AFAN的金融科技

经典机器学习算法概览及金融场景应用介绍

——基于sklearn的代码实践

24/06/14 20:00 UTC+8

直播进入/回放见星球→

June

知识星球会员直播

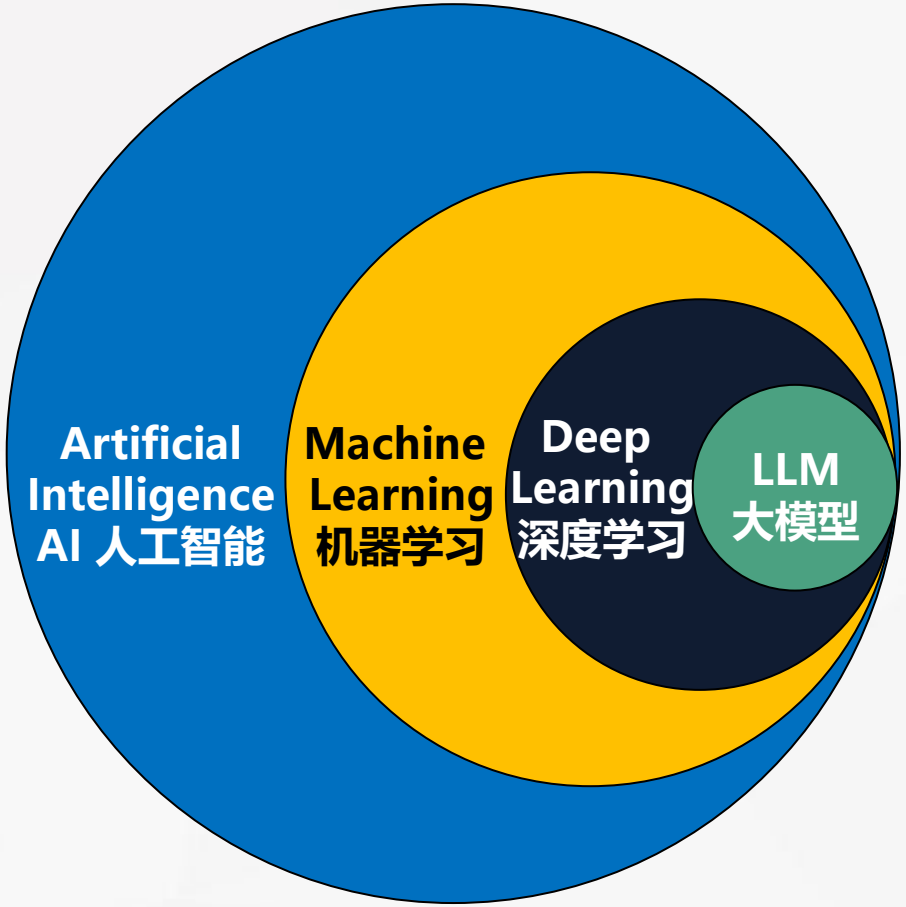
会员权益如下，快来加入吧：
1、每月至少1次的线上群体直播交流
2、不定期的金融科技专业话题分享

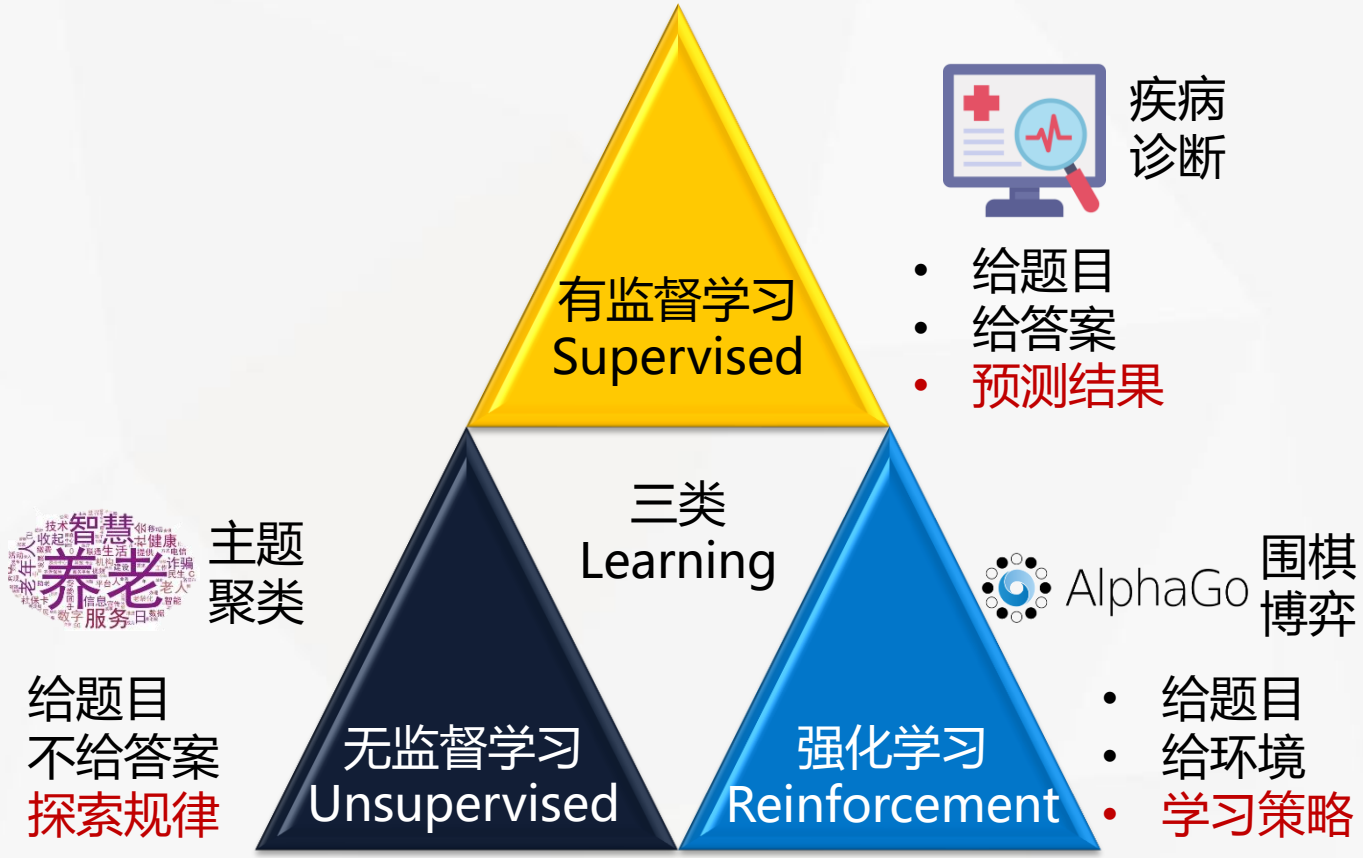
AFAN的金融科技

微信扫码加入星球

AI技术发展回顾

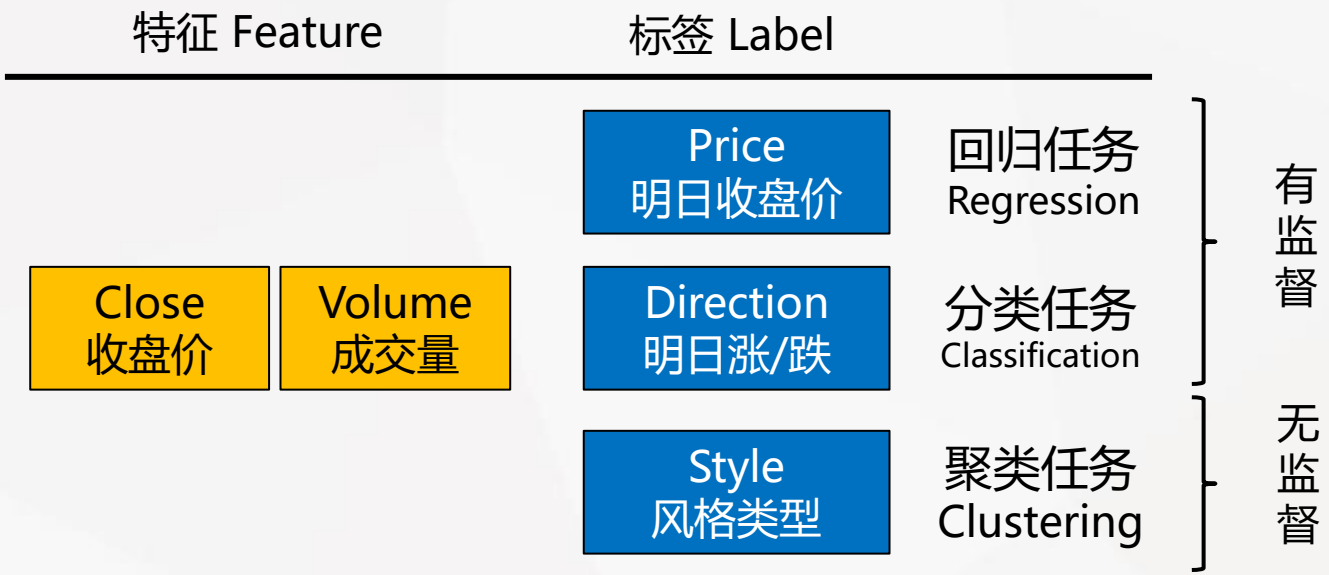
- AI基础——AI和机器学习





AI技术发展回顾

- **AI基础——有监督学习和AI数据场景**



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

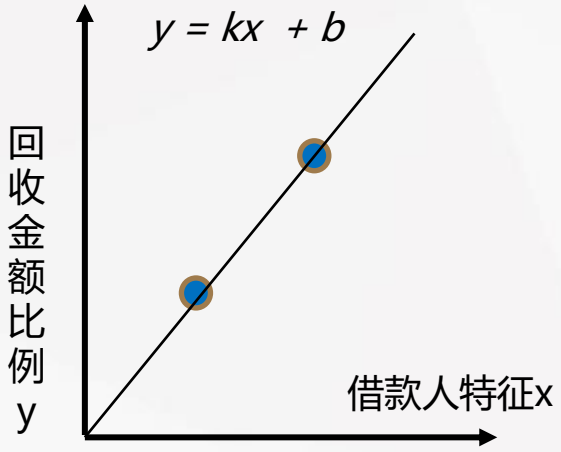
音频数据：Speech
Automatic Speech Recognition
自动语音识别

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

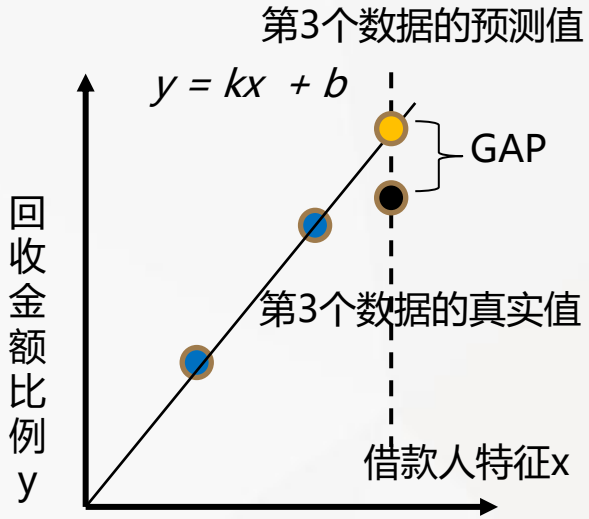
图网络数据：Graph
Graph Computing
图计算

基础机器学习算法

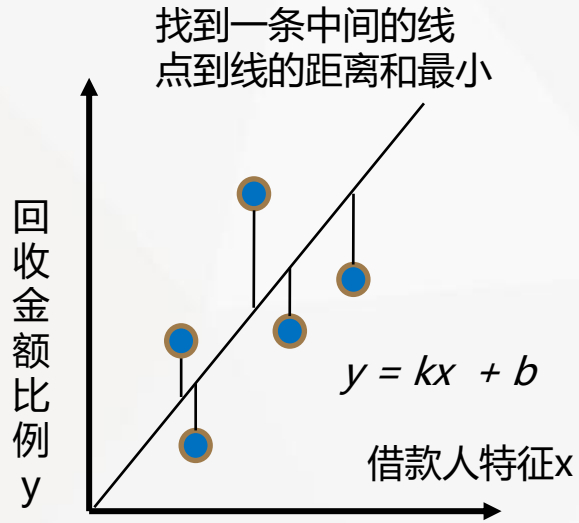
- 机器学习算法——最简单的线性回归



构建线性模型
基于2个数据点



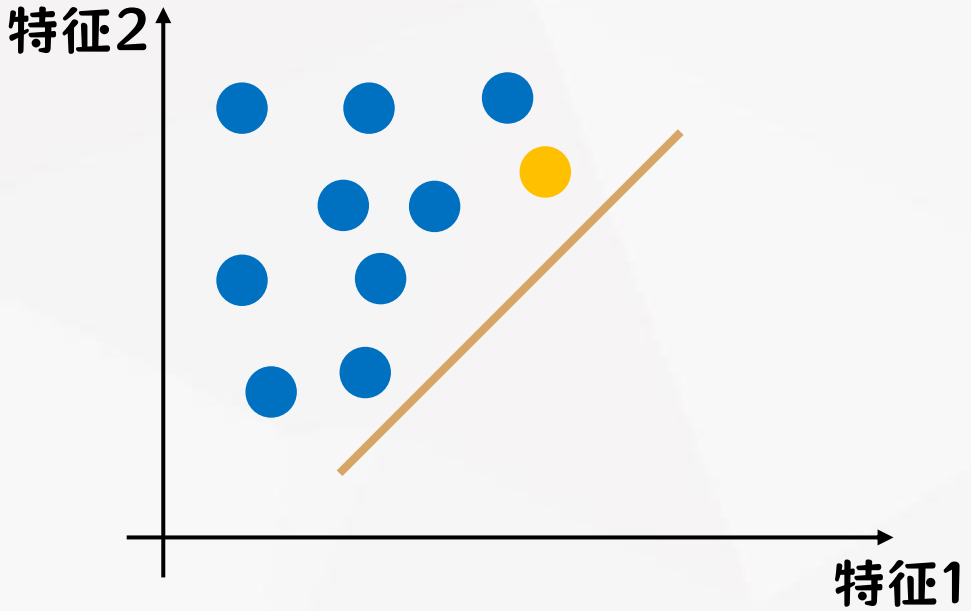
评估模型的好坏
预测和真实对比



线性回归模型
有较多数据点时

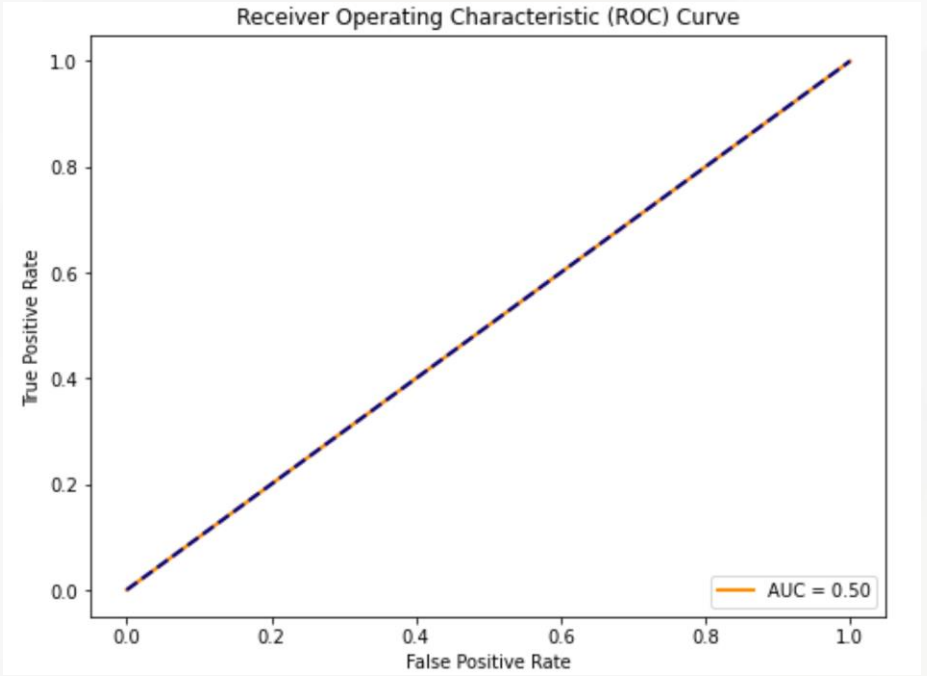
机器学习的常见任务

- 分类标签不均衡的评价指标



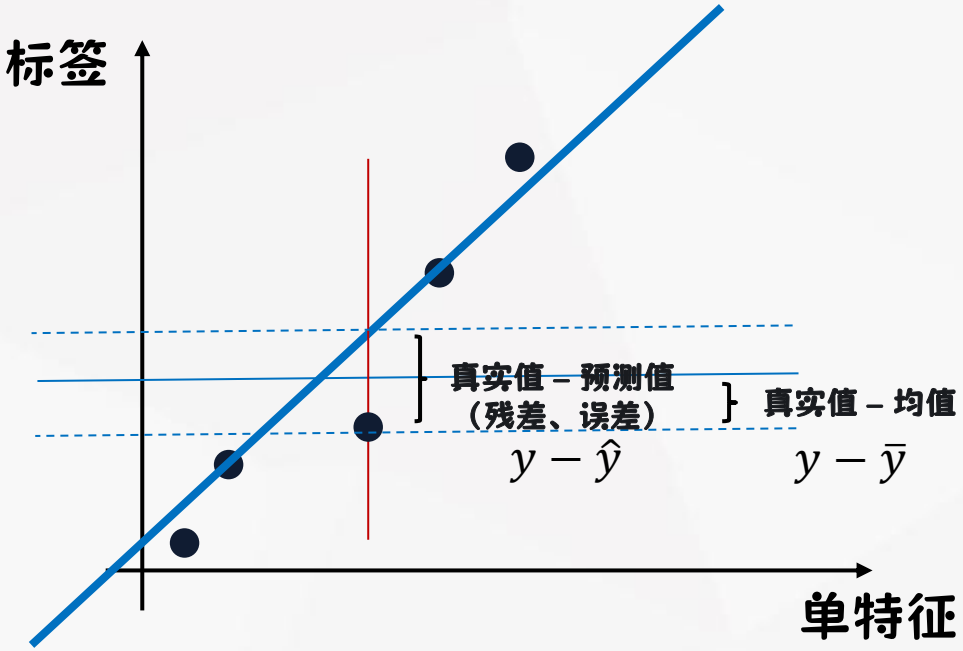
准确率 Accuracy = 90%

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

MSE 均方误差
Mean Square Error

$$= \frac{\sum(y - \hat{y})^2}{n}$$

RMSE 均方根误差
Root Mean Square Error

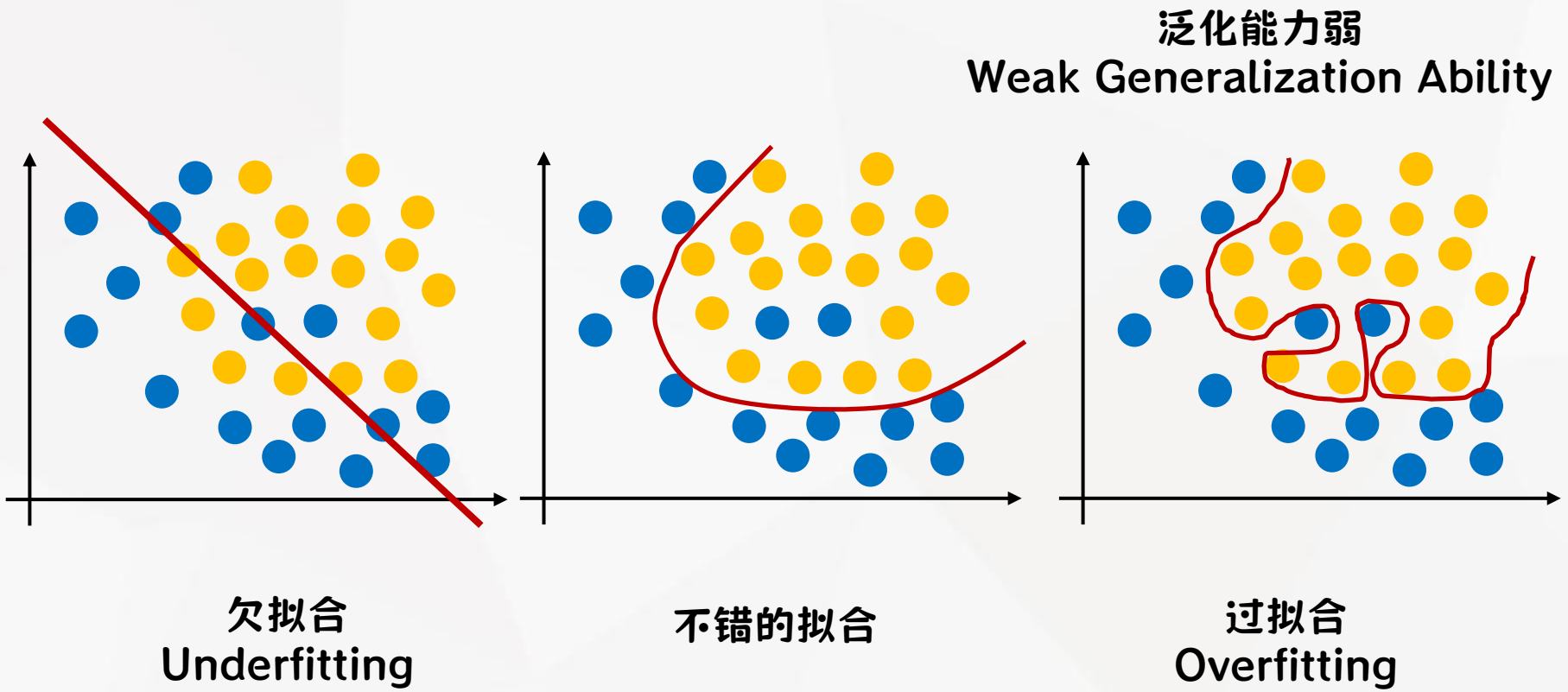
$$= \sqrt{MSE}$$

R方
R-squared
$(-\infty, 1)$

$$= 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

机器学习的常见问题

- 过拟合和模型泛化

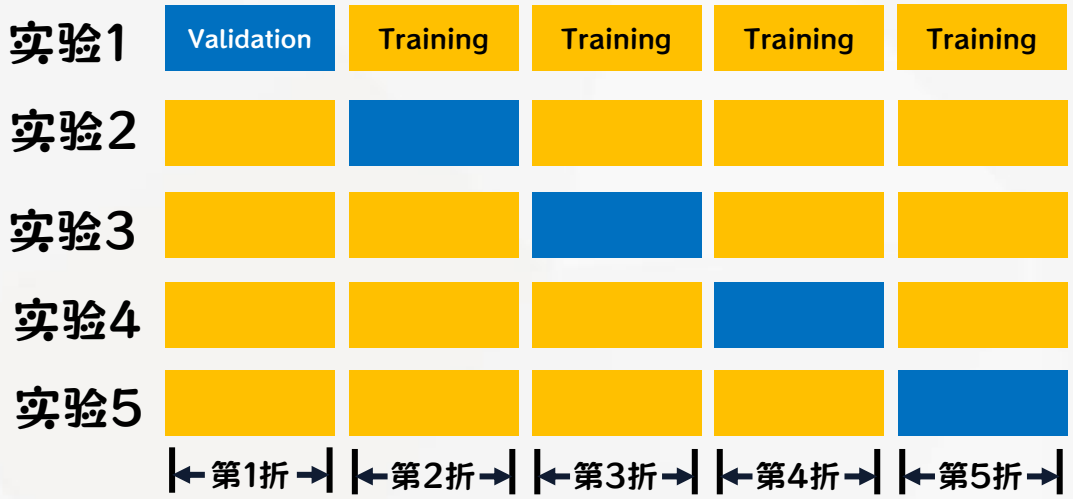


机器学习的常见问题

- 过拟合和模型泛化

留出法
Holdout

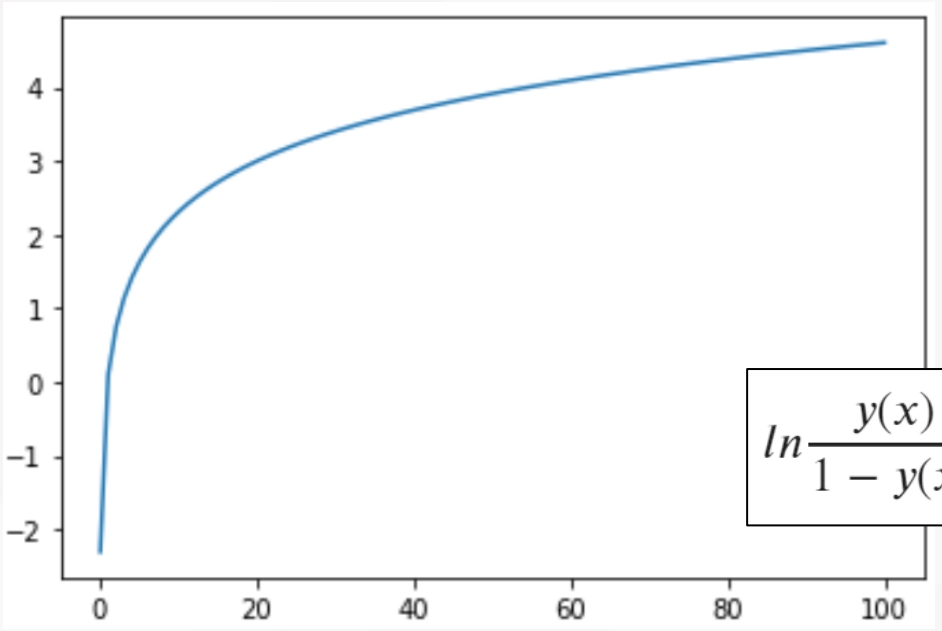
验证集分割
Validation Split



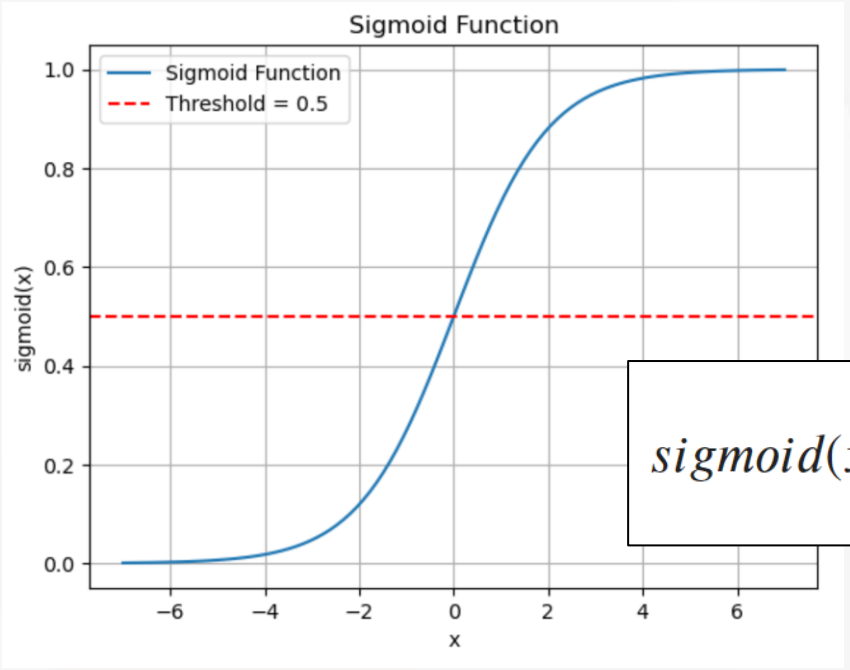
K折交叉验证
K-fold Cross Validation

基础机器学习算法

- 逻辑回归算法（Logistic Regression，分类算法）



$$ln\frac{y(x)}{1-y(x)}$$



$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

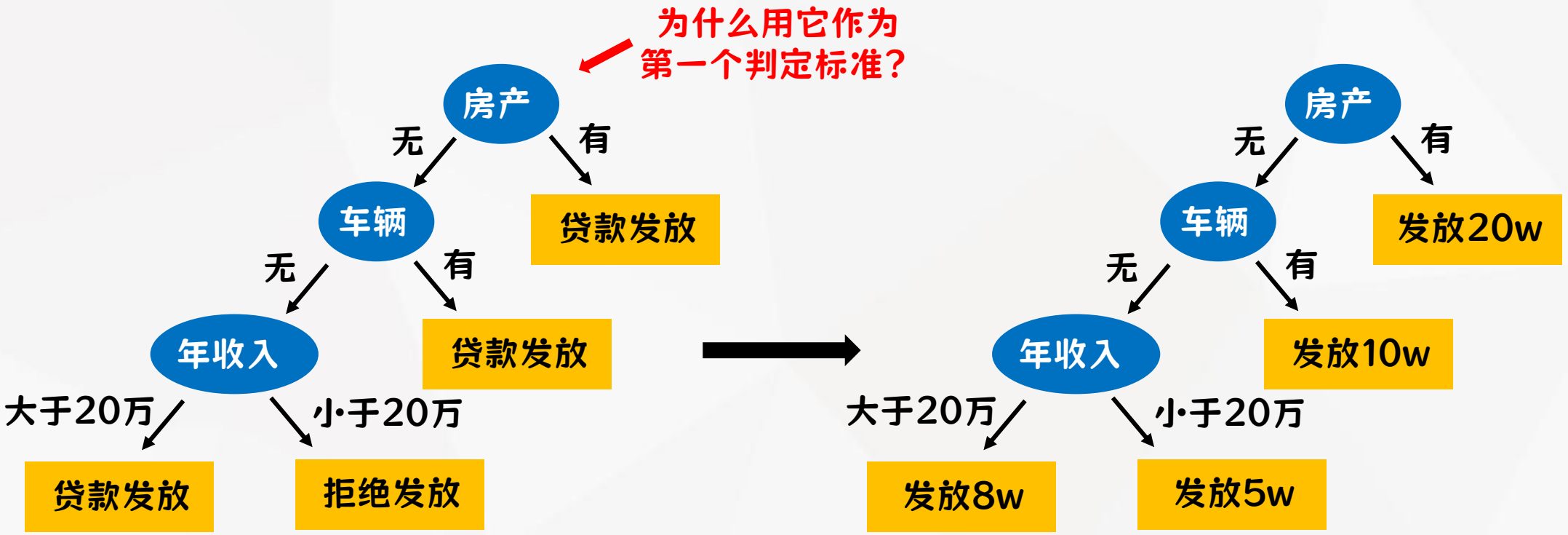
扩充定义域

关联回归方程

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

基础机器学习算法

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

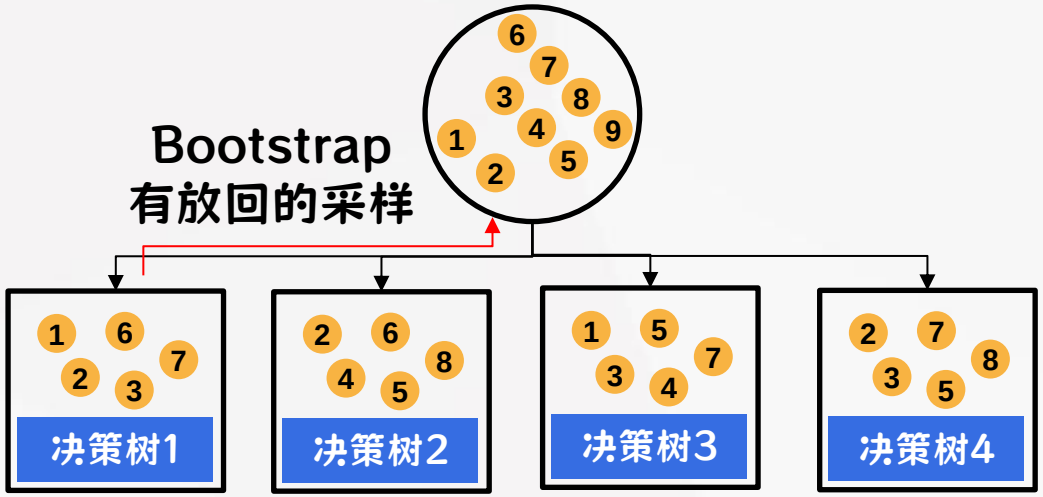


决策树符合人类的决策思路

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



- 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

X_1	X_2	X_3	X_4	Y
A1	B1	C1	D1	1
A2	B2	C2	D2	0
A3	B3	C3	D3	1
A4	B4	C4	D4	1
A5	B5	C5	D5	0

X_1	X_2	X_4	Y
A1	B1	D1	1
A4	B4	D4	1
A5	B5	D5	0

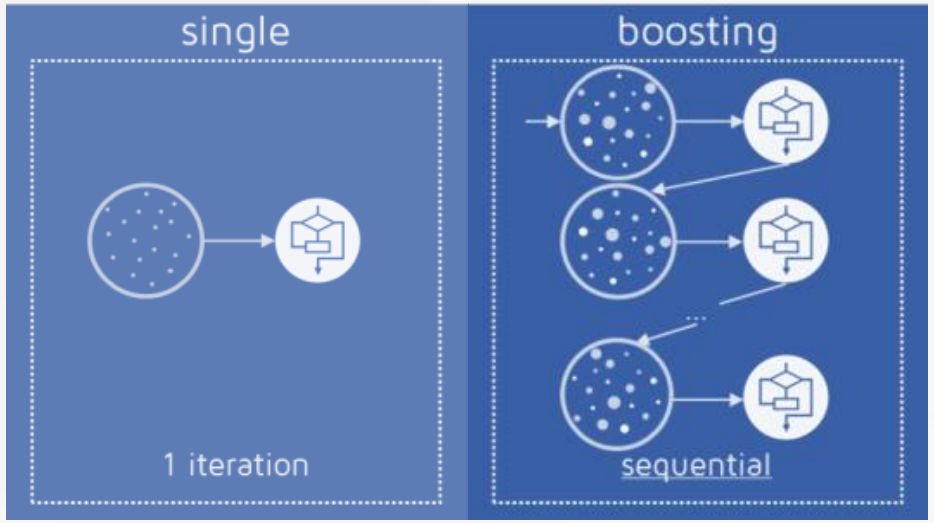
X_3	X_4	Y
C3	D3	1
C4	D4	1
C5	D5	0

X_1	X_2	X_3	Y
A1	B1	C1	1
A2	B2	C2	0
A5	B5	C5	0

随机2：单棵树所需特征随机选取

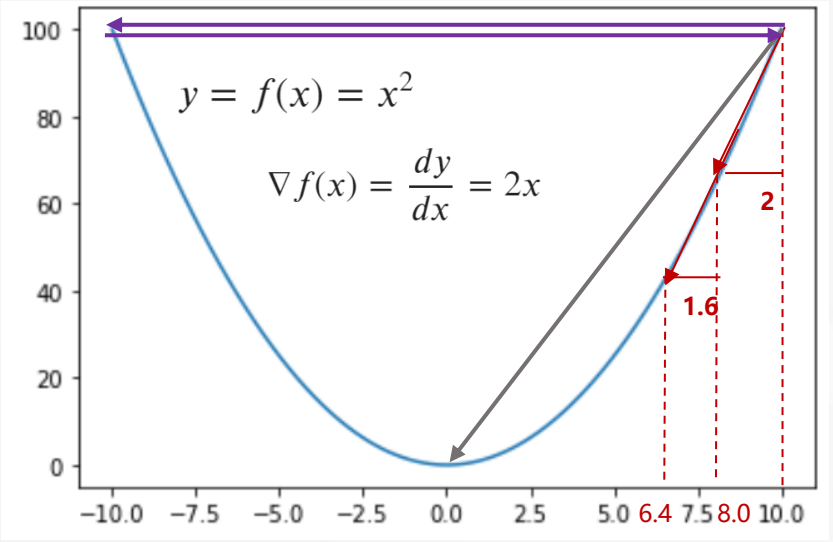
集成机器学习算法

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



$y - \hat{y}$

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

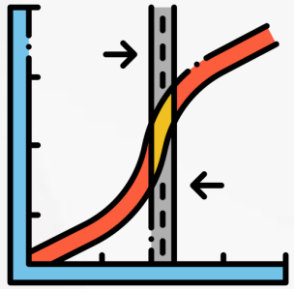
$x[-\alpha * \nabla f(x)]$	$f(x)$	$\nabla f(x)$	$\alpha * \nabla f(x)$	
10	100	20	2	$\alpha = 0.1$
8	64	16	1.6	
10	100	20	10	$\alpha = 0.5$
0	0	0	0	
10	100	20	20	
-10	100	-20	-20	$\alpha = 1.0$
10	100	20	20	

梯度下降和学习率参数

经典机器学习算法概览

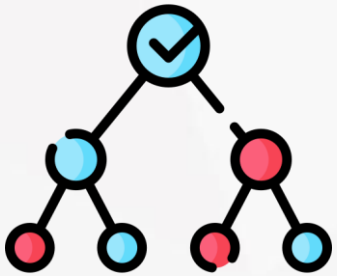
- 基础+集成机器学习算法

线性回归



逻辑回归

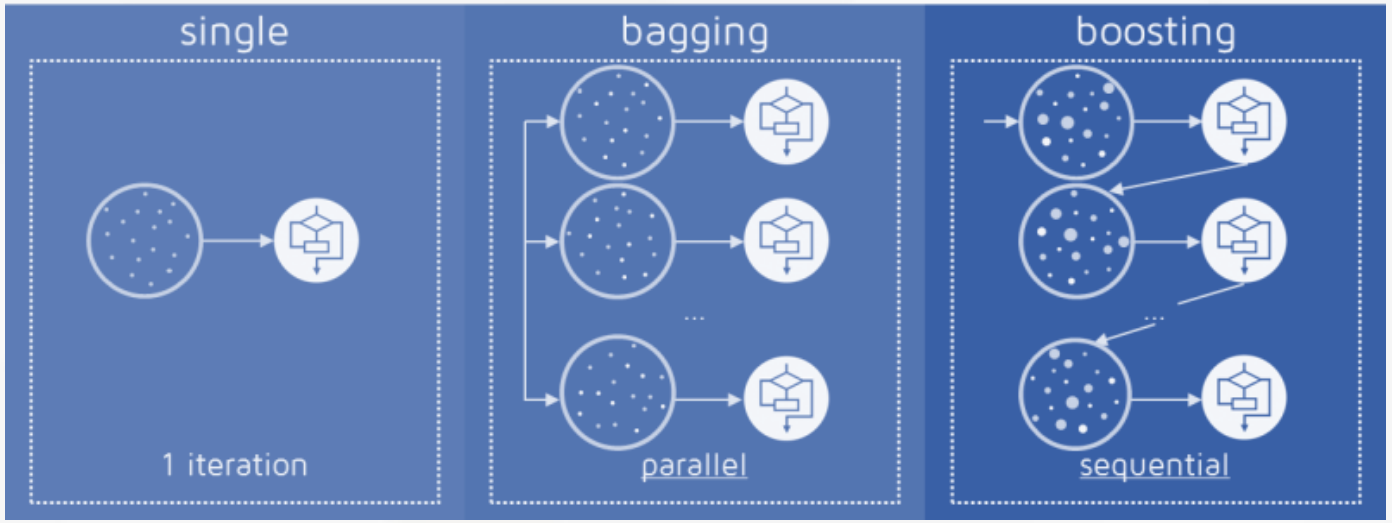
K近邻



决策树

基础机器学习算法

决策树　　随机森林　　GBDT、XGBoost
LightGBM、CatBoost



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)
predict预测	预测结果：Algo.predict(x_test) 分类概率：Algo.predict_proba(x_test)
score打分	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 = ?**

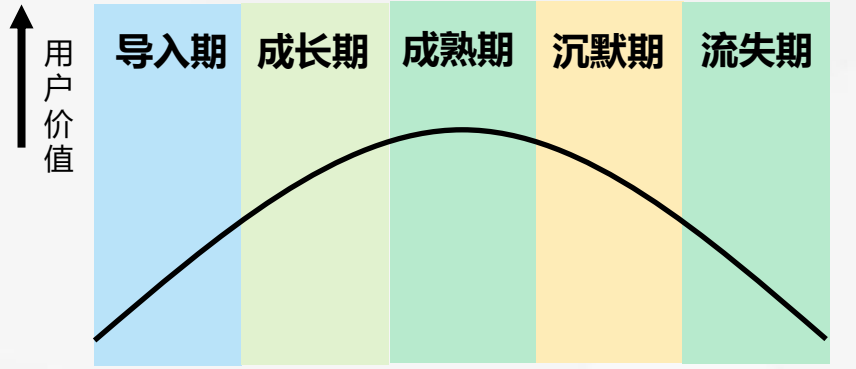
在金融场景中，往往需要先通过运营手段获取客户，然后根据客户的需求，可以通过信贷业务对客户发放贷款赚取利息收入，也可以向客户发售理财产品通过投资业务赚取投资收益，并且全程需要保证合规性，减少欺诈和洗钱带来的不良影响。这整个过程中都可以用AI进行场景赋能。

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。



AI有效赋能用户生命周期各环节运营

客户基本信息
- ◆年龄性别
- ◆籍贯住所
- ◆职业单位

用户行为数据
- ◆App埋点
- ◆借贷行为
- ◆产品购买

产品数据
- ◆投资产品
- ◆信贷产品
- ◆保险产品

银行交易流水
- ◆IP电话
- ◆交易日期
- ◆交易金额
- ◆对方账户

外部数据
- ◆购物消费
- ◆旅游消费
- ◆其他行为分

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，制定对应政策进行营销

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

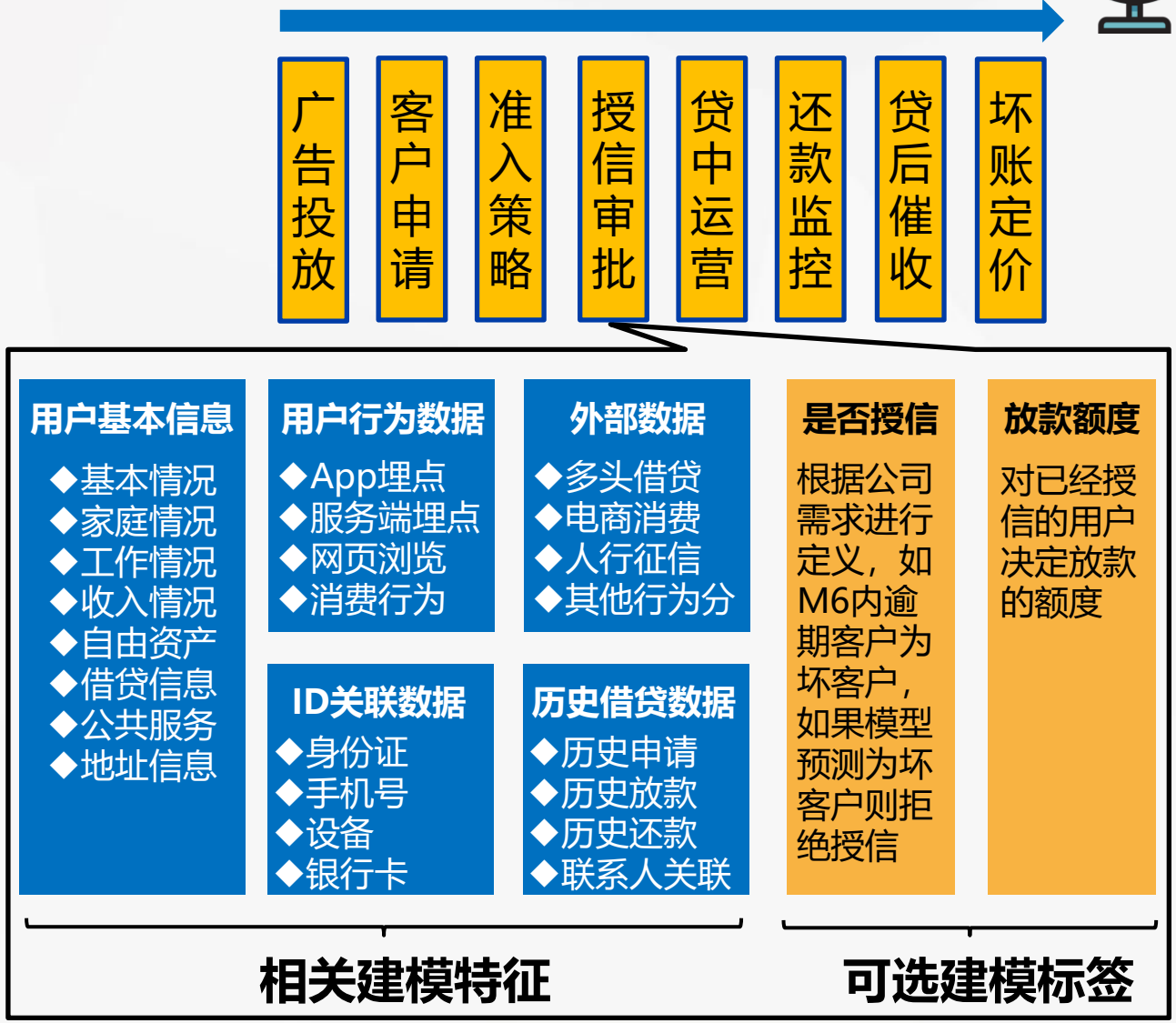
可选建模标签

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。

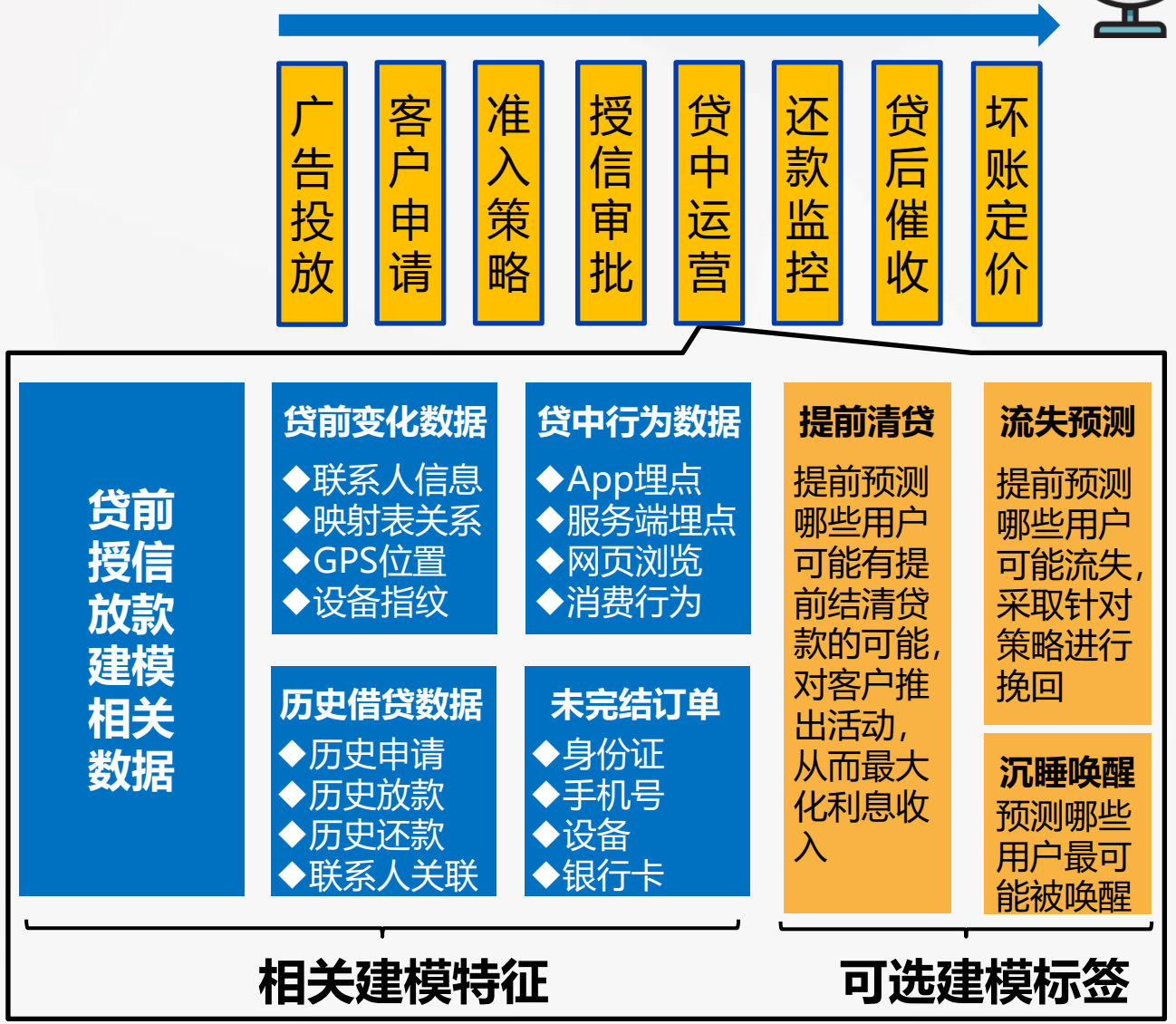


消费金融信贷业务场景

➢ 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。

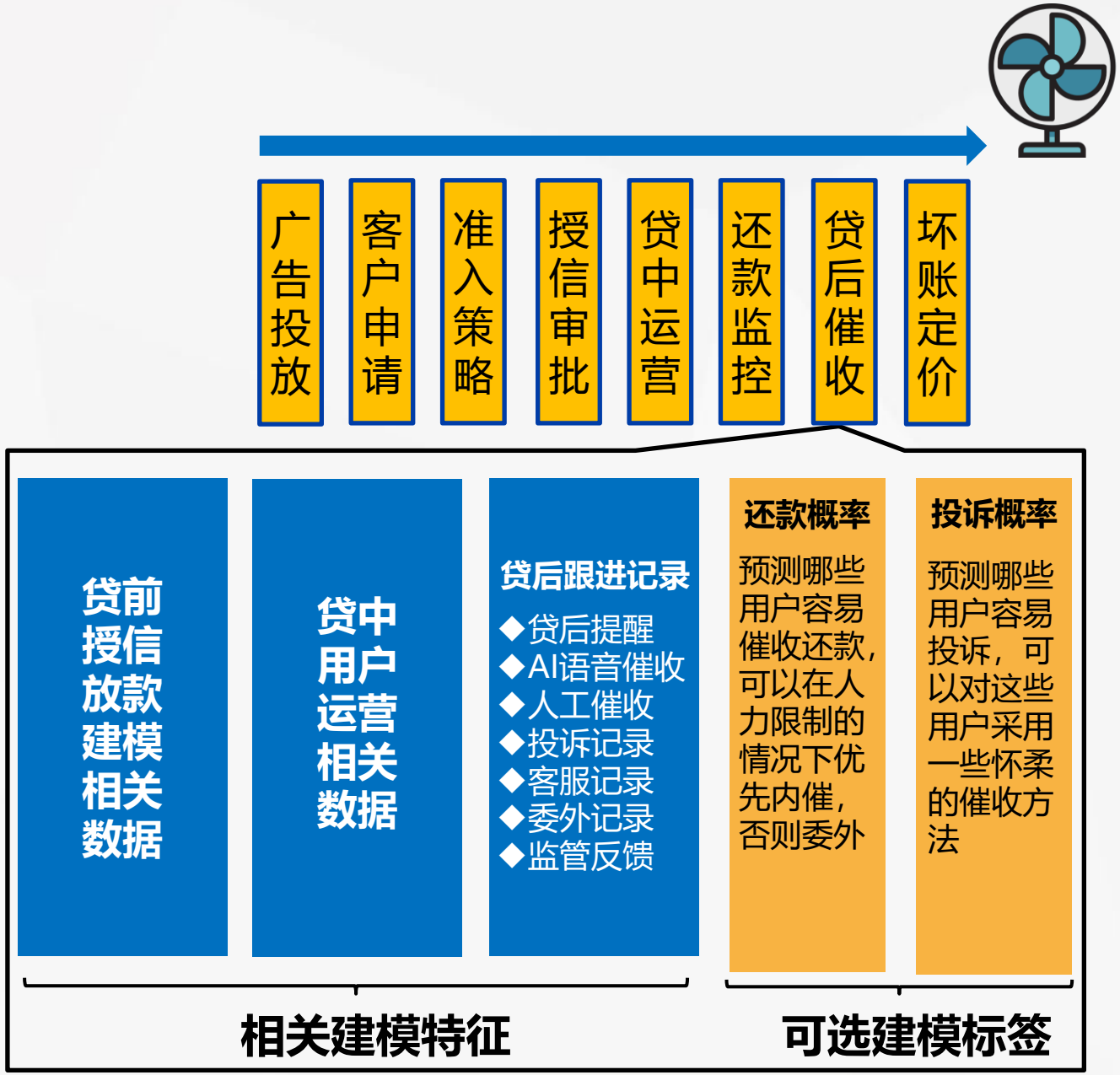


消费金融信贷业务场景

➢ 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。

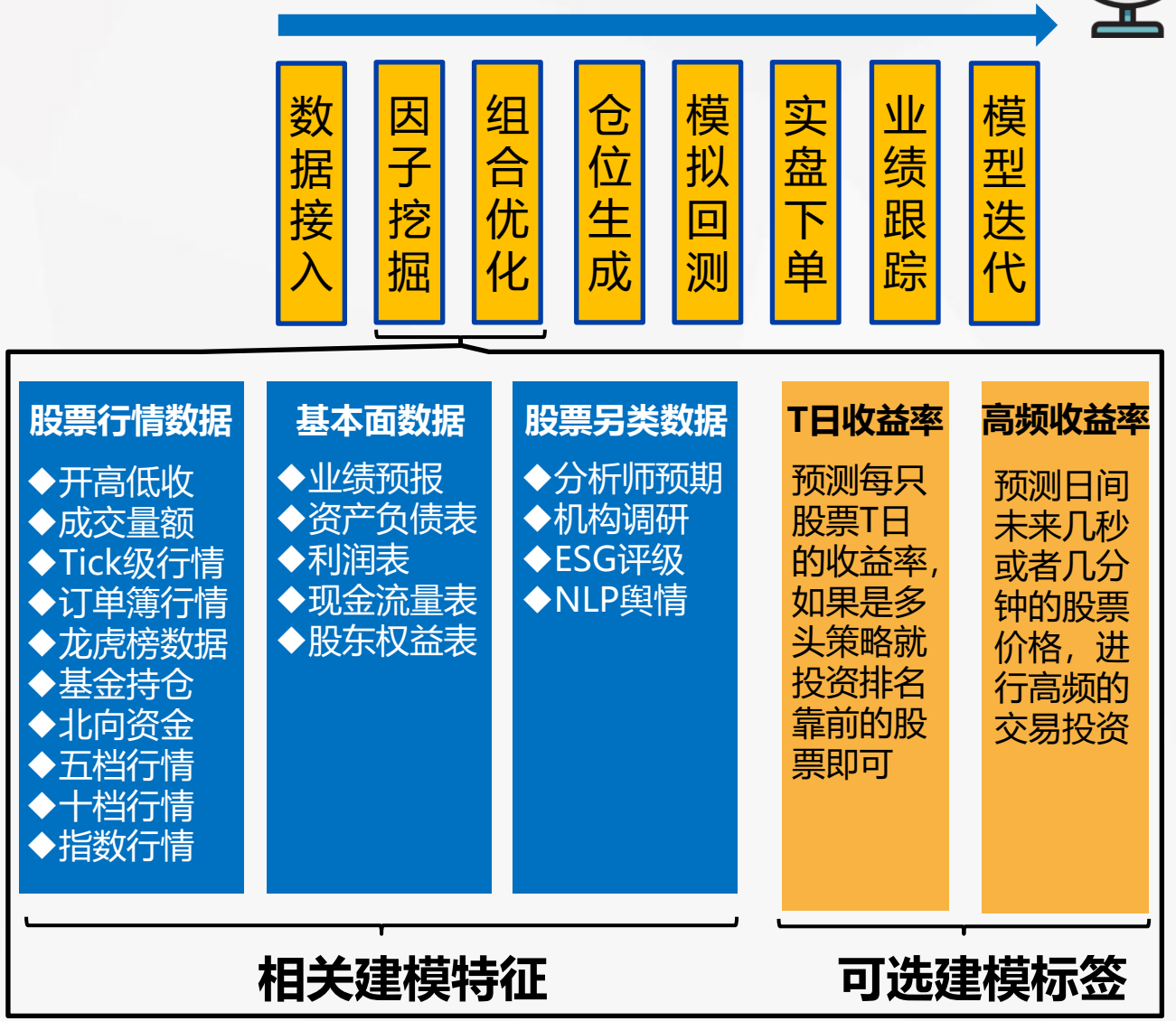


财富管理投资场景

➢ 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。

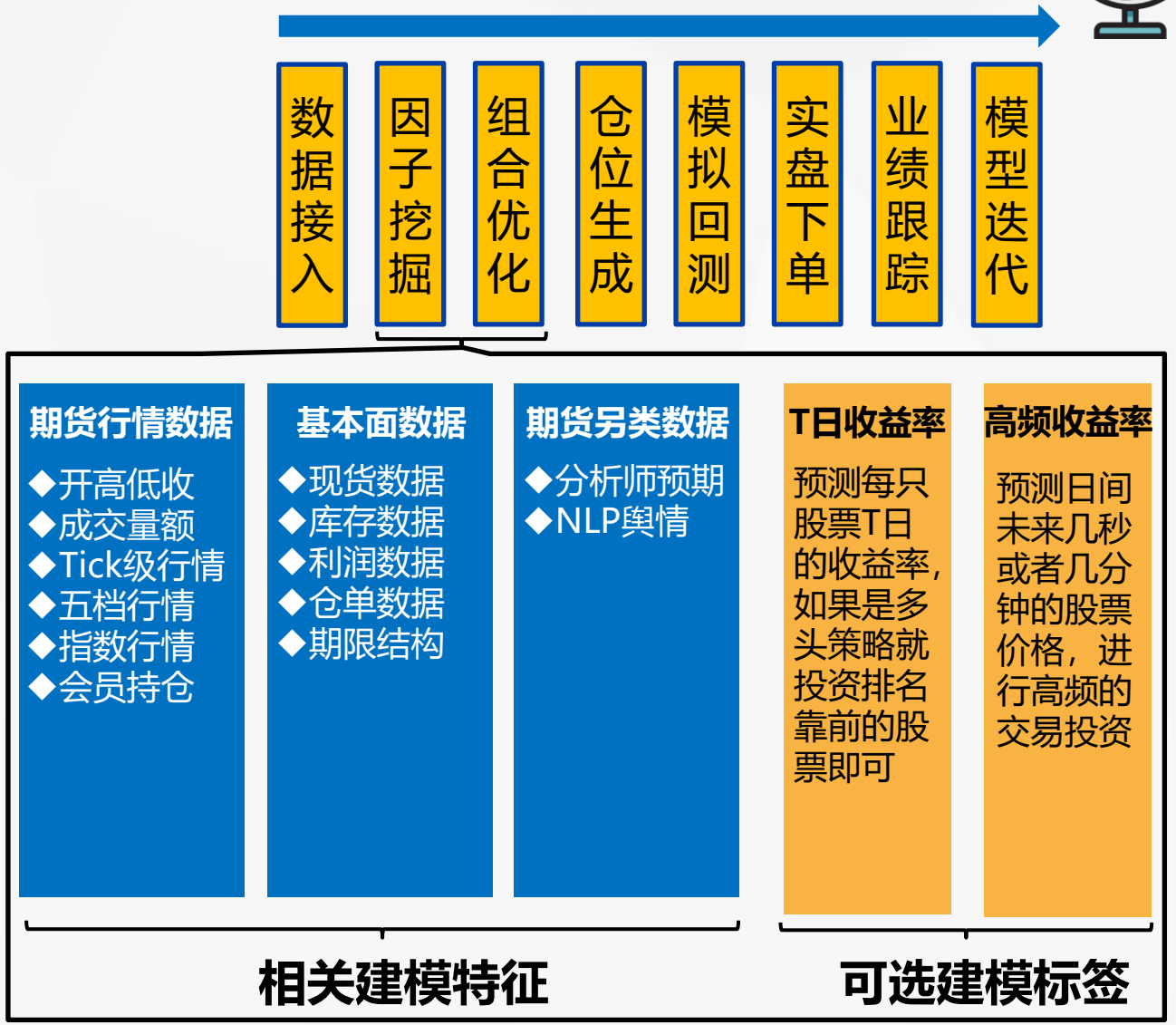


财富管理投资场景

➢ 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。

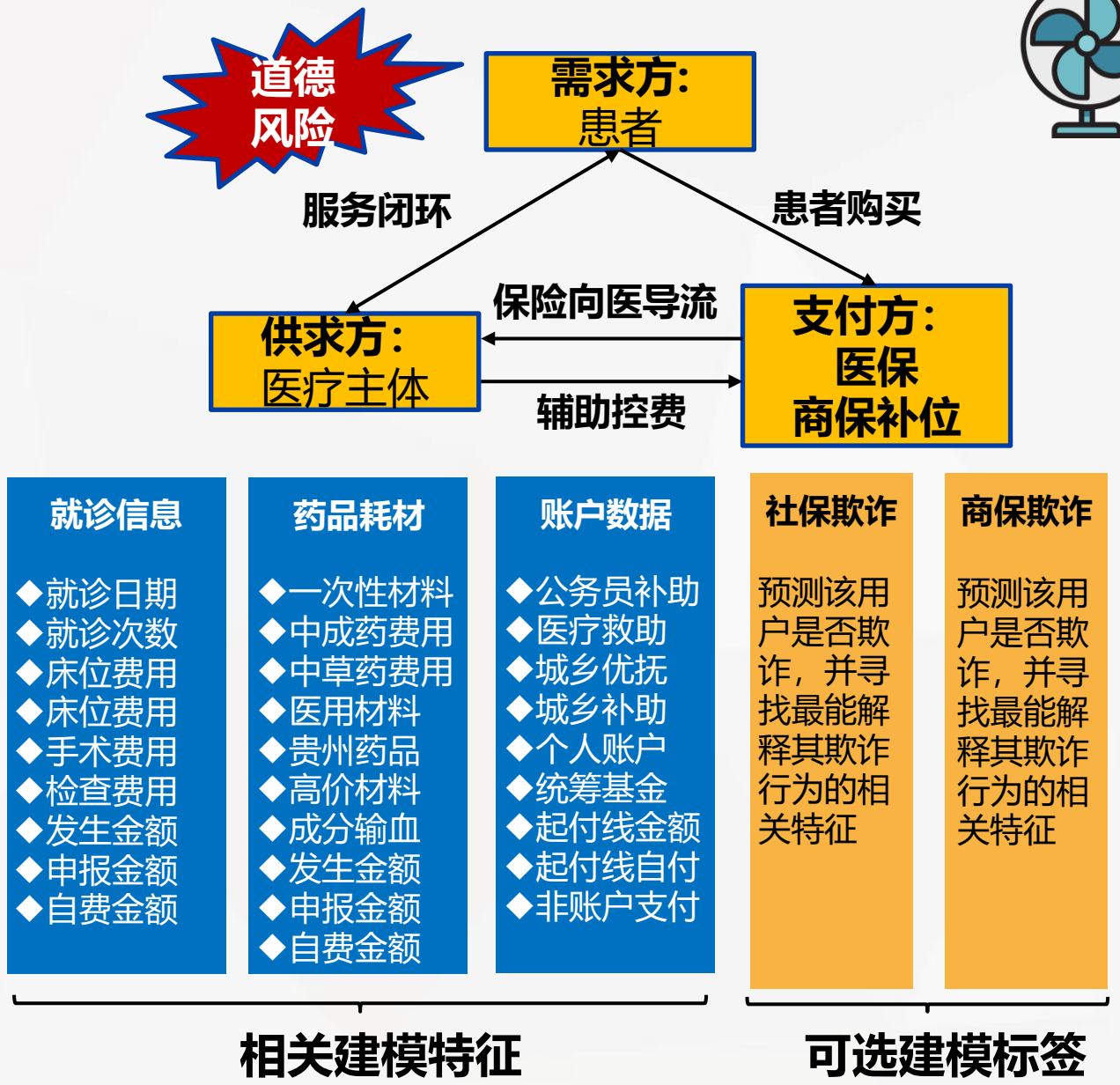


金融合规场景

➢ 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。



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。