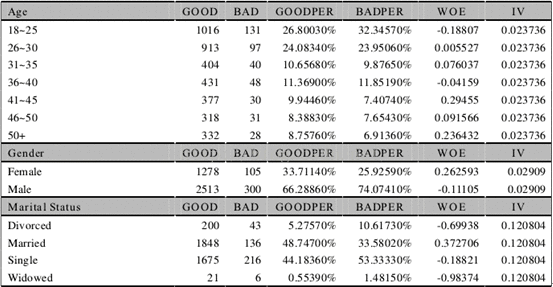
# 评分卡的相关指标含义：

|  |  |  |  |
| --- | --- | --- | --- |
| 公式 | http://img.blog.csdn.net/20150720231601100?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  woe=ln(odds)，beita为回归系数，altha为截距，n为变量个数，offset为偏移量（视风险偏好而定），比例因子factor。  http://img.blog.csdn.net/20150720231615349?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  总评分。或去掉负号。    Logistic Regression with Weight of Evidence    比例因子和偏移量为：  令好坏比为50，对应的评分为600；在些基础上评分值增加20分，**e.g.**   |  |  | | --- | --- | | 600 = log(50) \* factor + offset  620 = log(100) \* factor + offset | Factor = 20 / log(2)  Offset = 600 – factor \* log(50) | |
| WoE | Weight of Evidence 证据权重  http://img.blog.csdn.net/20150720231632642?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  其中，http://img.blog.csdn.net/20150720231644655?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center，odds为好坏比，即WOE=ln(odds)  e.g.  http://img.blog.csdn.net/20150720231659501?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  ID是自变量，有A1/A2/A3三个取值；因变量有1和0两种取值，相对应的是样本数量（如 Y=1|ID=A1 的样本数是10）。当自变量取值Ai时，编码为相应的WOEi。    Odds = 1, WoE = 0; Odds > 1, Pgood > Pbad, WoE > 0.  WoE值越高，说明这个分组的风险机率越低。（前提：good attribute 作分子。WoE可少于零） |
| IV | Information Value 信息价值  http://img.blog.csdn.net/20150720231718309?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  相当于自变量WOE值的一个加权求和，其值的大小决定了自变量对于目标变量的影响程度（与信息熵的公式稍似）。  一般应大于0.02，默认选IV大于0.1的变量进模型，但具体要结合实际。如果IV大于0.5，就是过预测（over-predicting）变量。   |  |  | | --- | --- | | **Information Value** | **Predictive Power** | | < 0.02 | useless for prediction | | 0.02 to 0.1 | Weak predictor | | 0.1 to 0.3 | Medium predictor | | 0.3 to 0.5 | Strong predictor | | >0.5 | too good to be true | |
| ROC/AUC | ROC曲线、受试者工作特征曲线、感受性曲线（receiver operating characteristic curve），用于二分问题。  判别：（1）ROC曲线越靠近左上角,试验的准确性就越高；（2）或算出线下面积（AUC, Area Under Curve）比较，AUC最大，则最佳。  AUC在 0.5～0.7时有较低准确性， 0.7～0.8时有一定准确性, 0.8~0.9则高，AUC在0.9以上时有非常高准确性。AUC=0.5时，说明诊断方法完全不起作用，无诊断价值。AUC<0.5不符合真实情况，在实际中极少出现。 |
| GINI | 见“CART”-“GINI指数” |
| 分组 | Segmentation 模型分组  将总体划分为几个子总体，然后对不同的子总体分别建立评分卡。视需要而是否使用。 |
| binning | Binning means the process of transforming a **numeric** characteristic into a **categorical** one as well as re-grouping and consolidating categorical characteristics.  等距、等频、按需、优化离散(常用的检验指标有卡方，信息增益，基尼指数，或WOE) |
| Cutoff | Cutoff临界分值  若申请人的信用得分在该分数上，则评分系统建议批准申请；反之。  选择方法：（1）申请批准比率与模型使用前的比率一致；（2）坏帐率与模型使用前的比率一致；  http://img.blog.csdn.net/20150720231746828?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center |
| 人工修正 | HSOs高端人工修正：超过临界值，但仍然不通过申请；  LSOs低端人工修正：反之。 |
| 属性 | Attribute 如维度“性别”有两个属性：“男”、“女” |
| 权重 | Score weight如未婚15分、已婚30分 |
| 拒绝推断 | Reject Inference  一般来说，信用评分的训练样本均来自于以前申请贷款被接受的那些人，而评分模型则是要应用到所有申请的人，因此如直接应用模型到被拒绝者，很有可能会出现问题。  Reject Inference对模型修正来解决这个问题。 |
| 模型验证 | Validation  留下一部分样本做最后的验证。 |
| PSI | 稳定度分析指标  PSI衡量两组样本评分是否有显著差异。  公式：http://img.blog.csdn.net/20150720231759742?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  判断：index <= 0.1，无差异；0.1< index <= 0.25，需进一步判断；0.25 <= index，有显著位移，模型需调整。 |
| LTV | LTV 贷放成数 Loan to Value  贷款额与抵押价值的比率，用于额度管理，如：   |  |  | | --- | --- | | 评分分级 | 最高贷放成数LTV | | A | 85% | | B | 75% |   个人信用评分>=80，且LTV<=85，接受贷款；否则反之。 |

例子



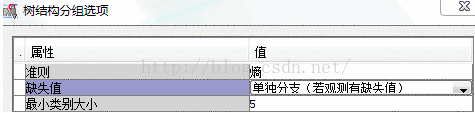
# 分箱的方法

监督离散化（supervised discretizaion），使用递归划分（Recursive Partitioning）将连续变量分为分段，背后是一种基于条件推断查找最佳分组的算法（Conditional Inference Tree）还有mdlp分箱评分算法评分函数，K2,binning,mdlp等

1、监督离散化（supervised discretization）

考虑类别信息（已知X 的值和Y 的值）。检验方法如：卡方检验（ChiMerge 慢、Chi-square、Chi2、CAIM、CACC、ameva），信息增益，基尼指数，最短描述长度原则（MDLP，基于熵），WoE等。

（1）最优准则：基于“树结构准则”查找最佳分组 （条件推理树Conditional Inference Trees, initially excludes missing values(NA) to compute the cutpoints, adding them back later in the process for thecalculation of the Information Value）。



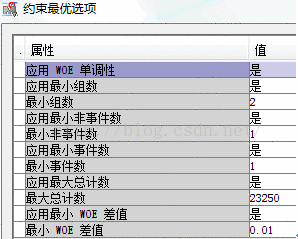
准则：

（a）熵；

（b）目标和分支节点使用Pearson卡方统计量的p值；

（2）单调事件率MonotonicEvent Rate：要求各组的单调事件率呈单调。

（3）约束最优ConstrainedOptimal Binning：基于预定义的约束创建分组。



2、无监督离散化（unsupervised discretization）

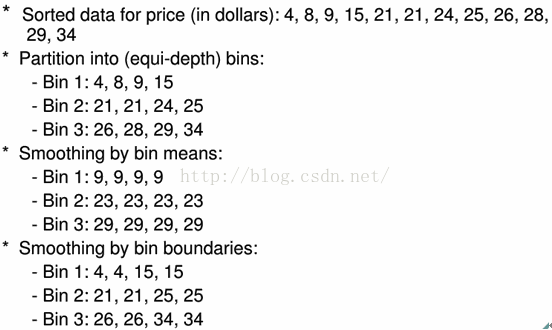
不考虑类别信息（已知X的值而未知Y的值）。分箱（binning/split bin）方法如：等宽、等频、聚类。

（1）等宽/等距：将数值属性的值域[ , ]等分为K个区间，即 (Xmax - X min) / K。前提：数值大概服从均匀分布。

（2）等频/等深/分位数（equi-depth）：和等宽类似，但它不是要求每个区间宽度一样，而是要求落在每个区间的对象数目相等。即共有M个点，每个区域含有M / K 个点。

分箱后再作数据平滑处理（smoothing）。

e.g.



3、人工定区间，按照需要而定。

Structure:

