法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

违约预测模型的后续工作

目录

从概率到分数

模型的验证监控

评分卡的其他细节

申请评分卡的使用

从概率到分数

评分卡分数的计算

- 评分卡模型用分数衡量逾期率的大小
- 分数的计算

$$score = Base\ Point + \frac{PDO}{\ln(2)}(-y)$$

其中，$y = \text{logit(p)} = \log(\frac{p}{1-p})$

- Base point: 基准分，无实际意义
- PDO: points to double odds, 好坏比每升高1倍，评分增加一个PDO的单位

从概率到分数

评分卡分数的计算(续)

PDO的证明：

$$y = \text{logit(p)} = \log\left(\frac{p}{1-p}\right) = \log(\frac{p_{bad}}{p_{good}})$$

当好坏比升高一倍时，

$$-y' = -\text{logit(p')} = \log\left(\frac{1-p'}{p'}\right) = \log\left(2\times\frac{p_{good}}{p_{bad}}\right) = \log(2) + \log\left(\frac{p_{good}}{p_{bad}}\right) = \log(2) - y$$

$$score' = base\ point + \frac{PDO}{log(2)}\times(-y') = base\ point + \frac{PDO}{log(2)}\times(\log(2) - y$$

$$) = base\ point + \frac{PDO}{log(2)}(-y) + PDO = score + PDO$$

注：也可以用其他的好坏比率，比如PTO(point to triple odds)，表示好坏比升高两倍，分数上升PTO个单位

从概率到分数

分数的分级(Pooling)

在评级模型中，得到分数后需要对分数进行分级(pooling)操作，将评分人群划分有限的几个组别

划分的方法

将分数视为连续变量，采用监督式方法例如best-KS或者ChiMerge进行有序划分，且一般划分为10组左右。

实际违约率

将评分卡结果进行分层后，每层对应一个实际违约率

$$assigned\ PD_i = \frac{\#\{第i^{th}层中，在表现期内违约的样本\}}{\{第i^{th}层的总样本\}}$$

同时，获取过去较长时间内(比如5～10)的长期实际违约率(long run PD)，以此为基准，得到较准率

$$scaling = \max\{1, \frac{long\ run\ PD}{total\ actual\ PD}\}$$

从概率到分数

分数的分级(续)

预期违约率

将建模样本的实际违约率乘以较准率，得到指预期违约率

$$assigned\ PD_i = actual\ PD_i\ in\ trainign\ data \times scaling$$

注：

- 预期违约率在评分卡模型生存周期内是固定的，而实际逾期率是变化的
- 在评分卡生存周期内，预期违约率要求不低于实际违约率。当这一条件不满足时，需要做假设检验(见本节课第二部分)

目录

从概率到分数

模型的验证与监控

评分卡的其他细节

申请评分卡的使用

模型的验证与监控

模型的验证

模型的验证(model validation)

评分卡模型训练完之后，需要在验证集上进行验证。通常，需要选择跟训练样本所在的日期不同的日期的申请样本做为验证集，称为OOT(out of date test)。这是为了验证模型在时间上的效力跟稳定性。

模型的监控(model monitoring)

模型在部署并执行后，需要定期对模型的表现进行监控，以保证模型的各项性能不会出现恶化。当某项指标持续恶化时，需要按需对模型进行调整甚至重新开发。模型的监控与验证基本是一致的，主要包含了对模型稳定性、准确性和排序性的监测。

模型的验证与监控

模型对违约与非违约人群的区分度

申请评分卡的目的

- 尽可能地区分出潜在的逾期人群和非逾期人群

区分人群的手段

- 分数的高、低

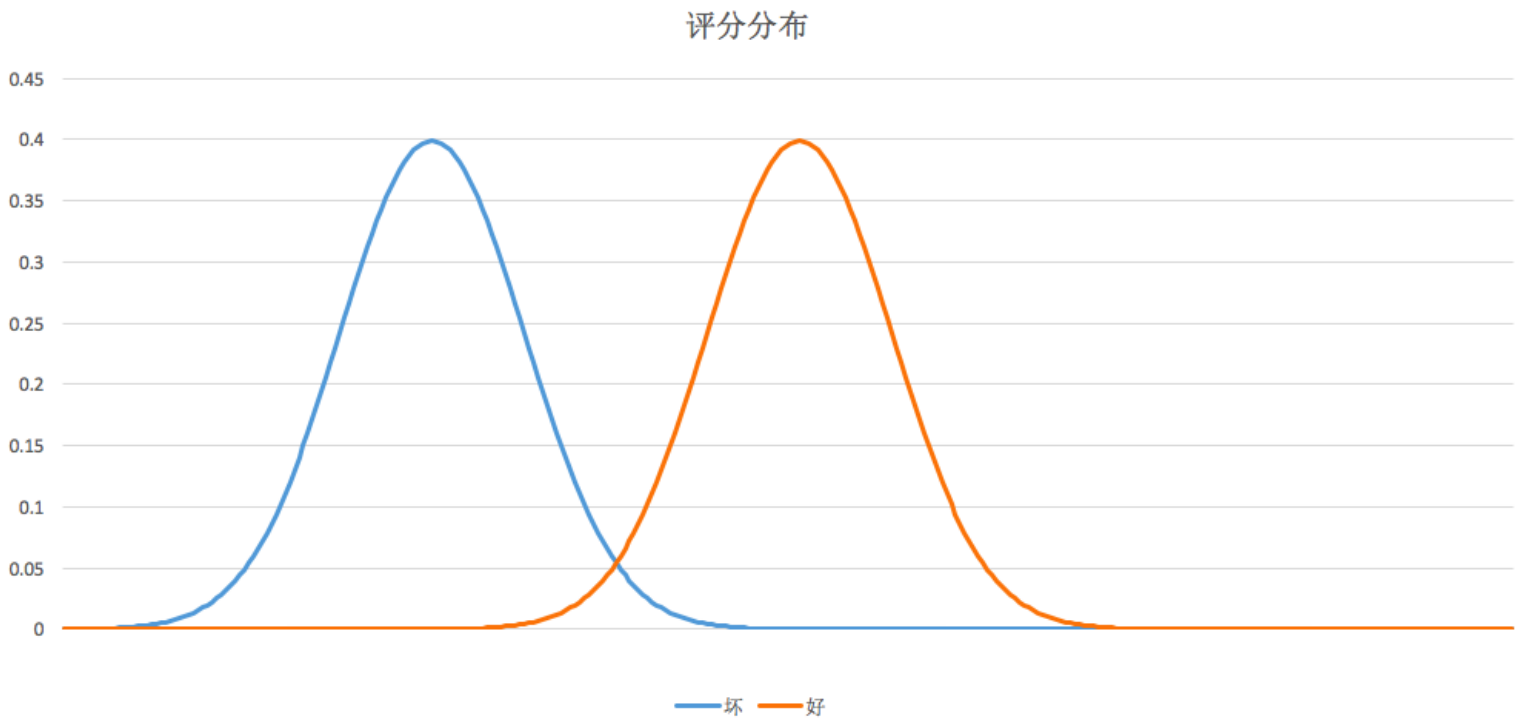
区分度的衡量

- KS，阈值=30%，KS值越高表明区分能力越强(见)，参见第6节课的介绍
- Gini Score
- Divergence Score

模型的验证与监控

模型对违约与非违约人群的区分度(续)

通常来讲，评分分布会出现双峰



模型的验证与监控

Gini Score

都是衡量好坏两批人群分数分布的差异性

Gini Score公式

$$Gini = \sum \frac{n_i}{N}\left(1 - p_i^2 - (1 - p_i)^2\right) = 2\sum \frac{n_i}{N}(1 - p_i)p_i$$

其中n_i是评分卡分组后的每组样本量，N是总样本量，p_i是每组的实际逾期率

Gini越低，表示划分后纯度越高

注：Gini Score 同时也依赖于分组方式。需要固定分组方式。

模型的验证与监控

Divergence Score

Divergence Score公式

- 基于评分群体的统计，与分组无关

$$Divergence = \frac{(\mu_{good} - \mu_{bad})^2}{\frac{1}{2} \times (var_{good} + var_{bad})}$$

模型的验证与监控

模型的准确度

评分模型通过分数的高低来判断申请者信用资质的好坏，意味着表现期内逾期人群的分数需要集中在低分段、非逾期人群的分数需要集中在高分段，形成一个“有序”的结果

AR

坏样本累计速度相对于全体样本累积速度的差异，参见第6节课的介绍

Kendal’s Tau

实际逾期率与分数的单调性，即分数越高，实际逾期率越低，参见第6节课的介绍

模型的验证与监控

模型的稳定性

评分卡结果，要求在人群分布不变的情况下保持一定的稳定性。可以用PSI衡量评分分布的波动：

$$PSI = \sum (C_i - B_i)\log(\frac{C_i}{B_i})$$

C_i是现阶段第i^{th}个组的人数占全部人数的比例

B_i是模型开发阶段阶段第i^{th}个组的人数占全部人数的比例

注：

- 该指标同时依赖于分数的划分方式
- 当PSI超过阈值(通常是25%)时，表明人群分布发生变化，或者评分稳定性减弱，需要重新评估模型的有效性

模型的验证与监控

预期违约率的保守性

从风险评估的角度，预期违约率需要比实际违约率高一些，称为保守估计(conservative estimate)。在监控工作中，当发现分组后第i组的预期违约率低于实际违约率时，要做二项检验

$$H0: Assigned\ PD_i \geq Actual\ PD_i$$

against

$$H1: Assigned\ PD_i < Actual\ PD_i$$

计算二项分布的p值：

$$p - value = Binomial(d, n, p)$$

d: 某个分组的实际违约人数

n: 该组的所有人数

p: 预期违约率

当有若干个分组出现assigned PD < actual PD 时，意味着现有的评分卡不足以反应真实的违约率，可能会影响模型在授信、调额方面的使用。

目录

从概率到分数

模型的验证与监控

评分卡的其他细节

申请评分卡的使用

评分卡的其他细节

模型的部署

基于逻辑回归的评分卡模型在完成了开发、验证和审计后，可以进入到部署阶段。不同的使用场景，应该选择不同的部署方式。

实时计算

用于线上申请行为，且模型部分依赖于三方数据。当申请进件信息传入到部署模型的服务器时，服务器会从后台数据库里实时查询相关信息(包括调用三方数据)，将数据转换成特征、完成分箱操作和WOE编码，带入模型。

- 优点：

准确度较高

- 缺点：

变量计算不宜涵盖太长的时间切片，且本机构、第三方数据源接口不能有延时

评分卡的其他细节

模型的部署(续)

非实时计算

用于线下申请行为。当申请进件信息传入到部署模型的服务器时，服务器会根据传入的数据计算分数。

- 优点：
  - 服务器并发压力小
  - 可人工干预
  - 特征跨度不受限制

缺点：

- 准确度较差，不能抓住突发事件(比如近期的多头)

评分卡的其他细节

拒绝推断

评分卡模型在开发过程中，选取的数据都是历史申请准入后、有实际表现的数据。而在使用时，被准入的客户可以观测到实际表现，被拒绝的客户则无法推断。换言之，我们可以推断评分卡准入客户的好坏情况，却无法推断拒绝客户。

目前尚未有很好的办法解决这个难题，一般可以借鉴的有：

方法一

在审核阶段，随机抽取少量低分段人群给予准入，以此来推断评分卡在低分段人群的表现

代价：会有违约损失

方法二

跟踪被拒绝掉的客群在其他平台上的表现

代价：跟踪的成本极高

目录

从概率到分数

模型的验证与监控

评分卡的其他细节

申请评分卡的使用

申请评分卡的使用

准入与拒绝

业务人员、风控人员根据评分卡的结果，对于申请进件准入或者拒绝。一般可以根据2条原则进行准入分的设定：

对于非首次使用评分卡的机构

- 当以提高业务量为目标时，在不降低坏账率的前提下，降低现有的准入分
- 当以降低违约率为目标时，在保持跟之前的人数一样多的情况下，提高准入分

对于首次使用评分卡的机构

领导决定通过率！

申请评分卡的使用

授信额度

预先设定好基础额度base limit(B)，盖帽额度hat limit(H)，托底额度 floor limit (F)。评分最高的区间对应的预期违约率是P_{min}，评分最低的区间对应的预期违约率是P_{max}，占比最高的区间对应的预期违约率是P_0，某一条进件对应的预期违约率是P_1，则该进件对应的授信度是：

如果$P_1 > P_0$

score	最低分	本次进件分	众数分
违约概率	P_{max}	P_1	P_0
调节因子	F/B	f=1+(F/B-1)/($P_{max} - P_0$)*($P_1 - P_0$)	1
额度	F	B*f	B

申请评分卡的使用

授信额度(续)

如果$P_1 < P_0$

score	众数分	本次进件分	最高分
违约概率	P_0	P_1	P_{min}
调节因子	1	f=1+(H/B-1)/($P_{min} - P_0$)*($P_1 - P_0$)	H/B
额度	B	B*f	H

申请评分卡的使用

利率定价

在利率定价模型中，

年利率 = 基础利率×渠道调节系数×客户信用调节系数×产品调节系数

其中客户信用调节系数受到预期违约率的影响

客户信用等级调节系数

客户基准逾期率	客户整体逾期率加权
客户信用最大调整系数	按信用等级前X%客户的加权逾期率，运算得出最大、小调整系数
客户信用调整系数	根据客户逾期率及最大/小加权逾期率的差距，得出该客户信用调整系数

例如：

由于每个客户的违约概率差异较大，容易出现极值，需要固定调整系数上下限

客户整体加权逾期率为1.5%，前5%客户加权逾期率为0.5%，则当客户逾期率为1.5%时，信用等级调节系数为1，当客户逾期率<=0.5%时，信用等级调节系数为0.33

逾期率极大值	0.50%	客户信用最大调整系数	0.3
客户基准逾期率	1.50%	基准值	1
逾期率极小值	2.25%	客户信用最小调整系数	1.5

	客户逾期率	过程	最终系数
当客户逾期率 <基准逾期率	0.80%	'=1+(0.3-1)/(0.5%-1.5%)*(0.8%-1.5%)	0.51
当客户逾期率>基准逾期率	2.10%	=1+(1-1.5)/(1.5%-2.25%)*(2.1%-1.5%)	1.4

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop