这个主创值不值？

—基于社交媒体的电影主创票房贡献度分析与推荐模型

作品文档

POPCORN 团队

2017 年 11 月 10 日

解题思路

为了评估主创团队对电影票房的贡献，我们将电影票房作为贡献度的拟合值。通过建立**主创画像**寻找主创与影响电影票房的主要因素的数学关系，进而求得**主创对于电影票房的贡献值**，根据主创贡献值给出**主创推荐模块**。

1）主创画像：

主创画像我们选取每部电影演员、导演的多维度变量（导演包括电影节奏/剪辑/分镜/情节组织、角色塑造、整体评价；演员包括演技、颜值、身材、声线、气质、台词功底、整体评价）来对主创进行画像建模，对微博电影评论做文本情感分析，分为 Sentiment 和 Aspect 两部分，分别使用 CNN+Dense 模型和 CNN 模型，最后将情感分析的结果与主创基本信息通过主创姓名映射来做匹配，最终得到主创的用户画像结果。

2）电影票房预测模型：

为了评估主创团队对电影票房的贡献，我们将电影票房作为贡献度的拟合值。通过建立主创画像寻找主创与影响电影票房的主要因素的数学关系，从而从多个维度分析主创对电影票房的贡献度。

从票房预测回归模型相关性检测发现影响票房的主要因素有电影的排片率、各电影网络平台想看平均想看数量、各电影垂直网站平均评分，为了找出主创特别是导演和演员对电影票房的贡献影响，我们需要在主创与排片率、想看数量、评分之间建立模型，希望用主创的相关画像信息估计出上述三个变量，进而通过这三个维度间接反应主创对票房的贡献影响。

3）主创推荐模块：

该模型可以对主创团队进行对某电影票房贡献值的评估，同时可对特定的要求作出最优的主创团队的推荐。我们团队的主要工作包括：

（1）主创团队的评估：输入导演、演员、题材等条件，得到预期的票房和主创团体对电影票房的贡献值。例如输入导演为冯小刚、演员为范冰冰、赵丽颖、胡歌和陈学冬，题材为 剧情和冒险，则估计的票房值为 529404248。

（2）主创团队的推荐：自定义输入影片题材、预算等部分变量值，系统分析不同主创团体时，预期的电影票，得到最优主创团体选择方案，以最少的成本

获得最大的收益。例如输入影片题材为冒险，预算为 100000000，是否续集等变量，系统推荐出导演为王岳伦，演员为黄晋琨、赵奔等演员。

对于发行公司、制作公司可通过该模型计算出明星对某部电影票房的贡献值，来选择最合适的主创团队，可提供给导演来帮助他们挑选合适的且对票房贡献大的演员。

目录

第一章　引言

近几年随着电影市场的热度提升，大量资本涌入电影行业。但中国电影产业尚处于发展阶段，市场的不成熟使得电影投资呈现出高风险高回报的特点。电影的创作最重要的三大部分由技术制作团队，导演和演员组成，演员是影响影片最终质量和品位的关键。不用名导演和名演员，怕票房没有保证；用了名导演和名演员，电影投资成本又飙升，投资风险急剧扩大，但要寻找出一个兼顾票房和名演员之间的平衡点是比较困难的。

在演员、导演、道具、后期制作、宣发等几项主要电影制作成本中，明星身价飙升，制作成本随之大幅提升，在明星演员身价水涨船高的情形下，现今电影蓬勃发展，带动的是随着电影好评而大红大紫的演员，明星光环越亮的演员片酬相对不便宜，在电影成本有限的情况下，我们能通过微博等社交媒体对明星的讨论度，帮助电影团队节省时间成本，收集相关信息作为拍摄续集的依据，电影公司可以透过我们团队所提供的数据资料去了解什么类型（喜剧、动作、冒险...）的电影会增加观赏意愿，不但能准确的锁定电影族群，更能从中了解观众喜爱的宣传方式，方便电影公司未来作出更有力的决策。

第二章　数据说明

2.1　数据源与数据采集

2.1.1 数据来源

我们文本数据主要是从新浪微博、豆瓣电影、中国票房、娱票儿票房分析、百度百科这五大平台中获取，具体爬取的信息如下表 2.1 所示：

表 2.1 数据来源

平台	数据类型	具体数据、来源网址	数据总量
新浪微博	主创微博基本信息	用户微博链接、用户 id、用户昵称、用户头像、关注数、粉丝数、爬取时间、爬取的链接	157204
	主创微博内容	微博内容、发布工具、转发数、评论数、点赞数、是否长微博、微博配图、视频的地址、是否转发、转发的原微博、原微博的转发数、热门评论、评论	
	电影话题微博内容	电影话题链接、话题、发表微博用户链接及用户的 id、昵称、头像、粉丝数、关注数、微博发布时间、微博内容、发布工具、转发数、评论数、点赞数、是否长微博、微博配图、视频的地址、是否转发、转发的原微博、原微博的转发数、热门评论、评论	509125
	来源网址	http://weibo.com	
豆瓣电影	电影基本信息	电影 id、电影名称、上映年份、导演、演员、类型、制片国家/地区、语言、片长、评分、评分人数、想看人数、看过人数、tags、剧情简介、问题个数	95073
	电影评论	短评内容	4454
	来源网址	http://m.douban.com/movie/recent/	
中国票房	电影基本信息	电影名称、上映时间、国家及地区、片长、类型、累计票房、导演、主演、制作公司、发行公司	3255
	来源网址	http://m.cbooo.cn /	
娱票儿票	电影基本信息	电影名称、上映时间、国家及地区、片长、类型、	2365

房分析		累计票房、剧情简介、制式、导演、演员、制片公司、发行公司、	
	电影票房信息	首日票房、首周票房、累计票房、上映前后每天趋势（当日票房、票房占比、拍片占比、场均人次、场次、人次、上座率、座位、排座占比）	
	影评相关信息	评分、电影口碑，不同平台（豆瓣、时光网、格瓦拉、猫眼、糯米、淘票票）的评分、评分人数、想看人数、评论数，用户评论关键词占比	
	来源网址	https://piaofang.wepiao.com/	
百度百科	明星基本信息	主创个人页面链接、中文名、外文名、别名、国籍、民族、星座、血型、身高、体重、出生地、出生日期、职业、毕业院校、经纪公司、代表作品、获奖经历	11164
	来源网址	http://wapbaike.baidu.com/	

2.1.2 爬虫原理

我们使用 Scrapy 爬虫的框架。Scrapy 是使用 Python 语言编写的开源爬虫框架，可对互联网中的网页内容进行抓取，并从中提取出结构化数据，提取到的数据可用于资料收集、舆情分析、数据挖掘等多个领域。框架中的各组成部分及功能如下图 2.1 所示。

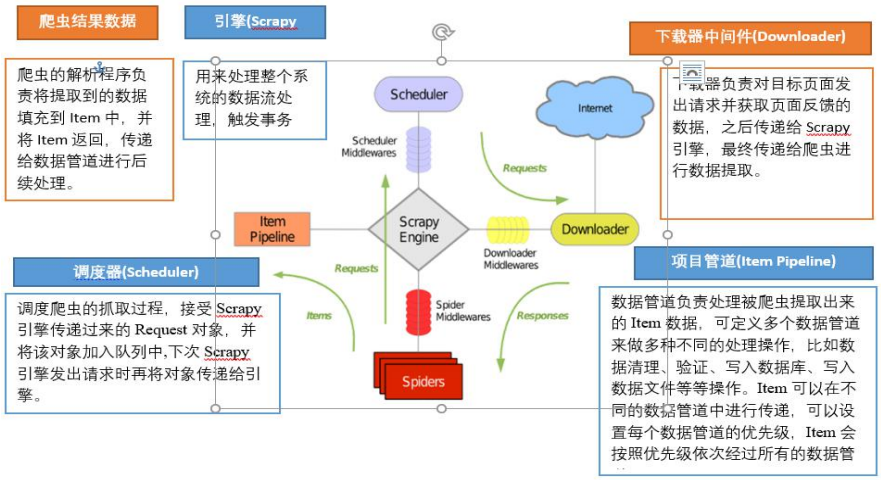


图 2.1 Scrapy 框架介绍

Scrapy 流程图如下图 2.2 所示：

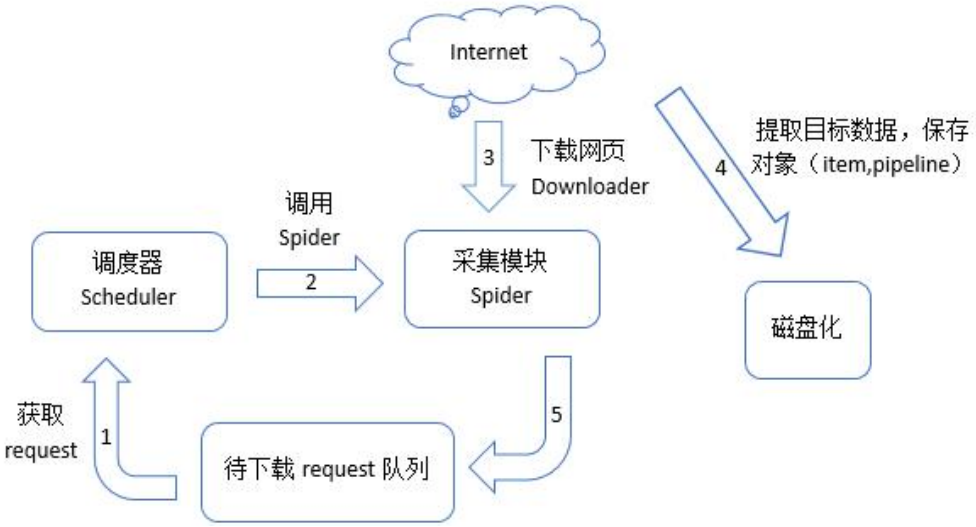


图 2.2 Scrapy 流程介绍

以下，我们以爬取新浪微博电影的影评为例，说明获取数据的过程。

1、获取新浪微博的 cookie:

2、页面的跳转和解释：

进入微博电影分类的页面做起始页：

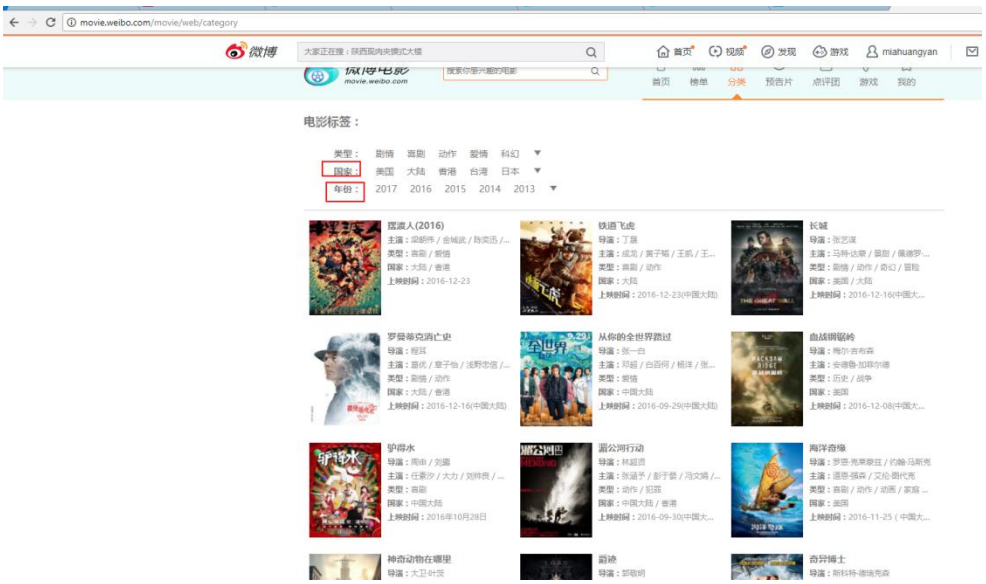


图 2.3 微博电影页面图

通过选择国家和年份向该页面发起 request 获取 html 源码，获取电影列表 id，进而转到电影详情页，如下图 2.4：

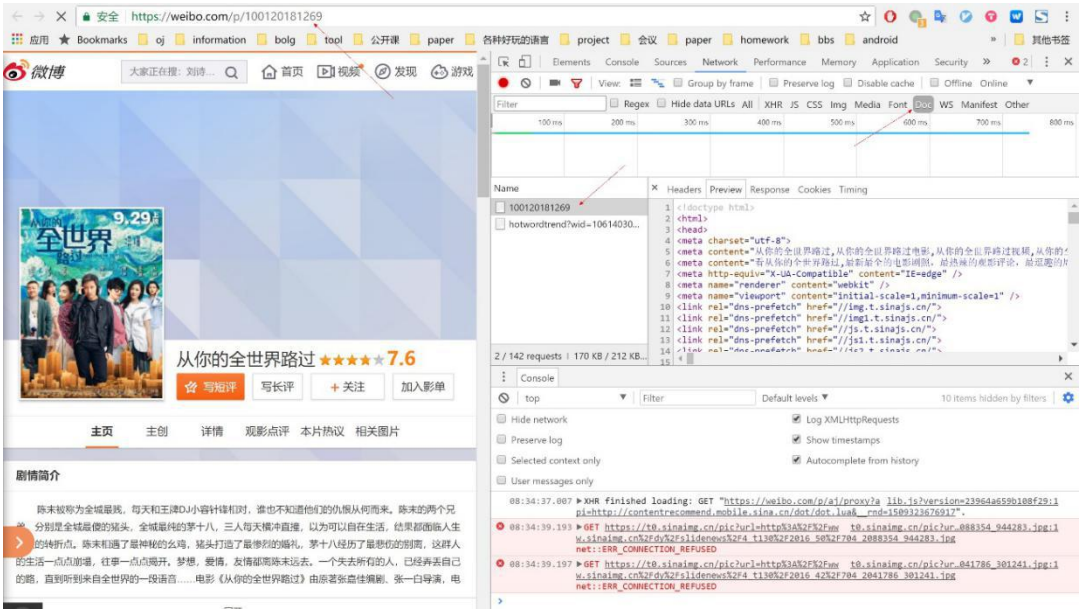


图 2.4 电影微博源码图

然后该页面的源码请求方式在如图红色箭头所指位置，通过分析可以知道该电影的关注量，点评数，视频播放数量在 domid 值为 Pl_Core_T8CustomTriColumn__16 的 script 脚本中，利用正则表达式就可以提取下图 2.5 的数值信息。

图 2.5 电影微博数据图

则当没有更多的点击项时，直接提取当前显示的评论信息，反之通过更多提取点击更多后跳转到的 url，而其在源码中的位置有唯一的标识：class="WB_cardmoreWB_cardmore_noborder S_txt1 clearfix 通过该标识正则提取即可,如下图 2.6。

图 2.6 话题评论图

当点击更多后，变化跳转到如下页面,如下图 2.7：

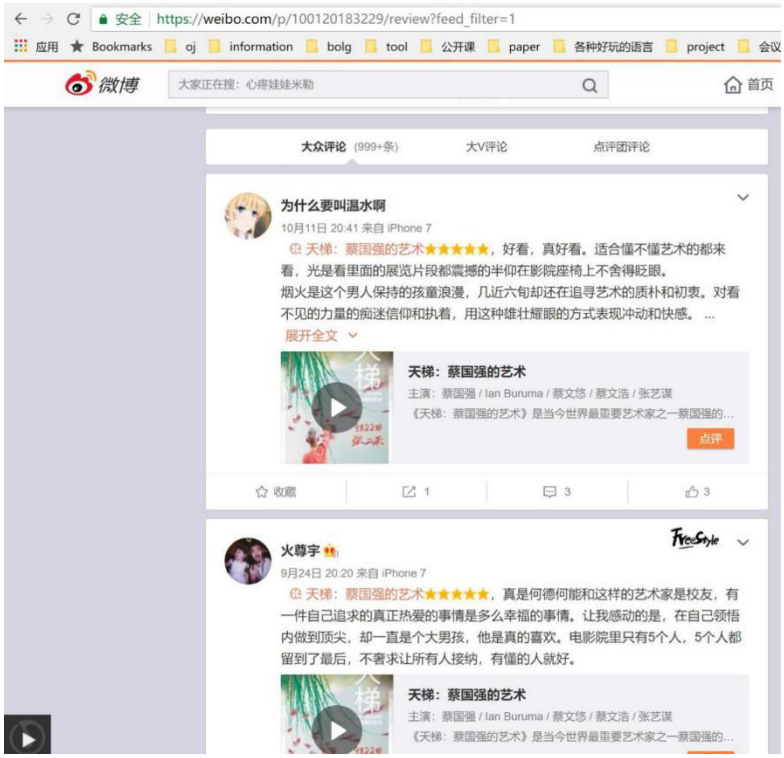


图 2.7 大众点评页面图

通过分析发现，每个页面最多动态加载两次 pagebar,如下图 2.8：

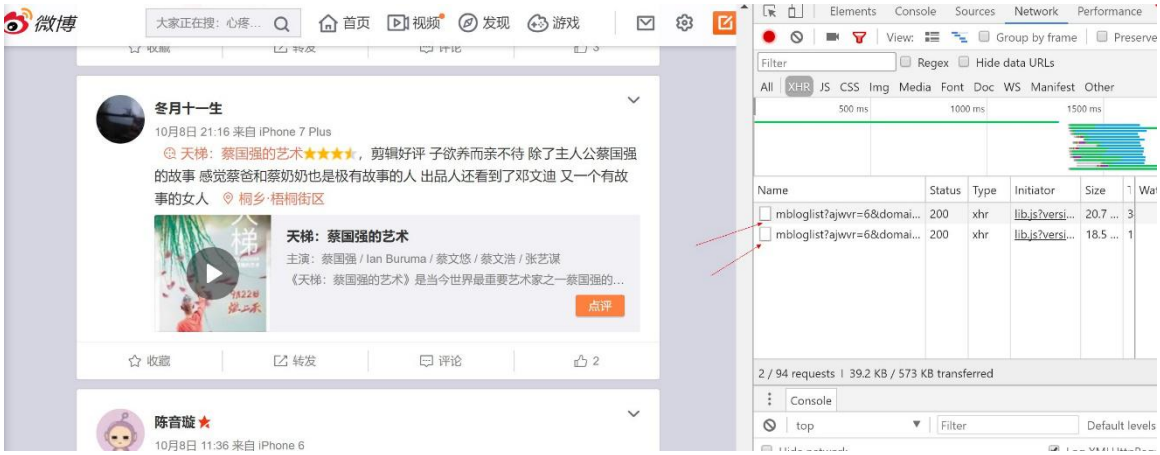


图 2.8 话题微博源码图

通过分析可以发现这些参数的信息可以在返回的页面中用标识 pids=Pl_Core_MixedFeed__57 提取对应的 script 来解析， 而 page 和 pre_page 的值是一样的， current_page 每页加 3，则每次加载后获取参数，然后构造 get 请求参数， 在获取两次 pagebar 生成的评论信息，接着跳转进入下一页，直到没有

下一页为止。

2.1.3 数据存储

为了更好的管理数据，我们把从五个不同平台爬取的数据存储在数据库中，对数据进行清洗和筛选，整理成建模所需的数据，存储在不同的数据库表中。

1、数据库系统 MongoDB

1）Mongodb：2007 年 10 月由 10gen 团队所开发，2009 年 2 月首度推出[1]。MongoDB 作为一款性能优良，功能丰富，支持海量数据存储的产品受到很多商家的青睐，是一个基于分布式文件存储的数据库，由 C++语言编写，旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。

2）特点：面向集合存储，易存储对象类型的数据，支持 RUBY，PYTHON，JAVA,C++，PHP 等多种语言，文件存储格式为 BSON（一种 JSON 的扩展），支持完全索引，包含内部对象，支持查询，支持复制和故障恢复。

3）优点：高性能、易部署、易使用，存储数据非常方便。

2、数据库表设计

通过对推荐系统功能与采用的模型进行数据分析，得到影响主创贡献度的关键因素，进而设计出本推荐系统的数据库表。主要的数据库表如下：

1）新浪微博-用户信息（Information）

表 2.2 新浪微博-用户信息

_id	NickName	Gender	Province	BriefIntroduction	Birthday	Num_Tweets	URL
用户 ID	昵称	性别	所在省	简介	生日	微博数	首页链接
Num_Follows	Num_Fans	VIPlevel	Sentiment	SexOrientation	city	Authentication	
关注数	粉丝数	会员等级	感情状况	性取向	所在城市	认证	

2）新浪微博-微博内容（Tweets）

表 2.3 新浪微博-微博内容

_id	ID	Content	PubTime	Co_oridinates	Tools	Like	Comment	Transfer
用户 ID-微博 ID	用户 ID	微博内容	发表时间	定位坐标	发表工具/平台	点赞数	评论数	转载数

3）豆瓣电影-电影基本信息

表 2.4 豆瓣电影-电影基本信息

subject_id	name	year	directors	actors	genres	channel	
电影 id	电影名称	上映年份	导演	演员	类型		
runtime	languages	discussion	countries	average	vote	wish	watched
片长	语言		制片国家	评分	评价人数	想看数	看过数
summary	tags	stars	review	question	image	summary	tags
剧情简介	标签		短评数		相关图片	剧情简介	标签

4）豆瓣电影-电影评论

表 2.5 豆瓣电影-电影评论

url	movie_name	short
电影首页链接	电影名称	短评内容

⇩

short_name	short_stars	short_time	short_supporter	short_comment
评论用户名	评分	评论时间	支持者	评论内容

5）中国票房-电影基本信息

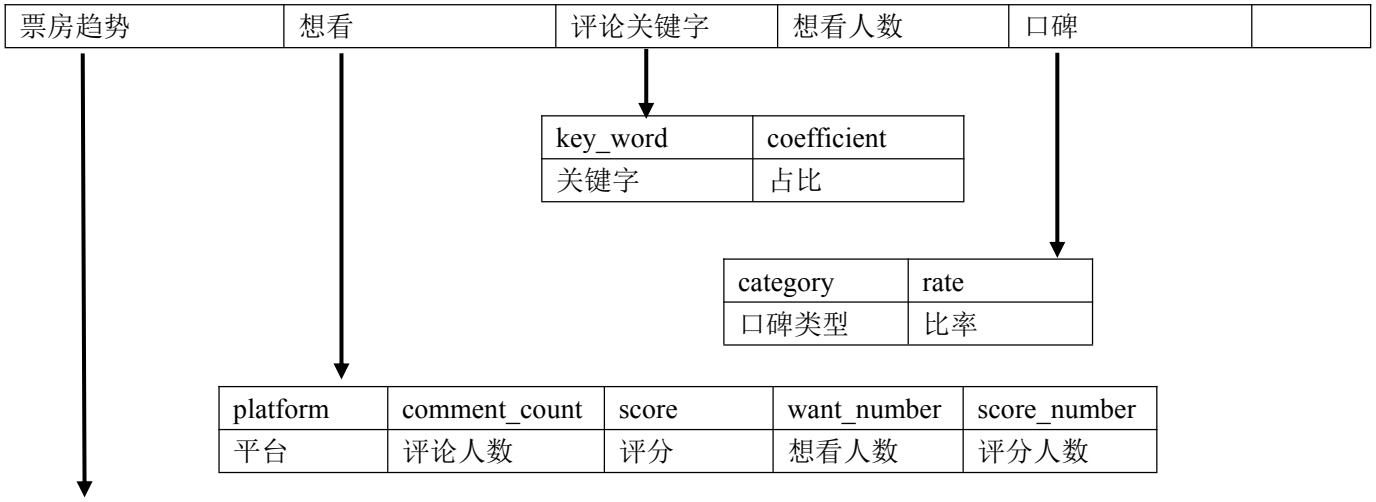
表 2.6 新浪微博-用户信息

movie_url	movie_name	show_time	country	movie_type	movie_time
电影首页链接	电影名称	上映时间	制片国家/地区	电影类型	片长
director	actor	production_company	distribution_company	box_office	source_platform
导演	演员	制作公司	发行公司	累计票房	数据平台

6）娱票儿票房分析-电影基本信息

表 2.7 新浪微博-用户信息

movie_name	release_date	format	plot	district	director
电影名称	上映时间	形式	电影简介	制片国家/地区	导演
actor	score	waati	movie_type	movie_url	
演员	评分	片长	电影类型	首页链接	
distribution_company	production_company	total_boxoffice	firstday_boxoffice	irstweek_boxoffice	
发行公司	制作公司	累计票房	首日票房	首周票房	
movie_trend	movie_want	movie_comment	want_count	word_of_mouth	



7）百度百科-明星基本信息

表 2.8 新浪微博-用户信息

url	chinese_name	english_name	other_name	country	nation
首页链接	中文名	外文名	别名	国籍	民族
constellation	blood	height	weight	birthplace	birthdate
星座	血型	身高	体重	出生地	出生日期
graduate_institutions	brokerage_agency	representative_works	awards	occupation	
毕业院校	经纪公司	代表作品	获奖情况	职业	

time	reward
获奖时间	获奖内容

第三章　主创贡献度计算模型与技术架构

我们提出了电影主创贡献度计算模型，如下图 3.1 所示即为模型的具体流程，一共分为网络爬虫获取数据、数据预处理、主创画像、票房预测、推荐系统及可视化 5 部分。

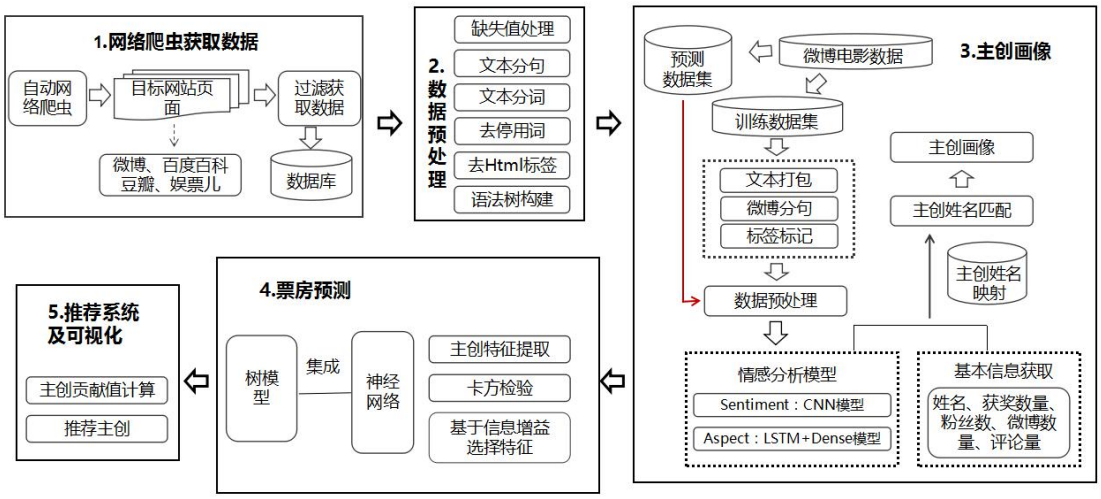


图 3.1 电影主创贡献度计算模型

接下来，我们分别对主创画像、票房预测和推荐系统进行详细介绍。其中模型的数据采集部分在第二章已经介绍过(详见 2.1 数据源与数据采集)。第四章主创画像，主要介绍了如何从多个维度进行画像建模以及主创画像结果分析。第五章票房预测，主要介绍了通过主创画像对电影票房的预测，得到主创对电影票房的贡献度。第六章推荐系统及可视化，介绍了主创推荐模型以及结果的展示。

我们系统的架构如下图 3.2 所示，底层为数据库，我们选取 Mongodb 来存储数据，有主创画像、电影票房预测、主创推荐三个算法模块，豆瓣、百度百科、余票儿、中国票房和新浪微博五个数据采集模块，可视化部分为主创画像与推荐系统。

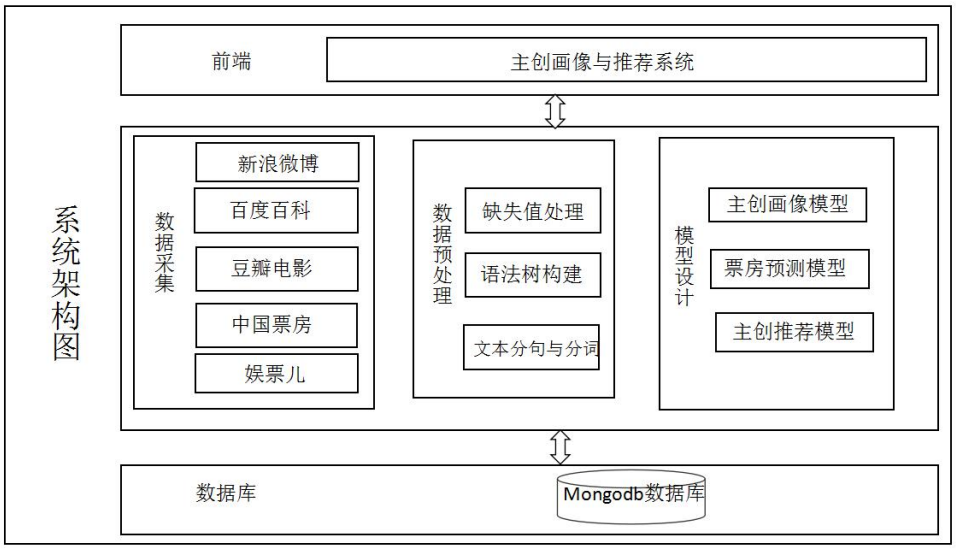


图 3.2 系统架构图

开发工具：编程语言：Python 3.5.1　　Python 2.7

机器学习及数据分析工具：scikit-learn SAS

开发平台：PyCharm

爬虫框架：Scrapy

第四章　主创画像模型

主创画像我们选取每部电影演员、导演的多维度变量（导演包括电影节奏/剪辑/分镜/情节组织、角色塑造、整体评价；演员包括演技、颜值、身材、声线、气质、台词功底、整体评价）来对主创进行画像建模，对微博电影评论做文本情感分析，分为 Sentiment 和 Aspect 两部分，使用 CNN+Dense 模型，最后将情感分析的结果与主创基本信息通过主创姓名映射来做匹配，最终得到主创的用户画像结果。

4.1　主创画像维度

主创画像信息主要来源于微博电影话题数据，通过对电影话题微博进行基于 aspect 细粒度的情感分析，从而产生主创画像的评价维度。

1）知名度

明星拥有了知名度才具备其个人品牌价值及社会影响力,知名度是衡量一个明星被公众知晓和了解的程度，是一个量的指标，公众对某个明星印象越深，印象越好才会对其作品或相联系的产品越多关注。知名度又分为人气热度和曝光度对其进斤衡量，人气热度可通过微博粉丝量、关注量来衡量，曝光度为报纸、电视、网络、杂志封面的曝光率加权所得。但曝光率也不是越高越好，一些电影明星会保持较低的曝光率以提升公众的期待，而真人秀会大大增加明星曝光率，这能让一些不为大众所熟知的普通明星迅速收获大量粉丝，但曝光的明星新闻若是负面新闻则会削减其个人影响力，因此，我们目前只选取人气热度来衡量明星的知名度。

数据来源及变量设置：数据来源于主创微博主页的粉丝数量，若未找到相关微博，则设为 0。

2）美誉度

美誉度是一个明星受到公众的信任、好感、接纳、欢迎的程度，是评价一个明星声誉好坏的指标，侧重于质的评价。公众对一个明星的好感的发展程度是分四个阶段喜爱、支持、赞许、信赖，所这四个情感程度分别去衡量大众对其喜爱

的明星的声誉评价。声誉评价的好坏是非常重要的一个定性的影响因素，因为高的知名度、关法度并不一定就说明这个明星有更高的价值，只有正面的积极的个人形象才能带来的其品牌价值的提升，才是真正的价值。究其原因，我们可以通过明星的获奖数量和主创参与全部电影的平均票房来定义明星的美誉度。

数量来源及变量设置：根据主创姓名匹配百度百科页面的获奖记录，统计获奖数量；通过查询娱票儿历史票房记录，匹配主创姓名，将主创参与过的相关作品的和平均票房作为主创影响力的拟合数据。

3）形象健康指数

形象健康指数是一个明星其个人形象给公众留下的正面印象的程度，健康在此时一个好印象的形容词。包括社会责任感和婚姻爱情家庭责任感。社会责任感是明星对社会事件的态度和行为，衡量明星是否做了一个公众人物对社会应尽的责任。比如，是否在突发性灾难出现时表现出对事件的关屯、和尽绵薄之力。《中国慈善家》每年会发布中国慈善名人排行榜，其中包括各行各业的名人，此榜单不是只考量捐款数额，而是综合个人捐款，间接募款，公益行动和公益影响力等方面，经过细致考量得到的结果。此榜单的排名可用来衡量明星的社会责任感。

数据来源及变量设置：统计 2013-2016 中国慈善家名人榜，设置虚拟变量来衡量其形象健康指数，如某明星榜上有名则对其设置为 1，相反则设置为 0。

4）粉丝忠诚度

粉丝是崇拜追捧某个明星的一个群体，大多是年轻人。大众对明星的喜爱递进顺序是从知道到偏好再到忠诚最后是目标。当大众对明星的喜爱达到忠诚和目标层次就可称得上是粉丝了。这里讨论的只是粉丝这个群体，粉丝对于明星的个人品牌价值的构建是有很大的帮助和提升的，当然也存在一些非理性的粉丝会对明星造成一些负面影响。在此将粉丝忠诚度分关注度和认同度 2 个方面去衡量。粉丝会持续的关注明星，转发和讨论明星的新闻，帮助明星扩大影响，所以粉丝对明星的关注度关系到明星的传播能力。粉丝对明星会有不同程度的认同度，认同度是粉丝与明星具有共同的认识和评价程度，粉丝对明星的认同度越高，越是会认为明星的一切都是对的，坚决支持其所有的言行做法，甚至于出现一些失去理智的做法。在此，我们通过微博转发数和主创相关微博评论情感分析来衡量粉

丝的忠诚度。

数据来源及变量设置：对每个明星所有微博的转发数求均值，对主创相关微博做情感分析，求出积极评价率，选取这两个数值作为粉丝忠诚度变量的最终取值。

5）商业价值

商业价值指事物在生产、消费、交易中的经济价值，明星的商业价值则对应的是其作品的经济价值。本文中我们使用明星的个人收入和广告曝光度来评估明星的商业价值。明星的个人收入一定程度上反映了明星的商业价值，广告虽然是企业主为自身产品而做的营销，但产品广告也同样增加了明星的曝光率，属于双赢，而且产品本身的质量也会对明星有所影响，明星代言好的产品会提升其形象，但若是代言的产品出现质量问题，则会对其声誉有负面影响并且降低其商业价值。福布斯中国名人榜则可以看到名人的年收入与其排名，该榜单是对中国名人调研和系统评估得到，故可通过福布斯中国名人榜的年收入和广告曝光度排名（目前数据未使用）来定义明星的商业价值。

数据来源及变量设置：统计 2011-2017 福布斯中国名人榜，获取榜上所有星的收入。若某明星多次榜上有名则取其平均值，若没有上榜则设置为 nan。通过对对应主创人员匹配的到平局收入作为此项的拟合。

6）个人综合潜力

成为明星需要先天条件和后天积累，一个人的综合素质对一个明星是否能够持续走红起到至关重要的作用。我们从技能指数、形象指数、整体评价来衡量一个明星的综合潜力。技能指数是一个明星在其领域的专业技能，比如唱功（声线）、演技、台词功底等等，以及是否多才多艺，多方位全能艺人的发展机会更多，更有发展潜力。形象指数对于明星是至关重要的，并不只是要美，有独特的个人风格会更具有可识别性，更容易被大众记住。然而形象好的明星自然在观众心中都是最美丽的，无论是长相还是身体对他们来说的是觉得不错的。因此，此处可用观众对演员的颜值和身材评价来映射出明星的形象指数。

数据来源及变量设置：从电影微博内容中获取演员演技、颜值、身材、声线、整体评价的评分。

综上所述，最后确定的主创画像 aspect 维度如下表：

表 4.1

电影					导演			演员						
音效	特效	电影深度	宣传	整体评价	电影节奏/剪辑/分镜/情节组织	角色塑造	整体评价	演技	颜值	身材	声线	气质	台词功底	整体评价

4.2 主创画像构建流程

如下图 4.1 所示即为主创画像部分的流程图，将从微博获取的数据分为训练数据集和预测数据集两部分，首先需要对训练数据集打标签,其次将训练数据导入 aspect sentiment 模型进行训练，未打标签的预测数据作为模型的应用数据集，利用主创姓名映射表匹配将两个数据集的结果 最终得到主创的用户画像结果。

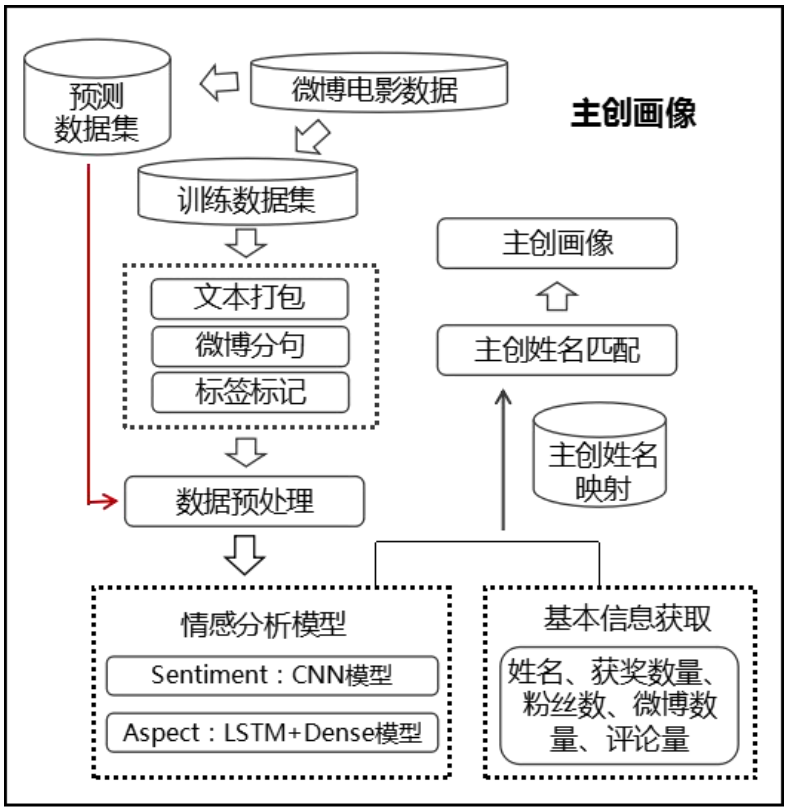


图 4.1 主创画像部分流程图

4.3 训练数据集

构建过程首先需要收集电影话题下的微博，通过人工对这些数据进行判断，根据 aspect 的分类提取每条的标签特征并对其进行评分。

评分分为 5 个等级：5（非常好）、4（好）、3（一般）、2（差）、1（非常差）。若影评对于同一个特征有不同的评价，则用分号将分值分开。比如下表 4-1，电影《第 101 次求婚》的影评中，可从中提取到黄渤的演技非常好，然而林志玲的演技则较差，所以对于演技此特征有两个评分，我们按顺序用分号分开。其中还提到了背景音乐（即电影音效），演员身材以及电影结局，其评价都较好。

表 4.2 电影标签举例

电影名称	序号	微博内容	特征	评分
第 101 次求婚	1	黄渤演的超级好的，只可惜林志玲全程尬演，结局也算满意了，结婚现场的背景音乐都听哭了，赵又廷超帅的，身材超好，给男神打 call。	音效	4
			演技	5;2
			颜值	5
			身材	4
			整体评价	4
	……			
建军大业	1	#欢乐暑期档#之#电影建军大业# 刘伟强的电影真不错 情节跌宕起伏 各位老戏骨都好带感 各个人物诠释的都不错	整体评价（电影）	4
			电影节奏/剪辑/分镜/情节组织	4
			角色塑造	4
			整体评价（导演）	5
			整体评价（演员）	4
			剧情吸引程度/剧本	4
	……			

总共打了 6000 多条带标签的微博语句作为训练集

4.4 Joint Aspect Sentiment Model

由于 sentiment 是跟 aspect 依依相关的，在建立分类模型的时候，应该将他们统一看待，因此一个可行的思路就是将 aspect 分类和 sentiment 分类统一起来，训练一个分类器。

4.4.1 训练标签准备

根据前文所描述的主创画像维度可是，我们将 aspect 分类成 16 个维度，每个维度具有 5 个情感标签，因此需要将人工的标记的标签集转换成如下图表所示：

每个标签的是由主创画像维度及其对应的情感分数所构成的。

4.4.2 模型架构

由于文本的情感和维度分类大多是由于局部的关键词所决定的，并且 CNN 模型在文本分类的问题上取得了不错的成绩，而且由于其 filter 和 pooling 具备能够捕捉文本局部特征的特点，因此将其作为神经网络模型的第一层。接下来连接标准化层和全连接层，最后在连接输出层。

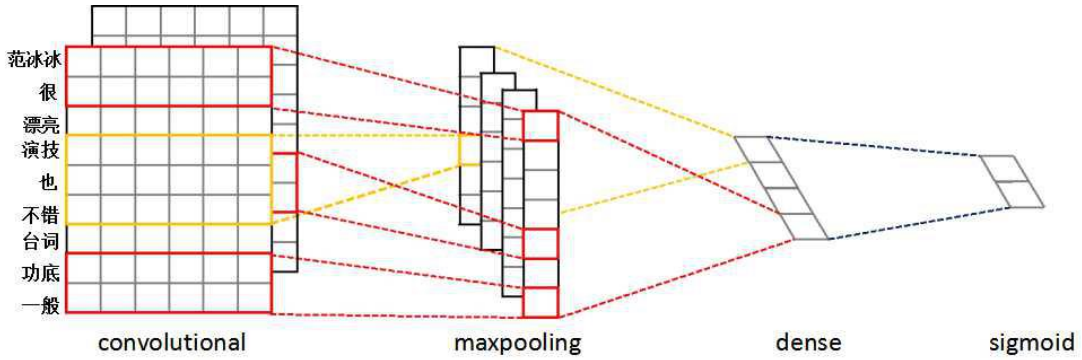


图 4.2 模型框架图

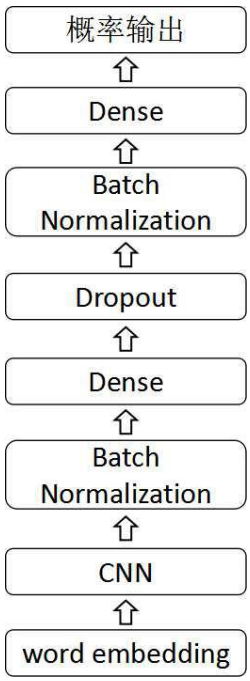


图 4.3 神经网络模型流程图

其中 BatchNormalization 层的作用是归一化输入数据，可以加速网络收敛，Dropout 层的作用主要是随机断开网络的连接，防止过拟合。

4.5 实验结果分析

4.5.1 词向量

我们从互联网上下载了一个基于 word2vec 训练的词向量版本，它是由新闻预料和小说预料训练而成的，经测试，其训练效果不错，例如输入微博之后，与其相关性最高的词为:

[('朋友圈', 0.886117696762085), ('微信', 0.8786516189575195), ('论坛', 0.8700537085533142), ('私信', 0.8430871367454529), ('推特', 0.84183269739151), ('围脖', 0.8334218263626099), ('新闻', 0.8333889842033386), ('官方论坛', 0.8325998187065125), ('网站', 0.8300749063491821), ('学校论坛', 0.8264977931976318)]。

4.5.2 数据预处理

1）分本分句

为了将整段的文本数据切分为以句为单位，我们需要对文本做分句处理，该项目利用哈尔滨工业大学的自然语言处理平台 LTP-Cloud，自动分句功能，LTP 的分句是根据中文标点里的句号、问号、感叹号、分号、省略号将文本段落切分为句子。

2）文本分词

将分句之后的微博语句通过 ltp 分词工具进行分词，LTP-Cloud，其后端依托于语言技术平台，语言云为用户提供了包括分词、词性标注、依存句法分析、命名实体识别、语义角色标注在内的丰富高效的自然语言处理服务。

3）去停用词

为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些无用信息词语，如语气词、标点符号、助词等等，该项目利用哈尔滨工业大学停用词表，顺序扫描分词结果，若词语在停用词表中，则去掉该词，若不在，则保留。

4）表情符号、话题符号去除

微博文本当中经常出现【害羞】、【高兴】以及#...#的格式的话题，在训练模型和预测的时候需先用正则表达式去掉。

4.5.3 模型训练

我们采用 keras 深度学习框架对模型进行训练，CNN 窗口大小为 4，设置 batch_size 的大小为 400，损失函数选择 binary_crossentropy,优化器才有 Adam，具有自适应调节学习速率，收敛快等优点。由于训练数据的大小不算多，因此模型采用 K 折交叉训练的方法对模型进行训练，这样可以在一定程度上避免过拟合的情况，最后预测的结果对 K 折模型取平均所得。模型的训练情况如图所示:

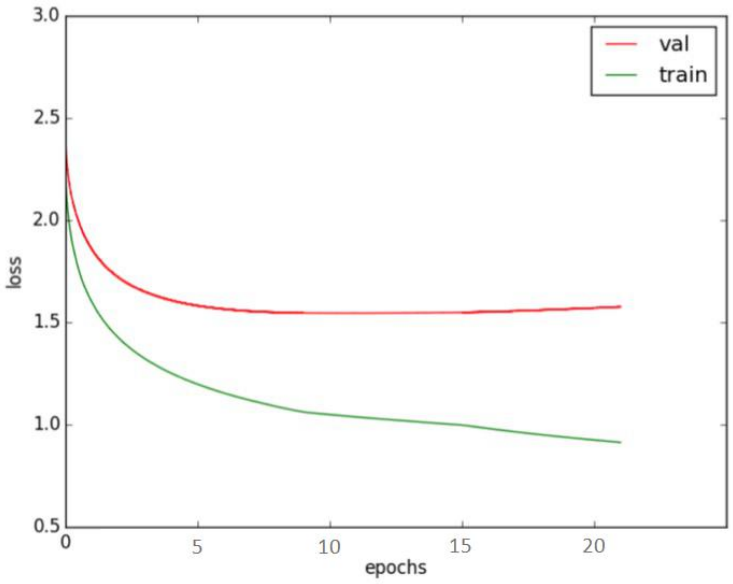


图 4.4 模型训练过程图

通过交叉验证，模型的 F 值为 0.345 。鉴于是纯数据驱动的模型，没有添加人工特征，模型的准确率没有很高，有改进的工具，比如加入人工特征或与常规模型进行集成。

4.5.4 主创姓名匹配

模型训练完成之后将其对剩余的微博电影微博做预测分析，得到预测电影微博的 aspect 和情感分数。

为了将模型预测出的每个标签维度的情感分数与主创个人相匹配，尽可能准

确构建主创的个人画像而不是整体画像，我们通过统计主创的称呼别称设计，主创映射表，建立规则匹配微博电影话题数据，规则映射表举例如下：

表 4.3 规则匹配示例表

姓名	规则匹配
范冰冰	范冰冰；冰冰；冰冰姐；范爷

映射表的构建我们采用叠字匹配和人工收集的方式相结合，模型训练完成之后可以得到如下所示的主创画像表：

表 4.4 主创画像示例表

字段	值	含义
name	付辛博	姓名
award_count	10	获奖数量
influence	66825356.0	所演出的所有平均票房
figure	3.65	身材分数
actor_overall	4.2	演员整体评价
voice	3	演员声线
line_ability	3	台词功底
num_fan	6764972.0	粉丝数量
num_follow	303.0	关注人数

第五章　电影票房预测模型

为了评估主创团队对电影票房的贡献，我们将电影票房作为贡献度的拟合值。通过建立主创画像寻找主创与影响电影票房的主要因素的数学关系，从而从多个维度分析主创对电影票房的贡献度。

5.1　电影票房收入影响因素分析

对电影票房进行预测，首先需要对影响电影票房的因素进行研究。国外研究者对电影票房因素的划分虽然很有启发性，但是中国的电影市场和电影观众的特性毕竟与北美不同，而国内目前对影响因素的研究大都过于细化或分散，不成系统。在总结国内外研究的基础上，根据电影生产周期的制片、宣传、发行三个阶段，提出以下八个影响电影票房的因素。

1）排片率

排片率是指同档期内一部电影的播映比率，是某部影片影院播放场次的比例。

假设某部电影的排片率为 10%，这意味着影院放映 100 场电影就有 10 部是该影片。排片率对电影作品的收入十分重要，排片率过低，则会限制观众影院消费的欲望。

数据来源及变量设置：娱票儿票房分析网页中获取到电影每天在全国影院的排片率，求平均值，作为该电影的排片率变量。

2）　网络口碑

随着互联网的快速发展，各类电影社区也纷纷发展和壮大起来，网上电影爱好者大军不断膨胀，人们第一时间便可以从网络上获得电影口碑，通过电影评分判断电影的品质，通过电影关注人数判断电影的热度，以此来决定是否到影院观看，电影网络口碑对于中国电影行业已经产生了重要影响。

网络口碑是指网络用户产生的内容，如用户对电影的评分、评论一部电影的文字等等。对网络口碑的量化研究，研究者初步形成了一些概念上的共识，从方便获取数据的考虑出发，一般搜集网站上现成的数据，主要从“值”和“量”两个方面考察。“值”意味着用户对影片的态度和口碑的说服力。像国外的、烂番琉等网

站都提供“打分”的功能，将所有用户的评分合成为一个分数，来代表观众对电影的态度。“量“意味着有多少人在关注、讨论该影片。

① 电影评分

电影评分反映了公众对一部电影的评价。一般认为电影评分的作用有两个：一是在电影上映初期影响消费者观影的决定，二是预测观众是否喜欢一部电影。现在的好多观众已经习惯在观影之前上专业的电影网站查看评分，观影后也会在上面分享经验和看法。可见，影评在一定程度上会影响其他观影者的观影意愿，进而对电影票房产生影响。

数据来源及变量设置：对余票儿、豆瓣、猫眼、淘票票、糯米、时光网、微博七个平台的电影评分求均值，作为电影评分变量的最终取值。

② 关注人数（想看人数）

以往的网络口碑研究认为，网站上“看过”、“想看”、“短评”的数量代表着观众对电影的知晓度和关注度，它们像广告一样，起着告知和传播的作用。如果一部电影能够得到观众的关注和肯定，就算没有大规模的市场营销，观众也会自发的为电影做宣传，从而对电影票房产生积极的影响。

数据来源及变量设置：对余票儿、豆瓣、猫眼、淘票票、糯米、时光网六个平台的电影想看人数求均值，作为电影关注人数变量的最终取值。

3） 实时票房占比

实时票房占比是指当天一部电影的票房占所有电影总票房的比率。假设某部电影当天的票房占比是 30%，当天电影总票房若有 1000 万则 300 万是由该电影贡献的。这意味的该电影在同档期电影的竞争力，占比越大，则竞争力越大，同时也表明观众的热爱度，实时票房占比也会影响到该电影的拍片率，影院为了获得更大的收入，可能会按照当天的票房占比来决定次日拍片率，我们也把实时票房占比作为影响票房的一个重要因素考虑。

数据来源及变量设置：求每日实时票房占比的平均值作为变量值，该数据可通过娱票儿票房分析网页直接获得。

4）档期

档期是电影产业化发展到一定程度的必然产物。电影《甲方乙方》拉开了我国电影档期的序幕。十余年来，贺岁档、暑期档、国庆档、情人节成为吸金效应

最强、竞争最激烈的档期。2017 年贺岁档为例，21 部影片共斩获约巧亿票房，其中《时间瑜伽》《乘风破浪》《西游降度篇》吸金 43 亿，这足以看出优质档期的票房引发的吸引力和竞争力。

数据来源及变量设置：本文拟打算将档期因素分为是否为贺岁档、暑期档等热门档期，设定为虚拟变量，若是黄金档期则为 1，否则为 0。具体档期的归属将根据电影的上映日期来判断，如下表所示，而电影的上映日期从豆瓣以及娱票儿票房分析网上获得。

表 5.1 档期变量

黄金档期	时间段（业内较为认可的标准）
贺岁档	每年的 11 月 20 至 2 月底
暑假档	每年的 6 月 1 日至 8 月 31 日
其他热门档期	仅劳动节档（5 月 1 日至 5 月 3 日）
	国庆节档（10 月 1 日至 10 月 7 日）

5）电影类型

电影类型反映了观众的观影偏好，往往决定影片的题材、受众和投资。综合

学者们的研究可以发现，影片类型对票房的影响较大且复杂，而消费者的观影决策受电影类型的影响也较大。研究不同电影类型对电影票房的影响可以为制作方的题材选择提供相应的依据，具有一定的市场价值。

数据来源及变量设置：基于我国的电影市场现状和行业通用标准，我们将影片类型分为类：剧情片、动作片、爱情片、喜剧片、奇幻片、冒险片，犯罪片，悬疑片和惊悚片，其他电影类型的变量不予统计，设定为虚拟变量，将对应类型的赋值为 1，不是则为 0，由于一部电影可能分属于多种不同类型，我们对一部电影的类型最多取前两种主要类型，有关电影类型的相关数据我们从豆瓣网和娱票儿票房分析网上获得。

6）故事熟悉程度

不管是在西方还是国内，众多制片方之所以热衷拍摄续集系列电影，原因在于续集是经典影片的延续，是电影票房的坚实保障，同时拍摄续集比制作全新的电影更加安全，不仅节约了营销、宣传成本，还在商业上为资金回笼增加了保险

系数。但一部系列电影是否能持续火热终究是由观众决定的，电影续集大行其道，续集电影同样存在颠覆整个系列的风险。

数据来源及变量设置：我们把续集系列作为故事熟悉程度的考察量，设定为虚拟变量，当电影是续集系列电影（不包含第一部）时，续集系列变量为 1，否则为 0，电影续集系列信息在百度百科中获得。

7）发行公司

发行因素在巴瑞李特曼研究的电影票房影响因素模型中是一个很重要的变量，然而国内学者对电影票房影响因素的定量分析中少有考虑进来。随着中国电影市场的日渐规范，一部电影的发行公司的营销能力占据着越来越重要的地位，有时凭借着强大的营销能力便能够轻松吸引观众走进电影院。因此，我们将发行公司的营销能力作为可能影响电影票房的因素之一研究。

数据来源及变量设置：鉴于发行公司的营销能力难以衡量，发行公司内部数据无法获得，我们拟用发行公司的市场份额大小的划分来衡量发行公司的营销能力。我国电影发行市场较为特殊，两大国营企业中影集团和华夏电影制片厂垄断较为严重，其市场份额占据了半壁江山。而行业内公认的五大民营企业翘楚光线影业、博纳影业、万达影视、乐视影业和华谊兄弟的市场份额近年来稳步递增，发行影片的数量和票房也比较多，其发行营销能力得到了认证。因此，我们拟定若一部电影的发行公司是以上七大发行公司之一，则发行公司的营销能力认定为比较强，设定为虚拟变量，赋值为 1 否则认为其发行公司的营销能力较弱，赋值为 0，电影发行公司数据我们从豆瓣网和娱票儿票房分析网上获得。

8）语言（国家）

不同观众对于不同国家所产电影的偏好不同，美国大片的影迷很多，对于国产当红小生青春片的追捧也大有人在，日本动漫系列很受欢迎等等，可见，电影的国家也是影响票房的一个因素。

数据来源及变量设置：我们把语言变量设定为虚拟变量，我们将影片语言分为三类：汉语、英语和其他语言，设定为虚拟变量，将对应类型的赋值为 1，不是则为 0，相关数据可来源于豆瓣网、中国票房网等。

5.2 自变量选取

通过 5.1 中对影响电影票房因素的分析，我们选取如下表所示变量作为模型的初始自变量:

表 5.2 自变量介绍

变量符号	变量含义	变量符号	变量含义
X1	排片率	X11	喜剧片
X2	口碑—想看人数	X12	奇幻片
X3	口碑—评分	X13	冒险片
X4	实时票房占比	X14	犯罪片
X5	汉语	X15	悬疑片
X6	英语	X16	惊悚片
X7	其他语言	X17	发行公司
X8	剧情片	X18	档期
X9	动作片	X19	故事熟悉程度
X10	爱情片		

5.3 相关性检验

5.3.1 连续型自变量相关性检验——Pearson 相关系数

Pearson 相关系数用于双变量正态分布的数据，是最常见的用来描述变量线性相关的统计量，两个连续变量间呈线性相关时，使用 Pearson 积差相关系数，连续性变量和票房的相关性通过皮尔逊相关系数来衡量，皮尔逊相关系数为：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

皮尔逊相关系数的取值介于-1 和 1 之间，当皮尔逊系数的取值接近于-1 或 1 时，表示两个变量之间有强相关性；当取值接近于 0 时，表示两个变量不是线性相关的；当取值大于 0 时，表示正相关；当取值小于 0 时，表示负相关。

表 5.3 Pearson 相关系数

Pearson 相关系数，N=2321 Prob > |r| under H0: Rho=0				
	schedule_rate	box_office_rate	want_count	score
box_office	0.77249	0.79477	0.73294	0.25054
box_office	<.0001	<.0001	<.0001	<.0001

由表 5.3 可以看出变量排片率（schedule_rate）X1，实时票房占比（box_office_rate）X4，想看人数（want_count）X2 均与因变量电影票房有显著性相关关系，且皮尔逊系数均大于 0.7，具有较强正相关性，score 变量有显著性的较弱正相关性 （p<0.0001,皮尔逊系数大于 0.25），故我们考虑暂时保留这 4 个连续变量。

5.3.2 离散型自变量相关性检验——Spearman 相关系数

Spearman 相关系数又称秩相关系数，是利用两变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，当两变量不符合双变量正态分布的假设时，需用 Spearman 秩相关来描述变量间的相互变化关系。

表 5.4 Spearman 相关系数表

Spearman 相关系数								
	is_hot_schedule	is_series	english	chuinese	other_language	plot	action	love
box_office	0.03851	0.28670	0.39536	-0.12254	0.03352	-0.18974	0.34577	-0.08536
box_office	0.0636	<.0001	<.0001	<.0001	0.1065	<.0001	<.0001	<.0001
	2321	2364	2321	2321	2321	2321	2321	2321
	comedy	fantasy	adventure	crime	suspense	thriller	company	comedy
box_office	0.05892	0.19881	0.27838	0.13899	0.02348	0.08328	0.16482	0.05892
box_office	0.0045	<.0001	<.0001	<.0001	0.2583	<.0001	<.0001	0.0045
Prob > |r| under H0: Rho=0								

由表 5.4 可以看出在这 15 个离散变量中，spearman 相关系数值都很小，但是 p 值大多满足<0.05，故我们推测这些离散变量可能与因变量相关。剔除明显未通过显著性检验的变量 other_language 和 suspense，其余变量暂时都保留。

如下表所示即为通过相关性检验的全部变量

表 5.5 通过相关性检验的变量

变量	定义
action	动作片
adventure	冒险片
box_office_rate	实时票房占比
chinese	汉语
comedy	喜剧片
company	发行公司
crime	犯罪片
english	英语
fantasy	奇幻片
is_hot_schedule	故事熟悉程度
is_series	是否续集
love	爱情片
plot	剧情片
schedule_rate	排片率
score	评分
thriller	惊悚片
want_count	想看人数

5.4 回归分析及其结果

选取通过相关性检验的变量如上表所示共 18 个，我们采用了多元线性回归、逐步回归和套索回归三种不同的方法对现有的变量做回归分析，并对变量做共线性诊断，最后对三种不同模型的结果作比较分析。

5.4.1 模型原理介绍

1）逐步回归模型：基本思想是将变量逐个引入模型，每引入一个解释变量后都要进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解

释变量由于后面解释变量的引入变得不再显著时，则将其删除。以确保每次引入新的变量之前回归方程中只包含显著性变量。这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止。以保证最后所得到的解释变量集是最优的。

2）套索回归模型：通过给回归估计上增加一个偏差度，惩罚回归系数的绝对值大小，来降低标准误差，如下为 Lasso Regression 公式：

$$y = \arg\min_{\beta \in R^P} \underbrace{\| y - X\beta \|_2^2}_{Loss} + \lambda \underbrace{\| \beta \|_1}_{Penalty}$$

它使用的惩罚函数是绝对值，而不是平方。这导致惩罚（或等于约束估计的绝对值之和）值使一些参数估计结果等于零。使用惩罚值越大，进一步估计会使得缩小值趋近于零。它能够减少变化程度并提高线性回归模型的精度，并能够有效的解决模型过拟合问题。

5.4.2 不同回归模型预测结果比较

1）探究共线性诊断对模型结果影响，以多元线性回归模型为例，如下表 5.6 所示，可以看出剔除带来共线性影响的变量后，评价拟合优度的重要指标 R 方反而较之前小幅减小，但是共线性诊断过后加强了模型的稳定性，大大提高了模型的泛化能力，减少了更换样本对模型结果造成的影响，故我们选择了牺牲一点点拟合度来提高模型的稳定性，以下 2）中三个方法均是在诊断并修正共线性问题之后的结果。

表 5.6 共线性诊断前后 多元线性模型结果比较

模型类型	选取变量个数	拟合度偏 R 方	模型稳定型
共线性诊断前	17	0.8057	很弱
共线性诊断后	16	0.7667	较强

2）如下表 5.7 为三种不同类型回归的结果比较，可以看出，三个模型的拟合度偏 R 方几乎无区别，模型稳定性也相同，但是自变量的个数却相差很大，而使用最少的预测变量数来最大化预测能力，可以减少冗余变量，达到精简模型的效果，同时可以提高线性回归模型的精度，并能够有效的解决模型过拟合问题，提高模型性能，故我们选择套索回归来进行票房预测。

表 5.7 三种回归模型结果比较

模型类型	拟合度偏 R 方	自变量个数	模型性能	模型稳定型
多元线性回归模型	0.7667	16	较差	较强
逐步回归模型	0.7671	10	较强	较强
套索回归模型	0.7654	5	强	较强

5.5 主创贡献度建模

通过上文的相关性分析可以发现，排片率与评分与票房的相关性较高。因此打算通过建立主创画像，寻找主创与排片率和评分之间的关系，通过画像数据对电影排片率及评分数量进行预测，从而间接影响电影票房。

基于上述陈述，我们通过以下两点体现主创贡献度分析：

1) 基于主创画像对排片率和评分进行回归预测，在通过排片率和评分对票房进行预测，从而间接影响票房。
2) 对于已上映电影可以通过更换主创信息预测出不同的票房，从而体现主创的价值。

想看数量的相关系数虽然也较高，但是经过实验发现想看数量这一特征与主创画像数据的相关性很差，因此舍去这一特征，之后预测票房所用到想看数量都取平均值代替。

5.5.1 排片率预测模型

1、训练数据

我们模型需要一个导演和四个主要演员的主创画像数据，包括主创画像模型预测的数据以及从百度百科上爬取的数据，并与 1846 部电影的排片率进行拼接，从而得到了 1846 条训练数据。

2、 输入特征

输入特征如表 5.8 所示

表 5.8 主创画像特征表

	变量名	定义
director	director_average_reports_count	导演微博平均转发数
	director_influence	导演的影响力
	director_overall	导演的整体评价
	director_average_followers_count	导演微博粉丝数
	director_num_fan	导演微博粉丝数
	director_num_follow	导演微博关注数
	director_average_attitudes_count	导演微博平均点赞数
	director_average_comments_count	导演微博平均评论数
	director_award_count	导演获奖数
actor1	actor1_average_attitudes_count	演员微博平均点赞数
	actor1_average_comments_count	演员微博平均评论数
	actor1_average_reports_count	演员微博平均转发数
	actor1_acting	演员演技
	actor1_appearance	演员颜值
	actor1_figure	演员演员身材
	actor1_line_ability	演员的台词功底
	actor1_overall	演员的整体评价
	actor1_voice	演员声线
	actor1_influence	演员出演电影的平均票房
	actor1_num_fans	演员微博粉丝数
	actor1_num_follow	演员微博关注数
	actor1_award_count	演员获奖数
actor2	...	
actor3	...	
actor4	...	

actor2 actor3 actor4 对应的特征属性根跟 actor1 相同，总共 61 个特征。

3、模型选择

我们选择梯度提升树（XGBoost）模型来拟合主创画像数据与排片率的关系，选择该模型的原因是提升树模型有很大的拟合能力，对噪声数据有比较强的鲁棒性。

4、特征筛选

我们通过基于 filtering 的方法，利用 XGBoost 本身的 feature_importance 值作为指标进行筛选。在学习率为 0.05,迭代次数为 2000，max_depth 为 4 的参数环境下运行模型，得到如下所示的特征重要性图：

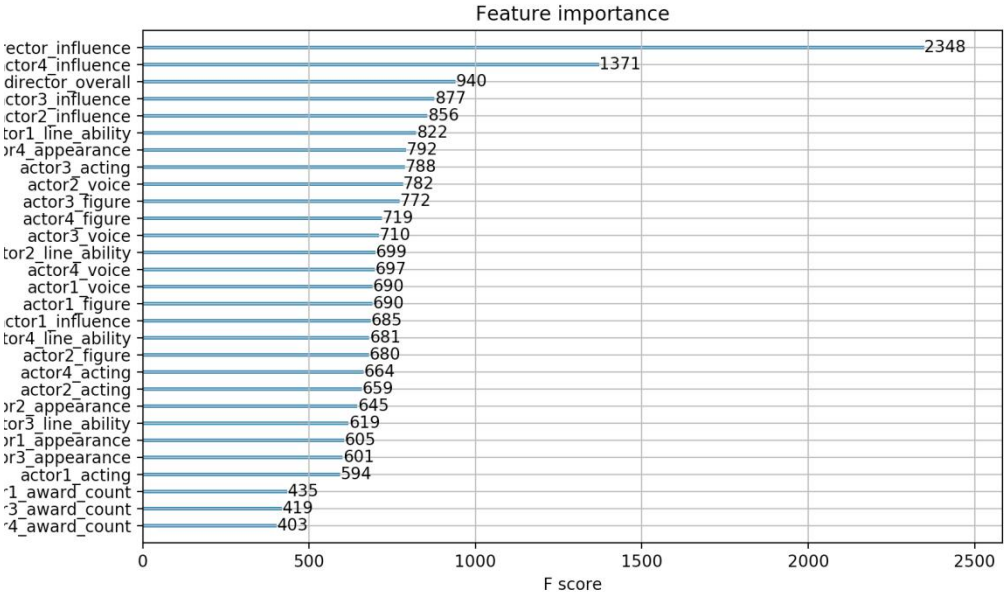


图 5.1 排片率特征重要性图

经过筛选后特征表 5.9 如下所示：

表 5.9 排片率模型特征筛选子表

director_influence	导演导的电影的平均票房
director_overall	导演的整体评价
actor1_line_ability	演员的台词功底
actor1_figure	演员身材
actor1_voice	演员声线
actor1_influence	演员出演电影的平均票房
actor1_acting	演员演技
actor1_appearance	演员颜值
actor1_award_count	演员获奖数
actor2_acting	演员演技
actor2_influence	演员出演电影的平均票房
actor2_voice	演员声线
actor2_line_ability	演员的台词功底
actor2_figure	演员身材
actor2_appearance	演员颜值
actor3_appearance	演员颜值
actor3_figure	演员身材
actor3_voice	演员声线
actor3_influence	演员出演电影的平均票房
actor3_acting	演员演技
actor3_line_ability	演员的台词功底
actor3_award_count	演员获奖数
actor4_award_count	演员获奖数
actor4_appearance	演员颜值

actor4_influence	演员出演电影的平均票房
actor4_line_ability	演员的台词功底
actor4_figure	演员身材
actor4_acting	演员演技
actor4_voice	演员声线

5、模型训练

将训练数据以 8：2 的比例随机分成训练集和验证集模型超参按照上节所示，设置验证集防止过拟合，评测标准采用 RMSE 均方根误差，训练过程如下图 5.2 所示:

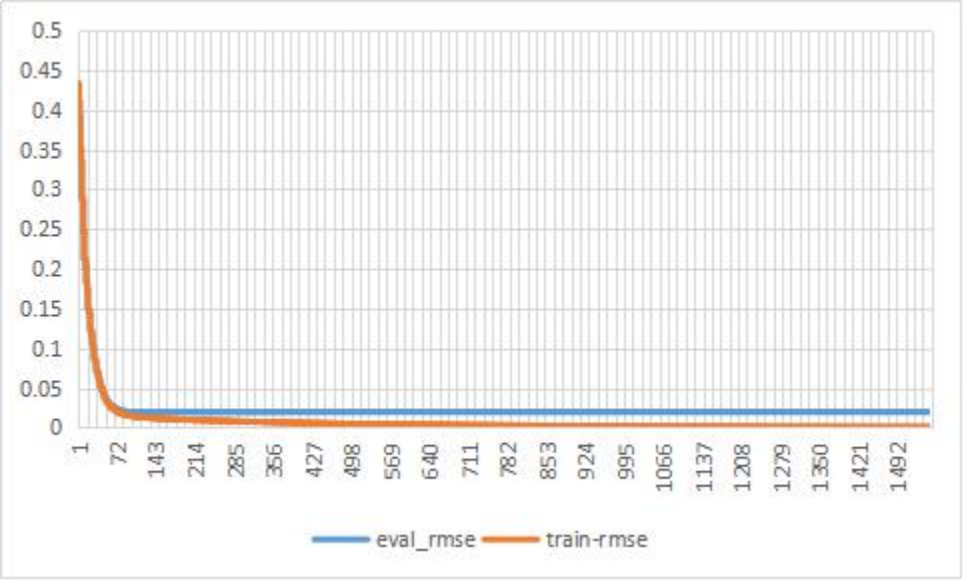


图 5.2 排片率模型训练过程

6、模型结果

XBGBoost 在验证集上的均方根误差只有 0.01979 ,可见拟合效果令人满意。经过交叉验证集验证，最佳迭代次数为 1759。

5.5.2 评分预测模型

1、训练数据

评分预测模型的整体思路与排片率预测模型相似，我们提取出 1846 部电影的评分与一个导演和四个演员的的主创画像特征共计 1846 条训练数据。

2、模型选择

同排片率预测模型一样，评分预测模型也选择梯度提升树模型（XGBOOST），学习率设置为 0.08，评测标准为 RMSE，max_depth 为 4，subsample 为 0.8，其

中设置较小的树深度和 subsample。

3、特征选择

总体输入特征同上节特征输入部分，利用 XGBoost 提供的 feature importance 指标做特征的 filtering，每个特征的 feature importance 情况如图 5.3 所示:

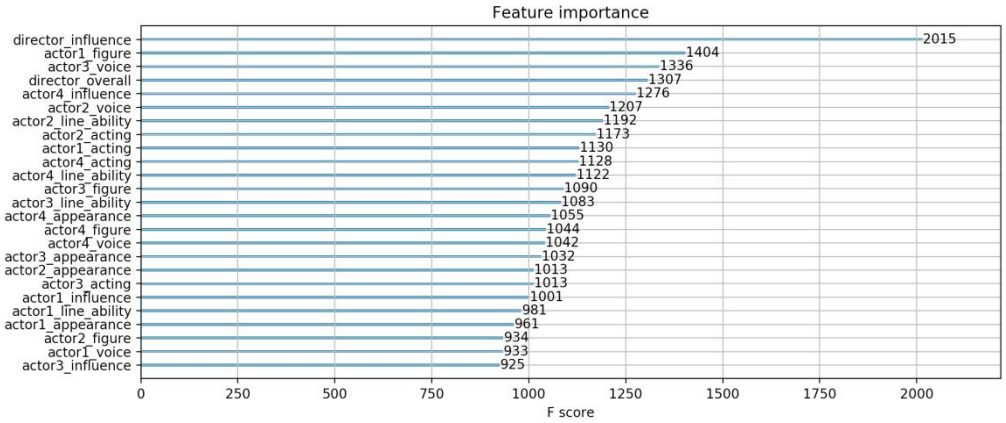


图 5.3 评分模型特征重要性图

筛选后的特征如下表 5.10：

表 5.10 评分模型特征筛选表

director_influence	导演导的电影的平均票房
director_influence	导演的整体评价
director_influence	演员声线
director_influence	演员身材
director_influence	演员出演电影的平均票房
director_influence	演员演技
director_influence	演员的台词功底
director_influence	演员颜值
director_influence	演员颜值
director_influence	演员身材
director_influence	演员声线
director_influence	演员的台词功底
director_influence	演员演技
director_influence	演员出演电影的平均票房
director_influence	演员演技
director_influence	演员声线
director_influence	演员身材
director_influence	演员颜值
director_influence	演员的台词功底
director_influence	演员出演电影的平均票房
director_influence	演员演技
director_influence	演员颜值

director_influence	演员身材
director_influence	演员的台词功底
director_influence	演员声线

4、模型训练

将训练数据以 8：2 的比例随机分成训练集和验证集模型超参按照上节所示，设置验证集防止过拟合，评测标准采用 RMSE 均方根误差，训练过程如下所示，

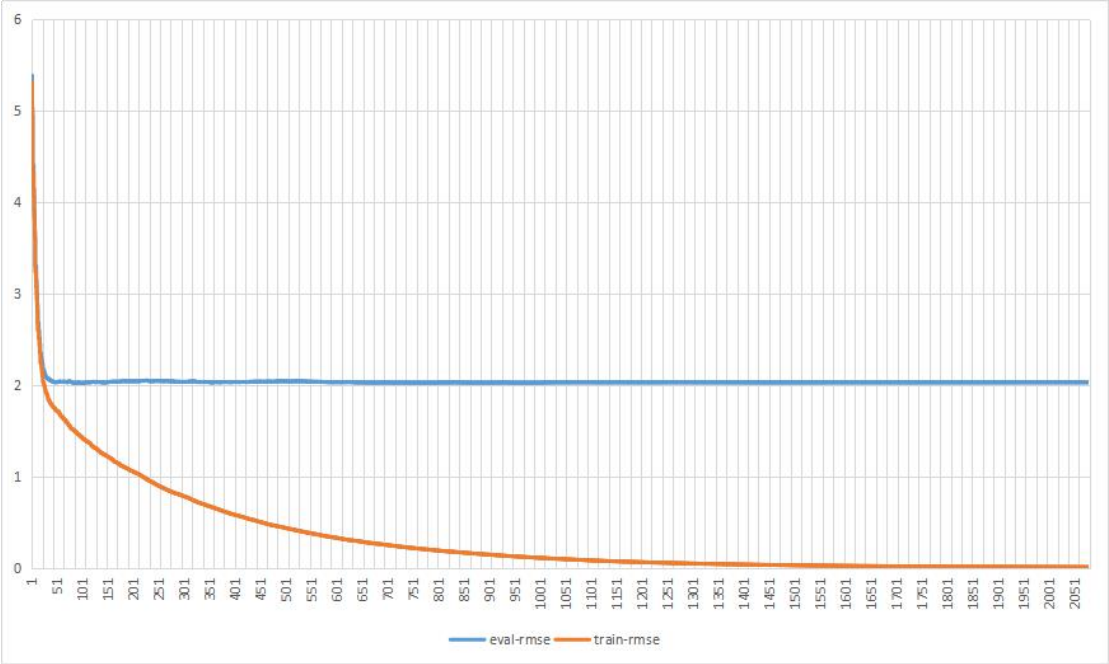


图 5. 4 评分模型训练过程图

5、模型结果

从训练过程可以发现，当迭代次数为 2073 趟时，验证集损失不在降低，模型在验证集上的损失误差为 2.03603。

第六章　主创推荐系统及其可视化

6.1 主创推荐模块

推荐模块是上述所有模型的应用，主要功能是根据主创画像推荐给电影发行商主要导演和四个主要演员，在给定的参数输入下，根据主创画像数据从所有主创中找出模型所预测的电影票房最高的一组主创。

1、模型输入

表 6.1 主创推荐输入特征表

字段名称	解释	类型
movie_budget	电影预算	int
movie_type	电影	list 例如[‘剧情’,’动作’…]
is_hot_schedule	是否是黄金期	binary int
is_serie	是否为续作	binary int
production_company	发行公司	list
iter	迭代次数	int

2、推荐模型算法

由于推荐主创组合可能性太高，找出全局预测票房最高的主创组合复杂度太高，因此我们采用一种近似的随机搜索算法，算法说明如下：

在一定的迭代次数下，随机从数据中找出一个导演和四个演员的画像数据，将数据输入票房预测模型，若所得出的票房大于原先预测的票房，则更新主创团队信息。完成迭代之后，输出预测票房最高的主创团队。

6.2 可视化模块

1、框架介绍

Flask 是一个使用 Python 编写的轻量级 Web 应用框架。Flask 基于 Werkzeug WSGI 工具箱和 Jinja2 模板引擎,使用 BSD 授权。Flask 没有默认使用的数据库、窗体验证工具，保留了扩增的弹性，可以用 Flask-extension 加入这些功能 ORM、窗体验证工具、文件上传、各种开放式身份验证技术。

2、交互说明

1)数据交互:

在处理和验证表单输入，本系统采用的 WTForms 框架。使用 WTForms, 制定表单域的 HTML 代码, 进一步将代码和显示独立开来，形成松耦合。当包含域的表单进行验证时，WTForms 中的验证器(Validators)为域(field)验证, 通过验证的数据调用(calling)域, 提供对应关键词参数, 它们会在输出中作为 HTML 属性注入，完成模板的渲染。

2)数据库交互:

Python 中使用 Pymongo 来操作 MongoDB 数据库，MongoEngine 把数据库操作部分作为 model 抽离出来。MongoEngine 是一个对象文档映射器（ODM），相当于一个基于 SQL 的对象关系映射器（ORM）。

3、功能介绍

针对本系统主要为制片方，导演或者编剧提供决策建议，因此本系统主要实现以下三个功能：主创画像，票房预测 主创推荐。

以下图 6.1 是系统的入口。

图 6.1 系统入口界面

点击进入主创画像，在主创画像的网页中，用户在搜索框输入演员或者导演的名字，则会出现主创的画像信息和过往主创导演或者主演的电影的电影票房记录。其中，演员的画像信息包括（整体评价、身材、声线、台词功底、演技、颜值、气质、粉丝数），导演的画像信息包括（电影节奏/剪辑/分镜/情节组织、角色塑造、粉丝数、整体评价）。以下图 6.2 是主创画像页面。

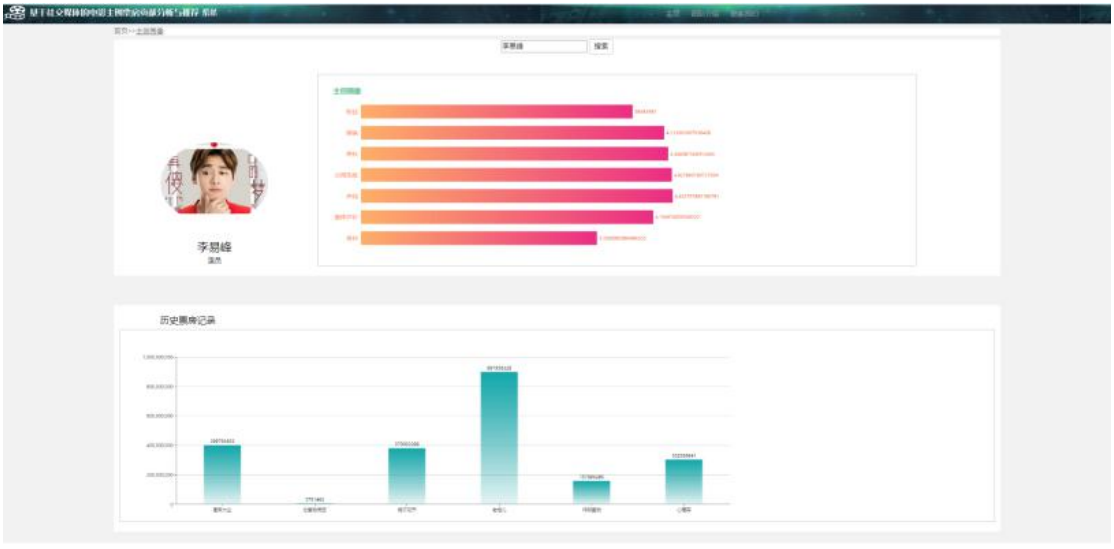


图 6.2 主创画像界面

进入票房预测，在该网页中，用户输入以下信息：电影预算、电影类型、是否黄金档、是否续集、发行公司、主创名单（1 名导演和 4 名演员），则实现对应条件下的票房预测信息。以下图 6.3 是电影票房的预测界面。

图 6.3 电影票房的预测界面

进入主创推荐：在该网页中，用户输入以下信息：电影预算、电影类型、是否黄金档、是否续集、发行公司，则实现对应条件，主创名单（1 名导演和 4 名演员）的推荐。以下图 6.4 是主创推荐的界面：

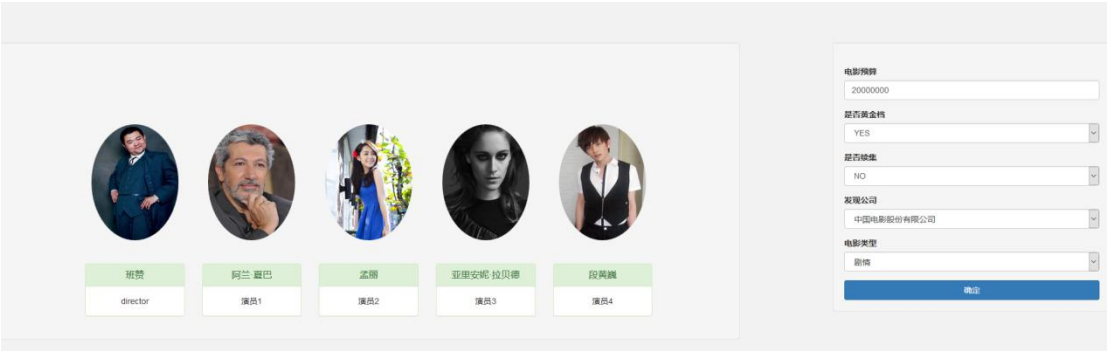


图 6.4 主创的推荐界面

第七章　两地电影属性关注度和情感分析比较

为了比较两岸观影用户对电影属性的关注度差异和好评率的不同，我们基于台湾的 PTT 影评数据和大陆的微博话题数据做了关注分析和情感分析的比较。

我们选取了电影的基本元素如音乐、剧本、特效及与电影相关的人物如导演、演员等作为我们研究的属性对象，其中关注分析研究的是两岸用户对于这些属性的声量差异，情感分析研究的是两岸用户对于这些属性的正面情感的比率。

7.1　属性情感分析步骤简介

情感分析主要分为如下图 7.1 中的 6 个步骤：

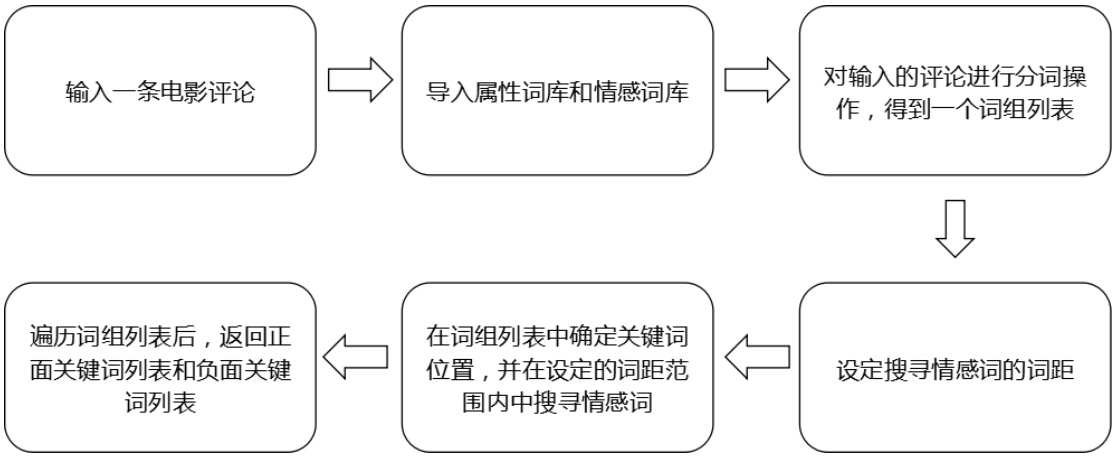


图 7.1 影评数据情感分析流程

1．从数据库提取一篇文章及文章标题与作者，先将标题与所有 2014-2017 的电影名称进行比对，如果标题有提到任何电影的话就保留；

2．如果标题没有提及电影的话，接着比对文章，看文章是否有提及电影，如果有的话就保留，没有的话就跳下一篇文章；

3．将保留的文章进行与六大面向的属性词库进行比对，如果没有提及任何属性的话就跳下一篇。如果有提及属性的话，与情感词库进行比对，在属性词前后距离 3 的范围内如果有任何情感词的话就将数据记录进数据库当中。

4．结束比对跳回步骤 1。

7.2 电影属性关注度分析

通过匹配电影属性词库，对每个属性词出现的频率计数，以此来拟合属性的关注度，部分属性词库如下表 7.1 所示：

表 7.1 电影属性词库部分

音乐	音效
音效	
配乐	
爵士	
音乐风格	
…	
表现	特效
演技	
战斗	
…	
…	

电影关注度分析比较，如下图 7.2 所示：

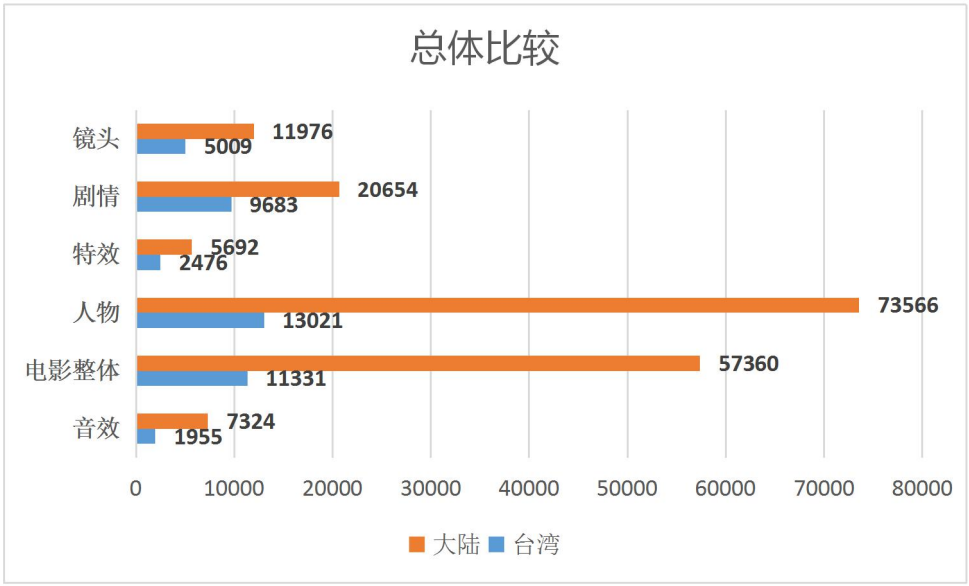


图 7.2 两地电影属性关注度

1) 大陆观众和台湾观众都比较关注电影中的人物和电影的整体；

2) 大陆和台湾关注的方面的顺序也大致相同，台湾数量上比大陆数量少许多，可能是因为台湾影评数量较少的原因。

虽然两岸观众关注的角度大致相同，但是考虑到两岸对电影类型的喜爱差异

与大陆和台湾上映的电影的差异，故我们选取了大陆和台湾都有爬取到的电影，一共 150 部左右，选取其中较为典型的 4 部电影（如表 7.2）来做比较。

表 7.2 电影关注度比较

编号	电影名称	台湾	大陆
1	横冲直撞好莱坞	2354	3461
2	美国队长	7238	3312
3	白日焰火	789	11388
4	摆渡人	955	19678

其中电影《横冲直撞好莱坞》台湾与大陆总关注度相近，大陆略多，电影《美国队长》台湾关注度远大于大陆，而《白日焰火》和《摆渡人》则是大陆远大于台湾。以下为每部电影的大属性关注度两地比较。

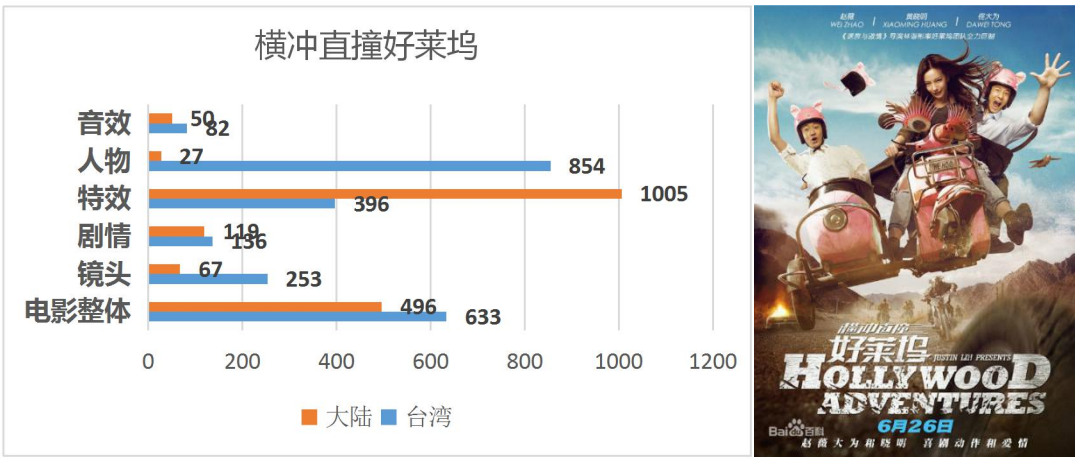


图 7.3 横冲直撞好莱坞——两地属性关注度比较

由上图 7.3 所示，电影《横冲直撞好莱坞》两地总体关注度相差不大，但是所关注的方面却大有不认同，显然，台湾最关注人物，而大陆则更关注特效。

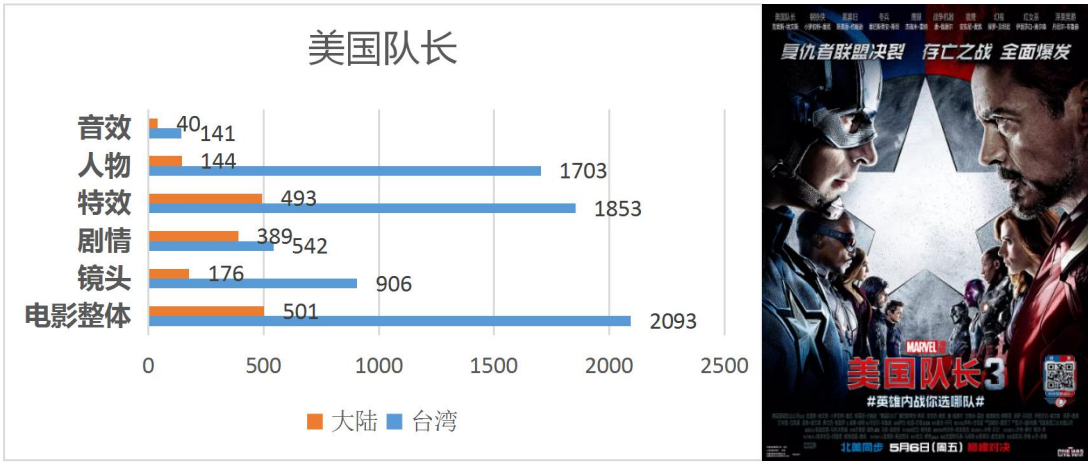


图 7.4 美国队长——两地属性关注度

由上图 7.4 所示，电影《美国队长》在台湾更受欢迎，总体关注度远大于大陆，但是两地最关注的都是电影整体和特效，在此之外，大陆会更关注剧情，而台湾则更关注镜头。

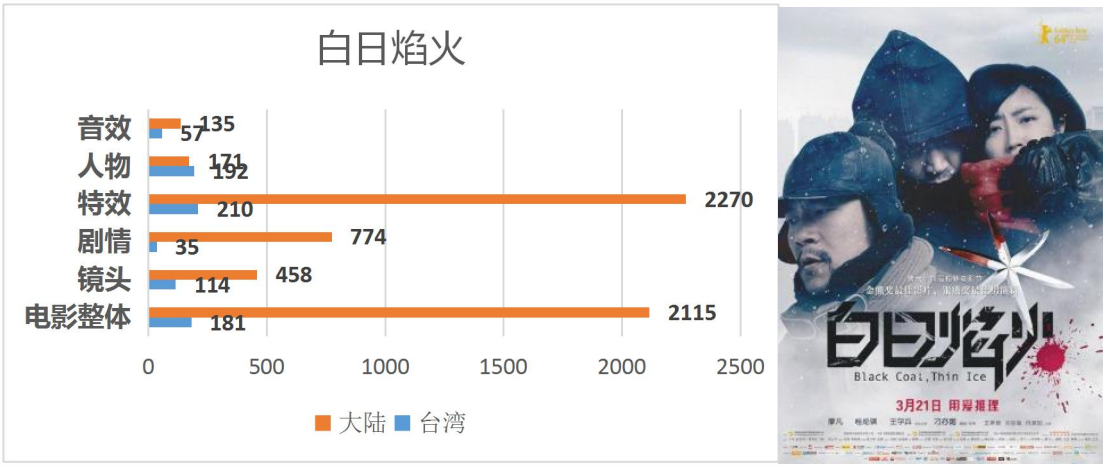


图 7.5 白色焰火——两地属性关注度

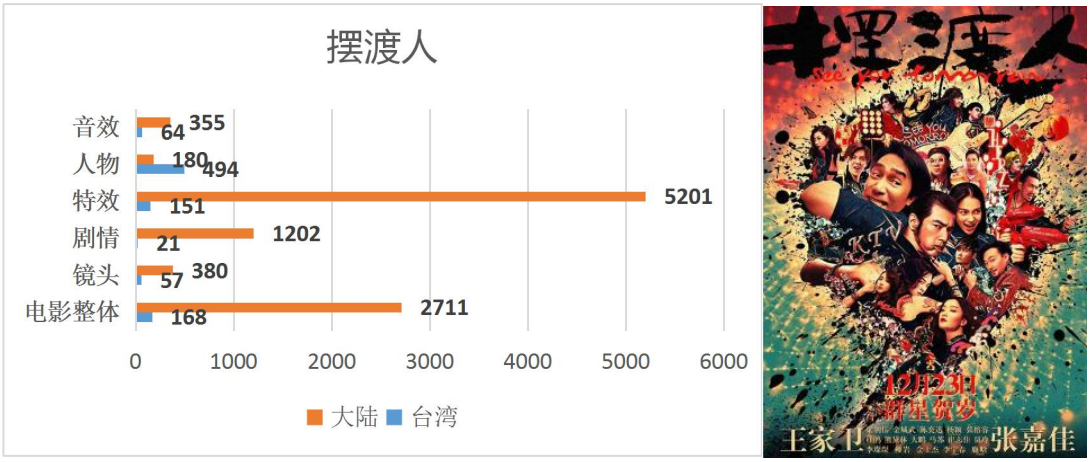


图 7.6 摆渡人——两地属性关注度

由上图 7.5 和 7.6 所示，电影《白色焰火》和《摆渡人》在大陆更受欢迎，总体关注度远大于台湾，大陆均是最关注特效，依次为电影整体，而台湾则是最关注人物、电影整体和特效。

7.3 电影属性情感分析比较

如下图 7.7 和 7.8 所示，为大陆和台湾电影属性情感分析结果表：

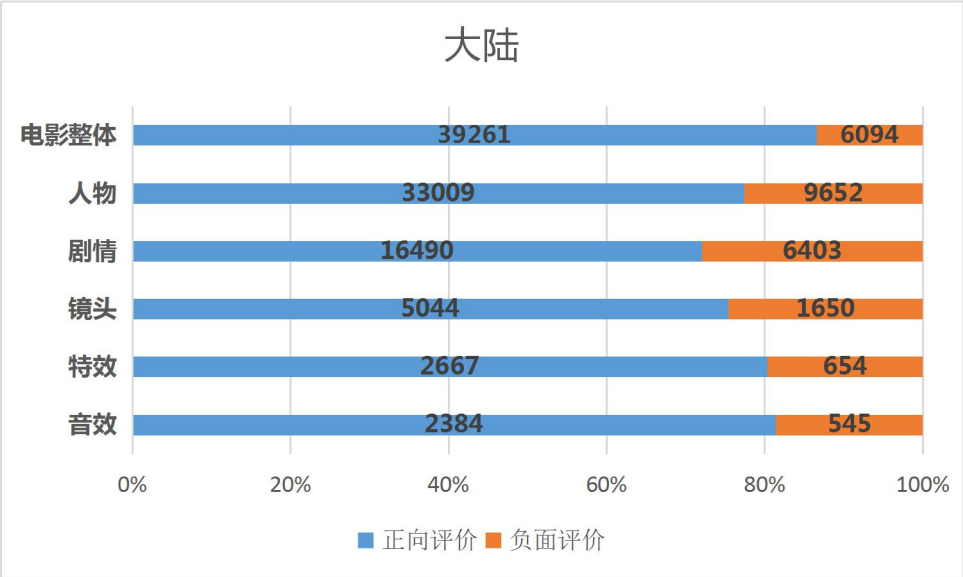


图 7.7 大陆电影属性情感分析结果

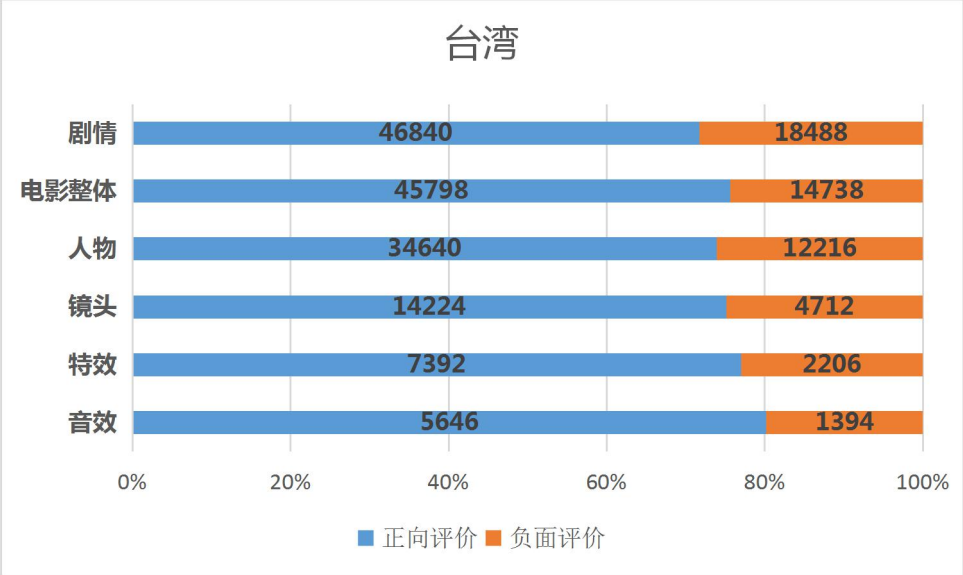


图 7.8 台湾电影属性情感分析结果

1）相同点：

A. 两岸对电影每个方面的评价都较好，只有百分之二十的不足；

B. 两地负面情感最大的属性均为剧情，表明观众对电影编剧编写的故事情节的要求相对较高；紧跟着就是人物，表明观众对于现有电影的演员选择不满意，且演员的演技及其他方面没得到观众的认可，制片人等电影创作者可以将重点放在编剧、导演、演员等（主创团体）的选择，尽可能提高观众的喜爱度。

2）不同点：

A.台湾正面情感最大的属性为音效，表明观众对于现有电影的音效较为满意，在

六大属性中最能满足大众需求；

B.大陆正面情感最大的属性为电影整体，表明观众对于现有电影都较为满意；

C.整体来看，大陆对于电影整体、音效这两部分的正向情感均大于 80；而台湾对于电影的六大属性的正面评价均在 80%以下，与大陆形成鲜明对比。

3)建议：

虽然两地对于电影不同属性的情感评价差异不是很大，但是可以看出两岸对人物与剧情都比较严格，制作者应该针对两岸地区观众对电影类型的喜爱与角度的做不同探讨，选出较优的主创团体，以求得最好票房结果。

第八章　总结

我们通过以下两点体现主创的贡献价值：

基于主创画像对排片率和评分进行回归预测，在通过排片率和评分对票房进行预测，从而间接影响票房。

对于已上映电影可以通过更换主创信息预测出不同的票房，从而体现主创的价值。

其中我们模型的重点在于通过微博数据建立主创画像，其中涉及到对细粒度属性的情感分析这一自然语言处理任务，通过人工打标签的方式获得训练数据喂入神将网络模型当中进行训练,从而得到主创画像数据。由于主创画像的模型目前是纯数据驱动的，模型效果非常依赖于 embedding 的质量，模型的效果还有很大提升的空间，例如可以加入人工特征以及继续优化模型参数等等。

得到主创画像的数据之后，通过主创画像预测排片率以及评分，通过预测的排片率和评分预测电影票房，从而使主创信息与电影票房建立了关系。预测排片率和预测评分都采用了拟合性能较好、模型鲁棒性较强的梯度提升树模型。其中预测排片率的模型效果好，但是由于目前训练的数据量还不够多，使得模型存在过拟合的风险，以后改进的通过增加训练数据的方式。

最后电影票房预测采用 Lasso 回归，模型的 R 平方是 0.7654，票房模型也有较大的提升空间，比如构建更多的特征等等。

推荐模块是其他模型的的综合应用，目的是尽可能找出票房贡献最高的主创阵容，由于主创组合方式有很多种，找出全局最优的主创阵容复杂度很高，因此采用随机优化的方式近似这个全局最优的主创阵容，具体可以查看相关环节。

团队介绍

POPCORN 团队来自于台湾淡江大学与中国科学院深圳先进技术研究院，本团队依托淡江大学信息管理所和深圳市高性能数据挖掘重点实验室，该两岸单元合作研究与实作网络意见探勘(Opinion Mining)之相关议题与技术研发。自 2010 年「部落格分析系统」之专题系统开发开始，针对情感分析(Sentiment Analysis)之理论与基础架构进行研发，以及网络意见探勘之应用领域进行系统与专题开发。双方就技术研发与应用系统开发已有初步成果，并共同表多篇论文。

指导老师： 萧瑞祥

博士，现为淡江大学资管系教授，中国科学院深圳先进技术研究院客座研究员。曾任淡江大学资管系系主任、创新育成中心主任、代理研发长、台湾信息管理学会秘书长、创新资服竞赛副主任委员等。目前主要研究专长领域为：社群网络意见探勘、信息教育、信息策略与管理、电子商务等。在信息安全、互联网舆

情分析、云计算应用等领域，具有一定的知名度，发表学术论文 60 多篇，20 多项专利。先后承担台湾科技部「台湾企业导入个人资料保护与管理制度及隐私标章之研究」计划，国科会「信息安全治理之实践性研究」计划，国科会「产学合作计划－校园云端服务管理系统整合研发」计划，国科会「信息安全管理系统遵循性辅助工具设计之研究」计划等重大项目。

队长：车丹丹

中国科学院深圳先进技术研究院计算机技术 2017 级硕士研究生，研究方向为数据分析和数据挖掘，初步掌握网络爬虫框架及工具。曾获得美国数学建模大赛二等奖，挑战杯全国大学生大赛三等奖，市场调查分析河北省二等奖等，并多次获得综合测评和创新创业奖学金。

成员：马强

中国科学院先进技术研究所数据挖掘实验室客座研究生，香港城市大学 2016 级研究生，主要研究方向数据挖掘、自然语言处理、情感分析，熟悉 java

及 python 开发，有较多的项目经验。

成员：黄妍明

中国科学院先进技术研究所数据挖掘实验室客座研究生，华南师范大学软件工程专业 2015 级研究生，研究方向自然语言处理，数据挖掘，机器学习。熟悉网络爬虫框架及工具，TensorFlow，theano 等框架，曾获得国家奖学金，中国大学生服务外包比赛三等奖，挑战杯软件比赛优秀奖等。

成员：钟琼丽

中国科学院先进技术研究所数据挖掘实验室客座研究生，汕头大学计算机技术专业 2016 级研究生，研究方向是数据挖掘、自然语言处理，在深度学习、机器学习领域内学习，主要从事网络数据采集与分布式数据处理、存储，曾多次获得综合测评一等奖及三好学生。

成员：陳竑嘉

台湾淡江大学资讯管理学系 2016 级硕士研究生。研究方向为舆情分析、文字探勘。熟悉使用 python 及 C#，主要应用于网络爬虫及情感分析。