Building an Early Warning Model for Fiscal Stress

comparing Logistic Regression with Random Forest

replication paper

Bela Tim Koch
17-734-377

Machine Learning in Economics (458657)
Spring Semester 2022

Department of Economics
University of Bern

1 Introduction and Literature Review

Since the latest Great Recession with its corresponding deterioration of public finances, the monitoring and prevention of fiscal crises has become increasingly prominent in the political debate, leading to a rising demand for the development of reliable and early indicators that signal possible fiscal stress. In order to be able to assess countries' vulnerability to fiscal distress ex-ante, the literature is increasingly devoted to the development of early warning systems for fiscal stress, which builds upon early warning systems for banking and currency crises (Honda, Tapsoba, and Issifou 2022). The standard tool used in the literature for early warning systems are the signaling approach as well as discrete dependent variable models, such as logistic regression (Jarmulska 2020).

As an alternative to the traditional methods, early warning models based on machine learning techniques are proposed, claiming a possible improvement of prediction accuracy (Beutel, List, and Schweinitz 2019). For example, when predicting the build-up of banking crises, the early warning system developed by Casabianca et al. (2019), which builds upon a supervised machine learning algorithm (i.e. Adaptive Boosting) outperforms the traditional approach of using a logit model. Another example which shows that using machine learning could drastically improve prediction accuracy is the early warning system developed by Samitas, Kampouris, and Kenourgios (2020), which reaches an accuracy of 98.8% when predicting the risk of contagion inside a financial network using a quadratic support vector machine.

However, a disadvantage of many of these machine learning methods compared to more traditional approaches is the difficulty of understanding how the results were obtained, which is why they are often referred to as a black box (Ghoddusi, Creamer, and Rafizadeh 2019). Consequently, the researcher is confronted with a trade-off between prediction and interpretation. Ghoddusi, Creamer, and Rafizadeh (2019) argue, that emphasis should be placed on interpretation for scientific research or policy decisions, since understanding the relationship and behavior among different variables is more important than prediction accuracy. In contrast, more emphasis is to be placed on prediction accuracy in specific industrial applications.

This paper aims to replicate some of the work done by Jarmulska (2020). In particular, attention will be paid on the comparison of the traditional method, i.e. a logit model with a least absolute shrinkage and selection operator, and a model based on machine learning, i.e. an implementation of the random forest algorithm. As the results obtained by a machine learning model are often criticized for being difficult to interpret, ways of interpreting the results obtained by the early warning model based on random forest are presented.

2 Model Description

2.1 Performance Metrics

Jarmulska (2020) uses sensitivity, specificity, their average as well as the area under receiver operating curve (AUROC) as measures to assess the effectiveness of the early warning models. Since sensitivity corresponds to the proportion of stress episodes correctly classified whereby specificity corresponds to the proportion of tranquil episodes correctly classified, these metrics are dependent on the threshold which determines whether a period is classified as a stress or tranquil episode (Jarmulska 2020). In this paper, this threshold is specified by maximizing the weighted sum of sensitivity and specificity. In contrast, the AUROC is a robust measure, since all possible thresholds are considered in the calculation of the AUROC. This measure represents the area under the receiver operating curve (ROC), which displays the trade-off between the true positive rate (i.e. sensitivity) and the false positive rate (i.e. 1 - specificity). Theoretically, the AUROC can be between 0 (worst possible classifier) and 1 (perfect classifier), whereby random guessing would result in to a value of 0.5 (Fawcett 2006).

2.2 Logit Model with LASSO Penalization

Jarmulska (2020) implemented two versions of discrete dependent variable model (logit regression), first a standard logit model with ordinary least squares estimates and second a logit model with a least absolute shrinkage and selection operator (LASSO) penalization. These models are often used as the standard econometric approach, which is why they are used as the benchmark in this study when assessing the usefulness of the random forest model.

Ordinary least squares estimates often have low bias but large variance, reducing prediction accuracy. Sometimes prediction accuracy can be improved by shrinking some coefficients towards zero to sacrifice bias in order to reduce variance of the predicted values (Tibshirani 1996). To do so, LASSO penalization as proposed in Tibshirani (1996) can be applied. Because this study focuses on the comparison of prediction accuracy of the traditional approach of using logit regression versus a random forest model, only the logit LASSO model is considered in this replication.

Following Hastie et al. (2009), the LASSO problem in the Lagrangian form is given as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{1}$$

whereby λ corresponds to the penalization parameter. As can be seen in Equation (1), the higher λ, the stronger a high coefficient is penalized and therefore the higher the number of coefficients shrunk to zero. Here, λ is chosen by 5-fold cross-validation, maximizing the AUROC.

2.3 Random Forest

As an alternative to the logit model, Jarmulska (2020) applied the random forest algorithm following Breiman (2001) as an ensemble of multiple classification trees for binary classification. A single tree divides the predictor space into distinct and non-overlapping regions, whereby an observation is classified depending on the region it falls into, i.e. at the terminal node. Each non-terminal node corresponds to a question with a binary response, what determines the structure of the tree. Since each terminal node assigns the same class to all observations within this node, minimizing the classification error at the level of the terminal nodes results in a minimization of the tree's overall classification error (Jarmulska 2020). To measure the precision of the fit, the Gini Index, which is used as a loss function in classification and regression trees, can be used (Jarmulska 2020):

$$g(w) = \sum_{k \neq j} p_{wk} p_{wj} = \sum_{k} p_{wk}(1 - p_{wk}) \tag{2}$$

whereby p_{wj} corresponds to the probability distribution of class j in node w.

However, such decision trees might suffer from overfitting, which can be counteracted by bootstrapping the training set and averaging all the predictions (James et al. 2013). If these so-called bagged trees are all influenced by some strong predictors, the single trees might be correlated resulting again in overfitting. To decorrelate the single trees, the non-terminal nodes can be forced to consider only a subset of the predictors, increasing the difference between the single trees, through which the problem of overfitting can be reduced (James et al. 2013). This ensemble method corresponds to the random forest algorithm. In this application, 10'000 trees are grown to construct the random forest. The number of variables considered at each non-terminal node is determined by the square-root of the number of variables considered by the random forest algorithm, which in this application corresponds to four.

3 Data Description

3.1 Dependent Variable

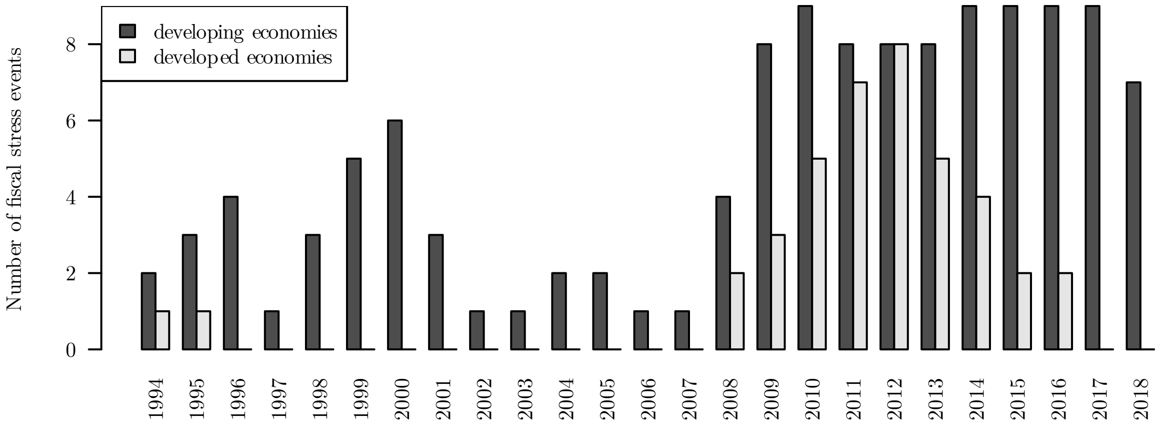
The binary dependent variable to be predicted follows the definition in Dobrescu et al. (2011) and takes the value of 1 in the case of a fiscal stress event in the next period and 0 otherwise. According to the definition provided by Dobrescu et al. (2011), an economy faces a fiscal stress event, if at least one of the following four conditions are fulfilled: (1) the economy fails to service debt as payments come due as well as if debt exchanges are distressed (2) the economy receives a large support program by the International Monetary Fund (3) hyperinflation is prevalent in the economy (inflation exceeding 35% for advanced economies, 500% for emerging economies) (4) the economy faces extreme financing constraints, i.e. the sovereign spread exceeds 1000 basis points or 2 standard deviations from the country average.

For the analysis, Jarmulska (2020) considers 43 economies (19 developing economies, 24 developed economies) for an observation period for years 1994-2018. 16.5% of the recorded observation are classified as fiscal stress events, whereby these stress events are not equally distributed across country groups and over time. Figure 1 displays the distribution of the recorded stress events over time for developing and developed economies. The data shows that developing economies are more prone to fiscal stress events with 29% of all observations classified as fiscal stress events compared to 7.1% for developed economies.

3.2 Explanatory Variables

Annual frequency data lagged 2 years in regard to the dependent variable, hence for an observation period for years 1992-2016, are used to train and test the early warning models. Jarmulska (2020) chose a lag of two years to simulate the reality of how the early warning models could be used in practice, as the data becomes available with a delay of up to one and a half years, leaving the decision-makers half a year time to react on the results of the models.

Figure 1: Distribution of Stress Periods



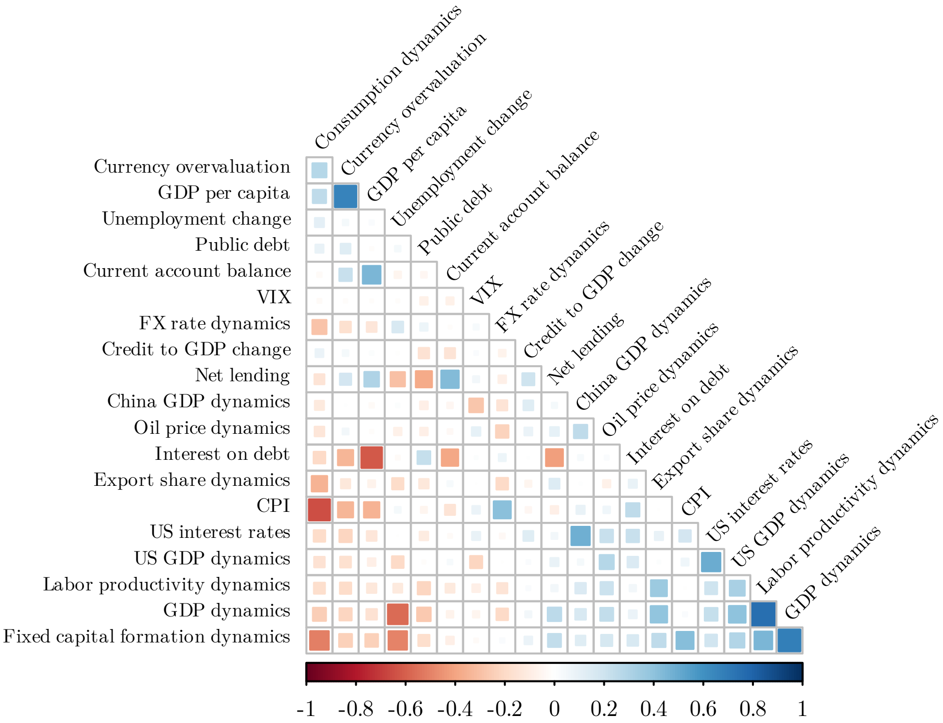
All variables used for building the early warning model are listed in Table 1 including their means with the distinction whether the economy is in a fiscal stress period or not. In addition, using a Wilcoxon test with statistical significance at 0.05%, Table 1 shows if the means in tranquil periods and in stress periods are statistically different and therefore indicating that the observed variables behave differently in stress periods, hence giving the variables potentially explanatory power.

Table 1: Means of Explanatory Variables

Variable	All periods	Tranquil periods	Stress periods	P-value	Significance
Competitiveness and domestic demand					
Current account balance	-0.52	0.28	-4.57	0.00	yes
CPI	4.26	3.68	7.18	0.00	yes
Credit to GDP change	1.42	1.58	0.58	0.40	no
Unemployment change	-0.04	-0.14	0.48	0.00	yes
Consumption dynamics	-4.21	-3.80	-6.31	0.00	yes
Export share dynamics	0.60	0.78	-0.36	0.08	no
Financial					
Fixed capital formation dynamics	7.82	8.00	6.96	0.34	no
FX rate dynamics	1.87	0.69	7.82	0.00	yes
Fiscal					
GDP dynamics	2.90	3.14	1.71	0.00	yes
China GDP dynamics	9.62	9.63	9.59	0.55	no
US GDP dynamics	2.46	2.58	1.82	0.00	yes
Labor market					
Labor productivity dynamics	1.76	1.90	1.05	0.00	yes
GDP per capita	26.13	28.08	16.28	0.00	yes
Macroeconomic and global economy					
Interest on debt	3.58	3.32	4.92	0.00	yes
US interest rates	4.27	4.40	3.64	0.00	yes
Net lending	-2.47	-2.06	-4.55	0.00	yes
Oil price dynamics	5.05	5.87	0.89	0.03	yes
Currency overvaluation	-33.76	-31.59	-44.73	0.00	yes
Public debt	58.51	56.69	67.71	0.00	yes
VIX	20.17	20.07	20.70	0.74	no

Pairwise correlation is also examined and visualized in Figure 2. Accordingly, the pairwise correlations are low in most cases, which is why it can be presumed that the variables contain different information and could thus be relevant for the early warning model. However, some variables are highly correlated, which is problematic for econometric models such as the logit model and therefore should be excluded in the model specification, while high pairwise correlations are unproblematic for the random forest model and the variables concerned can be kept in the model specification (Jarmulska 2020).

Figure 2: Pairwise Correlation



4 Empirical Results

4.1 Performance

For each year in the interval from 2006 to 2016, both the logit model and the random forest model are implemented recursively, trying to classify the state of the economy two years later. For the logit model, all variables with pairwise correlations exceeding 65% are excluded to avoid multicollinearity and biases. The models are fitted twice using either a binary variable as an explanatory variable to indicate whether an economy is developed respectively developing or GDP per capita as a continuous measure for the state of development of an economy. If the binary variable is used, interaction terms are added to the logit model, while the random forest model takes interactions into account by construction, thus making further adjustments redundant (Jarmulska 2020).

The performance of the models measured using sensitivity (% of correctly classified stress episodes) and specificity (% of correctly classified tranquil episodes) depends on the threshold chosen, whereby a period is classified as a stress episode if the predicted dependent variable exceeds this threshold. Following Jarmulska (2020), this threshold is determined by maximizing the weighted average of sensitivity and specificity, weighting sensitivity by factor 1, 1.5 and 2 relatively to sensitivity. Table 2 compares the performance of the various models summarized as the average performance of the recursively trained models by year on the corresponding test data, using a threshold determined by the maximized weighted average using factor 1.5 (note that the AUROC is not dependent on the threshold chosen (Jarmulska 2020)). Thereby it can be seen that the model using random forest outperforms the logit LASSO approach in both specifications and concerning all performance metrics.

Table 2: Average Prediction Accuracy of Early Warning Models for Years 2009-2018

	Logit LASSO		Random Forest	
	advanced dummy	GDP per capita	advanced dummy	GDP per capita
sensitivity (% of correctly classified stress episodes)	89.76	76.23	87.69	87.9
specificity (% of correctly classified tranquil episodes)	53.39	69.85	68	69.58
Average of sensitivity and specificity (%)	71.58	73.04	77.84	78.74
AUROC	0.85	0.86	0.88	0.89

4.2 Interpretability of Random Forest Algorithm

4.2.1 Interpretation of Logit LASSO vs Random Forest

As seen in Table 2, the model using random forest outperforms the traditional logit model. However, depending on the application, performance is not the only relevant factor when choosing a model. Another central criterion in many use cases is the interpretability of the model, whereby the logit model benefits from the possibility of applying classical econometric methods (such as marginal effects, Wald test). But there are also various methods that can be used to interpret the random forest model. These methods are discussed in the following chapters.

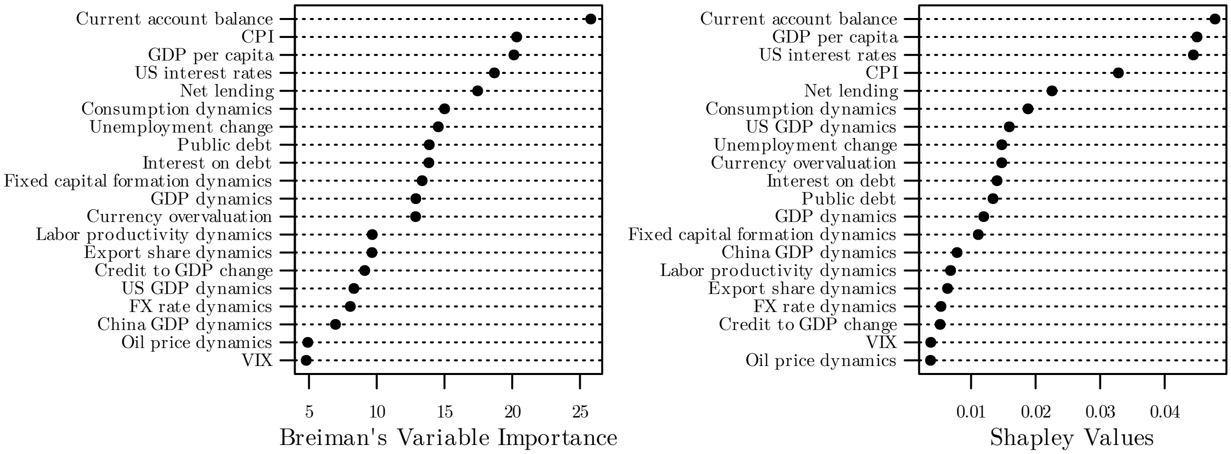
4.2.2 Variable Importance (Breiman 2001) and Shapley Values (Shapley 1953)

According to Jarmulska (2020) following Breiman (2001), variable importance corresponds to the average improvement of a model's performance caused by the addition of the respective variable. This improvement can be quantified either as the mean increase in performance or the mean decrease in impurity measured by the Gini index (see Equation (2)). When correlation is present between the different variables considered, it is important to be aware that the variable importance might be biased due to unrealistic observations caused by permutation when calculating the variable importance. In addition, the magnitude of the results will be decreased when correlation is present, since the actual importance is distributed between the correlated variables.

As Štrumbelj and Kononenko (2014) showed, Shapley values (see Shapley (1953)) - originally a solution concept in cooperative game theory - can be used to interpret the results of a machine learning algorithm by determining which variables are influential in a prediction. By assigning a quantitative value to each variable, Shapley values show the importance of a variable on the result of the prediction over a collection of observations (Ma and Tourani 2020). Further technical information can be found for example in Štrumbelj and Kononenko (2014) or in Ghorbani and Zou (2019).

Figure 3 shows the variable importance and the Shapley values for all variables considered in the model, whereby a higher value indicates a higher influence on the prediction. In line with the results obtained by Jarmulska (2020), both Breiman's variable importance and Shapley values similarly rank the importance of the variables.

Figure 3: Variable Importance and Shapley Values of Predictors used



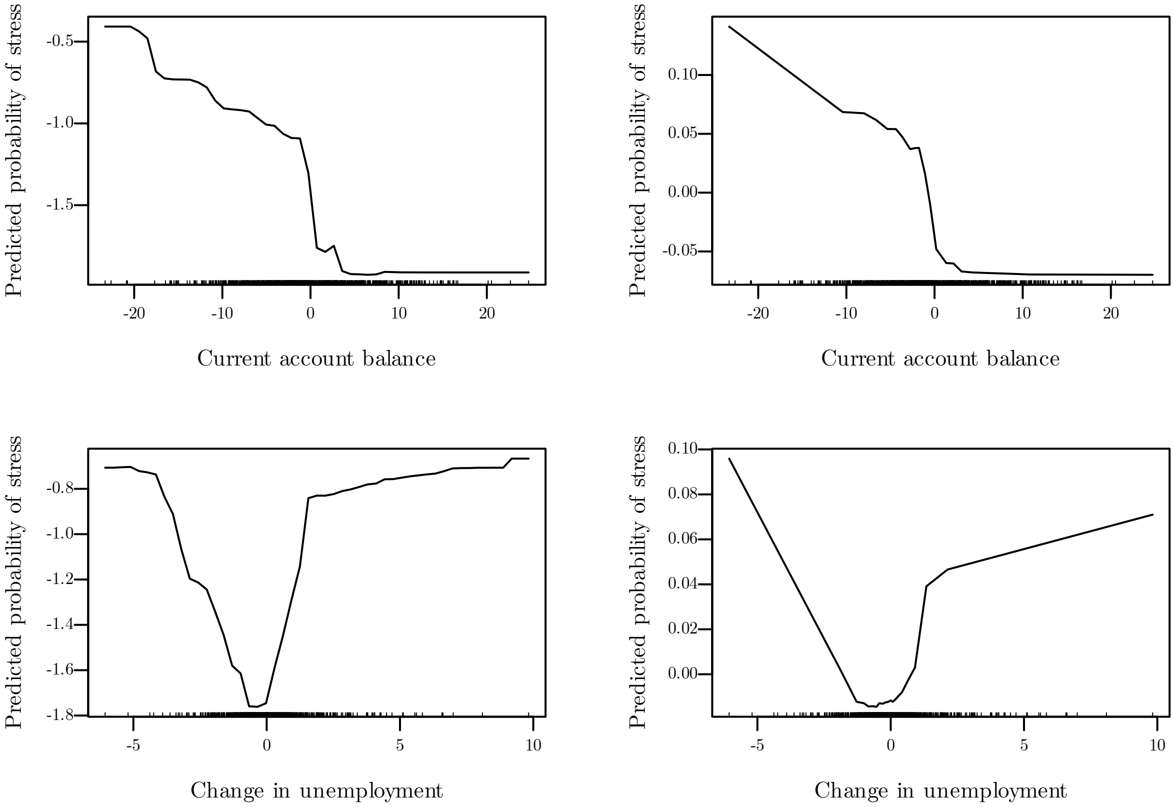
4.2.3 Partial Dependence Plots (Friedman 2001) and Accumulated Local Effects Plots (Apley and Zhu 2020)

In this application of predicting probabilities, partial dependence plots according to Friedman (2001) show the effect of a variable at different values on the predicted probability. To do this, all other independent variables are averaged and a model is fitted, depending on the varying variable to be considered. It should be noted that certain artificially created observations could bias the results, especially in the case of correlated variables, since these observations might be implausible in reality (Jarmulska 2020).

As an alternative to partial dependence plots, accumulated local effects plots following Apley and Zhu (2020) can be used, for which correlation among the independent variables does not lead to biases. Reason for this is, that accumulated local effects plots display the effect of a change in the variable to be considered on the predicted probability only in a small interval, accumulated over a grid of such small intervals, hence making pure effect of changes in this variable visible (Jarmulska 2020).

The partial dependence plots (left-hand side) and the local effects plots (right-hand side) are visible in Figure 4 for the variables *current account balance* and *unemployment changes*, whereby a similar shape of the curves is recognizable when comparing partial dependence against local effects. For many independent variables used in this model, an U-shaped curve can be detected when examining these plots, indicating that both up- and downshifts in the corresponding variable contribute to higher probability of fiscal stress. However, these U-shaped curves are strongly influenced by the tails of the distribution, potentially problematic for both partial dependence and accumulated local effects plots and should therefore be interpreted with caution (Jarmulska 2020). The rug plot on the x-axis in Figure 4 indicates the distribution of the variable considered.

Figure 4: Partial Dependence (LHS) and Accumulated Local Effects (RHS) Plots



5 Conclusion

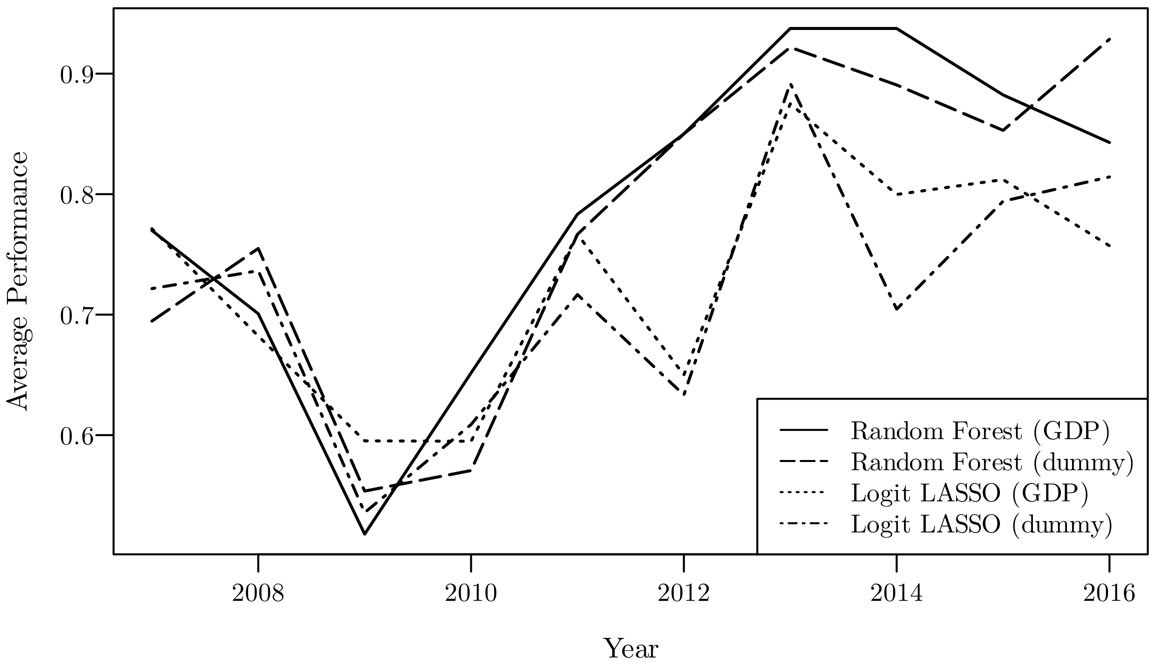
As obtained in Table 2, the random forest model with an average of sensitivity and specificity of 77-79% outperforms the logit LASSO model with an average of sensitivity and specificity of 71-73%. Jarmulska (2020) also attempts to estimate the probability of the occurrence of the first year of a fiscal stress episode only (i.e. excluding ongoing stress episodes), whereby prediction accuracy dropped by around 10%, indicating that predicting the first year of a fiscal stress episode is more difficult than predicting ongoing stress. However, Jarmulska (2020) claims that the objective of an early warning model is not to forecast a fiscal crisis, but rather to warn from a heightened level of vulnerability. Thus, the results of the proposed models can still be helpful and it is still worth to further develop these tools.

Examining the prediction accuracy over time, a drop can be observed during the sovereign debt crisis (2010-2012, see Figure 5). This illustrates that a purely quantitative model, such as the proposed logit LASSO and the random forest model, are unable to directly consider qualitative factors like the reputation or credibility of an economy (Jarmulska 2020). The strength of fiscal institutions is also a key factor influencing the probability of fiscal stress episodes, which is not considered in the models shown. Therefore, Jarmulska (2020) proposes the construction of an index which enables to consider these missing qualitative variables in the models.

Despite the accusation of being a black-box, section 4.2 showed that various methods can be used to provide interpretation of the results obtained by the machine learning approach. To further explore the usability and potential applications of machine learning in this and related questions, applying algorithms other than random forest could provide further valuable insights.

Appendix

Figure 5: Prediction Accuracy over Time



References

The code and data used for this project can be found in the corresponding GitHub-repository: `https://github.com/bt-koch/ML-in-Economics`.

Apley, Daniel W., and Jingyu Zhu. 2020. “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (4): 1059–86.

Beutel, Johannes, Sophia List, and Gregor von Schweinitz. 2019. “An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?” IWH Discussion Papers 2/2019. Halle Institute for Economic Research (IWH).

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Casabianca, Elizabeth Jane, Michele Catalano, Lorenzo Forni, Elena Giarda, Simone Passeri, et al. 2019. “An Early Warning System for Banking Crises: From Regression-Based Analysis to Machine Learning Techniques.” *EconPapers. Orebro: Orebro University.*

Dobrescu, Gabriela, Iva Petrova, Nazim Belhocine, and Emanuele Baldacci. 2011. “Assessing Fiscal Stress.” *IMF Working Papers* 11: 100.

Fawcett, Tom. 2006. “An Introduction to ROC Analysis.” *Pattern Recognition Letters* 27 (8): 861–74.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*, 1189–1232.

Ghoddusi, Hamed, Germán G. Creamer, and Nima Rafizadeh. 2019. “Machine Learning in Energy Economics and Finance: A Review.” *Energy Economics* 81: 709–27.

Ghorbani, Amirata, and James Zou. 2019. “Data Shapley: Equitable Valuation of Data for Machine Learning.” In *International Conference on Machine Learning*, 2242–51. PMLR.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Vol. 2. Springer.

Honda, Jiro, René Tapsoba, and Ismael Issifou. 2022. “When Do We Repair the Roof? Insights from Responses to Fiscal Crisis Early Warning Signals.” *International Economics.*

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Vol. 112. Springer.

Jarmulska, Barbara. 2020. “Random Forest Versus Logit Models: Which Offers Better Early Warning of Fiscal Stress?” *ECB Working Paper Series No 2408 / May 2020.*

Ma, Sisi, and Roshan Tourani. 2020. “Predictive and Causal Implications of Using Shapley Value for Model Interpretation.” In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, 23–38. PMLR.

Samitas, Aristeidis, Elias Kampouris, and Dimitris Kenourgios. 2020. “Machine Learning as an Early Warning System to Predict Financial Crisis.” *International Review of Financial Analysis* 71: 101507.

Shapley, L. 1953. “A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317.” In *Classics in Game Theory*, edited by Harold William Kuhn, 69–79. Princeton University Press.

Štrumbelj, Erik, and Igor Kononenko. 2014. “Explaining Prediction Models and Individual Predictions with Feature Contributions.” *Knowledge and Information Systems* 41 (3): 647–65.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.