**特征工程说明文档**

**1. 概述与目标**

本特征工程流程旨在将以单个航班为单位的、包含丰富气象雷达信息的原始流水数据，转换为一个专属于广州白云国际机场（ZGGG）的、以小时为单位的、结构化的多元时间序列数据集。

最终输出的数据文件 final\_zggg\_model\_input\_v4.csv 是后续所有模型训练和评估的唯一数据源。其核心目标是为每一个小时构建一个全面的特征向量，该向量不仅包含该小时的计划与实际交通流量，还包含一个经过“消毒”处理的、客观反映该小时机场天气状况的综合天气向量，以彻底杜绝“数据泄露”风险。

**2. 输入数据**

* **文件**: cleaned\_zggg\_full\_features\_data.csv
* **结构**: 每一行代表一个计划从ZGGG出发或抵达ZGGG的航班记录。
* **核心字段**:
  + dep\_airport\_code, arr\_airport\_code: 起降机场四字码。
  + scheduled\_dep\_time, scheduled\_arr\_time: 计划离港/到港时间。
  + actual\_dep\_time, actual\_arr\_time: 实际起飞/落地时间。
  + **基础气象特征**: 如 wind\_spd\_dep, vis\_arr 等，后缀 \_dep 表示起飞机场天气，\_arr 表示到达机场天气。
  + **雷达图特征**: 如 pct\_dbz10\_dep, max\_dbz\_arr 等，后缀规则同上。

**3. 特征工程执行流程**

整个流程围绕构建一个以小时为索引的DataFrame展开，并逐步填充各类特征。

**3.1. 构建时间序列骨架**

* 首先，脚本会扫描输入数据中的所有时间戳，找到最早和最晚的时间点。
* 然后，以小时为频率（freq='H'），创建一个覆盖整个时间范围的、连续无间断的时间索引。这个索引构成了我们最终数据集的“骨架”，确保了时间的连续性。

**3.2. 聚合交通流量特征**

脚本会计算四类核心的流量特征，并将它们聚合到小时骨架上：

* **planned\_departures**: 通过统计在每个小时内，有多少航班的scheduled\_dep\_time落入该时段，来计算**计划出港数**。
* **planned\_arrivals**: 通过统计scheduled\_arr\_time，计算**计划进港数**。
* **actual\_departures**: 通过统计actual\_dep\_time，计算**实际出港数**。
* **actual\_arrivals**: 通过统计actual\_arr\_time，计算**实际进港数**。

其中，“计划”流量将作为模型的核心输入特征，而“实际”流量将作为模型需要预测的目标（Target）。

**3.3. 构建独立的“ZGGG天气日志” (核心)**

这是为防止“数据泄露”而设计的关键步骤。其目标是创建一个独立于任何特定航班的、客观的ZGGG机场逐小时天气记录。

1. **数据分离**: 首先，将所有航班记录拆分为两个独立的组：
   * df\_dep: 所有从ZGGG出发的航班。
   * df\_arr: 所有抵达ZGGG的航班。
2. **天气信息提取与标准化**:
   * 从df\_dep中，提取所有以\_dep为后缀的天气特征列（基础+雷达），并将它们的时间戳统一为scheduled\_dep\_time。然后，将这些列的\_dep后缀去掉，得到标准化的天气特征名（如wind\_spd, max\_dbz）。
   * 从df\_arr中，提取所有以\_arr为后缀的天气特征列，并将它们的时间戳统一为scheduled\_arr\_time。同样，将\_arr后缀去掉。
3. **合并与聚合**:
   * 将上述两组标准化后的天气记录**合并**成一个庞大的“ZGGG天气观测日志”。
   * 对这个日志，以**小时**为单位进行resample（重采样）和mean（取平均值）操作。

经过这一步，我们就得到了一个全新的、以小时为索引的数据表。其中每一行都代表了ZGGG在该小时内，综合了所有出港和进港航班天气观测值的**平均天气状况**。这个过程彻底斩断了单个航班的“实际”动态与其附带天气特征之间的“秘密通道”。

**3.4. 合并与最终处理**

* 将步骤3.2中计算出的四类流量特征，和步骤3.3中计算出的“ZGGG天气日志”，全部合并（join）到我们在3.1中创建的时间序列骨架上。
* 对合并后的数据进行最后的清理，使用fillna(0)将所有缺失值（例如，某个小时没有任何航班，流量为NaN）填充为0。
* 最后，增加hour\_of\_day（小时）和day\_of\_week（星期几）两个时间上下文特征，以帮助模型理解周期性规律。

**4. 输出数据**

* **文件**: final\_zggg\_model\_input\_v4.csv
* **结构**: 一个规整的表格。每一行代表ZGGG机场的一个小时。
* **列 (特征)**: 包括该小时的计划出港/进港数、实际出港/进港数、一个包含所有天气指标（基础+雷达）的综合天气向量、以及小时和星期这两个时间特征。