

[在此处键入]

电力变压器状态预测系统设计

成绩	
----	--

南京工程学院

通信与人工智能学院、集成电路学院

# 实训报告

题 目	电力变压器状态预测系统设计
周 次	第一周
专 业	人工智能
班 级	人工智能 221
学 生 姓 名	彭 鑫
学 号	208221220
指 导 教 师	杨 帆
时 间	2025/9/7

## 摘要

本周完成了 ETT 数据集时间序列分析项目的第一阶段工作, 主要包括数据预处理与探索性分析 (EDA)。通过对 ETTh1、ETTh2、ETTm1、ETTm2 四个数据集的深入分析, 完成了数据清洗、特征缩放、相关性分析和时序特性研究, 为下一阶段的模型构建奠定了坚实基础。

- 1. 本周存在的问题..... 4
  - 1.1 项目前期准备工作..... 4
  - 1.2 数据预处理工作..... 4
  - 1.3 探索性数据分析（EDA）..... 4
  - 1.4 周期性模式识别..... 6
  - 1.5 特征关系可视化..... 6
  - 1.6 数据输出与文档整理..... 7
- 2. 本周存在的问题..... 7
  - 2.1 技术问题..... 7
  - 2.2 分析深度问题..... 8
- 3. 下周的工作计划..... 8
  - 3.1 模型设计与实现..... 8
  - 3.2 特征工程优化..... 8
  - 3.3 模型评估体系..... 8
  - 3.4 代码重构与文档..... 9

## 1. 本周存在的问题

### 1.1 项目前期准备工作

1.1.1 项目环境搭建，创建了名为 ETD-Env 的 Conda 虚拟环境，安装了 pandas、numpy、matplotlib、seaborn、scikit-learn、scipy、statsmodels 等必要的科学数据科学库。

1.1.2 数据集获取与了解，从 ETDataset 官方仓库下载了 ETT 数据集，包含 ETTh1.csv、ETTh2.csv、ETTm1.csv、ETTm2.csv 四个数据文件，初步了解了数据集的背景和字段含义。

1.1.3 项目目录结构设计，建立了规范的项目文件夹结构，包括 notebooks、outputs/cleaned、outputs/figures 等目录。

### 1.2 数据预处理工作

完成了对四个 ETDataset 的全面数据预处理工作：

1.2.1 缺失值处理：对数据集进行了缺失值检测，发现数据完整性较好，无明显缺失值问题。

1.2.2 异常值检测与处理：采用统计方法识别异常值，实施了两种处理策略：截尾处理（Winsorizing）和异常行删除，默认采用截尾处理方式以保持数据完整性。

1.2.3 特征缩放：实现了 StandardScaler 和 MinMaxScaler 两种缩放方法，根据数据特性选择合适的缩放策略，确保各特征在相同量级上进行分析。

### 1.3 探索性数据分析（EDA）

深入开展了多维度的探索性数据分析：

1.3.1 相关性分析：通过热力图(如图 1)分析各特征间的相关关系，发现 HUFL、HULL、MUFL、MULL 之间存在强正相关性（相关系数 $>0.8$ ），而 OT 与 LUFL、LULL 呈现负相关关系（相关系数约 $-0.3$ 至 $-0.5$ ）。

[在此处键入]

电力变压器状态预测系统设计

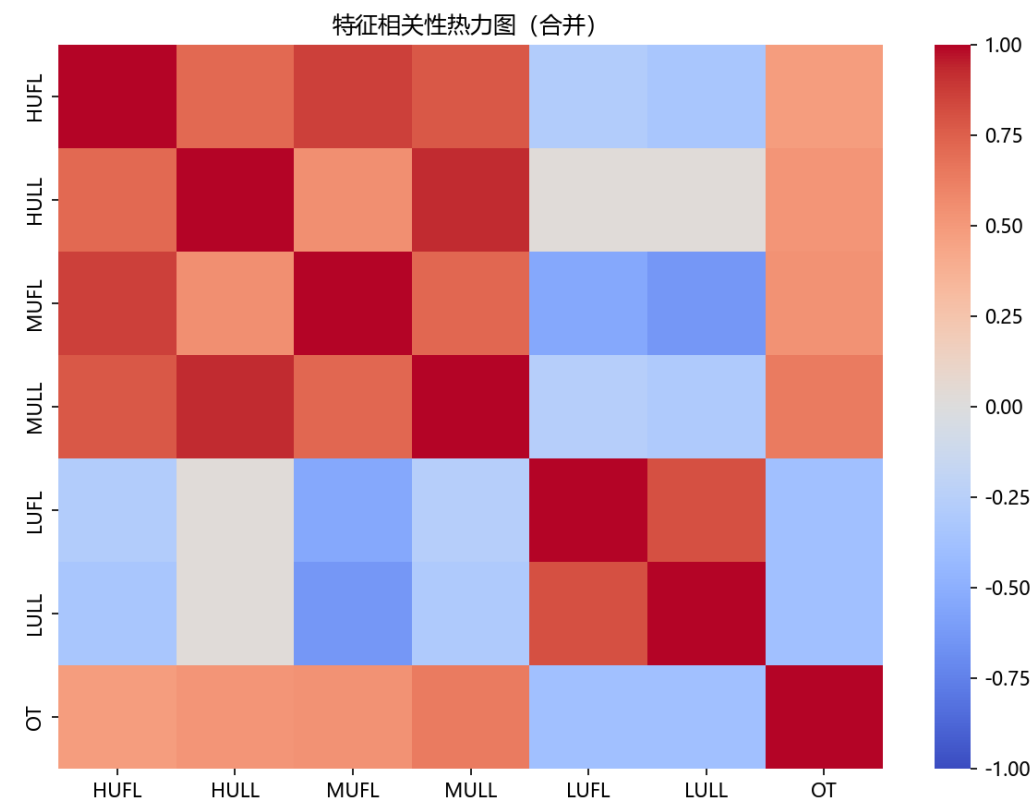


图 1 特征相关热力图

1.3.2 时序特性分析:

– 对 OT 进行 ACF/PACF (如图 2) 分析, 发现明显的自相关性和周期性特征

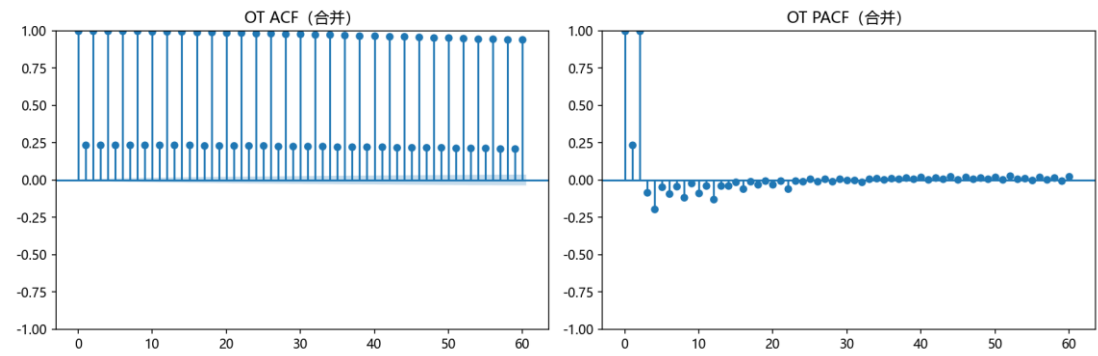


图 2 时序分析图

- ACF 显示强烈的周期性衰减模式, 表明时间序列具有长期依赖性
- PACF 在滞后 1 阶后快速衰减, 适合 AR(1) 建模

1.3.3 分布特性分析: 通过箱线图和直方图 (如图 3) 分析 OT 的统计分布, 发现其呈现近似正态分布, 峰值集中在 10-20 区间, 存在少量异常值。

[在此处键入]

## 电力变压器状态预测系统设计

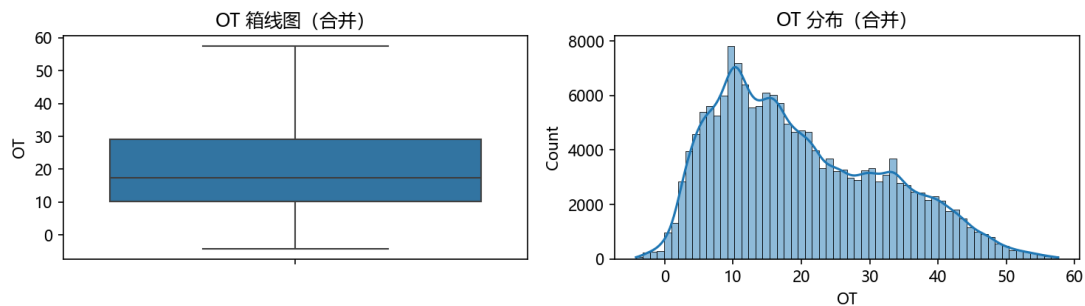


图 3 箱线图与直方图

### 1.4 周期性模式识别

1.4.1 日周期性分析：通过按小时平均分析（如图 4），发现 OT 具有明显的日周期模式，凌晨 5-6 点达到最低值（约 17.5），下午 15 点左右达到峰值（约 23.3），符合电力系统负荷变化规律。

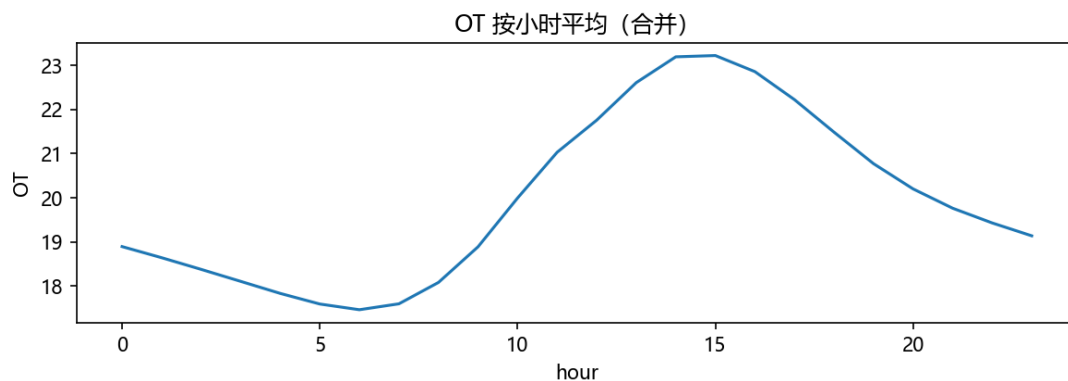


图 4 小时平均分布图

1.4.2 长期趋势分析：时序概览图（如图 5）显示 2016-2018 年间存在季节性变化和长期趋势，为季节性建模提供依据。

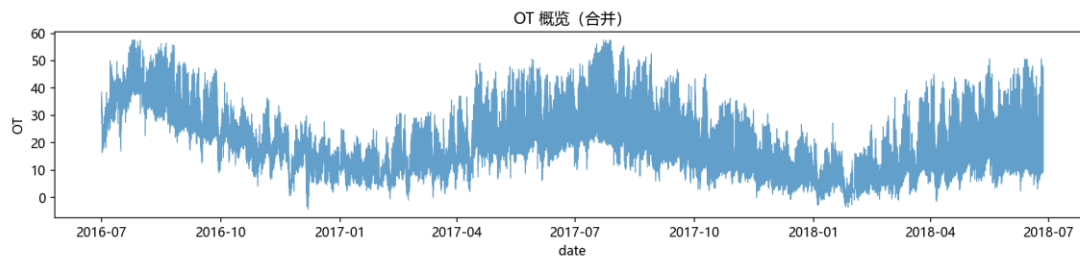


图 5 时序概览图

### 1.5 特征关系可视化

#### 1.5.1 散点图分析：

[在此处键入]

## 电力变压器状态预测系统设计

绘制 OT 与各负荷指标的散点图(如图 6)，揭示了不同特征与目标变量间的关系模式：

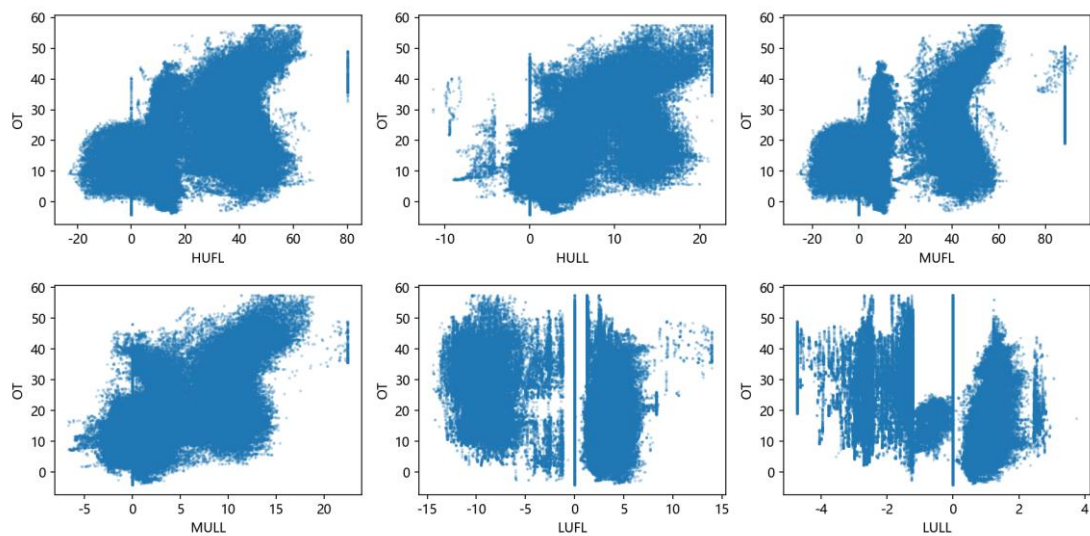


图 6 散点图

- HUFL、HULL、MUFL 与 OT 呈现较强的正线性关系
- LUFL、LULL 与 OT 的关系呈现复杂的非线性特征，存在多个聚类区域

## 1.6 数据输出与文档整理

1.6.1 数据文件输出：生成了清洗后的数据文件（\_clean.csv）和缩放后的数据文件（\_scaled.csv），为后续建模提供标准化输入。

1.6.2 可视化图表：生成了包括相关性热力图、ACF/PACF 图、分布图、周期性分析图、散点图等在内的完整可视化分析结果。

## 2. 本周存在的问题

### 2.1 技术问题

2.1.1 非线性关系处理：LUFL 和 LULL 与 OT 之间存在复杂的非线性关系，简单的线性分析无法完全揭示其规律，需要在后续建模中考虑非线性变换方法。

2.1.2 季节性特征提取：虽然发现了明显的周期性模式，但对于多层次周期性（日、周、月、季）的量化分析还需进一步深化。

## 2.2 分析深度问题

2.2.1 异常值成因分析：识别出异常值但未深入分析其产生原因，可能影响对数据质量的准确判断。

2.2.2 特征工程不够充分：当前主要进行了基础的缩放处理，缺乏基于领域知识的高级特征构建。

## 3. 下周的工作计划

### 3.1 模型设计与实现

3.1.1 基准模型构建：实现传统时间序列模型（ARIMA、SARIMA）作为基准，验证数据的可预测性。

3.1.2 机器学习模型：开发基于机器学习的预测模型，包括随机森林、支持向量机、梯度提升等算法。

3.1.3 深度学习模型：设计并实现 LSTM、GRU 等循环神经网络模型，充分利用时序数据的长期依赖特性。

### 3.2 特征工程优化

3.2.1 时间特征构建：基于周期性分析结果，构建小时、日、周、月等多层次时间特征。

3.2.2 滞后特征创建：根据 ACF/PACF 分析结果，创建合适的滞后特征以捕获自相关性。

3.2.3 非线性变换：针对 LUFL、LULL 等特征的非线性关系，尝试多项式变换、对数变换等方法。

### 3.3 模型评估体系

3.3.1 评估指标设计：建立包括 MAE、RMSE、MAPE 等多维度评估指标体系。

3.3.2 交叉验证策略：设计适合时间序列的交叉验证方法，避免数据泄漏问题。

3.3.3 可视化评估：开发预测结果的可视化展示系统，直观评估模型性能。



### 3.4 代码重构与文档

3.4.1 模块化重构：将分析代码重构为可复用的 Python 模块，提高代码质量。

3.4.2 自动化流程：建立从数据预处理到模型训练的自动化工作流。

3.4.3 技术文档：完善技术文档，为项目后续阶段和成果展示做准备。