·应用与服役·

基于SMOGN-XGBoost的钢包下渣剩余钢水量预测

樊士茜[1]，段豪剑[1]，谢忠研[1]，任　英[1]，张立峰[2]，尹　青[3]，吴小林[3]，赵德利[4]，李亚辉[4]
（1 北京科技大学冶金与生态工程学院，北京 100083；2 北方工业大学机械与材料工程学院，北京 100144；
3 江阴兴澄特种钢铁有限公司，江阴 214429；4 中国第一重型机械股份公司，齐齐哈尔 161042）

摘　要：钢包结构直接影响炼钢工艺的效率、质量和经济性。为进一步优化钢包结构设计，基于钢包下渣水模拟数据，深入探讨了不同机器学习算法在预测开始下渣时剩余钢水量的效能，并针对钢包底部结构变量对下渣剩余钢水量的影响进行了预测分析。首先，采用SMOGN技术对钢包下渣水模拟数据进行过采样预处理，以平衡数据分布，构建包含训练集和测试集的剩余水量特征集。在此基础上，分别测试了LASSO，SVR，ElasticNet，MLP以及XGBoost五种机器学习模型对剩余水量的预测能力。通过决策系数、均方误差和平均绝对误差三个指标进行评估，结果表明，XGBoost模型的预测效果最优，是剩余钢水量预测模型的首选。最后，采用XGBoost模型分析了钢包模型底部结构变量，包括水口直径、水口凸起高度、钢包底部台阶高度和钢包底部台阶与水口距离等对钢包下渣剩余水量的影响。结果表明，当水口直径超过ϕ40 mm时，剩余水量显著降低。降低水口凸起高度，以及增加钢包底部台阶高度，会显著降低钢包内剩余水量：当水口凸起高度超过26 mm时，剩余水量则将超过20 L；而当台阶高度超过11 mm且水口凸起高度低于11 mm时，剩余水量将减少到10 L以下。当台阶与水口距离增大时，剩余水量先减少，在距离大于100 mm后趋于稳定。研究结果为钢铁企业优化钢包结构、降低钢液浪费方面提供了重要参考，具有实际指导意义。

关键词：钢包下渣；机器学习；钢包底部结构；XGBoost；回归预测
DOI：10. 20057/j. 1003-8620. 2025-00077　　**中图分类号：**TF769. 2；TP181

Prediction of Remaining Molten Steel Volume during the Process of Ladle Slag Carry-over Based on SMOGN-XGBoost

Fan Shixi[1], Duan Haojian[1], Xie Zhongyan[1], Ren Ying[1], Zhang Lifeng[2], Yin Qing[3], Wu Xiaolin[3], Zhao Deli[4], Li Yahui[4]
(1 School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, China; 2 School of Mechanical and Materials Engineering, North China University of Technology, Beijing 100144, China; 3 Jiangyin Xingcheng Special Steel Co., Ltd., Jiangyin 214429, China; 4 China First Heavy Machinery Corporation, Qiqihaer 161042, China)

Abstract: The structure of the ladle directly affects the efficiency, quality, and economic aspects of the steelmaking process. To further optimize the ladle structure design, the current study deeply explored the effectiveness of different machine learning algorithms in predicting the remaining molten steel volume at the onset of the process of ladle slag carry-over based on water modeling data. Predictive analysis was conducted to explorer the impact of bottom structural variables of the ladle on the remaining molten steel volume during the process of ladle slag carry-over. Firstly, the SMOGN technique was employed to preprocess the water modeling data through oversampling, aiming to balance the data distribution and construct a feature set for residual steel volume, which included both training and testing datasets. Based on this foundation, the predictive capabilities of five machine learning models, namely LASSO, SVR, ElasticNet, MLP, and XGBoost, for the residual steel volume were tested. The evaluation was carried out through three metrics: the coefficient of determination, mean squared error, and mean absolute error. The results revealed that the XGBoost model outperformed the others in predictive accuracy, establishing it as the preferred model for forecasting the residual steel volume. Finally, the XGBoost model was utilized to analyze the impact of bottom structural variables of the ladle model on the residual steel volume. These variables included the diameter of nozzle, the height of nozzle, the height of steps, and the distance between nozzle and steps. The results indicated that when the nozzle diameter exceeded ϕ40 mm, there was a significant reduction in the residual steel volume. Reducing the height of nozzle and in-

基金项目：国家重点研发计划(No.2023YFB3709900)；国家自然科学基金(No.52474341、No.U22A20171)
作者简介：樊士茜(2001—)，女，硕士；　**E-mail：**fanshiximetal@sina.com；　**收稿日期：**2025-03-25
通信作者：段豪剑(1990—)，男，博士，副教授；　**E-mail：**duanhaojian@ustb.edu.cn

creasing the height of steps significantly decreased the residual steel volume in the ladle: when the nozzle height exceeded 26 mm, the residual volume surpassed 20 liters; whereas, when the step height exceeded 11 mm and the nozzle height was below 11 mm, the residual volume was reduced to less than 10 liters. As the distance between nozzle and steps increased, the residual steel volume initially decreased and then stabilized after the distance exceeded 100 mm. The findings of this research provided significant reference for steel enterprises to optimize ladle structure and reduce molten steel wastage, offering practical guidance for the industry.
Key Words: Ladle Slagging; Machine Learning; Ladle Bottom Structure; XGBoost; Regression Predictive Model

钢材生产需要尽量减少钢中夹杂物,特别是大颗粒外来夹杂物[1-2]。在生产过程中,由于追求钢液的收得率,会导致钢包下渣,进而影响钢液的洁净度。采用钢包下渣检测技术虽能够显著降低下渣量,但同时也降低了钢液收得率,造成资源浪费。因此,研究钢包下渣过程并采取措施控制下渣,对提高钢液洁净度和收得率至关重要。物理模拟[3-4]和数值模拟是研究钢包下渣的常用方法。物理模拟是通过物理模型和采用必要的测试手段,观察和分析特定过程[5-7]。物理模型的建立是以相似原理为基础,保证其规律与实际工况基本一致。Yin 等[8]通过建立水模型研究了动态变化模式下运行参数对涡流形成和炉渣携带的影响。Merder 等[9]探讨了不同气体流量和喷嘴布置对钢包内流场和传质过程的优化效果,为优化 Fe-Si 合金精炼工艺提供了理论依据。数值模拟则是利用计算机技术对物理过程进行模拟分析的研究方法[10]。具体而言,通过运用计算流体力学、传热学和冶金反应工程学等核心原理,对冶金过程中的流体流动[11-12]、热量传递[13]和化学反应等过程[14-15]进行预测,以此分析和优化冶金过程。Duan 等[16]通过数值模拟分析了钢包内钢液的温度分布、流动特性以及保温过程中热量损失的影响,证明了钢包壁和渣层的热传导对钢液温度均匀性有显著影响,同时,自然对流在维持钢液温度分布中起关键作用。Huang 等[17]建立了包含流体动力学、夹杂物动力学以及气泡-夹杂物相互作用的综合模型,采用计算流体动力学(CFD)技术模拟了钢液的流动和夹杂物的运动轨迹,为理解夹杂物的去除机制提供了重要的理论依据,同时,对钢包精炼工艺的优化提供了有价值的指导。Ji 等[18]利用大涡模拟(LES)、体积法(VOF)和离散相模型(DPM)耦合模型,优化了涌流辅助下渣工艺的操作参数,揭示了不同操作条件下钢包内流场和夹杂物去除效率的变化规律,减少钢液损失。

机器学习是一种基于数据和统计学的算法,通过计算机自主学习数据特征和规律,进而实现针对特定问题的预测、分类、聚类、降维等功能[19]。与传统的统计学方法相比,机器学习更加高效、准确,可以处理大量的、复杂的、高维度的数据,自动寻找数据中的关联和规律,从而发掘出有价值的信息[20]。近年来,机器学习技术在钢铁生产过程中得到广泛应用,涵盖炼铁、炼钢、铸造和轧制等多个环节[21-24]。Zhang 等[25]建立了合金成分和性能的“白盒”模型,结合了物理化学因素筛选与 Shapley 分析,用痕量 Cr 代替 Co,并保持了合金的力学性能与电导率。针对经验模型受限于钢种的问题,Wentzien 等[26]基于两个数据集,共计 1 800 个钢种,开发了用于预测马氏体转变温度的机器学习模型,显著提高了预测精度。Zhou 等[27]建立了基于 BP 神经网络的转炉终点磷含量预测模型,为实际生产提供了有益参考。

结合传统的物理模拟方法和新兴的机器学习技术,深入探讨了钢包底部结构对漩涡下渣现象的影响。首先,通过水模型实验模拟钢包下渣过程,获取不同工况下的剩余水量数据;随后,对实验数据进行系统化处理,并以此为基础评估了五种机器学习模型对剩余水量的预测效能;通过对比,选取表现最优的极限梯度提升(Extreme gradient boosting,XGBoost)模型,预测分析了水口直径、水口凸起高度、钢包底部台阶高度及台阶与水口距离等变量对剩余水量的影响;最后,使用机器学习模型预测双变量对剩余钢水量的影响。采用水模拟和机器学习结合应用的方法,有效地缓解了物理模型成本高,且受限于固定工况的缺点,发挥机器学习的高效建模优势;水模拟实验的实际数据,可以支持模型不断迭代,形成闭环优化;同时,以分析单变量影响的方式,对机器学习模型的预测性能和水模型的实验结果进行对比验证,保留水模拟的物理可信度,增加了模型和实验结果的可靠性。

1　水模型实验

由于水与钢液的流动特性相似且水易于观察,水模型已成为研究钢包内钢液流动行为的重要手

段。传统研究多聚焦于理想平整底部或单一冲击区凸起模型，考虑到改变钢包底部结构会对流场产生影响，以及钢包经过长期使用后，底部可能因耐火材料侵蚀或钢渣堆积形成非计划性台阶，为了降低下渣时的钢液损失，探究台阶的形成是否会对钢包下渣造成不利影响，需对钢包进行底部结构设计进行研究。以国内某钢厂钢包为原型，利用亚克力有机玻璃构建了一个相似比为 1∶5 的钢包物理模型，如图 1 所示，其中台阶 1 高度恒定为台阶 2 高度的两倍，具体结构参数见表 1。为优化钢包的底部结构，对水口直径、水口凸起高度、钢包底部台阶高度及台阶与水口距离等参数进行调整。其中，钢包底部台阶高度即为台阶 2 本身高度及台阶 1、2 的高度差；台阶与水口距离则为水口圆心到台阶 2 较近一侧的垂直距离，并通过在底部平面内旋转台阶 1、2 实现距离的改变。基于以上参数，构建多种实验工况，并在每种工况下分别进行三次重复实验，以减少偶然误差。

为模拟钢包内漩涡下渣现象及其导致的钢渣进入中间包的过程，实验操作如下：首先，向钢包模型注入 400 mm 高度的水；然后，在水面覆盖一层厚度为 20 mm 的食用油（模拟钢渣），并静置超过 30 min，以确保钢包模型内流场达到稳定状态；随后，打开水口，利用高速摄像机记录油层漩涡穿透水口（油被排出）的整个过程；最后，通过逐帧分析视频，识别油相开始进入水口的瞬间，并据此读取钢包内剩余水位的刻度，进而计算出钢包内的剩余水量。

2 数据预处理

基于水模拟实验，共收集了 69 个样本数据。这些数据涵盖了四个关键输入变量：水口直径、水口凸起高度、台阶高度以及台阶与水口的距离。各输入变量的详细统计分析和分布如图 2 所示。

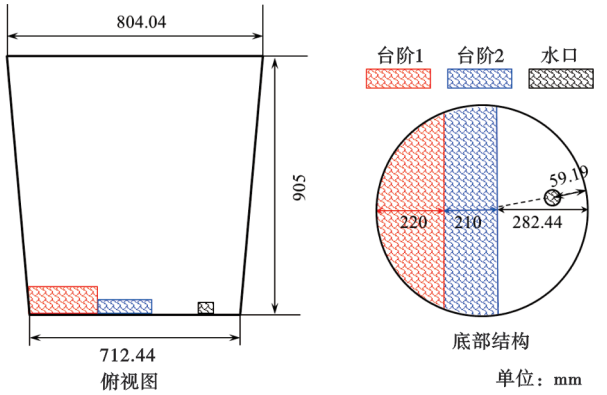

图 1 水模型结构示意图

Fig. 1 Schematic diagram of water model structure

表 1 水模型结构参数

Table 1 Parameters of water model structure

名称	原型尺寸/mm	5∶1 水模型尺寸/mm
钢包高度	4 525	905
钢包上直径	4 020.2	804.04
钢包下直径	3 562.2	712.44
底部台阶高度	470	94
抑制坝长度	706	141.22
抑制坝厚度	220	44
水口直径	250	50

由于样本中剩余水量较高的数据相对较少，这种数据的不平衡性可能会影响模型在预测高剩余水量时的性能[28]。鉴于此，采用 SMOGN[29] 方法对数据进行采样预处理。SMOGN 方法是由 Branco 提出的用于解决回归问题不平衡数据的预处理方法，其关键思想是结合随机欠采样与两种过采样方法：SmoteR 与高斯噪声。根据目标变量值构建两个分区：BinsR 分区和 BinsN 分区，其中 BinsN 分区的数据点分布为稀疏分布，BinsN 分区的数据点分布为密集分布；采用随机欠采样对所述 BinsN 分区中包含的数据点进行处理，使用过采样对所述 BinsR 中的分区进行处理。基于 BinsR 分区中每一个数据

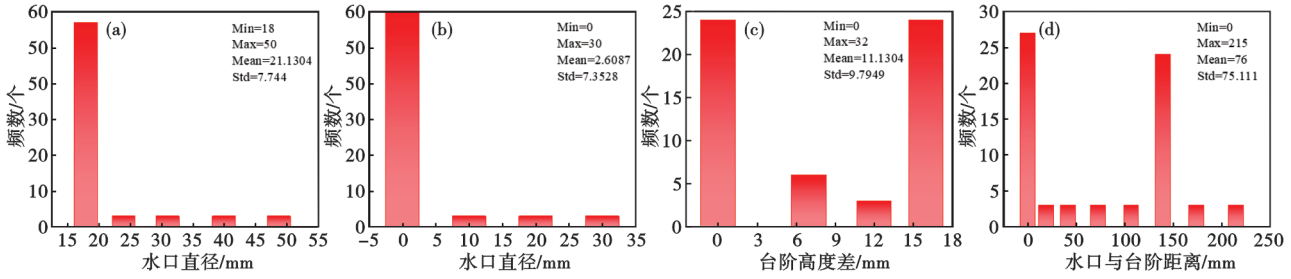

图 2 输入变量的统计结果与分布：(a)水口直径，(b)水口高度，(c)台阶高度差，(d)水口与台阶距离

Fig. 2 Statistical results and distribution of input variables : (a) diameter of nozzle, (b) height of nozzle, (c) difference in step height, (d) distance between nozzle and step

点,计算当前数据点与所述BinsR分区内剩余水模拟数据点之间的欧式距离,取中位数的一半作为“安全距离”;通过K最近邻算法对水模拟数据进行分类,设定最近邻个数为2;若当前数据点与最近邻点的距离在“安全距离”内时,使用SmoteR生成新的合成样本,若当前数据点与最近邻点的距离超过“安全距离”时,引入高斯噪声生成一个新的样本。过采样前后输出变量的分布情况对比如图3所示。经过过采样处理,样本中剩余水量较高的数据得到了有效补充,使得各区间内的数据点数量更加均衡,为模型的训练提供了稳定的基础。

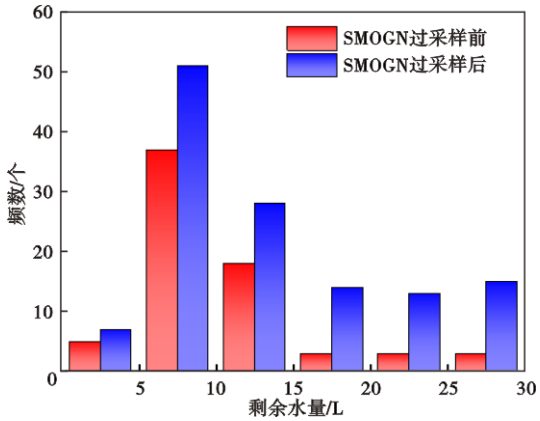

图 3　过采样前后数据分布对比

Fig. 3　Comparison of data distribution before and after oversampling

将过采样后的数据归一化为[0,1],并按照7:3的比例将数据划分为训练集和测试集。过采样前后的数据变化以及数据集的具体数量见表2。为了消除数据划分过程中可能存在的随机性对模型评估的影响,本研究在训练和测试不同模型时使用相同的训练集和测试集,从而保证模型评估的一致性和准确性。

表 2　过采样前后数据变化以及数据集数量

Table 2　Changes in data before and after oversampling and number of data sets

原始数据数量/个	过采样后数量/个	训练集数量/个	测试集数量/个
69	128	90	38

3　机器学习模型

采用最小绝对收缩和选择算子回归(Least absolute shrinkage and selection operator,LASSO)、弹性网络回归(Elastic net regression,ElasticNet)、支持向量机(Support vector regression,SVR)、多层感知机(Multilayer perceptron,MLP)以及极限梯度提升(Extreme gradient boosting,XGBoost)五种机器学习模型预测钢包水模型开始下渣时的剩余水量。

LASSO回归是一种常见的线性回归方法,最早由R. Tibshirani[30]提出,广泛应用在变量选择和参数估计中,其基本思想是对参数进行压缩,进而选择重要的变量,定义为式(1)。

$$\hat{\beta}(\mathrm{LASSO}) = \arg\min_{\beta}\left\{\left\| y - \sum_{i=1}^{p} x_i\beta_i \right\|^2 + \lambda\sum_{i=1}^{p}|\beta_i|\right\} \quad (1)$$

式中,λ为非正则参数,$\lambda\sum_{i=1}^{p}|\beta_i|$为惩罚项。$\lambda$越大,惩罚越强,更多系数被压缩为零。在建立LASSO模型时,使用5折交叉验证,确认了模型的λ为0.001 7,避免了λ过大造成的欠拟合以及过小造成的过拟合问题。

LASSO回归实际上是基于Frank提出的Birdge Regression[31],其中惩罚项的指数为1时,即得到LASSO回归,因此,一般也将LASSO称为L_1正则化。实际上,当β的指数为2时,即得到我们通常所说的岭回归,一般称为L_2正则化。ElasticNet则结合了LASSO回归和岭回归的特点[32],在建立模型时同时考虑L_1正则项和L_2正则项,前者可以实现特征选择和稀疏性,后者可以处理多重共线性问题,通过结合这两个正则项,可以得到更加稳健的模型。结合5折交叉验证,确定了ElasticNet模型的λ为0.001 67,L_1/L_2权重比为0.1。

SVR一种使用支持向量机(SVM)原理的回归模型[33]。它通过寻找一个最优的超平面来预测数据,这个超平面能够在一定程度上容忍训练数据中的噪声和异常值。在最小化预测误差的同时,SVR还会控制模型的复杂度,以避免过拟合,其回归函数为式(2)。

$$f(x) = \omega\varphi(x) + b = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i,x) + b \quad (2)$$

式中,$K(x_i,x)$为核函数,常用核函数有多项式核函数、径向基核函数、多层感知核函数等,建立时选取的核函数为多项式核函数。

MLP是一种人工神经网络,主要由输入层、隐藏层和输出层三层感知结构构成[34]。一个基本的三层神经网络结构如图4所示。每一层感知器都包含若干的神经元,而不同层之间的神经元通过权重w以及偏差b进行连接。为了调整权重,使网络能

够学习从输入到输出的映射关系，MLP采用了反向传播算法进行训练。通过改变MLP的网络结构，最后优化模型的隐含层数为2层，对应神经元数量分别为3和2，并选择Adam优化器对学习率进行自适应调节。

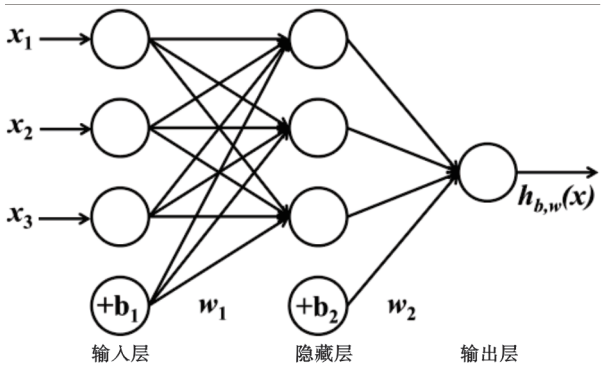

图 4 三层人工神经网络示意图

Fig. 4 Schematic diagram of a three-layer artificial neural network

XGBoost最初是由Tianqi Chen[35]作为分布式(深度)机器学习社区(DMLC)组的一部分的一个研究项目开始的，是一个可拓展的Tree boosting算法，被广泛用于数据科学领域。XGBoost中的X代表eXtreme(极致)，标志其能够更快的、更高效率的训练模型，可以看作梯度提升树(Gradient Boosting Decision Tree，GBDT)的一个改进版本。如式(3)所示，XGBoost是由k个基模型组成的一个加法运算式，其损失函数L可表示为式(4)。

$$\hat{y}_i = \sum_{t=1}^{k} f_t(x_i) \tag{3}$$

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \tag{4}$$

式中，f_k为第k个基模型，$\hat{y}_i$为第i个样本的预测值，y_i为对应的真实值，n为样本数量。模型的预测精度由模型的偏差和方差共同决定，损失函数代表了模型的偏差，想要将方差最小化，则需要简单的模型，所以目标函数Obj可以由模型的损失函数L与抑制模型复杂度的正则项Ω表示为式(5)。由于boosting模型是前向加法，以第t步的模型为例，模型对第i个样本x_i的预测为式(6)。

$$Obj = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{t=1}^{k} \Omega(f_t) \tag{5}$$

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{6}$$

式中，$\hat{y}_i^{t-1}$是由第t-1个模型给出的预测值，$f_t(x_i)$是加入的新模型的预测值，此时，目标函数就可以写成式(7)，求此时最优化目标函数，就相当于求解$f_t(x_i)$。根据泰勒公式，将函数$f(x+\Delta x)$在点x处进行泰勒的二阶展开，可得到式(8)。把$\hat{y}_i^{t-1}$视为x，$f_t(x_i)$视为Δx，则可以将目标函数写为式(9)。

$$Obj^{(t)} = \sum_{i=1}^{n} l(yi, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(yi, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \tag{7}$$

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{8}$$

$$Obj^{(t)} = \sum_{i=1}^{n} \left[l(yi, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \tag{9}$$

式中，g_i为损失函数的一阶导，h_i为损失函数的二阶导，由于在t步时$\hat{y}_i^{t-1}$其实是一个已知的值，所以$l(yi, \hat{y}_i^{t-1})$是一个常数，其对函数的优化不会产生影响，因此，目标函数可以写成式(10)。

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \tag{10}$$

模型性能通过决策系数R^2，均方误差MSE和平均绝对误差MAE三项指标进行评价，其公式定义如式(13)。

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{11}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\bar{y}_i - y_i)^2} \tag{13}$$

式中，y_i和$\hat{y}_i$分别剩余水量的实测值和预测值；$\bar{y}_i$是剩余水量的实测值的平均值；n为样本数据总数。R^2的取值范围为[0，1]，模型的R^2越高，MSE值和MAE值越低，模型的拟合效果越好。在建立XGBoost模型时，对eta(学习率)、n_estimators(迭代次数)、max_depth(树的最大深度)进行网格搜索，并取subsample(每棵树训练时随机采样的样本比例)=0.6，以提高模型的抗过拟合能力。最终选取的三个超参数的值分别为0.05、50、6。

机器学习部分中使用的LASSO，SVR，Elastic-Net，MLP以及XGBoost模型均基于Python的scikit-learn库实现的，并在一台配置为12 th Gen Intel(R) Core (TM) i9-12900H@2.50 GHz、16 GB内存的Win11系统的计算机上运行。

4　模型对比

图 5 对比了五种机器学习模型在预测剩余钢水量时的决策系数 R^2, 均方误差 MSE 和平均绝对误差 MAE。结果显示,除 LASSO 与 SVR 外,其他模型的 R^2 值均超过 0.95,显示出较高的拟合优度。在 MSE 和 MAE 指标方面,SVR 模型的预测误差较大,其 MSE 值与 MAE 值显著高于其他四种模型。相比之下,MLP、ElasticNet 与 XGBoost 模型均展现出较好的预测性能,MSE 值与 MAE 值均小于 1。其中,XGBoost 模型的 MSE 值和 MAE 值更接近于零,且 R^2 更接近于 1,这表明 XGBoost 模型在预测剩余钢水量方面具有更卓越的性能。

图 6 直观地展示了 XGBoost 模型对测试集中预测值与实际测量值之间的比较。数据点越接近对角线,表明模型的预测精度越高。从图 6 中可以看出,绝大多数数据点紧密分布在对角线附近,表明模型预测值与实际值高度一致。具体来说,在 10% 的相对误差范围内,模型的命中率达到 100%;而在 5% 的相对误差范围内,命中率依然高达 92.31%,进一步证明了 XGBoost 模型在预测剩余钢水量方面的优异性能。

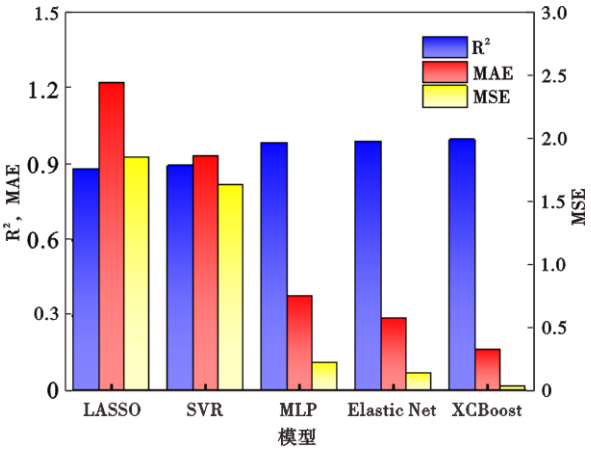

图 5　五种模型预测误差对比

Fig. 5　Comparison of prediction errors of five models

5　钢包底部结构对剩余钢水量的影响

5.1　单变量分析

基于以上分析,利用 XGBoost 模型研究了各输入变量对钢包内剩余水量的影响。采用了单一变量原则,改变所研究变量的取值范围,同时将其他输入变量取众数,具体的变量取值见表 3。

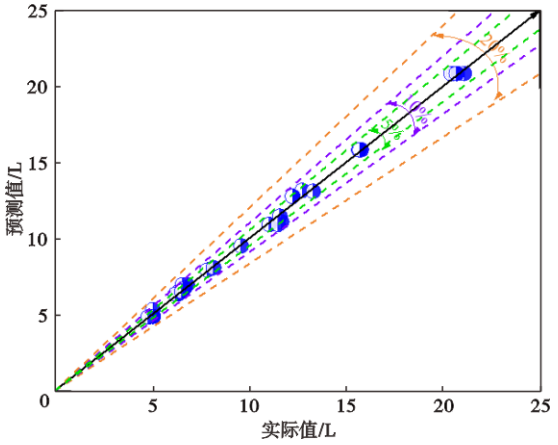

图 6　XGBoost 的预测剩余水量与实际剩余水量的对比

Fig. 6　XGBoost's predicted residual volume vs. actual residual volume

表 3　单变量研究范围取值

Table 3　The range of values for single variable studies

研究序号	水口直径/mm	水口凸起高度/mm	台阶高度/mm	台阶与水口距离/mm
1	15~55	0	16	141
2	18	0~40	16	141
3	18	0	1~41	141
4	18	0	16	0~220

图 7 对比了实验数据与模型预测值在单变量条件下的变化趋势。图中,红色空心点线代表预测值,黑色实心散点代表实验值。可以看出,水口直径超过 ϕ40 mm 时,剩余水量显著降低;随着水口凸起高度的降低和台阶高度的增高,剩余水量逐渐减少;当台阶与水口距离增大时,剩余水量先减少,在距离大于 100 mm 后趋于稳定。这些发现为钢包底部结构优化提供了重要依据,有助于减少生产中的钢水浪费。然而,图 7 显示,在台阶与水口距离的影响下,预测值普遍高于实验值,而其他三个变量的预测值与实验值基本一致。为了探究产生这一现象的原因,采用由 Lundberg 等[36]提出的 SHAP (SHapley Additive explained)方法,计算各个输入变量对模型输出的贡献,SHAP 值越大,则输入变量的贡献越高。由图 8 可知,水口凸起高度和台阶高度是影响钢包内剩余水量的主要因素;而另外两个变量,尤其是台阶与水口距离,对模型的贡献值较小,说明模型未能完全捕捉到变量与预测结果之间的深层关联,从而导致预测值与实验值之间的差距较大。

此外,图 7 中阶梯状趋势的形成的原因可归因于以下两点:一是模型的结构特性,XGBoost 模型是

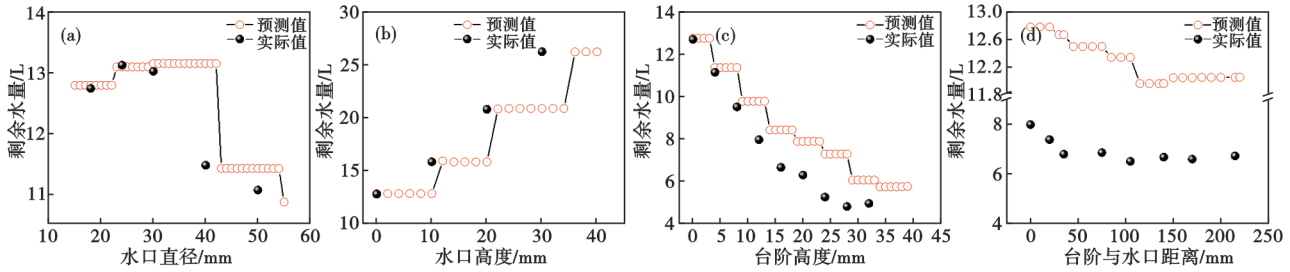

图 7 单变量影响分析:(a)水口直径,(b)水口高度,(c)台阶高度差,(d)水口与台阶距离

Fig. 7 Univariate impact analysis : (a) diameter of nozzle, (b) height of nozzle, (c) difference in step height, (d) distance between nozzle and step

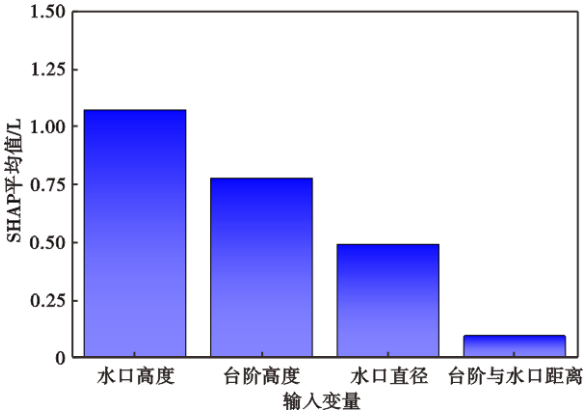

图 8 基于SHAP平均值的变量影响情况

Fig. 8 Impact of variables based on mean value of SHAP

基于决策树构建的,每个决策树都会根据特定的特征指标对输入数据进行层层的划分,以此来构建预测模型。在这个过程中,每个决策节点都对应着数据的一个子集,而这些子集在预测时可能会呈现出分段式的结果。二是实验数据的局限性,由于训练数据点的数量有限,数据分布不够密集和紧凑,某些数据区域内的样本点较为稀疏,这种数据的不均匀分布导致模型在训练时未能对这些稀疏区域的数据进行有效的划分和学习。因此,当模型对这些数据进行预测时,模型倾向于将这些数据点归类到同一决策树的“叶子”节点上,从而在预测结果中形成明显的阶梯状。

5.2 双变量分析

为了降低实验成本,使用提出的有效预测模型,采用与训练SMOGN-XGBoost模型时相同的输入变量,并同时改变其中两个变量,将其他输入变量取众数。由于台阶与水口距离对剩余水量的影响较小,预测误差相对较大,故对另外三个变量做双变量分析,具体的变量取值见表4。

双变量分析结果如图9所示。首先改变贡献值排在前两位的特征,水口凸起高度和台阶高度进行分析。分析表明,当台阶高度大于11 mm且水口凸起高度小于11 mm时,剩余水量会降至10 L以下;另一方面,当水口凸起高度超过26 mm时,剩余水量则会增加到20 L以上。当水口高度为0 mm,台阶高度高于28 mm时可以明显减少剩余水量;若台阶高度低于8 mm,且水口直径小于ϕ42 mm时,会使剩余水量显著增加。当台阶高度固定为16 mm时,水口高度低于10 mm,且水口直径大于ϕ42 mm时,剩余水量会低于12 L。

表 4 双变量研究范围取值

Table 4 The range of values for the bivariate study

研究序号	水口直径/mm	水口凸起高度/mm	台阶高度/mm	台阶与水口距离/mm
1	18	0~40	0~40	141
2	15~55	0	0~40	141
3	15~55	0~40	16	141

进行双变量分析时,固定不同的变量所影响的剩余钢水量的变化范围也不同,如图10所示。当固定某个变量,改变其他两个变量时,若剩余水量的变化量越小,则说明该变量对其影响越大。分析可得,固定水口高度时,剩余水量的变化量最小,即水口高度对剩余水量的影响效果最大,其次是台阶高度、水口直径,分析结果与图8的SHAP值反应的结果一致。综上可以得出以下结论:在生产过程中,应注重水口高度对钢包下渣的影响,将水口凸起高度控制在较低的水平,同时适当增加台阶高度,避免水口直径过小,可以有效减少钢水浪费,提高生产效率。

6 结论

结合物理模拟方法和机器学习方法,系统研究了钢包底部结构对浇铸过程中开始卷渣时剩余水量的影响,得到以下结论:

1)在五种机器学习模型中,XGBoost展现出了

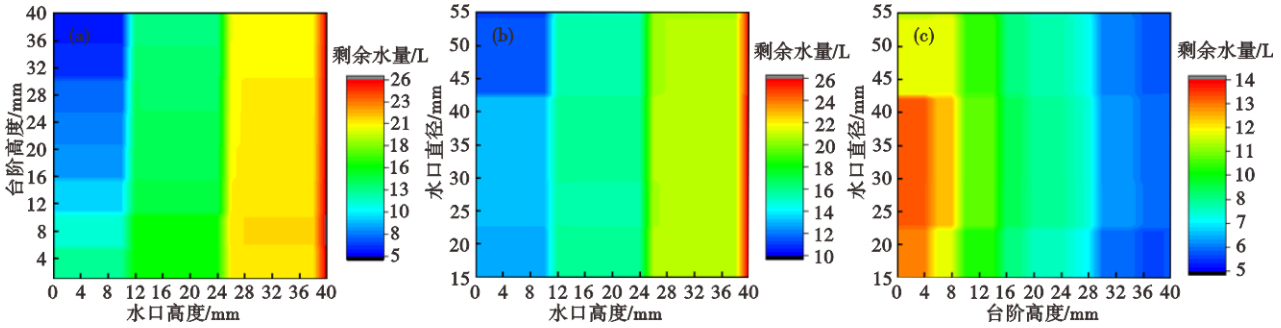

图 9　双变量影响分析:(a)水口高度与台阶高度,(b)水口高度与水口直径,(c)台阶高度与水口直径

Fig. 9　Bivariate impact analysis : (a) height of nozzle and step height, (b) height of nozzle and diameter of nozzle, (c) step height and diameter of nozzle

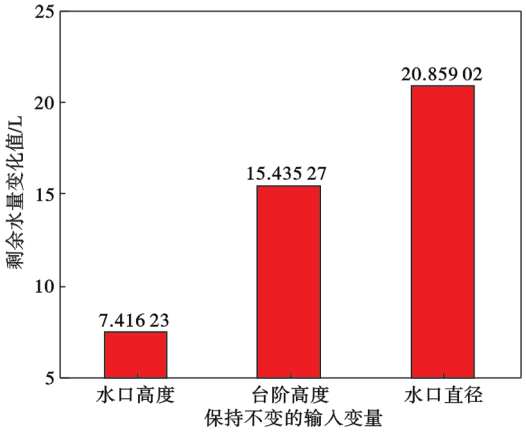

图 10　固定变量影响剩余水量变化值对比

Fig. 10　Comparison of invariant features affecting the value of change in remaining molten steel

最佳的拟合效果。具体来说,在10%的相对误差范围内,其命中率达到100%;而在5%的相对误差范围内,其命中率仍高达92.31%。表明,XGBoost模型在预测剩余水量方面具有很高的准确性和可靠性。

2)基于XGBoost模型的单变量分析表明,水口直径超过ϕ40 mm时,剩余水量显著降低;随着水口凸起高度的降低和台阶高度的增高,剩余水量逐渐减少;当台阶与水口距离增大时,剩余水量先减少,在距离大于100 mm后趋于稳定。进一步通过SHAP方法分析可知,影响钢包内剩余水量的两个主要因素为水口凸起高度和台阶高度。

3)固定台阶与水口距离的情况下进行双变量分析研究发现,当水口直径为ϕ18 mm,在台阶高度大于11 mm且水口凸起高度小于11 mm的条件下,剩余水量会降至10 L以下;若水口凸起高度超过26 mm,剩余水量则会增至20 L以上。当水口高度为0 mm,在台阶高度高于28 mm时可以明显减少剩余水量;若台阶高度低于8 mm,且水口直径小于ϕ42 mm时,会使剩余水量显著增加。当台阶高度固定为16 mm时,水口高度低于10 mm,且水口直径大于ϕ42 mm时,剩余水量会低于12 L。

参考文献

[1] Zhang L F, Thomas B G, Wang X H, et al. Evaluation and control of steel cleanliness - Review[J]. 85 th Steelmaking Conference Proceedings, 2002, 85431-452.

[2] Zhang L, Thomas B. State of the art in evaluation and control of steel cleanliness[J]. Isij International, 2003, 43(3): 271-291.

[3] Li Z W, Ouyang W, Wang Z L, et al. Physical simulation study on flow field characteristics of molten steel in 70 t ladle bottom argon blowing process[J]. Metals, 2023, 13(4): 639.

[4] Jardón-Pérez L E, Conejo A N, Amaro-Villeda A M, et al. Analysis of the Effect of Gas Injection System on the Heating Rate of a Gas Stirred Steel Ladle Assisted by Physical Modeling and PIV-PLIF Measurements [J]. ISIJ International, 2023, 63 (3): 484-491.

[5] 王家辉, 张　华, 方　庆, 等. 顶旋型湍流抑制器优化中间包流场的物理模拟[J]. 钢铁, 2023, 58(2): 72-82.

[6] 赵　烁, 卢中阳, 张　泽, 等. 双水口钢包浇注末期汇流旋涡形成的影响因素[J]. 钢铁, 2024, 59(5): 71-79.

[7] 杨晓江, 周泉林, 张　全, 等. 复吹转炉渣金间传质物理模拟及应用[J]. 钢铁, 2022, 57(12): 57-65.

[8] Yin Y B, Yang J H, Zhang J M, et al. Physical modeling of slag carryover in the last stage of ladle teeming during continuous casting with dynamic change of slide gate opening[J]. Journal of Materials Research and Technology, 2023, 23: 1781-1791.

[9] Merder T, Kozłowski S, Pieprzyca J, et al. Physical modeling of two-phase liquid - gas processes occurring in the refining ladle for Fe - Si alloy refining process[J]. Scientific Reports, 2024, 14: 17565.

[10] 李　嘉, 刘　雨, 王长军, 等. 气雾化制备18Ni250钢粉体的

数值模拟与试验验证[J]. 钢铁, 2024, 59(6): 112-121.
[11] 陈宏亮, 刘珍童, 周秋月, 等. 中间包内钢液流动和二次氧化的数值模拟[J]. 中国冶金, 2023, 33(7): 40-50.
[12] Li L, Tan D P, Yin Z C, et al. Investigation on the multiphase vortex and its fluid-solid vibration characters for sustainability production[J]. Renewable Energy, 2021, 175: 887-909.
[13] 张 振, 唐 珏, 储满生, 等. 高炉冷却板挂渣能力数值模拟分析[J]. 钢铁, 2024, 59(11): 54-64.
[14] 吴正义, 武 豪, 江雪婷, 等. 浇注末期堵流操作对中间包内钢液特性的影响分析[J]. 钢铁, 2024, 59(8): 58-69.
[15] Zhang L F, Thomas B G. Numerical simulation on inclusion transport in continuous casting mold[J]. Journal of University of Science and Technology Beijing, Mineral, Metallurgy, Material, 2006, 13(4): 293-300.
[16] Duan H J, Huang C D, Zhang L F. Numerical simulation of transient flow and heat transfer in a steel ladle during holding period [J]. Metallurgical and Materials Transactions B, 2024, 55(4): 2273-2288.
[17] Huang C D, Duan H J, Zhang L F. Modeling of motion of inclusions in argon-stirred steel ladles [J]. Steel Research International, 2024, 95(11): 2300537.
[18] Ji J H, Li D Q, Du H X, et al. Optimize the operation parameters in surging auxiliary slag skimming process based on large eddy simulation - volume of fluid - discrete phase model coupled model [J]. Steel Research International, 2024, 95(3): 2300417.
[19] Thomas R N, Gupta R. A survey on machine learning approaches and its techniques: [C]. 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science. 2020: 1.
[20] Li Y F, Chang J t, Kong C, et al. Recent progress of machine learning in flow modeling and active flow control [J]. Chinese Journal of Aeronautics, 2022, 35(4): 14-44.
[21] Zhang R H, Yang J. State of the art in applications of machine learning in steelmaking process modeling[J]. International Journal of Minerals, Metallurgy and Materials, 2023, 30(11): 2055-2075.
[22] 李维刚, 刘玮汲, 谢 璐, 等. 基于图卷积网络的热轧带钢轧制力预测[J]. 钢铁, 2023, 58(3): 89-96.
[23] 梁 印, 朱航宇, 罗林根, 等. 基于深度学习钢中非金属夹杂物图像识别[J]. 钢铁, 2023, 58(12): 62-70.
[24] 张 振, 唐 珏, 储满生, 等. 基于EEMD和机器学习的烧结矿FeO成分长短期综合预报[J]. 钢铁, 2023, 58(8): 32-40.
[25] Zhang H T, Fu H D, Li W D, et al. Empowering the sustainable development of high-end alloys *via* interpretive machine learning [J]. Advanced Materials, 2024, 36(48): 2404478.
[26] Wentzien M, Koch M, Friedrich T, et al. Machine learning-based prediction of the martensite start temperature[J]. Steel Research International, 2024, 95(10): 2400210.
[27] Zhou K X, Lin W H, Sun J K, et al. Prediction model of end-point phosphorus content for BOF based on monotone-constrained BP neural network[J]. 钢铁研究学报:英文版, 2022, 29(5): 751-760.
[28] Mitra R, McGough S F, Chakraborti T, et al. Learning from data with structured missingness [J]. Nature Machine Intelligence, 2023, 5(1): 13-23.
[29] Branco P, Torgo L, Ribeiro R. SMOGN: a pre-processing approach for imbalanced regression [C]. First international workshop on learning with imbalanced domains: Theory and applications[C]. Ljubljana: CEUR-WS, 2017: 36.
[30] Tibshirani R. Regression shrinkage and selection *via* the lasso [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.
[31] Frank L, Friedman J. A statistical view of some chemometrics regression tools[J]. Technometrics, 1993, 35(2): 109-135.
[32] Zou H, Hastie T. Addendum: Regularization and variable selection *via* the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(5): 768.
[33] Thissen U, Pepers M, B Ü, et al. Comparing support vector machines to PLS for spectral regression applications[J]. Chemometrics and Intelligent Laboratory Systems, 2004, 73(2): 169-179.
[34] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.
[35] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system [C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.
[36] Lundberg S, Lee S I. A unified approach to interpreting model predictions[EB/OL]. 2017: 1705.07874. https://arxiv.org/abs/1705.07874v2.