****

基于机器学习实现2022年美国大学生篮球比赛预测

专 业： 计算机科学与技术

队 名： LKWY\_Epiphany队

队 长： 李庆森

姓 名： 孔钰瑶 王浩 杨晓达

指导教师： 孙仁诚

2022年 12月

摘 要

随着现在竞技体育的流行与普及，本文主要是通过历年各个队伍的比赛数据，预测2022年美国男子大学生篮球锦标赛结果，综合利用NCAA历史比赛的数据以及其他数据来源，根据比赛的机制与规则来选择各个队伍之间比赛的赢家与输家。

基于此问题，我们首先分析过去几年各个场次比赛统计信息，理解并总结各个文件以及相关参数的含义，提取有效的文件，来为模型的建立提供数据支持。

其次，进行数据的预处理，之后对数据进行了可视化操作，通过绘制各个数据之间的散点图，直观地感受数据之间相关性， 同时验证并处理数据中的空值，无效值或异常值，之后根据数据之间的规律以及篮球比赛的规则及特性进行特征工程，从数据统计信息中，进行数据加工处理，提取出衡量队伍各方面能力的参数，比如计算命中率，失误率等，使数据能够更好地进行预测，并将多个相关文件进行合并，最后进行模型建立，我们建立LightGBM模型和随机森林模型等，并进行参数调整，绘制各个特征的贡献率，剔除贡献率低的特征，不断优化训练模型。

关键词：NCAA比赛 特征工程 LightGBM模型 随机森林模型

目录

[摘 要 1](#_Toc122376503)

[第一章 绪论 3](#_Toc122376504)

[1.1 论文研究背景与意义 3](#_Toc122376505)

[1.2 研究问题以及难点 4](#_Toc122376506)

[1.2.1问题描述 4](#_Toc122376507)

[1.2.2问题背景 4](#_Toc122376508)

[1.2.3 问题所在难点 5](#_Toc122376509)

[第二章 相关研究与技术 6](#_Toc122376510)

[2.1 随机森林模型 6](#_Toc122376511)

[2.1.1随机森林基本原理 6](#_Toc122376512)

[2.1.2随机森林算法 6](#_Toc122376513)

[2.2 XGBoost算法 7](#_Toc122376514)

[2.2.1原理 7](#_Toc122376515)

[2.3 LightGBM提升学习模型 8](#_Toc122376516)

[2.3.1 原理 8](#_Toc122376517)

[第三章 数据探索与问题求解 9](#_Toc122376518)

[3.1 数据说明 9](#_Toc122376519)

[3.1.1 数据的来源 9](#_Toc122376520)

[3.1.2 主要数据说明 9](#_Toc122376521)

[3.2 数据的处理 11](#_Toc122376522)

[3.2.1 数据预处理 11](#_Toc122376523)

[3.2.2 构建特征工程 11](#_Toc122376524)

[3.3模型建立 14](#_Toc122376525)

[3.5参数调优 16](#_Toc122376526)

[3.5.1关于ligntGBM的参数调优 16](#_Toc122376527)

[第四章 总结与展望 19](#_Toc122376528)

[4.1总结 19](#_Toc122376529)

[4.2展望 19](#_Toc122376530)

# 第一章 绪论

## 1.1 论文研究背景与意义

竞技体育是受关注度和普及度最高的项目之一，其最大的特点就是不可预测性.因为高强度，快节奏的比赛，往往在一分钟其至一秒钟的时问内，比赛的结果都有可能不同，影响胜负的因素很多，虽然说实力因素是主导因素，但是临场的发挥,心理因素，运气成分，一些突发的情况等也是决定比赛胜负的重要因素，而这些因素也是随机，不确定的，导致了竞技比赛的赛果难以预测。

近年来，体育竞技的分析已经演变成为一个重要的领域。球队的老板每年在各自球队身上花费成千上万的钱财，希望避免在那些对球队实力没有帮助的球员身上耗费资金。分析师花费大量的时间试图预测每年哪一支球队将赢得冠军，从而在商业上获取更高的经济效益，球迷也想在整个赛季赢得比赛竞猜的胜利。使用统计学竞技体育已经从比赛转变成为一门学科，研究者使用大量的概率统计模型去对比赛进行分析，能够比较准确的预测出系列比赛之间的胜负，赛季的冠军，由于体育竞技多为团队之间的对抗，队员之间的配合也十分的重要，将队员的个人数据统计因素考虑进去，可以更加准确的分析出队员对胜负的影响，对球队的页献。

## 1.2 研究问题以及难点

### 1.2.1问题描述

本次的项目题目选取于<<March Machine Learning Mania 2022 - Men’s Predict the 2022 College Men’s Basketball Tournament>>



图1.1 kaggle网站竞赛标题

主要是通过历年各个队伍的比赛数据，预测2022年美国男子大学生篮球锦标赛结果，综合利用NCAA历史比赛的数据以及其他数据来源，根据比赛的机制与规则来选择各个队伍之间比赛的赢家与输家。

此次项目主要分为两个阶段：

·在比赛的第一阶段，主要为模型构建阶段：我们需要根据历史比赛数据，构建和测试模型

·在比赛的第二阶段：我们将利用阶段一所构建的模型来预测2022年锦标赛的结果。

### 1.2.2问题背景

在进行问题求解前我们需要简单的了解比赛的相关规则：

NCAA有着悠久的历史，同时又有着一些很“独到”的传统。无论是球队常规赛分区、季候赛球队的选拔分区方式，还是季后赛残忍的单场淘汰制和那明显有别于FIBA、NBA(微博)的比赛规则都看起来很是有自己的传统特点。NCAA“疯狂”的赛制如下：

NCAA的篮球赛季开始于每年的11月下旬，通常在3月初左右结束常规赛季，然后进行一个月的“疯狂三月”。在常规赛季中，每间院校都会与各自联盟内的成员交战两次，主客场各一；每间院校也会选择与联盟外的一些学校交战，来提升自己的实力并为自己宣传造势。

当常规赛季结束，一般来说，绝大多数联盟都会取该联盟的前八名球队来进行单淘汰赛制的联盟冠军战，而由联盟冠军战中决出来的冠军，基本上都可以自动晋级NCAA锦标赛。除掉各联盟的冠军自动晋级外，NCAA锦标赛的剩余名额由委员会来决定，称为扩大邀请。邀请资格基本上由战绩、联盟强弱程度等方面来决定。

等64强名单确定之后，所有入围的球队将分为东、西、南和中西四个分区，公认的四强为各区第一种子，其它各校则由委员会按实力强弱排为2-16号种子。此后采取首尾捉对法来确定对阵（类似于NBA季后赛的对阵排列方法，如第1对第16，第2对第15，以此类推），采取单场淘汰赛制而进行决斗——输了一场就回家，淘汰速度之快非常惊人，这也是“疯狂三月”的真谛。

NCAA锦标赛分为三个阶段，第一阶段是第一和第二轮，从64强迅速淘汰到16强。16强也被称为甜蜜16（Sweet Sixteen），所有的16强球队进入到第二阶段，都会获得更高一个级别的关注度，这也是有志于希望选秀的球员得以吸引更高曝光率和更多球探关注的舞台。第二阶段将从甜蜜16淘汰到精英八强（Elite Eight），然后淘汰选出最终四强（Final Four）。

最终，在事先选好的决战地点，最终四强将经过半决赛和决赛来决出冠军。一场定胜负的赛制让比赛显得尤为激烈而残酷，却又有无与伦比的魅力。四强战是NCAA最大的舞台，对于那些并不是传统名校的球队来说，能够进入锦标赛就已经非常值得庆祝，而晋级四强战则更是天大的荣耀。

NCAA“特殊”的比赛规则：

1 参赛球员的球衣号码可以从以下37个号码中选择：0-5、00、10-15、20-25、30-35、40-45以及50-55；0和00只能有一个。

2 比赛分为上下半场，各20分钟，中间休息15分钟（NCAA男子1级比赛，中场休息可延长至20分钟）。如果需要进行加时赛，则每个加时赛时常为5分钟，中间休息1分钟。

3 比赛中每次进攻时间为35秒（NBA、FIBA均为24秒），进攻方底线发球后10秒内必须将球推进至前场（NBA、FIBA则均为8秒）。

4 当一支球队的半场普通犯规累计至7次，每犯规一次（含第7次），对方将获得“1+1”罚球机会，即：如果第一球罚进，将再获得一次罚球机会。（普通犯规是指除恶意犯规、故意犯规、阻挡对方投篮动作犯规之外的个人犯规。

5 当一支球队的半场普通犯规次数累计至10次，每犯规一次（含第10次），对方将获得2次罚球机会。

### 1.2.3 问题所在难点

首先是对数据文件、属性变量和比赛机制的理解。

其次是对数据的预处理，数据量过大，总共有40个文件，每个文件的属性过多，如何有效选取合适的文件，并从多个属性中提取出能够衡量队伍各方面能力的参数，包括选择哪些指标以及新变量如何构造等，来进行预测模型的构建。

# 第二章 相关研究与技术

## 2.1 随机森林模型

### 2.1.1随机森林基本原理

随机森林由LeoBreiman（2001）提出，它通过自助法（bootstrap）重采样技术，从原始训练样本集N中有放回地重复随机抽取k个样本生成新的训练样本集合，然后根据自助样本集生成k个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。特征选择采用随机的方法去分裂每一个节点，然后比较不同情况下产生的误差。能够检测到的内在估计误差、分类能力和相关性决定选择特征的数目。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类。

### 2.1.2随机森林算法

决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

随机森林是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类，然后看看哪一类被选择最多，就预测这个样本为那一类。

在建立每一棵决策树的过程中，有两点需要注意采样与完全分裂。首先是两个随机采样的过程，random forest对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个，那么采样的样本也为N个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现over-fitting。然后进行列采样，从M个feature中，选择m个（m << M）。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。一般很多的决策树算法都一个重要的步骤——剪枝，但是这里不这样干，由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现over-fitting。

决策树中分裂属性的两个选择度量：

1）信息增益

随机森林模型任意样本分类的期望信息：

* I(s1,s2,……,sm)=∑Pi log2(pi)(i=1..m) 其中，数据集为S，m为S的分类数目，Pi≈|Si/|S|，Ci为某分类标号，Pi为任意样本属于Ci的概率，si为分类Ci上的样本数
* I(s1,s2,……,sm)越小，s1,s2,……,sm就越有序（越纯），分类效果就越好。
* 由属性A划分为子集的熵：
* A为属性，具有V个不同的取值， S被A 划分为V 个子集s1,s2,……,sv，sij是子集sj中类Ci的样本数。E(A)= ∑(s1j+ ……+smj)/s \* I(s1j,……,smj)
* 信息增益：Gain(A)= I(s1,s2,……,sm)
* 分裂属性选择规则：选择具有最大信息增益的属性为分裂属性

2）基尼指数

* 集合T包含N个类别的记录，那么其Gini指标就是pj 类别j出现的频率
* 如果集合T分成m部分 N1 , N2 ,…, Nm 。那么这个分割的Gini就是
* 分裂属性选择规则：选择具有最小Ginisplit的属性为分裂属性（对于每个属性都要遍历所有可能的分割方法）。

## 2.2 XGBoost算法

### 2.2.1原理

XGBoost属于集成学习Boosting，是在GBDT的基础上对Boosting算法进行的改进，并加入了模型复杂度的正则项。GBDT是用模型在数据上的负梯度作为残差的近似值，从而拟合残差。XGBoost也是拟合数据残差，并用泰勒展开式对模型损失残差的近似，同时在损失函数上添加了正则化项。



L1正则化项：



L2正则化项：



## 2.3 LightGBM提升学习模型

### 2.3.1 原理

lightGBM，它是微软出的新的boosting框架，基本原理与XGBoost一样，使用基于学习算法的决策树，只是在框架上做了一优化（重点在模型的训练速度的优化）。直方图算法的基本思想是先把连续的浮点特征值离散化成k个整数，同时构造一个宽度为k的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

# 第三章 数据探索与问题求解

## 3.1 数据说明

### 3.1.1 数据的来源

来自于kaggle链接如下:

[March Machine Learning Mania 2022 - Men’s | Kaggle](https://www.kaggle.com/competitions/mens-march-mania-2022/data)

### 3.1.2 主要数据说明

此平台提供了40个文件，每个文件都有数个属性，且不同文件之间存在共同的标签和属性。

图示, 文本

描述已自动生成

图3.1 数据信息分析图

本次主要使用的文件的说明如下：

***MNCAATourneySeeds.csv***

* Season：比赛年份
* Seed：第一个字符（W、X、Y、Z）表示团队所在的区域后面两个数字（01、02、...、15 或 16）表示所在区域的种子

***MNCAATourneySeeds.csv***

* Season：赛事举办的年份
* Seed：由3-4个标识符组成，其中第一个字符要么是W,X,Y或Z（标识团队在的区域）和接下来的两位数字（01，02，…，15或16）表示该区域内的种子
* TeamID：标识团队的id号

***MNCAATourneySlots.csv***

该文件标识了在锦标赛进行时，根据种子队相互配对的机制。

* Season：决赛年份
* Slot：对于入围游戏，它是一个三字符字符串，标识获胜团队完成的种子，例如 W16 或 Z13。对于常规锦标赛游戏，它是一个四字符字符串，其中前两个字符告诉您游戏是哪一轮（R1、R2、R3、R4、R5 或 R6），后两个字符告诉您受青睐球队的预期种子。
* StrongSeed：这表示在这场比赛中参加比赛的预期更强种子球队。

***MNCAATourneyDetailedResults.csv***

* Season：季节
* WTeamID：这标识了赢得比赛的球队的ID号
* WScore：这标识了获胜团队的得分数
* LTeamID：这标识了输掉比赛的球队的ID号
* LScore：这标识了失败的球队获得的分数
* WLoc：这标识了获胜团队的“位置”。
* NumOT：这表示游戏中的加班次数，整数 0 或更高
* WFGM：现场进球（由获胜团队）
* WFGA：尝试的现场进球（由获胜团队）
* WFGM3：三分球（由获胜团队制作）
* WFGA3：三分球尝试（由获胜团队）
* WFTM：罚球（由获胜球队）
* WFTA：罚球尝试（由获胜球队）
* WOR：进攻篮板（由获胜球队拉动）
* WDR：防守篮板（由获胜球队拉动）
* WAst：（由获胜球队提供）

## 3.2 数据的处理

### 3.2.1 数据预处理

1）验证数据中时候参在空值，无效值或异常值，结果发现原始数据中不存在以上问题数据。

2）为方便模型处理，转变主客场特征的数据类型

3）初步思路：

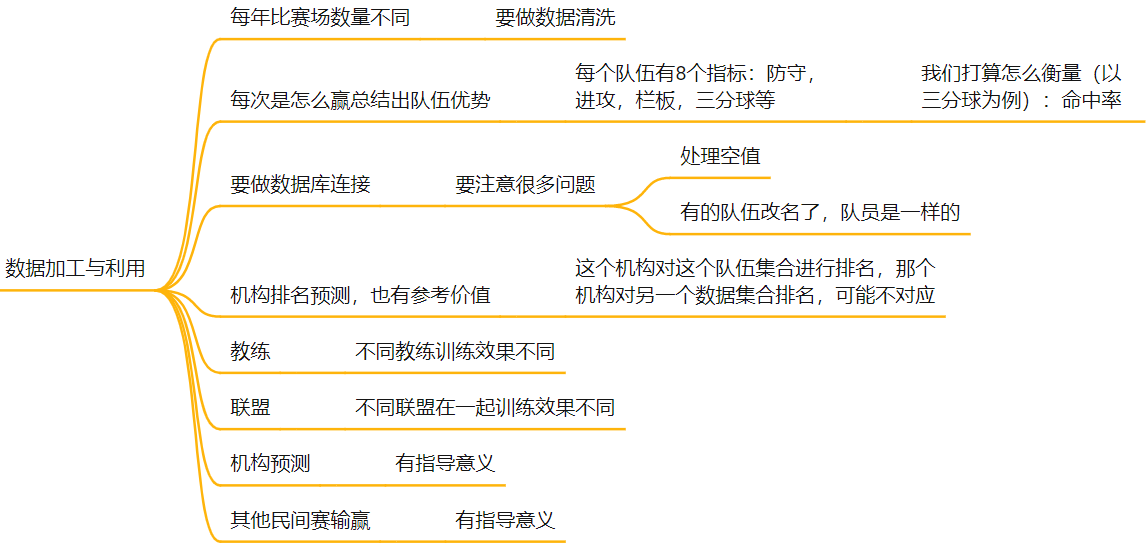


图3.2 数据处理初步思路

### 3.2.2 构建特征工程

1.原始数据特征及含义：

|  |  |
| --- | --- |
| 特征名 | 含义 |
| DayNum | 日期 |
| Score | 得分 |
| loc | 主客场 |
| FGM | 2分球  命中数 |
| FGM3 | 3分球  投篮出手数 |
| FGA | 2分球  投篮出手数 |
| FGA3 | 3分球  投篮出手数 |
| FTM | 罚球数 |
| OR | 进攻篮板 |
| DR | 防守篮板 |
| Ast | 助攻 |
| TO | 失误 |
| Stl | 抄截 |
| Blk | 盖帽 |
| PF | 犯规 |

2.构造新的特征特征。

命中率：

进攻失误率：

2分球命中率：

3分球命中率：

传球成功率：

分差：

3.进行聚合操作构造新的特征

|  |  |
| --- | --- |
| **特征名** | **含义** |
| WNum | 比赛场次 |
| AvgWpts | 平均分差 |
| TotalPoints | 总得分 |
| LDefense\_3pct | 防守三分球率 |

胜率：

场均得分：

胜场得分率：

3分球防守率：



4.对构造出的新的特征进行可视化

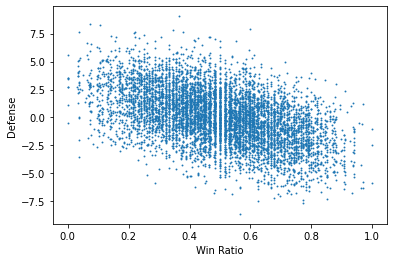


图3.3 Defense和Win Ratio相关性散点图

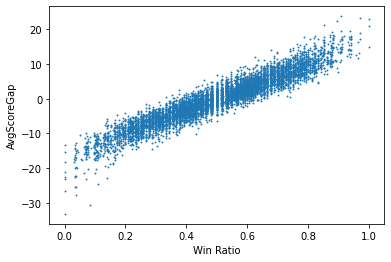


图3.4 AvgScoreGap和Win Ratio相关性散点图

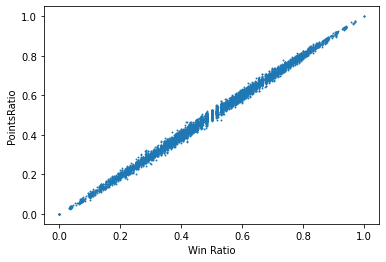


图3.5 PointsRatio和Win Ratio相关性散点图

5.最后进行数据合并，将所有构造的新特征合并为一个大表

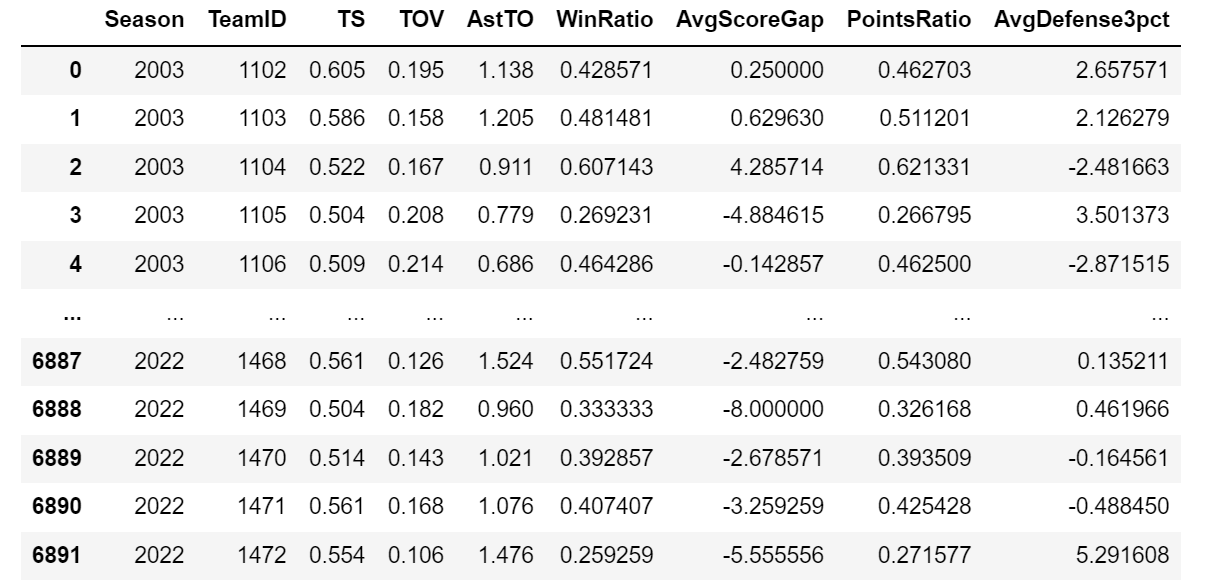


图3.6 训练特征数据总览图

## 3.3模型建立

1. 经过数据处理后，使用以下数据和标签建模：

features =["SeedA","SeedB",WinPercentageA',WinPercentageB',AvgScoringMarginA',AvgScoringMarginB',ComputerRankA",ComputerRankB]

2. 归一化用于产生规整数据

在数学建模过程中通常将数值过大的数据进行归一化处理，使得数据控制在[0,1]的范围内，作为模型的输入输出参数，计算方法如下：



有时候模型的输入拥有多个属性特征参数，归一化处理能消除量纲不同带来的数值大小差异，将不同类型的数据标准化处理。

3. 高斯化用于产生规整数据

4. 生成整合的数据样式如下

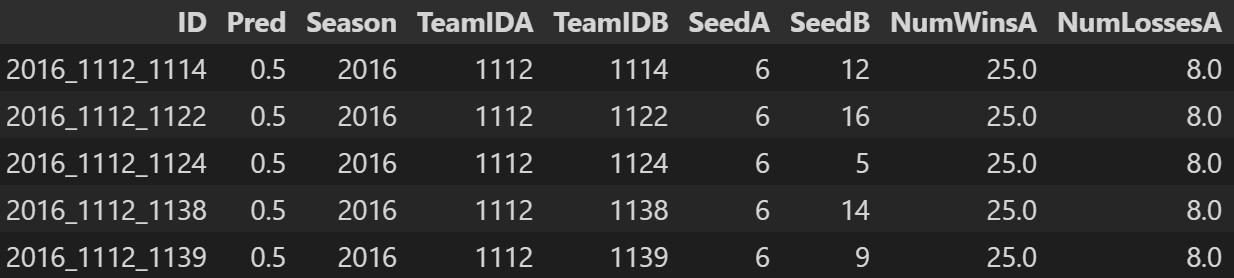


图3.7 整合数据样式图

使用lightGBM建模

lightgbmc\_model = LGBMClassifier(boosting\_type='gbdt', objective='binary',random\_state=42)

其中，GBDT是梯度提升树，表示使用基于树的模型进行计算，采用二分类模型，随机数设置为42

根据mode是否等于reg决定选用弹性回归网络还是XGB

1. if mode == "reg":
2. model = ElasticNet(alpha=1, l1\_ratio=0.5)*#弹性网络回归*
3. else:
4. model = xgb.XGBClassifier()

其中，弹性网络回归alpha=1, ratio=0.5

采用对数损失函数

绘制图像

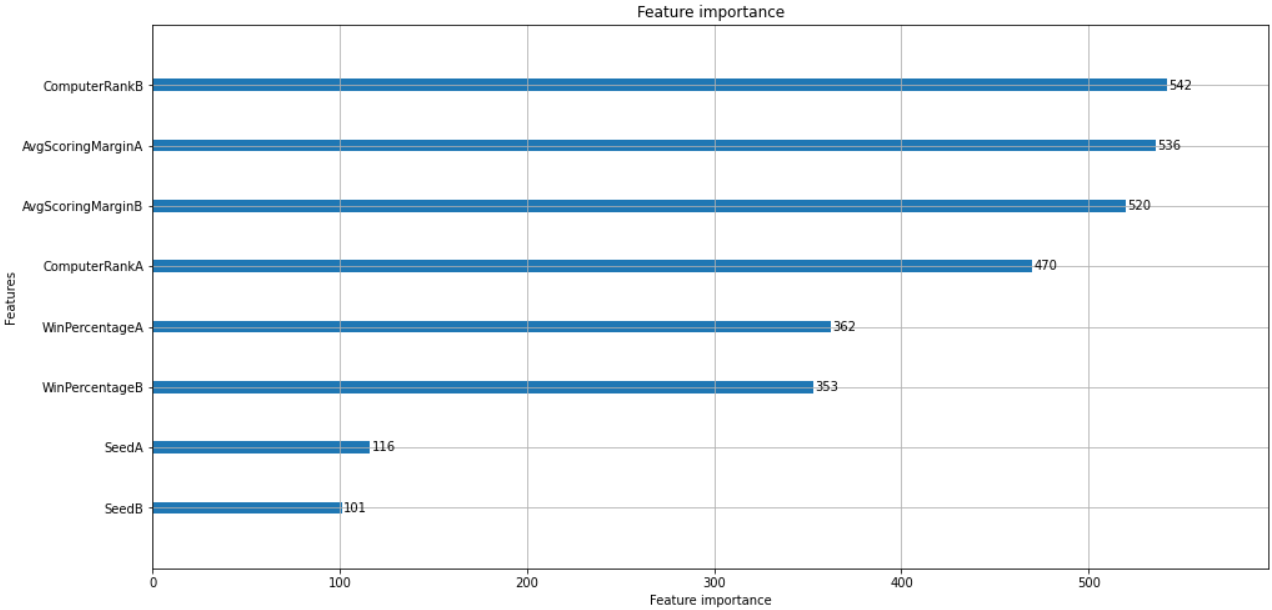
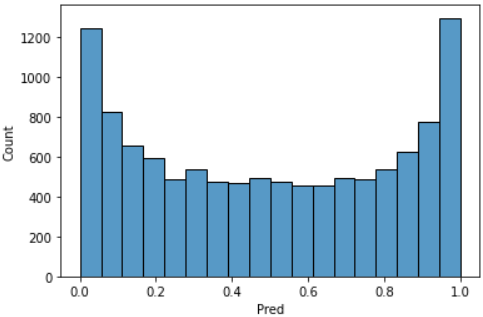


图3.8 数据特征贡献率图



？

## 3.5参数调优

### 3.5.1关于ligntGBM的参数调优

可调的参数有：

（1）boosting\_type

用于指定弱学习器的类型，默认值为 gbdt，表示使用基于树的模型进行计算。还可以选择为 gblinear 表示使用线性模型作为弱学习器

可选的参数值有：

* Gbdt：使用梯度提升树
* Rf：使用随机森林
* Goss：使用单边梯度抽样算法，速度很快，但是可能欠拟合。

（2）Objective

用于指定学习任务及相应的学习目标，常用的可选参数值如下：

* Regression:使用L2正则项的回归模型（默认值）。
* regression\_l1:使用L1正则项的回归模型。
* Mape:平均绝对百分比误差。
* Binary：二分类。
* Multiclass：多分类。

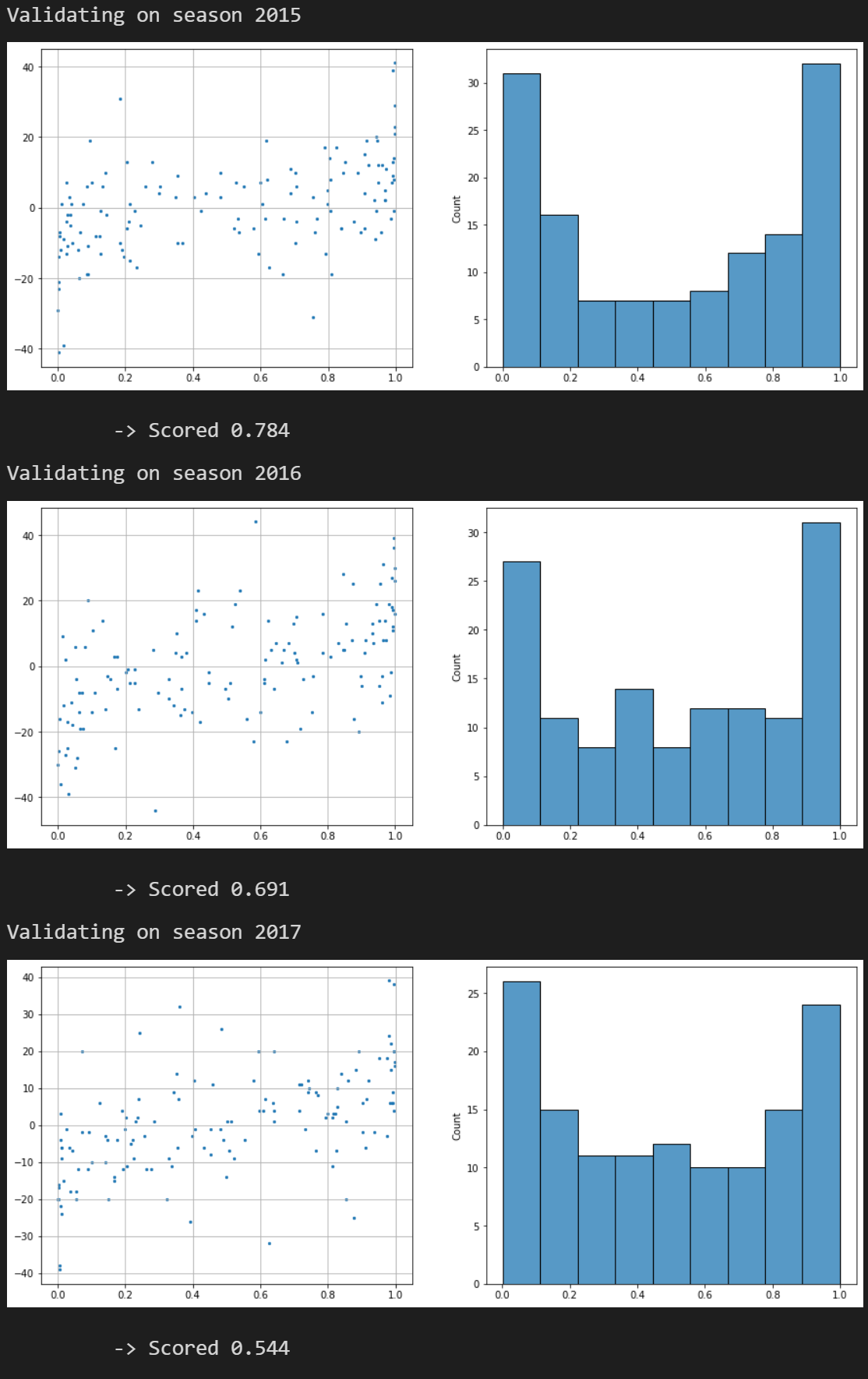
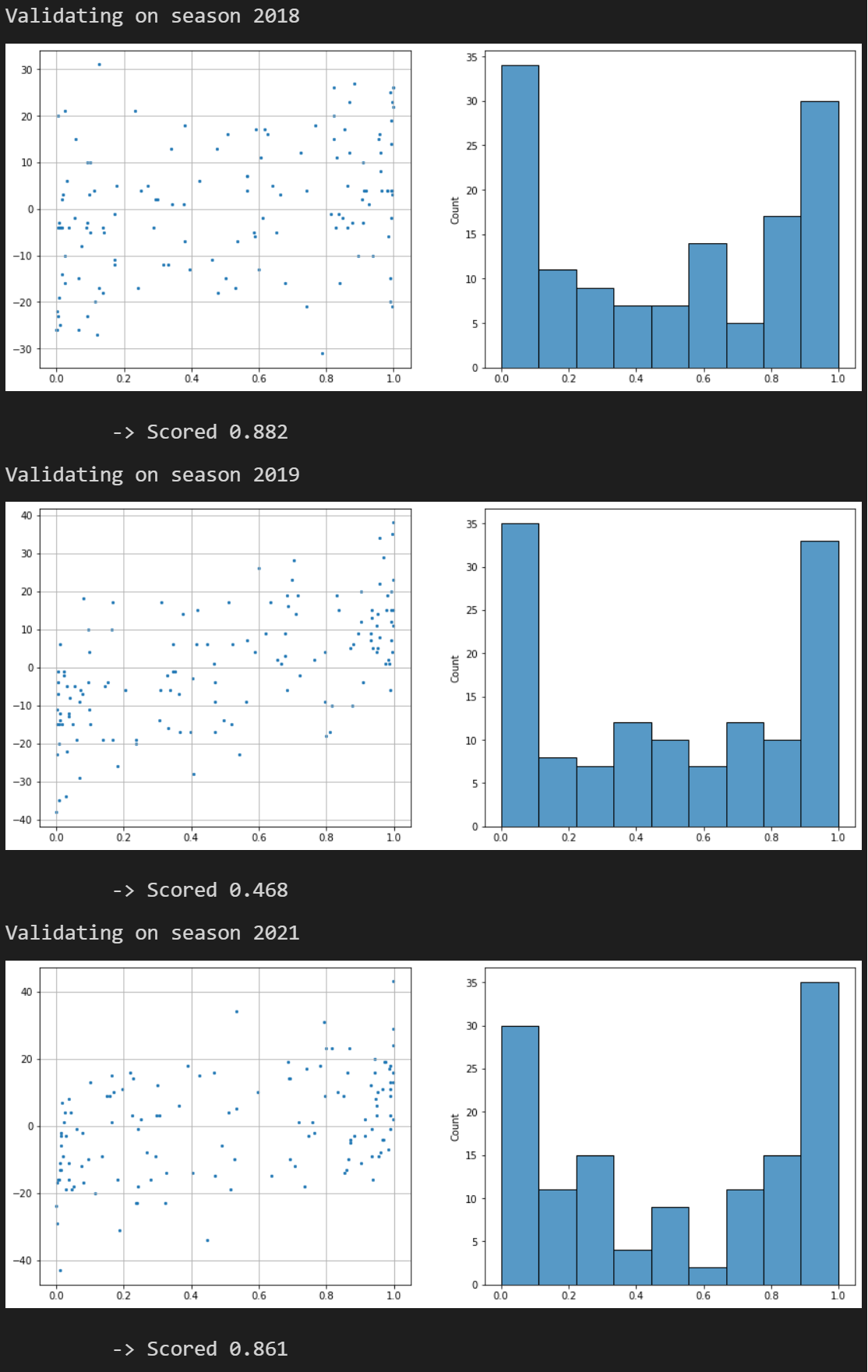
（3）random\_state

指定随机数种子

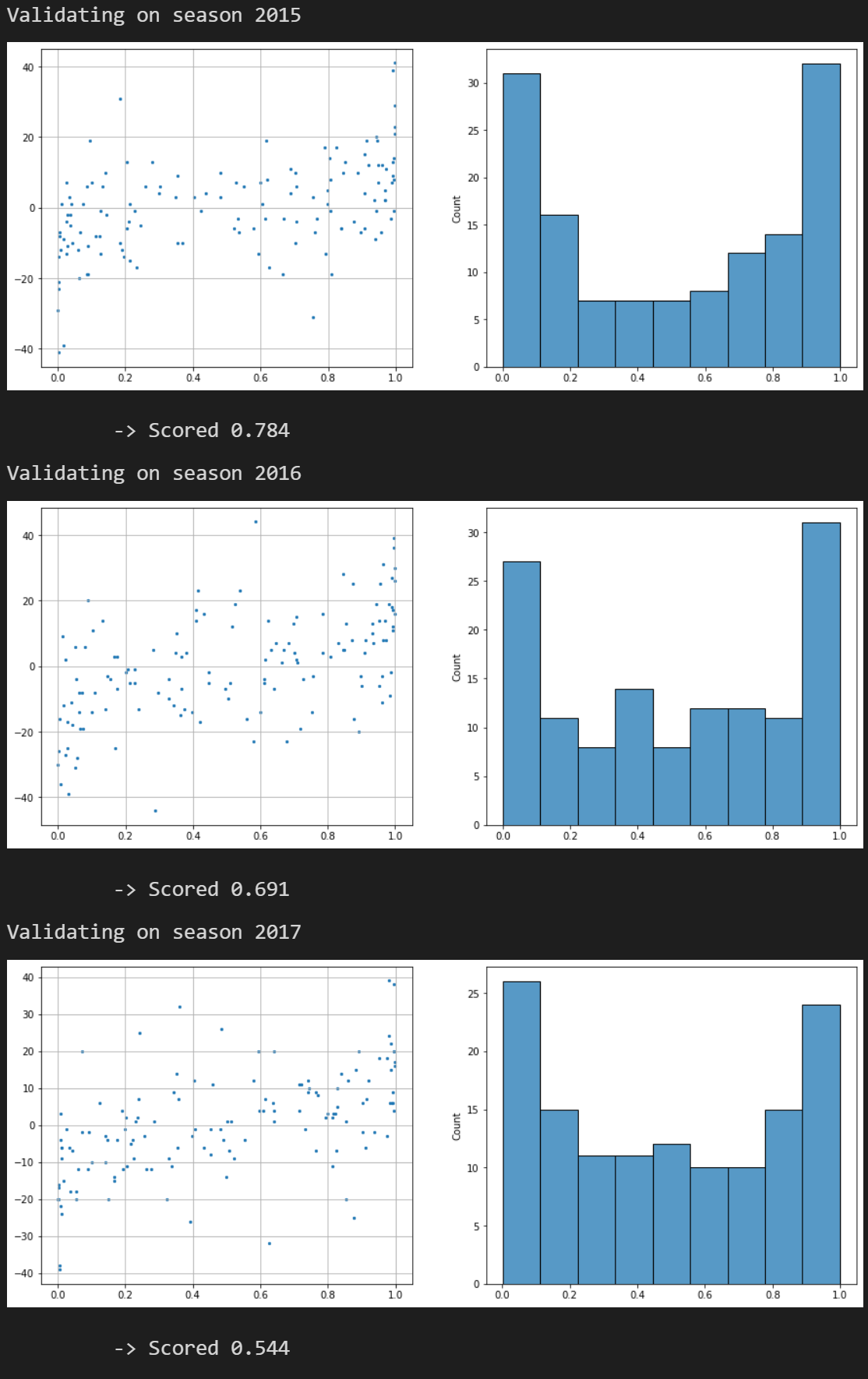
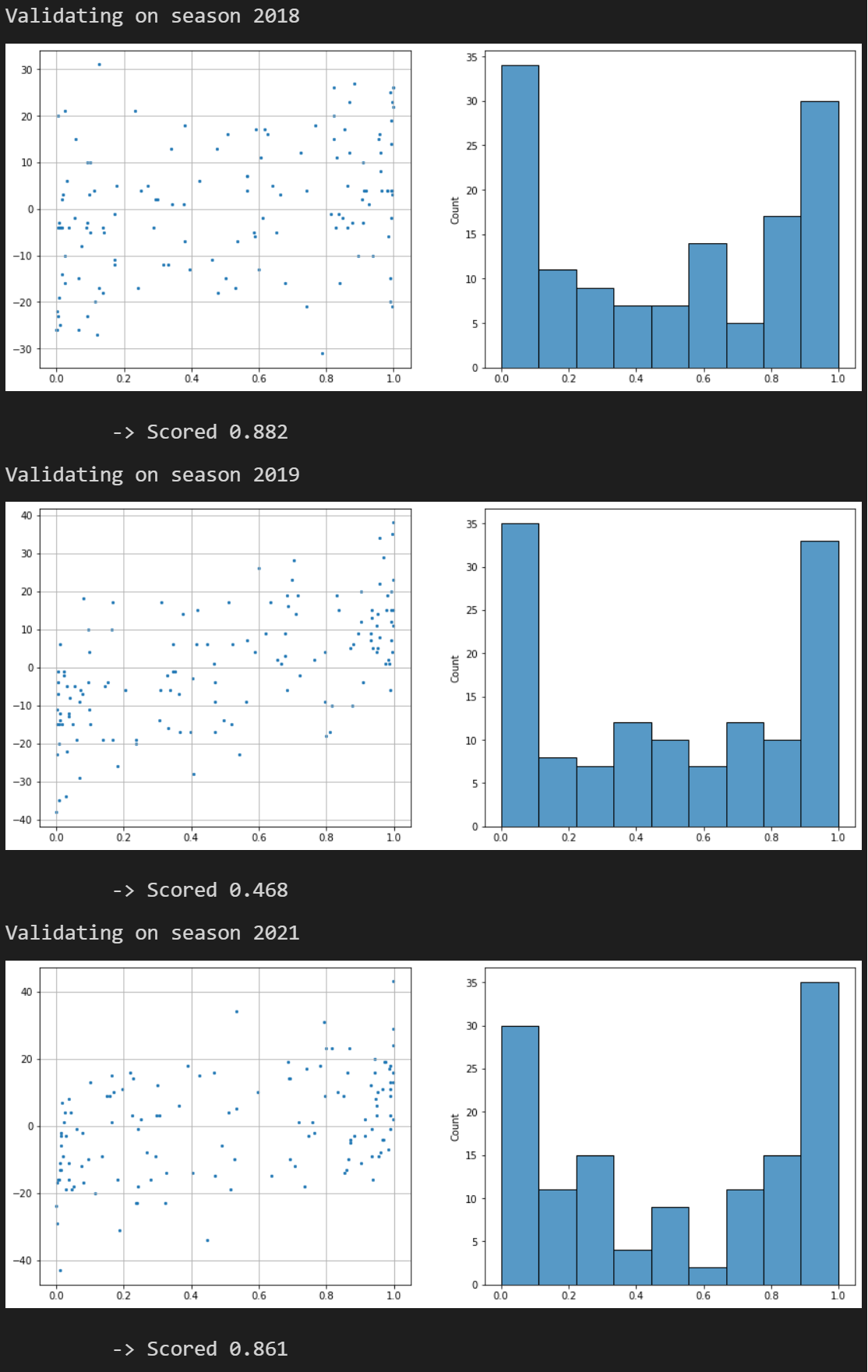
首先尝试了，树形默认设置，采用二分类模型，随机数设置为42，

1. lightgbmc\_model = LGBMClassifier(boosting\_type='gbdt', objective='binary',random\_state=42)

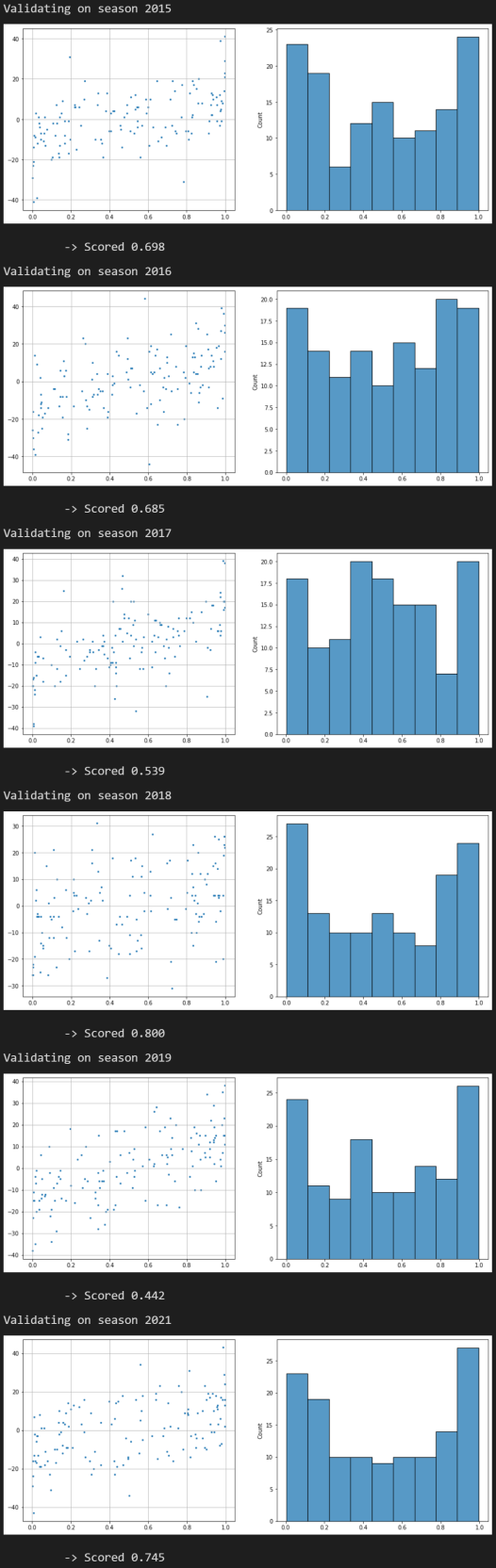
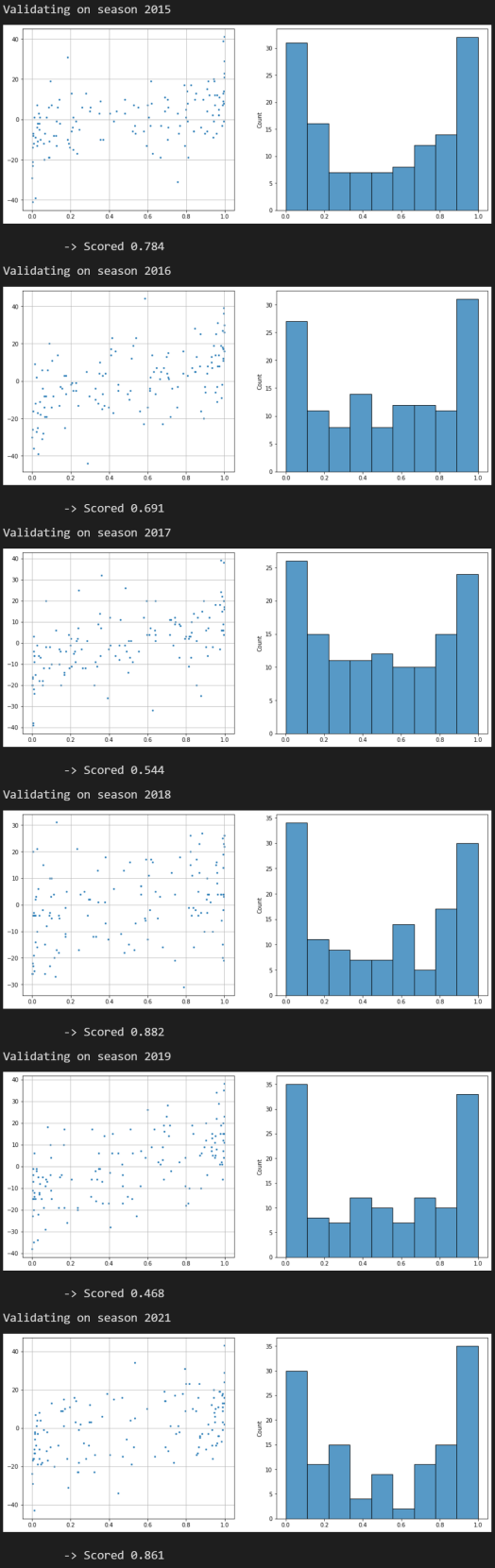
效果如下：

如若全改用默认值，随机数为40，结果变化不大，从消融对比分析角度看，这里参数的选择影响不大，所以接下来从另一个角度考虑

对比LightGBM和XGBoost

对比两者发现，得分发生了变化，有的预测的更准了，但是有的预测的更加糟糕了，为了模型的鲁棒性，我们还是选择使用最初的模型

# 第四章 总结与展望

## 4.1总结

1. 针对于较多的数据集以及过多的属性标签，采用了优先选择多表根据共同属性进行了合并操作，方便后续数据的操作和模型的构建与训练

2. 针对于选择哪些指标对比赛结果更具有参考性，我们查阅了之前篮球比赛的相关资料以及衡量能力的各个指标，采取了构造新的特征的思路，构造了命中率、进攻失误率、传球成果率，以及之前比赛输赢队伍的分差、团队获胜率等等作为主要建模特征指标

3. 针对如何选取更好的参数来进行模型的训练，多次调整模型参数，并且对比多个模型来进行模型的选取

## 4.2展望

1. 准确率较高，达到80%左右，后续可以进一步考虑更多的因素，诸如，城市、教练等可能因素的影响加入的训练模型当中，构造更多的决策变量，来提高准确率。

2. 不断更新预测模型的数据，随着历史数据的不断更新，不断迭代修正原始模型，符合机器学习中的反向传播算法理论，准确率会不断提高，并稳定在高准确率

3. 后续可以用到其他比赛预测当中，不会掺杂个人情绪的影响，完全有程序输出结果，具有一定的参考性。