|  |  |  |
| --- | --- | --- |
| 学号 | 姓名 | 任务 |
| 41721086 | 韩申 | 独立完成作业 |

非公企业成长性综合评价及分类

摘要：对企业成长性的研究，可以为企业分析其发展的优、劣势提供帮助。本文运用主成分分析法对非公有上市公司的成长性进行综合评价并按照评价结果优劣将非公有上市公司分为4类。并通过比较不同机器学习算法对非公有上市公司的企业成长性进行分类的效果。筛选出最优的分类模型。

关键词：主成分分析法、支持向量机、logistic模型、决策树

## 问题分析与细化

本文主要解决以下两个问题：

1. 通过数理统计的方法构建描述非公企业的成长性的指标，并根据该指标对企业进行分类。
2. 按照以上的分类结果。结合机器学习算法，使用更加简单的方式（选择尽可能少的变量）对非公企业的成长性进行有监督的分类。

问题一分析：

非公企业的成长性是指企业在发展过程中所具备的整合其内外部经济资源、提高其自身价值的能力，这是一种长期的、持续性的能力。一般而言上市公司的成长性可由以下五个一级指标来判断：盈利能力，偿债能力，营运能力，企业税负，发展能力。选取这些一级指标下属的一些有具体数学表达的二级指标来对一级指标进行描述从而间接的得到非公企业的成长性指标。由于成长性指标并没有统一标准，且市场上缺乏成长性具有鲜明特点的非公企业，因此不能按照得出成长性指标对非公企业成长性进行绝对划分。但是可以根据构建的成长性指标将所有的非公企业划分为4类，使得不同类之间的成长性指标有明显差异。

问题二分析：

通过判断上个问题所构造的成长性指标与各个变量之间的相关性、不同变量之间的相关性对变量进行筛选，获得与成长性相关性强，而各个变量之间相关性较弱的变量组合作为最终机器学习模型所需要的变量。根据这个变量组合以及问题一中的分类结果构造有监督的分类模型。通过对比不同模型之间的预测准确率来获得最好的企业成长性分类模型。

## 数据来源及变量说明

选择五个一级指标：盈利能力，偿债能力，营运能力，企业税负，发展能力下的18个二级指标，进行成长性指标的构建。从wind数据库获得我国所有上市公司2010年到2018年年报以及2019年中期报告中对应的18个财务指标作为原始数据。

选择指标如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 一级指标 | 二级指标 | 一级指标 | 二级指标 |
| 营运能力 | 总资产周转率/次 | 发展能力 | 营业收入增长率/% |
| 总资产报酬率/% | 每股净资产增长率/% |
| 净资产收益率/% | 净利润增长率/% |
| 应收账款周转率/次 | 所有者权益合计/元 |
| 存货周转率/次 | 总资产(同比增长率)/% |
| 偿债能力 | 流动比率/% | 盈利能力 | 营业利润/营业总收入/% |
| 速动比率/% | 销售净利率/% |
| 资产负债率/% | 企业税负 | 税金及附加/元 |
| 经营活动现金流量净额/负债合计 | 年末所得税率/% |

表1：一级指标、二级指标对应表

为了使数据变量更容易表达以及和数据表格中对于对18个变量按照如下吮吸进行编码：

|  |  |  |  |
| --- | --- | --- | --- |
| 编码 | 指标 | 编码 | 指标 |
| X1 | 总资产周转率/次 | X10 | 营业收入(N年,增长率)/% |
| X2 | 总资产报酬率/% | X11 | 每股净资产(相对年初增长率)/% |
| X3 | 净资产收益率/% | X12 | 净利润(增长率)/% |
| X4 | 流动比率/ | X13 | 应收账款周转率/次 |
| X5 | 速动比率/ | X14 | 存货周转率/次 |
| X6 | 资产负债率/% | X15 | 营业利润/营业总收入/% |
| X7 | 经营活动现金流量净额/负债合计 | X16 | 销售净利率/% |
| X8 | 所有者权益合计/元 | X17 | 税金及附加/元 |
| X9 | 总资产(同比增长率)/% | X18 | 年末所得税率/% |

表2：二级指标编码对应表

## 数据处理

观察原始数据发现如下问题：

1. 原始数据不知包含非公企业也包括许多上市的国有企业。
2. 许多上市公司相关变量的数据存在缺失。
3. 数据单位不统一，且数量级相差较大，如所有者权益合计数量级到109而其他比率指标在101~102。

处理方式：

对于问题1由于研究对象是非公企业，所以直接将国有企业的数据删除。

对于问题2可以选择删除存在缺失数据的样本或者对缺失的数据进行数理补齐，由于获得了10年的样本数据，因此数据量较大，个别缺失数据对总体特征影响不会很大。并且数理补齐缺失数据存在一定的‘失真’有可能会影响到模型质量。如:在支持向量机模型中，补齐的数据如果被选为支持向量，就会对这个模型造成较大的影响。因此对于存在数据缺失的样本也进行删除。

对于问题3变量数量级差异过大会导致数量级大的变量在构建成长性指标的过程中占有过大的权重，从而使得成长性指标几乎由数量级大的变量决定，数量级小的变量在成长性指标中体现不出来。针对这个问题能够采取对数据标准化或归一化处理来消除量纲，缩小数量级。考虑到选择的可能存在一些某些指标非常好，或非常差的离群样本。因此样本的最大值与最小值不一定能够包含在所选择的样本中，会对模型的预测造成较大的影响。因此选择对数据进行标准化处理。

经过数据处理后得到了，不含国有企业数据、不含缺失数据、且进行了标准化的14667个18维的样本数据其中由693个ST企业。并将此数据作为基础构造成长性指标，建立分类模型。

## 成长性指标构造

### 构造标准

首先列出两个成长性指标的构造标准：

1. 成长性指标应当能够尽可能的涵盖上述的18个二级指标的主要特征。
2. 其中成长性指标越差的样本是ST企业的可能性更大。
3. 成长性指标尽可能和所选变量呈线性关系

### 指标构造

主成分分析法来构造成长性指标：

（一）：相关性检验：

要进行主成分分析首先需要对选择的18个指标进行相关性检验。如果相关性太差则无法使用某一主成分来表示高维数据的特征。

利用SPSS对经过数据处理的数据进行KMO以及巴特利特检验。

|  |  |  |
| --- | --- | --- |
| **KMO和巴特利特检验** | | |
| KMO取样适切性量数。 | | .526 |
| 巴特利特球形度检验 | 近似卡方 | 122802.654 |
| 自由度 | 153 |
| 显著性 | .000 |

表3：变量KMO巴特利特检验结果

得到KMO值为0.526，大于0.5，表明主成分分析尚可接受，已达到主成分分析的可行性标准。Bartlett球形检验值为122802.654，检验显著性水平为0．000，表明研究中使用的变量间具有较强的相关性，说明主成分分析是有效的。

（二）对变量进行进一步的处理：得到各主成分的特征值及其方程贡献率：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 总方差解释 | | | | | | | | | |
| 成分 | 初始特征值 | | | 提取载荷平方和 | | | 旋转载荷平方和 | | |
|  | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % |
| 1 | 2.544 | 14.135 | 14.135 | 2.544 | 14.135 | 14.135 | 2.308 | 12.823 | 12.823 |
| 2 | 2.093 | 11.627 | 25.762 | 2.093 | 11.627 | 25.762 | 1.948 | 10.825 | 23.648 |
| 3 | 1.762 | 9.788 | 35.550 | 1.762 | 9.788 | 35.550 | 1.786 | 9.921 | 33.568 |
| 4 | 1.480 | 8.225 | 43.775 | 1.480 | 8.225 | 43.775 | 1.642 | 9.120 | 42.689 |
| 5 | 1.376 | 7.643 | 51.417 | 1.376 | 7.643 | 51.417 | 1.503 | 8.352 | 51.041 |
| 6 | 1.050 | 5.835 | 57.253 | 1.050 | 5.835 | 57.253 | 1.107 | 6.151 | 57.192 |
| 7 | 1.015 | 5.638 | 62.891 | 1.015 | 5.638 | 62.891 | 1.026 | 5.699 | 62.891 |
| 8 | 1.000 | 5.556 | 68.447 |  |  |  |  |  |  |
| 9 | .978 | 5.432 | 73.879 |  |  |  |  |  |  |
| 10 | .973 | 5.407 | 79.286 |  |  |  |  |  |  |
| 11 | .940 | 5.225 | 84.511 |  |  |  |  |  |  |
| 12 | .845 | 4.697 | 89.207 |  |  |  |  |  |  |
| 13 | .755 | 4.196 | 93.403 |  |  |  |  |  |  |
| 14 | .501 | 2.782 | 96.185 |  |  |  |  |  |  |
| 15 | .369 | 2.048 | 98.232 |  |  |  |  |  |  |
| 16 | .250 | 1.388 | 99.621 |  |  |  |  |  |  |
| 17 | .061 | .340 | 99.961 |  |  |  |  |  |  |
| 18 | .007 | .039 | 100.000 |  |  |  |  |  |  |
| 提取方法：主成分分析法。 | | | | | | | | | |

表4：主成分的特征值及其方程贡献率

得到特征值大于1的7个主成分以及各自的权重。（第i 个主成分的权重为该主成分方差百分比除以七个主成分累计的方差百分比。记为ai.）各个主成分的各个主成分的系数如下：

（三）用最大方差法对主成分进行旋转后得到的旋转后的主成分系数矩阵如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 旋转后的成分矩阵 | | | | | | | |
| 成分 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 总资产周转率 | -.085 | -.028 | -.136 | .277 | .079 | .547 | .013 |
| 总资产报酬率 | .091 | .144 | -.015 | .878 | .039 | .026 | .037 |
| 净资产收益率 | .009 | .001 | .036 | .841 | .008 | -.062 | .016 |
| 流动比率 | .938 | -.003 | .004 | -.081 | .128 | -.029 | .002 |
| 速动比率 | .945 | -.002 | -.001 | -.079 | .137 | -.034 | .000 |
| 资产负债率 | -.345 | -.048 | .094 | -.098 | .012 | .349 | -.015 |
| 经营活动产生的现金流量净额除总负责 | .578 | .031 | -.032 | .180 | -.088 | .064 | -.009 |
| 所有者权益合计 | -.044 | .009 | .927 | .003 | .002 | .016 | -.003 |
| 总资产 | .131 | .000 | -.001 | .035 | .847 | -.052 | .010 |
| 营业收入 | .006 | -.044 | .008 | -.052 | -.027 | .024 | .813 |
| 每股净资产 | -.018 | .096 | -.002 | .014 | .851 | .025 | .020 |
| 净利润 | -.005 | .077 | .001 | .079 | .043 | -.056 | .597 |
| 应收账款周转率 | .083 | .020 | -.020 | -.032 | -.042 | .581 | .007 |
| 存货周转率 | .036 | .009 | -.009 | -.011 | -.012 | .216 | -.034 |
| 营业利润除营业总收入 | .023 | .978 | .004 | .035 | .053 | .004 | .031 |
| 销售净利率 | .027 | .974 | .005 | .108 | .054 | .000 | .031 |
| 税金及附加 | -.018 | -.001 | .922 | .014 | -.004 | .012 | .011 |
| 年末所得税率 | -.193 | -.013 | .212 | -.089 | -.003 | .532 | .050 |

表5：旋转后的成分矩阵

（得到f1=-.085\*X1+.091\*X2+……-.193X18  …… f7=.013\* X1+.037\*X2+……+.050 X18）

（四）构造成长性指标：

通过对主成分进行加权求和得到成长性指标F。

F=a1\*f1+a2\*f2+a3\*f3+a4\*f4+a5\*f5+a6\*f6+a7\*f7

### 指标检验

该指标虽然能够在一定程度上描述18个变量的特征，但是是否能使得得分越低分的企业是ST企业概率越高还需要验证。采用如下方式进行验证：

首先计算所有样本的成长性指标F。之后取所有样本F值得10十分位数。将样本按照得分高低将样本均分为10组统计每组中ST企业得数量。（Fi表示第i组的得分，F1<-.4007,-.4007<= F2<-.3150,……0.2200<= F9<0.5222, F10>=0.5222）

|  |  |  |
| --- | --- | --- |
| 百分位数 | 10 | -.4007 |
| 20 | -.3150 |
| 30 | -.2464 |
| 40 | -.1816 |
| 50 | -.1122 |
| 60 | -.0372 |
| 70 | .0664 |
| 80 | .2200 |
| 90 | .5222 |

表6：成长性指标F的10分位点

此时得到每个组中ST企业数量如下表所示：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1组 | 2组 | 3组 | 4组 | 5组 | 6组 | 7组 | 8组 | 9组 | 10组 |
| 数量 | 235 | 109 | 77 | 59 | 39 | 43 | 39 | 34 | 29 | 29 |
| 合计 | 693 | | | | | | | | | |

表7：ST企业随F的分布

观察到构造的成长性指标基本满足先前提出的标准。

### 样本分类

根据构造的成长性指标对样本进行分类：

以组内差距最小，组间差距最大为原则。使用SPSS中的K均值分类方法，设置最大迭代次数为100，将样本按照成长性指标分为4组。此时迭代未达到最大迭代次数，但是迭代收敛，因此可以采用该分组结果作为组内差距最小，组间差距最大的分组结果。

（所得结果记录在文件kind.csv以及数据集data.sav中，kmeans迭代详细情况记录在：附件一.doc中）

其最终的迭代中心，以及各组中样本个数如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 最终聚类中心 | | | | |
|  | 聚类 | | | |
| 1 | 2 | 3 | 4 |
| 得分 | -9.11 | -.10 | 22.61 | 1.74 |

表8：各个分类的聚类中心

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | 1 | 14.000 |
| 2 | 13823.000 |
| 3 | 5.000 |
| 4 | 825.000 |
| 有效 | | 14667.000 |
| 缺失 | | .000 |

表9：各个分类中企业个数

不难看出：第一类中的企业成长性很差，第二类较差，第三类企业成长性非常好，第四类企业成长性则一般。并且发现，成长性极差，或者极好的企业都是少数，极差仅14个，而极好5个，据大部分企业成长性处于中下等水平，较少的优秀企业成长性比较优秀。这也与股市中流传的一句话 “7赔2平1赚” 相呼应。即：成长性较好的企业是较少的，不然股民们只要长线投资就能大概率的赚。

## 变量选择

### 选择标准：

1. 所选变量与企业种类是相关的。
2. 所选的变量组中任意变量之间不能存在显著线性相关关系。
3. 所选变量尽可能涵盖所有变量信息，并且尽可能少

### 选择方法及结果：

运用假设检验的方法来判断变量与变量之间是否相关。

对于选择标准1：

由于企业种类是离散的，而变量的值是连续的。离散变量与连续变量相关可以表达为：在不同的离散种类类别上，连续变量的分布是不同的。做出如下假设检验：

H0：在不同的离散种类类别上，连续变量的分布是相同的。

运用SPSS对该命题进行Kruskal-Wallis独立样本检验得到：在95%的显著性水平下，对18个变量中的每一个变量，都能够拒绝H0，即可以认为所有的变量都和企业的种类显著关。

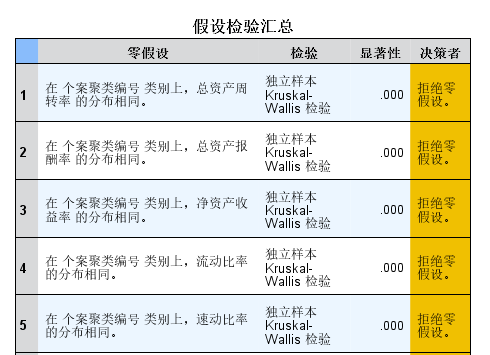


图1：部分变量Kruskal-Wallis检验结果

（详细的结果见文件：附件2.doc）

对于选择标准2：

通过多元线性回归模型逐步的筛出线性无关的变量，过程如下

假设H0：Xi与剩余的18-i个变量中的某些存在线性关系。运用spss多元线性回归得到：



图2：X1和剩余变量的线性回归检验

通过对变量系数的t检验得到如下结论：

在95%的显著性水平下认为X1总资产周转率和税金及附加和营业收入存在线性关系。因此删去X1保留X10和X17。

之后在剩余的变量中依次选择X2、X3等变量直到剩余的变量都需要保留。（具体过程见文件：附件3.docx 和 附件4.doc）

机器学习算法最终保留变量为X10、X11、X13、X14、X15、X16、X17、X18。

对于选择标准3

上述保留变量能够线性的表示被剔除变量因此包含了大部分变量的信息。对上述变量进行KMO 和巴特利特检验其相关性如下KMO值为0.509，虽然大于0.5，但是相关性十分弱。因此无需对上述保留的变量继续去除相关性。如果继续剔除变量将会损失部分变量的信息，因此不能继续剔除，此时选择变量达到尽可能少。

|  |  |  |
| --- | --- | --- |
| **KMO 和巴特利特检验** | | |
| KMO 取样适切性量数。 | | .509 |
| 巴特利特球形度检验 | 近似卡方 | 30622.325 |
| 自由度 | 28 |
| 显著性 | .000 |

表10：选择变量的相关性检验

## 模型算法介绍

### Logistic简介

定义二项logistic回归模型的条件分布如下：



其中x∈Rn是输入，Y∈{0,1}是输出，W∈Rn和b∈R是参数，w称为权重，b称为偏置。得到概率之后，可以通过设定一个阈值将样本分成两类。如：阈值为0.5的时候，当大于0.5则为一类，小于0.5为另一类。

为了进行多元分类，将权重向量和输入向量进行扩充：w = (w1,w2, …, wn, b)T，x = (x1,x2, …, xn, 1)T所以，logistic回归模型变为：



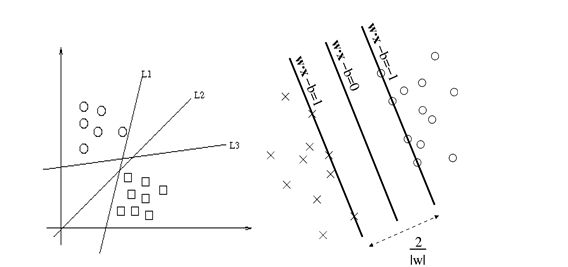
此时选择概率最大的一类作为x的结果。

能够通过梯度下降等一些数值求解的方法求解w的值

通过python来建立logistic模型并进行十折交叉验证，获得logistic模型下平均的预测准确率。

### 支持向量机简介

对于线性可分的二维数据，我们只需要使用一条直线ax1+bx2=c利用sign（ax1+bx2-c）就能把两类点区分开来。而在跟高维的空间，我们只需要一个超平面wx+b=0（w、x、b均为高维向量）我们也能通过函数sign（wx+b）把数据分为两类。而对于线性不可分的数据，我们使用核函数，把非线性可分的数据转换为线性可分即可。常用的核函数包括（线性核、高斯核、多项式核）



本文中使用python建立高斯核的SVM模型并通过网格搜索的方法寻找最佳的惩罚系数和高斯核系数。并进行十折交叉验证、获得最优参数下的预测准确率。

（PS：python中高斯核的支持向量机模型有两个重要的参数C和gamma：他们的作用如下：

C：惩罚系数，用来控制损失函数的惩罚系数，类似于LR中的正则化系数。C越大，相当于惩罚松弛变量，希望松弛变量接近0，即对误分类的惩罚增大，趋向于对训练集全分对的情况，这样会出现训练集测试时准确率很高，但泛化能力弱，容易导致过拟合。 C值小，对误分类的惩罚减小，容错能力增强，泛化能力较强，但也可能欠拟合。

gamma：核函数系数，gamma越大，σ越小，使得高斯分布又高又瘦，造成模型只能作用于支持向量附近，可能导致过拟合；反之，gamma越小，σ越大，高斯分布会过于平滑，在训练集上分类效果不佳，可能导致欠拟合。）

### 决策树简介

利用一种树形图作为分析工具。其基本原理是用决策点代表决策问题，用方案分枝代表可供选择的方案，用概率分枝代表方案可能出现的各种结果，经过对各种方案在各种结果条件下[损益值](https://baike.baidu.com/item/%E6%8D%9F%E7%9B%8A%E5%80%BC)的计算比较，为决策者提供决策依据。

本文中使用python选择信息熵作为决策树构建依据，并通过对决策树层数的最大值对决策树进行预剪枝，建立决策树模型，进行十折交叉验证并计算平均预测准确率。

## 结果与解释

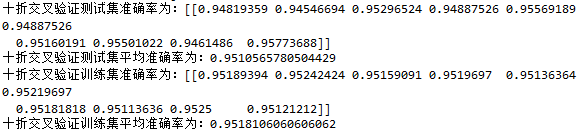
### Logistic模型预测结果

十折交叉验证下训练集和测试集的准确率，以及平均准确率如下：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| （%） | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 平均 |
| 训练集 | 95.18 | 95.24 | 95.16 | 95.20 | 95.14 | 95.22 | 95.18 | 95.11 | 95.25 | 95.12 | 95.18 |
| 测试集 | 94.82 | 94.55 | 95.30 | 94.89 | 95.57 | 94.89 | 95.16 | 95.50 | 94.61 | 95.77 | 95.10 |

表11：logistic模型十折交叉验证下训练集和测试集的准确率

程序运行原始结果如下：



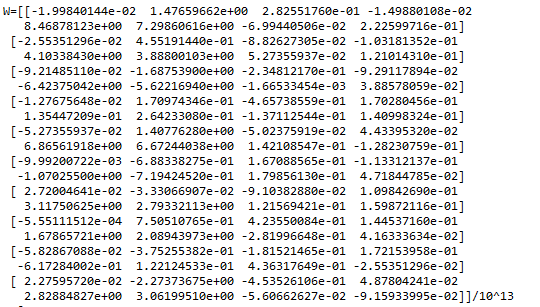
此时十折交叉验证所训练出的参数W如下：

W1=[ -1.99840144e-02、1.47659662e+00、2.82551760e-01、-1.49880108e-02、 8.46878123e+00、7.29860616e+00、 -6.99440506e-02、2.22599716e-01]/10^13

…

W10=[ 2.27595720e-02、-2.27373675e+00、-4.53526106e-01、4.87804241e-02、2.82884827e+00 3.06199510e+00、-5.60662627e-02、-9.15933995e-02] /10^13

程序运行结果如下：



**结果解释：**

Logistic模型虽然由于数据标准化导致训练的参数数量级过小，但是对于测试集的预测结果是十分令人满意的，其对企业分类的准确率高达95%，其原因可能是成长性指标本身就是各个变量的线性组合，企业种类由成长性指标得到，而logistic模型也正好是一个线性回归模型，因此对该问题有较好的预测结果。但是由于企业种类1和种类3所占样本数过少，因此对于这两类企业而言logistic模型预测的结果缺乏可信度，还需要结合其他指标进行经济学判断。但对于占大多数的2类企业和4类企业logistic模型的预测是十分优秀的。

### SVM模型预测结果

对合适的参数进行网格搜索，选择平均测试集预测准确率最高的以组参数作为最终的参数：（详细情况见文件：附件5.docx）

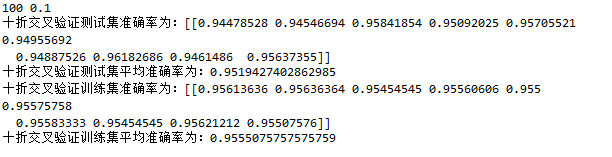
网格搜索如下所示：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| C | gamma | 测试平均 | 训练平均 | C | gamma | 测试平均 | 训练平均 |
| 1 | 0.00001 | 94.08% | 94.26% | 100 | 0.00001 | 94.29% | 94.50% |
| 0.0001 | 94.18% | 94.36% | 0.0001 | 94.35% | 94.58% |
| 0.001 | 94.29% | 94.52% | 0.001 | 94.49% | 94.69% |
| 0.01 | 94.47% | 94.67% | 0.01 | 94.87% | 95.19% |
| 0.1 | 94.90% | 95.20% | 0.1 | 95.19% | 95.55% |
| 10 | 0.00001 | 94.16% | 94.35% | 1000 | 0.00001 | 94.34% | 94.58% |
| 0.0001 | 94.28% | 94.49% | 0.0001 | 94.44% | 94.66% |
| 0.001 | 94.43% | 94.64% | 0.001 | 94.63% | 94.88% |
| 0.01 | 94.63% | 94.89% | 0.01 | 95.04% | 95.34% |
| 0.1 | 95.17% | 95.41% | 0.1 | 95.15% | 95.69% |

表12：高斯SVM模型参数搜索

得到在选择的参数中当C取100，gamma取0.1时模型在测试集预测的平均准确率最高：

在这组参数下：程序运行结果如下：



此时十折交叉验证结果如下：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| （%） | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 平均 |
| 训练集 | 95.61 | 95.64 | 95.45 | 95.56 | 95.5 | 95.58 | 95.58 | 95.45 | 95.62 | 95.51 | 95.55 |
| 测试集 | 94.48 | 94.55 | 95.84 | 95.09 | 95.71 | 94.96 | 94.89 | 96.18 | 94.61 | 95.64 | 95.19 |

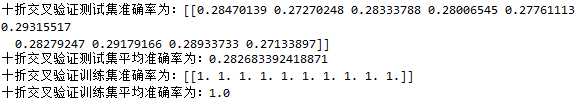
表13：SVM模型十折交叉验证下训练集和测试集的准确率

**结果解释：**

高斯核的SVM模型对于测试集的预测同样达到95%，具有令人满意的预测能力，其原因可能是，虽然高斯核的SVM模型不是线性回归模型，但是由于gamma值取得较小因此经过核函数变换后数据之间的线性关系没有发生太大变化，因此SVM模型预测结果同样令人满意。同logistic模型一样对于企业种类1和种类3而言SVM模型预测的结果也缺乏可信度，还需要结合其他指标进行经济学判断。但对于占大多数的2类企业和4类企业logistic模型的预测是比较可靠的。但是相比于logistic模型而言，高斯核SVM模型训练复杂度更高，并且网格搜索会很大的增加模训练的时间，对于更大样本容量的训练高斯核SVM模型在训练成本上是高于logistic模型的。

### 决策树模型预测结果及解释

在决策树模型中：在不进行预剪枝的情况下：程序运行结果如下：



在没有进行剪枝的情况下决策树模型出现了过拟合现象，使得模型在训练集准确率过高，而在测试集过低。

通过限制叶子节点样本数、限制决策树最大深度等方式对决策树模型进行剪枝得到的模型虽然在训练集上模型准确率有所降低，但是在测试集上的准确率却始终在28%左右。

**结果解释：**

由于企业种类仅仅分为四类，决策树模型在测试集准确率几乎和随机猜测准确率25%接近。于是可以认为决策树模型不适用于根据企业成长性对企业进行分类。

### 总结：

通过对以上三个模型预测结果的展示以及分析得出结论：在本文构造的成长性指标下，把企业划分为四类，并以此作为标签的情况下： 对于未知类型的企业进行分类时，logistic模型要优于高斯核的SVM模型。而决策树模型在此情况下，预测能力和随机猜测相差不大。因此在本文所描述的企业分类方式下，应当采用logistic模型对未知类型企业进行分类。

## 后续可以开展的工作

1. 针对指标构造：

根据不同的标准，可以构造不同的成长性指标。后续可以尝试寻找一些非线性的成长性指标来对企业进行描述，使得该非线性指标能将成长性良好的企业和成长性不好的企业更明显的区分开（两者指标差异较大）。此时再用机器学习模型对未知类型的企业进行预测，或许高斯核SVM在选择适当参数后，模型的预测效果要优于logistic模型。

1. 时间序列分析：

获得一些企业再更多年份的数据，对企业的数据进行时间序列分析。通过时间序列分析预测企业未来某些指标的值，在使用机器学习方法对预测值分类，最终得到未来企业成长性处于什么地位，实现对企业成长性的预测。

1. 使用更多的机器学习对未知类型企业进行分类，尝试找到比logistic模型更优的机器学习模型
2. 针对变量选择：

如果能够采取别的方式构造企业成长性指标，我们可以通过主成分分析法，提取主成分作为机器学习模型训练的指标。

## 参考文献：

[1]王志瑛,侯亭羽.基于财务视角的山西省高新技术企业成长性研究[J].经营与管理,2020(04):48-51.

[2]李飞,朱永明.云南省上市公司成长性综合评价与分析[J].商丘师范学院学报,2019,35(10):76-82.

[3]姜丽君,王莎莎.基于财务报表视角下对承德露露成长性的分析研究[J].中国乡镇企业会计,2019(02):71-75.

[4]卫萍. 上市公司信用债违约风险评估[D].华东理工大学,2017.

[5]赵序海,曾富全,何雪梅.ST上市公司成长性判断与血质指标设计与检验——基于1998~2013年中国ST上市公司的经验证据[J].经济研究参考,2015(41):66-72.

[6]赵君. 我国省域经济综合竞争力实证分析[D].辽宁师范大学,2015.

[7]马立新.湘西民族地区财产保险市场发展的实证分析[J].价值工程,2013,32(13):179-181.

**所有附件、程序、数据、参考文献可以在压缩包：（韩申+非公企业成长性综合评价及分类.zip）中找到**