豆瓣电影票房预测方法与电影推荐

2022 年编程基础课程期末报告

学号：2022103795

姓名：吕明倩

日期：2022 年 1 月 13 日

目　录

第一章　背景介绍

1.1 问题陈述

一部成功的电影从前期筹备到最终成片离不开所有人员的共同努力，电影公司制作一部新电影推向市场时，通常需要通过调研来发掘观众喜好的电影类型以及电影内容，从而创作出更贴合市场的电影作品，以期望收获更高收益，这时往期电影数据便成为上述分析的有利基础。通过对往期电影市场的数据进行探索，不仅更容易对当下电影预算、题材、票房等方面的问题做出合理的决策，而且能实现对未来趋势的合理预测。

所以本文首先聚焦电影票房的预测问题，考虑电影的发行国家、语言、年代、题材、色彩、分类等级等定性指标，以及预算、评分、评论数、$FaceBook$ 喜爱数等定量指标，分别构建多元线性回归、岭回归、$Lasso$ 回归、决策树、随机森林等预测模型，实现对电影票房的合理预测，并对不同模型表现效果进行综合评价。

此外，伴随着现代生活节奏的加快，看电影成为不少年轻人休闲放松的主要方式，选择一部更适合自己的电影更容易获得观影幸福感。因此本文的第二个目标就是基于 KNN 算法，从电影相似度和用户相似度两种角度分别建立推荐系统，完成更有效的电影推荐。

1.2 数据集简介

本文选择豆瓣 5043 部电影信息作为电影票房预测数据集（$movie\_data$），数据集共包含 28 个字段，各字段名称以及意义解释如表 1 所示：

字段名称	意义	字段名称	意义
movie_title	电影题目	director_name	导演名字
language	语言	director_facebook_likes	导演 FaceBook 粉丝数
country	发行国家	actor_1_name	演员 1 名字
content_rating	分级	actor_1_facebook_likes	演员 1FaceBook 粉丝数
title_year	发行年月	actor_2_name	演员 2 名字
color	色彩	actor_2_facebook_likes	演员 2FaceBook 粉丝数
duration	片长	actor_3_name	演员 3 名字
genres	题材类型	actor_3_facebook_likes	演员 3FaceBook 粉丝数
plot_keywords	剧情关键词	num_voted_users	投票人数
budget	制作成本	num_user_for_reviews	用户评论数
gross	总收入	num_critic_for_reviews	评论家点评数
aspect_ratio	画布比例	movie_facebook_likes	电影 FaceBook 点赞数
facenumber_in_poster	海报中人数	cast_total_facebook_likes	FaceBook 上投喜爱的总数
imdb_score	豆瓣评分	movie_imdb_link	电影数据链接

表 1：电影数据集字段名称及意义

其中除了描述电影属性的变量外，本文用 $budget$ 来表示电影总投入，用 $gross$ 来表示电影总票房收益，用导演与演员的 $FaceBook$ 粉丝数来反应观众对电影的期待度，用投票

人数、评论数、点赞数、豆瓣评分等字段来综合反应电影的口碑情况。

此外，数据表 $user\_movie$ 记录了 610 个用户的电影打分记录，数据表 $movie\_info$ 记录了 193609 部电影的名称与题材，可以作为构建电影推荐系统的数据基础。

1.3 数据清洗

在导入数据集后，通过 $df.describe()$ 和 $df.info()$ 描述数据基本信息。首先 $df.describe()$ 统计了电影片长、制作成本、总收入等 15 个数值变量的取值情况，发现数值全部为正值，且均在合理取值范围内，因此对各数值变量暂不需要进行数据规范化处理。其次 $df.info()$ 显示数据几乎全部字段均存在不同程度的缺失，因此需要进行缺失值处理。

字段名称	非空数	字段类型	字段名称	非空数	字段类型
movie_title	5043	object	director_name	4939	object
language	5031	object	director_facebook_likes	4939	float64
country	5038	object	actor_1_name	5036	object
content_rating	4740	object	actor_1_facebook_likes	5036	float64
title_year	4935	object	actor_2_name	5030	object
color	5024	object	actor_2_facebook_likes	5030	float64
duration	5028	float64	actor_3_name	5020	object
genres	5043	object	actor_3_facebook_likes	5020	float64
plot_keywords	4890	object	num_voted_users	5043	int64
budget	4551	float64	num_user_for_reviews	5022	float64
gross	4159	float64	num_critic_for_reviews	4993	float64
aspect_ratio	4714	float64	movie_facebook_likes	5043	int64
facenumber_in_poster	5043	int64	cast_total_facebook_likes	5043	int64
imdb_score	5043	float64	movie_imdb_link	5043	object

表 2：电影数据集缺失值情况统计表

一、缺失值处理

针对电影语言数据项缺失，但电影发行国家字段完整的情况，筛选出符合上述缺失类型的 10 部电影，它们的发行国家均为 “USA”，因此将他们的电影语言均设为 “$English$”。

此外，根据电影的普遍属性，将部分字段的缺失值设为默认值，比如电影颜色默认为 “$Color$”，电影分级默认为 “G”，缺失的导演名字以及演员名字默认为 “$unknown$”，缺失的导演以及演员们的 $FaceBook$ 粉丝数如果在其他电影中出现过，则用相同数值补充，如果未出现过，则设为默认值 “0”，用户评论数和评论家点评数如果存在缺失值也设为默认值 “0”。完成上述处理后，对剩下仍含有缺失值的电影数据进行删失处理，获得不含任何缺失值的 3797 条电影数据。

二、异常值处理

对于电影语言为 “$None$” 的影片由于其发行国家为 “USA” ，所以将其语言修改为 “$English$”。根据美国电影协会的分类标准，将影片分为“G”、“PG”、“PG-13”、“R”、“NC-17” 五类，删除影片等级为其他的电影数据，最终得到包含 3707 条完整数据的数据集。

第二章　探索性数据分析

2.1　电影分类情况

首先按电影的语言可以将影片分类，绝大多数影片的语言为英语，共有 3554 部英语电影。除表中统计数据外，本数据集还包含俄语、阿拉伯语等 18 种语言的影片各一部。

发行国家	English	French	Spanish	Mandarin	German	Japanese	Cantonese	Hindi	Korean
影片数目	3554	32	23	14	11	10	8	7	5
发行国家	Portuguese	Italian	Dutch	Thai	Norwegian	Dari	Persian	Danish	Aboriginal
影片数目	4	4	3	3	3	2	2	2	2

表 3：按语言分类结果

相似地可以根据发行国家将所有影片分类，对于本数据集美国发行的电影数目最多，共有 2955 部影片，其次是英国发行了 309 部影片，具体结果汇总如表 4 所示。除此之外，还有波兰、希腊等 17 个国家各发行的一部影片也存在于本数据集中。

发行国家	USA	UK	France	Germany	Canada	Australia	Spain
影片数目	2955	309	97	81	58	39	20
发行国家	Japan	China	Hong Kong	New Zealand	India	South Korea	Ireland
影片数目	15	15	13	11	10	8	7
发行国家	Italy	Mexico	Denmark	Brazil	Thailand	South Africa	Argentina
影片数目	7	6	6	4	4	3	3
发行国家	Czech Republic	Iran	Netherlands	Norway	Russia	Iceland	Hungary
影片数目	3	3	3	3	3	2	2

表 4：按发行国家分类结果

根据美国电影协会的分类标准将电影分为五个等级（表 5），根据电影色彩将影片分为彩色片和黑白片（表 6）。

电影分级	R	PG-13	PG	G	NC-17
影片数目	1702	1313	569	117	6

表 5：按电影分级分类结果

电影色彩	Color	Black and White
影片数目	3594	113

表 6：按电影颜色分类结果

通过对数据集部分字段的极值分析可知，票房最高的电影为 2009 年上映的《阿凡达》，参与评分人数最多以及豆瓣评分最高的电影是 1994 年上映的《肖申克的救赎》，获得最多人评论的电影是 2001 年上映的《指环王》，*FaceBook* 上点赞数最多的影片是 2014 年上映的《星际穿越》，最受欢迎的导演是 *JosephGordon-Levitt*，他自导自演了影片《唐璜》。

2.2 数据集可视化分析

首先分别统计每年的电影数量以及电影票房随年份增长的变化情况，注意到从 1990 年开始电影事业繁荣发展，经过十几年的爆发式增长，在 2002 年达到最大值，一年共发行了 189 部影片，并在未来十年维持在每年发行 160 部影片左右。而关于每年的电影总票房，则是在 2012 年达到当年 112 亿元的最高总票房。

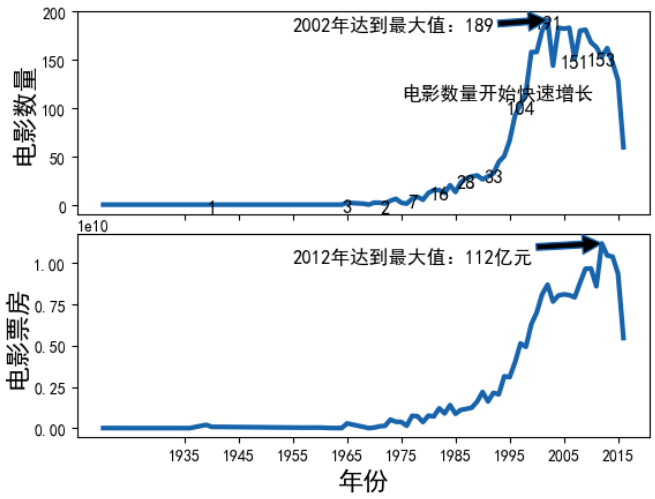


图 7：电影数量和电影票房随年份的变化情况

其次观察不用月份之间电影数目与票房的差异，发行电影相对较多的月份是 5 月和 1 月，在 9 月发行的电影总数目最少。而电影总票房则是在 12 月和 1 月数值较高，同样电影总票房的最小值也是出现在 9 月。由此可见，9 月是一年之中的电影淡季，发行的电影数目以及电影票房均相对较少。

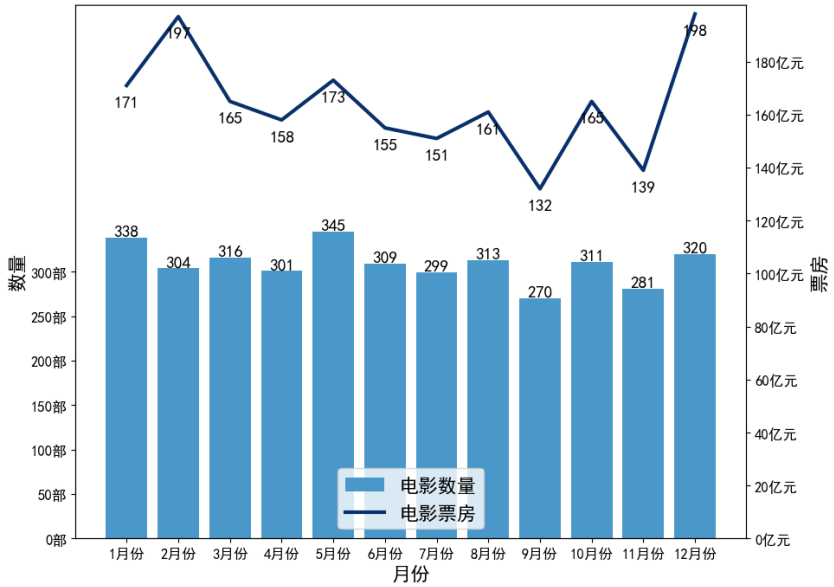
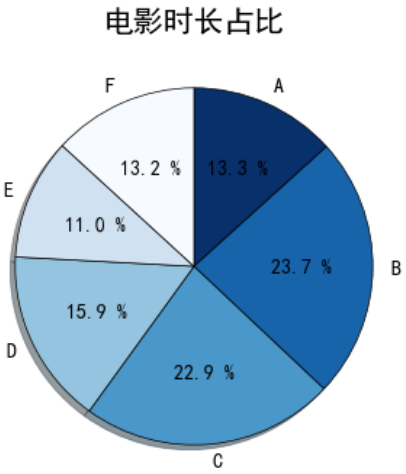


图 8：电影数量和电影票房按月份的统计情况

根据影片时长将电影划分为 6 个等级，片长不足 90 分钟的划分为一类，片长超过 130 分钟的划分为一类，此外片长在 90 分钟至 130 分钟之间的影片按每 10 分钟一级划分为四个等级。注意到片长在 90 至 100 分钟的影片数目最多，占总影片数目的 23.7%，但总体而言六组影片的数目差异并不大，可用作分类指标探究片长与口碑收益的关系。



时长	等级
(0,90]	A
(90,100]	B
(100,110]	C
(110,120]	D
(120,130]	E
(130,1000]	F

图 9：按时长分类的等级饼图（左）以及划分标准（右）

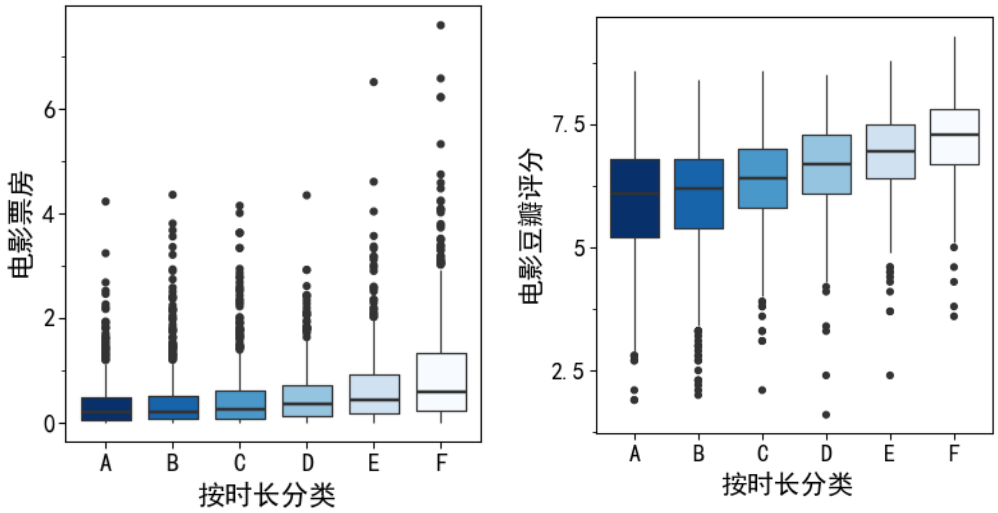


图 10：各时长分组的电影票房（左）和豆瓣评分（右）的箱线图

利用时长分组统计电影票房与豆瓣评分的分布情况，不难发现，随着影片时长的增加，电影票房呈现出上升趋势，并且电影票房的方差也存在增加现象，电影票房最高的电影时长超过 130 分钟。而对于豆瓣评分，也具有随着影片时长的增加而提高的趋势，但是豆瓣评分的方差却随着影片时长的增加而减小，这意味着对于时长越低的电影，豆瓣评分的差异化越大，时长越长的电影更可能收获高评分。

通过电影关键词和电影题材的词云图来直观展示文本信息，$love$、$death$、$murder$、$friend$ 在关键词中出现频率最高，这说明以爱情、朋友为主题的情感类电影和以死亡、谋杀为主题的悬疑类电影数据居多。$Sci-Fi$、$Comedy$、$Drama$ 等题材的电影数目相对较多，这说明在观测时段内发行带有科幻因素的电影以及喜剧、戏剧的电影更多。

图 11：电影关键词（左）和电影题材（右）的词云图

提取每部电影的题材信息，按照题材对电影进行分类，统计各题材电影数目以及总票房分布，发现戏剧题材的电影数目最多，喜剧题材的电影数目紧随其后，并且上述两种题材的电影总票房也分别为第四名和第二名，占有相当大的比重。对于题材为冒险类和动作类的电影，虽然电影数目相对较少，但却取得了相当高的总票房，这说明对于冒险类和动作类的电影更容易获得高票房。

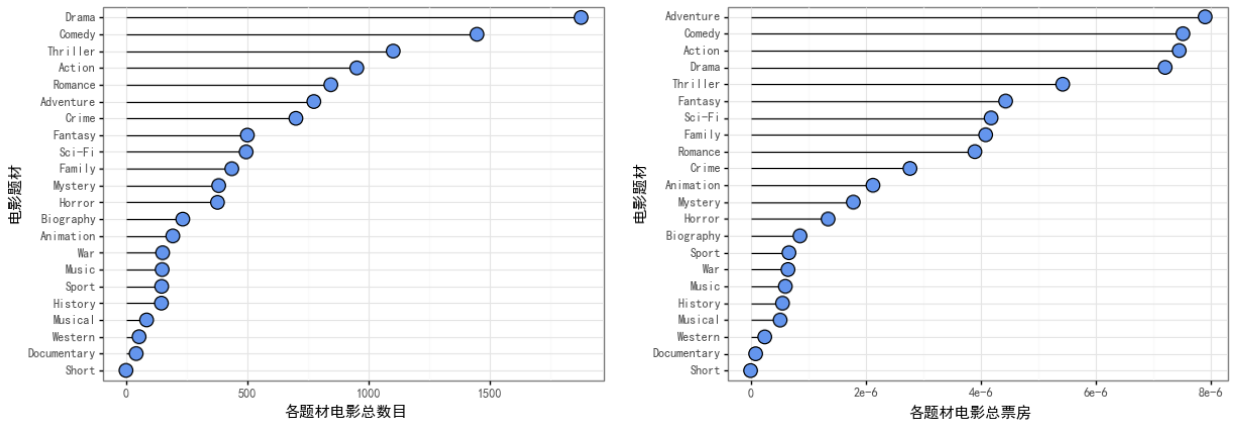


图 12：按电影题材分类的电影总数目（左）和电影总票房（右）的克利夫兰点图

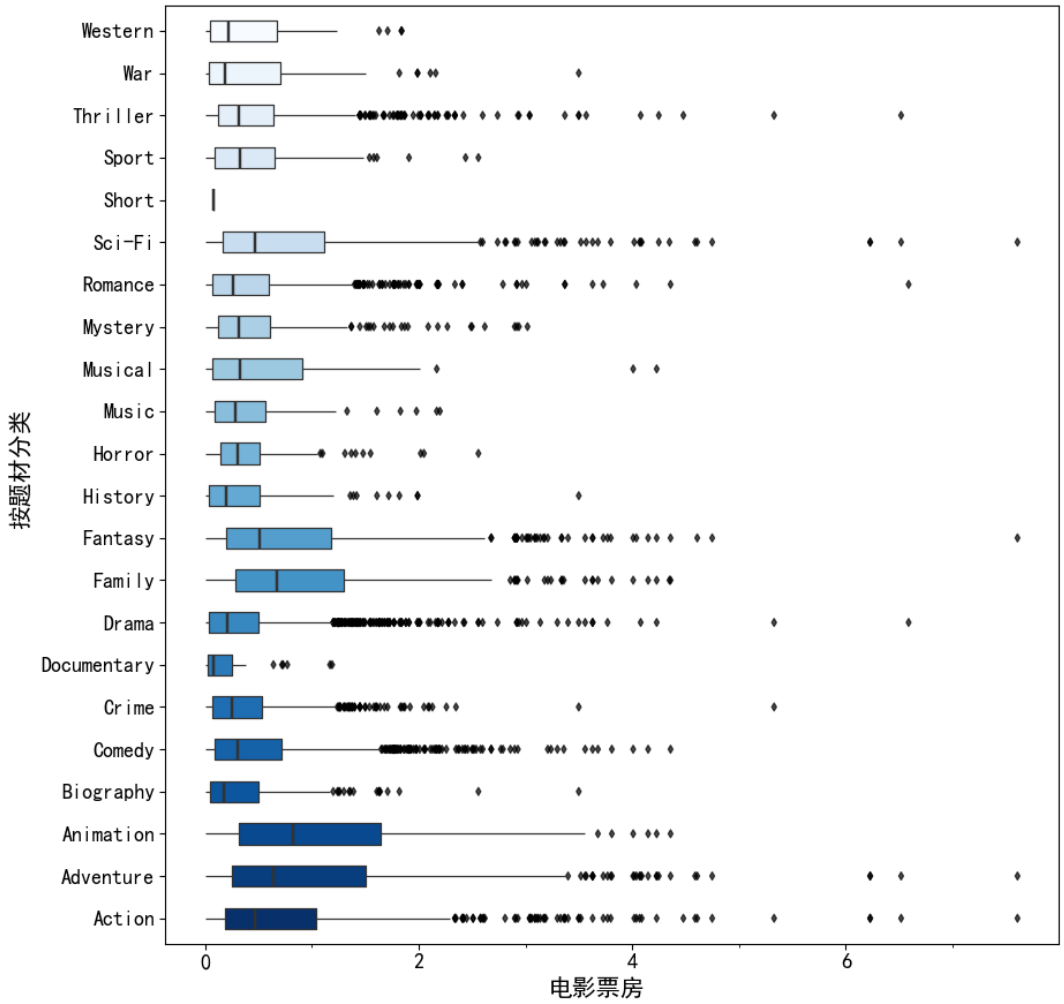


图 13：各题材电影的票房分布箱线图

用箱线图绘制各题材电影票房的详细分布情况，发现存在部分科幻幻想类以及冒险动作类的电影票房异常高于票房平均水平，这说明以上类型的电影更容易出现爆款电影，与此相比，纪录片、历史片、自传等题材的电影票房水平普遍较低，观看此类型电影的群体数目相对较少。

关注不同类型题材电影的豆瓣评分分布情况，音乐类和家庭类影片的评分两极化相对较大，票房水平相对较低的纪录片、历史片、自传等题材电影的豆瓣评分普遍高于其他类型的影片，这说明这三类电影的影片质量相对较高，观众对此类型影片的认可度也相对更高，但高口碑低收益的创作环境并不利于此类影片的长久发展。

最后检验部分指标间的相关性，发现对票房影响较大的是评分人数和评论数，说明观众评价对电影票房具有较大的影响，$FaceBook$ 上的电影喜爱总数与主演的演员影响力关系密切，此外评分人数与评论数、评论家点评数与电影点赞数也同样具备超过 0.7 的相关性，而电影预算与其他指标的相关性均不足 0.1，可以近似视为与其他指标独立。

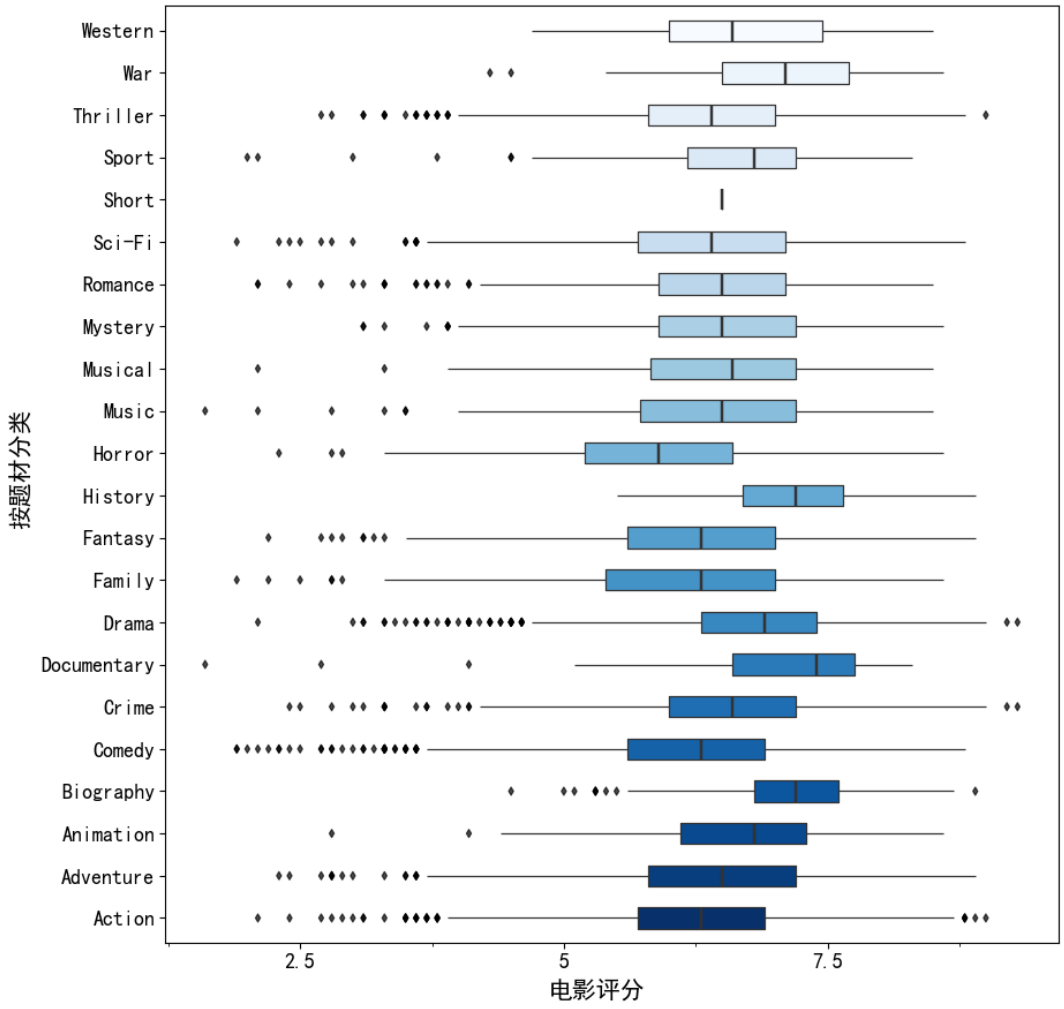


图 14：各题材电影评分分布箱线图

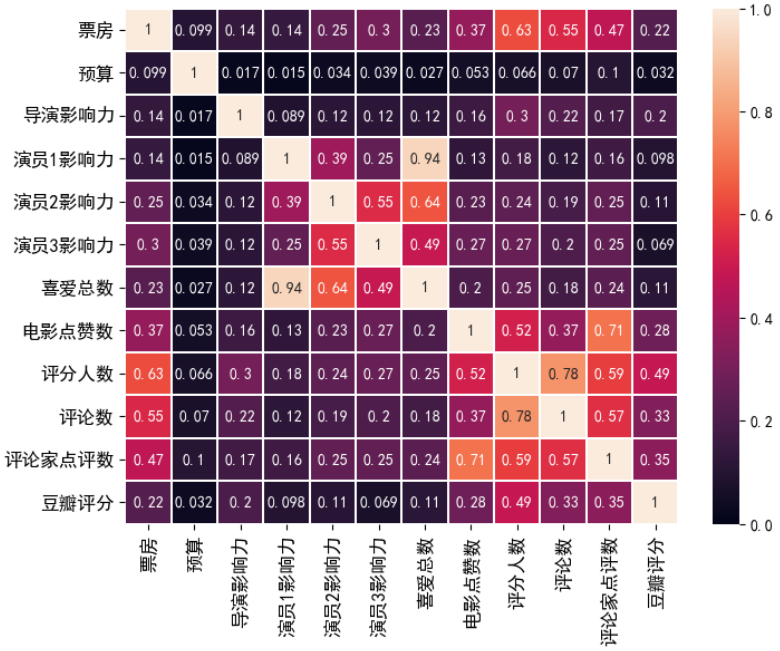


图 15：部分指标之间的相关性

第三章　票房预测问题

3.1　数据特征化处理

对于数据中的定性指标需要进行量化处理，需要变动的 12 个指标如表 16 所示，将“$gross$”作为因变量，并对其进行 $log1p$ 变换，提高响应变量的正态性，余下所有量化指标构成解释矩阵。

指标变量	调整方式
movie_title	计算电影名称长度
language	转化为分类哑变量
country	转化为分类哑变量
content_rating	转化为分类哑变量
title_year	分理出年月信息
color	转化为分类哑变量
genre_count	计算类型数目，按类型拆分计算平均票房和平均预算
keyword_count	计算关键字数
director_name	删除
actor_1_name	删除
actor_2_name	删除
actor_3_name	删除

表 16：特征化处理方式

将数据集分为训练集和测试集，其中预留 10% 用作预测，采用十折交叉验证方法依次构建多元线性回归模型、岭回归模型、$Lasso$ 回归模型、决策树模型和随机森林模型，完成票房预测问题，用均方根误差（$RMSE$）衡量模型拟合和预测效果。

3.2　多元线性回归模型

首先利用多元线性回归模型预测电影票房，每折交叉验证的训练集和验证集 $RMSE$ 差异化并不大，仅对于第一折和第十折存在验证集 $RMSE$ 略微偏高的情况，训练集和验证集的 $RMSE$ 均值分别为 1.51 和 1.59，预测集的 $RMSE$ 为 1.48，具备良好的预测效果。

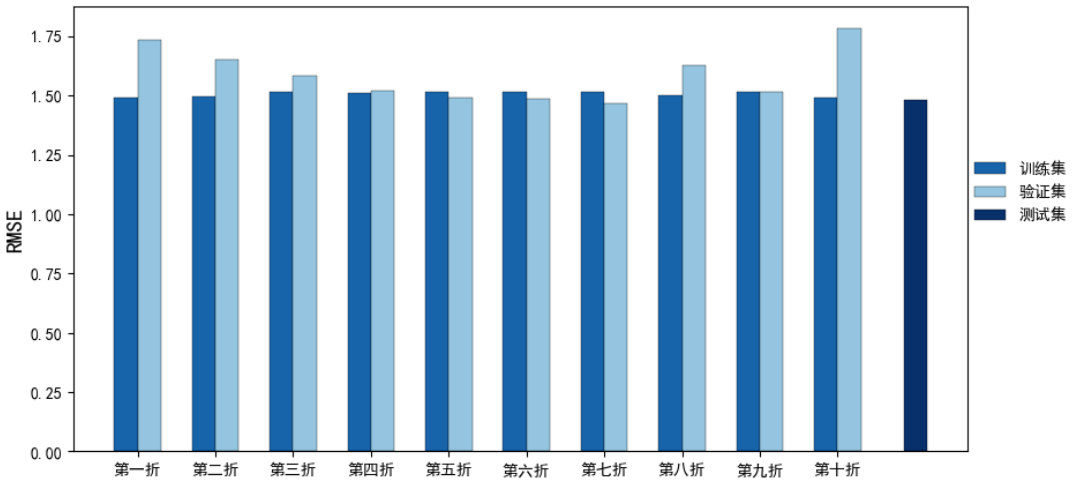


图 17：多元线性回归模型 RSME

3.3 正则化回归模型

考虑到图 15 所示的量化指标之间部分存在较强的相关性，可以在多元线性回归模型中引入正则化项，当引入 $L2$ 正则化项时，构建岭回归模型来降低模型的共线性，当引入 $L1$ 正则化项时，构建 $Lasso$ 回归模型在降低共线性的同时简化模型输入变量个数，达到降低模型复杂度的目的。

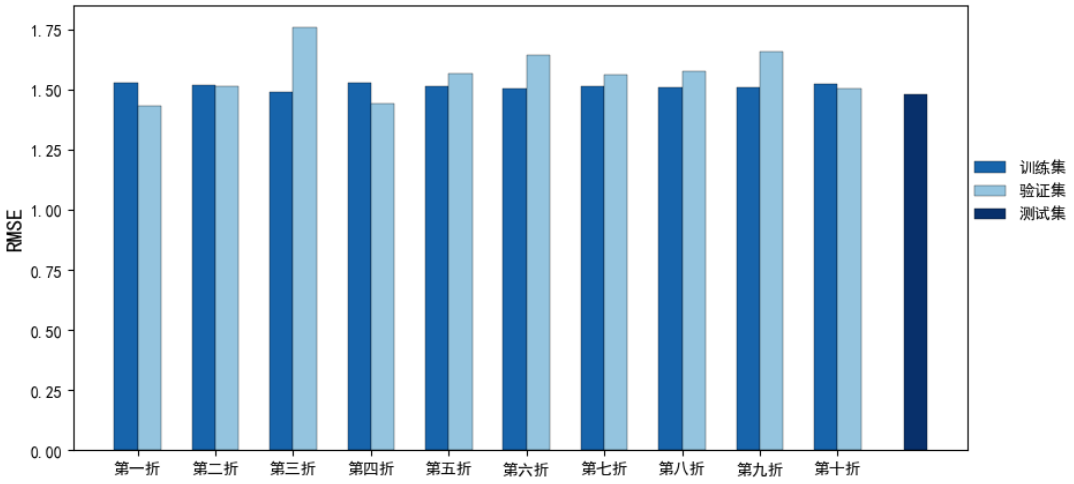


图 18：岭回归模型 RSME

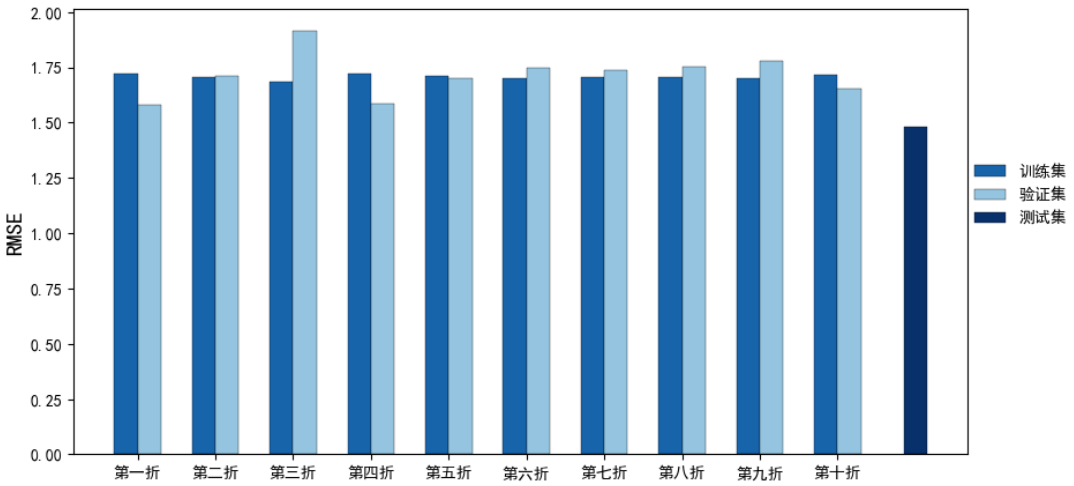


图 19：Lasso 回归模型 RSME

与多元线性回归模型相比，岭回归模型的训练集 $RMSE$ 几乎相同，但验证集和测试集的 $RMSE$ 略有降低，训练集、验证集和测试集的 $RMSE$ 分别为 1.52、1.57 和 1.47，这说明岭回归模型确实可以在一定程度上优化预测模型。另一方面，$Lasso$ 模型的训练集、验证集和测试集的 $RMSE$ 分别为 1.71、1.72 和 1.60，均方误差比多元线性回归模型略高，该模型牺牲部分预测精度来实现模型复杂度的降低，这一点也是具备现实意义的，用显著减少的变量个数达到理想范围内的预测精度。

3.4 决策树与随机森林

最后跳出回归模型，考虑采用树模型解决预测问题，分别构建决策树和随机森林模型。与回归模型相比，树模型具有更好的拟合效果和预测表现，对于决策树模型其训练集、验证集和测试集的 $RMSE$ 分别为 0.98、1.19 和 1.19，对于随机森林模型其训练集、验证集和测试集的 $RMSE$ 分别为 0.45、1.19 和 1.20。注意到决策树模型和随机森林模型具有相似的验证集和测试集均方根误差，但随机森林模型的训练集 $RMSE$ 仅为决策树模型的一半，这意味着通过增加多棵决策树构建随机森林可能存在过拟合的现象，仅使用决策树模型已经足够达到良好的预测效果。

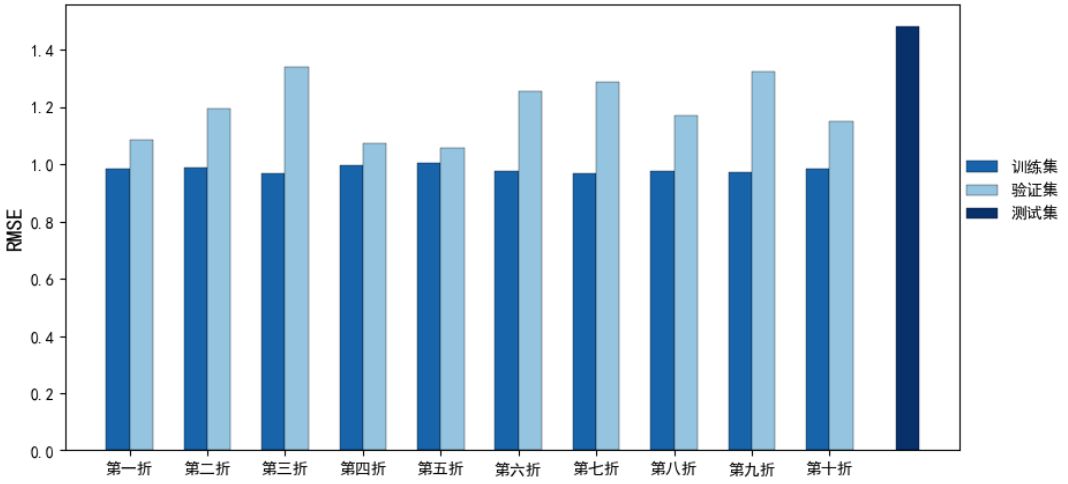


图 20：决策树模型 RSME

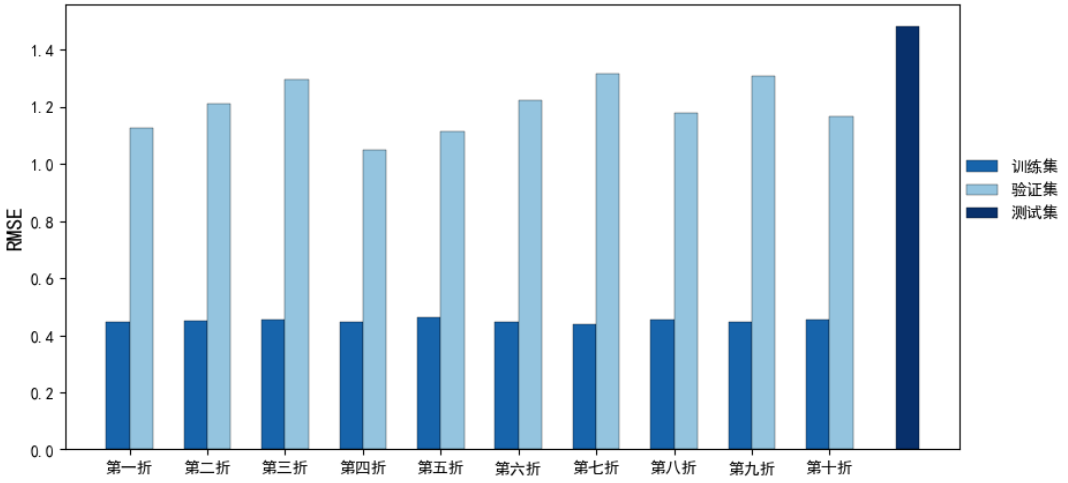


图 21：随机森林模型 RSME

第四章　电影推荐算法

4.1 针对电影相似度的推荐算法

将全部用户对某一电影的评分作为该电影的特征向量，使用 KNN 算法，对于用户输入的电影，选出与之最相似的十部电影作为推荐结果。

KNN 算法的基本原理是每当预测一个新值的类别时，在给定距离定义下寻找与其相近的 K 个值点，根据它们的类别来判断预测值点所属的分类，根据这一原理定义不同电影之间的相似程度，从而给出电影推荐，该算法可以采用以下三种模式。

$mode = 0 : KNNBaseline$

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v) \cdot (r_{vi} - b_{vi})}{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v)}$$
$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j) \cdot (r_{uj} - b_{uj})}{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j)}$$

$mode = 1 : KNNWithMeans$

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v) \cdot (r_{vi} - \mu_v)}{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v)}$$
$$\hat{r}_{ui} = \mu_i + \frac{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j) \cdot (r_{uj} - \mu_j)}{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j)}$$

$mode = 2 : KNNBasic$

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v) \cdot r_{vi}}{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v)}$$
$$\hat{r}_{ui} = \frac{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j) \cdot r_{uj}}{\sum_{j \in N_u^k(i)} \mathrm{sim}(i,j)}$$

定义类结构 $Movie\_KNN\_recommender$ 用作电影推荐，该类除了包含初始化函数之外，还含有寻找 KNN 相邻电影的函数 $search\_movie\_neighbors(self, movieID, num = 10)$ 以及输出电影名称的函数 $recommend\_movies(self, movieID, num = 10)$。

示例：考虑《钢铁侠》的电影推荐。依据该方法得到的推荐电影有同为漫威系列的《钢铁侠 2》、《复仇者联盟》、《战警》、《战警：第一站》、《蜘蛛侠》、《黑客帝国》、《暗黑骑士》、《星际迷航》等类型十分相似的电影，也有《美食总动员》和《宿醉》两部类型差异化略大的影片。

4.2 针对用户相似度的推荐算法

除了根据电影类型的相似度进行推荐之外，也可以考虑用户之间的相似性与差异度，实现对于不同用户的个性化推荐，丰富用户的推荐结果。将一个用户对于全部电影的评分作为该用户的特征向量，同样使用 KNN 算法，为每一个用户推荐电影。该算法同样具有 $KNNBaseline$、$KNNWithMeans$、$KNNBasic$ 三种模式，通过选择出与当前用户最相似的 10 个用户，分析他们看过且当前用户没看过的电影，在其中选择评分最高的 10 部影片作为最终推荐。

首先将包含用户和电影信息的数据集 $user\_movie$ 按 9∶ 1 拆分成训练集和测试集，定义类结构 $Personal\_KNN\_recommender$，该类除了包含初始化函数外，定义用来寻找相似用户的函数 $search\_user\_neighbors(self, usrID, num = 10)$，将相似用户看过的电影作为推荐电影的基础集合，定义函数 $recommend\_movies(self, usrID, num)$ 在基础集合中选择评分最高的电影完成推荐。

示例：考虑对用户 ID 为 66 的用户进行电影推荐。该用户一共对 345 部电影进行过评分，并对其中的 59 部作品打出过最高分 5 分。依据上述算法为该用户推荐的 10 部电影依次为《拯救大兵瑞恩》、《教父》、《美国 X 档案》、《谋杀绿脚趾》、《异形》、《天使爱美丽》、《杀出个黎明》、《死亡幻觉》、《绿里奇迹》、《教父 2》，不难发现，对该用户的电影推荐基本都带有悬疑色彩。

第五章　结论与总结

本文首先通过探索性数据分析发掘数据集的统计性信息与内在规律，之后解决了票房预测问题并完成了电影推荐算法。完整代码以及原始数据集已经上传至 $Github$，链接为 https://github.com/QforeverQ/movies。

在对原始数据集进行数据清洗后，在探索性数据分析的过程中，首先根据部分定性指标对电影进行分类，发现数据集中以美国发行的英语电影居多，绝大部分电影是彩色电影，并且适合 17 岁以上的观众观看。此外还可以发现，票房最高的电影为《阿凡达》，评分人数最多以及豆瓣评分最高的电影是《肖申克的救赎》，讨论度最高的电影是《指环王》，$FaceBook$ 上最受欢迎的影片是《星际穿越》等等。

统计每年的电影信息，发现 1990 年到 2000 年是电影飞速发展的黄金时期，在此期间

每年发行的电影数目增长迅速，电影票房也实现质的飞跃。关于不同月份之间电影市场的差异化，发现冬季电影票房普遍高于其他季节，每年的 9 月是发行电影数目最少、票房最低的淡季。为了探究电影时长与电影票房以及豆瓣评分之间的关联，将数据集按片长划分为 6 个等级后统计发现，随着影片时长的增加，电影票房和豆瓣评分均有增长，这意味着时长长更有可能产出收益与口碑双丰收的电影。

关注数据集中包含的文本类信息，通过词云图发现，数据集电影中出现最多的关键词是 “爱”、“谋杀”、“死亡”，涉及较多的题材类型包括 “科幻”、“喜剧”、“戏剧” 等等，这反映出情感类电影和悬疑类电影的数据居多，带有科幻色彩的电影以及喜剧或戏剧具有比较高的发行量。提取每部电影的题材信息，汇总不同类型题材电影的差异性，不难发现，纪录片、历史片、自传等题材的电影票房水平普遍较低，但这类电影的口碑一般远高于其他类型的影片，科幻幻想类和冒险动作类的电影更容易获得超高票房。

在解决票房预测问题之前，首先需要对数据集进行特征化处理，从文本信息中获取更多的数量指标，并对定性指标引入哑变量之后，依次建立回归模型和树模型实现票房预测。结果发现，在多元线性回归模型中加入 $L2$ 正则项构建岭回归模型能够实现预测精度的改善，在 $RMSE$ 作为衡量标准的情况下，树模型的拟合和预测效果明显优于回归模型。

最后关于电影推荐算法，以用户给观看过的电影的打分记录为分析基础，从电影相似度和用户相似度两个角度，基于 KNN 算法建立推荐模型。针对电影相似度的推荐算法是将全部用户对某一电影的评分作为该电影的特征向量，并以《钢铁侠》为例进行电影推荐，10 部推荐影片中有 8 部与《钢铁侠》同类型的漫威电影。针对用户相似度的推荐算法则是将一个用户对于全部电影的评分作为该用户的特征向量，例如为用户 ID 为 66 的用户推荐 10 部电影，大部分推荐电影都属于悬疑题材，这种方法将用户之间的差异性考虑进来，实现用户的个性化电影推荐。