Advances in Applied Mathematics 应用数学进展, 2023, 12(4), 1772-1784
Published Online April 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.124184

Hans汉斯

国产电影票房影响因素分析及票房预测

栾瑶瑶，洪晓晴，潘　珈，李　敏*

青岛大学数学与统计学院，山东　青岛

收稿日期：2023年3月24日；录用日期：2023年4月18日；发布日期：2023年4月28日

摘　要

2021年我国发布的《“十四五”文化产业发展规划》中指出，要推进文化产业创新发展，而电影作为文化传播的重要载体，对文化产业的影响不容忽视。随着人们越来越重视精神层面的追求，电影作为大众的主要娱乐方式之一，其市场规模逐渐扩大，竞争也日趋激烈。如何使得票房利润最大化是一个非常值得研究的问题。本文将研究影响票房的因素并构建合理的票房预测模型。首先爬取电影网站2012~2021年票房为一千万以上的国产电影为研究样本，共805部。根据电影类型、演员影响力、导演影响力、上映档期、电影时长、电影评分、总场次、首周票房、平均票价、场均人次等多个变量，分别建立Lasso、随机森林、BP神经网络三种票房预测模型，并筛选出对总票房有显著影响的因素。通过评价指标进行比较得出基于BP神经网络得到的模型可以较好地预测电影票房。同时得到电影票房的影响因素错综复杂，其中上映档期、评分、电影类型都起到了重要作用。

关键词

电影票房，Lasso，随机森林，神经网络，影响因素

Analysis of Influential Factors and Box Office Prediction of Domestic Films

Yaoyao Luan, Xiaoqing Hong, Jia Pan, Min Li*

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Mar. 24th, 2023; accepted: Apr. 18th, 2023; published: Apr. 28th, 2023

Abstract

In the “14th Five-Year Plan for the Development of Cultural Industries” released in 2021, it was pointed out that the innovative development of cultural industries should be promoted, and the

*通讯作者。

文章引用：栾瑶瑶，洪晓晴，潘珈，李敏. 国产电影票房影响因素分析及票房预测[J]. 应用数学进展，2023，12(4): 1772-1784. DOI: 10.12677/aam.2023.124184

influence of movies, as an important carrier of cultural communication, on cultural industries cannot be ignored. As people increasingly value the pursuit of spirituality, movies, as one of the main forms of entertainment for the masses, are gradually expanding their market, and the competition is becoming more intense. How to maximize box office profits is an issue worth studying. In this paper, we will study the factors that influence the box office and construct a reasonable box office prediction model. Firstly, we crawl the movie websites to find domestic movies with a box office of more than 10 million from 2012 to 2021, which are 805 movies in total. Three box office prediction models, Lasso, Random Forest and BP Neural Network, were built based on several variables, such as movie genre, actor influence, director influence, release schedule, movie duration, movie rating, total number of scenes, first week box office, average ticket price and average attendance, and the factors that have significant influence on total box office were screened out. By comparing the evaluation metrics, the model obtained based on the BP Neural Network can better predict the box office. It is also obtained that the influencing factors of movie box office are intricate, among which the release schedule, rating, and movie genre all play an important role.

Keywords

Movie Box Office, Lasso, Random Forest, Neural Network, Influence Factor

1. 引言

经济社会的飞速发展和人民生活水平的日益增长，人们对生活方式的追求已不仅仅是物质层面的需要，而是对精神世界的追求与向往。电影作为其中一种精神文化娱乐产品，深受大众的倾心。近几年，我国影视市场的行业发展势头尤为强劲。根据中国统计年鉴[1]和国家新闻出版广电总局电影局的数据，2010 年全年电影票房达到了 101.72 亿元，首次突破百亿元大关；2021 年全年国产票房达到了 393.35 亿元。图 1 为我国 2012~2021 年十年间电影的年度总票房走势。忽略 2020 年电影行业受到疫情的影响，可以认为我国电影票房整体呈上升趋势。

本文研究旨在研究出影响电影票房的主要因素，构造良好的电影预测模型。可以帮助电影出版方合理改善电影制作的投资规划；同时可为影院的放映策略提供科学的依据；此外给电影投资者提供更合理的收益参考。另外研究电影票房并提高电影行业的经济效益，对我国经济发展也有着重要影响。同样电影作为文化交流的场所，更是承担着文化输出这一重要角色。因此本文的研究对于我国电影行业的健康发展有着重要的现实意义。

目前国内外许多学者对电影票房的影响因素和预测展开了深入研究。例如，2018 年 Michael Siering 等[2]研究了网络舆论的影响，发现网络评论对于影片票房具有很大影响。电影票房的发展趋势具有一定特殊性，因此许多学者考虑建立时间序列模型预测电影票房，2019 年田源[3]通过 ARIMA 模型预测中国电影票房，效果较好。杨朝强等[4]尝试建立长短期记忆网络模型(LSTM)进行预测，预测结果相对比较准确。

由于本文选取的电影票房数据量大且影响因素复杂多样，机器学习在处理此类问题时通常会表现更好地预测效果。有学者考虑通过机器学习算法进行电影票房的预测与分析。2021 年宋玉萍等人[5]对 2015~2019 年国庆档上映的 558 部电影的首周票房进行分析，选取 17 个变量，建立线性回归和随机森林

回归模型对首周票房展开预测。同年李旺泽等人[6]以 2011~2018 年间票房过亿的国产电影展开研究，选取演员、导演、评分、首日票房等变量，建立随机森林预测模型，准确度达 85%。综上，本文通过爬虫技术获取相关数据，通过 python 软件对电影票房的影响因素进行可视化分析，并建立神经网络等模型进行票房预测，进一步完善电影票房预测机制。

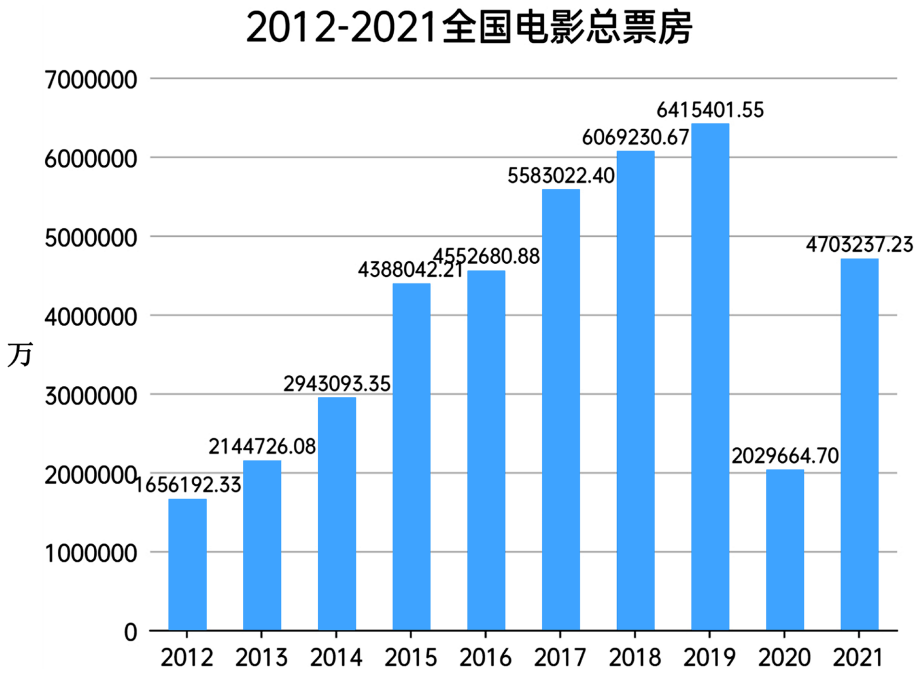

Figure 1. Total box office of domestic films from 2012 to 2021
图 1. 2012-2021 年度国产电影总票房

2. 数据处理与分析

2.1. 数据来源与处理

2.1.1. 数据来源

本文选取了 2012 年至 2021 年全国公开上映的电影为研究对象，选取票房 1000 万以上的国产电影展开研究。影响电影票房的因素有很多，结合相关文献，本文考虑将以下因素作为样本特征：电影时长，电影类型、上映日期、导演、演员、电影评分、总票房、总场次、首周票房、平均票价、场均人次[7]。通过网络爬虫技术、八爪鱼软件进行网页原始数据的获取，结合艺恩网、猫眼票房、Mtime 时光网、豆瓣电影和微博的数据综合得到最终数据。

2.1.2. 数据转换与定义变量

1) 上映档期

为分析电影上映日期对电影票房的影响，本文将电影上映日期划分为以下几类：春节档(除夕至一月初七)，暑期档(7.21 至 8.20)，国庆档(9.30 至 10.7)，七夕档(7.7 至 7.9)，元旦档(12.31 至 1.3)，五一档(4.30 至 5.3)，情人档(2.14)。将上映档期设为虚拟变量，如果该电影上映日期属于上述几个上映档期，则设为 1，否则设为 0。

2) 电影时长

电影时长也会对电影的票房产生一定的影响。本文将该变量处理为名义变量，分为以下三档：电影

时长在 90 分钟以下的为第一类，90~120 分钟为第二类，120 分钟以上为第三类，对其进行哑变量处理。

3) 电影类型

电影类型是基于电影叙事元素和情感反应进行相似分类的电影类别。电影的类型也决定了观看电影的人群，进而影响电影的票房。本文根据时光网、猫眼电影、豆瓣电影等多个网站上的共 34 个电影类型，进一步处理。发现“歌舞”、“戏曲”这两类电影也都属于“音乐类”，故将这两类都归属于“音乐类”。综上所述，共总结出 20 个电影类型。将每一类都设为哑变量，即某个电影属于其中一类，则赋值为 1，否则赋值为0。例如，2021 年上映的电影《你好，李焕英》，电影类型为剧情和喜剧，则赋值为(1,1,0,0,0,...,0)。

4) 导演影响力

导演对于一部电影是否成功有着重要的影响。中国电影史上涌现出许多著名的导演，例如陈凯歌、张艺谋、李少红等，他们执导拍摄的电影大多票房都比较高，这就叫做导演影响力。因为许多网站的主观思想或者评价标准的不同，所以从网页上获取的著名导演的数据可能不准确。故本文衡量导演影响力的标准如下：首先，获取该导演获奖或提名的次数，其次，获奖一次记为 1，提名一次记为 0.5，计算每个导演的获奖分数；将导演的获奖分数作为导演影响力变量的值，部分结果如表 1 所示。

Table 1. Director influence variable value
表 1. 导演影响力变量值

电影名称	导演	导演获奖	导演提名	获奖分数
长津湖	陈凯歌	8	14	15
战狼 2	吴京	1	8	5
流浪地球	郭帆	0	1	0.5
红海行动	林超贤	3	13	9.5
美人鱼	周星驰	6	28	20

5) 主演影响力

原始数据获得每部电影的两位主要演员，为了使该变量可以预测电影票房，对该变量进行数值化，如何衡量一个演员的影响力是该变量的关键。本文参考了演员的微博粉丝数以及该电影上映前的演员百度指数等指标作为衡量演员影响力的标准。综合考虑演员多方面因素将主演影响力分成 0~7 档，数值越高说明该部电影的主演影响力越大。例如：电影《长津湖》的两个主演，吴京的微博粉丝数为：1443.9 万；易烊千玺的微博粉丝数为：8961 万，两者的百度搜索指数高达日均 7952、17,436，故将这部电影的主演影响力划分为 7 档，记为 7。

6) 总场次、首周票房、平均票价、场次人均，作为连续型变量进行分析，不进行特别处理。

综上所述，本文包括 29 个自变量，总票房作为因变量。

2.2. 描述性统计分析

本文的数据丰富多样，为尽可能了解各变量与电影票房之间的关系，本小节将对个别因素进行描述性统计分析，对数据进行一个可视化展示。

2.2.1. 上映档期分析

近年来，越来越多的电影选择在春节、跨年、七夕等节假期上映。为了更加直观形象地说明各个上映档期与票房之间的关系，将各个上映档期的票房进行均值处理，绘制各上映档期平均票房的条形图。从图 2 可以看出，春节档的平均票房最高，其次是国庆档、元旦档、暑期档。这表明电影的上映档期和

总票房之间存在一定的关系。

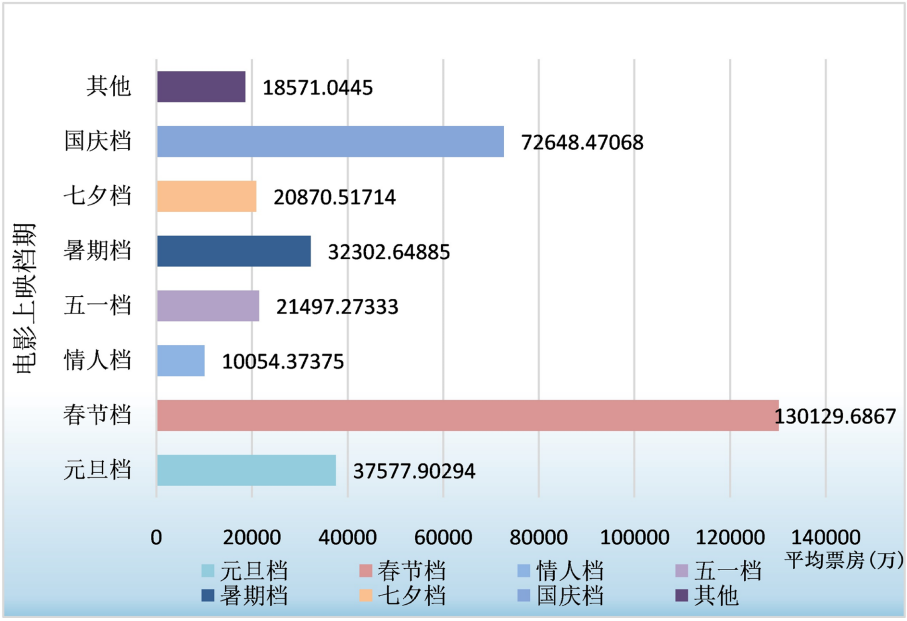

Figure 2. Bar chart of average box office by release period
图 2. 各上映档期平均票房条形图

2.2.2. 导演和票房

图 3 为导演是否获奖或提名与对数票房的关系图。本文选择时光网中第一顺位导演的获奖和提名数量化后的总和作为衡量影响票房的指标。从下图可以看出，得分高的导演，其电影的平均票房也更高，但这也并不是绝对的。因为老一代的著名导演，其获奖次数可能比较多，而现在观看电影的主流人群是年轻人，他们更喜欢看一些话题、观点更加新颖的电影，这可能导致获奖次数很多的导演近几年的票房并没有很高。且近几年涌现了一大批新生导演。这些年轻导演的获奖次数较少，电影却很受观众的喜爱。

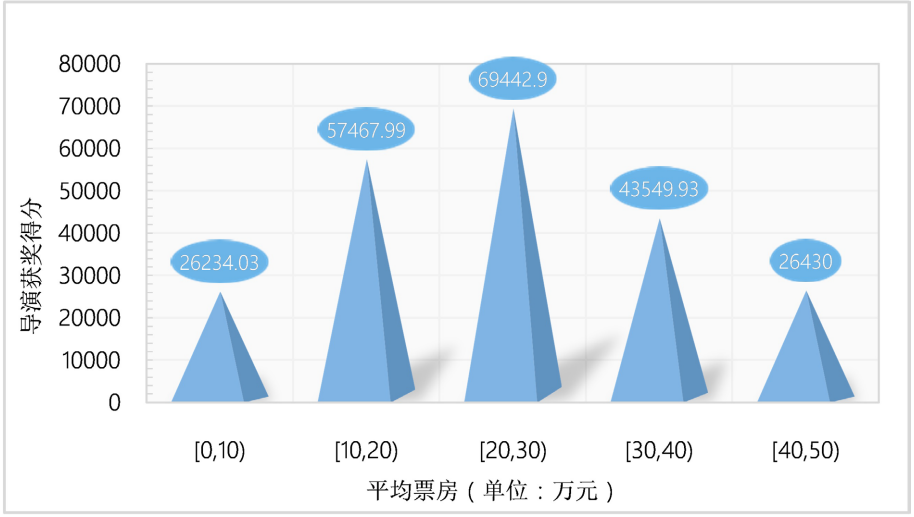

Figure 3. Chart of director's score and box office
图 3. 导演得分与票房关系图

2.2.3. 电影类型

电影类型可以直接反映一部电影的内容与情感。观影人会选择自己感兴趣的电影类型进行观看。随着中国电影市场的不断发展，近年来涌现了不同种类的影片，且题材、内容都更加新颖。采用词云图来展示，通过图 4 可以看出，科幻、戏剧、青春、灾难、战争、动作等类型的电影票房较高，受观众的喜爱度较高。

Figure 4. Film genre word cloud map
图 4. 电影类型词云图

3. 模型介绍

本文采用 Lasso、随机森林、BP 神经网络三种模型对数据进行建模分析。本章将对这三种模型进行简要介绍。

3.1. Lasso 模型

Lasso 回归方法是一种连续性收缩可以用于变量选择的高效方法，采用 L1 正则化思想，使得在满足回归系数的绝对值之和小于某个大于 0 的常数的条件下，最小化残差平方和。

$$\hat{\beta} = \arg\min_{\beta}\left\{\left\|y - X\beta\right\|^2 + \lambda\left|\beta\right|_1\right\} \tag{1}$$

Lasso 可以有选择的把变量放入模型从而得到更好的性能参数[8]。该方法通过参数 λ 调节模型的复杂度，λ 越大对变量多的线性回归模型的惩罚力度就随之越大，经过多次调试最终获得一个变量较少的模型，避免过拟合的发生[9]。

3.2. 随机森林模型

随机森林是一种由决策树构成的集成算法，属于集成学习中的 Bagging 方法[10]。它是使用了 CART 决策树作为弱学习器的集成方法。其中，Gini 指数作为因素的重要性评价指标，Gini 指数计算公式如下：假设一棵树中的每个节点 k，则 $G_k = 2\hat{p}_k\left(1-\hat{p}_k\right)$，其中 $\hat{p}_k$ 表示样本在节点 k 属于任何一类的概率值。节点的重要性程度由节点分裂前后 Gini 指数的变化程度来衡量：

$$I_{\Delta k} = G_k - G_{k1} - G_{k2} \tag{2}$$

其中 G_{k1} 和 G_{k2} 分别表示节点 k 产生的子节点，对每棵树进行迭代递归，最后采用随机抽样的方式，选定样本和变量，最终产生森林，该森林包含 T 棵树，定义变量 X_i 在第 t 棵树中出现次数为 N，则其重要性有下式表示：

$$I_{it} = \frac{1}{n}\sum_{t=1}^{T}\sum_{j=1}^{N} I_{\Delta j} \tag{3}$$

3.3. BP 神经网络模型

BP 神经网络是一种简单的人工神经网络。它是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一，能学习和存贮大量的输入－输出模式映射关系[11]。

一个最简单的三层 BP 如图 5 所示。

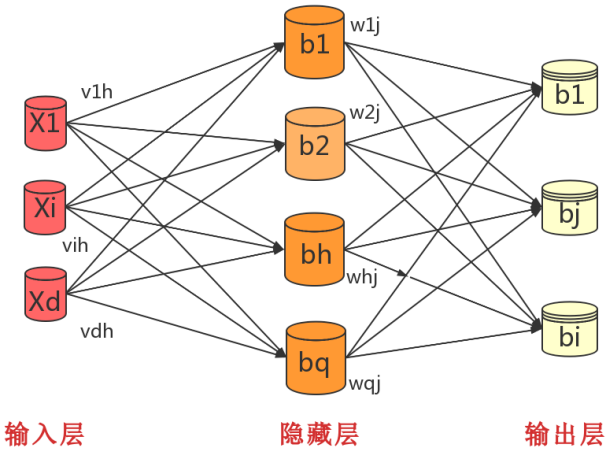

Figure 5. Three-layer BP neural network diagram
图 5. 三层 BP 神经网络图

其中，第 q 个隐藏神经元的输入：$\alpha_h = \sum_{i=1}^{d} V_{ih} * X_I$，第 j 个输出神经元的输入：$\beta_j = \sum_{h=1}^{q} W_{hj} * b_h$。

具体流程为：

第一步，网络初始化。

给各连接权值分别赋一个区间$(-1,1)$内的随机数，设定误差函数 e，给定计算精度值 ε 和最大学习次数 M。

第二步，随机选取第 k 个输入样本及对应期望输出。

$$x(k) = (x_1(k), x_2(k), \cdots, x_n(k)) \tag{4}$$

$$d_o(k) = (d_1(k), d_2(k), \cdots, d_q(k)) \tag{5}$$

第三步，计算隐含层各神经元的输入和输出。

$$hi_k(k) = \sum_{i=1}^{n} w_{ih} x_i(k) - b_h \tag{6}$$

$$ho_h(k) = f(hi_h(k)) \tag{7}$$

$$ho_h(k) = \sum_{h=1}^{p} w_{ho} ho_h(k) - b_0 \tag{8}$$

$$yo_o(k) = f(yi_o(k)) \tag{9}$$

第四步，利用网络期望输出和实际输出，计算误差函数对输出层的各神经元的偏导数 $\delta_o(k)$。

$$\frac{\partial e}{\partial w_{ho}} = \frac{\partial e}{\partial yi_o} \frac{\partial yi_o}{\partial w_{ho}} = -(d_o(k) - yo_o(k)) f(yi_o(k)) - \delta_o(k) \tag{10}$$

$$\frac{\partial e}{\partial hi_h(k)}=\frac{\partial\left(\frac{1}{2}\sum_{o=1}^{q}\left(d_o(k)-yo_o(k)\right)^2\right)}{\partial ho_h(k)}\frac{\partial ho_h(k)}{\partial hi_h(k)}=\left(\sum_{o=1}^{q}\delta_o(k)w_{ho}\right)f'\left(yi_o(k)\right)-\delta_h(k) \tag{11}$$

第五步，利用输出层各神经元的 $\delta_o(k)$ 和隐含层各神经元的输出来修正连接权值 $w_{ho}(k)$。

$$\Delta w_{ho}(k)=-\mu\frac{\partial e}{\partial w_{ho}}=\mu\delta_o(k)ho_h(k) \tag{12}$$

$$w_{ho}^{N+1}=w_{ho}^{N}+\eta_o(k)ho_h(k) \tag{13}$$

第六步，利用隐含层各神经元的 $\delta_h(k)$ 和输入层各神经元的输入修正连接权。

$$\Delta w_{ih}(k)=-\mu\frac{\partial e}{\partial w_{ih}}=-\mu\frac{\partial e}{\partial hi_h(k)}\frac{\partial hi_h(k)}{\partial w_{ih}} \tag{14}$$

$$w_{ih}^{N+1}=w_{ih}^{N}+\eta\delta_h(k)x_i(k) \tag{15}$$

第七步：计算全局误差。

$$E=\frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{q}\left(d_o(k)-yo_o(k)\right)^2 \tag{16}$$

第八步：判断模型合理性。

当误差达到预设精度或学习次数大于设定的最大次数，则结束算法。否则，选取下一个学习样本及对应的期望输出，返回到第三步，进入下一轮学习。

4. 结果分析

本章对模型建立的过程中，都先进行了划分训练集与测试集的操作。根据票房随机将 80%样本划分为训练集，剩余 20%样本为测试集。模型的建立都基于训练集进行，剩余的测试集用来检验并且评估该模型预测的可行性。

4.1. 影响因素分析

影响因素分析是分析问题的重要手段，可以清晰指出与问题相关的变量的重要程度。得到的特征重要性值越大，表明该特征对预测结果的影响越大。反之，重要性越小，表明该特征对预测的影响越小。

4.1.1. 基于 Lasso 的影响因素分析

本文利用 Lasso 对变量进行筛选，选择出与票房相关的显著变量，剔除系数为零的变量。结果显示：首周票房对整体票房有正相关作用，可见若一部电影首周票房高的话会促进后续几周的票房走势。电影的类型也会对总票房起到很大的作用，例如电影类型为战争类的回归系数为 0.28079，喜剧类、动作类、恐怖类为正相关。上映档期也对总票房产生至关重要的影响。

4.1.2. 基于随机森林的影响因素分析

建立随机森林模型得到预测变量的相对重要性的综合得分，并绘制特征重要性排序图。如图 6 所示，各个特征按重要程度排序，前十名依次为：首周票房，场次，电影评分，平均票价，主演影响力，电影时长，动作类型，喜剧类型，犯罪类型，导演的影响力得分。显然，首周票房对于票房的影响力远远大于其他特征，对票房的影响最大，这说明首周票房的高低可以从侧面反映出该部电影的未来票房走势。这也提示各大影院可以根据电影的首周票房制定合理的后续排片策略。在所有电影类型中，动作类型重要程度最高，其次是喜剧类型，这两种类型在观影者中是最受欢迎的。

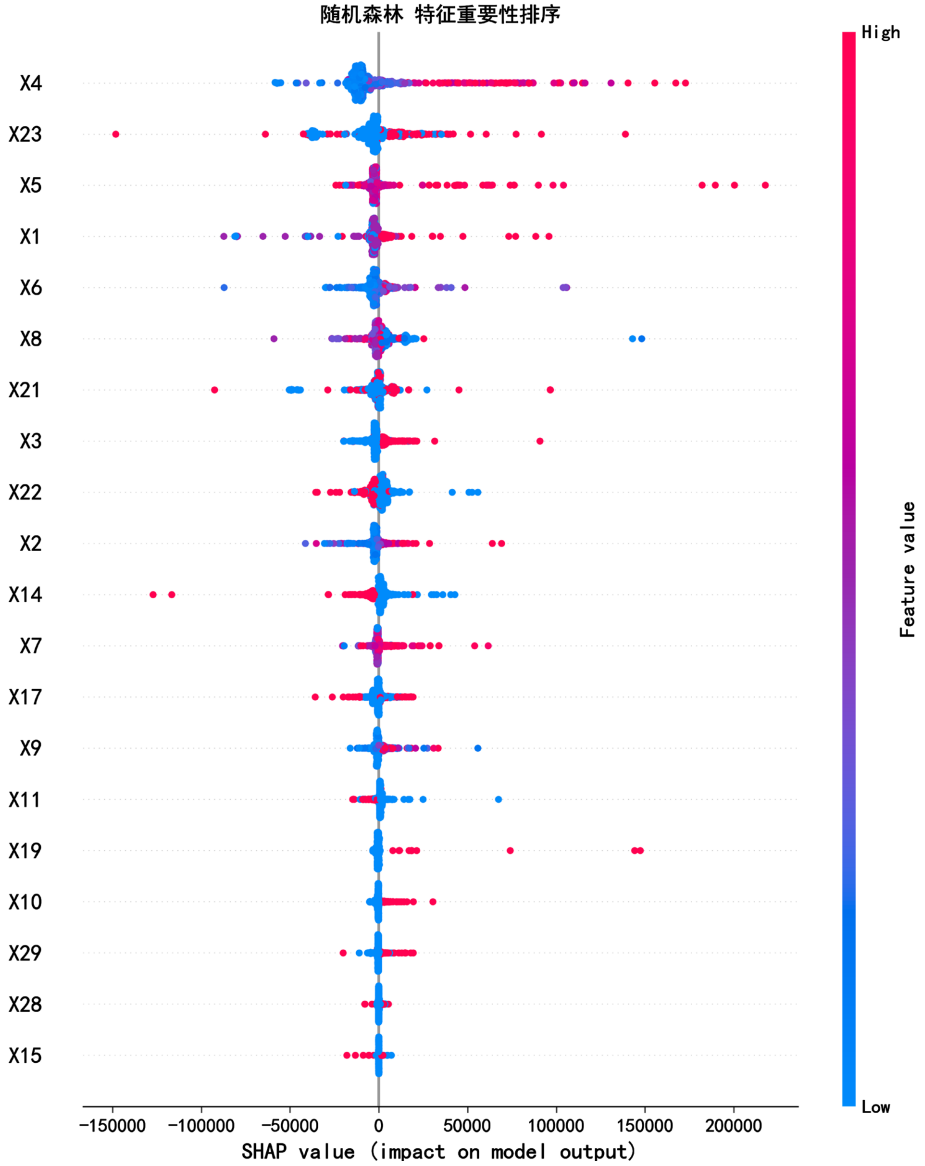

Figure 6. Random forest feature importance map
图 6. 随机森林特征重要性图

4.1.3. 影响因素小结

根据 Lasso 以及随机森林模型对原始变量进行重要性分析，可以得出不同因素对整体模型的效果产生不同的影响。有的因素在模型中起着重要作用，比如首周票房、电影评分；反之有的因素对模型的影响非常微小，比如电影类型中的奇幻类、运动类等。根据影响因素的重要性程度分析可以为电影市场提供一些可以对的建议。动作类、喜剧类等电影题材比较受大众亲睐；在考虑电影选角的时候，可以考虑一些流量明星，发挥明星的粉丝效应；在选电影上映时间的时候可以多考虑节假日等等。

本文在后续的 BP 神经网络建模中，根据 Lasso 和随机森林模型的变量重要性分析，选择排名前 20 变量进行建模分析，进一步提高模型的准确性。选择的 20 个变量为：首周票房，场次，电影评分，平均票价，主演影响力，电影时长，动作类型，喜剧类型，犯罪类型，导演的影响力得分，爱情类型，剧情类型，上映档期，历史类型，战争类型，冒险类型，古装类型，恐怖类型，灾难类型，科幻类型。

4.2. 预测结果分析

4.2.1. 基于 Lasso 建模

首先根据交叉验证方法选择最优的 λ 值。绘制 Lasso 回归拟合 MSE 图，图中有两条虚线，一个是均方误差最小时的 λ 值，一个是距离均方误差最小时的 λ 值，分别为 0.1590 和 0.2306。本文选择使用均方误差最小时的 λ 值进行预测。使用 cv.glmnet 函数进行分析，得到部分预测结果如表 2 所示。根据预测结果与真实结果比较，可以看出 Lasso 回归模型效果一般。相对误差超过 30%。下面将探索新的模型来更好的对票房进行预测，提高模型的准确度。

Table 2. Lasso forecast results (partial)
表 2. Lasso 预测结果(部分)

电影名称	真实值	预测值	相对误差
夏洛特烦恼	144,782.38	93,985.65	−35.08%
流浪地球	468,682.3	324,914.63	−30.67%
湄公河行动	118,859.38	86,761.99	−27.00%
妖猫传	53,003.07	71,193.70	34.32%
拆弹专家	40,049.45	57,832.48	44.40%

4.2.2. 基于随机森林建模

参数的取值会影响随机森林模型的预测效果，因此本文对随机森林模型的参数进行调优选择。参数 ntree 为树的数量，该值过小会导致模型错误率增加，数量过多也会导致模型复杂化，所以在保证效能的情况下我们尽量减少决策树的数量，减少运行时间。从图 7 中可以看出，在树的数量大于 80 的时候错误率基本趋于稳定，因此可以认为树的数量在 80 时的错误率水平最低，且在其前后的错误率水平基本一致。本文取参数 ntree = 80，此时错误率达到动态最低。

基于上述参数取值，使用 R 软件 caret 包的 predict 函数对测试集的票房数据进行预测。预测结果如表 3 所示。

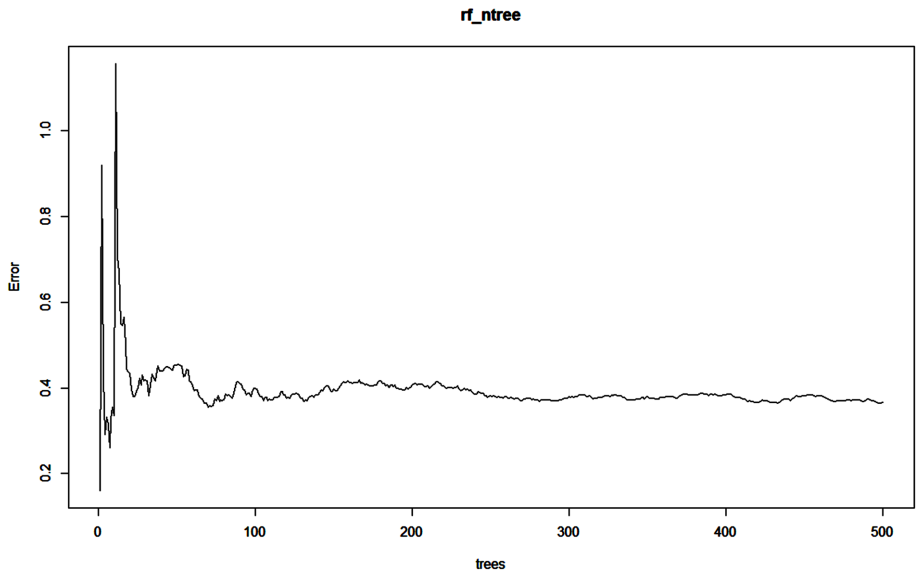

Figure 7. Random forest parameter selection graph
图 7. 随机森林参数选择图

Table 3. Random forest prediction results (partial)
表 3. 随机森林预测结果(部分)

电影名称	真实值	预测值	相对误差
爱情公寓	55,542.65	62,855.24	13.17%
大闹天竺	75,792.68	63,116.50	−16.72%
分手大师	66,544.06	52,417.28	−21.23%
宫锁沉香	5447.74	6496.79	19.26%
七月与安生	16,695.17	19,407.28	16.24%

从上述结果中可以看出随机森林模型效果比 Lasso 模型有较大提升。相对误差基本不超过 20%。下面将探索新的模型来更好的对票房进行预测，提高模型的准确度。

4.2.3. 基于 BP 神经网络建模

首先对模型进行参数选择，确定隐藏层节点个数。隐含层节点数的选择很重要，如果数目太少，网络将不能建立复杂的判断界，导致容错性差；若数目过大，导致模型训练时间过长，同时网络的泛化能力降低。利用公式确定隐藏层节点个数：

$$h = \sqrt{m+n} + a \tag{17}$$

h 为隐含层节点的数目，m 和 n 分别是输入层和输出层节点的数目，a 为 1~10 之间的调节常数。

本文根据上述经验公式，大概确定节点数为 7~15；依次选 7~15 的隐神经元训练，通过不断调整节点数，选取最优的模型。本文选用均方根误差为判断最优模型的标准，经过多次模型模拟，确定本模型的最优隐藏节点数为 8。其次需要确定学习速率。学习速率一般选取为 0.01~0.8，学习速率过大可能导致

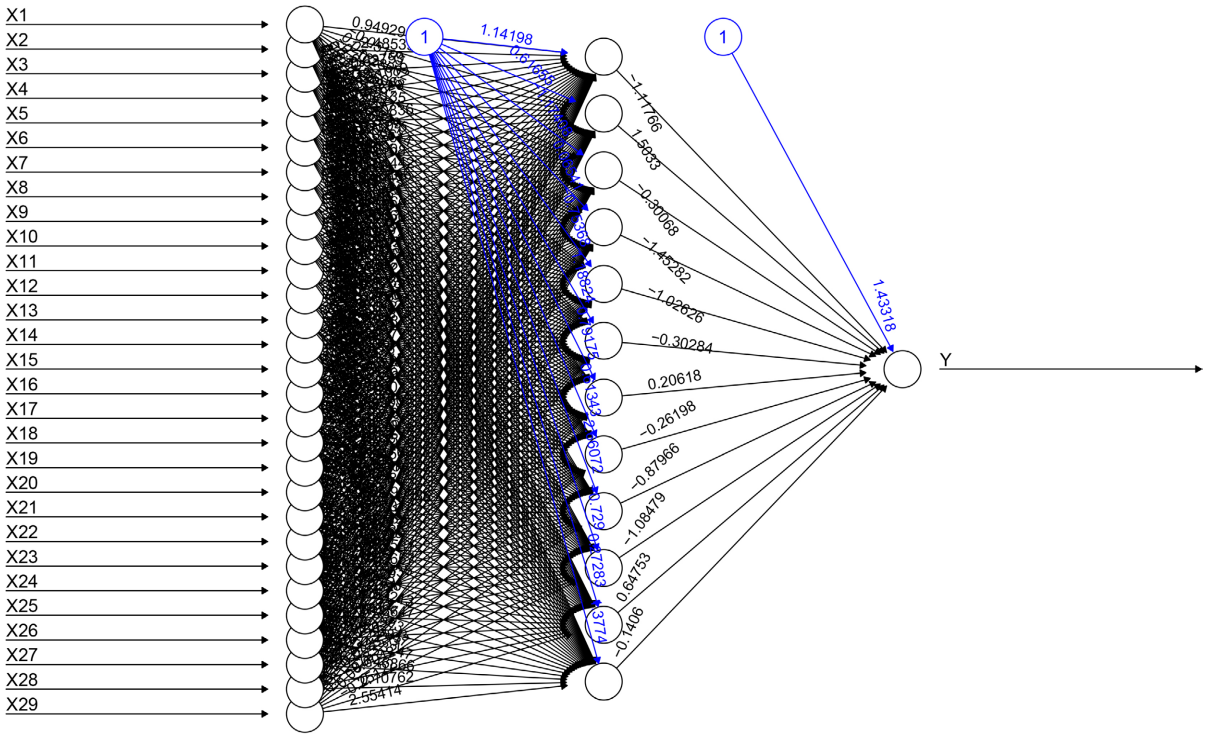

Figure 8. Three-layer BP neural network diagram
图 8. 三层 BP 神经网络图

系统不稳定，学习速率过小导致收敛太慢。本文模型选取的学习速率为 0.01。

确定上述参数后，模型模拟得到如下结果，根据 R 软件计算出真实值与预测值的相关性，得到模型预测的准确度达到 86.7%。神经网络建模图如图 8 所示。

表 4 为本模型预测的部分电影票房与真实票房的对比，可以看出 BP 神经网络模型的预测值与真实值的相对误差较小，测试集中各电影的票房预测偏差大部分小于 15%，模型预测较为准确。

Table 4. Neural network prediction results (partial)
表 4. 神经网络预测结果(partial)

电影名称	真实值	预测值	相对误差
奇门遁甲	29,964.44	33,902.33	13.14%
大闹天竺	75,792.68	68,581.27	−9.51%
归来	29,147.01	24,629.49	−15.50%
七月与安生	16,695.17	14,234.55	−14.74%
妖猫传	53,003.07	50,096.98	−5.48%

4.2.4. 预测结果比较

这一章我们采用了三个模型进行建模比较分析与预测。比较三个模型的均方根误差可知，Lasso 回归模型预测效果相比最差、随机森林其次、神经网络的预测效果最佳。我们选取了 10 部电影进行比较。表 5 我们将展示电影在不同模型下的预测效果比较。

Table 5. Comparison of results of different models
表 5. 不同模型结果比较

电影名称	实际票房	Lasso 预测值	随机森林预测值	BP 神经网络预测值	Lasso 相对误差	随机森林相对误差	BP 神经网络相对误差
冰封：重生之门	14,269.53	2963.86	13,231.38	14,231.38	53.92%	−0.27%	−0.27%
归来	29,147.01	52,026.17	26,355.59	25,819.9	78.5%	9.58%	−11.41%
七月与安生	16,695.17	25,191.92	19,407.28	15,400.63	50.89%	16.24%	−7.75%
妖猫传	53,003.07	71,193.7	51,100.69	50,096.98	34.32%	3.59%	−5.48%
奇门遁甲	29,964.44	38,218.43	35,824.67	33,902.33	27.55%	19.56%	13.14%
流浪地球	468,682.3	324,914.6	346,688.29	420,367.59	−30.67%	−26.03%	10.31%
小猫巴克里	8224.43	1925.625	2142.52	9680.46	57.31%	75.03%	15.04%
神探驾到	3740.71	2740.155	4292.92	3966.30	−26.75%	14.76%	6.03%
素人特工	2248.15	3328.643	3323.07	2987.12	48.00%	47.81%	24.74%
拆弹专家	40,049.45	57,832.48	44,271.35	46,915.2	44.40%	10.54%	17.14%

通过上表我们可以直观的看出结果，基于 BP 神经网络建模的模型对票房的预测有较好的效果，相对误差基本小于 15%；基于随机森林建模的模型对票房的预测效果一般，相对误差在 15%~30%；基于 Lasso 回归建模的模型对票房的预测效果最差，相对误差基本高于 40%。

5. 结论与展望

本文用三个模型对 2012~2021 年十年间国产票房超过 1000 万以上的 805 部电影进行处理分析。基于

结果分析与评估，BP 神经网络建模的模型对票房的预测有较好的效果。影响电影票房高低的因素非常复杂，我们选取的 29 个因素中，上映档期、评分、首周票房、电影类型为战争类、喜剧类等因素对票房起重要作用。电影出版方在上映电影时可以选择节假日票房上映档期来提高收益。电影制作商等可以选择知名度较高的演员来参演，利用主演的影响力来提高收益；在电影选材方面可以选择大众喜欢的类型来制作，例如喜剧片、战争片等等。对于即将上演的电影，影院可以通过相关的模型进行票房预测来安排排片的情况，在上映一周后可以加入首周票房这一因素更新票房预测来重新安排排片情况以获得利益最大化。当然本文的研究还存在很多不足，如下所示：

1) 在考虑导演影响力时，以导演的获奖次数和提名次数作为衡量导演影响力的标准，但近年来，国产电影市场出现了一大批优秀的年轻导演，他们可能没有获奖，但拍出来的电影也很受年轻人喜欢，所以仅以历年获奖次数作为衡量标准并不严谨。

2) 考虑的变量个数不足，影响电影票房的因素还有很多，本文只是选取了可以搜集到数据的一部分，电影的制作、宣传、投资成本都能影响电影票房。

3) 本文建立的模型也存在不足，Lasso 模型可以比较好地解释变量，但在预测方面效果不好，而随机森林和 BP 神经网络也存在缺点，都需要进一步优化，选择更优模型进行预测。

基金项目

山东省自然科学基金青年项目(ZR2021QA053)。

参考文献

[1] 国家统计局. 中国统计年鉴 2019 [M]. 北京: 中国统计出版社, 2019.

[2] Siering, M., Muntermann, J., Rajagopalan, B., *et al*. (2018) Explaining and Predicting Online Review Helpfulness: The Role of Content and Reviewer-Related Signals. *Decision Support Systems*, **108**, 1-12. https://doi.org/10.1016/j.dss.2018.01.004

[3] 田源. 中国电影票房季节性分析和预测——基于季节趋势模型和季节 ARIMA 模型[J]. 现代商业, 2019, 538(21): 47-50.

[4] 杨朝强, 蒋卫丽, 邵党国. 基于 LSTM 模型的电影票房预测算法[J]. 数据通信, 2019, 192(5): 34-37.

[5] 宋玉萍, 朱家明, 杨琴. 基于随机森林回归模型的国产电影首周票房预测分析[J]. 高师理科学刊, 2021, 41(1): 1-26.

[6] 李旺泽, 郑列. 基于随机森林回归模型的国产电影票房预测[J]. 湖北工业大学学报, 2021, 35(1): 114-117.

[7] 李振兴. 机器学习在电影票房预测中的应用研究[D]: [硕士学位论文]. 西安: 西安石油大学, 2020.

[8] 孙昕. 基于 Lasso 方法的中国股市时滞性回归分析[D]: [硕士学位论文]. 大连: 大连理工大学, 2017.

[9] 崔凝凝, 唐嘉庚. 基于回归分析的中国电影票房影响因素研究[J]. 江苏商论, 2012, 334(8): 35-39.

[10] 张鑫, 郭振宇. 基于随机森林的影片票房预测[J]. 现代电影技术, 2016, 452(3): 13-17+37.

[11] 张溪源. 基于 BP 神经网络的电影票房预测研究[D]: [硕士学位论文]. 乌鲁木齐: 新疆财经大学, 2020.