1、项目概述

在数据已经成为战略资源及经济资产的今天，通过数据挖掘和机器学习方法来分析海量数据，鼓励学科交叉跨界合作，探索以大数据为基础，涉及政府治理、产业升级等的计算方法及解决方案已经成为时代的发展的迫切需求。

1.1项目研究的方向及意义

企业营业退出风险预测系统采用了的数据挖掘技术、分类预测技术和机器学习技术，主要基于历史数据潜在的规律进行预测未来企业发展的风险情况。所以，企业或者政府可以一定的参考该系统的预测值进行宏观调控，从而保证企业和整个市场处于健康稳定的发展态势。

1.2 项目研究的背景

传统的企业评价主要基于企业的财务信息，借贷记录信息等来判断企业经营状况，以及是否可能违约等信用信息。对于财务健全、在传统银行借贷领域留有记录的大中型企业，这种评价方式无疑较为客观合理。然而，对于更大量的中小微企业，既无法公开获得企业真实财务信息，也无这些企业的公开信用信息，在强变量缺失的情况下，如何利用弱变量客观公正评价企业经营状况，正是本系统需要解决的主要问题。

系统所用到训练数据是从全国2000多万企业抽取部分企业（脱敏后），该数据包括企业主体在多方面留下的行为足迹信息。我们团队只要通过数据挖掘的技术和机器学习的算法，针对企业未来是否会经营不善构建预测模型，输出风险预测概率值。

1.3项目的基本业务描述

本系统以企业为中心，围绕企业主体在多方面留下的行为足迹信息构建训练数据集，以企业在未来两年内是否因经营不善退出市场作为目标变量进行预测。

主要利用训练数据集中企业信息数据，构建算法模型，并利用该算法模型对验证数据集中企业，给出预测结果以及风险概率值。

系统的预测结果以AUC值作为主要评估标准，在AUC值（保留到小数点后3位数字）相同的情况下，则以F1-score（命中率及覆盖率）进行辅助评估。

2、系统介绍

2.1数据介绍

程序所需要的有两种数据：

(1)企业身份信息（已脱敏）及企业在一定时间范围内的行为数据。该数据对训练集和评测集都是相同的。

(2)目标数据。目标值为该企业在2017年8月时的经营状况：停业1，正常营业0。

2.1.1目标数据

以下数据仅训练集提供。文件名：train.csv。

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识

name	type	字段名称	说明
TARGET	varchar2(2)	目标标签	1停业，0正常。评测数据无该字段

以下为评测数据。文件名：evaluation_public.csv。

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识

2.1.2训练及测试数据

以下数据训练集和评测集格式相同。通过EID关联，自行匹配分离出训练用数据和评测用数据。

1、企业基本信息数据1entbase.csv

所有数据表中，灰色底纹的字段为主键列。

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
RGYEAR	varchar2(4)	成立年度	
HY	varchar2(8)	行业大类	
ZCZB	number	注册资本	人民币：万元。已取整
ETYPE	varchar2(8)	企业类型	
MPNUM	number		已经计算完成的身份指标。空均意味着为0
INUM	number		已经计算完成的身份指标。空均意味着为0
FINZB	number		已经计算完成的身份指标。空均意味着为0
FSTINUM	number		已经计算完成的身份指标。空均意味着为0
TZINUM	number		已经计算完成的身份指标。空均意味着为0

2、变更数据2alter.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
ALTERNO	varchar2(50)	变更事项代码	
ALTDATE	date	变更时间	到月度
ALTBE	varchar2(4000)	变更前	仅ALTNO 为“05”和“27”有数据
ALTAF	varchar2(4000)	变更后	仅ALTNO 为“05”和“27”有数据

3、分支机构数据3branch.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
TYPECODE	varchar2(50)	分支ID	分支唯一标识
IFHOME	varchar2(50)	分支机构是否在同一个省	1为本省0为外省
B_REYEAR	date	分支成立年度	
B_ENDYEAR	date	分支关停年度	空则为该分支机构仍然正常经营

4、投资数据4invest.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
BTEID	varchar2(50)	被投企业ID	和EID共享ID
IFHOME	varchar2(50)	分支机构是否在同一个省	1为本省0为外省
BTBL	number	持股比例	
BTYEAR	varchar2(50)	被投企业成立年度	
BTENDYEAR	varchar2(50)	被投企业停业年度	空则为该被投企业仍然正常经营

5、权利数据5right.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
RIGHTTYPE	varchar2(50)	权利类型	
TYPECODE	varchar2(100)	权利ID	权利唯一标识号
ASKDATE	date	申请日期	到月度
FBDATE	date	权利赋予日期	到月度

6、项目数据6project.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
TYPECODE	varchar2(100)	项目ID	项目唯一标识号

name	type	字段名称	说明
IFHOME	varchar2(50)	项目地是否是企业登记地	1为本省0为外省
DJDATE	date	中标日期	到月度

7、被执行数据7lawsuit.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
TYPECODE	varchar2(100)	案件ID	被执行案件唯一标识号
LAWDATE	date	案发日期	到月度
LAWAMOUNT	number	标的金额（元）	空为无标的金额案件

8、失信数据8breakfaith.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
TYPECODE	varchar2(100)	失信ID	失信唯一标识号
FBDATE	date	失信列入日期	到月度
SXENDDATE	date	失信结束日期	到月度。空为未结束

9、招聘数据9recruit.csv

name	type	字段名称	说明
EID	varchar2(50)	企业ID	企业唯一标识
WZCODE	varchar2(50)	招聘网站代码	
RECRNUM	number	招聘职位数量	近半年内
RECDATE	date	最近一次招聘日期	月度

2.2程序介绍

该程序主要利用训练数据集中企业信息数据，构建算法模型，并利用该算法模型对验证数据集中企业，给出预测结果以及风险概率值。

2.2.1程序的用途及功能

该项目主要目的是根据透明且公开的企业行为数据，结合数据挖掘以及机器学习的方法来对企业经营退出的风险做一个预测，为企业的自身发展以及投资等利益行为做一个风险判断。

2.2.2 程序关键技术的特点

1.针对原始数据集做深度的理解和挖掘，充分利用了所有提供的数据；

2.使用了机器学习中的xgboost算法进行模型的训练，通过编程实现了复杂的机器学习算法并将其运用到实际应用中；

3.针对本项目在特征选择以及特征工程方面做了一定的创新，使得分类的结果大幅提高。

3、市场调研与分析

随着市场经济的发展和经济全球化的推动，公司面临的市场环境越来越复杂，存在的风险较以往更加严峻，更隐秘性和摧毁力。因此正确的有效的了解这些风险发生的可能性，将有利于我们进一步提高对风险的防范，这样才能做到未雨绸缪，防患于未然，使企业在汹涌澎湃的市场中持续健康的发展，从而取得预期的经济效益，实现持续的良性发展。

3.1市场定位

现如今大多数企业都存在或多或少的企业潜在风险，所以针对基本上所有的企业都需要进行企业风险预测。

3.2目标客户

比如像房地产行业的公司、移动互联网的公司、外汇中小企业都是我们潜在的顾客。

3.2市场前景

伴随着国内经济持续快速的发展的同时，各行各业都存在一些风险值。所以该系统如果成熟了，将拥有巨大的国内及国际市场。

4、项目可行性分析

4.1技术可行性分析

企业营业退出风险预测系统主要是基于历史数据信息发现潜在的规律，从而预测潜在的风险。该系统有自己独特的优越性：

1、简单性：在实现预测功能的同时，尽量让该平台操作简单易懂。

2、针对性：该平台设计使主要针对一些已经有一定的历史发展的公司进行未来两年的风险预测，所以具有专业突出和很强的针对性。

3、实用性：该平台能给企业未来的发展提供一定的建议，具有良好的使用性。

4、技术可行性：该系统的数据是基于该公司前几年的历史数据，然后对未来几年的发展进行预测。

4.2经济可行性分析

经济可行性：由于未来社会的发展，各行各业都将都需要对各自企业未来的发展提供一定的建议。所以给系统存在着很大的市场。而且该系统的维护相对简单，而且是机器学习经验丰富的软件开发人员一起共同研究的，这将为以后系统的顺利开发提供了有力的技术条件。

5、项目推广途径

5.1运营方式

运营方式采用线上免费查询该企业在未来两年内是否因经营不善退出市场的风险值，通过该平台再根据每个公司的情况，给出一些相应的建议（建议就需要收取一定的费用）。

5.2 推广平台

可以通过微信，支付宝，广告等进行推广。

6、融资方案及回报

人力成本：对于该软件系统的日常管理，维护与更新所支付专业人员的薪资，约占成本的45%。
办公场地：用于办公，约占成本的22%。
推广费用：用于系统的推广，比如广告等约占成本的25%。
其他费用：约占8%。

6.1融资方案

前期资金可以向亲朋好友借，但是这种凑资方案所能凑到的钱是有限的，不能满足较大数目的资金需求。所以这个时候就需要风险投资了。

因为我们是大学生创业项目，所以我们的技术创新效率高，有更多的活力，更加能适应市场的变化，并且我们的规模小，需要的资金量也小，风险投资公司所冒的风险也就很有限，我们的发展余地也大，所以风险投资者可能会接受我们。

6.2 回报方案

1、兼并收购，根据投资者的意愿，我们可以与大的行业公司并购，进行资本结合，以获取现金回报。

2、持续经营，过可持续发展规模化经营，严格管理，取得运营收益，收回投资，从而取得效益增长。

3、公开上市，此方式最能体现我们的市场价值，对风险投资者也是最理想的退出并获取高额回报的方式，如果状况良好，我们也在在3-5年之内上市，投资人股权不高于50%。