**数据建模笔试题**

**1、题目**

根据某客户2006年⽤电量数据（⻅附件），⽤机器学习模型来预测未来⼀⼩时的电量和未来6⼩时的⽤电量。数据⽂件中“time”列是时间戳（⽇期及相应的⼩时、分钟），“KWH”是每个时间戳对应的那个⼩时⾥的⽤电量（单位：度）。具体来讲，在每⼀个⼩时t, 我们需要预测（t+1）⼩时的⽤电量q(t+1)。可以使⽤任意开源的Python packages, such as sklearn, tensorflow, Keras, 等等。使⽤Mean Absolute Percentage Error来展示所建模型的预测效果。

**2、问题**

1. 使⽤XGBoost模型来建⽴预测模型。

2. 使⽤Long Short Term Memory模型来建⽴预测模型。

3. 描述⼀下模型所⽤到d的feature set。上述两个模型哪个更适合解决这个问题？

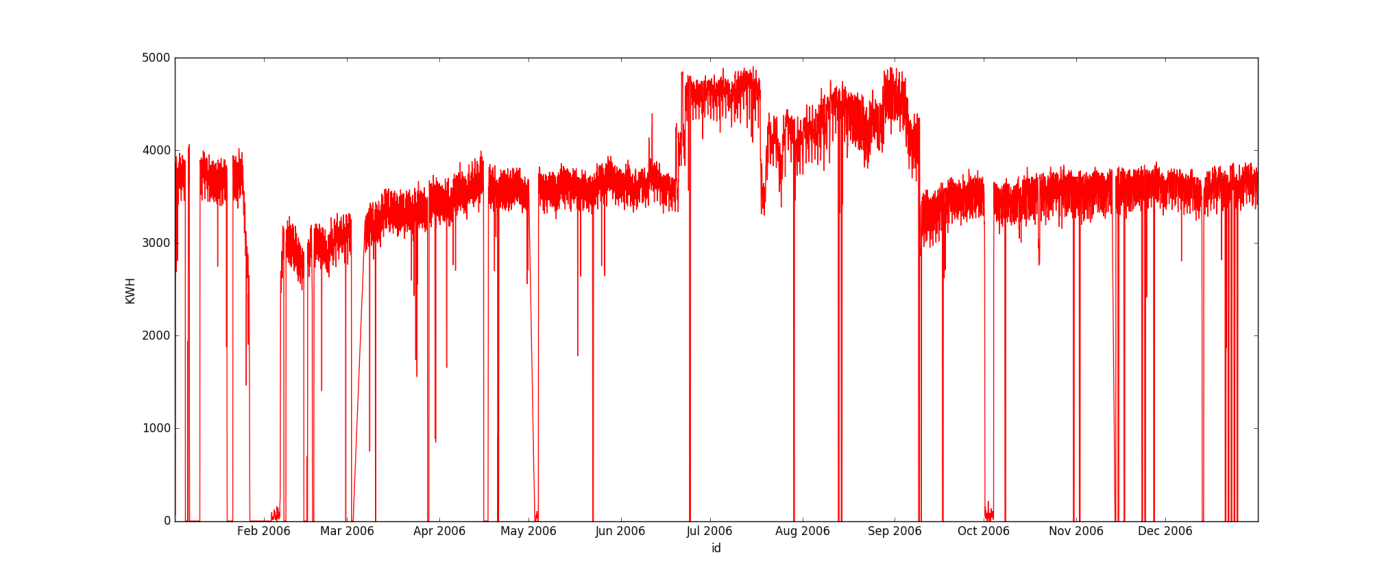
如果有兴趣，还可以解决在⼩时t预测⼀个6个元素的向量[q(t+1), q(t+2), q(t+3), q(t+4), q(t+5), q(t+6)]的问题。

**3、数据建模**

|  |  |
| --- | --- |
| 统计字段 | 数据 |
| 数据样本 | 8544条数据 |
| 开始时间 | 2006-01-02 00:59:00 |
| 结束时间 | 2006-12-31 23:59:00 |
| 最大KWH | 4900 |
| 最小KWH | 0 |

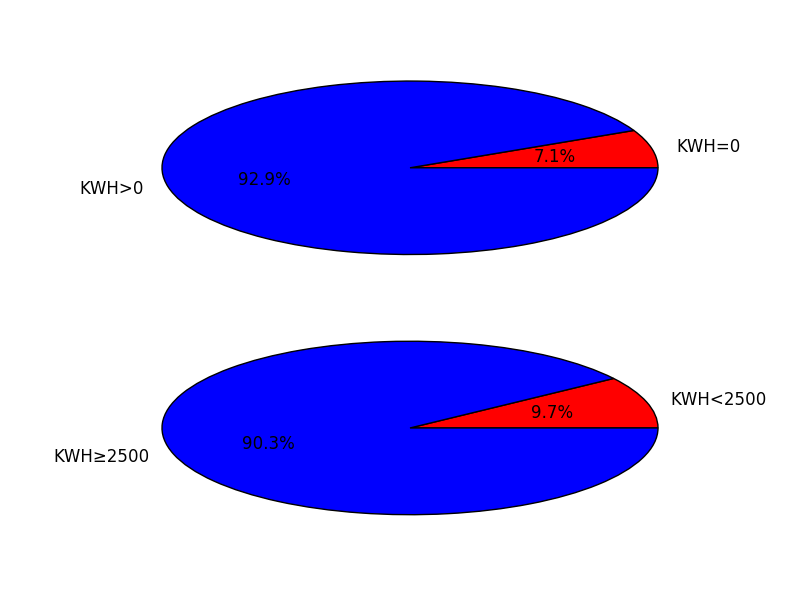
原始数据分析：

根据原始数据得到的时间与KWH的变化图。其中原始数据中大部分分布在一定的波动范围，但存在少量的数据在2000以下甚至为0的情况。具体走势图如下图一

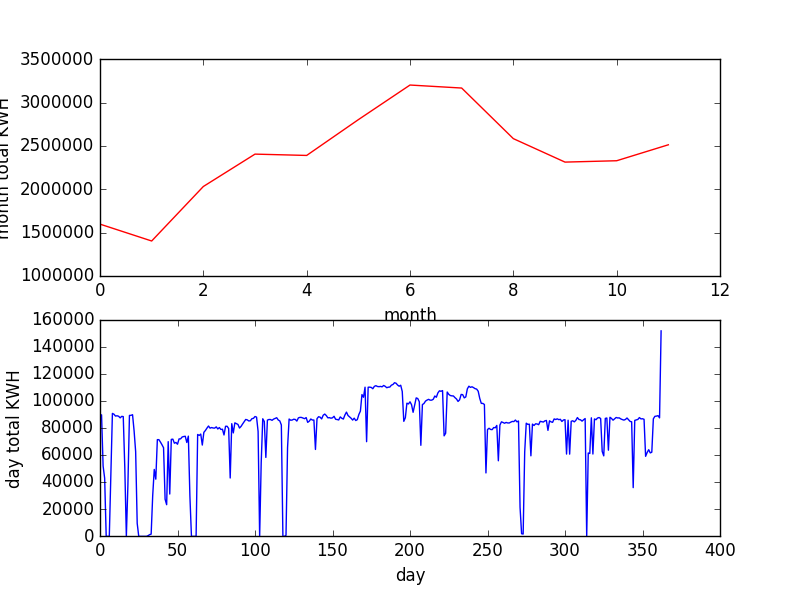


图一 时间与KWH

分析用电量为0和小于2500所占比例。根据图一选取了2500这个特殊的值进行比较。根据图二采用0和2500作为阈值进行数据统计，电量为0的存在7.1%和电量小于2500的所占9.3%。如果业务中不可能或者极小可能存在电量为0的情况，则可以考虑0为缺失或者异常值，需要进一步进行处理，可以考虑采用前七天不存在异常值的均值或者采用其他方式进行异常值替换。如果业务中存在电量为0的值，则不需要进行处理。



图二 比例关系



图三 月和日的电量

图一显示的是每个小时的用电量的变化，图三显示的是每个月和每天的电量变化。图三显示存在一些天的用电量等于0或者近似于0的用电量。而且用电量在1到6月份总体呈现上升的趋势，6月到12月先下降后平稳，根据该客户在2006年用电的数据可以看出该客户在6月左右的用电量达到了一年的峰值，二在2月达到的了低谷值。可以猜测该客户的用电分布。根据用电分布可以提供电量。图一与图三进行比较可得，存在某些天用电量近似于0的情况，大多数天的用电量保持在8000左右。每天的用电量波动导致了每个月差距相对较大。

1. **解决问题**

|  |  |
| --- | --- |
| 特征 | 标签 |
| t0,t1,t2.......t72 | t73 |
| t1,t2,t3.......t73 | t74 |
| t2,t3,t4.......t74 | t75 |
| ...... | ...... |
| t8471,t8472,t8473......t8543 | t8544 |

选取了预测目标的前三天的每个小时作为特征，即3\*24=72维特征。将12月前的数据用来训练模型，12月份的数据当作测试，即31\*24=744个测试样本。测试结果进行均值求解。

4.1 评估

本次采用Mean Absolute Percentage Error，即MAPE来评估模型，真实值与预测值的差绝对值与真实值和的比值。



选用Mean Absolute Error，即MAE来评估回归模型，真实值与预测值的差绝对的平均值。m为测试样本的数量。



**4.2 XGBoost模型**

模型在python3.6上进行测试，参数eta:0.05,objective:reg:linear,eval\_metric:mae,num\_boost\_round:1500

MAE：118

MAPE：3.497%

**4.3 keras的lstm模型**

LSTM（Long Short-Term Memory）是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。

MAE：173

MAPE：5.117%

**4.4 XGBoost与lstm模型的选择**

针对本次2006年的8544条数据样本，XGBoost在最后结果的表现上相对于lstm较好，主要原因可能由于样本数据过于少，影响到keras的lstm的学习，导致学习结果欠拟合，而数据量较少可以是XGBoost学习效果很好，很快达到收敛，致使结果相对于lstm要好。数据量较小时选择XGBoost模型，数据量规模较大可以选在DL的lstm模型。

**4.5 预测更长时间**

一种方式，预测未来的1小时到6小时，可以考虑建立一个模型来进行预测，同样选取相同的数据特征，但是在预测目标的时候选择未来的第6个小时作为目标进行训练，已同样的方式进行测试集构建并预测结果，这样的方式只需要构建一个模型便能实现结果的预测。

另一种方式，将未来的1小时到6小时当作6个预测目标，训练6个模型进行预测，测试集采用同样的方式进行结果的预测。当然随着距离预测目标的时间长短，其可信度存在一定的偏差。