某商业银行客户流失预测

- 背景描述
- 现状分析
  - 业务现状
  - 数据现状
- 需求分析
- 工作目标
- 分析思路
- 实施过程
  - 数据接入
  - 数据探索
  - 数据预处理
  - 相关性分析
  - 模型训练
  - 模型评估
- 业务分析
- 总结

背景描述

现在银行产品同众化现象普遍存在，客户选择产品和服务的途径越来越多，客户对产品的忠诚度越来越低，所以客户流失已经成为银行业最关注的问题之一。而获得新客的成本远高于维护老客户成本。

根据总行领导指示，需要坚持以客户为中心，重点做好营销和转化指导方针，加强客户群体建设和提高客户与我行粘度，加强对客户的跟踪和营销管理，减少不必要的客户流失。

据调研表明，商业银行客户流失较为严重。国内商业银行，客户流失率可达20%甚至更高。而获得新客户的成本，可达维护现有客户的5倍。

因此，从海量客户交易记录中挖掘出对流失有影响的信息，建立高效的客户流失预警体系尤为重要。

现状分析

业务现状

选取近一年的零售客户某业务的高价值客群进行分析，月均新增为1. 00%，月流失率为8. 00%，高价值客户流失严重，建立流失预警机制刻不容缓。

数据现状

本次挖掘的客群为零售客户某业务线条的高价值客群，这部分客群的数据主要存放在CRM系统当中，存放了客户的个人信息、账户信息、存款类信息、消费和交易类信息、理财和基金类信息。

需求分析

通过对某业务线条的高价值客群进行分析，发现流失率非常严重，需要建立高价值客群的流失预警模型，寻找客户流失的原因，指导业务加强客户维护，提供客户对我行产品的粘度。

工作目标

通过对零售客户某业务线条的高价值客群进行流失分析，建立流失预警模型，挖掘出流失的主要因子，指导业务人员维系客户关系。

分析思路

通过收集高价值客群的个人信息、账户类信息、交易类信息等维度数据，以及结合第三方数据，利用随机森林算法构建客户流失预警模型，并输出影响客户流失的主要因素。

实施过程

数据来源于CRM系统中客户基本信息表、账单表等；第三方数据，数据时间窗为近一年的数据，客群为高价值客群，本次案例已获取到部分数据总共100000条数据。

数据维度信息包含：

- 银行自有字段：账户类信息、个人类信息、存款类信息、消费、交易类信息、理财、基金类信息、柜台服务、网银类信息；
- 外部三方数据：外呼客服数据、资产类数据、其他消费类数据；

本次案例流失定义为：3个月内没有与银行业务任何往来的客户。

本案例只提取到部分特征字段仅供参考，字段说明如下：

字段名称	类型	字段说明
客户id	字符串	
年龄	整型	
卡龄	整型	
是否代发客户	整型	
月均代发金额	整型	
最多代发金额	整型	
性别	整型	取值为{0,1}，0表示男，1表示女

月均AUM	整型	
月初AUM	整型	
卡等级	整型	取值[0,19],共有20个等级。
是否个贷	整型	
1年内购买理财	整型	
下载手机银行	整型	
是否领取APP权益	整型	
是否登录APP	整型	
是否持有信用卡	整型	
是否关联还款	整型	
是否流失	整型	取值为{1,0}，1表示流失，0表示未流失。

数据接入

在实验中添加 数据源 节点，将数据接入，部分数据输出结果如图：

当前显示 100 条 / 总共有 100000 条数据　提示:点击单元格可查看超出的内容

客户id	年龄	卡龄	是否代发客户	月均代发金额	最多代发金额	性别
183826d7-2d3e-4854-8748-97856b3c3eb5	44	5	1	1471	3389	0
7937c18c-0288-4981-89c2-e49e2b658e52	55	4	0	0	0	1
6223accb-1ca0-4cad-bb7f-b4a4b4e29568	59	19	1	4810	5951	0
0fbe3b10-4fc9-4475-a7b2-2757d55006c1	19	0	0	0	0	0
271648c5-9e76-4a7d-bc1d-f250cbc51809	45	14	1	54032	122441	1
d33f98b2-8170-460b-b3ba-d972120a4241	58	11	1	31115	233516	0
a1abace9-51d3-43ab-8fad-f6627217484b	52	0	0	0	0	1
2430b84d-0bb2-4e9a-8a3d-848d4f59a164	12	0	0	0	0	0
6e8921f1-9fb5-40a5-a22d-c620aee9cde1	41	8	0	0	0	0
ae2f0c38-4ad2-4002-ab91-7000a53f730f	68	6	0	0	0	1
492e3688-6003-4ce2-810f-5b9774d6a7be	23	9	1	84936	274059	0

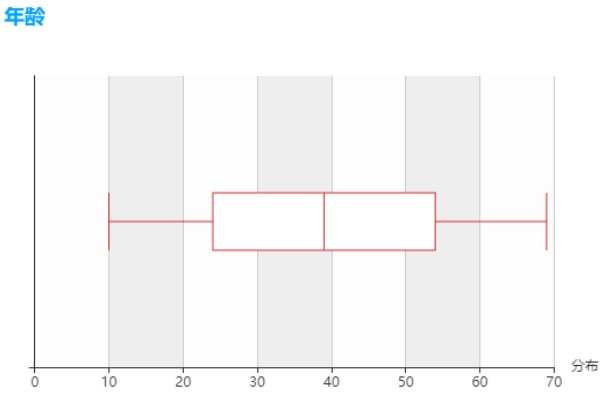
注意：表头中◇表示特征列，*表示标签列　　表头真名　表头别名

数据探索

本案例数据探索是针对客户流失数据探索分析各特征分布情况以及相关性情况。我们接入一个 全表统计 节点对流失数据进行统计分布情况，如图：

查看全量统计

列名	直方图	箱线图	最大值	最小值	平均值	标准差
年龄	查看	查看	69	10	39.45499	17.3418...
卡龄	查看	查看	19	0	7.89727	6.34545...
是否代发...	查看	查看	1	0	0.37482	0.48407...
月均代发...	查看	查看	99994	0	6690.79...	18392.4...
最多代发...	查看	查看	398944	0	24780.0...	69565.1...
性别	查看	查看	1	0	0.49707	0.49999...
是否流失	查看	查看	1	0	0.5012	0.50000...
月均AUM	查看	查看	9999821	0	751248....	2103124...
月初AUM	查看	查看	23940661	0	1810959...	5030793...
卡等级	查看	查看	4	1	1.6289	1.04930...
是否个贷	查看	查看	1	0	0.49848	0.50000...



通过全表统计分析发现年龄的最小值为10，按照常规而言未成年人无法办理银行信贷业务应该进行过滤处理。

我们通过 聚合 节点探索流失的整体数据分布情况，输出结果如图：

当前显示 2 条 / 总共有 2 条数据 提示:点击单元格可查看超出的内容

# Group_是否流失	# Count_是否流失
0	49880
1	50120

注意：表头中◇表示特征列，*表示标签列

表头真名 表头别名

流失的男女比例输出结果如图：

当前显示 2 条 / 总共有 2 条数据　提示:点击单元格可查看超出的内容

# Group_性别	# Count_是否流失
0	50293
1	49707

注意：表头中◇表示特征列，*表示标签列

表头真名　表头别名

分析发现客户流失数据和男女比例存在一个平衡的状态，分布比较合理。

我们再通过 过滤与映射 节点将流失的数据进行过滤，如图：

筛选器

输入关键字搜索列

STRING
- 客户id

LONG
- 年龄
- 卡龄
- 是否代发客户
- 月均代发金额
- 最多代发金额
- 性别

```
1 [是否流失] = 1
```

注意：在IE中，使用微软输入法输入中文会有问题，请切换输入法。

确定　取消

通过 聚合 节点查看男女的流失比例，输出结果如图：

当前显示 2 条 / 总共有 2 条数据　提示:点击单元格可查看超出的内容

# Group_性别	# Count_是否流失
0	25108
1	25012

注意：表头中◇表示特征列，*表示标签列

表头真名　表头别名

数据预处理

通过数据探索分析中我们需要过滤出年龄>18的流失数据，如图：

筛选器

输入关键字搜索列

STRING
- 客户id

LONG
- 年龄
- 卡龄
- 是否代发客户
- 月均代发金额
- 最多代发金额
- 性别
- 是否流失

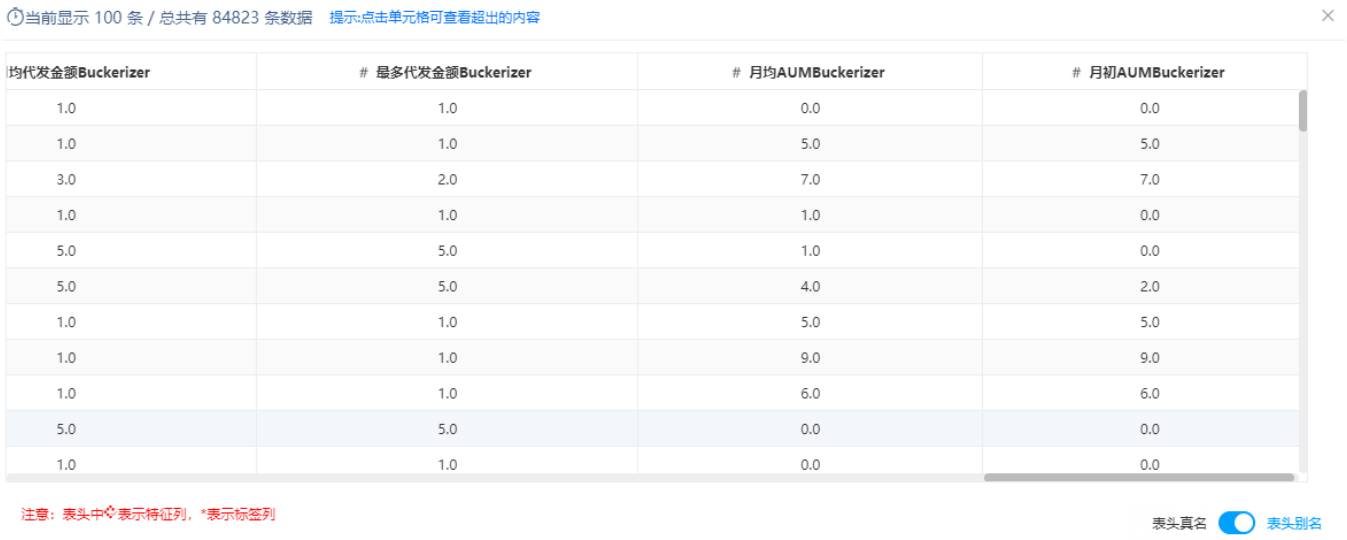
```
1 [年龄] > 18
```

注意：在IE中，使用微软输入法输入中文会有问题，请切换输入法。

确定　取消

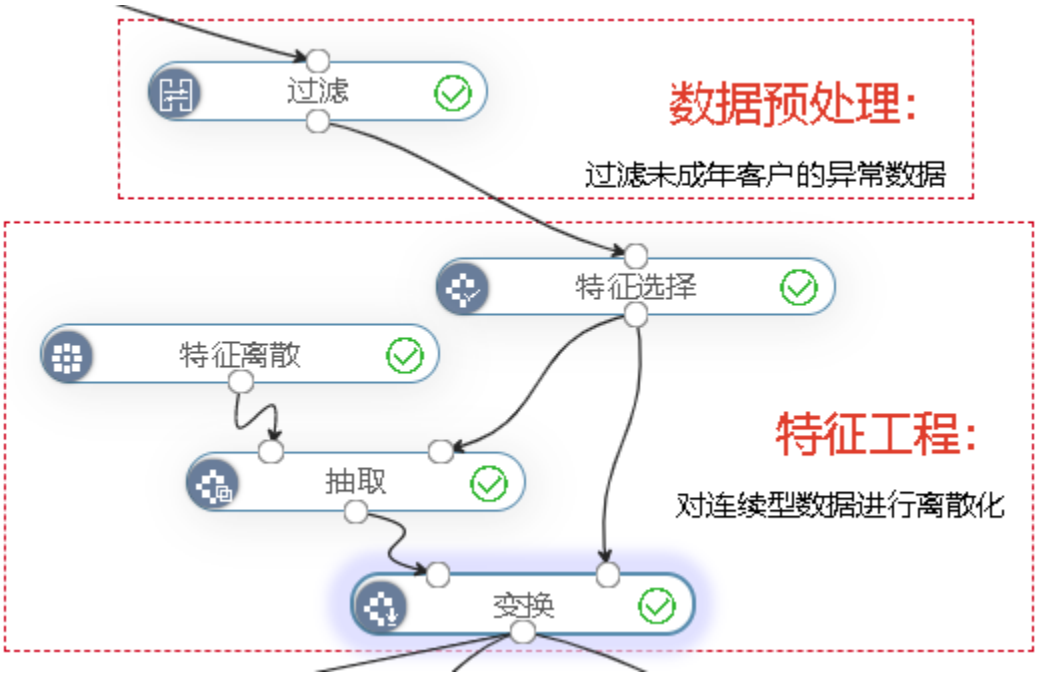
我们根据特征选择出连续性特征字段进行 特征离散，方便模型的准确程度，如图所示：

变换后的结果默认添加后缀Buckerizer，如图所示：



均代发金额Buckerizer	# 最多代发金额Buckerizer	# 月均AUMBuckerizer	# 月初AUMBuckerizer
1.0	1.0	0.0	0.0
1.0	1.0	5.0	5.0
3.0	2.0	7.0	7.0
1.0	1.0	1.0	0.0
5.0	5.0	1.0	0.0
5.0	5.0	4.0	2.0
1.0	1.0	5.0	5.0
1.0	1.0	9.0	9.0
1.0	1.0	6.0	6.0
5.0	5.0	0.0	0.0
1.0	1.0	0.0	0.0

整个数据预处理流程图：



相关性分析

我们通过相关性节点将各特征指标数据进行 相关性分析，方便特征选择进入模型训练，如图：

查看分析结果

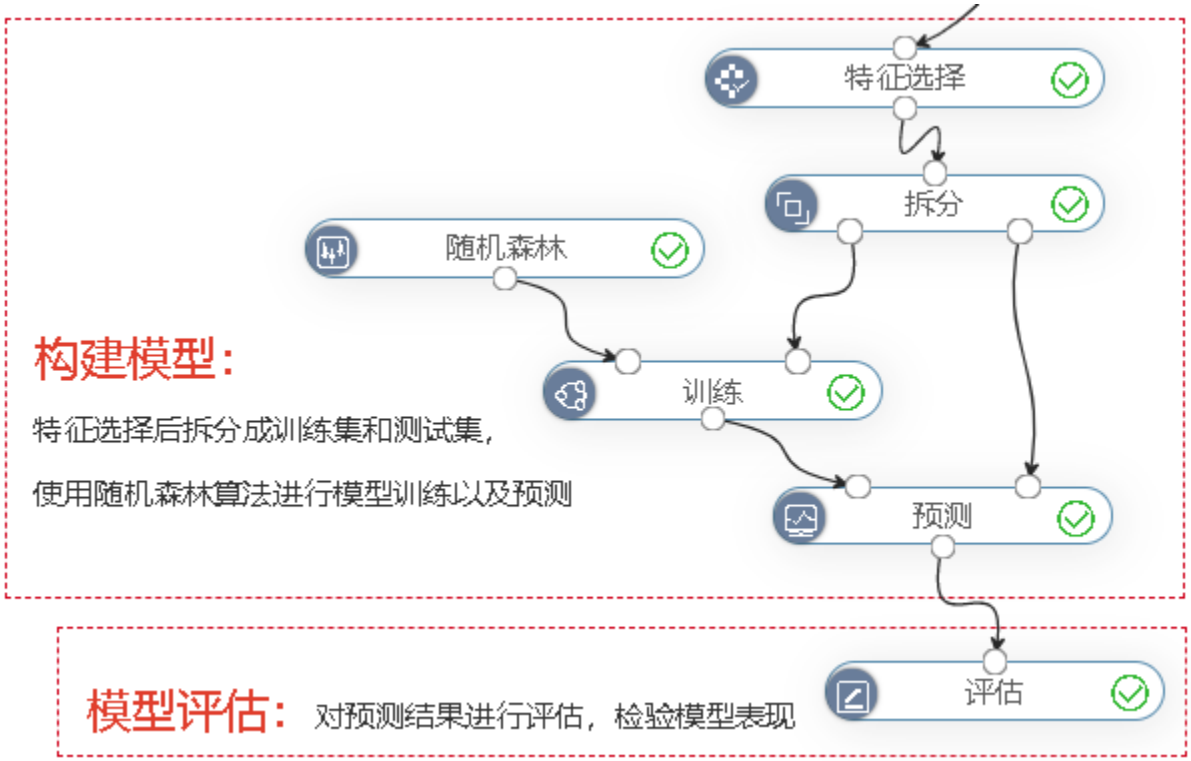
-1.00 1.00

	性别	是否流失	卡等级	是否个贷	1年内购买理财	下载手机银行	是否领取App权...	是否登录App	是否持有信用卡	是否关联还款	年龄Bucker...	卡龄Bucker...	月均代发金额Bu...	最多代发金额Bu...	月均AUMBuc...	月初AUMBuc...	是否代发客户
是否代发客户	-0.00	0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.00	-0.00	-0.00	-0.12	0.03	0.80	0.79	-0.01	-0.01	1.00
月初AUMBuckerizer	-0.00	-0.82	0.77	0.00	0.00	0.00	0.00	0.00	-0.00	-0.01	0.00	0.01	0.03	0.03	0.96	1.00	-0.01
月均AUMBuckerizer	-0.00	-0.81	0.77	0.00	0.00	0.00	0.00	0.00	-0.00	-0.01	0.00	0.00	0.07	0.07	1.00	0.96	-0.01
最多代发金额Buckerizer	-0.00	0.01	0.17	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.09	0.02	0.93	1.00	0.07	0.03	0.79
月均代发金额Buckerizer	-0.00	0.01	0.16	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.09	0.02	1.00	0.93	0.07	0.03	0.80
卡龄Buckerizer	-0.00	-0.00	0.00	0.01	0.01	-0.00	-0.01	-0.00	0.00	-0.00	0.06	1.00	0.02	0.02	0.00	0.01	0.03
年龄Buckerizer	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.01	0.00	-0.00	1.00	0.06	-0.09	-0.09	0.00	0.00	-0.12
是否关联还款	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.58	1.00	-0.00	-0.00	-0.00	0.00	-0.01	-0.01	-0.00
是否持有信用卡	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	1.00	0.58	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00
是否登录App	-0.01	-0.00	-0.00	0.00	-0.00	0.77	0.45	1.00	0.00	0.00	-0.01	-0.00	-0.00	0.00	0.00	0.00	-0.00
是否领取App权益	-0.00	-0.00	-0.00	-0.00	0.00	0.58	1.00	0.45	-0.00	-0.00	-0.00	-0.01	-0.00	-0.00	0.00	0.00	-0.01
下载手机银行	-0.01	-0.00	0.00	-0.00	-0.00	1.00	0.58	0.77	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00
1年内购买理财	0.01	-0.00	0.00	-0.00	1.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.01	-0.00	-0.00	0.00	0.00	-0.00
是否个贷	0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.01	-0.00	-0.00	0.00	0.00	-0.00
卡等级	-0.00	-0.60	1.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.16	0.17	0.77	0.77	-0.01
是否流失	0.00	1.00	-0.60	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.01	0.01	-0.81	-0.82	0.01
性别	1.00	0.00	-0.00	0.00	0.01	-0.01	-0.00	-0.01	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00

通过分析发现：是否代发客户、卡等级、月均代发金额、最多代发金额、月均AUM、月初AUM与是否流失都具有相关性，其他特征与是否流失相关性为0。因此我们通过特征选择出具有相关的特征，如图所示，标签列为是否流失。

模型训练

本案例采样 随机森林 算法进行模型训练，通过 拆分 节点将数据按照比例7:3拆分成训练集和验证集。整个模型训练流程如图所示：



参数配置如图：

模型评估

我们通过 评估 节点对数据进行评估，如模型训练流程图所示，评估结果如图：

指标	值		
confusion matrix(混淆矩阵)	真实|预测	0	1
	0	11703	825
	1	318	12444
accuracy(准确率)	0.9548042704626335		
roc曲线	查看ROC曲线		
auc	0.9690043714397567		
ks	查看KS曲线		
weighted precision(加权精确率)	0.9555203911895991		
weighted recall(加权召回率)	0.9548042704626334		
weighted F1 score(加权F1分数)	0.9547777000125989		
Class 0.0 precision(精确率)	0.9735462939855253		
Class 0.0 recall(召回率)	0.9341475095785441		
Class 0.0 F1 score(F1分数)	0.9534400586581938		
Class 0.0 falsePositiveRate(假阳率)	0.024917724494593323		
Class 0.0 truePositiveRate(真阳率)	0.9341475095785441		
Class 1.0 precision(精确率)	0.9378250056522722		
Class 1.0 recall(召回率)	0.9750822755054067		
Class 1.0 F1 score(F1分数)	0.9560908147977412		

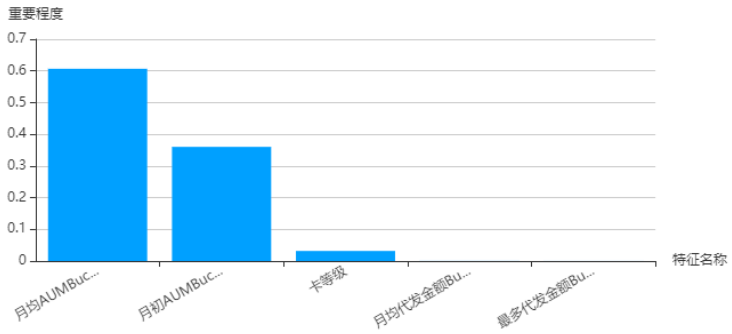
我们发现评估结果中F1得分为0.95，说明模型预测的效果比较好的。

业务分析

我们通过 随机森林特征选择 节点输出重要性较高的5个特征，结果如图：

查看分析结果

特征名称	重要程度
月均AUMBuckerizer	0.6066567076914587
月初AUMBuckerizer	0.3602443530594919
卡等级	0.03178563444406004
月均代发金额Buckerizer	0.0007770334956873465
最多代发金额Buckerizer	0.0005362713093019175



通过对某业务线条高价值客群进行流失预警分析，发现影响客户流失的主要因素为：月均AUM、月初AUM、卡等级等。主要原因可能为产品缺乏竞争力、活动较少等。

因此，我们可以采取相关的措施建议，如：加强客户关系维系、产品跟进、维护访问、追踪制度、扩大销售、机制维护等。

总结

本案例结合银行客户流失数据预测案例，重点介绍了随机森林在实际案例中的应用。本案例通过客户的交易信息数据挖掘出对流失影响的信息，从而加强对客户的跟踪和营销，减少不必要的客户流失。