风控指标

WOE & IV

WOE（Weight of Evidence）常用于特征变换，IV（Information Value）则用来衡量特征的预测能力

> *1. WOE describes the **relationship** between a predictive variable and a binary target variable.*
>
> *2. IV measures the **strength** of that relationship.*

$$WOE_i = \ln\left(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T}\right) = \ln\left(\frac{Bad_i}{Bad_T}\right) - \ln\left(\frac{Good_i}{Good_T}\right)$$

IV 可认为是**WOE的加权和**

$$\begin{aligned} IV_i &= \left(\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}\right) * WOE_i \\ &= \left(\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}\right) * \ln\left(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T}\right) \\ IV &= \sum_{i=1}^{n} IV_i \end{aligned}$$

WOE和IV的计算步骤

- **step 1**. 对于连续型变量，进行分箱（binning），可以选择等频、等距，或者自定义间隔；对于离散型变量，如果分箱太多，则进行分箱合并。
- **step 2**. 统计每个分箱里的好人数(bin_goods)和坏人数(bin_bads)。
- **step 3**. 分别除以总的好人数(total_goods)和坏人数(total_bads)，得到每个分箱内的边际好人占比(margin_good_rate)和边际坏人占比(margin_bad_rate)。
- **step 4**. 计算每个分箱里的 $WOE = \ln\left(\frac{\text{margin}_{-badrate}}{\text{margin}_{-goodrate}}\right)$
- **step 5**. 检查每个分箱（除null分箱外）里woe值是否满足**单调性**，若不满足，返回step1。注意⚠️：null分箱由于有明确的业务解释，因此不需要考虑满足单调性。
- **step 6**. 计算每个分箱里的IV，最终求和，即得到最终的IV。 备注：好人 = 正常用户，坏人 = 逾期用户

buckect	min_score	max_score	obs	bad	good	bad_rate	good_rate	margin_bad_rat	margin_good_r	odds(bad/good)	woe	iv
1	0	18	1390	70	1320	5.0%	95.0%	39.8%	15.1%	0.053	0.9692	0.2392
2	18	23	1070	33	1037	3.1%	96.9%	18.8%	11.9%	0.032	0.4585	0.0316
3	23	28	1162	20	1142	1.7%	98.3%	11.4%	13.1%	0.018	-0.1387	0.0023
4	28	34	1162	15	1147	1.3%	98.7%	8.5%	13.1%	0.013	-0.4308	0.0198
5	34	44	1212	12	1200	1.0%	99.0%	6.8%	13.7%	0.010	-0.6991	0.0482
6	44	100	1153	9	1144	0.8%	99.2%	5.1%	13.1%	0.008	-0.9390	0.0748
7	null	null	1775	17	1758	1.0%	99.0%	9.7%	20.1%	0.010	-0.7325	0.0765
总计	0	100	8924	176	8748	2.0%	98.0%	100.0%	100.0%	0.020	0.0000	0.2624

从相对熵角度理解IV

我们把PSI、IV的计算公式放在一起进行对比

$$IV = \sum_{i=1}^{n} \left(\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T} \right) * \ln \left(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T} \right)$$

$$PSI = \sum_{i=1}^{n} \left(\frac{Actual_i}{Actual_T} - \frac{Expect_i}{Expect_T} \right) * \ln \left(\frac{Actual_i}{Actual_T} / \frac{Expect_i}{Expect_T} \right)$$

我们会发现两者形式上是完全一致的，这主要是因为它们背后的**支撑理论都是相对熵**。我们可以归纳为：

> PSI衡量预期分布和实际分布之间的**差异性**，IV把这两个分布具体化为好人分布和坏人分布。IV指标是在**从信息熵上比较好人分布和坏人分布之间的差异性**。

IV 越大越好，PSI 越小越好。

KS

KS统计量是基于经验累积分布函数（Empirical Cumulative Distribution Function，ECDF）建立的，一般定义为：

$$ks = \max \{|cum\left(bad\_\text{rate}\right) - \text{cum}(\text{good\_rate})|\}$$

KS的计算过程及业务分析

- **step 1**. 对变量进行分箱（binning），可以选择等频、等距，或者自定义距离。
- **step 2**. 计算每个分箱区间的好账户数(goods)和坏账户数(bads)。
- **step 3**. 计算每个分箱区间的累计好账户数占总好账户数比率(cum_good_rate)和累计坏账户数占总坏账户数比率(cum_bad_rate)。
- **step 4**. 计算每个分箱区间累计坏账户占比与累计好账户占比差的**绝对值**，得到KS曲线。也就是: $ks = |\text{cum\_goodrate} - \text{cum\_badrate}|$
- **step 5.** 在这些绝对值中**取最大值**，得到此变量最终的KS值。

bucket	min_score	max_score	total	total_rate	goods	bads	bad_rate	cum_bad_rate	cum_good_rate	ks（%）	max_ks
1	0.03	0.23	90	10.1%	2	88	97.78%	25.73%	0.36%	25.4	
2	0.23	0.39	89	10.0%	24	65	73.03%	44.74%	4.74%	40.0	
3	0.39	0.53	89	10.0%	40	49	55.06%	59.06%	12.02%	47.0	
4	0.53	0.60	90	10.1%	45	45	50.00%	72.22%	20.22%	52.0	
5	0.60	0.65	88	9.9%	52	36	40.91%	82.75%	29.69%	53.1	<<<<<<
6	0.65	0.71	89	10.0%	63	26	29.21%	90.35%	41.17%	49.2	
7	0.71	0.79	89	10.0%	67	22	24.72%	96.78%	53.37%	43.4	
8	0.79	0.85	106	11.9%	97	9	8.49%	99.42%	71.68%	27.4	
9	0.85	0.90	85	9.5%	83	2	2.35%	100.00%	86.15%	13.8	
10	0.90	0.98	76	8.5%	76	0	0.00%	100.00%	100.00%	0.0	

PSI

在风控中，**稳定性压倒一切**。原因在于，一套风控模型正式上线运行后往往需要很久（通常一年以上）才会被替换下线。如果模型不稳定，意味着模型不可控，对于业务本身而言就是一种不确定性风险，直接影响决策的合理性，这是不可接受的。

稳定性是有参照的，因此**需要有两个分布——实际分布**（actual）**和预期分布**（expected）。其中，在建模时通常以训练样本（In the Sample, INS）作为预期分布，而验证样本通常作为实际分布。验证样本一般包括样本外（Out of Sample，OOS）和跨时间样本（Out of Time，OOT）。

$$psi = \sum_{i=1}^{n} (A_i - E_i) * \ln(A_i / E_i)$$

> **PSI = SUM((实际占比 - 预期占比) * ln(实际占比 / 预期占比))**

PSI 的计算过程

- **step 1**. 将**变量预期分布**（excepted）进行**分箱**（binning）离散化，统计各个分箱里的样本占比。 注意⚠️： a) 分箱可以是等频、等距或其他方式，分箱方式不同，将导致计算结果略微有差异； b) 对于**连续型**变量（特征变量、模型分数等），分箱数需要设置合理，一般设为10或20；对于离散型变量，如果分箱太多可以提前考虑合并小分箱；分箱数太多，可能会导致每个分箱内的样本量太少而失去统计意义；分箱数太少，又会导致计算结果精度降低。
- **step 2**. 按相同分箱区间，对**实际分布（actual）**统计各分箱内的样本占比。
- **step 3**. 计算各分箱内的**A - E**和**Ln(A / E)**，计算**index = (实际占比 - 预期占比) * ln(实际占比 / 预期占比)** 。
- **step4**. 将各分箱的index进行求和，即得到最终的PSI。

Score Range	Actucal %	Expected %	A - E	A / E	Ln(A / E)	Index
000-169	7%	8%	-1%	0.8750	-0.13353	0.0013
170-179	8%	10%	-2%	0.8000	-0.22314	0.0045
180-189	7%	9%	-2%	0.7778	-0.25131	0.0050
190-199	9%	13%	-4%	0.6923	-0.36772	0.0147
200-209	11%	11%	0%	1.0000	0.00000	0.0000
210-219	11%	10%	1%	1.1000	0.09531	0.0010
220-229	10%	9%	1%	1.1111	0.10536	0.0011
230-239	12%	10%	2%	1.2000	0.18232	0.0036
240-249	11%	11%	0%	1.0000	0.00000	0.0000
250+	14%	9%	5%	1.5556	0.44183	0.0221
					Population Stability Index =	0.0533

在计算得到PSI指标后，这个数字又代表什么业务含义呢？**PSI数值越小，两个分布之间的差异就越小，代表越稳定。**

PSI范围	稳定性	建议事项
0～0.1	好	没有变化或者很少变化
0.1～0.25	略不稳定	有变化，继续监控后续变化
大于0.25	不稳定	发生大变化，进行特征项分析

相对熵与PSI之间的关系

接下来，我们从数学上来分析相对熵和PSI之间的关系。

$$
\begin{aligned}
psi &= \sum_{i=1}^{n}(A_i - E_i) * \ln(A_i/E_i) \\
&= \sum_{i=1}^{n} A_i * \ln(A_i/E_i) + \sum_{i=1}^{n} E_i * \ln(E_i/A_i)
\end{aligned}
$$

将PSI计算公式变形后可以分解为2项，其中：

> 第1项：实际分布（A）与预期分布（E）之间的KL散度—— $KL(A||E)$
>
> 第2项：预期分布（E）与实际分布（A）之间的KL散度—— $KL(E||A)$

因此，**PSI本质上是实际分布（A）与预期分布（E）的KL散度的一个对称化操作**。其**双向**计算相对熵，并把两部分相对熵相加，从而更为全面地描述两个分布的差异。

CSI

评分卡中从WOE分箱到区间赋分

已知：

$Odds = \frac{p(Y=Bad|X)}{p(Y=\text{Good}|X)} = \frac{p}{1-p}$

$A > 0, B > 0$

then:

$$
\begin{aligned}
\text{credit\_score} &= A - B * \ln(Odds) \\
&= A - B * [\beta_0 + \beta_1 * WOE(x)] \\
&= \underbrace{(A - B * \beta_0)}_{\text{base\_score}} + \underbrace{(-B * \beta_1 * \begin{cases} woe_1, x \in bin_1 \\ woe_2, x \in bin_2 \\ \cdots \\ woe_m, x \in bin_m \end{cases})}_{\text{partial\_score}}
\end{aligned}
$$

特征稳定性指标（CSI）计算方法

$$CSI = \sum_{i=1}^{n} (\text{Distr\_}A_i - \text{Distr\_}E_i) * \text{partial\_score}_i$$

> **含义为：**CSI = SUM((每个分箱内实际占比 - 每个分箱内预期占比） * 分箱分值)

现以实际数据为例展示上述公式。如下图所示，最终CSI的计算结果为0.36，我们可以得到哪些信息呢?

> 1. **符号为正**：表示当前样本相对于开发样本往高分段偏移。反之，说明往低分段偏移。
> 2. **绝对值大小**：表示该特征维度的稳定性，数值越大，特征稳定性越差。

#1	#2	#3	#4	#5	#6	#7	#8	#9
Characteristic	Attribute	Develop Sample		Current Sample		Delta Ratio %	Partial Score	Shift Score
		cnt	ratio %	cnt	ratio %	#6 - #4		#7 * #8
var	a1	5298	24.4%	4265	21.1%	-3.3%	17	-0.56
var	a2	5308	24.5%	4853	24.0%	-0.5%	19	-0.09
var	a3	3410	15.7%	3287	16.3%	0.5%	26	0.14
var	a4	3665	16.9%	4272	21.1%	4.2%	30	1.27
var	a5	4000	18.4%	3522	17.4%	-1.07	[illegible]	[illegible]
var	Total	21681	100.0%	20199	100.0%	0.0%		0.36

PSI 和 CSI 的比较

1. PSI 是一个广泛应用的变量稳定性指标，可用来计算连续性、离散性变量。但其**无法反映很多细节原因，比如分布是右偏还是左偏，从而引起psi过大**。(即 PSI >=0, 而 CSI 有正负)
2. CSI 目前是出现在评分卡中，主要是为了衡量分数往高分偏移还是低分偏移，这个是 PSI 无法体现的。
3. 实际模型监控中，优先参考 PSI 看宏观。当不稳定性时，再参考 CSI 看细节。