### 1.案例大背景

某市税务局早期是利用传统的风险管理模型来防范发票虚开风险，该传统模型过分的依赖于税务干部业务水平和工作经验的主观性判断，防范效果不稳定且效果一般。因此税务局就想将数据安全开放给第三方的科研机构，通过第三方使用“机器学习”算法智能化高效化的预测纳税人虚开发票的风险指数。

### 2.任务介绍

案例的任务：第三方数据分析师根据税务局提供的企业发票虚开事实行为的一系列特征，以及历史案发数据标识，使用模型算法提取虚开发票特征，训练机器学习模型，解决识别虚开发票效率低、准确率不高的问题。

该任务预测的标签是：该企业是否有发票虚开的行为。是的标签为1， 否的标签为0。

另外，税务局指定了评估模型的指标，计算公式如下所示

其中N表示测试集样本的数量，yi表示第i个样本的真实值，pi表示第i个样本的预测概率值。T表示推理部分耗时，r表示运行时间得分，分为5个区间：T1(1, T>600s), T2(3, 360s<=T<600s), T3(5, 120s<=T<360s), T4(7, 60s<=T<120s), T5(10, T<60s)。

该得分值越大表示模型的效果越好。

### 数据集简介

数据集包括三个表：企业基本信息、企业的开票信息以及企业的纳税情况。各自表的字段含义如下所示;

#### 3.1企业基本信息表说明

|  |  |  |
| --- | --- | --- |
| 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 |

#### 3.2企业的开票信息表

|  |  |  |
| --- | --- | --- |
| 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 |  |

#### 3.3企业的纳税情况表

|  |  |  |
| --- | --- | --- |
| 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 |

#### 任务理解

根据领域知识考虑以下的情况会存在发票虚开的问题：

##### 4.1发票顶额开具

公司每月的发票顶额开具，且多次增量增版，或者短时间内开具增值谁发票金额突增

##### 4.2公司法人情况异常

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 。

##### 4.3资金周转异常

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

##### 4.4短期暴增发票

成立时间短，短期内大量开票。

##### 4.5高危行业检测

公司所属行业属于虚开高危行业

公司注册地址存在异常。

##### 4.6报税异常检测

是否申报报税

是否存在入库税款。

#### 5数据集字段探索

发票核定：是指主管税务机关根据纳税人的经营范围、经营规模、申报纳税等信息，对纳税人提出的领购发票的种类、面额、数量、每月开票限额、开票项目、购票方式进行核定的过程，包括初始核定和变更核定。

2017表示年份

Q1-4表示某一年的四个季度

字段含义解释

|  |  |  |
| --- | --- | --- |
| 字段名称 | 含义说明 | 备注 |
| fphdsl | 发票核定数量 | 公司最多能开的发票数量 |
| kps | 开票数 | 公司实际开具的发票数量 |
| kpje | 开票金额 | 公司实际开的发票总金额 |
| kpse | 开票税额 | 应缴纳的税额 |
| kpsl | 开票税率 | 应缴纳的税额占总额的比例 |
| rkse | 入库税额 |  |
| jxje | 进项金额 |  |
| jxsl | 进项税率 |  |
| zjnsrsbh | 主键：纳税人识别号 |  |
| nsrmc | 纳税人名称 |  |
| hyml | 行业门类 |  |
| hydl | 行业大类 |  |
| hyzl | 行业中类 |  |
| hy | 行业 |  |
| zcdz | 注册地址 |  |
| zczb | 注册资本 |  |
| scjydz | 生产经营地址 |  |
| fddbrxm | 法定代表人姓名 |  |
| fddbrmp | 法定代表人身份证号，经过了md5编码 |  |
| djrq | 登记日期 |  |
| xzjd | 乡镇街道 |  |
| jyfw | 经营范围 |  |
| cyrs | 从业人数 |  |
| frsjh | 法人手机号 |  |
| bsrxm | 办税人姓名 |  |
| bsrxmmp | 办税人手机号md5编码 |  |
|  |  |  |