数学建模美赛C题编程工作报告

一、工作背景

在2025年数学建模美赛中，我们选择了C题——“Models for Olympic Medal Tables”。本题要求我们利用提供的数据文件，开发一个模型来预测各国在奥运会中的奖牌数量，并分析影响奖牌数量的因素。根据题目要求，所有分析必须仅基于提供的数据文件，这要求我们在数据处理、特征工程和模型构建过程中严格遵循数据限制。

二、数据处理与特征工程

（一）数据预处理

1. 数据读取与清洗

• 使用pandas库读取了所有提供的数据文件：

• summerOly\_medal\_counts.csv：包含所有夏季奥运会的奖牌统计。

• summerOly\_hosts.csv：包含历届奥运会的主办国信息。

• summerOly\_programs.csv：包含各届奥运会的项目设置。

• summerOly\_athletes.csv：包含所有参赛运动员的信息及其比赛结果。

• 检查并处理了数据中的缺失值和异常值。例如，summerOly\_programs.csv中存在缺失值，通过机器学习模型进行了填充。

• 确保所有数据的完整性和准确性，为后续分析打下基础。

2. 特征提取

• 国家级特征：

• 历史奖牌总数、金牌数、奖牌增长率（基于summerOly\_medal\_counts.csv）。

• 是否为主办国（基于summerOly\_hosts.csv）。

• 运动员级特征：

• 运动员的历史成绩、参赛次数、项目类型（基于summerOly\_athletes.csv）。

• 项目特征：

• 每个项目的参赛国家数量、项目的变化趋势（基于summerOly\_programs.csv）。

（二）特征工程

• 构建了特征矩阵，将国家级特征、运动员级特征和项目特征整合为模型输入。

• 特别关注了项目类型和主办国效应，将其作为重要特征纳入模型。

三、模型构建与实现

（一）动态预测模型

1. 机器学习模型：

• 使用XGBoost实现奖牌预测模型，输入包括国家级特征、运动员级特征和项目特征。

• 通过交叉验证评估模型性能，优化超参数。

2. 贝叶斯模型：

• 使用历史奖牌数据作为先验分布，结合运动员特征进行贝叶斯更新。

• 输出预测的置信区间，评估不确定性。

（二）时间序列趋势分析

• 使用LSTM + Transformer模型，基于summerOly\_medal\_counts.csv中的历史奖牌数据，捕捉奖牌总数的整体增长趋势。

• 输出趋势曲线，作为背景参考。

（三）综合集成模型

• 将运动员驱动模型和时间序列趋势模型的输出按权重融合（80%运动员驱动模型 + 20%时间序列趋势模型）。

• 生成最终的奖牌预测结果。

四、模型评估与优化

（一）评价指标

• 使用RMSE、MAE、精度与召回率等指标评估模型性能。

• 引入Top-N准确率和Kendall-Tau相关系数，评估预测排名的准确性。

（二）不确定性分析

• 使用贝叶斯框架或蒙特卡洛方法生成预测的置信区间。

• 分析预测结果的不确定性来源。

（三）敏感性分析

• 使用SHAP或LIME工具分析特征对预测结果的重要性。

• 探讨不同特征（如运动员历史成绩、项目类型等）对预测结果的影响。

五、主场效应与教练效应分析

（一）主场效应

• 使用summerOly\_hosts.csv中的主办国信息，分析主办国在奖牌数量上的变化趋势。

• 通过统计方法验证主场优势的存在性。

（二）教练效应

• 从summerOly\_athletes.csv中提取教练信息，分析教练国籍与运动员成绩的关系。

• 选择三个国家，分析引入“伟大教练”后奖牌数量的变化。

六、总结与展望

（一）今日工作总结

• 完成了数据预处理、特征工程和动态预测模型的初步实现。

• 实现了时间序列趋势分析，并构建了综合集成模型。

• 进行了初步的模型评估和不确定性分析。

（二）后续工作计划

• 进一步优化模型结构，提升模型性能。

• 完成主场效应与教练效应的量化分析。

• 撰写完整的解决方案报告，并整理代码。

（三）注意事项

• 严格遵循题目对数据文件的限制，确保所有分析仅基于提供的数据。

• 定期与团队成员沟通，确保模型实现与论文撰写同步进行。

----

七、C.pdf文件内容详细解读

1. 文件内容概述

C.pdf文件包含了一系列Python代码片段和数据处理流程，主要目的是对2025年数学建模美赛C题提供的数据文件进行预处理、分析和初步建模。文件内容分为以下几个部分：

(1) 数据读取与编码检测

• 使用chardet库检测CSV文件的编码格式，并通过pandas库读取数据。

• 读取了以下五个数据文件：

• data\_dictionary.csv：数据字典，包含变量解释和示例。

• summerOly\_medal\_counts.csv：夏季奥运会奖牌统计。

• summerOly\_hosts.csv：历届奥运会主办国信息。

• summerOly\_programs.csv：各届奥运会的项目设置。

• summerOly\_athletes.csv：参赛运动员信息及其比赛结果。

(2) 缺失值检查与处理

• 编写了check\_missing\_values函数，检查每个文件的缺失值情况。

• 发现summerOly\_programs.csv文件存在缺失值，其他文件数据完整。

• 对summerOly\_programs.csv中的缺失值进行了处理：

• 使用线性回归、K近邻回归或随机森林回归对缺失的“Events”值进行预测填充。

(3) 数据转换与特征工程

• 对summerOly\_programs.csv文件进行了长格式到宽格式的转换，便于后续分析。

• 对奖牌数据进行了透视表处理，生成了按年份和国家划分的奖牌统计表。

• 对运动员数据进行了宽格式转换，统计了运动员连续参加奥运会的年数和参赛年份间隔。

(4) 初步模型实现

• 使用RandomForestRegressor、LinearRegression和KNeighborsRegressor对缺失数据进行了预测填充。

• 实现了奖牌数据的透视表生成，并将其保存为CSV文件。

• 对运动员数据进行了初步分析，包括连续参赛年数和参赛间隔的统计。

2. 文件内容的详细解读

(1) 数据读取与编码检测

代码中使用了chardet库来检测文件的编码格式，确保在读取时不会出现编码错误。这一部分的代码展示了如何处理不同编码格式的文件，例如：

def read\_csv\_with\_detected\_encoding(file\_path):

with open(file\_path, 'rb') as f:

result = chardet.detect(f.read())

encoding = result['encoding']

return pd.read\_csv(file\_path, encoding=encoding)

这一部分的目的是确保数据的正确读取，为后续分析打下基础。

(2) 缺失值检查与处理

在summerOly\_programs.csv文件中发现了缺失值，代码中使用了以下方法进行处理：

• 填充策略：对于“Discipline”列的缺失值，使用“Sport”列的值进行填充。

• 预测填充：对于“Events”列的缺失值，根据年份和运动项目，使用线性回归、K近邻回归或随机森林回归进行预测填充。

data['Discipline'] = data['Discipline'].fillna(data['Sport'])

model = RandomForestRegressor(n\_estimators=100, random\_state=42)

model.fit(X\_known, y\_known)

predicted\_events = model.predict(X\_missing)

这一部分展示了如何处理缺失值，确保数据的完整性。

(3) 数据转换与特征工程

代码中对数据进行了多种转换和处理，以便于后续的分析和建模：

• 长格式到宽格式的转换：将summerOly\_programs.csv文件从长格式转换为宽格式，便于分析各届奥运会的项目设置。

• 奖牌数据的透视表生成：生成了按年份和国家划分的奖牌统计表，并将其保存为CSV文件。

• 运动员数据的宽格式转换：统计了运动员连续参加奥运会的年数和参赛年份间隔。

data\_long = data.melt(id\_vars=['Sport', 'Discipline', 'Code', 'Sports Governing Body'],

value\_vars=years, var\_name='Year', value\_name='Events')

pivot\_df = data.pivot\_table(index=['Name', 'Sex', 'Team', 'NOC', 'City', 'Sport', 'Event'],

columns='Year', values='Medal', aggfunc='first').reset\_index()

这些操作展示了如何通过数据转换提取有价值的特征。

(4) 初步模型实现

代码中实现了以下模型：

• 随机森林回归：用于预测缺失的“Events”值。

• 线性回归和K近邻回归：在数据量较小时用于预测缺失值。

• 奖牌数据的透视表生成：生成了金、银、铜和总奖牌数量的表格，并将其保存为CSV文件。

model = RandomForestRegressor(n\_estimators=100, random\_state=42)

model.fit(X\_known, y\_known)

predicted\_events = model.predict(X\_missing)

这些模型的实现展示了如何利用机器学习方法处理数据中的缺失值。

3. 文件内容的总结

C.pdf文件展示了从数据读取、预处理、特征工程到初步建模的完整流程。文件中的代码片段和分析方法为解决2025年数学建模美赛C题提供了重要的基础。通过这些代码，我们可以：

• 确保数据的正确读取和完整性。

• 提取有价值的特征，为后续建模提供支持。

• 使用机器学习方法处理缺失值，提升数据质量。

• 生成奖牌统计表和运动员参赛信息，为进一步分析提供数据支持。

----

八、今日工作报告总结

1. 工作内容回顾

• 数据预处理：完成了数据的读取、编码检测、缺失值检查与处理。

• 特征工程：提取了国家级特征、运动员级特征和项目特征。

• 初步建模：实现了动态预测模型和时间序列趋势分析。

• 模型评估：进行了初步的模型评估和不确定性分析。

2. C.pdf文件内容的贡献

C.pdf文件中的代码和分析方法为今日的工作提供了重要的参考：

• 数据处理：通过编码检测和缺失值处理，确保了数据的完整性和准确性。

• 特征提取：通过数据转换和透视表生成，提取了有价值的特征。

• 初步建模：通过机器学习方法处理缺失值，为后续建模提供了支持。

3. 后续工作计划

• 模型优化：进一步优化模型结构，提升预测精度。

• 主场效应与教练效应分析：量化主场效应和教练效应的具体贡献。

• 报告撰写：整理分析结果，撰写完整的解决方案报告。

4. 注意事项

• 数据限制：严格遵循题目要求，仅使用提供的数据文件。

• 团队协作：与建模手和论文手保持密切沟通，确保工作同步进行。

----