|  |  |  |
| --- | --- | --- |
| **Problem Chosen**  C | **2024**  **MCM/ICM**  **Summary Sheet** | **Team Control Number**  2419698 |

**谁能在比赛中获胜**

**Summary**

2023年温布尔登网球男子单人决赛中，20岁的西班牙新星卡洛斯·阿尔卡拉斯战胜了传奇老将诺瓦克·德约科维奇。这场传奇比赛在进行中经历了多次局势转变，其中“势头”也就是胜利带来的心理效应得到广泛关注。我们利用赛中后两轮的数据建立了特征系数并建立模型分析该场比赛。

第一个问题中，在充分挖掘数据后构建了涵盖选手个人技术、实时心理状态以及身体疲惫程度三方面的特征指标体系。进行数据预处理以及标准化处理后，对16个特征变量进行了logistic回归，证明搭建的特征变量对选手得分point表现有显著影响。然后通过五折验证法各自得出LGBM、XGBOOST、SVC、MLP、LR的accuracy、recall、precision、fl和auc。其中LGBM其accuracy、precision、recall和fl以及auc分别为0.69、0.69、0.7、0.69和0.77，综合表现最好。最后选取LGBM模型对选手数据进行可视化。最后发现发球者和得分者与选手实际得分情况大体一致。选手胜利时势头较高，输的时候势头较小。

第二个问题中，通过斯皮尔曼相关系数分析训练好的LGBM模型所输出的概率与选手实际得分情况证明了二者之间显著相关，也就是势头越大，选手的表现越好，越容易取得得分。随之采用多项式回归，势头的p值同样小于0.05，其二次项系数为0.02，一次项系数为0.9021，因推得势头对结果的影响是显著的。

第三个问题里，将探讨焦点聚焦在预测选手的单局game胜负与否。首先通过以game为粒度的逻辑回归证明本文所搭建的特征变量体系对预测选手单局game胜负有显著影响。随后同样以五折验证法计算分别得出accuracy、recall、precision、fl和auc，其中SVC的accuracy，recall，precision，f1和auc分别为0.71, 0.75, 0.7, 0.72和0.75，优于其他模型。因此选择SVC以单局game的胜负为粒度去训练，最后通过可视化对比模型输出的势头与选手实际胜负情况发现较为一致。

问题四中，通过随机选择了4场比赛进行测试，将其余数据作为训练集。以确保我们所提出的模型的普适性和可靠性。通过roc-auc进行可视化，我们清洗地呈现了模型在比赛之间的显著差异，从而验证了该模型在不同比赛场景下的稳健性。

最后，我们对模型进行了评估，总体得出本文数据挖掘后所构建的特征变量显著影响选手赛事表现，根据特征变量的涵盖方面即选手个人技术、实时心理状态、势头、身体疲惫程度对网球教练提出一定建议。

**Keywords:** Logistic Regression；LGBM; polynomial regression；SVC

Contents

[1 Introduction 3](#_Toc19220)

[1.1 Background 3](#_Toc17950)

[1.2 Restatement of the Problem 3](#_Toc8602)

[1.3 Our Work 3](#_Toc26505)

[2 Assumptions and Justifications 3](#_Toc4614)

[3 Notations 3](#_Toc19813)

[4 Problem 1: Model establishment and solution 3](#_Toc7089)

[4.1 Abc 3](#_Toc23744)

[4.1.1 Abc 3](#_Toc20010)

[5 Problem 2: Model Establishment and Solution 4](#_Toc29401)

[6 Problem 3: Model Establishment and Solution 4](#_Toc5619)

[7 Sensitivity Analysis 4](#_Toc17766)

[8 Strengths and Weaknesses 4](#_Toc18492)

[8.1 Strengths 4](#_Toc14811)

[8.2 Weaknesses 4](#_Toc11657)

[References 5](#_Toc5129)

[Appendices 6](#_Toc2655)

[Report on Use of AI 7](#_Toc7820)

# Introduction

## Background

The men's singles final of the 2023 Wimbledon Open ended with Novak Djokovic losing at Wimbledon for the first time in 10 years. The final, which began with the gimmick of "Rising Star vs. Veteran", had many twists and turns before 20-year-old Spanish star Carlos Alcaraz bucked the trend and came out on top to win the title.

Multiple turnovers are often recognized as having a significant impact on the outcome of a sporting event. Multiple consecutive victories usually have a positive psychological effect on a player, also known as "momentum". One dictionary definition of the word "momentum" is "power or force gained through movement or a series of events". For the aforementioned final, official scoring data was collected for each point in time after the second round of the tournament. This data can be used to analyze the flow of scoring in the Wimbledon final and how momentum shifted between the two players, thus helping players and coaches conduct post-match reviews and analyze the causes of success or failure.2023 Wimbledon Open Men's Singles Final ended with Novak Djokovic losing at Wimbledon for the first time in 10 years. The final, which began with the gimmick of "rising star vs. veteran", had many twists and turns before 20-year-old Spanish star Carlos Alcaraz bucked the trend and came out on top to win the title.

Multiple turnovers are often recognized as having a significant impact on the outcome of a sporting event. Multiple consecutive victories usually have a positive psychological effect on a player, also known as "momentum". One dictionary definition of the word "momentum" is "power or force gained through movement or a series of events". For the aforementioned final, official scoring data was collected for each point in time after the second round of the tournament. This data can be used to analyze the flow of scoring in a Wimbledon final and the shifts in momentum between two players, thus helping players and coaches to conduct post-match reviews and analyze the reasons for success or failure.

打网球的人

描述已自动生成

Figure 1 Competition display

## Restatement of the Problem

Wimbledon officials have collected player scoring data for the full 2023 final, which includes the respective plate scores and position scores of both players at various points in the match. By analyzing the above background, we summarize the tasks to be solved as follows:

·Identify and build a system of factors influencing match scoring;

## Our Work

图示

描述已自动生成 Figure 2 Workflow

# Assumptions and Justifications

-Assumed that the player who serves the ball in a match is more likely to score points than the receiver.

-Assumed that a player's physical fatigue will have an effect on the outcome of the match.

-Assumed that a player's mental state plays a significant role in individual scoring performance.

-Assumed that "momentum" will affect a player's game performance.

-Assumed that a player's individual technical ability has a direct impact on the game.

# Notations

Table 1: Notations

|  |  |
| --- | --- |
| **Symbol** | **Description** |
| X1 | The number of games won in the current set |
| X2 | The score lead progress in the current game |
| X3 | Whether it is the server |
| X4 | Whether or not the previous point is scored |
| X5 | Set lead progress for this match |
| X6 | Whether or not the game serves a point (no touch) |
| X7 | Whether or not the game is returned for a score (no touch) |
| X8  X9  X10  X11  X12  X13  X14  X15  X16 | Whether or not there is a double fault in this game |
| Whether or not there is an unforced error in this game |
| Net approaches and ratio of points won at the net |
| Ratio of scoring opportunities to actual points scored on the opponent’s serve in this set |
| Total mileage run within this match |
| Total mileage run within the last three points |
| Player's movement mileage for the previous point |
| Real-time serve speed |
| Whether or not it's a real-time speed matching interaction term for the server and the serve |

# Problem 1: Modeling the Scoring Process for LGBM-based

## Feature engineering and data preprocessing

### Feature engineering

After fully understanding as well as mining the data from the last two rounds of the Wimbledon men's singles final, we believe that a tennis player's possible scores are related to factors such as the player's fatigue level, individual technical qualities as well as the player's mental state, in addition to the initiative of the serve.

-Individual technical ability: calculated using past or real-time player scores

-Player fatigue: Calculated using the player's real-time total mileage of the running map

-Mental state in real time: whether there was a service error, whether the player won without overtime, whether the player scored on the opponent's serve, etc., so as to establish the following index system:

Table 2: Notations

|  |  |
| --- | --- |
| **Symbol** | **Description** |
| X1 | The number of games won in the current set |
| X2 | The score lead progress in the current game |
| X3 | Whether it is the server |
| X4 | Whether or not the previous point is scored |
| X5 | Set lead progress for this match |
| X6 | Whether or not the game serves a point (no touch) |
| X7 | Whether or not the game is returned for a score (no touch) |
| X8 | Whether or not there is a double fault in this game |
| X9 | Whether or not there is an unforced error in this game |
| X10 | Net approaches and ratio of points won at the net |
| X11 | Ratio of scoring opportunities to actual points scored on the opponent’s serve in this set |
| X12 | Total mileage run within this match |
| X13 | Total mileage run within the last three points |
| X14 | Player's movement mileage for the previous point |
| X15 | Real-time serve speed |
| X16 | Whether or not it's a real-time speed matching interaction term for the server and the serve |

### 

### Data Preprocessing

为了更好的挖掘现有数据中所能提取的关键特征，我们对官方给定的原始数据进行了以下处理：

**异常值和缺失值处理**

In order to deal with missing values and outliers in the raw data, a series of data processing methods are used: filtering, feature engineering, conditional logic, ratio calculation, cumulative summation, outlier handling and indexing operations. These steps eliminate redundant information and facilitate the identification and extraction of relevant information from the dataset.

Step 1: Filter specific rows from a large dataset using conditional indexing. For example, df[df.match\_id==match\_id] selects all rows from df that match a given match\_id.

Step 2.

-Calculate new features (e.g., scoring differential, serve-side identification, aces, winning points, double faults, and percentage of points at the net, etc.) to reflect a player's match performance.

-Creates binary features (0 or 1) to indicate whether certain conditions are met or not, usually used to reflect the presence or absence of events.

-Calculate cumulative features, such as the sum or average of distance run.

Step 3:Apply logical operations to determine the value of the feature, e.g. 1 if condition else 0 is used to create binary features.

Step 4:Calculate proportional features, such as the ratio of the number of net points won by a player in a point or set to the total number of net points scored.

Step 5:The sum() function is used to calculate the cumulative statistics of a player under certain conditions (e.g., match, set, or game).

Step 6:When the denominator is zero, the eigenvalue is set to zero to avoid the error of dividing by zero.

Step 7:Use index.tolist().index() and iloc to locate and extract data from specific rows.

表格

描述已自动生成图3 数据展示

**标准化处理**

在搭建的16个特征系数中，由于不同系数之间总量纲较大，因而对所有特征进行了标准化处理。

## 基于特征工程的logistic检验

根据上述特征指标可以计算一个选手每一场match中每一个set的每一个game的每一个point得分情况以及对应选手指标具体时点表现。为了验证所搭建的特征对得分的影响有显著影响，同时考虑到每一个point得分情况是一个二分类变量，我们基于二分类Logistic回归的方式来检验每个自变量关于得分情况的影响性。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification table a** | | | | | |
|  | actual measurement | | prediction | | |
| label | | percentage of correct |
| 0 | 1 |
| Step 1 | label | 0 | 271 | 517 | 34.4 |
| 1 | 184 | 1058 | 85.2 |
| overall percentage | |  |  | 65.5 |
| a. The cut-off value is .500. | | | | | |

表3 logistic检验结果

可以看到,logistic回归的总体准确率为65.5。关于选手实际没有得分的样本，模型分类准确率为34.4；关于选手实际得分的样本，模型分类准确率为85.2%，远高于没有得分的样本。因而推知当前模型偏好于将样本分类到实际得分（label为1）的情况。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the equation** | | | | | | | |
|  | | B | Standard Deviation | Wald | DOF | Significance Level | Exp(B) |
| Step 1a | x1 | .016 | .176 | .009 | 1 | .926 | 1.017 |
| x2 | -1.141 | .363 | 9.881 | 1 | .002 | .319 |
| x3 | .689 | 1.286 | .287 | 1 | .592 | 1.992 |
| x4 | .170 | .166 | 1.048 | 1 | .306 | 1.186 |
| x5 | .032 | .203 | .024 | 1 | .876 | 1.032 |
| x6 | .689 | .127 | 29.400 | 1 | .000 | 1.991 |
| x7 | .456 | .128 | 12.697 | 1 | .000 | 1.577 |
| x8 | .472 | .153 | 9.491 | 1 | .002 | 1.603 |
| x9 | -.535 | .109 | 24.287 | 1 | .000 | .586 |
| x10 | .914 | .124 | 54.371 | 1 | .000 | 2.493 |
| x11 | .073 | .140 | .273 | 1 | .601 | 1.076 |
| x12 | .120 | .288 | .174 | 1 | .676 | 1.128 |
| x13 | -.164 | .482 | .116 | 1 | .734 | .849 |
| x14 | -1.445 | .657 | 4.842 | 1 | .028 | .236 |
| x15 | .511 | .626 | .668 | 1 | .414 | 1.667 |
| x16 | -.876 | 1.693 | .268 | 1 | .605 | .416 |
| Constant | -.072 | .346 | .043 | 1 | .835 | .931 |
| a. Variables entered in step 1：x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16。 | | | | | | | |

表4 指标影响程度

### 其中一半自变量的p值都小于0.05的。如果p值小于0.05，则有理由说明自变量x是显著影响因变量y的，所以在特征变量体系中能够显著影响选手实时得分情况的自变量包括x2、x6、x7、x8、x9、x10、x14。上述七个自变量对应了选手个人能力、选手疲惫情况以及心理状态，因而可知该三个因素显著影响选手实时得分。

## 五折验证法下LGBM动态捕捉得分流程

### 模型选取

机器学习模型选取中选用了集成树模型LGBM和XGBOOST，并使用了支持向量机、感知机网络和逻辑回归作为对比算法。然后使用5折交叉验证法验证模型，使用基于混淆矩阵的accruacy、precision、recall和fl以及auc和roc曲线进行结果评价。基于混淆矩阵的结果如表4所示，基于auc和roc曲线的训练集和测试集结果分别如图4和图5所示。

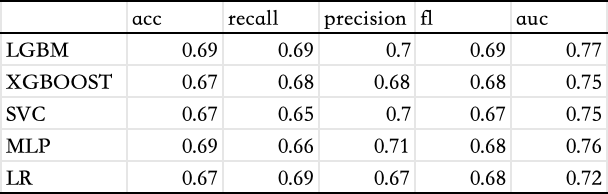


表5 交叉验证法结果

图表

描述已自动生成

图4 训练集ROC曲线

图表, 折线图

描述已自动生成

图5 测试集ROC曲线

LGBM其accuracy、precision、recall和fl以及auc分别为0.69、0.69、0.7、0.69和0.77，综合表现最好。其次是感知机神经网络，为0.69、0.66、0.71、0.68和0.76。二者模型指标差异较小，说明模型并没有非常明显的正负样本判别偏好，因而说明LGBM效果最好。

ROC曲线反映了不同阙值不同的precision和recall指标的变化。通过roc曲线图可以验证LGBM表现最好。

### 模型应用与训练

据此采用LGBM模型选取2023年温布尔登男单决赛选手Carlos Aloanaz进行数据训练与可视化。结果如图6所示。

图表

描述已自动生成

图6 选手表现预估

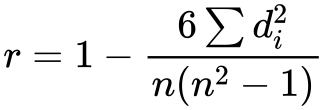
由于影响选手实际得分因素过多，本模型或将存在较多噪音，准确预测或只能达到其中0.7部分。根据分析，我们发现发球者和得分者与选手实际得分情况大体一致。选手胜利时势头较高，输的时候势头较小。因而说明本文提出的模型具有有效性。

# 问题二: Model Establishment and Solution

模型二的搭建目的是为了具体量化“势头”在比赛中的实际影响，旨在揭示球员在比赛过程中所呈现的波动和成功是否呈现一种非随机的趋势。对此通过模型一LGBM对比赛得分捕捉流程的应用以产生选手得分的概率输出。通过对比概率输出与实际选手得分情况进行关联程度分析。

## 斯皮尔曼相关性分析

斯皮尔曼相关系数，是度量两个变量的依赖性的非参数（不依赖于数据分布的特定假设）统计度量方法。他的取值范围在-1到1之间，其中：1表示正相关；-1表示负相关；0表示两个变量之间没有线性关系。斯皮尔曼相关系数(通常用符号r表示)的计算公式如下：



因为需要证明的是模型输出的“势头”能否影响到选手真实表现，所以我们预先证明其相关性是显著的，结果如表6所示。

|  |  |
| --- | --- |
| Correlation Coefficient | P-value |
| 0.479 | 0.000 |

表6 交叉验证法结果

可以看到P值小于0.05.势头和选手表现是显著相关的，并且这种关系是一种正相关关系，即势头越大，选手的表现越好，越容易取得得分。

## 多项式回归结果

同样的我们使用多项式回归分析势头是否能够显著影响选手得分，结果如图7所示。

表格

描述已自动生成

图7 选手表现预估

可以看到，势头的p值同样小于0.05，其二次项系数为0.02，一次项系数为0.9021，势头对结果的影响是显著的。

# Problem 3: Model Establishment and Solution

## 以game为粒度的统计logistic检验

为实现比赛预测，本文将采用选手单局game的胜负与否为粒度，以此降低事件的高随机性并在利用模型一的基础上改进从而提高效率

因此和问题一同样的，我们采用逻辑回归检验特征变量体系对选手胜负是否显著。结果如表7和表8所示。

表7 逻辑回归的分类性能

表格

描述已自动生成

相较基于point的逻辑回归，基于game的逻辑回归的准确性从65.5变为了71.4，基于game的模型结果更加能够反映选手的真实波动状况。接着，逻辑回归检验的结果如表8所示。

表8 逻辑回归的分类性能

表格

描述已自动生成

对比问题1的结果，该结果中许多原本不显著的变量也开始显著，其中包括x4，x11与x13。因此可知，特征变量体系对影响选手的单局胜负影响显著，并且影响优于得分表现。

## 模型选择

由于二元逻辑回归的性能有局限性，因此同样采用了多算法对比，并使用accuracy，recall，precision，f1和auc进行评价，使用5折交叉验证法验证，结果如下：

表9 以game为粒度的5折交叉验证算法结果表格

描述已自动生成

其中表现最好的是SVC，其accuracy，recall，precision，f1和auc分别为0.71, 0.75, 0.7, 0.72和0.75。SVC的recall指标最优，表明模型有一定偏好。

图表

描述已自动生成图表, 折线图

描述已自动生成

图8 训练集ROC曲线 图9 测试集ROC曲线

ROC曲线显示出SVC效果最好。此外LGBM、XGB的训练集得分远高于测试集，建议选取经典模型。

## 模型应用

图表

描述已自动生成图表, 直方图

描述已自动生成

图10 特征得分 图11 比赛走势

通过选取2023年温布尔登男单决赛中Carlos Alcaraz选手的实时表现，并将其可视化，结果如图11所示。在此次决赛中，对手第一局完胜，第二局Carlos Alcaraz险胜，第三局Carlos Alcaraz完胜，第四局对手完胜，第五局获得最终胜利。预测准确率对比得分流程模型更具优势。

# 预测评价与泛化性测试

问题四中，通过随机选择了4场比赛进行测试，将其余数据作为训练集。通过roc-auc可视化发现了模型在不同比赛之间存在差异，从而验证了该模型在不同比赛场景下的稳健性。关于模型性能差异的原因，本文提出选手个人能力或为异常重要因素。

## 四场比赛的选手波动预测效果评价

随机选取了2023年温布尔登男单比赛中的四场比赛，分别为’2023-wimbledon-1305’,’2023-wimbledon-1314’,’2023-wimbledon-1602’,’2023-wimbledon-1302’。利用LGBM模型对波动和roc曲线进行可视化，从而观察具体选手波动情况与模型的预测效果，结果分别如图8和图9所示。

可以看到，比赛与比赛之间是存在明显差异的，甚至对于某些比赛的auc可以达到0.9以上，这说明模型的效果已经很不错了。但是，我们在实验过程中发现，对于一些比赛，又会存在0.7多一点的auc值，这意味着模型在某些比赛上的表现是不够好的。

关于模型表现不佳问题，本文认为其实提出的指标体系是有一定缺陷的，这一缺陷主要来自于模型对于选手不具有先验知识，因此对于两个选手刚开始进行比赛的时候，模型的预测可能是不够精准的，因为此刻模型并不知道哪个选手的能力更加突出，而随着时间的推移，由于模型已经回去两个选手的实时比赛情况，例如能力比较强的选手大概率他的比赛set和game得分也是高于能力弱的选手的，这时模型就通过后验知识能够在后面的比赛中预测逐渐变得精准。

因此，如果在考虑纳入未来模型因素的话，一个十分重要的因素就是选手的个人能力，而这一版可以通过选手在过去的比赛中的表现得到。

例如，我们在指标体系的基础上进行扩展，将选手的能力根据今年比赛的成绩来分类，以每个选手赢得的总set数量为依据，得到每个选手的能力指标和对手能力指标。具体的我们使用每个选手的得分总数（能力）作为新的自变量，构建了选手个人能力和对手能力的指标。分析其与选手表现之间的关系。

结果发现，这两个变量基本上是显著影响game获胜情况的，特别的，对手能力相对于个人能力会有更突出的重要性。

## 其他体育比赛泛化性评价

虽然不同体育项目具有独特的规则和比赛特征，但本文章所提出的影响因素或将在不同体育中存在共通之处。以下是说明其适用性的一些观点：

1. 身体疲惫程度：体育赛事的第一耗费是个人体能。体能的过度消耗会影响运动员的比赛表现。
2. 个人技术能力：体育选手的个人素质是比赛表现的最基础因素。
3. 心理状态，其中主要为“势头”：选手的心理状态在各体育赛事中均有显著影响。其中“势头”，也就是胜利带来的心理效应是

综上，不同的体育赛事有着不同的规则，但本文章所搭建的特征变量体系以选手及全体体育赛事为参考而搭建，进而可以进行泛化对其他类别的体育比赛选手表现进行捕捉量化。

# Memo

**收件人：**网球选手及教练团队

**主题：**关于提升网球选手心理、技术与体能的综合建议

**正文：**

尊敬的选手和教练，

近期本组织通过对温网男子单打决赛后两轮的数据进行建模可视化后，充分了解到本文“势头”特征变量体系能够有效喜欢比赛得分过程并预测选手比赛表现，具备一定有效性。从而也可推知特征变量所涵盖的三方面即选手心理状态、个人能力以及体能训练对赛事表现有着举足轻重轻重的地方。

以下是我们针对这三个方面提出的综合建议，希望能为您的训练和比赛带来帮助：

一、心理训练的重要性

• 选手应培养冷静、自信的心态，尤其在关键时刻能够顶住压力，发挥最佳水平。

• 教练应协助选手调整心态，通过心理训练提升选手的抗压能力和自信心。

• 选手需学会在失败中保持冷静和积极，总结经验教训，不断自我提升。

二、持续精进技术能力

• 选手应致力于提高技术水平和比赛智商，包括击球质量、移动速度及战术运用。

• 教练应根据选手特点制定个性化训练计划，帮助选手有针对性地提升技术水平。

• 选手需学会根据对手特点和场地条件灵活调整战术，合理分配体力和精力。

三、强化体能训练的科学性

• 选手应通过有氧训练、力量训练、柔韧性训练等方式提升身体素质，为比赛做好准备。

• 教练应根据选手的体能状况和训练目标制定科学训练计划，确保训练效果与安全性。

四、赛中战术策略调整

• 选手应根据比赛中的实时数据反馈，灵活调整战术策略，提高应变能力。

• 教练在比赛中应密切观察选手表现，及时给予指导和建议，帮助选手做出正确决策。

五、赛后数据分析

• 选手和教练应深入分析比赛数据，了解选手的优点和不足，制定针对性的训练计划。

• 通过数据分析，教练可为下一场比赛制定更合适的战术策略。

希望选手和教练团队认真考虑并执行以上建议，共同努力提升选手的综合实力。祝训练与比赛顺利！

# Strengths and Weaknesses

## Strengths

* LightGBM (LGBM) 模型对大规模数据集具有更准确更高效的数据捕捉能力，有助于提高数据处理效率；
* 斯皮尔曼等级相关系数适合非参数性质数据，有较高稳健性并且结果呈现简单直观；
* 多项式回归可以拟合非线性关系，符合本次数据结构；
* 支持向量机 (SVC)在高维空间表现良好，泛化能力强；

## Weaknesses

* LightGBM (LGBM)需要调整许多参数来获得最佳性能，这可能需要较多的机器学习知识和实验；
* 支持向量机 (SVC)在大数据集上可能会消耗大量内存；
* 多项式回归可能在数据范围外表现不好，外推能力不强；

# References

[1] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[2] Aziz R M, Baluch M F, Patel S, et al. LGBM: a machine learning approach for Ethereum fraud detection[J]. International Journal of Information Technology, 2022, 14(7): 3321-3331.

[3]Hammouri Z A A, Delgado M F, Cernadas E, et al. Fast SVC for large-scale classification problems[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2021 (01): 1-1.

[4]Bujang S D A, Selamat A, Ibrahim R, et al. Multiclass prediction model for student grade prediction using machine learning[J]. IEEE Access, 2021, 9: 95608-95621.

[5]Massaoudi M, Refaat S S, Chihi I, et al. A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting[J]. Energy, 2021, 214: 118874.

[6]LaValley M P. Logistic regression[J]. Circulation, 2008, 117(18): 2395-2399.

[7]Sperandei S. Understanding logistic regression analysis[J]. Biochemia medica, 2014, 24(1): 12-18.

[8]Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[9]Carleo G, Cirac I, Cranmer K, et al. Machine learning and the physical sciences[J]. Reviews of Modern Physics, 2019, 91(4): 045002.

[10]Osman M, He J, Mokbal F M M, et al. Ml-lgbm: A machine learning model based on light gradient boosting machine for the detection of version number attacks in rpl-based networks[J]. IEEE Access, 2021, 9: 83654-83665.

[11]Barzegar R, Razzagh S, Quilty J, et al. Improving GALDIT-based groundwater vulnerability predictive map\*\* using coupled resampling algorithms and machine learning models[J]. Journal of hydrology, 2021, 598: 126370.

[12]Ntakolia C, Kokkotis C, Karlsson P, et al. An explainable machine learning model for material backorder prediction in inventory management[J]. Sensors, 2021, 21(23): 7926.

[13]Yakar M, Etiz D, Metintas M, et al. Prediction of radiation pneumonitis with machine learning in stage III lung cancer: a pilot study[J]. Technology in cancer research & treatment, 2021, 20: 15330338211016373.

[14]Fei B, Liu J. Binary tree of SVM: a new fast multiclass training and classification algorithm[J]. IEEE transactions on neural networks, 2006, 17(3): 696-704.

[15]Shawe-Taylor J, Sun S. A review of optimization methodologies in support vector machines[J]. Neurocomputing, 2011, 74(17): 3609-3618.

[16]Li X Z, Kong J M. Application of GA–SVM method with parameter optimization for landslide development prediction[J]. Natural Hazards and Earth System Sciences, 2014, 14(3): 525-533.

[17]Li H, Chung F, Wang S. A SVM based classification method for homogeneous data[J]. Applied Soft Computing, 2015, 36: 228-235.

# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: LGBM模型的训练，用以建立网球比赛得分流程并可视化 |
| 代码如下：  index = df[df.match\_id=='2023-wimbledon-1701'].reset\_index(drop=True).index  test = dataset.iloc[index]  train = dataset.drop(index,axis=0)  model = LGBMClassifier(random\_state=30)  model.fit(train[columns].values,train['label'].values)  pred = model.predict\_proba(test[columns].values)  pred = pd.DataFrame({'实时得分':pred[:,1]})  match1 = pred.iloc[:45]  match2 = pred.iloc[45:126]  match3 = pred.iloc[126:195]  match4 = pred.iloc[195:259]  match5 = pred.iloc[259:]  plt.figure(figsize=(12, 6), dpi=80, facecolor='w')  plt.plot(match1.index,match1.values,color='purple')  plt.plot(match2.index,match2.values)  plt.plot(match3.index,match3.values)  plt.plot(match4.index,match4.values)  plt.plot(match5.index,match5.values)  plt.xlabel("Points")  plt.ylabel("Performance")  ##plt.savefig('问题1\\经典对决实时走势.png',dpi=500)  plt.show() |
| Appendix 2 |
| Introduce: 五折检验法代码 |
| 代码如下：  import warnings  warnings.filterwarnings("ignore")  def function(model):  auc = round(cross\_val\_score(model,dataset[columns].values,dataset['label'].values, cv=5,scoring='roc\_auc').mean(),2)  acc = round(cross\_val\_score(model,dataset[columns].values,dataset['label'].values, cv=5,scoring='accuracy').mean(),2)  recall = round(cross\_val\_score(model,dataset[columns].values,dataset['label'].values, cv=5,scoring='recall').mean(),2)  precision = round(cross\_val\_score(model,dataset[columns].values,dataset['label'].values, cv=5,scoring='precision').mean(),2)  f1 = round(cross\_val\_score(model,dataset[columns].values,dataset['label'].values, cv=5,scoring='f1').mean(),2)  return acc,recall,precision,f1,auc  model = LGBMClassifier(random\_state=30,force\_col\_wise=True)  print(f'LGBMClassifier acc,recall,precision,f1,auc :{function(model)}')  model = XGBClassifier(random\_state=50)  print(f'XGBClassifier acc,recall,precision,f1,auc :{function(model)}')  model = SVC(random\_state=50)  print(f'SVC acc,recall,precision,f1,auc :{function(model)}')  model = MLPClassifier(random\_state=60)  print(f'MLPClassifier acc,recall,precision,f1,auc :{function(model)}')  model = LogisticRegression(random\_state=50)  print(f'LogisticRegression acc,recall,precision,f1,auc :{function(model)}')  from sklearn import metrics  def f(model\_list,name\_list,types='train'):  plt.figure(figsize=(8, 7), dpi=80, facecolor='w') # dpi:每英寸长度的像素点数  plt.xlim((-0.01, 1.02)) # x,y 轴刻度的范围  plt.ylim((-0.01, 1.02))  plt.xticks(np.arange(0, 1.1, 0.1)) #绘制刻度  plt.yticks(np.arange(0, 1.1, 0.1))    if types == 'test':  for model,name in zip(model\_list,name\_list):  ytest\_prob = model.predict\_proba(xvalid)[:,1]  fpr, tpr, \_ = metrics.roc\_curve(yvalid, ytest\_prob)  auc = metrics.auc(fpr, tpr)  plt.plot(fpr, tpr, '-', lw=2, label=f'{name} AUC:%.4f' % auc) # 绘制AUC  else:  for model,name in zip(model\_list,name\_list):  ytest\_prob = model.predict\_proba(xtrain)[:,1]  fpr, tpr, \_ = metrics.roc\_curve(ytrain, ytest\_prob)  auc = metrics.auc(fpr, tpr)  plt.plot(fpr, tpr, '-', lw=2, label=f'{name} AUC:%.4f' % auc) # 绘制AUC  plt.legend(loc='upper left',fontsize=15) # 设置显示标签的位置  plt.xlabel('False Positive Rate', fontsize=14) #绘制x,y 坐标轴对应的标签  plt.ylabel('True Positive Rate', fontsize=14)  plt.tick\_params(labelsize=23)  plt.grid(True, ls=':') # 绘制网格作为底板;b是否显示网格线；ls表示line style  #plt.savefig(f'roc\_auc({types}(采样前)).png',dpi=500)  plt.show()    xtrain, xvalid, ytrain, yvalid = train\_test\_split(dataset[columns].values,dataset['label'].values,random\_state=620,test\_size=0.2)  model1 = LGBMClassifier(random\_state=30)  model2 = XGBClassifier(random\_state=50)  model3 = SVC(probability=True,random\_state=50)  model4 = MLPClassifier(random\_state=60)  model5 = LogisticRegression(random\_state=50)  model1.fit(xtrain,ytrain)  model2.fit(xtrain,ytrain)  model3.fit(xtrain,ytrain)  model4.fit(xtrain,ytrain)  model5.fit(xtrain,ytrain)  f([model1,model2,model3,model4,model5],['LGBM','XGB','SVC','MLP','LR'],'test')  f([model1,model2,model3,model4,model5],['LGBM','XGB','SVC','MLP','LR'],'train') |

Report on Use of AI

1. OpenAl Enie (Nov 5, 2023 version, Ermie 4.0)

Queryl: <insert the exact wording of any subsequent input into the Al tool>

Output: <insert the complete output from the second query>