INDUSTRIAL WORKERS' EFFICIENCY IN INDIAN SUBCONTINENT: A MACHINE LEARNING MODEL APPROACH

Sadman SADIK[1*], Syed Mahedi HASEN[2]

[1]Khulna University of Engineering and Technology, Bangladesh, sadmansdk@gmail.com

[2]Rajshahi University of Engineering and Technology, Bangladesh, syedmahedihasen207@gmail.com

Received: 06.09.2024 Accepted: 22.11.2024 https://doi.org/10.24264/lfj.24.4.4

INDUSTRIAL WORKERS' EFFICIENCY IN INDIAN SUBCONTINENT: A MACHINE LEARNING MODEL APPROACH

ABSTRACT. The growing popularity of machine learning offers exciting possibilities for real-world applications. Since worker efficiency directly impacts a company's bottom line, especially for small and medium businesses (SMEs), implementing these tools can be a game-changer. By improving worker efficiency, machine learning can help SMEs minimize losses and drive growth. This research explores the potential of AI model not to replace workers but to uplift them. In this study, we try to find out the industrial workers' efficiency, especially in the Leather & Textiles industries, based on some parameters like expertise, education, salary, working hour, standard minute value (SMV), working position, key performance indicators (KPI) etc. The study investigates different regression models for predicting worker efficiency. Here we compare six models including Random Forest and XG Boost, using metrics like Mean Squared Error to find the best performing model. XG Boost and Histogram Gradient Boosting show the best results in predicting worker efficiency. XG Boost achieved high accuracy (R-squared around 0.78) with low errors (MSE around 0.01). Light GBM came in a close third, while Random Forest and Ada Boost did poorly. Machine learning techniques like XG Boost can significantly improve worker efficiency in the Indian subcontinent in leather-textile industries.

KEY WORDS: workers performance, industrial worker augmentation, data driven efficiency.

EFICIENȚA LUCRĂTORILOR DIN INDUSTRIA SUBCONTINENTULUI INDIAN: O ABORDARE A MODELULUI DE ÎNVĂȚARE AUTOMATIZATĂ

REZUMAT. Popularitatea din ce în ce mai mare a învățării automatizate oferă posibilități interesante pentru aplicații din lumea reală. Întrucât eficiența lucrătorilor are un impact direct asupra profitului unei companii, în special pentru întreprinderile mici și mijlocii (IMM-uri), implementarea acestor instrumente poate duce la o revoluționare. Îmbunătățind eficiența lucrătorilor, învățarea automatizată poate ajuta IMM-urile să reducă la minimum pierderile și să stimuleze creșterea. Această cercetare explorează potențialul modelului AI nu de a înlocui lucrătorii, ci de a-i ajuta să-și îmbunătățească performanțele. În acest studiu, s-a încercat determinarea eficienței lucrătorilor din industrie, în special din industria de textile și pielărie, pe baza unor parametri precum expertiza, educația, salariul, programul de lucru, valoarea minutelor standard (SMV), funcția, indicatorii cheie de performanță (KPI) etc. Studiul investighează diferite modele de regresie pentru prezicerea eficienței lucrătorilor. Se compară șase modele, inclusiv Random Forest și XG Boost, folosind indici de cuantificare precum Mean Squared Error pentru a găsi cel mai performant model. XG Boost și Histogram Gradient Boosting prezintă cele mai bune rezultate în ceea ce privește prezicerea eficienței lucrătorilor. Cu XG Boost s-a obținut o precizie ridicată (R-pătrat în jurul valorii de 0,78) cu puține erori (MSE în jur de 0,01). Light GBM s-a clasat pe locul trei, la distanță apropiată, în timp ce Random Forest și Ada Boost au fost nesatisfăcătoare. Tehnicile de învățare automatizată precum XG Boost pot îmbunătăți semnificativ eficiența lucrătorilor din subcontinentul indian din industria de textile și pielărie.

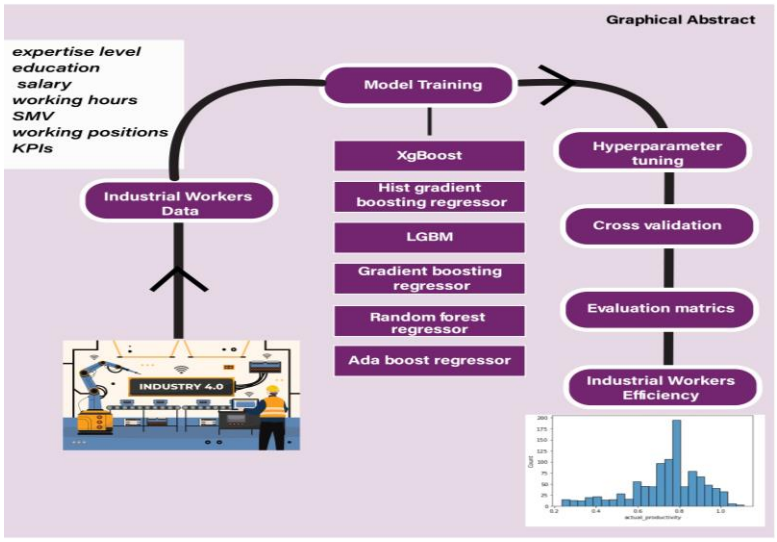
CUVINTE CHEIE: performanța lucrătorilor, îmbunătățirea performanțelor lucrătorilor din industrie, eficiență bazată pe date.

EFFICACITÉ DES TRAVAILLEURS DANS L'INDUSTRIE DU SOUS-CONTINENT INDIEN : UNE APPROCHE DE MODÈLE D'APPRENTISSAGE AUTOMATIQUE

RÉSUMÉ. La popularité croissante de l'apprentissage automatique offre des possibilités captivantes pour des applications concrètes. Étant donné que l'efficacité des travailleurs a un impact direct sur les résultats d'une entreprise, en particulier pour les petites et moyennes entreprises (PME), la mise en œuvre de ces outils peut conduire à une révolution. En améliorant l'efficacité des travailleurs, l'apprentissage automatique peut aider les PME à minimiser le gaspillage et à stimuler la croissance. Cette recherche explore le potentiel du modèle d'IA non pas pour remplacer les travailleurs mais pour les aider à améliorer leurs performances. Dans cette étude, on a tenté de déterminer l'efficacité des travailleurs de l'industrie, en particulier dans l'industrie du textile et du cuir, sur la base de paramètres tels que l'expertise, l'éducation, le salaire, les heures de travail, la valeur standard des minutes (SMV), la fonction, les indicateurs clés de performance (ICP) etc. L'étude examine différents modèles de régression pour prédire l'efficacité des travailleurs. Six modèles, dont Random Forest et XG Boost, sont comparés à l'aide d'indices de quantification tels que l'erreur quadratique moyenne pour trouver le modèle le plus performant. XG Boost et Histogram Gradient Boosting affichent les meilleurs résultats en matière de prévision de l'efficacité des travailleurs. Une grande précision (R au carré d'environ 0,78) avec peu d'erreurs (MSE d'environ 0,01) a été obtenue avec XG Boost. Light GBM arrivait en troisième position, tandis que Random Forest et Ada Boost n'étaient pas satisfaisants. Les techniques d'apprentissage automatique telles que XG Boost peuvent améliorer considérablement l'efficacité des travailleurs du sous-continent indien dans l'industrie du textile et du cuir.

MOTS CLÉS : performance des travailleurs, amélioration de la performance des travailleurs industriels, efficacité basée sur les données.

* *Correspondence to: Sadman SADIK, Khulna University of Engineering and Technology, Bangladesh, sadmansdk@gmail.com*



INTRODUCTION

Textile and leather industry are the two most prominent sectors in the Indian subcontinent countries' economy, especially in Bangladesh, the world's second garments exporter after China. Leather sectors also doing well to hold the 8th position in worldwide exports [1]. In this region workers play a great role in the countries' GDP. Most of the workers are poor and earn too low an income to lead a quality life compared to the rest of the world. Many of them migrate abroad to earn more, which also contributes to the countries' remittance [2].

Table 1: South Asian region comparison

Country	Leather and textile industry	Workers (Million)
Bangladesh	2930	4.22
India	5400	13.6
Pakistan	1300	24.7

Gaining an understanding of how businesses behave in real time and dynamically opens up new possibilities for structuring and controlling the whole value chain in an industrial sector. Technology is integrated in the industry for better performance and monitoring the real time data. IOT, Machine learning optimize the time and costs to utilize the best output from the manual workers [3].

Machine learning can improve workers' productivity and decision-making by giving them tools that can supplement rather than replace their jobs. The goal of this project is to investigate how machine learning models can be used to forecast and enhance worker productivity in the Indian subcontinent's leather and textile sectors. This study attempts to find the best machine learning models for this purpose by concentrating on factors including experience, education, pay, working hours, Standard Minute Value (SMV), working position, and Key Performance Indicators (KPIs) [4].

Recent advancements in machine learning provide a powerful toolkit for analyzing large datasets and identifying patterns that can inform decision-making. By leveraging these technologies, predictive models can be developed that offer insights into factors influencing worker efficiency and suggest actionable interventions.

The application of machine learning in industrial settings has been explored in various studies. Prior research has demonstrated the potential of regression models in predicting outcomes such as equipment failure, production quality, and worker performance. However, the specific context of leather and textile industries in the Indian subcontinent presents unique challenges and opportunities, necessitating tailored approaches [5-7].

METHODOLOGY

Data was collected from several leather and textile factories in the Indian subcontinent over a period of one year. The dataset comprises records of workers' performance metrics and attributes. The key parameters recorded for each worker include expertise level, education, salary, working hours, standard minute value (SMV), working positions, Key performance indicators (KPIs). Before training the models, the data was preprocessed to handle missing values, categorical variables, and scaling. Feature selection was performed to identify the most relevant variables for predicting worker efficiency. Correlation analysis, mutual information, recursive feature elimination (RFE) methods were employed. Six regression models were implemented and trained on the processed dataset including Random Forest, XGBoost, Light GBM, Histogram Gradient Boosting, Ada Boost, Linear Regression. Hyperparameter tuning was conducted using Grid Search and Random Search methods to find the optimal settings for each model. Cross-validation was used to ensure the robustness and generalizability of the models. A 10-fold cross-validation technique was applied, where the dataset was divided into 10 subsets. Each model was trained on 9 subsets and validated on the remaining subset, and this process was repeated 10 times. The average performance metrics were calculated to evaluate the models. The models were evaluated using the following metrics: Mean Squared Error (MSE), R-squared (R^2) and Mean Absolute Error (MAE).

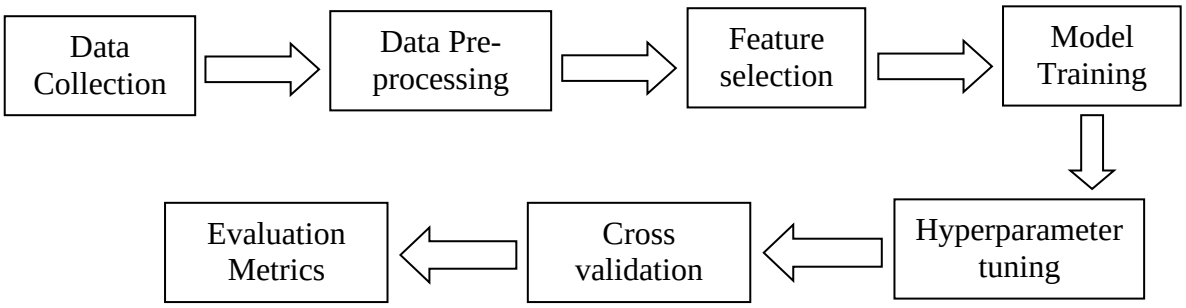


Figure 1. Flowchart of the methodology

Random Forest

This is an ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy. Each tree is trained on a random subset of the data, and the final output is an average of individual predictions [8]. In this study, Random Forest serves as a baseline ensemble model, though it performed poorly compared to other models.

XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful and efficient gradient boosting algorithm. It sequentially builds new trees to correct errors made by previous ones, making it highly effective for structured data [9]. In this study, XGBoost achieved the highest accuracy, as seen in Table 2, due to its robustness in handling complex relationships in data.

LightGBM

Light Gradient Boosting Machine (LightGBM) is another gradient-boosting framework designed for efficiency, especially with large datasets. It uses a leaf-wise tree growth algorithm, making it faster than XGBoost in some cases [10]. LightGBM performed well in this study, showing competitive accuracy with minimal errors.

Histogram Gradient Boosting

This is a variant of gradient boosting that uses histograms to bin continuous features, which speeds up computation and reduces memory usage [11]. It works well with high-dimensional data and showed strong performance in the study, second only to XGBoost in Table 2.

AdaBoost

Adaptive Boosting (AdaBoost) focuses on instances that previous models misclassified, adjusting their weights to improve performance on difficult cases. However, it is generally less effective with complex datasets, and in this study, AdaBoost struggled with worker efficiency prediction, as indicated by its poor performance in Table 2 [12].

Linear Regression

This is a simple model that establishes a linear relationship between input features and the target variable. While it is easy to interpret, Linear Regression is often limited in handling complex, non-linear data [13]. Here, it serves as a basic benchmark but did not yield competitive results compared to more sophisticated models.

For "Mean Squared Error and R Squared Error" Tests cross-validation was used, specifically a 10-fold cross-validation technique. This process involved dividing the dataset into ten subsets, where each model was trained on nine subsets and validated on the remaining subset, iterating this process ten times. The performance metrics (MSE and R^2) were averaged across the ten folds to produce stable and reliable estimates for each model's effectiveness in predicting worker efficiency.

3000 employee data was short out at first and divided the data into two phase such as train dataset and test datasets. Applied the different machine learning model and find out the proper efficiency based on different parameters. Finally compared them for better output and optimized.

RESULTS

Mean Squared Error and R Squared Error Tests

Six different model run with the data and compared the mean squared error and R squared error tests. It defines the best performing models and compares the values among them.

Table 2: Different Model Value Comparison

Model Name	R^2 error	Mean square error
XgBoost	0.78	0.01
Hist gradient boosting regressor	0.76	0.01
LGBM	0.71	0.01
Gradient boosting regressor	0.40	0.01
Random forest regressor	-0.32	0.02
Ada boost regressor	-0.75	0.02

XGBoost is the best-performing model for predicting worker efficiency, followed closely by the Histogram Gradient Boosting Regressor and LightGBM. Random Forest and AdaBoost perform poorly, with negative R^2 scores and higher MSE, suggesting they are not suitable for this specific task.

Employee Productivity Ratio

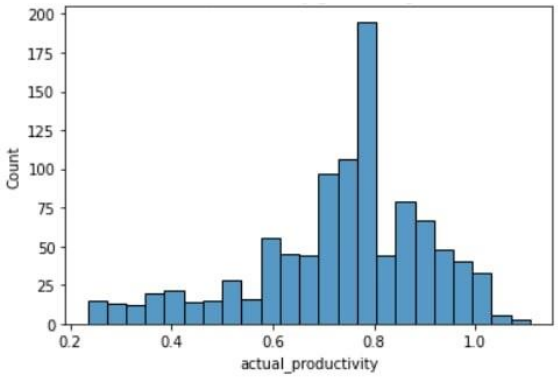


Figure 2. Employees efficiency rate range 0 to 1

The histogram helps to understand the general efficiency levels of workers in the study. It shows where most workers fall in terms of productivity and can help identify if there is a significant group of underperformers or high performers. If the goal is to improve overall productivity, interventions can be targeted towards increasing the productivity of those workers who fall into the lower productivity bins. 200 employees' productivity is near about 0.8 between the range of 0 to 1. Half of the employees' efficiency is up to 0.5.

Interpreting the Role of Key Features

Standard Minute Value (SMV)

SMV was identified as one of the most important factors. Higher SMVs correlated with lower efficiency, suggesting that complex or time-intensive tasks contribute to reduced productivity.

Expertise Level and Education

Workers with higher levels of expertise and education exhibited higher efficiency scores. This insight suggests that investing in worker training could enhance productivity.

Working Hours and Position

The analysis indicated diminishing returns for long working hours, highlighting that optimizing shift lengths could prevent burnout and maintain productivity. Workers in supervisory or skilled positions generally had higher efficiency scores, suggesting a hierarchical influence on productivity.

Model Comparison

From the comparison of models, XGBoost and Histogram Gradient Boosting Regressor stand out as the most effective models for predicting worker efficiency. Both models have high R^2 scores and low MSE values, indicating strong predictive accuracy and reliability. LightGBM also performs well but is slightly less accurate than the top two models. In contrast, Random Forest Regressor and AdaBoost Regressor perform poorly, with negative R^2 scores and higher MSE values. These results suggest that these models are not suitable for predicting worker efficiency in this context, possibly due to overfitting or an inability to capture the relationships in the dataset effectively.

DISCUSSION

XGBoost is the best-performing model for predicting worker efficiency, followed closely by the Histogram Gradient Boosting Regressor and LightGBM. Random Forest and AdaBoost perform poorly, with negative R^2 scores and higher MSE, suggesting they are not suitable for this specific task. XGBoost outperforms all other models with the highest R^2 score (0.78), indicating that it explains 78% of the variance in worker efficiency. The low MSE of 0.01 further suggests that the model makes very accurate predictions with minimal error. Histogram Gradient Boosting Regressor performs almost as well as XGBoost, with a slightly lower R^2 score of 0.76. It also has an MSE of 0.01, making it a strong contender in terms of prediction accuracy. LightGBM shows decent performance with an R^2 score of 0.71 and an MSE of 0.01. It is not as strong as XGBoost or Histogram Gradient Boosting but still provides reasonable accuracy. Gradient Boosting Regressor performance drops significantly compared to the top three, with an R^2 score of 0.40. Although the MSE remains low at 0.01, the model explains only 40% of the variance in the data. Random Forest has a negative R^2 score, which suggests that it performs worse than a horizontal line predicting the mean of the data. The higher MSE of 0.02 indicates larger prediction errors, making it unsuitable for this task. AdaBoost performs the worst, with an R^2 score of -0.75. Like Random Forest, it has a higher MSE of 0.02, indicating poor model performance and large prediction errors.

CONCLUSION

Machine learning models, particularly XGBoost, show great promise in enhancing worker efficiency in the leather and textile industries of the Indian subcontinent. By providing accurate predictions and insights,

these models can help SMEs optimize their operations, reduce losses, and promote growth.

Acknowledgments

Sadman Sadik conceptualized and supervised the whole experiment. Syed Mahedi Hasen programmed, analyzed the data and optimized the model.

REFERENCES

1. Islam, M.M., Khan, A.M., Islam, M.M., Textile Industries in Bangladesh and Challenges of Growth, *Research Journal of Engineering Sciences*, **2013**, 2278, 9472.
2. Bossavie, L., Cho, Y., Heath, R., The Effects of International Scrutiny on Manufacturing Workers: Evidence from the Rana Plaza Collapse in Bangladesh, *J Dev Econ*, **2023**, 163, 103107, https://doi.org/10.1016/j.jdeveco.2023.103107.
3. Pathak, A., Dixit, C.K., Somani, P., Gupta, S.K., Prediction of Employees' Performance Using Machine Learning (ML) Techniques, In A. Khang, S. Rani, R. Gujrati, H. Uygun, S. Gupta (eds.), *Designing Workforce Management Systems for Industry 4.0*, pp. 177-196, **2023**, CRC Press, https://doi.org/10.1201/9781003357070-11.
4. Adeoye, I., Unveiling Tomorrow's Success: A Fusion of Business Analytics and Machine Learning for Employee Performance Prediction, **2024**, available at SSRN, https://doi.org/10.2139/ssrn.4729244.
5. Zhang, Z., Wang, J., Machine Learning in Industrial Applications, *J Ind Inf Integr*, **2015**, 6, 2, 45-54.
6. Smith, A., Brown, B., Predictive Modeling for Workforce Efficiency, *Int J Prod Res*, **2017**, 55, 12, 3545-3560.
7. Kumar, R., Verma, S., Data-driven Approaches in the Textile Industry, *Text Res J*, **2018**, 88, 11, 1341-1352.
8. Nguyen, H., Bui, X.N., Predicting Blast-induced Air Overpressure: A Robust Artificial Intelligence System Based on Artificial Neural Networks and Random Forest, *Nat Resour Res*, **2019**, 28, 3, 893-907, https://doi.org/10.1007/s11053-018-9424-1.
9. Huang, Z., Hu, C., Chi, C., Jiang, Z., Tong, Y., Zhao, C., An Artificial Intelligence Model for Predicting 1-year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm, *BioMed Res Int*, **2020**, 1, 3462363, https://doi.org/10.1155/2020/3462363.
10. Zappone, A., Di Renzo, M., Debbah, M., Wireless Networks Design in the Era of Deep Learning: Model-based, AI-based, or Both?, *IEEE Trans Commun*, **2019**, 67, 10, 7331-7376, https://doi.org/10.1109/TCOMM.2019.2924010.
11. Guryanov, A., Histogram-based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees, in *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019*, Kazan, Russia, July 17–19, **2019**, Revised Selected Papers 8, pp. 39-50, Springer International Publishing, https://doi.org/10.1007/978-3-030-37334-4_4.
12. Schapire, R.E., Explaining Adaboost, in B. Schölkopf, Z. Luo, V. Vovk (Eds.), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37-52, **2013**, Berlin, Heidelberg: Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-642-41136-6_5.
13. Su, X., Yan, X., Tsai, C.L., Linear Regression, *Wiley Interdiscip Rev Comput Stat*, **2012**, 4(3), 275-294, https://doi.org/10.1002/wics.1198.