Forecasting High-Level Travel Advisories with Machine Learning for Peace Data

Mahda Soltani, Jeremy Springman, Erik Wibbels

University of Pennsylvania

The Challenge

U.S. Department of State (DOS) travel advisories are issued in response to conditions that pose significant risks to U.S. citizens abroad. These risks are often caused by a major domestic crisis, including civil unrest, armed conflict, natural disasters, or extreme repression.

Serious travel advisories impose significant costs on the U.S. government, which must ensure that US citizens in impacted countries are informed of threats to their security, facilitate last-minute travel, and in the extreme, close embassies and evacuate staff.

DOS staff often have little advanced warning of crises that prompt an advisory, which increases the cost of their response and hampers effectiveness. For this reason, DOS engaged with the Machine Learning for Peace (MLP) team to test whether MLP's high-frequency data tracking domestic political conditions can forecast the onset of high-level travel advisories.

Initial results are very promising. We can forecast the onset of serious travel warnings three and six months in the future with a high degree of accuracy. Furthermore, our models provide useful information about the political conditions on the ground that are predicting of future travel advisories.

Approach

We collected publicly available travel advisory data posted by DOS and archived on web.archive.org for 60 developing countries for 2012-2023. The key target variable we seek to predict is the onset of a level 3 or 4 travel advisory.[1] For each country, we forecast whether a new travel advisory will be issued (referred to as an onset) for each month. Onsets are rare and therefore difficult to predict (see Figure 1). However, this provides much more useful information than predicting the continuation of already existing advisories.

We combine our onset measure with monthly data on 20 Civic Space events from MLP. MLP is a flexible research infrastructure collecting data across 63 developing countries. Combining human-supervised web scraping of more than 300 high-quality domestic media outlets publishing in 40 languages with machine learning tools, MLP tracks major political events and measures their importance for domestic politics. MLP also uses predictive analytics to detect historical patterns and forecast future shifts in political conditions.

[1] The system for announcing travel advisories changed in 2018. We count all travel advisories as "serious" pre-2018; thereafter, we code level 3 and 4 advisories as serious.

Figure 1: High-Level Travel Warning Frequency. No onset includes country-months with no travel advisory or already existing advisory.

Conflict Onset	Count
No onset	8557
Onset	123

We then build models that use previous values of our political data to predict whether an onset will occur several months later. We measure the performance of our model in two ways: how often we correctly predict whether an onset occurs in a specific country-month, and how often we correctly predict whether an onset occurs within each country-quarter (three-month window).

Modeling Approach

To forecast advisory onsets, we combine LightGBM, a widely used machine learning approach to classification tasks, with temporal cross-validation, a process that helps to detect patterns in historical data that are likely to generalize to future data. We use Optuna for hyperparameter tuning, ensuring that our model is well calibrated for accuracy.

We assess the performance of our models using two standard metrics. ROC-AUC evaluates the model's capacity to distinguish 'true' positive onsets (travel warnings that the model predicted and which actually happened in real life) from 'false' positive onsets (i.e. travel warnings the model predicted to happen but that did *not* occur). AUPRC reflects the relationship between 'precision' and 'recall'.[2] AUCPR is designed specifically for predicting rare events like serious travel advisories and reveals whether most cases are identified without generating a lot of false alarms. Importantly, these scores are calculated by training the model on historical chunks of our data and then testing its ability to predict onsets in later months of data that the model has never seen.

Results

Figure 2 presents performance metrics for models predicting whether an onset will occur for each specific month looking six months in the future. The ROC-AUC of 0.87 reports that, when asked to make predictions about new data that the model was not trained on, 87% of onset months received a higher probability of onset than non-onset months. This score reflects that the model generates many true positives with few false positives. Across our sample, the probabilities assigned by our model to true positive onset months were roughly three times higher than those assigned to true negative months.

[2] Precision is the share of the model's predicted positives that are true positives. Recall is the share of the true positives in the data that the model correctly classified as true positives.

The AUCPR of 0.31 also indicates solid performance. For example, a model that knows the correct number of months with an onset but then randomly guesses the months in which those onsets occur achieves an AUCPR of only .01.

Figure 2: ROC-AUC and AUPRC scores when predicting the specific months in which advisories will occur. The distance between solid and dashed lines captures the improvements of predictions from our model over random chance.

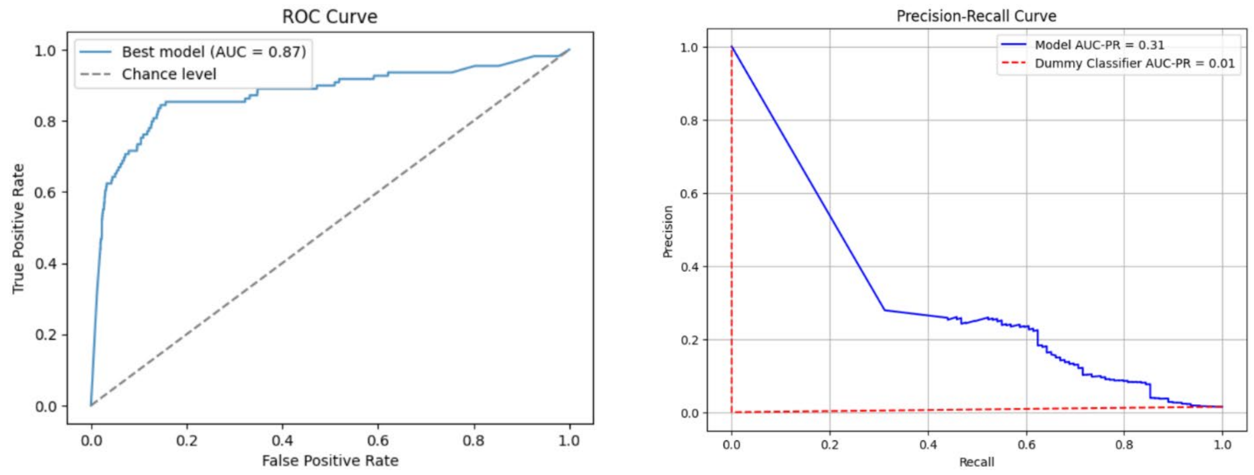
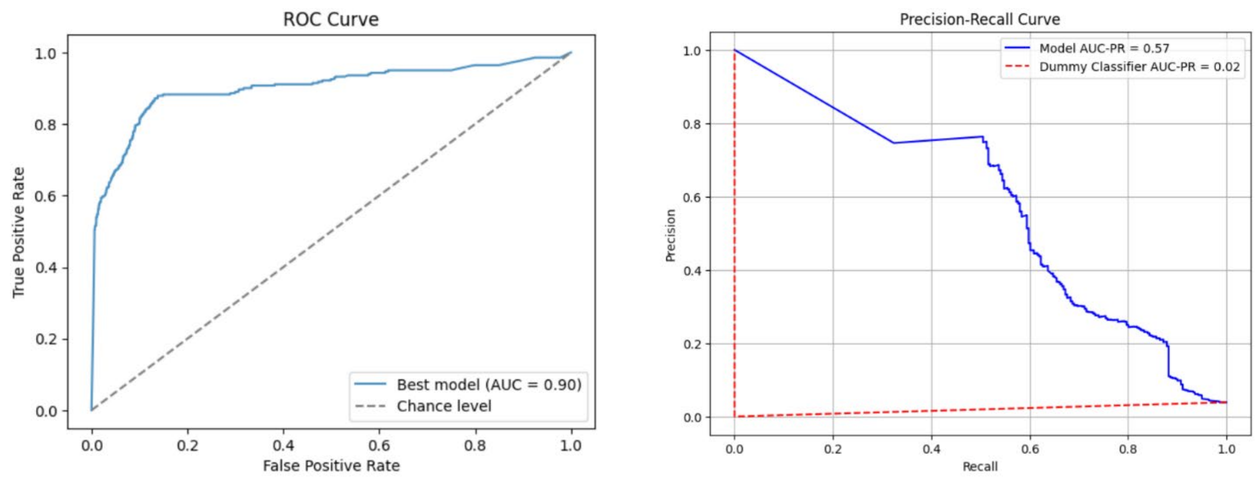


Figure 3 presents the results when trying to predict onsets within a 3-month window. The ROC-AUC score improves slightly from 0.87 to 0.9. However, the AUCPR score jumps from 0.31 to 0.57 (the model that randomly guesses improves from .01 to only .02).

Figure 3: ROC-AUC and AUPRC scores when predicting the quarter (3-month window) in which advisories will occur. The distance between solid and dashed lines captures the improvements of predictions from our model over random chance.



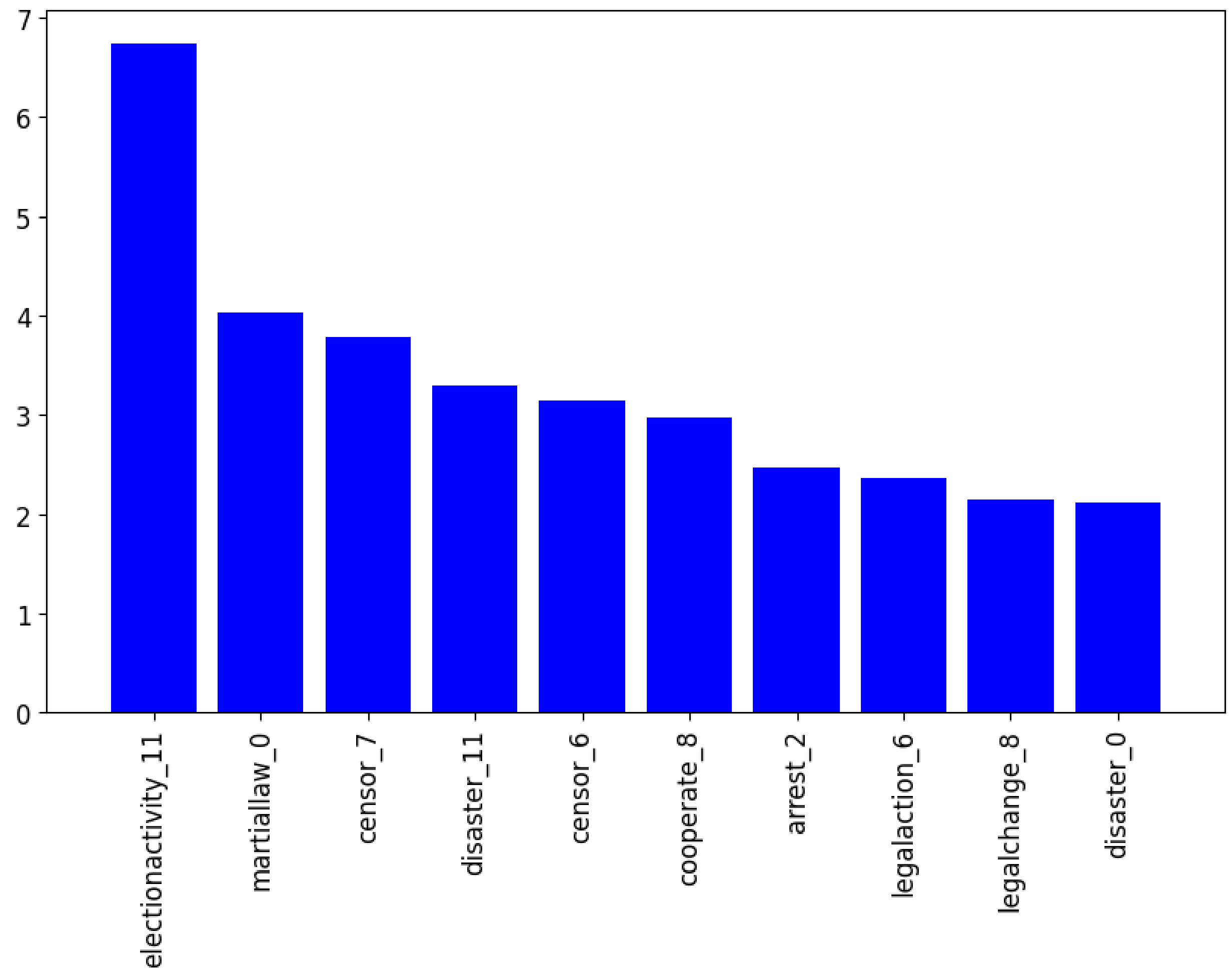
Importantly, the way we communicate results of forecasting models should reflect the needs of decision makers. For example, we may want to reduce the risk of false positives by requiring a relatively large increase in the probability of onset before we issue an early warning (precision). If models regularly make falsely predict onsets, policymakers may quickly lose trust that they can act based on the information being provided. However, lower tolerance for false positives also increases the number of real onsets that we will fail to provide advanced warning of (recall).

The second panel in Figures 2 and 3 illustrates this trade-off by showing how recall changes when we adjust the model to increase the precision of our forecasts. For example, we can deploy the model so that roughly 85% of the onsets that our model predicts will be true onsets (0.85 precision). However, this adjustment will only successfully predict about 20% of the total onsets that we should expect to occur (0.2 recall). In this scenario, decision makers can be very confident that onsets we predict will happen, but they should expect many onsets to happen that we do not predict in advance. If we allow the precision to drop to just below 80%, we should be able to successfully provide advanced warning of around 50% of the total onsets that occur in the real world. These trade-offs are critical when deploying the results from forecasting models in decision-making.

Finally, an analysis of our models can provide insight into the current political events that are strongly associated with future travel advisory onsets. Figure 4 plots the 10 variables that have the greatest effect on the predictions of model about where onsets will occur six months into the future. For each variable, the appended numbers reflect the number of months into the future these variables predict an onset. In other words, "electionactivity_11" tells us that increases in election activity right now increase the probability of an onset in 11 months.

Several key predictors emerge as consistent and influential across various forecast horizons. Notably, factors such as states of emergency (labeled martial_law), protests, election activity, and disasters are consistently associated with predictions of an onset, providing valuable insights into the conditions that often precede serious travel advisories.

Figure 4: Top 10 variables that are most highly predictive of onsets six-month in the future. Appended numbers reflect the number of months into the future these variables predict advisory onset.



Key Takeaways

Our forecasting models demonstrate an impressive ability to detect historical patterns that predict future onset of high-level travel warnings. Our performance metrics are similar to those achieved by the very best conflict forecasting projects and other successful rare event prediction tasks. This is strong evidence that a pipeline to regularly ingest DOS travel advisory data and generate up-to-date forecasts could help decision makers within USG agencies plan more effectively for crises that will affect U.S. citizens abroad. Our models also provide evidence about the political events that can serve as an early warning sign of future instability.

There are two questions for policymakers that should guide attempts to deploy these models for decision making. First, what is the most actionable "forecast window" over which to make predictions—is it 1 month, 3 months, 6 months, or some other time frame? Second, how do policymakers evaluate the tradeoff between true positives (correctly predicting a travel advisory), false positives (incorrectly predicting a travel advisory), and false negatives (failure to predict a travel advisory that actually happens)? False positives can be costly if they lead to unnecessary expenditures or undermine trust. However, more false negatives will result in more unforeseen crises. Forecasting models can be tuned to accommodate policymaker preferences over these tradeoffs.

Acknowledgements

This study was funded under the United States Agency for International Development (USAID) Center for Democracy, Human Rights.