基于机器学习方法的空气质量预测与影响因素识别

李佳成[1],梁龙跃[1,2]
(1. 贵州大学 经济学院,贵州 贵阳 550025;
2. 贵州大学 马克思主义经济学发展与应用研究中心,贵州 贵阳 550025)

摘　要:空气质量指数(AQI)的精准预测及影响因素识别,对空气污染防护和治理具有重要现实意义。选取北京市 2014 年第一季度至 2022 年第二季度 AQI 作为研究对象,探究六大污染物、五个气象因子和十四个经济变量对空气质量影响。选用 DT,RF,GBDT 和 XGBoost 模型对 AQI 进行预测,并使用稳定性选择方法定量分析各个变量对 AQI 的贡献。结果表明:四种模型方法均有良好的预测效果,其中 XGBoost 和 RF 的预测效果最优;六大污染物中 PM2.5,PM10 浓度和气象因素中的风速和气压对 AQI 影响较大;十四个经济变量对 AQI 的影响差异较大,其中城镇居民人均可支配收入、第三产业 GDP 和规模以上工业总产值等对 AQI 影响较大,而第一产业 GDP 和公路货物运输量等影响较小。

关键词:空气质量;影响因素;定量分析;机器学习;稳定性选择
中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1673-629X(2024)01-0164-07
doi:10.3969/j.issn.1673-629X.2024.01.024

Air Quality Prediction and Influencing Factor Identification Based on Machine Learning Methods

LI Jia-cheng[1],LIANG Long-yue[1,2]
(1. School of Economics,Guizhou University,Guiyang 550025,China;
2. Center for Development and Application of Marxist Economics,Guizhou University,Guiyang 550025,China)

Abstract:The accurate prediction of air quality index (AQI) and the identification of influencing factors are of great practical significance for air pollution prevention and control. The AQI of Beijing from the first quarter of 2014 to the second quarter of 2022 was selected as the research object to explore the influence of six major pollutants,five meteorological factors and fourteen economic variables on air quality. The DT,RF,GBDT and XGBoost models were selected to predict AQI,and the contribution of each variable to AQI was quantitatively analyzed using the stability selection method. The results show that the four model methods have excellent prediction effects,and XGBoost and RF have the best prediction effects; among the six major pollutants,PM2.5,PM10 concentration and meteorological factors,such as wind speed and pressure,have a greater influence on AQI;the influence of fourteen economic variables on AQI is quite different,among which the per capita disposable income of urban residents,tertiary industry GDP and gross industrial output value above designated size have a greater influence on AQI,while the primary industry GDP and road cargo transportation volume have a small influence.

Key words:air quality;influencing factors;quantitative analysis;machine learning;selection of stability

1　引言及文献综述

改革开放以来,随着中国工业、制造业和城市化进程的飞速发展,空气质量受到严重污染。空气中的污染物主要包括二氧化硫(SO_2)、一氧化碳(CO)、二氧化氮(NO_2)、臭氧(O_3)等气体污染物以及 PM2.5 和 PM10 等小型可吸入颗粒污染物,对身体健康和生态环境具有严重危害(吴春芳等,2021[1])。党和国家一直高度重视空气质量污染问题,“十四五”规划明确指

收稿日期:2023-02-10　　**修回日期**:2023-06-15
基金项目:国家自然科学基金项目(52000045);贵州省省级科技计划项目资助(黔科合基础-ZK[2022]一般 076);贵州省教育厅人文社会科学研究基地项目(23RWJD030);贵州大学经济学院创新基金资助项目(CJ2022107)
作者简介:李佳成(1996-),男,硕士研究生,研究方向为人工智能算法、数量经济学;通讯作者:梁龙跃(1986-),男,博士,硕导,研究方向为人工智能算法、数量经济学。

出要深入打好污染防治攻坚战,建立健全环境治理体系,不断改善空气质量。党的二十大报告指出要广泛形成绿色生产生活方式,碳排放达峰后稳中有降,生态环境根本好转,美丽中国目标基本实现。因此,对空气质量进行精准预测并识别其影响因素,对区域空气质量治理具有前瞻性和针对性的指导意义。

空气质量问题长期受学者的关注,研究范围主要涉及空气质量预测和影响因素分析两个方面。空气质量预测主要采取数值预测模型、统计预报模型和机器学习模型。数值预测模型预测精度高,现用于国内多个城市环境监测中心,但其专业性要求高,各模型适用范围单一(王自发等,2006[2];王哲等,2014[3])。统计预报模型在国内外均有应用,但受线性统计关系限制,难以模拟复杂多变的大气污染变化(沈劲等,2015[4];Bai 等,2016[5])。机器学习模型能够解决统计模型的非线性问题,具有较高预测精度和较强泛化能力等优点(杨思琦等,2017[6];陈建坤等,2022[7]),深度学习模型属于机器学习方法的一种,预测精度高,但存在模型结构复杂、解释性较差等不足(李栋等,2020[8];蒲国林等,2018[9])。空气质量影响因素现有大量研究,有学者在全国层面上运用回归模型研究不同城市的空气质量影响因素(陈永林等,2015[10];赵艳艳,2021[11]),有学者运用统计学和 OLS 方法研究城市群层面上的空气质量影响因素(李慧等,2021[12];金自恒等,2022[13]),还有学者在省级(直辖市)层面上运用空间计量和面板计量方法探究空气质量影响因素(刘利等,2021[14];丁镭等,2016[15])。上述学者大多基于计量方法分析空气质量影响因素,也有少量学者选用机器学习算法分析大气能见度和 PM2.5 浓度影响因素(张杨等,2021[16];夏晓圣等,2020[17]),但大多依据单个机器学习模型计算结果进行分析,极可能出现计算结果稳健性较差的情况。该文将多种机器学习算法与特征稳定性选择方法相结合,定量研究多个影响因素对空气质量的贡献。

文章的贡献主要有以下三个方面:(1)将空气质量预测和影响因素分析相结合,弥补现有研究中空气质量预测和影响因素单独研究的不足;(2)选用预测精度高的决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)和极端梯度提升(Extreme Gradient Boosting, XGBoost)模型预测北京市空气质量;(3)将 Meinshausen 和 Bühlmann(2010)[18]提出的稳定性选择方法和机器学习相结合,首次将该方法运用于空气质量影响因素定量研究中,为北京市政府和环境保护部门改善空气质量提供有效建议。

2　研究方法及模型构建

现有文献针对空气质量影响因素分析大多采用计量的回归方法度量,其计算结果可能和真实结果有较大差异;此外也有学者选用机器学习算法中的树类模型方法(Ng,2014[19]),因为树类模型不仅能考虑较多变量,并且能够直接算出变量的特征重要性,具有较高预测精度。但是如果只基于单个方法计算结果得出变量重要性,容易出现结果稳健性较差的情况。基于此,文章采用多种机器学习方法对空气质量进行预测,并将机器学习方法作为稳定性选择算法的基模型,定量识别多个因素对空气质量影响。

2.1　机器学习方法选择和介绍

考虑到深度学习方法虽然具有较高的预测精度,但存在模型原理复杂、解释能力弱等不足,为了实现空气质量预测和定量识别其影响因素,选择 DT, RF, GBDT 和 XGBoost 四种机器方法对 AQI 进行预测,并将其作为稳定性选择算法的基模型定性分析空气质量影响因素,接下来逐一介绍模型和算法。

DT 是一种非参数的有监督学习算法,一般具有运行速度快,无需对参数做假设等优点;但也存在容易过拟合,计算结果方差偏大等缺点。RF 由 Breiman(2010)[20]提出,是重抽样自举法的集成学习方法的一种,主要用于分类及回归分析,具有训练速度快和预测精度高的优点,但也存在噪声过大,造成回归过程中产生过拟合的可能。GBDT 是 Hastie 等(2009)[21]提出的一种基于梯度提升树的集成算法,具有并行运算、复杂度可控、泛化能力强的优点,但也存在训练阶段为串行结构速度较慢的不足。XGBoost 是 Chen 和 Guestrin(2016)[22]提出的集成方法,是对传统 GBDT 的提升和优化,属于 Boosting 集成学习算法,具有预测精度高、不易过拟合和扩展性强等优点,但也存在对噪声敏感和训练时间较长等不足。

2.2　模型最优参数

机器学习建模和预测分析过程中,模型参数对模型的预测效果及性能具有重要影响,采取时间序列交叉验证和网格搜索方法选择模型的最优参数,得出最佳预测效果。具体方法如图 1 所示。

文章选取时间序列递增窗口的方式。假设原始数据具有 100 个样本,先将数据集划分为训练集(前 80 个样本)及测试集(后 20 个样本),训练集用于交叉验证网格搜索调参,按比例分别被划分为训练子集和验证子集,验证子集用于评估模型性能,测试集作为测试集保留,不用于交叉验证过程,避免最终预测时数据提前“泄露”。在此过程中模型会将各个参数可能的取值进行排列组合,列出所有可能的组合结果,生成“网格”之后再选定使均方误差最小的参数组合作为该方

法的最优参数。

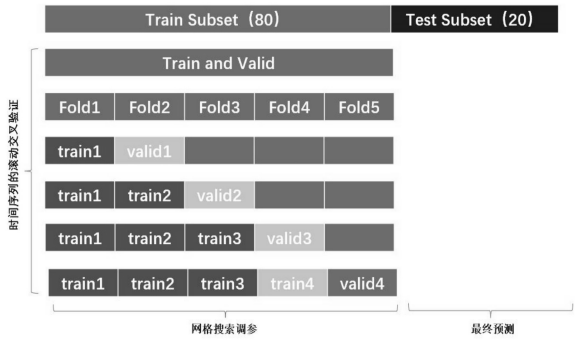

图 1 交叉验证和网格搜索方法

通过运用交叉验证和网格搜索方法，对 DT，RF，GBDT 和 XGBoost 模型的超参数分别进行运算，得到各个模型的最优参数。其中，DT 最优参数中最大深度=15，最大叶子节点数=5，叶子节点最少样本数=1；RF 最优参数中最大深度=5，弱学习器个数=200，GBDT 最优参数中最大深度=15，弱学习器个数=500，学习率=0.5，随机采样比例=0.7；XGBoost 最优参数中最大深度=10，弱学习器个数=150，学习率=0.05，随机采样比例=0.8。

2.3 评价指标选取

为了客观衡量四种非线性机器学习算法的预测性能，采用均方根误差（Root Mean Square Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）以及精度（Direction Accuracy，DA）作为模型模拟结果与实际值吻合程度的衡量指标，其中 RMSE 与 MAE 值越小表明模型预测效果越好，DA 值越大表明模型性能越好，DA 表示模型在预测趋势上的准确性。各评估指标数学表达式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{1}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{2}$$

$$\mathrm{DA}=\frac{1}{N}\sum_{i}^{N}\partial_i,\quad \partial_i=\begin{cases}1,(y_{i+1}-y_i)*(\hat{y}_{i+1}-y_i)\geqslant 0\\0,(y_{i+1}-y_i)*(\hat{y}_{i+1}-y_i)<0\end{cases} \tag{3}$$

其中，y_i 为时间序列真实值，$\hat{y}_i$ 为时间序列预测值，N 表示样本数量。

2.4 稳定性选择方法计算空气质量影响因素

稳定性选择方法是 Meinshausen 和 Bühlmann（2010）[18] 提出的，主要思想是利用重采样技术抽取不同的数据子集，然后对数据子集进行训练，得到多个变量的筛选结果，利用这些结果计算出各个特征的重要性大小。稳定性选择方法是采用重采样的思想，根据重采样得到的多个样本训练多个模型，因此最后集成的重要性排序结果要比单个模型的更优，理论上能够得到更加稳健的结果。

鉴于此，文章使用稳定性选择方法计算污染物、气象因素和经济因素对空气质量的重要性得分，下面介绍算法过程。

稳定性选择算法：

输入：

(X,y)：训练数据集，其中 $X\in R^{n*p}$，$y\in R^{n*1}$；

M：集成的大小；

Estimator：模型方法；

Λ：模型的 N（文中 $N=50$）个正则化参数组成的集合，即 $\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_N\}$；

π_{thr}：预先设定的阈值；

输出：选择变量的指标集 S^{stable}

稳定性选择算法的主要步骤为：

1. For $=1,2,\cdots,M$

a. 从 (X,y) 样本集中进行有放回随机抽样，以生成样本量为 $n/2$ 的子集 $(X^{(m)},y^{(m)})$

b. 对于 Λ 中的每个参数（$\lambda_N\in\Lambda$），在 $(X^{(m)},y^{(m)})$ 中运行相应模型的稳定性选择方法，并记变量选择结果为 $\hat{S}^{\lambda_*}_{[n/2],m}(n=1,2,\cdots,N)$；

For 循环结束

2. 计算影响因素被选为重要变量的概率：

$\widehat{\prod}j=\max\limits_{\lambda_*\in\Lambda}\{\widehat{\prod}J^{\lambda_*}\},j=1,2,\cdots,p$；其中，$\widehat{\prod}J^{\lambda_*}=\frac{1}{M}\sum\limits_{m=1}^{M}I\{j\in\hat{S}^{\lambda_*}_{[n/2],m},m\}$，$I\{\cdot\}$ 代表指示函数，当满足条件时，$I=1$，否则 $I=0$

3. 确定最后的选择集

$S^{\mathrm{stable}}=\{j:\max\limits_{\lambda_*\in\Lambda}\{\widehat{\prod}J^{\lambda_*}\}\geqslant\pi_{\mathrm{thr}}\}$，其中 π_{thr} 为预先设定的阈值，文中选取 0.7

稳定性选择的具体做法是通过计算单个因素在空气质量影响因素集合中被选出的次数，从而计算每个因素的重要性得分，由于稳定性选择模型能够自动识别重要性强弱的因素，因此重要性强的影响因素会有更大的概率被选中，所以其得分会接近重采样的次数 N，即 50 分；重要性相对较弱的影响因素得分会介于 0 至 50 分之间，而不相关的影响因素分数则接近 0 分，最后对每个因素的贡献率大小进行排序，即可确定空气质量的重要影响因素。

3 数据来源及预处理

3.1 数据来源

文章使用的六大污染物（SO_2，CO，NO_2，O_3，PM2.5 和 PM10）以及 AQI 来源于中国空气质量在线监测分析平台（https://www.aqistudy.cn），气象数据（平均温度、风速、湿度、气压和降雨量）来源于慧聚数据网（http://hz.hjhj-e.com），经济数据来源于北京市统计

年鉴和同花顺 iFinD 数据库。　　变量说明见表 1。

表 1　北京市空气质量影响因素指标体系

数据分类	选取指标(单位)
空气质量	AQI(无量纲),SO_2,CO,NO_2,O_3,PM2.5 和 PM10 除 CO 浓度单位为 mg/m^3,其他浓度均为 ug/m^3
气象因素	平均温度(℃)、风速(m/s)、湿度(%)、气压(hpa)、降雨量(nm)
经济因素	第一产业 GDP、规模以上工业总产值、建筑业企业:签订合同额、第二产 GDP、进出口、房地产开发投资、社会消费品零售总额、金融业 GDP、第三产业 GDP、公路货物运输量、公路旅客运输量、北京市 GDP、城镇居民人均可支配收入、一般公共预算支出 其中:城镇居民人均可支配收入单位为元,公路旅客运输量单位为万吨,公路旅客运输量为万人,剩余变量单位均为亿元

3.2　数据标准化处理

考虑不同指标的量纲不同,为提升模型拟合精度,文章对数据进行最大最小标准化处理,将数据置于统一量纲之中,其公式表示为:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

其中,X 表示某列原始数据,$X_{\max}$ 和 $X_{\min}$ 分别为该列数据的最大值和最小值,X^* 表示归一化后获取的新数据,经过该方法处理的数据将被映射到[0,1]之间。

4　实证结果分析

4.1　预测效果分析

文章采取 DT,RF,GBDT 和 XGBoost 回归算法对 AQI 数值进行预测,以 2014 年第一季度至 2020 年第四季度时间段作为训练集,2021 年第一季度至 2022 年第二季度数据为测试集,经过时间序列的交叉验证和网格调参选出最优参数,然后进行预测,根据预测结果得出两条结论:(1)四种机器学习算法都具有良好的预测性能,其预测结果的 RMSE 值分别为 0.171,0.155,0.169 和 0.153;MAE 值分别为 0.140,0.127,0.125和 0.119;DA 值分别为 0.6,1.0,0.8 和 1.0。(2)在四种机器学习算法中,XGBoost 和 RF 在预测效果上优于 DT 和 GBDT。

4.2　稳定性选择算法对 AQI 影响因素定量识别

文章以四种机器学习方法作为稳定性选择算法的基模型,运用 python 定量识别六大污染物、气象因素和经济因素(参见表 1)对空气质量的影响,具体影响效果为各个变量对 AQI 的贡献率大小,参考四种机器学习方法预测效果,其中 RF 和 XGBoost 两种模型的预测效果最优,因此将着重分析 RF 和 XGBoost 两种模型所计算的各个影响因素对 AQI 的平均贡献率(下文简称两种模型),同时为了保证结果更加稳健可靠,还将 DT,RF,GBDT,XGBoost 四种模型的计算结果取平均值作为对照。下面分别讨论六大污染物、气象因素和经济因素对 AQI 的贡献率。

4.2.1　六大污染物对 AQI 的贡献

首先,分析六大污染物对 AQI 的贡献,其中两种模型对 AQI 的平均贡献率大小如表 2 所示,四种模型作为对照,具体结果如表 2 所示。

表 2　污染物对空气质量的影响及其贡献率

RF 和 XGBoost 两种方法			四种机器学习方法		
排序	六大污染物指标名称	贡献率/%	排序	六大污染物指标名称	贡献率/%
1	PM2.5	14.800	1	PM2.5	14.672
2	PM10	12.547	2	PM10	14.176
3	O_3	6.909	3	O_3	7.345
4	NO_2	3.656	4	NO_2	3.414
5	CO	2.137	5	CO	2.891
6	SO_2	1.946	6	SO_2	1.194

如表 2 所示,六大污染物对 AQI 的贡献率排名分别是 PM2.5,PM10,O_3,NO_2,CO,SO_2,其中 PM2.5 和 PM10 的贡献之和达到 27.35%。PM2.5 和 PM10 属于小型颗粒污染物,均为 AQI 的主要影响因素;当 AQI 较大时,PM2.5 和 PM10 数值也呈现较高的状态,空气会出现污染状态,北京市可能伴随雾霾和沙尘暴,

对人们出行和身体健康具有重大危害。O_3,NO_2,SO_2 和 CO 作为气体污染物,对 AQI 的影响之和为 14.65%,也是现阶段国内面临的污染难题之一,需要针对各自污染源进行逐一解决。将表 2 的平均结果进行对比,所得结论也基本相同。

4.2.2 气象因素对 AQI 的贡献

气象条件作为一种自然因素,对空气质量也具有重要影响。有学者研究表明 AQI 与风速、气温、湿度以及降雨量呈负相关,与气压呈正相关(何振芳等,2021[23],周兆媛等,2014[24])。文章分析气象因素对 AQI 的贡献,选择两种模型对 AQI 的平均贡献率进行分析,四种机器学习方法结果作为对照,如表 3 所示。

表 3 气象因素对空气质量的影响及其贡献率

RF 和 XGBoost 两种方法			四种机器学习方法		
排序	气象因素指标名称	贡献率/%	排序	气象因素指标名称	贡献率/%
1	气压	3.942	1	气压	4.586
2	风速	1.076	2	风速	1.613
3	降雨量	0.607	3	降雨量	0.651
4	平均温度	0.260	4	平均温度	0.336
5	湿度	0	5	湿度	0.158

如表 3 所示,气象因素对 AQI 贡献率之和为 5.88%,其中气压和风速对 AQI 贡献率在五个气象因素中最高,分别为 3.94% 和 1.08%,其原因可能有以下两个方面:(1)选取的季度数据中,气压数据的均值为 1 012.14,其标准差为 8.12,气压值波动平缓并且其值高于标准气压值,导致气压对 AQI 贡献较大;(2)北京市地处平原,夏季和冬季风大,有利于 AQI 挥发,所以对 AQI 影响较大。剩余三个气象因素对 AQI 贡献较低的原因如下:①选取平均温度,其值低于白天平均温度,并且每日平均温度波动较小,导致模型较难识别其重要性,所以贡献率较低;②湿度对 AQI 的影响主要是通过影响降雨来降低 AQI,而北京市近几年季度降雨量均值较小,导致贡献率也比较低。对比四种机器学习方法,结果也类似。

4.2.3 经济因素对 AQI 的贡献

经济建设与环境保护协调发展是国家一直以来遵循的原则,定量识别经济因素对空气质量的影响,为空气质量改善提出有效建议具有重大的现实意义。文章分析经济因素对 AQI 的贡献,选择两种机器学习算法对 AQI 的平均贡献率进行分析,四种算法结果作为对照,结果如表 4 所示。从表 4 可以看出,空气质量影响因素的前七个经济变量分别为城镇居民人均可支配收入、第三产业 GDP、规模以上工业总产值、北京市 GDP、第二产业 GDP、社会消费品零售总额和公路旅客运输量,其贡献率分别是 12.80%、8.14%、6.28%、6.12%、5.16%、4.53% 和 3.47%。

对空气质量影响最重要的是城镇居民人均可支配收入,有学者研究发现随着北京市居民收入提高,空气质量改善(王会等,2018[25]);第三产业 GDP 对空气质量影响次之,其 GDP 值越高,越能改善空气质量状况;北京市规模以上工业总产值和北京市 GDP 也是影响空气质量的重要因素之一,但北京市工业在 2013 年已经进入后工业化时代,其工业主要为偏高端的污染相对较小的产业类型,因此其规模以上工业总产值越大,越能改善空气质量状况(王丽,2021[26]),也有学者研究指出北京市已进入经济与环境协调发展后期阶段,随着北京市 GDP 的不断增加,空气质量会不断改善(吴玉萍等,2002[27]);第二产业主要指加工制造业等企业,通常消耗大量能源并排放污染气体,是导致北京市空气质量恶化的重要影响因素,政府应该对第二产业的发展与排污进行重点审核,倒逼企业向绿色可持续发展转型;社会消费品零售总额主要包括社会生活消费品,其值越大,表明社会零售商品生产消耗越大,对北京市空气质量具有严重危害;公路旅客运输量和公路货物运输量可代表北京市公路运输状况,公路运输会消耗大量化石能源并排放尾气,是导致空气质量污染的重要原因;建筑业企业在施工过程中会导致废水、粉尘和废弃物等污染物,建筑业企业:签订合同额越大,对北京市空气质量的污染就越严重,政府应该重点管控该行业的污染排放;金融业是指经营金融商品的特殊行业,一般不会造成空气质量污染;此外房地产开发投资、进出口、一般公共预算支出和第一产业均会对北京市 AQI 造成一定的影响,由于定量分析中这几个经济变量对 AQI 影响较小,因此不在一一展开论述。对比四种机器学习方法的平均结果,各个经济因素对北京市 AQI 影响与 RF 和 XGBoost 两种方法的结

果相似,表明结果稳健可靠。

表 4　经济因素对空气质量的影响及其贡献率

RF 和 XGBoost 两种方法			四种机器学习方法		
排序	经济因素指标名称	贡献率/%	排序	经济因素指标名称	贡献率/%
1	城镇居民人均可支配收入	12.287	1	城镇居民人均可支配收入	10.365
2	第三产业 GDP	8.142	2	规模以上工业总产值	9.615
3	规模以上工业总产值	6.281	3	第三产业 GDP	6.336
4	北京市 GDP	6.129	4	北京市 GDP	4.710
5	第二产业 GDP	5.158	5	第二产业 GDP	4.335
6	社会消费品零售总额	4.527	6	社会消费品零售总额	4.258
7	公路旅客运输量	3.470	7	公路旅客运输量	2.862
8	建筑业企业:签订合同额	1.942	8	金融业 GDP	1.801
9	金融业 GDP	1.924	9	进出口	1.460
10	房地产开发投资	0.953	10	建筑业企业:签订合同额	1.336
11	进出口	0.578	11	房地产开发投资	0.967
12	一般公共预算支出	0.470	12	一般公共预算支出	0.630
13	第一产业 GDP	0.173	13	第一产业 GDP	0.150
14	公路货物运输量	0.087	14	公路货物运输量	0.138

5　结论及建议

5.1　结　论

(1)在空气质量预测中,选用 DT,RF,GBDT 和 XGBoost 模型,经过时间序列交叉验证和网格调参选取了模型的最优超参数对北京市 AQI 进行预测,四种机器学习算法均表现出良好的预测性能,RMSE 值均低于 0.172,MAE 值均低于 0.141,DA 值高于 0.6。

(2)六大污染物对空气质量的影响较大,RF 和 XGBoost 方法计算六大污染物对空气质量的平均贡献率为 42.00%,PM2.5 和 PM10 的贡献率分别为 27.35% 和 28.85%,气象因素中气压和风速对空气质量影响分别为 3.94% 和 1.08%,对 AQI 的影响较大。

(3)经济因素在两种方法计算中对空气质量影响的贡献率为 52.12%,其中城镇居民人均可支配收入、第三产业 GDP、规模以上工业总产值、北京市 GDP、第二产业 GDP 和社会消费品零售总额和公路旅客运输量是空气质量重要影响因素(四种机器学习方法结果类似),其贡献率之和为 46%,是北京市政府改善空气质量的重点关注对象。

5.2　政策建议

(1)高度重视高能耗、高污染行业的发展。实证结果表明建筑业企业和第二产业 GDP 对北京市空气质量具有重要影响,北京市政府应合理规划市内建筑业企业和第二产业的空间布局,因地制宜推进企业进入产业园区,集中建设污染排放处理设备,推进企业转向绿色可持续发展,倒逼企业结构升级转型,改善北京市空气质量(倪琳和郭小雨,2022[28])。

(2)充分发挥“有为政府”职能,推动北京市经济发展。实证结果表明,城镇居民人均可支配收入、第三产业 GDP、北京市 GDP 和金融业 GDP 对空气质量具有重要影响作用。北京市政府应立足北京市发展现状和发展优势条件,大力推进地区经济发展,针对第三产业、金融业和规模以上工业等提供一系列发展政策和方针,提高北京市 GDP 和人均可支配收入,达到改善空气质量的目的(Hao 等,2020[29];陈文和王晨宇,2021[30];周侗等,2022[31];王会等,2018[25])。

(3)优化交通运输。实证结果表明公路旅客运输量和公路货物运输量对于北京市 AQI 具有重要影响。北京市机动车大多行驶在城市人口密集区域,容易造成城市拥堵,进一步增加机动车能源消耗并增加汽车尾气排放,导致城市空气质量恶化(Lu 等,2017[32])。北京市政府应优化交通运输现状,大力推广新能源汽车,全市除公交车采用新能源汽车外,鼓励企事业单位购买新能源车辆;对个人和家庭提高新能源汽车购买补贴,减少汽车化石能源消费,改善空气质量。

参考文献:

[1] 吴春芳,林　勇,乐建培,等. 空气污染物 PM2.5 含量与某城区呼吸系统和心血管系统主要疾病指标阳性检出病例数的相关性研究[J]. 标记免疫分析床,2017,24(8):929-933.

[2] 王自发,谢付莹,王喜全,等. 嵌套网格空气质量预报模式系统的发展与应用[J]. 大气科学,2006(5):778-790.

[3] 王　哲,王自发,李　杰,等. 气象—化学双向耦合模式

(WRF-NAQPMS)研制及其在京津冀秋季重霾模拟中的应用[J]. 气候与环境研究,2014,19(2):153-163.
[4] 沈 劲,钟流举,何芳芳,等. 基于聚类与多元回归的空气质量预报模型开发[J]. 环境科学与技术,2015,38(2):63-66.
[5] BAI Y,LI Y,WANG X,et al. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions[J]. Atmospheric Pollution Research,2016,7(3):557-566.
[6] 杨思琪,赵丽华. 随机森林算法在城市空气质量预测中的应用[J]. 统计与决策,2017(20):83-86.
[7] 陈建坤,牟凤云,张用川,等. 基于多机器学习模型的逐小时 PM_(2.5)浓度预测对比[J]. 南京林业大学学报:自然科学版,2022,46(5):152-160.
[8] 李 栋,张 蕾,郭茂祖,等. 基于时空卷积残差网络的空气质量预测[J]. 计算机技术与发展,2020,30(6):124-129.
[9] 蒲国林,刘笃晋. 基于改进神经网络的环境空气质量预测[J]. 计算机技术与发展,2018,28(9):181-184.
[10] 陈永林,谢炳庚,杨 勇. 全国主要城市群空气质量空间分布及影响因素分析[J]. 干旱区资源与环境,2015,29(11):99-103.
[11] 赵艳艳,张晓平,陈明星,等. 中国城市空气质量的区域差异及归因分析[J]. 地理学报,2021,76(11):2814-2829.
[12] 李 慧,王淑兰,张文杰,等. 京津冀及周边地区“2+26”城市空气质量特征及其影响因素[J]. 环境科学研究,2021,34(1):172-184.
[13] 金自恒,高锡章,李宝林,等. 川渝地区空气质量时空分布格局及影响因素[J]. 生态学报,2022,42(11):4379-4388.
[14] 刘 利,邓宇宸,吴 丹,等. 广东省城市环境空气质量时空特征及经济影响因素分析[J]. 中国环境监测,2021,37(3):40-50.
[15] 丁 镭,刘 超,黄亚林,等. 湖北省城市环境空气质量时空演化格局及影响因素[J]. 经济地理,2016,36(3):170-178.
[16] 张 杨,张福浩,陈 才,等. 京津冀大气能见度特征分析及影响因素研究[J]. 测绘科学,2021,46(7):196-204.
[17] 夏晓圣,陈菁菁,王佳佳,等. 基于随机森林模型的中国 PM_(2.5)浓度影响因素分析[J]. 环境科学,2020,41(5):2057-2065.
[18] MEINSHAUSEN N,BÜHLMANN P. Stability selection[J]. Journal of the Royal Statistical Society:Series B (Statistical Methodology),2010,72(4):417-473.
[19] NG S. Boosting recessions[J]. Canadian Journal of Economics/Revue Canadienne D′économique,2014,47(1):1-34.
[20] BREIMAN L. Random forests[J]. Machine Learning,2001,45(1):5-32.
[21] HASTIE T,TIBSHIRANI R,FRIEDMAN J H,et al. The elements of statistical learning:data mining,inference,and prediction[M]. New York:Springer,2009.
[22] CHEN T,GUESTRIN C. Xgboost:a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco:ACM,2016:785-794.
[23] 何振芳,郭庆春,刘加珍,等. 河北省大气污染时空变化特征及其影响因素[J]. 自然资源学报,2021,36(2):411-419.
[24] 周兆媛,张时煌,高庆先,等. 京津冀地区气象要素对空气质量的影响及未来变化趋势分析[J]. 资源科学,2014,36(1):191-199.
[25] 王 会,宋璨江,赵 昭,等. 北京市居民改善空气质量的支付意愿及其影响因素分析[J]. 干旱区资源与环境,2018,32(8):16-22.
[26] 王 丽. 经济与自然结合视角的北京雾霾问题探讨[J]. 宏观经济研究,2021(5):142-154.
[27] 吴玉萍,董锁成,宋键峰. 北京市经济增长与环境污染水平计量模型研究[J]. 地理研究,2002(2):239-246.
[28] 倪 琳,郭小雨. 产业结构升级、城镇化发展与空气质量——来自中部地区的经验证据[J]. 生态经济,2022,38(5):183-189.
[29] HAO Y,ZHENG S,ZHAO M,et al. Reexamining the relationships among urbanization,industrial structure,and environmental pollution in China—New evidence using the dynamic threshold panel model[J]. Energy Reports,2020,6:28-39.
[30] 陈 文,王晨宇. 空气污染、金融发展与企业社会责任履行[J]. 中国人口·资源与环境,2021,31(7):91-106.
[31] 周 侗,张帅倩,闫金伟,等. 长江经济带三大城市群 PM_(2.5)时空分布特征及影响因素研究[J]. 长江流域资源与环境,2022,31(4):878-889.
[32] LU X,LIN C,LI Y,et al. Assessment of health burden caused by particulate matter in southern China using high-resolution satellite observation[J]. Environment International,2017,98:160-170.