第 8 卷◆第 11 期◆版本 1.0◆2025 年
文章类型：论文| 刊号（ISSN）：3082-8295(O) / 2630-4759(P)

基于因子分析和机器学习的生物医药公司价值预测模型

何群
贵州财经大学
DOI:10.32629/ej.v8i11.3113

[摘　要] 高质量发展是未来建设与经济发展必须贯彻的发展理念。本文选取我国461家生物医药上市公司作为研究对象,基于因子分析利用机器学习算法对生物医药公司构造价值预测模型。实证结果表明：营收规模与成长和综合盈利能力为影响价值分类的主要因子。三种机器学习算法中,多分类逻辑回归综合表现最优,测试准确率达到0.985,XGBoost次之；该研究中“因子分析+逻辑回归”模型的预测效果最优。为生物医药公司高质量发展提供参考。
[关键词] 生物医药；因子分析；机器学习；价值预测
中图分类号：TP242　文献标识码：A

Value Prediction Model for Biopharmaceutical Companies Based on Factor Analysis and Machine Learning

Qun He
Guizhou University of Finance and Economics

[Abstract] High-quality development is an essential concept for future construction and economic growth. This study selects 461 listed biomedical companies in China as research objects, employing factor analysis to reduce data dimensionality and machine learning algorithms to construct a value prediction model. Empirical results indicate that revenue scale and growth, along with comprehensive profitability, are the primary factors influencing value classification. Among the three machine learning algorithms, multiclass logistic regression performs the best with a test accuracy of 0.985, followed by XGBoost. The "factor analysis + logistic regression" model demonstrates optimal predictive effectiveness, providing a reference for promoting high-quality development in biomedical companies.
[Key words] Biomedical; Factor Analysis; Machine Learning; Value Prediction

引言

生物技术进入黄金10年,中国生物科技行业开启了从量变到质变的进程。相关部门出台政策提出要加大对生物医药高质量发展的支持,强调技术、资金、人才投入以补短板,行业迎来机遇与挑战。作为全球第二大医药市场,2024年中国规模以上医药企业营收达29762.7亿元,利润4050.9亿元,规模以上医药工业增加值同比增长3.4%,产业潜力吸引了大量投资者。

医药公司的数据多呈现高维度、非线性的分布特征。传统统计方法难以处理指标之间的多重共线性、复杂性等,存在明显的局限性。相比之下,机器学习就具有明显优势,机器学习是包含了多种技术、具有深度的理论基础和实践价值的先进体系,如今成为数据科学研究与应用的主流分析工具。

因此,本研究首先利用因子分析法对医药上市公司的财务数据进行降维与核心特征提取,得到综合财务评价模型；再提取公因子及综合财务得分为特征变量,引入多分类逻辑回归、XGBoost、随机森林三类机器学习算法构建价值预测模型,最后通过对比不同模型的性能指标,筛选最优预测方案。

1 文献综述

逻辑回归是一种用于分类的比较常见的统计学方法,通过Sigmoid函数的组合来建模事件的概率,函数的输出范围为[0-1],表示事件发生的概率。它的参数可解释性强、计算效率高,能满足医药行业的需求[1]。XGBoost是通过梯度下降优化损失函数、引入L1/L2正则化项控制模型复杂度的一种方法,在处理数据缺失与异常上表现比较突出,可以自动处理缺失值,适配数据缺口问题,同时可识别特征因子[2],为特征筛选提供量化依据。随机森林是基于Bagging框架与决策树的集成原理,通过随机抽样和特征选择双重机制,显著降低数据的过拟合风险[3]。

目前,已有相关研究利用“因子分析+”模型在多个领域证实了其可以提升预测精度和模型适配性的作用。在环境污染领

域,Jiaqiang等人[4]采用因子分析和机器学习模型结合的方法来预测污水处理厂出水中总氮和硝态氮的浓度,发现该模型显著提高了预测的精度。在岩土工程领域,简典[5]采用因子分析和集成机器学习的方法来预测岩质边坡稳定性,发现基于集成方法相比单一的模型具有更好的适配性。基于此,本文针对生物医药公司数据,运用“因子分析+多机器学习算法”的框架展开研究,利用因子分析提取的公因子,再结合多类机器学习算法构建预测模型,以期为生物医药公司优化财务策略、提升综合绩效,提供有价值的决策参考。

2 数据来源

本文数据来源于锐思数据库和各公司财务报表,研究对象是我国的461家生物医药上市公司,排除存在投资风险的公司和信息不完整的公司。基于全面性与可度量性原则,为全面反映出生物医药上市公司经济发展状况,本文选择15项代表性财务指标对财务绩效评价,具体如下：X_1：利润总额增长率(%)；X_2：营业利润增长率(%)；X_3：净资产收益率(%)；X_4：净利润增长率(%)；X_5：成本费用利润率(%)；X_6：营业利润率(%)；X_7：销售净利润率(%)；X_8：资产净利率(%)；X_9：净资产(元)；X_{10}：每股营业收入(元)；X_{11}：每股收益(元)；X_{12}：流动比率；X_{13}：速动比率；X_{14}：营业收入增长率(%)；X_{15}：总资产增长率(%)。

表1 特征根与方差贡献率

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	4.045	26.964	26.964	4.045	26.964	26.964	2.785	18.568	18.568
2	2.704	18.024	44.989	2.704	18.024	44.989	2.517	16.782	35.35
3	1.813	12.085	57.073	1.813	12.085	57.073	2.051	13.674	49.024
4	1.549	10.328	67.401	1.549	10.328	67.401	1.909	12.727	61.751
5	1.088	7.253	74.654	1.088	7.253	74.654	1.874	12.494	74.245
6	0.966	6.44	81.093	-	-	-	-	-	-
7	0.912	6.08	87.174	-	-	-	-	-	-
8	0.575	3.835	91.009	-	-	-	-	-	-
9	0.491	3.274	94.283	-	-	-	-	-	-
10	0.302	2.011	96.294	-	-	-	-	-	-
11	0.252	1.678	97.972	-	-	-	-	-	-
12	0.172	1.15	99.121	-	-	-	-	-	-
13	0.129	0.861	99.983	-	-	-	-	-	-
14	0.003	0.017	100	-	-	-	-	-	-
15	2.63E-09	1.75E-08	100	-	-	-	-	-	-

3 实证分析

3.1因子分析

(1)检验变量。因子分析的前提是所选指标间须存在一定相关性,以便依据因子贡献率提取公因子。因此,因子分析前,应对原始变量进行相关分析[6],确定各变量之间的相关性,以便确定是否适合因子分析方法。输出结果：KMO检验结果为0.756,表示基本上可以进行因子分析;Bartlett检验的结果中检验的p值为0.00＜0.05,表示拒绝原假设,适合进行因子分析。

(2)因子提取和因子旋转。对所选取样本的标准化数据进行因子分析,本研究因子提取方法为“主成分法”,采取方差最大化正交旋转,得出的各公因子方差贡献见表1。前5个因子的特征值均大于1提取作为公因子,累计方差贡献率可以解释74.245%的方差。

经过旋转后的因子载荷系数已经明显分化开来,并且对之前的指标体系分类进行了相关修正。根据载荷绝对值≥0.4确定每个公因子的核心关联指标：X3、X5、X8、X9、X11、X15归为公共因子F1,命名为综合盈利能力因子；X12、X13归为公共因子F2,命名为短期偿债能力因子；X1、X4归为公共因子F3,命名为利润成长能力因子；X6、X7归为公共因子F4,命名为销售盈利效率因子；X10、X14归为公共因子F5,命名为营收规模与成长因子。

(3)因子得分模型。根据因子得分系数矩阵可得出旋转后五个因子的因子得分表达式,再根据公因子特征值的贡献率,得出生物医药上市公司经济发展的综合得分表达式：

F=F1*18.568%+F2*16.782%+F3*13.674%+F4*12.272%+F5*12.494%

3.2机器学习

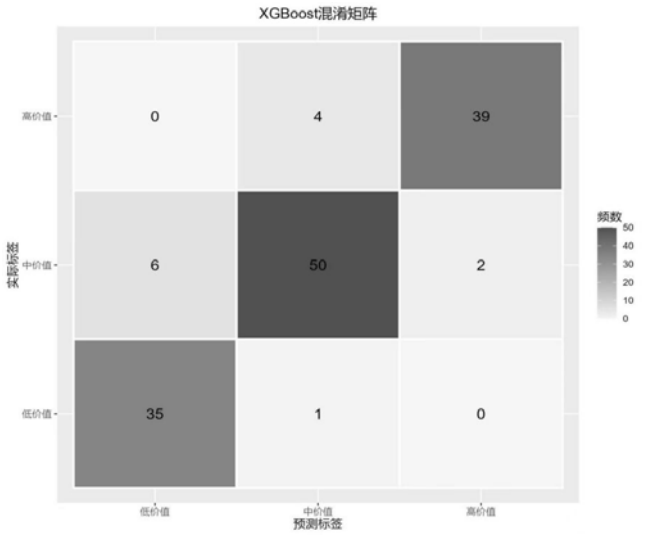

图1 XGBoost混淆矩阵图

将综合财务得分F样本划分为高价值(F≥70%分位数)、中价值(30%分位数≤F<70%分位数)、低价值(F<30%分位数)；然后选用多分类逻辑回归、XGBoost、随机森林三类算法,数据按7:3比例划分训练集与测试集,交叉验证优化模型超参数,以准确

率、宏F1分数、交叉验证误差为核心评估指标,最后根据评价指标结果对比三种机器学习模型预测性能。

(1)XGBoost结果。XGBoost模型对高中低三类价值的预测结果如图1：图中颜色越深就代表频数越高,整体来看模型误判情况较少,分类表现较好。

(2)随机森林。图2显示了通过随机森林得到的重要性特征,在对医药上市公司投资价值分类中,营收规模与成长和综合盈利能力是最重要的因子,对模型决策影响最大。

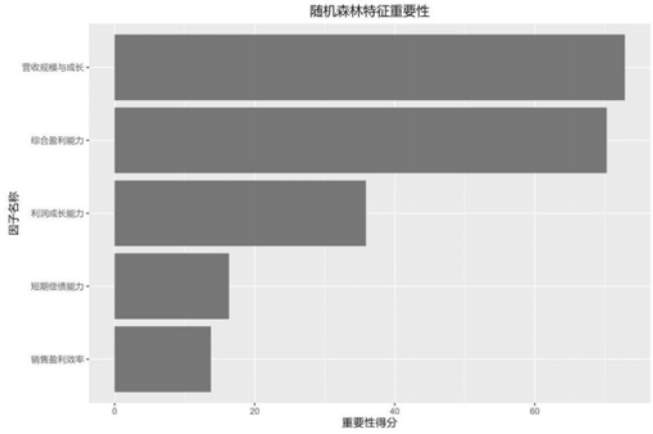

图2　随机森林特征重要性图

(3)三种模型结果对比。图3显示逻辑回归的测试准确率为0.985、宏F1分数为0.986,两个指标均为最高,且交叉验证误差为最低。该研究中“因子分析+逻辑回归”模型的预测结果最优。

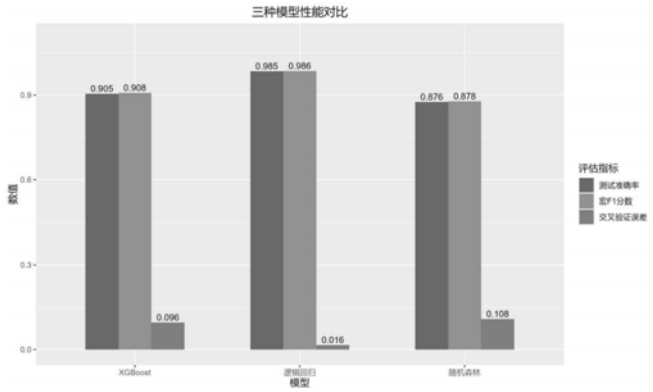

图3　模型性能对比图

4 结论

本研究利用我国461家生物医药上市公司财务数据,针对综合评价与投资价值预测问题,运用因子分析与多分类机器学习算法展开实证分析。结果显示：综合盈利能力、短期偿债能力、利润成长能力因子、销售盈利效率因子、营收规模与成长5个公因子中,营收规模与成长和综合盈利能力为影响价值分类的主要因子。三种机器学习算法中,逻辑回归综合表现最优,测试准确率达到0.985,该研究中“因子分析+逻辑回归”模型的预测效果最优。

[参考文献]

[1]Chen J,Xu X,Xu X,et al.Prediction of preeclampsia using maternal circulating mRNAs in early pregnancy[J].Archives of Gynecology and Obstetrics,2024:1-9.

[2]李占山,刘兆赓.基于XGBoost的特征选择算法[J].通信学报,2019,40(10):101-108.

[3]吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(03):37-41.

[4]Jiaqiang L,Lili D,Hongyong L,et al.Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning[J].Bioresource technology,2023,393130008-130008.

[5]简典.基于因子分析以及集成机器学习的岩质边坡稳定性预测模型[D].重庆交通大学,2024.

[6]刘照德,詹秋泉,田国梁.因子分析综合评价研究综述[J].统计与决策,2019,35(19):68-73.

作者简介：

何群(1999--),女,汉族,贵州毕节人,硕士研究生,研究方向：生物医学统计。