

# 经知識學刀掌語過光

AFAN的金融科技

——其于sklegrn的代码实践

24/05/14 20:00 UTC+B

直播进入/回放见星球>



器直员会郑星职限

会员权益如下,快来加入吧:

- 、每月至少1次的线上群体直播交流
- 2、不定期的金融科技专业话题分享

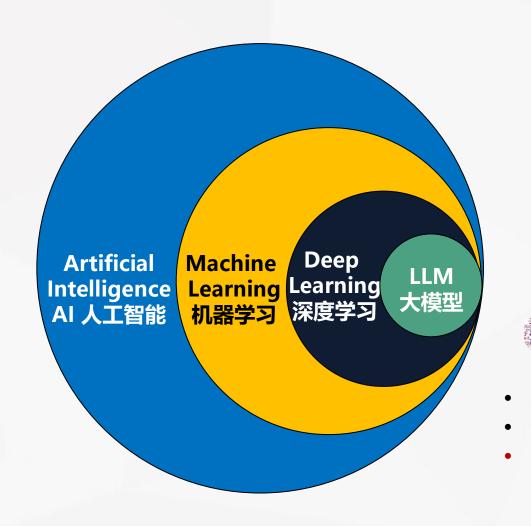


微信扫码加入星球



# AI技术发展回顾

➤ AI基础——AI和机器学习



疾病 诊断 给题目 有监督学习 给答案 Supervised 预测结果 三类 主题聚类 围棋 AlphaGo 博弈 Learning 给题目 给题目 无监督学习 强化学习 不给答案 给环境 Unsupervised Reinforcement 探索规律 学习策略



# AI技术发展回顾

#### > AI基础——有监督学习和AI数据场景

特征 Feature

标签 Label

Close 收盘价

Volume 成交量 Price 明日收盘价

Direction 明日涨/跌

Style 风格类型 分类任务

回归任务

Regression

Classification

聚类任务 Clustering 有监督

无监督

日期	股票编号	收盘价	成交量	明日收盘价	明日涨/跌
2024/1/2	Α	30	100	→ 33	→涨
2024/1/3	А	33	110	32	跌
2024/1/4	Α	32	105	35	→涨
2024/1/5	А	35	115	空	空

忽略 Ignore

特征 Feature

标签 Label

图像数据: Image/Video Computer Vision 计算机视觉

文本数据: Text Nature Language Process 自然语言处理

音频数据: Speech
Automatic Speech Recognition
自动语音识别

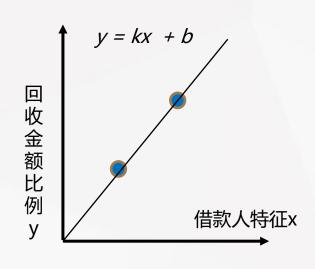
表格数据: Table
Structure Data Model
结构化数据建模

图网络数据: Graph Graph Computing 图计算

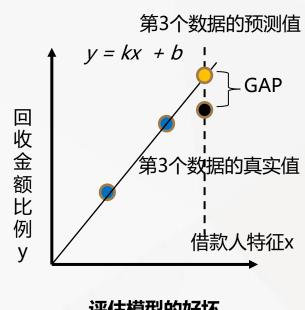


# 基础机器学习算法

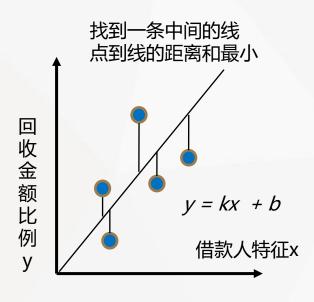
▶ 机器学习算法——最简单的线性回归



**构建线性模型** 基于2个数据点



评估模型的好坏预测和真实对比

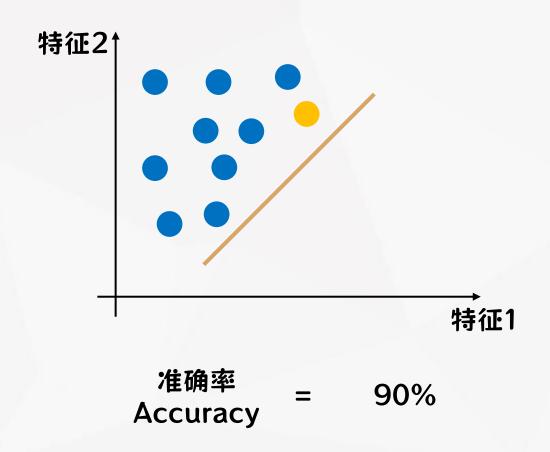


**线性回归模型** 有较多数据点时

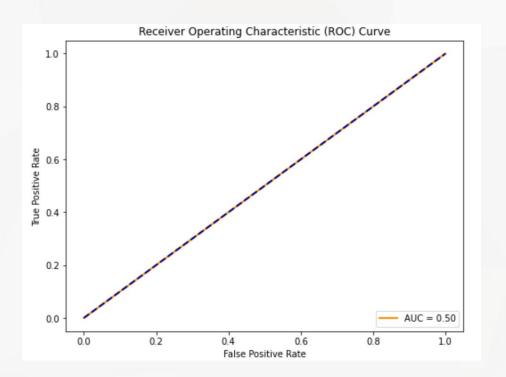


# 机器学习的常见任务

> 分类标签不均衡的评价指标



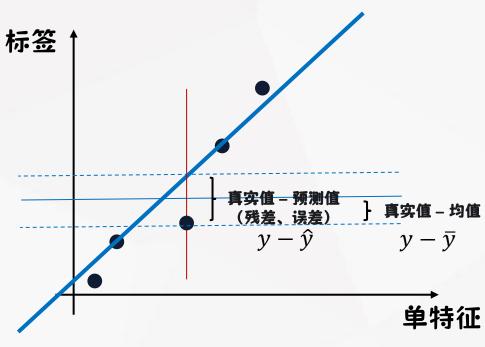
AUC = 0.5





# 机器学习的常见任务

> 回归任务及其评价指标



回归任务 Regression

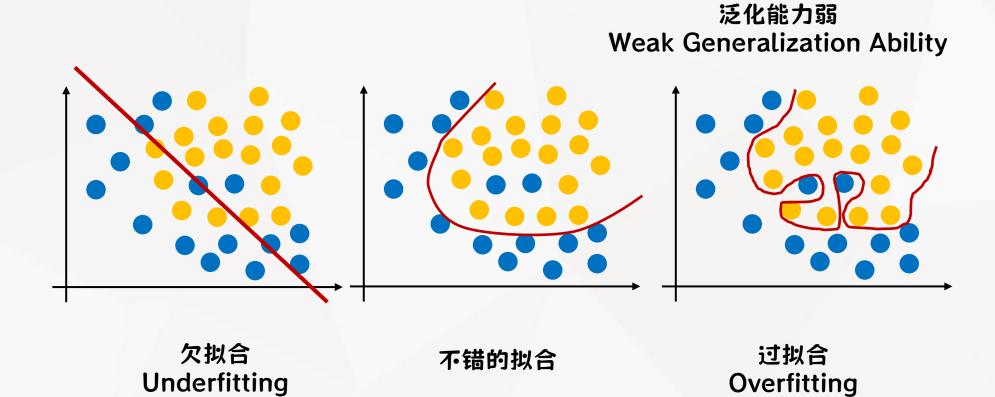
R-squared

 $(-\infty, 1)$ 



# 机器学习的常见问题

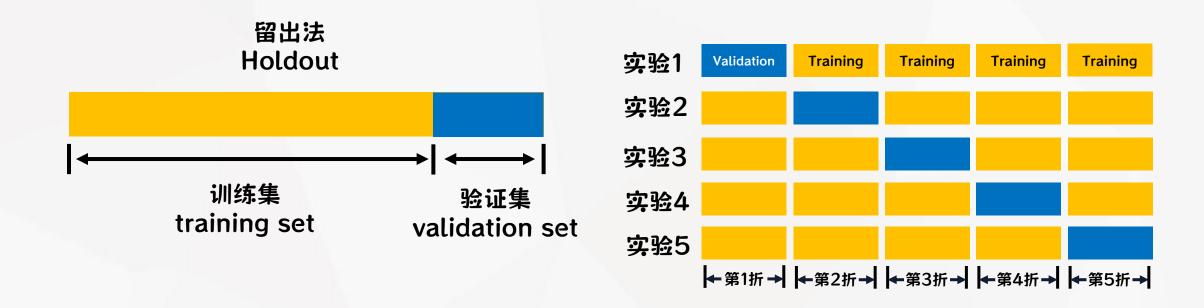
> 过拟合和模型泛化





# 机器学习的常见问题

> 过拟合和模型泛化

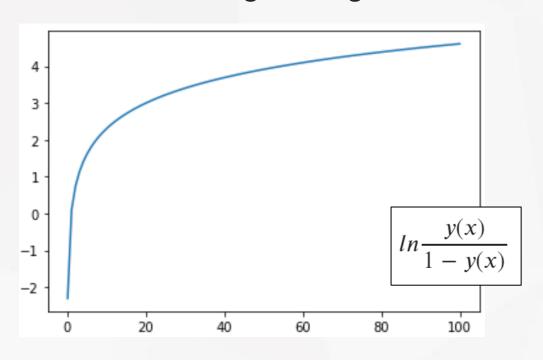


验证集分割 Validation Split K折交叉验证 K-fold Cross Validation



# 基础机器学习算法

#### > 逻辑回归算法(Logistic Regression,分类算法)



Sigmoid Function

1.0 Sigmoid Function

0.8 O.6 O.6 O.2 Sigmoid Function

0.2 Sigmoid Function

0.4 O.2 Sigmoid Function

1.0 Sigmoid Function

1.1 Sigmoid Function

1.2 Sigmoid Function

1.3 Sigmoid Function

1.4 Sigmoid Function

1.5 Sigmoid Function

1.6 Sigmoid Function

1.7 Sigmoid Function

1.8 Sigmoid Function

1.9 Sigmoid Function

1.0 Sigmo

扩充定义域

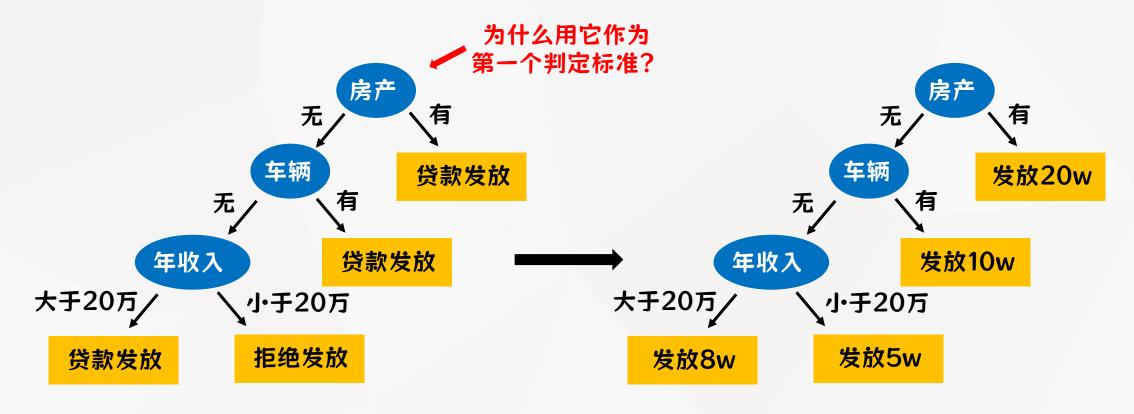
关联回归方程

$$ln\frac{y(x)}{1-y(x)} = ln\frac{\frac{1}{1+e^{-w^{T}x}}}{1-\frac{1}{1+e^{-w^{T}x}}} = ln(e^{w^{T}x}) = w^{T}x$$



# 基础机器学习算法

> 决策/回归树算法 (Decision/Regression Tree, 分类/回归算法)



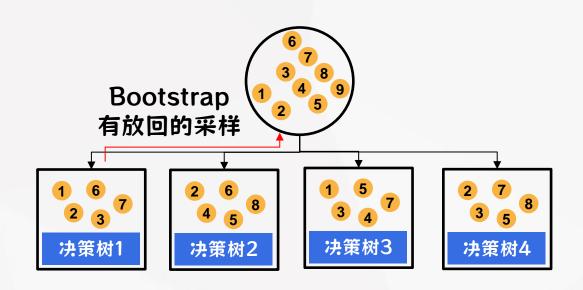
决策树符合人类的决策思路

回归树的预测结果是连续数值



# 集成机器学习算法

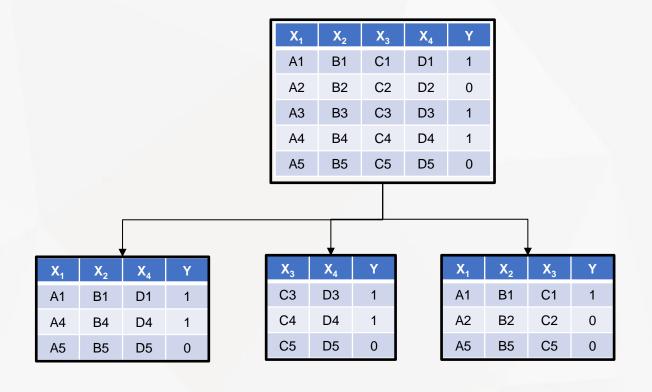
➤ 随机森林——Bagging类算法



9 Out of Bag Data 袋外数据

$$1 - (1 - \frac{1}{m})^n$$
$$1 - \frac{1}{e} = 63.2\%$$

随机1: 单棵树的训练数据集随机选取

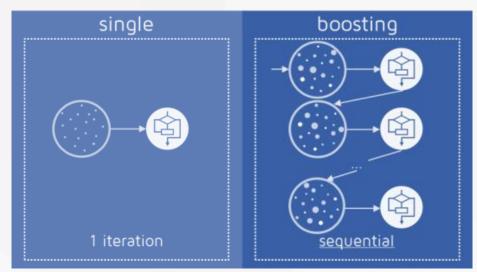


随机2: 单棵树所需特征随机选取



# 集成机器学习算法

➤ GBDT (梯度提升树, Gradient Boosting Decision Tree) ——Boosting类算法



y − ŷ 残差

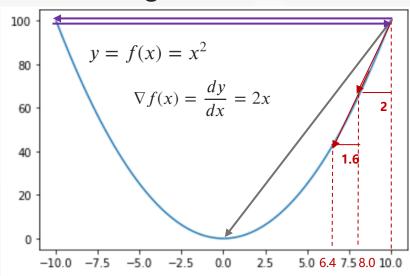
分类任务

回归任务

概率残差

数值残差

基学习器串行训练,不断优化残差目标



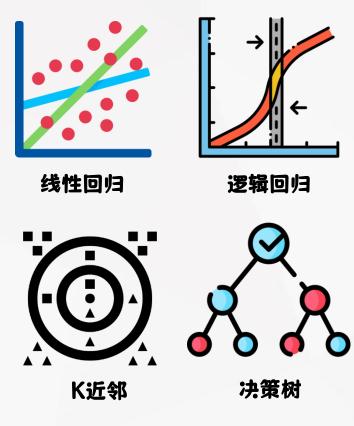
$x[-\alpha]$	$\nabla f(x)$	f(x)	$\nabla f(x)$	$\alpha * \nabla f(x)$	)
	10	100	20	2	$\alpha = 0.1$
	8	64	16	1.6	$\alpha = 0.1$
	10	100	20	10	0 F
	0	0	0	0	$\alpha = 0.5$
	10	100	20	20	
	-10	100	-20	-20	$\alpha = 1.0$
	10	100	20	20	
			·		

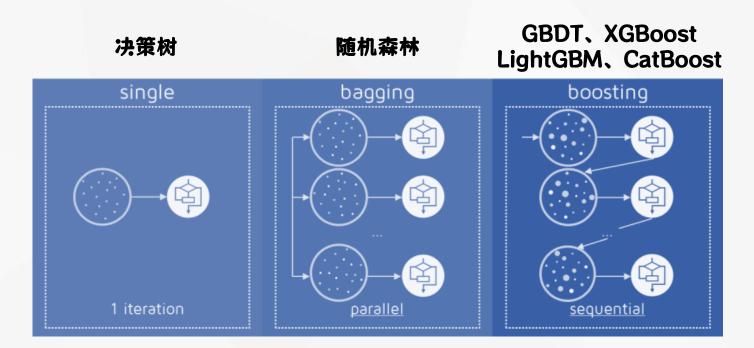
梯度下降和学习率参数



# 经典机器学习算法概览

> 基础+集成机器学习算法





基础机器学习算法

集成机器学习算法



# 机器学习代码实践的核心步骤

> Sklearn调用方式

分割训练集 train\_test\_split

fit训练 Algo.fit(x\_train, y\_train)

分类概率: Algo.predict\_proba(x\_test)

测试集打分: Algo.score(x\_test, y\_test)

训练集打分: Algo.score(x\_train, y\_train)



# 金融中的应用场景概述

➢ 金融场景 + AI技术 = ?

在金融场景中,往往需要先通过运营手段获取客户,然后根据客户的需求,可以通过信贷业务对客户发放贷款赚取利息收入,也可以向客户发售理财产品通过投资业务赚取投资收益,并且全程需要保证合规性,减少欺诈和洗钱带来的不良影响。这整个过程中都可以用AI进行场景赋能。

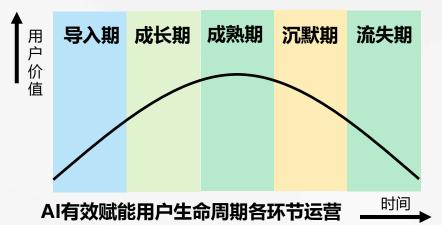


# 金融产品销售场景

#### > 客户价值营销模型

业务介绍:在金融产品销售中,对客户全 生命周期的营销和维护至关重要, 如果能 精准推荐提升用户对产品的购买意愿,并 很好的维护住用户,将产生巨大商业价值。

技术介绍:可以整合客户的基本信息、银 行交易流水、行为数据、产品数据等形成 一张特征宽表, 再构建机器学习模型实现 精准营销,并提前预判即将流失的客户进 行挽回。



#### 客户基本信息

- ◆年龄性别
- ◆籍贯住所
- ◆职业单位

#### 银行交易流水

- ◆IP电话
- ◆交易日期
- ◆交易金额
- ◆对方账户

#### 用户行为数据

- **◆App埋点**
- ◆借贷行为
- ◆产品购买

#### 外部数据

- ◆购物消费
- ◆旅游消费
- ◆其他行为分

#### 产品数据

- ◆投资产品
- ◆信贷产品
- ◆保险产品

#### 客服交互数据

- ◆用户咨询
- ◆用户投诉
- ◆短信促销
- ◆AI语音营销

#### 产品推荐

预测该类 用户最可 能购买的 产品类别。 制定对应 政策讲行 营销

#### 流失预测

提前预测 哪些用户 可能流失 采取针对 策略进行 挽回

相关建模特征

# 消费金融信贷业务场景

> 贷前风控授信放款模型

**业务介绍**:在消费金融类的信贷业务场景中,贷前的风控授信模型是最重要的模型,该模型的预测效果的好坏将直接影响到贷款的回收效果。

技术介绍:利用用户基本信息、行为数据、历史借贷以及外部数据等建模特征,定义是否授信以及放款额度标签,可以通过历史数据构建更加科学的授信放款机器学习模型,降低过审用户的违约率,减少公司的坏账损失。

# 客 准 授 还 贷 坏 户 入 信 中 款 后 坏

监

运

#### 用户基本信息

- ◆基本情况
- ◆家庭情况
- ◆工作情况
- ◆收入情况
- ◆自由资产 ◆借贷信息
- ◆公共服务
- ◆地址信息

#### 用户行为数据

策

◆App埋点

投

- ◆服务端埋点
- ◆网页浏览
- ◆消费行为

#### ID关联数据

- ◆身份证
- ◆手机号
- ◆设备
- ◆银行卡

#### 外部数据

- ◆多头借贷
- ◆电商消费
- ◆人行征信
- ◆其他行为分

#### 历史借贷数据

- ◆历史申请
- ◆历史放款
- ◆历史还款
- ◆联系人关联

#### 是否授信

#### 放款额度

定

对已经授 信的用户 决定放款 的额度

#### 相关建模特征

# 消费金融信贷业务场景

> 贷中用户运营相关模型

业务介绍: 在消费金融类的信贷业务场景 中,已经放款的用户如果提前清贷,那利 息收入就会减少;此外,当前行业获客成 本较高, 老客复贷是非常的优质案源, 所 以需要防范用户流失,以及唤醒沉睡用户。

技术介绍:除利用贷前授信放款模型中的 相关数据外,结合贷中的行为数据和相关 数据变化,可以构建精细化运营模型,提 升用户的LTV (生命周期价值) , 并挽回 更多的老客户,提升公司复贷业务规模。



投 放

准入 客户 策

贷 授 运 营

贷 后 还 款 监

定

贷前 授信 放款 建模 相关 数据

#### 贷前变化数据

- ◆联系人信息
- ◆映射表关系
- ◆GPS位置
- ◆设备指纹

#### 历史借贷数据

- ◆历史申请
- ◆历史放款
- ◆历史还款

#### 贷中行为数据

- ◆App埋点
- ◆服务端埋点
- ◆网页浏览
- ◆消费行为

#### 未完结订单

- ◆身份证
- ◆手机号
- ◆设备
- ◆银行卡

#### 提前清贷

从而最大

化利息收

提前预测 提前预测 哪些用户 哪些用户 可能流失 可能有提 前结清贷 采取针对 款的可能 策略进行 对客户推 挽回 出活动。

# 沉睡唤醒

流失预测

预测哪些 用户最可

相关建模特征

# 消费金融信贷业务场景

> 贷后用户催收投诉模型

**业务介绍**:在消费金融类的信贷业务场景中,如果客户逾期将会进行贷后的催收,但是由于人力资源有限,往往会选择将一部分难催的案件进行委外催收,并且催收不当如果导致客户投诉也会有舆论风险。

**技术介绍**:除利用贷前贷中相关数据之外,还可以使用贷后跟进的相关记录数据,构建模型来预测哪些案件的还款概率更大,以此更精准地将容易案件内催、难催案件委外,有效提升回收率;并能构建投诉模型,减少舆论风险和监管压力。

准入 贷 后 贷 客 授 还款 策 定 运 投 申 监 营 放 略 控

贷授放建相数

贷用运相数据

#### 贷后跟进记录

- ◆贷后提醒
- ◆AI语音催收
- ◆人工催收
- ◆投诉记录
- ◆客服记录 ◆委外记录
- ◆监管反馈

#### 还款概率

预测哪些 用分子 以对 用一些 的 化 以 用 一 些 收 方 的 化 收 方

投诉概率

相关建模特征

可选建模标签

法

# 财富管理投资场景

> 股票量化投资模型

**业务介绍**:在股票投资的经典量化策略中,量化研究员需要进行因子的挖掘和组合优化,如果是中低频就对未来T日预测,如果是中高频就对未来几秒几分钟预测,以此投资以期获取收益。

技术介绍:可以接入股票的行情、基本面以及另类数据,构建机器学习模型来预测未来T日或高频收益率,如果具备AutoFE自动化特征工程构造能力,甚至可以跳过因子挖掘步骤,一步到位实现组合优化。

数据接入

因子组合挖掘化

仓位生成

模拟回测

业绩跟踪

模型迭代

#### 股票行情数据

- ◆开高低收
- ◆成交量额
- ◆Tick级行情
- ◆订单簿行情
- ◆龙虎榜数据
- ◆基金持仓
- ◆北向资金
- ◆五档行情
- ◆十档行情
- ◆指数行情

#### 基本面数据

- ◆业绩预报
- ◆资产负债表
- ◆利润表
- ◆现金流量表
- ◆股东权益表

#### 股票另类数据

- ◆分析师预期
- ◆机构调研
- ◆ESG评级
- ◆NLP與情

#### T日收益率

#### 高频收益率

相关建模特征

# 财富管理投资场景

> 期货量化投资模型

业务介绍:在期货投资的经典量化策略中, 量化研究员需要进行因子的挖掘和组合优 化,如果是中低频就对未来T日预测,如 果是中高频就对未来几秒几分钟预测,以 此投资以期获取收益。

技术介绍:可以接入期货的行情、基本面 以及另类数据,构建机器学习模型来预测 未来T日或高频收益率,如果具备AutoFE 自动化特征工程构造能力, 甚至可以跳过 因子挖掘步骤,一步到位实现组合优化。



数 据 接

组 合 因 子 优 掘

仓 位 成

期货另类数据

◆分析师预期

◆NLP與情

回

型选

#### 期货行情数据

- ◆开高低收
- ◆成交量额
- ◆Tick级行情
- ◆五档行情
- ◆指数行情
- ◆会员持仓

#### 基本面数据

- ◆现货数据
- ◆库存数据
- ◆利润数据
- ◆仓单数据
- ◆期限结构

#### T日收益率

实

单

股票T日 的收益率, 如果是多 头策略就 靠前的股 票即可

预测每只

#### 高频收益率

预测日间 未来几秒 或者几分 钟的股票 价格,进 行高频的 交易投资

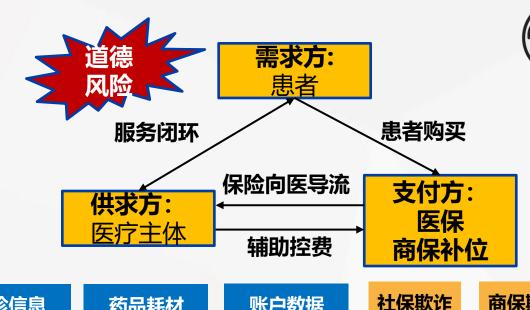
相关建模特征

# 金融合规场景

健康保险理赔反欺诈模型

业务介绍:健康保险由于保险标的和事故 的特殊性, 在医疗行业信息不对称的情况 下,会因为骗赔型的道德风险而产生保险 欺诈,造成保险公司和社保基金的损失。

技术介绍:可以整合理赔人员的就诊信息、 药品耗材以及账户数据构建机器学习模型 来预测是否有社保或医保的理赔欺诈,如 果具备AutoFE自动化特征工程构造能力, 甚至可以特征的多层构造,产生能够描述 出欺诈人员采用的欺诈模式的特征。



#### 就诊信息

- ◆就诊日期
- ◆就诊次数
- ◆床位费用
- ◆床位费用
- ◆手术费用
- ◆检查费用 ◆发生金额
- ◆申报金额
- ◆自费金额

#### 药品耗材

- ◆一次性材料
- ◆中成药费用
- ◆中草药费用
- ◆医用材料
- ◆贵州药品
- ◆高价材料
- ◆成分输血
- ◆发生金额 ◆申报金额
- ◆自费金额

#### 账户数据

- ◆公务员补助
- ◆医疗救助 ◆城乡优抚
- ◆城乡补助
- ◆个人账户
- ◆统筹基金
- ◆起付线金额
- ◆起付线自付 ◆非账户支付

# 预测该用

户是否欺 诈,并寻 找最能解 释其欺诈 行为的相 关特征

#### 商保欺诈

预测该用 户是否欺 诈,并寻 找最能解 释其欺诈 行为的相 关特征

相关建模特征

# 金融合规场景

> 银行交易反洗钱模型

**业务介绍**: 反洗钱是保护金融系统免受滥用的关键应用,有助于减少犯罪和保护消费者。很多金融机构也承担着反洗钱的职责。

技术介绍:可以整合客户信息、交易信息、 图谱网络信息等进行机器学习建模,如果 具备AutoFE自动化特征工程构造能力,甚 至可以特征的多层构造,产生能够描述出 洗钱规则的关键特征。

#### 上游涉罪

#### 洗钱通道

#### 下游账户

- ◆ 网络赌博 ◆
  - ◆ 非法传销
  - 腐败资金 ◆ 非法炒汇 毒品买卖 ◆ 非法集资
- ▶ 电信诈骗 ◆ 涉税资金
- ◆ 地下钱庄
- ◆ 珠宝拍卖
- ◆ 虚假投资
- ◆ 网银支付

- ◆ 个人账户
- ◆ 对公账户

#### 客户信息

- ◆年龄性别
- ◆籍贯住所
- ◆职业单位
- ◆指纹信息
- ◆人像信息
- ◆公司信息
- ◆行为序列 ◆其他KYC

- 交易信息
- ◆IP电话
- ◆交易日期
- ◆交易金额
- ◆对方账户
- ◆交易频次
- ◆大额交易
- ◆可疑交易 ◆沉默账户

#### 图谱网络

- ◆交易节点
- ◆交易入度 ◆交易出度
- ◆关联信息

#### 个人洗钱

预测该用 户是否涉 及到洗钱 操作当中

### 团伙发现

预测是否 属于某个 洗钱团伙

相关建模特征