International Journal of Operations Research and Artificial Intelligence

www.ijorai.reapress.com

Int. J. Oper. Res. Artif. Intell. Vol. 1, No. 2 (2025) 54–60.

Paper Type: Original Article

Predicting Super-Efficiency of Commercial Bank Branches Using Regression Models

Fateme Gerami[1,*], Shahla Gerami[1]

[1] Department of Electrical Engineering, Faculty of Electrical Engineering, Jundi Shapur University of Technology, Dezful, Iran; fatemee.geramii@gmail.com.

Citation:

Received: 14 November 2024 *Revised: 12 January 2025* *Accepted: 17 March 2025*	Gerami, F., & Gerami, Sh. (2025). Predicting super-efficiency of commercial bank branches using regression models. *International journal of operations research and artificial intelligence*, *1*(2), 54-60.

Abstract

This study focuses on predicting the super-efficiency scores of commercial bank branches by employing various regression models. The analysis is conducted on a dataset comprising 375 bank branches from the fiscal year 2017, utilizing a range of financial, operational, and cost-related indicators as input features. A suite of regression techniques, including linear regression, ensemble methods such as Random Forest and XGBoost, as well as neural network models, is implemented to estimate the super-efficiency values. Model performance is assessed through metrics including Mean Absolute Error (MAE) and the coefficient of determination (R^2). The findings reveal that non-linear models, especially ensemble-based algorithms, outperform linear models in terms of accuracy and generalizability. This regression framework offers a robust decision-support tool for evaluating and benchmarking the operational efficiency of bank branches.

Keywords: Data envelopment analysis, Machine learning, Commercial banks, Bank branch performance evaluation.

1 | Introduction

Efficiency and productivity are critical factors in enhancing the performance and service quality within the banking sector. Each bank branch, as an independent decision-making unit, is influenced by a variety of financial, operational, and cost-related factors that determine its effectiveness and overall efficiency. Accurately measuring efficiency and forecasting the future performance of bank branches is essential for managers and policymakers to improve resource allocation and optimize service delivery.

One of the widely used methods for efficiency assessment is Data Envelopment Analysis (DEA), a non-parametric approach that evaluates the relative efficiency of Decision-Making Units (DMUs) based on their inputs and outputs. DEA not only provides an efficiency score but also facilitates the identification of super-

Corresponding Author: fatemee.geramii@gmail.com

https://doi.org/10.48314/ijorai.v1i2.61

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

efficiency, which highlights branches performing beyond the conventional efficiency frontier. In this study, super-efficiency scores for bank branches were initially calculated using DEA and used as the target labels for subsequent regression modeling.

With recent advances in Machine Learning (ML), regression models have gained prominence in predicting performance indices within banking. These models are capable of capturing complex and nonlinear relationships between various financial and operational features of bank branches and their super-efficiency scores. The accuracy and generalizability of these predictive models depend heavily on the rigorous preprocessing of data, including normalization and partitioning into training, validation, and test subsets. These preprocessing steps ensure model robustness and reduce overfitting.

Furthermore, selecting appropriate optimization functions during the training of regression models plays a pivotal role in improving convergence rates and overall prediction accuracy. This research investigates multiple regression models, including linear regression, ridge regression, random forest, XGBoost, and Multilayer Perceptron (MLP) neural networks, to develop a comprehensive framework for predicting bank branch super-efficiency.

The primary goal of this study is to establish an effective and accurate regression-based approach for forecasting super-efficiency scores, which can serve as a valuable decision-support tool for bank branch performance evaluation and management. This study seeks to integrate DEA techniques with ML models to predict the super-efficiency scores of commercial bank branches. While traditional methods such as DEA are widely used for measuring the relative efficiency of DMU—particularly in the financial sector—the increasing complexity and nonlinearity of modern datasets necessitate the adoption of more advanced approaches like ML. By employing sophisticated regression models such as Random Forest, XGBoost, MLP, and a stacking ensemble, the study demonstrates a significant improvement in prediction accuracy compared to basic linear models. This innovative approach can serve as a decision-support tool for bank managers in assessing branch performance and optimizing resource allocation.

In the context of developing DEA models, the following references provide diverse foundational and methodological contributions. To begin with, a series of studies by Mozaffari et al. [1] exhibit a strong focus on extending DEA-R (Russell) models in various dimensions, such as cost efficiency, revenue efficiency, and super-efficiency [2]. These works contribute not only to the theoretical and mathematical depth of DEA models but also introduce new frameworks—such as slacks-based models and efficient frontier analysis—that enable more accurate performance evaluation of DMU [3]. Furthermore, the application of DEA-R models in real-world environments like the petrochemical industry—under fuzzy settings and with undesirable outputs—demonstrates the model's flexibility in handling complex and uncertain data [4].

Subsequently, research conducted by Noura et al. [5] addresses key topics such as congestion in DEA and super-efficiency through social effectiveness, which enriches the scope of DEA beyond conventional efficiency metrics [6]. These studies have helped incorporate more realistic factors, such as social roles and resource constraints, into performance measurement frameworks. Additional contributions include the analysis of supply chains using Sub-DMUs in DEA [7] and the evaluation of productivity indicators in the oil industry through Multi-Attribute Decision-Making (MADM) approaches [8], reflecting a productive intersection between DEA methodology and industrial decision support.

Finally, the integration of fuzzy logic into decision-making—particularly through the synergy of fuzzy AHP and fuzzy TOPSIS for ranking factors influencing employee turnover intention—shows a trend towards merging quantitative and behavioral perspectives in analytical models [9]. This approach enhances DEA's role from a purely technical evaluation tool to a hybrid, intelligent decision-support framework suitable for analyzing human-centric and uncertain environments. Collectively, these references represent a coherent pathway from theoretical advancement to interdisciplinary applications in the field of DEA.

The proposed hybrid methodology is inspired by foundational work on statistical frontier models [10], dynamic environmental indices in DEA [11], and ML techniques for dynamic inefficiency analysis [12].

Additionally, studies such as Guerrero et al. [13] and Guillen et al. [14] have shown that combining DEA with ML can enhance predictive capabilities and support more effective performance evaluation in organizational settings. Building on this body of research, the current study offers a precise, flexible, and scalable model for efficiency analysis in the banking industry.

The framework of the article is as follows:

In Section 2, various regression models, including linear regression, ridge regression, random forest, XGBoost, and MLP neural networks, are introduced, and their characteristics are explained.

Section 3 details the methodology, including labeling data using DEA, data preprocessing, standardization, and splitting into training, validation, and test sets. The training process, hyperparameter optimization, and model evaluation metrics are also described.

Section 4 presents the case study on 375 commercial bank branches, including data feature analysis, correlation examination, and feature variability assessment. And finally, Section 5 provides the overall conclusion, emphasizing the importance of ensemble models for improving the prediction accuracy of bank branch super-efficiency, and offers suggestions for future research.

2 | Regression Models

In this study, five distinct regression models were selected to predict the super-efficiency scores of bank branches, each chosen for its ability to capture different aspects of the relationship between input features and the target variable. These models encompass both linear and nonlinear approaches, allowing for a comprehensive exploration of potential dependencies within the data.

Linear regression serves as the most fundamental model, assuming a linear relationship between the explanatory variables and the outcome. Due to its simplicity and interpretability, it is often employed as a baseline against which the performance of more complex models can be compared. Ridge regression extends the linear regression framework by incorporating an L2 regularization term. This penalization helps to mitigate the risk of overfitting and addresses multicollinearity among input features, thereby improving the stability and generalizability of the model in high-dimensional settings.

Random forest regression is an ensemble learning method that aggregates the predictions of multiple decision trees constructed on bootstrapped subsets of the data. This approach effectively models nonlinear interactions and variable importance while enhancing robustness against noise and outliers. XGBoost Regression utilizes gradient boosting to iteratively combine weak learners, focusing sequentially on samples that were previously mispredicted. This method is highly efficient computationally and delivers state-of-the-art accuracy by optimizing both the model structure and learning process.

MLP regression, a type of feedforward artificial neural network, employs multiple layers of interconnected neurons to capture complex nonlinear relationships in the data. Through backpropagation and weight adjustment, MLP models can learn intricate patterns that traditional regression methods might overlook.

Collectively, these models provide a spectrum of analytical capabilities, facilitating a thorough evaluation of their predictive power in modeling the super-efficiency of bank branches.

3 | Methodology

The first phase of this study involved labeling the target variable by applying DEA, a widely used non-parametric technique for measuring relative efficiency. DEA evaluates each bank branch's efficiency based on multiple input variables—including financial, operational, and cost-related indicators—and corresponding output performance metrics. This process yielded a super-efficiency score for each branch, representing its performance relative to the efficiency frontier and enabling identification of units exceeding the typical efficiency boundary. These super-efficiency scores were utilized as the ground truth labels for subsequent regression-based predictive modeling.

Following the labeling step, the dataset comprised 375 bank branches, each described by 21 financial, operational, and cost-related features serving as explanatory variables. To prepare the data for modeling, a rigorous preprocessing pipeline was implemented. Initially, normalization was conducted using standardization to adjust all input features to a common scale with zero mean and unit variance. This step mitigated the impact of differing units and value ranges across features, ensuring equitable contribution to the model training. Next, the dataset was partitioned into training, validation, and testing subsets with proportions of 60%, 20%, and 20%, respectively. This data splitting strategy was designed to enable robust model training, hyperparameter tuning, and unbiased evaluation, thereby enhancing the generalizability and preventing overfitting of the predictive models.

The modeling phase involved implementing five distinct regression algorithms selected to capture diverse linear and nonlinear dependencies between the input features and the super-efficiency target variable. Linear Regression was employed as a baseline model to represent linear relationships. Ridge regression introduced L2 regularization to manage multicollinearity and reduce overfitting. Random Forest regressor, an ensemble technique aggregating multiple decision trees, was utilized to model complex nonlinear interactions and improve prediction stability. XGBoost regressor, leveraging gradient boosting methods with advanced optimization strategies, provided a powerful framework for accurate and efficient learning. Finally, an MLP neural network was applied to capture deep nonlinear patterns within the data through its multi-layered architecture.

Each model was trained using appropriate optimization algorithms tailored to the model architecture, such as variants of stochastic gradient descent and second-order optimization methods. These optimization functions were crucial in ensuring efficient convergence and fine-tuning of model parameters to maximize predictive accuracy.

For model evaluation, the trained algorithms were assessed on the unseen test dataset using standard regression metrics. Mean Absolute Error (MAE) quantified the average absolute deviation between predicted and actual super-efficiency scores, while the coefficient of determination (R^2) measured the proportion of variance in the target variable explained by the model. These performance indicators facilitated comprehensive comparison and informed the selection of the most effective model for super-efficiency prediction of bank branches.

4 | Case Study: Efficiency Analysis of 375 Commercial Bank Branches

The dataset comprises 375 commercial bank branches, each described by 22 numerical features related to financial, operational, and cost parameters. Initial examination of the dataset reveals no missing values, ensuring data completeness and reliability for subsequent modeling tasks. Descriptive statistics provide insight into the distribution and variability of the features. For example, total_deposits ranges widely from as low as 6.38 to over 3.2 million units, with a mean around 118,000 and a high standard deviation indicating substantial dispersion among branches. Similar patterns of wide variation are observed in fixed_assets, personnel_expenses, and gross_loans. This heterogeneity suggests the need for normalization or scaling prior to modeling to harmonize feature scales.

Correlation analysis highlights strong positive relationships among many financial variables. Notably, total_deposits exhibits very high correlation with gross_loans (0.978), total_assets (0.970), and net_interest_income (0.960), reflecting the interconnected nature of banking financial indicators. conversely, cost-related features such as price_of_labour and price_of_funds tend to show weak or negative correlations with these financial indicators.

The target variable, Eff_AP, representing super-efficiency scores, shows weak positive correlations (Less than 0.1) with most features, implying that the prediction task may involve complex, potentially nonlinear interactions that cannot be captured by simple linear models alone.

Table 1. Summary statistics of key bank branch variables, including mean, median, min, max, and standard deviation.

Feature	Mean	Median (50%)	Min	Max	Std Dev
Total_deposits	118,465.9	15,456.1	6.38	3,266,469	340,842
Fixed_assets	1,276.1	152.6	0.03	38,046.6	3,854.6
Personnel_expenses	1,263.1	197.2	1.19	17,653.7	2,756.4
Non_performing_loans	3,257.2	378.6	0.70	63,155.6	7,905.6
Gross_loans	95,143.0	12,642.7	25.23	2,185,860	239,667
Total_securities	36,835.2	3,999.2	5.72	944,889.6	109,428
Price_of_funds	0.0603	0.0181	0.00025	6.8593	0.3698
Price_of_capital	8.0336	2.7474	0.2662	480.7083	29.147
Price_of_labour	0.0106	0.0091	0.0011	0.0522	0.0064
Price_of_loans	0.1053	0.0783	0.0166	1.4559	0.1118
Total_interest_expenses	2,767.7	281.2	0.05	52,140.2	6,977.4
Non_interest_expenses	2,404.6	382.5	2.08	34,884.0	5,119.1
Total_Assets	180,516.0	22,350.8	52.43	4,006,242	460,427
Net_interest_income	3,142.5	488.1	1.45	80,227.0	8,311.9
Other_interest_income	1,941.7	167.1	0.07	44,367.9	5,639.9
Non_interest_income	1,629.7	191.9	0.08	25,492.0	3,591.3
LLPGL	0.9348	0.5500	0.0100	9.4500	1.1874
NPLGL	5.8181	2.8400	0.0500	63.4000	8.5938
LLP	714.57	74.46	0.05	19,891.3	2,215.4
Eff_AP (Target variable)	1.1829	0.9394	0.3031	11.4965	1.0842

The table provides a summary of descriptive statistics for the key features of data from 375 commercial bank branches. For each variable, central measures such as mean and median are reported alongside minimum, maximum, and standard deviation values. These statistics offer a general understanding of the data distribution, variability, and potential outliers.

The results indicate that certain variables, such as total deposits, gross loans, and total assets, exhibit very high means and standard deviations, reflecting substantial diversity in the size and performance of bank branches. This significant variability may arise from differences in operational scale, geographic location, and economic conditions among branches.

The model's target variable, Efficiency (Eff_AP), also shows a wide range of values with a mean around 1.18 and a standard deviation slightly above 1. This highlights significant variation in branch performance

Boxplots and histograms (Not shown here) further confirm the presence of skewed distributions and potential outliers in key variables, emphasizing the importance of careful preprocessing, such as outlier handling and robust scaling, before applying regression models.

In summary, the data exploration phase reveals rich and varied patterns within the dataset, providing a solid foundation for informed model selection and preprocessing strategies in the following phases.

This code implements a comprehensive predictive modeling pipeline to analyze the efficiency of 375 commercial bank branches using the target variable Eff_AP. The process begins by loading the dataset from an Excel file and separating input features from the target variable. The data is then split into training, validation, and test subsets to facilitate model training, hyperparameter tuning, and unbiased performance evaluation, respectively.

For models sensitive to feature scaling, the input data is standardized, whereas the Random Forest model is trained on the raw, unscaled data. Five regression models—linear regression, ridge regression, Random Forest, XGBoost, and Multi-Layer Perceptron (MLP —are defined and trained. Hyperparameters for Ridge, Random Forest, XGBoost, and MLP models are optimized using Grid Search combined with cross-validation to select the best parameter configurations.

Following training, each model's predictions on the validation set are evaluated using MAE and the coefficient of determination (R^2). The optimized base models are then integrated into a stacking regressor ensemble, where predictions from the base models, along with the original input features, are passed to a final linear regression meta-model for training.

Finally, the performance of the ensemble model is assessed on both the validation and test datasets, demonstrating improved prediction accuracy compared to the individual base models.

Table 2. Performance comparison of regression models on the validation set using MAE and R^2 metrics.

Model	Validation MAE	Validation R^2
Linear regression	0.6872	0.0079
Ridge regression	0.6942	0.0010
Random forest	0.4389	0.5653
XGBoost	0.3706	0.6344
MLP regressor	0.7107	0.0237
Stacking regressor	0.1976	0.9370

The table above compares the performance of various regression models based on two key metrics: MAE and coefficient of determination (R^2) on the validation dataset. These metrics serve as indicators of model prediction accuracy.

MAE represents the average absolute difference between predicted and actual values; lower values indicate better accuracy.

R^2 measures the proportion of variance explained by the model, ranging from 0 to 1, where values closer to 1 denote superior model performance.

In *Table 2*, Simple linear models such as linear regression and ridge regression exhibit relatively poor performance (High MAE and near-zero R^2), indicating their limited capability to capture the complexity of the data.

Tree-based models like Random Forest and boosting models such as XGBoost demonstrate significantly better results, with XGBoost achieving the best MAE and R^2 scores.

The MLP neural network performed worse than tree-based models.

Stacking regressor, an ensemble learning method, notably outperformed all individual models by achieving a substantially lower MAE and a much higher R^2. This highlights that combining multiple models can significantly enhance prediction accuracy.

5 | Conclusion

The study results indicate that advanced regression models significantly enhance the prediction accuracy of super-efficiency scores compared to traditional linear methods. Among the tested models, ensemble approaches such as Random Forest and XGBoost demonstrated superior performance, reflected in lower MAE and higher R^2 values. Notably, the stacking regressor, which integrates multiple base models, achieved the best predictive accuracy, underscoring the benefits of combining diverse algorithms.

These findings emphasize the importance of leveraging non-linear and ensemble modeling techniques in efficiency analysis within the banking sector. By accurately forecasting branch performance, the proposed approach can aid managers and decision-makers in identifying high-performing units, optimizing resource allocation, and guiding strategic improvements. Future research may explore incorporating additional data sources or employing advanced deep learning architectures to refine predictive capabilities further.

Funding

This research received no external funding and was conducted using independent academic resources.

Data Availability

The data used in this study are not publicly available due to confidentiality agreements, but can be provided upon reasonable request to the corresponding author.

References

[1] Mozaffari, M. R., Kamyab, P., Jablonsky, J., & Gerami, J. (2014). Cost and revenue efficiency in DEA-R models. *Computers & industrial engineering*, *78*, 188–194. https://doi.org/10.1016/j.cie.2014.10.001

[2] Gerami, J., Mozaffari, M. R., Wanke, P. F., & Correa, H. (2022). A novel slacks-based model for efficiency and super-efficiency in DEA-R. *Operational research*, *22*(4), 3373–3410. https://doi.org/10.1007/s12351-021-00679-6

[3] Mozaffari, M. R., Dadkhah, F., Jablonsky, J., & Wanke, P. F. (2020). Finding efficient surfaces in DEA-R models. *Applied mathematics and computation*, *386*, 125497. https://doi.org/10.1016/j.amc.2020.125497

[4] Mozaffari, M. R., Mohammadi, S., Wanke, P. F., & Correa, H. L. (2021). Towards greener petrochemical production: Two-stage network data envelopment analysis in a fully fuzzy environment in the presence of undesirable outputs. *Expert systems with applications*, *164*, 113903. https://doi.org/10.1016/j.eswa.2020.113903

[5] Noura, A. A., Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., Rashidi, S. F., & Parker, B. R. (2010). A new method for measuring congestion in data envelopment analysis. *Socio-economic planning sciences*, *44*(4), 240–246. https://doi.org/10.1016/j.seps.2010.06.003

[6] Noura, A. A., Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., & Fanati Rashidi, S. (2011). Super-efficiency in DEA by effectiveness of each unit in society. *Applied mathematics letters*, *24*(5), 623–626. https://doi.org/10.1016/j.aml.2010.11.025

[7] Rashidi, S. F., & Barati, R. (2014). On the comparison of supply chain with sub-Dmus in Dea. *Advances in environmental biology*, 2387–2391. https://b2n.ir/mx3313

[8] Rashidi, S. F. (2015). Evaluation of productivity indicators in the oil industry by using multi-attribute decision making approach (MADM). *International journal of advanced and applied sciences*, *2*(6), 25–31. https://b2n.ir/zn7859

[9] Barati, R., & Fanati Rashidi, S. (2024). Fuzzy AHP and fuzzy TOPSIS synergy for ranking the factor influencing employee turnover intention in the Iran hotel industry. *Journal of applied research on industrial engineering*, *11*(1), 57–75. https://doi.org/10.22105/jarie.2022.336603.1464

[10] Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, *6*(1), 21–37. https://doi.org/10.1016/0304-4076(77)90052-5

[11] Aparicio, J., Barbero, J., Kapelko, M., Pastor, J. T., & Zofío, J. L. (2017). Testing the consistency and feasibility of the standard Malmquist-Luenberger index: Environmental productivity in world air emissions. *Journal of environmental management*, *196*, 148–160. https://doi.org/10.1016/j.jenvman.2017.03.007

[12] Aparicio, J., Esteve, M., & Kapelko, M. (2023). Measuring dynamic inefficiency through machine learning techniques. *Expert systems with applications*, *228*, 120417. https://doi.org/10.1016/j.eswa.2023.120417

[13] Guerrero, N. M., Aparicio, J., & Valero-Carreras, D. (2022). Combining data envelopment analysis and machine learning. *Mathematics, 10*(6), 909. https://doi.org/10.3390/math10060909

[14] Guillen, M. D., Aparicio, J., & Esteve, M. (2023). Gradient tree boosting and the estimation of production frontiers. *Expert systems with applications*, *214*, 119134. https://doi.org/10.1016/j.eswa.2022.119134