第十一届中国大学生服务外包创新创业大赛

e企查

基于加权多层K-Means的企业分类系统

大熊维尼队

助力企业分类，实现客群挖掘

目录

CONTENT

你的企业画像构建专家

01

项目背景

助力企业分类 实现客群挖掘

金融科技场景中，企业数据量大且来源广泛，信息维度丰富，在分析企业还款能力、信用水平过程中，面临巨大的挑战。

企业数据量和维度的丰富，使得市场需要一种高效的企业画像评估工具，并对企业实现有效归类。

金融信贷机构

- 难以找到心仪投资企业
- 企业信息众多，难以筛选
- 很难快速对企业画像进行评估
- 企业评估结果难以推广

企业数据管理者

- 数据量巨大，维度大，处理难
- 人工处理效率低，出错率高
- 难以建立一种可量化的评估指标
- 难以建立适应数据变化的模型

你的企业画像构建专家

01

项目背景-现状

金融信贷机构

Q：如何尽快的提取企业的关键信息？

Q：如何减轻企业查询工作量？

Q：如何更加便捷地获取企业详细信息？

Q：如何准确推广得到更多潜在目标企业？

算法难题

Q：如何建立高效准确的无监督训练模型？

Q：如何解决数据度量不统一和鲁棒性问题？

Q：如何解决预处理过程中数据缺失严重，且聚类效率低下的问题？

解决方案及亮点

A：系统筛选核心评估建模指标+丰富的企业标签

A：批量查询+多角度条件筛选

A：企业数据可视化+生成pdf报告

A：智能推荐算法，挖掘潜在客群

A：通过实验对比，确定K-Means作为无监督模型，并在其基础上改进，提出**加权多层K-Means算法**

A：使用**等级化算法**，对**一维属性使用K-Means并根据簇中心值排序**，转化一定范围的等级，方便建模比较。

A：将属性归类到7个指定模块下，每个模块中的属性构建加权新指标——模块总分，在尽可能保证企业信息完整的前提下，实现**数据合理降维**，**提升聚类效率**，也**解决不同企业数据缺失不一致的情况，并确保聚类准确性。**

02 方案设计-算法架构

加权多层K-Means体系

①首先将企业数据，根据属性特点，将分入到风险、投资、创新、品牌、招聘、信用、基本7个二级指标当中，这里称作模块，并进行预处理。

②等级化算法使用K-Means算法对数据赋予等级。

③企业原属性等级化，生成原属性等级，即加权多层第一层。

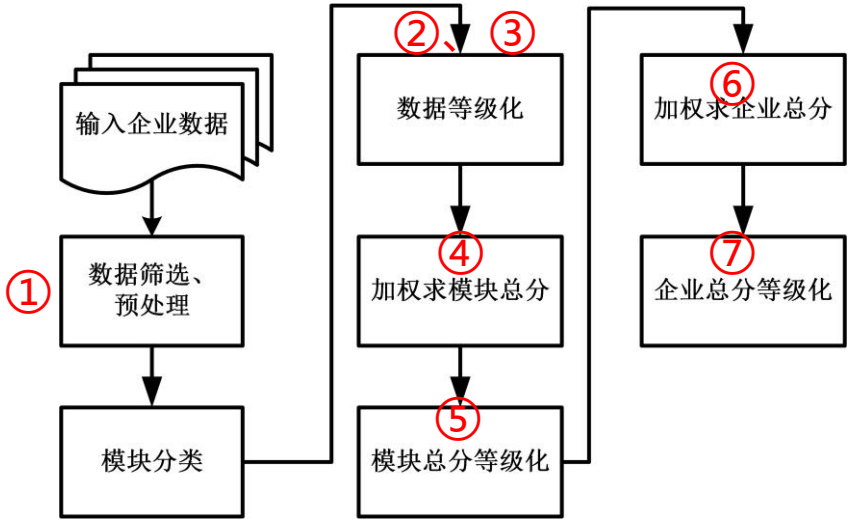
④根据每个模块内的属性等级加权求和，获得模块评分。即加权多层中的第一次加权。

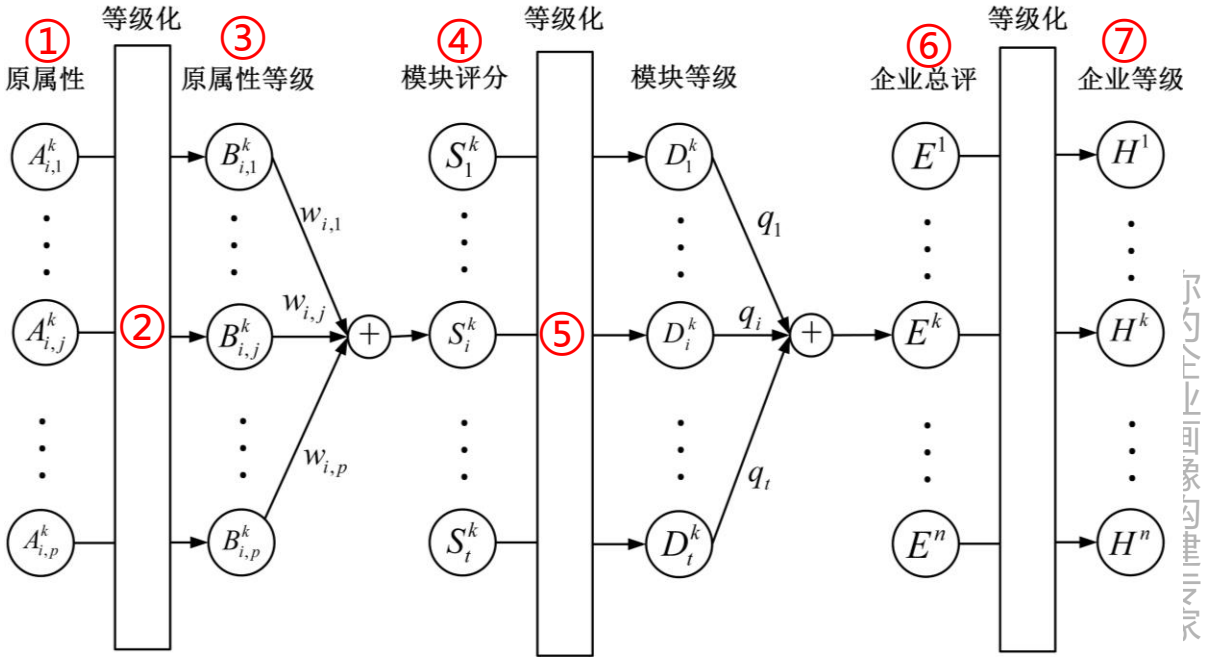
⑤不同的模块评分，再次等级化，获得模块等级。对应加权多层的第二层。

⑥所有模块的等级再次加权求和，获得企业总评分。对应加权多层的第二次加权。

⑦企业总评分再次等级化，得到企业等级。对应加权多层的第三层。

⑧7个模块的等级，即7×n的矩阵，使用K-Means加权聚类，获得一个聚类标签。





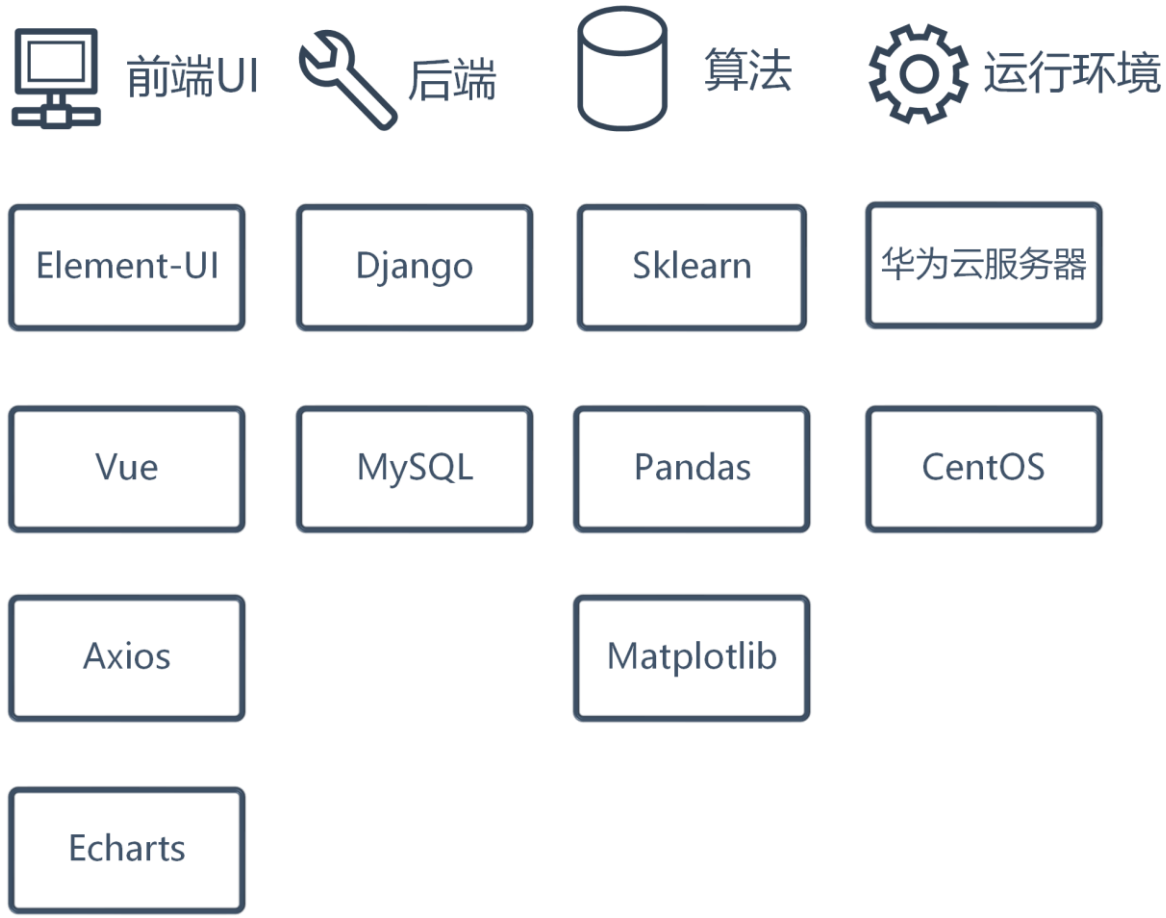
助力企业分类 实现客群挖掘

02

方案设计-技术路线

e企查整体采用Python的技术路线，**搭建基于加权多层K-Means的企业画像构建系统**

- 后端使用Django+MySQL的架构
- 算法使用Pandas+Matplotlib进行数据处理分析，使用Sklearn作为机器学习框架
- 前端采用Vue，Axios等最新技术
- 系统部署在华为云上



助力企业分类 实现客群挖掘

你的企业画像构建专家

02

方案设计-系统功能

系统管理

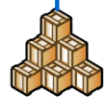
- 模型属性管理
- 机器学习模型权重调整
- 机器学习模型测试

数据管理

- csv批量导入数据
- 数据管理驾驶舱

普通用户

- 企业批量、筛选搜索
- 智能推荐相似企业
- 获取企业标签
- 数据可视化
- 企业pdf报告生成



未来拓展

- 形成企业社区生态
- APP端构建
- 系统优化

助力企业分类 实现客群挖掘

你的企业画像构建专家

03 成果展示-前台查询企业

①专供查询用户使用

②精确、提示查询

③热搜企业

④批量上传查询企业

⑤条件筛选查询

助力企业分类，实现客群挖掘

你的企业画像构建专家

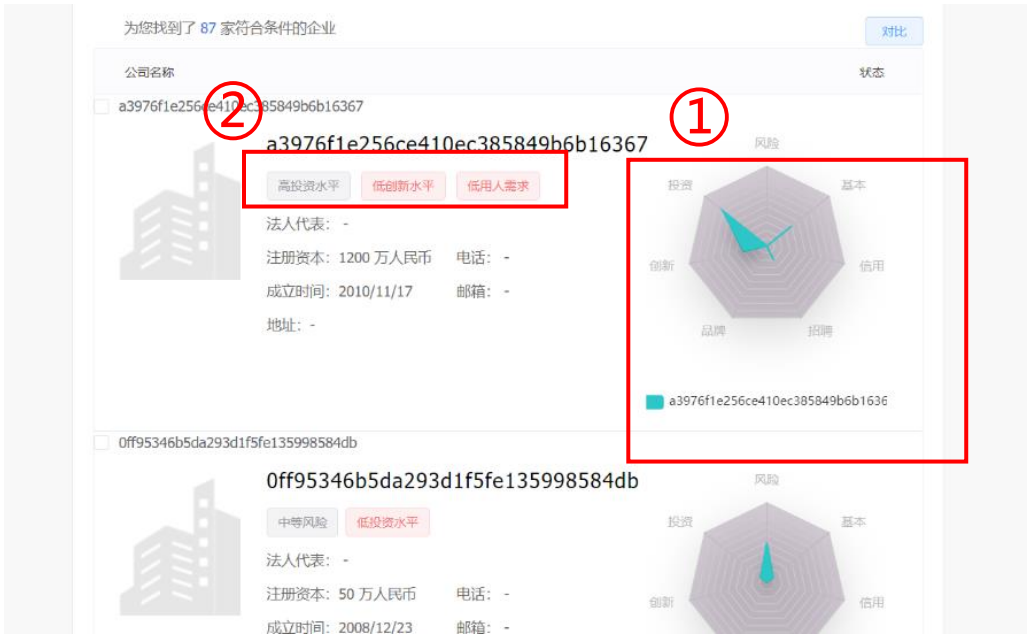
03 成果展示-前台辅助分析

①**雷达图**全方面展示企业能力

②提供**企业标签**，概括企业特征

③**对比图**分析企业能力差距

④**PDF评估报告**全面给出企业信息



助力企业分类，实现客群挖掘

你的企业画像构建专家

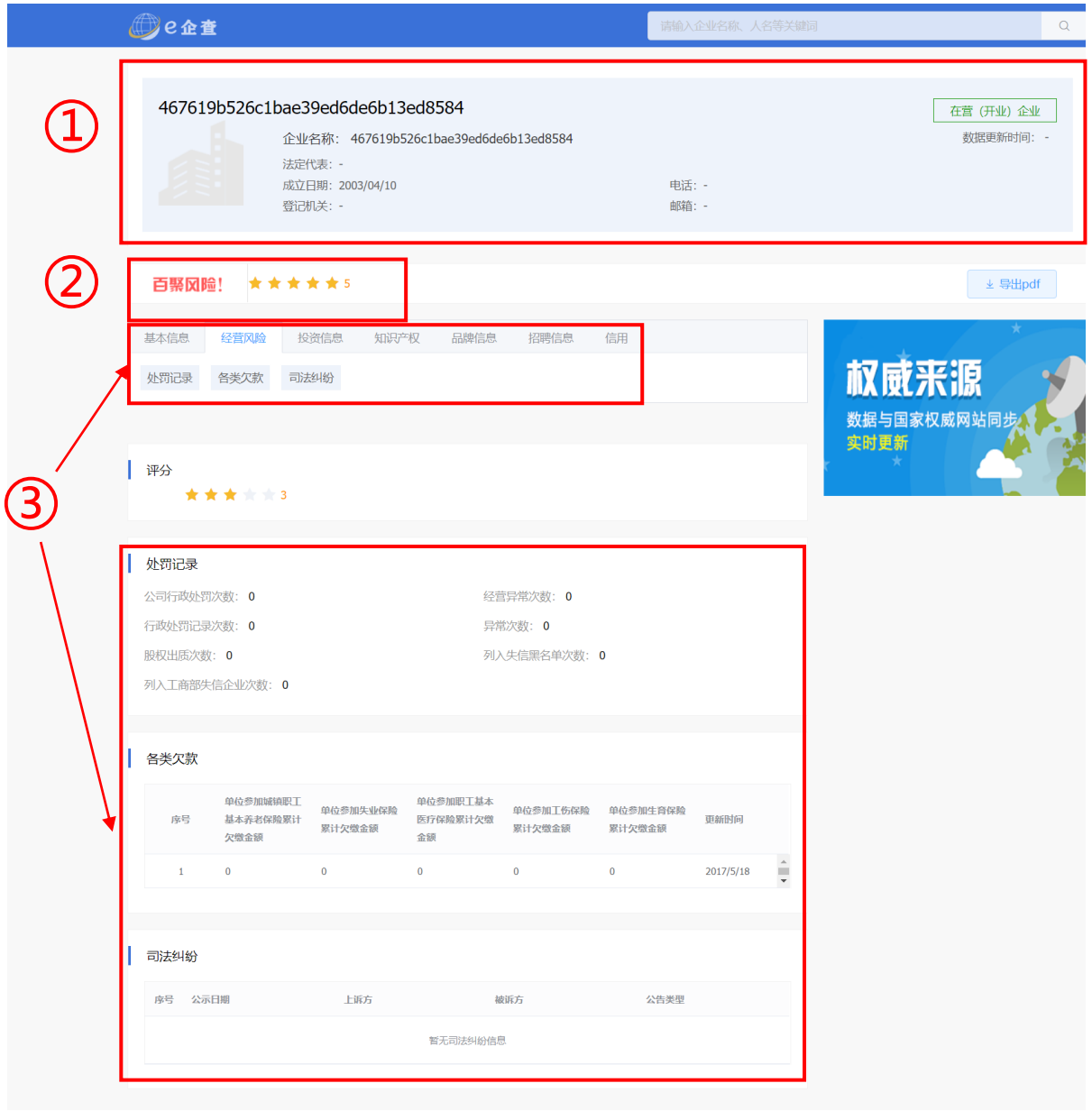
03 成果展示-企业详细信息

助力企业分类，实现客群挖掘

①流畅的用户体验，查询企业详细信息

②星级标注，显示指定模块信息的等级和企业总评等级

③切换模块，显示不同类型详细信息



你的企业画像构建专家

03 成果展示-企业智能推荐

助力企业分类、实现客群挖掘

精准推荐目标企业的相似企业

①**雷达图**体现企业各维度水平相似

②**标签**体现企业相似

目标企业

相似企业

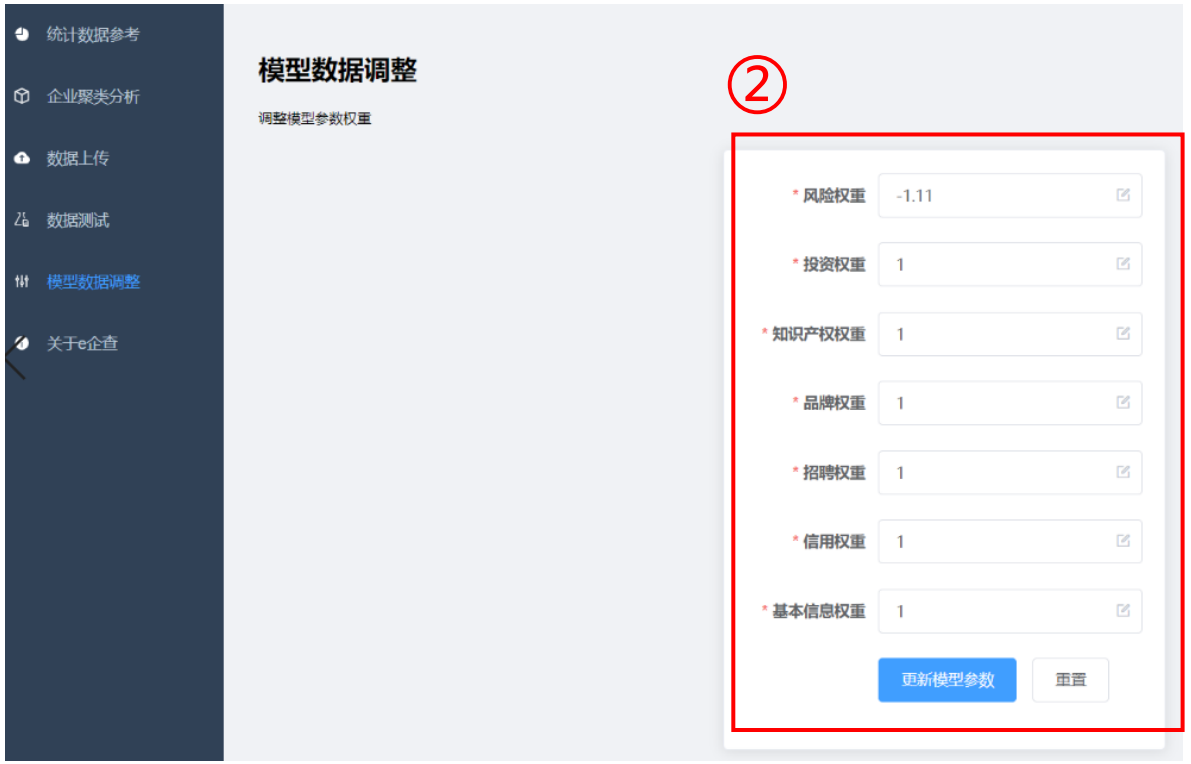
你的企业画像构建专家

03 成果展示-数据、无监督模型管理

①

①批量上传企业数据，易于企业数据更新

②无监督模型权重调整



助力企业分类 实现客群挖掘

你的企业画像构建专家

03 成果展示-数据驾驶舱

①企业各簇以及各等级总体分布

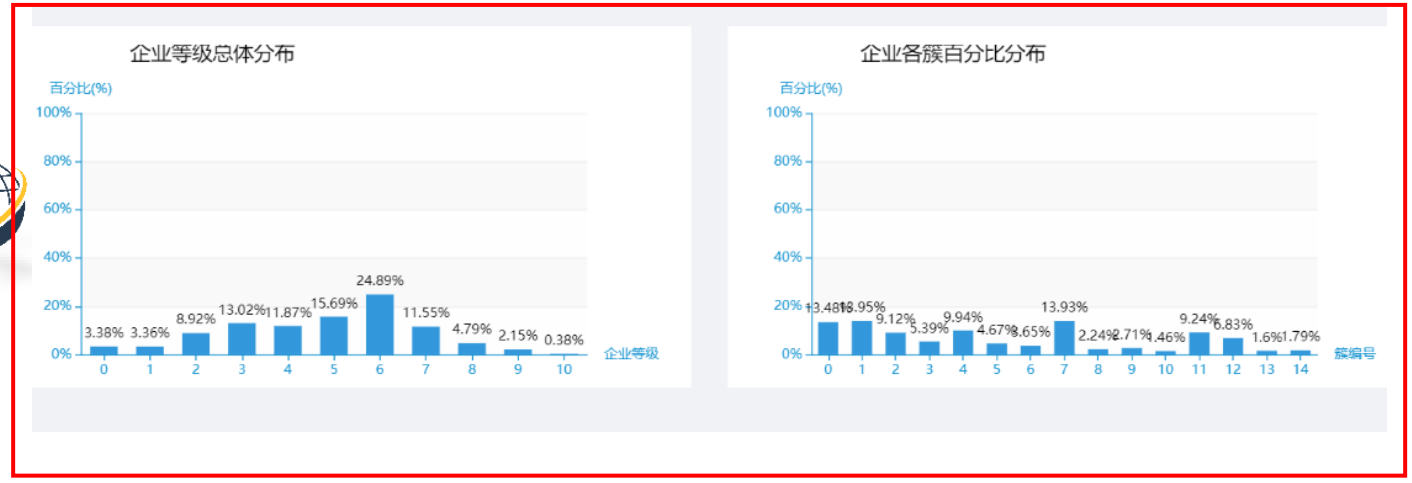
②企业各簇内，各模块内等级分布情况统计，帮助分析每个簇内部模块等级分布特征

③PCA降维可视化，一眼了解聚类结果

助力企业分类，实现客群挖掘

③PCA可视化（二维）

①左图展示企业0-10个等级的百分比分布，右图展示企业15个簇的百分比分布



②例如1号簇，企业综合等级分布

你的企业画像构建专家

03 成果展示-模型训练、预测

管理员测试模型的专用界面

①训练模块

用户可以通过上传自己的数据集，训练生成新模型，

②预测模块

用户可以选择系统本身的模型，或者自己新训练的模型进行预测。

训练和预测都会返回带标签的数据，以及簇描述文件，并返回数据处理用时，和训练或预测的总时间。

③详细的使用说明

03 成果展示-训练、预测结果分析

助力企业分类，实现客群挖掘

生成的训练或测试结果，是csv文件，包含每个模块属性的标签，按照0-10划分其等级。以ent_module.csv为例，前7列对应7个模块的等级，第八列表示企业加权总分，第九列表示企业综合等级，第十列，是加权聚类后划分的簇标签。此外还会生成pic，每张里面的图，描述了指定簇属性分布情况，概括簇特征。

训练或预测结果文件夹

pic	2020/5/22 13:58	文件夹	
base_module.csv	2020/5/22 13:58	Microsoft Excel ...	86 KB
brand_module.csv	2020/5/22 13:58	Microsoft Excel ...	29 KB
creativity_module.csv	2020/5/22 13:58	Microsoft Excel ...	168 KB
credit_module.csv	2020/5/22 13:58	Microsoft Excel ...	104 KB
ent_module.csv	2020/5/22 13:58	Microsoft Excel ...	13,248 KB
investment_module.csv	2020/5/22 13:58	Microsoft Excel ...	515 KB
recruit_module.csv	2020/5/22 13:58	Microsoft Excel ...	200 KB
risk_module.csv	2020/5/22 13:58	Microsoft Excel ...	623 KB

企业7个模块的评分等级

企业综合等级

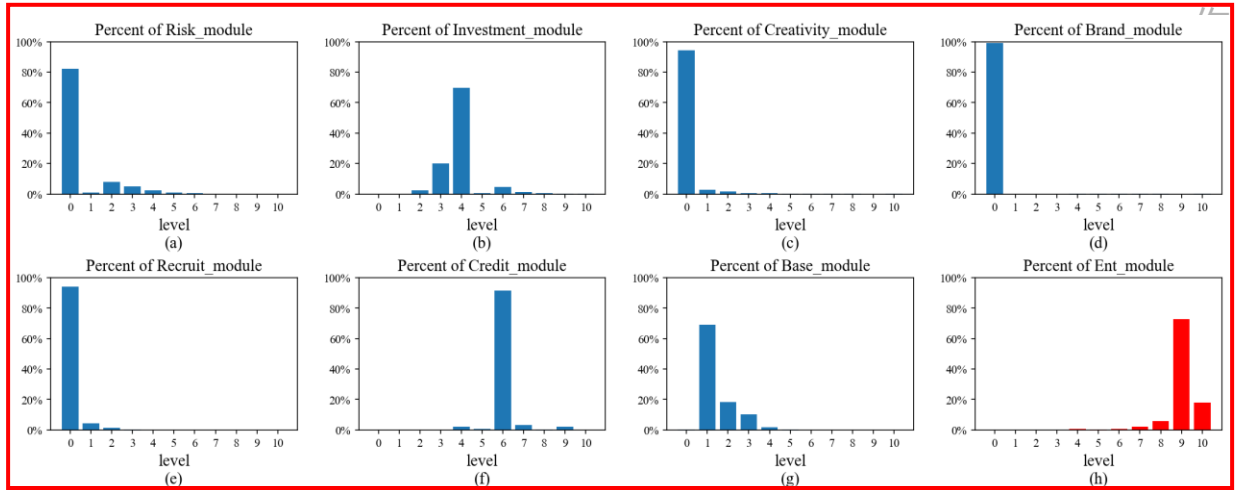
	entname	risk_modul	investment	creativity_r	brand_mo	recruit_mo	credit_mo	base_mod	ent_modul	ent_modul	ent_inner_ty
2	467619b52	3	8	10	8	2	9	3	36.67	10	7
3	c326dd469	1	10	6	10	2	7	2	35.89	10	7
4	7c31a233a	0	10	7	0	5	7	3	32	10	13
5	ce3265a93	0	9	4	0	1	9	4	27	10	7

簇特征描述图片，描述7个模块和总分的百分比分布

7号簇的各模块及总分分布

企业加权总分

企业簇标签



04 创新优势——算法特色与亮点

①**核心指标全面领先**(CH、SH指数越大越优，DB指数越小越优，测试数据为19w条企业数据)

模型	加权多层 K-Means	K-Means	Birch（层次聚类）	GMM (高斯混合聚类)
CH	136499.42	80648.91	28242.63	26089.34
DB	0.87	0.928	1.16	2.72
SH	0.56	0.55	0.42	0.26
训练时间(s)	11.43	14.75	19.67	30.97
预测时间(s)	0.23	0.25	0.26	0.29

助力企业分类 实现客群挖掘

你的企业画像构建专家

04 创新优势——算法特色

②等级化算法：

使用K-Means对一维属性聚类，并根据簇中心值排序。统一度量标准，将类似金额（范围在0-2亿），次数指标（例如行政处罚在0-20次），量化指标转化成属性内等级0-K，既方便了后续的处理，也便于直接了解该属性，在所有企业中的总体水平；

③模块分类：

将类似的企业属性，归入到同一模块下，并通过加权的方式构建一个新的指标——模块评分。实现数据的合理降维，缓解数据缺失问题。

④加权K-Means：

可以根据实际使用需要，更改不同属性的权重，调整不同属性对最终聚类结果的影响。

⑤扩展性良好：

用户可以通过自定义筛选条件，与聚类结果相结合，进一步做分类、智能推荐。

04 创新优势——系统功能技术特色

助力企业分类，实现客群挖掘

系统特色

前后台分离，术业有专攻。

①**前台部分**：符合用户使用习惯，结合本系统特色，保障用户体验。

②**后台部分**：专做数据维护和模型维护，提升使用体验同时也确保系统的安全性。

功能特色

①企业**智能推荐**，利用算法结果，综合各类信息，自动生成企业标签。

②支持**批量操作**，减少操作负担。

③**管理驾驶舱**，以图表，在线pdf方式直观反映模型、企业数据状况。

④**丰富筛选模式**，助力找到心仪企业。

技术特色

①整体采用**Restful**风格进行数据传输，减少系统数据传输负担。

②前端采用**Vue**技术，并使用Axios动态生成网页界面。

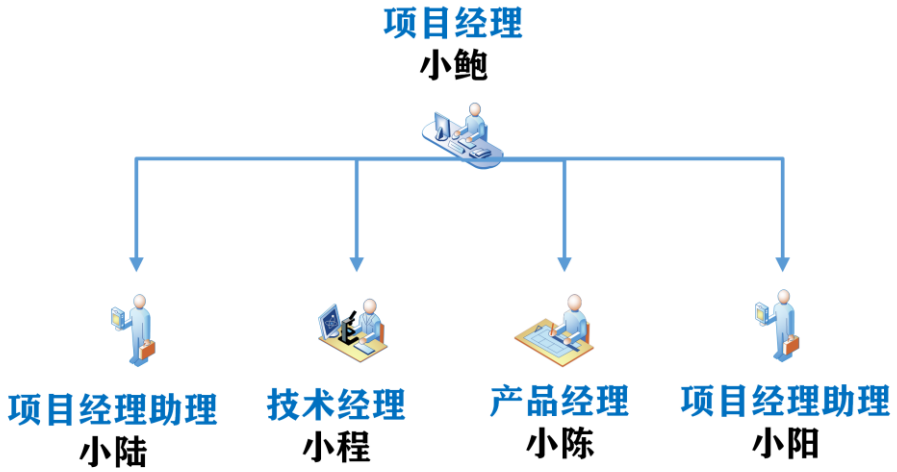
③使用**Echarts**进行数据可视化，具有良好的交互性。

④后端使用**Django**，具有优秀的**ORM**框架，与数据库进行交互，同时也能与机器学习模型进行无缝衔接。

你的企业画像构建专家

05 商业考量-组织结构

精简的团队架构。团队成员有长期合作经验，技术扎实，内部分工明确，能够形成良好的沟通与反馈。



角色	团队职责
小鲍 项目经理	1、制定项目规划，确定项目整体方案和框架 2、前端开发、UI设计、服务器管理 3、前端技术方案标准制定 4、制定用户使用手册
小程 技术经理	1、后端技术方案可行性、先进性以及标准的制定 2、后端开发、数据库维护、服务器管理 3、网页测试
小陈 产品经理	1、提炼项目需求与亮点，项目方案撰写 2、算法设计及算法、业务部分文档撰写负责 3、项目推广
小陆 项目经理助理	1、负责维护客户关系，持续于用户沟通需求 2、协助商业方案制定 3、协助项目经理进行项目规划 4、负责项目对外公关、宣传工作 5、一定的技术支持
小阳 项目经理助理	

05

商业考量-业务模式

助力企业分类，实现客群挖掘

Step 1

捕捉机会

微小企业数据的多元化，人工处理劳动成本、出错率高

Step 2

制定对策

设计加权多层K-Means算法，准确分类企业和画像构建

Step 3

建设能力

开发“e企查”平台，管理企业数据和模型，提供良好使用体验

Step 4

实现卓越

进一步提升算法性能，扩充业务范围，增强市场影响力

你的企业画像构建专家

05 商业考量-可行性分析

助力企业分类 实现客群挖掘

技术可行性：

通过实验对比确定K-Means模型，并在其基础上针对企业画像和分类，进行改进提出加权多层K-Means，各项指标领先传统聚类模型。前后端技术均使用当下最先进的技术，利于运行和维护。且团队成员拥有合作经验，能团结合作攻克各项难题。

商业可行性：

“e企查”在算法基础上，开发了丰富的可视化和辅助模块功能，在此基础上可以拓展会员服务，广告收入进行盈利。具有①社会背景的支撑性，②盈利点选择余地大③解放金融机构企业归类劳动力等优势

使用可行性：

1、项目自身而言的使用效益①技术优势性；②宣传创新性

2、使用者而言的使用效益

团队致力于开发具有良好交互的网页平台，且丰富的可视化和辅助评估，都可以带给用户舒适的体验。

法律可行性：

1、企业所有数据源自于国家公开的商业查询数据，有助于避免人为篡改数据行为。

2、本项目各类成果拥有自主知识产权。

你的企业画像构建专家

感谢观看

大熊维尼队：鲍锋雄　程凯
陈振乾　陆纪慧　阳璐