Predicting Health-Related Absenteeism with Machine Learning: A Case Study

Aleksander Piciga
ap7377@student.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Matjaž Kukar
matjaz.kukar@fri.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Abstract

Health-related absenteeism, or sick leave, is a complex issue with significant financial and operational implications for businesses. We present a machine learning approach to predict employee absenteeism in a Slovenian company. The study involved preprocessing and augmenting the dataset by incorporating domain knowledge, and evaluating various machine learning models. Gradient Boosted Regression Trees emerged as the most effective model, significantly outperforming the baseline model which merely predicted the previous year's absenteeism rate. Key attributes influencing absenteeism were identified, notably including current absenteeism, performance evaluations, and various job type and location-related features. Results highlight the potential of machine learning in proactively managing absenteeism and offer recommendations for future research, such as modeling absenteeism as a time series and incorporating additional data sources. We also show that the current data is not detailed and granular enough to further improve the results.

Keywords

absenteeism, data analysis, data augmentation, machine learning

1 Introduction

Absenteeism — temporary absence from work due to health reasons — is awidespread issue. In Slovenia, it has been on the rise since 2014 (Figure 1), with an average of 56,128 individuals absent daily in 2022, representing approximately 5.91% of the workforce [8]. This carries substantial financial burdens: direct costs like sick pay and indirect costs from overstaffing, reduced productivity and service quality [2]. The complexity of absenteeism, rooted in personal and organizational factors, makes it challenging to predict and manage effectively [10].

Recent years have witnessed a growing interest in leveraging artificial intelligence (AI) and machine learning (ML) to address the absenteeism challenge [5]. Various machine learning techniques, including neural networks, decision trees, random forests, and gradient boosting, have been employed to predict absenteeism and identify its underlying causes [3, 9]. These studies have demonstrated the potential of machine learning in providing valuable insights for proactive absenteeism management.

This paper presents a case study conducted in collaboration with a Slovenian IT company[1] aiming to improve absenteeism prediction and management. The study includes preprocessing and augmenting the company's employee data by incorporating domain knowledge, and evaluating various machine learning models. The findings highlight key attributes influencing absenteeism and offer recommendations for future research and interventions.

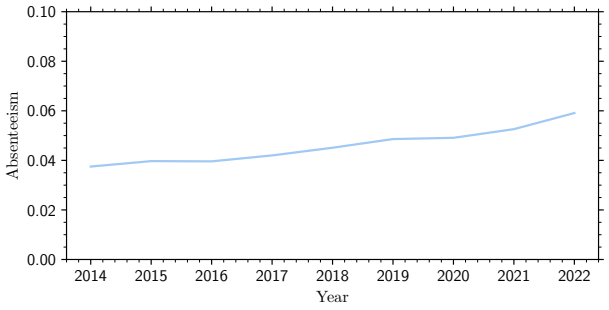

Figure 1: The increase in absenteeism rate in Slovenia between 2014 and 2022 [8]. We can observe a steady increase throughout the years.

The significance of our work extends beyond Company X, offering a blueprint for organizations tackling absenteeism. By showcasing machine learning's efficacy in predicting absenteeism and revealing its drivers, we contribute to the broader field and pave the way for data-driven interventions promoting a healthier, more productive workforce. This aligns with the growing trend of using AI and ML to address complex organizational challenges. Insights from such analyses can aid in strategic workforce planning, optimize resource allocation, and ultimately contribute to a more sustainable and resilient organization.

In section 2 we detail the data and preprocessing, section 3 outlines the methodology, section 4 presents the results, and section 5 discusses the findings and concludes the study.

2 Materials

The data used in our work spanned six years, from 2017 to 2022, and initially comprised 13,798 instances (aggregated employee records) with up to 49 attributes each. They include demographic details, work-related factors, performance evaluations and the current year's absenteeism rate for each employee, but no particulars about sick leave and other personal data.

The initial dataset required substantial preprocessing to prepare it for analysis and machine learning [6]. The data cleaning process involved addressing inconsistencies in attribute values, such as removing extraneous spaces and converting text to lowercase for uniformity. A significant challenge in the dataset was the presence of missing values, denoted by '/'. Their meaning and handling were discussed with a company representative to determine their origins and ensure appropriate treatment. In some

[1]The company asked to remain anonymous, so it is referred to as Company X.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.7260

cases, missing values were imputed based on the average values of similar instances. For example, missing values in 'Kilometers to work' were attributed to errors in data entry and were imputed using the average value for employees living in the same location and working at the same place. On the other hand, missing values in performance evaluations were due to employee's absence on evaluation days.

The target variable — health-related absenteeism rate in the following year — is a continuous variable ranging from 0 to 1. It signifies the proportion of workdays an employee is absent due to health reasons compared to the total number of workdays. The distribution of this target variable is heavily skewed to the right, with most values clustered near zero, indicating that the majority of employees have low absenteeism rates. However, there exist some outliers with extremely high absenteeism rates (Figure 2).

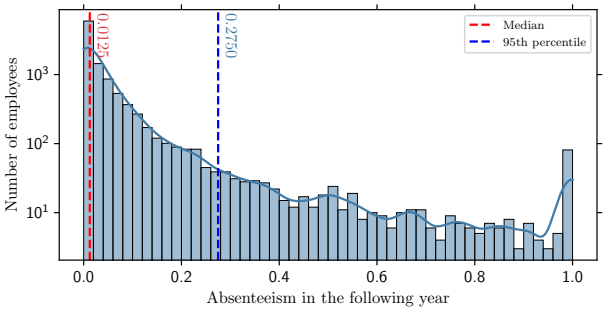

Figure 2: Log-distribution of the target variable. Most workers have very little absence, causing a right-tailed distribution with an "outlier" spike on the right.

The skewed distribution of the target variable has implications for the statistical analysis and machine learning modeling. Therefore, non-parametric statistical tests, such as the Spearman's rank correlation and Kruskal-Wallis test, were employed in EDA and data preprocessing. Additionally, the presence of outliers necessitates careful consideration during model building and evaluation.

The final dataset, comprising 10,347 instances and 42 attributes, serves as the foundation for the subsequent machine learning, where various models are trained to predict absenteeism rates.

3 Methods

The research methodology encompassed a multi-faceted approach, integrating exploratory data analysis, feature engineering, and the application of diverse machine learning models. The ultimate goal was to establish a robust predictive framework for health-related absenteeism, while also ensuring model interpretability to observe actionable insights.

3.1 Exploratory Data Analysis (EDA)

The initial phase involved a thorough EDA to understand the underlying data distribution, identify potential outliers, and uncover preliminary relationships between attributes and the target variable (absenteeism in the following year). Given the skewed nature of the target variable, visualizations like histograms and box plots were complemented by non-parametric statistical tests. The Spearman's rank correlation coefficient was employed to assess monotonic relationships between continuous attributes and the target variable, while the Kruskal-Wallis test was utilized to discern statistically significant differences across groups defined by categorical attributes.

3.2 Data augmentation/Feature Engineering

The original dataset underwent a series of transformations to enhance its suitability for machine learning. This included data cleaning, handling missing values, and the creation of new attributes based on domain knowledge and insights from the EDA. New attributes were engineered based on domain knowledge and statistical analysis. These included indicators for elevated absenteeism, receipt of bonuses or awards, high and low performance evaluations, and absenteeism rates within the employee's team and job type. External factors, such as average absenteeism rates in the employee's residential and work locations, were also incorporated. The feature engineering process was iterative, involving close collaboration with domain experts to ensure the derived attributes were meaningful and captured relevant aspects of employee behavior and organizational dynamics.

3.3 Machine Learning Models

Several well-known machine learning models were employed for absenteeism prediction, including Decision Trees, Linear Regression with L1 regularization, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Gradient Boosted Regression Trees (GBRT), and Random Forest. Hyperparameter optimization was conducted by using Optuna toolkit [1] to optimize model performance.

3.4 Model Evaluation and Selection

Model evaluation was performed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). The models were trained on past years' data and tested on the subsequent year, with the training set size increasing each year. The MAE provided an intuitive measure of the average prediction error, while the RMSE penalized larger errors more severely. The R2 quantified the proportion of variance in the target variable explained by the model. The models were also compared against a baseline model that simply predicted the previous year's absenteeism, to gauge the added value of the machine learning approach. A baseline model predicting the previous year's absenteeism rate was used for comparison.

3.5 Model Interpretation

SHAP (SHapley Additive exPlanations) values [4, 7] were calculated to interpret model predictions and assess attribute importance. SHAP values provide insights into the contribution of each attribute to the model's output, aiding in understanding the factors driving absenteeism. SHAP values provide a unified framework for interpreting any machine learning model, quantifying the contribution of each feature to the model's prediction for a given instance. By analyzing the SHAP values, it was possible to identify the most influential attributes and their directional impact on absenteeism.

3.6 Data Splitting

To ensure robust model evaluation and mitigate the risk of overfitting, the dataset was split into training and testing sets in a prequential manner (year after year). The models were trained on the training set and their performance was assessed on the unseen testing set for the subsequent year. This comprehensive methodological framework enabled a systematic exploration of

the factors influencing health-related absenteeism and the development of a predictive model to proactively manage this critical issue.

4 Results

The primary objective of our work was to develop machine learning models capable of predicting health-related absenteeism in the subsequent year. The models were evaluated using three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The baseline model, which simply predicted the previous year's absenteeism, served as a benchmark for comparison (Table 1).

Table 1: Model performance averaged year-over-year.

Model	RMSE	MAE	R^2
Random Forest	0.107	0.052	0.344
GBRT	0.108	0.051	0.333
Linear Regression	0.108	0.051	0.331
Regression Decision Tree	0.112	0.051	0.281
KNN	0.121	0.057	0.173
SVR	0.117	0.075	0.215
Baseline Model	0.121	0.051	0.156

As we can see, all machine learning models outperform the baseline model in terms of RMSE and R^2. This indicates their superior ability to explain the variance in the target variable (absenteeism in the following year). While the MAE remains relatively consistent across models, the improvement in RMSE and R^2suggests that the models are particularly effective in handling larger deviations in absenteeism predictions.

To establish the statistical significance of the model improvements, we conducted a paired T-test comparing the predictions of each model against the baseline model. All the selected models demonstrated statistically significant improvements ($p < 0.05$) in RMSE and R^2; this ensures that their superior performance is statistically substantiated and not due to chance.

4.1 Performance Trends and Impact of Additional Data per Employee

To gain deeper insights into model behavior, we examined their performance trends over the years. Figure 3 illustrates the evolution of MAE for each model.

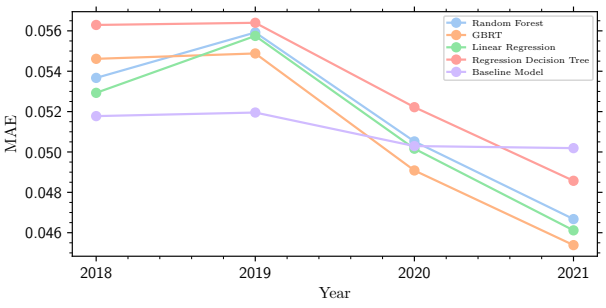

Figure 3: MAE trend over time with additional training data from past years.

Among the evaluated models, GBRT exhibited the best performance, achieving an MAE of 0.045, RMSE of 0.10, and $R^2$of 0.40 on the latest year's data. These results were statistically significantly better than the baseline model, demonstrating the effectiveness of GBRT in capturing the complex patterns underlying absenteeism.

Figure 4 reveals a general trend of MAE improvement for most models in later years, surpassing the baseline model in the final year. This suggests that the models benefit from the increasing amount of training data available in later years. RMSE and R^2charts (not shown) exhibit almost identical properties. It is clear that ML models profit tremendously from increasing amounts of data, as can be expected.

Given the observed performance gains in later years with larger training sets, we explored the impact of incorporating data from previous years. Figure 4 showcases the change in MAE for the final year when models were trained on data from the past year and the past three years, respectively.

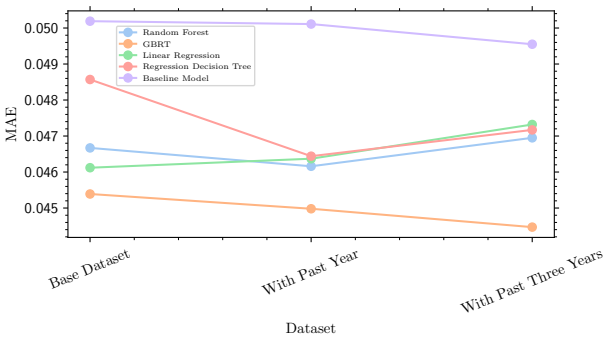

Figure 4: Impact of additional attributes from past years on MAE.

The GBRT model exhibited notable improvement with the inclusion of additional data, achieving an MAE of 0.044, RMSE of 0.093, and R^2 of 0.36. This underscores the value of historical data in enhancing the predictive capabilities of machine learning models for absenteeism and suggests that including even more historical data per employee would be beneficial.

4.2 Interpretability and Additional Insights

Analysis of SHAP values yielded the following key attributes influencing absenteeism:

- Current absenteeism rate
- Performance evaluations
- With respect to the employee's job type and location:
  - Absenteeism rate
  - Proportion of employees with elevated absenteeism
  - Proportion of employees without bonuses

Our findings suggest that absenteeism is influenced by a combination of individual factors (current absenteeism, performance evaluations) and organizational factors (job type, location, bonuses).

Additionally, a rather simple EDA visualisation of functional grouping of employees was quite surprising (Figure 5). Its interpretation can be quite speculative, possibly related to increased job satisfaction or engagement in certain groups. Another, somewhat surprising finding from EDA is that the COVID-19 pandemic did not significantly influence absenteeism rates in 2020, but it may have in 2021 (Figure 6).

Finally, t-SNE visualization of the full dataset shows that employees cannot easily be separated in clusters with similar absenteeism (Figure 7). We can identify some distinct subgroups

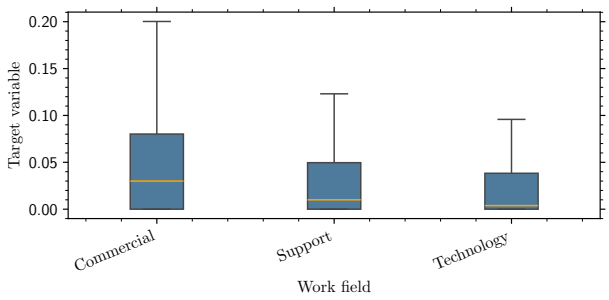

Figure 5: Target variable according to functional partitioning within the company.

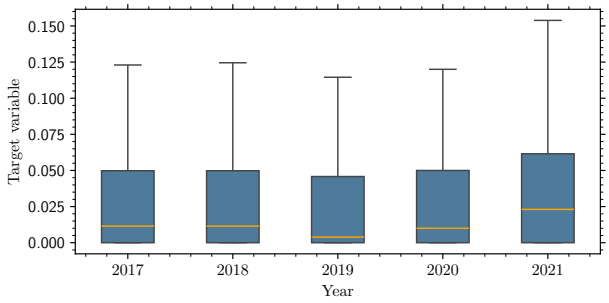

Figure 6: Target variable by year. Note the sharp increase in 2021, possibly attributable to the COVID-19 pandemic.

(like the cluster of red dots on the left), however most data points are quite intermingled. This suggests that with our current set of attributes, we shouldn't anticipate a significant improvement in predictive performance.

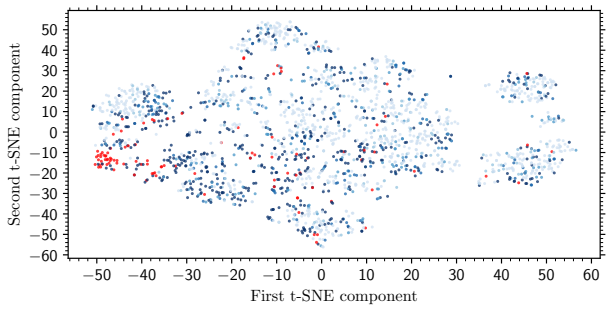

Figure 7: Data visualized in 2D space with t-SNE projection. Red dots represent examples with absenteeism in the next year above 0.25. Blue shades depict examples with absenteeism between 0 (light blue) and 0.25 (dark blue).

5 Discussion and Conclusion

Our work successfully demonstrates the application of machine learning to predict health-related absenteeism. The GBRT model's superior performance highlights its ability to capture complex data relationships, outperforming simpler models and the baseline. Also, identifying key attributes influencing absenteeism, such as current absenteeism, denied bonuses, work type and location, and performance evaluations, provides valuable insights.

The findings align with existing literature highlighting the multifactorial nature of absenteeism. The strong influence of current absenteeism on future absenteeism emphasizes its predictive power, suggesting that past behavior can be a significant indicator of future trends. The negative correlation between performance evaluations and absenteeism suggests that employees with higher evaluations tend to be less absent, potentially due to increased job satisfaction or engagement. The impact of denied bonuses on absenteeism points to the potential role of financial incentives and recognition in influencing employee attendance.

The limitations of our work include the relatively short time span and the potential influence of unmeasured external factors. Future research could address these limitations by: modeling absenteeism as a time series to capture its dynamic nature, incorporating additional data sources such as employee surveys, participation in wellness programs, and (within legal limits) health and personal circumstances data analyzing absenteeism at a finer granularity (e.g., monthly or daily), exploring the inclusion of employee health records and workplace environmental factors in predictive models, and conducting longitudinal studies to track absenteeism patterns over extended periods.

While quantitative improvements of ML model predictions are not overwhelming, the gained insights can enable targeted interventions to reduce absenteeism and promote a healthier workforce. By leveraging ML and data-driven insights, organizations can proactively manage absenteeism, thus improving productivity, financial stability, and employee well-being.

Acknowledgements

The authors sincerely thank to Company X for providing the data, domain expertise and several fruitful discussions.. The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-209).

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2623–2631. ISBN: 9781450362016. DOI: 10.1145/3292500.3330701.

[2] M. Bregant, E. Boštjančič, J. Buzeti, M. Ceglar Ključevšek, A. Hiršl, M. Klun, T. Kozjek, N. Tomaževič, and J. Stare. 2012. *Izboljševanje delovnega okolja z inovativnimi rešitami*. Združenje delodajalcev Slovenije.

[3] B. Hu. 2021. The application of machine learning in predicting absenteeism at work. In *2021 2nd International Conference on Computing and Data Science (CDS)*, 270–276. DOI: 10.1109/CDS52072.2021.00054.

[4] Y. Meng, N. Yang, Z. Qian, and G. Zhang. 2021. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16, 3, 466–490. DOI: 10.3390/jtaer16030029.

[5] I. H. Montano, G. Marques, S. G. Alonso, M. López Coronado, and I. de la Torre Díez. 2020. Predicting absenteeism and temporary disability using machine learning: a systematic review and analysis. *Journal of Medical Systems*, 44, 9, (Aug. 2020), 162. DOI: 10.1007/s10916-020-01626-2.

[6] A. Piciga. 2024. *Napovedovanje zdravstvenega absentizma s strojnim učenjem*. Bachelor's Thesis. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=160413.

[7] E. Štrumbelj and I. Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 3, (Dec. 2014), 647–665. DOI: 10.1007/s10115-013-0679-x.

[8] M. Zaletel, D. Vardič, and M. Hladnik. 2024. Zdravstveni statistični letopis Slovenije 2022. (2024). Retrieved June 5, 2024 from https://nijz.si/publikacije/zdravstveni-statisticni-letopis-2022/.

[9] W. Zaman, S. Zaidi, A. I. Abdullah, and B. Touhid. 2019. Predicting absenteeism at work using tree-based learners. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. Association for Computing Machinery, 7–11. DOI: 10.1145/3310986.3310994.

[10] S. Zupanc. 2011. *Absentizem, kolegialnost in obremenjenost posameznikov*. Bachelor's Thesis. Univerza v Ljubljani. http://www.cek.ef.uni-lj.si/UPES/zupanc1175.pdf.