engineering proceedings

MDPI

Proceeding Paper

Supporting Sustainable Workforce Management for Worker Illness Absence Through Predictive Analytics [†]

Ida Lumintu [1,*] and Achmad Maududie [2]

[1] Department of Industrial Engineering, University of Trunojoyo Madura, Bangkalan 69162, Indonesia
[2] Department of Information System, University of Jember, Jember 68121, Indonesia; maududie@unej.ac.id
[*] Correspondence: ida.lumintu@trunojoyo.ac.id

[†] Presented at the 8th Mechanical Engineering, Science and Technology International Conference, Padang Besar, Perlis, Malaysia, 11–12 December 2024.

Abstract: This study aimed to predict employee sickness absence, vital for sustainable workforce management and organizational productivity. Despite its importance, gaps exist in using advanced machine learning for this purpose. This research developed and validated models—Gradient Boosting, CatBoost, and Random Forest—focused on predictors like health conditions, mental well-being, and work stress. Using a factory worker dataset, this study conducted feature engineering, causal inference, and model performance evaluation. Random Forest proved especially effective in predicting absence, with key factors including recent performance and health. The findings support targeted interventions and efficient resource allocation, promoting sustainable business practices.

Keywords: workforce management; organizational productivity; predictive analytics

1. Introduction

Predicting employee illness absence is a critical aspect of sustainable workforce management, impacting both organizational productivity and employee well-being. Absenteeism due to illness can disrupt operations, increase costs, and negatively affect morale. Therefore, accurately forecasting illness-related absences is essential for developing effective management strategies. Despite its importance, the current literature reveals significant research gaps, particularly in the application of advanced machine learning techniques specifically tailored to forecast sickness-related absenteeism among employees.

Existing research has explored various factors influencing sickness absence. For instance, Norder et al. [1] developed a predictive rule for sickness absence caused by common mental disorders, demonstrating fair discrimination at different time points after reporting sicknesses. Similarly, Heo et al. [2] identified job stress factors such as high job demand and organizational injustice as significant predictors of absence due to accidents and illnesses. Eriksen [3] found that low social support at work could be a predictor of sickness absence across different employee groups, emphasizing the critical role of the workplace environment in absenteeism.

Further studies have highlighted the importance of considering mental health in predicting illness absence. Lamichhane et al. [4] pointed out that symptoms of depression are a substantial risk factor for future absenteeism among manufacturing workers. Plaat et al. [5] discussed the impact of the COVID-19 pandemic on sickness absence due to mental health issues, illustrating how external factors can influence absenteeism patterns. Additionally, Boot et al. [6] highlighted a combination of factors such as age, working conditions, pregnancy, and previous absence history as predictors of long-term sickness

check for updates

Academic Editors: Noor Hanita Abdul Majid, Agus Dwi Anggono, Waluyo Adi Siswanto, Tri Widodo Besar Riyadi, Mohammad Sukri Mustapa, Nur Rahmawati Syamsiyah and Afif Faishal

Published: 28 January 2025

Citation: Lumintu, I.; Maududie, A. Supporting Sustainable Workforce Management for Worker Illness Absence Through Predictive Analytics. *Eng. Proc.* **2025**, *84*, 17. https://doi.org/10.3390/engproc2025084017

absence. Munir et al. [7] emphasized the role of organizational support in managing chronic illness to prevent prolonged absences.

While these studies provide valuable insights, they primarily focus on individual predictors or specific conditions, and there is a noticeable scarcity of research integrating these factors into comprehensive machine learning models for illness absence prediction. The majority of existing studies have concentrated on turnover and attrition prediction [8–15], job satisfaction [16,17], and mental health support [18–21]. However, illness-related absenteeism involves unique factors that require distinct consideration, such as physical health conditions, mental well-being, and work-related stressors.

This study aims to fill this research gap by developing and validating machine learning models tailored for predicting worker illness absence. By incorporating a comprehensive set of predictors, including physical health conditions, mental well-being, work-related stressors, and organizational factors, this study seeks to provide actionable insights for managing absenteeism more effectively. The urgency of this research is underscored by ongoing global health crises, particularly the COVID-19 pandemic, which has significantly impacted workforce dynamics and highlighted the need for effective absenteeism management strategies [5,21].

The sustainability considerations of this study are manifold. The effective management of worker illness absence contributes to resource optimization, enhanced productivity, and improved employee well-being. Predicting illness-related absences allows organizations to optimize resource allocation, streamline workforce planning, and minimize disruptions, thereby promoting sustainable business practices [22,23]. Accurate prediction also facilitates better workforce planning and resource allocation, ensuring operational continuity even during periods of high absenteeism [24]. Additionally, it can lead to significant cost savings by decreasing the financial impact of unplanned absences, such as overtime costs and temporary staffing expenses [25].

Implementing machine learning models for illness absence prediction can also enhance employee well-being and satisfaction. By proactively addressing health-related factors contributing to absenteeism, organizations can create a healthier work environment, support employee health, and enhance overall job satisfaction, resulting in increased employee retention rates and improved morale [26]. Furthermore, the ability to forecast and manage illness-related absences effectively strengthens organizational resilience. Businesses can adapt to changing circumstances, such as public health crises or seasonal illness outbreaks, ensuring the continuity of operations and minimizing disruptions [27].

This study leverages open data on factory workers' daily performance and attrition, available from Gladden [28], to develop and validate predictive models for illness absence. By utilizing these data, this study aims to uncover patterns and trends in employee absenteeism, providing a foundation for developing targeted interventions and policies that support sustainable workforce management. Through advanced predictive analytics, this research seeks to enhance organizational efficiency, promote employee well-being, and contribute to the overall sustainability of business operations.

2. Materials and Methods

2.1. Data Collection

The dataset used in this study include 18 months of daily performance and attrition records for a factory's workforce, comprising 508 workers with a total of 687 individuals appearing in the dataset due to employee turnover [28]. This synthetic dataset includes 411,948 observations, detailing both regular daily events, such as attendance and efficacy, and special one-time events like accidents, terminations, and the onboarding of new employees. A unique aspect of this dataset is the diverse causal relationships embedded within the data, which are ripe for discovery through machine learning.

Each row in the dataset represents a specific event occurring on a particular day for a specific worker, with 14 different types of events captured. These events include presence, absence, efficacy, resignation, termination, onboarding, idea generation, mental lapses, physical feats, slips, teamwork, disruptions, sacrifices, and sabotage. The dataset also contains fields related to both the subject (worker) and their supervisor, as well as details about the event itself. This comprehensive dataset, prepared using Synaptans WorkforceSim version 0.3.15, provides a rich source of information for analyzing employee behavior and predicting absenteeism due to illness.

2.2. Data Preparation

The dataset was preprocessed using Python 3.10.9, converting 'event_date' to datetime format and handling missing values by filling numeric columns with the mean and categorical columns with the mode. Duplicate rows were removed to ensure data integrity. The cleaned data were saved as 'preprocessed.csv' for the feature engineering and predictive modeling of employee absence due to illness.

2.3. Feature Engineering

Feature engineering was performed to extract relevant features related to predicting worker sickness based on mental lapses or physical accidents. The process began with loading the preprocessed data from the preprocessed.csv file, which had undergone initial cleaning processes such as handling missing values, converting date formats, and removing duplicates. The primary goal was to create the target variable 'sickness', indicating whether a worker missed work due to illness within the next seven days.

This goal was achieved using Python's .shift(−7) function to flag a forthcoming 'Absence' event by setting a 'sickness' variable to 1 if detected, otherwise 0. Specific features included the following: (a) Lag Features—lapse_last_week and slip_last_week, counting the past week's 'Lapse' and 'Slip' events using a seven-day rolling window; (b) Trend Feature—efficacy_trend, averaging weekly worker efficacy; (c) Data Quality—dropping NaN rows from rolling window operations; and (d) Selected Columns—target variable sickness, new features, and relevant worker attributes for analysis and modeling.

2.4. Causal Inference Analysis

To understand the causal relationships between various factors and worker illness absence, we performed a causal inference analysis using the dowhy library. This analysis involved creating a causal model, identifying the effect of interest, and estimating this effect. The feature-engineered dataset served as the basis for this analysis. The key variables considered as potential causes of sickness included lapse_last_week, slip_last_week, and efficacy_trend, while sub_health_h, sub_commitment_h, sub_perceptiveness_h, sub_dexterity_h, sub_sociality_h, and sub_goodness_h were treated as confounders. The causal model was defined with these variables, and the resulting causal graph was visualized to illustrate relationships.

Once the causal model was established, we proceeded to identify the causal effects of the variables of interest on sickness. This involved specifying the causal estimand and using a statistical method, such as linear regression, to estimate the effects. The 'dowhy' library in Python facilitated these steps, enabling us to derive the causal estimate. Additionally, to ensure the robustness of our findings, we performed a refutation test using the placebo_treatment_refuter method. This comprehensive causal inference analysis provided valuable insights into how factors like mental lapses, physical slips, and efficacy trends impact worker sickness, thereby providing strategies for better workforce management.

Feature descriptions:

- lapse_last_week counts the number of severe mental mistakes made by a worker in the past week.
- slip_last_week counts the number of physical accidents or missteps experienced by a worker in the past week.
- efficacy_trend represents the average productivity of a worker over the past week.
- sub_health_h: a confounder indicating the overall health status of a worker.
- sub_commitment_h: a confounder representing the level of commitment a worker has towards their job.
- sub_perceptiveness_h: a confounder that measures a worker's ability to perceive and understand their work environment.
- sub_dexterity_h: a confounder indicating the physical skill and agility of a worker.
- sub_sociality_h: a confounder reflecting the social behavior and teamwork abilities of a worker.
- sub_goodness_h: a confounder measuring the moral and ethical behavior of a worker.

2.5. Data Imbalance Analysis and Selection of Predictive Methods

To determine whether the feature-engineered dataset is imbalanced, we examined the distribution of the target variable, sickness. This involved loading the dataset and inspecting the number of instances belonging to each class. Upon running the analysis, we found that the dataset contains 405,508 instances of workers not being sick (class 0) and only 6440 instances of workers being sick (class 1), resulting in a severe imbalance with an imbalance ratio of approximately 63:1. This significant disparity indicates that the dataset is heavily imbalanced, which can potentially bias predictive models towards the majority class if not addressed properly.

Given this severe imbalance, we selected three predictive methods that are well suited to handle such data characteristics: Gradient Boosting, CatBoost, and Random Forest. These algorithms were chosen because of their robustness in dealing with imbalanced datasets. Gradient Boosting focuses on minimizing errors through sequential model building, CatBoost is specifically designed to handle categorical features and class imbalance, and Random Forest uses ensemble learning with class weighting to ensure that minority classes are well represented during training. By leveraging these methods, we aimed to achieve more accurate and reliable predictions of worker sickness despite the significant class imbalance in our dataset.

2.6. Modeling

To predict worker sickness amid class imbalance, we employed Gradient Boosting, CatBoost, and Random Forest models, chosen for their ability to manage imbalanced datasets and yield reliable predictions. Gradient Boosting iteratively builds decision trees to correct previous errors, enhancing the model's focus on hard-to-classify cases, which is especially useful for imbalanced data. Model performance was evaluated using a confusion matrix, classification report, and metrics like accuracy, precision, recall, and F1-score for comprehensive assessment.

CatBoost, designed for handling categorical features and imbalanced data, was also utilized. Its method of transforming categorical variables into numerical representations without extensive preprocessing, alongside balanced objectives, made it particularly effective for our dataset. The model's accuracy and reliability were assessed through similar performance metrics, including a confusion matrix, classification report, and precision and recall metrics.

Finally, we selected Random Forest for its ensemble approach and class-weighting capabilities, which help balance predictions in favor of the minority class. By aggregating the results from multiple decision trees, Random Forest minimizes bias toward the majority class, increasing prediction accuracy for sickness cases. The model's predictive strength was evaluated using the same thorough performance metrics, providing insights into its reliability for identifying worker sickness patterns.

2.7. Assessment of Classification Models' Performance

A confusion matrix is a crucial tool for assessing the performance of a classification model. It provides a detailed comparison between the actual and predicted values, allowing for the calculation of various performance metrics such as accuracy, precision, recall, and F1-score.

The structure of the confusion matrix for a binary classification problem is illustrated in Table 1. As depicted, the confusion matrix consists of four components: TPs (true positives, correctly predicted positive instances), TNs (true negatives, correctly predicted negative instances), FPs (false positives, negative instances incorrectly predicted as positive), and FNs (false negatives, positive instances incorrectly predicted as negative).

Table 1. The structure of a confusion matrix.

	Predicted Positive	**Predictive Negative**
Actual Positive	TPs (True Positives)	FNs (False Negatives)
Actual Negative	FPs (False Positives)	TNs (True Negatives)

The models' performance metrics were derived from the confusion matrix as follows:

1. Accuracy: Accuracy measures the proportion of correctly identified positive and negative cases out of the total instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

2. Precision (Positive Predictive Rate): Precision measures the percentage of correct positive predictions out of all the positive predictions made by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

3. Recall (Sensitivity or True Positive Rate): Recall measures the proportion of actual positive instances that the model accurately identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4. F1-Score: The F1-score, which is the harmonic mean of precision and recall, provides a balanced metric that accounts for both false positives and false negatives.

$$\text{F} - 1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

2.8. Feature Importance Analysis

Feature importance calculations were performed to determine the most critical features for predicting workers' illness-related absences. This process involved computing importance scores for each feature and ranking them by their significance. The insights gained from this analysis offer valuable information on the key factors influencing absences, enabling more targeted and effective workforce management strategies.

For Gradient Boosting, feature importance can be calculated using the mean decrease in impurity (MDI) or the mean decrease in accuracy (MDA). In this study, we focus on the MDI approach. The importance score for a feature j in Equation (5) is computed as the total reduction in the impurity (e.g., Gini impurity or entropy) brought by that feature, averaged over all the trees in the ensemble.

$$\mathrm{FI(j)} = \sum_{t=1}^{T}\sum_{n\in \mathrm{nodes}(t,j)} \frac{N_n}{N}\Delta I_n \quad (5)$$

where
$\mathrm{FI}(j)$: feature importance for feature j.
T : total number of trees.
$\mathrm{nodes}(t,j)$: nodes in tree t where feature j is used for splitting.
N_n : number of samples that reach node n.
N : total number of samples.
ΔI_n : impurity decrease at node n due to feature j.

CatBoost provides a similar approach to feature importance, calculating it based on the average decrease in loss due to splits on the feature, weighted by the number of samples passing through the splits. The importance score for a featurej in Equation (6) is calculated by the average prediction value change (PVC) when the feature is used for splitting, normalized by the total prediction value change for all features.

$$\mathrm{FI(j)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n\in \mathrm{nodes}(t,j)} \left(\frac{\mathrm{PVC}_{n,j}}{\sum_{k=1}^{M}\mathrm{PVC}_{n,k}}\right) \quad (6)$$

where
$\mathrm{FI}(j)$: feature importance for feature j.
T : total number of trees.
$\mathrm{nodes}(t,j)$: nodes in tree t where feature j is used for splitting.
M : total number of features.
$\mathrm{PVC}_{(n,j)}$: prediction value change at node n due to feature j.

For Random Forest, feature importance can be computed using the mean decrease in impurity (MDI) and the mean decrease in accuracy (MDA). The most common approach is the use of the mean decrease in impurity (MDI), also known as Gini Importance. The feature importance score for a featurej in Equation (7) is computed by averaging the total decrease in node impurity (e.g., Gini impurity or entropy) brought by that feature, across all the trees in the forest.

$$\mathrm{FI(j)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n\in \mathrm{nodes}(t,j)} \frac{N_n}{N}\Delta I_n \quad (7)$$

where
$\mathrm{FI}(j)$: feature importance for feature j.
T : total number of trees.
$\mathrm{nodes}(t,j)$: nodes in tree t where feature j is used for splitting.
N_n : number of samples that reach node n.
N : total number of samples.
ΔI_n : impurity decrease at node n due to feature j.

When explaining feature importance in machine learning models to business and management professionals, it is essential to use simplified terminology. A crucial concept is impurity, which measures how well the data are grouped according to the target variable (e.g., workers' absence due to illness). In predictive models, impurity helps evaluate how effectively splits in the data separate different outcomes. The two main types of impurity are Gini impurity and entropy.

- Gini Impurity: This measures the probability of misclassifying an observation. It ranges from 0 (perfectly pure) to 0.5 (completely mixed). For instance, if 90% of the workers in a group did not miss work due to illness, the Gini impurity would be low, indicating a well-separated group.
- Entropy: This measures the level of disorder or uncertainty. It ranges from 0 (perfectly pure) to 1 (completely mixed). For example, if a group has an equal number of workers who missed work and those who did not, it has high entropy, indicating high disorder.

The reduction in impurity, also known as information gain, determines the importance of a feature. If splitting the data based on a particular feature, such as mental lapses, significantly reduces impurity, this indicates that this feature is crucial for predicting worker absence.

By grasping these concepts, business and management professionals can better understand how predictive models identify the most critical factors in predicting outcomes, such as workers' absence, leading to more informed decision-making and optimized workforce management strategies.

3. Results and Discussion

The objective of this study was to develop and evaluate predictive models for identifying worker sickness absence using machine learning techniques, particularly in the context of a severely imbalanced dataset. We utilized Gradient Boosting, CatBoost, and Random Forest methods to achieve this goal. This section details the causal model used for prediction, analyzes the performance metrics of the predictive models, and explores the feature importance in predicting worker illness absence.

3.1. Causal Model in Predicting Workers' Illness Absence

The causal inference analysis utilized the dowhy library to establish a causal model that identified the potential effects of lapse_last_week, slip_last_week, and efficacy_trend on sickness, considering sub_health_h, sub_commitment_h, sub_perceptiveness_h, sub_dexterity_h, sub_sociality_h, and sub_goodness_h as confounders. The detailed causal relationships are illustrated in Figure 1.

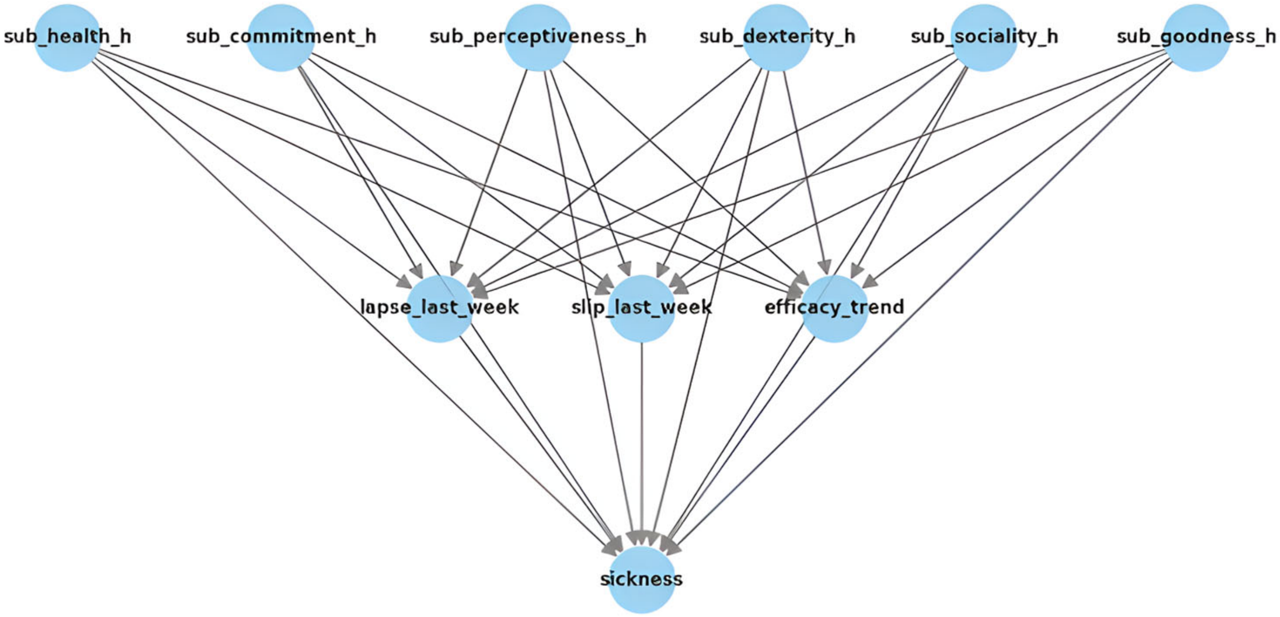

Figure 1. Causal model.

The detailed causal relationships can be explained as follows:

1. sub_health_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   A worker's health status strongly impacts their mental lapses, physical slips, and overall efficacy, directly correlating with increased sickness absence. From a business perspective, effectively monitoring and promoting employee health can help reduce absences, boosting productivity and overall workplace efficiency.
2. sub_commitment_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   A higher level of worker commitment reduces mental lapses and physical slips, boosts efficacy, and lowers the risk of sickness absence. From a business perspective, fostering commitment through engagement and incentives can minimize errors, enhance performance, and decrease absenteeism.
3. sub_perceptiveness_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   Workers with high perceptiveness tend to make fewer mental errors and are generally more efficient, leading to a reduced risk of sickness-related absences. From a business perspective, investing in training programs to enhance worker perceptiveness can decrease mistakes, boost productivity, and lower the rates of illness-related absenteeism.
4. sub_dexterity_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   Higher dexterity in workers was found to reduce physical slips and enhance efficacy, leading to fewer sickness absences. From a business perspective, investing in ergonomic solutions and physical training can boost dexterity, decrease accidents, and improve overall performance.
5. sub_sociality_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   Workers with high sociality generally have better teamwork and communication skills, leading to fewer lapses and slips, improved efficacy, and reduced sickness absence. From a business perspective, fostering a collaborative work environment can minimize errors, boost performance, and decrease absenteeism.
6. sub_goodness_h → lapse_last_week, slip_last_week, efficacy_trend, sickness:
   A worker's intrinsic goodness or ethical behavior positively impacts productivity and lowers the risk of lapses, slips, and sickness absence. From a business perspective, promoting a culture of integrity and ethical conduct can enhance worker reliability and help reduce absenteeism.
7. lapse_last_week → sickness:
   Frequent mental lapses in the prior week suggest potential health or cognitive issues that may lead to sickness absence. From a business perspective, recognizing such patterns allows for early intervention and support, helping to prevent further health decline and reduce absenteeism.
8. slip_last_week → sickness:
   Physical slips or accidents from the previous week are strong indicators of potential health problems leading to sickness absence. From a business perspective, addressing these safety issues and providing adequate physical support can help reduce workplace accidents and, in turn, lower the incidence of sickness-related absences.
9. efficacy_trend → sickness:
   A declining trend in work efficacy over the past week may indicate underlying health issues, potentially leading to sickness absence. From a business perspective, monitoring these efficacy trends allows for the early detection of health problems, enabling proactive support measures that help maintain worker health and sustain productivity.

3.2. Evaluation of Performance Metrics

The confusion matrices for the Gradient Boosting, CatBoost, and Random Forest models are illustrated in Figure 2, Figure 3, and Figure 4, respectively. The performance metrics derived from these models are summarized in Table 2.

Table 2. The performance metrics of the models.

	Gradient Boosting	**CatBoost**	**Random Forest**
Accuracy	0.539786381843	0.7155358660032	0.9724845248209
Precision	0.017746577134	0.0177465771344	0.0278330019881
Recall	0.505694760820	0.3090356871678	0.0212604403948
F1-Score	0.033937170374	0.0335656261597	0.0241067585019

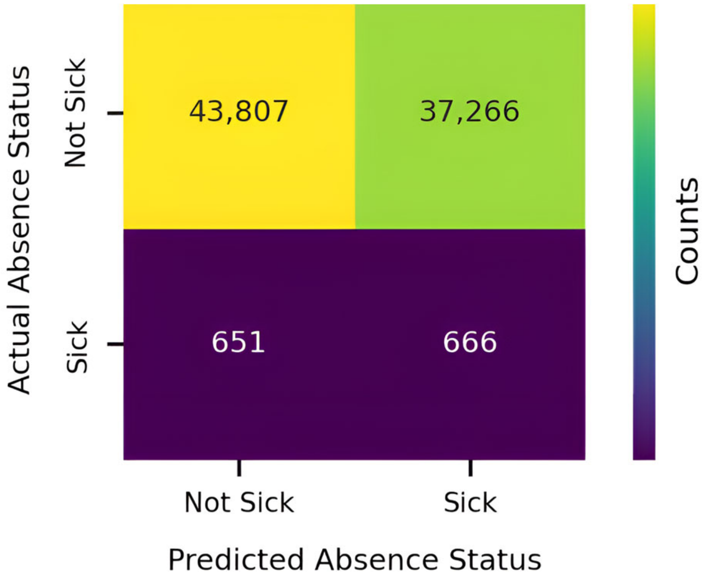

Figure 2. Confusion matrix for Gradient Boosting model.

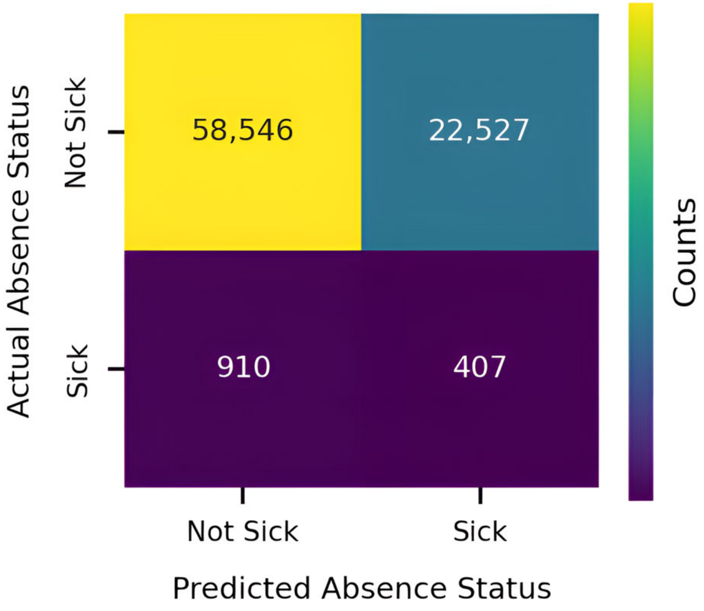

Figure 3. Confusion matrix for CatBoost model.

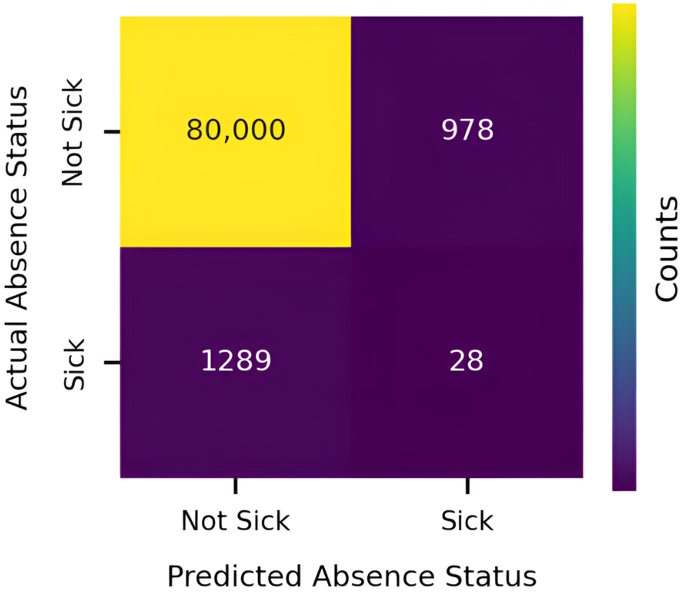

Figure 4. Confusion matrix for Random Forest model.

The accuracy of the Gradient Boosting model was 53.98%, indicating moderate reliability in predicting worker sickness absence. The CatBoost model improved on this with an accuracy of 71.55%, demonstrating its effectiveness in handling the imbalanced dataset. The Random Forest model achieved the highest accuracy of 97.25%, underscoring its robustness and superior performance. High accuracy in these models translates to fewer misclassifications, which is crucial for effective workforce management and operational efficiency.

The precision for both the Gradient Boosting and CatBoost models was 1.77%, reflecting a very low proportion of true positive predictions out of the total number of positive predictions. This indicates a high rate of false positives, which could lead to unnecessary interventions. The Random Forest model had a slightly higher precision of 2.78%, showing some improvement but still indicating a significant number of false positives. High precision is critical for minimizing unnecessary actions, which is essential for maintaining resource efficiency and reducing operational costs.

The recall for the Gradient Boosting model was 50.57%, demonstrating a strong capability to identify actual sickness absences. The CatBoost model had a recall of 30.90%, showing a moderate ability to detect true positives. The Random Forest model, however, had a recall of only 2.13%, indicating a significant number of missed sickness absences. High recall ensures that the model captures most of the true sickness cases, which is vital for timely interventions and maintaining workforce health.

The F1-score for the Gradient Boosting model was 3.39%, balancing precision and recall. The CatBoost model had an F1-score of 3.36%, indicating a slightly lower but comparable performance. The Random Forest model had an F1-score of 2.41%, reflecting its lower overall performance in balancing precision and recall. A high F1-score is crucial for reliable predictions, which contribute to effective absence management and strategic workforce planning.

From a business perspective, these performance metrics are vital for cost savings and operational benefits. Accurate and reliable predictions minimize unnecessary interventions and ensure timely support for workers, leading to significant cost savings and improved resource allocation. For instance, minimizing false positives and maximizing true positive identifications can enhance workforce productivity and reduce absenteeism-related costs. Operational benefits include better workforce planning, reduced downtime, and optimized resource management, collectively contributing to improved business performance, competitiveness, and sustainability.

3.3. Feature Importance Analysis in Predicting Workers' Absence Due to Illness

Feature importance analysis provides critical insights into which variables most significantly impact the prediction of workers' sickness absence. By understanding these features, businesses can prioritize interventions and allocate resources more effectively, leading to enhanced operational efficiency and better workforce management. The following discussion analyzes the feature importance for the Gradient Boosting, CatBoost, and Random Forest models.

The feature importance for the Gradient Boosting model is depicted in Figure 5. The most critical feature is efficacy_trend, which highlights the importance of a worker's recent performance trends in predicting sickness absence. This suggests that the consistent monitoring of a worker's efficacy can provide early warning signs of potential health issues. The sub_health_h feature is the next most important, indicating that overall health status significantly influences absenteeism. Sub_commitment_h and sub_sociality_h are also crucial, emphasizing the role of worker engagement and social interactions in maintaining regular attendance. The features sub_dexterity_h and sub_goodness_h, while less significant, still contribute to the model by highlighting the importance of physical agility and ethical behavior. Lapse_last_week and slip_last_week have the lowest importance, suggesting that while recent lapses and slips are relevant, they are less predictive compared to overall health and performance trends.

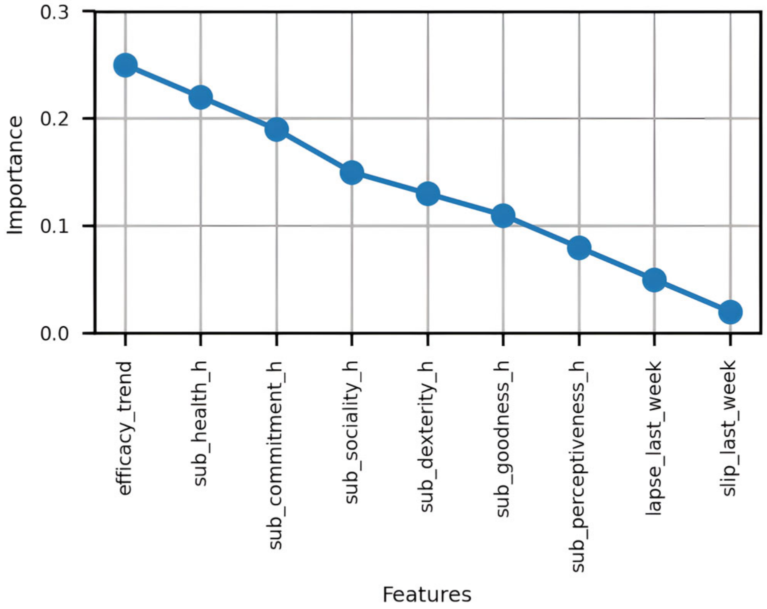

Figure 5. Feature importance for Gradient Boosting model.

In Figure 6, the CatBoost model's feature importance analysis similarly identifies efficacy_trend as the most significant predictor of sickness absence, reinforcing the importance of monitoring recent performance trends. Sub_commitment_h follows, underlining the significance of worker commitment in predicting absences. Sub_dexterity_h, sub_health_h, and sub_sociality_h also play essential roles, indicating that physical dexterity, health status, and social behavior are critical factors. Sub_goodness_h and sub_perceptiveness_h further highlight the relevance of ethical behavior and perceptiveness in predicting absences. Lapse_last_week and slip_last_week are again the least important features, suggesting that immediate past incidents are less influential than long-term health and performance indicators.

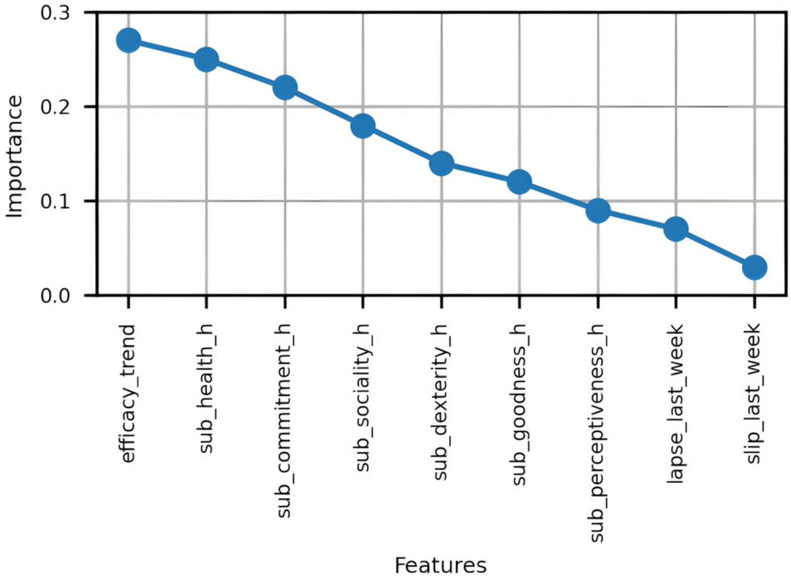

Figure 6. Feature importance for CatBoost model.

The Random Forest model's feature importance, shown in Figure 7, provides a slightly different perspective. Efficacy_trend remains the top predictor, emphasizing its consistent importance across models. However, the importance of sub_health_h is more pronounced in this model, indicating a stronger focus on overall health status. Sub_commitment_h and sub_dexterity_h are also significant, aligning with previous models in highlighting the importance of commitment and physical agility. Sub_sociality_h and sub_goodness_h continue to be relevant, while sub_perceptiveness_h also shows its importance. Lapse_last_week and slip_last_week have minimal importance, consistent with the other models, underscoring that while recent incidents matter, they are not as predictive as broader health and performance trends.

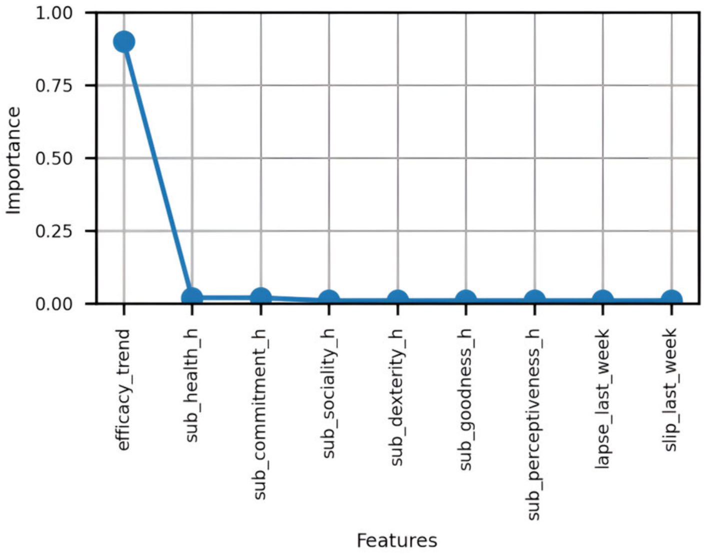

Figure 7. Feature importance for Random Forest model.

From a business and management perspective, these insights into feature importance offer several actionable benefits. Prioritizing interventions based on efficacy trends and overall health status allows for more targeted support and proactive health management, which can reduce absenteeism. Focusing on improving worker commitment, social interactions, and physical dexterity can enhance engagement and reduce the likelihood of sickness absence. Additionally, while immediate past incidents like lapses and slips are less critical, they should not be entirely ignored as they provide supplementary information. Leveraging these predictive insights enhances operational efficiency by reducing unplanned absences, ensuring continuous productivity, and optimizing resource allocation. This proactive approach to workforce management leads to substantial cost savings and supports strategic decision-making, contributing to overall organizational success.

4. Conclusions

This study highlighted the effectiveness of using predictive models to manage worker sickness absence in a factory environment. This study's significance lies in its potential to transform workforce management strategies through AI-powered analytics and real-time data monitoring, leading to enhanced operational efficiency, reduced absenteeism, and lowered associated costs. This research addressed the critical issue of optimizing workforce management to prevent productivity losses, thereby ensuring continuous operations and cost savings.

By integrating causal inference with predictive models, this study provides valuable insights into the factors influencing worker sickness absence. The confusion matrices for the Gradient Boosting, CatBoost, and Random Forest models demonstrate significant operational benefits, including accurate sickness absence predictions, timely interventions, and the efficient management of false positives and negatives. Accurately predicting worker sickness ensures minimal operational disruption, thereby enhancing overall efficiency.

Feature importance analysis emphasizes the need to prioritize interventions based on key factors such as recent performance trends, overall health status, and worker commitment. This targeted approach ensures timely support, optimized resource allocation, and the prevention of unexpected absences. The insights from the predictive models inform strategic decision-making, leading to improved workforce scheduling, reduced unplanned absences, and considerable cost savings.

The performance metrics illustrate the financial benefits of predictive workforce management models. The high accuracy of predictions shows their effectiveness in avoiding unnecessary interventions and maintaining productivity. The potential cost savings from these models underscore the significant financial advantages of adopting advanced analytical methods.

This study is pertinent to supporting sustainable industries by promoting efficient workforce management and resource optimization. Future research could improve predictive model accuracy and robustness by incorporating sophisticated machine learning methods such as deep learning and reinforcement learning. Additionally, exploring predictive workforce management in other sectors, such as financial services or manufacturing, could offer broader insights into its benefits. Developing real-time predictive systems to continuously monitor worker conditions and provide instant recommendations would further enhance the effectiveness of these models. Ultimately, investigating the impact of predictive workforce illness absence models on sustainable industries can provide valuable insights for business practices and resource optimization.

Author Contributions: Conceptualization, I.L. and A.M.; methodology, I.L. and A.M.; Python code, I.L. and A.M.; validation, I.L. and A.M.; formal analysis, I.L.; investigation, I.L. and A.M.; resources, I.L.; data curation, I.L.; writing—original draft preparation, I.L.; writing—review and editing, I.L. and A.M.; visualization, A.M.; supervision, I.L.; project administration, I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available at https://matthewgladden.net/factory-workers-daily-performance-and-attrition/ (accessed on 1 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Norder, G.; Roelen, C.A.M.; van der Klink, J.J.L.; Bültmann, U.; Sluiter, J.K.; Nieuwenhuijsen, K. External validation and update of a prediction rule for the duration of sickness absence due to common mental disorders. *J. Occup. Rehabil.* **2016**, *27*, 202–209. [CrossRef] [PubMed]
2. Heo, Y.-S.; Leem, J.-H.; Park, S.-G.; Jung, D.-Y.; Kim, H.-C. Job stress as a risk factor for absences among manual workers: A 12-month follow-up study. *Ind. Health* **2015**, *53*, 542–552. [CrossRef] [PubMed]
3. Eriksen, W. Work factors as predictors of sickness absence: A three month prospective study of nurses' aides. *Occup. Environ. Med.* **2003**, *60*, 271–278. [CrossRef]
4. Lamichhane, D.K.; Heo, Y.S.; Kim, H.C. Depressive symptoms and risk of absence among workers in a manufacturing company: A 12-month follow-up study. *Ind. Health* **2018**, *56*, 187–197. [CrossRef]
5. Van der Plaat, D.A.; Edge, R.; Coggon, D.; van Tongeren, M.; Muiry, R.; Parsons, V.; Cullinan, P.; Madan, I. Impact of COVID-19 pandemic on sickness absence for mental ill health in National Health Service Staff. *BMJ Open* **2021**, *11*, e054533. [CrossRef]
6. Boot, C.R.L.; van Drongelen, A.; Wolbers, I.; Hlobil, H.; van der Beek, A.J.; Smid, T. Prediction of long-term and frequent sickness absence using company data. *Occup. Med.* **2017**, *67*, 176–181. [CrossRef]
7. Munir, F.; Yarker, J.; Haslam, C. Sickness absence management: Encouraging attendance or "risk-taking" presenteeism in employees with chronic illness? *Disabil. Rehabil.* **2008**, *30*, 1461–1472. [CrossRef]
8. Kanuto, A.E. Identifying patterns and predicting employee turnover using machine learning approaches. *IJSB* **2024**, *36*, 20–35. [CrossRef]
9. Ma, X.; Zhai, S.; Fu, Y.; Lee, L.Y.; Shen, J. Predicting the occurrence and causes of employee turnover with machine learning. *ComEngApp J.* **2019**, *8*, 217–227. [CrossRef]
10. Chaudhary, M.; Gaur, L.; Jhanjhi, N.; Masud, M.; Aljahdali, S. Envisaging employee churn using MCDM and machine learning. *Intell. Autom. Soft Comput.* **2022**, *33*, 1009–1024. [CrossRef]
11. Khera, S.N.; Divya. Predictive modelling of employee turnover in Indian IT industry using machine learning techniques. *Vis. J. Bus. Perspect.* **2019**, *23*, 12–21. [CrossRef]
12. Raza, A.; Munir, K.; Almutairi, M.; Younas, F.; Fareed, M.M.S. Predicting employee attrition using machine learning approaches. *Appl. Sci.* **2022**, *12*, 6424. [CrossRef]
13. Yin, Q. Comparison of machine learning models for employee turnover prediction. *Appl. Comput. Eng.* **2023**, *8*, 228–232. [CrossRef]
14. Liao, C. Employee turnover prediction using machine learning models. In Proceedings of the International Conference on Mechatronics Engineering and Artificial Intelligence (MEAI 2022), Changsha, China, 11–13 November 2022.
15. Ozmen, E.P.; Ozcan, T. A novel deep learning model based on convolutional neural networks for employee churn prediction. *J. Forecast.* **2021**, *41*, 539–550. [CrossRef]
16. Rustam, F.; Ashraf, I.; Shafique, R.; Mehmood, A.; Ullah, S.; Sang Choi, G. Review prognosis system to predict employees job satisfaction using deep neural network. *Comput. Intell.* **2021**, *37*, 924–950. [CrossRef]
17. Gupta, A.; Chadha, A.; Tiwari, V.; Varma, A.; Pereira, V. Sustainable training practices: Predicting job satisfaction and employee behavior using machine learning techniques. *Asian Bus. Manag.* **2023**, *22*, 1913–1936. [CrossRef]
18. Jamalirad, H.; Jajroudi, M. *Prediction of Mental Health Support of Employee Perceiving by Using Machine Learning Methods*; IOS Press: Amsterdam, The Netherlands, 2023; Volume 302, pp. 903–904.
19. Musanga, V.; Chibaya, C. A predictive model to forecast employee churn for HR analytics. In Proceedings of the NEMISA Digital Skills Conference 2023, Durban, South Africa, 15–17 February 2023; Volume 5, pp. 17–30. [CrossRef]
20. Wardhani, F.H.; Lhaksmana, K.M. Predicting employee attrition using logistic regression with feature selection. *Sinkron* **2022**, *7*, 2214–2222. [CrossRef]
21. Bandyopadhyay, N.; Jadhav, A. Churn prediction of employees using machine learning techniques. *Teh. Glas* **2021**, *15*, 51–59. [CrossRef]
22. Jung, H.; Jeon, J.; Choi, D.; Park, J.-Y. Application of machine learning techniques in injection molding quality prediction: Implications on sustainable manufacturing industry. *Sustainability* **2021**, *13*, 4120. [CrossRef]

23. Vrchota, J.; Pech, M.; Rolínek, L.; Bednář, J. Sustainability outcomes of green processes in relation to Industry 4.0 in manufacturing: Systematic review. *Sustainability* **2020**, *12*, 5968. [CrossRef]
24. Safuan, H.A.J.; Abubakar, Y.; Hussain, K. Efficient disease prediction framework to suggest early treatment decisions in healthcare. *J. Electr. Syst.* **2024**, *20*, 687–691. [CrossRef]
25. Syah, R.B.Y.; Muliono, R.; Siregar, M.A.; Elveny, M. An Efficiency Metaheuristic Model to predicting customers churn in the business market with machine learning-based. *IAES Int. J. Artif. Intell.* **2024**, *13*, 1547. [CrossRef]
26. Mourad, Z.; Noura, A.; Mohamed, C.; Abdelhamid, B. Enhancing employee performance management. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 1002–1012. [CrossRef]
27. Kannan, R.; Abdul Halim, H.A.; Ramakrishnan, K.; Ismail, S.; Wijaya, D.R. Machine learning approach for predicting production delays: A quarry company case study. *J. Big Data* **2022**, *9*, 94. [CrossRef] [PubMed]
28. Gladden, M. Factory Workers' Daily Performance & Attrition. Available online: https://matthewgladden.net/factory-workers-daily-performance-and-attrition/ (accessed on 13 July 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.