Computational Materials Science 186 (2021) 110053

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

A guided analytics tool for feature selection in steel manufacturing with an application to blast furnace top gas efficiency

Stefan Stein [a,*], Chenlei Leng [a], Steve Thornton [b], Michel Randrianandrasana [b]

[a] Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom
[b] Tata Steel, R&D, Data Science & Analytics, No 9 Sir William Lyons Road, University of Warwick Science Park, Coventry CV4 7EZ, United Kingdom

ARTICLE INFO

Keywords:
Guided analytics
Autonomous analytics
Data science
Process optimization
Feature engineering
Steelmaking

ABSTRACT

In knowledge intensive industries such as steel manufacturing, application of data analytics to optimise process performance, requires effective knowledge transfer between domain experts and data scientists. This is often an inefficient path to follow, requiring much iteration whilst being suboptimal with regard to organisational knowledge capture for the long term. With the 'initial Guided Analytics for parameter Testing and controlband Extraction (iGATE)' tool we created a feature selection framework that finds influential process parameters and their optimal control bands and which can easily be made available to process operators in the form of guided analytics tool, while allowing them to modify the analysis according to their expertise. The method is embedded in a work flow whereby the extracted parameters and control bands are verified by the domain expert and a report of the analysis is automatically generated. The approach allows us to combine the power of suitable statistical analysis with process-expertise, whilst dramatically reducing the time needed for conducting the feature selection. We regard this application as a stepping stone to gain user confidence in advance of introduction of more autonomous analytics approaches. We present the statistical foundations of iGATE and illustrate its effectiveness in the form of a case study of Tata Steel blast furnace data. We have made the iGATE core functionality freely available in the `igate` package for the R programming language.

1. Introduction

Over the past decade or so Data Science has become an increasingly important topic in all aspects of business and industry. This reflects the increasing availability and power of computing resource and associated big data technologies over the same period. Data from manufacturing and business processes is being increasingly recognised as holding enormous business development potential. The vehicle for realisation of this potential is systematic data analysis and this has evolved from the traditional niche domain of the statistician to an organised *Advanced Analytics* business function commanding an increasingly prominent position on the senior management agenda [1]. Leveraging state of the art data science methods and machine learning algorithms, advanced analytics present many opportunities, including optimisation of manufacturing, maximisation of equipment effectiveness, enhanced logistics for customer service and increased precision in sales and marketing functions.

In materials science machine learning has successfully been used for the prediction of material properties on the microscopic and macroscopic scale, the discovery of new materials and the optimisation of process parameters in material synthesis [2]. We refer the reader to [3] for an overview of recent advancements and challenges for machine learning in materials science. It has also played an important role in finding new solid state electrolyte materials, in the development of battery management systems and in rechargeable battery science in general, where it was proven to be superior to traditional methods in terms of time efficiency and prediction accuracy [4,5]. In materials science traditional trial-and-error experimental methods have long be supplemented with theoretical computational simulations and to gain fundamental understanding of the properties and structures of materials of interest, modern materials science requires a close integration between computation and experiment [2]. However, modelling basic physiochemical properties of some complex materials such as rechargeable batteries is sophisticated and the resource consumption of a single computation can be large, limiting the applicability of theoretical simulations [5]. When combined with theoretical computational methods, machine learning models have successfully resolved some of these difficulties of modelling the relationships between materials

* Corresponding author.
E-mail addresses: s.stein@warwick.ac.uk (S. Stein), c.leng@warwick.ac.uk (C. Leng).

https://doi.org/10.1016/j.commatsci.2020.110053
Received 14 July 2020; Received in revised form 3 September 2020; Accepted 4 September 2020
Available online 18 September 2020

properties and physical factors and the corresponding experimental results have proven to be reliable [2].

In steel manufacturing in particular, even small improvements to stability, yield or quality make big differences to costs and profitability, making application of advanced analytics a lucrative endeavour. It must be recognized, however, that any machine learning project will only be of value to the company, if the knowledge gained from it can be transformed into actionable business insights. In the context of manufacturing, this means that the insights gained must be moved to the shop floor, where the actual wealth is created by educating the workforce about the findings and by providing them with actionable knowledge [6]. Tata Steel are making progress into this area by researching applications of *guided analytics* tools.

The premise of guided analytics is that successful application of data science requires effective knowledge transfer between domain experts having expertise about the physical problem under study and data scientists having knowledge of statistical and machine learning techniques [7]. Despite the many successes machine learning has had in many different industries and its ability to find relevant parameters and patterns, it was found that it is essential to find an approach that combines machine learning with domain expert knowledge. Otherwise, machine learning algorithms may produce a model that conflicts with expert knowledge [5]. On the other hand, eliciting expert knowledge usually requires much iteration between domain experts and data scientists and is an inefficient path to follow. It also tends to be suboptimal with regard to organisational knowledge capture for the long term, meaning that unless results of an analysis are stored in such a way that they are easily accessible and interpretable by other data science teams in the future, any knowledge gained might be lost when employees involved in the analysis leave or move to a different position within the company. It is also imperative to accurately and transparently document the assumptions underlying the analysis as assumptions may become outdated and invalid over time.

The 'initial Guided Analytics for parameter Testing and controlband Extraction (iGATE)' framework is a guided analytics tool that was designed to provide a standardized, expert knowledge driven analytics procedure that captures any expert feedback for future analyses and automates as much of the repetitive tasks of a data analysis as possible.

Especially during the feature engineering phase, before the actual analysis even begins, it is crucial to incorporate feedback from domain experts when deciding which features to include in the analysis. It has been recognized in the machine learning community that "coming up with features is difficult, time-consuming, requires expert knowledge" [8]. Systematic feature selection is also necessary to combat the curse of dimensionality, which was first described in [9], and domain expertise is needed for identifying so-called *target leakage*. Target leakage refers to using illegitimate variables to predict the values of the target variable whose distribution we want to understand [10]. In the context of manufacturing, it usually occurs, when using variables as predictors that may be highly correlated with the target variable, but that cannot be physically controlled during the manufacturing process. Any sensible tool would identify such variables as a strong predictors for the target variable, in terms of actionable business insight, however, these findings would be useless. Many experts consider target leakage one of the most insidious problems of automated machine learning [11], which illustrates that while autonomous feature engineering and by extension autonomous machine learning might be the ultimate goal in manufacturing, it is often not yet feasible and can only work for very specific and well defined tasks. Hence, improvements to the feature engineering process can be expected to go a long way in improving the results of the overall analytics project and several approaches to incorporate domain expert knowledge into a statistical or machine learning model have been proposed in the literatures. For example, [12] proposed an interactive approach using expert knowledge to identify the edges of a Bayesian network to encourage active interaction between domain experts and data scientists to ultimately improve performance and [13] used a fuzzy rule-based classification system to integrate domain expertise and data.

The DML $-\mathrm{FS}_{\mathrm{dek}}$ method presented in [14] is a feature selection scheme based on the combination of expert knowledge and the outcome of several machine learning procedures. In DML $-\mathrm{FS}_{\mathrm{dek}}$ each feature is given a score based on how important it is rated to the problem under study by various experts and the output of several machine learning methods. A feature is kept in the final model if its score exceeds a specified threshold. While we see the merit in this approach, for iGATE we have decided to focus on the situation in which it is not feasible to have experts review all of the potential features. In industrial contexts, such as modern factories, often hundreds or even thousands of process parameters are captured automatically by sensors and asking a domain expert to review them all for their suitability for a data science project is infeasible.

Therefore, we created a standardized, structured way of reviewing process parameters that required fewer iterations, required the domain expert to review fewer parameters, and stored any decisions made by domain experts and data scientists in a transparent way for future reference. Having one standardized approach for this works, because in industry the early stages of data science projects tend to follow similar patterns. For example, once a dataset has been assembled, one might look at the same type of summary statistics and plots during the initial stages of each data science project. Producing these plots and summary statistics is a repetitive procedure that can be automated to save time. We consider iGATE as a go-to tool that can be used in the initial stages of any analytics project. It automates the repetitive steps of the initial data analysis, while leaving enough flexibility for domain experts to modify the analysis according to their expertise. iGATE is a middle ground between autonomous and manual feature selection. It systematically compares good and bad products by applying statistical hypothesis tests to find a small set of potentially influential variables for review by the domain expert. iGATE was selected as a first implementation for guided analytics as the concepts involved are easily grasped by the domain experts, who possibly have had no prior statistical training. It also provides an initial estimation of favourable controlbands that under regular manufacturing conditions will result in good product quality and has been robustified to handle incomplete data. These controlbands, once validated, can immediately be translated into precise, actionable instructions for process operators. It thus combats concerns about so-called "black-box" models, which rarely enjoy the confidence of decision-makers as they lack explanatory power. The iGATE framework is schematically visualized in Fig. 1.

Depending on the industrial context, procuring samples may be very expensive and hence we chose techniques that have reasonable statistical power even for small sample sizes. Additionally, traditional statistical analysis approaches frequently rely heavily on assuming that features are distributed normally. There are many problems arising from over-reliance on normality of the data and we refer the reader to [15] and the references therein for examples. By using non-parametric statistical methods, we have robustified iGATE against different underlying distributions. While each individual method used in iGATE already existed beforehand, we combined them in a novel way to create the iGATE framework. iGATE also allows for automated, standardized reporting of the conducted analyses, aiding long term knowledge capture. The current implementation of iGATE can be considered a skeleton pipeline for advanced analytics projects to which new steps and methods can be added as user confidence in the use of guided analytics tools increases.

Steel manufacturing, like many other process industries, can be considered particularly knowledge intensive; that is, a high degree of human judgement is involved in decision making processes, making it ideal for the application of guided analytics. This is due in part to typically high degrees of legacy in information systems, but also because of practical limitations to implementation of robust sensor and actuator technologies. It is essential also of course to ensure the safety of

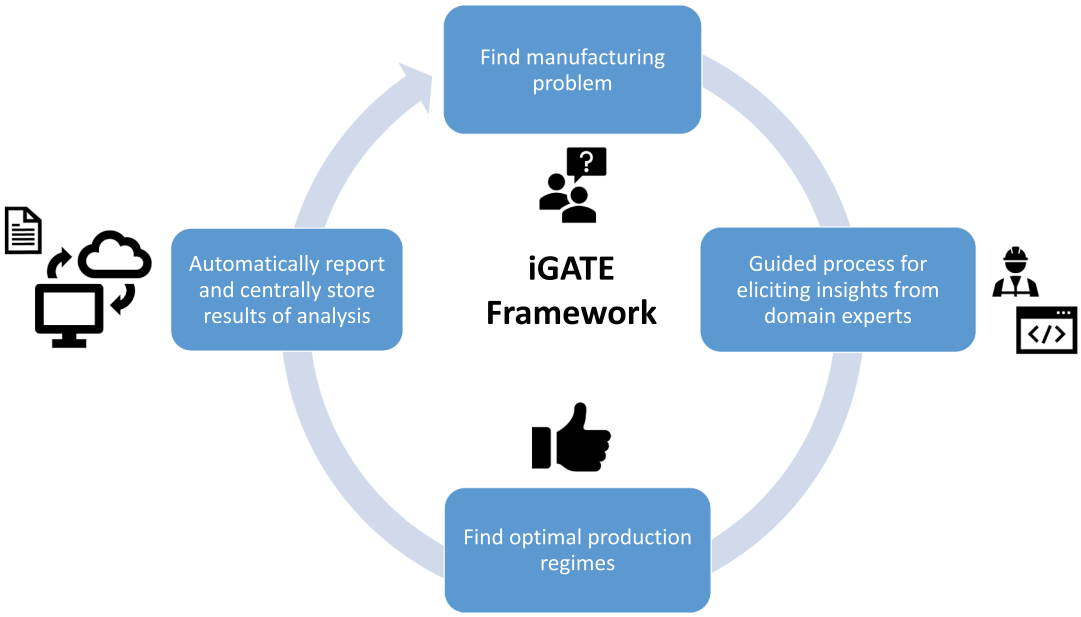

Fig. 1. Overview of the iGATE framework: Once a manufacturing problem has been found, iGATE will automatically find potentially influential process parameters using hypothesis tests. These are then systematically reviewed by a domain expert for their suitability for the problem under study. The expert is guided through this process by the application and encouraged to comment on their decision to include or discard parameters. For the retained parameters, iGATE finds optimal production regimes and validates these findings. At the end, a report of the analysis is automatically produced and centrally stored. This helps long-term knowledge retention within a company and can inform future data science projects.

personnel and operations and thus removal of the human links in the chain must not be taken lightly. Nevertheless there are many opportunities still for analytics autonomy to be implemented within the communications infrastructure where human consultation bottlenecks can safely be removed to streamline decision taking. At Tata Steel in Europe iGATE has been made available as a web based guidance tool providing on-demand analytics functionality, streamlining overall analytics usage. The tool has proven especially useful as an early step in Advanced Analytics projects, as a first consideration of dimensionality reduction, initial indication of possible relationships, and elicitation of expert knowledge regarding the relevance of data items in the context of the physical problem under investigation. The value to the company in particular lies in the domain expert being guided by the application in the decision process and by having them comment on their decisions, thus being able to capture and centrally store their expert judgement in a structured way.

After running iGATE it is possible to employ more powerful, but possibly less transparent machine learning algorithms to find correlations in the features selected by iGATE. In Section 4 we present an application of iGATE to blast furnace data from Tata Steel. Blast furnace steelmaking accounted for roughly 70% of steel produced worldwide in 2015 [16] and therefore, improving efficiency of blast furnaces has attracted a lot of attention as a lucrative area of application for advanced analytics technologies. Tata Steel also actively invests into this research [17,18]. The blast furnace is particularly interesting for guided analytics, as accurate measurement of modelling parameters is a major bottleneck according to [17] and thus data tends to be inherently messy and values need to be placed into context by domain experts. Data can be messy either because it is missing altogether or because it has been recorded incorrectly. Especially in the latter case the insights from a domain expert can be of great help. For a long time, the blast furnace itself has been considered a black-box [19]. With advances in machine learning, new approaches are suggested for modelling the blast furnace process. In [17], the authors use adaptive neural networks to predict various parameters associated with blast furnace performance and [20] model the thermal state change of the blast furnace hearth with support vector machines.

In summary, our contributions are.

1. We provide a fast framework for initial data exploration, feature selection and expert knowledge elicitation that is applicable beyond the immediate steel manufacturing application presented here.
2. iGATE works with messy data with potentially many missing or misrecorded values.
3. The results are easy to interpret and explain.
4. We provide a standardized way of documenting the analysis, the assumptions underlying it and its results, aiding effective knowledge capture.
5. We have published the technical components of the iGATE workflow, including the automatic reporting, as the `igate` package [21] for the R programming language on the Comprehensive R Archive Network (CRAN).[1]

The rest of the paper is structured as follows. Section 2 explains the mathematical foundations of the iGATE framework. Section 3 shows how iGATE can be extended to categorical target variables and in Section 4 we apply iGATE to blast furnace data provided by Tata Steel to find manufacturing parameters influencing top gas efficiency.

2. The iGATE methodology

2.1. iGATE overview

The main idea of iGATE is to compare the best products with the worst products and determine those production parameters in which they differ significantly, which are then concluded to be potentially influential for the product quality. This allows us to automatically exclude many parameters that are irrelevant to the problem under investigation. Since real-world data often contains missing values and might have outliers or wrongly recorded observations, care was taken to robustify iGATE against messy data. iGATE iteratively applies the

[1] https://cran.r-project.org.

Tukey-Duckworth test as proposed in [22], which performs well even for small sample-sizes. We explain how it works below. This statistical hypothesis test was also chosen for its ease of interpretation, making it possible to effectively explain any findings to persons with non-technical background. While we recognize that there are a variety of possibly better known statistical tools available for quantifying importance of covariates on a target variable, we also had to realize that typical modelling assumptions made by these tools, most notably the Gaussianity of the data, will frequently be violated in messy data sets. As a non-parametric hypothesis test the Tukey-Duckworth test is robust against different underlying distributions. We consider this choice a trade-off between theoretical statistical power and physical reality.

Having identified a manufacturing problem to be investigated, a data set is assembled for a typical period of operation, i.e. excluding known disturbances such as maintenance or equipment failures. This data set includes the *target variable*, i.e. the variable whose variation we want to explain and which typically gives a quantification of the quality of a product, as well as a number of features we consider potentially influential for the value of the target (called *suspected sources of variation*; SSVs). These can be automatically collected sensor data or some preselection of parameters chosen by domain experts that has yet to be refined. iGATE can be used with continuous target variables as well as categorical targets. We explain the general concept for continuous targets in this section and show how it can be generalized to categorical targets in the next section. The iGATE procedure consists of the following steps (detailed explanations follow below):

1. Select the 8 Best of the Best (BOB) and 8 Worst of the Worst (WOW) products.
2. Perform the Tukey-Duckworth test for each SSV.
3. For each SSV selected by the said test, perform unpaired Wilcoxon rank sum test.
4. Extract upper/ lower control bands for kept parameters.
5. Perform sanity check via regression plot; based on whether a trend is discernible or not and expert judgement, decide which parameters to keep.
6. Validate choice of parameters and control limits, the exact details of which are explained below.
7. Report findings in standardized format.

2.2. Products selection

When running iGATE with default settings, a box-plot approach is used for outlier detection and all observations that lie beyond 1.5 times the interquartile range of the 25th and the 75th quantile are removed before the analysis. This is justified, because we want to understand the behaviour of the target under regular production conditions. If one is interested in the insights outliers provide into the behaviour of the target, outlier removal can be switched off.

From the remaining dataset we select the 8 BOB and 8 WOW. If many samples are readily available, the number of BOB/ WOW can be specified manually by the user. When selecting the 8 best and 8 worst products in step 1, we might select observations that contain missing values for a lot of the SSVs, making analysis of those SSVs impossible. Hence, rather than selecting the same 8 BOB and 8 WOW for each SSV, we decided to select them dynamically: For the SSV we are currently investigating, we first remove those observations that contain missing values for that SSV and then select our BOB and WOW from the remaining complete cases. We conduct the analysis with the 16 selected observations and determine whether or not the current SSV is influential for the target variable. For the next SSV we repeat the process, starting again with the full data set. The user may choose to perform outlier removal for each SSV before selecting the BOB and WOW. If ties occur when selecting BOB and WOW, we select from the tied observations at random. Note, that the implicit assumption in testing each SSV in turn is that all the SSVs are independent. Of course, this is rarely the case in a real world applications. However, if two SSVs are highly correlated, we may expect that they both will be picked up by the hypothesis tests. In the further analysis following the use of iGATE, this correlation structure may then be further exploited and analysed.

2.3. The Tukey-Duckworth hypothesis test

The Tukey-Duckworth test used in step 2 is a distribution free hypothesis test pioneered in [22]. It tests whether or not two samples come from the same distribution and works as follows. After selecting the 16 BOB and WOW, we are left with 16 observations of the SSV under consideration. Denote this vector as $X = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_8$ being the values of the SSV of the BOB and $X_9, \ldots, X_{16}$ the values of the WOW respectively. Define the vector of labels $v = (v_1, \ldots, v_n)$, where $v_i = \text{BOB}$ if X_i is a value corresponding to a BOB and $v_i = \text{WOW}$ otherwise. Consider the order statistics $X_{(i)}$, where $X_{(i)}$ the i-th smallest entry of X. The *rank of* X_i is

$$R_i = \sum_{j=1}^{n} 1\left(X_j \leqslant X_i\right), \tag{1}$$

where $1(X_j \leqslant X_i)$ denotes the indicator function for the event that X_j is smaller than X_i. That is, R_i denotes the position of X_i in the ordered vector $\overline{X} = (X_{(1)}, \ldots, X_{(n)})$. We now consider the label vector ordered according to the ranks R_i,

$$\overline{v} = (v_{(1)}, \ldots, v_{(n)}),$$

where $v_{(i)}$ is the label of $X_{(i)}$. Our count summary statistic s is given as

$$s = \begin{cases} 0, & \text{if} v_{(1)} = v_{(n)}, \\ s_l + s_u, & \text{otherwise}, \end{cases}$$

where s_l and s_u are defined as the *lower* and *upper counts* defined as

$$s_l = \sum_{j=1}^{n} 1\left(v_{(1)} = \ldots = v_{(j)}\right), \quad s_u = \sum_{j=1}^{n} 1\left(v_{(n)} = \ldots = v_{v_{(n-(j-1))}}\right),$$

i.e. s_l counts how many of the entries of $\overline{X}$ at the *lower end* have the same label as $X_{(1)}$ and s_u counts how many entries of $\overline{X}$ at the *upper end* have the same label as $X_{(n)}$. If ties occur we take the average over all the possible values of s for each of the ties.

If the distribution of the BOB and WOW differ significantly in the current SSV, the BOB will cluster on one end of $\overline{X}$ and the WOW on the other. Hence, s will be large, if the distributions between BOB and WOW are significantly different and small otherwise. See Fig. 2 for an example. If the total count s is 6 or larger, we keep the SSV as potentially influential, if it is less we discard it. The critical value 6 for the summary statistic corresponds to a p-value of roughly 0.05 and is independent of the sample size 16, see [22]. The Tukey-Duckworth test is used as a preliminary step to strongly reduce the number of parameters under consideration and was chosen for its easy interpretability.

2.4. The Wilcoxon rank sum test

For those SSVs selected by the Tukey-Duckworth test, the two-sample Wilcoxon rank sum test described in [23] is performed and its result is recorded as an additional measure for the difference between BOB and WOW. It serves as a possibly more widely known alternative to the Tukey-Duckworth test, that might, however, be harder to explain to non-statisticians. Given the rank vector $R = (R_1, \ldots, R_n)$ from Eq. (1), we calculate the summary statistics

$$W_{\text{BOB}} = \sum_{i=1}^{8} R_i, \quad W_{\text{WOW}} = \sum_{i=9}^{16} R_i.$$

That is, we sum up all the ranks of the BOB and all the ranks of the

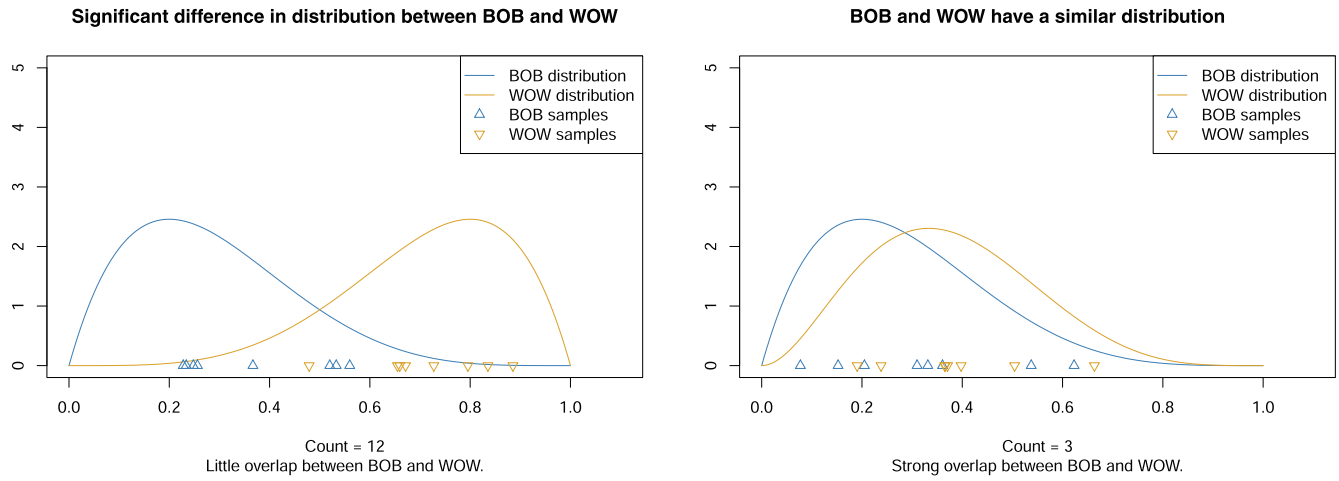

Fig. 2. Count summary statistic example for not normally distributed data. If the distribution of a specific SSV differs significantly between good products and bad products, good products will cluster at one end of the sorted vector $\overline{X}$ and bad products at the other (left plot; BOB are sampled from a Beta(2,5) distribution, WOW from a Beta(5,2) distribution). If the distribution of that SSV is similar for good and bad products, no such overlap will be observable (right plot; BOB sampled from a Beta(2,5) distribution, WOW from a Beta(3,5) distribution).

WOW. Simple algebra shows that $W_{\text{BOB}} + W_{\text{WOW}} = \sum_{i=1}^{16} i = 136$ and the values of W_{BOB} and W_{WOW} must lie between $\sum_{i=1}^{8} i = 36$ and $\sum_{i=9}^{16} = 100$. Under the null-hypothesis that the samples from the BOB and the WOW come from the same distribution, we expect W_{BOB} and W_{WOW} to take similar values. If they come from significantly different distributions, W_{BOB} and W_{WOW} will produce significantly different values and we keep the SSV under study as potentially influential. Notice that this hypothesis test is frequently also referred to as the Mann-Whitney U test, which uses slightly different, but equivalent, test statistics and was introduced independently from [23] in [24].

Since this is a multiple testing problem, we adjust the p-values using the Bonferroni-Holm procedure presented in [25]. The main function of these steps is to facilitate dimensionality reduction in the data to generate a manageable population for expert consideration. These two tests are preferred over the t-test, because they are distribution free and, while the t-test may be optimal for normally distributed data, for non-normal data it can get arbitrarily weak. Since we are using small sample sizes, normality of the data often cannot be tested reliably and hence is not a valid assumption.

2.5. Controlband extraction

For those SSVs retained after conducting the above hypothesis tests, control bands are extracted in step 4 to be used later for the validation in step 6. This is done as follows. If $s_l = k > 0$, then $v_{(1)} = \ldots = v_{(k)}$, but $v_{(k)} \neq v_{(k+1)}$. The control band for the group with label $v_{(1)}$ is then given as $I_l = [X_{(1)}, X_{(k)}]$ and we conjecture, that if the SSV under consideration is kept within I_l during production, we are more likely to obtain a good target value if $v_{(1)} = \text{BOB}$, and a bad target value if $v_{(1)} = \text{WOW}$. Similarly, if $s_u = k > 0$, we define the control band corresponding to the group of $v_{(n)}$ (which is necessarily different from $v_{(1)}$ by construction) as $I_u = [X_{(n-k+1)}, X_{(n)}]$.

2.6. Sanity check via regression plots

In step 5 a regression plot of for each SSV against the target is produced. These plots can be used by the domain expert as a sanity check for the selection done by the hypothesis tests, see Fig. 3. They can now review these plots, the extracted controlbands, the count summary statistic and the adjusted p-value together with their domain expertise to make a final decision on which parameters to keep and which to discard. If a trend can be seen in the plots and the order of magnitude of the extracted control bands align with their expertise, an SSV will be kept, otherwise discarded. Note that manual inspection of regression plots for all SSVs is often not feasible for processes with hundreds of parameters, whereas in iGATE the user will only have to check regression plots for those SSVs that passed the hypothesis tests. At this point, control bands may also be adjusted manually based on their expert knowledge.

2.7. Validation of controlbands

For the validation step, the production period from which the

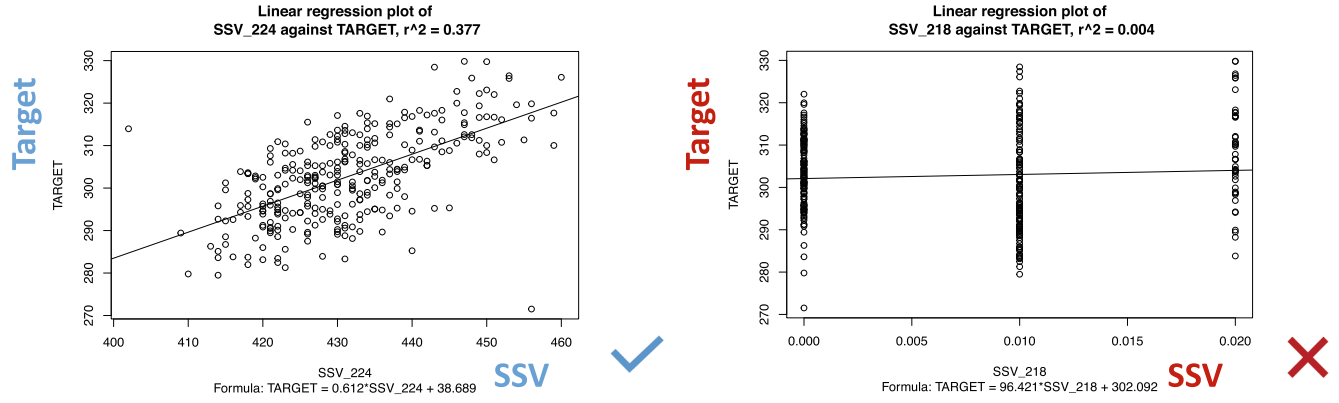

Fig. 3. Counts example: In this case the user might decide to keep the SSV on the left and discard the one on the right.

validation data is selected is dependent on the business situation, but should be from a period of operation consistent with that from which the initial population was drawn, i.e. similar product types, similar level of equipment status etc. The validation step extracts from the validation sample all the records for which any of the retained SSVs lies within these bands. The expectation is that if the SSV lies within the good band, then the target should also correspond to the good performance, and vice versa. The application gives feedback on the extent to which this criterion is satisfied, such has how many observations fall within the good/ bad band of each individual SSV, in order to help the user conclude the exploration and make recommendations for subsequent improvements.

2.8. Automatic report generation

In the last step, a report of the conducted analysis and its findings is automatically created. We provide a visual template of the report generated by iGATE in Fig. 4. For reasons of space, we omit the text and full lists of SSVs and plots, limiting ourselves to presenting the conceptual idea of what the report will look like. It starts with the "Overview" section containing the metadata of the analysis, such as when, which data set was used with which target variable. It follows with an outline of the analysis, describing the techniques used and showing a box-plot of the target variable (in case of a continuous target). In the "Results" section the retained SSVs together with their count summary statistics, their adjusted p-values and extracted control-bands are shown. This section also contains any comments made by domain experts about the SSVs. This is followed by the "Validation" section. If validation of the

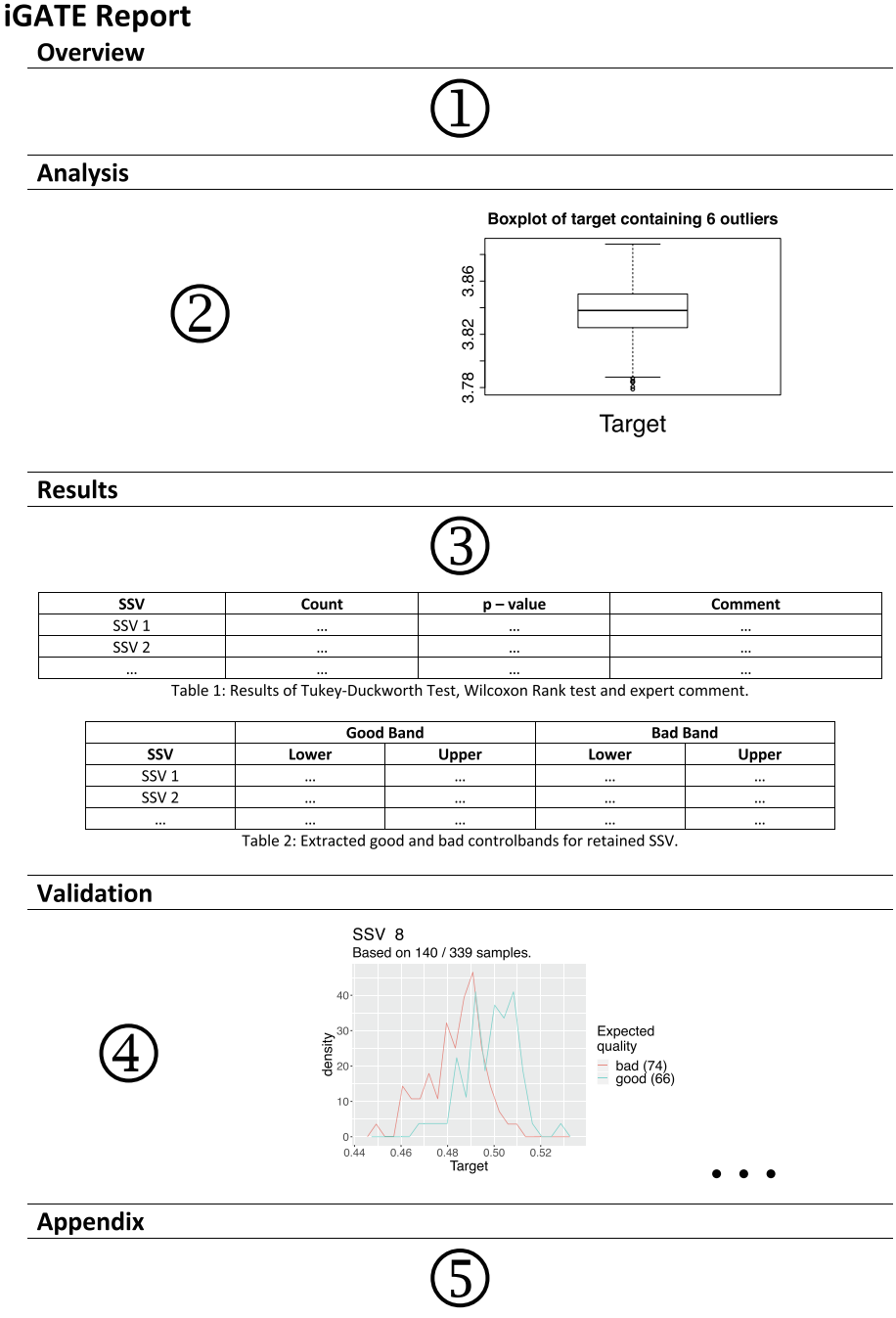

Fig. 4. A schematic representation of the report generated by iGATE. The content of each section is a combination of text, plots and tables and is automatically generated by iGATE. It is described in the following. 1) In the "Overview" section an overview of the conducted analysis is presented, including the date of the analysis, the name of the data set used, the target variable and the desired target values (high/ low) or best category. 2) The "Analysis" section contains a detailed description of the methods used, such as which hypothesis tests were used and what plots were created. In case of a continuous target variable it also contains a box-plot of the target. 3) It follows the "Results" section which gives an overview of the obtained results. A table with the SSVs selected by iGATE, the obtained count statistics, p-values and expert comments is presented, followed by another table containing the extracted controlbands for the retained SSVs. 4) The "Validation" section contains summary statistics about the validation results, such as how many samples of the validation set fall within the extracted controlbands and the distribution of the target variable amongst these samples. 5) The appendix contains a possibly long list of all the SSVs that were studied and the produced regression/ frequency plots for future reference.

results has been conducted, the results of it are presented here, listing how many observations fall into each of the extracted control bands etc. The appendix of the report contains a list of all the SSVs that were analysed as well as all the regression plots, such that if at a later stage the results of the analysis are reviewed by a different data scientist, it is clear to them how data decisions were taken in the original analysis.

3. Extending iGATE to categorical Target Variables

Using iGATE with categorical target variables is analogous to the continuous target case. The main difference is that when selecting the eight best and eight worst observations, this selection is unlikely to be unique. In this case, eight observations are selected from the best and from the worst category at random. Especially in the case of few categories with many observations in each category, this can be problematic, however, as the variance of each SSV amongst the observations within each category can still be very large and we might obtain a different result every time we run iGATE. To robustify against this, we implemented a multiple sampling approach. In this ensemble method we run iGATE with categorical target 50 times and only return those SSVs that come up as influential in at least 50% of the runs. This prevents a scenario in which we obtain a different outcome every time the analysis is conducted. The rest of the analysis follows the same steps as in the case with continuous target, the only difference being that in step 5, we do not produce regression plots. Instead, a normalized frequency plot for each retained SSV, split up by the various categories, is created. If there is a clear separation between the density curve for the best category and the curve for the worst category, the SSV is kept as potentially influential. If there is no strong distinction between the two curves, the SSV is discarded. See Fig. 5 for an example.

4. An Application to Blast Furnace top Gas Efficiency

We applied iGATE to blast furnace data provided by Tata Steel. For reasons of confidentiality we suppress the real variable names and simply refer to them generically by "SSV i" with $i = 1,\ldots,218$. We chose as target variable *top gas efficiency*, abbreviated as η_{CO}. The efficiency of a blast furnace is the amount of reductant (i.e. coke and other injectants containing carbon) used per tonne of hot metal produced. As a proxy for the efficiency of the furnace, the chemical decomposition of the *top gas*, i.e. the gas escaping at the top of the furnace, can be studied. More precisely, the top gas efficiency η_{CO} measures how efficiently the oxygen from the burden in the blast furnace is removed. It is calculated as

$$\eta_{CO} = \frac{CO_2}{CO + CO_2}.$$

That is, an increase in η_{CO} means, more CO_2 is produced rather than CO, meaning, the oxygen is removed using less gas and less coke, making the furnace more efficient. Typical values for η_{CO} are in the range of $45\% - -50\%$ [16]. It was known beforehand that η_{CO} is a definite indicator for process stability and thus, better understanding of η_{CO} would lead to better process control. Also, since it is negatively correlated with the amount of coke used, i.e. the higher η_{CO}, the less coke is needed to fuel the furnace, improvements to η_{CO} will have a direct, quantifiable business impact.

The data under study spanned around five years of daily mean η_{CO} values. In total, it contained 1692 observations and 218 potentially influential features. In total, 2.4% of all entries were missing or wrongly recorded, cf. Fig. 6. The missing data required no additional preprocessing as iGATE handles missing values automatically. Of this data set, we randomly selected 80% of the observations as training data, while we retained 20% for validation, leaving us with 1353 observations for training and 339 observations for validation.

Running iGATE on the training data returned 88 potentially influential features for η_{CO}. Upon presenting these variables to domain experts, several variables corresponding to target leakage were identified and removed from further analysis. For example, iGATE identified the coke rate (the amount of coke burned per tonne of hot metal produced) as a significant predictor for the values of η_{CO}. While coke rate is highly correlated with η_{CO}, it is a quality measure in and off itself and cannot be controlled directly. In another instance iGATE found the chemical decomposition of the coke, to be precise the concentration of a specific chemical element, to be influential. High concentrations of it were found to result in worse performance. A domain expert explained that this SSV can be interpreted as an indicator for the type of coke that is being used and that it was known that certain types of coke performed better than others, suggesting separate analyses for different coke types might be sensible. While it is statistical best practice to account for different group effects such as this one, it would have been difficult to find the right groups and parameters to adjust for without this expert's insight. This is especially true in a case like this, where the group membership of the samples is only encoded implicitly in their chemical decomposition. The experts also confirmed some of the selected SSVs. For example, a certain temperature setting was found to produce bad results when it was too

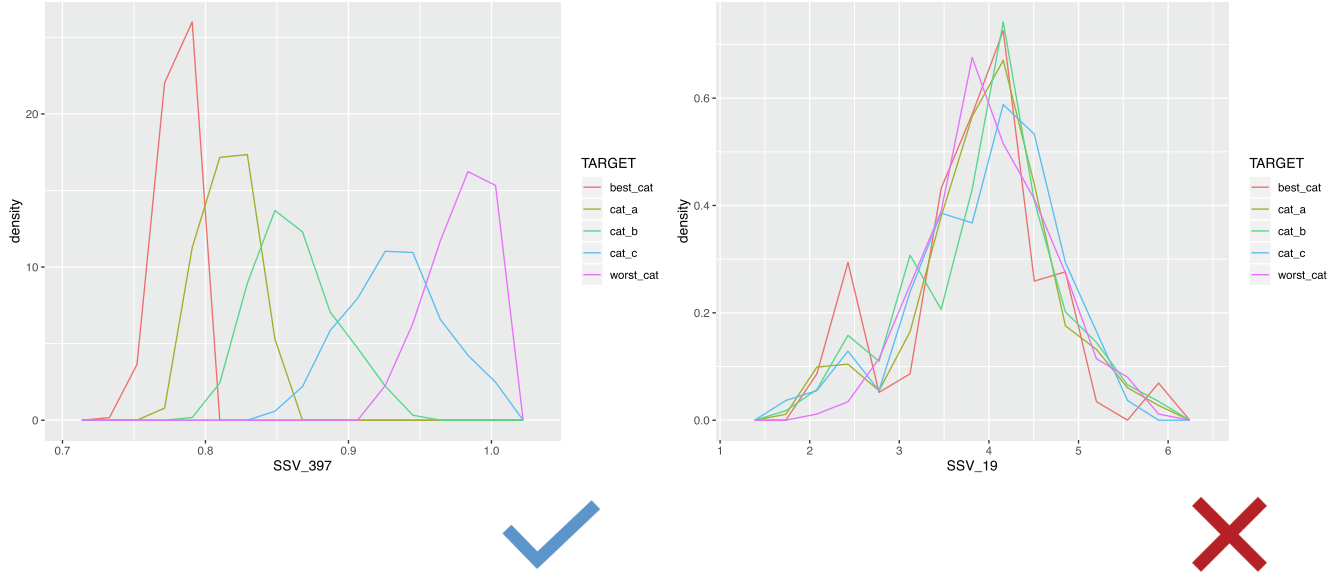

Fig. 5. Frequency polygon example: In this case the user might decide to keep the SSV on the left and discard the one on the right.

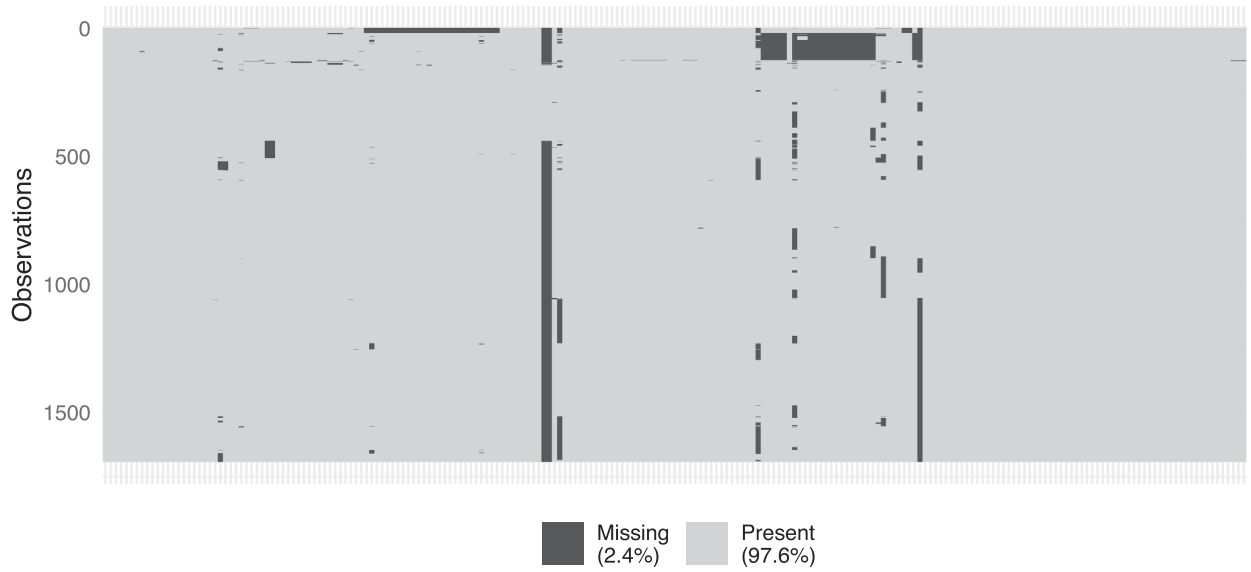

Fig. 6. Visualisation of the position of missing values in the blast furnace data. Each tick along the x-axis represents one feature, while the y-axis enumerates the rows in the data frame. Grey cells correspond to correctly recorded data. Black cells correspond to missing data. A total of 2.4% of data is missing.

high, which was interpreted to mean that if the temperature is too high, more fuel is used, producing lower values of η_{CO}. Having an expert confirm such findings and recording their comment on it can be equally valuable for the long-term knowledge capture within the company as it creates a knowledge pool that future data science projects can build on. This once more illustrates that expert feedback is essential for successful data science projects as fully autonomous approaches would not have been able to provide the necessary context to these findings. After incorporating the expert feedback, we retained 16 potentially influential variables for further analysis.

The iGATE framework does not yet employ any statistical regression modelling, hence there is no explicit loss function that can be used for validation purposes. Instead, in the validation step we try to gauge how well the control bands extracted by iGATE capture the differences

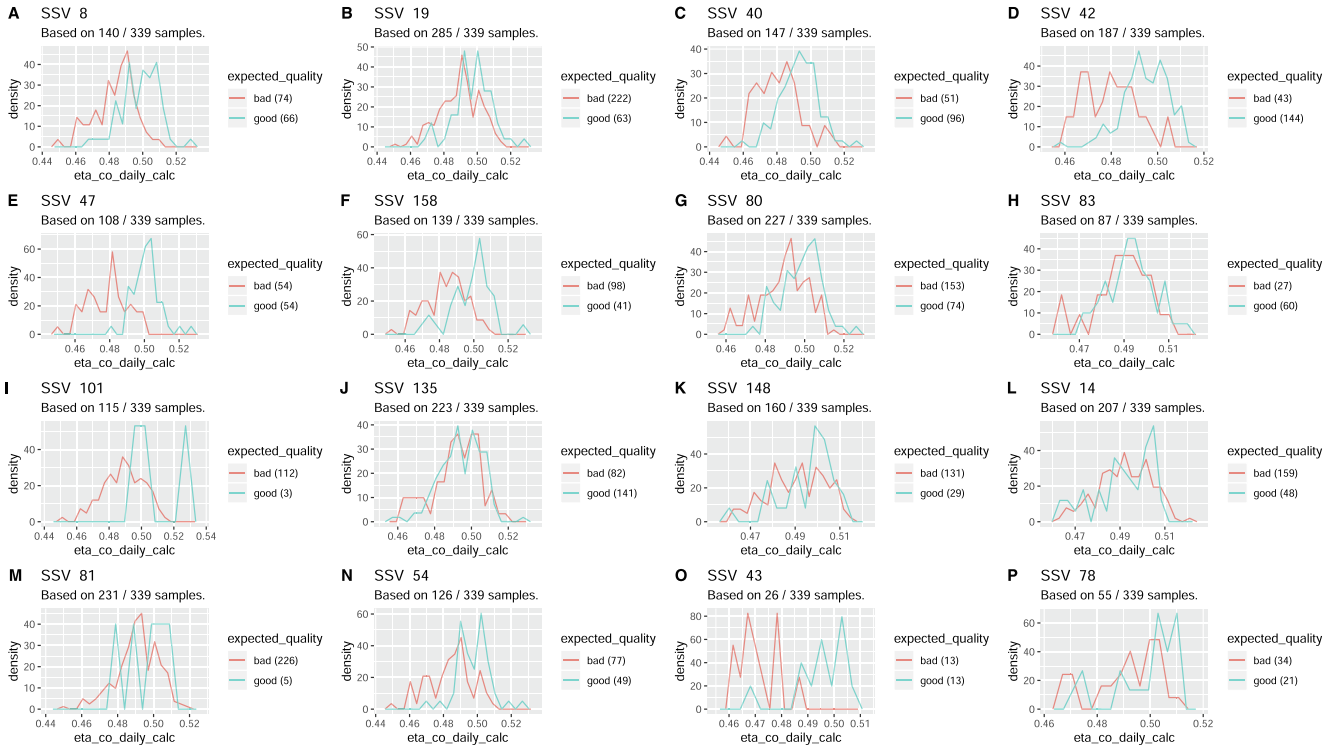

Fig. 7. Validation results. For each of the 16 retained SSVs we checked which of the 339 observations retained for validation had values for that SSV within the extracted good or bad control bands. We then plotted the η_{CO} value of these variables as a frequency plot with normalized density curve. In most cases we indeed observe that the mode of the η_{CO} values of the observations we expect to have good η_{CO} values is to the right of the mode of the observations we expect to have bad η_{CO} values, indicating that there indeed is a difference in distribution for that SSV between good and bad η_{CO} values. This is particularly prominent for plots **A, D, E, F, O**. In plots **B** or **J**, for example, no such difference is apparent, suggesting that these SSVs by themselves might not be significantly influential to the target variable after all.

between good and bad products. To that end, we took the 339 observations retained for validation, and for each retained SSV extracted those observations that fell into any of the good or bad control bands. The results are displayed in Fig. 7. It shows frequency plots of η_{CO} for those observations that fell into either of the control bands. The plots have been normalized to have density one to account for differing group sizes. We see that in most cases those observations that we would expect to yield a good, i.e. high, value of η_{CO} based on the extracted control bands indeed have higher η_{CO} values than those we would expect to yield bad η_{CO} values. This is particularly prominent in the plots **A, D, E, F, O**. There are several plots in which the distribution of η_{CO} overlaps strongly between those observations we would expect to have good quality and those we would expect to have bad quality, e.g. subplots **B** or **J**. This means that these variables on their own might not be impacting η_{CO} significantly after all and further analysis is needed.

5. Conclusion

With iGATE we created a guided analytics framework that presents a middle ground between autonomous and manual feature selection. It is fast, easy to explain to people without statistical training and the controlbands extracted by it can be translated into actionable instructions for process operators. The automated reporting feature is an integral part of iGATE that promotes knowledge capture within a company. We recognize that there are statistically more powerful tools available for assessing the influence of covariates on a target variable, but chose the tools used in iGATE for their easy interpretability and robustness against messy data. Much of the value of the traditional manual approach of domain experts and data scientists exchanging information lies in its interactivity and mutual guidance, an element which was retained in iGATE but significantly streamlined via the automation. While the methods used in iGATE already existed beforehand, novelty was added by combining them in this manner and extending them to categorical target variables.

The emphasis on explainable results seems justified to us as there commonly are concerns about basing business decisions with far-reaching consequences on results obtained from “black-box” models. We consider iGATE as a stepping stone in fostering user confidence in the use of guided analytics tools.

Disclosure Statement

Data for the case study in Section 4 was provided by Tata Steel. Steve Thornton and Michel Randrianandrasana are employed by Tata Steel.

Funding

This work was supported by Tata Steel. The authors Stein and Leng are supported by an EPSRC Industrial CASE training grant (EP/R51214X/1) with project reference number 1935144.

Data Availability

The raw data required to reproduce the findings in Section 4 cannot be shared at this time due to legal reasons. The processed data required to reproduce these findings cannot be shared at this time due to legal reasons.

CRediT authorship contribution statement

Stefan Stein: Methodology, Software, Validation, Writing - original draft, Visualization. **Chenlei Leng:** Conceptualization, Methodology, Supervision. **Steve Thornton:** Conceptualization, Supervision, Writing - original draft. **Michel Randrianandrasana:** Software, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the help and support received from blast furnace domain experts working at the Port Talbot blast furnace site. The authors also gratefully acknowledge the management of Tata Steel for providing the necessary data and permission to publish this work. The authors also thank the reviewer for their comments and suggestions.

References

[1] W.A. Jensen, Statistics = analytics? Qual. Eng. 32 (2) (2020) 133–144, https://doi.org/10.1080/08982112.2019.1633670.
[2] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Materiomics 3(3) (2017) 159–177, high-throughput Experimental and Modeling Research toward Advanced Batteries.https://doi.org/10.1016/j.jmat.2017.08.002.
[3] D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science, Annu. Rev. Mater. Res. 50 (1) (2020) 71–103, https://doi.org/10.1146/annurev-matsci-070218-010015.
[4] S. Shi, J. Gao, Y. Liu, Y. Zhao, Q. Wu, W. Ju, C. Ouyang, R. Xiao, Multi-scale computation methods: their applications in lithium-ion battery research and development, Chin. Phys. B 25 (1) (2016), 018212, https://doi.org/10.1088/1674-1056/25/1/018212.
[5] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater. 31 (2020) 434–450, https://doi.org/10.1016/j.ensm.2020.06.033.
[6] J.K. Brimacombe, The challenge of quality in continuous casting processes, Metall. Mater. Trans. A 30 (8) (1999) 1899–1912, https://doi.org/10.1007/s11661-999-0001-4.
[7] C.M. Anderson-Cook, L. Lu, P.A. Parker, Effective interdisciplinary collaboration between statisticians and other subject matter experts, Qual. Eng. 31 (1) (2019) 164–176, https://doi.org/10.1080/08982112.2018.1530357.
[8] A. Ng, Machine learning and ai via brain simulations, 2013.https://forum.stanford.edu/events/2011/2011slides/plenary/2011plenaryNg.pdf (accessed 04-Oct-2019).
[9] R. Bellman, Adaptive control processes, Princeton University Press, 1961.https://doi.org/10.1515/9781400874668.
[10] S. Kaufman, S. Rosset, C. Perlich, Leakage in data mining: formulation, detection, and avoidance, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 6, 2011, pp. 556–563, https://doi.org/10.1145/2020408.2020496.
[11] K. Larsen, D. Becker, Automated Machine Learning for Business, Oxford University Press, 2018 (in press).
[12] A.R. Masegosa, S. Moral, An interactive approach for bayesian network learning using domain/expert knowledge, Int. J. Approximate Reason. 54 (8) (2013) 1168–1181, https://doi.org/10.1016/j.ijar.2013.03.009.
[13] W. Tang, K.Z. Mao, L.O. Mak, G.W. Ng, Adaptive fuzzy rule-based classification system integrating both expert knowledge and data, in: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, vol. 1, 2012, pp. 814–821.https://doi.org/10.1109/ICTAI.2012.114.
[14] Y. Liu, J.-M. Wu, M. Avdeev, S.-Q. Shi, Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties, Adv. Theory Simul. 3 (2) (2020) 1900215, https://doi.org/10.1002/adts.201900215.
[15] S.T. Bakir, A distribution-free shewhart quality control chart based on signed-ranks, Qual. Eng. 16 (4) (2004) 613–623, https://doi.org/10.1081/QEN-120038022.
[16] M. Geerdes, R. Chaigneau, I. Kurunov, Modern Blast Furnace Ironmaking: An Introduction, third ed., IOS Press, 2015 https://doi.org/10.3233/978-1-61499-499-2-i.
[17] A. Agarwal, U. Tewary, F. Pettersson, S. Das, H. Saxén, N. Chakraborti, Analysing blast furnace data using evolutionary neural network and multiobjective genetic algorithms, Ironmaking Steelmaking 37 (5) (2010) 353–359, https://doi.org/10.1179/030192310X12683075004672.
[18] M. Guha, Revealing cohesive zone shape and location inside blast furnace, Ironmaking Steelmaking 45 (9) (2018) 787–792, https://doi.org/10.1080/03019233.2017.1338385.
[19] Y. Omori, Blast Furnace Phenomena and Modelling, Elsevier, 1987.https://doi.org/10.1007/978-94-009-3431-3.
[20] C. Gao, L. Jian, S. Luo, Modeling of the thermal state change of blast furnace hearth with support vector machines, IEEE Trans. Industr. Electron. 59 (2) (2012) 1134–1145, https://doi.org/10.1109/TIE.2011.2159693.
[21] S. Stein, igate: Guided Analytics for Testing Manufacturing Parameters, University of Warwick, r package version 0.3.3 (2019). URL:https://CRAN.R-project.org/package=igate.

[22] J.W. Tukey, A quick, compact, two-sample test to duckworth's specifications, Technometrics 1 (1) (1959) 31–48, https://doi.org/10.2307/1266308.
[23] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (6) (1945) 80–83, https://doi.org/10.2307/3001968.
[24] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1) (1947) 50–60, https://doi.org/10.1214/aoms/1177730491.
[25] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (2) (1979) 65–70. URL: http://www.jstor.org/stable/4615733.