1. **前言**

在信息化、智能化时代，随着社会征信体系的逐步完善，企业的信用、风险评估和僵尸企业识别也越来越受到政府、银行等机构的重视。长期以来，针对僵尸企业的分类方法存在指标模糊、主观性强等缺点，在评估过程中需要频繁的人工干预，分类效率较差，准度较低。为了解决这些问题，在本项目中使用大数据分析手段对海量企业信息进行比较与整合，通过数据挖掘和机器学习算法，对企业进行全方位、自动化的信用评估和风险测算，提取构造有效特征来完成僵尸企业分类任务。在具体实现上，项目使用半监督方法处理数据标签缺失，对主流的梯度提升树等机器学习方法进行融合调优，来构建准确、高效的分类模型，并设计了功能突出、操作便捷、界面精致的Web端应用供用户使用。通过僵尸企业的评估与识别，不仅能促使企业自检，推动其自我升级改造，还能优化银行等信贷机构的资金投入，对市场秩序和市场机制的维护有着重要的意义和价值。

1. **创意描述**

* 在处理训练数据集中缺失标签的样本时，我们使用半监督的PU-Learning方法来进行填充。该方法通过设置间谍样本来训练一个分类器，并使用这个分类器从无标签样本中寻找可靠负例，之后即可使用正例和这些可靠负例进行后续工作。
* 在对数据特征进行处理时，我们查找了一些在经济学中常用的指标，并结合国家发布的企业分类相关文件，构造了多个能够用来描述企业经营状况的特征。此外，我们通过随机组合的方法构造了一系列交叉特征，用这些特征来捕捉各特征之间存在的非线性关系；通过不同标准对数据进行分类，构造偏离值特征，用这些特征来描述样本与同类别其他样本之间的对比情况。这些生成特征对模型的训练和解释起到了重要的作用。
* 项目在过滤类、包装类和嵌入类三种方法的基础上进行优化，缩短了特征选择的实验时间。通过对不同特征集的分析综合，选取出对模型分类影响最大的特征子集，大大减少了特征数量，提升了模型运行效率。
* 在Web端以色彩鲜明的图表形式对模型分类结果进行展示，能够有效呈现企业的经营状况和画像标签。

1. **功能简介**

* 僵尸企业分类：用户可在Web端上传4份csv文件进行单条或多条企业数据的分类预测，并将结果下载保存到本地。
* 企业画像信息查询：用户可在Web端输入企业ID进行信息查询，前端页面将展示企业的基本信息、画像标签及与同类企业的经营状况比较。

1. **特色综述**

针对传统企业分类存在的主观性强、效率低下等问题，本项目中使用数据挖掘与机器学习方法来设计分类模型，所得模型预测精度高（在验证集上F-score为99.90%，AUC为99.93%）、运行速度快、泛化性能好。在系统的交互逻辑设计上遵循简单高效的原则，使用户通过简单的操作即可完成系统的基本功能。在系统的结果呈现上，我们使用图片与表格取代字符数据，方便了用户对模型预测结果信息的获取和认知。

1. **开发工具与技术**

开发工具：Visual Studio Code，Jupyter Notebook，Safari浏览器。

开发技术：

* 模型部分：Python，Numpy，Pandas，Matplotlib，Scikit-Learn，XGBoost，LightGBM。
* Web前端部分：HTML，CSS，JavaScript，Vue，Element-UI，Axios，Chart。
* Web后端部分：Nginx，WSGI，Flask。

1. **应用对象**

政府与市场监管部门，银行等信贷机构。

1. **应用环境**

阿里云学生服务器（共享计算型），单核CPU，2GB内存，安装Ubuntu16.04操作系统。

1. **结语**

本项目立足当前企业信用体系建设背景，针对僵尸企业分类中存在的问题，提出了一种基于数据挖掘和机器学习技术的自动化分类模型，该模型具有精度高、速度快等优点。为了方便结果展示，项目还开发了一款操作简单、界面友好的Web端应用，用户可以方便地获取到僵尸企业分类结果和企业画像标签等信息。本系统便捷高效、可扩展性强，可被用于企业信用评估和风险测算相关的各个领域，起到投资引导和市场监管等作用，具有广阔的应用前景和价值。