Hans Journal of Data Mining 数据挖掘, 2021, 11(3), 167-180
Published Online July 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.113015

Hans汉斯

基于网络信贷平台的客户信息挖掘

史佳璐，向永宏，芦俊丽*

云南民族大学数学与计算机科学学院，云南 昆明
Email: 1370816221@qq.com, 2938343077@qq.com, *754490047@qq.com

收稿日期：2021年5月22日；录用日期：2021年6月22日；发布日期：2021年6月30日

摘 要

随着我国网络支付的迅速发展，网络信贷平台在日常生活中初露头角，然而优质客源对网络信贷平台的发展至关重要。本文对某网络信贷平台的数据进行了分析和挖掘。首先，从客户贷款特征数据中对客户价值进行聚类分析，并针对可发展客户给出相应决策。其次，根据客户的借款类型数据，对客户等级进行重新评价，指明提高平台用户质量、数量的方向。最后，利用已认证客户的信息，研究已认证客户喜好的认证方式与客户属性的关系，按客户喜好推荐认证方式以提高客户的认证率，从而增加客户源的稳定性。综上，本文从三个方面进行了研究，对网络信贷平台的发展具有重要意义。

关键词

网络信贷平台，客户价值聚类，客户分群，认证方式推荐

Customer Information Mining Based on Internet Credit Platform

Jialu Shi, Yonghong Xiang, Junli Lu*

School of Mathematics and Computer Science, Yunnan Minzu University, Kunming Yunnan
Email: 1370816221@qq.com, 2938343077@qq.com, *754490047@qq.com

Received: May 22nd, 2021; accepted: Jun. 22nd, 2021; published: Jun. 30th, 2021

Abstract

With the rapid development of online payment in China, online credit platforms are emerging in daily life. However, high-quality customer sources are crucial to the development of online credit platforms. This article analyzes and mines the data of an online credit platform. Firstly, the clus-

*通讯作者。

文章引用: 史佳璐, 向永宏, 芦俊丽. 基于网络信贷平台的客户信息挖掘[J]. 数据挖掘, 2021, 11(3): 167-180.
DOI: 10.12677/hjdm.2021.113015

tering analysis of customer value is carried out from customer loan characteristic data, and the corresponding decision is given for the developing customer. Secondly, according to the type of loan data of customers, the level of customers is reevaluated, and the direction of improving the quality and quantity of platform users is pointed out. Finally, using the information of the certified customers, the relationship between the preferred authentication methods of the certified customers and the attributes of the customers is studied, and the authentication methods are recommended according to the preferences of the customers to improve the certification rate of the customers, so as to increase the stability of the customer source. The objective is to increase the customer's authentication rate and the stability of the customer source. In summary, this article conducts research in three aspects above, which is significant to the development of online credit platforms.

Keywords

Online Credit Platform, Customer Value Clustering, Customer Grouping, Authentication Mode Recommendation

1. 引言

金融是现代经济的核心，而贷款在金融行业中是一个热门业务。因此，贷款公司在金融市场发挥了重要作用。与此同时，随着我国网络支付的迅速发展，网络信贷平台[1]在日常生活中初露头角。因其具有门槛低、成本低等特点，吸引了大量的客户。但是，对于网络信贷平台来说，盲目信贷无形中为公司增加了风险，所以，对客户的信用等级预测变得至关重要。同时，为了使信贷公司的盈利更上一层楼，对贷款客户的价值分类和优先发展客户群的研究是很有必要的。另外，由于众多的网络贷款平台 APP 同时涌现在人们的视线中，采用特定的策略争夺客户对公司的发展具有促进作用。

本文将根据某网络信贷平台的数据，对客户价值进行分类分析，帮助公司决策；基于客户的借款类型数据，对借款类型的客户进行分群，寻找出潜在客户；利用已认证客户的信息，研究已认证客户喜欢的认证方式与客户属性的联系，按客户喜好推荐认证方式以提高客户的认证率。综上所述，本文将从客户价值分类、潜在客户群挖掘、认证方式推荐三个方面进行研究。这一研究对网络信贷平台的发展具有重要意义。

2. 客户价值分类

2.1. 构建贷款客户价值分析关键特征

2.1.1. 构建 TAPNO 模型

基于某网络信贷平台的数据，将客户的历史成功借款次数 T、历史成功借款金额 A、总待还本金 P、历史正常还款期数 N、历史逾期还款期数 O 这 5 个特征值作为网络信贷平台识别客户价值的特征(如表 1 所示)，记为 TAPNO 模型。

2.1.2. 箱线图识别异常值并处理

箱线图[2]，亦称为箱须图，能提供有关数据位置与分散情况的关键信息。箱线图利用数据中的统计量：最小值、下四分位数、中位数、上四分位数和最大值描述数据。

Table 1. Meaning of characteristics
表 1. 特征的含义

模型	网络信贷平台 TAPNO 模型
T	历史成功借款次数
A	历史成功借款金额
P	总待还本金
N	历史正常还款期数
O	历史逾期还款期数

根据 TAPNO 模型，选择与 TAPNO 特征相关的数据，其中不包括历史逾期还款期数，因为该特征在总体基数下，波动较小。因此，对其余 4 个特征利用箱线图识别异常值(如图 1~4)。对于异常值，使用相对应的下四分位数和上四分位数进行填充。

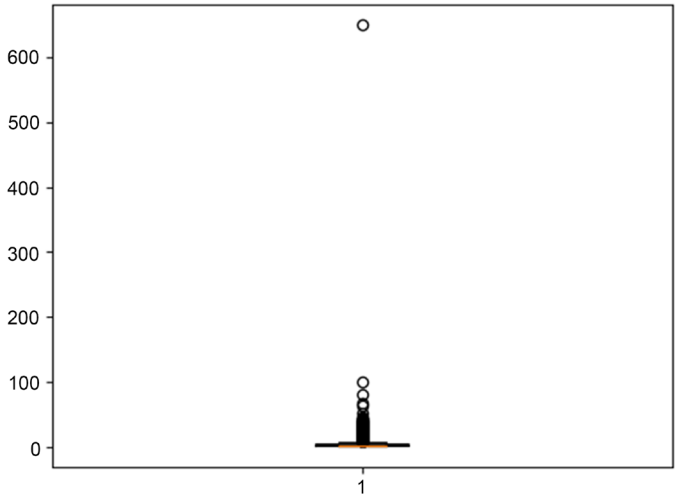

Figure 1. Abnormal data identification on historical successful borrowing counts T
图 1. 历史成功借款次数 T 异常数据识别

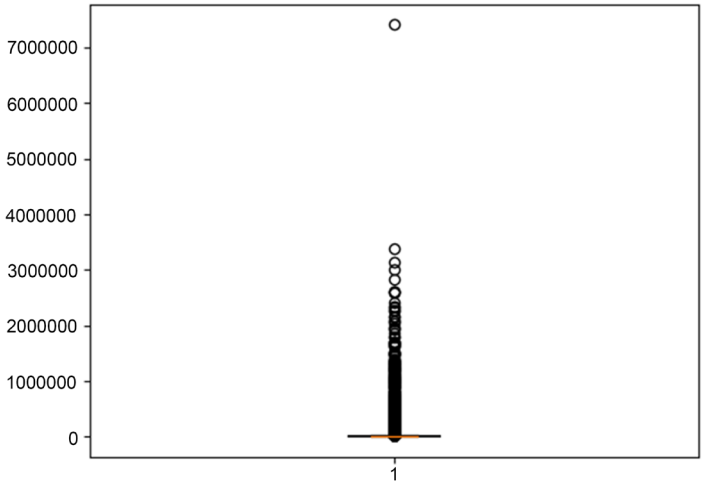

Figure 2. Abnormal data identification on historical successful borrowing amount A
图 2. 历史成功借款金额 A 异常数据识别

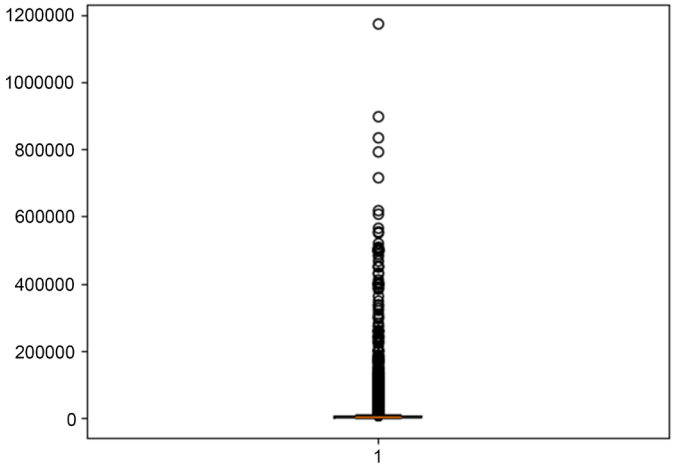

Figure 3. Abnormal data identification on total outstanding principal P
图 3. 总待还本金 P 异常数据识别

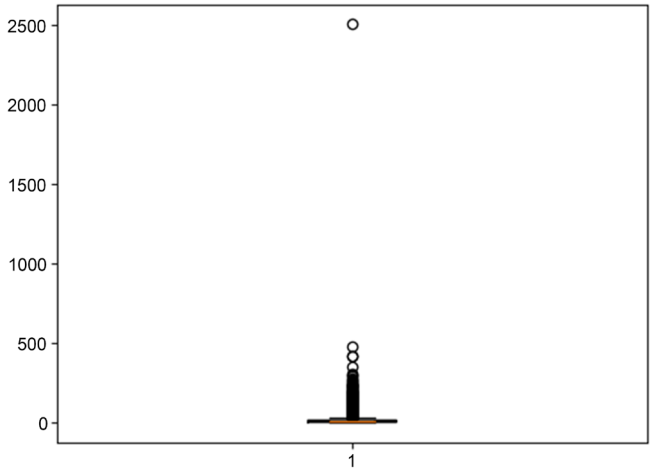

Figure 4. Abnormal data identification on historical normal repayment periods N
图 4. 历史正常还款期数 N 异常数据识别

2.1.3. 标准化 TAPNO 模型的 5 个特征

由于 5 个特征的数据取值范围差异较大(如表 2)，为了消除数量级数据带来的影响，需要对数据进行标准化处理，即标准差标准化[3]。

Table 2. Characteristic range of TAPNO
表 2. TAPNO 特征取值范围

特征名称	T	A	P	N	O
最大值	649	7,405,926.0	1,172,652.87	2507	60
最小值	9	26,722.0	13,120.68	37	0

经过标准差标准化处理的数据均值为 0，标准差为 1。其中 $\bar{X}$ 为原始数据的均值，δ 为原始数据的标准差，公式如下：

$$X* = \frac{X - \bar{X}}{\delta}$$

2.2. 使用 K-Means 算法对客户聚类

2.2.1. K-Means 算法原理

K-Means 算法[4]是一种最经典也是使用最广泛的聚类方法。K-Means 的思想很简单：对于一个聚类任务，首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心[5]；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度，分别将它们分配给与其最相似的聚类；然后再计算每个所获新聚类的聚类中心；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

2.2.2. 确定聚类结果

对数据进行聚类(如图 5)，其中，将所有客户聚为 4 个客户群，customer[0]为客户群 1，customer[1]为客户群 2，customer[2]为客户群 3，customer[3]为客户群 4，values 为每个客户群的聚类中心。

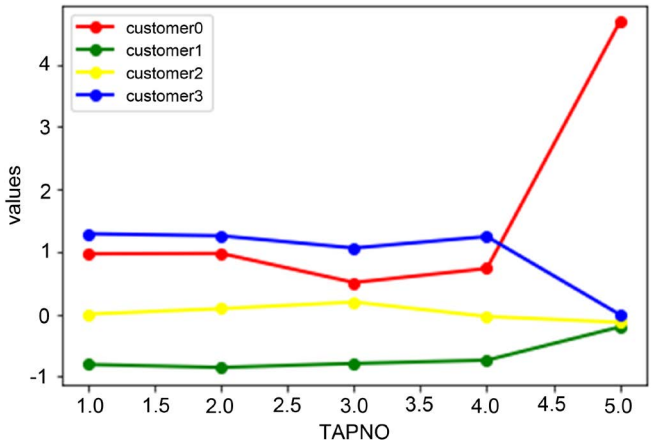

Figure 5. Customer clustering results
图 5. 客户聚类结果

2.2.3. 分析聚类结果

对聚类结果进行特征分析，如图 6~10 雷达图所示。客户群 1 的 5 个特征均较大，其中历史逾期还款

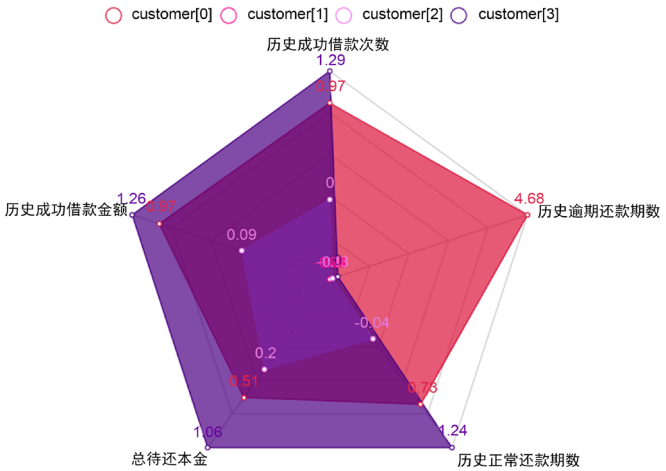

Figure 6. Radar chart of customer group characteristics analysis
图 6. 客户群特征分析雷达图

次数最多；客户群 2 的 5 个特征均最小；客户群 3 除历史逾期还款次数较少外，其余 4 个特征均处于中等水平；客户群 4 的历史逾期还款次数非常少，其余 4 个特征均最大。

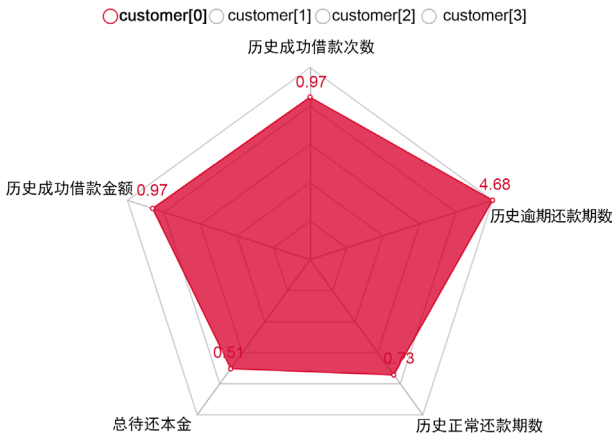

Figure 7. Radar chart of customer group 1 characteristic
图 7. 客户群 1 特征雷达图

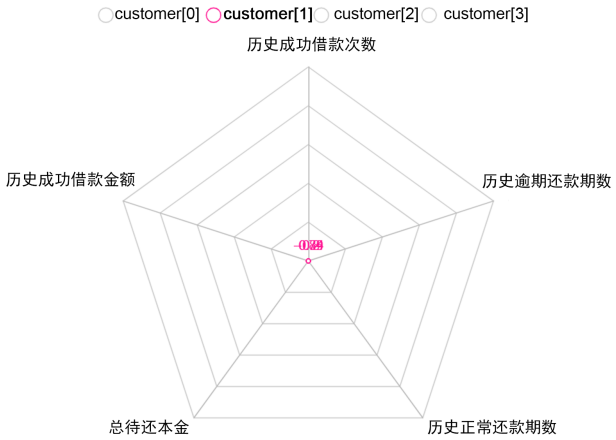

Figure 8. Radar chart of customer group 2 characteristic
图 8. 客户群 2 特征雷达图

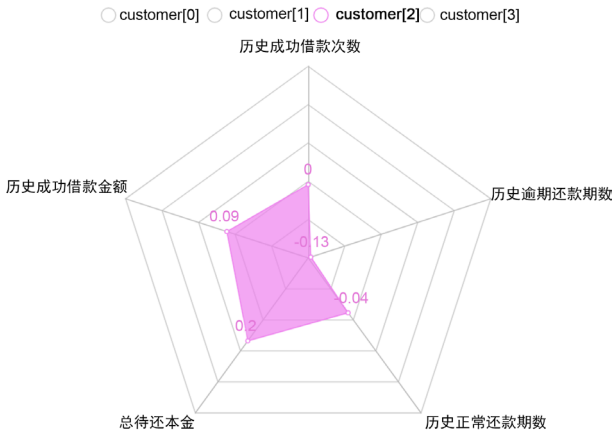

Figure 9. Radar chart of customer group 3 characteristic
图 9. 客户群 3 特征雷达图

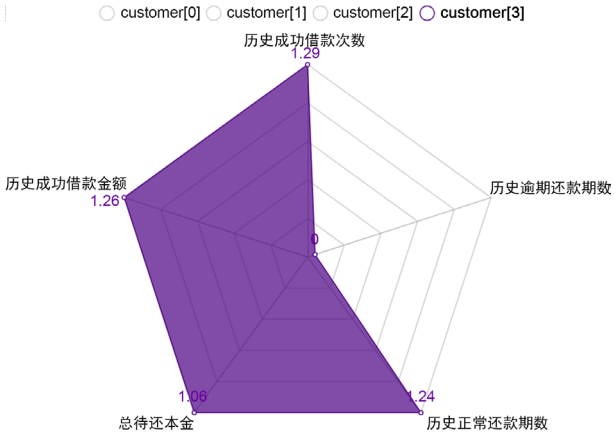

Figure 10. Radar chart of customer group 4 characteristic
图 10. 客户群 4 特征雷达图

综合业务分析，客户群 1 属于可发展客户，该客户其他方面均比较优秀，但是逾期情况较多，调整相关策略后可上升空间大；客户群 2 属于低价值客户，各方面都非常平庸；客户群 3 属于一般客户，虽然其他方面平庸，但逾期次数少；客户群 4 属于高价值客户，逾期次数少，同时经常借款，而且数额大。

因此，客户群价值排名如下：客户群 4——高价值用户；客户群 1——可发展用户；客户群 3——一般用户；客户群 2——低价值用户。

2.3. 可发展客户分析

对比可发展用户和高价值用户的特征分析，可见，尽管可发展客户的历史成功借款次数、历史成功借款金额、总待还本金、历史正常还款期数等特征较高价值用户略有不足，但差距不大。然而，可发展用户的逾期还款次数和高价值用户相比，差距非常大。因此，对于可发展用户，可采取措施降低该类客户的逾期次数，例如：当接近可发展用户的贷款期限时间时，便向此类用户推送提醒信息。

3. 潜在客户群挖掘

3.1. 初始评级分析

由于在客户首次贷款时，该网络信贷平台便对该客户的价值进行了初始评级[6]，评级分布如下(图 11)，

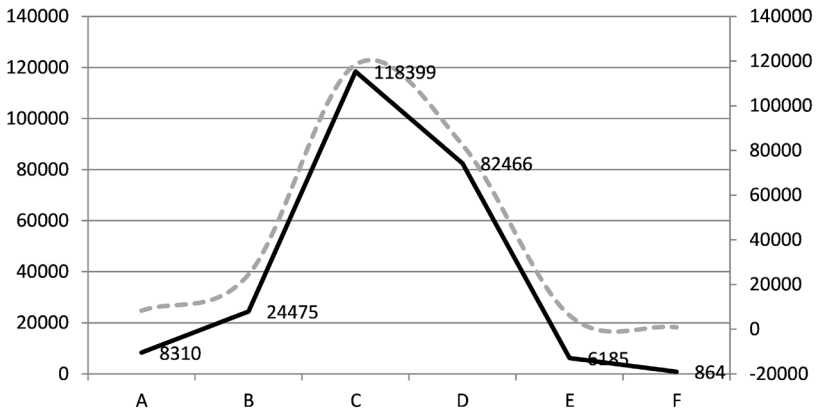

Figure 11. Initial rating distribution
图 11. 初始评级分布

基本呈现正态分布[7]。其中 C 和 D 等级人数较高，客户评级中等，评级最高和最低的人数均较少。考虑到该平台的发展，结合借贷类型，希望某类借贷类型的客户评级基本处于 C、D 等级，因为该类客户有向 A 和 B 等级转换的潜力。

3.2. 客户借贷类型的初始评级分析

根据客户的借贷类型，统计不同借贷类型中各个评级的数量(如图 12)，并计算每个类型各个评级的人数在该类型客户中的占比情况(如表 3)。从整体上看，借贷类型为电商的客户人数较少。

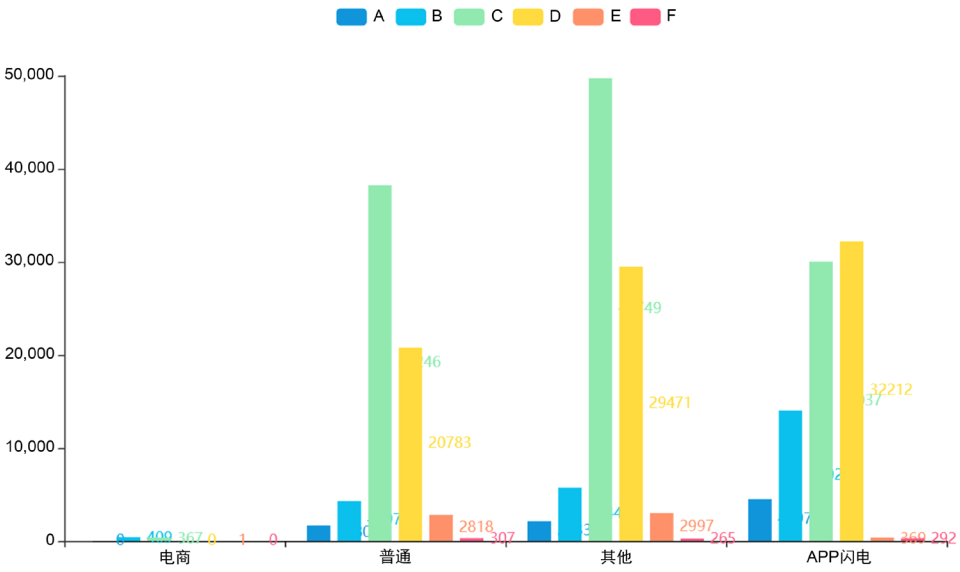

Figure 12. Distribution of ratings for different customer types
图 12. 不同客户类型的评级等级分布图

Table 3. Table of the number of people of different types and ratings
表 3. 不同类型各个评级的人数情况表

贷款类型 \ 初始评级	A	B	C	D	E	F
电商	0 (0.000)	409 (0.526)	367 (0.472)	0 (0.000)	1 (0.001)	0 (0.000)
普通	1680 (0.025)	4297 (0.063)	38246 (0.561)	20783 (0.305)	2818 (0.041)	307 (0.005)
其他	2133 (0.024)	5744 (0.064)	49749 (0.551)	29471 (0.326)	2997 (0.033)	265 (0.003)
闪电 APP	4497 (0.055)	14025 (0.172)	30037 (0.369)	32212 (0.396)	369 (0.005)	292 (0.004)

根据每个贷款类型中各个评级所占的比率，可视化出每种贷款类型各个评级分布比例图(图 13)，分析该图初步得出以下结论：

1) 电商客户大致分布在 B 和 C 等级，质量较高，但因其人数较少，因此可以通过某种方式吸引电商客户群体在该平台贷款。

2) 闪电 APP 客户大致分布在 C 和 D 等级，而且闪电 APP 客户人数较多，因此，该类客户群体属于最具潜力的客户群体，可以优先发展该类客户向更高等级提升。

3) 普通和其他客户群分布属于正态分布，且最多价值评级在 C 等级，符合正常情况。

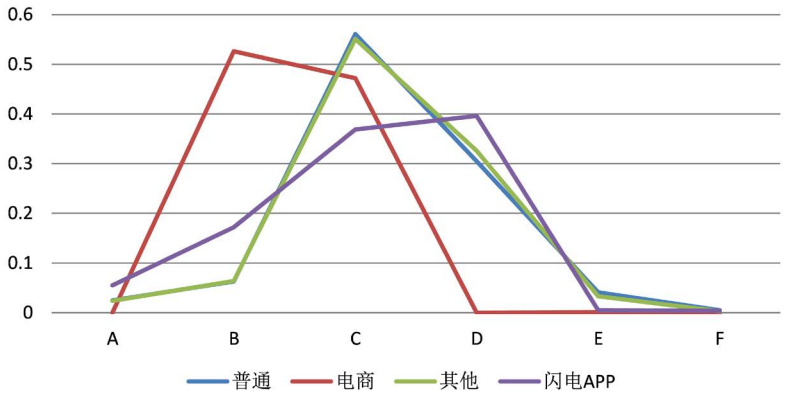

Figure 13. Distribution ratio of each rating for each loan type
图 13. 每种贷款类型各个评级分布比例图

3.3. 当前评级分析

3.3.1. 建立线性回归模型

随着客户在该平台贷款次数的增多，该客户的行为逐渐影响着其价值等级的变化，而客户的行为体现在 TAPNO 模型的 5 个特征和借款金额、借款期限、借款利率共 8 个特征中。因此，利用这 8 个特征并采用线性回归[8]的方式，可预测每位客户的当前价值评级。

回归是估计输入数据与连续值输出数据之间关系的过程。线性回归对一个或多个自变量和因变量之间的线性关系进行建模，要求实际输出与线性方程预测的输出的残差平方和最小，其计算公式如下，其中，θ 是线性回归直线的系数参数：

$$h_\theta(x)=\sum_{i=0}^{n}\theta_i x_i$$

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}\left(h_\theta\left(x^i\right)-y^i\right)^2$$

3.3.2. 线性回归模型结果分析

根据回归模型，数据集由 TAPNO 模型的 5 个特征和借款金额、借款期限、借款利率、初始评级构

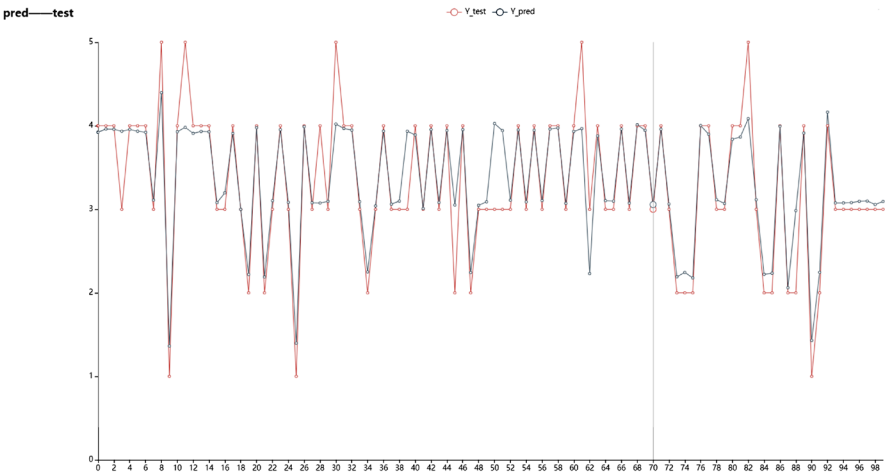

Figure 14. Comparison of current and initial rating (partial)
图 14. 当前评级与初始评级对比图(部分)

成，随机选取数据集的 80%作为训练数据集，接下来利用训练数据集训练线性回归器[9]，之后用模型对所有数据进行预测，最终得到当前价值评级与初始评级对比图(图 14，只展示部分)。

3.4. 客户借贷类型的当前评级分析

根据预测得到的每个客户的当前评级数据，统计预测后不同借贷类型中各个评级的数量，并计算每个类型各个评级的人数在该类型客户中的占比情况(如表 4)。

Table 4. Table of the number of people of different types and ratings after prediction
表 4. 预测后不同类型各个评级的人数情况表

初始评级 / 贷款类型	A	B	C	D	E	F
电商	181 (0.233)	475 (0.611)	103 (0.133)	7 (0.009)	2 (0.003)	9 (0.012)
普通	1904 (0.028)	8555 (0.126)	33697 (0.495)	23562 (0.346)	342 (0.005)	71 (0.001)
其他	2152 (0.024)	7268 (0.080)	49464 (0.547)	30281 (0.326)	1176 (0.013)	18 (0.000)
闪电 APP	4486 (0.055)	14143 (0.174)	30078 (0.369)	32305 (0.397)	420 (0.005)	0 (0.000)

根据每个贷款类型中各个当前评级所占的比率，可视化出每种贷款类型当前评级分布比例图(图 15)，分析该图得出以下结论：

1) 随着客户行为的增加，电商客户群的当前价值评级明显提高，从 B 和 C 等级向 B 靠拢，说明电商客户属于优质客户源。

2) 闪电 APP 客户群的价值仍大部分分布在 C 和 D 等级，然而，有向 D 等级偏移的趋势，该趋势与初步结论有差距。所以，应对闪电 APP 客户群的行为做出针对性措施，例如，通过预计还款时间消息推送，减少该客户群的预期还款次数等。但是客户群仍属于有潜力客户的性质没有变，因此，网络信贷平台针对做出相应的管控举措。

3) 普通客户群中初始评级在 E 等级的客户向 D 靠拢，说明普通客户属于可提升发展客户源。

4) 其他客户群与初步结论相符。

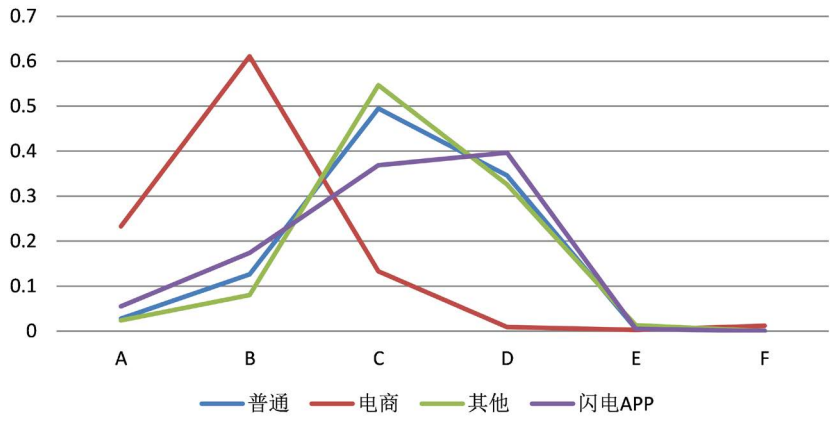

Figure 15. Distribution ratio of current ratings for each loan type
图 15. 每种贷款类型当前各个评级分布比例图

4. 认证方式推荐

4.1. 认证比例分析

尽管网络借贷平台逐渐出现在日常生活中，但根据该平台借贷人员中已认证人数与未认证人数的比例可视化结果(图 16)表明，该网络借贷平台已认证人数和未认证人数之比接近 6:4。此比例证明该平台上未认证客户仍有许多，即五分之二的客户属于临时客户，甚至有丢失此类客户的风险。为增加平台客户稳定性，则希望认证用户的数量越多越好。

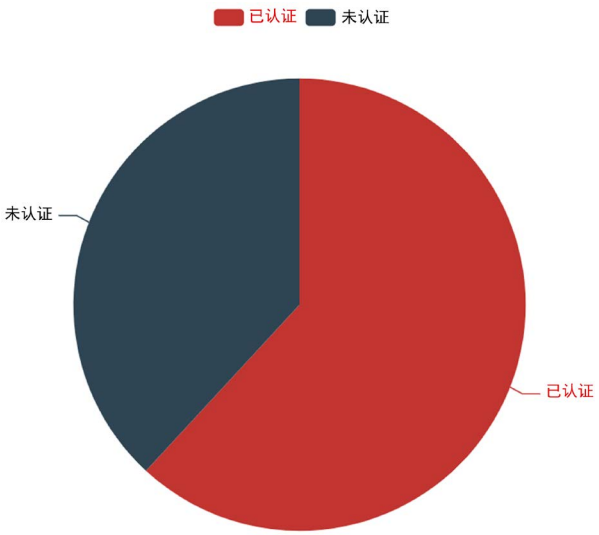

Figure 16. The proportion of people who have been certified and not certified
图 16. 已认证与未认证人数比例分布图

4.2. 认证方式与借贷人员属性联系分析

基于某网络信贷平台的数据，借贷人员属性包括年龄和性别。其中，将年龄按照年龄段划分为 4 个阶段，划分依据参考《信用卡年轻消费群体洞察报告》[10]，即 18~24 岁、25~35 岁、36~46 岁、47~56 岁。另外，该平台的认证方式共有六种：手机认证、户口认证、视频认证、学历认证、征信认证和淘宝认证。

分别计算不同年龄段不同性别的客户选择各种认证方式的数量，数量多极为喜欢，数量低则不喜欢。可视化计算结果如下图 17~20 所示，数据展示：

1) 18~24 岁的男性客户更倾向于手机认证，其次喜欢学历认证；该年龄段的女性客户优选学历认证，其次是手机认证；无论性别，该年龄段的客户选择淘宝认证的人数非常少。

2) 25~35 岁的男性客户优选手机认证，其次喜欢学历认证；该年龄段的女性客户选择学历认证和手机认证的人数差距不大；无论性别，该年龄段的客户亦选择淘宝认证的人数非常少。

3) 36~46 岁的男性和女性客户均更喜欢手机认证；无论性别，该年龄段的客户选择学历认证的人数明显下降。

4) 47~56 岁的男性和女性客户均非常喜欢手机认证；无论性别，该年龄段的客户选择其他认证的人数都很少。

5) 无论什么年龄段最受欢迎的认证方式是手机认证，最不受欢迎的认证方式为淘宝认证。

6) 综合来看，已认证人数中，男性多于女性。

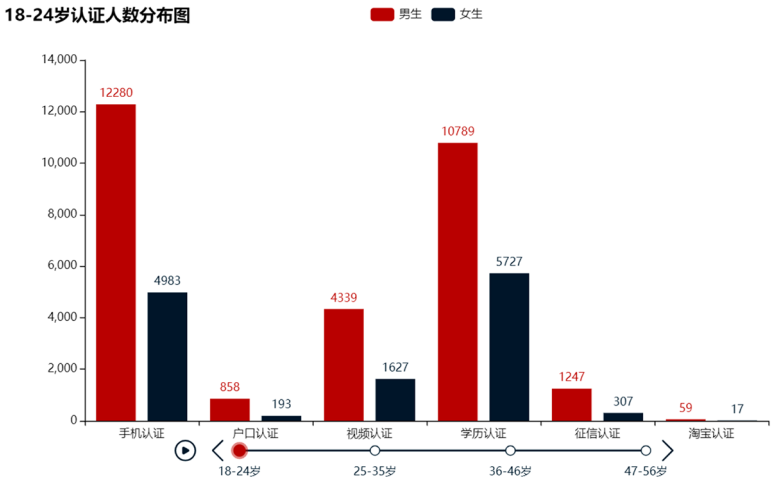

Figure 17. Distribution of people aged 18~24
图 17. 18~24 岁人数分布图

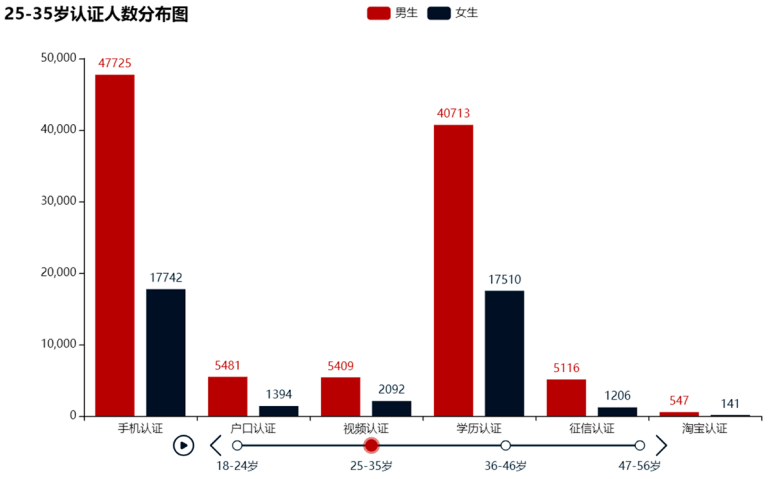

Figure 18. Distribution of people aged 25~35
图 18. 25~35 岁人数分布图

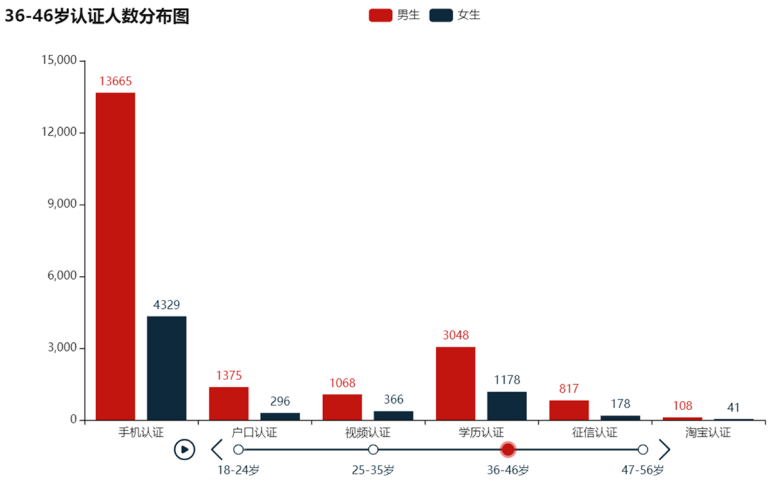

Figure 19. Distribution of people aged 36~46
图 19. 36~46 岁人数分布图

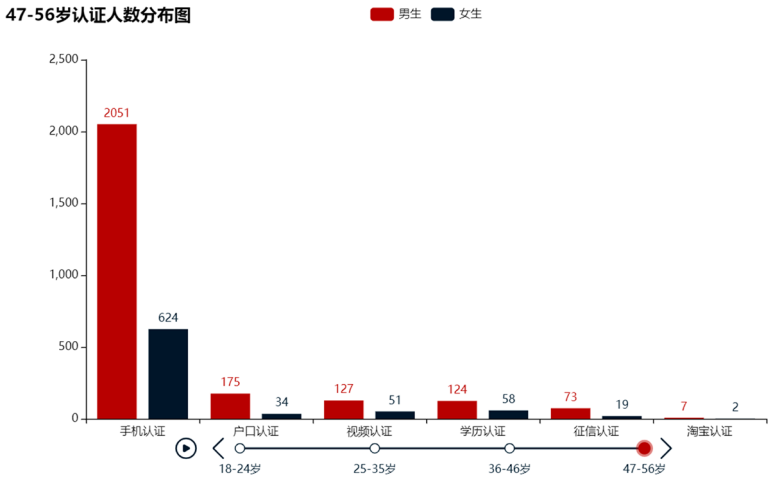

Figure 20. Distribution of people aged 47~56
图 20. 47~56 岁人数分布图

4.3. 认证方式推荐

根据认证方式与借贷人员属性分析，将根据不同年龄段、性别对未认证和未注册人员采用不同的认证方式推荐。

针对已首次借贷却未认证人员：

1) 若该人员属于 18~24 岁的男性客户则依次轮流推送手机认证、学历认证信息；若属于该年龄段的女性客户则依次轮流推送学历认证、手机认证信息。

2) 若该人员属于 25~35 岁的男性客户则依次轮流推送手机认证、学历认证信息；若属于该年龄段的女性客户则同时推送学历认证和手机认证信息。

3) 若该人员属于 36~56 岁的男性和女性客户则推送手机认证信息。

4) 由于选择淘宝认证的人数较少，因此可以考虑移除淘宝认证方法。

针对未首次借贷得潜在认证人员：

1) 在该类人员注册该平台时便推荐请求认证，认证方式均在注册页面。区别在于针对不同年龄段不同性别的客户，认证方式的顺序有所不同。考虑到认证方式的多样性和人类的情绪化，在注册界面推荐的认证方式不多于三个。

2) 若该类人员属于 18~24 岁的男性客户则在注册认证界面依次推送手机认证、学历认证和视频认证；若属于该年龄段的女性客户则依次轮流推送学历认证、手机认证和视频认证。

3) 若该类人员属于 25~46 岁的男性客户则在注册认证界面依次推送手机认证、学历认证和户口认证；若属于该年龄段的女性客户则依次轮流推送学历认证、手机认证和视频认证。

4) 若该类人员属于 47~56 岁的男性客户则在注册认证界面依次推送手机认证、户口认证和视频认证；若属于该年龄段的女性客户则依次轮流推送手机认证、学历认证和视频认证。

5) 若该类人员在注册认证界面均未选择任何一种认证方法，处理方式则与已首次借贷却未认证人员相同。

5. 结语

通过不同角度对某网络信贷平台的客户进行分析挖掘，对该平台的发展提供了进一步的方向，推动了网络信贷平台的普及和发展，促进了互联网金融[11]的应用。另一方面，由于借贷人员通过个人信誉借

贷，有助于体现借贷人员自身的信用价值，对建设社会信用体系有一定的促进作用。

本文研究的主要成果结论与建议如下：

1) 根据客户的历史成功借款次数 T、历史成功借款金额 A、总待还本金 P、历史正常还款期数 N、历史逾期还款期数 O 建立 TAPNO 模型，根据模型将客户聚类为高价值用户、可发展用户、一般用户、低价值用户，其中可发展用户为重点研究用户。针对可发展客户，可降低其逾期贷款次数，从而促进其向高价值客户发展。

2) 由于电商客户的初始评级大致分布在 B 和 C 等级，质量较高。随着该类客户贷款行为的改变，电商客户群的当前价值评级明显提高，从 B 和 C 等级向 B 靠拢。但是因其贷款人数较少，说明电商客户属于优质客户源。因此，可以通过某种方式吸引大量电商客户群体在该平台贷款，有利于提高平台的客户源的质量和数量。

3) 由于闪电 APP 客户的初始评级大致分布在 C 和 D 等级，而且闪电 APP 客户人数较多，因此，该类客户群体属于最具潜力的客户群体。但是，随着该类客户行为的增加，该类客户群有从 C 和 D 等级向 D 等级偏移的趋势。因此，针对该类客户群可在完整的监督举措管控下，优先发展该类客户向更高等级提升。

4) 普通客户群中初始评级在 E 等级的客户向 D 靠拢，且人数较多，说明普通客户属于可提升发展客户源。

5) 客户选择的认证方式与借贷人员属性的年龄段和性别有联系，不同年龄段和性别喜欢的认证方式不同。可以根据此结论分情况选择不同的认证方式推荐，提高客户的认证率，达到稳定客户源的目的。

参考文献

[1] 张文瑶. 中国电子商务平台小额信贷颠覆性创新研究[D]: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2018.

[2] 黄红梅, 张良均. Python 数据分析与应用[M]. 北京: 人民邮电出版社, 2018: 71-72.

[3] 余本国. Python 编程与数据分析应用(微课版) [M]. 北京: 人民邮电出版社, 2020: 163.

[4] 王振武. 数据挖掘算法原理与实现[M]. 第 2 版. 北京: 清华大学出版社, 2017: 159-160.

[5] 田腾浩. 优化初始聚类中心的 K-Means 算法[J]. 网络安全技术与应用, 2014(9): 42-43.

[6] 360 百科. 客户信用评级[EB/OL]. https://baike.so.com/doc/9066169-9397326.html, 2021-03-14.

[7] 盛骤, 谢式千, 潘承毅. 概率论与数理统计[M]. 第 4 版. 北京: 高等教育出版社, 2019: 46-50.

[8] Prateek Joshi, 主编. Python 机器学习经典实例[M]. 陶俊杰, 陈小莉, 译. 北京: 人民邮电出版社, 2017: 6-7.

[9] 简书. 线性回归器[EB/OL]. https://www.jianshu.com/p/4bfba8d0c2cf, 2017-11-18.

[10] CSDN. 信用卡年轻消费群体数据分析和洞察报告[EB/OL]. https://blog.csdn.net/yuanziok/article/details/73232146?utmsource=app, 2017-06-14.

[11] 迟春静. 互联网金融的机遇与挑战[J]. 科技创业月刊, 2019, 32(2): 42-44.