Operations Research Perspectives 7 (2020) 100147

Contents lists available at ScienceDirect

Operations Research Perspectives

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

Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty

Alberto Garre[a], Mari Carmen Ruiz[b], Eloy Hontoria[c,*]

[a] Food Microbiology, Wageningen University & Research, P.O. Box 17, 6700 AA, Wageningen, the Netherlands
[b] Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena Antiguo Hospital de Marina (ETSII), Av. Dr. Fleming S/N, 30202, Cartagena, Spain
[c] Departamento de Economía de la Empresa, Universidad Politécnica de Cartagena Antiguo Hospital de Marina (ETSII), Av. Dr. Fleming S/N, 30202, Cartagena, Spain

ARTICLE INFO

Keywords:
Output uncertainty
Waste management
Empirical study
Production planning
Sustainability

ABSTRACT

Food production is a complex process where uncertainty is very relevant (e.g. stochastic yield and demand, variability in raw materials and ingredients…), resulting in differences between planned production and actual output. These discrepancies have an economic cost for the company (e.g. waste disposal), as well as an environmental impact (food waste and increased carbon footprint). This research aims to develop tools based on data analytics to predict the magnitude of these discrepancies, improving enterprise profitability while, at the same time, reducing environmental impact aiding food waste management.

A food company that produces liquid products based on fruits and vegetables was analyzed. Data was gathered on 1,795 batches, including the characteristics of the product (recipe, components used…) and the difference between the input and the output weight. Machine Learning (ML) algorithms were used to predict deviations in production, reducing uncertainties related to the amount of waste produced. The ML models had greater predictive capacity than a linear model with stepwise parameter selection. Then, uncertainty is included in the predictions using a normal distribution based on the residuals of the model. Furthermore, we also demonstrate that ML models can be used as a tool to identify possible production anomalies.

This research shows innovative ways to deal with uncertainty in production planning using modern methods in the field of operation research. These tools improve classical methods and provide production managers with valuable information to assess the economic benefits of improved machinery or process controls. As a consequence, accurate predictive models can potentially improve the profitability of food companies, also reducing their environmental impact.

1. Introduction

Food waste is a global issue that has raised social and governmental awareness in the last years. One third of food produced for human consumption worldwide is lost, equivalent to approximately 1300 million tons each year. This implies that the production of 30% of the agricultural surface of the planet (about 1400 million hectares) and 250 million m^3 of water are wasted [1]. It is worth noting that a fourth of the food that is currently wasted could feed all the undernourished people in the world [2]. Food waste also has a societal and environmental impact being an important contributor to climate change. For these reasons, governments and other agencies have dedicated efforts towards reducing the amount of food that is wasted worldwide.

Food waste involves every step of the food chain, including food manufacturing [3]. Food production is, in most cases, complex and wastage may be produced in each individual step of production. On top of that, due to the stochasticity of consumer demand, some food is wasted because of the excess produce that expires before it can be sold. Several management strategies have been suggested to tackle this situation, although their application to food industries is problematic. Make to Stock (MTS) strategies are not appropriate for many products due to their short shelf life and/or the need for product customization (different recipes). Multiple production runs [4] are, in most situations, economically unviable due to high set up, cleaning and production costs. New and Mapes [5] suggest Make to Order (MTO) strategies, where fixed buffer stocks, mean yield rate and a service level yield rate are defined for each single batch production, together with safety stocks. However, MTO strategies are also problematic for food products with short shelf life.

Due to the uncertainty of the production process, production is

* Corresponding author.
E-mail address: eloy.hontoria@upct.es (E. Hontoria).

https://doi.org/10.1016/j.orp.2020.100147
Received 29 July 2019; Received in revised form 15 January 2020; Accepted 21 February 2020
Available online 22 February 2020

usually planned according to an output larger than the amount demanded by the costumer. The overproduction results in several additional costs (packaging, internal and external transportation, storage…). If no similar order is received in a short time frame, the extra produce can rarely be sold and must usually be discarded due to the short shelf-life of many food products. When production is smaller than the order, the producer must negotiate new conditions with the customer. In the case of rigid demand, production runs are repeated until the order is met, whereas for non-rigid demand the company must pay a penalty [6]. Hence, underproductions have more negative consequences than overproductions. Accordingly, production is usually planned in excess and overproductions are very likely to occur.

A transition to circular economies has been suggested as a strategy to reduce the amount of food waste. They consist in the construction of symbiotic relationships between industries, where the residuals of one industry (instead of being wasted) are used as prime materials by other industries, thus enabling a reduction of the amount of produce that is wasted [7,8]. The transition to circular economies requires several technological advancements, including better tools to predict the amount of food that is wasted during production. In addition, several approaches have linked inventory management with economic and environmental objectives [9] and lot sizing with the reduction of CO_2 emissions in inventory transportation [10].

Another key factor is how yield uncertainty is modeled. Mula et al. [11] classified different approaches to tackle uncertainty in manufacturing systems where most of them were carried out through fuzzy modeling. Yano and Lee [12] carried out an extensive review of the mathematical models applicable to assist production planning under uncertainty, focusing in those cases where a statistical approach had been applied. They acknowledged that a probabilistic description of the possible output for each batch size, although ideal, requires much experimentation and data collection, as well as advanced mathematical models. Consequently, most research works have applied a deterministic approach [13], without considering the inherent uncertainty of the process. Graves [14] argued that “these simplifications were necessary in order to keep the models tractable”. As an example of a stochastic model with strong simplifications, Silver [15] analyzed the dependence of the batch size with the standard deviation of the output. Another example is the work by New and Mapes [5], who addressed uncertain production losses considering a statistical model that relates the quantities of inputs and outputs to a random yield factor.

The definition of a predictive model for food production has several pitfalls. Efendigil et al. [16] pointed out the following: (1) lack of expertise might cause a misspecification of the model function, resulting in a poor regression, (2) a large amount of data is often required to guarantee an accurate prediction, (3) non-linear patterns are difficult to capture and (4) outliers can bias the estimation of the model parameters. Several of these issues can be mitigated through the application of modern technologies for data collection and analysis. The widespread application of sensors to measure various attributes of the installation has provided further insight into the production process, reducing uncertainty [17]. However, the widespread use of sensors usually results in large datasets whose analysis can be difficult. Indeed, “classical” statistical methods are usually not suitable for datasets of high dimensionality and may slow down operations if the appropriate tools are not applied [18]. Analysis techniques based on machine learning have proved invaluable for analyzing this sort of datasets. The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience [19]. Hence, the predictive power of machine learning algorithms is reliant in the volume and quality of the data used for “training” them. For this reason, they have gained popularity during the last years due to the increased use of sensors, in many cases surpassing classical statistical methods [20]. In the context of food production, machine learning techniques have already been applied to several case studies, including the prediction of crop yield [21], effluent concentration in wastewater plants [22], or the amount of food waste in households [23].

Nevertheless, despite their great potential, ML algorithms are rarely used in actual food industries for production planning. Instead, (simple) decision rules based on experience are commonly used. In this research work, we use an empirical dataset gathered from an actual food industry to assess the application of these advanced data analytics technique in this context. Furthermore, we explore ways to model yield uncertainty, so it can be included in model predictions. The study is based on a large empirical dataset, where the production output has been measured for more than a thousand of batches (described in Section 2). A fraction of the dataset gathered has been used to train several machine learning algorithms, whereas the remaining data points have been used for model validation. Moreover, a linear model with stepwise selection has been used as benchmark. Then, statistical techniques are used to compare among the tested models and identify the most accurate one (Section 3). Then, we define a stochastic model to describe the uncertainty of the predictions of the model with the best predictive power. The managerial applications of the model developed, as well as its environmental impact, are discussed in Section 4. Finally, the article ends with a set of conclusions drawn from the study in Section 5.

2. Materials and methods

2.1. Description of the food industry analyzed

A food industry located in the south part of Spain was analyzed in this work. It produces liquid products based on fruits and vegetables with different recipes, i.e. products are not standardized and are adapted to customer orders following a typical MTO strategy. As a consequence, there exists a high variability between orders in relation to quantities, recipes and packaging requirements. The typical shelf-life of these products usually ranges between one week and one month. On relation to the production plant, it is a serial system constituted of several machineries connected through pipes with different diameters, i.e. the output of one machine serves as input for the following one. Raw materials (mainly vegetables and fruits) are washed, cleaned and filtered before cutting and mixing using mixers. Almost-finished products are processed in pasteurization tanks, each with a maximum capacity of 20 Tons. The number of pasteurization tanks is adjusted according to the produced amount (with a maximum of nine tanks). After pasteurization, the product remains in buffer tank(s) until packaging. Then, the filling process begins after the product has been homogenized. The products are served in different packaging formats (bottles, barrels and boxes with different capacities) to meet the customer needs. Once production is completed, raw materials, semi-finished and finished products remain in machineries and pipes generating wastes which cannot be recovered. Therefore, the machinery and pipes must be cleaned using pressurized water. The remaining product must be wasted because it is mixed with water.

Currently, due to lack of knowledge, production is planned based on the experience of production managers. Only two variables (the amount of tanks to be used and the final packaging format) are currently used for the production planning.

2.2. Data collection

Data associated to waste generation corresponding to actual production was collected. No data was gathered under defective conditions (e.g. production process out of control, or defective inputs or outputs). Data belonging to former batches was discarded. During a period of ten months, daily records were taken, accounting for a total of 1795 batches using the Scada software. Factors that could be relevant for waste production were identified together with the managers of the plant, resulting in twenty variables. Factors related to the characteristics of the production (planned amount, number of tanks, filler used…), the

type of packaging and the characteristics of the product (pH, water activity…). Supplementary Material 1 summarizes the variables considered.

2.3. Predictive models

The regression problem can be described as finding a function f(X, θ) that maps a vector of regressors (X) of length n to an output variable (Y). The algebraic form of the function, f, depends on the problem analyzed. The model function, f, usually depends on some unknown model parameters (θ) that must be estimated using observations. In this study, X stands for the operation variables described in Supplementary material 1, whereas Y is the proportion of production losses ($Y=\frac{\text{output weight}}{\text{input weight}}$). Nine different regression models have been applied in this study: linear model with stepwise selection, regression tree, bagged tree, random forest, gradient boosting, ridge regression, lasso regression, elastic net, and spline regression. This section presents a brief description of the models used. All the functions required for the model fitting have been implemented in R version 3.5.1 [24], using the functions included in the *caret* package [25]. The R code implemented for model training and validation has been uploaded to the GitHub page of one of the co-authors (https://github.com/albgarre/ML-foodWaste).

2.3.1. Linear regression model with stepwise selection

The linear regression model [26] describes Y as a linear combination of the n regressors, as shown in Eq. (1).

$$Y=\beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{1}$$

This model has $n + 1$ model parameters (β_i) that must be estimated using observations. In linear regression, these are estimated as the vector of parameters $\hat{\beta}_i$ that minimize the square distance between the vector of model predictions $\hat{Y}$ and the observations Y. This is equivalent to the maximization of the log-likelihood when the residuals are assumed normally distributed.

However, due to correlations between regressors or the use of regressors without an effect on Y, linear regression models are prone to overfitting [20]. Consequently, a stepwise model selection algorithm based on the Akaike Information Criterion (AIC) has been used [27]. Briefly, this model selection algorithm begins by fitting the linear model with every single regressor X_i. Then, the model fitting is repeated omitting/adding one regressor. If the AIC calculated for the new model is lower than the one for the original model, the new model is kept. This algorithm is repeated until no improvement in the AIC is attained by adding/removing regressors.

2.3.2. Tree-based models: regression tree, bootstrap aggregated (bagged) tree, random forest

Tree-based models are commonly used for classification and regression. They define a recursive binary partition, where an observation is classified in different categories according to the values of the regressors. If the output variable (Y) is continuous, they are denominated regression trees. The complexity of the model is given by the number of branches (number of partitions) and the number of regressors considered. High values of both will result in overfitting.

One of the advantages of tree-based models with respect to other ML algorithms is their ease of interpretation. However, in many situations they lack the predictive power of other models. The aggregation of several regression trees of low complexity has been proposed as a strategy to improve the model accuracy. Bootstrap aggregated (bagged) trees are an example of such strategy. Bagging is a general clustering technique that can be applied to combine regression or classification models [28]. Briefly, in bagged trees, a large number of tree-based models are fitted to subsamples of the training data. The fitted models are considered as a bootstrap sample of the "actual" model that correctly best fits the data (among those that can be described as a combination of tree-based models). Predictions are calculated as a combination of the predictions made by each of the tree-based models fitted.

Bagged trees, although superior to simple tree-based models, have the limitation that the models fitted are not uncorrelated. If one regressor has high importance in the classification, it will be included in every model. The random forest algorithm is a modification of bagged trees, where features used for the construction of the tree are selected randomly [29]. This results in trees with lower correlation. The implementation of the random forest algorithm included in the *ranger* R package [30] has been used in this analysis.

2.3.3. (Stochastic) gradient boosting

Boosting applies a similar philosophy to the one used in bagging, where several weak predictors are combined. Boosting, however, applies the models recursively, so that each model improves the predictions of the previous ones. There is a large plethora of boosting techniques. In the case of gradient boosting, each model is fitted to the residuals of the previous ones [31]. Then, prediction are calculated as the combination of the model predictions weighted by the mean difference between the model predictions and the observations on each step. In this study we have applied stochastic gradient boosting. This algorithm is a modification of the gradient boosting algorithm that increases efficiency [32]. It does so by using a subsample of the training set on each iteration instead of the complete set, reducing the correlation of the models.

2.3.4. Regularization: ridge regression, the lasso, elastic net

Regularization is an extension of linear regression aiming to reduce overfitting. In linear regression, parameter estimates are calculated as those that minimize the RMSE of the model predictions. Regularization introduces a regularization function $\varepsilon(\lambda, \beta)$ that penalizes models with many model parameters, mitigating overfitting (Eq. (2)).

$$\max_{\beta_i}(\text{RMSE}+\varepsilon(\lambda, \beta)) \tag{2}$$

The most popular algorithms for regularization are ridge regression, the lasso and the elastic net. Each one of them proposes a different form for the regularization function. Ridge regression, proposes as regularization function the sum of the square of the coefficients ($\varepsilon(\lambda, \beta)=\lambda\cdot\sum_i \beta_i^2$), whereas lasso regression uses the sum of the absolute values ($\varepsilon(\lambda, \beta)=\lambda\cdot\sum_i |\beta_i|$). Therefore, both algorithms favor models with parameters with low absolute values. The penalty introduced in ridge regression tends to reduce the absolute value of every model parameters, whereas the lasso tends to set model parameters to zero. The parameter λ is a weight for the penalty function that controls the bias-variance trade-off. A value of $\lambda = 0$ results in the same model as the linear regression model, whereas higher values will increase the relevance of the penalty, reducing variance.

The elastic net is a combination of ridge regression and the lasso. In this model, the penalty function is the sum of the penalty functions proposed by ridge and lasso ($\varepsilon(\lambda, \beta) = \lambda_1\cdot\sum_i \beta_i^2 + \lambda_2\cdot\sum_i |\beta_i|$). With this, the elastic net model aims at combining the advantages of both the ridge regression and the lasso. This model has two λparameters (λ_1 and λ_2), each one weighting a penalty function.

2.3.5. Spline regression

Spline regression is a non-parametric regression technique. In this study, multivariate adaptive regression splines (MARS) have been applied, using the implementation included in the *earth* R package. This method was proposed by Friedman [33] for the analysis of non-linear multidimensional data. Briefly, MARS describes the data using splines as basis functions. It is based on a recursive partition of the design space, which defines the knot position and product degree of the splines.

2.4. Model training, validation and testing

The dataset analyzed has been randomly divided in a training and a test set. A 70% of the observations have been included in the training set, which has been used for parameter estimation and validation. In order to minimize overfitting, *k*-fold cross-validation has been applied [20]. The training set is divided in *k* subsamples of equal size and the model parameters are estimated *k* times. On each iteration, a single subsample is retained and the remaining *k-1* subsamples are used for parameter estimation. Predictions are, then, calculated as an average of the *k* estimations. In this study, 10-fold cross-validation has been applied.

The remaining observations (30%) have been set apart for testing, and have been used for evaluating the predictive accuracy of the model fitted. The accuracy has been evaluated using the Root Mean Squared Error (*RMSE*), the coefficient of determination R^2 and the Mean Absolute Error (*MAE*). Because these observations have not been included for parameter estimation, the accuracy of the model predictions are an estimate of its accuracy in actual industrial settings.

2.5. Outlier detection

The dataset analyzed comprises ad-hoc data gathered during production, where outliers are common [34]. They can be related to human error in the input of data or errors during operation that passed unnoticed to the plant managers. Outliers can have a big, negative, impact on regression models, so they have been removed before training the regression models. The 1.5*IQR rule has been applied to the response variable (the ratio between the output and the input weight), which states that data points further than 1.5 times the interquartile range (IQR) from the 0.25 and 0.75 quantiles are outliers [35].

3. Results

The dataset comprises 1795 batches, gathered in the timespan of 10 months. The average production loss in the dataset is 0.8 tons, with a standard deviation of 0.69 tons. In 40 of the records (2.5% of the total) the production loss is negative, indicating a final production below the ordered amount (underproduction). This is especially undesirable for the producer, due to the associated reputation damage and added economical costs. Fig. 1 illustrates the high uncertainty of the production loss with respect to the planned production. Although a positive correlation can be observed, this trend is not strong and the data has high scatter. Therefore, previous knowledge cannot be easily incorporated in the predictions, and decision must be made in an environment with high uncertainty.

The dataset has been analyzed using the regression models described in the Materials and Methods sections. Outliers usually have a detrimental effect on the accuracy of these models, so they have been identified according to the 1.5*IQR rule. A total of 90 observations (~5% of the total) were labeled as outliers and omitted from the analysis. Table 1 reports the *RMSE, MAE* and R^2 calculated for the training and test sets for each predictive model. For every predictive model tested, the R^2 of the training set was greater than the one of the test set, with the exception of the regression tree. This can be attributed to the overfitting of the training set, inevitable when ML algorithms are applied [36]. Therefore, it is essential for a correct evaluation of the actual accuracy of empirical models the use of a dataset that was not involved in parameter estimation.

The linear regression model with AIC parameter selection has higher R^2 than the elastic net and the lasso regression in both the training and test sets. The values of R^2 calculated for ridge regression are very similar (<10% difference) to those calculated for the linear regression model. The remaining ML models applied have higher R^2 in both the training and the test sets than those calculated for the linear

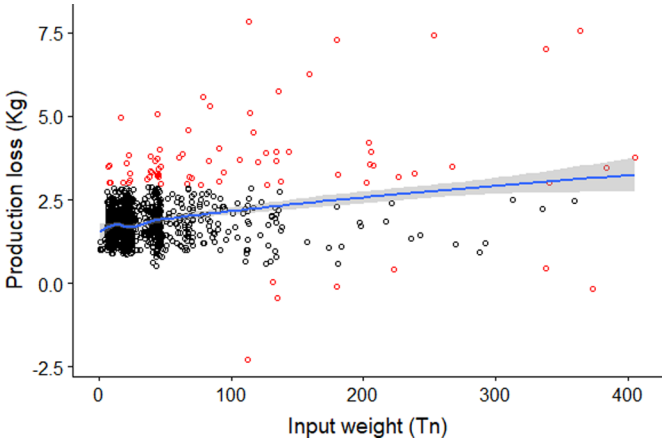

Fig. 1. Scatter plot of the relationship between the input weight and the production loss. Outliers identified according to the 1.5*IQR rule are marked in red. The solid line is a model fitted using local polynomial regression.

Table 1
RMSE, MAE and R^2 calculated for the training and test sets for each predictive model.

	Training			Test		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Linear model	0.018	0.014	0.442	0.019	0.015	0.421
Regression tree	0.016	0.012	0.553	0.016	0.013	0.586
Bagged tree	0.016	0.012	0.599	0.016	0.013	0.577
Random forest	0.014	0.010	0.705	0.016	0.012	0.596
Gradient boosting	0.014	0.011	0.654	0.016	0.012	0.623
Lasso	0.019	0.014	0.425	0.020	0.015	0.401
Ridge regression	0.018	0.014	0.448	0.019	0.015	0.426
Elastic net	0.019	0.014	0.428	0.020	0.015	0.404
Spline	0.016	0.012	0.591	0.016	0.012	0.594

model. The tree-based models clearly outperform the linear model in both the training and test sets. The model with the highest precision among those applied in this research is the gradient boosting algorithm ($R^2 = 0.654$ for the training test, $R^2 = 0.623$ for the test set). Spline regression has a precision close to the one of Random Forest, with $R^2 = 0.594$ in the test set.

The lasso, ridge regression and the elastic net are modifications of linear regression that introduce an additional term in the cost function to penalize models with have a very high number of parameters. As a consequence, the model parameters are shrunk towards zero. The linear regression model with AIC parameter selection also aims to reduce the number of model parameters. This is done iteratively, eliminating or adding one model parameter in the direction that increases the AIC. In that sense, regularization (lasso, ridge regression and elastic net) has several similarities with the AIC model selection. This can explain the similar result obtained for these algorithms.

On the other hand, tree-based models and spline regression follow an entirely different modeling approach to the one used by the linear model. The selection of the relevant features is not done according to the absolute values of the coefficients of a linear model. Furthermore, these models introduce a non-linear relationship between the regressors and the response. In the case studied here, this modeling approach is more precise than the linear model. This may indicate that the bias of the linear model is not related to a model overfitting (over-parameterization). Instead, there is a non-linear relationship between the output variable (the fraction of food loss) and the predictors that can only be described using the ML models. Although the non-linearity could be introduced in the linear model using polynomial regression and/or variable transformations, this would require to fit and compare different model formulations. ML algorithms provide a more convenient

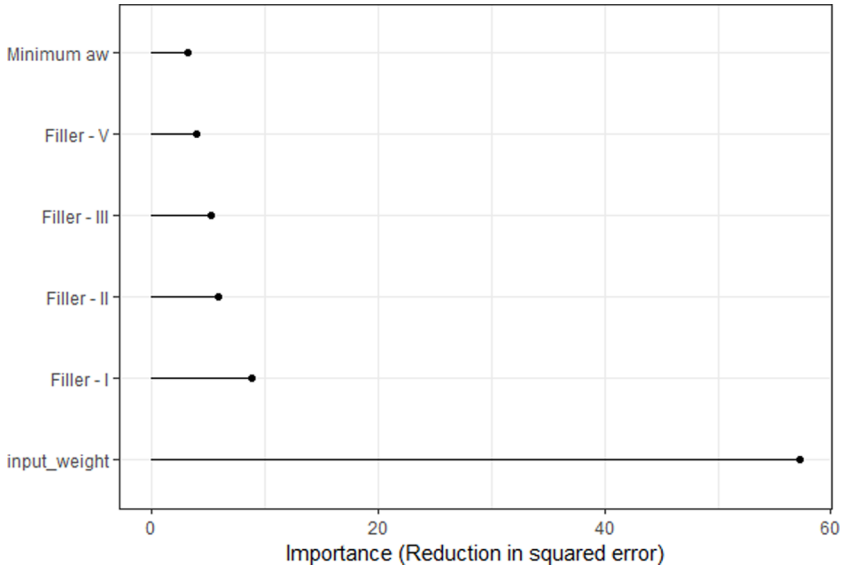

Fig. 2. Variable importance plot of the gradient boosting model.

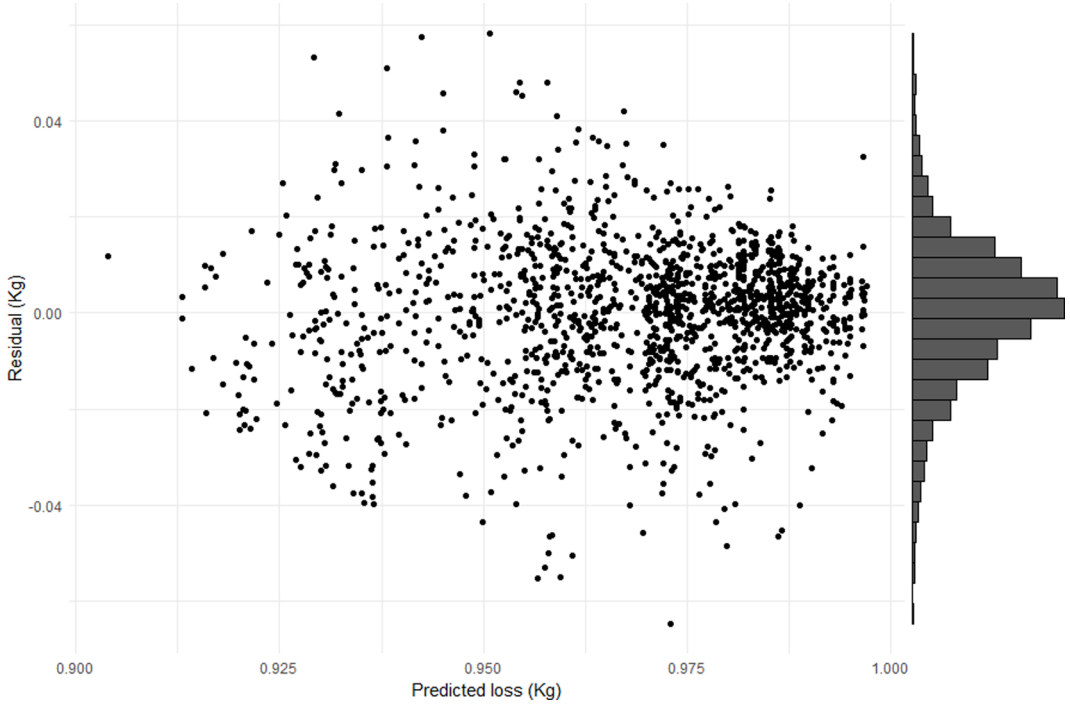

Fig. 3. Scatter plot of the residuals of the gradient boosting model. The right hand axis shows a histogram of the residuals.

approach, where the non-linear relationship is learnt from the data, without the requirement from the user to define additional hypotheses to describe such relationships.

The importance of the model variables most relevant for the gradient boosting model is illustrated in Fig. 2 using a variable importance plot. The total weight of the raw materials input is the variable with the highest importance. The configuration of the production line is largely dependent on the input weight (e.g. smaller batches require fewer tanks). Furthermore, due to the viscosity of the product, there is a loss in the piping of the installation (e.g. in valves and shoulders) whose contribution to the fraction of waste is larger for small batches. Therefore, the strong relationship between this predictor and the proportion of food waste is expected for the system analyzed.

The variable with the 6th highest importance for the gradient boosting model is the minimum water activity of the ingredients used in the recipe. The company analyzed uses a broad range of ingredients, with different physical properties. It is unlikely that the water activity of the ingredients directly affects food loss. However, it seems to be a valuable instrumental variable [37], that comprises products whose physical properties have a significant impact on the food loss of the final product.

The filler selected also has a very strong impact on the amount of food waste, according to the gradient boosting algorithm. This result was unexpected, because the characteristics of the fillers used in the production plant are identical (technical data not disclosed due to confidentiality), so the selection of the filler should not have a major impact on the production. Furthermore, the fillers are not selected according to the recipe. Instead, they are rotated, trying to balance production load between them. Hence, it is unlikely that the filler serves as an instrumental variable, in a similar way to the minimum water

activity of the ingredients. In light of these results, the quality managers of the plant planned a detailed audit of the component, which pinpointed an error in the calibration of one of the fillers.

Besides identifying the variables most relevant for the amount of production loss the gradient boosting model fitted to the data can be used to predict the amount of food loss. This information would be valuable for production managers in the company, as it would improve production planning and waste management. However, the predictions of the gradient boosting model do not include the uncertainty in production loss. Fig. 3 depicts the residuals of the gradient boosting model for the training set. They are normally distributed ($p < 0.05$; Shapiro-Wilk test) with zero mean and standard deviation of 0.015 Tn, as well as homoscedastic. This information can be incorporated as an error model [38] in the predictions of the gradient boosting algorithm using a normal distribution, thus, including uncertainty in the model predictions.

Among all the model predictors, the total input of raw material is the easier to manipulate during production planning. Fig. 4 illustrates the predicted food loss for two different recipes as a function of the input weight. The dashed line represents the expected loss, whereas the shaded areas indicate prediction intervals constructed at the 90% confidence level. This plot shows that the expected amount of food waste varies between recipes. Although there is a positive correlation between both variables (as already pointed out based on Fig. 1), the relationship between both variables is far from linear. The non-linearity of this relationship can be attributed to the complexity of the production line analyzed in this study. Increasing the production requires several changes in the installation (different tanks, connections…) that have a strong influence in the amount of food loss in a complex way.

Moreover, Fig. 4 illustrates that the uncertainty (width of the prediction interval) is higher for large batches than for smaller ones. The reason for this is that the model has been fitted to the fraction of food waste, not to its magnitude. Then, the calculation of the food loss requires a non-linear transformation that increases the variance of the response. This result is also reasonable from a practical point of view. Large batches are more complex and have higher uncertainty in the amount of food waste. Therefore, the model can be used to predict the amount of food waste (including uncertainty) for different production settings, and to adjust the production and waste management plans accordingly.

4. Discussion

The value of data has grown during the last years, gaining a central role for modern societies [18]. The development of (economic) sensors and new technologies to store and analyze that data has been named by many authors as the “new industrial revolution”, leading to the so-called Industry 4.0 [39,40]. In the context of food production, the application of technologies based on data can achieve improvements that involve every relevant stakeholder, from primary producers to consumers [41,42]. These technologies can support smart manufacturing through the whole product lifecycle [43], as well as provide new insights of every step of the food chain, providing a better understanding of each step from primary production to consumption [44,45]. At a larger scale, technologies based on Big data can revolutionize industrial production in practically every sector [46,47]. As an example, a system based on Big Data has already been implemented by the China Food and Drug Administration for food safety management [48]. Recent examples of the application of big data in food production include the use of radio-frequency identification (RFID) sensors to support food manufacturing [49–51], supply chain management [52,53].

This research work has illustrated that the analysis of historical data can provide added value to the food industry. Nevertheless, this data is usually of high dimensionality and complexity, limiting the application of classical statistical models. Hence, in order for the data to be useful for industries, they must be analyzed using advanced data analysis tools. In the case study analyzed in this research, we have shown how ML algorithms provide estimates of parameters of the production process that are more reliable than those obtained using classical statistical models. The application of these ML models could aid production planning of the analyzed industry, reducing both economic cost and food waste.

Mismatches between the weight of the final product and the costumer order are a big concern for the analyzed food industry. They include, among others, transportation costs, holding costs, warehousing costs, shortage costs and obsolescence costs [5]. Furthermore, due to the perishable nature of the product, overproductions usually become food waste, which has an associated environmental impact. For these

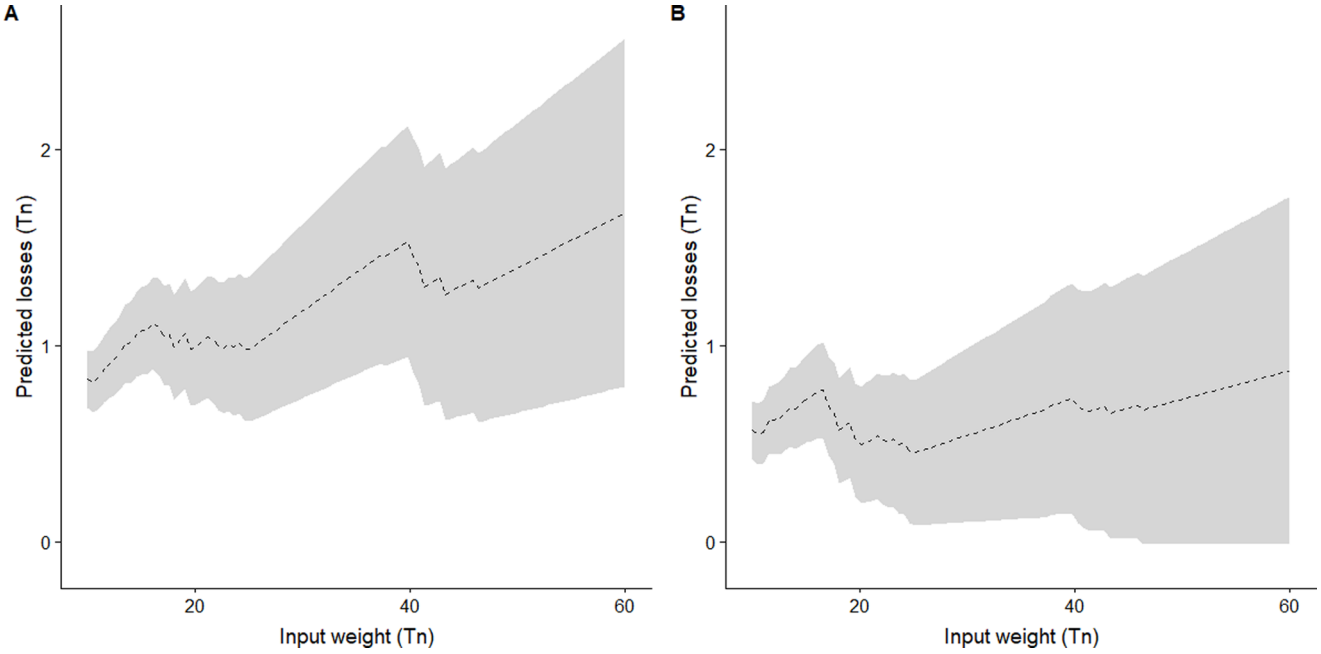

Fig. 4. Predicted food loss as a function of the input weight for two different recipes (differentiated by colours). The dashed line represents the model prediction of the gradient boosting model and the dashed area the 90% confidence interval.

reasons, accurate prediction tools are of great interest for the managers of this industry. Because of the characteristics of this industry, production cannot be halted once it has started. Therefore, it has to be planned only based on prior information, without any possible feedback during production. Before this study was conducted, due to the complexity of the process, production planning is carried out according to only two variables: the number of tanks to be used and the final packaging format. Consequently, decisions were made in an environment with high uncertainty and, due to the economical and image cost of overproductions, production was planned following an overly conservative approach that resulted in overstock that became food waste. The advanced data analytics tools explored in this work provide a probabilistic description of the amount of food loss, reducing uncertainty. This enables a more rational production planning that corrects some of the biases of classical models [16]. This, in turn, can potentially reduce the risks associated to underproduction, as well as reduce the environmental impact of the company.

It can be argued that the continuous prediction of the ML models developed in this study (Fig. 4) would hardly be used under actual settings. Instead, a qualitative response variable (e.g. low/intermediate/high waste) could be more easily understood and applied by the relevant stakeholders of a food industry. Nevertheless, the continuous output variable of the model can easily be transformed to a categorical one. For instance, one could define that losses lower than 1 Tn are defined as “Low waste”, between 1 and 2 Tn as “Intermediate waste” and higher than 3 Tn as “High waste”. Then, production could be planned according to these limits in a similar fashion as was shown in the results section. If, on the other hand, the models had been trained and validated using a categorical variable, they could not adapt to changes in the categories (e.g. number of levels and/or limits). This modeling approach would require the whole modeling process to begin from scratch. Furthermore, the uncertainty of the model predictions is very relevant in this study (Fig. 4), as is common in food industries [54]. The use of a continuous output variable enables to model uncertainty using a continuous error model. If, instead, we had used a categorical variable, the model for uncertainty would be more complex and less informative than the continuous one.

One of the main arguments against the use of ML models is model interpretation. In linear models, the importance of each one of the input variables can be assessed by evaluating and comparing the model coefficients. Some simple ML algorithms (e.g. regression trees) are easy to interpret in a similar fashion, but this is not the case for more complex algorithms that are usually applied as “black boxes” [55,56]. Due to the high predictive power of the more complex ML algorithms, several research efforts have been dedicated towards developing algorithms to better understand them. An example of such algorithm is the estimation of the “variable importance”, a parameter with an interpretation similar to the parameter values in a linear model without interactions [57]. It can be visualized as a “Variable Importance Plot” like the one depicted in Fig. 2. Indeed, this tool was used to identify a deficiency in one of the filler machines, extending the application of the model beyond its predictive power. As already mentioned, the unexpected high importance of this variable led the managers of the production plant to audit these components in detail, identifying a calibration error in one of them. Therefore, the application of Big Data and ML does not only reduce the uncertainty of production planning, but also provides valuable information about the performance of individual components, that can be used to detect deficiencies and support maintenance. Hence, it provides valuable information for all the stakeholders involved in production management and planning. Furthermore, this information can be incorporated to quality assurance programs, as a detection method of deviations from standard production.

5. Conclusions

This study illustrates the added value that the application of advanced analysis to historical data can bring to the food industry. ML methods have provided valuable information, outperforming classical statistical methods for predicting the amount of food waste. Moreover, we have included a stochastic model to describe the uncertainty of the model predictions. Besides their predictive power, ML models can provide insight into the production system, opening ways for future improvement policies (e.g. detection of anomalies). Therefore, we have illustrated that the application of ML to food industries can reduce production uncertainty, which may result in improved customer service, increased profitability and reduced wastes and CO_2 emissions.

CRediT authorship contribution statement

Alberto Garre: Software, Formal analysis, Visualization. **Mari Carmen Ruiz:** Conceptualization, Data curation, Supervision. **Eloy Hontoria:** Conceptualization, Data curation, Supervision, Investigation.

Declaration of Competing Interest

None.

Acknowledgments

Eloy Hontoria is grateful to Project RTI2018-099139-B-C21 financed by FEDER/Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación. Alberto Garre (20900/PD/18) is grateful to the Seneca foundation for awarding him a post-doctoral grant.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.orp.2020.100147.

References

[1] Gustavsson J, Cederberg C, Sonesson U, Van Otterdijk R, Meybeck A. Global food losses and food waste. FAO Rome 2011.
[2] Basher SA, Raboy D, Kaitibie S, Hossain I. Understanding challenges to food security in dry arab micro-states: evidence from qatari micro-data. Rochester: NY: Social Science Research Network; 2013.
[3] Irani Z, Sharif AM. Sustainable food security futures: perspectives on food waste and information across the food supply chain. J Enterp Inf Manag 2016;29. https://doi.org/10.1108/JEIM-12-2015-0117. 171–8.
[4] Barad M, Braha D. Control limits for multi-stage manufacturing processes with binomial yield (Single and multiple production runs). J Oper Res Soc 1996;47:98–112. https://doi.org/10.1057/jors.1996.9.
[5] New C, Mapes J. MRP with high uncertain yield losses. J Oper Manag 1984;4:315–30. https://doi.org/10.1016/0272-6963(84)90019-6.
[6] Choi S, Jeon S, Kim J, Park K. A newsvendor analysis of a binomial yield production process. Eur J Oper Res 2019;273:983–91. https://doi.org/10.1016/j.ejor.2018.09.029.
[7] Lieder M, Rashid A. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J Clean Prod 2016;115:36–51. https://doi.org/10.1016/j.jclepro.2015.12.042.
[8] Mirabella N, Castellani V, Sala S. Current options for the valorization of food manufacturing waste: a review. J Clean Prod 2014;65:28–41. https://doi.org/10.1016/j.jclepro.2013.10.051.
[9] Bonney M, Jaber MY. Environmentally responsible inventory models: non-classical models for a non-classical era. Int J Prod Econ 2011;133:43–53.
[10] Wahab MIM, Mamun SMH, Ongkunaruk P. EOQ models for a coordinated two-level international supply chain considering imperfect items and environmental impact. Int J Prod Econ 2011;134:151–8. https://doi.org/10.1016/j.ijpe.2011.06.008.
[11] Mula J, Poler R, García-Sabater JP, Lario FC. Models for production planning under uncertainty: a review. Int J Prod Econ 2006;103:271–85.
[12] Yano CA, Lee HL. Lot sizing with random yields: a review. Oper Res 1995;43:311–34. https://doi.org/10.1287/opre.43.2.311.
[13] Hegseth MA. The challenge of operation yield. Prod Inventory Manag 1984;25:4–10.
[14] Graves SC. Uncertainty and production planning. editors In: Kempf KG, Keskinocak P, Uzsoy R, editors. Plan. prod. invent. ext. enterp. state art handb, 1. Boston, MAUS: Springer; 2011. p. 83–101. https://doi.org/10.1007/978-1-4419-6485-4_5.
[15] Silver EA. A simple method of determining order quantities in joint replenishments under deterministic demand. Manag Sci 1976;22:1351–61. https://doi.org/10.1287/mnsc.22.12.1351.
[16] Efendigil T, Önüt S, Kahraman C. A decision support system for demand forecasting

with artificial neural networks and neuro-fuzzy models: a comparative analysis. Expert Syst Appl 2009;36:6697–707. https://doi.org/10.1016/j.eswa.2008.08.058.
[17] Azizi A., Ping L.W. Production throughput modeling under five uncertain variables using bayesian inference, 2012.
[18] Gobble MM. Big Data. The next big thing in innovation. Res-Technol Manag 2013;56. https://doi.org/10.5437/08956308x5601005. 64–7.
[19] Murphy KP. Machine learning: a probabilistic perspective. edición: 1. Cambridge, MA: MIT Press Ltd; 2012.
[20] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications. In: Edición R, editor. 1st edSpringer; 2013. 2013, Corr. 6th printing 2016.
[21] Gandhi N, Armstrong LJ, Petkar O, Tripathy AK. Rice crop yield prediction in India using support vector machines. 2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE, Khon Kaen: IEEE 2016:1–5. https://doi.org/10.1109/JCSSE.2016.7748856.
[22] Guo H, Jeong K, Lim J, Jo J, Kim YM, Park J, et al. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. J Environ Sci 2015;32:90–101. https://doi.org/10.1016/j.jes.2015.01.007.
[23] Grainger MJ, Aramyan L, Piras S, Quested TE, Righi S, Setti M, et al. Model selection and averaging in the assessment of the drivers of household food waste to reduce the probability of false positives. PLoS ONE 2018;13:e0192075https://doi.org/10.1371/journal.pone.0192075.
[24] R. Core Team. R: a language and environment for statistical computing. Vienna: Austria: R Foundation for Statistical Computing; 2016.
[25] Wing MKC from J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., et al. Caret: classification and regression training. 2018.
[26] Draper NR. Smith H. applied regression analysis. New York: Wiley-Blackwell; 1998. 3rd Edition. 3Rev Ed edition.
[27] Venables W.N., Ripley B.D.Modern applied statistics with s. 4th ed. New York: Springer-Verlag; 2002.
[28] Leisch F. Bagged clustering1999.
[29] Breiman L. Random forests. mach learn2001;45:5–32. 10.1023/A:1010933404324.
[30] Wright MN, Ranger ZA. A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 2017;77:1–17. https://doi.org/10.18637/jss.v077.i01.
[31] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–232.
[32] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367–78. https://doi.org/10.1016/s0167-9473(01)00065-2.
[33] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991:1–67.
[34] Maddala GS. Introduction to econometrics. 2nd ed New YorkToronto: Macmillan Pub. Co.; 1992. New YorkMaxwell Macmillan Canada ; Maxwell Macmillan International.
[35] Upton GJG, Cook I. Understanding statistics. Oxford: Oxford University Press; 1996.
[36] Box GEP, Hunter JS, Hunter WG. Statistics for experimenters: design, innovation, and discovery. Hoboken, N.J: Wiley-Blackwell; 2005.
[37] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91:444–55. https://doi.org/10.1080/01621459.1996.10476902.
[38] Goodchild MF, Guoqing S, Shiren Y. Development and test of an error model for categorical data. Int J Geogr Inf Syst 1992;6:87–103. https://doi.org/10.1080/02693799208901898.
[39] Kumar R, Singh SP, Lamba K. Sustainable robust layout using big data approach: a key towards industry 4.0. J Clean Prod 2018;204:643–59. https://doi.org/10.1016/j.jclepro.2018.08.327.
[40] Lee J, Kao H-A, Yang S. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia CIRP 2014;16:3–8. https://doi.org/10.1016/j.procir.2014.02.001.
[41] Bronson K, Knezevic I. Big data in food and agriculture. Big Data Soc 2016;3. https://doi.org/10.1177/2053951716648174. 2053951716648174.
[42] Xu F, Li Y, Feng L. The influence of big data system for used product management on manufacturing–remanufacturing operations. J Clean Prod 2019;209:782–94. https://doi.org/10.1016/j.jclepro.2018.10.240.
[43] Ren S, Zhang Y, Liu Y, Sakao T, Huisingh D, Almeida CMVB. A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: a framework, challenges and future research directions. J Clean Prod 2019;210:1343–65. https://doi.org/10.1016/j.jclepro.2018.11.025.
[44] King T, Cole M, Farber JM, Eisenbrand G, Zabaras D, Fox EM, et al. Food safety for food security: relationship between global megatrends and developments in food safety. Trends Food Sci Technol 2017;68:160–75. https://doi.org/10.1016/j.tifs.2017.08.014.
[45] Wolfert S, Ge L, Verdouw C, Bogaardt M-J. Big data in smart farming – A review. Agric Syst 2017;153:69–80. https://doi.org/10.1016/j.agsy.2017.01.023.
[46] Fosso Wamba S, Akter S, Edwards A, Chopin G, Gnanzou D. How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. Int J Prod Econ 2015;165:234–46. https://doi.org/10.1016/j.ijpe.2014.12.031.
[47] Garre A, Fernandez PS, Brereton P, Elliott C, Mojtahed V. The use of trade data to predict the source and spread of food safety outbreaks: an innovative mathematical modelling approach. Food Res Int 2019;123:712–21. https://doi.org/10.1016/j.foodres.2019.06.007.
[48] Chen K, Tan H, Gao J, Lu Y. Big data based design of food safety cloud platform. Appl Mech Mater 2014:536–7. https://doi.org/10.4028/www.scientific.net/AMM.536-537.583. 583–7.
[49] Fosso Wamba S. Achieving supply chain integration using RFID technology: the case of emerging intelligent B-to-B e-commerce processes in a living laboratory. Bus Process Manag J 2012;18:58–81. https://doi.org/10.1108/14637151211215019.
[50] Ngai EWT, Chau DCK, Poon JKL, Chan AYM, Chan BCM, Wu WWS. Implementing an RFID-based manufacturing process management system: lessons learned and success factors. J Eng Technol Manag 2012;29. https://doi.org/10.1016/j.jengtecman.2011.09.009. 112–30.
[51] Wamba SF, Chatfield AT. A contingency model for creating value from RFID supply chain network projects in logistics and manufacturing environments. Eur J Inf Syst 2009;18:615–36. https://doi.org/10.1057/ejis.2009.44.
[52] Kaur H, Singh SP. Heuristic modeling for sustainable procurement and logistics in a supply chain using big data. Comput Oper Res 2018;98:301–21. https://doi.org/10.1016/j.cor.2017.05.008.
[53] Tan KH, Zhan Y, Ji G, Ye F, Chang C. Harvesting big data to enhance supply chain innovation capabilities: an analytic infrastructure based on deduction graph. Int J Prod Econ 2015;165:223–33. https://doi.org/10.1016/j.ijpe.2014.12.034.
[54] Thompson KM. Variability and uncertainty meet risk management and risk communication. Risk Anal 2002;22:647–54. https://doi.org/10.1111/0272-4332.00044.
[55] Molnar C. Interpretable machine learning. christoph molnar; 2019.
[56] Alvarez-Melis D., Jaakkola T.S.On the robustness of interpretability methods. ArXiv180608049 Cs Stat2018.
[57] Nathans LL, Oswald FL, Nimon K. Interpreting multiple linear regression: a guidebook of variable importance. Pract Assess Res Eval 2012;17.