Journal of Turkish

Operations Management

Data driven approach for weight restricted data envelopment analysis models with a single output

Şenol Kurt[1*], Mustafa Kerem Yüksel[2], Burcu Dinçergök[3]

[1]Department of Business, Atılım University, Ankara
e-mail: kurtsenol@student.atilim.edu.tr, ORCID No: https://orcid.org/0000-0002-4526-1592
[2]Department of Economics, Bilkent University, Ankara
e-mail: mkeremyuksel@gmail.com, ORCID No: https://orcid.org/0000-0002-7051-6526
[3] Department of Business, Atılım University, Ankara
e-mail: burcu.dincergok@atilim.edu.tr, ORCID No: https://orcid.org/0000-0002-7050-8163
*Corresponding Author

Article Info

Article History:
Received: 29.07.2023
Revised: 23.08.2023
Accepted: 29.08.2023

Keywords

Data Envelopment Analysis,
Machine Learning,
Gradient Boosting Tree

Abstract

This study aims to explore whether a machine learning algorithm can be used to make improvements in assessing unit efficiencies via a data envelopment analysis (DEA) model. In this study, a DEA model is used to calculate the efficiency scores of Decision Making Units (DMUs). Subsequently, an ML algorithm is trained with the aim of predicting a single output using inputs. Ranking of input features based on relative feature importance values obtained from the trained ML model is fed to the DEA model as weight restrictions. As a result, the two DEA models are compared with each other. ML-based insights (feature importance ranking) improve the DEA model in the direction of fewer zero weights. The additional weight restrictions are data dependent and hence realistic. As a novel approach, this study proposes the use of machine learning-based feature importance values to overcome a limitation of a DEA model.

1. Introduction

Every organization, whether for-profit (such as bank, manufacturing or transportation company, etc.) or nonprofit (such as university, hospital, NGO, etc.) conducts operations to achieve certain goals and objectives. To do so, the organization converts inputs into outputs. It is imperative to measure this performance quantitatively to assess how efficiently inputs are transformed into outputs (performance measurement). Additionally, this process can pinpoint areas where an organization can enhance its operations, thereby guiding the organization towards increasing productivity.

An important family of tools for efficiency analysis, especially in the context of multiple inputs and multiple outputs, is Data Envelopment Analysis (DEA) (Ghiyasi et al., 2021). DEA is a nonparametric method based on linear programming (LP) to measure the (relative) performance and assign an efficiency score (ES) to each Decision-Making Unit (DMU) relying on a common set of input and output features. DEA was originally proposed by Charnes et al. (1978, 1981) under the constant returns to scale (CRS) setting (CCR model) where it is assumed that there exists proportionality between inputs and outputs. The CCR model was later extended by Banker et al. (1984) which allows for variable returns to scale (VRS). DEA aims to construct Production Possibility Set (PPS) instead of using functional forms that connect inputs to outputs. As the initial stage in the analysis, the PPS can be defined as the minimum set enveloping assessed observed units along with all feasible input and output combinations (Thanassoulis, 2001; Atici, 2012). PPS is used to calculate the relative efficiency of a DMU by comparing its input and output combination to a set of all possible combinations. The nonparametric nature of

DEA requires no functional relationship among input and output variables. In general, ES in a DEA model ranges from 0 to 1, which refers to the relative position of DMUs off the efficient frontier which is characterized by the DMUs with ES of 1 (and hence called efficient DMUs). A DMU with a smaller score is called to be inefficient (Cooper et al., 2011).

Meanwhile, the rapid growth in the amount of data and the increased computing power for processing vast amounts of data has recently made Machine Learning (ML) a widely used tool. ML aims to reveal patterns inside the data. It is possible that the rather complex (nonlinear) relationship between multiple independent and dependent variable might not be captured by rule-based solutions. In these situations, ML algorithms can help. Numerous studies aim to integrate DEA and ML, as both methodologies assist decision-makers quantitatively by providing valuable insights, especially when there are multiple variables and complex relationships among them (Emrouznejad & Shale, 2009; Jomthanachai et al., 2021; Zhu et al., 2018, Appiahene et al., 2020; Farahmand et al., 2014; Salehi et al., 2019; Koronakos & Sotiropoulos, 2020). However, this integration is not straightforward as there are some differences between the two methodologies. When an ML model is trained on a dataset, it can be used for unseen data. In DEA, however, the efficient frontier (and ES of DMUs) is calculated for given DMUs but an additional DMU may require a recalculation. Moreover, such a shift in the efficient frontier alters the ES of almost all the DMUs. This creates difficulty in replacing one method with the other blindly but does not prevent exploiting the best of both worlds to develop "integrated" tools.

For instance, some studies try to find and calibrate the "*best*" ML algorithm to predict ES which can otherwise be calculated by DEA (Jomthanachai et al., 2021; Koronakos & Sotiropoulos, 2020). To that end, a subset of DMUs is chosen and their ESs are calculated then these are used to train an ML algorithm using DEA inputs and outputs as the features of the ML model to predict the ES of DMUs. The trained ML model can predict the ES of a newly added DMU without the need for any calculation (Tayal et al., 2020). Since ES is a continuous variable, the ML model should be a regression algorithm. On the other hand, some studies use information from the DEA model to classify whether DMU demonstrates efficiency or not; and develop and train an ML model (classification algorithm) to predict the category to which the next DMU belongs (Singpai & Wu, 2020; Hoz et al. 2021).

Unsupervised ML algorithms, such as clustering and dimensionality reduction, are also used together with DEA (Hoz et al. 2021; Tayal et al., 2020). In general, the aim of utilizing clustering algorithms is to group DMUs based on their similarities for a given set of features. Since DEA requires a set of homogenous DMUs for ES calculation, clustering is used to create homogenous subgroups within a given set of DMUs. Then (possibly different) DEA models are employed for each subset of DMUs. Therefore, the efficiencies for each subgroup are evaluated within the subgroup.

Both ML and DEA broadly suffer from high dimensionality, which leads to poor performance (Chen et al., 2022; Kumar et al., 2021). "*The curse of dimensionality*" in ML is due to the fact that as the number of features increases, the observed data in the feature space is not distributed so that the feature space is represented 'fairly'. In other words, as the dimension increases, insufficient empirical evidence exist that explain the relationship between features and the output confidently in almost everywhere in the feature space except for a certain region the observed data is present. Dimensionality reduction techniques such as principal component analysis and t-distributed Stochastic Neighbor Embedding (Tayal et al., 2020; Lin, 2021) address this problem. A similar problem manifests itself in DEA as well. Due to the relative character of the DEA, with too many inputs and outputs, there may exist virtually efficient DMUs whose efficiency are only due to some insignificant inputs and outputs. Dimensionality reduction can be employed with DEA when the number of inputs and outputs is large.

Feature selection techniques are also used to tackle the problem of high dimensionality. Since datasets may consist of irrelevant and noisy data, selecting appropriate features generally increase ML performance. It is also applicable to DEA models. While calculating ES with DEA, using only 'appropriate' input and output features produces more simple, interpretable, and reliable results (Chen et al., 2022; Kumar et al., 2021).

ML is also used to interpret the DEA results, rather than predicting the ES of DMUs. Employing decision tree-based ML models, such as random forests and gradient boosting trees, can identify features that have more impact on predicting the target variable (Adler & Painsky, 2022). In DEA, the target variable can either be the ES (continuous) or the status of DMUs (discrete). In both cases, decision tree-based ML models can identify key factors that affect the performance of DMUs. Identifying these factors can help decision-makers on setting priorities and allocate resources for improving efficiency.

DEA methodology doesn't require predefined weights for input and output features to calculate ES. Due to its weight flexibility feature, it can result in undesired situations where the weights are unreasonable and are not in

line with expert views on the production process (Cooper et al., 2011). A DEA model may put too much weight on a few inputs and outputs while ignoring most of them which is a limitation of the DEA methodology. It may show a prevalence of zero weights leading to concerns about ES calculation (Forsund, 2013). Putting too much weight on a few inputs and outputs may result in defining an inefficient DMU as a virtually efficient one. Moreover, that DMU appears as a reference to other DMUs by being on the efficient frontier.

Applying weight restrictions may help to enhance discrimination and reduce weight dispersion. Price information, expert opinions, value information, and managerial goals are sources for imposing weight restrictions in DEA (Cooper et al., 2011). Rather than using subjective information, this study proposes to apply weight restrictions for single output datasets by using information obtained from the data itself. ML algorithms have the capability to recognize patterns and extract insightful information from complicated data. The motivation of this study is to integrate ML and DEA methodologies in a manner whereby acquired information from the dataset via an ML algorithm is used to overcome the limitation of traditional DEA models related to a substantial proportion of zero weight assignment. This study will also aim to explore whether an ML algorithm can be used to make improvements in a DEA model in the sense that certain inputs and outputs cannot be ignored, so that the virtually efficient DMUs are avoided, and more precise and realistic efficiency assessments can be made.

Decision tree-based ML models such as random forests, and gradient boosting trees can identify features that have more impact on predicting the target variable (Adler & Painsky, 2022). Previous studies apply feature importance values to identify and explain key factors that have an impact on the efficiency of DMUs (Aydın & Yurdakul, 2020; Nandy & Singh, 2020a; 2020b; Thaker et al., 2021; Rebai et al., 2020; Xu et al., 2021). However, this study proposes a novel approach to incorporating feature importance to DEA. The utilization of feature importance ranking as an additional weight restriction is proposed to improve discrimination power and weight dispersion of a DEA model.

The paper is organized as follows. Section 2 reviews the literature on studies integrating DEA and ML. The methodology and the model design are explained in Section 3. Section 4 contains the results of the study and Section 5 includes conclusions and recommendations.

2. Literature review

ML and DEA are two well-studied distinct areas of research and application. Recently, numerous studies have been conducted with the aim of integrating ML and DEA. In supervised ML, the goal is to get a predictive model which is trained on input features and an output (target) feature. The model learns the relationship between the inputs and the output during the training process. Then, the trained model can predict the target variable from given input features. DEA, on the other hand, relies on a mathematical model to calculate the relative efficiency scores of DMUs where each ES comes from a distinct optimization problem, different for each DMU.

Both methodologies are used to analyze data quantitatively and try to get insights from the data to help the decision-making process. They are used to assist decision-makers by providing useful outputs which cannot be obtained from the data at first glance. However, an ML model can be used when new data comes in whereas a DEA model must be rebuilt when new data is added to the dataset. An ML model is built to make predictions for unseen data. On the other hand, a DEA model is only built to make calculations for the DMUs to which the dataset belongs.

In an effort to integrate ML and DEA, one approach has been training multiple ML (regression) models where the features are the inputs and outputs of the DEA model and the target variable is the ES, and then predicting the ES using trained models on the test data and comparing various ML algorithms based on their prediction capabilities. The main motivation in this line of research is to replace the DEA methodology with an ML model so that additional DMUs do not require a costly re-computation. Studies employ different datasets, such as schools, hospitals, farms, etc., and tried to build a regression model to predict the ES of DMUs as close as possible to the ES calculated by a DEA model (Emrouznejad & Shale, 2009; Jomthanachai et al., 2021; Zhu et al., 2018, Appiahene et al., 2020; Farahmand et al., 2014; Salehi et al., 2019; Koronakos & Sotiropoulos, 2020).

A second approach involves training multiple ML (classification) models, where the target variable is whether a member belongs to efficiency tier (ET) or not based on DEA results. Subsequently, the model predicts the ET with trained models using test data, and finally, various ML algorithms are compared based on their prediction capabilities. In other words, predicting ET, rather than ES of DMUs is also a widely studied subject (Singpai & Wu, 2020; Hoz et al. 2021; Song & Zhang, 2009; Hong et al., 1999; Gupta et al., 2016; Mirmozaffari et al., 2020).

A third approach involves clustering DMUs with an unsupervised ML model to form homogeneous sets of DMUs, then, calculating ES separately for each cluster (subgroup) using DEA (Hoz et al. 2021; Özsoy & Örkcü, 2021; Aydın & Yurdakul, 2020; Mirmozaffari et al., 2020; Tayal et al., 2020). Özsoy and Örkcü (2021) initially employs an ML algorithm to cluster 43 Turkish airports into three groups: big-scale, middle-scale, and small-scale. Then they employ DEA to measure the performance of each group. Mirmozaffari et al. (2020) combine the clustering algorithm and DEA to study cement companies from developing countries. On the other hand, Aydın and Yurdakul (2020) applied two clustering algorithms, k-means, and hierarchical clustering, to divide the countries into three groups, then employ DEA models to measure the performance of 142 countries against COVID-19.

A fourth approach has been employing an ML algorithm for feature selection or dimensionality reduction, then building a DEA model only with selected features or in reduced dimensions (Tayal et al., 2020; Chen et al., 2022; Kumar et al., 2021; Zhang et al., 2015; Lin, 2021). In particular, Tayal et al. (2020) employ Principal Component Analysis to a set of factors influencing facility layouts for dimensionality reduction before utilizing a DEA model. Lin (2021) employs multiple DEA models by combining inputs and outputs in various combinations. Then the data is analyzed using t-distributed stochastic neighbor embedding (t-SNE) which is a dimension reduction technique to reveal the main characteristics of the data. Chen et al. (2022) employ Least Absolute Shrinkage and Selection Operator (LASSO) algorithm for feature selection and then employs a DEA model using only selected features. Kumar et al. (2021) also propose a feature selection methodology to select appropriate inputs and outputs for the DEA model.

Another approach is building an ML algorithm to predict ES or ET, then analyzing input and output features based on their feature importance (Aydın & Yurdakul, 2020; Nandy & Singh, 2020a; 2020b; Thaker et al., 2021; Rebai et al., 2020; Xu et al., 2021). Nandy and Singh (2020a) employ a DEA model to measure paddy producers' efficiency and then employ support vector machine and random forest algorithms to predict whether a DMU is efficient or inefficient. They utilize the trained ML model to identify key factors influencing performance. Nandy and Singh (2020b) employ random forest and logistic regression for predicting the ET of DMUs. In doing so, they seek to determine significant environmental factors that affect farmers' performance. Thaker et al. (2021) employ DEA to measure Indian banks' performance. Thereafter, they use random forests to examine the factors such as corporate governance, bank characteristics, etc. that have an impact on bank efficiency. Rebai et al. (2020) similarly employ DEA to calculate efficiency scores of Tunisian secondary schools, then utilize regression tree and random forest algorithms to find out key factors that influence academic achievement. Xu et al. (2021) also apply DEA to measure U.S. states' COVID-19 response performance, then employ four different ML models (classification and regression tree, random forest, boosted tree, and logistic regression) to predict whether a state is efficient or inefficient, to find out influential factors on performance.

Kongar & Adebayo (2021) classify unstructured data (text) with an ML model to form input and output features and then build a DEA model to calculate ES to assess the impact of social media marketing on business performance.

There are also studies in the other direction, which use DEA results as feedback to ML models. To cite a few, Kheirkhah et al. (2013) employ DEA to measure the performance of ANNs which have different number layers and nodes, and Mousavi et al. (2019) employ DEA to calculate the ES of DMUs, then use ES as an input feature to an ML algorithm.

In this study, on the other hand, a novel approach is proposed in incorporating feature importance to DEA. Rather than using feature importance only for the identification of key factors that affect the performance of DMUs, using feature importance ranking as an additional weight restriction to a DEA model is suggested. Due to the weight flexibility feature of a DEA model, it can assign zero weights to most of the input and output features. It means that the employed model ignores most of the inputs and outputs and put too much weight on a few. It can lead to undesired ES and efficiency discrimination among DMUs. Additional weight restrictions prevent a DEA model from assigning a large number of zero weights to the features and hence, avoid virtually efficient DMUs.

3. Model

As one of the weaknesses of the DEA approach is that the individual ES of DMUs requires a weight assignment that is most favorable to the DMU itself. This yields unreasonable neglect of certain inputs (and outputs) via zero weights. In this study, the aim is to develop a method (Figure 1) to assess and rank the relative weights of the inputs in a DEA setting and use this additional information for a weight-restricted version of the same DEA model. Python programming language (*gurobi* package for DEA and *scikit-learn* package for ML model) is used to develop models.

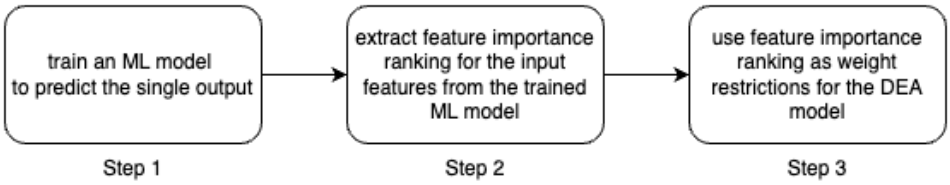

Figure 1. Proposed Methodology

For comparison, this study consists of two phases. Firstly, the input-oriented CCR model (see Charnes et al., 1978) is employed to calculate the ES of DMUs. The mathematical model in multiplier form for DMU r is given in Equation 1 to Equation 4:

$$\max \sum_j \lambda_j y_{rj} \tag{1}$$

subject to

$$\sum_i \mu_i x_{ri} = 1 \tag{2}$$

$$\sum_j \lambda_j y_{sj} \leq \sum_i \mu_i x_{si}, \quad \forall s \tag{3}$$

$$\lambda_j \geq 0, \ \forall j; \text{ and } \mu_i \geq 0, \ \forall i \tag{4}$$

where the objective value is the efficiency score of DMU r, inputs, and outputs of DMU s are denoted by $x_{si}$and y_{sj} respectively, and input and output weights are denoted by μ_i and λ_j, respectively. The analysis is restricted to have only one output.

In the next phase, the proposed approach is applied. Initially, feature importance ranking from an ML algorithm is extracted where inputs are the features, and the output is the target variable. This information yields in which order the inputs have a say on the output which cannot be neglected for an arbitrary DMU.

Then, the ES of each DMU is calculated with additional constraints that take into account the importance of the inputs. More specifically, the same mathematical problem is solved with the following additional constraints (Equation 5):

$$\mu_i \geq \mu_j \text{ if } \mathrm{r}(\mu_i) \leq r(\mu_j), \forall i, j \tag{5}$$

where $\mathrm{r}(\mu_i)$ is the importance ranking of weight μ_i of input i. In other words, if an input has a higher importance in determining the output, its weight should not be less than the weight of an input with less importance.

3.1. Dataset

The energy plant dataset used by Khezrimotlagh et al. (2019) is a good fit for the purpose of this study. The year 2020 data which contains 1644 energy plants (DMUs) is used in this study. In addition to the 3 inputs (number of generators, nameplate capacity, heat input), 5 of the 6 outputs are treated as undesired outputs since they are related to gas emissions (NOx, SO2, CO2, CH4, and N2O emissions) and hence are used as desired inputs (Zhou & Liu, 2015). Finally, a DEA model is built with 8 inputs and 1 output. The single output in the dataset is net electricity generated.

3.2. ML model

The objective of the ML model is to predict the single output using 8 inputs. As the target variable, net electricity generated is a continuous variable, a supervised regression ML model should be used. Since it produces competitive, highly robust, and interpretable results (Friedman, 2001), the gradient boosting tree regressor is selected as the predictive ML model.

The gradient boosting tree algorithm is a decision tree-based ML algorithm that can be used for both classification and regression problems (Natekin & Knoll, 2013). It is a kind of boosting algorithm that trains multiple ML models in a sequence. It tries to get more accurate predictions by combining them. It starts with a basic model and then the next model pays more attention to the wrong predictions the previous model has made by giving them more weights. This procedure is continued by the following models and at the end, they are combined to build a strong model that can make more accurate predictions.

Gradient boosting tree algorithm also produces relative feature importance (Adler & Painsky, 2022) which is a measure of how much it contributes to predicting the target variable. In other words, feature importance values reveal which features are more important, and which are less. As features are used to split data in decision tree algorithms, the closer the feature is to the root node, the higher relative importance the feature has. The features at the top of a decision tree contribute more than the features at the bottom to make predictions.

A gradient boosting tree is developed to predict net electricity generated values using 8 input features. 5-fold cross-validation is applied during the training process. Hyperparameters (subsample, minimum samples leaf, minimum samples split, maximum depth, number of estimators, learning rate) are tuned throughout the cross-validation using a grid search process. The hyperparameters that produce the least average root mean squared error (RMSE) at the end of the 5-fold cross-validation process, are selected for the model.

One of the advantages of the gradient boosting tree algorithm is that it can produce relative feature importance, which is used in this study, as well. Feature importance values are provided relatively, and it refers to a measure that indicates how much it contributes to predicting the target variable (Adler & Painsky, 2022). It helps the model developer to get insight from the dataset. Especially, when there is a large number of features, it is not easy to understand how the fitted model makes predictions. Feature importance values reveal which features are more important, and which are less. As features are used to split data in decision tree algorithms, the closer the feature is to the root node, the higher relative importance the feature has (Kotsiantis, 2013). The features at the top of a decision tree contribute more than the features at the bottom to make predictions.

A gradient boosting tree is developed to predict net electricity generated values using 8 input features, namely number of generators, nameplate capacity, heat input, NOx, SO2, CO2, CH4, and N2O emissions. 5-fold cross-validation is applied during the training process. Hyperparameters (subsample, minimum samples leaf, minimum samples split, maximum depth, number of estimators, learning rate) are tuned throughout the cross-validation using a grid search process. The hyperparameters that produce the least average root mean squared error (RMSE) at the end of the 5-fold cross-validation process, are selected for the model.

4. Results and discussion

In this study, a real-world dataset is utilized. Initially, a DEA model (input-oriented CCR) is used to calculate the ES of DMUs, which will be called DEA-01 afterward. Then, to obtain relative feature importance values of input features, an ML algorithm (gradient boosting tree) is trained that aims to predict the single output (net electricity generated) using input features. The trained model produces an average R2 of 98.51 and an average RMSE of 91.82 for the 5-fold cross-validation process. Feature importance values are given in Table 1. The feature importance ranking of input features obtained from the trained ML model is used as weight restrictions to the second DEA model (DEA-02). As a result, the two DEA models are compared with each other. Especially, zero weight counts, efficient DMU counts, and ES values for each model are reported.

Table 1 reports the feature importance values as well as the count and percentage of zero weights for each input (feature) where the inputs are sorted in ascending order with respect to the zero weight counts. In DEA-01, there are 23 efficient plants (among 1644), where 61% of the input weights are found to be zero. The sparse character of the weights is also evident from the observation that in DEA-01, for 6 out of 8 inputs, zero weight is assigned for more than half of the DMUs. It is also important to observe that feature importance is not fully correlated with zero counts (see Table 1) since otherwise there is no need to augment an ML model.

Table 1. Count and Ratio of Zero Weights

Feature	*Feature Importance Value*	*DEA-01 Model*		*DEA-02 Model*	
		Count	*Percentage*	*Count*	*Percentage*
Annual N^2O emissions (lbs)	0.9880	17	1.03	0	0
Annual NOx emissions (tons)	0.0008	720	43.80	89	5.41
Annual CO^2 emissions (tons)	0.0033	822	50.00	43	2.62
Annual SO^2 emissions (tons)	0.0005	1040	63.26	1190	72.38
Nameplate capacity (MW)	0.0006	1209	73.54	367	22.32
Total annual heat input (MMBtu)	0.0024	1231	74.88	97	5.90
Number of generators	0.0004	1404	85.40	1447	88.02
Annual CH^4 emissions (lbs)	0.0039	1589	96.65	316	19.22

Figure 2 plots the ES obtained from the two DEA models. As the constraint set of DEA-02 is the subset of the constraint set of DEA-01, the ES in DEA-02 will obviously be smaller for almost all the DMUs. This is also evident in the density plots of the ES of the DMUs (Figure 3) where the ES shifts to the left in DEA-02. To put the change in ES in perspective, Figure 4 shows that the percentage decrease in ES increases as the ES increases.

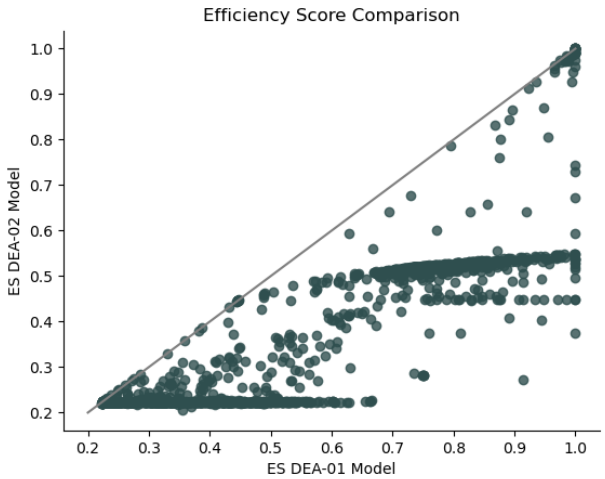

Figure 2. Comparison of ESs for the 2 DEA models

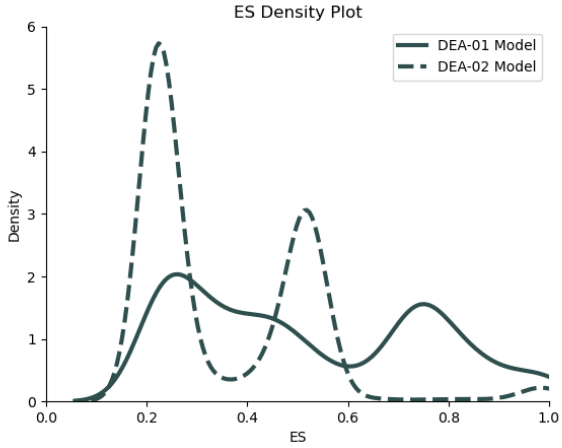

Figure 3. Density plots for the ESs in the 2 DEA models

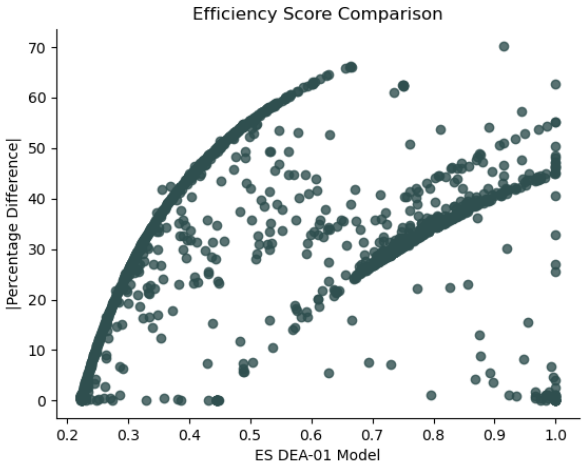

Figure 4. Percentage decrease in ES

In DEA-02, there are only 6 efficient plants. The ratio of the zero input weights is dramatically reduced to 27%. Moreover, for only 2 of the inputs, zero weight is assigned to more than 25% of the DMUs. Needless to say, these two are the ones with the least importance. In other words, at least 6 inputs are taken into account in ES calculation for almost all of the DMUs. Figure 5 plots the zero weight counts in each model where the size indicates the number of DMUs. The most frequently observed change is having 5 zero weighted inputs in DEA-01, but only 2 zero weights in DEA-02. The most radical change is observed in DMUs with 7 zero weights in DEA-01 and none in DEA-02. All in all, zero weight counts decrease for 1604 plants; remain the same for 32 plants, and increase for only 8 plants.

The results indicate that the ML-based DEA improves on the zero weight counts and thus has more discriminatory power against the virtually efficient DMUs. Hence the efficient frontier is more realistic.

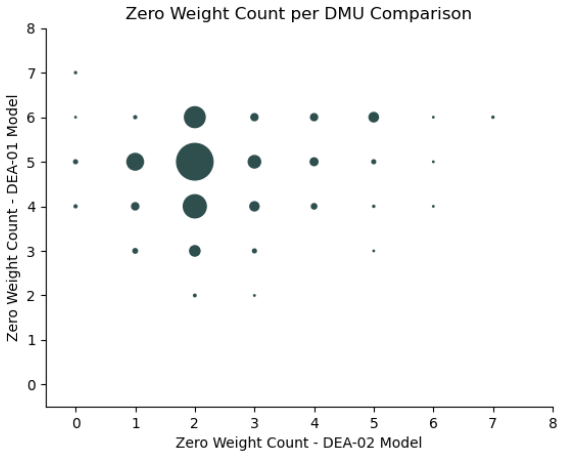

Figure 5. Comparison of zero weight counts in each model

5. Conclusion

DEA and ML are two widely used methodologies that aim to gain insights from the data. Several studies integrate two methodologies. These studies generally focus on predicting the ES or ET of newly added DMUs, finding out key factors that have an impact on the performance of DMUs, dimensionality reduction, and feature selection. This study proposes a novel approach to the integration of the two methodologies.

Since DEA methodology doesn't require predefined weights for input and output features to calculate ES, it may put too much weight on a few inputs and outputs while ignoring most of them by assigning zero weights. Putting too much weight on a few inputs and outputs may result in defining an inefficient DMU as a virtually efficient one. Moreover, that DMU appears as a reference to other DMUs by being on the efficient frontier. Applying weight restrictions may help to overcome this limitation. Although price information, expert opinions, value information, and managerial goals can be used for imposing weight restrictions in DEA (Cooper et al., 2011), this study proposes to apply weight restrictions for single output datasets by using information obtained from the data itself. ML algorithms have the capability to recognize patterns and extract insightful information from complicated data. Gradient boosting trees, a decision tree-based ML model, can rank features with respect to their impact on predicting the target variable. Rather than using this information only for the identification of key factors that affect the performance of DMUs, this study aims to use it as a weight restriction for a newly proposed DEA model to make a more precise and realistic efficiency assessment.

In this study, a real-world dataset about energy plants is used. The dataset contains 8 inputs and a single output. Initially, an input-oriented CCR model is developed to measure the ES of energy plants. Then, a gradient boosting tree algorithm is trained that aims to predict the single output using inputs to obtain relative feature importance values of input features. The implied ranking is later used as weight restrictions for a second DEA model. As a result, the proposed DEA model which takes the input-output relations into account improves on the zero counts and thus avoids virtually efficient DMUs as much as possible. It can be concluded that the proposed approach leads to a DEA model that has more discriminatory power and less zero weights. The approach uses information obtained from the data itself rather than relying on subjective judgments.

The proposed approach is examined on a dataset that contains only one output. Future research may be related to a dataset where there is more than one single output. The study may focus on whether feature importance ranking remains the same while predicting different outputs. Different ML algorithms can also be utilized to obtain feature importance ranking which may lead to differences in terms of importance ranking. The proposed approach may also be applied to datasets from different domains to observe its robustness and generalizability.

Conflicts of Interest

The authors declared that there is no conflict of interest.

Contribution of Authors

This study is based on the Şenol KURT's doctoral thesis. Doç.Dr. Burcu DİNÇERGÖK and Dr. Mustafa Kerem YÜKSEL are the thesis supervisors.

References

Appiahene, P., Missah, Y. M., & Najim, U. (2020). Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks. *Advances in fuzzy systems*, 2020, 1-12. doi: https://doi.org/10.1155/2020/8581202

Adler, A. I., & Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*, 24(5), 687. doi: https://doi.org/10.3390/e24050687

Atici, K. B. (2012). Using data envelopment analysis for the efficiency and elasticity evaluation of agricultural farms (Doctoral dissertation, University of Warwick). Retrieved from: https://wrap.warwick.ac.uk/54354/2/WRAP_THESIS_Atici_2012.pdf

Aydin, N., & Yurdakul, G. (2020). Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. *Applied Soft Computing*, 97, 106792. doi: https://doi.org/10.1016/j.asoc.2020.106792

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9), 1078-1092. doi: https://doi.org/10.1287/mnsc.30.9.1078

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, 2(6), 429-444. doi: https://doi.org/10.1016/0377-2217(78)90138-8

Charnes, A., Cooper, W. W., & Rhodes, E. (1981). Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. *Management science*, 27(6), 668-697. doi: https://doi.org/10.1287/mnsc.27.6.668

Chen, Y., Tsionas, M. G., & Zelenyuk, V. (2021). LASSO + DEA for small and big wide data. *Omega*, 102, 102419. doi: https://doi.org/10.1016/j.omega.2021.102419

Cooper, W. W., Seiford, L. M., & Zhu, J. (Eds.). (2011). *Handbook on data envelopment analysis* (2nd ed.). Springer.

De La Hoz, E., Zuluaga, R., & Mendoza, A. (2021). Assessing and Classification of Academic Efficiency in Engineering Teaching Programs. *Journal on Efficiency and Responsibility in Education and Science*, 14(1), 41-52. Retrieved from: https://hdl.handle.net/20.500.12834/880

Emrouznejad, A., & Shale, E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. *Computers & Industrial Engineering*, 56(1), 249-254. doi: https://doi.org/10.1016/j.cie.2008.05.012

Farahmand, M., Desa, M. I., & Nilashi, M. (2014). A combined data envelopment analysis and support vector regression for efficiency evaluation of large decision making units. *International journal of engineering and technology (IJET)*, 2310-2321. Retrieved from: https://www.researchgate.net/publication/288995583_A_Combined_Data_Envelopment_Analysis_and_Support_Vector_Regression_for_Efficiency_Evaluation_of_Large_Decision_Making_Units

Førsund, F. R. (2013). Weight restrictions in DEA: misplaced emphasis?. *Journal of Productivity Analysis*, 40, 271-283. Retrieved from: https://link.springer.com/article/10.1007/s11123-012-0296-9

Gupta, A., Kohli, M., & Malhotra, N. (2016, July). Classification based on Data Envelopment Analysis and supervised learning: A case study on energy performance of residential buildings. In 2016 *IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)* (pp. 1-5). IEEE. doi: 10.1109/ICPEICES.2016.7853706

Ghiyasi, M., Rouyendegh, B. D., & Özdemir, Y. S. (2021). Local and global energy efficiency analysis for energy production based on multi-plant generalized production technology. *IEEE Access*, 9, 58208-58215. doi: 10.1109/ACCESS.2021.3072493

Hong, H. K., Ha, S. H., Shin, C. K., Park, S. C., & Kim, S. H. (1999). Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning. *Expert Systems with Applications*, 16(3), 283-296. doi: https://doi.org/10.1016/S0957-4174(98)00077-3

Jomthanachai, S., Wong, W. P., & Lim, C. P. (2021). An Application of Data Envelopment Analysis and Machine Learning Approach to Risk Management. *IEEE Access*, 9, 85978-85994. doi: 10.1109/ACCESS.2021.3087623

Kheirkhah, A., Azadeh, A., Saberi, M., Azaron, A., & Shakouri, H. (2013). Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis. *Computers & Industrial Engineering*, 64(1), 425-441. doi: https://doi.org/10.1016/j.cie.2012.09.017

Khezrimotlagh, D., Zhu, J., Cook, W. D., & Toloo, M. (2019). Data envelopment analysis and big data. *European Journal of Operational Research*, 274(3), 1047-1054. doi: https://doi.org/10.1016/j.ejor.2018.10.044

Kongar, E., & Adebayo, O. (2021). Impact of Social Media Marketing on Business Performance: A Hybrid Performance Measurement Approach Using Data Analytics and Machine Learning. *IEEE Engineering Management Review*, 49(1), 133-147. doi: 10.1109/EMR.2021.3055036

Koronakos, G., & Sotiropoulos, D. N. (2020, July). A Neural Network approach for Non-parametric Performance Assessment. In 2020 *11th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-8). IEEE. doi: 10.1109/IISA50023.2020.9284346

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39, 261-283. doi: Retrieved from: https://link.springer.com/article/10.1007/s10462-011-9272-4

Kumar, A., Shrivastav, S. K., & Mukherjee, K. (2022). Performance evaluation of Indian banks using feature selection data envelopment analysis: A machine learning perspective. *Journal of Public Affairs*, e2686. doi: https://doi.org/10.1002/pa.2686

Lin, S. J. (2021). Integrated artificial intelligence and visualization technique for enhanced management decision in today's turbulent business environments. *Cybernetics and Systems*, 52(4), 274-292. doi: https://doi.org/10.1080/01969722.2021.1881244

Mirmozaffari, M., Shadkam, E., Khalili, S. M., Kabirifar, K., Yazdani, R., & Gashteroodkhani, T. A. (2021). A novel artificial intelligent approach: comparison of machine learning tools and algorithms based on optimization DEA Malmquist productivity index for eco-efficiency evaluation. *International Journal of Energy Sector Management*, *15*(3), 523-550. Retrieved from: https://www.emerald.com/insight/content/doi/10.1108/IJESM-02-2020-0003/full/html

Mirmozaffari, M., Yazdani, M., Boskabadi, A., Ahady Dolatsara, H., Kabirifar, K., & Amiri Golilarz, N. (2020). A novel machine learning approach combined with optimization models for eco-efficiency evaluation. *Applied Sciences,* 10(15), 5210. doi: https://doi.org/10.3390/app10155210

Mousavi, M. M., Ouenniche, J., & Tone, K. (2019). A comparative analysis of two-stage distress prediction models. *Expert Systems with Applications*, 119, 322-341. doi: https://doi.org/10.1016/j.eswa.2018.10.053

Nandy, A., & Singh, P. K. (2020). Application of fuzzy DEA and machine learning algorithms in efficiency estimation of paddy producers of rural Eastern India. Benchmarking: An International Journal, 28(1), 229-248. Retrieved from: https://www.emerald.com/insight/content/doi/10.1108/BIJ-01-2020-0012/full/html

Nandy, A., & Singh, P. K. (2020). Farm efficiency estimation using a hybrid approach of machine-learning and data envelopment analysis: Evidence from rural eastern India. *Journal of Cleaner Production*, 267, 122106. doi: https://doi.org/10.1016/j.jclepro.2020.122106

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

Özsoy, V. S., & Örkcü, H. H. (2021). Structural and operational management of Turkish airports: a bootstrap data envelopment analysis of efficiency. *Utilities Policy*, 69, 101180. doi: https://doi.org/10.1016/j.jup.2021.101180

Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, 70, 100724. doi: https://doi.org/10.1016/j.seps.2019.06.009

Salehi, V., Veitch, B., & Musharraf, M. (2020). Measuring and improving adaptive capacity in resilient systems by means of an integrated DEA-Machine learning approach. *Applied ergonomics*, 82, 102975. doi: https://doi.org/10.1016/j.apergo.2019.102975

Sarkis, J. (2007). *Preparing your data for DEA*. In Modeling data irregularities and structural complexities in data envelopment analysis (pp. 305-320). Springer, Boston, MA.

Singpai, B., & Wu, D. (2020). Using a DEA–AutoML Approach to Track SDG Achievements. *Sustainability*, 12(23), 10124. doi: https://doi.org/10.3390/su122310124

Song, J., & Zhang, Z. (2009, January). Oil refining enterprise performance evaluation based on DEA and SVM. In 2009 *Second International Workshop on Knowledge Discovery and Data Mining* (pp. 401-404). IEEE. doi: 10.1109/WKDD.2009.43

Tayal, A., Kose, U., Solanki, A., Nayyar, A., & Saucedo, J. A. M. (2020). Efficiency analysis for stochastic dynamic facility layout problem using meta-heuristic, data envelopment analysis and machine learning. *Computational Intelligence*, 36(1), 172-202. doi: https://doi.org/10.1111/coin.12251

Tayal, A., Solanki, A., & Singh, S. P. (2020). Integrated frame work for identifying sustainable manufacturing layouts based on big data, machine learning, meta-heuristic and data envelopment analysis. *Sustainable Cities and Society*, 62, 102383. doi: https://doi.org/10.1016/j.scs.2020.102383

Thaker, K., Charles, V., Pant, A., & Gherman, T. (2021). A DEA and random forest regression approach to studying bank efficiency and corporate governance. *Journal of the Operational Research Society*, 1-28. doi: https://doi.org/10.1080/01605682.2021.1907239

Thanassoulis, E. (2001). *Introduction to the theory and application of data envelopment analysis*. Dordrecht: Kluwer Academic Publishers.

Xu, Y., Park, Y. S., & Park, J. D. (2021). Measuring the Response Performance of US States against COVID-19 Using an Integrated DEA, CART, and Logistic Regression Approach. *In Healthcare* (Vol. 9, No. 3, p. 268). MDPI. doi: https://doi.org/10.3390/healthcare9030268

Zhang, Y., Yang, C., Yang, A., Xiong, C., Zhou, X., & Zhang, Z. (2015). Feature selection for classification with class-separability strategy and data envelopment analysis. *Neurocomputing*, 166, 172-184. doi: https://doi.org/10.1016/j.neucom.2015.03.081

Zhou, Z., & Liu, W. (2015). DEA models with undesirable inputs, intermediates, and outputs. *Data envelopment analysis: A handbook of models and methods*, 415-446. Springer

Zhu, N., Zhu, C., & Emrouznejad, A. (2020). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies. *Journal of Management Science and Engineering*. doi: https://doi.org/10.1016/j.jmse.2020.10.001

Zhu, J. (2009). *Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets* (Vol. 2). New York: Springer.