赛题及数据集

基于企业发票虚开事实行为特征，根据历史案发数据标识，使用模型算法提取虚开发票特征，解决识别发票虚开的难题。

# 虚开发票特征提示

1、公司每月开票顶额开具，多次增量增版，短时间内开具增值税发票金额突增。

2、短期内多户企业登记法人为同一人，法人手机号等地址信息类同，法人、财务等交叉担任，法人户籍非本地、法人设立异常集中、法人存在非正常企业 。

3、资金或存货周转次数平均每月超过多次，进项和销项数据异常（有进无销或有销无进,购销背离、购销品名不一致）。

4、成立时间短，短期内大量开票。

5、公司所属行业属于虚开高危行业，或者公司注册地址存在异常。

6、是否申报报税、是否存在入库税款。

7、企业名称不符合日常命名规范（语义识别）。

数据

比赛数据（脱敏后）抽取的是一段时间范围内，企业基本信息、企业开票数据，参赛选手根据特征字段信息进行建模，预测存在发票虚开的企业数据情况

1、训练集：5000条企业数据,每行数据代表一个训练样本，各字段之间由逗号分隔，企业纳税人识别号

（NSRSBH）为唯一识别主键

2、测试集:200条虚开发票企业数据标识，每行数据代表一个测试样本，格式为"NSRSBH, result"，代表纳税人识别号和预测结果。

|  |
| --- |
| 训练集名称 |
| 企业基本信息 |
| 企业开票信息 |
| 企业纳税情况 |

4、各字段数据描述如下：

企业基本信息

|  |  |  |
| --- | --- | --- |
| 字段 | 字段中文说明 | 示例 |

|  |  |  |
| --- | --- | --- |
| 字段 | 字段中文说明 | 示例 |
| NSRSBH | 主键:纳税人识别号  (md5加密) | de7fb199806bacebe5fe190b79a08d0b |
| ZCDZ | 注册地址 | 芜湖市镜湖区九华中路２９６号－１ |
| SCJYDZ | 生产经营地址 | 芜湖市镜湖区九华中路296号-1 |
| FDDBRXM | 法定代表人姓名 | 丁爱道 |
| FDDBRSFZHM | 法定代表人身份证号码  (md5) | 4d8a6b88135623dea87377923428d6c9 |
| DJRQ | 登记日期 | 41297 |
| JYFW | 经营范围 | 投资咨询服务（证券、期货除外），劳务派遣， 房屋、汽车中介服务，家政服务，  建筑物清洁及维护；景观照明、中央空调、 排烟管道设计、安装、维护、清洁；物业管理  （凭资质证经营）；石材铺设、翻新、养护。 |
| SJZB | 实缴资本 | 50万 |
| ZCZB | 注册资本 | 100000 |
| TZZE | 投资总额 | 0 |
| CYRS | 从业人数 | 2 |
| KYSLRQ | 开业设立日期 | 2017-11-25 |
| ZCDLXDH | 注册地联系电话(md5) | 5df0552d2eb095cd504f7a06ae8e543d |
| BSRXM | 办税人姓名 | 张俊生 |
| BSRYDDH | 办税人移动电话(md5) | 5df0552d2eb095cd504f7a06ae8e543d |
| CWFZRXM | 财务负责人姓名 | 张俊生 |
| CWFZRYDDH | 财务负责人移动电话  (md5) | 5df0552d2eb095cd504f7a06ae8e543d |
| HYDL | 行业大类 | 零售业 |
| HYZL | 行业中类 | 综合零售 |
| HY | 行业 | 其他综合零售 |

# 企业开票信息

|  |  |  |
| --- | --- | --- |
| 字段 | 字段中文说明 | 示例 |
| NSRSBH | 主键:纳税人识别号(md5加密) | de7fb199806bacebe5fe190b79a08d0b |
| YF | 月份 | 2017-12 |
| YJXZE | 月开票进项总额 | 30000 |
| YJXCS | 月开票进项次数 | 10 |
| YXXZE | 月开票销项总额 | 20000 |
| YXXCS | 月开票销项次数 | 5 |
| KPZD | 开票额度 | 30000 |
| GMQD | 购买商品汇总清单 |  |
| XSQD | 销售商品汇总清单 |  |

税务登记信息

|  |  |  |
| --- | --- | --- |
| 字段 | 字段中文说明 | 示例 |
| NSRSBH | 主键:纳税人识别号  (md5加密) | de7fb199806bacebe5fe190b79a08d0b |
| START\_DATE | 起始时间 | 2015/9/1 |
| END\_DATE | 终止时间 | 2015/9/30 |
| TAX\_CATEGORIES | 税种 | 个人所得税 |
| TAX\_ITEMS | 税目 | 增值税附征 |
| TAXATION\_BASIS | 计税依据 | 6607.3 |
| TAX\_RATE | 税率 | 0.2 |
| DEDUCTION | 扣除数 | 555 |
| TAX\_AMOUNT | 税额 | 766.46 |
| OPEN\_INVOICE\_DATE | 开票日期 | 2015/10/14 |
| STORAGE\_DATE | 入库日期 | 2015/10/15 |

评分标准

算法得分**100**分

评分通过logarithmic loss（记为logloss）评估模型效果，logloss越小越好。

1 *N*

*Score* = − *sum*(*y* = 1) ∑ (*yi* ∗ *log*(*pi*)) ∗ 100 × 0.9 + *Tr*

*i i*=1

其中N表示测试集样本数量，yi表示测试集中第i个样本的真实值，pi表示第 i个样本的风险概览。T 代表运行运行时间,r代表运行得分区间系数,分为5个区间系

数:T1(1,T>600s),T2(3,360s<=T<600s),T3(5,120s<=T<360s),T4(7,60s<=T<120s),T5(10,T<60s)

参赛步骤

1、选手使用jupyter环境开发进行开发，提交代码集为Jupyter Notebook数据格式

2、选手根据我们提供的pip安装类库，根据主办方提供的入参格式，本地/线上调试，得到Pandas库的DataFrame的数据结构（训练集）

3、根据DataFrame训练集和主办方提供的测试集数据进行数据模型设计、训练

4、根据主办方提供的入参格式得到样本集数据，和模型进行训练，得到结果集，提交结果集DataFrame的数据结构

5、根据主办方提供的入参格式在线平台进行结果集评分测算，打榜日期截止2019年12月13日12时