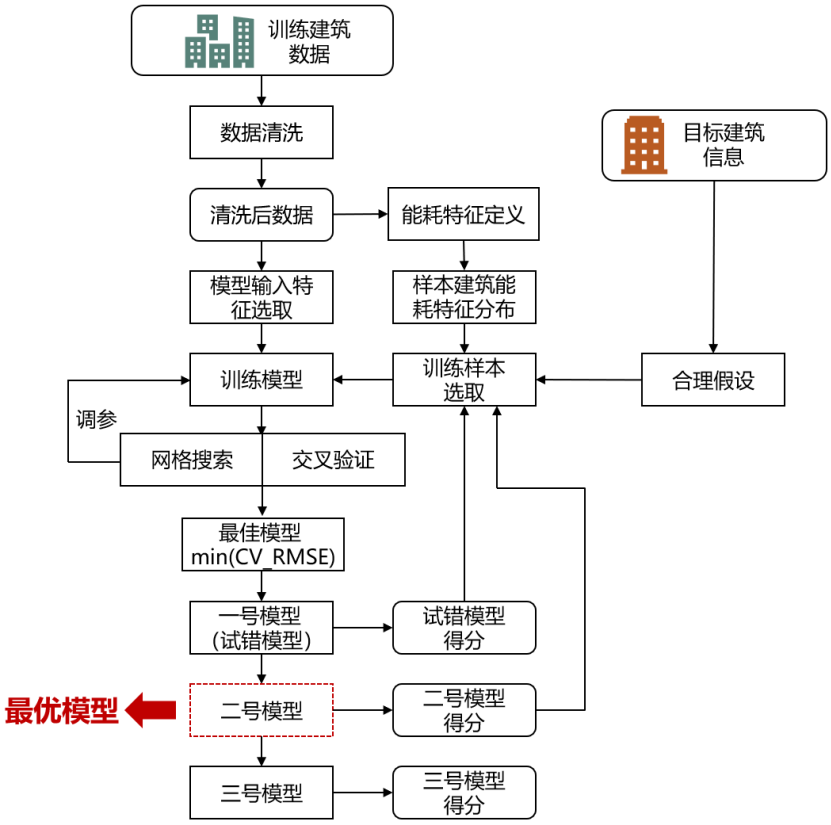

预测主流程示意图

本次预测任务的主要流程包括：

1） 数据清洗：对样本建筑的数据集进行数据清洗，数据清洗的主要内容包括离群值、死值（长时间数据保持不变）的处理以及可疑数据的判断。

2） 训练样本选取方法：定义了三个照明插座和空调能耗特征指标，通过分析样本建筑能耗的特征分布，结合目标建筑能耗特征的合理假设，分别确定了两种模型的训练样本的选取原则。

3） 模型输入特征选取：根据已有参考文献确定待输入变量，通过模型试训练选取预测结果最优的输入特征组合。

4） 模型建立：采用 Xgboost 算法分别建立照明插座和空调能耗模型，采用 5-折交叉验证和网格搜索方法寻找训练结果最佳（CV_RMSE 最小）的超参数取值组合，确定最终能耗模型。

5） 输出结果：输入目标建筑的模型输入特征，输出目标建筑的照明插座和空调能耗预测值。

1 目标建筑信息分析

分析目标建筑图纸发现，目标建筑由两部分组成，一部分是 30 层的塔楼，面积约占总面积的 75%；另一部分是 7 层的裙楼，面积约占总面积的 25%。图纸信息表明裙楼部分为商业用途，基于该信息做出以下推断：

目标建筑并非纯办公建筑，较高可能性为办公（塔楼）加商场（裙楼）的综合体。

基于该推断，考虑到商业区域的运营特征，该建筑的能耗可能有以下特征：

1. 节假日有一定量且稳定的能耗；
2. 工作时间结束后的 18 点至 22 点间，仍有一定量且稳定的能耗。
3. 能耗水平可能较高于一般办公建筑，但考虑到入住率、设备能耗水平、管理水平未知，该假设待验证。

此外，赛题信息表明目标建筑冷冻侧采用二次泵变流量系统，冷却水系统也为变流量，表明目标建筑在空调系统自控上可能较一般建筑更完善。

小结

通过分析目标建筑图纸，对目标建筑的能耗特征进行了推断，这些假设将作为我们分析训练建筑能耗特征并选取训练样本的依据。

2 数据清洗

2.1 去除离群值

对每一栋建筑，空调与照明能耗分开，若能耗数据低于该建筑 3 年能耗数据 0.005 分位数乘以 0.8，或高于 0.995 分位数乘以 1.2，则认定为离群值并移除。照明能耗去除了约 0.04%的数据，空调能耗去除了约 0.09%的数据。

2.2 去除死值

通过作图发现训练建筑能耗数据中存在死值，即较长的一段时间区间中能耗数据没有变化。计算每栋建筑每天能耗的标准差，若标准差小于 1，则认为是死值并移除当日数据。在去除异常值的前提下，照明能耗又去除了 4.0%的死值，空调能耗去除了 8.7%的死值。

2.3 可疑数据

除了异常值与死值外，我们还通过作图对能耗数据趋势的合理性进行判断，识别可疑的能耗数据。图 1 是训练建筑在三年中空调逐时平均能耗强度，观察图形有以下发现：

1、建筑 11 有非常特殊的用能特征；

2、建筑 19 的能耗在 2017 年较 2016、2015 年有非常大幅度的下滑；

3、建筑 5 三年的能耗水平间存在非常突出的差异；

4、多栋建筑在 2015 的能耗显著低于 2016 与 2017 年；

5、……（篇幅限制，照明能耗图形与结果未列出）

可疑数据并非可确认的错误数据，因此我们未将其去除，但是将尽可能避免在模型训练中使用可疑数据。

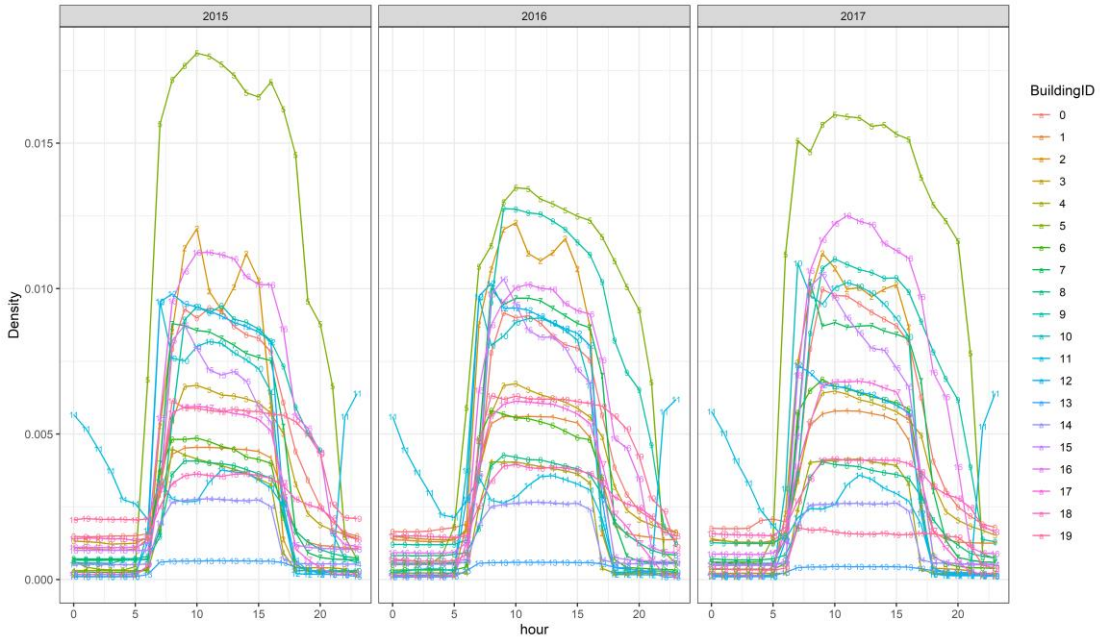

图 1 逐时平均空调能耗强度

2.4 小结

在进行数据清洗前，我们作图观察了各建筑能耗数据的分布，发现离群值的占比较小，因此增加了系数来缩窄离群值判定范围。死值占比较高，但由于死值往往连续分布在一个较长时间区间内，难以进行数据填充，因此选择直接剔除。此外，避免使用可疑数据。

3 训练建筑能耗特征分析

为了更清晰的认识训练建筑用能特征，我们提取了三个指标对训练建筑的用能特征进行分类：

1. 工作时间能耗强度：8 点至 18 点间的能耗强度；
2. 节假日相对能耗水平：节假日 8 点至 18 点间能耗强度与非节假日同时间能耗强度比值；
3. 傍晚非工作时间相对能耗水平：17 点至 21 点能耗强度与 8 点至 18 点能耗强度比值。

3.1 工作时间能耗强度

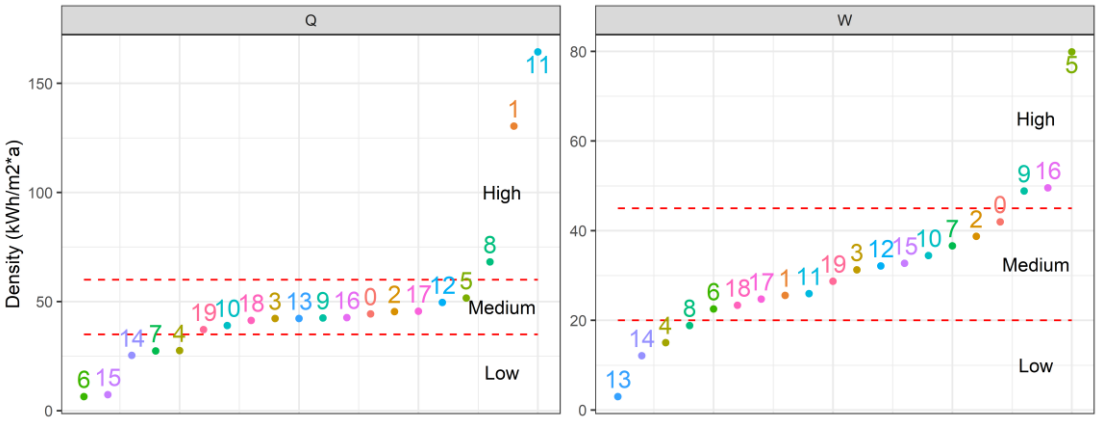

图 2 工作时间（8 点至 18 点）能耗强度，Q 为照明，W 为空调

图 2 按照建筑工作时间能耗强度依次排序，并将照明和空调能耗分别划分为低、中、高三级。训练建筑工作时间空调能耗强度的分布差异较大，照明能耗强度相对集中。通过模型试训练发现，赛题所提供的建筑信息并不足以解释建筑间能耗水平的差异，具体表现为当训练集包括能耗水平差异较大的建筑时，模型在训练集的准确度较差。一些可解释训练建筑间能耗水平差异的重要信息未知，例如入住率（直接影响全年能耗水平）、设备能耗水平（例如是否采用 LED 灯具，水系统是否变流量，空调热源是否为其他能源）、建筑能源管理水平（影响过渡季节空调能耗，非工作时间段能耗）、是否有商业区域等等，这些信息对模型预测准确度有非常重要的影响。缺少这些信息使我们无法判断预测建筑处于哪个区间内，因此我们采用试错模型来代替缺失信息对目标建筑能耗水平区间进行估计，具体在第 5 节进行讨论。

3.2 节假日与傍晚非工作时间相对能耗水平

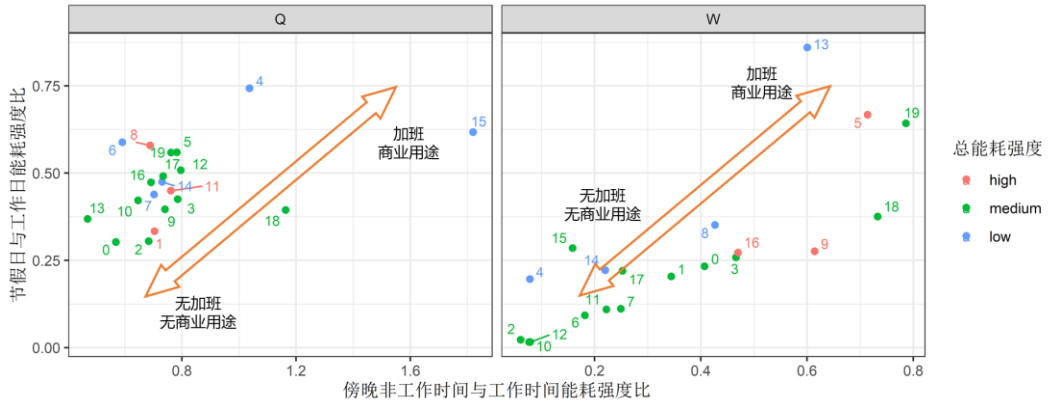

图 3 节假日与傍晚非工作时间相对能耗水平，总能耗强度按照低中高用不同颜色表示，Q 为照明，W 为空调

节假日与傍晚非工作时间相对能耗水平与建筑是否有加班情况或商业用途关系紧密。节假日与傍晚非工作时间相对能耗水平较高的建筑，较大概率有加班情况或商业用途，较低则相反。

训练建筑在空调能耗上显示出了明显的差异：2、10、12 号建筑在节假日与傍晚非工作时间基本无空调能耗；5、19、13 号建筑在节假日和傍晚非工作时间空调能耗相对水平达到

60%-80%。对照明能耗，除了 4、18 与 15 号建筑外，其他训练建筑的能耗特征分布较集中。

3.3 小结

通过计算三项能耗特征指标，我们发现训练建筑间的空调能耗特征存在较大差异。首先，由于入住率、设备数量及运行情况、建筑用能管理水平等信息未知，我们无法判断目标建筑的能耗水平区间；其次，节假日与傍晚非工作时间用能水平的结果表明，训练建筑的使用模式也存在较大差异，部分建筑例如 2、10、12 号建筑为典型的纯办公建筑，且运行时间固定，基本无加班情况；部分建筑例如 13、18、19 号建筑，有非常明显且稳定的加班情况或者商业用途。结合目标建筑的信息，分析得到的训练建筑能耗特征可以作为选择模型训练样本的依据。

4 模型建立

4.1 模型选择

在进行建筑能耗预测前，首先需要确定预测的方法，即白箱法和黑箱法。白箱法通过输入给定的建筑信息，通过能耗模拟软件来建立能耗模型输出预测能耗。这种方法结果的准确性非常依赖于所获取的建筑信息，但根据本次比赛中提供的建筑信息，我们认为尚不足以建立一个较为可靠的白箱预测模型，比如目标建筑的实际入住率、照明插座设备数量及控制方式、空调系统运行管理方式等信息均未知，会对白箱模型预测的能耗产生较大影响。黑箱法是通过数据驱动的方法来建立能耗预测模型，即通过建立建筑实际能耗数据与潜在的能耗影响因素的黑箱模型。相比白箱法，黑箱法建立能耗预测模型更为高效、灵活，并且近年来随着人工智能技术的发展，基于数据驱动的建筑能耗预测研究也取得了较多成果，基于黑箱法的能耗预测模型也表现出较高的可靠性(Deb et al., 2017)。因此，本次能耗预测模型的建立采用黑箱法。

采用黑箱法建立能耗预测模型时，模型预测的准确性将受到两个因素的影响：模型建立的算法和模型的输入特征(Fan et al., 2017)。本次建立的黑箱能耗预测模型采用了极限梯度提升算法 Xgboost，该算法可以看作是一种数值优化方法，旨在建立一个最小化损失函数的加性模型(T. Chen & Guestrin, 2016)。Xgboost 是基于梯度提升算法改进的算法，提高了梯度提升机（GBM）的计算效率并更好地解决了过拟合的问题。Xgboost 被广泛应用于多种工程领域以及机器学习竞赛中，并表现出较准确的预测效果和灵活性。在建筑长期或短期的能耗预测领域，与其他机器学习的方法相比，如支持向量机、人工神经网络、随机森林等，Xgboost 在基于大样本数据集建立预测模型时表现出更准确的预测效果以及更短的耗时(Fan et al., 2017; Touzani et al., 2018)。我们对比随机森林与 Xgboost 的结果也证明了这一点。

GBM 的核心思想为：首先通过一个决策树初始化模型最大程度地减少损失函数（回归问题上损失函数常为均方误差），在每次迭代步骤中添加一个新的决策树（即“weak learner”）来拟合现有的残差，将新的决策树添加到上一个决策树模型并更新加性模型的残差。该算法将一直进行迭代直到最大迭代次数满足为止，最终获得一个预测性能较好的学习器（即“strong learner”）。通过将决策树与残差进行拟合，可以在模型效果不佳的区域改进模型的预测效果。与 GBM 相比，Xgboost 在损失函数中加入了正则化项，从权衡偏差-方差的角度，该正则化项能够减少模型的方差，减少模型的复杂度从而避免过拟合；另一个较大的改进是采用了二阶泰勒展开式求解损失函数，加快了优化效率。详细的算法介绍可进一步阅读文献(T. Chen & Guestrin, 2016)。

本次照明插座和空调能耗预测模型建立的过程如图 4 所示。预测模型的建立基于 R 编程语言的开源包 RMV2.0，由美国劳伦斯伯克利实验室开发（https://github.com/samirtouzani/GBMbaseline）。

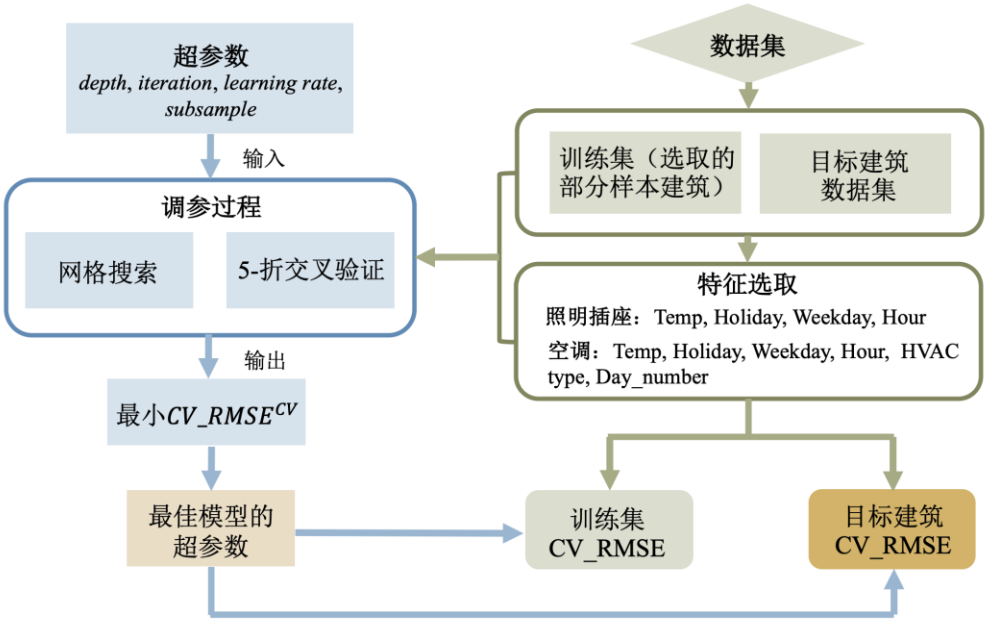

图 4 照明插座和空调能耗模型建立流程

4.2 特征选取

根据目前对办公建筑照明插座和空调逐时能耗的研究，空调逐时能耗的变化通常受到几个方面因素的影响：室外空气温度、节假日、周期特性（如每日或每周）、人员密度等(D. Chen et al., 2012)；而对于照明插座的逐时能耗，则主要受到节假日、周期特性（如每日或每周）、用户行为等因素影响(Anand et al., 2019)。因此，我们选择分别对照明插座和空调能耗建立模型。本次比赛提供的与建筑能耗相关的信息有限，主要包括室外气象参数、建筑面积、楼层、空调形式等，其中建筑面积、楼层这些变量主要影响建筑总能耗强度（年总能耗与总面积之比），与能耗数据的时序变化特征相关性不大。需要说明的是，照明能耗本身与气象参数相关性不大，但是由于照明能耗监测数据中可能混入空调能耗，因而温度也可能对照明能耗数据产生影响。

为了掌握逐时照明插座和空调能耗数据与其影响因素之间的关系，考虑纳入照明插座能耗模型的输入特征为：Holiday、Weekday、Hour、Temp；空调能耗模型的输入特征为：Temp、HVAC type、Holiday、Weekday、Hour、Day_number。由于采用了单位面积逐时能耗强度（逐时能耗/该建筑总面积）作为预测模型的输出变量，因此无需再使用建筑面积特征。各变量的名称和含义如表 1 所示。

表 1 考虑纳入照明插座和空调能耗模型的输入特征

变量	解释
Holiday*	中国法定节假日，包括周末，用 0，1 表示
Weekday	星期，表示为 1-7，分别代表一周 7 天
Hour	小时，表示为 1-24，分别代表一天 24 小时
Temp	室外空气温度
HVAC type	空调形式
Day_number	日编号，周期特性，表示为 1-365，分别代表一年 365 天

*Holiday 使用“RQuantLib”中“isHoliday”命令提取，“isHoliday”无法识别调班日期，因此调班日期手动添加。

在特征选取过程中，我们基于不同特征组合建立了多个模型，选取最优进而确定了表 1 的结果。在这一过程中，多个同类别的特征被用来横向对比。例如，我们比较了 Day_number、Week_number（一年中的第几周）、Month_number（一年中的第几月）以及 Season（季节）后，发现 Day_number 对模型的预测准确度提升更高，因此选择使用 Day_number。我们认为原因是 Day_number 相比同类特征有更高的解析度，可以为模型提供更多的信息。此外，在已经包含 Day_number 特征的情况下，其他同类别特征也无法提供更多的预测效力，无需再纳入考虑。

空调能耗模型的输入特征没有选用其他的室外气象参数如相对湿度、露点温度、大气压等，是因为：1）经验证，在室外空气温度基础上再添加相对湿度作为输入变量对模型的准确性几乎没有提高；2）露点温度表示在空气中水汽含量不变，保持气压一定的情况下，使空气冷却达到饱和时的温度，其简易计算公式可通过空气温度和相对湿度计算得到。根据提供的气象数据计算出露点温度与空气温度的皮尔逊相关系数为 0.8，因此未采用露点温度作为输入变量；3）已有研究未表明大气压对建筑空调能耗有明显影响。

4.3 调节超参数

Xgboost 算法有四个关键的超参数需要优化，分别为：1）决策树的深度（*depth*），用于控制回归树的复杂度，防止过拟合；2）最大迭代次数（*iteration*），对应回归树最大的数量；3）学习率（*learning rate*），用于给新生成的回归树的叶子结点的输出乘以一个收缩系数，防止某个回归树的影响过大，给后续迭代的回归树更多的学习空间，通常取值在 0-1 之间，数值越小，生成的决策树越多，模型复杂度也越高；4）子采样比例（*subsample*），通常默认为 0.5，即每次迭代过程中随机选取 50%的数据来训练数据，能够有效减少计算成本。以上四个待调超参数在建立模型时的取值区间及间隔如表 2 所示。

表 2 待调超参数的取值

超参数名称	取值
depth	[3,7]，间隔为 1
iteration	[50,300]，间隔为 25
learning rate	0.05, 0.1
subsample	0.5

机器学习的调参问题可描述为如何选取最佳的超参数组合来避免过拟合的问题，同时又能保证预测结果的准确性最高。网格搜索是机器学习领域中最常用并且最容易理解的调参方法。该方法通过遍历待调超参数可能取值的组合，分别对每一个超参数取值的组合建立预测模型，并计算模型准确性指标，最终选取准确性最高的超参数组合结果。本次预测任务采用变异系数（CV_RMSE）作为评价调参过程模型准确性的指标。

我们希望训练得到的预测模型对未知的数据集具有更好的泛化能力，因此需要对模型进行验证来确定对于未知数据集的预测效果。若采用全部的数据集作为训练集，并采用该训练集上的预测准确度来衡量模型的表现，不能真实地反映预测模型在未知数据集上的预测效果。因此，我们采用 k-折交叉验证方法调参。

k-折交叉验证方法是将样本数据集随机划分为 k 个大小大致相等的子样本，依次取一份作为测试集，其它份（k-1）作为数据集进行训练，然后取 k 次结果的平均值作为此模型的结果，如公式（1）。为了避免随机划分数据集时出现测试集和训练集的统计特征出现较大差

异，比如测试集大多为低用电时段的时间序列用电数据，而训练集相反，在随机划分数据集时，我们将时间序列按照“周”来分段，按照周为单位随机抽取数据。

$$CV\_RMSE^{CV} = \frac{1}{k}\sum_{i=1}^{k} CV\_RMSE_i \quad (1)$$

结合网格搜索的调参方法，对每个超参数的取值组合采用 k-折交叉验证方法来估计该模型的CV_RMSE^{CV}，最后最优模型对应CV_RMSE^{CV}最小的预测模型。k 通常取值 5 或 10，k 值越高，计算耗时也越长，因此本次预测任务中选择 k=5，即 5-折交叉验证。调参的结果在表 3 中一并展示。

4.4 预测模型结果

三次提交结果的照明插座和空调能耗预测模型的信息如表 3 所示，其中包括了采用的训练样本建筑、调参结果以及准确性结果。第二次提交的照明插座和空调能耗模型预测的结果均为最佳，在目标建筑 100%的数据集上计算得到的 CV_RMSE 分别为 0.2172 和 0.7893。在计算成本上，训练单个建筑一年的能耗数据大约需要 1min （包括 k-折交叉验证调参的时间），可见基于 Xgboost 算法建立预测模型的计算效率较高。

表 3 三次提交的预测模型结果，最优模型加粗标示

模型	提交次数	训练样本建筑	选取年份	模型输入变量	模型调参结果	耗时（min）	训练集 CV_RMSE	目标建筑 CV_RMSE
照明能耗模型	第一次	0，2，3，9，10，12，16，17	2017	Temp, Holiday, Weekday, Hour	depth=4, iteration=175, learning rate=0.1, subsample=0.5	9	0.19	0.2172
	第二次	**0，2，3，9，10，12，16，17**	**2017**	**Temp, Holiday, Weekday, Hour**	**depth=4, iteration=175, learning rate=0.1, subsample=0.5**	**9**	**0.19**	**0.2172**
	第三次	2，3，10，12，16，17	2016，2017	Temp, Holiday, Weekday, Hour	depth=4, iteration=275, learning rate=0.1, subsample=0.5	11	0.2	0.2436
空调能耗模型	第一次	16	2016，2017	Temp, Weekday, Hour, Day_number	depth=7, iteration=200, learning rate=0.05, subsample=0.5	2.6	0.21	1.6868
	第二次	**1，3，6，17，19**	**2016，2017（19 号无 2017 年）**	**Temp, Weekday, Hour, Day_number, HVAC type**	**depth=7, iteration=300, learning rate=0.05, subsample=0.5**	**14**	**0.52**	**0.7893**
	第三次	17，18，19	2016，2017（19 号无 2017 年）	Temp, Weekday, Hour, Day_number	depth=6, iteration=175, learning rate=0.05, subsample=0.5	7	0.66	0.8696

5 三次提交模型训练样本选取的讨论

本节对三次提交模型结果所采用的训练建筑样本依次进行了讨论。

5.1 一号模型——试错模型

在我们所提交的三次结果中，第一次的空调能耗模型最特殊，我们称之为“试错模型”。通过第三节的分析发现，由于缺少入住率、设备用能水平、建筑能源管理水平等关键信息，无法解释训练建筑之间所存在的显著的空调能耗水平差异。“试错模型”存在的意义即代替未知的上述建筑信息，对目标建筑的空调能耗水平区间进行估计。我们在“试错模型”训练样本的选取上，遵循两个原则：

1. 由于得分为 CV_RSME 指标，该指标无法表明偏差的方向，因此所选取训练样本的空调能耗应为训练建筑中的高水平或低水平，从而锁定偏差的方向；否则，可能需要第二个“试错模型”才能确定正确的区间。
2. “试错模型”的空调能耗时间分布特征应与目标建筑尽量吻合，例如节假日有一定量且稳定的能耗，傍晚非工作时间有一定量且稳定的能耗。减少形态特征偏差（shape bias，例如上升时间、峰值时长、下降时间）在总偏差中的贡献率，增加量级偏差（magnitude bias）在总偏差中的贡献率，尽可能准确地依据提交结果估计目标建筑能耗区间。

基于上述原则，我们选取 16 号建筑的 2016 年与 2017 年空调能耗数据作为一号模型（“试错模型”）训练样本，如图所示。16 号建筑的空调能耗水平仅次于 5 号建筑，而 5 号建筑由于 3 年能耗水平波动较大，被认为是可疑数据。此外，16 号建筑节假日与傍晚非工作时间能耗水平均位于中上水平。

对于照明插座能耗，大部分建筑能耗强度处于“medium”的水平，且节假日与傍晚非工作时间相对能耗水平在建筑间差异相对较小（除去离群的 3 栋建筑），我们认为目标建筑照明插座的能耗特征处于图中高密度区域的可能性较高，于是选取 0，2，3，9，10，12，16，17 建筑 2017 年的能耗数据作为照明能耗模型训练样本。这些建筑分布在能耗水平最集中的区间，且节假日与傍晚非工作时间相对能耗水平两个特征也差异不大。

“试错模型”所选择的训练建筑样本如图 5 所示。

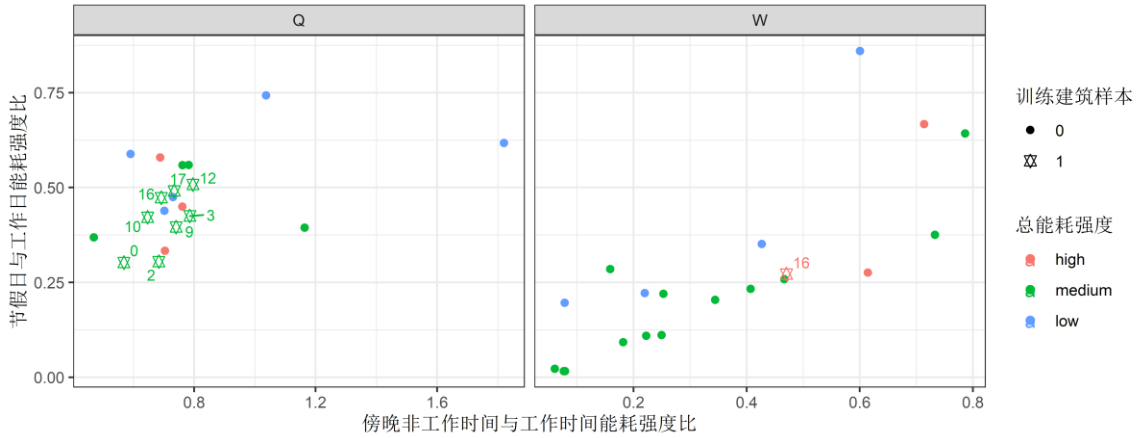

图 5 试错模型训练建筑样本，星号表示，Q 为照明，W 为空调

5.2 二号模型

提交“试错模型”后，我们得到基于 75%数据计算的 CV_RSME 为 1.2394。根据照明能耗预测模型在训练集上的表现，我们假设照明部分误差为 0.3，计算得到空调部分误差约为 1.64。考虑到 16 号建筑已经是训练建筑中空调能耗水平第二高的，目标建筑不太可能高出“试错模型”并达到 1.64 的差距，因此我们认为目标建筑能耗水平小于 16 号建筑。通过计算发现，当目标建筑能耗为“试错模型”结果的一半时，CV_RSME 的结果为 1.55。“试错模型”的空调能耗强度为 49kWh/m^2，我们推断目标建筑的空调能耗区间在 25kWh/m^2 至 35kWh/m^2 之间。

基于上述推断，在该能耗强度区间内选取 1，3，6，17 号建筑 2016 与 2017 年，19 号建筑 2016 年的空调能耗数据作为空调模型训练样本（19 号建筑 2017 年数据的大幅下滑存疑）。11 号建筑虽然处于这个区间但是因为用能特征异常，被认定为可疑数据。所选的 4 栋建筑中，19 号有非常显著的加班或商业用途，1、3、17 有一定程度的加班或商业用途，6 号则在节假日和非工时间能耗水平较低。这样选取的原因是，即使空调能耗强度大致处于同一水平的建筑，其余两项特征仍然存在较大差异，比如图中能耗强度为“medium”的建筑仍然分布分散。因此，为了能够让空调能耗预测模型有更好的泛化能力，我们选取了其他两个特征有一定差异的建筑作为训练样本。

由于缺少信息，本次提交未对照明能耗预测模型的训练样本进行调整，二号模型所选择的训练建筑样本如图 6 所示。

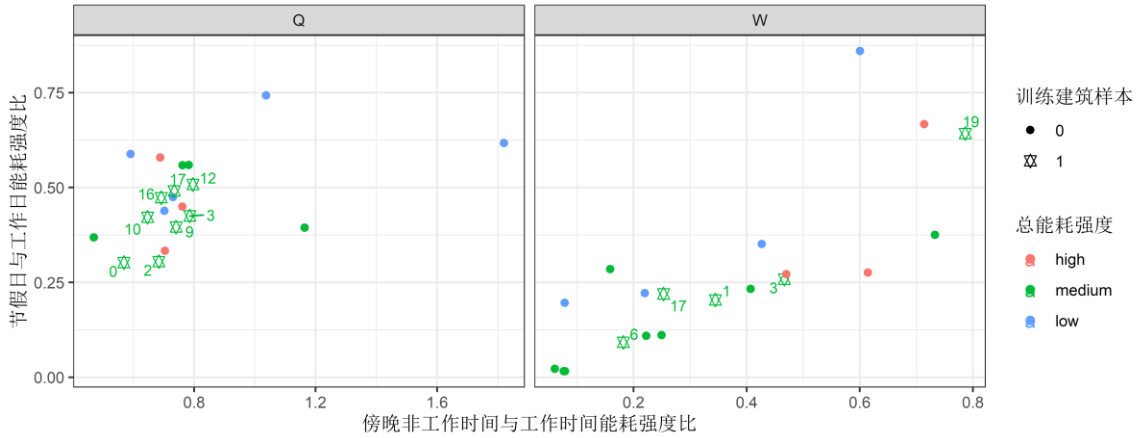

图 6 二号模型训练建筑样本，星号表示，Q 为照明，W 为空调

5.3 三号模型

第二次结果的加权 CV_RSME 值为 0.6144，相比“试错模型”有显著的提升，证明我们的推断方向正确。同样假设照明部分误差在 0.3，则空调部分误差应在 0.75 左右。

第二次提交结果的空调能耗强度为 28 kWh/m^2·a，参考《2017 年度上海市国家机关办公建筑和大型公共建筑能耗监测及分析报告》，空调能耗占比约 30%，推算得到总能耗强度约为 94.50 kWh/m^2·a。这一能耗强度在办公建筑与综合建筑中均处于较低水平，目标建筑在设备用能水平与设备运行管理上可能处于较高水平。

观察空调能耗数据发现，由于在第二次模型训练中使用了 1 号和 3 号建筑的数据，而 1 号和 3 号建筑的过渡季节能耗水平偏高，造成第二次提交结果的过渡季节能耗也偏高。训练建筑间的过渡季节用能存在较大的随机性，同一建筑在不同年份的用能水平有较大的差异，同一类

型的空调系统间也存在较大差异。我们认为过渡季节空调能耗水平应当与建筑用能管理水平、空调系统自控等因素密切相关。考虑到目标建筑用能水平较低，可能得益于其较高的能耗管理水平，有较大概率在过渡季节全新风工况运行；另外，通过赛题信息得知目标建筑冷冻侧为二次泵变流量系统，冷却侧也采用了变流量系统，目标建筑的空调自控系统可能较一般建筑更加完善，末端 VAV 系统也有较大概率可变频运行，过渡季节的能耗水平可能不会太高，因此我们第三次结果的改进主要针对过渡季节能耗的修正。

此外，虽然目标建筑仅有 25%的商业区域，考虑到商业建筑的能耗水平普遍高出办公建筑较多，因此我们还尝试继续增加节假日的能耗强度。

基于上述假设，在空调能耗模型训练样本中，我们去掉了过渡季节能耗偏高的 1 号和 3 号建筑，节假日能耗较低的 6 号建筑，并加入了节假日与非工作时间能耗较高、且过渡季节能耗水平较低的 18 号建筑。

对于照明能耗模型，同样出于增加节假日能耗水平的目的，我们去掉了 0 号与 2 号两栋节假日能耗水平较低的建筑。

三号模型所选择的训练建筑样本如图 7 所示。最终得分结果表明，我们第三次模型所依据的假设可能是错误的，加权 CV_RSME 上升到了 0.6818，空调和照明部分误差都分别增加。

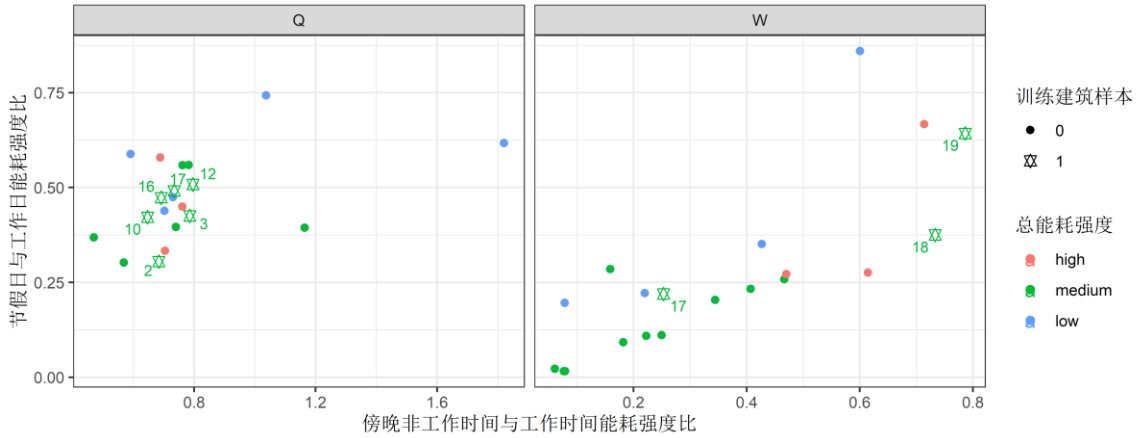

图 7 三号模型训练建筑样本，星号表示，Q 为照明，W 为空调

5.4 小结

对于照明插座能耗，当能耗强度处于中间水平，样本建筑其他两项特征差异不大，且该区间建筑数量最多，我们假定目标建筑的能耗特征较大概率与这些建筑相似，由此选取了部分能耗强度位于中间区域、其他两项特征差异较小的建筑作为训练样本。

空调能耗预测模型的建立相对复杂，从样本建筑的空调能耗特征上看，尽管能耗强度相似，但在其他两项特征上样本建筑仍然存在较大差异，这些差异可能会对预测模型的准确性有较大影响。我们三次建立预测模型的思路是，首先建立“试错模型”来确定能耗强度的大致区间，选取一个能耗强度偏高、其他两项能耗特征较符合办公及商业功能的样本建筑（16 号）来建立“试错”模型，依据得分判断目标建筑大致的能耗强度区间。基于推断的能耗强度区间，重新筛选能耗强度处于该区间的样本建筑，挑选节假日与傍晚非工作时间相对能耗水平有一定差异的建筑，从而保证模型的泛化能力，这次建立的预测模型取得了显著的提升。三号模型由于缺少进一步改进的信息，仅依据目标建筑在过渡季节能耗强度较低的假设，通过重新筛选样本建

筑，降低模型在过渡季节的空调能耗强度，然而该模型预测结果相对于第二次降低了，说明我们关于过渡季节空调运行情况的假设可能有误。

6 创新点

基于照明插座和空调能耗预测结果得分均为最佳的二号模型，我们总结了本次建立能耗模型的方法的创新之处如下：

1） 定义了照明插座和空调能耗的三个特征，找出相似的或有代表性的样本建筑，为训练样本建筑的选择提供依据。

2） 结合样本建筑照明插座和空调能耗的特征分布，分别建立了两种模型的训练样本的选取原则。照明插座能耗模型的训练样本选取多个相似的建筑，原因是样本建筑的照明插座能耗特征差异不大；对于空调能耗模型的训练样本，由于空调能耗特征存在较大差异，因此选取多个代表性建筑，即能耗强度相近，但其他两个特征有一定差异的建筑，保证训练样本的特征具有一定的多样性，使建立的模型具有更好的泛化能力。

3） 照明插座和空调能耗模型建立均采用了目前机器学习算法中表现卓越的 Xgboost 算法，采用 5-折交叉验证和网格搜索的方法调节模型的超参数，权衡了模型的偏差和方差，避免过拟合的同时提高准确性。

参考文献

Anand, P., Cheong, D., Sekhar, C., Santamouris, M., & Kondepudi, S. (2019). Energy saving estimation for plug and lighting load using occupancy analysis. *Renewable Energy*, *143*, 1143–1161. https://doi.org/10.1016/j.renene.2019.05.089

Chen, D., Wang, X., & Ren, Z. (2012). Selection of climatic variables and time scales for future weather preparation in building heating and cooling energy predictions. *Energy and Buildings*, *51*, 223–233. https://doi.org/10.1016/j.enbuild.2012.05.017

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, *74*(March), 902–924. https://doi.org/10.1016/j.rser.2017.02.085

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy*, *195*, 222–233. https://doi.org/10.1016/j.apenergy.2017.03.064

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039

上海市住房和城乡建设管理委员会. 2017年度上海市国家机关办公建筑和大型公共建筑能耗监测及分析报告. 上海，2018年. https://max.book118.com/html/2018/0606/171057075.shtm