**Report on Use of AI**

1. OpenAI *ChatGPT* (May 14, 2024 version, ChatGPT-4o)

Query1: *Request on translation of problem*

Output1: *(in Chinese)*

2025 MCM 问题 C: 奥运奖牌榜模型

任务背景：

2024年巴黎夏季奥运会结束后，全球体育迷们不仅关注各项比赛的结果，还热衷于分析各国的 奖牌榜。2024年奥运会的奖牌榜数据显示，美国以126枚奖牌位居总数榜首，中国与美国并列以40枚金牌排名第一。法国作为东道国，在金牌榜上排名第五，但在总奖牌数上排名第四。而英国则凭借14枚金牌，在总奖牌榜上排名第三。

奖牌榜的排名总是受到高度关注，但其他一些小国家的奖牌数同样有其重要性。例如，阿尔巴尼亚、佛得角、多米尼加和圣卢西亚在巴黎奥运会上获得了他们国家的首枚奖牌，其中多米尼加和圣卢西亚分别获得了金牌。全球还有60多个国家尚未赢得奥运奖牌。

预测未来奥运奖牌数量是一项常见的活动，通常基于运动员的表现和历史数据，而不仅仅是历史奖牌统计。对于即将到来的2028年洛杉矶夏季奥运会，预测的奖牌榜尤其引人注目。

数据：

提供了历届夏季奥运会的 奖牌榜、 东道国、以及每届奥运会各项比赛（按运动项目划分）的数据。

还提供了 所有奥运运动员 的数据，包括他们的运动项目和成绩（奖牌类型或无奖牌）。

模型要求：

奖牌数模型：

开发一个 奖牌数预测模型，至少包括金牌和总奖牌数。

评估模型预测的 不确定性/精度，并度量模型的性能。

基于该模型，预测 2028年洛杉矶奥运会 的奖牌榜，包含预测区间。哪些国家有可能表现更好？哪些国家会比2024年表现差？

模型应包括尚未获得奖牌的国家，预测哪些国家将在下一届奥运会获得他们的首枚奖牌，并评估这些估算的 概率。

模型应考虑 比赛数量和类型 对奖牌数的影响。探索不同比赛项目之间的关系，哪些项目对不同国家更为重要？主办国选择的比赛项目如何影响结果？

“伟大教练”效应：

运动员可能为不同国家参赛，但由于国籍要求，他们不容易更换国家。然而，教练可以从一个国家转到另一个国家，因为他们不需要是公民。某些教练可能会带来 “伟大教练”效应，比如郎平和贝拉·卡罗伊。分析数据，探讨是否存在这种效应，并估算其对奖牌数的影响。

选择三个国家，分析他们在哪些运动项目上应投资“伟大教练”，并估算这一投资的潜在影响。

其他见解：

模型应揭示一些 原创性见解，并说明这些见解如何为各国的奥委会提供有价值的参考。

提交要求：

一页摘要。

目录。

完整解决方案（包括模型、假设、结果等）。

参考文献列表。

AI使用报告（如使用了AI，这部分不计入总页面限制）。

数据文件：

data\_dictionary.csv：数据库描述与示例

summerOly\_athletes.csv：包含所有运动员、运动项目、年份和成绩（奖牌或无奖牌）

summerOly\_medal\_counts.csv：1896年至2024年间所有夏季奥运会的国家奖牌统计

summerOly\_hosts.csv：所有夏季奥运会的东道国列表

summerOly\_programs.csv：所有夏季奥运会的各项比赛数量（按运动项目分类）

术语表：

国际奥委会（IOC）：是负责组织夏季和冬季奥运会的国际非政府体育管理机构。

项目（Program）：奥运会的所有运动竞赛项目，由IOC为每届奥运会确定。

运动（Sport）：IOC定义的奥林匹克运动，通常由一个国际运动联合会（IF）管理。

学科（Discipline）：一个运动分支，包含一个或多个竞赛项目。

竞赛项目（Event）：每个学科中的具体赛事。

参考文献：

奥运会官网奖牌榜

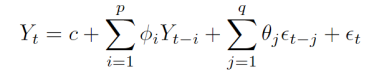
郎平简介

贝拉·卡罗伊简介

这个问题要求你开发一个数学模型，用于预测奥运奖牌的分布，分析各国的奖牌潜力，并为未来奥运会的奖牌分布提供有价值的预测与建议。

Query2: In problem (2), we will use ARIMA model and try to give the formula of ARIMA model and its applicable conditions.

Output2:

ARIMA (Autoregressive Integrated Moving Average) is a widely used time series analysis method suitable for processing time-dependent continuous data. The form of the ARIMA model is: 

The ARIMA model provides more accurate predictions by adjusting the values of p and q to capture autocorrelations and random fluctuations in time series. The goal of this logistic regression model is to learn the regression coefficients β0, β1,..., βn from the given training data. The learning process of regression coefficients is achieved by maximizing the likelihood function. The purpose of maximizing the likelihood function is to make the probability value predicted by the model as close as possible to the actual observed label.

Query3: The following is a summary of model fitting for a Poisson regression model. Use the data to generate a sensitivity heat map and a sensitivity uncertainty map for the model (use python’s matplotlib). Note that all the data you need is in this text. *Following is the model abstract*

Output3:

To generate sensitivity heat maps and sensitivity uncertainty maps, it is first necessary to clarify how to extract information from the provided Poisson regression model results. Sensitivity heat map: This graph usually shows the size of the influence of each independent variable on the target variable. From the regression coefficient (coef) and its standard error (std err), we can calculate the effect of each coefficient. We will use the standard error of the regression coefficient to measure the uncertainty. Sensitivity Uncertainty plot: This is a plot that shows the sensitivity of the model's output to the input parameters, and is usually plotted based on the confidence interval ([0.025, 0.975]) of the regression coefficient. Here is the Python code that generates these two diagrams:

“””

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import seaborn as sns
6. # 模型系数和标准误差
7. coefficients = {
8. 'Intercept': 6.5920,
9. 'C(Sport\_FE)[T.Gymnastics]': 0.2366,
10. 'C(Sport\_FE)[T.Volleyball]': 1.0672,
11. 'Coach': -0.0910,
12. 'Year\_FE': -0.0026
13. }
15. std\_errors = {
16. 'Intercept': 0.585,
17. 'C(Sport\_FE)[T.Gymnastics]': 0.253,
18. 'C(Sport\_FE)[T.Volleyball]': 0.256,
19. 'Coach': 0.197,
20. 'Year\_FE': 0.000
21. }
23. # 系数的置信区间 [0.025, 0.975]
24. ci\_lower = {
25. 'Intercept': 5.446,
26. 'C(Sport\_FE)[T.Gymnastics]': -0.259,
27. 'C(Sport\_FE)[T.Volleyball]': 0.566,
28. 'Coach': -0.478,
29. 'Year\_FE': -0.003
30. }
32. ci\_upper = {
33. 'Intercept': 7.738,
34. 'C(Sport\_FE)[T.Gymnastics]': 0.732,
35. 'C(Sport\_FE)[T.Volleyball]': 1.568,
36. 'Coach': 0.296,
37. 'Year\_FE': -0.002
38. }
40. # 将数据整理为 DataFrame
41. df = pd.DataFrame({
42. 'Coefficient': coefficients.values(),
43. 'Std\_Error': std\_errors.values(),
44. 'CI\_Lower': ci\_lower.values(),
45. 'CI\_Upper': ci\_upper.values()
46. }, index=coefficients.keys())
48. # 灵敏度热力图
49. sns.set(style="whitegrid")
50. plt.figure(figsize=(8, 6))
51. sns.heatmap(df[['Coefficient']], annot=True, cmap="coolwarm", cbar\_kws={'label': 'Coefficient Value'})
52. plt.title('Sensitivity Heatmap: Coefficients')
53. plt.tight\_layout()
54. plt.show()
56. # 灵敏度-不确定性图
57. plt.figure(figsize=(8, 6))
58. for feature in df.index:
59. plt.plot([df.loc[feature, 'CI\_Lower'], df.loc[feature, 'CI\_Upper']], [feature, feature], label=feature, marker='o')
60. plt.scatter(df.loc[feature, 'Coefficient'], feature, color='black')
62. plt.xlabel('Coefficient Value')
63. plt.ylabel('Features')
64. plt.title('Sensitivity-Uncertainty Plot')
65. plt.grid(True)
66. plt.tight\_layout()
67. plt.show()

“””

Sensitivity heatmap: seaborn.heatmap is used to draw the heatmap of the regression coefficient, showing the influence of each independent variable on the target variable. Through Coefficient column, the heat map shows the influence degree of each independent variable. Sensitivity Uncertainty plot: This is a representation of the sensitivity and uncertainty of the model output through the confidence interval of the regression coefficient. The confidence interval for each variable is represented by a line at either end, while the position of the regression coefficient is indicated by a black dot. Sensitivity thermogram: Shows the magnitude of the regression coefficient to help identify which variables have the greatest influence on the target variable. Sensitivity Uncertainty plot: Confidence intervals and actual estimates for regression coefficients are shown to help understand the uncertainty of these estimates. This code will generate the corresponding visualization based on the coefficient data of the Poisson regression model that you provided.

1. Kimi (Jan 20, 2025 version, k1.5)

Kimi is a LLM that experts in reading long text files. Therefore, we took advantage of it to read long technological essays.

Query1: Please analyze this article introducing the random forest model and answer my questions about the applicable conditions of the random forest model. And how to use the model correctly?

Output1:

**随机森林模型的适用条件**

随机森林模型适用于以下几种情况：

* **大数据集**：随机森林在处理大数据时表现优异，能够高效地处理大量的数据。
* **高维数据**：随机森林可以处理多达几千个解释变量，适用于高维数据集。
* **非线性关系**：随机森林能够捕捉变量之间的非线性关系，适用于变量之间存在复杂非线性关系的数据。
* **变量交互作用**：随机森林可以建模变量之间的交互作用，适用于变量之间存在交互作用的数据。
* **离群值不敏感**：随机森林对离群值不敏感，适用于数据中存在离群值的情况。
* **不需要变量选择**：随机森林不需要进行变量选择，适用于变量较多且难以选择的情况。

**如何正确使用随机森林模型**

1. **数据准备**：
   * **数据清洗**：处理缺失值，可以使用随机森林自带的缺失值处理方法（如中位数或众数填充）。
   * **数据标准化**：虽然随机森林对数据的分布要求不高，但标准化可以提高模型的稳定性和预测精度。
2. **模型参数设置**：
   * **ntree**：指定生成的分类树的数量。一般情况下，分类树的数量越大越好，但会增加计算负荷。可以通过交叉验证或袋外误差来确定最优的树的数量。
   * **mtry**：指定每个节点用来二分数据的自变量的个数。对于分类问题，默认值是自变量总数的平方根；对于回归问题，默认值是自变量总数的1/3。可以根据具体问题调整mtry的值，以优化模型性能。
3. **模型训练**：
   * 使用随机森林算法对数据进行训练，生成模型。可以使用R语言中的randomForest包或其他支持随机森林的工具。
4. **模型评估**：
   * **误差评估**：随机森林通过袋外误差（out-of-bag error）来评估模型的误差。对于分类问题，误差是分类的错误率；对于回归问题，误差是残差的方差。
   * **变量重要性**：随机森林可以评估每个自变量对因变量的重要性，可以通过importance参数来获取变量的重要性。
5. **模型优化**：
   * **参数调优**：通过调整ntree和mtry等参数，优化模型的性能。
   * **交叉验证**：使用交叉验证来评估模型的泛化能力，避免过拟合。
6. **模型应用**：
   * **预测**：使用训练好的模型对新的数据进行预测。
   * **解释**：通过变量重要性和部分依赖图等工具，解释模型的结果，理解变量对因变量的影响。

**注意事项**

* **类别不平衡**：随机森林在处理类别不平衡的数据时，可能会倾向于多数类。可以通过调整样本权重或使用过采样/欠采样方法来解决。
* **模型解释**：虽然随机森林的预测性能很好，但模型的解释性相对较弱。可以通过部分依赖图等工具来辅助解释模型结果。
* **计算资源**：随机森林需要较大的计算资源，尤其是在处理大数据集和高维数据时。确保有足够的计算资源来运行模型。

通过以上步骤和注意事项，可以正确地使用随机森林模型，充分发挥其在分类和回归分析中的优势。

1. Tabnine

*Auto-completions for code used in preparing our models.*