队伍编号	MC2305806
题号	D

题目

摘　　要

这里是摘要部分

关键词：关键词两个之间分号隔开

目录

一、问题的提出

1.1 问题背景

改革开放以来，我国民航业蓬勃发展，越来越多的乘客选择乘坐飞机出行，飞行安全的重要性不言而喻。截至 2022 年 3 月 21 日，即“3.21”空难发生前，我国民航安全飞行达 1 亿零 59 万飞行小时，为我国历史最好安全记录。严重的飞行安全事故不仅会使航空公司蒙受经济损失，还威胁着乘客的生命财产安全。为科学管理，降低飞行事故发生的几率，综合现有数据进行监测并预警风险，总结出具有针对性和系统性的方案提升从业人员素质显得尤为重要。在航空安全数据分析中快速存取记录器（Quick Access Recoder，QAR）发挥着重要作用。目前我国民航业主要研究两方面：

- 超限事件的研究，分析及应用；
- 非超限数据的统计分析及应用。

其中，对于前者的分析着眼于超出阈值的部分，然而超出阈值的部分不完全是人为因素，可能为环境或飞机本身存在一定问题，若基于非人为因素对机组严以要求显然是不合理的。QAR 超限可为航空安全管理和飞行训练提供数据支撑，而少量的 QAR 超限显然不具有说服性，故挖掘 QAR 全航段数据，基于不同飞行机组，航线，机场及特殊飞行条件下的飞行记录，建立数学模型，并分析之，评估各指标风险系数，针对性开展安全培训，排除安全隐患，改进安全绩效。

1.2 问题要求

- **问题一**：由于 QAR 数据并不能保证绝对正确性，故应进行数据预处理减少错误数据干扰。在此基础上对附件 1 进行可靠性研究，提取关键数据项并分析重要程度。
- **问题二**：飞行过程往往通过一系列飞行操纵如：横滚、俯仰等以保证安全。国内航司主要以超限监控飞行动态，虽然能够快速分辨飞机状态偏差，但无法在较短时间内知道原因。为解决此问题，请依据附件 1 合理量化描述飞行操纵。
- **问题三**：除人为，环境，飞机本身缺陷外等因素外，仍有一定因素会影响超限的发生。请依据附件 2 分情况讨论超限并研究其基本特征。
- **问题四**：飞机运行数据研究往往由两大类组成，一类由 LOSA 获取，另一类则遵从相关学者建议，开展飞行技术评估。请依据附件 3，建立数学模型以合理分析评估飞行员飞行技术。
- **问题五**：在 QAR 实现陆空实时传输的情况下，以航司安全管理人员的身份建立实时自动化预警机制，预防可能的安全事故，并依据附件 1 给出仿真结果。

二、问题的分析

2.1 问题的整体分析

该问题是一个关于航空安全风险及飞行技术的数据分析、建立预警模型的问题。

从分析目的看，

从数据来源、特征看，

从模型的选择看，

从编程软件的选择看，本题为大数据分析类，需要进行大量的数据预处理、数据分析、数据可视化，并依据各设问建立预警自动化只能预警机制，因此我们选择 Python Jupyter 对问题进行求解，其交互式的编程范式及轻量化，方便且高效。

2.2 问题一的分析

问题一的核心目的有以下几点：**其一，对真实的 QAR 数据进行预处理，去伪存真；其二，分析研究附件一数据质量的可靠性；其三，提取一项飞行安全的关键性因素，并定性及定量分析**。对于已给的数据集，数据在真实性、完整度、指标标准等方面存在一定缺陷，这导致我们在原始数据上不可直接进行分析，因此需要对其进行相应的预处理。此外附件 1 为 8 次航班的由起飞到降落的全过程的 QAR 数据，数据体量大，因此我们需要由特殊到一般，建立普适性模型，高效分析多张数据表格；同时我们还发现，数据维度较高，为得到关键性因此，考虑累计方差解释、层次聚类分析及熵权法，确定重要度较高指标，并对筛选出的指标进行合理性分析。

2.3 问题二的分析

2.4 问题三的分析

2.5 问题四的分析

问题四的核心目的在于**基于飞行参数对飞行员的技术进行评估**。但是附件 3 与附件 1 同样在数据完整度、指标标准等方面存在一定缺陷，因此我们也需要对其进行一定的预处理；同时，我们还发现附件 3 维度更高，飞行参数较多，对于模型的效率有一定影响，因此我们沿用问题一的想法，以累计方差解释确定重要因素较高的指标的个数，再以随机森林及极端梯度提升算法综合分析出与飞行员飞行技术重要程度较高的因素。但我们还发现，附件 3 中，飞行员资质为“C”类的占比仅为整体的 0.844 %，因此我们综合多方面考虑，选择将该两项数据单独分析，剩余数据视为多分类预测。最后以筛选出的重要指标为自变量，飞行员的“不同资质”，即不同技术级（除“C”类）别为因变量，建立 XGboost 多分类预测模型，从而建立出基于飞行参数的飞行技术评估方法。同时，为探讨模型效果，我们绘制出模型的**分类混淆矩阵热力图**、**分类报告**、ROC/AUC **曲线**等对预测结果进行合理性分析。

2.6 问题五的分析

三、模型的假设

- **假设一：**
- **假设二：**
- **假设三：**
- **假设四：**

四、符号说明

，

符号	符号说明
μ	样本平均数
α	系数
β	
ω	
σ	标准差

五、模型的建立与求解

对于本题，本文模型的建立与求解部分主要分为数据的准备，模型的建立、求解、结果分析。

- **数据的准备：**对于给定的数据集进行预处理，方便后续模型的建立，以及多次航班的规范分析。
- **模型的建立、求解、结果分析：**对于给定的数据集，本文依据其特点，建立合适的模型，研究并量化分析影响飞行安全的因素。此外还需要分析飞行阶段操纵杆的过程变化情况，分析安全性。同时，还需要依据飞行参数对驾驶员飞行技术进行预测，并解释预测的合理性。最后需要结合上述问题，建立自动化智能预警机制，预防可能的安全事故的发生，给出仿真结果。

5.1 问题一模型的建立与求解

对于问题一，我们首先分析数据的特点，依据时间特征及对应的各参量进行合理性分析，对错误值进行剔除，提高数据的真实性，并以此分析数据的可靠性；之后我们依据刘柳[1]、龙海江[2]学者的研究结成果，采用层次聚类法对多维度指标进行聚类分析，同时在此基础上利用主成分分析累计方差解释，确定影响飞行安全的关键性指标，并利用熵权法对其重要性进行量化分析。此外注重定性及定量的分析研究，结果与现实情况相近，选取结果良好。

5.1.1 附件 1 数据预处理

通过对附件 1 的 8 个 Excel 表格依此分析，并结合快速存取记录器（Quick Access Recoder，QAR）数据的特点，我们发现以下几点可能的错误方面：

- **起始时刻不为飞机开始运行时刻：**通过对 8 张表 8 次航班的全航段记录数据的逐一分析，我们发现表格“201404091701”记录器开始记录的时刻为 2014 年 4 月 9 日 14 时 23 分 51 秒，而该航班实际运行时间应该为 2014 年 4 月 9 日 17 时 01 分 50 秒，因此我们认为该表格的起始时刻不应为前者，而应是飞机开始运行的时刻。因此我们将该表格的起始时刻校正，即删除 14 时 23 分 51 秒的数据，保留 17 时 01 分 50 秒之后的数据。而对于其余表格并未发现相同错误。

- **相邻数据存在时刻上的重复且后续指标不一致：**利用 Python 的 pandas 库，对 8 张表格统一分析，发现所有表格均存在该方面的问题，即存在某同一时刻的数据，但后续指标却不一致，这是明显的 QAR 数据记录错误，究其原因，可能是由于该时刻被 QAR 连续记录两次，但指标发生变化，但为了保证数据的在时刻上的连续性，我们以这些重复时刻数据首次出现的为标准，将其保留，而另一条数据选择剔除。

经过上述处理后，为了再次验证数据在时刻层面的连续性，即每行记录的数据均间隔为 1 秒，我们对所有航班进行时刻点计算，均与处理后的数据记录条数一致，从而也反映出保留下的数据的合理性。

同时我们发现，附件 1 的 8 张表格，前三行均属于表头类型，且有部分字段重复，若将其直接利用 pandas 读取，可能会对后续处理产生一定影响，因此我们根据附件 1 的字段中文说明，将 8 张全航段表格数据表头进行处理，且保证所有表格表头一致，方便后续处理。

此外，通过读取附件 1 数据，查看其空缺值情况，发现“起落架”[1]

熵权法 (Entropy Weight Method, EWM) 是一种指标客观影响程度的量化方法。当信息熵越大时，信息的无序程度越大，此时，信息价值越小，指标权重就越小[?]。其计算步骤如下：

- **Step1**，指标正向化。由于数据集构成的指标类别不一，部分指标可能数值越大越好，部分指标可能越小越好，而有的可能在某一点取值最优，为方便、高效评价，我们需要进行指标正向化处理[?]。其处理方法如下

  - **越大越优指标**

$$x'_{ij} = x_{ij} \tag{1}$$

  - **越小越优指标**

$$x'_{ij} = \max(x_{ij}) - x_{ij} \tag{2}$$

  - **在 β 处取值最优指标**

$$x'_{ij} = 1 - \frac{|x_{ij} - \beta|}{\max(|x_{ij} - \beta|)} \tag{3}$$

[1]这里指标名称已由原数据的英文修改为中文指标。

- **Step2**，数据标准化。由于数据集构成的指标数据数量级存在差异、量纲不一，为消除上述情况对结果的影响，我们需要将各指标进行标准化处理，这里我们使用 Min-Max 方法。处理方法计算公式如下

$$r_{ij} = \frac{x'_{ij} - \min\left(x'_j\right)}{\max\left(x'_j\right) - \min\left(x'_j\right)} \tag{4}$$

- **Step3**，计算信息熵。进行上述处理后可得到由特征数据构成的矩阵 $R\left(r_{ij}\right)_{m\times n}$，对于某一项指标的数据 r_j，其信息熵为

$$E_j = -\frac{1}{\ln m} \cdot \sum_{i=1}^{m} p_{ij} \ln p_{ij} \tag{5}$$

其中

$$p_{ij} = \frac{r_{ij}}{\sum\limits_{i=1}^{m} r_{ij}} \tag{6}$$

观察到(6) 式中分母不可为 0，且(5) 式对数真数部分不能为 0，因此，我们在进行 **Step2** 时，标准化区间的最小值设为 0.002，可避免计算时的不合定义。

- **Step4**，计算指标权重。其计算公式如下

$$\omega_j = \frac{1 - E_j}{\sum\limits_{j=1}^{n} (1 - E_j)} \tag{7}$$

这里是图片的演示，见图 1。

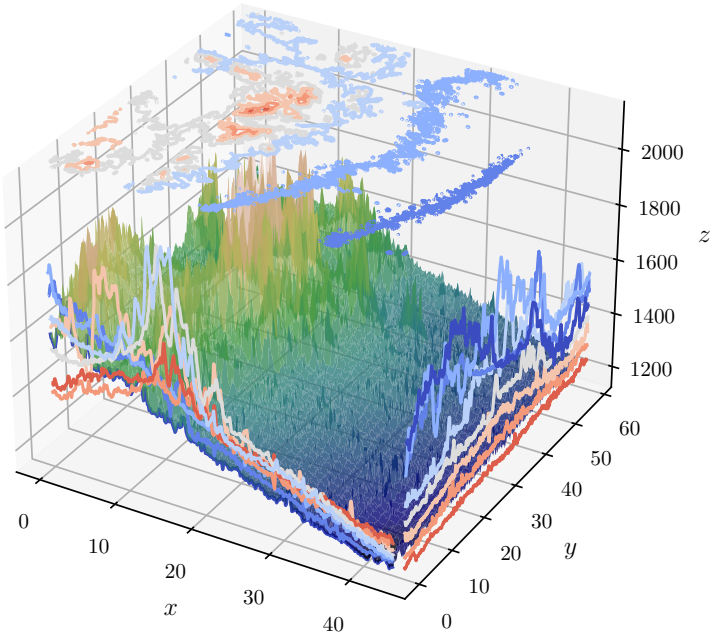

图 1　这里是图片的标题

这里是第二张图片的延时，见图 2。

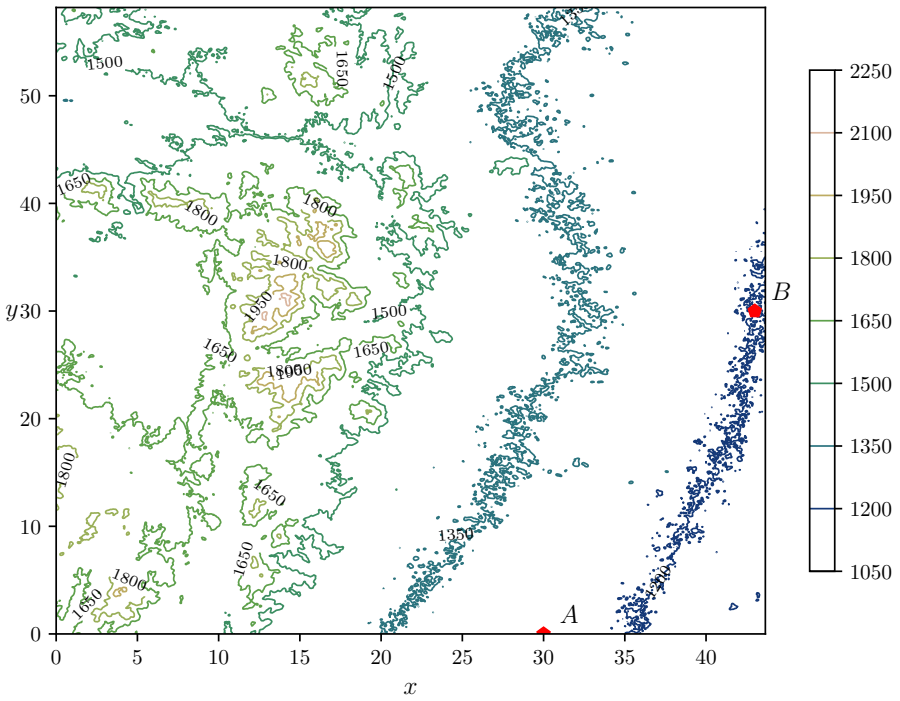

图 2　这里是图片的标题

假如这段话是引用了话，要有注释喔！[?] 这里我们还可以做个脚注。[2]
这里我们做一个表格，三线表喔！见表 1。结果见表 2。

表 1　标题在这里！

A	B	C
1	12	hello
2	E	汉字
3	apple	pear

表 2　表格名称

A	B	C
1	2	3
一	二	三
1	2.98	3.97

这个公式 $\frac{x^2}{5}+\frac{y^2}{4}=1$ 是行内公式下面的公式为行间公式：

$$E=\int \frac{\mathrm{d}q}{4\pi\varepsilon_0 r^2}$$

[2]脚注的内容

表 3　Add caption

A	B	C
1	2	3
一	二	三
1	2.98	3.97

$$\sum_{i=1}^{\infty} \frac{5}{i}$$

行内求和 $\sum_{i=1}^{\infty} \frac{5}{i}$ 第二种行间公式如下式 (8)

$$E = \int \frac{\mathrm{d}q}{4\pi\varepsilon_0 r^2} \tag{8}$$

六、模型的评价与推广

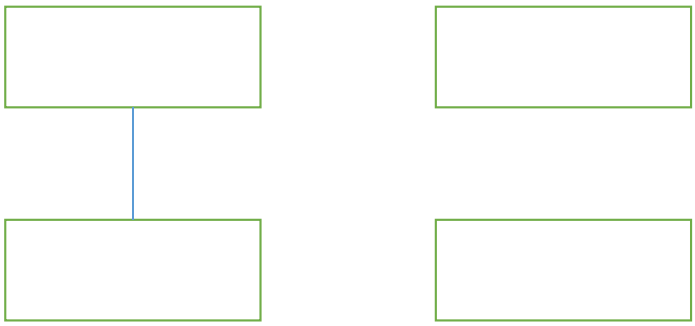

图 3　这里是图片的标题

6.1 模型的评价

- **模型的优点：**
  1. ……
  2. ……
- **模型的缺点：**
  1. ……
  2. ……
  3. ……
- **模型的改进：**
  1. ……
  2. ……
  3. ……

6.2 模型的推广

参考文献

[1] 刘柳. 基于 QAR 数据的着陆阶段飞行风险研究 [D]. 重庆大学,2018.

[2] 龙海江. 基于 QAR 数据的重着陆分析研究 [D]. 中国民用航空飞行学院,2020.DOI:10.27722/d.cnki.gzgmh.2020.000089.

附　录

[A] 图示

[B] 支撑文件列表

支撑文件列表如下（列表中不包含原始数据集）：

[C] 使用的软件、环境

为解决该问题，我们所使用的主要软件有：

Python 环境下所用使用到的库及其版本如下：

[D] 问题解决源程序

D.1

```
1 import numpy as np
```

D.2

D.3

D.4