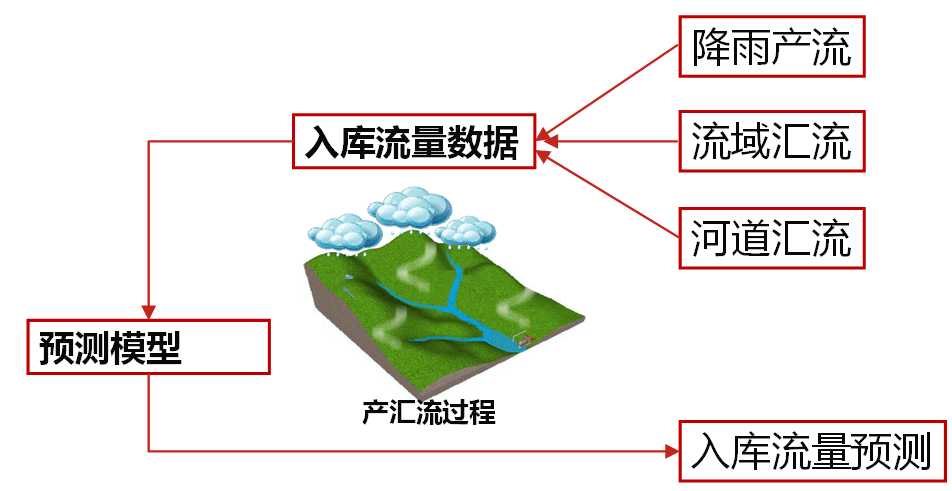
水电站入库流量预测报告

##### **前言**

##### 该文档主要是介绍通过机器学习模型LightGBM进行水电站流量入库预测。 对于水电站来说，发电是主要经济效益来源，而水就是生产的原料。对进入水电站水库的入库流量进行精准预测，能够帮助水电站对防洪、发电计划调度工作进行合理安排，实现避免洪涝灾害和提升发电经济效益的目的。



##### **目标**

##### 基于历史数据和当前观测信息，对电站未来7日入库流量进行预测（每3小时一个预测值，共56个待预测值）。

##### **数据解析**

竞赛主办方共提供了4类数据，包括历史入库流量数据、环境数据、降雨预报数据以及遥测站降雨观测数据。数据均为时序数据。

其中入库流量数据包含时间和流量两个字段。环境数据提供了温度、风速、方向三个字段。天气预报包含了未来五天的降雨情况。遥测站数据则包括了39个点的降雨量。

初赛提供：2013年-2018年的历史数据

决赛提供：2019年数据

数据维度：3小时为一个粒度点

数据缺失：初赛数据在14年缺少部分数据，决赛未提供18年数据

综上述，经过对数据的了解和分析，影响模型预测主要归纳为一下四个方面：

* 历史数据存在样本缺失
* 使用何种模型进行预测，NN还是回归
* 如何选取、构造特征，使用特征
* 数据的准确性

##### **赛题分析与模型选择**

从数据表现来看，是一个完完全全的时序题，针对时序题的做法有很多，找周期拟合、使用NN模型，本人尝试过LSTM、GRU、RNN、CNN等，通过线下拟合，自划分样本进行测试，可以观测到拟合效果非常好（如图4-1），但是反馈则是，只是存在部分段分数很高，部分分段很低，导致结果评分为BR，模型稳定性差。

遂转换思路，将问题转换成线性拟合问题，将时序数据看成一个单独的点，构造特征将时序保留，进行回归预测，重新构造测试数据，预测的输出作为下一次预测的输入，进行预测。最终选择竞赛界比较通用的LightGBM模型进行线性拟合。得到的表现却是各段分数平平如其，虽然分数较低，但是每一段的偏差相差不大，模型表现较为稳定。相对NN鲁棒性更强，这也是为什么在决赛选择LightGBM的原因。各个特征的重要程度表现如图4-2所示。

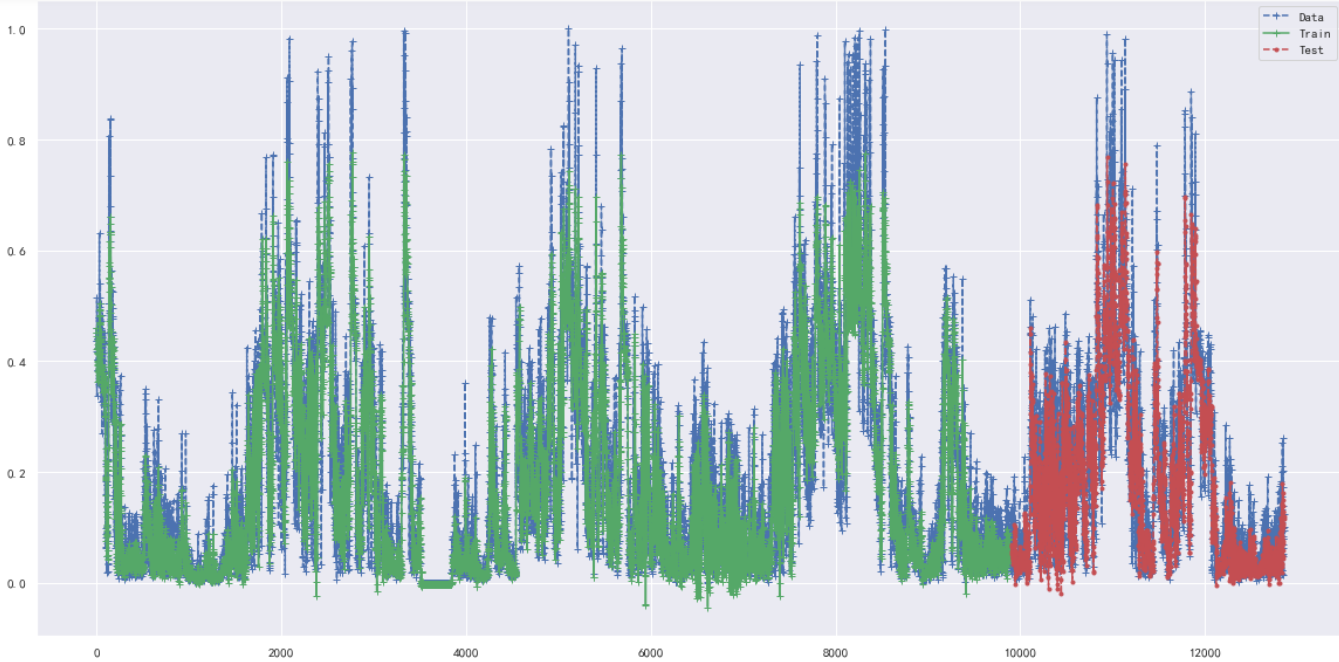


图4-1 cnn-gru拟合

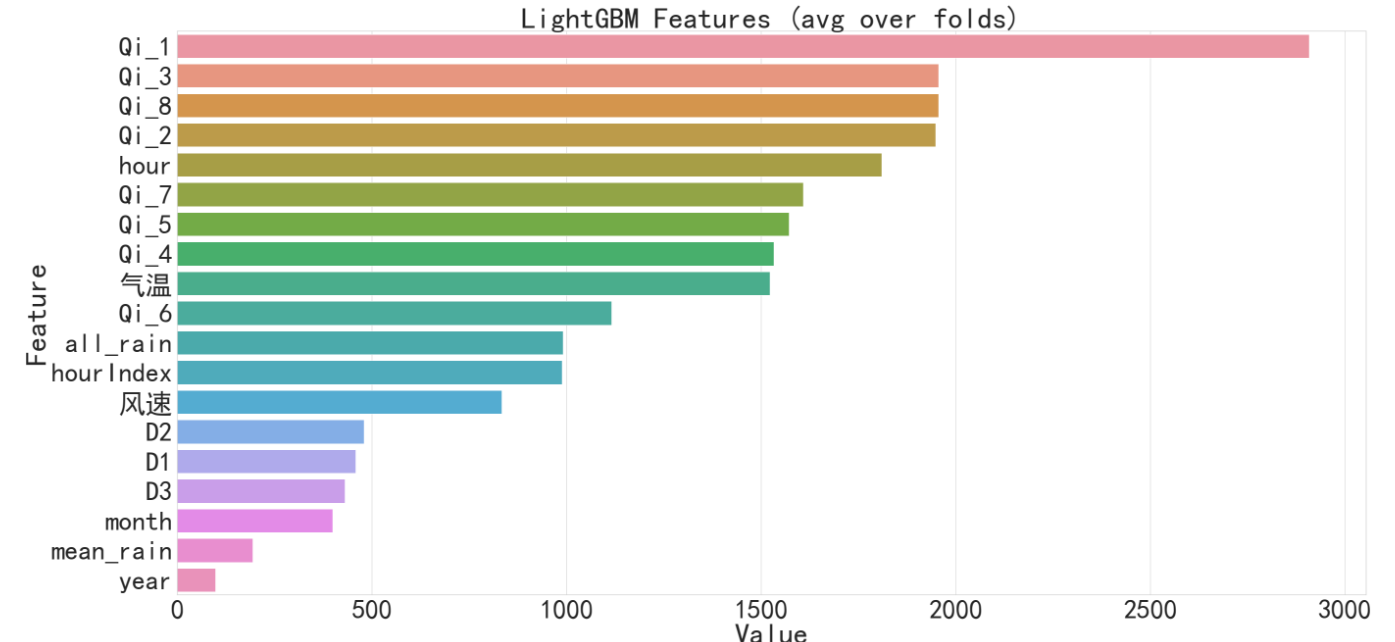
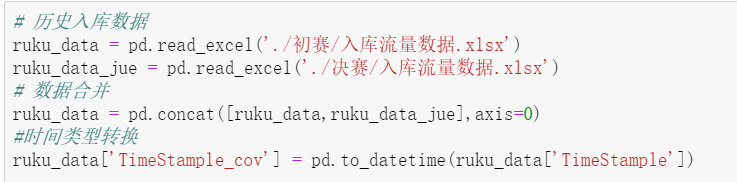


图4-2 特征重要度

##### **方法**

* 数据预处理

将初赛、复赛数据读入，相同类别数据进行concat合并。如历史入库数据。并将时间转换成datetime格式。其他三类数据类似处理。

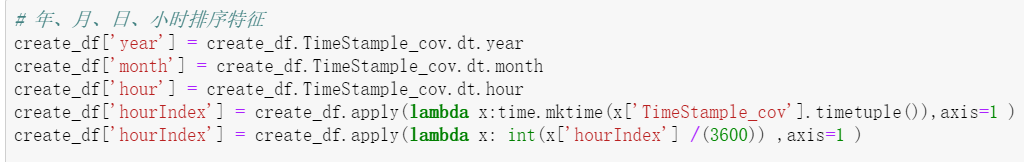


* 特征工程、训练集、测试集构造

1. 遥测站数据处理和特征工程
2. 39个遥测站数据直接求和，而且发现遥测站的数据更像是一个类别数据，和QI也存在一定的相关性。
3. 将原始的天数据转换成入库流量一直的时序数据3H粒度数据，方便关联
4. 天气预报数据
5. 这里使用的前期预报不是未来五天，而是前三天的一个天气预报作为特征输入。



1. 环境数据
2. 环境数据使用当天数据，考虑到风向数据分布不一致的问题，将其剔除，只是用温度和风速作为特征输入。
3. 入流流量数据
4. 历史8个点的时刻数据作为特征输入Q1-Q8
5. 保留时序特征
   1. 构造年、月、小时、小时IDX特征（保留时序，作为也可以理解为相近数据的权重）



1. 数据构造

按照待预测的时间段进行测试集构造。



* 模型构建

这里使用的是五折的交叉验证，对最终结果也是5折之后的平均结果。



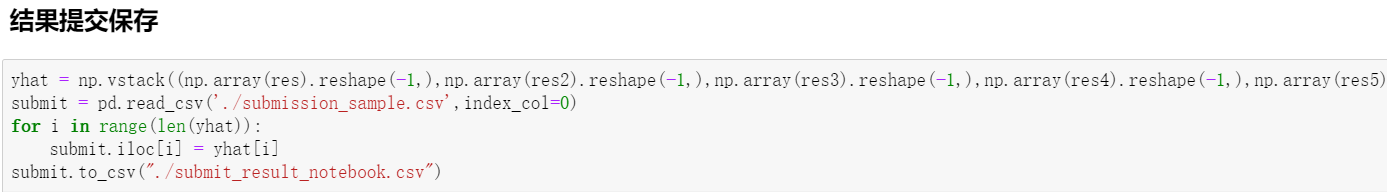
* 结果预测

对5段结果进行分别预测。每一次的输出作为下一次的输入，进行构造Q1-Q8的特征更新。五段预测方式一致。



* 结果提交

将5段结果数据进行拼接，保存至csv进行提交。



##### **总结**

从模型表现来看，最终结果五段结果均为负数，但是整体偏差不大，鲁棒性比较强，最终五段的平均在-75左右，其他朋友的模型肯定都比这个模型更加精致，从表现来看，他们在其他几段预测的结果表现都很不错，比这个模型更强，但是在第四段出现了意外，但这也是数据中不可计算的意外。他们的方案更加值得学习，共同进步，共同学习。