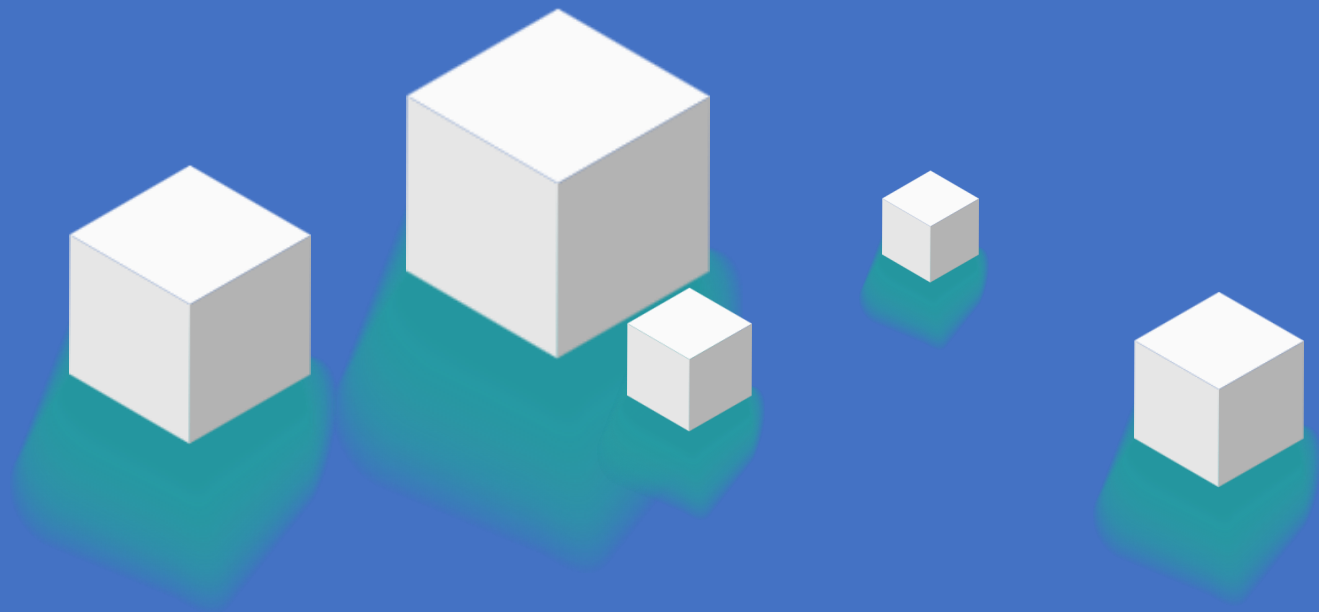
不同领域的AI算法



>> 今天的学习目标

不同领域的AI算法

- 金融行业的应用场景

个人金融、公司金融、金融机构

渠道运营、风险管理、资金交易、数字财务

- 制造行业：缺陷检测
- 快消行业：供应链补货
- AI大赛：二手车价格预测

如何通过特征工程，进行模型调优

金融行业的应用场景

Fintech应用场景

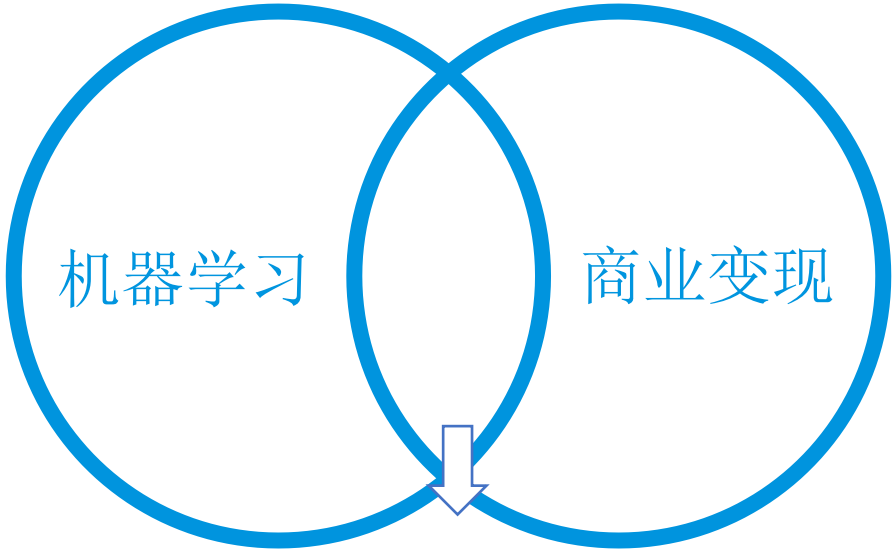
Fintech：

Financial + technology，通过科技手段，使得金融服务变得更加效率

金融业务：保险、银行、券商、基金需要有科技为其支撑。此外，互联网公司也在开展金融业务，如蚂蚁金服

高盛集团的总裁宣称：We are a tech firm.

近几年，摩根大通集团约有1/3的分析师和员工被逼迫加入编程课程学习



需求驱动技术发展

需求驱动人才发展

银行不同部门的应用场景

1、个人金融：

个人金融部门利用数据算法为客户提供个性化服务，比如客户画像与精准营销、家庭财务规划

2、公司金融：

公司金融部门通过数据算法优化企业金融服务，比如：企业信用评估、供应链金融

3、金融机构：

对金融机构客户提供服务，运用银行自身资源，开展负债业务、资产业务和中间业务等，以获取综合效益。

4、渠道运营：

渠道运营部门利用数据算法提升客户体验和运营效率，比如客户行为分析、网点运营管理

5、风险管理：

风险管理是银行的核心职能之一，主要涉及：风险建模，市场风险预测，合规与监控

6、资金交易：

资金交易部门利用数据算法提升交易效率和风险管理，比如算法交易、市场趋势预测

7、数字财务：

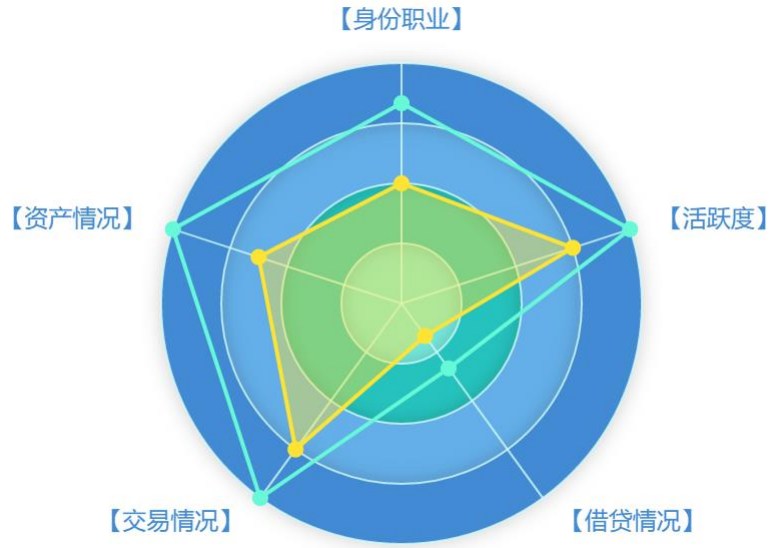
数字财务部门通过数据算法优化财务管理，比如财务预测与决策，智能融资决策

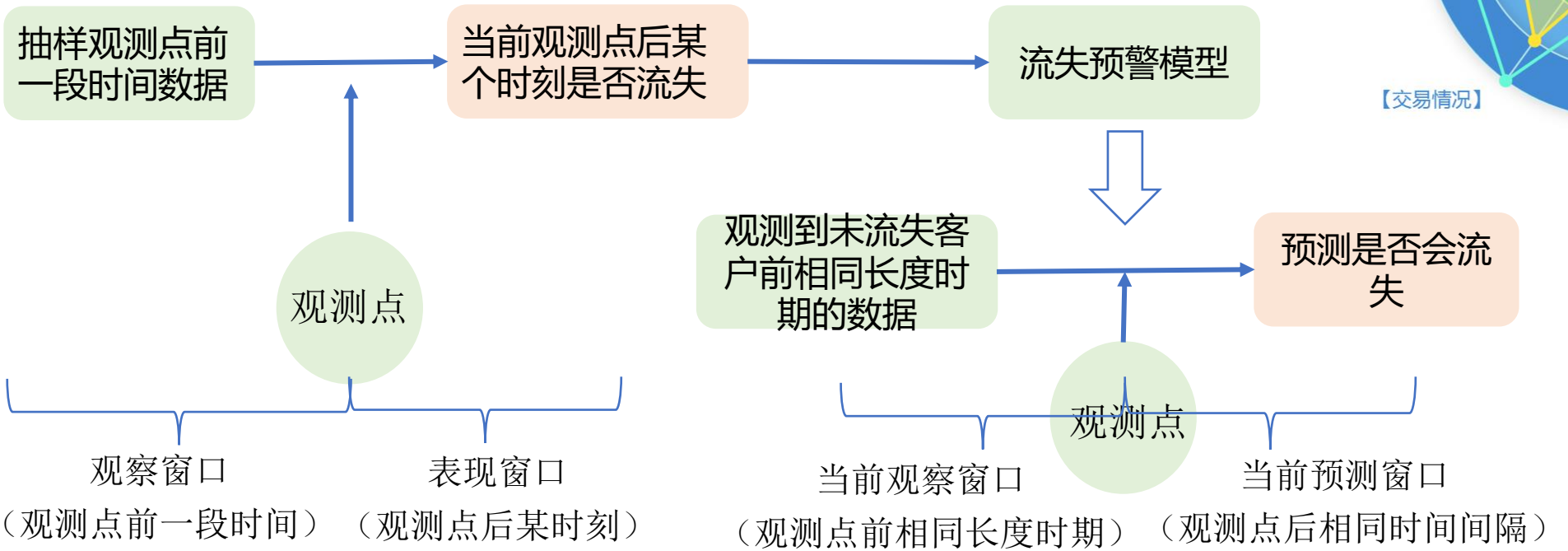
个人金融

客户流失分析与预警

场景：客户流失分析与预警

- 新客获取成本是老客户的5倍
- 同时老客户的价值 大于一般的新客户（习惯需要培养，且价值需要提升）





手机银行贷款页面优化

场景：手机银行贷款页面优化

- 手机银行作为重要的媒介触点，针对不同地域、不同客群、不同行为习惯的用户，匹配贷款产品
- 千篇一律 or 猜你喜欢

标签落地应用（易车案例）

标签生产 → 易知平台 → 易车管理后台

BEFORE

信息流TOP位置
全量+推荐

赋能业务场景多样化

NOW

信息流TOP位置
人群包+推荐

人群包+推荐 VS 全量 + 推荐

曝光量	减少18%
点击量	提升5%
点击率	提升30%

提供半自动化运营工具
赋能运营手段多样化

因客定价

场景：因客定价

- 利率可以基于借款人的信用、贡献度等因素进行差别化定价

优质客户 => 利率低一些

普通客户 => 利率高一些

优质客户的衡量可以参考信用评分卡

第一部分:基本身份信息					
0-18	0分	婚姻情况	未婚	5分	
18-24	1分		已婚	9分	
25-29	8分		已婚有子女	10分	
30-44	9分		已婚无子女	6分	
45-54	8分		离异有子女	4分	
55-48	3分		离异无子女	3分	
60以上	1分				
本地户籍	10分	现居住地址（年）	1年以内	2分	
其他户籍	2分		1-3年（含）	6分	
研究生及以上学	10分		3-5年（含）	8分	
大学本科	8分		5年以上	10分	
大专	7分				

第二部分:工作信息					
工作单位	公务员事业单位教师银行从业人员				10分
	国有企业大型上市公司员工				8分
	解放军中国武警部队公安检查				5分
	一般上市公司 股份制公司 私营企业员工				6分
	个体户以及个体户从业者				3分
	自由职业人员				0分
职位	高级领导	10分	职称	高级职称	10分
	中级领导	8分		中级职称	6分
	普通员工	3分		初级职称	5分
现单位工作时间	0-6个月	0分			
	6个月-一年	1分			
	1-3年	4分			
	3-5年	7分			
	5年上	8分			
	无单位	0分			
第三部分:资产信息					
房产信息	自有全款	6分	固有资产,房产车产 存款总值（万元）	50万以下	1分
	自有按揭	10分		50-100万	2分
	自建房	4分		100-200万	6分
	租房	0分		200万以上	7分
保险信息	购买人寿医疗 健康 分红 理财保险1年				6分
	购买人寿医疗健康分红理财保险1-2年				8分
	购买人寿医疗 健康分红 理财保险2年以上				10分
	参加5险				9分
	参加5险一金				10分
第四部分:负债比					
负债比	0-10%	9分	家庭年收入	5万以下	0分
	10%-20%	8分		6-9万	1分
	20%-40%	5分		10-15万	2分
	40%-60%	4分		16-20万	4分
	60%-80%	1分		20-30万	5分
	80%以上	0分		30万以上	8分

因客定价

某银行推出个人消费贷，采用“因客定价”策略，根据借款人的信用状况、贡献度等因素，差异化设定利率。

1. 优质客户：低利率

客户信息

姓名：张先生

信用评分：850分（分数区间：350-950分）

贡献度：高（长期存款客户，持有银行理财产品）

职业：大型企业高管

收入：稳定且高于平均水平

定价逻辑

信用评分高，违约风险低。

对银行贡献度高，属于高价值客户。

职业和收入稳定，还款能力强。

利率设定

基准利率：5%

优惠利率：4.2%（基于客户信用和贡献度下调0.8个百分点）

因客定价

2. 普通客户：较高利率

客户信息

姓名：李先生

信用评分：650分（分数区间：350-950分）

贡献度：中（仅持有普通储蓄账户）

职业：自由职业者

收入：波动较大

定价逻辑

信用评分中等，有一定违约风险。

对银行贡献度一般，未达到优质客户标准。

职业和收入稳定性较差，还款能力存在不确定性。

利率设定

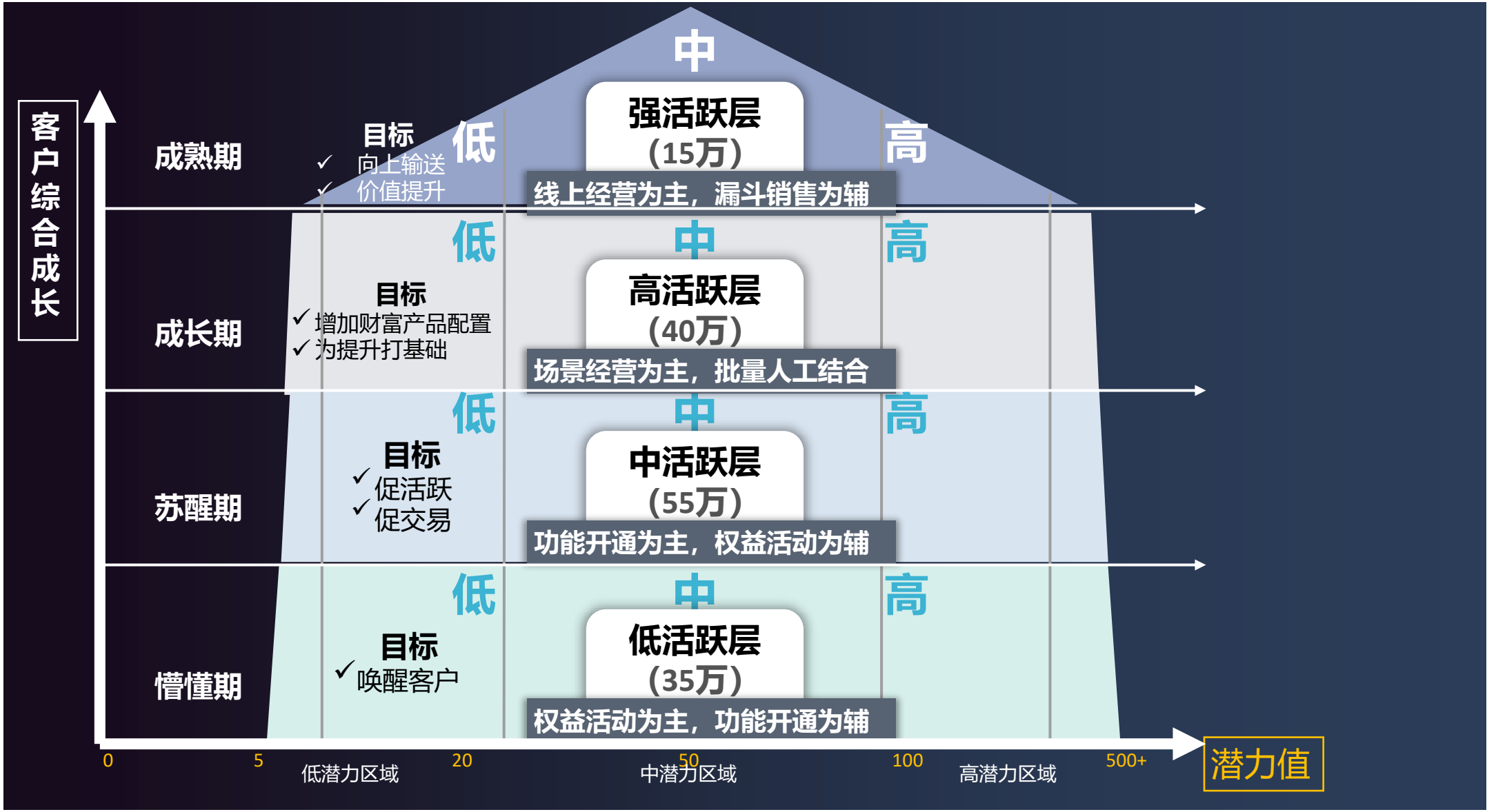
基准利率：5%

实际利率：6.5%（基于客户信用和贡献度上浮1.5个百分点）

长尾客群营销

经营客群	客群大类	经营目标	经营事件	经营策略	配置产品
成熟期	理财产品到期类	客群维稳	理财到期	结合客户产品持仓及风险情况，优先转化，其次续接	权益基金/分行卫星池产品
	单一现金持仓	客群维稳	产品转化	利用产品转化提高整体收益率	权益类基全/黄金账户
	资产缺配	客群维稳	资产配置	用料学的资产配置理念引导客户进行全面的配雪	金生利/权益类基金/债券基金等
	预警流失模型客户	客群维稳	资产配置	定期回访切入，了解有户的后续资金安排情况，有针处生进行给出合理建议，运用资产配置进行产品的领定。	资产配置的科学理念的应用
	存款偏好客户	存款提升	定期回访	介给我行的结构性存款或者大题存款，进行新资金的引入，增加客户的存款和AUM	结构性存款/大额存单
成长期	存款偏好客户	存款提升	定期回访	介绍我行的结构性存款或者大额存款，进行新资金的引入，增加客户的存款和AUM	结构性存款/大额存单
	潜力资产模型客户	客群提升	定明回访	介绍达标客户的权益体系，邀请客户升级卡等级	高收益理财产品/卡权益体系
	三方客群	客群提升	三方回流	分析目前市场行情，建议客户合理进行资产的分布，降低整体的风脸	权益类基金
	国有行净转出	客群提升	资金转出	了解客户转出的原因，用招行的理财产品收益优势授国转出的资金	高收益理财产品
	国有银行主账户	客群提升	服务邀约	综合介绍我行的财富管理优势，邀约客户体验。	高收益理财产品
苏醒期	资产临界不达标	客群提升	定期回访	介绍达标客户的权益体系，激请客户达标享受更多的除	高收益理财产品/增值服务
	高有效代发	代发留存	代发工资/奖金	洁合客户产品持仓及风险情况，制定周期性的定投计划或者财富产品的买入，实现资产的不断票积	基金定投/摩揭智投/高收益理财产品
	大额流入	大额留存	大额转入	了解客户后续的资全使用安排，用财富产品锁定资金	高收益理财产品/流动性理财产品
	大额流出	资金挽回	大额转出	了解客户转出的原因，给予合理化建议，邀约客户转回。	高收益理财产品/流动性理财产品
懵懂期	礼品型客户	粘度提升	权益活动	利用营销活动提升客户的信任度和粘度，先建立关系	活动
	活动型客户	粘度提升	综合活动	利用活动提升客户的信任度和粘度，逐渐挖掘客户潜力。	活动

长尾客群营销（基于客户的潜力值及活跃度进行细分）



公司金融

贷款商机挖掘

建立分行级存量客户数据库

一个数据库能装入分行XXX万零售客户，

一期建设若干授信敏感变量，包含个人反洗钱九要素，分行近一年12个月没有AUM数，四个季度结息数，以及是否为小微企业相关标识和信用卡持有情况

由于信用卡本身做了征信查询授权，就可以利用信用卡的贷后征信查询做信用模型，数据库的变量做白名单筛选。将总行产品预审模型到分行数据库跑数，生成最终准客户名单。

最终挑选出存量客户中符合贷款产品的准客户，交由三支队伍中的存量经营团队线下营销

分行存量零售客户
XXX万

小微标识客户名单
XX万

白名单

贷款商机挖掘

零售大表：客户姓名，年龄，收入，人生阶段，资产层级，职业……

XXX万零售客户

（包括信用卡和非信用卡客群）

企业客户表：客户姓名，年龄，企业名称，行业经营范围，企业成立时间，持股比例，是否专精特新企业纳税评级

XX万小微客户

字段所属表：所属库名，表名，字段名称，中文名称，字段类型，字段描述，字段备注，脱敏类型，安全等级，允许空值，统计口径是否主键，是否码值

意愿表：号码，姓名，性别，年龄，是否纳税A评级，营销响应分

筛选营销响应分 > 520的客户

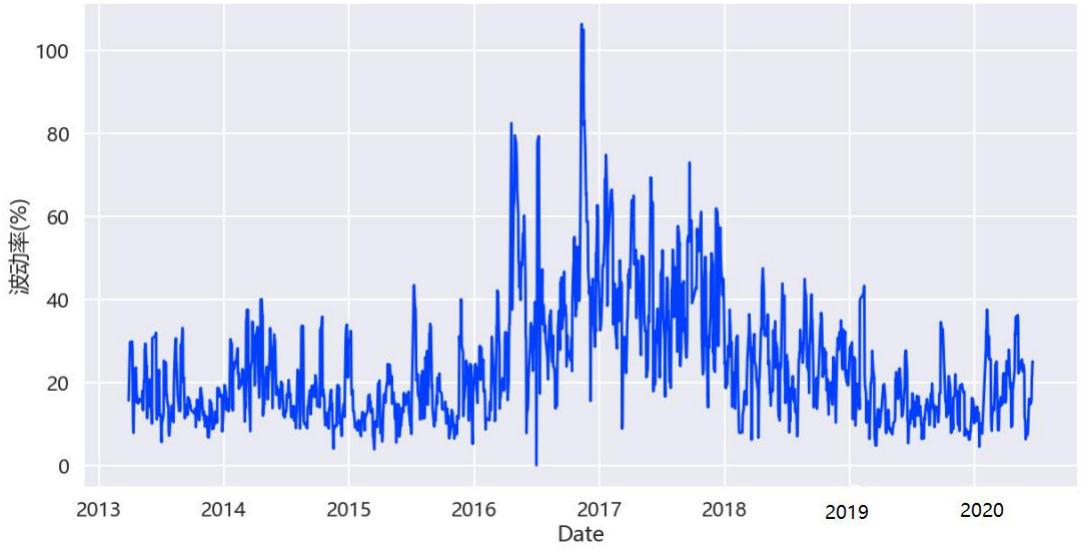
159XXXX	张飞	男	50	A	615
158XXXX	cy	男	38	A	625
……	……	……	……	……	……

金融机构

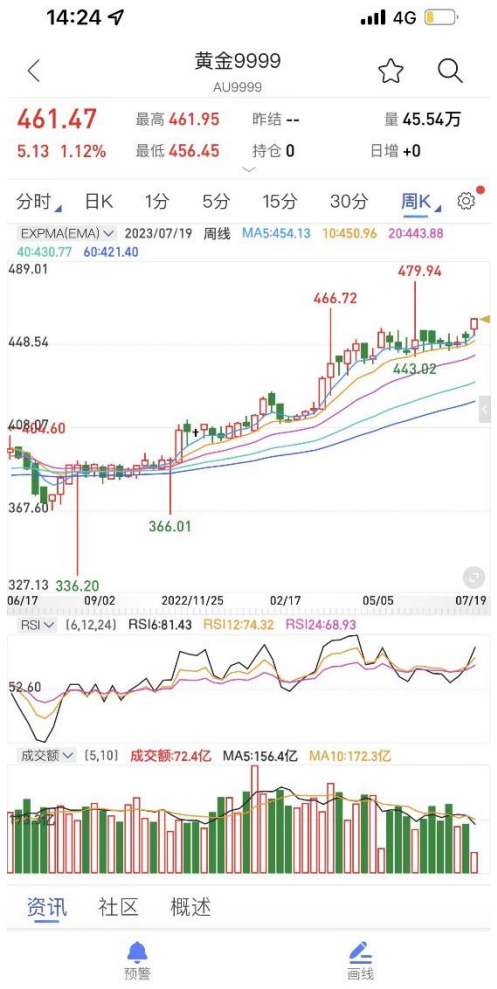
金融机构大客户与交易相关性分析

场景：金融机构大客户与交易相关性

- 市场上的波动会影响到金融机构大客户在我行的结算余额
- 对当前市场波动率进行判断，并基于结算余额与市场波动率的相关性进行分析和预警



黑色系商品期货波动率

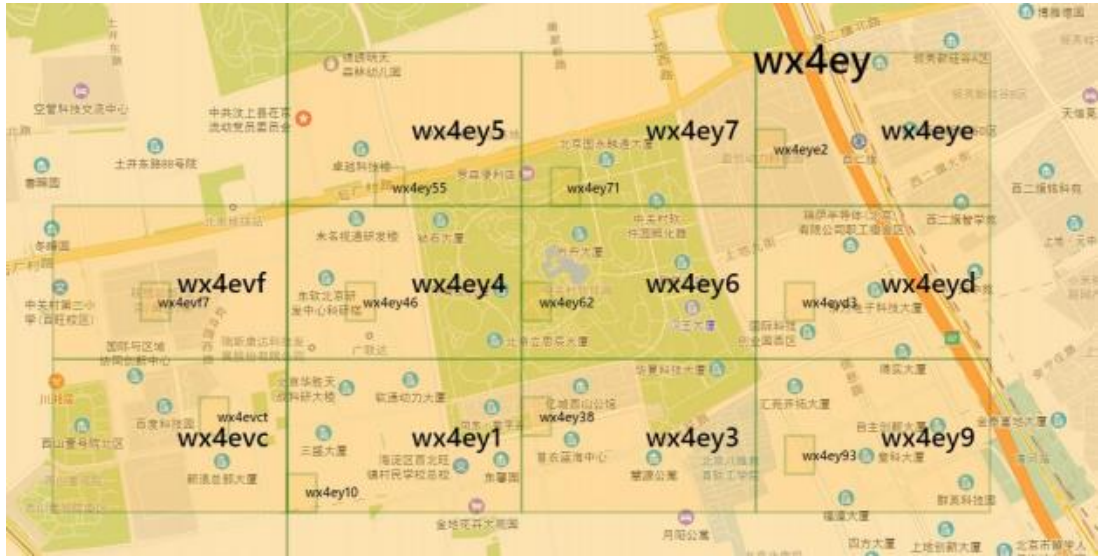


黄金期货主连

沪镍期货主连

渠道运营

商圈生意机会地图



基于GeoHash编码的区域，匹配用户归属算法，确定每个GeoHash片区内的选址相关基础数据：

常住人口数量及趋势

常住人群画像

区域消费偏好

区域消费等级

区域人口密度热力

用户画像

商品画像

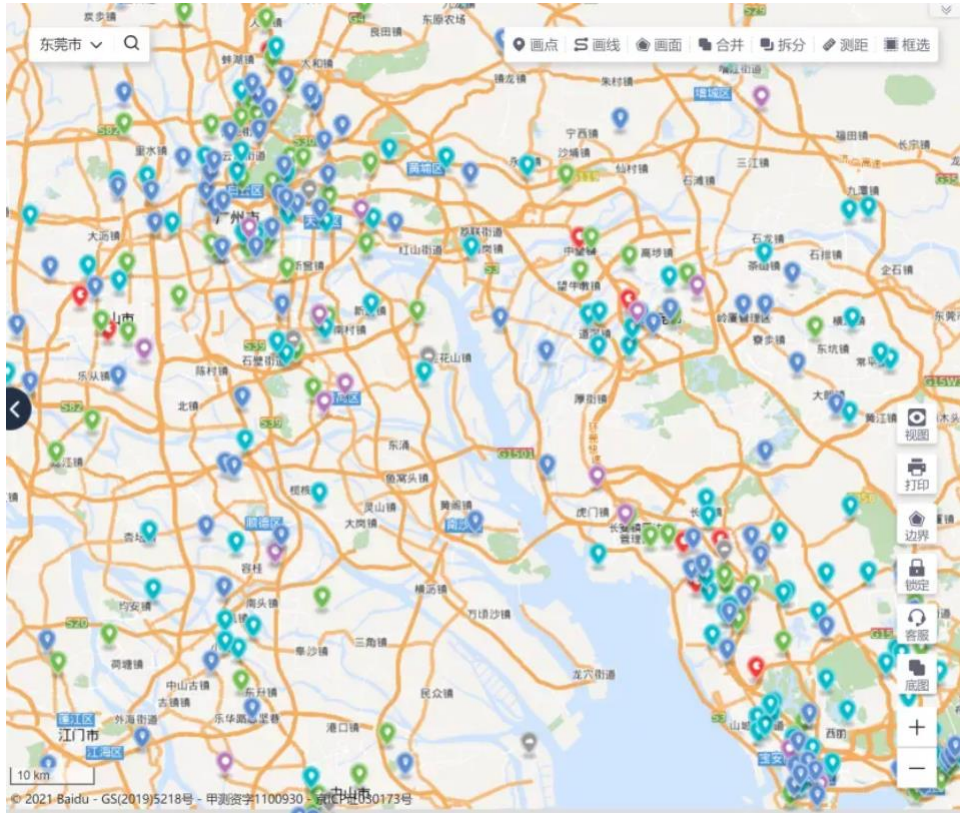
商品购买行为

站内访问行为

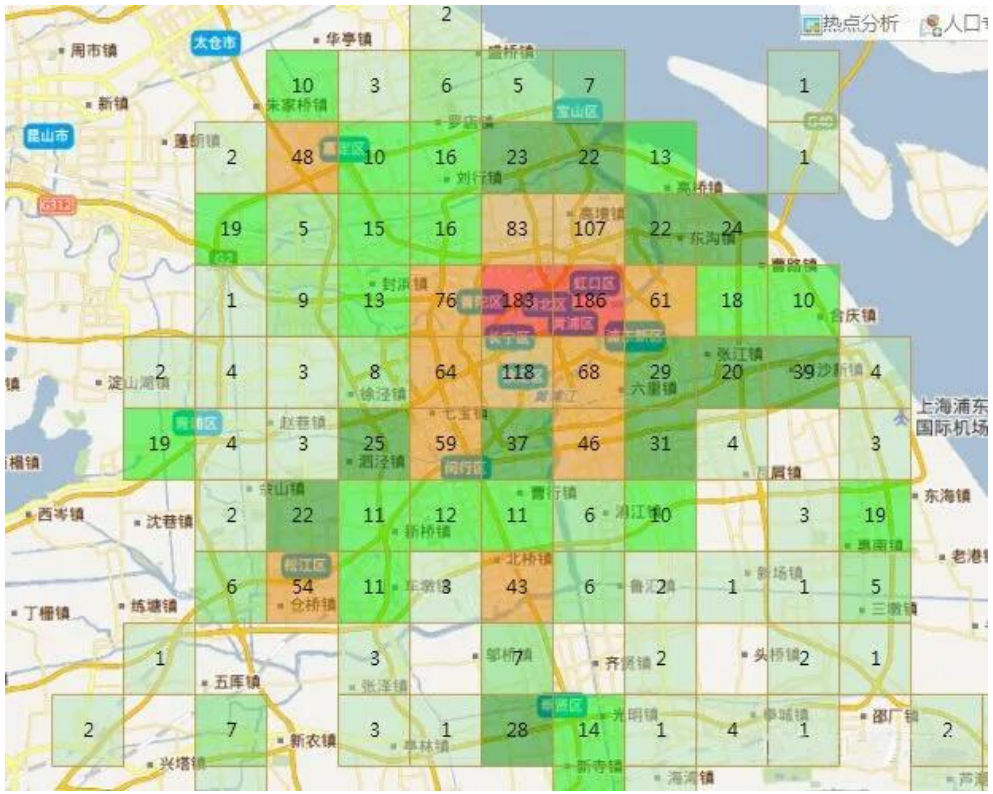
LBS Data

POI Data

商圈生意机会地图



基于GeoHash编码的区域，计算区域的竞争压力指数（网点密集度）



结合人口密度分布，计算区域网点的生意机会地图

风险管理

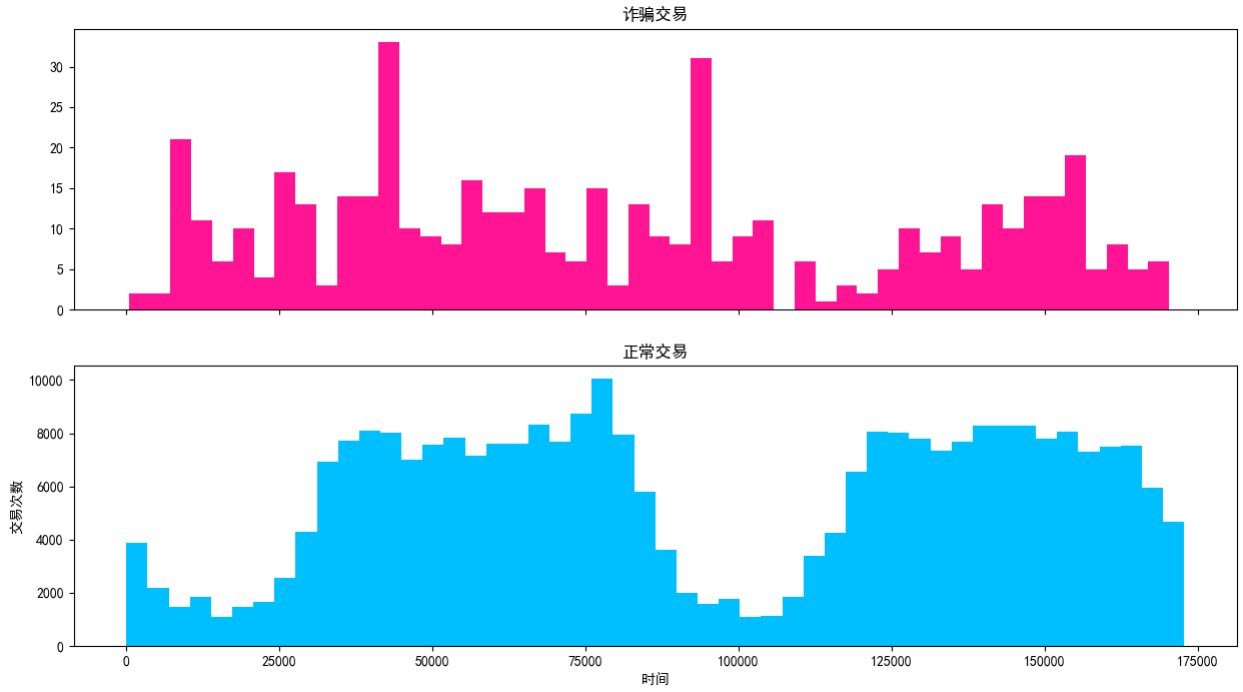
金融风控模型

场景：金融数据分析与风险控制

在信贷领域中存在信用卡违约和欺诈的风险，通过用户行为数据，分析申请借款用户的信用状况，来判断是否存在逾期。

通过分析交易时间、交易金额、收款方等多维度数据，我们还可以对信用卡会否被盗刷进行预测，防止信用卡被盗刷的风险

此外，通过预测模型，我们还可以分析出哪些因素容易导致违约，从而加强产品的设计

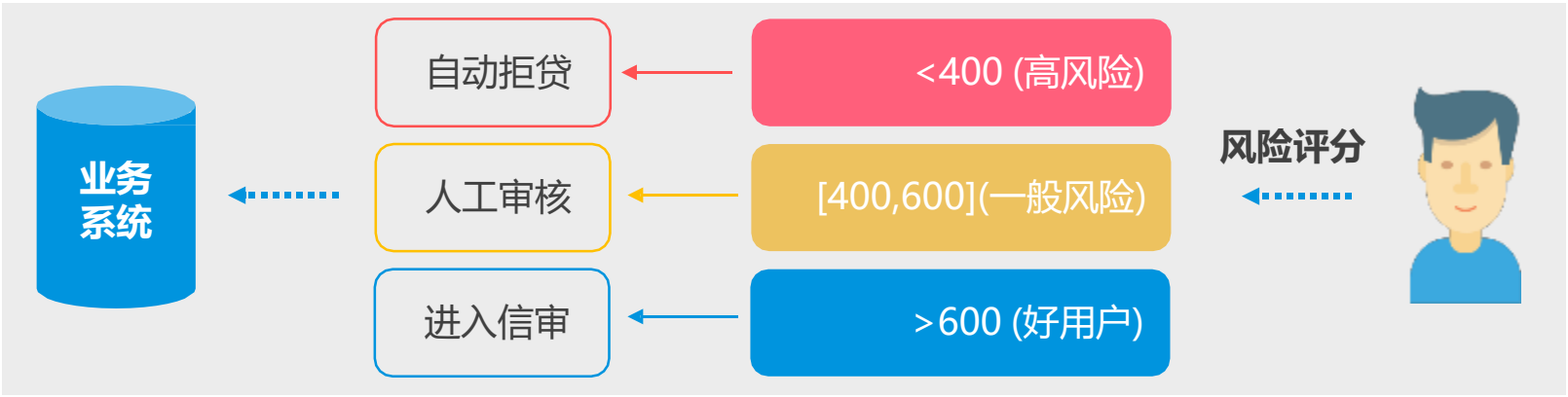


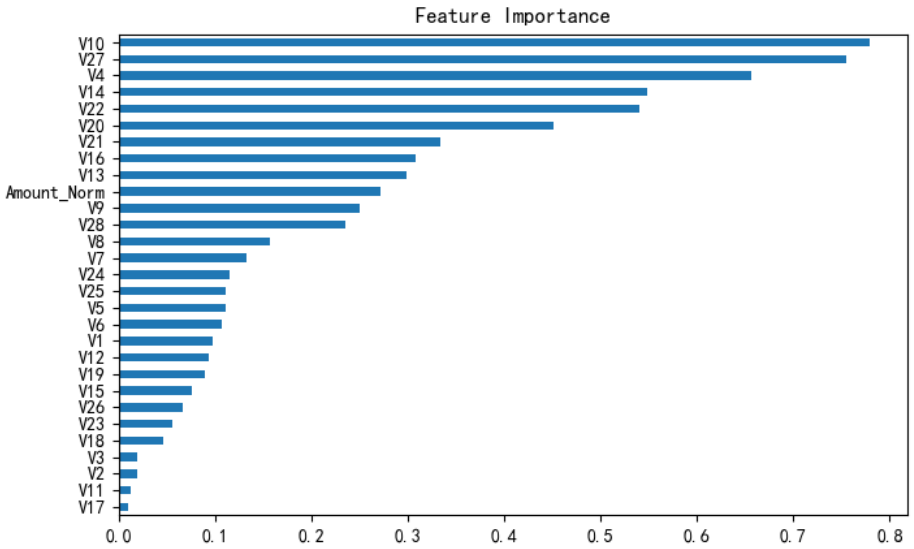
金融风控模型

1 SDK 采集行为数据

2 结合全网数据绘制关系图谱

3 机器学习完善欺诈风险模型





金融风控模型

评分卡模型：

- 评分卡模型是常用的金融风控手段之一

风控，就是风险控制，我们采取各种措施和方法，减少风险发生的可能性，或风险发生时造成的损失

- 根据客户的各种属性和行为数据，利用信用评分模型，对客户的信用进行评分，从而决定是否给予授信，授信的额度和利率，减少在金融交易中存在的交易风险
- 按照不同的业务阶段，可以划分为三种：

贷前：申请评分卡（Application score card），称为A卡

贷中：行为评分卡（Behavior score card），称为B卡

贷后：催收评分卡（Collection score card），称为C卡

Thinking： A卡B卡C卡的数据源有何区别？

1）外部征信数据是A卡中用户的主要数据

2）对于复贷用户，已经有了历史平台表现，所以B卡通常不会再次查询用户的外部数据，而是利用历史平台表现 => 节省成本

3）C卡不会再次查询用户外部征信数据，主要利用历史贷款过程中，催收人员记录的用户表现作为主要数据

金融风控模型

评分卡模型：

- 客户评分 = 基准分 + 年龄评分 + 性别评分 + 婚姻状况评分 + 学历评分 + 贷款申请次数

Thinking: 某客户年龄为27岁，性别为男，婚姻状况为已婚，学历为本科，贷款申请次数为1次，那么他的评分=?

650 (基准分) + 8(年龄评分) + 4(性别评分) + 8(婚姻评分) + 12(学历评分) + 13(贷款申请次数) = 695

Thinking：评分卡的最高分和最低分是多少?

最低分：650-8-2-3+1-8=630

最高分：650+13+4+8+12+13=700

变量名称	变量范围	得分
基准分	-	650
年龄	age<18	-8
	18<=age<25	-2
	25<=age<35	8
	35<=age<55	13
	55<=age	5
性别	男	4
	女	2
	未知	-2
婚姻状况	已婚	8
	未婚	-2
	未知	-3
学历	本科及以上	12
	大专	8
	其他	1
贷款申请次数（二代征信）	>3次	-8
	=3次	0
	=2次	5
	<=1次	13

Case：基于评分卡的风控模型开发

Project：基于评分卡的风控模型开发

- 数据集GiveMeSomeCredit，15万样本数据

https://www.kaggle.com/c/GiveMeSomeCredit/data

– 基本属性：包括了借款人当时的年龄

– 偿债能力：包括了借款人的月收入、负债比率

– 信用往来：两年内35-59天逾期次数、两年内60-89天逾期次数、两年内90天或高于90天逾期的次数

– 财产状况：包括了开放式信贷和贷款数量、不动产贷款或额度数量。

– 其他因素：包括了借款人的家属数量

– 时间窗口：自变量的观察窗口为过去两年，因变量表现窗口为未来两年

字段	说明	类型
SeriousDlqin2yrs	90天以上逾期的人，未来2年违约	Y/N
Age	年龄	整数
RevolvingUtilizationOfUnsecuredLines	除房地产和汽车贷款等无分期付款债务外，信用卡和个人信用额度的总余额除以信贷限额	百分比
DebtRatio	债务比（每月偿还的债务，赡养费，生活费除以每月的总收入）	百分比
MonthlyIncome	每月收入	实数
NumberOfOpenCreditLinesAndLoans	公开贷款(如汽车贷款或抵押贷款)和信用额度(如信用卡)的数量	整数
NumberRealEstateLoansOrLines	抵押贷款和房地产贷款的额度（包括房屋净值信贷）	整数
NumberOfTime30-59DaysPastDueNotWorse	借款人逾期30-59天的次数，但在过去两年没有更糟	整数
NumberOfTime60-89DaysPastDueNotWorse	借款人逾期60-89天的次数，但在过去两年没有更糟	整数
NumberOfTimes90DaysLate	借款人逾期90天（或以上）的次数	整数
NumberOfDependents	除自己(配偶、子女等)以外的家庭受养人人数	整数

资金交易

外汇市场预测

场景：外汇市场预测

- 对外汇市场进行预测，进行金融产品设计

比如两个月 美债和人民币倒挂激增，再看欧元，是否做个三角套汇（是否有套利空间，如果银行推广这个商品 银行是有收益的，对客户也是有收益的）

1）波动规律分析法

通过外汇市场行情的波动进行分析，认为之前的行情变化规律和后市行情的变化规律有一定的参照性

2）基本面分析法

适合中长期行情，货币的强弱反映该国经济的好坏，尽管中间会有暂时性的波动，但从长期来看，它的价位会最终回归到与经济状况相匹配的位置上。

3）技术分析法

比如MACD、RSI、移动平均线等指标作为分析依据

期货套利模型

场景：期货套利模型

- 影响某个品种的涨跌因素非常多（自身基本面、宏观经济、原材料价格等）
- 如果只做单边交易 => 遇到突发事件，会造成剧烈波动，导致平仓
- 套利交易可以针对单品种跨期，跨品种交易

两个合约的相关性极高，一多一空，可以对冲掉90%以上不确定因素 => 只要核心逻辑正确的，大概率是赚钱的

Thinking：空沪金 & 多伦敦金，在初期各投资50%在国内和国外，某一时期能稳定获利，需要注意什么？

数字财务

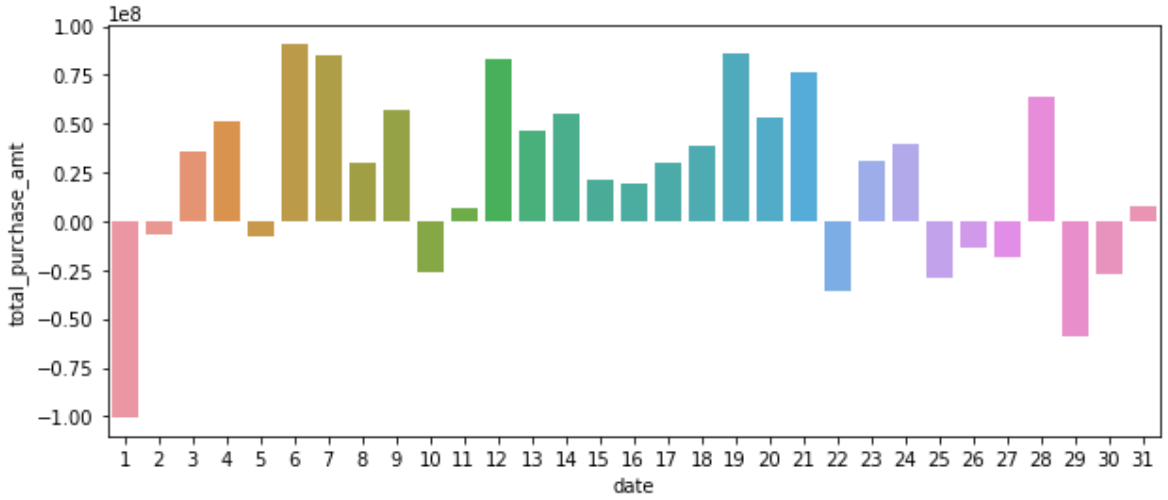
资金流入流出预测

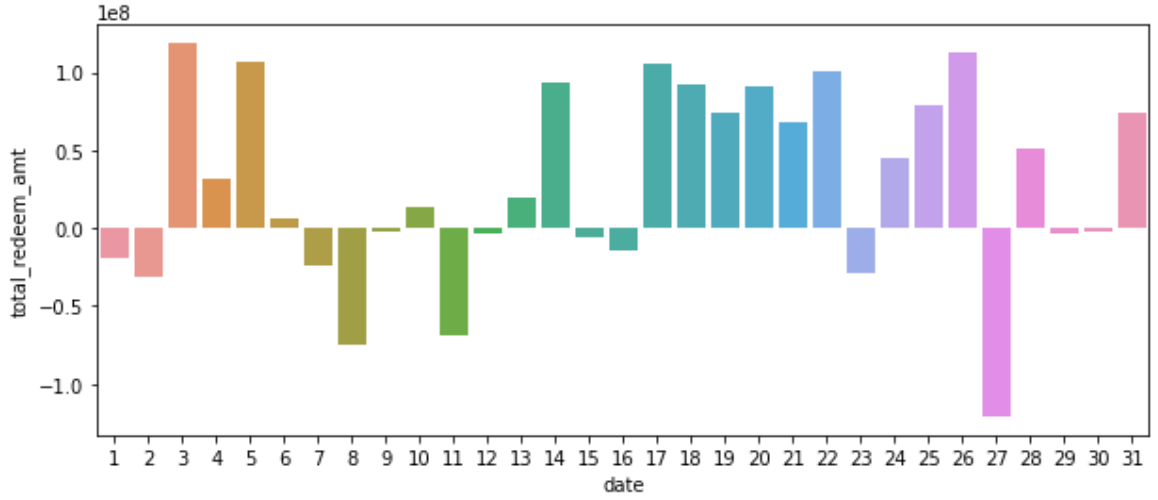
场景：资金流入流出预测

- https://tianchi.aliyun.com/competition/entrance/231573/information
- 数据集一共包括4张表：用户基本信息数据、用户申购赎回数据、收益率表和银行间拆借利率表

2.8万用户，284万行为数据，294天拆解利率，427天收益率

2013-07-01到2014-08-31，预测2014年9月的申购和赎回

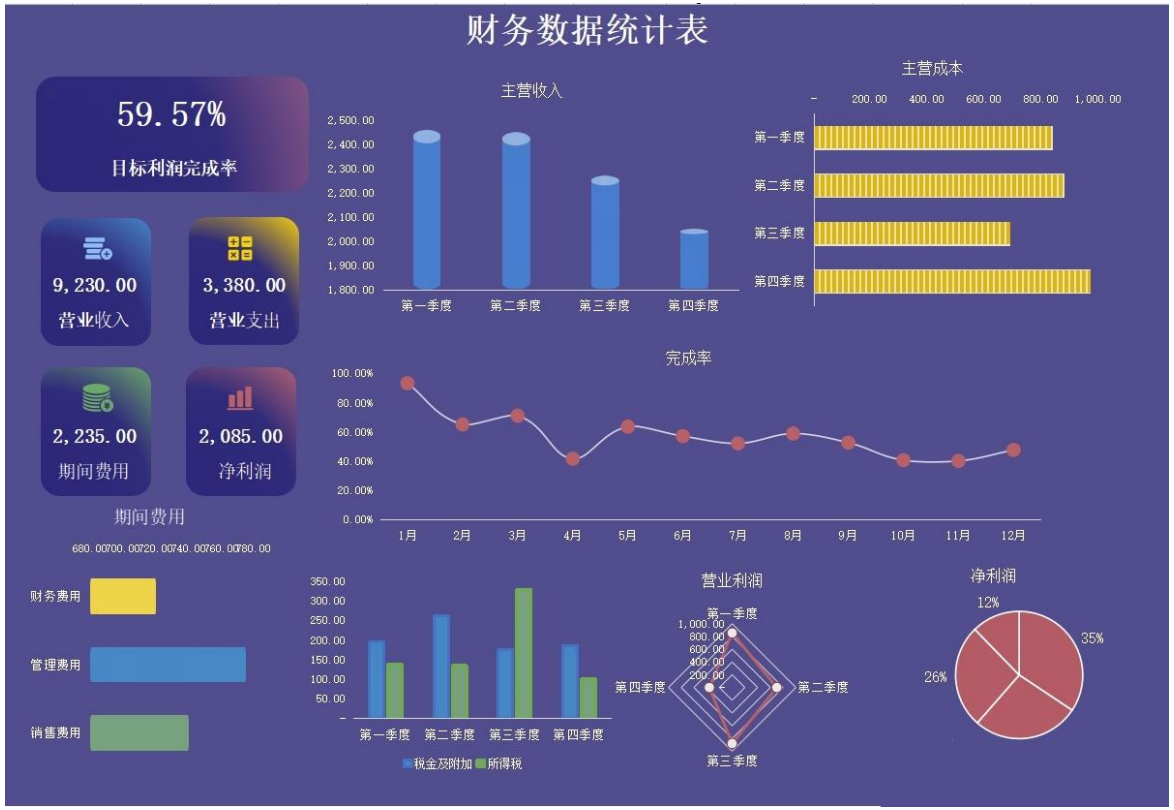


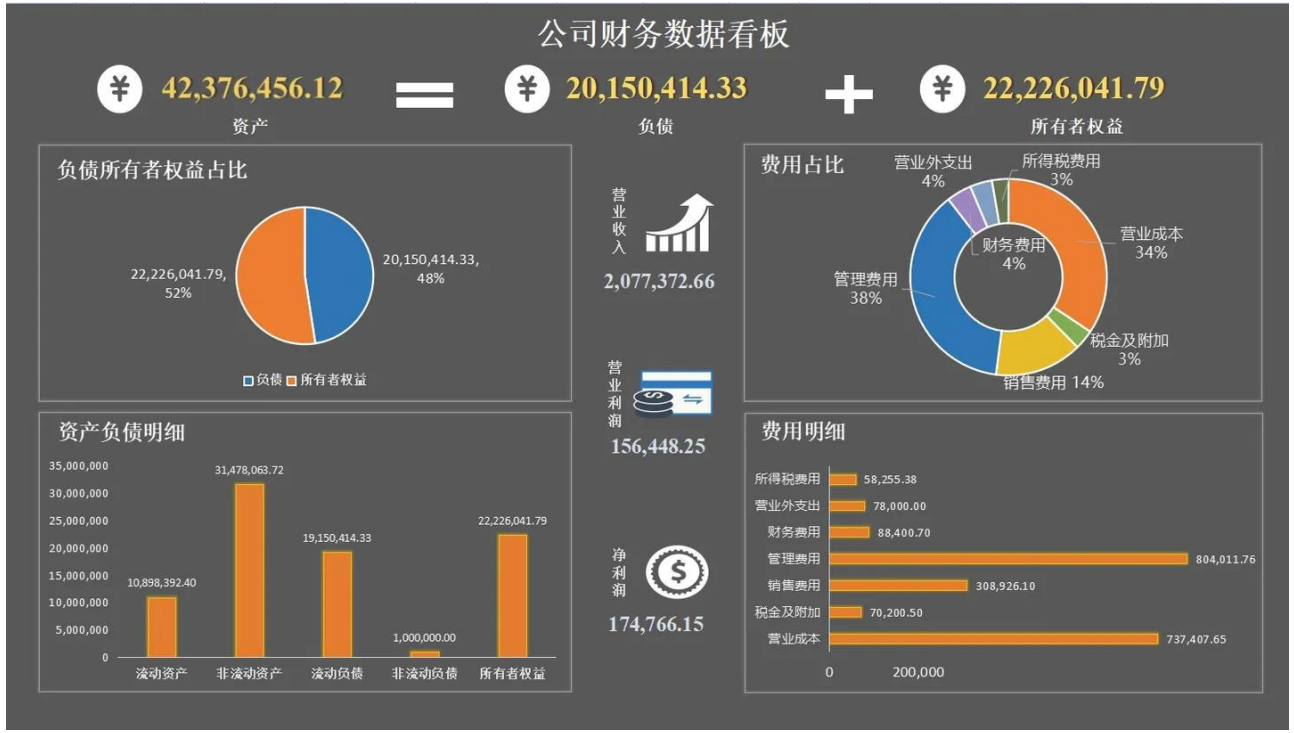


财务数据可视化报表

场景：财务数据可视化报表

数据提取 ⇨ 数据分析 ⇨ BI看板 ⇨ 运营建议





银行数智化面临的挑战

反欺诈

营销：个人或团体的薅羊毛，甚至职业化羊毛党

申请：从个人/企业主体，转变为黑产主体

交易：从一方欺诈演变为更为隐蔽的三方欺诈

授信

贷前信审：难以挖掘边界客户，不利于精准审核/预授信/授信

贷中风控：处理更多数据维度时难以做到实时化的再审核与放款

贷后管理：人工检测难以完成大批量、复杂行为数据下的风险状态评估

运营

操作风险：人员操作、制度漏洞、监管缺失等造成的操作风险，迫切需要从运营、数据安全等层面加强对风险管控

流动性风险：银行产品竞争日趋激烈，风险的实时识别、资金流入流出的预测成为新挑战

制造行业应用场景

螺栓扭矩曲线识别与缺陷检测

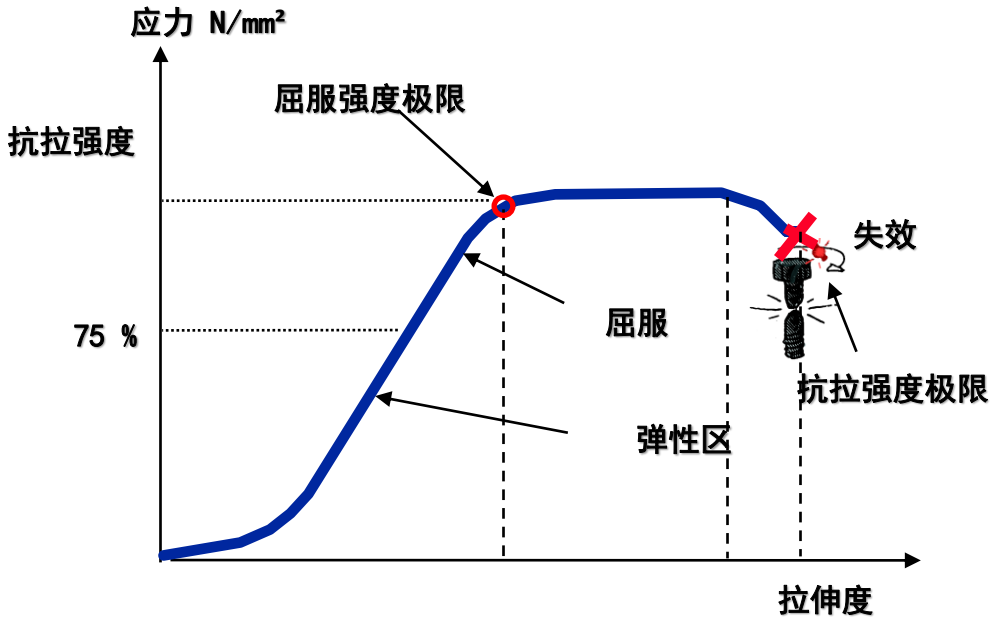
- 业务场景：

螺栓是否拧紧对于汽车安全至关重要

现在的方法：技术人员只检测最终点是否在某个范围之内

可能的问题：

实际上，螺栓需要经过一个平台期，再最终达到规定的范围窗口内，安全质量会更高。否则只检测最终点，而没有达到平台期，可能多年之后会有安全隐患



螺栓扭矩曲线识别与缺陷检测

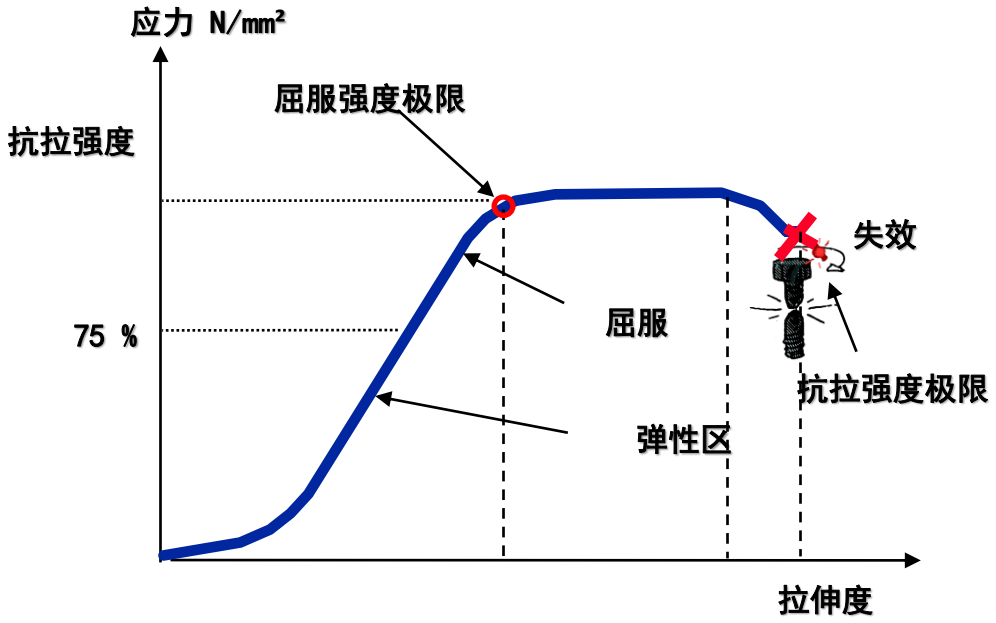
- 提出的数据分析解决方案：

对完整的扭矩曲线进行分析，检测：

1）是否存在平台期，以及平台期的长度

2）是否在平台期之前，存在曲线抖动，计算抖动的次数（扭矩爬坡的抖动情况也是螺栓拧紧质量的重要特征）

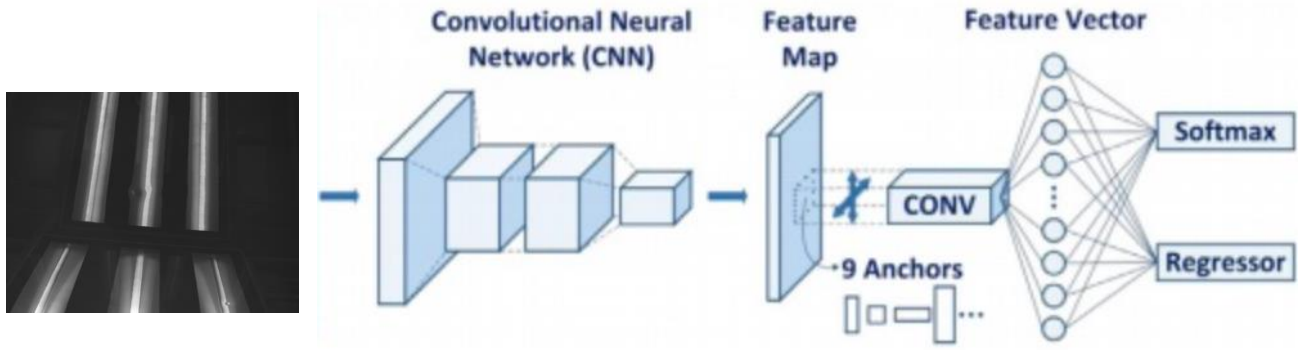
通过数据分析可以提取螺栓扭矩中的重要特征，从而对原有的螺栓拧紧标准进行更新（之前只对最终点设立标准）

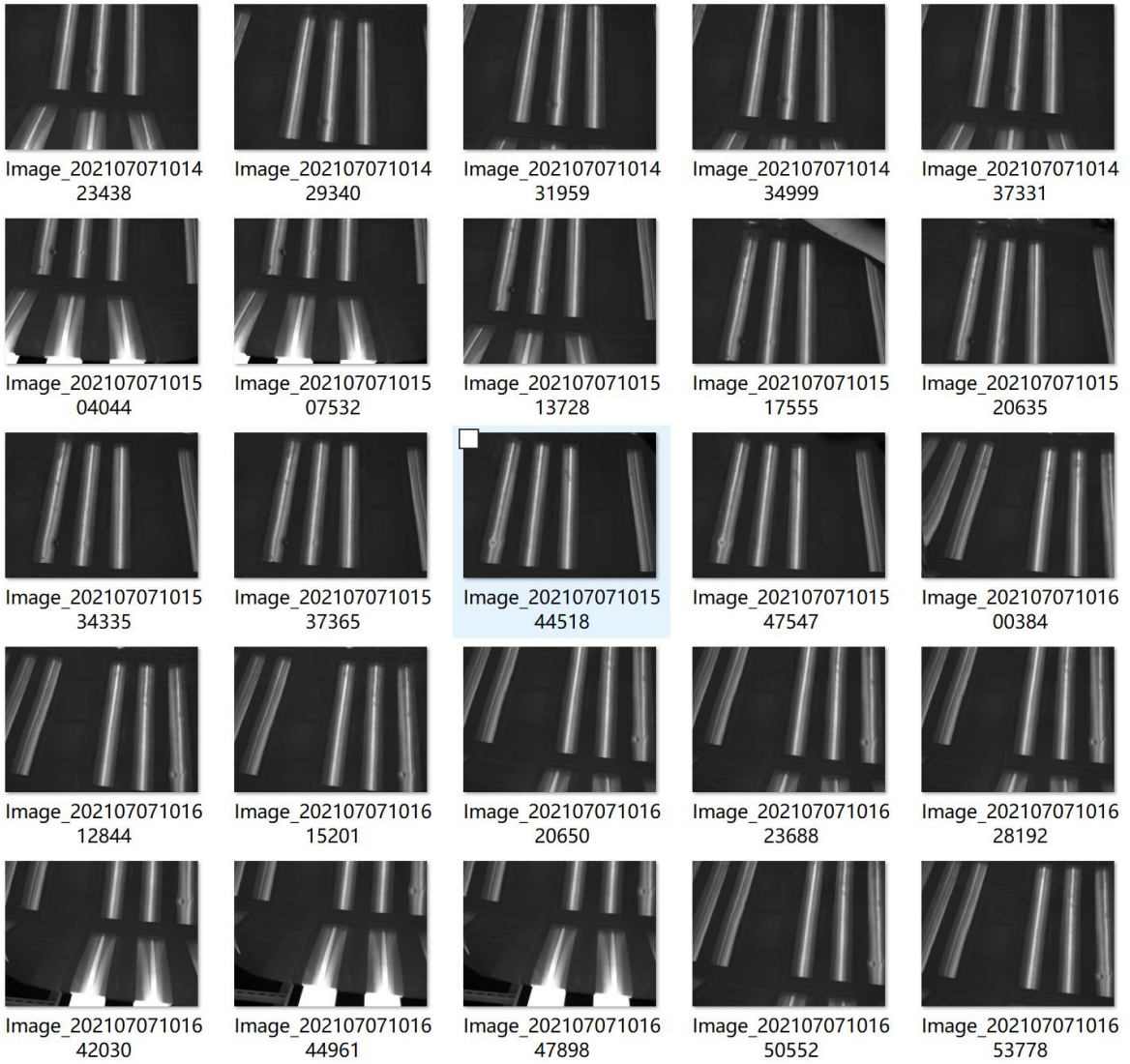


车身表面缺陷检测

缺陷检测：

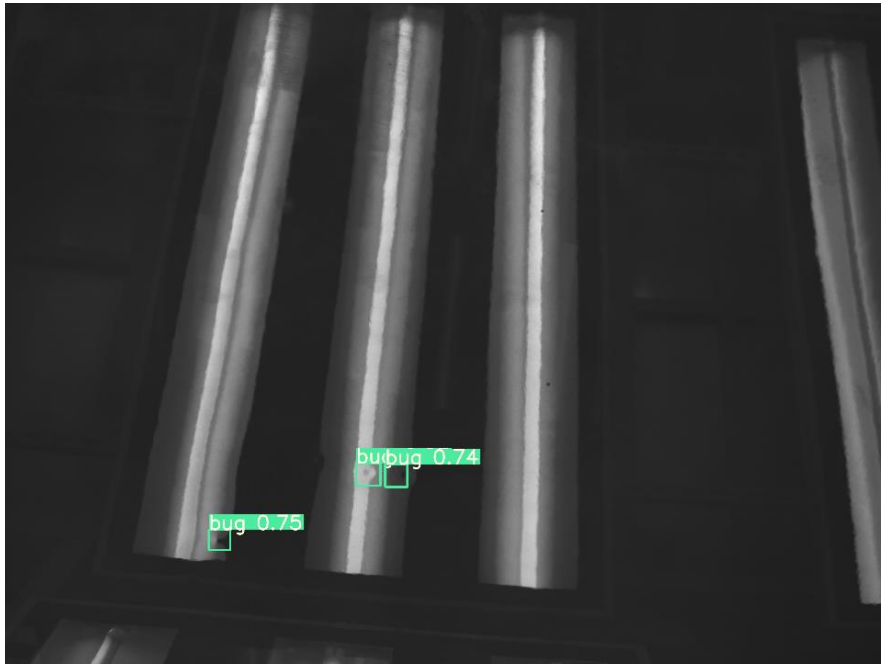
- 缺陷样本少的情况 => 机器视觉算法
- 缺陷样本多的情况 => 神经网络算法





车身表面缺陷检测

机器视觉算法



神经网络算法

快消行业应用场景

什么时候有个精准的预估？？？

业务对于未来门店订货的期望

市场角度

降低缺货损失

减轻清货压力

更加精准的预估

营运角度

提升订货效率

接货前库房存量低

AI应用场景思考

在日常工作中，您认为有哪些分析式AI与生成式AI的应用场景

- 客户流失分析与预警
- 因客定价
- 长尾客群营销
- 贷款商机挖掘
- 商圈生意机会地图
- 基于评分卡的风控模型开发
- 期货套利模型
- 资金流入流出预测
- 对公授信客户逾期分析
- 百万客群经营
- 反洗钱
- 贷款集中度分析
- 手机银行贷款页面优化
- 贷款质量迁徙率与客户标签分析
- 专精特新客群营销
- 公司/机构财报分析
- 营业厅智能推荐
- 缺陷检测
- 供应链补货预测

……

AI大赛：二手车价格预测（进阶）

Project：二手车交易价格预测

- 数据集：

used_car_train_20200313.csv

used_car_testB_20200313.csv

数据来自某交易平台的二手车交易记录

https://tianchi.aliyun.com/competition/ntrance/231784/introduction

ToDo：给你一辆车的各个属性（除了price字段），预测它的价格

Field	Description
SaleID	交易ID，唯一编码
name	汽车交易名称，已脱敏
regDate	汽车注册日期，例如20160101，2016年01月01日
model	车型编码，已脱敏
brand	汽车品牌，已脱敏
bodyType	车身类型：豪华轿车：0，微型车：1，厢型车：2，大巴车：3，敞篷车：4，双门汽车：5，商务车：6，搅拌车：7
fuelType	燃油类型：汽油：0，柴油：1，液化石油气：2，天然气：3，混合动力：4，其他：5，电动：6
gearbox	变速箱：手动：0，自动：1
power	发动机功率：范围 [0, 600]
kilometer	汽车已行驶公里，单位万km
notRepairedDamage	汽车有尚未修复的损坏：是：0，否：1
regionCode	地区编码，已脱敏
seller	销售方：个体：0，非个体：1
offerType	报价类型：提供：0，请求：1
creatDate	汽车上线时间，即开始售卖时间
price	二手车交易价格（预测目标）
v系列特征	匿名特征，包含v0-14在内15个匿名特征

Project：二手车交易价格预测

- 评价标准MAE(Mean Absolute Error)：

若真实值为$y=(y_1,y_2,\cdots,y_n)$，模型的预测值为$\hat{y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n)$，那么该模型的MAE计算公式为

$$MAE=\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n}.$$

例如，真实值$y=(15,20,12)$，预测值$\hat{y}=(17,24,9)$，那么这个预测结果的MAE为

$$MAE=\frac{|15-17|+|20-24|+|12-9|}{3}=3.$$

MAE是L1 loss，MAE越小模型越准确

- 提交结果，与sample_submit.csv中的格式一致

```
SaleID,price
150000,687
150001,1250
150002,2580
150003,1178
```

Project：二手车交易价格预测

- Step1，数据加载

原始数据是用空格分隔

- Step2，数据探索

数据整体情况

查看缺失值，缺失值可视化

查看label的分布（该项目中price为label）

- Step3，特征选择

- Step4，模型训练

使用XGBoost，超参数设置

祖传参数：

```
model = xgb.XGBRegressor(
        n_estimator=2000, learning_rate=0.1,
        subsample=0.8, colsample_bytree=0.8,
        min_child_samples=3, eval_metric='auc',
reg_lambda=0.5,
        max_depth=15, tree_method='gpu_hist' )
```

- Step5，模型预测

Project：二手车交易价格预测

Thinking：如何进行特征工程？

1. 数据预处理

加载训练集和测试集数据，合并数据，便于统一进行特征工程。

2. 特征工程

- 创建时间特征（如车龄、注册季节等）。
- 创建车辆特征（如功率、品牌-车型组合、异常值处理等）。
- 创建统计特征（如品牌均价、品牌价格比等）。
- 编码分类特征（如频率编码、CatBoost 分类特征标记）。

3. 特征选择与数据准备

- 删除无用特征：SaleID, name, offerType, seller（所有车offerType=0,seller只有1个为1，其他都为0）
- 确认所有分类特征被正确标记。
- 转换分类特征为 category 类型。

Project：二手车交易价格预测

```
def create_time_features(data):
    print("创建时间特征...")
    # 转换日期格式
    data['regDate'] = pd.to_datetime(data['regDate'], format='%Y%m%d', errors='coerce')
    data['creatDate'] = pd.to_datetime(data['creatDate'], format='%Y%m%d', errors='coerce')
    # 处理无效日期
    data.loc[data['regDate'].isnull(), 'regDate'] = pd.to_datetime('20160101', format='%Y%m%d')
    data.loc[data['creatDate'].isnull(), 'creatDate'] = pd.to_datetime('20160101', format='%Y%m%d')


    # 车辆年龄（天数）
    data['vehicle_age_days'] = (data['creatDate'] - data['regDate']).dt.days
    # 修复异常值
    data.loc[data['vehicle_age_days'] < 0, 'vehicle_age_days'] = 0
    # 车辆年龄（年）
    data['vehicle_age_years'] = data['vehicle_age_days'] / 365
```

Project：二手车交易价格预测

```
# 注册年份和月份
data['reg_year'] = data['regDate'].dt.year
data['reg_month'] = data['regDate'].dt.month
data['reg_day'] = data['regDate'].dt.day
# 创建年份和月份
data['creat_year'] = data['creatDate'].dt.year
data['creat_month'] = data['creatDate'].dt.month
data['creat_day'] = data['creatDate'].dt.day
# 是否为新车（使用年限<1年）
data['is_new_car'] = (data['vehicle_age_years'] < 1).astype(int)
# 季节特征
data['reg_season'] = data['reg_month'].apply(lambda x: (x%12 + 3)//3)
data['creat_season'] = data['creat_month'].apply(lambda x: (x%12 + 3)//3)
# 每年行驶的公里数
data['km_per_year'] = data['kilometer'] / (data['vehicle_age_years'] + 0.1)
```

```
# 车龄分段
data['age_segment'] = pd.cut(data['vehicle_age_years'],
                             bins=[0, 1, 3, 5, 10, 100],
                             labels=['0-1年', '1-3年', '3-5年', '5-10年', '10年以上'])
return data
```

Project：二手车交易价格预测

Thinking：时间特征对建模的重要性？

- 反映业务本质

时间特征往往与业务场景密切相关。例如在二手车价格预测中，车辆的使用年限（车龄）直接影响其残值，通常车龄越大，价格越低。

- 揭示数据变化趋势

时间特征可以帮助模型捕捉数据随时间变化的趋势和周期性。例如某些车型在特定年份或季节更受欢迎，价格可能更高。

- 辅助异常检测

时间特征有助于发现数据中的异常点，比如注册日期晚于创建日期、车龄为负等，这些异常会影响模型效果。

Project：二手车交易价格预测

```
def create_car_features(data):
    print("创建车辆特征...")
    # 缺失值处理
    numerical_features = ['power', 'kilometer', 'v_0', 'v_1', 'v_2', 'v_3', 'v_4', 'v_5', 'v_6', 'v_7', 'v_8', 'v_9', 'v_10', 'v_11', 'v_12', 'v_13', 'v_14']
    for feature in numerical_features:
        # 标记缺失值
        data[f'{feature}_missing'] = data[feature].isnull().astype(int)
        # 填充缺失值
        data[feature] = data[feature].fillna(data[feature].median())

    # 将model转换为数值型特征
    data['model_num'] = data['model'].astype('category').cat.codes
    # 品牌与车型组合
    data['brand_model'] = data['brand'].astype(str) + '_' + data['model'].astype(str)
```

对每个数值型特征，先生成一个“是否缺失”的二值特征，再用中位数填充缺失值。

缺失本身可能蕴含信息（如某些车型的某项参数经常缺失），模型可以利用“缺失标记”捕捉这种隐含信号。

品牌与车型组合，细化区分不同细分市场

Project：二手车交易价格预测

```
# 相对年份特征
current_year = datetime.datetime.now().year
data['car_age_from_now'] = current_year - data['reg_year']


# 处理异常值
numerical_cols = ['power', 'kilometer', 'v_0']
for col in numerical_cols:
    Q1 = data[col].quantile(0.05)
    Q3 = data[col].quantile(0.95)
    IQR = Q3 - Q1
    data[f'{col}_outlier'] = ((data[col] < (Q1 - 1.5 * IQR)) | (data[col] > (Q3 + 1.5 * IQR))).astype(int)
    data[col] = data[col].clip(Q1 - 1.5 * IQR, Q3 + 1.5 * IQR)


return data
```

对关键数值特征（如功率、公里数、排量）进行异常值检测，并生成异常标记，同时对异常值进行截断处理。

异常值往往会影响模型的稳定性，通过标记和修正，既保留了异常信息，又减少了极端值对模型的干扰。

Project：二手车交易价格预测

```
def create_statistical_features(data, train_idx):
    print("创建统计特征...")
    # 仅使用训练集数据创建统计特征
    train_data = data.iloc[train_idx].reset_index(drop=True)
    # 品牌级别统计
    brand_stats = train_data.groupby('brand').agg(
        brand_price_mean=('price', 'mean'),
        brand_price_median=('price', 'median'),
        brand_price_std=('price', 'std'),
        brand_price_count=('price', 'count')
    ).reset_index()
    # 合并统计特征
    data = data.merge(brand_stats, on='brand', how='left')
    # 相对价格特征（相对于平均价格）
    data['brand_price_ratio'] = data['brand_price_mean'] / data['brand_price_mean'].mean()
    return data
```

只用训练集数据做统计，避免把测试集泄漏到训练过程

对每个品牌，统计其在训练集中的平均价格、中位数、标准差、样本数。

品牌均价/中位价：反映该品牌车辆的市场定位和价值区间。

品牌价格标准差：反映该品牌车型价格的离散程度，间接体现品牌丰富度或车型跨度。

品牌样本数：衡量品牌在数据集中的流行度。

Project：二手车交易价格预测

```
def encode_categorical_features(data):
    print("编码分类特征...")
    # 目标编码的替代方案 - 频率编码
    categorical_cols = ['model', 'brand', 'bodyType', 'fuelType', 'gearbox', 'notRepairedDamage']
    for col in categorical_cols:
        # 填充缺失值
        data[col] = data[col].fillna('未知')
        # 频率编码
        freq_encoding = data.groupby(col).size() / len(data)
        data[f'{col}_freq'] = data[col].map(freq_encoding)

    # 将分类变量转换为CatBoost可以识别的格式
    for col in categorical_cols:
        data[col] = data[col].astype('str')

    return data, categorical_cols
```

将所有分类特征的缺失值统一填充为 ’未知’，
避免 NaN 导致模型报错。

频率编码能反映某一类别的“流行度”或“稀有度”，有助于模型捕捉类别分布信息。

对高基数类别（如车型、品牌）尤其有效，能缓解 one-hot 编码维度爆炸的问题。

频率编码是目标编码的无监督替代方案，避免了数据泄漏风险。

Project：二手车交易价格预测

```
def feature_selection(data, categorical_cols):
    print("特征选择和最终数据准备...")
    # 删除不再需要的列, 所有车offerType=0,seller只有1个为1，其他都为0
    drop_cols = ['regDate', 'creatDate', 'price', 'SaleID', 'name', 'offerType', 'seller', 'source']
    data = data.drop(drop_cols, axis=1, errors='ignore')


    # 确保所有分类特征都被正确标记
    if 'age_segment' not in categorical_cols and 'age_segment' in data.columns:
        categorical_cols.append('age_segment')
    if 'brand_model' not in categorical_cols and 'brand_model' in data.columns:
        categorical_cols.append('brand_model')
    # 转换分类特征
    for col in categorical_cols:
        if col in data.columns:
            data[col] = data[col].astype('category')
    return data, categorical_cols
```

确保衍生出来的分段特征（如车龄分段、品牌-车型组合）被纳入分类特征列表。这些衍生特征往往有很强的业务意义，能提升模型对类别信息的理解。

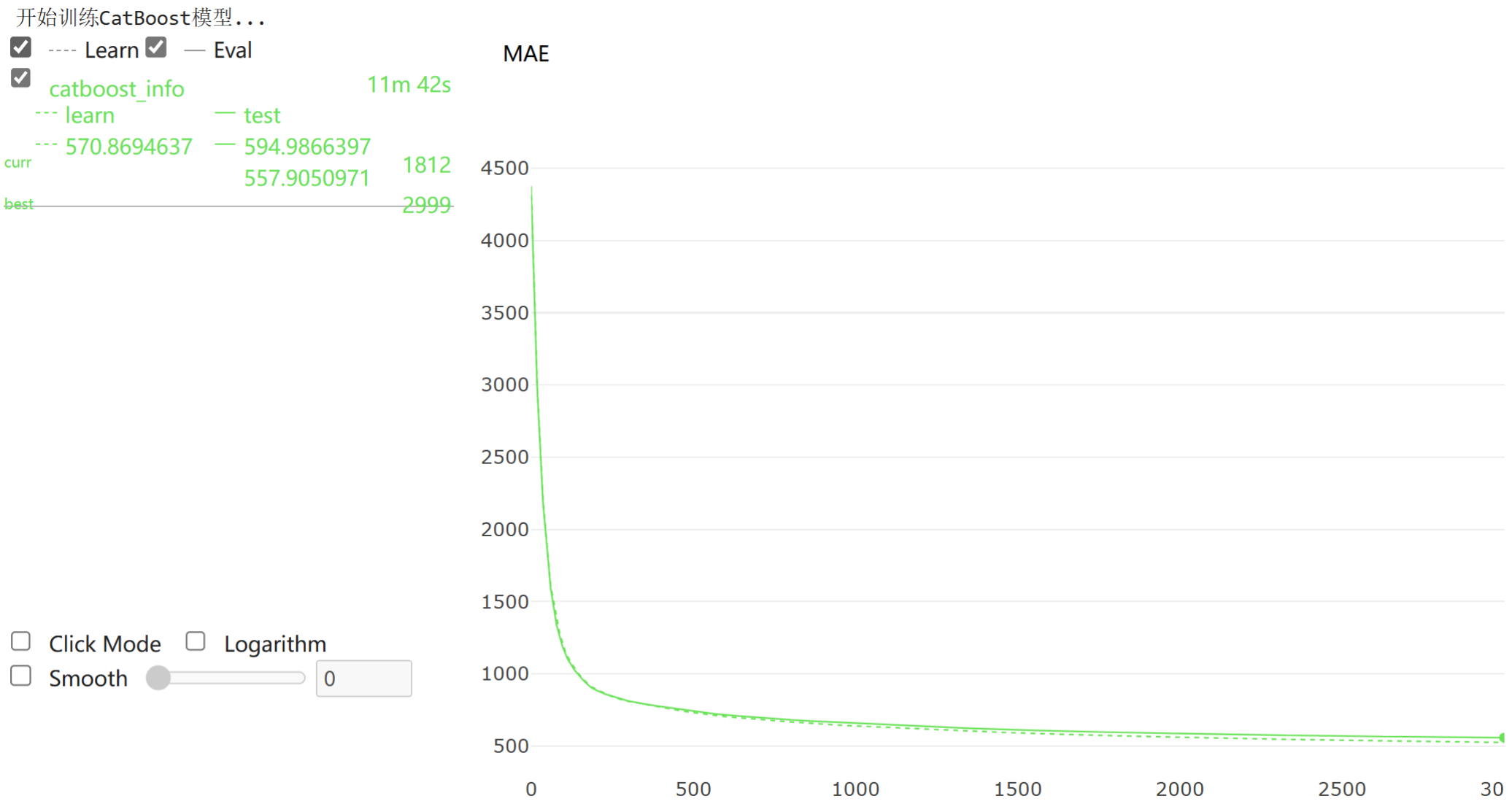
category 类型不仅节省内存，还能让 CatBoost 等模型自动识别并高效处理类别特征。

CatBoost 能自动学习类别特征之间的高阶交互（如品牌+车型+车龄分段），无需手动构造大量交叉特征。

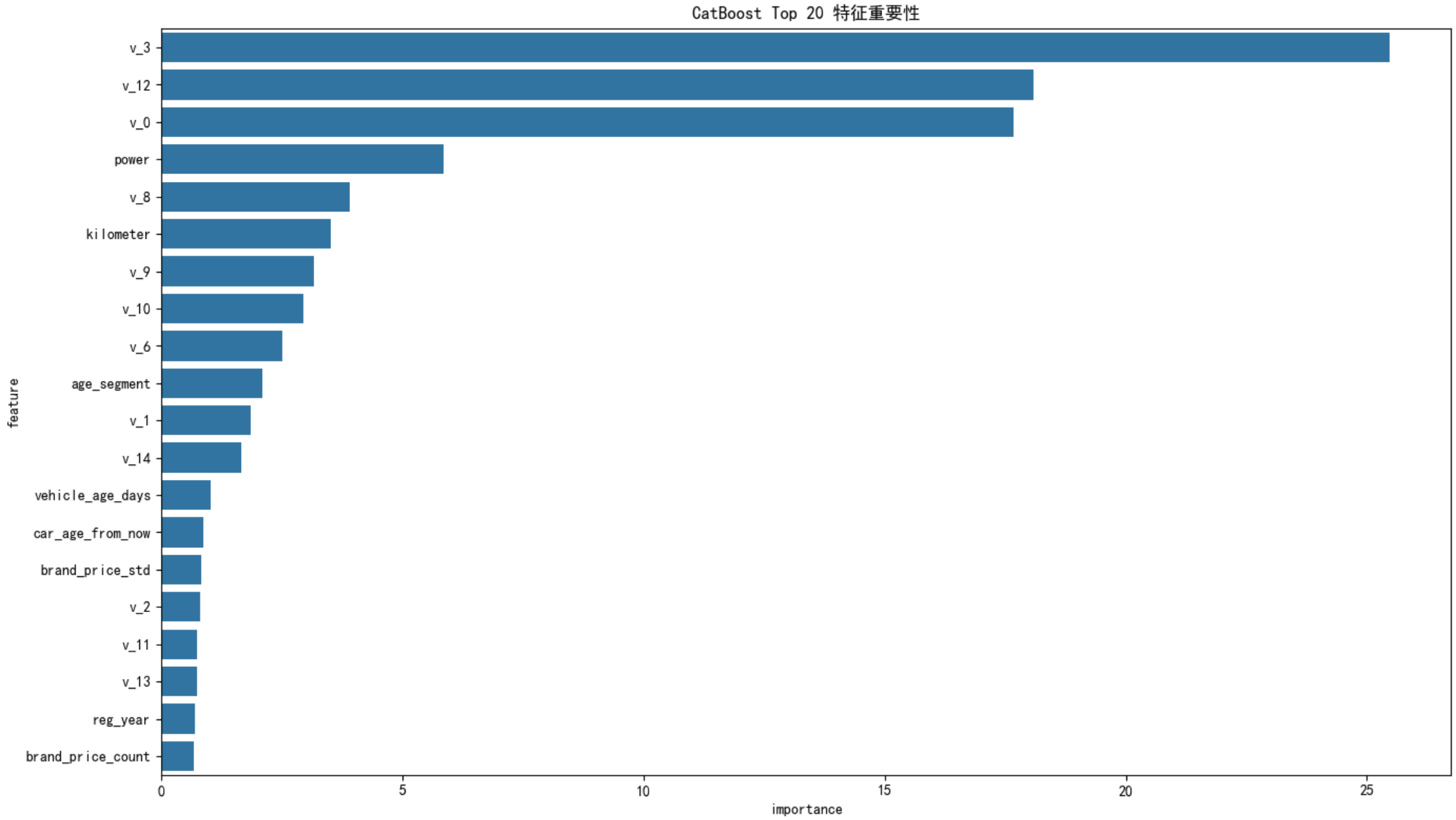
Project：二手车交易价格预测

```
def train_catboost_model(X_train, X_val, y_train, y_val, cat_features):
    print("\n开始训练CatBoost模型...")
    # 创建数据池
    train_pool = Pool(X_train, y_train, cat_features=cat_features)
    val_pool = Pool(X_val, y_val, cat_features=cat_features)
    # 设置模型参数
    params = {
        'iterations': 3000,
        'learning_rate': 0.03,
        'depth': 6,
        'l2_leaf_reg': 3,
        'bootstrap_type': 'Bayesian',
        'random_seed': 42,
        'od_type': 'Iter',
        'od_wait': 100,
        'verbose': 100,
        'loss_function': 'MAE',
        'eval_metric': 'MAE',
        'task_type': 'CPU',
        'thread_count': -1
    }
    # 创建模型
    model = CatBoostRegressor(**params)
    # 训练模型
    model.fit(
        train_pool,
        eval_set=val_pool,
        use_best_model=True,
        plot=True
    )
    model.save_model('processed_data/fe_catboost_model.cbm')
    return model
```

Project：二手车交易价格预测



Project：二手车交易价格预测



Summary（特征工程）

《二手车价格预测》项目中的特征工程：

1. 时间特征处理

日期格式转换：将注册日期、创建日期转为 day_diff，或者年、月、日等衍生特征。

衍生时间特征：如车龄（天/年）、注册/创建的年、月、日、季节、是否新车、每年行驶公里数、车龄分段等。

异常值修正：如注册日期晚于创建日期的情况归为合理默认值。

2. 车辆属性特征

缺失值处理与标记：对数值型特征填充中位数，并生成缺失标记特征。

特征交互与衍生：如品牌-车型组合、车辆绝对年龄等。

异常值检测与修正：对功率、公里数、排量等做异常值标记和截断，减少极端值影响。

Summary（特征工程）

3. 统计特征

品牌级别统计：基于训练集，统计每个品牌的均价、中位价、标准差、样本数。

相对价格特征：如品牌均价与全体均价的比值，反映品牌市场定位。

防止数据泄漏：统计特征只用训练集数据计算，保证评估公正性。

4. 分类特征编码

缺失值统一填充：所有分类特征缺失值填充为 '未知'。

频率编码：为每个类别特征生成其出现频率的数值特征，反映类别流行度。

类型转换：所有分类特征转为字符串或 category 类型，便于 CatBoost 原生处理。

5. 特征选择与类型标准化

删除无用特征：如原始日期、ID、价格、冗余字段等。

补充衍生类别特征：如车龄分段、品牌-车型组合等，纳入分类特征列表。

类型标准化：所有分类特征转为 category 类型，节省内存并提升模型兼容性。

Summary（特征工程）

Thinking：特征工程的价值？

特征工程决策模型上限，而模型（调参）只是把上限跑出来而已。

丰富的数据表达：通过多维度、多层次的特征衍生，极大提升了模型对业务本质的理解能力。

高效的类别特征利用：充分发挥 CatBoost 对类别特征的原生支持，避免 one-hot 维度爆炸，提升模型性能。

异常与缺失处理：通过缺失标记和异常值修正，提升模型鲁棒性和泛化能力。

Cursor进行特征工程，存在概率性的情况 => 可能处理会有问题，比如：

将 power+model进行数值相加，实际上没有物理含义，增加了噪音。

没有将训练集和测试集合并，单独处理的特征工程，可能测试集的编码缺失的情况

对kilometers的分箱可能是 [0, 5, 10, 20]，并不能反应整体的数据集的情况等

=> 需要从业务角度，对特征工程进行验证

打卡：二手车价格预测

针对AI大赛：二手车价格预测，编写AI算法，进行预测，挑战分数 < 550

https://tianchi.aliyun.com/competition/entrance/231784/information

训练集： used_car_train_20200313.csv

测试集： used_car_testB_20200421.csv

- 选择适合的模型
- 特征工程

时间特征处理

车辆属性特征

统计特征

分类特征编码

异常值处理


Thank You
Using data to solve problems