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 high amount of computer power in the cloud environment nowadays and the developments in the field of machine learning are providing easy access to high-performance services. Machine learning-supported tools are enabling companies to analyze and evaluate information in a quick and effective way (Tambe et al., 2019). We see this in the form of applications, software, and solutions that are common in business or that automate the different decision-making processes, such as programs that create and post job descriptions, application tracking systems that identify key words to place candidates in the right openings, tools for scheduling interviews with your online calendar, chatbots for screening, etc (Rąb-Kettler & Lehnervp, 2019). Human resources practices are not being oblivious to these developments. Experts in these practices are realizing the advantages of data-driven decision making (Fallucchi et al., 2020). Large amounts of human resources data can be analyzed in a short time and empirical inferences can be made, enabling experts to better understand employees and help anticipate issues and patterns (Merlin & Jayam, 2018). Being able to predict the best suited personnel for positioning or that will turn over is of particular interest to human resource departments and companies in general. Making the wrong decision when giving a promotion or demotion can cause waste of time and energy, as well as compromise the perceived organizational justice and support, resulting in more turnover. This is why personnel placement processes are some of the most pivotal in human resources (Merlin & Jayam, 2018).

When making decisions for placement using traditional methods, there is a high probability of this being affected by subjective factors that can cause biased choices from time to time (Fallucchi et al., 2020). With machine learning, on the other hand, these decisions are based on a bit more objective foundation than most other recommended methods, since it is based solely on the patterns the algorithm finds on the data (though there can still be bias in the development or implementation of particular algorithms, this is minimized in comparison with traditional methods). Not only this, but the decisions made in the personnel placement process can be explained to the candidates with their reasons , providing them with confidence in the results and diminishing the chances of low perceived organizational justice/support and high turnover (Jha et al., 2020).

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.

What Constitutes Machine Learning

Definitions of what constitutes machine learning (ML) and the differences with statistical modeling have been discussed at length in the literature (Breiman, 2001), yet the distinction is not clear-cut (Moons, 2014). The seminal reference on this issue is Breiman's review of the "two cultures" (Breiman, 2001). Breiman contrasts theory-based models such as regression with empirical algorithms such as decision trees, artificial neural networks, support vector machines, and random forests.

Theory-based models

Theory-based models are models that are based on theory and assumptions, such as traditional linear regression, and benefit from human intervention and subject knowledge for model specification. The analysis in this approach starts with assuming a stochastic data model for the inside of the black box. Usually, research that uses this approach starts by assuming that the data are generated by a particular model. This model is used as a template for statistical analysis. When faced with an applied problem, researchers that use this approach come up with a data model by looking at the literature developed by previous scholars or by their own theorizing, or some combination of both. This enables them to develop a reasonably good parametric class of models for a complex mechanism devised by nature, and then parameters are estimated and conclusions are drawn. However, these conclusions are about the model's mechanism, not about nature's mechanism, and therefore if the model is a poor emulation of nature, the conclusion could be wrong. Breiman (2001) criticized this approach, pointing out that: "A few decades ago, the commitment to data models was such that even simple precaution such as residual analysis or goodness-of-fit tests were not used. The belief in the infallibility of the data models was almost religious. It is a strange phenomenon – once a model is made, then it becomes truth and the conclusions from it are infallible (p. 202)." He concludes by using the following old saying: "If all a man has is a hammer, then every problem looks like a nail (p. 202)." To solve a wide range of problems, such as is the case in the social sciences with the abundance of variables, a larger set of tools is needed. The rapidly increasing ability of computers to store and manipulate data can provide us with more varied tools. Empirical-based models In the mid-1980s, with the development of neural networks and decision trees, a new community of researchers appeared focused on predictive accuracy (Cristianini, 2002). They began using these algorithms on working in complex prediction problems where it was obvious that data models were not applicable, such as speech recognition, image recognition, handwriting recognition, times series analysis, or financial market analysis. The approach is that nature produces data in a black box whose insides are complex and partly unknowable. The goal is not to explain the patterns in this data, but to predict them based on input; not to focus on data models, but on the characteristics of the algorithms (Breiman, 2001). Within psychology, this is the same approach as dustbowl empiricism (Schoenfeldt, 1999). A useful definition of machine learning is that it focuses on models that directly and automatically learn from data (Mitchell & Mitchell, 1997). For example, machine learning performs modeling more automatically than regression regarding the inclusion of nonlinear associations and interaction terms

(Boulesteix, 2014). To do so, machine learning algorithms are often highly flexible algorithms that require penalization to avoid overfitting (Deo, & Nallamothu, 2016). Some researchers describe the distinction between statistical modeling and machine learning as a continuum (Beam, & Kohane, 2018). Other researchers label any method that deviates from basic regression models as machine learning (He, & Garcia, 2009), such as penalized regression (e.g., LASSO, elastic net) or generalized additive models (GAM). We note that these methods do not belong to machine learning by using the "automatic learning from data" definition, and did not classify these as machine learning in this study

Data Hungriness

The concept of data hungriness refers to the sample size needed for a method to generate a prediction model with a good predictive accuracy (van der Ploeg et al., 2014). The data hungriness of a predictive modeling technique is defined as the minimum number of events per variable at which the optimism of the generated model is less than 0.01. Optimism is defined as the difference between error on our sample data and the error when applying our model to another dataset. Every machine learning model has some amount of error in its predictions. This error usually comes from two different sources: bias and variance. Bias is the tendency of the model to underfit, and variance is the tendency to overfit. The relationship between these two sources of error is known as the bias-variance tradeoff, and developers of machine learning models have to find the balance between the two.

To test if this trade-off has been done in a way that minimizes error, it is a good idea to measure performance using data that the model has never seen before. The performance of the model on this "test data" will be a more accurate predictor of the model's performance in the real world, which is the fundamental basis for cross-validation . The model's optimism, therefore, is the difference between the training error estimated from the data used to build the model and the test error estimated by applying the model into out-of-sample data. The sample size needed to minimize this difference is what we call data hungriness. Machine learning algorithms require big sample sizes to minimize this difference (van der Ploeg et al., 2014).

Parsimony

Parsimony is defined as the sample size and number of variables that a dataset must have in order to maximize the predictive accuracy of a model (Sanchez-Pint et al., 2018). A model is considered parsimonious when it both uses the least amount of variables possible (sparsity) and has good prediction accuracy. Typically, parsimony is reported as the performance metric of models that are sparse.

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/