个人消费信用贷款违约风险的因素研究

辣条

摘要

本研究旨在探究个人消费信用贷款违约风险的因素，并提出一种基于 Logistic 回归模型的预测方法。通过对数据集进行数据预处理、描述统计和可视化分析，我们找到了影响个人消费信用贷款违约风险的关键因素，并使用 Logistic 回归模型对其进行预测。结果表明，模型在预测无违约情况时表现良好，但在有违约情况时存在偏差。在未来的研究中，可以考虑增加更多的特征或采用其他算法来进一步提高模型性能。总之，该预测方法对金融机构提供了一种有用的决策支持工具，可以帮助其更好地评估个人消费信用贷款的违约风险。

关键词：逻辑回归 贷款违约 个人征信　预测 混淆矩阵

Abstract

The aim of this study is to explore the factors that affect the risk of personal consumer credit loan default and propose a prediction method based on the logistic regression model. By preprocessing, descriptive statistics, and visualization analysis of the dataset, we found key factors that affect the risk of personal consumer credit loan default and predicted them using the logistic regression model. The results showed that the model performed well in predicting no default but had bias in predicting default. In future studies, more features can be added or other algorithms can be used to further improve the model performance. In summary, this prediction method provides a useful decision support tool for financial institutions to better evaluate the risk of personal consumer credit loan default.

Key Words: logistic regression；loan default；personal credit； prediction；confusion matrix.

目录

第一章　　绪论

1.1 研究背景与意义

1.1.1 研究背景

根据 2023 年国家统计局发布的我国经济数据。2022 年，实物商品网上零售额比上年增长 6.2%；2022 年，商品零售额比上年增长 0.5%，新能源乘用车零售约 567 万辆，比上年增长 90%。种种数据显示**我国消费市场规模大潜力足，消费持续升级发展**。消费发展长期向好基本面没有改变。从近期情况看，随着疫情防控政策优化落实、消费场景有序恢复，前期受疫情影响较大的餐饮等接触型服务消费开始有所改善。从长期发展潜力看，我国总人口超过 14 亿人，城镇化率稳步提升，乡村市场蕴藏较大潜力，有力支撑我国消费市场稳定恢复发展。消费升级趋势没有改变。疫情发生前几年，居民消费已呈现出升级发展态势，即使受到疫情冲击，消费升级发展趋势没有改变。当前，居民品质化需求持续增加，绿色环保理念更加深入人心，服务消费意愿依然强烈。伴随消费场景创新拓展，居民收入稳步增长及市场供给不断完善，消费结构将持续优化升级。尽管 2022 年疫情对消费市场恢复产生了较大影响，但也要看到，消费市场韧性犹在、潜力较大。进入 2023 年，随着疫情防控优化调整措施进一步落实，扩大内需战略深入实施，以国内大循环为主体、国内国际双循环相互促进的新发展格局加快构建，消费市场稳定恢复的基础将更加牢固，消费市场有望恢复向好。

根据中国人民银行公布的数据，我国自 2013 年，人民币消费贷款余额始终是增长趋势，个人短期消费贷款除了 2020 年受疫情影响外相对前一年减少外，其他年份也是增长的趋势。2020 年金融机构贷款投向统计报告显示，2020 年末，普惠小微贷款余额 15.1 万亿元，同比增长 30.3%，增速比 2019 年末高 7.2 个百分点；全年增加 3.52 万亿元，同比多增 1.43 万亿元。个人短期消费贷款目前大

约在十万亿左右。而目前的金融业发展始终秉持底线思维，行业发展重点深入统筹重大金融风险 防范化解工作。因此在不断探索多种市场的形势的同时，也要坚持法制化，健全金融风险管控模式，符合审慎监管的要求，坚决遏制在金融市场发展中的各类风险反弹回溯，防止在处置其他领域风险过程中引发次生金融风险。

1.1.2 研究意义

个人全息信息是指个人在社会、经济、金融等各个方面活动中形成的完整而多维度的数据信息。这些数据可以反映个人的信用历史、还款能力、职业稳定性、收入水平、消费习惯等方面的情况。

对于贷款机构来说，了解客户的个人全息信息能够帮助其更好地评估贷款申请人的信用风险，制定更加合理的贷款政策和利率，并且加强违约管理和防范风险。因此，分析、评估个人全息信息对于贷款机构具有重要意义。

与传统的分析方法相比，以个人全息信息为基础的分析方法具有以下几个优势和不同点：

a. 数据来源更加全面和可靠。个人全息信息覆盖了多个领域，包括金融、教育、医疗、社保等，数据来源更加广泛，更加全面和可靠。
b. 分析效果更加准确和精细。使用个人全息信息进行分析，能够充分了解客户的信用历史、还款能力、职业稳定性、收入水平、消费习惯等方面的情况，从而能够提高风险评估的准确性和精细度。
c. 综合考虑多个因素。以个人全息信息为基础的分析方法可以结合多个因素进行综合评估，例如个人信用记录、财务状况、行业趋势等，这样能够更好地发现潜在的风险和机会。

我的研究内容主要是对影响个人消费信用贷款违约的因素进行分析，并建立相应的预测模型。通过深入研究影响个人消费信用贷款违约的各种因素，如贷款人的年龄、年收入、工作时长等，可以帮助贷款机构更好地了解客户的信用风险，制定更加合理的贷款政策和利率。同时，在建立预测模型的过程中，我将使用含有全息信息的数据集进行研究，从而能够更加准确地预测个人消费信用贷款违约率。并基于我的实证结果，为助贷款机构降低个人消费信用贷款违约率，保护金

融市场稳定，提高个人信用贷款的可持续性而提出一些有效的建议和措施。

1.2 研究方法

本文主要的研究方法是文献研究法、跨学科研究法、实证研究法，描述统计分析法、Logistic 回归研究法，通过利用这些方法，研究对个人消费信用贷款违约风险的预测模型的搭建，及个人消费信用贷款违约风险因素分析。

1.2.1 文献研究法

在中国知网、万方数据库、电子科技大学图书馆等渠道广泛收集、查阅、鉴别、整理有关个人消费信用贷款违约的相关研究成果，形成对该领域的科学认识与研究方法，总结已有的文献，提出相应的不足，启发选题与研究思路。。

1.2.2 Logistic 回归研究法

Logistic 回归研究法主要应用于分析一个二元因变量与多个自变量之间的关系，在本研究中，分析个人消费信用贷款是否违约（是或否）与其它多个因素（如收入水平、负债情况、工作时长等）之间的关系。

1.2.3 实证研究方法

建立个人消费信用贷款违约的风险预测模型。利用收集到的个人全息信息数据集，采用描述性统计分析法和 Logistic 回归研究法，分别对数据进行初步分析和处理。然后，根据实证分析的结果，建立个人消费信用贷款违约的风险预测模型，并对该模型的精度和稳定性进行验证和评估。

1.3 国内外文献综述

1.3.1 个人消费信用贷款的认识理论的相关文献

当全球社会进入相对稳定发展的阶段时，繁荣生产的商品催生出消费贷款，在之前的贷款发展中，抵押物始终是不可或缺的核心，而随着社会发展稳定，贷款人能够通过财务信息去确定借款人的责任，那么抵押物不再成为消费贷款不可获取的基础。弗兰科·莫迪利安尼(1954)提出的经典理论——生命周期理论，把个人生命周期作为自然人安排其生命周期内的储蓄和消费关系的前提做出总结，总效用最大化的实现成为了家庭消费规划的最终目标。但是家庭收入很难保证维持较长时间稳态，当消费和总收入存在不匹配时期，收入无法满足当期消费，消费贷款就成为收入短期中断的补充。Japelli 和 Pagano(1989)[1]对消费者借贷较少的国家进行研究，发现消费对当前收入波动的过度敏感程度明显更高，贷款的确为生活带来便利。

同时 $Dean.M.Maki$[2]研究发现，消费者当期消费开始需要从过去收入的积累中获得，一旦再无法满足，那么就必须使用未来的收入作为消费，因此个人消费信用贷款的需求出现了。最初的个人消费信用贷款首先是为了满足家庭必须消费，并且当时的学者和金融从业人士认为对未来收入的消费会导致未来社会经历的减缓，但是 Maki 的研究指出消费信贷对社会有积极作用。随着全球进入发展阶段，各国经济发展也进入平稳期，个人参与社会中的生产方式逐渐清晰，个人收入越来越成为可预测的一个指标，再伴随商品生产迫切需要消费实现再生产，助推了不需要抵押物的个人消费信用贷款的飞速发展。随着互联网发展，Dinh 和 Kleimeier(2007)[3]研究表明，已经有越来越多的个人消费贷款业务，已经脱离线下金融机构，转为了通过互联网办理。目前我国个人信用贷款发展迅速，几本涵盖了生活消费项目，行业从粗放竞争，开始进入深耕存量发展的环节，但是个人消费贷款还存在着一些问题，个人征信的系统的纳入范围还不够全面，个人消费信用贷款亟待个人征信系统的完善。

1.3.2 对于影响个人消费信用贷款的各种因素的相关文献及成果

个人消费信用贷款的发展离不开对违约风险的度量，只有能够准确把握坏账概率，才能助推金融机构健全发展相关业务。自从个人消费信用贷款出现以来，如何评价违约因素的方法就在不断出现新的，而且随着科技进步对于个人信用的评判维度越广,个人消费信用贷款的影响因素，也随着不断发展的技术在被甄别凸显出来。国内外个人消费信用贷款的发展时间不同，我国个人征信起步较慢，但是随着互联网的发展，我国个人消费信用贷款的发展更加繁荣。很多国内外学者对个人消费信用贷款的违约因素做出了研究。经济整体环境直接对贷款的发放产生影响，Rajian(1994)[4]对银行信贷政策的研究中，发现贷款的条件与经济发展相关，经济情况好，银行信贷发放要求低。一旦经济不景气，信贷服务收紧，整个市场都出现融资困难，在经济下行期间，对消费 贷款发放情况的研究中，指出信用好的人获得概率较高。

Lehnert(2014) [5]通过估算不同年龄的家庭在房价冲击中的消费弹性来检验房产价格冲击对信贷市场的影响，发现年轻的家庭面临着更快的预期收入增长，因此愿意比年长的家庭借更多的钱。

Fagereng 和 Halvorsen(2016) [6]在研究债务对消费的作用时发现，高债务家庭消费支出增长较低。危机过后，消费增长趋于平稳，这在很大程度上反映了高负债家庭的经常性反应。尽管如此，危机后的一种更强的关系表明预防性储蓄可能发挥了作用。

Chen(2002) [7]研究发现，对于不同教育背景的大学生，对于个人财务知识了解程度和也不同。Jonathan 和 Tony(2003) [7]对个人消费信用贷款发展相对完善的国家研究，得出贷款的发放和申请者的年龄相关的结论。Lusardi 和 Mitchell(2007) [8]对于婴儿潮退休的一代人的退休账户状况的研究反映，学历和收入对于个人财务情况有着很明显的相关性。李广明等(2011)[8]在对消费贷款中具有拖欠贷款行为借款者的基本特征研究发现，信用贷款借款利率越高、借款期限越短，借款人信用违约概率越高；借款利率高，借款人容易采取投机行为到期

后成本和利息金额过大，借款人无法及时还款。而借款期限短，到期后借款人资金周转紧张，信用违约风险较高。

基于上述学者研究，本文最终选用选取年龄、教育、工龄、收入、负债率和信用卡负债作为自变量，这些因素反映了客户的经济状况以及还款能力等方面的信息，是评估客户风险的重要依据。具体而言，年龄可能与个人的收入、职业稳定性等相关；教育水平可能与客户的职业、职业水平等有关；工龄可能反映客户的工作稳定性；收入则是评估借款人还款能力的一个重要指标；负债率和信用卡负债则是评估客户的偿债能力的重要指标。

而违约作为本研究中的因变量是因为个人消费信用贷款的违约问题直接影响金融机构的资产回收率和资产质量。

1.4 论文的创新与不足

本文采用了 Logistic 回归模型和实证分析方法，应用于个人消费信用贷款违约风险的预测模型搭建和影响因素分析，具有以下创新点：

首先，本研究采用了 Logistic 回归模型作为主要的预测模型，相对于其他常见的预测模型（如决策树、神经网络等），Logistic 回归模型具有参数简单、易解释、计算速度快等优点，同时在二分类问题中表现良好。本研究还将跨学科研究法、实证研究法等多种方法相结合，从不同角度深入探讨个人消费信用贷款违约的风险评估模型及其影响因素，具有较强的可操作性和实用价值。

其次，本研究所选择的变量涵盖了客户个人信息和经济状况等多个方面，例如年龄、教育、工龄、收入、负债率和信用卡负债等，能够全面反映客户的经济状况和还款能力等方面的信息。通过 Logistic 回归模型的建立和实证分析方法的运用，本研究建立了个人消费信用贷款违约的风险预测模型，并对该模型的精度和稳定性进行验证和评估，为金融机构提供有力的决策支持和帮助。

然而，本研究也存在一些不足之处。虽然 Logistic 回归模型是一种较为常见且易操作的预测模型，但现今已经出现了更加先进、准确的模型和技术，例如深度学习等。其次，本研究所使用的数据样本相对较少，仅有 699 个样本数据，可

能对研究结果的偏差有一定影响。因此未来工作可以考虑使用更多、更全面的数据集，采用更先进的模型和技术，以提高预测模型的精度和可靠性。

第二章　问题研究与假设提出

2.1 影响个人消费信用贷款违约风险因素分析

个人消费信用贷款违约的风险受到多个因素的影响，其中较为重要的因素包括客户的个人信息和经济状况这两方面。在本研究中，我们将采用 Logistic 回归模型对这些因素进行分析。

2.1.1 基本信息

1. 年龄：年龄是一个重要的影响因素，年龄越大，违约率越低。这可能与客户的收入水平、职业稳定性、偿债意识等方面有关。
2. 教育：教育水平较高的客户，其违约率通常更低。这可能反映出这部分客户更加注重自己的信用记录，有更强的理财意识，并且对于自己的还款责任有更强的意识和承担能力。
3. 工龄：工龄对于个人消费信用贷款违约风险也有一定的影响。工作稳定性较好的客户，其违约率相对较低。

2.1.2 经济情况

4. 收入：收入水平是评估客户还款能力的一个重要指标，同时也是影响违约率的重要因素。收入越高，违约率越低。
5. 负债率：负债率反映出客户总负债与总资产的比例，是评估客户偿债能力的重要指标。负债率越高，违约率也就越高。
6. 信用卡负债：信用卡负债是客户信用记录中的一个重要组成部分，反映客户的信用情况。信用卡负债越高，个人消费信用贷款违约风险也越高。

综上所述，根据我们的经验，初步选取以上六个变量作为影响因素进行研究。

2.2 变量选择与模型构建

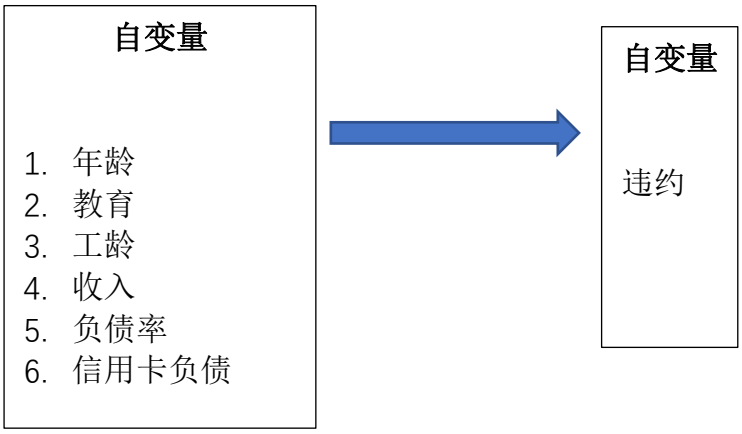


表 2.2 变量的量化说明

序号	变量类型	具体变量	变量量化及方式说明
Y	因变量	是否违约	未违约=0，违约=1
X1	个人信息	年龄	借款人年龄（岁）
X2		学历	初中及以下=1；高中=2；本科/专科=3；硕士=4；硕士以上=5
X3		工龄	工作时间年数（年）
X4	经济状况	收入	借款人年收入（万元）
X5		负债率	借款人目前的负债率
X6		信用卡负债	指信用卡中消费负债（万元）

逻辑回归方程为：

$$logit\big(P(Y=1)\big)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6$$

其中 P(Y=1)表示违约的概率，也就是因变量 Y 等于 1 的概率。logit 函数则是对违约概率的对数转换，使得方程结果可以被限制在 0 到 1 之间。β0、β1、β2、β3、β4、β5 和 β6 分别是模型的系数，它们表示每个自变量对因变量“违约”的影响。

2.3 假设检验

本文主要研究两个问题：

1.年龄、教育、工龄等因素是否与违约率有关系？

2.是否可以根据年龄、收入、负债率、信用卡负债等因素来预测违约？

因此做出如下假设：

表 2-3　假设提出

研究维度	假设	具体内容
个人信息	假设 H1	年龄对违约率没有显著影响
	假设 H2	教育对违约率没有显著影响
	假设 H3	工龄对违约率没有显著影响
经济状况	假设 H4	收入对违约率没有显著影响
	假设 H5	负债率对违约率没有显著影响
	假设 H6	信用卡负债对违约率没有显著影响

第三章 数据分析与预测过程

3.1 数据分析

3.1.1 数据预处理

首先对收集到的数据进行预处理，包括对空缺值和重复值的去除。

年龄	教育	工龄	收入	负债率	信用卡负债	其他负债	违约
41	3	17	176	9.3	11.36	5.01	1
27	1	10	31	17.3	1.36	4	0
40	1	15	55	5.5	0.86	2.17	0
41	1	15	120	2.9	2.66	0.82	0
24	2	2	28	17.3	1.79	3.06	1
41	2	5	25	10.2	0.39	2.16	0
39	1	20	67	30.6	3.83	16.67	0
43	1	12	38	3.6	0.13	1.24	0
24	1	3	19	24.4	1.36	3.28	1
36	1	0	25	19.7	2.78	2.15	0
27	1	0	16	1.7	0.18	0.09	0
25	1	4	23	5.2	0.25	0.94	0
52	1	24	64	10	3.93	2.47	0
37	1	6	29	16.3	1.72	3.01	0
48	1	22	100	9.1	3.7	5.4	0
36	2	9	49	8.6	0.82	3.4	1
36	2	13	41	16.4	2.92	3.81	1
43	1	23	72	7.6	1.18	4.29	0

又因为年龄、教育、工龄、收入、负债率等变量都是数字型，但它们的类型可能不一定都是 float。可能会存在整数型、布尔型、字符串型等不同类型的数据，因此接下来对数据进行了 float 的转化，以确保数据的一致性和准确性。

最终保留正确格式的数据 700 条。

3.1.2 描述统计

为了初步了解各个影响因素的数据分布和差异情况，接下来对数据进行了描述统计分析，结果如下：

	年龄	教育	工龄	收入	负债率	信用卡负债	其他负债	违约
count	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000
mean	34.860000	1.722857	8.388571	45.601429	10.260571	1.553457	3.058229	0.261429
std	7.997342	0.928206	6.658039	36.814226	6.827234	2.117209	3.287524	0.439727
min	20.000000	1.000000	0.000000	14.000000	0.400000	0.010000	0.050000	0.000000
25%	29.000000	1.000000	3.000000	24.000000	5.000000	0.370000	1.047500	0.000000
50%	34.000000	1.000000	7.000000	34.000000	8.600000	0.855000	1.985000	0.000000
75%	40.000000	2.000000	12.000000	55.000000	14.125000	1.905000	3.927500	1.000000
max	56.000000	5.000000	31.000000	446.000000	41.300000	20.560000	27.030000	1.000000

根据表格我们可以得出：

1. 最终分析的样本人数为 700 人。

1）自变量：贷款人个人信息：

2.平均年龄为 34.86 岁，标准差为 7.99 岁，最小值为 20 岁，最大值为 56 岁，我们所收集到的数据看出，几乎都处于劳动年龄范围内；

3.教育程度的平均值为 1.72，标准差为 0.93，表示大多数样本被调查者受过高中教育。

4.工龄的平均值为 8.39 年，标准差为 6.66 年，最小值为 0 年，最大值为 31 年。

2）自变量：贷款人经济状况：

5.收入的平均值为 45.60 千元，标准差为 36.81 千元，最小值为 14 千元，最大值为 446 千元。

6.负债率的平均值为 10.26%，标准差为 6.83%，最小值为 0.4%，最大值为 41.3%。

7.信用卡负债的平均值为 1.55 千元，标准差为 2.12 千元，最小值为 0.01 千元，最大值为 20.56 千元。

3）因变量：

违约指标的平均值为 0.26，标准差为 0.44，最小值为 0，最大值为 1。

其中重点分析标准差：

1.年龄和工龄相对较为分散：年龄和工作年限的标准差分别为 7.997342 和 6.658039，说明被调查者的年龄和工龄有较大的差异。年龄和工作年限可能会影响拖欠行为。

2.收入的波动性很大：收入的标准差为 36.814226，说明被调查者的收入差异非常大，有些人收入很高，而有些人收入很低。这表明收入是一个重要的指标，

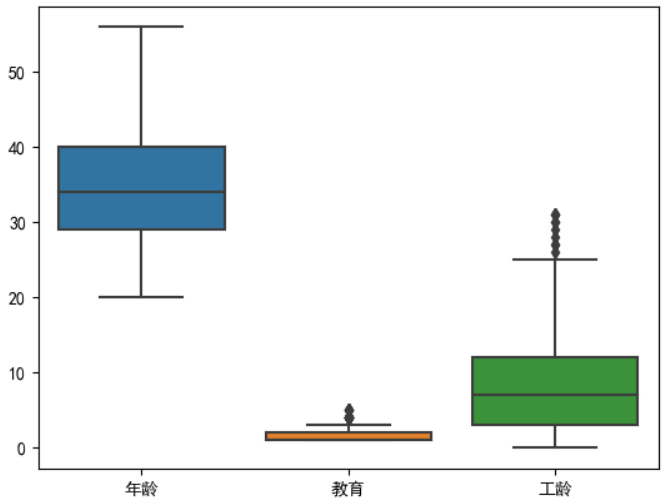
与违约行为可能存在着某种关系。

3.负债率、信用卡负债和其他负债的波动性适中：三个负债指标的标准差分别为 6.827234、2.117209 和 3.287524，说明这些指标在样本中的分布较为均匀，整体变化范围不是特别大，因此这些指标都是比较重要的财务状况指标。

3.1.3 可视化分析

为了更加直观的看到变量的分布和大概范围，接下来又进行了单变量和多变量的可视化分析：

1） 绘制箱线图



从图中可以看出：

a. 年龄值大部分在 35 岁左右，教育程度在 1~2 左右，工龄大部分 8~9 左右，和之前描述统计结果一致。通过图中可以更加直观的发现年龄和工龄的差距较大，并且教育和工龄都有异常值。

b. 因此，接下来考虑是否去除这些异常值。

异常值是指低于下限或高于上限的任何值。通过计算教育数据的四分位距（IQR）来确定异常值的上下限。其中四分位距是将教育数据按大小排序并

教育的异常值：

	年龄	教育	工龄	收入	负债率	信用卡负债	其他负债	违约
25	25.0	4.0	0.0	32.0	17.6	2.14	3.49	0.0
73	43.0	4.0	1.0	26.0	10.6	1.52	1.24	0.0
83	35.0	4.0	4.0	29.0	11.0	1.84	1.35	0.0
122	34.0	4.0	6.0	27.0	35.3	1.98	7.55	1.0
130	26.0	4.0	1.0	27.0	2.9	0.31	0.47	0.0
146	28.0	4.0	1.0	26.0	12.4	0.38	2.85	0.0
147	30.0	4.0	2.0	25.0	10.0	1.77	0.73	0.0
168	41.0	4.0	14.0	44.0	1.7	0.35	0.39	0.0
184	28.0	4.0	0.0	29.0	24.2	1.42	5.59	0.0
197	33.0	4.0	9.0	28.0	4.3	0.38	0.83	0.0
202	48.0	4.0	3.0	45.0	9.8	0.97	3.44	0.0
219	27.0	4.0	0.0	70.0	8.0	1.62	3.98	1.0
227	40.0	4.0	5.0	75.0	1.9	0.88	0.54	0.0
291	33.0	4.0	9.0	32.0	5.5	0.50	1.26	0.0
314	31.0	4.0	1.0	29.0	11.1	1.07	2.15	0.0
317	23.0	4.0	0.0	23.0	6.7	0.47	1.07	0.0

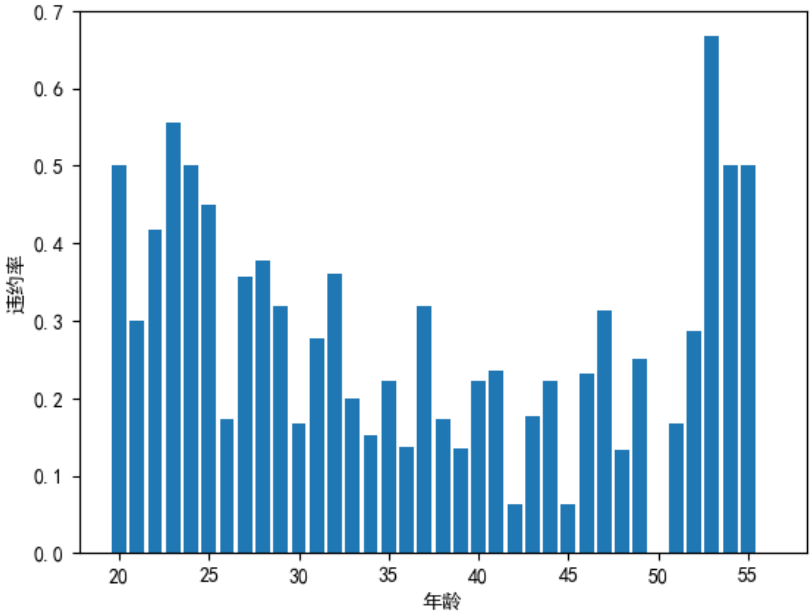
将其分为四个等份，然后计算第一和第三四分位数。上下限被定义为 Q3 + 1.5 * IQR 和 Q1 - 1.5 * IQR，其中 Q1 和 Q3 分别为教育数据的 25%和 75%分位数。最后统计结果 43 个。异常值比例很小（如小于 5%），对于整个结果影响不大，因此不剔除。

工龄类似，只有 6 个异常值，影响也不大，不剔除。

工龄的异常值：

	年龄	教育	工龄	收入	负债率	信用卡负债	其他负债	违约
300	47.0	1.0	29.0	129.0	25.3	20.56	12.08	1.0
528	51.0	2.0	31.0	249.0	7.8	4.27	15.15	0.0
622	48.0	2.0	30.0	148.0	7.2	3.97	6.68	0.0
632	47.0	1.0	31.0	136.0	23.1	14.23	17.18	1.0
675	48.0	1.0	30.0	101.0	6.4	1.87	4.59	0.0
691	47.0	1.0	31.0	253.0	7.2	9.31	8.91	0.0

2） 按照年龄分组，并计算每个年龄段的违约率——条形图



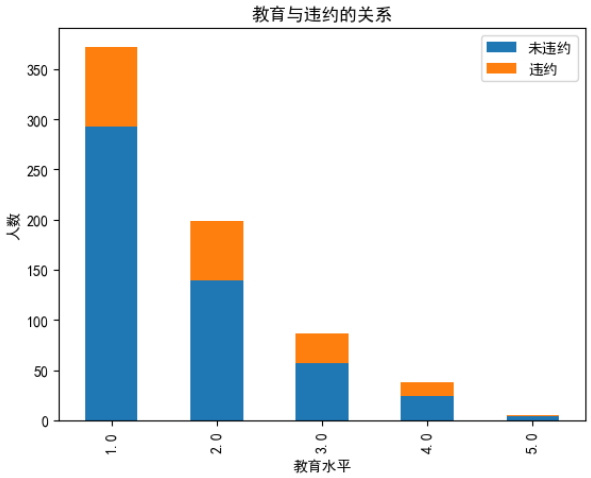
从条形图中看出：20-25 岁和 50-55 岁年龄组的违约率似乎比其他年龄组要高一些，根据经验推测原因：

1.缺乏财务意识：20-25 岁的人通常刚刚开始独立生活，他们可能缺乏足够的财务知识和经验来管理自己的财务，导致无法有效地还款。而 50-55 岁的人则可能由于家庭责任、子女教育等原因，也可能会忽略个人财务规划和管理，从而导致违约。

2.支付能力降低：20-25 岁的人通常处于学习阶段或刚刚步入职场，收入相对较低，同时面临着房租、交通、食品、日常开销等支出压力，很难保证还款能力。而 50-55 岁的人可能已经进入退休或者工资收入减少的状态，在养老金或其他被动收入来源不稳定的情况下，支付能力也可能会降低。

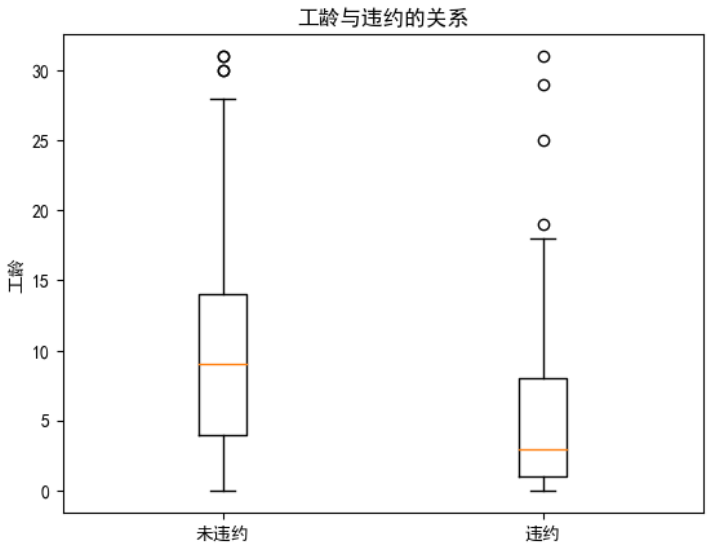
3.生活方式：某些年龄段的人更可能有一种消费倾向，例如在 20-25 岁的年轻人中，可能存在一种“享受当下”的思想，即通过消费来满足自己的需求，而不考虑未来可能出现的经济压力。50-55 岁的人则可能会有一种“补偿式消费”的倾向，即在退休前将生活方式提升到更高的水平，导致消费增加和资金短缺。

3） 按教育水平分组，并计算每种水平的违约人数和非违约人数——堆积柱状图



从图中可以明显发现：学历越低的人，违约越多，违约占比越高

4）工龄与违约的负相关关系：违约的工龄相比未违约的工龄平均较短



3.1.4 相关性检验

接下来对各个因素与因变量“违约“进行相关性分析：

	收入	负债率	信用卡负债	年龄	教育	工龄	违约
收入	1.000000	-0.026777	0.570217	0.478710	0.235190	0.619681	-0.070970
负债率	-0.026777	1.000000	0.501732	0.016398	0.008838	-0.031182	0.389575
信用卡负债	0.570217	0.501732	1.000000	0.295137	0.088245	0.403701	0.244739
年龄	0.478710	0.016398	0.295137	1.000000	0.022325	0.536497	-0.137657
教育	0.235190	0.008838	0.088245	0.022325	1.000000	-0.153621	0.114676
工龄	0.619681	-0.031182	0.403701	0.536497	-0.153621	1.000000	-0.282978
违约	-0.070970	0.389575	0.244739	-0.137657	0.114676	-0.282978	1.000000

a. 收入和违约之间的相关系数为-0.07，表示收入较低的人更容易违约。

b. 工作年限和违约之间的相关系数为-0.28，表明拥有更长工作经历的人可能更少违约。

c. 教育和违约之间的相关系数为 0.11，说明教育程度较高的人较少违约。

3.1.5 Logistic 回归模型

因为我的因变量“违约“是一个二元变量，只有 1 和 0 的取值，因此应该建立逻辑回归模型。

Logit Regression Results

Dep. Variable:	违约	**No. Observations:**	700
Model:	Logit	**Df Residuals:**	693
Method:	MLE	**Df Model:**	6
Date:	Thu, 22 Jun 2023	**Pseudo R-squ.:**	0.2861
Time:	22:28:23	**Log-Likelihood:**	-287.11
converged:	True	**LL-Null:**	-402.18
Covariance Type:	nonrobust	**LLR p-value:**	7.144e-47

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.9097	0.536	-1.697	0.090	-1.960	0.141
年龄	-0.0094	0.015	-0.629	0.529	-0.039	0.020
教育	0.0585	0.120	0.489	0.625	-0.176	0.293
工龄	-0.2299	0.030	-7.714	0.000	-0.288	-0.171
收入	-0.0052	0.006	-0.878	0.380	-0.017	0.006
负债率	0.0815	0.020	4.164	0.000	0.043	0.120
信用卡负债	0.5528	0.100	5.502	0.000	0.356	0.750

在回归模型中，如果是分析影响因素时，R 平方值不是特别重要，只需要重点分析显著性即可。

a. 年龄的系数为-0.0094，p 值为 0.529，大于 0.05 的显著性水平，因此接受原假设 H1，认为年龄对违约率没有显著影响。

b. 教育的系数为 0.0585，p 值为 0.625，大于 0.05 的显著性水平，因此接受原假设 H2，认为教育对违约率没有显著影响；

c. 收入的系数为-0.0052，p 值为 0.380，大于 0.05 的显著性水平，因此接受原假设 H3，说明收入对违约率没有显著影响。

d. 工龄的系数为-0.2299，p 值小于 0.05 的显著性水平，可以拒绝原假设，并认为工龄与违约率之间存在显著负相关关系。

e. 因为负债率的系数为 0.0815，p 值小于 0.05 的显著性水平，可以拒绝原假设 H4，并认为负债率与违约率之间存在显著正相关关系。

f. 信用卡负债对违约率有显著影响。因为信用卡负债的系数为 0.5528，p 值小于 0.05 的显著性水平，可以拒绝原假设，并认为信用卡负债与违约率之间存在显著正相关关系。

反思：这个结果和可视化分析的结果不同，可能是因为可视化分析是假定前提：自变量和因变量之间是线性相关的，而回归模型的假设是非线性相关的。

最终得出假设检验验证：

研究维度	假设	具体内容	验证结果
个人信息	假设 H1	年龄对违约率没有显著影响	支持
	假设 H2	教育对违约率没有显著影响	支持
	假设 H3	工龄对违约率没有显著影响	不支持，为负相关
经济状况	假设 H4	收入对违约率没有显著影响	支持
	假设 H5	负债率对违约率没有显著影响	不支持，为正相关
	假设 H6	信用卡负债对违约率没有显著影响	不支持，为正相关

3.2 预测

接下来研究第二个问题：是否能够根据年龄、收入、负债率、信用卡负债等因素是来预测是否会发生违约？

首要考虑：这六个自变量之间是不是存在多重共线性问题，考虑要删除一个或多个自变量。因此接下来初步检测计算自变量之间的相关性，通过相关系数矩阵可以看到：

	年龄	教育	工龄	收入	负债率	信用卡负债
年龄	1.000000	0.022325	0.536497	0.478710	0.016398	0.295137
教育	0.022325	1.000000	-0.153621	0.235190	0.008838	0.088245
工龄	0.536497	-0.153621	1.000000	0.619681	-0.031182	0.403701
收入	0.478710	0.235190	0.619681	1.000000	-0.026777	0.570217
负债率	0.016398	0.008838	-0.031182	-0.026777	1.000000	0.501732
信用卡负债	0.295137	0.088245	0.403701	0.570217	0.501732	1.000000

年龄和工龄之间的相关系数为 0.54，这意味着它们之间有一个较强的正相关性，即随着一个人的年龄增长，他们的工龄也可能增加。相似地，收入和信用卡负债之间的相关系数为 0.57，这意味着随着收入水平的提高，信用卡负债也会增加。而其他因素相关性可能不是很强。

因为年龄、工龄等多个变量之间存在相关性，这可能会导致模型的过拟合和不稳定性。因此，接下来使用 PCA 来减少数据的维度，提高模型的泛化能力。将数据集转化为 4 个主成分：

```
array([[130.68755028,   9.55288106,   2.57055808,   2.30350604],
       [-15.11891543,   2.97096388,   7.36972539,  -5.89666815],
       [ 10.57905132,  -4.90295876,  -6.00922657,  -3.12522492],
       ...,
       [-12.8989591 ,  -2.31102057,  -3.63527381,  -7.46723379],
       [ 33.32107415,  -8.38093307,  -3.56006851,  -3.52140194],
       [ -0.91827159,  -4.43276593,   3.7189276 ,  -2.7513294 ]])
```

接下来将现有的数据集划分为训练集和测试集，训练集用于训练模型，测试集用于评估模型的泛化能力。这里选择 20%作为测试并选用逻辑回归模型进行训练、在测试集的基础上进行预测（二元变量-违约）。

```
accuracy
```

```
0.8214285714285714
```

得出模型准确率为：82%，较优；

在模型预测之后，本研究还采用了分类报告和混淆矩阵进行对模型的预测效果进行评估。分类报告可以提供模型的精确度、召回率和 F1 得分等指标，以评估模型的预测效果。混淆矩阵可以提供模型的真阳性率、假阳性率、真阴性率和假阴性率等指标，以评估模型的分类效果。最终分类报告和混淆矩阵结果为：

	precision	recall	f1-score	support
0.0	0.82	0.97	0.89	102
1.0	0.85	0.45	0.59	38
accuracy			0.83	140
macro avg	0.84	0.71	0.74	140
weighted avg	0.83	0.83	0.81	140

```
[[99  3]
 [21 17]]
```

1） 分类报告

a. 精确率是指正确预测为正例的样本占所有预测为正例样本的比例，召回率是指正确预测为正例的样本占所有实际为正例样本的比例，F1 得分则是精确率和召回率的调和平均值。

b. 从报告中可以看出：对于标签 0.0（没有违约），模型的精确度、召回率和 F1 得分都较高，说明该类别的预测结果比较准确。

c. 而对于标签 1.0（有违约），模型的精确度和 F1 得分也较高，但召回率较低，说明该类别的预测结果可能会漏报一些违约情况。

2） 混淆矩阵

第一行是标签为 0.0 的样本数量和预测正确的数量，即有 102 个样本中，有 99 个被正确地预测为 0.0（真阴性），而有 3 个被错误地预测为 1.0（假阳性）；第二行表示真实标签为 1.0 的样本数量和预测情况，即有 38 个样本中有 17 个被正确地预测为 1.0（真阳性），而有 21 个被错误地预测为 0.0（假阴性）。从混淆矩阵中可以看出，模型在绝大多数情况下都能够准确地预测标签 0.0，但对于标签 1.0，模型存在一些假阳性和假阴性。

综上，模型总体预测能力较好，可以进行修改与实际应用。

第四章 总结与建议

4.1 总结

本文主要研究个人消费信用贷款违约风险的因素，并使用 Logistic 回归模型进行预测。在第一章中介绍了研究背景和意义，提出了研究方法和国内外文献综述。接着分析了影响个人消费信用贷款违约风险的因素，并提出了假设检验。在现有的数据集下，本研究先对数据集进行了数据预处理，留下 700 条有效样本，接着进行描述统计得到初步的信息，可视化分析、相关性检验和 Logistic 回归模型构建，并对模型进行预测。最后总结了研究结果，并提出了改进建议。

通过研究，我们发现影响个人消费信用贷款违约风险的因素主要包括个人信用情况和个人消费信用贷款情况，在 Logistic 回归模型构建时，我们使用 PCA 降维和训练集测试集划分等技术来优化模型，最终得到准确率为 82%的分类模型。然而，模型在预测有违约情况的召回率还可以进一步提高。因此，在之后的研究中，或许可以考虑增加更多的特征或采用其他算法来进一步提高模型性能。

总之，本研究为了解个人消费信用贷款违约风险的影响因素提供了一种基于 Logistic 回归模型的预测方法。该方法在实践中具有一定的可行性和参考价值，并可为相关金融机构提供决策支持。

4.2 建议

4.2.1 个人消费信用贷款机构的建议：

（1）找准服务定位，促进消费金融助推社会发展 个人消费金融贷款的主要使用者是个人消费和小微企业生产经营，因此大力发展个人消费信用贷款，不仅可以促进社会稳定，而且降低了借贷门槛，满足了个人 和小微企业的融资需求。作为传统银行业务的补充，个人消费信用贷款将大大提高普惠金融的发展，而一直以来个人征信系统之间的壁垒，还在阻碍个人消费信用贷款的发展，而现在个

人消费金融已经不能被忽视，行业只有不断更新个人征信系统的信息，在行业发展的过程中，推动我国个人征信系统向着更完善的方向发展。

（2）拥抱数据和算法，推动良好业态发展 积极促进数据挖掘方法和分析侦测技术的发展与应用，依托互联网技术以及全息数据分析，建立起需求方与金融机构之间的匹配桥梁。通过人工智能降低个人消费贷款违约率，维持行业良好健康发展，为维护金融稳定性添砖加瓦。

4.2.2 对政府监管部门的建议

建立信息整合平台，推动个人征信系统发展目前我国个人征信系统发展并不完善，各个征信系统之间存在信息共享阻碍、信息更新迟缓等等状况未得到根本改善，金融机构无法及时准确识别出借款人的信用状况，由此引发的谨慎审核，阻碍了个人消费信用贷款的发展。对于个人信用贷款行业发展来说，行业监管部门在对违法违规经营相关业务的金融机构进行严厉处罚，给与处罚结果公示，构建清朗的行业经营环境。推动个人征信数据信息流动，使得行业能够整体降低安全风险，降低行业对于违约风险的管理难度，降低个人消费信用贷款申请人的经济负担。

文献综述

[1] Jappelli, T., & Pagano, M. (1989). Consumption and Capital Market Imperfection: An International Comparison. The American Economic Review, 79(5), 1088-1105.

[2] Maki, D. M. (2000). The Growth of Consumer Credit and the Household Debt Service. Retrieved from https://www.bostonfed.org/publications/rdrs/2000/spring/dean-m-maki.aspx

[3] Dinh, T. H. T., & Kleimeier, S. (2007). A credit scoring model for Vietnam's retail banking market. International Review of Financial Analysis, 16(5), 471-495.

[4] Raghuram G. Rajan, Why Bank Credit Policies Fluctuate: A Theory and Some Evidence[J], The Quarterly Journal of Economics, Oxford University Press, 1994, vol. 109(2), pages 399-441.

[5] Lehnert, A. Housing, Consumption, and Credit Constraints. [J] SSRN Electronic Journal.2004.

[6] Aron, J., & Muellbauer, J. Wealth, Credit Conditions, and Consumption: Evidence from South Africa. Review of Income and Wealth[J],2013, 59, S161–S196.

[7] Fagereng, A., & Halvorsen, E. Debt and Household Consumption Responses[J]. SSRN Electronic Journal. 2016.