Predicting Financial Crises

Mouktar ABDILLAHI

1 Introduction

Before 2008, certain economists predicted the global financial crisis and the US housing bubble and today. However, most economists did not see it coming and they have seen theories they built for years get destroyed in few months. Hence, today, forecasters might be anxious to not be caught out the same way next time and some of them think that the next one is due by 2020. Even though this can be an overreaction as they might be over-cooking the warnings, we can take from this that there is a serious concern about predicting the next financial crisis. Having an accurate estimate of the probability of a financial crisis is very valuable today and this is problem tackled in this paper. [2] Following the work done by Schularick and Taylor in 2012 and [1]the work done by Daniel Fricke, I will be trying to replicate the work done in predicting financial crisis and extend it as more Machine Learning models will be applied to the data and other forms of data processing is applied.

Predicting financial crises is predicting the future in some way, and when judging the results of our Machine Learning models, the focus will be on the out-of-sample performance and you will see that these performances are highly variable depending on the ML method used and the data on which the model is applied. Data availability is also a great concern and even though the dataset on which this study is built goes back to 1870, only yearly values are reported and the low amount of data can be a concern here.

This paper is organized as follows: the first part will be explaining the methodology used in this study and the dataset on which it is applied while the second part will present the results of this study and the last part will be coming back on these results and giving insights on how valuable these results actually are.

2 Methodology and Data

The dataset comes from the Schularick and Taylor and is covering 14 developed countries such as the United States or France and goes from 1870 to 2008.

2.1 Two studies in one

Both Schularick-Taylor and Fricke's works were focusing on credit growth and as explained in the Schularick and Taylor, “financial crises are credit booms gone wrong”. The importance of credit growth is clearly explained in that paper and to replicate that study, we focused on predicting the probability of a crisis using 5 lags of credit growth. This study will be called **Study A**.

Besides this study, using all the other features available in the dataset was very tempting so we have decided to also use it. The second study was using all the features available in the dataset besides 'govass' which had more than 60% missing data and we have added the credit growth as another feature because we were convinced of its importance thanks to Schularick-Taylor. This study will be called **Study B**.

2.2 Challenge 1: Missing Data

As mentionned in the introduction, the dataset has a quite low amount of observations when it comes to using Machine Learning on it. In order to not face under-fitting, we had to keep as much observations as possible. The decision has been made to try two different ways when dealing with the missing data.

For the sake of replicability, we use the same data filters as the two previous studies and that is excluding world wars and their aftermaths (1914-1926 and 1939-1948) - this led to eliminating most of the missing data and for the few remaining, we also excluded those observations. It left us with only 1007 observations for the Study B and with 1587 for the Study A. It also creates one other problem as the countries are not on the same level when it comes to missing data and wars, so at the end we are left with real differences in the amount of observations by country for each study (you will be able to see it in Table 1). For the rest of the study, this will be called **Processing 1**.

	USA	SWE	NOR	NLD	JPN	ITA	GBP	FRA	ESP	DNK	DEU	CHE	CAN	AUS
Obs. A	100	132	127	103	102	129	117	107	97	113	107	97	133	123
Obs. B	87	114	50	75	51	81	97	47	58	50	92	50	57	87

Table 1: Differences in amount of observations

In order to extend the study and because we wanted to keep as much observations as possible, we decided to fill the missing data with the previous not NaN value and we did that for all features used depending on the fact that we were in Study A or Study B. For the rest of the study this will be called **Processing 2**.

2.3 Challenge 2: Imbalanced data

The next challenge faced was the fact that the dataset was really imbalanced as we only had 79 financial crises for 1946 observations, so that represented only 4%. The problem with an imbalanced dataset is that we can get an accuracy of 90% and maybe more because the classifier used classifies every observation as 'No Financial Crisis' and even though the accuracy score might be fantastic, the classifier is of no utility for us. When addressing imbalanced data, a first advice is to compare AUC Scores and not accuracy score, it is what we did here but this will not be enough.

[3] In order to solve this problem, we have used the SMOTE algorithm (Synthetic Minority Oversampling Technique). SMOTE uses k-nearest neighbors to create synthetic examples of the minority class and the results of that for each ML model will be between brackets in Table 2, 3, 4 and 5.

2.4 ML models

This study has been done on Python using Machine Learning models available in the scikit-learn package. The list of the ML models used in this paper goes as follows: Logistic Regression, Random Forest, Extra-Trees, Decision Tree, AdaBoost, Gradient Boosting, XGBoost, K-Nearest-Neighbors, Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, simple Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD).

In order to produce the out-of-sample performance, the typical method is by K-fold cross-validation. However, in order to be realistic, the past has to be used to predict the future so shuffling the data is not the way the cross-validation has to be done. We did what is called Time Series Cross-Validation and the figure 2 illustrates the out-of-sample validation approach.

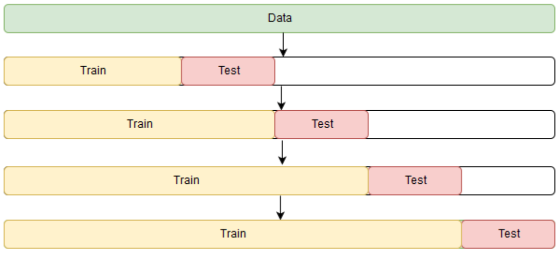


Figure 1: Time Series Cross Validation Method for K = 4

3 Results

In the next four following table, we will be reporting the results of the different Machine Learning models and the different approaches tested. A row called "Ratio - Crisis/No Crisis" is in every table and tells us how imbalanced is the data for each K and this is because the financial crises are not smoothly spread in the timescale from 1870 to 2008. When this ratio is way lower than the 4% ratio of the entire dataset, we have noticed that the AUC Score is very good generally and this is because classifying every observations as "No Financial Crisis" will not be reall costly in term of AUC Score. So, the results for that case are just ignored.

Model	**K = 1**	**K = 2**	**K = 3**
Ratio - Crisis/No Crisis	5.32%	0.51%	4.76%
Logistic Regression	0.736 (0.694)	0.655 (0.685)	0.607 (0.650)
Random Forest	0.657 (0.688)	0.752 (0.699)	0.634 (0.670)
Extra-Trees	0.632 (0.681)	0.674 (0.716)	0.604 (0.618)
Decision Tree	0.536 (0.560)	0.428 (0.494)	0.5 (0.520)
AdaBoost	0.637 (0.551)	0.944 (0.858)	0.660 (0.627)
Gradient Boosting	0.662 (0.676)	0.883 (0.835)	0.616 (0.638)
XGBoost	0.658 (0.645)	0.910 (0.810)	0.590 (0.633)
KNN	0.548 (0.545)	0.495 (0.495)	0.485 (0.532)
Naive Bayes	0.654 (0.654)	0.406 (0.390)	0.381 (0.361)
Linear Discriminant Analysis	0.696 (0.685)	0.321 (0.386)	0.461 (0.468))
Quadratic Discriminant Analysis	0.682 (0.681)	0.557 (0.514)	0.465 (0.439)
SVM	0.293 (0.779)	0.336 (0.444)	0.487 (0.494)
SGD	0.687 (0.633)	0.387 (0.635)	0.636 (0.736)

Table 2: Out-of-Sample performance (AUC) - Study A X Processing 1

Model	**K = 1**	**K = 2**	**K = 3**
Ratio - Crisis/No Crisis	2.87%	3.29%	5.02%
Logistic Regression	0.775 (0.850)	0.617 (0.532)	0.499 (0.524)
Random Forest	0.541 (0.594)	0.678 (0.532)	0.511 (0.583)
Extra-Trees	0.292 (0.428)	0.704 (0.582)	0.532 (0.578)
Decision Tree	0.429 (0.379)	0.496 (0.5)	0.492 (0.492)
AdaBoost	0.682 (0.842)	0.548 (0.468)	0.593 (0.651)
Gradient Boosting	0.254 (0.428)	0.679 (0.459)	0.556 (0.539)
XGBoost	0.690 (0.624)	0.622 (0.472)	0.548 (0.573)
KNN	0.567 (0.582)	0.5 (0.5)	0.487 (0.525)
Naive Bayes	0.558 (0.633)	0.484 (0.444)	0.467 (0.460)
Linear Discriminant Analysis	0.524 (0.521)	0.625 (0.631)	0.575 (0.592)
Quadratic Discriminant Analysis	0.532 (0.503)	0.598 (0.607)	0.566 (0.557)
SVM	0.464 (0.517)	0.707 (0.739)	0.636 (0.775)
SGD	0.558 (0.558)	0.623 (0.570)	0.611 (0.550)

Table 3: Out-of-Sample performance (AUC) - Study B X Processing 1

Model	**K = 1**	**K = 2**	**K = 3**	**K = 4**
Ratio - Crisis/No Crisis	5.71%	3.05%	0.54%	5.08%
Logistic Regression	0.497 (0.456)	0.801 (0.797)	0.564 (0.615)	0.637 (0.668)
Random Forest	0.541 (0.483)	0.627 (0.660)	0.770 (0.728)	0.561 (0.654)
Extra-Trees	0.537 (0.484)	0.539 (0.628)	0.726 (0.646)	0.620 (0.579)
Decision Tree	0.496 (0.463)	0.485 (0.541)	0.499 (0.5)	0.5 (0.493)
AdaBoost	0.432 (0.452)	0.546 (0.665)	0.809 (0.922)	0.674 (0.618)
Gradient Boosting	0.544 (0.444)	0.606 (0.753)	0.803 (0.835)	0.624 (0.642)
XGBoost	0.540 (0.460)	0.532 (0.762)	0.882 (0.878)	0.591 (0.663)
KNN	0.522 (0.531)	0.595 (0.680)	0.498 (0.499)	0.489 (0.484)
Naive Bayes	0.459 (0.502)	0.743 (0.728)	0.318 (0.343)	0.377 (0.364)
Linear Discriminant Analysis	0.455 (0.433)	0.606 (0.621)	0.314 (0.320)	0.489 (0.492)
Quadratic Discriminant Analysis	0.465 (0.479)	0.776 (0.739)	0.530 (0.492)	0.557 (0.587)
SVM	0.514 (0.538)	0.160 (0.867)	0.527 (0.411)	0.497 (0.500)
SGD	0.425 (0.346)	0.794 (0.604)	0.286 (0.497)	0.689 (0.637)

Table 4: Out-of-Sample performance (AUC) - Study A X Processing 2

For each table, the five highest AUC scores are highlighted (ignoring the columns with very low "Ratio - Crisis/No Crisis"). The scores using the SMOTE method are highlighted in green and the scores without over-sampling in green. You will notice that generally, the best working ML model for the SMOTE is not the best one without the SMOTE and vice-versa.

Model Ratio - Crisis/No Crisis	K = 1 6.92%	K = 2 3.21%	K = 3 0.52%	K = 4 4.89%
Logistic Regression	0.483 (0.523)	0.700 (0.665)	0.781 (0.733)	0.494 (0.491)
Random Forest	0.519 (0.518)	0.684 (0.618)	0.697 (0.597)	0.640 (0.533)
Extra-Trees	0.533 (0.507)	0.579 (0.494)	0.764 (0.5)	0.614 (0.527)
Decision Tree	0.488 (0.506)	0.479 (0.472)	0.733 (0.409)	0.496 (0.493)
AdaBoost	0.622 (0.634)	0.613 (0.626)	0.807 (0.501)	0.546 (0.507)
Gradient Boosting	0.547 (0.485)	0.603 (0.686)	0.882 (0.464)	0.541 (0.535)
XGBoost	0.460 (0.477)	0.609 (0.709)	0.751 (0.555)	0.656 (0.505)
KNN	0.459 (0.449)	0.577 (0.551)	0.701 (0.683)	0.513 (0.512)
Naive Bayes	0.534 (0.577)	0.678 (0.670)	0.318 (0.291)	0.497 (0.462)
Linear Discriminant Analysis	0.586 (0.595)	0.508 (0.553)	0.305 (0.203)	0.468 (0.496)
Quadratic Discriminant Analysis	0.374 (0.388)	0.564 (0.571)	0.359 (0.368)	0.510 (0.538)
SVM	0.455 (0.574)	0.281 (0.736)	0.956 (0.103)	0.480 (0.482)
SGD	0.550 (0.450)	0.525 (0.625)	0.203 (0.232)	0.492 (0.499)

Table 5: Out-of-Sample performance (AUC) - Study B X Processing 2

4 Comments on the results

4.1 Plots

To get a better grasp of how valuable these results can be, let's focus on the highest score of each table. First we can produce the confusion matrices for these scores to see how people were classified (we used a threshold of 0.5 on the probability of financial crisis, for the sake of simplicity).

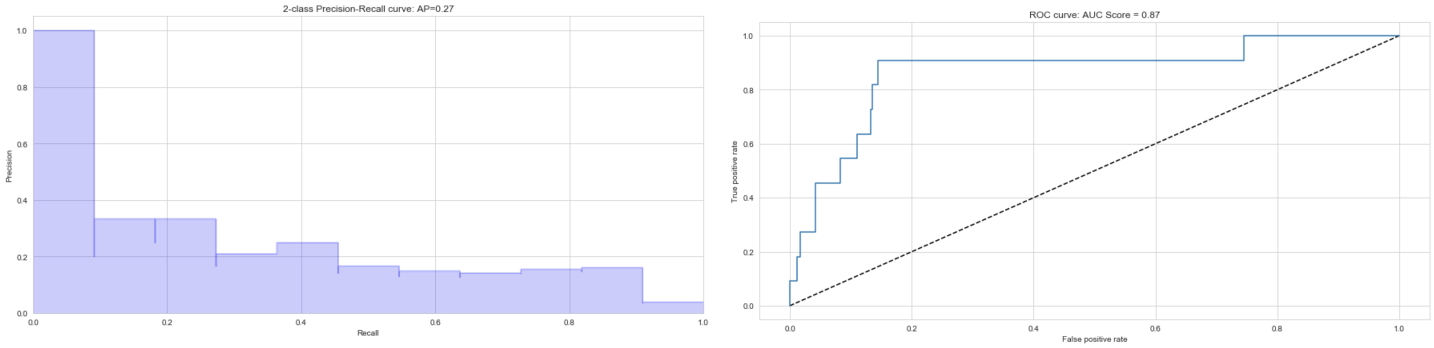
		Model Prediction	
		$\overline{\textbf{Fin. Crisis}}$	**Fin. Crisis**
Actual - Study A X Processing 1	$\overline{\textbf{Fin. Crisis}}$	366	11
	Fin. Crisis	15	5
Actual - Study A X Processing 1	$\overline{\textbf{Fin. Crisis}}$	233	11
	Fin. Crisis	4	3
Actual - Study A X Processing 1	$\overline{\textbf{Fin. Crisis}}$	359	2
	Fin. Crisis	10	1
Actual - Study A X Processing 1	$\overline{\textbf{Fin. Crisis}}$	355	19
	Fin. Crisis	8	4

Table 6: Confusion Matrices

As you can see, even though an AUC Score of 0.8 can look very good, the classifier still makes many mistakes when using 0.5 as a threshold. Maybe a better choice of threshold could improve the results we have seen in these confusion matrices.

Besides this, the ROC-AUC-Curve is a well-known performance measurement method for classification problem at various thresholds settings. In order to visualize the results of this study, out of the 20 highlighted scores, we will plot the ROC-AUC-Curve for the 2 highest AUC Scores.

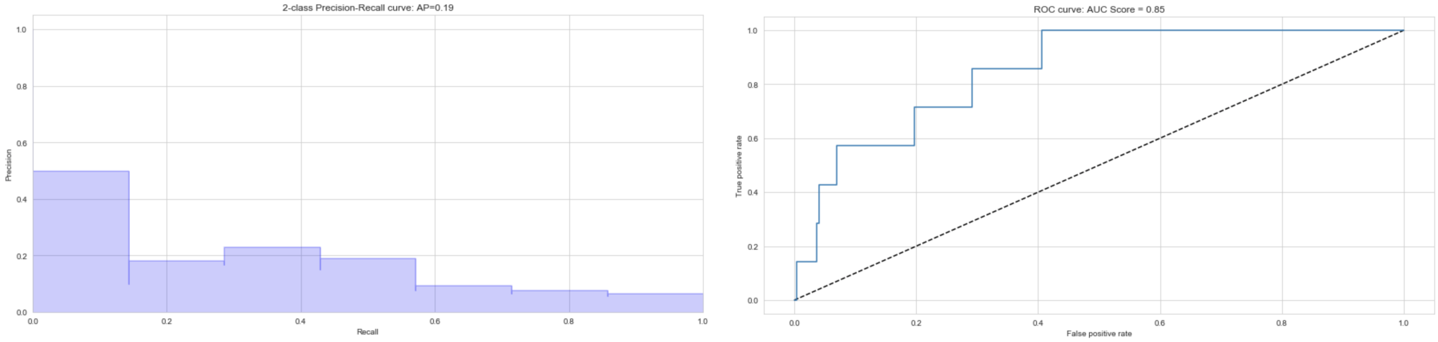
For the imbalanced classes problem, there are metrics that have been designed to tell you a more truthful story when working with imbalanced classes. We have, for example, the Precision (a measure of a classifier's exactness) and the Recall (a measure of a classifier's completeness). In order to visualize the results, we will plot the precision-recall curve for the same examples as the ROC-AUC-Curves. All these plots are in the following figures:



(a) Precision-Recall Curve

(b) ROC AUC Curve

Figure 2: Plots for SVM on Study A X Processing 2 (using SMOTE and K = 2)



(a) Precision-Recall Curve

(b) ROC AUC Curve

Figure 3: Plots for Logistic Regression on Study B X Processing 1 (using SMOTE and K = 1)

You can notice the average precision-recall scores are very low, even for the best methods and classifiers (0.27 and 0.19) so these Machine Learning methods can certainly spot few financial crises but are not very reliable.

4.2 Observations

1. We have AUC Scores that can be very high, reaching 0.867 but when a model is working well for a certain K, generally it does not work very well for the other Ks. Indeed, we do not have one row having high-scores for every K and this can be explained because the cross-validation is not stratified. That means that for each K, we have the same ratio Crisis/No Crisis, and as this cannot be the case in this study, it might explain that. However, other reasons could also exist.

2. When we combine the 20 highest AUCs, 12 of them come from the SMOTE method and the rest is computed without oversampling. We can say that the SMOTE method is valuable to a certain point even though its efficiency on this dataset is not overwhelming.

5 Conclusion

In this study, we have shown that many paths can be followed when trying to predict financial crises. ML methods can be useful for different prediction problems, but they may not be really reliable to predict accurately financial crises. If we had to select one ML algorithm it would be SVM using SMOTE to oversample and to fix a little bit the imbalanced data issue. We have made this choice because in three out of the four tables, the highest AUC Score is obtained doing so. Even though the results have many weaknesses, it is a good starting point and financial crises forecaster should certainly focus on these ML methods.

References

[1] Fricke, D. *Financial crisis prediction: a model comparison*, (2017).

[2] Schularick, M. and A. Taylor. *Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870-2008*, (American Economic Review 102, no. 2 (2012)).

[3] Anna Vasilyeva, *Using SMOTEBoost and RUSBoost to deal with class imbalance*, available on Medium at `https://medium.com/urbint-engineering/using-smoteboost-and-rusboost-to-deal-with-class-imbalance-c18f8bf5b805`