SMARTBI
思迈特软件

如何利用机器学习提高新冠病毒感染临床诊断的准确率

Smartbi，更聪明的大数据分析软件

Smartbi大数据产品研发部

目　录

CONTENTS

01需求场景

- 新冠病毒肆虐
- 确诊速度低，假阴性病例
- 医生大面积感染

需求场景

背景

2019新型冠状病毒（2019-nCoV），因2019年武汉病毒性肺炎病例而被发现，2020年1月12日被世界卫生组织命名。新型冠状病毒是以前从未在人体中发现的**冠状病毒新毒株。**

从湖北爆发的新冠肺炎病毒已经蔓延到我国很多地方，海外部分国家也出现了感染者，世卫组织(WHO)也紧急宣布，将中国疫情列为**国际关注突发公共卫生事件(PHEIC)。**

人感染了冠状病毒后常见体征有**呼吸道症状、发热、咳嗽、气促和呼吸困难**等。在较严重病例中，感染可导致**肺炎、严重急性呼吸综合征、肾衰竭**，甚至死亡。这些症状和季节性流感症状类似，不易区分；目前对于新型冠状病毒所致疾病没有特异治疗方法。

中国科学家在极短时间内完成了新冠病毒的基因测序并开放共享给世界，为世界各地快速开发诊断病毒识别的**核酸检测**赢得了宝贵的时间。

需求场景

核酸检测

利用从病患上呼吸道采集的咽拭子样本，检测样本中是否存在新冠病毒核酸。

临床诊断

依靠资深的临床医生的丰富经验，针对病患的临床表现、检查化验以及流行病学调查等数据做出诊断结果。

需求场景

问题

诊断准确性低

新冠病毒诊断的核酸检测准确性不高，核酸检测常出现多次“假阴性”，导致病患久久不能确诊治疗，病情迅速恶化甚至死亡。

诊断速度慢

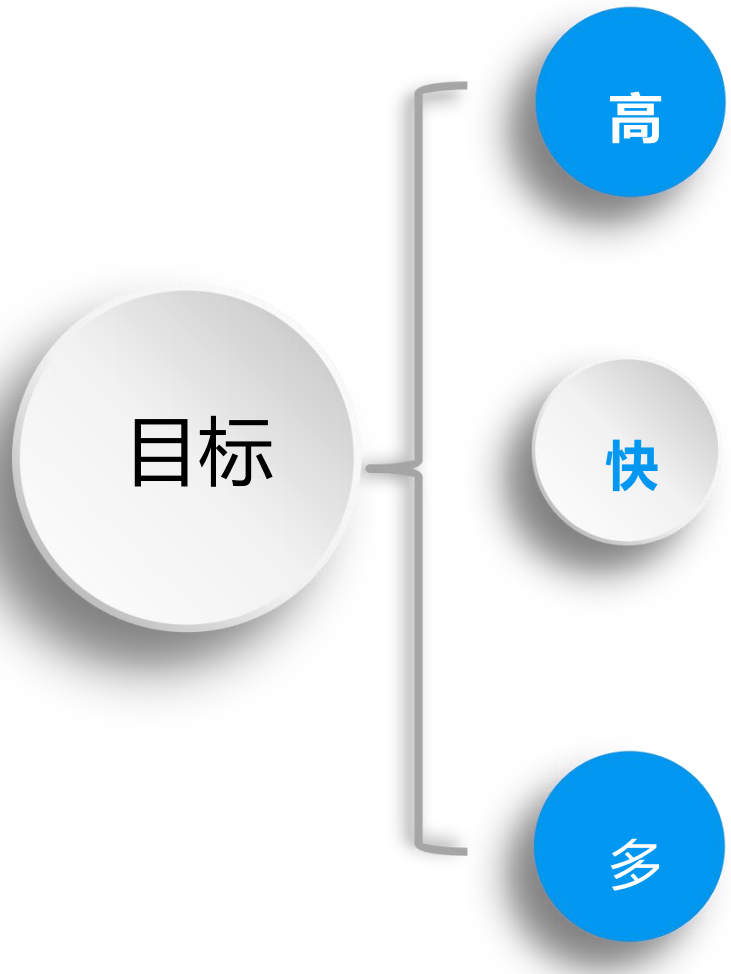
新冠病毒积累的疑似病例比较多，如不能快速确诊将导致病毒的进一步传播。**李文亮医生**从1月12日因高度怀疑住院到2月1日才确诊，竟然用了20天之久，第3次核酸检测阳性才确诊。

临床诊断经验的医生少

新冠病毒感染临床诊断需要经验丰富的资深医生，而大面积的医护人员感染和高强度的工作，导致一般的医生已经不堪重负，更不用说资深的医生了。中疾控报告：逾3000名医务人员感染新冠病毒。

02目标及方法

目标定义



提高诊断新冠病毒感染的**准确率**，及早对病人进行隔离、治疗。

加快新冠病毒感染诊断的**速度**，尽快排除积累的疑似病人堰塞湖危险。

让**更多**非资深的**一般医生**具有诊断识别新冠病毒感染者的能力。

解决方法

高

高准确率

机器学习能够大数据量不断的学习迭代和调优诊断模型，能够大幅度的提高病患的诊断准确率。

快

快速度

将病患数据批量发给训练好的诊断模型，模型能够立即给出诊断所有结果。

多

多医生

在诊断模型的辅助下，只需一般的医生即可完成对新冠病毒感染的诊断识别，使更多的医生具有临床诊断新冠病毒的能力。诊断模型相当于多个资深医生。

模型选择

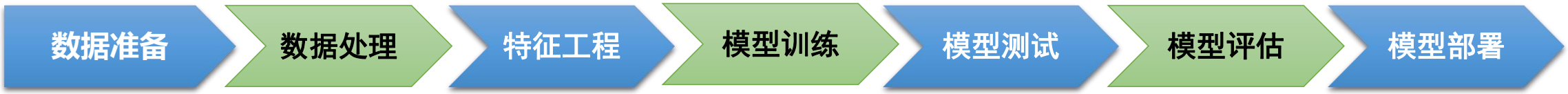
分类模型

由于已经具有了医生的诊断数据，并且有了诊断结果，因此可以利用机器学习中的监督学习模型进行训练。诊断结果是感染或者没有感染这两种情况之一，所以就是一个二分类模型。

03实施过程

SMARTBI 思迈特软件

模型构建流程



数据准备	数据处理	特征工程	模型训练	模型测试	模型评估	模型部署
数据包含流行病学、临床表现、检查化验三个维度，26特征	• 不平衡数据处理； • 数值化处理； • 离散化处理；	• 特征选择； • 特征分析	• 逻辑回归； • GBDT；	• 低方差； • 低偏差；	• 准确率； • Auc； • Roc；	• RESTFUL服务； • Json数据格式； • Spark、Python服务；

数据准备

- 样本构建：根据具体的业务场景，构建数据样本
- 维度选择：流行病学史调查数据、临床表现数据、检查化验数据

当前显示 100 条 / 总共有 100 条数据　提示:点击单元格可查看超出的内容

乏力	干咳	发病天数	鼻塞	流涕	咽痛	腹泻	呼吸困难	血氧浓度	外周血白细胞总数	淋巴细胞计数	肝酶	LDH	肌酶	肌红蛋白
是	是	5	是	否	是	是	否	低	低	减少	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
是	是	1	是	否	是	否	否	低	低	减少	正常	正常	正常	正常
否	是	8	否	是	是	否	否	正常	正常	减少	正常	正常	正常	正常
否	是	6	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	6	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	4	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	10	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	13	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	2	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常
否	是	12	否	是	是	否	否	正常	正常	正常	正常	正常	正常	正常

表头真名　表头别名

数据处理

- 不平衡数据处理；
- 数值化处理；
- 离散化处理

分箱设置

字段	区间
idnew_table	
体温	-INF,37.2,INF
发病天数	-INF,3,7,14,INF
疫区或病患社区旅行史Index	
病患接触史Index	
疫区人员接触史Index	
是否有聚集性活动Index	
乏力Index	
干咳Index	

* 以逗号分隔离散区间、负无穷为-INF、正无穷为INF。示例：-INF,30,60,90,INF

确定 取消

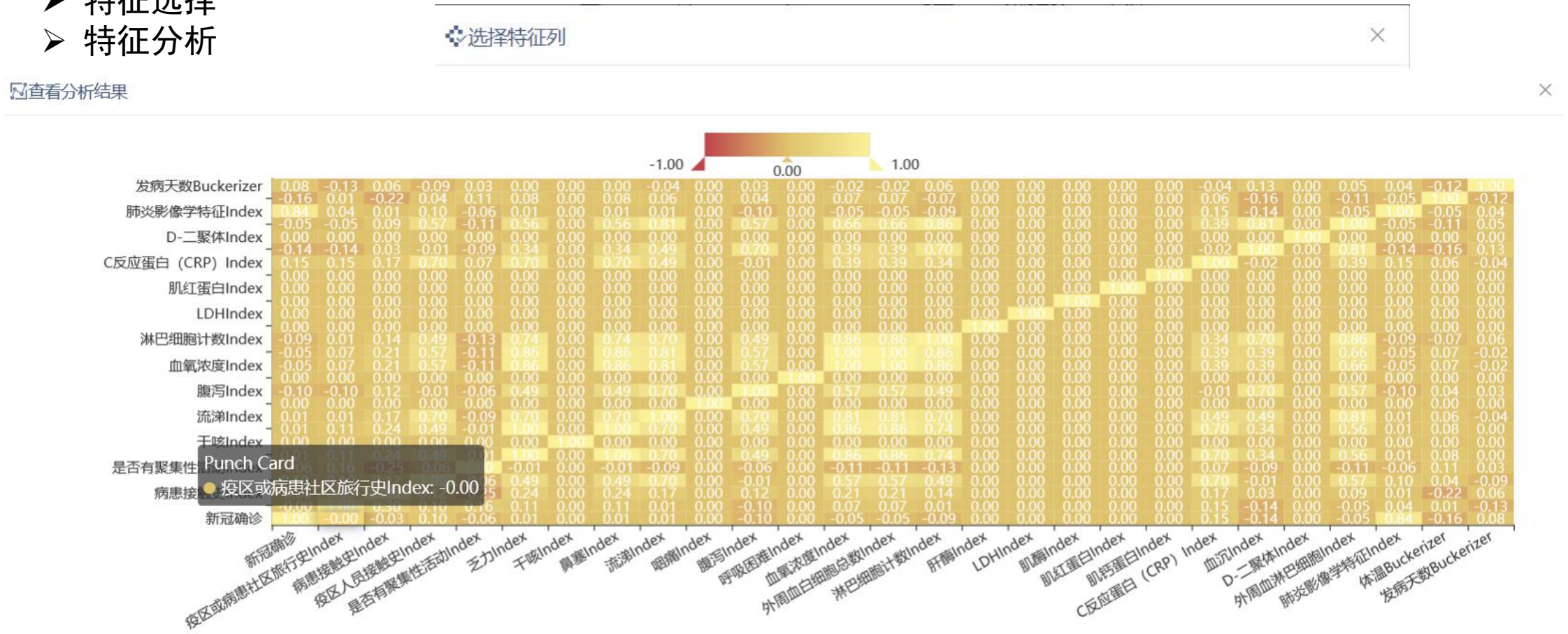
0.0 0.0 0.0

体温Buckerizer	发病天数Buckerizer
1.0	1.0
1.0	0.0
1.0	0.0
0.0	2.0
1.0	1.0
0.0	0.0
1.0	1.0
1.0	0.0
0.0	1.0

表头真名 表头别名

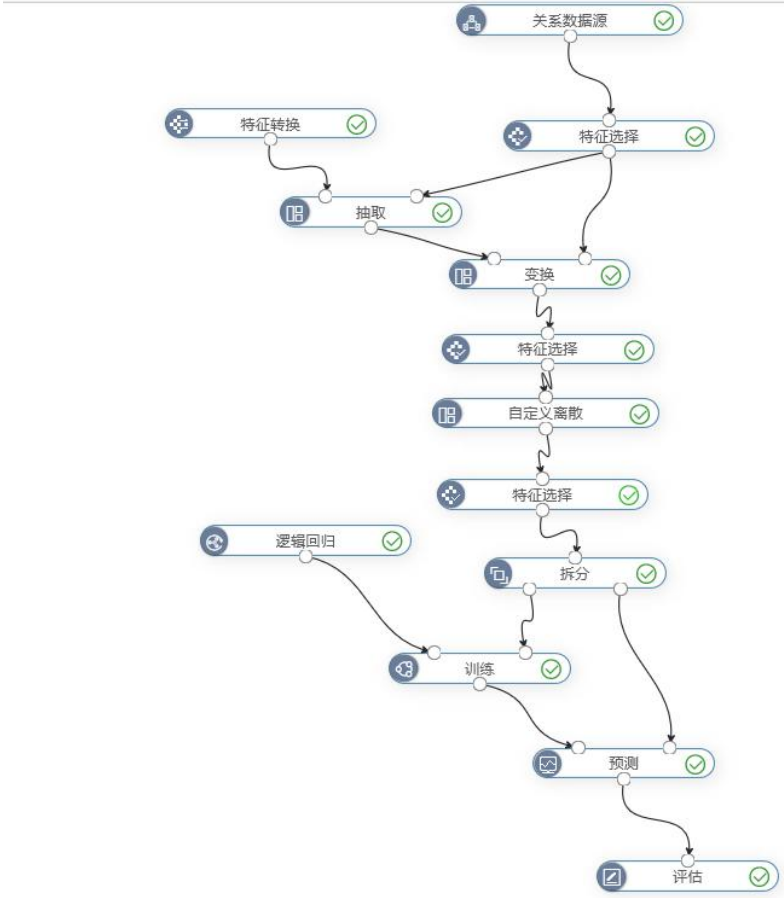
特征工程

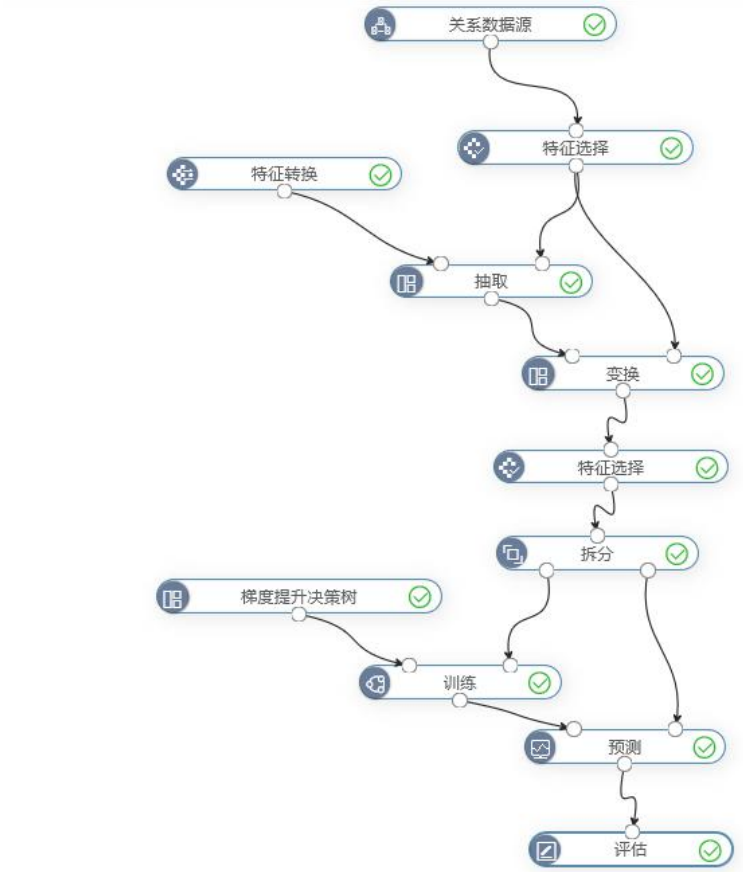
- 特征选择
- 特征分析



模型训练

- Logistic Regression
- GBDT





模型测试

- 低方差
- 低偏差

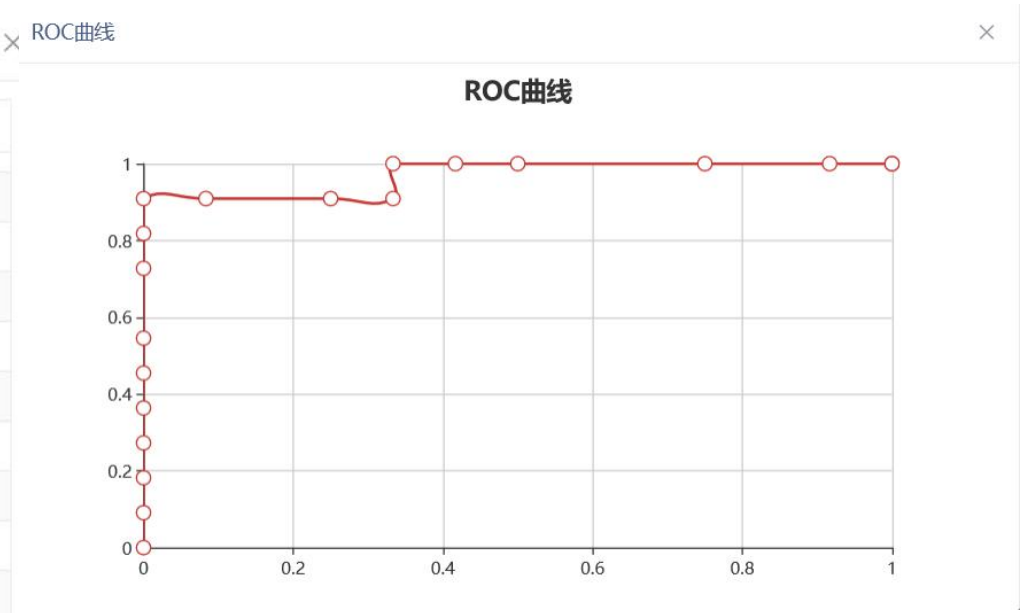
温Buckerizer	发病天数Buckerizer	features	featuresNormalized	rawPrediction	probability	prediction
1.0	1.0	(26,[24,25],[1.0,1.0])	(26,[24,25],[1.0,1.0])	[-2.3534121431771786,2.35...	[0.08679493899986007...	1.0
1.0	2.0	(26,[24,25],[1.0,2.0])	(26,[24,25],[1.0,2.0])	[-2.0054521821422675,2.00...	[0.11863166531474971...	1.0
0.0	2.0	(26,[25],[2.0])	(26,[25],[2.0])	[-0.9801860861535616,0.98...	[0.2728548616590124,...	1.0
1.0	0.0	(26,[23,24],[1.0,1.0])	(26,[23,24],[1.0,1.0])	[1.7332008351982036,-1.73...	[0.8498213844195708,...	0.0
1.0	2.0	(26,[23,24,25],[1.0,1.0,2.0])	(26,[23,24,25],[1.0,1.0,2.0])	[2.429120757268026,-2.429...	[0.9190211225011286,...	0.0
1.0	2.0	(26,[23,24,25],[1.0,1.0,2.0])	(26,[23,24,25],[1.0,1.0,2.0])	[2.429120757268026,-2.429...	[0.9190211225011286,...	0.0
1.0	0.0	(26,[24],[1.0])	(26,[24],[1.0])	[-2.7013721042120897,2.70...	[0.06289243985992832...	1.0
1.0	0.0	(26,[24],[1.0])	(26,[24],[1.0])	[-2.7013721042120897,2.70...	[0.06289243985992832...	1.0
1.0	0.0	(26,[23,24],[1.0,1.0])	(26,[23,24],[1.0,1.0])	[1.7332008351982036,-1.73...	[0.8498213844195708,...	0.0
1.0	0.0	(26,[23,24],[1.0,1.0])	(26,[23,24],[1.0,1.0])	[1.7332008351982036,-1.73...	[0.8498213844195708,...	0.0
1.0	0.0	(26,[24],[1.0])	(26,[24],[1.0])	[-2.7013721042120897,2.70...	[0.06289243985992832...	1.0
1.0	0.0	(26,[23,24],[1.0,1.0])	(26,[23,24],[1.0,1.0])	[1.7332008351982036,-1.73...	[0.8498213844195708,...	0.0
1.0	1.0	(26,[23,24,25],[1.0,1.0,1.0])	(26,[23,24,25],[1.0,1.0,1.0])	[2.0811607962331147,-2.08...	[0.8890585783117562,...	0.0
1.0	0.0	(26,[24],[1.0])	(26,[24],[1.0])	[-2.7013721042120897,2.70...	[0.06289243985992832...	1.0
1.0	0.0	(26,[24],[1.0])	(26,[24],[1.0])	[-2.7013721042120897,2.70...	[0.06289243985992832...	1.0

模型评估

- 准确率
- ROC
- F1加权数

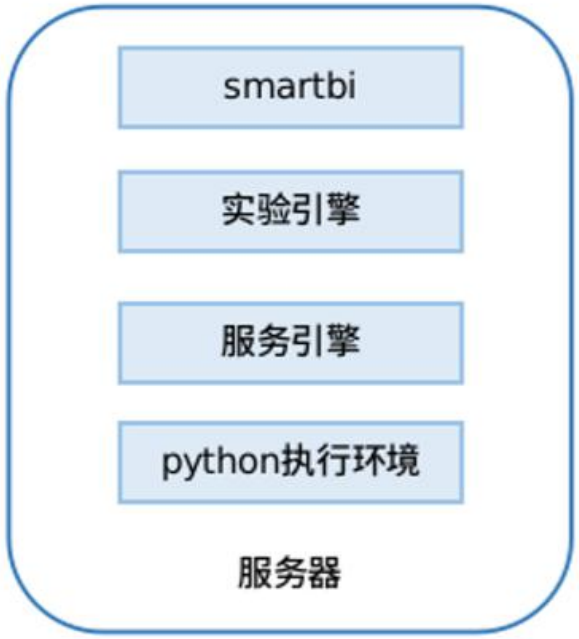
查看分析结果

指标	值
accuracy(准确率)	0.9565217391304348
roc曲线	查看ROC曲线
auc	1.0
ks	查看KS曲线
weighted precision(加权精确率)	0.959866220735786
weighted recall(加权召回率)	0.9565217391304348
weighted F1 score(加权F1分数)	0.9563561076604555
Class 0.0 precision(精确率)	1.0
Class 0.0 recall(召回率)	0.9090909090909091

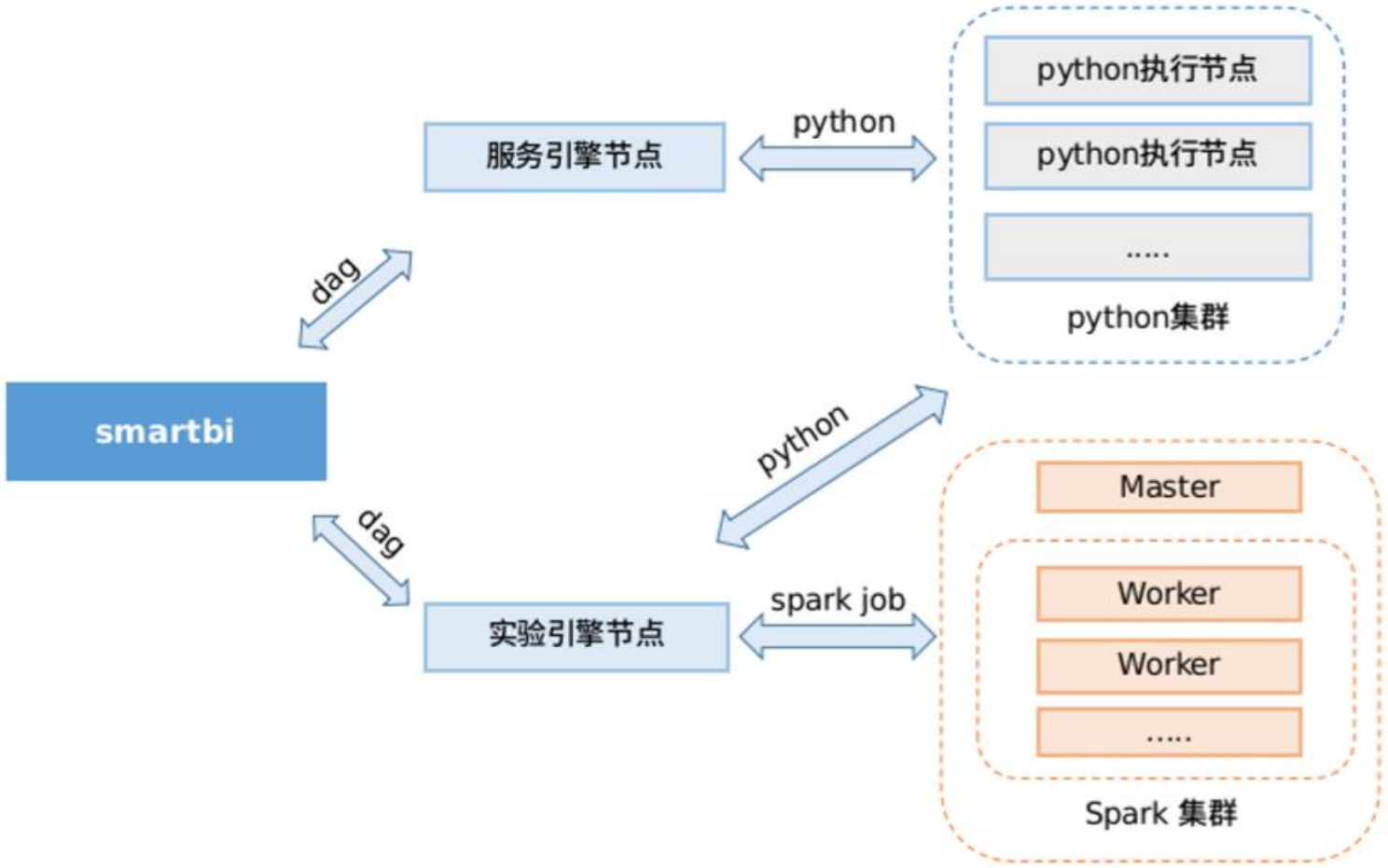


模型部署

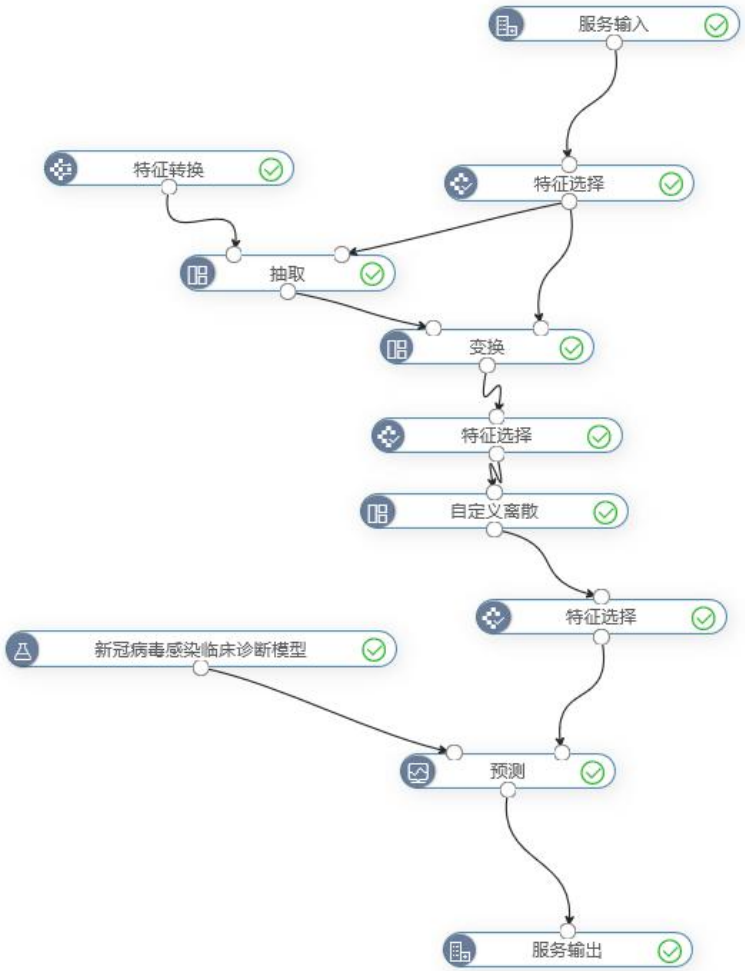
单节点部署



典型部署方式



模型部署



服务ID I4028818901705135135ec00017051c404d20006

服务名称 新冠病毒患者临床确诊模型2

服务别名 新冠病毒患者临床确诊模型2

RESTFUL https://192.168.1.9:8900/api/v1/services/I4028818901705135135ec00017051c404d20006

相关实例 I4028818901705135135ec000170524866ae0014

当前显示 1 条 / 总共有 1 条数据 提示:点击单元格可查看超出的内容

病患号	疫区或病患社区旅行史	病患接触史	疫区人员接触史	是否有聚集性活动	体温	乏力	干咳	发病天数	鼻塞	流涕	咽痛	腹泻	呼吸困难
病患125	无	有	无	无	38.5	否	是	4.0	否	是	是	否	否

当前显示 1 条 / 总共有 1 条数据 提示:点击单元格可查看超出的内容

温Buckerizer	发病天数Buckerizer	features	featuresNormalized	rawPrediction	probability	prediction
1.0	1.0	(26,[24,25],[1.0,1.0])	(26,[24,25],[1.0,1.0])	[-2.337561815884919,2.337...	[0.08805951674106194...	1.0

04工具支撑

SMARTBI 思迈特软件

全流程支持


定制管理
服务监控
11
全部
11
运行中
0
异常
0
已下线
立即刷新
10
秒
搜索
名称
描述
状态
操作
aaabbb
运行中
服务发布测试1
运行中
服务测试
测试服务,参数：a,b。计算：a+b
运行中
银行客户流失服务发布
运行中
场景三-服务-删除
运行中
abcde
abcde
运行中
服务测试
运行中
服务发布
运行中
部署服务

全可视化操作

拖拉拽完成建模

挖掘过程可视化

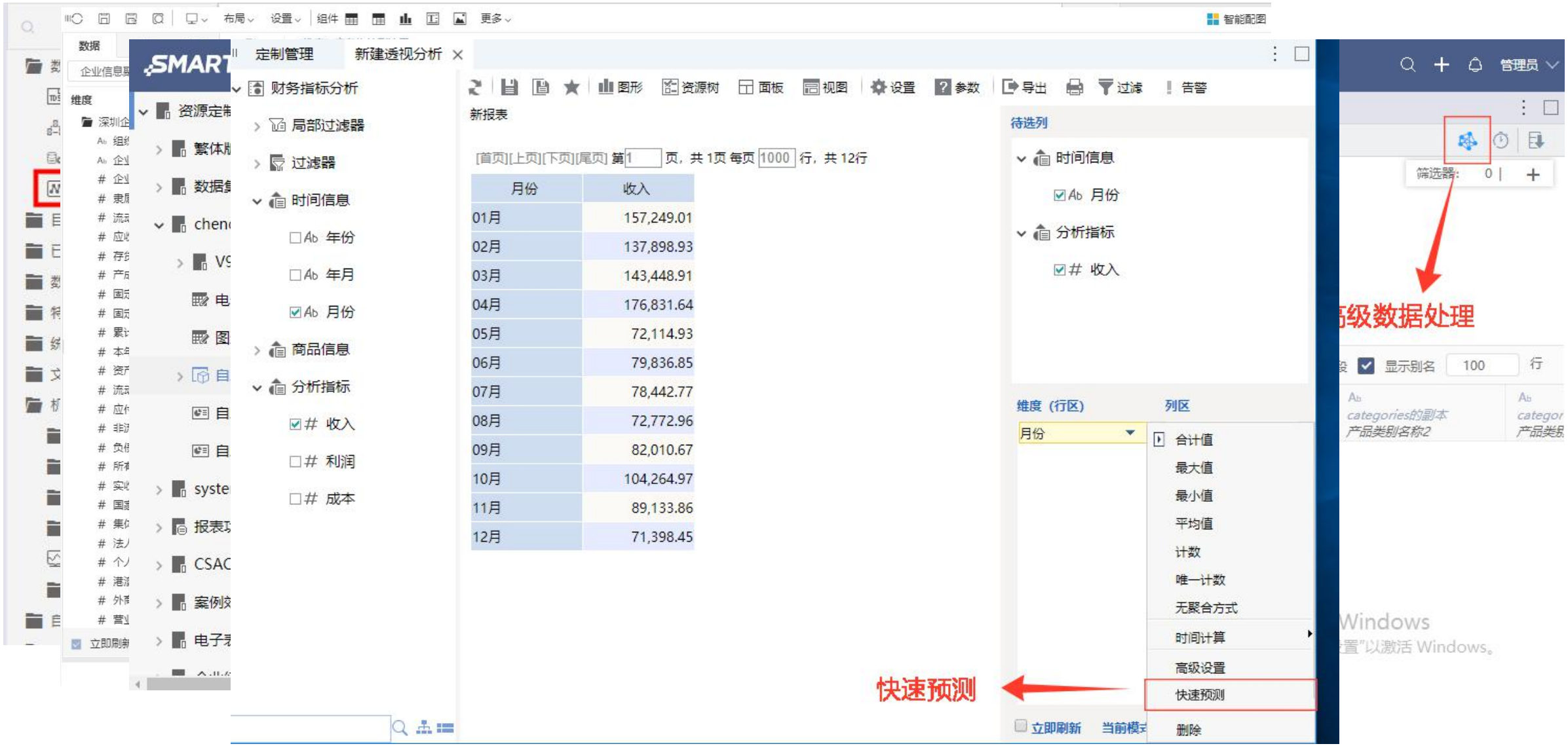
挖掘结果可视化

业务人员高度参与

模型参数智能、自动推荐


实验管理
模型管理
Demo
运行成功：50秒
参数
属性
帮助
数据源
目标源
已训练模型
数据预处理
特征工程
统计分析
文本分析
机器学习
分类算法
二分类算法
支持向量机
梯度提升决策树
多分类算法
逻辑回归
朴素贝叶斯
决策树
随机森林
回归算法
示例数据源
自动调参设置
* 拆分比例
0.7
评估标准
accuracy
参数
范围
elasticNet
0
1
regParam
0
100
maxIter
50
500
tol
0.00000001
0.000001
模型参数自动调整/推荐
取消
确定
评估
归一化
方法选择
StandardScaler
参数
单位标准差归一化
平均数据中心化
自动调参设置
自动调参设置
启用自动调参
最大迭代数 范围是>=0的正整数
40
混合参数 范围是[0,1]的数
0
正则参数 范围是>=0的数
0
收敛阈值 范围是>=0的数
0.000001
分类阈值(请用英文逗号隔开,且数量与分类数相同)

挖掘与BI无缝对接


定制管理
新建透视分析
财务指标分析
局部过滤器
过滤器
时间信息
年份
年月
月份
商品信息
分析指标
收入
利润
成本
图形
资源树
面板
视图
设置
参数
导出
过滤
告警
新报表
[首页][上页][下页][尾页] 第1 页，共 1页 每页 1000 行，共 12行
月份 收入
01月 157,249.01
02月 137,898.93
03月 143,448.91
04月 176,831.64
05月 72,114.93
06月 79,836.85
07月 78,442.77
08月 72,772.96
09月 82,010.67
10月 104,264.97
11月 89,133.86
12月 71,398.45
待选列
时间信息
月份
分析指标
收入
维度（行区）
列区
月份
合计值
最大值
最小值
平均值
计数
唯一计数
无聚合方式
时间计算
高级设置
快速预测
删除
快速预测
立即刷新
管理员
筛选器: 0
高级数据处理
显示别名 100 行

联系我们

S 广州思迈特软件有限公司

M 电话：020-85648869

A 北京、上海、武汉等办事处

R http://www.smartbi.com.cn

T sales@smartbi.com.cn