基于机器学习的票房预测模型及大数据分析

摘要

“互联网+”时代催生创业者。国内企业做大做强后，总想要开拓海外市场。同时随着人们经济水平的提高，人们也在考虑精神享受，比如看电影。对于一个考虑海外开拓电影市场的商人来说，一个亟待解决的重要问题是怎样拍商业电影才能赚钱、这其中是否存在一定的方法或统一公式可以进行有效预测？如果不能解决，那么就算投入再高也可能失败。

基于此背景，我们需要从已经获得的来自 Kaggle 的美国地区 1916-2017 年近 5000 部电影的数据中运用数据挖掘的方法做分析，并解决不同维度研究电影、机器学习预测模型、最优电影拍摄方案寻找这 3 个问题。

针对问题一，我们首先通过数据统计分析、离群点分析对数据集进行描述，并给定数据预处理方案，完成数据清洗、数据规约的工作，也即完成数据预处理，得到了能够使用的数据集；其次通过查阅相关文献、建立相关性度量模型、观察关系图等方法建立了票房影响因素函数模型；最后从不同 6 个维度分析了电影的影响变量，完成数据挖掘、数据可视化。

针对问题二，我们首先通过信息增益，选择了与平均评分、投票人数最相关的几个变量；然后通过反向传播、随机梯度下降完成神经网络模型搭建，并使用处理后的数据集完成神经网络训练；最后利用训练好的模型对用于预测的数据集进行预测。

针对问题三，我们结合对问题一、问题二的分析、解答过程进行讨论，最终得出结论：如果一部电影预算越大，风格越接近歌剧、戏剧片、惊悚片、动作传记，导演的其他作品越有名气，发行时间越接近 4、5、6 月时，为原创，则该部电影越可能赚到钱。

本文提供了一个数据预处理的一个范例，建立了票房影响因素函数模型，为海外投资电影的商人提供了一种预测电影票房的方法，具有一定参考价值，使用更完整、优质的数据集所训练出的模型将有可能准确判断一部电影票房高低与否。

关键词：Sprearman 秩相关系数，信息增益，神经网络，数据挖掘

一、问题重述

对于一个考虑海外开拓电影市场的商人来说，怎样拍商业电影才能赚钱是一个亟待解决的重要问题，如果不能找到一定的方法或统一公式进行有效预测，那么就算投入再高也可能失败。目前已经获得了来自 Kaggle 的美国地区 1916-2017 年近 5000 部电影的数据，需要从中运用数据挖掘的方法做分析，并解决如下 3 个问题：

• 从不同维度研究电影，分析电影票房的影响因素。可以考虑的维度包括观众喜欢什么电影类型、有什么主题关键词、电影风格随时间如何变换、拍原创电影好还是改编电影好等等。

• 选择合适的指标进行特征提取，建立机器学习的预测模型，根据 1000 部电影的基本信息来预测它们的平均评价、投票人数。

• 从数据的分析与研究中，尝试分析是否能找到一个确定拍摄一部电影最有可能赚到钱的模式或决策方法，并给出理由与依据。

二、问题分析

问题给出了过去一段时间内电影类别和评价数据作为训练集和测试机，需要基于分析各因素对电影盈利的影响，并建立和训练模型，对未来的电影的收入情况做出预测，并分析能提高电影盈利的模式或决策方法。

针对问题一，需要通过数据清洗、数据挖掘、数据分析、数据可视化来研究电影票房的影响因素有哪些，而可以分析的维度有：观众喜欢什么电影类型？有什么主题关键词？电影风格随时间是如何变化的？电影预算高低是否影响票房？高票房或者高评分的导演有哪些？电影的发行时间最好选在啥时候？拍原创电影好还是改编电影好？这些维度中，可能的分析方法有统计分析、关联规则挖掘、分类等等，可以建立票房预测模型。

针对问题二，核心在于指标的选取和机器学习预测模型的建立指标选取时可以通过信息增益选择最相关的变量。为了避免过拟合、提高泛化能力，可以考虑加入验证集，使用交叉验证的方法得到更好的预测模型。

针对问题三，一方面可以考虑优化，考察预测模型的输入是多少时具有最高输出，若最优解是存在的并且可解的，那么找到一个确定拍摄一部电影最有可能赚到钱的模式或决策方法是可能的，优化的方法即为所求方法；否则，则不能。另一方面可以结合问题一、二分析过程中得到的结论来讨论。

三、基本假设与符号说明

3.1 基本假设

1.电影数据真实可靠，能够反映每部电影的真实特点、口碑和票房情况；

2.数据集具有普遍性，能够反映当时的真实电影市场情况；

3.数据集中针对不同电影的评分均基于相同或相似标准；

4.观众的评分近似符合正态分布。

3.2 符号说明

变量名	变量解释
x_1	budget，预算
x_2	genres，电影类型
x_3	keywords，电影主题关键词
x_4	popularity，流行程度
x_5	release_date，发行日期
x_6	revenue，总收入
x_7	Runtime，电影时长
x_8	vote_average，平均评分
x_9	vote_count，总投票数
x_{10}	cast，演员
x_{11}	crew，职员
x_{12}	adapted，是否被改编
r_s	斯皮尔曼相关系数
x_i	变量 x 第 i 个值的降序排列秩
y_i	变量 y 第 i 个值的降序排列秩
n	样本容量
t	服从 t 分布的检验变量
$\overline{X}$	样本均值

S	样本标准差
μ	总体均值
σ^2	总体方差
u_α	α 上分位数
T	阈值，意味着评分大于它时观众喜欢这部电影
η	学习率
ω_i	权重
b_i	偏置

四、模型建立与求解

4.1 问题一：电影票房的影响因素

4.1.1 数据预处理

a. 数据统计分析

通过统计分析，数据集 tmdb_5000_movies（简称为数据集 1）中的每个数据对象共有 20 项变量，共有 4803 条数据。变量名、统计非缺失数量和数据类型如表 1 所示。

表 1 数据集 1 的描述

序号	变量名	非缺失计数	数据类型	序号	变量名	非缺失计数	数据类型
1	budget	4803	int64	11	production_countries	4803	json object
2	genres	4803	json object	12	release_date	4802	date(y/m/d)
3	homepage	1712	str	13	revenue	4803	int64
4	id	4803	int64	14	runtime	4801	float64
5	keywords	4803	json object	15	spoken_languages	4803	json object
6	original_language	4803	str	16	status	4803	str
7	original_title	4803	str	17	tagline	3959	str
8	overview	4800	str	18	title	4803	str
9	popularity	4803	float64	19	vote_average	4803	float64
10	production_companies	4803	json object	20	vote_count	4803	int64

同理，通过统计分析，也可以得到数据集 tmdb_5000_credits（简称为数据集

2）中的每个数据对象共有 4 项变量，共有 4803 条数据。变量名、统计非缺失数量和数据类型如表 2 所示。

表 2 数据集 2 的描述

序号	变量名	非缺失计数	数据类型
1	movie_id	4803	int64
2	title	4803	json object
3	cast	4803	json object
4	crew	4803	json object

经过进一步分析可以发现，尽管有些数据对象的某个变量有值，但该值为 0，不符合常识，如 budget 有 1025 个 0 值、runtime 有 35 个 0 值、vote_count 有 62 个 0 值，等等。鉴于 0 值太多（如 budget 的 0 值超过 20%），不管用什么数学方法完成缺失值填充都将可能造成数据挖掘出的结论出现问题；而根据 homepage 或其他方式查找相关资料将耗时太大，人力不可能在短时间内完成缺失值填充，这里选择将这种数据对象直接删除，剩下的数据足够完成进一步的数据挖掘。另外，两个表描述的是同样的 4803 部电影，他们依靠 id(movie_id)或 title 联系，有必要将两个表合并成一个表，以方便下一步的处理。

b. 数据预处理方案

在 a.数据统计分析中，我们给出了预处理的一般要求，下面列出表格，对每个变量规定更为具体的处理标准，如表 3。

表 3 数据预处理标准

变量	操作
budget	删除该栏 0 值或缺失值的对象
genres	提取该列表中每个字典“name”键对应的值
homepage	删除该栏
id	与 movie_id 匹配，之后删除
keywords	提取该列表中每个字典“name”键对应的值； 提取是否改编的信息
original_language	删除
original_title	删除（title 完全能代替）
overview	提取是否改编的信息，删除
popularity	删除该栏 0 值或缺失值的对象
production_companies	删除
production_countries	删除
release_date	删除该栏 0 值或缺失值的对象
revenue	删除该栏 0 值或缺失值的对象
runtime	删除该栏 0 值或缺失值的对象

spoken_languages	删除
status	删除
tagline	提取是否改编的信息，之后删除
title（数据集 1）	保留
vote_average	删除该栏缺失值的对象
vote_count	删除该栏 0 值或缺失值的对象
movie_id	与 id 匹配，完成数据集合并，之后删除
title（数据集 2）	删除
cast	提取该列表中每个字典“character”键对应的值
crew	提取该列表中每个“job”对应 director 的字典的“name”对应值
adapted	为新增变量，其值取决于 keywords、overview、tagline 中提取到的是否改编的信息

c. 预处理结果

经过预处理得到的数据集命名为 tmdb_5000_movies&credits_ap（简称数据集 3），其中每个数据对象共有 13 项变量，共有 3129 条数据。变量名、统计非缺失数量和数据类型如表 4 所示，完整表格见附件 Xtmdb_5000_movies&credits_ap.csv。

表 4 数据集 3 的描述

序号	变量名	非缺失计数	数据类型	序号	变量名	非缺失计数	数据类型
1	budget	3129	int64	8	title	3129	str
2	genres	3129	list	9	vote_average	3129	float64
3	keywords	3129	list	10	vote_count	3129	int64
4	popularity	3129	float64	11	cast	3129	list
5	release_date	3129	date	12	crew	3129	list
6	revenue	3129	int64	13	adapted	3129	int64
7	runtime	3129	float64				

可见经过处理后所有数据对象都没有属性缺失。

为方便阅读，在此对目前所有数据集做一总结，如表 5 所示。

表 5 数据集总结

数据集名称	数据集简称	数据条数	数据的属性数	数据集描述
tmdb_5000_movies	数据集 1	4803	20	题目所提供的电影基本信息
tmdb_5000_credits	数据集 2	4803	4	题目所提供的演职员信息
tmdb_5000_movies&credits_ap	数据集 3	3129	13	经数据集 1、2 清洗、合并而成

tmdb_1000_predict	数据集 4	1000	18	题目所提供的用于预测的电影基本信息

4.1.2 票房影响因素模型建立

票房（revenue）常常被作为一个电影成功与否的最关键因素，是我们的重点研究对象。

a. 总述

经过 4.1.1 的数据预处理，homepage、original_title 等无关变量已经剔除，但在得到数据集 3 中数据的 13 个变量中还有变量需要剔除，也即 title。

综合查阅相关参考文献[1,2,3]，可以认为流派 genres、主题关键词 keywords、演员 cast、职员 crew 等 4 个 list 类型的变量和发行日期 release_date 这个 date 类型变量都会对票房产生一定的影响。

b. 相关性度量模型

对于预算 budget、流行度 popularity、时长 runtime、平均评分 vote_average、投票人数 vote_count、是否改编 adapted 这些数值型变量，可以利用 Spearman 秩相关性分析从统计学角度分析二者联系建立相关性度量模型。

斯皮尔曼相关系数被定义成等级变量之间的皮尔逊相关系数。对于样本容量为 n 的样本，n 个原始数据被转换成等级数据，相关系数 ρ 为

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

原始数据依据其在总体数据中平均的降序位置，被分配了一个相应的等级。若几个数据数值相同，则取其等级的平均值作为几个数据的等级。

实际应用中，变量间的连结是无关紧要的，于是可以通过简单的步骤计算 ρ. 被观测的两个变量的等级的差值，则 ρ 为

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2}$$

其中 $d_i = (x_i - y_i)$，x_i, y_i 分别表示两个变量按降序排列的秩，n 为样本容量。

进行显著性检验时，令

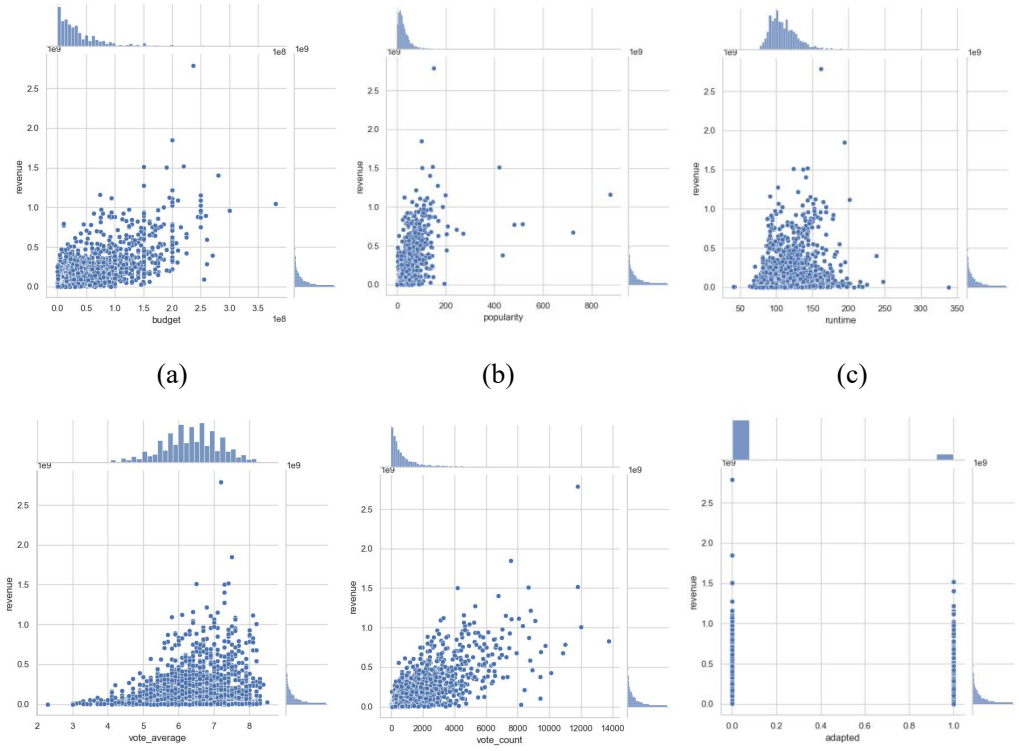
原假设：H0：X 和 Y 相互独立，H1：X 和 Y 正相关，或

原假设：H0：X 和 Y 相互独立，H1：X 和 Y 负相关

则假设检验的拒绝域分别为：$W=\{\rho_s >= c_\alpha\}$，$W=\{\rho_s <= d_\alpha\}$ ，c_α，d_α 为临界值，在零假设成立时，$t=r_s\sqrt{\frac{n-2}{1-r^2}}$，服从自由度为 v=n-2 的 t 分布。

在显著水平为α时，统计量的值落在否定域$\{t\,|\,|t|>\frac{t_\alpha}{2(n-2)}\}$中，拒绝零假设，Spearman 等级相关系数显著；否则，接受零假设，则 Spearman 秩相关系数不显著。[7]

在计算 Spearman 秩相关系数前，我们先通过两变量的散点图对变量关系有一个初步的认识，如图 1 所示。



(a) (b) (c)

(d) (e) (f)

图 1　6 个数值型变量与 revenue 的关系图

由上图可以初步得到 budget、popularity、vote_count 和 revenue 之间有较强相关关系，其余变量与 revenue 之间关系不明显，下面通过模型进行验证。

使用 python 的 scipy 库，容易计算出 budget、popularity、runtime、vote_average、vote_count、adapted 分别与 revenue 的 Spearman 秩相关系数，如表。

表 6 变量与 revenue 的 Spearman 秩相关系数

变量	**budget**	**popularity**	**runtime**	**vote_average**	**vote_count**	**adapted**
Spearman 系数	0.6807	0.6969	0.2258	0.1221	0.7428	0.1254
p 值	0	0	1.899e-37	7.156e-12	0	1.938e-12

根据如上计算结果同样可以看出与 revenue 具有很强相关关系的变量有 budget、popularity、vote_count，其余变量可暂不考虑。

c. 票房影响因素函数模型

根据以上讨论，给予变量符号，如表 7。

表 7 变量及其符号

变量名	变量符号	是否纳入模型	变量名	变量符号	是否纳入模型
budget	x_1	是	runtime	x_7	否
genres	x_2	是	vote_average	x_8	否
keywords	x_3	是	vote_count	x_9	是
popularity	x_4	是	cast	x_{10}	是
release_date	x_5	是	crew	x_{11}	是
revenue	x_6	是	adapted	x_{12}	否

故可以建立如下的票房影响因素函数模型：

$$revenue = x_6 = f(x_1, x_2, x_3, x_4, x_5, x_9, x_{10}, x_{11}) \tag{3}$$

其中 f 的具体表达式有赖于 4.1.3 的分析，以及类似 4.2 神经网络的求解。

4.1.3 不同维度的电影影响变量分析

（一）观众喜欢什么电影类型？有什么主题关键词？

判断观众喜欢一部电影与否，最直接的一个变量就是电影平均评分（vote_average, x_9 ,取值 0~10），很显然 x_9 越高表示观众越喜欢一部电影，否则表示观众越不喜欢一部电影，但选取一个阈值 T 来作为喜欢与不喜欢的分界线不能仅凭主观，而应该考虑 vote_average 的分布。作出 vote_average 的频数分布直方图（bins=20）如图 2 所示，可以得出 vote_average 的分布接近正态分布。

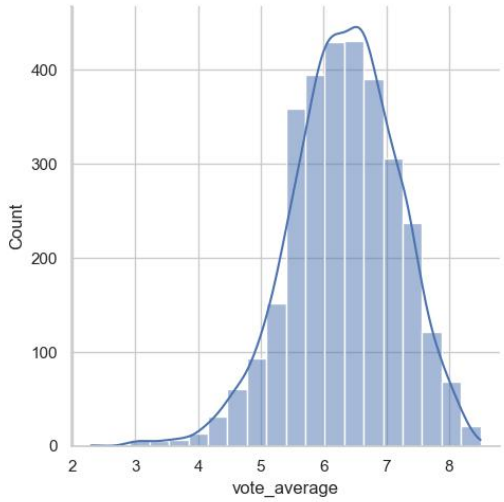


图 2 vote_average 的频数分布直方图

考察所有 vote_averag 取值，容易计算出样本均值 $\overline{X}$ 为 6.327261，样本标准差 S 为 0.851808，若近似认为 vote_averag 服从正态分布、样本均值等于总体均值 μ、样本方差等于总体方差 σ^2，则其概率密度函数为：

$$f_{x_9}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

取该正态分布的 0.3 上分位数 $u_\alpha=u_{0.3}$ 作为阈值 T，意味着我们认为评分最高的 30%观众喜欢该电影。由概率论知识，$P\{X>u_\alpha\}=1-F(u_\alpha)=1-\Phi(\frac{u_\alpha-\mu}{\sigma})=0.3$，查表得 $\frac{u_\alpha-\mu}{\sigma}=0.53$，解得 $T=u_{0.3}=6.78$.[7]

a. 电影类型

关于观众喜欢什么电影类型，由于各电影类型的数量大有不同，不宜直接比较各种类型电影被喜欢的数量，而应该比较各种类型电影被喜欢的比例，基于此可以对数据进行处理，列出各种类型电影被喜欢的比例（前 9）如表 8 所示。

表 8 电影类型与被喜欢的情况

电影类型	Documentary	War	History	Western	Drama	Music	Animation	Romance	Crime
被喜欢的比例	0.6765	0.5424	0.5385	0.5179	0.4545	0.3981	0.3667	0.3345	0.3333
被喜欢的数量	23	64	77	29	634	43	66	185	169
该类型电影总数	34	118	143	56	1395	108	180	553	507

在数据集的所有电影中，被喜欢的电影（评分大于 6.78）的仅占 0.3，而在

上表的 9 种类型电影中，被喜欢的电影均占 0.3 以上，故可认为这些电影类型（Documentary、War、History、Western、Drama、Music、Animation、Romance、Crime）是观众喜欢的。

b. 主题关键词

由于电影的主题关键词非常多，而不像 a 中电影类型十分有限，在这里我们采取不同的处理方式，即选择比较观众喜欢的电影中主题关键字的出现次数进行排序，同时去除被喜欢比例低于 0.3 的关键词。对数据进行处理，列出观众喜欢的电影中各种主题关键词的出现次数（前 5）如表 9 所示。

表 9 主题关键词与被喜欢的情况

主题关键词	based on novel	biography	independent film	dystopia	friendship
被喜欢的次数	82	48	43	41	39
关键词出现总次数	175	79	127	130	79
被喜欢的比例	0.4686	0.6076	0.3386	0.3154	0.4937

在上表的 5 个关键词中，被喜欢的关键词比例均占 0.3 以上，且按被喜欢次数排序，故可认为这些主题关键词类型（based on novel、biography、independent film、dystopia、friendship）是观众最喜欢的。

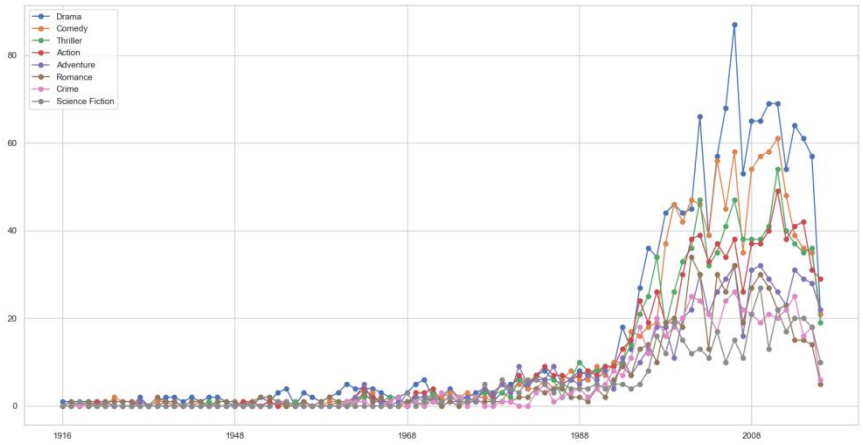
（二）电影风格随时间是如何变化的？

首先需要统计各类风格电影的数量，选取较热门的电影进行分析，这一点有别于（一）a。统计结果（前 16）如表 10 所示。

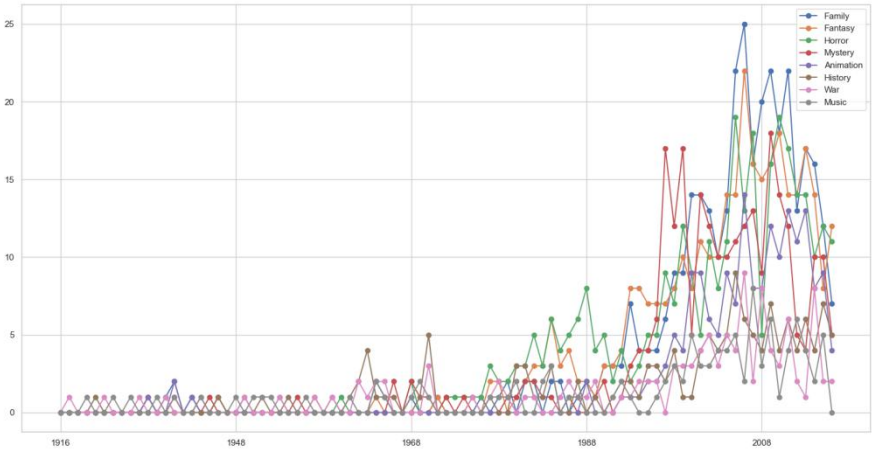
表 10 电影风格统计

电影类型	出现次数	电影类型	出现次数
Drama	1395	Family	353
Comedy	1062	Fantasy	335
Thriller	915	Horror	329
Action	893	Mystery	261
Adventure	653	Animation	180
Romance	553	History	143
Crime	507	War	118
Science Fiction	426	Music	108

考察这 16 种类型电影随时间的变化情况，如图 3 所示，为避免重合过多影响分析，将 Drama~Science Fiction、Family~Music 分开作图。



(a)



(b)

图 3 电影风格随时间的变化情况

可见，几乎所有类型的电影均有一种相似的趋势，即从 1915 年开始逐渐增长，2008 年前增长由为快速，而 2008 年后普遍有下降趋势。而 drama 在所有年份都几乎是最热的电影类型，Science Fiction 在其他电影最热时表现出低谷。

（三）电影预算高低是否影响票房？

在 4.1.2 中我们已经分析过预算高低与票房高低有一定相关关系。其关系图如图 4 所示，容易计算二者 Spearman 秩相关系数为 0.6807，皮尔逊相关系数 0.7044，说明二者的确有一定相关性，故可以认为一定程度上电影预算高低是否影响票房，预算越高则票房也越高。

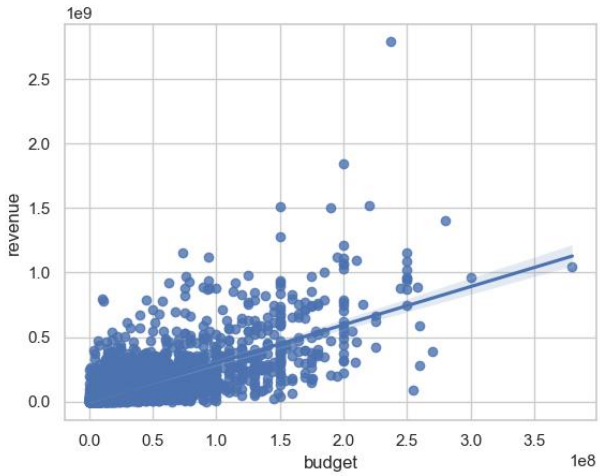


图 4 票房与电影预算关系图

（四）高票房或者高评分的导演有哪些？

a. 高票房

首先需要考察票房的分布如图 5 所示，类似幂律分布，低票房的是多数，高票房的是少数。计算 0.25 上分位数约为 1.501 亿元，接下来的处理我们认为票房收入大于 1501 万元则是高票房。

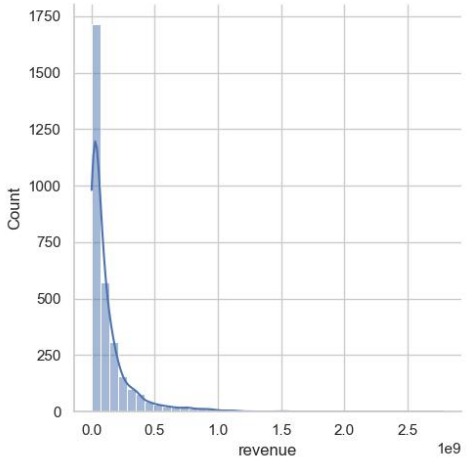


图 5 票房的频数分布直方图

可以通过数据处理，找到所有导演的电影平均票房，并排序。为避免偶然性，仅考察有 7 部电影及以上的导演，高票房的导演如表 10。

表 10 高票房导演

导演	票房	导演	票房
James Cameron	840509919	Tim Burton	238387017
Peter Jackson	722071424	Jay Roach	231091784

Christopher Nolan	528435404	Wolfgang Petersen	230716583
Michael Bay	486043719	Shawn Levy	221858706
Gore Verbinski	453859051	Brett Ratner	220502662
Chris Columbus	413959055	David Fincher	213934632
Sam Raimi	391612682	Garry Marshall	205884428
Sam Mendes	387649829	Marc Forster	201298420
Roland Emmerich	369602630	Ridley Scott	199347374
Zack Snyder	353742483	Ang Lee	198517099
Bryan Singer	351029375	Doug Liman	191737236
Steven Spielberg	338792339	Tom Shadyac	189882125
Barry Sonnenfeld	313205533	Todd Phillips	188627625
Lilly Wachowski	291790424	Edward Zwick	162208269
Lana Wachowski	291790424	Dennis Dugan	161853893
Robert Zemeckis	276201692	Tony Scott	155660061
M. Night Shyamalan	272483881	Simon West	154448457
Martin Campbell	264906421	Guy Ritchie	154368841
Ron Howard	261580958	Quentin Tarantino	153096869

b. 高评分

可以通过数据处理，找到所有导演的电影平均评分，并排序。为避免偶然性，仅考察有 7 部电影及以上的导演。若取阈值为 6.9，即认为 6.9 以上的评分为高评分，则高评分的导演如表 11。

表 11 高评分导演

导演	电影评分	导演	电影评分
Christopher Nolan	7.80	Alfred Hitchcock	7.05
Quentin Tarantino	7.55	Danny Boyle	7.01
Wes Anderson	7.41	Richard Linklater	7.01
Martin Scorsese	7.38	Lasse Hallström	6.99
David Fincher	7.34	Steven Spielberg	6.97
Peter Jackson	7.33	Clint Eastwood	6.92
James Cameron	7.33	Robert Zemeckis	6.91
Francis Ford Coppola	7.25	Sam Mendes	6.91

c. 总结

结合 a、b 的分析，找到高票房导演、高评分导演的交集，也即作品票房、评分都高的导演，如表 12 所示。

表 12 高票房、高评分导演

导演	票房	电影评分	导演	票房	电影评分
Peter Jackson	722071424	7.33	Steven Spielberg	338792339	6.97

Robert Zemeckis	276201692	6.91	Peter Jackson	722071424	7.33
Christopher Nolan	528435404	7.80	Sam Mendes	387649829	6.91
Quentin Tarantino	153096869	7.55	James Cameron	840509919	7.33

其中诺兰（Christopher Nolan）、斯皮尔伯格（Steven Spielberg）、皮特（Peter Jackson）都是我们熟知的优秀导演，说明我们的数据处理结果是和实际较为相符。

（五）电影的发行时间应选在什么时间？

一部电影制作的发行年份波动不大，未来不同年份发行是的收入难以预测，也不必预测；不同天发行的收入不一定具有周期性，且 3000 量级的数据太少，不能支持相关结论。故考虑发行时间时最应该考虑的是月份。我们统计历年来每年在某一月份的平均收入，并绘制折线图如图 6 所示。可以看出早年间（1968 年前）2、3、8 月都有时会出现票房高峰；1968 年后之后，4、5、6 月有时出现高峰，该几组曲线也几乎处于上方。故选择发行时间时可以考虑在 4、5、6 月发行。

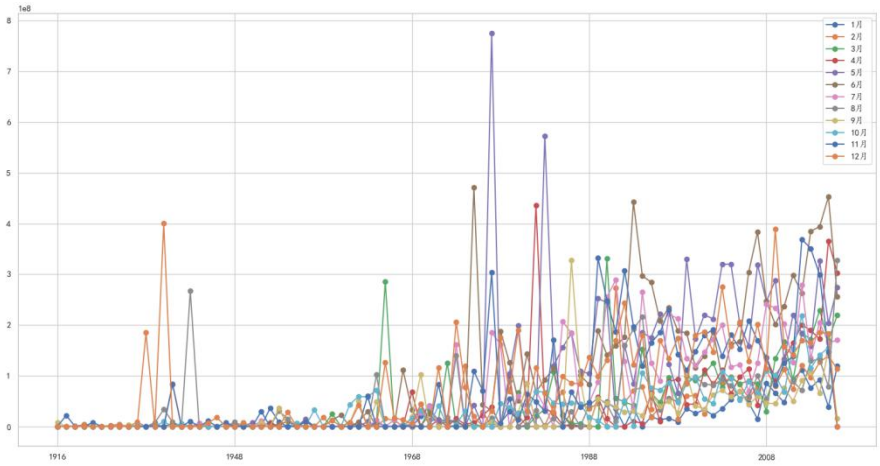


图 6 月份平均票房随年份的变化

（六）拍原创电影好还是改编电影好？

在 4.1.2 的分析中我们认为是否原创与票房收入相关性较弱，但没有考虑年份变化，现重新处理，统计每年原创电影、改编电影的平均收入，并绘制折线图，如图 7 所示，可见在 1978 年前改编电影普遍比原创电影有更好的票房（除 1965 年原创电影有极大的票房外），而 1978 年后原创电影普遍比改编电影有更好的

票房，并且近年来原创电影的平均票房还呈上升趋势。故近年来，拍摄原创电影也许能有更好的票房收入。

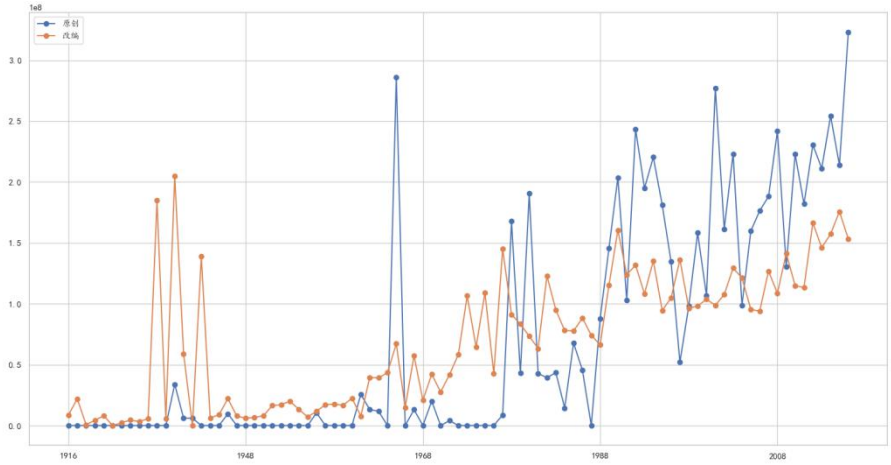


图 7 原创（改编）电影平均票房随年份的变化

4.2 问题二：电影评价的预测模型

4.2.1 特征筛选

为了降低预测模型的复杂度，提高模型的泛化能力，需要首先对预测指标进行筛选，选择的标准为能与电影票房、口碑等有更高的相关性。

信息熵刻画了一个系统的混乱程度，其定义如下：

$$\mathrm{H(Y)} = E\left[-\log p_i\right] = -\sum_{i=1}^{n} p_i \log p_i \tag{5}$$

而条件信息熵则刻画了在已知 X 的基础上还需要多少信息来描述 Y，其定义如下：

$$\mathrm{H}(Y \mid X) = -\sum_{x \in X}\left[p(x)\sum_{y \in Y} p(y \mid x)\log p(y \mid x)\right] \tag{6}$$

原 Y 的信息熵减去在 X 条件下 Y 的信息熵即得到了信息增益(information gain, IG)，它刻画了在已知 X 的基础上需要节约多少信息来描述 Y，其定义如下：

$$IG(Y \mid X) = H(Y) - H(Y \mid X) \tag{7}$$

在本问题中，基于 IG 可以刻画不同变量对预测结果的相关性。

4.2.2 进行交叉验证的神经网络模型

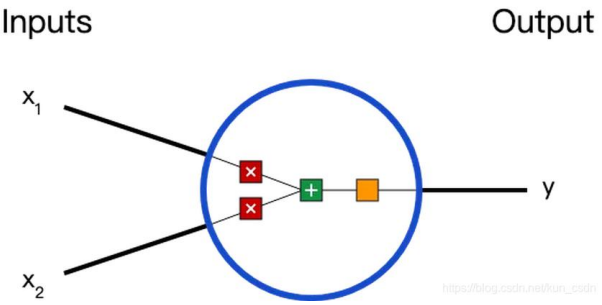


图 8 神经元模型

神经网络的基本单元是神经元。神经元先获得输入，然后执行某些数学运算后，再产生一个输出。在其中，输入到输出的过程用下面的函数刻画：

$$y = f(\omega_1 x_1 + \omega_2 x_2 + b) \tag{8}$$

f 为激活函数，这里我们使用 sigmoid 函数 $f(x) = \dfrac{1}{1+e^{-x}}$ 。

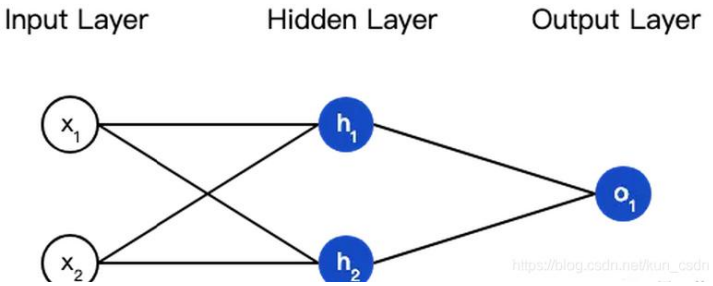


图 9 神经网络模型

如图为神经网络的一个简单举例，实际操作过程中输入输出根据场景要求，隐层节点数、网络层数可自行调试。[5,8]

同时根据一般要求，将数据集 3 按 7：3 的比例分成两个数据集，分别为测试集和验证集。

4.2.3 模型求解算法

定义损失，如均方误差 $MSE = \dfrac{1}{n}\sum_{i=1}^{n}(y_{true} - y_{pred})^2$ ，模型训练过程中力求损失最小化。完整的损失函数的自变量应该包括多个权重、偏置：

$$L = L(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, b_1, b_2, b_3) \tag{9}$$

考虑一个权重变化时损失函数的变化，则应向后计算偏导数

$$\frac{\partial L}{\partial \omega_1}=\frac{\partial L}{\partial y_{pred}}\frac{\partial y_{pred}}{\partial h_1}\frac{\partial h_1}{\partial \omega_1} \tag{9}$$

于是可以根据该偏导数来更新权重，以使损失减小，此时使用随机梯度下降（SGD）：$\omega_1 \leftarrow \omega_1-\eta\frac{\partial L}{\partial \omega_1}$，以完成权重的更新，其中$\eta$为学习率，反映训练网络时的速率快慢。多次计算偏导数，更新权重，计算损失，当损失减小到不能再小时完成训练，模型求解完毕。

4.3 问题三：最大化电影盈利的决策方法

我们结合对问题一、问题二的分析、解答过程进行讨论，最终得出结论：如果一部电影预算越大，风格越接近歌剧、戏剧片、惊悚片、动作传记，导演的其他作品越有名气，发行时间越接近 4、5、6 月时，为原创，则该部电影越可能赚到钱。

五、模型评价和改进方向

5.1 模型评价

5.1.1 优点

1.从图形可视化、统计分析两个方面完成大数据分析，对单一变量的分析可靠性高

2.神经网络隐藏层能够将数据的特征抽象出来，以一种简单的方式完成了对结果的预测

3.相关性度量模型迁移性强，在多种场景下都适用

5.1.2 缺点

1.没有使用多元回归分析，对变量间的相互影响考虑较少

2.数据预处理时对有缺失值的数据全部进行删除，元素的减少对结果预测可能造成影响

5.2 模型改进方向

1.向深度学习改进，提高模型精度，完成更准确的预测

2.尝试收集更多、更优质的数据对模型重新训练

3.考虑变量间的相关关系，探究变量间如何相互影响，可能有助于得到新的结论

六、参考文献

[1] 张明坤, 于辉. 浅析影响电影成功的因素[J]. 新西部:中旬・理论, 2012.

[2] 杨照帅. 电影成功因素探析——以威漫公司科幻系列电影为例[J]. 美与时代(下), 2014(10):71-72.

[3] 何萍. 影响电影票房的几大因素分析[J]. 中国电影市场, 2011, 000(011):8-10.

[4] 徐全智, 吕恕. 国家工科数学课程教学基地系列教材 概率论与数理统计[M]. 高等教育出版社, 2004.

[5] 李宗坤, 郑晶星, 周晶. 误差反向传播神经网络模型的改进及其应用[J]. 水利学报, 2003, 34(7).

[6] 姜启源. 数学模型(第二版)[M]. 高等教育出版社, 1987.

[7] 马燕. 卫生统计学[M]. 人民卫生出版社, 2002.

[8] 周志华. 《机器学习》[J]. 中国民商, 2016, 03(No.21):93-93.

七、附件

附件一 程序文件夹

内含 DataAna.py（数据初分析）、DataPrepro.py（数据预处理）、DataAna_ap.py、DataAna_ap2.py、……DataAna_ap6.py（分布针对问题一中 1、2、……6 个维度的处理）、network.py（神经网络模型搭建与预测）

附件二 数据文件夹

内含 tmdb_5000_movies&credits_ap.csv（数据清理、数据规约后的数据集）、tmdb_1000_predict.csv（预测结果）