扬州分析案例分享

分析组：华德禹 周志昊

目录

CONTENT

1 客户需求

背景、现状、目标

2 客户数据

数据源、数据结构

3 方案设计

原始方案、创新方案、方案对比

4 思维拓展

嵌入机器学习平台

1

客户需求

- 背景
- 现状
- 目标

■ 江苏省创新大赛

扬州公安局科技信息化处广陵分局-公寓楼住宅实战管理

■ 参赛方案建设背景

公寓式住宅租期灵活，在发展中逐渐形成了人员结构复杂、流动性强、治安问题突出等特点，传统的警务工作模式无法适应公寓楼管理的需求。针对此现状，以智能门禁系统建设为契机，创新公寓楼住宅实有人口管控手段，以便及时发现违法人员或帮扶对象。

■ 现状

1. 已经对各类警务数据综合分析得出特殊群体的特征规律及行为模式,并且据此构建完成公寓楼住户标签体系
2. 现有方案中以构建隐性涉毒人员标签为例，以基础属性、活动轨迹、特殊行为三类属性为特征构建完成隐性涉毒人员预测模型，预测住户中潜在的涉毒人员，但是预测精确度较差，不足30%

■ 目标

1. 通过使用机器学习相关算法构建预测模型，提高预测精确度

2

客户数据

- 数据源
- 数据结构

■ 数据源

1. 警务基础平台数据（人员基本信息、网吧上网信息、旅馆住宿信息等）
2. 蛛网数据（犯罪记录、犯罪人员基本信息、手机通话记录等）
3. 情报平台数据（民航数据、铁路数据等）

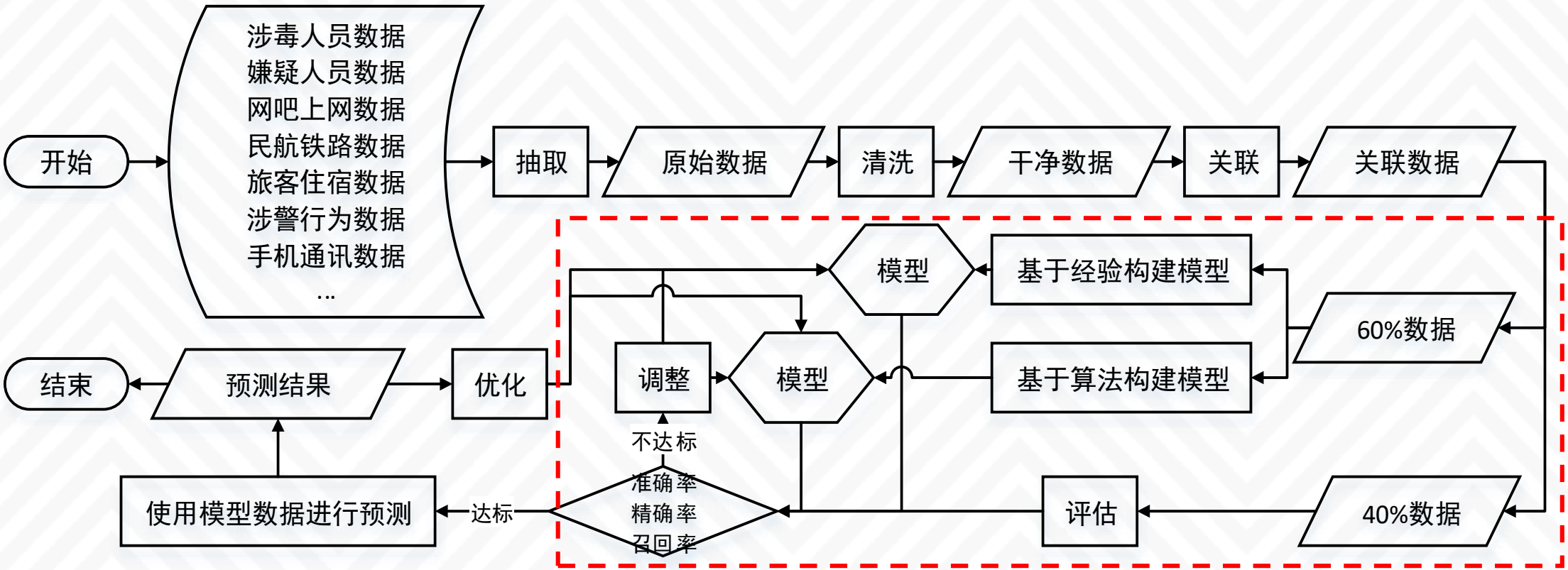
■ 数据结构

1. 自然特征：年龄、性别、民族、学历…
2. 社会特征：家庭情况、工作情况…
3. 行为特征：旅馆入住时间、入住时长、网吧上线时间、下线时间、是否有涉警行为、通话时间、通话时长…

3

方案设计

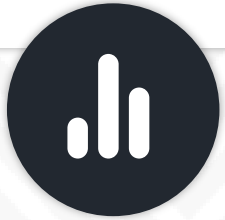
- 原始方案
- 创新方案
- 方案对比


开始
涉毒人员数据
嫌疑人员数据
网吧上网数据
民航铁路数据
旅客住宿数据
涉警行为数据
手机通讯数据
...
抽取
原始数据
清洗
干净数据
关联
关联数据
模型
基于经验构建模型
60%数据
基于算法构建模型
调整
模型
优化
结束
预测结果
不达标
准确率
精确率
召回率
达标
使用模型数据进行预测
评估
40%数据

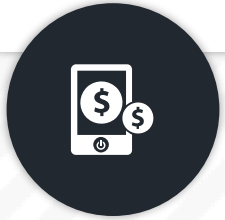
■ 原始方案

1、特征选取

3、模型评估



2、构建模型



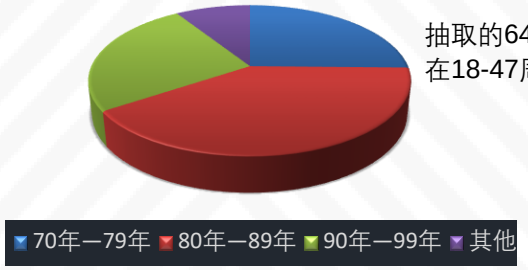
■ 特征选择

以2017年首次因涉毒被打击处理的共计644人作为隐性涉毒人员样本，研究其2016年全年的行为轨迹，再与同时间段内扬州市所有人历史轨迹作对比，分析得出以下6点差异较大的特征

1. 年龄是否在18-47周岁
2. 2016年内是否正常缴纳社保
3. 2016年内是否有多次登记入住时间在22时-次日3时的记录
4. 2016年内是否有多次在网吧下线时间在6时-8时之间的记录
5. 2016年内是否有涉警行为
6. 是否被多名显性涉毒人员存为手机联系人

原始方案-以预测隐性涉毒人员为例

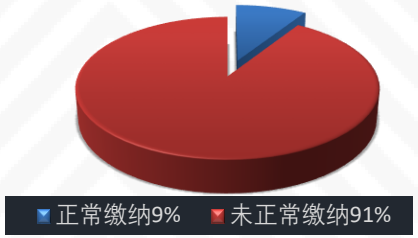
年龄结构



抽取的644名隐性涉毒人员样本中，年龄集中在18-47周岁之间的共计592人，占比92%。

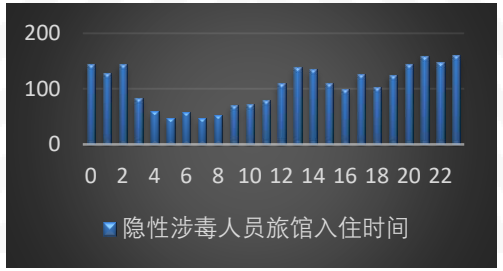
社保缴纳记录

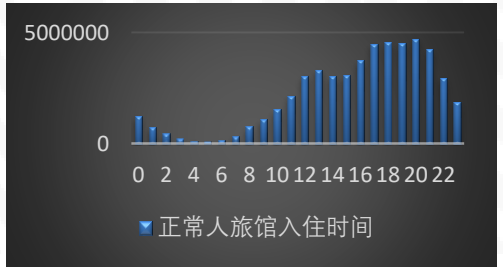
抽取的644名隐性涉毒人员中仅有58人在2016年正常缴纳社保，无社保缴纳记录及未连续缴纳社保的人员占比91%



登记旅馆入住时间

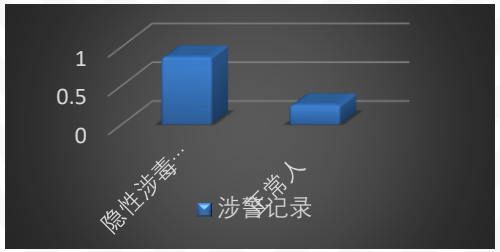
涉毒人员的登记入住时间在22时-次日3时之间达到峰值，明显有异于常人





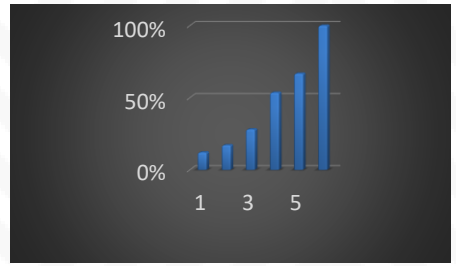
涉警行为

644名隐性涉毒人员在2016年共计涉警记录561人次，人均0.87次涉警记录而扬州市同年龄段的1273849名实有人口中，人均涉警0.25次



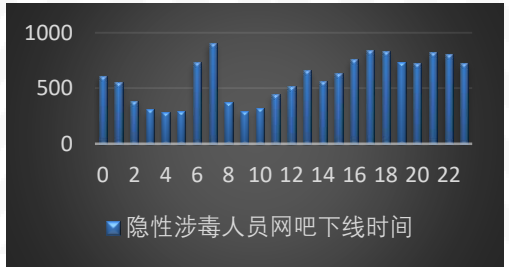
特定通联行为

被多名显性涉毒人员存为手机联系人



网吧上网下线时间

涉毒人员的下线时间在6时-8时之间达到峰值，明显有异于常人



■ 构建模型

依据蛛网抽取的211名涉毒人员通讯录，将其中被三个涉毒人员存为手机联系人的194人设为样本A，其比中的显性涉毒人员共49人，比中涉毒人员机率至少25.3%。

代码	特征	特征参数
A	被多名显性涉毒人员 存为手机联系人	被3名显性涉毒人员 存为手机联系人
B	是否正常缴纳社保	未正常缴纳
C	网吧下线时间	6时—8时
D	入住旅馆时间	22时—次日3时
E	是否涉警人员	是

■ 模型评估

用其他指标与样本A进行排列组合后，验证出比中显性涉毒人员机率的浮动情况如下，最高比中比例不超过30%

代码	特征	特征参数
A	被多名显性涉毒人员存为手机联系人	被3名显性涉毒人员存为手机联系人
B	是否正常缴纳社保	未正常缴纳
C	网吧下线时间	6时—8时
D	入住旅馆时间	22时—次日3时
E	是否涉警人员	是

组合方式	比中显性涉毒比例	组合方式	比中显性涉毒比例
样本A	25.3%	A+C	16% ↓
A+B	27.9% ↑	A+B+C	17%↓
A+D	27.3% ↑	A+C+D	16.3%↓
A+E	27.1% ↑	A+C+E	15.3%↓
A+B+D	29.8% ↑	A+B+C+D	10%↓
A+B+E	29% ↑	A+B+C+E	16.2%↓
A+D+E	26.8% ↑	A+C+D+E	15.3%↓
A+B+D+E	28.5%↑	A+B+C+D+E	15.7%↓

■ 创新方案

1、决策树

2、采样

3、特征选取

4、模型构建

5、模型评估

■ 机器学习算法-决策树

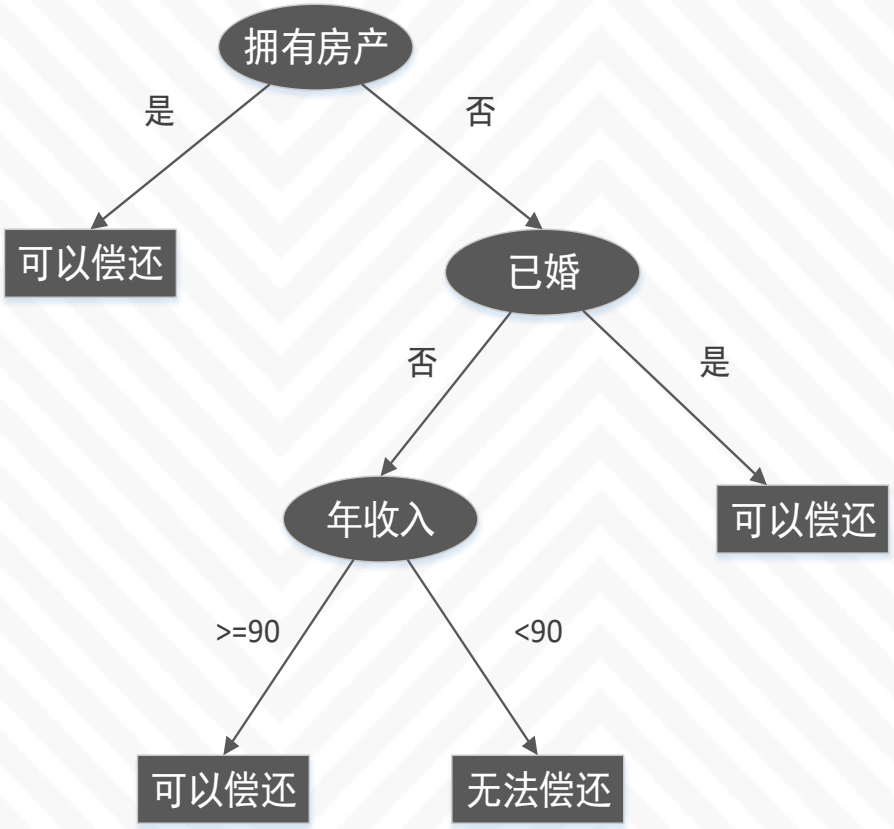
决策树算法是一种基本的分类方法，属于有监督学习，它是基于实例特征对实例进行分类的过程，也可以认为决策树就是很多if-then的规则集合，主要有以下几个特点：

1. 可读性强：模型本身由数据属性作为树枝节点构成，易于理解
2. 时间复杂度低：决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度
3. 对数据要求低：能够同时处理数据型和常规型属性
4. 随机森林：随机森林是用训练数据随机的计算出许多决策树，形成了一个森林，然后用这个森林对未知数据进行预测，选取投票最多的分类，此算法的错误率能够得到进一步降低

■ 决策树-举例说明

ID	拥有房产（是/否）	婚姻情况（单身，已婚，离婚）	年收入（单位：万元）	可以偿还债务（是/否）
1	是	单身	125	是
2	否	已婚	100	是
3	否	单身	70	否
4	是	已婚	120	是
5	否	离婚	95	是
6	否	已婚	60	是
7	是	离婚	220	是
8	否	单身	85	否
9	否	已婚	75	是
10	否	单身	90	是

ID	拥有房产	婚姻情况	年收入	可以偿还债务
11	否	单身	55	?

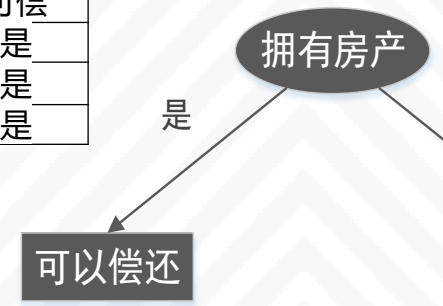


■ 决策树-如何构建-理解纯度

纯度：

如果一个分割点可以将当前的所有节点分为两类，使得每一类都很“纯”，也就是同一类的记录较多，那么就是一个好分割点。比如，“拥有房产”，可以将记录分成了两类，“是”的节点全部都可以偿还债务，非常“纯”，“否”的节点，可以偿还贷款和无法偿还贷款的人都有，不是很“纯”，但是两个节点加起来的纯度之和与原始节点的纯度之差最大，所以按照这种方法分割。构建决策树采用贪心算法，只考虑当前纯度差最大的情况作为分割点。

ID	婚姻	收入	可偿
1	单身	125	是
4	已婚	120	是
7	离婚	220	是



ID	婚姻	收入	可偿
2	已婚	100	是
3	单身	70	否
5	离婚	95	是
6	已婚	60	是
8	单身	85	否
9	已婚	75	是
10	单身	90	是

ID	收入	可偿
2	100	是
6	60	是
9	75	是

ID	收入	可偿
3	70	否
5	95	是
8	85	否
10	90	是

ID	收入	可偿
5	95	是
10	90	是

ID	收入	可偿
3	70	否
8	85	否

决策树-如何构建-量化纯度

量化纯度：三种量化方法

1、Gini不纯度：

$$Gini = 1 - \sum_{i=1}^{n} P(i)^2$$

2、熵（Entropy):

$$Entropy = -\sum_{i=1}^{n} P(i) * \log_2 P(i)$$

3、错误率:

$$Error = 1 - \max\{P(i) \mid i \in [1, n]\}$$

三个公式均是值越大，表示越“不纯”，越小表示越“纯”。

■ 决策树-如何构建-信息增益

信息增益（Information Gain）：

信息增益指划分前样本数据集的不纯度和划分后样本数据集的不纯度的差值。假设划分前的样本数据为S，并用属性A来划分样本集S，则按属性A划分S的信息增益Gain（S,A）为样本S的不纯度减去按属性A划分S后样本子集的不纯度，信息增益越大，越适合作为分割节点。

$$IG(S,A) = I(S) - I(S \mid A) = I(S) - \sum_{j=1}^{k} \frac{|A_j|}{|S|} * I(A_j)$$

其中，I代表不纯度（也就是前面三个公式的任意一种），K代表分割的节点数，一般K = 2。Aj表示子节点中的各记录。

提示：S可以看成例子中的样本总体，A可以看成拥有房产。

■ 决策树-如何构建-ID3算法

ID	性别	学生	少数民族	电脑
1	男	是	否	有
2	女	否	否	有
3	男	是	否	有
4	男	是	否	有
5	男	否	否	无
6	男	否	是	无

信息增益

$$IG(S,A)=I(S)-I(S\mid A)=I(S)-\sum_{j=1}^{k}\frac{|A_j|}{|S|}*I(A_j)$$

ID3算法用熵量化不纯度

$$Entropy=-\sum_{i=1}^{n}P(i)*\log_2 P(i)$$

ID	性别	电脑
1	男	有
3	男	有
4	男	有
5	男	无
6	男	无

ID	性别	电脑
2	女	有

$$I(S)=-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6}=0.918$$

$$I(S\mid A_{性别=男})=-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971$$

$$I(S\mid A_{性别=女})=-\frac{1}{1}\log_2 1-0\log_2 0=0$$

$$I(S\mid A_{性别})=\frac{5}{6}*I(S\mid A_{性别=男})+\frac{1}{6}*I(S\mid A_{性别=女})=0.809$$

$$IG(S,A_{性别})=I(S)-I(S\mid A_{性别})=0.109$$

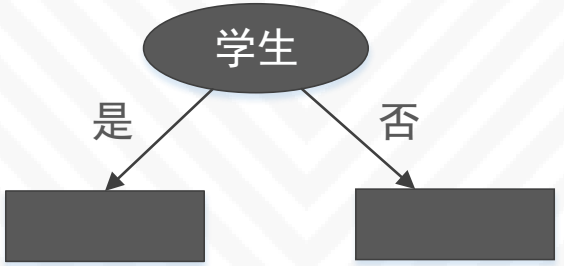
同理得到：

$$IG(S,A_{学生})=0.459$$

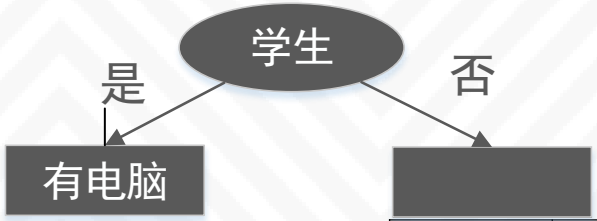
$$IG(S,A_{民族})=0.316$$

∵ 0.459>0.316>0.109

∴ 以学生为分割节点



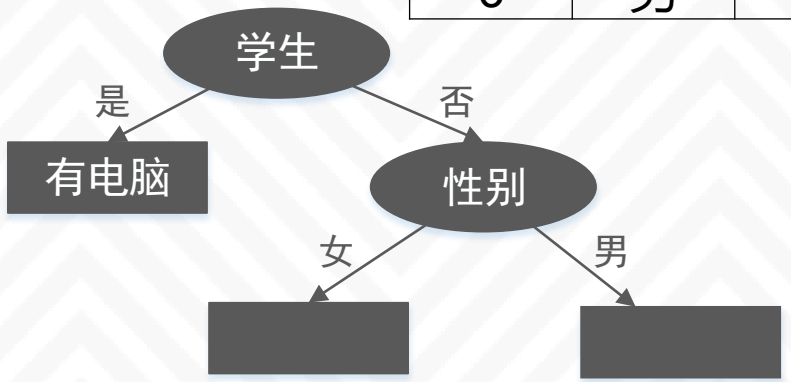
■ 决策树-如何构建-ID3算法



ID	性别	少数民族	电脑
2	女	否	有
5	男	否	无
6	男	是	无

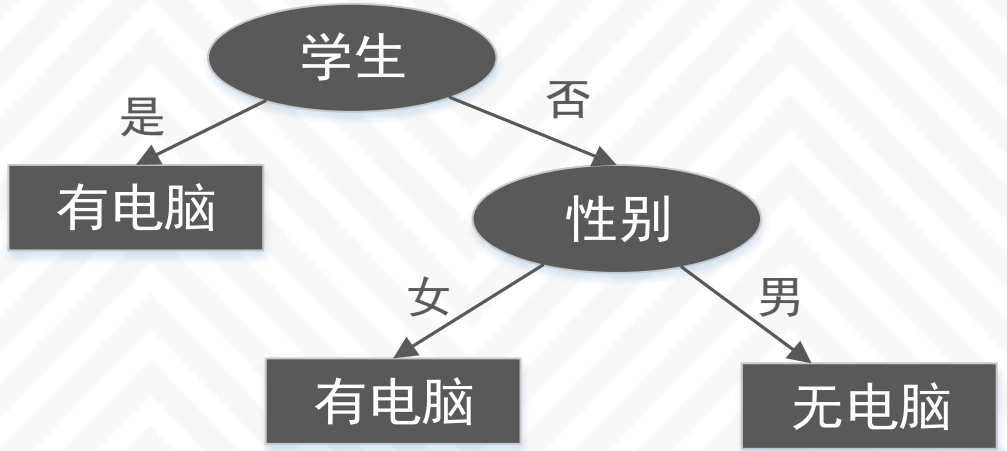
$$IG(A_{性别}) = 0.918$$

$$IG(A_{民族}) = 0.318$$



ID	少数民族	电脑
2	否	有

ID	少数民族	电脑
5	否	无
6	是	无



决策树-其他算法

C4.5算法：
使用信息增益率（Information Gain Ratio）量化不纯度：

$$IG(S,A) = I(S) - I(S \mid A) = I(S) - \sum_{j=1}^{k} \frac{|A_j|}{|S|} * I(A_j)$$

$$IGR(S,A) = \frac{IG(S,A)}{I(A)} = \frac{I(S) - I(S \mid A)}{I(A)}$$

CART算法：
使用Gini系数量化不纯度：

$$Gini = 1 - \sum_{i=1}^{n} P(i)^2$$

■ 采样

关联数据库中嫌疑人信息表和案件信息表，得到2017年所有隐性涉毒人员相关数据，并将该数据与旅馆住宿信息、手机通讯录信息、涉警行为信息等数据关联，作为正样本，将未登记在嫌疑人信息表和违法信息犯罪表中的正常人员数据作为负样本。然而，对于现有数据而言，正负样本严重不平衡，即正常人占比远大于隐性吸毒人员的占比，因此需要在两者特征属性分布基本不变的情况下对正样本采取上采样（即模拟生成和当前稀有样本临近的样本），对负样本进行下采样（即对负样本聚类，在每个类别中上按比例抽取部分样本）。最终采集正负样本各3631个，总计7262个样本，样本比例1:1。

■ 特征选择

1) 年龄：通过人员基本信息获得年龄特征
2) 性别：通过人员基本信息得到性别特征
3) 是否常驻居民：通过关联人员基本信息得到是否常驻居民特征
4) 住店入住时间：通过关联旅店住宿信息得到入住时间
5) 住店时长：通过关联旅店住宿信息得到住店时长
6) 手机联系人数量：通过关联手机通讯录信息得到手机联系人总数
7) 手机中是否存有显性涉毒人员手机号码：通过关联手机通讯录信息得到手机中存有显性涉毒人员手机号码数量
8) 是否被显性涉毒人员存为联系人：通过关联手机通讯录信息得到被显性涉毒人员存为联系人的人数
9) 住户用电量：通过关联住户信息及电表信息得到住户从入住期间平均用电量
10) 住户用水量：通过关联住户信息及水表信息得到住户从入住期间平均用水量
11) 住户燃气使用量：通过关联住户信息及燃气使用信息得到住户从入住期间平均煤气使用量
12) 网吧上线时间：通过关联网吧上网信息得到网吧上线时间
13) 网吧下线时间：通过关联网吧上网信息得到网吧下线时间
14) 网吧上网时长：通过关联网吧上网信息得到网吧上网时长
15) 2016年内乘坐飞机次数：通过关联民航信息得到本年度通过航空出行次数
16) 2016年内乘坐火车次数：通过关联铁路信息得到本年度通过火车出行次数
17) 2016年内乘坐客运次数：通过关联客运信息得到本年度通过汽车出行次数
18) 2016年内出行总次数：通过关联民航、铁路、客运信息得到本年度出行总次数。

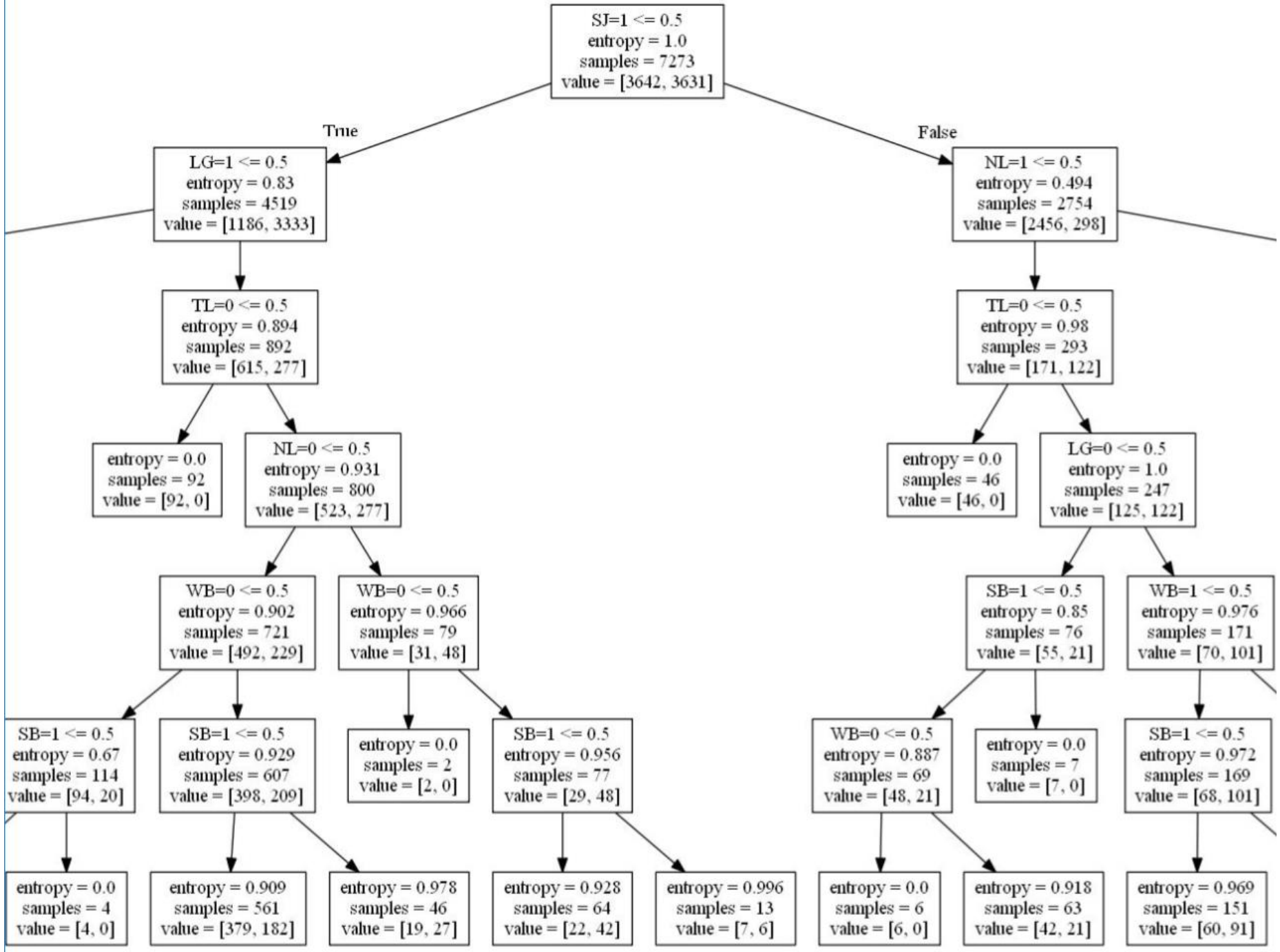
■ 特征选择

我们根据客户要求，为加快进度，使用原始方案中最初的6个特征构建模型：

1. 年龄是否在18-47周岁
2. 2016年内是否正常缴纳社保
3. 2016年内是否有多次登记入住时间在22时-次日3时的记录
4. 2016年内是否有多次下线时间在6时-8时之间的记录
5. 2016年内是否有涉警行为
6. 是否被多名显性涉毒人员存为手机联系人

■ 构建模型

以年龄是否在18-47周岁（NL）、是否正常缴纳社保（SB）、网吧下线时间（WB）、入住旅馆时间（LG）、是否涉警人员（SJ）、被显性涉毒人员存为手机联系人（TL）为特征，以是否为隐性涉毒人员（SD）为标签，使用7262中的60%的样本作为训练集，其中正样本2274个，负样本2511个，通过计算分割节点之后和之前的信息增益来选择最优分割点，最终可以构建如下决策树模型：


SJ=1 <= 0.5
entropy = 1.0
samples = 7273
value = [3642, 3631]
True
False
LG=1 <= 0.5
entropy = 0.83
samples = 4519
value = [1186, 3333]
NL=1 <= 0.5
entropy = 0.494
samples = 2754
value = [2456, 298]
TL=0 <= 0.5
entropy = 0.894
samples = 892
value = [615, 277]
TL=0 <= 0.5
entropy = 0.98
samples = 293
value = [171, 122]
entropy = 0.0
samples = 92
value = [92, 0]
NL=0 <= 0.5
entropy = 0.931
samples = 800
value = [523, 277]
entropy = 0.0
samples = 46
value = [46, 0]
LG=0 <= 0.5
entropy = 1.0
samples = 247
value = [125, 122]
WB=0 <= 0.5
entropy = 0.902
samples = 721
value = [492, 229]
WB=0 <= 0.5
entropy = 0.966
samples = 79
value = [31, 48]
SB=1 <= 0.5
entropy = 0.85
samples = 76
value = [55, 21]
WB=1 <= 0.5
entropy = 0.976
samples = 171
value = [70, 101]
SB=1 <= 0.5
entropy = 0.67
samples = 114
value = [94, 20]
SB=1 <= 0.5
entropy = 0.929
samples = 607
value = [398, 209]
entropy = 0.0
samples = 2
value = [2, 0]
SB=1 <= 0.5
entropy = 0.956
samples = 77
value = [29, 48]
WB=0 <= 0.5
entropy = 0.887
samples = 69
value = [48, 21]
entropy = 0.0
samples = 7
value = [7, 0]
SB=1 <= 0.5
entropy = 0.972
samples = 169
value = [68, 101]
entropy = 0.0
samples = 4
value = [4, 0]
entropy = 0.909
samples = 561
value = [379, 182]
entropy = 0.978
samples = 46
value = [19, 27]
entropy = 0.928
samples = 64
value = [22, 42]
entropy = 0.996
samples = 13
value = [7, 6]
entropy = 0.0
samples = 6
value = [6, 0]
entropy = 0.918
samples = 63
value = [42, 21]
entropy = 0.969
samples = 151
value = [60, 91]

■ 模型评估

从7262个样本中抽取40%数据作为测试集，其中正样本1368个，负样本1120个，使用决策树模型对该测试集进行预测，将预测结果与实际比对，计算出精确率89%，召回率83%

把正类预测为正类(TP)、把负类预测为正类(FP)、把正类预测为负类(FN)

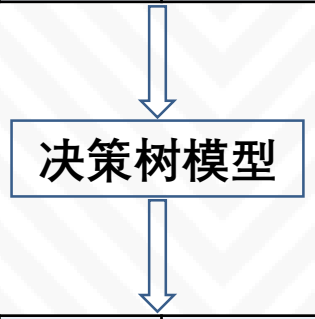
精确率P=TP/(TP+FP)

召回率R=TP/(TP+FN)

■ 使用模型预测

年龄是否在18-47周岁（NL）、是否正常缴纳社保（SB）、网吧下线时间（WB）、入住旅馆时间（LG）、是否涉警人员（SJ）、被显性涉毒人员存为手机联系人（TL）、是否为隐性涉毒人员（SD）

ID	NL	SJ	SB	TL	LG	WB	SD
1	1	1	0	0	1	0	?



ID	NL	SJ	SB	TL	LG	WB	SD
1	1	1	0	0	1	0	1

■ 方案比较

	原始方案	创新方案
精确率	<30%	89%
通用性	低	高
科学理论依据	不足	充足

思维拓展

- 分类模型组合
- 嵌入机器学习平台

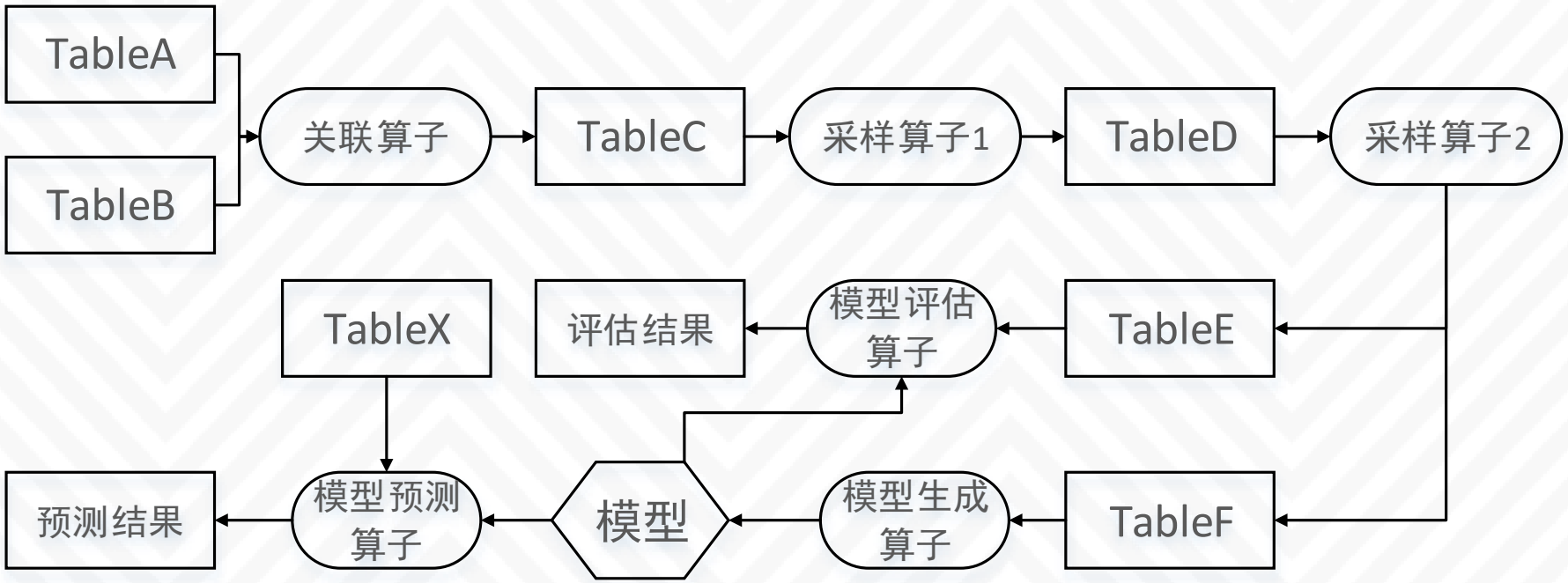
■ 多个分类模型组合

公交一卡通数据->小偷扒手预测模型

手机通话记录->传销人员预测模型

手机通话记录->电话诈骗人员预测模型

■ 将决策树分类算法嵌入到机器学习平台



谢谢观看