

Teradata™ Warehouse Miner Training Workshop

案例-1 客户流失预测分析

CMCC HQ

ZhangLei

Sunstone.Zhang@TERADATA-NCR.COM

You've never seen your business like this before.

专题概要

- 为什么进行客户流失预测？
- 移动效劳的竞争进入白热化状态
- 月平均流失率6.5% (即随机抽样10000人中只能找到650个真正有流失倾向的客户)
- 月租型，流失预测模型结果
- 找出流失倾向最高的客户，
- 前10000人名单中，约8000人〔80%〕下个月会流失
- 找出导致客户流失的原因以辅助设计和执行客户挽留的营销活动。
- 数据挖掘自动化机制更有效地利用挖掘的结果

分析模型的种类



根据业务目标
对对象属性的
初步划分

年龄分组；

商业用户/
个人用户；

客户价值

根据业务目标
对对象属性的
描述

保有期(与客
户价值有关)；

信用评分

根据相关属性
将对象划分到
已定类别，以
便针对性对待

流失倾向评分

使用量预测

现实模拟以改
善预测能力、
可控性和实施
效果

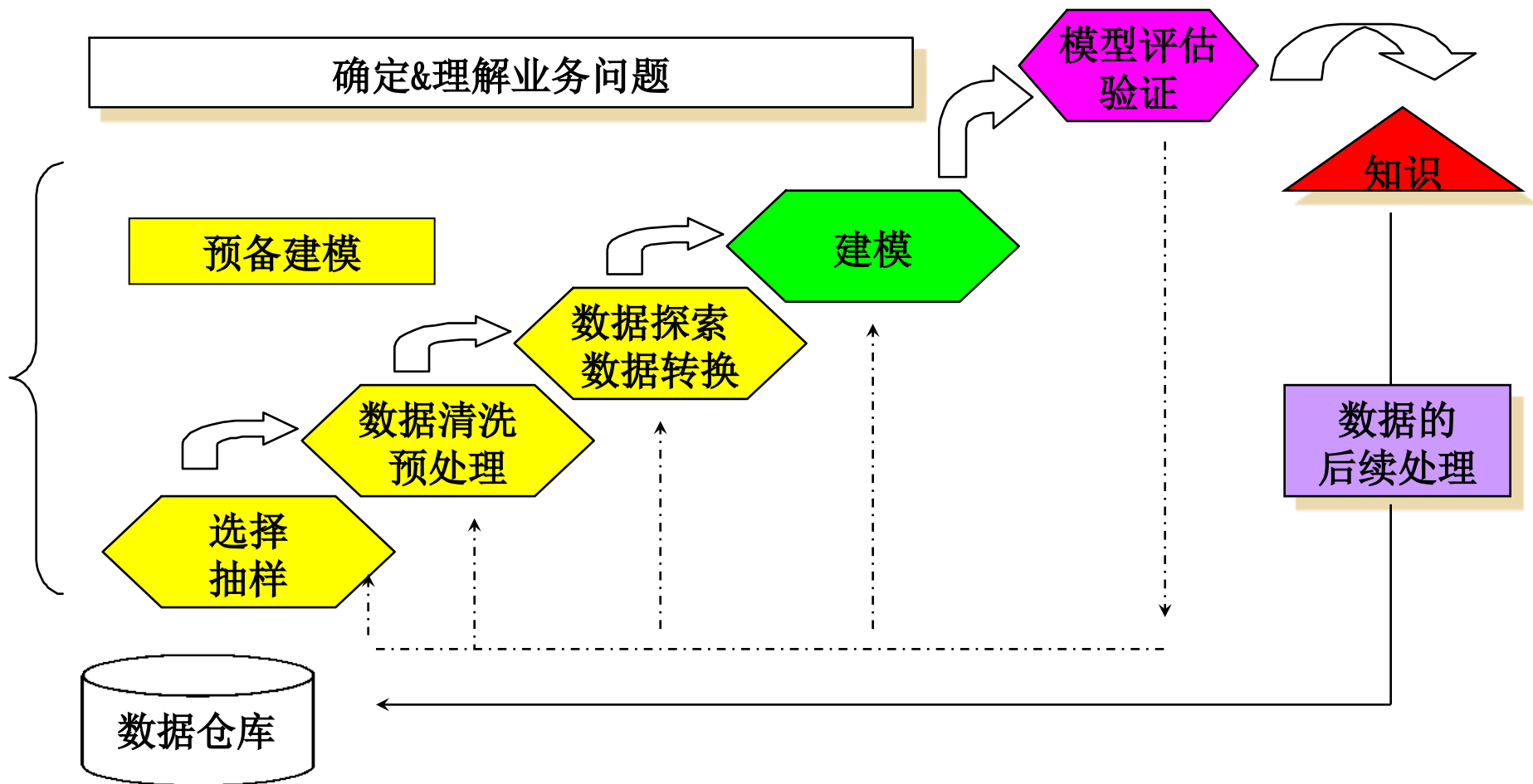
购置倾向评分

挖掘结果的发
布：

调度执行
收集响应
监控模型性能
记录结果



数据挖掘方法论



• Partly adapted from Fayyad U. M., Piatetsky-Shapiro, P. Smith (1995), 'From Knowledge Discovery to Data Mining: An Overview' in Advanced Knowledge Discovery and Data Mining.

业务问题定义

在这个数据挖掘的初始阶段，需明确阐述工程目标和客户业务需求。基于客户响应（如客户流失或产品购置趋势）特性，可以从概念上定义响应变量，与待选的预测变量没有直接的派生关系（直接的派生关系意指响应变量可由一或多个预测变量直接计算出来）。最终，为了这些目标可以必要地调整工程方案进程。包括

明确业务目标

定义响应变量

工程方案必要的调整

选取和抽样

搜寻并检查客户数据，创立一个数据映射概念图，将客户数据与建模相关的各个属性对应起来。数据能被整合到一个适当的程度，省略不适当的记录（如假设分析仅针对居民客户，那么省略非居民客户）、不完整的数据记录、训练数据、试验数据等等。包括：

数据来源

数据映射

准备数据评估

数据的必要聚合

数据抽样

探索型数据分析

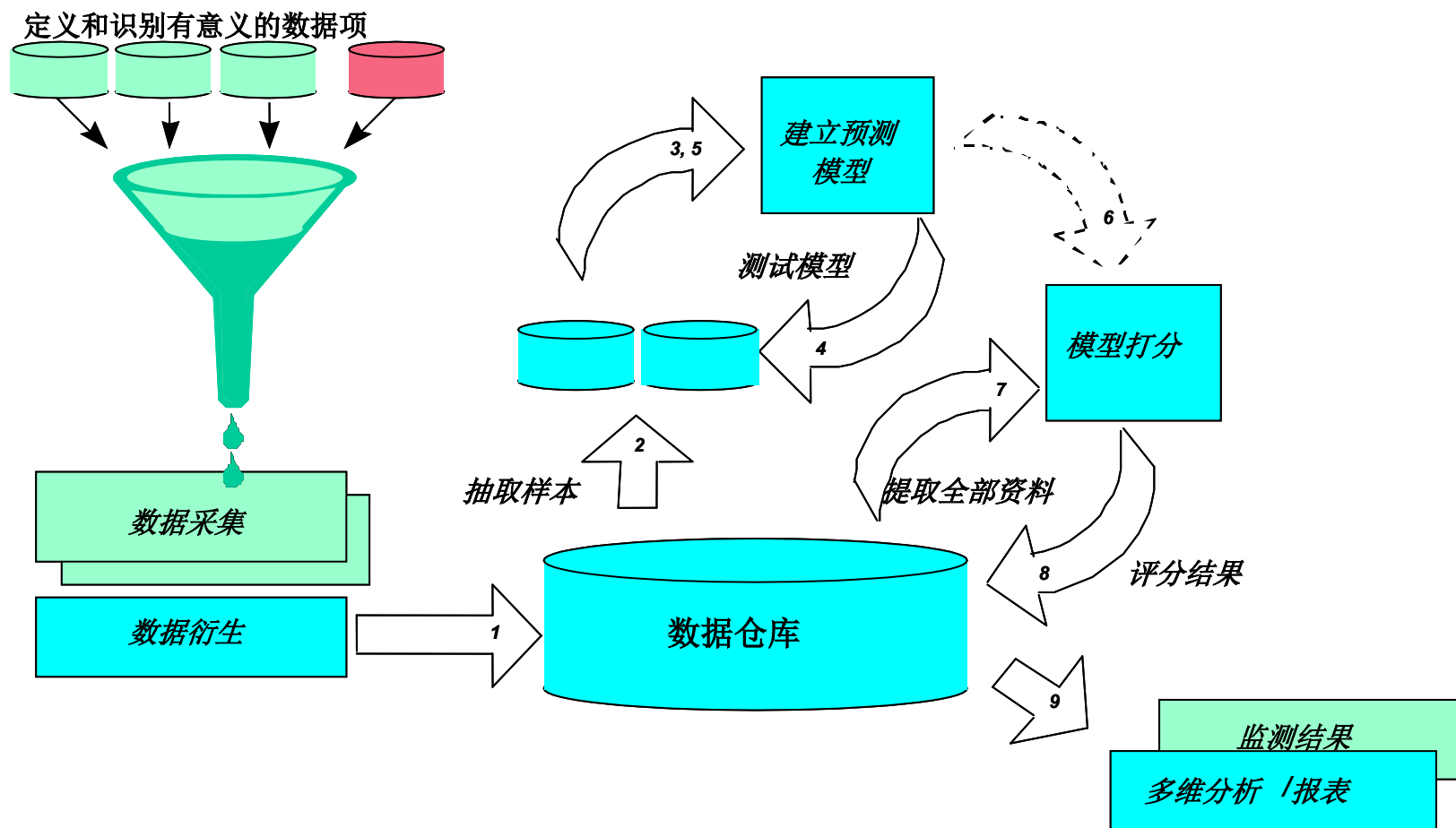
核查目前的数据源，探索在每个待选的预测变量和响应变量之间是否存在关系。数值分析是全面理解数据的第一步，随之进行的统计分析便于更好地了解有关数据的分布。包括：

- 数据质量检查
- 数据的必要整理
- 通过图形化呈现工具和其他的统计方法理解数据
- 分析待选预测变量和响应变量之间的关系
- 数据转换以辅助数据的分析
- 数据派生为建立模型做准备
- 整理和呈现数据探索的发现

建模

实施

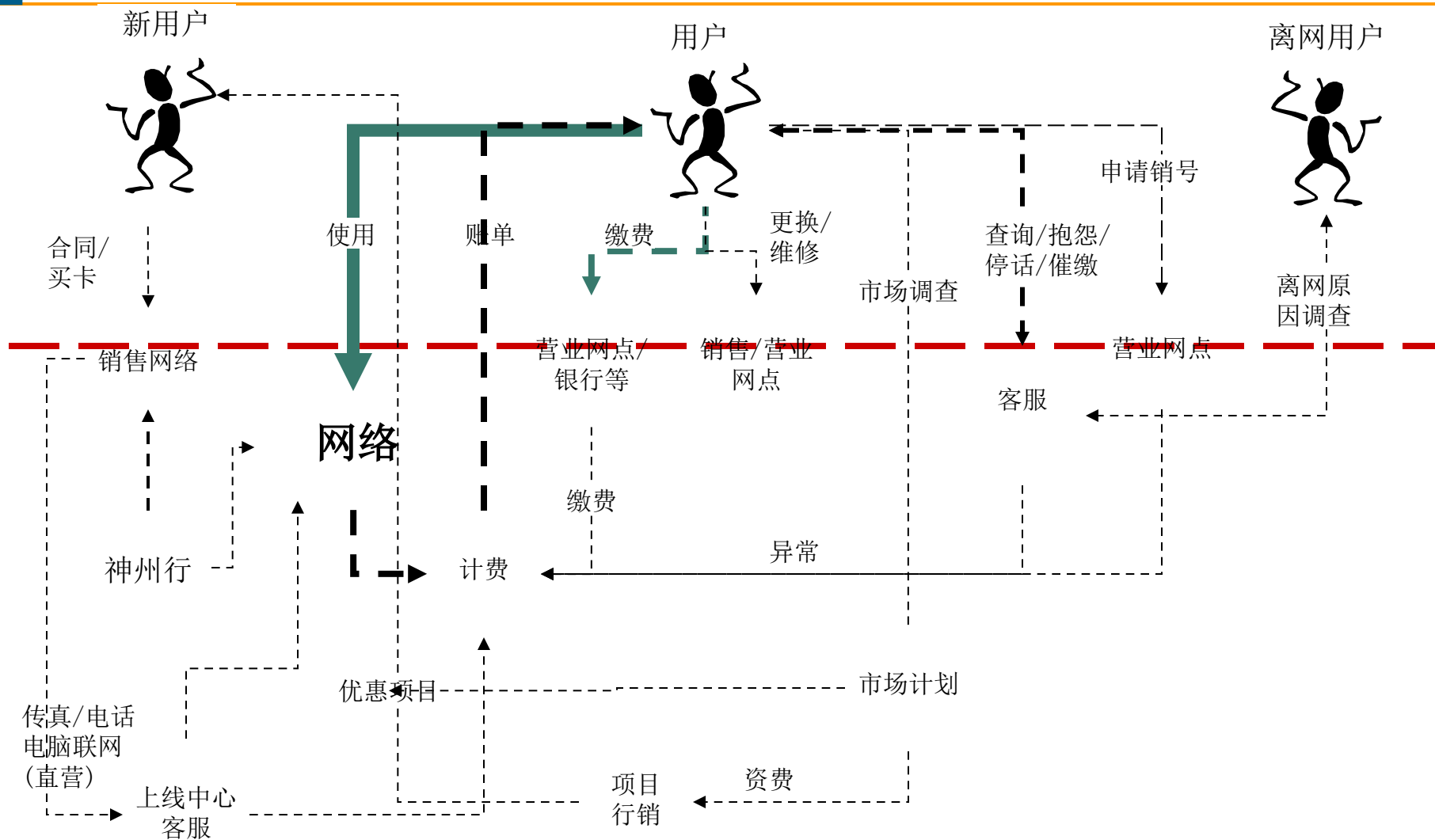
具体实施过程



客户流失的定义

- 以号码为单位
- 用户在网时间至少3个月
- 分析流失前6个月的数据以了解可能流失原因
- 正常用户
- 状态为‘正常’ 或 ‘停话’ 或 ‘强开’
- 假设状态为停话，停话原因属于自愿停话：
 - 报停双停 报停单停 报停完全
 - 挂失双停 挂失单停 挂失完全
- 假设状态为停话，停话原因属于非自愿停话：
 - 冒高双停 冒高单停 欠停双停
 - 欠停单停
 - 强关双停 强关单停 强关完全 预拆
- 最近停话期间少于3个月

业务理解



分析月租型流失用户的数据来源



数据质量问题

所需的分析数据	数据质量问题	原因
<ul style="list-style-type: none">□ 六个月通话的移动差□ 六个月通话的移动平均	缺乏数据	数据仓库中只有 2002 年 9, 10, 11, 12 月 4 个月的话单数据
<ul style="list-style-type: none">□ 已兑换次数□ 兑换类别□ 已兑换分数	缺乏数据	积分奖励计划从 2002 年 12 月才开始
<ul style="list-style-type: none">□ 催缴次数□ 催缴成功次数	缺乏数据	数据仓库中没有催缴记录
<ul style="list-style-type: none">□ 查询销号离网次数	缺乏数据	数据仓库中没有关于该项查询的分类
<ul style="list-style-type: none">□ 通话异常中断次数	异常数据	174 万客户在过去 3 个月内通话全部正常结束
<ul style="list-style-type: none">□ 客户状态变化	缺乏数据	数据库中只有 2002 年 11 月 20 日之后的客户状态连续变化的记录

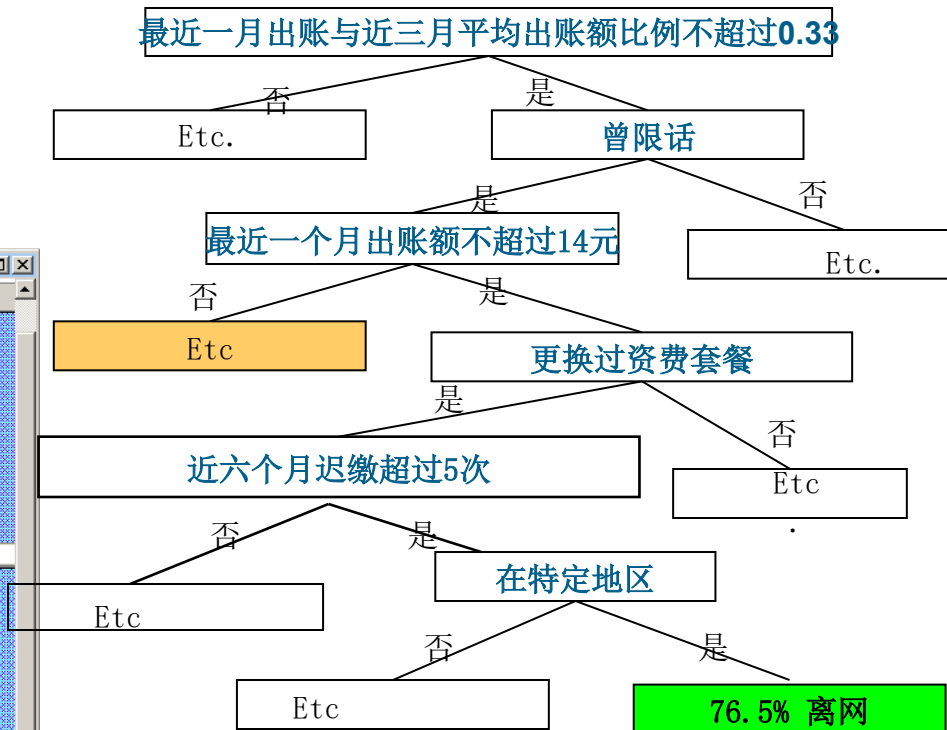
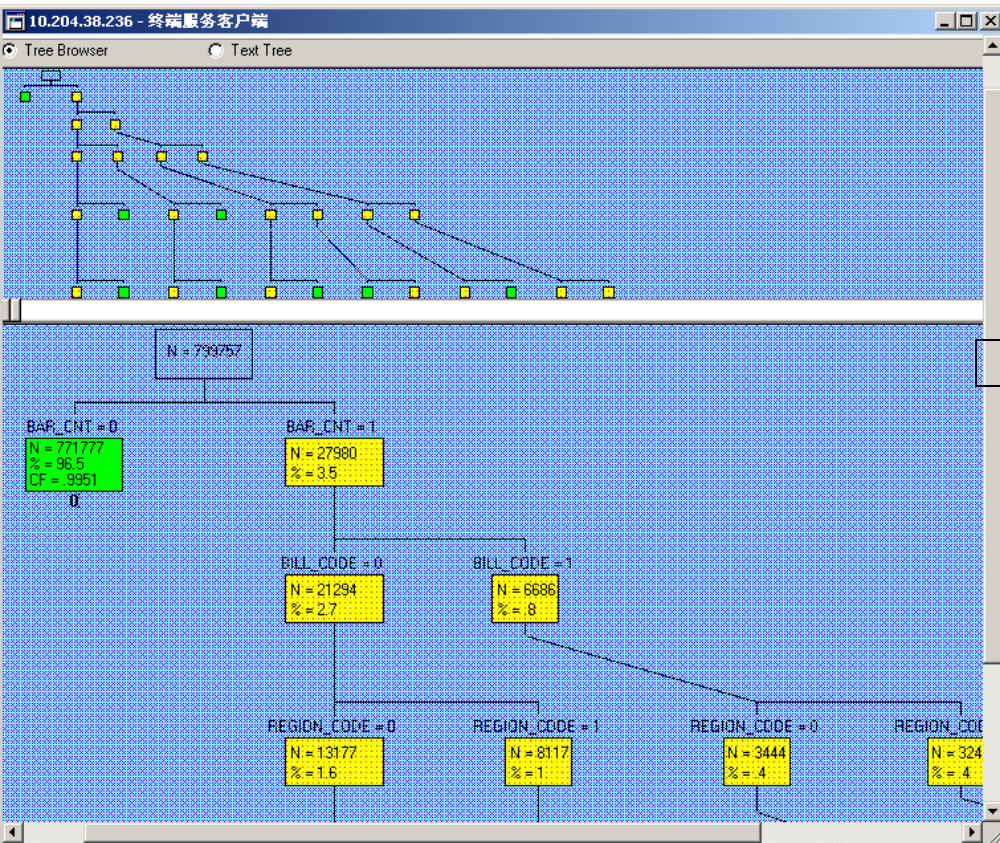
流失客户特征...

- 1 50元月租全球通, 20元月租全球通, 20元月租套餐, 20元月租亲情卡, 事发, 0元月租惠通卡, 30元月租套餐等资费套餐的客户。
 - 有过资费调整的客户。
 - 在J地区, S地区, X地区, JZ地区, Y地区, L地区的用户流失倾向较高。
 - 年龄不超过30岁的年轻族群流失比例高。
 - 最近一个月缴费额不超过前两个月平均缴费额
 - 在网时间少于17个月
 - 六个月内迟缴次数超过5次

流失客户特征（续）

- 近六个月内有限话记录的客户。
- 被叫次数不超过主叫次数的34%。
- 近三个月主叫固网时长比例不超过8%或超过80%。
- 近三个月内不曾不超过2个
- 三个月内被叫次数不超过20次。
- 三个月内主叫次数不超过45次。
- 三个月内被叫不同号码不超过17个
- 三个月内主叫不同号码不超过11个
- 三个月内主叫秒数不超过2465秒
- 三个月内发送短消息不超过2个

模型结果



模型评分



最近一月出账与近三月平均出账额比例不超过0.33

否

是

Etc.

曾限话

最近一个月出账额不超过14元

是

否

否

是

Etc

更换过资费套餐

Etc.

近六个月迟缴超过5次

是

否

否

是

Etc.

Etc

在特定地区

否

是

Etc.

76.5% 离网

- 最近一月出账与近三月平均出账额比例不超过0.33
- 曾限话
- 最近一个月出账额不超过14元
- 更换过资费套餐
- 近六个月迟缴超过5次
- 在特定地区

得分: 0.765分



最有可能销号的客户群组 #1

最近一个月缴费额不超过前两个月平均缴费额
曾限话

最近一个月出账金额不超过14元

曾更换资费方案

六个月内迟缴5次以上

属于〔J地区，S地区，X地区，
JZ地区，Y地区，L地区〕

资费属于低档普通套餐

近三个月被叫次数不超过20次

最近一个月缴费减少
曾限话
低档资费套餐
经常迟缴
更换过资费方案

最有可能销号的客户群组 #2

最近一个月缴费额不超过前两个月平均缴费额

曾限话

最近一个月出账金额不超过14元

曾更换资费方案

不属于〔J地区，S地区，X地区，
JZ地区，Y地区，L地区〕

通过直属营业厅入网

六个月内迟缴5次以上

近三个月被叫不同号码数不超过11个

欠费金额在-246.37和1431.87之间

被叫费用不超过13.67元

低档资费套餐

更换过SIM卡

最近一个月缴费减少

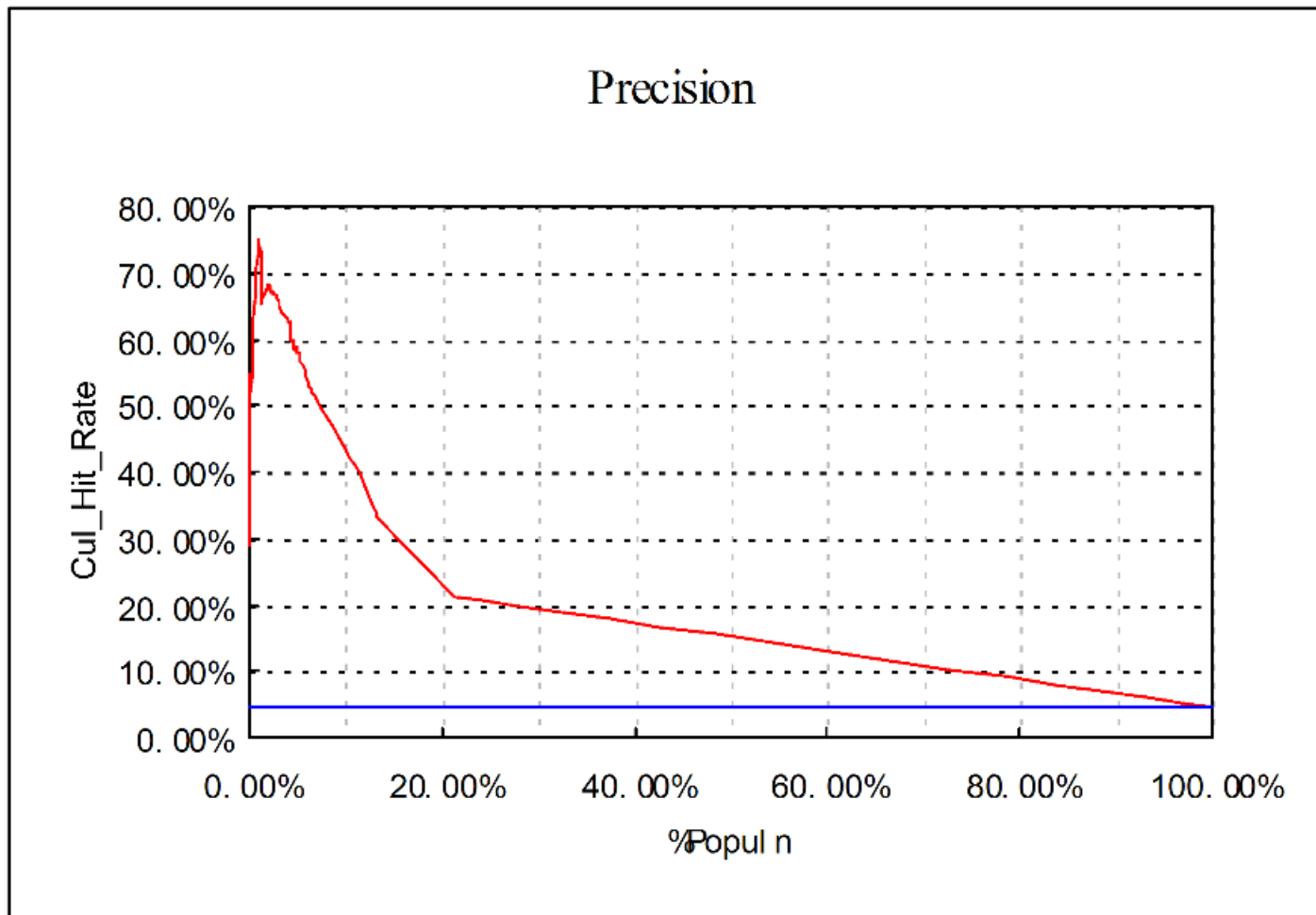
曾限话

曾更换资费方案

普通低档资费套餐

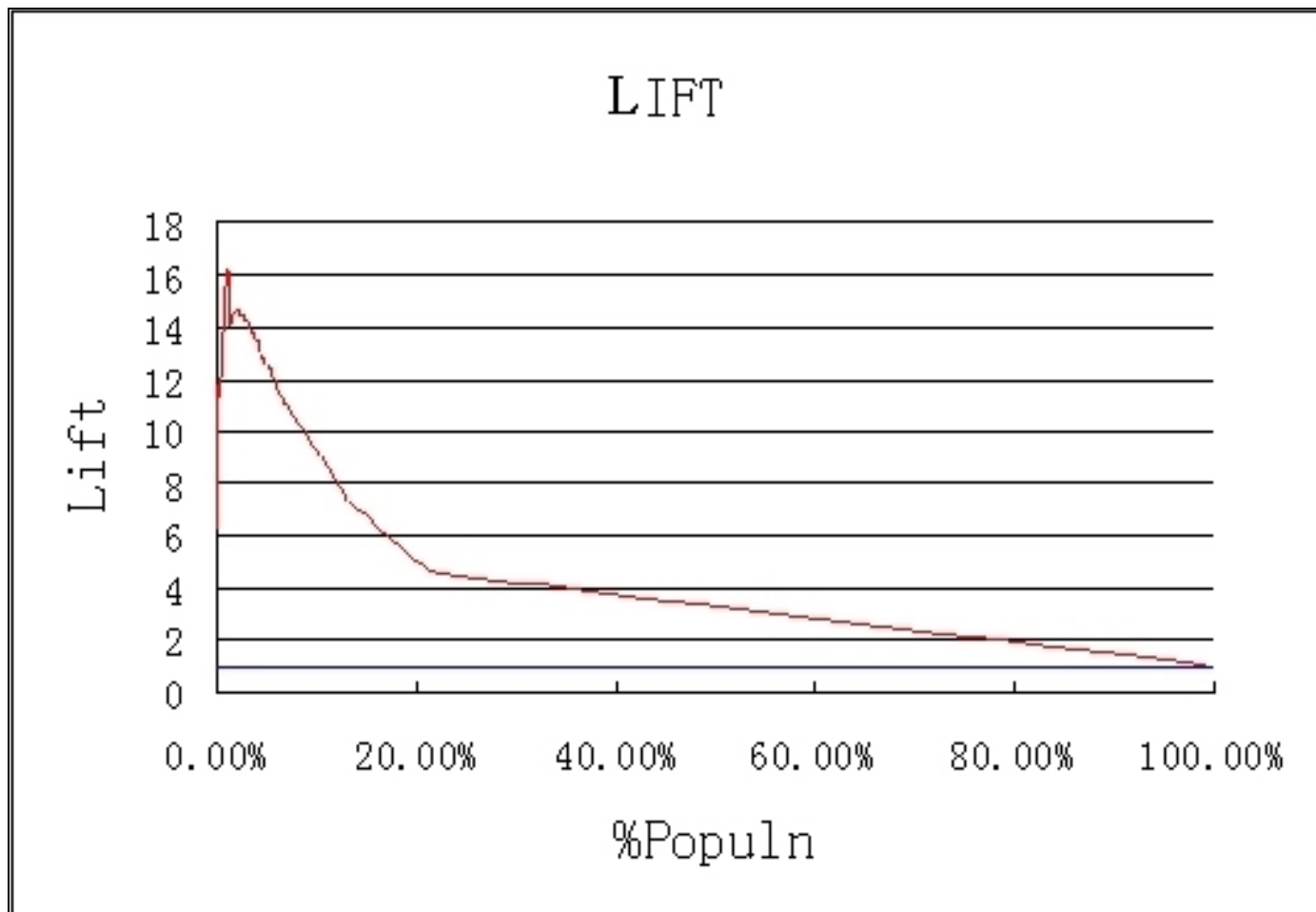
缴费行为较差

模型评估 — 命中率



X轴：按离网倾向评分从大到小排序后的客户占目标客户人数的百分比；
Y轴：前x%的客户中被准确预测为离网的客户占这批客户的百分比，即命中率

模型评估 — LIFT



X轴：按离网倾向评分从大到小排序后的离网客户占离网总人数的百分比；
Y轴：（前x%的客户中的命中率）/离网率

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/557010020062010003>