Received: 20 April 2022 | Accepted: 17 September 2022

DOI: 10.1111/apps.12435

METHOD

APPLIED PSYCHOLOGY IAAP

Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers

Abhishek Sheetal[1,2] | **Zhou Jiang**[3] | **Lee Di Milia**[1]

[1]School of Business and Law, Central Queensland University, Norman Gardens, Queensland, Australia

[2]Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, China

[3]Department of Business, Graduate School of Business and Law, RMIT University, Melbourne, Victoria, Australia

Correspondence
Zhou Jiang, Department of Business (or MBA Department), Graduate School of Business and Law, RMIT University, Melbourne, Victoria, Australia.
Email: dr.zhou.jiang@gmail.com

Abstract

This article introduces the research community to the power of machine learning over traditional approaches when analyzing longitudinal data. Although traditional approaches work well with small to medium datasets, machine learning models are more appropriate as the available data becomes larger and more complex. Additionally, machine learning methods are ideal for analyzing longitudinal data because they do not make any assumptions about the distribution of the dependent and independent variables or the homogeneity of the underlying population. They can also analyze cases with partial information. In this article, we use the Household, Income, and Labour Dynamics in Australia (HILDA) survey to illustrate the benefits of machine learning. Using a machine learning algorithm, we analyze the relationship between job-related variables and neuroticism across 13 years of the HILDA survey. We suggest that the results produced by machine learning can be used to generate generalizable rules from the data to augment our theoretical understanding of the domain. With a technical guide, this article offers critical information and best-practice recommendations that can assist social science researchers in conducting machine learning analysis with longitudinal data.

KEYWORDS

Big Five personality, longitudinal data, machine learning, neuroticism, Solomonoff induction, XGBoost

INTRODUCTION

Researchers typically use advanced regression-based methods to analyze longitudinal data, such as cross-lagged models, latent growth models, and autoregressive models (Chan, 2004; Kelloway & Francis, 2012; Liu et al., 2016). As these models are subsets of structural equation models (Selig & Little, 2012), they inherit several assumptions about the nature of the data, such as homoscedasticity, normal distribution, independence of observations, and lack of multicollinearity. These assumptions are often not met in cross-sectional datasets, let alone longitudinal data (Erceg-Hurn & Mirosevich, 2008). Additionally, researchers may sample somewhat different populations at different time points in longitudinal data collection efforts. For example, the Household, Income, and Labour Dynamics in Australia (HILDA) survey added new households in every wave who might or might not be sampled from the same population as the previous households in the dataset, thereby violating the homogeneity assumption of regressions. If researchers run regression-based approaches when these assumptions are violated, they can no longer guarantee the Type I error rate, which is the foundation of null hypothesis significance testing.

In addition, although longitudinal research designs are powerful because they allow researchers to make causal statements, longitudinal datasets often have problematic features. Many longitudinal studies insert or delete questions of interest in response to contemporary understanding of the constructs under investigation or to capture current phenomena, such as COVID-19. Furthermore, when participants are added as the study proceeds, participants recruited in subsequent waves were not represented in earlier waves. Similarly, participants drop out across successive waves, which means that participants recruited in previous waves were not represented in later ones. Thus, data are missing across periods, and they are not missing sparsely and at random; instead, data are missing systematically and in substantial volume. Researchers typically exclude cases with any missing data (e.g. see Wu et al., 2020). However, individuals without missing data are virtually always guaranteed to be non-representative subsets of the entire dataset. For example, in the HILDA survey, individuals without missing data would have to participate in every wave conducted, which is highly unlikely. Suppose researchers drop a substantial proportion of the observations due to missing values. In that case, they cannot generalize from the sample to the population, which is a crucial requirement for applied psychology research (Kitchenham & Pfleeger, 2002).

Given these issues, we argue that machine learning methods are ideal for analyzing longitudinal data. Unlike regression-based methods, many machine learning methods, such as random forest, gradient boosting, and deep learning, make no assumptions about the data. That is, these methods can be used to analyze data even if the observations are non-independent (e.g. individuals are clustered within households); the dependent variable is non-normally distributed; there is a high degree of multicollinearity among the predictor variables; the data come from different subpopulations, or there is a large volume of systematically missing values. This article demonstrates how industrial and organizational (IO) psychologists can use a flexible machine learning method, extreme gradient boosting (XGBoost), to analyze longitudinal data.

In addition to addressing common challenges encountered in longitudinal data analyses, machine learning methods can also help extract maximum information from the data. For example, researchers typically measure constructs using multiple intercorrelated items and then average the items to form a scale; in doing so, we lose discrete information at the item level, which may help explain the outcome variable above and beyond the scale average. For example, an inspection of the correlation between individual items and the outcome variable often identifies items with a stronger correlation than the scale average. In regression models, the inclusion of multiple intercorrelated items as predictors would lead to multicollinearity and is therefore problematic. However, machine learning models can easily handle multicollinearity, so the researcher can choose to use item-level scores and quantify the total impact of all items belonging to a single scale on the dependent variable. Thus, machine learning methods can maximize predictive power by retaining individual items without sacrificing interpretability. Researchers can compare the impact of all items belonging to different scales.

In addition, machine learning has implications for how we construe knowledge. The bulk of management and organizational research relies on hypothetico-deductive reasoning (H-D) supported by null-hypothesis-significance-testing (NHST). We argue that this well accepted combination should be the default method for future research. However, it is based on two assumptions. Firstly, the data meets the assumptions of NHST must be respected. Secondly, most of the analysis in research starts with an assumption of linearity—that the relationship between predictor and outcome is linear. If these two data assumptions cannot be met, then researchers have an option to switch to approximate[1] Solomonoff induction (Choudhury et al., 2021; Shrestha et al., 2021; Solomonoff, 1967) or to Peirce's abduction (Peirce, 1903, as cited in the Peirce Edition Project, 1998; Behfar & Okhuysen, 2018; Sheetal et al., 2020). Solomonoff induction is similar to ordinary least squares (OLS) where a known set of variables are already established based on past literature. Machine learning focuses on building a relationship and learning the interaction between the variables. However, Solomonoff induction deviates from OLS by learning any non-linearity that might exist.

The second choice to researchers is Peirce's abduction. In his seminal research, the 20th century philosopher argued that "Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis" (Burks, 1946, p. 303). In an abductive process, a researcher does not know the set of predictors, and focuses more on predictive accuracy and generalization. If a strong predictive accuracy is accomplished, the researcher then unravels the machine learning model for most logical explanations. This most logical explanation can sometimes create a novel hypothesis. The researcher needs to verify this hypothesis in a secondary controlled study. In this article, we focus on Solomonoff induction based scientific reasoning with a known set of predictors. However, abduction-based articles do also exist in literature (e.g. see Sheetal et al., 2020).

BACKGROUND ON MACHINE LEARNING

There has been an increasing interest in machine learning in the behavioral and organizational sciences in recent years. However, most machine learning methods were developed decades ago. For example, the random forest algorithm is over 20 years old (Breiman, 2001). Neural networks, which are responsible for numerous breakthroughs in recent years, were created over 60 years ago (Rosenblatt, 1958). The potential value of these algorithms can now be realized

because the computational power of commercially available hardware has just matched the demands of machine learning algorithms (Rosett & Hagerty, 2021). In the absence of affordable computational power, machine learning remained impractical for social science researchers. Hence, when researchers consider which machine learning method to use, they need to consider the computational power at their disposal. In terms of hardware, performing machine learning analyses on graphics card processing units (GPU) can be as much as 200 times faster than performing the same analyses on the computer's central processing unit (CPU) (Shi et al., 2016).

Such characteristics indicate some important differences between machine learning and traditional statistical analyses. Table 1 summarizes some of the attributes that differentiate these two methods. This comparison applies to any analytical problem, whether longitudinal or cross-sectional. The key difference is that traditional analyses are ideally suited to test prespecified hypotheses and assess whether a linear relationship exists between individual predictors and the outcome. In contrast, machine learning is ideally suited to accurately predicting the outcome and engaging in abductive reasoning, identifying new patterns in the data that can be subsequently verified and help move the field forward. Since traditional methods are still largely the default in the social sciences such as management and applied psychology, researchers may continue to use conventional statistical analyses. However, as illustrated in this article, machine learning is a powerful tool to help us explore, understand, and interpret datasets as well as make more reliable predictions. While it may be challenging to take researchers away from the 'default' conventional statistical approaches, the advantages of machine learning methods should be at least equally valued and endorsed. Similar to traditional statistical methods, some machine learning methods do make assumptions about the data (see Table 2). However, commonly used methods such as gradient boosting, random forest, and neural networks do not make any assumptions about the data distribution.

Although many traditional statistical methods can be run using a single function call, machine learning models require programming skills. However, most machine learning software and codes are in the open-source domain. Researchers can freely copy and edit codes that others have written and posted online (e.g. Github and StackOverflow). As most codes have often gone through multiple rounds of review before and after they are posted online, reusing peer-reviewed codes can reduce coding errors by as much as 50% (Cusumano, 1989). Numerous reference books provide guidelines for designing machine learning models and provide code snippets (Gareth et al., 2013; Goodfellow et al., 2016; Kuhn & Johnson, 2013; Kuhn & Johnson, 2019; Molnar, 2020). Hence, researchers do not need to write the entire program from scratch. This article presents the machine learning model development process suggested by Kuhn and Johnson (2013) for the specific use case of analyzing longitudinal data in the behavioral sciences. In addition, we explain any deviations from procedures suggested by Kuhn and Johnson (2013). This article cannot answer each problem encountered in longitudinal data analysis; however, it provides a guide to ask the right questions at each step of the machine learning modeling process.

Significant groundwork has already been completed to simplify machine learning development for non-computer science researchers, such as hiding away complicated matrix multiplications performed on graphics cards. Even "cross-validation," a key component of machine learning models (Hay, 1950; Refaeilzadeh et al., 2016), now appears as an optional parameter in many machine learning function calls. This is made possible via a programming style called object-oriented programming, in which the software exists in layers (Stroustrup, 1988). The lowest layer manages the matrix multiplications in graphics cards. The topmost layer presents a

TABLE 1 Comparison of attributes between traditional statistical and machine learning based analyses

Attribute	Traditional statistics	Machine learning
Most common usage	Assessing whether relationships can be generalized from the sample to the population	Accurately predicting or classifying future observations
Main goal	Testing whether pre-specified relationships exist in the data	Identifying patterns in the data without pre-conception
Ideally suited for	Deductive hypothesis testing	Abductive hypothesis generation (see Peirce, 1903, as cited in the Peirce Edition Project, 1998, and examples in Sheetal et al., 2020)
Shape of relationships between variables	Fits data onto predefined shapes specified in the statistical model	Learn the true shape of the relationship between variables
Mathematical proofs	The regression line is proven to be the best linear fit	The results are suggestive and cannot be proven to be optimal (Reyzin, 2019)
How to trust the analysis	Standard robustness tests	Test model on unseen (i.e. new) data. Test the generalizability via secondary analysis
Communicating the results to target audience	Standard equations, beta values	Shapley values (e.g. see Mokhtari et al., 2019)
Researcher skills needed	Training in statistics	Training in data science and programming
Researcher's experience needed	Experience in statistical models	Experience in analyzing diverse datasets
Computational power needed	Generally most modern laptops can do the analysis	Requires high-end computing environment
Model reuse	Need to build different models for each objective	One algorithm can be reused for different objectives
Number of predictors	Limited by multicollinearity. Adding more predictors to the model might break the model	Limited by computational power. Adding more predictors does not break model
Number of observations	Limited by availability; needing to adjust alpha based on number of observations (Maier & Lakens, 2022)	Limited by availability and computational power; more data is generally better
General pattern of results	Low predictability, high explainability (London, 2019)	High predictability, low explainability. Even though advances are continually happening to explain ML models, explainability is limited

simplified function call with inputs such as the dependent and independent variables. It hides away the other complex details from the researcher. These "machine learning frameworks" can now be readily used by researchers with minimal experience with machine learning. Popular frameworks include "caret" (Kuhn & Johnson, 2013), "mlr" (Bischl et al., 2016), and

TABLE 2 Assumptions made by various analytical methods

	Assumption about the data	Reference
Linear regression	Homoskedasticity, lack of multi-collinearity, and independently, identically, and normally distributed errors	Ezekiel (1925)
Logistic regression	Homoskedasticity, lack of multi-collinearity, no outliers, linear relationships in the logit metric	Stoltzfus (2011)
K-nearest neighbor	Independence of observations; similar observations are closer to each other in a measurable distance space	Mack and Rosenblatt (1979)
Support vector machines	Clear boundaries between groups, relatively small datasets	Tong et al. (2009)
Decision trees	Continuous variables can be discretized into meaningful buckets	Brodley and Utgoff (1995)
Naive Bayes	Independence of predictors	Lewis (1998)
LASSO	Sparsity (only a few predictors are relevant), irrelevant and relevant predictors are uncorrelated	Tibshirani (1996)
Random forest	No missing data, requires hyperparameter search	Breiman (2001)
Neural networks	No missing data, requires hyperparameter search	LeCun et al. (2015)
Gradient boosting	No formal assumptions, requires hyperparameter search	Friedman (2002)

"tidymodels" (Kuhn & Wickham, 2020). The python-based "scikit learn" is the most versatile framework and is popular among experienced programmers who seek fine-grained control over various machine learning tools. We use the "caret" framework in this article because of its ease of use.

Currently, the caret package is the most commonly used machine learning framework (Kuhn & Johnson, 2013). To date, there are 238 standard models or algorithms supported by caret (Kuhn, 2019), including those listed in Table 3. Researchers can switch between different models by passing the parameter method = "lasso" or method = "rf" to the caret function. However, different models have very different run times. Some models, such as random forest and LASSO, only run on the CPU. Otherwise, algorithms such as XGBoost and deep learning can run either on the CPU or the GPU. The computing time for XGBoost and deep learning is generally much shorter because they can run in parallel on the GPU (Cai et al., 2014).

Unlike regression-based methods with closed-form solutions, all machine learning algorithms only provide approximate solutions (Reyzin, 2019). Thus, it is impossible to prove whether a given machine learning model can solve a problem, whether a specific algorithm is superior to another, or whether a particular solution is optimal. As it is always possible that the researchers constructed a suboptimal model, the machine learning model developed needs to be compared against a baseline. We cannot use standard regressions as a baseline because the assumptions for ordinary least squares (OLS) regression are typically not met. However, researchers can use Bayesian models (i.e. a Naive Bayes model for a binary dependent variable and a Bayesian regression for a continuous dependent variable), which merely assume the independence of predictors (Jaya et al., 2019). As Bayesian regressions do not model interactions between predictors, machine learning models that can learn interaction patterns should be

TABLE 3 Past empirical or methodological literature using machine learning

Article	What this article is about
Simester et al. (2020)	Demonstrates that machine learning performs better with non-ideal, noisy, real-world datasets.
Grushka-Cockayne et al. (2017)	Demonstrates that machine learning is prone to overfitting and steps to reduce overfitting. These steps are part of this study.
Chari et al. (2008)	Not on a management topic. This is a research note alerting researchers on the problem of multicollinearity in statistical analysis and some tools to detect them.
Metcalf et al. (2019)	Research on "human-in-the-loop". This is the most cutting-edge research that showcases how machine learning can tap on to the expert knowledge of a few senior managers and augment its internal knowledge.
Kitchin and McArdle (2016), Tonidandel et al. (2018), and Wenzel and Van Quaquebeke (2018)	Summarize the definition of big-data and highlight how literature is sometimes misappropriating the phrase big-data.
L'Heureux et al. (2017)	A basic science article that demonstrates challenges in big-data analysis in great detail and builds a framework for any researcher interested in actual big-data research.
Kuhn and Johnson (2013)	One of the most authoritative methodological textbooks in machine learning.

superior (Jiang et al., 2008; Lewis, 1998). Bayesian regressions are not practical for very large datasets. However, researchers can compress large datasets using techniques such as SMOTE (Chawla et al., 2002). Organizational researchers have compared multiple machine learning models against each other to demonstrate "inter-model agreement" (Sajjadiani et al., 2019, p. 9). However, all the machine learning models may agree on a poor solution. Hence, a non-machine learning baseline model is essential.

Regression-based methods rely on null hypothesis significance testing, which provides p-values. In contrast, machine learning methods are rooted in Bayesian statistics. The parameter to be estimated is a random variable. Hence, the results of Bayesian analyses are in the form of whether X is more likely than Y or vice-versa.

In Table 3, we summarize some methodological and empirical literature in management that uses machine learning. Since this paper is more oriented to provide a technical tutorial that guides researchers in implementing machine learning, we are not going to elaborate on this line of literature in detail. We recommended that interested readers refer to these articles to gain more insights when needed.

DEMONSTRATION AND CASE ILLUSTRATION

This section demonstrates how to use machine learning to augment traditional analyses. In this demonstration, we have used the case of a longitudinal problem. We have attempted to encapsulate the longitudinal problem-specific steps in the data manipulation subsection and the bulk of the machine learning-specific steps in their subsection. Machine learning nearly always

requires custom software. However, with this partitioned approach, researchers can reduce the complexity of the overall machine learning software development and reuse portions of the steps for multiple projects.

Data

We use the HILDA dataset as an illustration (Wooden et al., 2002). HILDA is an annual household panel survey in Australia that started in 2001 and continues to date. HILDA sought to sample the same households across successive waves, and therefore, this database contains a large amount of demographic, sociological, and economic information about each household. However, some households dropped out, and new households were enrolled over the years. IO psychologists recently used HILDA to identify the effect of chronic job insecurity on employees' Big Five personality traits (Wu et al., 2020), but the personality data were collected at three time-points only: 2005, 2009, and 2013.

Our illustration uses 13 years of data from 2001 until 2013. Table 4 provides the total number of participants in each survey year and the number of items per year included in the analysis. Figure 1 indicates the number of waves that each participant was observed.

Overall, 19,664 individuals had non-missing values across the Big Five items in at least one of the three years. Of these, at least one item from all Big Five scales was recorded for 7289 individuals in all three years, for 3687 in two of the three years, and 8688 in only one year, leading to 37,929 observations. In contrast, Wu et al. (2020) focused only on 1046 cases that completed the personality measures in all three years; they only included participants who had complete responses to all the variables in their research model, provided complete gender and age demographic data, and were employees across all three waves. This means that a majority of the potentially usable data was dropped. Unlike regressions, several machine learning algorithms,

TABLE 4 Descriptive information about the HILDA survey

Year	Number of participants	Number of survey items
2001	19,914	3400
2002	18,295	4488
2003	17,690	4479
2004	17,209	4277
2005[†]	17,467	4712
2006	17,453	4977
2007	17,280	4894
2008	17,144	5017
2009[†]	17,632	5110
2010	17,855	5377
2011	23,415	5572
2012	23,182	5429
2013[a]	23,299	5354

[a]Big Five data were collected.

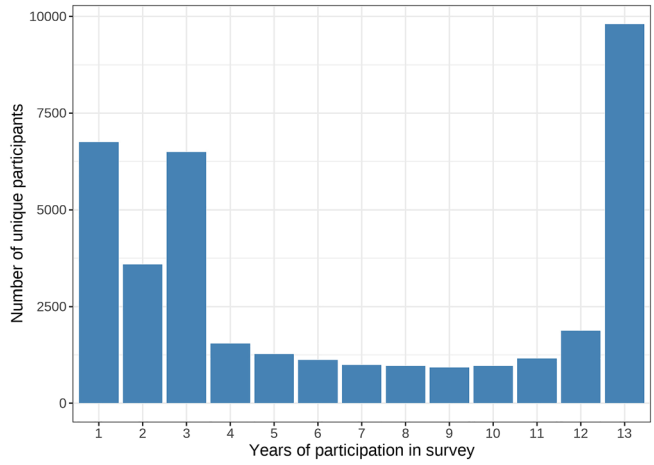

FIGURE 1 Number of years unique individual represented in the HILDA dataset [Color figure can be viewed at wileyonlinelibrary.com]

including XGBoost, can handle responses with missing values, allowing us to retain the maximum number of participants possible.

Unlike Wu et al. (2020), we do not restrict the data to active employees; instead, we retained those who were employed full-time (employment status = 1), employed part-time (employment status = 2), unemployed but looking for a full-time position (employment status = 3) and unemployed but looking for a part-time position (employment status = 4). We did so because we believe the job-related variables are relevant to all these participants.

HILDA administered 28 personality items to measure the Big-Five personality traits; we computed scale scores by taking the mean of the relevant items used to measure each of the five personality traits (alpha = .74–.80). As potential predictors, we included a total of 26 variables: gender, age, employment status (four categorical response options), job satisfaction (one item), job insecurity (four items), job stress (two items), job control (six items), job repetitiveness (two items), job complexity (three items), job initiative (one item), time demand (three items), and perceptions of fair pay (one item). In addition, we one-hot coded employment status. One-hot coding is similar to dummy coding, except one-hot coding creates an indicator variable for all response options. In contrast, dummy coding does not create indicator variables for the reference category. Consistent with our earlier argument, other than the dependent variable, we retained individual items as predictors rather than averaging items to form scales. Finally, we excluded cases if all independent variables were missing; a dependent variable cannot be predicted without an independent variable. This procedure resulted in a sample of 35,852 cases (observations) across 19,013 individuals.

Note that our goal is not to replicate the analysis of Wu et al. (2020) using machine learning; we use Wu et al. (2020) as an inspiration to demonstrate how machine learning methods can be used to analyze longitudinal data. Whereas Wu et al. (2020) focused on changes in personality traits using latent growth modeling based on four job-related variables in a small subset of the overall sample, we test whether there is a relationship between individuals' past job variables and their current personality by drawing from the largest usable sample. For illustration, we focus on neuroticism only because Wu et al. (2020) found the strongest effects on neuroticism; however, analogous models can be used to analyze the other four personality traits in the same way.

Exploratory data analysis

In this section, we assess whether the dataset we constructed above meets the assumptions of regression-based methods. For example, we assess whether the dependent variables are normally distributed, the presence of multicollinearity, homoscedasticity, and whether the data are drawn from a single homogeneous sample. In essence, we conducted exploratory data analysis (Behrens, 1997) following the procedure Wickham and Grolemund (2016) outlined.

Figure 2 shows the distribution of Big Five personality traits in the sample. We find that the distribution of neuroticism is right-skewed, such that most people rate low on neuroticism and very few people rate high on neuroticism. An Anderson-Darling test of normality confirmed that the distribution is indeed non-normal, A = 149.27, $p < 2.2E{-}16$. Thus, a key assumption of regression-based models is violated.

Next, we assess whether two other assumptions of regression-based models are met: lack of multicollinearity and homoscedasticity. In the presence of multicollinearity, "the variances of the estimates may be so large as to cast into doubt all our results" (Rockwell, 1975, p. 309). We used the mctest package in R (Imdadullah et al., 2016), which performs six tests for multicollinearity; the results are reported in Table 5. Five of the six tests indicated multicollinearity in the data, which means that regression-based models are inappropriate. Homoscedasticity assumes that the residual error term is randomly distributed with reference to the independent variables. Heteroscedasticity artificially lowers the *p*-values, thereby increasing Type I error (Astivia & Zumbo, 2019). We used the Breush–Pagan test to detect heteroscedasticity using the lmtest package in R (Hothorn et al., 2015), which indicated that the homoscedasticity assumption had been violated, BP = 347.48, $df = 302$, $p = 0.037$. Given the violation of two key assumptions, we cannot use regression-based methods to analyze the HILDA dataset.

Next, we tested whether the data are drawn from a single, homogeneous population by examining the proportion of observations that are outliers in an N-dimensional space spanning all the variables. Typically, IO psychologists assess outliers on just the dependent variable; however, it is advisable to confirm whether there are outliers on predictor variables to identify whether participants are drawn from a single homogeneous population. This is particularly a

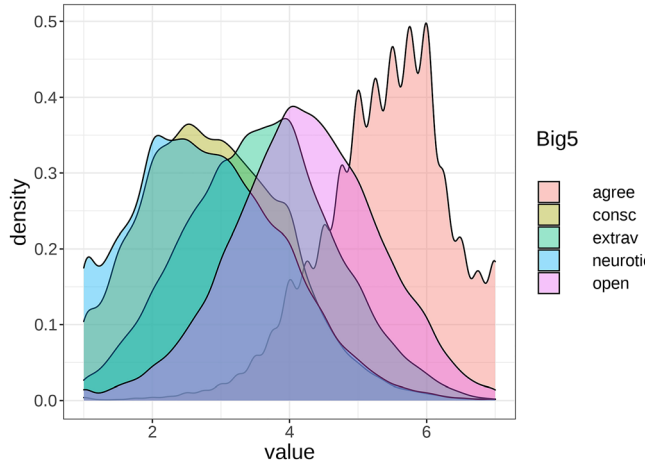

FIGURE 2 Histogram of the Big Five personality traits in the HILDA dataset [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Multicollinearity diagnostic statistics

	MC results	Detection
Determinant |X'X|	0.0000	1
Farrar chi-square	378045.7288	1
Red Indicator	0.1966	0
Sum of lambda inverse	3097.5969	1
Theil's method	94.7925	1
Condition number	1416.7249	1

Note: 1 indicates that collinearity is detected by the test; 0 otherwise.

concern in longitudinal datasets because the underlying population might qualitatively change over time, or external shocks might cause a permanent shift in the underlying psychological relationships (Tsay et al., 2000). To identify multivariate outliers, we use the performance package in R (Lüdecke et al., 2021), which simultaneously implements eight outlier detection algorithms: (1) z-scores (Iglewicz & Hoaglin, 1993), (2) interquartile range Mahalanobis distance (Cabana et al., 2021), (3) robust Mahalanobis distance (Gnanadesikan & Kettenring, 1972), (4) minimum covariance determinant (Leys et al., 2018); (5) invariant coordinate selection (Archimbaud et al., 2018), (6) ordering points to identify the clustering structure (Ankerst et al., 1999), (7) isolation forest, (Liu et al., 2008), and (8) local outlier factor (Breunig et al., 2000). Overall, 16,312 participants, or about 45% of the sample, were classified as outliers by a majority of these algorithms, which means that there is substantial heterogeneity in the data. In such cases, according to Vicari and Vichi (2013), hierarchical models must be used. Given that we do not know the relevant hierarchical structures in this case, using regression-based models would be inappropriate. However, many machine learning models make no assumptions about normal distribution, multicollinearity, or homoscedasticity; they also automatically try to learn the patterns across different subpopulations and can thus handle heterogeneous data.

Additionally, as we mentioned earlier, regression-based approaches have clear drawbacks when it comes to missing values, such as deleting any observation (e.g. a participant) that has even a single missing value. Although multiple imputation methods may be used to replace data missing at random, the reality is that data are rarely missing randomly in longitudinal datasets. For example, 96.60% of the HILDA observations had at least one missing value in our data frame. Not surprisingly, although 7289 HILDA respondents completed personality trait measures in all three years, Wu et al. (2020) only used 1046 valid samples. It is unlikely that these 1046 respondents are representative samples of the 19,664 respondents who completed at least one personality measure. Thus, dropping observations with missing values compromises generalizability. We used the MissMech package in R (Jamshidian et al., 2014) to test the hypothesis that the data are missing completely at random in our HILDA data frame, which was rejected. When data are not missing at random, excluding observations with missing values "may produce significant distortion in estimating" outcomes (Frankel et al., 2012, p. 1). We may observe different results when systematically missing values when analyzing the full dataset versus a small subset without any missing values. The machine learning algorithm that we used, XGBoost, analyzes all observations, including observations with partial data, thereby ensuring that the results can be generalized to the overall population.

To summarize, in this section, we have showed that our sample violated multiple assumptions that need to be met before we can use regression-based models and that the sample has values missing not at random. Fortunately, the machine learning method we use does not make any of these assumptions that apply to regression-based approaches and can learn the underlying patterns in the data even when data are not missing randomly.

Following the exploratory data analysis, we proceed to build machine learning models. Figure 3 provides a flowchart illustrating our methodological procedure of machine learning analysis. Annotated code for the model can be found in Appendix S1.

Method

To choose a machine learning model, we refer to Table 2. Many of the assumptions listed in the table are likely violated in the HILDA dataset. Therefore, we narrowed our choice to the three assumption-free models. As HILDA has a large proportion of missing values, we are left with gradient boosting, which can handle missing values (e.g. Sheetal & Savani, 2021). We used the XGBoost implementation of gradient boosting (Chen & Guestrin, 2016), which constructs boosted decision trees around the missing values and does not require missing values to be imputed. XGBoost can also utilize video graphics cards and thus completes the computations in significantly less time than other implementations of gradient boosting. XGBoost is currently the method of choice in most machine learning competitions (Nielsen, 2016).

The project was performed on an 18-core (36-thread) Intel(R) Xeon(R) W-2195 2.30 GHz CPU with 128 GB RAM and an Nvidia RTX 3090 graphics card. This graphics card has 10,496 cores, which allows for parallel operations. The graphics card interfaced with the machine learning software using NVidia's Cuda 11.0 drivers and libraries. We used a version of XGBoost implemented to run on the graphics card (GPU). The standard installation of XGBoost performs the matrix multiplications on the CPU by default. As we wished to perform the matrix multiplications on the GPU instead, we custom-installed the source code and linked it with the graphics card (see XGBoost Developers, 2022).

Machine learning models built explicitly for time-series analyses seek to predict a single variable by analyzing its history (e.g. stock prices and other economic variables for which there is data for many periods) (Ahmed et al., 2010). However, there are only a few data points in most applied psychology longitudinal studies. Moreover, researchers are not typically interested in predicting the value of the dependent variable in the next period but in modeling the relationship between predictors and the dependent variable over time. Thus, instead of using a particular machine learning model built explicitly for time series data (e.g. recurrent neural networks), we use a generic algorithm and manually model the concept of time.

Using the HILDA data, we constructed a data frame with all participant-year combinations with non-missing values for all Big Five personality traits and at least one non-missing predictor variable, which resulted in a total of 35,852 rows. This restructuring process allows us to treat each value of the dependent variable as an independent outcome that is not autocorrelated with the prior period's dependent variable from the same participant. HILDA started in 2001 and personality was measured for some participants in 2013. As we had prior data for the same individual for a maximum of 12 years, we included 13 sets of columns representing the current year and 12 prior years. As HILDA did not have any data prior to 2001, all columns referring to variables prior to 2001 were blank; however, we needed to include these in the data frame for consistency. This meant that for dependent variables measured in 2005/2009/2013, we had 4/8/12

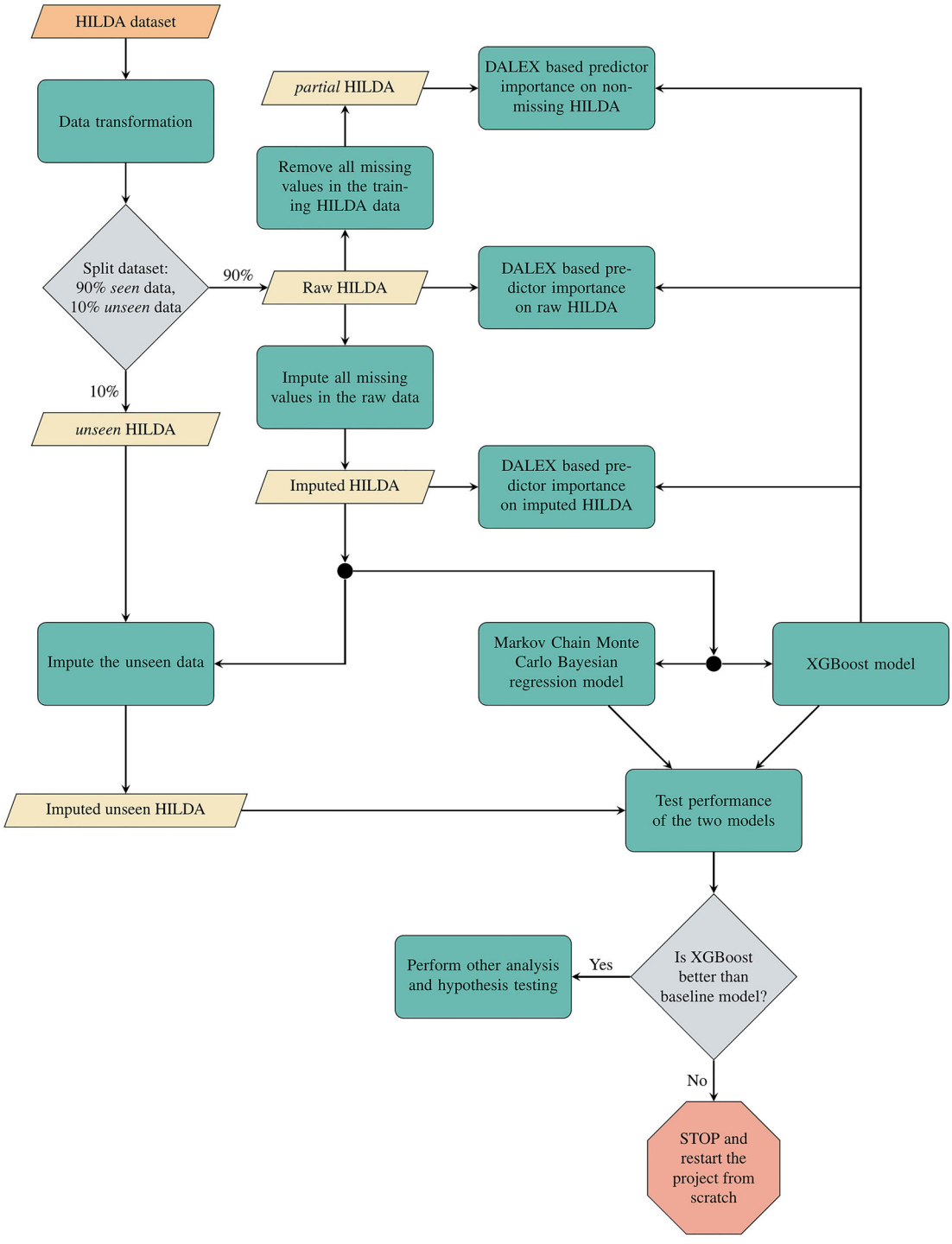

FIGURE 3 A flowchart illustrating the process of the machine learning analysis [Color figure can be viewed at wileyonlinelibrary.com]

prior years of independent variables with non-missing values, respectively; the columns representing the remaining 8/4/0 years, respectively, contained only missing values. For each of the 13 years in the data frame, we created 26 variables to represent the 26 predictor variables (i.e. all job-related items included in HILDA), which yielded 338 columns. We then deleted all blank or constant columns, which yielded a final set of 302 columns. The goal of the model was to learn to predict individuals' neuroticism in a given year based on the individual's job-related variables in the prior 12 years (whenever available).

As machine learning models tend to overfit the data they are exposed to, researchers typically undertake cross-validation; that is, the model is trained (or developed) on one sample (typically 80% or 90% of all the data) and tested on the other sample (the remaining 20% or 10%), and then these two samples are reshuffled. This reshuffling is needed because the model tries to improve its predictions across successive iterations. This process continues until the model has reached an asymptote in terms of accuracy in the validation data. However, suppose the model used the same training and validation data split across all iterations. In that case, the patterns it picks up might be specific to this training data and this validation data. To avoid this problem, the training-validation split is reshuffled after every iteration. Nevertheless, cross-validation can still overfit the data because the model has been exposed to the entire dataset across several iterations (Ng, 1997). Therefore, machine learning models should be tested on *out-of-sample*, *out-of-bag*, or *unseen* data, that is, data not previously used for training or validating the model at any stage (Kuhn & Johnson, 2013, Chapter 4) (see also Montgomery & Drake, 1991). We set aside a randomly selected 10% of the individuals as the *unseen data*; the remaining 90% of the individuals were designated as the *training data* on which the model was built. Figure 4 illustrates the various data partitions.

We earlier said that the performance of our XGBoost model should be compared against that of a Bayesian regression; however, Bayesian regressions cannot work with missing data. We first imputed the training data using a multiple-imputation method based on the random forest algorithm, using the package in R (Mayer, 2019). After we imputed the training data, we appended the unseen data to the imputed trained data and imputed missing values in the unseen data. This way, no information leaked from the unseen data to the training data. Just

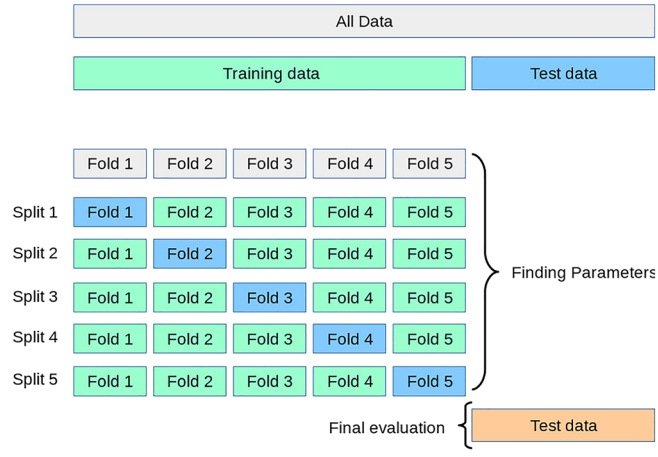

FIGURE 4 Illustration of how the data is partitioned for the machine learning analysis (Scikit-learning developers, 2022) [Color figure can be viewed at wileyonlinelibrary.com]

the information about the distribution of the predictors passed from the training data to the unseen data. To create the baseline model, we used the Bayesian regression package in R (Stan Development Team, 2018). We ran 3000 iterations on six chains to predict neuroticism using all predictor variables in the prior 12 years; this analysis was performed on the 90% seen data. Once the Bayesian simulation was complete, we tested it on the unseen data.

To ensure comparability with the Bayesian regression, we built the XGBoost model on the imputed HILDA training data (although XGBoost could also be built on the unimputed data). Researchers need to tune a number of XGBoost parameters for it to minimize the mean square error (MSE). We focused on five parameters, which resulted in a five-dimensional search space. We set the range for each parameter based on experience with similar datasets because it is computationally expensive to exhaustively test every combination of these five parameter values. We used the parameter search package (Bischl et al., 2017), which sought to optimize this search procedure. Table 6 provides the pre-set range of all parameters and the final parameter values.

Each parameter combination was tested using 10-fold cross-validation repeated five times. In other words, the model split the training data into 10 random subsets. Then, using the set of parameters provided, it built a model in nine subsets and used the 10th subset to validate the model. This loop was repeated 10 times such that each subset was used once as the validation data. The 10 subsets were then reshuffled, and this whole process was repeated five times. The average MSE from this procedure indexed the accuracy of that particular parameter combination.

Once the hyperparameter search process was completed, the XGBoost model was frozen. We then fed the imputed unseen data to the trained Bayesian regression and the final XGBoost models. Next, we removed the dependent variable, neuroticism, and asked the two models to predict neuroticism in the unseen data. We then compared the correlation between the actual neuroticism score and each model's predicted neuroticism score.

Once a machine learning model is built, the next step is to query the model to understand why the model is making the predictions that it is making. We used the package DALEX (Biecek, 2018) to identify which predictor variables have the biggest total effect on the dependent variable. DALEX randomly permutes each predictor (or group of predictors) one at a time and assesses its impact on the model's MSE. Based on Bayesian statistics, DALEX does not indicate whether individual predictors are statistically significant; however, DALEX rank orders predictors in terms of their importance and provides confidence intervals. For scales with more than one item, we used the grouped feature importance function from DALEX, which simultaneously shuffles the values of all items belonging to the scale. We conducted this analysis on three different versions of the seen dataset—(1) the imputed seen dataset on which the model

TABLE 6 Hyperparameters used in XGBoost and their final chosen value

Parameter	Type	Minimum	Maximum	Final chosen
max_depth	Integer	3	23	16
eta	Numeric	0.000001	0.999999	0.0427
min_child_weight	Numeric	0.0	2.0	1.8297
gamma	Numeric	0.0	20.0	0.00012
nrounds	Integer	500	2000	1882

was built (32,253 observations), (2) a version of the seen dataset with only complete observations (so no imputation was needed; 1149 observations), (3) and the unimputed seen data with missing values (i.e. the untouched dataset, 32,253 observations). We did so to assess how altering the dataset can change the contribution of the predictor variables. Of the three datasets, the final version is unaltered—we retained as many of the observations as possible and did not impute any missing values.

Results

The Bayesian regression on the seen data took 8 h to run on the CPU. We then presented the model with the imputed unseen data. Finally, we asked it to predict individuals' neuroticism based on all the independent variables. The Bayesian regression outputs three numbers for each observation—the median estimate of the dependent variables and the upper and lower bound of the 95% confidence interval. The correlation between the actual neuroticism and the median predicted neuroticism was $r = .56$ (see Figure 5).

The XGBoost model on the seen data took 36 hours to run on the GPU. We then presented the model with the imputed unseen data and asked it to predict individuals' neuroticism. The XGBoost model made a single best prediction for each observation. The correlation between the actual neuroticism and the predicted neuroticism was $r = .75$ (see Figure 6). Thus, the XGBoost model could predict people's neuroticism with very high accuracy based on job-related variables and substantially exceeds the accuracy of the baseline Bayesian regression.

Next, we assessed the relative contribution of the predictor variables to the XGBoost model's prediction in three different versions of the seen dataset. The length of each bar indicates the

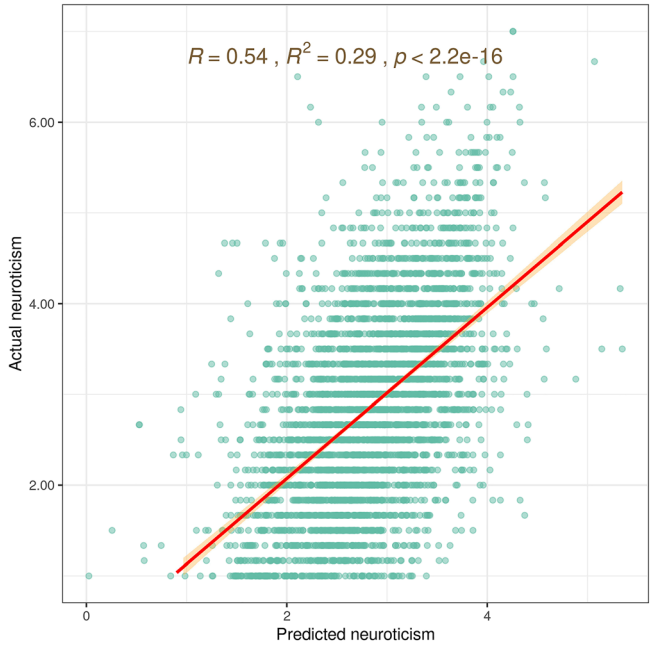

FIGURE 5 Correlation between the actual neuroticism and the predicted neuroticism per the Bayesian regression in the unseen data [Color figure can be viewed at wileyonlinelibrary.com]

change in MSE when each variable (or group of variables for multi-item scales) is shuffled one at a time. The error bars indicate the range of the MSE across different iterations. Note that the error bars are the smallest in the large imputed data. The error bars are larger both in the small data without missing values and in the large data with missing values.

The top psychological predictors of neuroticism were job insecurity, job stress, and job repetitiveness in all three datasets. However, in the untouched dataset (i.e. no imputation, observations with partial data retained), employment status made the biggest contribution to predicting neuroticism (Figure 7), more so than job insecurity. However, in the shrunk dataset with only complete observations (Figure 8) and in the imputed dataset (Figure 9), employment status mattered much less. This finding suggests the importance of observations with missing data—throwing them away can alter the pattern of results. In the current case, the XGBoost model was trained on imputed data because we wanted to compare it with the Bayesian regression. However, the variable importance can still be computed on the raw unimputed data.

DISCUSSION

This article illustrates how work and organizational psychologists can use machine learning methods to analyze longitudinal data. Although regression-based methods make numerous assumptions, most researchers do not verify whether their data actually meets the necessary assumptions. However, simply ignoring a problem does not make it go away—researchers are obligated to verify whether their data meet the assumptions before running a model. We have demonstrated that researchers can fruitfully analyze the data using machine learning even in cases where most regression-based assumptions are violated.

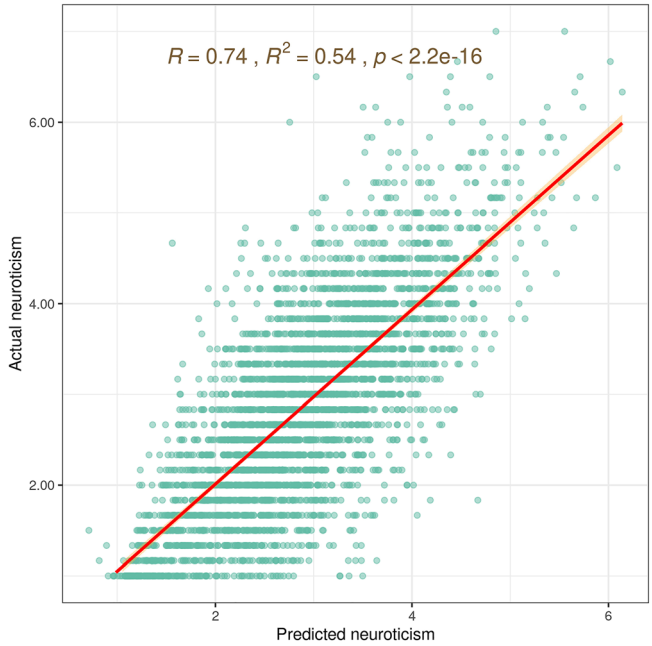

FIGURE 6 Correlation between the actual neuroticism and the predicted neuroticism per the XGBoost model in the unseen data [Color figure can be viewed at wileyonlinelibrary.com]

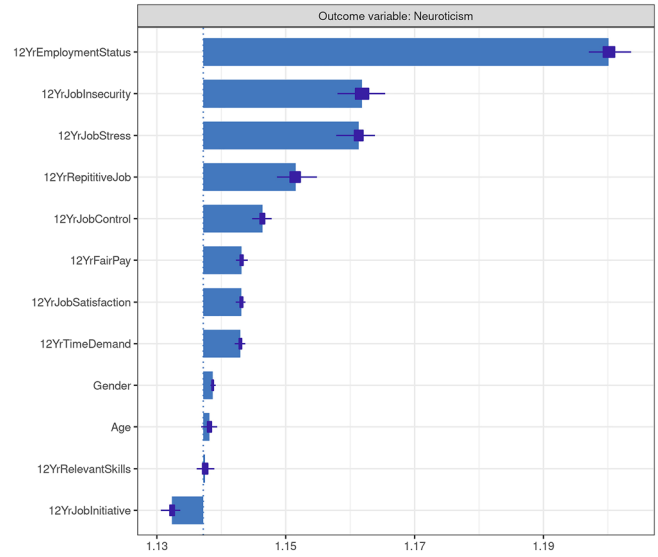

FIGURE 7 Rank ordering of the predictor variables in terms of their contribution to predicting job performance in the unimputed seen data (32,253 observations). *Note*: The X-axis indicates the changes in MSE when each variable or group of variables is shuffled. The error bars indicate the range of the MSE across different iterations. [Color figure can be viewed at wileyonlinelibrary.com]

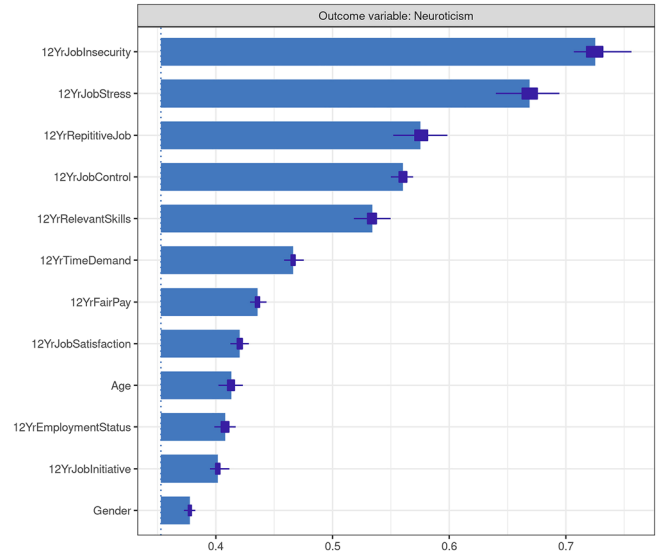

FIGURE 8 Rank ordering of the predictor variables in terms of contributions to predicting neuroticism in the seen data without missing values (1149 observations). *Note*: The X-axis indicates the changes in MSE when each variable or group of variables is shuffled. The error bars indicate the range of the MSE across different iterations. [Color figure can be viewed at wileyonlinelibrary.com]

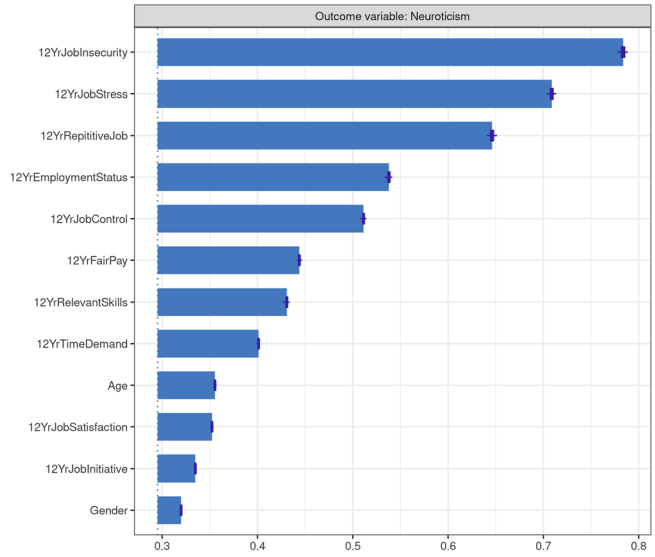

FIGURE 9 Rank ordering of the predictor variables in terms of contributions to predicting neuroticism in the imputed seen data (32,253 observations). *Note*: The X-axis indicates the changes in MSE when each variable or group of variables is shuffled. The error bars indicate the range of the MSE across different iterations. [Color figure can be viewed at wileyonlinelibrary.com]

Our XGBoost model could predict people's neuroticism based on their job-related variables over the past 12 years with very high accuracy, $r = .75$. This statistic was obtained from the 'unseen' data to which the model was never exposed. This high accuracy attests to the power of machine learning methods in picking up underlying patterns. In contrast, researchers do not test their model on new data when using traditional methods. As a result, they typically do not report the accuracy of the model's prediction.

Researchers typically delete observations with missing values because regression-based methods would automatically drop observations with even a single missing value. However, in the current case, the sample size drops by about 96% if we exclude observations with missing values, which is clearly a huge amount of data to drop. One solution is to impute missing values using a machine learning multiple imputation algorithm, such as missRanger. Another option is to use a machine learning model that can automatically handle missing values, such as XGBoost.

The standard method of calculating a scale score risks discarding useful information at the item level. The advantage of machine learning methods is that they can handle collinearity. Furthermore, there is no need to sacrifice interpretability with the grouped variable importance that we demonstrated above. If researchers do not care about individual items but only the overarching construct, then they can group the items belonging to a given scale when assessing the relative importance of different predictors.

Building a machine learning model with hyperparameters, such as XGBoost, requires some intelligent guesswork in terms of the possible range of each hyperparameter. As an initial guess, researchers can use the range we used in this paper. However, when the hyperparameter search process is running, researchers should pay attention to the loss value that would be displayed on the screen after each iteration. If the loss value is generally decreasing toward an asymptote,

then it means that the hyperparameter ranges are appropriate. On the other hand, if the loss values are haphazardly jumping around, then researchers need to either adjust the hyperparameter ranges or adjust some of the fixed parameters. Unfortunately, again, there are no rules about how to identify the appropriate hyperparameter range—the only solution is through trial and error.

While we believe machine learning provides several benefits, others have argued that traditional regression-based methods are superior when analyzing psychological data (Jacobucci & Grimm, 2020). However, these results were based upon simulated data, which met the assumptions required for regression models by design. Our argument is that most real-world data violate multiple regression assumptions (Erceg-Hurn & Mirosevich, 2008), and thus machine learning is a safer approach than regression to establish relationships among variables. While there are claims that traditional methods may still work better than machine learning even for real-world datasets (Nusinovici et al., 2020), researchers making these claims used one method with one set of hyperparameters. Nonetheless, as illustrated in this article, there are dozens of machine learning models, each with an infinite number of parameter combinations. We are not seeking to divide the literature to use one tool or another. The goal of our article was to highlight the advantages of machine learning when drawing on large longitudinal samples—that is, data that does not meet the assumptions of parametric statistics, data that has missing data not at random, and data that is not equally collected across data waves.

In addition, we would like to clearly highlight two of the drawbacks of machine learning that researchers must be aware of that are not commonly discussed in other machine learning articles in management and organizational sciences. Firstly, Solomonoff induction is an unsolvable problem (Solomonoff, 1967). Machine learning uses approximation techniques and even though it builds a set of plausible rules from the data, because of the unsolvable nature of the problem one cannot prove those rules are final. Machine learning is merely performing a Bayesian update of a priori rules using the provided data. New data and more sophisticated tools could upend past rules. Moreover, a non-peer-reviewed method to build a model could easily lead to non-generalizability of the generated rules. Secondly, machine learning is still struggling with the topic of causality (Schölkopf et al., 2021). All inferences from machine learning are strictly correlational. Both these drawbacks do not exist in traditional OLS methods. Hence in addition to highlighting the benefits of machine learning, we conclude with a general guidance that researchers must follow which using machine learning.

BEST-PRACTICE RECOMMENDATIONS FOR MACHINE LEARNING ANALYSIS OF LONGITUDINAL DATA

This section provides a less technical summary of how to proceed with machine learning based research and reemphasizes some points mentioned earlier. This non-technical note should be used by both researchers as well as readers who may not be interested in the details of the machine learning method but can quickly use these dos and don'ts to judge the quality of a machine learning article. We recommend that the following points be paid particular attention to.

First, given the correlational nature of math involved in machine learning, researchers cannot make causal claims from machine learning. Researchers in computer science and mathematics are still debating how to infer causality from machine learning. Until that debate is settled, inferences of causality must be supported by a follow-up study based on traditional methodologies. We recommend that wherever possible researchers use experiments to generate additional evidence, which can be used in conjunction with results of machine learning to

facilitate causal inferences. Some recent studies (e.g. Sheetal et al., 2020, 2022) have set examples researchers can refer to when designing and implementing machine learning projects that involve casual inference.

Second, machine learning findings are very informative but may also be argued to be merely suggestive. Machine learning is searching for patterns, and such search in a dataset can be an extremely hard computational problem. Mathematical advances in algorithms could produce better search results in future. Similarly, a stronger computer in future will allow researchers to widen the search parameters while keeping the search time under control. In addition, similar to traditional statistical research, data from new demographic groups that were not represented in previously collected data could alter the patterns that were produced by previous machine learning models. Hence, researchers should be aware of these aspects when making theoretical claims and generalizations in their articles.

Third, machine learning requires complex programming. A typical machine learning software program ranges 1000–3000 lines of codes (e.g. in R or Python). Machine learning programming is similar as that of a surgeon in the operating room. There are best practices, but the surgeon has the right to deviate from the best practices based on the patient's condition. There will be more exceptions than rules. Having said that, it is now established that one simple test/check is enough and that supersedes all best practices. The phrase that the "proof of the pudding is in the eating" (Cabitza & Zeitoun, 2019) summarizes the catch all for several best programming practices in machine learning. If the pudding tastes well in the test, then the method is likely to be just fine. This is similar to that of the patient who has recovered after the surgery, then most likely the surgeon's methods were sound. In practice, if the machine learning model performs similarly well in a secondary dataset that it was unexposed to before, then most likely the method followed in the model development is sound; a practice known as unseen testing. For the study illustrated in this paper, we had set aside 10% of the participants as unseen. And we tested our model against these 10% of the new participants (See Figures 5 and 6). Readers, reviewers, and editors should be cautious of articles that do not report unseen testing on a secondary dataset and may reasonably cast doubt on the results (e.g. possibly overfitted results with lower generalizability). This proof of the pudding test is a standard procedure recommended by the creators of modern machine learning tools (Kuhn & Johnson, 2013).

Finally, because of the unprovability nature of machine learning (Reyzin, 2019), one cannot just blindly trust any machine learning model without a reference point. Accuracy numbers of machine learning models from the proof of pudding test above are still not believable if the taster lacks a comparison point. In machine learning that is referred to as a baseline model, researchers must also report a non-machine learning based baseline model in the proof of the pudding test. Alternatively, researchers can use performance numbers from past research as a baseline point to compare. In this paper, we reported results of a Bayesian model for the illustrative study. Setting a traditional model as a baseline is important to help assess the performance of machine learning models. Researchers must not stick to under-fitting or underperforming machine learning models. In articles without reporting such a traditional baseline model, there might not be a way to rule out the possibility that their machine learning model is under-fitting or underperforming.

ACKNOWLEDGEMENTS

Open access publishing facilitated by RMIT University, as part of the Wiley - RMIT University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

We have no conflict of interest to disclose.

ETHICS STATEMENT

Data analyses in this article were conducted under the internationally recognized ethical regulations and norms.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://melbourneinstitute.unimelb.edu.au/hilda.

ORCID

Abhishek Sheetal https://orcid.org/0000-0002-4585-5358
Zhou Jiang https://orcid.org/0000-0002-6249-2659
Lee Di Milia https://orcid.org/0000-0001-7681-5589

ENDNOTE

[1] Solomonoff induction has no guarantee of convergence (Solomonoff, 1967). Hence machine learning tools based inductive reasoning is called approximate, "good enough" generalizations only to be upended by better quality data in future.

REFERENCES

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5–6), 594–621. https://doi.org/10.1080/07474938.2010.481556

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Record*, *28*(2), 49–60. https://doi.org/10.1145/304181.304187

Archimbaud, A., Nordhausen, K., & Ruiz-Gazen, A. (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, *128*, 184–199. https://doi.org/10.1016/j.csda.2018.06.011

Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in R and SPSS. *Practical Assessment, Research and Evaluation*, *24*(1), 1. https://doi.org/10.7275/q5xr-fr95

Behfar, K., & Okhuysen, G. A. (2018). Perspective—Discovery within validation logic: Deliberately surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry. *Organization Science*, *29*(2), 323–340. https://doi.org/10.1287/orsc.2017.1193

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160. https://psycnet.apa.org/doi/10.1037/1082-989X.2.2.131

Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. *The Journal of Machine Learning Research*, *19*(1), 3245–3249.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). Machine learning in R. *The Journal of Machine Learning Research*, *17*(1), 5938–5942.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. *arXiv preprint arXiv:1703.03373*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, *29*(2), 93–104. https://doi.org/10.1145/342009.335388

Brodley, C. E., & Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning*, *19*(1), 45–77. https://doi.org/10.1007/BF00994660

Burks, A. W. (1946). Peirce's theory of abduction. *Philosophy of Science*, *13*(4), 301–306. https://doi.org/10.1086/286904

Cabana, E., Lillo, R. E., & Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, *62*(4), 1583–1609. https://doi.org/10.1007/s00362-019-01148-1

Cabitza, F., & Zeitoun, J. D. (2019). The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence. *Annals of Translational Medicine*, *7*(8), 1–9. https://doi.org/10.21037/atm.2019.04.07

Cai, Z., Gao, Z. J., Luo, S., Perez, L. L., Vagena, Z., & Jermaine, C. (2014). A comparison of platforms for implementing and running very large scale machine learning algorithms. *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 1371–1382. https://doi.org/10.1145/2588555.2593680

Chan, D. (2004). Longitudinal modeling. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 412–430). Blackwell Publishing.

Chari, M. D., Devaraj, S., & David, P. (2008). Research note—The impact of information technology investments and diversification strategies on firm performance. *Management Science*, *54*(1), 224–234. https://doi.org/10.1287/mnsc.1070.0743

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57. https://doi.org/10.1002/smj.3215

Cusumano, M. A. (1989). The software factory: A historical interpretation. *IEEE Software*, *6*(2), 23–30. https://doi.org/10.1109/MS.1989.1430446

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Ezekiel, M. (1925). The assumptions implied in the multiple regression equation. *Journal of the American Statistical Association*, *20*(151), 405–408. https://doi.org/10.1080/01621459.1925.10503505

Frankel, M. R., Battaglia, M. P., Balluz, L., & Strine, T. (2012). When data are not missing at random: Implications for measuring health conditions in the behavioral risk factor surveillance system. *BMJ Open*, *2*(4), e000696. https://doi.org/10.1136/bmjopen-2011-000696

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: With applications in R*. Springer.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124. https://doi.org/10.2307/2528963

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl, K. C. Jr. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130. https://doi.org/10.1287/mnsc.2015.2389

Hay, E. N. (1950). Cross-validation of clerical aptitude tests. *Journal of Applied Psychology*, *34*(3), 153–158. https://doi.org/10.1037/h0053979

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., Mitchell, D., & Zeileis, M. (2015). Package 'Lmtest': Testing linear regression models. CRAN. https://cran.r-project.org/web/packages/lmtest/lmtest.pdf

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Asq Press.

Imdadullah, M., Aslam, M., & Altaf, S. (2016). Mctest: An R package for detection of collinearity among regressors. *The R Journal*, *8*(2), 495–505. https://doi.org/10.32614/RJ-2016-062

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *Journal of Statistical Software*, *56*, 1–31. https://doi.org/10.18637/jss.v056.i06

Jaya, I. G. N. M., Tantular, B., & Andriyana, Y. (2019). A Bayesian approach on multicollinearity problem with an informative prior. *Journal of Physics: Conference Series*, *1265*(1), 012021. https://doi.org/10.1088/1742-6596/1265/1/012021

Jiang, L., Zhang, H., & Cai, Z. (2008). A novel Bayes model: Hidden naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, *21*(10), 1361–1371. https://doi.org/10.1109/TKDE.2008.234

Kelloway, E. K., & Francis, L. (2012). Longitudinal research and data analysis. In R. R. Sinclair, M. Wang, & L. E. Tetrick (Eds.), *Research methods in occupational health psychology* (pp. 398–418). Routledge.

Kitchenham, B., & Pfleeger, S. L. (2002). Principles of survey research: Part 5: Populations and samples. *ACM SIGSOFT Software Engineering Notes*, *27*(5), 17–20. https://doi.org/10.1145/571681.571686

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1–10. https://doi.org/10.1177/2053951716631130

Kuhn, M. (2019). 6 available models: The caret package. Github. https://topepo.github.io/caret/available-models.html

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. CRAN. https://cran.r-project.org/web/packages/tidymodels/citation.html

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4–15). Springer. https://link.springer.com/content/pdf/10.1007/bfb0026666.pdf

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *Ieee Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE.

Liu, Y., Mo, S., Song, Y., & Wang, M. (2016). Longitudinal analysis in occupational health psychology: A review and tutorial of three longitudinal modeling techniques. *Applied Psychology*, *65*(2), 379–411. https://doi.org/10.1111/apps.12055

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Mack, Y. P., & Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, *9*(1), 1–15. https://doi.org/10.1016/0047-259X(79)90065-4

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 1–14. https://doi.org/10.1177/25152459221080396

Mayer, M. (2019). missRanger: Fast imputation of missing values. CRAN. https://cran.hafro.is/web/packages/missRanger/missRanger.pdf

Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping humans in the loop: Pooling knowledge through artificial swarm intelligence to improve business decision making. *California Management Review*, *61*(4), 84–109. https://doi.org/10.1177/0008125619862256

Mokhtari, K. E., Higdon, B. P., & Başar, A. (2019, November). Interpreting financial time series with SHAP values. In T. Pakfetrat, G. V. Jourdan, K. Kontogiannis, & R. Enenkel (Eds.), *Proceedings of the 29th Annual*

International Conference on Computer Science and Software Engineering (pp. 166–172). Association for Computing Machinery.

Molnar, C. (2020). Interpretable machine learning. Leanpub.

Montgomery, G. J., & Drake, K. C. (1991). Abductive reasoning networks. *Neurocomputing*, *2*(3), 97–104. https://doi.org/10.1016/0925-2312(91)90055-G

Ng, A. Y. (1997). Preventing "overfitting" of cross-validation data. *International Conference on Machine Learning*, *97*, 245–253. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.6720&rep=rep1&type=pdf

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win "every" machine learning competition? [Master's thesis, Norwegian University of Science and Technology]. NTNU Open. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-validation. In L. Liu & M. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532–538). Springer. https://doi.org/10.1007/978-1-4899-7993-3_565-2

Reyzin, L. (2019). Unprovability comes to machine learning. *Nature*, *565*, 166–167. https://doi.org/10.1038/d41586-019-00012-4

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods & Research*, *3*(3), 308–320. https://doi.org/10.1177/004912417500300304

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosett, C. M., & Hagerty, A. (2021). Why now? Computers enable a future with machine learning. In C. M. Rosett & A. Hagerty (Eds.), *Introducing HR analytics with machine learning: Empowering practitioners, psychologists, and organizations* (pp. 95–105). Springer. https://doi.org/10.1007/978-3-030-67626-1_7

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Towards causal representation learning. *arXiv preprint arXiv:2102.11107*.

Scikit-learning developers. (2022). Scikit-Learning 1.1.1. https://scikit-learn.org/stable/modules/cross_validation.html

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). The Guilford Press.

Sheetal, A., Chaudhury, S. H., & Savani, K. (2022). A deep learning model identifies emphasis on hard work as an important predictor of income inequality. *Scientific Reports*, *12*(1), 1–11. https://doi.org/10.1038/s41598-022-13902-x

Sheetal, A., Feng, Z., & Savani, K. (2020). Using machine learning to generate novel hypotheses: Increasing optimism about COVID-19 makes people less willing to justify unethical behaviors. *Psychological Science*, *31*(10), 1222–1235. https://doi.org/10.1177/0956797620959594

Sheetal, A., & Savani, K. (2021). A machine learning model of cultural change: Role of prosociality, political attitudes, and Protestant work ethic. *American Psychologist*, *76*(6), 997–1012. https://doi.org/10.1037/amp0000868

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Shrestha, Y. R., He, V. F., Puranam, P., & von Krogh, G. (2021). Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science*, *32*(3), 856–880. https://doi.org/10.1287/orsc.2020.1382

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2020). Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science*, *66*(6), 2495–2522. https://doi.org/10.1287/mnsc.2019.3308

Solomonoff, R. J. (1967). *Inductive inference research: Status, Spring 1967 (RTB 154)*. Rockford Research, Inc. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.8743&rep=rep1&type=pdf

Stan Development Team. (2018). RStan: The r Interface to Stan. R Package Version 2.18.2. https://mc-stan.org
Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x
Stroustrup, B. (1988). What is object-oriented programming? *IEEE Software*, *5*(3), 10–20. https://doi.org/10.1109/52.2020
The Peirce Edition Project. (1998). *The essential Peirce: Selected philosophical writings volume 2 (1893–1913)*. Indiana University Press.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B: Methodological*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Tong, H., Chen, D., & Peng, L. (2009). Analysis of support vector machines regression. *Foundations of Computational Mathematics*, *9*(2), 243–257. https://doi.org/10.1007/s10208-008-9026-0
Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547. https://doi.org/10.1177/1094428116677299
Tsay, R. S., Pena, D., & Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, *87*(4), 789–804. https://doi.org/10.1093/biomet/87.4.789
Vicari, D., & Vichi, M. (2013). Multivariate linear regression for heterogeneous data. *Journal of Applied Statistics*, *40*(6), 1209–1230. https://doi.org/10.1080/02664763.2013.784896
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. https://doi.org/10.1177/1094428117718627
Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media Inc.
Wooden, M., Freidin, S., & Watson, N. (2002). The household, income and labour dynamics in Australia (HILDA) survey: Wave 1. *The Australian Economic Review*, *35*(3), 339–348. https://doi.org/10.1111/1467-8462.00252
Wu, C. H., Wang, Y., Parker, S. K., & Griffin, M. A. (2020). Effects of chronic job insecurity on big five personality change. *Journal of Applied Psychology*, *105*(11), 1308–1326. https://doi.org/10.1037/apl0000488
XGBoost Developers. (2022). Building from source—Xgboost 1.5.2 documentation. XGBoost. https://xgboost.readthedocs.io/en/stable/build.html

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sheetal, A., Jiang, Z., & Di Milia, L. (2023). Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology*, *72*(3), 1339–1364. https://doi.org/10.1111/apps.12435