The Leadership Quarterly 33 (2022) 101515

Contents lists available at ScienceDirect

The Leadership Quarterly

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

Full Length Article

Opening the black box: Uncovering the leader trait paradigm through machine learning

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

[a] Vrije Universiteit Amsterdam, The Netherlands
[b] Harvard University, USA
[c] University of Otago, New Zealand
[d] Independent researcher

ARTICLE INFO

Article history:
Received 9 November 2019
Received in revised form 14 February 2021
Accepted 24 February 2021

Available online 16 March 2021

Keywords:
Leader trait paradigm
Machine learning
Complexity
Interpretability
Personality

ABSTRACT

Understanding the traits that define a leader is a perennial quest. An ongoing debate surrounds the complexity required to unravel the leader trait paradigm. With the advancement of machine learning, scholars are now better equipped to model leadership as an outcome of complex patterns in traits. However, interpreting those models is often harder. In this paper, we guide researchers in the application of machine learning techniques to uncover complex relationships. Specifically, we demonstrate how applying machine learning can help to assess the complexity of a relationship and show techniques that help interpret the outcomes of "black box" machine learning algorithms. While demonstrating techniques to uncover complex relationships, we are using the Big Five Inventory and need for cognition to predict leadership role occupancy. Among our sample ($n = 3385$), we find that the leader trait paradigm can benefit from modeling complexity beyond linear effects and generate several interpretable results.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

The leader trait paradigm is a perennial topic in the study of leadership, From Sun Tzu and Plato to contemporary academics and practitioners, the exploration of leader-defining traits is perpetual. This exploration is understandable given that finding solutions to the leader trait paradigm generates powerful outcomes for predicting who is likely to occupy leadership positions, how they are likely to perform, and what impact they will have on society. The ubiquity of psychometric assessment in the leadership selection process is a testament to the importance of the leader trait paradigm.

In spite of its great importance, the continual search for the traits that define a leader – across cultures and time – suggests that the leader trait paradigm is elusive. We have an increasingly large collection of leader trait observations with a disproportionate lack of shared understanding. There is confusion surrounding the level of complexity required to model the phenomenon (e.g., linear, additive effects versus an interactionist perspective) and limited capacity for handling the reservoir of data (e.g., the tradition of using linear models; Spisak, van der Laken, & Doornenbal, 2019). Further, as complexity problems are solved with analytical advancements, issues about "black box" interpretability emerge. In short, despite the impressive amount of philosophical thinking and empirical investigation, the search for the traits that define a leader remains. It is thus not surprising that Burns referred to leadership as "one of the most observed and least understood phenomena on earth" (Burns, 1978, p. 2).

In this paper, we propose that the relationship between traits and leadership can be advanced through the use of machine learning techniques – that is, a computational process for automatically "learning" patterns in data and improving performance on tasks such as prediction (Kuhn & Johnson, 2013). Machine learning is already adding value in marketing (Wedel & Kannan, 2016), human resource management (Garcia-Arroyo & Osca, in press; Strohmeier & Piazza, 2013), and logistics (Govindan, Cheng, Mishra, & Shukla, 2018; Wang, Gunasekaran, Ngai, & Papadopoulos, 2016). In general, wherever there is data, people are applying machine learning to explore patterns and make predictions. This explains why machine learning "now permeates our existence" (Kuhn & Johnson, 2013, p. 2), and is poised to deliver a new level of leader trait clarity.

Accordingly, we apply machine learning techniques to uncover the traits that predict leadership role occupancy – that is, whether an individual occupies a leadership role. Our aim is to guide researchers in applying machine learning techniques to uncover complex relationships. We start with a brief overview of what we know regarding the leader trait paradigm. Next, we highlight important issues obstructing the leader trait paradigm, followed by how these obstructions lead to different dilemmas such as choosing between simplicity versus complexity. We then describe our empirical approach for solving the dilemmas by

* Corresponding author at: Department of Management and Organisation, Vrije Universiteit Amsterdam, De Boelelaan, Amsterdam 1105 1081 HV, The Netherlands.
E-mail address: b.m.doornenbal@vu.nl (B.M. Doornenbal).

https://doi.org/10.1016/j.leaqua.2021.101515
1048-9843/© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

applying machine learning techniques, noting that non-parametric techniques often suffer from reduced interpretability. Following that, we use a large database (n = 3385) of trait variables and leadership role occupancy to compare the predictive performance of traditional (parametric) linear models (LM) versus the non-parametric technique of random forest (RF) analysis. Finally, we demonstrate how to incorporate recent analytical advancements to open up the black box of RF for model interpretability. We believe that a better understanding of the use of non-parametric machine learning, while maintaining model interpretability, can help to advance what we know about leadership and how we come to know it.

Adding non-parametric machine learning, which we denote in this paper as algorithmic machine learning, can advance the development of theory (Kolkman & van Witteloostuijn, 2019). Conventionally, scholars *inductively* build theory based on explanations found in specific observations and – subsequently – *deductively* test the theory by examining the consistency of the found explanations in a larger set of observations (Eisenhardt & Graebner, 2007; Mantere & Ketokivi, 2013). The machine learning techniques demonstrated in this paper fit within the tradition of *abduction*, which is a process of deriving hypotheses by appraising observations in light of the theory (Mantere & Ketokivi, 2013). We demonstrate how to apply algorithmic machine learning techniques that can quantitatively uncover (complex) relationships that are key in generating high predictive performance. Such results add value in subsequent inductive theory-building and can be translated into hypotheses for deductive tests of the theory (Kolkman & van Witteloostuijn, 2019). For example, machine learning might provide (inductive) insights that contribute to interactionist theories of the leader trait paradigm and provide a map for (deductively) testing unexpected interactions. Machine learning may also establish (in an empirically robust way) that adding complexity does not necessarily help to predict leadership based on traits. Either way, the main point is that machine learning can help shed additional light on theoretical uncertainties.

Dilemmas obstructing the leader trait paradigm

Although the amount of observations about the leader trait paradigm outweighs the unified understanding of the phenomenon, scholars and practitioners typically appreciate that traits matter (De Vries, 2012 ; Judge, Bono, Ilies, & Gerhardt, 2002 ; Zaccaro, 2007). Traits are "individual characteristics that (a) are measurable (b) vary across individuals, (c) exhibit temporal and situational stability, and (d) predict attitudes, decisions, or behaviors and consequently outcomes" (Antonakis, 2011, p. 270). Traits also shed light on deep, temporal processes such as leadership effectiveness, as well as shallow, thinner-sliced moments of leader emergence (Judge et al., 2002; Judge, Piccolo, & Kosalka, 2009) and leadership role occupancy (De Neve, Mikhaylov, Dawes, Christakis, & Fowler, 2013;). Traits also form a constellation of empirically relevant leadership variables, including Big Five traits (Judge & Bono, 2000), lower-level personality traits, such as need for cognition (Judge et al., 2009), and stereotypically gendered traits (Eagly & Johnson, 1990). The leader trait paradigm also adds value to a diverse range of topics such as evolutionary theory (Van Vugt, Hogan, & Kaiser, 2008) and leadership ethics (Babalola, Bligh, Ogunfowora, Guo, & Garba, 2019). The diversity of these individual differences, the role they play in various aspects of leadership, and the associated perspectives from which they are observed, are thus clear signs of the leader trait paradigm's relevance./ce:para>However, despite the broad and longstanding appreciation for traits, scholars and practitioners lack a valid and shared network of understanding about their impact. One current debate is about the level of complexity needed to study the traits that best define a leader. Scholars debate whether individual traits and linear additive effects suffice for understanding leadership. Some argue that a perspective on non-linearities (Vergauwe, Wille, Hofmans, Kaiser, & De Fruyt, 2018) and interactions (Jensen & Patel, 2011) of traits is necessary, whereas others argue that the context moderates the impact of traits (Phaneuf, Boudrias, Rousseau, & Brunelle, 2016; Tett & Burnett, 2003). Relatedly, the leader trait conversation is now dividing between focusing on interpretable results through simplicity versus increasing predictive performance at the cost of interpretability (e.g., Spisak et al., 2019).

This debate concerning leader trait simplicity and interpretability versus complexity and predictability is a dilemma inhibiting the transition from observations about the leader trait paradigm to a unified understanding of the phenomenon (i.e., the most observed, least understood criticism of Burns). If a model of the leader trait paradigm is too simplistic, then errors associated with "bias" occur. Here, the model *underfits*. That is, it is unable to capture the essence of a relationship because it ignores important predictors and/or (non-linear) effects, which give rise to the interactionist argument mentioned above. Conversely, if a model of the leader trait paradigm is too complex, then problems associated with "variance" occur. Here the model *overfits*. That is, it captures noise along with real patterns and the model does not generalize to unseen data. Hence, we are left with a picture of the leader trait paradigm that either overlooks important features or overvalues noise. This problem is commonly referred to as the bias-variance dilemma (Belkin, Hsu, Ma, & Mandal, 2019; Geman, Bienenstock, & Doursat, 1992).

To manage the bias-variance dilemma, and find a better balance between simple and complex representations, scholars can use a resampling technique known as cross-validation. Here, machine learning models are run through multiple training and validation iterations. During these iterations, models are fitted on training data and subsequently tested on (unseen) test data. The more the prediction error on the test set outweighs the prediction error on the training set, the greater the *overfit*. If this is the case, then a more simplistic model of the leader trait paradigm is perhaps a better alternative.

In the process of managing the bias-variance dilemma, scholars can get a sense of the *underfit* by comparing different modeling techniques. More specifically, by comparing the predictive performances of simple and complex models, one can get a sense of the extent to which complex effects are present in the data. For instance, one could compare the performance of a LM with only linear additive effects between traits and leadership with that of a RF, which is free to explore complex interactions and non-linear relationships beyond linear additive effects. The more the complex model (e.g., RF) outperforms the more simplistic model (e.g., LM) in terms of predictive performance, the greater the *underfit* of the simplistic model. If this is the case, then a complex model of the leader trait paradigm is perhaps a better alternative.

As we deal with the bias-variance dilemma, another issue arises where tradeoff problems between explanation and prediction emerge (Shmueli, 2010). If we choose a model that is explanation-oriented, then the predictive performance might be poor. For example, one may identify a significant pathway between extraversion and leadership role occupancy, but by focusing solely on linear additive effects or simple two-way interactions (as in much of the literature) we might miss important other aspects that explain whether a person occupies a leadership role. This is similar to bias where important features of reality (e.g., complex interactions and non-linear relationships) are overlooked and our ability to understand the (potentially) complex reality of the leader trait paradigm is reduced.

In contrast, if we select a model best at predicting leadership role occupancy, then we might have difficulties interpreting the predictions. In this scenario, the model accurately represents a complex reality of the relationship between traits and leadership role occupancy, but the underlying pathways and relationships are hard to interpret. Algorithmic models, utilizing cross-validation and the freedom to explore complex interactions and non-linear relationships, work to minimize the error associated with the bias-variance dilemma. This, however, may reduce our understanding of the *causal mechanisms* explaining leader trait

paradigm (i.e., the "gold standard" of leadership research; Antonakis, Bendahan, Jacquart, & Lalive, 2010).

Collectively, simplicity versus complexity and explanation versus prediction represent two significant and interconnected barriers to knowledge generation. The transition from discrete leader trait observations to a unified understanding of the leader trait paradigm is first blocked by difficulties in finding the right balance between simple and complex models. Second, though this longstanding problem can now be minimized with machine learning techniques, an obstruction emerges when the optimal balance between simplicity and complexity lands in a space exceedingly low on interpretably – thus hindering a deep understanding of the causal mechanisms between input and output. In short, we have a complexity dilemma and solving this can lead to an interpretability dilemma.

To show how algorithmic machine learning can help to resolve these dilemmas, and address the "most observed, least understood" criticism of Burns (1978), we leverage in this paper a large dataset of leader trait variables. We use this dataset as the "fuel" for both a simple and complex machine learning "engine." Specifically, as described above, we compare the predictive performance of a LM that solely tests linear additive effects with the predictive performance of a RF that allows more flexibility. Testing how well these engines perform results in information on the complexity of the leader trait paradigm. For instance, if the RF outperforms the LM in terms of predictive performance, this suggests that the true leader trait relationship is likely more complex than linear additive effects.[1] Thus, the first research question we focus on is:

Research question 1: To what degree do non-linear effects and interactions explain the trait-leadership role occupancy relationship?

Next, to address the interpretability dilemma, we open up the RF black box using various analytical procedures. By demonstrating these procedures, we guide scholars on how to get a sense of the complexity modeled by machine learning models. This step has the potential to increase the understanding of the leader trait paradigm, inspire theory-building, and provide input for subsequent hypothesis-testing. Here, we try to answer the following three research questions:

Research question 2a: How important is each trait in predicting leadership role occupancy within the RF?

Research question 2b: What is the shape of the relationship between traits and leadership role occupancy within the RF?

Research question 2c: What interactions are leveraged by the RF for predicting leadership role occupancy?

Finally, it is important to note that our machine learning approach is data-driven, not hypothesis-driven. Instead of developing hypotheses, which is common from an explanation perspective, we allow the machine learning models to explore patterns, which is common from a prediction perspective (Shmueli, 2010). This predictive style of research is central to our undertaking as it provides the necessary freedom for discovery.

Methods

Participants

In this paper we make use of data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands). The LISS panel is a representative sample of Dutch individuals (Scherpenzeel & Das, 2010) and is based on a true probability sample of households drawn from the population register. The data we used ($n = 3642$) were gathered in 2013. In our sample, 58.4% of the participants were women and the minimal age was 18 ($M = 53.73$, $SD = 16.07$).

[1] Note that the difference in model performance might change when non-linear effects are modeled through the LM as well.

Measures

In our analysis, we used measures for Big Five Inventory (BFI) personality traits, need for cognition (NFC), and leadership role occupancy (see Table 1). The BFI personality traits were measured with the 50-item Five Factor Model International Personality Item Pool (IPIP; Goldberg, 1999) and NFC with an 18-item NFC scale (Cacioppo & Petty, 1984). The items were translated into Dutch by professional translators and the survey was conducted in Dutch. The internal consistency of the measures ranged from 0.77 to 0.90. The correlations among the BFI traits as well as the correlations between NFC and openness, conscientiousness, and neuroticism are in line with correlations found in previous research (Furnham & Thorne, 2013; Van der Linden, te Nijenhuis, & Bakker, 2010). In our study, NFC shows a positive correlation with extraversion and agreeableness where previous studies have respectively reported inconsistent and often non-significant correlations (Furnham & Thorne, 2013). The correlation between NFC and leadership role occupancy, which, to the best of our knowledge, has not been directly tested before, was positive. The correlations between leadership role occupancy and openness, conscientiousness, and extraversion were positive, whereas the correlations between leadership role occupancy and agreeableness and neuroticism were negative. Compared with meta-analyses that tested the relationship between BFI and emergence – the latter measured in part as leadership role occupancy (Barling & Weatherhead, 2016) – the correlations are mostly in the same direction except for agreeableness (e.g., Ensari, Riggio, Christian, & Carslaw, 2011; Ilies, Gerhardt, & Le, 2004). In the meta-analysis of Ilies and colleagues, for example, agreeableness had a small overall positive correlation ($r = 0.05$), whereas we find a small negative correlation ($r = -0.05$).

As a measure of leadership role occupancy, we utilized the question: "What is your current profession/What profession did you exercise in your last job? (If you are retired, unemployed or currently do not have a job: please refer to the last job you had)." We classified individuals as occupying a leadership role (10.84%) when they answered the question with "higher supervisory profession (e.g. manager, director, owner of large company, supervisory civil servant)". We classified individuals as not occupying a leadership role when they answered the question with one of the following answers: "higher academic or independent profession (e.g. architect, physician, scholar, academic instructor, engineer)" (7.71%), "intermediate academic or independent profession (e.g. teacher, artist, nurse, social worker, policy assistant)" (29.2%), "other mental work (e.g. administrative assistant, accountant, sales assistant, family carer)" (31.8%), "semi-skilled manual work (e.g. driver, factory worker)" (8.68%), "unskilled and trained manual work (e.g. cleaner, packer)" (9.63%), and "agrarian profession (e.g. farm worker, independent agriculturalist)" (2.04%).

Analytical procedure

We first compare the predictive performance of a LM with the predictive performance of a RF. Both models use the BFI and NFC scales as traits to predict leadership role occupancy. We form the LM by conducting a logistic regression analysis. In this analysis, we fit straight lines through the data that describe the relationship between traits and leadership role occupancy. The model chooses the lines (i.e., coefficients, parameters) such that it minimizes the sum of squared residuals and assumes a binomial distribution (individuals are either in a leadership role or are not in a leadership role). While fitting the model, we only apply linear additive effects. The predictions of the LM are log odds (i.e., logits).

The RF is formed by iteratively splitting the data into subsamples based on demarcations (i.e., cut-points that progressively minimize error in prediction). For each demarcation, the RF picks a personality trait value that best divides the sample in terms of leadership role occupancy, hence minimizes the sum of squared residuals (Breiman, Friedman, Stone, & Olshen, 1984). The partitioning of the RF results in

Table 1
Means, standard deviations, Cronbach's alphas, and correlations among the variables.

Variables	M	SD	1	2	3	4	5	6	7	8	9
1. Gender	0.58	–	–								
2. Age	53.73	16.07	−0.03	–							
3. Leadership role occupancy	0.11	–	−0.25***	0.11***	–						
4. Openness	3.46	0.50	−0.13***	−0.13***	0.14***	(0.77)					
5. Conscientiousness	3.75	0.51	0.10***	0.12***	0.07***	0.24***	(0.81)				
6. Extraversion	3.22	0.65	−0.01	−0.02	0.10***	0.34***	0.11***	(0.77)			
7. Agreeableness	3.88	0.51	0.30***	0.08***	−0.05**	0.25***	0.32***	0.33***	(0.87)		
8. Neuroticism	2.49	0.70	0.15***	−0.14***	−0.11***	−0.16***	−0.20***	−0.24***	−0.06***	(0.88)	
9. NFC	4.28	0.95	−0.19***	−0.08***	0.20***	0.61***	0.19***	0.23***	0.11***	−0.25***	(0.90)

Note. N = 3385. NFC = Need for cognition. We report the internal consistencies (Cronbach's alphas) on the diagonal.
Gender dummy coded, 0 = male, 1 = female.
** $p < .01$.
*** $p < .001$.

a tree-like structure that models the underlying relationships in a dataset. Rather than using one tree to predict, multiple trees are formed that are all used for prediction based on a majority vote.

Following procedures described by Kuhn and Johnson (2013), we built the LM and RF based on normalized variables (normalized after splitting the data). The final RF is grown with 500 trees, 2 candidate variables for each split, and a minimal of 400 observations in each terminal node. These hyperparameters are the optimal settings based on a grid search approach.[2] In the next section, we report the results of our analysis in order to answer the four research questions. Each time before providing the results, we briefly describe the procedure we applied specific to the output.

It should also be noted that both the LM and RF models are incorporating a machine learning approach – more on that below. Further, it is important to stress that this is *not* a competition between LM and RF. One can always increase the number of variables (i.e., dimensionality) until LM is no longer able to compete (e.g., Spisak et al., 2019). Rather, the current goal is providing a "training" example of how to better calibrate the tradeoffs between simplicity versus complexity and explanation versus prediction in the exploration of leadership. Scholar and practitioners have a variety new tools to further this search. Here we introduce some of these advancements.

Results

Research question 1: To what degree do non-linear effects and interactions explain the trait-leadership role occupancy relationship?

Algorithmic machine learning models, such as the RF, can help to explore the extent to which the true relationship between personality traits and leadership role occupancy comprises non-linear and interaction effects. More specifically, if a RF outperforms a LM with only linear additive effects in terms of predictive performance, there may be interactions or non-linear effects present in the data that help the RF outperform the LM. Again, the RF is not limited to fitting straight lines through the data, but can use any potential binary splits of the data, also on multiple variables, to predict.

Utilizing machine learning techniques also encourages a move away from fitting models on the full sample. Although more common in the broader literature, for example in marketing modeling (Cooil, Winer, & Rados, 1987) and decision-making (Puelz & Sobol, 1995), leadership scholars often do not split their data into training, validation, and test sets when developing and evaluating models. This means that often the same data (i.e., all the sampled data) are used to develop the model and then to evaluate the predictive performance of that model. However, models developed in this way are prone to overfit, especially when modeling techniques are not limited to fitting straight lines through data. Inherently, models will try to minimize the error of their predictions in the data they are given – they are optimizing for the available data. As a result, the model and its evaluated predictive performance will likely be overestimates of the actual model's performance when given new data.

To avoid overfit, and to estimate actual relationships, scholars often conduct meta-analyses. In modeling complex relationships through flexible machine learning techniques, cross-validation is often used for model development. As mentioned, cross-validation implies that a model is iteratively trained and validated on samples of the same development data to optimize the model for (predictive) performance on new data. To compare the predictive performance of our RF and LM, we separate a test set for model evaluation. This test set is a (random) sample of the data that is not used for model development (neither training nor validation), but is used later to evaluate the performance of the models once they are developed and optimized.

Applying these two steps (i.e., cross-validation and a test set), we first removed a random 1/3 of our data as a test set. This test data contained 1127 observations of which 129 (11.4%) occupied a leadership role. We used the remaining 2/3 of the observations as model development data. This set of 2258 observations contained 238 (10.5%) individuals in a leadership position. We used these data for repeated cross-validation. In our cross-validation procedure, we randomly split these 2258 observations in ten approximately equal parts, and iteratively used nine as training data and the remaining one as validation data until all parts had functioned as validation data once. We repeated this procedure ten times to decrease the influence of chance. Finally, the performance of the RF and LM was evaluated on the test set.

Cross-validation set

In the model development data, the RF was better than the LM in predicting leadership role occupancy based on personality traits (i.e., BFI and NFC; see Fig. 1a). Fig. 1a shows how many individuals in a leadership role each model would correctly identify as occupying a leadership role (y-axis) if we were to check the top *x* observations with the highest predicted probabilities (x-axis). The steeper the model lines start out from the origin, the better the model is in predicting leadership role occupancy. Since the development data contained 238 individuals *actually* in a leadership role, we added a dashed line on the x-axis at 238 *predicted* leaders in Fig. 1a. At the dashed line, a perfect model would have identified all individuals with actual leadership roles. Hence, a perfect model in Fig. 1a would have a slope of 1 and levels flat at the intersect at 238 on both the x- and y-axis.

It is important to first note that both the LM and RF were better at identifying leadership role occupancy in our development data than a naïve, chance-based approach (i.e., their lines in Fig. 1a were steeper

[2] We explored 66 combinations of hyperparameters containing six values for the number of candidate variables for each split (i.e., 1, 2, 3, 4, 5, and 6) and 29 values for the minimal terminal node size (1, 5, 10, 25, 50, 100, 200, 300, 400, 500, and 600).

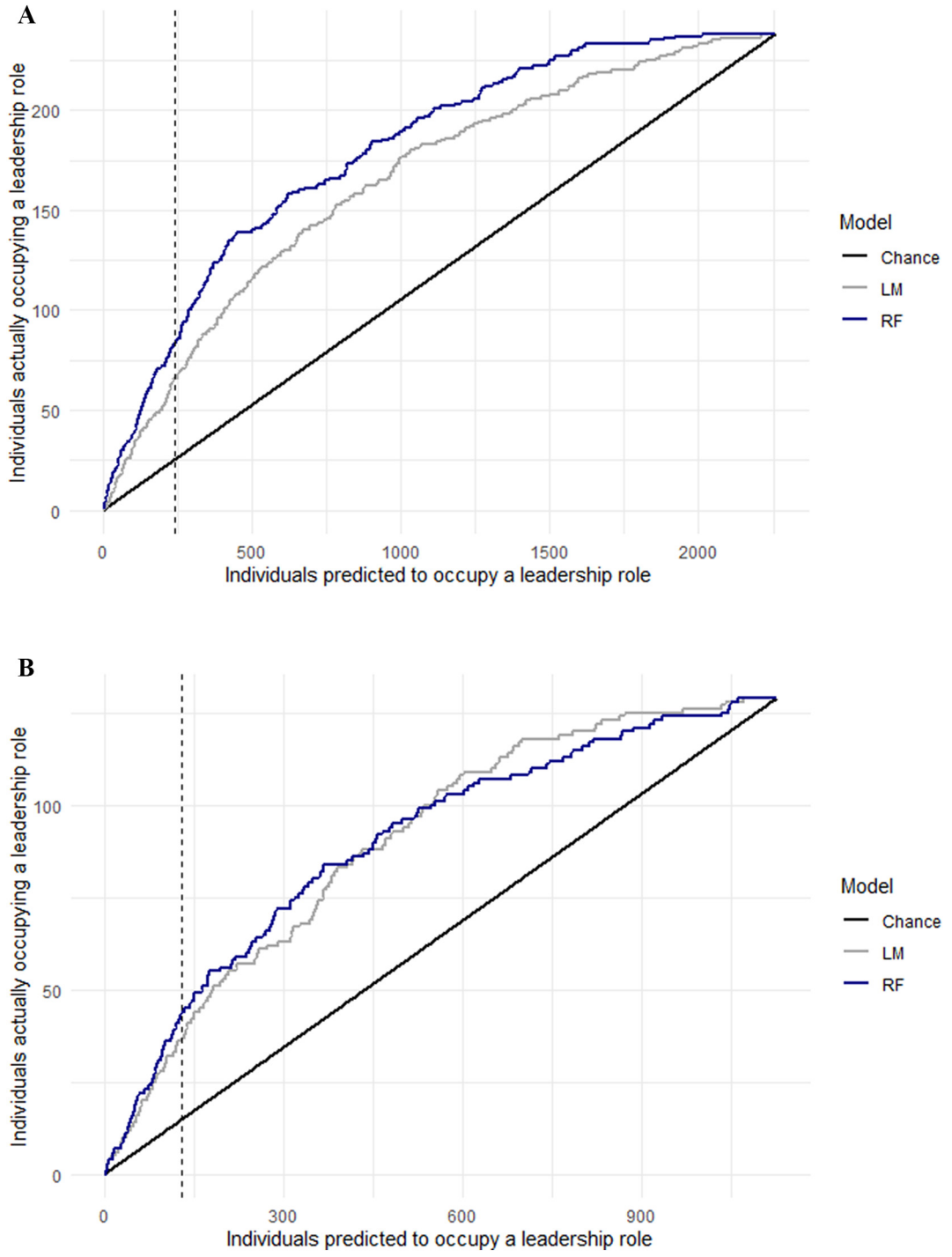

Fig. 1. (A) A cumulative gain chart, based on the cross-validation set, describing the number of individuals that are actually in a leadership role of the individuals that are likely to be in a leadership role identified by the model. *Note.* The dashed line intercepts the x-axis at 238, the number of individuals occupying a leadership role in the development data. (B) A cumulative gain chart, based on the test set, describing the number of individuals that is actually in a leadership role of the individuals that are likely to be in a leadership role identified by the model. *Note.* The dashed line intercepts the x-axis at 129, the number of individuals occupying a leadership role in the test set.

than chance). If we were to pick observations at random, we would find one leader approximately every ten guesses, just by chance. Thus, a naïve baseline expectation is finding about 25 (10.5%) individuals occupying a leadership role after 238 guesses. Both the LM and the RF models were better than this baseline, which is indicative that there is value in trait information for predicting leadership role occupancy. That said, we also found that RF outperformed LM. Of the top 238 observations predicted to be occupying a leadership role by the RF, 83 (34.9%) actually were in a leadership position. For the LM, 66 (27%) of the 238 individuals with leadership roles were identified.

These percentages represent the *sensitivity* of the models, the proportion of individuals actually occupying a leadership role that are correctly identified as such (also called the *true positive rate*, the *recall*, or the *probability of detection*). In this train-test development data, the RF

was thus more sensitive than the LM. The RF could better leverage the BFI and NFC personality trait data to predict leadership role occupancy. Collectively, the development data results suggest that (a) leader trait information is important for predicting leadership role occupancy and (b) more complexity than linear additive patterns exist in the trait-leadership role occupancy relationship.

Test set

If this development dataset was all we had, we would have concluded that the RF produced a considerably more predictive model than the LM. However, as we discussed before, it is desirable to evaluate a model's predictive performance on data other than those used to develop that model. The predictive performance in the development data will likely be an overestimate because models inherently overfit the data they are given – models may mistake random noise in the development data for relevant patterns. Techniques that have more freedom in modeling patterns – such as RF – are more prone to overfit. Flexible as they are, RF can fit complex interactions and non-linearities that are specific to the development dataset but may not occur in other samples of the same population. Hence, while RF better predicted leadership role occupancy in data used for model development, the leveraged BFI and NFC patterns might not generalize to new data. If the trait-leadership role occupancy patterns extracted from this development dataset are used to make predictions for unseen data, these predictions may be far off.

Accordingly, we assess the predictive performance of our models on new, unseen data - i.e., a test set. Note that our test data had not been used in developing the current models. Only now, we input this validation data into our cross-validated models and evaluate the output predictions. The results are visualized in the same way as before (see Fig. 1b). Since the test data contained 129 (11.4%) individuals with a leadership role, we are again using a dashed line to highlight how the models perform among the 129 individuals they predicted to be the most likely to have a leadership role. This time, the RF detected 44 (34.1%) of the individuals actually occupying leadership roles, whereas the LM detected 36 (27.9%) of them. Both outcomes are considerably higher than the naïve baseline (15 detected by chance, or 11.4%).

Finally, note that to the far right of the dashed line in Fig. 1b (i.e., the line indicating the predictive value of our models), the LM starts to overtake the RF. The LM crosses over when the vast majority of individuals in leadership roles (approximately 100) are identified, and both models predict that individuals are unlikely to be in a leadership role. This suggests that modeling more complex effects such as interactions and non-linearities is suitable when the goal is to *precisely* identify the majority of individuals occupying a leadership role (and avoid false positives) rather than broadly identifying (or *recalling*) all individuals with leadership roles (and avoid false negatives).

Collectively, the transition from the development data results to the test data results suggest that (a) trait information is important for predicting leadership role occupancy, (b) complexity beyond linear patterns in BFI and NFC has an advantage in precisely identifying the majority of individuals with leadership roles, and (c) the fact that complexity had an advantage justifies further exploration of RF (i.e., opening up the black box).

Research question 2a: How important is each trait in predicting leadership role occupancy within the RF?

One important goal in model interpretation is understanding the influence of each of the predictor variables on the target variable. To estimate which predictor variable is most strongly related to the target variable, scholars often focus on the standardized regression coefficients (β) when using a LM approach. Here, BFI and NFC predictor variables with regression coefficients further from zero have a relatively stronger association with the target variable and are thus seen as more important. In Table 2, we provide the coefficients of the LM. The Breusch-Pagan test indicated the existence of heteroskedasticity. Hence, we used the Huber-White sandwich estimator to compute the standard errors. As reported in this table, the LM approach suggests that NFC ($\beta = 0.528$, $p < .001$), extraversion ($\beta = 0.312$, $p < .001$), and agreeableness ($\beta = -0.368$, $p < .01$) are the most important traits for leadership role occupancy, while conscientiousness ($\beta = 0.145$, $p = .058$) and neuroticism ($\beta = -0.153$, $p = .052$) are marginally significant. The unstandardized regression coefficients (B) represent the slope of the line between the predictor variables and the dependent variable. For example, for every increase in NFC of one, the likelihood of leadership role occupancy increases by 0.558 ($SE = 0.102$, $p < .001$). The next step is searching for unique patterns in the RF black box.

Table 2
Linear modeling results focused at the relationship between BFI and NFC traits and leadership role occupancy.

Variable	B	SE	β	p
Openness	0.147	0.099	0.076	0.428
Conscientiousness	0.286	0.076	0.145	0.058
Extraversion	0.477***	0.080	0.312	<0.001
Agreeableness	−0.733***	0.070	−0.368	<0.001
Neuroticism	−0.218	0.075	−0.153	0.052
NFC	0.558***	0.075	0.528	<0.001
Constant	−4.517***	0.090	−2.381	<0.001
Chi-square	125.740			
df	6			

Note. Results are based on the cross-validation set. $N = 2258$. * $p < .05$. ** $p < .01$. *** $p < .001$.

The interpretation of algorithmic machine learning models, like the RF, is often considered harder (i.e., a black box). Fortunately, scholars have developed numerous ways that allow us to peek inside these more complex models and interpret them. For instance, one way to interpret the relative importance of predictor variables is via perturbation (Breiman, 2001; Fisher, Rudin, & Dominici, 2019). In machine learning, perturbation refers to adding random noise to the original values of the predictor variables. If small, random changes are made to the values of one of the predictor variables for all observations, this will (potentially) influence the predictions. The new predictions are likely to be more inaccurate, as the model was optimized to minimize the errors.

In other words, adding random changes to the predictor variables will result in changes (often increments) to the residual errors on the target variable (i.e., predicted value minus observed value). If we observe many or large changes in residual errors due to the random noise added to a specific predictor variable, we know that the original value on this predictor variable was quite relevant to the predicted value on the target variable. Hence, we can conclude that this variable is important for predicting the target variable in the current model. By running such a perturbation process iteratively for each of the predictors, and by calculating the sum of the changes in the residual errors for each of the perturbated predictor variables, the predictor importance can be estimated.

In order to explore the importance of each trait in predicting leadership role occupancy on the test data, we examined the increment in the prediction error after perturbing the predictors (Fisher, Rudin, & Dominici, 2019). We perturbed single predictor variables by randomly changing their values 200 times and keeping all other values constant to calculate the "loss drop" in the test set. The loss drop refers to the increase in the quadratic mean of the difference between the predicted likelihood (ranging from 0% to 100%) and the actual leadership role occupancy (0% or 100%). Note that 11.4% of the individuals in the test set were in a leadership role, which means that if a model were to predict 0% (i.e. no leadership role occupancy) for each individual, the root mean squared error (RMSE) would have been 11.4%. Thus, the RMSE will increase when the perturbations result in a decrease (i.e., drop) in predictive performance.

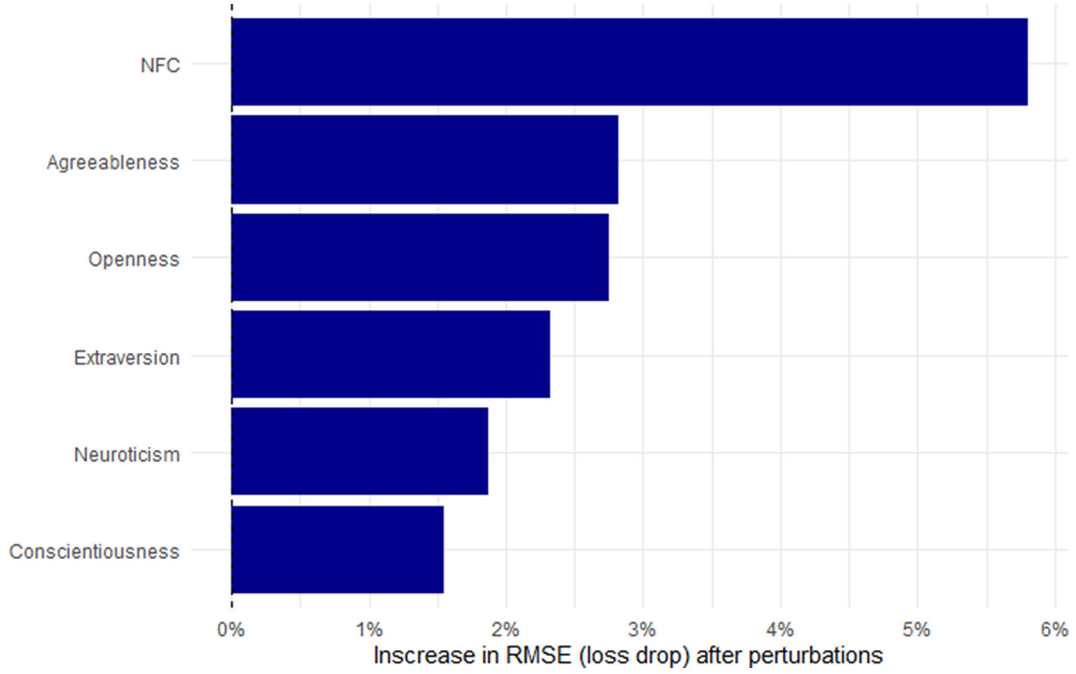

Fig. 2. Importance of traits in predicting leadership role occupancy. *Note.* The size of the bars refers to the increase in the RMSE introduced by randomly perturbing the predictors. *NFC* = Need for cognition.

We visualized the results in Fig. 2, in which we ordered the BFI and NFC traits from most important to least important. The predictions become most inaccurate upon perturbing NFC: the RMSE of the RF model increases by about 5.8%. NFC thus seems most important in predicting leadership role occupancy. Agreeableness and openness are the second most important traits in the RF (loss drop of 2.8%). The least important traits in the RF were – in descending order – extraversion (loss drop of 2.3%), neuroticism (loss drop of 1.9%), and conscientiousness (loss drop of 1.6%).

Research question 2b: What is the shape of the relationship between traits and leadership role occupancy within the RF?

Beyond identifying the importance of each predictor, another common goal in interpreting models is understanding the direction of the relationship between a predictor and a target variable. The LM offers this understanding through the sign of the regression coefficients: positive coefficients indicate that high values on the predictor variable result in high values on the target variable, whereas negative coefficients indicate that high values on the predictor variable result in low values on the target variable.

Researchers have developed alternative approaches to gain similar insights in the direction of relationships contained in algorithmic machine learning models, such as the RF. Because RFs may include complex, non-linear, and multi-variable relations, understanding directionality is more elaborate than simply quantifying whether high values on the predictor variable result in high or low values on the target variable. The technique that we use to examine the shape of the relationship is dubbed the "What-If" approach, or the *Individual Conditional Expectations* (Goldstein, Kapelner, Bleich, & Pitkin, 2015). This technique follows the ceteris paribus principle, which is Latin for "all other things held constant." It works by using the RF model in multiple prediction rounds. During each round, an artificial dataset is simulated where the values on a predictor variable of interest are manipulated while all other predictor variables are kept constant at their average value (i.e. ceteris paribus).

The resulting dataset is artificial in the sense that it consists of many simulated observations, with values running from the minimum to the maximum observed value on the predictor variable of interest, and constant values on all other predictor variables. By running these artificial observations through the RF model, the predicted values for the target variable can be obtained. This allows to inspect, across the whole range of values of a predictor, how this predictor variable relates to the target variable for the average observation in the dataset. By repeating this process for all predictor variables, one gets a basic understanding of the direction of the relationships between the predictors and the target variable.

Fig. 3 demonstrates the results of this ceteris paribus, what-if approach for the RF model. The figure shows the average leadership role occupancy likelihood as a function of each BFI and NFC predictor. For instance, for the personality trait openness, Fig. 3 demonstrates that the RF leverages a U-shaped relationship between openness scores and leadership role occupancy. In this case, the RF predicts the highest likelihood of leadership role occupancy for individuals that score either low (<2.5) or high (>5) on openness. A similar U-shaped relationship is leveraged for conscientiousness, where a higher likelihood is predicted for individuals who are low (<2.5) on conscientiousness and high (>4.8) on conscientiousness. In contrast, the relationships between leadership role occupancy and agreeableness and neuroticism in the RF model seems more linear. As individuals score higher on each of these traits, the predicted likelihood decreases. The relationships between extraversion and leadership role occupancy and NFC and leadership role occupancy discovered by the RF are quite different once more. The effect of extraversion seems almost quadratic. For extraversion, the RF predicts a stable low 8% leadership role occupancy for individuals with extraversion scores up to 2.5. The leadership role occupancy likelihood then increases somewhat linearly with extraversion, up to scores of 4.5, at which point the likelihood jumps up very strongly, up to nearly 20%. Note that 11.4% was used as cut-off value for predicting leadership role occupancy, because the test data contained 1127 observations of which 129 (11.4%) were in a leadership role. For NFC, the leveraged relationship seems to follow a sigmoid curve, where again the leadership role occupancy likelihood is low and stable for low values of NFC. Then,

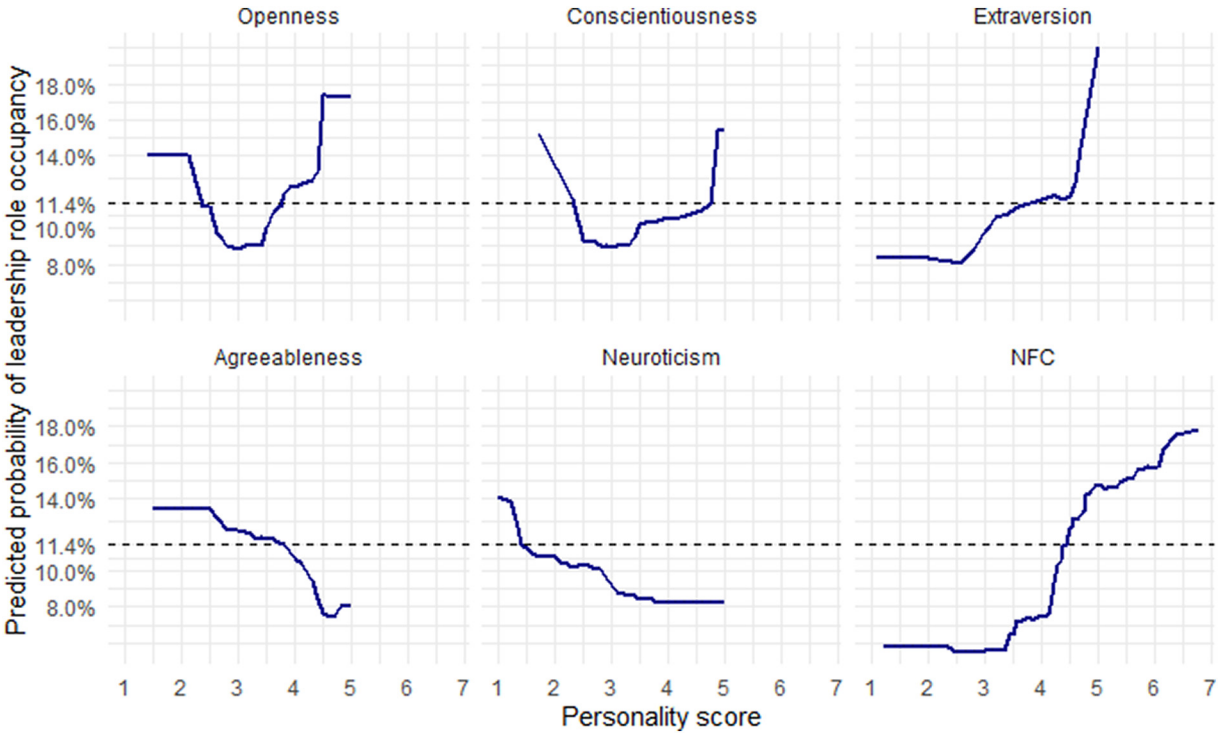

Fig. 3. Accumulated local effect plots showing average predictions for leadership role occupancy on the test set. The dashed line is the cut-off value we used. Values above the line suggest leadership role occupancy.

the likelihood rises quickly for NFC scores ranging between 4 and 5. Finally, the leadership role occupancy likelihood predicted by the RF rises only linearly for individuals with scores above 5 for NFC.

In conclusion, through a ceterus paribus – or what-if – approach, scholars are able to open up the black box of machine learning models such as the RF. Fig. 3 demonstrates that the RF model clearly leverages more complex effects than simple linear changes. The RF helps to find the U-shaped effects of openness and conscientiousness, or the respective quadratic or sigmoid effects of extraversion and NFC. Such findings can be used as input for inductively building theory on the (non)linearity of the relationship between leadership and traits, and in turn can inspire deductive tests of the relationship.

Research question 2c: What interactions are leveraged by the RF for predicting leadership role occupancy?

To better understand the leader trait paradigm, scholars have also started to unravel what combinations of traits are important for leadership (Jensen & Patel, 2011). In this endeavor, LMs have helped in testing interaction effects (King, George, & Hebl, 2005). Now, building on existing research, algorithmic machine learning models such as RF can explore the impact of combinations of multiple variables with increased flexibility. Hence, these models might provide valuable insights about what combinations of traits are important for leadership role occupancy (and leadership in general).

To get a sense of the interactions within RF, scholars can use the H-statistic (Friedman & Popescu, 2008). In brief, this statistic uses partial dependency decomposition – similar to the "what-if" approach we described – to measure to what extent interactions explain variance in the predicted outcome. It generates a value between 0 and 1 for each pair of variables. A value of 0 suggests no interaction between the two variables, and a value of 1 suggests no main effects (i.e., the prediction is solely based on the interaction). The H-statistic, when computed for single variables, estimates the extent to which that variable has an interaction effect with any other variable on the target variable.

The importance of interactions between traits in predicting leadership role occupancy is visualized in Fig. 4a. This figure shows that NFC has the strongest interaction effect (H-statistic), meaning that interactions between NFC and the other traits contribute the most to leadership role occupancy predictions. To better understand the interactions between NFC and the other traits, we computed the H-statistic for each pair with NFC. As illustrated in Fig. 4b, the interaction between openness and NFC contributed the most in predicting leadership role occupancy, which we explore next.

To further explore the interaction between BFI traits and NFC, we present the RF predictions on the test set against BFI and NFC scores in Fig. 5. In this figure, each person is denoted as either a blue circle or a gray triangle. Blue circles are used when the RF predicts *leadership role occupancy* (N = 129) and as gray triangles when the RF predicts *no leadership role occupancy*. As illustrated in Fig. 5, across all interactions, the RF frequently predicted leadership role occupancy for individuals scoring higher (at least 4) on NFC. Further, this higher NFC resulted often in predicted leadership role occupancy for individuals higher on openness (>3.5), conscientiousness (>3.0), and extraversion (>3.0); lower on neuroticism (>3.0); and not too high on agreeableness (<4.5). Note that these findings are largely in line with the accumulated local effects plotted in Fig. 3, in which we visualized the effects of each trait separately.

Discussion

This paper demonstrates how scholars can use algorithmic machine learning techniques to uncover complex relationships. In demonstrating the techniques, we provided an assessment of the leader trait paradigm through cross-validation, out-of-sample prediction, and interpretation methods. Specifically, we (a) explored *complexity* beyond linear additive trait effects and (b) showed how to gain *interpretability* of black box algorithmic techniques. Exploring complexity and gaining interpretability as demonstrated in this paper fits within the tradition of *abduction*, which is a process of deriving hypotheses by appraising observations in light of the theory (Mantere & Ketokivi, 2013). As

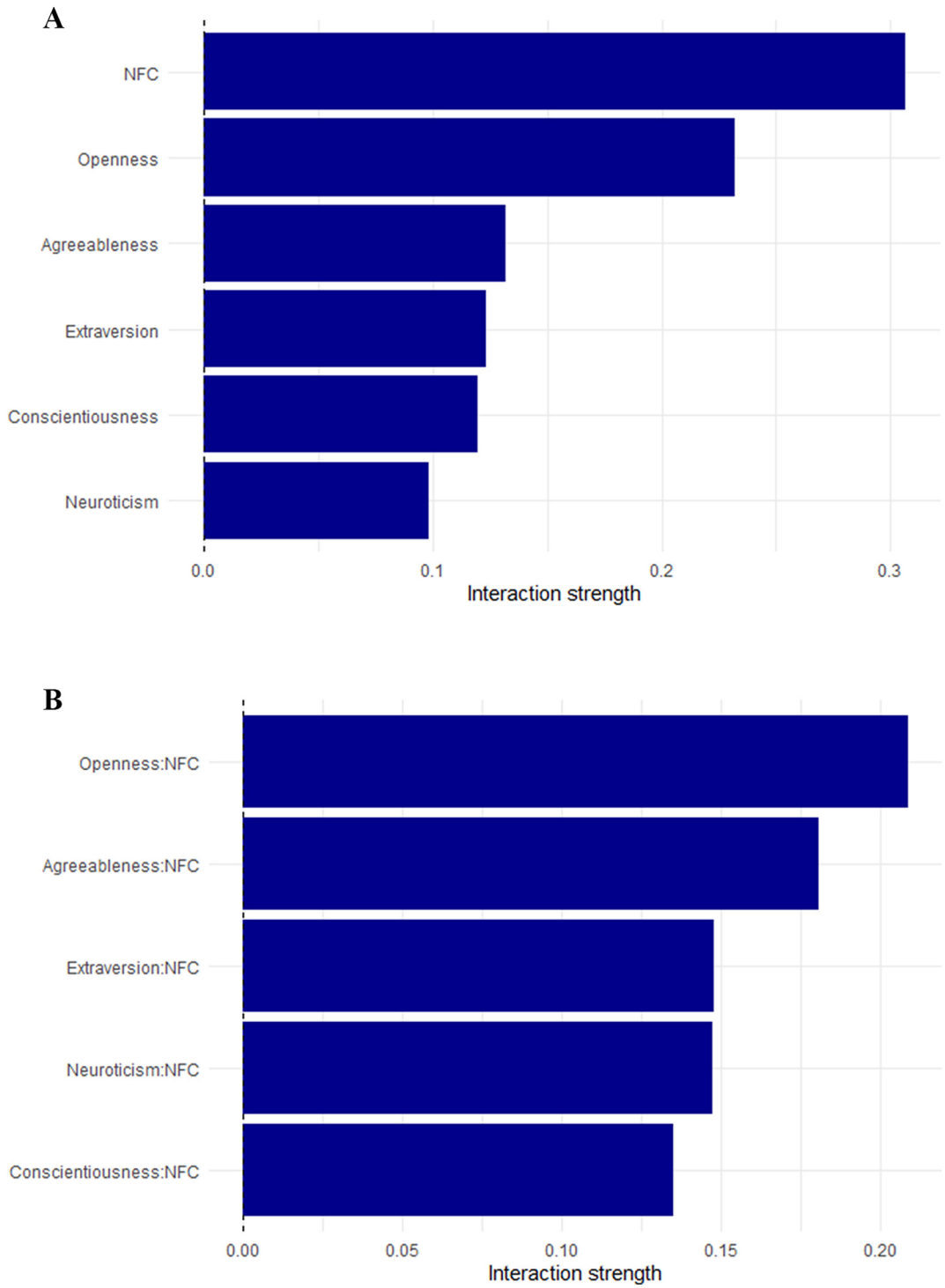

Fig. 4. (A) The interaction strength (H-statistic) for each variable with all other variables, predicting leadership role occupancy. (B) The interaction strength (H-statistic) for need for cognition (NFC) with all other variables, predicting leadership role occupancy.

described in more detail next, our findings provide (inductive) insights relevant to the interactionist theories of the leader trait paradigm and provide directions for (deductively) testing non-linear effects.

Answering our first research question, focusing on the degree to which non-linear effects and interactions help explain the trait-leadership role occupancy relationship, we compared the predictive performance of a linear model (LM) with the predictive performance of a random forest (RF) model. While comparing the two models, we offered more flexibility to the RF to fit complexity beyond linear additive effects. Our results suggested that the RF outperformed the LM in the cross-validation step and indicated that the RF had some advantage in predicting leadership role occupancy in the test set. These results suggest that modeling *complexity* beyond linear and additive effects of traits has *some* added value in predicting leadership role occupancy (RQ 1), supporting the interactionist accounts of the leader-trait paradigm.

Finding supporting evidence for leader-trait complexity raised the problem of *interpretability* (i.e., predictive complexity inhibits

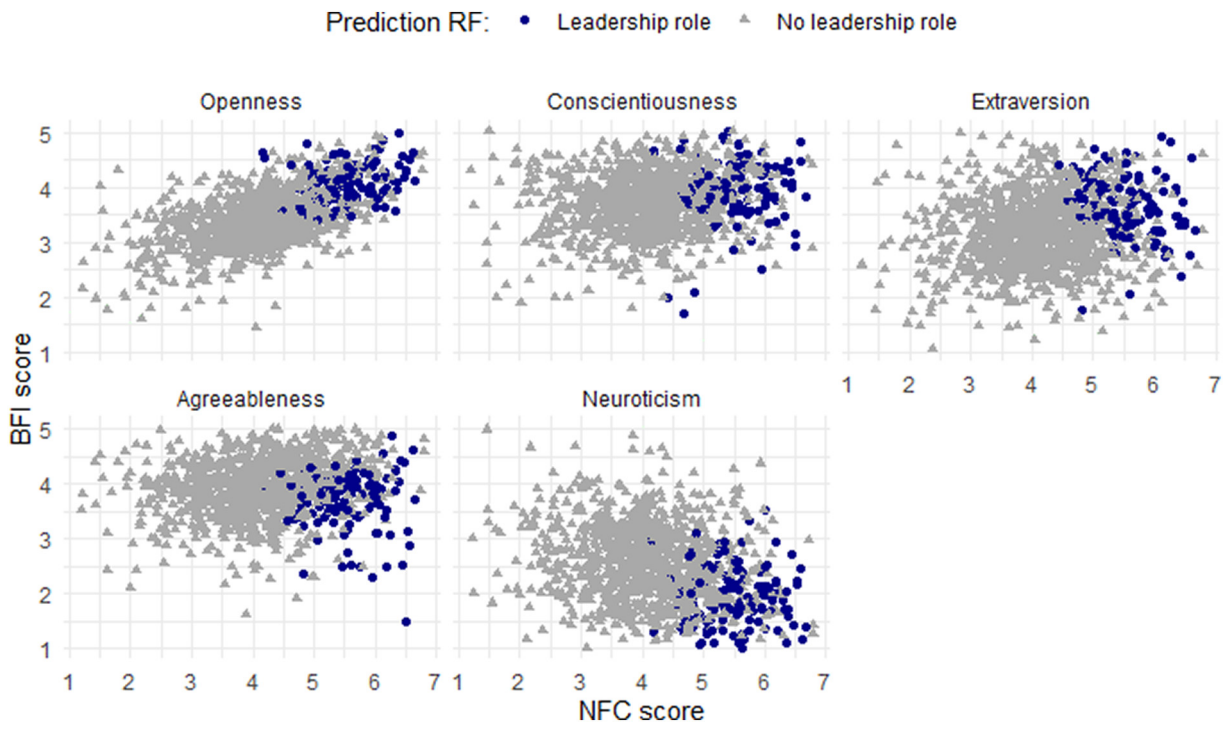

Fig. 5. Leadership role occupancy predictions by the RF as a function of NFC and each BFI trait. Points refer to persons.

explanatory clarity). To solve this dilemma, we introduced three methods for opening the RF black box. Addressing our second set of research question, we first explored the importance of each trait in the RF (RQ 2a) and found that NFC was most important for leadership role occupancy. To the best of our knowledge, this is the first time that NFC has been empirically tested in relation to leadership role occupancy. Second, we explored the shape of each factor's relationship with leadership role occupancy (RQ 2b), which revealed novel (inductive) insights relevant to the interactionist theories, such as the U-shaped effects of openness and conscientiousness, as well as the respective quadratic and sigmoid effects of extraversion and NFC. The U-shaped relationships of openness and conscientiousness, for example, suggests four paths to leadership roles: being extremely (a) curious and task-driven, (b) creative and easy-going, (c) cautious and efficient, and (d) conservative and unorganized. Third, we unraveled the interaction strengths of each trait combination (RQ 2c) and demonstrated that NFC had the strongest interaction effect, followed by openness. Here, again, NFC appears to be an important line of future research. Based on a visual inspection, we demonstrated a higher likelihood of leadership role occupancy for, among others, those scoring high on *both* NFC and openness. This is not surprising considering the creativity path to leadership role occupancy we uncovered in our RQ 2b results.

Our findings suggest that further exploration of the leader trait paradigm could benefit from (deductively) testing non-linear effects. In studying these effects, scholars should consider cross-validation to avoid overfit and biased results. Moving forward, scholars can use algorithmic machine learning to find novel insights for deductive testing using conventional analytical approaches. Important consideration should be given to the predictor variables selected for the models. As reported in the correlation table (Table 1), leadership occupancy was more common among men and older individuals. In a post-hoc analysis, we examined whether personality yielded different effects when gender and age were included in the models. We found similar effects of personality, contingent on gender and age, such that the likelihood of leadership role occupancy was higher for older men. The more variables included in the algorithmic machine learning models, the more complex patterns can be explored. Insights derived from such models may change if important predictor variables are omitted. Using the output of algorithmic machine learning as input for conventional analytical approaches is needed to test the validity of findings and results in clear parametric outcomes – suitable for meta-analytical tests. The takeaway here is that, again, LM and algorithmic methods such as RF are not in competition. Rather, they are complementary tools for scientific discovery.

The methods demonstrated in this paper are also relevant to practitioners and offer great opportunities for further exploration of the leader trait paradigm. Models like the ones we developed could inspire organizations in the selection of future leaders and the succession planning for current leaders. We found modeling more complex effects through algorithmic models suitable when the goal is to precisely identify the majority of actual leaders rather than broadly identifying (or recalling) all leaders. Thus, RF potentially reduces the trial and error costs of leader selection. As the study of machine learning and leadership matures, efforts can be made to further reduce such costs by incorporating trait data with other relevant factors, such as performance measures. Increasing data dimensionality will likely add to the RF advantage when it comes to identifying high potentials faster and with greater accuracy (Spisak et al., 2019).

Incorporating algorithmic machine learning techniques into decision-making also allows for a better understanding of (a) who has a personality profile similar to individuals in leadership positions and (b) how effective they will likely be in a leadership position. It is important to make this distinction between leadership role occupancy and effectiveness given that the ability to occupy a leadership position does not necessarily translate into effectiveness. With machine learning techniques, similar to those we highlighted in this paper, academics and practitioners can explore the potentially complex differences between leadership role occupancy and leadership effectiveness to reduce the chance of making false positives and false negatives (i.e., selecting leaders low on effectiveness, or *not* selecting effective leaders who are

Table 3
An overview of the demonstrated statistical/machine learning concepts.

Concept (what)	Explanation (how)	Practical use (why)	Further readings
Cross-validation and a holdout (validation) sample	A validation method in which part of the data (e.g., 80%) is used to iteratively fit (e.g., train) a model and validate its performance. The remaining data (e.g., 20%) is subsequently used to assess how well the optimized model performs on "unseen" data.	A model that is optimized for predictive performance on new data, as well as an estimate of how well your model can predict new data.	Arlot and Celisse (2010); Friedman, Hastie, & Tibshirani (2001) (Friedman, Hastie, & Tibshirani, 2001); Koul, Becchio, and Cavallo (2018); James et al. (2013); Kuhn and Johnson (2013); Zhang (1993)
Algorithmic machine learning such as random forest	A non-parametric statistical modeling algorithm that constructs many decision trees, each fitted based on a subset of the data and predictor variables. Because such algorithms are not bound to modeling linear effects, they can more freely model interactions and other complex patterns. The algorithm then makes a final prediction by averaging the predictions of the individual trees.	A model that is optimized for predictive performance on new data and which leverages the "wisdom of the crowd" by combining the predictions of many weaker models that each consider unique and/or overlapping predictive information residing in subsets of features and observations.	Breiman (2001); Biau and Scornet (2016); Friedman, Hastie, & Tibshirani (2001); James et al. (2013)
Area under the curve (AUC)	The area under the curve is a metric is computed using the receiver operating characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. Consequently, one can calculate the (proportional) area under this curve.	A model evaluation metric which quantifies how well a classification model performs for the full range of potential classification thresholds.	Bradley (1997); Hanley (2014); James et al. (2013)
Variable importance	The variable importance is a metric computed by randomly changing the variable values input into a predictive model, and averaging the consequent changes in the predictions on the outcome variable.	A metric to assess the predictive value of an independent variable in a non-parametric model.	Gregorutti, Michel, and Saint-Pierre (2017); Grömping (2009); Kuhn and Johnson (2013); James et al. (2013); Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008)
Ceteris Paribus; Individual conditional expectations; What-if approach	This approach requires the creation of artificial datasets in which values for all predictor variables but one are kept constant. For the non-constant variable, values ranging from the minimum to the maximum value are simulated. Next, these artificial datasets are input into a trained model to gather predictions for the artificial observations.	An approach that allows inspection of the ways in which variables influence the predictions of a non-parametric model.	Friedman (2001); Friedman and Popescu (2008); Friedman, Hastie, & Tibshirani (2001); Goldstein et al. (2015); Strobl et al. (2008)
H-statistic; Partial dependency decomposition	The H-statistic indicates to what extent interactions explain variance in the predicted outcome.	An approach, based on partial dependency decomposition, that allows inspection and quantification of the predictive value of interactions between variables in a non-parametric model.	Friedman (2001); Friedman and Popescu (2008); Friedman, Hastie, & Tibshirani (2001)

low on emergent qualities). Incorporating machine learning could help academics forge a new understanding of what makes an effective leader and help practitioners remove even more cost from the leader selection process.

Despite the added value of machine learning, it is not without limitations. Data, in particular related to the "four V's" of big data, is a perpetual concern: volume, variety, velocity, and veracity. First, though the amount of data we used here was large relative to typical datasets in leadership research, it was by no means "big data", the space where machine learning is at its best (i.e., volume). Second, in efforts to advance the leader-trait paradigm, we focused on traits. However, several other features such as height and intelligence (Ilies et al., 2004) are important for leadership as well (i.e., variety). Third, we studied the leader trait cross-sectionally and did not focus on changes in the data, such as changes in leadership role occupancy – which would especially be interesting to monitor as changes in contextual factors such as market competition and disruption alter who is preferred as a leader (velocity). Fourth, though our sample is from a trusted source, using validated measures, there is the perennial concern with data quality (i.e., veracity). Scholars will need to remain vigilant when it comes to quality as the hunt for compelling datasets intensifies.

That said, these limitations also represent exciting opportunities for future research. First, organizations and society are providing a never-ending stream of relevant data. Leadership scholars and practitioners can utilize these data streams to better explore and leverage complexity through machine learning. A good first step is perhaps teaming up with data scientists who are familiar with using large streams of data.

Second, machine learning can provide benefits even for those not concerned with big data. Specifically, we encourage the use of linear models in machine learning workflows. Though our LM was relatively simple, it utilized an iterated cross-validation approach finding the optimal predictive model. Such an approach strengthens analytical rigor and improves estimates of actual model performance. As we touched on above, cross-validation, for example, is already more common in other research fields, such as marketing (Cooil et al., 1987) and decision-making (Puelz & Sobol, 1995).

Finally, the use of algorithmic modeling in combination with analytical tools to interpret the black box will uncover novel patterns and relationships for future investigation. This output can provide the sort of unexpected connections necessary for advancing theory and developing new hypotheses using inherently more interpretable models. Our findings suggest, for example, that the combination between NFC and openness was important for the RF in predicting leadership role occupancy. Scholars who have more experience with conventional methods could then test this interaction. The same goes for the U-shape effects of openness and conscientiousness. Thus, algorithmic-driven output can be easily integrated into the broader leadership research community when interpretability is introduced. What is important to continually remember is that models such as LM an RF can work in unison, where RF can help to turn undiscovered complex patterns into theory, and LM can

help to test this theory through hypothesis (Kolkman & van Witteloostuijn, 2019).

In the present paper, we tested the complexity of the leader trait paradigm by comparing the predictive performance of a LM and a RF. We demonstrated how scholars can open the black box of RF models through a number of analytical techniques. While we provided a first look into the application of several techniques, we recommend readers to explore more in-depth resources regarding these techniques (see Table 3). For those interested in learning more about statistical and machine learning, we highly recommend the introductory though practical books written by James, Witten, Hastie, and Tibshirani (2013) and Kuhn and Johnson (2013). Moreover, we gladly refer to academic papers which provide in-depth insights into the application of machine learning in management research (e.g., Garcia-Arroyo & Osca, in press; George, Osinga, Lavie, & Scott, 2016; Wenzel & Van Quaquebeke, 2018). Collectively, applied to the complexity of the leader trait paradigm, we (a) demonstrated a way to test the degree to which complexity helps to explain a relationship while (b) maintaining a usable level of interpretability. This balance of complex yet interpretable results represents a significant step forward in the study of leadership.

Declaration of Competing Interest

None.

References

Antonakis, J. (2011). Predictors of leadership: The usual suspects and the suspect traits. In A. Bryman, D. Collinson, K. Grint, B. Jackson, & M. Uhl-Bien (Eds.), *Sage handbook of leadership* (pp. 269–285). London, England: Sage.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Babalola, M. T., Bligh, M. C., Ogunfowora, B., Guo, L., & Garba, O. A. (2019). The mind is willing, but the situation constrains: Why and when leader conscientiousness relates to ethical leadership. *Journal of Business Ethics*, *155*(1), 75–89.

Barling, J., & Weatherhead, J. G. (2016). Persistent exposure to poverty during childhood limits later leader emergence. *Journal of Applied Psychology*, *101*(9), 1305–1318.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Burns, J. M. (1978). *Leadership.* New York, NY: Harper and Row.

Cacioppo, J. T., & Petty, R. E. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. *Journal of Marketing Research*, *24*(3), 271–279.

De Neve, J. E., Mikhaylov, S., Dawes, C. T., Christakis, N. A., & Fowler, J. H. (2013). Born to lead? A twin design and genetic association study of leadership role occupancy. *The Leadership Quarterly*, *24*(1), 45–60.

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

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta- analysis. *Psychological Bulletin*, *108*(2), 233–256.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of Management Journal*, *50*(1), 25–32.

Ensari, N., Riggio, R. E., Christian, J., & Carslaw, G. (2011). Who emerges as a leader? Meta-analyses of individual differences as predictors of leadership emergence. *Personality and Individual Differences*, *51*(4), 532–536.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *Journal of Machine Learning Research*, *20*(177), 1–81 (arXiv: 1801.01489).

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

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954.

Furnham, A., & Thorne, J. D. (2013). Need for cognition: Its dimensionality and personality and intelligence correlates. *Journal of Individual Differences*, *34*, 230–240.

Garcia-Arroyo, J., & Osca, A. (in press). Big data contributions to human resource management: A systematic review. The International Journal of Human Resource Management.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

George, G., Osinga, E., Lavie, D., & Scott, B. (2016). Big data and data science methods for management research. *Academy of Management Journal*, *59*(5), 1493–1507.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Govindan, K., Cheng, T. C. E., Mishra, N., & Shukla, N. (2018). Big data analytics and application for logistics and supply chain management. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 343–349.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308–319.

Hanley, J. A. (2014). *Receiver operating characteristic (ROC) curves.* Wiley StatsRef: Statistics Reference Online.

Ilies, R., Gerhardt, M. W., & Le, H. (2004). Individual differences in leadership emergence: Integrating meta-analytic findings and behavioral genetics estimates. *International Journal of Selection and Assessment*, *12*(3), 207–219.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning. Vol. 112.*New York: Springer (p. 18).

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.

Judge, T. A., & Bono, J. E. (2000). Five-factor model of personality and transformational leadership. *Journal of Applied Psychology*, *85*(5), 751–765.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*(4), 765–780.

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.

Kolkman, D. A., & van Witteloostuijn, A. (2019). Data science in strategy: Machine learning and text analysis in the study of firm growth. Tinbergen Institute Discussion Paper 2019-066/VI. Available at SSRN: https://ssrn.com/abstract=3457271 or https://doi.org/10.2139/ssrn.3457271.

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. Frontiers in Psychology, 9, 1117, 1–4.

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

Mantere, S., & Ketokivi, M. (2013). Reasoning in organization science. *Academy of Management Review*, *38*(1), 70–89.

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.

Puelz, A. V., & Sobol, M. G. (1995). The accuracy of cross-validation results in forecasting. *Decision Sciences*, *26*(6), 803–818.

Scherpenzeel, A. C., & Das, M. (2010). "True" longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: advances in applied methods and research strategies* (pp. 77–104). Boca Raton: Taylor & Francis.

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

Spisak, B. R., van der Laken, P. A., & Doornenbal, B. M. (2019). Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning? *The Leadership Quarterly*, *30*(4), 417–426.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307–318.

Strohmeier, S., & Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, *40*(7), 2410–2420.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–516.

Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327.

Van Vugt, M., Hogan, R., & Kaiser, R. B. (2008). Leadership, followership, and evolution: Some lessons from the past. *American Psychologist*, *63*(3), 182–196.

Vergauwe, J., Wille, B., Hofmans, J., Kaiser, R. B., & De Fruyt, F. (2018). The double-edged sword of leader charisma: Understanding the curvilinear relationship between charismatic personality and leader effectiveness. *Journal of Personality and Social Psychology*, *114*(1), 110–130.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121.

Wenzel, R., & Van Quaquebeke, N. (2018). The double-edged sword of big data in organizational and management research: A review of opportunities and risks. *Organizational Research Methods*, *21*(3), 548–591.

Zaccaro, S. J. (2007). Trait-based perspectives of leadership. *American Psychologist*, *62*(1), 6–16.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 299–313.

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

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