Economia Aplicada, v. 27, n. 3, 2023, pp. 407-423

EARLY WARNING SYSTEMS VIA MACHINE LEARNING: A STUDY OF CURRENCY CRISES

Vitor Ostrensky
Leonardo Matsuno da Frota *

Resumo

Sistemas de Alerta Antecipado (EWS) para crises cambiais é um tópico essencial em macroeconomia. Recentemente, houve uma renovação dessa literatura com a introdução de métodos de aprendizado de máquina. No entanto, argumentamos que a maioria dos trabalhos publicados tem métricas de precisão excessivamente otimistas causadas pela desconsideração da autocorrelação ou atrasos na publicação de dados. Nossa contribuição é construir um Sistemas de Alerta Antecipado baseado em um conjunto de modelos de aprendizado de máquina apropriados para dados de séries temporais. Usando dados de 25 países entre 1995 e 2020, nossas descobertas são mais modestas do que trabalhos recentes, mas destacam a utilidade e as limitações dos sistemas de alerta antecipado na prática.

Palavras-chave: crises cambiais; previsão; machine learning.

Abstract

Early Warning Systems (EWS) for currency crises is an essential topic in macroeconomics. Recently, there has been a renewal of this literature with the introduction of machine learning methods. However, we argue that most published works have overly optimistic accuracy metrics caused by disregarding autocorrelation or data publication lags. Our contribution is to build an Early Warning System based on an ensemble of machine learning models appropriate for time series data. Using data from 25 countries between 1995 to 2020, our findings are more modest than recent works but highlight the usefulness and limitations of Early Warning Systems in practice.

Keywords: currency crises; forecasting; machine learning.

JEL classification: C53, G01

DOI: http://dx.doi.org/10.11606/1980-5330/ea189768

* Universidade Federal do Paraná

Recebido em 20 de agosto de 2021 . Aceito em 13 de julho de 2022.

1 Introduction

Currency crises take a heavy toll on any country's asset markets and the real economy, with heavy losses in output and productivity, especially in emerging markets (NAKATANI, 2018). Therefore, predicting or anticipating such crises is relevant to policymakers and market participants. However, predicting the timing of significant currency depreciation is very challenging. Nevertheless, there is no shortage of attempts. In particular, there is much discussion on Early Warning Systems (EWS): tools that give advanced warning signals about an impending currency crisis Eichengreen, Rose, and Wyplosz (1995) and Kaminsky, Lizondo, and Reinhart (1998).

To build an EWS to predict a currency crisis, we need to know where to look for signals. The theoretical literature on currency crisis is typically divided into three main phases - or *generations* -, with each one having different explanations for what may cause a crisis. The first-generation begins with Krugman (1979) and Flood and Garber (1984). A crisis arises from the inconsistency of the conducted macroeconomic policy, specifically, the combination of fixed exchange rates with a public deficit financed by a monetary expansion in rising inflation. In the second generation models, a crisis arises when the social cost of maintaining the currency regime is higher than the benefit of keeping it (OBSTFELD, 1996). Last, the third generation is concentrated on the financial aspects of the economy. Variables like debt, capital inflows, private credit, and the behavior of asset markets come into play (KRUGMAN, 2003). Another concern is the contagion between countries. In this perspective, a crisis in one place can trigger crises in similar countries, especially those with geographical proximity and/or a strong commercial relationship (MASSON, 1998). The currency crises of the late 1990s are an example of this possibility.

Cuaresma *et al.* (2008) summarizes the relevant economic variables for each generation of currency crisis models: 1st generation - variables that show the fiscal deficits financed by domestic credit; in particular, high inflation (or high money growth), current account deficits, and real interest rate raises; 2nd generation - variables that try to infer the market sentiment and expectations that can lead to herding behavior or self-fulfilling crisis; finally, third-generation - variables that show liquidity problems or more broad balance sheet vulnerabilities in the financial sector.

Early Warning Systems are designed in a way that allows the monitoring of key indicators that are supposed to show out-of-the-ordinary behavior in the months or years preceding a crisis. The job then is to calculate threshold values for these indicators; when they go above it, a good Early Warning System should point out that there is trouble ahead (KAMINSKY; LIZONDO; REINHART, 1998). These models are constructed so that they could be able to set off an alarm before the occurrence of a currency crisis, giving policymakers, investors, and other stakeholders, sufficient time to implement adequate policies to avoid - or at least attenuate - the adverse effects of the crisis (CANDELON; DUMITRESCU; HURLIN, 2014).

Typically, these models were made with standard statistical or econometric tools, such as logit models (BUSSIERE; FRATZSCHER, 2006; CANDELON; DUMITRESCU; HURLIN, 2014; EMIN; AYTAÇ, 2016; BOONMAN *et al.*, 2019), signal approaches (EICHENGREEN; ROSE; WYPLOSZ, 1995) and Markov Switching (DU; YU; LAI, 2020). Recently, with the rise of machine learning tools and applications in finance and macroeconomics, there has been an opportunity to reassess the Early Warning System framework with novel techniques such as random forests and neural networks, which are remarkably efficient in making predictions in several areas[1]. In the last few years, some papers have used models from the machine learning framework to predict currency crises. The methodological approach is quite different among them, given that several types of models have been used, such as decision trees (KINKYO, 2020), support vector machine (RAMLI; ISMAIL; WOOI, 2015) and neural networks (SEVIM *et al.*, 2014; ALAMINOS *et al.*, 2019).

However, there is still room for advances in terms of the methodological approach to Early Warning Systems. One of the most recurring problems is the frequency of the data used. The foreign exchange market is quite volatile and sensitive to shocks, which makes forecasts on an annual, half-yearly, or even quarterly basis very ineffective. In this respect, many papers do not consider the data publication lag in the analysis. Since data over a period (excluding financial market data) will only be available many weeks later, or even a couple of months later, the forecast cannot be made considering that the data was available on its reference date. Another issue in the literature is the indiscriminate use of the K-fold validation technique in macroeconomic and financial data, without accounting for the time structure of the series. A procedure that can lead to inaccurate estimates when the data is non-stationary Cerqueira, Torgo, and Mozetič (2020). Also, macroeconomic data usually shows some considerable percentage of missing observations for some countries and variables and there is little clarity on the major part of the literature on how they deal with this problem. Finally, many papers use accuracy as the model's performance metric. The problem is that accuracy is not easy to interpret with unbalanced classes, which is the case for currency crises.

Despite these methodological issues, most authors claim to have excellent forecasting results, so a question arises. Is it possible to predict exchange rate crises with such accuracy, or is this performance caused by methodological flaws? To answer this question, besides looking for alternatives to the mentioned problems, we combine standard econometric tools such as a logit model with novel machine learning algorithms such as random forests and neural networks to produce an Early Warning System for currency crises. Our results indicate that much of the literature may present models with overestimated forecast performance, resulting in Early Warning Systems that would probably not be workable in practice. Nevertheless, we show that it is possible to produce forecasts with practical applicability and predictive power, when the temporal structure of the data is taken into account.

The rest of this paper is organized as follows. Section II explains the methodological structure of the model as well as the data used in the models. Sections III gives model output examples and how to interpret them. Section IV concludes and Section V is devoted to the data appendix.

[1] It is worth mentioning that the theoretical development of these techniques is not recent. However, because of an increase in computational capacity in recent years, there has been a great popularization of them.

2 Data

Several different forecasting frequencies have been used in the literature: Annual Emin and Aytaç (2016) and Alaminos *et al.* (2019), semiannual Kinkyo (2020), quarterly Ramli, Ismail, and Wooi (2015), Chong and Yan (2018) and Boonman *et al.* (2019) and monthly Sevim *et al.* (2014) and Du, Yu, and Lai (2020). However, we did not find any published work with data frequency higher than monthly. In a forecasting model, the frequency of the data is directly linked to the objective and purpose of the forecast. A model with annual data, for example, most likely will have an annual update of forecasts when used in practice. This can be especially damaging to the model's forecasting ability in the event of currency crises, as, unlike other macroeconomic variables, substantial changes can occur suddenly. Countries can start a year with stability and, in a few months, grow into a situation of greater risk of strong exchange rate depreciation. An example of this is the exchange rates depreciation that occurred in several countries in April 2020, after the coronavirus outbreak. Therefore, as it is important for the usefulness of the models that they have higher frequencies, we choose to estimate our model with a monthly time frequency. However, this also requires that data be available more frequently for countries that are the focus of the forecast, which is not always true.

We based the choice of countries for the study on two criteria. The first is to be a country more vulnerable to this type of event. Therefore, with this criterion, countries in the Eurozone, the United Kingdom, the United States, Japan, and others without a history of major exchange rate depreciation were excluded. The second criterion was to select countries that had sufficient data for the forecast. In international macroeconomic databases, low-income and small countries typically have a higher proportion of missing data, so we chose to remove them from the analysis. The final sample has 25 countries:

America

- Argentina
- Brazil
- Chile
- Colombia
- Mexico
- Peru
- Uruguay

Asia

- China
- Korea
- India
- Myanmar
- Indonesia
- Israel
- Philippines
- Pakistan
- Thailand
- Turkey

Europe

- Czechia
- Hungary
- Poland
- Romania
- Russia

Africa

- Egypt
- Nigeria
- South Africa

Our sample corresponds to the period of 01/1995-12/2020, being divided between in-sample (training set), from the start to 12/2015, and out-of-sample (test set), from 01/2016 to 12/2020. As our time frequency is monthly, we mostly use variables with this frequency. However, variables with lower frequencies, such as GDP and debt data, were included. For this, we interpolate the quarterly and annual data using the Kalman filter method Gómez and Maravall (1994). Also, we use financial variables, such as the VIX and the oil price, that have higher frequencies. In these cases, we use the closing value of the previous month. Most of the variables underwent transformations to make them stationary, such as year-over-year percentage change. The indicators, the transformations, and the source of all variables are presented in Table A.2 in the appendix.

Most papers do not use real-time data, that is, considering the typical lag in making macroeconomic data available. This problem can seriously affect the forecast, since the model may end up being trained by observing data it would not observe in practice Bańbura and Rünstler (2011). However, it is not simple to build a vintage database for several countries with so many variables, since for each of these there will be a different lag for data dissemination, not to mention the issue of data revisions. Therefore, we consider an average difference of 2 months between the month of correspondence of the macroeconomic data and its release. Financial variables do not have a disclosure delay so they were not affected.

3 Method

3.1 Crisis definition

The first step in designing an Early Warning System is the precise definition of a currency crisis (BOONMAN *et al.*, 2019). We tested two different metrics that are used in the literature. First, the most straightforward option is to define it as a substantial devaluation of the nominal exchange rate in a relatively short period (i.e., 20 percent in a month). The nominal depreciation is a value that can be easily understandable and comparable between different countries or periods. However, it does not capture downward currency pressures that are dealt by the central banks' policies. For that reason, we also included the Exchange Market Pressure index Kaminsky, Lizondo, and Reinhart (1998).

From a macroeconomic point of view, it is more important to know whether there will be a crisis within a certain horizon than in a certain month because this period allows the authorities to prevent the crisis (CANDELON; DUMITRESCU; HURLIN, 2014). Therefore, our goal is to predict whether the month in question precedes a crisis, using a predetermined time interval, which in this case is 12 months. Thus, our first definition of the dependent variable is constructed as follows:

$$Y_{i,t} = \begin{cases} 1 & \text{if there is any month in next 12 with a depreciation above the defined threshold,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where $Y_{i,t}$ is our dependent variable for country i in period t.

It is also necessary to define the thresholds. As a way of measuring the effect of different crisis specifications, we used three different values: 10%, 15%, and 20%. There is no exact definition of an exchange rate crisis using the percentage devaluation of a currency in one month[2]. Thus, we are not

[2]Kaminsky, Lizondo, and Reinhart (1998) uses 25% in a year, for instance. We have tested larger

only concerned with capturing heavy crises but also with extraordinary depreciations.

However, as we mentioned above, sometimes countries spend a vast amount of resources in terms of international reserves to defend their currencies in some given range. This way, Kaminsky, Lizondo, and Reinhart (1998) incorporated this variable to create the concept of Exchange Market Pressure (EMP)[3] to gauge currency crises. So, for any country the EMP index is:

$$\boldsymbol{EMP}_t = \frac{\Delta e_t}{e_t} - \frac{\sigma_{\Delta e,t}}{\sigma_{e,t}} \tag{2}$$

where $\frac{\Delta e_t}{e_t}$ is the rate of change in the exchange rate in period t in relation to US dollars; $\sigma_{e,t}$ is the standard deviation of the rate of change calculated until period t; $\sigma_{r,t}$ is the standard deviation of rate of change in reserves calculated until period t; $\frac{\sigma_{e,t}}{\sigma_{r,t}}$ is the rate of change in reserves.

To obtain a dependent variable, we need a specific threshold. As the value of the EMP varies significantly between countries, the limit must be individual, even though the rule is the same for all. Following Kinkyo (2020), our dummy variable returns a value of 1 when the value of the EMP index exceeds the value of 95 percentile, calculated by country to date:

$$Y_{i,t} = \begin{cases} 1 & \text{if there is any month in next 12 with a EMP value above the 95th percentile,} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Using these definitions, we show the incidence of pre-crisis periods in our sample of countries in Figure 1. We can see that there is a correlation between countries, that is, some periods are characterized by several simultaneous crises. Two simple examples of this are the subprime crisis in 2008 and the beginning of the coronavirus pandemic in 2020. It is also interesting to note that some periods are characterized by crises only when we use the 10% threshold, as in the period from 2010 to 2012. While in the late 1990s, the share of crises with depreciation above 20% was higher.

3.2 Models

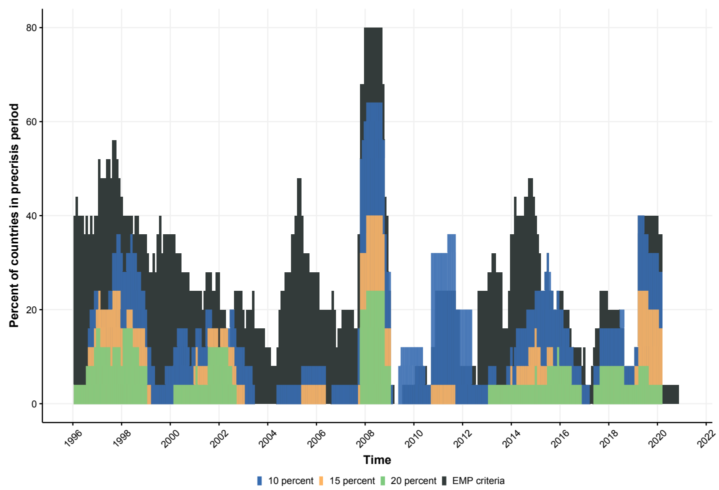
Our strategy for this forecasting model is to build an ensemble of different models. This type of approach is widely used in machine learning because it is proven to be quite effective in increasing the efficiency of the models and balancing their weaknesses (ZHANG; MA, 2012; AMPOMAH; QIN; NYAME, 2020). There are different ensemble methods, but we will use one of the simplest, soft voting. In this technique, we estimate the different models individually, which will return, for each month and country, a crisis probability value. Then, we average these multiple outputs to get a single probability value. There is also the possibility to include weights in each model, but here we used fixed weights.

To control the complexity of the model and to obtain a good generalization in out-of-sample settings, the hyperparameters for the models were obtained

thresholds, but the count and intensity of currency crisis have decreased since the nineties, so a 25% month depreciation has become rarer.

[3]There are different ways to build this index. For example, in Eichengreen, Rose, and Wyplosz (1995), there is the inclusion of the interest rate differential relative to the United States rate. However, we prefer to keep the simpler and more widespread version of the indicator.

Figure 1: Crises over time



through a grid-search algorithm - a standard approach in the early warning literature Holopainen and Sarlin (2017) and Beutel, List, and Von Schweinitz (2018). The optimized hyperparameters for the different models are shown in Table A.1.

We estimated four different models to get our Early Warning System. First, them logistic regression, which is the simplest and the most used in this type of forecast. So we can consider it our baseline estimation. The parameters are not country-specific, because we have unbalanced classes: there are many more "tranquil" periods than crisis periods. Estimating an individual model for each country could generate a loss of relevant information, and could lead to a problem of too many variables (features) for few observations. This approach is followed by the other models, as we estimate all the data together

The Random Forest is the second model (HO, 1998). This is an ensemble method that averages several decision trees using a bagging algorithm. Its primary goal is to reduce the issue of over-fitting, which is extremely prevalent when using deep decision trees. The random forest has recently attracted more interest in time series applications because it has shown to be quite successful in solving problems of this nature, such as in Lohrmann and Luukka (2019) and Kinkyo (2020).

Our third model is the XGBoost, which is also an ensemble method with decision trees. The principle is similar to that of the random forest but uses gradient boosting - a procedure that optimizes a loss function to aggregate the trees. This method has gained considerable relevance in recent years, as it is used to win several machine learning competitions. In terms of time series forecasting, it demonstrated excellent forecasting ability compared to other statistical and ML models (ABOLGHASEMI *et al.*, 2019), in addition to being used in the FFORMA model, it achieved second place in the M4 competition (MONTERO-MANSO *et al.*, 2020).

Finally, we will use a deep learning model, Long Short Term Memory (LSTM). This is a recurrent artificial neural network. This method makes it possible to process data sequentially. For this reason, it is used for time series problems, identifying relationships with lags between variables. Like the other two previous models, it was also shown that it outperforms traditional models in time series problems (SIAMI-NAMINI; NAMIN, 2018).

For the Random Forest, XGBoost, and LSTM models, we use the log loss criterion, also known as the cross-entropy loss, as a loss function to be minimized in the validation step. Most articles in the literature primarily use other metrics, such as AUC or accuracy. In contrast to these, which evaluate only the final classification of the model (in our case, crisis or non-crisis), the log loss penalizes the model based on the estimated probability. For example, an erroneous prediction that there is a 90% probability of a crisis has more penalties than a prediction that is also wrong, but which pointed out only a 60% probability of a crisis. Using the log loss is justified by the intrinsic difficulty of predicting a currency crisis. In this context, an increase in the probability of a crisis from 10% to 40% is relevant and informative regarding the heightened risk of a currency crisis, even though the crisis probability for the model is still less than 50% - below the threshold that would make the model indicate a currency crisis in the next 12 months.

3.3 Cross-Validation and Time Series

The standard approach to assess the generalizability of the results of machine learning applications is to use the K-fold cross-validation (CV) algorithm (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). However, time series data can present patterns - such as serial correlation and non-stationarity - that violate key hypotheses that are needed to guarantee the desirable statistical properties of the K-fold cross-validation method. Also, there is the basic intuition that we should not use data from the future to predict the past Bergmeir, Hyndman, and Koo (2018).

We are dealing with a set of macroeconomic and financial time series that in general are serially correlated and non-stationary. Therefore, we use a slightly different validation technique, called time series validation or forward validation, which respects the temporal structure of the data and has desirable statistical properties in the presence of non-stationary data Schnaubelt (2019).

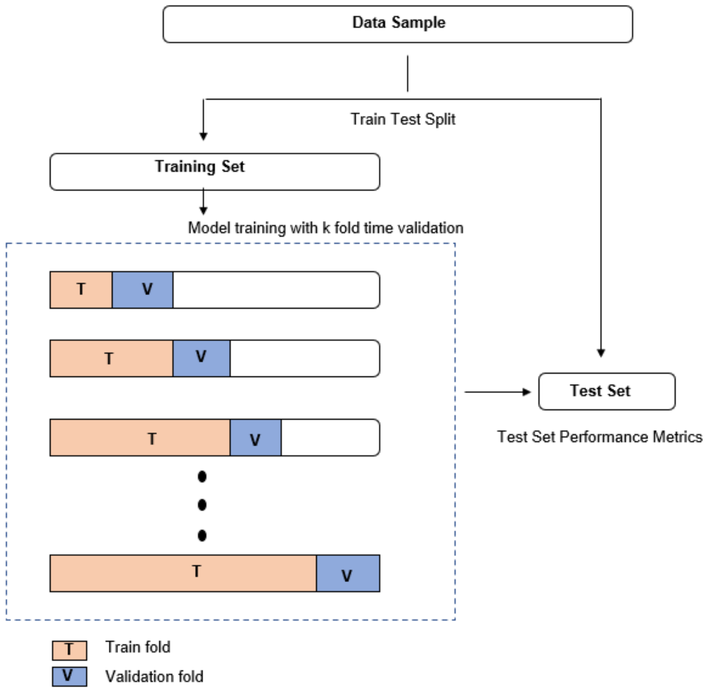
Instead of a random division, as done by K-fold, the data is split in an orderly fashion by time. Also, the data and validated parameters are only trained using previous data. We split the data using 5 sequential folds. The model metrics are then assessed in the test set. Figure 2 describes the validation and testing strategy.

3.4 Imputing missing data

The issue of missing observations is an important theme which is often overlooked in the Early Warning literature. As far as our knowledge goes, only Kinkyo (2020) details explicitly a method to correct this problem. Of course, most papers used lower time frequencies, which can mitigate the missing observations problem. Using fewer predictive variables or countries can also alleviate the problem, however, it is unlikely that the issue will be completely eliminated, and at limit, it makes the estimation of machine learning models impossible.

The incidence of missing observations in the sample is nearly 10%. This number varies considerably concerning each country and each variable. Some

Figure 2: Time validation method



series start after the first month (January 1995), in others there is no data for some countries. Attributing the missing observations is a challenge, especially in this case. Removing variables or countries with a lot of missing data can compromise the model by removing important information. The opposite case is equally damaging, since imputing a large chunk of data can cause a lot of noise to be included. We, thus, removed many countries from the sample as they did not have enough data. Following, Kinkyo (2020), we use the K—Nearest Neighbors method to impute the missing data. It is a multivariate impute method, which models each feature with missing values as a function of the other ones.

4 Results

4.1 Metrics evaluation

Forecasting a currency crisis is an imbalanced classification problem, that is, the distribution of the dependent variable between 0 (tranquil periods) and 1 (crisis periods) is far from balanced, which would be 50% of the observations for each. Naturally, the more unbalanced the problem is, the easier it is to achieve greater accuracy, even if the model is not necessarily better. For example, if we have a sample with only 1% of pre-crisis periods, a model that does not predict any of them may have 99% of accuracy by predicting only "tranquil periods". In summary, this is why in this case it is much more important to analyze the value of other metrics, such as the AUC. However, many studies, such as (SEVIM *et al.*, 2014; ALAMINOS *et al.*, 2019) use accuracy with-

Table 1: Metrics results

		10% drop	15% drop	20% drop	EMP index criteria
Accuracy	Logit	0.843	0.832	0.957	0.829
	Random Forest	0.786	0.932	0.957	0.720
	XGBoost	0.850	0.929	0.964	0.776
	LSTM	0.825	0.671	0.894	0.596
	Ensemble	0.839	0.879	0.952	0.596
AUC	Logit	0.537	0.546	0.500	0.500
	Random Forest	0.684	0.611	0.652	0.476
	XGBoost	0.629	0.627	0.677	0.501
	LSTM	0.497	0.528	0.619	0.499
	Ensemble	0.637	0.636	0.663	0.484
Log loss	Logit	0.387	0.241	0.178	0.484
	Random Forest	0.463	0.374	0.207	0.645
	XGBoost	0.347	0.255	0.140	0.481
	LSTM	1.144	0.495	0.248	2.681
	Ensemble	0.350	0.242	0.149	0.516

out any further discussion, which can make the reported results misleading in relation to the quality of the model.

It is necessary to define a probability value that separates the forecasts between tranquil periods and pre-crisis periods. As we are using log loss to validate the parameters, the default value of 50% may not be the most appropriate depending on the criteria you want to optimize. Similar to Candelon, Dumitrescu, and Hurlin (2012), we use an optimal cut-off value based on maximizing AUC on the validation set.

To evaluate the models, we only return the results of the out-of-sample metrics, as shown in Table 1. The log loss, which was used as a loss function in the validation set, has a rather simple interpretation. Nonetheless, for completeness, we also include the results of the AUC and the accuracy of all models. For each metric, the table shows the values for each of the models individually and the value of the ensemble by soft voting.

The estimated model using the dependent variable based on the EMP index has a much worse performance than models based on the simple nominal depreciation rate. This may be due to monetary policy, given that the central bank can avoid a strong devaluation using monetary reserves. Thus, this variable appears to have more noise than the variable with simple depreciation, which makes it less efficient to be used in the forecast.

An important point of these results is to show the effectiveness of using the model ensemble. This model returned the best AUC for the estimation with a 15% threshold, the second-best for 20% and 10%. It is worth mentioning that the adopted ensemble method was simple, in addition to being affected by the poor individual performance of Logit and LSTM. Some more robust ensemble methods, such as boosting, may bring even better results. Regarding the individual models, the consistent performance of the random forest algorithm is noteworthy, a not-so-complex model that provided results similar to the more complex XGBoost. On the other hand, the LSTM has failed in this regard. One possible explanation is the low number of observations in relation to what is necessary for deep learning models.

At first, the results of the performance metrics may seem discouraging. There are studies with accuracy close to 0.99 (ALAMINOS *et al.*, 2019), others with AUC above 0.95 (RAMLI; ISMAIL; WOOI, 2015), despite the differences in the dependent variable definition. However, we seek to use parameters that would allow our forecast to be performed with this metric in real policy-making and investment settings. If it were possible to predict a currency crisis with such precision, months before it occurred, using public databases and widely used methods, we would be facing a clear violation of the efficient markets hypothesis (MALKIEL; FAMA, 1970). Therefore, our forecast, with an AUC of 0.636 to 0.663, depending on the threshold used, can be considered adequate and realistic.

4.2 Some practical examples

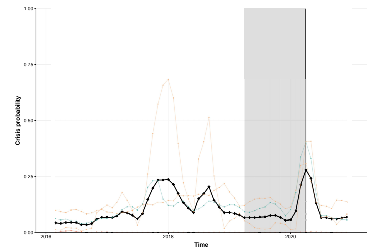
The values of the performance metrics are very important, but to evaluate an Early Warning System it is also interesting to understand how it behaves in showing us the probability of a crisis. For this, we selected some countries to monitor the performance of the model in the out-of-sample period. In all the graphs that are shown in this section, the black line is the probability estimated by the ensemble of the four models (logit, random forest, XGBoost and LSTM). The colored lines represent the individual probabilities estimated by these models. Furthermore, the vertical line represents that there was a depreciation above the specific threshold in that month and the shaded area represents the period 12 months before the crisis.

It is interesting to note that in April 2020 we had several major currency devaluations in developing countries caused by the worsening of the coronavirus pandemic. One of the countries most affected was Mexico, which has the probability shown in Figure 3. It is possible to see that the model detects an increase in instability in February and starts to rise. The same is true for Russia, which also suffered a sharp depreciation that month. However, in both cases, there is an increase in probability in previous periods that is not followed by a crisis. For the same period, the model for Brazil has an increase in probability, but one that is much less marked and evident than in the previous two cases. Also, there is a period with a high probability of depreciation, but without an event. The case of Argentina is more emblematic, given that it is a country that has been going through several exchange rate crises in the last 20 years. In this country, the model behaves very clearly and correctly, reflecting the political and economic conditions that the country has been through. In the last individual case, Egypt, unlike the other four, we chose to show the model with a 20% threshold. This case behaves in the ideal way of an Early Warning System, with very low probabilities in tranquil periods, but with a substantial increase in the probability right before a currency crisis.

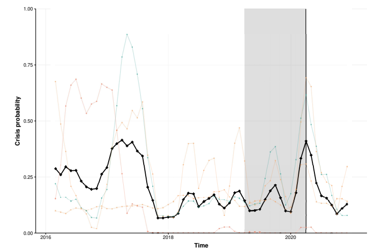
5 Conclusion

Our goal in this article was to build an Early Warning System for currency crises using machine learning models. Our contribution is to build a model that is applicable in policy making and investment settings and deals with recurrent problems in the literature such as: data with low frequency, that does not consider publication lag, misuse, and/or misreporting of validation techniques in time series data and inappropriate use of model evaluation metrics. Our model showed that although it is not possible to predict exchange rate

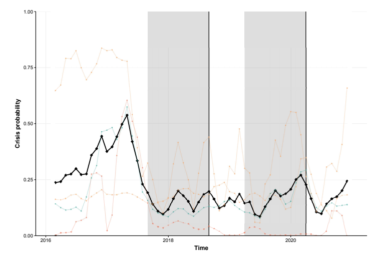
Figure 3: Probability of a crisis - Output of ensemble model



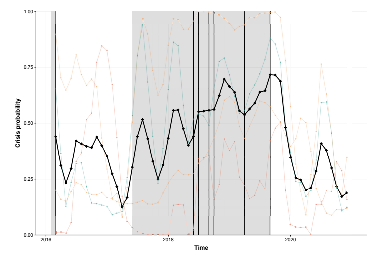
(a) Mexico - 10% depreciation threshold



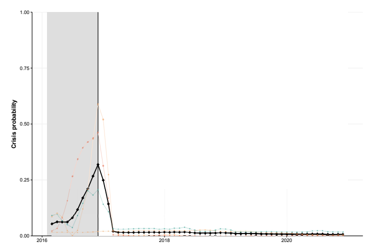
(b) Russia - 10% depreciation threshold



(c) Brazil - 10% depreciation threshold



(d) Argentina - 10% depreciation threshold



(e) Egypt - 20% depreciation threshold

crises with perfect accuracy, it is possible to build a model that returns useful probabilities for decision-making.

Still, there are limitations to our work. For example, due to the availability of data, we have chosen to work with a restricted sample of countries, almost all of which are middle-income. A wider sample - with the inclusion of low-income and frontier economies - would enrich the model predictions and robustness. Also due to data quality considerations, we started our dataset in 1995. A longer sample - reaching the 80s and 70s - would improve our understanding of how currency crises change as the global economy changes.

Finally, currency crises are notably very challenging to predict and the blind application of machine learning techniques and metrics to forecast them can provide misleading results and give potentially wrong signals to policymakers and other stakeholders. We do not consider that the objective of modeling a currency crisis should be to aim for unrealistic accuracy metrics rather it should be to understand the signals and thresholds that typically have preceded one. We argue that the usage of machine learning techniques - which are very powerful for prediction - should be coupled with an understanding of macroeconomic theory and policymaking objectives. While this approach, as we have shown in this paper, produces more sober results in terms of prediction accuracy it has the potential to provide more value in terms of guiding macro-prudential measures for policymakers or asset allocation strategies for investors.

References

ABOLGHASEMI, Mahdi *et al.* Machine learning applications in time series hierarchical forecasting. *arXiv preprint arXiv:1912.00370*, 2019.

ALAMINOS, David *et al.* Currency crises prediction using deep neural decision trees. *Applied Sciences*, v. 9, n. 23, p. 5227, 2019.

AMPOMAH, Ernest Kwame; QIN, Zhiguang; NYAME, Gabriel. Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. *Information*, v. 11, n. 6, p. 332, 2020.

BAŃBURA, Marta; RÜNSTLER, Gerhard. A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, v. 27, n. 2, p. 333–346, 2011.

BERGMEIR, Christoph; HYNDMAN, Rob J.; KOO, Bonsoo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, v. 120, p. 70–83, 2018.

BEUTEL, Johannes; LIST, Sophia; VON SCHWEINITZ, Gregor. An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions? *Deutsche Bundesbank Discussion Paper No. 48/2018*, 2018.

BOONMAN, Tjeerd M. *et al.* Early warning systems for currency crises with real-time data. *Open Economies Review*, v. 30, n. 4, p. 813–835, 2019.

BUSSIERE, Matthieu; FRATZSCHER, Marcel. Towards a new early warning system of financial crises. *Journal of International Money and Finance*, v. 25, n. 6, p. 953–973, 2006.

CANDELON, Bertrand; DUMITRESCU, Elena-Ivona; HURLIN, Christophe. Currency crisis early warning systems: Why they should be dynamic. *International Journal of Forecasting*, v. 30, n. 4, p. 1016–1029, 2014.

CANDELON, Bertrand; DUMITRESCU, Elena-Ivona; HURLIN, Christophe. How to evaluate an early-warning system: Toward a unified statistical framework for assessing financial crises forecasting methods. *IMF Economic Review*, v. 60, n. 1, p. 75–113, 2012.

CERQUEIRA, Vitor; TORGO, Luis; MOZETIČ, Igor. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, v. 109, n. 11, p. 1997–2028, 2020.

CHONG, Terence T. L.; YAN, Isabel K. Forecasting currency crises with threshold models. *International Economics*, v. 156, p. 156–174, 2018.

CUARESMA, Jesús Crespo *et al.* Determinants of currency crises: A conflict of generations? *Focus on European Economic Integration*, n. 1, p. 126–141, 2008.

DU, Jiangze; YU, Runfang; LAI, Kin Keung. Identification and prediction of currency crisis: Markov Switching-Based approach. *The Singapore Economic Review*, v. 65, n. 06, p. 1667–1698, 2020.

EICHENGREEN, Barry; ROSE, Andrew K.; WYPLOSZ, Charles. Exchange market mayhem: the antecedents and aftermath of speculative attacks. *Economic Policy*, v. 10, n. 21, p. 249–312, 1995.

EMIN, Dogus; AYTAÇ, Ayşegül. Predicting currency crises: how do indicators differ according to crisis definition? *Çankırı Karatekin University Journal of The Faculty of Economics and Administrative Sciences Y*, v. 6, n. 2, p. 85–102, 2016.

FLOOD, Robert P.; GARBER, Peter M. Collapsing exchange-rate regimes: Some linear examples. *Journal of International Economics*, v. 17, n. 1-2, p. 1–13, 1984.

GÓMEZ, Víctor; MARAVALL, Agustín. Estimation, prediction, and interpolation for nonstationary series with the Kalman filter. *Journal of the American Statistical Association*, v. 89, n. 426, p. 611–624, 1994.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The elements of statistical learning*. New York: Springer, 2009.

HO, Tin Kam. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, v. 20, n. 8, p. 832–844, 1998.

HOLOPAINEN, Markus; SARLIN, Peter. Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance*, v. 17, n. 12, p. 1933–1963, 2017.

KAMINSKY, Graciela; LIZONDO, Saul; REINHART, Carmen M. Leading indicators of currency crises. *IMF Staff Papers*, v. 45, n. 1, p. 1–48, 1998.

KINKYO, Takuji. A bi-annual forecasting model of currency crises. *Applied Economics Letters*, v. 27, n. 4, p. 255–261, 2020.

KRUGMAN, Paul. A model of balance-of-payments crises. *Journal of Money, Credit and Banking*, v. 11, n. 3, p. 311–325, 1979.

KRUGMAN, Paul. Crises: The Next Generation? *In: Economic Policy in the International Economy: Essays in Honor of Assaf Razin*. Organized by: E. Helpman and E. Sadka. Cambridge: Cambridge University Press, 2003. p. 15–32.

LOHRMANN, Christoph; LUUKKA, Pasi. Classification of intraday S&P500 returns with a Random Forest. *International Journal of Forecasting*, v. 35, n. 1, p. 390–407, 2019.

MALKIEL, Burton G.; FAMA, Eugene F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, v. 25, n. 2, p. 383–417, 1970.

MASSON, Paul R. *Contagion: monsoonal effects, spillovers, and jumps between multiple equilibria*. Washington, DC: International Monetary Fund, 1998.

MONTERO-MANSO, Pablo *et al.* FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, v. 36, n. 1, p. 86–92, 2020.

NAKATANI, Ryota. *Output costs of currency crises: shocks, policies and cycles*. Munich: Munich Personal RePEc Archive, 2018. (MPRA Paper No. 83549).

OBSTFELD, Maurice. Models of currency crises with self-fulfilling features. *European Economic Review*, v. 40, n. 3-5, p. 1037–1047, 1996.

RAMLI, Nor Azuana; ISMAIL, Mohd Tahir; WOOI, Hooy Chee. Measuring the accuracy of currency crisis prediction with combined classifiers in designing early warning system. *Machine Learning*, v. 101, n. 1, p. 85–103, 2015.

SCHNAUBELT, Matthias. *A comparison of machine learning model validation schemes for non-stationary time series data*. Nürnberg: Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. (FAU Discussion Papers in Economics No. 11/2019).

SEVIM, Cuneyt *et al.* Developing an early warning system to predict currency crises. *European Journal of Operational Research*, v. 237, n. 3, p. 1095–1104, 2014.

SIAMI-NAMINI, Sima; NAMIN, Akbar Siami. Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv preprint arXiv:1803.06386*, 2018.

ZHANG, Cha; MA, Yunqian. *Ensemble machine learning: methods and applications*. New York: Springer, 2012.

Appendix A

Table A.1: Hyperparameter Tuning via Grid-Search Algorithm

		10% drop	15% drop	20% drop
XGBOOST	Max Depth	3	2	3
	Number of Estimators	100	50	50
	Learning Rate	0.01	0.1	0.05
	Minimum Child Weight	10	5	7
Random Forest	Max Depth	20	50	60
	Min Samples Leaf	2	2	2
	Min Samples Split	5	2	5
	Number of Estimators	600	1000	200
LSTM	Hidden Layer	3	3	2
	Neurons per layer	64	32	8
	Dropout rate	0.1	0.01	0.1
	Learning Rate	0.001	0.001	0.001
	Batch Size	5	10	5
	Epoch Size	100	300	300

Table A.2: Metrics results

Data (Source or Distributor)	Transformations applied
Nominal Exchange Rate - Local Currency relative to USD (Fixer.io)	Month-over-month percentage change
Real Effective Exchange Rate - Based on consumer price index (CPI) (World Bank)	Hodrick–Prescott filter
Total Reserves excluding Gold, Foreign Exchange, US Dollars (World Bank)	Ratio in relation to GDP and monthly percentage change
Value of Exports, Free on board (FOB), US Dollars (World Bank)	year-over-year percentage change
Value of Imports, Free on board (FOB), US Dollars (World Bank)	year-over-year percentage change
Claims on private sector (IMF)	year-over-year percentage change
Money Market Rate (IMF)	year-over-year percentage change
Lending Rate (IMF)	year-over-year percentage change
Net Foreign Assets (IMF)	Ratio in relation to total reserves and Monetary Base, then year-over-year percentage change
Broad Money Liabilities (IMF)	Ratio in relation to total reserves and Monetary Base, then year-over-year percentage change
Consumer Prices (World Bank)	year-over-year percentage change
Monetary Base (IMF)	Used in other variable calculations
Balance of payments on goods & services (IMF)	year-over-year percentage change
Balance of Payments (current account) (IMF)	Ratio in relation to GDP and year-over-year percentage change
Commodity net export price index (IMF)	year-over-year percentage change
Coupcast (probability of political coup) (OEF)	No transformation applied
Election Month (Binary) (OEF)	Transformed to indicate a future election in the next 6 months (1 if positive, 0 if negative)
General government gross debt (IMF)	year-over-year percentage change
Stock market index (US Dollars) (World Bank)	year-over-year percentage change
Import Coverage (World Bank)	Level
Gross Domestic Product (World Bank)	year-over-year percentage change
Unemployment (World Bank)	year-over-year percentage change
Industrial Production Index (World Bank)	year-over-year percentage change
General government net lending/borrowing (IMF)	year-over-year percentage change
West Texas Intermediate 40 API Midland Texas; US$ per barrel (IMF)	year-over-year percentage change
NASDAQ Emerging Markets Index (NQEM) (NASDAQ)	No transformation applied
Gold Spot Prices - USD - Daily (Perth Mint)	No transformation applied
MSCI Emerging Markets Index Futures (Wiki Continuous Futures)	No transformation applied
US Dollar Index Futures (Wiki Continuous Futures)	No transformation applied
VIX (Yahoo Finance)	year-over-year percentage change