数据科学概论大作业 分析报告

小组成员：龚舒凯、郑楚睿、徐瀚臣、郭尔夫

关键词： 聚类分析、决策树、随机森林、主成分分析、神经网络、深度学习、比赛结果预测

1 背景概述

1.1 行业发展背景

NBA 作为全球最知名的职业篮球联赛之一，每年都吸引着数百万球迷的关注。同时，NBA 也是一个数据丰富的领域，每场比赛都蕴含着大量的统计数据，如得分、篮板、助攻、抢断等等。这些统计数据往往反映了比赛走势、球员状态，甚至决定了球队的输赢情况。因此，无论是对球队、球迷还是博彩业，数据科学在体育行业中的应用越来越受到重视。

1.2 课题概述

在这样的行业发展背景下，我们小组所担任的角色是**篮球彩票的数据分析团队**，希望通过数据科学的方法来预测 NBA 球队在 2022-2023 季后赛中的输赢情况，从而为购买彩票的散户和中国体育彩票**提供购买彩票和设定赔率的参考**。

概括而言，首先，我们利用**聚类分析**将球队分为几种不同的类型，并结合搜集的球队技术统计数据，训练**随机森林模型和神经网络模型** [2] 这两种模型来判断比赛的输赢概率，平行比较两种模型的准确性和稳定性，并选取表现更佳的模型来预测季后赛的淘汰结果。

1.3 数据来源

Basketball Reference[1] 下属于 Sports Reference，是与 NBA 官方合作的第三方权威技术统计数据平台，网站涵盖从 1996 年以来 NBA 比赛中得分、篮板、抢断等多重维度数据。由此，我们从 Basketball Reference 上扒取了 NBA 中东部西部共 30 支球队的命中数、三分数、罚球数、篮板数等篮球技术统计数据，作为我们数据分析的基础。在以下的部分，各队用下表所示代号表示：

球队名称	简称	球队名称	简称	球队名称	简称	球队名称	简称	球队名称	简称
勇士	GSW	快船	LAC	国王	SAC	太阳	PHX	湖人	LAL
火箭	HOU	马刺	SAS	独行侠	DAL	灰熊	MEM	鹈鹕	NOP
雷霆	OKC	开拓者	POR	爵士	UTA	掘金	DEN	森林狼	MIN
热火	MIA	老鹰	ATL	黄蜂	CHA	奇才	WAS	魔术	ORL
骑士	CLE	步行者	IND	活塞	DET	公牛	CHI	雄鹿	MIL
猛龙	TOR	凯尔特人	BOS	尼克斯	NYK	篮网	BKN	76 人	PHI

2 数据概述

2.1 数据选择

在聚类分析和训练模型前，我们选取了 2022-2023 赛季 NBA 的全部常规赛数据进行训练。我们注意到 NBA 的一个特点：赛季中期交易日前后队伍的实力变化可能极大，球队在中期交易日前后会出现明显的实力变化，以整个赛季的平均状态不能反映季后赛之前各支球队的真实状态，且由于到常规赛赛季末倒数 5 场比赛，部分球队提前获取季后赛席位后，球队的求胜欲降低，可能会对后续的季后赛胜负判断产生混淆。

因此在聚类分析和季后赛预测时，我们**把这赛季常规赛每支球队的交易截止日后以及去掉倒数五场比赛后的平均数据**作为聚类的依据，而不是整个赛季的平均值。在训练模型时，我们仍然选用全部的常规赛原始数据进行训练。

2.2 数据处理

我们综合考虑反映球队实力与状态的多元数据，考虑的指标可分为**数量型与比率型**，包括**出手数，命中率，三分数，主客场优势**等等。

然后，我们将上述指标作如下处理：**对于数量型指标**，我们将两个队某个变量的做差比较。比如在勇士 vs 湖人的某场比赛中，勇士篮板 100，湖人篮板 90，则勇士的篮板差为 $+10$；湖人的篮板差为 -10。**对于比率型指标**，我们求算两个队关于某个变量的比值。例如勇士的命中率为 0.5，湖人的命中率为 0.43，则勇士对湖人的命中率比为 $\frac{0.5}{0.43} = 1.16$

这样处理的目的是将对垒双方的数据向量 $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}$ 统一成一个数据向量 $\boldsymbol{\Delta\alpha}$，从而便于输入我们的模型进行计算。原始指标与处理后的指标如下表所示：

原始指标	原始指标简称	处理后指标	处理后指标简称	原始指标	原始指标简称	处理后指标	处理后指标简称
命中数	FG	命中数差	Δ FG	抢断	STL	抢断差	Δ STL
出手数	FGA	出手数差	Δ FGA	盖帽	BLK	盖帽差	Δ BLK
命中率	FG%	命中率比	Δ FG%	失误	TOV	失误差	Δ TOV
三分数	3P	三分数差	Δ 3P	犯规	PF	犯规差	Δ PF
三分出手	3PA	三分出手差	Δ 3PA	进攻效率	ORtg	进攻效率比	Δ ORtg
三分命中率	3P%	三分命中率比	Δ 3P%	防守效率	DRtg	防守效率比	Δ DRtg
罚球命中	FT	罚球差	Δ FT	回合数	Pace	回合数差	Δ Pace
罚球数	FTA	罚球数差	Δ FTA	有效命中率	eFG%	有效命中率比	Δ eFG%
罚球命中率	FT%	罚球命中率比	Δ FT%	失误率	TOV%	失误率比	Δ TOV%
进攻篮板	ORB	进攻篮板差	Δ ORB	进攻篮板率	ORB%	进攻篮板率比	Δ ORB%
总篮板	TRB	总篮板差	Δ TRB	防守篮板率	DRB%	防守篮板率比	Δ DRB%
助攻	AST	助攻差	Δ AST	罚球数占比	FT/FGA%	罚球数占比之比	Δ FT/FGA%

处理过后的数据集如下图所示，第一列即对垒双方的输赢情况，其他列为对垒双方的数据差值/比值

赢球结果	命中数差	出手数差	命中率比	三分数差	三分出手差	三分命中率比	罚球命中差	罚球数差	罚球命中率比	...	犯规差	进攻效率	防守效率	回合数	有效命中率比	失误率比	进攻篮板率比	防守篮板率比	罚球数占比之比	主客场
0	-5	-10	0.991507	6	13	1.098901	2	12	0.715278	...	-7	109.6	111.7	95.8	1.081081	1.053571	0.725806	0.941040	1.250000	0
1	11	11	1.162791	8	6	1.597070	11	12	1.051768	...	-5	137.5	96.3	99.6	1.219713	0.775000	2.821429	1.421488	1.381743	1
0	-3	6	0.863558	0	12	0.693694	4	6	0.980609	...	-5	122.9	125.1	90.3	0.870821	0.741259	2.503759	1.299850	1.210526	0
1	-2	-12	1.095238	8	5	1.457726	0	5	0.856667	...	-5	125.2	121.3	93.3	1.195446	1.376147	0.592715	0.850183	1.148276	0
0	-6	6	0.809717	3	15	1.000000	3	3	1.025840	...	1	102.3	107.9	98.2	0.846743	0.992187	1.224490	1.057895	1.051852	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	-7	-9	0.944551	3	5	1.064039	2	0	1.153740	...	1	110.8	119.9	98.4	0.996650	1.669903	0.771429	0.953757	1.283784	0
0	0	7	0.930018	-10	-15	0.877737	-2	-8	1.218430	...	2	117.9	129.0	108.6	0.854790	0.523810	0.636704	0.883133	0.821622	1
0	2	10	0.935547	-7	-2	0.494226	-15	-14	0.878857	...	3	111.6	129.0	103.9	0.867572	0.860294	0.811688	0.922667	0.510791	0
1	-4	-14	1.056948	-8	-14	0.664577	22	25	1.344768	...	-11	111.3	105.4	102.5	0.982524	1.029851	0.992565	0.997271	4.859155	1
0	-3	7	0.865031	8	24	1.087413	-7	-7	0.866947	...	6	106.5	111.4	102.3	0.948276	0.789062	1.034615	1.012312	0.603604	1

3 模型建立

3.1 模型 1：球队类型判断模型 [3]

NBA 中的不同球队因球员属性和战术打法而可以分为不同的类型。显然，球队类型也会反映在球队的技术数据之中（三分数、盖帽数等）。**我们希望提炼出 30 支球队的球队类型，借由数学模型探索球队之间的相克关系，进而有利于后续的输赢判断。**

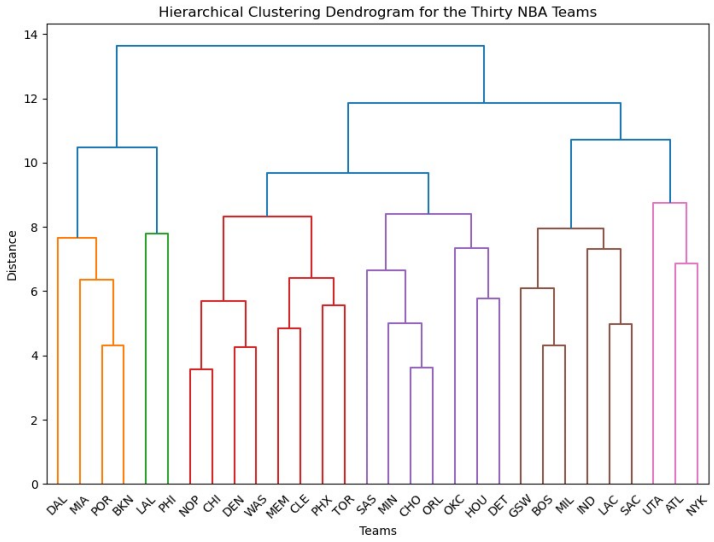
我们将**把这赛季常规赛每支球队的交易截止日后以及去掉倒数五场比赛后的**平均数据作为聚类的依据。为了方便聚类和后续的模型建模与分析，我们**运用 z-score 方法，对各项数据进行了标准化**。即

$$z = \frac{x - \mu}{\sigma}$$

30 支球队的平均数据如下图所示：

Team	Tm	Opp	FG	FGA	FG%	3P	3PA	3P%	FT	...	TRB%	AST%	STL%	BLK%	ofeFG%	ofTOV%	ofORB%	ofFT/FGA	dfDRB%	dfFT/FGA
DAL	0.373345	0.491133	-0.629480	-0.897691	0.101978	2.002741	1.742537	1.564234	0.084402	...	-1.970524	-0.609327	-1.377710	-1.335569	1.214715	-0.658044	-2.434353	0.298891	-0.451685	0.580091
DEN	-0.386157	-0.945883	0.396753	-0.473309	1.000566	-0.670972	-0.640505	-0.633213	-0.909317	...	0.791750	2.108232	-0.285425	-0.475751	0.455118	-0.094872	0.404239	-0.701543	0.483216	-0.901134
GSW	0.461518	0.131613	0.204007	0.365741	-0.046301	1.845541	1.980871	0.764114	-1.039650	...	0.478766	2.008667	0.019117	-0.721380	0.855028	1.652902	0.668608	-0.958038	-0.231713	-0.373199
HOU	-0.878808	1.399229	-0.330894	0.417001	-0.931851	-1.639709	-1.159758	-2.279376	0.185088	...	1.093986	-2.068480	0.450722	-0.932494	-1.544047	0.630131	2.221393	0.067651	0.122558	1.148743
LAC	1.235483	0.752885	0.722940	-0.710725	1.730368	0.499585	-0.172808	1.591710	0.921986	...	-0.363395	0.734184	0.971331	-1.340232	1.612061	0.948039	-0.615378	0.898402	-0.578461	-0.830074
LAL	-0.088514	-0.915427	-0.832720	-0.755241	-0.234114	-0.767846	-0.706792	-0.395301	1.958023	...	0.824646	-0.003877	-0.658473	-0.318746	-0.465606	0.707810	0.415010	1.912753	0.206868	-2.549856
MEM	0.628068	-0.401367	1.018333	1.144763	0.044146	0.389134	0.578284	-0.314030	-0.788932	...	-0.816531	-0.145930	1.030416	1.213287	0.094152	-1.889724	-0.526929	-0.908987	-0.915771	0.659281
MIN	0.050043	0.290969	0.430208	0.120231	0.417204	-0.178531	-0.469262	0.582384	-0.482499	...	-0.224271	0.939403	0.466473	0.421557	0.222095	0.545259	-0.471770	-0.480473	0.096406	1.514892
NOP	-0.919862	-1.512872	-0.524097	-0.757467	0.123082	-0.438475	-0.855940	1.029291	-0.651037	...	0.365330	0.706472	0.784476	-0.601213	-0.056396	0.941494	-0.568346	-0.357011	1.050784	0.749899
OKC	0.206796	-0.013745	-0.179206	1.371388	-1.625021	-0.188805	0.314640	-1.112316	0.924308	...	-0.600648	-1.405842	0.864240	-1.705063	-1.452013	-1.391285	0.176588	0.418453	-1.036487	0.398655
PHX	0.135123	-0.692083	0.759917	0.962122	-0.047022	-0.529486	-0.630038	-0.384297	-0.559107	...	0.632570	0.868605	0.075842	1.645396	-0.395610	-0.965693	0.756268	-0.762480	0.629772	2.265983
POR	-1.289350	1.328165	-1.911645	-0.695887	-1.770285	0.418857	0.862016	-0.791822	0.206976	...	-2.255284	-0.495633	0.699421	0.580465	-1.021877	0.513612	-1.551395	0.387212	-1.514455	0.319131
SAC	2.244576	1.149121	1.305742	0.110789	1.477602	1.360072	1.105573	1.195371	1.279492	...	0.584745	0.574721	-0.412940	-1.383527	1.758314	-0.971113	-0.219966	1.011662	1.404661	-0.447377
SAS	-0.971180	1.581965	-0.130164	1.678275	-1.836326	-0.113949	0.188091	-0.594763	-1.741064	...	-0.643358	1.030992	-0.245634	-0.969604	-1.605855	0.526559	0.054964	-1.830444	-0.205313	0.046094
UTA	0.200918	1.024113	0.160895	0.753836	-0.561304	-0.336758	0.199691	-1.284589	0.452122	...	1.462015	0.002735	-2.699045	2.510059	-0.667334	1.816028	1.310585	0.131965	0.482748	-0.537995
ATL	1.574180	1.460141	1.902231	1.722791	0.592835	-0.767846	-1.016135	0.188667	0.732290	...	1.130995	-1.506107	-0.454541	-0.013300	-0.090496	-1.255739	1.785953	0.149549	0.862610	-0.680604
BKN	-0.743515	-0.352914	-1.568351	-0.569085	-1.460571	0.597192	0.841929	0.025470	0.610910	...	-1.460255	0.211944	0.266007	0.542966	-0.699275	-0.913442	-0.989575	0.625386	-1.273448	-0.473440
BOS	1.091793	-0.539803	0.823305	0.817641	0.150647	2.138364	2.165673	1.089851	-1.040645	...	0.355872	0.267038	-0.399827	0.520518	1.039095	-1.044280	-0.612814	-1.088549	1.427016	-0.508381
CHI	-0.802298	-1.515779	-0.083403	-0.979842	1.017142	-0.928568	-1.372682	0.498018	-0.726253	...	0.099990	-0.350604	1.163357	0.398547	0.383041	0.153465	-0.714925	-0.479792	0.640126	-0.605758
CHO	-1.103066	-0.414933	-0.840247	-0.445858	-0.612557	-0.870774	-0.844339	-0.444325	-0.099458	...	-0.418325	0.430100	-0.139937	0.666676	-0.870981	0.909775	-1.117647	-0.063626	-0.041026	-0.352750
CLE	-0.420729	-1.477527	0.011140	-0.481119	0.495662	-0.690092	-0.678882	-0.687095	-0.292994	...	-1.292197	-1.010297	0.872986	0.582118	0.059031	-1.418820	0.087923	-0.189198	-1.544530	0.882449
DET	-1.889767	0.011451	-1.314472	-0.508180	-1.185391	-0.997921	-0.867540	-1.193750	-0.834897	...	-0.660114	-0.345089	-0.714431	-0.628189	-1.311841	1.430870	0.546411	-0.564405	-1.789398	0.908680
IND	1.030725	2.111392	1.346459	0.675934	0.931080	0.298975	0.153290	0.378276	-0.421212	...	-0.904062	0.632324	-0.087710	0.870497	0.724621	-0.363511	-0.064282	-0.468296	-1.504150	0.785184
MIA	-1.258559	-0.286003	-2.187649	-2.061030	-0.739471	-0.307696	-0.231731	0.078121	1.410820	...	-0.756439	0.191284	-0.439619	-1.129464	-0.454950	0.807066	-0.429713	1.800345	-0.692435	-0.703707
MIL	1.748660	0.414485	1.525861	1.025381	0.900069	1.775088	1.745851	0.919226	-0.537219	...	0.660164	0.385162	-1.976576	-0.270217	1.392892	-0.349486	-0.402018	-0.595558	0.450429	-1.541720
NYK	0.858298	-0.553508	0.371662	-0.321215	0.864752	0.629557	0.663705	0.565001	0.613000	...	1.969340	-2.061290	-0.698763	-0.349434	1.004355	-0.481324	0.930921	0.523146	2.125381	-0.193969
ORL	-0.416948	-0.204787	-0.105072	0.417001	-0.538477	-1.034248	-0.728888	-1.246019	0.228864	...	0.275687	-0.675387	0.415905	0.583320	-0.964332	0.638762	-0.038894	0.078892	1.127248	0.522858
PHI	0.401458	-1.127736	-1.025813	-2.228125	1.019168	0.074534	-0.588627	1.730553	2.578502	...	0.654801	0.287079	0.392549	0.730272	1.060594	-0.191126	-0.251831	2.938043	0.533211	0.454005
TOR	-0.639667	-1.139786	0.266279	0.971962	-0.663811	-1.023350	-0.797938	-1.113004	-0.903845	...	0.756078	-0.583712	2.565406	1.218192	-1.117146	-1.011482	1.855039	-1.007856	0.669634	0.921912
WAS	-0.432566	-0.052507	0.417481	-0.670081	1.386201	-0.544615	-0.780285	0.273613	-1.158651	...	0.228666	-0.117368	-0.447620	-0.309686	0.842645	0.778171	-0.204066	-0.786094	-0.533795	-1.457892

根据以上数据，**我们采用层次聚类的方法**，将不同距离程度的样本根据层次划分为树状结构。选用的距离为常见的 **Euclid 距离**（`metric='euclidean'`），计算点与类/类与类之间距离用的是最长距离法（`method='complete'`），层次聚类的结果如下图所示



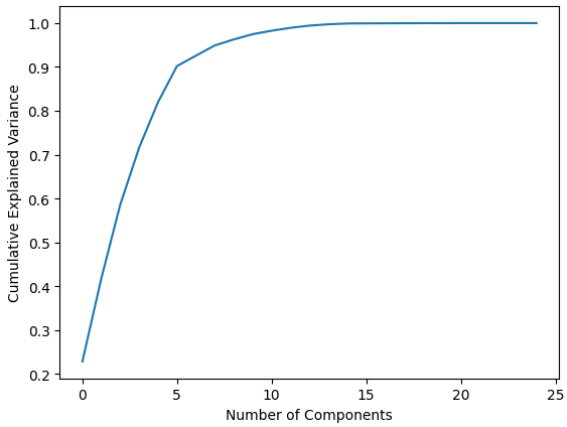
可以看到球队被分为 6 类，以下是我们对聚类结果的分析：

- 橙色：拥有持球大核、进攻能力较强防守能力较弱、运动能力普遍强的**攻强守弱型球队**
- 绿色：禁区护框与得分能力出众、造取犯规能力十分强的**阵地战防守罚球大队**
- 红色：运动天赋与能力出众、中距离投射稳定、三分能力较强、防守能力强大的**运动战攻防一体球队**
- 紫色：在本赛季进攻与防守方面都没什么建树、只是**争取选秀权的无欲无求队**
- 棕色：得分爆发性极为强大、传控能力出众、三分能力极为强大的**得分爆发型魔球球队**
- 粉色：拥有传切核心的控卫持球的**攻守平衡型球队**

3.2 主成分分析

将3.1中获得的球队类型结果合并入2.2的数据集后，我们基于这个新的数据集`dataset`先执行**主成分分析**。这是因为数据集`dataset`每行数据共有 25 个指标，数据维数过高可能在后续的模型训练出现**过拟合问题**。

首先，我们计算数据集各个主成分的方差占比，画出悬崖碎石图，如下图所示：



注意到悬崖碎石图的拐点处在 $n = 5$ 处，因此我们设置主成分个数为 5 个，并将数据集`dataset`转换为主成分，只保留前 5 个主成分为特征列。**仅保留主成分**的数据集`dataset`如图所示：

	PC1	PC2	PC3	PC4	PC5
0	-8.607509	0.493938	-6.866054	-15.663203	11.649705
1	17.519192	38.574854	1.632111	0.570833	16.742645
2	0.119878	5.002896	12.318611	2.997708	18.156003
3	-2.985700	-0.686626	11.117947	-18.908414	-2.273564
4	0.491177	-5.513319	-14.930629	-2.177799	16.353758
...	...	...	...	...	...
2454	-8.124001	-6.291519	-0.444290	-8.734699	-0.680837
2455	-6.452315	-13.895971	10.169374	14.496563	-7.266748
2456	5.186828	-22.794387	6.058760	14.924989	-5.588972
2457	-32.207033	24.554553	-10.582861	-7.052683	2.879501
2458	18.915332	-14.026568	-9.817424	-6.795701	13.083532

3.3 模型 2：随机森林模型的训练 [4]

训练的第一个模型是**随机森林模型**，随机森林是包含多个决策树的分类器，在 NBA 季后赛结果预测的主题下，每棵树模型输出的是赢（0）或输（1）的结果。

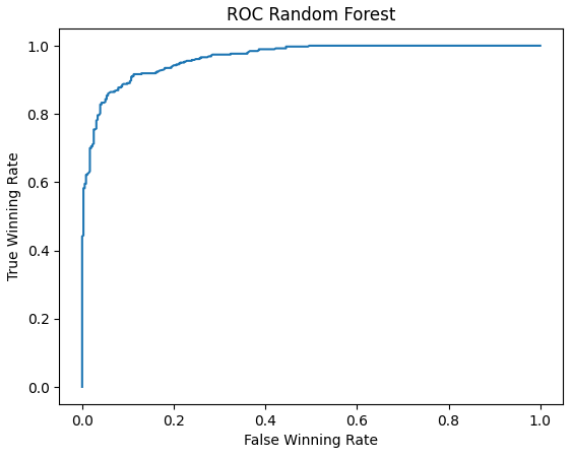
随机森林模型训练过程如下：

- 对 `dataset` 中的 2459 个样本有放回地随机抽取 2459 个、对 5 个主成分特征随机采样 m 个 $(m < 5)$，从而对每棵树 i 都得到一个数据集：$Z = \{(\boldsymbol{y_i}, \boldsymbol{X})\}$
- 其中 $\boldsymbol{X} = (\boldsymbol{PC_1}, \boldsymbol{PC_2}, ..., \boldsymbol{PC_m})$，矩阵 $\boldsymbol{X}$ 的每个列向量 $\boldsymbol{PC_i}$ 都是训练这棵树 i 时随机采样的特征，列向量 $\boldsymbol{y_i}$ 是输赢结果，元素为赢（记为 1）或输（记为 0）。
- 通过对每一种情况都建立一个决策树模型，将 $\boldsymbol{y_i}$ 和 $\boldsymbol{X}$ 输入模型后得到决策树 i 的投票结果
- **综合多个决策树，以决策树的投票结果作为两支球队对垒的输赢情况。**

随机森林模型训练完成后，我们在测试集上测试模型效果。与预测能力相关的评价指标如下表所示：

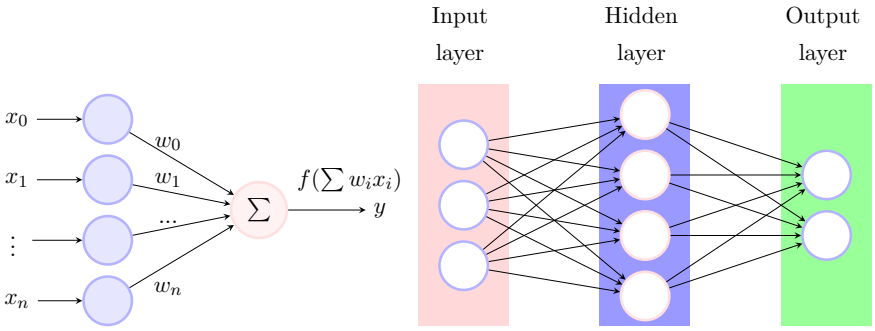
	precision	recall	f1-score	support
0	0.93	0.93	0.92	364
1	0.93	0.93	0.92	374
accuracy			0.93	738
macro avg	0.93	0.93	0.93	738
weighted avg	0.93	0.93	0.93	738

最后绘制出 ROC 曲线：



3.4 模型 3：神经网络模型的训练 [5]

我们希望基于主成分分析后的`dataset`，模拟一个理性人对球队对垒输赢情况的判断，故而构建了**神经网络模型**。在有了多个感知机的线性汇总、并选定了合适的激活函数的情况下，可以把多个人工神经元按照一定的层次结构连接起来，得到一个预测 NBA 季后赛输赢结果的人工神经网络 (ANN)。具体而言，我们选择的是前馈神经网络，并设置合理数量的隐藏层训练一个**多层感知机（MLP）**。

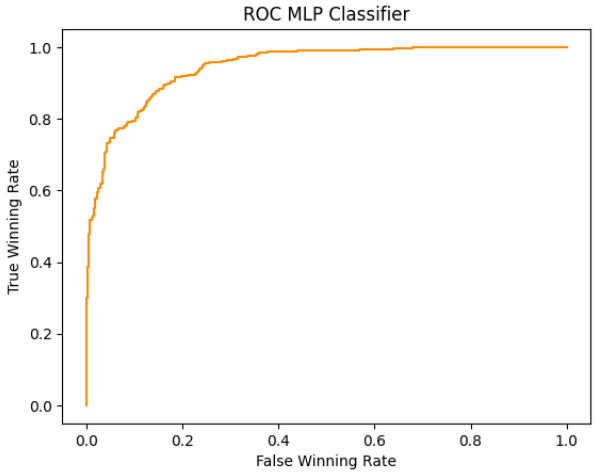


多层感知机神经网络模型训练过程如下：

- 首先划分训练集和测试集，选取 30% 的数据作为训练集
- 建立神经网络模型，经过调参，选择了 2 个隐藏层，每个隐藏层包含 10 个神经元；
- 在训练集上以之前得到的主成分指标 PC1、PC2、PC3、PC4、PC5 训练多层感知机模型；
- 在测试集上测试模型效果。与预测能力相关的评价指标如下表所示：

	precision	recall	f1-score	support
0	0.90	0.94	0.92	365
1	0.94	0.90	0.92	373
accuracy			0.92	738
macro avg	0.92	0.92	0.92	738
weighted avg	0.92	0.92	0.92	738

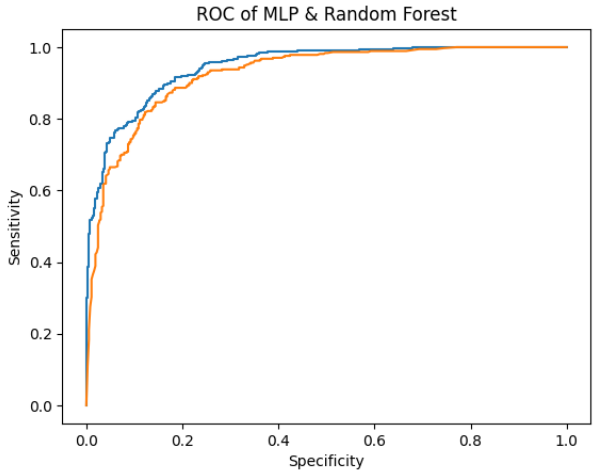
随后，我们绘制出 ROC 曲线，如下图所示：



ROC 曲线下的面积 AUC 很大，且没有明显过拟合现象，说明分类效果较好。综合上表和图可以看出，神经网络模型在测试集上的结果还是比较理想的。

3.5 模型的比对与选取

随机森林模型与神经网络模型在测试集上都具有良好的表现。我们将随机森林模型和神经网络模型的 ROC 曲线叠放来比较两个模型的优劣。如下图所示，蓝线为随机森林模型（RF）的 ROC 曲线，橙线为神经网络模型（MLP）的 ROC 曲线



注意到蓝线（RF）的 ROC 曲线在橙线（MLP）之外，说明**相同的灵敏度下随机森林模型的特异性有微弱的优势、相同特异性下随机森林模型的灵敏度同样有微弱的优势**，因此通过对模型的综合评估与比较，我们**选定随机森林模型**进行后续的预测。

4 预测效果分析与课题结论

4.1 预测效果分析

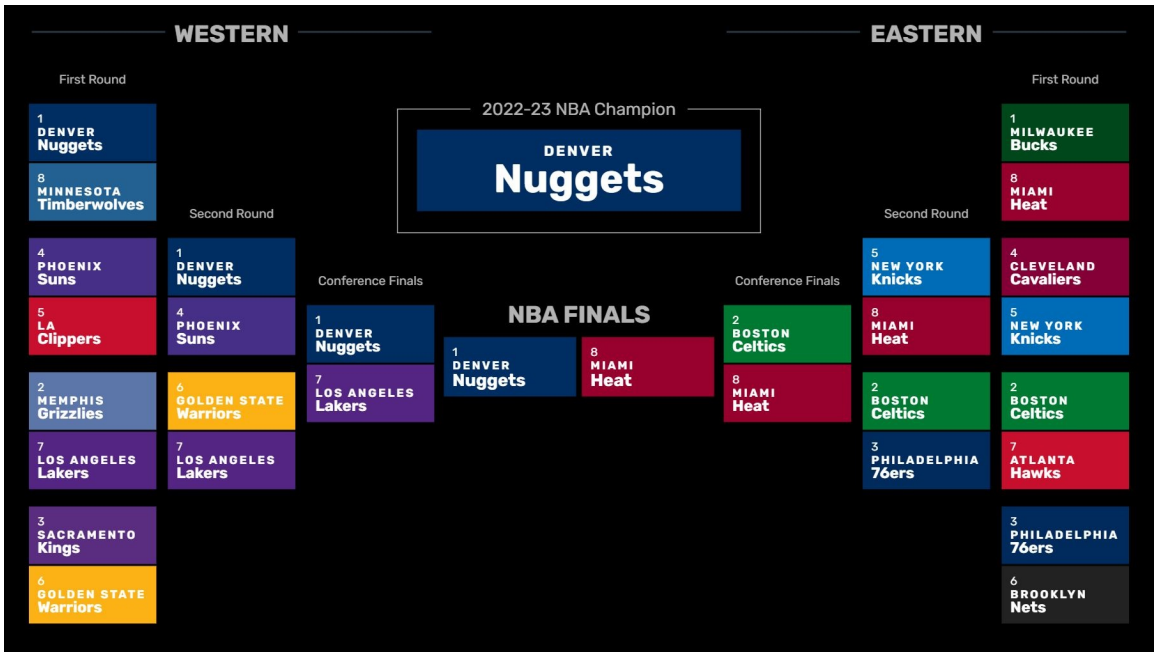
预测分为两个部分：

- 第一部分：根据现实中 NBA 季后赛的对阵情况，对每轮对阵的胜负结果进行预测
- 第二部分：根据模型运行结果晋级淘汰，最终得出模型预测的总冠军

对于第一部分，将对阵双方的数据进行主成分分析后输入到随机森林模型后，可得双方输赢概率。由于对垒双方 i, j 各有一组数据向量 $\boldsymbol{\Delta\alpha}_{i\to j}, \boldsymbol{\Delta\alpha}_{j\to i}$（如 GSW 对战 SAC 的比赛会生成 GSW 相对 SAC 的数据向量 $\boldsymbol{\Delta\alpha}_{GSW\to SAC}$，以及 SAC 相对 GSW 的数据向量 $\boldsymbol{\Delta\alpha}_{SAC\to GSW}$），从而对应生成两组输赢概率的向量 $\boldsymbol{p_1}, \boldsymbol{p_2}$，我们取 $\boldsymbol{p_1}$ 向量中队伍 i 赢的概率与 $\boldsymbol{p_2}$ 向量中队伍 i 赢的概率的平均值作为队伍 i 的最终获胜概率。季后赛所有比赛各队伍的获胜情况如下：

```
[[0.6555, 0.34450000000000003], GSW vs SAC, SAC win 65%
[0.532, 0.46799999999999997], LAL vs MEM, MEM win 53.2%
[0.5415, 0.4585], PHX vs LAC, PHX win 54.1%
[0.38449999999999995, 0.6155], DEN vs MIN, DEN win 61.6%
[0.16, 0.84], BOS vs ATL, BOS win 84%
[0.844, 0.15600000000000003], CLE vs NYK, NYK wins 84.4%
[0.020499999999999997, 0.9795], MIL vs MIA, MIL wins 98.0%
[0.1505, 0.8495], PHI vs BKN, PHI wins 85.0%
[0.615, 0.385], GSW vs LAL, LAL wins 61.5%
[0.4515, 0.5485], DEN vs PHX, DEN wins 54.9%
[0.726, 0.274], PHI vs BOS, BOS wins 72.6%
[0.973, 0.027000000000000024], MIA vs NYK, NYK wins 97.3%
[0.582,0.41800000000000004], LAL vs DEN, DEN wins 58.2%
[0.0375, 0.9625],BOS vs MIA, BOS 96.3%
[0.022, 0.978]], DEN vs MIA, DEN 97.8%
```

2022-2023 赛季 NBA 真实的晋级情况如下：



对比容易得知，模型做出完全正确预测的比赛如下：

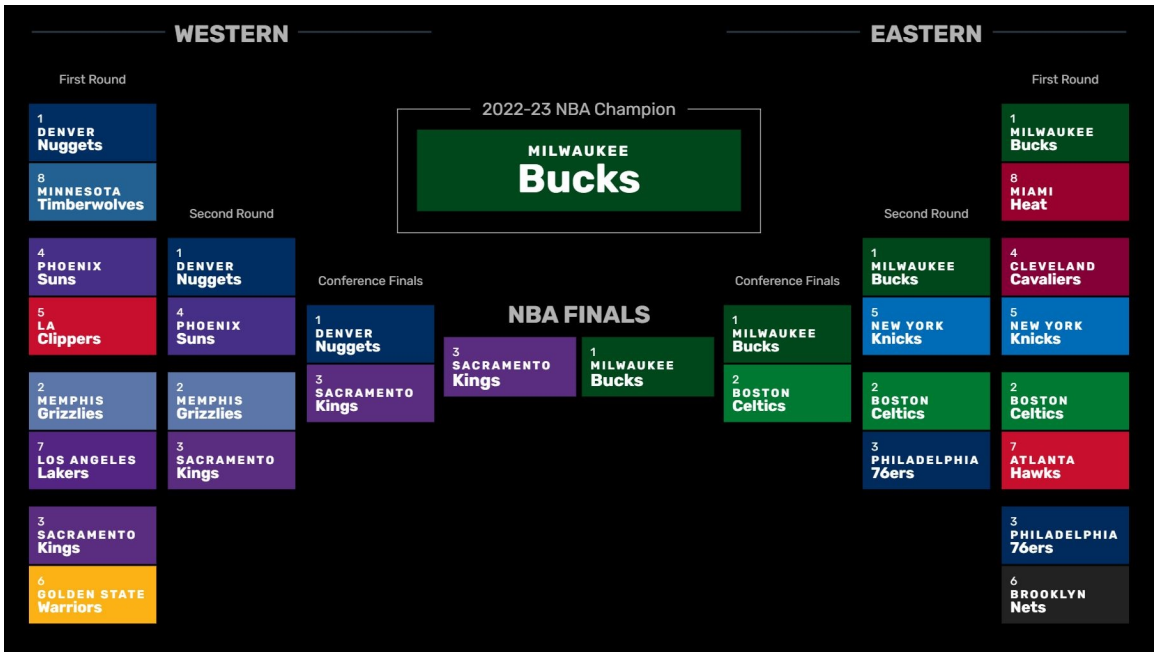
- 西部：（第一轮）DEN vs MIN ， PHX vs LAC；（第二轮）DEN vs PHX，GSW vs LAL，（西决）DEN vs LAL **准确率：5/7**
- 东部：（第一轮）CLE vs NYK，BOS vs ATL，PHI vs BKN；（第二轮）BOS vs PHI，**准确率：4/7**
- 总决赛：DEN vs MIA，**准确率：1**

如果依据第一部分的胜负预测数据给出我们的赔率，赔率对真实结果的准确率将达到 $\frac{10}{15}=\frac{2}{3}$，超过按照 ESPN 给出赔率对真实晋级结果的准确率 $\frac{9}{15}$。**可见在篮球彩票的赔率设定上，我们的模型有着较好的表现。**

对于第二部分，我们根据上述概率得出模型的预测晋级情况：

```
[[0.6555, 0.34450000000000003], GSW vs SAC, SAC win 65%
[0.532, 0.46799999999999997], LAL vs MEM, MEM win 53.2%
[0.5415, 0.4585], PHX vs LAC, PHX win 54.1%
[0.38449999999999995, 0.6155], DEN vs MIN, DEN win 61.6%
[0.16, 0.84], BOS vs ATL, BOS win 84%
[0.844, 0.15600000000000003], CLE vs NYK, NYK wins 84.4%
[0.020499999999999997, 0.9795], MIL vs MIA, MIL wins 98.0%
[0.1505, 0.8495], PHI vs BKN, PHI wins 85.0%
[0.6595, 0.3405], MEM vs SAC, SAC wins 66.0%
[0.43200000000000005, 0.568],DEN vs PHX, DEN wins 56.8%
```

```
[0.242, 0.758], MIL vs NYK, MIL wins 75.8%
[0.29550000000000004, 0.7044999999999999], BOS vs PHI, BOS wins 70.4%
[0.8554999999999999, 0.14450000000000007], DEN vs SAC, SAC wins 85.5%
[0.48150000000000004, 0.5185], MIL vs BOS, MIL wins 51.9%
[0.4535, 0.5465]] MIL vs SAC, MIL wins 54.7%
```



对比容易得知，模型做出完全正确预测的比赛如下：

- 西部：（第一轮）DEN vs MIN ， PHX vs LAC；（第二轮）DEN vs PHX，**准确率：3/7**
- 东部：（第一轮）CLE vs NYK，BOS vs ATL，PHI vs BKN；（第二轮）BOS vs PHI，**准确率：4/7**

最终模型预测的总冠军为 MIL（雄鹿），但在现实情况中，MIL 在首轮对阵中就被 MIA 淘汰，我们对原因进行了分析：一方面，MIL 队主将在系列赛第一场受了伤，缺席了比赛，复出后依然受到伤病困扰，没能发挥出最高水平；另一方面，MIA 队全员发挥神勇，主将巴特勒超水平发挥，以东部第八的身份连续过关斩将进入总决赛，这是外界都没有预测到的。根据该轮对决赛前胜率，外界认为 MIA 战胜 MIL 的概率仅为 3%，因此，模型预测结果仍然与现实中人们的普遍预期**（反映在赔率上）**相吻合。

CLE 与 NYK 的对阵我们实现了精确的预测，**在与盘口赔率结果相反的情况下我们与实际结果实现了意外的匹配**。这也体现出来了我们对各个球队的各项数据分析是比较到位的，我们成功预测出了隐藏在偶然性中的某些必然性。（NYK 的数据经合理建模后表现实际上强于 CLE，但基于常规赛排名以及大众对球队的印象，CLE 都优于 NYK）

4.2 模型的局限性与改进方案

4.2.1 模型的局限性反思

1. 数据的局限性：通过常规赛数据进行建模有不能更好的拟合季后赛球队表现的可能性，有些球队季后赛水平远超常规赛（如 MIA），但我们不能补充更有效的数据来衡量这些球队的实力（各队每赛季的都会出现较大的实力变化，不能挑选前几年各支球队的季后赛数据作为预测参考数据），他们的坚韧程度、教练季后赛职教水平（如热火队斯波教练）都远超常规赛，但这些能力无法精确量化。

2. 篮球比赛的非量化因素

- **1. 巨星效应**：SAC 对战 GSW 与 LAL 对战 MEM 这两轮比赛的现实结果与我们的预测结果产生了偏差。**最大的原因就是超级巨星给球队带来的效应无法被算法合理的量化**，但季后赛往往是超级巨星（例如：GSW 的 Steven Curry 和湖人队的 LeBron James）的舞台，他们在季后赛的超常发挥往往能打破他们在常规赛给人的印象，他们的能力决定着这两支球队的命运。
- **2. 吹罚尺度**：在季后赛裁判对某些球队存在明显的吹罚偏袒问题，直接影响到这轮系列赛的胜负，我们的数据也无法量化裁判的吹罚尺度。
- **3. 伤病问题**：常规赛中时常有当家球星陷入伤病麻烦的问题发生，对该球队在常规赛的成绩造成影响，导致在数据建模的球队在季后赛中表现的实力远不如现实季后赛中该球队的表现。相反，若在季后赛中出现当家球星陷入伤病麻烦但其常规赛十分健康的情况，由于我们的训练集以常规赛中后期数据为基础，其在数据建模模拟的季后赛中表现优异但在真实季后赛中十分羸弱。
  - GSW 的当家球星 Steven Curry 与 LAL 队的当家球星 LeBron James 在常规赛中缺席场次较多，这也是为何 GSW 与 LAL 在模型预测中第一轮被淘汰的原因

3. 存在一定的模型过拟合问题：尽管随机森林模型和神经网络模型在测试集和季后赛数据集下表现出了较好的预测能力，但由于数据量的限制（2459 条数据），仍然存在过拟合的风险。

4.2.2 改进方案

作为一个篮球彩票的数据分析团队，我们的任务是提供合理的赔率值，**更关注实力对比下的正常对阵情况，而非爆冷结果**。由此，可能的改进方案如下所列：

- **特征工程**：考虑增加更多与超级巨星、裁判吹罚和伤病相关的数据指标。例如，球员个人数据（得分、篮板、助攻等）可以作为超级巨星效应的指标；裁判数据（罚球次数、犯规次数等）可以用于量化裁判吹罚尺度；伤病数据（球员受伤次数、伤停时间等）可以用于估计球队的健康状况。
- **动态更新**：保持对数据的实时追踪和更新模型，及时获取最新数据并更新模型，以确保模型的准确性和可靠性。
- **引入其他预测模型**：如深度学习模型（如 LSTM 或 Transformer）或梯度提升树模型（如 XGBoost 或 LightGBM）。这些模型能够更好地捕捉复杂的非线性关系。

- **数据集扩大与交叉验证**：扩大样本数据范围，采用过往几年的季后赛数据作为训练集样本，让我们的模型**更加拥有 NBA 季后赛的样本特征（坚韧程度、防守强度较常规赛大幅度提升）**。同时执行交叉验证，使得最终预测结果趋于稳定。

5 参考文献与相关网站

[1] https://www.basketball-reference.com/

[2] https://github.com/Matheuskempa/My_Udacity_Capstone

[3] Shkedy Z, Sengupta R, Perualila N J. Identification of Local Patterns in the NBA Performance In-dicators[M]//Applied Biclustering Methods for Big and High-Dimensional Data Using R. Chapman and Hall/CRC, 2016: 323-344.

[4] Cheng G, Zhang Z, Kyebambe M N, et al. Predicting the outcome of NBA playoffs based on the maximum entropy principle[J]. Entropy, 2016, 18(12): 450.

[5] Khanmohammadi R, Saba-Sadiya S, Esfandiarpour S, et al. MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs[J]. arXiv preprint arXiv:2210.17060, 2022.