数据挖掘提纲

- 数据挖掘技术和开发流程说明
- 数据分析案例——额度统一管理

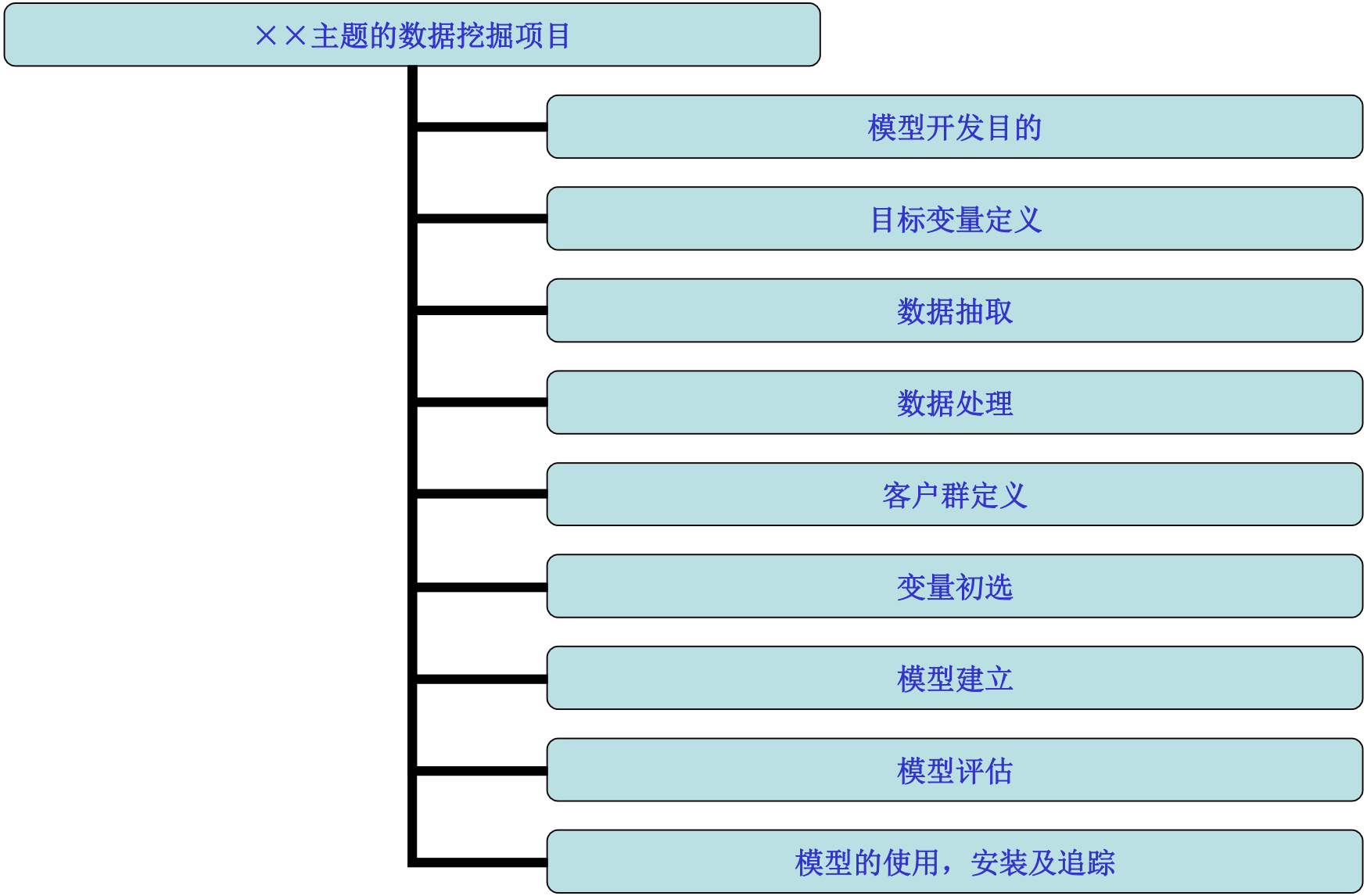
- 历史客户数据提取需求的确定

项目出发的出发点是运用过去的行为模式来预测将来的行为模式。这里需要根据数据基本信息，与相关方面的管理人员进行交流沟通，确定构建数据挖掘模型所需的客户样本、时间段和变量的选择。

- 需要资源支持

有关方面根据确定的数据提取需求提供数据及相关的说明资料。

项目开发的流程



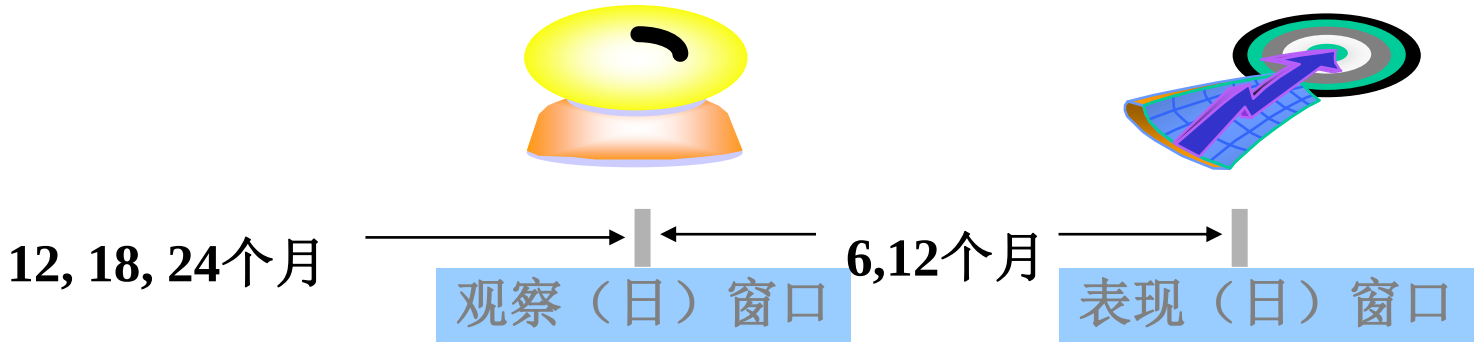
项目的需求和条件

- 业务需求
  - 详尽的数据字典，包括数据表的功能、逻辑关系
  - 数据流的生成图，以及数据表的存储方式与维护特点
  - 现有业务处理流程图
  - 信用卡产品特征指标
  - ……
- 数据需求
  - 客户/帐户/卡片信息表/关系表
  - 客户申请信息（年龄、性别、财产等）
  - 月帐表/月帐明细表（余额、消费、提现等）
  - 利润数据表（利息、各种费用、罚款等）
  - 产品线信息/汇率表
  - ……
- 数据质量评估
  - 一定规模的个人客户的数量
  - 一定时间长度的历史数据的积累

项目所涉及的技术问题

- 样板选择（Sample bias 样本偏差：拒绝推断）
- 模型选择（Logistic，linear ， classification tree, Linear programming ,Neural network, Generic Algorithms, Expert system）
- 变量筛选（IV，VARCLASS， Multi-collinearity, Stability checking）
- 变量转型transformation(函数，Classing..)
- 参数拟合(direction, significance)
- 结果效验（KS， Lorentz，Lift：Cross-Sample, out-of time）
- 安装检查（开发/实施测试）
- 表现追踪（KS， Lift， Freq， PSI。。。）

观测窗口和表现窗口图示



- **表现窗口是需要预测的——目标变量。**

  申请评分的表现窗口通常是账户设立后6- 18个月

- **观察窗口是用来预测的——表现变量**

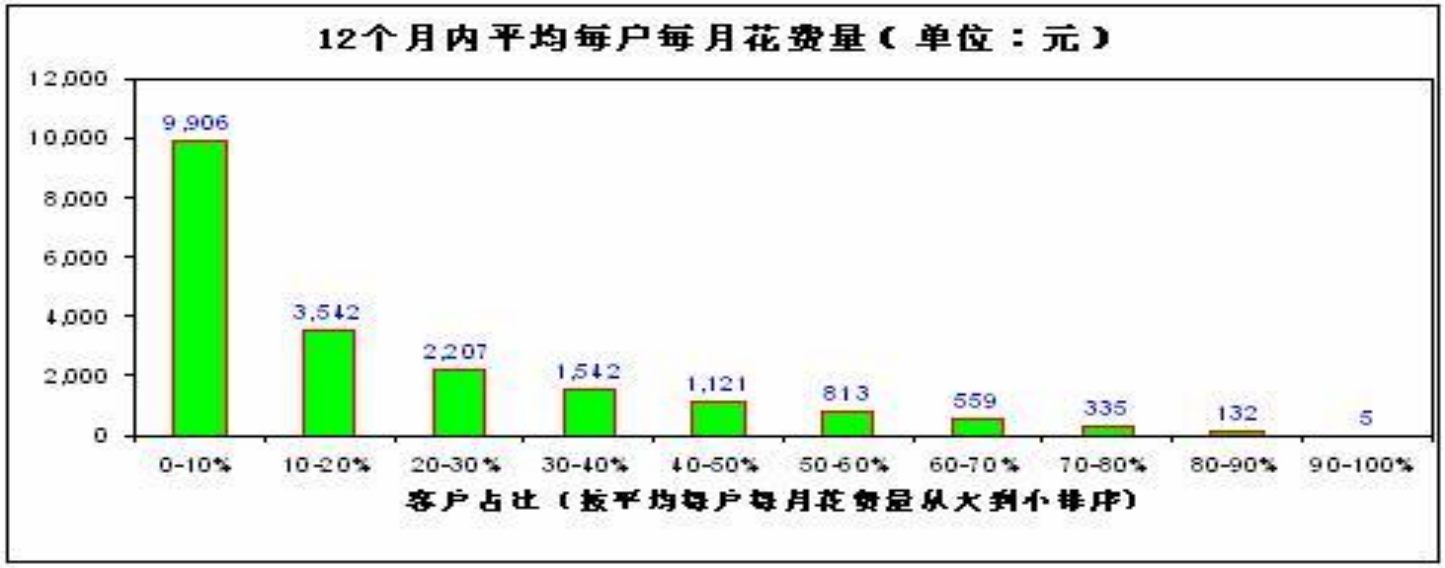
目标变量的确定

- 工作范围界定

  经验证明，能否将预测模型成功地应用到管理和生产中，在很大程度起决于我们对Y变量的定义。

- 目标变量制定原则：业务主导与数据分析相结合
  - 符合业务的要求，与开发目标相符合
  - Good/Bad客户具有较好的区别
  - 稳定性
- 定义入组标准/排斥标准
- 账户表现
  - “好”客户－未来12个月内，没有拖欠，拖欠不超过30天？
  - “坏”客户－未来12个月内，拖欠60+天，90+天，180+天？
  - Indeterminate－不在好坏的客户定义之内

例:花费预测目标变量（Y）的确定



确定花费预测的目标变量为高花费的可能性，取值1为是，0为否。然后需要对高花费作出分析及标准确定，参见上图客户花费情况。图中显示，12个月内平均每户每月花费量超过1500元的占40%左右，而超过2000元的占30%左右，两者都符合Logistic回归的要求作为Y＝1，但由于考虑到有大量的不活动客户，所以经过客户的同意，把超过2000元的行为设成Y＝1

本阶段的主要工作是诊断获取的数据，保证分析数据的质量达到分析标准，具体包括：

- 数据文件格式转换，将原始数据转为SAS数据集；
- 初步诊断原始数据，主要包括所提取的客户数、帐户数、子帐户和数据的起止时间是否符合数据提取要求；
- 检验每一个变量的数值是否合理，对分类变量进行频率分析，看其分布是否合理；对连续变量作单变量分析，计算最小值、最大值、均值、方差等。确定数据没有异常；
- 针对发现的问题与有关方面管理人员进行沟通，找出解决方案；
- 制定数据清理规则，清洗原始数据，合理处理异常值和缺失值等；
- 转换、生成新的分析变量，组合生成多种建模需要的包含复合信息的新变量。

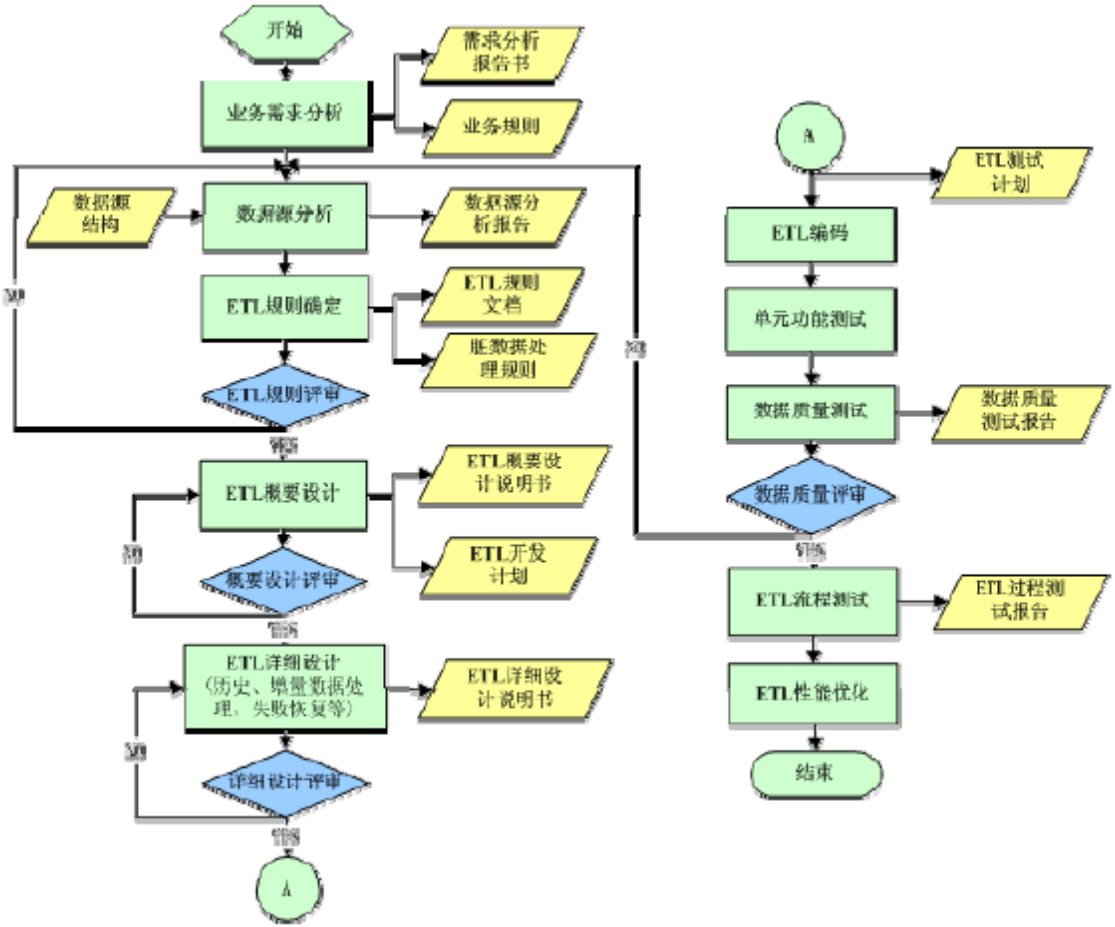
数据抽取与数据处理

对目标群体，全部抽取或随机抽取部分数据，进行如下处理：

- 数据检测（Data Exploring）
- 数据清理（Data Cleaning）
- 数据整合 （Data Integrate）
- 中间变量生成（Attributes）

单个评分卡需要的数据量问题：

- 变量系数的估计、构建稳健的模型等多方面原因均需要一定量的样本
- 没有明确的标准
- 同时需要注意BAD的个数



- 利用分类建立同质子群
- 分类基于以下考虑
  - 业务用途
  - 政策手册
  - 经验, 尝试和测试
  - 数据样本大小

观测变量初选（X）

- 工作范围界定

  分析每一个X变量与Y变量之间的关系，找出能够产生合理关系的X变量集，这些变量准备进入发展模型过程。

- 在变量较多的情况下，需要对变量进行初步筛选。

  利用“信息价值”搜索贡献最高的变量:

  $IV=\sum（G\%-B\%）\times LN（G\%/B\%）$

- 搜索最能鉴别未来对银行风险或利润贡献大小的变量
- 变量在业务上可解释，同时未来在系统上实施具有可行性
- 生成的变量集供建模使用

- **方法选择（Algorithm Selection）**

  逻辑斯蒂回归 （Logistic Regression）

  线性回归（Linear Regression）

  分类树 （Classification Tree）……

- **业界的标准：Logistic Regression**

  利用Logistic回归方法探寻存在于Y变量和X变量集之间的函数关系，包括函数形式，关键X变量和相应的参数（权重）。由于可能进入模型的X变量数量可能高达几百个，我们必须利用有效的迭代搜寻方法，将大量的X变量逐渐压缩到最具备稳定预测功能的十几个变量。其益处是：

  - ✓ 保持模型预测功能的在不同时间面对不同客户的稳定性；
  - ✓ 便于将模型植入计算机系统；
  - ✓ 降低管理成本；
  - ✓ 增加运算速度。

- **工作范围界定**

  对于经过反复迭代最后确定的模型，验证它们在具体应用时的稳定性。

  我们将进行两种测试：

  ✓ 对发展模型数据的验证；

  ✓ 对另外选取的不同客户不同时间数据的验证。

- **有效性检验**

  对目标变量有比较好的预测性

- **可靠性检验**

  对开发样本和校验样本（同时区，跨时区）的预测结果不应出现明显差异

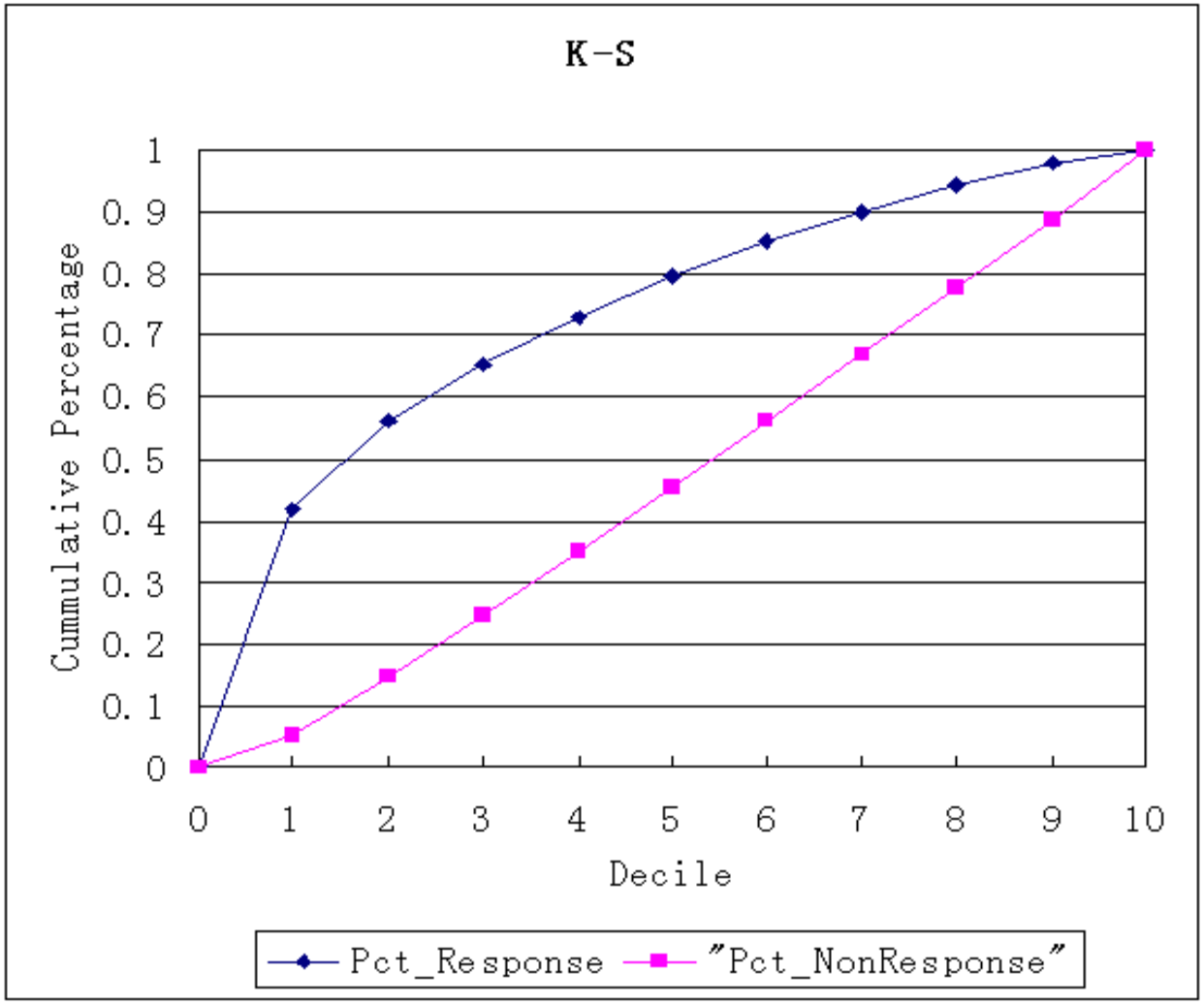
- **一致性检验**

  在模型开发样本所定义的时间段上和在交叉检验的样本所定义的时间段上应该有相似的预测能力

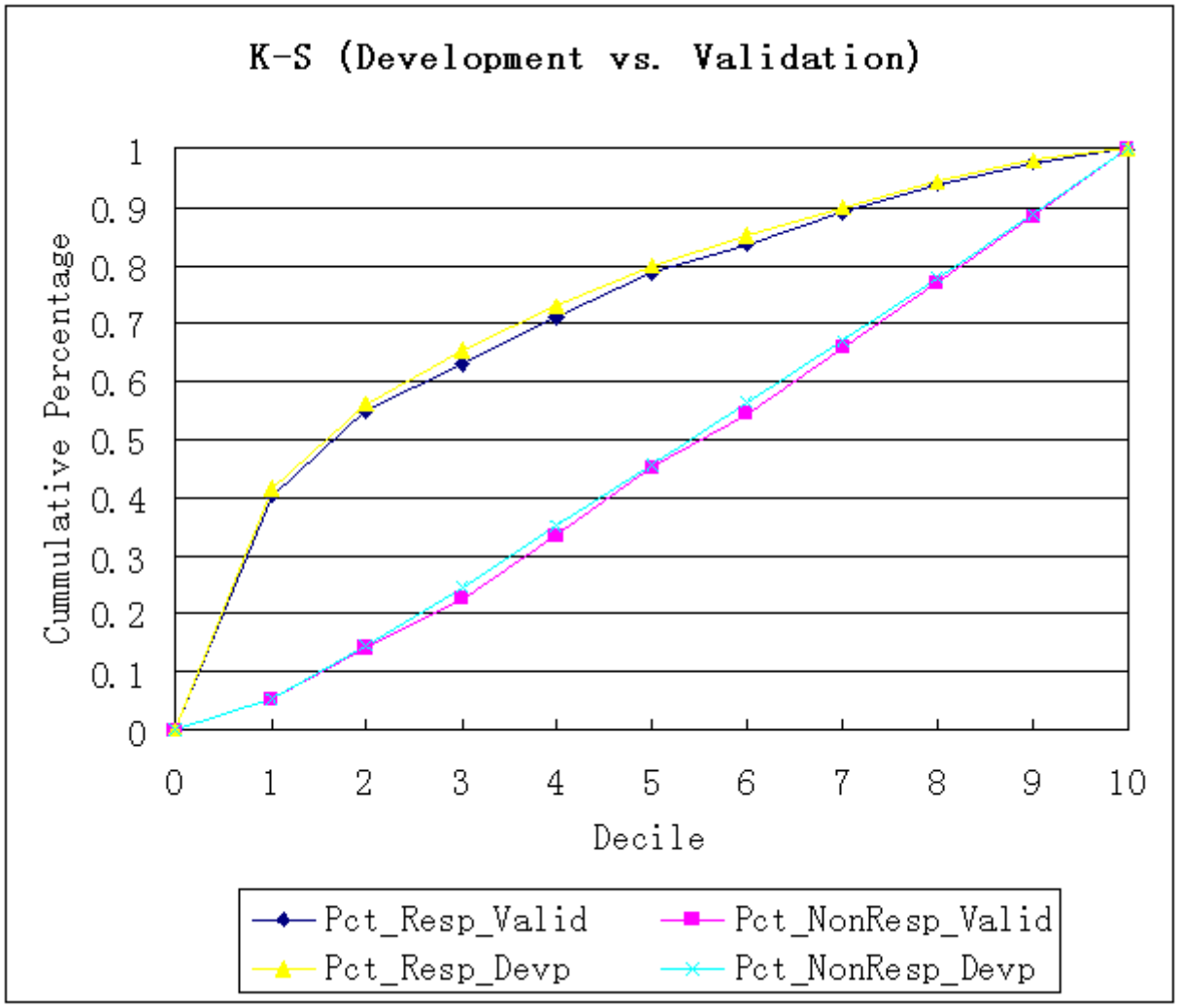
- **敏感性检验**

  模型对结果的预测能力不应该随一个或者少数几个数据的变动而发生巨大的改变。

模型评估指标示例：Lorentz曲线与KS

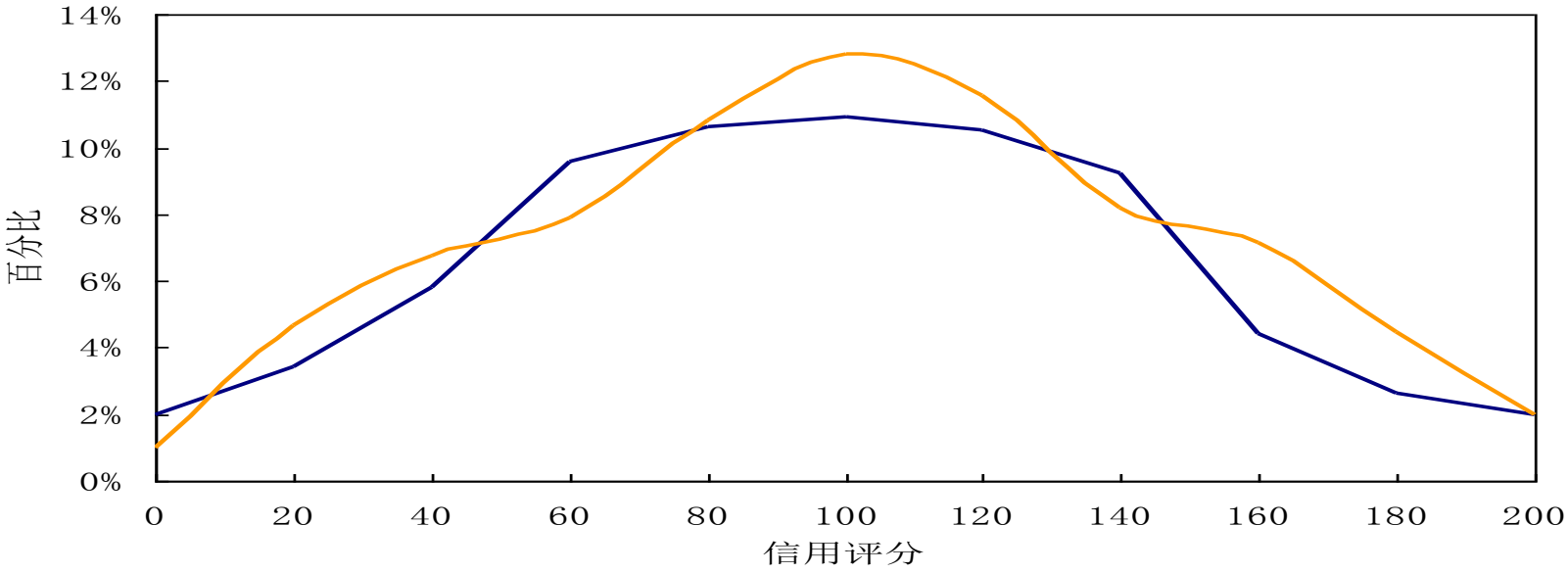


模型评估示例：可靠性检验



- 定期（每季度或半年）进行，以确保评分模型的预测能力不会随时间和外部环境变化而恶化。
- 影响评分预测力的因素：
  - 市场与经济环境发生变化
  - 消费者行为方式会逐渐改变
  - 信贷政策的调整引起批准人群改变
- 不同客户群的评分其预测力也不同
- 评分性能监控
  - KS、Lorentz等评估指标
  - 拖欠率分析
- 前端监控
  - 入组客户稳定性分析
  - 评分稳定性分析
  - 特征变量分析

评分监控示例:特征变量稳定性报告



A	B	C	D	E	F	G
			(C - B)	*(C / B)*	*Ln(E)*	*(D * F)*
特征变量	开发样本	当前情况	百分比差异	当前百分比/开发样本百分比		评分差异
0-4	0.127	0.150	0.023	1.181	0.166	**0.004**
5-6	0.150	0.223	0.073	1.487	0.397	**0.029**
7-10	0.188	0.197	0.009	1.048	0.047	**0.000**
11-14	0.191	0.210	0.019	1.099	0.095	**0.002**
15-17	0.167	0.153	-0.014	0.916	-0.088	**0.001**
18-25	0.170	0.057	-0.113	0.335	-1.093	**0.123**
25-High	0.007	0.010	0.003	1.429	0.357	**0.001**
总体	1.000	1.000		**变量稳定性**		**0.161**

数据挖掘提纲

- 数据挖掘技术和开发流程说明
- **数据分析案例——额度统一管理**

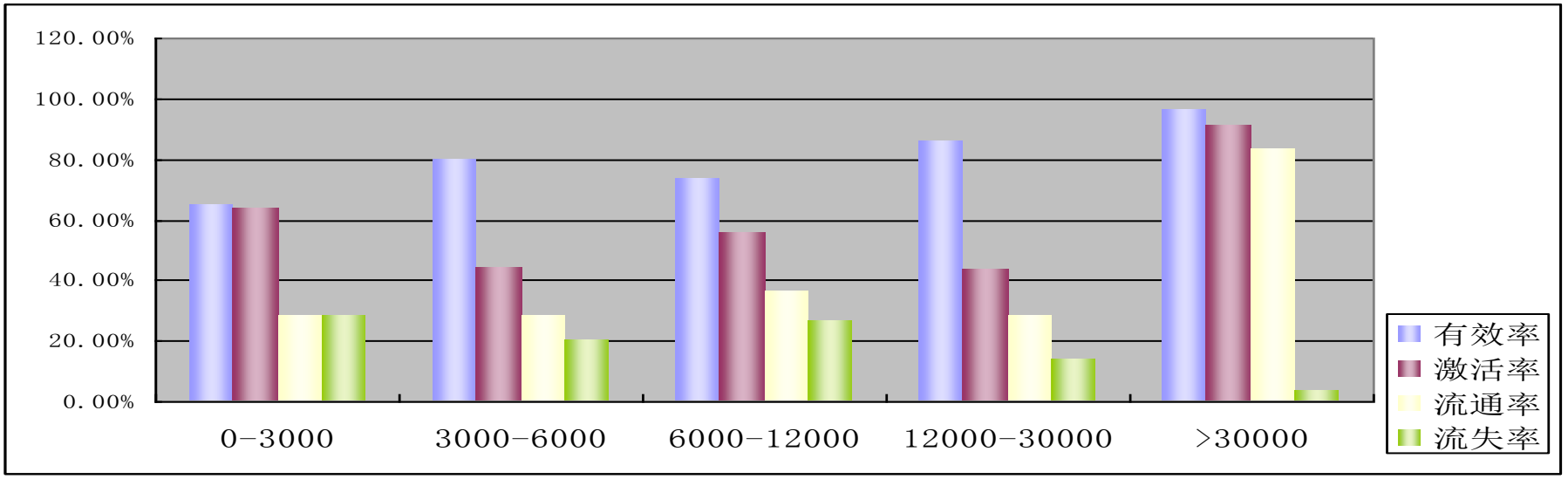
额度统一管理步骤

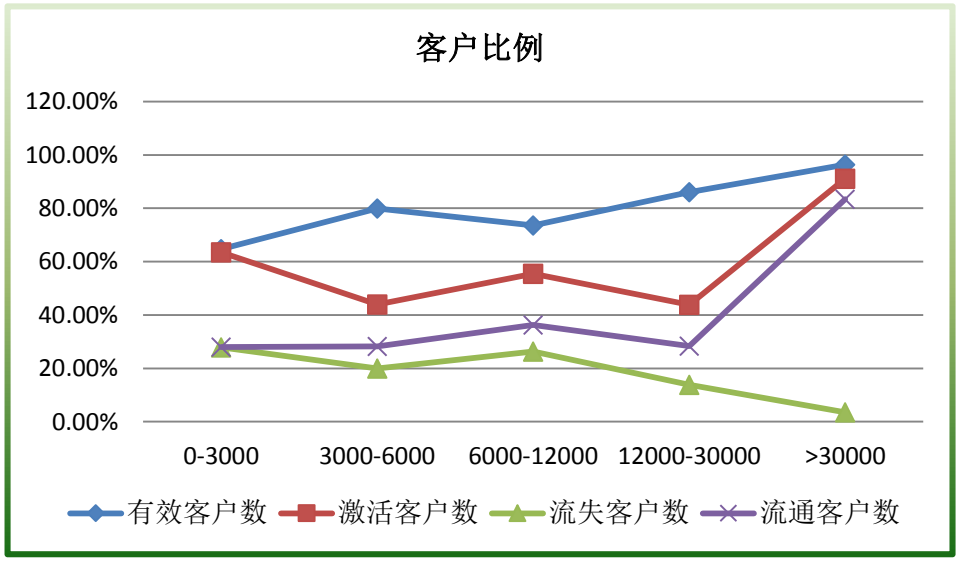
- 现有额度客群基本特征分析
- 已调额客户效果评价
- 客户特征与调额要素挖掘
- 调额方案设计
- 调额方案评估与调优

现有额度特群基本特征分析

- 授信额度区间分布分析
  - 累计客户占比
  - 有效率/激活率/流通率/流失率分析
- 授信额度使用情况分析
  - 正常/超额临时/未使用额度分布分析
  - 消费/取现/分期付款额度使用情况分析
  - 交易笔数/单比交易金额分析
  - 交易客户数/月均交易额度分析
  - 主要商户分析
- 不同额度使用率客户的收入/风险分析
  - 人均收入金额趋势分析
  - 贷款客户数和贷款金额分析
  - 逾期客户数和逾期金额分析
  - 循环信用客户数和循环信用率分析
  - 超额客户分析

分析图例和结果举例（一）





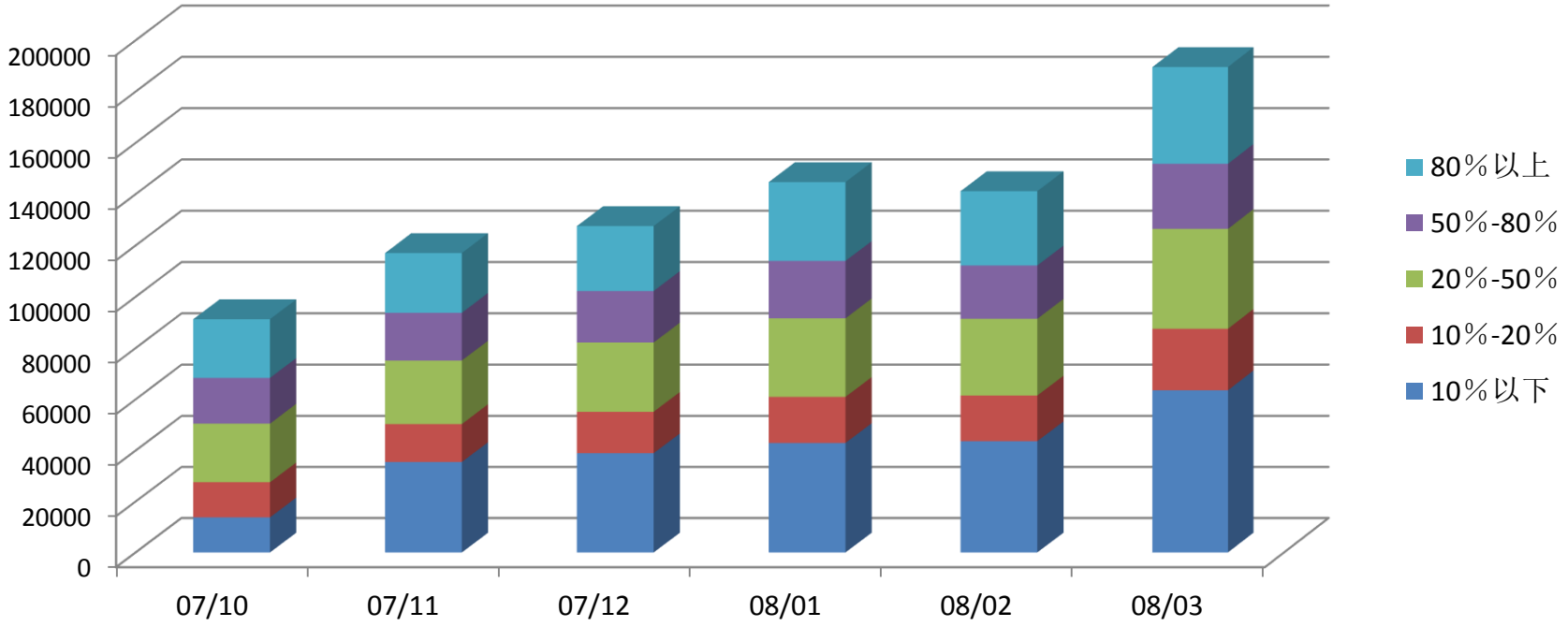
◆授信额度大于30000的高端客户，卡片激活及流通状况最好，且流失率最低。

◆在主要客户群中，额度为12000-30000元的客户，有效客户比例最高，但激活比例最低，因此在较高端的客户中，存在很高比例的潜在客户，同时也是高价值客户群体。

◆额度为6000-12000元的客户流通率最高，是较活跃客户，也是主要的利润来源群体。

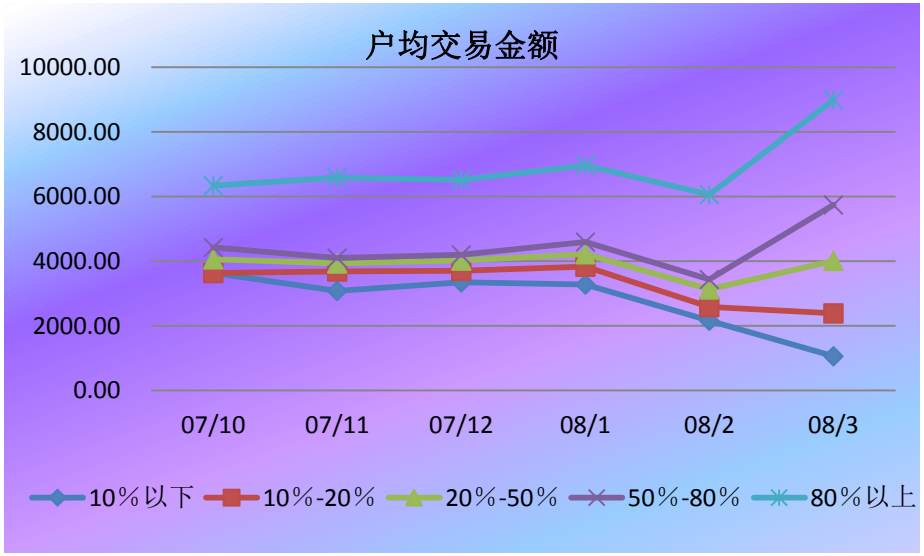
◆额度为3000-6000元客户，虽然绝对比例最高，但相对指标中等。

不同额度使用率的月度交易客户数



按照额度使用率分类，有图所示，额度使用率10%以下的交易客户数明显高于其他客户，且保持较高增长率，在2008年3月，额度使用率在10%以下的交易客户突破60000万，已达到使用率80%以上客户数的2倍。且额度使用率在20%-50%的交易客户也是当月交易客户的主要力量。进一步说明我行客户小额交易较多，中等消费客户占主导地位。

不同额度使用率的月度户均交易金额及笔数



- 额度使用率在80%以上的客户，月度户均交易金额均远远高于其他客户，月度户均交易金额为6505.65元，平均每笔交易金额1612.97元，该部分客户交易笔数少，且每笔交易金额大。额度使用率小于10%的客户，月度交易客户最多，但户均交易额每月仅2761.32元，平均每笔交易563.56元。
- 额度使用率较高的客户，交易笔数最少，往往是大额交易，以商旅客户为主；额度使用率较低的客户，交易笔数偏多，主要是普通购物消费，频率高且金额少

分析图例和结果举例（四）

序号	mcc描述	频率
1	杂货店	17.98%
2	百货商店	10.15%
3	帐篷店	9.07%
4	专业服务	5.40%
5	家庭用品店	4.95%
6	餐馆	4.46%
7	慈善 /社会服务组织	3.73%
8	政府服务	3.40%
9	非盈利性事业	3.30%
10	酒店 /汽车旅馆 /客栈	3.10%
11	航线，航空运输	1.90%
12	自动加油机	1.61%
13	医院	1.41%
14	药房	1.38%
15	建筑公司（房地产类）	1.04%
16	酒吧 /酒馆 /娱乐室 /迪厅	0.94%
17	电话服务 /设施	0.86%
18	按摩室	0.84%
19	理发 /美容店	0.81%
20	鞋店	0.80%

额度使用率>95%的客户在3月份的消费情况。

按mcc代码分类，消费客户数排名。前20位的商户类型如图所示。

排名前十位的商户，亦为全量客户的主要消费商户。在10-20的商户中，可以看出>95%的商户，主要消费类型如下：

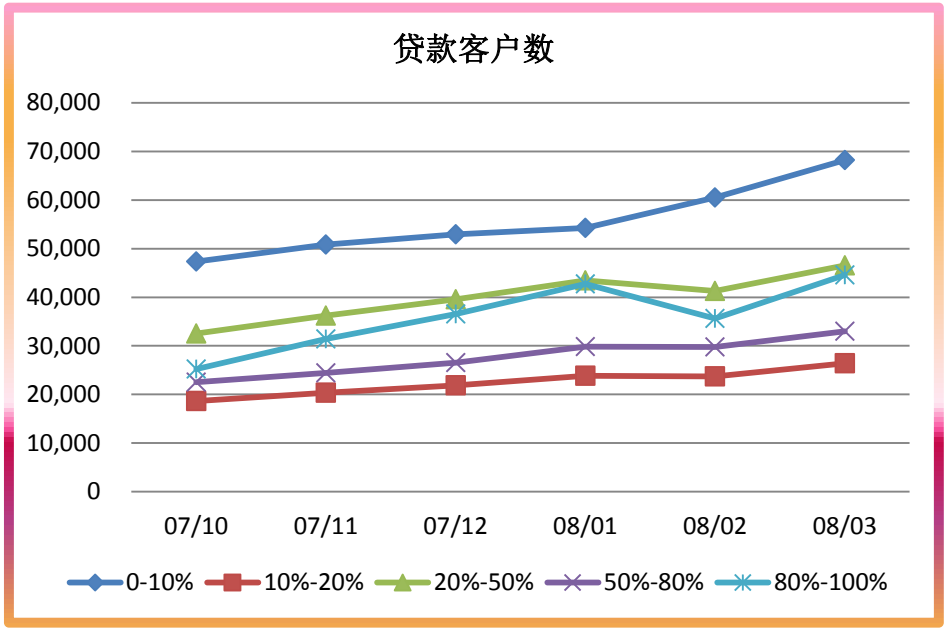
1：酒店住宿类

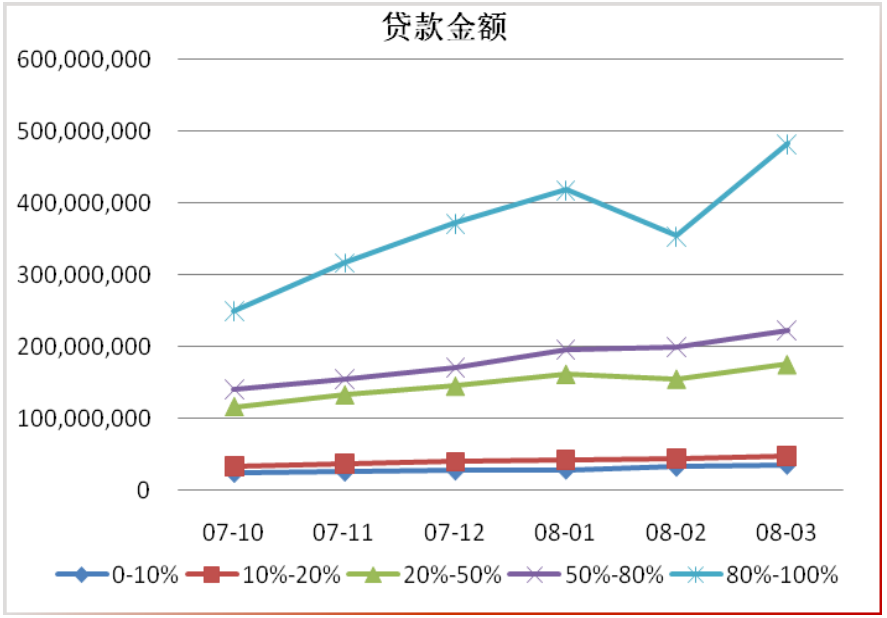
2：航空类

3：房地产类

4：娱乐消费类

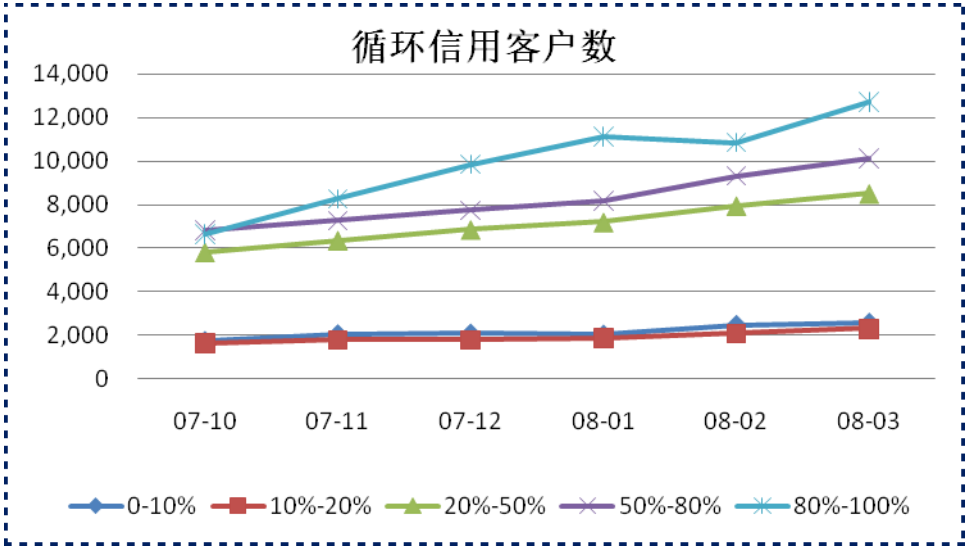
分析图例和结果举例（五）

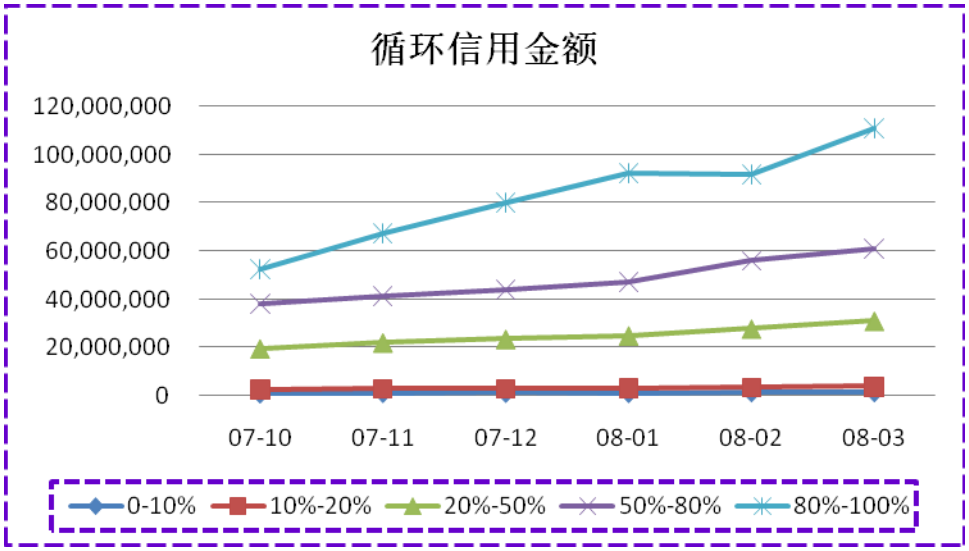




- 通过贷款客户数和贷款金额的变化情况可以看出，不同额度使用率的变化趋势几乎相同，因此可以看出对于贷款类指标，外部条件所带来的影响很小，主要影响因素是客户对额度的使用情况，即较高使用率的客户，贷款金额也较高，合理的授予客户额度，并采取有效措施促进客户充分使用额度，可以很好的提高贷款收入。

分析图例和结果举例（六）

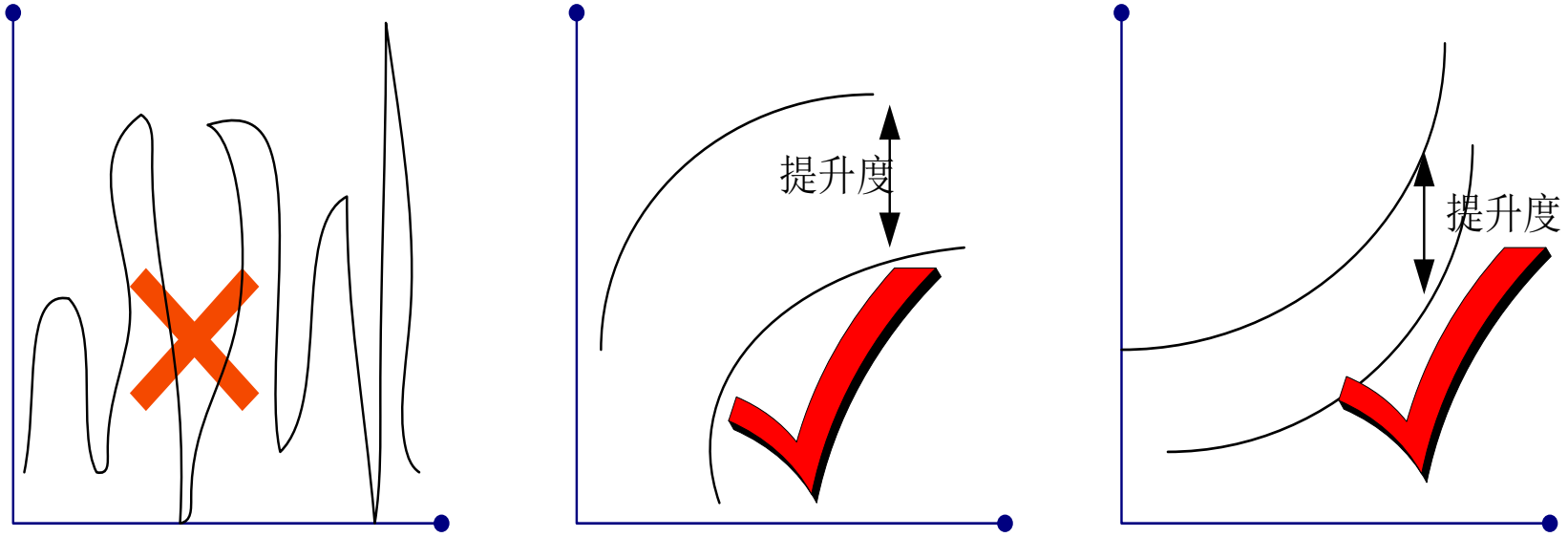




- 额度使用率在80%-100%之间的客户倾向于使用循环信用，主要因为对于信用卡使用率较高的客户，一般具有超前的消费观念，且高频率的使用状况表明这部分客户使用信用卡为习惯性消费而并非偶然性消费，额度使用率较高，且月均消费较高，促使了这部分客户更倾向于选择循环信用，以此来减轻短期经济压力。
- 额度使用率在20%以下的客户，几乎很少使用循环信用，该部分客户交易频繁，且金额较低，比较容易控制自身风险，较低的使用率也暴露了该客户群体消费观念略显保守。

已调额客户效果评价

- 行为特征稳定的客户群样本选择
  - 客户行为特征刻画（稳定性指标设计）
  - 客户月均历史收益滚动分析
  - 客户损失跟踪分析
  - 客户收益预测
- 样本客户群收益对比分析
  - 收益指标分析
  - 收益提升量化分析



客户特征与调额要素挖掘

- Logistic回归模型
- 双变量分析
- 时间序列分析模型

- 调额形式设计
  - 定期批量调额的条件指标设计
  - 实时的TRIGGER条件促发调额
  - 客户主动申请时的评分模型
- 调额方式设计
  - 据账单多长时间进行知会
  - 以短信/电话/账单形式通知
  - 条件触发的时间/频率/术语
  - 百分比调额/固定金额调额
  - 条件调额
  - 附属条件设置（赠品、积分）
  - 其他变异方式（利息减免）

调额方案评估与调优

- 量化指标的设定
  - 调额因素影响力分析
  - 调额因素响应模型
  - 调额竞争力提升度分析（结合征信报告）
- 调额方案比较分析
  - 调额表现周期分析
  - 客户对象随机样本的选择