## 逻辑回归模型开发流程（通用）

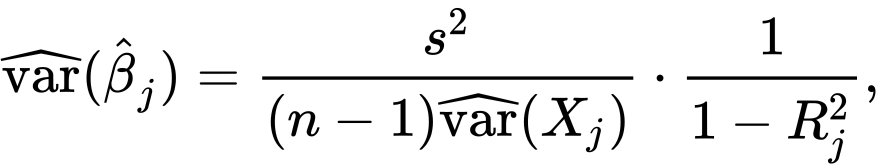
1. **数据准备-I：**时间维度数据合并 + 变量派生
2. 观察窗口历史数据合并（过去3/6/9/12/24 个月）
3. 派生时间维度衍生变量（总额，频数，增减趋势）
4. 派生业务维度衍生变量（消费偏好，可以交易，100% Uti within X days…）
5. **描述性分析**：单变量统计分布 + 识别缺失值和极端值
6. 统计分布：SAS proc univariat 可生成单变量统计指标, 包括矩和位置等

<https://www.stattutorials.com/SAS/TUTORIAL-PROC-UNIVARIATE.htm>

1. 缺失值处理：修复、替换、保留、删除。将缺失值作为一种数据特征保留，因为最终所有连续变量都要离散化，可以把缺失值作为离散的一类。(附录：缺失值处理方法)
2. 极端值处理：通过K-means聚类分析识别极端值，若极端值比例超过10%，则需将总体分为两个独立的数据集，即分群（segments），针对每个群开发单独的评分卡; 若小于10%，则在逻辑回归时需要考虑将极端数据剔除。

<https://towardsdatascience.com/k-means-clustering-in-sas-9d19efd4fb1b>

1. **特征分析：**定位预测力最强的变量，主要判断指标：
2. IV 信息价值 (针对逻辑回归)
3. IG 信息增益 (针对决策树)
4. R方 决定系数 (针对所有回归模型)
5. **数据准备-II:** 分箱，连续变量离散化 + 分类变量降基数
6. 粗分类 Coarse Classing:
   * + - 1. 数值变量：等高度分箱（10-20箱）
         2. 分类变量：保持其原有分类不变
         3. 计算每个bin的WOE
7. 精分类 Fine Classing: 将WOE相近的箱体合并起来，减少箱体数量，降低基数
8. **变量选择-I:** 选择预测力最强的变量，并去除多重共线性
9. 变量初选：预测力最强(IV, R^2)，且有业务意义的变量组合
10. 去除多重共线性：
11. 多重共线性诊断： VIF 方差膨胀因子>10 或 Tolerance < 0.1, 则说明该变量存在严重的多重共线性
12. 多重共线性消除: 制作两两变量间的皮尔森相关系数矩阵，定位相关变量组合，重新选择变量
13. 降低模型敏感度：通过引入正则项 Regularization
    1. L1正则：可以起到稀疏特征的筛选作用 (LASSO Regression)
    2. L2正则：无法筛选稀疏特征，但稳定性更强 (Ridge Regression)
    3. 同时运用L1+L2正则：结合两者优点 (弹性网络模型 Elastic Net)
14. **数据的归一化** Normalization：将数据映射到指定的范围，用于去除不同维度数据的量纲以及量纲单位
    1. Max-min 极差归一
    2. Z-score 标准差归一
    3. WOE 证据权重归一(针对逻辑回归)：用WOE值替代原数据集所有变量原有的值
15. **建立逻辑回归模型**：SAS proc logistic 或SAS EM logistic, 使用 stepwise 法再次选择变量
16. **判断模型是否成立**：四个判断指标：
17. 最大似然估计（判断单个变量的参数是否可用）: 单个变量的标准误越小越好，说明单个参数的估计置信区间不至于过大，是可靠可用的。判断标准是 Wald Chi-sqr P Value < 0.05；沃尔德方差公式：



1. 过拟合验：AIC SC 越小越好。本模型AIC为2700；两个模型应该选AIC小的那个；AIC大则有过拟合可能(变量太多)；

<https://en.wikipedia.org/wiki/Akaike_information_criterion>

1. 拟合不足检验：Hosmer Lemeshow Test; Chi P Value < 0.05, 否则说明欠拟合

<https://en.wikipedia.org/wiki/Hosmer%E2%80%93Lemeshow_test>

1. 全局零假设检验：假设所有参数都是0，相关关系是否为抽样误差偶然得到，标准为 Likelihood Ratio, Score, Wald的P Value < 0.05

<https://stats.idre.ucla.edu/sas/output/proc-logistic/>

1. **模型验证 Validation**: 检查模型的精确性和稳健性
2. 精确性 Accuracy: 模型对目标变量的预测是准确的，能够有力区分 Good & Bad客户。
3. **混淆矩阵** Confusion Matrix: 通过实际结果 \* 预测结果的矩阵，得到True Positive, False Positive, True Negative, False Negative 四组基本测试数据(何为预测positive取决于对预测概率的cutoff point选择)，并衍生出一系列衡量模型准确性的指标，其中最重要的3个指标：
   * + - 1. 测准率 Accuracy: (TP+TN)/Total;
         2. 命中率 Hit Rate: TP/(TP+FN); 也叫TPR, Sensitivity;
         3. 误报率 False Alarm： FP/(FP+TN); 也叫FPR;

<https://en.wikipedia.org/wiki/Confusion_matrix>

1. **ROC Curve:** 基于混淆矩阵, y轴为Sensitivity (hit rate), x轴为false alarm (1-specificity), 检查随着cutoff point的移动，两个准确率指标的增加变化。在任何cutoff point下，命中率比误报率越高越好。AUC/C值为ROC曲线下得面积，越大越好，最大值为1，Benchmark为0.25。

<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>

1. **KS Curve**: 衡量分离度，是最重要的验证指标；X轴为total population cumulative %, Y轴为good cumulative % 和 bad cumulative %两条曲线；两条曲线差异越大越好。Benchmark为30%。

<https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test>

1. **Lorenze Curve & Gini Coefficient**: 衡量数据纯度（分布均匀程度），x轴为total population cumulative %, Y轴为good cumulative %。两条曲线：1.均匀分布；2.实际分布；Gini为两条曲线下面积比A/(A+B), 当分布绝对不均匀时, Gini=1;

<https://en.wikipedia.org/wiki/Gini_coefficient>

1. **Lift Chart:** 检查responder rate提升倍数；Model selection VS random selection;

<https://docs.microsoft.com/en-us/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=asallproducts-allversions>

1. 稳健性 Stability: 确保模型在其他样本也适用，排除过拟合风险。
2. 交叉验证 Hold-out Validation： 训练数据集和验证数据集交叉检验KS值, Gini系数。
3. 超时验证 Out-of-time Validation：在建模窗口过去后，检查每个月的KS值,Gini系数。在超过2个月后，使用PSI检验模型是否随着时间发展有了重大变化。
4. 样本稳定性：Population shift index (PSI)
5. **评分刻度**:
6. 场景1.预测概率与Score反相关，如M2+%, 分数越高违约概率越低;
7. 基本公式： Score = A – B \* Ln(Odds)
8. 基本假设：
   * + - 1. 刻度 ScoreStep = 20; 每20分odds翻倍;
         2. 基点比率 Odds\_0 = 1/60;
         3. 基点得分 Score\_0 = 600；当Odds\_0时的Score;
9. 参数计算:
   * + - 1. Score\_0 = A – B \* Ln(Odds\_0)
         2. Score\_0 + ScoreStep = A + B \* Ln(2Odds\_0)
         3. B = ScoreStep/Ln(2)
         4. A = Score0 + B \* Ln(Odds\_0)
10. 分值分配：
    * + - 1. Score = A - B \* Ln(Odds)

= A - B \* (b0 + b1 \* WOE1 + b2 \* WOE2…)

= (A - B\*b0) - B \* b1 \* WOE1 - B \* b2 \* WOE2…

* + - * 1. Base Score = A - B \* b0
        2. Score of variable xn: -(B \* bn \* WOEn)

1. 场景2.预测概率与Score正相关, 如repay%, 分数越高越有可能还款;
2. 基本公式：Score = A + B \* Ln(Odds)
3. 基本假设（同上）
4. 得分计算（推导过程同上）
   * + - 1. B = ScoreStep/Ln(2)
         2. A = Score\_0 – B \* Ln(Odds\_0)
5. 参数分配
   * + - 1. Score = A + B \* Ln(Odds)

= A + B \* (b0 + b1 \* WOE1 + b2 \* WOE2…)

= (A + B\*b0) + B \* b1 \* WOE1 + B \* b2 \* WOE2…

* + - * 1. Base Score = A + B \* b0
        2. Score of variable xn: B \* bn \* WOE

## 附录

### 缺失值如何处理？

step1:调查缺失值的原因（1.系统问题?---最好替换补充； 2.无定义值?---最好设为"其他"）

step2:缺失值的处理：

1.替代法：用0，均值，中位数，众数（类别变量）替换；--- 当缺失值较少，影响小时 0.1-0.2%；优点：简单易施行；缺点：人为引入噪声

2.补齐法：用与之强相关的变量建立线性回归预测模型，补上预测值；--- 当缺失值较多，影响不容忽视；优点：能够尽可能准确地降低人为噪声；缺点：变量之间强相关，导致多重共线性，增大参数估计方差，致使模型不准

3.保留法：对于类别型变量，将其保留为[其他]类；对于数值型变量，将其按照类别变量处理，（连续变量视为离散变量）引入上千万甚至上亿维度；

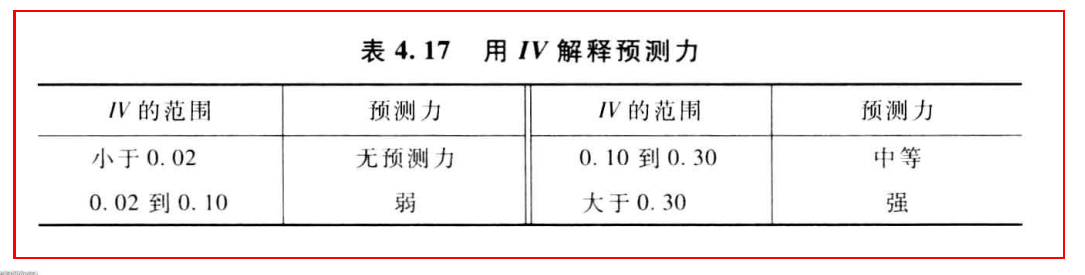
4.删除法：将有缺失值的数据删除掉，从样本中排除;

### IV、WOE的计算和范围

IV的计算：先对变量分箱，数值型变量离散化，分类型变量降基数（降维度），计算每一箱的WOE, 从而计算IV;

WOE = log(good\_pct / bad\_pct);

IV = Sum(WOE\*(good\_pct-bad\_pct));



### R方的计算

使用SAS proc corr 计算相关系数，再手工乘方计算决定系数；

### IG的计算

IG = 父节点信息熵 – Sum(子节点信息熵 \* 子节点数据量权重)；

信息熵 Entropy衡量数据的纯度/分布不均匀程度，Entropy越大，则数据纯度越低、分布越均匀。当不同类别数据完全均匀分布时，数据最不纯，Entropy达到最大值1。当所有数据全部集中在一个类别时，Entropy达到最小值0。决策树对数据的分类实际上是一个熵减过程，减少的熵也就是信息增益，代表信息分布由绝对均匀/混乱，向有序分布/秩序的发展。