

**Influence of Parsimony and Work-related Psychological Constructs in Predicting
Turnover Intention when Using Machine Learning VS Linear Regression**

Diego Figueiras¹

¹ Montclair State University

Author Note

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

ML VS REGRESSION

**Influence of Parsimony and Work-related Psychological Constructs in Predicting
Turnover Intention when Using Machine Learning VS Linear Regression**

**Influence of Parsimony and Work-related Psychological Constructs in Predicting
Turnover Intention when Using Machine Learning VS Linear Regression**

A DISSERTATION

Submitted to the Faculty of

Montclair State University in partial fulfillment

of the requirements

for the degree of Doctor of Philosophy

by

Diego Figueiras

Montclair State University

ML VS REGRESSION

Montclair, NJ

May 2024

Dissertation Chair: Dr. Michael Bixter

ML VS REGRESSION

**Influence of Parsimony and Work-related Psychological Constructs in Predicting
Turnover Intention when Using Machine Learning VS Linear Regression**

A DISSERTATION

Submitted to the Faculty of

Montclair State University in partial fulfillment

of the requirements

for the degree of Doctor of Philosophy

by

Diego Figueiras

Montclair State University

Montclair, NJ

May 2024

ML VS REGRESSION

Dissertation Chair: Dr. Michael Bixter

MONTCLAIR STATE UNIVERSITY

THE GRADUATE SCHOOL

DISSERTATION APPROVAL

We hereby approve the Dissertation

**Influence of Parsimony and Work-related Psychological Constructs in Predicting
Turnover Intention when Using Machine Learning VS Linear Regression**

of

Diego Figueiras

Candidate for the Degree:

Doctor of Philosophy

Graduate Programs:
Counseling

Dissertation Committee:

Certified by:

Dr. Michael Bixter
Dissertation Chair

Dr. Scott Herness
Vice Provost for Research and
Dean of the Graduate School

Dr. Kevin Askew

Date

Dr. John T. Kulas

Copyright©2024 by Diego Figueiras. All rights reserved.

Abstract

This dissertation explores the ongoing debate between traditional statistical regression models and machine learning (ML) algorithms in predictive modeling, focusing on the impact of sample size and the number of variables. Study 1 investigates the relationship between sample size and predictive accuracy, proposing hypotheses regarding the advantages of ML over regression as sample size increases. Additionally, the study examines the influence of the number of variables on predictive accuracy, emphasizing the trade-off between ML and regression models. Using data from the Federal Employee Viewpoint Survey, the research aims to contribute insights into the conditions favoring each modeling approach. Study 2 shifts the focus to incremental validity, exploring whether work-related psychological constructs enhance ML models' predictive accuracy in turnover intention compared to biodata alone. The proposed hypotheses suggest that incorporating psychological constructs will improve predictive accuracy, addressing the “garbage in garbage out” concern prevalent in ML applications. The methods involve diverse datasets, including responses from federal employees and an online survey through Amazon's MTurk, with machine learning algorithms such as Gradient Boosting Trees, Random Forest, Neural Networks, Support Vector Machines, and logistic regression. The dissertation seeks to advance understanding in the field, offering practical insights for researchers and practitioners navigating the dynamic landscape of predictive modeling.

Contents

	i
Chapter 1 (Heading 1)	2
Introduction	3
What Constitutes Machine Learning	6
Theory-based models	6
Empirical-based models	7
Common Terms in Machine Learning	7
Data Hungriness	8
Parsimony	9
Algorithms	9
Decision Trees	9
Classification and Regression Trees (CART)	9
Random Forests (RF)	10
Gradient Boosting Trees (GBT)	10
Bayesian Additive Regression Trees (BART)	11
Neural Networks (NN)	11
Support Vector Machines (SVM) classifiers	12
Evaluation metrics	12
Accuracy	12
Precision	12
Recall/Sensitivity	13
F1-score	13
Area Under the Receiver Operating Characteristics Curve (AUC-ROC)	13
Root Mean Square Error (RMSE)	13
Log-loss	14
Kappa Statistic	14
Geometric mean score (G-mean)	14
Rationale	14
Sample Size and Predictive Modeling	14
Hypotheses	16

Study 1: Sample Size and Number of Variables	16
Study 2: Incremental Validity	17
Methods	18
Study 1	18
Procedure and Participants	18
Data	18
Independent variables	18
Individual characteristics	18
Job characteristics	19
Organizational characteristics	19
Workplace environment	19
Dependent variable	19
Turnover Intention	19
Data analysis	20
Study 2	20
Procedure and Participants	20
Data	21
Work-related psychological constructs	21
Biodata	21
Data Analysis	21
Results	22
Sample Size 100,000 and EPV 10,000	22
Sample Size 10,000 and EPV 1,000	22
Sample Size 1,000 and EPV 100	23
Sample Size 200 and EPV 20	23
Discussion	24
References	25
Appendix	31

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 (Rab-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 regression-based 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 and number of variables. By examining the strengths and limitations of these modeling approaches, this study seeks to shed light on which combination of methods and data characteristics offers greater accuracy and reliability in predicting turnover within resource-constrained environments. Hopefully, this will inform practices in the decision-making process when using these algorithms.

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 (or the logit of the outcome variable in the case of binary logistic regression). 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 without having to specify a theoretical model, 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 small enough.

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 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 the field of I-O psychology, such as job satisfaction, organizational commitment, and conscientiousness, 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 by considering individual attitudes, perceptions, and experiences within the organizational context, as has been demonstrated before (Lee et al., 2021; Meyer et al., 2004).

This second study aims to address this research question by employing comprehensive data. By leveraging turnover intention 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

The delineation between machine learning (ML) and statistical modeling has been extensively explored in the literature (Breiman, 2001b). However, the distinction remains ambiguous, as noted by Collins et al. (2014). Breiman’s seminal work on the “two cultures” serves as a key reference in this discussion (Breiman, 2001b). Breiman highlights the contrast between theory-driven models like regression and 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).

An effective characterization of machine learning involves its emphasis on models that autonomously glean insights from data (T. M. Mitchell, 1997). Unlike regression, machine learning exhibits a higher degree of automation in modeling, particularly in handling nonlinear relationships and interaction terms, as exemplified by Boulesteix (Boulesteix & Schmid, 2014). Achieving this automation often entails the use of highly adaptable algorithms in machine learning, which may necessitate penalization techniques to prevent overfitting (Deo & Nallamothu, 2016). Some scholars depict the relationship between statistical modeling and machine learning as a continuous spectrum (Beam & Kohane, 2018). Meanwhile, others categorize any approach deviating from basic regression models as machine learning, encompassing methods like penalized regression (e.g., LASSO, elastic net) or generalized additive models (GAM) (He & Garcia, 2009). It is essential to highlight that, according to the “automatic learning from data” definition, these methods do not strictly fall under machine learning. In this study, we refrain from classifying them as such.

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 do not 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 the sample data and the error when applying the 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 is referred to as data hungriness. Machine learning algorithms generally 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, specifically turnover intention. 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). The way the algorithm works is that the input node moves to the hidden layer by being multiplied by the weight and having the bias added, each of which is associated with the input. This is done for every single value in the input (also called IV), and the results are put into the activation function formula associated to hidden layers. The activation function is what decides the predicted output, and the formula varies depending on which function is being used. Common activation functions are the sigmoid, soft plus, hyperbolic tangent, and rectified linear units (ReLU). The weights and bias come from fitting the NN to the data beforehand using a

method called “backpropagation.” After scaling the portion of the activation function estimated, the resulting functions are aggregated, then multiplied by the weights associated with the output and the bias also associated with the output is later added, resulting in a function that fits the data and that is used to make predictions based on the input. To model very complex neural networks with high predictive power, one can extend the model by adding more hidden layers, which is what we know as deep learning. 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 (SVMs) represent supervised machine learning algorithms frequently employed for classification tasks. These algorithms operate on the principle of identifying a hyperplane that optimally segregates the dataset into two distinct classes. The assurance of correct classification increases as data points move farther away from this hyperplane. The pivotal data points nearest to the hyperplane, known as support vectors, play a crucial role; their removal would impact the position of the dividing hyperplane, designating them as vital components of the dataset. The margin, denoting the distance between the hyperplane and the support vectors, is a key measure. The objective of SVM is to select a hyperplane with the widest possible margin, enhancing the likelihood of accurately classifying new data. Cross-validation is employed to determine this margin. SVM finds extensive application in tasks such as text classification and image recognition, excelling in managing outliers by permitting 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. However, 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 (Barandela et al., 2003; Kubat et al., 1997).

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 of the current dissertation proposal.

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). This is referred to 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 is worth noting, however, 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). 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). 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 dependent variables for personnel selection, only two articles out of the twenty-eight reviewed had data with sample sizes above 269,999, which is the sample size in Sanchez-Pinto, Venable, et al. (2018)'s study 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 similar? Knowing the threshold at which regression-based methods are

a better alternative is of interest when creating predictive models of turnover intention.

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

This study is based on prior research (Kirasich et al., 2018; Sanchez-Pinto, Venable, et al., 2018), which suggests that ML algorithms outperform regression models consistently when the sample size is large. 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). 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 outperform when the dataset is very large, the following hypotheses for Study 1 are proposed:

Hypothesis 1: As sample size increases, the relative advantage in predictive accuracy of turnover intention between ML and regression-based algorithms will increase.

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 2: As the EPV decreases, the relative advantage in predictive accuracy of turnover intention between ML and regression-based algorithms will decrease.

In conclusion, Study 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 literature review done on 28 articles that studied the predictive accuracy of machine learning models when predicting relevant variables from a job selection context, such as turnover or job satisfaction, noted that only 9% of the studies looked at independent variables that I/O psychologists typically use in predicting job performance in selection contexts, and that the literature has proved are relevant in predicting job performance/turnover, such as job satisfaction, commitment, conscientiousness, organizational identification, perceived organization support and justice, involvement, or engagement (Choi, 2011; Ertas, 2015; Franzen et al., 2022; Kim & Fernandez, 2017; Liss-Levinson et al., 2015, 2015; Maertz Jr et al., 2007; Meyer & Allen, 1991; Pitts et al., 2011; Reeve & Smith, 2001; Schaufeli et al., 2017; Shamir & Kark, 2004; Shore & Tetrick, 1991; Smith, 2005). None of the articles that were predicting turnover or turnover intention used any measure of any of these psychological constructs. The literature in the field of I/O psychology has much to contribute in using the appropriate predictors for building machine learning algorithms that want to predict job performance and turnover. This would help the “garbage in garbage out” criterion problem that is so common with ML.

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). However, when building machine learning models to predict turnover, most studies use data that could better be described as biodata. Biodata in the context of job selection refers to a comprehensive collection of an individual’s personal, educational, and professional information used for employment assessment and decision-making processes. This information typically includes details such as educational qualifications, work experience, skills, achievements, and other relevant data that assists employers in evaluating a candidate’s suitability for a specific position. Assessing whether work-related psychological constructs show incremental validity over biodata when establishing criterion-related validity with turnover intention is of interest in the study of machine learning in the I/O psychology

field. The proposed hypotheses are the following:

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

R (Version 4.3.1; R Core Team, 2023) and the R-packages *kableExtra* (Version 1.4.0; Zhu, 2024), *knitr* (Version 1.42; Xie, 2015), *papaja* (Version 0.1.2; Aust & Barth, 2023), and *tinylabels* (Version 0.2.4; Barth, 2023)

Study 1

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. Different samples will be randomly taken from this dataset to test for hypothesis 1. These will each have a sample of 50, 500, 5,000, 50,000, and the full 557,779.

Additionally, to test for hypothesis 2, samples with varying EPV will be randomly selected from the FAVS. These will be with 10, 50, 100, 200, and 500 EPV.

Data

The variables used in the FEVS are the following:

Independent variables

Individual characteristics

Several factors consistently linked to turnover intention, such as gender (Ertas, 2015 ; Ko & Hur, 2014; Wynen & Op de Beeck, 2014), minority status (Pink-Harper & Rauhaus, 2017; Weaver, 2015), educational achievement (Joo et al., 2012; Liss-Levinson et al., 2015), supervisory position (Ertas, 2015; Pink-Harper & Rauhaus, 2017), and organizational tenure (Lewis, 1992; Pitts et al., 2011; Wynen & Op de Beeck, 2014). Examples of survey items assessing these variables include “How long have you been with the Federal Government (excluding military service)?” and “What is your supervisory status?”

Job characteristics

Seven factors pertaining to job characteristics were chosen: job satisfaction (Ertas, 2015; Kim & Fernandez, 2017; Pitts et al., 2011), personal accomplishment (Ertas, 2015; Wynen & Op de Beeck, 2014), workload (Wynen & Op de Beeck, 2014, 2014), meaningfulness (Ertas, 2015; Kim & Fernandez, 2017; Weaver, 2015), and goal clarity (Ertas, 2015; Ko & Hur, 2014; Weaver, 2015). Sample survey items gauging these variables include “Considering everything, how satisfied are you with your job?”; “My work gives me a feeling of personal accomplishment”; and “My workload is reasonable.”

Organizational characteristics

Eleven factors linked to organizational policies were included: compensation [Wynen and Op de Beeck (2014); Wynen and Op de Beeck (2014)], family-friendly policies (Durst, 1999; Joo et al., 2012), training and skill development (Curry et al., 2005), diversity management (Choi, 2009), and promotion and advancement policy (Cotton & Tuttle, 1986; Johnson et al., 2000; Porter et al., 1974). Sample survey items evaluating these variables include “Considering everything, how satisfied are you with your pay?”; “My supervisor supports my need to balance work and other life issues”; and “How satisfied are you with your opportunity to get a better job in your organization?”

Workplace environment

Thirty-six factors pertaining to organizational environments were selected: organizational satisfaction (Liss-Levinson et al., 2015), coworker relations (Cotton & Tuttle, 1986), supervisory support (Maertz Jr et al., 2007; Smith, 2005), supervisory communication (Joo et al., 2012), managerial trustworthiness (Ko & Hur, 2014), organizational procedural justice (Choi, 2011), creativity (Ertas, 2015), employee empowerment (Kim & Fernandez, 2017; Pink-Harper & Rauhaus, 2017; Pitts et al., 2011), performance-oriented culture (Pitts et al., 2011), loyalty (Weaver, 2015), and organizational support (Liss-Levinson et al., 2015). Sample survey items assessing these variables include “Considering everything, how satisfied are you with your organization?”; “The people I work with cooperate to get the job done”; “My supervisor provides me with constructive suggestions to improve my job performance”; “My performance appraisal is a fair reflection of my performance”; and “I recommend my organization as a good place to work.”

Dependent variable

Turnover Intention

Some scholars advise against using turnover intention as a proxy for actual turnover in federal government agencies. Analyzing age and experience as the unit of analysis revealed strong correlations between turnover intention and actual turnover. Conversely, at the

organizational level, turnover intention exhibited a negative correlation with actual turnover (Cho & Lewis, 2012). Cohen et al. (2016) also observed similar findings, noting that the turnover intention rate is not significantly associated with actual turnover at the organizational level. Despite this caution, organizational-level research on turnover is recommended to utilize actual turnover, while individual-level studies should employ turnover intention at the employee level. Given that this study investigates employees' perceptions of organizational workplace environments and their relationship to turnover intention at the individual level, the use of turnover intention is deemed appropriate for the study's objectives.

Turnover intention is assessed using a single survey item: "Are you contemplating leaving your organization within the next year, and if so, why?" Response options include "No," "Yes, to take another job within the Federal Government," "Yes, to take another job outside the Federal Government," and "Yes, other." Turnover intention was recoded as a binary outcome, with "0" indicating "No" and "1" indicating "Yes, to take another job within the Federal Government," "Yes, to take another job outside the Federal Government," and "Yes, other."

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.

Study 2

Procedure and Participants

A Qualtrics survey including 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 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. The target sample size is 500 respondents.

Data

Work-related psychological constructs

The work-related psychological constructs will include the following FAVS measures: job satisfaction (Ertas, 2015; Kim & Fernandez, 2017; Pitts et al., 2011), organizational satisfaction (Liss-Levinson et al., 2015), supervisory support (Maertz Jr et al., 2007; Smith, 2005), managerial trustworthiness (Ko & Hur, 2014), employee empowerment (Kim & Fernandez, 2017; Pink-Harper & Rauhaus, 2017; Pitts et al., 2011), organizational support (Liss-Levinson et al., 2015), and organizational procedural justice (Choi, 2011).

Additionally, the following publicly available instruments will be added as independent variables: the 3-items Utrecht Work Engagement Scale (UWES-3) (Schaufeli et al., 2017), the refined 9-items Lodahl and Kejner's Job Involvement Scale (JI) (Reeve & Smith, 2001), the 18-items revised Three-Component Model (TCM) Employee Commitment Survey (Meyer & Allen, 1991), the 1-item Graphic Scale of Organizational Identification (GSOI) (Shamir & Kark, 2004), the 8-item Survey of Perceived Organizational Support (SPOS) (Shore & Tetrick, 1991), and the 28-items Concise Conscientiousness Measure (CCM-S) (Franzen et al., 2022).

Biodata

When measuring biodata, the following FAVS measures will be used: gender (Ertas, 2015 ; Ko & Hur, 2014; Wynen & Op de Beeck, 2014), minority status (Pink-Harper & Rauhaus, 2017; Weaver, 2015), educational achievement (Joo et al., 2012; Liss-Levinson et al., 2015), supervisory position (Ertas, 2015; Pink-Harper & Rauhaus, 2017), organizational tenure (Lewis, 1992; Pitts et al., 2011; Wynen & Op de Beeck, 2014), compensation [Wynen and Op de Beeck (2014); Wynen and Op de Beeck (2014)], family-friendly policies (Durst, 1999; Joo et al., 2012), training and skill development (Curry et al., 2005), diversity management (Choi, 2009), promotion and advancement policy (Cotton & Tuttle, 1986; Johnson et al., 2000; Porter et al., 1974), and coworker relations (Cotton & Tuttle, 1986).

Data Analysis

Several machine learning algorithms will be used to train a series of models using biodata features only. These models includes GBT, RF, NN, SVM, and multiple regression. A second series of models that include work-related psychological constructs only in the data will be trained. Predictive accuracy metrics will be compared for each dataset to test

hypothesis 3a. In addition, a baseline series of models will be trained using biodata features only. A second series of models that include work-related psychological constructs only in the data will be added to the models. Performance metrics of each model will be evaluated using accuracy, precision, recall, F1-scores, and AUC-ROC. Bootstrap resampling will be used as a statistical test to assess whether the improvement in performance when adding psychological constructs is statistically significant. An F-test will be used to determine if the change in R-squared is statistically significant for multiple regression. This is done to test for hypothesis 3b.

Results

Sample Size 100,000 and EPV 10,000

A logistic regression analysis was conducted to predict outcomes based on a set of predictors. Overall accuracy across categories was 0.76, as can be seen in table 1. The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.69, suggesting a moderate level of prediction consistency across categories. The performance of a xgboosting model across categories was 0.77, as can be seen in table ??(tab:xgb100k). The macro average precision was 0.74, with a recall of 0.69 and an F1 score of 0.70, suggesting moderate levels of predictive accuracy. The performance of a neural network model across categories was 0.76, as can be seen in table ??(tab:nn100k). The macro average precision was 0.74, with a recall of 0.69 and an F1 score of 0.71, suggesting moderate levels of predictive accuracy. The performance of a SVM model across categories was 0.76, as can be seen in table ??(tab:svm100k). The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.69, suggesting moderate levels of predictive accuracy. The performance of a random forest model across categories was 0.76, as can be seen in table ??(tab:rf100k). The macro average precision was 0.73, with a recall of 0.67 and an F1 score of 0.68, suggesting moderate levels of predictive accuracy.

Sample Size 10,000 and EPV 1,000

Note: nn10k was a bit different from nn100k

A logistic regression analysis was conducted to predict outcomes based on a set of predictors. Overall accuracy across categories was 0.76, as can be seen in Table 6. The macro average precision was 0.75, with a recall of 0.68 and an F1 score of 0.70, suggesting a moderate level of prediction consistency across categories. The performance of a xgboosting model across categories was 0.76, as can be seen in table ??(tab:xgb10k). The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.70, suggesting moderate levels of predictive accuracy. The performance of a neural network model across categories was 0.75, as can be seen in table ??(tab:nn10k). The macro average precision was 0.76, with a

recall of 0.65 and an F1 score of 0.66, suggesting moderate levels of predictive accuracy. The performance of a SVM model across categories was 0.76, as can be seen in table ??(tab:svm10k). The macro average precision was 0.75, with a recall of 0.68 and an F1 score of 0.69, suggesting moderate levels of predictive accuracy. The performance of a random forest model across categories was 0.75, as can be seen in table ??(tab:rf10k). The macro average precision was 0.73, with a recall of 0.66 and an F1 score of 0.68, suggesting moderate levels of predictive accuracy.

Sample Size 1,000 and EPV 100

A logistic regression analysis was conducted to predict outcomes based on a set of predictors. Overall accuracy across categories was 0.75, as can be seen in Table 11. The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.69, suggesting a moderate level of prediction consistency across categories. The performance of a xgboosting model across categories was 0.75, as can be seen in table ??(tab:xgb1k). The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.69, suggesting moderate levels of predictive accuracy. The performance of a neural network model across categories was 0.74, as can be seen in table ??(tab:nn1k). The macro average precision was 0.71, with a recall of 0.67 and an F1 score of 0.68, suggesting moderate levels of predictive accuracy. The performance of a SVM model across categories was 0.76, as can be seen in table ??(tab:svm1k). The macro average precision was 0.74, with a recall of 0.68 and an F1 score of 0.69, suggesting moderate levels of predictive accuracy. The performance of a random forest model across categories was 0.75, as can be seen in table ??(tab:rf1k). The macro average precision was 0.73, with a recall of 0.67 and an F1 score of 0.68, suggesting moderate levels of predictive accuracy.

Sample Size 200 and EPV 20

A logistic regression analysis was conducted to predict outcomes based on a set of predictors. Overall accuracy across categories was 0.77, as can be seen in Table 16. The macro average precision was 0.71, with a recall of 0.7 and an F1 score of 0.71, suggesting a moderate level of prediction consistency across categories. The performance of a xgboosting model across categories was 0.77, as can be seen in table ??(tab:xgb200). The macro average precision was 0.73, with a recall of 0.64 and an F1 score of 0.65, suggesting moderate levels of predictive accuracy. The performance of a neural network model across categories was 0.71, as can be seen in table ??(tab:nn200). The macro average precision was 0.64, with a recall of 0.65 and an F1 score of 0.64, suggesting moderate levels of predictive accuracy. The performance of a SVM model across categories was 0.77, as can be seen in table ??(tab:svm200). The macro average precision was 0.72, with a recall of 0.72 and an F1 score

of 0.72, suggesting moderate levels of predictive accuracy. The performance of a random forest model across categories was 0.77, as can be seen in table ??(tab:rf200). The macro average precision was 0.71, with a recall of 0.7 and an F1 score of 0.71, suggesting moderate levels of predictive accuracy.

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.
- Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown*. <https://github.com/crsh/papaja>
- Barandela, R., Valdovinos, R. M., & Sánchez, J. S. (2003). New applications of ensembles of classifiers. *Pattern Analysis & Applications*, 6, 245–256.
- Barth, M. (2023). *tinylabls: Lightweight variable labels*. <https://cran.r-project.org/package=tinylabls>
- 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*.
- Cho, Y. J., & Lewis, G. B. (2012). Turnover intention and turnover behavior: Implications for retaining federal employees. *Review of Public Personnel Administration*, 32(1), 4–23.
- Choi, S. (2009). Diversity in the US federal government: Diversity management and employee turnover in federal agencies. *Journal of Public Administration Research and Theory*, 19(3), 603–630.
- Choi, S. (2011). Organizational justice and employee work attitudes: The federal case. *The American Review of Public Administration*, 41(2), 185–204.
- Cohen, G., Blake, R. S., & Goodman, D. (2016). Does turnover intention matter? Evaluating the usefulness of turnover intention rate as a predictor of actual turnover rate. *Review of Public Personnel Administration*, 36(3), 240–263.
- 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.
- Cotton, J. L., & Tuttle, J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. *Academy of Management Review*, 11(1), 55–70.
- 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.
- Curry, D., McCarragher, T., & Dellmann-Jenkins, M. (2005). Training, transfer, and turnover: Exploring the relationship among transfer of learning factors and staff retention in child welfare. *Children and Youth Services Review*, 27(8), 931–948.
- 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.
- Durst, S. L. (1999). Assessing the effect of family friendly programs on public organizations. *Review of Public Personnel Administration*, 19(3), 19–33.
- Ertas, N. (2015). Turnover intentions and work motivations of millennial employees in federal service. *Public Personnel Management*, 44(3), 401–423.
- Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, 9(4), 86.
- Franzen, P., Arens, A. K., Greiff, S., Westhuizen, L. van der, Fischbach, A., Wollschläger, R., & Niepel, C. (2022). Developing and validating a short-form questionnaire for the assessment of seven facets of conscientiousness in large-scale

- assessments. *Journal of Personality Assessment*, 104(6), 759–773.
- 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.
- Johnson, J. T., Griffeth, R. W., & Griffin, M. (2000). Factors discriminating functional and dysfunctional salesforce turnover. *Journal of Business & Industrial Marketing*, 15(6), 399–415.
- Joo, Y. J., Lim, K. Y., & Kim, S. M. (2012). A model for predicting learning flow and achievement in corporate e-learning. *Educational Technology & Society*, 15(1), 313.
- 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.
- Kim, S. Y., & Fernandez, S. (2017). Employee empowerment and turnover intention in the US federal bureaucracy. *The American Review of Public Administration*, 47(1), 4–22.

- Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3), 9.
- Ko, J., & Hur, S. (2014). The impacts of employee benefits, procedural justice, and managerial trustworthiness on work attitudes: Integrated understanding based on social exchange theory. *Public Administration Review*, 74(2), 176–187.
- 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.
- Lewis, G. B. (1992). Men and women toward the top: Backgrounds, careers, and potential of federal middle managers. *Public Personnel Management*, 21(4), 473–492.
- Liss-Levinson, R., Bharthapudi, K., Leider, J. P., & Sellers, K. (2015). Loving and leaving public health: Predictors of intentions to quit among state health agency workers. *Journal of Public Health Management and Practice*, 21(Suppl 6), S91.
- 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.
- Maertz Jr, C. P., Griffeth, R. W., Campbell, N. S., & Allen, D. G. (2007). The effects of perceived organizational support and perceived supervisor support on employee turnover. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 28(8), 1059–1075.
- 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., & Allen, N. J. (1991). A three-component conceptualization of organizational commitment. *Human Resource Management Review*, 1(1), 61–89.
- 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., Sablinski, 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.
- Pink-Harper, S. A., & Rauhaus, B. (2017). Examining the impact of federal employee wellness programs and employee resilience in the federal workplace. *Journal of Health and Human Services Administration*, 40(3), 353–387.
- Pitts, D., Marvel, J., & Fernandez, S. (2011). So hard to say goodbye? Turnover intention among US federal employees. *Public Administration Review*, 71(5), 751–760.
- 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.
- Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology*, 59(5), 603.
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Răb-Kettler, K., & Lehnervp, B. (2019). Recruitment in the times of machine learning. *Management Systems in Production Engineering*.
- Reeve, C. L., & Smith, C. S. (2001). Refining lodahl and kejner’s job involvement scale with a convergent evidence approach: Applying multiple methods to multiple samples. *Organizational Research Methods*, 4(2), 91–111.
- 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.
- Schaufeli, W. B., Shimazu, A., Hakanen, J., Salanova, M., & De Witte, H. (2017). An ultra-short measure for work engagement. *European Journal of Psychological Assessment*.
- Schoenfeldt, L. F. (1999). From dust bowl empiricism to rational constructs in biographical data. *Human Resource Management Review*, 9(2), 147–167.
- Shamir, B., & Kark, R. (2004). A single-item graphic scale for the measurement of organizational identification. *Journal of Occupational and Organizational Psychology*, 77(1), 115–123.
- Shore, L. M., & Tetrick, L. E. (1991). A construct validity study of the survey of perceived organizational support. *Journal of Applied Psychology*, 76(5), 637.
- Smith, I. (2005). Achieving readiness for organisational change. *Library Management*, 26(6/7), 408–412.
- 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*.
- Weaver, T. L. (2015). Intent to exit: Why do US federal employees leave? *International Journal of Public Administration*, 38(6), 442–452.
- Wynen, J., & Op de Beeck, S. (2014). The impact of the financial and economic crisis on turnover intention in the US federal government. *Public Personnel Management*, 43(4), 565–585.
- Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. <https://yihui.org/knitr/>
- 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.
- Zhu, H. (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax*. <https://CRAN.R-project.org/package=kableExtra>

Appendix

Table 1*Logistic Regression Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.71	0.45	0.55	11812
1	0.78	0.91	0.84	24862
Accuracy	NA	NA	0.76	36674
Macro Avg	0.74	0.68	0.69	36674
Weighted Avg	0.75	0.76	0.75	36674

Table 2

	Precision	Recall	F1.Score	Support
0	0.70	0.48	0.57	11,812.00
1	0.78	0.90	0.84	24,862.00
Accuracy	NA	NA	0.77	36,674.00
Macro Avg	0.74	0.69	0.70	36,674.00
Weighted Avg	0.76	0.77	0.75	36,674.00

Note. Logistic Regression Evaluation Metrics

Table 3*Neural Network Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.69	0.49	0.57	11812
1	0.79	0.90	0.84	24862
Accuracy	NA	NA	0.76	36674
Macro Avg	0.74	0.69	0.71	36674
Weighted Avg	0.76	0.76	0.75	36674

Table 4*SVM Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.71	0.44	0.54	11812
1	0.77	0.92	0.84	24862
Accuracy	NA	NA	0.76	36674
Macro Avg	0.74	0.68	0.69	36674
Weighted Avg	0.75	0.76	0.74	36674

Table 5*Random Forest Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.70	0.43	0.53	11812
1	0.77	0.91	0.83	24862
Accuracy	NA	NA	0.76	36674
Macro Avg	0.73	0.67	0.68	36674
Weighted Avg	0.75	0.76	0.74	36674

Table 6*Logistic Regression Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.72	0.45	0.55	1079
1	0.77	0.91	0.84	2221
Accuracy	NA	NA	0.76	3300
Macro Avg	0.75	0.68	0.70	3300
Weighted Avg	0.76	0.76	0.74	3300

Table 7*xgboosting Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.71	0.46	0.56	1079
1	0.78	0.91	0.84	2221
Accuracy	NA	NA	0.76	3300
Macro Avg	0.74	0.68	0.70	3300
Weighted Avg	0.76	0.76	0.75	3300

Table 8*Neural Network Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.77	0.35	0.49	1079
1	0.75	0.95	0.84	2221
Accuracy	NA	NA	0.75	3300
Macro Avg	0.76	0.65	0.66	3300
Weighted Avg	0.76	0.75	0.72	3300

Table 9*SVM Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.72	0.44	0.54	1079
1	0.77	0.92	0.84	2221
Accuracy	NA	NA	0.76	3300
Macro Avg	0.75	0.68	0.69	3300
Weighted Avg	0.75	0.76	0.74	3300

Table 10*Random Forest Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.70	0.41	0.52	1079
1	0.76	0.91	0.83	2221
Accuracy	NA	NA	0.75	3300
Macro Avg	0.73	0.66	0.68	3300
Weighted Avg	0.74	0.75	0.73	3300

Table 11*Logistic Regression Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.70	0.45	0.55	110
1	0.77	0.90	0.83	220
Accuracy	NA	NA	0.75	330
Macro Avg	0.74	0.68	0.69	330
Weighted Avg	0.75	0.75	0.74	330

Table 12*xgboosting Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.72	0.44	0.54	110
1	0.76	0.91	0.83	220
Accuracy	NA	NA	0.75	330
Macro Avg	0.74	0.68	0.69	330
Weighted Avg	0.75	0.75	0.74	330

Table 13*Neural Network Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.66	0.46	0.55	110
1	0.77	0.88	0.82	220
Accuracy	NA	NA	0.74	330
Macro Avg	0.71	0.67	0.68	330
Weighted Avg	0.73	0.74	0.73	330

Table 14*SVM Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.71	0.45	0.56	110
1	0.77	0.91	0.83	220
Accuracy	NA	NA	0.76	330
Macro Avg	0.74	0.68	0.69	330
Weighted Avg	0.75	0.76	0.74	330

Table 15*Random Forest Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.69	0.43	0.53	110
1	0.76	0.90	0.83	220
Accuracy	NA	NA	0.75	330
Macro Avg	0.73	0.67	0.68	330
Weighted Avg	0.74	0.75	0.73	330

Table 16*Logistic Regression Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.59	0.56	0.57	18
1	0.84	0.85	0.85	48
Accuracy	NA	NA	0.77	66
Macro Avg	0.71	0.70	0.71	66
Weighted Avg	0.77	0.77	0.77	66

Table 17*xgboosting Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.67	0.33	0.44	18
1	0.79	0.94	0.86	48
Accuracy	NA	NA	0.77	66
Macro Avg	0.73	0.64	0.65	66
Weighted Avg	0.76	0.77	0.74	66

Table 18*Neural Network Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.47	0.50	0.49	18
1	0.81	0.79	0.80	48
Accuracy	NA	NA	0.71	66
Macro Avg	0.64	0.65	0.64	66
Weighted Avg	0.72	0.71	0.71	66

Table 19*SVM Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.58	0.61	0.59	18
1	0.85	0.83	0.84	48
Accuracy	NA	NA	0.77	66
Macro Avg	0.72	0.72	0.72	66
Weighted Avg	0.78	0.77	0.77	66

Table 20*Random Forest Predictive Metrics*

Class	Precision	Recall	F1-Score	Support
0	0.59	0.56	0.57	18
1	0.84	0.85	0.85	48
Accuracy	NA	NA	0.77	66
Macro Avg	0.71	0.70	0.71	66
Weighted Avg	0.77	0.77	0.77	66