The Leadership Quarterly 30 (2019) 417–426

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Short Communication

Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning?

Brian R. Spisak[a,b,*], Paul A. van der Laken[c], Brian M. Doornenbal[b]

[a] *University of Otago, New Zealand*
[b] *Vrije Universiteit Amsterdam, Netherlands*
[c] *Tilburg University, Netherlands*

ARTICLE INFO

Keywords:
Leader trait paradigm
Context
Machine learning

ABSTRACT

Using self-report personality data and 360-degree performance evaluations of 973 managers across various contexts, we investigated the leader trait paradigm using a range of machine learning methods. We found that a relatively simple linear ordinary least squares model incorporating direct effects of traits and context performed equally as well as our best performing complex machine learning alternatives (e.g., lasso and random forests) at predicting leader effectiveness under low-dimension conditions (i.e., a small number of predictors). We then increased dimensionality and found that newer machine learning methods excelled. Overall, our computationally intensive approach supports the argument that (a) direct effects (not interactions) of traits *and* context are important predictors of leader effectiveness and (b) appropriately matching combinations of methods, models, and data (from simple and conventional to complex and novel) creates a powerful machine learning engine for investigating leadership. We end with opportunities for future research, discuss practical implications, and provide a list of resources for those interested in learning more about this analytical future.

Uncovering which characteristics make some individuals more suitable as leaders than others is a perennial quest (Day & Antonakis, 2012; Tuncdogan, Acar, & Stam, 2017). One of the most longstanding pursuits within this vast body of work is predicting leadership effectiveness based on consistent traits such as personality factors (Judge, Piccolo, & Kosalka, 2009). Although researchers continue to find support for the relationship between personality traits and leadership (De Vries, 2012), some scholars question the relevance of the leader trait paradigm (Morgeson et al., 2007). Responding to this criticism, researchers advocate for a more complex perspective, where the context activates the effects of traits (Phaneuf, Boudrias, Rousseau, & Brunelle, 2016; Tett & Burnett, 2003), and the focus shifts away from direct main effects of separate traits towards nonlinear and interaction effects (Jensen & Patel, 2011). Answering this call for complexity, as we will demonstrate, brings clarity to the trait perspective, and broadly sets the stage for a new era of analytics in leadership research.

Scholars first provided support for advancements in leader trait research by demonstrating that pairs of traits, such as conscientiousness and emotional stability, interact on leadership outcomes (King, George, & Hebl, 2005). Leaders, however, display a broad range of traits – far beyond pairs (Judge et al., 2009) – and operate in diverse situations (Tett & Burnett, 2003). Thus, studies focusing on only two-way interactions and overlooking context are perhaps missing the actual complexity of traits. Fortunately, we are on the cusp of a methodological revolution commonly referred to as machine learning (ML) – i.e., an automated computational process for "learning" patterns in data from repeated experience to improve performance on tasks such as prediction (Kuhn & Johnson, 2013).

ML is a powerful engine for discovery and consequently an important way forward in the social sciences (e.g., Joel, Eastwick, & Finkel, 2017; Yarkoni & Westfall, 2017). Algorithmic ML methods such as random forests are ideally suited for situations that call for identifying optimal sets of potential covariates, examining the effects of multiple interactions simultaneously, and incorporating nonlinear relationships (Breiman, 2001; Efron, Hastie, Johnstone, & Tibshirani, 2004; Strobl, Malley, & Tutz, 2009). ML also benefits the external validity of findings by introducing cross-validation (of a magnitude greater than simple train/test resampling procedures), and has the added advantage of "naïve" data exploration without needing to specify a (complex) model beforehand. Likewise, ML can introduce "regularization" over many iterations for "shrinking" prediction error to drastically improve out-of-sample predictions. Thus, ML allows us to test and scrutinize competing leadership theories at a level never before possible (e.g., pitting the predictive performance of traditional leader

[*] Corresponding author at: Department of Management, University of Otago, PO Box 56, Dunedin 9054, New Zealand.
E-mail address: brian.spisak@otago.ac.nz (B.R. Spisak).

https://doi.org/10.1016/j.leaqua.2019.05.005
Received 30 April 2018; Received in revised form 14 May 2019; Accepted 22 May 2019
Available online 11 June 2019
1048-9843/ © 2019 Elsevier Inc. All rights reserved.

trait assumptions against the interactionist perspective). In other words, we can put the leader trait paradigm "through the (predictive) wringer" using ML.

However, though using a broader variety of ML potentially answers the call for advancement within the leader trait paradigm (Jensen & Patel, 2011; Judge et al., 2009; Tuncdogan et al., 2017), we argue that *prudence is necessary*. Many of the newer ML methods are designed for larger data volumes (i.e., larger sample sizes) and higher dimensionality (i.e., a larger number of features such as predictor variables) in a dataset than is the norm in leadership research. Moreover, while ML advancements can help with prediction, the outcomes are often less clear-cut to interpret than conventional approaches (Shmueli, 2010). This raises the question to what extent using a variety of ML can clarify the relationship between traits and leader effectiveness (or any other leadership relationships). Does the leader trait paradigm indeed require additional complexity to take its place in the pantheon of leadership literature, or is the longstanding methodological tradition of direct effects and simple two-way interactions sufficient for explaining effectiveness? To answer this question, we align with recent work stressing the combined value of "classical and new methods" in the reexamination of seemingly "elegant explanatory stor[ies]" (Yarkoni & Westfall, 2017, p. 1118).

Accordingly, with this short communication, we initiate the next generation response to the call for a focus on nonlinear effects and combinations of trait-context interactions (Jensen & Patel, 2011; Judge et al., 2009; Tuncdogan et al., 2017) by investigating the (potential) added value of ML beyond the commonly used ordinary least squares (OLS) regression approach. We accomplish this in the current study by analyzing multiple personality and contextual predictors of leader effectiveness across a range of analytical methods – from established and relatively simple to new and increasingly complex. The outcome of this investigation will provide answers to fiercely persistent questions regarding the leader trait paradigm, and more broadly help to clarify the (mis)application of ML. In particular, we address the following questions:

> (a) To what extent can personality predict leader effectiveness, (b) to what extent can personality-trait interactions add predictive validity, (c) to what extent are these personality-effectiveness relationships context-dependent, and (d) how do novel versus conventional analytical methods compare in terms of predictive performance?

Thus, the primary goals are exploring how well leader traits predict effectiveness, testing to what degree incorporating context helps in this exploration, and examining what (if anything) newer ML methods add. In short, we examine the extent to which predictive modeling supports or challenges existing explanatory accounts of the leader trait paradigm as we advance our analysis from simple to complex methods and models.

Finally, we do not make specific hypotheses. We, instead, utilize data-driven predictive modeling (as opposed to hypothesis-driven explanatory modeling), and explore predictive validity across methods as a straightforward way of comparing a simple versus an interactionist perspective of the leader trait paradigm (for a review of prediction versus explanation see Shmueli, 2010 and Yarkoni & Westfall, 2017).

Method

Participants

The respondents for our analysis were 973 leaders (after excluding 143 respondents from the original sample due to missing data). The leaders included 60.2% men and 24.3% women (15.5% declined to answer) with a mean age of 41.11 (*SD* = 7.96; 20.5% declined to answer). Most of these leaders (57.1%) worked in operational roles (i.e., line management) and the remaining 42.9% had a more strategic job position (i.e., top executives, board members, GMs, VPs, and divisional leaders). They also came from various employers with the majority employed in private sector organizations (67.7%) and the remainder in public/non-profit organizations (32.3%). As for company size, 41.2% of the respondents worked in organizations with a 1000+ employees, 40.3% with 200–999 employees, and 18.5% with 0–199 employees.

Table 1
Relationship between the Five-Factor Model (FFM) and the Hogan Personality Inventory (HPI).

FFM dimensions	HPI dimensions
Neuroticism	Adjustment
Extraversion	Ambition and Sociability
Agreeableness	Interpersonal Sensitivity
Conscientiousness	Prudence
Openness	Inquisitive and Learning Approach

Measures and procedure

Trait measures

We obtained secondary data from a dataset of leaders who completed the Hogan Personality Inventory (HPI) – a seven-factor assessment corresponding to the five-factor model (FFM) of personality (see Table 1; Hogan & Holland, 2003). This dataset included the following traits: adjustment (α = 0.76; 8 items; example item: I keep calm in a crisis), ambition (α = 0.63; 6 items; example item: I am a very ambitious person), sociability (α = 0.63; 5 items; example item: I am often the life of the party), interpersonal sensitivity (α = 0.63; 5 items; example item: I try to see the other person's point of view), prudence (α = 0.56; 7 items; example item: I strive for perfection in everything I do), inquisitive (α = 0.60; 6 items; example item: I like trying new, exotic types of food), and learning approach (α = 0.48; 4 items; example item: As a child, school was easy for me). These traits were measured on a scale ranging from very much disagree (=1) to very much agree (=5), and we included all these traits in our analysis. Though the alphas for several of the factors are low, existing meta-analysis finds average internal consistency reliability of the factors ranging from 0.71 (prudence) to 0.89 (adjustment; Hogan & Holland, 2003). Accordingly, we decided not to delete items (and obtain higher scores) to keep our findings more comparable to the large number of studies employing the HPI. Furthermore, the flexibility of our ML methods allowed us to simply add models incorporating all 41 HPI items (in addition to the HPI factor models).

Context and demographic measures

The dataset included several other relevant variables. We incorporated the following in our analysis to serve as an additional layer of complexity in the leader trait paradigm: sector (private vs. public sectors), leaders' job level (operational vs. strategic), organization size (small vs. medium vs. large), and gender (female vs. male vs. unknown). We included these variables because they are listed as relevant in the interactionist literature (i.e., the moderating effects of task, social, and organizational variables; Tett & Burnett, 2003).

Across sectors, different traits are needed as the focus shifts from shareholder to stakeholder orientations (Donaldson & Preston, 1995; Van der Wal, De Graaf, & Lasthuizen, 2008). Likewise, job-level is important because those rising to strategic levels of leadership experience distinct task demands (Gilboa, Shirom, Fried, & Cooper, 2008), and perhaps rely on different characteristics for effectiveness relative to those at the operational level of an organization (e.g., Resick, Whitman, Weingarden, & Hiller, 2009). Organization size is also important because the requirements for leading large organizations may differ significantly from small enterprises (Vaccaro, Jansen, Van Den Bosch, & Volberda, 2012). Finally, of the demographic variables, missing data for gender was minimal, and we included it as a relevant variable given that people often demand different behavior from men than from women (Eagly & Karau, 2002).

Effectiveness measures

The dataset included a 360-degree measure of leader effectiveness where four rater types (supervisors, peers, direct reports, and self) evaluated the target leader's performance on four dimensions (i.e., self-management, relationship management, working in the business, and working on the business) that were further divided into fourteen "themes." Respectively, theme items in these dimensions included integrity, communication, efficiency, and strategic planning. Leaders scoring high on self-management and relationship management exhibit above average intra- and interpersonal effectiveness (i.e., emotional and social skills), and leaders scoring high on working in the business and working on the business exhibit above average cognitive effectiveness for accomplishing and coordinating work-related tasks (i.e., operational and strategic competencies). Collectively, the measure captures a 360-perspective of (a) the types of leadership skills the leader possesses and (b) how effective they are (perceived) at applying those skills to the (leadership) context. These traits were measured on a scale ranging from least favorable (=1) to most favorable (=8), and we included all these traits in our analysis. We averaged the performance measures to generate the overall 360-degree score ($\alpha = 0.87$). For further details of the tool and research demonstrating its reliability and validity see Peter Berry Consultancy (2016). Likewise, a technical manual is available upon request from the consultancy for additional information such as evidence pertaining to construct and criterion validity (Peter Berry Consultancy & Hogan Assessment Systems, 2019).

We generated an overall 360-degree score to address the call for measurement "triangulation" of effectiveness relative to single measures (Richard, Devinney, Yip, & Johnson, 2009). Arguably, one reason for the single-dimension approach in applied research is simply the practical challenges with gathering a sufficient amount of multi-dimensional data. Though our dataset is not perfect (as we discuss below), one of its major strengths is the quantity of multi-dimensional performance data. We feel this represents a unique advantage given the observed value of 360-degree performance rating systems. As Oh and Berry (2009, p. 1510) state:

> Personality has been studied as one of the predictors [of managerial performance] and at the same time has often encountered skepticism for its low validity. Our results demonstrate that estimates of the validity of personality for predicting managerial performance are greater when managerial performance is assessed from multiple perspectives by utilizing the 360-degree performance rating system.

Simultaneously considering performance at multiple levels (i.e., 360-degree triangulation) provides a powerful lens for investigating the holistic relationship between traits and leader effectiveness in complex, multilevel organizations. Otherwise, the potential for selection bias exists where, for instance, a leader is effective when dealing with superiors at a strategic level but is generally ineffective with employees at an operational level.

The leaders completed the HPI and 360-degree evaluation as part of their employer's standard training and development practices, they provided informed consent, and were debriefed. All procedures in the current study received ethics approval. The archival data we analyzed is not under our direct control. Requests to access the data should be directed to Hogan Assessments. See Table 2 for means, standard deviations, and correlations among the study variables.

Analytical procedure

The purpose of our analysis was to explore the leader trait paradigm and the added value of newer and more complex ML (i.e., regularized regression and algorithmic methods) versus the commonly utilized OLS regression method. Conventional methods deliver interpretability, but lack the predictive potential of newer ML. Conversely, modern ML methods provide predictive power because they are free to explore complex interactions and nonlinear effects, but we cannot easily interpret the modeled relationships between variables (see Breiman, 2001). Then there are regularized regression methods (e.g., ridge regression) which optimize predictive generalizations yet provide somewhat interpretable results. Hence, we are exploring the benefits of complexity and flexibility versus interpretability and transparency along a continuum ranging from OLS regression and regularized regression to advanced algorithmic methods.

Analytical methods. In predicting leadership performance, we built and compared 32 different models using different combinations of methods and predictors. We selected five of the most commonly utilized methods (Kuhn & Johnson, 2013): OLS, ridge regression (RIDGE), least absolute shrinkage and selection operator (LASSO), gradient boosting machine (GBM), and random forests (RF).

OLS is probably the most frequently employed method in psychology research, hence we denote it as the conventional method. OLS estimates the parameters of a linear function of a set of predictor variables by the least squares principle (i.e., it seeks to minimize the sum of the squares of the differences between the observed outcome variable and the values predicted by the linear function, in other words, the sum of the squared errors). This method, however, struggles with multicollinearity and can easily overfit the training data at hand. Thus, OLS provides simplicity and clarity, but cannot fully appreciate underlying patterns in complex data.

To better manage the tradeoff between overfitting and underfitting, RIDGE and LASSO regression are frequently utilized. These methods are quite similar to OLS regression where they seek to minimize the sum of squared errors, but differ by also incorporating an additional penalty term to counteract overfit. For RIDGE regression, this penalty term, called the L2 penalty, consists of the sum of the squared values of the coefficients multiplied by a parameter λ. If this λ value is set to zero, the RIDGE model behaves the same as an OLS model (i.e., it minimizes the sum of squared errors). However, the higher the λ, the more a RIDGE model is inclined to decrease the largest of its coefficients in light of their predictive value. This simply means it will shrink the impact of overfitting in case of multicollinearity, and any irrelevant features on the trained model will have their coefficients minimalized (but not set to zero).

LASSO regression differs in that it employs the so-called L1 penalty term which consists of the sum of the absolute values of the coefficients multiplied by λ. Hence, for high values of λ, a LASSO model is inclined to decrease its coefficients to exactly zero (i.e., it removes irrelevant features from the model entirely rather than just minimizing their impact). Adding either one of these penalty terms for coefficient size decreases the variance of the model's predictions by introducing some bias in the model. This method of adding bias to reduce variance is called regularization, and hence we denote RIDGE and LASSO as regularized regression methods.

Finally, GBM and RF are tree-based ensemble models. Tree-based models do not model linear relationships like OLS and its variants. Rather, they iteratively split a data sample into subsamples based on optimal demarcations (i.e., cut-points that progressively minimize error in prediction). At every split, the model picks the optimal demarcations on one of the predictor variables (e.g., a personality factor) that splits the sample in a way that minimizes the sum of squared prediction errors (in case of regression) on the outcome variable in the subsamples (Breiman, Friedman, Stone, & Olshen, 1984). Simply put, both GBM and RF partition off outcome variable data based on the available predictor variables to create a predictive framework of reality. This results in a tree-like structure that models the underlying relationships in a dataset. One can then use this structure to make predictions for new data as it filters through the branches.

The ensemble part refers to the fact that GBM and RF models are comprised of many underlying "weak" models (e.g., high bias, low

Table 2
Means, standard deviations, and correlations among the study variables.

Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Hogan360	5.47	0.40												
2. Adjustment	3.46	0.74	.175***											
3. Ambition	4.15	0.63	.196***	.442$_{***}$										
4. Sociability	2.75	0.90	.083*	.087**	.390***									
5. Interpersonal sensitivity	3.69	0.58	.222***	.475***	.374***	.349***								
6. Prudence	2.88	0.59	.152***	.364***	.090**	−.259***	.294***							
7. Inquisitive	2.50	0.71	0.003	.151***	.213***	.356***	.187***	−.091**						
8. Learning approach	2.27	0.71	.105**	.210***	.245***	.155***	.152***	.029	.301***					
9. Sector[a]	0.68	0.47	−0.049	−0.055	−0.016	−0.059	−.099**	0.020	−0.043	0.003				
10. Job level[b]	0.43	0.50	0.007	0.045	.169***	0.035	−0.041	−0.041	−0.019	.067*	0.056			
11. Company size[c]	2.23	0.74	.147***	0.004	0.011	0.034	.064*	0.019	0.045	.074*	−.065*	−.165***		
12. Gender Male[d]	0.60	0.49	−.143***	0.024	0.027	0.046	−.100**	−0.033	.091**	−.073*	.090**	.114***	−.077*	
13. Gender Female[d]	0.24	0.43	.156***	−.068*	−0.051	−0.029	.142***	0.028	−.141***	.076*	−.107***	−.088**	.076*	−.696***

Note. a) Referent sector = non-profit, b) referent job-level = operational, c) company size: 1 ≤ 200, 2 = 200–999, 3 ≥ 999, and d) referent gender = other gender or unknown.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

variance models), whose predictions are aggregated to one single prediction. These underlying models are considered weak as they usually do not predict the data well individually, even though, collectively, their aggregated predictions are often quite accurate. Both GBM and RF also include many optimization and regularization parameters as well as complex meta-algorithms that help to avoid overfit. For instance, GBM and RF make use of bootstrap aggregation, or bagging, where each of the underlying weak models is trained on a different bootstrapped sample of the data (Breiman, 1996). Moreover, boosting lies at the core of the GBM algorithm, where each subsequent weak model is trained to learn from the errors of its predecessor (Freund, Schapire, & Abe, 1999). Because of their more iterative nature and computational complexities, we denote GBM and RF as algorithmic methods.

Regarding the value of using these various ML methods, Yarkoni and Westfall (2017) state:

> …the relative performance of different kinds of machine learning algorithms can potentially provide important insights into the nature of the data. For instance, if lasso regression outperforms ridge regression …, then one might conjecture that the underlying causal graph is relatively sparse [recalling that lasso removes irrelevant features]. If…random forests outperform standard regression models, then there may be relatively large low-order (e.g., two-way) interactions or other nonlinearities that the machine learning methods implicitly capture (p. 1116-1117).

Also, while the above is only a brief description of the complex inner workings of these five methods, more detailed outlines have been published (see Hoerl & Kennard, 1970 for RIDGE; Tibshirani, 1996 for LASSO; Friedman, 2001 for GBM; Breiman, 2001 for RF). Further, for broad overviews see Friedman, Hastie, and Tibshirani (2001) as well as James, Witten, Hastie, and Tibshirani (2013).

Leadership effectiveness predictors. Using HPI factors and context as predictors of leadership effectiveness, we investigated traits (e.g., OLS_t), traits and context variables (e.g., OLS_tc), traits and context variables and interactions between traits (e.g., OLS_t*t), and traits and context variables and interaction between traits and context variables (e.g., OLS_t*c). Further, RF and GBM methods may automatically include interactions between all the available predictors, thus only two different models were built for each of these methods: one with only trait data (e.g., RF_t) and one with trait and context variables (e.g., RF_t**c). We then repeated these same combinations for the HPI items (in addition to the HPI factor models).

Finally, it is important to note that the most complex OLS, LASSO, and RIDGE models included two-way interactions whereas GBM and RF are free to consider any n-way combinations of the predictors such as three-way interactions (e.g., trait by trait by context). As we will demonstrate, this freedom to explore complexity endows the algorithmic models with an advantage under the right circumstances.

Layers of analysis. Our analysis was conducted broadly across two layers: (a) the trait layer and (b) the trait and context layer. The trait layer is a predictive layer in which we investigated HPI factor direct effects on leader effectiveness. In the trait and context layer, we explored to what degree the predictive performance of OLS, regularized regression, and algorithmic methods change when complexity is added to the leader trait paradigm. Specifically, we started to increase dimensionality by incorporating interactions and organizational context into the HPI factor models (e.g., HPI factor and context interactions). We then generated a number of (tentative) explanatory insights gleaned from our predictive analysis of HPI factors and context. Finally, as mentioned above, we analyzed HPI survey items (in addition to factors) given that a significant advantage of our regularized and algorithmic methods as a means of prediction is the ability to utilize highly dimensional datasets. As predictors, we analyzed the 41 HPI items, the 6 context items, and their many interactions in this highly dimensional space.

Collectively, our analysis provided us with data ranging from lower dimensionality (i.e., HPI *factor* direct effects consisting of 7 predictors) to higher dimensionality (i.e., HPI *item* direct and interaction effects consisting of 1681 predictors). Hence, we investigated the tradeoff between interpretability and predictive performance by comparing multiple methods, models, and varying degrees of dimensionality.

Model construction and deployment. Following procedures outlined by Kuhn and Johnson (2013), we built the models based on normalized variables and complete cases. To predict performance, we used our five chosen methods (i.e., OLS, LASSO, RIDGE, GBM, and RF) in combination with each set of predictors for both HPI factors and HPI items (i.e., HPI, HPI and context, interactions between HPI, and interactions between HPI and context). Then, we had to determine the optimal parameters values for four of the methods, and separately for each model using these four methods. The two regularized methods required a λ-value to be set, and both GBM and RF have a range of different parameters to tune, including the minimal node size, the number of trees, or their maximum depth. Basically, different parameter settings can result in different models, and one wants to use values that result in the most predictive model. Hence, prior to our

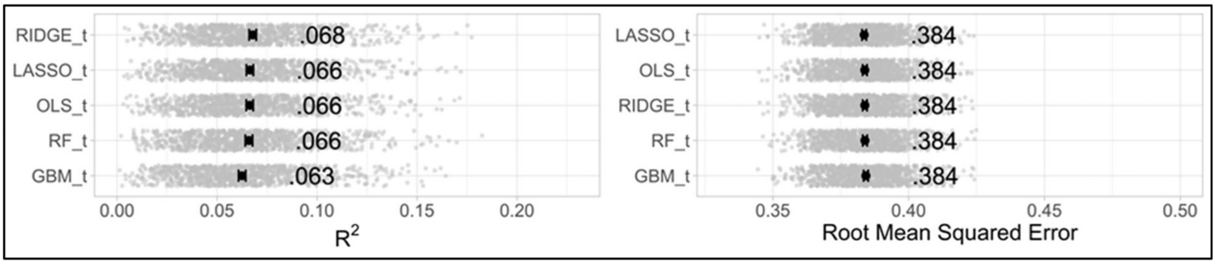

Fig. 1. Trait layer: Out-of-sample model performance for predicting the overall 360-degree effectiveness score. Dots represent the model performance in the 1000 test samples. The squares and error bars represent the average model performance and 95% confidence intervals. Models are sorted based on their average performance.

main analysis, we ran a so-called parameter tuning routine where we explored wide parameter grids for each of our models and, using cross-validation, we retrieved the best parameter settings among them. The parameter values input in these initial grids were determined based on common sense, experience, literature, and expert recommendations (e.g., Ridgeway, 2017 for GBM). The parameter settings we explored and chose for the models are included as online supplementary material.

Further, to get a robust sense of true model parameters, we used a Monte Carlo cross-validation routine, simulating many different situations by subsampling 1000 times from our dataset for each method and each set of predictors. In total, we generated 32,000 different models (i.e., [3 linear methods * 4 models] + [2 algorithmic methods * 2 models] * 1000 samples each * 2 for HPI factor data and HPI item data), training these on random 75% of the data, and assessing the out-of-sample predictive validity (i.e., model performance) using the remaining 25% test sample.

Whereas aggregating the external performance based on 50–200 of such datasets helps "to get a stable estimate of model performance" (Kuhn & Johnson, 2013; p. 72), the 1000 datasets we generated for each model added to our certainty in the performance estimates by pushing our cross-validated situations to an extreme. In fact, our approach was so computationally intensive that it could not be done on a personal computer, and we had to run the models in the cloud. Lastly, we compared model performance based on 95% confidence intervals, and conducted the analyses using *caret* (Kuhn, 2017) in R; with *elasticnet* (Zhou, 2013) for the LASSO and RIDGE models, *gbm* (Ridgeway, 2017) for the GBM models, and *ranger* (Wright, Wager, & Probst, 2018) for the RF models. The code for our procedure is available as supplementary material.

In comparing the performance of the different models, we calculated the root mean squared error (RMSE) and the coefficient of determination (R^2) resulting from the models' predictions in the test samples (i.e. out-of-sample performance). The RMSE indicates how far (on average) the model residuals are from zero. This distance is larger when the model predictions are further from the actual values. A smaller RMSE thus suggest a more accurate model. The out-of-sample R^2 indicates to what extent the models explain the variance in the outcome variable (and generalize from the training to the test samples). The out-of-sample R^2 is larger when the (root-squared) correlation between the model predictions and the actual values is stronger. Higher out-of-sample R^2 values thus suggest that the models explain more variance in the outcome variable in the test samples.

Finally, before reporting the main results, it is important to note that our cross-validation procedure has implications for the interpretation of the R^2 and RMSE values we find. Specifically, if our leadership effectiveness scores were randomly distributed (i.e., were completely up to chance), one would expect to find R^2 values close to zero and RMSE values close to the standard deviation of leadership effectiveness. In the online supplementary materials, we illustrate such an outcome and generate a baseline against which we can compare the predictive validity of our models. To create this baseline, we built similar models (using the same above procedures) focused at predicting a normally distributed random variable with the same mean ($M = 5.47$) and standard deviation ($SD = 0.40$) as our actual target variable. As expected, the baseline models highest R^2 was close to zero ($R^2 = 0.007$) and the lowest RMSE was similar to our standard deviation (RMSE = 0.411). This "chance" output, as we demonstrate below, is dissimilar to our main results, thus providing additional support to the validity of our findings.

Results

Trait layer

At the trait layer of analysis, we started by exploring only direct trait effects on 360-degree ratings of leader effectiveness (i.e., our lowest dimensional data). We found that the models included in the trait layer did *not* perform significantly different from each other. The performance of the models (including GBM and RF which automatically consider possible interactions) were almost identical out to three decimal points: ranging from an average RMSE of 0.384 (95% CI [0.383, 0.385]) and an average R^2 of 0.063 (95% CI [0.061; 0.064]) to an average RMSE of 0.384 (95% CI [0.383, 0.385]) and an average R^2 of 0.068 (95% CI [0.066, 0.070]). Fig. 1 provides an overview of the trait models' out-of-sample performance.

Overall, our simplest (and conventional) linear trait-based model OLS_t (RMSE = 0.384, 95% CI [0.383; 0.385]; $R^2 = 0.066$, 95% CI [0.065, 0.068]) predicted leadership performance equally as well as the more complex regularized models LASSO_t (RMSE = 0.384, 95% CI [0.383; 0.384]; $R^2 = 0.066$, 95% CI [0.065, 0.068]) and RIDGE_t (RMSE = 0.384, 95% CI [0.383; 0.385]; $R^2 = 0.068$, 95% CI [0.066, 0.070]) as well as the algorithmic models RF_t (RMSE = 0.384, 95% CI [0.383; 0.385]; $R^2 = 0.066$, 95% CI [0.064, 0.068]) and GBM_t (RMSE = 0.384, 95% CI [0.383; 0.384]; $R^2 = 0.063$, 95% CI [0.061, 0.064]). These more complex models are, however, harder to interpret (particularly the algorithmic alternatives) because they do not provide meaningful and clear parameters that explain the relationship between HPI factors and leader effectiveness. Considering this decreased transparency without improvements in model performance, the simple OLS alternative is arguably the better option in the low-dimensional context of trait direct effects.

A reason for the lack of added value is that our data – which represents a typical dataset found in psychological research – may simply not be the best sort of fuel for the modern ML engine (i.e., the dimensionality is too low). Though, by psychological standards, we have a sufficient number of variables and our sample size is fairly large, the strength of the ML engine is that it runs optimally on larger data volumes and a greater number of dimensions (Shmueli, 2010). Indeed, a recent publication demonstrated this ML complexity advantage (Joel et al., 2017), but what is not yet well-established empirically is to what degree more complex ML methods actually help with developing the domain of leadership science relative to more traditional alternatives. Does adding context and multiple interactions, as a way to increase dimensionality, improve our understanding of the leader trait paradigm (e.g., the trait activation hypothesis; Tett & Burnett, 2003), and when model complexity is increased, will newer ML methods then outperform conventional methods?

Trait and context layer

Thus, in the trait and context layer, we introduced more complexity (i.e., dimensionality) to our analysis by (a) allowing HPI factor interactions, (b) accounting for direct context effects, and (c) exploring factor-context interactions using the abovementioned contextual and demographic variables: sector (private and public), job level (strategic leaders and operational leaders), company size (small, medium, and large), and gender (female, male, and unknown). We then compared the predictive performance of the conventional OLS models versus the regularized and algorithmic newcomers (e.g., RIDGE and RF). Our results demonstrated a number of important outcomes.

First, and most important, adding context information significantly improved our ability to predict leadership effectiveness. In fact, the OLS model that included simple linear effects (not interactions) of trait and context variables performed among the best (i.e., OLS_tc; RMSE = 0.379, 95% CI [0.378; 0.380]; $R^2 = 0.090$, 95% CI [0.088; 0.092]). Other methods and models that performed equally well included the regularized RIDGE model incorporating trait and context interactions (i.e., RIDGE_t*c; RMSE = 0.380, 95% CI [0.379; 0.381]; $R^2 = 0.091$, 95% CI = [0.089; 0.092]), the regularized RIDGE model with direct linear trait and context effects (i.e., RIDGE_tc; RMSE = 0.379, 95% CI [0.378; 0.380]; $R^2 = 0.091$, 95% CI = [0.089; 0.092]), the regularized LASSO model also with direct linear trait and context effects (i.e., LASSO_tc; RMSE = 0.379, 95% CI = [0.378, 0.380]; $R^2 = 0.088$, 95% CI = [0.086, 0.090]), and the algorithmic RF model with (potentially complex) trait and context (interaction) effects (i.e., RF_t**c; RMSE = 0.379, 95% CI = [0.378, 0.380]; $R^2 = 0.089$, 95% CI = [0.087, 0.091]).

Fig. 2 provides an overview of the trait and context models' out-of-sample performance (top rows) compared to the direct effect trait layer analysis (bottom rows). Comparing the two layers, we see approximately a 35% improvement in explained out-of-sample variance due to the addition of contextual information (e.g., a 36.4% increase from a mean OLS_t $R^2 = 0.066$ to a mean OLS_tc $R^2 = 0.090$). Thus, including information about the context in which these leaders operate helped explain their performance.

Second, though direct effects of context improved predictive performance, adding interactions between HPI factor traits and context produced no additional predictive validity. For instance, the GBM and RF models did *not* outperform simple models based on linear direct effects of traits and context despite their freedom to account for any relationships explaining variance. Thus, given the loss of transparency associated with the more complex ML newcomers, the conventional (direct effects) OLS model incorporating traits and context is arguably a better alternative for balancing interpretability with predictive performance. Hence, the added value of our ML approach to the leader trait paradigm does not necessarily need to lie in the improved ability to predict leadership effectiveness through complexity. Rather, we thoroughly explored all possible complex relationships (to a degree beyond existing research) and concluded that in low-dimensional settings it is best to stick to simple linear models of traits and context.

In the next section, we briefly touch on the explanatory insights gleaned from our more transparent predictive models. Following that, we raise the dimensionality of our data to demonstrate how the tradeoff between interpretability and predictive performance shifts in favor of less transparent ML modeling of the leader trait paradigm.

Explanatory insights from the trait and context layer

To connect our predictive modeling with future explanatory research of the leader trait paradigm, we chose to average the coefficients for the 1000 iterations of our OLS trait and context direct effects models (i.e., OLS_tc). This was a logical choice given that the predictive performance of this relatively simple model was equivalent to the best-performing regularized and algorithmic alternatives while at the same time delivering the most interpretable outcomes. The following significant predictors of leader effectiveness emerged over the many iterations: (a) HPI ambition which corresponds to FFM extraversion ($b = 0.048$, $se = 0.015$, $p = .001$), (b) HPI interpersonal sensitivity which corresponds to FFM agreeableness ($b = 0.036$, $se = 0.016$, $p = .023$), (c) HPI prudence which corresponds to FFM conscientiousness ($b = 0.038$, $se = 0.014$, $p = .007$), (d) organizational size, where medium and large size had a positive effect on leader effectiveness (respectively, $b = 0.037$, $se = 0.017$, $p = .031$ and $b = 0.068$, $se = 0.017$, $p < .001$), and (e) gender, where being male had a negative effect on leader effectiveness ($b = -0.064$, $se = 0.015$, $p < .001$).

Also, regarding interpretability, as the methods increased in complexity (i.e., RIDGE, LASSO, GBM, and RF), it became increasingly difficult to extract meaningful averaged coefficients. We were able to retrieve coefficients for RIDGE and LASSO, but they varied greatly across the samples as these regularized methods worked towards their predictive goal. Further, GBM and RF were even less interpretable given their freedom to explore a variety of (potentially complex) effects over the 1000 iterations. Thus, this additional output is in no way definitive. It is simply a tangible suggestion for integrating predictive ML concepts into the (explanatory) leader trait literature. The fundamental takeaway is that conventional models incorporating simple direct effects of traits and contexts (not interactions) predicted leadership performance equally as well as much more complex alternatives while providing interpretable insights worthy of future explanatory research.

We can make this “keep it simple” claim because our data was of low dimensionality and newer ML methods are designed to handle highly dimensional datasets (i.e., the engine was not able demonstrate its full potential). As mentioned, the regularized linear methods can either shrink non-important parameters (RIDGE) or set them to zero (LASSO), and the algorithmic alternatives such as RF can capture complex nonlinear signals in the data. Utilizing these capabilities to predict leader effectiveness provides a significant advantage over more conventional (OLS) approaches when the appropriate, high-dimensional fuel is available (Breiman, 2001).

Predicting leadership effectiveness with high-dimensional data

To increase dimensionality and explore settings where the value of novel ML emerges, we investigated predictive performance in the trait and context layers using HPI items rather than factors. Utilizing various item and context models (some with over 1600 predictors) created the conditions for extracting more from the ML engine and allowed us to significantly improve our sense of externally valid parameters. The average external predictive validity (including confidence intervals) of the HPI *item* models in their test samples are demonstrated in Fig. 3.

This figure indeed gives a better sense of the true predictive validity of each method, for each set of predictors. For instance, comparing Figs. 2 with 3 suggests that the top-performing HPI *item* models outperformed the top-performing HPI *factor* models. Going further, when data dimensionality increased due to replacing HPI factors with HPI items, the regularized and algorithmic newcomers immediately started to outperform OLS. We observed that even with the simplest HPI *item* models incorporating only the direct effects of each item (i.e., 41 predictors), RIDGE and LASSO outperformed OLS because they were capable of ignoring variables that were less predictive (i.e., regularization by shrinking non-important parameters or setting them to zero). Additionally, when taking items *and* context into consideration (thus further increasing dimensionality), RIDGE, LASSO, and RF were able to focus on the most important predictors among the wide range of 47 predictors (i.e., HPI items and context), and hence we observed an

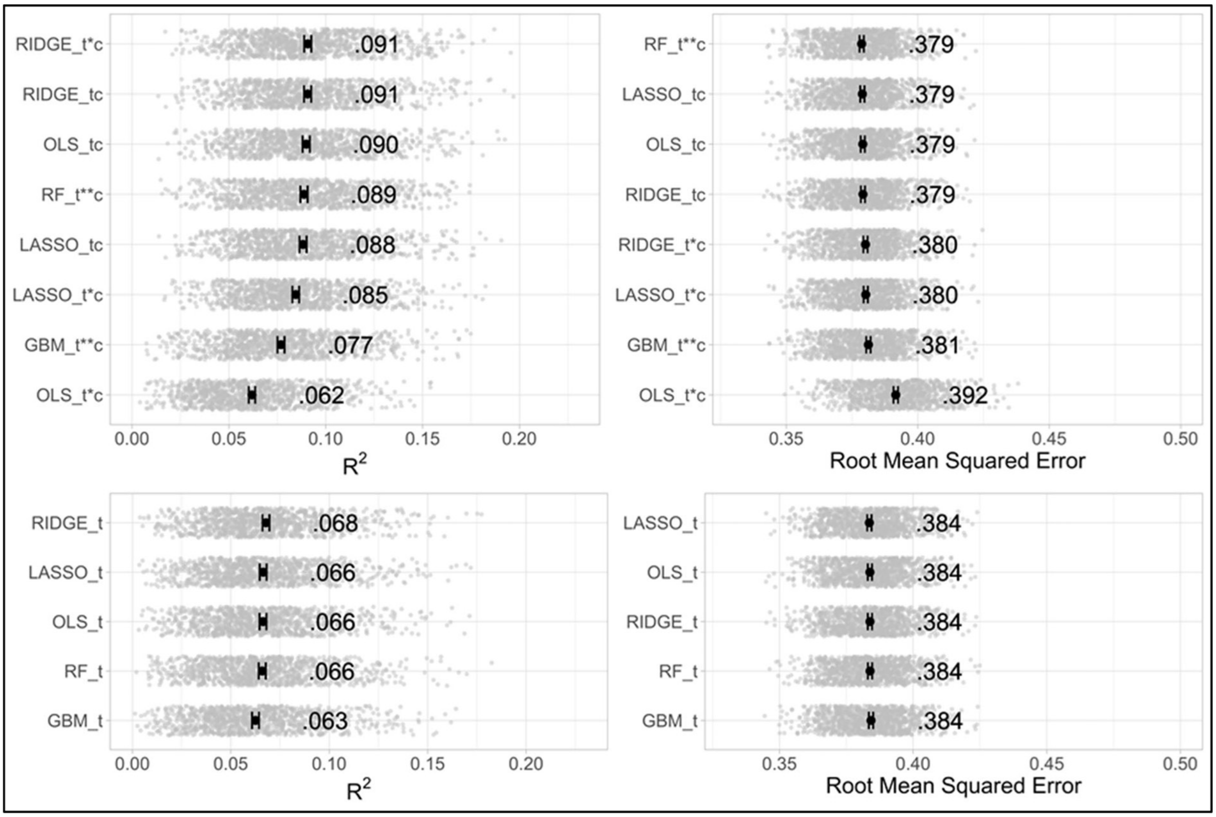

Fig. 2. Trait and context layer: Out-of-sample trait and context model performance (upper rows) for predicting the overall 360-degree effectiveness score compared to trait model performance in Fig. 1 (bottom rows). Dots represent the model performance in the 1000 test samples. The squares and error bars represent the average model performance and 95% confidence intervals. Models are sorted based on their average performance.

increase in their predictive performance when transitioning our analysis from HPI factors and context to the HPI items and context. Conversely, OLS suffered moving from the 13 HPI factors and context predictors (e.g., OLS_tc; RMSE = 0.379, 95% CI [0.378; 0.380]; R^2 = 0.090, 95% CI [0.088; 0.092]) to the 47 HPI items and context predictors (e.g., OLS_i_tc; RMSE = 0.384, 95% CI [0.384; 0.385]; R^2 = 0.085, 95% CI [0.083; 0.087]). Thus, as expected, once dimensionality was increased, the novel ML engine started to outperform OLS.

Interestingly, making the range of predictors much larger by adding interactions to the models worsened overall predictive performance. In general, almost none of our highly-dimensional data models performed better. For instance, when using data with the highest dimensionality (i.e., all 41*40 HPI item interactions plus 41 direct effects totaling 1681 predictors), the OLS engine entirely exploded across all 1000 iterations (i.e., RMSE = 84.403, 95% CI [58.550, 110.256]; R^2 = 0.006, 95% CI [0.005, 0.006]). Likewise, RIDGE and LASSO struggled under these high-dimension conditions (see Fig. 3). However, the exception to this decreased predictive performance was RF which rose to the top when it was set loose on all potentially complex HPI item and context effects. Though, it is important to keep in mind that RF is inherently one of the worst performing alternatives when it comes to interpretability.

In light of the differences in predictive performance between HPI factors and HPI items, our results underline that there are indeed situations in which more complex methods outperform conventional OLS. This consequently supports the idea that the data needs to match the method (i.e., the fuel needs to match the engine).

Discussion

The main purpose of our study was to investigate the leader trait paradigm and the added value of newer ML methods for predicting leader effectiveness versus conventional regression methods commonly utilized in leadership research (i.e., OLS). In particular, we examined the suggestion that interactions and nonlinear relationships among traits and context affect leadership outcomes. Through a variety of ML methods, we provided evidence that complex relationships are not adding much in terms of predictive performance in the current dataset.

However, though complexity in terms of interactions and nonlinear relationships did not add much, introducing complexity through increased dimensionality and then pairing it with the appropriate ML engine did. OLS was among the best-performing methods for predicting leader performance when the data had low-dimensionality, then the regularized regression methods RIDGE and LASSO emerged as data dimensionality increased, and finally the algorithmic RF alternative was best when the data was of the highest dimensionality. Key to this added ML value was having a variety of methods, models, and data to explore both interpretability and predictive performance.

Specifically, across our various fuel and engine setups, we observed that low-dimensional data paired with OLS provided us with the most straightforward and interpretable results and that high-dimensional data paired with the newer ML methods allowed us to extract more predictive value. Subsequently, by combining these various (simple to complex) methods and models, we found that (a) personality traits predicted leader effectiveness, (b) adding context improved our ability to predict effectiveness, (c) ML methods supported a direct effects approach to the leader trait paradigm (rather than an interactionist perspective), and (d) using various analytical methods (with the appropriate data) balanced interpretability with predictive performance.

These results have important theoretical implications regarding the mechanisms driving leader emergence versus leader effectiveness. Existing research supports a contingency model of leader *emergence* in which personal characteristics do have an interaction effect with context on leadership outcomes (e.g., younger leaders are voted for more

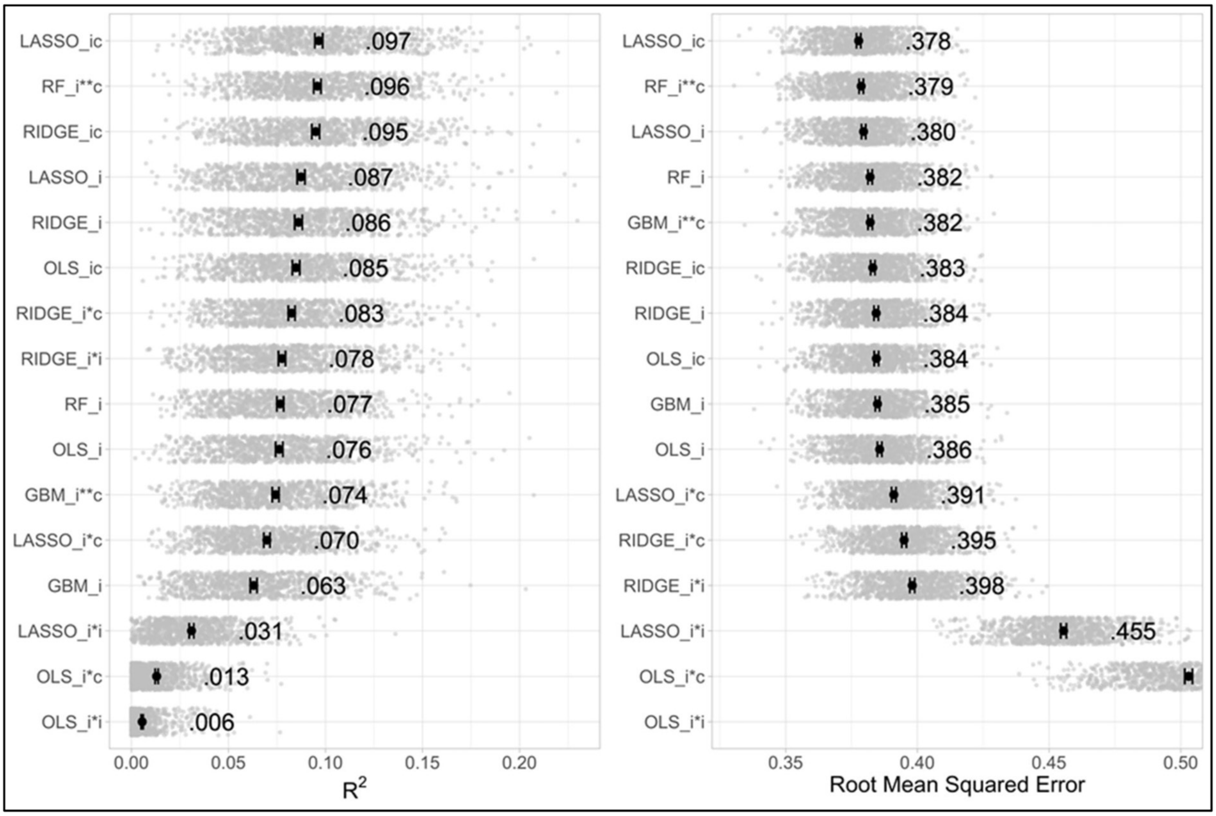

Fig. 3. HPI items and context layer: Out-of-sample HPI items and context model performance for predicting the overall 360-degree effectiveness score. Dots represent the model performance in the 1000 test samples. The squares and error bars represent the average model performance and 95% confidence intervals. Models are sorted based on their average performance.

often during times of change and older leaders for maintaining stability; Sharpanskykh & Spisak, 2011; Spisak, Grabo, Arvey, & van Vugt, 2014). However, our leader *effectiveness* findings tell a slightly different story. Specifically, models that included interactions did not improve our ability to predict effectiveness, but adding the direct effects of context did.

This predictive boost from context implies that leaders should choose their situations wisely when it comes to leading effectively. Many leaders, no matter how charismatic, transformative, extraverted, or generally adept at emergence, will struggle with effectiveness in difficult situations. Recall from our results that organizational size significantly influenced effectiveness ratings. This contextual emphasis derived from our intensive analysis consequently supports existing work on the “illusion of leadership” (Weber, Camerer, Rottenstreich, & Knez, 2001) and the “leader attribution error” (Hackman & Wageman, 2007) where the situation is a significant driver of effectiveness, but followers tend to mistakenly assign the majority of success or failure to the leader.

Such theoretical insights have significant applied implications. Shareholders, for example, may unnecessarily pay excessive amounts of money to CEOs thinking these leaders are the primary reason for increasing share value when deeper market forces are at play (Kolev, 2008). Likewise, in the political arena, a candidate many indeed “look the part” and emerge as the victor (Antonakis & Dalgas, 2009; Todorov, Mandisodza, Goren, & Hall, 2005), but if society does not have effective norms and prosocial policy in place, then leaders will remain largely ineffective at creating or maintaining positive outcomes. “Green leaders”, for instance, may contingently emerge to promote environmental sustainability (Spisak et al., 2014), but if the followers are not willing to adopt sustainable alternatives due to economic or ideological reasons, then green transitions will remain relatively slow (until the context demands change). Thus, our data supports the idea that deferring to a leader (as a person) for positive outcomes – rather than considering leadership as a process consisting of leaders, followers, and the situation – blurs an already murky understanding of effectiveness. Future ML research will need to replicate (and expand on) the current findings to better understand how *leader emergence* differs from *leadership effectiveness*.

In order to apply more complex ML approaches to such research, where scholars will continue to shift their focus away from direct main effects of separate traits towards the exploration of nonlinear and interaction effects between traits and context (Jensen & Patel, 2011; King et al., 2005), we will need to address fundamental limitations regarding the nature of our datasets (present offering included). Relative to what complex ML algorithms require, our fuel is frequently suboptimal (e.g., samples are too small, the number of variables are too few, scale reliability is too low, and/or excessive missing data). Though we worked to overcome these limitations through various computationally intensive techniques, future research will need to address this by gathering more applicable data if ML is to play an increasing role in leadership discovery.

Scholars can collect this leadership data through, for example, text mining, video recordings, and network analysis. Each of these techniques would result in highly dimensional data which could not be easily analyzed with conventional (OLS) methods without aggregating the data to a level where much of the underlying complexity is lost. Instead, the modern ML methods we propose here can extract insights from the data in its rawest form. Researchers can use text mining of conversations between leaders and followers to examine sentiment during interactions. We can also monitor the location of leaders and followers to see whether continuous co-location, frequent informal meetings, or any other geospatial patterns relate to effectiveness. In short, combining highly dimensional leadership data with ML unlocks vast opportunity for discovery. We suspect these advancements will fundamentally change how leadership research is conducted.

Perhaps a first step in this new direction is the curation of a special issue where multiple teams take different ML approaches to investigating the same dataset(s). This "different papers, same dataset" concept was used to set the benchmark for comparing multilevel methods in leadership (i.e., Bliese, Halverson, & Schriesheim, 2002), and it can potentially do the same for raising and addressing questions pertaining to ML. Scholars and practitioners, for example, will need to ask themselves if they have the right kind of data and if explanation or prediction is the primary concern. They will also need to consider additional variables such as transformational leadership, leader-member exchange, and IQ to further refine the fuel. Then, to fully integrate ML into the broader leadership community, they will also need to consider using methods such as structural equation modeling to confirm the causal order behind these predictive insights. Indeed, we can ramp up data volumes and add dimensionality to increase predictive performance, but how does that affect the relevance for leadership scholars and practitioners if we sacrifice too much interpretability?

Fortunately, for those interested in addressing these questions and integrating ML into their own leadership research, a repository of literature exists to guide them on their way. Among the number of excellent works we have cited, we particularly recommend Friedman et al. (2001), James et al. (2013), as well as Kuhn and Johnson (2013) for a broad introduction. Likewise, there is a small, but growing number of publications focused particularly on ML in Management and Psychological Sciences. We have already cited the contributions of Joel et al. (2017) as well as Yarkoni and Westfall (2017). In addition, readers may find the work of Putka, Beatty, and Reeder (2018) helpful for gaining a deeper understanding of prediction methods.

Finally, we encourage all scholars to acquire some level of appreciation for this extremely important analytical future. With the right fuel, ML will drive new theory by uncovering hidden complexities (or confirming established simplicity), it will elucidate blind spots between theory and reality, and even lead to new measures for explanatory modeling with smaller samples. However, as research inevitably incorporates ML advancement, we must also become increasingly aware of the pitfalls of this burgeoning approach. Inadvertently using the wrong fuel in the ML engine has the potential to damage both the engine and the operator.

Declaration of Competing Interest

None

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2019.05.005.

References

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play!. *Science, 323*(5918), 1183.

Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods in leadership: The articles, the model, and the data set. *The Leadership Quarterly, 13*(1), 3–14.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Boca Raton, FL: Chapman & Hall/CRC.

Day, D. V., & Antonakis, J. (2012). Leadership: Past, present, and future. In D. V. Day, & J. Antonakis (Eds.). *The nature of leadership* (pp. 3–25). Thousand Oaks, CA: Sage.

De Vries, R. E. (2012). Personality predictors of leadership styles and the self–other agreement problem. *The Leadership Quarterly, 23*(5), 809–821.

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review, 20*(1), 65–91.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407–499.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence, 14*(5), 771–780.

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 23*(5), 1189–1232.

Gilboa, S., Shirom, A., Fried, Y., & Cooper, C. (2008). A meta-analysis of work demand stressors and job performance: Examining main and moderating effects. *Personnel Psychology, 61*(2), 227–271.

Hackman, J., & Wageman, R. (2007). Asking the right questions about leadership. *American Psychologist, 62*(1), 43–47.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100–112.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York, NY: Springer.

Jensen, J. M., & Patel, P. C. (2011). Predicting counterproductive work behavior from the interaction of personality traits. *Personality and Individual Differences, 51*(4), 466–471.

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science, 28*(10), 1478–1489.

Judge, T. A., Piccolo, R. F., & Kosalka, T. (2009). The bright and dark sides of leader traits: A review and theoretical extension of the leader trait paradigm. *The Leadership Quarterly, 20*(6), 855–875.

King, E. B., George, J. M., & Hebl, M. R. (2005). Linking personality to helping behaviors at work: An interactional perspective. *Journal of Personality, 73*(3), 585–608.

Kolev, G. I. (2008). The stock market bubble, shareholders' attribution bias and excessive top CEO pay. *The Journal of Behavioral Finance, 9*(2), 62–71.

Kuhn, M. (2017). Package "caret": Classification and regression training. Comprehensive R archive network. Retrieved from https://github.com/topepo/caret.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York, NY: Springer.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*(4), 1029–1049.

Oh, I. S., & Berry, C. M. (2009). The five–factor model of personality and managerial performance: Validity gains through the use of 360-degree performance ratings. *Journal of Applied Psychology, 94*, 1498–1513.

Peter Berry Consultancy (2016). Bench strength of the leadership pipeline: Exploring 360° competencies that emerge at different leader levels [white paper]. Retrieved August 7, 2018, from https://peterberry.com.au/wp-content/uploads/2017/03/PBC-White-Paper_Benchstrength-of-the-Leadership-Pipeline-FINAL.pdf

Peter Berry Consultancy & Hogan Assessment Systems (2019). *Hogan 360° technical manual* (4th ed.). .

Phaneuf, J.-É., Boudrias, J.-S., Rousseau, V., & Brunelle, É. (2016). Personality and transformational leadership: The moderating effect of organizational context. *Personality and Individual Differences, 102*, 30–35.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods, 21*(3), 689–732.

Resick, C. J., Whitman, D. S., Weingarden, S. M., & Hiller, N. J. (2009). The bright-side and the dark-side of CEO personality: Examining core self-evaluations, narcissism, transformational leadership, and strategic influence. *Journal of Applied Psychology, 94*(6), 1365–1381.

Richard, P. J., Devinney, T. M., Yip, G. S., & Johnson, G. (2009). Measuring organizational performance: Towards methodological best practice. *Journal of Management, 35*(3), 718–804.

Ridgeway, G. (2017). Package "gbm": Generalized boosted regression models. R package version 2.1.3. Retrieved from https://github.com/harrysouthworth/gbm.

Sharpanskykh, A., & Spisak, B. R. (2011). An agent-based evolutionary model of leadership. In J. Zhan, (Ed.). *Proceeding of the 2011 IEEE international conference on privacy, security, risk, and trust, and IEEE international conference on social computing* (pp. 848–855). Boston: IEEE Computer Society Press.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.

Spisak, B. R., Grabo, A. E., Arvey, R. D., & van Vugt, M. (2014). The age of exploration and exploitation: Younger-looking leaders endorsed for change and older-looking leaders endorsed for stability. *The Leadership Quarterly, 25*(5), 805–816.

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*(4), 323–348.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500–517.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B Methodological, 58*(1), 267–288.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626.

Tuncdogan, A., Acar, O. A., & Stam, D. (2017). Individual differences as antecedents of leader behavior: Towards an understanding of multi-level outcomes. *The Leadership Quarterly, 28*(1), 40–64.

Vaccaro, I. G., Jansen, J. J., Van Den Bosch, F. A., & Volberda, H. W. (2012). Management innovation and leadership: The moderating role of organizational size. *Journal of Management Studies, 49*(1), 28–51.

Van der Wal, Z., De Graaf, G., & Lasthuizen, K. (2008). What's valued most? Similarities and differences between the organizational values of the public and private sector. *Public Administration, 86*(2), 465–482.

Weber, R., Camerer, C., Rottenstreich, Y., & Knez, M. (2001). The illusion of leadership: Misattribution of cause in coordination games. *Organization Science, 12*(5), 582–598.

Wright, M. N., Wager, S., & Probst, P. (2018). Package “ranger”: A fast implantation of random forests. R package version 0.9.0. Retrieved from https://github.com/imbs-hl/ranger.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.

Zhou, D. X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters, 83*(9), 2108–2112.