The title

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline. Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines. One sentence clearly stating the **general problem** being addressed by this particular study. One sentence summarizing the main result (with the words "**here we show**" or their equivalent). Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more **general context**. Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

The title

Introduction

Employee turnover is a critical concern for organizations, as it impacts productivity, performance, and overall organizational effectiveness (Blau & Boal, 2020; Griffeth et al., 2000). Accurate prediction of turnover is crucial for proactive human resource management and the implementation of effective retention strategies (Mitchell et al., 2001; Mobley et al., 1979). In recent years, the application of predictive modeling techniques has gained prominence in addressing this challenge. Specifically, the debate arises as to whether regression-based models or machine learning models are more effective in predicting turnover, particularly when working with small sample sizes.

The aim of this dissertation is to investigate and compare the predictive capabilities of regression-based models and machine learning models in the context of turnover prediction, focusing specifically on sample sizes below 200,000. By examining the strengths and limitations of these modeling approaches, this study seeks to shed light on which method offers greater accuracy and reliability in predicting turnover within resource-constrained environments.

Regression-based models, including linear regression, logistic regression, and Cox proportional hazards regression, have long been established as prominent tools in predictive modeling (Hosmer et al., 2013; Cox, 1972). These models are characterized by their simplicity, interpretability, and assumption of linearity between predictors and the outcome variable. The straightforward nature of regression-based models allows for the identification of significant predictors and estimation of their individual effects, facilitating an understanding of the underlying mechanisms driving turnover (Meyer et al., 2004; Hom et al., 2009).

Contrarily, machine learning models have garnered significant attention due to their ability to handle complex relationships and patterns in large datasets (Breiman, 2001; Hastie et al., 2009). Algorithms such as random forests, support vector machines, and artificial neural networks offer the potential to capture non-linear and interactive effects, making them valuable tools in predictive modeling (Niculescu-Mizil & Caruana, 2005; Zhang & Singer, 2010). Machine learning models have been increasingly applied to turnover prediction, displaying promising results in various studies (Biswas et al., 2019; Jaradeh & Dehghan, 2021).

While the application of machine learning models has gained momentum, their performance in the context of small sample sizes remains an open question. The literature suggests that machine learning models may face challenges, such as overfitting, when trained on limited data (Jiang et al., 2020; Varoquaux et al., 2018). Consequently, the predictive performance of these models might be compromised when sample sizes are below a certain threshold. Thus, it becomes imperative to evaluate whether regression-based models, with their simplicity and interpretability, outperform machine learning models when the sample size is below 200,000.

Additionally, the performance of machine learning models may be affected when the number of independent variables is small. In such scenarios, these models may face challenges such as overfitting or difficulty in identifying meaningful patterns (Varoquaux et al., 2018; Guyon & Elisseeff, 2003). On the other hand, regression-based models, with their simplicity and interpretability, may offer advantages in situations where the number of independent variables is limited, as they are less prone to overfitting and can provide transparent insights into the relationships between predictors and turnover (Pedersen & Skogstad, 2020; Hosmer et al., 2013).

Traditional variables in the demographic and biodata domain, such as age, gender, and education, have been commonly used in turnover prediction models (Hom et al., 2009; Meyer et al., 2004). However, the utilization of antecedent variables typically studied in

I-O psychology, such as job satisfaction, organizational commitment, and work-life balance, may provide deeper insights into the underlying factors contributing to turnover (Lee et al., 2019; Hom et al., 2009).

The application of I-O psychology antecedent variables in machine learning-based models holds promise for improving turnover prediction accuracy. These variables capture psychological and organizational aspects that directly impact employees' turnover intentions and behaviors (Lee et al., 2019; Meyer et al., 2004). By considering these variables in predictive models, organizations can gain a more comprehensive understanding of the complex dynamics that drive turnover and develop targeted interventions to mitigate it (Griffeth et al., 2000; Hom et al., 2009).

Contrarily, models relying solely on demographics and biodata may overlook critical factors contributing to turnover. While these variables provide basic demographic information, they may lack the depth and specificity necessary to capture the nuances and complexities of turnover behavior (Hom et al., 2009). Incorporating I-O psychology antecedent variables can offer a more nuanced and accurate prediction of turnover by considering individual attitudes, perceptions, and experiences within the organizational context (Meyer et al., 2004; Lee et al., 2019).

This study aims to address this research gap by employing a comprehensive dataset from multiple organizations. By leveraging turnover data and a restricted set of independent variables, along with relevant predictor variables such as demographics/biodata, job characteristics, employee engagement, and other I-O psychology antecedent variables of turnover, a comparative analysis will be conducted. The performance of regression-based models and machine learning models will be assessed using various metrics, including accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

The findings from this research will contribute to the existing literature on turnover prediction and provide valuable insights for practitioners and researchers alike. Understanding the relative performance of regression-based models and machine learning models when dealing with small sample sizes can guide decision-making regarding the choice of modeling techniques in resource-limited scenarios. Ultimately, this research aims to enhance our understanding of turnover prediction and inform effective retention strategies to mitigate the negative consequences of employee turnover. ## Hypothesis Hypothesis 1: Does regression cross-validate better when the sample size is below 200,000 compared to ML Hypothesis 2: Does regression cross-validate better when the number of variables is below 200 compared to ML Hypothesis 3: Do ML-based predictive models have better predictions of turnover when using IO IVs VS other IVs.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

Material

Procedure

Data analysis

We used R (Version 4.2.2; R Core Team, 2022) and the R-packages *papaja* (Version 0.1.1.9001; Aust & Barth, 2022), and *tinylabels* (Version 0.2.3; Barth, 2022) for all our analyses.

Results

Discussion

References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown.* Retrieved from https://github.com/crsh/papaja

Barth, M. (2022). *tinylabels: Lightweight variable labels.* Retrieved from https://cran.r-project.org/package=tinylabels

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/