企业名称：量子数聚

赛题名称：企业停业风险预测

# 一、赛题概述

## 1.1技术方向

|  |
| --- |
| 数据挖掘，分类预测，机器学习 |

## 1.2赛题背景

|  |
| --- |
| 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，正是本赛题需要解决的主要问题。  本次大赛从全国2000多万企业抽取部分企业（脱敏后），提供企业主体在多方面留下的行为足迹信息数据。参赛队伍需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任务描述

|  |
| --- |
| 本赛题以企业为中心，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。  参赛者需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。  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4位数字**）相同的情况下，则以F1-score（Precision及Recall）进行辅助评估。 |

## 1.4.赛程设置

# 二、数据介绍

## 2.1数据基本情况

本赛题提供两种数据

（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，正常经营0

该表仅训练数据有。

选手从数据中自行组成特征和数据格式，自由组合训练测试数据比例。

为保护数据安全，所有数据均已进行了采样和必要的脱敏处理。

数据中部分列存在空值或NULL，记录可能存在重复情况，请参赛者根据数据字段描述自行处理。

## 2.2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 |
| TARGET | varchar2(2) | 目标标签 | 1停业，0正常。评测数据无该字段 |
| **ENDDATE** | **date** | **停业时间** | **到月度。仅TARGET=1的企业有该字段** |

注：ENDDATE为企业停业时间，复赛新增字段。该停业时间仅为官方列入的停业时间。

一般情况下，企业的真实停业时间较官方时间要早；少数情况下，企业虽然已经在官方列入停业状态，但企业仍有可能发生一些行为（如投资，申请权利，招聘等）。

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 |

## 2.3训练及测试数据

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

所有数据表中，灰色底纹的字段为主键列

### 2.3.1 企业基本信息数据1entbase.csv

复赛为参赛者提供了两个不同省份的抽样数据，并在基本信息表中新增了标识企业所在省份的字段。

|  |  |  |  |
| --- | --- | --- | --- |
| **name** | **type** | **字段名称** | **说明** |
| EID | varchar2(50) | 企业ID | 企业唯一标识 |
| **PROV** | **varchar2(4)** | **省份代码** | **复赛新增** |
| 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 |  |
| FINZB | number |  |
| FSTINUM | number |  |
| TZINUM | number |  |

### 2.3.2 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 |

### 2.3.3 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 |

### 2.3.4 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) | 被投企业是否在同一个省 | 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 |

### 2.3.5 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 |

### 2.3.6 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 |
| DJDATE | date | 中标日期 | 到月度 |
| IFHOME | varchar2(50) | 项目地是否是企业登记地 | 1为本省0为外省 |

### 2.3.7 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 |

### 2.3.8 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 |

### 2.3.9 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 | 招聘网站标识 |
| POSCODE | varchar2(200) | 招聘职位代码 | 招聘职位代码 |
| RECDATE | date | 招聘日期 | 月度 |
| PNUM | varchar2(200) | 招聘人数 |  |

注：招聘数据9recruit.csv数据结构较初赛有变化

### 2.3.10 企业资质10qualification.csv（新增数据表）

|  |  |  |  |
| --- | --- | --- | --- |
| **name** | **type** | **字段名称** | **说明** |
| EID | varchar2(50) | 企业ID | 企业唯一标识 |
| ADDTYPE | varchar2(50) | 资质代码 |  |
| BEGINDATE | date | 资质授予日期 |  |
| EXPIRYDATE | date | 资质失效日期 | 空为失效日期缺失 |

# 3、作品提交和评价标准

## 3.1作品提交

提交CSV 结果文件

提交的CSV 文件中包含对所有企业的预测结果（0或1），以及为1的发生概率。

文件名：evaluation\_public.csv

|  |  |  |
| --- | --- | --- |
| **name** | **字段名称** | **说明** |
| EID | 企业ID | 评测数据集中的ID，需保证唯一，若有重复ID，则返回错误代码。 |
| FORTARGET | 预测结果 | 预测结果，0或1。若为空，则视为0 |
| PROB | 预测结果为1的概率 | 0-1之间的数，至少保留4位小数。若为空，则视为0 |

## 3.2评测规则

|  |
| --- |
| （按照最终大赛要求进行调整）  1）队伍人数上限：5人  2）数据使用：本赛题数据仅允许用于本次竞赛相关活动，禁止参赛者用做他用  3）外部数据：本赛题仅允许使用“竞赛数据”中提供的数据，不允许使用任何外部数据。 |

## 3.3评价标准

|  |
| --- |
| 本赛题以AUC值（AUC的评判使用预测概率）作为主要评测标准，F1分数（F1分数使用预测结果）作为辅助评测标准。  F1 =2\* Precision\*Recall / (Precision+Recall)  具体规则为：在AUC值（**保留到小数点后4位数字**）相同的情况下，以F1-score（Precision及Recall）进行辅助评估。  **评分公式如下：score=AUC取4位小数\*10000 + F1**  **评测得分的范围为【0,10001】**  注意：如果某些EID样本您没有给出预测结果（或预测概率），则都会被作为0对待。 |

# 4、答疑（FAQ）

**4.1.Q：训练数据的正、负样本？**

A：目标数据中TARGET=1为正样本，0为负样本。

**4.2.Q：数据有什么特殊特点？**

A：（1）企业基本信息数据是身份数据，一个企业只有一条记录；

（2）其他数据均为行为数据，只有有此项行为的企业才有数据。当然，企业可能有多个行为；

**4.3.Q：关于提交错误**

要求文件格式：

文件编码为 “无BOM-UTF8”、分隔符为英文半角逗号“,”的标准csv格式文件

文件大小不能超过20M（复赛）；文件大小不能超过8M（初赛）

文件记录数不能少于21万条（复赛）；文件记录数不能少于9万条（初赛）

不能有重复id

提交文件包含的字段必须是3列

预测值和预测概率必须能够转换为数值，不能含有非法字符和其他不能转换为数值的字符

预测值必须是0或者1

预测概率必须是0-1之间的数

以上规则任一条不符合，文件均不能正常提交