# 

**项目详细方案**

目录

[摘 要 1](#_Toc39689931)

[1.项目介绍 2](#_Toc39689932)

[1.1项目概述 2](#_Toc39689933)

[1.2项目背景 2](#_Toc39689934)

[1.3项目目标 4](#_Toc39689935)

[1.4项目意义 6](#_Toc39689939)

[1.5项目亮点 7](#_Toc39689940)

[2.需求分析 9](#_Toc39689946)

[2.1需求调研 9](#_Toc39689947)

[2.2功能需求 10](#_Toc39689948)

[2.3业务分析 11](#_Toc39689949)

[2.4市场分析 12](#_Toc39689952)

[3.算法模型 16](#_Toc39689955)

[3.1建模准备 16](#_Toc39689956)

[3.2模型特征选择 18](#_Toc39689960)

[3.3数据处理 20](#_Toc39689963)

[3.4模型概述 26](#_Toc39689969)

[3.5参数调优 32](#_Toc39689972)

[3.5模型结果 34](#_Toc39689973)

[4.系统设计与实现 36](#_Toc39689974)

[4.1系统总体设计 36](#_Toc39689975)

4[.2数据库设计 37](#_Toc39689976)

[4.3数据导入模块 37](#_Toc39689977)

[4.4 分类器模型模块设计 38](#_Toc39689978)

[4.5 分类器更新模块设计 40](#_Toc39689981)

[4.6 数据可视化模块设计 41](#_Toc39689982)

[5.可行性分析及风险管理 44](#_Toc39689983)

[5.1 SWOT分析 44](#_Toc39689984)

[5.2可行性分析 45](#_Toc39689985)

[5.3风险管理 47](#_Toc39689990)

[6.团队构成及项目管理 49](#_Toc39689994)

[6.1团队构成 49](#_Toc39689995)

[6.2团队分工 49](#_Toc39689996)

[6.3项目计划安排 50](#_Toc39689997)

[6.4项目管理 52](#_Toc39690002)

[7.参考文献 53](#_Toc39690006)

[8.附录 54](#_Toc39690007)

图表索引

[图1- 1 各地区“僵尸企业”清理压力及地方财政收入点图 3](#_Toc40453024)

[图1- 2分年份僵尸企业统计图 4](#_Toc40453025)

[图2- 1僵尸企业比例增长对地方和国家资产影响 9](#_Toc40453031)

[图2- 2系统功能需求图 11](#_Toc40453032)

[图2- 3系统业务流程图 12](#_Toc40453033)

[图2- 4部分中央出台的政策截图 13](#_Toc40453034)

[图3- 1数据处理流程图 20](#_Toc40651136)

[图3- 2多表融合与多表连接流程图 23](#_Toc40651137)

[图3- 3数据合并结果图 23](#_Toc40651138)

[图3- 4特征提取与新特征建立流程图 26](#_Toc40651139)

[图3- 5模型总体设计图 26](#_Toc40651140)

[图3- 6基于Stacking策略的XGBoost算法 29](#_Toc40651141)

[图3- 7损失函数变化图-改变最小子树权重 33](#_Toc40651142)

[图3- 8 损失函数变化图-改变数的最大深度 33](#_Toc40651143)

[图3- 9损失函数变化图-改变学习率 33](#_Toc40651144)

[图3- 10模型分数评估-最小子树权重 34](#_Toc40651145)

[图3- 11原始验证集上测试的模型分数结果截图 35](#_Toc40651146)

[图3- 12新数据集上测试的模型分数结果截图 35](#_Toc40651147)

[图4- 1系统总体架构图 36](#_Toc40651255)

[图4- 2数据库表结构图 37](#_Toc40651256)

[图4- 3首页界面截图 37](#_Toc40651257)

[图4- 4文件上传界面截图 38](#_Toc40651258)

[图4- 5预测结果界面截图 38](#_Toc40651259)

[图4- 6僵尸企业分类界面截图 39](#_Toc40651260)

[图4- 7非僵尸企业僵尸化等级预测结果界面截图 40](#_Toc40651261)

[图4- 8分类器更新模块逻辑图 41](#_Toc40651262)

[图4- 9僵尸企业列表界面截图 42](#_Toc40651263)

[图4- 10僵尸企业详情界面截图 42](#_Toc40651264)

[图4- 11僵尸企业大数据分析界面截图 43](#_Toc40651265)

[图6- 1团队成员结构图 49](#_Toc40651266)

[图6- 2任务安排甘特图 51](#_Toc40651267)

[表3- 1各省份标准统计表 18](#_Toc40651156)

[表3- 2 29个原始特征表 20](#_Toc40651157)

[表3- 3处理后保留特征列表 21](#_Toc40651158)

[表3- 4僵尸企业普遍存在的三个问题对应的原始特征 24](#_Toc40651159)

[表3- 5建立的两个新特征 24](#_Toc40651160)

[表3- 6各种融资方式的利率 25](#_Toc40651161)

[表3- 7提取出的融资额度特征 25](#_Toc40651162)

[表3- 8最终选择的特征 25](#_Toc40651163)

[表3- 9式(3.7)(3.8)参数说明表 27](#_Toc40651164)

[表3- 10修改后的FN-CHK模型公式参数说明表 28](#_Toc40651165)

[表3- 11式(3.13)(3.14)(3.15)中各参数说明表 30](#_Toc40651166)

[表3- 12僵尸企业分类标准表 30](#_Toc40651167)

[表3- 13式(3.16)(3.17)参数说明表 31](#_Toc40651168)

[表3- 14选取得预警体系指标 31](#_Toc40651169)

[表3- 15风险等级评估表 32](#_Toc40651170)

[表3- 16XGBoost参数说明 32](#_Toc40651171)

[表3- 17最终选择的参数值 34](#_Toc40651172)

[表3- 18原始验证集上测试的模型分数结果展示表 34](#_Toc40651173)

[表3- 19新数据集上测试的模型分数结果展示表 35](#_Toc40651174)

[表6- 1团队成员详细分工表 49](#_Toc40651273)

## 摘 要

近几年，“僵尸企业”频繁出现在中国官方文件以及各种会议中。据资料显示，从2001年开始我国上市公司中僵尸企业数量持续上升，且从2003年之后上市公司中僵尸企业的比例一直在13%左右波动。僵尸企业的存在破坏了市场机制，加剧了信贷资源错配，带来了严重的产能过剩问题，还对其他非僵尸企业产生了投资挤出效应。因此，如何依托大数据背景，结合人工智能算法准确、高效地识别出僵尸企业成了本项目着重解决的问题。

本系统通过数据治理，完善企业画像的各维度标签，达到对企业进行分类识别的目的。能根据用户提供的企业数据，准确、高效地识别出僵尸企业，同时根据僵尸企业的吸血表现把僵尸企业分为利息补贴、政府补贴、常青贷款三类；另一方面，对于非僵尸企业，系统根据用户提供的数据对其进行僵尸化风险等级预测。本项目的创新点主要体现在：

1. **识别标准**

目前已有的CHK标准、FN-CHK标准等均没有考虑我国的国情，同时也没有考虑我国省份的政策差异。因此，团队提出一个新的识别标准——ALG标准。

1. **模型融合**

本项目的模型在参考XGBoost算法的同时结合了统计模型，通过stacking模型融合提高了模型识别的准确性。

1. **数据处理**

在特征选择上结合特征相关性分析，增强特征相关性，从而提高了模型识别的准确性；在缺失值处理上，保留更多的数据对模型进行训练，提高了模型的查全率和查准率。

1. **预警功能及预测结果处理**

本项目除了实现了僵尸企业的识别和分类的基础功能外，同时增加了对非僵尸企业僵尸化风险等级预测的功能，在完善系统功能的同时提高了更友好的用户体验。除此之外，项目提供企业下载僵尸企业识别结果的功能，满足了用户对预测结果进行单独保存的需求，同时方便用户对预测结果进行后续的自定义处理。

本团队实现了一个优质的识别预测模型，模型具有准确性、高效的特点。本项目响应了中央出台的一系列政策，推动了互联网及大数据时代的发展，在政治层面的和时代层面都具有举足轻重的作用，同时对于我国市场机制以及经济发展也具有重大意义。

综上所述，本系统在解决政府以及社会高度关注的问题——僵尸企业对我国市场经济发展的危害方面具有举足轻重的作用。

# 1.项目介绍

## 1.1项目概述

僵尸企业画像及分类系统是一种结合人工智能算法，通过数据治理，并采用统计、分析、挖掘等技术，完善企业画像的各维度标签，达到对企业进行合理分类识别的目的。该项目涉及大数据、人工智能等许多前沿技术，具有创新性和适用性。

依托人工智能技术建立企业画像模型进行僵尸企业分类识别，瞄准了互联网步入大数据时代的时代背景，国家工商局发布依靠大数据时代前沿科技勾勒企业经营变化情况的政策支持。团队在充分了解市场需求、结合时代背景的情况下，完善了僵尸企业的识别标准，构建了识别僵尸企业的算法模型，搭建了可视化网页展示平台，最终形成了僵尸企业画像及分类系统。

本项目的开展将推动在互联网的大背景下，大数据在实际应用中的进一步发展和推广，为保持市场机制平衡、合理分配信贷资源提供了数据支撑，为识别僵尸企业提供了解决方案，对解决目前市场面临的产能过剩问题，保障我国经济动能转换以及高质量发展具有重要作用。同时，有利于大数据应用在我国各个领域的普及，减少人力、物力的大量投入，充分发挥人工智能的独特优势，极大提高市场上企业的质与量，推动我国市场的健康发展。

同时，本项目还充分发挥了数据可视化技术的优势，结合大数据对僵尸企业进行了存库和可视化的统计分析。在可视化页面清晰展示了全国各个地区的僵尸企业的分布情况、僵尸企业的行业分布以及僵尸企业的成因特征统计等。

## 1.2项目背景

在互联网逐渐步入大数据时代后，企业的工商、司法、经营、上市、知识产权、舆情等多维度数据的获取更加便捷。2016年9月，国家工商总局发布《工商总局关于新形势下推进监管方式改革创新的意见》，并指出要“依托大数据加强监管”，“充分发挥大数据在制定完善新型市场监管制度和政策中的作用，搜集掌握经营者、消费者和社会公众的反应，跟踪监测有关制度和政策的实施效果”，“在工商登记、企业监管、网络交易、竞争执法、消费维权等领域率先开展大数据示范应用”。通过对企业从业人数、成立年限、注册资本、营业收入、风险信息、行政处罚、纳税信用等级、黑名单、上市信息、电商信息等数据关联处理，按需进行权重分割，并对这些数据进行综合分析，构建出企业全息画像，整体评估一个企业综合价值，更好地勾勒出企业的经营变化情况。

僵尸企业是指缺乏盈利能力却能够以低于市场最优利率成本获得信贷资源，依靠外界输血而缺乏自生能力的企业，僵尸企业的存在破坏了市场机制，加剧了信贷资源错配，带来了严重的产能过剩问题，还对其他非僵尸企业产生了投资挤出效应。此外，由于我国金融制度不健全，僵尸企业往往会倾向于采用企业间商业信用的非正式金融方式融资，带来对僵尸企业合作关联企业的“传染”，甚至会拖累银行成为“僵尸银行”。僵尸企业如果不能够及时处理，经济动能转换以及高质量发展就无法保障，去杠杆工作的受阻将加大系统性金融风险，鉴于僵尸企业的危害性，自2015年以来，中央出台了僵尸企业处置的一系列政策指导文件，试图以处置僵尸企业作为深化供给侧结构性改革，调整经济结构以及促进经济行稳致远的重要抓手。

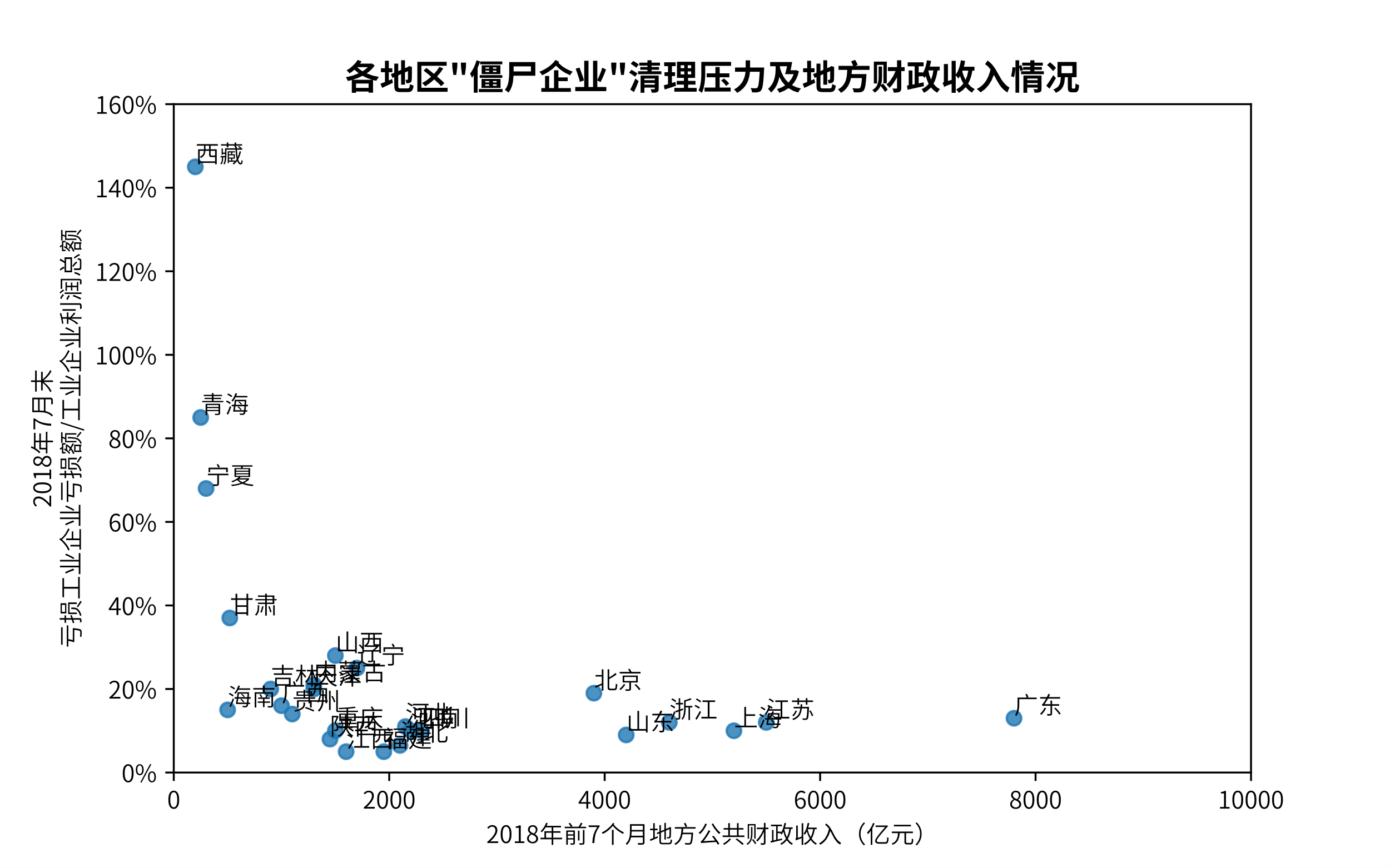


图1- 1 各地区“僵尸企业”清理压力及地方财政收入点图

20世纪90年代日本经济在资产价格泡沫破灭后，陷入了衰退状态，出现了“失去的十年”。学术界普遍认为，主要原因就是日本有大量的僵尸企业，占用了大量无效率的银行贷款，甚至挤出了优秀的企业，出现了“逆向淘汰”。

近几年，“僵尸企业”这个名词开始频繁出现在中国官方文件及会议中。据资料显示，从2001年开始我国上市公司中僵尸企业数量持续上升，且从2003年之后上市公司中僵尸企业的比例一直在13%左右波动，下图显示了从2000-2015年上市公司中僵尸企业的数量及所占比例。僵尸企业在市场上的频繁活动势必给我国市场机制带来危机，也会对我国经济发展产生严重威胁。

图1- 2分年份僵尸企业统计图

因此，如何根据多维度的企业画像，面向不同的应用场景对企业分类识别，分析挖掘人、自然人、物品（包括特种设备、产品、食品等）、案事件之间的显性和隐性关系，建立市场主体信用管理与风险评估模型，对企业的信用风险、经营风险、质量风险、安全风险、关联风险等进行全面评估及分级预警，为系统监管、精准监管、高效监管提供数据支持，如何利用人工智能等前沿科技来准确分类识别出僵尸企业，以保护市场机制，平衡信贷资源的分配，解决现存的产能过剩问题等越来越受到市场监管部门的重视。

## 1.3项目目标

随着我国经济的发展，现实中有的企业尽管出现亏空、甚至资不抵债，却由于可以获得政府或者金融机构的补贴而得以维持经营，或者由于本国破产程序成本过高而阻碍企业控制权发生转移，困而不变、僵而不死，从而变成“僵尸企业”。僵尸企业的存在不仅浪费社会资源、扰乱市场秩序、脱离资产监管而且严重损害职工权益，容易造成金融危机，因此，从市场上现存的企业中准确高效地识别出僵尸企业刻不容缓。

本项目旨在通过数据治理，采用统计、分析、挖掘等技术，完善企业画像的各维度标签，并采用人工智能算法建立僵尸企业的分类模型，达到对企业进行合理的分类识别。从而避免市场机制遭到破坏以及严重的产能过剩问题。

综上所述，只有尽早发现隐患问题，准确识别出僵尸企业并加以处置，我国经济动能转换以及高质量发展才能得到保障，才能规避系统性的金融风险。

### 1.3.1任务要求

**1.总体要求**

* 1. 对僵尸企业的表现形式有深入的了解和阐述；
  2. 依据用户期望的内容对数据的处理过程进行详细地描述；
  3. 建立模型并对模型有充分的认识和理解；
  4. 保证模型分类识别的精确性与高效性。

**2.技术要求和指标**

开发技术不限，建议使用Python和R语言。

### 1.3.2用户期望

**1.要求**

* 1. 多表数据融合；
  2. 数据特征提取；
  3. 数据预处理；
  4. 数据建模。

**2.目标**

追求模型分类识别的精确性与高效性。

### 1.3.3最终目标

本项目致力于提供一个优质的预测识别系统，系统通过对企业数据的分析实现了对僵尸企业的识别，根据僵尸企业的吸血表现对其进行分类，同时对于非僵尸企业进行了僵尸化风险等级预测。对于僵尸企业的识别，需要制定僵尸企业识别标准，经过一系列分析发现，僵尸企业的特点主要为：

* + 1. 亏损性：企业无望恢复生气，主要表现为以净利润等指标来衡量的财务业绩差和负债率很高；
    2. 吸血性：企业需要依靠长期（三年及以上）的吸血才能继续维持企业现状、继续生存，其主要表现为依靠财政补助、银行贷款等资源；
    3. 绑架性：企业规模庞大，员工多，一旦放弃救助就会导致社会危机，从而强制由银行不断提供贷款，实现银行和企业的绑架。

僵尸企业基本已经不产生效益，但是却依然占有着土地、资本、劳动力等生产活动中所需要的重要资源，从而会严重阻碍我国供给侧结构性改革，阻碍我国经济发展的活力。

因此，本项目通过数据治理，采用统计、分析、挖掘等技术，完善企业画像的各维度标签，并采用人工智能算法建立僵尸企业的分类模型，达到对企业进行合理的分类识别以及风险预测的目的，从而保障我国经济动能转换以及高质量发展，避免带来系统性的金融风险。

**1.战略层目标**

从战略层面来讲，本项目的预测分类模型需要具备高效且精确的特点，同时针对预测过程中的分类标准要具有一定的合理性和适用性。现在市场上的对于僵尸企业的绝大部分识别标准都不准确，例如：CHK标准、FN-CHK标准等，都存在其自身的缺点，同时也无法完全适应我国的实际情况。本项目需要结合CHK标准、FN-CHK标准以及我国各省颁布的对于僵尸企业的定义完善预测分类模型的评判标准，使得模型具有更高的适用性和分类的准确性，同时对于用户来讲，项目组设计了操作简单、界面美观的可视化web操作页面，在简单易懂、易操作的同时支持用户单个或是批量输入查询。以上都是本项目的优势，同时也是本项目所具备的市场竞争力。

**2.操作层目标**

从操作层来说，本项目主要实现以下要求：

1. 实现对于僵尸企业的分类识别模型；
2. 简洁的可操作web页面；
3. 保证模型的高效性和精确性，提高模型泛化能力；
4. 实现可视化web端，允许单个或者批量输入查询分类结果；
5. 系统的异常情况处理。

**3.用户层目标**

从用户层来说，用户是项目的主要接收对象，需要考虑用户对项目的接收程度。对用户的友好性主要体现在项目具有简洁美观且操作简便的web端页面，同时满足用户单个、批量输入查询分类结果的需求。除此之外，模型分类识别的高效性和精确性满足用户对于识别僵尸企业的需求。另外，项目也要从用户的角度进行进一步的完善，基本满足用户对僵尸企业画像及分类的期望，同时本项目考虑到数据安全及隐私，不会设计后台数据库对用户输入数据及查询结果进行保存，因此用户完全可以放心使用本项目提供的模型。

**4.企业层目标**

对于企业而言，本系统具备应有的市场竞争力，且满足客户的需求及客户对本系统的基本期望——操作简单、界面简洁。对于数据治理和模型标签处理以及僵尸企业的评判标准都有本项目的特色，模型在市场上具有一定的竞争力。因此，对于企业来说，本系统也是值得信赖和可靠的。

## 1.4项目意义

在互联网逐渐步入大数据时代后，本项目结合大数据，利用人工智能算法实现了对僵尸企业分类识别的目的，不仅对于我国市场机制以及经济发展具有重大意义，同时推动了互联网及大数据时代的发展，响应了中央出台的一系列政策，在时代层面和政治层面都也具有举足轻重的作用。

**1.市场意义**

僵尸企业缺乏盈利能力，但却以低于市场最低利率成本获得信贷资源，依靠外界输血。因此，僵尸企业的存在破坏了市场机制，加剧了信贷资源错配，带来了严重的产能过剩问题，还对其他非僵尸企业产生了投资挤出效应。

因此，本项目旨在通过对企业各方面数据的综合分析，构建企业全息画像，整体评估企业的综合价值，从而准确识别出僵尸企业，以达到保护我国市场机制，维护现市场上健康企业的权益，给中小型发展前景良好企业提供主动创新、提高效率的动力，维持市场优胜劣汰的机制和公平竞争的环境。

**2.时代意义**

在大数据背景下，本项目能够方便地获得企业从业人数、成立年限、注册资本、营业收入、风险信息、行政处罚、纳税信用登记、黑名单、上市信息、电商信息等数据。然后通过数据治理，并采用统计、分析、挖掘等技术，完善企业画像的各维度标签，采用人工智能算法建立僵尸企业的分类模型。

因此，本项目最终目标的实现，不仅达到了识别僵尸企业的目的，同时也展示了互联网时代人工智能技术越来越成熟，同时也揭示了在大数据背景下各种数据的获得、处理、分析越来越方便，这样的便捷给各行各业的发展都带来了巨大优势。这样的案例，对大数据时代的发展具有极大的推动作用，同时也提升了全民用户对人工智能的信任度和熟悉度，为未来人工智能的大量普及奠定了良好的用户基础。

**3.政治意义**

僵尸企业如果不及时处理，经济动能转换以及高质量发展就无法保障，去杠杆工作的受阻将加大系统性金融风险。鉴于僵尸企业存在的严重危害性，自2015年以来，中央出台了僵尸企业处置的一系列政策指导文件，试图以处置僵尸企业作为深化供给侧结构性改革以及调整经济结构。

然而，处置僵尸企业的前提是能够从市场上准确高效地分类识别出僵尸企业，因此，本项目基于人工智能算法的分类模型对于这一政策的成功实行显得尤为重要。该分类识别模型，不仅能够实现对我国市场机制的保护，同时也是在我国出台一系列有关僵尸企业处置问题的政策后，必然完成的一项目标。然而，面对市场上现存的众多企业以及各种复杂的企业数据，只有利用大数据分析等技术准确、高效地分类识别出僵尸企业，才能使得僵尸企业受到应有的处置，才能保证一系列的有关政策得以成功推行。

## 1.5项目亮点

### 1.5.1识别标准

无论是CHK标准、FN识别法还是FN-CHK标准，都无法适应我国的市场的实际情况，也没有考虑各个省份政策的差异。因此，本项目对各省份依据党中央国务院标准制定的僵尸企业评判标准进行了总结和分析，发现各省份主要从企业生产经营状况、负债率和欠息欠费方面认定僵尸企业。最终提出一个能够从我国国情出发，考虑政策因素的新的识别标准——ALG标准。

主要的识别依据是：

* + 连续三年亏损；
  + 资产负债率三年处于85%以上；
  + 长期（三年及以上）欠薪、欠税、欠息、欠费；
  + 满足以上两个及以上。

### 1.5.2模型融合

模型算法在参考XGBoost的同时，融合了统计模型，即采用数学统计方法来建立模型，经过数理统计法求各变量之间的函数关系。统计模型在对大量随机事件的规律性做出推断时仍然具有统计性，从而有针对性地提高了模型的准确率。除此之外，算法模型对剪枝操作进行了进一步的优化，使得模型收敛速度加快，从而确保模型识别的高效性。

### 1.5.3数据处理

本项目提出的模型在特征选择上结合了特征相关性分析，因此增强了特征之间的相关性，从而提高了模型识别的准确性。同时在缺失值处理上，通过处理争取保留了更多的数据来对模型进行训练，从而提高了模型的查全率和查准率。

### 1.5.4预警功能及预测结果处理

系统在实现了对僵尸企业分类识别功能的基础上，增加了对于非僵尸企业僵尸化的风险等级预测功能，对于非僵尸企业的预测，达到了未雨绸缪的效果，根据提供的风险等级预测结果提前对企业日常经营和经济情况进行调整和修订，避免企业进一步僵尸化从而危害市场。

同时，增加了对于僵尸企业系统提供成因分析，用户可以根据僵尸企业的成因分析结果，修改有关部门对市场上企业监管的侧重点，从而更好地维护市场机制，确保我国经济行稳致远地健康发展。

除此之外，系统再前端操作上提供了僵尸企业识别结果的excel表格的结果下载，满足了用户对预测结果进行后续自定义处理的需求。

系统增加对非僵尸企业风险等级预测功能，僵尸企业成因分析功能以及预测结果自定义处理，使得系统的功能更加完善，同时从用户角度出发确保了更友好的用户体验。

# 2.需求分析

## 2.1需求调研

研究发现，2011年和2012年我国僵尸企业比例约为3.3%，2013年后这一比例升至5%，2016年有所回落；2008年“四万亿”经济刺激计划造成固定资产投资过快增长，可能是2013年后僵尸企业比例上升的原因；我国东、中、西部地区的僵尸企业比例依次递增，黑色金属冶炼、造纸等传统产业的僵尸企业比例最高，国有僵尸企业比例是非国有企业的4倍。

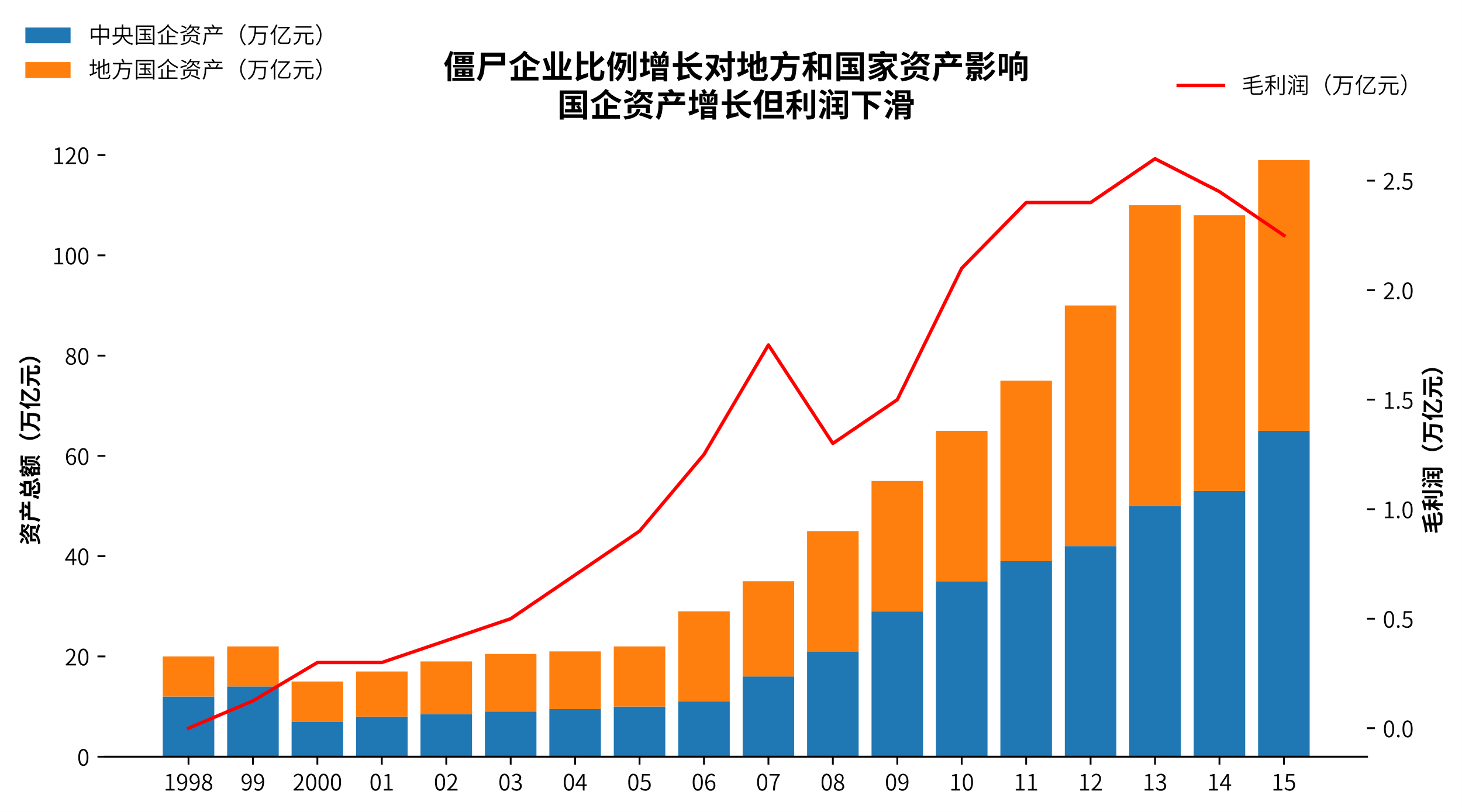


图2- 1僵尸企业比例增长对地方和国家资产影响

近年来，“僵尸企业”成了中国经济转型背景下的热点词汇。2015年11月4日，李克强总理在主持召开国务院常务会议时首次明确提出，“要加快推进僵尸企业重组整合或退出市场”；2016年2月1日，国务院印发了《关于钢铁行业化解过剩产能实现脱贫发展的意见》，决定在钢铁行业率先尝试化解过剩产能、清理僵尸企业；2016年9月，国家工商总局发布《工商总局关于新形势下推进监管方式改革创新的意见》，并指出要“依托大数据加强监管”，“充分发挥大数据在制定完善新型市场监管制度和政策中的作用，搜集掌握经营者、消费者和社会公众的反应，跟踪监测有关制度和政策的实施效果”等；2018年12月4日，国家发展和改革委员会等部门联合发布《关于进一步做好“僵尸企业”及去产能企业债务处置工作的通知》，要求2020年底前全部完成处置工作。自2015年以来，中央出台了僵尸企业处置的一系列政策指导文件，试图以处置僵尸企业作为深化供给侧结构性改革以及调整经济结构。

由此可见，僵尸企业是影响经济可持续发展、破坏金融系统稳定性和阻碍我国经济转型升级的顽疾。国务院、国家工商总局等都已经陆续出台了各种政策旨在能够准确地从海量市场中揪出僵尸企业使其尽快退出市场，加快僵尸企业的清理速度，从而保护市场公平竞争机制，推动我国市场经济行稳致远地发展。然而，目前相关研究还处于初步阶段，无论是僵尸企业的识别标准，还是僵尸企业的分类识别模型都还尚不成熟，因此，很难准确、快速地分类识别出僵尸企业以及进一步从市场中清理僵尸企业。

团队成员通过阅读服务外包命题赛题手册、僵尸企业相关论文，了解国内外有关僵尸企业的研究现状以确定本项目的开发需求。在互联网逐渐步入大数据时代后，获得企业的工商、司法、经营、上市、知识产权、舆情等多维度的数据越来越便捷。为了合理利用这些数据，更快地适应大数据时代，本项目根据获得的企业基本信息、经营状况信息、企业风险信息等，在企业真实数据的基础上为企业建立标签模型体系，将企业的具体行为属性进行标签化，最终形成一个多元化的企业标签对象，即构建企业画像，将零散的企业数据，通过数据治理、特征标签化等处理转换为普通用户能够看懂的信息。这种对数据的统一处理和对数据归纳整理是适应大数据时代必然进行转变。利用大数据及人工智能技术，结合企业画像的理论对企业进行全面分析，为企业成长、行业发展、政府监管等各方面提供可靠的数据分析是目前僵尸企业分类识别项目的市场需求。

## 2.2功能需求

本项目旨在利用人工智能算法，通过数据治理，采用统计、分析、挖掘等技术，完善企业画像的维度标签，达到对僵尸企业进行分类识别的目的。其主要功能需求包括：

* 根据用户提供的企业数据，对多个表格文件进行数据融合；
* 在众多的企业数据中，提取本项目模型需要的数据特征；
* 对用户提供的原始数据进行预处理，避免无效数据对最终结果产生影响；
* 建立算法模型，对用户提交的企业数据进行准确地僵尸企业分类识别；
* 对于模型识别出的僵尸企业，根据该企业数据进行进一步的僵尸企业分类和成因分析；
* web端可视化界面

·僵尸企业分类识别的可视化展示

·僵尸企业信息的数据分析

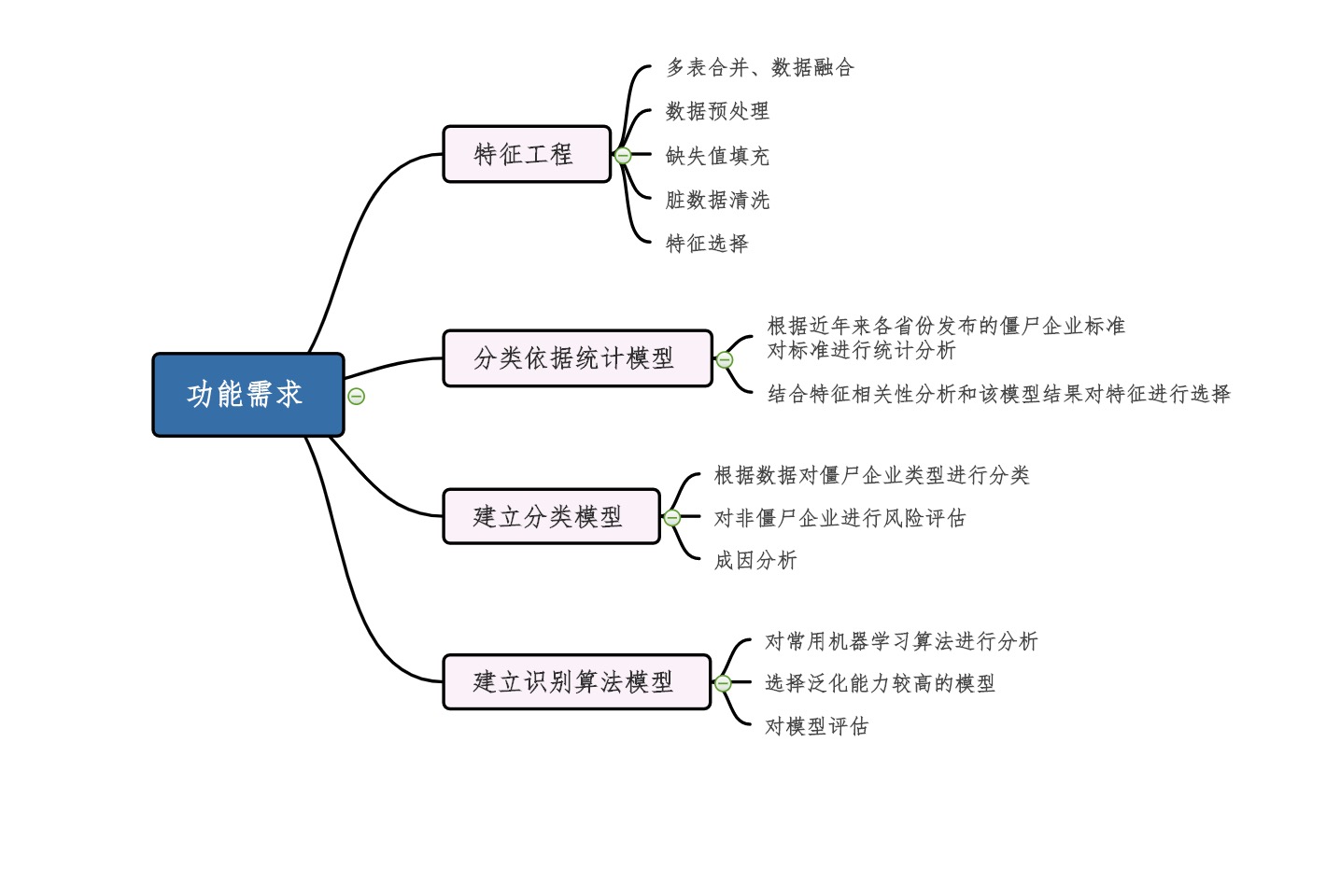


图2- 2系统功能需求图

## 2.3业务分析

### 2.3.1用户特点及其面临的问题

系统面向的用户是市场监督管理局。

市场监督管理局贯彻落实党中央关于市场监督管理工作的方针政策和决策部署，在履行职责过程中坚持和加强党对市场监督管理工作的集中统一领导。其职责包括了之前工商局、质监局、食品监督管理局的职能。作为维护市场公平、保护市场机制，推动我国经济发展的重要职能部门，是负责找出僵尸企业的重要职能部门。

僵尸企业缺乏盈利能力，但却以低于市场最低利率成本获得信贷资源，依靠外界输血，破坏了市场机制，加剧了信贷资源错配，带来了严重的产能过剩问题，打破了市场优胜劣汰的机制和公平竞争的市场环境。由此可见，僵尸企业的分类识别是市场监督管理局一直以来的职责所需，如何准确高效地识别出僵尸企业，令其退出市场是市场监督管理局亟待解决的问题。

### 2.3.2业务流程

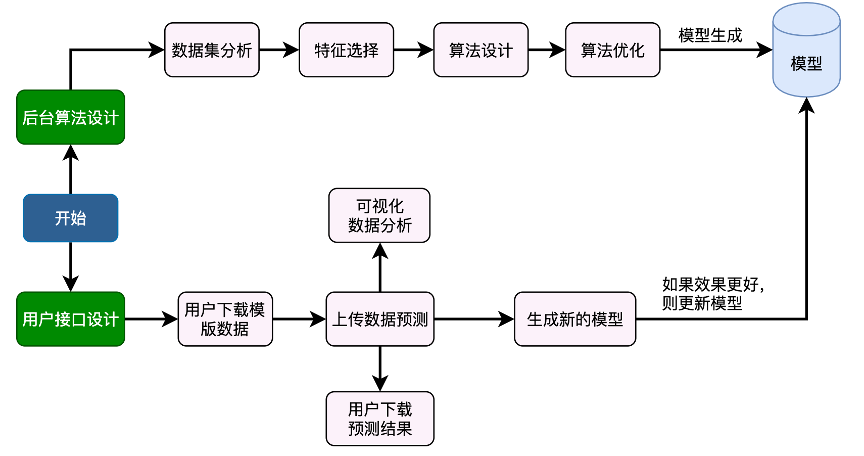


图2- 3系统业务流程图

## 2.4市场分析

### 2.4.1基于PEST模型的市场宏观分析

1. **政治环境**

近年来，“僵尸企业”成了中国经济转型背景下的热点词汇。2014年6月，国务院发布《关于加强金融监管防范金融风险工作情况的报告》，要求严格控制对高能耗、高排放企业和产能过剩行业的贷款，对经营难以为继且产品缺乏竞争力的僵尸企业和项目，要实施破产或兼并重组；2015年11月4日，李克强总理在主持召开国务院常务会议时首次明确提出，“要加快推进僵尸企业重组整合或退出市场”；2015年12月，李克强总理在国务院常务会议上首次对僵尸企业提出了具体的清理标准，即“要对持续亏损3年以上且不符合结构调整方向的企业采取资产重组、产权转让、关闭破产等方式予以‘出清’”；2016年初，国务院国有资产监督管理委员会再次明确表示，要用三年时间基本完成中央企业中345户大中型僵尸企业的清理；2016年2月1日，国务院印发了《关于钢铁行业化解过剩产能实现脱贫发展的意见》，决定在钢铁行业率先尝试化解过剩产能、清理僵尸企业；2016年9月，国家工商总局发布《工商总局关于新形势下推进监管方式改革创新的意见》，并指出要“依托大数据加强监管”，“充分发挥大数据在制定完善新型市场监管制度和政策中的作用，搜集掌握经营者、消费者和社会公众的反应，跟踪监测有关制度和政策的实施效果”等；2018年12月4日，国家发展和改革委员会等部门联合发布《关于进一步做好“僵尸企业”及去产能企业债务处置工作的通知》，要求2020年底前全部完成处置工作。



图2- 4部分中央出台的政策截图

随着僵尸企业的频繁活动，中央出台了一系列僵尸企业处置以及清理的政策指导文件，试图加剧僵尸企业退出市场的进度以保护市场机制和市场健康环境，确保我国经济行稳致远地发展。因此，从海量的企业中分类识别出僵尸企业的任务刻不容缓。

1. **经济环境**

“僵尸企业”大多是已停产半停产、连年亏损、资不抵债，靠政府补贴和银行续贷维持经营的企业，主要集中在一些产能过剩的行业。国务院国资委新闻发言人彭华岗介绍，2018年未纳入专项工作范围的僵尸特困企业比2017年减亏增利373亿元，比2015年减亏增利2007亿元。

僵尸企业的存在破坏了市场机制，加剧了信贷资源的错配，带来了严重的产能过剩问题，还对其他非僵尸企业产生了投资挤出效应。僵尸企业是影响经济可持续发展、破坏金融系统稳定性和阻碍我国经济转型升级的顽疾，若不及时处理，经济动能转换以及高质量发展就无法保障，去杠杆工作的受阻将加大系统性金融风险。

1. **社会环境**

从2016年开始，中央把供给侧结构性改革作为经济体制改革和宏观调控的主要方向。供给侧结构性改革的主要任务是“去产能、去库存、去杠杆、降成本、补短板”。五大任务之首就是去产能，去产能的重点就是处置僵尸企业，而处置僵尸企业的第一步就是从海量的企业中分类识别出僵尸企业。

随后，国家发展改革委等11个部门联合印发《关于进一步做好“僵尸企业”及去产能企业债务处置工作的通知》，明确原则上应在2020年底全部完成处置工作。专家分析，《通知》强调坚持市场化法治化原则，进行分类处置进程有望再提速。

研究发现，2011年和2012年我国僵尸企业比例约为3.3%，2013年后这一比例升至5%，2016年有所回落；2008年“四万亿”经济刺激计划造成固定资产投资过快增长，可能是2013年后僵尸企业比例上升的原因；我国东、中、西部地区的僵尸企业比例依次递增，黑色金属冶炼、造纸等传统产业的僵尸企业比例最高，国有企业僵尸比例是非国有企业的4倍。

面对社会、中央及市场变化的各种压力，市场监督管理局在加快僵尸企业处置速度的同时，不得不采用先进的技术手段尽快揪出僵尸企业，还社会一个公平竞争的市场环境、健康发展的市场交易平台。

1. **技术环境**

随着互联网逐渐步入大时代，大数据、人工智能等技术渐渐渗透于各行各业。在海量的企业中，如何通过数据治理，采用统计、分析、挖掘等技术，完善企业画像的各类维度标签，最后采用人工智能算法建立僵尸企业画像的分类模型，达到对企业进行合理分类识别的目的。

国内华南师范大学计算机专业，提出了一种基于大数据技术的企业画像的构建模型以及构建方法。将企业画像分为五大维度：企业属性维度、企业信用属性维度、企业交易特征维度、企业内外关联特征维度、企业评价信息维度。在数据处理方法上，文献提出三大部分内容：标签提取、关联分析、分类与预测。其中在标签提取上，采用k-means、LDA、NB、CNN等特征提取方法优缺点的比较；关联分析采用Apriori算法和GRI关联规则算法的对比分析表；最后在分类预测上，给出了KNN算法、SVM算法、决策树算法的优缺点对比。

然而，国外对于该行业的研究现状主要为：Turney将机器学习模型应用到关键词中，把关键词提取看作是一个分类问题，使用 C4.5决策树算法和遗传算法进行关键字提取；Mihalcea基于无监督学习的图模型理论，利用共现窗格中词与词的共现关系构造图模型，并依托谷歌公司的网页排序pageRank算法，提出了TextRank关键词提取算法；短文本分类方面，Sen等人抽取数据中的特征项，作为基本单位，构建了一种网络 模型分类，分类性能得到了一定的提高；Sriram等人以Twitter信息为源数据，提出了一种分类方法，在使用常规的特征工程进行特征扩展的情况下，分类的准确率达到了百分之八十左右。

无论国内还是国外，对于僵尸企业的研究投入都越来越大，随着互联网的发展，利用当下的最新科技解决现有的现实问题也是各行业升级发展的必然。因此，如何保证模型的高效性以及分类的准确性是团队考虑的主要问题，也是解决该问题的关键所在。

### 2.4.2市场及行业分析

僵尸企业分类识别模型主要针对市场监督管理部，面对自2012年以来，中央出台的一系列僵尸企业处置文件。至2020年，国务院资委已经全面梳理出央企需要专项处置和治理的“僵尸企业”和特困企业户共2041家，涉及资产更是高达三万亿元，因此对于僵尸企业的识别上，仍存在较大的市场。从各种数据来看，僵尸企业并不仅仅存在于中国，全球发达国家或发展中国家同样存在僵尸企业现象。经济学界第一个提出“僵尸企业”概念的人是美国经济学家肯尼。他指出，20世纪80年代，美国储蓄贷款业陷入危机，尤其表现为“僵尸”储贷机构尾大不掉，并且不断侵蚀健康机构的利润。如果把一般会计准则下有过严重经济损失的储贷机构都看作是僵尸企业的话，那么美国储蓄贷款业的僵尸机构有600多家，涉及资产1300亿美元，占行业总资产的12%。

由此可见，对于僵尸企业的识别模型不仅可以应用于中国，国外很多国家也具有这样的市场需求，但对于僵尸企业的识别和处置是政府长期以来的目标，因此，僵尸企业分类识别模型的用户对象比较单一。

随着“僵尸企业”问题的不断进化，越来越多的人专注对僵尸企业识别的研究，同时也出现了各种僵尸企业的判定标准。其中有：官方标准、CHK标准、FN-CHK标准以及各类FN-CHK修正标准等等。面对有限的市场容量以及越来越激烈的竞争现状，想要在该行业立足就必须提高模型的准确性和高效性，加大对模型实现的研究投入，只有技术上的突破和突出的模型亮点才是在竞争中立足的关键。

# 3.算法模型

## 3.1建模准备

### 3.1.1问题背景

命题单位已经多角度、多层次、多领域汇聚了约不同行业的5万家企业相关信息，其中定义为僵尸企业的有2千家。项目数据包括训练集、开发集和测试集，前期将发布训练集和开发集，测试集将在此后公布，直接用于模型评测。

企业数据一部分为通用的标签信息，其他的新增标签由参赛人员根据经验和数据本身特点，针对具体问题制订；然后通过数据治理，并采用统计、分析、挖掘等技术，完善企业画像的各维度标签；最后采用人工智能算法建立僵尸企业的分类模型，达到对企业进行合理的分类识别的目的。

### 3.1.2待解决的问题

智能高效是当今时代各行各业的主要奋斗目标！目前，僵尸企业的评判标准在各国各省都并不统一。对于我国各省来说，大部分省份都根据省内政策等特性制定了特定的僵尸企业判定标准。除了僵尸企业的标准不一，根据标准对僵尸企业的分类识别更是耗时耗力，面对当今日益增长的企业数量，从中准确高效地分类和识别出僵尸企业变得尤为重要。

根据命题企业给出的企业数据，其中主要包括四个方面：企业的基本信息、企业知识产权拥有情况、企业融资情况以及企业连续三年的盈利情况。在建模之前，需要先对企业提供的数据进行处理，包括无效数据的处理以及特征处理等，以保证数据的有效性以及最终模型的准确性。

1. **针对僵尸企业的识别**

需要拥有准确的评判标准实现二分类的问题，而该评判标准需要根据现有的各种现有标准结合僵尸企业特性进行制定。准确的评判标准就是实现从大量的企业中识别出僵尸企业。由于僵尸企业的形成与企业经营、国家政策等各方面息息相关，因此需要制定普遍适用的评判标准就需要全面考虑各方面的影响，同时忽略我国各个省份之间政策的细微差别，真正做到从大量企业中准确高效地分类识别僵尸企业。

1. **针对僵尸企业的分类**

在识别出来的僵尸企业中，根据僵尸企业不同种类，准确了解僵尸企业不同类别的具体特征，对僵尸企业进行分类处理。了解僵尸企业分类标准以及不同类别的具体特点就能实现对僵尸企业的分类。

1. **针对模型的准确性**

根据给出的数据以及僵尸企业的具体特性，制定更加全面准确的评判标准，同时对模型的参数进行调整，即可保证分类识别模型的准确性。

1. **针对模型的高效性**

对于分类模型采用的算法，在设计过程中使用巧妙的处理，避免过于复杂的计算导致模型分类速度变慢，同时对数据的特征进行正确处理也是确保分类识别模型高效性的办法。

### 3.1.3名词解释

1. **僵尸企业**：指缺乏盈利能力无望恢复生气却能够获得放贷者或政府的支持而免于倒闭，是指依靠外界输血而免于倒闭的负债企业。
2. **注册资本：**指合营企业在登记管理机构登记的资本总额，是总营各方已经缴纳的或合营者承诺一定要缴纳的出资额的总和。
3. **企业类型：**法律对不同类别企业的具体需求，如设立的条件、设立的程序、内部组织机构等来组建企业。企业主要分类有：合资、独资、国有、私营、全民所有制、集体所有制、股份制、有限责任等等。
4. **债券融资：**是有偿使用企业外部资金的一种融资方式。包括：银行贷款、银行短期融资（票据、应收账款、信用证等）、企业短期融资券、企业债券、金融租赁、政府贴息贷款等等。
5. **股权融资：**指企业的股东愿意让出部分企业所有权，通过企业增资的方式引进新的股东，同时使总股本增加的融资方式。
6. **内部融资：**指公司经营活动结果产生的资金，即公司内部融通资金，它主要由留存收益和折旧构成，指企业不断将自己的储蓄转化为投资的过程。
7. **项目融资：**指贷款人向特定的工程项目提供贷款协议融资，对于该项目所产生的现金流量享有偿债请求权，并以该项目资产作为附属担保的融资类型。
8. **政策融资：**以政府信用为担保的，政策性银行或其他银行对一定的项目提供的金融支持，主要以低利率甚至无息贷款的形式，其针对性强，发挥金融作用强。
9. **资产总额：**指企业资产负债表的资产总计项。即是指企业拥有或控制的全部资产，这些资产包括流动资产、长期投资、固定资产、无形及递延资产、其他长期资产等。
10. **负债总额：**指过去的交易、事项形成的现时义务，履行该义务预期会导致经济利益流出企业，包括流动负债和长期负债。
11. **营业总收入：**指企业在从事销售商品，提供劳务和让渡资产使用权等日常经营业务过程中所形成的经济利益的总流入。
12. **利润总额：**指税前利润，也就是企业在所得税前一定时期内经营活动的总成果。
13. **净利润：**指在利润总额中按规定缴纳了所得税以后公司的利润留存，一般也称为税后利润或净收入。
14. **纳税总额：**指应交税金科目和其他应交款科目和管理费用的二级科目税金的借方金额的总和。
15. **所有者权益合计：**指企业投资人对企业净资产的所有权。

## 3.2模型特征选择

### 3.2.1 僵尸企业认定标准发展历史

僵尸企业识别研究早期关注资产负债率和现金流 (Tosjitaka et al.，2003;Peek & Rosengren，2005)，Caballero从信贷的角度测算企业最优利率，通过企业最优利率和实际支付利率进行比较以识别20世纪90年代日本僵尸企业，从而建立起识别僵尸企业的**CHK标准**。CHK标准符合当时日本企业实际，在国际上得到了普遍认同，但该方法仅考虑企业银行信贷优惠，标准单一且未能考虑 企业借短还长、借新还旧行为，因而往往容易漏判或误判。Fukuda和Nakamura(2008)对此进行了修正，建立了**FN识别法**，该方法综合考虑资产负债率、持续信贷和盈利，使测度结果大大改善，但仅依靠一年数据进行盈利分析，同时忽略了优质企业往往容易获得低息贷款的情况，因而也不够准确，且仍不符合我国僵尸企业的实际，该方法并未考虑政府补助。我国僵尸企业主要集中在重化工业且以国有企业为主，要素效率低下，创新意识薄弱，通常具有长期的依赖性、亏损性和负增长性等共同特点。因此，我国对僵尸企业的识别及标准与国外有着较大的差异。

基于**FN-CHK标准**，董登新(2015)较早地提出了可考虑政府补助的识别标准，张栋等(2016)将扣除政府补助后的盈利纳入CHK标准，来识别中国钢铁上市僵尸企业，朱鹤、何帆(2016)则构建考察银行贷款优惠和政府补助的四种单条件识别方法和考察过度借贷、连续亏损等的三种多条件识别方法对其进行识别。进而，造血能力、输血标准、时间标准等被纳入僵尸指数模型指标体系当中（栾甫贵、刘梅，2018)，也有研究者从僵尸表现、财务状况(盈利能力、营运 能力、成长能力、偿债能力)、企业特征等方面建立僵尸企业预警体系(周琎，2018)。相对于学术界，政府则更倾向于采用简便的标准快速识别僵尸企业，国务院、工信部及各省分别提出了僵尸企业的识别方法及具体标准，浙江省早在本世纪初就在全国率先建立了零用电量、半年以上零税收、资产负债率大于85%且连年亏损等三类僵尸企业识别标准。

### 3.2.2部分省份的僵尸企业认定标准

根据党中央国务院的标准，各省份在此标准基础上根据各省政策特点进行了细化，见下表。

表3- 1各省份标准统计表

|  |  |
| --- | --- |
| 省份 | 标准 |
| 陕西省 | * 连续三年亏损；资产负债率三年处于85%以上； * 经营活动产生现金净流量连续三年为负数； * 连续三年扣除非经常性损益后的息税前利润小于借款费用； * 产能利用率低于50%； * 停产半年以上； * 无法偿还到期负债且持续一年以上。 |
| 青海省 | * 连续三年亏损 * 资产负债率三年处于85%以上； * 经营活动产生现金流量连续三年为负数； * 连续三年扣除非经常性损益后的息税前利润小于借款费用； * 产能利用率低于50%； * 停产半年以上（半年以上未交流转税）； * 无法偿还到期负债（主要指银行借款）且持续一年以上时间。 |
| 重庆省 | * 生产经营困难，资产负债率超过85% * 连续亏损3年及以上，长期（三年及以上）欠薪、欠税、欠费等无法经营的，主要靠政府贷款和银行贷款方式来支持经营的企业。 |
| 甘肃省 | * 连续三年亏损企业、资不抵债企业、停业企业、三级以下控股企业； * 连续三年无收益投资、余额300万元以下投资、三级以下企业的参股投资； * 现金流为负的企业或生产线，两年内不能投产的项目或投产后没有效益的“四类控股企业”和“三类参股投资”。 |
| 福建省 | * 主要指已停产半停产、连年亏损、资不抵债、靠政府补贴和银行续贷存在且不符合国家产业结构调整方向的规模以上工业企业。 * 连续1年以上停止生产经营活动且末缴增值税； * 或者资产负债率超过85%； * 主要靠政府或集团总部补贴和银行续贷维持生产经营，且剔除补贴后连续三年以圬损； * 资不抵债且解套无望、长期停产腹产无望的规模以上工业企业。 |
| 广东省 | * 主要靠政府补贴或银行续贷等方式维持生产经营、资产负债率超过85%； * 且连续亏损3年以上； * 连续三年以上欠薪欠款欠息欠费； * 生产经营困难造成停产半年以上或半停产1年以上的工业企业。 |
| 浙江省 | * 企业利润不够支付企业信贷资金利息的企业，主要标志是企业亏损、资产负债率高； * 企业停止经营活动半年以上，主要标志是半年以上未上交增值税； * 企业生产经营活动基本属于停顿状态，主要标志是企业申请暂停用电容量。 |
| 山东省 | * 一类是企业还在生产经营，但效率低下，利润不够支付企业信贷利息，只要标志是持续亏损，资产负债率高； * 二类是企业已停止经营活动半年以上，主要标志是半年以上未上交增值税； * 三类是企业生产经营活动基本属于停顿状态，主要标志是企业暂停用电容量。 |
| 湖北省 | * 生产经营困难造成停产半年以上或半停产1年以上； * 资产负债率高且连续亏损3年以上；主要靠政府补贴或银行续贷等方案维持生产经营； * 长期欠薪、欠税、欠息、欠费。 * 满足以上任何两项及以上调钱，中型及以上规模的工业企业。 |

根据上表不难发现，各省份主要是从企业生产经营状况、负债率和欠息欠费方面进行认定的。他们的共有特点有：

1. 连续三年亏损；
2. 资产负债率三年处于85%以上；
3. 长期（三年及以上）四欠；
4. 满足两个及以上。

利用这些特点可以在企业提供的数据集中提取出对僵尸企业识别算法更有帮助的特征，然后再根据这些特征的特点进行数据预处理。

## 3.3数据处理

数据处理流程图如下：

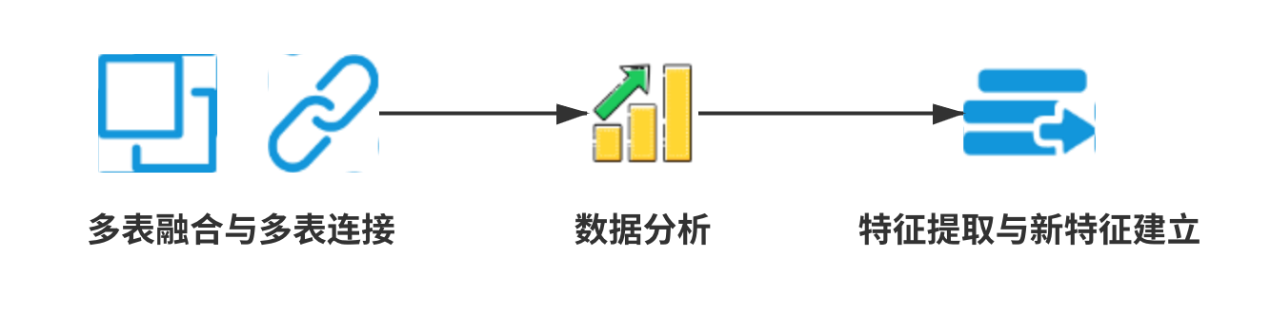


图3- 1数据处理流程图

### 3.3.1 数据概要

企业提供的数据集中包含了以下29个特征：

表3- 2 29个原始特征表

|  |  |
| --- | --- |
| 特征名称 | 特征描述 |
| ID | 企业ID，在数据合并中作为连接主键 |
| 注册时间 | 企业注册时间 |
| 注册资本 | 企业注册资本 |
| 行业 | 企业所属行业 |
| 区域 | 企业所在区域 |
| 企业类型 | 企业所属类别 |
| 控制人类型 | 自然人/法人 |
| 控制人ID | 控制人ID，在企业答疑中该特征被删去 |
| 控制人持股 | 控制人持股比例 |
| 专利 | 企业专利数量 |
| 商标 | 企业商标数量 |
| 著作权 | 企业著作权数量 |
| 从业人数 | 企业从业人数 |
| 资产总额 | 企业资产总额 |
| 负债总额 | 企业负债总额 |
| 营业总收入 | 企业营业总收入 |
| 主营业务收入 | 企业主营业务收入 |
| 利润总额 | 企业利润总额 |
| 净利润 | 企业净利润 |
| 纳税总额 | 企业纳税总额 |
| 所有者权益合计 | 企业所有者权益合计 |
| 债权融资额度 | 企业债权融资 |
| 债权融资成本 | 企业债券融资利息 |
| 股权融资额度 | 企业股权融资 |
| 股权融资成本 | 企业股权融资利息 |
| 内部融资和贸易融资额度 | 企业内部融资和贸易融资 |
| 内部融资和贸易融资成本 | 企业内部融资和贸易融资利息 |
| 项目融资和政策融资额度 | 企业项目融资和政策融资 |
| 项目融资和政策融资成本 | 企业项目融资和政策融资利息 |

原始数据集中数据众多，特征高达29个。所有的特征全部放入僵尸企业识别模型会影响模型的训练效果以及最终模型预测的准确率和高效性。因此依据特征相关性分析结果以及从部分省份僵尸企业认定标准中提取出的四个特点，决定将29个原始特征，减少至6个特征，，而被删除的其他23个特征，与预测标签之间的相关性在0.01以下。如果将相关性过低的特征引入模型，不仅会影响算法复杂度，更会影响算法准确性。因此将原有特征维度减少至6个是合理的。保留特征信息如下：

表3- 3处理后保留特征列表

|  |  |  |
| --- | --- | --- |
| 特征名称 | 特征类型 | 说明 |
| 亏损 | 新增特征 | 企业连续三年净利润信息合并为亏损特征 |
| 负债 | 新增特征 | 将企业连续三年负债信息合并为负债特征 |
| 债权融资信息 | 原有特征 | 企业原有融资信息 |
| 股权融资信息 | 原有特征 |
| 内部融资和贸易融资信息 | 原有特征 |
| 项目融资和政策融资信息 | 原有特征 |

### 3.3.2 多表融合与多表连接

1. **无用样本删除**

对数据集中原始标签为空值的数据进行删除操作。因为在建模准备过程中希望采用的两个算法模型：岭回归、XGBoost都属于监督学习，这两个算法都需要样本特征和样本标签的输入，因此对原始标签为空值的数据进行删除。

1. **多表融合与多表连接**

删除无用样本后，训练集样本数量只有5073条，与验证集的样本数量相差过大。因此，将训练集与验证集合并，共同进行后续的数据处理。数据合并的基本流程为：

1. 将训练集与验证集合并。对于企业基本信息，验证集中存在的“控制人ID”特征在训练集中并不存在，企业答疑中该特征也被删去，因此将该特征删除；
2. 以企业ID为主键，连接企业基本信息和企业专利信息，生成企业基本信息数据集。
3. 企业融资信息、企业年报两个数据集中，一个企业ID对应3条数据，所以要先对这两张表中的“year”特征进行缺失值填充。目前主流的缺失值填充方法，例如均值填充、中位数填充等并不适用于该特征的缺失值填充。通过观察发现，每个企业ID只有2015，2016，2017三年的信息，因此遍历有年份缺失的企业ID信息，找到缺失的年份进行填充。这样的不足：

* 如果一个企业ID对应的三条数据中，缺失了两个年份无法直接确定填充值；
* 遍历时间过长

改进方法：如果一个企业ID对应的三条数据中，缺失了两个年份，就对这两个年份进行排序，填充到空值中。对于遍历时间过长的问题，使用Python的内置函数map进行解决，多线程执行缺失值填充，加快填充速度。

1. 以用户ID、year两个特征作为主键，连接企业融资信息与企业年报，生成企业历年信息数据集；
2. 将企业基础信息中的原始标签添加到企业历年信息数据集中。

企业基本信息数据集、企业历年信息数据集中的特征如下：

多表融合与多表连接流程图如下：

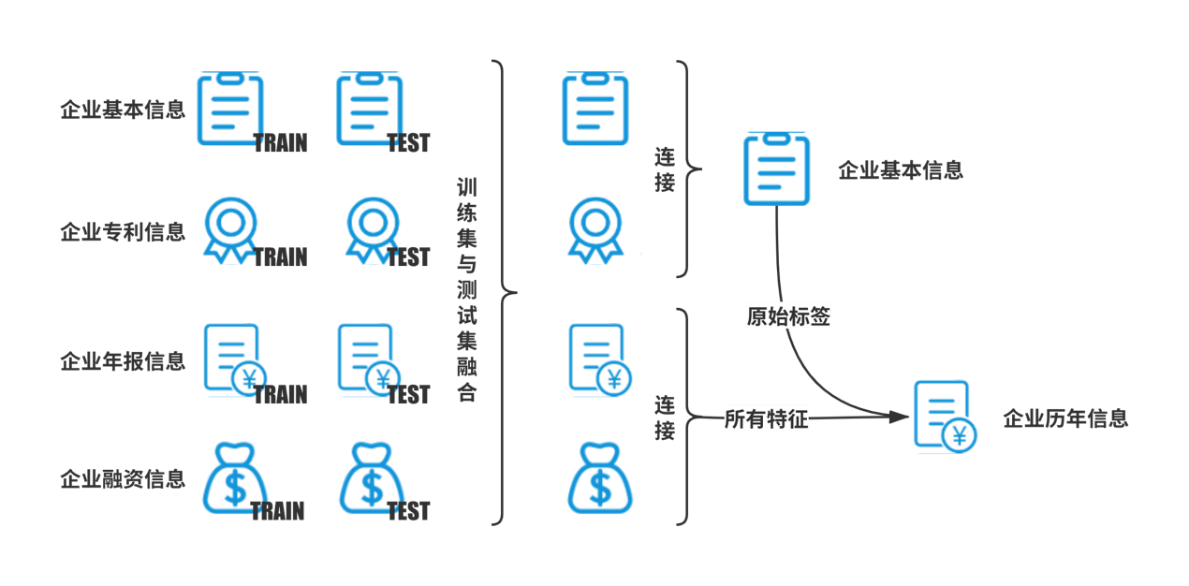


图3- 2多表融合与多表连接流程图

数据合并结束后，整个数据集分为企业基本信息和企业历年信息两张表，如下图所示：

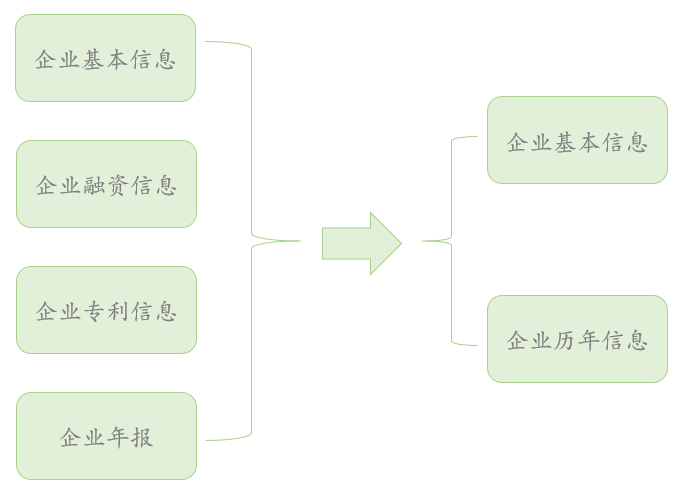


图3- 3数据合并结果图

### 3.3.3 数据分析

在数据分析中，某一特征的数据分布可以反映在数据集中的特征相关性，如果一个特征的数据分布十分均匀，那么这个特征在数据集中就是无用的，可以删除。

企业基本信息中，行业、区域、企业类型、控制人类型等特征，数据分布十分均匀，从而认为这些特征是无用的，在最后的特征提取与特征选择上不会用到这些数据。

团队对国际僵尸企业识别标准以及国内9个省份的僵尸企业识别标准进行了分析，发现这些标准共有的特征是：连续三年亏损，三年负债率大于0.85，长期四欠。对应到本项目数据集中的特征是净利润、负债总额、资产总额、融资信息。因此，在后续的特征提取与模型训练过程中，使用净利润、负债总额、资产总额、融资信息这些原始特征进行新特征建立与模型拟合。

### 3.3.4 缺失值填充

经过查阅资料和数据统计分析发现，缺失值可以通过经济学公式对数据进行填充，比如：资产总额-负债总额=所有者权益合计；当年净利润小于0，纳税总额均为0，反之亦然；融资额度信息要么同时为0，要么同时大于0，而且融资额度和融资成本之间存在着固定的比例关系。由此，我们对数据缺失值进行了填充，填充后存在nan值的样本仅有100多个。最后删除存在nan值的样本。这样的填充方法不仅保证了填充数据的准确性，还避免了部分样本的缺失，有利于增强模型的鲁棒性和泛化能力。

### 3.3.5 特征提取与新特征建立

对部分省份僵尸企业认定标准进行统计分析后，得出了僵尸企业普遍存在的三个问题：连续三年亏损、连续三年资产负债率达85%以上、长期四欠。满足这三个特点中的两个及以上就将该企业认定为僵尸企业。这三个特点对应了数据集中的以下特征：

表3- 4僵尸企业普遍存在的三个问题对应的原始特征

|  |  |
| --- | --- |
| 特征名称 | 特征描述 |
| 资产总额 | 企业年报中的资产总额 |
| 负债总额 | 企业年报中的负债总额 |
| 净利润 | 企业年报中的净利润 |
| 融资信息 | 四大融资方式的额度、成本信息 |

这些特征都在企业历年信息数据集中。但是原始数据集中存在的问题是：每一个企业对应了3条数据，这样的数据直接送入模型进行拟合不仅会导致模型过拟合，泛化能力减弱，也不符合僵尸企业的“长期性”，即连续三年亏损、连续三年资产负债率达85%以上、长期四欠这三个特点。因此我们需要对每个企业的三条数据进行合并并建立两个新特征：

表3- 5建立的两个新特征

|  |  |
| --- | --- |
| 特征名称 | 特征描述 |
| 亏损 | 对应的原始特征：净利润  是否连续三年亏损，是为1，不是为0 |
| 负债 | 对应的原始特征：资产总额、负债总额  是否连续三年负债率均值在85%以上，是为1，不是为0 |

再对数据进行分析，数据集中的融资信息，额度与成本之间存在固定的比例关系，比例关系如下：

表3- 6各种融资方式的利率

|  |  |
| --- | --- |
| 融资方式 | 利率 |
| 债权融资 | 8% |
| 股权融资 | 4% |
| 内部融资和贸易融资 | 6% |
| 项目融资和政策融资 | 6% |

由于融资成本与融资额度之间存在线性关系，因此融资信息只保留融资额度信息即可。再对连续三年各种融资方式的额度信息分别进行求和。提取出的特征信息如下：

表3- 7提取出的融资额度特征

|  |  |
| --- | --- |
| 特征名称 | 特征描述 |
| 债权融资额度 | 对应的原始特征：债权融资额度  对原有的三年债权融资额度进行求和 |
| 股权融资额度 | 对应的原始特征：股权融资额度  对原有的三年股权融资额度进行求和 |
| 内部融资和贸易融资额度 | 对应的原始特征：内部融资和贸易融资额度  对原有的三年内部融资和贸易融资额度进行求和 |
| 项目融资和政策融资额度 | 对应的原始特征：项目融资和政策融资额度  对原有的三年项目融资和政策融资额度进行求和 |

最后，经过特征提取与新特征建立，最终选择的特征信息如下：

表3- 8最终选择的特征

|  |  |
| --- | --- |
| 特征名称 | 特征描述 |
| 亏损 | 对应的原始特征：净利润  是否连续三年亏损，是为1，不是为0 |
| 负债 | 对应的原始特征：资产总额、负债总额  是否连续三年负债率均值在85%以上，是为1，不是为0 |
| 债权融资额度 | 对应的原始特征：债权融资额度  对原有的三年债权融资额度进行求和 |
| 股权融资额度 | 对应的原始特征：股权融资额度  对原有的三年股权融资额度进行求和 |
| 内部融资和贸易融资额度 | 对应的原始特征：内部融资和贸易融资额度  对原有的三年内部融资和贸易融资额度进行求和 |
| 项目融资和政策融资额度 | 对应的原始特征：项目融资和政策融资额度  对原有的三年项目融资和政策融资额度进行求和 |

特征提取与新特征建立流程图如下：



图3- 4特征提取与新特征建立流程图

## 3.4模型概述

### 3.4.1模型总体设计

本项目模型的总体设计如下图所示：

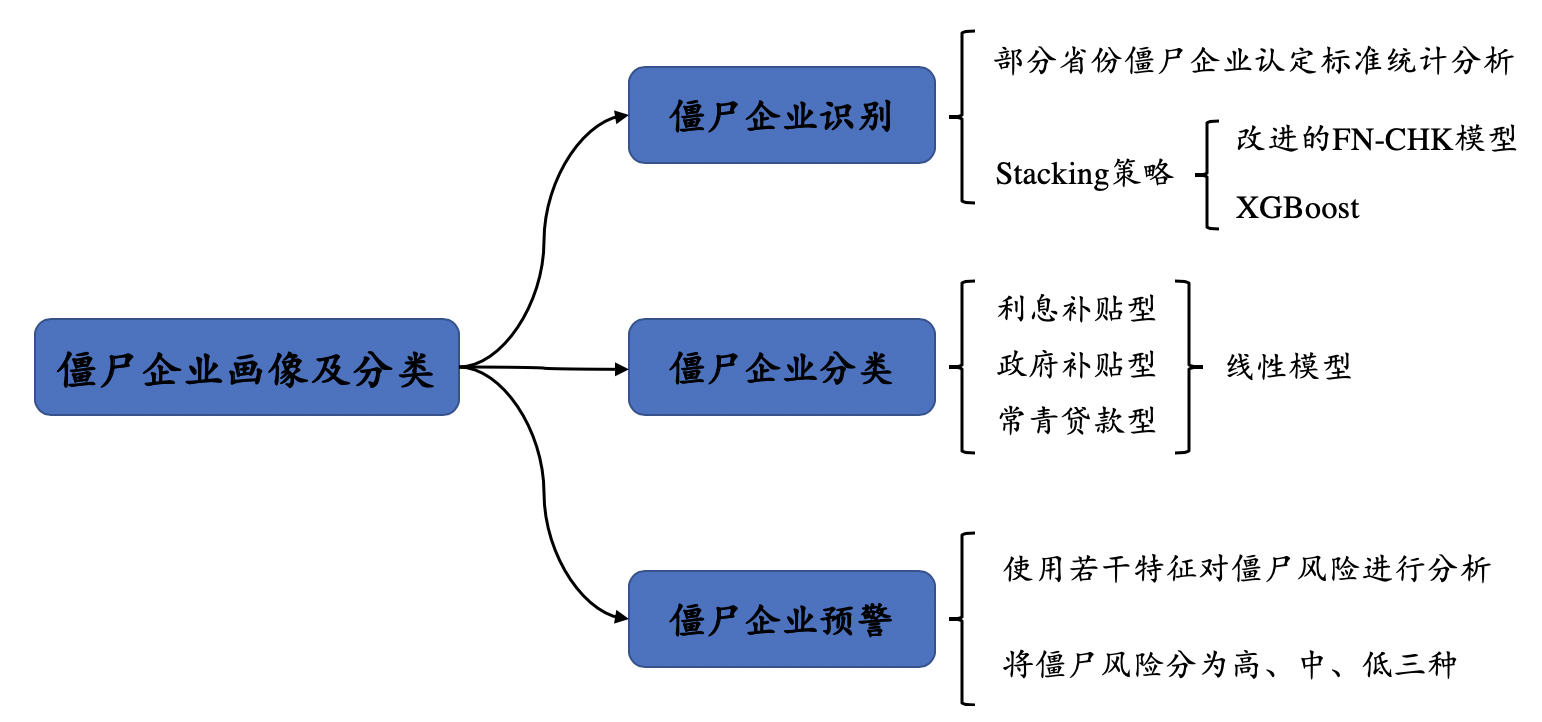


图3- 5模型总体设计图

### 3.4.2模型详细描述

接下来分别从僵尸企业识别、分类、预测对模型进行介绍。

**1.僵尸企业识别模型**

针对僵尸企业识别问题，为了使模型的准确性更高，防止模型过拟合，团队参考XGBoost确定了本系统的算法模型:基于Stacking策略的XGBoost算法。该算法将改进后的FN-CHK模型与XGBoost梯度提升树算法进行集成，模型训练时会考虑更多的特征维度，进而使得整个模型的准确率、泛化能力得到提升。

（1） XGBoost：XGBoost的核心思想是在梯度提升树的基础上加入了正则化惩罚项并使用二阶泰勒展开求出伪残差，相当于梯度提升树和岭回归的结合。

在目标函数中加入正则化项：

（1） XGBoost

XGBoost的核心思想是在梯度提升树的基础上加入了正则化惩罚项并使用二阶泰勒展开求出伪残差，相当于梯度提升树和岭回归的结合。

在目标函数中加入正则化项：

使用二阶泰勒展开，求伪残差：

适用于正则化项相关的lambda、gamma作为分割点的标准：

（2） 改进的FN-CHK模型：结合已有论文，对FN-CHK模型进行改进。算法模型如下：

原有的FN-CHK模型模型为：

表3- 9式(3.7)(3.8)参数说明表

|  |  |
| --- | --- |
| **参数** | **参数说明** |
|  | t-1年的最低短期借款利率 |
|  | 年末短期银行借款余额 |
|  | t-5到t-1年的最低长期借款利率 |
|  | 长期银行借款余额 |
|  | t-5到t-1年的最低可转债利率 |
|  | 未清偿债券余额 |
|  | 实际利息支出 |
|  | 利息支付下限 |
|  | 超额利息支付比例 |

FN-CHK模型假设的僵尸企业吸血性仅表现为对银行资金的依赖，然而，我国的实际情况是，大部分僵尸企业依靠政府补贴持续经营。除此之外，FN-CHK以一年的数据为根据判断僵尸企业的类别，这与僵尸企业的长期性不符。这些原因最终会导致低估僵尸企业数量，因此市场上很多僵尸企业没有被识别出，没有达到用户想要的效果。

参考已有论文，对FN-CHK模型进行改进，改进后的模型为：

与原有的FN-CHK模型相比，改进后的模型使用未清偿债券利率取代了原来的t-5到t-1年的最低可转债利率。改进后的模型打破了原有模型的局限性，将政府补贴、常青贷款也加入僵尸企业分类标准中，由此发现更多的僵尸企业。

由于题目提供的数据集中没有模型中对应的全部数据，因此团队针对数据集的特殊性对改进后的FN-CHK模型进一步进行修改。修改后的模型为：

表3- 10修改后的FN-CHK模型公式参数说明表

|  |  |
| --- | --- |
| **参数** | **参数说明** |
|  | 四种融资方式利率均值 |
|  | 每年四种融资方式额度之和 |
|  | 每年负债总额 |
|  | 每年四种融资方式成本之和 |
|  | 利息支付下限 |
|  | 实际支出利息 |
|  | 超额利息支付比例 |

最后运用Stacking策略将两个算法集成，算法集成的具体流程如下图所示：

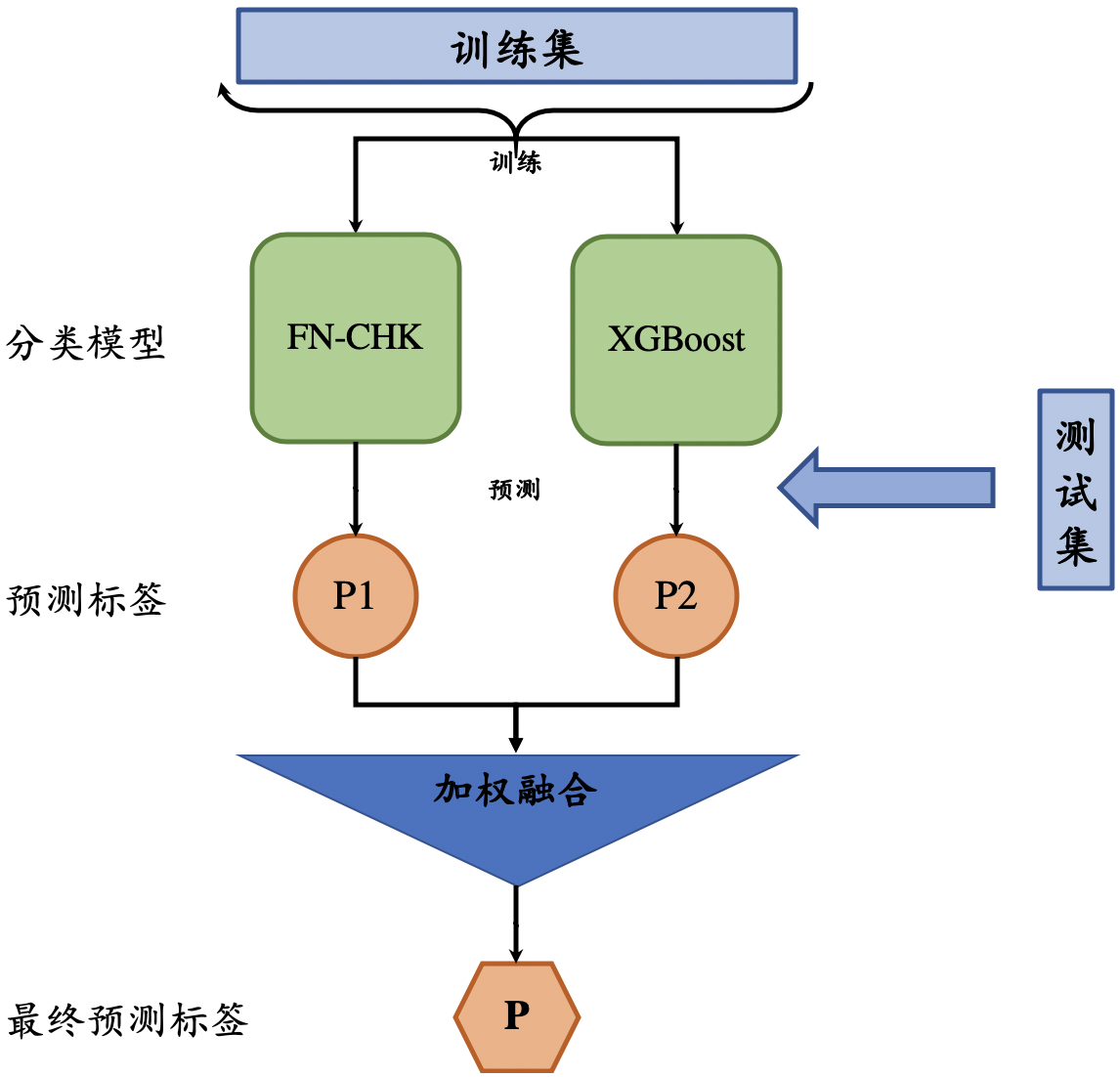


图3- 6基于Stacking策略的XGBoost算法

**2.僵尸企业分类模型**

针对僵尸企业的分类问题，结合现有论文资料发现目前僵尸企业大体分为三类：利息补贴型、政府补贴型、常青贷款型。僵尸企业分类模型通过以下方式对僵尸企业进行分类：

（1） 计算式(3.12)中的EIR指标区分第I类“利息补贴型”僵尸企业；

（2） 计算政府补贴依赖程度指标区分第II类“政府补贴型”僵尸企业；

（3）计算最低利息保障倍数指标区分第III类“常青贷款”型僵尸企业

各指标计算公式如下：

表3- 11式(3.13)(3.14)(3.15)中各参数说明表

|  |  |
| --- | --- |
| **参数** | **参数说明** |
|  | 实际利息支出，计算方式见式(3.11) |
|  | 利息支付下限，计算方式见式(3.10) |
|  | 项目融资和政府融资额度 |
|  | 净利润 |
|  | 企业息税前利润，即数据集中的利润总额 |
|  | 超额利息支付比例 |
|  | 政府补贴依赖程度 |
|  | 最低利息保障倍数 |

三类僵尸企业的具体表现和判别标准，如下表所示：

表3- 12僵尸企业分类标准表

|  |  |  |  |
| --- | --- | --- | --- |
| 类型 | 吸血表现 | 判别标准 | 僵尸性特征描述 |
| 第I类 | 利息补贴 |  | 日常经营活动亏损，需要依靠利息补贴来维持经营；  即便日常经营活动可能产生利润，但息税前利润不足以支付利息，需要利率折让才能够偿还利息 |
| 第II类 | 政府补贴 |  | 日常经营活动亏损，需要靠政府补贴来维持净利润为正，从而免于被“特别处理”；日常经营活动亏损，以政府补贴弥补超过50%的亏损，但最终净利润仍为负 |
| 第III类 | 常青贷款 | 借款总额增加 | 日常经营活动亏损，获得“常青贷款”，以“借新还旧”的形式维持经营；日常经营活动能够创造利润，但息税前利润不足以支付利息，只能以“常青贷款”来偿还旧债利息 |

(注：指负债率)

1. **僵尸企业风险预警模型**

针对非僵尸企业，本模型会从僵尸表现、财务状况、企业特征三个维度选取若干指标对其进行僵尸化风险预测，最后给出当前企业所属的风险等级。根据不同的吸血表现和风险阈值，分为高风险、中风险、低风险三类。

风险指数的计算公式如下：

表3- 13式(3.16)(3.17)参数说明表

|  |  |
| --- | --- |
| 参数 | 参数说明 |
|  | 采用逐步回归估计法计算出的偏置 |
|  | 采用逐步回归估计法计算出的特征参数  以消除特征之间的多重共线性 |
|  | 选取的预警体系指标，具体指标说明见表3-14 |
|  | 归一化函数 |
|  | 最小值 |
|  | 最大值 |
|  | 风险指数 |

表3- 14选取得预警体系指标

|  |  |  |  |
| --- | --- | --- | --- |
| **一级指标** | **二级指标** | **指标名称** | **指标说明** |
| 僵尸表现 | 僵尸性程度 | 超额利息支付比例 | 见式(3.13) |
| 最低利息保障倍数 | 见式(3.14) |
| 政府补贴依赖程度 | 见式(3.15) |
| 财务状况 | 盈利能力 | 资产收益率 | 净利润/平均总资产 |
| 偿债能力 | 资产负债率 | 总负债/总资产 |
| 营运能力 | 应收帐款周转率 | 营业收入/平均应收帐款 |
| 总资产周转率 | 销售收入/平均资产总额 |
| 成长能力 | 营业收入增长率 | 营业收入增加值/上期营业收入 |
| 净利润增长率 | 当期净利润增长值/上期净利润 |
| 总资产增长率 | 当期资产增长值/上期资产总额 |
| 企业特征 | 企业规模 | 总资产规模 | 总资产的对数值 |
| 员工规模 | 员工人数的对数值 |
| 所属地区 | 所属地区 | 所属地区分类代号 |

风险等级判别标准如下表所示：

表3- 15风险等级评估表

|  |  |  |
| --- | --- | --- |
| 风险等级 | 判别标准 | 风险描述 |
| 低风险 | 吸血表现为利息补贴 | 日常经营亏损，需要依靠利息补贴来维持净利润为正，以避免被识别成僵尸企业  风险较低 |
| 中风险 | 吸血表现为政府补贴 | 依靠政府补贴维持净利润为正，以避免被特殊处理  风险中等 |
| 高风险 | 吸血表现为常青贷款 | 依靠常青贷款，保证其他贷款利息的正常支付  风险较高 |

## 3.5参数调优

由于解决僵尸企业分类、非僵尸企业风险预警两个问题所使用的方法不属于机器学习的范畴，因此本章节仅对僵尸企业识别问题中所使用的XGBoost模型的参数调优进行说明。

对XGBoost模型准确度影响较大的参数共有3个，如下表所示：

表3- 16XGBoost参数说明

|  |  |
| --- | --- |
| 参数 | 参数说明 |
|  | 最小子树权重 |
|  | 树的最大深度 |
|  | 学习率 |

团队使用网格搜索法对参数进行调优并将模型拟合过程中损失函数的变化可视化。在1.5至1.9的范围内改变最小子树权重的值，得到的损失函数变化图如下图所示：

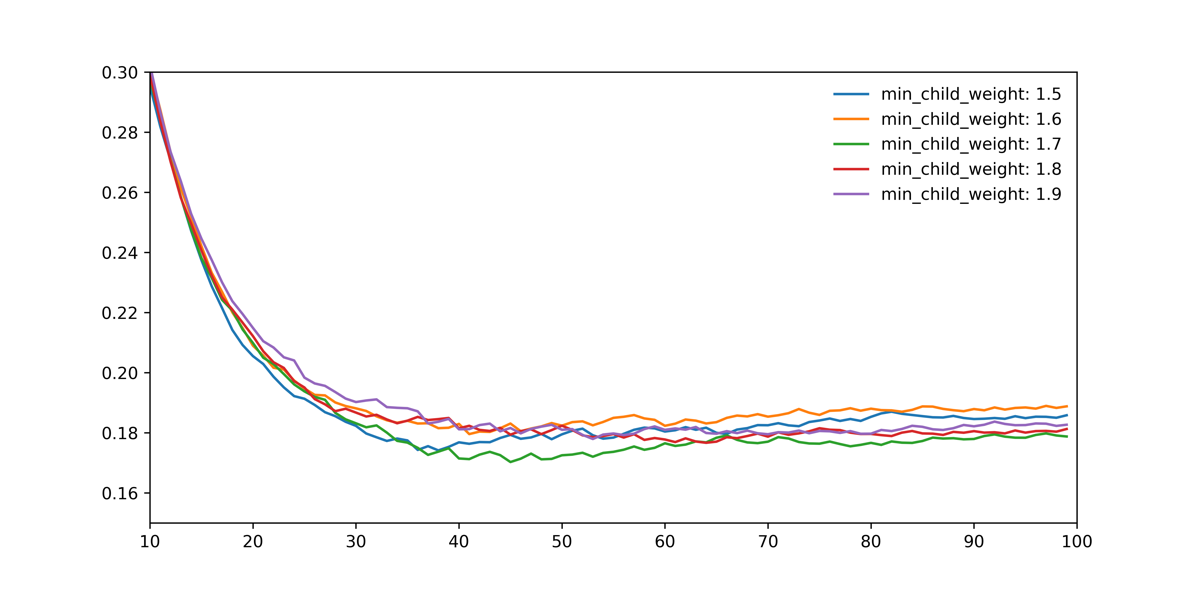


图3- 7损失函数变化图-改变最小子树权重

在5至9的范围内改变树最大深度的值，得到的损失函数变化图如下图所示：

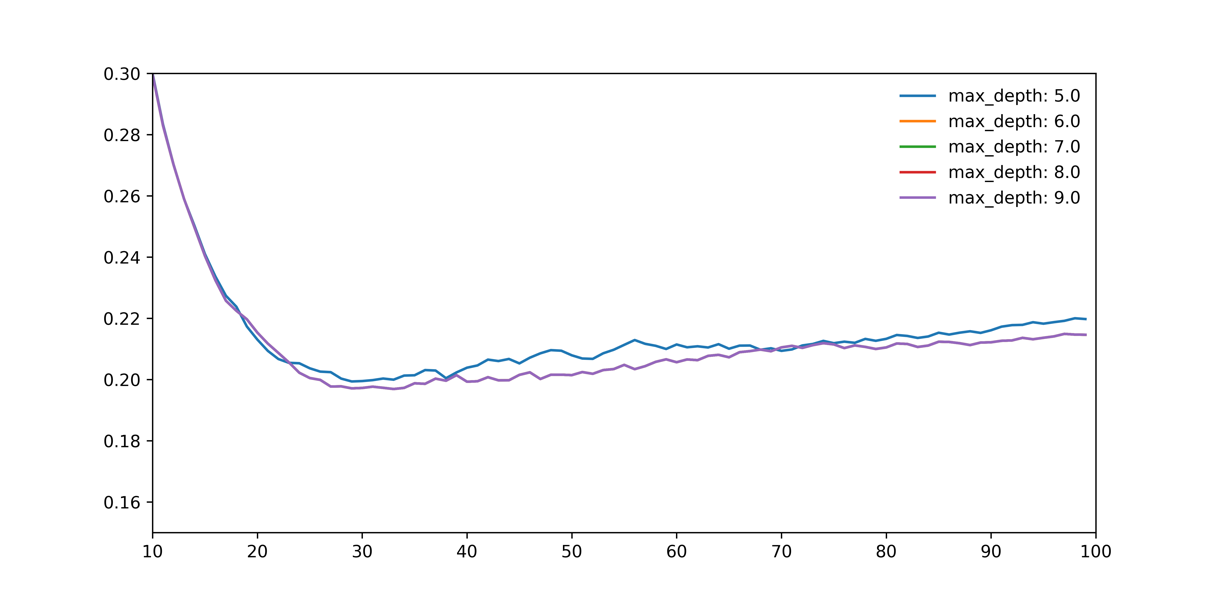


图3- 8 损失函数变化图-改变数的最大深度

在0.1至0.6的范围内改变学习率，得到的损失函数变化图如下图所示：

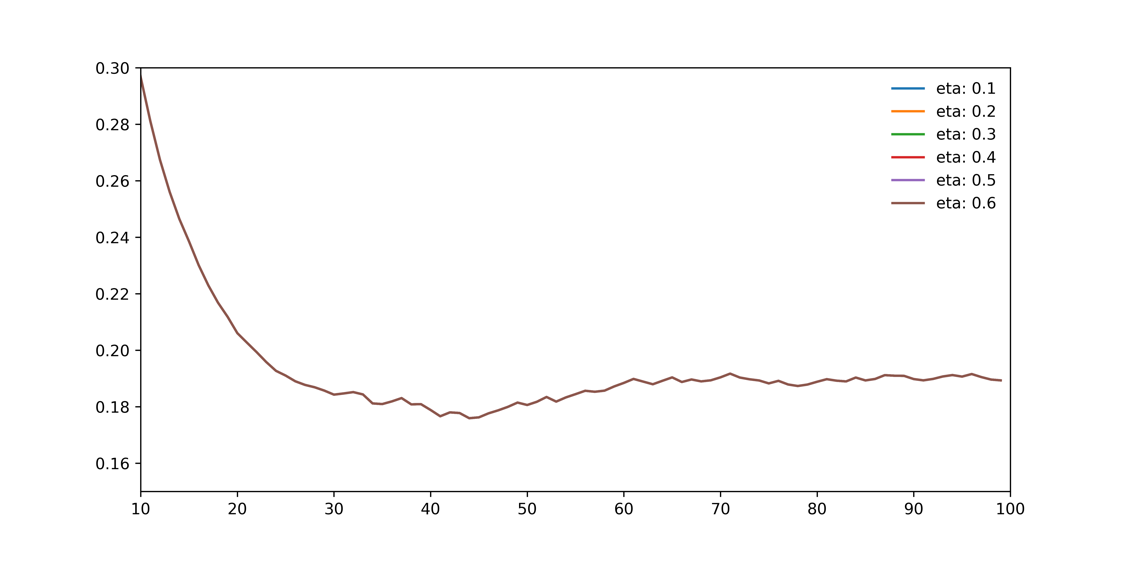


图3- 9损失函数变化图-改变学习率

在1.5至1.9的范围内改变最小子树权重的值，使用准确度和F1分数对模型进行评估。分数变化图如下图所示：

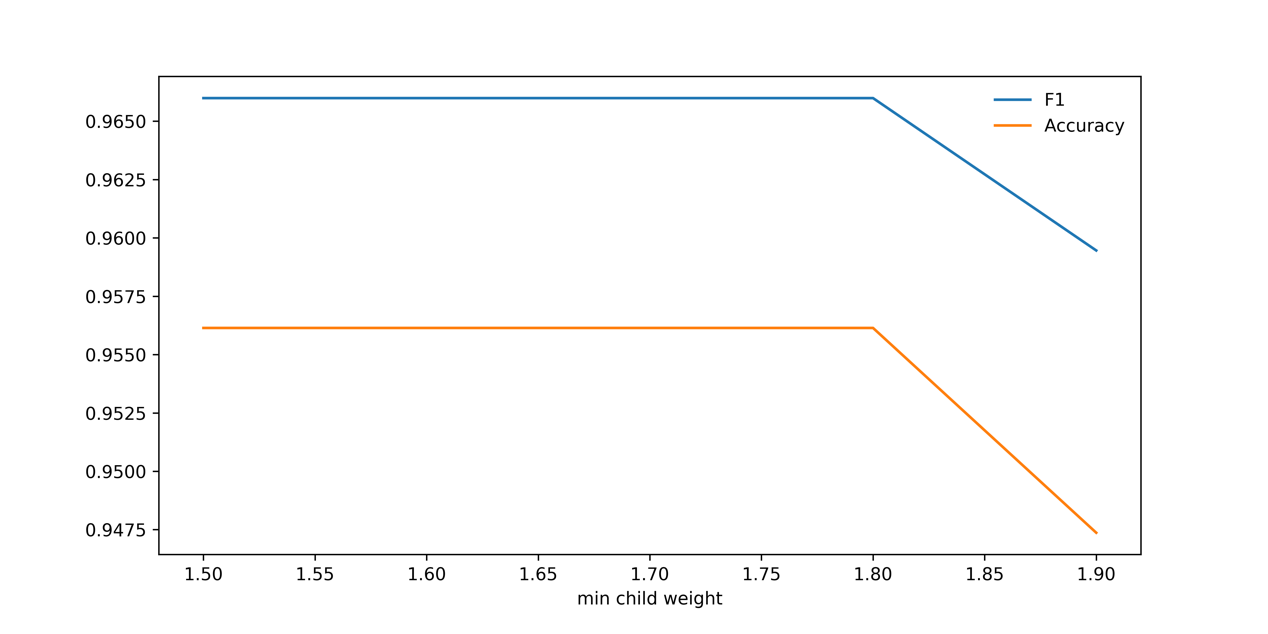


图3- 10模型分数评估-最小子树权重

改变树最大深度、学习率两个参数，模型分数均没有变化。

最后，团队选择的模型参数值如下表所示：

表3- 17最终选择的参数值

|  |  |
| --- | --- |
| 参数 | 参数值 |
|  | 1.9 |
|  | 6 |
|  | 0.5 |

## 3.5模型结果

由于僵尸企业分类与僵尸企业风险预警属于线性模型，没有数据拟合过程，企业提供的数据集也没有这两个功能对应的训练集于测试集，故本节仅针对僵尸企业识别功能进行模型结果分析。

下面将使用准确度、查全率、召回率、F1\_score等评估方法对僵尸企业识别模型进行评估。

表3- 18原始验证集上测试的模型分数结果展示表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类器 | 准确率 | 精确率 | 召回率 | F1值 |
| XGBClassifier | 100% | 100% | 100% | 100% |

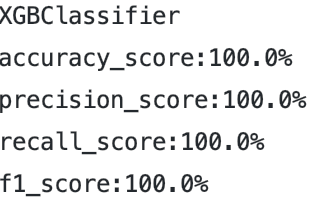


图3- 11原始验证集上测试的模型分数结果截图

可见，基于Stacking策略的XGBoost算法在验证集上效果非常好，准确率、查全率、召回率、F1\_score均为100%。但是这也表明模型很有可能发生了过拟合。为了判断算法模型是否过拟合，团队借助数据分析，结合僵尸企业的数据特点以及2010−2016年我国非金融上市公司的财务、债券开源数据集：Wind，创造了与原有数据集具有相同特征的新数据集，并使用相同的评估方法测试模型的泛化能力与鲁棒性。测试结果如下：

表3- 19新数据集上测试的模型分数结果展示表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类器 | 准确率 | 精确率 | 召回率 | F1值 |
| XGBClassifier | 99.99% | 100.00% | 99.97% | 99.98% |

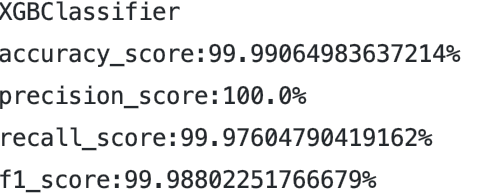


图3- 12新数据集上测试的模型分数结果截图

在新数据集上模型得分依然高达99.98%，这也与十折交叉验证的结果相匹配。这足以说明该算法模型没有过拟合，模型准确率与泛化能力都达到了题目的要求。

模型训练速度与预测速度也很快。使用10万余条数据、3万多个样本的原始数据集进行模型训练，全部步骤（包括多表融合、数据处理、特征提取与新特征建立、训练数据拟合、验证集测试等）仅需34.19秒。模型拟合速度仅需2.78秒。预测4个样本、12条数据的小验证集仅需0.6秒；预测10万个样本的大验证集仅需42.49秒。这些数据表明，该算法模型算法复杂度并不高，预测速度快，用户使用体验较好。

# 4.系统设计与实现

## 4.1系统总体设计

僵尸企业分类识别系统由4个子模块构成：分别是数据导入模块、分类器模型模块、分类器更新模块、数据可视化模块。

1. 数据导入模块：提供模板下载，实现文件的上传和预处理的功能。
2. 分类器模型模块：需要先使用部分数据进行模型初始化，然后利用模型对需要查询分类的企业进行识别，从而识别出僵尸企业并对这些僵尸企业进行分类。除此之外，对非僵尸企业根据输入的企业数据对其进行僵尸化风险等级的预测。
3. 分类器更新模块：利用用户上传的数据对模型进行再训练，并与现存模型进行比较，实现模型的不断更新。
4. 数据可视化模块：对系统预测出的僵尸企业数据进行一个可视化的展示，同时根据大数据统计对僵尸企业地域分布，企业类型分布以及僵尸企业特征等进行可视化展示。

系统的总体架构图如图4-1所示：

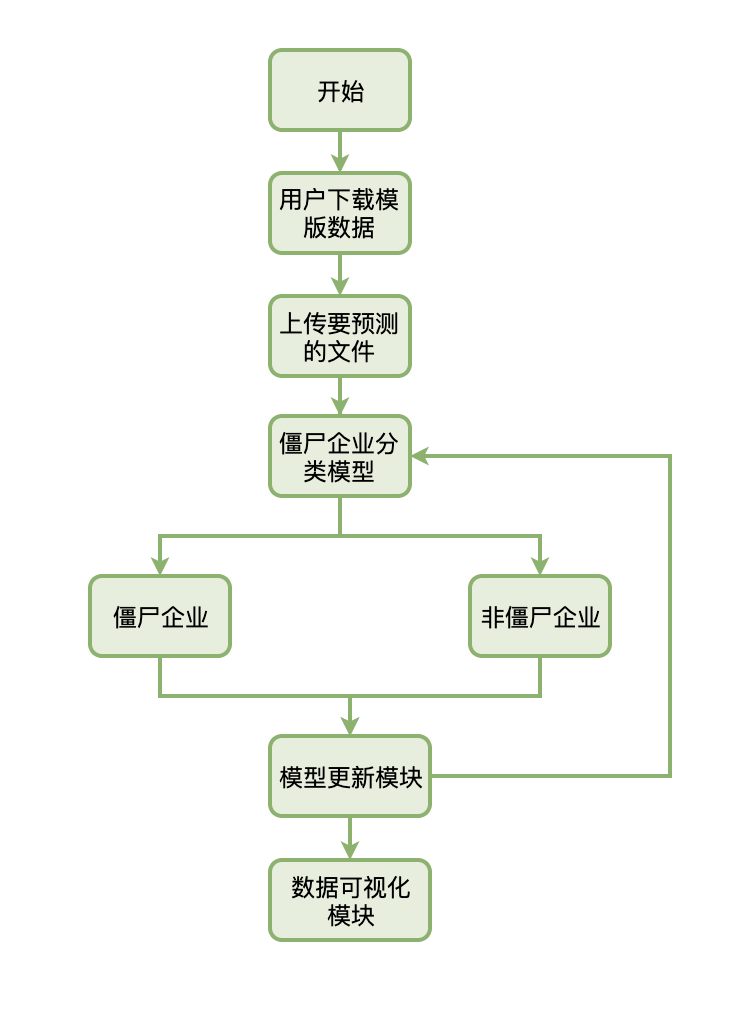


图4- 1系统总体架构图

## 4.2数据库设计



图4- 2数据库表结构图

## 4.3数据导入模块

在数据导入模块，系统提供了需要上传的文件模板下载功能。上传至系统进行预测的四个文件必须严格模板填写要求进行填写。系统首页即可看到模板下载按钮，首页界面如下图所示：



图4- 3首页界面截图

在点击开始预测按钮之后，系统会弹出上传文件的对话框，用户需要上传一个或者多个企业的数据，分别按照要求放在四个文件中上传至系统，界面显示如图所示:

屏幕的截图

描述已自动生成

图4- 4文件上传界面截图

这个页面中的选择文件按钮点击后会弹出文件选择框，用户需要一次性将填写好且符合模板要求的文件全部导入，点击开始上传后，即完成了企业数据的上传。

## 4.4 分类器模型模块设计

### 4.4.1僵尸企业的分类识别

根据用户导入的企业数据，系统模型会对其进行预测。如果导入的是多个企业数据，预测结果界面最多展示8条企业数据。用户可以自行点击详细预测信息按钮或下载预测结果按钮，查看其他结果。

电脑屏幕的截图

描述已自动生成

图4- 5预测结果界面截图

页面在展示企业的预测结果的同时，给出模型的正向和负向概率（即僵尸企业和不为僵尸企业的概率），用户可以根据这个概率进一步了解该企业被判定为僵尸企业的情况。

点击下载预测结果按钮之后，会下载.csv预测结果文件，文件内容包含所有判定为僵尸企业的企业id以及所属的僵尸企业类型，供用户参考。

点击详细预测信息按钮之后，可以看到僵尸企业的类别以及部分重要的僵尸企业判别数据。

电脑萤幕的截图

描述已自动生成

图4- 6僵尸企业分类界面截图

### 4.4.2非僵尸企业的等级预测

在预测结果界面，点击详细预测信息按钮，即可看到所有输入企业的预测详细情况。点击非僵尸企业分页可以查看其中非僵尸企业的部分重要信息展示、非僵尸企业的吸血类型以及非僵尸企业僵尸化的等级预测结果。具体界面展示如下图：

电脑萤幕画面

描述已自动生成

图4- 7非僵尸企业僵尸化等级预测结果界面截图

对于目前还不是僵尸企业的企业，系统会根据所给数据对企业现状进行分析，对其僵尸化风险进行预测，给出一个风险等级预测结果，供用户参考，以保证调整企业的生产避免其进一步僵尸化。

## 4.5 分类器更新模块设计

该系统的数据库会根据用户提供的预测数据进行不断完善。根据后期补充的数据对模型进行更新完善，通过补充数据对模型进行实时优化，从而保证模型的准确率和适用性得到进一步提升。分类器更新模块实现逻辑如图所示：

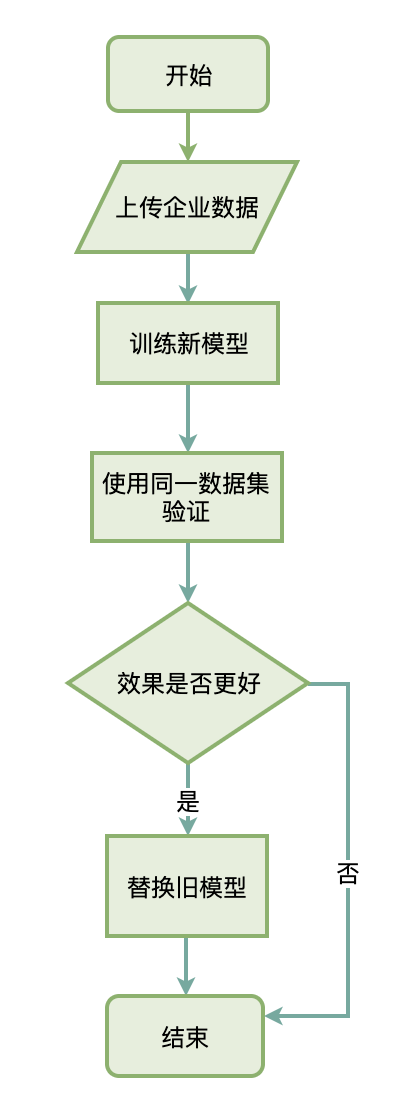


图4- 8分类器更新模块逻辑图

## 4.6 数据可视化模块设计

在数据可视化模块中，团队主要采用开源工具Echarts 对数据进行可视化展示。

系统会对在该系统中出现的所有僵尸企业数据进行展示，便于用户对这些僵尸企业的详细情况能够做进一步的了解和分析。

在企业列表页面展示了数据库中所有企业的概要信息，具体如图所示：

黑色的屏幕

描述已自动生成

图4- 9僵尸企业列表界面截图

点击企业列表中的企业，可以进入企业详细信息页面，同时在页面右上方提供搜索功能，输入企业ID或者企业名称可以快速查找特定企业。企业详细信息界面可以展示单个企业的具体信息，包括企业的基本信息，贷款信息，年报等。具体如下图所示：

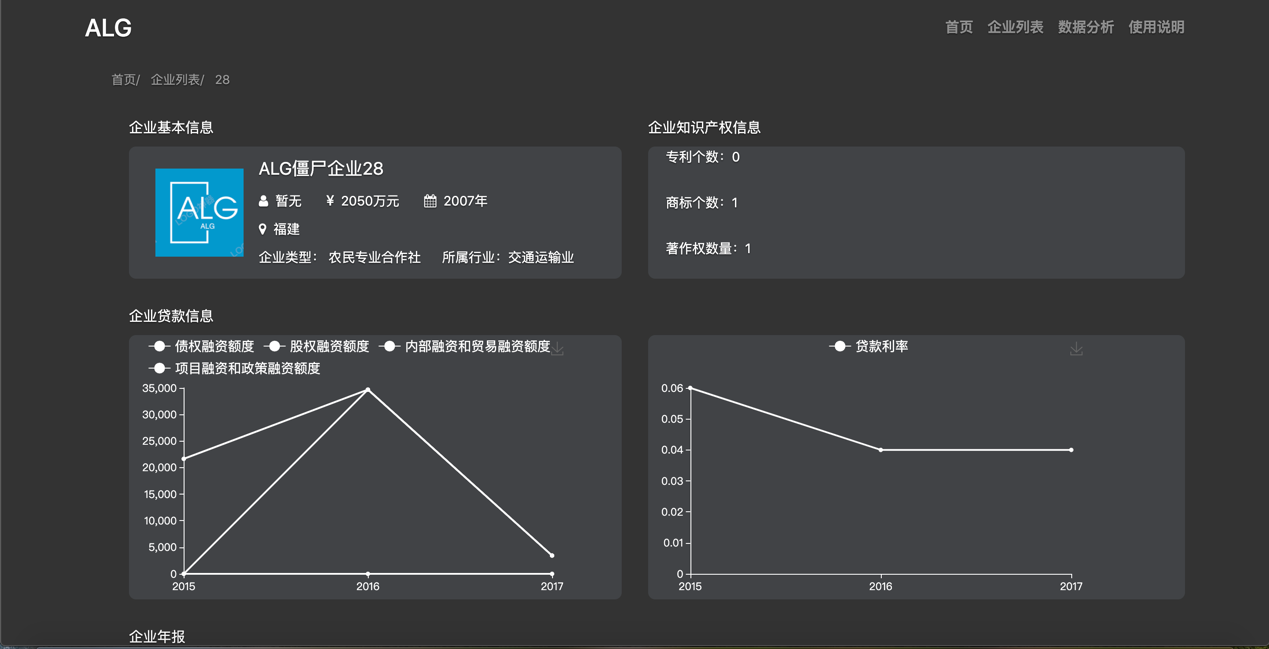


图4- 10僵尸企业详情界面截图

除此之外，系统利用大数据对僵尸企业的地域分布情况、僵尸企业行业分布、注册时间规律、企业类型以及僵尸企业的成因进行了简单分析和统计结果展示，便于用户清晰地了解僵尸企业各类特征和分布情况。界面详细展示情况如图所示：

电脑游戏的屏幕

描述已自动生成

图4- 11僵尸企业大数据分析界面截图

# 5.可行性分析及风险管理

## 5.1 SWOT分析

1. **优势（Strength）**
2. 团队经验丰富。团队成员都有参赛经验，且在比赛中获得不错成绩，各团队成员间配合默契，成员之间的沟通频繁且高效，成员专业能力较强，无论是在项目进度的把控上，还是对模型的实际开发都展示了团队卓越的能力。组内分工明确，对于项目出现问题，能及时给出好的解决方案。因此，团队具备充足的信心和实力。
3. 开发工具简单。使用大家熟知的Pycharm和Mysql，不管是Pycharm还是Mysql，发展至今技术已经非常成熟，因此不管是高效性、可靠性还是适用性都能得到很好的保证
4. 适应市场需求，具有一定市场竞争力。目前市场上绝大部分对于僵尸企业的识别标准都不准确，例如：CHK标准、FN-CHK标准等，都存在其自身的缺点，同时也无法完全适应我国的实际情况。因此，本项目提出的模型结合了现市场上同类产品优点，更好地适应了市场需求，同时增强了模型的泛化性，从而保证了模型的市场竞争力。
5. 前端界面简洁美观，同时利用大数据背景增加了僵尸企业分布以及僵尸企业特征分析功能。
6. Web端操作简单，附赠使用说明，上手简单易懂，降低了用户使用时的学习成本。
7. 增加了对企业的僵尸化风险评估功能。
8. **劣势（Weakness）**
9. 评估标准泛化性不高，对于僵尸企业的判定存在地域差别，政策差异，因此本项目的判定标准还能进一步完善。
10. 产品在初期阶段，数据不够完善，因此可能会导致系统的准确度受到影响。
11. **机会（Opporyunity）**
12. 当前市场上的各种僵尸企业预测系统仍存在各种各样的漏洞，本产品综合考虑了各种影响因素，制定了一套更加完善了僵尸企业判别系统。
13. 目前市场生这类产品的应用还不普遍，需求仍然紧缺，因此本产品应当抓住时机，抢占市场，为本产品赢得更加乐观的市场前景。
14. 随着政府对僵尸企业的重视力度不断加大，僵尸企业识别系统正好适应了当前政府的号召，利用大数据背景解决现实问题，具有很大的社会意义和实用价值。
15. **威胁（Threat）**
16. 一旦市场上出现较多同类产品的竞争，而产品面向的用户有限，因此可能会出现用户的流失，导致产品失去市场。
17. 随着僵尸企业整顿效果的好转，越来越多的僵尸企业退出市场，对于本产品的市场需求也会随之降低。

## 5.2可行性分析

### 5.2.1技术可行性

近几年随着我国对僵尸企业整治力度的加大，对僵尸企业的研究也越来越多，市场上也出现了各种僵尸企业的评判标准，除了官方和普遍认同的CHK标准、FN-CHK标准之外，还有很多地区政府自主制定的判定标准。因此，本产品结合了现市场上的僵尸企业判定标准，制定了更加完善、适用范围更广的识别标准。

此外，在数据库方面本项目采用Mysql数据库，此数据库管理系统软件是免费的，在法律意义上任何人都有权使用。因此，在数据库的使用上不存在太大的问题。对于前端团队具有专业的UI设计师，前端界面简洁大方且美观，具有专属本产品独有的特色，前端框架选择上使用了Layui框架，同时加上了Echarts软件的可视化界面。后端预测模型参考了技术成熟的XGBoost算法，实现了僵尸企业的识别和分类功能，同时保证了识别的准确性和高效性，并通过对大数据的分析实现企业目前的风险等级预测。

本项目组采用的Mysql数据库管理软件、Layui前端框架、Echarts可视化软件等技术,并参考XGBoost模型，这些技术框架都已经较为成熟。因此，对于本系统的功能实现和项目开发在技术上具有可行性。

### 5.2.2经济可行性

1. **成本低**

本产品使用的工具Pycharm和Mysql，都是已经非常完善且成熟的，且操作简单方便，直接使用即可。同时本产品的配置文件都已现有，我们在系统方面只需进行平台搭建，数据库整理，数据可视化。

1. **可用性高**

团队所做的是对僵尸企业的分类识别，对数据全方位分析以及可视化展示。再加上面对市场如此庞大的产品需求，而目前市场上同类产品还存在很多空缺，所以本产品有很好的前景。同时，本产品实现了僵尸企业识别的准确性和高效性，并对非僵尸企业进行了风险等级预测，因此本产品具有很高的实用性。

1. **竞争性强**

从市场现况来看，国内僵尸企业占10%，欧洲28%，美国33%。由此可以看出国内外对这类产品都有庞大的市场需求。目前国内市场上的绝大部分对于僵尸企业的识别标准或多或少都存在缺陷，例如：CHK标准、FN-CHK标准等，无法完全适应我国的实际情况。然而，本产品的判别标准结合了其他标准的共同点，减少了标准的受限性。因此，本产品具有绝对的市场竞争力。从技术层面来看，工具上我们利用技术成熟的Pycharm和Mysql，上手快；web界面设计我们采用了Layui和bootstarp结合，从而确保前端界面的简洁美观；后端模型参考XGBoost算法，确保了模型的准确性。因此，本产品在市场竞争上具有一定的优势。

### 5.2.3法律可行性

有关法律版权问题，奥力给团队一直奉行坚持原创、拒绝盗版和抄袭的原则。该系统是原创自主设计，不存在任何抄袭侵犯专利权、版权等问题，不涉及时代及国家敏感信息，因此在法律上是可行的。

### 5.2.4市场可行性

僵尸企业是影响经济可持续发展、破坏金融系统稳定性和阻碍我国经济转型升级的顽疾,但目前相关研究仍处于起步阶段。研究发现，2011年和2012年我国僵尸企业比例约为3.3%，2013年后这一比例升至5%，2016年有所回落；2008年"4万亿"经济刺激计划造成固定资产投资过快增长,我国东、中、西部地区的僵尸企业比例依次递增,黑色金属冶炼、造纸等传统产业的僵尸企业比例最高,国有企业僵尸比例是非国有企业的4倍。因此，市场急需一种对于僵尸企业的识别分类系统，以满足市场监督管理局对僵尸企业治理的需求。由此可见，本系统存在巨大的市场。

除此之外，对于僵尸企业的识别模型不仅可以应用于中国，国外很多国家也具有这样的市场需求，但对于僵尸企业的识别和处置是政府长期以来的目标，在产品成熟的条件下，可以向海外扩大产品市场。因此，本系统在市场推广上具有可行性。

## 5.3风险管理

### 5.3.1风险管理体制

本系统在研发和维护的过程中，团队采用以下风险管理及内控体系，并根据项目实际情况对其进行了完善。

1. 项目组在每年的年度计划中，必须包含风险管理体系，每个小项目的开发计划也必须包括风险管理计划。
2. 在制定项目目标时，必须同时进行风险识别、风险分析及风险控制责任、风险控制方案，以确定最终的风险应对行动策略和计划，达到将不确定变为确定的目标，减少项目的不确定性。
3. 项目组建立风险跟踪制度，定期对项目进行风险评估并进行项目风险管理计划调整，制定更加适合项目的风险管理计划。
4. 各分支管理人员应当在每次例会上报告近期故障以及预测下一步风险分析。
5. 项目风险应该进行定期小结以及各分支成员之间应该进行经验分享。
6. 项目经理不得以风险识别的结果评价开发团队成员个人的表现。

### 5.3.2风险及防范策略

1. 技术风险

面对一个市场竞争激烈的新兴产品，加上涉及到的技术广泛，在如今技术更新换代迅速的时代，系统涉及的技术性问题可能会渐渐被突破，从而使得产品失去竞争力。

防范策略：

将系统尽快投入市场，在试用的过程中对产品进行完善，保证系统的高准确率和高适用性。同时，团队成员需要及时学习新的技术，增强团队的专业性，并实现对产品的及时更新，以迎合新的市场需求。

1. 市场风险

一旦市场上出现较多同类产品的竞争，而产品面向的用户有限，从而导致产品的用户流失。另一方面，随着僵尸企业整顿效果的好转，越来越多的僵尸企业退出市场，对于本产品的市场需求也会随之降低。

防范策略：

及时更新系统，在功能上除了实现对僵尸企业的识别分类之外，增加对非僵尸企业风险等级预测的功能。通过增加系统功能，避免被市场淘汰的风险，同时增大产品适用范围。

### 5.3.3风险管理的意义

项目风险管理在一个项目中占据及其重要的地位，其具有的主要意义如下：

1. 增加项目成功的机率，使项目达到预期的结果。
2. 从项目进度、质量和成本目标来看，项目管理与风险管理的目标是一致的。通过风险管理以减少风险对项目进度、质量、成本的影响，最终实现项目目标。
3. 从计划职能看，项目计划考虑的是未来，从未来存在不确定因素，风险管理的职能之一是减少项目整个过程中的不确定性，有利于计划的准确性。
4. 从项目实施过程看，不少风险是在项目实施过程中由潜在变成现实的，风险管理就是在风险分析基础上拟定具体措施来消除、缓和及转移风险，并避免产生新的风险，因此有利于项目的实施。

# 6.团队构成及项目管理

## 6.1团队构成

本项目团队组成成员共5人，分别是项目经理、算法工程师、前端工程师、系统工程师。其组织结构主要如下图：

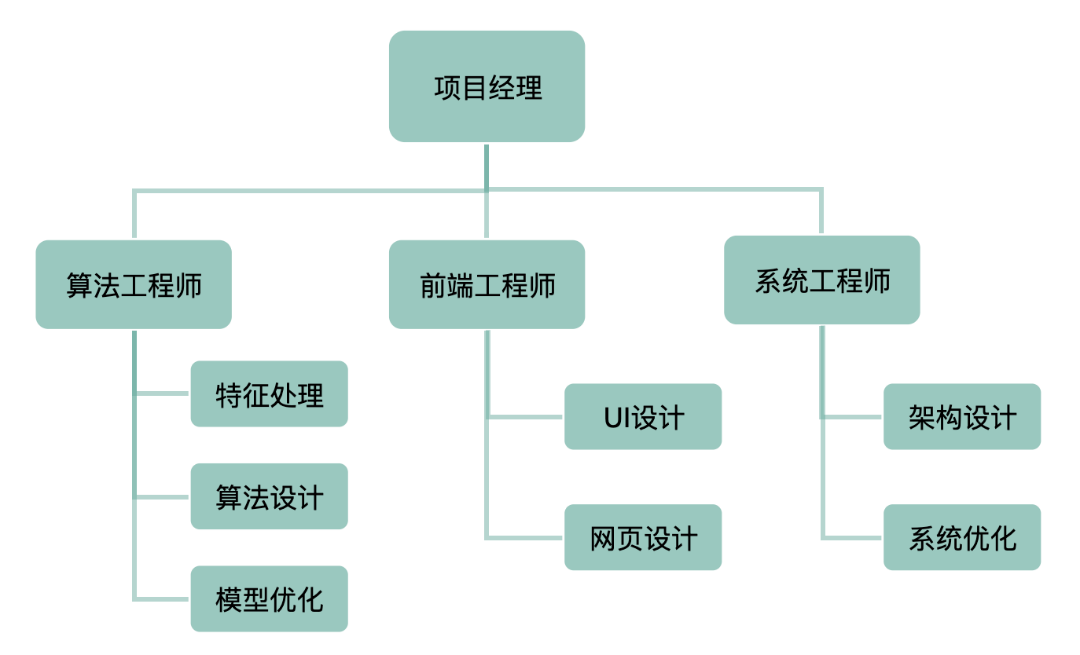


图6- 1团队成员结构图

## 6.2团队分工

结合项目的最终目标及任务安排，考虑团队成员的特长以及整个项目的功能需求等，对项目的分工进行了细化安排，具体如下表：

表6- 1团队成员详细分工表

|  |  |  |
| --- | --- | --- |
| **姓名** | **角色** | **职责** |
| QQ | 项目经理 | ①项目策划设计  ②客户需求分析  ③项目整体规划  ④项目范围管理  ⑤项目进度安排  ⑥项目人员调整  ⑦项目沟通管理  ⑧项目质量管理  ⑨文档撰写 |
| CJL | 系统工程师 | ①项目整体设计  ②资料收集和分析  ③技术方案设定  ④项目测试管理  ⑤技术问题处理  ⑥控制整体开发进度  ⑦算法调参及分析  ⑧文档撰写 |
| GZR、LYZ | 算法工程师 | ①算法设计  ②算法实现  ③模型建立  ④算法调参及分析  ⑥与前端交互  ⑥文档撰写 |
| GCQ | 前端设计师 | ①前端界面设计  ②企业数据展示  ③数据可视化展示  ④与后端联系  ⑤前端界面维护  ⑥文档撰写 |

## 6.3项目计划安排

### 6.3.1任务分解及进度安排

**1.任务分解**

本项目主要有三个需要完成的任务：企业画像模型的构建、僵尸企业分类识别算法的设计、数据的可视化展示。因此在结合整个项目实际情况及目标期望将项目任务分为以下几个部分：

1）项目整体管理；

2）分类识别算法的实现；

3）前端可视化界面实现；

4）项目成果测试和验收；

**2.进度安排**

按照比赛时间安排，我们对整个项目的项目开发周期进行了简单的计划和安排。首先将整个项目周期分为五个阶段：需求分析阶段、整体设计阶段、系统开发阶段、系统测试阶段和维护阶段。

为了按时且保质保量地完成计划任务，我们对各阶段任务进行了进一步细分：赛题分析、需求调研、资料查找和搜集、技术准备、前端需求确定、后端算法开发、前端界面设计、前端网页开发、前后端交互、系统测试、系统完善。

任务安排的具体甘特图如图所示：

图6- 2任务安排甘特图

### 6.3.2阶段性任务讨论

在春节假期期间，我们每周会进行线上的会议，针对近期完成的任务以及每个人负责模块的进度进行探讨。同时保证至少每个月与指导老师进行一次交流和汇报，同时完成该段时间的任务成果验收，然后进行下一阶段的任务计划安排。

### 6.3.3项目计划跟踪

项目经理定期对各阶段的项目计划的实现情况进行跟踪检查验收，其中包括各阶段的进度、工作量、工作成果及近期正在完成的任务等，从而了解项目的实际进展情况以及项目成员现阶段的具体情况，根据实际的情况对整体的实施计划进行调整。

### 6.3.4偏差控制

项目经理将跟踪验收得到的具体情况和实际数据与原先制定的需求中的数据进行对比分析，估算偏差。如果项目进展显著偏离需求，应保证及时发现项目问题并采取措施进行调整。同时注意文档和PPT跟踪的正确性，将偏差控制在可接受的范围内，避免文档、PPT以及最后的成果三者之间存在较大的出入，保证项目始终在规定的时限和计划轨道上。

## 6.4项目管理

### 6.4.1项目进度管理

项目经理定期召开阶段性周期会议，讨论项目的进展情况。其中包括每周的固定会议、指导老师的定期验收以及根据进度安排的甘特图参考比较当前项目进度。根据每阶段项目成员的完成情况和遇到的问题，对下阶段的任务安排进行及时的调整，保证项目进度能够尽可能的按照任务安排甘特图的时间定期高质量完成。

### 6.4.2项目质量管理

项目质量管理由项目经理和系统工程师在指导老师的指导下完成。其中主要包括以下几个方面：

* 根据对项目的计划跟踪，观察项目是否符合项目的需求，并且在任务安排甘特图的进度安排下完成；
* 项目的偏差控制，如果项目完成情况与实际需求出现偏差，应该在第一时间发现问题并且提出解决的方案，在保证不影响项目完成质量的情况下，调整任务安排，保证项目能够按期按质完成；

### 6.4.3项目沟通管理

根据实际情况，组织线上或者线下会议，对项目进度包括现阶段的任务完成情况以及下一阶段的计划安排进行讨论和交流。针对每次会议提出的问题需要及时提出解决的方案，同时各模块的衔接和配合也是会议沟通的目的。除此之外，为了保证会议沟通的高效和有序，按照会议安排相关人员需要提前准备汇报的提纲，保证会议的质量，同时减少不必要的无效沟通。每次会议结束需要对会议内容进行记录，同时项目经理对下次会议的时间和大致内容进行安排和计划制定。

# 7.参考文献

[1]田娟,朱定局,杨文翰.基于大数据平台的企业画像研究综述[J].计算机科学,2018,45(S2):58-62.

[2] 廖信林,贺雪莹,李敬.安徽省制造业僵尸企业的测度分析——基于修正后的FN-CHK模型[J].池州学院学报,2018,32(02):68-72.

[3] 中国僵尸企业研究报告——现状、原因和对策[EB/OL]. http://nads.ruc.edu.cn/upfile/file/20160727155621\_848924\_58213.pdf

[4]Zombie Firms and Performance of R&D Support Programs for Small and Medium Enterprises. 2018, 21(4):1474-1492.

[5] 刘阳. 一种企业画像系统的设计与实现[D].河北师范大学,2019.

[6]赵静,赵荔.“僵尸企业”预警框架构建[J].财会月刊,2017(33):29-34.

[7] 毛涓.国有“僵尸企业”的分类识别与风险防范[J].经营管理者,2019(12):78-79.

[8]卓越,王玉喜.僵尸企业识别标准的中国适用性检验与修正——基于中国工业企业数据的分析[J].经济评论,2019(03):123-137.

[9]孙博文,柳明,张伟广.僵尸企业识别研究综述:修正与异质特征——基于企业产品创新的视角[J].宏观质量研究,2019,7(03):79-98.

[10]吴立欧.僵尸企业治理的温州经验[J].法制博览,2019(34):31-33+36.

[11]鲍秀坤. 论“僵尸企业”如何认定[C]. 河南省法学会、山西省法学会、湖北省法学会、安徽省法学会、江西省法学会、湖南省法学会.第十二届“中部崛起法治论坛”论文汇编集.河南省法学会、山西省法学会、湖北省法学会、安徽省法学会、江西省法学会、湖南省法学会:河南省法学会,2019:293-298.

[12]彭良军.僵尸企业认定的经济标准与法律标准探究——基于会计准则[J].财会通讯,2018(22):27-31.

[13]Xinfeng Jiang,Sihai Li,Xianzhong Song. The mystery of zombie enterprises – “stiff but deathless”[J]. China Journal of Accounting Research,2017,10(4).

[14]张栋,谢志华,王靖雯.中国僵尸企业及其认定——基于钢铁业上市公司的探索性研究[J].中国工业经济,2016(11):90-107.

[15]邵慰,孙慧梦,俞晓波.我国僵尸企业的识别、形成机制及治理研究[J].贵州商学院学报,2019,32(03):57-63.

[16]胡雪璐.供给侧改革下我国“僵尸企业”的认定与应对办法[J].财会学习,2018(19):191+193.

[17]黄少卿,陈彦.中国僵尸企业的分布特征与分类处置[J].中国工业经济,2017(03):24-43.

[18]李强. 僵尸企业识别、成因及政策选择[D].东北财经大学,2017.

[19]王松子. 论“僵尸企业”的形成及其法律规制[D].西南政法大学,2017.

[20] 周琎, 冼国明, 明秀南. 僵尸企业的识别与预警—来自中国上市公司的证据[J]. 财经研究, 2018, 44(4): 130-142.

[21]李旭超. 市场扭曲、资源错配与中国全要素生产率[D].浙江大学,2017.

# 8.附录

### 会议纪要（一）

会议记录人：QQ 时间：2019年12月28日

**会议摘要：项目选择、赛题分析、确定团队及接下来的计划；**

具体内容：

1. 项目选择

根据“第十届中国大学生服务外包创新创业大赛企业命题赛题手册”中的14个赛题的具体要求结合自身能力，最终经过讨论决定选题A09。

1. 赛题分析

根据赛题背景，结合业务背景问题说明，对赛题的情况进行了简单的讨论。随后，对赛题的用户期望和任务要求进行了仔细研究和总结，确定赛题的任务范围以及最终目标。

1. 团队确定

根据赛题要求和A09题目具体的能力要求，我们最终确定了项目组成员，分别为：项目经理1名、系统工程师1名、算法工程师2名、前端工程师1名。

1. 下阶段安排

确定完赛题和最终的团队，接下来一个阶段就是对僵尸企业目前的研究状况和市场需 求进行调研。项目组成员分为两组：一组搜集现阶段僵尸企业的国内外研究现状，另一组通过多渠道地对僵尸企业的市场进行调研方便后续项目开发。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（二）

会议记录人：CJL 时间：2020年1月13日

**会议摘要：总结项目市场调研情况、国内外研究现状、制定下一阶段计划；**

具体内容：

1. 项目市场调研情况

市场上的企业越来越多，基于企业运营时产生的海量数据，可以利用大数据技术结合企业画像的理论来对企业进行全面分析，为企业成长、行业发展、政府监管等各方面提供可靠的数据分析。根据企业全息画像，面向不同的应用场景对企业分类识别，特别是识别现市场中的僵尸企业越来越受到市场监管部门和政府的重视。市场上对僵尸企业的定义和评判标准尚不统一，各类标准分类识别僵尸企业均有其局限性，尤其是标准的制定有其主观性，因此基于企业画像分类识别方法成为识别僵尸企业方法的首选。而从海量 的企业中准确地识别出僵尸企业，这对于市场监管部门和政府越来越重要。

1. 国内外研究现状

1）**国内现状：**

华南师范大学计算机，提出了一种基于大数据技术的企业画像的构建模型以及构建方 法。将企业画像分为五大维度：企业属性维度、企业信用属性维度、企业交易特征维 度、企业内外关联特征维度、企业评价信息维度。

在数据处理方法上，文献提出三大部分内容：

①标签提取：k-means、LDA、NB、CNN等特征提取方法优缺点的比较；

②关联分析：Apriori算法和GRI关联规则算法的对比分析表；

③分类预测：给出了KNN算法、SVM算法、决策树算法的优缺点对比。

2）**国外现状：**

①Turney将机器学习模型应用到关键词中，把关键词提取看作是一个分类问题，使用 C4.5决策树算法和遗传算法进行关键字提取；

②Mihalcea基于无监督学习的图模型理论，利用共现窗格中词与词的共现关系构造图 模型，并依托谷歌公司的网页排序pageRank算法，提出了TextRank关键词提取算 法；

③短文本分类方面，Sen等人抽取数据中的特征项，作为基本单位，构建了一种网络 模型分类，分类性能得到了一定的提高；

1. 下一阶段计划和任务安排

接下来一段时间，我们就要正式进入系统的开发阶段，首先需要完成的就是前端需求 的完善以及界面设计；其次是后端算法的设计和构想。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（三）

会议记录人：GZR 时间：2020年2月15日

**会议摘要：前端界面设计想法讨论、算法设计讨论、资料查找及文本撰写等工作安排；**

具体内容：

1. 前端界面设计想法讨论

根据本次系统开发的要求，前端需要实现单个和批量查询分类，这是本次前端需要完 成的最重要的核心任务。除此之外，为了系统的完善以及界面的美观，我们对前端界面的设计进行了讨论。最终决定，web端不涉及用户登录，属于一个开放网页的形式供用户分类查询，同时在显示页面除了查询结果，也会对查询企业的一些数据进行可视化展示和分析。

1. 算法设计讨论

根据赛题背景，结合业务背景问题说明，对赛题的情况进行了简单的讨论。随后，对赛题的用户期望和任务要求进行了仔细研究和总结，确定赛题的任务范围以及最终目标。

1. 资料查找和文本撰写等工作安排

根据安排，这段时间QQ、CJL两人负责继续查找相关的资料，为分类算法的评判标准 提供可信材料并且撰写部分文本内容。LYZ同学协助完成资料收集和查找工作。

在完成本阶段任务的同时做好记录，以备下次会议对自己所负责的任务进行进度汇报，并对过程中所遇到的问题进行讨论。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（四）

会议记录人：LYZ 时间：2020年2月22日

**会议摘要：前端需求确定、后端算法设计、前端界面设计（即上一次会议的任务的进度汇报）、过程中所遇问题的讨论；**

具体内容：

1. 上次任务的进度汇报

①前端需求：前端需求基本确定，按照上次讨论结果设计前端界面基本没有问题。

②前端界面：界面设计已经完成了首页，后续结果展示界面现在有3个版本，包括颜 色的选择和展示风格的选择需要大家进一步确定。

③后端算法设计：后端对企业提供的企业属性进行了分析，对各字段进行了充分了 解，算法具体设计暂时还没有开始。

④文本撰写：文本的第一章‘项目简介’具体内容已经全部完成，第二章‘最终目标 与问题分析’的部分内容已经完成。

⑤资料收集情况：通过资料的查找，发现大部分省份都公布了自己省的僵尸企业判断 标准，僵尸企业的判断标准涉及到很多方面，所以与当地的具体政策有关，所以各省 份的判断标准不同是合乎常理的，但是有的省份的僵尸企业判断标准并没有公布，因 此可能对本项目僵尸企业判别标准的最终确定带来一些困难。

1. 过程中遇到的问题讨论

①经过对企业提供的数据进行一些简单处理之后，发现很多的数据貌似与真实情况并 不符合。因此在数据处理上遇到一些困难，包括对NA值得处理，是删除还是全部置为0等处理。最终经过讨论，决定对NA值用均值进行填充。

②在本系统的僵尸企业判别标准的确定上，存在困难。各个省份都有自己的判别标 准，本系统是否需要分省份进行识别，还是仅用一个综合的判别标准适用于所有的省份。最后，经过对各个省份判别标准的分析，发现其中存在很多的共同点，若是综合一下共同点制定一个统一的标准对所有的企业进行识别也是可行的，从而简化了算法的设计。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（五）

会议记录人：GCQ 时间：2020年2月29日

**会议摘要：前端网页开发、后端算法设计及开发；**

具体内容：

1. 前端网页开发的进度

前端网页最终结果显示页面，经过这段时间的讨论最终确定了最后的界面。目前正在根据设计的界面进行具体的开发。

1. 后端算法开发进度

后端的算法已经对数据的各个字段进行了进一步的了解和讨论，同时各特征的处理也在有序进行，但是最终特征应该如何处理还在进一步讨论中，同时僵尸企业的判断标准也还在做进一步的完善。

1. 整体项目进度和计划

整个系统的前端和后端都已经进入了正式的开发阶段，文本的内容也在进一步地填充和完善，根据目前的进度来看，后端的算法开发需要加快进度。接下来一段时间，后端需要尽快处理好特征开始训练模型，包括是否借鉴已有的开源算法以及对开源算法的选择。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（六）

会议记录人：QQ 时间：2020年3月7日

**会议摘要：前端开发进度、后端算法架构、文本进度；**

具体内容：

1. 前端

前端页面的开发基本上大致完成，但是最后的企业各类数据可视化展示模块还不是特 别美观，需要做进一步的修改。除此之外，由于前端开发人员后期需要与后端交互，所以这方面的知识技能还需要进一步加强。

1. 后端算法

后端的特征处理基本已经完成，根据收集到的资料加上对僵尸企业特征的分析，最终 确定了特征处理的方案。具体特征处理详细情况见前面详细文本。另外经过多次实验 加上对于深度学习的了解和经验，最后项目组决定参考XGBoost算法。对于僵尸企业的评判标准也基本确定。

1. 文本进度

文本内容一直跟随前端和后端开发的进度进行内容的填充，每次会议都会对会议内容 进行记录，包括进度和最终选定方案的记录，方便接下来的文本内容撰写。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（七）

会议记录人：CJL 时间：2020年3月22日

**会议摘要：前端界面初步展示、后端算法、评判标准；**

具体内容：

1. 前端界面初步展示

在上次的基础上，前端界面的配色等进行了修改调整，同时前端的模板下载以及文件上传的细节处理已经完成。前端界面及各个界面跳转基本全部完成。前端与后端的交互在进行中，暂时还没有完成。下一阶段的计划就是接入后端算法，大打通整个系统。

1. 后端算法

在选择的XGBoost算法基础上，根据本项目的具体需要进行了简单调整，目前已经在算法识别的准确率上达到的99%以上，但是算法的广泛适用性仍然存在局限，还需要进一步修改。同时，算法预测的速度还较慢，算法效率还需要进一步优化。

1. 评判标准

评判标准结合各种资料，最终确定为：

* 连续三年亏损
* 资产负债率三年处于85%以上
* 长期（三年及以上）四欠

满足以上条件的两个及两个以上就被认为是僵尸企业。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（八）

会议记录人：GZR 时间：2020年3月29日

**会议摘要：后端算法、前端与后端的交互、文本；**

具体内容：

1. 后端算法

根据项目需求，在XGBoost算法的基础上，加入了统计模型，同时根据实际情况对算法进行了一些调整，在运行效率上也对算法进行了相应的调整。目前，在准确率上依然保持在99%,效率有了一定的提升，但是还需要进一步的优化，同时后期也团队也还在计划对算法进行更进一步的改进，以达到更加具有广泛性、准确性，高效性的算法。

1. 前端与后端的交互

经过进一步的学习和相应技术的了解，本阶段前端界面接入了后端算法，但是在模板下载和上传文件的时候，由于mac系统和windows系统的差异，还存在着一些bug，如：windows下载的模板压缩包没有后缀名无法自动识别打开，同时预测结果文件也存在同样的问题。除此之外，前端接入后端算法之后，预测数据的速度较慢这一缺点显示得更加明显，因此也强调了算法优化的必要性。

1. 文本

文本的第三章、第四章、第五章已经正式开始撰写，随着技术部分的完成，文本涉及的内容也需要进一步完善。根据团队实际情况和目前各部分的进度，对文本撰写任务进行了细分，同时也开始了对文本的第一轮检查和纠正，主要包括语句和错别字的检查。对于文本撰写。

1. 下一阶段任务

在下次会议前，文本上需要完成第三章和第五章的绝大部分内容。前端需要解决现在发现的bug,文件上传和文件下载的后缀格式问题。另外，后台算法继续优化。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（九）

会议记录人：LYZ 时间：2020年4月4日

**会议摘要：前端验收、后端算法进度汇报、文本进度验收；**

具体内容：

1. 前端验收

前端界面完成，同时成功接入后台算法，解决了上次会议发现的bug。目前前端实现了模板下载、文件上传，僵尸企业预测、僵尸企业特征展示、僵尸企业列表展示以及使用说明界面等功能。前端数据可视化方面，基本实现了任务要求完成了全部内容，经过团队讨论决定，针对题目提到的支持单条数据查询和批量查询，团队决定全部采用上传excel文件的形式进行预测，因为涉及到的数据较多，即便是单条数据预测仍然会涉及到很多数据的输入，这不仅给用户带来不便，同时也增加了数据输错的风险，因此，团队决定全部采用excel文件上传的方式进行数据上传，避免过多的手动输入导致出错。

1. 后端算法进度汇报

后端算法在运行效率的优化上遇到瓶颈，目前还一直处于攻克阶段，暂时没有突破性的进展，接下来也将继续想办法做进一步的优化。

1. 文本进度验收

对于项目的详细方案文本，除了第四章对模型的详细介绍还未完成之外，其他部分基本全部跟随项目开展进度完成。除此之外，还需要完成项目概要介绍，接下来的任务是开始撰写项目概要介绍以及开始准备项目ppt的制作以及ppt展示的策划。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（十）

会议记录人：GCQ 时间：2020年4月12日

**会议摘要：项目测试、任务验收、下一阶段任务；**

具体内容：

1. 项目测试

从需求分析、项目可行性分析到最终的项目实际开发，截止到目前，本项目的前端工作基本完成，共涉及到三个主要界面以及文件上传模板下载等基本功能。后端算法也基本完成，从特征处理，数据分析，基于XGBoost算法，结合统计模型完成了后端算法的设计，并根据最终的测试数据对算法做了进一步的优化和改进。至于文本工作，项目详细方案已经完成85%以上，还需要进一步的完善，至于项目概要介绍也还在进一步的完善。

1. 任务验收

根据任务要求和用户期望，最终完成了数据治理、duobiao2融合、数据特征提取、数据预处理、数据建模。最终采用人工智能算法建立了僵尸企业的分类模型，达到了对企业进行合理的分类识别的目的。

1. 下一阶段任务

根据目前的项目进度，接下来，团队的主要任务是完善需要提交的各种材料，包括项目详细文本、项目概要介绍、项目简介PPT以及企业所需要提交的材料，除此之外，还需要录制项目演示视频。然而，接下来的重点是项目概要介绍文本撰写以及项目简介ppt的制作。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ

### 会议纪要（十一）

会议记录人：QQ 时间：2020年5月14日

**会议摘要：项目完善、提交材料整理、录制项目演示视频、后期任务安排；**

具体内容：

1. 项目完善

对后期新增加的功能，非僵尸企业的僵尸化风险等级预测功能进行了进一步完善，对风险等级的划分标准进行了细微的调整。除此之外，对于僵尸企业的分类功能，根据本项目提出的ALG标准对分类功能做了进一步完善，同时对界面展示做了微调。

1. 提交材料整理

根据主办方要求，本项目需要提交的材料包括：项目概要介绍、项目简介PPT、项目详细方案、项目演示视频、企业要求提交材料。根据前期各文档的几个修改版本对整个项目的文档进行简单整理。同时检查最终版的文案材料以及简介PPT的动画和播放是否存在问题。

1. 演示视频的录制

打开系统的前端展示界面，录制项目的演示视频

1. 后期任务安排

接下来需要对视频进行简单的处理，同时检查演示视频搭配上旁白播放整个过程是否清晰。演示视频验收完成后，在比赛官网提交项目所有材料。

参与会议人员：

项目经理：QQ

系统工程师：CJL

算法工程师：GZR、LYZ

前端工程师：GCQ