WP/19/228

IMF Working Paper

Machine Learning and Causality:
The Impact of Financial Crises on Growth

by Andrew Tiffin

INTERNATIONAL MONETARY FUND

WP/19/228

IMF Working Paper

Middle East and Central Asia Department

Machine Learning and Causality: The Impact of Financial Crises on Growth

Prepared by Andrew Tiffin[1]

Authorized for distribution by Martin Cerisola

November 2019

***IMF Working Papers* describe research in progress by the author(s) and are published to elicit comments and to encourage debate.** The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

Abstract

Machine learning tools are well known for their success in prediction. But prediction is not causation, and causal discovery is at the core of most questions concerning economic policy. Recently, however, the literature has focused more on issues of causality. This paper gently introduces some leading work in this area, using a concrete example—assessing the impact of a hypothetical banking crisis on a country's growth. By enabling consideration of a rich set of potential nonlinearities, and by allowing individually-tailored policy assessments, machine learning can provide an invaluable complement to the skill set of economists within the Fund and beyond.

JEL Classification Numbers: C21; C40; E6; E44; E50; E60; F3;G01; G20.

Keywords: Supervised machine learning; causal inference; policy evaluation; counterfactual prediction; randomized experiments; treatment effects; banking crisis; financial crisis.

Author's E-Mail Address: atiffin@imf.org

[1] I am most grateful to the IMF's outstanding Vulnerability-Exercise Upgrade team and to colleagues in the Middle Eastern and Central Asia Department, for their insight and support. I am also particularly grateful to Sebastian Herrador Guzman, who first alerted me to the recent literature on Causal Forests. All errors and omissions, of course, are my own. The views expressed in this paper are those of the author and do not reflect the views of the IMF, the IMF Executive Board, or IMF Management.

Contents Page

Boxes

Figures

I. Introduction

> *The difficulty is to detach the framework of fact—of absolute fact—from the embellishments of theorists…*
>
> *The Memoirs of Sherlock Holmes*

Machine Learning (ML) is far from new. Indeed, research in this field has a pedigree that extends back more than fifty years, and often entails well-known methods from non-parametric statistics. But the pace of adoption of these techniques has picked up rapidly over the past 10 years, driven in large part by their impressive predictive capabilities.

Accurate prediction, of course, is important. But for empirical economists, prediction by itself is not always enough. Questions concerning the *effects* of a particular policy, for example, present a fundamentally different problem; as the answers require us to estimate *what would have happened* in the absence of that policy stance. This is the central challenge of causal inference and has perhaps been a key reason why machine learning hasn't made greater headway among economists. Predictions can be validated, and so lend themselves to ML techniques. Counterfactuals cannot, as we never get to see the path not taken.

Nonetheless, there is a swiftly expanding literature ("causal machine learning") that has endeavored to take the strengths and innovations of ML methods and apply them to causal inference problems—leading to more precise, less biased, and more reliable estimators of causal effects.

This paper will provide a gentle introduction to some leading research in this area, taking a concrete task as an example—assessing the impact of a hypothetical financial crisis on output growth. Focusing mainly on recent work on *Causal Forests* (Athey and others, 2018), the paper will use an ML approach to provide plausible estimates of the average impact of a crisis, as well as estimates of how that impact varies from country to country. Moreover, the paper will exploit recent advances in the interpretation of complex ("black box") models to show how machine learning can provide valuable insight into the factors underlying these different country-level effects, with particular attention to potential thresholds and interactions.

Section II of this paper will outline some ways in which the machine-learning literature has addressed issues of causal inference, while section III will focus in particular on the intuition behind the Causal Forest algorithm. Section IV will apply this tool to the estimation of the cost of financial crises, and will draw on recent advances in ML visualization, exploration, and explanation to investigate not only the different country-by-country cost of a crisis, but also the particular factors that may make a crisis worse in some countries compared to others. Section V briefly outlines ways in which the causal forest approach can be extended to other types of problems, and section VI then concludes.

II. A MACHINE-LEARNING APPROACH TO CAUSAL INFERENCE

> *More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history*
>
> *Gary King, 2007*

A. Estimating Causal Relationships: Predicting the Counterfactual

The theoretical gold standard for estimating the causal impact of an event (or policy intervention) is a randomized controlled experiment. But for most questions in economics, experiments are often impractical, unethical, or simply impossible. So, a large share of empirical work relies on observational data. Unfortunately, estimating the causal effect of an event or policy from this type of data is problematic, as we never see the counterfactual—we never see what *would have happened* had a different policy been chosen. This is what Holland (1986) calls the "fundamental problem of causal inference." (Box 1)

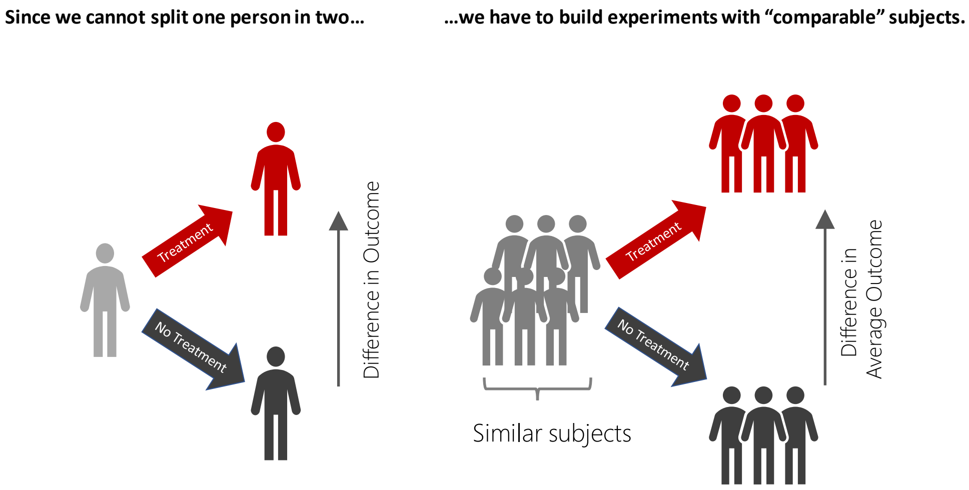
The literature on deriving causal relationships from observational data is vast, and essentially examines strategies to build a convincing proxy for this *un*observable counterfactual.[2] For example, with a large enough dataset, one strategy might be to simply measure the average outcome for those subjects that experienced the event and compare it against the average for those that did not. This strategy, however, demands that we observe all "confounding factors" (i.e., factors correlated with both the outcome and the likelihood of the event). *Conditional* on these observed confounders, the intervention ("treatment") is essentially as good as randomly assigned, and so the average difference between "treated" and "untreated" groups can be taken as a valid proxy for a causal effect.

In a standard regression framework, conditioning on these confounders is typically done by including a pre-identified set of "control" variables. But in a deeper sense, such conditioning strategies are essentially constructing a credible *prediction of the counterfactual*. For example, the standard "propensity score" approach of Rosenbaum and Rubin (1983) calculates the conditional probability of a particular treatment given a set of observed covariates. The "propensity score matching estimator" then predicts the *missing* counterfactual for each *actual* observation; by using the closest observation from the other group, according to its propensity score (Abadie and Imbens, 2006). What makes causal inference particularly challenging, however, is that these counterfactual predictions can never be validated—we are not predicting an outcome that will become known at some future point in time, we are instead predicting a potential outcome that will *never* be known. In this context, it is important that such predictions are at least as plausible and persuasive as possible.

[2] For a survey of this literature, see Imbens and Rubin (2015), and Morgan & Winship (2015).

Box 1. The Fundamental Problem of Causal Inference

Estimating causal relationships is difficult, as we never get to see the path not taken. Suppose we wanted to estimate the impact of a new medicine on a particular subject. To answer that question, we need to measure what happens when the subject takes the medicine, and compare that to what happens when she fails to take the medicine. But this is impossible. We can only ever see *one* of these outcomes, not both. Causal inference, therefore, essentially entails arriving at a plausible estimate of an unobserved (and fundamentally unobservable) counterfactual.


Since we cannot split one person in two…

…we have to build experiments with "comparable" subjects.

The best we can do is to construct (perhaps implicit) experiments in which similar people take different paths. An idealized experiment, for example, would entail identical twins, where the subjects are assumed to be the same in every particular. If one twin takes the medicine and the other does not, then the second (untreated) twin serves as the counterfactual. In the same vein, if we had a group of essentially similar subjects, then the average outcome for those who did not take the treatment would serve as the counterfactual for those who did. The difference in the average outcome between the treated and untreated subjects would then be the treatment effect for that particular group. We could then use that average to predict the likely effect for other subjects with the same characteristics. As an extension, if we wanted to estimate the likely effect for people with a different set of characteristics, we would need a different set of experiments; each with a matched set of subjects, all of whom share those characteristics. This is the core idea behind strategies to estimate *individual* treatment effects, both from specially designed experimental studies, and from observational data—both of which sort subjects into well-matched sub groups.

If, instead, we are interested in the *average* effect for a broader, more heterogenous population, estimation requires that the treatment be allocated randomly. This is the core idea behind the Randomized Controlled Experiment (RCT), where randomization ensures that both treated and untreated groups are alike *on average*, so that any idiosyncratic differences are cancelled out when comparing outcomes between the two groups. In observational studies, of course, treatment is almost never random, so we need to account statistically for all the ways the groups *might* be systematically different.

[1] Twin studies have long been a staple in public health and behavioral science. See McGue and others (2010).

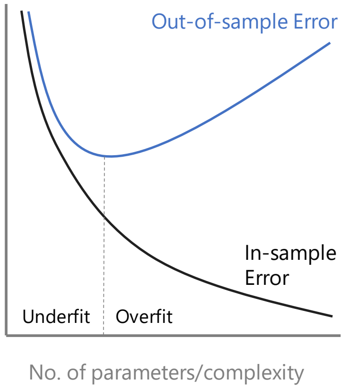
B. Machine Learning: Making Better Predictions

There is no broadly accepted consensus on the definition of machine learning. As a general guide, the field has its origins in computational statistics, and is chiefly concerned with the use of algorithms to identify patterns within a dataset (Kuhn and Johnson, 2016). The actual algorithms can range from the simplest OLS regression to the most-complex "deep learning" neural network; but ML is distinguished by its often single-minded focus on predictive performance—indeed, the essence of machine learning is the design of experiments to assess how well a model trained on one dataset will predict new data (Box 2).

In this regard, the growing popularity of ML techniques stem from their ability to discover complicated patterns that have not been specified in advance. In economics in particular, the world is complex, and everything is connected. For some problems, this is less of a concern, as we are able to build empirical models that can successfully focus on specific economic relationships. For other problems, simple models are often insufficient. As an illustration, predicting a financial crisis is a very challenging problem—the rare onset of a crisis is likely shaped by the interaction of a range of economic drivers, and there is no theoretical consensus on how these drivers come together to trigger a crisis. So, specifying a suitable model a priori is difficult. Instead, a useful predictive model should be able to efficiently sift through a broad range of potential independent variables, identifying the relationships, thresholds, and interactions that are most reliably and robustly informative.

Machine-learning does this well. Random Forests (RF), for example, has proven to be one of the most popular and successful general-purpose algorithms currently available (Box 3). Random Forests can handle a range of different predictor types (binary, categorical, numerical) and the algorithm implicitly incorporates an element of variable selection—automatically sifting through a broad range of covariates, and focusing on those with the greatest predictive power. Moreover, RF can also be applied to a wide range of modeling tasks, ranging from classification, to regression, to cluster analysis. Further, when applied to prediction problems, the performance of the procedure is generally impressive. Although there is no single algorithm that will dominate in all applications (known in machine-learning parlance as the "no free lunch theorem"), Random Forests will usually do well and will often take less time and effort to train than most alternatives.

Box 2. Better Predictions Through Regularization: Penalized Regression and LASSO



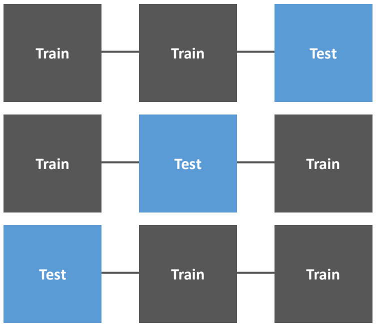
Fitting is easy. Prediction is hard. So, focusing on in-sample fit is often an inadequate guide to predictive performance. Indeed, within machine-learning it is stressed that in-sample fit tells us little of value, other the number of parameters in a model (it is always possible to boost the fit by adding parameters). The danger is that a model with a supposedly good in-sample fit may in fact be modeling idiosyncratic noise. When taken out of sample, then, the model will perform poorly. We say that such a model is "overfit," and this is the core issue that machine learning seeks to address.

Regularization is aimed at preventing overfitting. A penalized (or "regularized") regression seeks to balance in-sample fit against a penalty term that depends on the magnitude of the regression coefficients. A standard OLS regression with many covariates will typically produce parameters with a large variance, making the model unreliable. Adding a penalty term, however, causes the coefficients for unimportant variables to shrink towards zero, and so allows the model to identify those variables most strongly associated with the outcome. This LASSO (Least Absolute Shrinkage and Selection Operator) penalty, in particular, will often assign a coefficient of exactly zero, effectively dropping that variable from the regression. Within the penalty term, a *tuning parameter*, lambda (λ) is used to control the strength of the penalty. The higher the value of lambda, the greater the degree of regularization, and the more coefficients are reduced to zero. A key choice for the researcher, then, is to select the value of lambda that optimizes likely predictive performance.

$$\beta_{LASSO} = \underset{\beta_j}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{\left(y_i - \sum_j x_{ij}\beta_j \right)^2}_{RSS} + \underbrace{\lambda \sum_j |\beta_j|}_{LASSO\ Penalty}$$

Estimating future performance: Holdout validation. The process of predicting how a model will perform on new data is called model *validation*. In *holdout validation*, the data is split into a training and testing set. The model is generated using only the training set, and is then asked to make predictions using only the test set. Comparing these predictions with actual outcomes gives the *validation error*, which can then be used to choose between different models, or indeed to select the ideal value for the tuning parameter lambda (λ).

Estimating future performance: Cross validation. As an alternative, *cross validation* takes advantage of the entire data set. The basic idea is simple: (i) First divide the entire dataset into K folds (say, K=3), take one of those folds and set it aside as a test set. (ii) Using the remaining (2) folds as a training set, estimate the model, and then use the test set to determine the model's prediction error. (iii) Repeat this procedure using all combinations of the test and training sets, producing an array of (3) validation errors associated with that particular model, which then provides a gauge of its average out-of-sample performance. Once again, this metric can be used to help choose between different types of models or to find a tuning parameter value (λ) that optimizes out-of-sample performance. In machine-learning, these settings are called "hyperparameters," and are tuned to minimize the *cross-validation error*.



Train	Train	Test
Train	Test	Train
Test	Train	Train

[3] Although cross validation aims to provide a good estimate of out-of-sample performance, best practice in machine learning often employs an *additional* explicitly quarantined evaluation set, which then provides a final indicator as to how the model will likely perform when predicting future (unknown) data.

C. An Emerging Synthesis: "Causal" Machine Learning

Traditionally, prediction and causal inference have been treated as two very separate problems. Prediction asks *what usually happens,* given a particular set of circumstances. Causal inference asks instead *what would happen if we intervened in the system*? The focus of the second question is quite distinct, and for problems where we cannot run experiments, the challenge of teasing out an answer from observational data is arguably more difficult.[3]

Until recently, the divide between causal analysis and prediction mirrored a similar divide between traditional econometrics and machine learning.[4] Econometrics has generally focused on *explanation*, with particular attention to issues of causality, and a premium placed on parsimonious models that are easy to interpret. A "good" model in this framework is mostly assessed according to statistical significance and in-sample goodness-of-fit. Machine learning, on the other hand, has focused more on prediction, with emphasis instead on a model's accuracy rather than its interpretability. A "good" machine-learning model, then, is determined by looking at its likely out-of-sample success.

Over the past few years, however, the distinction between machine learning and econometrics has narrowed significantly. This is most obvious for applications such as time-series analysis and forecasting, which are centered almost entirely on making accurate predictions.[5] For causal analysis, however, the growing integration of machine learning stems from the fact that many estimation problems can be broken down into different steps, some of which resemble pure prediction problems.

Take a standard instrumental variables (IV) problem, for example, where we are interested in the impact on Y of an *endogenous* treatment variable X. In some situations, there may be a broad range of potential instruments available (which can sometimes exceed the number of observations) leaving us with the question as to which set of instruments to use in constructing the IV estimator. Typically, the researcher will be forced to choose a relatively parsimonious subset of instruments, and then justify that choice based on prior knowledge or ad-hoc intuition. Belloni and others (2012) outline a procedure in which the selection of instruments is instead determined optimally through a machine-learning algorithm such as a

[3] More formally, suppose we have two random variables *(Y,X)* distributed according to a joint probability distribution *p(y,x)*. **Prediction** draws from the conditional *p(y|x)*, which is the distribution of *Y* given that we *observe* that variable *X* takes value *x*. For **causal inference**, however, we want the distribution of *Y* if we were to *set* the value of *X* to *x*. This describes the distribution of *Y* we would observe if we intervened in the data generating process by artificially forcing the variable *X* to take value *x*, but otherwise simulating the rest of the variables according to the original process. Importantly, this data generating procedure is **not** the same as the joint distribution *p(x,y)*. See Pearl (2009).

[4] See Breiman (2001a) for a discussion on the different cultures associated with the two fields.

[5] For example, see Tiffin (2014) for an application of machine-learning techniques to nowcasting in data-poor emerging markets.

LASSO regression (Box 2)—in this case, the ML technique is introduced only in the first-stage regression, which is a simple predictive relationship.

A similar, if more complicated, opportunity arises in the case where the treatment variable is *exogenous*, but where there are a large number of potential confounding variables, only *some* of which are actually important. Often, of course, we may not know *a priori* which variables are important and which ones are not. Including *all* variables in the regression runs the risk of overfitting, where spurious relationships can potentially be given a causal interpretation. Failing to control for key confounders, on the other hand, will bias the estimated impact of the treatment.

One machine-learning solution (A) might be to run a LASSO-style regression that automatically selects the most important confounders (based on their ability to predict the outcome). In an alternative solution (B), Wyss and others (2014) outline a procedure that uses machine-learning to estimate the propensity score for each observation (with covariates in this case selected on their ability to predict the likelihood of inclusion in the treatment set). In both cases, the predictive strengths of machine-learning are exploited to improve those components of inference that can be reduced to pure prediction.

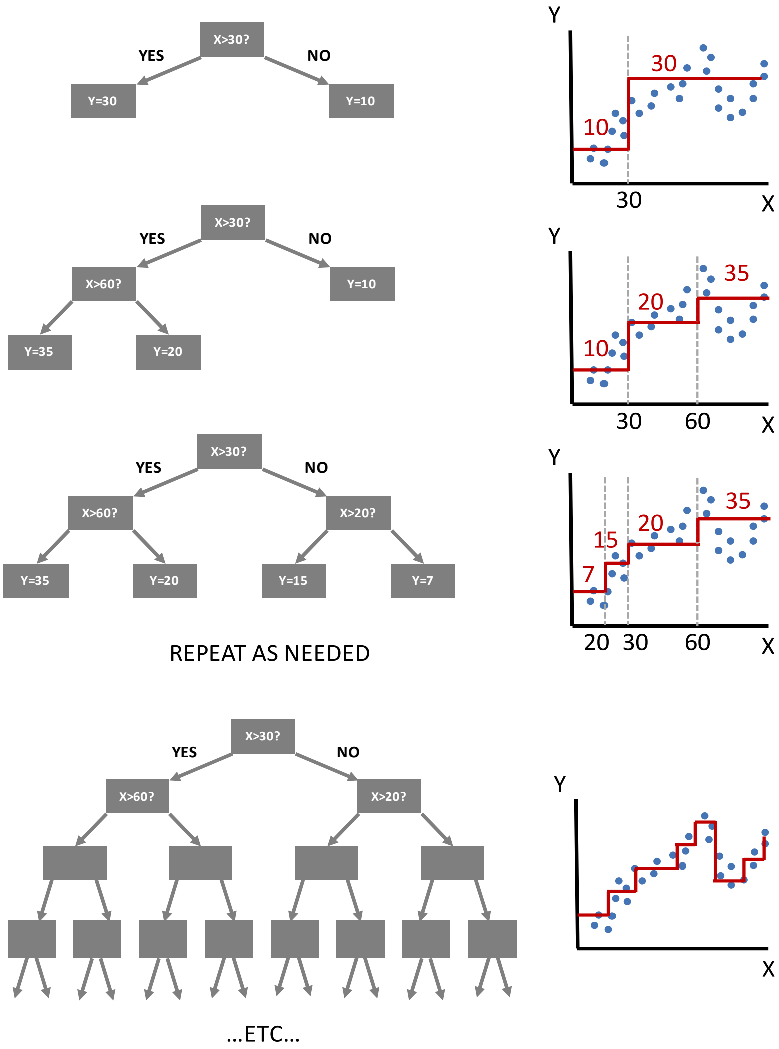
As an extension, Belloni and others (2014b) note that, ideally, potential confounders should be selected based on their association with *both* the outcome *and* the treatment variable (as it is this *joint* association that is the source of bias for the treatment effect). They therefore outline a double-selection procedure, where they first use a LASSO regression to select covariates correlated with the outcome, and use a similar LASSO regression to select covariates correlated with the treatment. They then take the union of both sets of covariates, and include them as controls in a standard OLS regression; showing that the resulting estimator of the treatment effect is potentially much improved over a naive estimate stemming from solution (A).

All of the machine-learning modifications outlined above are focused primarily on estimating the average treatment effect (ATE) for the sample as a whole. Often, however, we may also be interested in exploring whether that effect differs from subject to subject—determining, for example, whether financial crises tend to have more of an impact in some countries rather than others, and investigating the factors that may make that impact worse in a particular instance. It is in the estimation of these heterogeneous treatment effects (HTE) where the predictive strengths and flexibility of machine learning can be leveraged to their fullest extent.

Box 3. Better Predictions Through Ensembles: Decision Trees and Random Forests (Cont.)

Tree-based methods provide an intuitive, easy-to-implement way of modeling complex relationships. At core, these methods are based on the notion of a decision tree; which aims to deliver a structured set of yes/no questions to predict a particular outcome. One of the key attractions of decision trees is that they can take an extremely complex, non-linear problem, with a wide range of potential predictor variables, and reduce it to a procedure that is easily understood by a non-technical user. Imagine a flowchart, where each level is a question with a yes or no answer. Following the chart, and answering the questions one by one, eventually the chart will give a solution to the initial problem. That is a decision tree. The challenge for the algorithm is to come up with the right questions.

A traditional econometric method would usually center around a logit or probit model. But decision trees take a very different approach. Rather than fitting a linear regression, they are focused instead around the repeated partitioning of the predictor space into two sets, starting with an initial split that decreases the prediction error the most: i.e., the algorithm considers every possible split on every possible predictor variable, and chooses the one split on the one variable that best separates the sample into the two most dissimilar subsamples (based on the predicted outcome). These binary partitions then continue until the termination of the tree, and are recursive—i.e., each subsequent split only considers the subsample under which it falls, rather than the whole dataset. The result is an efficient set of yes/no questions that can quickly sort any individual instance into an appropriate group of similar observations.

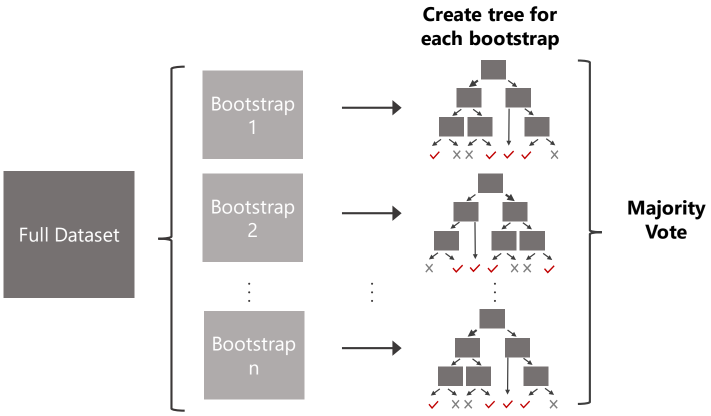


A regression tree is a type of decision tree, which is designed to approximate a continuous real-valued function. Essentially, by sorting the dataset into groups of similar observations, it provides a non-parametric estimate of the expected outcome for any individual within that group.

Box 3. Better Predictions Through Ensembles: Decision Trees and Random Forests (Concl.)

Decision trees are computationally efficient, and work well for problems where there are important nonlinearities and interactions. Trees tend not to work as well if the underlying relationship is linear, but even then, they can reveal aspects of the data that are not apparent from a traditional linear approach (Varian, 2014).

The Random Forest (RF) algorithm (Breiman, 2001b) is an "ensemble" technique that modifies the decision-tree approach to minimize the problem of overfitting. One problem with decision trees is that they often provide models that fit the training sample extremely well, but perform poorly when making out-of-sample predictions. A common solution is to shorten or "prune" the tree by imposing a penalty for an overly long/complex structure (analogous to the penalty term (λ) in Box 2). As an alternative solution, the RF algorithm modifies the decision-tree approach—instead of focusing on a single tree, it uses the data to grow numerous (unpruned) trees and then combines the results.



The first Random Forest modification is the use of bootstrap aggregation (or "bagging"). In bagging, an individual tree is built on a random sample of the dataset, roughly two thirds of the total observations—the remaining one-third are referred to as out-of bag (OOB) observations and can be used to gauge the accuracy of the tree. This is repeated hundreds or thousands of times. When asked to predict the most likely outcome of a new instance, then, the RF algorithm will feed that instance through each of these thousands of individual trees, and will aggregate their predictions; say by taking the average. The fact that none of the trees is pruned means that each individual tree is a weak model that will have a hard time distinguishing the dataset's underlying *signal* from statistical *noise*. However, by building a large ensemble of (weak) individual trees, the algorithm is essentially exploiting the law of large numbers to average out the noise, leaving only the signal.

The second modification is to take a random sample of the set of predictors at each split. In the case of highly correlated predictors, and particularly in the event of a single driving predictor, bagging by itself can be insufficient, as it may simply produce multiple versions of the same tree. To get around this problem, RF introduces an added element of randomization—at each split, the algorithm only considers a random subset of the available set of predictors (usually the total number of predictors divided by three). By randomizing the predictor space, the RF algorithm effectively guarantees that the trees that go into the final collection will be relatively diverse. Again, each tree on its own will be a weak model, as it is grown on a deliberately limited dataset. But the essence of the RF approach is that, by combining a large number of (uncorrelated) weak models, we can end up with an aggregate prediction that is surprisingly strong.

III. USING RANDOM FORESTS FOR CAUSAL INFERENCE: THE CAUSAL FOREST

> *Am I right in thinking that the method of multiple correlation analysis essentially depends on the economist having furnished, not merely a list of the significant causes, but a* ***complete*** *list? … If so, this means that the method is only applicable where the economist is able to provide beforehand a correct and indubitably complete analysis of the significant factors.*
>
> *J.M.Keynes, 1939*

A. The Standard Problem

Consider a researcher who wishes to study the impact of a binary treatment (W) on a particular outcome (Y), but who suspects that the estimate may be biased by a (potentially large) set of confounding variables (X). A simple solution would be to regress Y on a dummy version of W, and to control for enough confounding variables (X) so that, *conditional on these variables*, the likelihood of being in the treatment group can be assumed to be independent of the outcome. There is, of course, no way of testing this assumption, so the choice of control variables must be defended on theoretical grounds or on the basis of the researcher's domain expertise. The coefficient τ, then, is an estimate of the average treatment impact for the sample.

$$Y_i = \underbrace{\tau_{(i)}}_{\text{Interact?}} \times W_i + \overbrace{\boldsymbol{X_i}\beta}^{\substack{\text{How many?} \\ \text{Higher order terms?}}} + \varepsilon_i$$

But the potential confounding role of particular variables is not always obvious *a priori*, which means that there is always a chance that the wrong variables may be omitted and that the estimate of τ is biased. Similarly, there is also the possibility that the influence of these variables may be nonlinear, with potential threshold effects, changing slopes, or key interactions between the variables themselves. The researcher might try to allow for such effects within a linear regression; but all thresholds, interactions, and higher-order polynomial terms would then have to be specified in advance—and the cost of getting this wrong is, again, a biased estimate of the treatment effect (τ). Moreover, trying to include *all* potential interactions and thresholds would run the risk of overfitting, as the number of possibilities would in most cases be quite large.

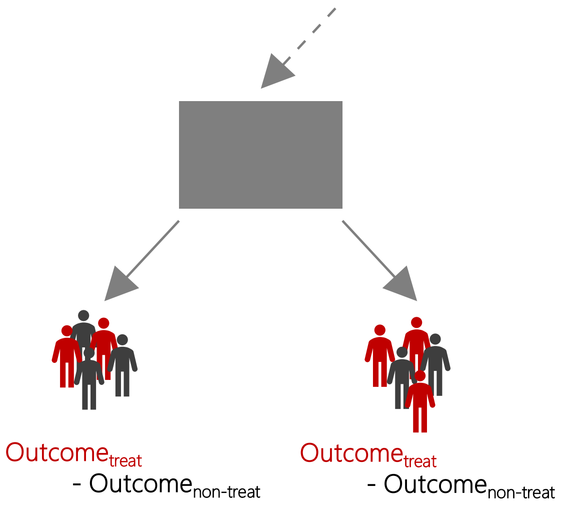
Also, suppose the researcher was additionally interested in how the impact of the treatment might change in different circumstances; perhaps to study how the costs and benefits of a particular policy might vary for different subjects, or perhaps to gauge differing degrees of vulnerability for subjects who may all be at risk of the same event. Within the linear regression framework, the researcher could interact the treatment dummy with certain variables of interest—if the interaction term is significant, then the treatment effect depends on the level of that variable. But again, the researcher would have to specify these interactions in advance, and the main determinants of potential heterogeneity may not always be obvious, *a priori*. Further, the factors that shape the impact of an event or policy may not act in a linear manner, so the researcher would have to specify any potential thresholds or interactions up front, with the same problems and risks.

B. The Causal Forest

One possible solution is to take advantage of machine learning's flexibility to identify well-matched subgroups. The estimated effect for any individual instance, then, would be the average effect for the subgroup to which they belong.

The *causal forest* is a random forest made up of *honest causal trees*.[6] These are similar to the decision trees outlined in Box 3. However, a typical regression tree recursively partitions the data by repeatedly asking, "what variable and what particular split would best separate the data into the two most dissimilar sets, by outcome?" The tree then arrives at a sequence of splits that efficiently sorts the whole dataset into bins ("leaves"), where the average outcome in each bin serves as the predicted outcome for all its members.

But we are not primarily interested in the *outcome* by itself. We are interested instead in the *impact of a treatment on that outcome* for an individual observation—and again, this impact can never be measured directly, as a single observation cannot both be treated and not treated. So, the causal tree takes a modified approach, and focuses on the average *difference in outcomes* between treated and non-treated observations within each leaf of the tree. Compared to a regular decision tree, the causal tree uses a splitting rule that explicitly balances two objectives: first, finding the splits where treatment effects differ most, and second, estimating the treatment effects most accurately. So, as the causal tree partitions the data it asks, "Where can we make a split that will produce the biggest difference in *treatment effects*, but which will still give an accurate estimate of the effect?" As a result, each terminal leaf of a causal tree represents a custom-made artificial experiment; in which the subjects of the experiment are as similar as possible; and where the average effect for that specific group helps predict the individual effect for future observations with the same features.


$\text{Outcome}_{\text{treat}}$
$- \text{Outcome}_{\text{non-treat}}$
$\text{Outcome}_{\text{treat}}$
$- \text{Outcome}_{\text{non-treat}}$

In addition, the procedure provides for valid confidence intervals and inference. Asymptotically, this requires that each tree is "honest." The basic idea is that you cannot use the same outcome data to *both* partition the tree *and* estimate the average impact—you have to choose one or the other. So, the causal-forest algorithm divides the data in two: half for determining the splits for the tree, and half for populating that tree with observations and then

[6] Athey & Imbens (2016); Athey, Tibshirani, & Wager (2018); Wager & Athey (2018).

estimating the treatment effect in each leaf. For an individual tree, it might seem that this procedure throws away half the data. But the RF algorithm typically builds thousands of individual trees, and uses a new bootstrap sample for each tree; so ultimately the algorithm ends up exploiting *all* the data for both splitting and estimation.[7]

IV. Case Study: The Impact of a Financial Crisis on Growth

> *It is essential to distinguish between triggers (the particular events or factors that touched off the crisis) and vulnerabilities (the structural weaknesses in the financial system and in regulation and supervision that propagated and amplified the initial shocks)… subprime losses were clearly not large enough on their own to account for the magnitude of the crisis. Rather, the system's vulnerabilities, together with gaps in the government's crisis-response toolkit, were the principal explanations of why the crisis was so severe.*
>
> *Ben Bernanke, 2010*

A. Risk vs. Vulnerability

Although often treated synonymously, crisis risk and crisis vulnerability are conceptually distinct. To take a common illustration, consider two neighboring houses on the Florida seaboard. Both face the same *risk* of a tropical storm or hurricane. But a sturdy house made of concrete is less *vulnerable* to a storm than a fragile house made of light weatherboard.

Of course, the situation for financial crises maybe less clear cut, as events do not always result from exogenous acts of nature, but may instead reflect the decisions of forward-looking actors—each of whom has information and incentives that are shaped by the actions of others. But the distinction is valuable nonetheless. Financial crises all reflect some confluence of an underlying economic vulnerability and a specific crisis trigger.[8] The underlying vulnerability can be something like an asset-price bubble or balance-sheet mismatch, but the trigger can be almost anything—political turmoil, terms of trade shocks, or even contagion from other countries (IMF, 2010).

Distinguishing empirically between triggers and vulnerabilities is difficult, and most studies implicitly combine the two—defining crises as those episodes in which an unpredictable trigger has *also* had a large impact, and so effectively exploring only those cases where underlying vulnerabilities are significant. This approach has formed the basis of the Early

[7] Wager & Athey (2018) also note that it is possible to build honest trees *without* double splitting, by training a *classification* tree that predicts the treatment *assignment* (W), rather than a *regression* tree that predicts the treatment effect (τ). The leaves of the resulting “propensity tree” include well-matched training observations based on their likelihood of inclusion in the treatment set, allowing for the *propensity score matching approach* of Rosenbaum and Rubin (1983), noted above.

[8] In other fields (e.g., security assessment), the trigger is often called a “threat,” so that “crisis risk” combines the likelihood of an event (threat level) with the impact of that event (vulnerability). (Crisis Risk = Threat × Vulnerability).

Warning Systems (EWS) literature, which accelerated sharply in the wake of the emerging-market turbulence of the 1990s.

Traditional early-warning models aim to anticipate crises ahead of time, and have typically relied on two approaches: discrete-choice (logit or probit) regression (see Eichengreen and Rose, 1998); or the signaling approach pioneered for the Fund by Kaminsky and Reinhart (1999). These have a number of advantages, including their ease of interpretation and widespread acceptance. But they also have a number of potential shortcomings: such as frequently large gaps between in-sample fit and out-of-sample predictive performance; or difficulty in coping with a large number of complex predictors.

As noted earlier, however, machine learning is well-suited to deal with these types of problems, and has featured more and more prominently in the crisis-prediction literature. The 2000s, for example, saw a marked pick up the exploration of non-linear decision trees for crisis forecasting, including in the Fund (Ghosh and Ghosh, 2003). And the use of machine learning tools has similarly flowered in the wake of the Global Financial Crisis, particularly in the examination of advanced-economy banking crises.[9]

But the above models focus primarily on *prediction*. What if, instead of concentrating on the likelihood of a crisis shock, we were instead interested in gauging a country's exposure to that shock *once it arrives*. The latter question is more closely associated with the concept of vulnerability; and in a forward-looking world where the probability of a trigger may not be entirely exogenous, the factors that shape the arrival of that trigger are perhaps not the same as those shaping the ultimate cost of the crisis. Of course, estimating the impact of a crisis requires that we also estimate what *would have happened* if the crisis had not occurred—which requires a more causal framework.

B. Estimating the Cost of a Crisis

In the economics literature, empirical studies into the impact of financial crises have typically taken one of two approaches: i) a dummy-variable approach; or ii) a comparison of actual post-crisis growth to a pre-crisis trend. The former generally draws on a cross-country panel regression, and uses the coefficient of a crisis dummy to indicate the growth cost of a crisis.[10] Adequately controlling for potential confounding variables, then, the estimated counterfactual is an average for the sample as a whole.

But some crises are more severe than others, so the second method takes a more country-specific approach; estimating the output loss for each country by measuring the gap between that country's actual growth and the rate that *would have been expected* based on prevailing trends. The identification of this trend, then, defines the counterfactual. Unfortunately, there

[9] See Savona and Vezzoli (2015), Alessi and Detken (2018).

[10] See Huchison and Noy (2002), or Demirgüç-Kunt and others (2006).

is no well-established method to do this. Some studies simply project a linear trend of recent GDP, excluding the years immediately prior to the crisis (Abiad and others, 2009). Some take the trend from an HP filter (Laeven and Valencia, 2018), while others use more detailed time-series forecasting techniques (Cerra and Saxena, 2008). In any event, results are often sensitive to the method chosen.

Moreover, few of these studies have then gone on to explore which factors are most important in shaping the scale of the output loss, and so have little insight to offer countries that have not (yet) experienced a crisis. The closest in spirit to the current paper is the study by Abiad and others (2009), who use Bayesian Model Averaging (BMA) to cope with a large number of potential confounding variables, in circumstances where theory provides an insufficient guide as to which variables belong in the real regression. The data and computational requirements of BMA, however, are formidable, so the authors use the procedure primarily as a robustness check on smaller-scale OLS regressions, which themselves consider only one or two variables at a time.[11] Further the authors do not consider potential non-linear links between any of their variables and crisis-related output losses.

In the sections below, we will use a causal-forest framework to estimate the impact of financial crises on growth. The "treatment" in this case is the crisis. The "outcome" is the short-run trajectory of growth immediately following the crisis. Considering a broad range of potential covariates, and considering also any potential nonlinearities or interactions, the causal forest provides an individual-country counterfactual—drawing on the experience of similar countries to estimate what *would have happened* to output growth in that country had the crisis been avoided. Similarly, for any country under consideration, the model provides an estimate of the hypothetical cost of a financial crisis, even if that crisis is itself unlikely.

C. Data

The financial-crisis "treatment" variable is taken directly from Laeven and Valencia (2018), who provide a standard crisis definition that can be applied consistently across a broad range of countries. Under their classification scheme, financial crises meet two sets of criteria: evidence of significant financial distress; and a significant policy intervention. The Laeven and Valencia dates are widely used in the literature, placing this model well in the broader context of research on financial crises. Crisis frequencies are broadly consistent across income groups, averaging one crisis every 30 to 45 years. But crisis patterns have varied considerably over time, with events in the 1980s and 1990s concentrated mainly among emerging markets (EM) and lower-income countries (LIC), and with crises since 2000 clustered mainly during the GFC and mostly among advanced economies (AE).

[11] Although they share many similarities, Bayesian Model Averaging is *not* equivalent to ensemble machine-learning (such as random forests), and typically performs less well at prediction. (Minka, 2000)

The outcome variable is the cumulative rate of output growth in the two years immediately following the crisis (including the crisis year)—this is typically the amount of time taken for growth to revert to pre-crisis trend (Demirgüç-Kunt and others, 2006).

For our set of possible confounding covariates, we make few a priori assumptions and instead consider a broad range of variables across a variety of sectors (financial, external, fiscal, real). While we have endeavored to ensure the widest possible coverage, data availability poses a significant constraint—there are many potentially informative variables with short histories (financial soundness indicators) or limited country coverage (housing prices), or both. Balancing these considerations, our final dataset includes 46 variables over 1985–2017, and covers 107 countries from both emerging markets and advanced economies, for a total of over 3300 observations. Details of the included variables are provided in Annex 1. A key feature of our dataset, however, is that crises represent a very small minority of our sample (around 2½ percent of all observations). Although the Random Forest algorithm is relatively robust to this problem, a large imbalance may nonetheless hamper the ability of the algorithm to accurately estimate the impact of a crisis. (More concretely, with an extremely imbalanced dataset, there is a chance that an individual bootstrap sample will contain few or even none of the minority class). We therefore improve the balance of the dataset by resampling, raising the overall proportion of crises from 2½ percent to almost 10 percent.[12] For the model, all covariates are lagged by one period to ensure that they are not themselves affected by the treatment.

D. Results: Explaining a "Black Box" Model

Individual Estimates

The estimated impact of a financial crisis for a sample of countries is presented in Figure 1—again, the aim of the model is not to predict the likelihood of a crisis, but instead to estimate the likely *cost* of a hypothetical crisis, *if* one were to occur. The figure shows the potential cost for these countries in 2015 (the cutoff year in the dataset), although it would also be possible to provide an estimate for any country, in any year.

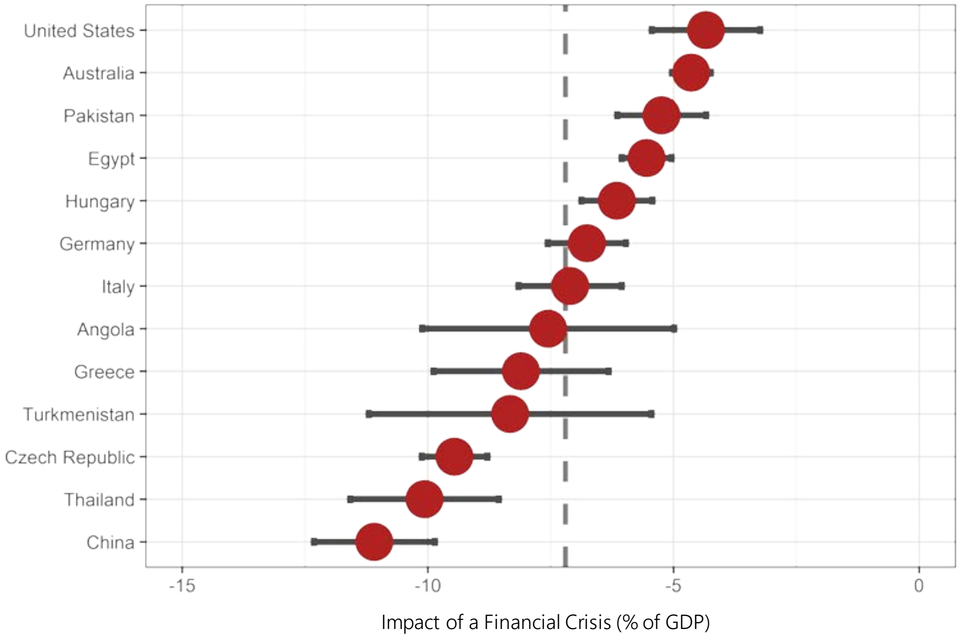
For the sample as a whole, the average cumulative cost of a crisis is estimated at 7.2 percentage points of growth over two years; which is broadly consistent with previous studies. Hutchison and Noy (2002) suggest a 2-year cost of 6–7 percentage points, while Demirgüç-Kunt and others (2006) estimate a 2-year cost of 7½ percentage points. The

[12] There are many ways of tackling class imbalance, but typical resampling solutions involve either: oversampling (increasing the number of minority-class observations) or downsampling (decreasing the number of majority-class observations). We use a synthetic oversampling technique (SMOTE), which creates synthetic samples that are similar to existing minority observations, rather than simply generating new copies that are exactly the same as the minority observations. This ensures that the new oversampled observations accurately reflect the topology of the dataset surrounding each minority observation, while also minimizing the risk of overfitting (Chawla and others, 2002).

baseline estimate in Abiad and others (2009) is slightly lower (4½ percentage points) but interestingly, when the counterfactual trend is calculated using the growth projections of Fund country desks, the estimate from the Abiad study is closer to ours, at almost 8 percentage points.

As illustrated by the figure, the cost of a hypothetical crisis varies significantly from country to country. Conditional on a crisis occurring, some countries would face a significant drop in growth, while others would get off relatively lightly. In addition, as noted above, the causal forest also provides confidence intervals for each estimate, and these intervals also vary from country to country.[13] Intuitively, the algorithm calculates a country's estimated crisis impact by gathering similar countries into a group, and then comparing those who had a crisis to those that did not. This process is repeated across thousands of bootstrapped samples, and the different estimates are then aggregated. If there are many countries with the same characteristics, then the groups for that country will be broadly the same for each sample, and so the model will be relatively confident about the final estimate. If the country has few comparable peers, however, then the bootstrapped groups for that country will tend to be more diverse, and so the confidence interval will be wider.

Figure 1. Causal Forest Predictions (Select Countries)



Source: IMF VE Database, author's calculations

[13] For a formal treatment of the calculation of confidence intervals within the causal forest framework, see Athey and others (2018).

Predicting the Severity of a Hypothetical Crisis

Having determined the individual treatment effect (ITE) of a crisis for each country, the next obvious question is *why* some countries are supposed to have a more costly crisis than others. But the causal forest model consists of thousands of individual trees, each of which is highly nonlinear, making it difficult to determine the key drivers behind any particular prediction.

In this context, a parallel strand of the machine learning literature has focused on methods to make complex ("black box") models as transparent and accessible as possible.[14] These can be applied to any model, simple or complex, and essentially rely on repeated simulation—altering the value of a particular input variable, and keeping track of what happens to the model's output. With enough well-chosen simulations it is possible to gain an intuitive understanding of: which variables in the model are most important; the effect of each variable in shaping any individual prediction; and the effect of each variable over a large number of possible predictions.

For our purposes, we are predicting the cost of a hypothetical crisis (rather than its likelihood) and want to explore which variables feature most prominently in determining that cost. There are a range of methods available, but an increasingly popular technique has its origins in cooperative game theory (Box 4). In particular, Shapley values capture the contribution of each individual covariate to the difference between the actual prediction, on the one hand, and the mean prediction for the sample, on the other.[15] These values are consistent and additive, and so allow for a complete breakdown of the factors behind any particular prediction. Moreover, Shapley values are the only estimators with a solid theoretical basis. In interpreting these results it should be stressed that the values are tailored to each specific instance. For example, two countries may have the same value for a particular covariate (e.g., current account deficit), but nonetheless may have different Shapley values, as that variable may matter *more* for one of the countries, owing to its own particular circumstances (e.g., high external debt).

As an illustration, we include Shapley-value breakdowns for a couple of countries in Figure 2. To repeat, the model says nothing about the likelihood of either of these countries experiencing a crisis. It simply outlines the impact of a hypothetical crisis, if it were indeed to occur. Compared to the "average" crisis, for example, the model suggests that Australia would experience a relatively mild episode, with an output loss of only 4.6 percentage points (compared to a sample average of 7.2 points). The chart shows the five most important factors that prompt the model to assign this lower estimate, chief among which are Australia's flexible exchange rate, low manufacturing share, and high trade openness. Other

[14] See Molnar (2019) for a survey.

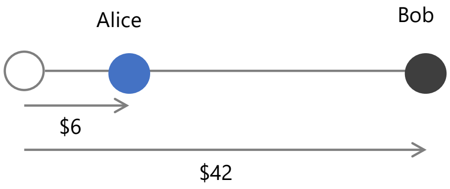
[15] Sometimes misinterpreted, the Shapley value is the average contribution of a particular covariate value to the prediction, across all possible coalitions of variables. It is *not* the difference in prediction when we remove the variable from the model.

mitigating factors include Australia's fiscal space (as proxied by moderate government consumption over the past five years) and relatively small output gap—again, these are all pre-crisis initial conditions.

Box 4. Assigning the Blame: Shapley Values

Prediction is hard. Explaining a prediction is even harder. This is particularly problematic in the case of complex models, where certain variables (e.g. oil prices) may only matter for certain types of countries (e.g. oil importers), or where the impact of a particular variable (current account deficit) may depend on the value of another variable (external debt), or indeed where the final prediction is the result of an *ensemble* of thousands of sub models. In such circumstances there is often some tension between a model's *accuracy* and its *interpretability*. In response, various methods have been proposed to help users interpret the predictions of complex models, and one of the most promising has its origins in cooperative game theory.

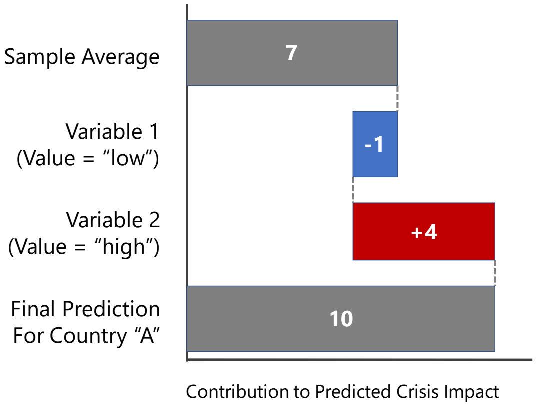
Shapley Values initially came out of a core question in game theory: in a coalition of multiple players with differing skill sets, what is the fairest way to allocate a collective payoff? One solution is to imagine the players joining in sequence, and then keeping track of their marginal contribution. But what if some players, say Alice and Bob, have similar skill sets? Then, it might be the case that Alice would have a higher marginal contribution if she joined the group before Bob, as she would be the first one to provide their overlapping skill set. When Bob joined, his marginal contribution would be lower. The Shapley Value concept was developed in response to this problem, and can be understood as *finding each player's marginal contribution, averaged over every possible sequence in which the players could have been added to the group*. To take the simplest example, suppose Alice and Bob are sharing a taxi, and Alice lives on the route to Bob's house. Their marginal contribution to the cost of the taxi ride will depend on the order in which their claims are considered. The Shapley Value, however, will average these contributions over all conceivable sequences (in this case there are only two) to arrive at the fairest possible allocation.



	Alice Pays	Bob Pays
Alice goes first	$6	$36
Bob goes first	$0	$42
Shapley	$\frac{\$0+\$6}{2} = \$3$	$\frac{\$36+\$42}{2} = \$39$

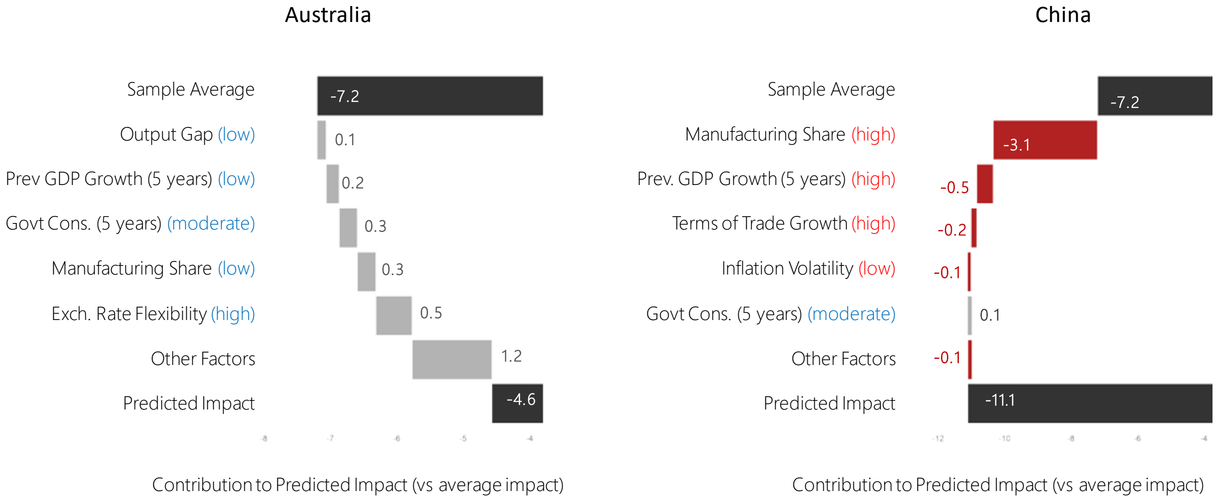
This concept can be used to explain the contributions of different variables to an individual prediction. The "payoff" is the actual prediction for a particular instance *less* the average prediction for the entire dataset. The "players" are the values of each variable that fed into that prediction, which "collaborate" to produce the overall payoff. Shapley Values divide this prediction (payoff) among the variables (players) in a way that fairly represents their respective contributions. In the case of a crisis, for example, where a country's estimated cost is higher than for the sample as a whole, Shapley Values will indicate which variables prompted the model to assign a higher cost for that country, and will provide a quantitative guide as to each variable's relative contribution.

Shapley Values: Impact of a Crisis (% GDP)



China, on the other hand, would be expected to experience a more severe crisis, mostly owing to the country's high manufacturing share (the model suggests that the manufacturing sector is relatively more sensitive to a disruption in financial flows). Additional worrisome factors include rapid GDP growth, a sharp increase in China's terms of trade, and volatile inflation, although the cost of the crisis may be ameliorated slightly by the potential availability of fiscal space.

Figure 2. Shapley Values: Impact of a Banking Crisis on Output Growth



Source: IMF VE Database, author's calculations

Aggregate Findings

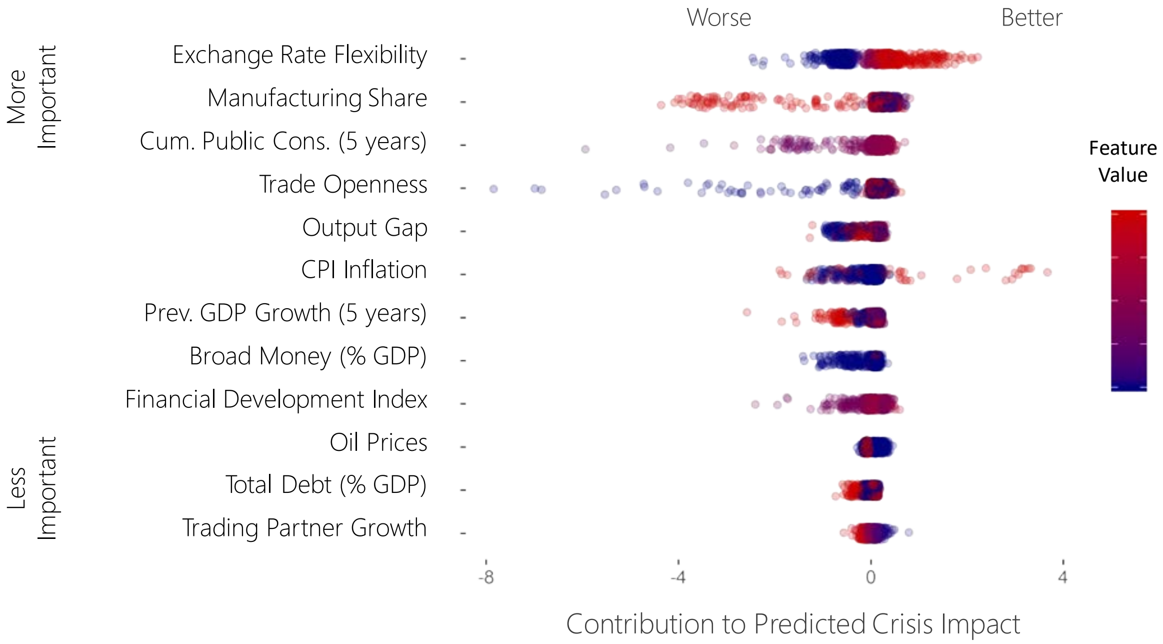
With similar breakdowns available for every observation, we can then explore which factors tend to be most important for the sample as a whole. For example, Figure 3 plots individual Shapley values for the entire dataset. Variables which are more important in driving the model's predictions will tend to have a wider variance in their individual contribution—unimportant variables will have little impact, either positive or negative, and so will have Shapley values clustered around zero.

The chart shows the 12 most important variables, with higher ranked variables placed at the top.[16] For example, it seems that exchange-rate flexibility is the *most* important factor overall in shaping the impact of a banking crisis, with greater flexibility associated with a smaller output loss (consistent, perhaps, with the shock-absorber role highlighted in Edwards and Yeyati, 2003).

[16] Variables are ranked according to the mean absolute value of their respective Shapley values.

Also important is the share of manufacturing in each country's economy, as well as the cumulative amount of public-consumption spending in the 5 years prior to the crisis, and the country's degree of trade openness. For manufacturing, the distribution seems somewhat skewed, suggesting that the impact of this variable may be nonlinear—with the manufacturing share only having an effect if it exceeds a certain threshold, at which point a larger share is associated with a more severe episode.[17] Similarly for trade openness, it appears that openness has little effect except for countries below a certain threshold, at which point relatively closed countries will tend to have more costly crises. Public spending is also skewed, but displays a less obvious mapping with the severity of a crisis. This is consistent with a more hump-shaped relationship, with both excessively high and excessively low levels of spending associated with more costly crises (perhaps owing to the fact that they may indicate a potential lack of fiscal space going forward).

Figure 3. Distribution of Shapley Values



Source: IMF VE Database, author's calculations

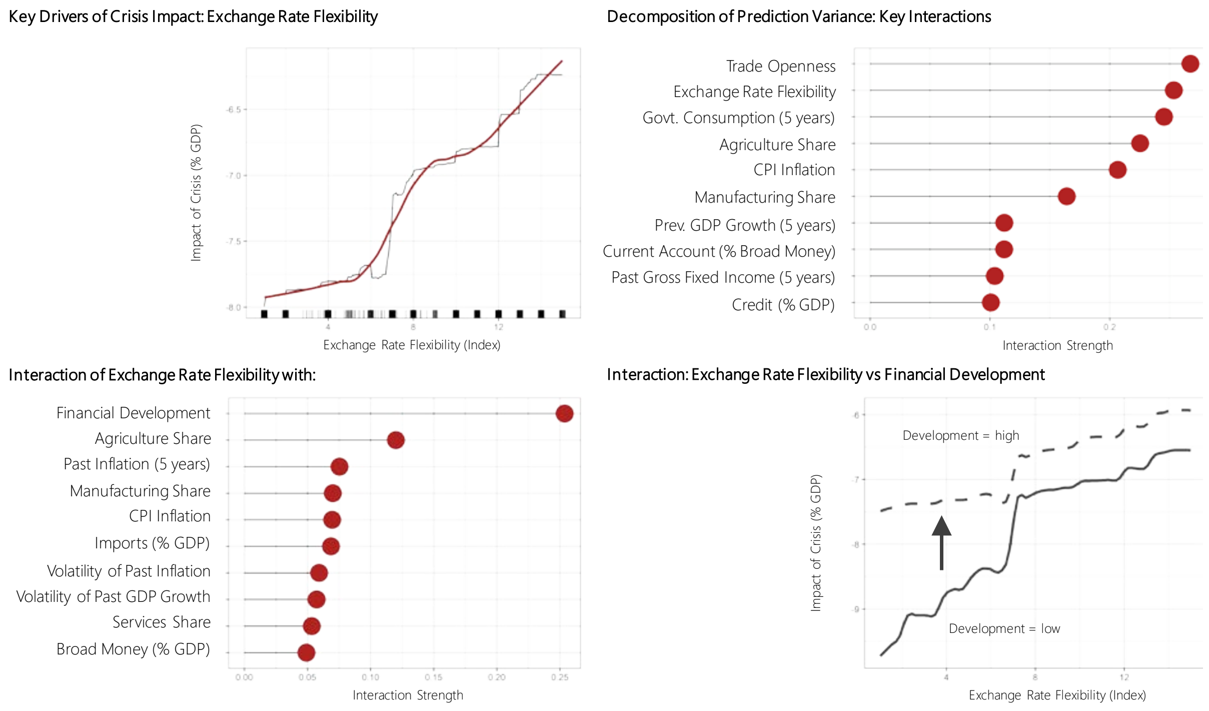
Having identified the most important variables for the model as a whole, the machine-learning literature has also developed techniques to summarize each relationship in more detail. A comprehensive exploration of each variable is beyond the scope of this paper, but we can

[17] The importance of a country's manufacturing share is consistent with the general argument of Rajan and Zingales (1998) on the role of financial development in promoting growth—that financial development disproportionately helps firms or industries that depend on external financing (i.e., non retained funds). Manufacturing firms are relatively exposed in this regard (Balta and Nikolov, 2013), and so may tend to suffer the most from a sharp disruption in the financial system.

illustrate some of these techniques by delving deeper into the relationship between exchange-rate flexibility and the likely cost of a crisis.

Figure 4 below (top left) presents a partial dependence plot (PDP) for the exchange-rate variable, which shows the effect of that variable on the predicted impact of a crisis—partial dependence works by marginalizing the model's output over the distribution of all non-exchange-rate variables. As suggested in Figure 3, a greater degree of flexibility is associated with a milder episode.

Figure 4. Exploring the Role of Exchange-Rate Flexibility



Key Drivers of Crisis Impact: Exchange Rate Flexibility

Decomposition of Prediction Variance: Key Interactions

Interaction of Exchange Rate Flexibility with:

Interaction: Exchange Rate Flexibility vs Financial Development

Sources: IMF VE Database; author's calculations.

We can also use simulation techniques to explore potential interactions and nonlinearities. Figure 4 (top right) summarizes the degree to which each variable interacts with others, based on the variance-decomposition procedure introduced by Friedman and Popescu (2008)[18,19] In our case, even though the partial dependence plot for exchange-rate flexibility

[18] Interactions between features are measured via the decomposition of the prediction function: If a feature j has no interaction with any other feature, the function can be expressed as the sum of the partial function depending only on j and the partial function depending on all features other than j. Any variance not explained can be attributed to the interaction and is used as a measure of interaction strength.

[19] The PDP procedure used in this paper, as well as all Shapley-value implementations and variance decompositions, are drawn from the "iml" package in R, authored by Molnar (2018).

appears relatively monotonic, the procedure suggests that the effect of this variable is nonetheless strongly influenced by the level of other variables. Digging deeper, we can use the same decomposition procedure to explore which interactions are most important in this regard (Figure 4, bottom left). The results suggest that the impact of flexibility on the cost of a crisis is shaped by a country's level of financial development, with the potential shock-absorber role of the exchange rate significantly reduced for countries with a lower level of development (Figure 4 bottom right).

V. Further Research

This paper is not intended as a major contribution to the literature on financial crises. Instead, the goal has been to provide a gentle introduction to some recent advances in machine learning—advances that promise to be directly applicable to the challenges routinely faced by policy-oriented economists. In our case, we have chosen financial crises as an illustrative application, but other possibilities are much more wide ranging, and will likely be limited only by the availability of suitable data

To the best of our knowledge, we are the first to apply the causal-forest algorithm to a macroeconomic dataset, but more micro-level data might allow for similar individual-level analysis of potential interventions (such as labor-market reforms, changes to tax policy, and so on).

We should also note here that the causal-forest algorithm is not limited to analysis of binary interventions. In the event of a continuous treatment variable, for example, the procedure is modified easily—in each leaf of the causal tree, instead of calculating the *difference* in outcomes, the algorithm instead calculates the *covariance* between the treatment variable and the outcome variable. The end result for the individual treatment effect, then, is an estimate of the partial impact of a one-unit increase in the treatment variable. Similarly, if the researcher suspects that the variable of interest is endogenous, the causal forest can accommodate an instrumental-variables approach, in which the quantity calculated in each leaf is simply the IV estimator.[20] In the latter case, the procedure is called "instrumental forests."

VI. Conclusion

Traditionally, prediction and causal inference have been treated as two very separate problems. Most of the prediction literature assumes that predictions are made by a passive observer who has no influence on the outcome. Causal inference, on the other, explicitly asks what would happen if we actively try to *intervene* in the system; or equivalently, *what interventions* most likely led to a particular outcome. The distinction is subtle, but

[20] This is $\left[{}^{cov(outcome, instrument)}\!\big/_{cov(treatment, instrument)}\right]$

fundamental—a crowing rooster may be a good *predictor* of an impending sunrise, but intervening on the rooster won't in itself stop the sun from rising.

In this context, the machine-learning literature has typically concerned itself mostly with prediction, but has recently focused more and more on issues of causality—prompted in part by a growing realization that many problems originally thought to be predictive (e.g., what happens to online sales if we increase a particular form of advertising) actually entail the discovery of a more causal relationship.

Causal discovery, of course, is a vital part of empirical economic analysis, and this paper has provided a gentle introduction to some recent advances in the growing field of "causal machine learning." Focusing mostly on the Causal Forest algorithm, and taking the costs of a financial crisis as an illustrative example, the paper has endeavored to show how such techniques can produce plausible results—the estimated average impact of a crisis, for example, is consistent with previous estimates, as is the potential role for exchange-rate flexibility or financial development in shaping the cost for any particular country. Further, these techniques can also allow for a significantly richer discussion of potential thresholds and non-linearities; to an extent that is usually not feasible using more traditional econometric methods. More generally, by enabling the consideration of a rich set of variables and interactions, and allowing for a more tailored assessment of the individual circumstances of each country, these techniques can provide an invaluable complement to the techniques currently employed by economists, both within the Fund and beyond.

References

Abadie, Alberto, and Guido W. Imbens, 2006, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, Vol. 74(1), pp. 235–267.

Abiad, Abdul, Petya Koeva Brooks, Irina Tytell, Daniel Leigh, and Ravi Balakrishnan, 2009, "What's the Damage? Medium-Term Output Dynamics After Banking Crises," IMF Working Paper No. 09/245 (Washington: International Monetary Fund).

Alessi, Lucia & Carsten Detken, 2018, "Identifying Excessive Credit Growth and Leverage," *Journal of Financial Stability*, Vol. 35(C), pp. 215–225.

Athey, Susan and Guido W. Imbens, 2016, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, Vol.113 (27), pp. 7353–7360.

———, 2017, "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, Vol. 31 (2), pp. 3–32.

Athey, Susan, Julie Tibshirani, and Stephan Wager, 2019, "Generalized Random Forests," *Annals of Statistics*, Vol.47 (2), pp. 1148–1178.

Balta, Narcissa and Plamen Nikolov, 2013, "Financial Dependence and Growth Since the Crisis," *Quarterly Report on the Euro Area*, Directorate General Economic and Financial, European Commission, Vol. 12 (3), pp. 7–18.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, 2012, "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, Vol. 80 (6), pp. 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, 2014a, "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28 (2), pp. 29–50.

———, 2014b, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, Vol. 81 (2), pp. 608–650.

Breiman, Leo, 2001a, "Statistical Modeling: The Two Cultures," *Statistical Science*, Vol. 16 (3), pp. 199–231.

———, 2001b, "Random Forests," *Machine Learning*, Vol. 45 (5), pp. 5–32.

Cerra, Valerie, and Sweta Chaman Saxena, 2008, "Growth Dynamics: The Myth of Economic Recovery," *American Economic Review*, Vol. 98 (1), pp. 439–57.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, 2002, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* Vol. 16, pp. 321–357.

Demirguc-Kunt, Asli, Enrica Detragiache, and Poonam Gupta, 2006, "Inside the Crisis: An Empirical Analysis of Banking Systems in Distress," *Journal of International Money and Finance*, Vol. 25 (5), pp. 702–718.

Edwards, Sebastian, and Eduado Levy Yeyati, 2005, ”Flexible Exchange Rates as Shock Absorbers," *European Economic Review*, Vol. 49 (8), pp. 2079–2105.

Friedman, Jerome H, and Bogdan E Popescu, "Predictive Learning Via Rule Ensembles," *The Annals of Applied Statistics,* Vol. 2 (3), pp. 916–54.

Ghosh, Swati R., and Atish R. Ghosh, 2002, "Structural Vulnerabilities and Currency Crises," IMF Working Paper No. 02/9 (Washington: International Monetary Fund).

Holland, P. W., 1986, Statistics and Causal Inference, *Journal of the American Statistical Association*, Vol. 81 (396), pp. 945–960.

Hutchison, Michael, and Ilan Noy, 2005, "How Bad Are Twins? Output Costs of Currency and Banking Crises," *Journal of Money, Credit and Banking*, Vol. 37 (4), pp. 725–52.

Imbens, Guido W., and Donald B. Rubin, 2015, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press).

IMF, 2010, *The IMF-FSB Early Warning Exercise: Design and Methodological Toolkit*, International Monetary Fund.

Kuhn, Max and Kjell. Johnson, 2013, *Applied Predictive Modeling*. New York: Springer.

Laeven, Luc, and Fabian Valencia, 2018, "Systemic Banking Crises Revisited," IMF Working Paper No. 18/206 (Washington: International Monetary Fund).

Minka, Thomas P., 2000, "Bayesian Model Averaging Is Not Model Combination," http://www.stat.cmu.edu/minka/papers/bma.html.

Morgan, Stephen L., and Christopher Winship, 2015, *Counterfactuals and Causal Inference: Methods and Principles for Social Research,* 2nd ed. (Cambridge University Press).

McGue, Matt, Merete Osler, and Kaare Christensen, 2010, "Causal Inference and Observational Research: The Utility of Twins," *Perspect Psychol Sci*. Vol. 5 (5), pp. 546–556.

Molnar, Christoph, 2018, "iml: Interpretable Machine Learning," R package version 0.5.1. https://CRAN.R-project.org/package=iml.

———, 2019, *Interpretable Machine Learning*, https://leanpub.com/interpretable-machine-learning.

Pearl, Judea, 2009, *Causality: Models, Reasoning, and Inference*. 2nd ed. (Cambridge University Press).

Rajan, Raghuram G., and Luigi Zingales, 1998, "Financial Dependence and Growth," *American Economic Review*, Vol. 88 (3), pp. 559–586.

Rosenbaum, Paul R., and Donald B. Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70 (1), pp. 41–55.

Savona, Roberto and Vezzoli, Marika, 2015, "Fitting and Forecasting Sovereign Defaults Using Multiple Risk Signals," *Oxford Bulletin of Economics and Statistics*, Vol. 77 (1), pp. 66–92.

Tiffin, Andrew J., 2016, "Seeing in the Dark; A Machine-Learning Approach to Nowcasting in Lebanon," IMF Working Paper No. 16/56 (Washington: International Monetary Fund).

Varian, Hal R. 2014, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, Vol. 28 (2), pp. 3–28.

Wager, Stephan, and Susan Athey, 2018, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, Vol. 113 (523), pp. 1228–1242.

Wyss, Richard, Alan R. Ellis, M. Alan Brookhart, Cynthia J. Girman, Michele Jonsson Funk, Robert LoCasale, and Til Stürmer, 2014, "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score," *American Journal of Epidemiology*, Vol. 180 (6), pp. 645–655.

ANNEX I

Data are taken from the Fund’s *Vulnerability Exercise Database*, with variables chosen to ensure a broad coverage: across sectors, across countries, and across time. The final dataset includes 46 variables over 1985-2017, and covers 107 countries from both emerging markets and advanced economies, for a total of over 3364 observations.

Table 1. Causal Forest: Features

External Variables

- Import Growth (goods and services, annual percent change)
- Export Growth (goods and services, annual percent change)
- Terms of Trade (annual percent change)
- Trading Partner GDP (annual percent change)
- Export Prices (annual percent change)
- Trade Openness (exports plus imports/GDP)
- Oil Prices (average Brent, WTI, Dubai)
- Financial Account, Net Financial Liabilities (percent of GDP)
- Oil Imports (percent of GDP)
- Oil Exports (percent of GDP)
- Goods and Services Imports (percent of GDP)
- Goods and Services Exports (percent of GDP)
- Current Account Balance (percent of GDP)
- Net FDI (percent of GDP)
- International Reserves (percent of IMF ARA metric)
- Gross Foreign Exchange Reserves (percent of GDP)
- Exchange Rate Flexibility (index)

Fiscal Variables

- Central Government Debt (percent of GDP)
- Public Consumption (real, 5-year cumulative growth, relative to GDP)
- External Debt (percent of GDP)
- Fiscal Balance (percent of GDP)

Financial Variables

- Current Account Balance (percent of broad money)
- Broad Money (percent of GDP)
- Private Debt, Loans, and Securities (percent of GDP)
- Federal Funds Rate (effective, percent)
- 3-month US T-bill Rate (percent)
- US Term Spread (bps)
- Financial Development (index)
- Bank Credit to Private Sector (percent of GDP)
- Credit Gap (deviation from trend, percent of GDP)

Real Variables

- Commodity Exporter (dummy)
- Consumer Price Inflation (annual average, percent)
- GDP Growth (real, 5-year cumulative)
- Output Gap (HP filter, percent of GDP)
- Consumer Price Inflation (annual percent, 5-year cumulative)
- Inflation Volatility (rolling, 4-year)
- Gross Fixed Capital Formation (real, 5-year cumulative growth, relative to GDP)
- Private Consumption (real, 5-year cumulative growth, relative to GDP)
- Domestic Demand (real, 5-year cumulative growth, relative to GDP)
- Income per capita (PPP, relative to USA)
- Gross Public Savings (percent of GDP)
- Gross Private Savings (percent of GDP)
- GDP Growth Volatility (5-year rolling)
- Services Value Added (percent of GDP)
- Manufacturing Value Added (percent of GDP)
- Agriculture Value Added (percent of GDP)