Advances in Social Sciences 社会科学前沿, 2019, 8(4), 489-495
Published Online April 2019 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2019.84069

Hans汉斯

Analysis and Forecast of the Influencing Factors of Chinese Box Office

Shanshan Ping

Jiangxi University of Finance and Economics, Nanchang Jiangxi
Email: 1183363468@qq.com

Received: Mar. 20th, 2019; accepted: Apr. 4th, 2019; published: Apr. 11th, 2019

Abstract

More and more people and investors focus on film industry in the context of promoting the development of film industry. Studying on the influence factors of box office and predicting box office can benefit the development of film industry and can be the base for investors to make right decisions. This article selects top 100 Chinese mainland domestic films since 2011, which excludes animated films. We select and define 18 indexes to build the indexation system of influence factors of revenues of releasing films. We use stepwise multiple regression method to analyze the influence factors of box office. From the result, the market factor has the best effect. We can predict box office from multiplied regression, Neural Networks, Random forest and SVR model. We can find the best fitting model by comparing these four models. We can find that the random forest model is the best model.

Keywords

Box Office, Multiple Regression, Neural Networks, SVR Model, Random Forest

我国电影票房影响因素分析及预测

平珊珊

江西财经大学，江西 南昌
Email: 1183363468@qq.com

收稿日期：2019年3月20日；录用日期：2019年4月4日；发布日期：2019年4月11日

摘 要

电影票房作为电影行业最为主要的收入来源，研究其影响因素并对其进行预测，有利于电影行业的发展

文章引用: 平珊珊. 我国电影票房影响因素分析及预测[J]. 社会科学前沿, 2019, 8(4): 489-495.
DOI: 10.12677/ass.2019.84069

和投资者做出正确投资决策。本文选择了2011年至2017年，除动画电影外的100部国产电影，在创作、市场和营销力度三个方面共选择18个指标构建指标体系。首先用逐步多元回归模型对电影票房影响因素进行分析，然后选择10部电影进行预测，通过多元回归、BP神经网络、随机森林和SVR模型构建票房预测模型，并比较几个模型的拟合程度和稳定性，发现随机森林模型的预测效果最好，因此用随机森林模型作为预测电影票房的模型更加合适。

关键词

电影票房，多元回归，神经网络，SVR模型，随机森林

1. 引言

近十几年来，我国国民的收入有了大幅提升，物资水平的丰富使得人们的精神文化需求增长迅速。因此以电影产业为代表的文化产业迅速发展，电影作为一种文化传播的途径，不仅能够满足人们对文化的需求，也是一国在世界上话语主导权的体现[1]。电影对于传播一个国家、一个社会的文化起着至关重要的作用[2]。很多国家和政府为了宣传本国的价值观和文化都对本国的电影行业给予了大力支持与引导。如今，中国已跻身世界第二大电影市场，在制作技术、画质等方面都有了很大的突破，2016 年电影总票房达到了 440.46 亿，同时，在全球经济一体化大趋势下的今天，电影作为文化软实力的一种重要表现形式，加强电影产业的发展是取得国际竞争优势的重要途径之一。

中国电影产业的发展对促进中国经济发展和中国传统文化传播都有着不容小觑的作用，对中国电影票房的研究则是较为直观的展现出电影在国民经济中的地位，能够对未来促进电影行业的稳健发展带来重要的参考。在我国的电影产业发展中，80%的电影产业收入来源于票房收入，电影票房对我国电影产业的促进作用显而易见。同时，在快餐文化盛行的当下，外国文化给我国传统文化带来一定程度的冲击，我国的电影文化也逐渐接受外来文化趋于国际化。在这种形式下，我国的传统文化更应该被大力的保护、宣扬和传承，并且电影作为文化传播的一种载体，增强我国的文化软实力、促进我国电影在世界范围内的传播、加强海外出口和推广力度是提升我国的国际竞争力至关重要的渠道。本文对中国电影票房的影响因素进行研究分析，从多个影响因子来论证对我国电影票房的影响，之后通过比较多元线性回归和三种机器学习模型的拟合精度，选出拟合精度较高的模型，给投资者提出了良好的预测票房模型，为投资者减少了风险，对我国电影产业起到了积极作用，对我国文化产业的发展及经济发展都具有一定的正面意义。

2. 建立指标体系

因各个国家或地区之间的经济发展状况存在差异，各国政府对有关电影产业的支持程度也不一样，各地区电影市场的发展也有所不同，所以对票房收入产生影响的因素也各不相同[3]。由于动画片受众大部分来自于低龄用户，故本文在剔除动画片后，选取了 2011~2017 年期间票房前 100 名的电影进行研究，从创作、市场和营销力度三个方面来阐述电影票房的影响因子，构建我国电影票房影响因素研究的指标体系[4]。本文具体用到的指标如表 1 所示：

Table 1. Indicator system table
表 1. 指标体系表

一级指标	二级指标	因变量
创作方面	电影类型	电影票房
	故事内容	
市场方面	导演评分	
	演员评分	
	技术效果	
	档期	
营销方面	发行公司	
	口碑评分	
	想看人数	
	百度指数	

3. 模型建立

为了探究什么样的模型能更好、更稳定地拟合出我国的电影票房，本文采用多元线性回归，BP 神经网络，SVR 模型，随机森林模型进行拟合，并选出最稳定、预测最准确的模型。

3.1. 多元线性回归

在多元线性回归之前，应该要对变量进行筛选，以消除多重共线性，本文采用逐步回归法来筛选变量。

逐步回归的基本思想是把自变量逐个地放入回归方程中，每次首先保留住对因变量最显著的那个变量，在继续放入自变量的同时，如果原来已经在回归方程中的自变量变得不显著了，要把这个不显著的自变量删除。反复进行上述的操作，直到回归方程中的自变量对因变量都有显著影响，同时又没有漏掉对因变量影响显著的自变量。

进行逐步回归之后，我们就可以得到多元线性回归的最优模型。利用 R 软件进行逐步回归的结果如表 2 所示：

Table 2. Step by step
表 2. 逐步回归

	系数	标准误	t 值	P 值
截距项	2.149	0.802	2.679	0.008908
是否是贺岁片	2.817	0.711	3.964	0.000156
百度指数	1.167e−05	1.788e−06	6.529	4.96e−09
是否 IP	−2.343	0.716	−3.274	0.001547
是否奇幻剧	3.171	0.996	2.666	0.00931
想看人数	1.432	0.841	1.703	0.00251

$R^2 = 0.5887$, $F = 23.27$, $P = 9.45e{-}15$.

上表的逐步回归结果显示，自变量的回归的系数都是显著的，对于本文的情况来说，上述结果已经是最优的线性回归方程。选取的变量为“是否是贺岁片”，“百度指数”，“是否 IP”，“是否是奇幻

剧”，“猫眼想看人数”这 5 个指标。接下来就用这 5 个指标来进行分析。从自变量的系数特点来看，只有是否是 IP 有负影响，也就是说 IP 剧会导致电影票房降低，这与实际生活中的情况不一样，说明可能这个线性模型的解释度不是很好，有可能是数据中的异常值导致的，需要进一步的优化。

3.2. BP 神经网络

在实际生活中，很多模型并不是线性的，并不能用线性回归模型来解决，而神经网络可以解决这个问题。神经网络作为一种经典的机器学习模型，处理非线性问题是它的一个重要功能。它的工作原理与人的神经系统类似，由多个多层次的神经元有机地组合在一起，把信息广泛分布式地存储在神经元中，能够自主学习。其算法的基本思想是：当一个信号传入系统的一个神经元时，首先乘以一个权值，再经过一个激活函数输出，进入下一个神经元，重复执行上述操作，直到输出层输出一个值，最后比较输出层的数据与实际的数据之间的差异，如果差异大于之前设定的误差值，那么就改变各神经元连接层的权值，直到使得误差小于设定的误差。

神经网络模型可以简单表达如公式(1)：

$$\mathrm{Y}=f_2\left(\sum_{i=I}^{N_i} w_j^2\cdot\left(f_1\left(\sum_{j=I}^{N_j} w_{ij}^1\cdot x_i\left(\mathrm{I}\right)+b_j^1\right)\right)+b_0^2\right) \tag{1}$$

其中，N_i 和 N_j 表示隐含层节点数与输入层节点数，f_1 和 f_2 分别为隐含层和输出层的传递函数，w_j^2 和 w_{ij}^1 分别表示为隐含层 m 个节点到单个输出节点的权重和输入层 i 个节点到隐含层 m 个节点的权重。b_j^1 和 b_0^2 分别表示第 m 个隐含层节点偏倚和输出层的偏倚[5]。

另外，隐含层有 3 个神经元节点，隐含层节点数是根据经验得出的，通常有下面的公式(2)：

$$m=\sqrt{n+l}+\alpha \tag{2}$$

上式中，n 可以理解为输入层节点数，l 为输出层节点数，α 可以取 1 到 10 的任意整数(本文取了 1)，m 为隐含层节点数，本文中 $n=1$，$l=1$，$\alpha=1$，所以 m 可以选择 3。

图 1 为迭代次数，可以看到迭代 116 次之后收敛。

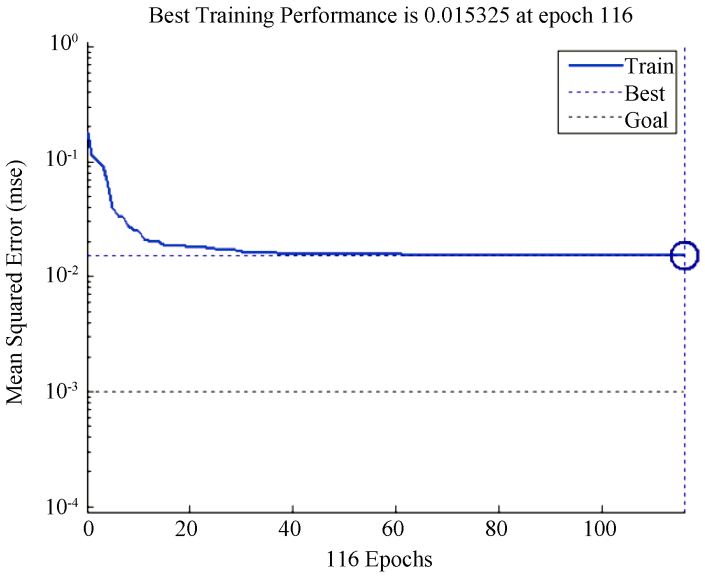

Figure 1. Neural network iterations
图 1. 神经网络迭代次数

3.3. SVR 支持向量机回归模型

SVR 就是指支持向量机回归模型，它是基于 SVM 模型的回归模型。SVM 即支持向量机也是机器学习模型，在判别、分类以及回归分析有很多的应用[6]。这种分类方法的原理是在一个空间内寻找到一个超平面，使得这个超平面能够把空间分成两部份，且这两部分到这个超平面的距离最大。如果数据线性不可分，那么就把原数据通过一个非线性映射转换成高维空间的数据，再在这个空间找到一个符合要求的超平面。

本文采用的 SVR 模型的基本形式可以解释如下：

设 S 为输入变量的值，Y 为相应的输出值。支持向量机的基本思想是寻找从输入空间到输出空间的一个非线性映射 $\varphi(x) \in R^d \to F$，将输入数据 x 映射到高维特征空间 F，并在特征空间中用公式(3)来估计函数：

$$f(x) = (w * \varphi(x)) + b, w \in F \tag{3}$$

w 是权值向量，b 是偏置项。w 和 b 通过最小化一个泛函公式(4)来估计：

$$R(w) = \frac{1}{2}\|w\|^2 + C\sum_i^n J^i(y_i, d_i) \tag{4}$$

其中 d_i 为 SVR 的实际输出，J 是 ε 不敏感损失函数。

通过 R 软件来运行 SVR 模型，模型参数表示如表 3：

Table 3. Basic parameters of SVR model
表 3. SVR 模型基本参数

SVM-Type	SVM-Kernel	cost	gamma	epsilon
eps-regression	radial	1	0.05882353	0.1

3.4. 随机森林模型

随机森林模型其实是从决策树模型发展而来的，是决策树模型的泛化版本。决策树是一种基本的分类器，一般是将特征分为两类，判定样本属于哪个类别的算法。而一个随机森林模型中有多个决策树，其分类的确定是由多个决策树分类结果的众数决定，可以形象的描述为投票决定。在处理回归问题时，就是把多个决策树的结果进行平均。

随机森林模型工作原理可用图 2 来表示：

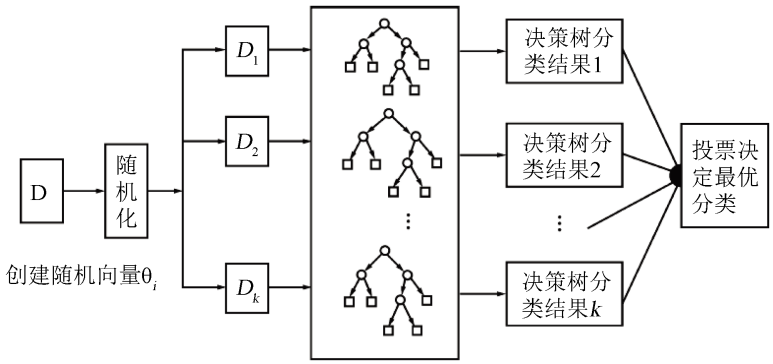

Figure 2. Random forest schematic
图 2. 随机森林原理图

本文对随机森林的模型主要参数设置为：ntree = 100，mtry = 2。然后得到了精确度递减特征值重要性水平，如图 3 所示：

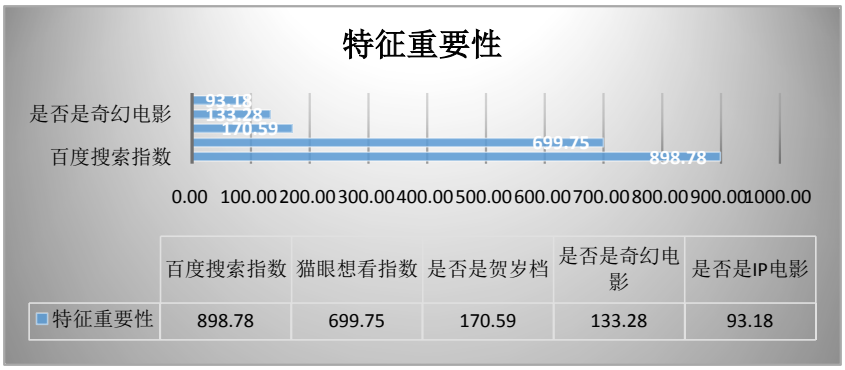

Figure 3. Accuracy decrement eigenvalue importance level
图 3. 精确度递减特征值重要性水平

从上图中可以看出“百度搜索指数”与“猫眼想看指数”这两个特征的重要性要明显大于其余 3 个指标，说明“百度搜索指数”与“猫眼想看指数”这两个指标最具分类能力。

4. 模型结果分析

4.1. 拟合度与稳定性对比

用多元线性回归，BP 神经网络，SVR 支持向量机回归和随机森林模型对数据进行拟合之后，我们可以得到每个模型的 MSE 与 NMSE，这 3 个模型的结果如表 4：

Table 4. MSE & NMSE
表 4. MSE 和 NMSE

	多元线性	BP 神经网络	SVR	随机森林
MSE	9.1949	4.0424	32.1390	4.6703
NMSE	0.3458	0.7839	1.2086	0.1756

从 MSE 的角度来看，SVR > 多元线性 > 随机森林 > BP 神经网络，说明 BP 神经网络的拟合度最好；从 NMSE 的角度来看，SVR > BP 神经网络 > 多元线性 > 随机森林，说明随机森林的稳定性最好。所以综合这两个指标，我们可以认为 SVR 模型拟合效果不是很好，而随机森林模型值得考虑。

4.2. 预测精确度对比

为了检测这 3 种模型的预测能力，本文选择了 10 部电影作为测试样本，得到预测的票房，然后与真实值相比较，得到偏差最小的模型，其结果如表 5。

总偏差是取真实票房值与预测值的差值的绝对值，再求和得到的[7]。那么可以看到总偏差值：SVR > BP 神经网络 > 多元线性 > 随机森林，说明随机森林模型的拟合效果更好。

5. 结论与建议

本文最初选择了 18 个与电影票房有关的变量，但是这些变量肯定有一些存在着共线性问题或者与因变量(电影票房)相关性不明显的变量。因此，我们在进行建模之前进行了变量筛选，剔除了 13 个变量，保留了 5

个变量，分别是“是否为贺岁片”，“百度指数”，“是否IP”，“是否是奇幻剧”，“猫眼想看人数”。

Table 5. Total deviation between predicted and actual values
表 5. 预测值与实际值的总偏差

电影名	票房真实值	多元线性	BP 神经网络	SVR 预测	随机森林预测
澳门风云	5.52	7.3503	7.8547	8.5214	8.0674
北京遇上西雅图	5.2	7.0053	6.6638	5.7055	6.6731
狄仁杰之神都龙王	6.02	4.1891	2.9331	3.3144	5.2376
风暴	3.14	6.4794	7.2816	7.9796	6.5112
一九四二(冯小刚)	3.72	3.5476	5.5963	5.9536	5.8967
建党伟业	4.23	3.5493	2.772	3.0035	4.1311
画皮 II	7.03	2.3572	3.7645	4.2976	4.4611
失恋 33 天	3.57	9.7334	7.1653	7.2171	9.6888
龙门飞甲	5.56	7.6592	8.0354	8.8565	8.9623
金陵十三钗	6.09	7.5555	6.7837	7.0633	7.4399
总偏差		**24.0599**	**24.3912**	**25.1618**	**23.8897**

对于剩下的5个变量，我们首先建立了多元线性回归模型，结果发现“是否是IP电影”这个自变量的系数为负数，不符合实际生活中的情况。我们猜测是由于数据中的异常值导致的这个问题。我们剔除了3个异常值之后，经过逐步回归筛选了7个变量：“是否贺岁片”、“百度指数”、“是否喜剧片”、“是否动作片”、“是否是IMAX”、“豆瓣评分”、“想看人数”，在新的数据的基础上建立了新的多元线性回归模型、BP神经网络模型、SVR模型和随机森林模型。剔除异常之后，多元线性回归模型的R^2变大了，而且自变量的系数也符合我们实际生活中的规律：7个变量的系数都为正，说明这些因素都对电影票房有正向的影响，影响力最大的是影片是否为贺岁档和是否是IMAX。

对于以上的结论，笔者给出以下建议：

1) 对于电影投资方，如果仅是出于票房考虑，应该投资贺岁档的电影，在技术上应该采用IMAX，这样可以更好地吸引观众，获得更高的票房[8]。

2) 如果想预测一部电影的票房，可以关注一下它的“猫眼想看指数”、“豆瓣评分”和“百度搜索指数”，并且看它是否是贺岁片、喜剧片、动作片，或者电影是否采用了IMAX技术。

3) 可以选用上述的7个指标建立随机森林模型来对电影票房进行预测。

参考文献

[1] 刘连涛. 我国3D电影票房影响因素[J]. 商, 2016(13): 201-202.

[2] 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(8): 35-39.

[3] Hennig-Thurau, T., Houston, M.B. and Walsh, G. (2007) Determinants of Motion Picture Box Office and Profitability: an Interrelationship Approach. *Review of Managerial Science*, **1**, 65-92.

[4] 陈然. 我国商业电影票房影响因素研究——基于多元线性回归和神经网络分析[D]: [硕士学位论文]. 昆明: 云南财经大学, 2016.

[5] 程相君, 王春宁, 陈生潭. 神经网络原理及其应用[M]. 北京: 国防工业出版社, 1995.

[6] 陈诗一. 非参数支持向量回归和分类理论及其在金融市场预测中的应用[M]. 北京: 北京大学出版社, 2008.

[7] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013(4): 1190-1197.

[8] 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, 2015.