Influence of Parsimony and Sparsity in Predicting Turnover when Using Machine Learning VS Linear Regression

Diego Figueiras1

1 Montclai State University

Influence of Parsimony and Sparsity in Predicting Turnover when Using Machine Learning VS Linear Regression

# Rationale for Project

The goal of this dissertation is to use advanced machine learning tools and compare them to regression when predicting one of the most well-known constructs in the industrial/organizational psychology literature: Employee Turnover. Employee turnover is a critical concern for organizations, as it impacts productivity, performance, and overall organizational effectiveness 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. 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.

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 Jr 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 and Caruana (2005); Kermany et al. (2018). Machine learning models have been increasingly applied to turnover prediction, displaying promising results in various studies (Biswas et al. (2020)).

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 260,000.

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.

# Methods

## Procedure and Participants

The data that will be used for this study are responses to the Federal Employee Viewpoint Survey (FEVS) publicly available at the website of the Office of Personnel Management. The dataset consists of 107 variables and a sample size of 557,779 federal employees that took the survey in 2022. Additional, simulated responses with artificial variables predetermined to correlate with the criterion will be generated, using different sample sizes. These will be 500, 10,000, 500,000, 1,000,000, and 20,000,000. The number of variables will also variate per predictive model, with these being 5 variables, 50, 500, 1000, and 10,000. A Qualtrics survey including all the measures in the FEVS and additional biographical information will be sent out via Amazon’s MTurk. Participants will be sourced to create a sample that closely mirrors the diverse working population across various industries. Quality assurance measures, including attention check items, tracking response times, and synonymous/antonymous items will be used. The R careless package will be used to control for careless responding, with participants that are responding inconsistently in a significant way with the synonymous/antonymous items being excluded from the study. The Cloud Research platform will be used to to help ensure the quality of the data. The following MTurk and Cloud Research inclusion criteria will be used for the online survey: approved 95% HIT (task) rate on MTurk, 100+ HITs approved, geographical locations restricted to the United States, Cloud Research approved participants, duplicate IP addresses blocked, and suspicious geocodes blocked. Additionally, participants will be compensated for their participation.

## Data analysis

Several machine learning and regression models will be trained, and the predictive accuracy measures of Turnover intention will be estimated for each combination of data, sample sizes, number of variables, and model. The algorithms being used will be determined by a prior literature review conducted by the author in which the most frequent and accurate algorithms used in predicting turnover were the following: Gradient Boosting Trees (GBT), Random Forest (RF), Neural Networks (NN), and Support Vector Machines (SVM). Additionally, a logistic regression will be used as a comparison.

# Discussion

As Richard Landers said in an interview with SIOP, “Let’s blend I-O psychology’s tried-and-true practices where we know what we’re measuring and we’re very confident in the kinds of recommendations we’re giving, and let’s figure out where the intersections are with some of the new stuff coming out, to figure out what is truly new and useful and what is just a faddish waste of time.” (Landers, 2019). Many of the new technological developments in the fields of data science and computers science are not backed by the expertise that I/O psychology has been developing over the years, yet they appear more attractive for many HR practices. It is important that we blend with these communities and contribute with our knowledge, otherwise, as Landers puts it, “they’re going to run away with the farm,” (Landers, 2019). A way in which the field of I/O psychology could contribute in the development of machine learning solutions for HR practices would be to build algorithms that use antecedents of job performance commonly studied by I/O psychology researchers as input variables. These could be variables such as organizational commitment, organizational engagement, organizational identification, perceived organizational support, perceived organizational justice, etc. Showing evidence of incremental predictive accuracy over algorithms built with other types of input would encourage better practices. Although machine learning algorithms have the advantage that they can learn from the data and look at associations between variables that would need to be specified if a regression technique was used instead, they are also very data hungry. This makes them viable for fortune 500 companies, for instance, but in most contexts there’s the possibility that regression methods will produce the same predictive accuracy (or perhaps even better) without having to sacrifice parsimony and sparsity. It is also important to be selective on the variables that are being fed to the algorithms, since using all and any data that gives good predictive performance could lead to the aforementioned GIGO issue (“garbage in, garbage out”). Input may be subjective (and/or biased), particularly if there are IVs such as “previous performance”. This GIGO issue might be exacerbated by peoples’ awe and wonder of ML. Using variables supported by previous research would be best practice in building these ML solutions.

**Itemized Budget and Budget Narrative**

The funding being requested for this dissertation project is solely for participant compensation. Due to the focus of the study being on current workers, we cannot use subject pools like MSU SONA. Instead, we will be recruiting participants through Amazon Mechanical Turk/Cloud Research, which requires paying participants. The target sample size is 500. With the current minimum wage in the USA as a reference ($7.25 per hour), and the additional MTurk research costs that factor into the payments, the projected costs are $1 to $1.50 per participant. That is a total projected cost of the project of $500 to $750. As a result, I am requesting the maximum $500 amount for this internal grant, which would be a tremendous help in covering the cost of the project.

# References

Biswas, S., Fernandez Castellanos, D., & Zaiser, M. (2020). Prediction of creep failure time using machine learning. *Scientific Reports*, *10*(1), 16910.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management*, *26*(3), 463–488.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hom, P. W., Tsui, A. S., Wu, J. B., Lee, T. W., Zhang, A. Y., Fu, P. P., & Li, L. (2009). Explaining employment relationships with social exchange and job embeddedness. *Journal of Applied Psychology*, *94*(2), 277.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122–1131.

Meyer, J. P., Becker, T. E., & Vandenberghe, C. (2004). Employee commitment and motivation: A conceptual analysis and integrative model. *Journal of Applied Psychology*, *89*(6), 991.

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal*, *44*(6), 1102–1121.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin*, *86*(3), 493.

Niculescu-Mizil, A., & Caruana, R. (2005). *Predicting good probabilities with supervised learning*. 625–632.