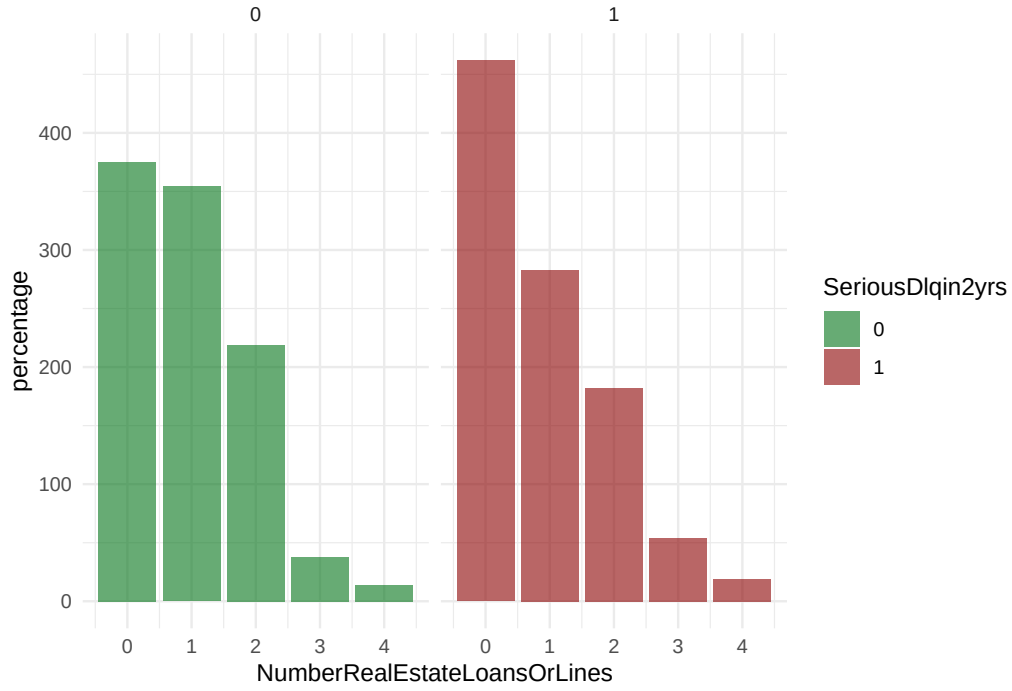

图 2: 信用卡逾期与否两类人群的债务数量（红色代表逾期）

通过统计图可知，在抵押贷款和房贷上，总体来看，大部分人的贷款数量都在 3 份以下。信用较好的持卡人相比信用较差的持卡人，没有贷款的比例更小、有 1-2 份贷款的比例更大，有更多贷款的比例更小。这表明，对于大多数持卡人而言，有 1-2 份贷款是合理的，这也是持卡人有一定财力进行还款的体现；然而，持卡人若同时背负过多份贷款，则会有很大的还款压力，信用卡还款逾期概率上升。

5.3 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。

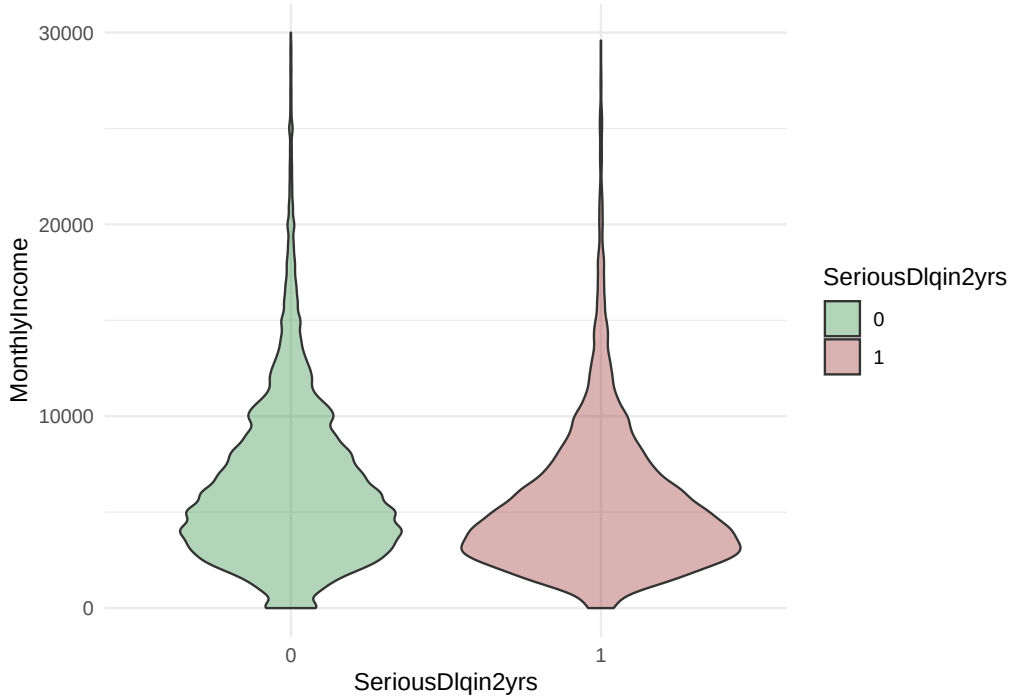

图 3: 信用卡逾期与否两类人群的月收入（红色代表逾期）

根据图可得，大部分信用卡适用人群的月收入在 10000 美元以内。图中信用卡逾期与否的两个图形对比后可得信用卡逾期的持卡人中，收入在 5000 元以下的人数比较多，这可能是由于收入较低的人群收入剩余较少，一旦不可控因素月开支增加，可能无法及时还款，收入 10000 元以上信用卡逾期人数少的原因正好与之相反。由月收入的密度分布可得，逾期的人群更加集中在 3000 美元的月收入左右；而按时还贷的人群工资范围则比较平均，集中性弱一些。其原因可能是国家政策规定工资不得低于一定数额，各地的最低工资标准虽然不同，但是低收入人群的工资范围是相似的，某一收入范围对于现在的生活开销来说比较吃力，每月开销需要控制才能不超于预算，该类人群信用卡逾期概率大。

6 Logit 回归

6.1 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）。

表 2: 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。自变量对因变量的正负向作用分析如下：

1. 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。
2. 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。
3. 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。
4. 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。
5. 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。
6. 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。

[3]模型详细见附录

7. 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。

6.2 预测

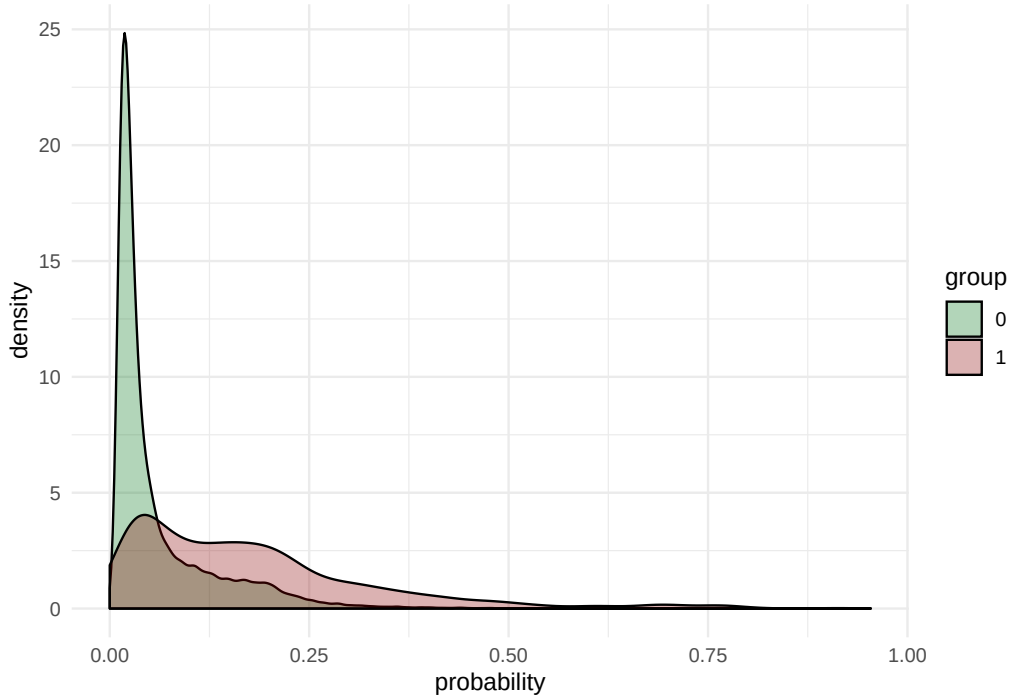

图 4: 预测的逾期概率值（红色代表已知为逾期）

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。

6.3 混淆矩阵与验证结果

灵敏度（Sensitivity）

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

特异度（Specificity）

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

假阳性率为 1 - 特异度

表 3: 混淆矩阵表

Prediction	Reference	Freq
0	0	27910
1	0	2003
0	1	68
1	1	86

表 4: 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）。

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	0.558	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**

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。

6.4 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) [4] [5] [6] 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。

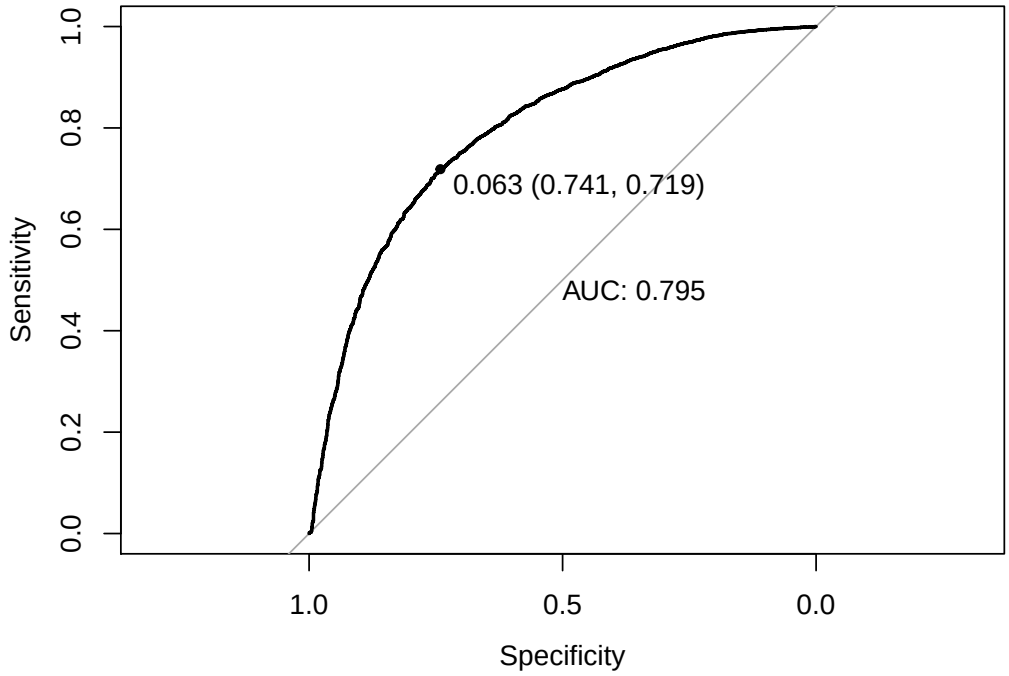

图 5: Logit 模型的 ROC 曲线

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。

7 模型选择

7.1 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。[4]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，使用 Kappa 和准确率作为模型的评价指标。

Kappa 统计量（Cohen 1960）[7] 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）

7.2 Logit 回归

表 6: 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。

7.3 线性判别分析（LDA）

Fisher（1936）[8] 和 Welch（1939）[9] 分析了获得最优判别准则的方式。

由贝叶斯法则：

[4]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。

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

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**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[5]，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。

表 7: 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

[5]一般要求数据集含有至少预测变量 5——10 倍的样本

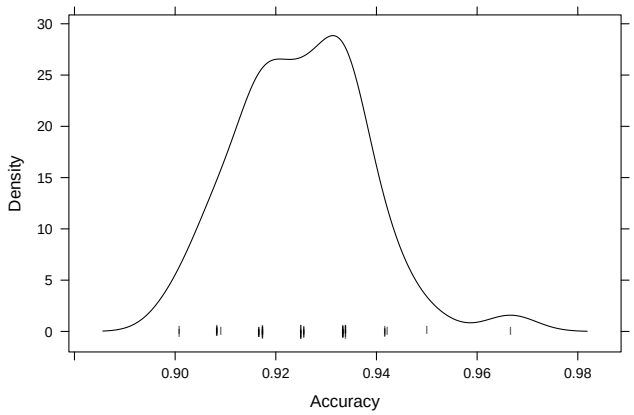

图 6: 在重抽样下 LDA 模型的准确率分布

7.4 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）[10] 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）[11] 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	0.930	0.000	0.004	0.000
2	0.930	0.000	0.004	0.000
3	0.931	0.021	0.005	0.062
4	0.930	0.018	0.006	0.056
5	0.930	0.025	0.007	0.084
6	0.930	0.024	0.008	0.086
7	0.930	0.024	0.008	0.086
8	0.930	0.024	0.007	0.085
9	0.930	0.024	0.007	0.085
10	0.930	0.024	0.007	0.085

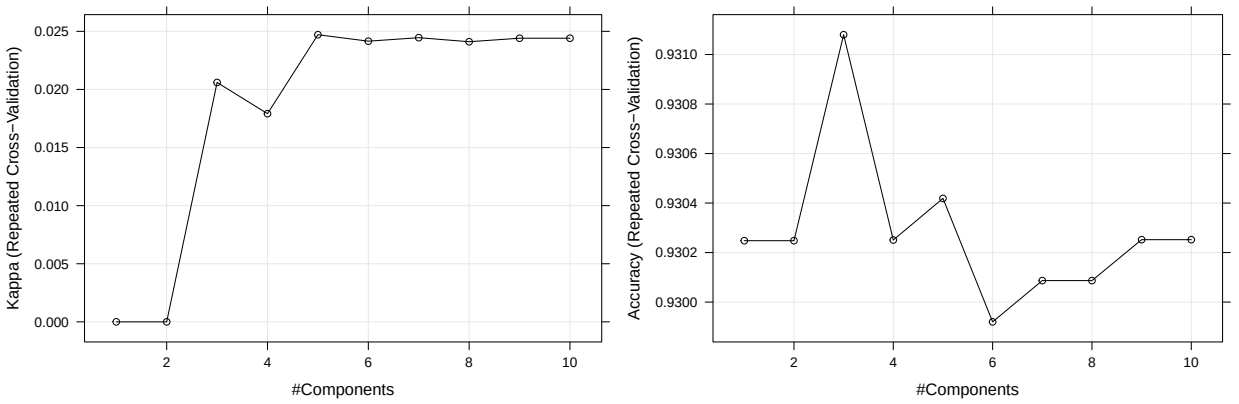

图 7: Kappa 和准确率指标随主成分个数的变化

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。

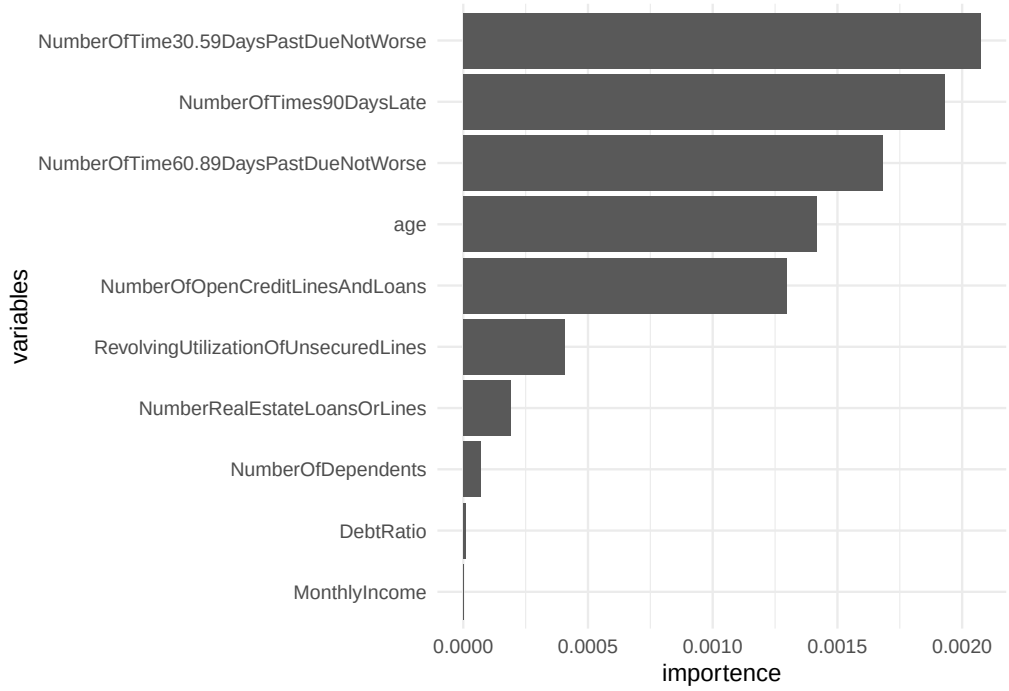

图 8: 变量重要程度

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。

7.5 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
0.149	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

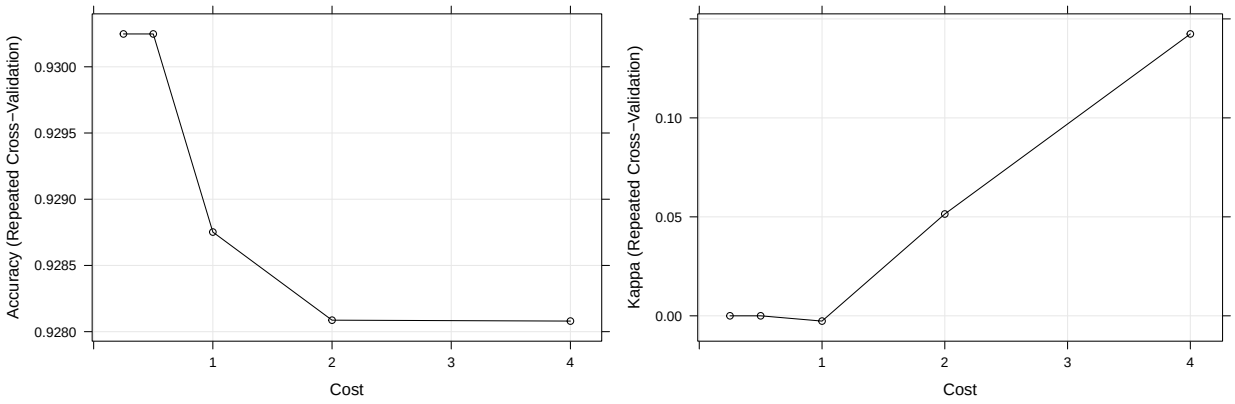

图 9: 调优参数不同取值下的准确率和 Kappa 指标变化

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。

7.6 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）[12] 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：

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

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

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 $f_1 = f_i + \lambda f_i^{(j)}$

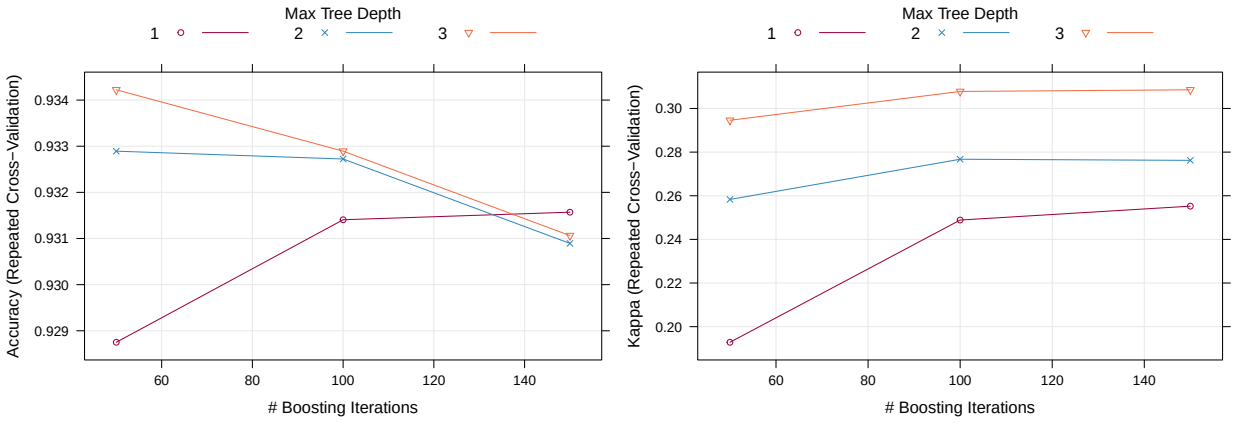

图 10: 调优参数和迭代次数不同取值下的准确率和 Kappa 指标变化

助推树的加深和迭代次数的增多一般引起 Kappa 指标的上升，引起的准确率变动并不大。

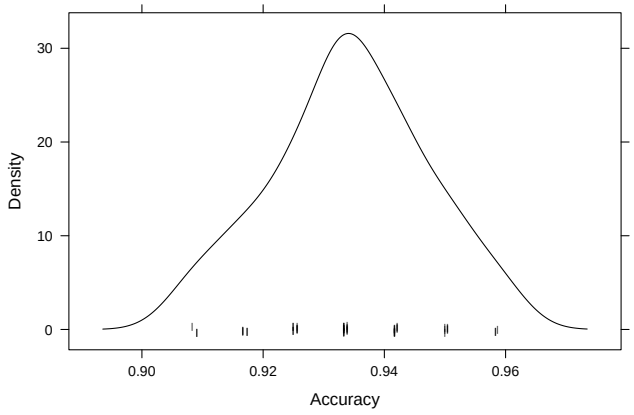

图 11: 在重抽样下 GBM 模型的准确率分布

7.7 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。[6] 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。[7]

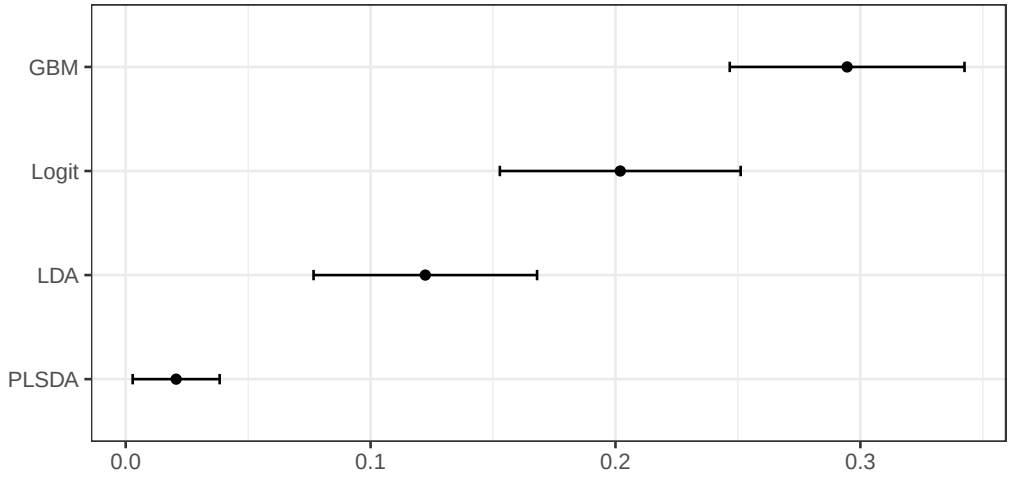

图 12: 模型间 Kappa 的比较（0.95 置信区间）

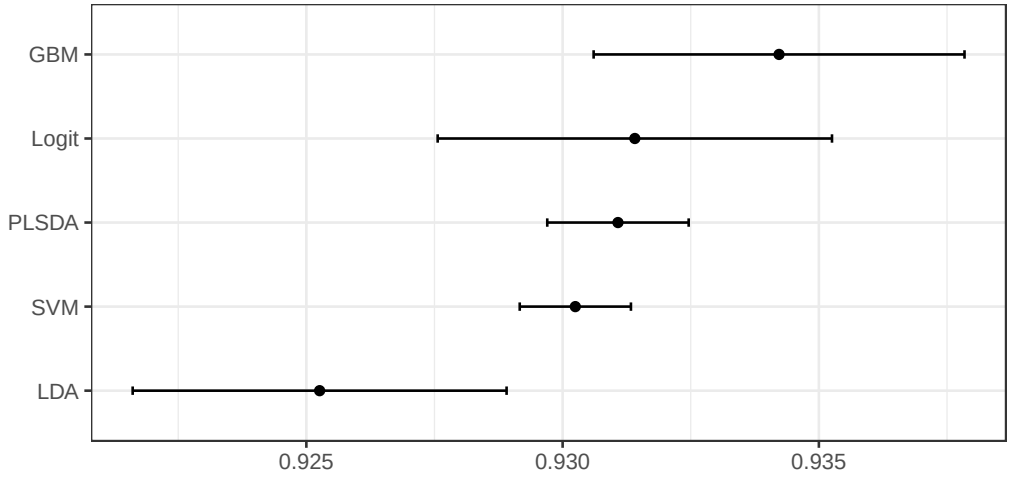

图 13: 模型间准确率的比较（0.95 置信区间）

[6] 具体模型比较数据见附录

[7] 重抽样 50 次：10 折交叉验证重复 5 次

8 总结

在此研究中，我们主要研究了商业银行信用卡逾期预测问题。

8.1 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。

8.2 模型选择

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。

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。

8.3 变量选择

根据 PLSDA 的结果，数据集中最重要的变量描述的是持卡人过去是否有信用卡逾期的先例，它们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，持卡人依旧可以通过合理规划及时还款。即这些变量并不直接反应持卡人的还贷习惯，影响较低。

9 参考文献

[1] 李延东, 郑小娟. 信用评分卡体系的发展及应用 [J]. 甘肃金融, 2016, 000(3): 53–55.

[2] 李冰. 我国商业银行信用卡风险管理分析 [J]. 商业故事, 2018(7).

[3] 叶纯青. 信用卡风控引入第三方征信机构 [J]. 金融科技时代, 2016(6): 85–85.

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

[5] 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.

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

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

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

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

[10] 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.

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

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

10 附录

10.1 模型间准确率和 Kappa 的比较

表 10: 模型间准确率的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.901	0.917	0.925	0.925	0.933	0.967	0

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
PLSDA	0.926	0.926	0.933	0.931	0.933	0.942	0
SVM	0.926	0.926	0.933	0.930	0.933	0.933	0
GBM	0.908	0.926	0.933	0.934	0.942	0.959	0
Logit	0.901	0.925	0.933	0.931	0.942	0.975	0

表 11: 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	

表 12: 模型间 Kappa 的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	-0.039	-0.015	0.118	0.122	0.188	0.732	0
PLSDA	0.000	0.000	0.000	0.021	0.000	0.211	0
SVM	0.000	0.000	0.000	0.000	0.000	0.000	0
GBM	-0.038	0.181	0.302	0.295	0.422	0.597	0
Logit	-0.038	0.105	0.183	0.202	0.322	0.757	0

表 13: 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	

10.2 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
```

10.3 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 ...
```

```
##  SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines      age
##  0:139974         Min.   :0.00000                     Min.   :  0.0
##  1: 10026         1st Qu.:0.02987                     1st Qu.: 41.0
##                   Median :0.15418                     Median : 52.0
##                   Mean   :0.31920                     Mean   : 52.3
##                   3rd Qu.:0.55905                     3rd Qu.: 63.0
##                   Max.   :1.00000                     Max.   :109.0
##
##  NumberOfTime30.59DaysPastDueNotWorse   DebtRatio      MonthlyIncome
##  Min.   : 0.000                       Min.   :0.0000   Min.   :      0
##  1st Qu.: 0.000                       1st Qu.:0.1751   1st Qu.:   3400
##  Median : 0.000                       Median :0.3665   Median :   5400
##  Mean   : 0.421                       Mean   :0.4663   Mean   :   6670
##  3rd Qu.: 0.000                       3rd Qu.:0.8683   3rd Qu.:   8249
##  Max.   :98.000                       Max.   :1.0000   Max.   :3008750
##                                                        NA's   :29731
##  NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
##  Min.   : 0.000                  Min.   : 0.000
##  1st Qu.: 5.000                  1st Qu.: 0.000
##  Median : 8.000                  Median : 0.000
##  Mean   : 8.453                  Mean   : 0.266
##  3rd Qu.:11.000                  3rd Qu.: 0.000
##  Max.   :58.000                  Max.   :98.000
##
##  NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
```

```
##  Min.   : 0.000              Min.   : 0.0000
##  1st Qu.: 0.000              1st Qu.: 0.0000
##  Median : 1.000              Median : 0.0000
##  Mean   : 1.018              Mean   : 0.2404
##  3rd Qu.: 2.000              3rd Qu.: 0.0000
##  Max.   :54.000              Max.   :98.0000
##
##  NumberOfDependents
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 0.757
##  3rd Qu.: 1.000
##  Max.   :20.000
##  NA's   :3924
```