****



**本 科 毕 业 论 文**

院 系 工程管理学院

专 业 计算机与金融工程

题 目 个人贷款贷中时期违约预测

年 级 2015 学 号 151278005

学生姓名 陈俊杰

第一指导教师 李心丹 职 称 教授

第二指导教师 丁晓平 职 称

提交日期

**南京大学本科生毕业论文（设计、作品）中文摘要**

题目：个人贷款贷中时期违约预测

院系：工程管理学院

专业：金融工程（计算机与金融工程方向）

本科生姓名：陈俊杰

指导教师（姓名、职称）：李心丹,教授；方立兵，讲师；丁晓平

摘要：

本文在以往有关个人贷款违约预测文献的基础上，结合商业银行的实际情况对个人贷款性质进行了分析，提出了利用贷中时期个人贷款的信息，对贷款进行违约预测的想法。对此，本论文假设，在贷中时期，个人贷款的信息会因为某些事件而发生改变。同时，对于变化的个人贷款信息而言，根据时间长度与风险之间的关系，若记录该个人贷款信息的时间点距离贷款到期日越近，则该款项信息对于贷款违约的预测力越强，用该信息来预测对应的个人贷款是否会违约的准确性越高；反之，则越低。最后，本文根据个人贷款信息记录的时间点与款项到期日的时间长度，对商业银行个人贷款数据集进行划分，并用利用划分之后的数据集进行实证分析，初步验证了以上假设。

关键词：商业银行；个人贷款；违约预测；机器学习

**南京大学本科生毕业论文（设计、作品）英文摘要**

THESIS：Prediction of Interim Default in Personal Loan

DEPARTMENT：School of Management and Engineering

SPECIALIZATION: Financial Engineering(FinTech)

UNDERGRADUATE: Chen Junjie

MENTOR: Li Xindan, Profession; Ding Xiaoping

ABSTRACT: Based on the previous literature on the prediction of individual loan default, this paper analyses the nature of individual loan in the light of the actual situation of commercial banks, and puts forward the idea of using the information of individual loan to predict the default of loan during the period of loan. In this regard, this paper assumes that during the loan period, personal loan information will change because of certain events. At the same time, for the changing personal loan information, according to the relationship between time length and risk, if the time point of recording the personal loan information is closer to the maturity date of the loan, the stronger the ability of the loan information to predict the default of the loan, the higher the accuracy of using this information to predict whether the corresponding personal loan will default; otherwise, the lower. Finally, according to the time point of personal loan information record and the time length of the maturity date, this paper divides the personal loan data set of commercial banks, and uses the data set after dividing to carry on the empirical analysis, preliminarily verifies the above hypothesis.

KEY WORDS: Commercial Bank; Personal Loan; Default Prediction; Machine Learning

目录

[第一章 绪论 4](#_Toc11270985)

[1.1. 引言 4](#_Toc11270986)

[1.1.1. 个贷增长趋势显著 4](#_Toc11270987)

[1.1.2. 个贷大幅增长带来的挑战 6](#_Toc11270988)

[1.1.3. 商业银行现行应对方案 6](#_Toc11270989)

[1.2. 文献综述 7](#_Toc11270990)

[1.3. 问题提出 10](#_Toc11270991)

[1.4. 论文结构安排 11](#_Toc11270992)

[第二章 过采样算法和模型评价指标 13](#_Toc11270993)

[2.1. 模型评价指标 13](#_Toc11270994)

[2.1.1. 混淆矩阵 13](#_Toc11270995)

[2.1.2. 精确率（precision）、召回率（recall）、F1值 13](#_Toc11270996)

[2.1.3. AUC和ROC 14](#_Toc11270997)

[第三章 数据选取、处理及描述 18](#_Toc11270998)

[3.1. 数据选取 18](#_Toc11270999)

[3.2. 原始数据预处理 24](#_Toc11271000)

[3.3. 数据描述性统计 29](#_Toc11271001)

[3.3.1. 离散值字段描述性统计 29](#_Toc11271002)

[3.3.2. 连续值字段描述性统计 31](#_Toc11271003)

[第四章 机器学习训练预测 34](#_Toc11271004)

[4.1. 数据集分组 34](#_Toc11271005)

[4.2. 机器学习模型训练预测 36](#_Toc11271006)

[4.2.1. 混淆矩阵 36](#_Toc11271007)

[第五章 总结与展望 40](#_Toc11271008)

[5.1. 总结 40](#_Toc11271009)

[5.2. 启示与展望 41](#_Toc11271010)

[参考文献： 42](#_Toc11271011)

[致谢 44](#_Toc11271012)

# 绪论

## 引言

### 个贷增长趋势显著

随着时代的发展，在国人收入水平节节攀升的同时，其消费水平也在不断上涨。对于新一代中国公民来说，基于对祖国进一步繁荣富强、对自身未来收入稳定的信赖，国人的储蓄比例逐年下降，消费比例逐年上升。更多的国人开始敢于消费，并敢于预支未来收入进行消费。可以说，国人的消费观念正从滞后消费逐渐转变为超前消费。所谓超前消费，就是在当前的收入水平不足以购买需要的产品或服务的情况下，消费者尝试向自然人或金融机构进行贷款，通过获取即期现金流、支付未来现金流的方式来满足自己当前对资金需求的消费行为。该消费行为使得整个国家的个人贷款总量迅速增长，给主营资产业务的商业银行带来了巨大的盈利机会。同时，因以资产业务为主要盈利模式，商业银行在过去几百年里不断优化自身的资产业务管理体系，在资产业务管理方面积累了丰富的经验，使其在一次次的金融浪潮中屹立不倒。因此，相比于自然人和其他金融机构而言，商业银行凭借其在资本业务管理方面成熟的管理模式，往往能从众多提供即期现金流的自然人和金融机构中脱颖而出，成为大部分贷款客户的合作对象。基于以上两点原因，商业银行资本业务规模越来越大，资本业务中的个人贷款总额以及个人贷款笔数也显现出逐年增长的趋势。如表1-1所示：

表1-1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 数额（亿） | 2013/12/31 | 2014/12/31 | 2015/12/31 |
| 招商银行 | 贷款余额 | 21970.94 | 25139.19 | 28242.86 |
| 个人贷款余额 | 8002.49 | 9713.27 | 12267.01 |
| 个贷占比 | 36.42% | 38.64% | 43.43% |
| 浦发银行 | 贷款余额 | 17674.94 | 20283.8 | 22455.18 |
| 个人贷款余额 | 3940.38 | 4647.82 | 5873.04 |
| 个贷占比 | 22.29% | 22.91% | 26.15% |
| 兴业银行 | 贷款余额 | 13570.57 | 15931.48 | 17794.08 |
| 个人贷款余额 | 3536.44 | 3859.5 | 5119.06 |
| 个贷占比 | 26.06% | 24.23% | 28.77% |
| 民生银行 | 贷款余额 | 15742.63 | 18126.66 | 20480.48 |
| 个人贷款余额 | 6055.29 | 6546.81 | 7280.28 |
| 个人贷款余额 | 38.46% | 36.12% | 35.55% |
| 平安银行 | 贷款余额 | 8472.89 | 10247.34 | 12161.38 |
| 个人贷款余额 | 3256.5 | 3849.95 | 4411.42 |
| 个贷占比 | 38.43% | 37.57% | 36.27% |
| 2016/12/31 | 2017/12/31 | 2018/12/31 | 5年累计增长 | 5年复合增长率 |
| 31616.81 | 35650.44 |  | 1.62 | 10.16% |
| 15405.94 | 17852.95 |  | 2.23 | 17.41% |
| 47.23% | 50.08% |  |  |  |
| 27628.06 | 31946 | 32448.22 | 1.81 | 12.57% |
| 9701.99 | 12321.31 | 12765.13 | 3.15 | 25.81% |
| 35.12% | 38.88% | 39.34% |  |  |
| 20798.14 | 24306.95 | 25500.2 | 1.79 | 12.36% |
| 7505.38 | 9108.24 | 9496.35 | 2.58 | 20.83% |
| 36.09% | 37.47% | 37.24% |  |  |
| 24615.86 | 28043.07 | 29065.09 | 1.78 | 12.24% |
| 9009.22 | 11058.27 | 11275.87 | 1.83 | 12.80% |
| 36.60% | 39.43% | 38.80% |  |  |
| 14758.01 | 17042.3 | 17725.28 | 2.01 | 15.00% |
| 5409.44 | 8490.35 | 9286.82 | 2.61 | 21.13% |
| 36.65% | 49.82% | 52.39% |  |  |

对于招商银行、浦发银行、兴业银行、民生银行和平安银行这五家股份制银行而言，其从2013年底到2018年3月份的贷款总额都呈现出大幅上升的趋势，其中增长幅度最小的是招商银行，5年复合增长率为10.16%；增长幅度最大的是平安银行，5年复合增长率为15%。在贷款总额大幅上升的情况下，相比于贷款总额，五家股份制银行的个人贷款总额上升趋势更为显著。在个贷方面，增长幅度最小的是民生银行，其5年复合增长率为12.80%；增长幅度最大的是浦发银行，其5年复合增长率为25.81%。由表1-1可以看出，近年来，个人贷款呈现出了逐年稳定大幅上升的趋势，是商业银行扩展并深化其资产业务的重要业务。

### 个贷大幅增长带来的挑战

然而，越来越多个人贷款涌入商业银行，不仅给商业银行带来了巨大的盈利机会，同时也对其资本业务风险管理提出了艰巨的挑战。

一方面，更多的个人贷款总额意味着需要投放更多的资金，导致商业银行自身的资本充足率下降。资本充足率下降带来的隐忧是，当突发情况发生时，商业银行可能没有足够的现金流加以应对，陷入流动性不足的困境，更甚则会导致银行倒闭。尽管在个人贷款总量增多时，商业银行能从中获取更多的预期利润，但是，可以预期，随之而来的是不良贷款笔数增多，更多贷款款项难以收回，潜在的资金损失更大。

另一方面，个人贷款具有贷款数额小、贷款笔数多的特性，同时，相比于企业贷款，个人贷款需要申请人提供的资料较少。企业贷款需要提供营业执照、税务登记证、组织机构代码证书、财务报表、公司章程、完税凭证和企业法人的个人凭证等大量资料，而个人贷款只需要提供个人资料即可。在商业银行要求将风险落实到每一笔贷款的情况下，个贷笔数多、借款人资料不全面的特性，给商业银行对其资产业务实施全面风险管理带来了巨大的挑战。

### 商业银行现行应对方案

为了应对个贷笔数增多导致不良贷款笔数相应增多，以及个贷借款人有效信息含量较低的问题，商业银行在个贷的贷前时期和贷中时期，分别进行风险防范和风险控制。

一方面，为了控制不良贷款发生，提高贷款款项的回收率，银行在贷前时期先对申请借款的客户进行信用风险评级，并以客户的信用风险等级为依据，来有效甄别优质客户和不良客户，向被识别为优质客户的贷款申请人发放贷款，而将被认定为不良客户的贷款申请人拒之门外。

在这一方面，商业银行的实际做法是从借款申请人的经营能力、盈利能力、偿债能力、发展能力，以及客户数值和信用状况等方面，对客户进行综合评价和等级信用的确定，给客户确定一个最高贷款额度，并在该最高贷款额度内，根据客户的实际需求进行贷款发放。

另一方面，为了及时发现借款人还款能力的变化，商业银行在贷后时期定期收集并更新借款人的个人信息，目的是能够及时识别借款人还款能力的变化。当借款人的还款能力出现显著下降时候，商业银行能够通过最新收集的数据来及时识别数据反映的借款人还款能力显著下降的信息，并采取相应的措施（如提前收回款项等），以减少该笔款项所造成的损失。

为此，商业银行以债务人的实际还款能力为衡量标准，对贷款进行了五级分类，该贷款分类法由早期的“一逾两呆”分类法演变而来，以解决“一逾两呆”分类法所存在的问题（如统一将未到期的贷款视为正常贷款，或者是把仅仅逾期一天的贷款视为不良贷款）。该贷款分类法分类标准如表1-1所示：

表1-1

|  |  |
| --- | --- |
| 类别 | 定义 |
| 正常 | 贷款损失概率为0 |
| 关注 | 贷款损失概率不超过5% |
| 次级 | 贷款损失概率在30%到50%之间 |
| 可疑 | 贷款损失概率在50到75%之间 |
| 损失 | 贷款损失概率在75%到100%之间 |

## 文献综述

在学术界，为了应对资产业务中贷款损失难以估计的问题，国内外专家学者不断对贷款的信用评估方法进行深入探索和研究，极大地推动了贷款信用评估方法的发展，为商业银行对贷款进行风险评估提供了坚实的理论基础。如今，信用评估方法大致可以分为三类，即定性分析、定量分析和利用机器学习算法进行分析。其中，定性分析是最早被专家学者用于信用评估的方法，该方法根据体现借款人的还款能力信息，依赖评估人的专业知识和评估经验对贷款进行评估。该方法虽然有评估人深厚的专业知识和丰富的评估经验作为支撑，但是在没有统一标准的情况下，该方法具有严重主观性的缺陷难以避免，该缺陷限制了定性分析在信用评估中的发展。之后，专家学者开始用定量分析方法对贷款信用评估进行分析研究。该方法利用统计学科的模型，如线性回归模型（LR）、线性判别分析模型（LDA）和Probit模型等，计算出每笔贷款的“信用值”，再根据贷款的“信用值”大小进行违约可能性分析。这种方法的优势是能够在一定程度上消除定性分析所具有的主观性过大、信用评估标准难以统一的问题，在理论上具有完全可行性。然而，在这类方法实际应用的过程中，该类方法的不足逐渐显现，原因是，几乎所有的定量分析方法，都对应用场景进行了严苛的限制，比如将评价指标限制为正态分布，而在实际情况中，评价指标往往不服从正态分布，导致定量分析方法无法应用于商业银行实际的信用评估工作中。随后，伴随着计算机领域人工智能的快速发展，大量机器学习算法成为专家学者的研究对象，被应用于商业银行的信用评估领域。与定性分析和传统定量分析相比，机器学习算法克服了定性分析主观性过大、判别标准难以统一的问题，而且不要求满足传统定量分析过于严苛的假设前提，同时，在实际研究中，专家学者们发现，基于机器学习进行信用评估的方法能大大提高贷款预测分类的准确率。因此，越来越多的机器学习算法被应用于商业银行的信用评估领域，关于如何将机器学习算法与贷款信用评估更有效结合的研究和探索工作也成为专家学者和银行从事人员关注的焦点。

由于上世纪美国银行资产业务的迅速发展，美国的商业银行急切需要一种全新的方法来预防和管理信用风险，于是，当时美国的专家学者和银行从业人员开始对信用评估展开深入研究。信用评估是商业银行根据借款客户的实际还款能力，对借款客户的信用情况进行评估的一种方法，具体的做法就是将所有贷款按照某一评价标准分成不同类别，比如按是否违约的标准，将每一笔贷款分成履约类和违约类，或者将贷款分成正常、关注、次级、可疑和损失五类等。在上世纪70年代，美国商业银行对于借款客户的信用评估主要依赖银行从业人员的专业知识和工作经验进行主观性判断。这一判断模式根据借款人的品行（Character）、资产规模（Capital）、收入水平（Capacity）、抵押（Collateral）以及条件（Condition），来评估借款客户的信用品质，因此又被称为5C系统。然而，该方法得出的信用评估分值往往基于银行从业人员的主观判断，并且在当时借款客户迅速增多的背景下，需要大量的人力对借款客户进行信用评估，从而导致该方法难以具有统一信用评估标准的弊端尽显。在这样的背景下，专家学者和商业银行工作人员开始转向定量分析。在1968年，纽约大学教授爱德华·阿特曼(Edward Altman)仔细对比了美国破产企业和非破产企业的财务数据，并通过数理统计方法，从22个传统的财务比率指标中筛选出5个影响企业未来是否破产的重要因素，构建了5变量Z-score模型。在1977年，阿特曼（Altman）、赫尔德门（Haldeman）和纳内亚南（Narayanan）将原始Z-score模型中的五个变量增加到了七个，并证明了第二代模型能大大提升对不良贷款人的识别精度，第二代模型称为ZETA信用风险模型（ZETA Credit Risk Model）[1]。除了ZETA模型之外，在同一时期较受欢迎的模型还有Logit模型，该模型假设所有贷款的违约概率服从特定的分布，并以财务数据为自变量，将预测得到的因变量以某种转换公式，转换为0到1之间的数字，并把这个数据作为对借款客户的违约概率的预测。在1977年，Martin利用多因子Logit模型准确预测了在1975年至1976年间，美国共23家银行的倒闭结果[2]。在1983年，Madalla通过实证研究调整了利用Logit模型得到的违约概率与该笔贷款履约或违约的关系，该学者通过大量的研究，表明当违约概率值P不小于0.551时，该笔贷款属于不良贷款；而当违约概率值小于0.551时，该笔贷款可视为正常贷款[3]。在1985年，West选取了大量财务比率因子，利用因子分析和Logit模型估计银行陷入财务困境的概率，实证结果表明，因子分析和Logit估计相结合是一种可靠的银行状况评价方法[4]。在2000年，Laitinen检验泰勒级数展开法是否可以用来解决与破产预测模型的函数形式相关的问题，他首先用Logit模型来描述破产风险模型，再用泰勒展开式来拟合Logit函数。通过实证表明，泰勒展开模型在破产预测中具有有效性，并说明了现金总资产比、现金流量总资产比和股东权益总资产比这三个指标是预测破产与否的重要因素[5]。上个世纪末，由于计算机领域机器学习算法的兴起，机器学习算法在信用评估领域方面的应用也受到了大量专家学者的青睐。在1996年，Henley和Hand将k最近邻模型（KNN）用于对消费贷款申请人的信用度进行评估。他们对KNN模型中的超参数k（最近邻个数）以及d（距离计算方式）进行了深入的讨论，并将KNN模型与传统线性模型、Logit模型和决策树模型进行了比较分析[6]。在2000年，West研究了物种神经网络模型，分别为多层感知器模型、混合专家模型，径向基函数模型、学习矢量量化模型和模糊自适应共振模型，分别用这五种模型对真实数据集进行训练测试，并将结果与传统方法中的线性判别分析、逻辑回归、K近邻、核密度估计以及决策树进行比较。实证结果表明，在信用评分应用中应同时考虑混合专家神经网络模型和径向基函数神经模型，而在传统模型中，逻辑回归是最适用的模型[7]。

我国从上世纪80年代末期，开始对贷款过程中存在的信用风险进行预防、控制和管理，由于我国研究资产业务中产生的信用风险的时间较短，目前我国在这方面的研究，主要集中于研究国外的信用评估模型是否适用于中国经济市场。在2002年，鲁炜、赵恒珩和刘冀云将KMV模型直接应用于评估中国上市公司的信用风险，初步证实了将KMV模型应用于中国上市公司的理论意义和现实意义[8]。在2004年，肖艳结合了中国上市公司的传统财务指标和现金流指标，利用逻辑回归模型构建了一个上市公司财务困境预警模型，实证研究表明逻辑回归模型能够较为准确地对上市公司的财务状况进行预测[9]。在2005年，针对我国上市公司在财务数据具有高维度性和高相关性的特点，梁琪在传统的逻辑回归分析中引入了主成分分析，对我国沪深两市上市公司的财务数据进行研究，实证表明，结合主成分分析的逻辑回归模型比简单的逻辑回归分析的预测准确率更高[10]。在2006年，马若微结合粗糙集理论（RS）和人工神经网络（ANN）对中国上市公司是否会陷入财务困境进行了预测，实证结果表明两者结合的模型具有较强的预测能力[11]。同年，肖文兵和费奇基于支持向量机（SVM）模型，提出了一种新的个人信用评估预测方法，并将支持向量机方法与线性判别分析、逻辑回归分析、k近邻模型、分类回归树模型进行了比较，实证研究表明，SVM具有很好的预测效果[12]。在2009年，彭建刚和吕志华针对原Credit Risk+模型中，行业风险因子之间相互独立的不合理假设，提出了基于行业特性的多元系统风险因子Credit Risk+模型[13]。

## 问题提出

以上所述，几乎绝大多数文献，在运用定性分析、定量分析或者是机器学习算法对个人贷款进行到期违约预测时，其训练和预测数据所采用的特征值都是借款客户在申请贷款时向商业银行上报的数据，即贷前数据，而需要进行预测的结果则是贷款客户能否如期足额偿还本息，即贷后结果。可以预见的是，在贷款发放日到还款日之间的贷中时期，能体现借款客户还款能力的特征信息会因某些事件而发生改变。以房产为抵押物的个人贷款为例，在贷中时期，该房产抵押物的价值会随房地产行情的好坏而发生变动，当房地产行情高涨时，房产抵押物价值上升，借款客户的还款能力也随之提升，借款客户到期足额还本付息的可能性更大；而当房地产行情萧条时，房产抵押物价值下降，借款客户的还款能力下跌，其到期无法偿还本息的可能性增大。以上例子，说明无论是用何种方法对个人贷款进行到期违约预测，都不应该只关注借款客户的贷前信息，而需要把注意力更多地放在贷后信息上，因为从理论上来说，在贷前时间点与贷款到期时间点之间有一段“不确定”的时间段，在这段“不确定”的时间段中，可能会发生使得借款客户还款能力变化的事件。根据金融风险原理，该段“不确定”的时间段越长，发生影响借款客户还款能力的事件的可能性越大。比如，当一笔贷款要求在发放时间点24小时后还本付息时，借款客户在贷前和贷款到期时间点的还款能力基本不会发生显著变化；而当一笔贷款的持续时间段为十年时，借款客户的还款能力很可能在贷款持续期内发生翻天覆地的变化。因此，根据贷款持续时间段长度与风险之间的关系，本文提出以下假设，当体现借款客户还款能力的信息所在的时间点距离还款日越远时，该信息对借款客户能否按时足额还本付息的解释力越弱；反之，当体现借款客户还款能力的信息所在的时间点距离还款日越近时，该体现还款能力的信息对借款客户能否按时足额还本付息的解释力越强。

为了验证该假设，本文舍弃了用贷前面板数据来预测借款客户能否如期偿还本息的传统模式，而是提出了一种新的预测借款客户能否按时还本付息的模式。该模式尝试跟踪每一笔贷款在贷中时期不同时间点，能体现借款客户实际还款能力的信息特征，根据距离贷款到期日的时间段长度，将数据集划分为互不相交的集合，通过多种机器学习算法进行预测分析，分别对不同集合中的数据集进行训练预测，并对不同集合中的预测结果进行对比。

## 论文结构安排

在第二章中，本文首先对预处理数据需要用到的标准化方法进行简要介绍，接下来介绍评价模型常用的方法，包括混淆矩阵、精确率、召回率、F1值和AUC（Area Under Curve）。

第三章以介绍数据处理的方式，详细地阐述了本文以数据驱动为基础的模型。在该章中，本文首先对原始数据的特征名称及意义进行简要介绍，然后详细介绍对原始数据进行预处理的过程，最后对处理完毕后的数据进行描述性统计。

在第四章中，对第三章中得到的原始数据集进行划分，得到的互不相交的数据集合，分别用决策树、随机森林、Logit模型和Catboost机器学习算法进行训练预测，计算不同数据集合的混淆矩阵、精确率、召回率、F1值和AUC结果，最后对模型评价指标进行详细的分析和对比。

第五章根据第四章中的对比分析结果，对全文进行总结。该章首先再次强调本文可能的贡献，接下来指出本文在数据选取和模型构建上所存在的不足，最后在本文工作的基础上，提出可能继续深入研究的方向。

# 过采样算法和模型评价指标

## 模型评价指标

本论文采用的模型评价指标有：混淆矩阵、精确率（precision）、召回率（recall）、F1值（F1-Score）和AUC（Area under Curve）。

### 混淆矩阵

对于模型的二分类问题（此处的二分类为违约、履约，或者是negative、positive），如果我们希望检验模型的预测结果好坏，需要知道以下四种情况的数量：

（1）数据真实标签是履约（Positive），而模型预测结果是履约（Positive）的数据量，即真正例（True Positive）数据量；

（2）数据真实标签是履约（Positive），而模型预测结果是违约（Negative）的数据量，即假反例（False Negative）数据量；

（3）数据真实标签是违约（Negative），而模型预测结果是履约（Positive）的数据量，即假正例（False Positive）数据量；

（4）数据真实标签是违约（Negative），而模型预测结果是违约（Negative）的数据量，即真反例（True Negative）数据量

由此得到的混淆矩阵如表2-2：

表2-2

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 真实值 | |
| 履约（Positive） | 违约（Negative） |
| 预测值 | 履约（Positive） | 真正例（TP） | 假反例（FP） |
| 违约（Negative） | 假正例（FN） | 真反例（TN） |

### 精确率（precision）、召回率（recall）、F1值

对于混淆矩阵来说，其显示的是真正例（TP）、假反例（FP）、假正例（FN）和真反例（TN）的个人，然而，在面对大量的数据时，我们更关心的是各个统计量的占比，而非绝对个数。因此，在混淆矩阵统计量的基础上，增加了精确率（Precision）、召回率（Recall）、和F1值（F1-Score），这三个统计量的定义和统计意义如下：

（1）精确率（Precision）。定义：在模型预测的所有履约的样本中，真正履约的样本的占比。公式：，表示为真正例（TP）占真正例（TP）与假反例（FP）之和的比例。该统计量的统计意义是：在模型判断为履约的样本中，真正履约的样本占比有多少，即模型是否预测得对。

（2）召回率（Recall）。定义：在所有真正履约的样本中，模型预测履约的样本的占比。公式：，表示为真正例（TP）占真正例（TP）与假正例（FN）之和的比例。该统计量的统计意义是：在真正履约的样本中，模型预测为履约的样本占比有多少，即模型是否预测得全。

（3）在实际情况中，精确率（Precision）和召回率（Recall）这两个统计量会相互“制约”，即若某一模型的准确率高，则其召回率就低；若召回率高，则其准确率就低。因此，在准确率（Precision）和召回率（Recall）的基础上，引入了综合考虑准确率和召回率的统计量，称为F1统计量（F1-Score）。该统计量的公式为：。

### AUC和ROC

在实际预测过程中，很多机器学习模型是为每一个测试样本产生一个实值，或预测该测试样本为正例的概率值，然后将该实值或概率值与一个二分类阈值进行比较，若大于分类阈值，则预测为正类，否则预测为反类。因此，根据机器学习模型的特性，我们可以将学习器对每一个测试样本的预测概率进行排序，将“最可能”是正类的测试样本排在最前面，而将“最不可能”是正类的测试样本排在最后面。因此，分类过程就可以通过在所有排好序的正类概率中选择某一个正类概率作为一个“截断点”（cut point）来进行，该“截断点”将测试样本分为两部分，预测概率不小于该“截断点”的测试样本判作正类，小于则判作反类。现在，我们得到了数据的真实类别，也得到了学习器预测得到的测试样本类别，可以根据以下公式分别计算出“真正例率”（True Positive Rate，简称TPR）和“假正例率”（False Positive Rate，简称FPR）：

假设测试样本数量为N，同时假设学习器对于每个测试样本的正类预测概率都不相同。因此，将这N个预测概率从小到大进行排列后。将这N个预测概率分别作为一个“截断点”，可以得到N组 TPR和FPR值。以FPR为横轴，TPR为纵轴，将N组TPR和FPR值绘制成ROC曲线，即“受试者工作特征”（Receiver Operating Characteristic）曲线，该曲线下的面积即为AUC（Area under Curve）值。AUC值介于0和1之间，其意义为：在测试样本中随机选择一个正样本和一个负样本，该学习器将真实正样本预测为正样本的概率为，将真实负样本预测为正样本的概率为时，的概率。可以预见，在一个“随机预测”的模型中，AUC值为0.5。因此，若该学习器的AUC值小于0.5，说明该学习器的性能甚至不如“随机预测”的模型。同时，对于一个学习器的性能而言，AUC越接近于1，说明其预测能力越强。

下面以一个简单的例子详细介绍AUC如何计算，假设测试样本中含有四个样例，其真实类别与学习器的预测概率如表2-3：

表2-3

|  |  |  |
| --- | --- | --- |
| 样本序号 | 真实类别 | 学习器预测为P概率 |
| 1 | N | 0.1 |
| 2 | N | 0.5 |
| 3 | P | 0.4 |
| 4 | P | 0.8 |

（1）“截断点”为0.1时，所有概率不小于0.1的样本都预测为正例，此时所有的样本都预测为正例，对应的混淆矩阵如表2-4：

表2-4

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 真实值 | |
| P | N |
| 预测值 | P | TP = 2 | FP = 2 |
| N | FN = 0 | TN = 0 |

此时，TPR = TP / (TP + FN) = 2 / 2 = 1，FPR = FP / (TN + FP) = 2 / 2 = 1。

（2）“截断点”为0.5时，所有概率不小于0.5的样本都预测为正例，此时1号样本预测为反例，2号样本预测为正例，3号样本预测为反例，4号样本预测为正例。因此对应的混淆矩阵如表2-5：

表2-5

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 真实值 | |
| P | N |
| 预测值 | P | TP = 1 | FP = 1 |
| N | FN = 1 | TN = 1 |

此时，TPR = TP / (TP + FN) = 1 / 2 = 0.5，FPR = FP / (TN + FP) = 1 / 2 = 0.5。

（3）“截断点”为0.4时，所有概率不小于0.4的样本都预测为正例，此时1号样本预测为反例，2号样本预测为正例，3号样本预测为正例，4号样本预测为正例。因此对应的混淆矩阵如表2-6：

表2-6

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 真实值 | |
| P | N |
| 预测值 | P | TP = 2 | FP = 1 |
| N | FN = 0 | TN = 1 |

此时，TPR = TP / (TP + FN) = 2 / 2 = 1，FPR = FP / (TN + FP) = 1 / 2 = 0.5。

（4）“截断点”为0.8时，所有概率不小于0.8的样本都预测为正例，此时1号样本预测为反例，2号样本预测为反例，3号样本预测为反例，4号样本预测为正例。因此对应的混淆矩阵如表2-7：

表2-7

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 真实值 | |
| P | N |
| 预测值 | P | TP = 1 | FP = 0 |
| N | FN = 1 | TN = 2 |

此时，TPR = TP / (TP + FN) = 1 / 2 = 0.5，FPR = FP / (TN + FP) = 0 / 2 = 0。

根据（FPR，TPR）对得到的ROC曲线如图2-4：

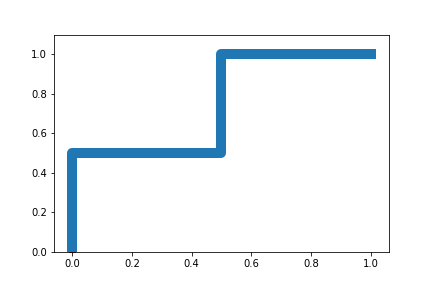


图2-4

该ROC曲线下的AUC值为0.75，说明在测试样本中随机选择一个正样本和一个负样本，该学习器将真实正样本预测为正样本的概率为，将真实负样本预测为正样本的概率为时，的概率为0.75。

# 数据选取、处理及描述

## 数据选取

本文需要根据个人贷款的客户信息，识别出能体现借款客户还款能力的信息，进而根据这方面的信息预测个人贷款到期时借款客户还本付息的可能性。因此，若要对个人贷款进行违履约预测，首先需要的是借款客户的基本信息，该基本信息应包括能体现借款客户绝对偿还能力的信息，如收入、支出、收入稳定性等信息。此外，借款客户的基本信息还应包括能体现债务人偿还意愿的信息，如已偿还贷款占总收入的比率等，但有关能体现债务人偿还意愿的信息不仅在理论上没有精确定义的指标，在实际操作过程中也难以收集。同时，有研究表明债务人的偿还意愿与偿还能力呈现出明显的正相关关系，即当债务人偿还能力高时，其偿还意愿也相对较高；反之，其偿还意愿也相对较低。根据以上理论，将能体现借款客户绝对偿还能力的信息，作为债务人的偿还能力和偿还意愿这两方面的替代，是合理的。因此，本文选取了能体现客户绝对还款能力的信息作为客户的基本信息。同时，可以意识到，只根据客户的基本信息无法准确预测出个人贷款的违履约情况，因为有关借款客户的很多基本信息涉及到个人隐私问题，如客户的过去经历和详细消费情况等。有关这方面的信息，若借款客户不愿意提供，商业银行也就无法得到。在这种情况下，商业银行只能得到有关客户年龄、性别、学历等非常笼统、不具有能展现客户独特个性的数据，用这样的基本信息数据进行贷款违履约预测，得到不尽如人意的预测效果也不足为奇。因此，对于商业银行来说，在手握客户基本信息数据之外，还需要更多维度的数据来描述借款客户，以让借款客户的形象更为具体，同时也让预测过程变得更为简单。在这种情况下，综合真实性和可获得性两方面考虑，不妨用借款客户与商业银行交互的信息，如借款客户在商业银行的信用卡持有数量和借款客户在商业银行的交易金额等信息，作为借款客户基本信息的补充，来进一步描述客户的偿还能力。

在实际中，商业银行往往是先确定借款客户的最高贷款额度，再在最高额度的基础上，根据借款客户的贷款需求分批次给借款客户发放贷款，其中，所有贷款的总额不超过该借款客户的最高贷款额度。根据上述商业银行的实际放贷流程，可以推断，商业银行给贷款客户发放的每一笔贷款都可能因贷款金额、贷款利率、贷款期限等而产生差异。对于不同类型的贷款而言，由于贷款款项信息中的贷款本金、贷款利率、还款方式等不同，借款客户的相对还款能力也会发生改变。例如，对于一个最高贷款额度为二十万的借款客户来说，考虑两笔除了贷款本金不同，其他贷款款项信息都相同的个人贷款，假设其中一笔的贷款本金为一万元，另一笔的贷款本金为二十万元，那么，对于前一笔贷款而言，借款客户能按期还本付息的可能性更大，其相对的偿还能力也会更强；而对于后一笔贷款而言，由于其贷款本金与最高贷款额度相同，在到期日借款客户无法偿还本息的可能性更大，借款客户的相对偿还能力就越弱。因此，若要评价一个借款客户的偿还能力，不仅仅需要得到能体现其绝对偿还能力的信息，而且需要获取影响其相对偿还能力的贷款款项信息。由于商业银行在确定发放贷款前，会与借款客户签订一份包含详尽贷款信息的合同，从这个角度来看，商业银行在获取借款客户的每一笔贷款款项信息上几乎没有任何困难。因此，在上一段中确定了使用借款客户的基本信息和借款客户与商业银行交互的信息共同反映债务人的偿还能力之后，该段确定将个人贷款的款项信息作为衡量借款客户相对偿还能力的补充，借款客户的贷款款项信息包括贷款本金、贷款利率、贷款期限等信息。

除了以上数据信息之外，商业银行为了及时获取和捕获借款客户贷中时期偿还能力的变化，还需要借助其他途径进行外部信息获取。例如，通过宏观分析，根据借款客户所处行业的总体收入变动，来对该借款客户的偿还能力进行二次评估，或者是通过大数据分析，收集借款客户近期的消费数据，对其还款能力进行再次评估。然而，在可获得性方面，该部分数据获取的难度较大，通常需要商业银行与其他数据公司进行合作，而且这部分数据往往涉及用户隐私，在一般情况下，数据公司不会提供该外部数据给商业银行使用。在对这部分数据进行信息挖掘的难易程度方面，由于这部分数据属于非结构化数据，若要对该数据进行分析，往往需要用到自然语言处理（NLP）等技术，在商业银行没有足够的数据分析基础的情况下，为了分析外部信息而发展NLP等技术需要耗费大量的资金。最后，在数据效率方面，这种外部数据所蕴含的信息量与数据量之比远远低于客户基本信息、或者是贷款款项信息所蕴含的信息量与数据量之比，导致商业银行从这部分外部信息中得到的有效信息远远小于预期，而且为了存储和处理这部分外部信息，还需要投入大量的资金，用于建设分布式存储集群和分布式数据处理集群。基于以上三点，尽管在理论上，外部信息由于其实时性和可靠性，能对借款客户的偿还能力起到很好的预测效果，但是在实际的数据收集和数据分析中操作难度较大，因此本论文不考虑外部信息给借款客户还款能力所带来的影响，而只考虑贷款款项信息、客户基本信息和客户与商业银行的交互信息。

本文所考虑到的影响借款客户在贷中的偿还能力的信息特征，及本文选取该信息特征与否的说明表格如表3-1所示：

表3-1

|  |  |  |
| --- | --- | --- |
| 贷中影响借款客户偿还能力的特征信息 | | 本文考虑情况 |
| 借款客户信息 | 借款客户基本信息 | ✓ |
| 借款客户与商业银行的交互信息 | ✓ |
| 贷款款项信息 | | ✓ |
| 外部信息 | | 🗶 |

根据以上论述，在实际选取数据字段过程中，选取了五张能体现借款客户基本信息、借款客户与商业银行交互信息以及贷款款项信息的宽表，分别为客户基本信息表、信用卡信息表、贷款信息表、客户交易汇总表（近3个月）以及签约信息表。其中，客户基本信息表中体现了借款客户的基本信息，信用卡信息表、客户交易汇总表和签约信息表共同体现了借款客户与商业银行的交互信息，贷款信息表包含了贷款款项信息。

如表3-2所示，在客户基本信息表中，选取的字段有客户号（cust\_no）、数据日期（data\_date）、是否为行员（banker\_or\_not）、性别（gender）、婚否（marital\_status）、学历（edu\_status）、职业类型（profession\_type）、生日（birth\_date）和客户等级（cust\_level）共九个字段。在该表中，客户号（cust\_no）和数据日期（data\_date）为主键，即以客户号和数据日期作为唯一标识，说明同一个客户在不同时期的基本数据会发生变化，符合本文的假设，即在贷中时期，影响借款客户还款能力的特征信息会随时间变化。该表中，是否为行员字段是离散值字段，取值为是（Y）或否（N）；性别字段为离散值字段，取值为男（M）或女（F）；婚否字段为离散值字段，取值为是（Y）或否（N）；学历字段为离散值字段，取值为文盲或半文盲、小学、初中、高中、中等专业学校或中等技术学校、技术学校、研究生、博士研究生或博士后；职业字段为离散值字段，取值为税务、合伙、独资、法律/司法、广告、（军人、警察、武警）、私营类职业、邮电通信、农业畜牧、文化/娱乐/体育、咨询、交通运输、健康/医疗服务、一般工商业、旅游/餐饮/娱乐、房地产/建筑/装修、社会服务、科研/教育、个体、银行/金融/证券/保险、商业/贸易、IT/网络/计算机、制造业、国家机关事业单位、自由职业、学生、工薪类职业或其他职业；生日字段为连续值字段；客户等级字段为离散值字段，取值为1、2、3、4、5或6。

表3-2

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 是否为行员 | 0-1 | 是 |
| 性别 | 0-1 | 是 |
| 婚否 | 0-1 | 是 |
| 学历 | 字符 | 是 |
| 职业类型 | 字符 | 是 |
| 生日 | 字符 | 否 |
| 客户等级 | 数值 | 是 |

如表3-3所示，在信用卡信息表中，选取的字段有客户号（cust\_no）、数据日期（data\_date）、普卡持有数（common\_card\_num）、金卡持有数（gold\_card\_num）、贵宾卡持有数（vip\_card\_num）、白金卡持有数（platinum\_card\_num）、购易贷卡持有数（gyd\_card\_num）共七个字段。在该表中，客户号（cust\_no）和数据日期（data\_date）为主键，即以客户号和数据日期作为唯一标识，说明同一个客户在不同时期的基本数据会发生变化，符合本文的假设，即在贷中时期，影响借款客户还款能力的特征信息会随时间变化。该表中，普卡持有数、金卡持有数、贵宾卡持有数、白金卡持有数、购易贷卡持有数这五个字段均是离散值字段，取值为自然数。

表3-3

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 普卡持有数 | 数值 | 是 |
| 金卡持有数 | 数值 | 是 |
| 贵宾卡持有数 | 数值 | 是 |
| 白金卡持有数 | 数值 | 是 |
| 购易贷卡持有数 | 数值 | 是 |

如表3-4所示，在贷款信息表中，选取的字段有客户号（cust\_no）、数据日期（data\_date）、授信额度（credit\_line）、可用授信额度（available\_credit\_line）、贷款日期（loan\_date）、还款周期（repayment\_period）、还款期数（repayment\_term）、贷款金额（loan\_amount）、贷款利息（loan\_interest）、每期还款金额（repayment\_amount）、逾期本金（overdue\_principal）、逾期利息（overdue\_interest）共十二个字段。在该表中，客户号（cust\_no）和数据日期（data\_date）为主键，即以客户号和数据日期作为唯一标识，说明同一个客户在不同时期的基本数据会发生变化，符合本文的假设，即在贷中时期，影响借款客户还款能力的特征信息会随时间变化。该表中，授信额度、可用授信额度、贷款金额、贷款利率、每期还款金额、逾期本金和逾期利息字段都为连续值字段，这些字段的取值范围是不小于零的有理数。贷款日期、还款周期、还款期数字段为离散值字段，这些字段的取值范围是自然数。

表3-4

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 授信额度 | 数值 | 否 |
| 可用授信额度 | 数值 | 否 |
| 贷款日期 | 字符 | 否 |
| 还款周期 | 数值 | 否 |
| 还款期数 | 数值 | 否 |
| 贷款金额 | 数值 | 否 |
| 贷款利息 | 数值 | 否 |
| 每期还款金额 | 数值 | 否 |
| 逾期本金 | 数值 | 否 |
| 逾期利息 | 数值 | 否 |

如表3-5所示，在贷款信息表中，选取的字段有客户号（cust\_no）、数据日期（data\_date）、近三月交易金额（tra\_amo\_3mon）、近三月他行转入金额（ban\_tra\_amo\_3mon）、近三月我行转出金额（tra\_ban\_amo\_3mon）、近三月转出合计增量（tot\_inc\_tra\_amo）、近三月转入合计增量（tot\_tra\_agg\_inc）共七个字段。在该表中，客户号（cust\_no）和数据日期（data\_date）为主键，即以客户号和数据日期作为唯一标识，说明同一个客户在不同时期的基本数据会发生变化，符合本文的假设，即在贷中时期，影响借款客户还款能力的特征信息会随时间变化。该表中，近三月交易金额、近三月他行转入金额、近三月我行转出金额字段为连续值字段，这些字段的取值范围是不小于零的有理数，而近三月转出合计增量、近三月转入合计增量字段为连续值字段，这两个字段的取值范围是有理数。

表3-5

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 近三月交易金额 | 数值 | 否 |
| 近三月他行转入金额 | 数值 | 否 |
| 近三月我行转出金额 | 数值 | 否 |
| 近三月转出合计增量 | 数值 | 否 |
| 近三月转入合计增量 | 数值 | 否 |

如表3-6所示，在签约信息表中，选取的字段有客户号（cust\_no）、数据日期（data\_date）、是否签约网上银行（net\_bank\_sign）、是否签约手机银行（mobile\_bank\_sign）、是否签约第三方存管（third\_party\_sign）、是否签约基金（fund\_sign）、是否签约理财（manage\_sign）共七个字段。在该表中，客户号（cust\_no）和数据日期（data\_date）为主键，即以客户号和数据日期作为唯一标识，说明同一个客户在不同时期的基本数据会发生变化，符合本文的假设，即在贷中时期，影响借款客户还款能力的特征信息会随时间变化。该表中，是否签约网上银行、是否签约手机银行、是否签约第三方存管、是否签约基金、是否签约理财字段都是离散值字段，取值范围为是（Y）或否（N）。

表3-6

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 是否签约网上银行 | 0-1 | 是 |
| 是否签约手机银行 | 0-1 | 是 |
| 是否签约第三方存管 | 0-1 | 是 |
| 是否签约基金 | 0-1 | 是 |
| 是否签约理财 | 0-1 | 是 |

从商业银行的数据库中获取客户基本信息表、信用卡信息表、贷款信息表、客户交易汇总表（近3个月）和签约信息表这五张表中的部分字段后，需要对这五张表中的某些字段进行预处理，筛选出这些字段中符合要求的数据，才能对其中的每一笔贷款进行违履约预测。

## 原始数据预处理

根据本章第一部分选取的五张数据表信息，对于其中的每一张表，其每一条数据的唯一标识都是客户号（cust\_no）和数据日期（data\_date），因此，若要综合考虑借款客户的基本信息、借款客户与商业银行交互的信息，以及贷款款项信息，应根据相同的客户号（cust\_no）和相同的数据日期（data\_date），将这五张数据表中的字段合并，得到一张能将第一部分中五张表的所有数据字段合并在一起的宽表，该表的主键同样是客户号（cust\_no）和数据日期（data\_date），并且该表的每一条数据，能同时包含该客户在数据记录时间点时的基本信息、与商业银行交互的信息，以及该客户的贷款款项信息。该合并表的基本结构如表3-7所示：

表3-7

|  |  |  |
| --- | --- | --- |
| 字段名 | 数据类型 | 是否离散值 |
| 客户号 | 字符 | 否 |
| 数据日期 | 字符 | 否 |
| 是否为行员 | 0-1 | 是 |
| 性别 | 0-1 | 是 |
| 婚否 | 0-1 | 是 |
| 学历 | 字符 | 是 |
| 职业类型 | 字符 | 是 |
| 生日 | 字符 | 否 |
| 客户等级 | 数值 | 是 |
| 普卡持有数 | 数值 | 是 |
| 金卡持有数 | 数值 | 是 |
| 贵宾卡持有数 | 数值 | 是 |
| 白金卡持有数 | 数值 | 是 |
| 购易贷卡持有数 | 数值 | 是 |
| 授信额度 | 数值 | 否 |
| 可用授信额度 | 数值 | 否 |
| 贷款日期 | 字符 | 否 |
| 还款周期 | 数值 | 否 |
| 还款期数 | 数值 | 否 |
| 贷款金额 | 数值 | 否 |
| 贷款利息 | 数值 | 否 |
| 每期还款金额 | 数值 | 否 |
| 逾期本金 | 数值 | 否 |
| 逾期利息 | 数值 | 否 |
| 近三月交易金额 | 数值 | 否 |
| 近三月他行转入金额 | 数值 | 否 |
| 近三月我行转出金额 | 数值 | 否 |
| 近三月转出合计增量 | 数值 | 否 |
| 近三月转入合计增量 | 数值 | 否 |
| 是否签约网上银行 | 0-1 | 是 |
| 是否签约手机银行 | 0-1 | 是 |
| 是否签约第三方存管 | 0-1 | 是 |
| 是否签约基金 | 0-1 | 是 |
| 是否签约理财 | 0-1 | 是 |

得到能同时体现借款客户的基本信息、借款客户与商业银行交互的信息、贷款款项信息的宽表后，由于该宽表中的某些特征字段的数据类型是非数值型，而机器学习算法是无法处理非数值型字段的，因此需要按照某种转换规则，将这部分特征字段转换成数值型字段，每一个非数值型字段的转换规则如下：

对于学历字段而言，文盲或半文盲对应的数值为0、小学对应的数值为1、初中对应的数值为2、高中对应的数值为3、中等专业学校或中等技术学校对应的数值为4、技术学校对应的数值为5、研究生对应的数值为6、博士研究生对应的数值为7、博士后对应的数值为8。原始取值类型与对应的数值如表3-8所示：

表3-8

|  |  |
| --- | --- |
| 学历字段 | |
| 原始取值类型 | 对应的数值 |
| 文盲或半文盲 | 0 |
| 小学 | 1 |
| 初中 | 2 |
| 高中 | 3 |
| 中等专业学校或中等技术学校 | 4 |
| 技术学校 | 5 |
| 研究生 | 6 |
| 博士研究生 | 7 |
| 博士后 | 8 |

对于职业类型字段而言，税务对应的数值为0、合伙对应的数值为1、独资对应的数值为2、法律/司法对应的数值为3、广告对应的数值为4、（军人、警察、武警）对应的数值为5、私营类职业对应的数值为6、邮电通信对应的数值为7、农业畜牧对应的数值为8、文化/娱乐/体育对应的数值为9、咨询对应的数值为10、交通运输对应的数值为11、健康/医疗服务对应的数值为12、一般工商业对应的数值为13、旅游/餐饮/娱乐对应的数值为14、房地产/建筑/装修对应的数值为15、社会服务对应的数值为16、科研/教育对应的数值为17、个体对应的数值为18、银行/金融/证券/保险对应的数值为19、商业/贸易对应的数值为20、IT/网络/计算机对应的数值为21、制造业对应的数值为22、国家机关事业单位对应的数值为23、自由职业对应的数值为24、学生对应的数值为25、工薪类职业对应的数值为26、其他职业对应的数值为27。原始取值类型与对应的数值如表3-9所示：

表3-9

|  |  |
| --- | --- |
| 职业类型字段 | |
| 税务 | 0 |
| 合伙 | 1 |
| 独资 | 2 |
| 法律/司法 | 3 |
| 广告 | 4 |
| 军人、警察、武警 | 5 |
| 私营类职业 | 6 |
| 邮电通信 | 7 |
| 农业畜牧 | 8 |
| 文化/娱乐/体育 | 9 |
| 咨询 | 10 |
| 交通运输 | 11 |
| 健康/医疗 | 12 |
| 一般工商业 | 13 |
| 旅游/餐饮/娱乐 | 14 |
| 房地产/建筑/装修 | 15 |
| 社会服务 | 16 |
| 科研/教育 | 17 |
| 个体 | 18 |
| 银行/金融/证券/保险 | 19 |
| 商业/贸易 | 20 |
| IT/网络/计算机 | 21 |
| 制造业 | 22 |
| 国家机关事业单位 | 23 |
| 自有职业 | 24 |
| 学生 | 25 |
| 工薪类职业 | 26 |
| 其他职业 | 27 |

对于生日字段而言，需要将字符转换成年份数值，例如，若生日字段的字符为“198101”，则应提取出其中的年份1981，用现在的年份2019减去1981，即可得到该借款客户的年龄。

对于剩下两个非数值型字段数据日期和贷款日期，由于该宽表中包含贷款款项信息，贷款款项信息只有在贷款真正发放之后才会生效，并被记录入库。因此，从理论上看，对于该宽表上的每一条信息，都应该有：该条数据的数据日期不早于贷款日期。在实际操作过程中，发现只有十几条数据的数据日期早于贷款日期，因此该宽表的实际数据情况与预期相符。

对于普卡持有数（common\_card\_num）、金卡持有数（gold\_card\_num）、贵宾卡持有数（vip\_card\_num）、白金卡持有数（platinum\_card\_num）、购易贷卡持有数（gyd\_card\_num）这五个字段，由于大部分数据的普卡持有数字段都大于零，而其他四个字段的数值都是零，因此可以认为在本数据集中的借款客户基本上为普通客户。于是，将这五个字段的数值进行累加，将五个字段简化为一个字段，该字段称为信用卡持有数（cre\_card\_num）。

对于可用授信额度（available\_credit\_line）字段，由于在该宽表中已经存在授信额度（credit\_line）字段，因此将该字段转化为可用授信额度比（cre\_rate）字段，具体的转化方法是，将可用授信额度字段的数值除以信用额度字段的数值。

对于近三月他行转入金额（ban\_tra\_amo\_3mon）、近三月我行转出金额（tra\_ban\_amo\_3mon）这两个字段，由于这两个字段的数值之和与近三月交易金额（tra\_amo\_3mon）字段的数值相同，因此将这两个字段分别转换为近三月他行转入金额占比（last\_3mon\_in\_rate）字段和近三月我行转出金额占比（last\_3mon\_out\_rate），具体方法是分别用近三月他行转入金额字段和近三月我行转出金额字段的数值处以近三月交易金额字段的数值。

将非数值型数据转化为数值型数据之后，由于本文希望用每笔贷款真实的违履约情况作为该笔个人贷款的标签，而个人贷款只有在到期日才能观察到其是否违约。因此，需要筛选贷款到期日符合一定条件的个人贷款。例如，若某笔个人贷款的到期日为2020年1月1日，则该笔个人贷款不应纳入本次实验的考虑范围中，这是因为2020年1月1日还没有到来，无法得到该笔个人贷款真实的违履约标志。同时，观察到合并后宽表中的数据日期字段的区间段是从2016年10月至2018年11月，以月为间隔。因此，符合本文实验要求的个人贷款，其到期日应在2016年10月和2018年11月之间。然而，对于合并后得到的宽表而言，其体现贷款款项信息的字段中，并没有个人贷款到期日字段。因此，本文用贷款日期字段和还款周期字段，来计算每一笔个人贷款的到期日。其中，贷款日期字段表示该笔贷款的发放日期，例如，若贷款日期字段的值为“201701”，则说明该笔贷款是在2017年1月发放的。而还款周期则表示该笔个人贷款的存续期，例如，若还款周期字段的数值为36，则说明该笔贷款的存续期为36个月，即还款日期应为贷款日期的36个月之后。因此，可以利用贷款日期和还款周期这两个字段计算出该笔贷款的还款日期。综上所述，利用贷款日期和还款周期计算出每笔贷款的还款日期，并筛选还款日期在2016年10月到2018年11月之间的个人贷款。筛选符合要求的个人贷款的规则如表3-10所示：

表3-10

|  |  |  |
| --- | --- | --- |
| 序号 | 规则描述 | 所用字段 |
| 1 | 数据日期晚于贷款日期 | 数据日期、贷款日期 |
| 2 | 还款日期在2016年10月至2018年11月之间 | 贷款日期、还款周期 |

经过上述筛选，得到的所有贷款发放日和贷款到期日均在2016年10月至2018年11月之间的个人贷款。其中符合条件的数据条数为104200条。注意到还款周期（loan\_term）字段，还款周期为12，即个人贷款的存续期为12个月的数据量为90712条，占经过筛选后数据总量的87%。因此，筛选出还款周期为12的数据，不再考虑还款周期差异对个人贷款违履约的影响。经过此次筛选后，数据总量为90712条。

对数据进行条件筛选后，需要确定每一条数据的违履约标签。然而，注意到合并后的宽表中，在能体现贷款款项信息的字段中，并没有没有违履约标签字段，因此需要将该宽表中的某些字段转化成违履约标签字段。注意到在合并后得到的宽表中，含有逾期本金（overdue\_amt）字段和逾期利息（overdue\_int）字段，判断这两个字段的中的数值是否大于零，可以得知在数据日期（data\_date）所对应日期的一个月之内，该笔个人贷款是否违约。因此，将判断该笔个人贷款是否违约的问题，转换为判断该笔贷款的逾期本金或者是逾期利息是否大于零的问题，若这两个字段中有一个字段的数值大于零，则判定该笔贷款违约。

综上所述，对合并后的宽表进行预处理后，得到的宽表如表3-11所示：

表3-11

|  |  |  |
| --- | --- | --- |
| 字段名 | 作用 | 来源表 |
| 客户号 | 主键 | 所有表 |
| 贷款日期 | 主键 | 贷款信息表 |
| 数据日期 | 主键 | 所有表 |
| 是否违约 | 标签 | 贷款信息表 |
| 是否为行员 | 特征 | 客户基本信息表 |
| 性别 | 特征 |
| 婚否 | 特征 |
| 学历 | 特征 |
| 职业类型 | 特征 |
| 年龄 | 特征 |
| 客户等级 | 特征 |
| 信用卡持有数 | 特征 | 信用卡信息表 |
| 授信额度 | 特征 | 贷款信息表 |
| 可用授信额度比 | 特征 |
| 贷款期数 | 特征 |
| 贷款金额 | 特征 |
| 贷款利息 | 特征 |
| 每期还款金额 | 特征 |
| 近三月交易金额 | 特征 | 客户交易汇总表（近3个月） |
| 近三月他行转入金额占比 | 特征 |
| 近三月他行转出金额占比 | 特征 |
| 近三月转出合计增量 | 特征 |
| 近三月转入合计增量 | 特征 |
| 是否签约网上银行 | 特征 | 签约信息表 |
| 是否签约手机银行 | 特征 |
| 是否签约第三方存管 | 特征 |
| 是否签约基金 | 特征 |
| 是否签约理财 | 特征 |

## 数据描述性统计

### 离散值字段描述性统计

在本章第二部分得到的宽表中，离散值字段为是否为行员、性别、婚否、客户等级、是否签约网上银行、是否签约手机银行、是否签约第三方存管、是否签约基金和是否签约存管字段，一下列表为这些字段的描述性统计，如表3-12所示：

表3-12

|  |  |  |
| --- | --- | --- |
| 是否为行员 | | |
| 取值 | 数量 | 占比 |
| 是 | 5954 | 93.44% |
| 否 | 85758 | 6.56 |
| 性别 | | |
| 取值 | 数量 | 占比 |
| 男 | 30536 | 33.66% |
| 女 | 60176 | 66.34% |
| 婚否 | | |
| 取值 | 数量 | 占比 |
| 是 | 73501 | 24.41% |
| 否 | 17211 | 75.59% |
| 客户等级 | | |
| 取值 | 数量 | 占比 |
| 1 | 70618 | 83.81% |
| 2 | 9136 | 10.07% |
| 3 | 2579 | 2.84% |
| 4 | 2369 | 2.61% |
| 5 | 382 | 0.42% |
| 6 | 228 | 0.25% |
| 是否签约网上银行 | | |
| 取值 | 数量 | 占比 |
| 是 | 83292 | 91.8% |
| 否 | 7420 | 8.2% |
| 是否签约手机银行 | | |
| 取值 | 数量 | 占比 |
| 是 | 80992 | 89.28% |
| 否 | 9720 | 10.72% |
| 是否签约第三方存管 | | |
| 取值 | 数量 | 占比 |
| 是 | 84321 | 92.95% |
| 否 | 6391 | 7.05% |
| 是否签约基金 | | |
| 取值 | 数量 | 占比 |
| 是 | 90712 | 100% |
| 否 | 0 | 0 |
| 是否签约理财 | | |
| 取值 | 数量 | 占比 |
| 是 | 73362 | 80.87% |
| 否 | 17350 | 19.13% |

由以上列表可得，对于是否为行员字段，非行员的数据量占总数据量的93.44%，而是行员的数据量仅占总数据量的6.56%，由此可知，该批次贷款主要面向的客户是非银行工作人员客户。对于性别字段，66.34%为女性借款客户，而男性借款客户仅有33.66%，可知本批次贷款更加青睐女性群体。对于婚否字段，已婚的数据量占总数据量的75.59%，而未婚的则占24.41%，由此可得商业银行更愿意向已婚人士发放贷款，或者说已婚人士对于贷款的需求程度更高。对于客户等级字段，客户等级最高的1级客户占所有客户的83.83%，而等级最低的客户则占0.25%，说明商业银行会向少数等级偏低的客户发放贷款。对于是否签约网上银行、是否签约手机银行、是否签约第三方存管和是否签约理财这四个字段，取值为是的占比都超过80%，而只有是否签约基金字段的取值全部为是，因此该字段对于预测个人贷款违履约没有任何贡献，故将该字段剔除。

### 连续值字段描述性统计

连续值字段的描述性统计如表3-13所示：

表3-13

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 字段 | 均值 | 标准差 | 0 | 25% | 50% | 75% | 最大值 |
| 年龄 | 40.1 | 7.5 | 23 | 34 | 39 | 46 | 61 |
| 信用卡持有数 | 1.25 | 0.46 | 1 | 1 | 1 | 1 | 4 |
| 授信额度 | 63.01万 | 78.35万 | 10万 | 20万 | 30万 | 70万 | 500万 |
| 可用授信额度占比 | 0.1172 | 0.21 | 0 | 0 | 0 | 0.15 | 1 |
| 还款期数 | 10.29 | 2.77% | 1 | 10 | 10 | 12 | 12 |
| 贷款金额 | 26万 | 38万 | 0 | 5万 | 13.1万 | 29.8万 | 475万 |
| 贷款利息 | 5.83% | 1.24 | 0.001% | 4.83% | 4.89% | 4.90% | 18% |
| 每期还款金额 | 72 | 810.28 | 0 | 0 | 0 | 0 | 24072.86 |
| 近三月交易金额 | 40.1万 | 434万 | 0 | 1万 | 12.93万 | 65.08万 | 411.84万 |
| 近三月他行转入金额占比 | 0.63 | 0.26 | 0 | 0.5 | 0.52 | 0.998 | 1 |
| 近三月我行转出金额占比 | 0.32 | 0.24 | 0 | 0 | 0.43 | 0.49 | 1 |
| 近三月转出合计增量 | -13123 | 106万 | -109万 | -1万 | 0 | 11461 | 133万 |
| 近三月转入合计增量 | -12097 | 111万 | -113万 | -1万 | 0 | 14700 | 137万 |

对于年龄字段，最小年龄为23，最大年龄为61，该字段说明这批贷款主要面向有还款能力的未退休成年人；对于信用卡持有数字段，所有借款客户的信用卡持有数量都不小于一，说明只有拥有信用卡的客户才有可能成功申请贷款，而绝大部分的借款客户的信用卡持有数都为1，则说明商业银行会根据客户的情况发放适合客户的信用卡。对于授信额度字段，观察到最低授信额度为10万，25%、50%和75%分位数分别为10万、20万、30万和70万，据此推测商业银行的授信额度梯度为10万。对于可用授信额度占比字段，大部分数值都为零，说明借款客户申请的贷款金额，往往与其最高额度相同；对于还款期数字段和每期还款金额字段，在每期还款金额字段中，绝大部分数值为0，而均值仅仅为72，说明在该批次贷款中，尽管每笔贷款都有相应的期数，但是几乎所有贷款每期的还款额都是0，说明绝大多数的贷款只有在款项到期日才要求还本付息，而在款项到期日之前，则不需要偿还本金或支付利息。因此，将还款期数字段和每期还款金额字段剔除，并将所有贷款视为贷款期数为1的贷款，即只有在贷款到期日才需要还本付息，在贷款到期日之前不需要付任何费用；对于贷款金额字段，发现有异常值零，应将该数据从数据集中剔除；对于贷款利息字段，其最小值为0.001%，即该笔贷款很可能是无利息贷款或者是“人情”贷款，不收取任何贷款费用，同时，观察到绝大多数的贷款利息在4.9%左右，可以推断出该批次贷款可能是相同类型贷款；对于近三月交易金额、近三月转出合计增量和近三月转入合计增量这三个字段，观察到该字段的标准差极大，说明用户与该商业银行之间的亲和度差异极大；对于近三月他行转入金额占比和近三月我行转出金额占比这两个字段，相对而言，由他行转入金额的占比稍大，但基本上相差无几。

经过以上描述性统计分析，将贷款期数、每期还款金额和是否签约基金字段剔除，原因是贷款期数和每期还款金额这两个字段相互矛盾，可能会对贷款的违履约预测产生不利的影响，而在该数据集中，所有的数据在是否签约基金字段的值都为是，因此该字段对贷款的违履约预测不会产生任何影响。经过描述性统计分析后，得到的宽表字段如表3-14所示：

表3-14

|  |  |  |
| --- | --- | --- |
| 字段名 | 作用 | 来源表 |
| 客户号 | 主键 | 所有表 |
| 贷款日期 | 主键 | 贷款信息表 |
| 数据日期 | 主键 | 所有表 |
| 是否违约 | 标签 | 贷款信息表 |
| 是否为行员 | 特征 | 客户信息表 |
| 性别 | 特征 |
| 婚否 | 特征 |
| 学历 | 特征 |
| 职业类型 | 特征 |
| 年龄 | 特征 |
| 客户等级 | 特征 |
| 信用卡持有数 | 特征 | 信用卡信息表 |
| 授信额度 | 特征 | 贷款信息表 |
| 可用授信额度比 | 特征 |
| 贷款金额 | 特征 |
| 贷款利息 | 特征 |
| 近三月交易金额 | 特征 | 客户交易汇总表（近3个月） |
| 近三月他行转入金额占比 | 特征 |
| 近三月他行转出金额占比 | 特征 |
| 近三月转出合计增量 | 特征 |
| 近三月转入合计增量 | 特征 |
| 是否签约网上银行 | 特征 | 签约信息表 |
| 是否签约手机银行 | 特征 |
| 是否签约第三方存管 | 特征 |
| 是否签约理财 | 特征 |

# 机器学习训练预测

## 数据集分组

由于本文需要对贷中时期的个人贷款进行违履约预测，并验证如下假设：当个人贷款的数据日期距离款项到期日越近时，该笔款项的信息对于个人贷款违履约的预测力越强，预测准确率越高；反之，则越弱，预测准确率越低。因此，需要按数据日期到款项到期日的时间段长度对数据集进行划分，再通过对比不同时间段长度的个人贷款信息的预测力强弱，来验证该假设是否成立。由于在预处理中，所有数据的还款周期均为12个月，即对于该数据集中的每一条数据，其数据日期到款项到期日之间的时间长度不小于0个月，同时不大于12个月。利用数据日期、贷款日期和还款周期这三个字段得到一个距离还款日长度字段后，对该字段进行描述性统计处理，得到的结果如表4-1所示：

表4-1

|  |  |
| --- | --- |
| 属性 | 数值 |
| count | 90712 |
| mean | 4.526072 |
| std | 3.193074 |
| min | 0 |
| 25% | 2 |
| 50% | 4 |
| 75% | 7 |
| max | 12 |

由该描述性统计可以看出，最小值为距离到期日0个月，25%分位数为距离到期日2个月，50%分位数为4个月，75%分位数为7个月，最大值为12个月。因此，根据使得数据集尽可能均匀划分的原则，将第三章中得到的数据集划分为四部分，分别为距离贷款到期日0、1和2月的数据集，称为➀号数据集；距离贷款到期日3和4月的数据集，称为➁号数据集；距离贷款到期日5、6和7月的数据集，称为➂号数据集；距离贷款到期日8、9、10、11和12月的数据集，称为➃号数据集。结合第三章数据预处理过程，可知数据日期距离贷款到期日的取值范围为从0到12的整数，因此，该步骤将数据集划分为互不相交的四个数据集。数据划分对应表格如表4-2所示：

表4-2

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 数据日期距离到期日月份（长度） | 数据量 | 占比 |
| ➀ | 0、1、2 | 30754 | 33.90% |
| ➁ | 3、4 | 25590 | 28.21% |
| ➂ | 5、6、7 | 15553 | 17.15% |
| ➃ | 8、9、10、11、12 | 18815 | 20.74% |

这四个数据集的标签数量及占比对应表格如表4-3所示：

表4-3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 履约数量 | 违约数量 | 履约占比 | 违约占比 |
| ➀ | 30571 | 183 | 99.40% | 0.60% |
| ➁ | 25587 | 3 | 99.99% | 0.01% |
| ➂ | 15542 | 11 | 99.93% | 0.07% |
| ➃ | 18807 | 8 | 99.96% | 0.04% |

考虑到在第三章得到的宽表中，每一条数据的主键是客户号、贷款日期和数据日期，说明只有同时得到客户号、贷款日期和数据日期这三个字段的数据，才能在宽表中得到唯一的一条数据。因此，有理由推测，对于同一个客户的同一笔贷款，在还款周期（12个月）内，因其数据日期不同而被记录多次，而且，当该笔贷款的数据日期不同时，该笔贷款的客户基本信息、客户也商业银行的交互信息也产生了相应的变化。为了验证同一个客户的同一笔贷款在贷款存续期（12个月）内的不同时间点都有记录，本文记录了每一条数据的客户号和贷款日期，结果发现，在90712条数据中，有19424条相同的（客户号，贷款日期）数据，说明对于同一个客户的同一笔贷款，在12个月的贷款存续期内，平均而言，会按月为间隔，被记录4.7次。

由于同一个客户的同一笔贷款在贷款存续期（12个月）内被记录了多次，结合第三章中大大部分个人贷款只有在款项到期日才需要还本付息的结论，对上表中➁、➂和➃号数据集中违约数据量极低的合理推测是：这些违约的贷款要求分期还本付息，因此在贷款到期日前就有逾期本金和逾期利息的记录。然而，该数据集中的大部分个人贷款只需要在贷款到期日还本付息，其在未到期前不要求借款客户还本付息，因此也就不存在违约记录。因此对于不需要分期偿付的个人贷款，其在到期日之前均标记为履约。这就解释了表中➁、➂和➃号数据集中违约数据量极低的现象。

理清表中➁、➂和➃号数据集中违约数据量极低的缘由后，需要对数据的违约字段进行调整，理由是：对于不需要在贷款到期日之前还本付息的个人贷款来说，其在贷款到期日之前不会产生违约信号，统一被标记为履约，而只有在贷款到期日，该笔贷款才会标上真实的违履约标记。因此，需要对数据集中19424条以（客户号，贷款日期）对为唯一标识的个人贷款的违履约标记进行调整，具体方法是：对于每一条以（客户号，贷款日期）对唯一表示的个人贷款，若该贷款某一条数据的违履约标记为违约，则该笔贷款在贷款日期字段上的数据的违履约标记均为违约。对数据集中每一条数据的违履约字段进行调整后，得到的四个数据集的标签数量及占比对应表格如表4-4所示：

表4-4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 履约数量 | 违约数量 | 履约占比 | 违约占比 |
| ➀ | 30222 | 532 | 98.27% | 1.73% |
| ➁ | 25328 | 262 | 98.97% | 1.03% |
| ➂ | 15392 | 161 | 98.96% | 1.04% |
| ➃ | 18583 | 232 | 98.76% | 1.24% |

根据以上表格，经过违履约标记调整后，不同数据集的违约率大约为1.24%，较为合理。

## 机器学习模型训练预测

对上一节得到的四个数据集分别随机划分为训练集和测试集，其中，训练集占70%，测试集占30%，然后，分别用决策树、随机森林、Logit模型和catboost进行训练，得到学习器，再用学习器对测试集进行预测，最后用模型评价指标对模型进行评价。

模型评价指标结果如下：

### 混淆矩阵

表4-5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 决策树 | ➀ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 9018 | 64 |
| 违约 | 27 | 118 |
| ➁ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 7576 | 26 |
| 违约 | 21 | 54 |
| ➂ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 4597 | 36 |
| 违约 | 11 | 22 |
| ➃ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 5529 | 34 |
| 违约 | 32 | 50 |
| 随机森林 | ➀ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 9016 | 66 |
| 违约 | 25 | 120 |
| ➁ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 7564 | 38 |
| 违约 | 18 | 57 |
| ➂ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 4590 | 26 |
| 违约 | 15 | 35 |
| ➃ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 5523 | 40 |
| 违约 | 28 | 54 |
| Logit | ➀ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 9021 | 61 |
| 违约 | 29 | 116 |
| ➁ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 7568 | 36 |
| 违约 | 18 | 55 |
| ➂ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 4586 | 28 |
| 违约 | 15 | 37 |
| ➃ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 5536 | 47 |
| 违约 | 20 | 42 |
| Catboost | ➀ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 9022 | 58 |
| 违约 | 20 | 125 |
| ➁ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 7570 | 37 |
| 违约 | 15 | 55 |
| ➂ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 4596 | 34 |
| 违约 | 9 | 27 |
| ➃ | 真实值\预测值 | 履约 | 违约 |
| 履约 | 5554 | 25 |
| 违约 | 18 | 48 |

对于表4-5，由于在4个数据集中，履约样本和违约样本的数据量各不相同，根据混淆矩阵中的绝对数值难以得出结论，因此，将混淆矩阵中的绝对数值转换为准确率、召回率和F1值，如下表（由于测试集中存在严重类别不平衡问题，导致履约类的Precision、Recall和F1-Score都接近1，因此不再关注履约类，仅关注违约类），结果如表4-6所示：

表4-6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 违约类 | | Precision | Recall | F1-Score |
| 决策树 | ➀ | 0.65 | 0.81 | 0.72 |
| ➁ | 0.68 | 0.72 | 0.70 |
| ➂ | 0.38 | 0.67 | 0.48 |
| ➃ | 0.60 | 0.61 | 0.60 |
| 随机森林 | ➀ | 0.65 | 0.83 | 0.73 |
| ➁ | 0.60 | 0.76 | 0.67 |
| ➂ | 0.57 | 0.70 | 0.63 |
| ➃ | 0.57 | 0.66 | 0.61 |
| Logit | ➀ | 0.66 | 0.80 | 0.72 |
| ➁ | 0.60 | 0.75 | 0.67 |
| ➂ | 0.57 | 0.71 | 0.63 |
| ➃ | 0.47 | 0.68 | 0.56 |
| Catboost | ➀ | 0.68 | 0.86 | 0.76 |
| ➁ | 0.60 | 0.79 | 0.68 |
| ➂ | 0.44 | 0.75 | 0.56 |
| ➃ | 0.66 | 0.73 | 0.69 |

对于商业银行而言，其更关注的是：模型是否正确判断了所有可能违约的个人贷款。因此，对于违约贷款是否“找得全”的问题，应主要关注召回率（Recall）指标，对于每一个数据集下，每一个算法的召回率对比如图4-1所示：

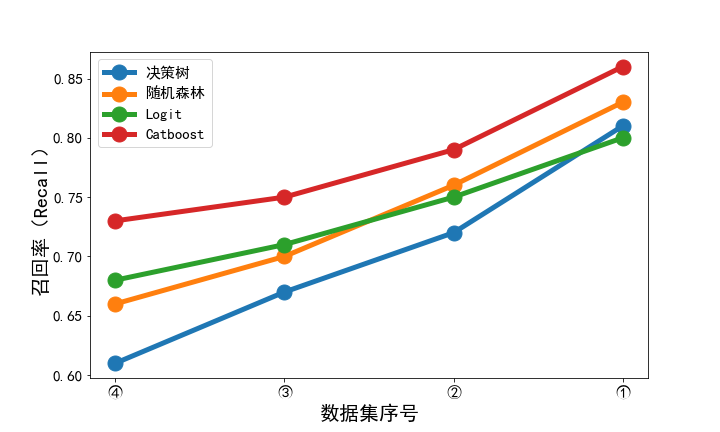


图4-1

由上图可见，对于违约类数据，对于每一个模型而言，都有➃数据集的召回率小于➂数据集的召回率小于➁数据集的召回率小于➀数据集的召回率。因此，可以说明，当个人贷款的真实信息越接近与贷款到期日时，其预测能力越强。

# 总结与展望

## 总结

本文在前人研究的基础上，结合商业银行的实际情况对个人贷款进行了分析。通过阅读文献，发现现有文献中对个人贷款违履约情况研究的不足，提出了在贷中时期，根据个人贷款的信息，对贷款进行违履约预测的想法。同时，结合时间长度与风险之间的关系，提出了距离贷款到期日的时间长度与个人贷款信息预测力之间的关系，即当记录个人贷款信息的时间点与贷款到期日的时间长度越近时，该个人款项信息的预测力越强，用该信息来预测对应的个人贷款是否会违约的准确性越高；反之，则越低。

为了验证该假设，本文采用了商业银行有关个人贷款的原始数据，该原始数据涵盖了个人贷款债务人的基本信息、在贷中时期债务人与商业银行的交互信息、以及个人贷款款项的基本信息。获取到该原始数据后，本文首先筛选出其中在理论上，对预测个人贷款违履约具有贡献性的字段，再通过描述性统计分析，剔除其中数据分布不合理的字段，得到用于实验的数据集。接着，按距离个人贷款到期日的时间长短，将该数据集划分为四个互不相交的数据集，利用先前研究表明，适用于个人贷款违履约预测的决策树、随机森林、Logit模型、以及近几年提出的Catboost模型，对四个数据集分别进行训练及预测。将四个数据集的预测结果进行对比，得到以下结论：随着个人贷款信息的记录时间越接近于贷款到期日，该信息对贷款是否会违约的预测力越强、准确率越高。

综上所述，文本可能的贡献在于，舍弃了以往论文用贷前数据预测贷后个人贷款是否会违约的做法，而用贷后不断发生变化的个人贷款信息来对款项违履约进行预测，并通过实证研究，证实了当个人款项信息距离贷款到期日越远时，用该信息来预测个人贷款是否会产生违约的准确率越低，该信息的价值也就越低。因此，商业银行不能只关注债务人的贷前信息，而是需要在贷后实时收集债务人信息，分析其还款能力的变化情况，以减少贷款所带来的风险。

## 启示与展望

本文通过研究发现，不同时间点的贷后信息对个人贷款的预测能力不同，对于更接近贷款到期日的贷后信息，其预测能力越强。基于这个结论，若要基于本文进行进一步的分析，可以通过对比贷后不同时间点的个人贷款信息的差异，找出同一笔贷款在不同时间点上贷款信息的差异，该差异很可能就是影响个人贷款违约的重要因素。进一步地，深入分析导致个人贷款信息变化的因素，也许能够对进一步深化个人贷款风险管理产生促进行性作用。

另一方面，由于本文利用的个人贷款数据都是如性别、年龄等非常大众化的数据，而基于贷款信息作违履约预测，从理论上来说，需要能区分每一个债务人的信息。因此，对于商业银行而言，需要挖掘并收集能更多地体现客户还款能力的信息。在这方面，需要国家、商业银行和借款客户三者之间共同的努力。第一，需要国家政策的支持，出台新的政策进行宏观调控；第二，需要银行完善债务人信息体制，一个合理的做法是商业银行与债务人签订信息保障合约；第三，商业银行完善债权人的个人信息核实机制。

# 参考文献：

[1] Altman E I , Haldeman R G , Narayanan P . ZETATM analysis A new model to identify bankruptcy risk of corporations[J]. Journal of Banking & Finance, 1977, 1(1):0-54.

[2] Martin D . Early warning of bank failure : A logit regression approach[J]. Journal of Banking & Finance, 1977, 1(3):249-276.

[3] Mcdermed A A. Maddala, G. S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press, 1983, xi + 401 pp. $39.50[J]. Cambridge England Cambridge University Press, 1983, 79(387):80-81.

[4] West R C . A factor-analytic approach to bank condition[J]. Journal of Banking & Finance, 1985, 9(2):0-266.

[5] Laitinen E K , Laitinen T . Bankruptcy prediction Application of the Taylor's expansion in logistic regression[J]. International Review of Financial Analysis, 2000, 9(4):327-349.

[6] HENLEY, W. E , D. J . A k-nearest-neighbour classifier for assessing consumer credit risk[J]. Journal of the Royal Statistical Society, 1996, 45(1):77-95.

[7] West D . Neural network credit scoring models[J]. Computers & Operations Research, 2000, 27(11-12):1131-1152.

[8] 鲁炜, 赵恒珩, 刘冀云. KMV模型关系函数推测及其在中国股市的验证[J]. 运筹与管理, 2003, 12(3).

[9] 肖艳. 上市公司财务困境预警的Logit模型实证研究[J]. 湖南工程学院学报（社会科学版）, 2004, 14(4):11-14.

[10] 梁琪. 企业经营管理预警：主成分分析在logistic回归方法中的应用[J]. 管理工程学报, 2005, 19(1):100-103.

[11] 马若微. 基于RS与ANN的上市公司财务困境预测模型的实证研究[J]. 南开管理评论, 2006(3):85-91.

[12] 肖文兵, 费　奇. 基于支持向量机的个人信用评估模型及最优参数选择研究[J]. 系统工程理论与实践, 2006, 26(10):73-79.

[13] 彭建刚, 吕志华. 基于行业特性的多元系统风险因子CreditRisk+模型[J]. 中国管理科学, 2009, V17(3):56-64.

# 致谢

本文从论文选题，理论模型设定，编程实现和论文撰写过程，都是在李心丹、方立兵老师的指导下完成。李老师在传授系统知识的同时，启发我独立思考和解决问题的能力，鼓励创新型思维和敢于实践的思想。让我接受前沿的机器学习技术，为今后的学习和研究打下了坚实的基础。另外，李老师严谨的科研态度，发散性思维以及孜孜不倦的探索精神是我学习的目标和榜样，指导我在接下来的研究生活中不断开拓进取。我的毕业论文凝聚了李老师的辛勤的果实，在此真诚的感激我的指导老师李心丹。

此外，还得感谢李老师课题组的研究生学长们，他们给予我巨大的帮助，帮助我在初步探索科研的道路上少走弯路，多培养研究性思维。感谢大学四年的同学，和他们四年的相处中，不仅收获了知识，更收获了无价的友谊。感谢母校南京大学给予我良好的生活条件和浓厚的学习氛围，让我在学习和生活中感受到关怀。

最后感谢我的父母亲，是他们无微不至得关怀和奉献让我能踏入南京大学的学习平台，感受世界的乐趣，让我树立正确的世界观人生观和价值观，明确未来的奋斗目标。为家庭增添荣誉，为学校增添荣誉，为祖国增添荣誉。