逻辑回归

移动运营商的困惑——客户流失

1.5%客户
流失率→
流失预警?

流失的定义

☆ 公司认为只要**符合以下两条中的一条**即被认为是流失或离网

数据背景

入网时间、当月花费、话费情况等等

通话详单，量级最大且最重要。平均每月500万条左右

随机选取5万个左右（月均花费大于80）的客户，收集期：2014年3月到8月。ID已脱敏。数据按月份统计。

表1 客户基础通讯表字段解释（部分示例）

字段中文名称	说明	取数日期
设备编码	脱敏后的用以唯一识别客户的 ID 号	取当前状态
入网时间	格式：YY-YY-MM，用户使用服务开始时间	取当前状态
拆机时间	格式：YY-YY-MM，用户拆除服务时间	取当前状态
客户状态	正常/停机/拆机，其中停机分为客户主动申请停机和欠费停机；在表中‘1’表示正常，‘2’表示拆机，‘3’表示停机	取当前状态
号码等级	号码费别，由 0 到 9，从低到高共十个等级，用于区分号码资源的优劣情况	取当前状态
客户分群	公众或政企，用于描述客户级分群，客户级属性为政企或者加入政企集团网 30 天以上，则客户分群为政企；其余为公众，注释：在表中 1 表示政企，2 表示公众	取当前状态
托收方式	现金或银行划扣，标注用户缴费方式，托收是指从绑定银行账号中扣费；在表中 1 表示现金，2 表示银行划扣	取当前状态
入网渠道	设备发展渠道，注释：1 表示自有直销渠道，2 表示自有实体渠道，3 表示社会实体渠道，4 表示社会电子渠道，5 表示社会直销渠道，6 表示自有电子渠道	取当前状态
当前月消费	单位：元，当前实际出账费用	取当前状态
是否融合	办理了多业务融合优惠套餐，1 表示是，0 表示否	取当前状态
下月状态	指下个月的保有状态，分稳定、流失和不稳定三种状态	取当前状态
是否延迟缴费	在当前月消费的下月，超过缴费期时间，未及时缴费的判断为延迟缴费	统计月数据
本地语音通话费	单位：元，号码归属地市内语音通话费	统计月数据
长途语音通话费	单位：元，在归属地拨打市外语音通话费	统计月数据
省内语音漫游费	单位：元，在省内（除归属地）漫游拨打电话的语音通信费	统计月数据
省际语音漫游费	单位：元，在省外漫游拨打电话的语音通信费	统计月数据
国际语音漫游费	单位：元，在国外漫游拨打电话的语音通信费	统计月数据
短信费	单位：元，发送短信的费用	统计月数据
主叫次数	单位：次，号码作为主叫的通话次数	统计月数据
被叫次数	单位：次，号码作为被叫的通话次数	统计月数据

点对点通信与社交网络

表2 客户点对点通信数据

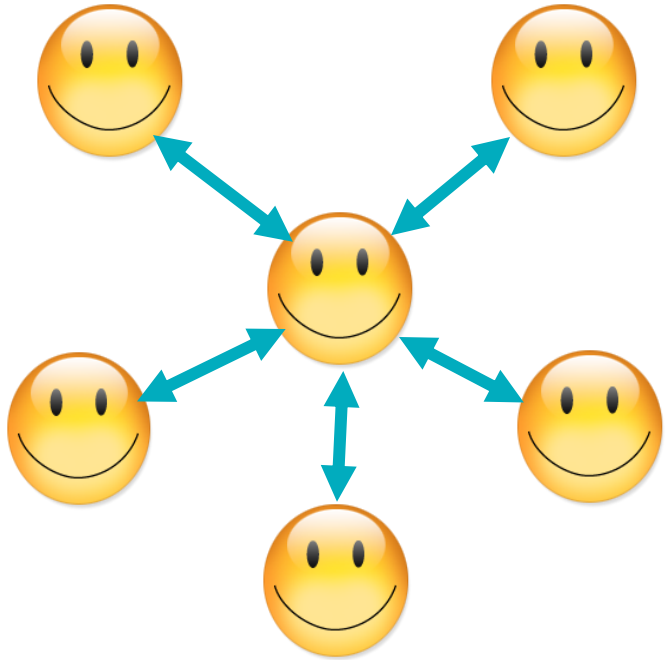
字段中文名称	说明	取数日期
目标设备编码	脱敏后的用以唯一识别客户的 ID 号	统计月数据
对方设备编码	通过【目标设备编码】进行统计的被叫及主叫	统计月数据
通话次数	两个号码之间当月通话的次数	统计月数据
通话时长	单位：分钟，每次通话的时间	统计月数据
呼叫类型	主叫：目标设备编码拨打对方设备编码 被叫：对方设备编码拨打目标设备编码	统计月数据


变量构建

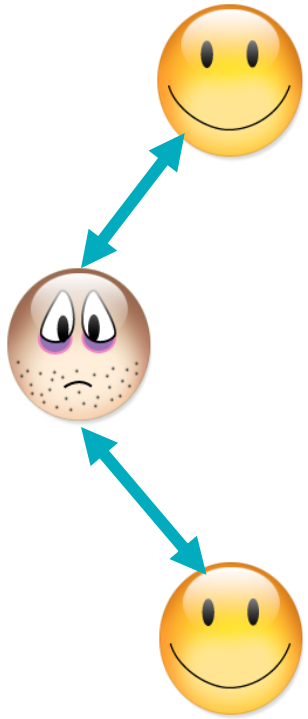
如何利用通话详单数据？

社交网络变量的产生——个体的度

通话人数
degree



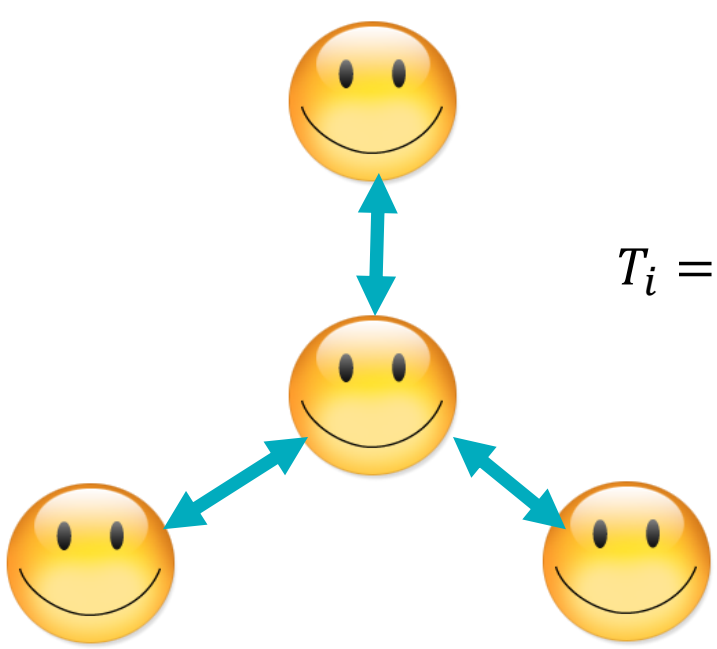
$$D_i = \sum_{j \neq i} a_{ij}$$

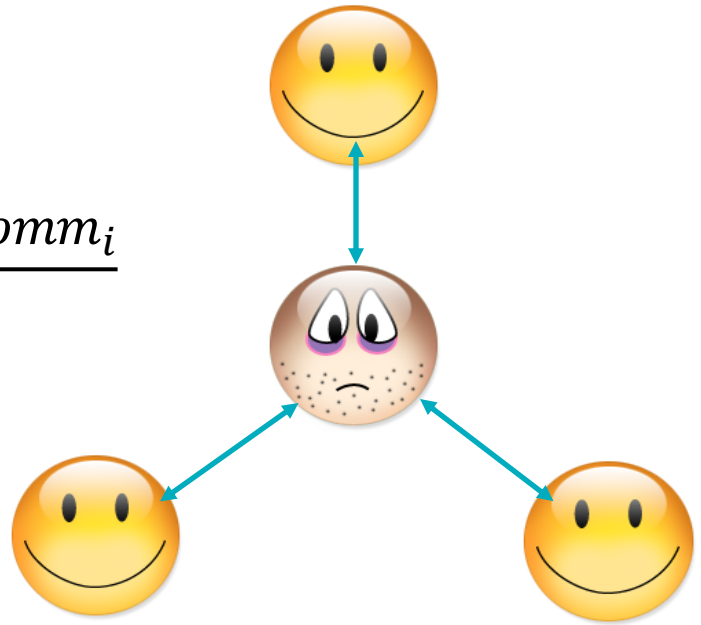


社交网络变量的产生——联系的强度

人均通话时长
Minutes per person

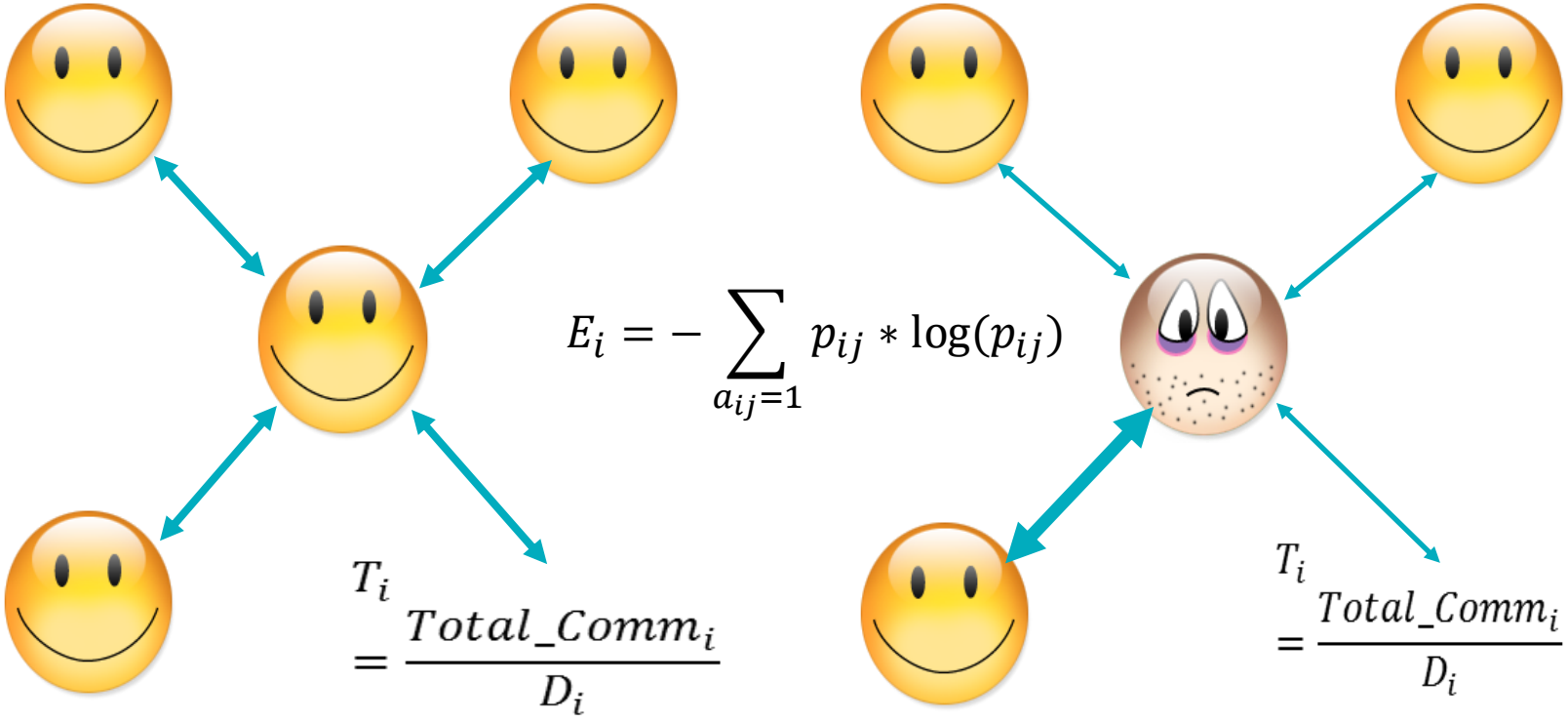
$$T_i = \frac{Total\_Comm_i}{D_i}$$





社交网络变量的产生——个体的信息熵

人均通话时长分布
entropy



$$E_i = -\sum_{a_{ij}=1} p_{ij} * \log(p_{ij})$$

$$T_i = \frac{Total\_Comm_i}{D_i}$$

$$T_i = \frac{Total\_Comm_i}{D_i}$$

变量结构

☆ 因变量：用户是否流失或离网（流失为1）

☆ 自变量

✓ 核心指标（Key variables）

- 通话人数：统计月拨打/接听电话的人数
- 通话人数变化率：当月比上月通话人数的变化
- 人均通话时长：通话总分钟数/通话总人数
- 人均通话时长的分布：用户与每个人的通话时长在总通话时长中的分布情况，该值越大说明用户的通话越分散，该值越小，说明用户的通话越集中

✓ 传统指标（Traditional variables）

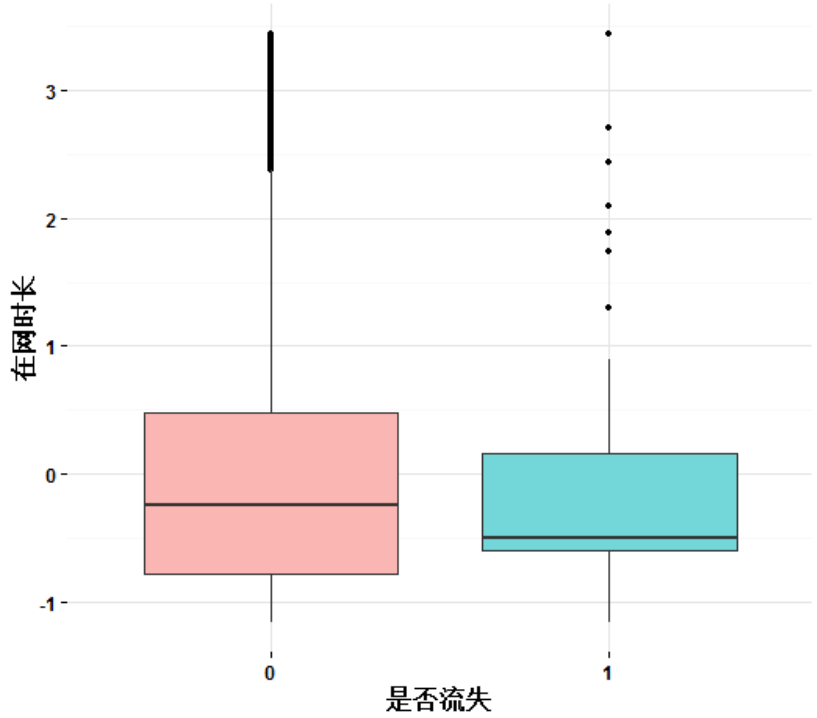
- 在网时长：统计月时间减去入网时间
- 当月花费：统计月的消费金额
- 花费变量率：当月比上月花费的变化

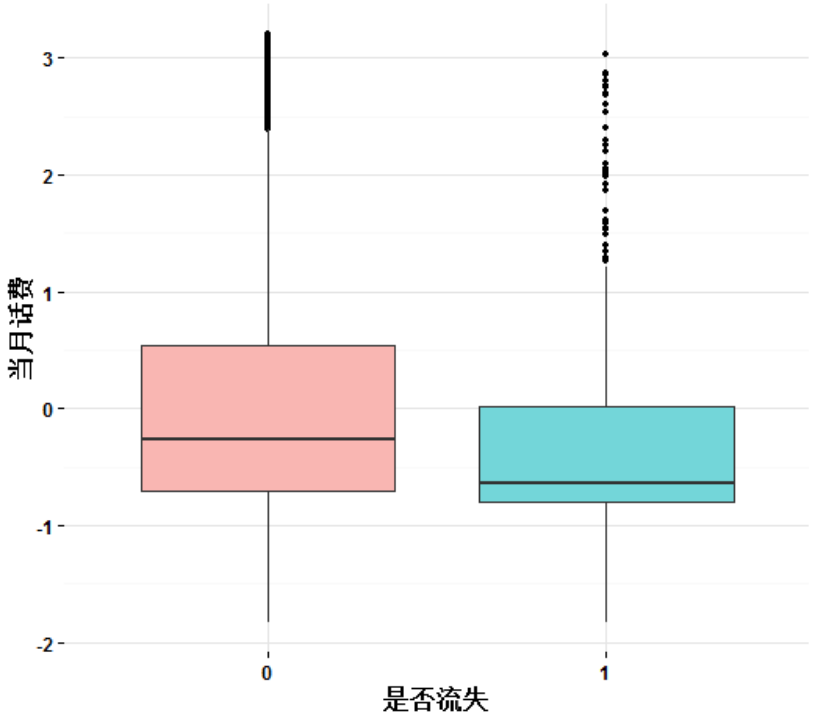
变量介绍

	变量名	详细说明	备注
因变量（下月）	是否流失	1=流失；0=不流失	流失率1.27%
自变量（当月）	在网时长	连续变量，单位：天	数据截取日减去入网时间
	当月花费	连续变量，单位：元	统计当月的总花费
	个体的度	连续变量，单位：人数	$D_i = \sum_{j \neq i} a_{ij}$
	联系强度	连续变量：分钟/人	$T_i = \frac{Total\_Comm_i}{D_i}$
	个体信息熵	连续变量	$E_i = -\sum_{a_{ij}=1} p_{ij} * \log(p_{ij})$
	个体度的变化	连续变量，单位：%	（当月个体的度-上月个体的度）/上月个体的度
	花费的变化	连续变量，单位：%	（当月花费-上月花费）/上月花费

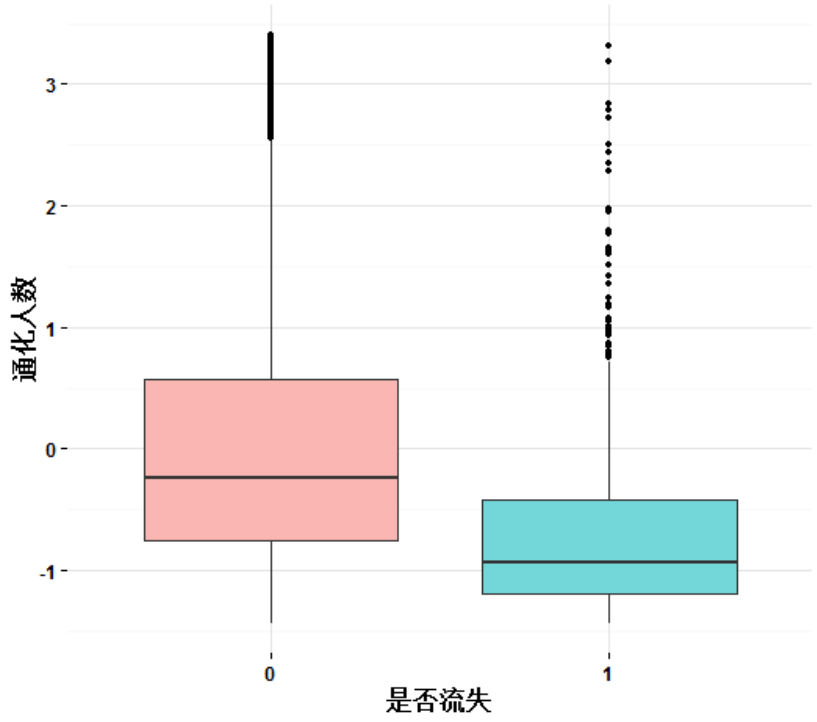
注：由于本案例关注的是【预警】模型，所以在后续的建模中我们关注的是当月的一些自变量是否会对下月的流失产生影响，这样模型可以做到提前预警。

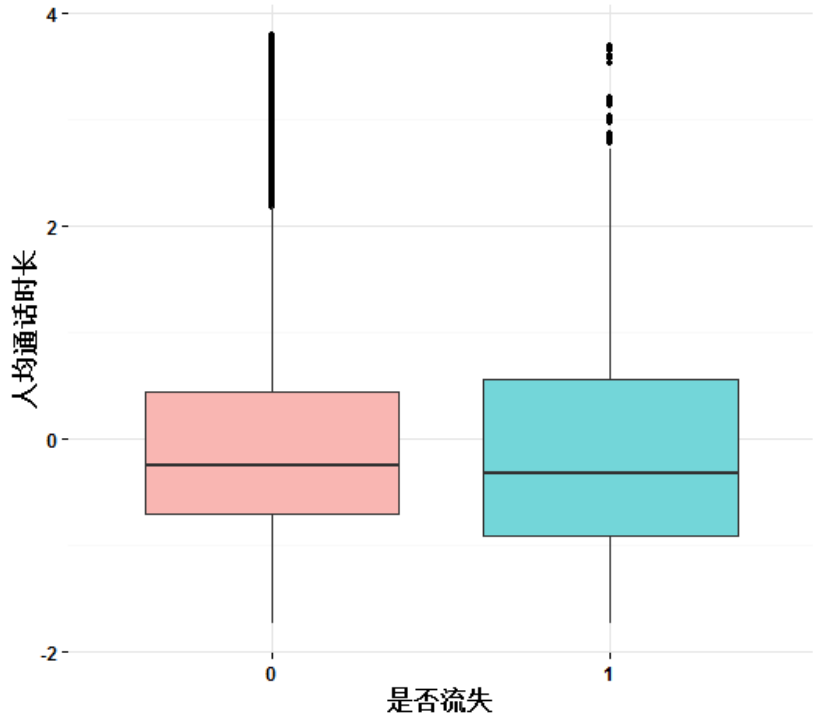
描述分析



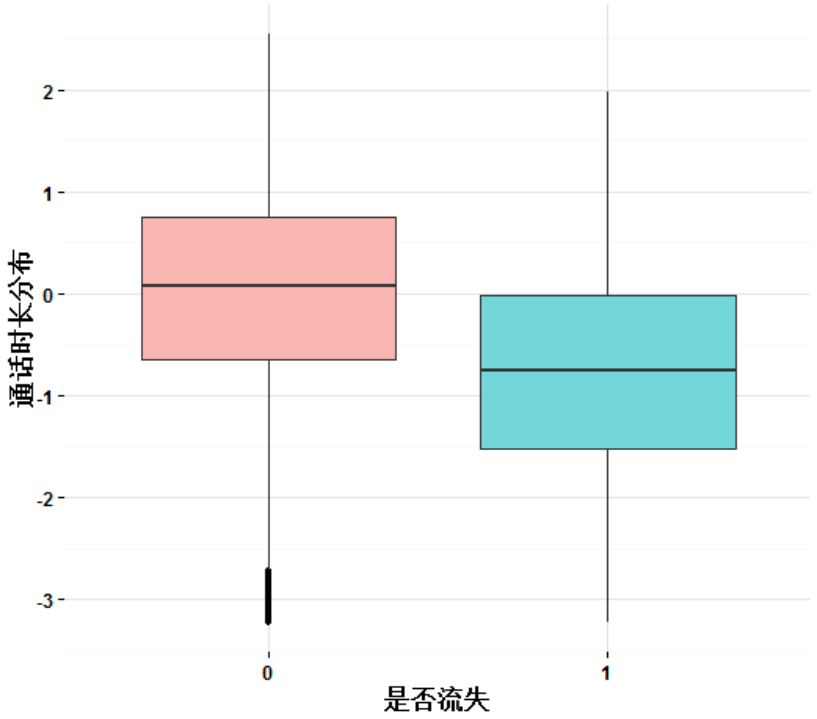


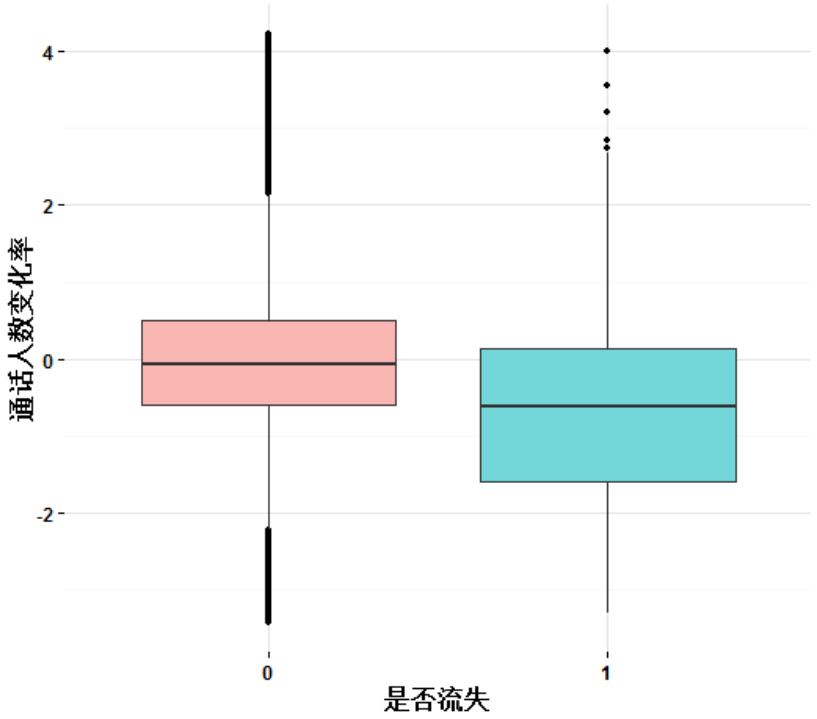
描述分析





描述分析

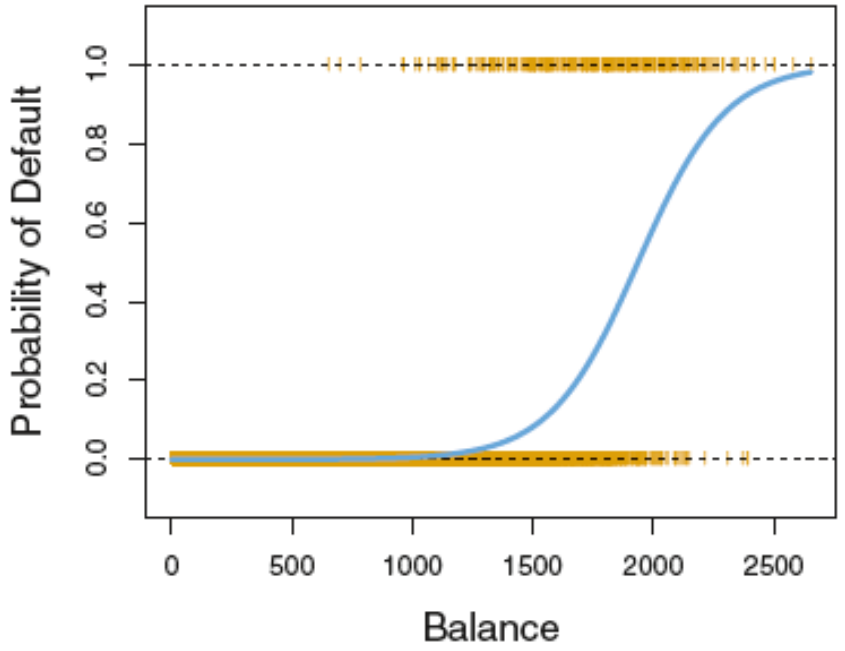




回归模型

逻辑回归

- Logistic Function: $P(Y=1|X) \stackrel{\text{def}}{=} p(X) = \frac{e^{X'\beta}}{1+e^{X}}$
- Log odds: $\log\left(\frac{p(X)}{1-p(X)}\right) = X'\beta$
- β的解读：
  - 符号
  - log odds 的变化
  - 对于$p(X)$ 的影响：与X有关



Probit模型和Logistic模型

$$F_{\varepsilon}(t)=\Phi(t)$$

$$F_{\varepsilon}(t)=\frac{\exp(t)}{1+\exp(t)}$$

似然函数

$$P(Y_i|X_i) = \begin{cases} \dfrac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}, & \text{如果} Y_i = 1 \\ \dfrac{1}{1+\exp(X_i'\beta)}, & \text{如果} Y_i = 0 \end{cases}$$

$$P(Y_i|X_i) = \left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\}^{Y_i} \left\{\frac{1}{1+\exp(X_i'\beta)}\right\}^{1-Y_i}$$

似然函数

$$\prod_{i=1}^{n} P(Y_i|X_i) = \prod_{i=1}^{n} \left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\}^{Y_i} \left\{\frac{1}{1+\exp(X_i'\beta)}\right\}^{1-Y_i}$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \log\{P(Y_i|X_i)\} = \sum_{i=1}^{n} \left[Y_i \log\left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\} + (1-Y_i)\left\{\frac{1}{1+\exp(X_i'\beta)}\right\}\right]$$

最大化似然函数得到估计值

似然比检验

- $LR= -2 \max_{\beta_0} L(\beta_0 \cdot \beta_1=0)-(-2 \max_{\beta_0, \beta_1} L(\beta_0 \cdot \beta_1))$

  $=Deviance_0-Deviance_1$

- 在原假设$\beta_1=0$下, LR 近似服从自由度为DF的卡方分布, 而 DF 是包含在β_1中的变量个数。

R函数 glm

☆R中的广义线性回归语句*glm*

☆语法为：*glm(formula, **family = gaussian**, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = glm.control(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)*

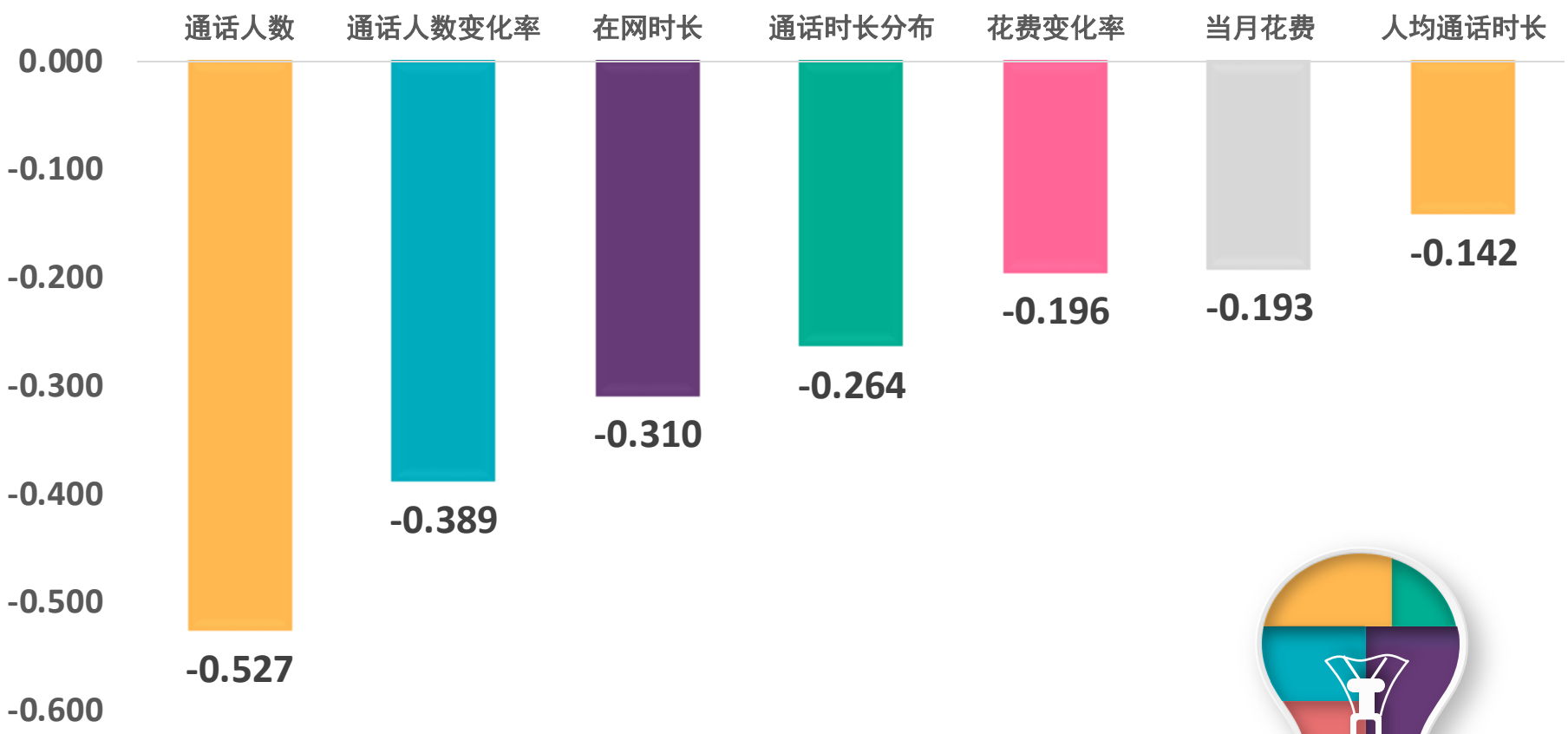
☆与lm不同之处就在于参数***family***

☆这个参数的作用在于定义一个族以及连接函数，使用该连接函数将因变量的期望与自变量联系起来

☆family= binomial(link=logit)表示引用了二项分布族binomial中的logit连接函数

回归结果

	估计值	标准误差	P值
常数项	-4.712	0.061	<0.001
在网时长	-0.310	0.057	<0.001
当月花费	-0.193	0.052	<0.001
通话人数	-0.527	0.111	<0.001
人均通话时长	-0.142	0.040	<0.001
通话时长分布	-0.264	0.074	<0.001
花费变化率	-0.196	0.043	<0.001
通话人数变化率	-0.389	0.043	<0.001


通话人数
通话人数变化率
在网时长
通话时长分布
花费变化率
当月花费
人均通话时长
0.000
-0.100
-0.200
-0.300
-0.400
-0.500
-0.600
-0.527
-0.389
-0.310
-0.264
-0.196
-0.193
-0.142


模型选择与预测

模型的选择

☆最小化KL距离 $AIC = -2 \times \log\{L(\beta_0, \beta_1)\} + 2 \times df$

☆最大化后验概率 $BIC = -2 \times \log\{L(\beta_0, \beta_1)\} + \log(n) \times df$

☆R函数 step, BIC需设置k=log(n)

☆BIC选择标准更严格，变量更少。

模型预测

$$P(Y_i^* = 1|X_i^*) \approx p(X_i^{*\prime}\hat{\beta}) = \frac{\exp(X_i^{*\prime}\hat{\beta})}{1 + \exp(X_i^{*\prime}\hat{\beta})}$$

$$\hat{Y}_i^* = \begin{cases} 1, & \text{如果 } p(X_i^{*\prime}\hat{\beta}) > \alpha \\ 0, & \text{如果 } p(X_i^{*\prime}\hat{\beta}) \leq \alpha \end{cases}$$

预测评估

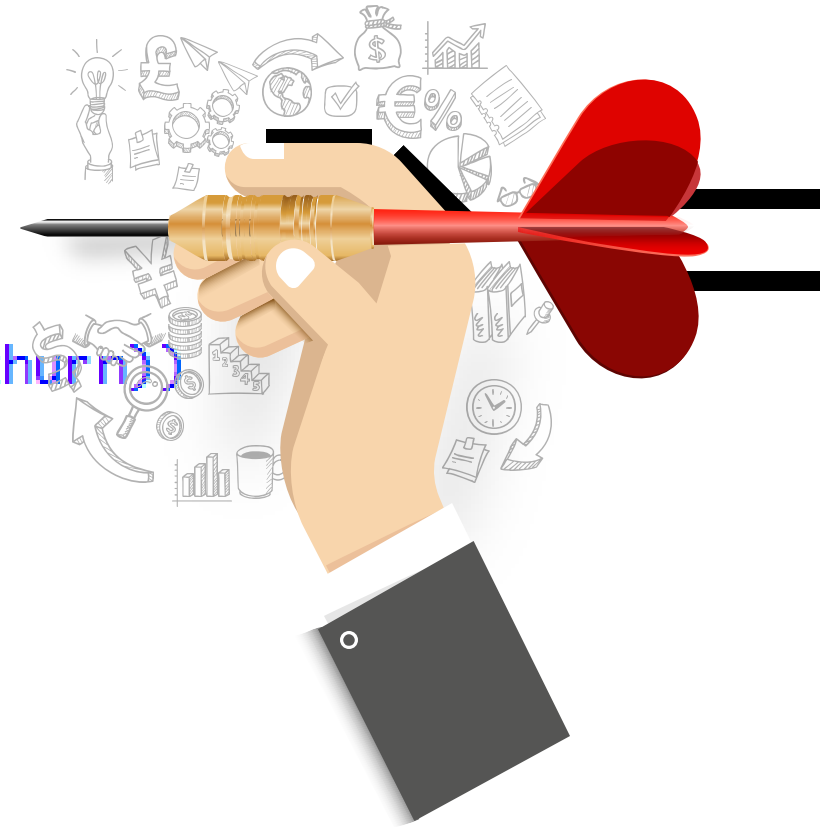
$$\mathrm{MCR} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}\left(Y_i^* \neq \hat{Y}_i^*\right)$$

R预测结果

```
> ypre1=1*(Yhat>0.5)
> table(ypre1,dat2$churn)

ypre1     0     1
    0 46001   447
> ypre2=1*(Yhat>mean(dat2$churn))
> table(ypre2,dat2$churn)

ypre2     0     1
    0 25927    90
    1 20074   357
```



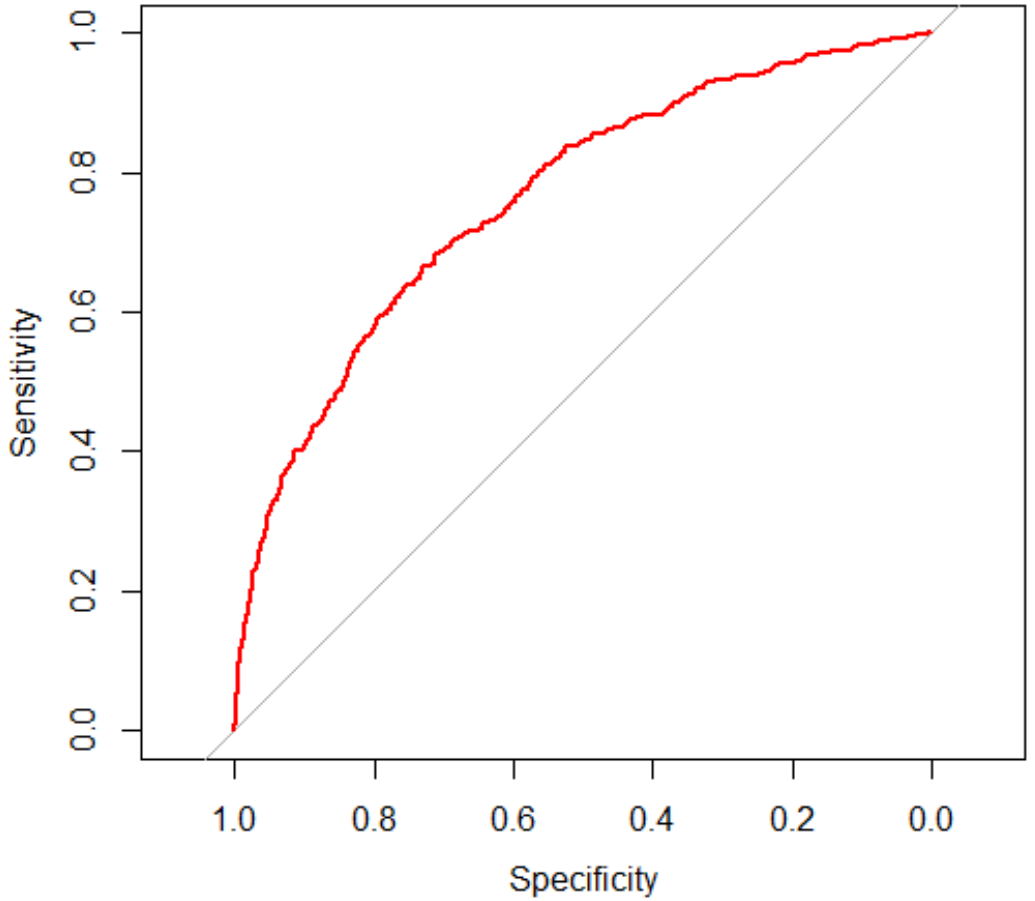
定义两种分类错误P(state=1|X)>α

		True Response	
		0	1
Prediction	0	a	b
	1	c	d

False Positive Rate（FPR） = c/(a+c)

True Positive Rate（TPR） = d/(b+d)

ROC与AUC



TPR

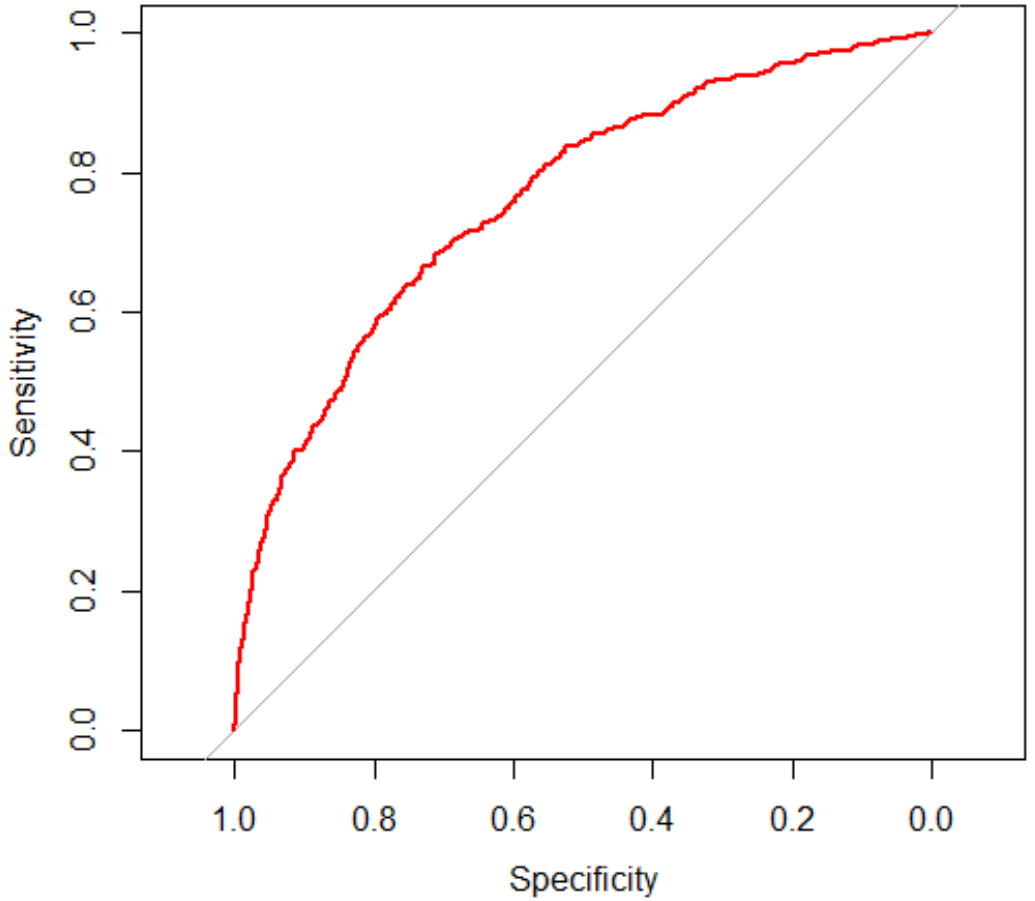
AUC估计值：0.7622

1-FPR

AUC的计算

$$\widehat{AUC} = \frac{\sum_{i\in D_1}\sum_{j\in D_0}\{I[p(X_i^*\hat{\beta}) > p(X_j^*\hat{\beta})] + 0.5I[p(X_i^*\hat{\beta}) = p(X_j^*\hat{\beta})]\}}{n_1 n_0}$$

ROC与AUC

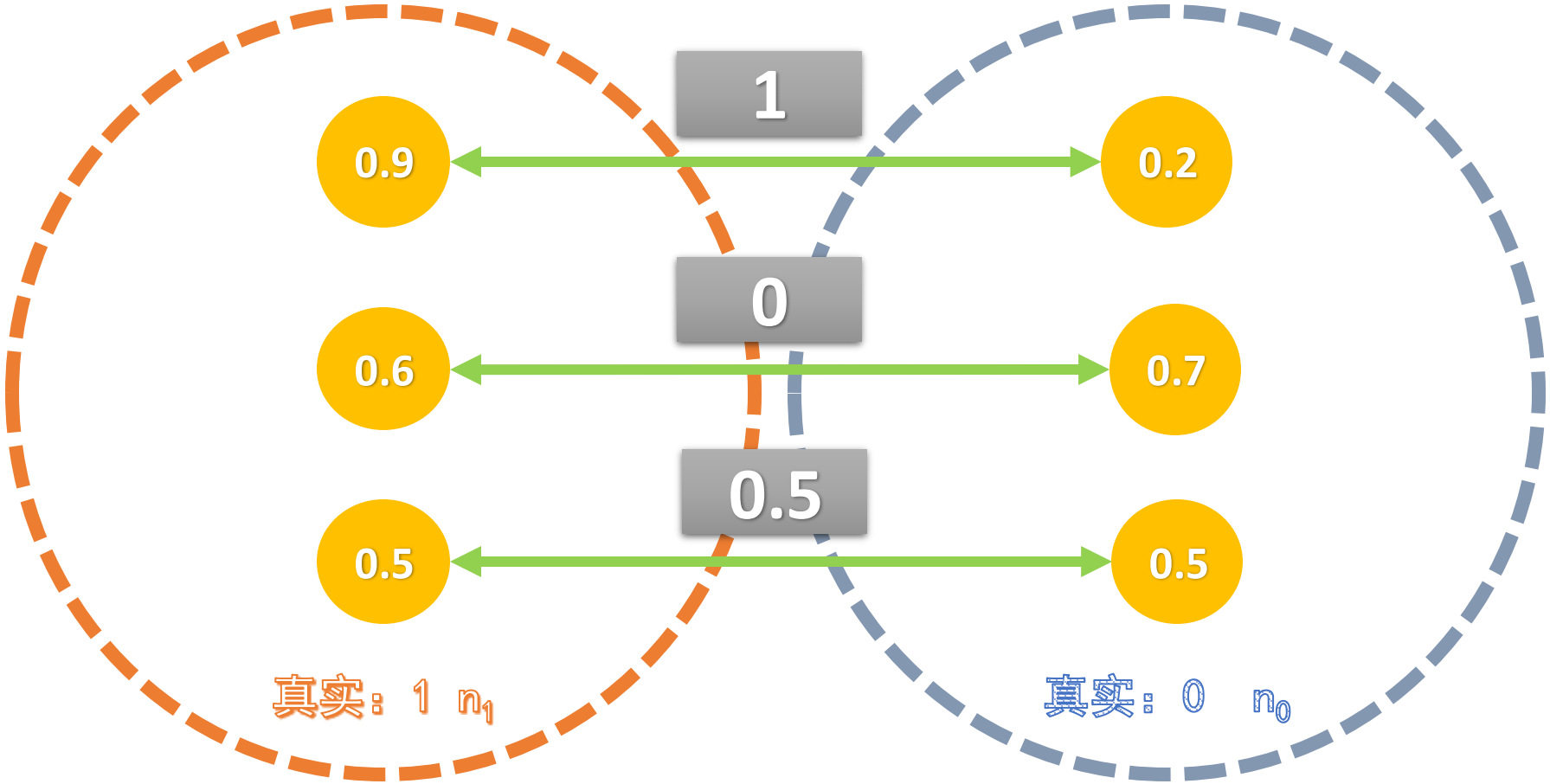


TPR

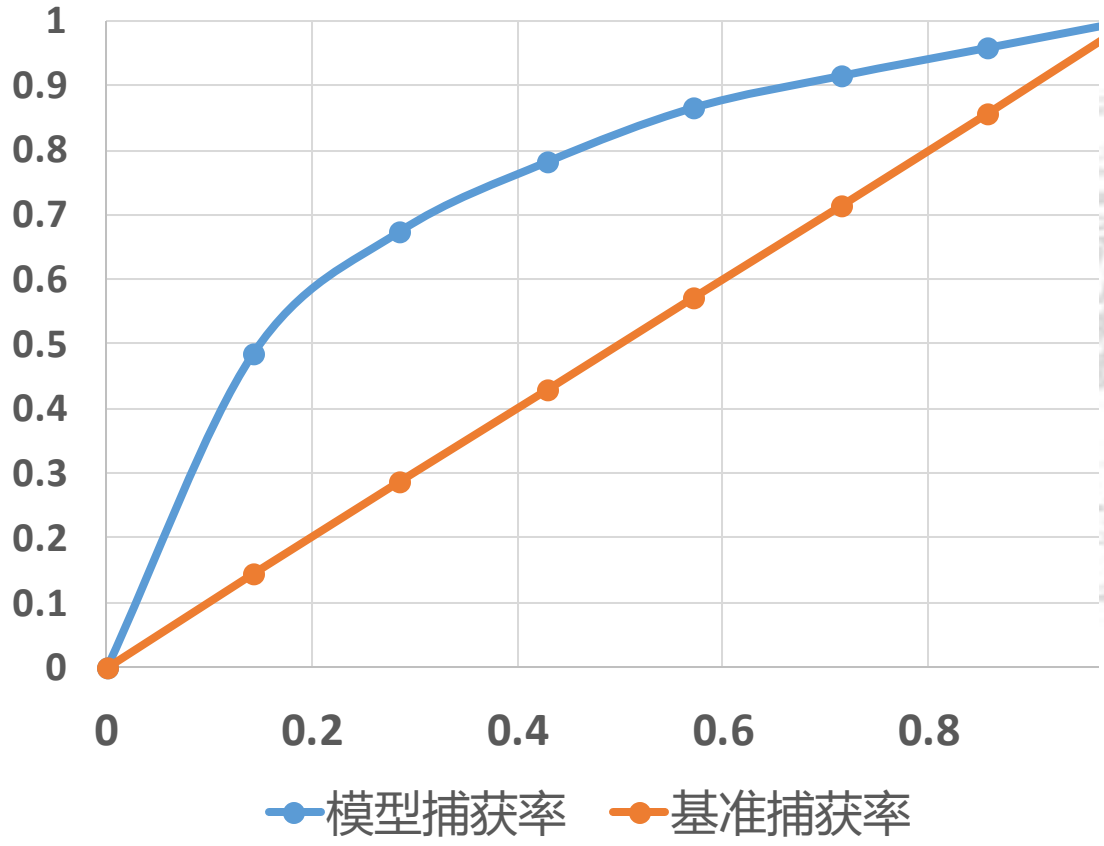
AUC估计值：0.7622

1-FPR

AUC的计算



覆盖率与捕获率曲线



- ✓ 为了评估模型的预测效果，我们提出了**覆盖率-捕获率**曲线
- ✓ 根据模型给出每个样本的预测流失概率值
- ✓ 按照预测值**从高到低**对样本进行排序
- ✓ 例如只覆盖前10%的样本，计算对应的真实流失的样本数占所有流失样本数的比例，记为捕获率
- ✓ 以此类推，可以覆盖不同比例的样本，就可以计算不同的覆盖率对应的捕获率，从而得到覆盖率捕获率曲线
- ✓ 如果在**较低的覆盖率**情况可以获得**较高的捕获率**，那么说明模型的精度比较高
- ✓ 本案例中**20%的覆盖率**差不多可以达到**60%的捕获率**，说明覆盖预测概率值最高的前20%人，可以抓住60%的流失客户

训练集与测试集的划分

A 10折交叉验证

B 随机拆分训练集测试集

C In-sample和Out-of-Sample的结果

小结：模型建立