负荷预测实验

李浩宇

摘要

此实验目的在于展示一套结合机器学习的负荷预测流程，其中包括对数据的预处理（清洗转换）、特征处理、模型调试等。此次负荷预测实验数据来源于 2012 年 kaggle 的电能负荷预测比赛，该比赛内容是对美国内 20 个区域的每小时负荷(hourly load)进行预测。本文将主要配合程序简单解释此次实验的三个核心部分，分别是负荷数据特征描述，负荷数据特征化，负荷预测。实验最后部分将对 xgboost 算法于区域 1（zone1）进行负荷预测并对其精度进行评价和作出一些提高负荷预测性能的建议。此实验于 Anconda 平台下使用编译语言 python3.6 进行编写。

所有程序、数据、文章所用原图均在附件文件夹中。

关键词：负荷数据特征化，负荷预测，xgboost 算法

1　第一部分 数据特征描述

此次实验目的在于说明进行负荷预测的流程，加上缺少真实的用户负荷数据，所以采用了在与实际负荷预测需求时间粒度（小时）相同的 2012 年 kaggle 的电能负荷预测比赛的数据集进行实验。

1.1　数据集描述

此数据集分别包含在 2004 年 1 月 1 日至 2008 年 6 月 30 日内美

国某 20 个区域的每个小时的负荷与 11 个测量点的每个小时的气温。原始比赛要求对此期间内 20 个区域的 8 周的缺失进行回测填补处理，再利用完整的数据集对 2008 年 7 月 1 日至 2008 年 7 月 7 日的每小时负荷进行预测；为了更贴近实际应用，同时缺失数比例较少，我把缺失数据删除掉后的数据集作为实验数据集。为训练模型以及检查预测效果，我将按时间的顺序把前 95%数据集划分作为训练数据，后 5%的数据作为测试数据用以检验模型预测的精确性。由于每个区域的每小时负荷预测可以用相似的方法，故本次实验只对区域 1（zone1）的每小时负荷进行特征分析以及预测。

数据集来源可见：https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting

1.2 数据预处理

由于数据集并不适合直接建模，所以需要进行转换清洗，转换成可用于建模的数据。所用到程序为 data_transfrom.py 和 eda.py

表 1.1：主要数据集列表

原始数据集（负荷）	Load_history.csv
原始数据集（温度）	temperature_history.csv
处理后导入模型数据集	Load_history_augmented.csv

1.3 数据特征描述

本节主要描述内容为：

1) 20 个区域的日负荷曲线；

2) 20 个区域的负荷与 11 个测量点温度之间的关系；

3) 区域 1(zone1)的负荷特征。

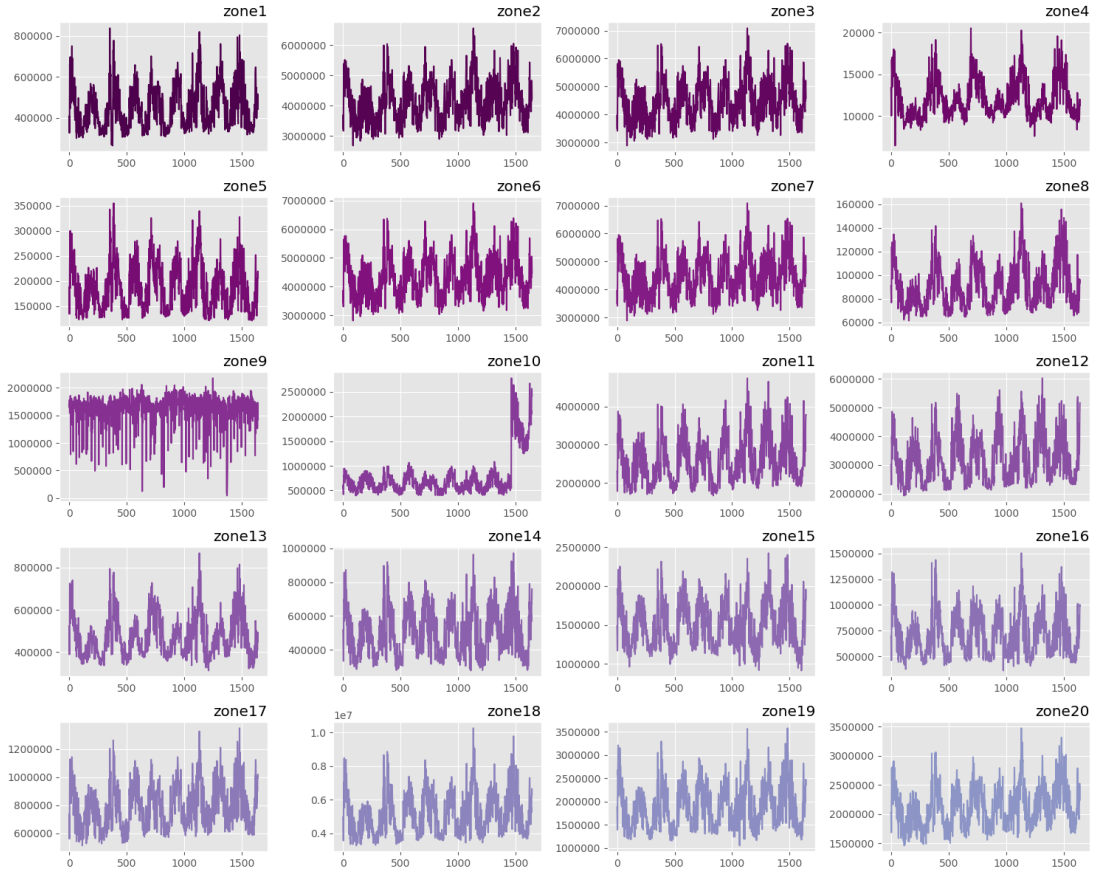

图 1.1　20 个区域日负荷曲线

如上图所示，20 个区域内的日负荷多呈一种有规律波动的时间序列。

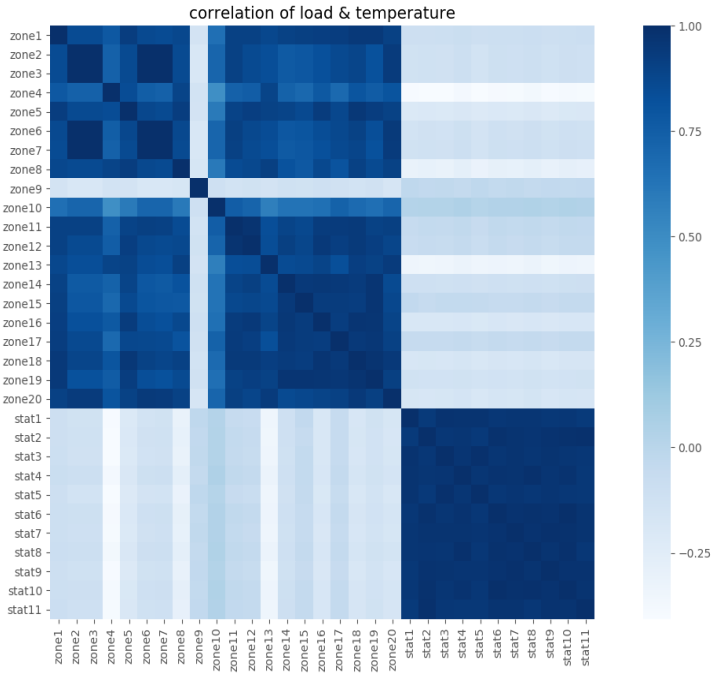

图 1.2　20 个区域的负荷与 11 个测量点温度之间的关系

如上图所示，大部分区域之间的负荷有着较强的相关性，区域 9 除外；所有温度监测点检查的温度之间也呈高度相关的关系。同时区域与单个温度测量点间的相关性并不强，说明可能了下面几种情况：

1) 存在一个区域内有多个温度测量点；

2) 一个温度测量点跨几个物理区域的边缘；

3) 有些区域不存在温度测量点。

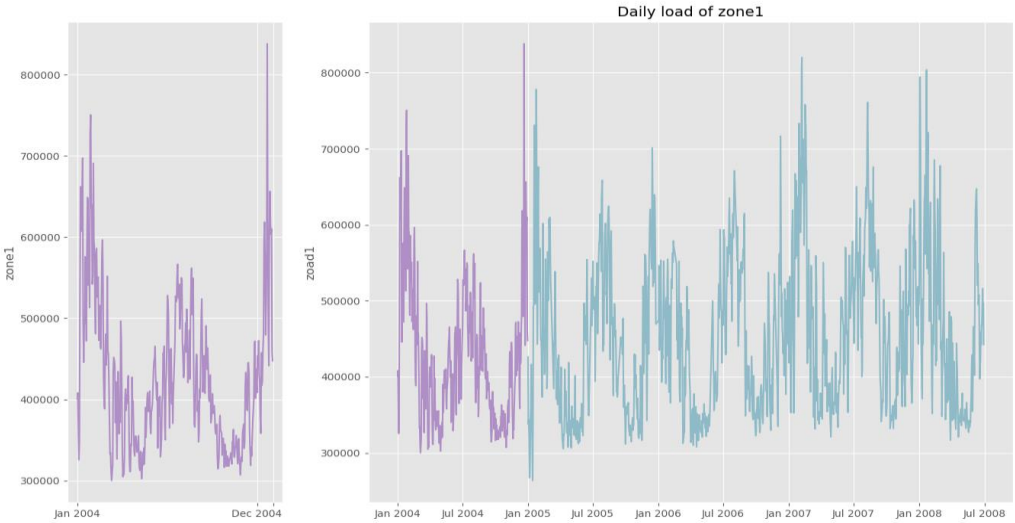

图 1.3　区域 1(zone1)的日负荷特征

从图三可以看出，区域 1 的日负荷在每一年内的变化都是呈三峰两谷的形态，中间的峰位随着年份推移变高。

2　第二部分 数据特征化

在本节我将罗列本次实验所用到的变量特征化的图示，最后以表格形式列出特征变量的目录。

2.1　区域 1 每小时负荷的分布

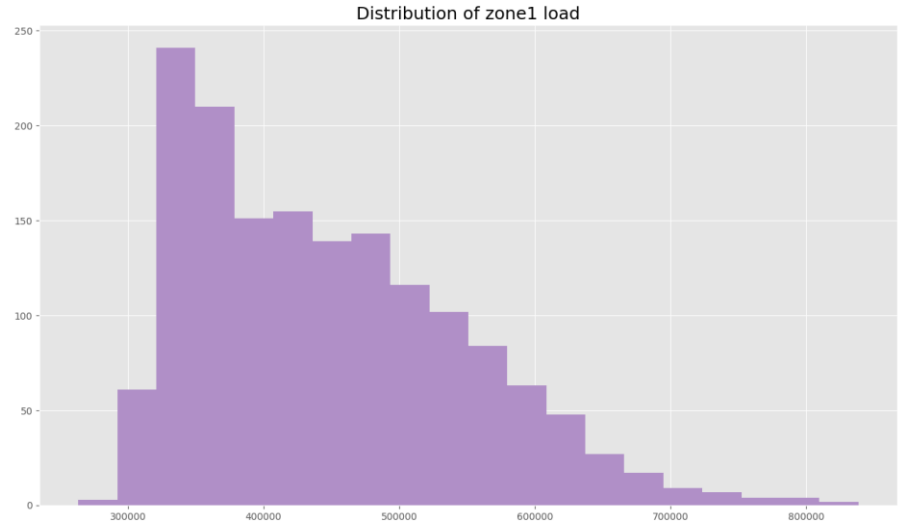

图 2.1　区域 1 每小时负荷的分布

由上图我们发现，区域 1 每小时的负荷分布是右偏的，即它的众数小于其平均数，若果我们单纯使用移动平均值作为其每个小时的负荷预测往往是会低估真是的负荷。

2.2　区域 1 每小时负荷日内特征

在本节，我将对区域 1 分别季节、周、日内的每小时负荷进行特征分析，目的是确定是否需要加入小时、周、季节等的特征变量。

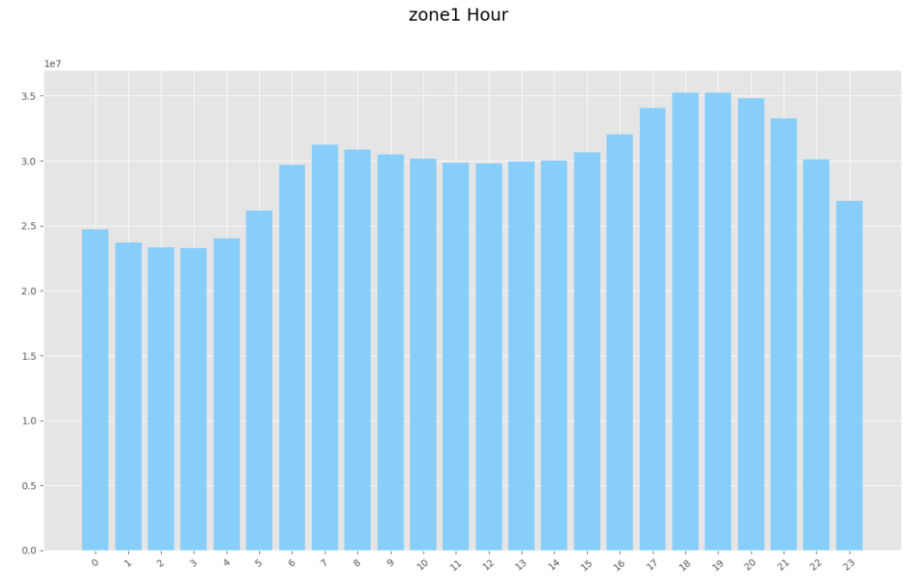

图 2.2　区域 1 每小时负荷的日内形态

鉴于累加值与平均值图示形态相同，在此图示只用了每小时负荷的累加值。由上图可以发现，日内负荷基本呈现两峰三谷的形态。

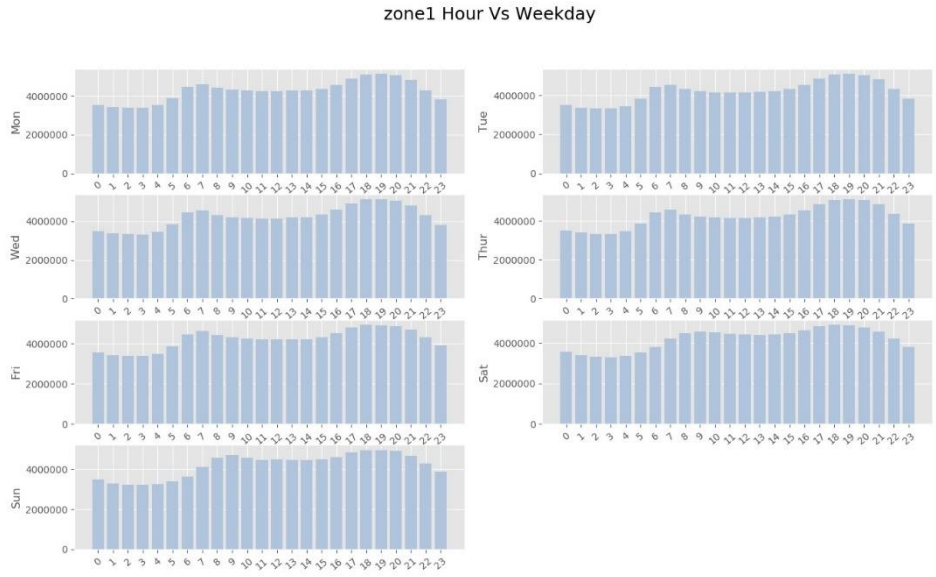

图 2.3　区域 1 每小时负荷的周形态

上图可见，每周不同的星期每小时负荷形态基本呈现两峰三谷的形态，但个体间还是有一点差异。如周一到周五第一个峰段突出较明

显，周日第一个谷段更深等特征。

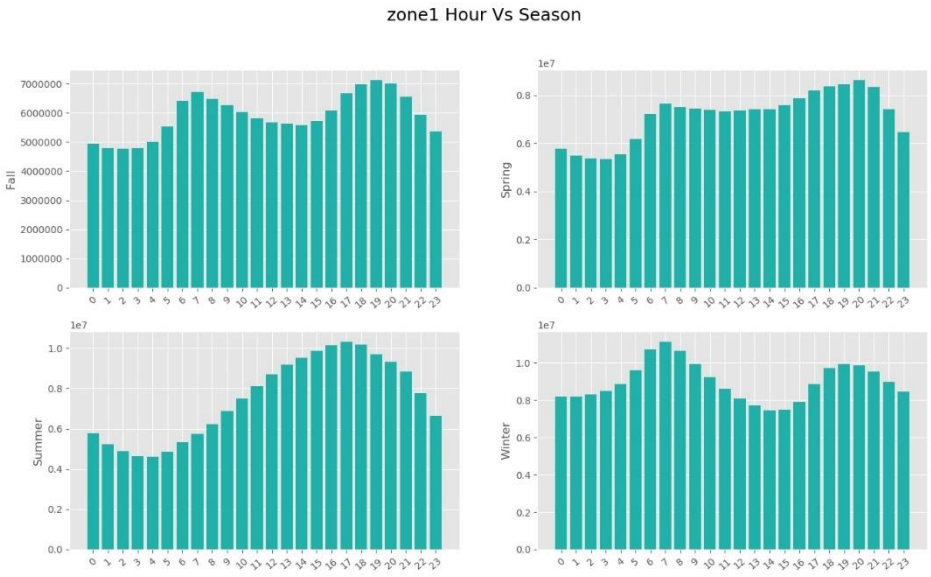

图 2.4　区域 1 每小时负荷的季节形态

上图可以看出，不同季节的日内每小时特征差异是十分明显的。

2.3　区域 1 负荷季节特征

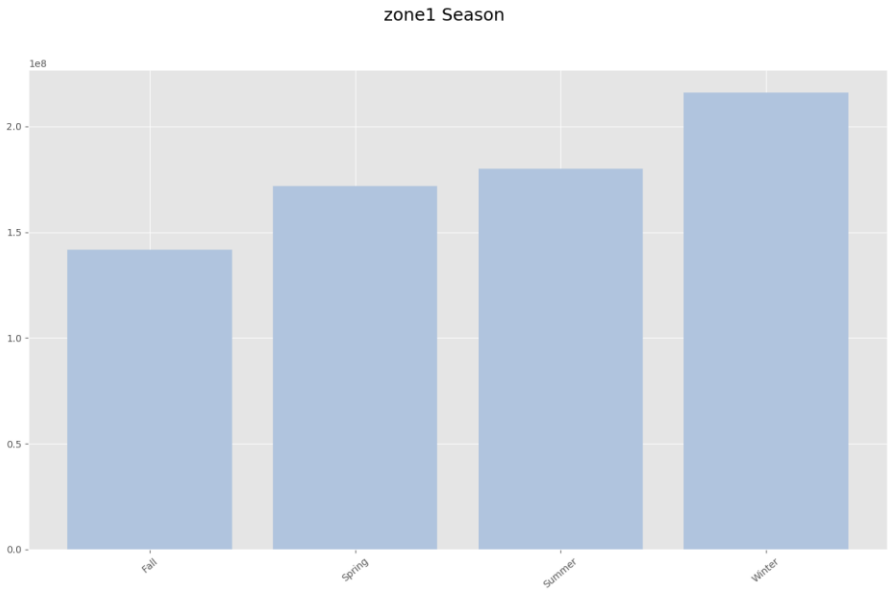

图 2.5　区域 1 四季节负荷形态

区域 1 在每个季节的负荷特征是存在非常明显的差异，秋季

低，春季次之，再者夏季，冬季最高。

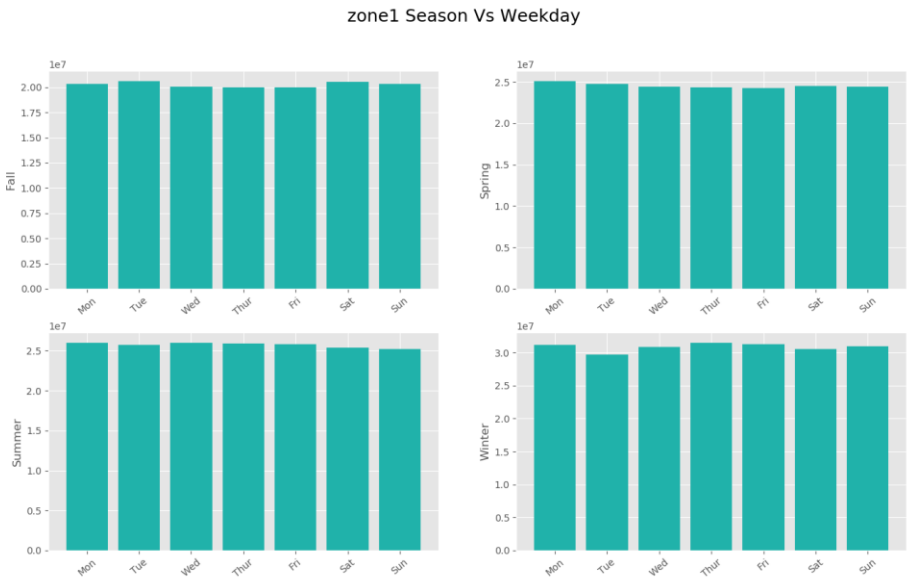

图 2.6　区域 1 四季节负荷的周形态

区域 1 四季节负荷的周形态并不是特别明显，但不同季节间的周负荷特征还是存在一定的差异。

2.4　特征化列表

经过上述章节分析不同特征的负荷形态，目的是为了尽可能在已知信息下提取（拓展）一些对负荷预测有关的特征。特征化工程在机器学习中占有决定性的作用，算法的性能决定预测精度的下限，特征化决定预测精度的上限。

表 2.1：特征化变量

特征	属性
Weekday	Mon, Tue, Wens, Thur, Fri, Sat, Sun.
Season	Spring, Summer, Fall, Winter.
Holiday	True ,False.

Hour	0，1，2，3，4，5, …, 23.
Weekend	True ,False.
Month	1, 2, 3, …, 12

受制于比赛数据的信息，所以此次实验的特征化工程并不完整，在实际的应用当中，我们可以获得更多影响负荷的特征。

本节需要用到的程序为 eda.py 以及 data_describe.py。

3 第三部分 负荷预测

3.1 Xgboost 算法

在此实验中，我是用的核心算法是 Xgboost，Xgboost 本质是一种树模型，是特征（feature）到结果/标签（label）之间的映射，是 GBDT 的优化形式。各种实验证明 Xgboost 在分类/回归中具有非常良好的性能。由于本文目的不是为了介绍该算法，所以此节仅做简述。在这里，可以把 Xgboost 算法想象成一个黑盒子，往里面输入数据就可得出结果。

3.2 Xgboost 算法调参

Xgboost 里面存在大量需调整的参数，这些参数直接影响 Xgboost 的性能，这里主要采用 CV 法进行调参。鉴于实验设备的性能，调参时间过长，某些参数的选择并未做到特别细致。

3.3 负荷预测

最后导入模型的数据集为“Load_history_augmented.csv”文件，在把里面特征变量转化为虚拟变量并划分训练集以及测试集后，就可以把训练集导入 Xgboost 模型中调参，在调参结束训练模型后最终用测

试集来检验负荷预测的精度。

我把数据集分为了三部分，训练集中再包含了一个训练集 B 及测试集 B，训练集 B 作用是训练模型，测试集 B 是检验模型性能，详细如下表。

表 3.1: 数据集划分

训练集（时间划分）95%		测试集（时间划分）5%
训练集 B（随机划分）	测试集 B（随机划分）	测试集
80%	20%	

训练集划分时间段为“2004-01-01 0:00”至“2008-04-11 21：00”，测试集划分时间段为“2008-04-11 22:00”至“2008-06-30 5:00”。

3.4 预测结果

为了检验预测效果我使用了平均绝对误差 MAE（Mean Absolute Error ）这个指标。平均绝对误差（MAE）是绝对误差的平均值，平均绝对误差能更好地反映预测值误差的实际情况。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f_i - y_i|$$

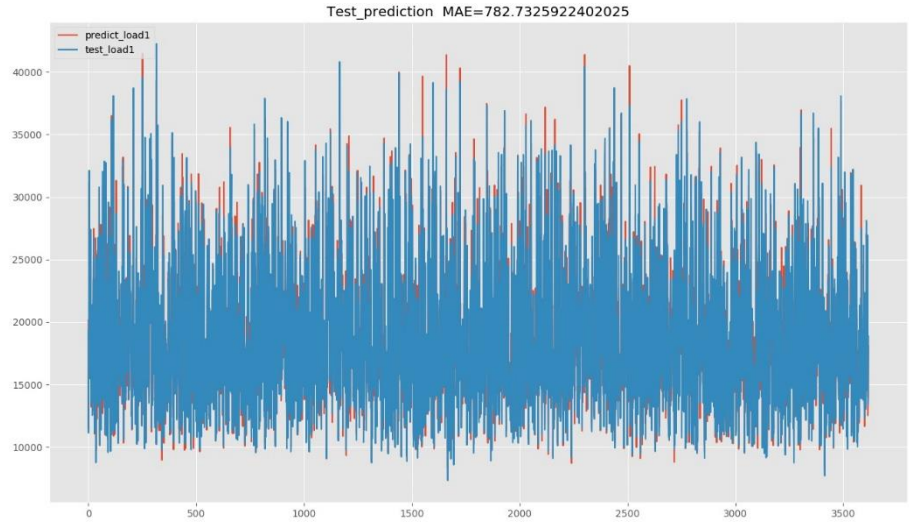

图 3.1 训练集内的测试集 B 预测效果

测试集 B 是由训练集中随机划分出来的数据集，用以检验训练集内模型预测的精确度。MAE=782.732，模型预测性能尚佳。

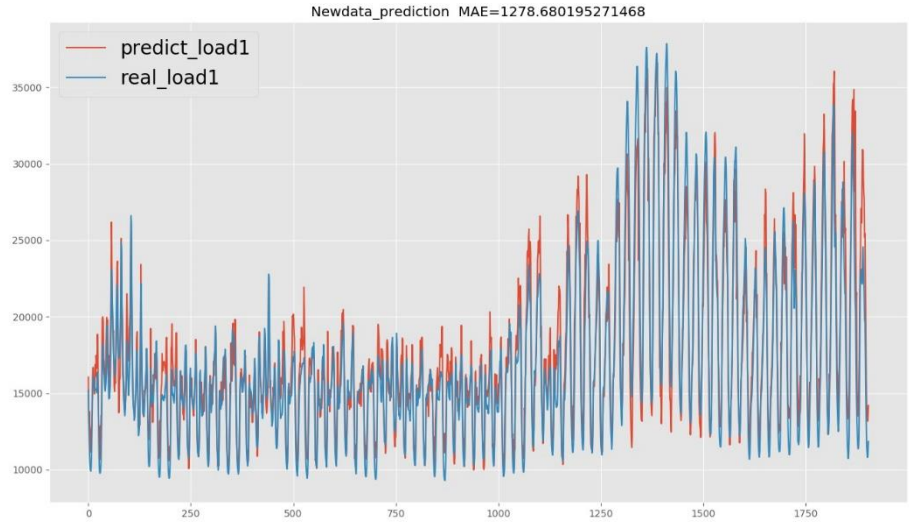

图 3.2 测试集预测效果

划分测试集的时间段为 2008 年 4 月 11 日 22 时至 2008 年 6 月 30 日 5 时。测试集预测的 MAE=1278.68，预测效果理想，但模型可

能存在过拟合，这里猜测该问题来源于模型调参。

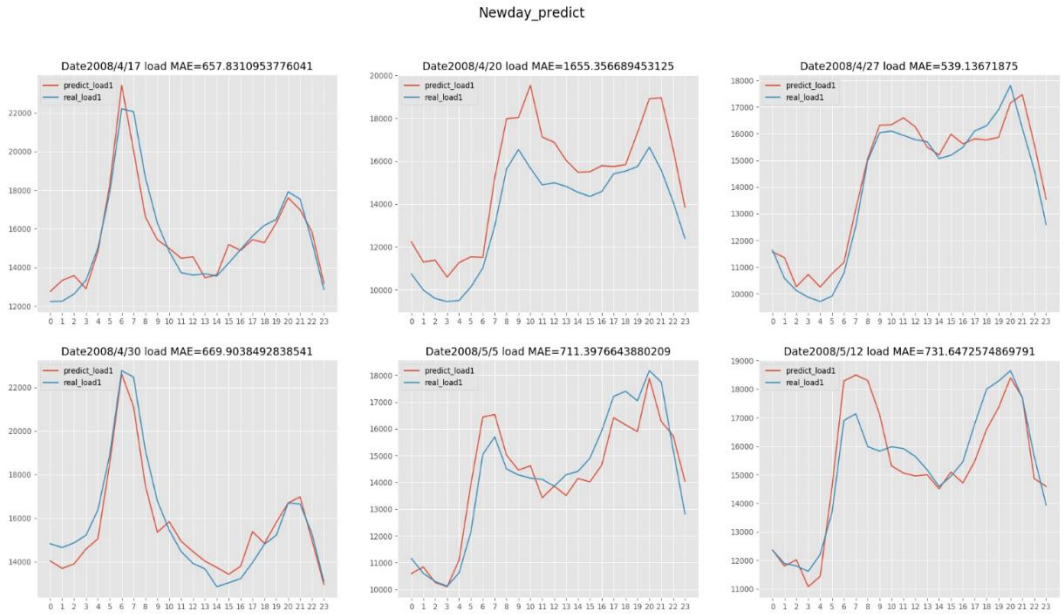

图 3.3 测试集某六天的负荷预测效果

我随机抽取了某六天观察其负荷预测效果，可以看到负荷预测值的变化趋势与实际负荷值基本吻合，预测效果达到预期期望。其中 2008-4-20 日存在较大的 MAE，其他日期的 MAE 值均低于平均水平。不可忽略某些天存在较大的预测误差（MAE），这是由于该些天的特征提取不佳所造成的原因。

本节需用到程序 load_xgb_pre.py。

4 第四部分 结论与展望

4.1 结论

本文描述了一种日内 24 个时间点负荷预测的流程，从数据处理到特征化再到训练模型最后到负荷预测，使用的 Xgboost 算法在负荷预测中也有着不俗的表现。由于数据来源于 Kaggle 比赛，出自比赛的性质，许多影响区域的负荷因素并不完整，所以影响预测精度的进

一步提高。

4.2 展望

1) 数据采集：在现实情况下我们可以收集到比此实验更完整的数据及数据特征，有助于提高日内负荷预测的精度。
2) 模型调参：在模型调参的工作中做得更加细致，不仅可以尽可能减少出现过拟合的情况，还能提高模型的表现。
3) 模型表现：要想让模型的表现有一个质的飞跃，需要依靠其他的手段，诸如，特征工程(feature egineering)，模型组合(ensemble of model)，以及堆叠(stacking)。这些是一个非常庞大的工程。
4) 特征工程：提取 T+1 日可知的影响负荷的特征（如每小时的天气，天气转移，温度，节日，客户生产计划等），使负荷预测有实际落地的效用。
5) 设备性能 :考虑到电力现货初阶段售电公司每天需要报 24 个时间点的负荷量，提高负荷预测精度，缩短预测系统运行时间是非常重要的。

最后，希望此小项目能对电力现货的准备带来一些帮助。