电影大数据

摘要 对于观众来说，电影评分对于观众选择电影来说是有很重要的参考建议，建立合理的评分模型是非常必要的。而对于电影投资方来说，票房的走势才是更重要的。本文先重点研究了现有网站的评分模型，分析了利弊性，建立了合适的、完整的、系统的电影评分体系，对此模型的的效果进行了模拟，从而达到优化的目的。同样，对于电影的票房预测，本文重点分析了影响电影票房的各个指标，建立了符合现实、有预测价值的票房预测模型，对其预测效果进行合理评判。

首先，对于电影评分体系，我们先爬取了豆瓣和 IMDB 的电影的具体数据，通过数据可视化分析出了影响两个网站排名的的具体因素。然后分析了豆瓣、IMDB 的电影评分算法，得出了他们评分排名系统的优劣性，找到了影响电影评分的两个主要因素，即评分人数和实际评分。因此我们根据这两个影响参数通过**威尔逊区间**算法得出了电影评分的初步估计量，然后再根据**贝叶斯平均**算法修正这个初步估计量，再结合实际给出了电影的评分。据此模型，我们实现了对豆瓣和 IMDB 的 TOP50 预测。

其次，对于票房预测体系，我们进行**多元线性回归**分析。我们将综合采用电影类型，明星效应，导演效应，电影档期，电影时长，评论人数，电影评分，制作技术八个指标来进行建模。

关键字：威尔逊区间，贝叶斯平均，多元线性回归。

一 、问题的重述

1.1 问题背景

电影评分对于观众选择电影有很重要的参考建议，如今比较热门的电影评分网站包括：豆瓣，IMDB，猫眼电影，Rotten Tomatoes 等等，每个网站都有自己不同的评分模型和规则，目前的评分模型有直接求平均分的，但这样可能会存在恶意刷分的现象；也有给不同的用户分配不一样的权重的，大 V 的权重较高，甚至某些网站只允许大 V 影评人评分；还有通过贝叶斯统计算法来进行综合评分的。每种评分模型都有各自的优缺点，大部分网站也没有确切给出评分模型的具体细节。而对于电影出品方、投资方来说，更关注的不是评分，而是电影的票房。如果能够在电影出品前期准确预测该电影的票房，会对电影出品方、投资方提供更好的参考。如何建立合理的电影评分体系以及如何预测电影的票房走势，都是如今大数据时代的需求。

1.2 问题重述

对于观众来说，当我们开始习惯于用评分来决定自己是否去看一部电影时，电影评分对我们的价值不言而喻。

对于电影投资方来说，票房的预测是至关重要的，投资方与制片方需要预测参考。

我们需要就以上两个关键点进行切入，为科学决策提供具体数学依据，为此请完成以下几个任务

1. 请分析比较现有的多家网站的电影评分模型，在此基础上给出你们团队认为合理的电影评分模型；
2. 豆瓣 TOP50 电影和 IMDB 的 TOP50 电影有很多相同也有很多不同的电影，请基于你们的分析，给出影响两者排名不同的具体因素，并且给出你们的电影评分模型下 TOP50 电影；
3. 请你们团队建立电影票房预测模型，并预测今年如下两部电影的全球票

房：Nolan 的《信条 Tenet》和娄烨的《兰心大剧院》（可假设完全不受疫情影响，电影正常上映，电影院正常开放）；

4. 依据你们团队的票房预测模型，分析影响电影票房的最重要的因素。

二 、问题分析

2.1 对问题一的分析

题目中要求我们对一些热门的电影评分网站包括：豆瓣，IMDB，猫眼电影的评分模型和规则进行模型分析。虽然目前大部分网站没有开源详细的评分模型细节，但是我们通过网上的现有的数据进行分析。

首先，我们整理了现有的主流电影评价网站的评分机制。包括各个网站的打分人群，打分机制和记分机制……我们横向对比了同一电影不同电影评价网站的评分，纵向总结了各个网站的评分模型参数。

接下来，我们将会对上述电影评价网站的打分模型进行分析，并且抽样网站的电影数据进行模型的简单验证。根据豆瓣 CEO 的回答，豆瓣的评分模型是基于部分滤波算法结合平均算法，IMDB 的评分模型是基于贝叶斯平均算法。根据猫眼电影官方发布，猫眼电影评分分为两部分，观众评分和专业评分。观众评分部分也是基于部分滤波算法结合平均算法。专业评分部分是实名制打分和评价，最终分数并列显示在观众评分旁边。专业评价体系有详细的评价指标。本文只考虑观众评分。因此可得豆瓣与猫眼的评分模型是基于平均算法，IMDB 的评分模型是基于贝叶斯平均算法。

最后，我们使用更能准确反映评价差异的威尔逊区间算法作为初步分数拟定，然后通过贝叶斯平均算法对评分进行修正补偿。

2.2 对问题二的分析

题目中要求我们对豆瓣电影和 IMND 电影的 TOP50 电影类型进行分析。目前网络上流行着对 TOP250 进行分析，因为我们将扩大数据量，对 TOP250 的数据进行分析。

首先，我们通过爬虫爬取了两个网站 TOP250 榜单的数据，并进行数据的冲洗，分类与整理以及数据的可视化。

接下来我们按照电影的年份分布与制片国家地区进行分类与统计。我们将从用户群体层面，文化层面与历史数据角度进行分类分析。

最后我们将结合我们的模型计算得出我们模型下的 TOP50 电影。

2.3 对问题三的分析

题目中要求我们建立电影票房预测模型，并预测题目给出的两部电影。

首先，一部电影包含了众多的属性，例如电影类型、导演、主演、上映时间、电影时长、故事情节等等。在一些电影中主演的流量可以带动大批粉丝进行观影，对票房有极大的促进作用。

因为电影票房受到诸多要素的影响，并且他们之间的关系难以直观的体现，因此我们使用多元线性回归算法。

线性回归模型其本质上是用一条曲线去拟合一个或者多个自变量 x 与因变量 y 之间关系的模型，若曲线是一条直线或超平面时是线性回归，否则是非线性回归。

模型方法确定之后我们需要进行数据的收集清洗与标准化，并且确定好我们一些指标的赋值。

指标的选取和赋值方法需要区分数值类与非数值类，如电影类型和电影时长。电影类型不可量化，我们可以通过一些词频或者是 TOP250 分布来量化，电影时长则是可量化的因为可以标准化之后使用。

最后我们根据各个变量的系数表进行分析得出结论。

2.4 对问题四的分析

根据问题 3 的求解，我们基于多元回归分析模型得到各个变量的系数表，我们可以根据各个信息要素线性拟合 R 方进相关性的分析。

多元回归分析模型有几个适用条件

（1）线性趋势。因变量与自变量之间存在线性关系，一般通过散点图(简单线性相关)或散点图矩阵(多元线性回归)进行简单判断。此外，残差分析还可以检验线性趋势，部分残差图是一种更专业的判断方法。如果关系明显不是线性的，则应进行变量变换校正或其他分析。

（2）独立性。因变量各观测间相互独立，即任意两个观测的残差的协方差为 0。可用 Durbin-Watson 检验是否存在自相关。

（3）正态性。对自变量的任一个线性组合，因变量均服从正态分布。此处正态分布意为对某个自变量取多个相同的值，对应的多个因变量观测值呈正态分布。这里的正态分布是指如果一个自变量取多个相同的值，那么多个因变量

的观测值就是正态分布的。但是在实际的样本中，自变量的固定值通常只有少数甚至只有 1，而相应的因变量的随机观测值只有少数甚至 1，所以没有办法直接进行调查。在模型中转换为考察残差是否符合正态分布。

（4）方差齐性。同正态分布类似，模型需要利用残差图考察残差是否满足方差齐性。方差不齐可进行加权的最小二乘法。

（5）各自变量间不存在多重共线。存在多重共线可导致结果与客观事实不符、估计方程不稳定等诸多问题。逐步回归可以限制有较强关系的自变量进入方程，如存在多重共线，可以剔除某个造成共线性的自变量，或合并自变量，也可改用领回归、主成分回归、偏最小二乘法回归等。多重共线可以利用容差、方差膨胀因子、特征根、条件指数、方差比例、相关系数以及残差图等多种方法考察。

（6）因变量为连续变量，自变量不受限制。在实际应用中，当自变量为分类变量时，可以采用最优尺度回归(分类回归)。

建立多元线性回归模型并不困难，但需要验证多元线性分析的条件以及所建立的模型是否能够最优地拟合数据。

所以我们基于多元线性回归算法，提取影响电影票房的因素，并建立变量，然后爬取数据，清洗数据，正向化数据，最后进行线性回归分析，并对结果进行分析，得出最后的结论。

三 、问题一、二的求解

3.1 现有平台评分算法分析

3.1.1 豆瓣、猫眼使用的算法简介

豆瓣电影的评分基于用户打分。每一部电影的分数，其关键因素为平均分数，但不简简单单是，抛开其他影响因素，我们将豆瓣的评分算法简化为一个单一的求均值算法。简单来说就是这么一个程序：把豆瓣用户的打分（一到五星换算为零到十分）加起来，再除以用户数。这个分数完全来自程序的计算，中间没有编辑审核，每过几分钟，程序会自动重跑一遍，以便把最新的分数加进来。

3.1.1.1 算法实现

$$\mu = \frac{x_1 + x_2 + \cdots x_n}{n}$$

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

参数符号	说明
μ	总体均值
n	总体例数
x	样本值
σ^2	总体方差
σ	标准差

表 1

3.1.1.2 算法分析

平均算法在数据量较大的情况下比较能够接近真实值，如豆瓣前 250 的电影排行榜的前几位的电影，四星以上的评价占据了百分之 90 以上。但在样本数量较少的时候数据则出现极大的不可靠性，如以下的这三条在豆瓣上的评分数据。

电影名	评价人数	5 星	4 星	3 星	2 星	1 星	评价得分
堂吉诃德 Don Quixote-Mariinsky Theater	138	69.7%	23.7%	6.6%	0.0%	0.0%	9.3
西部浪子 Nevada Smith	150	16.7%	38.9%	37.5%	6.9%	0.0%	7.3
天启 The Apocalypse	108	1.6%	4.8%	12.9%	27.4%	53.2%	3.5

表 2

但可以发现 TOP250 榜内未出现因评分人数少且评分高的电影，可见豆瓣在 TOP250 榜的电影数据有在评价人数上做限制。

同时因为平均算法不能防止一些“水军”和“黑子”的刷分操作，因此一些电影会出现评分两级分化，也就是方差会特别大。

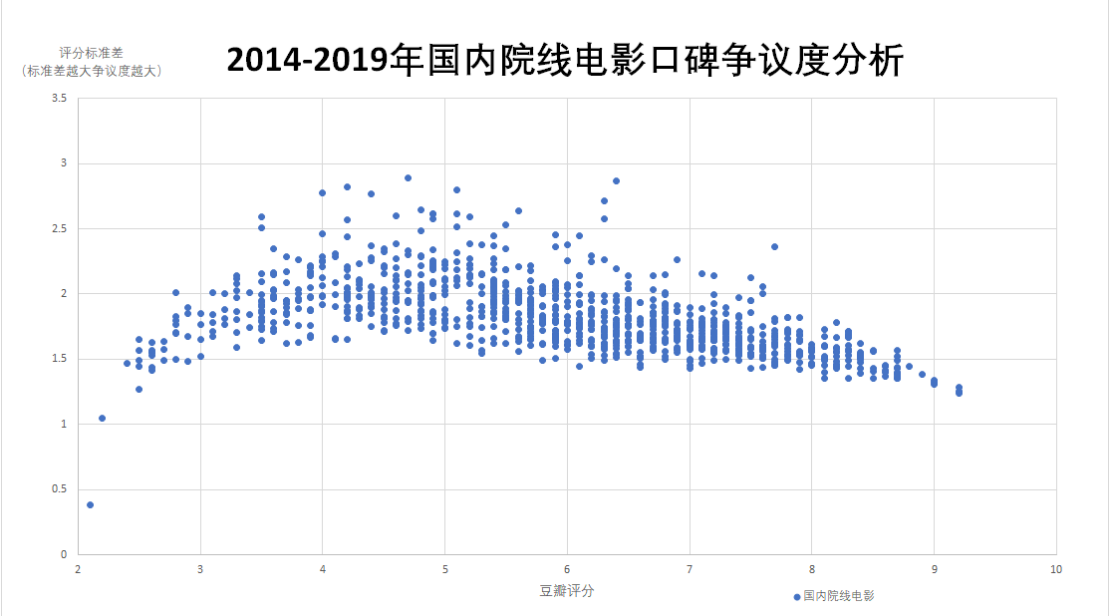

图 1

如图一，以评分标准差为纵坐标，豆瓣评分为横坐标作图，得到如上的散点图，标准差越大，代表电影的评分争议度越大。可以看出对于绝对的烂片和绝对的好片，用户的争议度很小。

3.1.2 IMDB 网站 TOP250 使用的算法简介

IMDB.COM 它所特有的电影评分系统深受影迷的欢迎，注册的用户可以给任何一部影片打分并加以评述，而网站又会根据影片所得平均分、选票的数目等计算得出影片的加权平均分并以此进行 TOP250（最佳 250 部影片）和 Bottom100（最差 100 部影片）的排行。分值系统采用 10 分制，最低为 1 分，最高为 10 分。然后再根据贝叶斯统计的算法得出加权分。

3.1.2.1 算法实现

公式如下：

$$WR = \frac{v}{v+m}R + \frac{m}{v+m}C$$

参数符号	说明
WR	加权得分
R	该电影的用户投票的平均得分
v	该电影的投票人数
m	排名前 250 名的电影的最低投票数
C	所有电影的平均得分

表 3

3.1.2.2 算法分析

利用先检概率的补偿让少量的评价产生的结果不至于达到不可靠，在评价数量增多后减少补偿值的比例，最终使得随着评价数量的增加而评分逐渐逼近实际值。 这样的做法对投票人数较少的电影进行了补偿，减小了不同电影之间投票人数的差异，使得投票人数较少的电影也可能排名靠前。当评价数据成正态分布时，得出的结果和普通平均分布所生成的结果是一样的，但其含义却很不一样。

实际分析如下：

1.IMDB 为每部电影增加了 25000 张选票，并且这些选票的评分都为 6.9。

2.假设所有电影都至少有 25000 张选票，那么就都具备了进入前 250 名的评选条件

3.假设这 25000 张选票的评分是所有电影的平均得分（即假设这部电影具有平均水准）

4.用现有的观众投票进行修正

5. 长期来看，v/(v+m)这部分的权重将越来越大，得分将慢慢接近

真实情况。

3.1.3 评分模型的客观性与真实性

首先，将 IMDb 与豆瓣进行对比，它们同采用十分制，也都来源于大众打分，存在较强的可比性。但是豆瓣采用的是 5 星制，IMDB 采用的是 10 星制，两相对比，豆瓣评分产生的误差肯定大于 IMDB.

其次，根据 DT 财经提供的 2014 年至 2019 年在中国大陆上映的电影数据，有 1128 部在两个网站上都获取了有效评分，我们将他们进行了比对

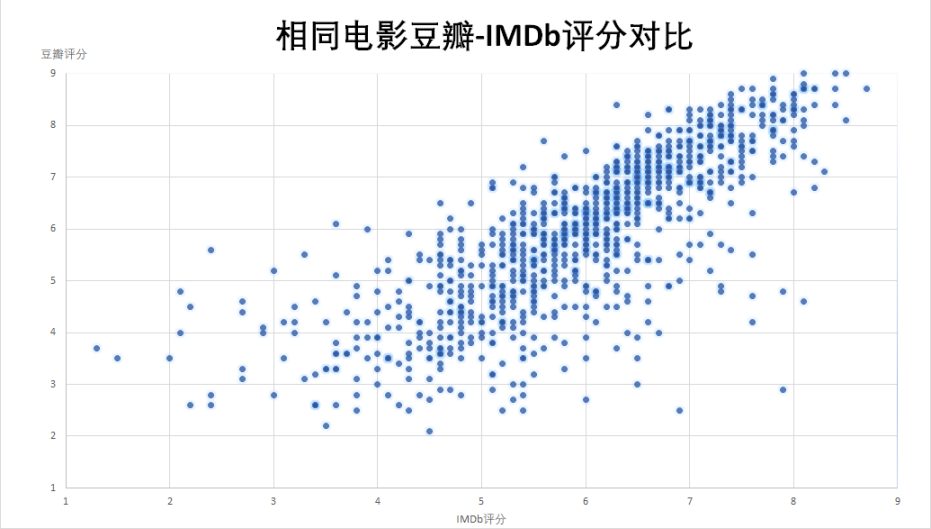

图 2

如图 2 所示，以豆瓣评分为横坐标，IMDB 评分为纵坐标制图，每个圆点代表一部电影，基本上看去，豆瓣评分越高的电影，IMDB 的评分也越高我们，进一步利用最小二乘法对两组数据进行了相关性检验，相关系数为 0.65，说明同一部电影的豆瓣评分和 IMDb 评分存在 65%左右的高度相关。

综上，对于大部分电影来说，表现的效果在两个网站上基本一致，表明两个网站的评分模型对于大部分电影来说是适用的。而对于偏离回归线较远的个别电影点，可能是由于中美的文化差异或者其他原因造成的，这里不做分析。

3.2 评分模型的建立

3.2.1 威尔逊区间算法构建评分体系

1） 威尔逊区间法是基于二项分布的一种算法，其结果和好评率和评价次数有

关，它可以解决投票人数过少、导致结果不可信的问题。能够保证排名的可信度。

2） 计算公式如下：

$$s = p_{min} \cdot S_{max},\ p_{min} = \frac{\hat{p} + \frac{1}{2n}K^2 - K\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{K^2}{4n^2}}}{1 + \frac{1}{n}K^2},\ K = z_{1-2/\alpha}$$

参数符号	说明
Smax	最大评分
p	好评率（平均值/总分值）
n	该电影的投票人数
k	统计量常数

表 4

计算步骤如下：

1. 计算每个电影评分的“好评率”
2. 计算每个“好评率”的置信区间
3. 根据置信区间的下限乘以最大评分得出对应电影评分

3） 以豆瓣的评分机制为例，用户评价打分为一星至五星对应 0-10 分。取豆瓣排名靠前的电影做统计，那么在 95%的置信水平下的评分为：

	A	B	C	D	E	F	G	H	I	J	Y	Z	AA	AB
1	电影名称	评分	评价人数	5星人数	4星人数	3星人数	2星人数	1星人数	短评数量		好评率	pmin	s威尔逊	差值
2	肖申克的救赎 The Shawshank Redemption	9.6	1353097	84.1%	14.1%	1.6%	0.1%	0.1%	265551		0.96	0.959657	9.596573	0.003427
3	霸王别姬	9.6	993975	81.3%	16.2%	2.3%	0.2%	0.1%	214889		0.96	0.9596	9.595995	0.004005
4	这个杀手不太冷 Léon	9.4	1226642	74.1%	22.4%	3.3%	0.2%	0.1%	224430		0.94	0.939565	9.395652	0.004348
5	阿甘正传 Forrest Gump	9.4	1059710	75.5%	21.1%	3.1%	0.2%	0.1%	175071		0.94	0.939529	9.39529	0.00471
6	盗梦空间 Inception	9.3	1065943	69.7%	25.7%	4.2%	0.3%	0.2%	219175		0.93	0.929501	9.295011	0.004989
7	泰坦尼克号 Titanic	9.3	999308	71.7%	23.6%	4.4%	0.3%	0.1%	154447		0.93	0.929479	9.294785	0.005215
8	千与千寻 千と千尋の神隠し	9.3	989858	69.9%	25.5%	4.3%	0.2%	0.1%	161403		0.93	0.929477	9.294772	0.005228
9	三傻大闹宝莱坞 3 Idiots	9.2	959379	67.3%	25.9%	5.8%	0.6%	0.3%	206296		0.92	0.919442	9.19442	0.00558
10	美丽人生 La vita è bella	9.5	619758	79.3%	17.9%	2.5%	0.2%	0.1%	130088		0.95	0.949432	9.494324	0.005676
11	我不是药神	9.0	1028147	56.9%	35.3%	7.1%	0.5%	0.2%	338347		0.9	0.899412	8.994118	0.005882
12	疯狂动物城 Zootopia	9.2	830619	64.1%	30.2%	5.3%	0.3%	0.1%	191286		0.92	0.9194	9.194004	0.005996
13	辛德勒的名单 Schindler's List	9.5	549988	76.8%	20.3%	2.7%	0.1%	0.1%	85127		0.95	0.949385	9.493846	0.006154
14	忠犬八公的故事 Hachi: A Dog's Tale	9.3	696461	70.9%	23.8%	4.8%	0.3%	0.1%	158059		0.93	0.929381	9.293809	0.006191
15	机器人总动员 WALL·E	9.3	710838	71.0%	24.3%	4.4%	0.2%	0.1%	124537		0.93	0.929381	9.293809	0.006191
16	海上钢琴师 La leggenda del pianista sull'oce	9.2	787260	67.4%	26.4%	5.5%	0.5%	0.2%	144252		0.92	0.919379	9.193787	0.006213
17	星际穿越 Interstellar	9.2	752167	68.9%	24.9%	5.3%	0.5%	0.4%	195515		0.92	0.919371	9.193714	0.006286
18	大话西游之大圣娶亲 西遊記大結局之仙履奇	9.2	740074	67.2%	25.9%	6.1%	0.5%	0.3%	123287		0.92	0.919356	9.193558	0.006442
19	放牛班的春天 Les choristes	9.3	658711	67.7%	27.8%	4.2%	0.2%	0.1%	112295		0.93	0.929355	9.293549	0.006451
20	楚门的世界 The Truman Show	9.2	722952	65.7%	29.2%	4.8%	0.3%	0.1%	140093		0.92	0.919352	9.193521	0.006479
21	怦然心动 Flipped	9.0	852841	58.9%	32.9%	7.6%	0.5%	0.1%	212829		0.9	0.899349	8.993493	0.006507
22	少年派的奇幻漂流 Life of Pi	9.0	776772	60.7%	31.7%	6.9%	0.5%	0.2%	194849		0.9	0.899318	8.993176	0.006824
23	龙猫 となりのトトロ	9.2	656007	64.3%	29.4%	5.9%	0.3%	0.1%	121058		0.92	0.919317	9.193172	0.006828
24	当幸福来敲门 The Pursuit of Happyness	9.0	780274	59.3%	33.1%	7.0%	0.4%	0.1%	124714		0.9	0.899309	8.993091	0.006909
25	摔跤吧！爸爸 Dangal	9.0	741030	60.1%	32.9%	6.5%	0.4%	0.2%	188959		0.9	0.899301	8.993012	0.006988
26	寻梦环游记 Coco	9.0	698989	59.9%	32.3%	7.3%	0.5%	0.1%	214791		0.9	0.899283	8.992834	0.007166
27	让子弹飞	8.7	850778	51.3%	36.5%	10.7%	1.1%	0.4%	166256		0.87	0.869268	8.692675	0.007325
28	无间道 無間道	9.1	610668	63.0%	31.1%	5.6%	0.2%	0.1%	85649		0.91	0.909245	9.092446	0.007554
29	飞屋环游记 Up	8.9	698074	56.1%	35.5%	7.9%	0.4%	0.1%	114670		0.89	0.889239	8.89239	0.00761
30	触不可及 Intouchables	9.2	506046	65.3%	29.4%	4.9%	0.3%	0.1%	122690		0.92	0.919228	9.192275	0.007725
31	流浪地球	7.9	1058032	30.5%	40.2%	22.8%	4.5%	2.0%	480724		0.79	0.789219	7.892194	0.007806
32	头号玩家 Ready Player One	8.7	726319	49.2%	37.5%	11.7%	1.2%	0.4%	221058		0.87	0.869214	8.692141	0.007859

图 3

由图 3 可以看出当评分人数较多时，威尔逊区间算出来的评分和平均算法算出来的评分基本无差值。

4）再随机抽取豆瓣评论量较少的电影数据做统计，在 95%的置信水平下采用威尔逊区间算法的评分为：

1	电影名称	评分	评价人数	5星人数	4星人数	3星人数	2星人数	1星人数	短评数量	好评率	pmin	s	差值
36506	七骗 Seven Times Lucky	6.4	30	0.0%	28.6%	61.9%	9.5%	0.0%	2	0.64	0.178278	1.782785	4.617215
36507	暴风猎人 Storm Chasers: Revenge of the Tw	5.5	31	4.3%	21.7%	34.8%	21.7%	17.4%	1	0.55	0.088027	0.880272	4.619728
36508	闪电十一人GO VS 纸箱战机W イナズマイレ	6.4	29	17.4%	21.7%	39.1%	8.7%	13.0%	2	0.64	0.177827	1.778273	4.621727
36509	人鬼情深 Saaya	7.6	36	32.0%	28.0%	32.0%	4.0%	4.0%	3	0.76	0.291154	2.911543	4.688457
36510	摇滚梦 Starstruck	6.6	32	0.0%	48.0%	40.0%	4.0%	8.0%	2	0.66	0.186682	1.866823	4.733177
36511	星期五杀手 หมาแก่ อันตราย	5.7	37	4.3%	13.0%	47.8%	34.8%	0.0%	1	0.57	0.093803	0.938026	4.761974
36512	鳄鱼威利 Wally Gator	7.8	39	18.5%	51.9%	29.6%	0.0%	0.0%	3	0.78	0.302659	3.026588	4.773412
36513	女雕刻家 The Sculptress	7.9	41	18.2%	59.1%	22.7%	0.0%	0.0%	3	0.79	0.308551	3.085507	4.814493
36514	史上最强弟子兼一：铁桥下的破坏神 史上最	6.8	45	14.3%	28.6%	39.3%	17.9%	0.0%	2	0.68	0.198529	1.985285	4.814715
36515	布莱希特的最后夏天 Abschied - Brechts let	8.0	49	29.4%	41.2%	29.4%	0.0%	0.0%	3	0.8	0.315952	3.159516	4.840484
36516	汤姆和罗拉 Tom et Lola	8.0	41	28.0%	44.0%	28.0%	0.0%	0.0%	3	0.8	0.313853	3.138528	4.861472
36517	2012: The War for Souls	6.0	95	14.3%	25.0%	21.4%	25.0%	14.3%	1	0.6	0.10733	1.073304	4.926696
36518	保卫察里津 Oborona Tsaritsyna	7.0	48	21.9%	21.9%	43.8%	9.4%	3.1%	2	0.7	0.206909	2.069088	4.930912
36519	Gone in 60 Seconds: The Ride	7.1	29	14.8%	37.0%	37.0%	11.1%	0.0%	2	0.71	0.20484	2.048395	5.051605
36520	宇宙战舰大和号2199 第五章 "望乡的银河间	8.6	66	54.3%	28.3%	13.0%	0.0%	4.3%	3	0.86	0.351003	3.510033	5.089967
36521	生命三乐章3：安魂曲 Après la vie	7.3	43	3.8%	61.5%	30.8%	3.8%	0.0%	2	0.73	0.217288	2.17288	5.12712
36522	普林歌莎 Princesa	7.4	40	20.7%	41.4%	31.0%	3.4%	3.4%	2	0.74	0.220437	2.204373	5.195627
36523	葛雷奥特曼 戈迪斯的反击 ウルトラマンG	7.5	54	28.6%	28.6%	34.3%	5.7%	2.9%	2	0.75	0.227582	2.275824	5.224176
36524	广岛 Hiroshima	8.9	43	57.1%	28.6%	14.3%	0.0%	0.0%	3	0.89	0.3644	3.644005	5.255995
36525	欲望都市：告别 Sex and the City: A Farewe	9.2	65	72.7%	18.2%	7.3%	1.8%	0.0%	3	0.92	0.384682	3.846818	5.353182
36526	Take That - The Ultimate Tour [2006]	9.2	33	73.1%	15.4%	7.7%	3.8%	0.0%	3	0.92	0.380991	3.809912	5.390088
36527	以假乱真 Machan	7.7	30	21.7%	52.2%	17.4%	4.3%	4.3%	2	0.77	0.229522	2.29522	5.40478
36528	最后的幸存者 Last Man Standing	6.8	190	11.1%	41.7%	30.6%	11.1%	5.6%	1	0.68	0.128679	1.286793	5.513207
36529	史上最强弟子兼一：来自俄罗斯的战士 史上	6.8	54	11.8%	26.5%	50.0%	11.8%	0.0%	1	0.68	0.121069	1.210687	5.589313
36530	Deacon's Dilemma	8.0	24	25.0%	50.0%	25.0%	0.0%	0.0%	2	0.8	0.240038	2.400381	5.599619
36531	福星小子 OVA 愤怒的夏贝特 うる星やつら	6.8	36	11.1%	29.6%	51.9%	3.7%	3.7%	1	0.68	0.118123	1.181226	5.618774
36532	Io non protesto, io amo	8.3	153	35.7%	46.4%	14.3%	3.6%	0.0%	2	0.83	0.26764	2.676401	5.623599
36533	Andrew Lloyd Webber: The Premiere Collect	8.4	39	59.3%	11.1%	25.9%	0.0%	3.7%	2	0.84	0.261604	2.616042	5.783958
36534	Sarah Brightman: Diva - The Video Collectio	8.9	42	57.1%	32.1%	10.7%	0.0%	0.0%	2	0.89	0.284385	2.843854	6.056146
36535	母亲的勇气 Mutters Courage	7.4	40	17.2%	41.4%	37.9%	3.4%	0.0%	1	0.74	0.132984	1.329836	6.070164
36536	Green Day: International Supervideos	9.0	36	63.0%	22.2%	14.8%	0.0%	0.0%	2	0.9	0.288199	2.881992	6.118008

图 4

由图 4 显然可以看出当评论人数较少时，威尔逊区间算法与豆瓣自身的平均值算法存在较大差异，相对于常用的平均值法，威尔逊区间算法更准确的反映了评价差异。

5）所以根据以上两种分析，综合我们得到的豆瓣 3 万条电影数据，以评分人数为横坐标，威尔逊算法得出的评分和平均算法得出的评分作为纵坐标画图 5，得到如下的对比分析图。很明显的表示出了威尔逊算法与平均值算法的差异性。

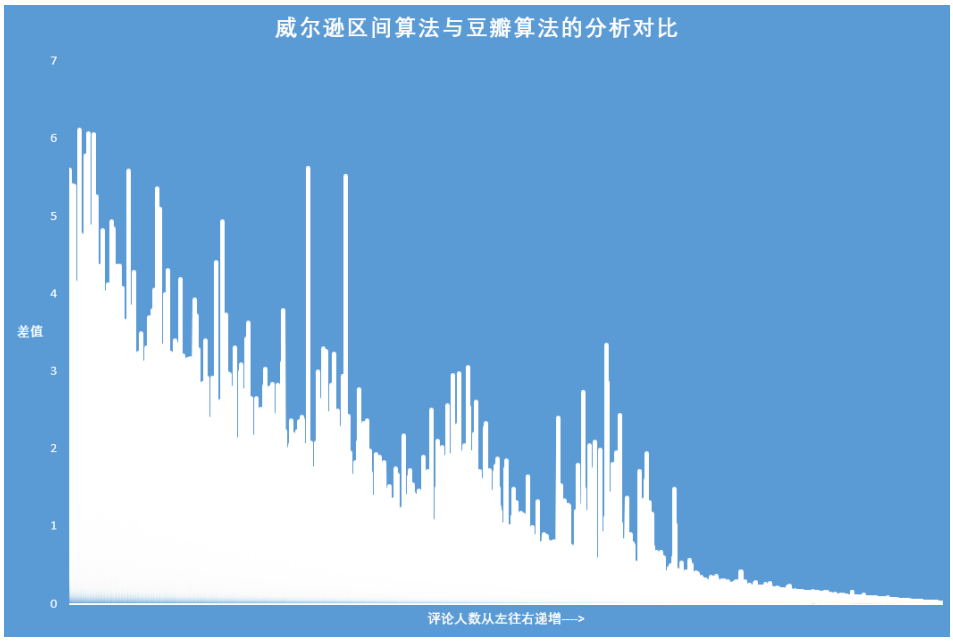

图 5

虽然威尔逊区间得出的效果较好，但不可否认的是，由于威尔逊区间的算法和评论人数挂钩，那么就会同时引入一个新的问题，举例子来说，如果只有 10 个人投票，“威尔逊区间的下限值”会将赞成票的比例大幅度拉低，排行前列的会是哪些投票人数比较多的电影而投票人数比较少的电影排名也许会长期靠后。

3.2.2 贝叶斯平均算法构建评分修正体系

由于威尔逊区间起到的作用是降低“赞成票比例”的作用，会使对应电影的得分变小、排名下降。所以对于评分数较少的电影依然存在一定的局限性，所以我们采用贝叶斯算法来起到修正的作用。根据上文中 IMDB 使用的贝叶斯算法分析来看，严格来说，贝叶斯平均法并不是一个评分模型，而是平衡模型。其核心是为冷门题目提供一个补偿值使其不至于因为少量评价而产生不可靠的评分，而在评价数量多后减小补偿值的比例，最终使得随着评价数量的增加而评分逐渐逼近实际值。

相比于 IMDB 的算法，我们根据实际的模型进行了简化。

计算公式如下：

$$\bar{x} = \frac{CM + ns}{n + C}$$

参数符号	说明
C	补偿评价数
M	评价评分
S	计算出的评分
n	评价数量

表 5

同样以豆瓣的评分机制为例，取豆瓣的 3 万条电影数据，算出的统计结果中排名靠前的电影如下图 6 所示（其中 S 为威尔逊区间算法得出的评分值，M 取 4，C 取 1000）

电影名称	评分	评价人数	5星人数	4星人数	3星人数	2星人数	1星人数	短评数量	好评率	pmin	s威尔逊	差值	贝叶斯	差值
肖申克的救赎 The Shawshank Rede	9.6	1353097	84.1%	14.1%	1.6%	0.1%	0.1%	265551	0.96	0.959657	9.596573	0.003427	9.596084	0.003916
霸王别姬	9.6	993975	81.3%	16.2%	2.3%	0.2%	0.1%	214889	0.96	0.9596	9.595995	0.004005	9.59533	0.00467
控方证人 Witness for the Prosecutic	9.6	167791	81.2%	16.9%	1.7%	0.1%	0.1%	49388	0.96	0.958999	9.589991	0.010009	9.586061	0.013939
剧院魅影：25周年纪念演出 The Ph	9.7	7695	87.6%	10.9%	1.2%	0.2%	0.2%	2257	0.97	0.964788	9.647878	0.052122	9.562891	0.137109
是，大臣 1984圣诞特辑 Yes, Minist	9.8	3397	92.6%	5.8%	0.9%	0.1%	0.6%	607	0.98	0.969687	9.696866	0.103134	9.508214	0.291786
美丽人生 La vita è bella	9.5	619758	79.3%	17.9%	2.5%	0.2%	0.1%	130088	0.95	0.949432	9.494324	0.005676	9.493296	0.006704
辛德勒的名单 Schindler's List	9.5	549988	76.8%	20.3%	2.7%	0.1%	0.1%	85127	0.95	0.949385	9.493846	0.006154	9.492688	0.007312
悲惨世界：25周年纪念演唱会 Les I	9.6	4827	81.3%	16.5%	1.9%	0.2%	0.1%	1539	0.96	0.952478	9.524779	0.075221	9.397519	0.202481
这个杀手不太冷 Léon	9.4	1226642	74.1%	22.4%	3.3%	0.2%	0.1%	224430	0.94	0.939565	9.395652	0.004348	9.395152	0.004848
阿甘正传 Forrest Gump	9.4	1059710	75.5%	21.1%	3.1%	0.2%	0.1%	175071	0.94	0.939529	9.39529	0.00471	9.394711	0.005289
新世纪福音战士剧场版：Air/真心	9.4	35010	76.7%	18.1%	4.4%	0.5%	0.4%	5262	0.94	0.936883	9.368826	0.031174	9.351595	0.048405
背靠背，脸对脸	9.4	30989	72.6%	24.6%	2.6%	0.1%	0.1%	9727	0.94	0.937047	9.37047	0.02953	9.351007	0.048993
灿烂人生 La meglio gioventù	9.4	30627	74.8%	20.0%	4.3%	0.6%	0.3%	9650	0.94	0.937029	9.370292	0.029708	9.350603	0.049397
茶馆	9.4	25800	74.1%	22.0%	3.5%	0.2%	0.1%	6358	0.94	0.93661	9.366097	0.033903	9.342796	0.057204
夏目友人帐 第六季 特别篇 铃响的	9.6	3234	81.5%	15.3%	3.0%	0.1%	0.1%	596	0.96	0.947705	9.477055	0.122945	9.294736	0.305264
盗梦空间 Inception	9.3	1065943	69.7%	25.7%	4.2%	0.3%	0.2%	219175	0.93	0.929501	9.295011	0.004989	9.294458	0.005542
泰坦尼克号 Titanic	9.3	999308	71.7%	23.6%	4.4%	0.3%	0.1%	154447	0.93	0.929479	9.294785	0.005215	9.294195	0.005805
千与千寻 千と千尋の神隠し	9.3	989858	69.9%	25.5%	4.3%	0.2%	0.1%	161403	0.93	0.929477	9.294772	0.005228	9.294176	0.005824
机器人总动员 WALL·E	9.3	710838	71.0%	24.3%	4.4%	0.2%	0.1%	124537	0.93	0.929381	9.293809	0.006191	9.29298	0.00702
忠犬八公的故事 Hachi: A Dog's Tal	9.3	696461	70.9%	23.8%	4.8%	0.3%	0.1%	158059	0.93	0.929381	9.293809	0.006191	9.292963	0.007037
放牛班的春天 Les choristes	9.3	658711	67.7%	27.8%	4.2%	0.2%	0.1%	112295	0.93	0.929355	9.293549	0.006451	9.292655	0.007345
熔炉 도가니	9.3	420120	68.6%	26.8%	4.2%	0.3%	0.1%	108278	0.93	0.9292	9.292	0.008	9.290599	0.009401
大闹天宫	9.3	159678	71.7%	22.3%	5.3%	0.4%	0.2%	17101	0.93	0.928552	9.285519	0.014481	9.281846	0.018154
福尔摩斯二世 Sherlock Jr.	9.4	8381	75.4%	21.3%	3.0%	0.2%	0.2%	2944	0.94	0.933921	9.339209	0.060791	9.269577	0.130423
城市之光 City Lights	9.3	58794	68.1%	27.1%	4.4%	0.2%	0.1%	9863	0.93	0.927609	9.276087	0.023913	9.266175	0.033825
唐顿庄园：2015圣诞特别篇 Downt	9.4	8165	76.5%	19.4%	3.7%	0.3%	0.0%	2289	0.94	0.933512	9.335119	0.064881	9.263815	0.136185
摇滚莫扎特 Mozart L'Opéra Rock	9.4	7737	76.0%	19.8%	3.5%	0.5%	0.1%	2863	0.94	0.933698	9.336977	0.063023	9.261792	0.138208
福音战士新剧场版：破 エヴァンゲ	9.3	43213	70.6%	22.8%	5.7%	0.5%	0.3%	8146	0.93	0.927201	9.272011	0.027989	9.258567	0.041433
银魂完结篇：直到永远的万事屋 劇	9.3	28196	73.2%	19.0%	6.8%	0.6%	0.4%	8404	0.93	0.926666	9.266656	0.033344	9.246158	0.053842
十二怒汉 Studio One: Twelve Angry	9.4	7958	75.5%	21.0%	3.1%	0.3%	0.2%	722	0.94	0.930069	9.300692	0.099308	9.228596	0.171404
憨豆先生精选辑 The Best Bits of Mr	9.6	2907	84.9%	12.2%	2.2%	0.3%	0.5%	328	0.96	0.942468	9.424682	0.175318	9.227414	0.372586
极品基老伴：完结篇 Vicious Series	9.3	16425	72.1%	22.4%	4.9%	0.4%	0.3%	4227	0.93	0.925372	9.253721	0.046279	9.218931	0.081069
空之境界 第五章 矛盾螺旋 劇場版	9.3	16700	70.2%	24.2%	4.9%	0.5%	0.3%	2566	0.93	0.924859	9.248589	0.051411	9.214437	0.085563
夏目友人帐 五 特别篇：一夜酒杯	9.5	3651	77.1%	19.1%	3.3%	0.2%	0.2%	520	0.95	0.936518	9.365182	0.134818	9.209303	0.290697
三傻大闹宝莱坞 3 Idiots	9.2	959379	67.3%	25.9%	5.8%	0.6%	0.3%	206296	0.92	0.919442	9.19442	0.00558	9.193831	0.006169
疯狂动物城 Zootopia	9.2	830619	64.1%	30.2%	5.3%	0.3%	0.1%	191286	0.92	0.9194	9.194004	0.005996	9.193324	0.006676

图 6

仔细分析我们使用的公式，我们为每一部电影都增加了一个固定的选票 C，同时这些选票的评分都为我们的评分补偿值 M。这样做一个很显著的效果就是拉近了不同电影之间投票人数的差异。然后在公式中，我们使用经过皮尔逊区间算法算出的评分 S 进行修正，长期来看 S 这部分的权重将越来越大，得分会慢慢接近于真实的情况。

3.2.3 模型建立步骤

①：根据电影的数据（用户打分）利用威尔逊区间算法算出在一定置信区间下威尔逊置信区间的下限值。

②：将评分的最大值乘以利用威尔逊区间算出的下限值，得到该电影的初步评

分 S。

③：拟定贝叶斯的补偿评价数 C 和补偿评分 M，根据贝叶斯平均算法算出修正后的电影评分值。

至此我们建立了关于网站电影评分模型，如图 7

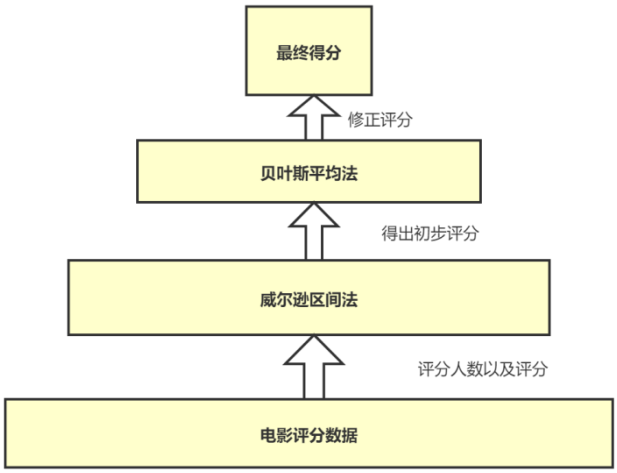

图 7

3.3 模型的结论与修正

3.3.1 模型结论

1） 在我们爬取的豆瓣 9 万多条电影数据中随机抽取 3 万条做验证，得到的对比效果如图 8

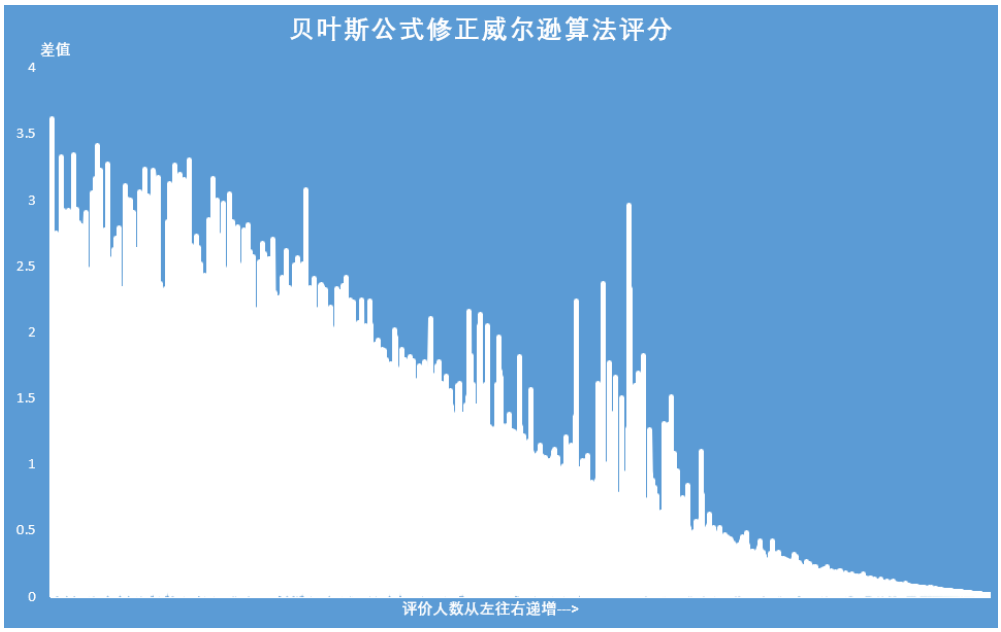

图 8

由上图可知，经过修正后的评分与原来威尔逊区间算法得出的评分相对于平均值算法的差值明显减小，随着评分人数的增加，模型的真实性越来越高。随着时间线的推移，模型会不断对评分分数进行修正，使评分的真实性不断上升。

2）由于 IMDB 的爬取数据时，电影的具体信息存在 URL 的不断跳转，因此爬取效率十分低下，我们收集到的数据样本较少，这里我们取爬取到的 TOP250 为例，验证我们的模型与 IMDB 模型对于电影排行前列的影响，得到的效果如下图

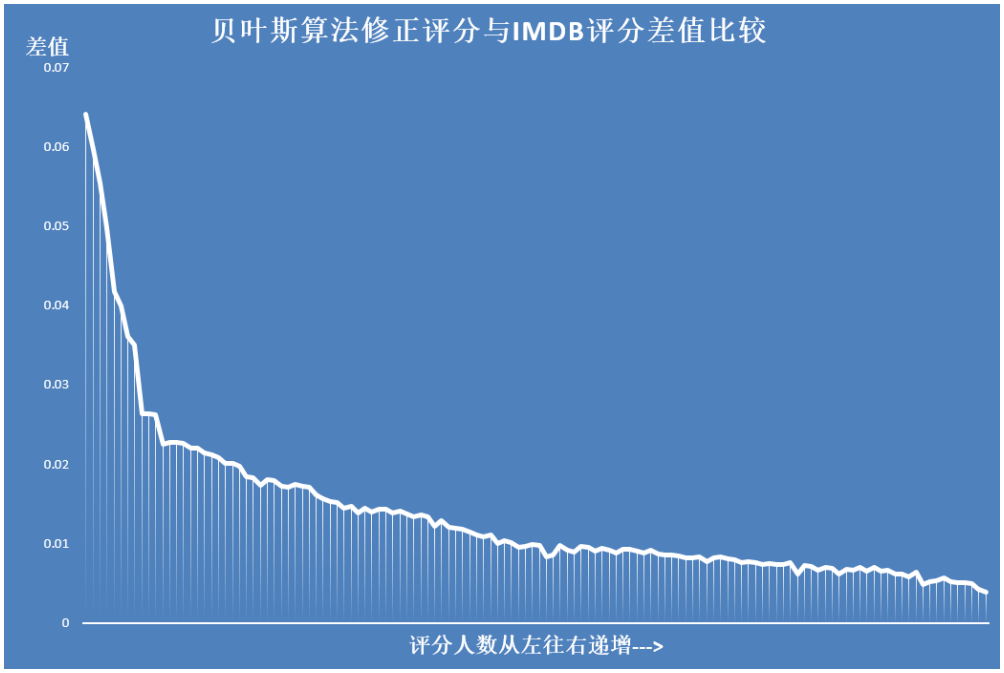

图 9

由此可以看出，对于公认的好电影，我们建立的模型得到的评分与 IMDB 的算法得出来的评分差值很小，说明我们的模型对于排名靠前，评价人数较多的电影具有非常好的适应性。

3.3.2 模型修正

1）豆瓣采用 5 星 10 分制，IMDB 采用 10 星 10 分制，对于我们模型来说，采用评价划分更为细致的 10 星 10 分制作为计算威尔逊评分的入口值更合适。

2）尽管使用威尔逊区间算法改善了对于那些评论数较少引起的评分不可信问题，但是在豆瓣 TOP250 的排名中我们会发现，采用威尔逊区间-贝叶斯平均后得到的结果中，会存在一些评分人数相比于其他电影较少的电影，但是排名却比较高的情况，原因是对于某些有特殊意义的电影，真爱粉会集中拉高此电影的分数，并且评论人数也不是特别低，因此针对此种情况，需要对靠前榜单中的电影进行滤波，设定一个进入能进入 TOP250 的最低评分人数。

3.4 豆瓣与 IMDB 排名影响因素分析

3.4.1 差异性描述

根据网络数据分析，我们可以发现优秀的欧美电影在两个排行榜中都能出现，如肖申克的救赎，星际穿越，指环王 3。但一些优秀的亚洲电影，如中国国产与港产电影与日本的动漫电影，大部分只出现在豆瓣的 TOP250 里。

3.4.2 豆瓣与 IMDBTOP250 数据对比

3.4.2.1 发行年份分布

根据图 10，我们可以看到大部分的豆瓣电影是分布在 90 年代之后。年代比较久远的电影上榜的数量较少

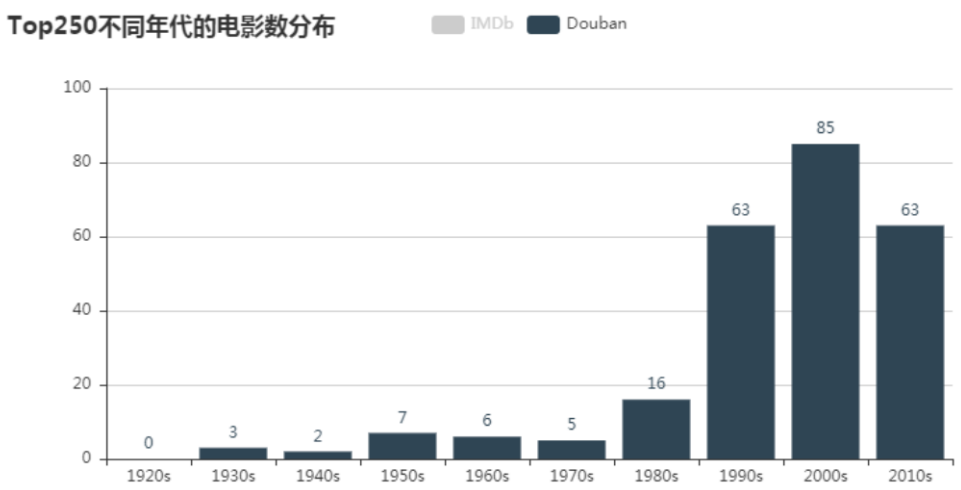

图 10

根据图 11，我们可以看到 IMDb 电影总体上也大部分分布于 90 年代之后，但对一些年代比较久远的电影也有收录。数量比豆瓣多很多。

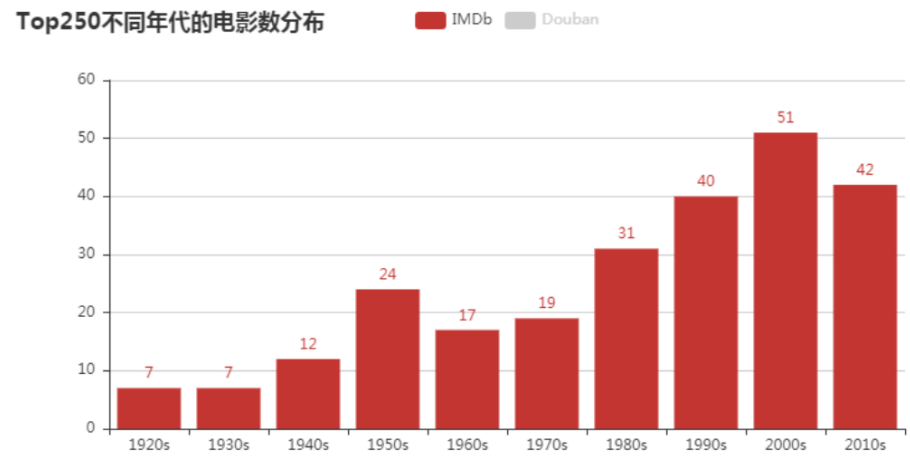

图 11

3.4.2.2 国家地区分布

根据图 12,我们可以得出由制片国家分布来看，2 个榜单中美国电影占比最高。

在豆瓣榜单中，中国电影占比较高；在 IMDb 榜单中，只有 1 部香港王家卫的《In the Mood of Love》(《花样年华》)。

上面统计数据中，豆瓣的中国电影数有 35 部，包括了港台。

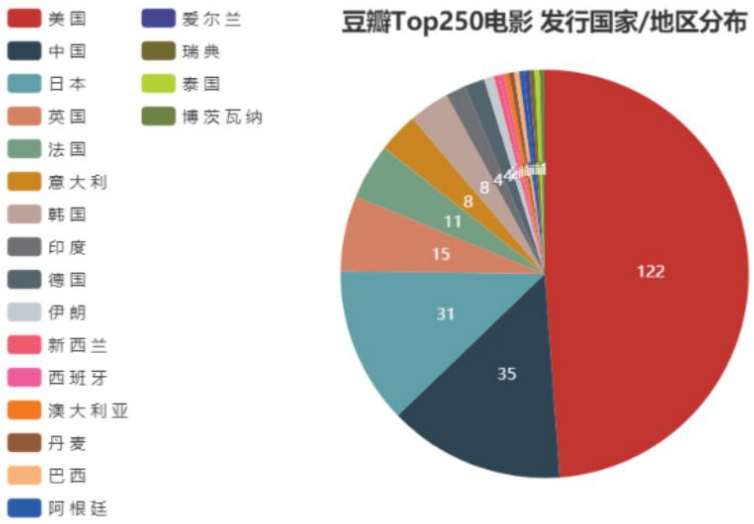

图 12

3.4.3 影响因素分析

3.4.3.1 用户群体

从用户群体的居住地域分析，豆瓣网的用户几乎全是中国的用户，因此对一些本土的电影关注会更多，因此一些优秀的中国电影能够得到用户的好评，并上榜。

但 IMDb 面向的是全球范围内的用户，用户的母语主要是英语，同时还能支持多种其他语种。在使用体验上会优于国人。并且对亚洲的电影关注并没有豆瓣用户高。所以这是产生差异的原因之一。

3.4.3.2 文化差异

对于一些电影所传达的价值观与文化表达，不同人群接受的程度是不一样的，正是因为用户群体有着中西方明显的差异，所以在一些电影鉴赏的过程中得到的反馈也不相同。

3.4.3.3 电影营销

对电影来说，本身就是一种大众传播的工具；而电影之所以存在或存在得

好不好的支点，完全视它与观众之间的互动关系是否顺畅良好，因而营销传播做的好不好，会直接影响影片传播。由于语言和地域差异，中国观众们对豆瓣的接受程度较高就并不足为奇。同时在网站宣传方面，豆瓣面向中国受众的宣传也会比面向全球受众的宣传要充分和专业。

3.5 评分模型下的 TOP50

3.5.1 PS:具体数据祥见附录

四、 问题三、四求解

4.1 电影票房预测模型的建立

4.1.1 研究方法介绍

电影票房收多个因素的影响，因此我们进行多元回归分析，我们把包括两个或两个以上自变量的回归称为多元线性回归 。

多元线性回归的基本原理和基本计算过程与一元线性回归相同，但由于自变量个数多，计算相当麻烦，一般在实际中应用时都要借助统计软件。这里只介绍多元线性回归的一些基本问题。

但由于各个自变量的单位可能不一样，比如说一个消费水平的关系式中，工资水平、受教育程度、职业、地区、家庭负担等等因素都会影响到消费水平，而这些影响因素（自变量）的单位显然是不同的，因此自变量前系数的大小并不能说明该因素的重要程度，更简单地来说，同样工资收入，如果用元为单位就比用百元为单位所得的回归系数要小，但是工资水平对消费的影响程度并没有变，所以得想办法将各个自变量化到统一的单位上来。前面学到的标准分就有这个功能，具体到这里来说，就是将所有变量包括因变量都先转化为标准分，再进行线性回归，此时得到的回归系数就能反映对应自变量的重要程度。这时的回归方程称为标准回归方程，回归系数称为标准回归系数，表示如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

4.1.2 影响因素的描述

影响电影票房的因素有很多，电影类型，电影故事梗概，品牌价值，明星效应，导演效应，电影档期，发行公司，宣传力度，电影评分，因为电影故事梗概无论是过去还是现在都难以用机械的标准去科学得衡量，因此从客观性考虑本次模型将不考虑故事梗概。

品牌价值也是难以统一衡量，这考虑到一个电影公司的出品，但会存在一些少量的小型电影公司投资拍摄出高质量票房高的电影，并且难以对电影公司的等级进行过评定，因此本次模型中也不考虑该指标。

因此综合考量，我们将采用电影类型，明星效应，导演效应，电影档期，电影时长，评论人数，电影评分，制作技术八个指标来进行建模。

指标	变量名
电影类型	x1
制片公司	x2
导演效应	x3
电影档期	x4
电影时长	x5
制作技术	x6
评论人数	x7
电影评分	x8

表 6

4.1.3 影响因素的赋值

4.1.3.1 电影类型

电影类型方面我们根据豆瓣票房 TOP300 的类型进行整理如图 13，可以看出电影类型的分布规律。因为一部电影可能符合多种电影类型，假如用词频作为

变量值，这样会到时误差比较大，因此我们将 x1 变量拆分为一个多元向量。

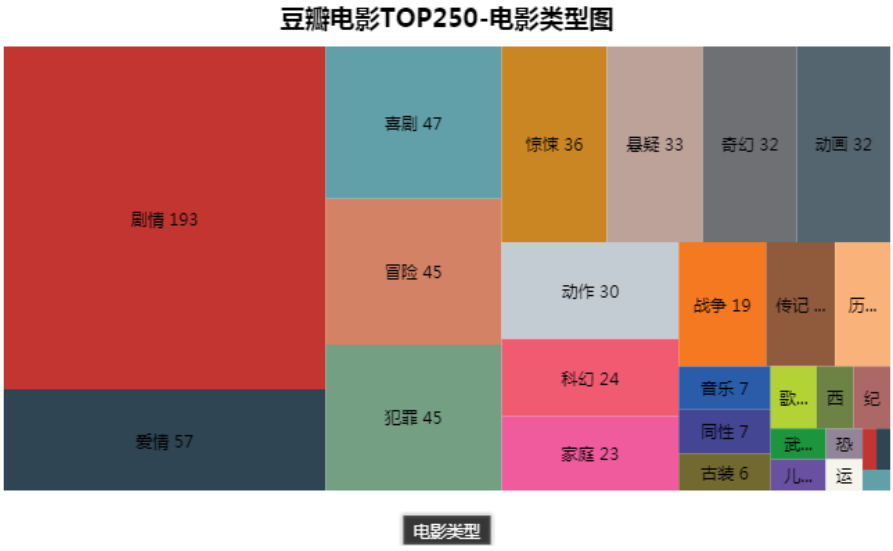

图 13

本模型将影片类型分为 17 类。每一个类型都设为一个虚拟变量，动画 x1.1、记录 x1.2、歌舞 x1.3、犯罪 x1.4、历史 x1.5、幻想 x1.6、体育 x1.7、西部或近代 x1.8、动作 x1.9、冒险 x1.10、恐怖 x1.11、推理 x1.12、剧情 x1.13、爱情 x1.14、家庭理论 x1.15、喜剧 x1.16、战争 x1.16、战争 x1.17。如表 7……豆瓣 top250 电影类型的数据透视表

类型	剧情	爱情	喜剧	犯罪	冒险	惊悚	悬疑	动画	奇幻	动作	科幻	家庭	战争	传记	历史
词频	193	58	46	45	45	36	33	32	32	30	24	23	19	15	12

类型	同性	古装	音乐	歌舞	纪录	儿童	武侠	恐怖	运动	情色	灾难
词频	8	6	6	5	4	3	3	2	2	2	1

表 7

- 这里规定一部电影符合的类型最多不大于 5 个。
- 如《星际穿越》符合剧情,科幻,悬疑,家庭,冒险，则得分 295
- 如《迁徙的鸟 Le peuple migrateur》符合纪录片，则得分 4

4.1.3.2 制片公司

电影的值作、发行、放映是电影产业链中最为重要的三个环节。电影的制

作发行。放一次电影产业链中最为关键的三个关键环节，处于中间环节的电影发行更是重中之重，不仅承担着承上启下的衔接配合作用，更是作为投资成本获取实际利益利润的重要途径。因为一些中小型公司上映的作品数量非常有限。也会在票房上显示。因此，我们根据。猫眼官方发布的数据。排出了发片前十的公司。

电影的制作、发行、放映是电影产业链中最为重要的三个关键环节。处于中间环节的电影发行更是重中之重，不仅承担着承上启下的衔接配合作用，更是收回投资成本、获取实际利润的重要途径。电影发行有三种主要的工作：一是与制片方、放映方进行洽谈、对接和结算等工作；二是进行调度和发放影片等工作；三是对影片进行宣传及推广等工作[49]。随着电影营销水平的不断提高，电影发行的工作也在不断扩大，包含了放映范围、档期选择、宣传策略、分账比例等内容，将电影产品的商业价值在各个节点内实现利益最大化，从而延长电影的生命长度并扩展利润空间。随着互联网的迅速发展，以淘宝、猫眼、支付宝、大众点评网为代表的互联网企业大批进军电影行业，与电影的制作方、发行方、放映方开展各种线上虚拟与线下实体形式的合作，将线上虚拟与线下实体的有效结

合，对电影发行产生深刻的影响。

这次根据猫眼电影总结出排名前十电影公司。若电影的出片有这几个公司，则 x2=1。

名称	公司
1	中国电影股份有限公司
2	华夏电影发行有限公司
3	北京光线传媒股份有限公司
4	华谊兄弟传媒股份有限公司
5	五洲电影发行有限公司
6	万达映射传媒有限公司
7	博纳影业集团股份有限公司
8	乐视影业（北京）有限公司
9	中影数字电影发展（北京）有限公司
10	福建恒业影业有限公司

表 8

4.1.3.3 导演效应

优秀的导演往往作品都会受人期待，导演的工作是把剧本通过一些列的艺术手法来塑造荧幕形象。

衡量导演效应可以通过以下三种方法进行衡量

第一种是通过奖项进行衡量如中国电影金鸡奖，大众电影百花奖、中国电影华表奖、香港电影金像奖、中国电视剧飞天奖、中国电视金鹰奖等等。通过获奖一次将值加一可以衡量一个导演的成就，但由于数据种类多，多多奖项之间的价值比较没有明确体系的衡量，加上数据不全所以将不采取这种方法。

第二种是将导演效应设为数值变量，根据专业电影网站上观众对导演的评分值作为导演效应的值。但由于这类数据影响力较小，数据量不大，因此也不采取这种方法。

第三种是根据导演最佳作品的平均分，暂未将商业价值潜力、艺术影响力、社会责任、公众形象等等要素计入。

本次模型将采用第三种模式。

4.1.3.4 电影档期

美国电影五大档期：

1.冬季档期：新年的第一个周五开始至春假前。

2.春季档期：春假开始至亡将士纪念日周末为止。

3.夏季档期：从每年 5 月下旬至 9 月 4 日为止，它的时间跨度长，又正值暑期，自然是所片商最为重视的档期。

4. 秋季档期：9 月 5 日起至感恩节前，其目标观众多为女性

5.岁末假日档期：从感恩节前的一个周末直到新年，虽然没有暑期档时间那么长，但此档 因为横跨感恩节和圣诞节两大节日，也是片商必争档期。

此外还有：新年、3 月复活节、5 月阵亡将士纪念日、7 月独立纪念日、9 月劳工节、10 月哥伦布发现新大陆纪念日、11 月的感恩节、12 月的圣诞节

美国的暑期档是电影营销者通过各种手段营造出来的。

中国电影经过多年的发展已经慢慢形成了以下几个电影档期，中国电影业最大的档期是贺岁档，其次是暑期档。中国特色电影档期：

1、贺岁档，贺岁档泛指每年 11 月初到次年 3 月初的电影档期，大约在八九十天左右。

2、五一档：一般泛指每年五一期间的电影档期。

3、暑期档：一般泛指每年 6 月-9 月的电影档期。

4、国庆档：一般泛指每年国庆期间的电影档期。

其中含有情人节档、三八档、清明档、愚人档、端午档、七夕档、光棍节档、双 12 档。

本次模型以中国档期为例子若上映的电影能在其中的一个档期内则赋值 1

若一部电影能在两个档期内上映则 x4 值为 2。

因为考虑中国春节人民消费强度加强，所以春假元旦的变量值加 1。即若有一电影是在春节档期内上映，则 x4=2。

档期	时间定义	赋值
元旦	1.1-1.3	2
春节	1.27-2.2	2
清明	4.4	1
端午	5.1-5.3	1
中秋	5.28-5.30	1
国庆	10.1-10.7	1
白色情人节	2.14	1
妇女节	3.8	1
儿童节	6.1	1
暑期档	30-33 周	1
光棍节	11.11	1
双十二档	12.12	1
圣诞节	12.25	1

表 9

4.1.3.3 电影时长

时间，时长短了观众觉得仍有余力不过瘾，时长长了会让人不耐烦、感觉疲惫。短了会让人意犹未尽。但通过数据显示一些高分电影普遍电影的时长会稍长一些。未了验证这一影响因素，我们决定把电影时长也作为其中一个影响因素进行分析。

电影时长已经是一个数值类型，不需要做特别的转换，我们将采用分钟模式下的电影时长对 x5 进行赋值。如《肖申克的救赎》x5 的值为 142。

4.1.3.4 制作技术

3D 技术。由于人们生活水平的提高，消费者在看电影 的同时，开始追求更多社交等方面的需求，注重环境的优雅、服务的优良和消费高品位。到拥有豪华空间、声光效果更佳的豪华影院看电影，已经成为了一种时尚消费，一种品味的表现。这种消费者需求的变化与影院经营理念的改变使得 3D 电影成为电影的一大卖点。

IMAX 电影技术为观众提供了比普通影片更广阔的视野，更好的体验。3D+IMAX 就是两种技术同时运用。两种技术的结合可能会有 1+1>2 的效果，也有可能是 1+1<2 的效果。这需要我们去验证。所以 X6 的赋值规律是，若使用 IMAX 技术或 3D 技术则加 1。

若同时使用 IMAX 技术和 3D 技术 x6=2

若没使用任何新技术则 x6=0

若使用了其中的任一中技术则 x6=1

4.1.3.5 评论人数

豆瓣评分、IMDB 评分、微博话题讨论、猫眼“想看”人数，这些反应电影受关注程度会对票房有一定的影响

评价人数可以反应一部电影的受关注程度。因此评价人数我们可以参考豆瓣评分人数作为我们的参考指标。

因为评分人数已经是一个数据值，所以 x7 的取值等于评价人数值。

4.1.3.6 评论分数

电影的评分在一定程度上反映电影的质量和口碑。如今正处于互联网飞速发展的时代，中国的电影市场已经进入了口碑时代，中国电影观众也非常关注电影产品的质量，

观看电影的大众，大部分都会去考虑这部电影的影评分数，来判断这部电影是否值得一看。 因因此我们 x8 的取值为豆瓣和 IMDB 评价分数的平均值。如《肖申克的救赎》豆瓣 9.7，IMDB9.3，所以 x8 取值为 9.5。

4.2 样本获取

4.2.1 样本数据收集

因为猫眼票房的数据在疫情期间暂时进行了封闭，所以我们只能从其他网站爬取数据，本次数据来源我们爬取了国外的电影数据库网，The Movie Database（TMDb）。

我们通过装入数据集，然后通过 json 工具包进行处理，把每一条数据拆分成拥有名字、评分、票房、导演等属性的数据。

同时我们其他因素的数据来源则是微博公开数据，百度指数数据、豆瓣数据。

4.2.2 样本数据处理

我们以建立数据集后，对数据进行了去重、填充等数据清洗操作。同时，我们进行了电影的聚类统计，最终将影片类型分为 17 类。同向比较了电影主角影响力，导演影响力对电影票房的正相关影响。我们环向统计了美国五大档期，中国四大档期下各电影票房的排名。

4.2.3 样本数据分析

我们通过 IBM spass21 软件的多元回归分析的相关计算，为了更为精确的研究评价量的重要因素（去除量纲的影响）我们考虑使用标准化回归系数。标准化系数的绝对值越大，说明对因变量的影响就越大。那么我们根据 SPISS 计算得出的回归线性方程组为:

$$y=0.389x1+0.394x2-0.308x3-0.058x4-$$
$$0.038x5+0.134x6+0.375x7+0.128x8-30.947$$

系数 a						
模型		未标准化系数		标准化系数	t	显著性
		B	标准误差	Beta		
1	(常量)	-30.947	17.727		-1.746	.088
	日期	5.676	1.975	.389	2.875	.006
	导演	6.485	2.612	.394	2.483	.017
	发行公司	-6.476	3.008	-.308	-2.153	.037
	评分	-.820	2.330	-.058	-.352	.727
	电影时长	-.021	.083	-.038	-.250	.803
	类型	.021	.021	.134	.988	.329
	技术	3.771	1.437	.375	2.624	.012
	评分人数	.000	.000	-.128	-.956	.345

a. 因变量：票房/亿

表 10

4.2.4 公式与系数解释

通过以上的线性公式以及 SPSS 计算结果可以看过，豆瓣电影评分和豆瓣网站提供的各个信息要素线性。拟合程度 R2 达到 0.254，也就是说票房的 25.4%可以用该模型来解释。根据分析系数，我们可以看到和票房成正相关的因素是档期、导演影响、类型、和电影技术。

首先，在我们算出来的未标准化参数中，评分和电影时长的显著性远远高于 0.05，说明对于在我们分析的数据中，评分和发行公司对票房的影响不具有明显的统计学意义，所以本模型暂不考虑这两个系数。

其次，这里可以看到发行公司系数为-3.008，实际上意味着发行公司和票房有巨大的影响，但是设定一套统一的发行公司赋值方法成为关键。因为没有

统一的计算方法所以也导致本次模型设定该参数超出我们的预期。

显然的，值得一提的是电影技术与导演效应。在我们生活中我们知道 3D 电影和 IMAX 电影是能带给我们更好的观影体验。人们更加愿意多付一部分钱去获取更好得观影体验。这个现象与观点在本次模型中得到了验证。

众所周知，导演的影响也是如此。一些受人崇拜的导演会有自己相对应的粉丝，如“诺兰粉”，在电影上映之前，会有很多人慕名前去观看，不过最后电影的作品好与坏，都能吸引到人们为此买单。

最后，电影评分与电影时长在系数上体现的就是对票房影响不明显，所以这也解释了为什么出现高分电影底票房和低分电影高票房的情况。

4.3 影响票房的最重要的因素

基于多元回归线性分析结果，在我们构建的模型下，影响电影票房最重要的因素是电影制片导演。

也就是说，在本次分析中，导演效应的显著性和相关性最高。

4.4 模型的结论与评价

模型		平方和	自由度	均方	F	显著性
1	回归	1997.504	8	249.688	3.084	.008b
	残差	3319.147	41	80.955		
	总计	5316.651	49			

a. 因变量：票房/亿

b. 预测变量：(常量)，评分人数，发行公司，日期，导演，类型，技术，电影时长，评分

表 11

由表 10 可以看出 ，这里的显著性为 0.008，小于 0.05，也就是由自变量电影的八个特征和因变量“电影实际票房”建立的线性关系回归模型的统计学意义较为明显。

模型摘要				
模型	R	R 方	调整后 R 方	标准估算的误差
1	.613a	.376	.254	8.99749

a. 预测变量：(常量)，评分人数，发行公司，日期，导演，类型，技术，电影时长，评分

表 12

由于猫眼专业版的电影票房数据被屏蔽了，我们得到的数据相对局限，因此我们的模型是基于豆瓣统计的票房 TOP300 来统计分析的，而在票房排行榜前列的电影影响他们票房的特征往往都具有相似性以及整体的局限性，所以当样本数据较少时，分析出的模型具有较大的误差和聚类性，这也就是我们的模型出现拟合优度较低的原因。

为了能普遍的反映电影特征和电影票房之间的适应关系，需要通过大样本来进行特征性矫正，通过分析大样本中我们给出的八个特征值对于电影票房的相关性，这样综合对比才能建立出一个良好基于多元线性回归的票房预测函数。

4.5 Nolan 的《信条 Tenet》和《兰心大剧院》票房预测

1）《信条 Tenet》

档期	导演	发行公司	评分	时长	类型	技术	评分人数
1	9.5	1	7.1	132	223	4	123516

预计票房 y=58.2325 亿

2）《兰心大剧院》

档期	导演	发行公司	评分	时长	类型	技术	评分人数
1	6.5	1	7.8	126	193	1	1129

预计票房 y=21.8505 亿

五、 参考文献

[1]从豆瓣电影评分算法说起，聊聊排名算法
https://zhuanlan.zhihu.com/p/20683599
[2]基于用户投票的排名算法（六）：贝叶斯平均
http://www.ruanyifeng.com/blog/2012/03/ranking_algorithm_bayesian_average.html
[3]如何不按平均分排序
https://www.evanmiller.org/how-not-to-sort-by-average-rating.html
[4]平珊珊. (2019). 我国电影票房影响因素分析及预测. 社会科学前沿, 008(004), P.489-495.
[5]裴培, & 蒋垠茏. (2014). 国内外电影票房影响因素分析. 合作经济与科技(02), 20-23.
[6]任丹. (2015). 基于多元线性回归模型的电影票房预测系统设计与实现. (Doctoral dissertation).
[7]汤子涵. (2019). 基于多元回归和神经网络的我国电影票房的研究. (Doctoral dissertation).
[8]焦亚男. (2018). 豆瓣评分与电影票房关系探析. 青年记者.

六、 附录

6.1 评分模型下的 TOP50

6.1.1 豆瓣 TOP50

电影名称	评分	评价人数	好评率	pmin	威尔逊	贝叶斯修正
肖申克的救赎 The Shawshank Redemption	9.6	1353097	0.96	0.959657	9.596573	9.596084
霸王别姬	9.6	993975	0.96	0.9596	9.595995	9.59533
美丽人生 La vita è bella	9.5	619758	0.95	0.949432	9.494324	9.493296
辛德勒的名单 Schindler's List	9.5	549988	0.95	0.949385	9.493846	9.492688
这个杀手不太冷 Léon	9.4	1226642	0.94	0.939565	9.395652	9.395152
阿甘正传 Forrest Gump	9.4	1059710	0.94	0.939529	9.39529	9.394711
盗梦空间 Inception	9.3	1065943	0.93	0.929501	9.295011	9.294458
泰坦尼克号 Titanic	9.3	999308	0.93	0.929479	9.294785	9.294195
千与千寻 千と千尋の神隠し	9.3	989858	0.93	0.929477	9.294772	9.294176
机器人总动员 WALL·E	9.3	710838	0.93	0.929381	9.293809	9.29298
忠犬八公的故事 Hachi: A Dog's Tale	9.3	696461	0.93	0.929381	9.293809	9.292963
放牛班的春天 Les choristes	9.3	658711	0.93	0.929355	9.293549	9.292655
熔炉 도가니	9.3	420120	0.93	0.9292	9.292	9.290599
三傻大闹宝莱坞 3 Idiots	9.2	959379	0.92	0.919442	9.19442	9.193831

疯狂动物城 Zootopia	9.2	830619	0.92	0.9194	9.194004	9.193324
海上钢琴师 La leggenda del pianista sull'oceano	9.2	787260	0.92	0.919379	9.193787	9.193069
星际穿越 Interstellar	9.2	752167	0.92	0.919371	9.193714	9.192962
大话西游之大圣娶亲 西遊記大結局之仙履奇緣	9.2	740074	0.92	0.919356	9.193558	9.192795
楚门的世界 The Truman Show	9.2	722952	0.92	0.919352	9.193521	9.192739
龙猫 となりのトトロ	9.2	656007	0.92	0.919317	9.193172	9.192311
触不可及 Intouchables	9.2	506046	0.92	0.919228	9.192275	9.191159
教父 The Godfather	9.2	484682	0.92	0.919195	9.191951	9.190786
活着	9.2	394431	0.92	0.919114	9.191143	9.189712
天堂电影院 Nuovo Cinema Paradiso	9.2	389047	0.92	0.919102	9.191023	9.189572
乱世佳人 Gone with the Wind	9.2	359630	0.92	0.919068	9.19068	9.189111
鬼子来了	9.2	323798	0.92	0.919013	9.190131	9.188389
辩护人 변호인	9.2	304899	0.92	0.918994	9.189942	9.188092
无间道 無間道	9.1	610668	0.91	0.909245	9.092446	9.09156
蝙蝠侠：黑暗骑士 The Dark Knight	9.1	491052	0.91	0.909163	9.091627	9.090526
天空之城 天空の城ラピュタ	9.1	457971	0.91	0.909116	9.091161	9.089981
指环王 3：王者无敌 The Lord of the Rings: The Return of the King	9.1	389293	0.91	0.909026	9.090265	9.088877
两杆大烟枪 Lock,	9.1	346569	0.91	0.909	9.090004	9.088445

Stock and Two Smoking Barrels						
飞越疯人院 One Flew Over the Cuckoo's Nest	9.1	346577	0.91	0.909	9.089997	9.088438
窃听风暴 Das Leben der Anderen	9.1	308929	0.91	0.908946	9.08946	9.087712
我不是药神	9	1028147	0.9	0.899412	8.994118	8.993614
怦然心动 Flipped	9	852841	0.9	0.899349	8.993493	8.992887
少年派的奇幻漂流 Life of Pi	9	776772	0.9	0.899318	8.993176	8.99251
当幸福来敲门 The Pursuit of Happyness	9	780274	0.9	0.899309	8.993091	8.992428
摔跤吧！爸爸 Dangal	9	741030	0.9	0.899301	8.993012	8.992315
寻梦环游记 Coco	9	698989	0.9	0.899283	8.992834	8.992094
罗马假日 Roman Holiday	9	522263	0.9	0.89915	8.991505	8.990516
搏击俱乐部 Fight Club	9	502473	0.9	0.89914	8.991402	8.990374
哈尔的移动城堡 ハウルの動く城	9	504070	0.9	0.89913	8.991301	8.990276
闻香识女人 Scent of a Woman	9	441999	0.9	0.89908	8.990798	8.98963
指环王 1：魔戒再现 The Lord of the Rings: The Fellowship of the Ring	9	409312	0.9	0.899006	8.990062	8.988801
死亡诗社 Dead Poets Society	9	377121	0.9	0.899005	8.99005	8.988681
狮子王 The Lion King	9	390862	0.9	0.898985	8.989854	8.988533
指环王 2：双塔奇兵	9	365426	0.9	0.898917	8.98917	8.987758

The Lord of the Rings: The Two Towers						
音乐之声 The Sound of Music	9	306340	0.9	0.898849	8.988492	8.986809
飞屋环游记 Up	8.9	698074	0.89	0.889239	8.89239	8.891683

6.1.2 IMDB 网站 TOP50

电影名	评分	评分总人数	好评率	威尔逊算法	贝叶斯修正
The Shawshank Redemption	8.5	22,523	0.85	8.45277	8.43574
The Godfather	8.3	25,546	0.83	8.25344	8.24027
The Godfather: Part II	8.3	28,764	0.83	8.25615	8.24441
The Dark Knight	8.5	33,800	0.85	8.46153	8.45009
12 Angry Men	8.3	46,461	0.83	8.26557	8.25823
Schindler's List	8.2	50,481	0.82	8.16624	8.15996
The Lord of the Rings: The Return of the King	8.2	60,242	0.82	8.16912	8.16383
Pulp Fiction	8.2	63,452	0.82	8.16991	8.16489
Il buono, il brutto, il cattivo	8.4	101,418	0.84	8.37731	8.37367
The Lord of the Rings: The Fellowship of the Ring	8.2	104,923	0.82	8.17664	8.17358
Fight Club	8.2	106,230	0.82	8.17678	8.17376
Forrest Gump	8.3	136,475	0.83	8.27998	8.27745
Inception	8.2	136,760	0.82	8.17955	8.17720
Star Wars:	8.2	137,622	0.82	8.17961	8.17728

Episode V - The Empire Strikes Back					
The Lord of the Rings: The Two Towers	8.2	137,945	0.82	8.17964	8.17731
The Matrix	8.3	141,790	0.83	8.28036	8.27792
Goodfellas	8.2	144,853	0.82	8.18013	8.17791
One Flew Over the Cuckoo's Nest	8.2	153,419	0.82	8.18070	8.17860
Shichinin no samurai	8.2	155,326	0.82	8.18081	8.17874
Se7en	8.3	156,747	0.83	8.28132	8.27912
La vita è bella	8.5	158,965	0.85	8.48236	8.47988
Cidade de Deus	8.3	167,748	0.83	8.28195	8.27989
The Silence of the Lambs	8.4	168,653	0.84	8.38243	8.38023
It's a Wonderful Life	8.4	191,672	0.84	8.38352	8.38159
Star Wars	8.4	193,442	0.84	8.38360	8.38168
Saving Private Ryan	8.5	206,621	0.85	8.48454	8.48263
Gisaengchung	8.2	209,307	0.82	8.18348	8.18194
Sen to Chihiro no kamikakushi	8.2	210,477	0.82	8.18353	8.18200
The Green Mile	8.4	217,364	0.84	8.38453	8.38282
Interstellar	8.4	218,477	0.84	8.38457	8.38287
Léon	8.2	221,265	0.82	8.18394	8.18248
The Usual Suspects	8.3	222,508	0.83	8.28434	8.28278
Seppuku	8.2	232,582	0.82	8.18433	8.18295
The Lion King	8.2	256,825	0.82	8.18509	8.18384
American History	8.3	263,144	0.83	8.28560	8.28428

X					
The Pianist	8.2	285,140	0.82	8.18586	8.18472
Back to the Future	8.2	286,763	0.82	8.18590	8.18477
Terminator 2: Judgment Day	8.5	289,435	0.85	8.48694	8.48558
Modern Times	8.3	296,534	0.83	8.28644	8.28527
Psycho	8.6	302,319	0.86	8.58759	8.58620
Gladiator	8.3	306,158	0.83	8.28665	8.28552
City Lights	8.4	316,384	0.84	8.38718	8.38601
The Departed	8.2	320,370	0.82	8.18666	8.18565
The Intouchables	8.2	320,752	0.82	8.18667	8.18566
Whiplash	8.4	321,537	0.84	8.38729	8.38613
The Prestige	8.3	324,235	0.83	8.28703	8.28596
Once Upon a Time in the West	8.3	337,416	0.83	8.28729	8.28626
Hotaru no haka	8.4	343,701	0.84	8.38771	8.38662
Casablanca	8.3	343,864	0.83	8.28741	8.28640
Rear Window	8.3	355,759	0.83	8.28762	8.28664

6.2 代码

6.2.1 评分模型算法

```
1.  //C++代码
2.  float Score_cal(pos, n, confidence)//pos 为总分，n 为评分人数，confidence 为置信水平
3.  {
4.  float z=1.96; //置信水平为 95%，对应的 Z 为 1.96，根据需求更改
5.  float C=1000; //贝叶斯平均补偿人数
6.  float M=4;    //贝叶斯平均补偿分数
7.  if(n == 0)
8.  return 0;
9.  else
10. {
11. float phat = 1.0*pos/n;
```

```
12. float S=(phat+z*z/(2*n)-z* Math.sqrt((phat*(1-
    phat)+z*z/(4*n))/n))/(1+z*z/n);
13. float score=(C*M+n*S)/(n+C);//最终得分
14. return score;
15. }
16. }
```

//EXCEL 代码

```
1.  =IFERROR((([@[Up Votes]] + 1.9208) / ([@[Up Votes]] + [@[Down Votes]]) - 1.9
    6 *
2.  SQRT(([@[Up Votes]] *  [@[Down Votes]]) / ([@[Up Votes]] +  [@[Down Votes]])
     + 0.9604) /
3.  ([@[Up Votes]] +  [@[Down Votes]])) / (1 + 3.8416 / ([@[Up Votes]] +  [@[Dow
    n Votes]])),0)
```