基于贝叶斯优化XGBoost（Bayes-XGBoost）碳排放量预测

## 1 实验结果与分析

## 1.1 Bayes-XGBoost模型预测分析

将碳排放序列中的前t期的值组成的矩阵作为输入特征数据，第t+1时期的数据值作为输出数据，以上两组数据构成了监督学习的一个样本，t为模型训练的时间窗口。本文经过对比调试，选择时间窗口t=8。在模型训练过程中，使用一种称为前向验证的技术，在前向验证中，首先通过选择一个切点（本文使用除过去8年外，所有数据用于训练，最近3年用于测试），通过对训练数据集进行训练并预测测试数据集的第一步来评估模型。然后将来自测试集的真实观测值添加到训练数据集中，重新拟合模型，再让模型预测测试数据集中的第二步。对整个测试数据集重复此过程将为整个测试数据集提供一步式预测，可以从中计算出误差度量以评估模型的性能。在模型预测时，采用前向滚动窗口预测方法[1]，即在进行样本外预测时，将样本内数据作为特征数据传入模型得到的一步预测值扩充入样本数据内，并再一次使用扩充后的样本重复以上过程。

对江西省及11个市区的碳排放强度按照上述方法进行训练并预测未来30年的碳排放强度，变动趋势如图1所示，根据预测结果，江西省及11个市区的碳排放强度均在2030年前达到峰值。2020年后，江西省及11个市区碳排放均在攀升一段时间后呈现震荡回落趋势，其中新余市和鹰潭市震荡幅度较大，其他地区均在一段时间后趋于较小振幅的回落趋势。模型预测2030年江西省碳排放总量将达到4698.33×10^4 t 。

以江西省碳排放预测为例，使用机器学习可解释性框架SHAP[2]计算出最终模型的前40个样本的特征分布。由于该数据的前一步的输出作为后一步最后一个特征输入，所以按照时间顺序列样本的分布图如图二所示，x轴刻度为2009年以后的样本按照时间顺序的编号，y轴为碳排放量，图中有8条曲线，分别代表前8年的特征输入，红色区域代表输出增大区域，蓝色区域代表输出值减少区域，当红色区域面积大于蓝色区域是，输出碳排放量为增大趋势，当红色区域面积小于蓝色区域时，输出碳排放量为减少趋势，当红色区域面积与蓝色区域面积相当时，表示输出碳排放量区域稳定变化趋势。由图可知，2009年以后江西省碳排放量先是增大，然后减少，最后趋于稳定，并且缓慢减少，这与前一步分析结果一致。

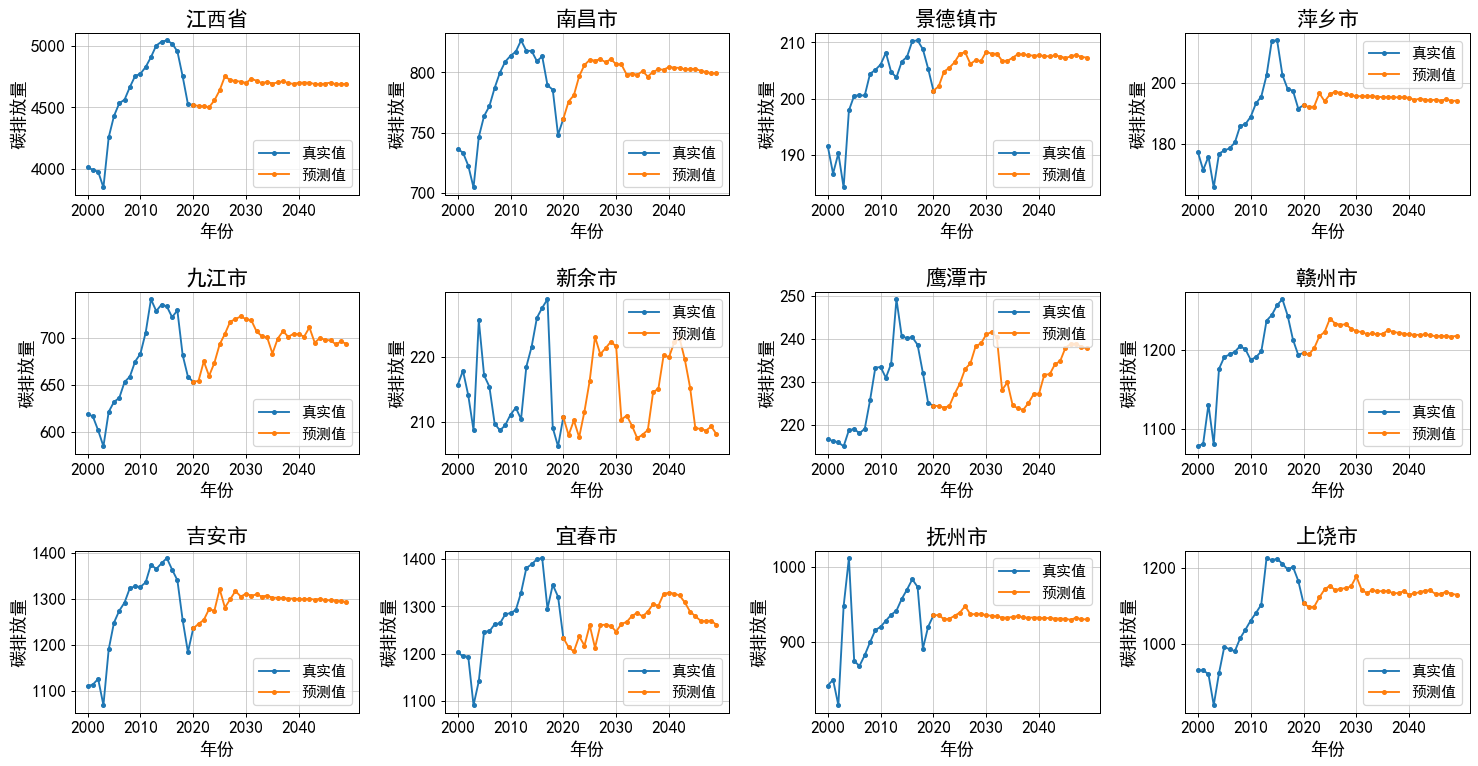


图1 江西省及11个市区碳排放量预测

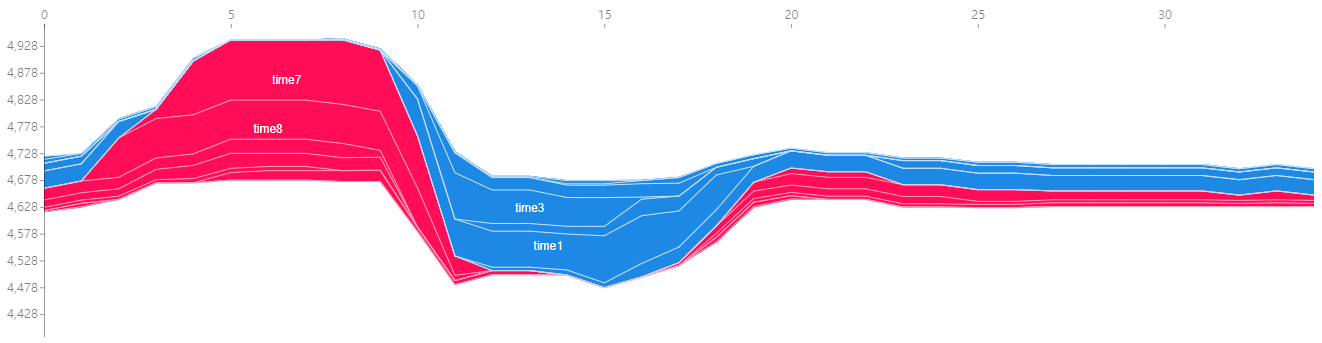


图2 江西省碳排放量预测模型在2009年后SHAP特征分布图

[1] 张劲帆, 刚健华, 钱宗鑫, 等. 基于混频向量自回归模型的宏观经济预测[J]. 金融研究, 2018, 457(7): 34-48.

[2] Lundberg, S.M., Nair, B., Vavilala, M.S. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2, 749–760 (2018).