

## 开课吧数据竞赛第六课-钟老师-20191124

笔记本： 开课吧-小钟讲课

创建时间： 2019/11/17 星期日 22:50

更新时间： 2019/11/24 星期日 0:55

作者： 你看起来好像很好吃n\_n

URL: <https://www.kaggle.com/c/LANL-Earthquake-Prediction>

---

# 开课吧-数据竞赛及相关问题 从小工到专家

---

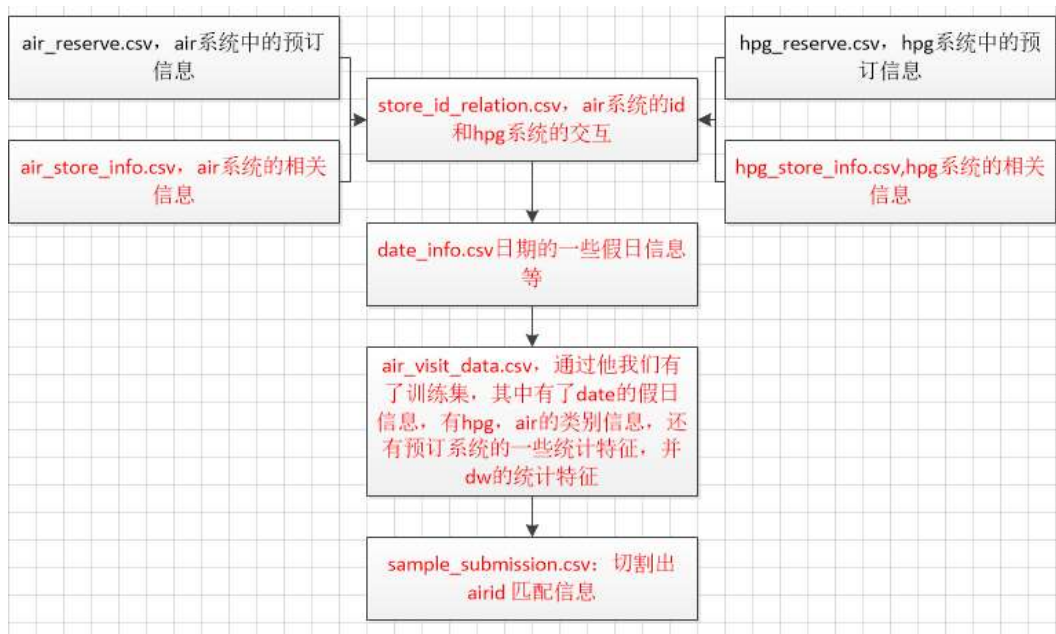
时间： 2019-11-24

## 1. 特征重要性

1.1 xgboost 特征重要性, lightgbm特征重要性, catboost特征重要性

## 2. Recruit Restaurant Visitor Forecasting-Predict how many future visitors a restaurant will receive招 募餐厅游客预测-预测餐厅将来会接待多少游客

### 2.1 回顾上一节课知识点-如何构建一个时间序列模型-baseline



## 2. 2回顾上一节课知识点-回顾Stacking

训练集	M1	P1	合并	训练集	P1	→ ←	LABEL
测试集		T1		训练集	P2	→M4←	
训练集	M2	P2		训练集	P3	→ ↓ ←	
测试集		T2		测试集	T1	↓	<u>PRED</u>
训练集	M3	P3		测试集	T2	→M4→	
测试集		T3		测试集	T3		



## 2. 3回顾上一周做的特征工程

1. 原始特征: air\_store\_info中的air\_store的类别信息;
2. hpg\_store\_info中的hpg\_store的类别信息(通过store\_id\_ration.csv交互获得作为air\_store\_id的信息) ;
3. air\_reserve.csv中和hpg\_reserve.csv的预订信息, 预订信息的统计特征, 预订时间-到访时间的sum,mean, 预订到访人数的mean,sum他们都可以匹配到一个air\_store\_id里面;
4. 通过历史信息可以知道dayofweek的统计信息。

## 2. 4进一步的特征工程

在上一周当中，我们使用了dayofweek进行统计预订时间-到访时间的sum,mean, 预订到访人数的mean,sum, 那么其实就是借用了dayofweek, train和test的共同特点air\_store\_id和dayofweek共同的特点。

### 2.4.1进一步时序提分1

一般在一个机器学习模型当中，增加训练集，往往模型会更优更具有鲁棒性，而时序题，往往是增加窗口来叠加训练集。

### 2.4.2进一步时序提分2

增加强特：时序特征往往增加不同窗口的中位数，最大值，最小值。

## 3. LANL Earthquake Prediction-Can you predict upcoming laboratory earthquakes?您能预测即将发生的实验室地震吗？

### 3. 1数据格式

### 3. 2和其他机器学习问题数据格式的异同点

### 3. 3傅里叶变换的作用

### 3. 4建模的要点