某银行信用卡评分分析

背景描述及需求

银行在市场经济中起着至关重要的作用。他们决定谁能获得资金，以什么条件获得资金，并决定投资决策的成败。为了让市场和社会发挥作用，个人和企业需要获得信贷。信用评分算法是银行用来决定贷款是否应该发放的一种方法，它对违约概率进行猜测。为了推进信用卡业务良性发展，减少坏账风险，各大银行都进行了信用卡客户风险识别相关工作，建立了相应的客户风险评分模型。

本案例分析的是通过预测某人在未来两年内遭遇财务困境的可能性，来提高自己在信用评分方面的水平。主要应用于相关融资类业务中新用户的主体评级，适用于个人和机构融资主体。本案例定义逾期90天以上就算作坏客户。

信用卡客户评分数据挖掘主要包括以下步骤：

1) 从银行获取信用卡相关信息；

2) 数据探索：探索整体数据分布和探索不同变量之间的关系；

3) 数据预处理工作：包括数据清洗、数据离散化、处理样本不平衡问题等操作；

4) 构建信用评分卡模型，计算各指标的分值及综合评分；

5) 根据评分结果，分析该银行的客户的信用风险情况。

实施过程

本案例数据集来源于kaggle赛题数据，共计15万条客户数据，包括信用客户和逾期客户，并对数据进行人工标注，标注分为两类，分别为：0（信用客户）和1（逾期客户）。字段说明见表2-1。

字段名称	类型	字段说明
SeriousDlqin2yrs	整型	好坏客户。取值为{0,1}
RevolvingUtilizationOfUnsecuredLines	浮点型	可用额度比值
age	整型	年龄
NumberOfTime30-59DaysPastDueNotWorse	整型	逾期30-59天笔数
DebtRatio	浮点型	负债率
MonthlyIncome	整型	月收入
NumberOfOpenCreditLinesAndLoans	整型	信贷数量
NumberOfTimes90DaysLate	整型	逾期90天笔数
NumberRealEstateLoansOrLines	整型	固定资产贷款量
NumberOfTime60-89DaysPastDueNotWorse	整型	逾期60-89天笔数
NumberOfDependents	整型	家属数量

表2-1 字段说明

数据接入

在实验中添加 数据源 节点，将评分卡客户数据读取进来，部分数据如图2-1所示：

图2-1 评分卡客户数据

为了方便理解本数据集每个特征的含义，使用 元数据编辑 节点，添加中文字段别名，更改后的输出如图2-2所示，流程图如图2-3所示。

图2-2 元数据编辑

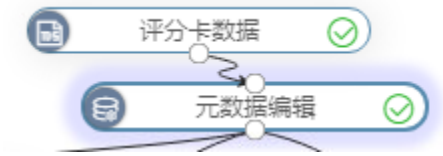


图2-3 流程图

数据探索

本案例的探索分析是对数据进行缺失值、重复值与异常值分析，分析出数据的规律以及异常值。为了查看整体数据集数值型数据的情况，我们接入一个全表统计节点，选中所有数值型字段如图2-4所示，输出结果如图2-5所示，可以看到部分数据（月收入、家属数量）存在缺失值。可以看到“月收入”缺失达到近20%，“家属数量”缺失较少仅有2.6%的缺失。

图2-4 选择列

查看全量统计

	方差	总和	行数	唯一值	缺失值	偏度	峰度
...	0.06237...	10026	150000	2	0	3.46882...	10.0327..
...	62377.7...	907265....	150000	11985	0	97.6305...	14544.2..
...	218.208...	7844281	150000	86	0	0.18899...	-0.4946...
...	17.5794...	63155	150000	16	0	22.5968...	522.359..
...	4152704...	5.29507...	150000	19868	0	95.1568...	13733.8..
...	2.06918...	8022208...	150000	13594	29731	114.038...	19503.8..
...	26.4808...	1267914	150000	58	0	1.21530...	3.09092..
...	17.3830...	39896	150000	19	0	23.0871...	537.720..
...	1.27638...	152736	150000	28	0	3.48244...	60.4747..
...	17.2655...	36058	150000	13	0	23.3315...	545.664..
...	1.24341...	110612	150000	13	3924	1.58822...	3.00151..

图2-5 数据缺失情况

为了统计所有数据中好坏客户的分布情况，选择 聚合 节点，选择分组计数，如图2-6所示，输出结果分布情况图2-7所示，发现0类样本占有较大的比例，则需要考虑到样本不平衡问题。

图2-6 聚合

图2-7好坏客户分布情况

通过 全表统计 节点查看所有数据的分布情况，查看各指标的直方图、箱线图分布情况，如图2-8所示。发现“年龄”的最小值居然是0，但是根据我们的常识，小于18岁是不能在银行办理信用卡或是贷款业务的。以及看到三个逾期天数指标（逾期30-59天、逾期60-80天，逾期90天）是存在比较严重的离群值的。

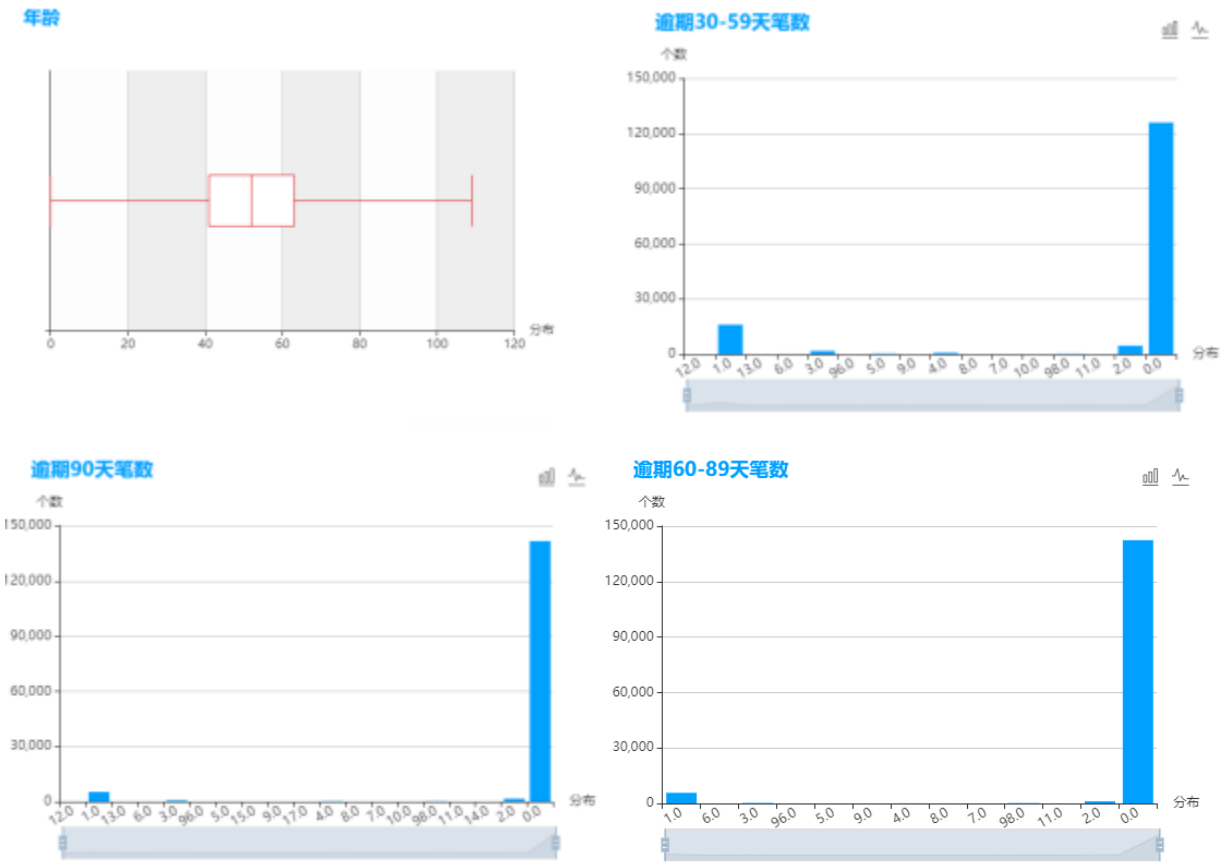


图2-8 直方图、箱线图

数据预处理

通过数据探索发现，月收入、家属数量这两个字段数据有部分空值、三个逾期天数指标存在异常值和部分数据可能有重复值。以及好坏客户的数据比例存在明显的不平衡现象，如果将这些数据直接进入模型，必然会对分析造成很大的影响，得到的结果的质量也必然是存在问题的。那么，在利用到数据之前就必须先进行数据预处理，把无价值的指标及数据去除。

去重复值

通过 去除重复值 节点将重复行的数据进行给去除，去除后结果如图2-9所示：

当前显示 100 条 / 总共有 149389 条数据 提示:点击单元格可查看超出的内容

# 好坏客户	# 可用额度比值	# 年龄	# 逾期30-59天笔数	# 负债率	# 月收入	# 信贷数量	# 逾期90天笔数
0	0.0344	69	0	0.0424	2500	17	0
0	0.834	38	0	0.3937	13000	13	0
0	0.0	76	0	0.0025	7704	3	0
0	0.4124	28	0	0.0373	4991	5	0
0	1.0	41	0	0.2526	4663	4	0
0	0.0316	39	0	0.6447	3900	14	0
0	0.0923	61	0	0.2294	7500	23	0
0	0.6416	58	0	0.3869	5000	7	0
0	0.0788	30	0	2129.0	null	6	0
1	1.0276	54	4	1.056	6250	17	1
1	1.0	34	0	0.2543	6166	5	1

注意：表头中◇表示特征列，*表示标签列 表头真名 表头别名

图2-9 去除重复值

空值处理

由于“家属数量”缺失较少，可直接使用中位数进行填充。根据“月收入”分布情况来看可将这个特征用平均值来填充，流程图如图2-10所示。

图2-10 空值处理

异常值处理

根据探索分析发现年龄的最小值为0，通常我们知道年龄小于18岁是不能办理银行信用卡或者贷款业务的，并且发现三个逾期天数指标（逾期30-59天、逾期60-80天，逾期90天）是存在比较严重的离群值的。通过 行选择 节点筛选出年龄<18的数据分析发现仅有年龄=0的这一条数据，如图2-11所示。因此需要将年龄<18的数据进行删除过滤，如图2-12所示。

图2-11 异常值

图2-12 过滤

通过行选择节点筛选出发现三个逾期指标出现的情况发生在相同的行，维度都是（225,11）。因此将其中一个异常指标过滤删除即可，如图2-13所示。

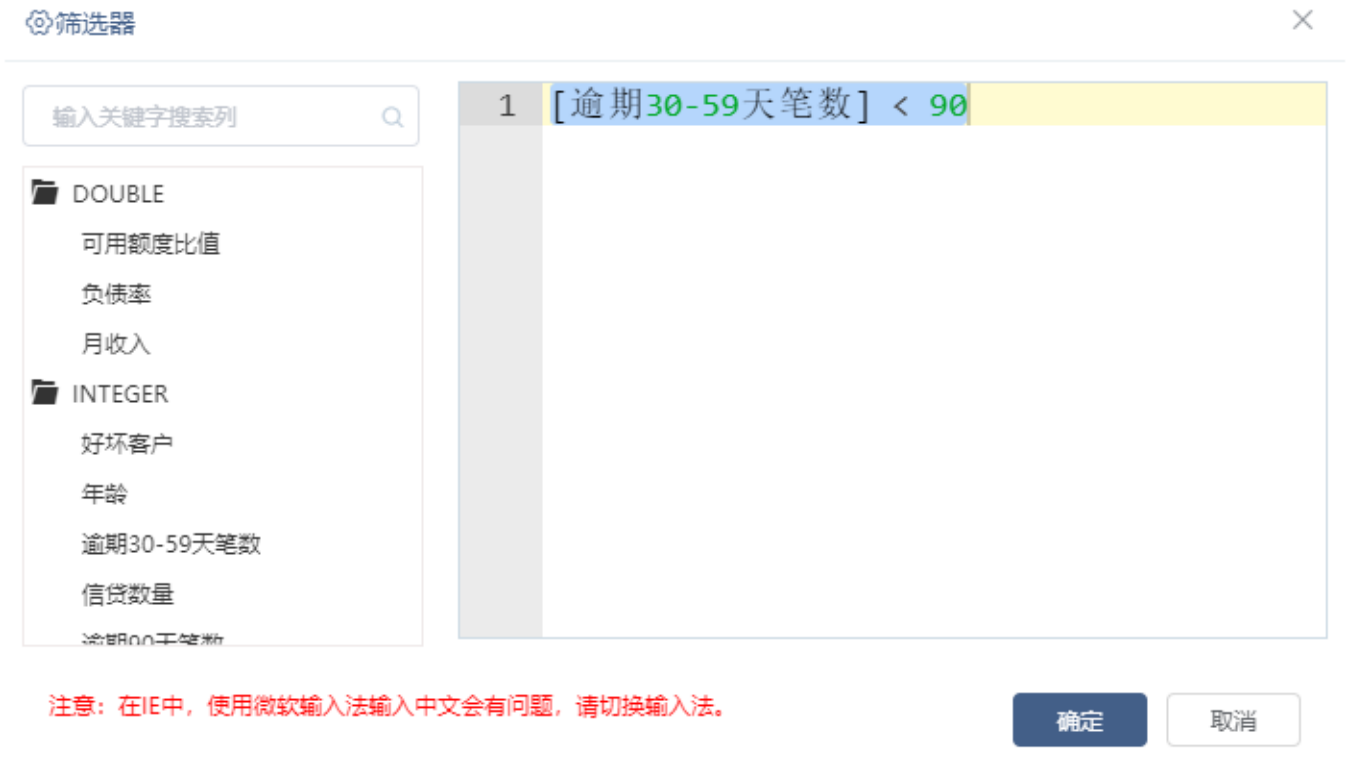


图2-13过滤

处理样本不平衡

通过图2-7所示发现，0:1=139974:10026，是存在严重的样本不平衡的。这是在金融风控中非常常见的，因为会存在严重违约的用户毕竟是少数。本案例采取下采样的方法处理数据不平衡。如图2-14所示。处理不平衡数据后通过 聚合 节点分析发现1类和0类数据达到平衡状态，如图2-15所示。

图2-14 下采样设置

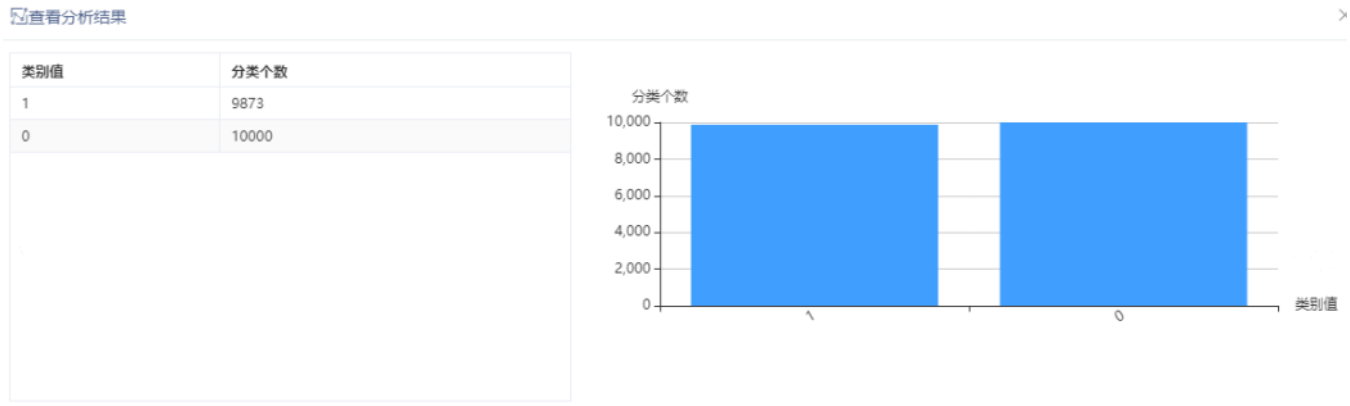


图2-15 下采样后处理结果

WOE编码

在建立模型前，需要对连续变量离散化，特征离散化后，模型会更稳定，降低了模型过拟合的风险。连续变量是在任意两个值之间具有无限个值的数值变量。连续变量可以是数值变量，也可以是日期/时间变量。例如，零件的长度，或者收到付款的日期和时间。因此，我们使用WOE编码节点操作分箱设置，如图2-16所示。

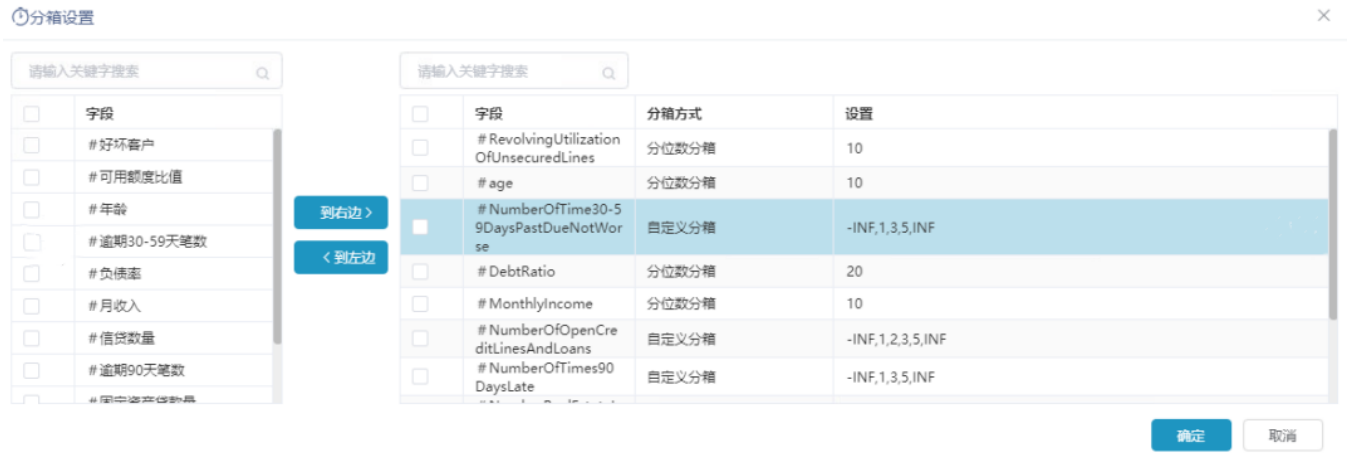


图2-16分箱

分箱设置后通过WOE编码节点计算出IV值以及WOE值，通过查看IV值筛选重要特征建立评分卡模型，将WOE编码后的特征输入模型。WOE编码后输出结果如图2-17所示。

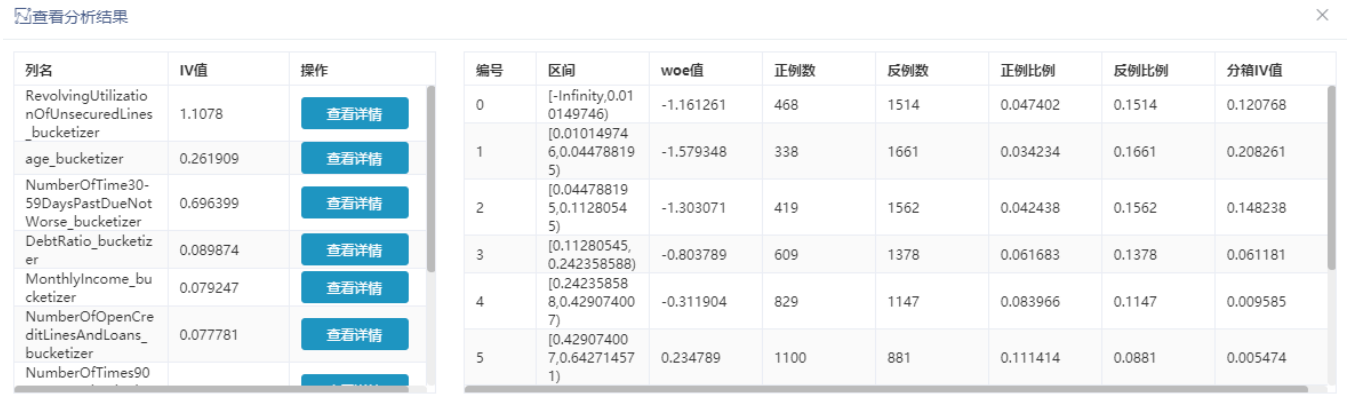


图2-17 WOE编码

整个的数据预处理流程如图2-18所示。

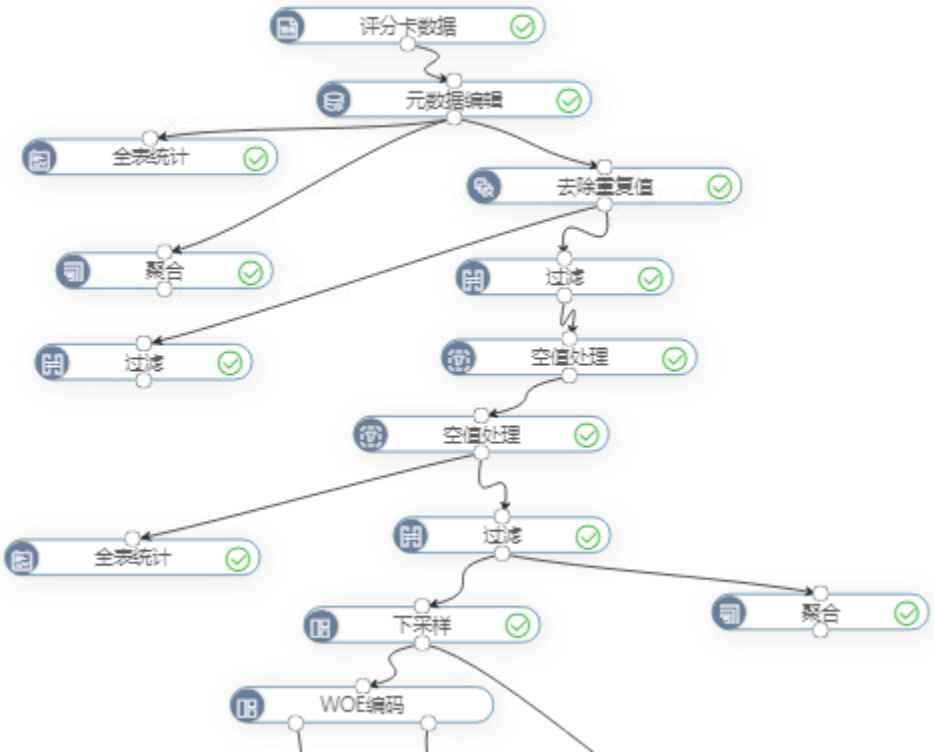


图2-18 数据预处理

特征选择

特征选择

我们根据IV值判断选择需要进入评分卡模型的WOE编码后的特征列，如图2-19所示。

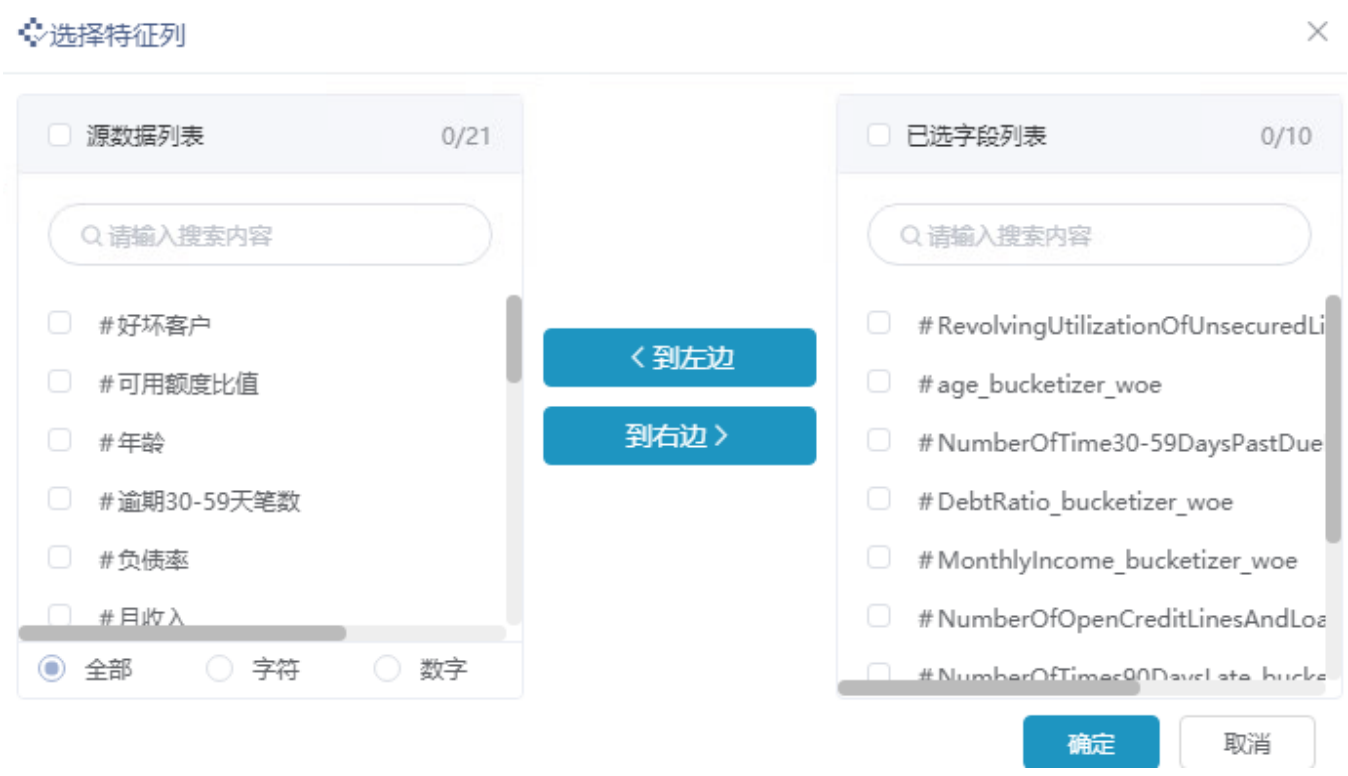


图2-19 特征选择

模型建立

本案例采取逻辑回归模型，整体的流程图如图2-20所示。

逻辑回归具有以下优势：

1、逻辑回归经过信贷历史的反复验证是有效的；

2、模型比较稳定相对成熟；

3、建模过程透明而不是黑箱；

4、不太容易过拟合；

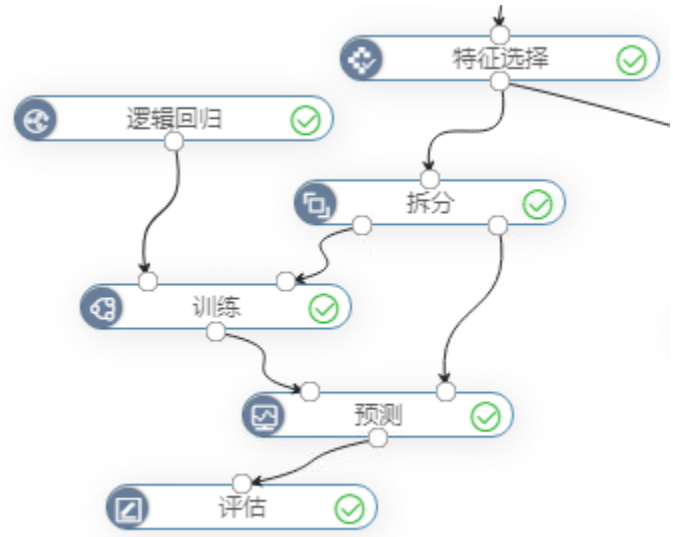
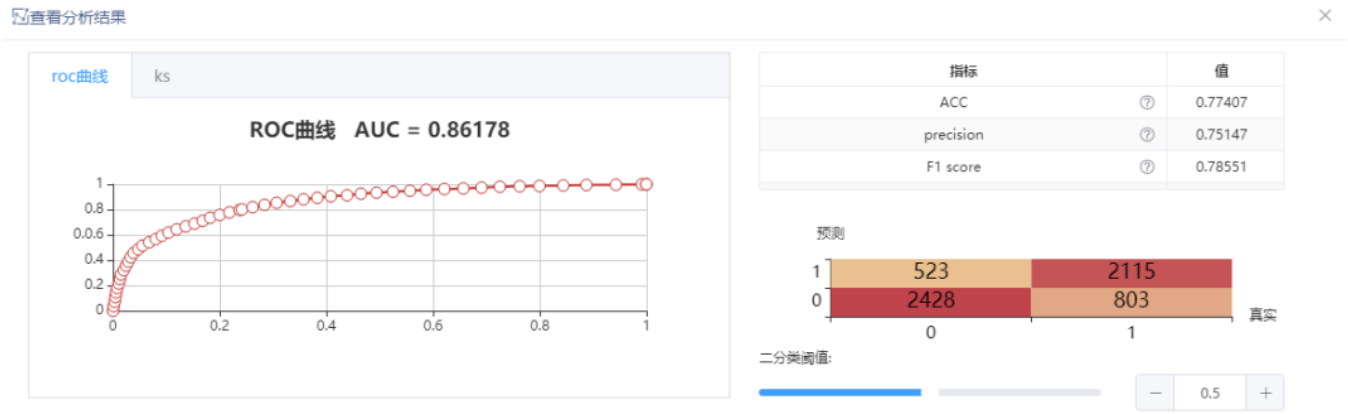


图2-20 模型训练及预测评估

通常而言，评分卡模型一般采用roc或ks曲线来评价模型的好坏。本案例的评估结果如图2-21所示，发现该模型的auc取值为0.86178，ks的最大取值为0.56，说明该模型的效果是不错的。



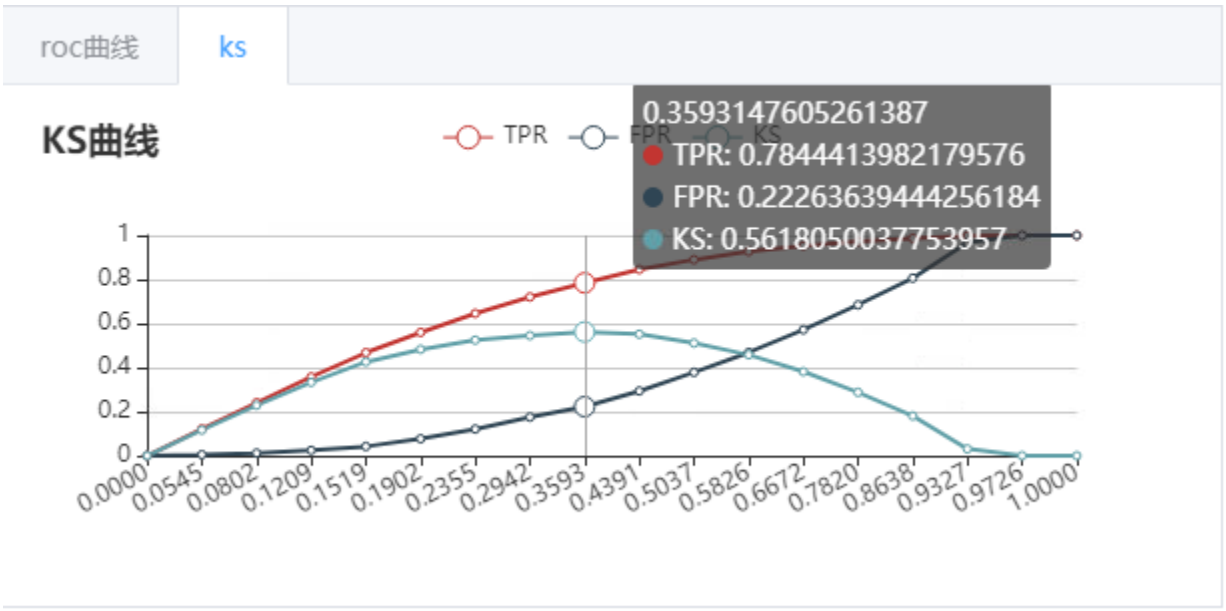


图2-21 评估结果

计算评分

模型系数

通过逻辑回归模型训练后接入 模型系数 节点，输出的模型系数如图2-22所示。

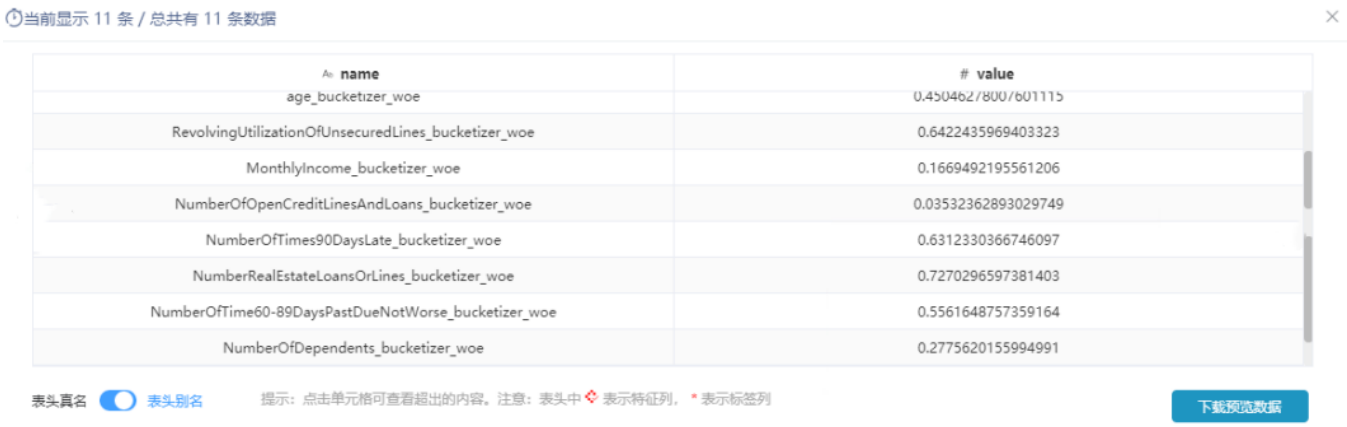


图2-22 模型系数

评分卡构建

我们需要将逻辑回归转换为对应的分数，（0-999分）。接入评分卡构建节点，设置好相应的参数，我们设置基础分值为1000，好坏比为0.2，PDO为20（每高20分好坏比翻一倍），如图2-23所示评分卡构建模型如图2-24所示。

图2-23 评分卡构建

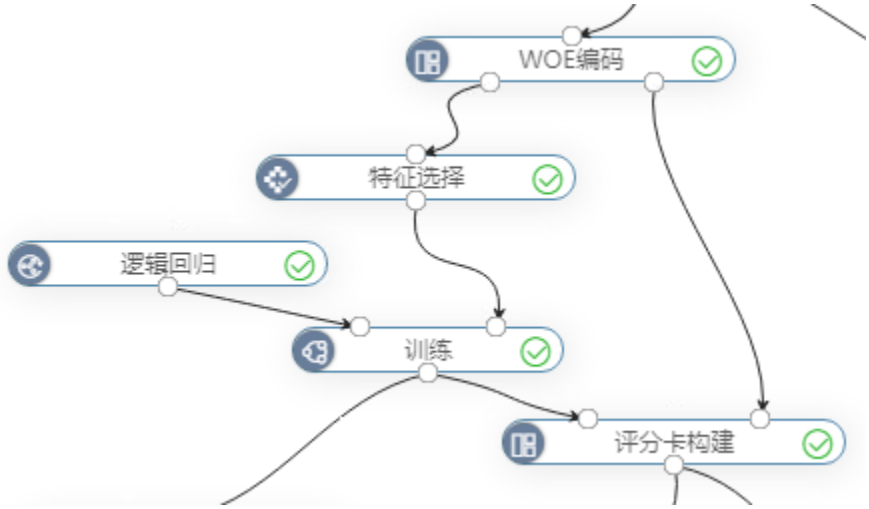


图2-24 评分卡模型

评分卡输出

我们在评分卡模型构建后接入评分卡输出节点，评分卡表输出结果如图2-25所示。

当前显示 75 条 / 总共有 75 条数据

	# bucketizer	block	# woe	# iv	# unscaled	# scaled
etizer	1.0	(0.010149746,0.044788195]	-1.5793478797655431	0.2082610528927908	-1.0143260631207098	29.267263622173477
etizer	2.0	(0.044788195,0.11280545]	-1.3030700761127605	0.1482385875285127	-0.8368884127479719	24.1474952569787
etizer	3.0	(0.11280545,0.242358588]	-0.8037888494989738	0.06118169129192435	-0.5162282418827523	14.895198490621501
etizer	4.0	(0.242358588,0.429074007]	-0.31190362762934265	0.009585929771378213	-0.20031810770740702	5.779958811794747
etizer	5.0	(0.429074007,0.642714571]	0.23478916721501555	0.005474102418242384	0.15079183927479672	-4.350932774565497
etizer	6.0	(0.642714571,0.853658537]	0.7630433889801422	0.055731041396373145	0.49005973076014764	-14.140134866140913
etizer	7.0	(0.853658537,0.987636364]	1.2309697963446826	0.13549841745167288	0.7905824697293172	-22.8113881698447
etizer	8.0	(0.987636364,0.9999999]	1.3459658701706754	0.022185547394674075	0.8644379618173389	-24.94240721340076
etizer	9.0	(0.9999999,Infinity)	1.4594506813521857	0.34087536481198866	0.9373228551486464	-27.04542069669673

表头真名 表头别名 提示：点击单元格可查看超出的内容。注意：表头中 表示特征列，* 表示标签列 下载预览数据

图2-25 评分卡输出

评分预测

我们对数据集进行特征选择后计算得分，接入评分预测节点将得分结果输出，如图2-26所示，分值越高则违约风险越大。评分卡构建预测流程如图2-27所示。

当前显示 100 条 / 总共有 19873 条数据

# NumberOfTimes90DaysLate	# NumberRealEstateLoansOrLines	# NumberOfTime60-89DaysPastDueNotWorse	# NumberOfDependents	# score
0	3	0	1	923.8473740002283
0	1	1	0	962.0656386152801
0	1	0	2	897.8783199236022
0	1	0	0	980.3858122152925
1	0	0	0	955.1955380191106
1	0	0	0	945.277610195087
0	0	0	1	928.9095619579987
0	0	0	0	925.6965503998678

表头真名 表头别名 提示：点击单元格可查看超出的内容。注意：表头中 表示特征列，* 表示标签列 下载预览数据

图2-26 评分预测

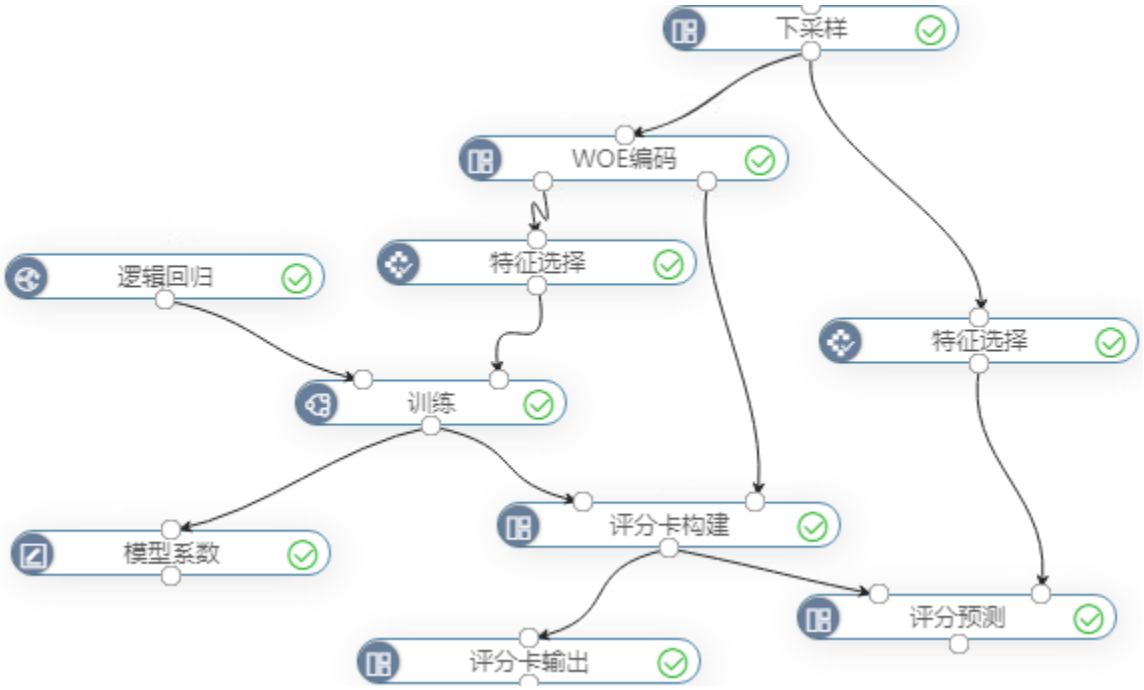


图2-27 评分卡构建及预测

总结

本章结合信用卡评分的案例，重点介绍了数据挖掘算法中逻辑回归分类算法在实际案例中的应用。本案例研究客户信用问题，从分析出客户的信用分值中可挖掘出该客户的违约风险程度，并针对违约客户的这些客群中采取相应的措施。