Hans Journal of Data Mining 数据挖掘, 2021, 11(2), 11-20
Published Online April 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.112002

Hans汉斯

基于随机森林方法的国产电影票房研究

郭　萱，鲍　勇，于水源

北京科技大学天津学院，天津
Email: 1540473460@qq.com, 778547694@qq.com, 392996541@qq.com

收稿日期：2021年2月17日；录用日期：2021年3月22日；发布日期：2021年3月29日

摘　要

随着经济的发展和人民生活水平的提高，电影行业迅速发展。电影票房影响因素的研究及预测，对提高国产电影质量十分必要。目前学者们多采用神经网络方法对电影票房进行研究，神经网络方法未给出变量重要性排序，预测结果不够稳健。本文依据2014~2018年225部国产影片的相关数据，采用随机森林方法建立电影票房预测模型。得到了影响我国国产电影票房的因素主要有首周末票房、首映日票房、百度指数、豆瓣评分和点映票房。同时本文采用线性回归模型和神经网络模型建立电影票房的预测模型，应用三种方法对2019年12部国产电影票房进行预测。结果表明：随机森林在电影票房预测方面更加精确稳健，对《飞驰人生》、《银河补习班》等八部影片的预测误差在10%左右。神经网络和线性回归模型预测误差较大。

关键词

电影票房预测，影响因素，随机森林，神经网络，线性回归

Research on Box Office of Domestic Films Based on Random Forest Method

Xuan Guo, Yong Bao, Shuiyuan Yu

Tianjin College, University of Science and Technology Beijing, Tianjin
Email: 1540473460@qq.com, 778547694@qq.com, 392996541@qq.com

Received: Feb. 17th, 2021; accepted: Mar. 22nd, 2021; published: Mar. 29th, 2021

Abstract

With the development of economy and the improvement of people's living standard, the film industry develops rapidly. It is necessary to study and forecast the influencing factors of film box of-

文章引用: 郭萱, 鲍勇, 于水源. 基于随机森林方法的国产电影票房研究[J]. 数据挖掘, 2021, 11(2): 11-20.
DOI: 10.12677/hjdm.2021.112002

fice to improve the quality of domestic films. At present, most scholars use the neural network method to study the box office of films. The neural network method does not give the order of importance of variables, and the prediction results are not robust enough. Based on the relevant data of 225 domestic films from 2014 to 2018, this paper adopts the random forest method to establish the box office prediction model. The main factors that influence the box office of domestic films in China are the box office of the first weekend, the first day box office, baidu index, douban score and the advance screenings box office. At the same time, this paper adopts linear regression model and neural network model to establish the box office prediction model, and applies three methods to predict the box office of 12 domestic films in 2019. The results show that the random forest is more accurate and stable in the prediction of box office, and the prediction error of eight films such as "Pegasus" and "Looking Up" is around 10%. The prediction error of neural network and linear regression model is large.

Keywords

Box Office Forecast, Affecting Factors, Random Forests, Neural Network, Linear Regression

1. 引言

近年来中国电影业发展迅速，但国产电影票房收入在总票房中占比不大，国产电影想要在国际电影市场占有一席之地仍需更大的努力，分析影响电影票房的因素，预测电影票房，提高电影质量势在必行。2014 年杨威[1]使用新媒体微博数据作为研究对象，利用神经网络方法建立票房预测模型，并与支持向量机模型和线性回归模型进行预测精度对比，结果表明神经网络模型准确度高于其他模型。2017 年张雪[2]使用多元线性回归、BP 神经网络和卷积神经网络建立票房预测模型，结果表明[2]：多元回归模型预测效果精确度较低，BP 神经网络和卷积神经网络预测效果都比回归好。2018 年郭萱[3]针对 2014~2016 年 173 部国产电影采用随机森林方法进行电影票房影响因素分析，兼顾数值预测方法与分类预测方法提供合理的电影票房预测方案。2019 年鲁月[4]基于随机森林构建票房组合预测模型并与基于 BP 神经网络、k-均值[4]和局部 BP 神经网络的国产电影票房预测模型进行对比，结果表明基于随机森林因素筛选的国产电影票房组合模型在一定程度上提高了票房的预测精度。

随机森林方法提出至今，已经被广泛应用于机器学习、生物医学、生物信息学和数据挖掘等众多领域。该方法不仅可以进行分类和回归预测，同时可以给出变量重要性排序[5]。相比于参数模型中假设较多，参数估计数值不稳定的问题，随机森林方法可以更好地解决噪声问题以及数据中的异常值问题、能更好地对大规模数据进行处理[6]、具有良好的解释性及学习过程快速。本文采用随机森林方法对国产电影票房影响因素进行分析，并对 2019 年 12 部影片的票房进行预测。首先根据问题实际背景给出七个影响国产电影票房的因素，分别为：档期、是否有续集、首映日票房、点映票房、首周末票房、百度指数和豆瓣评分。基于 2014~2018 年 225 部影片的相关数据，采用随机森林方法建立回归模型，得到影响国产电影票房的主要因素并给出 2019 年 12 部影片电影票房的预测值和预测误差。同时采用电影票房领域应用较多的神经网络方法和线性回归方法对 2019 年 12 部影片进行预测。将随机森林预测结果与神经网络和线性回归模型预测结果进行对比分析。在变量选择方面随机森林具有一定的优势，在预测方面随机

森林方法比其他两种方法更为精确。

2. 随机森林方法介绍

随机森林是机器学习算法之一，由多个决策树分类器组合而成。随机森林[7]的基本思想是每次随机选取一些特征，独立建立树，重复这个过程，保证每次建立树时变量选取的可能性一致，如此建立许多彼此独立的树，最终的分类结果由产生的这些树共同决定。将分类树替换成回归树，把类别替换为每个回归树预测值的加权平均，就可以将随机森林树转换成随机森林回归算法。随机森林流程图如图 1 所示。

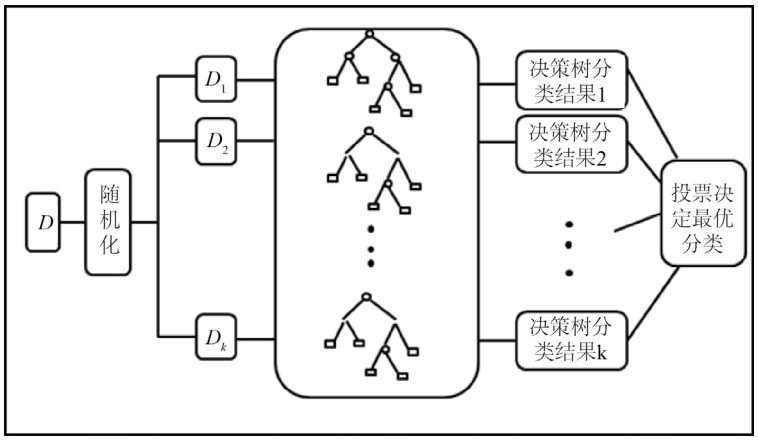

Figure 1. Random forest flow
图 1. 随机森林流程图

2.1. 随机森林算法

随机森林算法为[8]：

(1) 对于 $b=1,\cdots,B$

(a) 通过 bootstrap 抽样方式产生 b 个样本子集。

(b) 对每个 bootstrap 样本建立随机森林树 T_b，每个叶子节点递归地重复以下步骤，直到叶子节点包含的数据量为 $n_{\min}$ 为止。

① 从 p 个自变量中随机选择 mtry 个自变量。在使用随机森林做回归时 mtry 默认值为 $p/3$，使用随机森林做分类时默认值为 $\sqrt{p}$，其中 p 为自变量个数。

② 在 mtry 个自变量中选择最好分裂变量和分裂点。

③ 将节点拆分为两个叶子节点。

(2) 输出集成树 $\{T_b\}_1^B$

(3) 预测

(a) 对于回归问题，待测样本 x 的预测为：

$$\widehat{f_{rf}^B}(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x)$$

(b) 对于分类问题，设 $\widehat{C}_b(x)$ 是第 b 棵树的类预测。

$$\widehat{C_{rf}^B}(x)=\text{majorityvote}\left\{\widehat{C}_b(x)\right\}_1^B$$

majorityvote 为多数票。

2.2. 变量重要性排序

随机森林可以给出变量重要性的排序，其具体过程如下[9]：

(1) 对每个 bootstrap 抽取的样本建立一个回归树模型，同时使用该模型对相应的袋外数据 OOB(out-of-bag)进行预测，得到 B 个袋外数据的残差均方，记为 $MSE_1, MSE_2, \cdots, MSE_B$。

(2) 变量 X_i 在 B 个 OOB 样本中随机置换，得到新的 OOB 样本，然后用已建立的随机森林模型预测新的 OOB 样本，得到随机置换后的 OOB 残差均方如下：

$$\begin{bmatrix} MSE_{11} & MSE_{12} & \cdots & MSE_{1B} \\ MSE_{21} & MSE_{22} & \cdots & MSE_{2B} \\ MSE_{31} & MSE_{32} & \cdots & MSE_{3B} \\ \vdots & \vdots & \ddots & \vdots \\ MSE_{p1} & MSE_{p2} & \cdots & MSE_{pB} \end{bmatrix}$$

(3) 用 $MSE_1, MSE_2, \cdots, MSE_B$ 与如上矩阵对应的第 i 列向量相减，平均后再除以标准误则得到变量 X_i 的重要性排序，即

$$score_i = \left(\sum_{j=1}^{B} \left(MSE_j - MSE_{ij} \right) / b \right) \Big/ S_E, (1 \le i \le p)$$

随机森林方法通过在 OOB 样本中随机地置换变量，计算预测精度下降程度来衡量变量的重要性，其数值越大说明变量越重要。

3. 中国电影票房影响因素及预测的实证分析

3.1. 中国电影票房影响因素指标的选取以及数据来源

本文根据实际问题背景以及数据获取难易程度选择影响中国电影票房的 7 个因素，分别为：档期、是否有续集、首映日票房、点映票房、首周末票房、百度指数和豆瓣评分。

① 档期：中国电影目前主要有暑期档和贺岁档两大特殊档期，本文将档期分为三类[10]：贺岁档为每年 12、1、2 月份；暑期档为每年 6、7、8 月份；其余月份为其他档期记为第一类。

② 续集：漫威系列电影的成功启示我们是否有续集可能会吸引特定的观众带来源源不断的票房。国内《人在囧途》、《叶问》、《战狼》等影片的成功预示着续集有可能成为影响电影票房的因素。

③ 首映日票房：首映日票房整体上可以反映观众对于一部电影的关注度，可以反映电影上映前电影的宣传效果。《美人鱼》上映当天票房达到 2.72 亿元，最终票房大卖。发行商可以根据首映日票房进一步明确影片定位，调整营销策略。

④ 点映票房：点映是电影上映前，制作团队在个别城市、个别影院对影片提前放映。点映在好莱坞有半个多世纪的历史，中国电影点映始于张艺谋导演的作品《英雄》[11]。点映一方面可以满足观众的好奇心，为电影的正式上映积累大量的口碑，另一方面可以通过观众的反馈调整上映期间的场次，适当改变营销方案。

⑤ 首周末票房：电影上映一周的首周末票房可以检验这部电影是否被观众认可，可以为接下来一段时间的排片宣传提供一定的参考。

⑥ 百度指数：百度指数是当前互联网时代重要的统计分析平台之一，是众多企业营销决策的重要依据。百度指数里可以看到以电影名为关键词的搜索量规模大小，电影上映前百度指数是指以该电影为关键词的预告片以及宣传片的搜索量。本文统计了一部电影上映前四周的百度指数，由于搜索量波动较大，

选择电影上映前四周的平均百度指数作为研究变量。

⑦ 豆瓣评分：豆瓣电影是中国最大最权威的电影分享与评论社区[12]，电影上映后，观众会通过自己的综合观感在豆瓣电影给出综合评分以及评论。豆瓣评分代表着电影口碑。一部电影的评分会随着上映期间观众的评价不断更新，无法动态收集，本文采用电影上映后的综合评分作为研究变量。

各影响因素指标具体如表 1。

根据《艺恩数据》及《中国电影票房数据库》，得到我国 2014~2018 年上映的 225 部影片的观测数据，部分数据如表 2 所示。

Table 1. Index selection
表 1. 指标选取

变量	类型	指标选取
x_1：档期	分类变量	贺岁档记为 3，暑期档记为 2，其他记为 1
x_2：是否有续集	分类变量	有续集记为 1，无续集记为 0
x_3：首映日票房	数值型变量	上映当天总票房
x_4：点映票房	数值型变量	上映前提前点映总票房
x_5：首周末票房	数值型变量	上映后第一个周末总票房
x_6：百度指数	数值型变量	上映前四周平均百度指数
x_7：豆瓣评分	数值型变量	豆瓣评分

Table 2. Some data of 225 films from 2014 to 2018
表 2. 2014~2018 年 225 部影片部分数据

影片	y (亿)	x_1	x_2	x_3 (亿)	x_4 (万)	x_5 (亿)	x_6	x_7
《心花路放》	11.70	1	0	1.03	3695.5	2.78	171661.25	7
《西游记之大闹天宫》	10.46	3	0	1.29	159.7	3.2	23957.25	4.1
《捉妖记》	24.4	2	0	1.63	1128.5	5.02	167710.5	6.8
…								
《寻龙诀》	16.83	3	0	1.62	941.90	5.92	73390.5	7.6

进行数据分析前，为了消除量纲以及数据数量级大小的影响，对于数值型变量的观测数据进行标准化处理，即因变量票房，自变量首映日票房、点映票房、首周末票房、百度指数和豆瓣评分标准化处理 z 分数表示为公式(1)：

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

$\bar{x}$ 表示本组数据的平均数，s 表示本组数据的标准差。

3.2. 随机森林模型的建立

建立票房回归预测模型，随机森林模型中有两个参数 mtry 和 ntree，其中 mtry 表示每一步分裂选择的自变量个数，ntree 为随机森林中树的个数，R 软件 randomforest()函数默认 mtry = $p/3$，p 为自变量个数，默认树的个数 ntree = 500，Gareth James 等[13]指出随机森林里参数取默认值也可以取得较稳健的效果。

模型建立过程：

(1) 对于 $b=1,\cdots,500$

(a) 通过 bootstrap 抽样方式产生 500 个样本子集。

(b) 对每个 bootstrap 样本建立随机森林树 T_b，每个叶子节点递归地重复以下步骤，直到叶子节点包含的数据量为 5 为止。

① 从 7 个自变量 $x_1,x_2,\cdots,x_7$ 中随机选择 2 个自变量。

② 在 2 个自变量中选择最好分裂变量和分裂点。

③ 将节点拆分为两个叶子节点。

(2) 输出集成树 $\{T_b\}_1^B$

(3) 预测

对于回归问题，待测样本 x 的预测为：

$$\widehat{f_{rf}^B}(x)=\frac{1}{500}\sum_{b=1}^{500}T_b(x)$$

3.3. 变量重要性排序

使用上述建立的随机森林模型，可以采用 R 软件计算得出影响电影票房的变量重要性排序如表 3 所示。随机森林方法通过在 OOB(out-of-bag)样本中随机地置换变量，计算变量重要性。该数值越大，变量越重要，对电影票房的影响越大。

Table 3. Order of importance of variables
表 3. 变量重要性排序

变量	节点纯度
档期	3.34913
是否有续集	1.76756
首映日票房	44.11292
点映票房	13.79366
首周末票房	60.88706
百度指数	20.88351
豆瓣评分	14.74532

通过计算得出的变量重要性排序可知，首周末票房、首映日票房、百度指数、豆瓣评分和点映票房为影响电影票房的重要因素，档期和续集对于电影票房的影响可以忽略不计。R 软件中 plot()函数可以给出变量重要性排序图，如图 2 所示。

3.4. 三种方法预测结果比较

除了随机森林模型外，神经网络在分类数据预测方面有较好的效果，线性回归模型是传统的预测模型，这两种方法在电影票房的预测方面都有提及，本文选取 2019 年 12 部影片使用三种模型进行电影票房的预测，预测误差定义如公式(2)：

$$\text{预测误差}=\frac{|\text{实际值}-\text{预测值}|}{\text{实际值}} \tag{2}$$

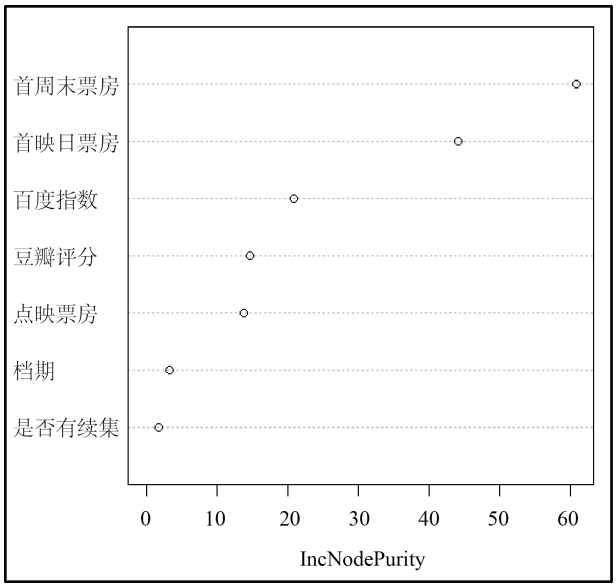

Figure 2. Order of importance of variables
图 2. 变量重要性排序

采用三种方法对影片预测数值如表 4，影片预测对比折线图如图 3，影片预测误差如表 5。

Table 4. Box office estimates for 12 films in 2019
表 4. 2019 年 12 部影片票房预测结果

影片	实际票房(亿)	随机森林预测票房(亿)	神经网络预测票房(亿)	线性回归预测票房(亿)
《中国机长》	28.84	20.20	13.70	24.93
《飞驰人生》	17.03	18.40	13.20	6.18
《扫毒 2 天地对决》	12.85	10.50	12.60	13.65
《叶问 4》	11.72	9.30	12.70	11.31
《攀登者》	10.88	9.93	13.70	4.43
《比悲伤更悲伤的故事》	9.46	6.61	5.80	11.86
《银河补习班》	8.64	8.57	5.80	9.68
《反贪风暴 4》	7.88	7.48	5.81	8.93
《熊出没原始时代》	7.09	7.27	5.80	6.37
《使徒行者谍影行动》	6.93	9.82	5.83	3.27
《老师好》	3.50	3.11	5.80	3.53
《追龙 2》	3.04	4.14	5.80	4.33

表中可以看出，随机森林回归模型显示出良好的预测精度。影片的预测误差多在 20%左右。神经网络和线性回归模型对于部分影片的预测误差较小，但其预测误差范围较大，对某些影片的预测误差太大。12 部影片中，随机森林对电影《银河补习班》的预测误差达到 0.81%，对电影《熊出没原始时代》的预测误差为 2.54%。说明随机森林选择出的影响电影票房的重要因素，首周末票房度、首映日票房、百度指数、豆瓣评分和点映票房是决定这两部影片总票房收入的重要因素。

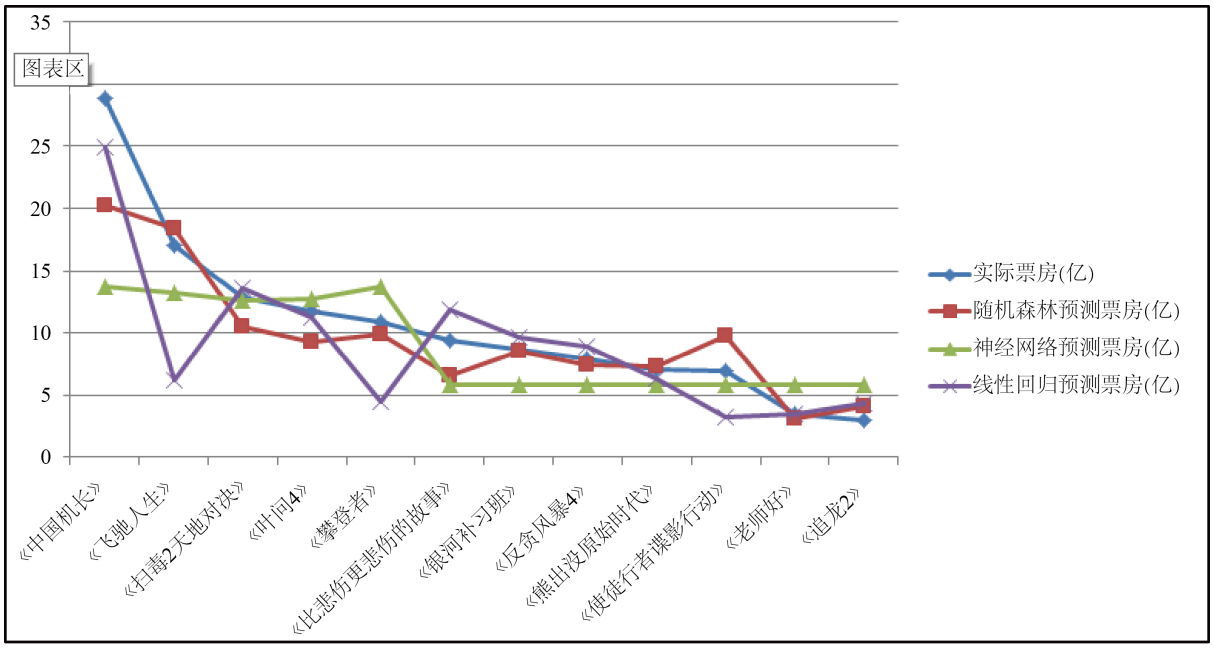

Figure 3. Line chart of box office predictions for 12 films in 2019
图 3. 2019 年 12 部影片票房预测结果折线图

Table 5. Box office error for 12 films in 2019
表 5. 2019 年 12 部影片票房预测误差

影片	随机森林预测误差(%)	神经网络预测误差(%)	线性回归预测误差(%)
《中国机长》	29.96%	52.50%	13.56%
《飞驰人生》	8.04%	22.49%	63.71%
《扫毒 2 天地对决》	18.29%	1.95%	6.23%
《叶问 4》	20.65%	8.36%	3.50%
《攀登者》	8.73%	25.92%	59.28%
《比悲伤更悲伤的故事》	30.13%	38.69%	25.37%
《银河补习班》	0.81%	32.87%	12.04%
《反贪风暴 4》	5.08%	26.27%	13.32%
《熊出没原始时代》	2.54%	18.19%	10.16%
《使徒行者谍影行动》	41.70%	15.87%	52.81%
《老师好》	11.14%	65.71%	0.86%
《追龙 2》	36.18%	90.79%	42.43%

电影《银河补习班》2019 年 7 月 18 日上映，总票房 8.78 亿元，首周末票房 2.70 亿元，首映日票房 0.65 亿元，上映前四周平均百度指数 52717.5，豆瓣评分 6.3 分，点映票房 0.99 亿元。电影《熊出没原始时代》2019 年 2 月 5 日上映，总票房 7.18 亿元，首周末票房 1.90 亿元，首映日票房 0.74 亿元，上映前四周平均百度指数 1303.75，豆瓣评分 6.7 分，点映票房 0.40 亿元。

两部影片的数据对比表明，首周末票房和首映日票房都取得极大成功的基础上，在 2019 年同档次总票房收入的影片中，这两部影片有一个共同的特点即点映票房较高。《银河补习班》在上映前多地超前

点映，观众的反响和点映现场的效果反馈其不仅在故事发展上吸引观众眼球，人员演技及影片质量同样好评如潮，使得该片的首映日票房和首周末票房取得不错的成绩。但是因为点映效果良好，提高了观众对这部影片的期待，使得其后续豆瓣评分成绩一般。《熊出没之原始时代》有别于《银河补习班》，其不仅是依靠点映票房取得总票房的好成绩，这部影片有续集，连续多年积累了一定的口碑，传播较为广泛，使得其在上映之后取得不错的票房。

故首周末票房、首映日票房、百度指数、豆瓣评分和点映票房对电影票房总收入影响较大，制片方、营销方和院线可以根据影片的实际市场情况，采取合适的营销方式来提高票房收入。

4. 结论

本文依据 2014~2018 年 225 部国产电影票房数据，运用随机森林方法对影响国产电影票房的因素进行分析，最终得出影响电影票房的主要因素有首周末票房、首映日票房、百度指数、豆瓣评分和点映票房五个因素，档期以及是否有续集这两个因素对电影票房的影响可以忽略不计。

首周末票房可以检验一部电影的口碑，反映电影的火爆程度以及在后续电影放映中的竞争态势，发行方可根据首周末票房来调整营销策略。首映日票房是电影本身类型、导演、演员类型以及上映前电影宣传情况的一个综合体现，提高首映日票房需要提高电影本身的制作，需要加强电影的宣传。百度指数可以洞察电影上映前观众对影片的兴趣、监测舆情动向、定位受众的特征。提高百度指数多在电影制作拍摄包括电影制作完成后要实时宣传，引起观众兴趣。豆瓣评分会动态影响电影票房，豆瓣评分较高的电影会吸引一部分观众，豆瓣评分较低会让一部分本来要去看电影的观众选择放弃观看。电影点映属于电影宣传环节，点映过程中可以收集观众对影片的初步评价，如果电影有所不足可以调整营销策略弥补电影本身的不足。

在分析影响电影票房因素的基础上，本文采用随机森林模型、神经网络模型和线性回归模型对 2019 年 12 部影片进行了预测。预测结果表明，随机森林对于《银河补习班》、《熊出没原始时代》、《反贪风暴 4》这三部影片的预测误差在 5%左右取得较好的效果，部分影片预测误差较大，但整体来讲随机森林预测票房较为稳健。神经网络和线性回归模型对于部分影片的预测效果良好，针对大部分影片其预测误差波动较大，就本文的研究而言，针对票房数据建立的三种预测模型，随机森林取得良好的效果。

电影票房影响因素分析及预测中，演员阵容、导演、发行商和新浪微博的宣传力度等是否会影响电影票房的收入，如何将这些变量量化纳入模型本文没有提及，有待继续探索研究。

参考文献

[1] 杨威. 基于微博数据的电影票房预测模型研究[D]: [硕士学位论文]. 安徽: 安徽大学计算机应用技术专业, 2014.

[2] 张雪. 基于深度学习卷积神经网络的电影票房预测[D]: [硕士学位论文]. 北京: 首都经济贸易大学统计学院, 2017.

[3] 郭萱. 基于随机森林的电影票房预测研究[D]: [硕士学位论文]. 北京: 中国石油大学(北京)数学系, 2018.

[4] 鲁月. 基于随机森林因素筛选的国产电影票房组合预测模型研究[D]: [硕士学位论文]. 江苏: 南京航空航天大学经济与管理学院, 2019.

[5] Boulesteix, A.L., Janitza, S., Kruppa, J., *et al.* (2012) Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery*, **2**, 493-507. https://doi.org/10.1002/widm.1072

[6] 曹正凤. 随机森林算法优化研究[D]: [博士学位论文]. 北京: 首都经济贸易大学统计学院, 2014.

[7] 王星. 大数据分析: 方法与应用[M]. 北京: 清华大学出版社, 2013: 63-65.

[8] Hastie, T., Tibshirani, R. and Friedman, J. (2008) The Elements of Statistical Learning. Stanford, California, August, 588-590.

[9] 徐戈. 基于随机森林模型的房产价格评估[J]. 统计与决策, 2014(17): 22-25.

[10] 聂鸿迪. 中国电影票房的影响因素及其实证研究[D]: [硕士学位论文]. 北京: 北京交通大学经济管理学院, 2015.

[11] 郎倩雯. 中国电影公关营销策略研究[D]: [硕士学位论文]. 浙江: 浙江大学传媒与国际文化学院, 2011.

[12] 宋恩梅, 朱梦娴. 社会化媒体信息分布规律研究: 以电影评论为例[J]. 信息资源管理学报, 2015(3): 25-36.

[13] James, G., Witten, D., Hastie, T., *et al.* (2013) An Introduction to Statistical Learning: With Applications in R. Springer, New York, 320-321. https://doi.org/10.1007/978-1-4614-7138-7_1