doi:10. 16055/j. issn. 1672-058X. 2021. 0005. 009

基于梯度提升决策树的焦炭质量预测模型研究*

程泽凯[1], 闫小利[1**], 程旺生[2], 袁志祥[1,3]
(1. 安徽工业大学 计算机科学与技术学院, 安徽 马鞍山 243002; 2. 马鞍山钢铁股份有限公司 制造部, 安徽 马鞍山 243000;
3. 工业互联网智能应用与安全安徽省工程实验室, 安徽 马鞍山 243002)

摘　要:焦炭是高炉炼铁的重要原料,其质量是影响铁水质量和高炉顺行的重要因素,针对焦炭质量存在检验难、滞后性、预测误差大等问题,提出一种基于梯度提升决策树算法的焦炭预测模型;结合专家经验与相关性分析方法,深入研究配合煤质量对焦炭质量的影响;最后利用配合煤质量指标对焦炭质量指标灰分、硫分、耐磨强度、抗碎强度进行建模预测;根据某焦化厂历史生产数据对模型进行评估,实验结果表明:基于梯度提升决策树的焦炭质量预测模型相较于线性回归模型、随机森林模型,决策树模型误差小、准确率高,可以为焦化厂配煤炼焦提供一定的理论依据。

关键词:焦炭质量;预测模型;梯度提升决策树
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1672-058X(2021)05-0055-06

0　引　言

焦炭由配合煤在约 1 000 ℃的高温条件下经干馏而获得,在高炉炼铁中起着燃料、还原剂、增碳剂及骨架的作用[1],其质量的好坏直接影响高炉运行状态和焦化厂的经济效益。随着一系列的环保政策法规逐渐完善出台,高炉朝着现代化、大型化建设,对焦炭生产的环保标准和质量要求日益提高,同时国内优质炼焦煤较少、价格昂贵且地区分布不均匀[2]。如何提高焦炭质量和产量,降低炼焦成本成了炼焦行业目前急需解决的问题之一。同时焦炭质量检测难,存在很大的滞后性,焦炉炼焦具有非线性、时变缓慢、高延迟、工况复杂的特点[3],建立焦炭质量预测模型具有重大意义。

目前,学者基本采用加权平均、专家经验、线性回归和神经网络等方法对焦炭质量进行预测。曾令鹏等[4]利用韶钢 2016 年的焦炭质量数据建立了线性回归焦炭质量预测模型,对焦炭灰分、硫分、耐磨强度、抗碎强度进行预测,预测模型已应用于韶钢实际生产, 实现了焦炭质量预测自动化;刘春梅[5]利用 BP 神经网络,通过炼焦煤质量数据对焦炭质量预测,平均准确率达到了 95%;陶文华等[6]利用主元分析法确定焦炭质量预测模型输入变量,利用差分算法对神经网络初始权值和阈值优化,建立了基于 DE-BP 优化的焦炭质量预测模型,该模型收敛精度快,预测精度高,可以为焦炭生产提供参考价值;袁正波等[7]利用遗传算法对支持向量机进行参数寻优,建立了基于 GA-SVM 的焦炭质量预测模型,与 BP 神经网络相比,其误差更小、模型的泛化能力更好。由于加权平均法误差较大,专家经验预测的准确性主要取决于专家的生产实践经验和丰富的专

收稿日期:2020-08-11;**修回日期**:2020-12-31.
* **基金项目**:国家重点研发计划项目(2016YFF020440508).
作者简介:程泽凯(1975—),男,安徽巢湖人,副教授,硕士,从事人工智能、数据挖掘和机器学习研究.
** **通讯作者**:闫小利(1995—),女,重庆人,硕士研究生,从事计算机应用、机器学习研究. Email: lilyyan0707@163. com.

业知识,但主观性太强、普适性差、不能进行定量分析,有时难以保证预测的准确性[8]。线性回归方法简单且容易实现,但对于复杂的非线性数据处理能力差,因此预测误差较大。传统的神经网络训练速度慢,容易陷入局部极小点而无法达到全局最优解,预测精度低[9]。支持向量机由 Vapnik[10] 等在20世纪70年代提出,具有良好的学习能力、泛化能力,可以解决高维问题,在小样本情况下预测误差较小、准确率高,可以避免选择神经网络结构和易陷入局部极小点等问题[11],但 SVM 对参数调节和函数选择敏感。

梯度提升决策树属于机器学习算法,最早由 Firedman 教授[12] 提出,是 Boosting 算法的一种,它将多个性能较差的弱学习器通过某种方式集成起来得到一个强学习器模型,由分类回归树(CART)、梯度提升(Gradiant Boosting)、缩减(Shrinkage)组成。梯度提升决策树算法可以灵活处理各种类型的数据,在相对少的调参时间情况下,预测的准确率也比较高,算法使用一些健壮的损失函数,对异常值的鲁棒性非常强,因此,近年来梯度提升决策树广泛应用于预测研究领域。路志英等[13] 利用 Focal Loss 改进梯度提升决策树算法,并对天津强对流灾害进行预测,实验结果表明:基于 Focal Loss 改进的梯度提升决策树模型效果优于逻辑回归、梯度提升决策树、随机森林与多层感知机模型;徐永瑞等[14] 利用梯度提升决策树对电力系统负荷进行预测,实验结果表明该模型的精度、泛化能力以及运算速度均优于 LSTM。

综上所述,本文提出了基于梯度提升决策树的焦炭质量预测模型,并对某焦化厂历史生产数据进行训练。实验结果表明:该模型相比于线性回归、随机森林、决策树模型预测精度高、误差小,对焦化厂生产具有一定的指导意义。

1 焦炭质量预测模型

1.1 梯度提升决策树算法原理

假设梯度提升决策树的训练集为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$,$x_i\in X\subseteq R^n$,$y_i\in Y\subseteq R^n$,梯度提升决策树采用迭代的思想,每轮迭代产生一个 CART 回归树 $T(x,\Theta_m)$,Θ 表示第 m 棵 CART 回归树参数,$m=1,2,\cdots,M$,每棵 CART 回归树在上一个 CART 回归树的残差基础上往残差 $g_{m,i}$ 减小的方向梯度迭代,更新 CART 回归树 $f_m(x)$,使损失函数 $L(Y,f(x))$ 最小,预测的结果 $f_M(x)$ 为初始值加上各 CART 回归树的残差。具体计算如式(1)(2)(3)所示:

$$g_{m,i}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)}=f_{m-1}(x) \tag{1}$$

$$f_m(x)=f_{m-1}(x)+T(x,\Theta_m) \tag{2}$$

$$f_M(x)=\sum_{m=1}^{M}T(x,\Theta_m) \tag{3}$$

1.2 数据预处理

本文所有数据来自某焦化厂历史生产数据,由于采集的数据规模大,量纲不同(如焦炭质量硫分为0.7%,耐磨强度为5.5%,抗碎强度为89.4%),存在缺失值和异常值等情况,所以需要对数据进行预处理操作,这样建立的模型才能准确、真实地反应生产情况。

整个数据预处理流程:删除缺失数据,如果一组测量数据中某个测量值残余误差的绝对值 $x_i>3\sigma$,则该测量值为异常值,应剔除,其中 σ 代表标准差。利用式(4)对输入变量、输出变量归一化。式(4)如下所示:

$$\widetilde{V}=\frac{V-V_{\min}}{V_{\max}-V_{\min}} \tag{4}$$

$\widetilde{V}$ 表示归一化后的值,V 为归一化前的样本值,$V_{\max}$ 和 $V_{\min}$ 为样本中的最大值和最小值。归一化后的数据分布在[0,1]区间,消除了量纲对模型的影响,提升了模型的收敛速度和精度。为了反映真实的生产数据,还需根据式(5)把预测结果反归一化。经过数据预处理后,获得580组数据。

$$V=\widetilde{V}(V_{\max}-V_{\min})+V_{\min} \tag{5}$$

1.3 焦炭质量预测模型输入输出变量的确定

焦炭影响高炉运行状态的主要质量指标[15] 有焦炭灰分、焦炭硫分、焦炭强度等。焦炭的灰分含量增高会使高炉冶炼中的炉渣量增高,导致料柱透气性和透液性变差,焦比增高;含硫量高的焦炭使高炉

利用率和钢铁质量下降;焦炭的冷态强度要求高且均匀稳定,则高炉冶炼强度、焦炭负荷、喷煤比得到提高,降低了生铁成本。因此选取焦炭灰分、硫分、耐磨强度、抗碎强度作为模型的输出。

影响焦炭质量的主要因素[16]有单种煤性质(水分、硫分、黏结指数、胶质层最大厚度等)、煤场管理、配煤工艺(煤预处理工艺、装炉煤的配合与粉碎工艺等)、炼焦工艺(焦炉的加热控制、焦炉压力制度)等。配煤炼焦包括一系列复杂的化学、物理反应,但最主要的影响因素是配合煤的质量。配合煤质量指标众多且复杂,根据理论,将所有可以影响焦炭质量的指标都作为模型输入,可能会得到准确率较高的预测模型,但会增加模型的复杂度和训练时间。本文依据专家经验及变量皮尔逊相关性分析结果选取模型输入变量,降低了模型输入维度,提高了模型的收敛速度。皮尔逊相关系数广泛用于衡量两个连续变量之间的相关程度,式(6)中,$i=1,2,\cdots,N$,$E(X)$,$E(Y)$分别代表 X 与 Y 的均值,$X=\{x_1,x_2,\cdots,x_N\}$,$Y=\{y_1,y_2,\cdots,y_N\}$,有

$$r_{X,Y}=\frac{\sum_{i=1}^{N}(x_i-E(X))(y_i-E(Y))}{\sqrt{\sum_{i=1}^{N}(x_i-E(X))^2}\sqrt{\sum_{i=1}^{N}(y_i-E(Y))^2}} \tag{6}$$

分析结果如表 1 所示。

表 1　焦炭质量数据相关性分析结果

Table 1　Correlation analysis results of coke quality data

焦　炭		配合煤						
		水分	灰分	硫分	挥发分	黏结指数	胶质层最大厚度	煤的最终收缩度
灰分	皮尔逊相关系数	0.032	0.375**	−0.127**	−0.052	0.032	−0.092*	0.009
	sig(双侧)	0.459	0.000	0.003	0.229	0.460	0.032	0.838
硫分	皮尔逊相关系数	−0.092*	−0.143**	0.190**	0.401**	0.026	0.183**	0.105*
	sig(双侧)	0.033	0.001	0.000	0.000	0.552	0.000	0.015
耐磨强度	皮尔逊相关系数	−0.186**	−0.169**	0.176**	0.062	−0.129**	0.316**	−0.094*
	sig(双侧)	0.000	0.000	0.000	0.153	0.003	0.000	0.030
抗碎强度	皮尔逊相关系数	0.159**	0.280**	−0.322**	−0.161**	0.082	−0.493**	0.138**
	sig(双侧)	0.000	0.000	0.000	0.000	0.059	0.000	0.001

注:**表示 sig<0.01,为相关性高度显著;* 表示 0.01<sig<0.05,为相关性显著。

其中,皮尔逊相关系数的取值范围为(−1,1),$r>1$ 表示正相关,$r<0$ 表示负相关,$r=0$ 表示零相关,r 的绝对值越大表示相关程度越高。因此选择配合煤水分、灰分、硫分、挥发分、黏结指数、胶质层最大厚度、煤的最终收缩度作为模型的输入。

2　梯度提升决策树模型训练及实验结果

本文利用 Python 3 的 Pandas 数据处理包和 Scikit-learn 机器学习包进行数据分析建模。

网格搜索法是一种穷举搜索方法,通过循环遍历,在候选参数集中选取不同的参数进行训练,选取误差最小的参数作为模型的最终参数。本文利用网格搜索法来确定模型相关参数,当梯度提升决策树的学习率为 0.01,损失函数为平方损失函数,弱学习器的数目为 100, CART 最大深度为 5 时,误差最小。

将数据集中前 530 组数据划分为训练集,剩下的 50 组为测试集,对线性回归、决策树、随机森林、梯度提升决策树这 4 个模型进行训练,为了比较各种模型的性能优劣,采用平均绝对误差 MAE、其值均方根误差 RMSE 衡量模型的预测精度和泛化能力,如式(7)、(8)所示:

$$R_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}|y_i-f_i| \tag{7}$$

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-f_i)^2} \tag{8}$$

其中,y 代表样本观测值,f 代表预测值。实验结果如表 2 和图 1—图 4 所示。

表2 模型性能比较结果

Table 2 Model performance comparison results

模 型	评价指标	灰分	硫分	耐磨强度	抗碎强度
梯度提升决策树	R_{MAE}	0.129 3	0.014 2	0.109 4	0.129 9
	R_{RMSE}	0.156 8	0.017 7	0.141 1	0.168 6
线性回归	R_{MAE}	0.134 2	0.015 0	0.121 2	0.627 9
	R_{RMSE}	0.155 8	0.018 7	0.143 7	0.735 5
随机森林	R_{MAE}	0.129 5	0.015 5	0.112 2	0.277 6
	R_{RMSE}	0.159 9	0.020 3	0.151 3	0.330 3
决策树	R_{MAE}	0.141 8	0.020 8	0.140 6	0.284 2
	R_{RMSE}	0.180 2	0.029 5	0.182 0	0.313 4

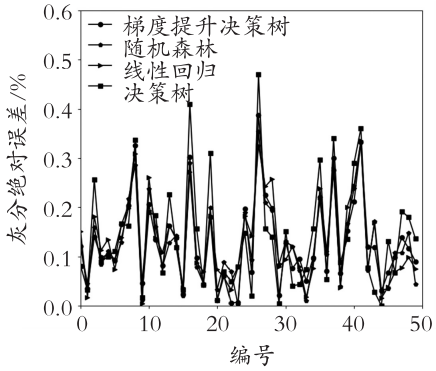

图1 灰分预测绝对误差对比

Fig. 1 Comparison of absolute error of ash prediction

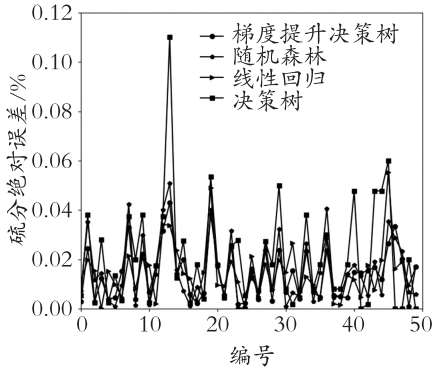

图2 硫分预测绝对误差对比

Fig. 2 Comparison of error of sulfur prediction

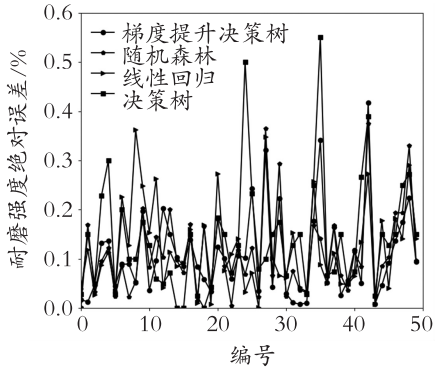

图3 耐磨强度预测绝对误差对比

Fig. 3 Comparison of absolute error of M_{10} prediction

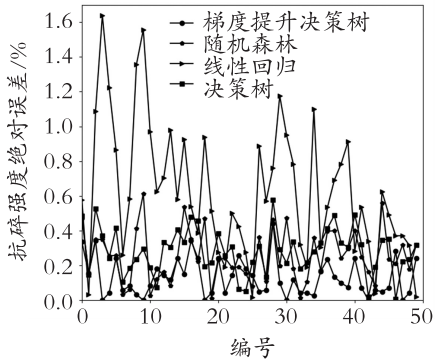

图4 抗碎强度预测绝对误差对比

Fig. 4 Comparison of absolute error of M_{40} prediction

由表2和图1—图4可知:梯度提升决策树模型预测焦炭各质量指标的R_{MAE}、R_{RMSE}均为最小,相比于其他3种模型更适合焦炭质量预测;线性回归模型对焦炭质量指标灰分、硫分、耐磨强度拟合较好,抗碎强度预测误差较大,但梯度提升决策树的预测精度更高;非线性3种算法拟合误差由小到大顺序为梯度提升决策树、随机森林、决策树,其中梯度提升决策树、随机森林都属于集成学习算法,相比于单一决策树预测精度高、误差小且不容易过拟合;由于存在焦炉炉况波动,数据波动较大,随机森林对波动值不怎么敏感,所以预测误差比梯度提升决策树模型大。

3 结论与讨论

梯度提升决策树、线性回归、随机森林、决策树

均属于机器学习算法,在训练样本数量有限的情况下,基于梯度提升决策树的焦炭质量预测模型相比于线性回归模型、随机森林模型、决策树模型误差更小,准确率更高。梯度提升决策树建立的焦炭质量预测模型拥有较好的泛化能力和鲁棒性,能够较为准确地预测焦炭质量,可以为配煤炼焦提供一定的理论依据。目前,只考虑了配煤工艺对焦炭质量的影响,后续工作将研究配煤工艺与炼焦工艺对焦炭质量的影响,以提高预测精度与准确率。

参考文献(References):

[1] 薛改凤,项茹,陈鹏,等. 炼焦煤质量指标评价体系的研究[J]. 武汉科技大学学报,2009,32(1):36—40
XUE G F, XIANG R, CHEN P, et al. Research on the Evaluation System of Coking Coal Quality Index [J]. Journal of Wuhan University of Science and Technology, 2009,32(1):36—40 (in Chinese)

[2] 邓小利,徐飞,王遂正. 中国稀缺炼焦煤资源分布特征[J]. 中国煤炭地质,2018,30(6):26—29
DENG X L, XU F, WANG S Z. Distribution Characteristics of Scarce Coking Coal Resources in China[J]. Coal Geology of China, 2018, 30(6):26—29 (in Chinese)

[3] 孙立迹. 焦炉全自动测温与加热优化控制技术应用[D]. 天津: 天津大学, 2013
SUN L J. Application of Automatic Temperature Measurement and Heating Optimization Control Technology in Coke Oven[D]. Tianjin:Tianjin University, 2013 (in Chinese)

[4] 曾令鹏,刘克辉,范国光. 韶钢焦炭质量预测模型的研究[J]. 燃料与化工,2018,49(3):13—15
ZENG L P, LIU K H, FAN G G. Study on the Coke Quality Prediction Model of Shaoguan Steel[J]. Fuel & Chemical Industry, 2018, 49(3): 13—15 (in Chinese)

[5] 刘春梅. 基于 BP 神经网络的炼焦煤质量预测研究[J]. 煤炭技术,2012,31(4):247—249
LIU C M. Prediction of Coking Coal Quality Based on BP Neural Network [J]. Coal Technology, 2012, 31 (4): 247—249 (in Chinese)

[6] 陶文华,袁正波. 焦炭质量的 DE-BP 神经网络预测模型研究[J]. 系统仿真学报,2018,30(5):1650-1656
TAO W H, YUAN Z B. Research on DE - BP Neural Network Prediction Model of Coke Quality[J]. Journal of System Simulation, 2018, 30(5): 1650 - 1656 (in Chinese)

[7] 袁正波,陶文华,王志峰. 高炉焦炭质量的 GA-SVM 模型预测[J]. 测控技术,2017,36(11):57—60,65
YUAN Z B, TAO W H, WANG Z F. GA - SVM Model Prediction of Coke Quality in Blast Furnace [J]. Measurement and Control Technology, 2017, 36 (11): 57—60,65 (in Chinese)

[8] 田陆峰,李志凯,张丽. 影响焦炭性质的因素分析及质量预测模型研究现状[J]. 煤质技术,2014(1):41—43
TIAN L F, LI Z K, ZHANG L. Analysis of Factors Affecting Coke Properties and Research Status of Quality Prediction Models[J]. Coal Quality Technology,2014(1): 41—43 (in Chinese)

[9] 覃光华,李祚泳. BP 网络过拟合问题研究及应用[J]. 武汉大学学报(工学版),2006(6):55—58
QIN G H, LI Z Y. Research and Application of BP Network Overfitting Problem [J]. Journal of Wuhan University (Engineering Science Edition), 2006 (6): 55—58 (in Chinese)

[10] VAPNIK V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000
VAPNIK V N. The Essence of Statistical Learning Theory[M]. ZHANG X G, Translated. Beijing: Tsinghua University Press, 2000 (in Chinese)

[11] 苟雪银,郭立新,张连波. 支持向量机和神经网络在粗糙面参数反演中的比较[J]. 计算物理,2014,31(1): 75—84
GOU X Y, GUO L X, ZHANG L B. Comparison of Support Vector Machine and Neural Network in Rough Surface Parameter Inversion[J]. Computational Physics, 2014,31(1):75—84 (in Chinese)

[12] FRIEDMAN J H . Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5): 1189—1232

[13] 路志英,汪永清,孙晓磊,等. 基于 Focal Loss 改进的 GBDT 模型对天津强对流灾害的预报[J]. 灾害学, 2020,35(3):34—37,50
LU Z Y, WANG Y Q, SUN X L, et al. Forecast of Tianjin Severe Convective Disaster Based on Improved GBDT Model of Focal Loss[J]. Journal of Catastrophe Science, 2020, 35(3):34—37,50 (in Chinese)

[14] 徐永瑞,左丰恺,朱新山,等. 改进 GBDT 算法的负荷预测研究[J/OL]. 电力系统及其自动化学报:1—8[2020

-09-21]. https://doi.org/10.19635/j.cnki.csu-epsa.000618

XU Y R, ZUO F K, ZHU X S, et al. Research on Load Forecasting with Improved GBDT Algorithm [J/OL]. Journal of Electric Power System and Automation: 1—8 [2020-09-21]. https://doi.org/10.19635/j.cnki.csu-epsa.000618 (in Chinese)

[15] 纪增辉.论焦炭质量对高炉冶炼的影响[J].中国金属通报,2019(8):207—208

JI Z H. Discussion on the Influence of Coke Quality on Blast Furnace Smelting [J]. China Metal Bulletin, 2019(8): 207—208 (in Chinese)

[16] 阳春华,沈德耀,吴敏,等.焦炉配煤专家系统的定性定量综合设计方法[J].自动化学报,2000(2):226—232

YANG C H, SHEN D Y, WU M, et al. Qualitative and Quantitative Comprehensive Design Method of Coke Oven Coal Blending Expert System [J]. Acta Automatica Sinica, 2000(2):226—232 (in Chinese)

Research on Coke Quality Prediction Model Based on Gradient Boosting Decision Tree

CHENG Ze-kai[1], YAN Xiao-li[1], CHENG Wang-sheng[2], YUAN Zhi-xiang[1,3]

(1. School of Computer Science and Technology, Anhui University of Technology, Anhui Maanshan 243002, China; 2. Manufacturing Department, Maanshan Iron and Steel Co., Ltd, Anhui Maanshan 243000, China; 3. Anhui Key Laboratory of Industrial Internet Intelligence Application and Security, Anhui Maanshan 243002, China)

Abstract: Coke is an important raw material for blast furnace ironmaking, and its quality is an important factor affecting the quality of molten iron and the smooth operation of blast furnace. In order to solve the problems of difficult inspection, hysteresis, and large prediction errors in coke quality, a coke prediction model based on gradient boosting decision tree algorithm is proposed. Combined with expert experience and correlation analysis, the influence of mixed coal quality on coke quality is studied. Finally, the mixed coal quality parameters are used to predict the ash content, sulfur content, M10 and M40 of coke quality parameters. The model is evaluated based on the historical production data of a coking plant. The experimental results show that the coke quality prediction model based on the gradient boosting decision tree has less error and higher accuracy than the linear regression model, random forest model, and decision tree model. It can provide a certain theoretical basis for coal blending and coking of the coking plant.

Key words: coke quality; prediction model; gradient boosting decision tree

责任编辑:李翠薇

引用本文/Cite this paper:

程泽凯,闫小利,程旺生,等.基于梯度提升决策树的焦炭质量预测模型研究[J].重庆工商大学学报(自然科学版),2021,38(5):55—60

CHENG Z K, YAN X L, CHENG W S, et al. Research on Coke Quality Prediction Model Based on Gradient Boosting Decision Tree[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2021, 38(5):55—60