New Approaches to Forecasting Civil Violence and Geopolitical Crises

Mahda Soltani* Jeremy Springman* Erik Wibbels*

April 6, 2025

Abstract

Forecasting the onset of geopolitical crises, such as armed conflict and state failure, presents a significant challenge due to the rarity of these events, the difficulty of predicting onsets as distinct from ongoing crises, and the lack of high-frequency data on covariates that might capture rapidly evolving political and social dynamics. This study addresses these challenges by introducing three key innovations. First, we leverage data from the Machine Learning for Peace (MLP) project, which provides monthly data on a wide range of civic space activity that we expect to enhance the accuracy of onset forecasts. Second, we propose a novel forecasting target: U.S. Department of State travel advisories, which provide a broader measure of geopolitical risk by encompassing diverse crises, including political violence, health emergencies, and natural disasters. Third, we introduce methodological advancements, including temporal cross-validation tailored to time-series data, to improve model robustness and interpretability. Our findings demonstrate that the combination of high-frequency data on civic space and our modeling approach significantly improve the predictive performance of both conflict and State Department travel warning onset forecasts.

1 Introduction

The capacity to anticipate the onset of geopolitical crises is a long-standing goal of academic and policy research (cites). The ability to forecast events such as civil conflict and state failure not only informs theoretical debates about the drivers of instability (cites) but also shapes the international community's capacity to respond effectively to emerging crises (cites). As indicated by the global refugee crisis, the consequences of

*PDRI-DevLab@Penn, University of Pennsylvania

civil instability in one country can ripple across borders. This interdependence has intensified both scholarly efforts to better understand the dynamics of conflict and policymakers' reliance on more sophisticated early-warning systems (cites to EWSs). Researchers and decision-makers alike seek accurate forecasting models to preempt and mitigate the far-reaching consequences of geopolitical disruptions, deepening our academic understanding of these crises while providing practical tools for coordinated, timely responses Dowding 2021; Lewis-Beck and Stegmaier 2014; Schrodt 2014.

Forecasting efforts, however, face two core challenges. First, accurately predicting the onset of events – as distinct from the continuation or escalation of ongoing crises – is inherently difficult. Civil war (and other crisis) onsets are rare and complex phenomena that often result from a confluence of factors, making them harder to model than the persistence of conflicts once they have started (cite to Bazzi et al). Forecasting onsets is a persistent academic challenge, while being crucial for citizens and policymakers who would benefit from advance warning to plan and implement preventive measures. Conflict onset also necessitates different data inputs and methodological approaches than those used for predicting the continuation of established events (cite). Since onsets are temporally sensitive and result from complex dynamics, they require modeling approaches that can capture temporal dependencies and adapt to evolving social dynamics Bergmeir and Benítez 2012. Second, there has historically been a lack of fine-grained, high-frequency data on civic and political indicators that could improve onset forecasts. Datasets aggregated at broader temporal intervals, such as annually, may overlook the weekly and monthly changes in political and social environments that serve as proximate indicators of impending crises Schrodt 2014; Cederman and Gleditsch 2009. Additionally, models focusing primarily on macro-level events might obscure localized dynamics critical to understanding conflict onset, affecting the timeliness and accuracy of forecasts Pierskalla and Hollenbach 2013; Salehyan et al. 2012.

To demonstrate the importance of addressing these issues, we focus on the prediction of armed conflict – one of the most prevalent tasks in the forecasting literature. While many studies have achieved reasonable accuracy in forecasting the incidence of conflict, predicting the onset of conflict has proven more challenging. We improve upon standard approaches in three ways. First, we use data from the Machine Learning for Peace (MLP) project to introduce a host of new monthly civic and political covariates (cite). The MLP dataset draws on high-quality, national media reporting and two machine classifiers to capture high-frequency measures of civic space activity, such as arrests, defamation cases, declarations of martial law, purges of the bureaucracy, etc. which we expect to provide leverage in predicting conflict onset.

Second, for the first time in conflict forecasting we implement temporal cross-validation, which is specifically adapted for panel data and particularly effective for time-series forecasting (cites). Unlike traditional cross-validation, temporal cross-validation respects the temporal integrity of the data both in model training and performance assessment. This approach can mitigate the risk of overfitting to patterns that might not accurately reflect temporal processes; it also avoids reliance on performance metrics tuned to particular temporal windows of data when conflict onsets might be rare, which is common with conflict onsets. Our approach delivers both robust predictive performance and a high degree of interpretability, making it a valuable approach for both academics and policymakers.

Third, we combine the MLP data and these methodological innovations to forecast a new high-frequency target variable, namely high-level U.S. Department of State (DOS) travel warnings, which inform U.S. citizens about potential risks abroad U.S. Department of State. Such warnings are important because they prompt the DOS to mobilize extensive resources to locate U.S. citizens in the affected country, coordinate travel arrangements, and, in extreme cases, close embassies. Beyond their real-world importance, travel advisories have a number of analytically attractive features. First, they are issued by a common source (the U.S. Department of State) that attempts to harmonize what kinds of events warrant elevated travel warnings. Second, they summarize a broad range of political and social crises, encompassing not only political violence but also other threats such as natural disasters, health crises, crime, and social unrest. Forecasting severe travel warnings, thus, represents an attempt to target a wider-range of geopolitical crises than is typical in conflict forecasting. As far as we know, we are the first to systematically collect this data, which is well archived on the internet but difficult to scrape and aggregate.

Our analysis yields three key findings. First, the inclusion of granular, monthly data on civic dynamics significantly enhances the ability to predict the onset of crises. New predictor variables such as declarations of martial law, lagged electoral activity, and censorship all contribute to forecast accuracy. These results underscore the value of introducing more nuanced civic covariates and echo recent work by (cite to JPR paper) on the role of protest activity in foreshadowing civil conflict. Second, the implementation of temporal cross-validation provides more accurate and realistic assessments of model performance compared to traditional evaluation methods. Our models demonstrate strong predictive performance with high interpretability of key predictors. Third and finally, we show that high-level DOS travel advisories are amenable to forecasting. Like conflict onsets, these are rare events, but they have the advantage of capturing a broader range of security threats and offering a new line of inquiry in conflict forecasting. These three findings also combine to offer a new round of insights for policymakers intent on reducing the costs and incidence of geopolitical crises.

In the sections that follow we begin by situating our study in the broader research on conflict forecasting. This review highlights the state of the literature and identifies gaps that our research aims to address. Thereafter, we describe our three innovations in greater detail. We then outline our data and modeling choices before presenting the results and discussing feature importance. We conclude with a discussion of the implications of our findings for academic forecasters and policy practitioners, including limitations and avenues for future research.

2 Literature Review

Predicting geopolitical crises—including civil conflicts, state failures, and major political upheavals—has a long history in both academic and policy arenas (cites). Accurate forecasts serve a dual role: they deepen theoretical understanding of crisis dynamics and inform pragmatic strategies for prevention, resource allocation, and peacekeeping (Redl and Hlatshwayo 2021; Duursma and Karlsrud 2019). Despite sustained attention, crises remain notoriously difficult to anticipate, often emerging from complex, rapidly evolving social, economic, and political conditions.

Within this broader field, forecasting the incidence of armed conflict has been particularly prominent. By studying where conflicts already exist, researchers have identified core structural risk factors—such as poverty, weak governance, and ethnic grievances—that make some regions more prone to sustained or escalating violence (Fearon and Laitin 2003; Salehyan and Gleditsch 2006; Beck, King, and Zeng 2000). This focus on incidence reflects both practical and analytical considerations: active conflicts generate clearer and more regular data (e.g., battle events, casualties, troop movements) than brand-new conflicts, which tend to flare up suddenly and unpredictably. Over time, incidence forecasting has become integral to policy responses. International organizations and governments rely on these models to prioritize peacekeeping deployments and humanitarian relief, particularly in conflict "hot spots" with long-standing security issues. Consequently, the conventional wisdom in conflict forecasting often centers on understanding and projecting ongoing violence—where to expect continuations, escalations, or expansions to new locations (Gleditsch 2007; Hegre, Nygård, and Landsverk 2021).

Beginning in the early 2000s, the development of fine-grained data markedly enhanced conflict incidence forecasting. Projects like the Armed Conflict Location & Event Data Project (ACLED) gathered daily or weekly event-level information on political violence, making it possible to track the geographic and temporal progression of attacks, protests, and riots (Rød, Hegre, and Leis 2023). Meanwhile, advances in geospatial imagery allowed analysts to detect forced population displacement or devastated infrastructure, both of which help gauge the severity of active conflicts (Witmer 2015; Goodman, BenYishay, and Runfola 2024). Likewise, unstructured digital media—such as social media posts or online news—began to supply near-real-time evidence of events on the ground, capturing changing public sentiment or incident reports from conflict zones more promptly than official statistics (Mueller and Rauh 2017). These data innovations have given researchers more precise ways to model ongoing conflict dynamics, often down to the district or city level.

Consequently, risk assessments for active warfare—where violence is already visible—became increasingly detailed and spatially precise.

In parallel, researchers adopted novel computational methods to capitalize on these richer datasets. Ensemble learning (e.g., random forests, gradient boosting) and neural networks improved predictive accuracy by uncovering nonlinear interactions among multiple predictors (Brandt et al. 2022; Malone 2022). Spatio-temporal convolutional networks and recurrent neural networks (RNNs), in particular, handled sequential data adeptly, spotlighting how conflict activity evolves over time and migrates across regions (Hegre et al. 2021). These methodological advances, combined with probabilistic modeling techniques, have yielded sophisticated risk forecasts for areas already experiencing conflict. Policymakers increasingly consult these models to evaluate the likelihood of an ongoing war intensifying, spreading to neighboring districts, or subsiding. In short, the arrival of granular data and new machine-learning approaches solidified incidence forecasting as a core pillar of conflict research, providing ever more accurate near-term predictions of continued violence.

However, forecasting the onset of brand-new conflicts remains significantly more challenging. While incidence-focused models have grown increasingly precise, they do not necessarily address the rarity and suddenness of new outbreaks—a task that raises distinct analytical hurdles and requires different data inputs. Notwithstanding these advances, conflict onsets remain exceedingly rare, often under 1% of observations. Many studies—historically centered on active conflict—still validate models using conventional cross-validation or short out-of-sample tests that do not fully respect the temporal structure of data or the sparsity of new outbreaks (**cederman2017**; **blair2017**; Ward 2002). These methods may inadvertently produce inflated performance metrics, particularly if the validation windows contain few onsets or if future data sneak into the training process. As a result, while incidence forecasting enjoys considerable empirical support, the question of how to rigorously evaluate predictive models for rare new crises lingers.

A second gap relates to the limited availability of truly high-frequency civic data. While projects like ACLED provide aggregated information on violent events, tracking of non-violent indicators of social or political instability—such as waves of arrests or legal crackdowns—varies by dataset and may be underreported in some contexts Raleigh, Kishi, and Linke 2023. This leaves a gap for detecting subtle, fast-moving shifts in civic space that might presage new conflicts. Third, conflict studies overwhelmingly focus on armed conflict as their outcome of interest. This narrow lens can ignore broader geopolitical crises—from health emergencies to natural disasters and large-scale criminal violence—that sometimes destabilize governments as much as war does. Policymakers often need early-warning models capturing any crisis that fundamentally threatens stability, not solely those with defined fatality thresholds.

3 Three Innovations in Geopolitical Forecasting

We introduce three innovations to the conflict forecasting literature: first, we deploy a set of new monthly covariates measuring a rich array of civic activities across more than 60 countries; second, we introduce a new target variable to the conflict forecasting community, namely monthly high-level Department of State (DOS) travel warnings; and third, we introduce temporal cross-validation as an improved approach to assessing forecasting models of geopolitical crises.

First, we introduce a broader set of high-frequency measures of 'civic space' as predictors of conflict onset. We define civic space as the day-to-day practices through which citizens, civic actors and governments organize, communicate with, and act upon each other. As noted most recently by X, conflictual domestic dynamics are likely to be important precursors to broader civil conflicts. Of course, such changes in politics and society – like protest movements, waves of arrests, corruption scandals, press restrictions, government purges, and legal changes – often emerge quickly, and it is exactly this feature that seems likely to contribute to forecasting conflict *onset*. Unfortunately, most of the standard indicators that researchers have to track civic space are measured annually. The Varieties of Democracy project (**V-DEM**), the Civil Society Organization Sustainability Index (FHI360/USAID 2019) and the World Justice Project's Rule of Law Index (World Justice Project 2019) (among many others) provide indices capturing freedom of the press (Freedom House 2017), the rule of law (World Justice Project 2019), the ease of civic organizing (FHI360/USAID 2019; CIVICUS 2019) and beyond, but they are only available annually.

The Machine Learning for Peace (MLP) project addresses that limitation by applying recent advances in natural language processing (NLP) to a large corpus of more than 120 million digital news articles from international, regional, and mostly national sources across 64 countries. The project identifies 20 types of civic space events, ranging from protests and legal changes to bureaucratic purges, censorship and the mobilization of security forces to provide rich, monthly data on a wide range of civic space activities.[1] By providing monthly data, MLP echos projects that conflict forecasters are familiar with, including the Armed Conflict Location Event Data Project (ACLED), Uppsala Conflict Data Project Georeferenced Event Dataset (UCDP GED), and the Political Event Classification, Attributes, and Types dataset (Polecat). MLP expands upon those projects by providing data on many civic events beyond violence and protests and offers the opportunity to assess how a wider range of civic dynamics condition conflict onset.

Second, we introduce U.S. Department of State (DOS) travel advisories as a target variable. Armed conflict data, while valuable for analyzing violent instability, focuses narrowly on events that meet high-intensity thresholds, such as fatalities. This clear, narrow scope has advantages, but it comes at the cost

[1] See the Technical Report on the Production of Civic Space and RAI Event Count Data for details on the project's corpus, classifier, and classifier performance.

of excluding a wide range of instability with potentially significant geopolitical implications, ranging from drug cartel violence to terrorist attacks to disease outbreaks. Travel advisories, by contrast, provide a broad measure of geopolitical risk. Indeed, travel advisories are designed to communicate a wide range of risks to U.S. citizens in foreign countries, including political instability, terrorism, health crises, natural disasters, and crime. Analysts at U.S. embassies and consulates gather information from diverse sources and coordinate with consular affairs officers in Washington, DC to integrate this information with broader geopolitical insights to produce regular assessments of security conditions. To the extent travel advisories incorporate a broader range of social, political, economic and health risks, they might even be precursors to the armed conflicts that have received the lion's share of forecasting attention.

Beyond being informative about a broad range of geopolitical threats, travel advisories provide several attractive features as a target variable for forecasting. First, unlike datasets that aggregate information from many different sources (such as news media) travel advisories are issued by a single, centralized authority: the U.S. Department of State. This consistency ensures that the criteria and thresholds used to evaluate risks remain rather standardized across all advisories.[2] The uniformity of travel advisories enhances their reliability as a target variable, enabling researchers to focus on forecasting without the additional complexity of data harmonization. Second, in contrast to annual conflict datasets, travel advisories are updated monthly to ensure they reflect contemporary conditions. The DOS' commitment to timely updates ensures that travel advisories align closely with evolving risks, thereby providing a high-frequency target variable. This feature also aligns the outputs of such models with the immediate needs of policymakers and stakeholders, who rely on these advisories for decision-making. Lastly and relatedly, these high-level travel advisories have significant real-world implications because they prompt significant operational responses. Their issuance leads the DOS to mobilize extensive resources, including locating U.S. citizens in affected areas, coordinating emergency travel arrangements, and, in extreme cases, initiating embassy closures and relocating consular operations (U.S. Department of State, n.d.).

Third and finally, we introduce temporal cross-validation to geopolitical forecasting. We begin with a discussion of two limitations of the traditional approach to model training and evaluation in this area of forecasting and then introduce temporal cross-validation as a means of addressing those limitations. Out-of-sample evaluation involves training models on historical data and testing them on separate, future time periods (Hyndman and Athanasopoulos 2018) . In this approach, the dataset is typically divided into two segments: a training set, which includes historical observations used to develop and fit the predictive model,

[2] Our conversations with staff in DOS who work on travel advisories indicate that considerations such as the number of US citizens in a country and potential political implications of travel warnings *do* impact these decisions at the margins. Nevertheless, these considerations can be modeled whereas considerations such as the extent of urban bias of multiple media sources in a country are hard to discern and generally not.

and a test set, comprising unseen observations. The model uses patterns and relationships learned from the training data to generate predictions for the test set. Out-of-sample evaluation provides an indication of how well the model generalizes to unseen data.

Out-of-sample evaluation is particularly challenging in the context of rare-event forecasting, as standard models struggle with class imbalance and event rarity (Shyalika, Wickramarachchi, and Sheth 2024). Conflict onset (and high-level travel warnings) are very rare indeed. In UCDP Georeferenced Event Dataset (UCDP GED) v.24.1, conflict dataset, for instance, conflict onsets account for only 1% of the observations. The rarity of conflict onset events, especially when combined with the paucity of historical data, often results in short out-of-sample periods containing few or no instances of conflict onset (Ward, Greenhill, and Bakke 2010). The predominance of negative (no-conflict) cases in such datasets can bias probability estimates downward in traditional statistical models, leading to misleading inference about event likelihood (King and Zeng 2001). For example, if a model predicts no conflict during a two-year out-of-sample period and no conflict occurs, the performance metrics might suggest high accuracy, even though the model has not demonstrated its ability to predict actual conflict onsets.

Another widely used approach—k-fold cross-validation—poses challenges when applied to rare-event forecasting in time-ordered data. K-fold cross-validation partitions the dataset into k subsets, or 'folds,' where the model is trained on $k-1$ folds and tested on the remaining fold, repeating this process k times so that each fold serves as the test set once (Kohavi 1995). While this method is effective for evaluating model performance on cross-sectional datasets, it fails to account for the temporal dependencies inherent in time series data, leading to information leakage and misleading performance estimates (Bergmeir and Benítez 2012). Specifically, standard k-fold cross-validation assumes that data points are independently and identically distributed (i.i.d.), an assumption violated in time-ordered data where past observations influence future values. If training and test sets are randomly partitioned, the model is likely to learn from future observations, invalidating the predictive evaluation (Bergmeir and Benítez 2012). This leakage undermines the evaluation's realism and fails to simulate the conditions under which forecasting models operate in practice, where only past information is available for predicting future events. Thus, k-fold cross-validation is ill-suited for assessing models designed to forecast rare, temporally dependent phenomena like conflict onsets.

To address limitations in evaluation methods, we implement temporal cross-validation (CV), a method initially introduced by (Rodolfa, Lamba, and Ghani 2021) to evaluate predictive models specifically on time-ordered data while preserving the chronological sequence of observations. Temporal CV involves training the model on historical data up to a specified cutoff point and then testing it on subsequent data, ensuring that each test set represents genuinely unseen information. This process is repeated over successive time splits, with each iteration extending the training period further into the future. By employing these sequential, time-based splits, temporal CV establishes a rolling evaluation framework that captures both short-term and long-term predictive performance, offering a more comprehensive assessment of model robustness throughout the years.

The primary advantage of temporal cross-validation lies in its prevention of information leakage by preserving the temporal sequence of data, ensuring that future data points do not influence the model during training. This setup provides a realistic evaluation of model performance by simulating real-world conditions, where future information is inherently unavailable (Hyndman and Athanasopoulos 2018). For rare event forecasting, temporal cross-validation is particularly beneficial because its successive time splits capture trends across different years, reducing the risk of overfitting to a specific time period (Nicolò et al. 2024). This iterative structure evaluates the model across diverse temporal contexts, enhancing its adaptability to shifting socio-political conditions and increasing its reliability for forecasting scenarios where event patterns change over time.

4 Measures and Models

Our initial target variable is conflict onset, taken from UCDP/PRIO Armed Conflict Dataset (version 24.1).[3] In addition to include the MLP data as covariates, we also incorporate measures that capture the temporal and contextual background of conflict history in each country. Specifically, we include the number of months since the last conflict and a binary indicator marking whether a conflict occurred within the past 12 months. To capture underlying country-level characteristics related to conflict occurrence, we implement a Bayesian encoding technique that combines a country-specific average conflict rate with the global average rate, weighted by the number of observations for each country and a smoothing parameter, based on data up to the current date to avoid information leakage. In other words, for each observation i in country c, the smoothed encoding $E_{i,c}$ is calculated as:

$$E_{i,c} = \frac{n_i \cdot \mu_{c,i} + k \cdot \mu_g}{n_i + k}$$

where $\mu_{c,i}$ is the country-specific expanding mean up to observation i, given by

$$\mu_{c,i} = \frac{\sum_{j=1}^{n_i} y_{j,c}}{n_i},$$

and $y_{j,c}$ being the target variable for country c and n_i the number of prior observations in that country. The global cumulative mean up to observation i, μ_g, incorporates all data up to this point:

[3] They define as conflict as including at least 25 battle-related deaths.

$$\mu_g = \frac{\sum_{j=1}^{i} y_j}{i},$$

Here, k is the smoothing parameter, which mitigates the influence of isolated events in countries with limited conflict history by allowing the encoded value $E_{i,c}$ to converge toward the global mean μ_g when n_i is small. Conversely, for countries with ample historical data, the encoding reflects more specific trends by giving greater weight to the country-specific mean $\mu_{c,i}$.

To account for the economic factors that may influence conflict dynamics (Collier and Hoeffler, 2004; Hegre & Sambanis, 2006), we calculate the z-score across a range of *monthly* and *quarterly* economic indicators that characterize each country's unique economic landscape. Given that data availability varies across countries, the z-score incorporates available indicators, thereby normalizing each country's economic profile based on the indicators available for that country.

We first assess the model's performance by training it on historical data and testing it on separate, future time periods for forecast horizons of 3 and 6 months. This approach echoes standard practice in much of the conflict forecasting literature and provides an initial benchmark for the model's performance on both near-term (3 months) and mid-term (6 months) horizons. We then apply our preferred temporal CV to train and validate the models, leveraging all available data up to each time cut. To capture both immediate and gradual changes impacting conflict onset, I include 12-months of lagged independent variables, accounting for delayed effects and underlying trends as outlined in the literature (Wilkins, 2018; Guo, 2018; Hegre, 2017). For each time cut, we validate my models in 3-month intervals for both forecast horizon ($h = 3$ and $h = 6$) synchronizing each validation period with the MLP dataset's quarterly update cycle to ensure the model consistently incorporates fresh data and evaluates performance with the latest data.

As an example, consider Algeria with a 3-month horizon ($h = 3$). For the first time cut, the training data spans from January 2012 to January 2013, with lagged features extending through April 2013. The training target values (y_{train}) represent outcomes up to July 2013. To ensure temporal consistency with MLP's quarterly updates, the validation set for each time cut includes three consecutive monthly observations per country:

- **First validation row**: X_{val} includes data from May 2012 to April 2013 to predict y_{val} for July 2013.
- **Second validation row**: X_{val} spans June 2012 to May 2013 to predict y_{val} for August 2013.
- **Third validation row**: X_{val} covers July 2012 to June 2013 to predict y_{val} for September 2013.

This pattern continues for each subsequent time cut, with each training period expanding to include all data up to the time cut, while the validation set consistently includes three monthly observations aligned with the quarterly data release. Thus, for the final time cut, the training data for Algeria spans up to May 2022, with lagged features extending through April 2023, and targets up to July 2023. The validation set is structured as follows:

- **First validation row**: X_{val} includes data from August 2021 to July 2022 to predict y_{val} for October 2022.

- **Second validation row**: X_{val} spans September 2021 to August 2022 to predict y_{val} for November 2022.

- **Third validation row**: X_{val} covers October 2021 to September 2022 to predict y_{val} for December 2022.

Temporal cross-validation advances through time cuts to evaluate the model's performance in the face of evolving temporal patterns, which is crucial for rare-event forecasting. We apply the same method to the 6-month horizon (h = 6), with validation predictions extending six months into the future while maintaining the lagged structure for each forecast window.

We rely on LightGBM, a gradient-boosted tree-based machine learning algorithm. To do so, we prepare the data by transforming the panel data into a format suitable for the classifier's API. This involves creating a feature matrix X and a target vector y, where each row in X represents 12 months of lagged panel data preceding each target value. This approach ensures the model has access to sufficient historical context for making informed predictions, while strictly excluding any information from future periods to avoid data leakage while maintaining its temporal integrity.

Finally, we employ Optuna for hyperparameter tuning, an open-source framework that optimizes hyperparameters using the Tree-structured Parzen Estimator (TPE). Unlike traditional methods such as grid search or random search—which exhaustively or randomly explore hyperparameter values, Optuna's Bayesian optimization refines the search process iteratively based on previous trials. This method achieves superior performance with fewer evaluations, fine-tuning the model to better capture complex patterns in conflict onset (Akiba et al., 2019).

5 Results

5.1 Forecasting Conflict Onset

Table 1 presents the three primary performance metrics—ROC AUC, AUC PR, and Brier Score—for out-of-sample evaluation and temporal cross-validation across the 3- and 6-month forecast horizons. The ROC

AUC (Receiver Operating Characteristic Area Under the Curve) quantifies the model's ability to differentiate between positive and negative classes at various thresholds, with a comparison against the 50% chance line as a baseline (Fawcett, 2006). The AUC PR (Area Under the Precision-Recall Curve) assesses the trade-off between precision and recall across thresholds, offering insight into the model's capability to detect the minority class, here the conflict onset, particularly in imbalanced datasets. For added context, we benchmark our AUC PR against a dummy model that assigns 1s based on the distribution of positive instances within the data (Saito & Rehmsmeier, 2015). The Brier Score, in turn, evaluates the accuracy of probabilistic predictions by calculating the mean squared error between predicted probabilities and actual outcomes, serving as a crucial measure of model calibration quality (Brier, 1950). Together, these metrics provide a comprehensive evaluation of the models' classification performance.

Table 1: Performance Metrics

Evaluation Method	Horizon	ROC AUC	AUC PR	Brier Score
Out-of-Sample	3	1	1	0.02
Out-of-Sample	6	1	1	0.01
Temporal CV	3	0.68	0.03	0.07
Temporal CV	6	0.61	0.02	0.03

As described in Table 1, the standard, out-of-sample approach to evaluating model performance achieves near-perfect performance over both 3- and 6-month horizons, with ROC AUC and AUC PR scores both reaching 1, and low Brier Score of 0.02 and .01, respectively. Yet this remarkable performance is largely an artifact of the fact that conflict onset is very rare. In our test set of July through December of 2023, there was a single conflict onset across all our countries (Mali in September of 2023), and the model correctly forecast that one onset. This underscores the sensitivity of standard performance metrics in conflict forecasting to short test windows when the target is so rare.

In contrast, Figures 1 and 2 display the results for temporal cross-validation, revealing a notable decline in model performance. For the 3-month forecast horizon, temporal cross-validation yields an ROC AUC of 0.68, an AUC PR of 0.03, and a Brier Score of 0.070, indicating a substantial shift from the near-perfect performance observed in traditional out-of-sample evaluation (see Figure 1). Likewise, the 6-month results show an ROC AUC of 0.61, an AUC PR of 0.02, and a Brier Score of 0.03 (Figure 2). While these metrics are lower than those using the traditional out-of-sample approach, cross-validation provides a more thorough assessment of model performance over the entire temporal range of the data. By distributing evaluation across many temporal windows, each characterized by varying numbers of conflict onsets, temporal cross-validation provides a much fuller assessment of model performance. While the test set under the traditional approach includes a single conflict onset in 2023, temporal cross-validation encompasses 67 (3-month horizon) and

64 (6-month horizon) conflict onsets across validation windows. The increased frequency of conflict onsets underscores how temporal CV subjects the model to more dynamic and realistic assessment and, therefore, results in more modest performance metrics.

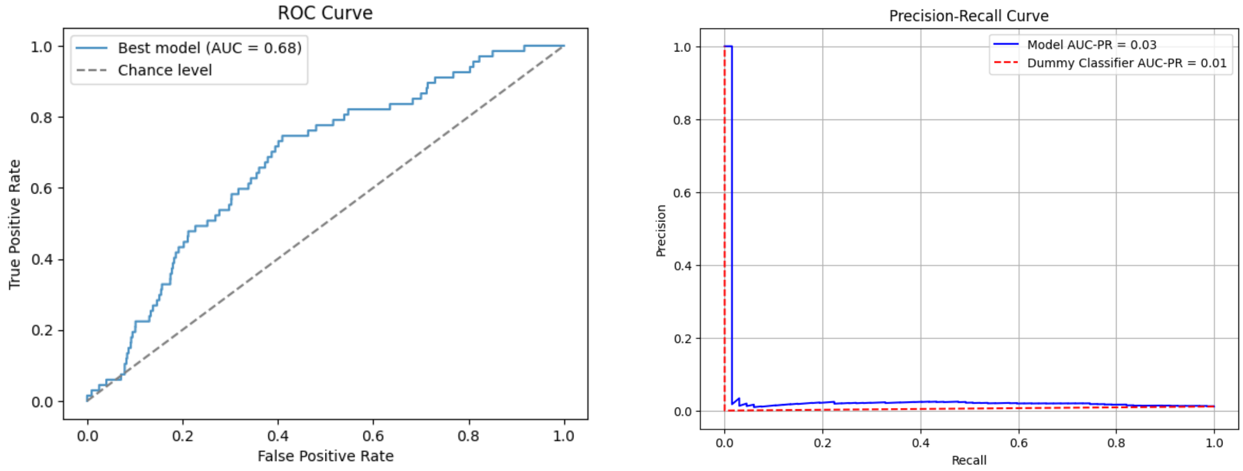


Figure 1: ROC AUC & AUC PR Curves for 3-month Temporal CV

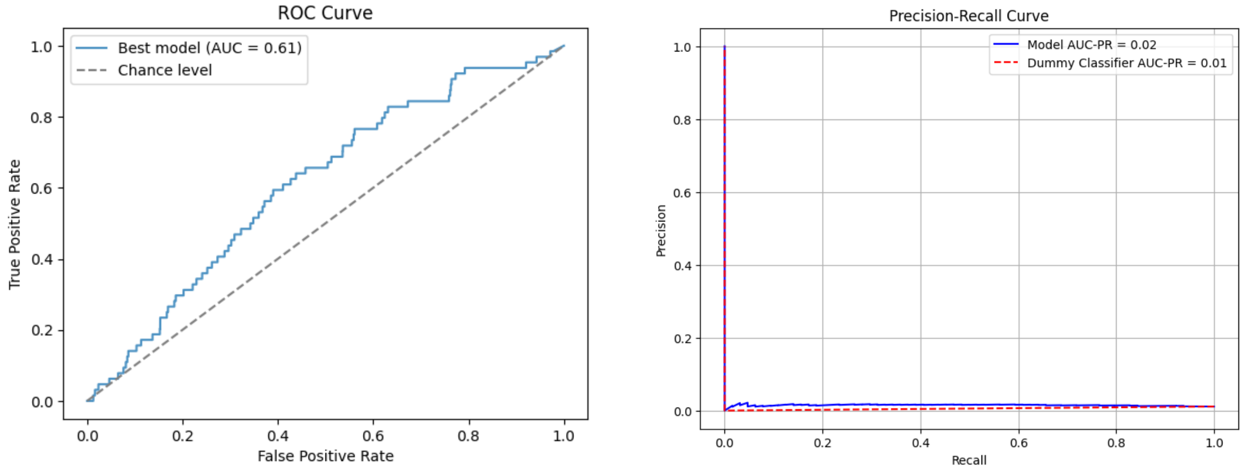


Figure 2: ROC AUC & AUC PR Curves for 6-month Temporal CV

The comparison brings forward an essential insight. While out-of-sample performance metrics can produce impressive results, they often convey a narrow understanding of model effectiveness, especially when applied to rare-event forecasting. Temporal cross-validation, with its ability to account for a wider range of events and contexts, uncovers potential vulnerabilities that may go unnoticed in simpler evaluations. By interpreting performance metrics with attention to the event distribution, we gain a more realistic and modest perspective on the model's predictive power.

5.2 Forecasting the Onset of Travel Advisories

We now shift the focus to U.S. Department of State (DoS) travel advisories as a target variable. Armed conflict data, while valuable for analyzing violent instability, focuses narrowly on events that meet high-intensity thresholds, such as fatalities. This limited scope often excludes critical early warning indicators of broader political instability. Travel advisories, by contrast, provide a richer measure of geopolitical risk, making them an intriguing new target for forecasting.

Unfortunately, there is no standard dataset on DOS travel advisories, and the DOS itself does not maintain historical data. Thus, we used the Internet Archive to compile monthly advisories issued by the DoS from 2012 to the end of 2023. The resulting data spans two distinct periods due to changes in the DoS advisory system. Before 2018, all advisories carried a general risk assessment without categorical differentiation, so we include all travel advisories from this period in my analysis. Beginning in 2018, the DoS adopted a four-level system to standardize risk communication: Level 1 (Exercise Normal Precautions), Level 2 (Exercise Increased Caution), Level 3 (Reconsider Travel), and Level 4 (Do Not Travel). These categories reflect increasing severity, with Levels 3 and 4 signaling serious risks that warrant heightened caution or the avoidance of travel altogether. For the post-2018 period, we focus exclusively on Levels 3 and 4, as these advisories represent the most severe warnings and map neatly onto travel warnings from the earlier period. We compile this information for each of the 60 countries included in the MLP dataset, encoding it as a binary indicator of whether or not a country had a travel advisory in a given month.

To model travel advisory *onsets*, we built upon the framework we develop above and deploy to model conflict onset. In addition to the MLP dataset and the standardized economic indicator z-scores that account for each country's economic dynamics, we include Bayesian encodings tailored to travel advisory onsets. Above this encoding was based on conflict onsets; here, we adapt it to reflect the historical travel advisory onset rate for each country, combined with the global average onset rate.

We include a convariate for continued advisories, which flags countries that already had an advisory onset in a previous period that continues into the current month. This measure allows the model to distinguish between the persistence of existing advisories and the onset of new ones, providing critical temporal context. Given the rarity of travel advisory onsets, which make up only 1.42% of the dataset (see Table 2), this indicator is essential for isolating the factors driving new advisories and improving predictive accuracy in identifying these rare events.

We also include a COVID-19 indicator variable, which addresses the unique impact of the pandemic on

Table 2: Summary of Travel Advisory Data

Column	1s (Count)	0s (Count)	1s (%)
Travel Advisory Incidence	2610	6070	30.07%
Travel Advisory Onsets	123	8557	1.42%

travel advisories. As shown in Figure 3, 2020 witnessed a significant spike in travel advisory onsets due to the global spread of COVID-19. To mitigate this, we include an indicator that takes a value of 1 for any country with a travel advisory explicitly linked to the pandemic, allowing the model to account for this systemic shock.

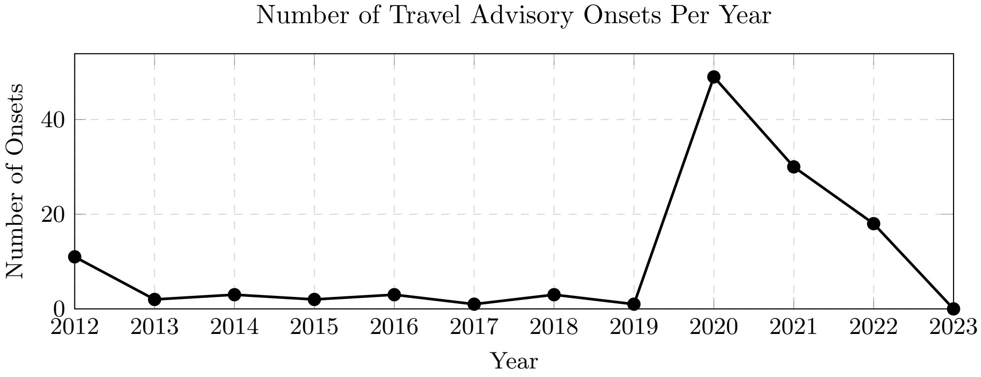


Figure 3: Number of Travel Advisory Onsets Per Year

Given the COVID shock, we also had to determine how much historical data to include in each training period for temporal CV to forecast future onsets. Using all available data might introduce noise if older observations have become obsolete or irrelevant, while using too little data could fail to capture meaningful long-term trends. To address this, we apply Bayesian Change Point Detection (BCPD) to identify shifts in the underlying trends in the target variable.

BCPD offers a probabilistic approach to identifying points in a time series where statistical properties, such as the mean or variance, shift. By modeling the data as a sequence of segments separated by these change points, BCPD estimates the posterior probability of shifts, enabling the detection of structural changes in trends over time (Adams & MacKay, 2007). This capability makes it particularly useful for addressing my problem of determining how much historical data remains relevant for predicting future outcomes. By pinpointing moments where trends change significantly—due to events like policy shifts, economic disruptions, or systemic shocks—BCPD helps identify segments of data that are most representative of the current dynamics, allowing us to avoid overfitting to outdated patterns while retaining the segments of data most valuable for forecasting. We apply BCPD to both travel advisory onset data and the incidence of travel advisories and rely on the red dashed lines in both plots, which represent significant structural shifts in the data, to inform our decision.

In Figure 4, which focuses exclusively on advisory onsets, the data exhibits stability across most of the observed timeline, with the exception of significant volatility during the 2020–2022 period that coincides with the COVID-19 pandemic. Outside this anomalous period, the onset data demonstrates a stable structure, suggesting that longer-term trends dominate the underlying dynamics.

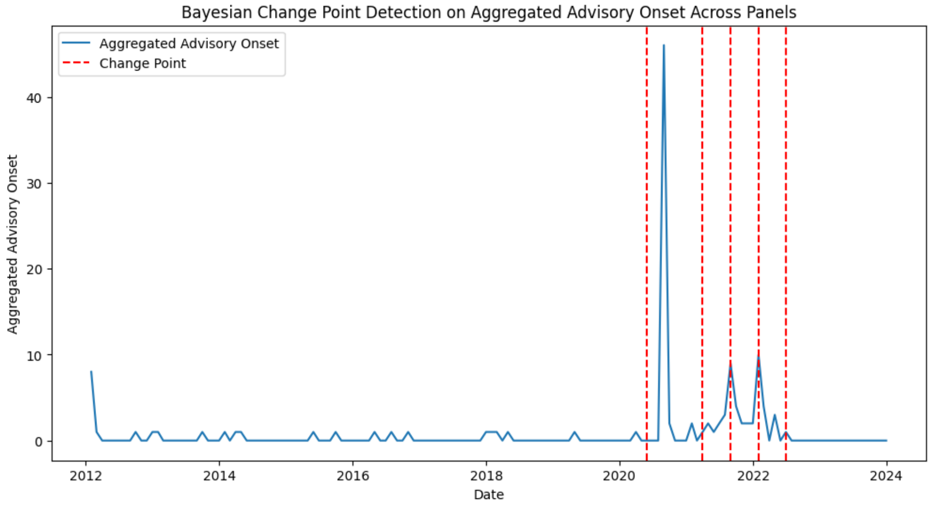


Figure 4: Onset BCPD

Figure 5, however, which considers the aggregated incidence of travel advisories, presents a different temporal dynamic. Here, the data shows periodic changes approximately every two years, indicating that travel advisories tend to follow a medium-term cyclical pattern.

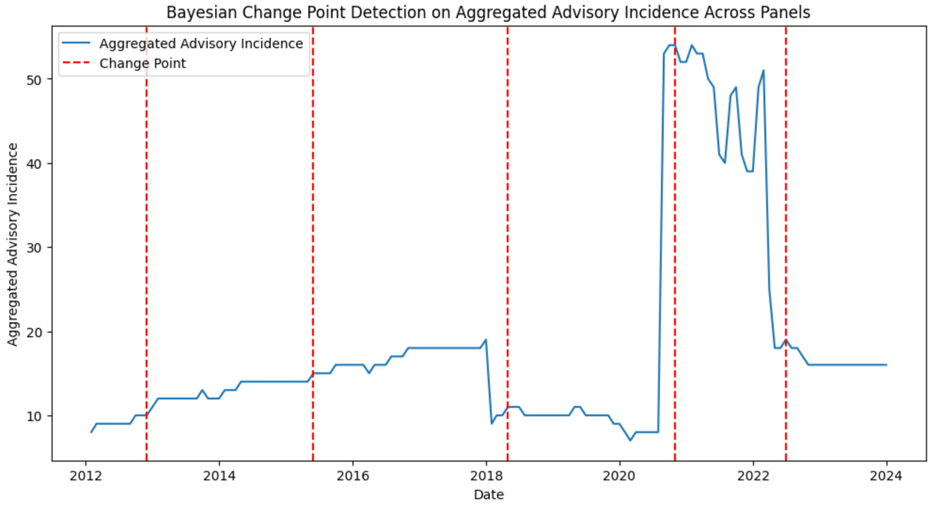


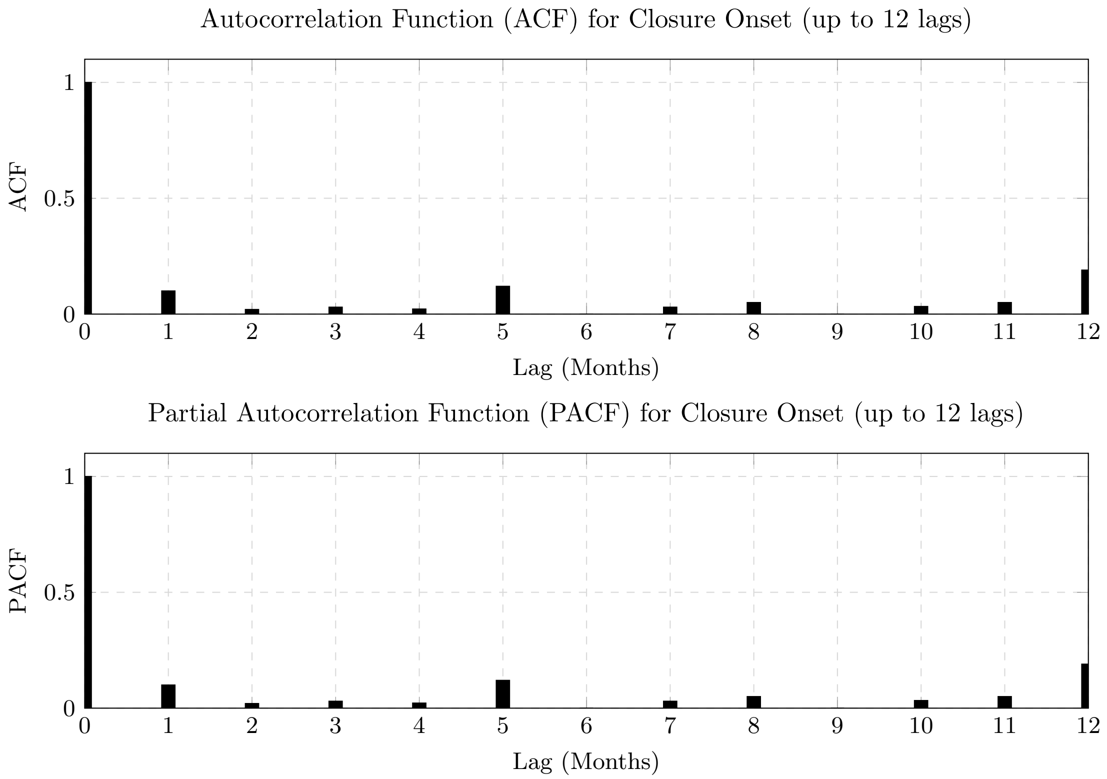
Figure 5: Incidence BCPD

This difference in temporal dynamics between the two plots suggests that the optimal training window must accommodate both the stability observed in onset data and the cyclical patterns seen in the aggregated incidence data. A training window that is too short (e.g., 2 years) might overemphasize recent fluctuations, especially during periods of volatility such as the COVID-19 pandemic, potentially reducing the model's ability to generalize to stable, long-term patterns. On the other hand, a window that is too long (e.g., 6+ years) risks diluting the impact of more recent cyclical changes, particularly the two-year periodicity evident in the aggregated data, and might introduce outdated information less relevant to current conditions. We tested 3-, 4-, and 5-year windows and evaluated their ability to balance responsiveness to recent changes and stability over time. Among these options, the 5-year training window consistently demonstrated the best performance across key predictive metrics.

We then turn to the number of lagged months to include as features in the model. While the training span focuses on the time frame of historical data used to train the model, the number of lagged months focuses on the temporal span of the model's covariates that directly influence the target variable. This is a critical consideration because lagged covariates can capture temporal dependencies and trends that may be important for predicting travel advisory onsets (Box et al., 2015; Hyndman & Athanasopoulos, 2018). Similar to the span of the training window, including too many lagged months can lead to issues such as overfitting, while using too few lagged months risks omitting relevant temporal information and degrading predictive accuracy.

To determine the most appropriate number of lagged months, we assessed the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for travel advisory onsets. The ACF measures the correlation between a time series and its lagged versions at various time steps, providing a comprehensive view of how the values in the series depend on their previous states (Box et al., 2015). The PACF, on the other hand, isolates the contribution of each individual lag by removing the effects of shorter lags. This method reveals the direct relationship between the target variable and a specific lagged value while accounting for intermediate lags (Hyndman & Athanasopoulos, 2018).

We limit the analysis to a maximum of 12 months due to the span of the dataset and the nature of travel advisory data, where older trends are less likely to be relevant for near-term forecasting. Figure 6 shows the ACF and PACF for travel advisory onset. The ACF at lag 12 shows a value close to 0.2, indicating a notable correlation between the current value of advisory onset and the value from 12 months ago and suggesting the presence of a seasonal or cyclical pattern. Similarly, the PACF at lag 12 demonstrates a significant value around 0.2, further confirming a direct and meaningful relationship between the current value and the value from 12 months ago, even after accounting for intermediate lags. Based on this analysis, we include 12 months of lagged values of advisory onset as features in the forecasting model.


Autocorrelation Function (ACF) for Closure Onset (up to 12 lags)

Partial Autocorrelation Function (PACF) for Closure Onset (up to 12 lags)

Figure 6: Autocorrelation and Partial Autocorrelation Functions for Closure Onset (up to 12 lags)

We forecast travel advisory onsets over standard 3- and 6-month horizons. Onsets are measured in two ways; in the first, the indicator variable takes on a value of 1 for the specific month of the onset. Our second measure takes on a value of 1 for the quarter that an onset happens, where we define the quarter as the month before, or, and after an onset. This rolling window expands the forecast target to include onsets that occur within a one-month margin before or after the exact forecast month, effectively capturing advisory onsets that may fall slightly outside rigid monthly boundaries.

To evaluate the models, we use temporal cross-validation to ensure robustness and reliability in my results. For the 3-month forecasts, the temporal cross-validation framework yielded 42 timecuts, with 111 onsets for the fixed window and 286 onsets for the rolling quarterly window. For the 6-month forecasts, there are 40 timecuts, with 109 onsets for the fixed month and 129 onsets for the rolling quarterly window. Consistent with our approach to conflict onset above, we use gradient-boosted trees (LightGBM) and employ Optuna for automated hyperparameter optimization.

Table 3 presents the performance of the model in terms of ROC AUC, AUC PR, and Brier Score across these horizons and onset measures. As a baseline comparison for AUC PR, the table also includes the performance of a dummy model that assigns positive predictions proportional to the distribution of positive instances in the dataset. The ROC AUC scores are consistently high across all forecasting scenarios, with values ranging from 0.87 to 0.90. These results indicate that the models possess strong discriminative power, effectively distinguishing between periods with and without travel advisory onsets. Notably, the rolling window approaches exhibit slightly higher ROC AUC scores (0.89 for the 3-month horizon and 0.90 for the 6-month horizon) compared to their fixed horizon counterparts (both at 0.87).

Table 3: Performance Metrics

Horizon	ROC AUC	AUC PR	Dummy AUC PR	Brier Score
3-Month Fixed	0.87	0.26	0.01	0.14
3-Month Rolling Window (±1 Month)	0.89	0.54	0.02	0.12
6-Month Fixed	0.87	0.31	0.01	0.14
6-Month Rolling Window (±1 Month)	0.90	0.57	0.02	0.13

The AUC PR scores highlight the challenges of predicting travel advisory onsets within a highly imbalanced dataset. The baseline dummy model achieves AUC PR values of only 0.01 for fixed horizons and 0.02 for rolling windows, emphasizing the difficulty of capturing meaningful patterns in such a sparse dataset. In contrast, my forecasting models consistently outperform the dummy model across all horizons and window types, indicating that the models successfully identify relevant signals in the data and capture the underlying dynamics of travel advisory onsets. This superior performance becomes even more pronounced when using rolling windows. While the fixed horizon models achieve AUC PR scores of 0.26 (3-month) and 0.31 (6-month), the rolling window models demonstrate significant improvements, with AUC PR scores of 0.54 (3-month) and 0.57 (6-month).

The Brier Scores, which measure the mean squared difference between predicted probabilities and actual outcomes, are relatively low across all models, ranging from 0.12 to 0.14. Lower Brier Scores indicate better calibration of predicted probabilities, suggesting that the models not only make accurate classifications but also provide reliable probability estimates.

The strong performance metrics of the models naturally lead to an exploration of which features contribute most to their success. LightGBM's feature importance rankings provide valuable insight into the relative contribution of the covariates in my models. Figures 7 display the top 10 most important features for the 3-month and 6-month horizons. The "_ number" suffix attached to each variable in the figures indicates the number of lagged months prior to the forecast period that the variable reflects. Across both horizons, MLP-derived covariates, such as "martiallaw," "electionactivity," "protest," and "censor," consistently rank among the most influential features. These covariates provide granular insights into political and societal dynamics: "martial law" captures instances of a state of emergency; "election activity" reflects electoral processes, including campaign events and voter registration drives; "protest" identifies public demonstrations of dissent or demand for action; and "censor" tracks efforts to suppress or restrict access to information, such as restrictions on journalist or website bans.

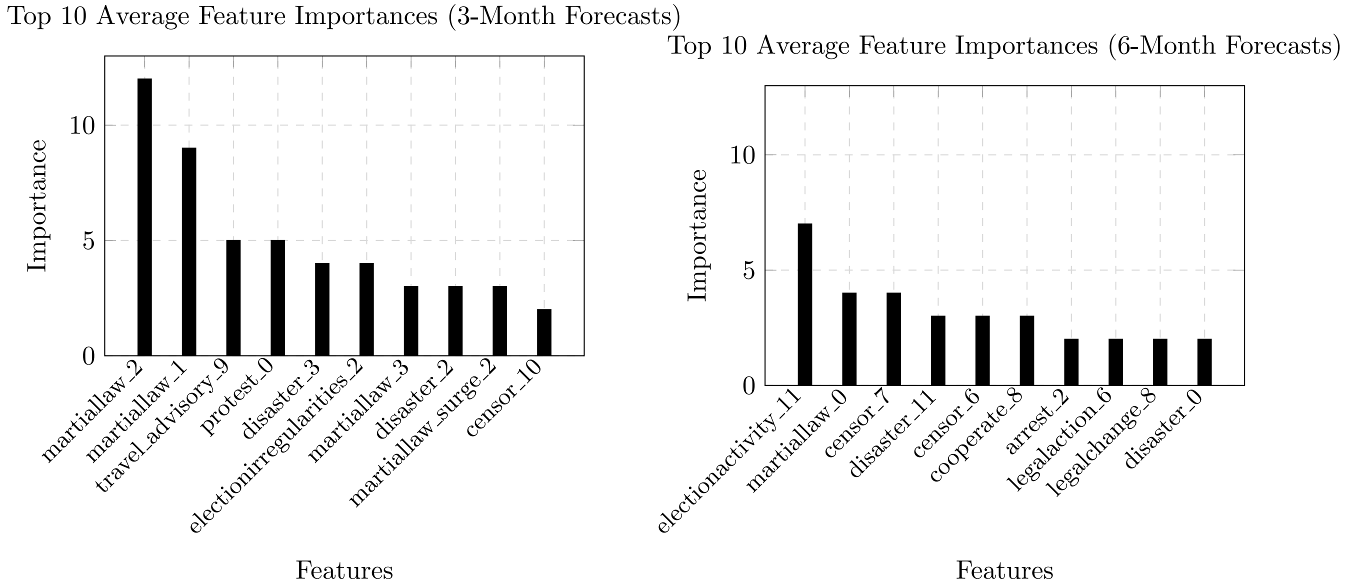


Figure 7: Top 10 Average Feature Importances for 3- & 6-Month Forecasts

Their prominence underscores the critical role these monthly indicators play in capturing dynamic patterns of instability, such as governance disruptions, electoral activity, or civil unrest, that correlate with travel advisory issuance. The high feature importance scores for MLP covariates also validate their inclusion as part of my predictive framework. These variables, by offering granular, time-sensitive insights into political and societal dynamics, augment the model's ability to interpret the underlying drivers of advisory issuance. Moreover, their performance suggests that my methodology successfully integrated these novel covariates in a way that materially contributed to the model's success. This finding not only emphasizes the value of leveraging MLP data but also supports the broader case for using such dynamic and high-frequency indicators in forecasting frameworks.

6 Conclusion

This study advances the field of conflict and crisis forecasting by introducing a novel approach that leverages high-frequency, granular data to improve the predictive accuracy of geopolitical crisis onset. By moving beyond traditional conflict datasets, which often emphasize the continuation of ongoing conflicts, and incorporating civic space activity and U.S. Department of State travel advisories as predictive targets, this research provides a more comprehensive framework for understanding and anticipating crises. Additionally, the implementation of temporal cross-validation tailored to panel data offers a robust evaluation method that addresses the unique challenges posed by rare-event forecasting, such as data imbalance and temporal dependencies.

The broader significance of this research lies in its potential to bridge the gap between theoretical inquiry and practical application. By expanding the scope of forecasting to include diverse indicators such as travel advisories, the study highlights the importance of adopting a holistic perspective on geopolitical risks, encompassing not only political violence but also health emergencies, natural disasters, and societal unrest to enhance the relevance of forecasting models for real-world decision-making. For policymakers, the ability to anticipate risks with greater precision facilitates the allocation of resources, informs preventive interventions, and supports efforts to maintain global stability.

Despite its contributions, it is imperative that I address the study's limitations. The rarity of both conflict and travel advisory onsets remains a fundamental challenge, as limited positive instances constrain model performance and generalizability. Additionally, the reliance on high-frequency media-sourced data introduces potential biases related to uneven coverage or exclusion of certain events, while the focus on U.S. Department of State travel advisories reflects the geopolitical priorities of a single actor. These factors underscore the need for caution in generalizing the findings across contexts. Furthermore, while temporal cross-validation provides a robust framework for evaluation, it is computationally intensive and may not fully capture the impact of unanticipated shocks that lack historical precedent.

Looking forward, this work sets the stage for innovative approaches to crisis forecasting. Expanding the dataset to include other actors' advisory systems or integrating regional spillover effects could provide richer insights into crisis dynamics. Further methodological advancements, such as network-based models or causal inference techniques, could uncover deeper relationships between predictors and outcomes, improving both interpretability and policy relevance. Ultimately, this study underscores the critical role of theory-informed and data-driven approaches in advancing our understanding of geopolitical risks and shaping the future of early-warning systems.