THE INTRODUCTION AND APPLICATION OF RECURSIVE PARTITIONING METHODS IN ORGANIZATIONAL SCIENCE

BY

JING JIN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor James Rounds, Chair
Professor Fritz Drasgow, Director of Research
Professor Lawrence J. Hubert
Professor Hua-Hua Chang
Associate Professor Daniel A. Newman

ABSTRACT

Traditionally, multiple linear regression has been widely used in the field of organizational science for predictive modeling. Despite its pervasive use, the classical regression model falls short in several aspects, including the lack of flexibility in handling complex nonlinear relationships and the strict assumptions imposed by parametric approaches. To overcome these limitations, the current study examined an alternative, nonparametric recursive partitioning method – Classification and Regression Trees (CART), and its advanced successor, random forests. Results from two Monte Carlo simulations (Study 1 and 2) showed that random forests consistently produced comparable predictive accuracy as the traditional methods when the data was structured in a linear or simple additive model, yet exhibited substantially more accurate results when the data was structured in a complex nonlinear manner. CART outperformed the traditional methods for evaluating model fit (i.e., resubstituition accuracy), but was not as effective when generalizability was evaluated, except when the data was structured in a nonlinear tree-like pattern. Two empirical studies were also conducted to illustrate the application of the two recursive partitioning methods for predicting employee turnover (Study 3) and job performance (Study 4). Practical guidance is provided regarding how the feature selection procedure of random forests and a single decision tree constructed by CART could be combined to explore complex relationships within the data and better facilitate model interpretation. Limitations and implications for future research are also discussed.

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

For several decades, multiple regression has been one of the most popular data analytic techniques in the field of organizational science (LeBreton, Tonidandel, & Krasikova, 2013). Researchers and practitioners have generally adopted the basic linear additive model to both make predictions and understand the functional relationship between the predictor variables and the criterion. Nevertheless, as theory and empirical evidence evolve, a growing number of researchers have begun to challenge the basic assumptions of multiple linear regression and have noticed that the complexities of the real world may go beyond simple linear and additive effects. There has been an emerging awareness of the existence of nonlinear or curvilinear relationships in psychological and organizational research. Theoretically, Pierce and Aguinis (2013) have proposed a meta-theoretical principle, which they called the "too-much-of-a-good-thing effect" to account for a variety of seemingly paradoxical results in organizational research. They contended that positive and monotonic relationships may become asymptotic or even negative after a certain inflection point, resulting in a curvilinear pattern. Grant and Schwartz (2011) have reviewed the effects of several psychological constructs on well-being and performance. Similarly, they have concluded that "there is no such thing as an unmitigated good" (p. 62) as positive phenomena become negative after reaching inflection points. As such, they called for more attention to advance our knowledge regarding nonmonotonic inverted-U-shaped effects in psychological research. Empirically, many studies have tested and supported a significant nonlinear/curvilinear link between personality traits and important work outcomes (e.g., Cucina & Vasilopoulos, 2005; Day & Silverman, 1989; LaHuis, Martin, & Avis, 2005; Le, Robbins, Ilies, Holland, & Westrick, 2011; Robie & Ryan, 1999; Vasilopoulos, Cucina, & Hunter, 2007; Whetzel, McDaniel, Yost, & Kim, 2010). In addition to the nonlinear higher-order effects of any

individual variable, multiple variables can also interact to affect the outcome. As noted by LeBreton, Tonidandel, and Krasikova (2013), "the world of psychology is not always a simple, main-effects world" (p. 3). When the effect of an individual variable is examined without being put into the context of possible covariates, the results can be biased.

In response to these significant limitations of simple linear additive regression, there have been some methodological developments in the field. For example, Anderson (1970) and Lynch (1985) argued that both multiplicative and multilinear models are legitimate ways to combine predictors. Methods such as polynomial regression (Edwards & Parry, 1993) and hierarchical multiple regression (Cortina, 1993) have been developed to explicitly incorporate appropriate power terms and multiplicative interactions into the linear regression model. However, the form of the model including interactive and higher-order components must be specified in advance, involving a significant numbers of assumptions and guesswork. Moreover, the complexities of the real data may go beyond the most sophisticated model that could be specified. Thus, a model determined a priori may fall short in detecting complex relationships. Another significant limitation of these classical regression methods is that they are bounded by the basic assumptions of a parametric approach such as normal distributions or residual independence. Given the lack of knowledge of the underlying data distribution, these assumptions may be considered too strict or even unrealistic by organizational scholars.

More recently, with the rapid growth of computer science, there has been some development and application of a set of advanced, nonparametric predictive modeling techniques under the umbrella of machine learning and data mining. These methods have been proposed to overcome the shortcomings of the traditional regression approach and to offer an inductive approach to uncover the underlying relationships in the data, and have been widely adopted in

fields such as applied statistics, bioinformatics, engineering, and financial analysis. Among various machine learning methods, standard recursive partitioning methods, such as classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984), as well as its methodologically improved adaptations such as random forests (Breiman, 2001) have received extensive attention and empirical support in a variety of scientific fields.

Compared to their popularity in other fields, recursive partitioning methods are rarely employed in organizational research. The current study intends to provide an application of these methods (i.e., CART and RF) to the field of organizational research, and addresses two important questions: When should we prefer the recursive partitioning methods over the traditional regression approaches? How should we best apply recursive partitioning methods to organizational research?

To accomplish these purposes, the current paper is organized in the following way. It begins with a review of several limitations associated with the traditional linear additive regression model for classification and prediction in organizational research. CART and random forests are then introduced as methods to address these limitations. Then, four studies are described that compared CART and random forests with traditional methods. In Study 1 and Study 2, two Monte Carlo simulations were conducted to compare the classification and prediction accuracy of CART, random forests, and several classical statistical methods (e.g., logistic regression, linear and quadratic discriminant analysis, and multiple linear regression), under several manipulated conditions. In Study 3 and Study 4, CART and random forests were applied to two empirical samples to better understand employee turnover and job performance.

1. Limitations of the Traditional Regression Approach

1.1 Nonlinearity in Organizational Research

A. Higher-Order Effects

Traditionally under the linear regression framework, each predictor is assumed to have a linear relationship with the criterion variable. Nevertheless, more and more researchers have challenged this assumption and highlighted the importance of investigating nonlinear relationships explicitly (e.g., Guion, 1991).

Nonlinear relationships occur when a predictor is not uniformly related to the criterion across the entire range of the predictor. One of the most common forms of nonlinear effect is a quadratic or cubic relationship. As summarized by Guion (1991), Murphy (1996) and Whetzel et al. (2010), possible quadratic relationships include: (1) a nonlinear but positive monotonic relationship; (2) an asymptotic relationship (J-shape or plateau-shape), which is a nonlinear, monotonic relationship up till a certain point, then no relationship after, and it is also considered as a "deficiency-sufficiency" form (Murphy, 1996); (3) a non-monotonic, U shape or inverted-U shape relationship, where moderate levels are worse or better than too much or too little of a characteristic. Nonlinear models can also take a cubic form. For example, Lord and Novick (1968) suggested that logistic relationship which takes the "S" shape could fit some test scores. An example of the S-curve model is provided by Sinclair and Lyne (1997).

Among the set of nonlinear relationships summarized above, the U shape or inverted-U shape relationship has received the most theoretical and empirical support. For example, evidence has accumulated that the relationship between personality and job performance is not a linear monotonic relationship. In the past two decades, personality inventories have received

growing popularity in making high-stakes selection and promotion decisions in organizational settings (Oswald & Hough, 2010). Contrary to its widespread application, the criterion-related validity evidence of personality measures is far from satisfactory. In a review paper, Arthur, Woehr and Graziano (2001) pointed out several major issues about using personality assessments for selection purposes, one of which was the "appropriateness of linear selection models" (p. 658). Indeed, for a majority of previous studies, one implicit assumption of modeling the personality-performance relationship was linearity. Nevertheless, this assumption has been criticized on the grounds that the omission of the nonlinear relationship between personality and performance could potentially underestimate the true validity of personality measures (e.g., Murphy, 1996; Murphy & Dzieweczynski, 2005; Ones, Viswesvaran, Dilchert, & Judge, 2007; Robie & Ryan, 1999).

For example, conscientiousness has consistently been shown to be the best predictor among the Big Five, and it is generally agreed that higher conscientiousness leads to better job performance. More recently, several researchers began to question this conclusion and argued that excessively conscientious people can be maladaptive. Overly conscientious employees could be considered as "rigid, inflexible, and compulsive perfectionists" (p. 114, Le et al., 2011), and pay too much attention to details that compromise their achievement of the main goal (Mount, Oh, & Burns, 2008). Supervisors who are overly conscientious may also be bounded by rules and regulations, and lack flexibility in handling diverse situations (Murphy & Dzieweczynski, 2005).

Similar arguments have been raised for other Big Five dimensions. For example, Emotional Stability was found to have a curvilinear relationship with job performance: as long as one passes the "critically unstable" range and possess "enough" emotional stability, the positive relationship between emotional stability and job performance disappeared (Barrick & Mount,

1991, p.20). This is also consistent with the effects of tension and stress on performance: moderate level of Anxiety (one facet of Emotional Stability) best facilitates performance. With regard to Agreeableness, highly agreeable employees may also lack the assertiveness to enforce some procedures. For example, a highly agreeable customer service representative may be too cooperative with customers to take actions that are in the company's best interests (Mount, Barrick, & Stewart, 1998), and could potentially be "giving away the store" (Whetzel et al., 2010). Murphy and Dzieweczynski (2005) also articulated that leaders at the high end of agreeableness may not be able to provide bad news and critical feedbacks effectively, which could impair their job performance.

In addition to the theoretical justifications summarized above, most empirical evidence also supports the existence of the nonlinear relationships between various selection predictors and job performance. For example, Day and Silverman (1989) found that impulse expression and several dimensions of job performance had a curvilinear relationship. LaHuis, Martin, and Avis (2005), using a situational judgment test and biodata items as a measure of conscientiousness, also found a significant inverted-U shape relationship between conscientiousness and job performance. More recently, Le et al. (2011) examined the relationships between two of the big five traits (conscientiousness and emotional stability) and different dimensions of job performance (i.e., task performance, OCB, and CWB), and found support for curvilinear relationships. They also demonstrated that job complexity was a significant moderator of the curvilinear relationship, so that the optimal level of conscientiousness for a certain job depends on the complexity of that job. Besides job performance, significant nonlinear relationships have been uncovered for other performance-related criteria such as training performance

(Vasilopoulos, Cucina, & Hunter, 2007) and college GPA (Cucina & Vasilopoulos, 2005; Robbins, Allen, Casillas, Peterson, & Le, 2006).

However, it should be noted that two studies failed to detect any significant nonlinear effect. Robie and Ryan (1999) found no evidence of a significant quadratic or cubic relationship between conscientiousness and job performance across four samples. Le et al. (2011) provided a plausible explanation for this result, stating that the lack of significant nonlinear relationship might be attributed to the complexity of those jobs being sampled, which created a higher threshold for the optimal level of conscientiousness. In another study, Whetzel et al. (2010) examined 32 occupational personality scales and found that nonlinearity was uncommon. In addition, they also showed that the incremental validities of quadratic and cubic terms over linear were modest to small. Part of the reason for the null findings in these studies might be attributed to the inherent limitations of the polynomial and hierarchical multiple regression approaches these researchers have adopted: (1) a prior model specified in advance may lack the flexibility to fully uncover all the complex relationships within the data; (2) the data may not necessarily follow a normal distribution or any other statistical assumptions posited by these methods.

B. Interaction Effects

In addition to the higher-order effects covered above, another source of nonlinearity may arise from the interaction among different predictors. Any predictor does not exist in a vacuum. It co-exists with a constellation of other predictors. Therefore, their joint effect usually provides additional information above and beyond each in isolation (Baron & Kenny, 1986). There could be two patterns of joint effect: a multiplicative effect or a compensatory effect. For the multiplicative effect, it is believed that the benefit of being high on two predictors is greater than

the additive effects of the two in combination; whereas for the compensatory effect, being low on one predictor could be made up by a high score on the other predictor. Theoretical and empirical evidence has also supported the existence of such interaction effects.

Personality and intelligence have been suggested as jointly influencing performance. Maier (1958) proposed that job performance was a multiplicative function of motivation and ability: $P = f\ (\ M \times A\ )$, the relationship between motivation and performance was moderated by ability level. Several empirical studies supported Maier's argument. Hollenbeck, Brief, Whitener and Pauli (1988) showed that self-esteem interacted with aptitude to predict performance of life insurance salespersons: self-esteem was positively related to sales volume for high aptitude salespersons, and negatively related to sales volume for low aptitude salespersons. Wright, Kacmar, McMahan and Deleeuw (1995) demonstrated that Achievement Need was positively related to job performance among warehouse workers who were high in cognitive ability, but the relationship was negative among those low in cognitive ability. Perry, Hunter, Witt and Harris (2010) also identified a multiplicative effect of cognitive ability and Achievement (a facet of conscientiousness) in predicting task performance. They reported that cognitive ability only predicted task performance positively for individuals with high Achievement; for low Achievement individuals, cognitive ability was negatively related to task performance. Besides their joint effect, personality and intelligence can also work in a compensatory manner where a low score in one factor is not necessarily disadvantageous as long as the score on the other factor is high. Using samples from the Navy and the Army, Perkins and Corr (2005) found that cognitive ability could serve as a buffer to neuroticism in that the negative effect of neuroticism on job performance only existed among less cognitively capable individuals. However, for the smarter employees, neuroticism was unrelated to job performance.

In addition to the personality – intelligence interaction, different personality attributes can also interact with each other to influence job performance, as suggested by several researchers (Barrick & Mount, 2005; Hogan, Hogan, & Roberts, 1996; John & Srivastava, 1999). For example, several empirical studies have explicitly examined the interactions among different Big Five dimensions. Witt (2002) found an interactive effect between extraversion and conscientiousness in predicting performance: extraversion was positively related to performance among highly conscientious workers but negatively or unrelated to performance among low conscientious workers. Judge and Erez (2007) found that for customer service employees, extraversion was only positively related to job performance for highly emotionally stable individuals; for customer service employees who scored at the lower end of emotional stability, extraversion was detrimental in that it had a negative relationship with job performance. In another example, Witt, Burke, Barrick and Mount (2002) showed that low levels of agreeableness consistently canceled the positive effects of conscientiousness on job performance across five groups of employees from diverse occupations. Thus, highly conscientious workers who lacked interpersonal sensitivity and competence could be ineffective and even yield dysfunctional outcomes, particularly in jobs that require substantial interactions and cooperation with others. Beyond the Big Five, social skill has also been shown to moderate the relationship between conscientiousness and performance across multiple samples (Witt & Ferris, 2003): conscientiousness was only positively related to performance for individuals with high social skill. Another study by Blickle et al. (2008) examined the moderation effect of political skill and found that, for workers with high political skill, agreeableness was positively related to job performance while conscientiousness was negatively related (Blickle et al., 2008).

C. Configural and Complex Interactions

When interaction effects are combined with higher-order effects, more complex multi-way interaction can emerge among various predictors. And the unique combination of predictors has been argued to carry additional information above and beyond individual predictors (Horst, 1954, 1968; Lee, 1961; Cronbach & Glaser, 1953). As noted by Bergman and Magnusson (1997), "there is a growing acceptance of a holistic, interactionistic view in which the individual is seen as an organized whole, functioning and developing as a totality" (p.291). This argument is also supported by a recent call for a paradigm shift in work psychology from the traditional variable-centric approach to a person-centric approach (Weiss & Rupp, 2011). Following the holistic approach, to represent the patterns of multiple subgroups, a set of configurations (or profiles), as represented by conceptually developed typologies, or empirically captured taxonomies, are usually identified.

The identification of configurations or profiles has been shown to provide a unique value for descriptive purposes (Bergman & Magnusson, 1997; Doty & Glick, 1994; Magnusson & Törestad, 1993). For instance, Asendorpf (2003) suggested that a clear advantage of configural types was their easiness to communicate to a wider audience regarding personality differences. Maertz and Campion (2004) classified employees who voluntarily left their jobs based on why and how they quit, and proposed four profiles of quitters: impulsive quitters, comparison quitters, preplanned quitters, and conditional quitters. Woo (2009) explored six profiles with distinctive combinations of personality characteristics and demonstrated that each profile was associated with a unique pattern of frequencies and reasons for voluntary turnover.

The profiles of different subgroups can also be identified to further improve prediction (Davison & Kuang, 2000). This is usually accomplished through cluster analysis, and the resulting cluster membership is used in multiple regression as an additional predictor. Despite its

theoretical promise, the majority of empirical evidence indicates that cluster membership (or typology) is either unrelated to the outcome or does not provide incremental information above and beyond a simple examination of the effects of individual predictors. For example, Asendorpf (2003) compared the predictive power of the configural type approach with the continuous dimension approach, and found the former was consistently less predictive than the latter in cross-sectional predictions. Within the educational domain, Schmitt et al. (2007) identified five groups of college students based on a combination of fourteen predictors, including measures such as biodata, situational judgment tests, and cognitive ability. However, the group memberships they identified did not offer significant incremental validity above and beyond individual variables in predicting GPA and other areas of college performance. Similarly, Shen (2011) found that personality profiles failed to demonstrate substantial incremental validity over personality levels in predicting OCB (organizational citizenship behavior) or CWB (counterproductive work behavior).

One plausible explanation for the lack of significant findings, as suggested by Shen (2011), is that in traditional cluster methods the specification of the relationship between subgroup membership and criterion was usually set as linear. Another limitation might be a failure to investigate non-linear or non-additive relationships between predictors and the criterion.

1.2 Strict Assumptions of Parametric Approach

Traditional parametric approaches, such as linear regression, or linear discriminant analysis, are subject to several strict assumptions, such as normality, independence, and homoscedasticity of error distribution (i.e., the variance of errors is the same across all levels of the predictor). Another strong assumption is that the predictor variables follow continuous and

multivariate normal distributions. However, as Hosmer and Lemeshow (1989) claimed, the normality assumption is rarely met in reality. For example, response distortion has been widely documented in personnel selection situations (Rosse, Stecher, Miller, & Levin, 1998; Morgeson et al., 2007). The questions included in many personality inventories are fairly transparent and easy to fake (Alliger, Lilienfeld, & Mitchell, 1995; Furnham, 1986), and the job application setting creates a strong situational demand that is likely to invoke high impression management, leading some applicants to answer in a social desirably manner and endorse items that will make them look good. This response distortion inevitably results in a negatively skewed distribution, which may interfere with the predictive validity of selection tests. Rosse et al. (1998) showed that personality scale scores were more negatively skewed for applicants (-.28) compared to incumbents (.04). Skewed distributions may further exacerbate the problem when the selection ratio is low and a top-down selection approach is applied, in which case a substantial number of top-scoring applicants may be individuals who consciously faked good. Since the strict assumptions imposed by the regression approach are usually unrealistic in data sets used by organizational scholars, biased estimates may be engendered when these assumptions are violated.

Summary

There has been substantial conceptual and empirical evidence to date to support the existence of nonlinear, interactive, or even complex configural relationships in organizational research. The omission of these relationships not only leads to underestimation of the linkage between the predictors and the criterion, but also poses serious risks when these results are applied to make high-stakes decisions. The majority of previous studies that have examined the nonlinear or interactive effect generally took the parametric regression-based approach, such as

polynomial regression (Edwards & Parry, 1993) or hierarchical multiple regression (Cortina, 1993). There methods create power or product terms explicitly in the regression equation and test for the significance of the nonlinear terms. However, one limitation of this approach is that the current theories and practices have not been thorough and precise enough to specify the exact nature of all the complex relationships a priori. Another limitation is that the main effect terms in a regression equation that make up an interaction must be entered before the inclusion of that interaction term, which may decrease statistical power for detecting significant interaction effects (Cohen, 1988; see also Crews, 2008; Kiernan, Kraemer, Winkleby, King, & Taylor, 2001). Additionally, traditional multiple regression, as a parametric approach, is subject to several strong assumptions about the underlying data distribution that are usually hard to meet in real situations. As a result, it is important to find an alternative solution that is more flexible in dealing with imperfect data and better at leveraging the information suggested by the data in an explorative and inductive manner. To address these issues, several nonparametric methods in the area of machine learning have been proposed, including recursive partitioning methods such as classification and regression trees (CART) and its successor random forests.

2. Classification and Regression Trees (CART)

2.1 Logic of CART

CART is a binary recursive partitioning method that builds classification trees and regression trees for predicting categorical and continuous variables, respectively (Breiman et al., 1984). CART starts by placing all observations into one node, and then examines exhaustively through all possible binary splits within a predictor and across all predictors to search for a best split. The best split is defined as the one that reduces the impurity of the data in that node the

most (Berk, 2008). For classification trees, several types of node impurity measures have been proposed, including the Gini index, the Bayes error, and the cross-entropy function (Breiman et al., 1984), among which the Gini index is most commonly used. The Gini impurity index can be computed by summing the probability of each observation being chosen multiplied by the probability of making mistakes in categorizing that observation randomly, and the equation is specified as: $\text{GDI} = \sum_{j=1}^{K} p_j(1-p_j) = 1 - \sum_{j=1}^{K} p_j^2$. For a binary outcome, the Gini index has a minimum of 0 when all cases belong to one category and a maximum of 0.5 when cases are equally distributed among all categories, therefore it could be considered as a measure of node balance. For regression trees, the impurity of a node is represented by the within-node sum of squares for the response variable: $i(\tau) = \sum(y_i - \bar{y}(\tau))^2$. Based on the best split, the total sample is then partitioned into two groups. Within each group, the values of the response variable are intended to be as homogeneous as possible. The same searching procedure is then applied to each subgroup separately, with a new optimal split found for each subset to further divide it into more homogenous subgroups. This partitioning of the data continues until a predetermined stopping point is reached, or until no further meaningful reduction in the node impurity can be found, resulting in a single tree. Any subsequent new observations can then be classified based on the set of split decisions (if-then logical conditions) derived from this original tree.

The output of a CART analysis is often displayed in a tree diagram. Figure 1 is a simple illustration. The full dataset is stored in the root node. Cases with variable *X* greater than C_1 are placed to the right branch and cases with *X* equal to or smaller than C_1 are placed to the left branch. The right subset can be divided again, with cases having variable $Y > C_2$ going to the right node and $Y <= C_2$ going to the left. From a regression point of view, all splits beyond the

initial split of the root node imply interaction effects (Berk, 2008). In the current example, there is an interaction effect between *X* and *Y* so that the influence of *Y* depends on the value of *X*. As CART can reuse the same predictor over and over, possibly with different thresholds or cutoff values, there are times when a subsequent partitioning uses the same predictor again, resulting in a more complicated curvilinear function. Berk (2008) named it as a "stagewise" process given that earlier stages will not be revisited once a later stage is set.

CART determines the predicted value based on a majority vote. For a classification tree, the label of the majority group within a terminal node is attached to that node. For example, if 80% of the cases in terminal node 3 are high performers and 20% are low performers, the label of terminal node 3 is then assigned as "high performers." When a new case is classified into this node, its predicted group will be "high performer." For a regression tree, the mean value of the response variable within a terminal node is calculated, and assigned as the predicted value for that group.

After a tree has been constructed, how should its classification or prediction accuracy be evaluated? In general, a training sample is used to develop a tree classifier (i.e., a set of decision rules), and a test sample runs through the classifier to make classifications/predictions. The discrepancy between the predicted and the observed outcomes is used to estimate the level of accuracy, typically indicated by error terms. Two common types of error are usually computed: Overall misclassification error in classification, or overall mean squared error (MSE) in regression, when the test sample is the same as the training sample used to develop the tree classifier, and the error rate represents how skillful the method has been in ***fitting*** the data. This is also called the resubstituition error (Breiman et al., 1984). Generalization error (i.e., cross-validation error, as introduced below) is derived when two different samples, a training sample

and a test sample, are used to build and test the tree structure respectively, indicating how skillful the method has been in ***forecasting*** (Berk, 2008). For a classification tree, a confusion table, which cross-tabulates the observed and predicted classes, is also critical for evaluating model accuracy from the following perspectives: Type I error (false positive), Type II error (false negative), true positive) and true negative.

Researchers such as Berk (2008) and Strobl, Malley and Tutz (2009) have claimed that the estimated resubstituition error rate is overly optimistic and yields lower than actual error rates. Indeed, Cureton (1950) already pointed out the problem with the resubstitution error in his classic article. Therefore, when multiple samples are available, or a large sample can be partitioned into several subsets, it is preferred to use different samples for training and testing to produce a more honest and accurate estimation. In the case when no test sample is available or training sample is limited in size, a widely-used alternative is the *k*-fold cross-validation procedure (Breiman et al., 1984). According to this method, the original sample is randomly partitioned into *k* equally sized subsamples. Each time, the combination of the (*k* -1) subsamples is used to grow the tree, with the one left-out subsample serving as the test sample. This process is then repeated *k* times until each subsample has been used exactly one time as the test sample. The average classification accuracy (or MSE) across the *k* results is then used as a single estimate of cross-validation accuracy. In practice, *k* usually takes the number of 10. However, if *k* equals to *N* (i.e., total sample size) and each test subsample includes only one case, this cross-validation method is also called the "leave-one-out" estimate.

2.2 Empirical Applications of CART

In the past three decades, CART has been used in various scientific areas, such as genetics, bioinformatics, ecology, finance, mechanical engineering, and neuroscience (e.g., De'ath & Fabricius, 2000; Frydman, Altman, & Kao, 1985; Reibnegger, Weiss, Werner-Felmayer, Judmaier, & Wachter, 1991). Within the broad discipline of psychology, there has been an increasing application of CART for predicting a variety of outcomes, such as violence (Steadman et al., 2000), sexual recidivism (Lloyd, 2008), suicidal ideation (Kerby, 2003), positive and negative affect (Gruenewald, Mroczek, Ryff, & Singer, 2008), perceived life stress (Scott, Jackson, & Bergeman, 2011), college admission (Sah, 1998), and student attrition (Windham, 1994). Two recent reviews have summarized the application of recursive partitioning methods in psychology (cf. Merkle & Shaffer, 2011; Strobl, Malley, & Tutz, 2009). Despite its popularity in other areas, the application of CART in I/O psychology and management science is rare. In the following section, I review the few studies that have either applied CART, or compared CART with traditional methods for predicting and/or explaining various organizational behaviors.

One area in I/O psychology that CART has been more frequently considered is to predict and explain employee attrition and job performance. For example, a military study conducted by Stark et al. (2011) compared traditional multiple logistic regression with CART in predicting attrition among first-term enlisted soldiers, and found CART performed slightly better than logistic regression in classifying "stayers" and "leavers". More importantly, CART was able to uncover nonlinear and configural relationships in the response data and identify different reasons of staying or leaving for soldiers. Gjurich (1999) predicted naval Surface Warfare Officer (SWO) retention level using logistic regression and CART, and identified the reasons for officers' attrition. Reis (2000) also applied CART in another military setting to select the most important

factors for predicting student aviators' performance in the aviation-training pipeline, including both a categorical outcome of completion/attrition and a continuous outcome of flight school grades. For a non-military setting, Chien and Chen (2008) proposed using a data mining framework to explore the relationship between personnel profiles and work behaviors, and developed decision trees for predicting job performance, retention and turnover reasons based on demographic information and work experience for a high-tech company. Beyond understanding attrition and performance, Keeney, Snell, Robison, Svyantek and Bott (2004) have applied CART to examine patterns of organizational climate and employee personality characteristics associated with three different organizations. Their findings supported the ASA (i.e., Attraction-Selection-Attrition) theory and the presence of a person-environment interaction. CART has also been used to shorten surveys and remove items that do not contribute. For example, in a study conducted by Lee and Drasgow (working paper), a 22-item retention likelihood scale was examined using CART, and the final rule incorporated only 4 useful items that still provided an accurate classification of stayers and leavers. Petway (2010) also applied CART to reduce the length of the Big Five Inventory.

2.3 Comparing CART with Classical Methods

As summarized above, standard parametric models such as linear and logistic regressions fall short in several aspects when used in organizational sciences. The introduction of CART may overcome some of the limitations associated with traditional methods, as described below.

A. Flexibility in handling complex relationships

CART provides researchers with a flexible tool for exploring and exploiting nonlinear and configural relationships as suggested by the data. It is especially suitable for conditions when

the researchers suspect that there are complex nonlinear or interaction effects among the predictors, or between the predictor variables and the criterion, but are not sure about the exact form a priori. In addition, compared to multiple regression's one-model-fits-all procedure, CART is able to specify an individual equation for each subgroup. In CART, each group of participants can be traced through a series of specific splits and corresponding splitting variables, with each set of splits representing a multiple regression equation. A sample regression equation is

$$f(X,Z) = \beta_0 + \beta_1 [I(x \leq c_1)] + \beta_2 [I(x > c_1 \,\&\, z \leq c_1)] + \beta_3 [I(x > c_1 \,\&\, z > c_2)]$$

where I represent the indicator function. As noted by Berk (2008, p.104), "In practice, there is no need to translate the partitioning into a regression model; the partitioning results stand on their own as a regression analysis. But if one wishes, the recursive partitioning can be seen as a special form of stepwise regression." This is especially suitable for identifying subgroups and their corresponding characteristics.

B. Not restricted by parametric assumptions

Another flexibility of CART is that it's a nonparametric method, which means it does not rely on assumptions about sampling observations from a normal distribution. As discussed above, it can be difficult to meet these assumptions with real data. In situations like these, CART shows superior flexibility in handling a wide variety of data types, without requiring any sort of data transformation prior to analysis. On a related note, extreme responses and outliers are more likely to happen in situations with seriously skewed distributions. While the parametric method can be greatly impacted by outliers if they are not explicitly detected and removed, CART just isolates the outliers into a separate node. Hence, CART's results are less likely to be affected by extreme responses.

C. Ease of handling missing data

When missing values exist for the independent variables, a common procedure embraced by many statistical models is pairwise or listwise deletion, excluding observations with one or more attributes missing. This treatment potentially reduces the amount of information being used for model estimation. Missing values are less of an issue for CART though, because CART examines one variable at a time, so only cases with missing values in that particular variable are temporarily removed, whereas all the other cases are retained for that investigation. When other variables are considered, the observations deleted temporarily in a previous step will be included again. This approach does not involve permanently deleting any cases when developing a predictive algorithm and therefore maximizes the information being used.

When CART is used to make predictions for a new observation with missing values in any of the splitting variables, a so-called *surrogate variable* is used to replace the missing variable (Hastie, Tibshirani, & Friedman, 2001). A surrogate variable in CART is the one drawn from the remaining set of variables that most closely mimics or predicts the splits of the primary variable that is missing. This is a much simpler approach than imputing missing values and is preferable when dealing with cases that have partially missing data.

D. White box approach

In most cases, the interpretation of the results derived from a decision tree (i.e., a set of if-then conditional statements) is straightforward and intuitively clear. Any hierarchical nature of the independent variables, their interactions, and their relative importance for different classes, are explicit within the tree structure. This white box approach can be of tremendous help for researchers to uncover the underlying relationships between the predictors and the outcome variable in a more explicit manner, potentially leading to an enhanced understanding and better

theory building. For example, each branch in the decision tree can be considered as an individual profile, and the set of predictors and their cut-off values help define the characteristics of that particular profile. Breiman et al. (1984, p.7) noted that "an important criterion for a good classification procedure is that it not only produces accurate classifiers (within the limits of the data) but that it also provides insight and understanding into the predictive structure of the data." Moreover, a tree diagram is often much easier to comprehend than the regression coefficients for practitioners. This approach is even more desirable when compared to other "black-box" advanced nonlinear classifiers, such as artificial neural networks (ANN; Bishop, 1995), whose internal patterns are never uncovered.

As a summary, the above section suggested that CART provides a valuable alternative in areas where the traditional regression approach falls short, and offers additional insights that are not otherwise revealed. However, CART is not without its shortcomings. In the following section, I discuss some caveats to note when applying CART.

2.4 Caveats for CART

A. Model Instability

CART may have limitations when stability and generalizability are considered. To make the terminal nodes as homogeneous as possible, CART can construct a very large tree with very few cases in each terminal node, which is usually called *overfitting*. Overfitting may engender three concerns: it causes the tree structure to be unstable, hard to generalize across samples, and difficult to interpret. Overfitting can be attributed to the following three reasons: first, a small sample size, especially when one or more terminal nodes include only a small number of observations; second, weak predictors, which is usually indicated by heterogeneous terminal

nodes; third, highly correlated predictors, which means selecting one of the two competing predictors rather than the other can change a tree structure substantially. To protect the tree from overfitting and strengthen the generalizability of the results, several remedies have been proposed. The first strategy is to set up a stopping rule, so that the tree will no longer be split after the stopping rule is achieved. A variety of heuristics or options can be used to determine the stopping rule, such as setting up a minimum leaf size (i.e., the number of cases within each node), specifying the homogeneity criteria, or defining the desired number of intermediate steps. The second approach to prohibit CART from constructing a large and unstable tree is through "pruning" (Breiman et al., 1984; Berk, 2008), usually in the backward direction. The underlying mechanism of pruning is to first grow a large tree, then cut off the branches that do not increase predictive accuracy sufficiently. The most widely used pruning strategy is the "minimal cost-complexity pruning" developed by Breiman et al. (1984), which seeks to find a balance between overall misclassification error and tree complexity (i.e., number of terminal nodes) by adding a penalty for complex trees. This process removes weak branches by combining nodes that do not reduce misclassification error sufficiently to justify the extra complexity added. As the value of the penalty parameter can vary across a range, a sequence of optimal trees can be generated. Therefore, a second stage following this pruning step is to carry out a cross validation to select a single best tree. The so called "one standard error rule" (Breiman et al., 1984) is generally used to make this best selection, that is, choose the smallest-sized tree whose cross-validation costs are within the one standard error range of the minimum cross-validation cost. In cases when a large number of predictors are included for a relatively small sample, a more advanced approach to increase model stability and generalizability is through resampling. By resampling, hundreds and thousands of trees are constructed to create a tree ensemble, and the results averaged across

the ensemble are then used as the decision rule. The two most popular ensemble methods – Tree Bagging and Random Forests – are discussed below.

B. Model Bias

Another problem with CART is that its "greedy" search mechanism may induce bias in variable selection (Doyle, 1973; Loh, 2002; Loh, 2009): with all other things equal, variables with a larger number of distinct values have a greater chance to be selected as the splitting variables (Hothorn, Hornik, & Zeileis, 2006). For example, a categorical variable X that takes n distinct values can have $(2^{n-1} - 1)$ binary splits. When n increases, this number grows exponentially. Thus, a categorical variable with more values has a larger opportunity to be chosen. This selection bias can potentially cause irrelevant variables to be selected, leading to erroneous inferences.

An alternative method for constructing classification and regression trees, called GUIDE (i.e., Generalized, Unbiased Interaction Detection and Estimation), has been developed by Loh (Loh, 2002; Loh, 2009) to overcome the problem of selection bias. The main argument of GUIDE is that compared to restricting splits to one variable on a single node, one can sometimes gain greater predictive accuracy by splitting on a linear combination of multiple variables. For classification trees, GUIDE involves three steps of chi-square significance tests. Step one is to test for main effects. It performs a chi-square test of independence of each predictor versus the outcome variable and calculates its significance probability. The first split is then based on the most significant variable. When no main effect achieves a specified level of significance, a second step is conducted to test for interaction effects. If an interaction test is significant, a two-level split search for the splitting points is performed. A two-level split is needed because the

best split of X_1 may depend on how X_2 is subsequently split; similarly, the best split of X_2 may also need to consider the best subsequent splits of X_1. Considering their joint effect, the pair of variables with the largest chi-square value is then selected. If no test in the first two steps is significant, a third step is performed to test for linear structure. However, Hastie et al. (2001) cautioned that even though this linear combination of splits might improve predictive power, it may hurt interpretability. Therefore GUIDE is limited to two variables at a time when fitting a linear model. With respect to regression trees, GUIDE helps solve three problems: first, it controls for selection bias by applying a chi-square test for signed residuals and a bootstrap calibration of significance probabilities; second, it also includes tests for significant pairwise interactions between regression variables; third, it fits more complex node models (e.g., multiple, stepwise linear, polynomial and ANCOVA) than the piecewise constant model of CART.

GUIDE, although helpful for reducing model bias and uncovering good splits hidden behind interactions, still faces the same problem of instability and lack of generalizability. To address these challenges, two ensemble methods based on the CART algorithm, Tree Bagging and Random Forests, have been proposed.

3. Tree Ensemble Methods

3.1 Tree Bagging

"Bagging," an abbreviation for "Bootstrap Aggregation," is built on the idea of bootstrap resampling and majority vote (Breiman, 1996). Take a classification tree as an example. Each time, a random sample of size *N* is drawn with replacement from the original data to construct a single decision tree without pruning. For each tree, a case is assigned to one of the terminal nodes and the category associated with that node is stored. After repeating this process a number

of times, hundreds or even thousands of trees are constructed. Finally, one can count how many trees have classified a case into each category, and assign the class membership based on majority vote. As for regression trees, the average response across multiple trees is calculated as the predicted outcome for each case. An even more accurate procedure based on bagging is called "out-of-bag" estimation: only observations not included in the bootstrap training sample are used to form the test sample. By aggregating a collection of fitted values, tree bagging is able to compensate for overfitting by reducing random variations and canceling out idiosyncrasy of the training data, which results in a more stable decision algorithm and a more honest estimate of model fit.

Tree bagging, nevertheless, is not without limitations. First, the model averaging process is a black-box approach with limited interpretability. Since bagging involves averaging across a large number of trees, there is no longer a single tree structure for interpretation, and no direct way to relate predictors to the response as in a standard decision tree. The output generated by a bagging procedure is merely the predicted class for each case or the predicted probability of membership in a class, yet the underlying mechanism regarding how each case is assigned is unknown. Another limitation exists when a set of predictors are correlated with each other and the fitted values are not fully independent (Berk, 2008). For example, when two important predictors are highly correlated, the inclusion of one could potentially obscure the effect of the other and prohibit its inclusion, leading to a biased estimate. The introduction of the Random Forests method provides a solution to these problems (Breiman, 2001).

3.2 Random Forests

Similar to tree bagging, random forests method also leverages the advantage of model averaging by constructing a number of trees from bootstrap samples, and determining the predicted values according to the majority vote (or average). A critical difference between tree bagging and random forests lies in the way they select predictors at each split. For Tree Bagging, all the predictors are examined before each node is split, whereas for random forests, a random sample of the predictors without replacement is taken before each split. This treatment prevents an important predictor from being outplayed by a strong competitor that is highly correlated, and gives each predictor a chance to make an impact in different contexts with different covariates. As a consequence, beyond reducing sampling error, the random forests method also has the following three advantages compared to CART and tree bagging: First, it is able to work with a very large number of predictors. Second, it is preferred for dealing with substantially correlated predictors. Third, it helps further reduce random variance beyond sampling of observations. Therefore, random forests has been claimed to be most suitable for "small *n* large *p* problems" (Strobl, Malley, & Tutz, 2009, p. 339) when there are too many predictors with too few observations. Instead of simultaneously considering all the predictors at the same time, random forests examines a subset of predictors at one time, which is easier to handle. This strategy demonstrates a clear advantage over classical methods such as logistic regression when the number of dimensions is high (e.g., Bureau et al., 2005). Nevertheless, as with bagging, the random forests method is also a black-box approach that is limited in interpretation, lacking a unique tree structure and corresponding statistical functions linking the predictors with the outcome.

3.3 Determining Variable Importance

Estimating the relative contribution of each predictor for overall predictive accuracy usually offers another lens to interpret a model. This is not uncommon given that regression weights are widely considered as the indicator of variable importance. The random forests method provides a mechanism to calculate variable importance that includes two variations. The first one is called raw importance (or Gini importance for classification trees), which evaluates the total decrease in node impurity by splitting on a variable for each tree and then averages over all trees. Every time a split of a node is made on a particular variable, the sum of the split criterion (i.e., an impurity index such as Gini for classification or MSE for regression) for the two child nodes is less than the parent node. Adding up the impurity decreases for each individual variable over all trees in the random forest gives the estimate of the variable importance. Although intuitive, this method has been shown to carry forward the bias of a single tree, which favors continuous variables or variables with many categories (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Thus, it is not recommended in situations when predictor variables vary in their number of categories or scale of measurement. The second approach is called permutation importance (Breiman, 2001). This is achieved by randomly permuting the values of any predictor across all of the observations in the training data set or out-of-bag data set, and comparing the difference in predictive accuracy (i.e., misclassification rate or MSE) between the real data and the permutation data across all trees. If the random permutation of a particular variable substantially decreases the predictive accuracy, then this predictor is considered as a reasonably important one. Otherwise, it is not considered as important. A limitation of the permutation importance method is that it severely over-estimates the importance of a predictor variable if it correlates with other variables (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008).

In simple multiple regression, the values of the standardized regression coefficients (or squared standardized regression coefficients) can be used to quantify the importance of different predictors and the strength of the connections between the predictors and the criterion. However, regression coefficients could be seriously distorted given the existence of multicollinearity among predictors. Recently, several comprehensive alternatives based on more solid statistical inferences such as dominance analysis (Azen, Budescu, & Reiser, 2001) and relative weights analysis (Johnson, 2000), have been proposed to overcome the multicollinearity problem. Dominance analysis addresses the problem of correlated predictors by examining the additional contribution of each predictor across all subset of regression models (Azen & Budescu, 2003; Budescu, 1993). Relative weights analysis, on the other hand, adopts a variable transformation approach by creating a set of new predictors that are orthogonal counterparts of the original ones (Johnson, 2000). Although developed based on different statistical rationales, past research has shown that the two methods tend to yield virtually identical results (LeBreton, Ployhart, & Ladd, 2004). However, dominance analysis is less preferred, especially when the number of predictors is large, mainly due to its computational burden (i.e., it computes all possible subsets of regression models). Indeed, Johnson and LeBreton (2004) noted that dominance analysis can handle no more than 10 predictors at a time. Therefore, relative weights analysis is generally preferred, particularly for high-dimensional data.

Compared to the traditional regression coefficient approach, both random forests and dominance analysis/relative weights are superior when multicollinearity is present (e.g., Bi & Chung, 2011). However, two unique features further differentiate random forests from dominance analysis/relative weights. First, both dominance analysis and relative weights are limited to the first-order linear regression model and do not include any interactive or higher

order polynomial effects. This means the variable importance generated from these two techniques reflects only the "main effects". Random forests, however, covers the impact of each predictor individually as well as its multivariate interactions with other predictors. Second, the foundational assumption of both dominance analysis and relative weights analysis is that the form and function of the regression model have been correctly specified (Budescu, 1993; Tonidandel & LeBreton, 2010). This could not be guaranteed particularly with high dimensional data that includes complex interactions. The random forests method offers more flexibility in cases when the regression model is not specified a priori.

4. Comparing Classical Methods, CART, and Random Forests

4.1 Brief Review of Several Classical Methods

Two of the most commonly adopted modeling methods for classification problems are Fisher's discriminant analysis (DA) and logistic regression (LR). The underlying mechanism of discriminant analysis is to find a linear combination of a set of continuous predictors that best differentiates two or more classes of objects or events based on ordinary least squares (OLS). Discriminant analysis formulates a function that predicts the probability of an individual falling into each category, and then assigns the group membership based on the highest probability. When group variances are equal, linear discriminant analysis (LDA) is used; otherwise, quadratic discriminant analysis (QDA) is applied. However, DA is based on OLS that assumes normal distributions of the error variances. When this requirement is not satisfied and the dependent variable is a binary outcome, logistic regression (LR) is usually preferred for modeling the binomial distribution. LR is a special form of the generalized linear model with the link function being a log transformation of the odds ratio (i.e., the probability that an observation belongs to

one category divided by the probability that it falls into the other category), which is often called a logit function, as shown in the formula below. It describes how the probably of a binary outcome is related to a linear combination of the predictor variables (Hosmer & Lemeshow, 1989). Unlike DA that uses OLS, LR obtains parameter estimates through maximum likelihood estimation.

$$\text{logit}\left(\pi(x_i)\right) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \alpha + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + e_i$$

or alternatively in terms of $\pi(x_i)$:

$$\pi(x_i) = \frac{\exp\{\alpha + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + e_i\}}{1 + \exp\{\alpha + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + e_i\}}$$

When the response variable is a continuous numerical one, multiple linear regression, which is another form of generalized linear model with the link function being "identity," is most commonly used (formula shown below):

$$Y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + e_i$$

where $e_i \sim N\ (0, \sigma^2)$ and independent.

4.2 Comparing Classical Methods, CART, and Random Forests

A number of previous empirical and simulation studies have compared the predictive accuracy of CART and the above mentioned classical methods. Results from empirical studies based on single dataset have been conflicting. For instance, several studies have found that traditional methods such as LDA, QDA, and LR performed comparably or better than CART (Arminger, Enache, & Bonne, 1997; Dudoit, Fridlyand, & Speed, 2002; Preatoni et al., 2005;

Ripley, 1994; Williams, Lee, Fisher, & Dickerman, 1999), whereas other studies (e.g., Grassi, Villani, & Marinoni, 2001) reported that CART provided higher classification accuracy than LDA. These conflicting results might be attributed to the idiosyncratic nature of the sample being used. A couple of simulation studies suggested that certain data and model distribution characteristics had differential impacts on the classification accuracy of traditional methods and CART. For example, Finch and Schneider (2006, 2007) conducted two Monte Carlo simulations that compared the cross-validated misclassification accuracy of LDA, QDA, LR, and CART under a variety of simulated conditions. Their results indicated that when the parametric assumptions such as normal distribution and equal covariance matrices of the two groups were met, LDA and LR performed as well as QDA and all were slightly better than CART. However, when the two groups' covariance matrices are not equal, regardless of the underlying distribution of the predictors, QDA and CART consistently outperformed LDA and LR. The more unequal, the better QDA and CART performed. They also found a significant main effect of the group size ratio: when the two groups were extremely unequal in size (e.g., 10:90), the misclassification rate was consistently higher for the smaller group regardless of data distribution, sample size, or the modeling method being used. However, Finch and Schneider (2006) showed that CART did offer some buffering effects in situations like that, showing a slightly lower misclassification rate for the smaller group and a higher classification error for the larger group. These authors noted that one limitation of their simulations was that the interaction effects among the variables were not considered at all. Therefore, a more comprehensive simulation was conducted to extend the comparison (Holden, Finch, & Kelly, 2011). Holden et al. (2011) showed that when the simulated models included interactions, linear models such as LDA, QDA, and LR consistently had difficulty in obtaining accurate classification, whereas CART always

produced the greatest accuracy. The superiority of CART was particularly notable when the relationship between group membership and the predictors was complex and nonlinear. Similar conclusions were also supported by several other simulation studies regarding the impact of nonlinear relationships (Curram & Mingers, 1994; West, Brockett, & Golden, 1997), with the results showing that LDA and LR performed as well as more advanced alternatives such as classification trees under linear conditions, but were not as effective when the relationships between the predictors and the criterion measure were nonlinear.

Prior simulation studies comparing CART with traditional multiple linear regression (MLR) for predicting continuous outcomes have demonstrated similar results to that when predicting a categorical criterion. Breiman et al. (1984) conducted several simulation and empirical studies comparing regression trees with linear regression, and concluded that regression trees could be much more accurate for nonlinear problems, but tended to be less accurate when the problem followed a good linear structure. Garson (1998) also showed that linear models such as MLR performed poorly for prediction when a number of interactions existed among the predictor variables, based on a Monte Carlo simulation. Empirical evidence showed that CART performed comparable or slightly better than multiple linear regression in terms of predictive accuracy (Finch et al., 2011; Kerby, 2003; Fu, Anderson, Courtney, & Hu, 2007), yet demonstrated additional advantages in other areas: CART displayed lower RMSE and higher R^2 results than did MLR, and offered cutoff scores to aid decision making in applied settings.

Unlike comparisons between CART and traditional methods which are bounded by sample and model characteristics, studies comparing random forests with other methods have generally shown that random forests possess higher levels of predictive accuracy than other

classical methods or a single decision tree. Halstead (2006) compared linear regression with random forests and two alternatives and found that random forests best predicted military recruiter's job performance while linear regression yielded the worst prediction. Based on the random forests variable importance selection mechanism, Halstead (2006) identified the best subset of features out of 260 predictors that produced the best model generalization. Sut and Simsek (2011) compared CART, random forests, and a set of other regression methods for predicting mortality in head injuries, and showed better performance for random forests than CART based on both overall accuracy and area under the ROC Curve. Loh (2009) also conducted a set of simulation studies comparing ensemble tree methods with several single-decision-tree methods such as CART and GUIDE, and found that in general, ensemble methods fit better than any of the single-tree methods, and random forests was the best for most samples. In addition to superior predictive accuracy, the random forests method also excels at finding important predictors. A simulation study conducted by Eliot (2011) showed that when there was no structure among the predictors, random forests performed better than LR and CART at finding the order of variable importance, especially when predictors were not highly correlated. However, when the data was structured in the way that one variable was only predictive of the outcome within a level of another variable, CART and LR described the data better than random forests. Eliot (2011) suggested that each method may be best suited for certain types of data and distributions, and combining them may provide additional insights.

5. The Current Study

Given previous theoretical and empirical evidence, it is expected that CART would outperform traditional classification and regression methods, such as DA, LR or MLR, when nonlinear, interactive, or complex configural relationships exist in the data. The classification

accuracy may also vary depending on parameters such as group size ratio (Finch & Schneider, 2006, 2007; Perlich, Provost, & Simonoff, 2003). However, although random forests is expected to perform better than CART and other traditional methods in a variety of conditions, there is a lack of simulation research comparing random forests with CART and other methods under nonlinear data structure or various group size ratios. As a consequence, the first purpose of the current study is to replicate and extend previous simulations by comparing CART and random forests with classical methods such as logistic regression, linear and quadratic discriminant analysis, and multiple linear regression for predicting both categorical and continuous outcomes under a variety of manipulated conditions. In addition to the set of parameters examined before, I also included some other factors that have theoretical or empirical relevance with organizational research. Using a series of Monte Carlo simulations, the first two studies examined the methodological strengths and weaknesses of CART and random forests compared to the set of commonly used regression-based methods, particularly in the context of organization research.

The second goal of the current research was to illustrate the empirical application of CART and random forests using real data, comparing their predictive accuracy with the same set of traditional methods for addressing employee turnover problem (a binary outcome) in Study 3 and job performance (a continuous outcome) in Study 4. The aim of these two studies was to offer practical guidance regarding how CART and random forests could be combined to enhance both model prediction and interpretation.

In a summary, the current paper provides an exposition and new findings regarding recursive partitioning methods for the researchers in the field of organizational sciences, and offers some empirical suggestions about when and how these methods should be applied.

CHAPTER 2: METHOD

STUDY 1

In Study 1, a Monte Carlo simulation was carried out to systematically compare the two decision tree methods, CART and random forests (RF), with three traditional classification methods: logistic regression (LR), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). In particular, the current study aimed at exploring if and when the decision tree methods were superior to the traditional methods. The manipulated parameters were selected to cover a variety of characteristics associated with the sample (base rate), distributions (level of skewness), relationships within the set of predictors (multicollinearity), and the relationships between the predictors and the outcome variable (variance explained and data complexity). These parameters were chosen based on their theoretical and practical relevance to personnel selection research (Aguinis, Culpepper, & Pierce, 2010; Bobko, Roth, & Potosky, 1999; Finch, Edwards, & Wallace, 2009).

Personnel selection was of particular interest in the current paper because, given the high-stakes context of personnel selection, predictive accuracy is usually the foremost concern – even a small improvement in predictive utility can lead to a substantial monetary saving, such as reducing training costs (White, Nord, Mael, & Young, 1993) or increasing return on investment (ROI). In addition, understanding how predictors and criterion are related could assist personnel psychologists to develop better selection decisions. Note that sample characteristics and sample size were beyond the scope of the current study. This is because previous literature (Fan & Wang, 1999; Lei & Koehly, 2003) found that sample size was not a significant factor in differentiating CART from LR or other traditional methods. Thus, the sample size was fixed to 1, 000 in the current study to make sure that when the base rate was set to be relatively small, there would be

sufficient cases in the minority group to yield meaningful results. Finally, the number of predictors was fixed at four to replicate what was done in previous research (e.g., Holden, Finch, & Kelly, 2011; De Corte, Lievens, & Sackett, 2006).

Manipulated Parameters

Multicollinearity. To examine whether the existence of multicollinearity would impact the accuracy of different classification methods, the intercorrelations among the predictors were manipulated. In simulation studies, the correlation matrix of the predictor variables is usually used to generate data with a specified structure of the predictors. Past simulations in the area of personnel selection (e.g., Bobko, Roth, & Potosky, 1999; De Corte, Sackett, & Lievens, 2011; Finch, Edwards, & Wallace, 2009) usually took the correlation matrix of the most commonly used selection predictors (e.g., conscientiousness, cognitive ability, biodata, structured interview and integrity test) derived from meta-analyses to represent the input matrix. Despite its high fidelity and relevance, this fixed set of intercorrelations is limited in its coverage of a full range of relationships. As indicated by Cohen (1988), the effect size of a Pearson correlation coefficient ranges from small (below 0.3) to large (above 0.7). As such, to increase the variety of input matrices, the elements in the simulated four by four matrix were randomly generated from two intervals, 0 – 0.3 and 0.3 – 0.7, to represent two scenarios: the four predictors were either relatively independent (i.e., low multicollinearity) or moderately to highly correlated (i.e., high multicollinearity), respectively. A correlation coefficient above 0.7 was not considered in the current study, given that this scenario is rare in organizational research when different predictors are used to assess different constructs.

Level of skewness. Skewness was manipulated in the current research to study violation of the normality assumption of the traditional parametric approaches. The data were generated as either positively skewed or negative skewed, with the level of skewness being set as either 0.3 or 0.7, so there were four levels of skewness: -0.7, -0.3, 0.3, and 0.7. Note that 0.3 and 0.7 were chosen based on extensive review of actual data from selection tests. I was given permission to access the database of a consulting firm, which included a variety of selection tools that measured many non-cognitive predictors. Data analyses showed that in general, the skewness ranged from 0.3 to 0.7 in real selection situations. As such, these numbers were chosen in the present study to represent slightly and substantially skewed data.

Base rate. As the criterion of interest of the current study was a binary variable, an important parameter to be considered was the base rate (i.e., prior probability distribution of the two classes). Within the realm of personnel selection, base rate could play out in the form of selection ratios (e.g., Komar, Brown, Komar, & Robie, 2008; Sackett & Roth, 1996), turnover rate, or success/failure rate. The focal interest of the current study was to understand how to better predict whether a selection decision is successful or not. So 1 was used to represent a failure and 0 representing a success. As it is rare to have 50% to 90% of the applicants being selected fail their job, only the first two levels of base rate, 0.1 and 0.3, were included in the current study. Based on the data generated from the input matrix as well as the functional relationships between the predictors and the criterion (i.e., the first four models described below), an array of continuous outcome variables (y) can be obtained. To convert the continuous variable into a dichotomous variable, the values of y were ranked in descending order, and the base rate was used to determine the cut-off point with either the top 10% or top 30% of the cases being assigned a 1, and the remaining assigned a 0.

Data complexity. The functional relationship between the set of predictors ($x_1 - x_4$) and the outcome variable (y) was also manipulated to affect data complexity. Five models linking the predictors with the outcome were defined. Model 1 was the most basic, a simple linear additive model. Models 2 to 4 were modified to incorporate nonlinear components in the forms of interaction, quadratic and cubic terms. Model 2 added the quadratic term of x_1^2 to the basic model. Model 3 was specified by adding an interaction term between x_1 and x_2. Model 4 was the most complex additive model: it included an interaction term, a quadratic term, and a cubic term. Regression weights for these four models were set to one to simply the relationship, and a specific error component was added to each model. More information about how the error component was created is introduced in the next paragraph. Despite the difference in the level of complexity among Models 2 to 4, they were essentially all under the same umbrella of polynomial regression. However, the actual relationship within the data may go beyond the form of a linear regression model or any type of polynomial regression. As such, I decided to further extend the functional relationship between the predictors and the criterion variable, bringing in a configural model structured in the form of a classification tree. A prototype tree (shown in Figure 2) was created to factor in various conditions, including the higher-order effect of a single predictor, and two-way and three-way interactions among various predictors. The mathematical expression of this model was mapped out as a set of stepwise regression functions, with a_1, a_2, b_1, b_2, c_1, c_2 and d_1 being a set of predetermined cut-off scores at different decision points. To make sure that the final base rate was 0.1 or 0.3 as defined, meaning the group size ratio of 1 versus 0 being either 1:9 or 3:7, one parameter (b_1) was freed up at the very beginning of model set-up with all the other parameters fixed. This extra degree of freedom was created to allow the base rate requirement to be met. After the value of $b_1$was determined, it was then incorporated back to

generate the final model. (A side note: as the base rates specified in the current study was a bit extreme, not any given value of a_1 could result in the expected base rate of 0.1 or 0.3. Thus, multiple values of a_1 were plotted against the final base rate with b_1 fixed, and the value of a_1 that yielded a base rate closest to 0.1 or 0.3 was used). The models are specified as follows:

1. $y_1 = x_1 + x_2 + x_3 + x_4 + e_1$

2. $y_2 = x_1 + x_2 + x_3 + x_4 + x_1^2 + e_2$

3. $y_3 = x_1 + x_2 + x_3 + x_4 + x_1 * x_2 + e_3$

4. $y_4 = x_1 + x_2 + x_3 + x_4 + x_1 * x_2 + x_3^2 + x_4^3 + e_4$

5. $y_5 = 1$, if

   $x_1 \geq a_1$ & $x_2 \geq b_1$,

   or $x_1 \geq a_1$ & $x_2 < b_1$ & $x_3 \geq c_1$,

   or $x_1 < a_1$ & $x_1 \geq a_2$ & $x_2 \geq b_2$,

   or $x_1 < a_1$ & $x_2 < b_2$ & $x_3 \geq c_2$ & $x_4 \geq d_1$.

   $y_5 = 0$, if

   $x_1 \geq a_1$ & $x_2 < b_1$ & $x_3 < c_1$,

   or $x_1 < a_2$ & $x_2 \geq b_2$,

   or $x_1 < a_1$ & $x_2 < b_2$ & $x_3 \geq c_2$ & $x_4 < d_1$,

   or $x_1 < a_1$ & $x_2 < b_2$ & $x_3 < c_2$.

Proportion of variance explained (R^2). Organizational researchers are usually interested in the proportion of criterion variance explained by the set of predictors, represented by the coefficient of determination R^2. The square root of R^2 can be conceptually understood as a Pearson correlation coefficient between the predictor composite and the criterion. To manipulate the composite correlation level, I chose 0.3 and 0.7 respectively indicating either a small or a large correlation. The manipulated R^2 was set to 0.09 and 0.49, accordingly. Based on the following equation: $R^2 = 1 - \frac{VAR_{err}}{VAR_{tot}}$, the error variance could be determined if the composite variance of predictors variance was known. To estimate the composite variance of the predictors, I first generated 10,000,000 sets of random data of x_1 through x_4 , then calculated y_1 through y_4 based on the above equations without adding any error variance. The variance of y_1 through y_4 were then computed and served as an approximate measure of composite predictor variance. Returning to the equation above, the level of error variance VAR_{err} could then be estimated for each equation. Lastly, the error component in each of the four equations was randomly generated from a normal distribution of (0, VAR_{err}).

Prediction method. Since the main goal of the present study was to compare the predictive accuracy of different classification methods, five methods, including classification tree (CART), random forests (RF), logistic regression (LR), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA), were applied to the set of random samples generated.

To sum up, the current study included a full-factorial design crossing 2 levels of multicollinearity (0.0 – 0.3 and 0.3 – 0.7), with 4 levels of skewness (-0.7, -0.3, 0.3, 0.7), 2 levels of base rate (0.1 and 0.3), 5 levels of data complexity, 2 levels of variance explained (0.09 and 0.49), and 5 prediction methods. This design created 800 unique combinations of parameter

values as summarized in Table 1. A pilot study was conducted to compare the results from 10 random samples with that from 100 random samples and no significant difference was found. Therefore, 10 random samples were drawn for each condition, producing a total of 8, 000 samples.

To evaluate classification accuracy, two measures were compared. First, an overall misclassification error (a.k.a., resubstitution error) was calculated when the same sample used to build the tree was also applied to fit the model. The resubstitution error represents how well the model *fit* the data. Second, a 10-fold cross-validation error, which is the mean error rate based on the 10-fold cross-validation procedure described above, was estimated. It is a measure of how well the method is at forecasting. Cureton (1950) clearly illustrated the danger of interpreting the resubstituition error rate, whereas the forecasting measure such as cross-validation error is likely to be a much more realistic indicator of the classification accuracy.

Data Analysis

All the analyses were conducted using MATLAB (2012 a) statistics toolbox. To generate the input correlation matrix, "RndomCorr" was first used to generate a random correlation matrix with four dimensions (Numpacharoen & Atsawarungruangkit, 2012), then I wrote a function to transform the entities in the matrix to match the set intercorrelation intervals (either 0.0 – 0.3 or 0.3 – 0.7). The lower triangular matrix was constructed using Cholesky factorization ("chol"), which was then being multiplied by a 4*1000 matrix of random numbers drawn from a Pearson distribution with standard deviation of 1, kurtosis of 3 (i.e., normal distribution) and skewness set at one of the four values (-0.7, -0.3, 0.3, and 0.7). Following this process, a predictor matrix that included 1,000 sets of $x_1 - x_4$ was generated, representing the non-normal data that satisfied the

predetermined parameters. Next, response variables y_1 to y_5 were computed based on the equations specified above for each set of $x_1 - x_4$. Those continuous y values were then converted to dichotomous response variables according to one of the base rates (0.1 and 0.3). This process was repeated 10 times to generate 10 random samples for each condition.

In the next step, different classification methods were applied to fit the data generated above. Classification trees were grown using the "ClassificationTree.fit" function. Random forests were constructed using "fitensemble". Logistic regression was applied using the "glmfit" function with "binomial distribution" and the "logit" link function. And linear and quadratic discriminant analyses were performed using the "ClassificationDiscriminant.fit" function.

To calculate various classification accuracy metrics, some MATLAB functions that are common to the four types of classifiers (CART, RF, LDA, and QDA) were applied: "resubLoss" was computed to calculate resubstitution error and "kfoldLoss" was used to estimate 10-fold cross-validation error. With regard to logistic regression, none of the above statistics were readily available. As a result, I wrote MATLAB functions to generate these metrics based on their definitions. In the current study, the ROC curve approach was applied and the optimal value that minimized the overall misclassification error was calculated and set as the cut-off point. Since the cut-off score was set up in a way to maximize the classification accuracy, it is noteworthy that any results generated from this approach represented the upper bound of the classification accuracy of LR, and should be interpreted with caution. To calculate 10-fold cross-validation error for logistic regression, the data was randomly ordered and 10 equal segments were created. For each trial, nine segments were used as the training data to build the LR classifier while the remaining one serving as the test data. This process was repeated until all of

the 10 segments were utilized as the test data for one time. Then the mean classification error was calculated over the 10 trials.

Results across 10 replications were averaged, and a factorial analysis of variance (ANOVA) was conducted to compare the two classification accuracy metrics under different simulated conditions. The main effect for each factor and interaction effects between these factors and the classification methods were evaluated. Complex three-way interaction effects were beyond the consideration of the current study given their difficulty of interpretation.

Results

Tables 2 and 3 list the raw results of the resubstitution error rate and 10-fold cross-validation error rate for the five classification methods grouped by the five manipulated parameters. Table 4 is a summary of the average resubstitution and cross-validation error rate for each of the data models (Y1 to Y5) under each of the analysis methods. As shown, both CART and RF exhibited smaller resubstitution error rates than the other three methods. This is because both CART and RF were fully grown without pruning, resulting in a large number of terminal nodes within each decision tree (i.e., more than one hundred for $N = 1000$ observations in most conditions). Additionally, RF has the seemingly remarkable error rate of 0.0 or almost 0.0 in all conditions. The explanation for this effect is that RF has generated its classification based on 100 trees. Some additional analyses showed that when only a single tree was grown, the resubstitution error rate of RF was similar to that of CART. However, as more trees were included in the RF, its resubstitution error decreased and approached 0 gradually.

With regard to cross-validation, among all the methods, CART has the highest cross-validation error rates across all the conditions for the four polynomial regression models (Y1 –

Y4). Given that CART has a much smaller resubstituition error rate, this is a clear indication of the existence of overfitting. Surprisingly, the cross-validation error rate of RF is smaller than CART by substantial margin, and comes close to LR and LDA in the condition that most favorable to the latter two methods (i.e., the linear additive model, Y1). This is because the greedy algorithm of RF has been balanced by its averaging mechanism that prevents it from over-fitting. As the data generation methods become more complex, RF becomes as effective as the three traditional methods (except for Y4). Post-hoc analyses showed that the difference between RF and the three traditional methods was not significant for model Y1 through Y4. However, when the data was structured in a configural pattern (Y5), RF was significantly more effective than the set of traditional methods. In this latter condition, CART's performance is similar to RF, yet the error rates are not quite as low.

Table 5 summarizes the means and standard deviations of the two classification accuracy metrics under different parameters. Main effects of all the parameters and interaction effects between classification method and all the other parameters are summarized in Table 6. According to the tables, there is a general pattern that higher intercorrelations among predictors (i.e., the existence of multicollinearity) result in slightly more accurate classifications, albeit the main effect was not significant. This pattern is consistent across different classification methods, as indicated by the non-significant interaction effects. The main effect of proportion of variance explained is significant for both metrics, so that a larger explained variance (i.e., a smaller error variance) is associated with a smaller resubstitution error and a smaller cross-validation error. Base rate also exhibits a significant main effect across both metrics. A larger base rate is associated with a higher error rate for both training and test data. This is because a smaller base rate presents a simpler classification problem in the sense that all cases can be classified as the

majority group resulting in a low error rate. Moreover, explained variance and base rate interact with classification methods significantly when resubstitution error was evaluated. According to Figures 3 and 4, the three traditional methods are more substantially influenced by the change of explained variance and base rate, while the two decision tree methods tend to be more stable. Regarding the parameter of skewness, there is a tendency that the more positively skewed a dataset is, the more accurate the classification will be, yet none of the main effects or interaction effects was significant. Data complexity has a significant impact on resubstitution error and cross-validation error, with the tree-like stepwise regression model (Y5) consistently yielding the most accurate predictions, whereas the remaining models (Y1 – Y4) do not differ notably from each other. The two significant interaction effects between data complexity and classification methods are illustrated in Figures 5 and 6. As shown in Figure 5, when resubstituition error is considered, random forests classification is most stable across different models. For the remaining methods, only Y5 could be significantly differentiated from all the other equations. When model forecasting capability is considered (i.e., cross-validation error), Figure 6 shows that CART does not predict as well as the three traditional methods for the polynomial regression models (Y1 – Y4), but is as good as random forests when the model is specified as a tree-like structure (Y5), and both CART and random forests are much better than the traditional methods under model Y5. Among the five methods, random forests consistently yields classification error rate that is close to optimal. CART manifests some advantage over the three traditional methods for configural data (i.e., Y5) but is the worst for polynomial data. The three traditional methods do not differ substantially.

Discussion

Results from the current study show that when different classification methods were compared on how well they fit (or overfit) the data (i.e., resubstituition accuracy), the two decision tree methods consistently outperformed all the three classical methods, across all the manipulated parameters; moreover, random forests was even more accurate than CART with the resubstitution error close to 0. On the other hand, when forecasting ability was compared, CART and random forests exhibited significantly higher cross-validation error rates than resubstitution error rates, whereas the three traditional methods did not show much difference. This result was consistent with Berk's (2008) claim that the resubstitution error rate is an overly optimistic measure for decision tree methods and thus cross-validation error rate is required to evaluate the true classification accuracy.

Additionally, there was a significant interaction effect between data complexity and classification method on predictive accuracy. When data was structured in a linear regression or any type of polynomial regression form, random forests performed comparably to the three traditional approaches. CART was significantly less accurate than the other four methods. Nevertheless, when the data was structured in a tree-like stepwise function, both random forests and CART significantly outperformed the remaining traditional methods, while random forests and CART did not differ significantly from each other. This result indicates that the disadvantage of CART only existed when traditional polynomial regression based models were used to generate the data; when a tree model (Y5) was specified, CART outperformed all the three traditional methods. Therefore, the current study suggests that the reason previous empirical and simulation studies that failed to show the advantage of CART is that inappropriate models were studied. If the data has a tree-like structure, CART would show some benefits, such as increased classification accuracy. Among the remaining set of parameters, proportion of variance

explained, base rate, and data complexity all displayed some significant main effects, so that a larger proportion of explained variance, a smaller base rate, and a more complex model (Y5) were all associated with better classification. However, interaction effects indicated that the resubstitution accuracy of the three classical methods showed a more substantial difference under different levels of explained variance and base rate, yet their decision-tree counterparts were less impacted by these parameters.

STUDY 2

In Study 1, classification accuracy was examined when the outcome was a dichotomous variable. To extend the comparison to a continuous outcome, a similar simulation was conducted in Study 2 to compare the predictive accuracy of CART (regression trees in this case), random forests (RF), and multiple linear regression (MLR). No previous simulation study has compared the predictive accuracy of the two decision tree methods with the multiple linear regression approach when predicting a continuous outcome. According to results from Study 1, it was expected that random forests would produce the highest level of accuracy, especially under the more complex polynomial regression and configural models. Depending on the metric being compared, multiple linear regression and CART may have advantages under certain conditions.

Manipulated Parameters

Three parameters (i.e. multicollinearity, level of skewness, and proportion of variance explained) were set similar to Study 1 with one exception: three levels of explained variance were considered in this study – in addition to the 0.09 and 0.49 conditions, another scenario with a much larger explained variance of 90% (i.e., 0.95*0.95 = 0.90) was also considered. This condition was added to allow the data structure to more closely mimic the pre-defined equations. As the outcome was no longer a binary variable for the present study, base rate was no longer applicable and was therefore not considered. The set-up of the data complexity was slightly different, as explained below.

Data complexity. Equation Y1 – Y4 were the same as in Study 1. Y6 was constituted by eight multiple conditional regressions, representing the eight branches specified in Figure 2:

$$y_6 = x_1 + x_2 + e_1, \text{ if } x_1 \geq a_1 \ \& \ x_2 \geq b_1;$$

$$= x_1 + x_2 + x_3 + e_2, \text{ if } x_1 \geq a_1 \,\&\, x_2 < b_1 \,\&\, x_3 \geq c_1;$$

$$= x_1 + x_2 + x_3 + e_3, \text{ if } x_1 \geq a_1 \,\&\, x_2 < b_1 \,\&\, x_3 < c_1;$$

$$= x_1 + x_1 + x_2 + e_4, \text{ if } x_1 < a_1 \,\&\, x_1 \geq a_2 \,\&\, x_2 \geq b_2;$$

$$= x_1 + x_2 + e_5, \text{ if } x_1 < a_2 \,\&\, x_2 \geq b_2;$$

$$= x_1 + x_2 + x_3 + x_4 + e_6, \text{ if } x_1 < a_1 \,\&\, x_2 < b_2 \,\&\, x_3 \geq c_2 \,\&\, x_4 \geq d_1;$$

$$= x_1 + x_2 + x_3 + x_4 + e_7, \text{ if } x_1 < a_1 \,\&\, x_2 < b_2 \,\&\, x_3 \geq c_2 \,\&\, x_4 < d_1;$$

$$= x_1 + x_2 + x_3 + e_8, \text{ if } x_1 < a_1 \,\&\, x_2 < b_2 \,\&\, x_3 < c_2.$$

Note that some of the equations above were identical. However, it is kept in this manner intentionally to represent the eight branches in the tree structure explicitly. In addition, since different models (Y1 through Y6) have different scale levels, to make the results comparable, standardized scores were calculated for all the response values generated from all the models.

Prediction method. Three methods were evaluated regarding their predictive accuracy: regression tree (CART), random forests (RF) and multiple linear regression (MLR).

To sum up, Study 2 included a full-factorial design crossing 2 levels of multicollinearity (0.0 – 0.3 and 0.3 – 0.7), with 4 levels of skewness (-0.7, -0.3, 0.3, 0.7), 5 levels of data complexity, 3 levels of variance explained (0.09, 0.49, and 0.90), and 3 prediction methods, resulting in 360 unique combinations of parameter values, as summarized in Table 1. Similar to Study 1, pilot study showed that the results from 10 random samples did not differ significant from 100 random samples. Therefore, a random sample of 10 was drawn for each condition, producing a total of 3,600 samples.

Two statistics were computed to evaluate predictive accuracy: (a) overall mean squared error (a.k.a., resubstitution MSE), which was the summed squared error between predicted and observed values divided by the number of cases; and (b) 10-fold cross-validation mean squared error, which was the mean MSE across 10 cross-validations.

Data Analysis

A process similar to that in Study 1 was used to generate non-normal data, and continuous response variables were created using the five equations specified in this study. Regression trees were grown using the "RegressionTree.fit" function, random forests were constructed using "fitensemble," and multiple linear regression was applied using the "glmfit" function with "link" parameter specified as "identity." To calculate overall MSE and cross-validation error, "resubLoss" and "kfoldLoss" were used, respectively.

10 iterations were averaged and one-way and two-way analyses of variance (ANOVA) were conducted to compare the two predictive accuracy statistics under a variety of simulated conditions. The main effect of each parameter and the interaction effect between these parameters with the prediction method were estimated.

Results

Table 7 and 8 show the resubstitution MSE and 10-fold cross-validation MSE for various predictive techniques grouped by the five manipulated parameters. The average resubstitution and cross-validation error rate for each of the data models (Y1 to Y6) under each of three analytical methods (i.e., MLR, CART, and RF) is summarized in Table 9. Unlike the classification problem where RF had the smallest resubstitution error, Table 9 show that CART had the smallest error for training sample, followed by RF, both were much smaller than MLR.

When cross-validation error was considered, MLR consistently showed better predictive accuracy than RF and CART under the simplest model (Y1), and when the explained variance was small (0.09) across all the other models. For all the other conditions, RF exhibited the best predictive accuracy. The comparison between MLR and CART showed that on average CART was not as accurate as MLR, but when the level of variance explained was large (0.90) and the data was structured via the most complex polynomial regression model (Y4) or configural tree-like model (Y6), CART significantly outperformed MLR, yet did not do as well as RF.

Table 10 is a summary of the means and standard deviations of the two predictive accuracy metrics for different parameters. Main effects of all the parameters and interaction effects between regression method and all the other parameters are summarized in Table 11. Similar to Study 1, higher predictor intercorrelations lead to a slightly smaller MSE for resubstitution and cross-validation, albeit the difference between the two levels of intercorrelations is not significant for either metric. The proportion of variance explained exhibits a significant main effect on both resubstitution and cross-validation errors, with a large variance associated with more accurate prediction. Moreover, proportion of variance explained also interacts significantly with regression method. As shown in Figures 7 and 8, an increase in explained variance remarkably reduces the resubstitution error of MLR and the cross-validation error of CART. Skewness, as in Study 1, does not show any significant main or interaction effects. Unlike what was found in Study 1, data complexity did not show any main or interaction effects for the regression problem. A further analysis indicated that the three-way interaction among proportion of variance explained, data complexity, and prediction method was significant for both resubstitution MSE, F (16, 315) = 18.54, $p < .001$), and cross-validation MSE, F (16, 315) = 8.71, $p < .001$. As shown in Figure 9, the difference between MLR and CART/RF

resubstitution error was larger when the model became more complex, especially when a large proportion of variance was explained. With regard to cross-validation error, Figure 10 shows that when the model is simple, RF and MRL did not show much difference; as the model gets more complex, decision tree methods, especially RF, began to show an advantage, especially when the variance explained was moderate to large. Unlike Study 1, the superiority of CART under the tree-like model (Y6) was only prominent when the error variance was small and the data more closely represent the structure being specified. Lastly, predictive method is another significant factor. When MLR was used to fit the data, the resubstitution MSE was significantly larger than when CART or random forests was used, and CART had a slightly smaller error than RF. When cross-validation error is considered, random forests presents a significantly better forecasting capability than CART and slightly more accurate prediction than MLR.

Discussion

Results from Study 2 showed that when different classification methods were compared on their resubstitution accuracy, the two decision tree methods consistently outperformed the linear model, across all the manipulated parameters. When CART and random forests were compared, results from Study 1 and Study 2 were slightly different: CART was a little more accurate than random forests for predicting a continuous outcome, while RF tended to capture the nuances of the original data better than CART for classification problems, though the different between the two methods was not significant. On the other hand, when forecasting ability was compared (i.e., cross-validation accuracy), Study 2 also showed that CART and random forests exhibited significantly higher cross-validation error than resubstitution error. The multiple linear regression method did not display much difference with regard to the two types of errors. Across various situations, random forests produced the most accurate cross-validated

prediction on average. However, proportion of variance explained, data complexity, and prediction method displayed a significant three-way interaction: when the explained variance was small, RF did not differ significantly from MLR; when the explained variance was moderate to large, RF was better than MLR for all the complex polynomial regression models (Y2 – Y4) as well as the configural tree-like model (Y6). Moreover, when the explained variance was large, CART showed substantial smaller cross-validation error rates than MLR when the data was structured in the most complex polynomial regression model (Y4) and configural tree-like pattern (Y6). The current study again shows that a predictive method is most advantageous when the data structure is aligned with the assumptions and specifications of the method. This conclusion holds especially when noise is not the dominant component in the data.

Similar to Study 1, the proportion of variance explained exhibited a significant main effect and an interaction effects with classification method. However, the existence of multicollinearity (i.e., higher predictor intercorrelations) tended to play in favor of classification problems and against regression problems, although the difference between the two levels of multicollinearity was not significant. Future study should investigate whether these results can be replicated.

To sum up, results of Study 2 further support the superiority of random forests method for making precise predictions across a range of conditions, especially when the data structure is more complex than the simple linear model, and the error variance is small. CART had the highest level of accuracy under resubstitution, but was less accurate for cross-validation. CART's cross-validation error was only smaller than traditional methods when the data was structured in a complex polynomial regression model or a configural tree-like pattern, and a large proportion of criterion variance was explained by the predictors.

STUDY 3

To supplement the simulations in Studies 1 and 2, the goal of Study 3 was to illustrate the empirical application of CART and random forests to actual data for predicting the dichotomous outcome of employee turnover. More specifically, I addressed two objectives in the current study: prediction and interpretation. The first objective was to examine which method offers the most accurate prediction (primarily forecasting) of both involuntary and voluntary employee turnover. To achieve this goal, the two decision tree methods (CART and random forests) were compared with the three traditional methods (LR, LDA, QDA) on various metrics. In addition, as multiple predictors were included, I also explored if and how predictive accuracy changes as a result of the number of predictors. The second objective was to understand how CART and random forests could be used to interpret the causes of employee turnover. As shown by previous studies, random forests exhibited superior predictive accuracy, yet lack a single tree diagram to interpret the functional relationship between the predictors and the outcome; CART was good at interpretation, but fell short at cross-validation accuracy. To balance the interpretive and predictive power of decision tree methods, the current study was designed to offer some practical guidance to combine the two methods through variable selection and decision tree construction.

Sample

The sample in Study 3 was from the employee work experience survey data collected by a large insurance company in the United States. The work experience survey was conducted with new employees when they reached their milestone tenure points (i.e., 3 months and 15 months). The current sample included 12,417 full records of current and past employees (8,144 cases for 3-month, and 4,273 for 15-month). By the time the data was aggregated, some employees were surveyed once (i.e., those who had not stayed up till 15 months), whereas others were surveyed

at both tenure points. For this reason, the actual number of unique employees included in the sample was smaller than 12,417. However, as both employee names and work IDs were detached from the dataset, the exact number of unique employees included was not identifiable. Gender and ethnicity information was not available either. Among the 12,417 cases, 570 (4.59%[1]) employees were terminated involuntarily, 2,742 (22.08%[2]) employees left voluntarily, and the remaining 9,105 cases represented employees who stayed in the organization for either 3 months or 15 months. Among the involuntary turnover, 425 (74.56%) of them were terminated at or before the 3-month time point, while the remaining 145 (25.44%) were terminated between the 3-month and the 15-month time points. 1,987 (72.47%) cases of the voluntary turnover left at or before the 3-month time point, while the remaining 755 (27.53%) employees left between 3 months and 15 months. The reasons for involuntary termination identified by the company included 9 categories, such as absent without leave, death, employment agreement expired, job requirements not met and misconduct. In the current study, I was only interested in the three categories under involuntary turnover that represent poor performance (i.e., absent without leave, job requirements not met, and misconduct), which added up to 416 cases. The reason for voluntary turnover defined by the organization included 10 categories, such as another opportunity, dissatisfaction, did not return from maternity leave, domestic duties, moved, regular retirement, return to school, etc. The current study only focused on the five categories under voluntary turnover that were not related to either retirement or family/external reason. These categories include another opportunity, became an agent, dissatisfaction, return to school, and no reason given, resulting in 1,997 voluntary turnover cases. In combination with the remaining 9,105 employees who stayed, a full record of 11,518 employees were included in the final

[1, 2] *Note: the actual percentage is larger than this given that the unique number of employees is smaller than 12,417*

analysis, covering three major work groups: 6,138 employees were from the paraprofessional/support group (53.29%), 4,353 (37.79%) were from the professional/technical group, and 892 (7.74%) were from the management group. With regard to tenure, 7,475 of them were from the 3-month group, and 4,043 were from the 15-month group.

Measures

In the work experience survey, employees were asked to evaluate various aspects of their workplace including how they viewed themselves, their supervisor, their job, and the organization, etc. The survey items were then grouped into nine factors based on factor analysis conducted in a previous study by the organization. A description of the nine factors follows.

Onboarding (ONB). The onboarding factor includes 9 items about employees' first impression and onboarding experience. A sample item is: "I felt welcomed into my team/unit when I started at [organization name]." Cronbach's alpha reliability of this factor was 0.84.

Supervisor (SUP). The supervisor factor consists of 9 items that cover employees' feelings about how they are treated by their supervisor. A sample item is: "My supervisor gives me feedback that helps to improve my job performance." Cronbach's alpha reliability was 0.93 for this factor.

Coworker (COW). 6 items were included to measure employees' evaluations of their coworkers, and their work relationship with their coworkers. One sample item is: "My coworkers treat each other with dignity and respect." Cronbach's alpha was 0.90 for this factor.

Senior Leadership (SL). This factor is made up of 6 items that asks employees to evaluate the behaviors of the senior leaders in the organization. One sample item is: "Senior leadership

considers the needs of employees when changes are made." Cronbach's alpha reliability was 0.91.

Financial Rewards (FW). This factor includes 7 items that ask employees if their compensation and benefits are fair and satisfactory. One sample item is: "My pay is fair for the work I do." Cronbach's alpha reliability of this factor was 0.86.

Job Fit (JF). Job fit is a measure of job satisfaction that consists of 5 items. One sample item is: "I like the kind of work I do." Cronbach's alpha reliability was 0.84 for job satisfaction.

Discretionary Effort (DE). Discretionary effort factor comprises 2 items that asks if employees put in extra effort at work. A sample item is: "I perform beyond what is typically expected or required." Cronbach's alpha reliability of this factor was 0.64.

Organizational Commitment (OC). This factor, although named as organization commitment, includes 5 items that actually measure employees' views about their development opportunities within this organization. A sample item is: "I feel I can grow and develop at [organization name]." Cronbach's alpha reliability of this factor was 0.81.

Intent to Stay (ITS). Intent to stay is a measure of employees' withdraw and turnover intention that is made up of 5 items. A sample item is: "I am seriously considering leaving this company." Cronbach's alpha reliability of this factor was 0.83.

All of the predictors included in the study were rated on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

Outcome Measure. The dependent variable in the current study was the employees' work status when the survey was collected including three categories – involuntary turnover, voluntary turnover, and staying. Leavers were coded as 1 and stayers as 0.

Data Analysis

The focal interest of the present study was to apply different classification methods for predicting a *binary* outcome in an applied setting. Since the outcome measure includes three categories, I decided to segregate the data into two groups. The first group compared involuntarily terminated versus stayed employees, and the second group contrasted voluntarily left with stayed employees. The analyses were organized around the two objectives of prediction and interpretation.

To meet the first objective, the five classification methods were applied to the data when all the nine predictors were included, and compared based on the following four classification accuracy metrics: overall misclassification rate (resubstitution error), 10-fold cross-validation error, and 10-fold cross-validation sensitivity and specificity. The same set of MATLAB functions and procedures as described in Study 1 were adopted in the current study. A second objective was to explore how the classification accuracy of various methods changed as the number of predictors changed.

Regarding model interpretation, a variable selection process was first applied. Variable selection is widely used in many domains where a large number of predictors are present, such as bioinformatics. To narrow the focus to a smaller subset of variables that maximizes model prediction, the variable selection process included two steps, a variable importance ranking and a backward elimination. First, variable importance was computed and ranked based on the

following five statistics: regression coefficients, squared regression coefficients and *p* values that were derived from logistic regression, relative weights that were generated from logistic relative weights analyses (Tonidandel & LeBreton, 2010), and predictor importance ranking that was estimated from random forest using the "predictorImportance" function of MATLAB. Note that this particular MATLAB function calculates the variable importance based on raw importance rather than permutation importance. I chose to use raw importance in the current study since all the predictors were continuous with approximately the same score range.

Second, based on the importance ranking derived from the first step, one predictor was eliminated at a time, starting from the least important one. As a result, the five classification methods under the designated 8 conditions with the number of predictors ranging from 2 to 9 were compared against the four accuracy metrics, resulting in 160 conditions.

Finally, based on the set of most important predictors identified, a single classification tree was built and pruned. To overcome the instability and enhance the theoretical relevance of the resulting tree diagram, a cost function was introduced to adjust the relative weights associated with sensitivity and specificity.

Results

Descriptive statistics of the nine predictors for the three groups of people (i.e., stayed, voluntary turnover, and involuntarily turnover) are summarized in Table 12. For almost all the predictors, stayed employees exhibited the highest mean ratings among the three groups (except for the factor of Discretionary Effort where stayed and involuntarily left employee were similar). Voluntarily left employees showed the lowest ratings (except for the Coworker factor where

involuntary turnover employees rated the lowest). Table 13 shows the means, standard deviations, and intercorrelations among the predictors for all employees.

When all the nine predictors were included to predict voluntary turnover, random forests generated the smallest resubstitution classification error rate (0.0161), followed by CART (0.0972). Random forests and CART both had much smaller resubstituition errors than the three traditional methods of LR (0.1782), LDA (0.1792), and QDA (0.1962). When evaluated against the cross-validation accuracy, LR (0.1789) and LDA (0.1792) had slightly smaller errors than RF (0.1869), followed by QDA (0.1974), and all four methods were much smaller than CART (0.2624). Sensitivity results averaged across 10-fold cross-validation showed that CART (0.2048) was most sensitive at classifying voluntarily left employees, followed by QDA (0.1402), RF (0.0481), and the two simpler methods had the smallest sensitivity: LR (0.0342), LDA (0.0371). The specificity level was opposite to the pattern of sensitivity with CART (0.8545) having the smallest specificity, and LR (0.9937) and LDA (0.9926) having the largest specificity.

When predicting involuntary turnover, RF (0.0062) showed the smallest resubstitution error, followed by CART (0.0323), both were smaller than the three traditional methods. However, CART showed some level of overfitting with the largest cross-validation error (0.0634), while RF (0.0438) had a smaller cross-validation error that was similar to the level of LR (0.0439) and LDA (0.0437). Both CART (0.0264) and QDA (0.0264) showed superior cross-validation sensitivity, while the remaining three methods were not able to successfully identify involuntary turnover. The zero sensitivity and near one specificity for the three traditional methods were likely due to the small base rate (4.37%) of the involuntary turnover sample.

When various predictors were ranked according to their importance derived from different approaches, the results showed a mixture of similarity and difference (see Tables 14 and 15). For the comparison between regression coefficients with relative weights, although both β^2 and relative weights indicated that ITS (Intent to Stay) and JF (Job Fit) were the two most important predictors for predicting both voluntary and involuntary turnover, the remaining order of the predictor importance was quite different for the two statistics. This difference between β^2 and relative weights might be attributed to multicollinearity among predictors. Indeed, the correlation matrix in Table 13 showed that the predictors themselves were all moderately to highly correlated. In situations when multicollinearity is present, regression coefficients may obscure the contribution of each individual predictor. As such, the ranking from relative weights analysis is preferred.

When comparing variable importance ranking from relative weights analysis with that from random forests, both perspectives indicated that Intent to Stay (ITS), Job Fit (JF), and Financial Rewards (FR) tended to be among the most important factors for both voluntary and involuntary turnover. Nevertheless, Onboarding (ONB) was the most important factor for predicting voluntary turnover and second most important factor for predicting involuntary turnover, according to random forests variable ranking. Yet it was listed at the bottom of the relative weights ranking, and was not significant in the logistic regression. The supervisor (SUP) factor showed a similar pattern. The difference between random forests and relative weights, variable-importance rankings suggests that there might be some complex interaction or higher order effects that relative weights do not capture. Given the presence of multicollinearity and perhaps complex nonlinearity, random forests may provide the most comprehensive and accurate

picture of variable importance. The variable selection and model construction process introduced in a later paragraph were therefore based on the order derived from random forests.

Tables 16 and 17 summarize the results comparing classification accuracy of the five methods when different numbers of predictors were included. As shown, LR and LDA exhibited the smallest cross-validation error rates across various situations, and their classification accuracies were relatively stable despite the number of predictors. However, their smallest cross-validation error was at the cost of sensitivity. Since the base rate was relative small in both scenarios (17.99% for voluntary turnover and 4.37% for involuntary turnover), LR and LDA simply classified almost all cases into the larger base rate group (i.e., "stayer") yet still attained the smallest cross-validation error. QDA had slightly larger cross-validation error rate than LR and LDA, and this error rate was impacted by the number of predictors included: the more predictors, the larger the cross-validation error. Unlike LR and LDA, QDA was able to identify some of the "leavers." There was also a general pattern that more predictors lead to a larger sensitivity for QDA. In comparison, the two decision tree methods (i.e., CART and RF) exhibited larger cross-validation error than resubstitution error, and the difference between the two metrics was more significant when more predictors were involved. CART showed the largest cross-validation error across all situations, but was among the best when sensitivity was evaluated. Similar to QDA, CART was also sensitive to the changing number of predictors – more predictors led to a larger cross-validation error and a larger sensitivity. With regard to RF, its cross-validation error was also influenced by how many predictors were used, but in the opposite direction as CART and QDA. When the number of predictors was small, RF had slightly larger cross-validation error than LR and LDA; but as the number of predictors got

larger, RF's cross-validation error approached that of to LR and LDA. The sensitivity of RF was moderate, larger than LR/LDA, but smaller than CART/QDA.

In summary, CART was most sensitive to the changing number of predictors, followed by RF and QDA; whereas the two simpler methods (LR and LDA) were more resistant to change. In general, LR and LDA showed the smallest cross-validation error rate, but fell short at sensitivity. CART and QDA were superior regarding sensitivity, but exhibited large cross-validation error when more predictors were included. RF showed relatively balanced cross-validation error and sensitivity, especially when the base rate was not extremely small, and RF was more advantageous when the number of predictor was large.

Understanding the functional relationship between the predictor variables and the outcome was another vital piece of information for model interpretation. However, the black box approach of random forests fell short without a specific tree diagram. As such, CART was used to grow a single classification tree that provided a visual representation of the specific process to enhance model interpretation. As CART may become unstable when the number of predictors is large, a subset of the nine predictors was selected to fit the decision tree. Information from both random forests and relative weights variable importance was taken into consideration to determine what predictors to include. As such, the top five predictors (i.e., ONB, FR, SUP, ITS, JF) for voluntary turnover and the top six predictors (i.e., FR, ONB, SUP, JF, ITS, and COW) for involuntary turnover were chosen as a starting point to build a decision tree for each scenario, respectively.

To understand voluntary turnover, the full tree was first grown without pruning. The resulting structure was quite unstable with 57 levels of partition and far too much detail for

meaningful interpretation. As a result, the full tree was pruned according to the "1 standard error rule" (Breiman et al., 1984). Figure 11 shows a clean, interpretable structure without removing any predictor. However, a further examination of the terminal nodes showed that there were several terminal nodes with very small sample sizes. For example, one of the terminal nodes contained only 9 cases, although 8 out of the 9 were voluntary turnovers. Another terminal node consisted of a relatively balanced distribution of stayed versus left employees (47:61) even though the node was assigned as "turnover." As Berk (2008) suggested, when the number of cases within each terminal is small, or the proportion of the two categories is similar, the classification tree may face a high risk of being unstable and hard to generalize, and interpretation should be carried out with great caution. To mitigate this risk, a cost function indexing the relative weights of false positive versus false negative was taken into consideration. From a theoretical standpoint, it is more detrimental if a potential quitter is not spotted, compared to incorrectly assigning a stayer to be a quitter. Therefore, I set the ratio of the cost of a false negative to a false positive to be 5 in the current study[2]. After the costs were factored in, a new classification tree was fully built that contains 102 levels. The full tree was then pruned with 98 levels trimmed off, resulting in a meaningful, interpretable structure as shown in Figure 12. The sample sizes in the terminal nodes of this tree were large enough to draw conclusions about the splitting decisions. A rough skim of the results indicated that the majority of the voluntary turnover cases were able to be captured by this tree diagram. This observation is further supported by the sensitivity and specificity levels. Indeed, the sensitivity level of this classification tree (0.6159) is higher than all the traditional models considered, and even better than when all the predictors were included. As expected, the specificity level was compromised

[2] 5 was chosen because the ratio between stayed versus voluntarily left employees was roughly 5.

(0.5736). However, as a high specificity was not as desirable as a high sensitivity in the current case, I relied on this final decision tree to explain the functional relationship between predictors and outcome. According to the tree diagram, three out of the original five predictors (i.e., Intent to Stay, Financial Rewards, and Supervisor) were retained in this model with some level of interactions, and five profiles of employee voluntary turnover were identified: (1) ITS < 3.775; (2) ITS ≥ 3.775 & FR < 3.365; (3) 3.775 ≤ ITS < 4.1 & FR ≥ 3.365; (4) ITS ≥ 4.1 & FR ≥ 3.365 & 4.415 ≤ SUP ≤ 4.435; (5) ITS ≥ 4.1 & FR ≥ 3.365 & 4.805 ≤ SUP ≤ 4.87 (shown in Figure 12).

Similar procedures were applied to understand involuntary turnover. The top six predictors were first included to build and prune a classification tree, resulting in a tree diagram displayed in Figure 13. The sample sizes of this tree were even smaller for the majority of the terminal notes, partly due to the smaller base rate of involuntary turnover. Therefore, a cost function was factored in with the ratio of costs between false negative to false positive set as 20[3]. A new classification tree was built with 98 levels of partitions. I then trimmed 96 levels off the full tree and came to a final interpretable tree, as shown in Figure 14. In this final tree, even the smallest node had a sample size of 123, and it classified the involuntarily terminated employees relatively more accurately. The sensitivity level of this final tree was 0.3726, more than 10 times larger than the sensitivity of the regular classification tree. The specificity was slightly lower (0.7915). The three terminal nodes categorized as "stayer" all had a much larger group of stayed employees compared to left employees (i.e., unbalanced nodes), further strengthening confidence in this classification decision. Three out of the original six predictors were used in the final model (i.e., Job Fit, Onboarding, and Financial Rewards), and there were four profiles of

[3] 20 was chosen because the ratio between stayed versus involuntarily left employees was roughly 20.

involuntary turnover: (1) JF < 2.9; (2) JF ≥ 2.9 & ONB < 3.055; (3) JF ≥ 2.9 & 4.435 ≤ ONB ≤ 4.565; (4) JF ≥ 2.9 & ONB ≥ 4.885 & FR < 4.365 (shown in Figure 14).

Discussion

Results of the current study were consistent with what was found in Study 1: CART tended to overfit and was most unstable with the largest cross-validation error. Random forests had the smallest resubstitution error, but showed similar level of cross-validation accuracy as the traditional methods. Moreover, the current study demonstrated that CART was most significantly impacted by number of predictors with more predictors leading to a less stable tree structure and a larger cross-validation error rate. QDA also show a slight increase in cross-validation when more predictors were included. However, the advantage of RF was more prominent when the number of predictors was larger – the more predictors, the smaller the cross-validation error. LR and LDA had the smallest cross-validation error regardless of the number of predictors, but fell short with regard to sensitivity. Therefore, if the main purpose of a study is to increase overall predictive accuracy, LR and LDA may be considered. If there are only a few predictors and the sensitivity of the prediction is of particular importance, CART and QDA are preferred. If there are a large number of predictors, RF should be applied. Note that the conclusions of the current study should be generalized with caution, given that the base rate of this sample was very small while the sample size was quite large.

With regard to variable importance rank, different techniques revealed some level of inconsistency. The difference between regression coefficients and relative weights might be attributed to predictor multicollinearity. The difference between relative weights and random forests may suggest the presence of complex interactive or higher-order effects. Random forests

tends to represent the most comprehensive and accurate picture of variable importance when multicollinearity and complex relationships exist in the data. Thus, when the relationships within the data were unclear, random forests may offer more accurate information about variable importance. Empirically, it is suggested that results from both random forests and relative weights be considered simultaneously – consistent findings between the two methods could further confirm the importance of certain variables. In addition, variables that are identified as important by random forests but not by relative weights may deserve some additional exploration and theoretical understanding.

To understand the functional relationship, a single classification tree was built based on the set of most important predictors. Results showed that when equal cost was applied, the classification tree yielded a couple of small-size or heterogeneous terminal nodes, which could impair the stability and generalizability of the decision algorithm. Berk (2008) suggested that a cost function should be introduced when constructing a decision tree, especially when one of the conditions had greater importance than the other. In the current study, correctly predicting voluntary or involuntary turnover is of particular importance. Therefore, the cost ratio was set other than 1:1. As shown in the results, introducing a cost ratio other than 1:1 inevitability increased the overall proportion of cases misclassified; however, a significant increase in sensitivity led to better identification of left employees, which was considered to be of more importance. The resulting decision tree diagrams provided a decision rule that could be used for future classification, and shed some light on the causes of voluntary and involuntary turnover. The present results showed how different configurations of work experience impacted employees' decision to stay or leave. According to the results, intent to stay was the most effective predictor of whether someone will stay or quit voluntarily. This finding is consistent with past research

that shows turnover intention is highly related to actual turnover behavior (Jackofsky, 1984; Mobley, Griffeth, Hand, & Meglino, 1979). Job satisfaction (called "Job Fit" in the data) was the best predictor for differentiating employees who stayed and those who were terminated involuntarily – that is, stayed employees usually were more satisfied. Financial reward emerged as an important factor for both cases so that employees who were not satisfied with their pay may leave, either voluntarily or involuntarily. A good working relationship with supervisor was vital to keeping employees from leaving unwantedly, and good onboarding experience and training may help employees be better equipped to handle their job and prevent them from being terminated involuntarily.

STUDY 4

The current study was designed to illustrate the empirical application of decision tree methods for predicting a continuous outcome – job performance. More specifically, nine predictors (including cognitive ability and bio-data measures) from the pre-employment selection inventory were examined and linked to supervisor-rated job performance. Similar to Study 3, the two decision tree methods were compared with the traditional linear regression approach when all, or a subset of predictors were included in the model. Additionally, a single regression tree was built based upon the most important subset of predictors, and the profiles of the high and low performers were uncovered.

Sample

The sample in Study 4 was from a concurrent validation study conducted by a large insurance company. Full information was available for 1,818 incumbents including selection test scores and their supervisor-rated job performance. These employees were from two major functional groups: 1,027 (56.49%) were from the paraprofessional/support (PPSP) group that included administrative/office/operations support, technical support and customer contact support job functions, and 791 (43.51%) were from the professional/technical (PRTC) group which included administrative/operations professional, information technical/technical professional and customer contact professional job groups. Gender and ethnicity were not available.

Measures

Background and Experience Questionnaire (BEQ). The BEQ is a biodata measure covering eight competency dimensions defined by the organization. These include: “Continuous

Learning Orientation (CLO)," "Dependability (DP)," "Developing and Maintaining Relationship (DMR)," "Flexibility (FL)," "Interpersonal Communication (IC)," "Leadership Skill (LD)," "Personal Motivation (PM)," and "Self-Management Skills (SM)." Weighted scores were used that ranged from 0 – 6.

Following Policies and Procedures (FPP). FPP is a measure of cognitive ability administered under proctored conditions. There were two different versions of the test, one for PPSP group and the other for PRTC group. Weighted scores ranged from 13.29 to 41.65.

Job Performance. The criterion measure in the current study was supervisor-rated job performance, a composite measure of 11 skill areas and 11 or 10 work behaviors (11 for PPSP and 10 for PRTC). The overall score ranged from 1 to 5.

As the original items and score calculations were kept by an external vendor, the number of items within each scale and Cronbach's alpha reliability were not available.

Data Analysis

To understand which regression method could maximize predictive accuracy, and how the nine selection predictors were related to job performance, a similar set of analyses were conducted as those in Study 3. For the first step, the three methods, including CART, random forests, and multiple linear regression, were used to fit the data. Their predictive accuracy was compared based on both resubstituition mean squared error and 10-fold cross validation mean squared error. Next, variable importance rankings were computed using the following five statistics: regression coefficients, squared regression coefficients and *p* values that were derived from multiple linear regression, relative weights that were derived from relative weights analysis, and random forests variable importance estimates. To understand how the number of predictors

would impact the predictive accuracy of various methods, eight models with the number of predictors, ranging from 2 to 9 were constructed and compared on resubstitution MSE and 10-fold cross-validation MSE. Lastly, the optimal subset of predictors was selected to build a single regression tree and then pruned to a meaningful level. The resulting decision algorithm represented the configural interactions among different selection predictors and a set of cut-off points. Cost functions are not applicable to regression trees. As Berk (2008) indicated, symmetric costs are often inappropriate for classification, yet they are required when the response variable is continuous. Thus, the cost function was not considered in the current study.

Results

Table 18 provides a summary of the descriptive statistics and intercorrelations among the 9 predictors and the criterion variable. The biodata predictors are moderately to highly intercorrelated. The Following Policies and Procedures (FPP) measure is negatively correlated with all the remaining predictors within each job family (professional vs. support); but when the two job families are combined, the overall correlation between FPP and the other selection predictors are positive. The criterion-related validity of the predictors is positive and significant, yet the absolute values are relatively small (i.e., below 0.2).

In the prediction of job performance, CART showed the smallest resubstitution MSE (0.1301), followed by random forests (0.1667), both were much smaller than the traditional linear regression (0.2509). In a comparison of cross-validation accuracies, random forests (0.2647) resulted in the most accurate prediction, followed by multiple linear regression (0.3125), and CART was least generalizable (0.4009).

Table 19 shows the results of variable importance ranking. The rank orders of variable importance derived from standardized regression coefficients and relative weights differ substantially. This finding was expected given the high intercorrelations among the predictor variables. A further comparison between relative weights and random forests showed that the top four predictors identified by relative weights analysis (i.e., FPP, DP, SM, and PM) were also among the top five from random forests analysis. One exception was CLO (Continuous Learning Orientation) – it was at the very bottom according to both multiple regression and relative weights, yet was ranked second for random forests. This inconsistency in ranking implied that the relationship between CLO and job performance might involve some higher-order or complex interaction components. As such, the subsequent feature selection process was based on the variable importance rank generated by random forests.

As shown in Table 20, CART was most sensitive to how many predictors were included in the model – the inclusion of more predictors was associated with a smaller resubstitution error rate yet a larger cross-validation error rate. The random forests method was significantly impacted by the number of predictors when examined against resubstitution error, but was generally stable with regard to cross-validation. Multiple linear regression showed the highest level of stability when the number of predictors changed. Although more predictors led to a slightly larger resubstitution error and a smaller cross-validation error, the difference was negligible. Across all three methods, random forests consistently demonstrated the highest level of cross-validation accuracy. The average cross-validation error of random forests (M = 0.2656) was 13% less than the average of multiple linear regression (M = 0.3069).

To illustrate how the set of selection predictors were related to job performance, a regression tree was built based on the five most important predictors derived from random

forests (i.e., Following Policies and Procedures, Continuous Learning Orientation, Dependability, Self-Management Skills, and Personal Motivation). The original tree was quite leafy with 189 levels. Therefore, I pruned 183 levels off for a meaningful structure, as shown in Figure 15. Only four out of the five predictors were selected in the final tree, with Continuous Learning Orientation not included. Unlike that in the classification scenario, no cost weight was applicable for regression trees. The cross-validation MSE for this tree was 0.3555. On the other hand, when the four predictors were entered into a multiple linear regression equation, the cross-validation MSE was 0.3075. Although the cross-validation error of the regression tree was slightly larger than the linear regression method, the tree structure in Figure 15 was informative for uncovering the underlying mechanism of how these predictors worked. For example, according to the decision tree, when an employee has a high cognitive ability score (i.e., FPP), he/she is very likely to perform well in the job. When an employee's cognitive ability score is relatively low, as long as he/she is high in Dependability, the performance score of this employee is still likely to be moderate to high. When an employee has low scores in both cognitive ability and Dependability, even if he/she has high Personal Motivation, this person is unlikely to be successful at work. The richness of the information offered by the decision tree diagram was beyond what could be told from a linear regression model.

Discussion

Results from Study 4 supported conclusions from the three previous studies. First, CART tended to overfit as demonstrated by a very low level of resubstituition error rate yet much higher cross-validation error rate. This pattern was strengthened when more predictors were included with the resubstituition error being even lower while cross-validation error increased. Second, random forests emerged as the most desirable approach, with the resubstituition error almost as

low as CART but cross-validation accuracy consistently higher than the other methods. The impact of predictor quantity on random forests was mainly reflected on resubstituition error with more predictors resulting in overfitting. Third, multiple linear regression displayed a relative stable pattern across changing numbers of predictors with more predictors not necessarily yielding better predictions. Fourth, variable importance rankings derived from various techniques did not quite converge, likely due to reasons such as multicollinearity or complex interactive or higher-order relationships among the predictors or between the predictors and the outcome variable. Fifth, the random forests method was able to identify a common set of important predictors as suggested by relative weights. Moreover, random forests also pointed out a unique variable (i.e., CLO) that might have been overlooked by other methods given its potential complex relationships with other variables. Finally, a simple regression tree was built based on the four most important predictors identified by both random forests and relative weights. Although its cross-validation error was slightly larger than its linear regression counterpart, the tree diagram provided some additional insights about how the selection predictors were related to job performance. As shown in Figure 15, cognitive ability (FPP) is the most important predictor, which is consistent with the conclusion from previous meta-analyses (Schmidt & Hunter, 1998). However, a low cognitive ability score does not always lead to poor performance – when an employee is high on Dependability, his/her performance may still be moderate to high.

CHAPTER 3: GENERAL DISCUSSION

The overarching goal of the current paper was to introduce two decision tree methods, Classification and Regression Trees (CART) and Random Forests (RF), and compare them with several traditional approaches through two simulations and two empirical applications. The second goal was to suggest when and how the decision tree methods should be applied.

Results from the two simulation studies showed that when resubstitution error and 10-fold cross-validation error were compared, the two decision tree methods consistently showed a large discrepancy with the error in the training sample significantly smaller than that of the test sample, yet the traditional methods (i.e., LR, LDA, QDA, or MLR) yielded similar levels of error rates. This result implies that the greedy algorithms of the decision tree methods capitalize on chance, so the resubstitution error rate is overly optimistic. Consistent with Cureton (1950), it is essential that the cross-validation error rate should be relied on to gauge the accuracy of a statistical model. A further comparison between CART and random forests indicated that CART tended to overfit with the cross-validation error substantially larger than the other methods in general, while random forests classification accuracy was not attenuated as much through cross-validation. This conclusion is consistent with previous findings that the random forests method is superior to CART with regard to stability and generalizability.

The current paper also showed that when the two decision tree methods were compared with the traditional methods on predictive accuracy (i.e., cross-validation error), the underlying structure of the data played a significant role. In general, when the data was structured in a simple linear additive manner, random forests performed comparably to the traditional methods, and CART was significantly inferior to all the other methods. However, when the data was structured in a configural tree-like pattern, both random forests and CART outperformed the

traditional methods. It should be noted that when the predicted outcome was a continuous variable, the proportion of variance explained moderated the conclusion above – the advantage of CART and random forests under the configural data structure was only prominent when the proportion of variance explained was moderate to large. This is because when predicting a continuous outcome, if the majority of the variance came from error, the data generated may represent more of a random distribution rather than a particular structure. Large error variance might be a lesser issue for predicting a binary outcome as the dichotomization process may round off the impact of the error. Additionally, it was found that when the variance explained was large, random forests was more accurate than linear regression even when the data was structured in a polynomial regression model (Y2, Y3 or Y4), and CART was better than linear regression under the most complex polynomial model (Y4).

The conclusions of the current research are not quite the same as previous simulation studies that also incorporated interaction or nonlinear relationships. A number of reasons may explain the divergence, including the difference in the set of manipulated parameters and analytic procedures. For example, the current study performed 10-fold cross-validation whereas previous simulations did not follow this approach for cross-validation. Also, the cut-off score for logistic regression was based on the ROC curve whereas previous studies used an arbitrary point of 0.5. Nevertheless, the present findings suggest that the inclusion of interaction or nonlinear effects does not always guarantee the superiority of CART. When interaction and high-order terms are included in a linear additive manner, the nonlinear model specified is nothing more than a polynomial regression. This type of relationship may still be an over simplification of the inherent structure of real data. Instead, a set of regression equations that represent multiple subgroups may better capture configural relationships in the data. This latter situation is the case

where CART and random forests show more value beyond the traditional approaches. The current findings imply that when a method is aligned with the scenario to which it best applies, its utility is maximized.

Regarding the remaining set of simulated parameters, the proportion of variance explained and the base rate both demonstrated significant main effects: a larger explained variance and a smaller base rate were associated with better predictions. In addition, interaction effects showed that the difference between various levels of the simulated parameters appeared to be more substantial for traditional methods. The other two factors, multicollinearity and skewness, did not show any significant main or interaction effects.

Study 3 and 4 provided two empirical illustrations of how decision tree methods can be used to enhance model prediction and interpretation. Similar to what was found in the two simulation studies, CART showed overfitting while random forests cross-validated well, primarily due to the model averaging feature of random forests. The number of predictors included in the model had the greatest impact on CART with more predictors leading to more serious overfitting, as reflected by the decreasing resubstitution error and increasing cross-validation error. However, the influence of predictor quantity on random forests cross-validation accuracy was in the opposite direction – more predictors slightly boosted the predictive accuracy of random forests. Unlike their decision tree counterparts, the set of traditional methods were more stable in response to a changing number of predictors, except that the classification error for QDA slightly increased when more predictors were involved.

In empirical settings, sensitivity or specificity sometimes has particular importance beyond the overall cross-validation error for a classification problem. For example, in Study 3,

identifying which employee may leave is critical for intervention purposes. When the set of methods were evaluated purely on cross-validation error, LR and LDA tended to have the best performance; yet, their overall classification accuracy was at the cost of low or near zero sensitivity. This is because when the base rate is low, simply placing everyone into the larger group (i.e., stayers) could still generate a great overall accuracy. However, this is counter to the purpose of identifying leavers for a turnover study, so these two methods may not be preferred. On the contrary, CART and QDA showed a higher sensitivity, but exhibited larger cross-validation errors when more predictors were included. Random forests showed relatively balanced cross-validation error and sensitivity, especially when the base rate was not extremely small, and the number of predictor was large.

In addition to examining the predictive accuracy of various models, the two empirical studies also introduced a process to understand the functional relationship between the predictors and the outcome, which included two steps: step one was variable selection, meaning that all the predictors were ranked according to their relative importance, generated from either random forests or relative weights. The information from both analyses was then integrated to determine the most important subset of predictors. Step two was to build and prune a single decision tree (either a classification tree or a regression tree) based on the predictors selected in step one. For classification problems, a cost function was incorporated to adjust the weights of different outcomes.

The current paper helps address two primary concerns of statistical modeling, prediction and interpretation, through the introduction of the two recursive partitioning methods. Regarding prediction, the primary concern is to minimize prediction error. Traditional multiple regression approaches fall short in two aspects: the multivariate nature of the prediction and model

uncertainty. CART is able to consider the multivariate nature of the prediction problem by design. Indeed, CART was shown to be quite sensitive to the sample used for developing the model by picking up the nuances in the data set and reclassifying at high levels of accuracy. However, unless the nature of the data is structured in a configural manner represented by a series of regression equations, the CART level of fit usually could not be carried over to a new sample, leading to a lack of generalizability. This deficiency in generalizability worsened when more predictors were included and the model was more complex. The random forests method, which shares the same mechanism of CART in dealing with the multivariate challenge, also addresses the model uncertainty issue through averaging across predictions from multiple models. Both simulation and empirical results of the current study supported this notion that RF generated similar or better cross-validation accuracy compared to traditional approaches, especially when the model was more complex than simple linear regression. This advantage can be strengthened when there are more predictors, or when the data inherently takes the form of a configural structure. As a summary, it is suggested that random forests should be considered to supplement, if not replace the traditional predictive models when the goal is to make accurate predictions, especially when there are multiple predictors and their relationship is unclear.

A second concern of predictive modeling is interpretation, which includes the understanding of the relative contribution of each variable for making the prediction, and the functional relationship between the predictors and the outcome. For the first goal, the random forests approach is able to integrate more complex interactive or higher-order effects, which may add additional value beyond both the standardized regression coefficient and relative weights approach. The current study showed that random forests and relative weights converged with respect to several important predictors, but random forests also picked out another few predictors

that might be overlooked. A similar conclusion was reached by Ostberg (2005) who found a discrepancy between the most important variables identified by regression and neural network models (another type of machine learning). Therefore, it is suggested that when the underlying structure of the data is unclear, rankings from both relative weights and random forests should be considered, compared, and integrated. Information from variable importance rankings can also be used to select the most important subset of variables for building a smaller, more stable tree structure. In terms of the second goal, the black-box approach of random forests could be supplemented by a single decision tree with the most important subset of the predictors pruned to the appropriate level. The resulting tree diagram could provide additional insights about the configural relationship between various predictors and the outcome. For example, in the case of the Study 3, the classification tree was able to portray several profiles of both voluntary and involuntary left employees. As Breiman et al. (1984, p. 265) noted, "tree structured regression offers an interesting alternative for looking at regression type problems. It has sometimes given clues to data structure not apparent from a linear regression analysis. Like any other tool, its greatest benefit lies in its intelligent and sensible application." In practice, when CART is used to understand the functional relationship between predictors and outcome variable, it is suggested that the results should be scrutinized against subject matter information to protect it from misleading interpretations (Berk, 2008). Additionally, Berk (2008) noted that a few factors might be related to unstable tree structure, such as small sample size in terminal nodes, or heterogeneous terminal nodes. Therefore, factors such as terminal node size or distribution of multiple categories should be evaluated. A cost function, which assigns differential weights for various categories, can be factored in when building a decision tree, to increase the theoretical and practical relevance of the model developed.

The current study does not suggest completely dropping regression as an important analytic method. As shown by some of the simulation and empirical results, traditional regression approaches showed merit under certain conditions, for example, when the relationships among the predictors were relatively straightforward. Whenever possible, it is suggested that a complex CART or random forests model should be compared with a simpler linear or logistic regression model, and a decision can be made about whether simpler models would be adequate. Additionally, decision tree methods can be used in combination with regression to take advantages of the strengths of both. For example, variable importance derived from relative weights analysis can be compared with the results from random forests to reinforce the importance of certain predictors.

Limitations and Future Directions

The results of the current study should be interpreted with its limitations in mind. For example, the examinations of the empirical data in Studies 3 and 4 were based on scale level rather than specific item level predictors, due to the unavailability of the item level data. As CART and random forests might show particular superiority for large-*p*-small-*n* scenarios, meaning that if more predictors were included, a study with item level data might show a more substantial difference between the decision tree methods and the traditional approaches. Future research is encouraged to replicate the current study using data sets with a larger number of predictors and relatively small sample sizes.

In addition, despite some general trends discovered in Study 1 and 2, it should be noted that the simulation results are bounded by the specific parameters manipulated, and the conclusions hold only within the ranges from which the parameter values were sampled. Further

generalization beyond the current parameters requires caution. For example, it was shown in the current study that skewness did not demonstrate any significant main or interaction effects on predictive accuracy. It was not clear if this was because skewness per se had no significant impact, or because the levels of skewness used in the current study (0.3 and 0.7) were not extreme enough to induce any large change. Similarly for Study 3, the base rate was quite small for the empirical samples. It is uncertain whether the same conclusion holds for a sample with a larger base rate. Future research might take into consideration a broader set of values, or investigate some other important parameters that are pertinent to organizational science, such as the level of range restriction.

Moreover, results from the logistic regression should be interpreted with caution because of the way the cut-off point was set. Note that the current paper applied the ROC curve approach and found the value that minimized the misclassification error to serve as the cut-off point. However, this procedure is not always followed in applied settings. Instead, an arbitrary cut-off point such as 0.5 might be chosen. Therefore, the results of the current study represent the upper bound of LR's predictive accuracy and might be overly optimistic. Future studies may try out different strategies for setting cut-off points and explore whether the results differ.

Theoretical and Practical Implications:

With the rapid growth of computer science in the past two decades, we have seen an increase in the development and application of advanced predictive modeling techniques in various fields such as applied statistics, bioinformatics, engineering, and finance. Organizational science as a field, however, has generally been slow in adopting these methods to fully take advantage of the methodological advances. In the past decade, only four papers published in

Organizational Research Methods (Detienne, Detienne & Joshi, 2003; Minbashian, Bright, & Bird, 2010; Palocsay & White, 2004; Somers & Casal, 2009) were either an introduction or application of artificial neural networks (another machine learning algorithm). No paper has introduced other popular algorithms, such as decision tree or support vector machines (SVM). Locke (2007) has argued that current researchers have placed too much emphasis on deductive approach of theory testing. Since theories in many areas of organizational science are fairly rudimentary, more studies need to be empirically driven. Instead of testing hypotheses based on not-so-valid theories, Locke (2007) suggested that an inductive approach may help develop more good theories. The empirical and inductive approach of machine learning fits this need.

Practically speaking, given the superior predictive power of random forests for complex models and a large number of predictors, it can be considered as a valuable off-the-shelf tool for exploring features of data sets. For example, in the high-stakes context of personnel selection, predictive accuracy is usually a foremost concern, and even a small improvement in predictive utility can lead to a substantial amount of monetary savings (White et al., 1993). Setting up a random forests software wizard to assist with making selection decisions may result in greater utility than would a traditional regression model. Several software packages are readily available to run these analyses, including R, MATLAB, SAS, Weka, etc.

No other topic in the talent management area has been more popular than big data and analytics in the past two years, as covered by various media including New York Times (e.g., Lohr, 2013), Harvard Business Review (McAfee & Brynjolfsson, 2012), and Forbes (Bersin, 2013). With the growing awareness by the corporate world, HR analytics has rocketed to the top of the corporate agenda for a growing number of organizations. As more and more data are being collected from various sources at different time points on an ongoing basis, there is a pressing

need for I/O Psychologists and HR practitioners to adapt to the changing nature of the incoming data in terms of volume and velocity. I/O psychologists need to better leverage the existing knowledge and best practices from business intelligence world, where artificial intelligence such as machine learning and data mining are regularly used. The current paper examined one of the machine learning methods, recursive partitioning, with the hope of bringing more attention to a broader array of predictive solutions. Future studies should explore other promising machine learning and data mining methods, such as support vector machines (SVM).

Conclusion

The current study introduced Classification and Regression Trees (CART) and random forests to the field of organizational science. Results of the current research suggest that the random forests method has promise for overcoming the methodological and practical limitations of the traditional multiple regression approach. As was shown, in linear or simple additive settings, applying random forests produces results comparable to the traditional methods; yet in complex or unknown settings, the adoption of random forests may substantially enhance predictive accuracy. Moreover, random forests provides additional insights for determining variable importance and feature selection. The combination of a single decision tree constructed by CART with a random forest can substantially enhance model interpretation, providing valuable insights regarding complex configural relationships or uncover profiles of subgroups.

REFERENCES

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648-680.

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1995). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7,* 32-39.

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review, 77,* 153-170.

Asendorpf, J. B. (2003). Head-to-head comparison of the predictive validity of personality types and dimensions. *European Journal of Personality, 17,* 327-346.

Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis and freeforward networks. *Computational Statistics, 12,* 293-310.

Arthur, W., Woehr, D. J., & Graziano, W. G. (2001). Personality testing in employment settings: Problems and issues in the application of typical selection practices. *Personnel Review, 30,* 657-676.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8,* 129-148.

Azen, R., Budescu, D. V., & Reiser, B. (2001) Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology, 54,* 201-225.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Barrick, M. R. & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Barrick, M. R., & Mount, K. M. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18,* 359-372.

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology, 9,* 291-319.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York, NY: Springer.

Bersin, J. (2013, Feb 17). *Big data in human resources: Talent analytics comes of age*. Retrieved from http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/

Bi, J., & Chung, J. (2011) Identification of drivers of overall liking: Determination of relative importance of regressor variables. *Journal of Sensory Studies, 26,* 245-254.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press.

Blickle, G., Meurs, J. A., Zettler, I., Solga, J., Noethen, D., Kramer, J., & Ferris, G. R. (2008). Personality, political skill, and job performance. *Journal of Vocational Behavior, 72,* 377-387.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561-589.

Breiman, L. (1996) Bagging Predictors. *Machine Learning, 24,* 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning. 45,* 5-32.

Breiman, L. Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114,* 542-551.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., & Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol, 28,* 171-182.

Chien, C. & Chen, L. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34,* 280-290.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Crews, H. (2008). *Fast FSR methods for second-order linear regression models.* (Unpublished Doctoral Dissertation). North Carolina State University, Raleigh, NC.

Cronbach, L. J., & Glaser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456-473.

Cucina, J. M., & Vasilopoulos, N. L. (2005). Nonlinear personality-performance relationships and the spurious moderating effects of traitedness. *Journal of Personality, 73,* 227-260.

Cureton, E. D. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10,* 94-96.

Curram, S. P., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *Journal of the Operational Research Society, 45,* 440-450.

Davison, M. L. & Kuang, H. (2000). Profile patterns: Research and professional interpretation. *School Psychology Quarterly, 15,* 457-464.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25-36.

De'ath, G. & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81*, 3178-3192.

Detienne, K. B., Detienne, D. H., & Joshi, S. A. (2003). Neural networks as statistical tools for business researchers. *Organizational Research Methods, 6*, 236-265.

De Corte, W., Lievens, F., & Sackett, P. R. (2006) Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology, 91*, 523-537.

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decision required for selection system development. *Journal of Applied Psychology, 96*, 907-926.

Doty, D. H., & Glick, W. H. (1994). Typologies as a unique form of theory building: Toward improved understanding and model. *Academy of Management Review, 19*, 230-251.

Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly, 24*, 465-467.

Dudoit, S., Fridlyand, D., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77-87.

Edwards, J. R. & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*, 1577-1613.

Eliot, M. (2011) *Tree-based methods and a mixed ridge estimator for analyzing longitudinal data with correlated predictors.* (Unpublished Doctoral Dissertation). University of Massachusetts Amherst, Amherst, MA.

Fan, X., & Wang, L. (1999) Comparing linear discriminant function with logistic regression for the two-group classification problem. *Journal of Experimental Education, 67*, 265-286.

Finch, W. H., Chang, M., Davis, A. S., Holden, J. E., Rothlisberg, B. A., & McIntosh, D. E. (2011) The prediction of intelligence in preschool children using alternative models to regression. *Behavior Research Methods, 43*, 942-952.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009) Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology, 94*, 318-340.

Finch, W. H., & Schneider, M. K. (2006) Misclassification rates for four methods of group classification: Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. *Educational and Psychological Measurement, 66*, 240-257.

Finch, H., & Schneider, M. K. (2007) Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression tress: Three- and give-group cases. *Methodology, 3,* 47-57.

Frydman, H., Altman, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance, 40,* 269-291.

Fu, S., Anderson, D., Courtney, M., & Hu, W. (2007). The relationship between culture, attitude, social networks and quality of life in midlife Australian and Taiwanese citizens. *Maturitas, 58,* 285-295.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7,* 385-400.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists.* London: Sage.

Gjurich, G. D. (1999). *A predictive model of surface warfare officer retention: Factors affecting turnover*. (Unpublished Master Thesis). Naval Postgraduate school, Monterey, CA.

Grant, A. M., & Schwartz, B. (2011). Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on Psychological Science, 6,* 61-76.

Grassi, M., Villani, S., & Marinoni, A. (2001). Classification methods for the identification of "case" in epidemiological diagnosis of asthma. *European Journal of Epidemiology, 17,* 19-29.

Gruenewald, T. L., Mroczek, D. K., Ryff, C. D., & Singer, B. H. (2008). Diverse pathways to positive and negative affect in adulthood and later life: An integrative approach using recursive partitioning. *Developmental Psychology, 44,* 330-343.

Guion, R. M. (1991). Personnel assessment, selection, and placement. *Handbook of Industrial and Organizational Psychology, 2,* 327-397.

Halstead, J. B. (2006). *Recruiter Selection Model.* Operations research center of excellence technical report NO: DSE-TR-0623. DTIC #: ADA443651.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* New York, NY: Springer.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decision: Questions and answers. *American Psychologist, 51,* 469-477.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement, 71,* 870-901.

Hollenbeck, J. R., Brief, A. P., Whitener, E. M., & Pauli, K. E. (1988). An empirical note on the interaction of personality and aptitude in personnel selection. *Journal of Management, 14,* 441-451.

Horst, P. (1954). Pattern analysis and configural scoring. *Journal of Clinical Psychology, 10,* 3-11.

Horst, P. (1968). Configural analysis and pattern recognition. *Journal of Clinical Psychology, 24,* 383-405.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*, 651-674.

Jackofsky, E. F. (1984). Turnover and Job Performance: An Integrated Process Model. *The Academy of Management Review, 9*, 74-83.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Second Edition* (102-138). New York, NY: The Guilford Press.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regressions. *Multivariate Behavioral Research, 35*, 1-19.

Johnson, J. W., & LeBreton, J. M. (2004) History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods, 7*, 238-257.

Judge, T. A., & Erez, A. (2007). Interaction and intersection: The constellation of emotional stability and extraversion in predicting performance. *Personnel Psychology, 60*, 573-395.

Keeney, M. J., Snell, A. F., Robison, S. J., Svyantek, D. V., & Bott, J. (2004). Personality and situational pattern differences across three work groups: A comparative examination of worker personality and organizational climate using three pattern-extraction analyses. *Organizational Analysis, 12*, 183-203.

Kerby, D. S. (2003). CART analysis with unit-weighted regression to predict suicidal ideation from Big Five traits. *Personality and Individual Differences, 35*, 249-261.

Kiernan, M., Kraemer, H. C., Winkleby, M. A., King, A. C., & Taylor, C. B. (2001). Do logistic regression and signal detection theory identify different subgroups at risk? Implications for the design of tailored interventions. *Psychological Methods, 6*, 35-48.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of Conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*, 140-154.

LaHuis, D. M., Martin, N. R., & Avis, J. M. (2005). Investigating nonlinear conscientiousness-job performance relations for clerical employees. *Human Performance, 18*, 199-212.

Le, H., Oh, I., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*, 113-133.

LeBreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods, 7,* 258-282.

LeBreton, J. M., Tonidandel, S., & Krasikova, D. V. (2013). Residualized relative importance analysis: A technique for the comprehensive decomposition of variance in higher order regression models. *Organizational Research Methods*, *16*, 449-473.

Lee, M. C. (1961). Interactions, configurations, and nonadditive models. *Educational and Psychological Measurement, 21,* 797-805.

Lee, W. C. & Drasgow, F. (working paper), Predicting employee retention: An examination of a retention likelihood scale.

Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification. *Journal of Experimental Education, 72,* 25-49.

Lloyd, M. D. (2008). *Incremental validity of commonly-used risk assessment measures in predicting serious sexual recidivism.* (Unpublished Doctoral Dissertation). University of Minnesota, Two Cities, MN.

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management, 33,* 867-890.

Loh, W. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*, 361-386.

Loh, W. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics, 3*, 1710-1737.

Lohr, S. (2013, April 20). *Big data, trying to build better workers*. Retrieved from http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html?pagewanted=all&_r=2&

Lord, F. M.., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Oxford England: Addison-Wesley.

Lynch, J. G. (1985). Uniqueness issues in the decompositional modeling of multiattribute overall evaluations: An information integration perspective. *Journal of Marketing Research, 22*, 1-19.

Maertz, C. P., & Campion, M. A. (2004). Profiles in quitting: Integrating process and content turnover theory. *Academy of Management Journal, 47*, 566-582.

Maier, N. R. F. (1958). *The appraisal interview: Objectives, methods and skills*. Oxford, Engand: Wiley.

Magnusson, D., & Törestad, B. (1993). A holistic view of personality: A model revisited. *Annual Review of Psychology, 44*, 427-452.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The management revolution. *Harvard Business Review, 90,* 60-68.

Merkle, E. C. & Shaffer, V. A. (2011). Binary recursive partitioning: Background methods, and application to psychology. *British Journal of Mathematical and Statistical Psychology, 64*, 161-181.

Minbashian, A., Bright, J. E. H., & Bird, K. D. (2010). A comparison of artificial neural networks and multiple regression in the context of research on personality and work performance. *Organizational Research Methods, 13,* 540-561.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin, 86,* 493-522.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selsection contexts. *Personnel Psychology, 60,* 683-729.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11,* 145-165.

Mount, M. K., Oh, I., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology, 61,* 113-139.

Murphy, K. R. (1996). Individual differences and behavior in organizations: Much more than g. In K. Murphy (Ed), *Individual differences and behavior in organizations* (pp. 3-30). San Francisco: Jossey-Bass.

Murphy, K. R. & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance, 18,* 343-357.

Numpacharoen, K., & Atsawarungruangkit, A. (2012, Oct 1). *A. generating random correlation matrices based on the boundaries of their coefficients*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2127689

Ones, D. S., Viswesvaran, C., Dilchert, S., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995-1027.

Ostberg, D. E. (2005). *A comparative analysis of artificial neutral networks, classification trees, and multivariate linear regression for predicting retail employee tenure and turnover*. (Unpublished Doctoral Dissertation). Portland State University, Portland, OR.

Oswald, F. L. & Hough, L. M. (2010). Personality and its assessment in organizations: Theoretical and empirical developments. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 153–184). Washington, DC: American Psychological Association.

Palocsay, S. W., & White, M. M. (2004). Neural network modeling in cross-cultural research: A comparison with multiple regression. *Organizational Research Methods, 7,* 389-399.

Perkins, A. M., & Corr, P. J. (2005). Cognitive ability as a buffer to neuroticism: Churchill's secret weapon? *Personality and Individual Differences, 40,* 39-51.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Three induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research, 4,* 211-255.

Perry, S. J., Hunter, E. M., Witt, L. A., & Harris, K. J. (2010). P = f (Conscientiousness x Ability): Examining the facets of conscientiousness. *Human Performance, 23,* 343-360.

Petway, K. T. (2010). *Applying adaptive methods and classical scale reduction techniques to data from the Big Five inventory.* (Unpublished Master Thesis). University of Southern California, Los Angeles, CA.

Pierce, J. R., & Aguinis, H. (2013). The too-much-of-a-good-thing effect in management. *Journal of Management, 39,* 313-338.

Preatoni, D. G., Nodari, M., Chirchella, R., Tosi, G., Wauters, L. A., & Martinoli, A. (2005). Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management, 69,* 1601-1614.

Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G., & Wachter, H. (1991). Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. *PNAS, 88,* 11426-11430.

Reis, P. M. (2000). *Determinants of flight training performance: An analysis of the impact of undergraduate academic background.* (Unpublished Master Thesis). Naval Postgraduate School, Monterey, CA.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological), 3*, 409-456.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98,* 598-616.

Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *International Journal of Selection and Assessment, 7,* 157-169.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634-644.

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 549-572.

Sah, K. I. (1998). *Binary decision tress in the strategic management of college admissions.* (Unpublished Doctoral Dissertation). Arizona State University, Phoenix, AZ.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmitt, N., Oswald, F. L., Kim, B. H., Imus, A., Merritt, S., Friede, A., & Shivpuri, S. (2007). The use of background and ability profile to predict college student outcomes. *Journal of Applied Psychology, 92,* 165-179.

Scott, S. B., Jackson, B. R., & Bergeman, C. S. (2011). What contributes to perceived stress in later life? A recursive partitioning approach. *Psychology and Aging, 26,* 830-843.

Shen, W. (2011). *The application of a person-oriented criterion-related configural approach to the relationship between personality traits and work behaviors.* (Unpublished Doctoral Dissertation). University of Minnesota, Twin Cities, MI.

Sinclair, R.R., & Lyne, R. (1997). *Non-linearity in personality-performance relations: Models, methods, and initial evidence.* In J. Hogan (Chair), Non-cognitive measures of job performance. Symposium conducted at the annual conference of the Southwestern Psychological Association, Fort Worth, TX.

Somers, M. J., & Casal, J. C. (2009). Using artificial neural networ12ks to model nonlinearity: The case of the job satisfaction – job performance relationship. *Organizational Research Methods, 12,* 403-417.

Stark, S., Chernyshenko, O. S., Drasgow, F., Lee, W. C., White, L. A., & Young, M. C. (2011). Optimizing prediction of attrition with the U.S. Army's assessment of individual motivation (AIM). *Military Psychology, 23,* 180-201.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H., & Banks, S. (2000). A classification tree approach to the development of the actuarial violence risk assessment tools. *Law and Human Behavior, 24,* 83-100.

Strobl, C., Boulesteix, AL., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustriations, sources, and a solution. *BMC Bioinformatics, 8,* 25.

Strobl, C., Boulesteix, AL., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9,* 307-317.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323-348.

Sut, N., & Simsek, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications, 38,* 15534-15539.

Tonidandel, S., & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weights analysis. *Organizational Research Methods, 13,* 767-781.

Vasilopoulos, N. L., Cucina, J. M., & Hunter, A. E. (2007). Personality and training proficiency: Issues of bandwidth-fidelity and curvilinearity. *Journal of Occupational and Organizational Psychology, 80,* 109-131.

Weiss, H. M., & Rupp, D. E. (2011). Experiencing work: An essay on a person-centric work psychology. *Industrial and Organizational Psychology, 4,* 82-97.

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science, 16,* 370-391.

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L. H. (1999). A comparison of statistical methods for prenatal screening for down syndrome. *Applied Stochastic Models in Business and Industry, 15,* 89-101.

Windham, P. W. (1994). *Factors influencing student attrition in public community colleges.* (Unpublished Doctoral Dissertation). Florida State University, Tallahassee, FL.

Witt, L. A. (2002). The interactive effects of extraversion and conscientiousness on performance. *Journal of Management, 28,* 835-851.

Witt, L. A., Burke, L. A., Barrick, M. R., & Mount, M. K. (2002). The interactive effects of conscientiousness and agreeableness on job performance. *Journal of Applied Psychology, 87,* 164-169.

Witt, L. A., & Ferris, G. R. (2003). Social skill as moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology, 88,* 809-820.

Whetzel, D. L., McDaniel, M. A., Yost, A. P., & Kim, N. (2010). Linearity of personality-performance relationships: A large-scale examination. *International Journal of Selection and Assessment, 18,* 310-320.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). The assessment of background and life experiences (ABLE). In T. Trent & J. H. Laurence (Eds), *Adaptability screening for the armed forces* (pp. 101-162). Washington DC: Office of the Assistant Secretary of Defense.

Woo, S. E. (2009). *Dispositional differences in frequencies of and reasons for voluntary turnover: A latent-class approach.* (Unpublished Doctoral Dissertation). University of Illinois, Urbana-Champaign, IL.

Wright, P. M., Kacmar, K. M., McMahan, G. C., & Deleeuw, K. (1995). P=f(M × A): Cognitive ability as a moderator of the relationship between personality and job performance. *Journal of Management, 21,* 1129-1139.

TABLES

Table 1

Simulation Conditions

Parameter	Levels
Multicollinearity	0.0 – 0.3, 0.3 – 0.7
Level of Skewness	-0.7, -0.3, 0.3, 0.7
Base Rate (classification only)	0.1, 0.3
Proportion of Variance Explained	0.09, 0.49, 0.90 (for regression only)
Data Complexity	Y1: Simple linear regression
	Y2: Polynomial regression with quadratic term
	Y3: Polynomial regression with interaction term
	Y4: Polynomial regression with complex nonlinear terms
	Y5: Tree-structured stepwise regressions for binary outcome
	Y6: Tree-structured conditional regression for continuous outcome
Prediction Method	Classification: CART, RF, LR, LDA, QDA.
	Regression: CART, RF, GLM.

Table 2

Resubstitution Error Rate for Predicting Binary Outcomes (Study 1)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
Y1	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.1001	0.1000	0.1000	0.0488	0.0000
				-0.3	0.0994	0.1000	0.0999	0.0449	0.0000
				0.3	0.0988	0.1000	0.1000	0.0441	0.0001
				0.7	0.0988	0.0993	0.0998	0.0493	0.0002
			0.3	-0.7	0.2940	0.2987	0.2976	0.1051	0.0000
				-0.3	0.2905	0.2959	0.2920	0.1006	0.0000
				0.3	0.2900	0.2941	0.2931	0.1040	0.0000
				0.7	0.2915	0.2967	0.2952	0.1048	0.0000
		0.7*0.7	0.1	-0.7	0.0952	0.0989	0.1036	0.0385	0.0001
				-0.3	0.0910	0.0954	0.0953	0.0399	0.0002
				0.3	0.0871	0.0906	0.0899	0.0353	0.0000
				0.7	0.0783	0.0822	0.0861	0.0335	0.0001
			0.3	-0.7	0.2131	0.2194	0.2319	0.0798	0.0000
				-0.3	0.2063	0.2125	0.2132	0.0818	0.0000
				0.3	0.2063	0.2118	0.2129	0.0763	0.0000
				0.7	0.1903	0.1979	0.2013	0.0721	0.0000
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.1007	0.1000	0.1000	0.0495	0.0002
				-0.3	0.0983	0.1000	0.0999	0.0478	0.0000
				0.3	0.0982	0.1001	0.0997	0.0462	0.0000
				0.7	0.0987	0.0995	0.1003	0.0450	0.0000
			0.3	-0.7	0.2926	0.2978	0.2974	0.1030	0.0000
				-0.3	0.2867	0.2946	0.2944	0.1006	0.0000
				0.3	0.2882	0.2933	0.2909	0.1032	0.0000
				0.7	0.2853	0.2898	0.2897	0.1064	0.0000
		0.7*0.7	0.1	-0.7	0.0950	0.0994	0.1024	0.0388	0.0001
				-0.3	0.0906	0.0958	0.0953	0.0374	0.0001
				0.3	0.0870	0.0897	0.0889	0.0359	0.0000
				0.7	0.0816	0.0850	0.0869	0.0336	0.0001

Table 2 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
			0.3	-0.7	0.2042	0.2121	0.2226	0.0739	0.0000
				-0.3	0.1990	0.2036	0.2071	0.0726	0.0000
				0.3	0.1977	0.2059	0.2034	0.0750	0.0000
				0.7	0.1968	0.2011	0.2073	0.0753	0.0000
Y2	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.0990	0.1000	0.0996	0.0459	0.0002
				-0.3	0.0992	0.1000	0.0993	0.0496	0.0000
				0.3	0.0972	0.0998	0.0982	0.0448	0.0000
				0.7	0.0981	0.1001	0.0995	0.0457	0.0001
			0.3	-0.7	0.2936	0.2975	0.2957	0.1002	0.0001
				-0.3	0.2892	0.2951	0.2863	0.1012	0.0000
				0.3	0.2897	0.2942	0.2858	0.1000	0.0000
				0.7	0.2863	0.2921	0.2891	0.1053	0.0002
		0.7*0.7	0.1	-0.7	0.0933	0.0990	0.0947	0.0367	0.0001
				-0.3	0.0879	0.0949	0.0901	0.0384	0.0000
				0.3	0.0797	0.0843	0.0792	0.0314	0.0000
				0.7	0.0693	0.0730	0.0720	0.0275	0.0001
			0.3	-0.7	0.2367	0.2427	0.2393	0.0725	0.0000
				-0.3	0.2259	0.2333	0.2150	0.0797	0.0000
				0.3	0.2080	0.2133	0.2005	0.0746	0.0000
				0.7	0.1969	0.2018	0.1956	0.0700	0.0000
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.0991	0.1000	0.1000	0.0440	0.0001
				-0.3	0.0983	0.0999	0.0998	0.0479	0.0000
				0.3	0.0972	0.0994	0.0991	0.0467	0.0000
				0.7	0.0987	0.1002	0.1003	0.0461	0.0000
			0.3	-0.7	0.2900	0.2967	0.2963	0.1032	0.0000
				-0.3	0.2849	0.2934	0.2921	0.0985	0.0000
				0.3	0.2828	0.2890	0.2835	0.1021	0.0000
				0.7	0.2789	0.2846	0.2830	0.1028	0.0000
		0.7*0.7	0.1	-0.7	0.0908	0.0984	0.0984	0.0360	0.0001
				-0.3	0.0787	0.0859	0.0832	0.0303	0.0000

Table 2 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
				0.3	0.0736	0.0791	0.0773	0.0291	0.0000
				0.7	0.0695	0.0724	0.0735	0.0295	0.0000
			0.3	-0.7	0.2043	0.2085	0.2190	0.0749	0.0001
				-0.3	0.1986	0.2072	0.2022	0.0742	0.0000
				0.3	0.2030	0.2097	0.2048	0.0746	0.0000
				0.7	0.1936	0.2005	0.1993	0.0692	0.0000
Y3	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.0997	0.1000	0.0999	0.0462	0.0003
				-0.3	0.0992	0.0999	0.0999	0.0458	0.0002
				0.3	0.0986	0.0999	0.0986	0.0456	0.0001
				0.7	0.0973	0.0989	0.0989	0.0479	0.0000
			0.3	-0.7	0.2864	0.2933	0.2921	0.1011	0.0000
				-0.3	0.2895	0.2942	0.2913	0.1013	0.0000
				0.3	0.2869	0.2920	0.2873	0.1029	0.0001
				0.7	0.2858	0.2905	0.2893	0.0993	0.0000
		0.7*0.7	0.1	-0.7	0.0855	0.0941	0.0945	0.0347	0.0002
				-0.3	0.0814	0.0889	0.0864	0.0345	0.0001
				0.3	0.0792	0.0836	0.0818	0.0315	0.0000
				0.7	0.0692	0.0721	0.0712	0.0289	0.0000
			0.3	-0.7	0.2049	0.2136	0.2113	0.0785	0.0000
				-0.3	0.2086	0.2145	0.2080	0.0783	0.0002
				0.3	0.1955	0.2015	0.1948	0.0749	0.0001
				0.7	0.1960	0.2016	0.1988	0.0693	0.0000
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.0995	0.1000	0.0999	0.0476	0.0000
				-0.3	0.0997	0.1000	0.1000	0.0453	0.0003
				0.3	0.0982	0.0995	0.0996	0.0441	0.0000
				0.7	0.0971	0.0994	0.1000	0.0462	0.0000
			0.3	-0.7	0.2869	0.2953	0.2906	0.1027	0.0001
				-0.3	0.2813	0.2899	0.2865	0.0996	0.0000
				0.3	0.2801	0.2856	0.2828	0.0999	0.0002
				0.7	0.2819	0.2865	0.2855	0.1029	0.0000

Table 2 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
		0.7*0.7	0.1	-0.7	0.0806	0.0943	0.0917	0.0330	0.0000
				-0.3	0.0785	0.0840	0.0830	0.0310	0.0000
				0.3	0.0763	0.0798	0.0799	0.0331	0.0000
				0.7	0.0716	0.0751	0.0758	0.0314	0.0000
			0.3	-0.7	0.1896	0.1961	0.2098	0.0708	0.0000
				-0.3	0.1857	0.1927	0.1923	0.0706	0.0001
				0.3	0.1897	0.1962	0.1939	0.0740	0.0001
				0.7	0.1960	0.2019	0.2049	0.0739	0.0000
Y4	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.0999	0.1000	0.1000	0.0456	0.0000
				-0.3	0.0989	0.1000	0.0997	0.0466	0.0001
				0.3	0.0980	0.0999	0.0996	0.0446	0.0000
				0.7	0.0956	0.0996	0.0968	0.0434	0.0000
			0.3	-0.7	0.2948	0.2981	0.2969	0.1037	0.0000
				-0.3	0.2898	0.2974	0.2939	0.0998	0.0000
				0.3	0.2837	0.2898	0.2870	0.1003	0.0000
				0.7	0.2875	0.2915	0.2920	0.0995	0.0000
		0.7*0.7	0.1	-0.7	0.0964	0.0998	0.0986	0.0420	0.0000
				-0.3	0.0872	0.0946	0.0926	0.0360	0.0000
				0.3	0.0759	0.0792	0.0746	0.0331	0.0001
				0.7	0.0743	0.0776	0.0712	0.0262	0.0001
			0.3	-0.7	0.2542	0.2596	0.2649	0.0874	0.0000
				-0.3	0.2421	0.2477	0.2409	0.0871	0.0000
				0.3	0.2284	0.2345	0.2220	0.0835	0.0001
				0.7	0.2235	0.2308	0.2218	0.0837	0.0000
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.0998	0.1000	0.1000	0.0484	0.0001
				-0.3	0.0988	0.1000	0.0999	0.0439	0.0001
				0.3	0.0965	0.0995	0.0983	0.0466	0.0001
				0.7	0.0956	0.0992	0.0972	0.0450	0.0000
			0.3	-0.7	0.2910	0.2999	0.2986	0.1043	0.0000
				-0.3	0.2853	0.2926	0.2899	0.1042	0.0000

Table 2 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
				0.3	0.2821	0.2868	0.2838	0.1008	0.0000
				0.7	0.2792	0.2854	0.2873	0.1055	0.0001
		0.7*0.7	0.1	-0.7	0.0845	0.0967	0.0917	0.0368	0.0000
				-0.3	0.0747	0.0836	0.0813	0.0296	0.0002
				0.3	0.0725	0.0784	0.0760	0.0315	0.0001
				0.7	0.0685	0.0738	0.0720	0.0266	0.0001
			0.3	-0.7	0.2304	0.2365	0.2397	0.0814	0.0000
				-0.3	0.2207	0.2268	0.2263	0.0831	0.0000
				0.3	0.2177	0.2262	0.2249	0.0838	0.0001
				0.7	0.2151	0.2208	0.2215	0.0828	0.0000
Y5	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.0589	0.0920	0.0698	0.0010	0.0000
				-0.3	0.0624	0.0734	0.0612	0.0021	0.0000
				0.3	0.0613	0.0644	0.0540	0.0020	0.0000
				0.7	0.0682	0.0725	0.0527	0.0025	0.0000
			0.3	-0.7	0.1572	0.1656	0.1241	0.0009	0.0000
				-0.3	0.1489	0.1573	0.1031	0.0030	0.0000
				0.3	0.1294	0.1342	0.1028	0.0037	0.0000
				0.7	0.1159	0.1198	0.0961	0.0011	0.0000
		0.7*0.7	0.1	-0.7	0.0589	0.0920	0.0698	0.0010	0.0000
				-0.3	0.0716	0.0795	0.0692	0.0029	0.0000
				0.3	0.0634	0.0672	0.0547	0.0034	0.0000
				0.7	0.0668	0.0701	0.0533	0.0026	0.0000
			0.3	-0.7	0.1527	0.1600	0.1283	0.0018	0.0000
				-0.3	0.1504	0.1579	0.1079	0.0033	0.0000
				0.3	0.1294	0.1342	0.1028	0.0037	0.0000
				0.7	0.1159	0.1198	0.0961	0.0011	0.0000
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.0463	0.0967	0.0560	0.0016	0.0000
				-0.3	0.0453	0.0648	0.0445	0.0017	0.0000
				0.3	0.0512	0.0567	0.0477	0.0014	0.0000
				0.7	0.0500	0.0515	0.0416	0.0022	0.0000

Table 2 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
			0.3	-0.7	0.0703	0.0790	0.0597	0.0026	0.0000
				-0.3	0.1086	0.1140	0.0841	0.0035	0.0000
				0.3	0.1013	0.1058	0.0822	0.0021	0.0000
				0.7	0.1233	0.1292	0.1084	0.0030	0.0000
		0.7*0.7	0.1	-0.7	0.0286	0.0955	0.0225	0.0005	0.0000
				-0.3	0.0335	0.0601	0.0326	0.0011	0.0000
				0.3	0.0376	0.0447	0.0358	0.0015	0.0000
				0.7	0.0394	0.0431	0.0353	0.0021	0.0000
			0.3	-0.7	0.1093	0.1204	0.0986	0.0013	0.0000
				-0.3	0.1086	0.1140	0.0841	0.0035	0.0000
				0.3	0.1293	0.1331	0.0997	0.0033	0.0000
				0.7	0.1217	0.1285	0.0992	0.0040	0.0000

Table 3

10-fold Cross-Validation Error Rate for Predicting Binary Outcomes (Study 1)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
Y1	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.1015	0.1000	0.1000	0.1704	0.1054
				-0.3	0.1015	0.1000	0.1001	0.1696	0.1061
				0.3	0.1010	0.1002	0.1015	0.1675	0.1063
				0.7	0.1004	0.0997	0.1028	0.1686	0.1065
			0.3	-0.7	0.3003	0.3005	0.3013	0.3974	0.3356
				-0.3	0.2981	0.2983	0.2982	0.3927	0.3313
				0.3	0.2963	0.2969	0.3012	0.3926	0.3297
				0.7	0.2986	0.2994	0.3041	0.4053	0.3411
		0.7*0.7	0.1	-0.7	0.0981	0.0991	0.1072	0.1550	0.1071
				-0.3	0.0944	0.0959	0.0975	0.1458	0.1048
				0.3	0.0909	0.0921	0.0959	0.1383	0.1016
				0.7	0.0827	0.0842	0.0897	0.1243	0.0956
			0.3	-0.7	0.2174	0.2202	0.2344	0.2945	0.2386
				-0.3	0.2141	0.2148	0.2193	0.2911	0.2416
				0.3	0.2141	0.2154	0.2185	0.2899	0.2401
				0.7	0.1977	0.2010	0.2063	0.2743	0.2228
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.1008	0.1000	0.1000	0.1724	0.1046
				-0.3	0.1009	0.1000	0.1000	0.1681	0.1057
				0.3	0.1002	0.1004	0.1019	0.1642	0.1074
				0.7	0.1009	0.0999	0.1037	0.1701	0.1045
			0.3	-0.7	0.2998	0.3003	0.3037	0.3940	0.3344
				-0.3	0.2954	0.2951	0.2999	0.3926	0.3289
				0.3	0.2952	0.2956	0.3003	0.4006	0.3270
				0.7	0.2931	0.2945	0.2999	0.4044	0.3255
		0.7*0.7	0.1	-0.7	0.0989	0.0995	0.1050	0.1531	0.1105
				-0.3	0.0950	0.0969	0.0980	0.1445	0.1061
				0.3	0.0911	0.0920	0.0926	0.1342	0.1018
				0.7	0.0871	0.0871	0.0911	0.1233	0.0947

Table 3 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
			0.3	-0.7	0.2114	0.2151	0.2276	0.2821	0.2376
				-0.3	0.2039	0.2065	0.2121	0.2785	0.2322
				0.3	0.2035	0.2074	0.2094	0.2811	0.2295
				0.7	0.2040	0.2043	0.2122	0.2760	0.2256
Y2	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.1017	0.1000	0.1002	0.1645	0.1037
				-0.3	0.1009	0.1000	0.0995	0.1660	0.1066
				0.3	0.1005	0.0997	0.1000	0.1596	0.1053
				0.7	0.0995	0.1004	0.1035	0.1669	0.1079
			0.3	-0.7	0.2977	0.2981	0.3013	0.3998	0.3324
				-0.3	0.2946	0.2975	0.2946	0.3924	0.3240
				0.3	0.2966	0.2969	0.2957	0.3895	0.3226
				0.7	0.2927	0.2957	0.2996	0.3869	0.3162
		0.7*0.7	0.1	-0.7	0.0967	0.0993	0.0969	0.1325	0.1011
				-0.3	0.0908	0.0952	0.0926	0.1325	0.0961
				0.3	0.0842	0.0856	0.0819	0.1216	0.0862
				0.7	0.0744	0.0745	0.0749	0.1089	0.0796
			0.3	-0.7	0.2430	0.2460	0.2434	0.2938	0.2460
				-0.3	0.2323	0.2375	0.2224	0.2955	0.2396
				0.3	0.2131	0.2146	0.2071	0.2716	0.2212
				0.7	0.2040	0.2048	0.2003	0.2671	0.2144
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.1009	0.1000	0.1000	0.1726	0.1057
				-0.3	0.0998	0.1000	0.0999	0.1658	0.1067
				0.3	0.0997	0.0997	0.1008	0.1709	0.1054
				0.7	0.1008	0.1003	0.1034	0.1638	0.1067
			0.3	-0.7	0.2967	0.2979	0.3000	0.3967	0.3341
				-0.3	0.2916	0.2966	0.2972	0.3796	0.3172
				0.3	0.2923	0.2902	0.2924	0.3950	0.3276
				0.7	0.2871	0.2850	0.2928	0.3888	0.3168
		0.7*0.7	0.1	-0.7	0.0937	0.0985	0.1028	0.1328	0.1042
				-0.3	0.0840	0.0871	0.0867	0.1207	0.0936

Table 3 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
				0.3	0.0782	0.0802	0.0806	0.1171	0.0841
				0.7	0.0744	0.0753	0.0762	0.1048	0.0753
			0.3	-0.7	0.2109	0.2113	0.2219	0.2717	0.2285
				-0.3	0.2038	0.2082	0.2060	0.2692	0.2205
				0.3	0.2071	0.2112	0.2101	0.2845	0.2315
				0.7	0.1996	0.2019	0.2034	0.2691	0.2168
Y3	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.1025	0.1000	0.1001	0.1726	0.1052
				-0.3	0.1015	0.0998	0.1005	0.1669	0.1067
				0.3	0.1009	0.0998	0.0999	0.1729	0.1061
				0.7	0.1003	0.0990	0.1015	0.1675	0.1076
			0.3	-0.7	0.2923	0.2957	0.2975	0.3920	0.3231
				-0.3	0.2969	0.2966	0.2962	0.3908	0.3237
				0.3	0.2954	0.2954	0.2971	0.3943	0.3291
				0.7	0.2925	0.2943	0.2983	0.3950	0.3333
		0.7*0.7	0.1	-0.7	0.0893	0.0939	0.0964	0.1322	0.1012
				-0.3	0.0855	0.0900	0.0889	0.1279	0.0912
				0.3	0.0844	0.0846	0.0854	0.1199	0.0918
				0.7	0.0725	0.0738	0.0744	0.1063	0.0808
			0.3	-0.7	0.2111	0.2142	0.2163	0.2847	0.2291
				-0.3	0.2149	0.2167	0.2123	0.2846	0.2255
				0.3	0.2018	0.2042	0.1997	0.2748	0.2189
				0.7	0.2019	0.2054	0.2064	0.2698	0.2188
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.1014	0.1000	0.1000	0.1661	0.1048
				-0.3	0.1011	0.1000	0.1003	0.1686	0.1057
				0.3	0.1009	0.0999	0.1015	0.1738	0.1060
				0.7	0.1007	0.0996	0.1027	0.1715	0.1057
			0.3	-0.7	0.2930	0.2970	0.2964	0.4039	0.3284
				-0.3	0.2907	0.2897	0.2921	0.3914	0.3195
				0.3	0.2884	0.2892	0.2911	0.3890	0.3212
				0.7	0.2894	0.2882	0.2928	0.3905	0.3219

Table 3 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
		0.7*0.7	0.1	-0.7	0.0847	0.0940	0.0954	0.1227	0.0930
				-0.3	0.0831	0.0849	0.0867	0.1257	0.0875
				0.3	0.0800	0.0822	0.0831	0.1178	0.0880
				0.7	0.0748	0.0764	0.0807	0.1148	0.0847
			0.3	-0.7	0.1952	0.1969	0.2136	0.2673	0.2229
				-0.3	0.1935	0.1942	0.1978	0.2671	0.2129
				0.3	0.1962	0.1982	0.1982	0.2660	0.2167
				0.7	0.2017	0.2035	0.2107	0.2715	0.2288
Y4	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.1024	0.1000	0.1000	0.1695	0.1048
				-0.3	0.1014	0.1000	0.0999	0.1769	0.1053
				0.3	0.1010	0.1000	0.1014	0.1643	0.1064
				0.7	0.1003	0.0996	0.0996	0.1615	0.1022
			0.3	-0.7	0.2995	0.2987	0.3028	0.4006	0.3345
				-0.3	0.2972	0.2986	0.3005	0.4043	0.3275
				0.3	0.2904	0.2923	0.2944	0.4055	0.3254
				0.7	0.2971	0.2941	0.3002	0.3937	0.3282
		0.7*0.7	0.1	-0.7	0.1009	0.0999	0.0997	0.1500	0.1078
				-0.3	0.0905	0.0950	0.0944	0.1393	0.0974
				0.3	0.0797	0.0809	0.0798	0.1238	0.0816
				0.7	0.0786	0.0794	0.0757	0.1110	0.0783
			0.3	-0.7	0.2611	0.2618	0.2703	0.3351	0.2815
				-0.3	0.2491	0.2500	0.2487	0.3236	0.2625
				0.3	0.2362	0.2369	0.2288	0.3020	0.2455
				0.7	0.2309	0.2330	0.2268	0.3037	0.2458
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.1015	0.1000	0.1001	0.1700	0.1066
				-0.3	0.1010	0.1001	0.1003	0.1657	0.1040
				0.3	0.0996	0.0996	0.1001	0.1613	0.1048
				0.7	0.0994	0.0998	0.1008	0.1643	0.1054
			0.3	-0.7	0.2994	0.2998	0.3026	0.3970	0.3366
				-0.3	0.2942	0.2937	0.2955	0.3974	0.3270

Table 3 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
				0.3	0.2916	0.2907	0.2927	0.3873	0.3226
				0.7	0.2878	0.2893	0.2963	0.3863	0.3224
		0.7*0.7	0.1	-0.7	0.0884	0.0968	0.0950	0.1307	0.0960
				-0.3	0.0802	0.0846	0.0838	0.1172	0.0828
				0.3	0.0779	0.0804	0.0798	0.1195	0.0829
				0.7	0.0738	0.0762	0.0749	0.1072	0.0772
			0.3	-0.7	0.2379	0.2391	0.2464	0.3153	0.2616
				-0.3	0.2263	0.2303	0.2341	0.3134	0.2455
				0.3	0.2246	0.2275	0.2295	0.3039	0.2440
				0.7	0.2228	0.2230	0.2282	0.3081	0.2396
Y5	0.0 – 0.3	0.3*0.3	0.1	-0.7	0.0631	0.0925	0.0745	0.0125	0.0111
				-0.3	0.0664	0.0742	0.0643	0.0135	0.0101
				0.3	0.0659	0.0664	0.0571	0.0120	0.0101
				0.7	0.0733	0.0740	0.0572	0.0124	0.0119
			0.3	-0.7	0.1620	0.1689	0.1289	0.0070	0.0069
				-0.3	0.1534	0.1578	0.1094	0.0149	0.0102
				0.3	0.1338	0.1348	0.1053	0.0145	0.0108
				0.7	0.1203	0.1215	0.1003	0.0118	0.0110
		0.7*0.7	0.1	-0.7	0.0631	0.0925	0.0745	0.0125	0.0111
				-0.3	0.0749	0.0807	0.0733	0.0130	0.0113
				0.3	0.0681	0.0698	0.0589	0.0155	0.0110
				0.7	0.0708	0.0725	0.0569	0.0154	0.0114
			0.3	-0.7	0.1591	0.1627	0.1330	0.0127	0.0079
				-0.3	0.1553	0.1592	0.1142	0.0137	0.0085
				0.3	0.1338	0.1348	0.1053	0.0145	0.0108
				0.7	0.1203	0.1215	0.1003	0.0118	0.0110
	0.3 – 0.7	0.3*0.3	0.1	-0.7	0.0520	0.0973	0.0629	0.0106	0.0126
				-0.3	0.0496	0.0666	0.0496	0.0137	0.0110
				0.3	0.0551	0.0583	0.0535	0.0133	0.0129
				0.7	0.0540	0.0532	0.0452	0.0159	0.0121

Table 3 (cont.)

Model	Multicollinearity	Variance	Base Rate	Skewness	Method				
					LR	LDA	QDA	CART	RF
			0.3	-0.7	0.0749	0.0799	0.0633	0.0110	0.0098
				-0.3	0.1135	0.1170	0.0868	0.0142	0.0106
				0.3	0.1067	0.1089	0.0865	0.0124	0.0108
				0.7	0.1295	0.1315	0.1109	0.0176	0.0121
		0.7*0.7	0.1	-0.7	0.0321	0.0953	0.0296	0.0044	0.0043
				-0.3	0.0386	0.0612	0.0373	0.0076	0.0043
				0.3	0.0429	0.0464	0.0407	0.0094	0.0065
				0.7	0.0433	0.0449	0.0393	0.0115	0.0111
			0.3	-0.7	0.1138	0.1220	0.1013	0.0106	0.0092
				-0.3	0.1135	0.1170	0.0868	0.0142	0.0106
				0.3	0.1354	0.1348	0.1045	0.0166	0.0113
				0.7	0.1272	0.1303	0.1036	0.0201	0.0152

Table 4

Mean and Standard Deviation of the Two Classification Accuracy Metrics by Model and Method (Study 1)

	Method				
	LR	LDA	QDA	CART	RF
Resubstitution					
Y1	0.170 (0.084)	0.174 (0.085)	0.175 (0.085)	0.066 (0.027)	0.000 (0.000)
Y2	0.169 (0.086)	0.173 (0.087)	0.170 (0.086)	0.064 (0.027)	0.000 (0.000)
Y3	0.164 (0.084)	0.169 (0.085)	0.168 (0.084)	0.063 (0.027)	0.000 (0.000)
Y4	0.173 (0.089)	0.178 (0.090)	0.176 (0.090)	0.066 (0.029)	0.000 (0.000)
Y5	0.088 (0.040)	0.100 (0.036)	0.074 (0.029)	0.002 (0.001)	0.000 (0.000)
Cross-Validation					
Y1	0.175 (0.086)	0.175 (0.086)	0.179 (0.087)	0.247 (0.104)	0.193 (0.097)
Y2	0.173 (0.087)	0.175 (0.088)	0.175 (0.088)	0.239 (0.107)	0.187 (0.097)
Y3	0.169 (0.086)	0.171 (0.086)	0.172 (0.086)	0.239 (0.106)	0.186 (0.097)
Y4	0.179 (0.091)	0.180 (0.091)	0.181 (0.092)	0.250 (0.112)	0.194 (0.103)
Y5	0.093 (0.040)	0.102 (0.036)	0.079 (0.029)	0.013 (0.003)	0.010 (0.002)

Note. SD is listed in parentheses within each cell.

Table 5

Mean and Standard Deviation of the Two Classification Accuracy Metrics for the Six Parameters (Study 1)

Parameter	Level	Resubstituition Error	Cross-Validation Error
Multicollinearity	0.0 – 0.3	0.106 (0.095)	0.168 (0.104)
	0.3 – 0.7	0.101 (0.093)	0.163 (0.104)
Proportion of Variance Explained	0.3*0.3	0.116 (0.106)	0.187 (0.118)
	0.7*0.7	0.091 (0.078)	0.144 (0.082)
Base Rate	0.1	0.058 (0.039)	0.093 (0.036)
	0.3	0.149 (0.109)	0.238 (0.098)
Skewness	-0.7	0.107 (0.096)	0.170 (0.105)
	-0.3	0.104 (0.094)	0.166 (0.103)
	0.3	0.102 (0.093)	0.163 (0.103)
	0.7	0.100 (0.094)	0.161 (0.104)
Data Complexity	Y1	0.117 (0.097)	0.194 (0.095)
	Y2	0.115 (0.098)	0.190 (0.096)
	Y3	0.113 (0.096)	0.187 (0.095)
	Y4	0.119 (0.101)	0.197 (0.101)
	Y5	0.053 (0.051)	0.059 (0.048)
Classification Method	LR	0.153 (0.084)	0.158 (0.086)
	LDA	0.159 (0.084)	0.160 (0.084)
	QDA	0.153 (0.087)	0.157 (0.088)
	CART	0.052 (0.035)	0.198 (0.133)
	RF	0.000 (0.000)	0.154 (0.113)

Note. SD is listed in parentheses within each cell.

Table 6

One-way and Two-way ANOVA Results Summary for Predicting Binary Outcomes (Study 1)

Parameter	ANOVA	Resubstituition Error	Cross-Validation Error
Multicollinearity	Main	F (1,798) = 0.48, p = .487	F (1,798) = 0.57, p = .450
	Interaction	F (4,790) = 0.15, p = .965	F (4,790) = 0.03, p = .998
Proportion of Variance Explained	Main	**F (1,798) = 14.43, p < .001**	**F (1,798) = 34.64, p < .001**
	Interaction	**F (4,790) = 2.57, p = .037**	F (4,790) = 0.41, p = .805
Base Rate	Main	**F (1,798) = 250.04, p < .001**	**F (1,798) = 770.06, p < .001**
	Interaction	**F (4,790) = 114.31, p < .001**	F (4,790) = 0.25, p = .908
Skewness	Main	F (3,796) = 0.21, p = .887	F (3,796) = 0.29, p = .831
	Interaction	F (12,780) = 0.06, p = 1.000	F (12,780) = 0.01, p = 1.000
Data Complexity	Main	**F (4,795) = 15.59, p < .001**	**F (4,795) = 71.17, p < .001**
	Interaction	**F (16,775) = 2.37, p = .002**	**F (16,775) = 4.25, p < .001**
Classification Method	Main	**F (4,795) = 185.87, p < .001**	**F (4,795) = 5.07, p < .001**

Note. Bold cells represent significant main or interaction effects.

Table 7

Resubstitution Mean Squared Error for Predicting Continuous Outcomes (Study 2)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
Y1	0.0 – 0.3	0.30*0.30	-0.7	0.9028	0.2786	0.4093
			-0.3	0.8958	0.2817	0.4098
			0.3	0.9116	0.2897	0.4140
			0.7	0.8939	0.2773	0.4070
		0.70*0.70	-0.7	0.5314	0.1726	0.2508
			-0.3	0.5055	0.1612	0.2391
			0.3	0.4960	0.1663	0.2361
			0.7	0.4695	0.1547	0.2237
		0.95*0.95	-0.7	0.1026	0.0539	0.0628
			-0.3	0.1192	0.0602	0.0726
			0.3	0.1052	0.0568	0.0662
			0.7	0.1061	0.0522	0.0650
	0.3 – 0.7	0.30*0.30	-0.7	0.8900	0.2790	0.4075
			-0.3	0.9274	0.2868	0.4199
			0.3	0.9280	0.2786	0.4194
			0.7	0.9086	0.2760	0.4130
		0.70*0.70	-0.7	0.5355	0.1688	0.2491
			-0.3	0.4896	0.1528	0.2275
			0.3	0.4952	0.1579	0.2313
			0.7	0.5369	0.1707	0.2478
		0.95*0.95	-0.7	0.0900	0.0394	0.0478
			-0.3	0.0967	0.0400	0.0510
			0.3	0.1091	0.0432	0.0568
			0.7	0.0949	0.0387	0.0500
Y2	0.0 – 0.3	0.30*0.30	-0.7	0.9248	0.2758	0.4125
			-0.3	0.9373	0.2806	0.4161
			0.3	0.9236	0.2858	0.4159
			0.7	0.9225	0.2820	0.4147

Table 7 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
		0.70*0.70	-0.7	0.6501	0.1693	0.2463
			-0.3	0.6324	0.1674	0.2382
			0.3	0.6154	0.1633	0.2424
			0.7	0.5568	0.1455	0.2218
		0.95*0.95	-0.7	0.3497	0.0571	0.0714
			-0.3	0.3764	0.0595	0.0812
			0.3	0.3068	0.0499	0.0681
			0.7	0.2489	0.0472	0.0636
	0.3 – 0.7	0.30*0.30	-0.7	0.9240	0.2857	0.4127
			-0.3	0.9332	0.2748	0.4205
			0.3	0.9245	0.2918	0.4149
			0.7	0.9137	0.2856	0.4131
		0.70*0.70	-0.7	0.6004	0.1622	0.2387
			-0.3	0.5907	0.1590	0.2373
			0.3	0.5805	0.1629	0.2406
			0.7	0.5839	0.1732	0.2487
		0.95*0.95	-0.7	0.2606	0.0435	0.0561
			-0.3	0.2688	0.0460	0.0606
			0.3	0.2426	0.0457	0.0634
			0.7	0.1919	0.0378	0.0517
Y3	0.0 – 0.3	0.30*0.30	-0.7	0.9179	0.2811	0.4110
			-0.3	0.9168	0.2810	0.4063
			0.3	0.9129	0.2815	0.4074
			0.7	0.9266	0.2781	0.4133
		0.70*0.70	-0.7	0.5829	0.1719	0.2405
			-0.3	0.5850	0.1612	0.2452
			0.3	0.5757	0.1642	0.2410
			0.7	0.5480	0.1508	0.2263
		0.95*0.95	-0.7	0.2534	0.0530	0.0666
			-0.3	0.2514	0.0573	0.0744

Table 7 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
			0.3	0.2477	0.0581	0.0731
			0.7	0.2409	0.0569	0.0717
	0.3 – 0.7	0.30*0.30	-0.7	0.9197	0.2797	0.4157
			-0.3	0.9230	0.2908	0.4119
			0.3	0.9278	0.2890	0.4185
			0.7	0.9054	0.2862	0.4136
		0.70*0.70	-0.7	0.5848	0.1645	0.2449
			-0.3	0.5661	0.1606	0.2378
			0.3	0.5841	0.1690	0.2525
			0.7	0.5929	0.1743	0.2577
		0.95*0.95	-0.7	0.2017	0.0411	0.0539
			-0.3	0.2136	0.0444	0.0590
			0.3	0.2018	0.0486	0.0640
			0.7	0.1685	0.0392	0.0491
Y4	0.0 – 0.3	0.30*0.30	-0.7	0.9256	0.2795	0.4170
			-0.3	0.9237	0.2725	0.4081
			0.3	0.9211	0.2688	0.4113
			0.7	0.9235	0.2841	0.4147
		0.70*0.70	-0.7	0.6522	0.1692	0.2475
			-0.3	0.6524	0.1635	0.2598
			0.3	0.6251	0.1644	0.2529
			0.7	0.6228	0.1690	0.2442
		0.95*0.95	-0.7	0.3984	0.0561	0.0842
			-0.3	0.3796	0.0522	0.1006
			0.3	0.3714	0.0497	0.1016
			0.7	0.3515	0.0562	0.0831
	0.3 – 0.7	0.30*0.30	-0.7	0.9321	0.2867	0.4129
			-0.3	0.9296	0.2875	0.4140
			0.3	0.9332	0.2785	0.4181
			0.7	0.9301	0.2730	0.4173

Table 7 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
		0.70*0.70	-0.7	0.6239	0.1677	0.2475
			-0.3	0.6326	0.1635	0.2507
			0.3	0.6273	0.1663	0.2519
			0.7	0.6623	0.1884	0.2716
		0.95*0.95	-0.7	0.3291	0.0478	0.0712
			-0.3	0.3115	0.0459	0.0844
			0.3	0.3188	0.0514	0.0800
			0.7	0.2953	0.0488	0.0766
Y6	0.0 – 0.3	0.30*0.30	-0.7	0.9224	0.2734	0.4031
			-0.3	0.9104	0.2635	0.3911
			0.3	0.9113	0.2543	0.3917
			0.7	0.8705	0.2581	0.3813
		0.70*0.70	-0.7	0.6419	0.1208	0.1838
			-0.3	0.6654	0.1300	0.1951
			0.3	0.6094	0.1128	0.1710
			0.7	0.5456	0.0930	0.1434
		0.95*0.95	-0.7	0.4653	0.0194	0.0328
			-0.3	0.4554	0.0212	0.0342
			0.3	0.4345	0.0177	0.0306
			0.7	0.3838	0.0145	0.0253
	0.3 – 0.7	0.30*0.30	-0.7	0.8917	0.2601	0.3854
			-0.3	0.8702	0.2534	0.3809
			0.3	0.8535	0.2543	0.3656
			0.7	0.8233	0.2419	0.3534
		0.70*0.70	-0.7	0.5219	0.0945	0.1470
			-0.3	0.5190	0.0933	0.1419
			0.3	0.4866	0.0851	0.1319
			0.7	0.4402	0.0718	0.1134
		0.95*0.95	-0.7	0.3335	0.0152	0.0245
			-0.3	0.3589	0.0162	0.0266

Table 7 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
			0.3	0.3242	0.0131	0.0226
			0.7	0.3073	0.0120	0.0201

Table 8

10-fold Cross-Validation Mean Squared Error for Predicting Continuous Outcomes (Study 2)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
Y1	0.0 – 0.3	0.30*0.30	-0.7	0.9120	1.5701	0.9747
			-0.3	0.9046	1.5780	0.9730
			0.3	0.9213	1.5779	0.9831
			0.7	0.9038	1.5654	0.9693
		0.70*0.70	-0.7	0.5364	0.9734	0.5921
			-0.3	0.5104	0.9037	0.5619
			0.3	0.5010	0.8924	0.5551
			0.7	0.4746	0.8543	0.5252
		0.95*0.95	-0.7	0.1035	0.2728	0.1511
			-0.3	0.1206	0.3066	0.1735
			0.3	0.1063	0.2837	0.1598
			0.7	0.1073	0.2779	0.1555
	0.3 – 0.7	0.30*0.30	-0.7	0.8989	1.5802	0.9644
			-0.3	0.9365	1.6153	0.9987
			0.3	0.9369	1.5730	0.9943
			0.7	0.9191	1.5698	0.9824
		0.70*0.70	-0.7	0.5410	0.9346	0.5853
			-0.3	0.4953	0.8678	0.5349
			0.3	0.5003	0.8805	0.5422
			0.7	0.5425	0.9382	0.5866
		0.95*0.95	-0.7	0.0910	0.2021	0.1131
			-0.3	0.0976	0.2043	0.1203
			0.3	0.1103	0.2378	0.1336
			0.7	0.0959	0.2046	0.1185
Y2	0.0 – 0.3	0.30*0.30	-0.7	0.9354	1.5641	0.9772
			-0.3	0.9485	1.5829	0.9883
			0.3	0.9335	1.6103	0.9929
			0.7	0.9316	1.5728	0.9876

Table 8 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
		0.70*0.70	-0.7	0.6577	0.9413	0.5791
			-0.3	0.6402	0.9207	0.5629
			0.3	0.6224	0.9284	0.5717
			0.7	0.5634	0.8434	0.5202
		0.95*0.95	-0.7	0.3546	0.3067	0.1718
			-0.3	0.3820	0.3257	0.1936
			0.3	0.3124	0.2600	0.1613
			0.7	0.2525	0.2473	0.1496
	0.3 – 0.7	0.30*0.30	-0.7	0.9336	1.5368	0.9788
			-0.3	0.9437	1.6153	0.9986
			0.3	0.9346	1.6066	0.9843
			0.7	0.9234	1.5548	0.9749
		0.70*0.70	-0.7	0.6077	0.9123	0.5629
			-0.3	0.5980	0.8862	0.5579
			0.3	0.5873	0.9265	0.5652
			0.7	0.5906	0.9344	0.5831
		0.95*0.95	-0.7	0.2648	0.2227	0.1329
			-0.3	0.2734	0.2348	0.1430
			0.3	0.2468	0.2386	0.1474
			0.7	0.1944	0.2019	0.1205
Y3	0.0 – 0.3	0.30*0.30	-0.7	0.9279	1.5775	0.9797
			-0.3	0.9268	1.5898	0.9675
			0.3	0.9237	1.6041	0.9659
			0.7	0.9376	1.5634	0.9756
		0.70*0.70	-0.7	0.5901	0.9261	0.5656
			-0.3	0.5928	0.9256	0.5748
			0.3	0.5818	0.8976	0.5650
			0.7	0.5547	0.8656	0.5308
		0.95*0.95	-0.7	0.2577	0.2766	0.1578
			-0.3	0.2548	0.3078	0.1766

Table 8 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
			0.3	0.2519	0.2998	0.1732
			0.7	0.2448	0.2885	0.1690
	0.3 – 0.7	0.30*0.30	-0.7	0.9313	1.6062	0.9875
			-0.3	0.9330	1.5686	0.9790
			0.3	0.9374	1.5986	0.9940
			0.7	0.9151	1.5747	0.9773
		0.70*0.70	-0.7	0.5911	0.9300	0.5767
			-0.3	0.5727	0.9190	0.5590
			0.3	0.5906	0.9591	0.5912
			0.7	0.5990	0.9759	0.6053
		0.95*0.95	-0.7	0.2044	0.2187	0.1261
			-0.3	0.2168	0.2396	0.1375
			0.3	0.2049	0.2527	0.1499
			0.7	0.1709	0.1978	0.1153
Y4	0.0 – 0.3	0.30*0.30	-0.7	0.9354	1.6128	0.9818
			-0.3	0.9347	1.5671	0.9640
			0.3	0.9304	1.5907	0.9824
			0.7	0.9333	1.5837	0.9812
		0.70*0.70	-0.7	0.6608	0.9082	0.5827
			-0.3	0.6611	0.9272	0.6027
			0.3	0.6334	0.8962	0.5854
			0.7	0.6304	0.9012	0.5702
		0.95*0.95	-0.7	0.4064	0.2540	0.1865
			-0.3	0.3866	0.2896	0.2199
			0.3	0.3788	0.2959	0.2137
			0.7	0.3577	0.2613	0.1850
	0.3 – 0.7	0.30*0.30	-0.7	0.9420	1.5850	0.9812
			-0.3	0.9405	1.6042	0.9813
			0.3	0.9440	1.6310	0.9926
			0.7	0.9411	1.5712	0.9821

Table 8 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
		0.70*0.70	-0.7	0.6310	0.9188	0.5793
			-0.3	0.6409	0.9251	0.5880
			0.3	0.6355	0.9143	0.5898
			0.7	0.6711	1.0388	0.6378
		0.95*0.95	-0.7	0.3347	0.2363	0.1609
			-0.3	0.3176	0.2542	0.1783
			0.3	0.3245	0.2591	0.1800
			0.7	0.3006	0.2454	0.1676
Y6	0.0 – 0.3	0.30*0.30	-0.7	0.9322	1.5516	0.9557
			-0.3	0.9208	1.5060	0.9233
			0.3	0.9214	1.4640	0.9217
			0.7	0.8798	1.4109	0.9027
		0.70*0.70	-0.7	0.6488	0.6729	0.4299
			-0.3	0.6720	0.7106	0.4578
			0.3	0.6152	0.6311	0.4015
			0.7	0.5519	0.5265	0.3353
		0.95*0.95	-0.7	0.4706	0.1146	0.0796
			-0.3	0.4605	0.1207	0.0827
			0.3	0.4389	0.1054	0.0737
			0.7	0.3877	0.0854	0.0611
	0.3 – 0.7	0.30*0.30	-0.7	0.9010	1.4624	0.9107
			-0.3	0.8788	1.4124	0.9031
			0.3	0.8618	1.3723	0.8609
			0.7	0.8315	1.3376	0.8333
		0.70*0.70	-0.7	0.5278	0.5403	0.3456
			-0.3	0.5248	0.5225	0.3332
			0.3	0.4915	0.4896	0.3096
			0.7	0.4454	0.4221	0.2652
		0.95*0.95	-0.7	0.3372	0.0881	0.0597
			-0.3	0.3628	0.0938	0.0638

Table 8 (cont.)

Model	Multicollinearity	Variance	Skewness	Method		
				MLR	CART	RF
			0.3	0.3273	0.0793	0.0550
			0.7	0.3103	0.0683	0.0491

Table 9

Mean and Standard Deviation of the Two Prediction Accuracy Metrics by Model and Method (Study 2)

	Method		
	MLR	CART	RF
Resubstitution			
Y1	0.506 (0.336)	0.164 (0.097)	0.237 (0.148)
Y2	0.602 (0.271)	0.165 (0.098)	0.240 (0.146)
Y3	0.573 (0.291)	0.166 (0.098)	0.240 (0.145)
Y4	0.636 (0.244)	0.166 (0.095)	0.251 (0.137)
Y6	0.606 (0.219)	0.125 (0.103)	0.187 (0.151)
Cross-Validation			
Y1	0.511 (0.339)	0.911 (0.556)	0.560 (0.351)
Y2	0.610 (0.274)	0.916 (0.554)	0.567 (0.348)
Y3	0.580 (0.294)	0.923 (0.553)	0.567 (0.346)
Y4	0.645 (0.246)	0.928 (0.556)	0.586 (0.332)
Y6	0.613 (0.222)	0.700 (0.573)	0.442 (0.356)

Note. SD is listed in parentheses within each cell.

Table 10

Mean and Standard Deviation of the Two Prediction Accuracy Metrics for the Five Parameters (Study 2)

Parameter	Level	Resubstitution MSE	Cross-Validation MSE
Multicollinearity	0.0 – 0.3	0.331 (0.265)	0.682 (0.427)
	0.3 – 0.7	0.317 (0.265)	0.659 (0.443)
Proportion of Variance Explained	0.30*0.30	0.532 (0.276)	1.148 (0.294)
	0.70*0.70	0.318 (0.191)	0.653 (0.176)
	0.95*0.95	0.123 (0.120)	0.211 (0.097)
Skewness	-0.7	0.329 (0.268)	0.678 (0.438)
	-0.3	0.328 (0.267)	0.679 (0.433)
	0.3	0.324 (0.265)	0.672 (0.436)
	0.7	0.316 (0.263)	0.654 (0.438)
Data Complexity	Y1	0.302 (0.262)	0.661 (0.457)
	Y2	0.336 (0.266)	0.697 (0.433)
	Y3	0.326 (0.263)	0.690 (0.440)
	Y4	0.351 (0.266)	0.720 (0.422)
	Y6	0.306 (0.270)	0.585 (0.418)
Regression Method	GLM	0.585 (0.274)	0.592 (0.277)
	CART	0.157 (0.098)	0.876 (0.556)
	RF	0.231 (0.145)	0.545 (0.345)

Note. SD is listed in parentheses within each cell.

Table 11

One-way and Two-way ANOVA Results Summary for Predicting Continuous Outcome (Study 2)

Parameter	ANOVA	Overall Misclassification Error	Cross-Validation Error
Multicollinearity	Main	F (1,358) = 0.24, p = .626	F (1,358) = 0.26, p = .608
	Interaction	F (2,354) = 0.19, p = .830	F (2,354) = 0.01, p = .988
Proportion of Variance Explained	Main	**F (2,357) = 118.70, p < .001**	**F (2,357) = 622.56, p < .001**
	Interaction	**F (4,351) = 213.36, p < .001**	**F (4,351) = 182.84, p < .001**
Skewness	Main	F (3,356) = 0.05, p = 0.986	F (3,356) = 0.06, p = .980
	Interaction	F (6,348) = 0.04, p = 1.000	F (6,348) = 0.00, p = 1.000
Data Complexity	Main	F (4,355) = 0.43, p = .788	F (4,355) = 1.05, p = .383
	Interaction	F (8,345) = 0.72, p = .671	F (8,345) = 0.52, p = .838
Regression Method	Main	**F (2,357) = 177.85, p < .001**	**F (2,357) = 22.88, p < .001**

Note. Bold cells represent significant main or interaction effects.

Table 12

Mean and Standard Deviation for Stayed, Voluntary, and Involuntary Turnover Employees (Study 3)

	Stayed (N = 9105)	Voluntary Turnover (N = 1997)	Involuntary Turnover (N = 416)
DE	4.48 (0.52)	4.43 (0.57)	4.49 (0.55)
ITS	4.40 (0.62)	4.09 (0.82)	4.26 (0.72)
JF	4.10 (0.63)	3.91 (0.77)	3.95 (0.76)
SL	4.17 (0.62)	4.12 (0.67)	4.15 (0.66)
OC	4.37 (0.55)	4.26 (0.67)	4.32 (0.65)
ONB	4.29 (0.53)	4.23 (0.57)	4.25 (0.61)
FR	3.97 (0.64)	3.84 (0.73)	3.88 (0.67)
SUP	4.39 (0.58)	4.31 (0.66)	4.32 (0.70)
COW	4.44 (0.53)	4.38 (0.58)	4.36 (0.63)

Note. DE = Discretionary Effort, ITS = Intent to Stay, JF = Job Fit, SL = Senior Leadership, OC = Organizational Commitment, ONB = Onboarding, FR = Financial Rewards, SUP = Supervisor, COW = Coworker.

Table 13

Mean, Standard Deviation, and Intercorrelations among Predictors (All Employees Included) (Study 3)

	M (SD)	1	2	3	4	5	6	7	8	9
1. DE	4.47 (0.53)	0.64	0.46**	0.52**	0.47**	0.56**	0.45**	0.34**	0.43**	0.38**
2. ITS	4.34 (0.68)	0.46**	0.83	0.54**	0.39**	0.59**	0.38**	0.38**	0.39**	0.34**
3. JF	4.06 (0.67)	0.52**	0.54**	0.84	0.58**	0.65**	0.53**	0.51**	0.60**	0.53**
4. SL	4.16 (0.63)	0.47**	0.39**	0.58**	0.91	0.63**	0.54**	0.52**	0.55**	0.47**
5. OC	4.35 (0.57)	0.56**	0.59**	0.65**	0.63**	0.81	0.57**	0.57**	0.56**	0.48**
6. ONB	4.28 (0.54)	0.45**	0.38**	0.53**	0.54**	0.57**	0.84	0.46**	0.63**	0.47**
7. FR	3.94 (0.66)	0.34**	0.38**	0.51**	0.52**	0.57**	0.46**	0.86	0.44**	0.38**
8. SUP	4.37 (0.60)	0.43**	0.39**	0.60**	0.55**	0.56**	0.63**	0.44**	0.93	0.51**
9. COW	4.42 (0.55)	0.38**	0.34**	0.53**	0.47**	0.48**	0.47**	0.38**	0.51**	0.90

Note. $N = 11,518$. DE = Discretionary Effort, ITS = Intent to Stay, JF = Job Fit, SL = Senior Leadership, OC = Organizational Commitment, ONB = Onboarding, FR = Financial Rewards, SUP = Supervisor, COW = Coworker. Cronbach's alpha reliabilities of the predictors are listed in the diagonal. ** $p < .01$.

Table 14

Relative Importance of Predictors for Predicting Voluntary Turnover (Study 3)

	RF	RW	β	β^2	p
ONB	10.7 E-06 (1)	6.83 E-04 (8)	0.0049 (6)	0.0000 (9)	0.9384
FR	10.3 E-06 (2)	25.35 E-04 (3)	-0.1450 (7)**	0.0210 (6)	0.0022
SUP	9.30 E-06 (3)	8.38 E-04 (7)	0.0379 (5)	0.0014 (8)	0.5221
ITS	8.49 E-06 (4)	253.94 E-04 (1)	-0.6859 (9)**	0.4705 (1)	0.0000
JF	8.02 E-06 (5)	58.91 E-04 (2)	-0.2909 (8)**	0.0846 (2)	0.0000
COW	7.90 E-06 (6)	6.22 E-04 (9)	0.0422 (4)	0.0018 (7)	0.4335
SL	7.51 E-06 (7)	13.51 E-04 (5)	0.2276 (1)**	0.0518 (3)	0.0001
OC	6.77 E-06 (8)	22.28 E-04 (4)	0.2039 (3)**	0.0416 (5)	0.0042
DE	3.45 E-06 (9)	12.60 E-04 (6)	0.2107 (2)**	0.0444 (4)	0.0001

Note. RF = Random Forests. RW = Relative Weights. β is logistic regression coefficient. p is significant value for linear regression coefficient. ** $p < .01$. The number in parentheses represents its relative ranking. ONB = Onboarding, FR = Financial Rewards, SUP = Supervisor, ITS = Intent to Stay, JF = Job Fit, COW = Coworker, SL = Senior Leadership, OC = Organizational Commitment, DE = Discretionary Effort.

Table 15

Relative Importance of Predictors for Predicting Involuntary Turnover (Study 3)

	RF	RW	β	β^2	p
FR	8.92 E-06 (1)	5.06 E-04 (4)	-0.1885 (6) *	0.0355 (4)	0.0458
ONB	8.29 E-06 (2)	1.20 E-04 (9)	-0.0421 (4)	0.0018 (9)	0.7347
SUP	7.27 E-06 (3)	2.17 E-04 (8)	-0.0646 (5)	0.0042 (8)	0.5622
JF	6.45 E-06 (4)	14.87 E-04 (1)	-0.2806 (8) **	0.0787 (2)	0.0080
ITS	6.40 E-06 (5)	13.69 E-04 (2)	-0.3265 (9) **	0.1066 (1)	0.0003
COW	6.33 E-06 (6)	4.47 E-04 (5)	-0.2723 (7) **	0.0741 (3)	0.0065
SL	5.73 E-06 (7)	2.69 E-04 (6)	0.1764 (1)	0.0311 (5)	0.1118
OC	5.29 E-06 (8)	2.50 E-04 (7)	0.1577 (2)	0.0249 (6)	0.2537
DE	3.16 E-06 (9)	5.21 E-04 (3)	0.0972 (3)	0.0094 (7)	0.3551

Note. RF = Random Forests. RW = Relative Weights. β is logistic regression coefficient. p is significant value for linear regression coefficient. ** $p < .01$. The number in parentheses represents its relative ranking. FR = Financial Rewards, ONB = Onboarding, SUP = Supervisor, JF = Job Fit, ITS = Intent to Stay, COW = Coworker, SL = Senior Leadership, OC = Organizational Commitment, DE = Discretionary Effort.

Table 16

Classification Error Rate for Predicting Voluntary Turnover (Study 3)

		TREE	RF	LR	LDA	QDA
Overall Misclassification Error	2 Predictor	0.1709	0.1656	0.1789	0.1799	0.1811
	3 Predictors	0.1468	0.1087	0.1789	0.1799	0.1854
	4 Predictors	0.1274	0.0640	0.1789	0.1794	0.1858
	5 Predictors	0.1181	0.0301	0.1786	0.1787	0.1894
	6 Predictors	0.1093	0.0203	0.1787	0.1789	0.1898
	7 Predictors	0.1058	0.0168	0.1783	0.1792	0.1904
	8 Predictors	0.1010	0.0163	0.1782	0.1783	0.1937
	9 Predictors	0.0972	0.0161	0.1782	0.1792	0.1962
Cross-Validation Error	2 Predictor	0.1937	0.1967	0.1792	0.1799	0.1819
	3 Predictors	0.2245	0.2123	0.1792	0.1799	0.1855
	4 Predictors	0.2340	0.2020	0.1794	0.1798	0.1861
	5 Predictors	0.2458	0.2005	0.1792	0.1790	0.1901
	6 Predictors	0.2519	0.1959	0.1791	0.1792	0.1909
	7 Predictors	0.2608	0.1900	0.1791	0.1792	0.1921
	8 Predictors	0.2649	0.1912	0.1788	0.1789	0.1946
	9 Predictors	0.2624	0.1869	0.1789	0.1792	0.1974
Sensitivity	2 Predictor	0.0270	0.0315	0.0072	0.0000	0.0135
	3 Predictors	0.0921	0.0771	0.0075	0.0000	0.0300
	4 Predictors	0.1652	0.0886	0.0321	0.0230	0.1062
	5 Predictors	0.1868	0.0906	0.0295	0.0285	0.1167
	6 Predictors	0.1813	0.0706	0.0263	0.0290	0.1247
	7 Predictors	0.1883	0.0621	0.0263	0.0305	0.1267
	8 Predictors	0.1938	0.0421	0.0342	0.0325	0.1357
	9 Predictors	0.2048	0.0481	0.0342	0.0371	0.1402
Specificity	2 Predictor	0.9773	0.9725	0.9993	1.0000	0.9945
	3 Predictors	0.9254	0.9435	0.9992	1.0000	0.9866
	4 Predictors	0.8977	0.9535	0.9936	0.9951	0.9691
	5 Predictors	0.8786	0.9550	0.9944	0.9948	0.9619
	6 Predictors	0.8724	0.9650	0.9952	0.9944	0.9593
	7 Predictors	0.8601	0.9741	0.9953	0.9942	0.9573
	8 Predictors	0.8538	0.9769	0.9939	0.9941	0.9522
	9 Predictors	0.8545	0.9809	0.9937	0.9926	0.9478

Table 17

Classification Error Rate for Predicting Involuntary Turnover (Study 3)

		TREE	RF	LR	LDA	QDA
Overall Misclassification Error	2 Predictor	0.0425	0.0402	0.0438	0.0437	0.0439
	3 Predictors	0.0413	0.0298	0.0435	0.0437	0.0450
	4 Predictors	0.0383	0.0213	0.0438	0.0437	0.0465
	5 Predictors	0.0374	0.0116	0.0436	0.0437	0.0474
	6 Predictors	0.0368	0.0068	0.0436	0.0437	0.0495
	7 Predictors	0.0352	0.0065	0.0435	0.0437	0.0493
	8 Predictors	0.0332	0.0062	0.0437	0.0437	0.0500
	9 Predictors	0.0323	0.0062	0.0438	0.0437	0.0522
Cross-Validation Error	2 Predictor	0.0456	0.0460	0.0439	0.0437	0.0440
	3 Predictors	0.0487	0.0486	0.0439	0.0437	0.0458
	4 Predictors	0.0546	0.0471	0.0439	0.0437	0.0473
	5 Predictors	0.0582	0.0469	0.0437	0.0437	0.0482
	6 Predictors	0.0618	0.0456	0.0437	0.0437	0.0502
	7 Predictors	0.0623	0.0448	0.0444	0.0437	0.0510
	8 Predictors	0.0610	0.0444	0.0441	0.0437	0.0508
	9 Predictors	0.0634	0.0438	0.0439	0.0437	0.0540
Sensitivity	2 Predictor	0.0120	0.0120	0.0000	0.0000	0.0000
	3 Predictors	0.0096	0.0192	0.0000	0.0000	0.0024
	4 Predictors	0.0168	0.0000	0.0000	0.0000	0.0144
	5 Predictors	0.0168	0.0000	0.0000	0.0000	0.0240
	6 Predictors	0.0216	0.0000	0.0000	0.0000	0.0288
	7 Predictors	0.0192	0.0000	0.0023	0.0000	0.0240
	8 Predictors	0.0288	0.0024	0.0000	0.0000	0.0337
	9 Predictors	0.0264	0.0000	0.0000	0.0000	0.0264
Specificity	2 Predictor	0.9975	0.9970	0.9998	1.0000	0.9997
	3 Predictors	0.9943	0.9940	0.9998	1.0000	0.9977
	4 Predictors	0.9878	0.9965	0.9998	1.0000	0.9956
	5 Predictors	0.9841	0.9966	1.0000	1.0000	0.9942
	6 Predictors	0.9801	0.9980	1.0000	1.0000	0.9919
	7 Predictors	0.9797	0.9988	0.9991	1.0000	0.9912
	8 Predictors	0.9806	0.9991	0.9996	1.0000	0.9910
	9 Predictors	0.9781	0.9999	0.9998	1.0000	0.9880

Table 18

Mean, Standard Deviation, and Intercorrelations between Predictors and Job Performance (Study4)

	M (SD)	1	2	3	4	5	6	7	8	9	10	11	12
1. CLO	4.14 (1.04)	1.00	0.48**	0.42**	0.38**	0.48**	0.45**	0.47**	0.45**	-0.14**	-0.09**	0.03	0.07**
2. LD	5.07 (0.98)	0.48**	1.00	0.51**	0.49**	0.61**	0.57**	0.61**	0.55**	-0.09**	-0.10**	0.11	0.12**
3. PM	4.60 (0.81)	0.42**	0.51**	1.00	0.50**	0.58**	0.56**	0.50**	0.53**	-0.10**	-0.09**	0.09**	0.13**
4. DP	5.01 (1.01)	0.38**	0.49**	0.50**	1.00	0.60**	0.53**	0.51**	0.58**	-0.12**	-0.08*	0.11**	0.17**
5. DMR	3.02 (0.58)	0.48**	0.61**	0.58**	0.60**	1.00	0.66**	0.65**	0.68**	-0.10**	-0.07	0.19**	0.13**
6. FL	2.83 (0.65)	0.45**	0.57**	0.56**	0.53**	0.66**	1.00	0.56**	0.66**	-0.07*	0.00	0.16**	0.11**
7. IC	3.16 (0.53)	0.47**	0.61**	0.50**	0.51**	0.65**	0.56**	1.00	0.55**	-0.08**	-0.05	0.12**	0.10**
8. SM	4.36 (0.90)	0.45**	0.55**	0.53**	0.58**	0.68**	0.66**	0.55**	1.00	-0.10**	-0.05	0.15**	0.16**
9. FPPS	26.62 (5.76)	-0.14**	-0.09**	-0.10**	-0.12**	-0.10**	-0.07*	-0.08**	-0.10**	1.00	---	---	0.18**
10. FPPP	30.14 (5.58)	-0.09**	-0.10**	-0.09**	-0.08*	-0.07	0.00	-0.05	-0.05	---	1.00	---	0.14**
11. FPP	28.15 (5.95)	0.03	0.11**	0.09**	0.11**	0.19**	0.16**	0.12**	0.15**	---	---	1.00	0.21**
12. PERF	3.53 (0.52)	0.07**	0.12**	0.13**	0.17**	0.13**	0.11**	0.10**	0.16**	0.18**	0.14**	0.14**	1.00

Note. N = 1,818 for most of the relationship among predictors, except for N = 1,027 between FPPS and all the predictors and N = 791 between FPPP and all the predictors; N = 1,815 for the relationship between predictors and job performance, except for N = 1,025 between FPPS and job performance and N = 790 between FPPP and job performance. ** p < .01. CLO = Continuous Learning Orientation, LD = Leadership Skill, PM = Personal Motivation, DP = Dependability, DMR = Development and Maintaining Relationship, FL = Flexibility, IC = Interpersonal Communication, SM = Self-Management Skills, FPPS = Following Policies and Procedures Support (Group), FPPP = Following Policies and Procedures Professional (Group), FPP = Following Policies and Procedures, PERF = Job Performance (composite). The number in parentheses represents its relative ranking.

Table 19

Relative Importance of Predictors for Predicting Job Performance (Study 4)

	RF	RW	β	β^2	p
FPP	5.93 E-05 (1)	0.0363 (1)	0.0436 (6) **	0.0019 (8)	0.0000
CLO	2.80 E-05 (2)	0.0008 (9)	0.0267 (7)	0.0007 (9)	0.1043
DP	2.58 E-05 (3)	0.0133 (2)	0.1160 (3) **	0.0135 (4)	0.0000
SM	2.53 E-05 (4)	0.0071 (3)	0.0619 (5) **	0.0038 (7)	0.0096
PM	2.51 E-05 (5)	0.0049 (4)	0.1840 (2) **	0.0339 (2)	0.0000
FL	2.22 E-05 (6)	0.0017 (7)	-0.1231 (9) **	0.0152 (3)	0.0002
IC	2.07 E-05 (7)	0.0014 (8)	0.2112 (1) **	0.0446 (1)	0.0000
LD	1.91 E-05 (8)	0.0038 (5)	0.0672 (4) **	0.0045 (6)	0.0009
DMR	1.64 E-05 (9)	0.0028 (6)	-0.0691 (8)	0.0048 (5)	0.0920

Note. RF = Random Forests. RW = Relative Weights. β is general linear regression coefficient. p is significant value for linear regression coefficient. ** $p < .01$. The number in parentheses represents its relative ranking. FPP = Following Policies and Procedures, CLO = Continuous Learning Orientation, DP = Dependability, SM = Self-Management Skills, PM = Personal Motivation, FL = Flexibility, IC = Interpersonal Communication, LD = Leadership Skill, DMR = Development and Maintaining Relationship.

Table 20

Mean Squared Error for Predicting Job Performance (Study 4)

	Method		
	CART	RF	MLR
Resubstitution			
2 Predictor	0.2404	0.2434	0.2578
3 Predictors	0.2170	0.2334	0.2526
4 Predictors	0.1839	0.2057	0.2520
5 Predictors	0.1689	0.1953	0.2518
6 Predictors	0.1533	0.1889	0.2514
7 Predictors	0.1403	0.1705	0.2512
8 Predictors	0.1303	0.1657	0.2510
9 Predictors	0.1301	0.1667	0.2509
Cross-Validation			
2 Predictor	0.2779	0.2629	0.2995
3 Predictors	0.3111	0.2613	0.3028
4 Predictors	0.3238	0.2703	0.3074
5 Predictors	0.3590	0.2683	0.3076
6 Predictors	0.3725	0.2668	0.3077
7 Predictors	0.3727	0.2650	0.3088
8 Predictors	0.3850	0.2651	0.3090
9 Predictors	0.4009	0.2647	0.3125

FIGURES

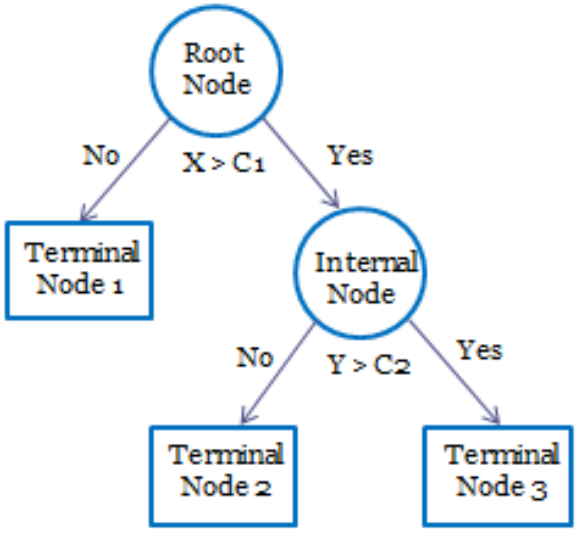

Figure 1. A Sample Tree Diagram

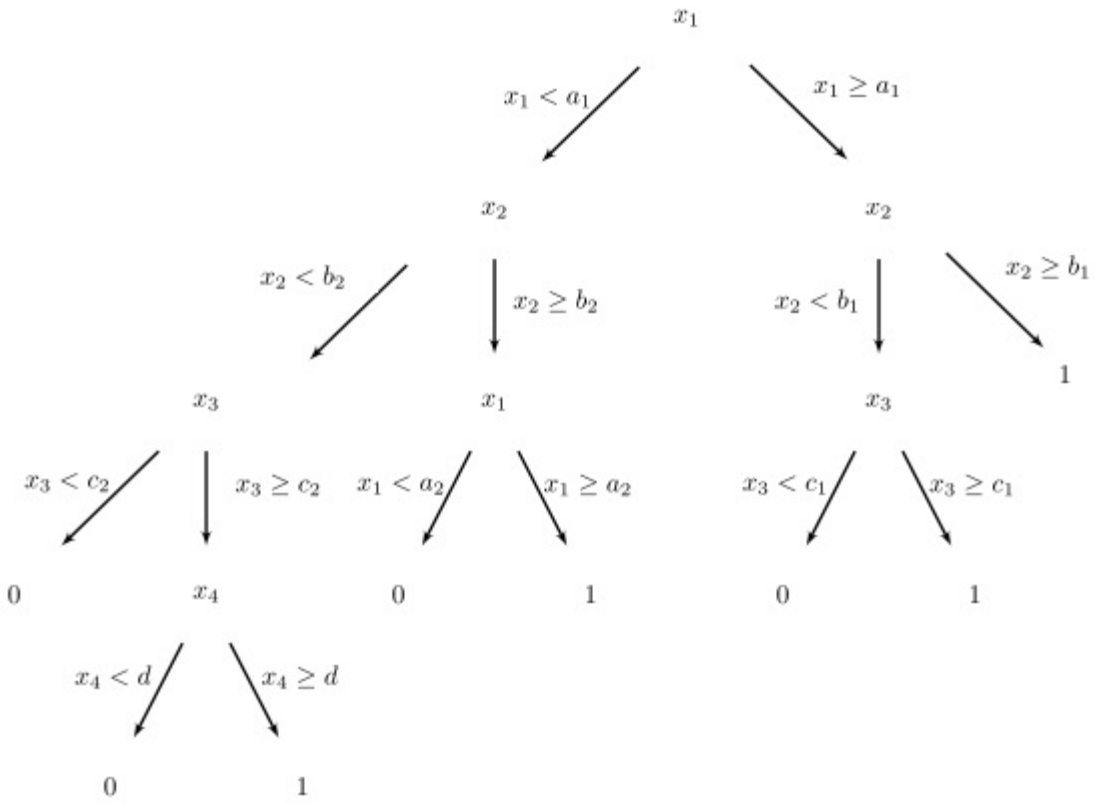

Figure 2. Prototype Tree Diagram for Simulation

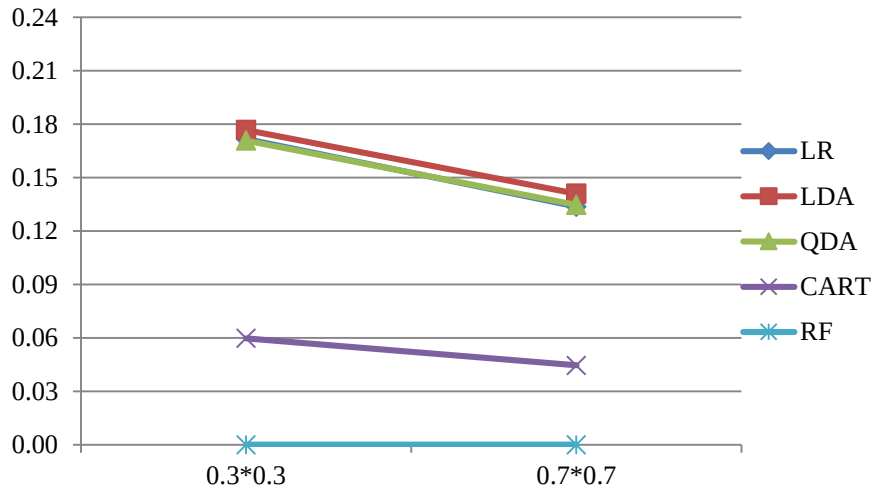

Figure 3. Interaction Effect between Proportion of Variance Explained and Classification Method on Resubstitution Error (Study 1)

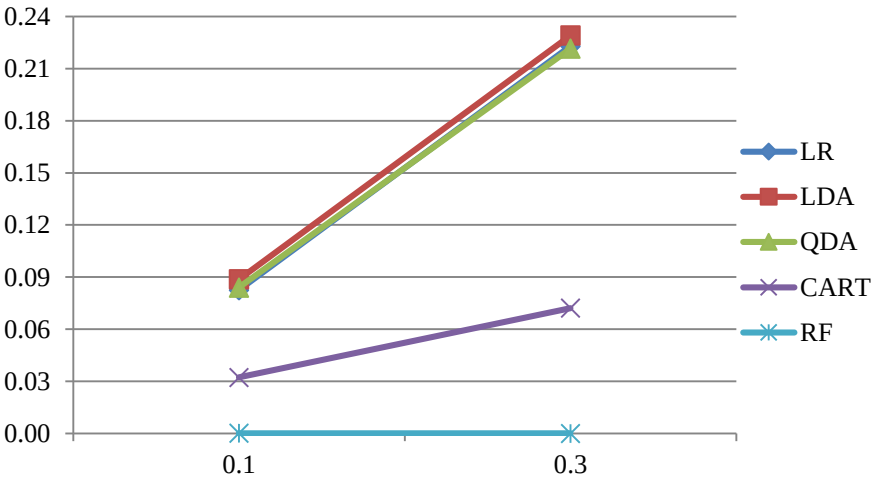

Figure 4. Interaction Effect between Base Rate and Classification Method on Resubstitution Error (Study 1)

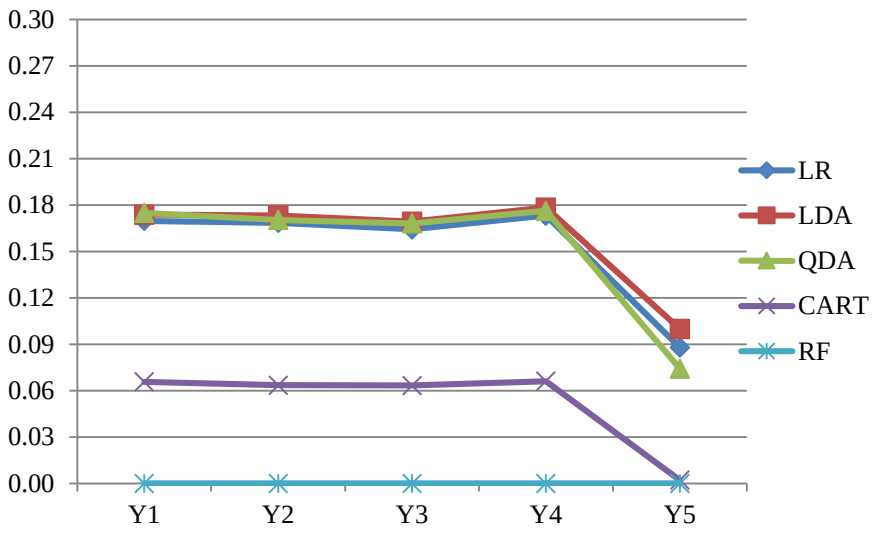

Figure 5. Interaction Effect between Data Complexity and Classification Method on Resubstitution Error (Study 1)

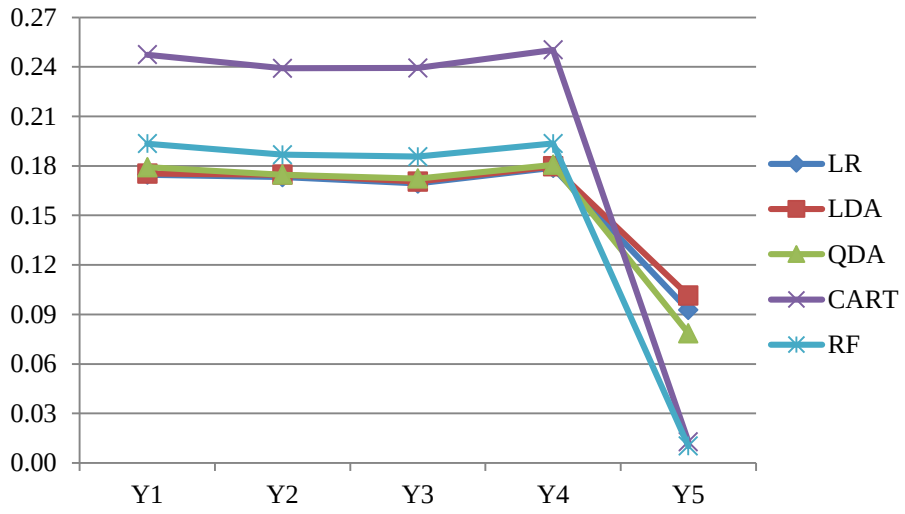

Figure 6. Interaction Effect between Data Complexity and Classification Method on Cross-Validation Error (Study 1)

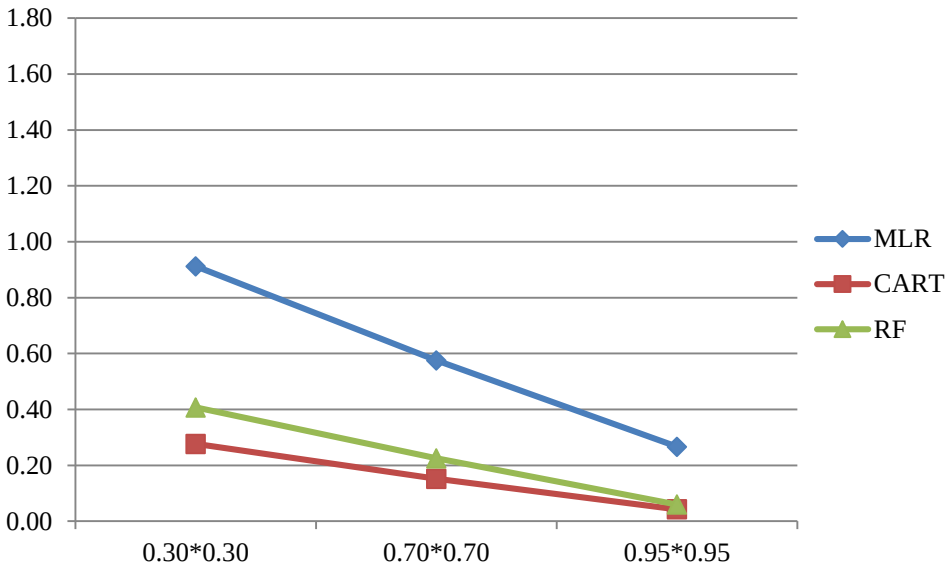

Figure 7. Interaction Effect between Proportion of Variance Explained and Prediction Method on Resubstitution MSE (Study 2)

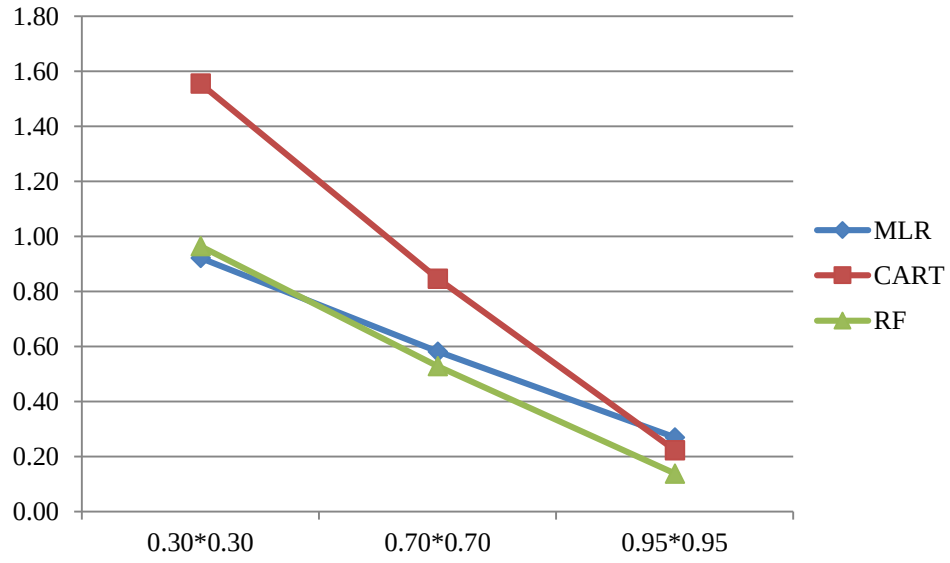

Figure 8. Interaction Effect between Proportion of Variance Explained and Prediction Method on Cross-Validation MSE (Study 2)

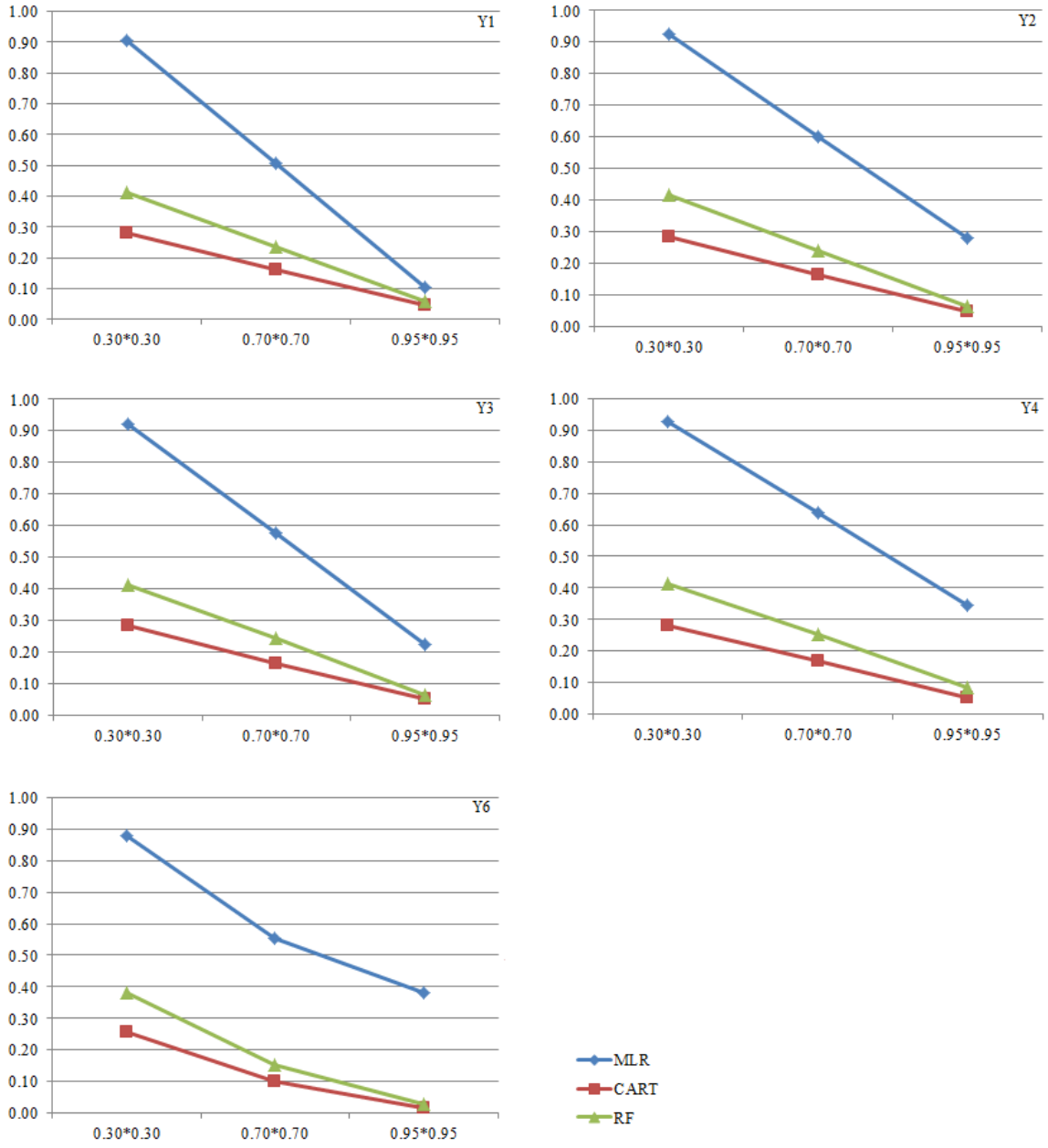

Figure 9. Three-Way Interaction Effect among Proportion of Variance Explained, Data Complexity and Prediction Method on Resubstitution MSE (Study 2)

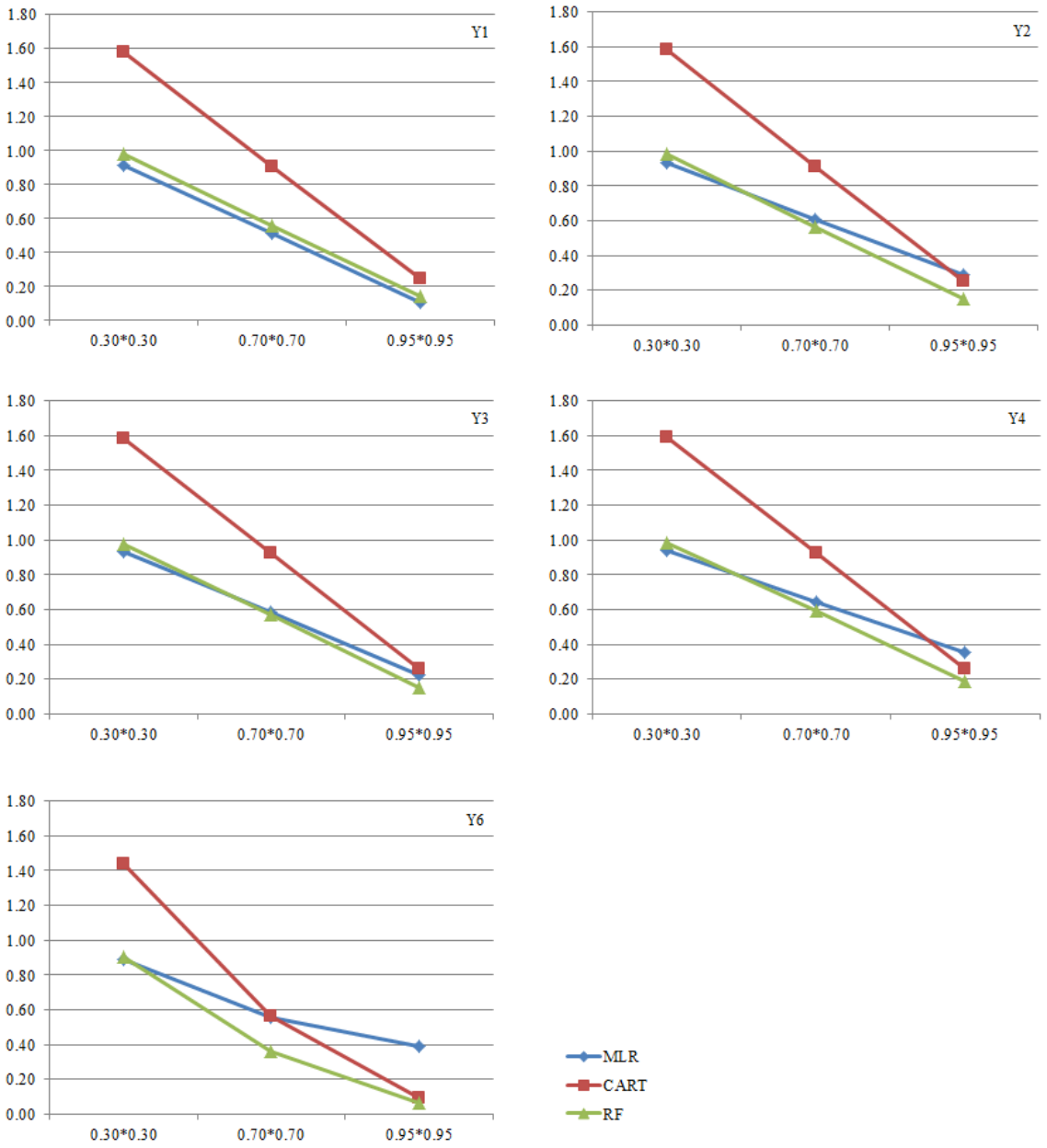

Figure 10. Three-Way Interaction Effect among Proportion of Variance Explained, Data Complexity and Prediction Method on Cross-Validation MSE (Study 2)

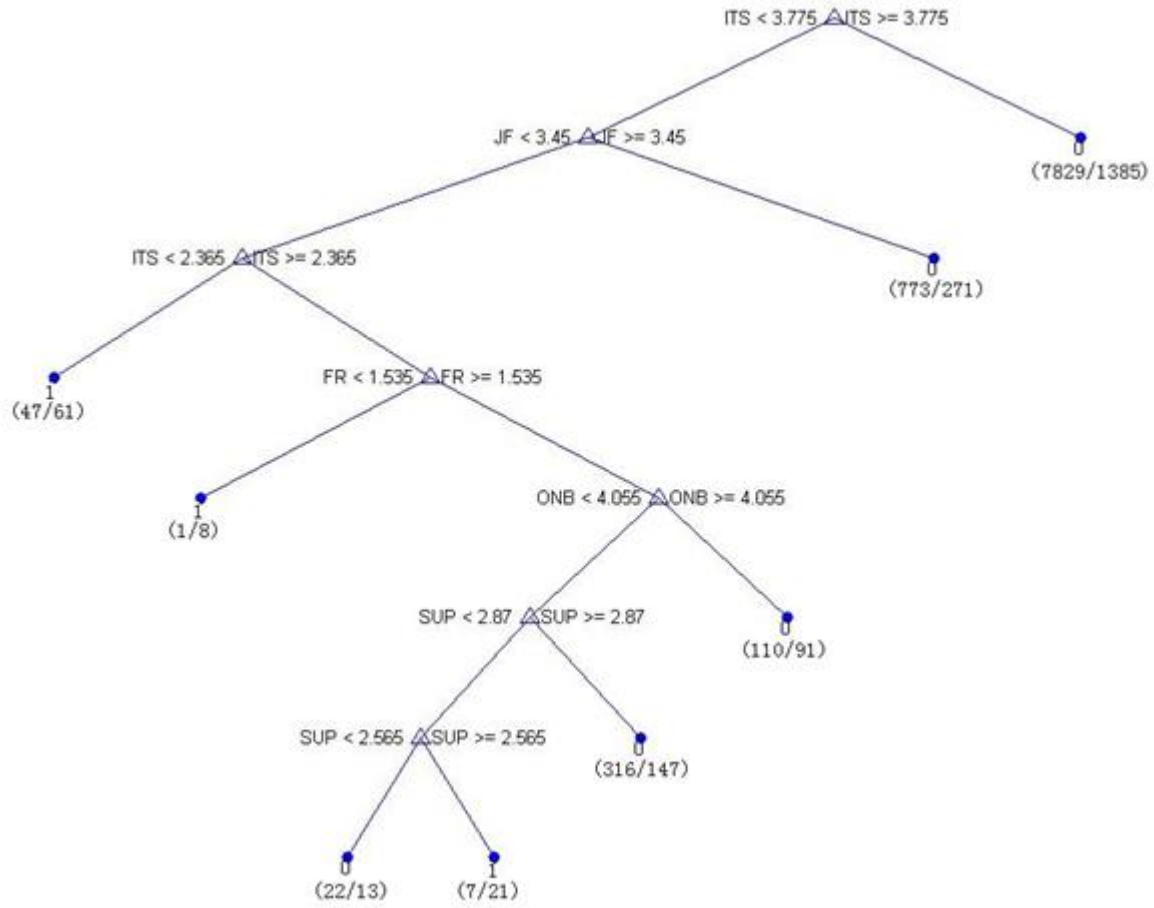

Figure 11. Initial Tree Diagram for Predicting Voluntary Turnover (Study 3)

Note. "0" = stay; "1" = voluntary turnover; number in parentheses represents number of stayed/left employees in each terminal node. ITS = Intent to Stay, JF = Job Fit, FR = Financial Rewards, ONB = Onboarding, SUP = Supervisor.

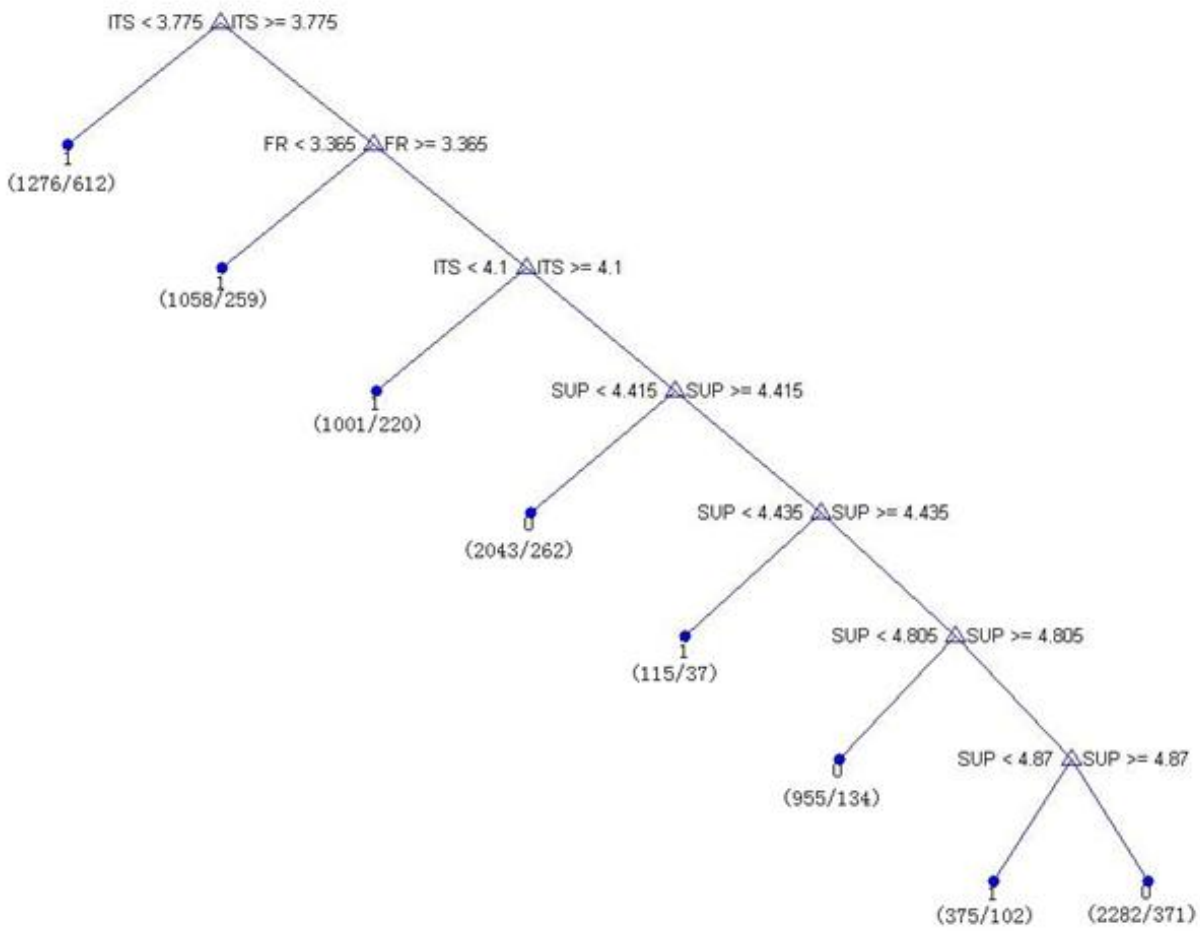

Figure 12. Final Tree Diagram for Predicting Voluntary Turnover (Study 3)

Note. "0" = stay; "1" = voluntary turnover; number in parentheses represents number of stayed/left employees in each terminal node. ITS = Intent to Stay, FR = Financial Rewards, SUP = Supervisor.

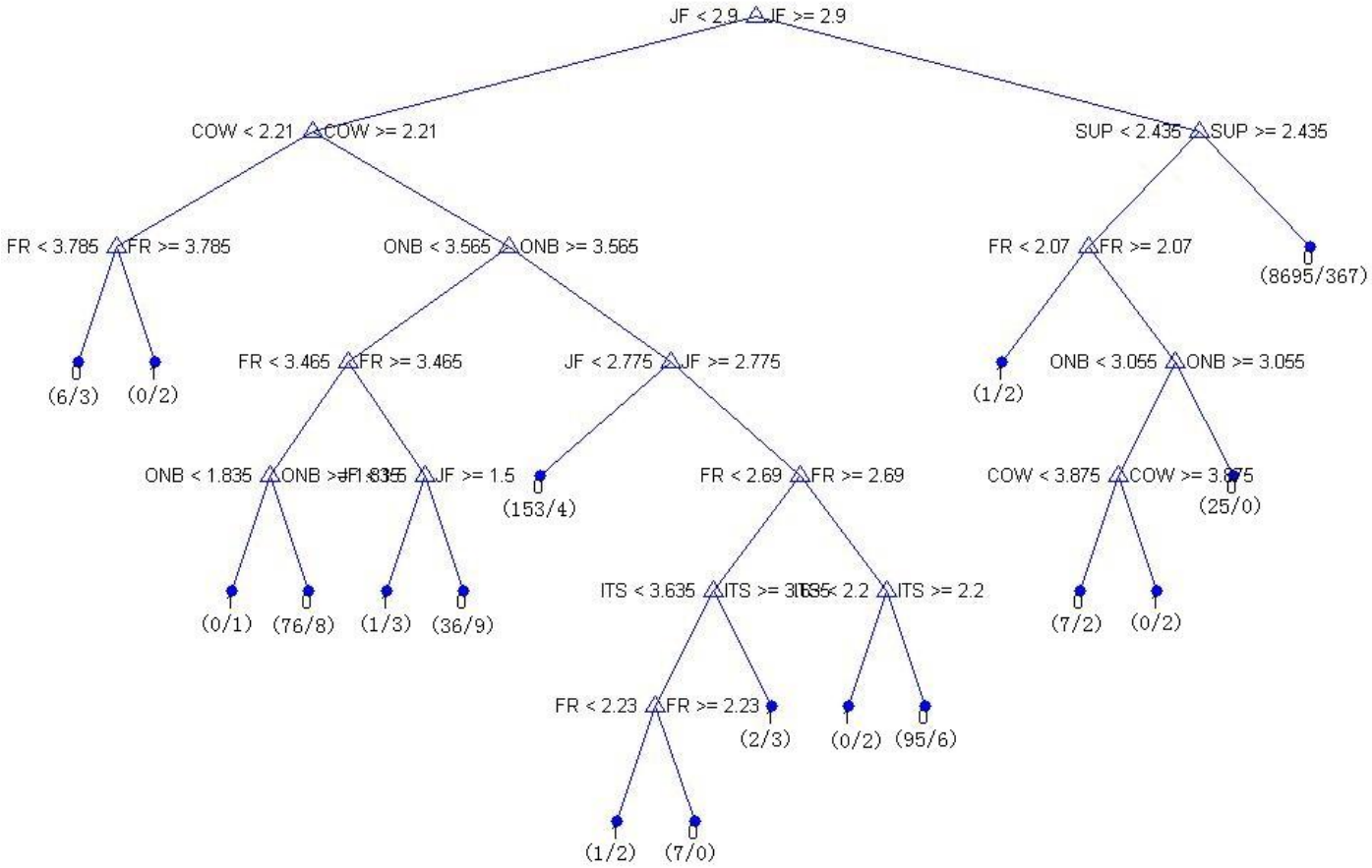

Figure 13. Initial Tree Diagram for Predicting Involuntary Turnover (Study 3)

Note. "0" = stay; "1" = involuntary turnover; number in parentheses represents number of stayed/left employees in each terminal node. JF = Job Fit, COW = Coworker, SUP = Supervisor, FR = Financial Rewards, ONB = Onboarding, ITS = Intent to Stay.

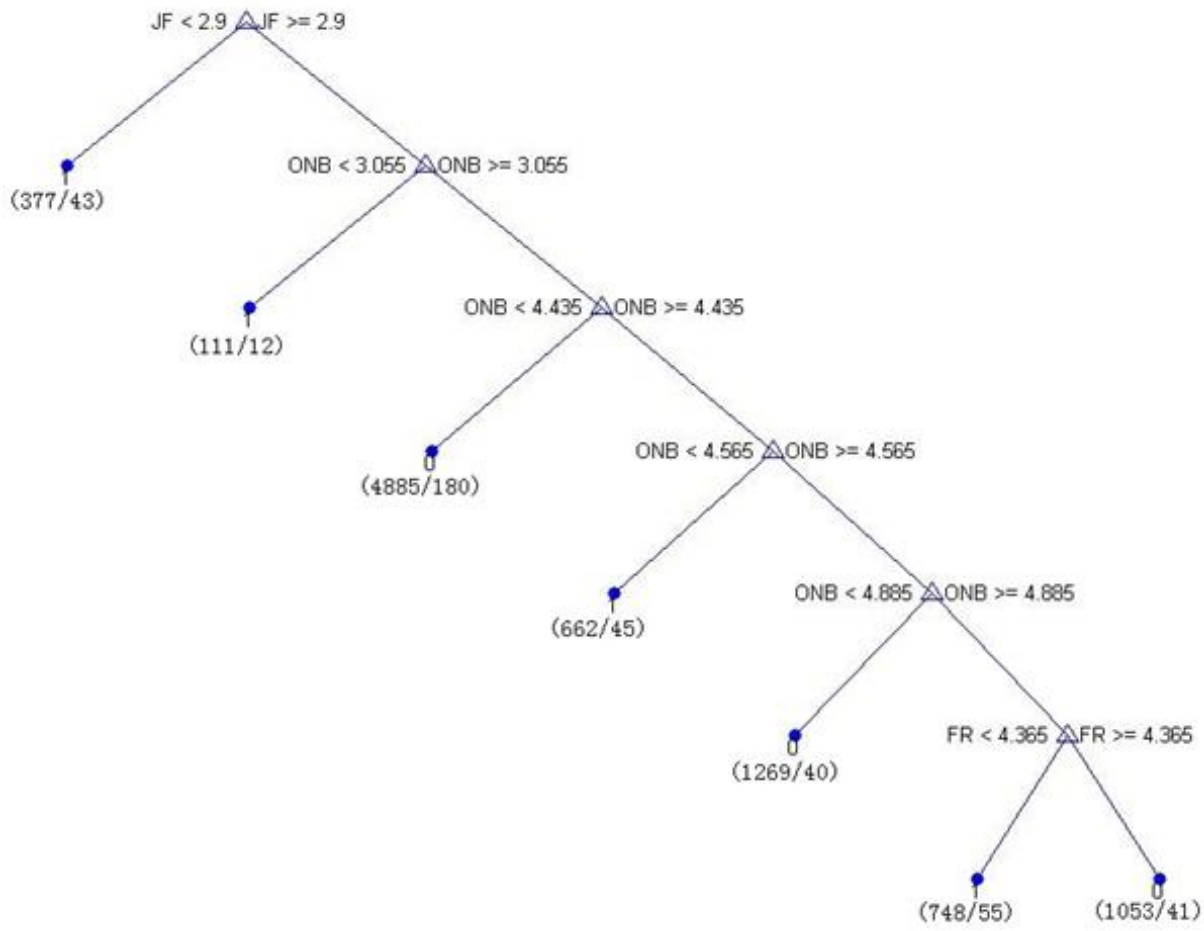

Figure 14. Final Tree Diagram for Predicting Involuntary Turnover (Study 3)

Note. "0" = stay; "1" = involuntary turnover; number in parentheses represents number of stayed/left employees in each terminal node. JF = Job Fit, ONB = Onboarding, FR = Financial Rewards.

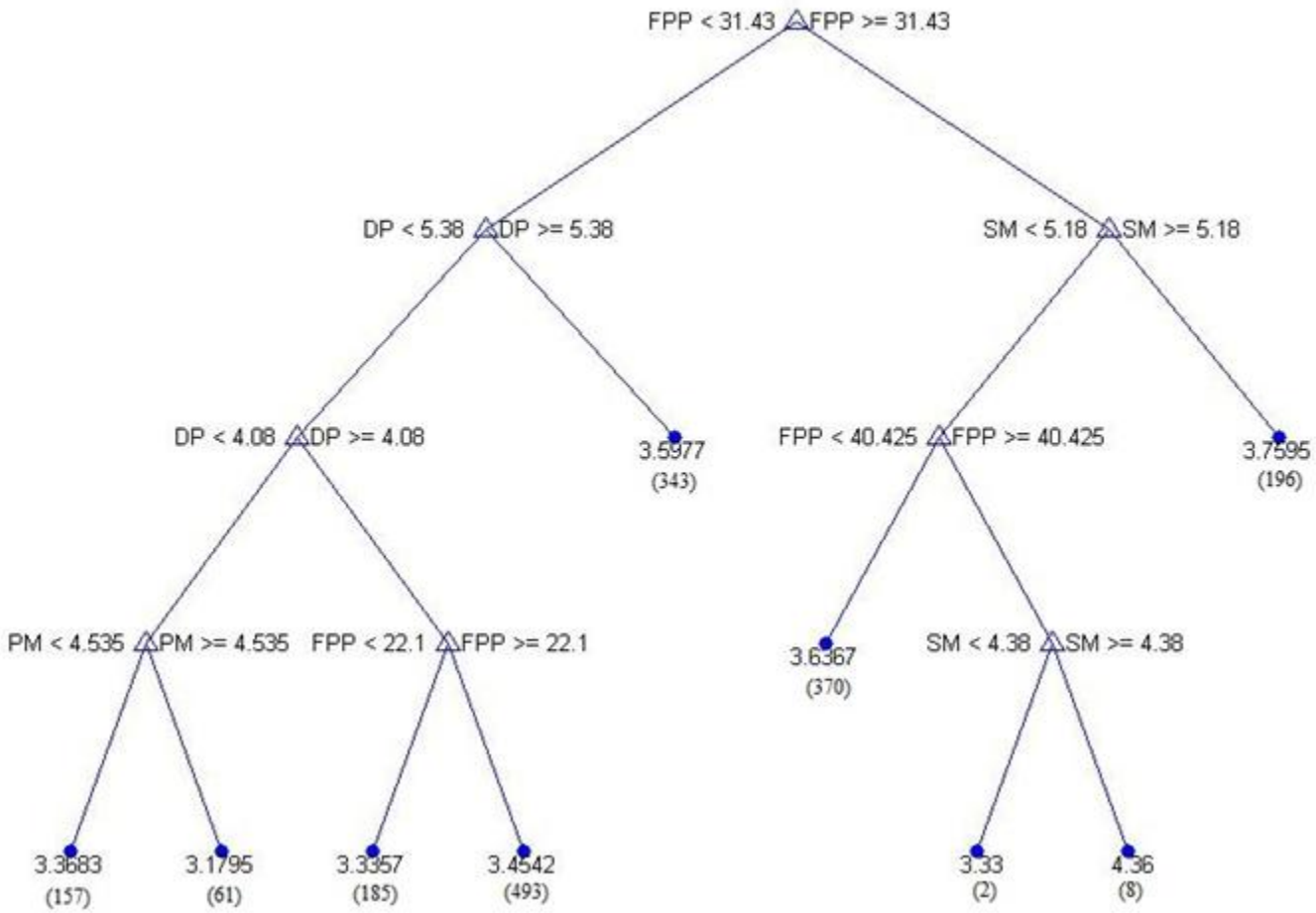

Figure 15. Regression Tree for Predicting Job Performance (Study 4)

Note. Numbers in terminal nodes represent the average response value in that node; numbers in parentheses represent node size. FPP = Following Policies and Procedures, DP = Dependability, SM = Self-Management Skills, PM = Personal Motivation.