一般的数据建模步骤为SEMMA(Sample，Explore，Modify，Model，Access)。

#### 1. 数据取样（Sample）

Sample步骤即为建模数据取样，数据宽表中的观测并非全部用于LOGISTIC建模输入，一般把数据集划分为建模数据集（Training Data）和验证数据集（Validation Data）。默认采用0.7/0.3的比例拆分原数据宽表，基于目标变量分层取样。

#### 2. 数据探索（Explore）

Explore即为数据探索，主要检查各变量字段的数值缺失，统计量检查及变量预筛选（初始变量数过多时如果不做筛选直接进入模型构建会大大延缓模型的收敛速度）。常用的变量筛选方法有单变量R-Square和IV（Information Value）值等方法，评分卡大多采用IV值筛选法，保留IV值在0.05以上的变量。

WOE（Weight of Evidence）和IV：

证据权重（WOE）是计算IV值的基础，用于衡量变量某个属性的风险，WOE越高表示该属性分组中观测是正例（目标值bad）的可能性越低，可以直观的认为WOE和目标变量的概率之间存在着某种单调的影响关系。其计算公式为：



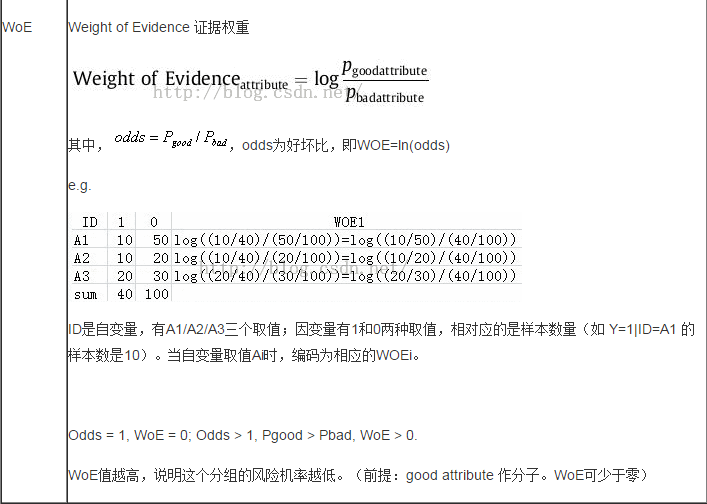
pct\_good为该属性分组中负例数（非再购）在总体负例的占比，pct\_bad为该属性分组中正例数（再购车原车）在总体正例的占比。

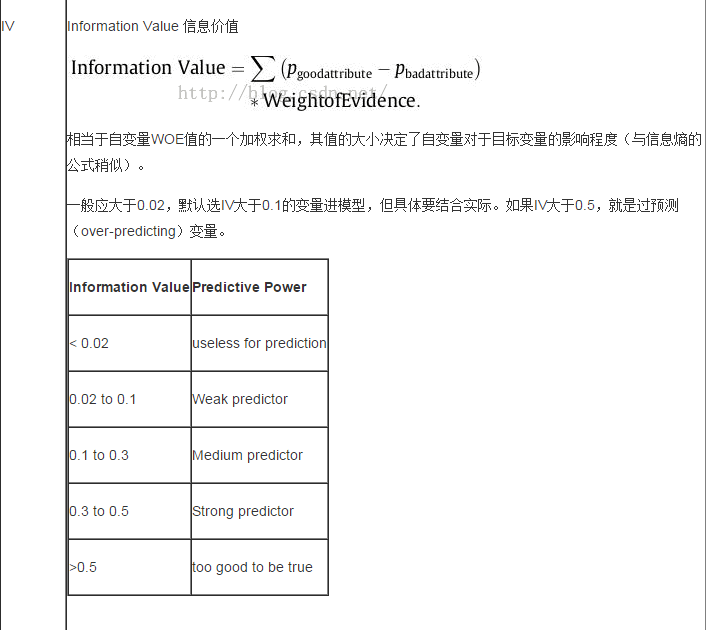
IV（信息价值）衡量了一个自变量对目标变量的区分程度，它是WOE的一个加权平均值，计算公式如下：



IV值越大表示目标正负例在该变量上的分布差异越大，该变量对目标变量的区分能力越好。

IV值计算一般应用于离散变量，对于区间型变量（连续型），计算这类变量的IV值需要先对变量进行分段，一般可按照变量取值从低到高等观测数划分20段（该值可适当调整），计算各段的WOE值和加总的IV值，剔除IV值低于0.05（该值可适当调高）的变量。经IV值筛选后剩余变量数尽量不要超过100。





#### 3.数据调整（Modify）

Modify步骤主要是对自变量字段进行必要的函数变换，统一自变量之间的测度，改善自变量的数值分布等。一般的数值变换采用WOE值对原变量的各分段值进行替换，替换后的变量值进入LOGISTIC建模。该方法类似于传统的连续变量分段打分替换（E.G. 等分100段，从高到低赋值100 - 1）。

进行WOE值替换前先进行分段调整，原20等分段基于取值区间合并后剩余的分段数仍有可能过多，分段数过多容易造成模型的解释效果不明显，影响业务理解的直观性。因而需要对变量分段做进一步的调整，合并剩余分段，保证各变量的最大分段数不超过10个。然后再次计算合并分段的WOE值，替换变量原区间段取值。

#### 4. 建模（Model）

Model步骤输入训练集（Training Data）中经过预筛选的变量，进行模型变量挑选和LOGISTIC回归参数估计。LOGISTIC回归的变量挑选常用的可分为前进法、后退法和逐步回归法。

前进法的主要思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。具体做法是首先对一个因变量y和k个自变量分别建立回归方程，并分别计算这k个回归方程的F值，选其最大者进行显著性检验，有意义则挑选该变量进入模型。再此基础上计算其他自变量的偏回归平方和，选取偏回归平方和最大者作F检验，显著则进入模型。重复上述步骤直到没有变量引入为止。偏回归平方和F值表示如下，服从F(1,n-k-1)分布：



后退法，是先将k个自变量全部放入回归方程，然后逐步剔除。首先挑选偏回归平方和最小的变量，作F检验求其相应的p-value，检查其是否显著（p-value小），如果p-value值过大则删除该变量。剔除变量后重新建立回归方程，重复上述过程直到没有新的变量再被剔除为止。

逐步回归法，前进法存在着这样的缺点当一个变量被引入方程时，这个变量就被保留在这个方程中了，当引入的变量导致其不显著时，它也不会被删除掉，后退法同样存在着这样的缺点，当一个变量被剔除时就永远的被排斥在方程以外了，而逐步回归法克除了两者的缺点。逐步回归的思想是有进有出。将变量一个一个的引入，每引入一个变量对后面的变量进行逐个检验，当变量由于后面变量的引入而不变的不显著时将其剔除，进行每一步都要进行显著性的检验，以保证每一个变量都是显著的。

逐步回归法对变量的挑选更加细致，但其运算量比前进法和后退法高，相应的模型收敛速度也较慢。一般选用逐步回归法作为LOGISTIC建模方法。

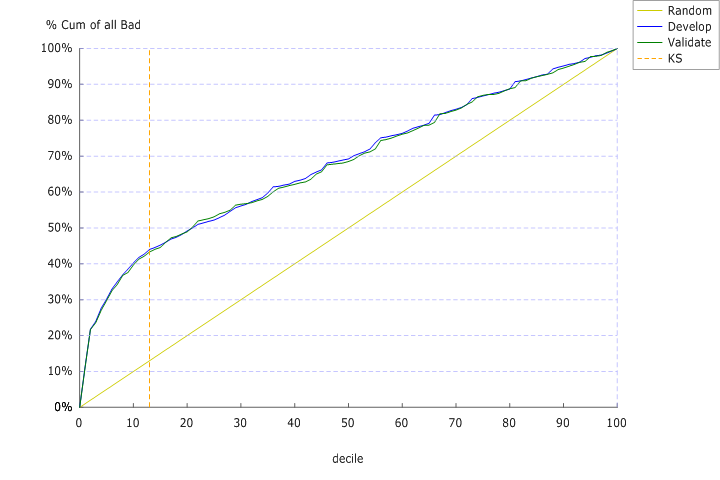
#### 5. 模型评估（Assess）

分类模型性能评估主要考虑其分类的准确性，特别是对‘正例’观测的识别率，常用的指标有ROC，GINI，AUC，KS等。

KS（Kolmogorov-Smirnov）值：在K-S检验中，先计算观察数据被比较的两套累积分布函数，然后求这两个累积分布函数的差的绝对值，其最大值即为K-S。一般认为KS值大于0.2的模型可用，0.3-0.5模型效果较好，0.5-0.7模型效果优秀，0.7以上模型可能存在过拟合。

AUC（Area Under Curve）值：模型曲线（一般是ROC曲线）与坐标轴之间围成的区域面积，其取值介于0和1之间。越接近1说明模型的效果越好。

GINI系数：模型曲线与45度线之间的面积/完美模型曲线与45度线之间的面积，其取值介于0和1之间，越接近1表示模型的预测能力越高。GINI系数和AUC之间存在如下的函数关系：AUC=GINI\*0.5+0.5。



如上所示的Lorenz图，横轴是样本群体通过逻辑回归参数计算出来的概率值，从高到低进行分段后的分位点值，纵轴为分段的累积正例覆盖率。45度线是基线（Base Line），代表随机筛选的情况下得出的正例覆盖率。Lorenz曲线与Baseline的最大纵轴截距差即为KS值。

对于模型的评估除了需检查是否适用于建模的样本外，还要考虑其应用的普遍性，所以还需要对其他不同批次的抽样样本应用模型，考察模型的稳定性，之后才能将其应用到实际中。一般采用验证数据集（Validation Data）对模型进行检验，并绘制其KS曲线与训练数据集的效果进行比较，两者越贴和表示模型的稳定性越好。

#### 评分（Score）

评分主要是基于估计的模型参数构建打分公式，保证形成的分值和模型估计概率值之间为单调的关联关系。

打分公式示例：

