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

Diego Figueiras1

1 Montclai State University

Author note

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

Correspondence concerning this article should be addressed to Diego Figueiras, Dickson Hall 226. E-mail: [figueirasd1@montclair.edu](mailto:figueirasd1@montclair.edu)

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

# Introduction

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 (T. R. 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.

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 (Macijauskien & Stankevi, 2021). 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 (Macijauskien & Stankevi, 2021).

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.

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 269,999. 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 (Cox, 1972; Hosmer Jr et al., 2013). 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 (Hom et al., 2009; Meyer et al., 2004).

Contrarily, machine learning models have garnered significant attention due to their ability to handle complex relationships and patterns in large datasets (Breiman, 2001a; 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 (Kermany et al., 2018; Niculescu-Mizil & Caruana, 2005). 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 (Varoquaux, 2018; Xu et al., 2020). 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 269,999.

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 (Guyon & Elisseeff, 2003; Varoquaux, 2018). 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 (Hosmer Jr 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., 2012; 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 (Hom et al., 2012; Lee et al., 2021).

The application of work-related psychological 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., 2021; 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., 2012).

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., 2012). Incorporating work-related psychological antecedent variables can offer a more nuanced and accurate prediction of turnover by considering individual attitudes, perceptions, and experiences within the organizational context, as had been proven before (Lee et al., 2021; Meyer et al., 2004). Using them when building machine learning complex models is congruent with previous research.

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 work-related psychological 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).

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, 2001b), yet the distinction is not clear-cut (Collins et al., 2014). The seminal reference on this issue is Breiman’s review of the ‘’two cultures’’ (Breiman, 2001b). 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 & Scholkopf, 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, 2001b). 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 (T. M. Mitchell, 1997). For example, machine learning performs modeling more automatically than regression regarding the inclusion of nonlinear associations and interaction terms Boulesteix and Schmid (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

## Common Terms in Machine Learning

* Supervised learning: these are algorithms that have a dependent variable they are trying to predict. If the dependent variable is discrete, it would be a classifier; if it is continuous, it would be regression-based (Hastie et al., 2009). For the purposes of this study we will be focusing on this type of machine learning.
* Unsupervised learning: these are algorithms that don’t have a dependent variable that they are trying to predict, but rather look for associations and clusters among the independent variables (Hastie et al., 2009)
* Cross-validation: this is a technique used to assess the performance and generalizability of a predictive model. It involves partitioning the dataset into multiple subsets, training the model on some of these subsets, and evaluating its performance on the remaining data. This process is repeated multiple times, with different subsets used for training and testing, allowing for a more robust evaluation of the model’s effectiveness across various data scenarios (Douglass, 2020).
* Regularization: in machine learning, this is a technique employed to prevent overfitting and enhance the generalization performance of a predictive model by adding a penalty term to the cost function. This penalty discourages the model from fitting the training data too closely and helps to control the complexity of the model, preventing it from becoming too intricate and specialized to the training set (Hastie et al., 2009).
* Ensemble: in machine learning, this involves combining predictions from multiple individual models to create a more robust and accurate overall prediction (Dietterich, 2000). By leveraging the diversity among the constituent models, ensemble methods aim to improve generalization performance, mitigate overfitting, and enhance predictive accuracy across a variety of scenarios.

## 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 (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 (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-Pinto, Luo, 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.

# Algorithms

In this study, various machine learning algorithms are described and assessed in their ability to predict dependent variables of interest in the field of I/O psychology. This section provides a general overview of the theory behind these algorithms.

## Decision Trees

Decision trees are a type of supervised machine learning method. They create classification or regression models following a tree-like structure (Mahesh, 2020). Each fork is a split in a predictor variables, and in which each end node contains a prediction for the outcome variable (Marsland, 2011). They are used to explain how the target variable’s values can be predicted based on other variables. Nodes are split into sub-nodes based on a threshold value of a variable (Mahesh, 2020). Decision trees are capable of handling missing values and mixed features, as well as to select variables automatically. On the downside, their predictive power is not as high as other algorithms and they are not stable with high model variance and small variations in the data. This could result in a large effect on the tree structure that is not meaningful. Understanding how they operate becomes important in order to understand other algorithms, such as random forest and gradient boosting trees (Alpaydin, 2020; Marsland, 2011).

## Classification and Regression Trees (CART)

A Classification and Regression Tree is an umbrella term that refers to the use of either classification and/or regression trees (Marsland, 2011). We already explained classification trees, so lets now explain regression. Regression trees are a type of decision tree. They are different from classification trees in that each leaf represents a numeric value, while classification trees have “true” or “false” in their leaves, or some other discrete category. The roots and branches of regression trees are typically ranges of the IVs, while the leaves are typically the average value of the DV in those ranges. The structure of these trees typically starts with the root being the lower threshold of the IVs, and its leaf the average DV at this range. Then the first branch is the upper threshold, with the leaf being the average DV at that range. The second branch is the middle threshold of the IVs, with the leaves being the average DV between the middle and the lower threshold and the average DV between the middle and the upper threshold. Because of this structure, regression trees are better at capturing non-linear associations than simple linear regression (Mahesh, 2020).

## Random Forests (RF)

Random forest are machine learning algorithms that take an ensemble approach that provides an improvement over decision trees (Mahesh, 2020). They are built by taking a random sample of the data and then building an ongoing series of decision trees on the subsets. They create many decision trees (hence “forest”) to improve predictive accuracy and if one or more of these smaller decision trees are not relevant, they get ignored in favor of the better ones. In other words, they combine a group of weak learners to form a stronger learner.

The way this algorithm works is that a number of decision trees are built on bootstrapped training sets and a random sample of IVs are chosen at each step as split variables from the full set of predictions in each decision tree. In this way, it is unlikely that all of the individual trees will be influenced by a few “noisy” predictors. Hundreds of decision trees are built this way, which is why these algorithms can be very slow and take a lot of computational memory. The algorithm keeps track of the outcome predicted by every decision tree and picks the most frequent one; this is called “bagging”. Variance is reduced by taking the average of the uncorrelated trees (“out of bag” samples), making the final result more reliable and less variable. They greatly help reduce overfitting and bias because of this, which are two of the greatest limitations of regular decision trees. Similarly, they can fill in missing data by computing the weighted averages across the hundreds of samples, in the case of continuous data, and frequencies, in the case of categorical data. (Alpaydin, 2020).

## Gradient Boosting Trees (GBT)

Gradient boosting trees are very similar to random forests but with the difference that the former learn sequentially instead of randomly. The ways the individual trees are built and their results are combined is different. Random forest builds independent decision trees and combines them in parallel, while gradient boosting trees use a method called “boosting”, which combines each learner sequentially so that each new tree corrects the errors of the previous one. A weak learner would be a decision tree with only one split, which is also called a “stump”. To evaluate how well each tree does, the algorithm uses a loss function (such as cross entropy, to name one). In classification trees, when the DV label and predictor do not agree, the loss function is close to 1; when they are in perfect agreement, the loss function is 0. In GBT, a series of trees are created and each of them tries to lower the loss function of the previous tree in the series. Trees are constantly added in this way until no further enhancement can be achieved. This makes predictions with GBT fast and memory-efficient, although they are hard to visualize and interpret. Compared to random forest, they have a lot of model capacity that enables them to model complex relationships and decision boundaries. As with previous decision trees methods, there is also the danger of overfitting (Mahesh, 2020).

## Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees are also an ensemble method that uses many decision trees as its building blocks, much like RF or GBT. It could be said that BART tries to capture the best of RF and GBT; they construct the trees by sampling randomly, much like bagging in RF, and they also try to capture signal not yet accounted for by the current model, much like boosting in GBT. It is considered a nonparametric function approach used to predict using regression trees. They rely on recursive binary partitioning of predictor space into hyperrectangles. Hyperrectangles are cubes used to classify data, since putting data in 3d space is a better way to classify mixed data. Once this is done, the trees are summed and regularized to avoid overfitting (Chipman et al., 2010; Marsland, 2011).

## Neural Networks (NN)

Neural networks were built with the intention of emulating the human nervous system. The input is associated with weights, bias, activation functions, and a computed output from all of this. Much like decision trees, NNs consist of nodes and connections between the nodes, with input nodes, output nodes, and different layers of nodes in between. A neural network may contain more than one layer between input and output to handle complex problems. These layers are referred to as hidden layers (Alpaydin, 2020). Due to the rapid development of hardware and continuous research on backpropagation techniques, NNs are the most studied area in machine learning (Alpaydin, 2020; Mahesh, 2020; Marsland, 2011).

## Support Vector Machines (SVM) classifiers

Support vector machines are supervised machine learning algorithms commonly used in classification problems. They are based on the idea of finding a hyperplane that best divides the dataset into two classes. The further from the hyperplane that the data points are, the more confident we can be that they have been correctly classified, and the support vectors would be the data points that are nearest to it and that, if removed, would alter the position of the dividing hyperplane. Therefore, these points can be considered critical elements of the dataset. The distance between the hyperplane and the support vectors is called margin, and the goal of this algorithm is to choose a hyperplane with the greatest possible margin in order to have greater chances of correctly classifying new data. To determine the margin, cross-validation is used. SVM is mostly used for text classification tasks and image recognition. They are great at handling outliers, since they allow misclassifications and overlapping classifications (Alpaydin, 2020; Mahesh, 2020; Marsland, 2011).

# Evaluation metrics

Performance metrics evaluate a model’s predictive performance and tell you how good or poor the performance of the model is. In this study, various performance metrics will be used. This section provides a general description of them.

## Accuracy

Accuracy is a prediction performance metric used for classification machine learning algorithms (those that predict a dependent variable that has categorical data rather than continuous). It is perhaps the simplest metric to use and implemented, and it’s defined as the number of correct predictions divided by the total number of predictions. This ratio is usually stated as a percentage (Marsland, 2011).

## Precision

Precision is also a ratio for classification algorithms, much like accuracy, but between true positives and total amount of positives predicted. This metric focuses on Type-I errors (Alpaydin, 2020); in other words, incorrectly labeling the dependent variable during prediction. A precision score towards 1 means that the model did not miss any true positives and is able to classify well between correct and incorrect labeling of the dependent variable. A low precision score of less than 0.5 means that the classifier has a high number of false positives, which can be an outcome of imbalance class. This metric cannot measure the existence of Type-II error (Marsland, 2011).

## Recall/Sensitivity

Recall, also called sensitivity, focuses on Type-II errors (Alpaydin, 2020). It is similar to precision but with a different denominator, which would be the number of true positives plus the number of false negatives. Recall towards 1 will mean that the model did not miss any true positives and is able to classify well between correctly and incorrectly labeling the dependent variable. A low recall score of less than 0.5 would mean that the classifier has a high number of false negatives, which can be an outcome of imbalance class. This metric cannot measure the existence of Type-I error (Alpaydin, 2020; Marsland, 2011).

## F1-score

This metric combines precision and recall, since it is the harmonic mean between the two (Alpaydin, 2020). A high F1 score means that there is high precision and high recall. A low F1 score has little to no meaning, since although it tells us that the precision and recall were low, it does not tell us which cases were incorrectly classified as true positives or false negatives. It is still useful in deducing the performance of the model (Marsland, 2011).

## Area Under the Receiver Operating Characteristics Curve (AUC-ROC)

AUC-ROC is a useful metric to compare different machine learning classifiers. The true positives ratio (TPR) and false positives ratio (FPR) are plotted in a graph under different thresholds. This resulting curve is called the Receiver Operating Characteristics Curve (ROC). The area under this curve is what we call AUC, which is equivalent to the probability that a randomly chosen true positive case is deemed to have a higher probability of being positive than negative than a randomly chosen negative case. In other words, a high AUC means that the probability of a randomly chosen positive example is indeed positive. This tells that the model does a good job at discriminating negatives and positives, with most true positives at one end and positives in the other (Alpaydin, 2020; Marsland, 2011).

## Root Mean Square Error (RMSE)

The root mean square error is another standard way of measuring the accuracy of a predictive model. It shows how far predictions fall from measured true values using Euclidean distance (Myers et al., 2013). In other words, it is simply the standard deviation of the residuals. It tells you how spread out the errors are and how concentrated the data is around the line of best fit. It indicates the absolute fit of the model to the data and the average model prediction error in units of the dependent variable. The lower the RMSE, the better the predictive accuracy.

## Log-loss

Log-loss is another important metric in assessing the performance of machine learning algorithms, especially classifiers. It indicates how close the prediction probability is to the true value, and the more the predicted probability is different from the actual value, the higher the log-loss (Alpaydin, 2020; Marsland, 2011).

## Kappa Statistic

Kappa statistic, also called Cohen’s kappa, is a metric used to assess the performance of a classification model. It is the same metric to compute inter-rater agreement, but in the context of machine learning, it can be used to compare the algorithm’s predictions on different classes or labels of a dependent variable (Alpaydin, 2020; Marsland, 2011).

## Geometric mean score (G-mean)

The geometric mean is another useful metric in assessing accuracy of performance. It is the root of the product of a class-wise sensitivity. It tried to maximize the accuracy on each of the classes and keep them balanced at the same time. Mathematically, it is basically the square root of the product of sensitivity and specificity (Kubat et al., 1997); Barandela et al. (2003)].

# Rationale

Predictive modeling has gained significant attention in recent years across various disciplines due to its potential to make data-driven decisions. One of the fundamental questions that researchers and practitioners encounter in predictive modeling is whether traditional statistical regression models or machine learning (ML) algorithms are more effective in producing accurate predictions. This debate is crucial as it guides the selection of appropriate modeling techniques for specific datasets, which can vary in terms of sample size and the number of variables. In this section, we will review the literature and rationale leading to the hypotheses for Study 1.

## Sample Size and Predictive Modeling

Predictive modeling techniques, such as regression models and ML algorithms, differ in their ability to handle different sample sizes effectively. Owing to its flexibility, ML methods are claimed to have better performance over traditional statistical modeling, and to better handle a larger number of potential predictors (Vijayakumar & Cheung, 2019). Many of these methods have the ability to capture non-linear associations without a theoretical specification by the researcher, as has been shown by simulation studies done by Miller et al. (2016), in which they simulated datasets with prescribed multivariate dependencies and non-linear effects, and they found out that random forest and boosting performed better than regression in selecting variables relevant to fit the model. Multiple linear regression models also have the inconvenience that they use assumptions when fitting the data. Namely, that there is statistical independence of the errors, a linear relationship between dependent and independent variables, homoscedasticity of the error, and normality of the error distribution (Myers et al., 2013). We refer to this as inconvenient because these assumptions may not always be met, which can lead to biased results that jeopardize predictive accuracy. Machine learning methods, on the other hand, are quite effective at handling non-linear and complex data without having to make any assumptions, even if datasets are noisy (Craninx et al., 2008). Alzate Vanegas et al. (2022) did a study that showed evidence of classification trees outperforming logistic regression when predicting attrition in the U.S. Marine Corps. Froud et al. (2021) found that machine learning outperformed regression for simulated non-linear heteroscedastic variables when predicting quality of life and academic performance of school children in Norway. Luu et al. (2020) found that machine learning (specifically, gradient boosting trees) outperforms logistic regression in predicting next-season NHL player injury. It has been shown that machine learning algorithms also require more data than traditional data modeling, since their data hungriness is so demanding (Ploeg et al., 2014). It is worth noting, also, that there are cases where regression models outperform machine learning in predictive accuracy (Sanchez-Pinto, Luo, et al., 2018; Sanchez-Pinto, Venable, et al., 2018). Furthermore, machine learning models are typically assessed in terms of discrimination performance (e.g., accuracy, area under the receiver operating characteristic [ROC] curve [AUC]), while the reliability predictions (calibration) are often not assessed (Van Calster et al., 2016).

In a preliminary systematic literature review conducted by the author of this dissertation, when looking at ML-based predictive models built to predict turnover, among other relevant I/O psychology dependent variables, only two articles out of the twenty-eight reviewed had data with sample sizes above 269,999, which is the threshold established by Sanchez-Pinto, Venable, et al. (2018) at which machine learning algorithms outperform traditional data modeling such as regression. Without meeting the data hungriness of machine learning algorithms, not only is the accuracy of the models jeopardized, but the outcomes might not be any better than regression, raising the question of why should the parsimony of these traditional models be sacrificed for more complex models if the results might be the same?

In the same literature review, when it comes to parsimony, the article with the lowest events per variable (EPV) was 1.62 and the highest was 22267.42. The median EPV was 44, which is small and more appropriate for predictive studies using regression, as stated by Sanchez-Pinto, Venable, et al. (2018), since their research shows that machine learning algorithms perform better than regression at a EPV above 200. Only ten out of the twenty-eight articles in this literature review had a EPV above 87. Using machine learning algorithms with such small number of variables could compromise the bias-variance tradeoff, overfitting the data.

# Hypotheses

## Study 1: Sample Size and Number of Variables

Since there is some recognition that regression models perform well with smaller datasets, where the number of observations is limited, and that ML algorithms tend to shine when the dataset is substantial, the following hypotheses to investigate this relationship are proposed:

Hypothesis 1a: If the sample size is below 269,999, predictive accuracy metrics will be better when running a regression model than when running a ML algorithm.

Hypothesis 1b: If the sample is above 269,999, predictive accuracy metrics will be worse when running a regression model than when running a ML algorithm.

The choice of the threshold, 269,999, is based on prior research (Kirasich et al., 2018; Sanchez-Pinto, Venable, et al., 2018), which suggests that this is the point at which ML algorithms begin to outperform regression models consistently. In their study, they compared machine learning methods with regression methods for variable selection for two differently sized datasets, focusing on parsimony in predictor selection, which balances predictive accuracy with the sparsity of predictors. Their findings were that for the smaller dataset (N=6,565), regression achieved the best predictive accuracy, while machine learning methods faring better only in the larger dataset (N=269,999).

Hypothesis 1c: there are no significant differences in predictive accuracy metrics between regression models and ML algorithms for certain sample sizes. This hypothesis is included to account for the potential influence of other factors, such as the complexity of the data or the specific predictive task.

In addition to sample size, the number of variables in a dataset is another critical factor affecting the choice between regression models and ML algorithms. Going back to the study done by Sanchez-Pinto, Venable, et al. (2018), predictive accuracy decreased for machine learning when the dataset had fewer variables compared to regression models, whose predictive accuracy remained the same or improved. Specifically, this happened when the EPV were below 200. This leads to the proposal of Hypothesis 2.

Hypothesis 2a: If the EPV is below 200, predictive accuracy metrics will be better when running a regression model than when running a ML algorithm.

Hypothesis 2b: If the EPV is above 200, predictive accuracy metrics will be worse when running a regression model than when running a ML algorithm.

Hypothesis 2c: there are no significant differences in predictive accuracy metrics between regression models and ML algorithms regardless of number of variables. This hypothesis is included to account for the potential influence of other factors, such as the complexity of the data or the specific predictive task.

In conclusion, Studies 1 aims to investigate the interplay between sample size and the number of variables in predictive modeling by testing the hypotheses outlined above. Through a systematic exploration of these factors, we hope to contribute to the understanding of when regression models or ML algorithms are more suitable for predictive tasks in real-world scenarios. Both simulated and real-world data will be used to test these two hypotheses.

## Study 2: Incremental Validity

The application of work-related psychological 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., 2021; Meyer et al., 2004). Assessing whether they show incremental validity when establishing criterion-related validity with turnover intention is of interest in the study of machine learning in the I/O psychology field. This is why the following hypotheses are proposed:

Hypothesis 3a: machine learning models trained with work-related psychological constructs will show better predictive accuracy metrics when predicting turnover intention than models trained using biodata.

Hypothesis 3b: the predictive accuracy of machine learning models will increase when work-related psychological constructs are added to the model.

# 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.

# Results

# 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.

# References

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Alzate Vanegas, J. M., Wine, W., & Drasgow, F. (2022). Predictions of attrition among US marine corps: Comparison of four predictive methods. *Military Psychology*, *34*(2), 147–166.

Barandela, R., Valdovinos, R. M., & Sánchez, J. S. (2003). New applications of ensembles of classifiers. *Pattern Analysis & Applications*, *6*, 245–256.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama*, *319*(13), 1317–1318.

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

Boulesteix, A.-L., & Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal*, *56*(4), 588–593.

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

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). *BART: Bayesian additive regression trees*.

Collins, G. S., Groot, J. A. de, Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., et al. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1), 1–11.

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

Craninx, M., Fievez, V., Vlaeminck, B., & De Baets, B. (2008). Artificial neural network models of the rumen fermentation pattern in dairy cattle. *Computers and Electronics in Agriculture*, *60*(2), 226–238.

Cristianini, N., & Scholkopf, B. (2002). Support vector machines and kernel methods: The new generation of learning machines. *Ai Magazine*, *23*(3), 31–31.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.

Deo, R. C., & Nallamothu, B. K. (2016). Learning about machine learning: The promise and pitfalls of big data and the electronic health record. In *Circulation: Cardiovascular Quality and Outcomes* (No. 6; Vol. 9, pp. 618–620). Am Heart Assoc.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.

Douglass, M. J. (2020). *Book review: Hands-on machine learning with scikit-learn, keras, and tensorflow, by aurélien géron: O’reilly media, 2019, 600 pp., ISBN: 978-1-492-03264-9*. Springer.

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

Froud, R., Hansen, S. H., Ruud, H. K., Foss, J., Ferguson, L., & Fredriksen, P. M. (2021). Relative performance of machine learning and linear regression in predicting quality of life and academic performance of school children in norway: Data analysis of a quasi-experimental study. *Journal of Medical Internet Research*, *23*(7), e22021.

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.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

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

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hom, P. W., Mitchell, T. R., Lee, T. W., & Griffeth, R. W. (2012). Reviewing employee turnover: Focusing on proximal withdrawal states and an expanded criterion. *Psychological Bulletin*, *138*(5), 831.

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.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *Icml*, *97*, 179.

Lee, G. R., Fernandez, S., & Lee, S. (2021). An overlooked cost of contracting out: Evidence from employee turnover intention in US federal agencies. *Public Personnel Management*, *50*(3), 381–407.

Luu, B. C., Wright, A. L., Haeberle, H. S., Karnuta, J. M., Schickendantz, M. S., Makhni, E. C., Nwachukwu, B. U., Williams III, R. J., & Ramkumar, P. N. (2020). Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: An analysis of 2322 players from 2007 to 2017. *Orthopaedic Journal of Sports Medicine*, *8*(9), 2325967120953404.

Macijauskien, I., & Stankevi, ivil. (2021). Artificial intelligence solution in human resources management: Case study of chatbot’s implication. *STRATEGICA*, 517.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Marsland, S. (2011). *Machine learning: An algorithmic perspective*. Chapman; Hall/CRC.

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.

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods*, *21*(4), 583.

Mitchell, T. M. (1997). *Machine learning*.

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.

Myers, J. L., Well, A. D., & Lorch Jr, R. F. (2013). *Research design and statistical analysis*. Routledge.

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

Ploeg, T. van der, Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 1–13.

Rąb-Kettler, K., & Lehnervp, B. (2019). Recruitment in the times of machine learning. *Management Systems in Production Engineering*.

Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big data and data science in critical care. *Chest*, *154*(5), 1239–1248.

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*, *116*, 10–17.

Schoenfeldt, L. F. (1999). From dust bowl empiricism to rational constructs in biographical data. *Human Resource Management Review*, *9*(2), 147–167.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15–42.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68–77.

Vijayakumar, R., & Cheung, M. W.-L. (2019). Replicability of machine learning models in the social sciences. *Zeitschrift Für Psychologie*.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J., et al. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, *6*(10), 1122–1129.