## 决策树模型开发流程（通用）

* 1. **数据准备：**时间维度数据合并 + 变量派生
     1. 观察窗口历史数据合并
     2. 派生两类衍生变量：
        1. 派生时间维度衍生变量（总额，频数，增减趋势）
        2. 派生业务维度衍生变量（消费偏好，疑似欺诈，100% Uti within X days…）

[最终原始数据集包含变量600+，400+衍生，100+原始]

* 1. **描述性分析：**单变量统计分布 + 识别缺失值和极端值
     1. 统计分布：生成Desc表，包含矩和位置等基本统计分布指标，具体：OBS, Mean, STD, Min P25 P50 P75 Max Missing%;
     2. 缺失值处理：若缺失值占比超过10%，需要进行处理。处理方式有：修复、替换、保留、删除。决策树对缺失值不敏感，可直接保留缺失值。
     3. 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模型; 若小于10%，则在运行模型前需要考虑将极端数据剔除。
  2. **特征分析：**定位预测力最强的变量，三大判断指标：
     1. IG 信息增益 （算法见附录，可基于Entropy, Gini, Chi-square…）
     2. IV 信息价值
     3. R2 决定系数
  3. **变量选择:** 选择预测力最强的变量，并去除多重共线性
     1. 变量初选：预测力最强(IG, IV, R2)，且有业务意义的变量组合（10-20个）
     2. 去除多重共线性：
        + 1. 多重共线性诊断：

VIF 方差膨胀因子>10 或 Tolerance < 0.1, 则说明该变量存在多重共线性；

Eigen Value 特征向量 和 Codition Index 差距极大，则说明存在多重共线性；

* + - * 1. 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
  1. **建立决策树模型：**数据导入SAS EM，建立决策树，SAS的决策树算法是SAS Tree Algorithms)，包含并验证了ID3,C4.5,CART,CHAID等经典决策树算法。可通过调整参数设置来模拟各种经典算法。
     1. 算法基础：所有决策树均是基于贪心算法，即每一步选取最优解，从根节点开始递归，触发终止条件结束。
     2. 两个关键：决策树算法的两个关键是：a.节点分裂条件; b.分裂终止条件;
     3. 主流决策树算法：
        + 1. Hunt算法: 最基础的算法，一切决策树算法的起源；本质是贪心算法，每一步选取最优解，从根节点递归到所有数据均归于同一个类，无法再细分则终止。
          2. CHAID算法，即Chi-square Automatic Interaction Detector，卡方自动交互检测算法。是很常用的算法。主要特征是使用Chi-square作为节点分裂条件。
          3. ID3 算法，即Interactive Dichotomiser 3，交互二分树第三代；特点是用Entropy信息熵、IG信息增益作为节点分裂条件。
          4. C4.5算法：是对ID3决策树改进，不再使用IG信息增益作为分裂依据，而改为使用IG Ratio信息增益率。原因是当节点分裂为很多个类，而个别类数据量特别少时，纯度极高，Entropy极低，造成信息增益被夸大虚高。使用信息增益率IG Ratio可以避免这个问题。
          5. CART算法，即Classification And Regression Tree, 分类回归树。主要特征是使用Gini系数作为节点分裂依据。
  2. **模型验证 Validation:** 检查模型的精确性和稳健性；
     1. **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。
        1. **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;
        2. **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。ROC曲线下得面积未AUC值（也叫c值）
        3. **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%。
        4. **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;
        5. **Lift Chart:** 检查responder rate提升倍数；Model selection VS random selection;
     2. **稳健性 Stability:** 确保模型在其他样本也适用，排除过拟合风险。
        1. **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。
        2. **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。在有超过2个月的验证数据积累后，使用PSI检验模型是否随着时间发展有了重大变化。

## 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
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，疑似欺诈，100% Uti within X days…）

[最终原始数据集包含变量600+，400+衍生，100+原始]

1. **描述性分析：**单变量统计分布 + 识别缺失值和极端值
2. 统计分布：生成Desc表，包含矩和位置等基本统计分布指标，具体：OBS, Mean, STD, Min P25 P50 P75 Max Missing%;
3. 缺失值处理：若缺失值占比超过10%，需要进行处理。处理方式有：修复、替换、保留、删除。本模型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)
4. 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。
5. **特征分析：**定位预测力最强的变量，两大判断指标：
6. IV 信息价值
7. R方 决定系数
8. **数据准备-II:** 分箱，连续变量离散化 + 分类变量降基数
9. 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
10. 精分类 Fine Classing:
    * + - 1. 数值变量：按照变量本身的数值升序排序，排序后WOE应该呈单增或单减(理想状况下)，将WOE相近的bin合并（差距15%以内）；
          2. 分类变量：按照WOE升序获降序排序，将WOE相近的合并（差距15%）；
11. **变量选择-I:** 选择预测力最强的变量，并去除多重共线性
12. 变量初选：预测力最强(IV, R^2)，且有业务意义的变量组合（10-20个）
13. 去除多重共线性：
    * + - 1. 多重共线性诊断：

VIF 方差膨胀因子>10 或 Tolerance < 0.1, 则说明该变量存在多重共线性；

Eigen Value 特征向量 和 Codition Index 差距极大，则说明存在多重共线性；

* + - * 1. 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

1. **WOE转换：**用WOE值替代原数据集所有变量原有的值。
2. **建立逻辑回归模型：**run proc logisttic, 使用 stepwise 法再次选择变量
3. **判断模型是否成立：**四个判断指标：
4. 最大似然估计: 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 Chisqr P Value < 0.05
5. 拟合度检验：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)；
6. 拟合不足检验：Hosmer Lemeshow Test; Chi P Value < 0.05, 否则说明欠拟合
7. 模型显著性检验：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
8. **模型验证 Validation:** 检查模型的精确性和稳健性；
9. **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。
   * + 1. **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;
       2. **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。
       3. **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%。
       4. **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;
       5. **Lift Chart:** 检查responder rate提升倍数；Model selection VS random selection;
10. **稳健性 Stability:** 确保模型在其他样本也适用，排除过拟合风险。
11. **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。
12. **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。
13. **评分卡刻度:**
14. 场景1.预测概率与Score反相关，如M2+%, 分数越高违约概率越低;
15. 基本公式： Score = A – B \* Ln(Odds)
16. 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;
17. 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)
18. 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

## 花旗银行大额存单目标客户定位模型（决策树）

1. **项目背景：**花旗银行在2017年9月获得大额存单发行销售资格。为了扩大存款规模，多元化客户资产配置，建立决策树模型，对大额存单目标客户进行精准定位识别，提高产品推介成功率。
2. **样本选取：**选取全体借记卡客户作为建模样本（19万），2017年10月作为观察月observation month，过去的12个月作为观察窗口observation window，之后的3个月作为表现窗口performance window，2019年1月为表现月。在2017年10月之后3个月内购买了大额存单的客户+购买了20万以上一年期以上定期存款的客户作为目标变量。
   1. 样本：全体借记卡客户（19万）；70%作为训练集，30%作为表现集。
   2. 观察月：2017年10月
   3. 观察窗口：12个月，2016年11月-2017年10月
   4. 表现月：2018年1月
   5. 表现窗口：3个月，2017年11月 – 2018年1月
3. **建模过程：**详见通用决策树建模流程
4. **模型结论：**选取决策树中response rate最高的5个叶节点作为目标客户群。
   1. 预测变量5个：
      1. 过去3个月定期存款金额
      2. 未来1个月产品到期金额
      3. AUM规模
      4. 定存、债券、储蓄类保险占AUM比例
      5. 过去半年手机银行登录次数
   2. 客户画像特点3个：
      1. 50岁以上女性。
      2. 偏好保守型产品。定存、债券、储蓄类保险。
      3. 有钱，且有闲钱。AUM高，未来一个月有产品到期。
5. 模型表现：
   1. 稳定性验证：
      1. Hold-out Validation: 训练与验证数据集KS均达到43，Gini系数达到81。
      2. Out of Time Validation: 表现期结束后，每月监测KS,Gini趋于稳定。
      3. PSI Population Stability Index: PSI值0.1, 模型随时间变化不大
   2. 产品转化率：模型于2018年3月投入运行，每月选取1000位客户交一线电销跟进，客户转化率在19%左右。过往存款产品转化率往往不到10%。

## 花旗银行信用卡-保险/基金交叉销售模型（决策树）

1. 项目背景：
   1. 业绩下滑：受到17年监管政策收严影响，花旗银行主打银保产品-具有理财属性的短期两全险遭到停售，只允许销售长期险种，造成保险中收业绩骤然下滑，由月3000万平台降至800万平台。作为财富管理营收三驾马车之一（基金，保险，存款），保险中收缺口亟待弥补。
   2. 上升瓶颈：花旗银行存量贵宾客户(Citi Gold)保险渗透率高达47%，基金渗透率高达56%，资产配置程度较高，存量客户资源开发已趋饱和，需要寻找新的增长点：1)新贵宾客户-量极小；2)信用卡客户-量大；信用卡交叉销售是重要的新增长点。
2. 样本选取：选取全体信用卡客户作为建模样本（46万），2017年12月观察月，之前12个月为观察窗口，之后的3个月为表现窗口，未来3个月刷卡购买我行或他行保险的客户作为目标变量。
   1. 样本：选取全体信用卡客户作为样本（46万）
   2. 观察月：2017年12月
   3. 观察窗口：过去12个月，即2017.1-2017.12
   4. 表现月：2018年3月
   5. 表现窗口：之后3个月，即2018.1-2018.3
   6. 目标变量：未来3个月通过信用卡刷卡购买了我行或他行保险的客户作为目标变量
3. 建模过程: 详见通用决策树建模流程
4. 模型结论：选取决策树中response rate最高的7个叶节点作为目标客群。
   1. 预测变量5个：
      1. 月收入1万以上
      2. 高刷卡消费(ANR)客户
      3. 高海外消费客户
      4. 有附属卡 – 可能有子女，有家庭保险需求
      5. 有车 – 可能有旅游保险需求
   2. 客户画像特征6个：
      1. 基本：35-40男性
      2. 家庭：已婚、育有子女、有车
      3. 阶层：处在事业发展期、财富积累阶段的中产阶级
      4. 行为模式：高消费欲望、高消费能力，喜爱出国旅游购物；海外刷卡消费金额是普通客户的10倍以上
      5. 产品偏好：存量已购保险客户更偏好性价比较高的重疾、健康险
      6. 从事行业：RG为C，行业类型为不稳定行业职业的客户反而更乐于购买保险。原因是保险销售本质上是贩卖焦虑，对未来不确定性越大，焦虑越严重的人群越容易购买保险。
5. 模型表现：
   1. 稳定性验证：
      1. Hold-out Validation: 训练与验证数据集KS分别达到49和51，Gini系数达到82。
      2. Out of Time Validation: 表现期结束后，每月监测KS,Gini趋于稳定。
      3. PSI Population Stability Index: PSI值0.1, 模型随时间变化不大
   2. 产品转化率：模型于2018年5月投入运行，每月形成目标客户清单交一线营销跟进，客户转化率在19%左右。

## 汇丰银行催收评分卡模型（逻辑回归）

1. **项目背景:** 汇丰银行信用卡业务自2016年启动以来，逾期迁徙率逐年走高，高于行业平均水平，呆账控制能力不足。催收能力低下的原因之一是催收策略不当，原有催收策略以客户逾期事件长短为主要决策依据，拖欠时间越长，催收力度越大，导致：
   1. **催收不及时：**对于高风险的早期逾期客户催收力度不足，行动滞后，错过催收最佳时间窗口。
   2. **催收使错力：**对于长期逾期的死硬客户催收过度，投入了大量催收资源在难度最大、最不容易还款的客户群体上，相当于从铁公鸡上拔毛，导致整体催收效果不佳。

为解决上述问题，建立催收评分模型，量化客户还款概率，定位最有可能/最不可能还款的客户群体；从以逾期时间长短为决策依据，转变为以客户还款倾向、风险程度为决策依据。

1. **样本选择、目标变量设置、模型选择：**选取全体逾期信用卡客户为样本（1.5万，3.3亿RMB）,2018年12月为cutoff point, 此前12个月为观察窗口，此后3个月为表现窗口，未来3个月至少还了最低还款额的客户为目标变量，建立逻辑回归模型.
2. **建模过程：详见逻辑回归通用建模流程**
3. 模型结论:
   1. 预测变量5个：
      1. 逾期期数
      2. 逾期金额
      3. 过去6个月Uti rate
      4. 过去6个月还款率
      5. 是否曾失联
   2. 高风险客户画像：
      1. 40-50岁男性
      2. 大专学历以下
      3. 东部地区
      4. 自雇
      5. 制造业
4. 模型表现:
5. Asdfasdf
6. Asdfasdfasdf

## 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

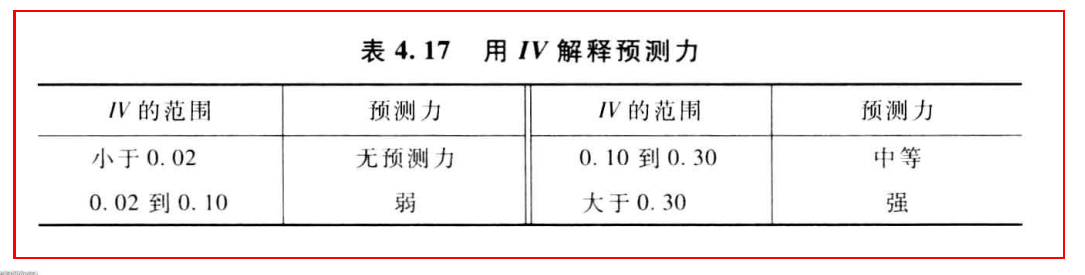
3.保留法：对于类别型变量，将其保留为[其他]类；对于数值型变量，将其按照类别变量处理，（连续变量视为离散变量）引入上千万甚至上亿维度；

### 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。