基于机器学习算法模型

京东注册场景数据测试报告

# 一、背景

在电商平台的用户注册的场景下，通过算法判别注册事件中出现的欺诈行为的风险。本次测试使用的算法包括专家经验的规则决策引擎，无监督机器学习算法决策系统，多维数据关联的复杂网络分析技术，用于甄别京东注册事件的数据异常情况，挖掘潜在的欺诈行为风险。

* 专家经验的规则引擎的决策引擎（基础）
* 无监督机器学习算法决策系统
* 多维数据的关系的复杂网络分析技术

# 二、数据源分析

## 1、数据源描述

**字段描述**：由于京东出于对数据控制的原因，提供的记录字段总共仅9个，详见表1。

**有效性分析**：手机号码只传前7位，事件标识和注册时间在数值上是重复的，设备字段缺失2.07%，操作系统类型字段缺失15.67%，注册来源为APP的注册时间缺乏实际意义。总体而言，数据分析样本实际有效性打了不少折扣。

备注：

1）记录总数为2222059条，

2）设备字段记录缺失原因是获取超时导致。

3）操作系统类型字段记录缺失原因是PC端注册无记录。

4）注册时长来源为APP该字段值无实际意义（因为只是服务器接口函数的调用时间），只有来源为PC的字段有实际意义。

**字段意义**：账户、设备、IP、手机号前7位、事件标识均是非类别型字段，取值倾向于唯一。

|  |  |  |
| --- | --- | --- |
| 字段 | 字段类型 | 备注 |
| 账户 | string |  |
| 设备 | string |  |
| ip | string |  |
| 注册来源 | string | 类别型字段 |
| 手机号前七位 | string |  |
| 注册时间 | string | 长度为13位的时间戳 |
| 注册时长 | float | 转换成float处理，ms |
| 事件标识 | string | 数值与注册时间无区别 |
| 操作系统类型 | string | 类别型 |

表1 数据源字段类型描述

图1 数据源字段缺失率分布图

|  |  |  |
| --- | --- | --- |
| 字段 | 字段缺失记录数 | 字段缺失率 |
| 账户 | 0 | 0.00% |
| 设备 | 46096 | 2.07% |
| ip | 0 | 0.00% |
| 注册来源 | 0 | 0.00% |
| 手机号前七位 | 0 | 0.00% |
| 注册时间 | 0 | 0.00% |
| 注册时长 | 0 | 0.00% |
| 事件标识 | 0 | 0.00% |
| 操作系统类型 | 348111 | 15.67% |

表2 数据源记录缺失统计

## 2、总体分析

**时间范围**：按与京东的约定，本次测试数据的样本以20170817传入的为准，其它日期传入的数据不计入测试样本。因此数据日期涉及到四天（20170417—20170420）的数据。

**数据总量**：共计2222059条注册记录。

|  |  |
| --- | --- |
| 日期 | 每天调用量 |
| 20170417 | 585707 |
| 20170418 | 582491 |
| 20170419 | 521934 |
| 20170420 | 531927 |
| 合计 | 2222059 |

表3 日调用量分布

图2 日调用量分布图

## 3、字段分析

**1、类别型字段**

（1）、注册来源

|  |  |  |
| --- | --- | --- |
| 注册来源 | 记录数 | 占比 |
| 主站 | 327546 | 14.74% |
| App | 1894513 | 85.26% |

表4 注册来源的分布比例

（2）操作系统类型

|  |  |  |  |
| --- | --- | --- | --- |
| 操作系统 | 记录数 | 占比 | 备注 |
| M | 973496 | 43.81% | M端是用浏览器中网页看的（无法区分pc、手机） |
| android | 781992 | 35.19% | 需合并 |
| 缺失 | 348111 | 15.67% | Pc端获取不到 |
| apple | 114960 | 5.17% |  |
| ipad | 2789 | 0.13% |  |
| ios | 679 | 0.03% |  |
| shangou-android | 14 | 0.00% | 脏数据 |
| 14 | 11 | 0.00% | 脏数据 |
| Android | 7 | 0.00% | 需合并 |

表5 操作系统类型分布比例

**2、数值型字段**

（1）注册时长

发现注册时长数值异常，最大和最小值范围相差非常大。见下表：

|  |  |
| --- | --- |
| 分位点 | 对应注册时长数值 |
| 最小值 | 2 |
| 25%分位数 | 6 |
| 中位数 | 11 |
| 75%分位数 | 17 |
| 最大值 | 3593666 |

表6 注册时长分位数对应表

经京东的业务专家解释：单位是一致，为毫秒，但通过PC和APP的注册时长含义不一致：

App注册时长：服务器端的接口函数的调用时间，即从函数接口调用，到返回的时间长度，和注册行为时间长度无关。

PC注册时长：从进入注册页面，到页面提交的时间长度，反映注册行为的时间长度。

# 三、判别模型构建与效果

## 1、特征构建

以用户注册事件作为主维度，从横向和纵向两个维度构建特征集合。

**横向特征**：构建设备、IP等字段关联的全局特征。（缺失手机号，邮箱等数据）。

**纵向特征**：构建事件的时序特征，以注册事件的时间为基准，向前回溯一段时间，或者N个事件的特征。

**关系图特征**：构建以设备、账号为节点，提取节点之间关联的复杂网络特征。

构造纵向特征共计13个，横向关联特征流量30个，关系图特征3个。

## 2、复杂网络关系挖掘

从复杂网络关系图挖掘设备关联账号的异常团簇，先前数据分析注册数据字段少，仅画出设备和账号之间的关联关系。图中绿色圈代表设备，其外围的红点代表账号。下图展示了设备关联账号超过100个数目的异常团簇，共计14个团，最大一个团关联12212个账号。并且可以看出，每个团簇之间独立，相互之间没有关联的网络连接，这因为注册场景时间短，设备与账号之间关联非常弱。



图8 设备与账号的网络关系图

## 3、专家经验规则决策引擎

根据专家经验对注册场景的欺诈的行为特征提炼，构建欺诈判别的经验模型。选择其中有意义的时间和关联维度特征。并依据数据源的分布情况，构造决策规则，以及设定对应的阈值条件。

（1）选择的特征及阈值，如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 特征名缩写 | 阈值 | 基础分 | 权重分 |
| v\_dev\_1min\_account\_cnt | 2 | 60 | 1 |
| v\_dev\_10min\_account\_cnt | 3 | 60 | 1 |
| v\_dev\_1h\_account\_cnt | 3 | 60 | 1 |
| v\_dev\_1day\_account\_cnt | 4 | 60 | 1 |
| v\_ip\_1min\_account\_cnt | 5 | 30 | 1 |
| v\_ip\_1min\_dev\_cnt | 5 | 30 | 1 |
| v\_ip\_10min\_account\_cnt | 10 | 30 | 0.5 |
| v\_ip\_10min\_dev\_cnt | 10 | 30 | 0.5 |
| 特征名缩写 | 阈值 | 基础分 | 权重分 |
| h\_ip\_risk\_label\_num | 3 | 30 | 10 |
| h\_ip\_isproxy\_http | 1 | 20 |  |
| h\_ip\_isproxy\_socks | 1 | 20 |  |
| h\_ip\_isproxy\_vpn | 1 | 20 |  |
| h\_ip\_type\_idc | 1 | 20 |  |
| h\_ip\_isforeign | 1 | 10 |  |
| h\_ip\_isreserve | 1 | 10 |  |
| h\_ip\_format\_abnormal | 1 | 10 |  |
| h\_dev\_no\_id | 1 | 20 |  |
| h\_is\_sensetive\_time | 1 | 5 |  |
| h\_is\_same\_addr\_mob\_ip | 1 | 8 |  |
| h\_ip\_10min\_cnt\_max | 10 | 20 | 1 |
| h\_ip\_10min\_account\_max | 10 | 20 | 5 |
| h\_dev\_10min\_cnt\_max | 3 | 30 | 2 |
| h\_dev\_10min\_account\_max | 3 | 30 | 10 |
| h\_dev\_account\_cnt | 4 | 30 | 10 |
| h\_mob\_isdata | 1 | 60 |  |

表7 JD注册场景特征与阈值表

（2）经过行业专家评审的风险分计算公式如下：

计算公式：基础分\*（值 > 阈值）+ 权重分\*（值-阈值）

（3）专家经验规则策略引擎对注册记录总量的判别结果如下：

选择风险分数为60分及以上的规则拒绝一共144446，其中主站被拒绝89908，app被拒绝386175条。



表8 记录风险分布表

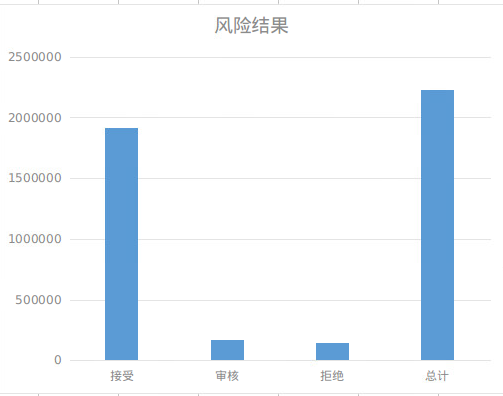


图9 记录风险分布图

（4）规则命中情况分析：

从下表可以分析出：

1、规则命中次数最多的分别是：设备ID缺失、IP命中风险标签的个数。

2、设备ID的缺失对结果有不少的扰动，其他两个规则命中，则说明规则引擎发现出现短时间内出现大量垃圾注册，以及发现可疑IP的情况。



表9 规则命中分布表

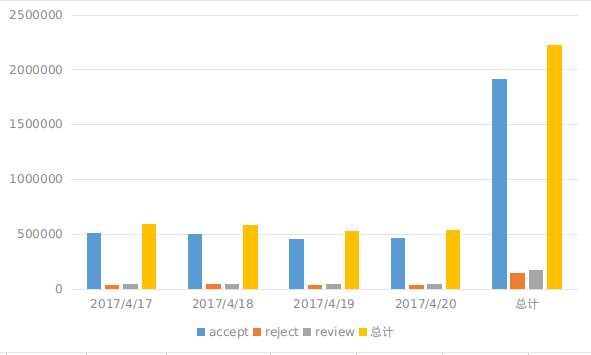
规则每日命中情况分析：从每日分布图来看，每日的异常占比相差不大。

图10 规则引擎命中每日分布图

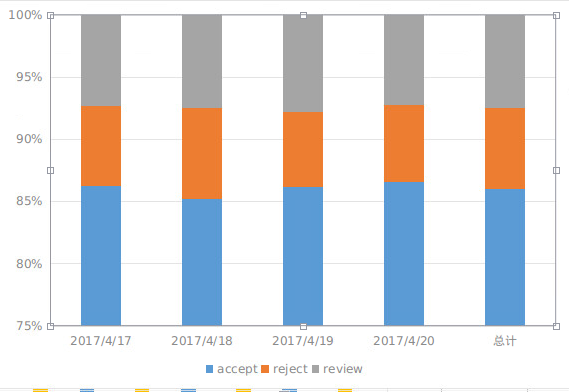


图11 规则引擎每日命中类别比例图

## 4、无监督机器学习算法决策系统

注册场景的记录数据，京东没有提供标签信息，因此无法采用有监督的机器学习算法来训练模型和预测结果，而采用无监督的机器学习算法。对于该测试项目，选择了孤立森林和混合高斯算法模型。通过对注册数据来训练，对正常的数据进行建模，学习数据的分布。然后使用模型寻找可疑的异常数据，提供给人工进一步核实。

**1）孤立森林模型（iForest）**

**输入特征变量：**

|  |  |
| --- | --- |
| 特征英文名 | 特征中文名 |
| v\_dev\_1min\_cnt | 设备1分钟内出现次数 |
| v\_dev\_10min\_cnt | 设备10分钟内出现次数 |
| v\_dev\_1hour\_cnt | 设备1小时内出现次数 |
| v\_dev\_1day\_cnt | 设备1天内出现次数 |
| v\_dev\_1min\_account\_cnt | 设备1分钟关联账户数 |
| v\_dev\_10min\_account\_cnt | 设备10分钟关联账户数 |
| v\_dev\_1hour\_account\_cnt | 设备1小时关联账户数 |
| v\_dev\_1day\_account\_cnt | 设备1天关联账户数 |
| v\_dev\_1min\_ip\_cnt | 设备1分钟关联ip数 |
| v\_dev\_10min\_ip\_cnt | 设备10分钟关联ip数 |
| v\_dev\_1hour\_ip\_cnt | 设备1小时关联ip数 |
| v\_dev\_1day\_ip\_cnt | 设备1天关联ip数 |
| v\_ip\_1min\_cnt | 1分钟内ip出现次数 |
| v\_ip\_10min\_cnt | 10分钟内ip出现次数 |
| v\_ip\_1hour\_cnt | 1小时内ip出现次数 |
| v\_ip\_1day\_cnt | 1天内ip出现次数 |
| v\_ip\_1min\_account\_cnt | 1分钟内ip关联账户数 |
| v\_ip\_10min\_account\_cnt | 10分钟内ip关联账户数 |
| v\_ip\_1hour\_account\_cnt | 1小时内ip关联账户数 |
| v\_ip\_1day\_account\_cnt | 1天内ip关联账户数 |
| v\_ip\_1min\_dev\_cnt | 1分钟内ip关联设备数 |
| v\_ip\_10min\_dev\_cnt | 10分钟内ip关联设备数 |
| v\_ip\_1hour\_dev\_cnt | 1小时内ip关联设备数 |
| v\_ip\_1day\_dev\_cnt | 1天内ip关联设备数 |
| v\_account\_1min\_cnt | 1分钟内账户出现次数 |
| v\_account\_10min\_cnt | 10分钟内账户出现次数 |
| v\_account\_1hour\_cnt | 1小时内账户出现次数 |
| v\_account\_1day\_cnt | 1天内账户出现次数 |
| v\_account\_1min\_dev\_cnt | 1分钟内账户关联设备数 |
| v\_account\_10min\_dev\_cnt | 10分钟内账户关联设备数 |
| v\_account\_1hour\_dev\_cnt | 1小时内账户关联设备数 |
| v\_account\_1day\_dev\_cnt | 1天内账户关联设备数 |
| v\_account\_1min\_ip\_cnt | 1分钟内账户关联ip数 |
| v\_account\_10min\_ip\_cnt | 10分钟内账户关联ip数 |
| v\_account\_1hour\_ip\_cnt | 1小时内账户关联ip数 |
| v\_account\_1day\_ip\_cnt | 1天内账户关联ip数 |
| h\_ip\_risk\_label\_num | ip命中风险标签的个数 |
| h\_ip\_wool\_occur\_days | ip发生薅羊毛的天数 |
| h\_ip\_signup\_occur\_days | ip发生垃圾注册的天数 |
| h\_ip\_scapling\_occur\_days | ip发生黄牛的天数 |
| h\_ip\_click\_occur\_days | ip发生虚假点击的天数 |
| h\_ip\_spam\_occur\_days | ip发生垃圾信息的天数 |
| h\_ip\_sms\_occur\_days | ip发生短信轰炸的天数 |
| h\_ip\_isproxy\_http | ip是否是http代理 |
| h\_ip\_isproxy\_socks | ip是否是socks代理 |
| h\_ip\_isproxy\_vpn | ip是否是vpn代理 |
| h\_ip\_third\_risk\_cnt | ip命中第三方情报种类 |
| h\_ip\_type\_idc | ip类型是否是机房 |
| h\_ip\_isforeign | 是否是国外ip |
| h\_ip\_isreserve | 是否是保留ip |
| h\_ip\_format\_abnormal | ip格式是否正确 |
| h\_dev\_no\_id | 设备缺失 |
| h\_sign\_time | 注册时长 |
| h\_is\_sensetive\_time | 是否是敏感时段注册 |
| h\_is\_same\_addr\_mob\_ip | ip和手机号注册地址是否一致 |
| h\_ip\_10min\_cnt\_max | ip10分钟最大注册次数 |
| h\_ip\_10min\_cnt\_mean | ip10分钟平均次数 |
| h\_ip\_10min\_account\_max | ip10分钟最大注册账户 |
| h\_ip\_10min\_account\_mean | ip10分钟平均注册账户数 |
| h\_dev\_10min\_cnt\_max | 设备10分钟最大注册次数 |
| h\_dev\_10min\_cnt\_mean | 设备10分钟平均注册次数 |
| h\_dev\_10min\_account\_max | 设备10分钟最大注册账户数 |
| h\_dev\_10min\_account\_mean | 设备10分钟平均注册账户数 |
| h\_dev\_account\_cnt | 设备注册账户数 |
| h\_mob\_isdata | 是否流量卡 |

**模型训练参数配置：**

由于PC接入和APP接入在关键特征注册时长有较大差异，因此模型和数据集采用分别训练的方式。

iTree树的数量：200棵

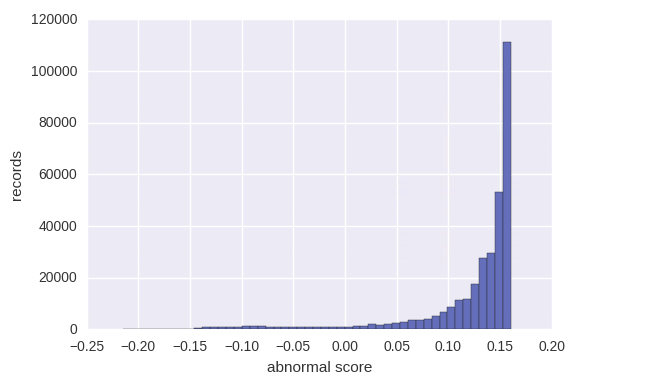
训练记录数量： PC为327546条，APP为947256条

预测记录数量：PC为327546条，APP为1894513条

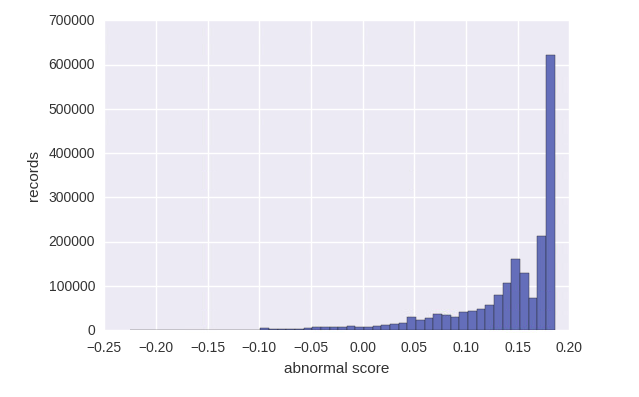
**预测结果：**

用模型来预测的结果如下：

|  |  |  |
| --- | --- | --- |
|  | PC渠道 | APP渠道 |
| 参数iTree树数 | 200 | 200 |
| 训练集记录数 | 327546 | 947256 |
| 预测记录数 | 327546 | 1894513 |
| 异常分阈值 | 0 | -0.08 |
| 异常记录数量 | 18647 | 8064 |
| 异常记录占比  （异常记录数／预测记录数） | 5.7% | 0.42% |



主站渠道--iForest 模型异常分分布



APP渠道--iForest 模型异常分分布

**预测效果评价：**

1）拒绝记录的交叉分析，分主站和APP两种渠道统计分析。

主站渠道的异常记录与规则引擎交集的数量为16040条，占异常记录的比率为 16040 ／18647 = 86%

APP渠道的异常记录於规则引擎交集的数量为5410条，占异常记录的比率为5410/8064 = 67%

2）预测试异常结果的人工评价，选取异常得分最低的记录6047条（异常分小于-0.1的部分，分数越低，记录越异常），抽样后人工打标，准确率81%，有部分ip及标签类的特征取值异常，但不足以把记录标黑。

从中可见，无监督机器学习筛选的专家经验规则引擎的的判别结果有较大的重合部分，可以作为预测的一种参考，相当于另外一个维度的专家意见。

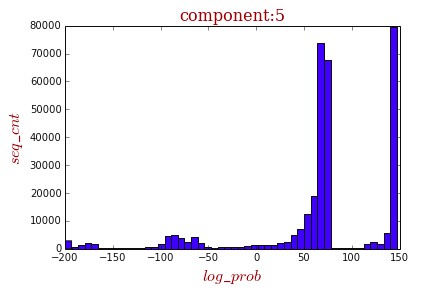
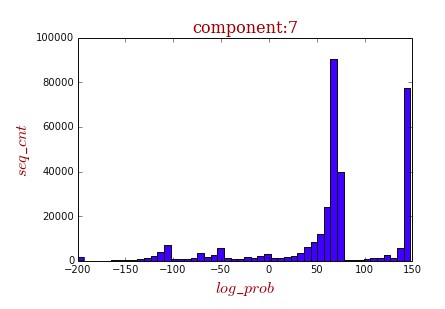
**2）混合高斯模型GMM**

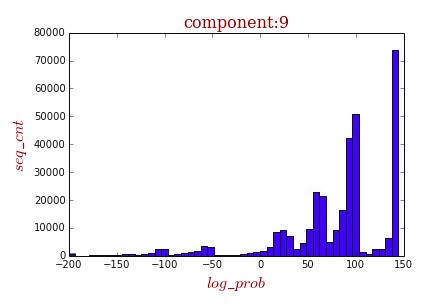
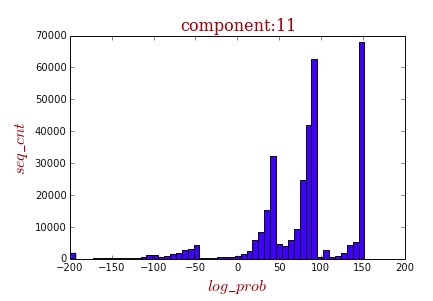
GMM(Gaussian mixture model) 混合高斯模型在机器学习、计算机视觉等领域有着广泛的应用。其典型的应用有概率密度估计、背景建模、聚类等。

其中是系数，=1;是高斯分布密度。

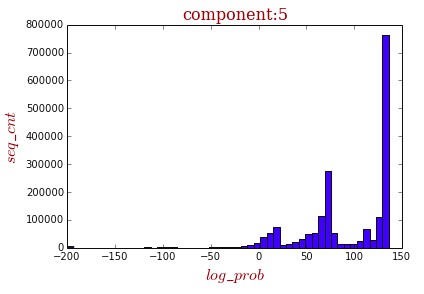
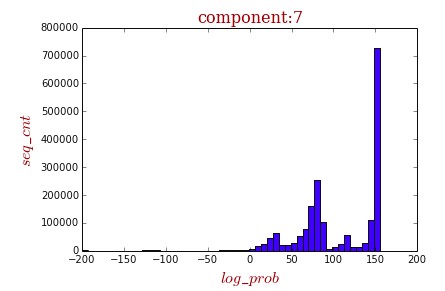
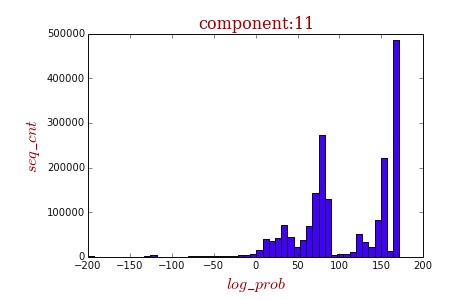
**模型参数选择：**

分别对主站渠道，APP渠道两种来源对注册记录的异常情况进行评估，分别设定K=5，7，9，11四种情况分析。如下

主站渠道K=5，7，9，11的记录分布

APP渠道K=5，7，9，11的记录分布

算法模型参数选择：将不同的K值得出的结果，根据分布：对于主站：取0为分界点；对于app，取10为分界点；

将模型结果和规则结果对比得出如下表格：

**主站结果：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阈值为0 |  |  |  |  |
| K值 | 拒绝数 | 规则重合数 | 拒绝占比 | 重合占比 |
| 5 | 41261 | 25984 | 12.60% | 62.97% |
| 7 | 45622 | 27006 | 13.93% | 59.20% |
| 9 | 25895 | 22238 | 7.91% | 85.87% |
| 11 | 23907 | 19638 | 7.30% | 82.14% |
| 阈值10 |  |  |  |  |
| K值 | 拒绝数 | 重合数 | 拒绝占比 | 重合占比 |
| 5 | 43311 | 26599 | 13.22% | 61.41% |
| 7 | 47371 | 27458 | 14.46% | 57.96% |
| 9 | 28657 | 23082 | 8.75% | 80.55% |
| 11 | 25455 | 20325 | 7.77% | 79.85% |

**APP结果：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阈值为10 |  |  |  |  |
| K值 | 拒绝数 | 规则重合数 | 拒绝占比 | 重合占比 |
| 5 | 113414 | 50293 | 5.99% | 44.34% |
| 7 | 39838 | 19523 | 2.10% | 49.01% |
| 9 | 42155 | 22633 | 2.23% | 53.69% |
| 11 | 60084 | 28814 | 3.17% | 47.96% |
| 阈值为20 |  |  |  |  |
| K值 | 拒绝数 | 重合数 | 拒绝占比 | 重合占比 |
| 5 | 230677 | 87036 | 12.18% | 37.73% |
| 7 | 69756 | 31971 | 3.68% | 45.83% |
| 9 | 70927 | 37346 | 3.74% | 52.65% |
| 11 | 112546 | 48324 | 5.94% | 42.94% |

结论：从与专家经验的匹配度来看， K=9，主站阈值取0，APP阈值取10，结果较好。

**预测效果评价：**

1）拒绝记录的交叉分析，分主站和APP两种渠道统计分析。

主站渠道的异常记录与规则引擎交集的数量为22238条，占异常记录的比率为 22238 ／25895 = 85.87%

APP渠道的异常记录於规则引擎交集的数量为22633条，占异常记录的比率为22633/42155 = 53.69%

从中可见，无监督机器学习筛选的专家经验规则引擎的的判别结果有较大的重合部分，可以作为预测的一种参考.

## 5、多模型决策系统

**几个模型对比分析**

专家经验的规则引擎决策模型，其效果依赖于行业专家的经验积累，以及对场景的欺诈特征的深刻理解，构造合理的判别规则，选取合适的阈值。对专家的先验知识积累依赖深。

无监督机器学习算法决策模型，其本质是挖掘输入的数据的异常特征，依赖于数据源的特点与分类。对行业领域内的数据分析专家经验的依赖较低。有利于解决无标签情况下的人工智能技术的运用，提高开发效率。

在有准确标签的情况下，可结合有监督的机器学习算法

**联合模型策略**

需要根据实际应用场景对模型的结果进行组合。比如在对拒绝准确率需求高的场景，可以采用各个模型的交集，保证拒绝的准确性，而对召回率需求高的场景，可以增加各个模型独立发现的部分进行补充。

1. 最小集合的组合策略，取专家规则引擎，几个无监督算法模型的预判结果的交集作为风险高的子集，剩余作为正常集合。拒绝与正常的比值在组合策略中最小，也就是保证每次拒绝的准确性，牺牲就是存在漏网的鱼。
2. 最大集合的组合策略，取专家规则引擎，几个无监督算法模型的预判结果的并集作为高风险的 子集，剩余的作为正常集合。拒绝与正常的比值在组合策略中最大。也就是把一切可疑的数据都抓住，牺牲就是每次拒绝的准确性，存在误判的情况。
3. 混合的组合策略，例如，以专家规则策略引擎作为基础，有条件的增加拒绝集合。又例如，可以认为每个模型都是一个专家意见，通过投票方式进行决策拒绝。

模型1 判别结果

模型2 判别结果

模型3 判别结果

几种模型判别结果集合示意图