Unsupervised Learning and Data Mining Approach for Fashion Employees's Productivity

D. Sierra-Porta[1,a,*]

[1] *Facultad de Ciencias Básicas. Universidad Tecnológica de Bolivar, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias, 130010, Colombia.*
[a] *D. Sierra-Porta (orcid=0000-0003-3461-1347)*
[*] *Corresponding author: dporta@utb.edu.co*

October 31, 2022

Abstract

The apparel industry is one of the most profitable industries worldwide. Until the end of the 1970s, companies engaged in the apparel business produced clothing in Europe and North America. These production units performed all the steps for manufacturing clothes, from cutting the fabric to finishing. Many workers were hired for production, and the companies were responsible for their wages, social security, and working conditions. Regardless of the market globalization and the international demand, the industry relies on human capital (employees) to ensure product effectiveness and efficiency and to achieve industrial objectives. Because of the dependence on manual labor, the production of a garment company depends on the productivity of employees working in various departments of the company. When employees do not meet the company's management objectives, some of the links in the production chain fail, thereby negatively affecting the quality and efficiency of the company. By analyzing data from a major garment company in Bangladesh, we created a predictive model for employee productivity in terms of various variables involved in the employee labor process. Data mining was used for data manipulation and cleaning, whereas random forest, gradient boosting, and extreme gradient boosting were applied for predictability estimates. The extreme gradient boosting model proved to be the most efficient in predicting employee productivity with a mean absolute error of 0.0174, root mean square error of 0.0497, mean absolute percentage error of 2.6602, correlation coefficient of 0.96 and Accuracy(%) of 97.34, between original and predicted data. These values are considerably lower than those previously reported in the literature. The proposed model is an important tool for decision makers to evaluate the actions to be taken by a company for profit maximization when certain variables are known to yield a certain performance.

Keywords: Machine Learning, Data Mining, Labor market productivity.

JEL codes: **E24:** Employment - Unemployment - Aggregate Labor Productivity. **F66:** Labor. **J24:** Human Capital - Skills - Occupational Choice - Labor Productivity. **C14:** Semiparametric and Nonparametric Methods. **C53:** Forecasting and Prediction Methods. **C55:** Large Data Sets: Modeling and Analysis. **C45:** Neural Networks and Related Topics.

1 Introduction

The garment industry presents some of the most significant opportunities and challenges worldwide because it is constantly changing. It is one of the largest industries on the planet and comprises designers, textile producers, models, retailers, and consumers, who have, as their only common element, any product for personal use, such as clothing, footwear, or accessories. The garment industry developed with capitalism [Brodkin, 2000, Mezzadri, 2010, Shamsuzzoha et al., 2013]; the companies involved in this industry can easily move their production centers to places where labor is cheap. As capitalism has developed, major changes have occurred in trade. This causes competition and globalization at the international level, which is technologically more competitive.

Manual work in this industry is fundamental for clothing manufacture, even at present. Currently, new technologies take precedence to guarantee the effectiveness and efficiency of the work [Yunus and Yamagata, 2012, Mark et al., 2021].

Manual work is still important for creating different garments. Although at present, we entrust everything to machinery, the mechanism is not perfect, and we can only create garments that we visualize by using hands. Many garments are marketed in this manner. This is becoming more prevalent, owing to the professionals working in this sector.

The textile industry is one of the most powerful sectors of the economy. Specifically, in terms of the cash flow of several countries, the industry generally accounts for at least 400 million euros each year; thus, more people depend on its activities, such as fashion [Beltrami et al., 2019, Cabigiosu, 2020, Beauloye, 2019]. However, other international markets report different astronomical values. Spain's fashion industry has a turnover of 14,688 million euros. This also includes the number of companies engaged in this activity, which was 19,470 in Spain. The fashion industry employs 300 million workers worldwide [Beauloye, 2019, Gazzola et al., 2020]. The latter is significant in terms of the importance of fashion in these countries.

Manual labor continues to be fundamental in clothing production. Although technology is employed in a large part of the production processes of this business, many people employed in this sector are required to use manual skills (for example, when sewing) to make products in precisely the same way they had imagined.

Companies in the fashion sector are increasingly demanding professionals who have design and decision-making skills that guarantee business sustainability over time and the optimisation of its profits. Although fashion is crucial for the industry because it is a differential factor, future professionals and entrepreneurs must have other skills pertaining to production, communication, and marketing strategies such that creative and beautiful products successfully reach the market.

Given this critical scenario, many studies have been conducted to analyze, in terms of data mining and artificial intelligence(AI), how the discipline can help address some of the main problems, processes, and methodologies of the industry to improve its understanding and add value to its knowledge and predictability [Imran et al., 2021] using prediction techniques [Al Imran et al., 2019, Balla et al., 2021, Dogan and Birant, 2021, Goyzueta et al., 2021, Obiedat and Toubasi, 2022].

The data mining process in the presence of missing data is particularly important; researchers and decision-makers regularly deal with this. Having a complete data archive is ideal; however, applying inappropriate imputation methods to achieve this can create more problems than solutions.

Over the last few decades, procedures that have better statistical properties than traditional options have been developed, such as listwise, pairwise, mean method, and hot-deck. Algorithms based on machine learning have been important and widely accepted in the recent decades [Lakshminarayan et al., 1996, Lakshminarayan et al., 1999, Jadhav et al., 2019]. Their implications in secondary data analysis should be evaluated with caution, particularly when other methodologies are subsequently applied for analysis or to establish prediction or classification strategies. Each situation is different, and the non-response rate and its spatial distribution change between surveys; thus, adopting the same imputation procedure for all variables in all surveys is not convenient.

In this study, our aim is to use certain machine learning tools mainly for generating models that can predict the behaviour and predictability of the productivity of employees in the garment industry. In our study, we used a methodology similar to the previous work but refined the data mining strategy and used a different algorithm for the results. Our results are comparable but more optimal than those obtained by Balla et al. [Balla et al., 2021], and we found an improvement over the results obtained by Imran et al. [Imran et al., 2021]. We used and compared three different regressors: random forest [Biau and Scornet, 2016, Shi and Horvath, 2006, Belgiu and Drăguţ, 2016], gradient boosting [Natekin and Knoll, 2013, Bentéjac et al., 2021], and extreme gradient boosting [Chen et al., 2015]. Selecting the best model to use for this particular data was weighted based on several error metrics, and the effect that the data imputation process can have was compared. A strategy was employed to convert variables into dummy variables or categorise them using numerical values. The following sections present information on the data used and data manipulation strategies. Finally, we present the results and conclusions derived from our findings.

2 Data description and methodology

2.1 Description of dataset used

In this study, we used a dataset called garments_worker_productivity, which was initially constructed from a survey conducted by the engineering department of a major garment company in Bangladesh. The principal reference for this dataset is found in [Imran et al., 2021, Al Imran et al., 2019] and available from UCI Machine Learning (Center for Machine Learning and Intelligent Systems) repository at https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees. This dataset includes essential attributes of the garment manufacturing process and employee productivity, collected manually and validated by industry experts.

The dataset contains 1197 instances and 15 attributes. One of the attributes has the employee evaluation date and information collected from January 2015 to March 2015. A complete description of these attributes is available in [Al Imran et al., 2019], each attribute is described in table 1 as follows:

Attribute	Description
Date	The date is in MM-DD-YYYY format
Day	Days of the week
Quarter	Part of this month. One month is divided into four parts
Department	The department is associated with the instance
team_no	The team number associated with the instance
no_of_workers	The number of workers on each team
no_of_style_change	The number of changes to a specific product style
targeted_productivity	The targeted productivity is set by the Authority for each team for each day
SMV	Standard Minute Value, this is the time allocated for a task
wip	Work in Progress includes the number of unfinished items for the product
over_time	Represents the amount of overtime by each team in minutes
incentive	Represents the number of financial incentives (in the UDB) that enable or motivate certain actions
idle_time	The length of time the product has stalled for several reasons
idle_men	The number of unemployed workers due to production disruptions
actual_productivity	The actual percentage of productivity generated by workers. It ranges from 0-1.

Table 1: Dataset Description.

2.2 Engineering data mining

One of the most important obstacles in this dataset is a variable called work in progress (wip), which includes the number of unfinished products). This variable contains 506 instances of missing data, which corresponds to 42.27% of the total data. This is a significant obstacle to developing AI methodologies for addressing problems and building results.

One of the strategies implemented is the consolidation of missing information. Here, we considered the application of the iterative imputation [Zhang et al., 2010] method but using Random Forest classifier [Pantanowitz and Marwala, 2009], widely used in data mining processes, as well as nearest neighbor imputation (KNN) [Beretta and Santaniello, 2016]. Although we used both techniques, the final results were very similar. Of particular interest is the ability of iterative imputers to mimic the behavior of missForest [Stekhoven, 2015], a popular imputation package for R. KNN imputation, alternative, learns from samples with missing values by using a distance metric that considers missing values, rather than imputing them.

Additionally, the dataset has 11 numerical variables and four categorical variables (date, quarter, department, and day). For the latter, we implemented a coding process for these variables. First, for the variable date, we divided its information into three columns: Year, Month, and Day. We used two strategies for the remaining categorical variables.

- **strategy 1 (s_1):** convert the categorical variables to dummy variables. Dummy variables are indicators of the presence or absence of a category in a categorical variable. The usual convention dictates that

0 represents absence while 1 represents presence. The conversion of categorical variables into dummy variables leads to forming the two-dimensional binary matrix where each column represents a particular category.

- **strategy 2** (s_2): A similar process to the previous one, but we code each element of the variable with a different digit; for example, for the days Monday to Tuesday, we assign numbers from 1 to 7 in a new variable.

Regardless of whether we use Strategy 1 or Strategy 2, we finally obtain a dataset containing either 28 or 18 attributes. In the following, because both strategies 1 and 2 provide results that do not differ significantly, we refer only to the results using strategy 1. Figure 1 shows the histograms for the original WIP variable in the dataset compared with the inputted variable for the various methods described earlier.

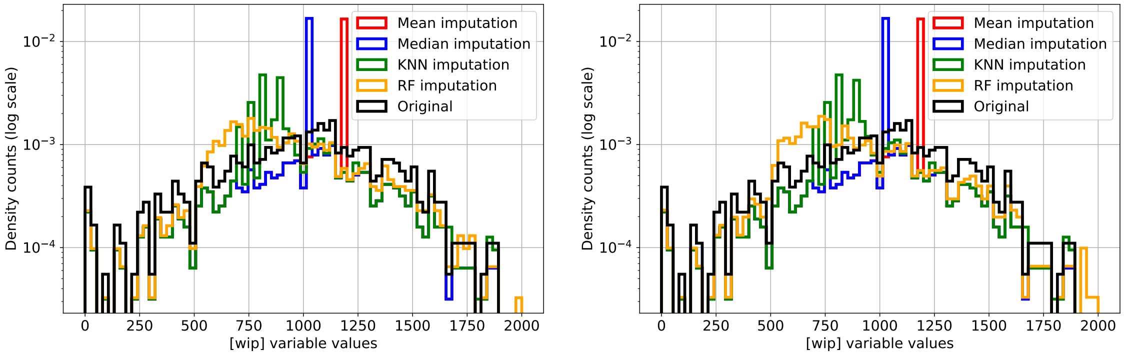

Figure 1: Original WIP variable histogram in the dataset than the imputed variable for the various methods: mean, median, Random Forest and KNN imputer. The last one produces better visual results than all the others leaving a distribution significant close to the original.

A link to the written code is provided at the end of this study for review and reproduction by anyone of interest.

2.3 Prediction algorithm

We used three different regressors to construct the model: random forest, gradient boosting, and extreme gradient boosting.

Random forest [Biau and Scornet, 2016, Belgiu and Drăguţ, 2016] is a flexible and easy-to-use machine learning algorithm that produces, even without parameter tuning, a significant result most of the time. It is the most widely used algorithm due to its simplicity and it is applicable for classification and regression tasks. Random forest is a supervised learning algorithm that, as you can already see from its name, creates a forest and makes it somewhat random. Meaning, the random forest creates multiple decision trees and combines them to obtain a more accurate and stable prediction. The more trees there are in the forest, the more robust the forest. This algorithm adds additional randomness to the model as it grows the trees. Instead of checking the most important feature when splitting a node, it looks for the best feature among a random subset of features. This results in wide diversity, which generally results in a better model. Thus, in random forests, the algorithm for splitting a node only considers a random subset of features. We can make the trees more random by using additional random thresholds for each feature instead of looking for the best possible thresholds, as in a normal decision tree.

A gradient boosting model [Natekin and Knoll, 2013, Bentéjac et al., 2021] is formed by a set of individual decision trees, trained sequentially, so that each new tree tries to improve the errors of the previous trees. The prediction of a new observation is obtained by aggregating the predictions of all individual trees that constitute the model.

Many predictive methods generate global models, where a single equation is applied to the entire sample space. When the use case involves multiple predictors that interact in a complex and nonlinear way, it is challenging to find a single global model that reflects the relationship between the variables. Tree-based statistical and machine learning methods encompass a set of nonparametric supervised techniques that manage to segment the space of predictors into simple regions, within which interactions are easier to handle. This feature gives them potential. Tree-based methods have become one of the benchmarks in the predictive field because of the good results they generate for a wide variety of problems. We explore how gradient boosting tree models are built and predicted throughout this study.

Gradient boosting is a generalization of the AdaBoost algorithm [Schapire, 2013] that allows any cost function as long as it is differentiable. The flexibility of this algorithm has made it possible to apply to boost to a multitude of problems (regression, multiple classifications, etc.), making it most successful machine-learning methods. Although there are several adaptations, the general idea of all of them is the same: to train models sequentially so that each model adjusts the residuals (errors) of the previous models. Although the boosting process (gradient boosting) achieves improved predictive capability over single-tree-based models, this has an associated cost, and the interpretability of the model is reduced. Because it combines multiple trees, a simple graphical representation of the model cannot be obtained. Visually identifying important predictors is not immediately possible.

Because of its good results, gradient boosting has become the reference algorithm for dealing with tabular data; hence, multiple implementations have been developed. Each has features that make them more suitable depending on the use case. Scikit-learn has two native implementations.

XGBoost [Chen et al., 2015] is an optimized distributed gradient boosting library that is designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the gradient boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT or GBM) that solves many data science problems quickly and accurately. The same code runs in a significant distributed environment (Hadoop, SGE, and MPI) and can solve problems beyond billions of examples.

Although individual analysis of hyperparameters is helpful in understanding their impact on the model and identifying ranges of interest, the final search should not be performed sequentially. Each hyperparameter interacts with the others. We resort to a grid or random search to analyse various hyperparameters combinations. More information on search strategies can be found in machine learning using Python and Scikit-learn.

When computational resources (or time) are limited, it is advisable to follow one of the following strategies to identify the optimal hyperparameters of a gradient boosting model: (i) fix the number of trees and optimize the learning rate and (ii) fix the learning rate and add as many trees as necessary but activate early stopping to avoid overfitting. Once the values of these hyperparameters are identified, the remaining parameters are refined.

2.4 Metrics for model evaluation

We used four metrics to evaluate the performance of our model and compared it with the results of previous studies. The metrics used were the mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), and root mean squared error (RMSE). We also used the correlation coefficient.

The MAE measures the extent to which estimates or forecasts differ from the true values. It is most often used in a time series but can be applied to any statistical estimation. It can be applied to two sets of numbers, where one set is real, and the other is an estimate, forecast, or prediction. The MAE can be expressed as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}, \tag{1}$$

where y_i represents the original data array and $\hat{y}_i$ represents the predicted data array from the model.

The MAPE is an indicator of demand forecast performance that measures the (absolute) error size in percentage terms. It estimates the magnitude of the percentage error makes it a frequently used indicator by forecasters because of its ease of interpretation. It is even helpful when the volume of product demand is unknown because it is a relative measure.

The formula for calculating the MAPE is

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\%. \tag{2}$$

The MSE indicates the closeness of a regression line to a set of points. This is determined by taking the distances from the points to the regression line (these distances are the ‘errors’) and squaring them. Squaring is necessary to remove negative signs. It also gives more weight to the larger differences. It is called the mean squared error, as it finds the average of a set of errors. The lower the MSE, the better is the forecast. The formula is

$$\text{MSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}. \tag{3}$$

The corresponding RMSE simply the root square for the last one, that is RMSE = $(\text{MSE})^{0.5}$.

The correlation coefficient is a specific measure that quantifies the strength of the linear relationship between two variables in correlation analysis. This coefficient is represented by R in the correlation reports. For two variables, the formula compares the distance of each data point to the mean of the variable and uses this comparison to show how well the relationship between the variables conforms to an imaginary line drawn between the data. This is what is meant by correlation examines linear relationships. The correlation only refers to the two given variables and does not provide information about relationships beyond bivariate data. This test cannot detect outliers in the data (and therefore biases the results) and cannot correctly detect curvilinear relationships.

3 Results and Discussions

The results presented in this section for a model generation were calculated on an HP HP HP ProBook 440 G8 Notebook PC, 11th Gen Intel(R) Core i5-1135G7 @ 2.40GHz × 8 with 8.0 GiB of RAM, in execution times under one minute.

The main objective has been to estimate (predict) the current productivity of employees using as a basis for prediction several of the factors that imply an improvement (or difficulty) in meeting the goals set by the company to achieve production objectives. In this case, some factors are divided into those inherent to the employees themselves, such as the time they dedicate to a task, the amount of work in process at the same time, or the department and team in which they work. Some other factors are more related to incentives, which include the production (productivity) target expected by the company, and finally, other factors involve the number of employees who are not working at the moment due to process failures or the amount of time a product has been out of production.

3.1 Prediction process

For regressors, the dataset was divided into two parts: a training set (898 instances, 75%) and a test set (299 instances, 25%). The metrics were evaluated using the full prediction dataset.

To find the best hyperparameters in the configuration of the best predictive model, a strategy of searching and evaluating all models with a permutation network of parameters was used, running the model for each combination, and then evaluating the best one using the NMSE as the estimator.

For the random forest, we combined two important hyperparameters: the number of estimators for the decision trees, $N_{estim} = [100, 200, 400, 600, 800, 1000]$, and the maximum depth of the tree, $max_{depth} = [10, 15, 20, 25, 30]$, which implies 30 possible candidates. For gradient and extreme gradient boosting, 100 candidates were used to compute the fraction of features (randomly selected) that will train each tree F_{bytree}=linspace(0.5, 0.9, 5), the number of estimators for the decision trees $N_{estim} = [100, 400, 600, 1000]$ and the maximum depth of the tree $max_{depth} = [10, 15, 20, 25, 30]$. This methodology was implemented in the imputation strategy and data-mining manipulations s_1 and s_2.

For the two strategies, the results of the most optimal hyperparameters are summarized in Table 2.

The obtained results for error metrics for the models are presented in Table 3, and for comparison with our results, we include references to error obtained by two preceding approaches.

To better understand the variation in the actual and predicted productivity of the model for all regressors, Figure 2 shows the relationship between the observed and predicted values in a scatter plot. The straight line represents a perfect prediction. We cannot obtain perfect accuracy, but the values are extremely close, as can be confirmed from Table 3. Most of the points exhibited minimal variation from the actual value. For most of the points, the model learned extremely well and accurately predicted the actual value.

Strategy	Algotithm	N_{estim}	max_{depth}	F_{bytree}	Lowest RMSE found
s_1	Random Forest	400	15	–	0.1195
	Gradient Boosting	100	10	–	0.1445
	Extreme Gradient Boosting	100	10	0.6	0.1222
s_2	Random Forest	600	15	–	0.1212
	Gradient Boosting	100	10	–	0.1421
	Extreme Gradient Boosting	100	10	0.8	0.1218

Table 2: Hyperparameters used for best model prediction concerning to s_1 and s_2 data mining strategies. Lowest RMSE found refer to best score found in terms of measure the distance between the model and the data, (NEG_MEAN_SQUARED_ERROR) which return the negated value of the metric.

Source	DM strategy	Algorithm	MAE	MSE	RMSE	MAPE	Accuracy (%)	R
Ref$^{(a)}$	–	DL	0.0890	0.0180	0.1341	15.949	–	–
Ref$^{(b)}$	–	RF	0.0787	0.0153	0.1236	–	–	0.71
This work	s_1	RF	0.0365	0.004	0.0629	6.3009	93.70	0.94
		GB	0.0182	0.0031	0.0559	2.7565	97.24	0.95
		XGB	**0.0174**	**0.0025**	**0.0497**	**2.6602**	**97.34**	**0.96**
This work	s_2	RF	0.0369	0.004	0.0635	6.4349	93.57	0.94
		GB	0.019	0.0034	0.0586	2.8508	97.15	0.94
		XGB	**0.0176**	**0.0029**	**0.0534**	**2.7207**	**97.28**	**0.95**

Table 3: Performance of the model metric compared with those of previous models. First value is for s_1 and second value for s_2. Ref$^{(a)}$: Performance metrics from [Al Imran et al., 2019] in which the algorithm is Deep Learning (DL). Ref$^{(b)}$: Performance metrics from [Balla et al., 2021] in which the algorithm is Random Forest (RF). DM refers to Data Mining strategy keep for handle missing data for predictions.

All the aforementioned analysis and error metrics show that our model has learned the underlying pattern of the data well and can predict the actual productivity of garment employees. Moreover, our model is superior to the results obtained by previous studies, reducing by as much as ∼78%, ∼79%, ∼81% the values for MAE, MSE and MAPE, respectively.

Considering the results in Table 3, we can further conclude that the strategy of constructing dummy variables for categorical variables with zero (absence) and one (presence) values are usually the same results to the first strategy of categorizing variables with values on a nominal scale, however strategy s1 seems to be more successful in absolute terms. The effect of strategy s_1 increases the predictive capacity and minimises the error, and allows the identification of other predictor variables to improve the model.

Figure 3 shows similar analysis for the prediction and feature importance of variables when strategy s_2 is used.

We tracked the number of instances that are at a relative error more significant than the critical error between the original value of the predicted variable, x_{pred}, and the predictor variable, x_{orig}, with $RE = |x_{pred} - x_{orig}| \times 100/x_{orig}$. We selected 5% as the critical error and count the number of instances above this critical error value that the imputation has influenced. The results suggest that 24%, 9%, and 8% of the data have a relative error greater than 5% for the random forest, gradient boosting, and extreme gradient boosting. This further confirms the goodness of fit and improved predictions of the variable in this study using gradient boosting.

3.2 Factors that matter for the estimation of an employee's output

One of the most important objectives of developing machine learning algorithms to estimate, in this case the current productivity of the company's employees, is precisely to infer the importance of certain metrics or variables that influence productivity.

The above analysis revealed that we have found a non-parametric model that can predict with a high degree of accuracy the current production of employees in terms of several variables or metrics of the company. The best performing model is the one developed under extreme gradient boosting regression models. The

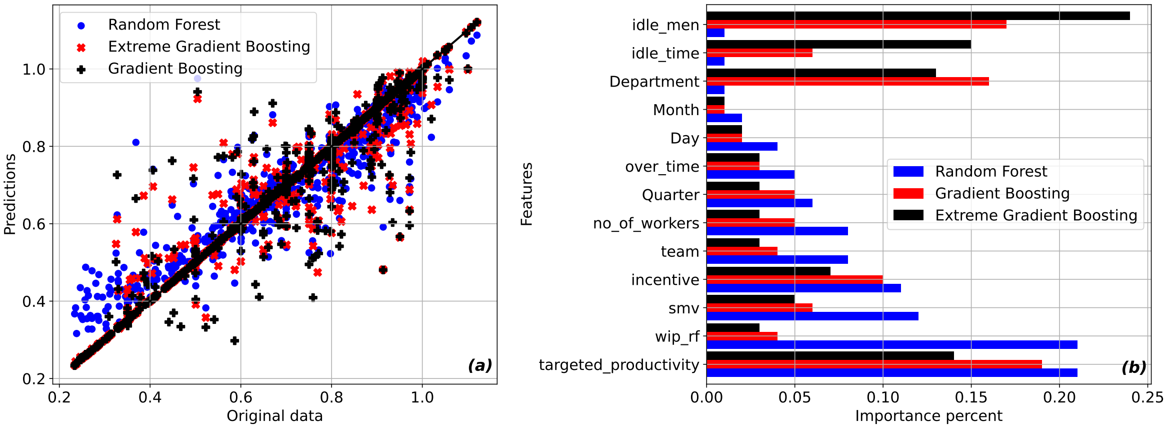

Figure 2: Prediction (a, left panel) and variable importance (b, right panel) in the model using strategy s_1 for all regressors considered.

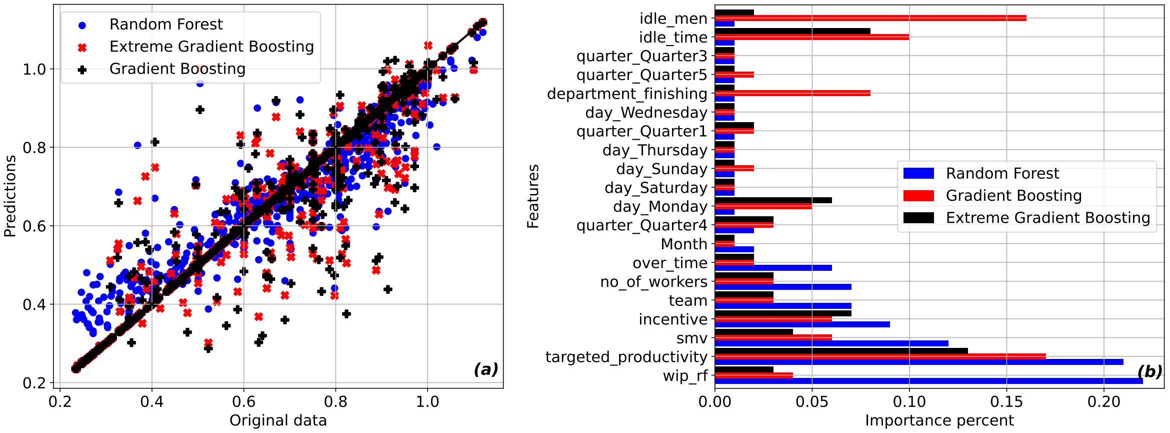

Figure 3: Prediction (a, left panel) and variable importance (b, right panel) in the model using strategy s_2 for all regressors considered.

model has been possible to reproduce after a data engineering process that has been done in two strategies. For each of the data mining strategies on the original data, we have found that the same model provides different significance for the predictor variables.

In strategy s_1 in which the non-numerical (categorical) variables have been coded, we have found that 6 variables collect about 80% of the explanation of employee productivity which are: the number of workers not working due to production delays, the time in which a product has been stagnant for various reasons, the expected productivity (goal) set by the company, the Department to which the worker belongs and the incentive (measured in money) offered for finishing tasks, with percentages of 24%, 15%, 14%, 13% and 7%, respectively.

This is in accord with expectations, at least for the first two contributions. Workers will have higher productivity to the extent that the company's processes are not stalled due to failures or delays in some part of the production process or when fewer employees are not working due to delays in the production line. However, although economic incentives have been studied in the past as a mechanism to increase workers' output (see for example: [Aguinis et al., 2013, Martha and Herbert, 2013, Yousaf et al., 2014]) in this case of the fashion industry, although it is an important variable it is not among the first in importance, in effect only contributing 7 percent. Workers in this business will produce almost regardless of the economic incentive. However, in the case of the data studied, the incentives offered have been generally low, ranging from USD 0 to USD 3600 for performing tasks with an average of approximately USD 40. In fact, for the

1197 workers in this company, only 3% receive incentives of more than USD 100. Moreover, this conclusion is much more powerful when we count the percentage of employees who have achieved a productivity higher than the expected target. Indeed, from the data we see that about 73% of employees produce more than what the company has set as a goal. In this case the incentives average about USD 45.

For the case of strategy s_2 in which the categorical variables have been coded using dummy variables, six variables account for 70% of the model's explanation. In this case, being working in the sewing department accounts for 30 percent of the significance followed by productivity goal (13 percent), time a product has been stagnant for various reasons (8 percent), economic incentive (7 percent), respectively.

4 Concluding remarks

The methodology used in this study has proven to help provide clues about how to solve a problem in an actual situation, in this case, to analyse employee productivity evaluations in an important fashion company. When applied to real data, our results indicated that a modelling process using random forest exhibits good performance in the prediction of employee productivity in terms of some initial variables. Approximately 23% of the forecast was highly influenced by the targeted productivity set by the authority for each team for each day; another 11% was the allocated time for a task, and another 11% was highly important for the amount of financial incentive in local money that enables or motivates a particular course of action and 20% of the associated team number with the instance and the number of workers in each team, to finally add 65% of the prediction capability with these five variables.

Moreover, the model predicted that the target plan for expected productivity set by the company management is the variable that best influences employee productivity at the time of evaluation. This variable contributed 20% to the estimation, at least twice as much as any other variable.

The model's performance was evaluated based on five evaluation metrics: MSE, RMSE, MAE, MAPE, and correlation coefficient. Our model improved the prediction capability by 70% compared with other neural network models in previous studies. For the case study in this work, gradient boosting was a more practical approach than neural networks and deep learning of random forest in terms of the metrics used to compare models.

The data engineering process is also vital for obtaining more accurate results. The iterative imputation method proved to be adequate. This showed that the data mining process for the initial data should not be performed lightly, and the initial data manipulation and the mining process are essential. This predictive performance can undoubtedly help manufacturers set precise targets, minimize production losses, and maximize profits.

Availability of data and materials

Finally, the model built and data, calculations, and algorithms used are available at https://github.com/sierraporta/Unsupervised_Learning_Productivity_Garment_Employees. Anyone can use this tool for the reproducibility of our results.

Author contributions

Author D. Sierra Porta is responsible for the research design, methodology design and application, study conceptualization, data analysis, data curation, writing and review of the paper.

Conflict of interest statement

The authors declare that there is no potential conflict of interest related to this article.

Funding

This research was not funded by any public or private entity and has been carried out by the author's work.

Acknowledgment

D. Sierra-Porta would like to thank A. Marrugo-Henández for kindly reviewing this manuscript and discussing the methodology and approach to the problem. His ideas allowed us to deepen some aspects and better redirect our methodology. We also thank the Universidad Tecnológica de Bolivar for allowing us to use the computer equipment on which the calculations were performed and the models for this study were generated.

References

[Aguinis et al., 2013] Aguinis, H., Joo, H., and Gottfredson, R. K. (2013). What monetary rewards can and cannot do: How to show employees the money. *Business Horizons*, 56(2):241–249.

[Al Imran et al., 2019] Al Imran, A., Amin, M. N., Rifat, M. R. I., and Mehreen, S. (2019). Deep neural network approach for predicting the productivity of garment employees. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1402–1407. IEEE. http://doi.org/10.1109/CoDIT.2019.8820486.

[Balla et al., 2021] Balla, I., Rahayu, S., and Purnama, J. J. (2021). Garment employee productivity prediction using random forest. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, 18(1):49–54. https://doi.org/10.33480/techno.v18i1.2210.

[Beauloye, 2019] Beauloye, F. E. (2019). Luxury resale: A secondhand strategy for brands. *Luxe Digital.* Available online: https://luxe.digital/business/digital-luxury-reports/luxury-resale-transformation/.

[Belgiu and Drăguţ, 2016] Belgiu, M. and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[Beltrami et al., 2019] Beltrami, M., Kim, D., and Rölkens, F. (2019). The state of fashion 2019. *Cámara de Comercio de Bogotá. Biblioteca Digital.* Available online: https://bibliotecadigital.ccb.org.co/handle/11520/26575.

[Bentéjac et al., 2021] Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

[Beretta and Santaniello, 2016] Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, 16(3):197–208. https://doi.org/10.1186/s12911-016-0318-z.

[Biau and Scornet, 2016] Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227. https://doi.org/10.1007/s11749-016-0481-7.

[Brodkin, 2000] Brodkin, K. (2000). Global capitalism: What's race got to do with it? *American Ethnologist*, 27(2):237–256. https://doi.org/10.1525/ae.2000.27.2.237.

[Cabigiosu, 2020] Cabigiosu, A. (2020). An overview of the luxury fashion industry. *Digitalization in the Luxury Fashion Industry*, pages 9–31. https://doi.org/10.1007/978-3-030-48810-9_2.

[Chen et al., 2015] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

[Dogan and Birant, 2021] Dogan, A. and Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 166:114060. https://doi.org/10.1016/j.eswa.2020.114060.

[Gazzola et al., 2020] Gazzola, P., Pavione, E., Pezzetti, R., and Grechi, D. (2020). Trends in the fashion industry. the perception of sustainability and circular economy: A gender/generation quantitative approach. *Sustainability*, 12(7):2809. https://doi.org/10.3390/su12072809.

[Goyzueta et al., 2021] Goyzueta, C. A. R., De la Cruz, J. E. C., and Machaca, W. A. M. (2021). Advantages of assembly machine learning models for predicting employee productivity in a garment manufacturing company. In *2021 IEEE Engineering International Research Conference (EIRCON)*, pages 1–4. IEEE. http://doi.org/10.1109/EIRCON52903.2021.9613559.

[Imran et al., 2021] Imran, A. A., Rahim, M. S., and Ahmed, T. (2021). Mining the productivity data of the garment industry. *International Journal of Business Intelligence and Data Mining*, 19(3):319–342.

[Jadhav et al., 2019] Jadhav, A., Pramod, D., and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933. https://doi.org/10.1080/08839514.2019.1637138.

[Lakshminarayan et al., 1996] Lakshminarayan, K., Harp, S. A., Goldman, R. P., Samad, T., et al. (1996). Imputation of missing data using machine learning techniques. In *KDD*, volume 96.

[Lakshminarayan et al., 1999] Lakshminarayan, K., Harp, S. A., and Samad, T. (1999). Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275. https://doi.org/10.1023/A:1008334909089.

[Mark et al., 2021] Mark, B. G., Rauch, E., and Matt, D. T. (2021). Worker assistance systems in manufacturing: A review of the state of the art and future directions. *Journal of Manufacturing Systems*, 59:228–250. https://doi.org/10.1016/j.jmsy.2021.02.017.

[Martha and Herbert, 2013] Martha, H. and Herbert, K. (2013). The impact of monetary and non-monetary rewards on motivation among lower level employees in selected retail shops. *African Journal of Business Management*, 7(38):3929–3935.

[Mezzadri, 2010] Mezzadri, A. (2010). Globalisation, informalisation and the state in the indian garment industry. *International Review of Sociology*, 20(3):491–511. https://doi.org/10.1080/03906701.2010.511910.

[Natekin and Knoll, 2013] Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21. http://doi.org/10.3389/fnbot.2013.00021.

[Obiedat and Toubasi, 2022] Obiedat, R. and Toubasi, S. A. (2022). A combined approach for predicting employees' productivity based on ensemble machine learning methods. *Informatica*, 46(5). https://doi.org/10.31449/inf.v46i5.3839.

[Pantanowitz and Marwala, 2009] Pantanowitz, A. and Marwala, T. (2009). Missing data imputation through the use of the random forest algorithm. In *Advances in computational intelligence*, pages 53–62. Springer.

[Schapire, 2013] Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference*, pages 37–52. Springer. https://doi.org/10.1007/978-3-642-41136-6_5.

[Shamsuzzoha et al., 2013] Shamsuzzoha, A., Kankaanpaa, T., Carneiro, L. M., Almeida, R., Chiodi, A., and Fornasiero, R. (2013). Dynamic and collaborative business networks in the fashion industry. *International Journal of Computer Integrated Manufacturing*, 26(1-2):125–139. https://doi.org/10.1080/0951192X.2012.681916.

[Shi and Horvath, 2006] Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138. https://doi.org/10.1198/106186006X94072.

[Stekhoven, 2015] Stekhoven, D. J. (2015). missforest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*, pages ascl–1505. https://doi.org/10.1093/bioinformatics/btr597.

[Yousaf et al., 2014] Yousaf, S., Latif, M., Aslam, S., and Saddiqui, A. (2014). Impact of financial and non-financial rewards on employee motivation. *Middle-East journal of scientific research*, 21(10):1776–1786.

[Yunus and Yamagata, 2012] Yunus, M. and Yamagata, T. (2012). The garment industry in bangladesh. *Dynamics of the Garment Industry in Low-Income Countries: Experience of Asia and Africa (Interim Report). Chousakenkyu Houkokusho, IDE-JETRO*, 6:29.

[Zhang et al., 2010] Zhang, S., Wu, X., and Zhu, M. (2010). Efficient missing data imputation for supervised learning. In *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*, pages 672–679. IEEE. http://doi.org/10.1109/COGINF.2010.5599826.