metals

MDPI

Article

Feature Engineering to Embed Process Knowledge: Analyzing the Energy Efficiency of Electric Arc Furnace Steelmaking

Quantum Zhuo [1], Mansour N. Al-Harbi [2] and Petrus C. Pistorius [1,*]

[1] Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; qzhuo@andrew.cmu.edu
[2] Hadeed, Tareeq 162 with 241, P.O. Box 11669, Jubail 31961, Saudi Arabia; harbimn@hadeed.com.sa
* Correspondence: pistorius@cmu.edu

Abstract: The importance of electric arc furnace (EAF) steelmaking is expected to increase worldwide as parts of the industry transition to lower carbon dioxide emissions. This work analyzed one year's operational data from an EAF plant that uses a large proportion of direct-reduced iron (DRI) in the furnace feed. The data were used to test different approaches to quantifying the effects of process conditions on specific electricity consumption (kWh per ton of crude steel). In previous work, inputs such as the proportion of DRI, fluxes, natural gas, and oxygen were linearly correlated with the specific electricity consumption. The current work has confirmed that conventional multiple linear regression (MLR) reproduces electricity consumption trends in EAF steelmaking, but many model coefficients deviated significantly from expected values and appeared unphysical. The implementation of engineered features—the slag volume and total carbon input—in an MLR model resulted in coefficients that were closer to expectations, but did not improve prediction accuracy. Further improvement was obtained by applying the engineered features to a non-linear machine-learned model (based on XGBoost), yielding both physically reasonable trends and smaller prediction errors. Trends from Shapley dependence analysis (applied to the XGBoost model) are quantitatively consistent with theoretical trends. These include the energy needed to melt slag, and the endothermic effect of carbon additions. The fitted models demonstrate the potential to diagnose poor slag foaming by showing an increase in electricity consumption with increased oxygen use. This example demonstrates that practically important steelmaking process insights inferred via a linear regression approach can be improved by applying Shapley analysis to a machine-learned model based on engineered features.

Keywords: electric arc furnace; multiple linear regression; XGBoost; Shapley; feature engineering

check for updates

Academic Editor: Changming Fang

Received: 4 November 2024
Revised: 19 December 2024
Accepted: 26 December 2024
Published: 28 December 2024

Citation: Zhuo, Q.; Al-Harbi, M.N.; Pistorius, P.C. Feature Engineering to Embed Process Knowledge: Analyzing the Energy Efficiency of Electric Arc Furnace Steelmaking. *Metals* **2025**, *15*, 13. https://doi.org/10.3390/met15010013

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Modern electric arc furnaces are flexible and efficient units for producing high-quality steel. As the steel industry evolves over the next decades—with the target of net zero carbon dioxide emissions—increasing attention will be paid to the energy consumption and emissions of electric arc furnaces.

In 2023, 29% of the world's steel production was produced using the electric arc furnace (EAF) route [1]. The share of EAF production is expected to grow as the steel industry changes to lower its carbon intensity [2]. The transition from the blast furnace–basic oxygen furnace (BF-BOF) route to a combination of ironmaking by direct reduction with green hydrogen and EAF steelmaking (H_2 DRI-EAF) will require much more electricity: around

3.5 MWh per ton of crude steel produced by the H_2 DRI-EAF route [3]. Current EAF production in the United States consumes approximately 440 kWh of electricity per ton of crude steel [4], with approximately 100 kg of Scope 1 CO_2 emissions from combustion within the furnaces [5].

The aim of the work presented here is to test two different approaches to quantifying the main process variables that affect the electricity consumption in EAF steelmaking for a furnace that uses direct-reduced iron that was produced with natural gas, additional carbon input as coal, oxygen injection, and natural gas burners. The first approach is based on linear regression of plant data, and the second on machine learning. The linear regression approach is similar to that followed in a large European Union research project, summarized in a report by Köhle et al. [6], which identified the main variables that affect electricity consumption as the amounts of direct-reduced iron (DRI), hot metal, and fluxes added to each heat, the usage of oxygen and natural gas, the tapping temperature, and the tap-to-tap time. Linear relationships were fitted to plant data and compared with theoretical values based on mass and energy balances. The equation of Köhle et al. is as follows [6]:

$$E/(\text{kWh/ton}) = 300 + 900(W_{Fe}/W_{LS} - 1) + 80(W_{DRI} + W_{HBI})/W_{LS} - 300(W_{HM}/W_{LS}) + 1600(W_{flux}/W_{LS}) + 0.7(T_{tap} - 1600) + 0.85\,t_{TT} - 8V_{NG} - 4.3V_{O2},$$

where E is the electricity consumption per tonne of taped steel, W_{Fe} is the mass of all metallic feed (ton), W_{LS} is the tap mass of liquid steel (ton), W_{DRI} and W_{HBI} (ton) are the masses of DRI and hot-briquetted iron in the charge, W_{HM} is the mass of hot metal in the furnace feed (ton), W_{flux} is the added mass of fluxes (ton), T_{tap} is the steel tapping temperature (°C), t_{TT} is the tap-to-tap time (minutes), V_{NG} is the amount of natural gas used in burners (Nm^3 per ton of steel), and V_{O2} is the amount of lance oxygen injected (Nm^3 per ton of steel).

Other multiple linear regression (MLR) approaches adopt a similar approach, using different sets of variables, and have been reviewed recently [7,8].

In this work, the idea that improved correlations would be obtained is tested by making two main changes to the approach of Köhle et al. First, an additional input variable is included—the total carbon input to the furnace (amounting to around 2.5% of the steel tap mass). As discussed below, the reaction of carbon with iron oxide is strongly endothermic and increases the energy requirement for steelmaking. Second, instead of masses of DRI and fluxes, the slag volume is calculated instead. The slag volume is the mass of slag relative to the steel mass; the energy required to heat and melt the slag constituents is a significant fraction of the process energy requirement, especially for heats that use a large proportion of DRI. The slag volume is more directly related to the energy consumption than the masses of DRI and fluxes; using a single variable (slag volume) instead of two correlated variables also avoids potential fitting problems caused by collinearity. The validity of these changes was tested in two ways. First, whether the coefficients of the linear regression are physically feasible (based on simple mass and energy balances), and second, whether the prediction errors (residual standard error) decreased.

The multiple linear regression approach would fully quantify the factors that affect EAF electricity consumption if the underlying relationships are linear, as well as if the variables are independent, which is not the case. An approach based on machine learning was also tested, since this readily handles non-linear behavior. The specific modeling method tested here is XGBoost [9], which fits relationships similar to decision trees, combining multiple trees to predict the target variable. Recent examples of the use of XGBoost serve to illustrate its utility in fitting non-linear data from diverse fields [10,11]. To reveal the underlying relationships between the input variables and the target variable in the fitted

XGBoost model, the SHAP approach (Shapley additive explanations) was used [12]. Similar to the approach to test MLR, measures of the success of the XGBoost–SHAP approach were the residual standard error and the physical feasibility of the modeled relationships.

The calculated variables (carbon input and slag volume) were used as inputs to the XGBoost model (instead of DRI and flux additions); the choice of such calculated variables, or "engineered features", has been stated to be the most effective method to include physical (and chemical) principles in machine learning models [13].

The novelty of this work is that it reduces the wide selection of potential MLR features into the most fundamental (independent) inputs and allows for the consideration of non-linear trends, such as slag foaming effects.

A preliminary analysis of the data used in this work was presented at the Molten 2024 conference [14], with a focus on the effect of slag volume on electricity consumption.

2. Methods

2.1. Plant Data

One year of data were used from an EAF plant that charges a mixture of scrap and cold DRI as metallic feed, with lime and doloma as fluxes. The AC furnace (year of construction: 2013) has a typical tap mass of 155 tons, utilizing eccentric bottom tapping, with a hot heel of 15–20 tons. The typical charge mixture is 80% DRI and 20% scrap. Natural gas is used in multiple oxyfuel burners and additional oxygen is injected through several injectors. The median DRI composition was 2.2 percent carbon, 90.3 percent total iron, 83.2 percent metallic iron, and 5.3 percent gangue. Figure 1 shows the distribution of DRI compositions (for a year of data).

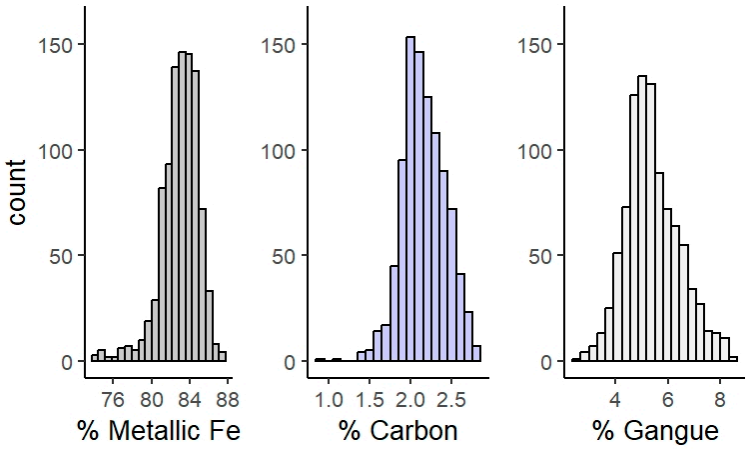


Figure 1. Distribution of compositions (mass percentages) of direct reduced iron (total number of observations: 947).

Additional carbon is fed through the top of the furnace and is injected. Slag analyses were recorded approximately three times per day; the average slag composition is in Table 1. Slag samples were taken once slag flushing started, about halfway through the heat. The iron oxide concentrations in the slag are high, similar to other operations that use a large proportion of DRI in the feed. The concentrations in Table 1 add to less than 100%, largely because the XRF analysis reports the iron oxide as Fe_2O_3, even though most of the iron in the slag is divalent. Slag samples were examined microscopically, and no significant metallic iron was detected.

Table 1. Average slag composition (mass percentages).

%CaO	$\%SiO_2$	%MgO	$\%Al_2O_3$	%MnO	%'FeO'
24.0	14.2	5.99	4.93	2.39	42.9

Some of the data entries appeared erroneous or related to unusual plant conditions; such outliers were identified by "modified Z-scores" [15,16] of greater than 3.5 and were removed from the dataset. The test for outliers was applied to the %CaO (slag analysis) and the tap-to-tap time. Details of the ranges of the values of input variables are given in Table 2.

Table 2. Variables used in scatter plot matrices.

Variable Name	Definition	Units	Range (Minimum—Lower Quartile—Median—Upper Quartile—Maximum)
DRI_ton_ton	Mass of DRI input relative to tap mass of steel	ton DRI/ton crude steel	0.41–0.70–0.73–0.77–1.16
flux_ton_ton	Total flux additions relative to tap mass of steel	ton fluxes/ton crude steel	0.019–0.037–0.039–0.041–0.055
T_tap_C	Tapping temperature of steel	°C	1578–1619–1632–1648–1714
TT_min	Tap-to-tap time	Minutes	54.5–65.9–73.8–85.4–133.7
NG_ton	Natural gas (NG) used in burners	Nm^3 NG/ton crude steel	0–0.96–1.2–1.4–2.5
O2_ton	Oxygen used in burners and lances	Nm^3 O_2/ton crude steel	15.0–19.8–21.3–22.4–27.3
Slag_vol	Mass of slag relative to tap mass of steel	kg slag/ton crude steel	69–113–123–135–186
C_kg_ton	Total carbon input relative to tap mass of steel	kg carbon/ton crude steel	15.5–23.7–25.9–28.5–39.8

After the removal of outliers, 699 of the original 806 entries remained, each referring to a different heat of steel. Of these, 12 heats featured "pour-back"—the return of some liquid steel to the EAF. The masses of returned liquid steel were small (median of 15 tons) compared with the tap mass (median of 154 tons for all heats). The small number of heats with pour-back, as well as the small mass of steel involved, implies that the dataset would allow limited evaluation at best of the decrease in electricity consumption due to the return of liquid steel.

For each entry, the slag volume was calculated from the analyzed CaO concentration in the slag together with the masses of lime and doloma added to the heat. The total carbon input was the sum of the carbon in DRI (based on the median composition), plus injected and charge carbon.

2.2. Calculation Approach

All calculations and data visualization were performed with RStudio [17], using the libraries ggplot2, ggthemes, xgboost, shapviz, olsrr, and psych.

For fitting the XGBoost model, the plant data were randomly split into training data (70%) and test data (30%). Similar to the example shown by Labram [10], different combinations of two XGBoost parameters were tested: learning rate and maximum depth. The chosen combination of these two parameters was the one that yielded the smallest root mean squared error (RMSE) upon applying the fitted XGBoost model to the test data. For these tests, the maximum tree depth varied from 4 to 10 (step size: 1), and the learning rate varied from 0.005 to 0.2 (in a 1–2–5 series). In most cases, the smallest RMSE was obtained with a learning rate of 0.01 and a maximum depth of 4. As an example, Figure 2 shows a contour plot of the RMSE values obtained for different combinations of learning rate and

maximum depth; in this example, the lowest RMSE was at a learning rate of 0.01 and a maximum depth of 4.

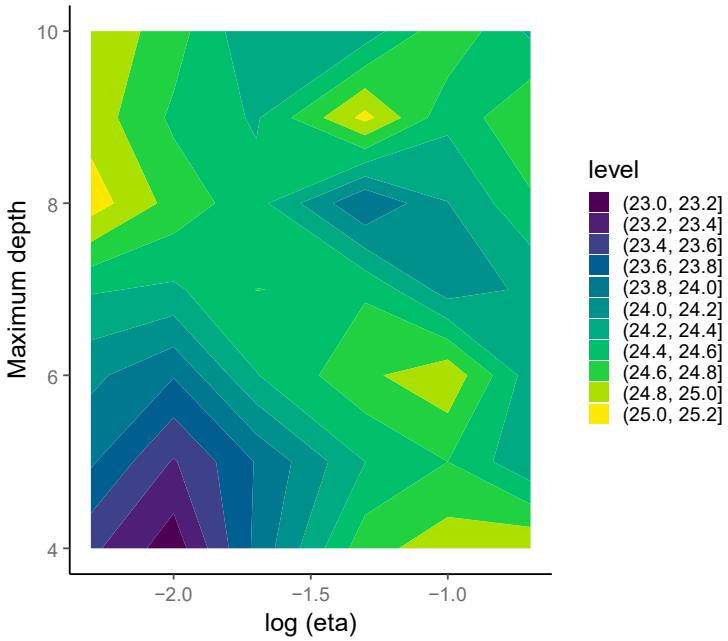


Figure 2. Contour plot of the RMSE values (in kWh/ton) obtained with test data after applying XGBoost models trained with different values of the learning rate ("eta") and the maximum tree depth; the chosen model is the one with the lowest RMSE value (at a learning rate of 0.01 and a maximum depth of 4, for this example).

To test for the effects of the random choice of training and test data on the outcome of the fitting procedure, the fitting procedure (with optimization of eta and maximum depth) was repeated 8 times.

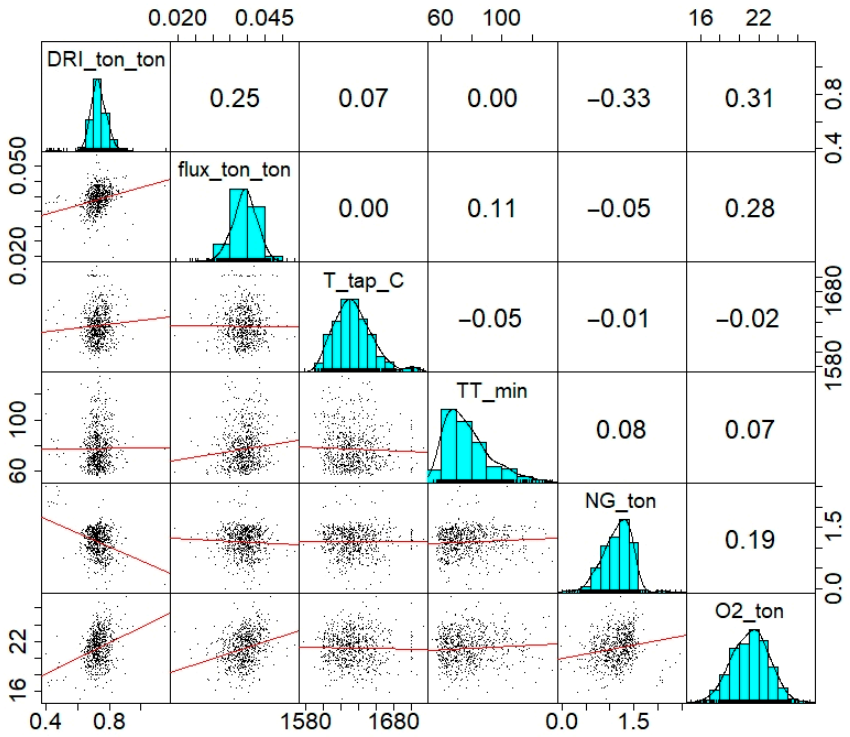
3. Results and Discussion

3.1. Distribution and Correlation of Plant Variables and Engineered Features

Figure 3 compares the distribution of input variables, with and without the engineered features. The figure is a pair of scatter plot matrices [16], with pairwise plots of variables shown below the diagonal of each matrix, the distribution of each variable on the diagonal, and Pearson correlations above the diagonal. The abbreviations of the variable names are explained in Table 2.

The scatter plot matrix shows that all the input variables (and engineered features) varied considerably over the nearly 700 heats, allowing the effects of energy consumption to be tested. The scatter plots do not show any clear relationships between the input variables, though instances of larger Pearson coefficients indicate that some collinearity might be present. The correlation coefficient of 0.25 between the DRI and flux inputs follows from the use of lime and doloma to flux the gangue in DRI. This relationship is removed in case b with the use of engineered features (slag volume and carbon input). However, invoking the engineered features replaces the correlation between flux input and DRI mass with a relatively strong correlation between the slag volume and carbon input—the DRI is the major source of both gangue and carbon in this operation. The mean ratio of DRI mass to steel tap mass is 0.73, and the average carbon concentration in the DRI is 2.16%. This means that the DRI contributes an average of 16 kg of carbon per ton of crude steel, out of the average total carbon input of 26 kg/ton.

(**a**)



(**b**)

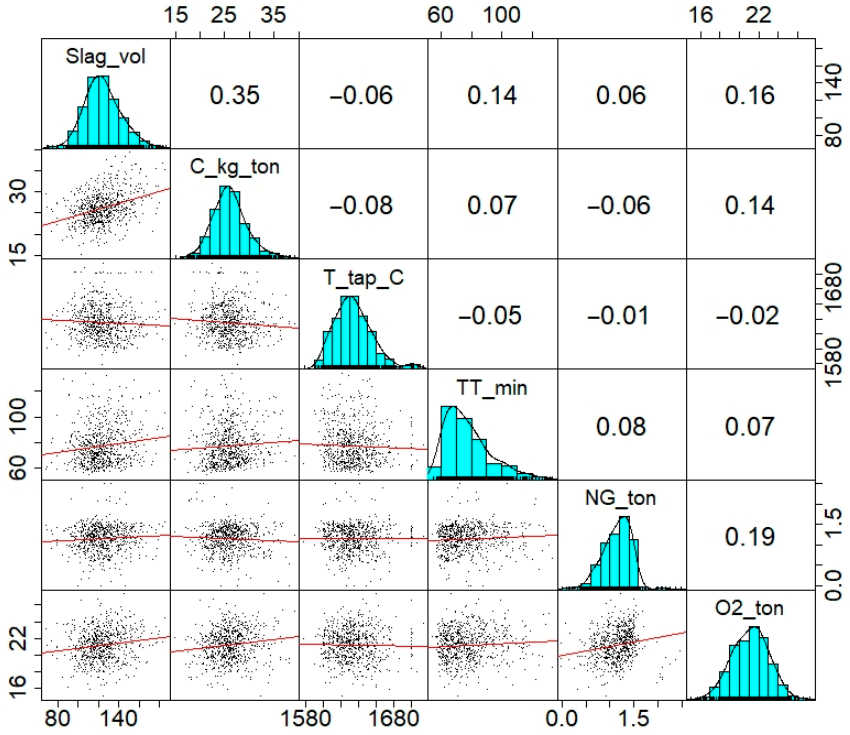


Figure 3. Scatter plot matrices of the variation of and correlation between input variables for (**a**) variables similar to those used by Köhle et al. [6] and (**b**) replacing the DRI and flux variables with the engineered features of slag volume and carbon input. Scatter plots are shown below the diagonal, variable distributions on the diagonal, and Pearson coefficients above the diagonal. See Table 2 for explanations of the variable names and their units.

Based on this comparison, it is not evident that replacing the Köhle-style input variables of DRI and flux masses with the engineered features of slag mass and total carbon input would eliminate collinearity, if that is indeed a problem for the fitted relationships.

Such issues were considered by evaluating the physical feasibility of the coefficients in the multiple linear regressions.

3.2. Physical Feasibility of Fitted Coefficients for Multiple Linear Regression

3.2.1. Calculation of Expected Relationships

The expected relationships between changes in input (and engineered) variables and the electrical energy required to produce one ton of crude steel were calculated from mass and energy balances. The underlying assumption is that the change in the energy requirement equals the change in electrical energy consumption. This is a reasonable assumption if the slag is foamed well, in which case the efficiency of heat transfer from the arc to the furnace charge is close to 100% [18]. The calculation approach is summarized in Figure 4.

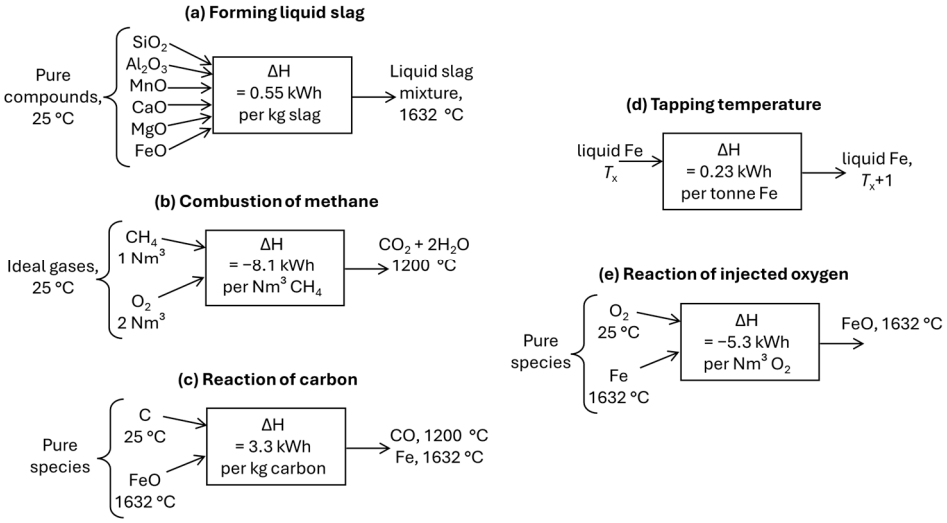


Figure 4. Illustration of the energy balances used to estimate the effects of different reactions and changes on the energy demand of EAF steelmaking: (**a**) Heating and melting pure oxides to form liquid slag; (**b**) combustion of methane with cold oxygen; (**c**) reaction of cold carbon with FeO; (**d**) a 1 °C increase in liquid steel temperature at tap; (**e**) reaction of cold injected oxygen with liquid iron.

The enthalpy changes for the reactions in Figure 4 were calculated with pure-species enthalpies from Kubaschewski et al. [19], and the model of Björkvall et al. [20] was used to calculate the enthalpy of the molten slag. These calculations required assumptions regarding the slag composition (taken to be that in Table 1), the temperature of gaseous products (taken to be 1200 °C based on results from Gandt et al. and Nakayama & Kubo [21,22]), and the slag and steel temperatures (taken to be 1632 °C, the median tapping temperature). If the actual conditions in the furnace were to depart from these assumptions, there would be some effect on the quantitative trends, but the calculated values provide a reasonable estimate.

3.2.2. Results of Multiple Linear Regression with Different Input Variables

Table 3 summarizes the results of the multiple linear regressions, using data from all the heats and the two sets of input variables to calculate the electricity consumption in kWh/ton liquid steel. The columns show the variables considered, their expected effect (taken from Köhle et al. [6] or the energy balances illustrated in Figure 4), the actual fitted coefficients with their standard errors, and the *p*-values to indicate significance. The last column is the variance inflation [23], which tests for collinearity; values of the variance inflation factor close to 1 (found for all of these variables) indicate that collinearity is not significant. This means that the use of the engineered features (slag volume and carbon

input) did not have any effect on collinearity, since collinearity appears not to play a role despite the (weak) correlation between some variables, as seen in Figure 3.

Table 3. Results of multiple linear regression to show the effect on electricity consumption (in kWh/ton) with (**a**) original input variables and (**b**) engineered features; units of the expected coefficients are shown in the first column of values.

(a) Original Input Variables; R^2 = 0.48; Residual Standard Error = 23.1 kWh/ton					
Variable	**Köhle Coeff.**	**Fitted Coeff.**	**Std Error**	***p*-Value**	**Variance Inflation**
DRI	80 kWh/ton	193	16	$<2 \times 10^{-16}$	1.4
Pour-back	−300 kWh/ton	−75	58	0.19	1.0
Fluxes	1600 kWh/ton	3323	237	$<2 \times 10^{-16}$	1.1
Tap temp.	0.7 kWh/ton.K	−0.04	0.04	0.31	1.0
Tap-to-tap	0.85 kWh/ton.min	0.35	0.06	4.4×10^{-9}	1.0
Natural gas	−8 kWh/Nm^3	10.4	3.0	0.0073	1.3
Oxygen	−4.3 kWh/Nm^3	1.21	0.53	0.023	1.3
(b) Engineered Variables; R^2 = 0.41; Residual Standard Error = 24.8 kWh/ton					
Variable	**Theoretical coeff.**	**Fitted coeff.**	**Std error**	***p*-value**	**Variance inflation**
Pour-back	−377 kWh/ton	−133	62	0.03	1.0
Slag volume	0.5 kWh/kg slag	0.31	0.06	9.8×10^{-8}	1.3
Total carbon	3.3 kWh/kg C	2.9	0.28	$<2 \times 10^{-16}$	1.3
Tap temp.	0.23 kWh/ton.K	0.07	0.04	0.13	1.0
Tap-to-tap	>0 kWh/ton.min	0.37	0.06	1.8×10^{-8}	1.0
Natural gas	−8.1 kWh/Nm^3	−9.4	2.9	0.0016	1.0
Oxygen	−5.3 kWh/Nm^3	4.7	0.5	$<2 \times 10^{-16}$	1.1

However, using the engineered features did improve the agreement between the fitted coefficients and their expected (physically feasible) values. For the original set of variables (Table 3a), all the input variables with significant effects on electricity consumption ($p < 0.05$) have quantitatively different coefficients than expected: the fitted effects of DRI and fluxes are much stronger than expected, and the apparent effect of the natural gas burners is to increase the electricity consumption (significantly so, the fitted coefficient for natural gas is about 10 kWh/Nm^3; given that the natural gas input varied from zero to 2.5 Nm^3/ton, the apparent effect on electricity consumption could be up to 25 kWh/ton). The natural gas burners help to melt scrap early in the heat; the fitted effect of the burners to increase electricity consumption does not appear feasible. In contrast, the effect of higher levels of oxygen injection to increase electricity consumption can be real, as discussed later.

The improved physical feasibility of the fitted coefficients for the engineered features is evident in Table 3b. Other than injected oxygen, all the statistically significant coefficients have the correct signs and are similar to the theoretical values. This approach to reducing the list of process variables into the most fundamental influencing variables increases the applicability to a wide variety of furnace operations and can be readily implemented in industry practice.

The increased electricity consumption with increased carbon input may appear surprising at first since the carbon is oxidized in the furnace and the oxidation of carbon by oxygen is exothermic. However, given the much larger concentration of iron than carbon in the metal bath, injected oxygen first reacts with iron to form iron oxide [24]. The oxidation of carbon hence proceeds by reactions with FeO, not with O_2. Whereas the reaction of carbon with oxygen to form carbon monoxide is exothermic, the reaction between iron oxide and carbon (yielding carbon monoxide) is endothermic, as also illustrated by Figure 4c. To further illustrate the balance between oxygen and carbon for this operation,

an approximate elemental balance (excluding oxidation by air, which is entrained into the furnace) is presented in Table 4. The balance shows that approximately half as much oxygen enters the furnace in the form of unreduced FeO in DRI, compared with injected O_2. The oxygen output terms show that the amount of FeO leaving in the slag is slightly less than the incoming FeO in the DRI (that is, a net reduction of FeO from the DRI occurs), but overall more oxygen (on a molar basis) enters the furnace than carbon (even without considering entrained air). The net effect is that oxidized iron reports to the slag. While unreduced FeO is a significant source of oxygen in this DRI-based operation, a similar balance—the net oxidation of iron—generally holds in other EAF operations, as evidenced by the typical FeO concentration in EAF slag of approximately 30% [25]. The availability of excess FeO to react with injected carbon means that an endothermic effect of carbon is also expected in scrap-based EAF operations, similar to the DRI-based operation in this study.

Table 4. Elements of an approximate oxygen and carbon balance (excluding false air), based on median values.

Input Streams		
Parameter	**Basis of Calculation**	**Amount (mol per kg Steel)**
Elemental oxygen (O) from injected O_2	21.3 Nm^3 O_2 per ton steel	1.90
Elemental oxygen (O) from FeO in DRI	7.1% of DRI is Fe in oxide form; 0.73 ton DRI per ton steel	0.93
Carbon: total from DRI, injected carbon, charge carbon, and natural gas	25.9 kg C in solid feed per ton steel; 1.2 Nm^3 natural gas per ton steel	2.21
Output Streams		
Parameter	**Basis of calculation**	**Amount (mol per kg steel)**
Elemental oxygen (O) in slag	123 kg slag per ton steel; 42.9% "FeO" in slag	0.73
Dissolved oxygen in steel	805 ppm	0.05
Summary		
(Total oxygen input) − (Total carbon input) = (1.90 + 0.93)–2.21 = 0.62 mol/kg steel		
(Total oxygen output as FeO and dissolved oxygen) = 0.73 + 0.05 = 0.78 mol/kg steel		

Of course, the reaction of carbon with iron oxide has several positive effects, including improving the yield of metallic iron and forming the carbon monoxide that foams the slag.

While the use of the more fundamental variables (including the two engineered features) improved the physical feasibility of the coefficients, the prediction accuracy did not increase and the residual standard error was similar for both cases. The next section discusses whether the use of a non-linear machine-learned model improved accuracy.

3.3. Non-linear Machine-Learned Model

Since the advantages of using the engineered features were evident from the linear regressions, the second set of input variables (those of Table 3b) were used with the machine-learning method (XGBoost). The XGBoost model did result in lower residual errors; for the eight repeats (different selections of training and testing data), the residual standard error varied from a minimum of 15.7 kWh/ton to a maximum of 20.4 kWh/ton, with an average of 18.6 kWh/ton (compared with 23.1 and 24.8 kWh/ton for the linear regressions). The improved fit is illustrated in Figure 5, which compares the actual energy consumption with the predictions of the different models.

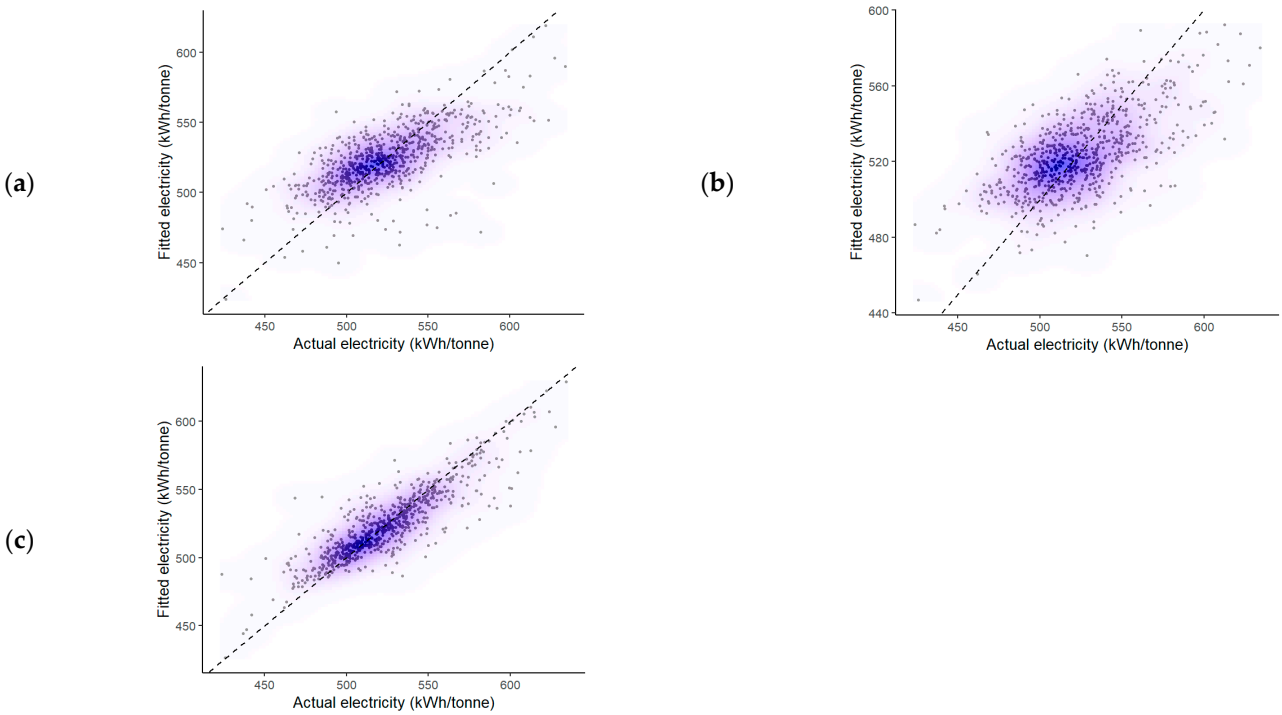


Figure 5. Comparison of fitted and actual electricity consumption for the three models; the shading indicates the distribution of points. The plots are for (**a**) multiple linear regression with original variables (RMSE = 23.1 kWh/ton); (**b**) multiple linear regression with engineered features (RMSE = 24.8 kWh/ton); and (**c**) XGBoost model with engineered features (RMSE = 18.6 kWh/ton on average).

In addition to smaller residual errors, the XGBoost model can reveal non-linear effects of input variables on electricity consumption by applying the Shapley method. This method calculates a weighted value to quantify the predicted impact on the output variable due to the inclusion of a specific input. The results of the Shapley analysis are shown in different ways in Figures 6 and 7.

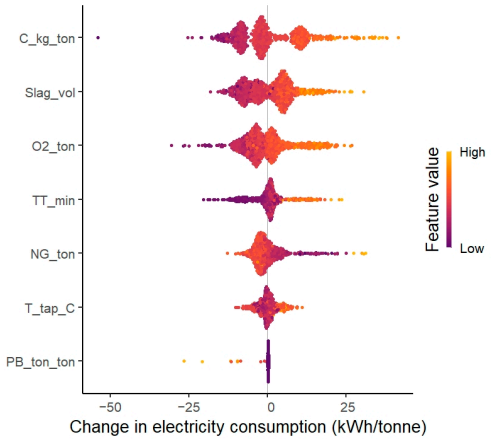


Figure 6. Beeswarm plot showing the range of effects of different input variables on electricity consumption (see Table 2 for an explanation of the variable names). The color scale shows the correlation between the change in the input variables ("features") and the change in electricity consumption.

The beeswarm plot in Figure 6 illustrates the relative importance of different input variables. The vertical position of variables in Figure 6 is chosen based on the mean absolute value of their Shapley values; that is, variables are shown in decreasing order of their effect on electricity consumption. The plot reveals that the total carbon input, injected oxygen,

and slag volume have the largest effect on electricity consumption—confirmation of the importance of the two engineered features. The clustering of the points in the beeswarm plot illustrates the distribution of the size of the effects (shown in more detail by the Shapley dependence plots in Figure 7). The color scale illustrates the direction of the effect. As examples, the beeswarm plot shows that higher values of carbon input, oxygen consumption, slag volume, and tap-to-tap time are all associated with higher electricity consumption. Of the variables considered, here, only natural gas usage decreased electricity consumption; the effects of tapping temperature and pour-back are small and not clear. The beeswarm plot illustrates how Shapley analysis provides a plant with a ranked list of inputs to target for energy efficiency.

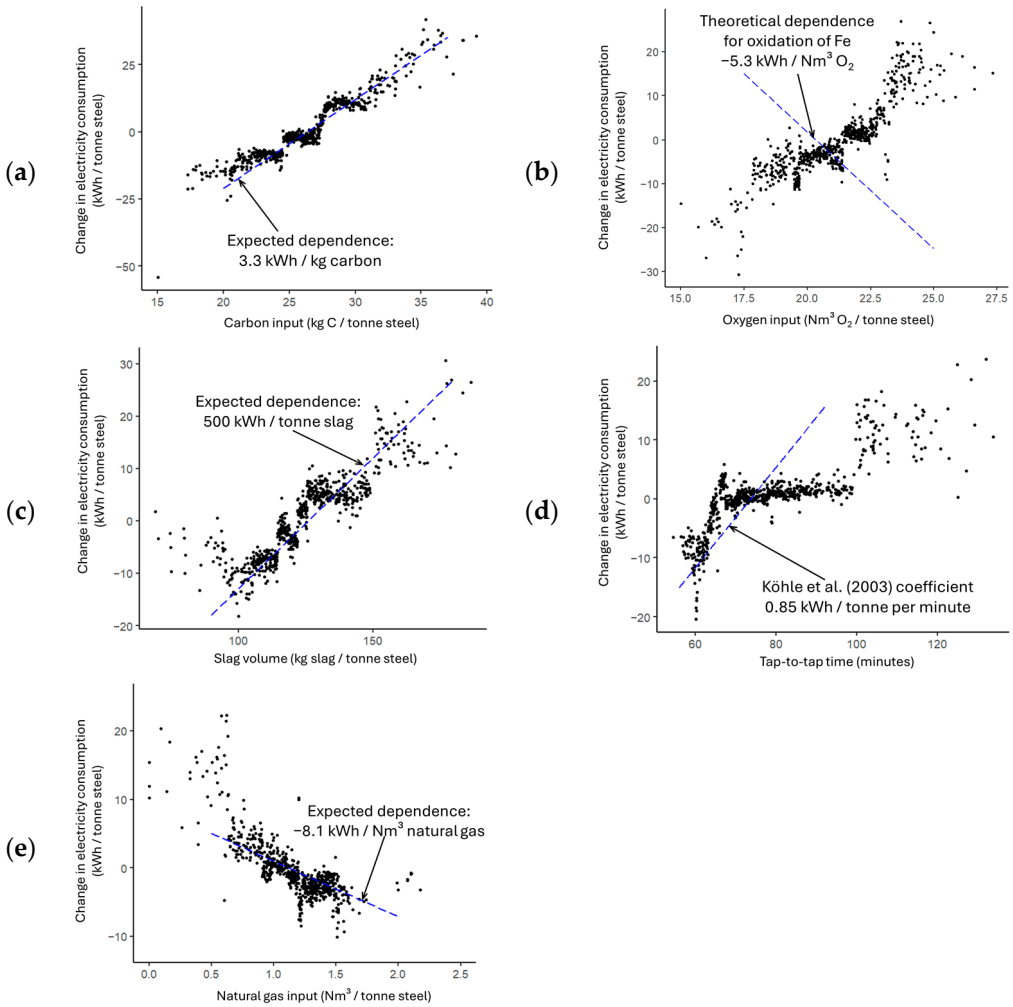


Figure 7. Shapley dependence plots showing the effects of the major input variables on electricity consumption, as fitted with the XGBoost model (data points): (**a**) total carbon input, (**b**) oxygen consumption, (**c**) slag volume, (**d**) tap-to-tap time, and (**e**) natural gas usage. The broken lines show the relationships expected from the mass and energy balances, or—in the case of tap-to-tap time—the literature value [6].

The Shapley dependence plots (Figure 7) further clarify the modeled effects of the different input variables (specifically, the total input of carbon, oxygen, and natural gas, slag volume, and tap-to-tap time), and compare the model trends with those expected from the mass and energy balances. While the effects are scattered (as expected, given that the RMSE is around 18 kWh/ton), the modeled effects of three of the input variables do broadly agree with expectations: increased carbon input is associated with higher electricity consumption, with a trend that is consistent with the calculated endothermicity of 3.3 kWh

per kilogram of carbon; the electricity consumption is higher for heats with larger slag volumes, with a trend similar to the calculated energy required to heat and melt the slag constituents (0.5 kWh per kilogram of slag); and heats in which more natural gas was used in the burners generally have lower electricity consumption.

The modeled trend of electricity consumption with tap-to-tap time (Figure 7d) shows an effect at shorter times that is steeper than the model coefficient of Köhle et al.; the value from Köhle et al. is 0.85 kWh/ton.minute, but the actual value for a furnace would depend on the average heat loss rate and heat size. Possible reasons for the plateau at tap-to-tap times longer than 80 min are not evident. However, it should be noted that heats with such long tap-to-tap times were unusual, as the median tap-to-tap time for all heats was 73.6 min, with the upper quartile at 85.3 min. Other unmodeled variables may have played a role in the heats with longer tap-to-tap times.

The Shapley analysis of the XGBoost model confirmed the effect of higher oxygen usage found by linear regression (Table 3b). Heats that used more oxygen generally had higher specific electricity consumption. All possible reactions of oxygen in the furnace are exothermic and should contribute positively to the energy balance, regardless of whether the oxygen reacts with molten iron (the dominant reaction), carbon, or carbon monoxide. However, oxygen can also affect electricity consumption by changing the slag's foaming behavior. The slag is foamed by carbon monoxide bubbles that result from the reaction of carbon (injected, in DRI, and in the metal bath) with iron oxide in the slag [18]. The iron oxide originates from FeO in the DRI, as well as from injected oxygen. While injected oxygen thus contributes to slag foaming by forming FeO, higher concentrations of iron oxide in the slag lower the foaming index [26]. The foaming index is the ratio of the foam depth to the gas flux; a slag with a lower foaming index might not cover the arcs, which would lead to lower thermal efficiency in the furnace. This possible link between the oxygen injection rate, the foaming index, and electricity consumption can be studied in the future if data on the rate of heat loss to the EAF cooling panels are available.

4. Conclusions

The analysis presented here serves two purposes:

First, as a demonstration of feasibility: The analysis of approximately 700 EAF heats demonstrates that a machine learning method can reveal significant and physically plausible trends when coupled with Shapley analysis. The tools needed for such analyses are readily available as libraries, for example, in RStudio, as used here. The calculations are rapid, taking only a few minutes when using a laptop computer for these examples. The use of engineered features with an XGBoost model resulted in improved estimation errors for electricity consumption (RMSE = 18.6 kWh/ton, compared with 23.1 kWh/ton for multiple linear regression without engineered features).

Second, for process implications: The main factors that affect electricity consumption are evident from the analysis. While the endothermicity of the slag-foaming reaction between carbon and iron oxide is not surprising, it is striking that this reaction is responsible for the largest variation in electricity consumption for these heats: the carbon input varied by approximately 24 kg of carbon per ton of steel between heats, with an expected effect on energy consumption of 3.3 kWh per kg of carbon. A similar endothermic effect of carbon additions is also expected in scrap-based operations, since in general, the slag of EAFs contains FeO that is available to react with carbon. This is the case whether the feed is mainly DRI or mainly scrap. From these results, it appears feasible to decrease the Scope 1 carbon footprint of EAF steelmaking (by injecting less oxygen and oxidizing less carbon) while maintaining adequate slag foaming. Heats that used more oxygen generally had

higher electricity consumption; a plausible reason is that high FeO concentrations in the slag (at high oxygen injection rates) decreased the effectiveness of slag foaming.

Since several process variables are modeled to have linear effects on the energy required for EAF steelmaking, multiple linear regression should be successful in fitting the relationship between such variables and electricity consumption. However, this work demonstrates that the variables to be included in the model need to be chosen carefully to avoid the fitting of unphysical relationships: feature engineering is essential. In addition, some effects might not be linear; a non-linear machine-learned model was shown to produce smaller prediction errors than multiple linear regression. One concern with machine-learned models is the lack of transparency; however, the results showed that Shapley analysis can reveal the underlying relationships, as demonstrated here.

Author Contributions: Conceptualization, Q.Z. and P.C.P.; methodology, Q.Z. and P.C.P.; software, Q.Z. and P.C.P.; validation, Q.Z. and P.C.P.; resources, M.N.A.-H. and P.C.P.; data curation, M.N.A.-H.; writing—original draft preparation, P.C.P.; writing—review and editing, Q.Z., P.C.P. and M.N.A.-H.; visualization, Q.Z. and P.C.P.; supervision, P.C.P.; project administration, P.C.P.; funding acquisition, P.C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the National Science Foundation under Grant No. OISE 2230743, and by the Center for Iron and Steelmaking Research at Carnegie Mellon University.

Data Availability Statement: Original plant data not available because of industry confidentiality.

Conflicts of Interest: Author M.N.A.-H. was employed by the company Hadeed. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. World Steel Association. *2024 World Steel in Figures*; World Steel Association: Brussels, Belgium, 2024.
2. International Energy Agency. *Iron and Steel Technology Roadmap: Towards More Sustainable Steelmaking*; OECD: Paris, France, 2020; ISBN 978-92-64-44114-9.
3. Pei, M.; Petäjäniemi, M.; Regnell, A.; Wijk, O. Toward a Fossil Free Future with HYBRIT: Development of Iron and Steelmaking Technology in Sweden and Finland. *Metals* **2020**, *10*, 972. [CrossRef]
4. Pistorius, P.C. Relating Reported Carbon Dioxide Emissions to Iron and Steelmaking Process Details. In *Extraction 2018*; Davis, B.R., Moats, M.S., Wang, S., Gregurek, D., Kapusta, J., Battle, T.P., Schlesinger, M.E., Alvear Flores, G.R., Jak, E., Goodall, G., Eds.; The Minerals, Metals & Materials Series; Springer International Publishing: Cham, Switzerland, 2018; pp. 631–638, ISBN 978-3-319-95021-1.
5. Pistorius, P.C. Decarbonization Priorities for Electric Furnace Steelmaking. In Proceedings of the 8th International Congress on the Science and Technology of Steelmaking, Montreal, QC, Canada, 2–4 August 2022; pp. 377–381.
6. Köhle, S.; Hoffman, J.; Baumert, J.C.; Picco, M.; Nyssen, P.; Filippini, E. *Improving the Productivity of Electric Arc Furnaces*; Technical Steel Research Primary Steelmaking; Office for Official Publ. of the European Communities: Luxembourg, 2003; ISBN 978-92-894-6136-8.
7. Carlsson, L.S.; Samuelsson, P.B.; Jönsson, P.G. Predicting the Electrical Energy Consumption of Electric Arc Furnaces Using Statistical Modeling. *Metals* **2019**, *9*, 959. [CrossRef]
8. Conejo, A.N. Predictive Models on Energy Consumption. In *Electric Arc Furnace: Methods to Decrease Energy Consumption*; Springer Nature: Singapore, 2024; pp. 105–130, ISBN 978-981-9740-52-9.
9. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: San Francisco, CA, USA, 2016; pp. 785–794.
10. Labram, A. Fitting Data with XGBoost | Institute and Faculty of Actuaries. Available online: https://web.archive.org/web/20240526140351/https://www.actuaries.org.uk/news-and-insights/news/article-fitting-data-xgboost (accessed on 2 November 2024).
11. Piraei, R.; Afzali, S.H.; Niazkar, M. Assessment of XGBoost to Estimate Total Sediment Loads in Rivers. *Water Resour. Manag.* **2023**, *37*, 5289–5306. [CrossRef]

12. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. Available online: http://arxiv.org/abs/1705.07874 (accessed on 3 November 2024).
13. Haghighatlari, M.; Hachmann, J. Advances of Machine Learning in Molecular Modeling and Simulation. *Curr. Opin. Chem. Eng.* **2019**, *23*, 51–57. [CrossRef]
14. Zhuo, Q.; Pistorius, P.C.; Al-Harbi, M.N. Slag Volume Effects on DRI-Based Electric Furnace Steelmaking. In Proceedings of the Molten 2024: Supporting the Transition to Sustainable Technologies, Brisbane, QLD, Australia, 17–19 June 2024; Hayes, P., Jak, E., Eds.; AusIMM: Brisbane, QLD, Australia, 2024; pp. 1337–1345.
15. Iglewicz, B.; Hoaglin, D.C. *How to Detect and Handle Outliers*; The ASQC Basic References in Quality Control: Statistical Techniques; ASQC Quality Press: Milwaukee, WI, USA, 1993; ISBN 978-0-87389-247-6.
16. NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151) 2020. Available online: https://www.itl.nist.gov/div898/handbook/ (accessed on 20 January 2024).
17. RStudio Team. *RStudio: Integrated Development Environment for R 2021*; RStudio, PBC: Boston, MA, USA, 2021. Available online: http://www.rstudio.com/ (accessed on 2 November 2024).
18. Conejo, A.N. Slag Foaming. In *Electric Arc Furnace: Methods to Decrease Energy Consumption*; Springer Nature: Singapore, 2024; pp. 233–461, ISBN 978-981-9740-52-9.
19. Kubaschewski, O.; Alcock, C.B.; Spencer, P.J. *Materials Thermochemistry*, 6th ed.; Pergamon Press: Oxford, UK, 1993; ISBN 978-0-08-041888-9.
20. Björkvall, J.; Sichen, D.; Seetharaman, S. Thermodynamic Model Calculations in Multicomponent Liquid Silicate Systems. *Ironmak. Steelmak.* **2001**, *28*, 250–257. [CrossRef]
21. Gandt, K.; Meier, T.; Echterhof, T.; Pfeifer, H. Heat Recovery from EAF Off-Gas for Steam Generation: Analytical Exergy Study of a Sample EAF Batch. *Ironmak. Steelmak.* **2016**, *43*, 581–587. [CrossRef]
22. Nakayama, M.; Kubo, H. Progress of Emission Control System in Electric Arc Furnace Melt-Shops. *NKK Tech. Rev.* **2001**, *84*, 16–23.
23. Collinearity Diagnostics. Available online: https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html (accessed on 2 November 2024).
24. Conejo, A.N. Oxygen Injection. In *Electric Arc Furnace: Methods to Decrease Energy Consumption*; Springer Nature: Singapore, 2024; pp. 131–166, ISBN 978-981-9740-52-9.
25. Teo, P.T.; Zakaria, S.K.; Salleh, S.Z.; Taib, M.A.A.; Mohd Sharif, N.; Abu Seman, A.; Mohamed, J.J.; Yusoff, M.; Yusoff, A.H.; Mohamad, M.; et al. Assessment of Electric Arc Furnace (EAF) Steel Slag Waste's Recycling Options into Value Added Green Products: A Review. *Metals* **2020**, *10*, 1347. [CrossRef]
26. Jung, S.-M.; Fruehan, R.J. Foaming Characteristics of BOF Slags. *ISIJ Int.* **2000**, *40*, 348–355. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.