(Ir)Reproducible Machine Learning: A Case Study

Sayash Kapoor, Arvind Narayanan
Princeton University

Date: August 2, 2021
A significantly updated version of this paper is available at reproducible.cs.princeton.edu.
Please read and cite the updated paper instead of this earlier draft.

***Abstract*— The use of Machine Learning (ML) methods for prediction and forecasting has become widespread across the quantitative sciences. However, there are many known methodological pitfalls in ML-based research. As a case study of these pitfalls, we examine the subfield of civil war prediction in Political Science. Our main finding is that several recent studies published in top Political Science journals claiming superior performance of ML models over Logistic Regression models fail to reproduce. Our results provide two reasons to be skeptical of the use of ML methods in this research area, by both questioning their usefulness and highlighting the pitfalls of applying them correctly. Results identifying pitfalls in studies that use ML methods have appeared in at least eight quantitative science fields. However, we go farther than most previous research to investigate whether the claims made in the reviewed studies survive once the errors are corrected. We argue that there is a reproducibility crisis brewing in research fields that use ML methods and discuss a few systemic interventions that could help resolve it.**

Over the last few years, there has been a marked shift towards the paradigm of predictive modeling in many quantitative science fields. Models which are better at prediction are thought to enable an improved understanding of scientific phenomena [1, 2]. This shift has been facilitated by the creation and widespread use of machine learning (ML) methods. Traditional statistical methods typically assume an underlying model for the data based on existing theory. In contrast, ML methods for prediction typically assume no underlying distribution from which the data is drawn and try to optimize predictive performance on *out-of-sample* data — data that the model was not trained on [3]. Leading scholars in numerous quantitative science fields have called for the adoption for ML methods [4–9].

However, pitfalls in using ML methods have led to exaggerated claims about their performance. Such errors can lead to a feedback loop of overoptimism about the paradigm of prediction — especially as non-replicable publications tend to be cited more often than replicable ones [10]. A slew of systematic studies have uncovered errors due to common pitfalls in published research that uses ML methods; we summarize reports from 8 fields in Table I.

These results highlight the importance of examining the reproducibility of findings in communities adopting ML methods. Here, we define a research finding as being reproducible when the code and data used to obtain the finding are available and the data is correctly analyzed [11, 12]. Note that there are many conflicting definitions of reproducibility across disciplines and there have been recent efforts to resolve these conflicts [11–15]. We situate our work in the field of applied ML research — where the goal is to use ML methods to study some scientific question. There is a much better known reproducibility crisis in research that uses traditional statistical methods [16]. There are also serious concerns about reproducibility in research that focuses on developing new ML methods [12].

Through a case study of the shift to predictive modeling, we highlight the various ways in which the use of ML methods can go awry, leading to findings that can't be reproduced when the errors are fixed. We focus on a subfield of political science that recently started emphasizing out-of-sample prediction performance: civil war research. While earlier papers in the field focused on explanation and understanding using traditional statistical methods, an influential paper by Ward et al. [17] made a strong case for using out-of-sample predictions in order to evaluate the performance of models of civil war onset. Since then, a paradigm of prediction focused on evaluating out-of-sample performance of models of civil war has emerged. While a predictive modeling approach, if performed correctly, can indeed help distill knowledge, our work offers a critique of the prediction paradigm in the field of civil war research by showcasing the irreproducibility of several peer-reviewed articles due to methodological errors.

Overview of methods and results

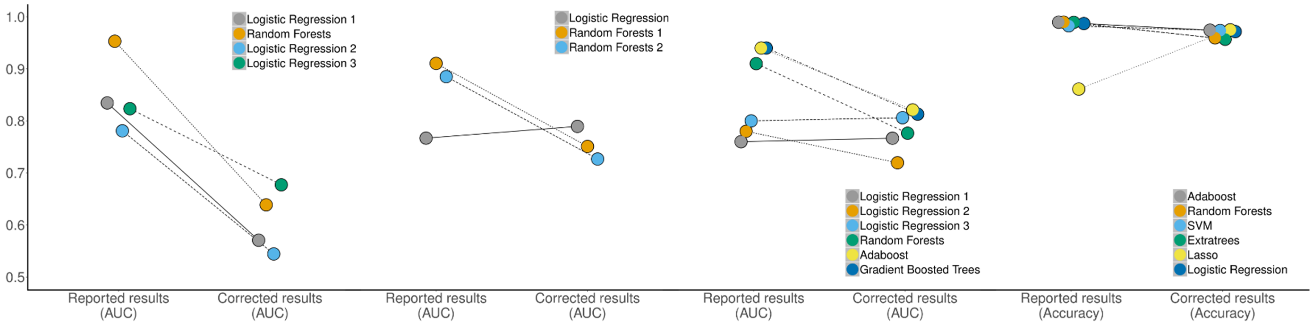
We review recent papers on civil war prediction. To find relevant papers, we used the search results from a dataset of academic literature [22] for papers with the terms *'civil' AND 'war' AND ('prediction' OR 'predicting' OR 'forecast')* in their title or abstract, as well as papers that were cited in a recent review of the field [23]. To keep the number of papers tractable, we limited ourselves to those that were published in the last 5 years, i.e. with a publication date of 2016 or later. This yielded 124 papers. We narrowed this list to the 15 papers that were focused on predicting civil war and evaluated performance using a train-test split. We share the full list of papers in the Supplement.

Out of the 15 papers that meet our inclusion criteria, 12 share the complete code and data for their results. For these 12 papers, we attempted to identify errors and reproducibility issues from the text and through reviewing the code provided with the papers. When we identified errors, we re-analyzed the data with the errors corrected. [1]

We found errors in 4 of the 12 papers — exactly the 4 papers that claimed superior performance of ML models over Logistic Regression for predicting civil war.[2] All the errors are different forms of leakage, a common type of pitfall in machine learning. None were inferrable from the text in the papers. We discovered them by reading the code provided with the papers. This highlights the importance of sharing reproduction materials with publications.

[1] When we did not identify errors or reproducibility issues from the text or the code, we did not attempt to execute the code and verify the reported results, except in one case discussed below.

[2] We use "ML models" as a shorthand for models other than Logistic Regression, specifically, Random Forests, Gradient-Boosted Trees, and AdaBoost. To be clear, all these models including Logistic Regression involve learning from data in the predictive modeling approach.



Paper	Muchlinski et al.	Colaresi and Mahmood	Wang	Kaufman et al.
Claim	Random Forests model drastically outperforms Logistic regression models	Random Forests models drastically outperform Logistic regression model	Adaboost and Gradient Boosted Trees drastically outperform other models	Adaboost outperforms other models
Error	Incorrect imputation	Incorrect reuse of an imputed dataset	Incorrect reuse of an imputed dataset, *k*-fold cross validation with temporal data	Data leakage due to proxy variables, *k*-fold cross validation with temporal data

Fig. 1: Results. A comparison of reported results vs. corrected results in the 4 papers on civil war prediction that compare the performance of ML models and Logistic Regression models. The main findings of each of these papers are invalid due to various forms of data leakage: Muchlinski et al. [18] impute the training and test data together, Colaresi and Mahmood [19] and Wang [20] incorrectly reuse an imputed dataset, and Kaufman et al. [21] use proxies for the target variable which causes data leakage. When we correct these errors, ML models do not perform substantively better than Logistic Regression models for civil war prediction in each case [18–21]. The metric for Kaufman et al. is accuracy; for all other papers, it is AUC.

When the errors are corrected, ML models perform no better than Logistic Regression models in every case except Wang [20], where the difference between the AUC of the ML models and Logistic Regression models drops from 0.14 to 0.015. All 4 of these papers were published in top-10 journals in the field of Political Science and International Relations according to SCImago Journal Rankings [24].

Note that we did not encounter issues with computational reproducibility in the narrow sense for these papers: we were able to reproduce the reported results in each case using the same code and data; the observed differences between the reported results and our reproduction using the same code and data are extremely small for 3 of the 4 papers with errors. While differences between the reported results and our reproduction for Muchlinski et al. [18] were substantial in some cases, the relative order of performance of the various models in our reproduction remained the same as that in the reported results. Details of our reproduction are included in the Supplement.

We found another reproducibility issue that is less severe but more pervasive, affecting 9 of the 12 papers: a lack of hypothesis testing or uncertainty quantification. As an illustration of why this is an important issue, we analyze one paper with a particularly small sample size — only 11 instances of civil war onset in the test dataset — and show that the claimed performance differences are not statistically significant and the reported AUC values have very large confidence intervals. Since we did not attempt to reproduce the findings of papers where we did not find errors by reading the text and code, our results represent a lower bound on the number of papers with errors.

Overall, this means that we identify reproducibility issues in 13 of the 15 papers: errors in 4 cases, incomplete code or data in 3 cases, and a lack of hypothesis testing or uncertainty quantification in 9 cases (including 3 of the 4 papers with errors).

DATA LEAKAGE IN STUDIES CLAIMING SUPERIORITY OF ML

Data leakage is a spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy. Since the spurious relationship won't be present when the model is deployed, leakage usually leads to inflated estimates of model performance. Figure 1 shows the results of our reproduction efforts for the 4 papers claiming superior performance for ML models. We find that data leakage leads to overoptimistic performance claims in each case. We now analyze the various errors present in these papers.

Imputing the training and test data together [18]: A common way to deal with missing values in the data is to use imputation methods to fill in missing values [25]. Imputing the training and test datasets together refers to using data from the training as well as the test datasets to create an imputation model that that fills in all missing values in the dataset. This is an erroneous imputation method for the predictive modeling paradigm, since it can lead to data leakage. This results in incorrect, overoptimistic performance claims.

Imputing the training and test datasets together is a well-known pitfall in the predictive modeling community — discussed in ML textbooks [26], blogs [27] and popular online forums [28]. In general, imputation methods for predictive modeling must maintain the *learn-predict* separation to get an accurate estimate of out-of-sample performance [29] — the data in the test set should never be used before the evaluation step, including in any data pre-processing step. This is a difference from explanatory modeling, where it is a standard practice to perform the imputation procedure on the entire dataset — data imputation is a data pre-processing step. Contrarily, in predictive modeling, the goal is to optimize predictive performance on a dataset that the model does not have access to during training, to evaluate how well the model generalizes when it is applied to a new dataset. Here, data imputation is a part of the modeling step [26]. If the training and out-of-sample test sets are imputed together, the imputation model makes use of data from the training set and well as the labels of the target variable from the test set to fill in the missing values in the test set, which means that correlations between the target variable and independent variables in the training set are replicated in the test set. The purportedly "out-of-sample" test set now consists of data that has similar correlations between the target and independent variables as those observed in the training data. This defeats the original purpose of using an out-of-sample dataset, which was to evaluate model performance on data that the model did not have access to during training. We demonstrate this phenomenon using a simulation in the Supplement.

Muchlinski et al. [18] claim that a Random Forests model vastly outperforms Logistic Regression models in terms of out-of-sample performance using the AUC metric [30]. They received two critiques of the methods used in their paper [20, 31]. In response, they published a reply with clarifications and revised code addressing both critiques [32]. We use the revised version of their code. We find that the error in their imputation methods exists in the revised code as well as the original code, and was not identified by the previous critiques.

The impact of leakage in Muchlinski et al.'s paper is especially severe because of the level of missingness in the test dataset which they use to evaluate the out-of-sample performance of their models. Surprisingly, over 95% of the data values in the out-of-sample test dataset are missing (which is not reported in the paper), and 70 of the 90 variables used in their model are missing for *all* instances in the out-of-sample test set.[3] All of these values are filled in using their incorrect imputation method. This means that 95% of their "out-of-sample" test set now consists of data which has similar correlations between the target and independent variables as those observed in the training data. Thus, the results reported in Muchlinski et al. do not reflect the true out-of-sample performance of their model.

In order to correct their results, we use a version of the data that does not use imputation, allowing us to impute the training and test data separately to avoid leakage. Specifically, their training set is a dataset by Hegre et al. [33] that was imputed by Muchlinski et al. The test set was released by Muchlinski et al. themselves. To be clear, Muchlinski et al. performed two imputation steps: one within the training dataset and one jointly on the train and test datasets. Only the latter leads to leakage. Details of our imputation methodology as well as a visualization of the data leakage in Muchlinski et al. are included in the Supplement.

Incorrect reuse of an imputed dataset [19, 20]: Colaresi and Mahmood [19] and Wang [20] compare the performance of ML models and Logistic Regression models and report that ML models vastly outperform Logistic Regression for predicting civil war onset. However, they re-use the imputed version of the Hegre et al. dataset [33] provided by Muchlinski et al. [18]. They use the imputed version of the dataset both for training and testing, either via a train-test split [19] or via k-fold cross-validation [20]; they do not use the out-of-sample test set provided by Muchlinski et al. This means that Colaresi and Mahmood's and Wang's studies are subject to exactly the same pitfall as above, albeit with a slightly different dataset. This leads to over-estimates of performance in both papers, similar to the ones in Muchlinski et al. It also showcases that by re-using datasets without carefully examining their construction, future research can lend credibility to exaggerated claims about the performance of ML models by "independently" validating their claims. Interestingly, one of these papers [20] aimed to highlight errors in an earlier version of the reproducibility materials provided by Muchlinski et al. [18]. Despite the resulting critical focus, they did not report errors in the imputation methodology.

Data leakage due to proxy variables [21]: Kaufman et al. [21] describe the benefits of using Adaboost models in a variety of settings. We limit ourselves to their claims about civil war prediction, which is not the primary focus of their paper; their other claims are outside the scope of our review. Specifically, they use a dataset created by Fearon and Laitin [42] in order to build models of civil war prediction. Critically, they use all variables in the dataset apart from the target variable as independent variables in their models. However, there are several proxy variables for the dependent variable in the dataset. The use of these as independent variables in the models created by Kaufman et al. causes data leakage and leads to over-optimistic performance claims. For example, the dataset contains variables such as *colwars, cowwars* and *sdwars* which represent instances of civil wars in other datasets. A model that uses these as independent variables would indeed attain high accuracy, but this would be a result of data leakage and would not be indicative of the model's out-of-sample predictive performance. We list all variables that cause data leakage in the Supplement and remove these from the list of independent variables in our corrected models. We find that once we remove the variables that cause leakage, ML models no longer outperform Logistic Regression models, and further, none of the models outperforms a baseline that predicts the outcome of the previous year — i.e., a baseline that predicts *war* if the outcome of the target variable was civil war in the previous year and predicts *peace* otherwise.

There are a few additional issues with their methods and results, listed below. Despite these issues, data leakage due to proxy variables causes the main difference between the original results and the corrected results. We describe the pitfalls pertaining to data leakage and cross-validation here, and defer a detailed description of the other issues to the Supplement.

- Kaufman et al. use an incorrect parameter selection technique when creating their baseline Lasso model that leads to the model always predicting *peace*. We correct this using a standard technique for parameter selection.
- It is unclear from their paper which target variable Kaufman et al. use. At different points in their paper, they mention that their main prediction task is *onset* as well as *occurrence*.
- They replace all missing values with 0 — which is not a standard way of imputing missing values, and don't report their imputation methodology in their paper.
- They ascribe significance to a small difference between the accuracy of Logistic Regression (98.7%) and AdaBoost (99.0%) in their paper without any statistical tests of significance for this difference [43].
- Their choice to use $k-$fold cross validation with temporal data may lead to over-optimistic performance claims [29, 44].

$k-$fold cross validation with temporal data [20, 21]: $k-$fold cross validation shuffles the dataset before it is divided into training and test datasets. When the dataset contains temporal

[3] While leakage is particularly serious in predictive modeling, a dataset with 95% of values missing is problematic even for explanatory modeling.

Field	Paper	Year	Num. papers reviewed	Num. papers with pitfalls	Pitfalls
Neuroimaging	Whelan et al. [34]	2014	—	4	Incorrect train-test split.
Autism diagnostics	Bone et al. [35]	2015	2	2	Biased evaluation, data leakage.
Bioinformatics	Blagus et al. [36]	2015	—	6	Data leakage.
Nutrition research	Ivanescu et al. [37]	2016	—	4	Incorrect train-test split.
Clinical epidemiology	Christodoulou et al. [38]	2019	71	48	Biased evaluation, data leakage.
Computer Security	Arp et al. [39]	2020	30	30	Multiple pitfalls.
Medicine	Vandewiele et al. [40]	2021	24	21	Incorrect train-test split, data leakage.
Radiology	Roberts et al. [41]	2021	62	62	Multiple pitfalls.

TABLE I: Recent results from eight fields in the quantitative sciences. In each case, the authors found pitfalls in papers applying ML methods in their fields.

data, the training dataset could contain data from a later date than the test dataset because of shuffling. Thus, during training the model has access to information that would not be available when the model is deployed. This is a data leakage that could lead to over-optimistic performance claims [29]. Both Wang [20] and Kaufman et al. [21] use $k-$fold cross validation to estimate model performance. Since there isn't consensus about the practical consequences of this pitfall [45], and to maintain comparability between the original and corrected results by testing on the same instances of civil war, we continue to use $k-$fold cross validation in the corrected results in Figure 1. We report results using an out-of-sample test set in the Supplement.

Lack of significance testing and uncertainty quantification

AUC-ROC (or any other metric) by itself, without understanding the variance, is inadequate for comparing predictive performance. For example, a classic paper on the ROC curve by Fawcett [30] cautions that "without a measure of variance we cannot compare the classifiers" in the context of comparing classifier performance on the basis of their ROC curves, advocating for variance estimation using multiple test sets or bootstrapping. Especially when sample sizes are small, significance testing and uncertainty quantification are important steps towards reproducibility across the statistical sciences, and machine learning is no exception [43, 46–52].

We found that 9 of the 12 papers for which complete code and data were available included no significance tests or uncertainty quantification for classifier performance comparison (see Supplement for details). As a heuristic, we suggest three factors that affect how likely it is that a lack of such tests will undermine the primary claims of the paper: when the primary claims are about out-of-sample performance, when the sample size or the number of positive instances in the test set is small, or when the reported performance difference between models is small.

We do not perform significance testing and uncertainty quantification for all 9 papers due to time limitations. However, as an illustration, we examine this issue in detail in the case of Blair and Sambanis [53] since their test dataset has a particularly small number of instances of civil war onset (only 11). Blair and Sambanis propose a model of civil war onset that uses theoretically informed features (called the *escalation* model) and report that it outperforms other baseline models of civil war onset using the AUC metric on an out-of-sample dataset.

We find that the performance of the model proposed by Blair and Sambanis — the *escalation* model — is not significantly better than any other model of civil war onset.[4] Further, all models have large confidence intervals for their out-of-sample performance. For example, while the smoothed AUC performance of the *escalation* model is 0.85, the 95% confidence interval calculated using bootstrapped test set resampling [54] is [0.66-0.95]. Similarly, while the next best model in their comparison is the *cameo* model, with an AUC metric of 0.82, the 95% confidence interval is [0.64-0.95]. Reporting these results without the large confidence intervals thus makes the performance estimates seem more certain than they are. Even this is an underestimate of the true uncertainty, both because it accounts for only one source of variance in classifier performance and because the data are not i.i.d — there are temporal as well as geospatial correlations in the dataset.

In general, there can be multiple sources of variance that affect the uncertainty estimates of model performance and there are ongoing research efforts to fully account for the variance of performance estimates of ML models [49, 50]. However, estimating and reporting bootstrapped confidence intervals for model performance is a well-known way to account for sample variance, and accounting for even one source of variance can be useful even if it is an underestimate [30, 54]. This does not yet appear to be the norm in most communities that develop or use ML methods.

In parallel to our work, Blair and Sambanis received a response from Beger et al. [55] highlighting various methodological issues, and published a rebuttal [56]. Neither of these papers discuss the role of significance testing and uncertainty quantification. Blair and Sambanis do report performance evaluations for a variety of different model specifications. However, the purpose of such robustness checks is to determine whether model performance sensitive to the parameter choices; it is unclear whether it helps deal with issues arising from sampling variance.

At any rate, Blair and Sambanis's results do in fact turn out to be highly sensitive to another modeling choice: the fact that they compute the AUC metric on the smoothed ROC curve instead of the empirical curve that their model produces (this issue was also pointed out by Beger, Morgan, and Ward [55]; Blair and Sambanis [56] discuss it in their rebuttal). Smoothing refers to a transformation of the ROC curve to make the predicted probabilities for the war and peace instances normally

[4] Z = 0.64, 1.09, 0.42, 0.67; p = 0.26, 0.14, 0.34, 0.25 for a one-tailed significance test comparing the smoothed AUC performance of the *escalation* model with other baseline models reported in their paper — *quad, goldstein, cameo* and *average* respectively. We implement the comparison test for smoothed ROC curves detailed by Robin et al. [54]. Note that we do not correct for multiple comparisons; such a correction would further reduce the significance of the results.

distributed instead of using the empirical ROC curve [54]. While the authors state that they smooth the ROC curves for ease of visual interpretation, they do not mention that they compute AUC values on the smoothed curves. The AUC metric is widely understood to refer to the empirical curve rather than the smoothed curve in machine learning [30, 55], so this comes as a surprise. Smoothing the curve alters the values of the AUC metric. We find that the performance difference highlighted by Blair and Sambanis [53] disappears when we compare the AUC metric for the empirical ROC curves instead of the smoothed curves. Smoothing the curve changes the value of the AUC metric by 0.05 — which is as much as the difference between the best and worst performing models compared in the paper. [5]

Discussion

Our results are best seen in light of reports about the pitfalls of adopting ML methods in various research communities. Each of the papers in Table I highlights pitfalls in published literature that uses ML methods.[6] Most of the reviews in Table I analyze the content of the papers to uncover pitfalls, and not the attached code. This leaves open the possibility that the claims in the original papers survive, at least in a weakened form, when the pitfalls are corrected. In contrast, we perform an in-depth code review to both discover and correct the errors in the code and show that when the errors are fixed, the original results cannot be reproduced. While there can be multiple reasons for being unable to reproduce a paper's findings, for example the lack of availability of reproduction materials [46] or a false positive finding [12], our focus is on pitfalls in performance evaluation that lead to claims not supported by the evidence. Beyond reproducibility, generalizability is also a serious concern for research that uses ML methods [57]. Another reason for overoptimism in research that uses ML methods that has been uncovered by systematic reviews is the use of weak baselines [58, 59].

Pitfalls in using ML methods should not be seen as problems being caused solely due to individual researchers not being trained in these methods, but also as a systemic feature of applied ML research as it stands today. Systemic causes have been recognized in past reproduction crises as well, and identifying issues as such calls for interventions that go beyond individual action or remediation. For example, incentives in academic publishing and a bias towards publishing 'novel' results means that false positives are more likely to be published and cited in scientific journals [10, 60].

None of the leakage errors we found were decipherable from reading the papers or re-running the code to verify that the reported results were produced — we needed to review the code in order to find the errors in each case. Thus, our work supports calls for incorporating code review into the peer-review process as one way to improve reproducibility [61, 62]. Still, we caution that pitfalls might be subtle and that peer reviewers may lack the time and incentives to uncover them. For example, one of the papers we reviewed had over 10,000 lines of code [53]. Similarly, while the pre-registration of research plans can improve the transparency of research methods and reduce false positive results [2], it does not guard against methodological pitfalls that might occur despite registered research plans, including leakage. There is no silver bullet for uncovering pitfalls before papers are published. Thus, our findings also support the usefulness of systematic, methodologically critical reviews of published research across fields adopting ML methods, and using in-depth code review to investigate their reproducibility.

More sweeping interventions are worth considering. In particular, the common task framework is a paradigm for ML research that addresses the root causes of the reproducibility failures that we identified as well as the vast majority of similar failures identified in previous research (Table I). The common task framework allows us to compare the performance of competing ML models using an agreed-upon training dataset and evaluation metrics, a secret holdout dataset, and a public leaderboard. Dataset creation and model evaluation is left to impartial third parties who have the expertise and incentives to avoid errors. This framework is most useful when a field has agreed upon a set of research problems to address [1]. The importance of the common task framework has been emphasized in recent literature in ML as well as social sciences [1, 2, 63–65]. Donoho called it the 'secret sauce' of the predictive culture, attributing most major successes of machine learning to the common task framework [66]. To be clear, the move to a common task framework should not come at the expense of domain expertise. ML experts working without domain knowledge are equally prone to errors, as highlighted by the failure of prospective validation in medical ML research [67]. Rather, we recognize that when communities adopt the common task framework, they should adapt it to their needs using domain expertise. We also caution that there are many downsides to a singular focus on optimizing a particular accuracy metric by a community to the exclusion of other scientific and normatively desirable properties of models [68, 69].

Finally, our results question the usefulness of ML models for civil war prediction, as they are shown to perform no better than Logistic Regression models across a variety of settings. The comparison between ML and Logistic Regression models has been a subject of debate in other Political Science subfields as well [70–72]. In each of the four papers that compare the performance of ML models and Logistic Regression models for civil war prediction [18–21], ML models substantially outperform Logistic Regression models in the presence of data leakage, but not when the leakage is fixed. We hypothesize that this is not a coincidence. Leakage can introduce complex patterns into the dataset that ML models might be better able to represent than Logistic Regression models. Thus, when there are widespread methodological pitfalls in the research literature that compares different types of models, it may lead to a systematic bias in favor of complex models.

Our finding about the questionable usefulness of ML models for civil war prediction is consistent with similar sobering findings in other tasks involving predicting social outcomes such as children's life outcomes [65] and recidivism [73]. This is in contrast to perception tasks such as image classification where true leaps in predictive accuracy have been achieved using complex models. For research on predicting social outcomes, we argue that claims of high accuracy should be treated as a-

[5] Here, we report results for the 1 month forecast in Blair and Sambanis [53]. Our findings generalize to the 6 month forecast reported in their paper as well, and are detailed in the Supplement.

[6] We used Google Scholar searches for terms related to overoptimism and Machine Learning to find the papers in Table I. We learned about the paper by Arp et al. [39] because it highlighted pitfalls in a paper authored by one of us. This further underscores our view that it is not sufficient for individual researchers to be more vigilant and systemic interventions are necessary.

priori unlikely and should be subjected to careful scrutiny. While there isn't a rigid threshold for what constitutes high accuracy, AUC values well over 0.9 for civil war onset prediction — which were common among the papers we reviewed — appear facially implausible.

In conclusion, our findings highlight the need for tempering the optimism about the paradigm of prediction and call for systemic interventions that can help distinguish true advances from false optimism.

ACKNOWLEDGEMENTS. We are grateful to Jessica Hullman, Matthew J. Salganik and Brandon Stewart for their valuable feedback on drafts of this paper. We thank Robert Blair, Aaron Kaufman, David Muchlinski and Yu Wang for quick and helpful responses to drafts of this paper.

References

[1] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. "Prediction and explanation in social systems". eng. In: *Science (New York, N.Y.)* 355.6324 (Feb. 2017), pp. 486–488.

[2] Jake M. Hofman et al. "Integrating explanation and prediction in computational social science". en. In: *Nature* 595.7866 (July 2021), pp. 181–188.

[3] Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (Aug. 2001), pp. 199–231.

[4] Susan Athey and Guido W. Imbens. "Machine Learning Methods That Economists Should Know About". In: *Annual Review of Economics* 11.1 (2019), pp. 685–725.

[5] Daniel R. Schrider and Andrew D. Kern. "Supervised Machine Learning for Population Genetics: A New Paradigm". en. In: *Trends in Genetics* 34.4 (Apr. 2018), pp. 301–312.

[6] John Joseph Valletta et al. "Applications of machine learning in animal behaviour studies". en. In: *Animal Behaviour* 124 (Feb. 2017), pp. 203–220.

[7] R. Iniesta, D. Stahl, and P. McGuffin. "Machine learning, statistical learning and the future of biological research in psychiatry". en. In: *Psychological Medicine* 46.12 (Sept. 2016), pp. 2455–2465.

[8] Scott Tonidandel, Eden B. King, and Jose M. Cortina. "Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science". en. In: *Organizational Research Methods* 21.3 (July 2018), pp. 525–547.

[9] Tal Yarkoni and Jacob Westfall. "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning". en. In: *Perspectives on Psychological Science* 12.6 (Nov. 2017), pp. 1100–1122.

[10] Marta Serra-Garcia and Uri Gneezy. "Nonreplicable publications are cited more than replicable ones". en. In: *Science Advances* 7.21 (May 2021), eabd1705.

[11] Jeffrey T. Leek and Roger D. Peng. "Opinion: Reproducible research can still be wrong: Adopting a prevention approach". en. In: *Proceedings of the National Academy of Sciences* 112.6 (Feb. 2015), pp. 1645–1646.

[12] Joelle Pineau et al. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)". In: *arXiv:2003.12206 [cs, stat]* (Dec. 2020).

[13] Engineering National Academies of Sciences. *Reproducibility and Replicability in Science.* en. May 2019.

[14] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. "A visual tool for defining reproducibility and replicability". en. In: *Nature Human Behaviour* 3.7 (July 2019), pp. 650–652.

[15] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. "What does research reproducibility mean?" en. In: *Science Translational Medicine* 8.341 (June 2016), 341ps12–341ps12.

[16] Open Science Collaboration. "Estimating the reproducibility of psychological science". en. In: *Science* 349.6251 (Aug. 2015).

[17] Michael D Ward, Brian D Greenhill, and Kristin M Bakke. "The perils of policy by p-value: Predicting civil conflicts". en. In: *Journal of Peace Research* 47.4 (July 2010), pp. 363–375.

[18] David Muchlinski et al. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data". en. In: *Political Analysis* 24.1 (2016), pp. 87–103.

[19] Michael Colaresi and Zuhaib Mahmood. "Do the robot: Lessons from machine learning to improve conflict forecasting". en. In: *Journal of Peace Research* 54.2 (Mar. 2017), pp. 193–214.

[20] Yu Wang. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment". en. In: *Political Analysis* 27.1 (Jan. 2019), pp. 107–110.

[21] Aaron Russell Kaufman, Peter Kraft, and Maya Sen. "Improving Supreme Court Forecasting Using Boosted Decision Trees". en. In: *Political Analysis* 27.3 (July 2019), pp. 381–387.

[22] Daniel W. Hook, Simon J. Porter, and Christian Herzog. "Dimensions: Building Context for Search and Evaluation". English. In: *Frontiers in Research Metrics and Analytics* 3 (2018).

[23] Corinne Bara. *Forecasting civil war and political violence.* en. May 2020.

[24] *Scimago Journal & Country Rank.*

[25] A. Rogier T. Donders et al. "Review: A gentle introduction to imputation of missing values". en. In: *Journal of Clinical Epidemiology* 59.10 (Oct. 2006), pp. 1087–1091.

[26] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling.* en. New York, 2013.

[27] Rayid Ghani, Joe Walsh, and Joan Wang. *Top 10 ways your Machine Learning models may have leakage (URL: http://www.rayidghani.com/2020/01/24/top-10-ways-your-machine-learning-models-may-have-leakage/).* en-US. Jan. 2020.

[28] *Imputation before or after splitting into train and test? (URL: https://stats.stackexchange.com/questions/95083/imputation-before-or-after-splitting-into-train-and-test).*

[29] Shachar Kaufman et al. "Leakage in data mining: Formulation, detection, and avoidance". In: *ACM Transactions on Knowledge Discovery from Data* 6.4 (Dec. 2012), 15:1–15:21.

[30] Tom Fawcett. "An introduction to ROC analysis". en. In: *Pattern Recognition Letters.* ROC Analysis in Pattern Recognition 27.8 (June 2006), pp. 861–874.

[31] Marcel Neunhoeffer and Sebastian Sternberg. "How Cross-Validation Can Go Wrong and What to Do About It". en. In: *Political Analysis* 27.1 (Jan. 2019), pp. 101–106.

[32] David Alan Muchlinski et al. "Seeing the Forest through the Trees". en. In: *Political Analysis* 27.1 (Jan. 2019), pp. 111–113.

[33] Håvard Hegre and Nicholas Sambanis. "Sensitivity Analysis of Empirical Results on Civil War Onset:" en. In: *Journal of Conflict Resolution* (2006).

[34] Robert Whelan and Hugh Garavan. "When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging". en. In: *Biological Psychiatry.* Mechanisms of Aging and Cognition 75.9 (May 2014), pp. 746–748.

[35] Daniel Bone et al. "Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises". en. In: *Journal of Autism and Developmental Disorders* 45.5 (May 2015), pp. 1121–1136.

[36] Rok Blagus and Lara Lusa. "Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models". In: *BMC Bioinformatics* 16.1 (Nov. 2015), p. 363.

[37] A. E. Ivanescu et al. "The importance of prediction model validation and assessment in obesity and nutrition research". en. In: *International Journal of Obesity* 40.6 (June 2016), pp. 887–894.

[38] Evangelia Christodoulou et al. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". en. In: *Journal of Clinical Epidemiology* 110 (June 2019), pp. 12–22.

[39] Daniel Arp et al. "Dos and Don'ts of Machine Learning in Computer Security". en. In: (Oct. 2020).

[40] Gilles Vandewiele et al. "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling". en. In: *Artificial Intelligence in Medicine* 111 (Jan. 2021), p. 101987.

[41] Michael Roberts et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". en. In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 199–217.

[42] James D. Fearon and David D. Laitin. "Ethnicity, Insurgency, and Civil War". en. In: *American Political Science Review* 97.01 (Feb. 2003), pp. 75–90.

[43] Thomas G Dietterich. "Approximate statistical tests for comparing supervised classification learning algorithms". In: *Neural computation* 10.7 (1998), pp. 1895–1923.

[44] David R. Roberts et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure”. en. In: *Ecography* 40.8 (2017), pp. 913–929.

[45] Christoph Bergmeir and José M. Benítez. “On the use of cross-validation for time series predictor evaluation”. en. In: *Information Sciences*. Data Mining for Software Trustworthiness 191 (May 2012), pp. 192–213.

[46] Matthew B. A. McDermott et al. “Reproducibility in machine learning for health research: Still a ways to go”. en. In: *Science Translational Medicine* 13.586 (Mar. 2021).

[47] Gaël Varoquaux. “Cross-validation failure: Small sample sizes lead to large error bars”. en. In: *NeuroImage*. New advances in encoding and decoding of brain signals 180 (Oct. 2018), pp. 68–77.

[48] Peter Henderson et al. “Deep Reinforcement Learning That Matters”. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018).

[49] Victoria Volodina and Peter Challenor. “The importance of uncertainty quantification in model reproducibility”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2197 (May 2021), p. 20200071.

[50] Xavier Bouthillier et al. “Accounting for Variance in Machine Learning Benchmarks”. en. In: (2020), p. 23.

[51] J. A. Hanley and B. J. McNeil. “A method of comparing the areas under receiver operating characteristic curves derived from the same cases”. eng. In: *Radiology* 148.3 (Sept. 1983), pp. 839–843.

[52] James Carpenter and John Bithell. “Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians”. en. In: *Statistics in Medicine* 19.9 (2000), pp. 1141–1164.

[53] Robert A. Blair and Nicholas Sambanis. “Forecasting Civil Wars: Theory and Structure in an Age of “Big Data” and Machine Learning”. en. In: *Journal of Conflict Resolution* 64.10 (Nov. 2020), pp. 1885–1915.

[54] Xavier Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12.1 (Mar. 2011), p. 77.

[55] Andreas Beger, Richard K. Morgan, and Michael D. Ward. “Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict”. en. In: *Journal of Conflict Resolution* (July 2021), p. 0022002720982358.

[56] Robert A. Blair and Nicholas Sambanis. “Is Theory Useful for Conflict Prediction? A Response to Beger, Morgan, and Ward”. en. In: *Journal of Conflict Resolution* (July 2021), p. 00220027211026748.

[57] Xin-Lu Cai et al. “Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data”. en. In: *Human Brain Mapping* 41.1 (2020), pp. 172–184.

[58] Orianna DeMasi, Konrad Kording, and Benjamin Recht. “Meaningless comparisons lead to false optimism in medical machine learning”. In: *PLoS ONE* 12.9 (Sept. 2017).

[59] Maurizio Ferrari Dacrema et al. “A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research”. In: *ACM Transactions on Information Systems* 39.2 (Jan. 2021), 20:1–20:49.

[60] John P. A. Ioannidis. “Why Most Published Research Findings Are False”. en. In: *PLOS Medicine* 2.8 (Aug. 2005), e124.

[61] William E. Davis et al. “Peer-Review Guidelines Promoting Replicability and Transparency in Psychological Science”. en. In: *Advances in Methods and Practices in Psychological Science* 1.4 (Dec. 2018), pp. 556–573.

[62] “Easing the burden of code review”. en. In: *Nature Methods* 15.9 (Sept. 2018), pp. 641–641.

[63] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. “Machine Learning for Social Science: An Agnostic Approach”. In: *Annual Review of Political Science* 24.1 (2021), pp. 395–419.

[64] Mark Liberman. “Fred Jelinek”. en. In: *Computational Linguistics* 36.4 (Dec. 2010), pp. 595–599.

[65] Matthew J. Salganik et al. “Measuring the predictability of life outcomes with a scientific mass collaboration”. en. In: *Proceedings of the National Academy of Sciences* 117.15 (Apr. 2020), pp. 8398–8403.

[66] David Donoho. “50 Years of Data Science”. In: *Journal of Computational and Graphical Statistics* 26.4 (Oct. 2017), pp. 745–766.

[67] Myura Nagendran et al. “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies”. en. In: *BMJ* 368 (Mar. 2020), p. m689.

[68] Amandalynne Paullada et al. “Data and its (dis) contents: A survey of dataset development and use in machine learning research”. In: *arXiv preprint arXiv:2012.05345* (2020).

[69] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. “Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers”. In: *arXiv:2106.15195 [cs]* (June 2021).

[70] Nathaniel Beck, Gary King, and Langche Zeng. “Improving Quantitative Studies of International Conflict: A Conjecture”. en. In: *American Political Science Review* 94.1 (Mar. 2000), pp. 21–35.

[71] Scott De Marchi, Christopher Gelpi, and Jeffrey D. Grynaviski. “Untangling Neural Nets”. en. In: *American Political Science Review* 98.2 (May 2004), pp. 371–378.

[72] Nathaniel Beck, Gary King, and Langche Zeng. “Theory and Evidence in International Conflict: A Response to de Marchi, Gelpi, and Grynaviski”. en. In: *American Political Science Review* 98.2 (May 2004), pp. 379–389.

[73] Julia Dressel and Hany Farid. “The accuracy, fairness, and limits of predicting recidivism”. In: *Science advances* 4.1 (2018).