Prediction in Political Economy: Pre-Registered Forecasts of Sovereign Debt Crises*

June 18, 2026

Abstract

Civil conflict research, election forecasting, and other political science sub-fields use prediction to complement hypothesis testing; political economy does not. This is a missed opportunity. Central banks and international financial institutions release outcome data on predictable schedules, which lets researchers pre-register forecasts and evaluate them once new data appear. We apply this design approach to sovereign debt crises. Using country-year data from 1960 to 2021, we trained a random forest classifier and derived four fixed-effects logistic regression models, and we pre-registered our predictions for 2022-2024 on OSF before the Bank of Canada released the outcome data. Our predictive performance is comparable to existing economics forecasts of sovereign debt distress, but under stricter conditions: ours is the first out-of-sample evaluation, with outcome data unavailable at the time of analysis. We offer a blueprint for prediction work that other political economy scholars can adapt to complement existing hypothesis testing approaches.

*The predictions from this study were pre-registered at OSF: https://osf.io/mvwjy/?view_only=7bba3fe5e3164d80a9b338bd83583fcc

Political economy research has focused almost exclusively on questions of *why*, by relying on hypothesis testing. Conversely, prediction-focuses research on questions of *when* are uncommon. In this way, prediction and explanation are treated as distinct scientific tasks rather than as complements (Cranmer and Desmarais, 2017). Other political science subfields have moved beyond this dichotomy. For example, civil conflict research has built shared evaluation standards over two decades through competitions, special issues, and replication exercises (Ward, Greenhill and Bakke, 2010; Blair and Sambanis, 2020; Colaresi and Mahmood, 2017; Hegre et al., 2013). Election forecasting has developed alongside, including pre-registered observational designs (Monogan III, 2013). Political economy has not.

This is surprising for two reasons. First, policymakers, market actors, and international organizations care about when economic events like crises occur, not only why. Second, observational data in political economy come out regularly. Central banks, national governments, and international financial institutions release economic data on predictable schedules. This creates an opportunity for prediction-focused research. When outcome data are released on a schedule, researchers can pre-register predictive models before the next release and evaluate them once the data appear.

We apply this framework to the study of sovereign debt crises - a discrete outcome with clear importance. Governments that default or require emergency international financing face prolonged recessions, political instability, and constrained policy autonomy (Sandleris, 2016; Mohseni-Cheraghlou, 2016). Knowing when these events are likely matters for the actors who manage them. Importantly, our study takes advantage of the fact that the Bank of Canada and the IMF regularly release data on debt distress. We trained a random forest model on country-year data from 1960 to 2021 using publicly available, annually updated economic and political indicators. We also estimate fixed-effects logistic regressions, motivated by recent work on the limits of model complexity (Morucci and Spirling, 2024). We deposited our predictions on OSF before the Bank of Canada released the 2022–2024 outcome data in October 2025. Of the 179 countries in our evaluation sample, 38 experienced a crisis.

This study makes two contributions. First, we establish a research practice for prediction in political economy. International organizations and central banks release outcome data on predictable schedules. We argue that using these data releases, political economy research can complement hypothesis testing with prediction-focused research. Our application focuses on sovereign debt crises, but this approach can be extended to other types of crises, expropriation, or sanctions.

Second, we show that simple fixed-effects regression models predict debt crisis onset about as well as a random forest in this setting. The barrier to prediction in political economy is not methodological sophistication.

Prediction Approaches

We predict debt-based financial crises for 2022-2024 using country-year data from 1960 to 2021. We exclude micro-states, countries that do not interact with international debt markets (e.g. North Korea), or lose statehood before 2021 (e.g. South Vietnam). Full sample details are in Appendix A1.

Our goal is to predict debt crisis onset in a three-year window, 2022–2024. We train and evaluate our prediction on country-year data from 1960 to 2021, then generate forecasts for 2022–2024 before observing any crisis outcomes in that period. We pre-registered these predictions and committed to evaluating them once the data were released in 2025. Because sovereign defaults are publicly visible events, researchers working in this area inevitably possess ambient knowledge of at least the prominent outcomes in their holdout window before conducting the formal evaluation. This is a design limitation that no retrospective train/test split can rule out.

We employ two modeling strategies to generate these predictions of debt crisis onset. The first is a random forest classifier. The second strategy estimate regression models with country fixed effects, using variable subsets informed by our random forest results. We pre-registered the random forest predictions in July 2023. After the random forest evaluation was complete, we developed

the regression strategy as a test to Morucci and Spirling's (2024) claims. We pre-registered those predictions separately in July 2025, before the Bank of Canada released the 2022–2024 outcome data in October 2025.

Target Variable

Our forecasting goal is to predict the onset of sovereign debt crises. Crises are not just defaults. A measure limited to formal default would systematically undercount distress by treating bailouts as non-events. Countries may receive large-scale external financing to avoid default. Therefore, we focus on non-payments of debt obligations, along with IMF financing, following Chakrabarti and Zeaiter (2014) (see also Manasse and Roubini (2009) and Savona and Vezzoli (2015)). The limitation of this approach is each condition of the measure has a distinct data-generating processes. However, even crisis-specific datasets construct their binary indicators from multiple underlying components, combined with coding judgment calls (for example, see Laeven and Valencia, 2020; Reinhart and Rogoff, 2008).

Pre-registration also constrains the choice of target variable to annually updated sources, ruling out datasets such as Laeven and Valencia (2020) and Reinhart and Rogoff (2008) are not released on a predictable annual schedule.

We create a composite debt distress indicator by combining default onset, significant escalation of defaulted debt, and IMF financing above 200 percent of annual quota. We use default data from the Bank of Canada's Database of Sovereign Defaults (Beers and Nadeau, 2015). Following Chakrabarti and Zeaiter (2014), we create a debt crisis variable with the following criteria: (1) if a government starts a default episode, meaning they default on some amount of debt (minimum of one million USD) where they had no default the previous year; or (2) the amount of defaulted debt increases by at least 30 percent from the previous year; or (3) government accepts IMF financing above the annual quota of 200 percent mark of General Resources Account.[1] The measure equals

[1]The IMF's Extended Fund Facility (EFF) provides short-term financial assistance to countries facing balance of

one if a government is in a default crisis (meeting one of our three criteria) and zero otherwise. Our main target is to predict whether a country starts a debt crisis in times $t+1$, $t+2$, or $t+3$.

Predictors

Based on our survey on financial crisis research, we use predictors routinely updated by their sources from four categories: economic fundamentals, political institutions, crisis history, and international relationships.

Random forest models (in R) require non-missing observations in each of the predictors. Given the large number of predictors in the models, pairwise deletion would severely limit the sample. As an alternative, we use imputation to address missingness using the *missForest* package in R (Stekhoven and Bühlmann, 2012). To avoid using the test data to build observations in the training data, we conduct two stages of imputation. First, we impute the training data only using the training sample. This imputed dataset is then used to build and evaluate the training random forest models. Then we impute the test data set using both the test and training dataset to improve the accuracy and efficiency of the imputed test data set.

Random Forest

Starting with the random forest approach, we build a forecasting model using only data from a training set. After we are satisfied with the model, we evaluate it on a separate test data set. Splitting the data into training and test portions allowed us to generalize our predictions and avoid overfitting (Colaresi and Mahmood, 2017).

We identified three possible strategies for splitting the data. The first, randomly selecting observations without regard for time or country, ignores the dependence structure inherent in country-year panels. The second, splitting by an arbitrary year, runs into a problem debt dynamics evolving

payments problems.

over time. The lending environment of the 1980s crises, the emerging market crises of the 1990s, the Eurozone crisis, and the post-COVID landscape are all different. A temporal cutoff risks overfitting.

Thus, we take the third option. We split the data into training and test portions by country. To ensure we have countries from each region, we stratify by region (see Appendix A1 for sample details). Country-splitting preserves the full 1960-2021 time series in the training data. This risks introducing dependencies across the test and training samples, but we expect that it limits the risk of overfitting on a specific temporal window. The testing dataset was not evaluated until we were satisfied that our prediction model was accurate.

One potential issue with machine learning approaches is overfitting, the training data. To avoid this problem, we further separate the training data into 10-fold partitions. Nine of the partitions are then used to build a forecasting model, which is evaluated on the remaining 10th fold (Colaresi and Mahmood, 2017).

To build the forecast, we use random forest models.[2] The model algorithm aggregate predictions across multiple decision trees, each trained on a bootstrapped sample of the data, with final classifications determined by majority vote.

To evaluate the random forest models, we follow Colaresi and Mahmood's (2017) approach. First, we compare the random forest models to a benchmark logistic regression that predicts debt crises, using a version of the model in Shea and Poast (2018, Table G.15, Model 1). We use this model as the benchmark because the authors were primarily concerned about the prediction of default, rather than hypothesis testing. We then build different versions, or "loops", of the random forest model, and compare them to the benchmark logit model.

To evaluate the predictions of the modeling approaches, we use the Receiver Operating Characteristic (ROC) of the predictions that balance the true positive rate (sensitivity) with the false positive rate (specificity).

[2]We use the randomForest package in R (RColorBrewer and Liaw, 2018).

After each iteration of the random forest, we examined the cases we had classified incorrectly on the 10th fold and determined whether input adjustments were needed before moving to the test data, using Colaresi and Mahmood's (2017) model criticism technique.

Loop 1 used only economic variables; GDP per capita and foreign reserves received the highest permutation importance scores. Loop 2 added political variables, which only marginally increased the AUC (0.771 to 0.782). Loops 3 and 4 introduced additional international and institutional variables with diminishing increases to the AUC. We stopped iterating at this point to avoid overfitting.

Table 1 summarizes the performance of each model on training and test data. Each iteration of the random forest model offers an incremental improvement. Based on the test-set AUC, we select the random forest from Loop 4 (AUC $= 0.793$) as the best performing model. The appendix details each iteration, including ROC curves, model criticism plots, and variable importance rankings.

Table 1: Prediction Performance on Training and Test Data

Model	AUC Train	AUC Test
Logit Model	0.787	0.720
Random Forest Loop 1	0.936	0.771
Random Forest Loop 2	0.945	0.782
Random Forest Loop 3	0.945	0.786
Random Forest Loop 4	0.946	0.793

Regression

The second strategy tests whether model complexity is necessary. Morucci and Spirling (2024) show that social science data have a low intrinsic dimensionality – that is, a small number of variables do most of the predictive work. If so, simpler models should perform comparably to more complex ones. We therefore estimate regression models with country fixed effects, using variable subsets informed by our random forest results.

We develop four regression models to forecast crisis onset using key predictors of debt crises, identified from the machine learning analysis. Each regression model takes the form:

$$\Pr(\text{Crisis}_{i,t+1,t+2,t+3} = 1) = \Lambda(\alpha_i + \boldsymbol{\beta}'\mathbf{X}_{it}) \tag{1}$$

where α_i are country fixed effects, $\mathbf{X}_{it}$ are predictor variables measured as of 2021, and Λ is the logistic function. The variable subsets for each model draw on the random forest's variable importance rankings but were specified and pre-registered before evaluation. The first model includes core economic fundamentals: GDP per capita, total reserves as a share of external debt, short-term external debt, external debt-to-reserves ratio, central government debt, and crisis history variables. The second focuses on political institutions: cumulative default propensity, rule of law, government grants and revenue, electoral democracy, and crisis history. The third combines the top economic and political predictors with regional indicators for Africa and East Asia. The fourth is a uses only GDP per capita, years since last crisis, and cumulative default propensity.

Prediction Evaluation

Table 2 reports the twenty countries with the highest predicted probability of crisis onset from the random forest model, along with corresponding prediction from the regression and actual realized outcomes. Of the 179 countries in the evaluation sample, 38 experienced a crisis during 2022–2024. Six of the twenty highest risk countries experienced a crisis; the full set of predictions is reported in the supplemental appendix .

The formal evaluation criteria were pre-registered before observing crisis outcomes. First, we derived an ROC curve for our predictions for 2022-2024, along with the corresponding AUC. We expected the AUC to be consistent with the test data set evaluation of 0.79. Given that we are only evaluating one year (n=179), we use a Wald confidence interval to provide a band of AUC that we would consider successful: $0.79 \pm \frac{1}{\sqrt{179}} = [0.72, 0.86]$. We apply the same benchmark to the regression models and compare their AUCs directly against the random forest.

Second, we evaluate the predictions using a Hosmer-Lemeshow goodness-of-fit test (Hosmer

Table 2: Highest-Risk Countries: Random Forest and Parsimonious FE Predictions

Country	Random Forest Pr(Crisis) Pr(Crisis)	Parsimonious FE Pr(Crisis)	Crisis Onset
Papua New Guinea	0.479	0.21	Yes
Haiti	0.434	0.74	No
Chad	0.431	0.49	No
Ecuador	0.398	0.51	No
St. Vincent & Grenadines	0.395	0.58	No
Afghanistan	0.385	0.52	No
Angola	0.375	0.32	No
Yemen	0.362	0.37	No
Republic of the Congo	0.352	0.61	No
Cameroon	0.346	0.63	No
Malawi	0.346	0.51	No
Benin	0.343	0.50	Yes
The Gambia	0.340	0.47	Yes
Equatorial Guinea	0.339	0.40	No
Burundi	0.337	0.45	No
Bhutan	0.332	0.10	No
El Salvador	0.331	0.23	Yes
Cape Verde	0.330	0.25	Yes
Pakistan	0.330	0.47	No
Eswatini	0.324	0.26	Yes

Notes: Top 20 Random Forrest predicted probabilities of debt crisis onset during 2022–2024, with regression predictions.

and Lemesbow, 1980). We sort predictions into 10 bins, then compare the observed number of crises per bin against the expected count using a Pearson χ^2 statistic. We also conduct the same test with 5 bins as a sensitivity check.

For some of the countries on the list, we knew at the time of pre-registration that they have already defaulted or entered into financing agreements with the IMF. For example, Zambia received IMF financing in 2022 and was nearing a restructuring deal at the time of pre-registration. Ghana and Sri Lanka also received IMF assistance in 2023. We marked these countries as already in crisis. The remaining countries act as our pre-registered predictions.

Table 3 reports the evaluation results for the random forest predictions, along with the four regression prediction.. The random forest achieves an AUC of 0.681, outside of the pre-registered

Table 3: Evaluation of Crisis Predictions, 2022–2024

	Random Forest	Economic FE	Political FE	Comprehensive FE	Parsimonious FE
AUC	0.681	0.713	0.713	0.708	0.736*
HL 10-bin (p-value)	0.117*	0.002	0.019	0.012	0.094*
HL 5-bin (p-value)	0.029	0.001	0.002	0.001	0.036

Notes: * indicates passing pre-registered benchmark for prediction success. AUC computed against crisis onset during 2022–2024. Pre-registered benchmark: AUC $\in [0.72, 0.86]$, derived from the test-set AUC of $0.793 \pm 1/\sqrt{179}$. Hosmer-Lemeshow p-values test the null of good calibration; $p \geq 0.05$ indicates failure to reject. The random forest evaluation sample ($N = 179$) excludes Zambia, Ghana, and Sri Lanka, whose crises were known at pre-registration. The fixed-effects sample ($N = 178$) additionally excludes Ethiopia, following the second pre-analysis plan.

bounds of $[0.72, 0.86]$. Out of the regression model, only the parsimonious specification passes the benchmark. This result is consistent with the low intrinsic dimensionality arguments Morucci and Spirling (2024).

Moving onto the Hosmer-Lemeshow calibration test, the random forest predictions pass the 10-bin test ($p = 0.117$), but not the 5-bin test. The regression models, by contrast, fail the tests at both the 10-bin and 5-bin level, except the parsimonious model, which passes the 10-bin model ($p = 0.094$).

We did not pre-register a rule for aggregating or evaluating across the three evaluation criteria. No model passes all three tests, though the parsimonious fixed-effects specification passes more tests than the others. The results do make clear that both modelling approaches fall short of the performance we anticipated. In addition, the gap between the test-set AUC and the pre-registered evaluation underscores why retrospective holdout performance should not be taken at face value.

Design Principles for Political Economy Prediction

We designed this study to establish a template for political economy prediction. Without an appropriate benchmark in the subfield, we did our best to base our approach on best practices from the conflict processes literature. We draw several lessons from the exercise that can inform future prediction work.

First, keep the target variable simple. Our target event – debt distress – was measured using three different components, following existing research (Chakrabarti and Zeaiter, 2014). These components each have their own data-generating processes, complicating the prediction exercise. Pooling these events inflates the positive class with heterogeneous cases — a country seeking precautionary IMF lending faces different dynamics than one in outright default. Future exercises should build separate prediction models for each component and evaluate each independently.

Second, keep the model simple. For this study, we started with the random forest models, and made them more complex. In hindsight, we would start with the simpler models, and pre-register our predictions for each iteration. The good news for scholars is that the barrier to entry for prediction drops considerably. Scholars who already estimate logistic regressions for hypothesis testing can extend their analysis to out-of-sample prediction with minimal additional investment.

Third, pre-register the evaluation protocol with the same specificity as the predictions themselves. We pre-registered three evaluative tests — AUC benchmarks and Hosmer-Lemeshow calibration— but we did not specify how to adjudicate mixed results across them. When one test passes and another fails, which takes precedence? We left that question unanswered. Conflict forecasting benefits from decades of shared evaluation standards built through competitions and special issues (Ward, Greenhill and Bakke, 2010; Hegre et al., 2013; Blair and Sambanis, 2020). Political economy research should build equivalent infrastructure. In the meantime, we recommend treating AUC as the primary evaluation criterion.

Finally, this study revealed a tension between model evaluation and domain knowledge. We allowed test-set AUC to dictate model choice, selecting the random forest from Loop 4 (AUC = 0.793) over simpler specifications. The out-of-sample results suggest we should have weighed substantive priors more heavily. Lo et al. (2015) shows that statistically significant predictors are not automatically good forecasting inputs, and vice versa — the variables that improve historical classification do not necessarily carry forward. Future prediction exercises should treat domain knowledge as a constraint on model selection, not just an input to variable choice.

Conclusion

Using random forest and regression models on country-year data from 1960 to 2021, we evaluated pre-registered forecasts of debt crisis onset for 2022–2024. While the prediction success was modest at best, we found that simpler models performed best.

To the best of our knowledge, this study is the first prediction-focused applications of machine learning to sovereign debt crises in political science. We hope that this establishes a template others can extend to another political economy outcomes, such as other types of crises (e.g. bank or currency), expropriation, or sanctions.

The lessons learned in this study point toward a future agenda for prediction work in political economy. Conflict forecasting built its evaluation infrastructure and norms over decades through competitions and special issues. Political economy prediction will improve faster if the subfield begins building equivalent shared standards.

References

Beers, David and Jean-Sébastien Nadeau. 2015. "Introducing a new database of sovereign defaults." *Available at SSRN 2609162* .

Blair, Robert A and Nicholas Sambanis. 2020. "Forecasting civil wars: Theory and structure in an age of Big Data and machine learning." *Journal of Conflict Resolution* 64(10):1885–1915.

Chakrabarti, Avik and Hussein Zeaiter. 2014. "The determinants of sovereign default: A sensitivity analysis." *International Review of Economics & Finance* 33:300–318.

Colaresi, Michael and Zuhaib Mahmood. 2017. "Do the robot: Lessons from machine learning to improve conflict forecasting." *Journal of Peace Research* 54(2):193–214.

Cranmer, Skyler J and Bruce A Desmarais. 2017. "What can we learn from predictive modeling?" *Political Analysis* 25(2):145–166.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand and Henrik Urdal. 2013. "Predicting armed conflict, 2010–2050." *International Studies Quarterly* 57(2):250–270.

Hosmer, David W and Stanley Lemesbow. 1980. "Goodness of fit tests for the multiple logistic regression model." *Communications in statistics-Theory and Methods* 9(10):1043–1069.

Laeven, Luc and Fabian Valencia. 2020. "Systemic banking crises database II." *IMF Economic Review* 68(2):307–361.

Lo, Adeline, Herman Chernoff, Tian Zheng and Shaw-Hwa Lo. 2015. "Why significant variables aren't automatically good predictors." *Proceedings of the National Academy of Sciences* 112(45):13892–13897.

Manasse, Paolo and Nouriel Roubini. 2009. "Rules of Thumb for Sovereign Debt Crises." *Journal of International Economics* 78:192–205.

Mohseni-Cheraghlou, Amin. 2016. "The aftermath of financial crises: a look on human and social wellbeing." *World Development* 87:88–106.

Monogan III, James E. 2013. "A case for registering studies of political outcomes: An application in the 2010 House elections." *Political Analysis* 21(1):21–37.

Morucci, Marco and Arthur Spirling. 2024. "Model complexity for supervised learning: why simple models almost always work best, and why it matters for applied research." *Department of Political Science, Michigan State University* .

RColorBrewer, Suggests and Maintainer Andy Liaw. 2018. "Package 'randomforest'." *University of California, Berkeley: Berkeley, CA, USA* .

Reinhart, Carmen M. and Kenneth S. Rogoff. 2008. *This Time is Different: Eight Centuries of Financial Folly*. Princeton University Press: Princeton, NJ.

Sandleris, Guido. 2016. "The costs of sovereign default: Theory and empirical evidence." *Economia* 16(2):1–27.

Savona, Roberto and Marika Vezzoli. 2015. "Fitting and forecasting sovereign defaults using multiple risk signals." *Oxford Bulletin of Economics and Statistics* 77(1):66–92.

Shea, Patrick E and Paul Poast. 2018. "War and default." *Journal of Conflict Resolution* 62(9):1876–1904.

Stekhoven, Daniel J and Peter Bühlmann. 2012. "MissForest–non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28(1):112–118.

Ward, Michael D, Brian D Greenhill and Kristin M Bakke. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of peace research* 47(4):363–375.