

# 華東理工大學

## 模式识别作业

题    目: 基于伦敦共享单车  
          数据的回归预测

年    级: 2019 级

学    院: 信息科学与工程学院

专    业: 控制科学与工程

学    号: Y20190056

姓    名: 董裕峰

指导老师: 赵海涛

2019 年 12 月 7 日

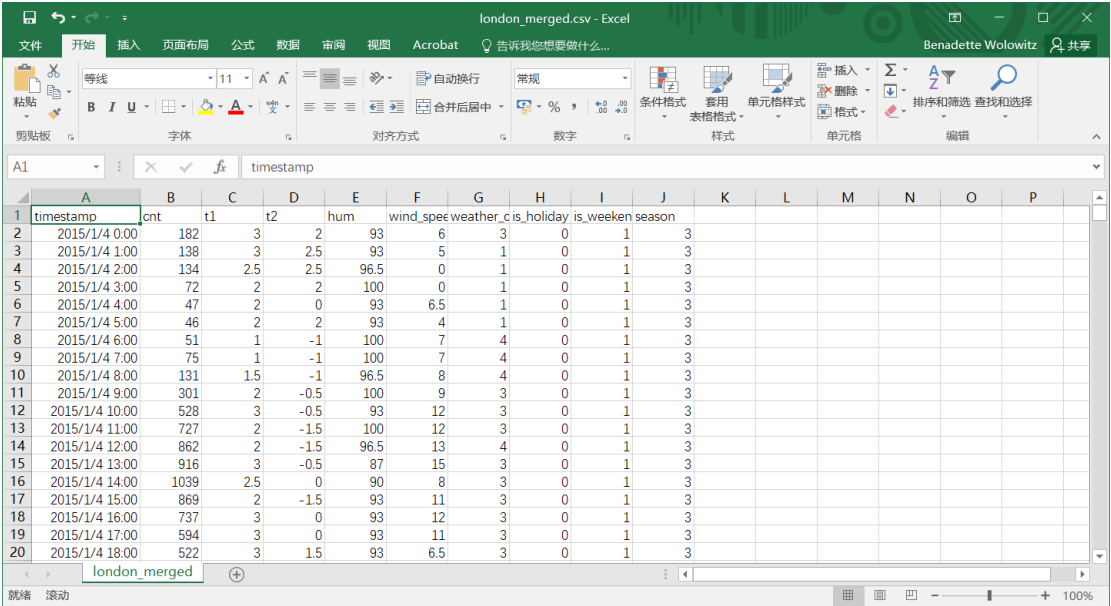
# 基于伦敦共享单车数据的回归预测

董裕峰

本课题采用的数据集取自 Kaggle 数据集的伦敦共享单车数据集，旨在通过对数据的分析处理，能够对未来的伦敦共享单车的数量进行预测。以下将分为数据集介绍、数据处理、回归预测以及总结四个部分进行介绍。

## 1 数据集介绍

伦敦共享单车数据集仅有一个名为 london\_merged.csv 的 CSV 文件，采集的是 2015-01-04 到 2017-01-03 之间每天的各整点时刻的相关数据信息。文件具体内容如图 1.1 所示。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	timestamp	cnt	t1	t2	hum	wind_speed	weather_code	is_holiday	is_weekend	season						
2	2015/1/4 0:00	182	3	2	93	6	3	0	1	3						
3	2015/1/4 1:00	138	3	2.5	93	5	1	0	1	3						
4	2015/1/4 2:00	134	2.5	2.5	96.5	0	1	0	1	3						
5	2015/1/4 3:00	72	2	2	100	0	1	0	1	3						
6	2015/1/4 4:00	47	2	0	93	6.5	1	0	1	3						
7	2015/1/4 5:00	46	2	2	93	4	1	0	1	3						
8	2015/1/4 6:00	51	1	-1	100	7	4	0	1	3						
9	2015/1/4 7:00	75	1	-1	100	7	4	0	1	3						
10	2015/1/4 8:00	131	1.5	-1	96.5	8	4	0	1	3						
11	2015/1/4 9:00	301	2	-0.5	100	9	3	0	1	3						
12	2015/1/4 10:00	528	3	-0.5	93	12	3	0	1	3						
13	2015/1/4 11:00	727	2	-1.5	100	12	3	0	1	3						
14	2015/1/4 12:00	862	2	-1.5	96.5	13	4	0	1	3						
15	2015/1/4 13:00	916	3	-0.5	87	15	3	0	1	3						
16	2015/1/4 14:00	1039	2.5	0	90	8	3	0	1	3						
17	2015/1/4 15:00	869	2	-1.5	93	11	3	0	1	3						
18	2015/1/4 16:00	737	3	0	93	12	3	0	1	3						
19	2015/1/4 17:00	594	3	0	93	11	3	0	1	3						
20	2015/1/4 18:00	522	3	1.5	93	6.5	3	0	1	3						

图 1.1 伦敦共享单车数据集

数据集中变量说明：

- “timestamp” —— 用于分组数据的 timestamp 字段
- “cnt” —— 新共享单车的数量
- “t1” —— 实际温度（单位：℃）
- “t2” —— 体感温度（单位：℃）
- “hum” —— 湿度百分比
- “wind\_speed” —— 风速（单位：km/h）
- “weather\_code” —— 天气类别
- “is\_holiday” —— 布尔值，1 假日/0 非假日

“is\_weekend”——布尔值，如果日期是周末，则为 1

“season”——季节类别，0—春季；1—夏季；2—秋季；3—冬季。

“weather\_code”类别说明：

1=清晰；大部分清晰，但有一些值与薄雾/雾/雾斑/雾接近

2=散云/少云

3=云层破碎

4=多云

7=雨/小雨/小雨

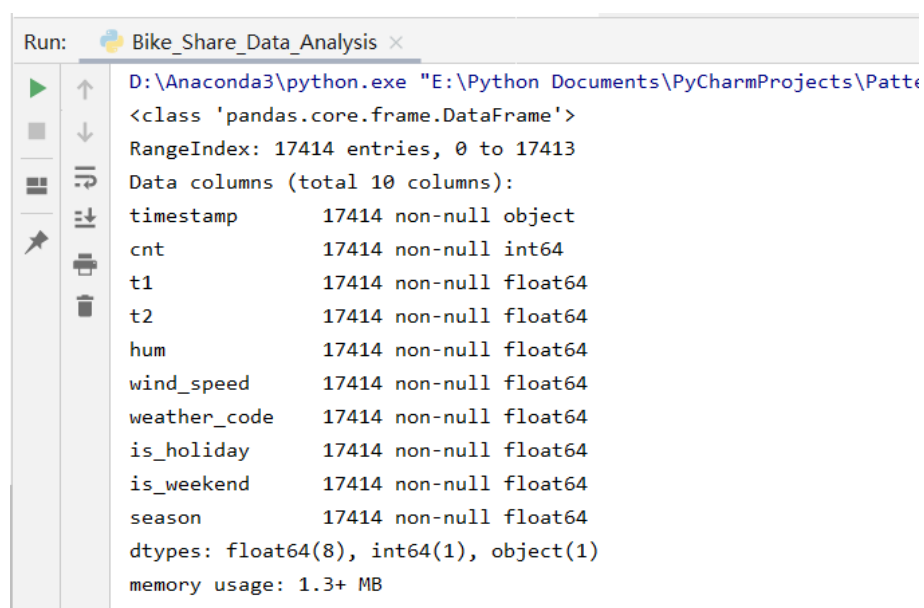
10=雷雨

26=降雪

94=冻雾

## 2 数据分析

(1) 读取数据，查看数据信息。



```
Run: Bike_Share_Data_Analysis x
D:\Anaconda3\python.exe "E:\Python Documents\PyCharmProjects\Pattern
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17414 entries, 0 to 17413
Data columns (total 10 columns):
timestamp      17414 non-null object
cnt            17414 non-null int64
t1             17414 non-null float64
t2            17414 non-null float64
hum            17414 non-null float64
wind_speed     17414 non-null float64
weather_code   17414 non-null float64
is_holiday     17414 non-null float64
is_weekend     17414 non-null float64
season         17414 non-null float64
dtypes: float64(8), int64(1), object(1)
memory usage: 1.3+ MB
```

图 2.1 共享单车数据集数据信息

从图中可看出，每个变量都没有缺失值，但是存在非数值变量“timestamp”。且一些整型变量如“weather\_code”、“is\_holiday”、“is\_weekend”在读入时被转换成了浮点型变量，不过只要数值大小一致，最终应该不会影响回归预测。

(2) 针对单车数量进行异常值的剔除。

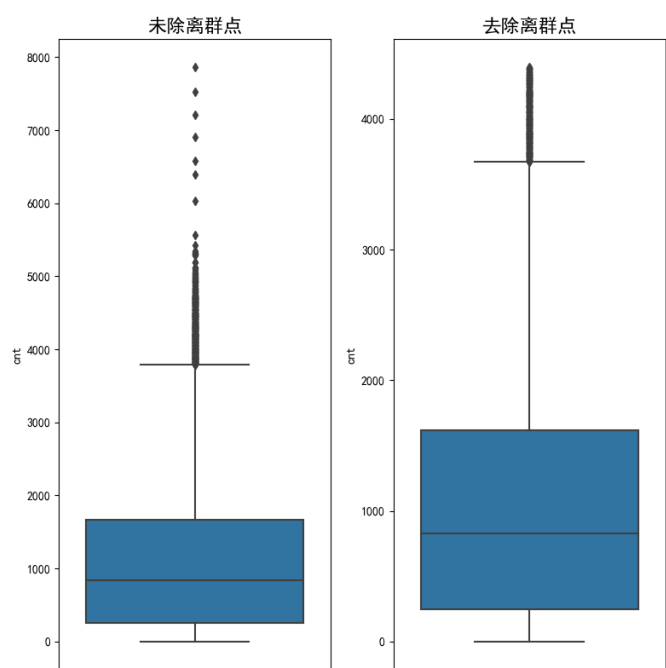


图 2.2 单车每小时数量箱型图

从图 2.2 左侧图中可看出，箱型图上方有较多异常值，但是考虑到白天与深夜的用车数量确实会存在较大差异，因而仅剔除在单车数量均值的三倍标准差之外的数据，即  $cnt.mean - 3 * cnt.\sigma < cnt < cnt.mean + 3 * cnt.\sigma$ ，从而得到图 2.2 右侧图。

### (3) 将采样时间转换成小时，对数据进行相关性分析

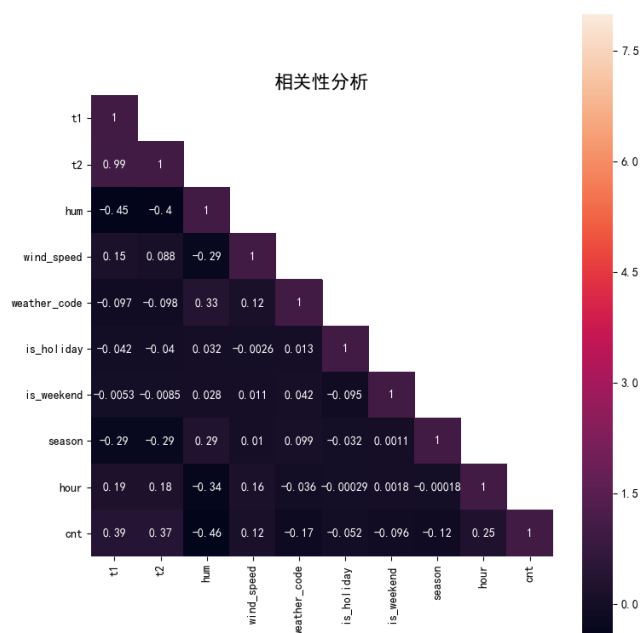


图 2.3 数据变量相关性分析

从图 2.3 中可看出，cnt 与 t1、t2、hum、wind\_speed、weather\_coder、season

和 hour 等具有较强关系，与 is\_holiday 及 is\_weekend 关联性较小，同时注意到实际温度 t1 与体感温度 t2 基本线性相关。

(4) 对数据分布分析

对单车数量的取值分布、取对数后的取值分布（由于数据存在零值，因而是加一后再取对数）以及单车数量随时间和季节的变化进行分析，得到图 2.4-图 2.6

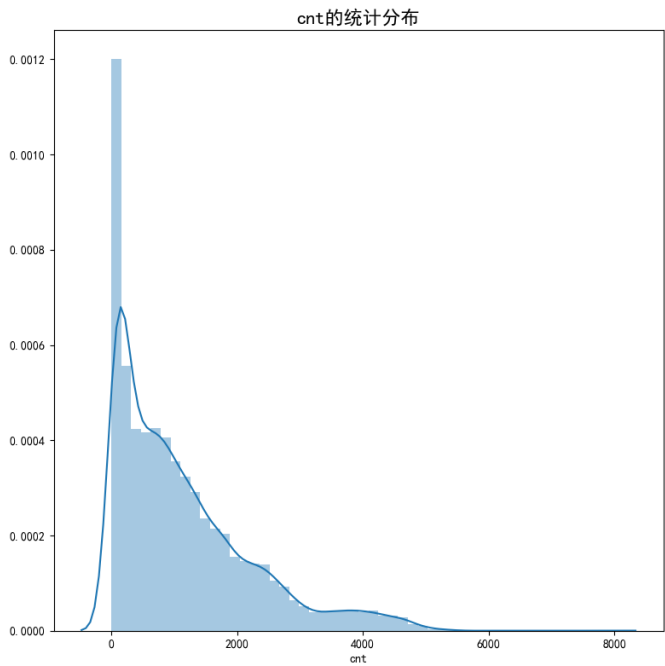


图 2.4 单车每小时数量取值分布

从图 2.4 可看出，原始的每小时单车数量在小范围内近似正态分布。

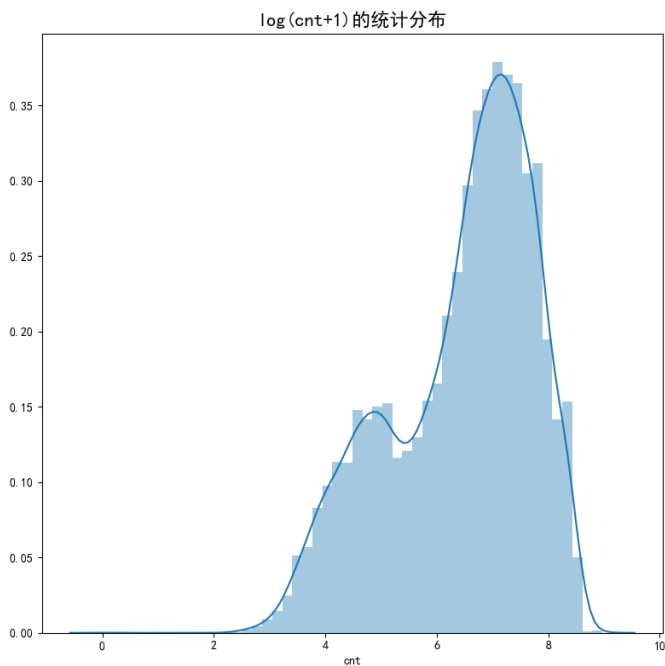


图 2.4 单车每小时数量对数取值分布

从图 2.5 可以看出，每小时单车数量取对数后的取值分布较均匀。

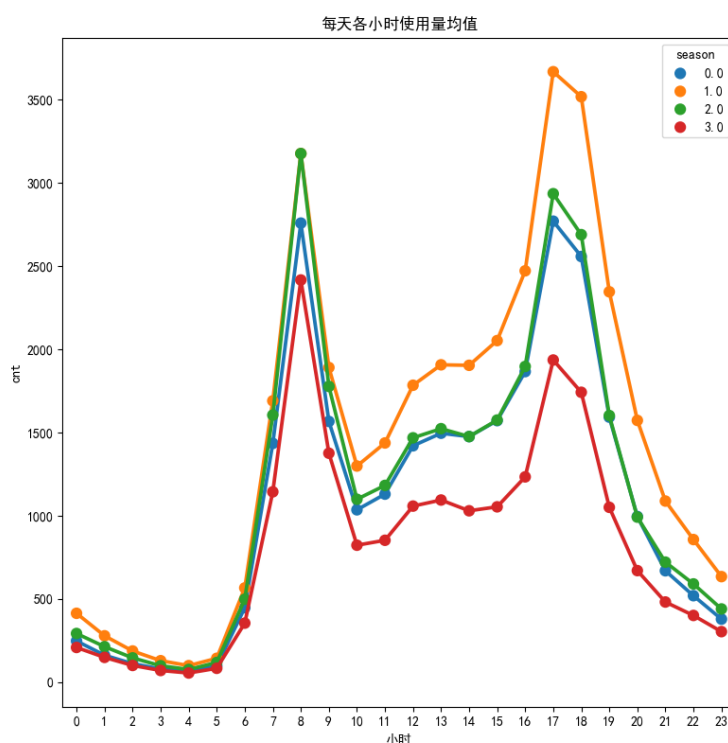


图 2.6 单车每小时数量随时间和季节变化曲线

从图 2.6 可看出，每小时单车数量随时间的变化趋势在四个季度中基本相同，而且可以看到，每小时单车数量总在上午 8 时和下午 17 时达到局部峰值。

### 3 回归预测

将原始数据分成两部分，2015-01-04 至 2016-01-03 的数据作为训练数据集，2016-01-04 至 2017-01-03 的数据作为测试数据集。本文采取了两种回归分析方法，一种是岭回归法，另一种是随机森林法。前者是在线性回归的基础上引入 L2 正则化，即待估计参数的平方和项，来减少过拟合现象。后者是一种结合自助采样法与决策树的集成学习方法。

在回归预测时，结合前面的相关性分析结果，针对不同因变量做了两组实验。一组以“t1”、“hum”、“wind\_speed”、“weather\_code”、“season”和“hour”为自变量，“cnt”为因变量，记为 E1 组；另一组以“t1”、“hum”、“wind\_speed”、“weather\_code”、“is\_holiday”、“is\_weekend”，“season”和“hour”为自变量，“cnt”为因变量，记为 E2 组。

E2 组的实验结果见图 3.1~图 3.4，E1 组的实验结果见图 3.5~图 3.8，每幅图中蓝色线条均代表实际值，黄色线条均代表预测值。

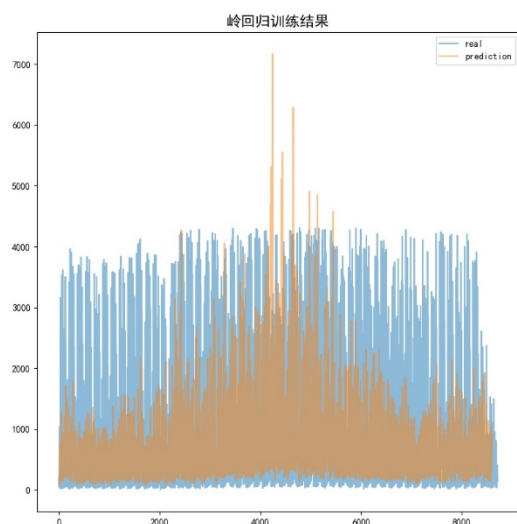


图 3.1 E1 组岭回归训练结果

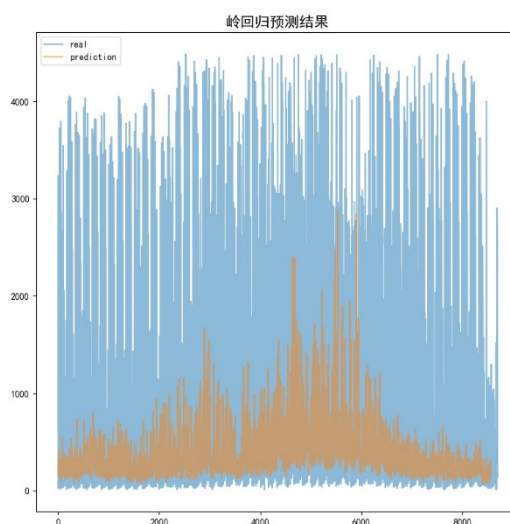


图 3.2 E1 组岭回归测试结果

从图 3.1 和图 3.2 中可看出，E1 组岭回归模型不管是在训练集还是测试集上的表现均不理想，预测值与实际值之间相差较大。

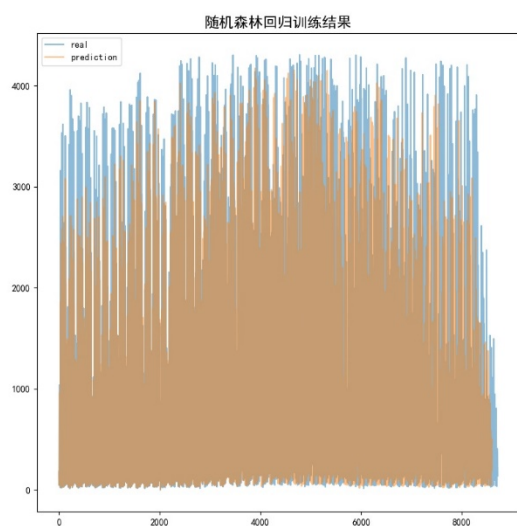


图 3.3 E1 组随机森林回归训练结果

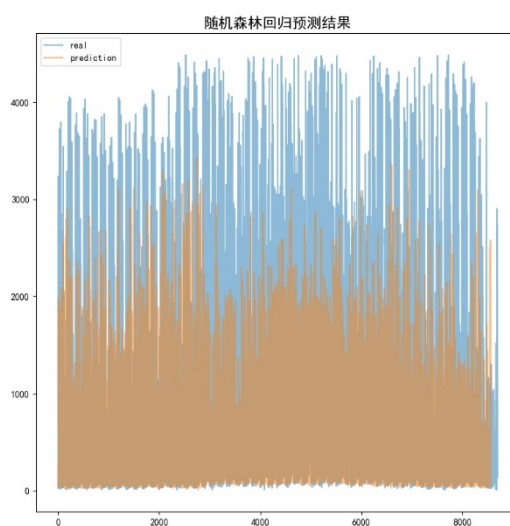


图 3.4 E1 组随机森林回归测试结果

从图 3.3 和图 3.4 中可看出，E1 组随机森林回归模型在训练集上达到了较高准确率，但是在测试集上表现不佳，预测值与实际值存在一定差距，不过能够在一定程度上反映单车数量的变化趋势且比岭回归模型的结果要好。

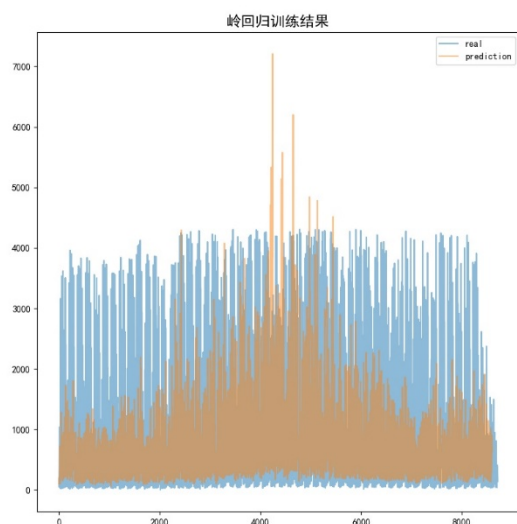


图 3.5 E2 组岭回归训练结果

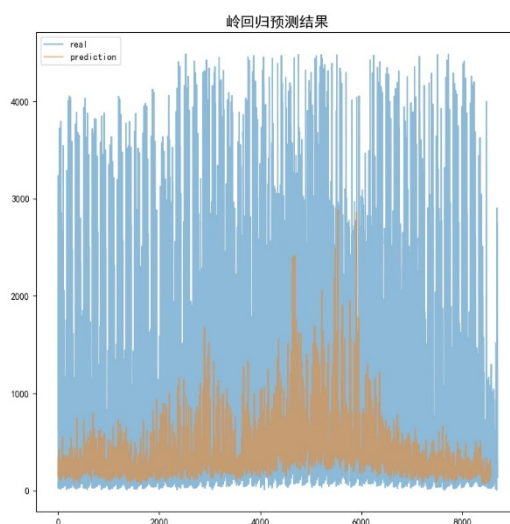


图 3.6 E2 组岭回归测试结果

从图 3.5 和图 3.6 中可看出，E2 组岭回归模型基本同 E1 组岭回归模型，不管是在训练集还是测试集上的表现均不理想，预测值与实际值之间相差较大。

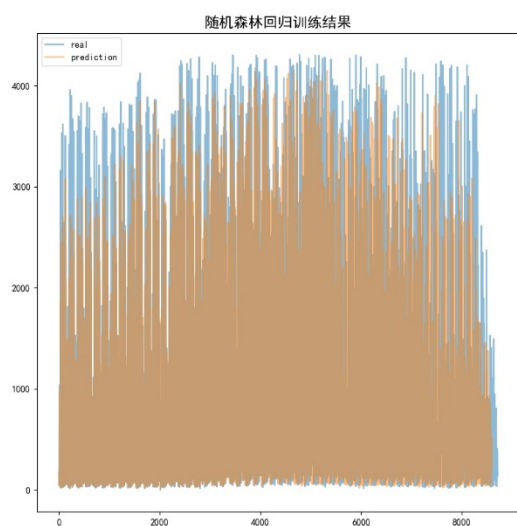


图 3.7 E2 组随机森林回归训练结果

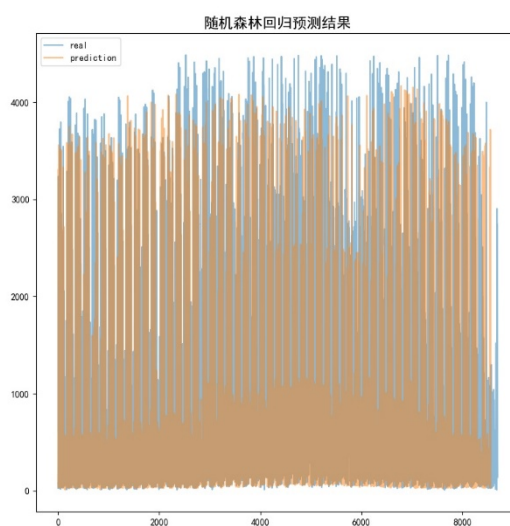


图 3.8 E2 组随机森林回归测试结果

从图 3.7 和图 3.8 中可看出，E2 组随机森林回归模型在训练集和测试集上的表现均较佳，能够较准确预测共享单车数量。

综合来看，在本回归预测任务中，随机森林模型要优于岭回归模型，且 E2 组的随机森林模型达到了最好的效果。

## 4 总结

通过前面的实验，不难发现，针对非线性的问题，不管是线性回归还是修正的线性回归（岭回归），其所得到的预测模型都存在一定的局限性，且不能够准确进行预测。与之相反，随机森林模型则能够达到较优的结果。因而，不难推知，随机森林在非线性的回归预测问题上比线性回归类方法有更好的表现。同时，注



意到在回归问题中，变量的选取对最终的预测结果也存在着较大影响。

总的来说，本次作业使我收获颇丰，一来是增强和加深了对模式识别所涉及的一些基本方法如线性回归和随机森林的认识和了解，再者是认识到了实际问题的复杂性。在很多时候求解实际问题并不是方法的简单套用，还要进行足够的数据分析和处理。

## 附件

- 附件 1 London 共享单车数据集 london\_merged.csv
- 附件 2 London 共享单车数据集数据说明 data description.txt
- 附件 3 数据分析程序 Bike\_Share\_Data\_Analysis.py
- 附件 4 回归预测程序 Bike\_Share\_Prediction.py