
# 信用评分

李志勇/张兴敏

西南财经大学

Southwestern University of Finance and Economics


# 第三章 特征工程

# 数据处理和变换


- Measurement scales: Nominal/Ordinal/Interval/Ratio
- 连续变量：可直接使用，也可转换为分类变量
- 分类变量：需要进行转换后使用
- 细分类：原始分类，提供最多的细节
- 粗分类/分箱/分组：把细分类合并成数量更少的粗分类组合，考虑将风险相似的组归合并在一起。
- 粗分类（ Coarse classification ）、分箱（Binning）

| Provides: | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| The “order” of values is known | | ✔ | ✔ | ✔ |
| “Counts,” aka “Frequency of Distribution” | ✔ | ✔ | ✔ | ✔ |
| Mode | ✔ | ✔ | ✔ | ✔ |
| Median | | ✔ | ✔ | ✔ |
| Mean | | | ✔ | ✔ |
| Can quantify the difference between each value | | | ✔ | ✔ |
| Can add or subtract values | | | ✔ | ✔ |
| Can multiple and divide values | | | | ✔ |
| Has “true zero” | | | | ✔ |

# 数据处理和变换


## 粗细分类

- 在细分类时，“年龄”或许是每年一组；但在粗分类时，可以把它们分为有意义且数量更少的组合，如18~30， 31~40，41~55，55+。粗分类取决于特征类型。
- 连续/顺序变量——暗含排序，数据间的风险关系通常应该是单调的。不考虑特殊编码组，把临近的组合并在一起。
- 离散/类别变量——不暗含排序。分组只能在经验或可用数据的基础上进行。在有必要分组且数量少的地方，需要人的经验判断。
- 当一个分类变量有太多的类别（属性）时，我们可以考虑粗分类，使模型更稳健，消除极端值影响，同时也能处理风险与该特征间的非单调关系。

# 数据处理和变换

- 细分类

表　工作时间细分类

| 时间 | 好人数量 | p(x\|G) | 坏人数量 | p(x\|B) | 好坏比率 | 证据权重 | 信息比率 |
|---|---|---|---|---|---|---|---|
| 001 | 136 | 0.0026 | 5 | 0.0016 | 27.2000 | 0.4735 | 1.6056 |
| 002 | 582 | 0.0112 | 62 | 0.0203 | 9.3871 | -0.5904 | 0.5541 |
| to 003 | 686 | 0.0132 | 60 | 0.0196 | 11.4333 | -0.3932 | 0.6749 |
| to 006 | 1113 | 0.0215 | 120 | 0.0392 | 9.2750 | -0.6024 | 0.5475 |
| to100 | 2778 | 0.0536 | 218 | 0.0713 | 12.7431 | -0.2847 | 0.7522 |
| to 106 | 1810 | 0.0349 | 161 | 0.0526 | 11.2422 | -0.4100 | 0.6636 |
| to 200 | 2856 | 0.0551 | 246 | 0.0804 | 11.6098 | -0.3779 | 0.6853 |
| to 206 | 1265 | 0.0244 | 89 | 0.0291 | 14.2135 | -0.1755 | 0.8390 |
| to 300 | 3974 | 0.0767 | 328 | 0.1072 | 12.1159 | -0.3352 | 0.7152 |
| to 400 | 4217 | 0.0814 | 271 | 0.0886 | 15.5609 | -0.0849 | 0.9186 |
| to 500 | 3883 | 0.0749 | 249 | 0.0814 | 15.5944 | -0.0828 | 0.9205 |
| to 700 | 5132 | 0.0990 | 287 | 0.0938 | 17.8815 | 0.0541 | 1.0555 |
| to 1,000 | 6323 | 0.1220 | 282 | 0.0922 | 22.4220 | 0.2803 | 1.3236 |
| to 1,500 | 6912 | 0.1334 | 277 | 0.0906 | 24.9531 | 0.3873 | 1.4730 |
| to 2,000 | 4281 | 0.0826 | 142 | 0.0464 | 30.1479 | 0.5764 | 1.7796 |
| to High | 5873 | 0.1133 | 262 | 0.0856 | 22.4160 | 0.2801 | 1.3232 |
| total | 51821 | | 3059 | | 16.9405 | | |

# 数据处理和变换


- 细分类

| | 好 | 坏 | 不定 | 好坏比率 | WoE | 信息值 |
|---|---|---|---|---|---|---|
| 缺失 | 12878 | 698 | 1252 | 18.5 | 0.09 | 0.113 |
| 自有房产 | 13856 | 703 | 1401 | 19.7 | 0.16 | 0.158 |
| 租房 | 9461 | 777 | 1262 | 12.2 | –0.33 | 0.000 |
| 父母同住 | 7234 | 527 | 919 | 13.7 | –0.21 | 0.005 |
| 配偶房产 | 1595 | 73 | 138 | 21.8 | 0.26 | 0.026 |
| 公司房产 | 599 | 45 | 119 | 13.3 | –0.24 | 0.000 |
| 共同所有 | 6173 | 245 | 493 | 25.2 | 0.40 | 0.146 |
| 合计 | 51797 | 3069 | 5584 | 16.9 | | 0.448 |

# 数据处理和变换

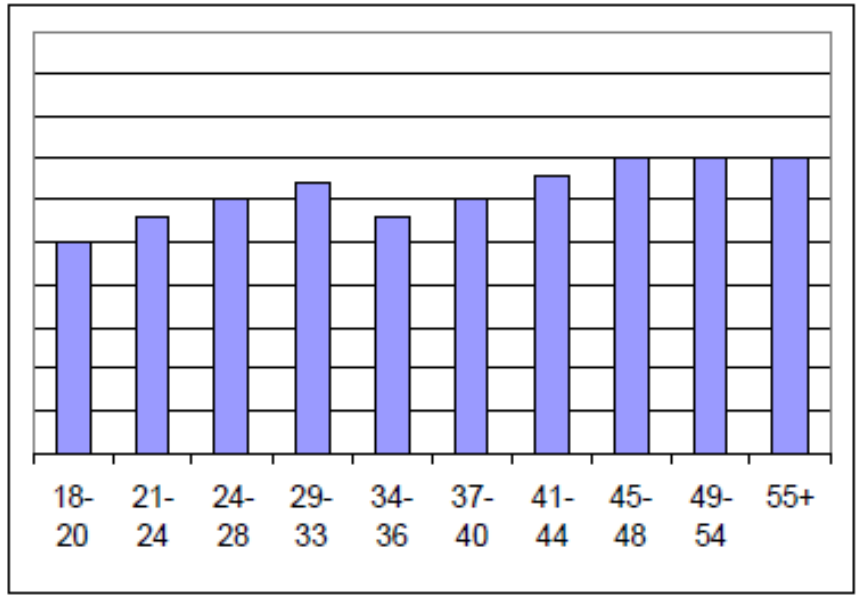

粗分类

- 合适数量：分类不宜太多或太少，太多使得模型复杂，太少导致信息损失，一般不超过10个。
- 足够样本：每类里面尽量包含5%以上样本
- 合理分组：1. 风险相似；2. 逻辑合理；3. 合理断点
- 合理关系：组间的风险差异直观
- 按照常识归类和根据好坏比率归类同样重要
- 粗分类既是一门技术，又是一门艺术

# 数据处理和变换


连续特征的粗分类（连续特征离散化）

- 像年龄、收入这样的连续特征，首先把特征值分出10-20个分位点，建立群体数量大致相等的子类
- 计算每个子类好坏比率
- 尝试将好坏比率相近又相邻的子类合并



# 数据处理和变换


分类特征进行粗分类后，仍无法直接使用，还需要进行转换

- 单变量方法：对数、指数、归一化、正态标准化（Z统计量）、多项式展开或这些方法的组合
- 单变量方法缺点：可能丢失数据中的某些关键信息；不被信息系统支持
- 多变量方法
- 虚拟变量——将分类特征转换为一些列的二元变量，能够完全充分表示该特征携带的信息
- 风险变量——为每个特征建立和目标变量的线性关系变量（证据权重、好人概率）

# 数据处理和变换


## 虚拟变量

- 虚拟变量又称哑变量（dummy variable or proxy），是属性的一个人工变量，是量化了的自变量，通常取值为0或1。
- 当一个分类变量有k个类别的时候，只需引入k-1个虚拟变量来充分表示这个特征包含的信息，否则容易出现虚拟变量陷阱问题。
- 粗分类后至少包含一个缺省组（null group）；为非缺省组设置二元虚拟变量。缺省组应为：（1）最接近平均风险的组；（2）观测数最多的组；（3）空白值组或缺失值组；或（4）数据不足或无法合并到其他属性的组。
- 缺点：类别很多时会引入过多自变量，自由度降低，容易过度拟合，模型复杂，计算时间更长。
- 优点：能针对每个类别给出统计结果

| | 虚拟变量示例 | |
|---|---|---|
| 粗分类 | 数量 | 违约率 |
| 是 | 30000 | 5.00% |
| 否 | 3000 | 12.00% |
| 缺失 | 2000 | 10.00% |
| 合计 | 35000 | 5.90% |

# 数据处理和变换

- 虚拟变量设置

| 序号 | 学历 | X1(本科) | X2（硕士） | X3（博士） |
|---|---|---|---|---|
| 1 | 本科 | 1 | 0 | 0 |
| 2 | 硕士 | 0 | 1 | 0 |
| 3 | 本科 | 1 | 0 | 0 |
| 4 | 本科 | 1 | 0 | 0 |
| 5 | 高中 | 0 | 0 | 0 |
| 6 | 硕士 | 0 | 1 | 0 |
| 7 | 博士 | 0 | 0 | 1 |

4类

3个虚拟变量

# 数据处理和变换

one-hot encoding （独热编码）vs. dummy encoding

小学 -> [1, 0, 0, 0, 0]
中学 -> [0, 1, 0, 0, 0]
大学 -> [0, 0, 1, 0, 0]
硕士 -> [0, 0, 0, 1, 0]
博士 -> [0, 0, 0, 0, 1]

小学 -> [1, 0, 0, 0]
中学 -> [0, 1, 0, 0]
大学 -> [0, 0, 1, 0]
硕士 -> [0, 0, 0, 1]
博士 -> [0, 0, 0, 0]

# 数据处理和变换


风险变量——**证据权重**weights of evidence

- 证据权重（WOE）用在某一分类特征变量X的每个类别上**x**

$$w(\mathbf{x}) = \ln(I(\mathbf{x})) = \ln\left(\frac{p(G\mid\mathbf{x})/p(B\mid\mathbf{x})}{p_G/p_B}\right) = \ln\left(\frac{g_i/b_i}{n_G/n_B}\right) = \ln\left(\frac{g_i n_B}{b_i n_G}\right)$$

- 其中，

  $b_i$ 和 $g_i$ 是该特征落在某类别 $i$ 中坏人和好人的数量

  $n_B$ 和 $n_G$ 是坏人和好人的总数量

- 证据权重将某个分类特征转换成数值量化的变量，可以直接放入模型

# 计算WOE


| 收入 | 好人数量 | 坏人数量 | 好坏比率 | 好人比例 | 坏人比例 |
|---|---|---|---|---|---|
| 低 | 5000 | 2000 | 2.5 | 14.3 | 33.3 |
| 中等 | 10000 | 2000 | 5.0 | 28.6 | 33.3 |
| 高 | 20000 | 2000 | 10.0 | 57.1 | 33.3 |
| 合计 | 35000 | 6000 | 5.8 | 100.0 | 100.0 |

# 数据处理和变换


## 证据权重

- 步骤：（1）考虑不同粗分类方式；（2）看WOE符号是否与直觉相符，选择符号正确的那种；（3）建模查看预测效果；（4）选择较好且符合逻辑的粗分类方式。
- 证据权重关注具体属性相对较少，但优点有：

（1）一对一转换，自由度更大，模型更稳健；

（2）属性间的关系确定、可比；

（3）正负号指示与风险的关系，直观便于理解；

（4）与逻辑回归评分卡自然契合

（5）变量显著性较好

# 数据处理和变换


## 信息值Information Value（IV）

- FICO公司采用信息值来度量某个特征的预测能力，信息值IV又叫Kullback散度。

$$\mathrm{IV}=\sum_{i=1}^{n}\left[\left(\frac{g_i}{n_G}-\frac{b_i}{n_B}\right)\times \mathrm{WoE}_i\right]$$

- 某特征的IV值小于0.1效果不太好，大于0.5存疑。

| 收入 | 好人数量 | 坏人数量 | 好坏比率 | 好人比例 | 坏人比例 | 证据权重 |
|---|---|---|---|---|---|---|
| 低 | 5000 | 2000 | 2.5 | 14.3 | 33.3 | −0.847 |
| 中等 | 10000 | 2000 | 5.0 | 28.6 | 33.3 | −0.154 |
| 高 | 20000 | 2000 | 10.0 | 57.1 | 33.3 | 0.539 |
| 合计 | 35000 | 6000 | 5.8 | 100.0 | 100.0 | 信息值 |

# 数据处理和变换

| 年龄组 | 实际好人数量 | 期望好人数量 | 好人%－坏人% | $\ln(\frac{好人\%}{坏人\%})$ |
|---|---|---|---|---|
| 30 岁以下 | 175 | 26 | －0. 066 | －0. 291 |
| 30 岁以上 | 725 | 74 | 0. 066 | 0. 085 |
| 40 岁以下 | 350 | 56 | －0. 171 | －0. 365 |
| 40 岁以上 | 550 | 44 | 0. 171 | 0. 329 |
| 50 岁以下 | 525 | 71 | －0. 127 | －0. 197 |
| 50 岁以上 | 375 | 29 | 0. 127 | 0. 362 |

三种年龄分组方式得到的IV值

0.019+0.006=0.025

0.062+0.056=0.119

0.025+0.046=0.071

# 风控模型开发流程


- 数据获取：从各数据源获取原始数据
- 数据处理：对原始数据进行整理变换
- 特征工程：确定有预测能力、符合逻辑、稳定可得、合规、有关、相关性低的特征作为备选变量
- 样本细分：根据市场、渠道、客户、数据、流程等因素对样本进行分层，分别建立评分卡
- 训练模型：用历史样本建立模型
- 拒绝推断：开发申请评分卡时对被拒绝的申请者的表现进行推断。
- 模型校准：确保分数在不同评分卡中有相同的含义，反映相应的违约概率。
- 验证交付：用保留样本和近期样本检验模型是否过度拟合或不稳定，然后准备投入使用。

# 数据准备


数据来源：客户、内部、外部

- 申请表格——采集并编译，现多为电子表格，或不需表格。
- 征信机构——查询、回溯检索。
- 内部系统——从内部系统获取行为数据。
- 表现数据——每个样本的好坏结果。
- 匹配合并——将所有数据表整合到一起。匹配键key？

# 数据准备


数据校验data validation：对数据的有效性进行查验

- 例如，没有150岁的人，没有21岁就已持有银行账户超过30年的人。
- 发现有特殊情况的，不应被简单理解和处理，而应用特殊代码如9999标记。如果存在缺失数据，最规范做法是把它们作为缺失值进行编码，而不是试图估算填入一个实值。
- 实际操作中，缺失数量不太多时，也可以替代为某个实值（平均值）
- 真实数据包含大量的缺失值、极端值、异常值

# 样本设计

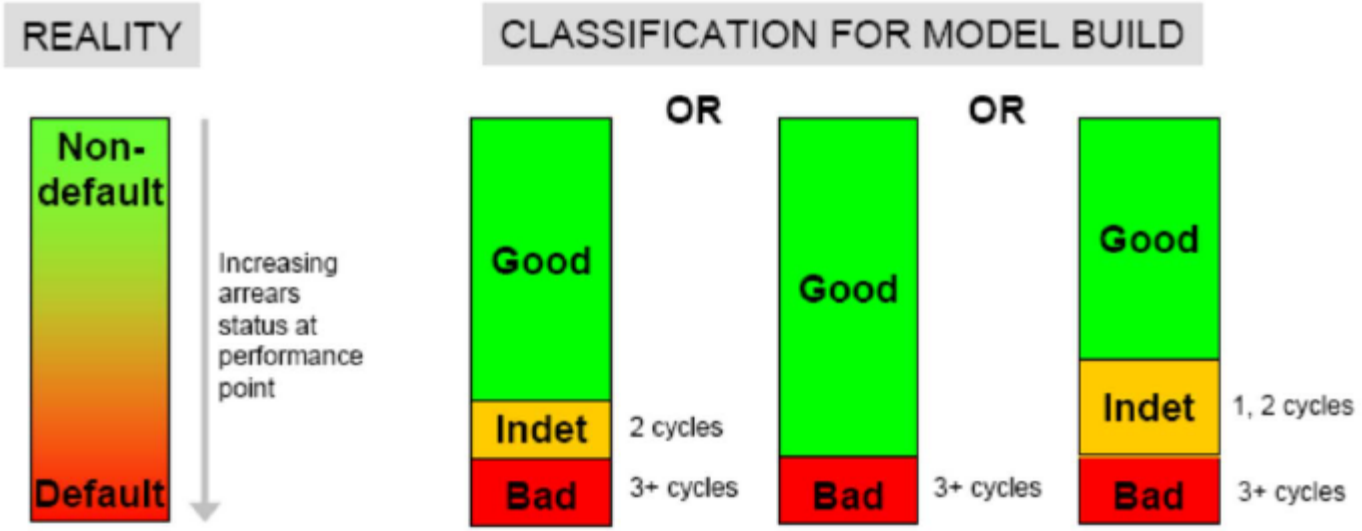
好坏定义

坏账户：default/charged off，90天/连续三期逾期（M3）——巴塞尔协议

好账户：Fully paid (prepayment)

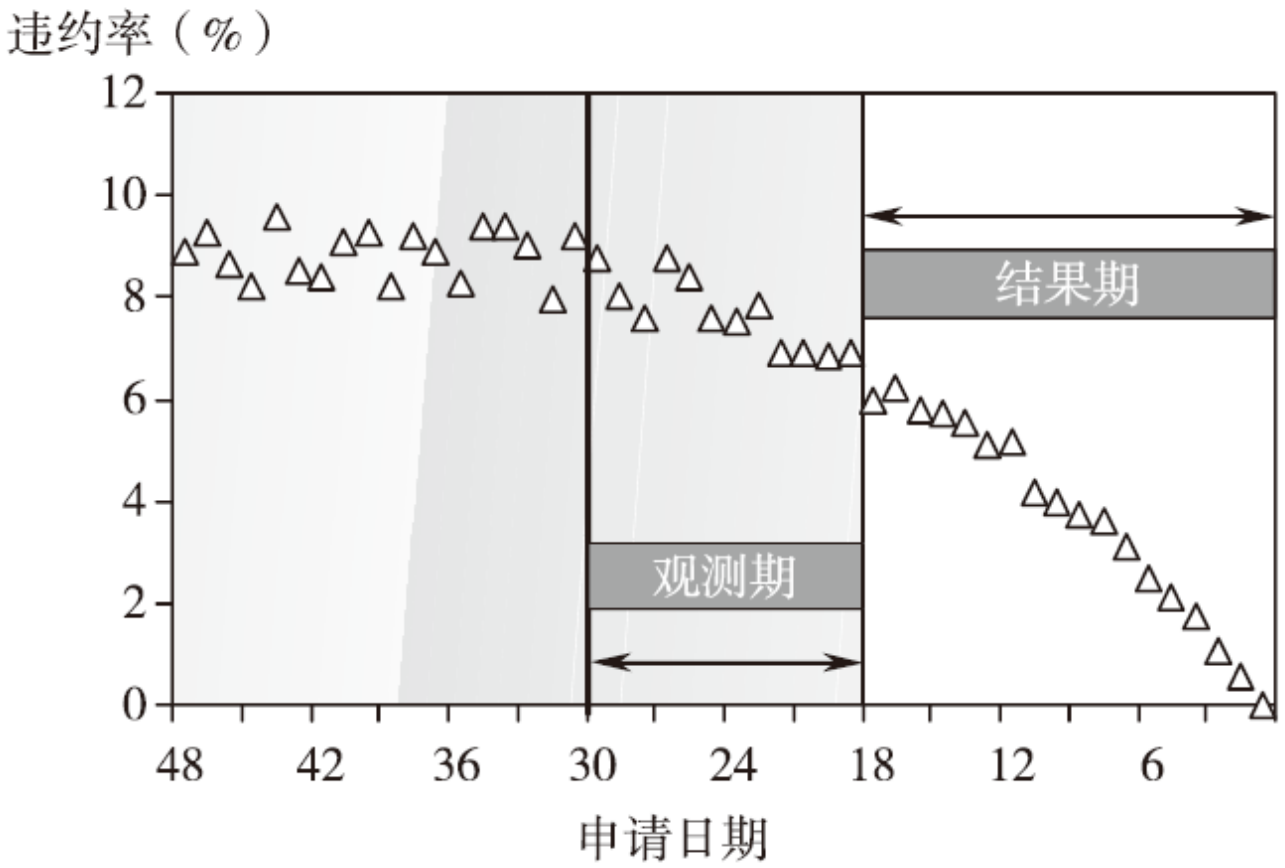
正常还款账户：Current; up-to-date

不确定的账户：M1 or M2



# 样本设计

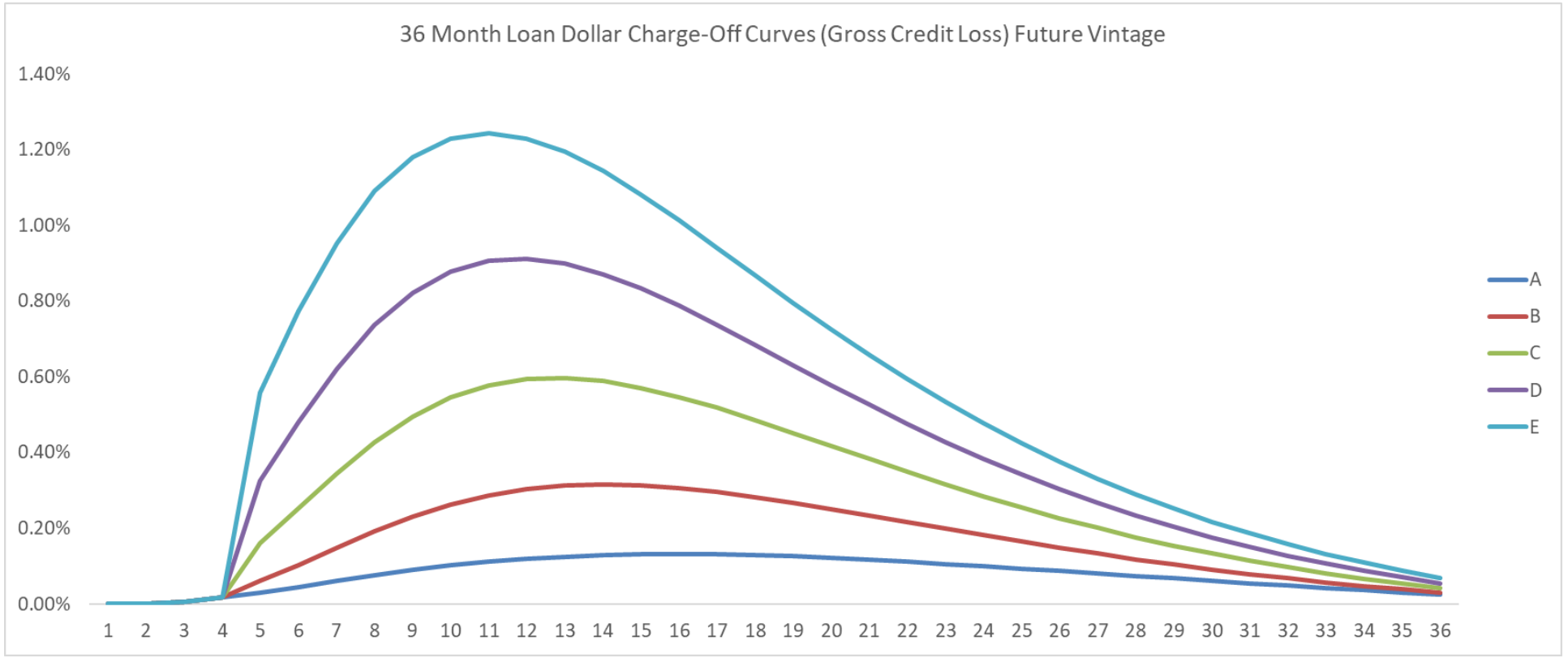
违约成熟



# 样本设计

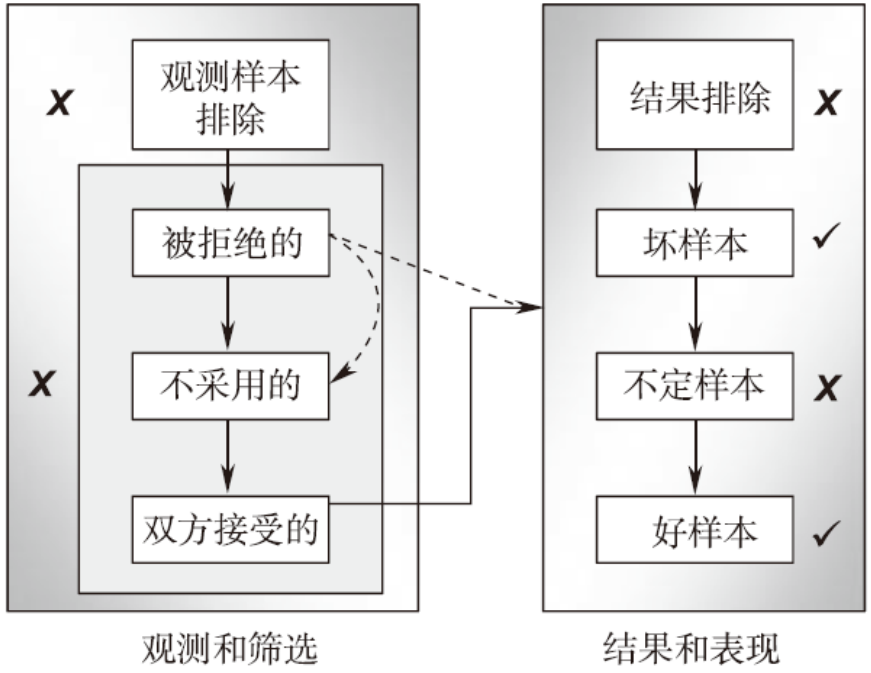
账龄分析vintage analysis

Month on Book



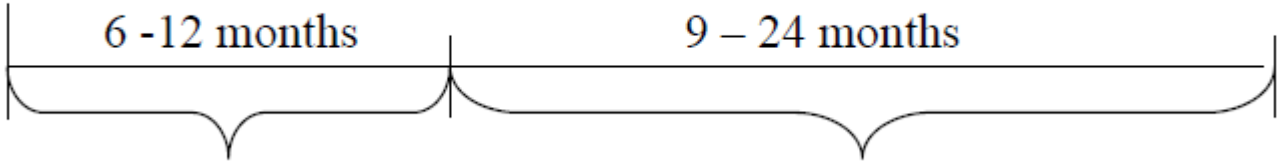
# 样本设计

- 训练样本training sample——用来开发预测模型的样本，这是最关键的部分，需要大量的观测。
- 保留样本（测试样本）hold-out or test sample——历史样本中单独保留的样本，用来验证模型，模型可能过度拟合，在训练样本表现很好但在其他样本上没用。

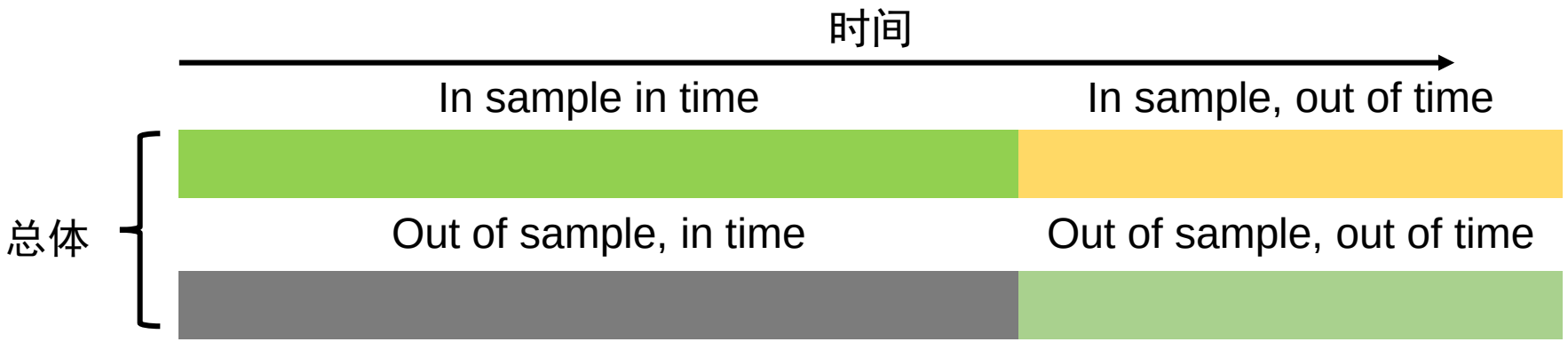


# 样本设计

时间窗口Observation window




抽样时期　　表现时期




近期样本：建模前三个月左右的样本，确保特征分布和评分卡稳定

# 样本设计


- 样本分层segmentation决定是否要细分总体并为每部分单独建模。
- 例如，为25岁以下和25岁及以上的人各建立一个评分卡，或者为那些高收入的人建立一个评分卡，也为低收入的人建立另一个。
- 建立多个评分卡会带来大量额外的工作，所以只在理由充分并且改进预测效果时使用。
- 样本分层的理由有如下几个：

  数据可得性、特征相关性、经营策略
- 通用评分卡generic model和分层评分卡segmented model：一般通用模型普适性好，专用模型预测性好。

# 特征工程


- 变量剔除：稳健的模型通常有10到20个特征变量，而可用的特征远远多于这个数量，需要剔除变量。
- 剔除的特征：区分好坏的能力偏小；与其他确定要使用的变量高度相关甚至存在共线性；在时间上不稳定。
- 多重共线性（multicollinearity）：严重危害模型！

  变量不显著；符号不确定；显著性不稳定；过度拟合
- 多重共线性解决办法：相关性分析、变量剔除、逐步回归、主成分分析

# 特征工程

| | Own phone | No phone | |
|---|---|---|---|
| Owner | 95% | 50% | 90% |
| Tenant | 75% | 65% | 70% |
| | 91% | 60% | |

- 自有房产的人比租房的人的好人比例更好（90% vs 70%），有电话的人比没电话的人的好人比例更高（91% vs 60%），
- 在线性模型里，有电话有房产的人的得分最高，得分最低的是没有电话的租房者。事实上，有房产但没有电话的人实际上更坏（50%）。
- 如果事先不知道这个现象，这种非线性关系很难被线性模型捕捉到。
- 决策树、神经网络、支持向量机、贝叶斯网络都能很好的应对这种非线性关系。建模人员可以先用决策树发现重要的相互作用，然后依次进行样本分层建模。
- 在这个例子里就是为自有房产和租房的人单独建模。一般年龄是主要的分层变量。