Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2515374

Foresting Greatness: Olympic Medal Predictions based on PRE Model

Summary

The Olympic Games, inspired by the ancient Greek Games, is the world's premier international multi-sport event. Held every four years, it features summer and winter editions, bringing together athletes from around the globe to compete in a wide range of sports. For every sports lover, the Stimulation of Olympic Medals is a very interesting topic, while it's crucial for countries to adjust their strategy. **In this paper, we will analyze the data of the Summer Olympic Games from 1896 to 2024, and use the data to simulate the results of the 2028 and 2032 Olympics using different methods from 'PRE' model mostly based on machine learning.**

First, we've examined, cleaned and transformed the raw data, setting the ice sports aside, combining various teams of one country, and mapping the countries which no longer exist to the current existing country. Then, we turn to machine learning, select the independent variables to use by **Correlation Coefficient Matrix** and **Principal Component Analysis**, split the data into training and testing parts, and use **EXtreme Gradient Boosting** methods to train features and target matrix. **Kolmogorov-Smirnov Test** is used to ensure the effectiveness of the model, while history data are reprocessed using feature importance and data in recent years are given higher weights to improve the accuracy of the model. In the end, we use the model to predict the results of the 2028 and 2032 Olympics, and the results are shown in multiple forms.

After that, we encode country labels, create feature dataset of each country,and filter out data for countries that won medals in 1896 to construct training and testing datasets of the feature of 'First Win Country'.**Random Forest** is used to train the model, and the results of the top ten countries most likely to win their first medal in 2028 are shown in the form of a bar chart.

By analyzing the data given, we've found that some specific sports and events play a significant role in the medal tally of some specific countries, for example, long distance race for Kenya. We calculate the **proportion** of medals from these events, further estimate their importance, and finally come to the result of the extent to which choosing these events impact countries performance in the medal list.

Finally, we combined methods used above and created a comprehensive prediction model called **'PRE'** model, naming after the primary methods we use. The model is specifically-tuned to calculate **Legendary Index** so as to detect **Great Couch Effect** using data of US gymnastics team coached by Bela Karolyi and Marta Karolyi. Lang Ping is successfully detected as a legendary coach, and we've found many more great teams and athletes, even controversial results(2024 Male Fencing,Italy) and decline of the great ranks(2024 Tennis,China).

Greatness is not born, but made. We hope that our model may help countries' Olympic committes to accommodate strategies and achieve better results in the future Olympic Games.

Keywords: Olympic Games; Data Analysis; Stimulation; Machine Learning;

Contents

1 Introduction

1.1 Background and Literature Review

The Olympic Games, often simply referred to as the Olympics, are the world's foremost international multi-sport events. With a history that dates back over 2,000 years to ancient Greece, the modern Olympics were revived in 1896 by Pierre de Coubertin. The Olympics are held every four years, with the Summer Olympics and the Winter Olympics alternating every two years. The Summer Olympics feature a vast array of sports, from athletics and swimming, which are considered the cornerstones of the Games, to more specialized sports like fencing, badminton, and gymnastics. Athletes from around the globe gather to compete at the highest level, showcasing their extraordinary skills, determination, and physical prowess.

Over the years, the Olympics have become a platform for athletes to break records, inspire generations, and promote cultural exchange. They have also had a significant impact on the host cities, driving urban development, improving infrastructure, and boosting the local economy. Despite facing various challenges, including political issues and the impact of global events, the Olympic Games continue to hold a special place in the hearts of people worldwide, symbolizing the power of sports to unite and uplift humanity.

The prediction of Olympic medal counts has long been a topic of interest among sports enthusiasts, statisticians, and researchers. Understanding how to forecast the number of medals a country or athlete might win is not only a matter of curiosity but also has implications for sports management, marketing, and national pride.

Traditional approaches are typically made closer to the start of an upcoming Olympic Games when information about the current athletes scheduled to compete becomes available. This approach allows for a more accurate assessment of a country's or athlete's medal prospects. For example, the virtual medal table forecast by Nielsen [1] provides a more real-time and data-driven prediction. By incorporating current athlete performance, injuries, and recent competition results, these modern models can better capture the dynamic nature of sports.

However, these data may be concealed and intentionally modified by countries' Olympic committees, which may produce misleading predictions.As a result, our model will be based on the historical data of the Olympic Games, which is more reliable and less likely to be manipulated. Research in this area has also explored the use of advanced statistical techniques. Machine learning algorithms, for instance, have been employed to analyze large datasets encompassing a wide range of variables related to athletes, sports, and countries. These algorithms can identify complex patterns and relationships that might not be apparent through traditional statistical methods.

In conclusion, the field of predicting Olympic medal counts has evolved significantly over the years. While historical contemporary methods still provide some basis for understanding trends, the focus has shifted towards historical data. The use of advanced statistical techniques and an increased awareness of external factors have improved the accuracy of these predictions. However, there is still room for further research, particularly in the use of machine learning methods and elements that can't be captured directly by historical data, such as the Great Coach Effect. As the Olympics continue to evolve, so too will the methods and models used to predict the medal counts, ensuring that this remains a vibrant and relevant area of study.

1.2 Restatement of the Problem

Considering the background, in this paper we are required to solve the following problems:

- **Task 1:** Develop a model for medal counts for each country, both **Gold** and **Total**, and use the model to predict various countries' performance in the 2028 Olympics. The model also includes estimates of the precision and measures of how well the model performs.
- **Task 2:** Develop a model for prediction of when a country will win its **first** medal in the Olympics, and the probability of winning it in the coming next LA Olympics. Also, we will evaluate this model.
- **Task 3:** Develop a model for estimation of the **"Great Coach"** effect, and further using this model to identify three countries suitable for imitating this strategy.
- **Task 4:** Calculating the **"Great Athlete"** effect, referring to the phenomenon that some great athletes won a large number of the (Gold) medals of a certain sport or event, and sometimes his/her country's medal tally greatly depend on him/her during his/her athlete career. We will also explain how this can inform country Olympic committees.

2 Assumptions and Notations

2.1 Assumptions and Justifications

To simplify problem, make it convenient to construct simulation model and ensure it's effectiveness, following basic assumptions are made, each of which is properly justified.

- **Data Preprocess and Country Mapping**.

  Data is cleaned and preprocessed to make it more suitable. We set the ice sports aside, combine various teams of one country, and map the countries which no longer exist to the current existing country. It's a rather difficult task since countries have changed a lot during the past century, but it's necessary to stick to the current world map.

- **Relationship between Medal counts and Historical data**.

  Most data provided is examined to be accurate and reliable, although some may not be the same as the data provided from other sources. Few of the data is examined and changed since it's too faraway from reality. Medal counts are also assumed to be related to multiple variables:years, gender, athletes and number of sports, diciplines, events,etc.

- **Countries that will possibly take part in the 2028 Olympics**.

  Countries that have taken part in 2024 Olympics will also take part in 2028 Olympics. There's no reason for them not to attend the Olympics, and it's impossible to estimate medal of new countries if they've never taken part in the Olympics using historical data. Russia is also assumed to participate in the 2028 Olympics, whether the athletes will be able to compete under the Russian flag or not, their country is called Russia.

- **Possible Sports in 2028**.

  There's said to be five new sports in the 2028 Olympics, but we set them aside because we don't have enough data to predict the medal counts of these events. Similarly, we set aside

the sports that will be removed from the Olympics in 2028, and assume that most sports are like that in 2024 or their most common appearance in history(Such as gymnastics).

Other details will be stated in the following sections when we discuss the specific models.

2.2 Notations

Symbols	Definitions	Symbols	Definitions
h	aaaaaaaaaaaaaaaaaa	a	aaaaaaaaaaaaaaaaaaaa
k			
c_p			
ρ			
T			

Define the Main parameters. Other Specific notations will be listed and explained later.

3 Data Preprocessing

3.1 Data Cleaning

3.1.1 Garbled Code

3.1.2 Null, Duplicated Values and Renaming

Listing 1: Data Cleaning

```
data.fillna(0, inplace=True)
data.drop_duplicates(inplace=True)
athletes.rename(columns={'Team': 'Country'}, inplace=True)
```

Team are renamed to Country in the athletes data so that there won't be Germany-1.

3.1.3 Data Splitting

Listing 2: Data Splitting

```
data.fillna(0, inplace=True)
data.drop_duplicates(inplace=True)
athletes.rename(columns={'Team': 'Country'}, inplace=True)
```

3.1.4 Remove ice sports and athletes playing ice sports

Listing 3: Data Splitting

```
ice_sports = ['Figure Skating', 'Ice Hockey']
programs = programs[~programs['Sport'].isin(ice_sports)]
athletes = athletes[~athletes['Sport'].isin(ice_sports)]
```

Before 1924, ice sports were held at the Summer Olympic Games. These two sports made their debuts at 1908 and 1920 Summer Olympics respectively, but in 1924 they were moved to the

first edition of the Winter Olympic Games and became permanent fixtures on the sports program for the Winter Olympics from then on. Since their data are few, we set them aside.

3.2 Country Mapping

4 Model Overview

In order to solve those problems, we will proceed as follows:

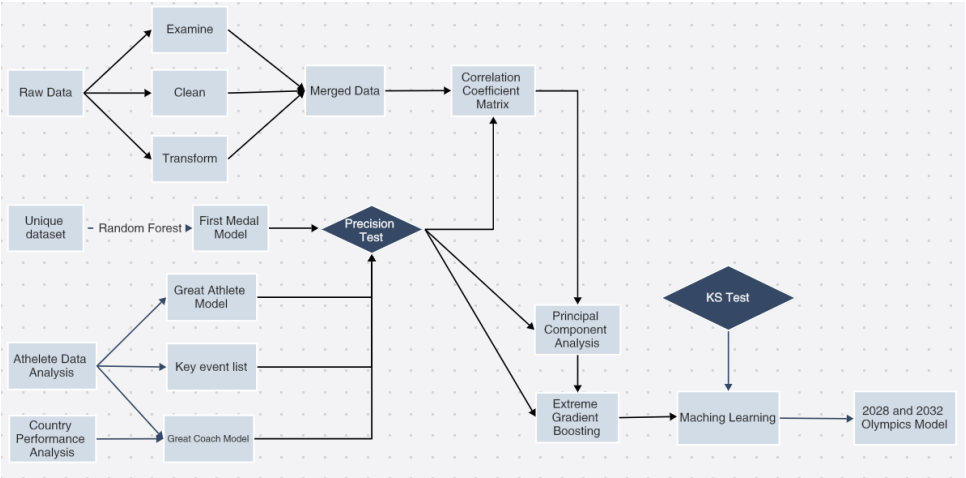

Figure 1: Flow chart of our work

- **Stating assumptions**. By stating our assumptions, we will narrow the focus of our approach towards the problems and provide some insight into Olympic medal issues.
- **Making notations**. We will give some notations which are important for us to clarify our models.
- **Presenting our model**. In order to investigate the problem deeper, we divide our model into three sub-models. One is a **medal tally** prediction model based on past data and host country effect. Another one is a non-linear model focusing on the analysis of data characters of the relevant period when a country won its **first** medal. The last is about certain **key sports/events**, **great coach** and **great athletes**
- **Defining evaluation criteria and comparing sub-models**. We define two main criteria to evaluate our model: the precision of our model and the ability to calculate probability of winning first medal.
- **Analysis of influencing factors**. In term of the impact of different factors on our model, we take those into consideration: medal tally, year and participation.
- **Model testing and sensitivity analysis**. With the criteria defined before, we evaluate the reliability of our model and do the sensitivity analysis.
- **Further discussion**. We discuss about different ways to arrange medal counts. Then we improve our model to apply them in reality.
- **Evaluating the model**. We discuss about the strengths and weaknesses of our model:

1) ...
2) ...
3) ...
4) ...

5 Task I : Predicting the Medal of the 2028 LA Olympics

6 Task II : Predicting First Medal Countries

7 Task III : Great Coach Effect

8 Task IV : Important Events and Great Athletes

9 Other Insights

9.1 Try to predict the medal of the 2032 Olympics

9.2 Greatness and the strategy of the countries' Olympic committees

In this part, we will focus on different distribution of inflow faucets. Then we discuss about the real-life application of our model.

- Different Distribution of Inflow Faucets

  In our before discussion, we assume there being just one entrance of inflow.

  From the simulating outcome, we find the temperature of bath water is hardly even. So we come up with the idea of adding more entrances.

  The simulation turns out to be as follows

  From the above figure, the more the entrances are, the evener the temperature will be. Recalling on the before simulation outcome, when there is only one entrance for inflow, the temperature of corners is quietly lower than the middle area.

  In conclusion, if we design more entrances, it will be easier to realize the goal to keep temperature even throughout the bathtub.

- Model Application

  Our before discussion is based on ideal assumptions. In reality, we have to make some corrections and improvement.

  1) Adding hot water continually with the mass flow of 0.16 kg/s. This way can ensure even mean temperature throughout the bathtub and waste less water.
  2) The manufacturers can design an intelligent control system to monitor the temperature so that users can get more enjoyable bath experience.
  3) We recommend users to add bubble additives to slow down the water being cooler and help cleanse. The additives with lower thermal conductivity are optimal.
  4) The study method of our establishing model can be applied in other area relative to convection heat transfer, such as air conditioners.

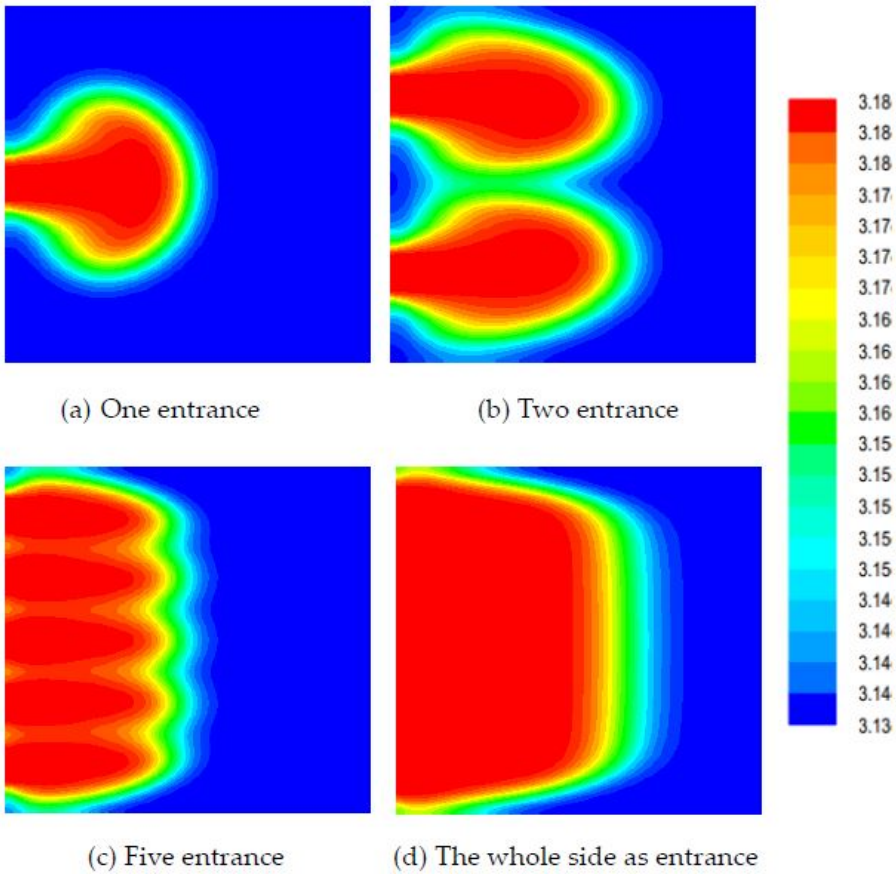

Figure 2: The simulation results of different ways of arranging entrances

10 Strength and Weakness

10.1 Strength

- We analyze the problem based on thermodynamic formulas and laws, so that the model we established is of great validity.
- Our model is fairly robust due to our careful corrections in consideration of real-life situations and detailed sensitivity analysis.
- Via Fluent software, we simulate the time field of different areas throughout the bathtub. The outcome is vivid for us to understand the changing process.
- We come up with various criteria to compare different situations, like water consumption and the time of adding hot water. Hence an overall comparison can be made according to these criteria.
- Besides common factors, we still consider other factors, such as evaporation and radiation heat transfer. The evaporation turns out to be the main reason of heat loss, which corresponds with other scientist's experimental outcome.

10.2 Weakness

- Having knowing the range of some parameters from others' essays, we choose a value from them to apply in our model. Those values may not be reasonable in reality.

- Although we investigate a lot in the influence of personal motions, they are so complicated that need to be studied further.
- Limited to time, we do not conduct sensitivity analysis for the influence of personal surface area.

10.3 Further Discussions

References

[1] Nielsen. *2024 virtual-medal-table-forecast*. https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/. 2024.

Appendices

Appendix A First appendix

In addition, your report must include a letter to the Chief Financial Officer (CFO) of the Goodgrant Foundation, Mr. Alpha Chiang, that describes the optimal investment strategy, your modeling approach and major results, and a brief discussion of your proposed concept of a return-on-investment (ROI). This letter should be no more than two pages in length.

Here are simulation programmes we used in our model as follow.

Input matlab source:

```
function [t,seat,aisle]=OI6Sim(n,target,seated)
pab=rand(1,n);
for i=1:n
    if pab(i)<0.4
       aisleTime(i)=0;
    else
        aisleTime(i)=trirnd(3.2,7.1,38.7);
    end
end
```

Appendix B Second appendix

some more text **Input C++ source:**

```
//============================================================================
// Name        : Sudoku.cpp
// Author      : wzlf11
// Version     : a.0
// Copyright   : Your copyright notice
// Description : Sudoku in C++.
//============================================================================

#include <iostream>
#include <cstdlib>
#include <ctime>

using namespace std;

int table[9][9];

int main() {

    for(int i = 0; i < 9; i++){
        table[0][i] = i + 1;
    }
```

```
    srand((unsigned int)time(NULL));

    shuffle((int *)&table[0], 9);

    while(!put_line(1))
    {
        shuffle((int *)&table[0], 9);
    }

    for(int x = 0; x < 9; x++){
        for(int y = 0; y < 9; y++){
            cout << table[x][y] << " ";
        }

        cout << endl;
    }

    return 0;
}
```

Report on Use of AI

1. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4,)

   Query1: <insert the exact wording you input into the AI tool>

   Output: <insert the complete output from the AI tool>

2. OpenAI Ernie (Nov 5, 2023 version, Ernie 4.0)

   Query1: <insert the exact wording of any subsequent input into the AI tool>

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

3. Github CoPilot (Feb 3, 2024 version)

   Query1: <insert the exact wording you input into the AI tool>

   Output: <insert the complete output from the AI tool>

4. Google Bard (Feb 2, 2024 version)

   Query1: <insert the exact wording of your query>

   Output: <insert the complete output from the AI tool>