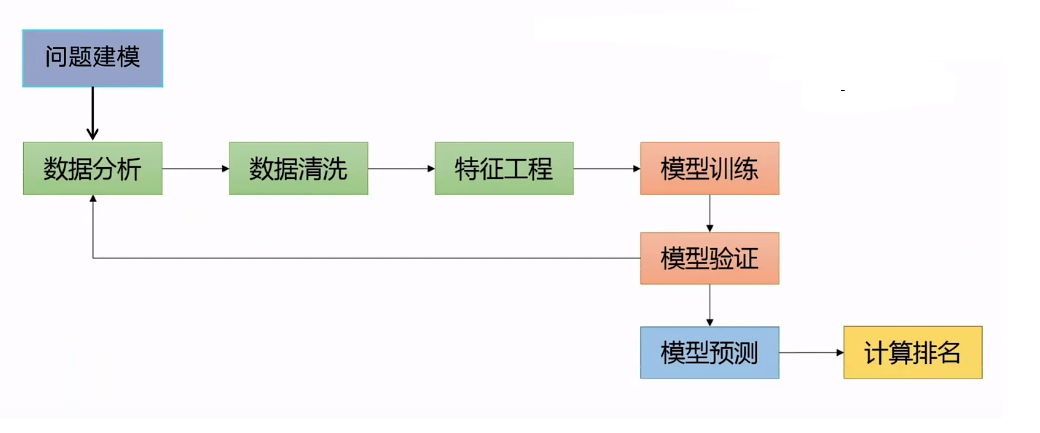
# 竞赛上分技巧

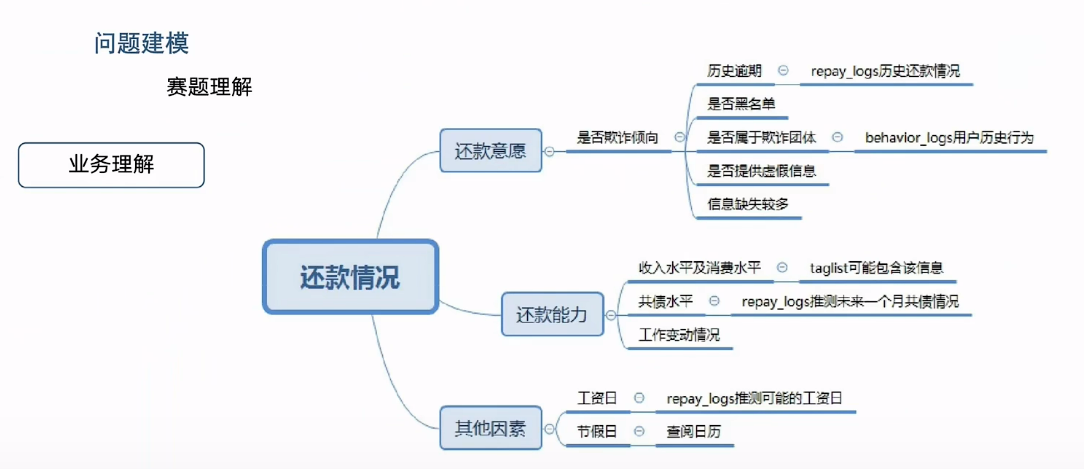
## 竞赛主要模块



开始可以参考别人的baseline也可以自己写（可以跳过数据分析的过程），在模型验证或者计算排名的时候发现效果不好，则对各个模块进行优化

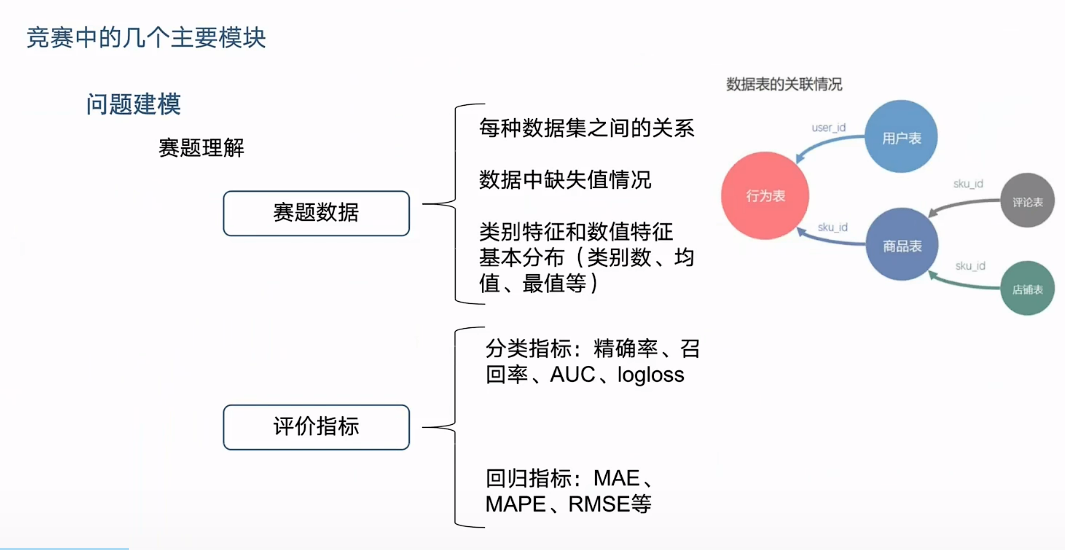
### 问题建模

#### 1.1业务理解



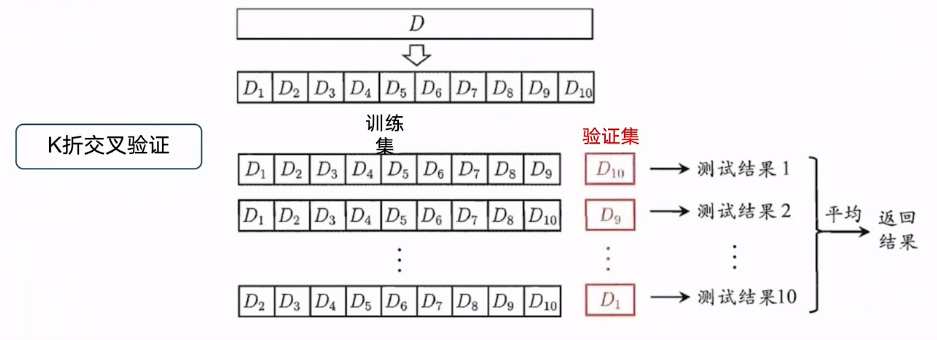
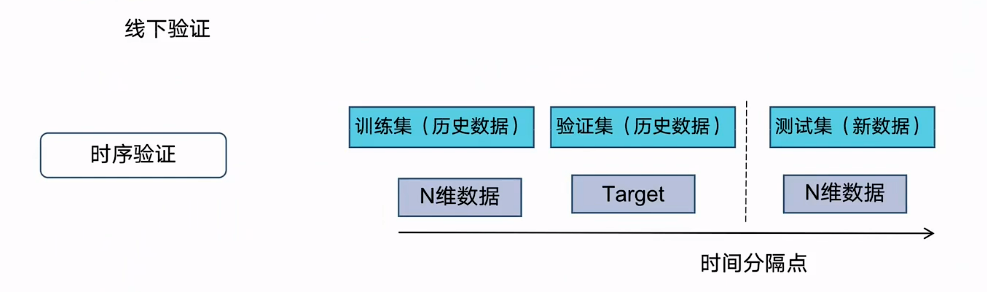
工业相关、金融风控等赛题，构造特征常见套路：统计（历史温度的变化、历史温度的趋势、和温度相关的字段有哪些）、聚合、趋势；或者需要围绕业务挖掘特征

#### 1.1.2数据理解



有多个数据文件时，先理清楚数据文件的关系（哪些数据没有价值、哪些数据有缺失值、哪些数据是类别特征（统计个数、属性数）、哪些是数值特征（最大值、最小值、中位数、方差））；确定评价指标：MAE、MAPE（回归）；精确率、召回率、AUC、logloss（分类）

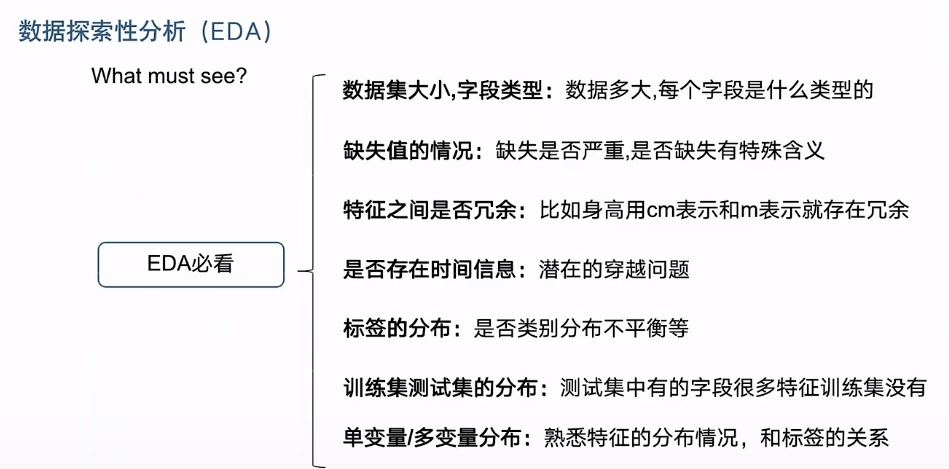
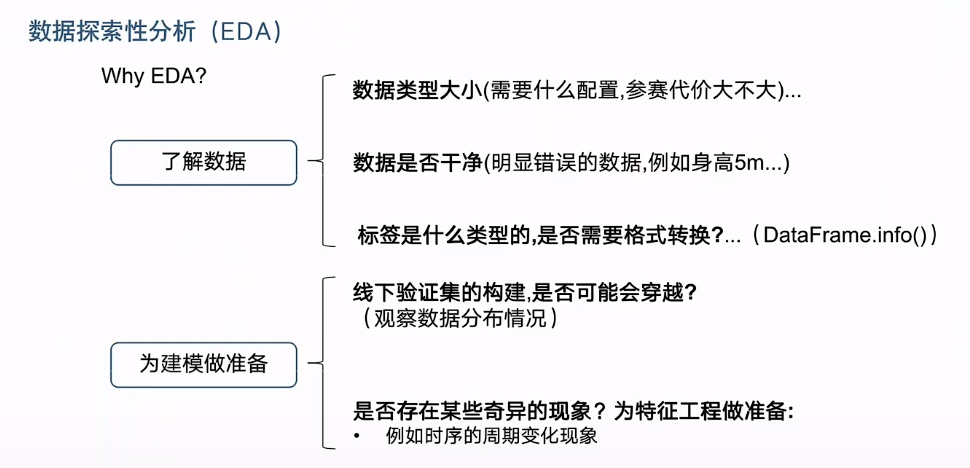
#### 1.1.3 线下验证



和时序相关的，选择按照时序进行预测，保证线上线下的一致性（和测试集时间比较接近的验证集去验证）；无关的可以选择K折交叉验证进行验算

### 数据探索性分析

了解数据是否存在异常数据（身高5cm）、训练集验证集数据分布是否一致（可能有些标签或者特征训练集出现，测试集没有出现）；lightgbm，XGboost不做缺失值处理，神经网络要做缺失值处理；

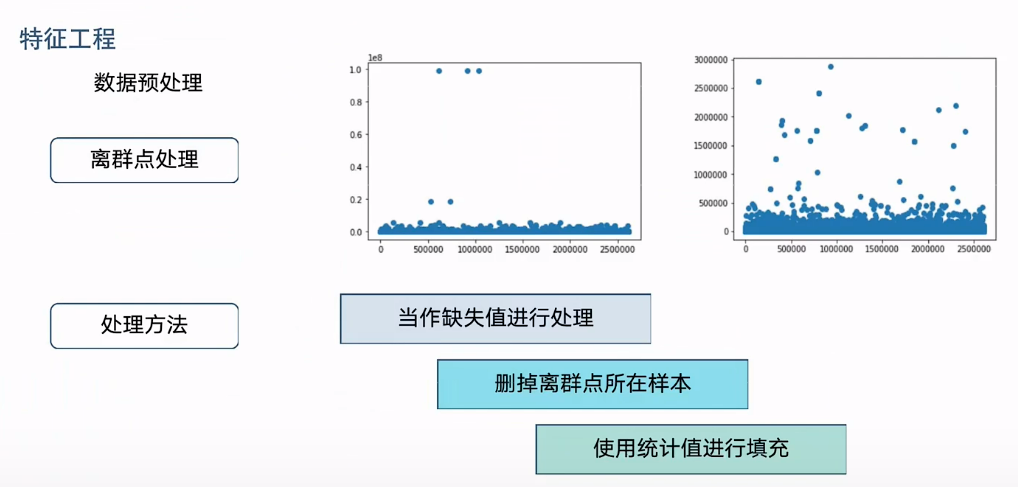


#### 1.2.1数据分析

离群点处理：过滤/当做缺失值处理（XGBoost可以处理缺失值）/使用统计值填充/删掉样本

缺失值处理：没有意义的数据设为np.nan；有特定的业务max(fea)+1/min(fea)-1(数模型一般不处理特征缺失)

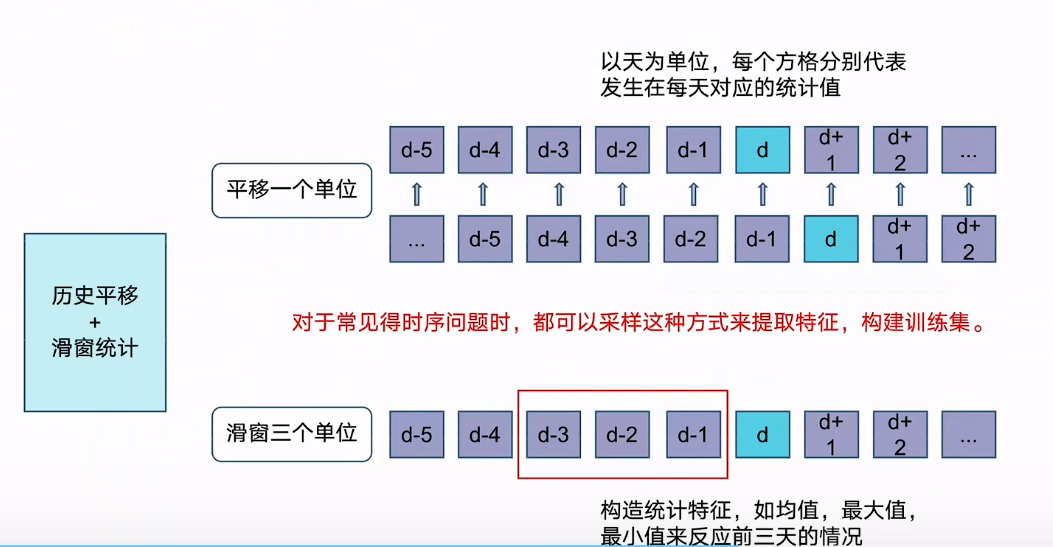
错误值处理：当做缺失值处理（-1、999都可能表示缺失值）、或者删除样本



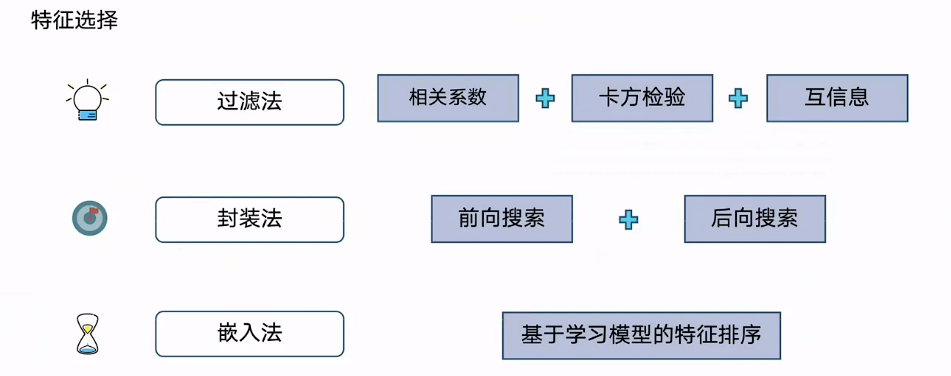
#### 1.2.2 特征工程

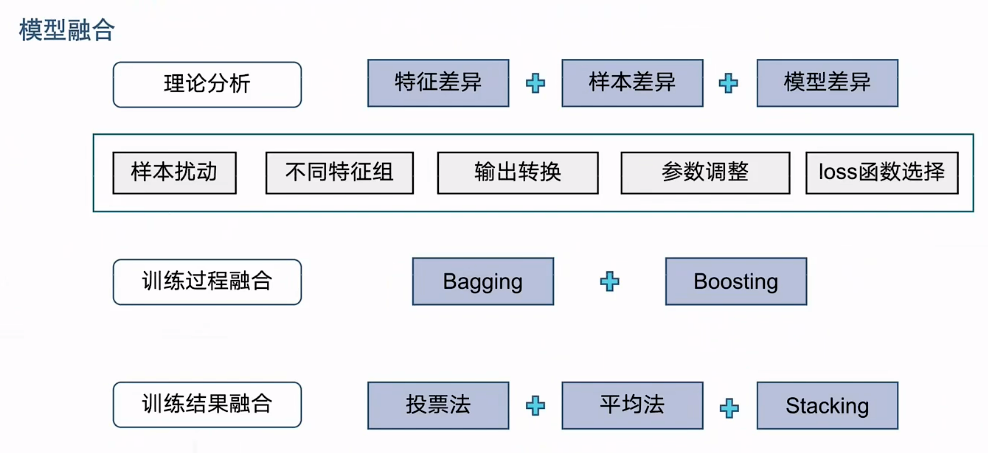
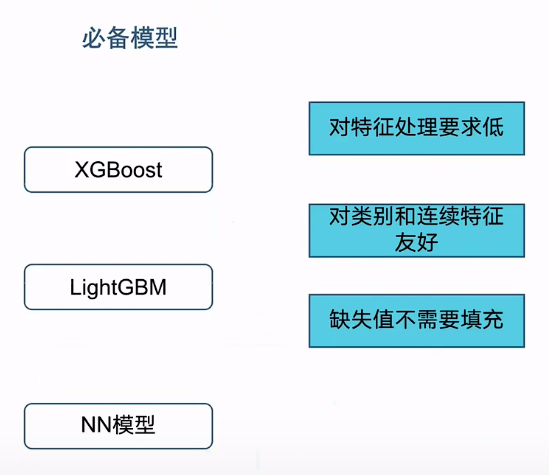
类别特征：string类型进行自然数编码->独热编码（count编码、目标编码）

数值特征：统计特征数据均值、方差、按照时间构建特征趋势、业务交叉构造、分桶、二值化

时间特征：拆分成日期变量（年、月、周、日、小时、分钟）；距离某天的时间差（是否某个特殊的日期，平滑处理）；时序相关特征（历史平移、滑窗统计）

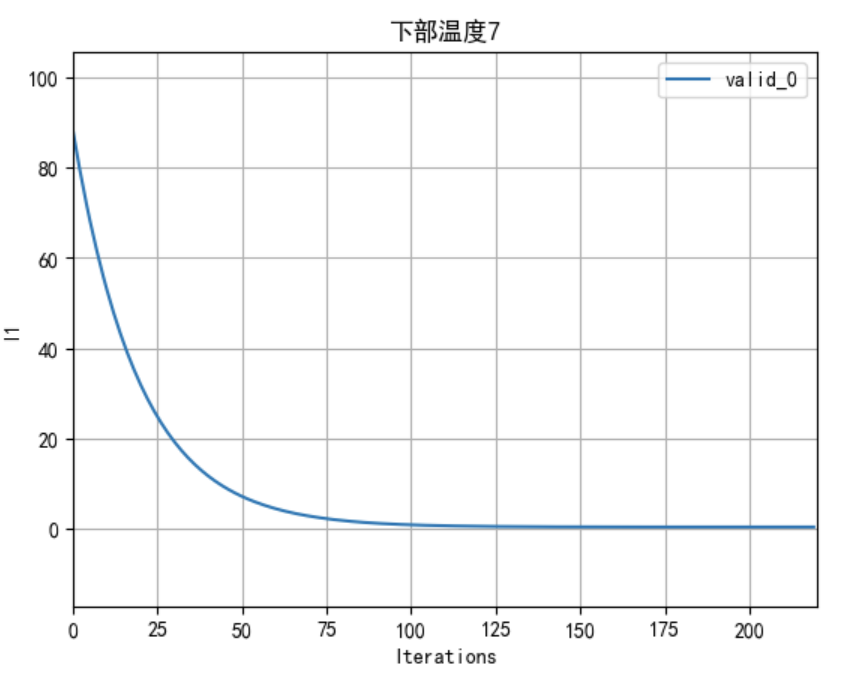
#### 1.2.3 特征选择

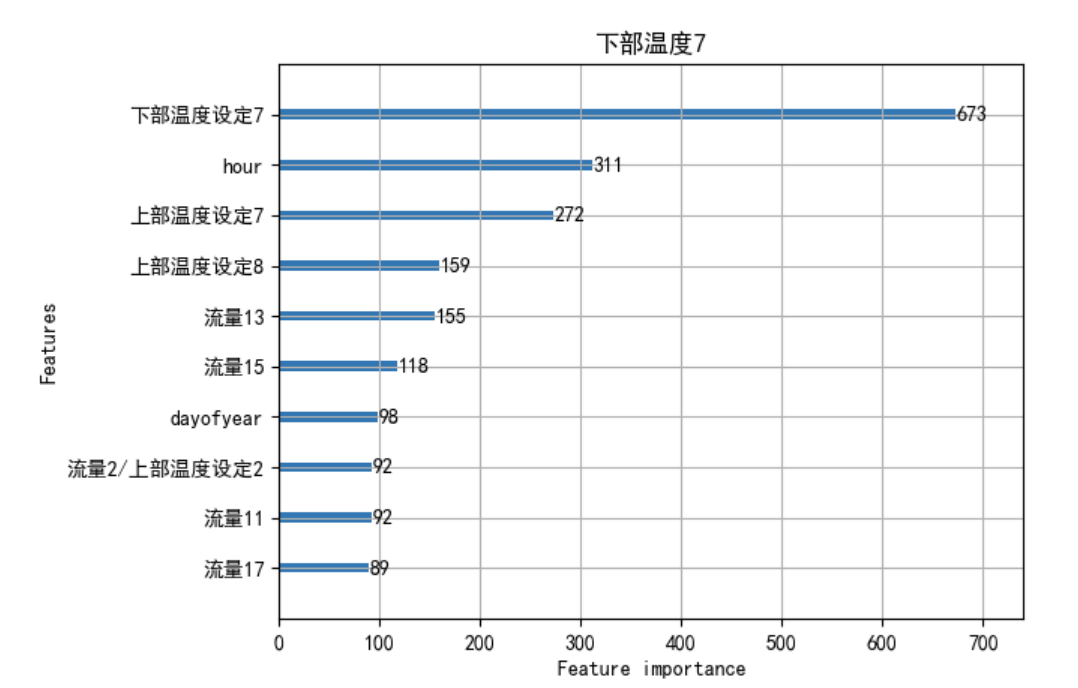


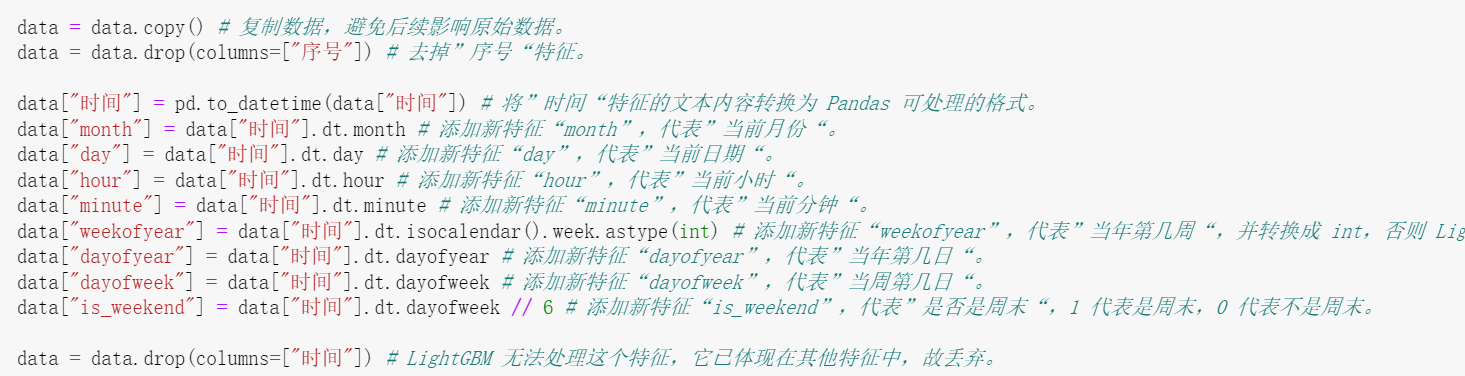


## 2.常见上分技巧

1. 通过绘制loss曲线观察模型训练中的loss收敛情况，并且绘制出特征重要性观察特征对于模型的影响

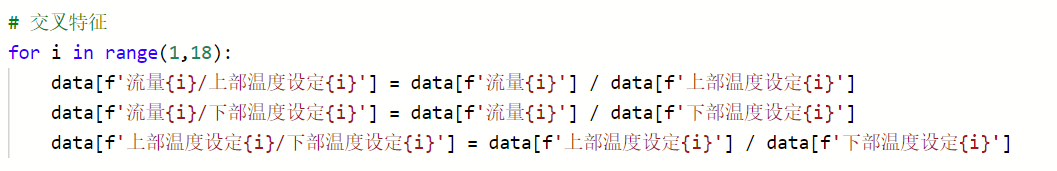




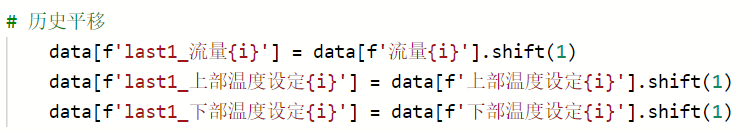
1. 特征工程处理，时间特征拆分代码如下

构建特征时，常用f来把变量字符串化，并且利用shift和diff很容易构建历史平移、差分、窗口统计数据

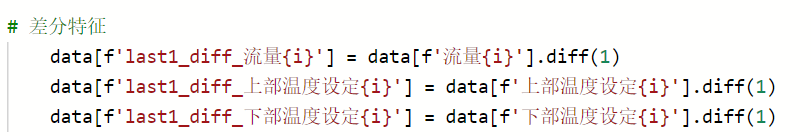
构建交叉特征



构建历史平移数据（不需要处理缺失值，shift是把数据整体下移）



构建差分数据（diff是把数据与上面的数据相减）



窗口统计数据

