ABSTRACT

Title of dissertation: COMPARING THE VALIDITY & FAIRNESS OF MACHINE LEARNING TO REGRESSION IN PERSONNEL SELECTION

Jordan Joy Epistola, Doctor of Philosophy, 2022

Dissertation directed by: Professor Paul Hanges
Department of Psychology

In the realm of personnel selection, several researchers have claimed that machine learning (ML) can generate predictions that can out-predict more conventional methods such as regression. However, high-profile misuses of ML in selection contexts have demonstrated that ML can also result in illegal discrimination and/or bias against minority groups when developed improperly. This dissertation examined the utility of ML in personnel selection by examining the validity and fairness of ML methods relative to regression. Studies One and Two predicted counterproductive work behavior in Hanges et al.'s (2021) sample of Military cadets/midshipmen, and Study Three predicted job performance ratings of employees in Patalano & Huebner's (2021) human resources dataset. Results revealed equivalent validity of ML to regression across all three studies. However, fairness was enhanced when ML was developed in accordance with employment law. Implications for the use of ML in personnel selection, as well as relevant legal considerations, are presented in my dissertation. Further, methods for further enhancing the legal defensibility of ML in the selection are discussed.

COMPARING THE VALIDITY & FAIRNESS OF MACHINE LEARNING TO REGRESSION IN PERSONNEL SELECTION

by

Jordan Joy Epistola

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Psychology.
2022

Advisory Committee:

Professor, Dr. Paul Hanges, Chair

Professor, Dr. Jeffrey Lucas

Professor, Dr. James Grand

Program Director, Dr. Juliet Aiken

Assistant Program Director, Dr. Kenneth Yusko

Assistant Research Professor, Dr. Brian Kim

Acknowledgements

First and foremost, I would like to thank my parents and my adviser for supporting me throughout my Ph.D. career and providing me with the opportunity to hone in on my skills and develop into an I/O Psychologist. I would not be in the position that I am in today if I had not received the support, and I am appreciative of all of the help.

I would also like to thank my dissertation committee and colleagues at the University of Maryland for contributing to the present research and challenging me to become a better thinker and researcher. I would also like to thank my friends for making the journey more enjoyable along the way, and my research assistants for exploring exciting ideas with me.

To the reader, I hope that you find this work as interesting and exciting as I do. I am passionate about this area of research and take pride in my work, and my dissertation is just the beginning of what is next to come in my I/O Psychology career.

Declarations

Funding: The survey data collected by Hanges et al. (2021) and used in Studies One and Two of the present research was funded by the Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Cooperative Agreement Number W911NF-15-2-0093)

Conflict of Interest: The survey data used in Studies One and Two of the present research was collected to partially fulfill the aforementioned grant. However, the analyses and results reported in this study extends the work of Hanges et al. (2021), and my findings were not reported as a part of the grant. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

Machine learning (ML) has recently captivated the interests of researchers and practitioners in industrial-organizational psychology (Lee et al., in press; Oswald et al., 2020). Indeed, ML was listed as the #1 trend in organizations by Society for Industrial-Organizational Psychology (SIOP) members in 2016, 2019, and 2020 (SIOP, 2016; 2019; 2020). ML is a subfield of computer science that uses computer algorithms to create estimations in an iterative fashion that either identify patterns among a set of variables or optimally predict some dependent variable or variables (Lee et al., in press). Depending on the researcher's goal for creating the ML estimation, specific rules can be programmed in the algorithm such that the predictive ability of a single or multiple variables is maximized within the designated constraints.

In a selection context, several researchers have claimed that properly developed ML will generate predictions that are more robust and that will out-predict more conventional estimation methods (Foster et al., 2020; Hastie et al., 2009). However, if developed improperly, ML could produce predictions that may actually harm the organization if used in an actual selection context. For example, improperly created ML algorithms could result in either illegal discrimination or they could incorporate predictors that have been either implicitly or explicitly ruled out by court decisions (Ajunwa et al., 2016; Barocas et al., 2019). The Amazon corporation, for instance, ended use of an automated resume-reviewing recruiting tool because it had bias against female applicants (Dastin, 2018). HireVue, a company that provides automated video interview services to companies, was scrutinized by various groups for "perpetuating discriminatory hiring practices" via its use of facial recognition and analysis software (Barrett, 2021; Harwell, 2019; 2019). Inappropriate and/or controversial use of ML in selection is more

likely to occur when it is applied without detailed consideration of the current practice and legal context and history surrounding selection (Tippins et al., 2021).

Industrial-Organizational (I/O) psychologists are positioned to offer significant value in the use of ML in selection as they can offer substantive expertise in how to properly build algorithms consistent with both the legal context and theories of human behavior in the workplace (Oswald et al., 2020). The realm of personnel selection, in particular, is a major area where ML can be utilized properly or improperly depending on one's knowledge of the domain (Tippins et al., 2021). In selection, the decision of who to hire or promote extends far beyond the predictive validity of algorithms. Organizational goals and issues regarding fairness must be considered to ensure that the selection system is strategically designed to promote hiring goals while being consistent with employment law. A failure to do so may result in hiring individuals lacking critical domain-general traits (e.g., cognitive ability, conscientiousness) that are helpful as the nature of a job changes overtime as well as other traits critical for job effectiveness (e.g., prosocial personality, commitment). In terms of fairness, the organization may face costly legal penalties and lawsuits should the selection system be found to violate employment law. For instance, the *Uniform Guidelines on Employee Selection Procedures* (1978) strongly recommend that practitioners evaluate the fairness of their selection procedures when it is feasible to do so. Further, the shifting burden of proof model, established in the *Griggs v. Duke Power (1971)* indicates that the plaintiff can win their EEO claim against a company by finding a different selection system that is equally as valid but fairer than the existing system. Finally, the Supreme Court case *Connecticut v. Teal (1982)* states that each component of the selection system must be job-related and legally defensible. Should each of these considerations be made when developing

the ML algorithm, it is possible that the result is an upgraded selection procedure that greatly advances an organizations ability to identify and retain quality talent.

Despite the optimistic claims of the potential of ML in personnel selection, recent work in the applied literature (e.g., Allen et al., 2020) casts doubt on its promise. Specifically, Allen et al. (2020) conducted a criterion-related validation (CRV) study where they compared the results of a regression analysis with two frequently used ML models (i.e., ridge regression and random forests regression). Using assessment data on a sample of 208 individuals from a U.S. telecommunications company, they concluded that traditional regression analysis generally outperformed ML in predicting job performance. Though these results are quite striking and intriguing, the generalizability of Allen et al.'s (2020) findings are questionable.

In particular, the sample used in Allen et al. (2020) was small and ML performs best when it is applied to big datasets (Foster et al., 2020). In small data sets ($n < 1000$), ML tends to overfit the data used for training the algorithm (Combrisson & Jerbi, 2015; Vabalas et al., 2019), suggesting that regression would outperform ML in Allen et al.'s small sample. Overfitting results when the algorithm not only captures relationships that are generalizable across samples but also capture relationships that are unique to the sample used to create the algorithm. Overfitting results in substantial loss in predictive power when the overfitted algorithm is used in a new dataset. It is more likely to occur with smaller samples, and certain ML algorithms are more prone to it than others depending on the structure and characteristics of the dataset (Hastie et al., 2009). Considering that ML was developed for use on data with sizes that exceed the capacity of traditional computers to process within an acceptable time and value, it is very likely that overfitting occurred in Allen et al.'s (2020) sample of 208 individuals. In addition, only two ML algorithms were evaluated in their study, raising the possibility that other ML algorithms

would have generated more valid predictions than traditional regression. When interested in prediction, it is recommended to test several different ML algorithms so that the algorithm that demonstrates the best accuracy for the prediction of interest is selected (Hastie et al., 2009). In the present dissertation, I seek to improve upon Allen et al.'s (2020) work by using larger datasets than considered in their study. Further, I compare the validity of a larger collection of ML algorithms than Allen et al. (2020). In this way, I hope to provide a clearer assessment of ML over traditional regression procedures.

However, there is more to consider than the validity of a selection system. It is also important to examine the fairness of the various estimation algorithms. To the best of my knowledge, little research has been conducted on the fairness of ML models relative to regression in personnel selection. Indeed, the dearth of literature on this topic is evidenced by Allen et al.'s (2020) omission of fairness in their analyses, as well as Tippins et al.'s (2021) recent call for research on the legality of ML methods in selection. Frankly, the lack of existing research in this area is quite concerning given the importance that fairness has both in terms of the legal and ethical context surrounding personnel selection (Aiken et al., 2013). To address Tippins et al.'s (2021) call, my dissertation also compared the fairness of the different ML algorithms to each other as well as to the fairness of the traditional regression method used to predict future performance.

Lastly, aside from the validity and fairness of the selection system it is also important to recognize the following four stages surrounding the use of a selection system. First, the *broader strategy stage* consists of considerations made during the planning of the selection system. This stage is guided by the principle of business necessity in that the criterion variable that is focused on (e.g., job performance, turnover, and/or diversity) is selected based on whether it best

advances the businesses' interests and/or ability to perform successfully. Second, the *engineering stage* consists of various engineering decisions made during the development of the selection system. Engineering decisions include the types of algorithms, predictor variables and tuning parameters being used, as well as transformation of any variables. These decisions are critical to consider as they alone can impact both the validity and fairness of the selection system. Third, the *application stage* consists of using the selection system for its intended purpose of selecting individuals to hire and/or promote. Fourth, the *reviewing stage* consists of whether the selection system is indeed fulfilling its designated business strategy purpose and/or if revisions are needed. These four selection stages are critical to recognize as each is subject to different recommendations and legal considerations. For instance, engineering decisions cannot be made during the application stage unless such decisions are made in a manner that did not consider protected class (*Hayden v. Nassau County*, 1996). In my dissertation, I focused specifically on the effect of common engineering decisions made during the engineering stage to allow for a more comprehensive examination of regression to machine learning.

In summary, my dissertation addressed the differential validity and differential fairness of several commonly used ML algorithms when different engineering decisions were made. Specifically, regression and six ML estimation methods were applied to predict two forms of job performance in three studies. Studies One and Two examined counterproductive work behaviors (CWB) in Hanges et al.'s (2021) total sample of 2026 cadets/midshipmen from the three largest U.S. Service Academies. CWB is a contextualized form of job performance (Motowildo & Kell 2013, Sackett, 2002), which consists of intentional behaviors that undermine the goals and interests of an organization such as aggression (physical and verbal), destruction of property, theft, and misuse of time and resources (Fox et al., 2001; Spector et al., 2006). Engineering

decisions pertaining to the concepts of oversampling (i.e., oversampling employees from underrepresented protected groups during the development of selection methods) and content validity (i.e., the use of predictors identified by subject matter experts to be relevant to the criterion) were also examined in Studies One and Two. Study Three then built upon the findings of the prior studies and examined job performance ratings of employees in a human resources dataset developed by Patalano & Huebner (2021). Further, ML interpretability methods were incorporated to further compare the predictor-criterion relationships between regression and machine learning. My dissertation provided important implications for the growing use of ML in personnel selection by examining the validity and fairness of these estimation methods relative to regression analysis. In the next section, I review the I/O Psychology literature on test validity and fairness and discuss the specifics of how validity and fairness were evaluated in my dissertation.

Chapter Two: Validity in Personnel Selection

Validity in Personnel Selection

In I/O Psychology, validity is conceptualized as the degree to which the evidence and theory support the interpretations of test scores for the proposed use of the test (Tippins et al., 2018). In personnel selection, this means that the selection procedure must be supported by some evidence for it to be considered as valid. According to the *Uniform Guidelines on Employee Selection Procedures* (1978), validity evidence is obtained by conducting one of the three kinds of validity strategies: *content-oriented*, *construct, and criterion-related validities.* In content validity, expert judgments are used to connect the test to the important features of the job. In *construct validity*, information is collected to show that the measure being used is assessing the latent variable that it was intended to measure. In particular, construct validity is evident when the assessment under investigation shows both convergent validity (i.e., the test is correlated with other measures that it should be related to) and discriminant validity (i.e., the test is not related to constructs that it is theoretically should not be). Finally, *criterion-related validity* regards evidence demonstrating the test in question actually has a relationship with some criterion of interest on the job. While the *Uniform Guidelines of 1978* treated validity via a trinitarian perspective, psychology has moved away and now takes a unitarian perspective in that all validation strategies provide evidence of the construct validity of the measure (Landy, 1986).

In general, it is recommended that job analysis precede the identification of tests in a selection context (e.g., *Guardians v. Civil Service, 1980; Moody v. Albemarle, 1974; PGA v. Martin, 2001*). Job analysis refers to the detailed examination of a job, "which may include the determination of what is done, how it is done, the context in which it is done (including the strategic importance of the job and its essential functions to organizational success), as well as

the KSAOs (knowledge, skills, abilities and other characteristics) needed to perform the job successfully" (Gutman et al., 2017). Job analysis is critical for content validity, and it might enhance the probability of finding predictors that have significant relationships with the criterion of interest (i.e., criterion-related validity). Conducting a job analysis would enhance the construct validity of an examine because it allows the logical identification of and creation of tests (i.e., content validity) that have the potential to best predict some criterion of interest (i.e., criterion-related validity). Of the content-oriented and criterion-related validity designs, the criterion-related validity approach is the strategy underlying the ML approach. For instance, machine learning can be easily applied during examination of the predictive ability of current selection procedures for actual job performance ratings.

Traditionally, selection procedures generate a predicted score of job performance using OLS regression methods. This process typically entails combining multiple components of the selection procedure (e.g., cognitive ability, personality score, interview score) in a regression equation where the components are regressed onto the criterion of interest (e.g., job performance). The beta coefficients of each component are determined via regressing the predictors on the criterion of interest and/or are selected by subject matter experts. The predicted score of the criterion is then used to inform hiring, firing, and promoting decisions. A significant relationship between the predicted score of the criterion and the actual criterion score provides evidence of criterion validity.

Establishing the validity for predictors is essential because if there is a claim of discrimination against the company, then evidence for validity is needed for the company to defend against these complaints. Indeed, the use of an invalid procedure (e.g., one that is equivalent to random acceptance of applicants) can lead to a gross misuse of human and

economic resources for an organization (Cascio & Aguinis, 2019). Valid selection procedures provide substantial benefits to organizations that contribute to their competitive advantage (Ployhart, 2012). For this reason, it is recommended that the Pearson correlation coefficient be used to communicate the utility of valid selection procedures, as opposed to the coefficient of determination or mean squared error which reduces their perceived importance (Schmidt et al., 1979). Valid selection procedures are highly discriminatory in their ability of differentiating between high and low performing candidates.

In my dissertation, I examined the possibility that machine learning methods may produce enhanced predictive validity than traditional regression methods. Specifically, I compared the criterion validity produced by six machine learning algorithms to that of OLS regression. The Pearson correlation coefficient between the predicted score of the criterion and the actual criterion score was generated for each of the estimation methods. Statistical significance tests of the difference between two correlations were conducted to determine whether a significant difference in validity existed between the estimations. In addition to validity, I also compared the fairness of the estimations produced by the seven methods. In particular, I conducted two important tests of fairness that were widely referenced in the fairness literature in selection- e.g., Cohen's d-statistic and Cleary's model of test bias -to assess whether the estimations produced by these methods were fair with regard to protected class. Fairness was evaluated between genders in Studies One and Two as this was the only protected variable reported by Hanges et al.'s (2021). In Study 3, fairness was examined between gender and race as both protected variables were available in Patalano & Huebner (2021) sample. In the next section, I review the literature on fairness and describe each of the specific tests I used to evaluate the fairness of the methods.

Chapter Three: Fairness in Personnel Selection

Fairness in Personnel Selection

The *Civil Rights Acts of 1964* and *1991* prohibit employment discrimination on the basis of protected class characteristics. Currently in the United States, an individual's race, color, religion, national origin, sex, disability, familial status, sexual orientation, and gender identity are recognized as protected characteristics (*Equal Employment Opportunity Commission*, 2021). Hiring, firing, and promoting on the basis of these protected characteristics is illegal (e.g., *McDonnell Douglas Corp v. Green, 1973*). A violation of disparate treatment is made out when an individual is intentionally discriminated against because of a protected characteristic (e.g., race, age, or gender). Disparate treatment is illegal and results in legal penalties and lawsuits for the organization[1]. Though employment law is quite explicit on prohibiting disparate treatment, the definition of what constitutes a fair procedure beyond this requirement are unclear. Indeed, no single definition of fairness exists as the notion of what one considers to be fair is highly subjective (Cascio & Aguinis, 2019; Landon & Arvey, 2007). In my dissertation, I draw upon outcome-focused and psychometric-focused perspectives of fairness as both have been referenced by the scientific community and the courts to evaluate fairness (e.g., *EEOC v. Dial Corporation, 2006; EEOC v. Ford Motor Company; 2008*).

Outcome-Focused Perspectives of Fairness. The outcome-focused fairness perspective is very popular, and it emphasizes the outcomes produced by the selection procedure on protected groups. The legal concept of adverse impact (AI) is emphasized from this perspective. AI refers to when selection policies, practices, rules, or other systems appear to be neutral but

[1] To demonstrate disparate treatment, the plaintiff has to argue that a) the plaintiff is a member of a minority group; b) the plaintiff applied for the job and was qualified for the job; c) despite the plaintiff's qualifications, they were not hired; and d) the job remained open or they hired someone else who had the same or less qualifications as the plaintiff. Disparate treatment, therefore, is a logical and not statistical argument.

result in a disproportion impact on a protected group (Hanges et al., 2013). The key distinctions between AI and disparate treatment are that a) AI is unintentional while the latter is motivated by discriminatory intent; b) AI uses statistical arguments to demonstrate its existence whereas disparate impact is a logical argument; and c) AI involves multiple individuals being affected while disparate impact can occur for a single individual. AI can be indicative of disparate treatment in many cases, but it is not always illegal in of itself. In particular, AI is allowed when the employer possesses sufficient evidence of the validity behind the targeted selection practice and when no equally valid alternative practices exist.

AI is a function of how the selection practice used, as opposed to an inherent property of the procedure itself (Hanges et al., 2013). The predominant method for detecting AI is the Four-Fifths Rule which states that AI occurs when the selection rate for a certain protected group is less than 80 percent of that of the group with the highest selection rate (e,g, *Uniform Guidelines on Employee Selection Procedures, 1978; Waisome v. Port Authority, 1991*). For instance, consider a situation where 100 candidates apply for a job, of which 50 are men and 50 are women. If ten candidates are hired and seven are men and three are women, AI would occur as the selection rate for women is less than 80 percent of selection rate for men (e.g., 14% for men and 6% for women). The Four-Fifths rule is the first analytic step for AI analyses.

More advanced statistical analyses, such as Cohen's effect size (d) statistic also exist and are preferred by the majority of the scientific community for evaluating AI (Gutman et al., 2017). These tests provide greater utility by providing information on the potential magnitude of AI produced by the procedure (i.e., Cohen's d statistic) as opposed to simply detecting whether AI was present (i.e., Four-Fifths Rule). Moreover, statistical estimates of AI are sufficient for legally establishing whether AI occurred from the selection procedure even when the Four-Fifths Rule

does not detect AI (*Isabel v. City of Memphis*, 2005; Murphy & Jacobs, 2012). Indeed, the Four-Fifths Rule has been criticized by legal scholars and statisticians to detect AI on its own as the Four-Fifths Rule is more prone to type I and type II errors than advanced statistical tests (Boardman, 1979; Gastwirth & Miao, 2009; Sullivan & Feinn, 2012).

Cohen's d-statistic is a standardized effect size measure that examines the magnitude of AI produced by the selection procedure between two groups in terms of standard deviations (Murphy & Jacobs, 2012). Conceptually, the d-statistic, shown in Equation 1, indicates how far the two-group means are from each other in standard deviation units. Analytically, this entails calculating the difference between two sample means of predicted scores divided by the pooled standard deviation of the entire dataset, where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s^2_p is the pooled standard deviation for both groups, s^2_1 and s^2_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes (Cohen, 1988).

$$\text{Cohen's d statistic} = \frac{\bar{x}_1 - \bar{x}_2}{s_p^2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\,s_1^2 - (n_2-1)\,s_2^2}{n_1+n_2-2}}} \qquad \text{Equation 1}$$

Cohen's d-statistic can be classified as small (d = 0.2), medium (d = 0.5), and large (d ≥ 0.8). However, the small, medium, and large classifications of Cohen's d-statistic should only be used as a general guide for interpreting the magnitude of the effect between two groups.[2] For an effect size of 0, the mean of group two is the 50th percentile of group one, and the distributions overlap completely indicating that there is no difference in the magnitude of predicted scores for the two groups of interest (Sullivan & Feinn, 2012). For a large effect of 0.8, the mean of group two increases to the 79th percentile of group one, indicating that the mean of predicted scores for

[2] The context surrounding the task at hand should inform the interpretation of effect size as the d-statistic does not consider other variables such as the reliability of the selection procedure nor diversity of the sample (Sullivan & Feinn, 2012).

group two is higher than 79% of people from group one. AI is inferred when the effect is large as individuals belonging to the group with the higher mean would be selected at disproportionately higher rate. The d-statistic quantifies the size of AI produced by a selection procedure between two groups and is preferred by many legal scholars and statisticians for examining AI (Finkelstein & Levin, 2015; Murphy & Jacobs, 2012).

Although statistical analyses such as the Cohen's d-statistic are sufficient for legally establishing whether AI occurred, it is recommended that they still be used alongside the Four-Fifths rule when determining whether AI occurred[3]. AI has routinely been referenced by the courts in determining the fairness of selection procedures (e.g., *EEOC v. Dial Corporation, 2005; EEOC v. Ford Motor Company; 2008; Griggs v. Duke Power Co., 1964*). When AI is produced from a selection procedure, organizations are required to show that the targeted procedure is supported by appropriate evidence of validity in order to legally justify its use. A failure to do so can result in costly legal penalties and lawsuits for the organization (*Moody v. Albemarle, 1974; Watson v. Fort Worth Bank, 1988; Meacham v. Knolls, 2006*). According to the shifting burden of proof model, AI is also allowed when the employer can demonstrate that no viable, less discriminatory alternative selection procedures exist at the time the selection procedure was used (Hanges et al., 2013).

Though AI has been referenced by the courts to evaluate fairness, it has not been endorsed as the sole criteria for establishing fairness. This is because AI reflects the consequences of the procedure, rather than a property unique to the procedure itself (Hanges et al., 2013). As such, AI does not directly examine whether the selection procedure is inherently fair or unfair, raising the possibility that it is inappropriate for determining the fairness. For this

[3] When there is strong adverse impact, the different approaches typically yield similar conclusions (Miao & Gastwirth, 2013).

reason, the courts and scientific community by and large explicitly reject AI (and other outcome focused definitions of fairness) as constituting fairness in of itself (Tippins et al., 2018). Instead, they endorse relying on psychometric perspectives of fairness which focus on the statistical properties of the selection procedure. In the next section, I briefly discuss this perspective and the methods used to examine it.

Psychometric Perspective of Fairness. The psychometric approach of fairness focuses on eliminating test bias of the selection procedure for different protected groups. This approach draws upon the principle of individual merit, whereby unfairness is determined when the selection procedure systematically overpredicts or underpredicts the performance of any individual or group of individuals (Aguinis et al., 2010; Schmidt & Hunter, 1974). The psychometric approach has received the broadest support by the scientific community (Aguinis et al., 2010; Tippins et al., 2018) and the courts (e.g., *Cormier v. P.P.G Indus*., 1983; *Hamer v. City of Atlanta*, 1989; *United States v. City of Erie*, 2005), though it is recommended that it be used to evaluate fairness conjointly with outcome-focused perspectives[4]. Analytically, test bias is defined as when, "members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of that group that obtains the lower score (Uniform Guidelines on Employee Selection Procedures, 1978)."

The literature on test bias distinguishes at least two of its forms: measurement bias and predictive bias. Measurement bias concerns group differences in the relationship between the

[4] It is recommended that multiple converging sources of evidence of fairness are gathered as each source contributes to a greater understanding of the fairness of the inferences that can be drawn from the selection procedure (Tippins et al., 2018).

selection procedure (e.g., items used to operationalize cognitive ability) and the latent variable to be measured (e.g., actual cognitive ability) (Millsap, 1995). Examinations of measurement bias include examinations of measurement invariance, where a series of confirmatory factor analyses are used to statistically infer equivalence in understandings of constructs among different groups (van de Shoot et al., 2012) and differential item functioning, where item-response theory is used to examine whether members from different groups on the same level of the latent trait have a different probability of giving a certain response to a particular item (Embretson & Reise, 2000).

Predictive bias involves group differences in the relationship between the selection procedure (e.g., predicted job performance) and the criterion variable of interest (e.g., actual job performance). Conceptually, predictive bias establishes fairness by examining whether a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest (Tippins et al., 2018). Analytically, tests of predictive bias entail using Cleary's (1968) model of test bias, which involve a series of moderated multiple regression models, where the criterion measure (i.e., job performance) is regressed on the predictor score (i.e., preemployment score), subgroup membership (i.e., gender) and an interaction term between the two (Aguinis et al., 2010; Cascio & Aguinis, 2019). Predictive bias occurs when a main effect of subgroup membership and/or interaction effect between the predictor score and subgroup membership is observed. The main effect of subgroup membership indicates differences in the starting intercept value of the regression line between the predictor and criterion for subgroups of interest; and the interaction between the predictor score and subgroup membership indicates differences in the slope for subgroups. Both slope and intercept differences between subgroups indicate predictive bias as the relationship between the predictor (i.e., preemployment score) and criterion (i.e., job performance) substantively differs for

subgroups. Figure 1 illustrates the three forms of predictive bias that can emerge between the predictor and criterion variable with gender as the subgroup of interest (Furr & Bacharach, 2014).

In my dissertation, I focused on predictive bias when examining the psychometric approach of fairness as I am interested in applying ML to personnel selection for the purposes of comparing its predictive validity relative to regression. Predictive bias is particularly relevant for supervised ML algorithms as it builds off their primary strength to enhance predictive ability of a target criterion variable by identifying and learning the most generalizable patterns in the dataset (Bzdok et al, 2018). The presence of predictive bias is important for organizations as the presence of it can result in the organization missing out on quality talent to hire, which may harm the organization's ability to achieve company goals and compete long-term. In addition, the organization may face costly legal penalties and/or lawsuits for using a selection system which unfairly predicts job performance for certain protected groups. Thus, my dissertation examined multiple definitions of fairness in the I/O Psychology literature alongside legal precedents in employment law that apply to the use of machine learning in personnel selection. In the next section, I present the machine learning algorithms I intend on using for my dissertation.

Chapter Four: Machine Learning Algorithms

Machine Learning Algorithms

Though several definitions of machine learning (ML) exist, it is widely agreed that ML refers to the application of computer algorithms that are programmed to use a variety of statistical methods to understand and learn from data[5] (Foster et al., 2020). The primary task of these algorithms consists of identifying meaningful relationships and structures of the data and adjusting themselves based on the information "learned" from the data to best predict and depict the data. To do this, the algorithms reference a series of steps and statistical methods that were programmed into the algorithm by a human. The exact steps and statistical methods referenced to learn the data vary depending on the particular algorithm. Broadly speaking, ML algorithms can be classified under the umbrella of unsupervised or supervised learning.

Unsupervised learning refers to algorithms which use statistical methods that do not have a specific target variable to estimate or predict (Hastie et al., 2009; James et al., 2013). Examples of the statistical methods referenced by these algorithms include k-means clustering, principal components analysis and association rules learning (Foster et al., 2020). These algorithms are often used for data exploration purposes as they can "learn" about the data without any prior knowledge of the structure of the data (e.g., number of variables, variables of interest, predictor/criterion designation, etc.). Insights on hidden clusters, groups, or patterns in the data can be learned from these algorithms. Though unsupervised learning algorithms allow for more complex processing tasks compared to supervised learning, they can also be more unpredictable and difficult to interpret due to their lack of structure.

[5] Given the purpose of this paper is to convey the utility of ML in personnel selection, I focus primarily on the intuition behind ML and its algorithms in this section, instead of the statistical theory and formulas that underlie each of the methods.

Supervised learning refers to algorithms that learn from the data using statistical methods which designate a specific target criterion variable. Analytically, this entails taking input pairs of data points (X, Y) where X are designated as the predictor variables and Y as the target criterion variable. Using these pairs of data points as training data, these algorithms reference the steps and statistical methods it was programmed to use to generate a model consisting of some subset of the predictor variables that best predicts the target variable (Hastie et al., 2009). Examples of methods referenced by these algorithms include regression, *k*-nearest neighbors, decision trees, etc. (James et al., 2013). Depending on if the target variable is categorical or numerical, the process of generating the model is referred to as classification-based or regression-based, respectively. This distinction is important as certain algorithms are more appropriate for classification- or regression-based tasks. In general, supervised learning algorithms are best suited for prediction purposes.

One key feature of supervised learning algorithms, which distinguish them from traditional statistical analysis, is the emphasis placed between balancing the validity of the algorithm (i.e., degree to which it predicts the target criterion in the known dataset) with the generalizability of the algorithm (i.e., degree to which it predicts the target criterion in future datasets) (Hastie et al., 2009). In ML, the primary goal is not to find the most optimal model which describes the relationships observed in the known dataset. Rather, the goal is to learn the most generalizable patterns in the dataset which enhance the overall predictive ability of the model for future datasets[6] (Bzdok et al, 2018). Analytically, this process entails incorporating steps and methods within the algorithm that are intended to reduce model overfitting during the model development process. Cross-validation, a resampling procedure which involves

[6] According to Bzdok et al. (2018), the difference between statistics and ML is their principal goal: statistics draws population inferences from a sample, while ML finds generalizable predictive patterns.

partitioning the data into subsets wherein the analysis is conducted on one subset (e.g., training set), and then validating the analysis on the other subsets (i.e., testing set), and then repeating the process on different subsets of the data, is frequently used during this process to reduce model overfitting (James et al., 2013). Model overfitting refers to when the algorithm predicts the criterion variable in the training data too well by learning patterns in the data that are based upon chance instead of meaningful relations between variables. It is most likely to occur with smaller sample sizes and results in algorithms that possess lower predictive validity in future datasets (James et al., 2013). The exact procedure that an ML algorithm uses to combat model overfitting varies for each algorithm.

In the next sections, I review the details underlying the ML algorithms I used to predict CWBs in the Military and performance ratings for employees. Since both CWBs and performance ratings were already measured as numerical variables in their respective datasets, I selected several supervised regression-based algorithms. In particular, I selected three regularization-based algorithms- lasso, ridge, and elastic net -and three tree-based algorithms- decision trees, random forests, and extreme gradient boosting. These six algorithms are widely used in ML and are all suitable for personnel selection due to their enhanced interpretability relative to other algorithms (James et al., 2013). I present the intuition behind the regularization-based algorithms first.

Regularization-Based Algorithms

Regularization-based algorithms refer to a class of supervised learning algorithms that try to minimize the chance of overfitting the data by adding a penalty term to the algorithm (referred to as regularization). This penalty term is done to decrease the complexity of the model by reducing the effect of irrelevant predictors (i.e., predictors that are more affected by chance) in predicting

the criterion of interest. The result is a model that is simpler and less prone to overfitting the training data. Indeed, regularized models are often more robust and generalizable in predicting the criterion, especially when a large number of predictors are incorporated in the model (James et al., 2013).

When applied to the ordinal least square's (OLS) regression model, regularization essentially works by shrinking the coefficient estimates of the regression model. To illustrate this, consider the format of the OLS regression equation (Cohen et al., 2003) which is shown in Equation 2. This equation consists of minimizing the residual sums of squares of prediction (i.e., a measure for the difference between the predicted values of the criterion its actual value) through a linear equation where *N* equals the number of observations, y_i the actual criterion values, B_0 the intercept, *p* the number of predictors, *Bj* the predictor coefficients, and *X* the predictors in the model:

$$Minimum\ (RSS) = \sum_{j=1}^{N}(y_i - B_0 - \sum_{j=1}^{p} B_j X_{ij})^2 \qquad \text{Equation 2.}$$

Analytically, regularization adds a penalty term to the OLS regression equation as shown in Equation 3. This penalty term affects how the minimum RSS is calculated, as well as the magnitude of the coefficients used the regression model (Hastie et al., 2009).

$$Minimum\ (RSS) = \sum_{j=1}^{N}(y_i - B_0 - \sum_{j=1}^{p} B_j X_{ij})^2 + Penalty\ Term \qquad \text{Equation 3.}$$

In terms of the regression coefficients, the penalty term shrinks or *regularizes* the coefficients. As a result, irrelevant predictors are impacted to a greater extent by the penalty term as their coefficient values are typically smaller than that of the more relevant predictors. Several variants of penalty terms exist, and each comprise a different regularization method. Again, these penalties are designed to enhance the parsimony and generalizability of the regression model.

The penalties unique to the three regularization regression methods used in this study, e.g., lasso ridge, and elastic net, are discussed next.

Lasso Regularization. Lasso regularization involves the use of the L1 regularization method to calculate the penalty term (Hastie, 2020). In L1 regularization, the penalty term is computed as the sum of the absolute magnitude of the predictor coefficients, as shown in Equation 4. In this equation, λ is a tuning parameter for the penalty term, *p* equals the number of predictors in the model and *B* represents the coefficient value of the predictor.

$$L1\ Penalty\ Term = \lambda \sum_{j=1}^{p} |B_j| \qquad \text{Equation 4}$$

As with OLS regression, lasso regularization aims to minimize the residual sums of squares to establish predictive validity. As shown in Equation 5, the key difference is that the lasso model also incorporates the L1 penalty term during this process such that the term is added to the regression equation.

$$Minimum\ (RSS) = \sum_{j=1}^{N} (y_i - B_0 - \sum_{j=1}^{p} B_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |B_j| \qquad \text{Equation 5}$$

The L1 penalty term can take a wide range of values as it is controlled by the λ tuning parameter. When λ = 0 the model is equivalent to the OLS regression model; and when | λ | → ∞, the penalty parameter increases resulting in the coefficients of the model being shrunken towards zero and/or equaling zero whereby they are eliminated from the model (Hastie et al., 2009). Accordingly, lasso regularization can perform variable selection processes as increases in the value of λ can result in predictors having coefficient values equal to zero. The optimal value of λ is generated via the cross-validation resampling procedure. Lasso regularization can produce models that are less prone to overfitting the data as irrelevant predictor variables are eliminated (Ranstam & Cook, 2018; Xu et al., 2010).

Ridge Regularization. Ridge regularization consists of using the L2 regularization method to calculate the penalty term as opposed to the L1 method. The L2 regularization method is computed as the sum of the square magnitude of the predictor coefficients, where λ is a tuning parameter for the penalty term, *p* equals the number of predictors in the model and *B* represents the coefficient value of the predictor (Hastie, 2020):

$$L2\ Penalty\ Term = \lambda \sum_{j=1}^{p} {B_j}^2 \qquad \text{Equation 6}$$

$$Minimum\ (RSS) = \sum_{j=1}^{N} (y_i - B_0 - \sum_{j=1}^{p} B_j X_{ij})^2 + \lambda \sum_{j=1}^{p} {B_j}^2 \qquad \text{Equation 7}$$

As evident when comparing Equation 4 with Equation 6, the difference between the lasso and ridge regularization method is that the squared sum of coefficients is used to calculate the penalty term for the ridge method as opposed to the absolute sum of coefficients for the lasso method. Conceptually, this distinction is significant because it can result in the coefficient estimates approaching zero instead of equaling zero when the tuning parameter $|\lambda| \rightarrow \infty$ for the ridge model (Hastie et al., 2009). Accordingly, ridge regularization does not perform variable selection processes as the ridge will always include all of the variables in the model even as $|\lambda| \rightarrow \infty$ (Hastie et al., 2009). Ridge regularization is most effective when the researcher desires all variables being included in the model and/or when there is multicollinearity among the predictors (Warton, 2012).

Elastic Net Regularization. The elastic net regularization method combines the lasso (L1) and ridge (L2) methods in computing the penalty parameter. Specifically, the elastic net incorporates an additional tuning parameter α that determines the extent to which the lasso and ridge methods are used to calculate the penalty term of the elastic net, where λ is a tuning parameter for the entire penalty term, *p* equals the number of predictors in the model and *B* represents the coefficient value of the predictor (Hastie, 2020 ; Zou & Hastie, 2005):

$$Elastic\ Net\ Penalty\ Term = \lambda \sum_{j=1}^{p} (\alpha |B_j| + (1-\alpha)\ {B_j}^2) \qquad \text{Equation 8}$$

$$Minimum\ (RSS) = \sum_{j=1}^{N} (y_i - B_0 - \sum_{j=1}^{p} B_j\ X_{ij})^2 + \lambda \sum_{j=1}^{p} (\alpha |B_j| + (1-\alpha)\ {B_j}^2) \text{ Equation 9}$$

As seen in Equation 8, the lasso and ridge methods are linearly combined in the elastic net penalty term. The additional tuning parameter α can be predetermined should the researcher desire the lasso or ridge method to have a greater contribution to calculating the penalty term. Conversely, α can be generated via the cross-validation resampling procedure should the researcher desire an optimal value of α for prediction of the target criterion variable. I chose to use the optimal value of α in my dissertation due to my interest in comparing the predictive validity of ML methods with regression. Conceptually, the elastic net can be thought of as an optimal compromise between the lasso and ridge methods. Indeed, work by Zou & Hastie (2005) reveal that the ridge portion of the elastic net method (i.e., the second term in Equation 8) encourages highly correlated predictors in the model to be averaged, and the lasso portion (i.e., first term in Equation 8) encourages greater refinement and distinction among the coefficients of the averaged predictors (Hastie et al., 2009).

Though the elastic net regularization method represents an advancement over the lasso and ridge methods, it is important to remember that it along with the other two regularization methods are all still subject to the linearity assumption of the OLS regression model. As one can imagine, this assumption of linearity can be problematic when the true relationship between the predictor variables and criterion variables are non-linear in nature. In such scenarios, non-linear ML methods such as decision trees can produce enhanced predictive ability of the criterion. Decision tree methods are widely used in machine learning as they are highly interpretable and capable of producing valid and generalizable predictions. In the next section, I review the details of the three decision tree methods I intend on using in my dissertation.

Decision Tree-Based Algorithms

Decision tree-based algorithms refer to a class of methods that rely upon the decision tree predictive model. The decision tree model works by generating a non-linear mapping of the predictor variables (e.g., cognitive ability, conscientiousness, motivation, interview performance) to the criterion variable (e.g., predicted job performance). Conceptually, this entails partitioning out the variability of the criterion variable into a set of smaller components whereby a simple model of the predictors is fitted for each component (Hastie et al., 2009). Specifically, recursive binary partitioning is conducted during this process such that each component separates the criterion variability into two partitions (i.e., binary) and then continues parsing out the variability into two partitions (i.e., recursive) until some stopping rule is reached. Typical stopping rules of recursive binary partitioning include either: (a) a minimum number of cases that can exist within each partition of the criterion variability; (b) the maximum number of times that the criterion variability can be partitioned; and/or (c) a minimum amount of criterion variability needed for each partition to occur (James et al., 2013). The stopping rules are intended to balance the validity of the decision tree model with the potential of overfitting it to the data.

Analytically, recursive binary partitioning begins by first selecting and identifying a predictor variable (e.g., cognitive ability) and a split-point on that predictor (e.g., cognitive ability score of 30 out of 50) that best explains the variability in the criterion variable (e.g., job performance). All possible variables (e.g., cognitive ability, conscientiousness, motivation, interview performance) and values for said variables are examined in order to determine which variables and split-points are optimal. The result of the first split is two regions of criterion variability, whereby the mean of the criterion in each region is modeled (Hastie et al., 2009). The process then repeats as one or both of these regions are split into two more regions based on

some predictor variable combination (e.g., conscientiousness) and their split-points (e.g., conscientious score of 80 out of 100) with the goal of optimally minimizing the remaining criterion variance. Again, the splitting process continues, until the designated stopping rule is reached.

Figure 2 shows an example of a decision tree model that is designed to predict employee job performance. The stopping rule in this example consists of a maximum number of times that the variability in job performance can be partitioned. Specifically, the allowed number of partitions or *depth* for the tree is set to 2. Cognitive ability and conscientiousness are identified as the predictors that best explain job performance in this example. The first split occurs for the cognitive ability predictor at the split-point of 30. This split results in separating employees who have cognitive ability scores less than 30 from those who have scores greater than or equal to 30. The next splits occur for the variables of cognitive ability and conscientiousness. For employees who previously had cognitive ability scores less than 30, the cognitive ability predictor is again referenced to split the group at scores of 20. Terminal node 1 depicts the distribution of job performance ratings for employees that had cognitive ability scores less than 20 and terminal node 2 shows the distribution for employees that had scores greater than or equal to 20 but less than 30. Regarding the group of employees that had cognitive ability scores greater than or equal to 30, the conscientiousness predictor is used to split the group at conscientious scores of 80. Terminal node 3 illustrates the distribution of scores for employees who had cognitive ability scores greater than or equal to 30 and conscientiousness scores less than 80; terminal node 4 illustrates the distribution for employees who instead had conscientious scores greater than or equal to 80.

I will use the stopping rule in my dissertation recommended by Hastie et al. (2009). Specifically, recursive binary partitioning of variability in the criterion variable continues until the minimum sample size of 5 is reached for each region in the split. The result of this stopping rule is usually a very large decision tree. The large decision tree is then pruned using a procedure referred to as cost-complexity pruning. Conceptually, cost-complexity pruning consists of calculating a complexity criterion parameter (C_p) which determines the optimal size of the decision tree by setting a minimum improvement value needed in predicting the criterion variable for each split of the model to be made. Each split of the large decision tree that does not meet the minimum improvement value designated by the C_p is removed or *pruned* from the final decision tree model. Thus, larger values of C_p result in smaller decision trees, while smaller C_p values result in larger trees. The optimal C_p value is determined using the cross-validation resampling procedure. The optimal C_p value maximizes the generalizable predictive ability of the model by addressing the interplay between model overfitting and validity in the known dataset.

Overall, the decision tree algorithm is widely used in ML due to its predictive accuracy and robustness (Foster et al., 2020). In addition, the algorithm is considered to be highly interpretable relative to other ML algorithms. For instance, the final decision tree model can be plotted to show the order, exact variables and accompanying split-points used to predict the criterion variable. The decision tree algorithm is also equipped with a measure of variable importance which shows the relative impact of each predictor variable in predicting the criterion variable within the model[7]. Figure 3 contains an example of a variable importance plot. As shown in this figure, this plot shows the relative contribution of each predictor in the prediction of the

[7] Variable importance is calculated by the sum of the decrease in error when split by a variable, where said value is then compared to other variables to determine the relative importance of variables to one another (Foster et al., 2016).

criterion. For example, in Figure 3, the two most important predictors are var1 and var2. There is a substantial drop in importance for v5 which is the third most important variable. Variable importance scores are very informative for theory-building efforts. The variable importance scores can also inform further analyses where predictors with high-ranking scores may be chosen for further investigation, or for building a more parsimonious model (Kazemitabar et al., 2017).

Lastly, decision tree algorithms can be further unpackaged by using model-agnostic methods for interpretability such as partial dependence (PD) plots and individual conditional expectation (ICE) plots. PD and ICE plots illustrate the directionality and (non)linearity between a single predictor and the criterion variable within an algorithm (Molnar, 2021). Figure 4 shows an example of a PD plot between the variable 1 predictor and the criterion variable. As shown in Figure 4, the relationship between variable 1 and the outcome appears to resemble a negative logarithmic relationship, where the magnitude of the relationship decreases as variable 1 increases from 0 to 30. Additional model-agnostic methods such as local surrogate models (LIME) can also be applied to depict interaction effects for decision tree algorithms (Molnar, 2021).

Though the decision tree algorithm is quite useful in prediction, several advancements in the algorithm have been made which demonstrate even greater predictive ability. Ensemble methods are one class of algorithms that build upon the decision tree model and produce more accurate predictions. Essentially, ensemble methods work by combining the results of multiple decision tree models (i.e., an *ensemble* of models) to reach a prediction as opposed to relying on one single model (Foster et al., 2020). In particular, the ensemble enhances prediction by addressing the randomness associated with building a decision tree model: e.g., the order and selection of predictors in the decision tree model may vary as a function of what variable was previously

used to partition the criterion. Several procedures for addressing the randomness of a single decision tree model exist for ensemble methods. In the next section, I discuss the logic behind the two decision tree-based ensemble algorithms used in this study.

Random Forests. The random forests algorithm develops an ensemble of decision tree models by building upon a bootstrapping process referred to in the ML literature as bagging. In bagging, each of the decision tree models is built from a sample drawn with replacement from the training set. Random forests take bagging a step further by randomizing the selection of predictors available for the decision tree model. Specifically, instead of selecting among all of the predictor variables to build the decision tree as done in bagging, a random limited subset of predictors is available to use at each step/split of the tree. By using both bagging and a random subset of predictors available to build the decision tree, the ensemble of tree models that are produced by the random forests algorithm approaches statistical independence from one another. Statistical independence is critical as the final step of the random forests algorithm consists of combining the result of each of the decision tree models to form a prediction. In particular, statistical independence of the models enhances the generalizable predictive ability of the algorithm as it is less prone to overfit the data relative to a single decision tree model or bagged ensemble of models. Indeed, the random forests algorithm generates predictions that are more valid and robust than single and bagged decision tree models (Foster et al., 2020). The random forests algorithm is conceptualized as a parallel ensemble method as each of the decision tree models are generated separately from one another, wherein the outputs of each model are eventually aggregated using the mean.

Extreme Gradient Boosting. As opposed to the parallel process utilized by the random forests algorithm, the extreme gradient boosting algorithm works by using a sequential ensemble

method process wherein each decision tree model is built in a stage-wise fashion to generate predictions of the criterion. In particular, the process referred to as boosting occurs, wherein the models are built sequentially by correcting the errors made by the previous models until no further improvements can be made (Chen & Guestrin, 2016). The "boosting" element of this process occurs by how the errors are corrected. Namely, the weights of the inaccurate predictions from the prior tree models are artificially boosted with a higher weight such that the next tree models are better positioned to address or "correct" these errors.

For extreme gradient boosting specifically, the errors are corrected via minimizing a differential loss function[8] which is designed to correct the errors of the prior model with the new model, while also accounting for not overfitting or "overcorrecting" the new model (Foster et al., 2020). Again, the learning process is sequential, whereby the algorithm generates each new model from the errors produced from the prior model until error can no longer be decreased[9]. When error can no longer be decreased, the final step of the extreme gradient boosting algorithm takes place. This step consists of a taking a weighted sum of each of the models produced by the algorithm, wherein the models that produce a smaller differential loss function are given a larger weight in producing the final prediction. The logic behind extreme gradient boosting and other boosting algorithms is to combine the results of a sequence of poor models- i.e., models that are only slightly better than random guessing, such as small decision trees -into one final ensemble algorithm that is highly predictive of the criterion variable (Chen & Guestrin, 2016).

[8] Chen & He (2021), the developers of the extreme gradient boosting algorithm, provide a detailed review of how the differential loss function is calculated and other technical features of the algorithm.

[9] More specifically, the extreme gradient boosting algorithm builds a sequence of models from the training dataset that tries to minimize the difference between the predicted value of the criterion with the actual criterion value that was produced from the prior model in a manner where this difference is minimized but model overfitting is also accounted for.

My dissertation used these decision tree ML algorithms, as well as the regularization-based algorithms discussed previously. Specifically, I conducted three studies that compared the validity and fairness of these estimation methods relative to that of regression in a Military sample (Studies One and Two) and human resources sample (Studies 3). Further, I examined how each of the seven estimation methods are affected when different practices are implemented with the intent of enhancing the legal defensibility of the estimation methods. For instance, I applied oversampling in Study One by manipulating the representation of male and female cadets/midshipmen in the training/validation datasets; focused on the content validity of the predictors referenced by the methods in Study Two; and focused on the legal defensibility and interpretability of the estimation methods in Study Three. In the next section, I discuss the importance of legal defensibility in selection, and then describe how validity and fairness are evaluated across the three studies in my dissertation.

Chapter Five: Legal Defensibility & Present Research

In the realm of personnel selection, legal defensibility of the selection method is almost as important as its own validity and fairness. From a practitioner's viewpoint, validity and fairness are simply forms of evidence accepted by the courts that provide legal defensibility of the method as well as a justification of its use in selection. For this reason, it is advised that practitioners seek to enhance the legal defensibility of their methods over and beyond minimum legal requirements wherever possible. Practices that increase both the actual and perceived legal defensibility for the selection methods are encouraged to be implemented[10]. Thus, the goal of the practitioner is multifaceted: to produce a legally defensible selection tool that is (a) valid, (b) fair and (c) not prone to potential legal challenges of ineffectiveness and/or bias from others in the first place.

As previously discussed, the Civil Rights Acts of 1964 and 1991 prohibits the use of protected characteristics (e.g., race, color, religion natural origin, sex, disability, familial status, sexual orientation, and gender identity) when selecting and/or assessing employees (*EEOC v. FAPs Inc*., 2014). As such, selection methods that advantage or disadvantage these characteristics, even with the intent of increasing validity and/or fairness, are illegal. For instance, the practices of racial quotas (i.e., numerical requirements for hiring individuals of a specific protected group); race-correcting (i.e., taking race into account to increase prediction); race-norming (i.e., applying a within-group score conversion to account for the protected class of the test-taker); and adversity scores (i.e., boosting scores for a protected group to account for systemic disadvantages associated with past and present discrimination) are outlawed by the

[10] Potential legal challenges can come at a cost to the company's reputation and/or finances, especially when the initial challenge is granted merit by the courts and legal counsel for the company is required.

courts when selecting individuals in an employment context[11] (Greenlaw & Jensen, 1996; Yuvraj, 2019). Although these practices can enhance prediction and address systemic inequities, they are inappropriate for use because they reward or penalize protected characteristics in the selection process. Accordingly, it is critical that practitioners implement legally defensible practices when developing selection methods.

I propose one legally defensible practice that can be applied to machine learning methods that does not reward or penalize protected characteristics. In particular, I propose that the distribution of employees from different protected backgrounds be considered when the training/validation datasets used to develop the estimations are created. My proposal is consistent with the current practice of oversampling in selection, i.e., oversampling employees from underrepresented protected groups during the development of selection methods. Typically, oversampling is done when few employees of a protected group are present in the organization. The logic behind oversampling is to ensure that employees from underrepresented protected groups are also examined by the practitioner. Oversampling enhances legal defensibility by fostering perceptions of fairness via the pursuit of more adequate representation of protected groups during the development of the selection method. Moreover, oversampling does not reward nor penalize protected characteristics, nor references them in the examination of job performance. Rather, it simply ensures adequate representation for protected groups in an attempt to make the selection method fairer and more generalizable for individuals of all protected backgrounds. I applied oversampling in Study One of my dissertation and examined

[11] In the higher education context, courts have recently ruled that it is legal for higher education professionals to consider protected characteristics in the admissions process (*Fisher v. University of Texas, 2013; Fisher v. University of Texas, 2016; Students for Fair Admissions v. President and Fellows of Harvard College, 2020; Students for Fair Admissions v. University of North Carolina, 2021*).

how the distribution of employees from different protected classes in the training/validation datasets impact estimation method validity and fairness in predicting Military CWBs.

Another practice that can be applied to enhance the legal defensibility of machine learning methods in selection consists of using variables that are supported via content validity. Content validity refers to the extent to which the variables referenced by the selection method are indeed measuring appropriate content of the job that the selection method was designed to predict. In the context of selection, content validity is typically established by subject matter experts who perform a work analysis (i.e., a detailed examination of a job, which may include the determination of what is done, how it is done, the context in which it is done, as well as the KSAOs to perform the job successfully) (Gutman et al., 2017). Content validity is critical for determining the legal defensibility of the selection method, especially when adverse impact is detected (*Bradley v. City of Lynn*, 2006; *Moody v. Albemarle*, 1973). Indeed, the courts have ruled that content validity is critical for establishing legal defensibility of the selection method, and that multiple sources of validity are preferred (*Guardians v. Civil Service*, 1980; *Gulino v. New York State Education Department*, 2012; *Smith v. City of Boston*, 2015).

In Study Two, I examined how the content validity of predictors affected the validity and fairness of the seven estimation methods in predicting Military CWBs. In Study Three, I applied my findings from Study One and Study Two to a human resources dataset developed by Patalano & Huebner (2021). Specifically, I used oversampling as well as content valid predictors to develop a legally defensible selection tool for employees. The validity and fairness of the seven estimation methods were again examined in Study Three. However, greater emphasis was placed in comparing regression to the six other machine learning estimation methods. For instance, I incorporated variable importance analyses to the machine learning estimation methods to show

how they can be communicated for legal defensibility purposes. In the next section, I present how validity and fairness were evaluated in all four studies of my dissertation.

Validity

The validity of the seven estimation methods used in all four studies of my dissertation (e.g., regression, decision tree, random forests, extreme gradient boosting, ridge regularization, lasso regularization, elastic net regularization) was examined using the Pearson product moment correlation coefficient between predicted criterion scores and actual criterion score. In particular, Studies One and Two examined predicted CWB scores with actual CWB scores, whereas Study 3 compared predicted job performance ratings with actual job performance ratings. The Pearson *r* coefficient was used to interpret the effectiveness of the selection methods, as opposed to other metrics such as the coefficient of determination or mean squared error[12]. Steiger's Z-Test for the difference between two dependent correlation coefficients was used to determine whether the validity coefficients of the machine learning estimation methods were significantly different than that of regression.

Fairness

Fairness was examined between genders in Studies One and Two as this was the only protected characteristic that was reported by all 2026 cadets/midshipmen in Hanges et al.'s (2021) sample. Study 3 examined fairness between racial minorities and majorities for emergency response workers. Two widely referenced fairness tests in the I/O psychology literature were conducted on the seven estimation methods of my dissertation. Specifically, Cohen's d-statistic was used to measure the outcome-focused perspective of fairness, and the Cleary Model was used to assess the psychometric-focused perspective of fairness.

[12] Schmidt et al. (1979) recommends using the Pearson *r* coefficient for communicating the importance of the selection method to those less familiar with statistics.

Cohen's d-Statistic. Cohen's d-statistic calculated the magnitude of AI produced by the selection procedure between two groups within a protected class in standard deviation units. Positive effect sizes indicated that the expected performance scores generated by the method was greater for men than women (Studies 1-3) and racial majority employees than racial minority employees (Study 3). Negative effect sizes indicated the opposite where expected scores were greater for women than men (Studies 1-3) and racial minority employees than racial majority employees (Study 3). AI was inferred when the effect size is large. The group negatively affected by AI was determined by the directionality of the effect size (e.g., women or racial minority employees are disproportionately affected when there is a positive large effect size).

Cleary Model of Test Bias. Cleary's model of test bias was conducted to examine whether predictive bias of the methods existed between the protected groups of interest. As discussed earlier, this test consisted of three moderated multiple regression models, where the criterion measure was regressed on the predictor score, subgroup membership and an interaction term between the two (Aguinis et al., 2010). Predictive bias occurred when a main effect of subgroup membership and/or interaction effect between the predictor score and subgroup membership was observed. The main effect of subgroup membership indicated differences in the slope of the regression line between the predictor and criterion for subgroups of interest; and the interaction between the predictor score and subgroup membership indicated differences in the starting intercept value for subgroups. Both slope and intercept differences between subgroups indicates predictive bias as the relationship between the predictor (i.e., preemployment score produced by the method) and criterion (i.e., expected performance score) substantively differs for subgroups. Validity and fairness were evaluated in the same manner across all four studies of my dissertation.

Chapter Six: Study One

In Study One, I examined the validity and fairness of predicting CWBs by the seven estimation methods when oversampling was applied to the training/validation dataset split. In particular, I examined four scenarios whereby the distribution of male and female cadets/midshipmen varied in the training/validation dataset split. This study had two objectives: first to examine the performance of the seven estimation methods in predicting CWBs in the Military, and second to examine the extent that the representation of male and female cadets/midshipmen from different Military Academies in the training/validation datasets influenced estimation method performance. Though my dissertation was exploratory, I anticipated that the manner by which male and female cadets/midshipmen were distributed in the training/validation datasets would impact the validity and fairness of the seven estimation methods.

Method

Participants and Procedure.

Study One utilized the total sample of military cadets/midshipmen that was reported in Hanges et al. (2021). As reported in their study, a total of 2,026 military cadets/midshipmen from the United States Army (N = 273), Navy (N = 260), and Air Force (N = 1493) completed questionnaires on a variety of psychological constructs that were theorized to affect CWBs. Five individual-level psychological variables reflecting stable differences in personality and behavior were used in Study One. These psychological variables were selected by subject matter experts and fall into the "Other Characteristic" category of the KSAO framework.

To maintain anonymity of the participants, the three Military Service Academies required participants to only provide demographic information regarding their gender, formal leadership

experience, socioeconomic status, and political orientation[13]. The majority of cadets/midshipmen were male (64.3%) and did not have formal leadership experience (63.5%). Approximately 3.3% of the cadets/midshipmen self-identified as coming from lower class Social Economic Status (SES) background, 11.8% self-identified as from a working-class SES background, 39.7% self-identified as from a middle-class SES background, 40.5% self-identified as from an upper-middle class SES background, and finally, 4.7% self-identified as from an upper-class SES background. In terms of political orientation, participants reported their beliefs towards social and economic issues separately. Regarding social issues, 4.7% of cadets/midshipmen self-identified as very liberal, 21.8% self-identified as liberal, 34.4% self-identified as moderate, 29.4% self-identified as conservative, and 9.8% self-identified as very conservative. For economic issues, 1.3% of cadets/midshipmen self-identified as very liberal, 7.6% self-identified as liberal, 33.6% self-identified as moderate, 43.7% self-identified as conservative, and 13.7% self-identified as very conservative.

Oversampling Manipulation

The oversampling manipulation consisted of splitting the training/validation datasets used to develop the estimation methods in four different ways. In particular, each of the conditions used a 75-25 ratio to split the training/validation datasets but varied in terms of how male and female cadets/midshipmen were distributed across the datasets. Condition one randomly separated male and female cadets/midshipmen in the training/validation datasets. Condition two randomly split the training/validation datasets in a manner where the distribution of men and women in the training/validation datasets were equal to that in the total dataset (e.g..

[13] A little over 25% of participants provided race/ethnicity information. Of the cadets/midshipmen who provided this information, the majority were White (64.7%), followed by Multiracial (12.0%), Asian-American (9.7%), African-American (7.1%), and Latino (6.2%). The remaining race/ethnicity categories each contained less than 2% of the respondents.

approximately 64.3% male in both datasets). Condition three randomly split the training/validation datasets whereby the distribution of cadets/midshipmen from each academy were equal to that in the total dataset. Lastly, both the proportions of men and women as well as cadets/midshipmen from each academy in the total dataset were accounted for when the training/validation datasets were split in condition four. The variables used to develop the seven estimation methods remained consistent across conditions.

Variables.

Predictors:

Prosocial personality. Prosocial personality refers to series of personality characteristics that are associated with an individual engaging in a pattern activity that is motivated to help others (Penner et al., 1995). The social responsibility characteristic of prosocial personality was measured in Hanges et al.'s (2021) study as this characteristic was highly relevant to the cadet/midshipmen population as it is in congruence with the culture of responsibility at the Academies. For instance, the item "no matter what my fellow cadets have done, there is never a reason for me taking advantage of them" reflected the norms of teamwork and camaraderie at the Academies; and the item "using social media to point out another cadet's problems/screw-ups/short-comings" referenced the norm of using social media to facilitate responsible behavior at the Academies. A single factor multigroup confirmatory factor model between men and women for this measure demonstrated excellent fit to the data[14] ($\chi^2(28) = 56.77$, CFI = .96, RMSEA = .03, SRMR = .03).

[14] Hu and Bentler's (1999) suggested values of model fit were used to evaluate model fit: e.g., SRMR values of approximately 0.08 or below, RMSEA values of approximately 0.06 or below, and CFI values of approximately 0.95 or above. Model fit was established when one or more of the aforementioned values met Hu and Bentler's (1999) guidelines.

Entitlement. Entitlement was measured using Campbell et al.'s (2010) measure of psychological entitlement. This personality measure assessed the degree to which individuals possess a stable and pervasive sense that he/she deserves more and is entitled to more in life than others. Prior work demonstrated that Campbell et al.'s (2010) measure is a unitary construct that has better construct validity than other entitlement measures. Excellent fit to the data was observed in a single factor confirmatory factor model between men and women for this measure ($\chi^2(4) = 37.12$, CFI = .96, RMSEA = .09, SRMR = .03).

Social dominance. Social dominance was measured using Ho et al.'s (2015) two factor measure of social dominance orientation- specifically, SDO-Dominance (SDO-D) and SDO-Egalitarianism (SDO-E). SDO-D assessed individuals' preference for social systems in which higher status groups forcefully oppress lower status groups. SDO-E assessed individuals' preference for social systems that maintain inequality through subtle ideologies and social policies that favor higher status groups and disfavor lower status groups. A two-factor model of SDO-D and SDO-E demonstrated good fit to the data for men and women ($\chi^2(36) = 268.77$, CFI = .91, RMSEA = .08, SRMR = .05).

Competitiveness. Competitiveness was measured using Newby & Klein's (2014) measure of competitive affectivity. According to Newby & Klein (2014), trait competitiveness can be split into four factors (i.e., general, dominant, affective, and personal enhancement). Competitive affectivity refers to the extent to which individuals derive feelings of superiority and powerfulness as a result of competition. Individuals who scored high on this dimension enjoy competition and feel personally affected by losing competitions. This dimension of competitiveness was theorized to be most related to aggression towards others in the Academies.

Good fit to the data was observed for this measure between men and women ($\chi^2(10) = 137.98$, CFI = .93, RMSEA = .11, SRMR = .06).

Conformity to masculine norms. Conformity to masculine norms was measured using Hsu & Iwamoto's (2014) eight factor measure of conformity to masculinity norms. Hsu & Iwamoto's (2014) measure is a refinement of Mahalik et al.'s (2003) original masculinity scale which measured: conformity to (a) violence norms; (b) promiscuity norms; (c) heterosexual presentation norms; (d) risk-taking norms; (e) emotional control norms; (f) power over women norms; (g) winning norms; and (h) self-reliance norms. The dimensions of (a) violence, (b) promiscuity, (c) heterosexual presentation, (d) risk-taking and (e) emotional control were selected due to their relevance to the cadet/military population. Conformity to masculine norms is an important predictor of individual behavior and well-being (Mahalik et al., 2003; Marrs, 2013). Moreover, it was appropriate and important to examine in the Military due to the historically male context of the Military and the intersectionality between masculinity narratives and the military role (Arkin & Dobrofsky, 1976; Hinojosa, 2010). A five-factor model of the selected masculinity norms demonstrated excellent fit to the data for men and women ($\chi^2(102) = 227.56$, CFI = .99, RMSEA = .03, SRMR = .03).

Formal Leadership Experience. Cadets/midshipmen reported whether they had experience holding a formal leadership position at their respective institutions. Roughly 36.5% of cadets/midshipmen reported having formal experience leading others at their institutions.

Criterion Variable: Counterproductive Work Behavior. Counterproductive work-behavior (CWB) was measured using self-reported instances of such behavior. Specifically, eleven items from Spector et al.'s (2006) CWB measure were selected due to their perceived relevance to the Military cadet population. Examples of CWB behaviors include acts targeted

against the organization (CWB-O) such as "purposely wasting materials/supplies" and "skipping required events"; and towards people (CWB-P) such as "making fun of someone's personal life" and "starting an argument with someone." A two-factor model depicting CWB-O's and CWB-P's as latent factors demonstrated acceptable but below average fit for men and women ($\chi^2(4) = 37.12$, CFI = .83, RMSEA = .09, SRMR = .07). The two-factor model was referenced by the majority of the CWB literature (Dalal, 2005).

Results

Validity Results

Table 1 provided the validity coefficients of the seven estimation methods for predicting CWBs. The sample size of the data used to statistically test these coefficients was 2023, and the sample included both male and female cadets/midshipmen as well as the four datasets generated by the oversampling manipulation. As expected, all of the estimation methods resulted in statistically significant prediction equations. As shown in Table 1, six of the seven estimation methods produced validity coefficients in the 0.40 range: (i.e., regression; random forests; extreme gradient boosting; ridge regularization; lasso regularization; elastic net and one (i.e., the decision tree method) produced a validity coefficient in the 0.30 range (i.e., $r(2020) = 0.34$, $p < 0.01$).

To compare the validity coefficient for the regression method to the validities of the different ML estimation methods, I used Steiger's Z-Test for comparing dependent correlations. As shown in the second row of Table 1, the regression method had significantly greater validity than that of the decision tree ($r = 0.34$), $z_H = 7.99$, $p < 0.01$ and random forests ($r = 0.41$), $z_H = 4.76$, $p < 0.01$ methods. No significant differences in validity were observed between regression and the other ML methods.

In summary, I did not find any of the ML estimation methods to have validity greater than the regression method. However, two of the ML methods (i.e., decision tree and random forest) exhibited significantly lower validity than the regression method.

Fairness Results

In the next set of analyses, I examined the adverse impact ratios for each estimation method. To identify the factors that affected the adverse impact of the different estimation methods, I conducted a repeated measures factorial ANOVA to evaluate the main and interactive effects of estimation method, oversampling manipulation, and gender on predicted scores. Estimation method was examined as a within-group variable in this analysis given that all seven methods were conducted using the same data, while oversampling and gender were examined as between-group variables. Table 2 presents the results of the repeated measures factorial ANOVA.

As shown in Table 2, both estimation method $F(6, 12090) = 21.48$, $p < 0.01$, $\eta^2 = 0.01$ and gender $F(1, 2015) = 81.51$, $p < 0.01$, $\eta^2 = 0.04$) had significant main effects on the predicted scores. No main effect of oversampling was found.

I conducted pairwise comparisons using the Bonferroni test to determine the source of the significant estimation method and gender main effects. With regard to the estimation method main effect, the average scores produced by regression (M = 1.84, SD = 0.01) was significantly lower than the average scores produced by the decision tree (M = 1.85, SD = 0.00) and random forests (M = 1.86, SD = 0.01) algorithms. However, the average regression scores were higher than the average scores produced by the extreme gradient boosting (M = 1.84, SD = 0.01). With regard to the gender main effect, the average score for men (M = 1.89, SD = 0.01) were significantly higher than the average score for women (M = 1.80, SD = 0.01).

As shown in Table 3, there was a significant interaction between estimation method and gender on predicted scores, $F(6, 12090) = 17.83$, $p < 0.01$, $\eta^2 = 0.01$. This significant interaction indicates that the mean difference between males and females differ as a function of the estimation method used. No additional interaction effects were found. Table 4 presents the predicted scores and Cohen's d-statistic for men and women across each of the seven estimation methods.

The d-statistics were in the 0.40 range for the regression (d = 0.44, 95% CI [0.34, 0.53]), extreme gradient boosting (d = 0.41, 95% CI [0.32, 0.50]), ridge regularization (d = 0.44, 95% CI [0.34, 0.53]), lasso regularization (d = 0.44, 95% CI [0.34, 0.53]), and elastic net regularization (d = 0.44, 95% CI [0.34, 0.53] estimation methods. Effect sizes in the 0.30 range were observed for the decision tree (d = 0.30, 95% CI [0.21, 0.40]) and random forests method (d = 0.36, 95% CI [0.26, 0.45]). The positive d-statistics reveal that predicted CWB scores were higher for male cadets/midshipmen.

The d-statistic for the decision-tree was significantly lower than that of regression, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization as indicated by its' confidence interval (95% CI [0.21, 0.40]). No differences in the d-statistic were observed between decision-tree and random forests. Since d-statistics closer to zero are evidence of less adverse impact than d-statistics greater than zero, the decision tree showed less adverse impact than the other estimation methods.

The Cleary Model of Test Bias was then used to determine whether each estimation method demonstrated gender test bias in its prediction of CWBs. The results of the three regression models used to detect slope and/or intercept bias for each method were shown in table 5. Gender test bias was observed across all seven of the methods used to predict CWBs. Slope

bias[15] was detected for the regression, random forests, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization estimations. Intercept bias[16] was detected for decision tree and random forests estimations.

The change in R^2 analyses for the three steps of the Cleary model revealed significant improvements in predicting actual CWBs via the addition of gender (Cleary Model 2) and then the interaction between gender and predicted CWBs (Cleary Model 3). Specifically, Cleary Model 2 enhanced the predictive ability for the decision tree ($\Delta R^2 = 0.005$, $F(1, 2020) = 11.68$), and random forests estimations ($\Delta R^2 = 0.003$, $F(1, 2020) = 6.48$). Cleary Model 3 enhanced the predictive ability for regression ($\Delta R^2 = 0.005$, $F(1, 2019) = 11.79$), random forests ($\Delta R^2 = 0.002$, $F(1, 2019) = 5.93$), extreme gradient boosting ($\Delta R^2 = 0.004$, $F(1, 2019) = 9.42$), ridge regularization ($\Delta R^2 = 0.005$, $F(1, 2019) = 11.72$), lasso regularization ($\Delta R^2 = 0.005$, $F(1, 2019) = 11.67$), and elastic net regularization estimations ($\Delta R^2 = 0.005$, $F(1, 2019) = 11.58$). Accordingly, all seven estimation methods demonstrated predictive bias for gender, making them inappropriate for use in a selection context.

Discussion

The present study focused on the impact that oversampling had on estimation method validity and fairness. Specifically, I examined whether the distribution of male and female cadets/midshipmen from different Military Academies in the training/validation dataset affected the estimation methods. Though I expected that the oversampling manipulation would impact the performance of the seven estimation methods, I found that neither validity nor fairness were affected. Instead, I found only the estimation method and gender to impact validity and fairness.

[15] A significant interaction effect of gender and predicted scores in the third Cleary model indicates slope bias.

[16] A significant main effect of gender in the second Cleary model indicates intercept bias.

Moreover, I found an interaction between estimation method and gender such that the difference in predicted scores between men and women differed for some estimation methods.

Regarding validity, I found four of the six machine learning estimation methods to produce validity coefficients that were equivalent to regression: e.g., extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization. The decision tree and random forests estimation methods produced significantly lower validity coefficients than regression. Thus, the decision tree and random forests estimations demonstrated less validity evidence than the other estimation methods. As such, neither the decision tree nor random forests estimations would be legally defensible for use over regression in selection, assuming that each estimation method was also deemed fair by the courts (Tippins et al., 2018).

On the note of fairness, all estimation methods demonstrated gender predictive bias. In particular, slope bias was detected for six of the seven estimation methods (e.g., regression, random forests, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization) and intercept bias was detected for two estimation methods (e.g., decision tree and random forests). Thus, all estimation methods would be inappropriate for selection as they exhibited the psychometric form of unfairness which the courts (*Cormier v. P.P.G Indus.*, 1983; *Hamer v. City of Atlanta*, 1989; *United States v. City of Erie*, 2005) and the scientific community (Aguinis et al., 2010; Tippins et al., 2018) typically view the presence of intercept and/or slope bias as an unacceptable condition in selection contexts.

Regarding the outcome-focused perspective of fairness, the adverse impact effect sizes between men and women for five of the six machine learning estimation methods (e.g., random forests, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization) were equivalent to regression. However, the observed adverse impact effect size

was nearly 0.1 units less than regression for the random forest's estimation method, and the decision tree estimation method produced a significantly lower adverse impact effect size than regression at nearly 0.15 units less. The lower adverse impact effect sizes exhibited by both methods provides significant practical implications, given the shifting burden of proof model.

According to the shifting burden of proof model, AI is defensible only when the employer can demonstrate that an equivalently valid and less discriminatory alternative selection method could not have reasonably existed and/or been used at the time the selection method in question was used (Hanges et al., 2013). Given that the validity for the decision tree and random forests estimations was significantly lower than the other estimation methods, the shifting burden of proof model would not be applicable here as the employer could state that the other methods were not equivalent. Indeed, the use of a less valid selection method for the intent of increasing fairness is explicitly outlawed by the courts, unless sufficient business justification can be made regarding how greater representation of underrepresented protected groups enhances the organizations' interests and ability to perform its duties successfully (*Detroit Police Officers association v. Young*, 1971; *Petit v. Chicago*, 2003; Riccucci & Saldivar, 2012). The shifting burden of proof model is a critical factor when deciding which selection method is most legally defensible and appropriate to implement.

Returning to the null effect of the oversampling manipulation on validity and fairness, several takeaways can still be taken from an applied perspective. In terms of legal defensibility, the mere attempt by the practitioner to use a fair representation of protected groups when developing the selection method is positive, because it fosters greater perceptions of fairness regarding the validity of the selection method. Again, perceived fairness is an important component for practitioners to consider as it is in their interest to produce selection methods that

are not prone to potential criticism and/or legal challenges from others in the first place. Indeed, HireVue was recently criticized for "perpetuating discriminatory hiring practices" via its use of facial recognition and analysis data- of which likely demonstrated criterion and some content validity but appeared biased to others (Barrett, 2021; Harwell, 2019; 2019). An explicit statement and/or brief explanation of how protected characteristics were specifically accounted for in the development of their selection method could potentially assuage these claims. In addition, the consideration of the distribution of protected characteristics in the training/validation datasets was consistent with current legal rulings on fairness. Again, this practice did not reward or penalize protected characteristics, nor did it reference them in a manner that produced different standards for members of different protected groups (e.g., race-correcting, race-norming). Rather, it simply sought adequate representation for protected groups in an attempt to identify a more valid selection method that was generalizable for people from different protected groups.

In terms of statistical explanations, it was possible that the effect of the oversampling manipulation was not strong enough. For instance, the distribution of male and female cadets/midshipmen from different Military Academies could be exaggerated such that an extremely low number of women (e.g., far beyond random chance) was included in the training dataset and/or only one Military Academy location comprised the training dataset. In this case, the estimation methods would be disproportionately developed with male training data and/or one Academy location, and CWB estimations could be less generalizable to women in the validation dataset and/or the other two Military Academies. Further, the total sample was predominately male and from one Military Academy, possibly indicating that the overall validity of the estimation in the validation dataset would remain high albeit they would not generalize well for women and the other two Military Academies. Adverse impact effect sizes between men

and women could be amplified in this situation, resulting in a significant effect of the oversampling manipulation. Though this condition could have strengthened my manipulation, I did not include it because it is inconsistent with current machine learning practice which often consists of a random distribution of protected groups in the training/validation datasets. Further, the practice of segregating protected groups from one another in the training/validation datasets would harm the legal defensibility of the estimation methods and expose them to potential critiques of bias for *purposely* relying upon one protected group to develop the estimations.

Lastly, all of the results of Study One could be severely impacted by range restriction of the sample. Consider the construct of CWBs, and how antithetical it is the culture of the Military Academies[17]. It is very possible that cadets/midshipmen at the academies engage in far less CWBs than enlisted military personnel, University students, or less educated young adult populations. As such, the validity coefficients and adverse effect sizes observed in Study One could likely depict decreased estimates. Future research could examine how to best correct for range restriction: e.g., correcting for range restriction during the training/validation process for each estimation method, and/or correcting the estimates that are produced from the process. In addition, future research could focus on the sampling strategy used to develop the estimation methods, as opposed to correcting for range restriction within the algorithm itself. A broader sampling strategy could be applied and include comparable populations that likely engage in different levels of CWB, such as enlisted military personnel and University ROTC students. A larger range of CWBs is an important factor to consider when developing selection methods to increase the effectiveness of the estimation methods. The key here is to focus on ensuring that

[17] The Military service academies are renowned for emphasizing personal ethics and honor. Moreover, admission into the service academies is highly competitive and consists of strict academic, physical and leadership requirements.

the inclusion of comparable populations to the specific population of interest is logical and legally defensible. In other words, the goal of selection is to develop a fair and valid selection method that is legally justified for use.

In Study Two, I focused on a different aspect of the selection method development process which consisted of the content validity of predictors used by the estimation methods. The content validity of predictors is critical for the legal defensibility of a selection method (Gutman et al., 2017). Indeed, the courts have ruled that criterion validity evidence alone is inadequate for establishing legal defensibility (*Gillespie v. State of Wisconsin*, 1985; *Police Officers v. City of Columbus*, 1990). Rather, the more sources of validity evidence a selection method possesses (e.g., criterion, content, face, etc.), the more evidence exists to support the claim that the selection method is legally defensible (*Guardians v. Civil Service*, 1980; *Gulino v. New York State Education Department*, 2012; *Smith v. City of Boston*, 2015). Given the sole focus on prediction (e.g., criterion validity) in machine learning practice[18], it is important that other forms of validity (e.g., content validity) are also considered when examining the potential of utility of machine learning estimation methods in the selection context specifically. Again, the use of an indefensible selection method can result in significant legal costs for an organization. In Study Two, I examined how the content validity of the predictors used by the machine learning estimation methods impacted their validity and fairness.

[18] It is common practice for ML practitioners to discount the content and/or construct validity of predictors they use to develop their algorithms. In fact, predictors are often regarded solely as additional sources of data available to the algorithm to parse through and reference in its estimation of the criterion of interest.

Chapter Seven: Study Two

In Study Two, I examined the impact that the content validity of predictors used by the seven estimation methods have on validity and fairness. In particular, I manipulated whether the methods were restricted to content valid predictors that were also supported by prior empirical evidence and/or whether they included content valid predictors that were not previously validated and/or predictors that were explicitly illegal in the context of selection (e.g., gender, religiosity). The primary purpose of this study was to examine how different criteria for the content validity of predictors affected the validity and fairness of the seven estimation methods in predicting Military CWBs.

Method

Participants and Procedure.

Study Two used the same sample of 2,026 military cadets/midshipmen that was used in Study One. Formal leadership experience and the five "Other Characteristic" psychological variables used in Study One were again referenced: e.g., prosocial personality, entitlement, social dominance, competitiveness, conformity to masculine norms. However, up to seven additional variables were included for use by the estimation methods depending on the condition of the content validity manipulation. The method by which the training/validation dataset was split was consistent across the manipulation and consisted of splitting the data whereby both the proportion of men and women and proportion of cadets/midshipmen from each academy in the total dataset were equal to that in the total dataset (e.g., condition four of Study One)[19].

[19] Though Study One did not reveal an impact of the training/validation dataset creation manipulation, I decided to create our training/validation dataset using the fourth condition of my prior manipulation which accounted for the gender and location of cadets/midshipmen. I made this decision because I believe that gender and location are theoretically meaningful categories that should still be considered when developing valid and fair predictions of cadets/midshipmen's CWBs. Moreover, I believe that considering these categories will increase the legal defensibility of the estimation methods.

Content Validity Manipulation.

The content validity manipulation consisted of three conditions wherein the predictors used by the estimation methods varied in terms of their content validity. Content validity was established by a team of subject matter experts. In particular, myself and three other graduate students with prior experience conducting research on unethical behavior at the Military Academies examined the individual-level variables used in Hanges et al.'s (2021) original study. We drew upon our experience directly working with cadet/midshipmen interviews, as well as our knowledge of the I/O psychology literature to select variables that were relevant to cadets/midshipmen's engagement in CWBs. In the first condition, only content valid predictors supported by empirical evidence were used. These predictors consisted of the five "Other Characteristic" psychological variables that demonstrated relationships with CWBs per the academic literature, as well as two legal biodata predictors theorized to strongly influence CWBs: e.g., formal leadership experience, and the Military Academy that the cadet/midshipman attended.

In the second condition, all of the content valid predictors supported by empirical evidence were used alongside three content valid predictors that did not exhibit prior empirical support. These predictors included two affective psychological variables that were developed specifically for predicting CWBs at the Military Academies: e.g., loyalty to fellow cadets/midshipmen and individual identification with the military. The third predictor consisted of intra-individual response variation (IRV), an index designed to detect insufficient effort responding (Dunn et al., 2018). IRV scores are related to individual differences in conscientiousness, agreeableness, and boredom proneness (Dunn et al., 2018).

In the third condition, explicitly illegal and/or controversial predictors were added alongside the content valid predictors with and without prior empirical support. The explicitly illegal and/or controversial predictors included two protected class characteristics (e.g., gender of cadet/midshipmen, religiosity) and three controversial predictors which are strongly associated with demographics (e.g., political orientation, socioeconomic status, parent's education attainment). The validity and fairness of the seven estimation methods were evaluated across the content validity manipulation.

Variables.

Academy Location. The location of cadets/midshipmen were provided by the respective institutions. Roughly, 13.5% of cadets/midshipmen were from the Army, 12.8% from the Navy, and 73.7% were from the Air Force.

Loyalty to Fellow Cadets/Midshipmen. Hanges et al. (2021) created a six-item measure of loyalty to fellow cadets/midshipmen. This measure was based upon information collected from cadet/midshipmen interviews as well as discussions with senior officers at the Academies. Sample items included: "I would sacrifice the mission of the academy if doing so protected a cadet/midshipman of my squadron" and "my first loyalty is to the cadets in my squadron." Given that this measure was not previously validated as a single factor measure, it was examined at the item-level as opposed to the variable-level.

Identification with the Military. Hanges et al. (2021) also created a six-item measure of identification with the Military. As with the loyalty to fellow cadets/midshipmen measure, these items were created based upon discussions with senior officers at the Academies, as well as the results of the interviews. Identification with the Military refers to the extent to which a cadet/midshipman personally identifies as a member of the military academy and branch. Sample

items of this measure include, “I identify strongly with the Academy” and “I am proud to be a member of the United States Military/Navy/Air Force.” This measure was examined at the item-level as opposed to the variable-level given that it was not previously validated as a single factor measure.

Intra-individual Response Variation. IRV is a statistical index designed to detect insufficient effort responding, i.e., reduced effort by participants when answering questionnaires often due to inattentiveness, fatigue, or speediness (Hong et al., 2019). IRV is calculated by taking the standard deviation of responses across a set of consecutive item responses for an individual (Dunn et al., 2018). IRV scores are related to individual differences in conscientiousness, agreeableness, and boredom proneness (Dunn et al., 2018).

Gender. The gender of all cadets/midshipmen was provided by the respective institutions. Gender reflected the assigned gender at birth for cadets/midshipmen. The majority of cadets/midship were male (64.3%).

Religiosity. Religiosity consisted of religious attendance and religious feelings. In terms attendance, 19.3% self-reported never attending religious services, 28.9% as seldom attending religious services, 17.7% as sometimes attending religious services, 24.5% as frequently attending religious services, and 9.7% as faithfully attending religious services. Regarding religious feelings, 19.9% preferred living their life without any particular faith/religion, 41.5% believed a person could be saved through faith/religion and 38.6% believed that their faith is the only true faith/religion that leads to salvation.

Political Orientation. Political orientation was comprised of two components: beliefs towards social issues and beliefs towards economic issues. Regarding social issues, 4.7% of cadets/midshipmen self-identified as very liberal, 21.8% self-identified as liberal, 34.4% self-

identified as moderate, 29.4% self-identified as conservative, and 9.8% self-identified as very conservative. For economic issues, 1.3% of cadets/midshipmen self-identified as very liberal, 7.6% self-identified as liberal, 33.6% self-identified as moderate, 43.7% self-identified as conservative, and 13.7% self-identified as very conservative.

Socioeconomic Status. Socioeconomic status (SES) was self-reported by cadets/midshipmen. Approximately 3.3% of the cadets/midshipmen self-identified as coming from lower class SES background, 11.8% self-identified as from a working-class SES background, 39.7% self-identified as from a middle-class SES background, 40.5% self-identified as from an upper-middle class SES background, and 4.7% self-identified as from an upper-class SES background.

Parent's Education Attainment. Parent's education attainment was also self-reported by cadets/midshipmen. Regarding mother's education attainment, 1.70% of cadets/midshipmen self-reported less than high school, 14.8% self-reported high school diploma, 12.5% self-reported associate degree, 42.8% self-reported bachelor degree, and 28.2% reported graduate degree or above. For father's education attainment, 2.2% of cadets/midshipmen self-reported less than high school, 15.9% high school diploma, 8.9% associate degree, 34.5% bachelor degree, and 38.5% graduate degree or above.

Results

Validity Results

Table 6 shows the validity coefficients of the seven estimation methods used to predict CWBs across gender and the three levels of the content validity manipulation. The sample size of the data used to statistically test these coefficients was 1510. The sample was comprised of both male and female data and the three content validity manipulations, which were created by using

only the stratified by gender and academy training/validation dataset method. As expected, all of the estimation methods resulted in statistically significant prediction equations. Six of the seven validity coefficients were in the 0.50 range: e.g., regression, $r(1507) = 0.53$, $p < 0.01$; random forests, $r(1507) = 0.51$, $p < 0.01$; extreme gradient boosting , $r(1507) = 0.53$, $p < 0.01$; ridge regularization, $r(1507) = 0.53$, $p < 0.01$; lasso regularization, $r(1507) = 0.53$, $p < 0.01$; and elastic net , $r(1507) = 0.53$, $p < 0.01$. The decision tree method had a validity coefficient of $r(1507) = 0.34$, $p < 0.01$. Steiger's Z-Test revealed that regression ($r = 0.53$) was significantly greater than that of the decision tree ($r = 0.36$), $z_H = -8.94$, $p < 0.01$. No significant differences in validity were observed between regression and the other methods.

Evidently, I again found that the validity of regression was equivalent to that produced by the ML estimation methods. However, only the decision tree estimation method exhibited significantly lower validity than the regression method in this Study.

Fairness Results

The repeated measures factorial ANOVA was used to assess the main and interactive effects of estimation method, content validity manipulation, and gender on predicted scores. Table 7 shows the main effects and Table 8 shows the interaction effects of the repeated measures factorial ANOVA while Table 7. Similar to Study One, estimation method was examined as a within-group variable whereas the content validity manipulation and gender were examined as between-group variables.

Main effects were found for method, $F(6, 9036) = 3.68$, $p < 0.01$, $\eta^2 = 0.00$ and gender, $F(1, 1506) = 40.89$, $p < 0.01$, $\eta^2 = 0.03$, on predicted scores. No main effect of the content validity manipulation was found. The Bonferroni test was used to assess pairwise comparisons on the method and gender factors. Relative to the average regression score (M = 1.86, SD =

0.00), the average scores were significantly higher for random forests (M = 1.87, SD = 0.00) and lower for extreme gradient boosting (M = 1.85, SD = 0.00) and ridge regularization (M = 1.86, SD = 0.00) algorithms. Regarding gender, the average scores were significantly higher for men (M = 1.89, SD = 0.00) than women (M = 1.81, SD = 0.00).

An interaction effect between estimation method and gender was found, $F(6, 9036) = 5.33$, $p < 0.01$, $\eta^2 = 0.00$. This means that gender mean differences on the predictor scores varied as a function of estimation method. No additional interaction effects were detected. With regard to the obtained estimation method by gender interaction, Table 9 presents the predicted scores and Cohen's d-statistic for men and women across each of the seven estimation methods.

Effect sizes of mean differences for men and women were in the 0.30 range for regression (d = 0.34, 95% CI [0.23, 0.45]), random forests (d = 0.33, 95% CI [0.22, 0.43]), extreme gradient boosting (d = 0.33, 95% CI [0.22, 0.44]), ridge regularization (d = 0.36, 95% CI [0.25, 0.47]), lasso regularization (d = 0.35, 95% CI [0.24, 0.46]), and elastic net regularization (d = 0.35, 95% CI [0.24, 0.46]). The d-statistic for the decision tree estimation method was in the 0.20 range (d = 0.27, 95% CI [0.16, 0.38]). No significant differences in effect sizes however were observed between any two methods. Thus, the decision tree estimation method showed smaller mean differences between males and females than any of the other estimation methods.

Predictive bias for the estimation methods was evaluated using the Cleary Model of Test bias. The results for each estimation method are shown in Table 10. Slope and/or intercept bias was observed for six of the seven estimation methods: e.g., regression, decision tree, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization. Specifically, slope bias was detected for regression, extreme gradient boosting, ridge

regularization, lasso regularization and elastic net regularization estimations. Intercept bias was detected for the decision tree estimation. Only the random forests estimation generated predictions that did not demonstrate slope nor intercept bias as evidenced by only the predicted CWB scores significantly predicting actual CWB scores in Cleary model 2 ($\beta = 0.51$, $t(1509) = 22.70$, $p < .01$) and Cleary model 3 ($\beta = 0.60$, $t(1508) = 8.80$, $p < .01$).

Additionally, incremental validity analyses did not reveal significant improvements in predicting actual CWBs via the addition of gender (Cleary Model 2) and then the interaction between gender and predicted CWBs (Cleary Model 3) for the random forests estimation. Indeed, Cleary Model 2 enhanced the predictive ability for the decision tree estimation ($\Delta R^2 = 0.004$, $F(1, 1509) = 7.09$). Cleary Model 3 enhanced predictive ability for regression ($\Delta R^2 = 0.004$, $F(1, 1508) = 8.87$), decision tree ($\Delta R^2 = 0.003$, $F(1, 1508) = 5.05$), extreme gradient boosting ($\Delta R^2 = 0.004$, $F(1, 1508) = 7.45$), ridge regularization ($\Delta R^2 = 0.004$, $F(1, 1508) = 8.84$), lasso regularization ($\Delta R^2 = 0.004$, $F(1, 1508) = 8.10$) and elastic net regularization estimations ($\Delta R^2 = 0.004$, $F(1, 1508) = 8.17$).

Therefore, six of the seven estimation methods- regression, decision tree, extreme gradient boosting, ridge regularization, lasso regularization and elastic net regularization - demonstrated predictive bias for men and women. Only the random forests estimation did not show predictive bias, making it appropriate for use in a selection context.

Discussion

Study Two examined how the content validity of predictors used by the estimation methods affected their validity and fairness. In particular, I examined how different criteria for the content validity of predictors, e.g., whether they were supported by prior empirical evidence and/or included illegal predictors, affected the performance of the estimation methods.

Interestingly, I found that neither estimation validity nor fairness were affected the content validity manipulation. As such, I reached a similar result to Study One where only the estimation method, gender, and the interaction between estimation method and gender affected the validity and fairness of seven estimation methods.

In terms of validity, I found five of the six machine learning methods to produce validity coefficients equivalent to regression: e.g., random forests, extreme gradient boosting, ridge regularization, lasso regularization, and elastic net regularization. Only the decision tree estimation method produced a significantly lower validity coefficient than regression, making it an inappropriate estimation method to use for selection over regression as well as the other five machine learning methods.

Regarding fairness, regression and five of the six machine learning methods demonstrated gender predictive bias. Specifically, slope bias was present for six estimation methods: e.g., regression, decision tree, extreme gradient boosting, ridge regularization, lasso regularization, and elastic net regularization. Intercept bias was detected for the decision tree estimation method. However, the random forests estimation method demonstrated neither slope nor intercept bias. Thus, the results of Study Two suggested that the random forests estimation method would be the only legally defensible estimation method for selection purposes given that it demonstrated validity and no predictive bias (*Cormier v. P.P.G Indus*., 1983; *Hamer v. City of Atlanta*, 1989; *United States v. City of Erie*, 2005).

Adverse impact effect sizes for regression, random forests and four other machine learning estimations (e.g., extreme gradient boosting, ridge regularization, lasso regularization, elastic net regularization) were all equivalent and in the 0.30 range. The adverse impact effect size for the decision tree estimation method was in the 0.20 range, however, it was not

significantly less than that of the other estimation methods. Thus, the shifting burden of proof model would not be applicable for the decision tree estimation method for two reasons: the adverse impact effect size was not significantly lower than that of the other estimation methods, and the validity was significantly worse than the other estimation methods (Hanges et al., 2013).

Combined, the results of Study Two revealed that the random forests estimation method would be the most appropriate estimation method to use in selection. Specifically, the random forests method demonstrated equivalent validity to regression and the other estimation methods; no predictive bias; and an adverse impact effect size that was not greater than any equivalently valid estimation method without predictive bias. From a practitioner's perspective, one of the key features of a selection method is its legal defensibility. As the courts have ruled, adverse impact evidence falls secondary to validity and predictive bias evidence when establishing legal defensibility (Aguinis et al., 2010; Tippins et al., 2018). Moreover, adverse impact is only considered indefensible by the courts when an equivalently valid alternative method existed at the time in which the selection method was used and showed lower adverse impact (Hanges et al., 2013).

The practice of reducing adverse impact at the cost to validity is explicitly outlawed by the courts. In particular, this is because doing so is considered a violation of *Title VII of the Civil Rights Act* as it uses applicants protected class information to inform the selection decision[20] (*Ricci v. DeStefano*, 1990). Selection methods can only be modified if the modification is made in a protected-class neutral fashion (*Hayden v. Nassau County*, 1996). Indeed, protected class information can only be explicitly considered in selection when sufficient business justification

[20] In *Ricci v. DeStefano* (1990), the court found that the practitioner's decision to discard some of the selection method results due to high adverse impact was inappropriate because the practitioner did not have strong evidence that the test was invalid and/or technically deficient.

can be made that protected characteristics are essential for performing the job. For instance, when greater representation of certain protected groups enhances the organizations' interests and ability to perform its duties successfully (*Detroit Police Officers association v. Young*, 1971; *Petit v. Chicago*, 2003; Riccucci & Saldivar, 2012) and/or when a particular job requires certain essential protected characteristics to perform the job adequately (e.g., bona fide occupational qualifications[21]).

Though the random forests estimation was found to be a more appropriate estimation method than regression and the other machine learning estimations, it should be noted that this finding occurred when the test/training datasets were distributed in a manner that considered both the gender and location of cadets/midshipmen and the content validity of predictors was then manipulated. As such, it is possible that these results would not generalize to situations wherein the test/training datasets were distributed randomly and/or in a manner that did not consider cadets/midshipmen's gender and location. Moreover, if the content validity manipulation was altered to include a condition whereby only non-content valid predictors was used, then it is possible that the observed result would not be reached. However, the addition of this possible condition would greatly harm the legal defensibility of the estimation methods and would be an unlikely situation in the practice of selection[22]. Nonetheless, the null finding of the

[21] A bona fide occupational qualification (BFOQ) is a very narrowly interpreted exception to Equal Employment Opportunity (EEO) laws. A BFOQ allows employers to base employment decisions for a particular job on such factors as sex, religion, or national origin if they are able to demonstrate that such factors are an essential qualification for performing a particular job. For instance, requiring male prison enforcement in all male maximum security prisons for safety reasons (*Dothard v. Rawlinson*, 1977) and requiring female nurses in obstetrical care for patient privacy and quality of care reasons (*Slivka v. Camden-Clark Memorial Hospital*, 2004).

[22] In selection, it is very uncommon that all of the predictors used by the estimation methods contain zero content validity. Again, the courts have ruled that criterion validity evidence alone is inadequate for establishing legal defensibility (*Gillespie v. State of Wisconsin*, 1985; *Police Officers v. City of Columbus*, 1990). More sources of validity evidence are conducive to support the claim that the selection method is legally defensible (*Guardians v. Civil Service*, 1980; *Gulino v. New York State Education Department*, 2012; *Smith v. City of Boston*, 2015).

content validity manipulation in the present study suggests that the use of content valid predictors supported by empirical evidence should be used in selection due to its equivalent validity and fairness, and enhanced legal defensibility.

Future research should focus on the specifics of how each engineering decision made during the estimation method development process impacts the subsequent validity and fairness of the estimation methods. Computational modeling approaches could be advantageous in this regard as they could examine the conditional effects of various engineering decisions. For instance, Monte Carlo simulations could be conducted whereby the relationships between the predictors and the criterion, as well as their relationships with different protected groups are manipulated. Such research could produce more generalizable results and identify the specific circumstances where the validity and fairness of machine learning estimation methods are greater than that of regression. Nonetheless, the primary aim of my dissertation was to examine whether machine learning estimations could *potentially* present an improvement in estimation validity or fairness, and the results of the random forests estimation method in Study Two supported the primary aim of my dissertation.

In Study Three, I built upon the results of Study One and Study Two by applying both practices into a human resources dataset. In particular, I applied oversampling and content valid predictors to develop legally defensible selection tools for employees present Patalano & Huebner's (2021) sample. Patalano & Huebner (2021) samples were made available to instructors and researchers to teach HR analytics. The criterion variable consisted of employee job performance. Fairness was assessed between male and female employees, and racial majority and minority employees. No manipulation affecting different configurations of the test/training datasets or predictors was present in either study. Rather, the legal defensibility of the estimation

methods was instead emphasized, and greater focus was placed on the interpretability of the machine learning methods. In the next section, I present the three methods I incorporated for machine learning interpretability, as well as the design for Study Three.

Chapter Eight: Study Three

In Study Three, I focused on how the estimation methods alone impact validity and fairness when oversampling was applied, and content valid predictors were used. Moreover, I incorporated three methods designed for machine learning interpretability: e.g., variable importance, partial dependence plots, and individual conditional expectation plots. These three methods revealed how the predictors used by the ML methods specifically relate to predicting employee job performance ratings. Patalano & Huebner's (2021) HR Metrics and Analytics dataset was used in Study Three, and fairness was evaluated for both gender and race as both protected information was provided. Study Three had three purposes: first to extend the present research to an actual selection focused dataset; second to demonstrate the usefulness of ML interpretability methods for the legal defensibility of the estimation methods; and third to again evaluate the validity and fairness of the estimation methods in predicting employee performance ratings.

Method

Participants and Procedure.

Study Three used the sample of 311 employees present in Patalano & Huebner's (2021) HR Metrics and Analytics dataset. This dataset is publicly available on Kaggle for instructors and researchers to use and was developed to simulate HR data and teach graduate-level HR professionals. It contains data on a variety of variables including protected characteristics, position and salary information, performance ratings, behavioral data, and attitudinal variables.

In terms of the variables used by the estimation methods, I selected variables that exhibited content validity in predicting job performance. In particular, myself and four subject matter experts examined the variables available in the dataset and identified five variables which

we deemed and agreed upon as being relevant to predicting job performance for employees. The three behavioral variables chosen to use in this study consisted of the number of special projects that the employee worked on during the last six months, the number of times that the employee was late to work during the last 30 days, and the number of times the employee was absent from work. Two attitudinal variables were selected and included employee engagement and employee satisfaction. The three behavioral variables can be classified as key components of successful performance of the job. Employee engagement and employee satisfaction can be regarded as belonging to the "Other" category of the KSAO framework.

Oversampling was applied in Study 3 and consisted of accounting for employee gender and race. In particular, the training/validation dataset was split whereby both the proportion of men and women, and proportion of racial minorities and majorities in the total dataset were equal to that in the total dataset.

Machine Learning Interpretability Methods.

Three machine learning interpretability methods were incorporated in Study Three: e.g., variable importance, partial dependence (PD) plots, and individual conditional effect (ICE) plots. These methods are capable of unpacking the black box often associated with ML methods (Molnar, 2021). I applied the three interpretability methods only to decision tree-based ensemble methods, since there exist better alternatives for interpreting decision-tree and regularization-based methods. For instance, the decision-tree method can be visually depicted as it consists of a single decision-tree model as opposed to an ensemble of models. Regularization-based methods can be interpreted similarly to regression as they also produce regression coefficients that reflect the directionality and magnitude of each predictor and the criterion variable.

Conversely, decision tree-based ensemble methods, such as random forests and extreme gradient boosting methods, cannot be visually depicted into a single tree diagram nor produce regression coefficients that reflect the directionality and magnitude of each predictor and the criterion variable. Methods designed to unbox these ML methods are needed to identify the relationship between each predictor and criterion. Variable importance, partial dependence plots (PDP), and individual conditional effect plots (ICE) are three methods that confer this information for decision tree-based ensemble methods, and I incorporated them in Study Three.

Variable Importance. Variable importance analysis depicts the relative impact of each predictor variable in predicting the criterion variable within a decision tree-based method (Foster et al., 2016). Conceptually, variable importance analysis works by calculating the extent to which the removal of each variable decreases the accuracy of the method in predicting the criterion. Figure 3 presents an example of a variable importance plot where the relative contribution of each predictor in the prediction of the criterion was shown. In Figure 3, the two most important predictors were var1 and var2, and a substantial drop in importance was observed for v5 which was ranked as the third most important variable. Variable importance scores can be treated similarly to an effect-size measure, and are critical for interpreting decision-tree methods.

Partial Dependence Plots. PD plots are a model-agnostic method that was incorporated to illustrate the predictor-criterion relationship for decision tree ensemble methods. As opposed to the variable importance analysis, PD plots illustrate the nature of the predictor-criterion relationship rather than the strength of their relationship. Specifically, PD plots show the predicted directionality and (non)linearity between a single predictor and the criterion, while simultaneously accounting for the predicted effects of the other predictors used by the estimation method (Molnar, 2021). PD plots depict a single predictor-criterion relationship for the entire

sample. Figure 4 presents a PD plot of variable 1 and the criterion, where the predictor-criterion relationship resembled a negative logarithmic relationship as variable 1 increased from 0 to 30.

Individual Conditional Effect Plots. ICE plots are another model-agnostic method that I included to show the predictor-criterion relationship for decision tree ensemble methods. Unlike PD plots, ICE plots depict the predicted predictor-criterion relationship for each data point in the sample separately, as opposed to showing the average predicted relationship and accounting for the average effect of the other predictors used by the estimation method (Molnar, 2021). The results of an ICE plot are separate predicted predictor-criterion lines for each data point within the sample rather than a single predicted average line.

Variables.

Gender. The majority of the employee sample in Patalano & Huebner's (2021) dataset was female (56.6%). Only male and female gender designations were listed for employees.

Racial Ethnicity. Racial ethnicity was described by Patalano & Huebner (2021) as the race that employees self-identified with. Approximately 60.1% of employees self-identified as White, 25.7% self-identified as Black or African American, 9.3% self-identified as Asian, 3.5% self-identified as two or more races, 1.0% identified as American Indian or Alaska Native, and 0.3% self-identified as Hispanic.

Number of Special Projects. This variable was described as the number of special projects that the employee worked on during the last six months. The mean number of special projects that employees worked on was 1.22 (SD = 2.35), and the range was from 0 to 8 projects.

Number of Times the Employee was Late to Work. This variable reflected the number of times that the employee was late to work during the last 30 days. The mean number of times employees were late to work was 0.41 (SD = 1.30), and the range was from 0 to 6 times.

Absences. Absences referred to the number of times that an employee was absent from work throughout their duration of their employment. The mean number of employee absences was 10.24 (SD = 5.85), and the range was from 1 to 20 times.

Employee Engagement. Employee engagement was described by Patalano & Huebner (2021) as employee's results from the last engagement survey, managed by an external partner. Employee engagement scores were scored from 1 to 5, and the observed range of scores was from 1.12 and 5.00. The mean engagement score for employees was 4.11 (SD = 0.79).

Employee Satisfaction. Patalano & Huebner (2021) described employee satisfaction as a basic satisfaction score between 1 and 5, as reported by employees on a recent employee satisfaction survey. The mean score for employee satisfaction was 3.89 (SD = 0.91), and the observed range was 1 to 5.

Results

Validity Results

Table 11 provides the validity coefficients of the seven estimation methods for predicting job performance. The sample size of the data used to statistically test these coefficients was 75. All seven estimation methods produced statistically significant prediction equations. As shown in Table 11, six of the seven estimation methods produced validity coefficients in the 0.60 range: (i.e., regression; random forests; extreme gradient boosting; ridge regularization; lasso regularization; elastic net and one (i.e., the decision tree method) produced a validity coefficient in the 0.70 range (i.e., $r(75) = 0.71$, $p < 0.01$).

Steiger's Z-test for comparing dependent correlations were used to compare the validity coefficient for the regression method to the validities of the different ML estimation methods. As shown in the second row of Table 11, no significant differences were observed between

regression and the other ML methods. As in Studies One and Two, I did not find any of the ML estimation methods to have validity greater than the regression method

Fairness Results between Male & Female Employees

The same repeated measures factorial ANOVA was performed to assess the main and interactive effects of estimation method and gender on predicted scores. Table 12 presents the results of the repeated measures factorial ANOVA. Consistent with Studies One and Two, a main effect of estimation method on predicted scores was detected, $F(6, 438) = 2.16$, $p < 0.05$, $\eta^2 = 0.03$. No main effect of gender nor interaction between estimation method and gender were found.

Bonferroni pairwise comparisons were used to assess pairwise comparisons on the estimation method factor. No significant differences were found in the predicted scores produced by regression (M = 2.96, SD = 0.05) and the machine learning estimations: e.g., decision tree (M = 2.96, SD = 0.04), random forests (M = 3.00, SD = 0.05), extreme gradient boosting (M = 2.96, SD = 0.05), ridge regularization (M = 2.96, SD = 0.05), lasso regularization (M = 2.95, SD = 0.05) and elastic net regularization (M = 2.95, SD = 0.05). Table 13 presents the predicted scores and Cohen's d-statistic for men and women across each of the seven estimation methods.

Effect sizes of mean differences for men and women were in the -0.30 range for regression (d = -0.30, 95% CI [-0.76, 0.16]), ridge regularization (d = -0.31, 95% CI [-0.77, 0.15]), lasso regularization (d = -0.32, 95% CI [-0.78, 0.14]), and elastic net regularization (d = -0.31, 95% CI [-0.77, 0.15]). The d-statistic was in the -0.40 range for the decision tree (d = -0.44, 95% CI [-0.90, 0.02]), random forests (d = -0.44, 95% CI [-0.90, 0.03]), and extreme gradient boosting (d = -0.42, 95% CI [-0.87, -0.05]) estimation methods. However, no significant differences in effect sizes were observed between any two estimation methods. The negative

effect size indicated that predicted job performance scores were greater for female employees than male employees.

The Cleary Model of Test Bias was used to examine whether each estimation method exhibited gender test bias in its prediction of job performance. The results of the three regression models (shown in table 14) were used to detect slope and/or intercept bias for each method. Gender test bias was observed for four estimation methods. Slope bias was detected for the random forests estimation method. Intercept bias was identified in the regression, random forests, lasso regularization and elastic net regularization estimations. Gender test bias was not detected for the decision tree, extreme gradient boosting and ridge regularization estimation methods.

Incremental validity analyses on the three Cleary models revealed a significant improvement in predicting job performance for only the random forests estimation method. In particular, the addition of the interaction between gender and predicted job performance (Cleary Model 3) enhanced the predictive ability for random forests ($\Delta R^2 = 0.038$, $F(1, 71) = 4.97$). No significant improvements were detected for the six other estimation methods. Accordingly, only the decision tree, extreme gradient boosting, and ridge regularization estimation methods would be appropriate for use in selection based on the fairness results for men and women.

Fairness Results between Racial Majority & Racial Minority Employees

Table 15 presents the results of the repeated measures factorial ANOVA analysis used to examine the main and interactive effects of estimation method and race on predicted scores. A main effect of estimation method was again found on predicted scores, $F(6, 439) = 2.27$, $p < 0.05$, $\eta^2 = 0.03$. No main effect of race nor interaction between estimation method and race were found. Bonferroni pairwise comparisons did not detect significant differences in the predicted

scores produced by regression and the machine learning estimation methods. Table 16 presents the predicted scores and Cohen's d-statistic for racial majorities and racial minorities across each of the seven estimation methods.

Effect sizes were in the 0.00 range for the seven estimation methods: e.g., regression (d = 0.02, 95% CI [-0.44, 0.48]), decision tree (d = 0.10, 95% CI [-0.46, 0.56]), random forests (d = 0.06, 95% CI [-0.40, 0.41]), and extreme gradient boosting (d = 0.08, 95% CI [-0.38, 0.54]), ridge regularization (d = 0.02, 95% CI [-0.44, 0.47]), lasso regularization (d = 0.02, 95% CI [-0.44, 0.48]), and elastic net regularization (d = 0.02, 95% CI [-0.44, 0.48]). No significant differences in effect sizes were observed between any two estimation methods.

The Cleary Model revealed test bias in the prediction of job performance between racial majorities and minorities in three estimation methods. Table 17 lists the results of the Cleary Model for each estimation method. Here, we see both slope and intercept biases for the decision tree, random forests, and extreme gradient boosting estimations. Regression and the three regularization-based machine learning estimation methods did not demonstrate either form of test bias for race[23].

Incremental validity analyses on the three Cleary models revealed significant improvements in predicting job performance for the decision tree, random forests, and extreme gradient boosting estimation methods. Specifically, the addition of the interaction between race and predicted job performance (Cleary Model 3) enhanced the prediction for the decision tree ($\Delta R^2 = 0.060$, $F(1, 71) = 9.95$), random forests ($\Delta R^2 = 0.036$, $F(1, 71) = 4.54$), and extreme

[23] A similar pattern in the presence of slope and/or intercept bias was observed between Black/African American and White employee specifically. Slope and intercept bias was detected for the decision tree and extreme gradient boosting estimations. Slope bias was detected for random forests estimation. Regression and the three regularization-based machine learning estimation methods did not demonstrate either form of test bias between Black/African American and White employees.

gradient boosting estimations ($\Delta R^2 = 0.046$, $F(1, 71) = 6.94$). No significant improvements were observed for regression, ridge regularization, lasso regularization, and elastic net regularization.

Machine Learning Interpretability Results

Combined, the results of the fairness analysis for gender and race, suggested that ridge regularization was the only estimation method appropriate for use in selection due to its validity in predicting job performance, as well as lack of predictive bias. Fortunately, ridge regularization is a highly interpretable ML method as it also produces coefficients that can be interpreted similarly to regression. Table 18 depicts the coefficients for regression and the ridge regularization estimations. Here, we see similar coefficients between the estimation methods, albeit some coefficients for regularization are smaller compared to regression.

The regression coefficients revealed two significant predictors of job performance (e.g., number of times employee was late to work, and absences) whereas the ridge regularization coefficients indicated three significant predictors of job performance (e.g., number of times employee was late to work, absences, and employee engagement). The negative relationship between the number of times the employee was late to work and job performance was nearly identical between regression ($b = -0.27$, $t(74) = -4.92$, $p < 0.01$) and ridge regularization ($b = -0.24$, $t(74) = -5.433$, $p < 0.01$). The positive relationship between absences and job performance for regression ($b = 0.02$, $t(74) = 2.32$, $p = 0.02$) and ridge regularization ($b = 0.02$, $t(74) = 2.35$, $p = 0.02$) was identical. Employee engagement significantly predicted job performance for only ridge regularization ($b = 0.17$, $t(74) = 2.37$, $p = 0.02$), though similar pattern was observed for regression ($b = 0.15$, $t(74) = 1.71$, $p = 0.09$). As with regression, beta coefficients offered important explanatory information for interpreting the predictors used by regularization-based methods.

Though, ridge regularization was the only estimation method to not demonstrate predictive bias in Study Three, I also conducted ML interpretability methods for the random forests estimation as it was the only estimation to not demonstrate predictive bias in Study Two. Accordingly, the variable importance, PD plot and ICE plot analyses for the random forests estimation in Study Three are presented below for illustrative purposes.

Figure 5 depicts the results of a variable importance analysis for the random forests estimation. Here, we see four of the five predictors with importance score greater than zero. The rank-order of the four important predictor variables consisted of (1) number of times the employee was late to work, (2) employee engagement, (3) employee satisfaction, and (4) absences. Further, the importance score of the number of times the employee was late to work was roughly triple that of employee engagement, whereas the importance score of employee engagement was double that of employee satisfaction and absences. Indeed, the rank-order and magnitude of the variable importance scores painted a similar picture to the beta coefficients for regression and ridge regularization, with the number of times the employee was late to work and employee engagement displaying the largest beta coefficient values, respectively. Variable importance scores allow us to determine the relative contribution of each predictor and can be treated as effect-size measures for predictors within ensemble-based ML methods.

PD plots were generated to show the nature of the predictor-criterion relationship for the two most important variables in the random forests estimation. Figure 6 presents the predicted directionality and (non)linearity between the number of times the employee was late to work and job performance scores. Here, we see a negative predictor-criterion relationship with the largest drop in predicted job performance occurring between 1 and 2 days late compared to other days.

Figure 7 presents a PD plot for employee engagement and job performance. As seen in Figure 7, the predictor-criterion relationship was positive. Job performance scores peaked and exceeded 3.1 at employee engagement scores of 3.8 and 4.7. However, job performance scores decreased and plateaued near 3.0 between employee engagement scores of 3.8 and 4.7; and decreased to 2.95 between employee engagement scores of 4.7 and 5.

In addition to PD plots, ICE plots were created to indicate the predictor-criterion relationship for the two most important variables in the random forests estimation. Figures 8 and 9 show the ICE plots for the number of times the employee was late to work and employee engagement, respectively. As seen in Figure 8, the same negative relationship between the number of times the employee was late to work and job performance was depicted by the red line in the ICE plot to the PD plot in Figure 6. However, unlike Figure 6, Figure 8 shows the predicted relationship between the variables for each case in the sample. Here, we see minor differences in the predicted relationship for each case as evidenced by the fluctuations between the averaged red line and each black line. In addition, the vertical axis of the ICE plot reflects the magnitude of the effect that the predictor has on the criterion as opposed to the PD plot which shows the predicted score of the criterion.

Figure 9 shows the averaged predicted relationship of employee engagement and job performance in red, and the predicted relationship for each case of the sample in black. Again, a positive predictor-criterion relationship was observed with job performance scores peaking near 3.8 and 4.7 for the majority of cases in the sample. However, there appeared to be substantive individual variation between employee engagement scores of 3.8 and 4.7 with some cases experiencing increases or decreases in job performance scores within the interval. Evidently, ICE plots provide additional information than PD plots by showing the predicted predictor-criterion

relationship for each data point in the sample separately, as opposed to only showing the average predicted relationship.

Discussion

Study Three focused on the validity and fairness of the estimation methods, as well as their interpretability. Specifically, I used oversampling and content valid predictors to develop the estimations, and then incorporated variable importance, partial dependence, and individual conditional expectation plot analyses to interpret the ML estimations. As opposed to Studies One and Two, the criterion variable consisted of employee job performance ratings, and both validity and fairness were examined across race and gender as both protected classes were provided in Patalano & Huebner's (2021) HR Metrics and Analytics dataset.

In terms of validity, I found the six ML estimation methods to be equivalent in predicting employee job performance ratings to regression. Indeed, no differences in validity were observed using Steiger's Z-test for comparing dependent correlations. Thus, when only the validity results are considered, all estimation methods appear appropriate to use for selection.

Regarding fairness, I found only the ridge regularization estimation to not produce predictive bias. In particular, neither slope nor intercept bias was detected for the ridge regularization estimations for gender and race. I found the regression, lasso regularization, and elastic net regularization estimations to show slope and intercept bias for gender; the extreme gradient boosting estimation to show slope and intercept bias for race; and the decision tree and random forests estimations to show slope and intercept bias for gender and race. As such, I reached a similar conclusion to Study Two where only one ML estimation method would be appropriate and legally defensible for selection purposes (*Cormier v. P.P.G Indus.*, 1983; *Hamer*

v. City of Atlanta, 1989; *United States v. City of Erie*, 2005). However, in Study Three, the only legally defensible method was ridge regularization as opposed to the random forests method.

In terms of adverse impact, no significant differences were observed in the adverse impact effect sizes of the estimations for gender nor race. For gender, predicted job performance scores were greater for women than men as indicated by adverse impact effect sizes in the -0.30 range for regression, ridge regularization, lasso regularization and the elastic net regularization estimations; and -0.40 range for the decision tree, random forests, and extreme gradient boosting estimations. For race, predicted job performance scores were slightly in favor for racial majorities over racial minorities with adverse impact effect sizes in the 0.00–0.10 range for all seven estimations. Because predictive bias was detected in all estimation methods except for ridge regularization, the shifting burden of proof model would not apply here as there is no evidence that an equivalently valid and psychometrically fair alternative to ridge regularization exists which shows a smaller adverse impact (Hanges et al., 2013). Only the ridge regularization estimation method would be recommended for use in selection and any decision to use the other estimation methods or even toss out the results of the ridge estimation after implementation would be unadvised (Aguinis et al., 2010; Gutman et al., 2017; *Ricci v. DeStefano*, 2009).

The ML interpretability analyses included in Study Three provided additional support to the legal defensibility of the ridge regularization estimation. In particular, I showed how the beta coefficients produced by the ridge estimation can be interpreted identically to that of regression. In addition, I showed how variable importance, PD and ICE plot analyses can be applied to interpret random forests and other ensemble tree-based estimation methods. As shown in the variable importance analysis, we saw which variables were most important for the random forests estimation in predicting job performance. The variable importance analysis provided us

with comparable information to an effect size. The PD and ICE plots showed us the directionality and nature of the predictor-criterion relationship. ML interpretability analyses provide a means to unpacking the black box often associated with ML methods.

From a legal perspective, the ML interpretability analyses provide additional evidence of criterion-related validity by showing the statistical relationship between each predictor and the criterion as opposed to only the overall validity of the estimation method in predicting job performance. For instance, in addition to observing a validity coefficient of 0.67 between the predicted and observed job performance scores for regression and ridge regularization, we also saw that the number of times the employee was late to work negatively predicted job performance and was the strongest predictor of job performance for both estimation methods. In addition, we saw that employee absences were a significant predictor of job performance for both estimations, while employee engagement was a significant predictor for the ridge regularization estimation only. Employee satisfaction and the number of special projects did not predict job performance for both estimations, however the inclusion of them as predictors were appropriate as they both were supported by content validity evidence.

On that note, the courts have consistently ruled that there is no preference for criterion-related validity evidence nor any alternative single form of validity evidence for that matter (*Gillespie v. State of Wisconsin,* 1985). Indeed, the courts have required that criterion-related validity evidence be preceded by work analysis (*Moody v. Albemarle*, 1973; *Guardians v. Civil Service*, 1980; *PGA v. Martin*, 2001) and have upheld the view that "the greater number of converging sources of validation evidence, the better (Gutman et al., 2017)." For instance, in *Police Officers v. City of Columbus* (1990), the courts ruled that selection procedures must measure a substantial and important part of the job reliably (i.e., content valid) as well as provide

adequate discrimination (i.e., criterion valid). In *Bradley v. City of Lynn* (2006), the courts ruled that the use of cognitive ability as the sole basis for selecting firefighters was insufficient as there was not enough evidence of job relatedness and/or examination of other alternative predictors. This ruling was made despite cognitive ability predicting firefighter performance, as well as meta-analytic support[24] of cognitive ability predicting job performance (Hunter, 1986; Van Iddekinge et al., 2017). Thus, when it comes to legal defensibility both content and criterion validity evidence are critical, making any effort to maximize criterion validity at the expense of omitting content validity highly questionable.

From a statistical viewpoint, the observed null effects of the content valid predictors can be viewed as type II errors stemming from model overfitting. This was possible since only 311 employees were present in Patalano & Huebner's (2021) sample and ML tends to overfit the data in small datasets ($n < 1000$) (Combrisson & Jerbi, 2015; Vabalas et al., 2019). In such case, it was possible that the null effects of employee satisfaction and number of special projects on performance ratings reflected an anomality specific to the training dataset[25]. If a larger training dataset was used, then the content valid predictors could emerge as statistically significant predictors of job performance given that a significant relationship between the content valid predictors and the criterion indeed exists. Moreover, if content validity was enhanced and the predictors were specifically developed based upon a rigorous job analysis as opposed to selecting

[24] The courts have ruled that meta-analytic support can be used alongside other forms of job relatedness to show the validity of selection methods (*Adams v. City of Chicago*, 1996; *Williams v. Ford Motor Co.*, 1999).

[25] When the entire dataset of 311 employees was analyzed, significant zero-order correlations with performance ratings were observed for the number of times the employee was late to work ($r(309) = -0.74$), absences ($r(309) = 0.05$), employee engagement ($r(309) = 0.55$), and employee satisfaction ($r(309) = 0.30$). No significant zero-order correlation was observed between performance ratings and the number of special projects.

available variables supported by content valid, then it is possible that the content valid predictors would also be supported by criterion validity.

Lastly, aside from evaluating the validity of the predictors used for selection, it is important to also consider the validity of the criterion variable. Consider a scenario where the performance appraisal system is deficient, and the criterion variable is contaminated with aspects other than the job itself (e.g., subjective ratings, rating biases, discrimination, etc.). In this case, content valid predictors may appear ineffective in predicting the criterion of interest. Insights gleaned regarding the usefulness of predictors from purely criterion validity approaches may be highly inappropriate and replicate existing deficiencies in the performance appraisal system. For instance, if racial discrimination affected performance ratings, then criterion-related validity methods would capture relationships that perpetuate existing racial discrimination if implemented. It is for this, and other reasons related to the history of employment discrimination in the United States, that the courts have recognized that multiple forms of validity exist, ruled that multiple forms of validity evidence are preferred, and required that all validity studies be preceded by work analysis (*Gillespie v. State of Wisconsin*, 1985; Gutman et al., 2017).

Chapter Nine: General Discussion & Implications

Although each study had its limitations, in combination, the reported results provided a detailed examination of the validity and fairness of machine learning estimation methods relative to regression in personnel selection. Following the results of Study One, I found that ML performed equivalently to regression in terms of validity and fairness, albeit all estimation methods were inappropriate for use in selection as they demonstrated predictive bias. In Study Two, I found that ML performed equivalently to regression in terms of validity, yet superior in terms of fairness. Specifically, the random forests estimation method did not demonstrate predictive bias, whereas the regression estimation did. Lastly, in Study Three, I detected a similar pattern where ML performed equivalently to regression in terms of validity but superior in terms of fairness. However, while the random forests and regression estimations exhibited predictive bias, the ridge regularization estimation was the only method that did not. Combined, the results of my dissertation revealed instances where ML provided utility over regression in selection.

Regarding the manipulations, I found that the oversampling manipulation, i.e., the manner by which male and female cadets/midshipmen were distributed in the training/validation datasets, did not affect estimation validity nor fairness. Further, the content validity manipulation did not impact the validity and fairness of the estimation methods. However, both oversampling and content valid predictors were incorporated into my subsequent studies on account of their importance for the legal defensibility of the estimation methods. Again, oversampling is consistent with current selection practice, and content validity is recognized by the courts as being equally important for legal defensibility as criterion validity. In Study Three, I focused further on the legal defensibility of the estimation methods by incorporating both oversampling and content valid predictors, as well as three ML interpretability methods: e.g., variable

importance, PD plots, and ICE plots. The ML interpretability methods unpacked the black box associated with ML estimation methods and provided additional evidence of criterion related validity by revealing the specific nature of the predictors and the criterion used by the estimation methods. Study Three thus presented a highly legally defensible use of ML in selection, which was critical given that the results revealed that ML was a more appropriate method to be used in selection than regression.

Across all three studies, several legal cases and considerations were discussed. First, the courts preference for a lack of predictive bias over reductions in adverse impact for constituting a fair selection method was discussed (*Cormier v. P.P.G. Indus.,* 1983; *Hamer v. City of Atlanta,* 1989; *United States v. City of Erie,* 2005). In particular, predictive bias is unaccepted by the courts and scientific community as it reflects a property inherent to the selection method itself; whereas adverse impact is allowed since it is an external feature reflecting the tests' consequences (Aguinis et al., 2010; Hanges et al., 2013; Tippins et al., 2018). According to the shifting burden of proof model, AI is defensible when the employer can demonstrate that an equivalently valid and fairer (i.e., lower AI) alternative selection method could not have reasonably existed and/or been used at the time that the selection method in question was used. The shifting burden of proof model is a critical factor when deciding which selection method is most legally defensible and appropriate to implement.

Indeed, the use of a less valid selection method for the intent of increasing fairness is explicitly outlawed by the courts as it violates *Title VII of the Civil Rights Act* by considering applicants protected class information to inform the selection decision (*Ricci v. DeStefano*, 1990). Selection methods can only be modified if the modification is made in a race-neutral fashion (*Hayden v. Nassau County,* 1996). Protected class information can only be explicitly

considered when sufficient business justification can be made regarding how greater representation of the protected group in question enhances the organizations' interests and ability to perform its duties successfully (*Detroit Police Officers Association v. Young,* 1971; *Petit v. Chicago*, 2003), and/or when a particular job requires certain essential protected characteristics to perform the job adequately (*Dothard v. Rawlinson*, 1977; *Slivka v. Camden-Clark Memorial Hospital*, 2004).

Lastly, the importance of content validity for selection methods was discussed. First, the courts have required that criterion-related validity evidence be preceded by work analysis (*Moody v. Albemarle,* 1973; *Guardians v. Civil Service,* 1980; *PGA v. Martin,* 2001). Second, the courts have ruled that there is no preference for criterion-related validity evidence over content validity evidence in terms of establishing legal defensibility of the selection method (*Gillespie v. State of Wisconsin*, 1985). The importance of content validity is highlighted in *Police Officers v. City of Columbus* (1990) and *Bradley v. City of Lynn* (2006). In *Police Officers v. City of Columbus* (1990), the courts ruled that selection methods must be content valid and measure a substantial part of the work being performed by the job, as well as be effective in predicting performance scores (i.e., criterion valid). In *Bradley v. City of Lynn* (2006), the courts ruled that cognitive ability was insufficient as being the sole basis for selecting firefighters- despite considerable criterion validity evidence -as cognitive ability on its own lacked sufficient content validity evidence for firefighters, and alternative predictors with lower AI could have also been used. Indeed, content validity is a critical feature for the legal defensibility of a selection method, making any selection method that lacks content validity for the sake of optimizing criterion validity highly questionable.

Though a thorough examination regarding the legal defensibility and context surrounding selection procedures was made, it is important to note that fairness is a social construct. Specifically, the notion of what constitutes as fair is subjective and value laden (Cascio & Aguinis, 2019; Landon & Arvery, 2007). For instance, fairness can be based upon equality norms, which emphasize that individuals be treated the same way regardless of individual difference or need; or equity norms, which recognize that individuals experience different circumstances and that such differences must be accounted for in order to provide equal opportunities for individuals. In my dissertation, I drew upon legal rulings on fairness to define whether the selection method was considered to be fair. As such, the courts' preference for equality-based definitions of fairness in the area of employment, such as the Cleary Model, was given precedence over equity-based definitions such as adverse impact. No challenge was made regarding the appropriateness of the courts' preference of fairness as the focus of my dissertation was on applying selection methods legally in employment contexts.

The appropriateness of equality v. equity-based definitions of fairness is significant because the latter definition has been preferred by the courts for selection in contexts outside employment. In the context of higher education, the courts have ruled in favor of several equity-based definitions of fairness regarding college admission affirmative action policies. In *Regents of the University of California v. Bakke* (1978), the Supreme Court held that the practice of considering applicants' racial background during the college admissions process was permissible but prohibited the use of racial quotas. In *Grutter v. Bollinger* (2003), the Supreme Court held that policies that favor underrepresented minority groups were appropriate, so long as other factors were considered and that protected class was only evaluated on an individual basis for every applicant. The logic behind *Grutter v. Bollinger* (2003) was that of business necessity as it

was in the universities' best interest to obtain the educational benefits that emerge from a diverse student body. Other affirmative action policies that favor underrepresented minority group policies have recently been supported by the courts (*SFFA v. President and Fellows of Harvard College*, 2019; *SFFA v. University of North Carolina*, 2021). Thus, the key takeaway here from a practitioner's perspective is that the appropriateness of equality v. equity-based definitions of fairness depends upon the legal rulings surrounding the context in which the selection method is to be used in. However, from an ethics perspective, the appropriateness of equality v. equity-based definitions largely depends upon one's own values on fairness.

Finally, although I demonstrated the utility of ML over regression in selection and discussed relevant legal considerations, it is critical to note that my dissertation was focused on the effect that two common engineering decisions had on the validity and fairness of algorithms during the engineering stage. Indeed, other less common engineering decisions could have been explored such as the use of item-level scores versus scale-level scores, standardized versus unstandardized scores, and/or dimension reduction techniques. Further, the other three stages of the selection system were not evaluated in this dissertation. For instance, the broader strategy stage could have been examined by emphasizing the need for diversity for business necessity purposes, manipulating characteristics of the dataset such as the presence and/or magnitude of interaction effects between predictor-criterion relationships for protected groups, and/or the optimization of different and/or multiple criterion variables simultaneously. Future research is needed on the impact of alternative engineering decisions to the performance of regression and ML during the engineering stage, as well as other decisions and factors influencing the broader strategy, application and/or reviewing stages of selection.

Lastly, it is important to note that my dissertation did not include a traditional selection dataset. In particular, Hanges et al.'s (2021) sample was intended to predict CWBs in the military as opposed to job performance, and Patalano & Huebner's (2021) sample was a teaching dataset designed to simulate HR data for HR professionals. As such, future research is needed to examine the validity and fairness of the estimation methods with a traditional selection dataset. In doing so, researchers can identify whether the utility of ML over regression in selection generalizes to a traditional selection context. Regardless, my dissertation has showed instances where ML exhibited enhanced fairness and legal defensibility over regression in nontraditional selection contexts, lending support to this dissertation's primary aim, which consisted of examining whether machine learning estimations could *potentially* present an improvement in estimation validity or fairness in selection.

Appendices

Table 1. Validity Coefficients & Steiger's Z Comparison of Dependent Correlations

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Validity	0.46 (0.42, 0.49)	0.34 (0.30, 0.38)	0.41 (0.37, 0.45)	0.45 (0.41, 0.48)	0.46 (0.42, 0.49)	0.46 (0.42, 0.49)	0.46 (0.42, 0.49)
Comparing Validity to Regression	-	$Z_H = 7.99$ ($p < 0.01$)	$Z_H = 4.76$ ($p < 0.01$)	$Z_H = 1.36$ ($p > 0.05$)	$Z_H = 0$ ($p > 0.05$)	$Z_H = 0$ ($p >0 .05$)	$Z_H = 0$ ($p > 0.05$)

Note: Sample size = 2023. The first row of this table provides the correlation of each estimation method with CWBs and the 95% confidence intervals for each correlation are in parentheses. The second row provides the results of the Steiger's Z-test for dependent correlations. In particular, the table shows the Z-test values comparing the validity of each ML estimation method to the validity of the regression method.

Table 2. Main Effects of Repeated Measures ANOVA on Predicted Criterion Scores

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method	6, 12090	0.72	0.12	21.48	<0.00	0.01
Gender	1, 2015	22.74	22.74	81.51	<0.00	0.04
Oversampling	3, 2015	0.26	0.09	0.31	0.82	0.00

Table 3. Interaction Effects of Repeated Measures ANOVA on Predicted Criterion Scores

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method x Gender	6, 12090	0.60	0.10	17.83	<0.00	0.01
Estimation Method x Oversampling	18, 12090	0.13	0.01	1.32	0.16	0.00
Oversampling x Gender	3, 2015	0.24	0.08	0.29	0.83	0.00
Estimation Method x Gender x Oversampling	18, 12090	0.74	0.00	0.73	0.78	0.00

Table 4. Predicted Criterion Means & Cohen's d-Statistic

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Overall	1.86	1.86	1.87	1.85	1.86	1.86	1.86
Men	1.89	1.88	1.90	1.88	1.89	1.89	1.89
Women	1.80	1.83	1.82	1.79	1.80	1.80	1.80
Cohen's d-statistic	0.44 (0.34, 0.53)	0.30** (0.21, 0.40)	0.36 (0.26, 0.45)	0.41 (0.32, 0.50)	0.44 (0.35, 0.53)	0.44 (0.35, 0.53)	0.44 (0.35, 0.53)

Table 5. Cleary Model Test Bias

Regression	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.208	-	1	2021	529.32**
• Predicted Score	0.98	0.04	0.46**					
Model 2:				0.208	0.001	1	2020	1.98
• Predicted Score	0.96	0.04	0.45**					
• Gender	-0.03	0.02	-0.03					
Model 3:				0.213	0.005	1	2019	11.79**
• Predicted Score	1.38	0.13	0.65**					
• Gender	0.54	0.17	0.54**					
• Predicted Score * Gender	-0.31	0.09	-0.57**					
Decision Tree	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.117	-	1	2021	268.39**
Predicted Score	0.91	0.06	0.34**					
Model 2:				0.122	0.005	1	2020	11.68**
Predicted Score	0.88	0.06	0.33**					
Gender	-0.07	0.02	-0.07**					
Model 3:				0.122	<0.000	1	2019	0.15
Predicted Score	0.94	0.17	0.36**					
Gender	0.02	0.23	0.02					
Predicted Score * Gender	-0.05	0.12	-0.09					
Random Forests	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.168	-	1	2021	408.79**
Predicted Score	0.83	0.04	0.41**					
Model 2:				0.171	0.003	1	2020	6.48*
Predicted Score	0.81	0.04	0.40**					
Gender	-0.05	0.02	-0.05*					
Model 3:				0.173	0.002	1	2019	5.93*
Predicted Score	1.10	0.12	0.54**					
Gender	0.35	0.17	0.35*					
Predicted Score * Gender	-0.22	0.09	-0.40*					
Extreme Gradient Boosting	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.200	-	1	2021	505.76**
Predicted Score	0.99	0.04	0.45**					
Model 2:				0.201	0.001	1	2020	3.00
Predicted Score	0.98	0.05	0.44**					
Gender	-0.04	0.02	-0.04					
Model 3:				0.205	0.004	1	2019	9.42**
Predicted Score	1.36	0.13	0.62**					
Gender	0.50	0.18	0.50**					
Predicted Score * Gender	-0.30	0.10	-0.53**					
Ridge Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.207	-	1	2021	528.36**
Predicted Score	1.01	0.04	0.46**					
Model 2:				0.208	0.001	1	2020	1.73
Predicted Score	1.00	0.05	0.45**					
Gender	-0.03	0.02	-0.03					
Model 3:				0.213	0.005	1	2019	11.72**
Predicted Score	1.44	0.14	0.65**					
Gender	0.57	0.18	0.56**					
Predicted Score * Gender	-0.33	0.10	-0.59**					
Lasso Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**

Model 1:				0.208	-	1	2021	529.61**
Predicted Score	0.99	0.04	0.46**					
Model 2:				0.208	0.001	1	2020	1.83
Predicted Score	0.98	0.04	0.45**					
Gender	-0.03	0.02	-0.03					
Model 3:				0.213	0.005	1	2019	11.67**
Predicted Score	1.40	0.13	0.65**					
Gender	0.55	0.17	0.55**					
Predicted Score * Gender	-0.32	0.09	-0.57**					
Elastic Net Regularization	**B**	**SE**	$\boldsymbol{\beta}$	$\mathbf{R^2}$	$\mathbf{\Delta R^2}$	$\mathbf{df_1}$	$\mathbf{df_2}$	**F Change**
Model 1:				0.208	-	1	2021	529.37**
Predicted Score	0.99	0.04	0.46**					
Model 2:				0.208	0.001	1	2020	1.81
Predicted Score	0.98	0.04	0.45**					
Gender	-0.03	0.02	-0.03					
Model 3:				0.213	0.005	1	2019	11.58**
Predicted Score	1.40	0.13	0.64**					
Gender	0.55	0.17	0.55**					
Predicted Score * Gender	-0.32	0.09	-0.57**					

Table 6. Validity Coefficients & Steiger's Z Comparison of Dependent Correlations

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Validity	0.53 (0.49, 0.57)	0.36 (0.31, 0.40)	0.51 (0.47, 0.55)	0.53 (0.50, 0.57)	0.53 (0.49, 0.57)	0.53 (0.49, 0.57)	0.53 (0.49, 0.57)
Comparing Validity to Regression	-	$Z_H = 8.94$ ($p < 0.01$)	$Z_H = 1.77$ ($p > 0.05$)	$Z_H = 0.23$ ($p > 0.05$)	$Z_H = 1.77$ ($p > 0.05$)	$Z_H = 0.51$ ($p > 0.05$)	$Z_H = 0.51$ ($p > 0.05$)

Note: Sample size = 1512. The first row of this table provides the correlation of each method with CWBs and the 95% confidence intervals for each correlation are in parentheses. The second row provides the results of the Steiger's Z-test for dependent correlations. In particular, the table shows the Z-test values comparing the validity of each ML method to the validity of the regression method.

Table 7. Main Effects of Repeated Measures ANOVA on Predicted Criterion Scores

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method	6, 9036	0.17	0.03	3.68	<0.00	0.00
Gender	1, 1506	1.98	1.98	40.89	<0.00	0.03
Content validity manipulation	2, 1506	0.01	0.00	0.08	0.92	0.00

Table 8. Interaction Effects of Repeated Measures ANOVA on Predicted Criterion Score

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method x Gender	6, 9036	0.25	0.04	5.33	<0.00	0.00
Estimation Method x Manipulation	12, 9036	0.03	0.00	0.37	0.98	0.00
Manipulation x Gender	2, 1506	0.00	0.00	0.01	0.99	0.00
Estimation Method x Gender x Manipulation	12, 9036	0.02	0.00	0.19	0.99	0.00

Table 9. Predicted Criterion Means & Cohen's D-Statistic

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Overall	1.86	1.86	1.87	1.85	1.86	1.86	1.86
Men	1.89	1.88	1.90	1.88	1.89	1.89	1.89
Women	1.80	1.82	1.82	1.80	1.80	1.80	1.80
Cohen's d-statistic	0.34 (0.23, 0.45)	0.27 (0.16, 0.38)	0.33 (0.22, 0.43)	0.33 (0.22, 0.44)	0.36 (0.25, 0.47)	0.35 (0.24, 0.46)	0.35 (0.24, 0.46)

Table 10. Cleary Model Test Bias

Regression	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.282	-	1	1510	592.53**
• Predicted Score	1.01	0.04	0.53**					
Model 2:				0.282	0.001	1	1509	1.40
• Predicted Score	1.00	0.04	0.53**					
• Gender	-0.03	0.02	-0.03					
Model 3:				0.287	0.004	1	1508	8.87**
• Predicted Score	1.36	0.13	0.71**					
• Gender	.46	0.17	0.44**					
• Predicted Score * Gender	-.27	0.09	-0.48**					
Decision Tree	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.128	-	1	1510	221.16**
Predicted Score	0.86	0.06	0.36					
Model 2:				0.132	0.004	1	1509	7.091**
Predicted Score	0.84	0.06	0.35**					
Gender	-0.07	0.03	-0.06**					
Model 3:				0.135	0.003	1	1508	5.053*
Predicted Score	0.45	0.18	0.19*					
Gender	-0.67	0.27	-0.63*					
Predicted Score * Gender	0.33	0.15	0.57*					
Random Forests	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.262	-	1	1510	536.67**
Predicted Score	1.09	0.05	0.51**					
Model 2:				0.263	0.001	1	1509	2.01
Predicted Score	1.08	0.05	0.51**					
Gender	-0.03	0.02	-0.03					
Model 3:				0.264	0.001	1	1508	1.96
Predicted Score	1.27	0.14	0.60**					
Gender	.024	0.20	0.23					
Predicted Score * Gender	-0.15	0.11	-0.27					
Extreme Gradient Boosting	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.284	-	1	1510	600.39**
Predicted Score	1.13	0.05	0.53**					
Model 2:				0.285	0.001	1	1509	1.52
Predicted Score	1.12	0.04	0.53**					
Gender	-0.03	0.02	-0.03					
Model 3:				0.289	0.004	1	1508	7.45**
Predicted Score	1.48	0.14	0.70**					
Gender	0.48	0.19	0.46*					
Predicted Score * Gender	-0.28	0.10	-0.49**					
Ridge Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.279	-	1	1510	585.09**
Predicted Score	1.09	0.05	0.53**					
Model 2:				0.280	<0.000	1	1509	0.97
Predicted Score	1.08	0.05	0.53**					
Gender	-0.02	0.02	-0.02					
Model 3:				0.284	0.004	1	1508	8.84**
Predicted Score	1.46	0.14	0.71**					
Gender	.051	0.18	0.48**					
Predicted Score * Gender	-0.29	0.10	-0.51**					
Lasso Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**

Model 1:				0.281	-	1	1510	591.04**
Predicted Score	1.06	0.04	0.53**					
Model 2:				0.282	0.001	1	1509	1.08
Predicted Score	1.05	0.04	0.53**					
Gender	-0.02	0.02	-0.02					
Model 3:				0.286	0.004	1	1508	8.10**
Predicted Score	1.41	0.13	0.70**					
Gender	0.47	0.18	0.45**					
Predicted Score * Gender	-0.27	0.10	-0.48**					
Elastic Net Regularization	**B**	**SE**	$\boldsymbol{\beta}$	$\mathbf{R^2}$	$\mathbf{\Delta R^2}$	$\mathbf{df_1}$	$\mathbf{df_2}$	**F Change**
Model 1:				0.281	-	1	1510	590.69**
Predicted Score	1.07	0.04	0.53**					
Model 2:				0.282	0.001	1	1509	1.07
Predicted Score	1.06	0.04	0.53**					
Gender	-0.02	0.02	-0.02					
Model 3:				0.286	0.004	1	1508	8.17**
Predicted Score	1.42	0.13	0.71**					
Gender	0.47	0.18	0.45**					
Predicted Score * Gender	-0.27	0.10	-0.48**					

Table 11. Validity Coefficients & Steiger's Z Comparison of Dependent Correlations

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Validity	0.67 (0.52, 0.78)	0.71 (0.58, 0.81)	0.64 (0.48, 0.76)	0.69 (0.55, 0.79)	0.67 (0.52, 0.78)	0.68 (0.54, 0.79)	0.67 (0.53, 0.78)
Comparing Validity to Regression	-	$Z_H = -1.10$ ($p > 0.5$)	$Z_H = 0.83$ ($p > 0.5$)	$Z_H = -0.62$ ($p > 0.5$)	$Z_H = \infty$ ($p > 0.5$)	$Z_H = -0.82$ ($p > 0.5$)	$Z_H = \infty$ ($p > 0.5$)

Note: Sample size = 75. The first row of this table provides the correlation of each estimation method with performance ratings and the 95% confidence intervals for each correlation are in parentheses. The second row provides the results of the Steiger's Z-test for dependent correlations. In particular, the table shows the Z-test values comparing the validity of each ML estimation method to the validity of the regression method.

Table 12. Main & Interaction Effects of Repeated Measures ANOVA on Predicted Criterion Scores

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method	6, 438	0.15	0.03	2.16	<0.05	0.03
Gender	1, 73	2.93	2.93	2.51	0.12	0.03
Estimation Method x Gender	6, 438	`0.43	0.07	0.61	0.73	0.08

Table 13. Predicted Criterion Means & Cohen's D-Statistic

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Overall	2.96	2.96	3.00	2.96	2.96	2.95	2.95
Men	2.89	2.89	2.91	2.88	2.89	2.88	2.89
Women	3.02	3.04	3.10	3.05	3.02	3.02	3.02
Cohen's d-statistic	-0.30 (-0.76, 0.16)	-0.44 (-0.90, 0.02)	-0.44 (-0.90, 0.03)	-0.42 (-0.87, 0.05)	-0.31 (-0.77, 0.15)	-0.32 (-0.78, 0.14)	-0.31 (-0.77, 0.15)

Table 14. Cleary Model Test Bias

Regression	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.450	-	1	73	59.81**
• Predicted Score	0.96	0.12	0.67**					
Model 2:				0.464		1	72	1.879
• Predicted Score	0.93	0.13	0.65**					
• Gender	-0.15	0.11	-0.12					
Model 3:				0.488		1	71	3.323
• Predicted Score	0.68	0.19	0.47**					
• Gender	-0.15	0.74	-1.16*					
• Predicted Score * Gender	0.45	0.25	1.04					
Decision Tree	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.506	-	1	73	74.85**
• Predicted Score	1.26	0.15	0.71**					
Model 2:				0.511	0.004	1	72	0.650
• Predicted Score	1.23	0.15	0.70**					
• Gender	-0.09	0.11	-0.07					
Model 3:				0.514	0.003	1	71	0.471
• Predicted Score	1.07	0.28	0.60**					
• Gender	-0.78	1.01	-0.60					
• Predicted Score * Gender	0.23	0.33	0.53					
Random Forests	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.406	-	1	73	49.839**
• Predicted Score	0.96	0.14	0.64**					
Model 2:				0.413	0.007	1	72	0.869
• Predicted Score	0.93	0.14	0.62**					
• Gender	-0.11	0.12	-0.09					
Model 3:				0.451	0.038	1	71	4.973*
• Predicted Score	0.54	0.22	0.36*					
• Gender	-1.99	0.85	-1.55*					
• Predicted Score * Gender	0.62	0.28	1.44*					
Extreme Gradient Boosting	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.478	-	1	73	66.939
• Predicted Score	1.08	0.13	0.69**					
Model 2:				0.485	0.006	1	72	0.869
• Predicted Score	1.05	0.14	0.68**					
• Gender	-0.10	0.11	-0.08					
Model 3:				0.491	0.006	1	71	0.869
• Predicted Score	0.86	0.25	0.55**					
• Gender	-0.93	0.89	-0.72					
• Predicted Score * Gender	0.28	0.29	0.64					
Ridge Regularization	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.451	-	1	73	59.860**
• Predicted Score	1.01	0.13	0.67**					
Model 2:				0.464	0.014	1	72	1.842
• Predicted Score	0.98	0.13	0.65**					
• Gender	-0.15	0.11	-0.12					
Model 3:				0.488	0.023	1	71	3.256
• Predicted Score	0.71	0.20	0.47**					
• Gender	-1.55	0.78	-1.21					
• Predicted Score * Gender	0.47	0.26	1.09					

Lasso Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.464	-	1	73	63.081**
• Predicted Score	1.01	0.13	.68**					
Model 2:				0.476	0.012	1	72	1.682
• Predicted Score	0.99	0.13	0.88**					
• Gender	-0.14	0.11	-0.11					
Model 3:				0.501	0.025	1	71	3.542
• Predicted Score	0.71	0.19	0.48**					
• Gender	-1.56	0.76	-1.21*					
• Predicted Score * Gender	0.48	0.26	1.10					
Elastic Net Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.453	-	1	73	60.509**
• Predicted Score	0.97	0.12	0.67**					
Model 2:				0.467	0..014	1	72	1.834
• Predicted Score	0.94	0.13	0.66**					
• Gender	-0.15	0.11	-0.02					
Model 3:				0.491	0.024	1	71	3.368
• Predicted Score	0.68	0.19	0.47**					
• Gender	-1.51	0.75	-1.17*					
• Predicted Score * Gender	0.46	0.25	1.05					

Table 15. Main & Interaction Effects of Repeated Measures ANOVA on Predicted Criterion Scores

	DF	SS	MS	F-Value	P-Value	Eta-Square
Estimation Method	6, 438	0.16	0.03	2.27	<0.05	0.03
Race	1, 73	0.04	0.04	0.04	0.85	0.00
Estimation Method x Race	6, 438	`0.02	0.00	0.26	0.96	0.00

Table 16. Predicted Criterion Means & Cohen's D-Statistic

	Regression	Decision Tree	Random Forests	Extreme Gradient Boosting	Ridge	Lasso	Elastic Net
Overall	2.96	2.96	3.00	2.96	2.96	2.95	2.95
Racial Majorities	2.97	2.99	3.03	2.99	2.97	2.96	2.97
Racial Minorities	2.96	2.95	3.00	2.95	2.96	2.95	2.96
Cohen's d-statistic	0.02 (-0.44, 0.48)	0.10 (-0.46, 0.56)	0.06 (-0.40, 0.41)	0.08 (-0.38, 0.54)	0.02 (-0.44, 0.47)	0.02 (-0.44, 0.48)	0.02 (-0.44, 0.48)

Table 17. Cleary Model Test Bias

Regression	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.450	-	1	73	59.81**
• Predicted Score	0.96	0.12	0.67**					
Model 2:				0.452	0.002	1	72	0.24
• Predicted Score	0.96	0.13	0.67**					
• Race	0.06	0.11	0.04					
Model 3:				0.460	0.008	1	71	1.09
• Predicted Score	1.07	0.16	0.75**					
• Race	0.84	0.76	0.65					
• Predicted Score * Race	-0.27	0.25	-0.62					
Decision Tree	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.506	-	1	73	74.85**
• Predicted Score	1.26	0.15	0.71**					
Model 2:				0.512	0.005	1	72	0.79
• Predicted Score	1.27	0.15	0.72**					
• Race	0.10	0.11	0.07					
Model 3:				0.572	0.060	1	71	9.95**
• Predicted Score	1.69	0.19	0.96**					
• Race	2.68	0.83	2.07**					
• Predicted Score * Race	-0.87	0.28	-2.01**					
Random Forests	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.406	-	1	73	49.84**
• Predicted Score	0.96	0.14	0.64**					
Model 2:				0.409	0.003	1	72	0.36
• Predicted Score	0.96	0.14	0.64**					
• Race	0.07	0.12	0.05					
Model 3:				0.444	0.036	1	71	4.54*
• Predicted Score	1.20	0.18	0.80**					
• Race	1.80	0.82	1.39*					
• Predicted Score * Race	-0.57	0.27	-1.36*					
Extreme Gradient Boosting	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.478	-	1	73	66.94**
• Predicted Score	1.08	0.13	0.69**					
Model 2:				0.482	0.004	1	72	0.57
• Predicted Score	1.08	0.13	0.68**					
• Race	0.08	0.11	0.06					
Model 3:				0.529	0.046	1	71	6.94**
• Predicted Score	1.40	0.18	0.90**					
• Race	2.08	0.76	1.61**					
• Predicted Score * Race	-0.67	0.26	-1.56**					
Ridge Regularization	**B**	**SE**	**β**	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.451	-	1	73	59.86**
• Predicted Score	1.01	0.13	0.67**					
Model 2:				0.452	0.002	1	72	0.23
• Predicted Score	1.01	0.13	0.67**					
• Race	0.05	0.11	0.04					
Model 3:				0.460	0.008	1	71	1.00
• Predicted Score	1.12	0.17	0.74**					
• Race	0.85	0.80	0.65					
• Predicted Score * Race	-0.27	0.27	-0.62					

Lasso Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.464	-	1	73	63.08
• Predicted Score	1.01	0.13	0.68**					
Model 2:				0.465	0.002	1	72	0.26
• Predicted Score	1.01	0.13	0.68**					
• Race	0.06	0.11	0.44					
Model 3:				0.475	0.010	1	71	1.29
• Predicted Score	1.13	0.17	0.76**					
• Race	0.93	0.78	0.72					
• Predicted Score * Race	-0.30	0.26	-0.69					
Elastic Net Regularization	**B**	**SE**	***β***	**R^2**	**ΔR^2**	**df_1**	**df_2**	**F Change**
Model 1:				0.453	-	1	73	60.51**
• Predicted Score	0.97	0.12	0.67**					
Model 2:				0.455	0.002	1	72	0.24
• Predicted Score	0.97	0.13	0.67**					
• Race	0.06	0.11	0.04					
Model 3:				0.456	0.001	1	71	0.10
• Predicted Score	0.97	0.13	0.67**					
• Race	0.03	0.14	0.02					
• Predicted Score * Race	0.02	0.06	0.03					

Table 18. Unstandardized Beta Coefficients for Regression & Ridge Regularization

	Regression	Ridge Regularization
Number of Special Projects	-0.03	-0.03
Number of Times Employee Late to Work	-0.27**	-0.24**
Absences	0.02*	0.02*
Employee Engagement	0.16	0.17*
Employee Satisfaction	-0.02	-0.01

Figure 1. Slope, Intercept and Slope & Intercept Bias (Furr & Bacharach, 2014)

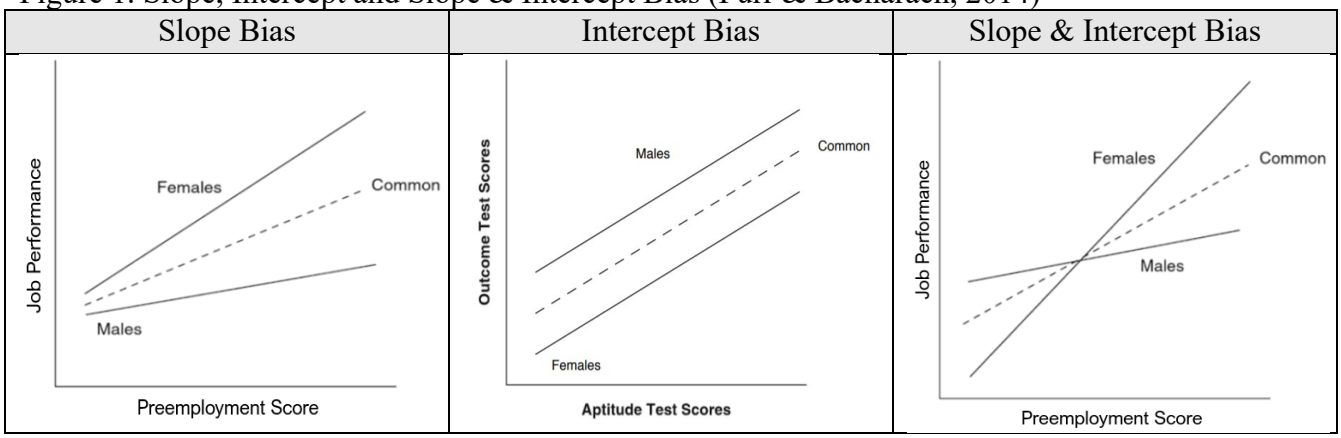


Figure 2. Example Decision Tree for Job Performance Ratings (Molnar, 2021)

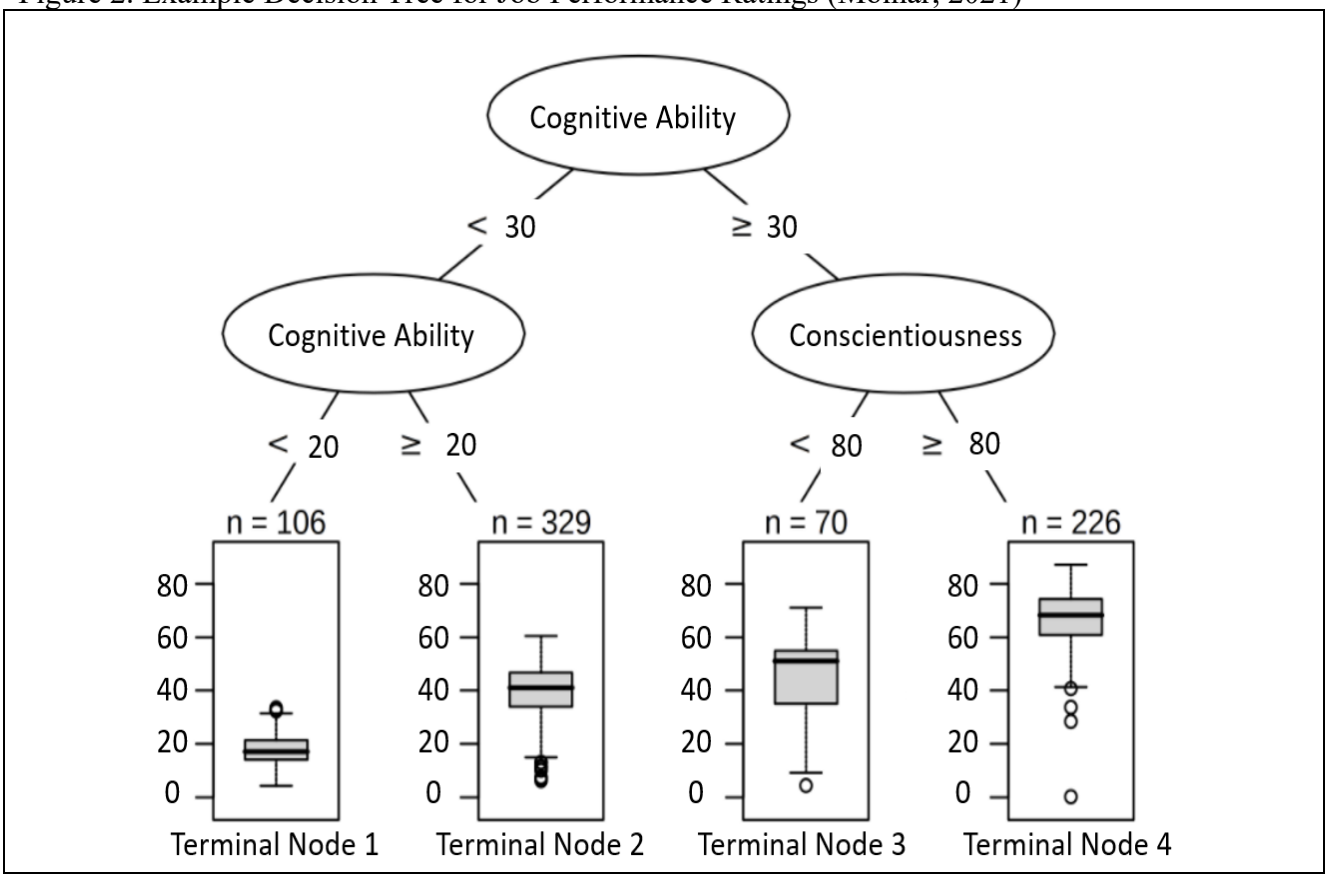


Figure 3. Illustration of a Variable Importance Plot.

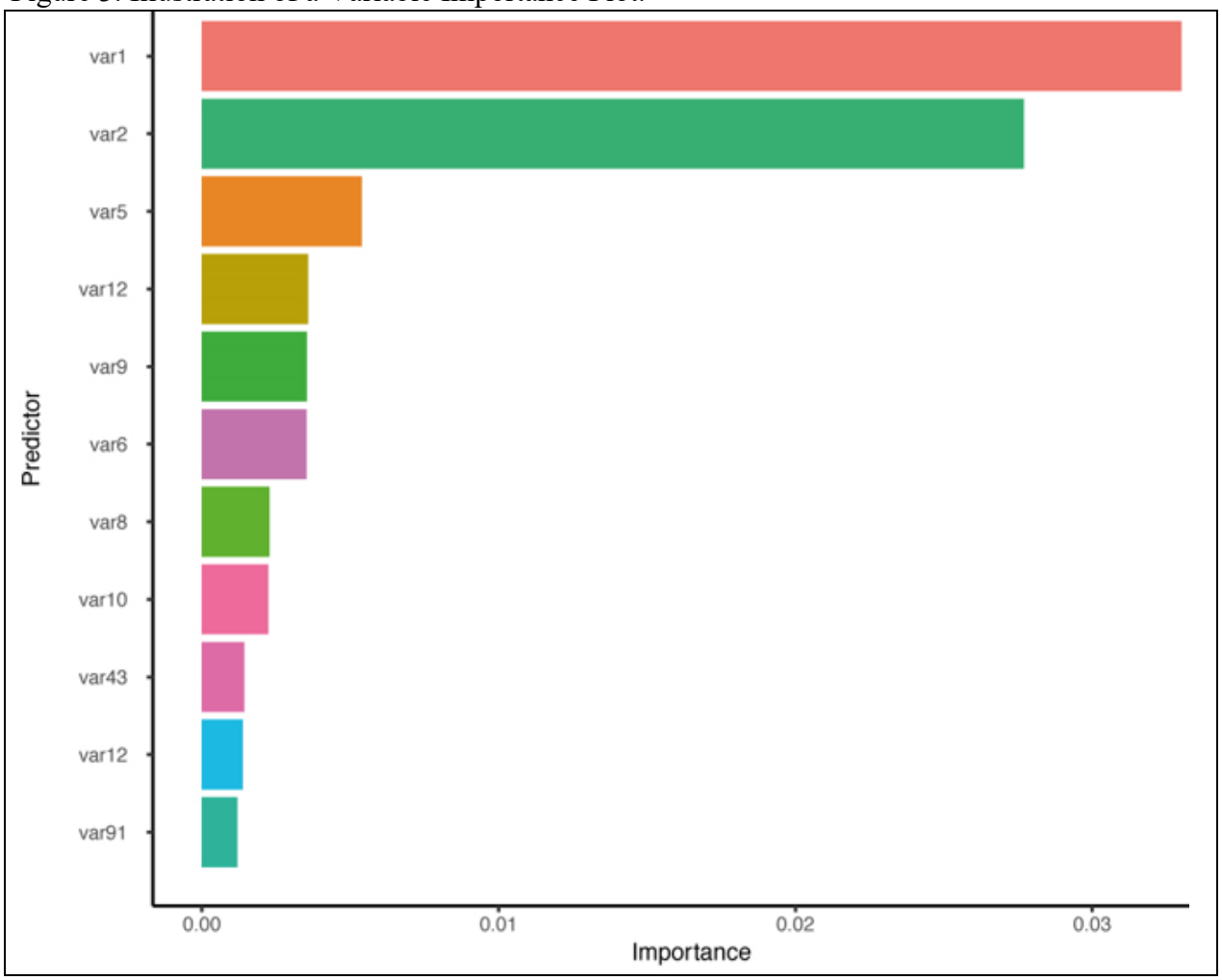


Figure 4. Partial Dependence Plot

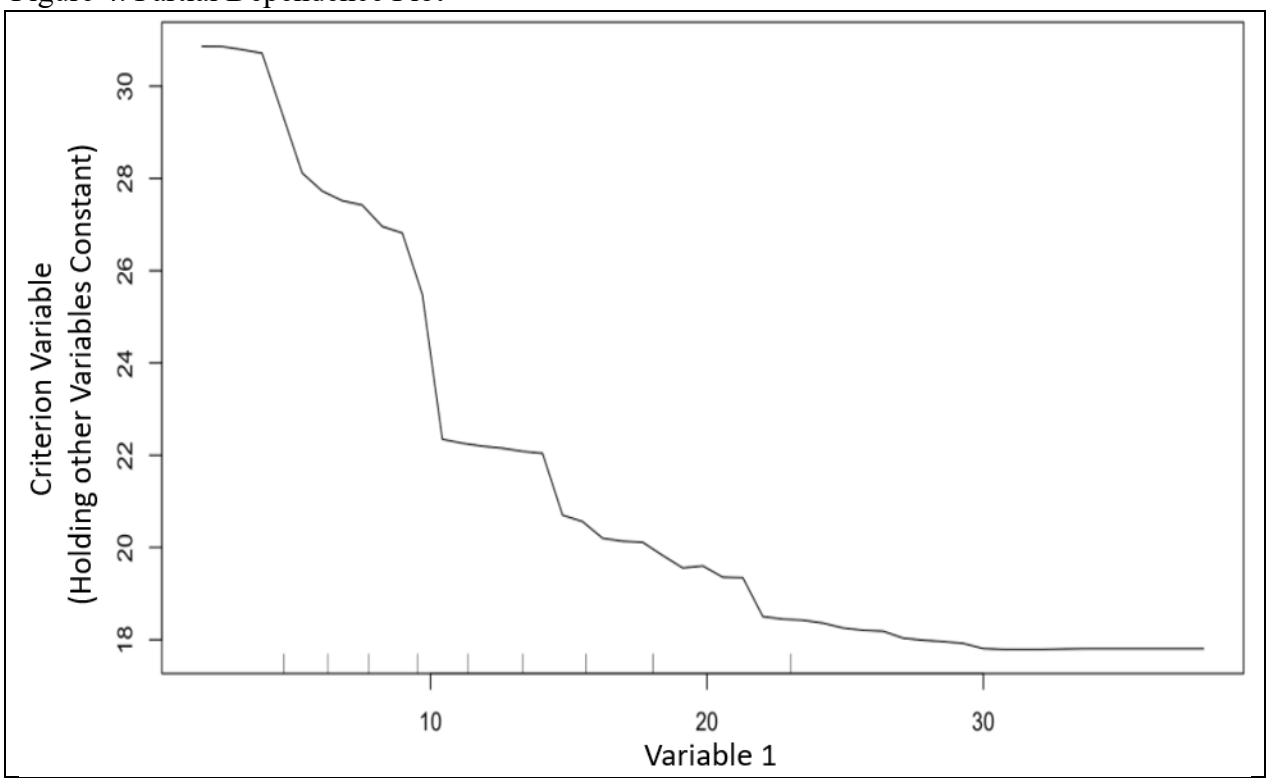


Figure 5. Variable Importance Analysis for Random Forests.

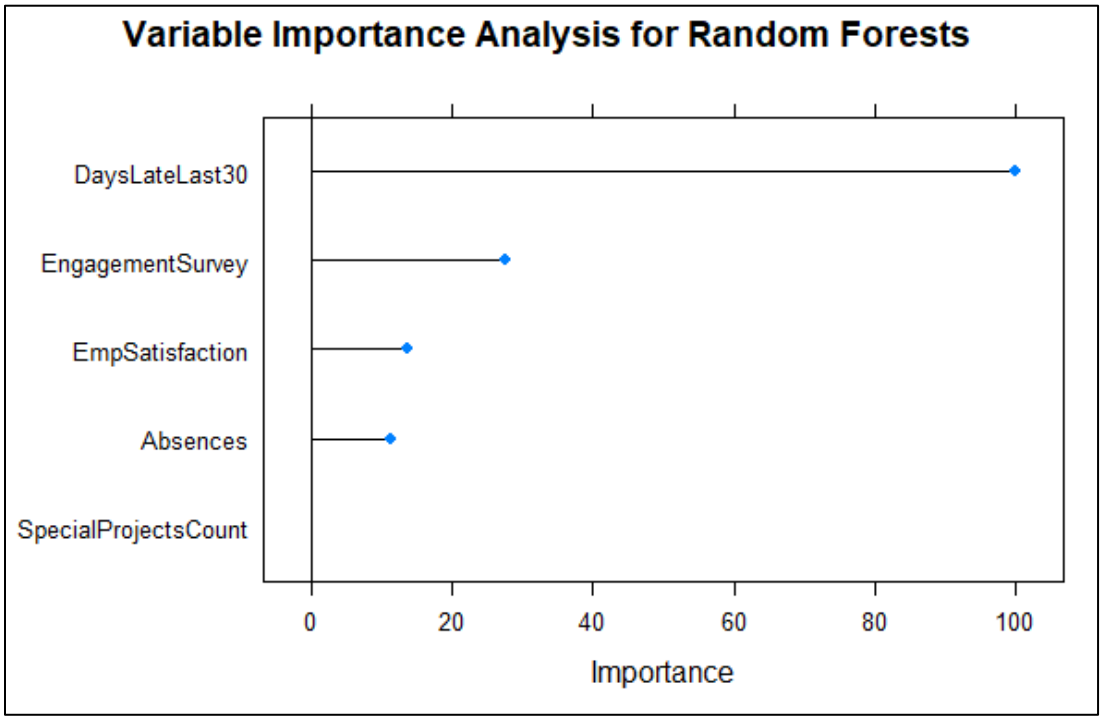


Figure 6. PD Plot for Number of Dates Late in Random Forests Estimation.

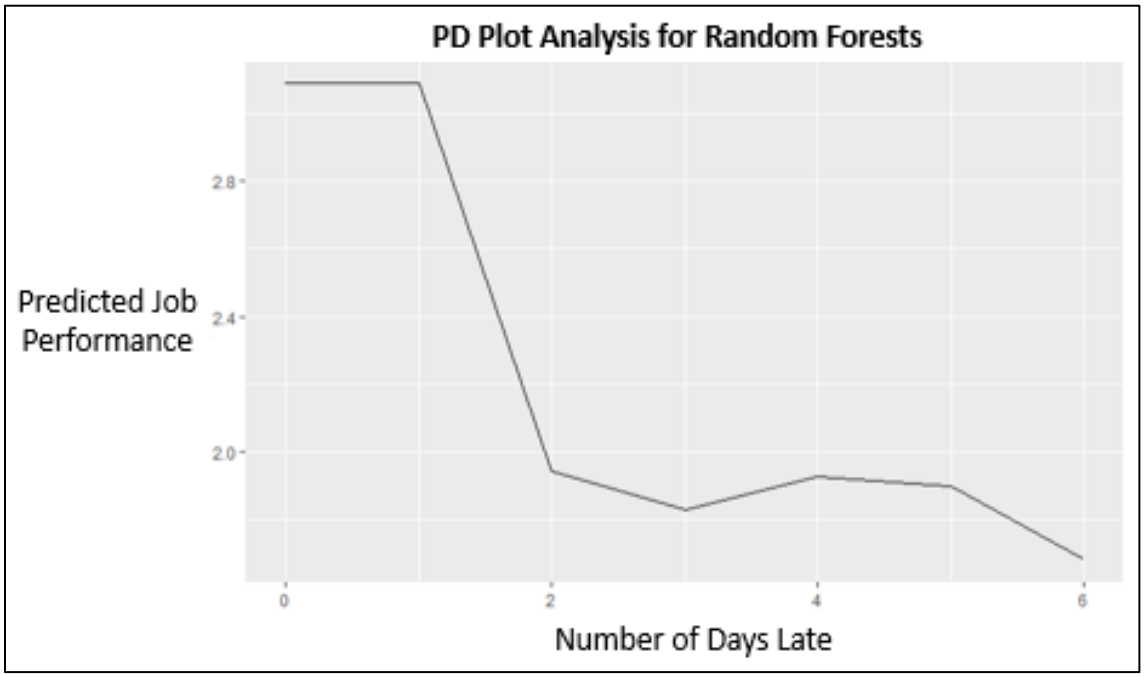


Figure 7. PD Plot for Employee Engagement in Random Forests Estimation.

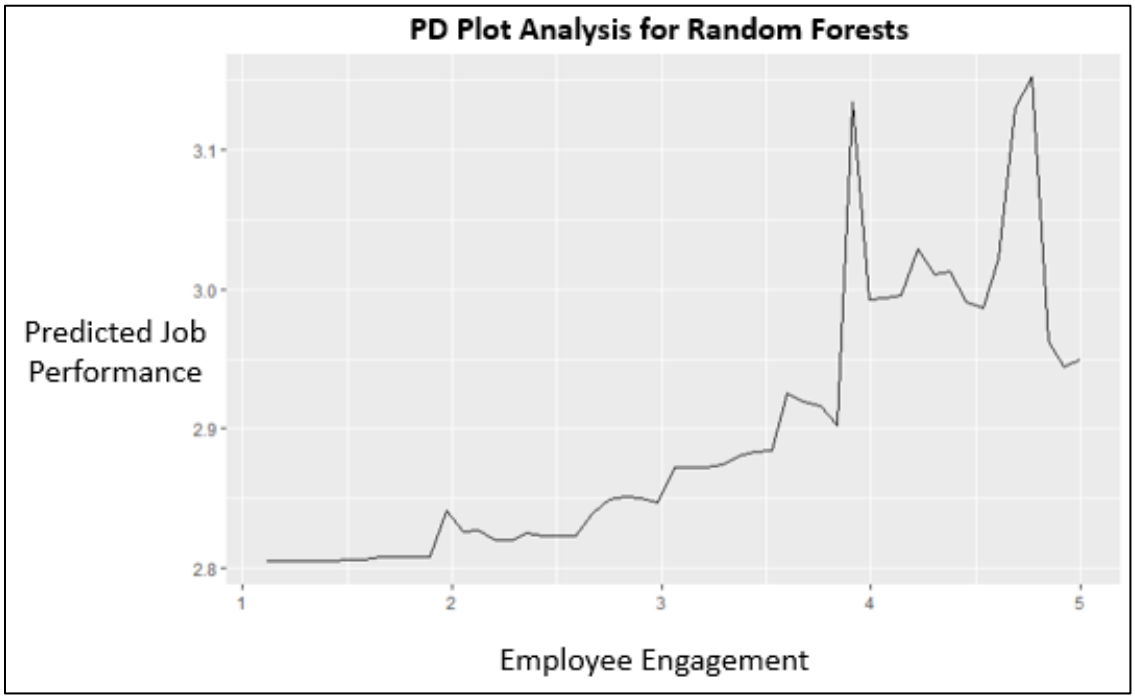


Figure 8. ICE Plot for Number of Dates Late in Random Forests Estimation.

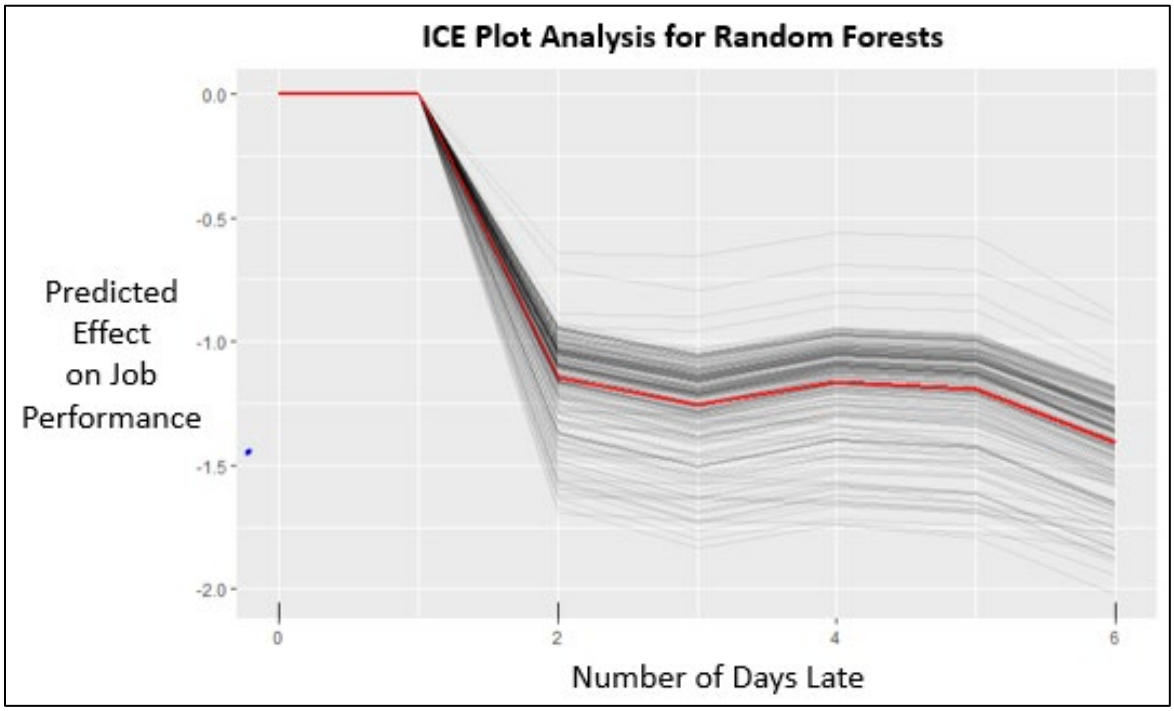
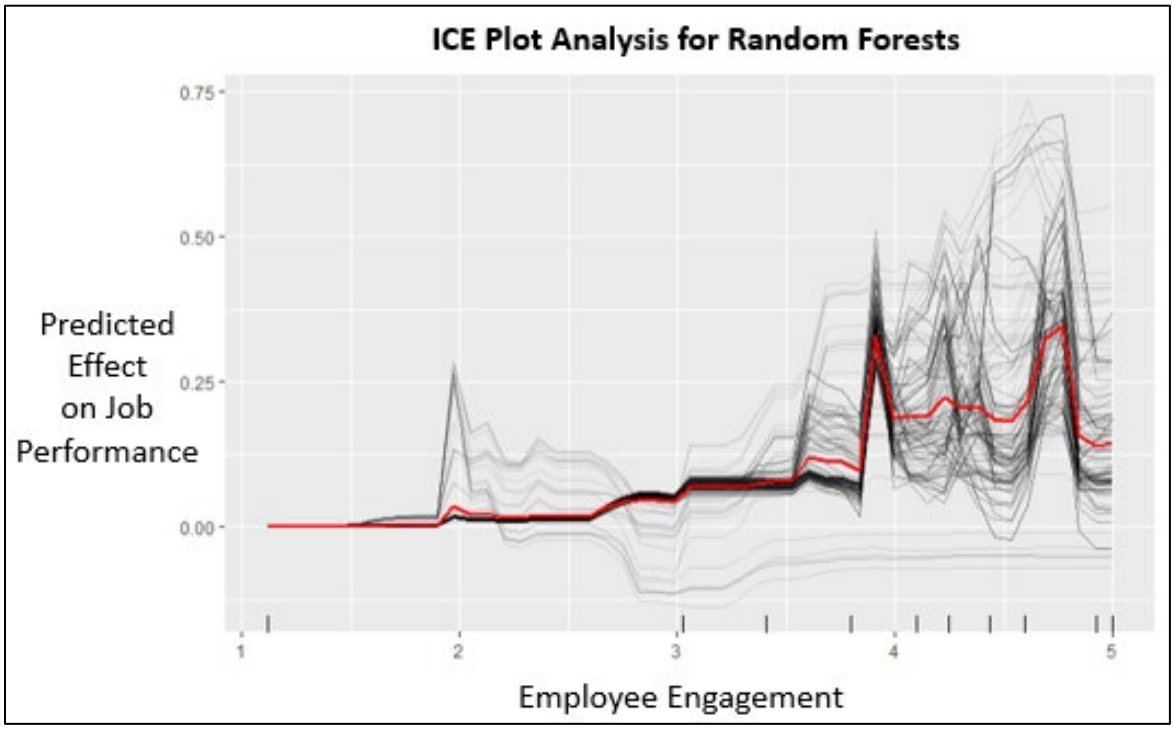


Figure 9. ICE Plot for Employee Engagement in Random Forests Estimation



Bibliography

Adams v. City of Chicago, 135 F.3d 1150 (1998).

Aguinis H, Culpepper S, Pierce C. (2010). Revival of test bias research in preemployment testing. *J Appl Psychol*. 2010 Jul;95(4):648-80. doi: 10.1037/a0018714. PMID: 20604587.

Aguinis, H., Culpepper, S., & Pierce, C. (2016). Differential Prediction Generalization in College Admissions Testing. *Journal of Educational Psychology*, 108, 1045-1059.

Aiken, J., Salmon, E.D. & Hanges, P.J. The Origins and Legacy of the Civil Rights Act of 1964. *J Bus Psychol* 28, 383–399 (2013). https://doi.org/10.1007/s10869-013-9291-z

Aiken, J. & Hanges, P. (2017). The Sum of the Parts: Methods of Combining Assessments for Employment Decisions.

Ajunwa, I., Friedler, S.A., Scheidegger, C., & Venkatasubramanian, S. (2016). Hiring by Algorithm: Predicting and Preventing Disparate Impact.

Allen, K. S., Affourtit, M., & Reddock, C. M. (202“) "The Machines Aren’t Taking Over (Yet): An Empirical Comparison of Traditional, Profiling, and Machine Learning Approaches to Criterion-Related Validati”n," *Personnel Assessment and Decisions*: Vol. 6 : Iss. 3 , Article 2. DOI: https://doi.org/10.25035/pad.2020.03.002

Arkin, William & Dobrofsky, Lynne. (2010). Military Socialization and Masculinity. *Journal of Social Issues*. 34. 1–1 - 168. 10.1111/j.1540-4560.1978.tb02546.x.

Barocas. S. & Hardt, M. & Narayanan, A. (2021). *Fairness and Machine Learning.* fairmlbook.org.

Barrett, Liam. (2021). Using Artificial Intelligence to Reimagine Enforcement of Workplace Discrimination Laws. *Georgetown Journal of Poverty Law & Policy.*

https://www.law.georgetown.edu/poverty-journal/blog/using-artificial-intelligence-to-reimagine-enforcement-of-workplace-discrimination-laws/#_ftn7

https://www.law.georgetown.edu/poverty-journal/blog/using-artificial-intelligence-to-reimagine-enforcement-of-workplace-discrimination-laws/#_ftn7

Boardman, A. E. (1979). Another analysis of the EE‘C 'four fifths' rule. *Management Science*, 25, 770-776.

Bradley v. City of Lynn, 443 F.Supp.2d 145 (2006).

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, 2, 171-183.

Bzdok, D., Altman, N. & Krzywinski, M. (2018). Statistics versus Machine Learning. *Nature Methods*. 15. 10.1038/nmeth.4642.

Campbell, W. K, Bonacci, A. M, Shelton, J., Exline, J. J & Bushman, B. J. (2010). Psychological entitlement: Interpersonal consequences and validation of a self-report measure. *Journal of Personality Assessment, 83*(1), 29-45.

Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.

Chen, T. & He, T. (2021). xgboost: eXtreme Gradient Boosting. Package Version: 1.3.2.1

Civil Rights Act of 1964, 7, 42 U.S.C., 2000e et seq (1964).

Civil Rights Act of 1991, 109, 42 U.S.C.. 2000e et seq (1991).

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Lawrence Erlbaum Associates Publishers.

Cormier v. P.P.G. Industries Inc., 702 F.2d 567 (1983).

Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods*. 2015 Jul 30;250:126-36. doi: 10.1016/j.jneumeth.2015.01.010. Epub 2015 Jan 14. PMID: 25596422.

Connecticut v. Teal, 457 U.S. 440 (1982).

Dalal, R. (2005). A Meta-Analysis of the Relationship Between Organizational Citizenship Behavior and Counterproductive Work Behavior. *The Journal of applied psychology.* 90. 1241-55. 10.1037/0021-9010.90.6.1241.

Dastin, J., Amazon Scraps Secret Recruiting Tool That Showed Bias Against Women. Reuters (Oct. 10, 2018), available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Detroit Police Officers' Association v. Young, 608 F.2d 671 (1979).

Dothard v. Rawlinson, 433 U.S. 321 (1977).

Dunn, A.M., Heggestad, E.D., Shanock, L.R. *et al.* Intra-individual Response Variability as an Indicator of Insufficient Effort Responding: Comparison to Other Indicators and Relationships with Individual Differences. *J Bus Psychol* **33,** 105–121 (2018). https://doi.org/10.1007/s10869-016-9479-0

EEOC v. Dial Corp., 469 F.3d 735 (2006).

EEOC v. FAPs Inc., Civil No. 10-3095 (JAP)(DEA).

EEOC v. Ford Motor Company, 1:04-cv-00845-SAS (2008).

Embretson, S. & Reise, S. (2000). Item Response Theory For Psychologists.

Equal Employment Opportunity Commission (EEOC). (2021). *Facts about Discrimination in Federal Government Employment Based on Marital Status, Political Affiliation, Status as a Parent, Sexual Orientation, and Gender Identity* | U.S. Equal Employment Opportunity Commission. U.S. Equal Employment Opportunity Commission. https://www.eeoc.gov/federal-sector/facts-about-discrimination-federal-government-employment-based-marital-status

Finkelstein, M. O., & Levin, B. A. (2016). *Statistics for lawyers*. New York: Springer.

Fisher v. University of Texas at Austin, 133 S.Ct. 2411 (2013).

Fisher v. University of Texas at Austin, 136 S.Ct. 2198 (2016).

Foster, I., Ghani, R., Jarmin, R., Kreuter, F. & Lane, J. (2020). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. 10.1201/9780429324383.

Fox, S., Spector, P. & Miles, D. (2001). Counterproductive Work Behavior (CWB) in Response to Job Stressors and Organizational Justice. *Journal of Vocational Behavior*. 59. 291-309. 10.1006/jvbe.2001.1803.

Furr, R. & Bacharach, Verne. (2008). Psychometrics: An Introduction.

Gastwirth, J. L. (1988). Statistical Reasoning in Law and Public Policy. Vol. 1. Statistical Concepts and Issues of Fairness. 486 Seiten, zahlr. Abb. und Tab. Academic Press, Inc., Boston, San Diego, New York u. a. 1988. Preis: 56,— £.

Gastwirth, J. L. & Miao, W. (2009). Formal statistical analysis of the data in disparate impact cases provides sounder inferences than the U. S. government's ‘four-fifths’ rule: an

examination of the statistical evidence in *Ricci vs. DeStefano, Law, Probability and Risk,* Volume 8, Issue 2, June 2009, Pages 171–191, https://doi.org/10.1093/lpr/mgp017

Gillespie v. State of Wisconsin, 771 F.2d 1035 (1985).

Greenlaw, P. S., & Jensen, S. S. (1996). Race-Norming and the Civil Rights Act of 1991. *Public Personnel Management*, *25*(1), 13–24. https://doi.org/10.1177/009102609602500102

Griggs v. Duke Power Co., 292 F.Supp. 243 (1968).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Grutter v. Bollinger, 539 U.S. 306 (2003).

Guardians v. Civil Service, 630 F.2d 79 (1980).

Guardians Assn. v. Civil Serv. Comm'n of New York City, 463 U.S. 582 (1983).

Gulino v. New York State Education Department, 460 F.3d 361 (2006).

Gutman, A., Outtz, J. L., Dunleavy, E. (2017). An updated sampler of legal principles in employment selection. In Farr, J. L. & Tippins, N. T. (Eds.), *Handbook of employee selection* (2nd ed., pp. 631–658). New York, NY: Routledge.CrossRef.

Hamer v. City of Atlanta, 872 F.2d 1521 (1989).

Harwell, D. (2019). A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job, The Washington Post, available at https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/?wpisrc=al_trending_now__alert-economy–alert-national&wpmk=1

Harwell, D. (2019). Rights Group Files Federal Complaint Against AI-Hiring Firm HireVue, Citing "Unfair and Deceptive Practices," The Washington Post, available at

https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

Hanges, P. J., & Gettman, H. J. (2004). *A comparison of test-focused and criterion-focused banding methods: Back to the future?* In H. Aguinis (Ed.), *Test-score banding in human resource selection: Technical, legal, and societal issues* (p. 29–48). Praeger Publishers/Greenwood Publishing Group.

Hanges, P.J., Lucas, J., Baxter, A., DeAngelis, K., Dobbs, J., McCone, D., Norton, M., Woodruff, T. Beavan, K., & Epistola, J. (2021). Organizational culture, ethical leadership, and trust. Report provided to US Army Research Institute for the Behavioral and Social Sciences.

Hanges, P. J., Salmon, E. D., & Aiken, J. R. (2013). Legal issues in industrial testing and assessment. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (vol. 1, pp. 693–711). Washington, DC: American Psychological Association.

Hastie, T. (2020) Ridge Regularization: An Essential Concept in Data Science, *Technometrics*, 62:4, 426-433, DOI: 10.1080/00401706.2020.1791959

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer.

Hayden v. Nassau County, 180 F.3d 42 (1999).

Hinojosa, R. (2010). Doing Hegemony: Military, Men, and Constructing a Hegemonic Masculinity. *The Journal of Men's Studies*, 18(2), 179–194. https://doi.org/10.3149/jms.1802.179

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and

measuring preferences for intergroup inequality using the new SDO7 scale. *Journal of Personality and Social Psychology, 109*(6), 1003-1028.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of Detecting Insufficient Effort Responding: Comparisons and Practical Recommendations. Educational and Psychological Measurement, 80(2), 312–345. https://doi.org/10.1177/0013164419865316

Hsu, K, & Iwamoto, D. K. (2014). Testing for Measurement Invariance in the Conformity to Masculine Norms-46 across White and Asian American College Men: Development and Validity of the CMNI-29. *Psychology of Men & Masculinity*, 15, 4, 397–406. doi:10.1037/a0034548.

Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, 6:1, 1-55, DOI: 10.1080/10705519909540118

Hunter, J. E. (1986). Cognitive ability, cognitive aptitude, job knowledge, and job performance. *Journal of Vocational Behavior, 29*(3), 340–362. https://doi.org/10.1016/0001-8791(86)90013-8

Isabel v. City of Memphis, 404 F.3d 404 (2005).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (1st ed. 2013, Corr. 7th printing 2017 ed.). Springer.

Kazemitabar, J., Amini, A., Bloniarz, A., & Talwalkar, A. (2017). Variable Importance using Decision Trees, Supplementary Material for NIPS 2017 paper.

Koenig, N. & Thompson, I. (2021). *SIOP Machine Learning Competition* [Conference presentation]. SIOP 2021 Convention, New Orleans, LA, United States.

Landon, Timothy & Arvey, Richard. (2007). Ratings of Test Fairness by Human Resource Professionals. *International Journal of Selection and Assessment*. 15. 10.1111/j.1468-2389.2007.00380.x.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183.

Lee, A., Inceoglu, I., Hauser, O., & Greene, M. (2020). Determining causal relationships in leadership research using machine learning: The powerful synergy of experiments and Data Science. The Leadership Quarterly, 101426. https://doi.org/10.1016/j.leaqua.2020.101426

Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

Mahalik, J. R., Locke, B. D., Ludlow, L. H., Diemer, M. A., Scott, R. P. J., Gottfried, M., & Freitas, G. (2003). Development of the Conformity to Masculine Norms Inventory. *Psychology of Men & Masculinity, 4*(1), 3–25. https://doi.org/10.1037/1524-9220.4.1.3

Marrs, Heath. (2013). Conformity to Masculine Norms and Intellectual Engagement. *Masculinities and Social Change*. 2. 226-244. 10.4471/MCS.2013.33.

McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973).

Meacham v. Knolls Atomic Power Laboratory, 461 F.3d 134 (2006).

Meier, P., Sacks, J., & Zabell, S. (1984). What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule. *American Bar Foundation Research Journal*, 9(1), 139-186. Retrieved April 9, 2021, from http://www.jstor.org/stable/828307

Miao, W. & Gastwirth, Joseph. (2013). Properties of statistical tests appropriate for the analysis of data in disparate impact cases. *Law, Probability and Risk*. 12. 37-61. 10.1093/lpr/mgs032.

Millsap, R.E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30, 577–605.

Molnar, C. (2021). Interpretable machine learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

Moody v. Albemarle Paper Co., 417 U.S. 622 (1974).

Motowidlo, S.J. & Kell, Harrison. (2013). Job performance. *Handbook of Psychology*, 12: Industrial and Organizational Psychology. 82-103.

Murphy, K., & Jacobs, R. (2012). Using effect size measures to reform the determination of adverse impact in equal employment litigation. Psychology, Public Policy and Law, 18, 477-499.

Newby, J. L., & Klein, R. G. (2014). Competitiveness reconceptualized: Psychometric development of the competitiveness orientation measure as a unified measure of trait competitiveness. *The Psychological Record, 64*(4), 879-895.

Oswald, F., Behrend, T., Putka, D., & Sinar, E. (2020). Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for Organizational Research and Practice. *Annual Review of Organizational Psychology and Organizational Behavior*. 7. 1-29. 10.1146/annurev-orgpsych-032117-104553.

Penner, L. A., Fritzsche, B. A., Craiger, J. P., & Freifeld, T. S. (1995). Measuring the prosocial personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment, Vol. 10* (pp. 147-163). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Petit v. Chicago, 352 F.3d 1111 (2003).

PGA Tour, Inc. v. Martin, 532 U.S. 661 (2001).

Ployhart, R. E. (2012). *Personnel selection and the competitive advantage of firms.* In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology: Vol. 27. International review of industrial and organizational psychology 2012* (p. 153–195). Wiley Blackwell.

Police Officers v. City of Columbus, 916 F.2d 1092 (1990).

Ranstam, J. & Cook, A. J. (2018). LASSO regression, *British Journal of Surgery*, Volume 105, Issue 10, September 2018, Page 1348, https://doi.org/10.1002/bjs.10895

Regents of the University of California v. Bakke, 438 U.S. 265 (1978).

Ricci v. DeStefano, 557 U.S. 557 (2009).

Riccucci, N. M. & Saldivar, K. (2012). The status of employment discrimination suits in police and fire departments across the United States. *Review of Public Personnel Administration*. DOI: 10.1177/0734371X12449839

Rogers, J. (2003). *Midshipmen military performance as an indicator of officer fleet performance*. Master's Thesis, Naval Postgraduate School. Calhoun: Institutional Archive of the Naval Postgraduate School.

Sackett, P. R. (2002). The structure of counterproductive work behaviors: Dimensionality and relationships with facets of job performance. *International Journal of Selection and Assessment*, 10, 5–11.

Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist, 29*(1), 1–8. https://doi.org/10.1037/h0035844

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, *64*(6), 609.

Section 6A, Uniform Guidelines on Employee Selection Procedures (1978); 43 FR __, (August 25, 1978); 29 CFR part 1607, section 6A.

SFFA v. President and Fellows of Harvard College, 980 F.3d 157 (2020).

Smith v. City of Boston, 144 F.Supp.3d 177 (2015).

Society for Industrial-Organizational Psychology (SIOP). (2015, December 16). *SIOP Announces Top 10 Workplace Trends for 2016.* https://www.siop.org/Research-Publications/Items-of-Interest/ArtMID/19366/ArticleID/1724/SIOP-Announces-Top-10-Workplace-Trends-for-2016

Society for Industrial-Organizational Psychology (SIOP). (2019, January 9). *It's the Same, Only Different: Top 10 Workplace Trends*. https://www.siop.org/Research-Publications/Items-of-Interest/ArtMID/19366/ArticleID/1639/It%E2%80%99s-the-Same-Only-Different?utm_source=Web&utm_medium=Article&utm_campaign=Top10

Society of Industrial-Organizational Psychology (SIOP). (2020, February 13). *Top 10 Workplace Trends for 2020*. Society of Industrial-Organizational Psychology (SIOP). https://www.siop.org/Research-Publications/Items-of-Interest/ArtMID/19366/ArticleID/3361/Top-10-Workplace-Trends-for-2020

Spector, P., Fox, S., Penney, L., Bruursema, K., Goh, A. & Kessler, S. R. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal?.*Journal of Vocational Behavior*. 68. 446-460. 10.1016/j.jvb.2005.10.005. https://scholarcommons.usf.edu/psy_facpub/710

Students for Fair Admission vs. University of North Carolina at Chapel Hill, No. 14-954 (M.D.N.C., 2021).

"Student" Gosset, W. G. (1908). "The probable error of a mean". *Biometrika*. 6 (1): 1–25. doi:10.1093/biomet/6.1.1. hdl:10338.dmlcz/143545.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education,* 4(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Tippins, N., Oswald, F. & McPhail, S. M. (2021). *Scientific, Legal, and Ethical Concerns about AI-Based Personnel Selection Tools: A Call to Action.* Manuscript submitted for publication.

Tippins, N. & Sackett, P. & Oswald, F. (2018). Principles for the Validation and Use of Personnel Selection Procedures. Industrial and Organizational Psychology. 11. 1-97. 10.1017/iop.2018.195.

US v. City of Erie, PA, 411 F.Supp.2d 524 (2005).

Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14(11): e0224365. https://doi.org/10.1371/journal.pone.0224365

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Van Iddekinge, C. H., Aguinis, H., Mackey, J. D., & DeOrtentiis, P. S. (2018). A Meta-Analysis of the Interactive, Additive, and Relative Effects of Cognitive Ability and Motivation on

Performance. *Journal of Management*, 44(1), 249–279. https://doi.org/10.1177/0149206317702220

Waisome v. Port Auth. of New York & New Jersey, 948 F.2d 1370 (1991).

Warton, D. (2008) Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices, Journal of the American Statistical Association, 103:481, 340-349, DOI: 10.1198/016214508000000021

Watson v. Fort Worth Bank & Trust, 487 U.S. 977 (1988).

Williams v. Ford Motor Co., 187 F.3d 533 (1999).

Xu, H., Caramanis, C., & Mannor, S. (2010). Robust Regression and Lasso. Information Theory, IEEE Transactions on. 56. 3561 - 3574. 10.1109/TIT.2010.2048503.

Yuvraj, J. (2019). "Racial Indirection". UC Davis Law Review. 52 (5): 74. SSRN 3312518.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 301-320. Retrieved April 6, 2021, from http://www.jstor.org/stable/3647580