



慧融宝

慧融宝

INTELLIGENT FINANCIAL TOOL

【金融科技服务】企业数据无监督分类

让小微企业不再是小微企业

二〇二〇年一月二十日，新型冠状病毒于国内爆发

国内经济需求和生产**骤降**，投资、消费受明显**冲击**

大量中小微企业不得不延迟返工甚至停工停产，却仍需支付**高昂**的成本

仅长三角地区，就约有**4328.1户**小微企业面临现金流**吃紧**的危险

不要紧=停业

中小微企业寿命对比

国家	寿命饼图	平均寿命	第一次获得贷款
美国		8年	> 成立二年零九个月
中国		3年	< 成立四年零四个月
日本		12年	> 成立三年零六个月

其中，成立**3年后**的小微企业**还正常营业**的约占**三分之一**

有相当一部分的小微企业只有熬过“死亡期”之后，才能获得贷款

缺口极大

STP分析：市场细分 Segmentation

细分变量：供应链环节需求



源头信贷市场

P2P中介市场

第三方金融辅助市场

上游市场明晰
但进入门槛高

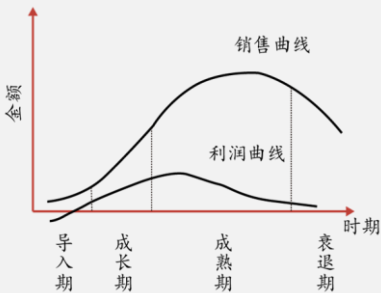
中下游需求可观
但存在体系短板



联结中下游
多方协作，大势所趋

17家民营银行总资产达到
8000亿元
新进入者对原先市场份额
的威胁较小

大量逐利投资者涌入P2P
模式
P2P平台良莠不齐



结合生命周期理论：已经历十年的导入期，处于迅速需求上升爬坡的成长期
竞争积聚程度低而市场容量大

STP分析：目标市场 Targeting

市场定位 Positioning

01




[劣势分析]

- 1) 起步晚——只抓最痛点——尽量放弃源头信贷型市场
- 2) 资源相对匮乏——集中精力——定位第三方金融辅助市场，放弃P2P中介型市场

[优势分析]

- 1) 没有“巨无霸”垄断市场——人人都可以是主角——建立核心优势
- 2) 聚焦“高精度”簇分类——新功能的突破——国家级重点实验室，集聚学术与应用人才
- 3) 公益向左，商业向右——商业功能再突破——社会责任与企业利益同步的可持续发展。

02

	算法精确度	费率成本
市场领先者  百融云创	领先：将近99%	费用较高：按项目收费
市场挑战者  摩羯智投	较高：97.8%	费率高：小投情况下3.5%费率
市场追随者  慧融宝	仍具优势：96.8%	低费率：费率在0.8%~1.2%浮动让利用户
↓ 依托于国家级重点实验室的第三方金融辅助产品 “低费率”、“高精度”、“人性化”		

问题分析

用户类型

直接用户

金融机构
正规信贷公司

间接用户

中小微企业

用户特点

放贷实力强、潜力大但惧贷、恐贷

人工审阅，时间、工作量成本高

数据源有限，缺少全局的数据分析

资产规模小、资金实力较弱

数量庞大，但需求满足率只有两成

资质审查困难、借贷困难

用户问题

问题一：资金流通效率不高，大量机会成本浪费

问题二：识客难、获客难、活客难

问题三：人工审阅导致资质审查、风险识别难

问题四：劣币驱逐良币导致市场信息混乱，信息跟进难

问题五：难以承担显著上浮的信贷成本

问题六：由于自身实力弱大多局限于单一渠道

问题七：疫情期间，短期偿债能力下降，增加了不良贷款的潜在风险

解决思路

对症下药

关键问题	目标对象	方式与途径	关键技术
问题一：资金流通效率低 问题六：融资渠道单一，可得性差	金融机构、公司 中小微企业	寻求一种无监督聚类方法，并训练模型，对中小微企业有效簇划分，并实现毫秒级响应，以此实现机构企业之间的双向配对	K-Means++
问题二：识客难、获客难、活客难 问题三：人为审阅导致风险评估困难 问题七：疫情使不良贷款潜在风险上升	金融机构、公司	选择合适的特征工程进行特征提取，利用PCA降维提高训练模型的效率和模型准确率，降低出错率，助力精准获客	离散型变量处理、分箱、交叉特征、特征选择、Z-Score标准化、PCA降维
问题四：劣币驱逐良币导致市场信息混乱，数据源有限，缺少全局分析	金融机构、公司 中小微企业	原始数据的集中整合，支持批量搜索、模糊搜索，使用可视化界面echarts全局展示	Echarts可视化展示，Redis缓存
问题五：难以承担显著上浮的信贷成本	金融机构、公司 中小微企业	使用springboot框架、阿里云ECS低成本搭建并运营平台，让利消费者	Springboot、云服务

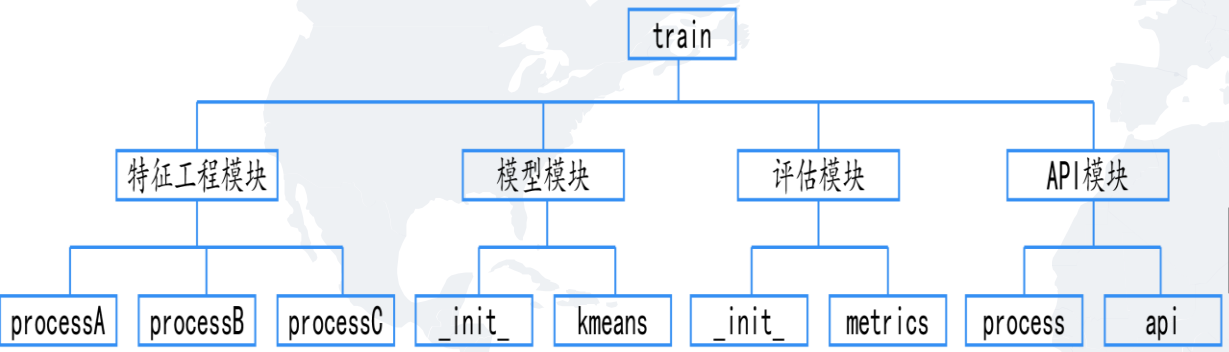
算法需求分析

算法模块

“模块化设计”方便调试和添加功能

算法数据流

严格按照
“数据流向”进行模型构建



时间复杂度



$O(l * n * k * m)$

空间复杂度



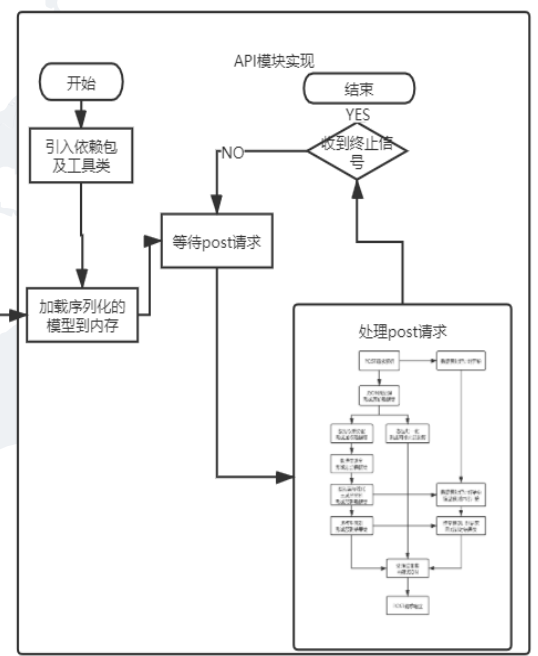
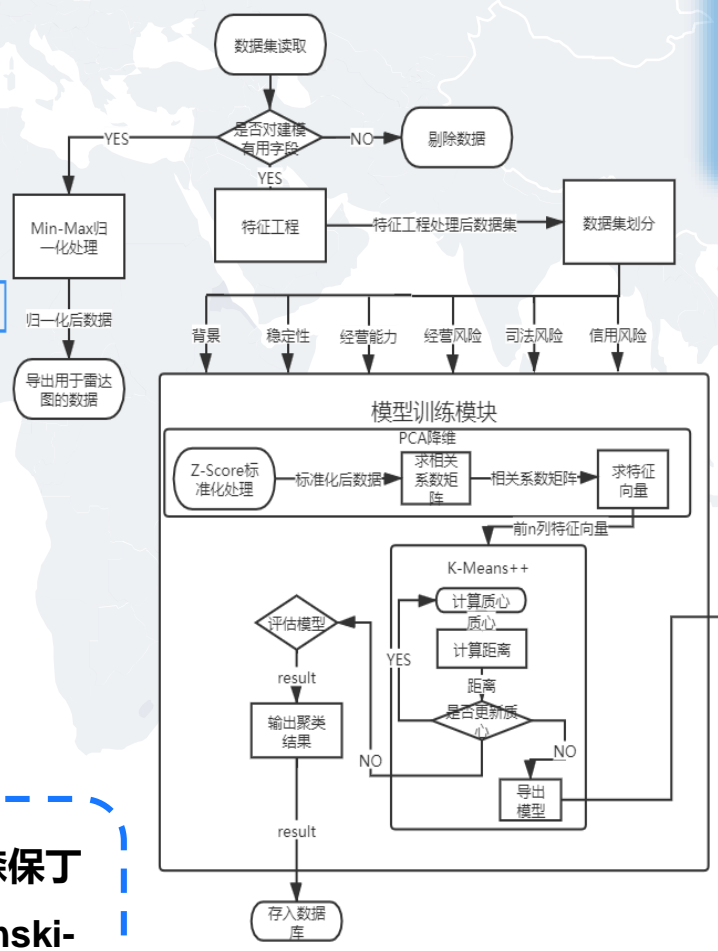
$O(n * m)$

性能评估
及稳定性



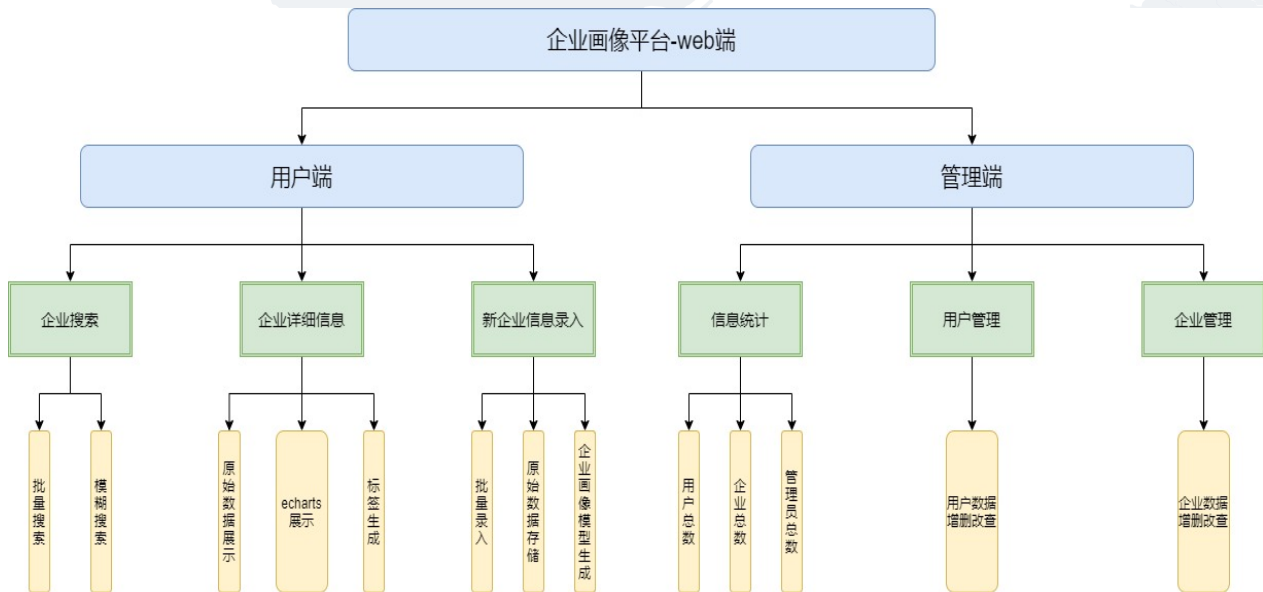
稳定性第一

l 为点与点之间距离的复杂度， n 为点数， k 为聚类中心个数， m 为迭代次数
 n 为点数， m 为迭代次数
轮廓系数接近1.00，戴维森保丁指数稳定在0.5以下，Calinski-Harabasz指数达到十万级

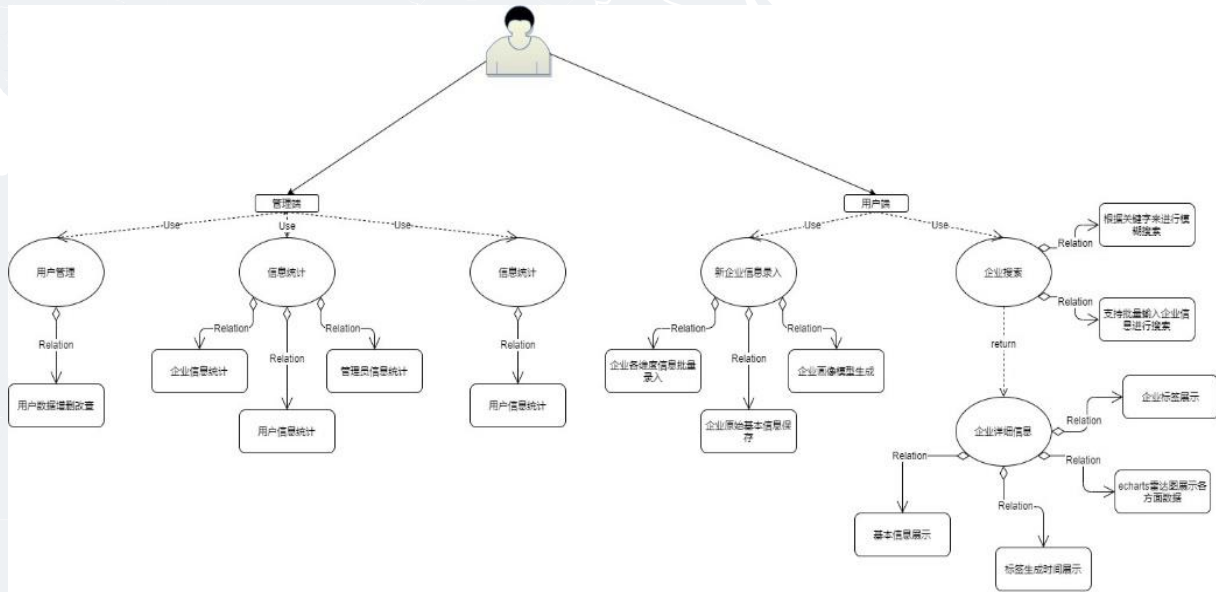


系统设计

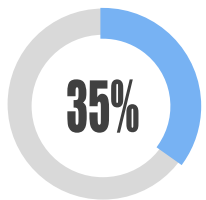
功能结构图



系统用例图

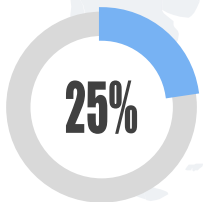


时间性需求



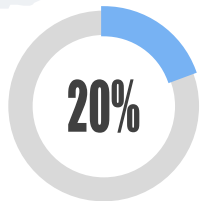
平均响应时间
达到毫秒级

故障处理需求



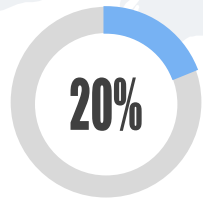
当抛出异常时，则进行降级，信息条目的内容替换为本地静态页面，进行的服务保持原样并进行缓存数据，在重连后进行恢复

并发性需求



估计用户规模为50万，峰值在线人数25万，峰值并发用户数10万

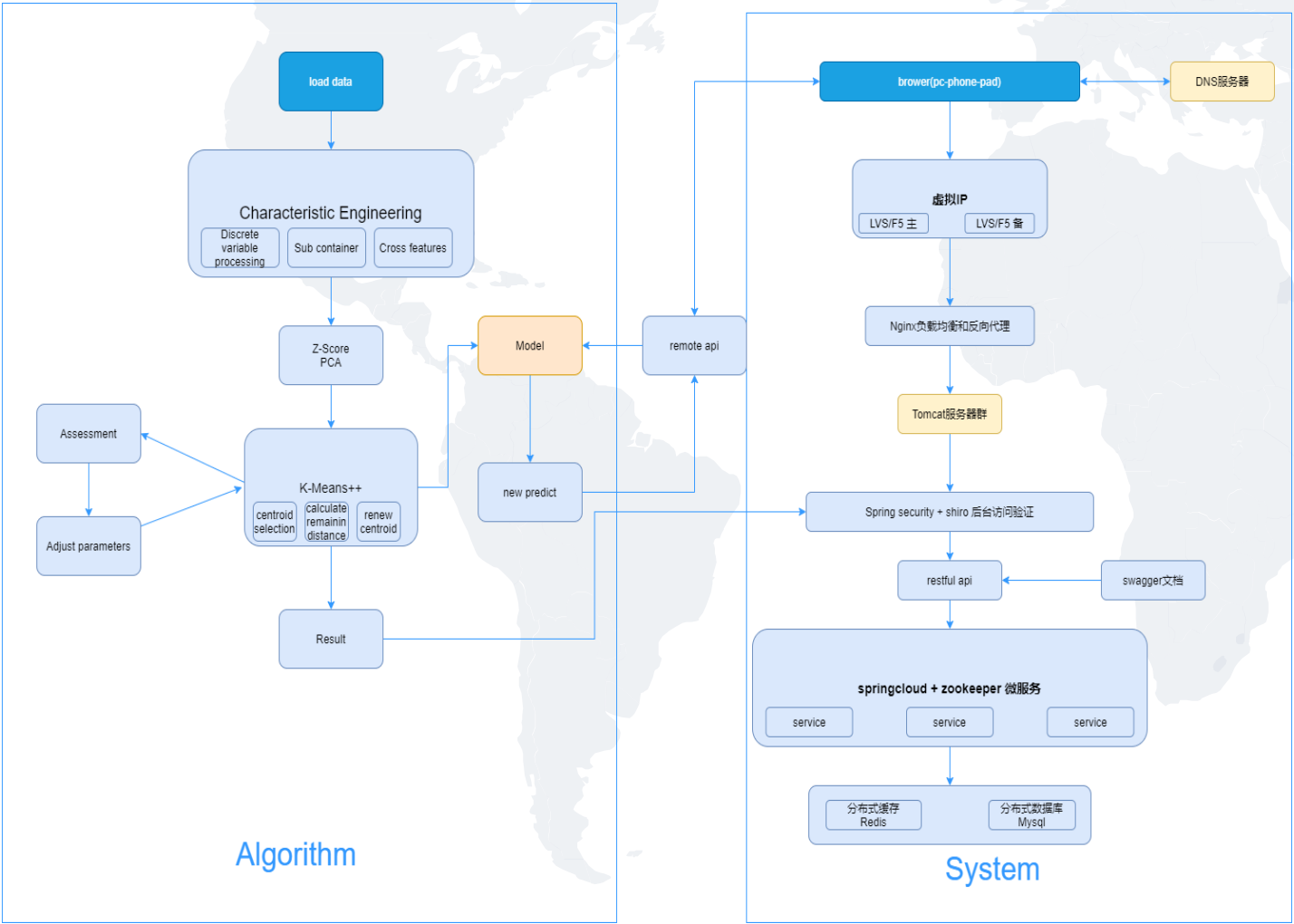
容量需求



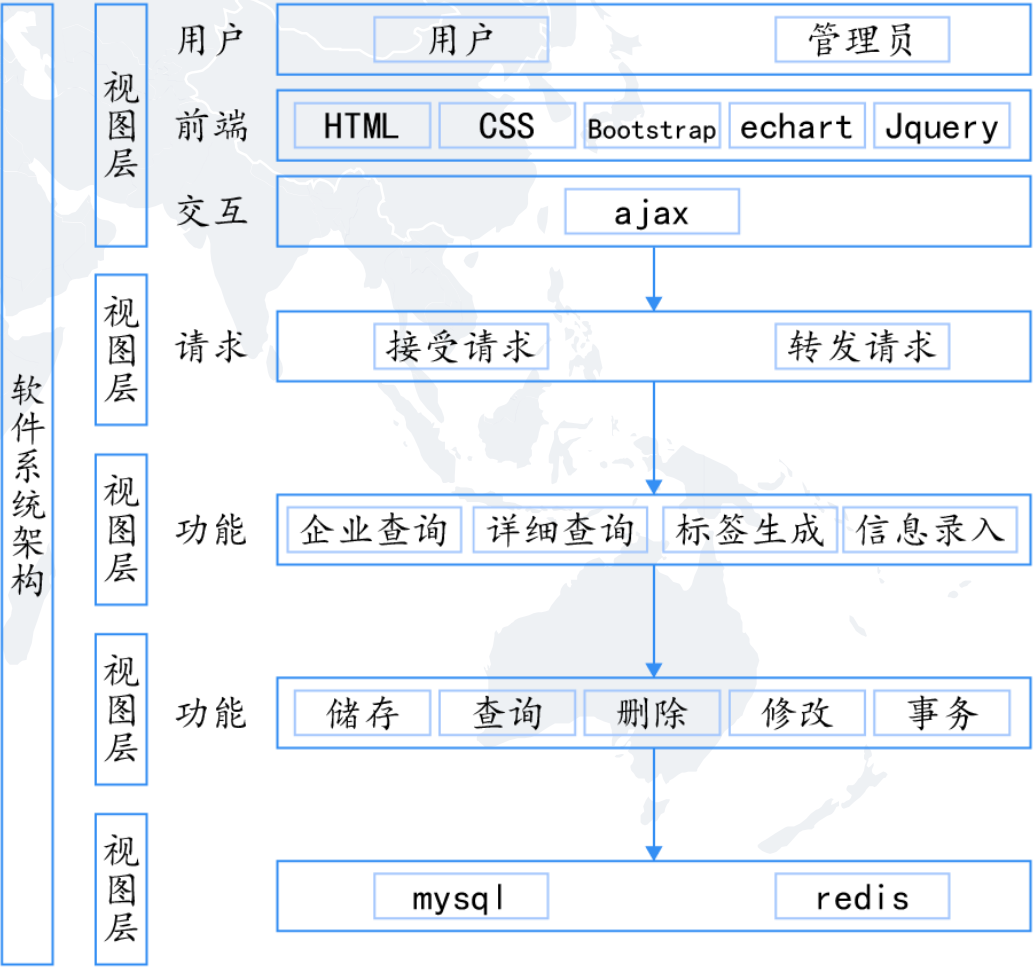
CPU占用率 $\leq 50\%$
内存占用率 $\leq 50\%$
服务器5G带宽

技术路线

技术框架



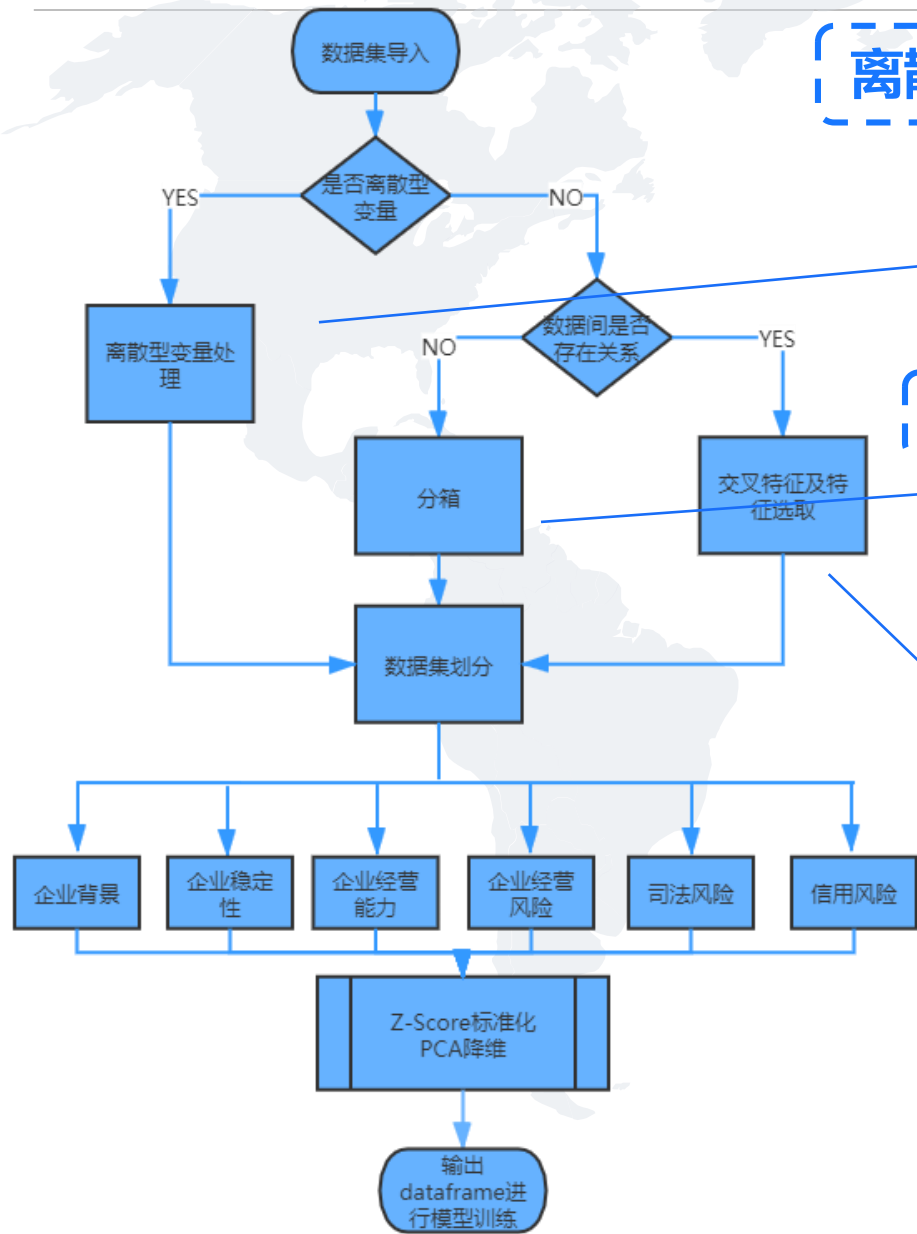
系统架构图



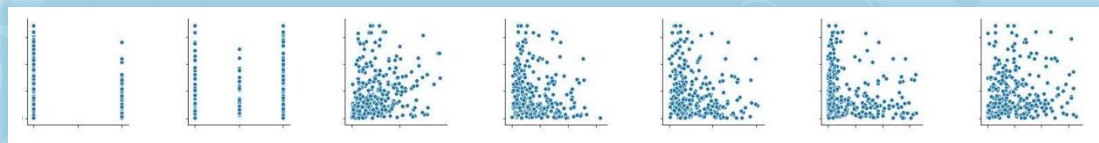
算法实现



特征工程

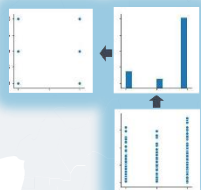


离散型变量处理



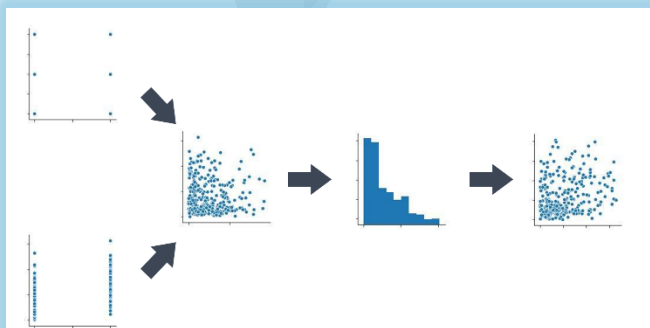
对于**企业状态**这类数据，具有明显的**离散性**，因此，我们根据其状态的不同对其进行了合理的**数值化处理**，从而使得该中文字段能够**正确入模**。

分箱



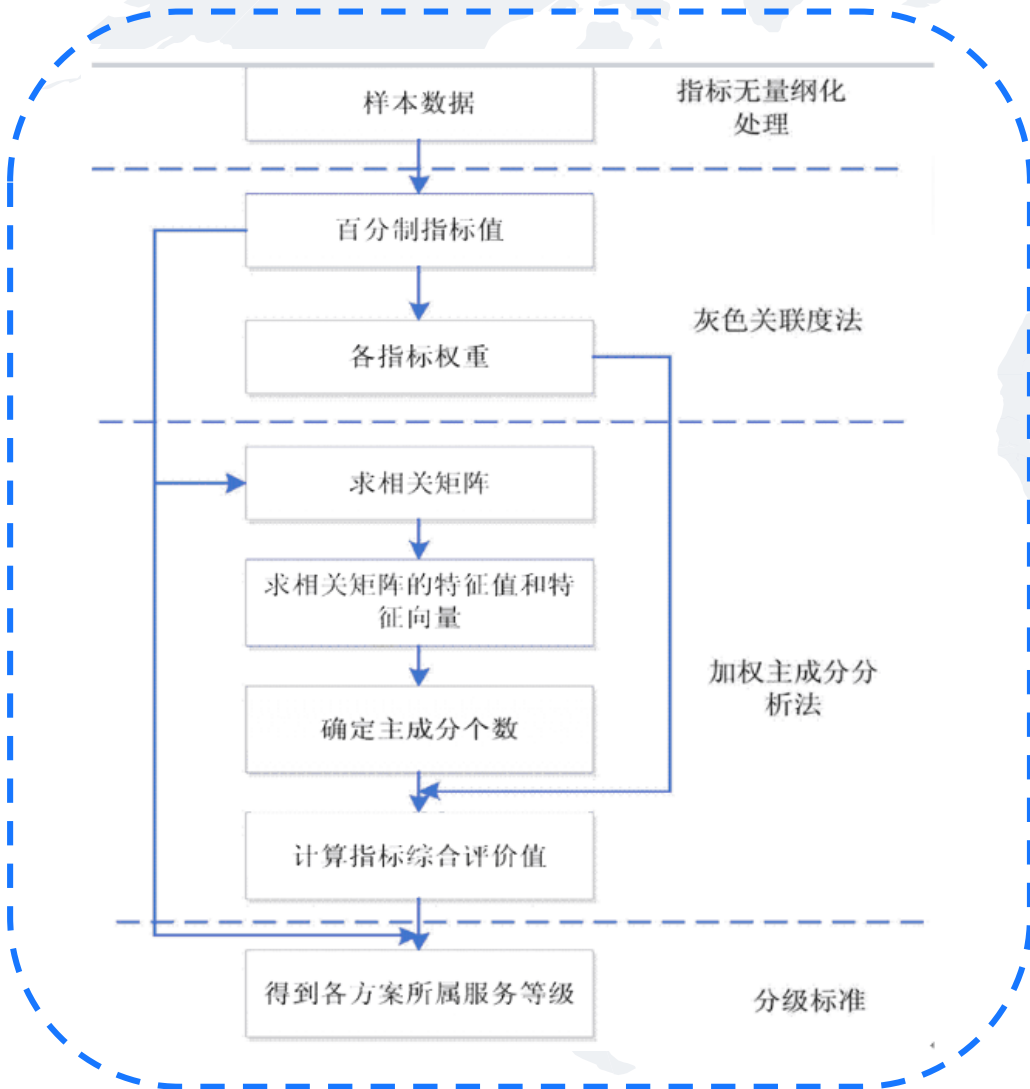
对于**注销时间**和**吊销时间**这类数据，因为只要记录在案就说明企业已被注销或者吊销，因此我们按其有无进行了**分箱操作**，使得模型能够**正确判别**其情况。

交叉特征及特征选取

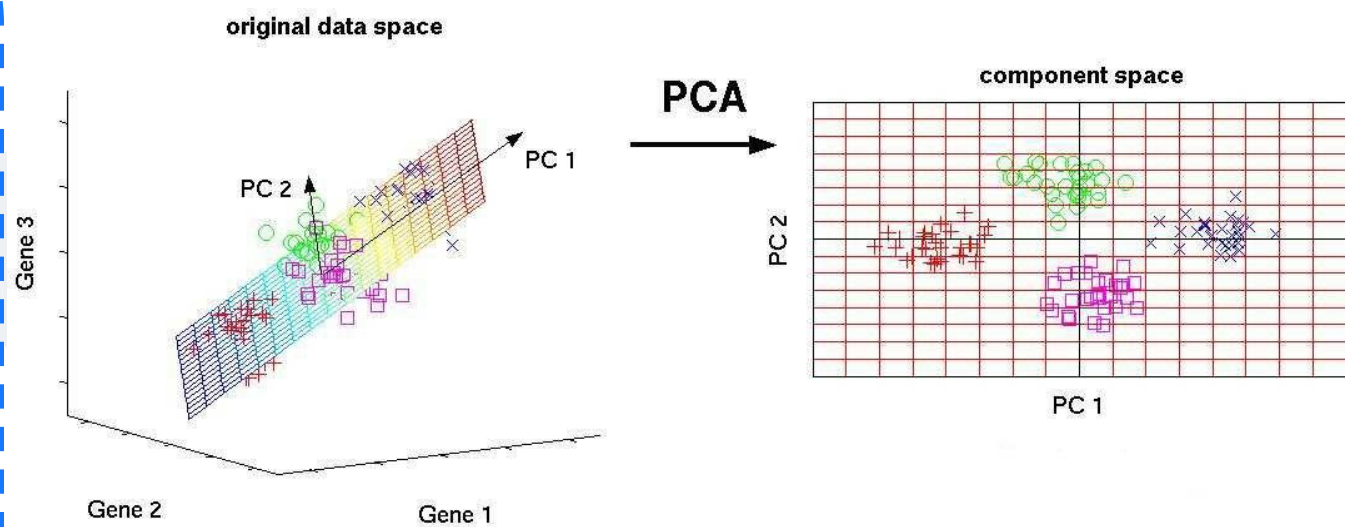


对于**企业年报出资信息**这类数据，其实缴出资时间、认缴出资时间、累计实缴额、累计认缴额之间有着较大的联系，因此我们根据其实缴出资时间和认缴出资时间、累计实缴额和累计认缴额之间的**关系**，将其**划分为不同权重特征**，并作了合理的**数值化操作**，使其可以作为**入模特征**。

主成分分析法降维

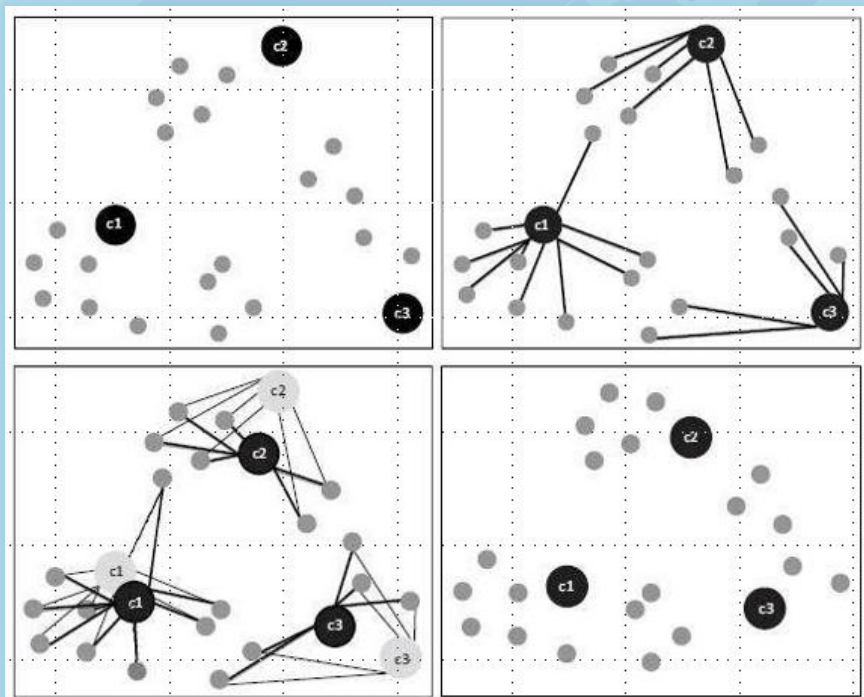


结果示意图



由分布在源空间的数据经过降维后到分布在新的合成空间的数据，使得数据的特征更加明显，有利于去除数据的噪声，使得模型训练更容易收敛，提高效率

核心算法分析——K-Means++算法



1. 随机选取一个样本作为第一个聚类中心 c_1 ，然后计算**每个样本与当前已有类聚中心最短距离**（即与最近一个聚类中心的距离），距离计算公式如下：

$$dist(X, C) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2}$$
$$D(x_i) = \min_{0 \leq j \leq k} dist(x_i, c_j)$$

其中 C 、 c_j 代表聚类中心点， k 表示当前聚类中心的数量， X 、 x_i 表示样本点。

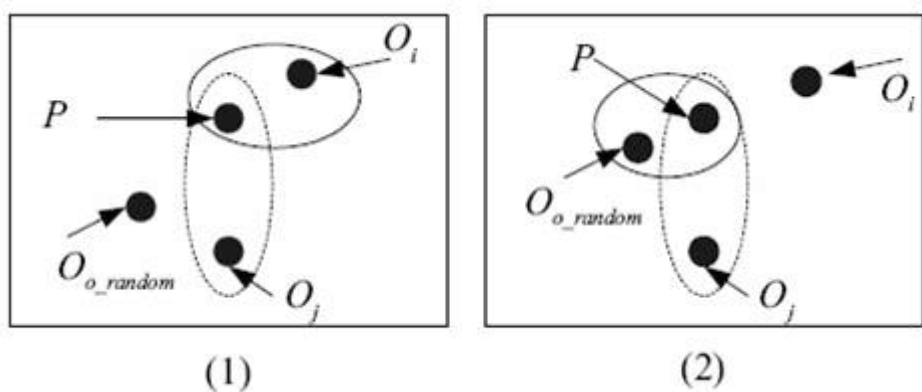
2. 接着计算每个样本被选为**下一个聚类中心的概率**，概率公式如下：

$$P(x_i) = \frac{D(x_i)^2}{\sum_{x_i \in X} D(x_i)^2}$$

这个**值**越大，表示被选取作为**聚类中心**的概率较大。

3. 最后，用**轮盘法**选出下一个聚类中心；

4. 然后重复这个步骤，直到**选出 k 个聚类中心**。这个优化在一定程度上提高了**时间效率**，并且使得最后聚类结果的合理程度也得到一定提升。



实验评估：与贝叶斯高斯混合、高斯混合、自组织映射算法比较

多指标第一

企业背景评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	1.78	1.17	4.34	0.60	65062.58	0:02:56.96
GaussianMixture	1.78	1.17	4.34	0.60	65062.58	0:02:40.60
Som	0.68	0.63	2.15	0.24	0.67	1:01:20.87
K-means++	5.36	0.48	36.22	0.67	264946.03	0:01:50.77

企业司法风险评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	8.57	0.51	158.27	1.00	237136.96	0:01:17.02
GaussianMixture	9.09	0.47	160.09	1.00	288956.69	0:00:44.86
Som	16.29	2.08	15.65	0.99	3807.80	0:59:37.28
K-means++	10.92	0.42	167.64	0.99	379310.38	0:02:51.55

企业经营风险评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	12.18	1.52	27.77	0.88	76013.14	0:04:38.60
GaussianMixture	3.08	1.68	4.84	0.88	22401.35	0:04:37.44
Som	12.12	27.66	21.48	0.55	5228.77	1:09:57.18
K-means++	6.95	0.39	52.30	0.90	401711.72	0:04:55.00

企业稳定性评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	0.64	0.67	2.07	0.60	72245.37	0:00:47.22
GaussianMixture	0.69	0.65	2.31	0.59	82740.20	0:00:17.52
Som	0.18	3.80	0.09	-0.01	0.01	0:58:15.83
K-means++	2.04	0.51	11.31	0.89	362918.39	0:00:04.77

企业经营能力评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	1.60	0.95	5.04	0.83	111806.43	0:04:32.65
GaussianMixture	2.09	0.96	6.01	0.86	115385.38	0:03:56.32
Som	0.77	1.88	0.82	-0.61	0.08	1:02:03.32
K-means++	2.03	0.40	13.72	0.87	310039.84	0:03:55.37

企业信用风险评估

Method	CP	DB	SP	SS	CH	TIME
BayesianGaussianMixture	1.65	0.66	40.68	0.99	220209.35	0:01:43.61
GaussianMixture	2.88	0.96	37.20	0.99	147948.82	0:01:39.94
Som	0.41	1.89	0.43	-0.82	0.05	0:58:48.22
K-means++	0.51	0.10	42.72	1.00	1851197.54	0:00:45.68

系统展示

批量查询



新企业预测



模糊搜索



企业画像及簇标签展示



原始数据分析



新企业信息创建



项目目标

定性目标

每一个企业簇群体形成明显的划分界限，响应速度到毫秒级
线性降低金融机构的人工、时间成本，整合多渠道数据
解决金融机构“寻客难、审核难、资金流通效率低”的痛点
缓解中小微企业“融资难、融资慢、融资贵”的烦恼

定量目标

新企业标签预测所需时间控制在1秒以内，预测精准性达96.8%以上，
六个维度的模型训练所需总时间节省到14分钟以内
模型轮廓系数接近1.00，实现4~6类合理有效的簇划分
数据的安全性保证达98.2%以上，系统使用过程的稳定性到99%以上

目标检测

指标评估：紧密性方法、间隔性方法、戴维森堡丁指数、轮廓系数、Calinski-Harabasz指数
测试人：小沈、小余、小宋
测试结果：均达到优秀水平

白盒测试：小方
黑盒测试：林老师
测试结果：均达到良好水平

获取使用体验
测试人：学校合作银行的25位员工、2名专家评估
测试结果：好评率为82.5%

实施计划

“慧融宝”甘特图

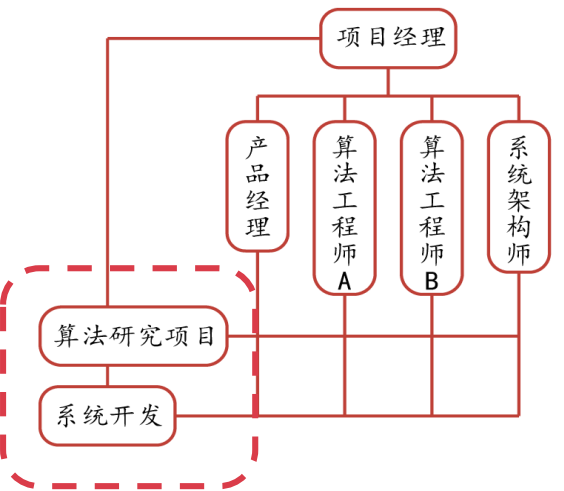
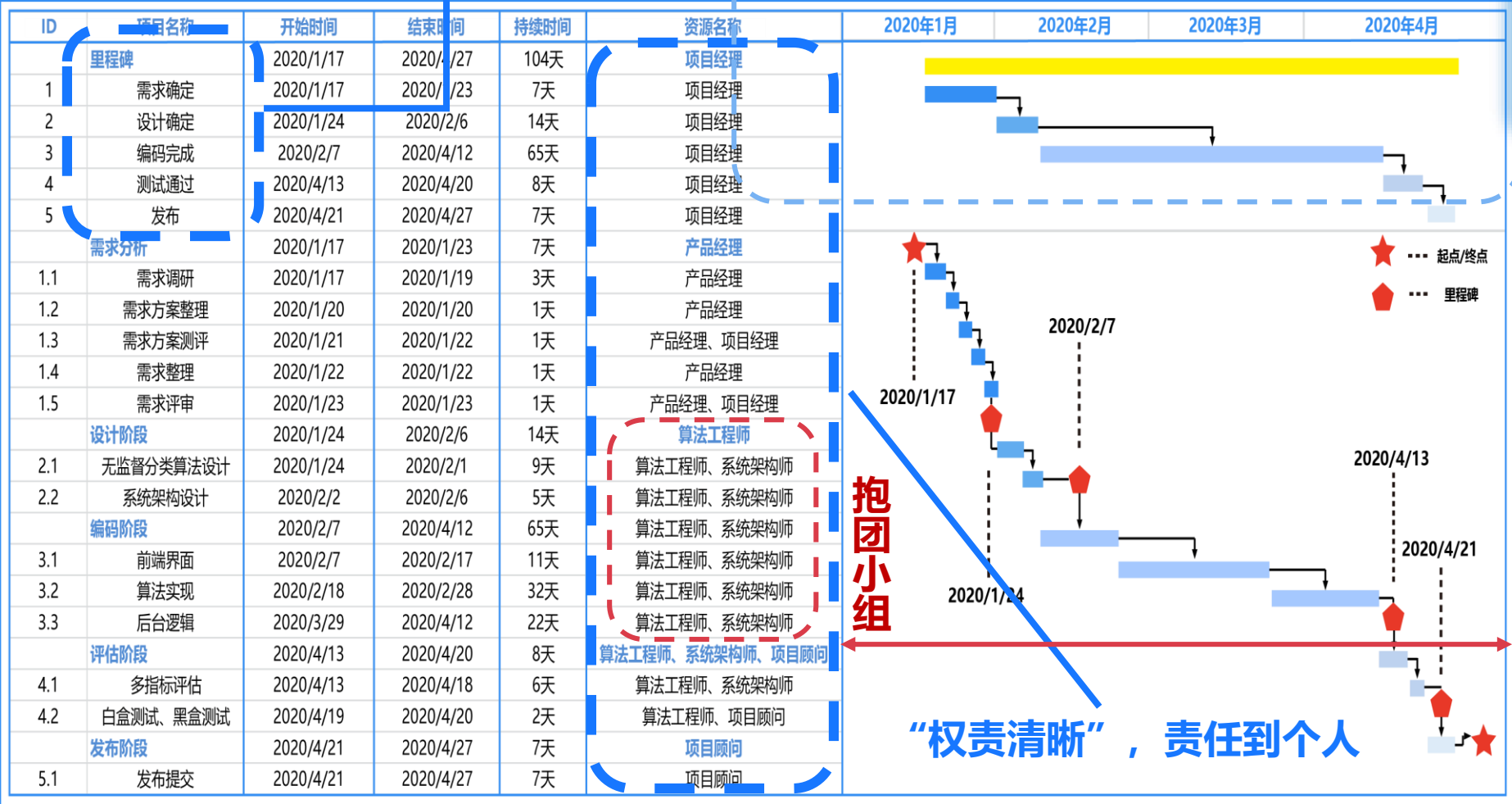
设立父任务、子任务

清晰的时间线

全周期项目制管理

闭环管控评估模式

搭配“矩阵式组织架构”



核心成员

4名来自国家级重点实验室


包揽华为、阿里巴巴、网易、海康威视等

各大
项目经理

国家级、
系统架构

国家级团体程序设计天
浙江省
一项软
阿里云python基础编程
python数据科学分析

100079
北京市丰台区乡路88号院石幢中心5号楼
102
北京中安华彬知识产权代理有限公司


2020R11L396336

软件登记受理通知书

流水号: 2020R11L396336
软件全称: 一种基于Kmeans算法的金融无监督分类平台
版本号: V2.0
登记类型: 计算机软件著作权登记申请
申请人:
代理人: 北京中安华彬知识产权代理有限公司

根据《计算机软件著作权登记办法》第十九条的规定, 申请人提出的上述软件登记申请中国版权保护中心予以受理。

受理号: 2020R11S0362244
受理日期: 2020年 04月 23日

经核实确认中国版权保护中心收到如下申请文件:

申请表 1份, 3页/份
源程序 1份, 66页/份
文档 1份, 13页/份
身份证复印件 1份, 1页/份
代理人身份证明文件 1份, 1页/份

中国版权保护中心软件登记部

2020年 04月 23日

注意事项: 申请人收到本通知后, 如需询问有关事宜请咨询受理号和流水号。

华为网络
一篇
网易m
国家级

立项
竞赛
情感

成本模型

构造性成本

采用COCOMO模型估算研发成本

L——源指令条数 1KDSI=1000DSI

E——开发工作量（以人月计） 1MM=1/12人年=19人日=152人时

D——开发进度。（以月计）

开发工作量：MM=a*(KDSI)^b

开发进度：TDKV=c*(MM)^d

经验常数a=2.4, b=1.05, c=2.5, d=0.38

预计源指令条数为6KDSI

开发工作量：MM=a*(KDSI)^b=2.4*(5K)^{1.05}=13人月

开发进度：TDKV=c*(MM)^d=2.5(13)^{0.38}=6月

高度相符

盈利可行性分析

投资净现值 $NPV = \sum_{k=0}^n \frac{NCF_k}{(1+i)^k} - \sum C = 51.135 \text{万元}$

远大于零，项目方案盈利能力很好

内涵报酬率 $\sum_{t=1}^n \frac{NCF_t}{(1+r)^t} - C = 48.67\%$

大于15%的资金成本率，收益能力好

动态回收期 $\sum_{k=0}^n I_k = \sum_{k=0}^n O_k$ ，回收期为2.56年

回收期为2.56年，方案可行