基于 Logistic 回归的信用评分模型

1. 数据集准备：

使用 GiveMeSomeCredit 数据集，共包含 150000 条数据，包含如下特征。使用时按照 7(105,000)： 3(45000)比例随机划分训练集和测试集(**原始数据集/cs-trainin.csv**)。

变量名	描述	类型
SeriousDlqin2yrs(目标值)	是否有超过 90 天或更长时间逾期未还的不良行为	**0(好)，1(坏)**
RevolvingUtilizationOfUnsecuredLines	可用额度比值	百分数
age	年龄	整数
NumberOfTime30-59DaysPastDueNotWorse	逾期 30-59 天笔数	整数
DebtRatio	还款率(每月偿还债务，赡养费，生活费用).	百分比
MonthlyIncome	月收入	实数
NumberOfOpenCreditLinesAndLoans	信贷数量	整数
NumberOfTimes90DaysLate	逾期 90 天笔数	整数
NumberRealEstateLoansOrLines	抵押贷款和房地产贷款	整数
NumberOfTime60-89DaysPastDueNotWorse	逾期 60-89 天笔数	整数
NumberOfDependents	家属数量	整数

表 1 数据集特征

1.1 缺失值统计与处理

训练集缺失值(**清洗后的数据**/missing_values.xlsx)

	count	ratio
SeriousDlqin2yrs	0	0
RevolvingUtilizationOfUnsecuredLines	0	0
age	0	0
NumberOfTime30-59DaysPastDueNotWorse	0	0
DebtRatio	0	0
MonthlyIncome	20729	19.74190476
NumberOfOpenCreditLinesAndLoans	0	0
NumberOfTimes90DaysLate	0	0
NumberRealEstateLoansOrLines	0	0
NumberOfTime60-89DaysPastDueNotWorse	0	0
NumberOfDependents	2734	2.603809524

表 2 训练集缺失值统计

测试集缺失值(**清洗后的数据**/missing_values.xlsx)

	count	ratio
SeriousDlqin2yrs	0	0
RevolvingUtilizationOfUnsecuredLines	0	0
age	0	0
NumberOfTime30-59DaysPastDueNotWorse	0	0
DebtRatio	0	0
MonthlyIncome	9002	20.00444444
NumberOfOpenCreditLinesAndLoans	0	0
NumberOfTimes90DaysLate	0	0
NumberRealEstateLoansOrLines	0	0
NumberOfTime60-89DaysPastDueNotWorse	0	0
NumberOfDependents	1190	2.644444444

表 3 测试集缺失值统计

此表为家属数量为空时，月收入和负债率的描述统计(**清洗后的数据/亲属数量为空**.xlsx)

	MonthlyIncome	DebtRatio
count	**0**	**2734**
mean		**1114.678493**
std		**4891.827226**
min		**0**
25%		**21**
50%		**347.5**
75%		**1556.75**
max		**220516**

表 4 家属数量为空

此表为月收入为空时，家属数量和负债率的基本统计(**清洗后的数据/月收入为空**.xlsx)

	NumberOfDependents	DebtRatio
count	**17995**	**20729**
mean	**0.313031398**	**1659.092431**
std	**0.807038551**	**3583.971493**
min	**0**	**0**
25%	**0**	**125**
50%	**0**	**1160**
75%	**0**	**2382**
max	**8**	**307001**

表 5 月收入为空

通过观察上述两张表，发现
(1) 月收入为空的样本通常有较高负债
(2) 家属数量为空的样本与月收入为空的样本基本是同一批样本
因此可以将两个特征的缺失值填入 0.

1.2 数据集不平衡

正负样本频数：

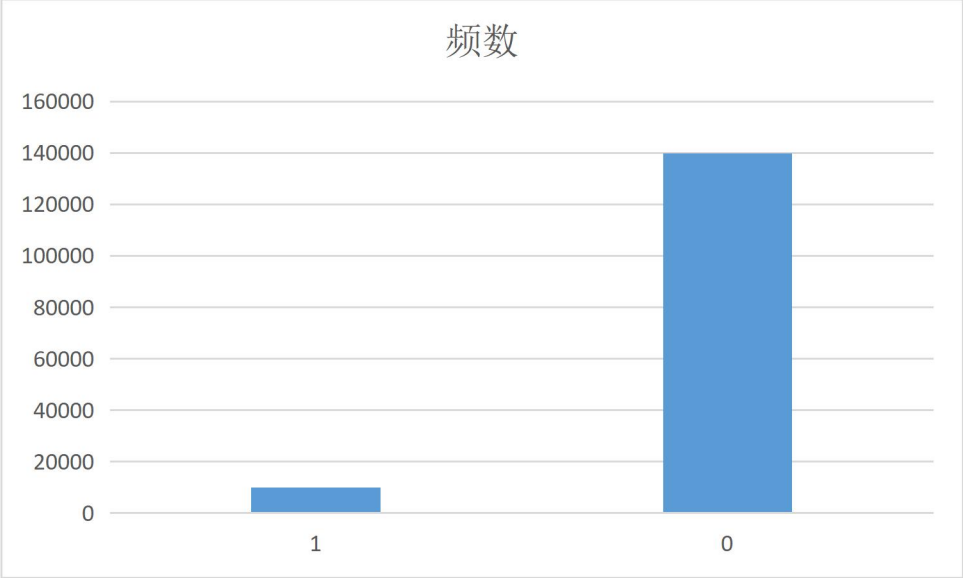

图 1 正负样本频数

不难发现，正样本和负样本极不平衡(1:14)。

2. 单因素分析

2.1 特征: RevolvingUtilizationOfUnsecuredLines

该特征总体数据分布

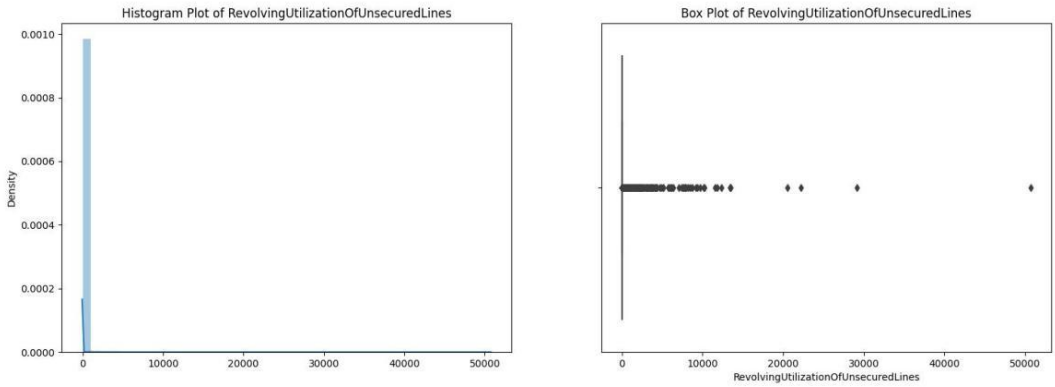

图 2 RevolvingUtilizationOfUnsecuredLines 的直方图和盒子图

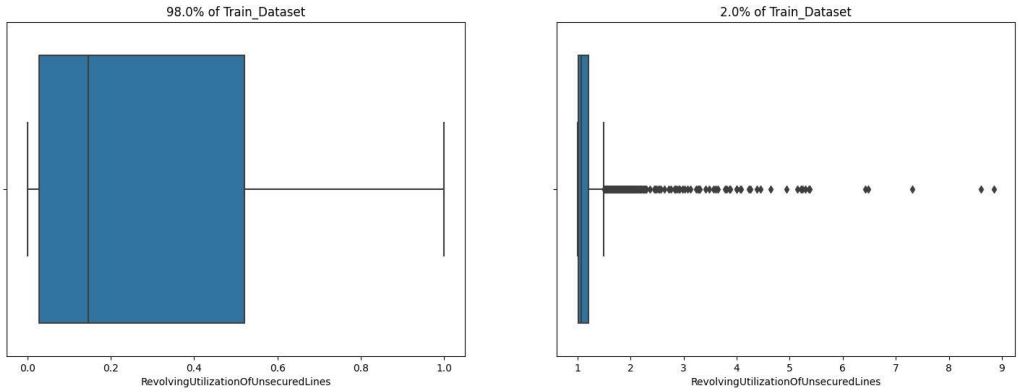

图 3 前 2%和前 98%的样本的盒子图

不难观察到 98%的值在 0 和 1 之间，且偏度靠左。此外 2%的值大于 1，推测可能是部分样本借款超过额度。最后，约 0.5%的值大于 10，这些值会被舍弃。

2.2 特征: DebtRatio

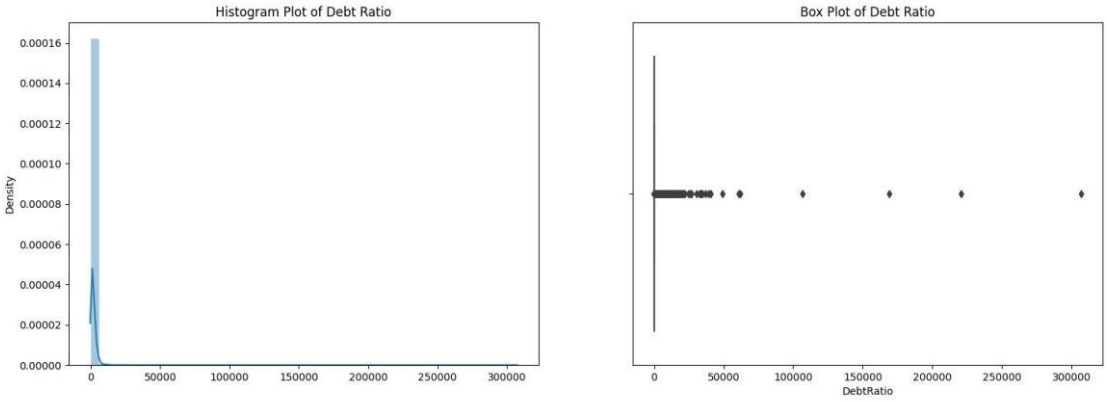

图 4 DebtRatio 的直方图和盒子图

DebtRatio 的描述统计(**单因素分析/DebtRatio/描述统计.xlsx**)

	DebtRatio	DebtRatio(1-10)	DebtRatio(>10)
count	**104834**	**4439**	**8**
mean	**348.8024786**	**2.528377365**	**42103.11326**
std	**1760.067961**	**2.161489571**	**107236.4486**
min	**0**	**1.000527426**	**10.13861386**
25%	**0.175224159**	**1.177700829**	**1097.75**
50%	**0.367014171**	**1.529646902**	**2159.819438**
75%	**0.869411321**	**3**	**7781.346462**
max	**307001**	**10**	**307001**

表 6 数据分布的描述统计

DebtRatio 的大致分布(**单因素分析/DebtRatio/分布.xlsx**)

	below 1	between 1 - 10	beyond 10
1	76.58393269	4.234313295	19.18175401

可以发现 76%的值在 0-1 之间，4%的值在 1-10 之间，其余的 20%的值相当高。

2.3 Age

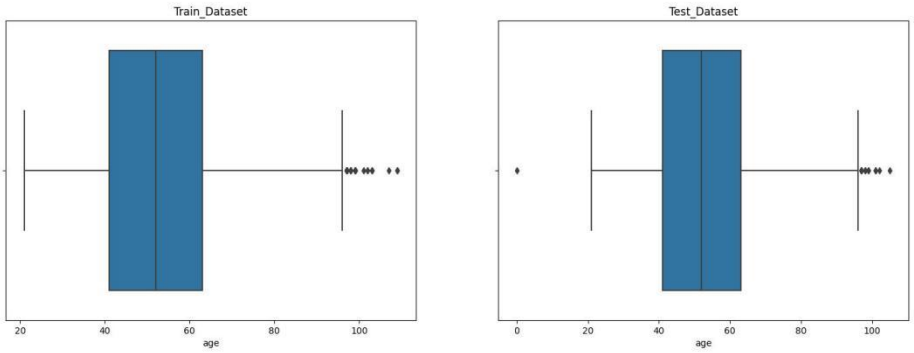

图 5 Age 盒子图

可以发现有一些 0 值，显然不合业务逻辑，将其替换为最小值 21 岁。

2.4 NumberOfOpenCreditLinesAndLoans

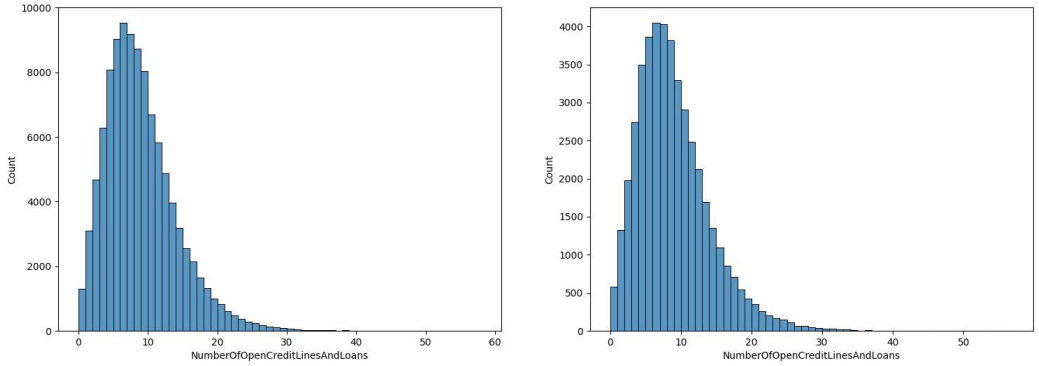

图 6 信贷数量直方图

可以发现该特征没有异常值。

2.5 NumberRealEstateLoansOrLines

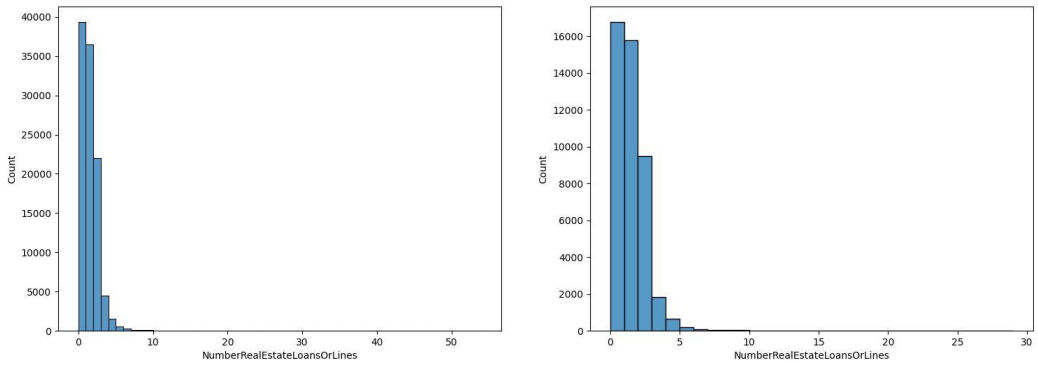

图 7 抵押贷款和房地产贷款直方图

该特征向右偏移，无异常值。

2.6 NumberOfDependents

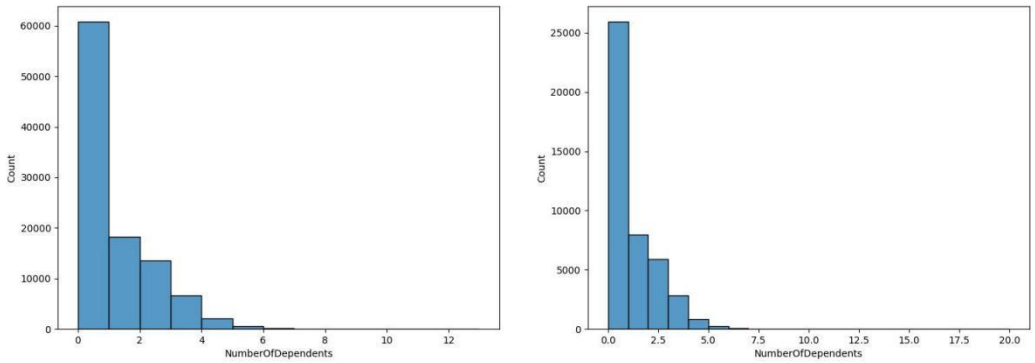

图 8 亲属数量直方图

无异常值。

2.7 NumberOfDaysPastDueNotWorse

该表在(**单因素分析/PastDue/逾期笔数的样本统计.xlsx**)

	30-59days	60-89days	90days
0	87993	99526	98974
1	11212	4006	3683
2	3258	779	1087
3	1225	212	454
4	539	66	215
5	233	25	90
98	196	196	196
6	102	12	57

7	39	5	29
8	19	2	12
9	9	1	16
10	3		8
96	3	3	3
12	2		
11	1	1	3
13			3
15			2
14			1
17			1

表 7 逾期数量统计

这些特征具有相似的分布。有两个不同的值(98 和 96)。一个借款人不可能在两年内表现出 98 或 96 次拖欠。还可以观察到这些值共享相同的对应索引，这可能表明数据输入错误。然而，它们不能被删除，因为它们拥有识别违约成员的高信息。这类借款人中有 55%违约，而全球违约率为 6%。最好我们保留它们，并为这些值分配一个单独的类。

3. 基础模型

3.1 共线性筛选

通过 VIF(方差膨胀因子)对模型中的特征进行检验，得到如下值(**基础模型/方差膨胀因子.xlsx**)

	feature	vif
0	RevolvingUtilizationOfUnsecuredLines	1.143330715
1	age	1.148324358
2	NumberOfTime30-59DaysPastDueNotWorse	43.82815993
3	DebtRatio	1.031416136
4	MonthlyIncome	1.029787995
5	NumberOfOpenCreditLinesAndLoans	1.299608566
6	NumberOfTimes90DaysLate	76.65267064
7	NumberRealEstateLoansOrLines	1.274164193
8	NumberOfTime60-89DaysPastDueNotWorse	98.07158333
9	NumberOfDependents	1.086482468
10	constant	20.8734827

表 8 VIF 检验共线性

可以观察到，NumberOfTime30-59DaysPastDueNotWorse，NumberOfTimes90DaysLate，NumberOfTime60-89DaysPastDueNotWorse 的 VIF 值非常大，可以考虑舍弃其中两个。

3.2 初版 LR 模型的构建(不做 WOE，也不通过 VIF 筛选特征)

直接将训练数据输入模型进行训练(使用**梯度下降法**训练)，然后测试。(**基础模型/基础 LR 模型**)

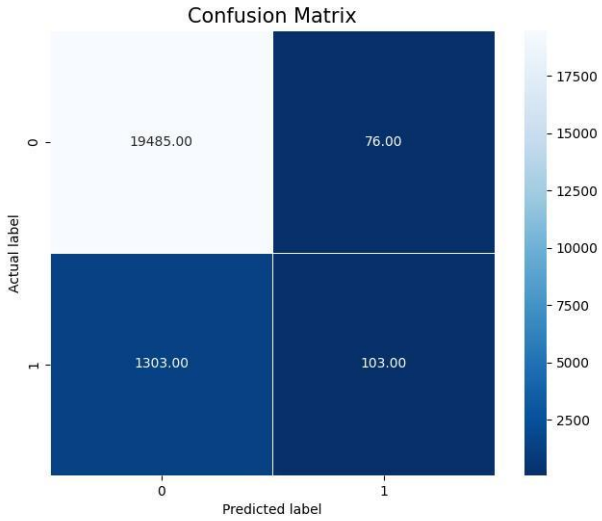

图 9 初版 LR 模型的混淆矩阵

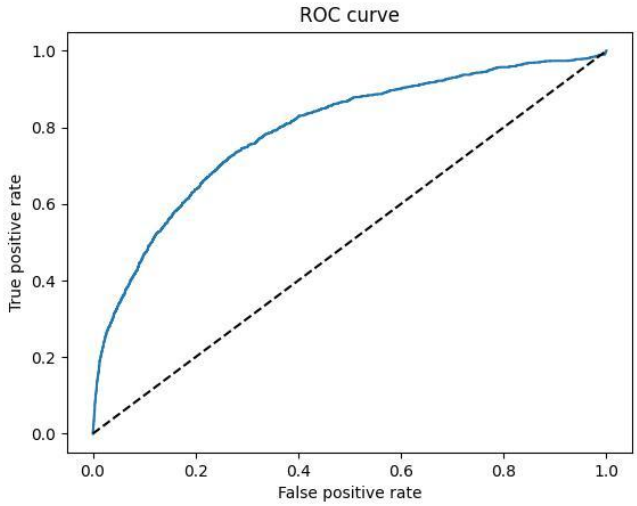

图 10 初版 LR 模型的 AUC 曲线

基准模型只有 0.7904 的 AUC 分数，不够强大。

4. WOE 转换

4.1 离散变量的 WOE 转换

特征中的离散变量(种类小于 50)包括逾期数量和亲属数量总共 4 个特征。
将单个值作为一个分箱计算 WOE，若有 WOE 相近的分箱则合并。

以下三个变量的 WOE 存放在 WOE/PastDue/ 文件夹下

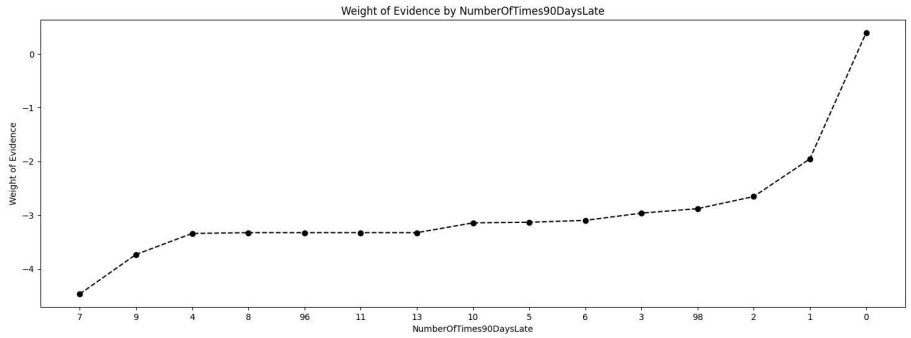

图 11 逾期 90 天的 WOE

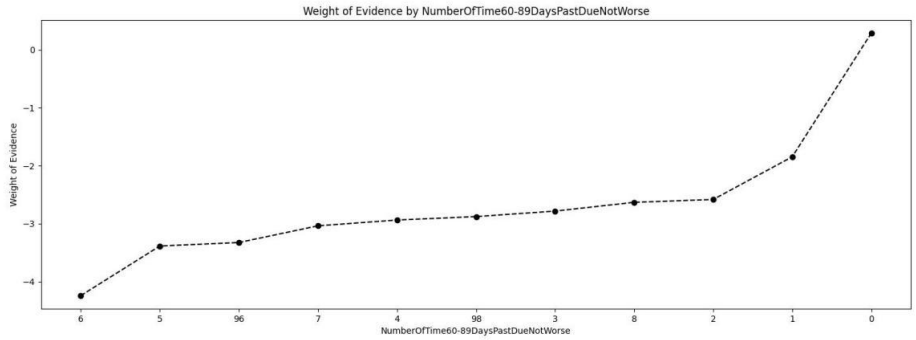

图 12 逾期 60-89 天的 WOE

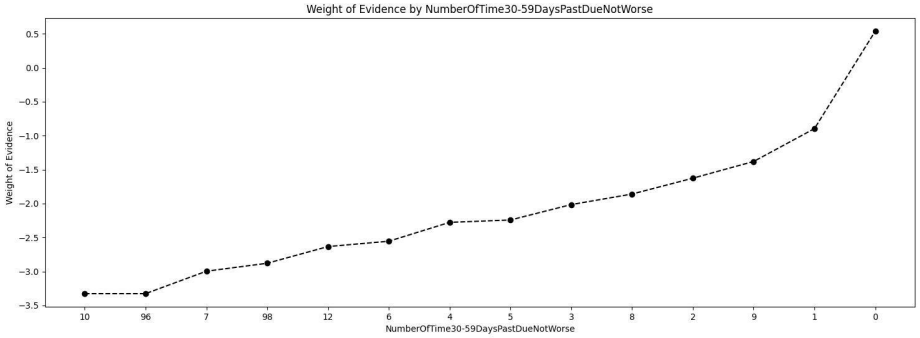

图 13 逾期 30-59 天的 WOE

亲属数量的 WOE 存放在 WOE/PastDue/NumberOfDependents.xlsx 中。

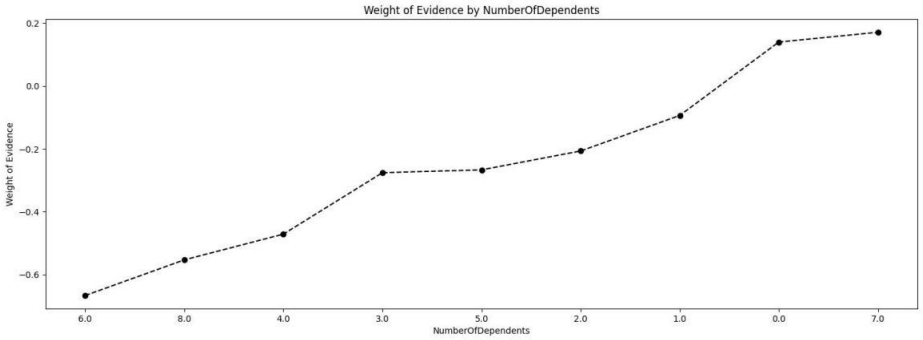

图 14 亲属数量的 WOE

连续变量的 WOE 转换

4.1.1 月收入(MonthlyIncome)

先绘制一个基本的数据分布：

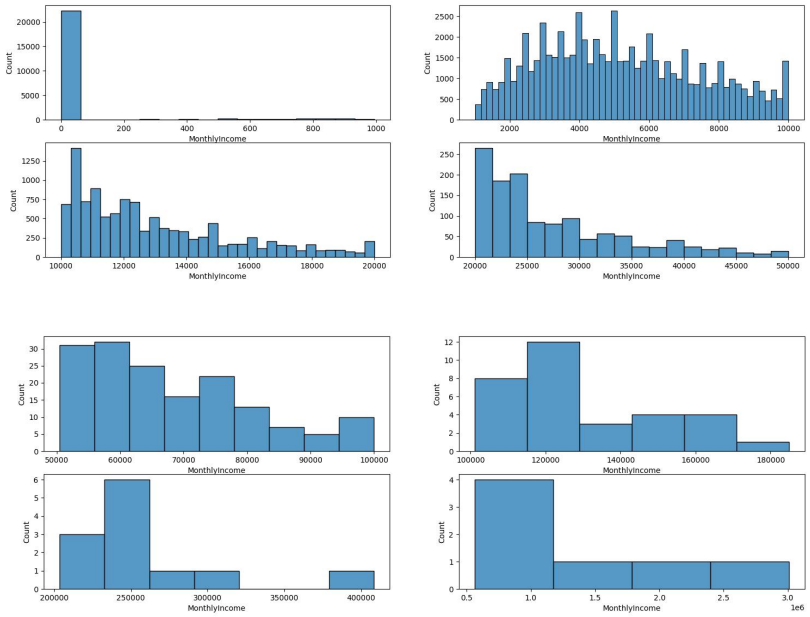

图 15 月收入的频数分布

将 MonthlyIncome 做手动分箱

区间	分箱数
[0, 1000]	1
(1000, 10000]	4(按照 4 分位数分箱)
(10000, 20000]	等距分为 4 箱
(20000, 50000]	2 箱([20000, 30000], [30000, 50000])
(50000, 70000]	1
(70000, 100000]	2
(100000, 140000]	3
(140000,200000]	4
(200000,500000]	5
(500000,3500000]	6

表 9 月收入分箱

MonthlyIncome 的 woe 值见 **WOE/MonthlyIncome/MonthlyIncome.xlsx** 表格。

4.1.2 DebtRatio

先观察数据分布

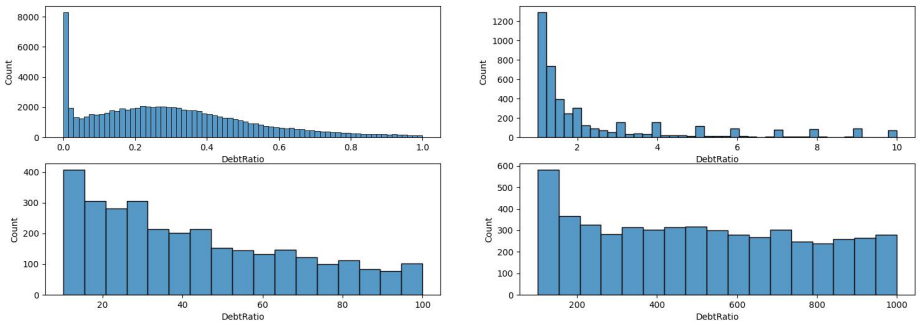

表 10 负债率分布

将 DebtRatio 做手动分箱

区间	分箱数
[0, 1)	等距分为 9 箱
[1, 10)	1
[10, 100)	1
[100, 1000)	1
[1000, max)	1

表 11 负债率手动分箱

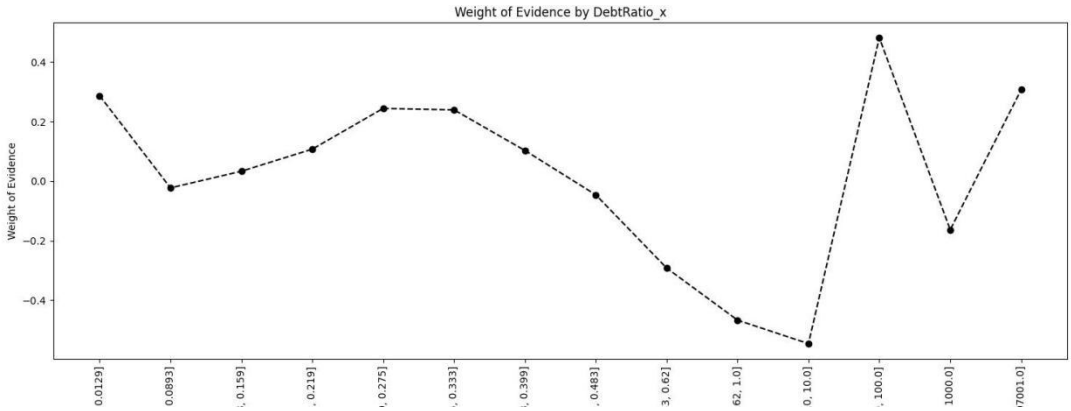

图 16 负债率 WOE 曲线

计算得到的 WOE 值见 **WOE 转换/DebtRatio/DebtRatio.xlsx** 中

4.1.3 RevolvingUtilizationOfUnsecuredLines

观察数据分布

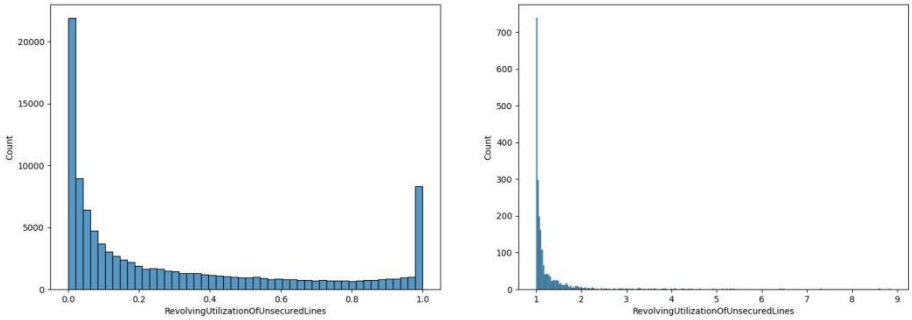

图 17 RevolvingUtilizationOfUnsecuredLines 的数据分布

分箱策略：

区间	分箱数
[0, 1]	49
(0, max)	9

表 12 分箱策略

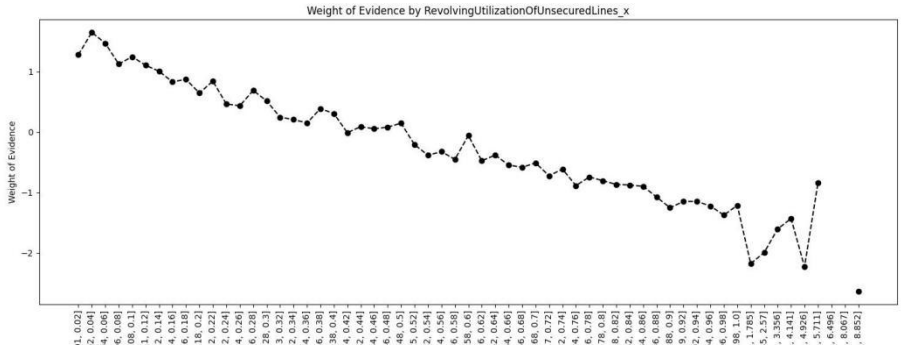

表 13 WOE 曲线

该特征的 WOE 值存放 **WOE 转换文件夹中的同名文件夹中。**

4.2.4 NumberOfOpenCreditLinesAndLoans

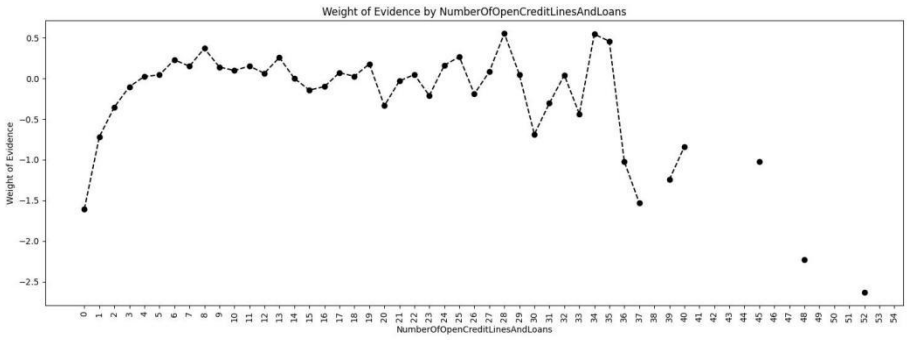

图 18 信贷数量 WOE 曲线

该特征的 WOE 值存放 **WOE 转换文件夹中的同名文件夹中。**

4.2.5 NumberRealEstateLoansOrLines

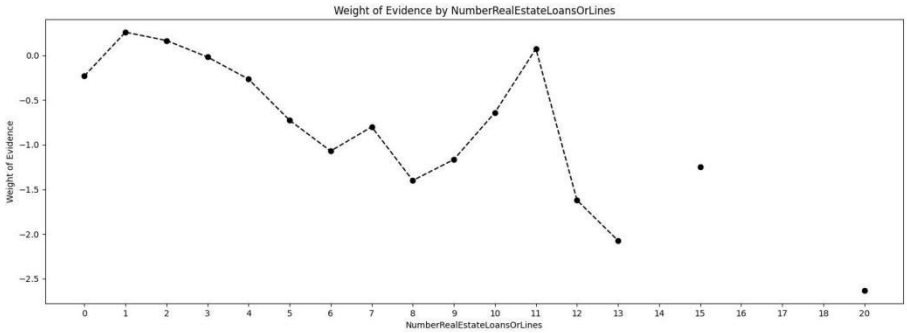

图 19 房贷抵押贷款 WOE 曲线

该特征的 WOE 值存放 **WOE 转换文件夹中的同名文件夹中。**

4.2.6 Age

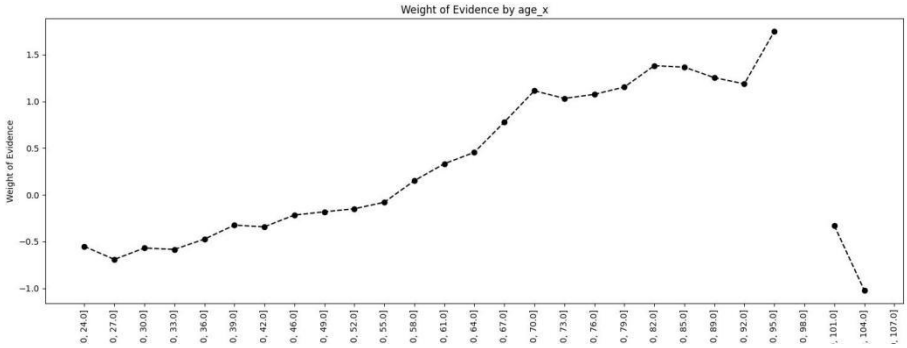

图 20 年龄的 WOE 曲线

其中 Age 变量在(104.0, 107.0]区间上无法求解 WOE，直接与后一个区间(107.0, 110.0]合并(在 Excel 中操作)。该特征的 WOE 值存放 **WOE 转换文件夹中的同名文件夹中。**

4.2 编码与建模

前一小节求出的 WOE 中包含许多正无穷和负无穷，将其分别替换为 1 和-1，然后将所有特征都替换为相应的 WOE 编码值。然后带入 LR 模型中进行训练和预测。

4.2.1 不筛选变量直接入模

不筛选特征直接带入得到如下结果(**最终模型/未做特征筛选的模型/**)

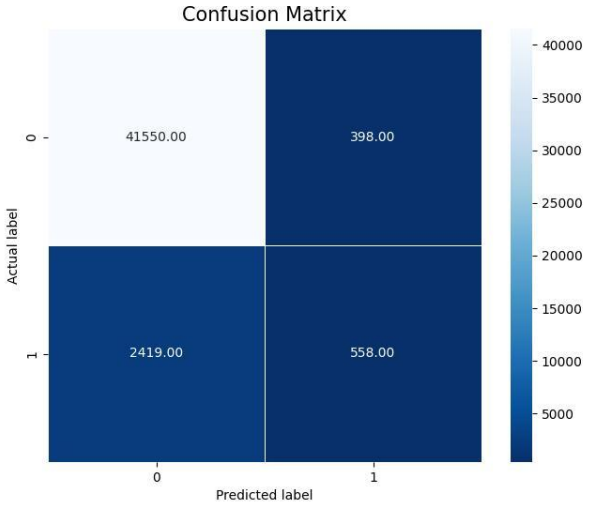

图 21 LR_WOE 模型的混淆矩阵

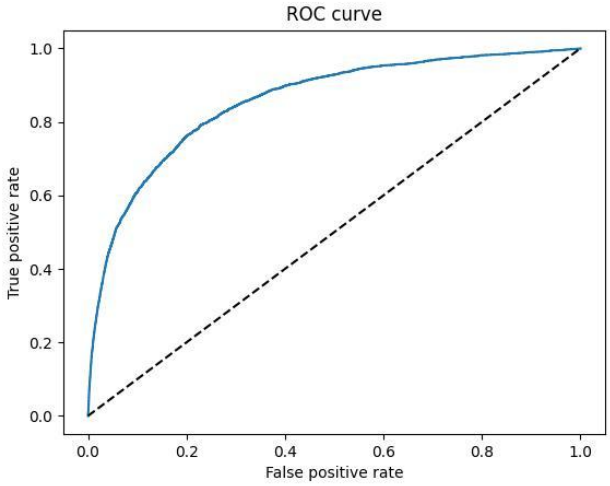

图 22 LR_WOE 模型的 ROC 曲线

最终得到的 ROC 分数为 0.856561344293088，显著优于基准模型的 0.7904。

最终得到的逻辑回归模型为。（**最终模型/未做特征筛选的模型/权重和评分表.xlsx**）

$$\beta = \begin{pmatrix} -0.496299193 \\ -0.419656597 \\ -0.522785723 \\ -0.884935759 \\ -0.219174254 \\ 0.112226311 \\ -0.521525077 \\ -0.592248789 \\ -0.373283042 \\ -0.271610416 \end{pmatrix}, b = -2.65221902$$

$$p = \text{sigmoid}(x\beta + b)$$

4.2.2 筛选部分变量入模

取出 VIF 值较大的两个特征 NumberOfTime60-89DaysPastDueNotWorse 和 NumberOfTimes90DaysLate 不入模。训练后得到的模型在测试集的预测结果为(**最终模型/特征筛选的模型/**)

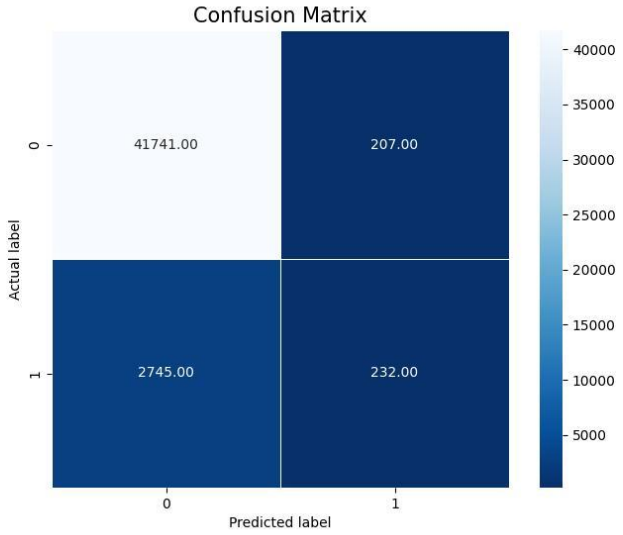

图 23 LR_WOE_FE 的混淆矩阵

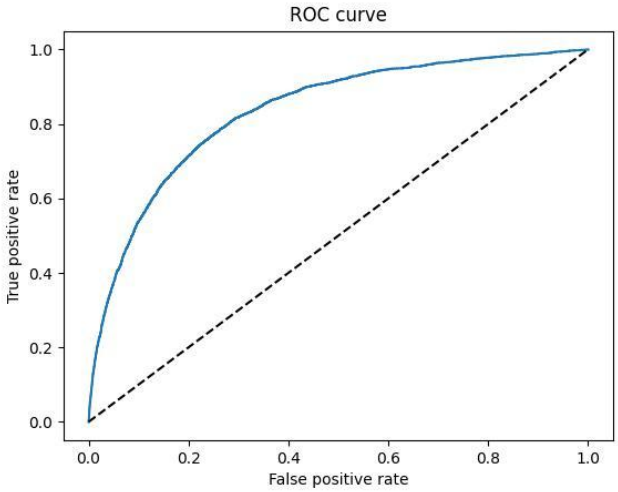

图 24 LR_WOE_FE

最终得到的 AUC 分数为 0.8338956794693009，同样显著优于基准模型的 0.7904，但相比与前一小节的模型性能有所下降。

最终得到的逻辑回归模型为。权重存放在(**最终模型/特征筛选的模型/权重.xlsx**)

$$\beta = \begin{pmatrix} \text{-6.16779236e - 01} \\ \text{-4.41427546e - 01} \\ \text{-7.52550804e - 01} \\ \text{-8.02165487e - 01} \\ \text{-3.84161140e - 04} \\ \text{-1.19759010e - 01} \\ \text{-6.47582839e - 0} \\ \text{-3.00387302e - 01} \end{pmatrix}, b = \text{- 2.69888313}$$

$$p = \text{sigmoid}(x\beta + b)$$

5. 模型转换

5.1 转换方式

LR 模型不够直观，需要转化为评分模型。
首先将客户逾期的概率表示为 p，则正常的概率为 1-p。二者的比值即 odds 的计算公式如下：

$$\text{Odds} = \frac{p}{1-p}$$

要将逻辑回归模型的参数转换为评分卡，需要如下的公式

$$\text{Score} = A - B\log(\text{Odds})$$

其中 A 和 B 是参数。式中的负号可以使逾期概率越低，得分越高。通常情况下，这是分值的理想变动方向，即高分值代表低风险，低分值代表高风险。公式中的 log(odds)为逻辑回归模型的计算结果

$$\log(\text{odds}) = x^T\beta + b$$

根据数据拟合后，得到的模型参数为 w。式子中的 A 和 B 可以通过两个已知的分数或假设的分值得到。
式中的常数 A、B 的值可以通过将两个已知或假设的分值带入计算得到。
通常情况下，需要有两个前提假设才能计算出 A、B 的值：
（1）给某个特定的比率设定特定的预期分值(P)；
（2）确定比率翻番的分数（PDO）
根据以上的分析，我们首先假设比率为 x 的特定点的分值为 P。则比率为 2x 的点的分值应该为 P-PDO。代入式中，可以得到如下两个等式：$P = A - B\log(x)$

$$P - \text{PDO} = A - B\log(2x)$$

解方程得到 $B = \frac{\text{PDO}}{\log(2)}$; $A = P + B\log(x)$

假定我们设定好坏比 1：1 时分数是 500，好坏比翻倍的分数（PDO）为 20 分，代入式中求得：B=20/log(2)，A=500。最终评分的分值为

$$\text{Score} = A - B\{x^T\beta + b\}$$

此时所有变量都是 WOE 编码后的值，可以将这些自变量中的每一个都写成如下形式

$$\text{Score} = A - B\{\beta_0 + (\beta_1\omega_{11})\delta_{11} + (\beta_1\omega_{12})\delta_{12} + \ldots + (\beta_x\omega_{x2})\delta_{x2} + \ldots\}$$

其中ω_{11}为x_i变量的第 j 个分组的 WOE，β_i为逻辑回归方差的系数，为已知变量。δ_{ij}为二元变量(0/1),表示变量x_i是否在第 j 个分组中，因为每个入模变量只会在某个记录上取一个值，例如：x_1变量的取值为第二分组，则δ_{12}取 1，δ_{11}，δ_{13}…δ_{1j}取值为 0.上面的式子可重新表示为

$$\text{Score} = (A - B\beta_0) - (B\beta_1\omega_{11})\delta_{11} - (B\beta_1\omega_{12})\delta_{12} - \ldots - (B\beta_x\omega_{x2})\delta_{x2} - \ldots$$

此式为最终的评分卡公式，转换为表格形式为

变量	取值(分箱)	分值
基准点	----	$(A - B\beta_0)$
x_1	1	$-B\beta_1\omega_{11}$
	2	$-B\beta_1\omega_{12}$
	…	…
	k_1	$-B\beta_1\omega_{1k_1}$
x_2	1	$-B\beta_2\omega_{11}$
	2	$-B\beta_2\omega_{12}$
	…	…
	k_2	$-B\beta_2\omega_{1k_1}$
…		
x_n	1	$-B\beta_n\omega_{11}$
	2	$-B\beta_n\omega_{12}$
	…	…
	k_n	$-B\beta_n\omega_{1k_1}$

详细评分表存放在(**最终模型/未做特征筛选的模型/权重和评分表.xlsx**)

5.2 转换结果

因为提取特征后模型的性能有所下降，所以只转换没有提取特征的模型的评分卡。在输出的 WOE 表格中输入公式进行计算，令 B=20/log(2)，A=500。计算得到的评分卡见 Excel

6. 模型评价

6.1 CAP 和 AR 分数

CAP（Cumulative Accuracy Profile）的步骤如下：首先，对于所有预测的结果，按照其概率 probability 值降序排列起来，然后我们顺序依次把样本划分到观察集中。假设所有的样本数量为 T，真实的正样本数量为 Tp，真实的负样本的数量为 Tn，Tp+Tn=T。每次都划分一个样本到观察集中，观察集中的样本个数为 O，观察集中包含的真正的正样本数为 Op。

在最理想的情况下，存在一个阈值，大于的样本全部为真实正样本，小于的全部为真实负样本，所以在这种情况下，probability>时，每增加一个样本到预测的正样本中，都是真实的正样本，直到阈值，此时 O=Op=Tp，然后再往里加样本时，所有的样本都是负样本，所以将一直保持 O=Op=Tp 的状态，直到所有样本添加完。

在最差的情况下，则是一个斜率为常数的直线。

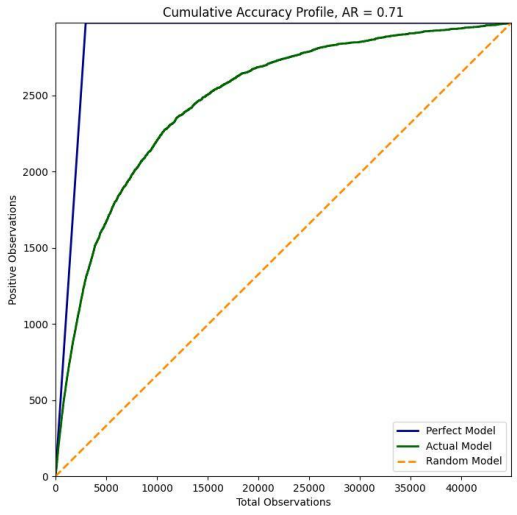

图 25 CAP

设 S 为完美模型和最差模型之间的面积，Sa 为实际模型与最差模型之间的面积，AR 分数为

$$AR = \frac{S}{S_a}$$

本模型的 AR 分数为 0.7131247866137768。

6.2 KS

6.2.1 基本指标介绍

在介绍 KS 之前，需要先介绍以下指标

混淆矩阵		预测值	
		Positive	Negative
真实值	Positive	TP	FN
	Negative	FP	TN

真实值是 positive，模型认为是 positive 的数量（True Positive=TP）
真实值是 positive，模型认为是 negative 的数量（False Negative=FN）：这就是统计学上的第一类错误（Type I Error）
真实值是 negative，模型认为是 positive 的数量（False Positive=FP）：这就是统计学上的第二类错误（Type II Error）

指标	公式	意义
准确率（ACC）	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	分类模型所有判断正确的结果占总结果的比例
精确率（PPV）	$Precision = \frac{TP}{TP+FP}$	在模型预测是Positive的所有结果中，模型预测正确的比例
真阳性率（TPR） 灵敏度	$Sensitivity = Recall = \frac{TP}{TP+FN}$	在真实值是Positive的所有结果中，模型预测错误的比例
真阴性率（TNR） 特异度	$Specificity = \frac{TN}{TN+FP}$	在真实值是Negative的所有结果中，模型预测正确的比例
假阳性率（FPR） 误诊率	$FPR = 1 - Specificity = \frac{FP}{TN+FP}$	在真实值是Negative的所有结果中，模型预测错误的比例
假阴性率（FNR） 漏诊率	$FNR = 1 - Sensitivity = \frac{FN}{TP+FN}$	在真实值是Positive的所有结果中，模型预测错误的比例
F1 Score	$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$	综合了Precision与Recall的产出的结果。F1-Score的取值范围从0到1的，1代表模型的输出最好，0代表模型的输出结果最差。

图 26 基本指标

KS(Kolmogorov-Smirnov)曲线（洛伦兹曲线）的纵轴是表示 TPR 和 FPR 的值，就是这两个值可以同时在一个纵轴上体现，横轴就是阈值，表示模型能够将正、负客户区分开的程度越大。两条曲线之间相距最远的地方对应的阈值，就是最能划分模型的阈值，即 KS=max(TPR-FPR)，KS 值越大，模型的区分度越好。

6.2.2 KS 曲线

K-S 曲线的做法：
(1)把模型对样本的输出概率(predict_proba)从大到小排序，计算对应不同阈值时，大于等于阈值的样本数占总样本的百分比 percentage

(2)计算阈值取每个概率时对应的 TPR 和 FPR 值，分别画(percentage, TPR)和(percentage, FPR)的曲线
(3) K-S 曲线上的 KS 值，即 max(TPR−FPR)，即两条曲线间的最大间隔距离。
本模型的 KS 曲线如下图

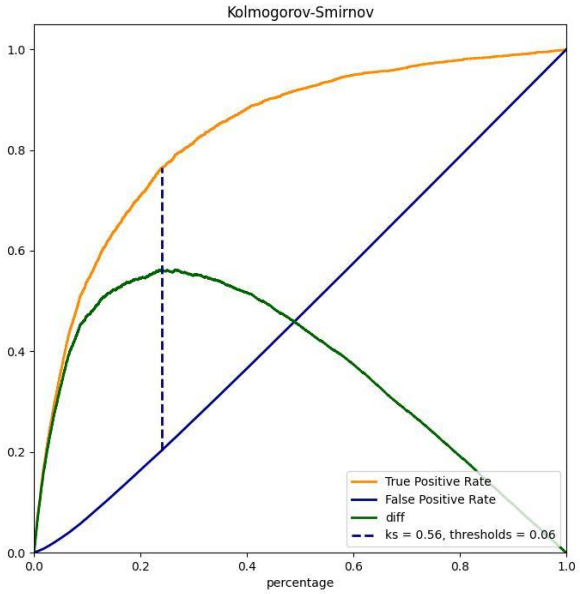

图 27 KS 曲线

6.3 ROC 和 AUC

实质上，ROC 曲线是多个混淆矩阵组合的结果。简单来说，模型对每个样本的预测结果为一个概率值，我们需要从中选取一个阈值来判断客户的好与坏。定好阈值后，超过此阈值认为客户为坏客户，低于此阈值定义为好客户。每一个阈值都有对应的混淆矩阵，都有对应的 TPR 和 FPR，取多个阈值就能得到多个 TPR 和 FPR，就能绘制出曲线。

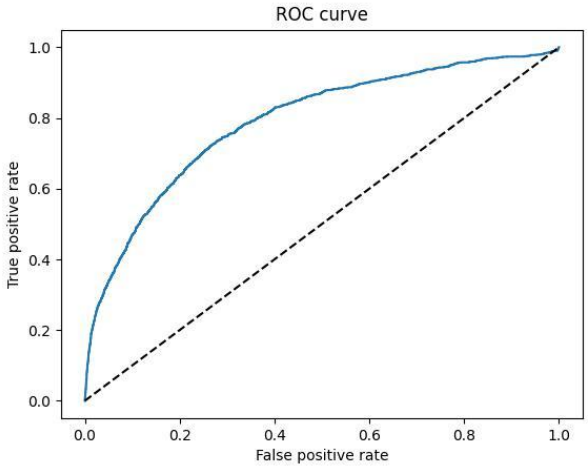

图 28 ROC 曲线

本模型的 AUC 分数为 0.856561344293088，大于 0.5，优于随机猜测。

6.4 PSI 稳定度检验

PSI 反映了验证样本在各分数段的分布与建模样本分布的稳定性。稳定性的检验是需要参照的，因此需要两个分布，即实际分布和预期分布。其中建模时将训练样本作为预期分布，测试或者验证样本作为实际分布。从视觉上理解，

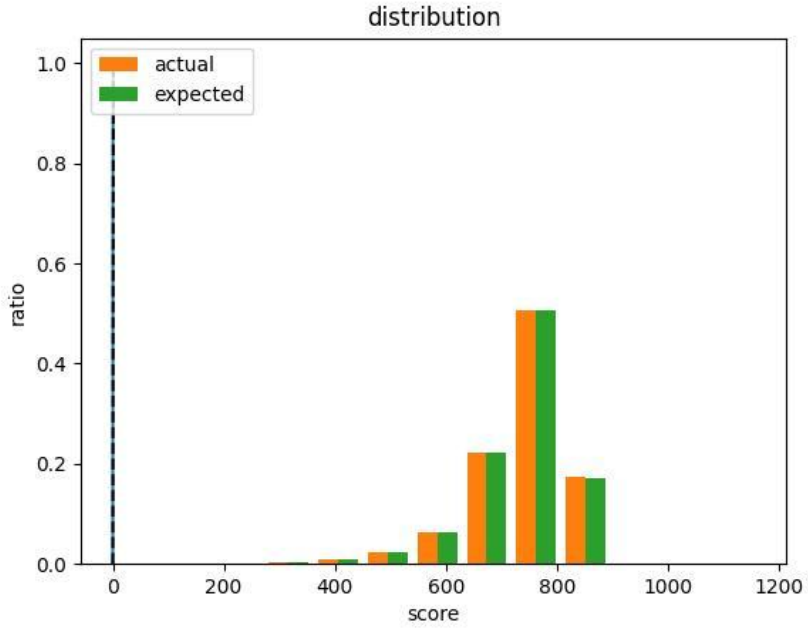

图 29 分布的对比

测试集和训练集的分布基本一致。PSI 的计算公式为

$$psi = \sum_{i=1}^{n} (A_i - E_i) * \ln(\frac{A_i}{E_i})$$

计算得到 PSI 的值为 0.0007621062834286972，说明训练集和测试集的得分分布一致。

7. 存在的问题

(1) WOE 分箱使用人工的方式而不是最有分享，部分 WOE 值为无穷大。
(2) 将无穷大的 WOE 直接替换为 1 或者-1 比较粗暴，应该有更加自适应的方法。
(3) 部分特征的评分卡不符合常理，比如年龄评分卡中，老年人的得分高于年轻人。
(4) 筛选特征时，没有根据 IV 值进行筛选，也没有求各个变量之间的相关性。