Routledge
Taylor & Francis Group

International Interactions

Empirical and Theoretical Research in International Relations

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gini20

Forecasting conflict in Africa with automated machine learning systems

Vito D'Orazio & Yu Lin

To cite this article: Vito D'Orazio & Yu Lin (2022): Forecasting conflict in Africa with automated machine learning systems, International Interactions, DOI: 10.1080/03050629.2022.2017290

To link to this article: https://doi.org/10.1080/03050629.2022.2017290

Published online: 15 Jan 2022.

INTERNATIONAL INTERACTIONS
https://doi.org/10.1080/03050629.2022.2017290

Routledge
Taylor & Francis Group

Forecasting conflict in Africa with automated machine learning systems

Vito D'Orazio[a] and Yu Lin[b]

[a]Associate Professor of Political Science, University of Texas at Dallas; [b]PhD Candidate in Computer Science, University of Texas at Dallas

ABSTRACT

The ViEWS problem is to forecast changes in the level of state-based violence for each of the next six months at the PRIO-GRID and country level. For this competition and toward the goal of improving sub-national and country level forecasts, we experiment with combinations of automated machine learning (autoML) systems and limited datasets that emphasize the endogenous nature of conflict. Two core findings emerge: autoML improves predictive performance and the Dynamics model performs best. The data used for the Dynamics model is limited to measures of state-based violence built from the event-level violence data plus those describing the spatial and temporal structure of the data. The intent is to capture spatial and temporal conflict dynamics while not overfitting to exogenous factors, which is especially problematic with flexible autoML algorithms and the types of highly disaggregate data used here. At the PGM level, this model won the ViEWS competition for "predictive accuracy" and split the win for "originality." Beyond the ViEWS competition, we expect conflict forecasting models that couple advanced autoML systems with variables that reflect a diverse set of conflict dynamics to have high predictive performance, especially at sub-national and sub-annual aggregations.

El problema del ViEWS es que predice los cambios en el nivel de violencia estatal de cada uno de los próximos seis meses a nivel de PRIO-GRID y de país. En el marco de esta competencia y con el objetivo de mejorar las predicciones a nivel regional y nacional, probamos combinaciones de sistemas de aprendizaje automático (autoML) y conjuntos de datos limitados que ponen de relieve la naturaleza endógena de los conflictos. Hay dos resultados principales: el autoML mejora el rendimiento predictivo y el modelo Dynamics es el que mejor funciona. Los datos utilizados para el modelo Dynamics se limitan a las medidas de la violencia a nivel estatal establecidas a partir de los datos de la violencia sobre eventos más los que

KEYWORDS

Conflict forecastingpredictive modelingautomated machine learning

CONTACT Vito D'Orazio dorazio@utdallas.edu Vito D'Orazio, The University of Texas at Dallas, 800 W Campbell Rd, GR 31, Richardson, TX 75080, USA

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (http://dx.doi.org/10.1080/03050629.2022.2101217).

Replication materials are available at http://dvn.iq.harvard.edu/dvn/dv/internationalinteractions. All questions regarding replication may be directed to Vito D'Orazio.

describen la estructura espacial y temporal de los datos. La intención es captar la dinámica espacial y temporal de los conflictos sin caer en el exceso de ajuste de los factores exógenos, lo que supone un problema, sobre todo con los algoritmos autoML flexibles y los tipos de datos altamente desagregados que se utilizan aquí. A nivel de PGM, este modelo ganó la competencia del ViEWS tanto por su "precisión predictiva" como por su "originalidad". Más allá de la competencia del ViEWS, esperamos que los modelos de previsión de conflictos que combinan sistemas avanzados de autoML con variables que reflejan un conjunto diverso de dinámicas de conflicto tengan un alto resultado predictivo, sobre todo en agregados regionales y semestrales.

La problématique du ViEWS (Violence early-warning system, système d'alerte précoce sur la violence) est de prévoir les évolutions du niveau de violence étatique pour chacun des six prochains mois au niveau de la grille PRIO et au niveau national. Pour ce concours et dans l'objectif d'améliorer les prévisions au niveau infranational et au niveau national, nous avons expérimenté des combinaisons de systèmes de machine learning automatisés (autoML) et de jeux de données limités mettant l'accent sur la nature endogène des conflits. Deux résultats fondamentaux sont apparus : l'autoML améliore les performances prédictives et le modèle Dynamiques est le plus efficace. Les données utilisées pour le modèle Dynamiques sont limitées aux mesures de la violence étatique établies à partir des données sur la violence au niveau des événements ainsi que de celles qui décrivent la structure spatiale et temporelle des données. L'objectif est de capturer les dynamiques spatiales et temporelles des conflits tout en évitant un ajustement excessif aux facteurs exogènes, ce qui est particulièrement problématique avec les algorithmes d'autoML flexibles et les types de données très désagrégées qui sont utilisés ici. Au niveau PGM, ce modèle a remporté le concours ViEWS à la fois dans les catégories « Précision prédictive » et « Originalité ». Au-delà du concours ViEWS, nous nous attendons à ce que les modèles de prévision des conflits qui allient des systèmes avancés d'autoML à des variables reflétant un ensemble diversifié de dynamiques de conflits aient de hautes performances prédictives, en particulier aux niveaux d'agrégation infranationaux et infra-annuels.

Introduction

Researchers have used many different algorithms to predict and forecast conflict, including common methods such as logistic regression (Goldstone et al. 2010; Ward, Greenhill, and Bakke 2010). To improve predictive performance, some have used machine learning algorithms such as neural networks (Beck, King, and Zeng 2000; King and Zeng 2001; Lagazio and Marwala 2006) and random forests (Colaresi and Mahmood 2017; Hegre et al. 2019; Hill and Jones 2014). Others have used network-based

techniques, which are generally less flexible but designed to account for known or expected dependencies in conflict data (Cranmer and Desmarais 2017; Dorff, Gallop, and Minhas 2020; Minhas, Hoff, and Ward 2016).

Recently, automated machine learning (autoML) algorithms have been developed and used to solve problems where the researcher values predictive performance but is less concerned with the choice of learning algorithm, hyperparameter tuning, and other modeling decisions that are considered secondary or even arbitrary (Hutter, Kotthoff, and Vanschoren 2019). The basic idea of autoML is to search over a large space of potential solutions, including different learning algorithms and hyperparameters, to find those with the best predictive performance. AutoML algorithms are highly flexible, which is promising for conflict forecasting because of the many nonlinear relationships in conflict data. However, this flexibility also means that autoML algorithms are prone to return solutions that are overfit and will not generalize well to new, out-of-sample data.

Researchers have also used many different predictors in their forecasting models. Generally, these variables are theoretically meaningful and have been selected from a larger set of potential inputs that could be expected to improve performance. However, increasing the number of predictors means the model is more likely to be overfit, especially when coupled with a flexible autoML algorithm. Data-driven methods for variable selection, such as forward and backward stepwise, exist but also run the risk of producing models that are not well-grounded in theories of conflict and are especially difficult to interpret. So, features are commonly grouped and selected by theory. For example, in models that forecast civil conflict there may be a set of predictors to represent grievances. By grouping variables by theory, interpretation is easier since the model as a whole is assumed to represent a familiar theory or class of theories.

One approach to variable selection that is grounded in theory, is easy to implement because it is not dependent on external data, and makes for an excellent benchmark model is to begin with very few or no exogenous variables and focus on modeling the endogenous nature of the conflict itself (Cranmer and Desmarais 2017). Spatial and temporal dependencies are a well known feature of conflict (Carter and Signorino 2010; Schutte and Weidmann 2011; Ward, Siverson, and Cao 2007). Models with variables that capture these conflict dynamics, and with a highly flexible machine learning algorithm, are expected to produce forecasts that are accurate, less overfit, and understandable simply as forecasts of future violence based on present and past violence.

In this paper, we experiment with autoML algorithms and limited datasets for the ViEWS conflict forecasting problem. In short, the ViEWS problem is to forecast changes in state-based violence for each of the next six

months (Hegre et al. 2019). These monthly forecasts are made at the PRIO-GRID level (PGM) and the country level (CM) (Tollefsen, Strand, and Buhaug 2012). The primary scoring criteria for the ViEWS problem is the mean squared error (MSE), although others may be used. As part of the ViEWS competition, forecasts were submitted in September, 2020 for each month, October through March, 2021. For more information on the competition, see Hegre, Vesco, and Colaresi (2022) and Vesco et al. (2022).

Our core experiments test three autoML systems: CMU, AlphaD3M (Drori et al. 2018), and H2O (H2O.ai 2017). CMU and AlphaD3M have been developed as part of the Data-Driven Discovery of Models (D3M) program funded by the Defense Advanced Research Projects Agency, while H2O is an off-the-shelf, publicly available autoML suite.[1] All three are open-source projects. Using each autoML system, we compare the full ViEWS model specification to three reduced datasets: *ViEWS-30*, *Dynamics*, and *Dynamics-Hurdle*. The ViEWS-30 model uses the 30 most important features as determined by the ViEWS team. The Dynamics model includes variables that are intended to capture spatial and temporal conflict dynamics. The Dynamics-Hurdle model is only for the PGM setting, and it uses the same variables as the Dynamics model but forecasts no conflict in all grids that experienced no conflict in the training data. For this study, the dependent variable is the natural log of state-based fatalities from the Georeferenced Event Data (Sundberg and Melander 2013). For the ViEWS problem, we forecast the level of violence and subtract the current level of violence to arrive at our predicted change.

Our results show two primary findings. One, autoML consistently improves predictive performance. This finding holds across all autoML algorithms tested. The stacked ensemble method from the H2O solver performs best. Within the stacked ensemble, the most informative base model is the gradient boosting machine. Across all autoML systems, boosting methods (e.g., gradient boosting machine) tend to perform better than bagging methods (e.g., random forest). If researchers have no particular reason to choose one learning algorithm over another, the results suggest using an autoML system that has boosting and stacked ensemble methods.

The second primary finding is that the Dynamics model performs best. This is the case for both the PGM and CM problems, but the finding is more pronounced and consistent for PGM. Looking at the importance of individual variables, the count of fatal events (*ged_count_sb*) is consistently among the top contributors. Other measures of *violence* within the grid or country at time *t*, such as the log fatalities (*ln_ged_best_sb*), are also important contributors. This

[1]CMU has been developed at Carnegie Melon and AlphaD3M has been developed at NYU. For updates on these systems or the D3M program, visit https://datadrivendiscovery.org/.

supports the notion that conflict at time t is the best predictor of future conflict, but also that multiple measures of conflict at time t are valuable.

At the PGM level, our model won the competition for "predictive accuracy" and, along with Lindholm et al. (2022), split the win for "originality" (Vesco et al. 2022). Beyond this competition, we expect conflict forecasting models that couple advanced autoML systems with variables that reflect a diverse set of conflict dynamics to have high predictive performance, especially at sub-national and sub-annual levels of aggregation. These models are relatively easy to build since external data is not needed. They also make excellent benchmarks for theory building and comparison because they start by modeling the endogenous nature of conflict, without assuming we know much about what that endogenous process is (Cranmer and Desmarais 2017). Given the exclusion of exogenous, theory-driven variables, Dynamics models are consistent with the suggestion that "less is more" (Ward 2016, 84). Fortunately, the ViEWS team has provided a large number of variable transformations to capture a diverse set of conflict dynamics. This blueprint will be helpful to use and expland upon when moving beyond GED and into other type of political violence.

The Problem and Motivation

The conflict forecasting problem can be described by the following formula:

$$\hat{y}_t = f_a(x_{t-k}, y_{t-k}; \theta_k) \tag{1}$$

where $\hat{y}_t$ is the predicted target value at time t. x_i and y_i denote corresponding exogenous and endogenous features at time t, respectively. a is the learning algorithm, such as random forest or gradient boosting. $f_a(\cdot; \theta_k)$ is the model parameterized by θ_k and k represents the time lag considered by this model. Both exogenous and endogenous variables are optional to the forecasting model.

Conflict researchers typically specify a, x, and y using a mix of insights from both substantive and statistical theory (Blair and Sambanis 2020). The algorithm's tuning parameters, θ, are generally fit to maximize predictive performance. While there may be some search over all these parameters, the search tends to be limited due to computational resources and time.

AutoML algorithms search over an expansive space (S) to identify the solution with the best predictive performance. That is, the autoML algorithm explores different combinations of learning algorithms, tuning parameters, and possibly feature sets including both exogenous and endogenous variables. Depending on the algorithm, a solution s might incorporate further modeling and data decisions, such as what do to with

missing values and outliers, whether to reduce data dimensionality, how to adjust for class imbalance, and alternative variable specifications. At a minimum, autoML algorithms search over a and the tuning parameters.

Performance measures such as the mean squared error and the continuous rank probability score are used to compare these solutions, with the expected test error rate as the arbiter (Brandt, Freeman, and Schrodt 2014; James et al. 2013). Using the expected test error rate, which is often estimated with methods such as k-fold cross-validation, along with standard data partitioning techniques, helps to mitigate overfitting (Hastie, Tibshirani, and Friedman 2011; Hyndman and Athanasopoulos 2018).

In contrast to inferential models, predictive models emphasize predictive performance, either in model selection, model assessment, or both. While "machine learning" is essentially just a set of algorithms, predictive modeling is more of a research design than a choice of algorithm. As described in D'Orazio (2020), conflict researchers have used this design to offer guidance for policy-makers (Goldsmith and Butcher 2018 Hegre et al. 2013; Rost, Schneider, and Kleibl 2009), to assess variable importance (Gelpi and Avdan 2018; Ward et al. 2010), and for testing and comparing theory (Chenoweth and Ulfelder 2017; Gleditsch and Ward 2013; Jones and Lupu 2018). Bowlsby et al. (2020) offer some caveats in using predictive models for theory testing. Each of these applications for predictive modeling in conflict research would benefit from expanding the search space S.

Figure 1 represents the motivation for autoML conceptually. The solution space S is represented by the 9×9 grid. Each cell in the grid is a potential solution, i.e., a fit model that will provide a prediction given a set of inputs. The white cells are potential solutions that have not been fit or tested, while the shaded cells are those that have been fit. For example, the researcher may select a learning algorithm and experiment with different feature sets. Here, the white cells may represent solutions using other learning algorithms, while the shaded cells are the models in the researcher's experiment. The outlined box consisting of six cells in the lower left of

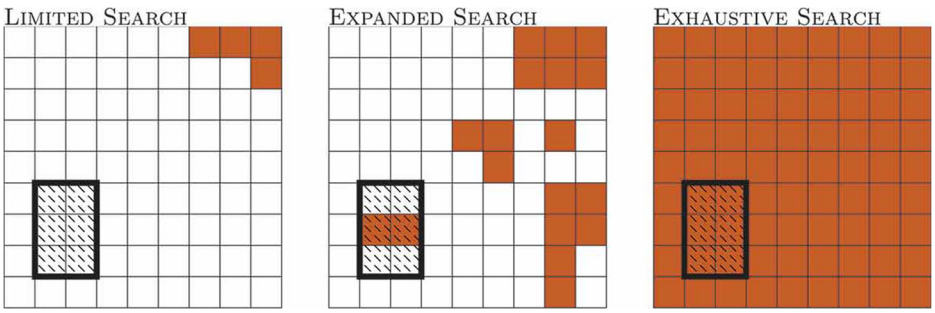

Figure 1. Example predictive modeling search types. Conflict researchers tend toward limited search, but would benefit by moving toward exhaustive search.

each grid represents solutions to avoid because they are not meaningful or helpful. For example, for some binary classification problems a model that always predicts 1 or 0 might have good predictive performance, but is not meaningful.

The Limited Search, on the left, represents a case where autoML is not used and all decisions have been specified by the researcher. On the right, the Exhaustive Search represents a full autoML search. Moving from a limited search to an exhaustive search, we expect to see improvements in predictive performance because more of S is covered. With predictive models, knowing that more of the search space has been assessed is a way to build confidence in any inferences drawn.

An AutoML search does not mean the research automatically accepts the model with highest predictive performance. Rather, it means a larger solution space has been explored in terms of models that have been fit. As algorithms search in increasingly exhaustive ways, there will be solutions found that are not helpful and are not meaningful. This can happen even when the researcher takes care in establishing the configurations of the data and the problem, such as omitting nonsense variables and making informed decisions about how to handle missingness. In a limited search, researchers avoid these models. With autoML, solutions require vetting and may be filtered after the search. Other solutions will be found that are acceptable, but may be overfit or less desirable because of a lack of interpretability. These solutions often require subject-matter expertise to identify.

General AutoML

AutoML systems solve a hyperparameter optimization (HPO) problem. A trained model, $f(\cdot;\theta)$, can be recognized as the "realization" of a learning algorithm with corresponding hyperparameters. Assume this model has M hyperparameters, which are denoted as $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$. The *hyperparameter configuration space* is represented as $S_{hp} = \lambda_1 \times \lambda_2 \times \ldots \lambda_m$. Let v be an instantiated vector of hyperparameters in the hyperparameter configuration space. Given a dataset D, the goal of the autoML system is to find

$$v^*_{hp} = \underset{v \in S_{hp}}{\operatorname{argmin}} L(v, f, D_{train}, D_{valid})$$

where $L(v, f, D_{train}, D_{valid})$ measures the loss of a model f, trained with hyper-parameter vector v.

Solutions for this HPO problem can be divided into three main categories: 1) *Model-Free Blackbox Optimization* (e.g. Grid Search, Random

Search); 2) *Bayesian Optimization* (e.g. meta-learning, Gaussian process); 3) *Neural Architecture Search*. The latter two optimization families have limited applicability or require extensive computational resources, so the *Model-Free Blackbox optimization* method is more common and is used by the autoML engines tested here. Grid Search and Random Search are further described in Bergstra and Bengio (2012).

The *hypterparameter configuration space* becomes very large, with cardinality $|S_{hp}| = |\lambda_1| \times |\lambda_2| \cdots |\lambda_m|$ where each $v \in S_{hp}$ indicates one hyperparameter configuration. The grid search algorithm iterates through all possible configurations and returns the best model. Obviously, a complete grid search becomes prohibitively costly very quickly. For practical use, this family of autoML systems have a time constraint, which means the system returns solutions and stops the process after the allotted time. Random search may help to find better solutions when the space becomes large. Here, each v_{hp} is sampled from the space. The stop criterion can either be a time allotment or a performance-metric-based criterion. If the latter, the search stops when the improvement between two hyperparameter vectors is less than a predefined threshold.

H2O and Stacking

To preview results, H2O outperforms AlphaD3M and CMU in most of our settings. We expect this is because H2O, unlike the others, uses an ensemble of base models as a potential solution. Specifically, H2O uses an ensemble algorithm called Stacking (Van der Laan, Polley, and Hubbard 2007). The method is conceptually simple: it trains a second-level meta-learner to find the optimal combination of the base learners (e.g., Lasso Regression, Logistic Regression, Random Forest, etc.). Mathematically, assume we have n data instances and K base learners. Each base learner is indexed by $k \in \{1, \ldots, K\}$, and represents a function $P_n \rightarrow \hat{\Psi}_k(P_n)$ from empirical probability distribution P_n to functions of the covariates. Follow the same convention, the meta learner can be defined as

$$\hat{\Psi}_{meta}(P_n) \equiv \hat{\Psi}_{\hat{K}(P_n)}(P_n)$$

where $\hat{K}(P_n)$ is an indicator function that select the best learner under cross-validation. The stacking algorithm first trains the base learners with cross-validation and build corresponding selectors. Next, it will bind the prediction vectors from each base learner together to form a new design matrix Z, and the meta learner is trained on this matrix.

Specifically, the meta learner is trained with V-fold cross-validation, where the n data points are splited into V disjoint blocks. For each $v \in \{1, \ldots, V\}$, the validation set $Val(v)$ is just the v-th block and the training

set $Tra(v)$ is the union of rest blocks. Let $\psi_{nkv} \equiv \hat{\Psi}_k(P_{n,\,Tra(v)})$ be the realization of the k-th estimator applied to the training set $P_{n,\,Tra(v)}$. It's noteworthy that the meta learner transforms any original data point (X_i, y_i) to (Z_i, y_i) before the training, where $Z_i \equiv (\psi_{nkv}(X_i) : k = 1, \ldots, K)$ is the vector consisting of the K predicted values according to the K trained estimators. Finally, H2O trains the meta learner on this constructed matrix Z using a Generalized Linear Model by default.

Data and Research Design

The ViEWS problem is to forecast changes in the log of state-based fatalities (*ln_ged_best_sb*) for each the next six months. These monthly forecasts are made at the PRIO-GRID level (PGM) and the country level (CM). The dependent variable is from the Georeferenced Event Dataset (Sundberg and Melander 2013). Forecasts are made for 1, 2, 3, 4, 5, and 6 months in advance. While the ViEWS problem is to forecast changes in log fatalities, our models forecast the level of log fatalities. The change is calculated by subtracting the current level from the forecasted level. The results reported in this paper are for the level of violence, not the change.

The data partitions for the reported results and for model selection are shown in Table 1. While ViEWS made global data available, the ViEWS problem is to forecast violence in Africa. So, we restrict our data to grids and countries in Africa, which is consistent with the ViEWS benchmark.

Mean squared error (MSE) is one of the primary scoring criteria for the ViEWS competition (Vesco et al. 2022). We use MSE for all model selection and evaluation results.

AutoML Systems

We use three different open-source and publicly available autoML systems. The CMU and AlphaD3M systems have been developed as part of DARPA's Data-Driven Discovery of Models (D3M) program (D3M 2021), while H2O is an off-the-shelf solver (H2O.ai 2017). Conceptually, autoML algorithms compile primitives to form a solution, given some input data and problem configurations. The primitives may include data manipulations, missing value imputations, learning algorithms, etc. The D3M solvers use primitives developed by the D3M program, and are different from

Table 1. Data partitions with corresponding month IDs

	Starts	Ends	Range of *month_id*
TRAIN	Jan, 1990	Dec, 2012	121→396
VAL	Jan, 2013	Dec, 2015	397→432
TEST	Jan, 2016	Dec, 2018	433→468

those used by H2O. However, both AlphaD3M and CMU have access to the same library of primitives.

The ViEWS problem can be structured as a regression task or a time series forecasting task. To solve the problem as a regression task, we shift the target column by k time units, where k ranges from 1 to 6. This requires a distinct autoML run for each k. We solve the problem as a regression task using all three autoML systems. In addition, we solve the problem as a time series forecasting task using CMU and AlphaD3M. For this setting, we specify the spatial and temporal features (grids or countries, and month ids) and provide only the state-based violence variable as the target. In terms of parsimony, the time-series forecasting setup is the simplest. H2O did not have a time series forecasting option.

Features and Data

We test these three autoML engines across four different feature and data settings: *ViEWS, ViEWS-30, Dynamics*, and *Dynamics-Hurdle*. We begin with the data compiled by the ViEWS team, which is publicly accessible at https://github.com/UppsalaConflictDataProgram/OpenViEWS2. The number of features used in each setting is shown in Table 2. The *ViEWS* model includes all the variables that the ViEWS teams uses in their benchmark model. This includes data from the GED plus other data resources such as REIGN (Bell, Besaw, and Frank 2021) and V-Dem (Coppedge et al. 2021). The *ViEWS-30* model includes the features that ViEWS has identified as the most important (Jansen2020).

The *Dynamics* model includes variables intended to capture spatial and temporal conflict dynamics, plus *country* and *month* indicators. The data used for this model is limited to measures of state-based violence built from the GED, and variables describing the spatial and temporal structure of the data. No additional data resources are used. In the ViEWS data, these are the features that end with *_sb*, to indicate they were built from the state-based violence data. For example, *tlag_1_ged_best_sb* is the 1 period time lag of the number of state-based fatalities. Table 3 shows the different classes of variable transformations. In addition to spatial and temporal lags, the transformations include dummy variables to capture thresholds, "time since" variables to measure peace spells, counts to capture the

Table 2. Number of features

	CM	PGM
ViEWS	797	41
ViEWS-30	32	32
Dynamics	96	42
Dynamics-Hurdle	–	42

Table 3. Variable classes

Class	Description
Violence	Thresholds, counts, measures of violence at time t.
Time since	Number of periods since violence has been observed.
Time since spatial	Number of periods since violence has been observed in a neighbor.
Time since decay	Function to weight temporal distance since violence.
Time lag	Violence at period $t - i$ where $i > 0$.
Spatial lag	Violence in neighbors, neighbors of neighbors, etc.
Space-time distance	Spatial and temporal distance to violence as a single number.
Time lag spatial lag	Violence in neighbors at period $t - i$ where $i > 0$.
Rolling max	Rolling maximum of violence in a time window.
Onset	Onset of violence in a time window.

number of fatal events as opposed to the number of fatalities, and space-time measures that reflect distance from a conflict event using both time and space. More details about the transformations may be found in the ViEWS documentation, Section 3.1.10 Transforms (https://views.pcr.uu.se/download/docs/views.pdf).

There are two main ideas underlying the Dynamics model. One, there has been extensive research suggesting the importance of spatial and temporal dependencies for understanding and forecasting all types of conflict (Boehmke, Chyzh, and Thies 2016; Dorff, Gallop, and Minhas 2020; Harff 2003; Sandler et al. 2007; Weidmann and Ward 2010). So, one underlying idea is to model conflict as an endogenous process, but without assuming we know what that endogenous process really looks like. The goal is for these features to capture the intricate and diverse ways that conflict is driven by conflict. The second consideration is to prevent overfitting to exogenous factors. Overfitting is always a concern in machine learning, and more of a concern with autoML. Since conflict is a rare event, and relatively rarer at disaggregate spatial and temporal levels, overfitting becomes even more of an issue.

The *Dynamics-Hurdle* model is only for the PGM forecasts. It contains the same predictors as the *Dynamics* model, but drops all grids that experienced no state-based violence in the training and validation data. Of the 10,677 total grids, 9,406 never experience any conflict and so we train these models without these grids, and set the predicted value to 0 in the test set. See the Appendix for more details.

Empirical Results

The primary empirical results are presented in Table 4. Each cell is an individual autoML run, and reports the MSE on the test set using the best performing model on the validation set. From these results we draw several conclusions.

One, every autoML run performs better than the the benchmark in the PGM setting, and nearly every run performs better than the benchmark for

Table 4. MSE for regression experiments

		PGM (MSE*100)						CM (MSE*10)					
Time lag		1	2	3	4	5	6	1	2	3	4	5	6
Benchmarks		2.96	3.05	3.11	3.19	3.25	3.29	8.57	7.93	7.49	7.74	7.21	7.92
AlphaD3M	Dynamics	2.27	1.84	1.82	1.85	2.19	2.18	5.07	5.19	6.13	6.37	6.26	6.64
	Dynamics-Hurdle	1.80	1.84	1.88	1.93	1.96	1.99	NA	NA	NA	NA	NA	NA
	ViEWS-30	2.71	2.26	2.24	2.39	2.35	2.38	4.92	5.47	6.47	6.02	8.53	6.29
	ViEWS	2.31	2.31	2.42	2.39	2.37	2.41	7.02	7.19	7.00	7.17	7.39	7.28
CMU	Dynamics	1.82	1.97	1.94	1.92	1.84	1.89	4.99	5.83	6.03	6.17	6.77	6.90
	Dynamics-Hurdle	1.81	1.86	1.89	1.93	1.95	1.98	NA	NA	NA	NA	NA	NA
	ViEWS-30	1.93	1.92	1.90	1.94	1.99	1.92	4.81	5.61	6.87	7.91	7.25	7.33
	ViEWS	2.20	2.28	2.31	2.30	2.26	2.31	8.11	8.35	8.26	8.17	8.46	8.26
H2O	Dynamics	1.73	1.78	1.79	1.79	1.83	1.86	4.82	5.37	5.96	6.63	7.25	7.35
	Dynamics-Hurdle	1.79	1.84	1.89	1.94	1.99	2.02	NA	NA	NA	NA	NA	NA
	ViEWS-30	1.79	1.86	1.91	1.97	1.99	1.94	5.11	5.15	6.30	7.14	7.63	7.98
	ViEWS	1.78	1.85	1.86	1.88	1.90	1.92	6.80	6.61	6.72	6.92	6.77	7.00

CM. Furthermore, many of the improvements are substantial. For example, the H2O *Dynamics* model for $k = 1$ represents a 42% reduction in MSE from the benchmark.

Two, for each autoML solver, the *Dynamics* models are generally an improvement over the others. This is consistently the case for H2O in the PGM setting. We speculate that this is true because the *Dynamics* approach models the conflict process itself and does not overfit to exogenous features that are in the *ViEWS* and *ViEWS-30* feature sets. While the *Dynamics-Hurdle* contains the same set of features, dropping the zero conflict cases decreases performance for H2O, improves performance for AlphaD3M, and is mixed for CMU.

Three, H2O out-performs CMU and AlphaD3M in both PGM and CM settings. We expect that this is because H2O includes a Stacked Ensemble as a solution, while CMU and AlphaD3M do not. In each run, the Stacked Ensemble is the best performing model for H2O. It also presents more intuitive results in that the MSE tends to increase consistently as the forecast period increases. Assuming that we should be better at forecasting the level of violence next month than the month after, this makes sense. CMU and AlphaD3M do not show the same consistency. For example, we can score each autoML solver 1 point if the MSE increases as the time-step increases, for a possible total of 7 points. H2O scores a 5 while CMU scores a 2 and AlphaD3M scores a 1.

There are other notable differences in our experience with the H2O and the D3M systems.[2] H2O is easier to install and is more user-friendly. It has simple options to reduce S by allowing the researcher to specify a, the learning algorithm. D3M systems have this feature, but it requires extensive expertise to use. This feature may be used for direct comparisons to

[2]Experiences are at or before the time of writing.

Table 5. MSE across time series specifications

		PGM
AlphaD3M	Markov Blanket MLP	0.014
	Gradient Boosting	0.015
CMU	Random Forest	0.029
	Vector Autoregression	0.190
H2O	NA	NA

existing findings, or because a researcher may see a benefit in some of the algorithm's statistical properties. On the other hand, the D3M systems are more capable than H2O in terms of the diversity of tasks they can solve. Both CMU and AlphaD3M have time-series forecasting options, while H2O does not.

We solved the ViEWS problem using the time-series forecasting setting, which means inputting to the autoML system a dataset that contains only three variables: the outcome, the time period, and the cross-section (i.e., *country* for CM and *grid id* for PGM). The results from these runs are shown in Table 5. Using MSE, the best performing solution is actually AlphaD3M's Markov blanket multilayer perceptron model in the time-series forecasting setting. However, when inspecting the predictions, they are constant within grid and only vary across grids. The predicted value does not change after observations of new or intensified conflict, which is not the case with the other PGM solutions. So, we do not expect these solutions to generalize well. This highlights a primary concern with using autoML, which is that the algorithms will find solutions that are overfit regardless of how careful one is to structure inputs. Thus, while it is good practice to always look beyond the performance measure and inspect the predictions, it is essential practice when using autoML.

Each of the autoML systems have features to explore the results, and we found D3M's Pipeline Profiler very effective (Ono et al. 2020). The tool visualizes the search space of the D3M systems, showing different primitives in different solutions. Figure 2 shows the search at the CM level for the CMU solver, with solutions 1 through 12 ordered from best to worst. One interesting finding from Figure 2 is that the random forest algorithm, a bagging method, is third best for CM. It is the eighth best for PGM (see Appendix). Boosting methods, including gradient boosting, xgboost, and adaboost, are the best performing. For the H2O solver, the best performing model is the stacked ensemble, but the best base learner is the gradient boosting machine. This suggests that, if researchers are going to use a single machine learning algorithm instead of an ensemble, better results may be obtained with boosting methods than bagging methods (Hastie, Tibshirani, and Friedman 2011).

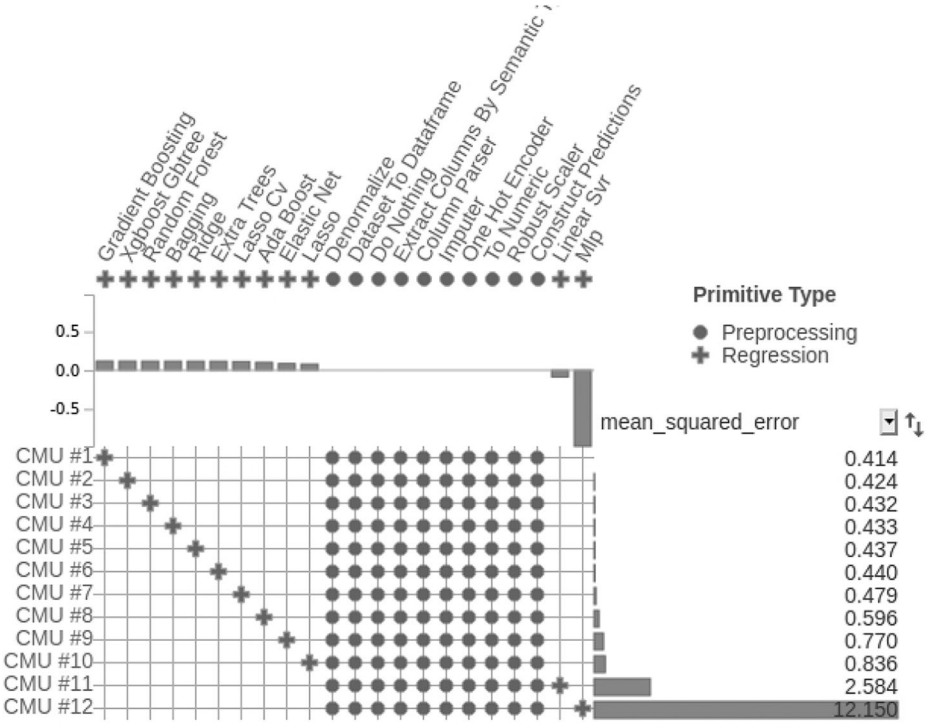

Figure 2. Pipeline profiler for solution comparison using CMU and the CM forecasting problem for 1 month out forecasts.

Forecast Exploration

Based on the empirical results, we selected the H2O and Dynamics model to produce the forecasts for the ViEWS competition. In this section we explore this model for both PGM and CM settings. Sections 5.1 and 5.2 describe the true, *a priori* forecasts that were submitted in September 2020 for October through March, 2021. To assess variable importance, Section 5.3 uses the partitions described in Table 1 and the leading contributor to the stacked ensemble, the gradient boosting machine.

PGM Forecasts

Figure 3 shows the level of forecasted violence across Africa in October 2020 and March 2021. While there are many grids with potential for violence, there are roughly seven clusters: (1) central Mali and Burkina Faso; (2) northwestern Nigeria; (3) Lake Chad region including Borno in northeastern Nigeria, Chad, and Cameroon; (4) western Cameroon; (5) the borders of the DRC and Uganda, Rwanda, and Burundi; (6) southern Somalia; and (7) the Cabo Delgado province in northwestern Mozambique.

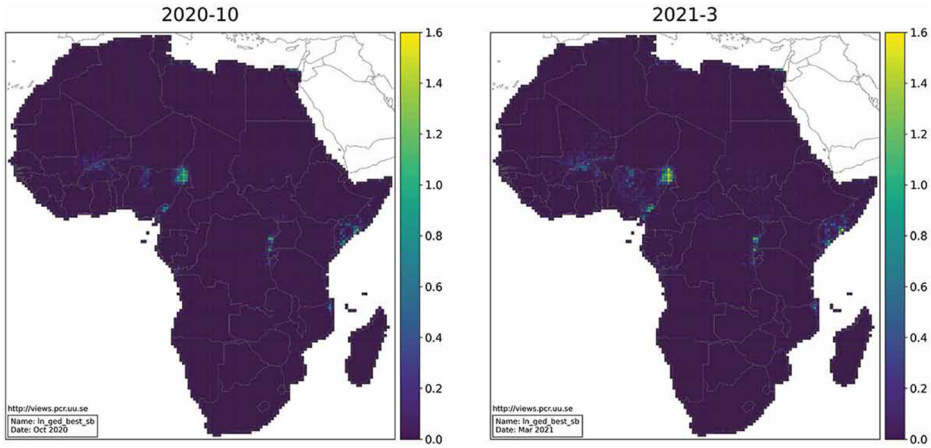

Figure 3. Oct and Mar PGM forecasts.

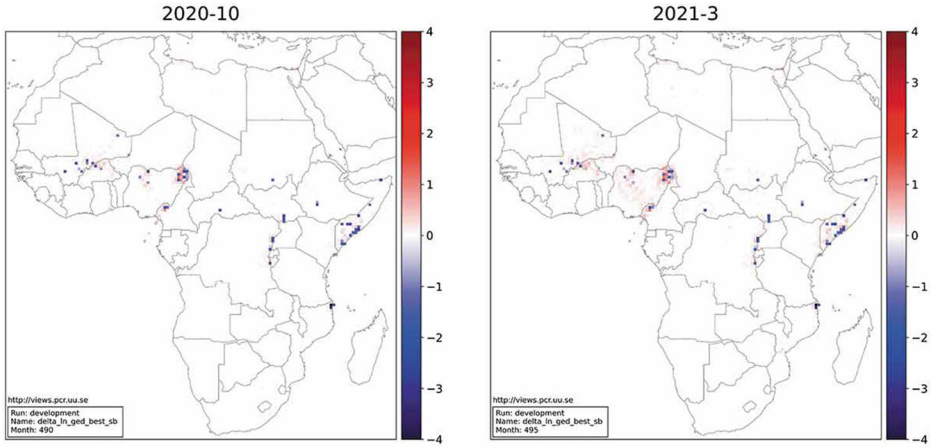

Figure 4. Oct and Mar PGM delta.

Comparing forecasted violence for October and March in Figure 3, violence is expected to increase and spread slightly. Figure 4, which shows the forecasted change in violence, reveals the expected spread in violence more clearly. Blue and red cells indicate expected decreases and increases, respectively. While many grids are blue, there are more red cells in March than in October.

Figure 5 focuses on the Lake Chad region and the Mozambique and Tanzania border area. This figure shows observed and forecasted violence from October 2019 to March 2021. With multiple grids in each region, the values are the sum of violence across grids. Points to the right of the vertical dashed line are forecasts, while points to the left are observed.

The blue line is Nigeria, Cameroon, and Chad in the areas near Lake Chad. This region has been the site of violence associated mainly with

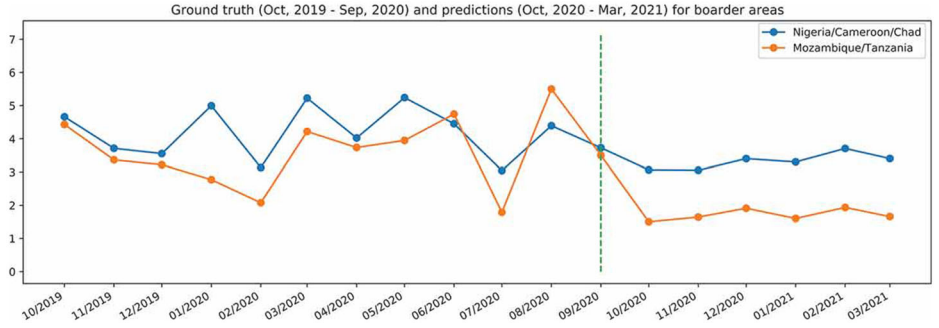

Figure 5. Figure PGM Border.

Boko Haram. The range of observed values is from 3.04 to 5.24. The forecasted values are within this range, but at the lower end, ranging from 3.06 to 3.71. Thus, the model forecasts a continuation of violence in this region, but toward the lower end of what has been seen over the last year.

The red line is along the eastern border of Mozambique and Tanzania and includes Mozambique's Cabo Delgado province. This region has seen a surge in extremist violence from groups potentially linked to Islamic State (Blake 2020; Campbell 2020). The observed amount of state-based violence in this region is slightly lower than that of the Lake Chad region, and the forecasted violence for October to March is lower as well. However, the forecast is at or below observed violence. This suggests a larger expected decrease over the next six months than in the Lake Chad region.

Both of these regions are conflict prone, as shown by the consistent presence of state-based violence in the first twelve months of Figure 5. Given the Dynamics model is focused on modeling the process itself, the forecast of continued violence is intuitive and suggests the model is behaving the way it is intended.

CM Forecasts

Figure 6 shows the forecasted level of conflict for each state in October and March for the CM models. Similar to the PGM model, the forecasted increase in violence is greater for March than October. For some states, however, this still represents a decrease from the observed level in September. As seen in Figure 7, which shows the change in violence, several states are blue. This includes places that have seen higher levels of state-based violence, including Nigeria, Somalia, and Mozambique. In general, North Africa is expected to see increased violence, as shown by red states of Libya, Chad, Niger, and Algeria. Egypt is white for October, which means no change, but red for March.

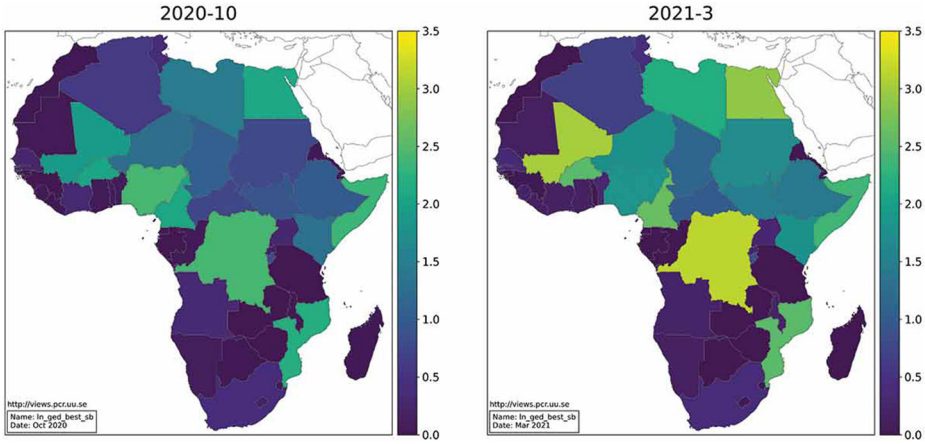

Figure 6. Oct and Mar CM forecasts.

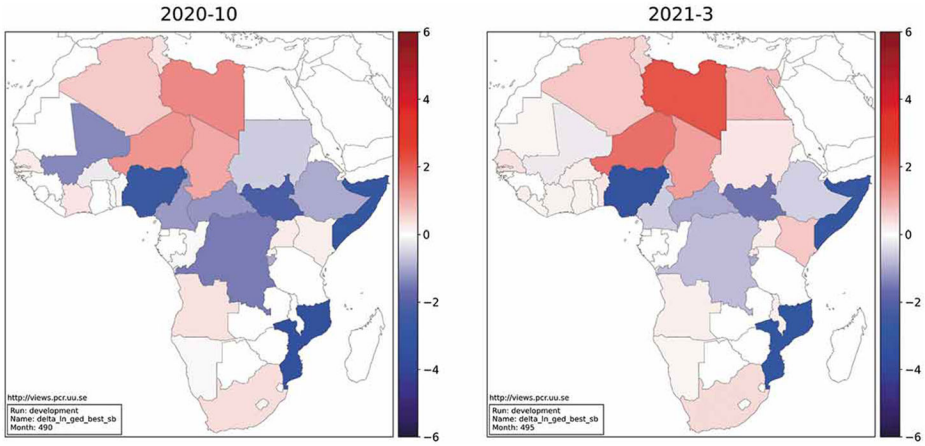

Figure 7. Oct and Mar CM delta.

Figure 8 shows observed and forecasted violence using the country-month model for five states: Nigeria, Cameroon, Chad, Mozambique, and Egypt. The observed violence and forecasted violence differs in some interesting ways, and these results are somewhat less intuitive than those presented for the PGM model. The observed violence varies within state, for example Chad has both no violence and, in April 2020, the highest level of violence at 6.96. These high levels of violence in Chad correspond to an offensive against Boko Haram and extremists by the Chadian government in the Lake Chad region (Harwood 2020). The forecasted violence has lower variance and shows slight trends by state. Nigeria is trending down while Chad and Egypt are trending up. Furthermore, forecasted violence across states is not drastically different. If Chad is excluded, the forecasts

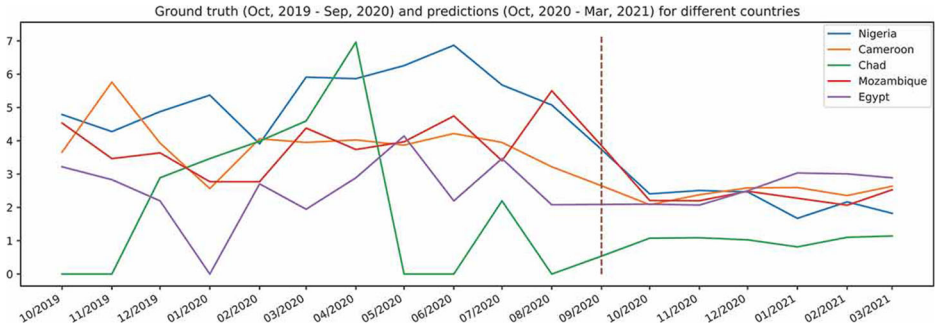

Figure 8. Figure CM.

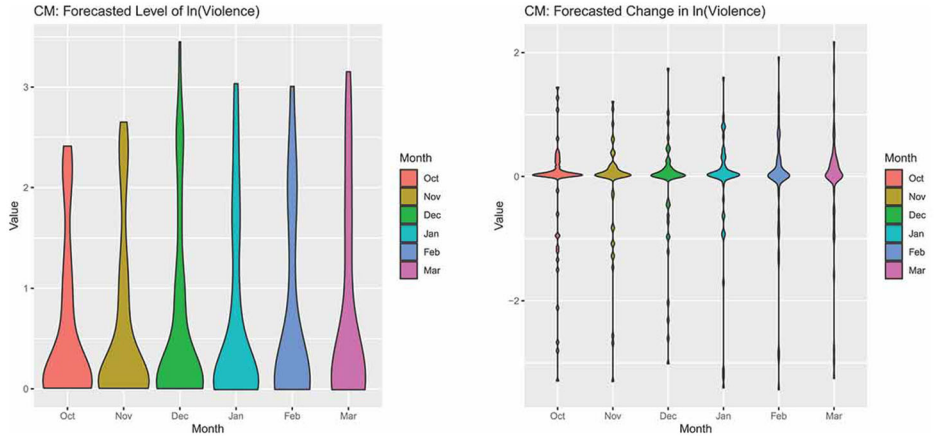

Figure 9. Oct and Mar CM delta.

for the other four states are extremely similar, especially for October through December.

Figure 9 shows the distributions of forecasted levels and change in violence for the CM models. Looking at the right panel, forecasted increases are smaller and fewer in number than forecasted decreases. However, no violence and no change in violence is the most common forecast. Looking at the left panel, conflict is generally expected to increase over these six months, as mass at zero decreases and shifts upward. While this is not true for every state and region, both CM and PGM models forecast more conflict on the continent in March than October. The PGM version of Figure 9 is in the Appendix, and results are similar to those discussed here.

Variable Importance

Variable importance is measured by calculating the relative influence of each variable. Assume each data point contains M features, $X_i = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}\}$

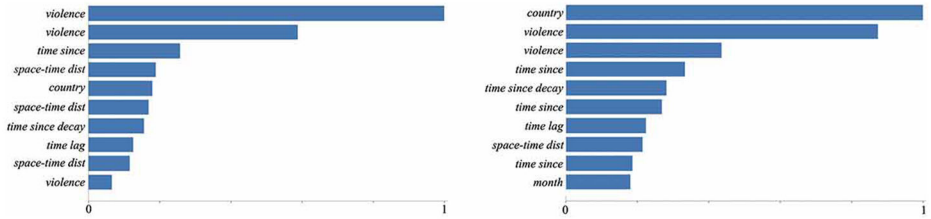

Figure 10. PGM Variable Importance for $s = 1$ (Left) and $s = 6$ (Right).

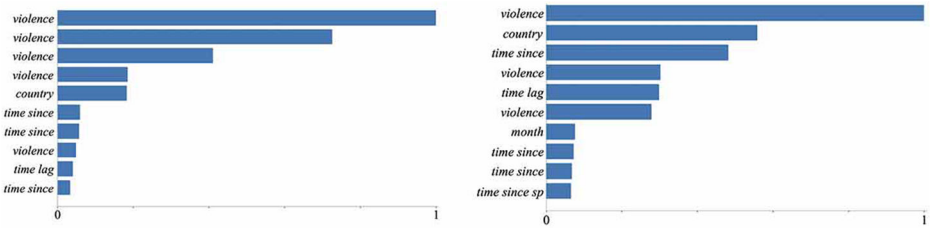

Figure 11. CM variable importance for $s = 1$ (Left) and $s = 6$ (Right).

and we want to assess the importance of a specific feature x_s. Its variable importance score is computed by the following formula.

$$VI(x_s) = \left[\sum_{i=0}^{N}(y_i - \bar{y})^2\right] \times N - \sum_{v \in x_s}\left[N_v \times \sum_{j=0}^{N_v}(y_j - {}_{\bar{y}}v)^2\right]$$

where N denotes the number of data points, v is a unique value in x_s, and N_v denotes the number of instances where $x_s = v$. Figures 10 and 11 show variable importance scores for the leading base model in the stacked ensemble, the gradient boosting machine, because H2O did not support variable importance for the stacked ensemble. See H2O.ai (2017) for more details on how variable importance is assessed with H2O.

Figure 10 shows the importance for the top ten variables for the $s = 1$ and $s = 6$ PGM models.[3] While each bar represents the importance score for an individual variable in the model, the figures are labeled with the class of variable transformation as described in Table 3. Looking at the left panel, the top two features have very high importance scores relative to the

[3]The top ten in the left panel are: *ged_count_sb, ln_ged_best_sb, time_since_greq_25_ged_best_sb, stdist_k1_t10_ged_dummy_sb, country_name, stdist_k1_t1_ged_dummy_sb, decay_12_time_since_ged_dummy_sb, tlag_2_ged_dummy_sb, stdist_k1_t001_ged_dummy_sb, greq_5_ged_best_sb*. The top ten in the right panel are: *country_name, ged_count_sb, greq_1_ged_best_sb, time_since_greq_25_ged_best_sb, decay_12_time_since_ged_dummy_sb, time_since_greq_500_ged_best_sb, tlag_5_ged_dummy_sb, stdist_k1_t001_ged_dummy_sb, time_since_ged_dummy_sb, month*.

others. These are *ged_count_sb* are *ln_ged_best_sb*, which are the count of fatal events and the log fatalities. They are both measures of violence within the grid at time t. Hirose, Imai, and Lyall (2017, 54) state, "the conflict modeling literature has demonstrated that a leading predictor of future violence is simply the prior distribution of violence," and the results here support this.

The plot in the right panel shows variable importance for the $s = 6$ PGM model. The top two features, *country* and *ged_count_sb*, have considerably higher importance scores than the rest. The importance of *country* could explain why we see more increases in violence for the March forecasts in Figure 4. Essentially, at six months out, the model forecasts conflict to occur within a state, but has greater uncertainty as to the location of the violence and so we see more light red cells. At one month out, however, the forecasted violence is more concentrated. For $s = 6$ forecasts, features 3 through 10 have greater importance than in the $s = 1$ model. This may also be related to the uncertainty of 6 month forecasts.

Variable importance for the top ten predictors in the CM models are shown in Figure 11.[4] As with PGM, for $s = 1$ the two most valuable contributors are *ged_count_sb* and *ln_ged_best_sb*. The top two for $s = 6$, however, are flipped: *ged_count_sb* is first and *country_name* is second. As with PGM, features 3 through 10 in the $s = 6$ model have more importance than in the $s = 1$ model.

One takeaway from these figures is the consistent importance of *ged_count_sb*, the count of fatal events. It is the most valuable predictor in three of the models and the second most important in the other. While we often see the total number of fatalities or the log fatalities, it is much less common to see the count of fatal events in forecasting models. On the whole, the *violence* class of features contributes most. This speaks to the value of measuring conflict in different ways and including those measures in the forecasting model.

Another takeaway is the lack of overall importance for spatial variables. There is not a single spatial lag in Figure 10, and only one in Figure 11, where it is the tenth most important for $s = 6$. In the PGM models, there are four space-time variables that may be capturing a spatial dynamic, but we do not know to what degree since the distance combines elements of

[4]The top ten in the left panel are: *ged_count_sb*, *ln_ged_best_sb*, *ged_best_sb*, *greq_5_ged_best_sb*, *country_name*, *time_since_greq_25_ged_best_sb*, *time_since_greq_5_ged_best_sb*, *ged_summy_sb*, *tlag_4_greq_25_ged_best_sb*, *time_since_greq_100_ged_best_sb*. The top ten in the right panel are: *ged_count_sb*, *country_name*, *time_since_greq_5_ged_best_sb*, *ln_ged_best_sb*, *tlag_1_greq_5_ged_best_sb*, *greq_5_ged_best_sb*, *month*, *time_since_greq_100_ged_best_sb*, *time_since_ged_dummy_sb*, *time_since_greq_500_splag_1_1_ged_ best_sb*.

both space and time. For the CM models, however, the spatial dynamic appears largely absent.

Conclusion

The ViEWS problem is to forecast changes in the level of state-based violence for each of the next six months at the PRIO-GRID and country level. For this competition and toward the goal of improving sub-national and country level forecasts, we experiment with combinations of automated machine learning (autoML) systems and limited datasets that emphasize the endogenous nature of conflict. Two core findings emerge from our analysis. One, autoML improves predictive performance, and the stacked ensemble method in H2O performs best. Two, the Dynamics model provides the most accurate forecasts. These findings hold at both the PGM and CM levels, although they are stronger at the PGM level. As a result of our experiments, we selected the Dynamics model with the H2O solver to enter into the ViEWS forecasting competition. At the PGM level, this model won for "predictive accuracy" and split the win with Lindholm et al. (2022) for "originality" (Vesco et al. 2022).

If researchers are interested in predictive performance and do not have a particular reason for choosing one learning algorithm over another, our results indicate that they should default to the autoML design and use a system with the stacked ensemble method. H2O is one such system, and is publicly available and open-source. As feature sets change from one forecasting problem to another, there is no guarantee that the stacked ensemble will perform best. However, the autoML design will provide the researcher with a greater opportunity to select a high performing model.

In addition to improved predictive performance, the Dynamics model has other advantages. The data used for this model is limited to measures of state-based violence built from the GED, and variables describing the spatial and temporal structure of the data. Therefore, the approach is applicable to any type of conflict and any level of aggregation. It is interpretable as a model that forecasts conflict as a function of present and past conflict. This is a simpler interpretation compared to models with different sets of exogenous variables, even when those variables are themed (Hegre et al. 2021). It makes for a good benchmark since it models the endogenous nature of conflict itself, without assuming we know what that endogenous process is (Cranmer and Desmarais 2017). It is also relatively easy to build, since no additional data resources with exogenous variables are required.

We expect conflict forecasting models that couple advanced autoML systems with variables that reflect a diverse set of conflict dynamics to have

high predictive performance, especially at sub-national and sub-annual levels of aggregation. Future research could explore this in applications to different types of political violence. The ViEWS data provide a number of conflict variables that we used in the Dynamics model, but they are not the only transformations possible. Future research could also explore other transformations.

Acknowledgments

We would like to thank Mike Colaresi, Håvard Hegre, and Paola Vesco for their work organizing the Violence Early Warning System prediction competition and workshop. We would also like to thank the competition scoring committee, Nils Weidmann, Adeline Lo, and Gregor Reisch, and all competition and workshop participants.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8750-17-2-0114. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

Beck, N., G. King, and L. Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94 (1): 21–35.

Bell, C., C. Besaw, and M. Frank. 2021. "The Rulers, Elections, and Irregular Governance (REIGN) Dataset." *One Earth Future*. Access 11 June 2020. https://oefdatascience.github.io/REIGN.github.io/.

Bergstra, J., and Y. Bengio. 2012. "Random Search for Hyper-parameter Optimization." *The Journal of Machine Learning Research* 13 (1): 281–305.

Blair, R. A., and N. Sambanis. 2020. "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning." *Journal of Conflict Resolution* 64 (10): 1885–915. doi:10.1177/0022002720918923

Blake, J. 2020. "Preventing the Next Boko Haram in Northern Mozambique." Accessed 2020 Dec 14. https://www.cfr.org/blog/preventing-next-boko-haram-northern-mozambique

Boehmke, F. J., O. Chyzh, and C. G. Thies. 2016. "Addressing Endogeneity in Actor-specific Network Measures." *Political Science Research and Methods* 4 (1): 123–49. doi:10.1017/psrm.2015.34

Bowlsby, D., E. Chenoweth, C. Hendrix, and J. D. Moyer. 2020. "The Future Is a Moving Target: Predicting Political Instability." *British Journal of Political Science* 50 (4): 1405–17. doi:10.1017/S0007123418000443

Brandt, P. T., J. R. Freeman, and P. A. Schrodt. 2014. "Evaluating Forecasts of Political Conflict Dynamics." *International Journal of Forecasting* 30 (4): 944–62. doi:10.1016/j.ijforecast.2014.03.014

Campbell, J. 2020. "The Military-First Approach in Northern Mozambique Is Bound to Fail." Accessed 2020 Dec 14. https://www.cfr.org/blog/military-first-approach-northern-mozambique-bound-fail

Carter, D. B., and C. S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18 (3): 271–92. doi:10.1093/pan/mpq013

Chenoweth, E., and J. Ulfelder. 2017. "Can Structural Conditions Explain the Onset of Nonviolent Uprisings?" *Journal of Conflict Resolution* 61 (2): 298–324. doi:10.1177/0022002715576574

Colaresi, M., and Z. Mahmood. 2017. "Do the Robot: Lessons from Machine Learning to Improve Conflict Forecasting." *Journal of Peace Research* 54 (2): 193–214. doi:10.1177/0022343316682065

Coppedge, M., J. Gerring, C. H. Knutsen, S. I. Lindberg, and J. Teorell. 2021. "V-Dem [Country–year/country–date] Dataset V11.1." *Varieties of Democracy Project.* doi: 10.23696/vdemds21.

Cranmer, S. J., and B. A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25 (2): 145–66. doi:10.1017/pan.2017.3

D'Orazio, V. 2020. "Conflict Forecasting and Prediction." *Oxford Research Encyclopedia of International Studies* doi: 10.1093/acrefore/9780190846626.013.514.

D3M. 2021. "Data-Driven Discovery of Models." https://datadrivendiscovery.org/

Dorff, C., M. Gallop, and S. Minhas. 2020. "Networks of Violence: Predicting Conflict in Nigeria." *The Journal of Politics* 82 (2): 476–93. doi:10.1086/706459

Drori, I., Y. Krishnamurthy, R. Rampin, R. D. P. Lourenco, J. P. Ono, K. Cho, C. Silva, and J. Freire. 2018. "AlphaD3M: Machine Learning Pipeline Synthesis." In *Proceedings of the ICML Workshop on Automatic Machine Learning* Stockholm, Sweden, pp. 1–8.

Gelpi, C., and N. Avdan. 2018. "Democracies at Risk? A Forecasting Analysis of Regime Type and the Risk of Terrorist Attack." *Conflict Management and Peace Science* 35 (1): 18–42. doi:10.1177/0738894215608998

Gleditsch, K. S., and M. D. Ward. 2013. "Forecasting Is Difficult, Especially about the Future: Using Contentious Issues to Forecast Interstate Disputes." *Journal of Peace Research* 50 (1): 17–31. doi:10.1177/0022343312449033

Goldsmith, B. E., and C. Butcher. 2018. "Genocide Forecasting: Past Accuracy and New Forecasts to 2020." *Journal of Genocide Research* 20 (1): 90–107. doi:10.1080/14623528.2017.1379631

Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54 (1): 190–208. doi:10.1111/j.1540-5907.2009.00426.x

H2O.ai. 2017. June. "H2O AutoML." H2O version 3.30.0.1.

Harff, B. 2003. "No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder since 1955." *American Political Science Review* 97 (1): 57–73. doi:10.1017/S0003055403000522

Harwood, A. 2020. "After Lake Chad Offensive, April One of Deadliest Months in Boko Haram Conflict." *Council on Foreign Relations*. Accessed 2020 Dec 14. https://www.cfr.org/blog/after-lake-chad-offensive-april-one-deadliest-months-boko-haram-conflict

Hastie, T., R. Tibshirani, and J. Friedman. 2011. *The Elements of Statistical Learning*. 2nd ed. New York, Springer.

Hegre, H., C. Bell, M. Colaresi, M. Croicu, F. Hoyles, R. Jansen, M. R. Leis, A. Lindqvist-mcgowan, D. Randahl, and E. G. Rød. 2021. "ViEWS2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System." *Journal of Peace Research* 58 (3): 599–611. doi:10.1177/0022343320962157.

Hegre, H., J. Karlsen, H. M. Nygård, H. Strand, and H. Urdal. 2013. "Predicting Armed Conflict, 2010–2050." *International Studies Quarterly* 57 (2): 250–70. doi:10.1111/isqu.12007

Hegre, H., M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Högbladh, and R. Jansen. 2019. "ViEWS: A Political Violence Early-warning System." *Journal of Peace Research* 56 (2): 155–74. doi:10.1177/0022343319823860.

Hegre, H., P. Vesco, and M. Colaresi. 2022. "Lessons from an Escalation Prediction Competition." *International Interactions* 48 (4).

Hill, D. W., and Z. M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108 (3): 661–87. doi:10.1017/S0003055414000306

Hirose, K., K. Imai, and J. Lyall. 2017. "Can Civilian Attitudes Predict Insurgent Violence? Ideology and Insurgent Tactical Choice in Civil War." *Journal of Peace Research* 54 (1): 47–63. doi:10.1177/0022343316675909

Hutter, F., L. Kotthoff, and J. Vanschoren. 2019. *Automated Machine Learning: Methods, Systems, Challenges*. Cham, Switzerland: Springer Nature.

Hyndman, R. J., and G. Athanasopoulos. 2018. *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. New York, NY: Springer.

Jansen, R., H. Hegre, M. Colaresi, and F. Hoyles. 2020. *Benchmark Models for the ViEWS Prediction Competition*. Uppsala, Sweden: ViEWS.

Jones, Z. M., and Y. Lupu. 2018. "Is There More Violence in the Middle?" *American Journal of Political Science* 62 (3): 652–67.

King, G., and L. Zeng. 2001. "Improving Forecasts of State Failure." *World Politics* 53 (4): 623–58. doi:10.1353/wp.2001.0018

Lagazio, M., and T. Marwala. 2006. "Assessing Different Bayesian Neural Network Models for Militarized Interstate Dispute: Outcomes and Variable Influences." *Social Science Computer Review* 24 (1): 119–31. doi:10.1177/0894439305281512

Lindholm, A., J. Hendriks, A. Wills, and T. B. Schön. 2022. "Predicting political violence using a state-space model." *International Interactions* 48 (4).

Minhas, S., P. D. Hoff, and M. D. Ward. 2016. "A New Approach to Analyzing Coevolving Longitudinal Networks in International Relations." *Journal of Peace Research* 53 (3): 491–505. doi:10.1177/0022343316630783

Ono, J. P., S. Castelo, R. Lopez, E. Bertini, J. Freire, and C. Silva. 2020. "PipelineProfiler: A Visual Analytics Tool for the Exploration of AutoML Pipelines." *IEEE Transactions on Visualization and Computer Graphics* 27 (2): 390–400. doi:10.1109/TVCG.2020.3030361

Rost, N., G. Schneider, and J. Kleibl. 2009. "A Global Risk Assessment Model for Civil Wars." *Social Science Research* 38 (4): 921–33. doi:10.1016/j.ssresearch.2009.06.007

Sandler, T. and W. Enders. 2007. "Applying Analytical Methods to Study Terrorism." *International Studies Perspectives* 8(3): 287–302.

Schutte, S., and N. B. Weidmann. 2011. "Diffusion Patterns of Violence in Civil Wars." *Political Geography* 30 (3): 143–52. doi:10.1016/j.polgeo.2011.03.005

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–32. doi:10.1177/0022343313484347

Tollefsen, A. F., H. Strand, and H. Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49 (2): 363–74. doi:10.1177/0022343311431287

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). doi:10.2202/1544-6115.1309.

Vesco, P., H. Hegre, M. Colaresi, R. B. Jansen, A. Lo, G. Reisch, and N. B. Weidmann. 2022. "United They Stand: Findings from an Escalation Prediction Competition." *International Interactions* 48 (4). doi:10.1080/03050629.2022.2029856

Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. "The Perils of Policy by P-value: Predicting Civil Conflicts." *Journal of Peace Research* 47 (4): 363–75. doi:10.1177/0022343309356491

Ward, M. D., R. M. Siverson, and X. Cao. 2007. "Disputes, Democracies, and Dependencies: A Reexamination of the Kantian Peace." *American Journal of Political Science* 51 (3): 583–601. doi:10.1111/j.1540-5907.2007.00269.x

Ward, M. D. 2016. "Can We Predict Politics? Toward What End?" *Journal of Global Security Studies* 1 (1): 80–91. doi:10.1093/jogss/ogv002

Weidmann, N. B., and M. D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6): 883–901. doi:10.1177/0022002710371669