《信息检索与数据挖掘》课程作业

——对电影票房数据集的数据挖掘

姓名：张雪文

学号：162110104

班级：1621001

目　　录

1. 数据预处理

1.1 数据集介绍

（1）数据集来源：本课程作业所使用的数据集爬取自艺恩娱数网（网址为：https://ys.endata.cn/DataMarket/Index），原始为 xlsx 类型文件，后处理成 csv 文件。记录了从 2023 年 8 月 1 日到 10 月 25 日的电影票房的相关信息。

（2）数据集规模：

数据集共有 11713 个样本，44 个特征。所有特征名称如下：

日期	排名	电影 ID	电影名称	影片英文名称
当前票房(万)	累计票房(万)	累计场次	累计人次(万)	天数
票房占比	当前场次	当前人次(万)	人次占比	场均人次
场均收入,	黄金场票房(万),	黄金场场次,	黄金场人次(万),	黄金场排座(万)
黄金场场均人次	票房环比	场次环比	人次环比	场次占比
上午场票房(万)	上午场场次	上午场人次(万)	下午场票房(万)	下午场场次
下午场人次(万)	加映场票房(万)	加映场场次	加映场人次(万)	上座率
黄金场票房占比	黄金场场次占比	黄金场人次占比	黄金场上座率	票房占全国比
当前排座(万)	排座占比			

其中，19 个特征数据类型为 float64 类型，21 个特征数据类型为 int64 类型，4 个特征数据为其他类型。

1.2 数据预处理

使用 python 相关库函数统计后，发现数据集无重复数据，无异常数据，但存在缺失值。另外，为了方便后期处理，需要对一些数据格式进行转换，并根据不同的应用进行数据归一化处理

（1）基于 KNN 插补算法的缺失值处理

进行缺失值统计时，结果如左图（由于排版问题一些特征的缺失值数量被裁剪掉，被裁剪掉的特征缺失值数量均为 0）。可以看到，缺失情况如下：

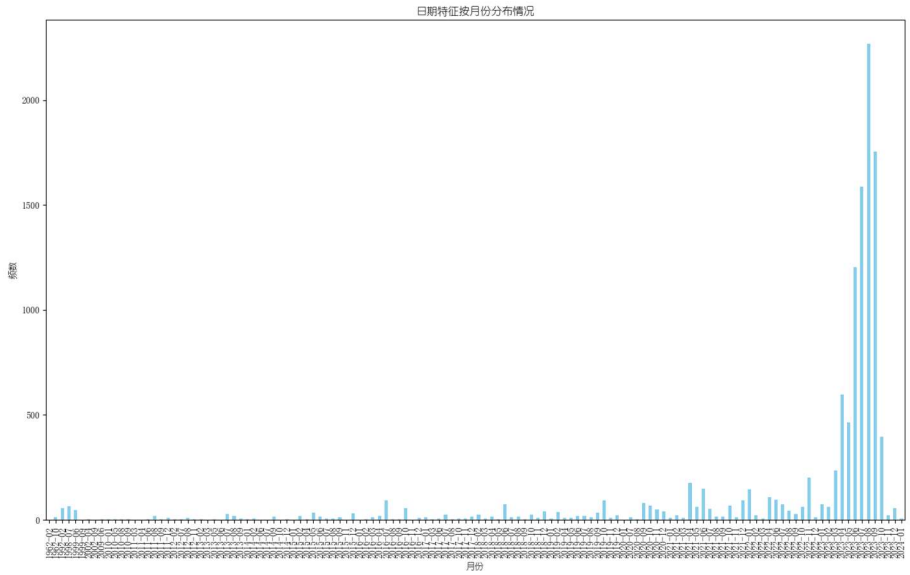
影片英文名称：缺失 196

累计上映天数：缺失 782

上映日期：缺失 782

经过分析，影片英文名和上映日期在后续处理中使用不到（上映日期可由累计上映天数反应），故只需对‘累计上映天数‘列进行数据填充。

开始尝试使用累计上映天数的众数或平均值进行填充，但经过统计发现（上映日期月份分布如下图所示），大多数电影上映月份都位于 2023 年 7 月至 9 月，累计上映天数的众数或平均值都会较小。而缺失累计上映天数的样本主要分为两种类型：2023 年近期的电影展演活动（如：2023-2024 环球经典 IP 影片复映活动、2023 成都·天府科幻电影展）和一些上映日期较为靠前的



老电影（如：老兵新传、苍穹），两极分化有些严重。

经过查阅资料和分析数据，最终使用 KNN 插补算法，完成对空缺数据的填补：

首先使用 MinMaxScaler 方法对下列特征进行归一化处理，然后通过使用欧氏距离计算不同样本之间的距离，找出与缺失值最接近的其他 5 个样本，用这些样本的平均值来替代缺失值，并进行取整操作。最终填补效果较为贴合实际。

```
[['排名','电影ID','累计上映天数','当前票房(万)','累计票房(万)','累计场次','累计人次(万)',
 '票房占比','当前场次','当前人次(万)','人次占比','场均人次','场均收入','黄金场票房(万)',
 '黄金场场次','黄金场人次(万)','黄金场排座(万)','黄金场场均人次','票房环比','场次环比',
 '人次环比','场次占比','上午场票房(万)','上午场场次','上午场人次(万)','下午场票房(万)','下午场场次',
 '下午场人次(万)','加映场票房(万)','加映场场次','加映场人次(万)','上座率','黄金场票房占比',
 '黄金场场次占比','黄金场人次占比','黄金场上座率','票房占全国比','当前排座(万)','排座占比']]  # 选择需
```

（2）将‘日期‘列字符数据进行转换

数据集的一个重要特征是时间，但是原数据时间类型为字符型（格式如：2023-10-25），故增加一列特征名为 date 的数据，将‘日期‘列数据转换为如 1025 的整型类型，计算方式为：月份*100+日期。

2.使用主成分分析对数据进行降维

2.1 算法思路

由于本项目是对电影票房相关信息的数据挖掘，与电影名等一些字符类型无关，故使用 41 个数值类型列作为初始数据矩阵（使用特征如下图所示）。

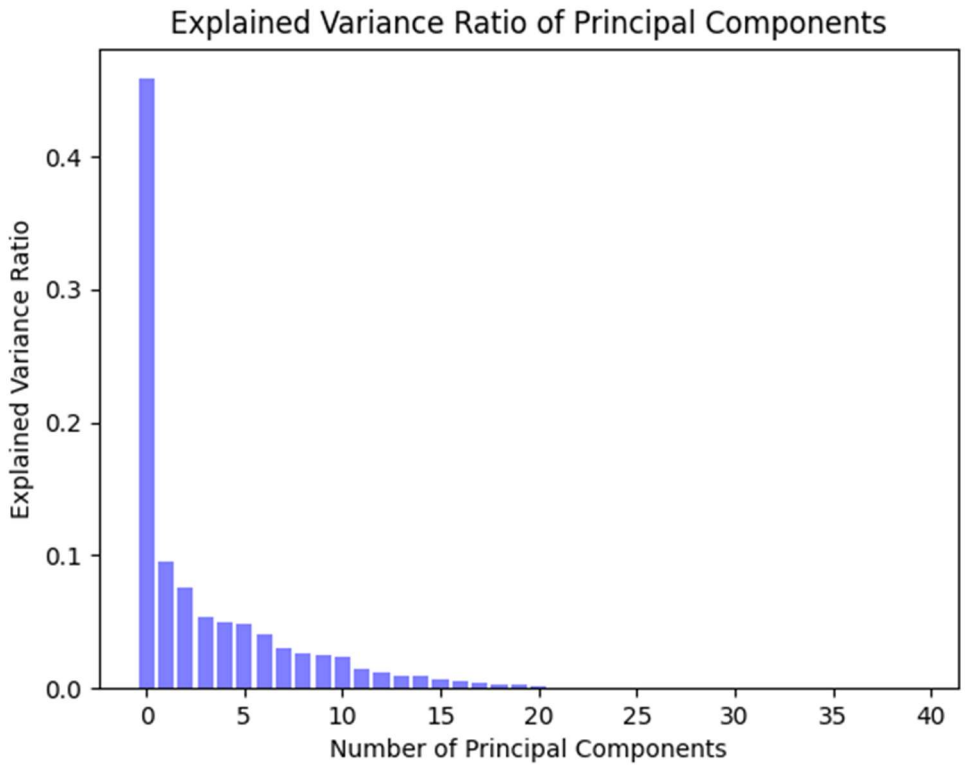
```
['date', '排名', '电影ID', '天数', '累计上映天数', '当前票房(万)', '累计票房(万)', '累计场次', '累计人次(万)',
 '票房占比', '当前场次', '当前人次(万)', '人次占比', '场均人次', '场均收入', '黄金场票房(万)',
 '黄金场场次', '黄金场人次(万)', '黄金场排座(万)', '黄金场场均人次', '票房环比', '场次环比',
 '人次环比', '场次占比', '上午场票房(万)', '上午场场次', '上午场人次(万)', '下午场票房(万)', '下午场场次',
 '下午场人次(万)', '加映场票房(万)', '加映场场次', '加映场人次(万)', '上座率', '黄金场票房占比',
 '黄金场场次占比', '黄金场人次占比', '黄金场上座率', '票房占全国比', '当前排座(万)', '排座占比']]
```

为了排除数据本身量级的不同对于降维结果的影响，首先使用 StandardScaler 对数据进行标准化数据归一化处理。

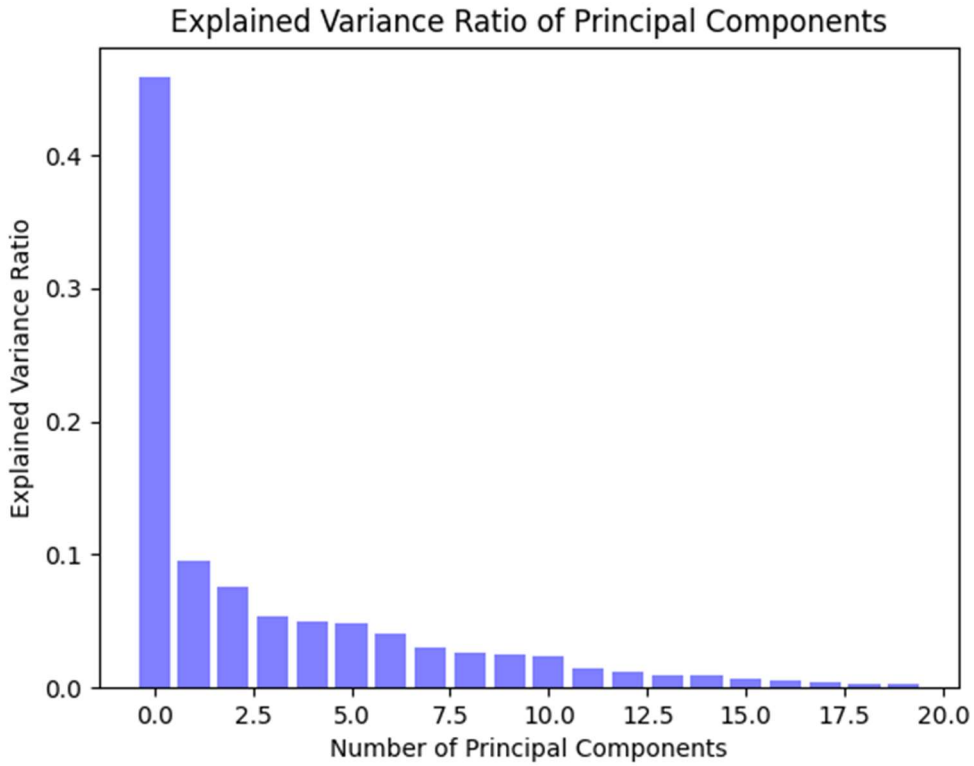
对数据进行主成分分析，选取特征值的贡献率之和超过 0.9 的前几项特征向量作为降维后的新数据矩阵。

2.2 结果分析

下图是 41 个特征值的贡献率：



可以看到，后 21 个特征值几乎没有贡献。只需关注前 20 个特征。前 20 个特征值的贡献率如下图：



具体数值依次为:

[0.46130282 0.0953302 0.07596288 0.05297236 0.05007483 0.04694502
0.03973764 0.03087882 0.02650493 0.0245756 0.02362317 0.01467788
0.01229844 0.00996891 0.00895871 0.00635422 0.0057132 0.0044052
0.00284691 0.00237473]

通过计算得到，前十个特征的贡献率的和为 0.9039715200000001，故取前十个作为主成分。

主成分分析后，数据矩阵维度降为 10，如下图:

	0	1	2	...	7	8	9
0	14.773635	2.611271	-2.533863	...	2.033682	0.481085	-2.006192
1	10.770290	1.413355	-1.930440	...	0.972586	-0.108617	-1.388226
2	13.346406	1.467778	-3.301268	...	1.524102	-0.632028	-1.568540
3	4.243579	1.419898	-3.301717	...	-0.141220	0.644709	-0.624020
4	6.092144	1.404821	-4.554240	...	0.206704	0.370448	-0.753119
...	...	...	...	...	...	...	...
12523	-0.991834	-1.574401	0.529582	...	1.127488	-0.854784	0.452037
12524	-0.993855	-1.584758	0.529529	...	1.130830	-0.854342	0.452544
12525	-0.994109	-1.590029	0.522480	...	1.144956	-0.855050	0.453199
12526	-0.996342	-1.591944	0.532085	...	1.167216	-0.855918	0.453530
12527	-0.997201	-1.599178	0.529806	...	1.177482	-0.856181	0.453548

[12528 rows x 10 columns]

3.聚类

3.1 算法思路

由于原始数据并没有对数据进行分类，且数据样本量较大，故使用聚类算法对数据进行分类。使用 K-means 算法进行聚类。、

K-means 算法有一个问题，参数 K 如何选择？本项目使用计算不同 K 值下的轮廓系数方法，通过比较得到 K 的最优值。

轮廓系数轮廓系数综合了样本与其所属簇内的相似度以及与最近的其他簇间的不相似度。它的计算方法如下：

1. 对于每个样本，计算与同簇其他样本的平均距离（a）。
2. 对于每个样本，计算与最近簇内样本所在簇的平均距离（b）。
3. 轮廓系数 s 计算公式如下：

$$s = \frac{b - a}{\max(a, b)}$$

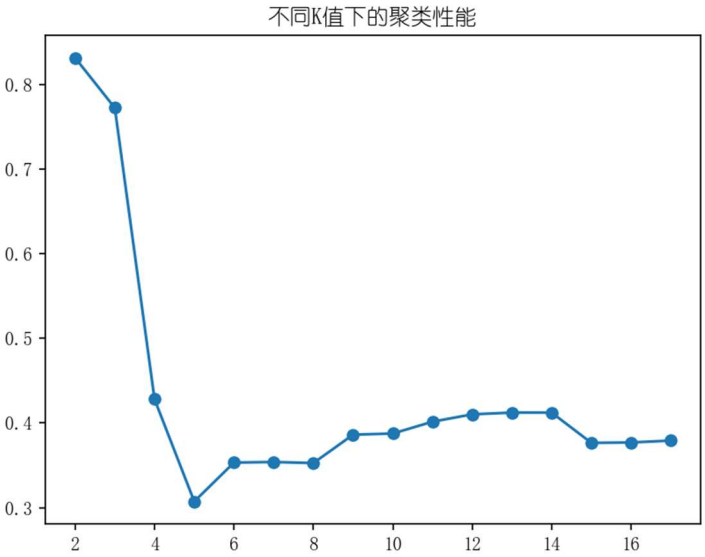
轮廓系数的取值范围在-1 到 1 之间。轮廓系数越接近 1，表示样本聚类越准确合理，簇内距离较小且簇间距离较大；接近-1 则表示聚类结果差，样本被错误地分配到了相邻簇；若值在 0 值附近，则说明样本在两个簇的边界上，样本聚类重叠。

故可通过计算不同 k 值下的轮廓系数，来找到 K 值的最优解。

本项目计算了在 k 分别取 2 到 18 的情况下的轮廓系数。

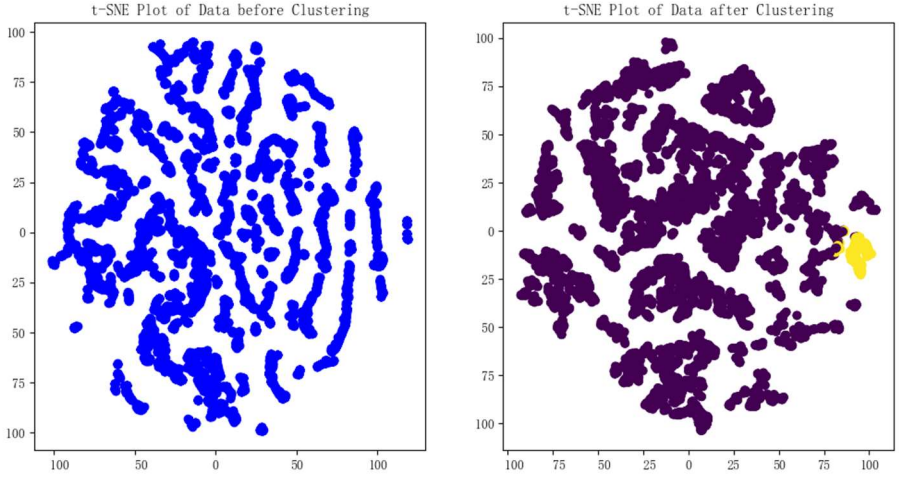
3.2 结果分析

下图为 k 取 2-18 间不同值时的轮廓系数曲线：



可以看到，当 K=2 时，轮廓系数最大，因此此时聚类效果最好。轮廓系数约为 0.83，接近于 1，表明此时聚类效果好，故最终使用 2-means 算法在上一步骤主成分分析降维后的数据矩阵上对样本进行聚类。

进一步分析聚类前后的 tsne 图（如下图），可以看到，聚类后的样本虽然被分为了两类，但是两类样本量非常不平衡。



将分类结果保存下来并进一步查看，并结合作者相关经验与生活体验和网上进一步资料查询，发现分类标签为 1 的样本基本上为爆款电影（相对）（如下图）：

Column1	日期	排名	电影ID	电影名称	影片英文名称	当前票房(万)	累计票房(万)
0	2023/10/25	1	39755	河边的错误	Only the River Flows	13046497.11	134312963.4
1	2023/10/25	2	39741	坚如磐石	Under The Light	5675035.26	1284150654
126	2023/10/24	1	39755	河边的错误	Only the River Flows	13835464.47	121266466.3
127	2023/10/24	2	39741	坚如磐石	Under The Light	5946112.5	1278475619
238	2023/10/23	1	39755	河边的错误	Only the River Flows	15708044.26	107431001.8
239	2023/10/23	2	39741	坚如磐石	Under The Light	6283470.14	1272529506
342	2023/10/22	1	39755	河边的错误	Only the River Flows	39507577.3	91722957.52
343	2023/10/22	2	39741	坚如磐石	Under The Light	14977387.72	1266246036
493	2023/10/21	1	39755	河边的错误	Only the River Flows	52207680.22	52215380.22
494	2023/10/21	2	39741	坚如磐石	Under The Light	21138408.82	1251268648
651	2023/10/20	1	39741	坚如磐石	Under The Light	13400721.62	1230130239
652	2023/10/20	2	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	12199202.1	712793029.3
795	2023/10/19	1	39741	坚如磐石	Under The Light	10079827.53	1216729518
796	2023/10/19	2	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	7955763.75	700593827.2
798	2023/10/19	4	39742	前任4：英年早婚	The Ex-file 4	6686487.86	916197698.5
918	2023/10/18	1	39741	坚如磐石	Under The Light	10782754.61	1206649690
919	2023/10/18	2	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	7922956.56	692638063.5
920	2023/10/18	3	39742	前任4：英年早婚	The Ex-file 4	7167664.28	909511210.6
1047	2023/10/17	1	39741	坚如磐石	Under The Light	11379368.99	1195866936
1048	2023/10/17	2	39742	前任4：英年早婚	The Ex-file 4	7810059.65	902343546.4
1049	2023/10/17	3	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	7561518.55	684715106.9
1166	2023/10/16	1	39741	坚如磐石	Under The Light	12243237.17	1184487567
1167	2023/10/16	2	39742	前任4：英年早婚	The Ex-file 4	8606179.01	894533486.7
1168	2023/10/16	3	39745	莫斯科行动	Moscow Mission	7532212.22	557109164.6
1169	2023/10/16	4	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	7385496.32	677153588.4
1278	2023/10/15	1	39741	坚如磐石	Under The Light	30730553.7	1172244330
1279	2023/10/15	2	39742	前任4：英年早婚	The Ex-file 4	20295989.86	885927307.7
1280	2023/10/15	3	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	20189238.42	669768092.1
1281	2023/10/15	4	39745	莫斯科行动	Moscow Mission	19721379.82	549576952.4
1427	2023/10/14	1	39741	坚如磐石	Under The Light	42481379.13	1141513776
1428	2023/10/14	2	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	28203049.85	649578853.6
1429	2023/10/14	3	39742	前任4：英年早婚	The Ex-file 4	27705822.95	865631317.8
1430	2023/10/14	4	39745	莫斯科行动	Moscow Mission	26418654.65	529855572.5
1576	2023/10/13	1	39741	坚如磐石	Under The Light	23169561	1099032397
1577	2023/10/13	2	39740	志愿军：雄兵出击	THE VOLUNTEERS：TO THE WAR	18303117.93	621375803.8

而分类标签为 1 的样本为相对非爆款电影（或者由于累计上映日期长在当日已不算爆款），如下图：

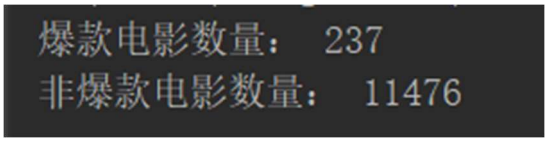
Column1	日期	排名	电影ID	电影名称	影片英文名称	当前票房(万)	累计票房(万)
17	2023/10/25	18	39654	孤注一掷	No More Bets	76418.34	3850005887
18	2023/10/25	19	39555	力量密码	The Source Of Power	67750	16131465.12
19	2023/10/25	20	39699	第八个嫌疑人	Dust To Dust	59049.7	439181448.9
20	2023/10/25	21	39763	白塔之光	The Shadowless Tower	49309	131007.5
21	2023/10/25	22	39406	望道	Manifesto	47670	80460071.53
22	2023/10/25	23	39770	拯救嫌疑人	Who's the suspect	47037	52237
23	2023/10/25	24	39737	贝肯熊：火星任务	BACKKOM BEAR: MARS MISSION	46432.25	72101885.82
24	2023/10/25	25	36770	苍穹	the sky	30640	16713300
25	2023/10/25	26	39609	封神第一部：朝歌风云	Fengshen Trilogy	28865.64	2634514048
26	2023/10/25	27	39722	敢死队4：最终章	Expend4bles	21477.59	156216924.3
27	2023/10/25	28	39733	功夫王之萌虎上山	KUNG FU TIGER	20436.02	4912403.56
28	2023/10/25	29	39633	热烈	One And Only	17103.31	913023574.4
29	2023/10/25	30	39759	故园飘梦	Lingering Dream of Homeland	16384.42	66855.79
30	2023/10/25	31	39679	青春破晓	LIGHTING THE YOUNG	10470	58367
31	2023/10/25	32	38829	邓小平小道	The Man Of People	9450	26655358.65
32	2023/10/25	33	39756	一个和四个	One and Four	9111.4	124616.4
33	2023/10/25	34	39778	少年先锋	YOUNG PIONEERS	8250	10440
34	2023/10/25	35	39736	我是哪吒2之英雄归来	I Am Nezha	7186.4	16661691.94
35	2023/10/25	36	39563	爸爸，我懂你了	Second Chance with Dad	6025	786860.89
36	2023/10/25	37	39731	看不见的顶峰	Invisible Summit	5775	393105.16
37	2023/10/25	38	39622	谁说我不靠谱	Aren't I Reliable	4300	1063326
38	2023/10/25	39	39758	樱桃阑下	Ying Tao Gu Xia	4290	23979
39	2023/10/25	40	39568	远山花开	Blossom in The Mountains	4250	1294693.09
40	2023/10/25	41	38935	井冈星火	SPARKS OF FIRE IN JINGGANGSHAN	4235	16236434.52
41	2023/10/25	42	39753	黄鹤楼之盐道迷局	THE CONSTABLE AND THE SNAKE CHARMER	3940.8	69856.26
42	2023/10/25	43	39661	萤火虫的天空	The sky of Firefly	3306	1448418.38
43	2023/10/25	44	39762	洋子的困惑	Yangzi's Confusion	2815.5	160403.91
44	2023/10/25	45	39174	以法之名	For Justice	2500	354395.9
45	2023/10/25	46	39730	火凤重天	Phoenix in Fire	2450	465890.73
46	2023/10/25	47	39751	余下都是春天	Tomrrow is another day	2060	51913
47	2023/10/25	48	39750	花开那年	The Year of Blossoms	2041	34766.18
48	2023/10/25	49	39586	幸福小马灯	Happiness of century lantern	1812	315515.52

故可基本认为聚类后，数据被分为了两类：是/否为爆款电影。故在原数据后再加上一列特征——‘是否为爆款电影‘作为数据标签，1 为是，0 为否，来进行下一步的分类。

4. 分类

4.1 基于随机森林算法对数据进行分类

通过进一步具体统计爆款电影数量和非爆款电影数量，得到如下结果：



可以看到，两个类别的样本量非常不平衡。故先使用 SMOTE 方法进行过采样来处理样本不平衡问题。

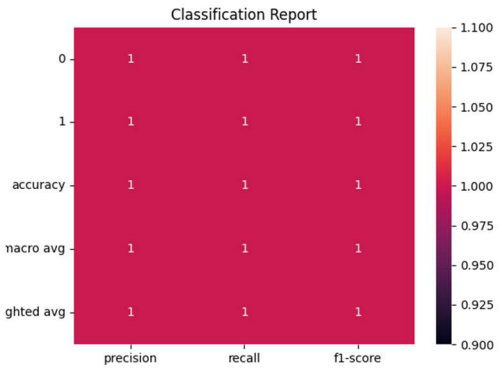
将样本按照 7：3 的比例划分为训练集和测试集。

接着，使用随机森林算法（此处使用决策树分类器作为基分类器），将‘是否为爆款电影‘作为目标变量，下列特征作为分类特征对样本进行分类：

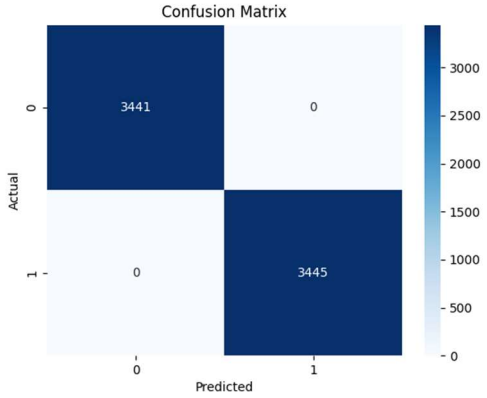
```
# 特征
X=data.loc[:,['date','排名','电影ID','天数','累计上映天数','当前票房(万)','累计票房(万)','累计场次','累计人次(万)',
              '票房占比','当前场次','当前人次(万)','人次占比','场均人次','场均收入','黄金场票房(万)',
              '黄金场场次','黄金场人次(万)','黄金场排座(万)','黄金场场均人次','票房环比','场次环比',
              '人次环比','场次占比','上午场票房(万)','上午场场次','上午场人次(万)','下午场票房(万)','下午场场次',
              '下午场人次(万)','加映场票房(万)','加映场场次','加映场人次(万)','上座率','黄金场票房占比',
              '黄金场场次占比','黄金场人次占比','黄金场上座率','票房占全国比','当前排座(万)','排座占比']]
```

4.2 结果分析

当按照 8：2 的比例将数据集划分为训练集和测试集进行分类时，分类结果如下：



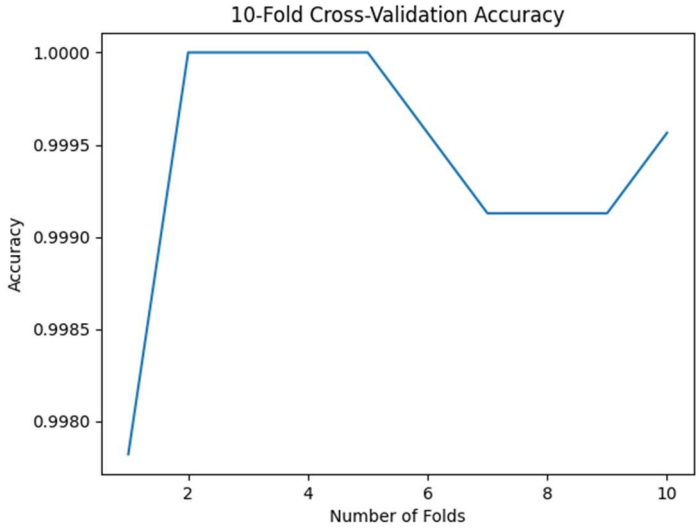
准确率等



混淆矩阵

可以看到，分类的准确率等于 100%，太过于理想了。初步猜测可能是由于样本总数只有一万多条，数据量过少导致的结果。

于是使用 10 折交叉验证的方式进行分类模型评估。得到的十次验证的准确率折线图如下图所示：



可以看到，10 次的准确率都较高，基本都在 0.9975 以上。

分析两次结果，可以看到，随即森林分类器的分类效果很好，甚至可以说是过于理想了。分析原因，可能是因为使用了过采样的方式对不平衡数据集进行了补充，导致分类器过拟合了。另外，由于标签是由上一步骤的聚类得到的，而上一步骤的聚类效果较好，导致数据很容易就可以根据标签被分类。