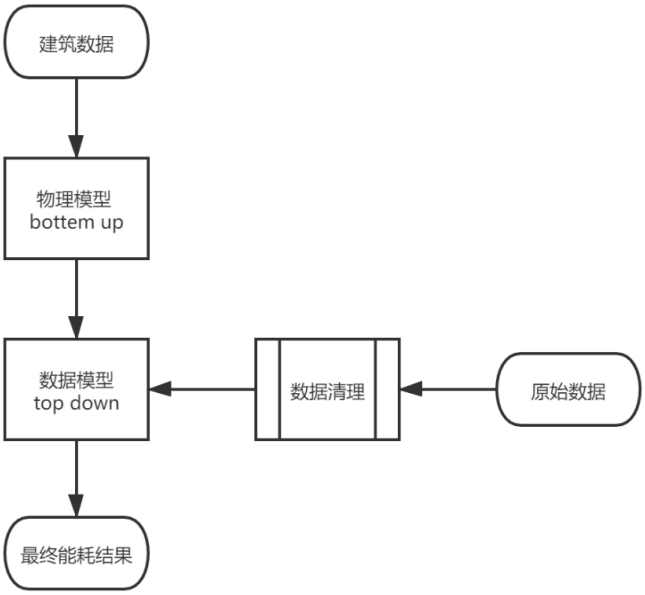
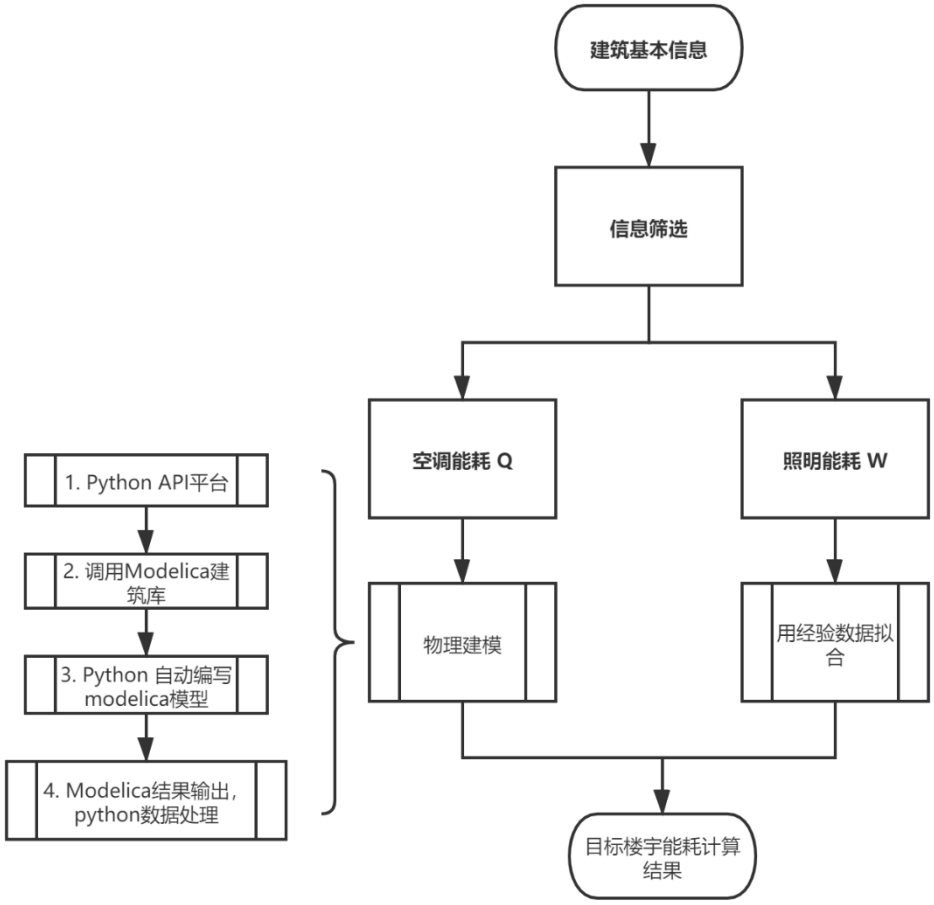
一，

本文采取了结合物理模型和数据模型的方式对目标楼宇进行能耗预测。第一部分在物理模型中通过建筑数据得到可信模拟结果。第二部分将第一部分结果与清洗过的其他楼宇历史数据共同导入数据模型中进行学习预测。从而得到最终预测数据。



二，第一部分：物理模型部分



1） 准备阶段

在众多建筑数据中选择关键性物理信息。应用集群算法及历史经验选出以下 7 个重要指标作为物理模型的输入变量：面积，楼层，U 值，窗墙比，HVAC 系统，天气数据及用途。【1】(De Jaeger, Reynders et al. 2020)

2） 建模阶段

空调能耗模拟分为以下 4 个小步骤：

1. 在 python 中建立 python 与 modelica 的接口
2. 用 python 调用 AixLib 建筑模拟库 【2】
3. 自动在 Modelica 后台中生成建筑模型
4. 用 python 读取并处理储存在 mat 文件中的模拟结果结果处理阶段

效果评价：优点——省去很多人力成本，不用对楼宇进行手动建模，适合区域性建模，能在半小时内完成上百栋楼的建模与模拟 【3】(Reinhart and Cerezo Davila 2016)

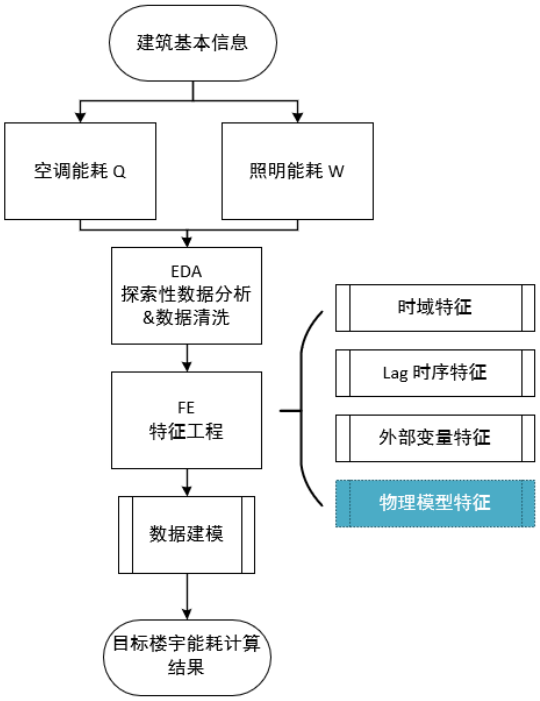
缺点——对商业建筑，非民用建筑精准度欠佳，因此需要利用其它模型修正，本文采取用数据模型修正。

照明能耗模拟直接用经验数值取标准负荷曲线进行拟合生成。

3） 数据整合处理阶段

将物理模型中输出的空调能耗结果根据设备 COP 转化为最终能源（电能）。

三，第二部分：数据模型部分



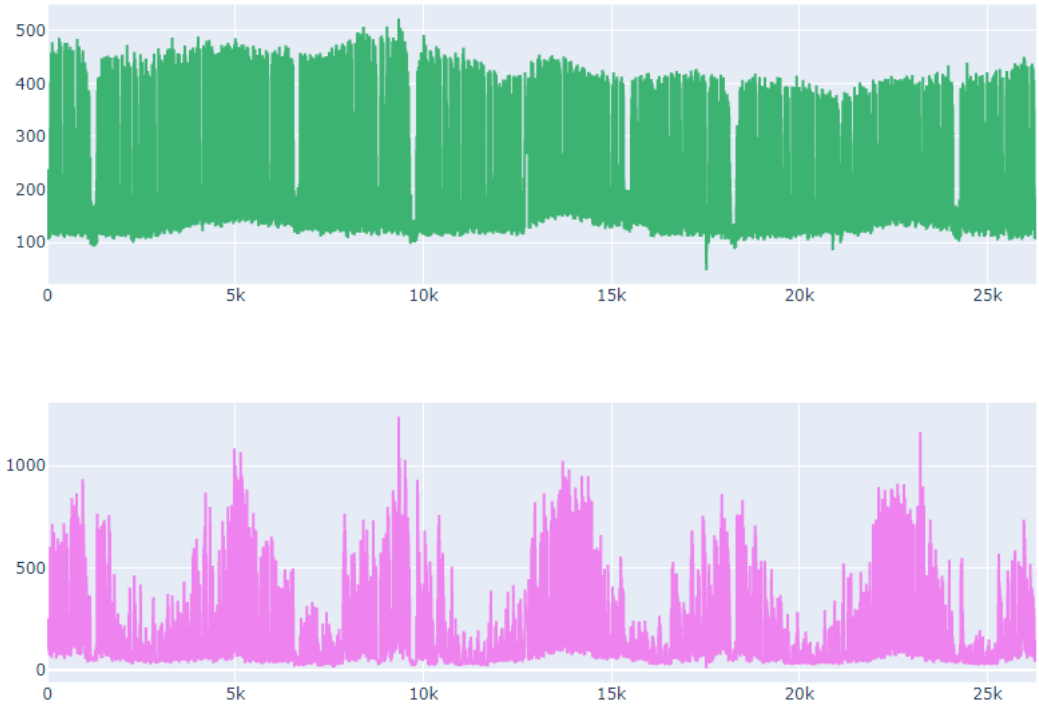
探索性数据分析 Exploratory Data Analysis

数据探索性分析是对数据进行建模预测的第一步。

数据清洗

对于训练数据，从 EDA 可视化中了解到数据中存在部分的异常值，
对于异常值的筛选：

- 选取数据平均标准差 3.5 倍进行标的
- 跳变异常值替换为均值
- 负异常值替换为零值



清洗后数据表现出稳定的周期性和趋势性，适合用 STL 时序方法进行分解分析。

相关性分析

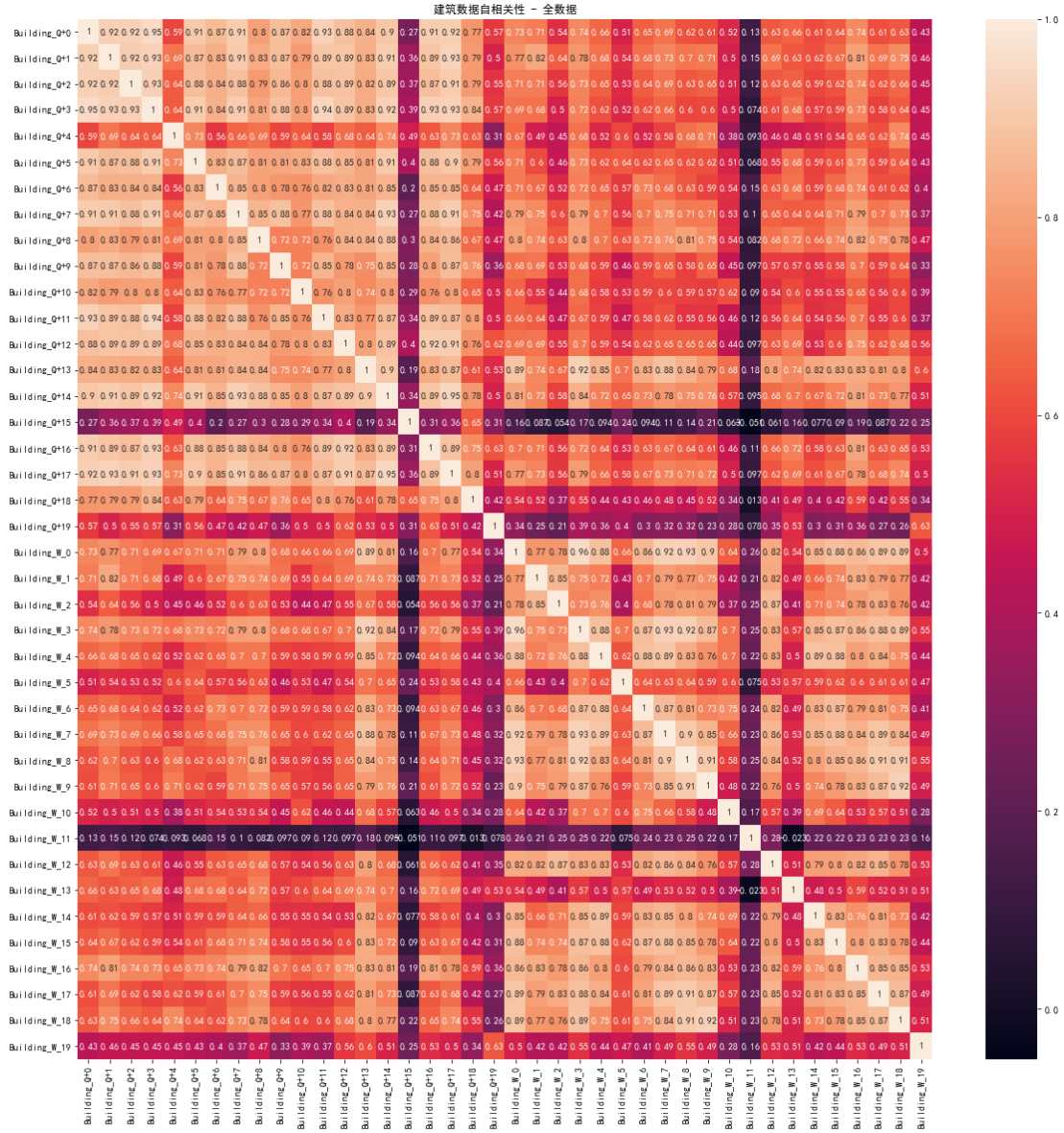
下图是针对 19 栋楼不同的能耗之间的数据进行 Pearson 矩相关系数分析，可以看出，空调能耗之间存在较高的矩相关性（左上矩阵），照明能耗存在较高的矩相关性（右下矩阵），其中，针对楼宇 4.10.13.15 分别存在着低相关性的异常现象，原因在于：1、部分楼存在某一时段的数据缺省；2、楼宇系统不同两点原因形成。另训练/测试集比例为 1/57（训练：19 栋楼，3 年数据；测试：一栋楼，1 年数据）为避免训练数据混入过多噪声。可只选取相似度最高，数据相对完整的楼宇数据作为训练集。

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

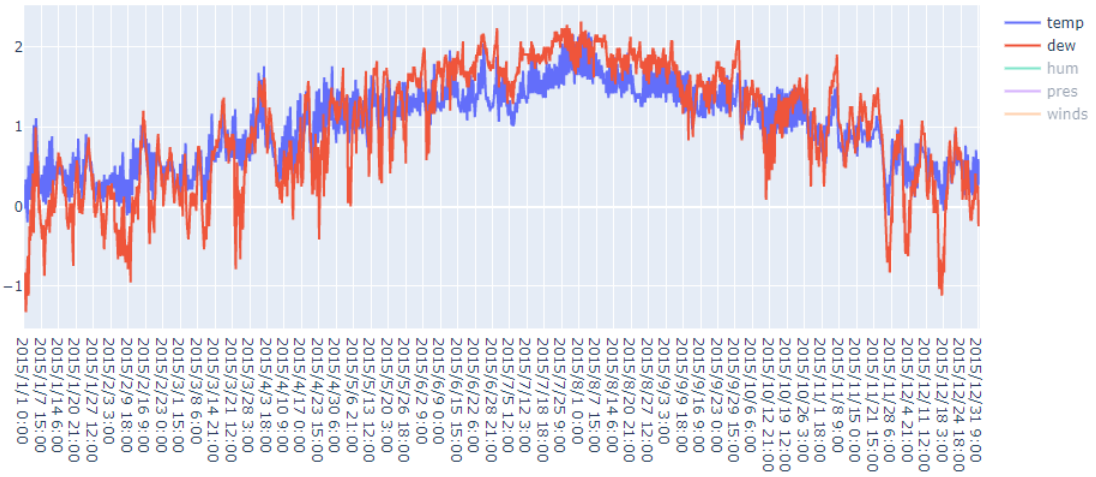
[1]

Pearson 矩相关系数

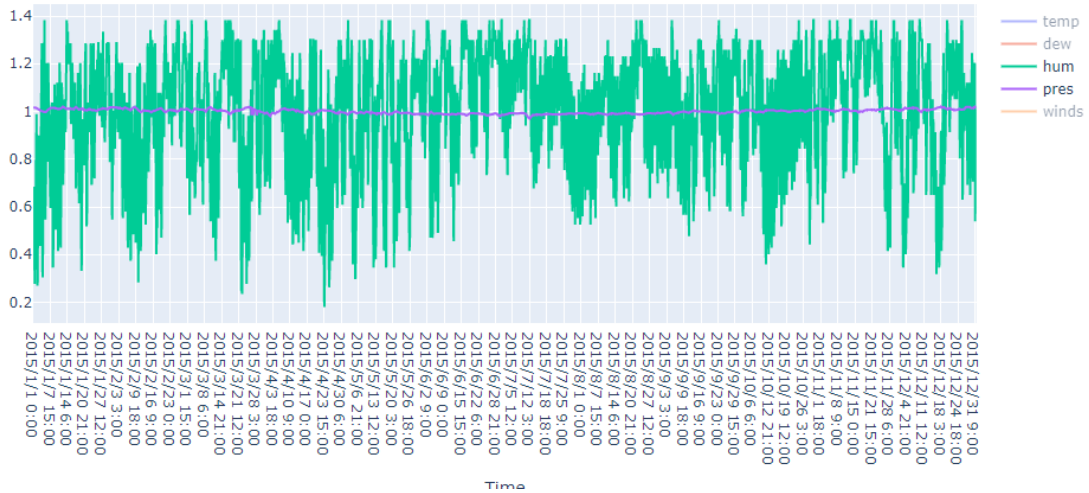
[1] Benesty, Jacob, et al. "Pearson correlation coefficient." Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009. 1-4.



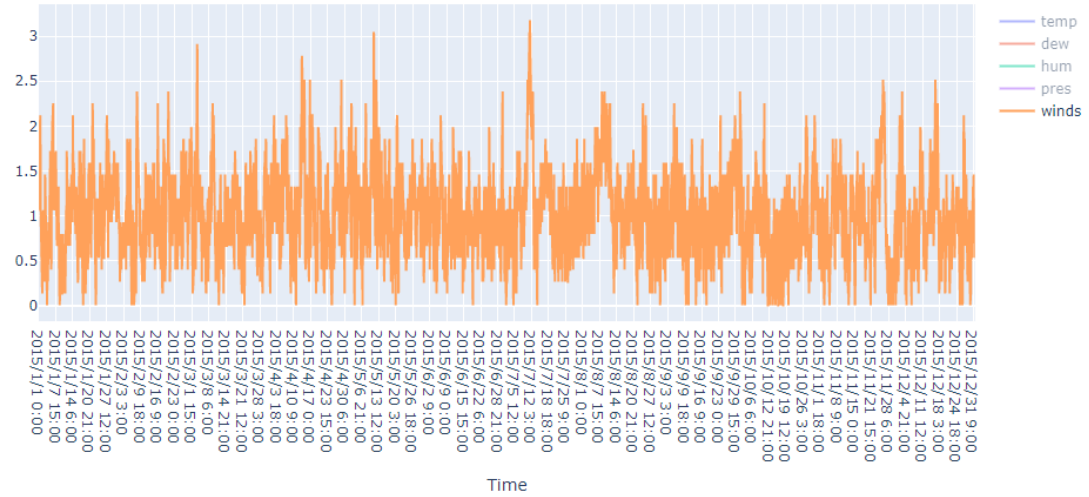
天气数据分析:



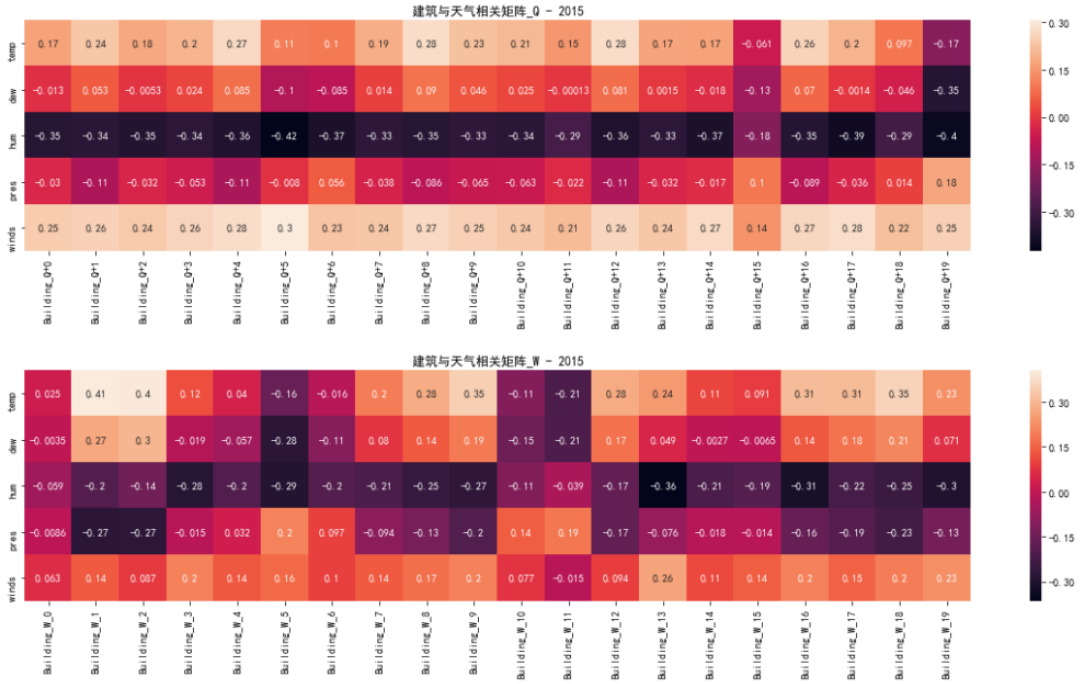
上图以 2015 年为例，气温与露点温度走势图（标准化后），可看出夏秋季气温较高，露点温度相对平稳，春冬季气温较低，露点温度波动幅度较大。



上图为湿度与气压标准化后走势，夏季气压较低，冬季气压较高，且气压变化幅度较低。同时，湿度分布较为随机



风速相对表现也较为随机，而且波动幅度较大。



以 2015 年为例进行皮尔逊矩相关分析，可以看出对应建筑（横轴）与天气数据（纵轴）的相关性，从而进一步排除异常值。总体而言，气温和风速对于建筑能耗有较强的正相关矩，同时，湿度与建筑能耗呈现负相关。相比而言，空调能耗 Q 的相关性表现较为明显，照明能耗表现的随机性较强。

特征工程

对于特征工程的建模将基于对数据 EDA 的结论驱动。特征工程主要由四部分组成：时域特征，时序特征，外部变量和物理建模输入

时域特征

对于楼宇能耗时间序列而言，存在明显的周期性，趋势性。对于 STL 传统时间序列建模成效较好。时域主要提取时间周期性特征，例如年、月、日、day of week，week of month，是否周末，是否节假日等

时序特征

楼宇能耗存在着较强的周期性，引入固定滞后项周期的值可强化模型的周期性特征。这里可以应用 day to day 模型或者 recursive 模型，对于本竞赛，我们团队使用的是 recursive 递归模型

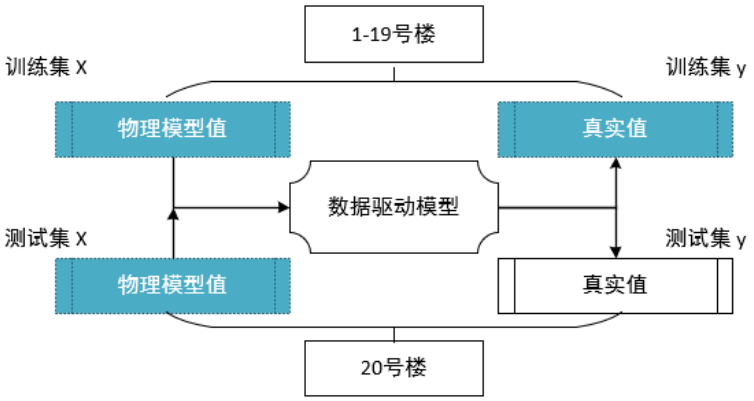
外部特征

外部特征在本竞赛中主要指天气数据。根据 EDA 和经验分析，外部天气数据（气温，露点温度）对于建筑物的空调耗能影响较大，时间（白天黑夜）对于建筑物照明耗能影响较大。同时，建筑的冷热负荷存在着较大的时滞性，所以在引入外部特征的时候，也加入了滞后项 lag 和滚动窗口项 rolling

物理模型数据输入：数据映射

对于本竞赛，难点在于对于目标预测建筑 20 号楼。训练集内无任何历史数据，对于数据驱动的预测模型而言造成了很大的制约。只有在建筑信息上（楼宇面积，地上地下层数和设备类型）与训练集有关联性。所以物理模型的驱动可以很好的解决数据联结的问题，通过同时

对 1-19 号楼的物理模型模拟结果和真实值的比较，可以给数据驱动模型提供较好的数据关联映射。另一方面，物理模型的输入也可以最大限度的利用本身建筑的元件参数。



在模型训练过程中，模型可以你和物理模型之和真实值之间的误差 pattern，从而强化 1-19 号楼和 20 号楼的能耗数据关联。

模型

本竞赛我们使用的是 Lightgbm 回归模型[2][3]，梯度提升（梯度增强）是一种用于回归和分类问题的机器学习技术，其产生的预测模型是弱预测模型的集成，如采用典型的决策树作为弱预测模型，这时则为梯度提升树（GBT 或 GBDT）。像其他提升方法一样，它以分阶段的方式构建模型，但它通过允许对任意可微分损失函数进行优化作为对一般提升方法的推广。模型是基于树状随机森林结构进行弱分类器的集合，核心梯度下降代码如下图：

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:
   1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (or weak learner, e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

四，方法特色

本文中使用的原创方法有以下三个特色

1. 结合了自下而上的物理模型及自上而下的数据模型对楼宇进行全方位模拟。
2. 物理建模部分不同于常见的精细建模方法，采用了 UBEM 建模思路，能快速高效的建立模型，生成模拟结果

---

[2] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems. 2017.

[3] https://github.com/microsoft/LightGBM

3. 数据建模与物理建模有机结合，通过数据建模对物理建模进行校核。

五，思考与局限性

根据最终结果来看，在 Q（空调能耗）精确度较高 0.2424，而在 W（照明能耗）上精确度低 4.141。可能原因是因为在物理模型阶段空调能耗采用了切实模型进行计算而照明能耗只采用了经验值，从而影响了第二阶段数据模型的学习与预测，导致 W 上精度偏低。

六，引用

【1】De Jaeger, I., et al. (2020). "A building clustering approach for urban energy simulations." Energy and Buildings **208**.

【2】Mueller, Dirk & Lauster, Moritz & Constantin, Ana & Fuchs, Marcus & Remmen, Peter. (2016). AixLib – An Open-Source Modelica Library within the IEA-EBC Annex 60 Framework.

【3】Reinhart, C. F. and C. Cerezo Davila (2016). "Urban building energy modeling – A review of a nascent field." Building and Environment **97**: 196-202.