UNIVERSITY OF TARTU
Faculty of Social Sciences
School of Economics and Business Administration

Master's thesis

EARLY WARNING SYSTEM FOR FINANCIAL CRISIS:APPLICATION OF RANDOM FOREST

Wanyama Geofrey

Supervisor: Mustafa Hakan Eratalay (Ph D)

Co-Supervisor: Luca Alfieri

Tartu 2020

Name and signature of supervisor . . . . . .

Allowed for defence on . . . . . .
(date)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced.

. . . . . .
(signature of author)

Abstract

The study identifies important variables in detecting the likely occurrence of a financial crisis 1 to 3 years from its onset . We do this by implementing random forest on Macroeconomic Historical time series data set for 16 developed countries from 1870-2016. By comparing the misclassification error for logistic regression to that obtained for random forest, we show that random forest outperforms logistic regression under the out-of-sample setting for long historical macroeconomic data set. Using the SMOTE technique, we show that minimising class imbalance in the data set improves the performance of random forest. The results show that important variables for detecting a financial crisis 1 to 3 years from its onset vary from country to country. Some similarities are however also observed. Credit and money price variables for instance emerge as very important predictors across a number of countries.

Keywords: Financial crisis, Random Forest, SMOTE, Historical Macroeconomic Data.

ACKNOWLEDGEMENTS

I wish to thank my supervisors Mustafa Hakan Eratalay and Luca Alfieri for their support and supervision.

I also thank Prof. Dr. Jaan Masso, the head of department for being such a supportive person throughout the course.

Finally, I thank my family for the support as well as all my classmates that made the two years bearable.

Contents

1 Introduction **5**

2 Literature review **7**

3 Data **11**

4 Methodology **12**

4.1 Target variable 12

4.2 Description of the models 13

4.2.1 Logistic Regression 13

4.2.2 Random Forest 13

4.3 Comparing Logistic regression and Random Forest 14

4.4 Fitting Random Forest Model 15

4.5 Boosting Random Forest using SMOTE 15

4.6 Variable Importance 16

5 Discussion of the results **16**

5.1 Comparing logistic regression and Random forest 16

5.2 Boasting prediction using SMOTE 17

5.3 Variable Importance 17

6 Conclusion **18**

7 References **20**

8 Appendix **24**

8.1 Table 1: Table showing Summary literature review 24

8.2 Table 2: Table showing Crisis years per country 1870-2008 25
8.3 Table 3: Variable names and description 26
8.4 Inspecting stationarity using Auto correlation Function (Before detrending) 27
8.5 Inspecting stationarity using Auto correlation Function (After detrending) 29
8.6 Table 4: Misclassification error for logistic regression and random forest on significant variables from imbalanced data 31
8.7 Table 5: Misclassification error for random forest before and after SMOTE 31
8.8 Variable Importance 32
8.9 Table 6: Variables included in each country model 48

1 Introduction

Early Warning system (EWS) for a long time broadly belonged to two categories; The signals approach pioneered by Kaminsky et al (1998) and the discrete binary dependent models. EWS based on these models however have overtime been failing short in identifying potential crisises prompting questions on the accuracy of these approaches in modeling crisis. The signals approach has for instance been criticized for not providing a way to aggregate the information provided by individual indicators (Demirgu c - Kunt and Detragiache 2005). Similarly, binomial discrete-dependent-variable models are inadequate in modeling tailed distributions associated with Financial crisis (Kumar et al, 2003), they are prone to post-crisis bias (Bussiere and Fratzscher, 2006).

There has thus existed a constant attempt to improve these methods and a desire to adopt new ones that improve predictions of crisises. In this effort, machine learning methods have started get traction as possible candidates for improving prediction. Previously, the adoption of machine learning methods such as random forest had been limited by the absence of large data set on which machine learning algorithms can be built. Overtime however, better data mining techniques and accumulation of data has made data more available which has seen the rise in the popularity and adoption of machine learning techniques.

In this study, we implement random forest to identify variables that are important in detecting the likely occurrence of a financial crisis 1 to 3 years from its onset in 16 developed countries. The choice of the algorithm is informed by it's ability to perform better than other techniques (Alessi and Detken, 2018; Tanaka et al., 2016; Holopainen and Sarlin, 2017), the easy with which it can be implemented and interpreted compared to other machine learning techniques that are more complicated such as NN, LSTM and which in some cases have more data requirements. Addition-

ally, unlike traditional econometric methods , the approach we propose is not limited by the distribution of the populations, it is more robust even with outliers and takes into account the interactions between multiple indicators.

By comparing the misclassification error of logistic regression and random forest fitted on only significant variables, the results show that random forest outperforms logistic regression when the two are applied to along historical macroeconomic data set under the out-of-sample setting.

To improve the performance of random forest, we minimise class imbalances in the data using the SMOTE technique which increases the decision space of the minority class by oversampling it using K-Nearest neighbours and bootstrapping. We show that complimenting random forest with techniques that minimise class imbalances within the data such as SMOTE improves the performance of random forest.

We thus contribute to the literature by proposing a random Forest based EWS. We extend and improve on related studies that have applied the same technique by using a large data spanning over 145 years provided by Jordà et al., (2019). We argue that previous studies that have employed the method did so on very limited data sample sizes often with very few crisis episodes unlike the data set used in this set which provides more than 90 crisis.

Additionally, our study is the first to our knowledge to minimise data imbalance in a historical macroeconomic data set used in this study by complimenting random forest with the SMOTE technique. This technique is an improvement from random sampling with replacement which has been widely used in previous studies because it doesn't propagate the bias of widening the decision space of the minority class on the same elements.

The rest of the paper is structured as follows; First we review previous related studies, we then discuss the data used in this study. The next section discusses the methodology adopted in this study followed by a discussion of the results and

the conclusion in section 6.

2 Literature review

Kaminsky et al (1998) are largely credited for pioneering early warning systems (EWS) for financial crisis following their seminal paper on the leading indicators of currency crises. They proposed a signals approach that involves monitoring the evolution of selected macroeconomic indicators and sending a signal when their values deviate from a set threshold value ("signal"). As an advantage, the signals approach provides a way to trace the root cause of the crisis to a single variable.

The approach however has its shortfalls. It was for instance criticised by Berg and Pattillo (1999) who argued that the approach yields very low explanatory power and commits high type I and type II errors. Moreover,Demirgüç-Kunt and Detragiache, 2005; Duca and Peltonen 2013 also noted that the signals approach doesn't provide a framework to evaluate the collective contribution of multiple variables in the prediction of crisis.

Following Berg and Pattillo (1999) seminal paper that advocated for the use of statistical models, many models in which a binary crisis indicator is simulated against macroeconomic variables have been used [Kumar et al, 2003; Berg and Coke, 2004; van den Berg et al., 2008; Jorda et al., 2010; Duca and Peltonen,2013; Candelon et al.,2014; Asanović, 2017; etc.].

For models under this category however, the logit model has been reported to perform better than its sister model the probit model. Probit models have been discredited as being poor at fitting fat tailed distribution such as those exhibited by crisises due to irregular occurrence (Kumar et al, 2003). Moreover Berg and Coke (2004) also showed that the ordinary probit models underestimate standard errors.

In an attempt to minimise the limitations associated to binomial discrete-dependent-variable models, some studies have advocated for further considerations when applying them. One such

consideration that emerges from the literature is the need to take into account the ability of crisis to persist (Bussiere and Fratzscher, 2006) and thus advocate multi-dynamic frame that takes into account the tranquil, pre-crisis, and post-crisis/recovery states.

Additionally,some studies have emphasised the heterogeneous nature of crisis across countries (Falcetti and Tudela, 2006; van den Berg et al, 2008) and cautioned against the adoption of panel data in EWS models as this poses the risk of perpetuating the assumption of constant and homogeneous crisis causing factors across countries. To take into account this heterogeneity, segmenting countries into clusters based on statistical methods has been recommended (Berg et al,2008)

Clustering however introduces limitations of it is own. First, if clustering is aimed at mimicking homogeneous crisis causing conditions among a group of countries (countries that have related conditions or economic behavior), it would be expected that such countries experience crisis simultaneously or within a close time period. There is however no sufficient evidence of a cluster of countries experiencing crisis simultaneously (Jordà et al.,2010). Second, it considerably limits the data left to work with. As such, generalizing findings to other countries may raise questions.

Additional caution regarding the adoption of binary-dependent models comes from Candelon et al (2014) who like Bussiere and Fratzscher (2006) observed the persistence of crisises and advocate for taking into account exogenous effect of the persistence. However, according to Jordà et al (2010), the occurrence of a crisis doesn't depend on the time since the last occurrence.

These contradictions perhaps point to the fact that modeling rare events such as financial crisis is not an easy task and consensus on the best method cannot easily be established. There has thus a need to always try out new ways of modeling financial crisis depending on the resources and opportunities that become available with time. One such resource and opportunity that has come with time is the ac-

cumulation of data spanning over a long horizon which permits the adoption of new techniques or improvement of the existing ones.

More recently, EWS based on nonparametric methods have emerged. Decision tress [Martinez,2016; Sevim et al., 2014; Holopainen and Sarlin, 2017], Artificial Neuron Networks [Aydin et al, 2015; Sevim et al., 2014] are among some of the techniques that have gained traction in the literature in recent years.

These methods are easy to explain (Sevim et al., 2014) and have the ability to consider indicators collectively (Alessi and Detken, 2018). However, besides being relatively harder to apply, Machine learning techniques also have additional requirements. They for instance require sufficiently large data to produce robust results (Martinez, 2016) in contrast to traditional econometric methods that perform well even with small datasets. This concern in part has informed the choice of the dataset used in this study because of its longevity (spans over 145 years). Additionally, despite the good performance of decision trees, their performance is not very robust with additional predicators (Alessi and Detken, 2018). They recommend aggregating multiple trees for better performance which is precisely what random forest does. Random forest from this perspective has three major advantages; it takes into account interactions between multiple indicators, it is less affected by outliers, and is not limited by the underlying distribution or assumptions made about the population.

Whereas Random Forest has been widely used in other fields such as political inteligence, it has not been widely used in macroeconomics studies mainly due to the frequency with which macroeconomic phenomenon are observed. Most macroeconomic indicators used in modeling macroeconomic phenomenon such as financial crisis are observed on annual basis which limits the data need for the application of such methods. There also concerns surrounding the "black box" nature of the method. There is some skepticism also as to whether methods such as

Random Forest improve predictions than the traditional econometric methods such as logit and probit models. The rationale here is that if Random forest doesn't significantly improve predictions, then it is not worthy in terms of the associated costs such as the large data requirements.

Studies aimed at comparing the performance of econometric methods (logit) and machine learning techniques (Random Forest) have concluded differently. The difference in results on which model is better can be attributed to different things; difference in the quality of data used is one such reason. Beutel et al, (2018) for instance favors logit model over Random Forest under the out-sample setting but use a relatively small data set spanning 45 years. This limited data set we argue; favors the logit model that performs relatively better even with small data sets than Random forest that requires relatively large data sets. The data set we employ in this study covers a period of 170 years and thought there is no standard threshold for "enough data', we use a relatively large data set. Secondly and perhaps most importantly, the disagreement over which model is better seems to stem from the difference in the model evaluation method adopted. Studies using the out-sample evaluation approach have generally concluded that the logit model is more robust and outperforms Random Forest in predicting Financial Crisis (Beutel et al, 2018; Daniel, 2017). In contrast, studies that have employed other methods notably the k-Fold cross validation have concluded in favor of Random forest (Alessi and Detken, 2018; Tanaka et al., 2016). Some studies however have criticized this approach arguing that it over-estimates the performance of machine learning techniques (Holopainen and Sarlin, 2017; Neunhoeffer and Sternberg ;2018).

But even in case where the same method of model evaluation has been used, contradicting results have been obtained. Holopainen and Sarlin (2017) for instance used the same out-of-sample approach used by Beutel et al, (2018) and concluded that Random Forest outperforms the logit model. This difference

in outcomes may in part be to the fact that the performance of machine learning techniques such as Random Forest depends on the choice of hyperparameters used which may vary depending on one's level of experience and expertise. A careful model specification is therefore crucial for attaining improved predictions from machine learning techniques.

3 Data

The data set used in this study comes from the Jordà-Schularick-Taylor Macrohistory Database provided by Jordà et al., (2017). It is an annual data set running from 1870-2016 and includes 16[1] developed countries namely Australia, Belgium, Denmark, Finland, France, Germany, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, UK, USA.

The data set has been consolidated from many sources to include extensive series on many macroeconomic indicators which makes it one of the longest running panel data set on macroeconomic variables and has widely been used in related studies. This data has therefore been chosen because of the longevity of the data series which enables working with random forest and the extensive nature of the variables available.

The target variable is a dummy variable coded by Schularick and Taylor (2012) who also extended on the previous studies by Bordo et al. (2001) as well as Reinhart and Rogoff (2009). The variable takes on the value of 1 if a crisis happened, otherwise it takes on 0. Table 2 in the appendix shows crisis considered for each country.

The data has missing information which is different for each country and for each variable. To overcome this problem, as a general rule of thumb, for each country, any series that is missing more than a quarter (15%) of the time under consideration is dropped. We then impute the remaining missing data using linear interpolation.

We perform the Augment Dicker-Fuller

[1]The original data sets has 17 countries but Canada is excluded from this study because of missing data

test for stationarity and consequently transform the series using lag differencing.

One common practice in the EWS literature is the splitting of data into pre and post world War II; however the method we adopt in this study requires large data and we thus don't split the data. Instead we follow Schularick and Taylor (2012) and exclude data covering the periods of the two World Wars.[2]

4 Methodology

We apply random forest to identify key variables for detecting the likely occurrence of a financial crisis in the next 1 to 3 years (4 to 12 quarters). To do this, we asses the relative importance of these variables in predicting the probability of a financial crisis happening in a given time period.

Financial crisises by their nature are very rare events and predicting the exact time when one will happen has proved very difficulty. In line with the standard practice, this study doesn't focus on predicting the exact time when the crisis will happen but the probability of happening in a given time range (1 to 3 years in this study). We then identify important variables in detecting financial crisis by assessing their impact on the Out-of-Bag error.

4.1 Target variable

A key desirable feature of an early warning system is the ability to detect a crisis in time to allow for the policy makers to make interventions or make policy changes. Therefore in choosing the window time frame, one must keep a balance so that it is long enough to allow policy interventions and close enough to permit the observation of evolution in the build up to the crisis (Beutel, List and von Schweinitz, 2019). To achieve this, we transform the original database financial crisis dummy variable $\tilde{C}$ into a new target variable. Our new target variable is a dummy variable which has value 1 if a crisis happened in the next n number of

[2]Excluded data from 1914-1919 for World War I and from 1939-1947 for world War II

years.

$$c_{t,n} = \begin{cases} 1, & \text{if } \tilde{C}_{t+n} = 1 \\ 0, & \text{otherwise} \end{cases}$$

where n=1,2,3

The crisis periods are then excluded from the data to minimise bias arising from the already existing imbalances in those periods.We therefore estimate the probability

$$P(c_{t,n}|X_t) \tag{1}$$

of a crisis happening in the next 1 to 3 years (where X_t is a vector of predictors).

4.2 Description of the models

4.2.1 Logistic Regression

Our benchmark model is logistic regression which we fit as follows:

$$Prob(Y_t = 1|X_t) = \frac{e^{X_t\beta}}{1 + e^{X_t\beta}} \tag{2}$$

where $Prob(Y = 1|X_t)$ is the probability of country being in a crisis one to three years from t and X_t) is a vector of predictors.

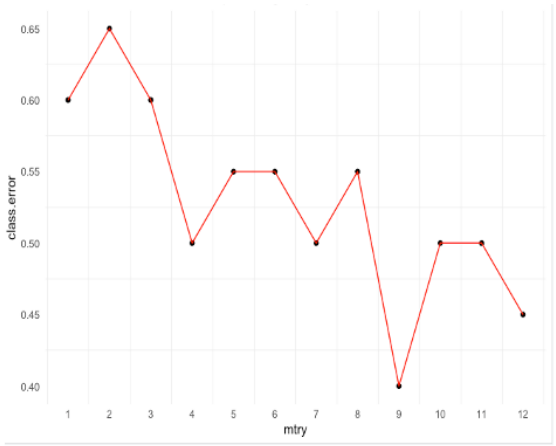


Figure 1: The figure shows the number of mtry that yeilds the least classification error

4.2.2 Random Forest

Random forest which was pioneered by Breiman (1996) randomly selects subsets of observations and estimates decision trees on them.

We implement the random forest algorithm using the "Random forest" Library in R software. The algorithms takes on three key hyperparameter that specify the number of trees to grow, number of variables to sample at each split and the minimal number of observations per terminal node. To optimise the performance of the algorithm, we seek to set the combination of hyparameters that minimise

the classification error. To do this, we run different models on the train data set using different combination of hyparameters and chose the combination that yield the least error rate. In figure 1 above, error rate is minimal when mtry equals to 9.

4.3 Comparing Logistic regression and Random Forest

To fit the model, first we split the data set into two mutually exclusive training and testing sets.The common practice is allocating 75% to the training set and 25% to the training set. The rationale behind allocating more data to the training set is to provide enough data for training the model. The test data set is used for validating the model.

We fit the logistic regression model using all variables available for each country and perform backward elimination based on a chosen level of significance. We then refit the model dropping a variable with maximum p -value greater than 0.05 until all the variables are significant at 5% level of significance.The misclasification error of the fitted logistic regression is then calculated

We then fit a random forest model on the train set containing the variables included in the logistic regression and it's misclassification error is obtained. The ME of the two models are compared and the model with the lowest ME value is considered to be better at fitting the crisis.Figure 2 illustrates the process of variable selection and model comparison.

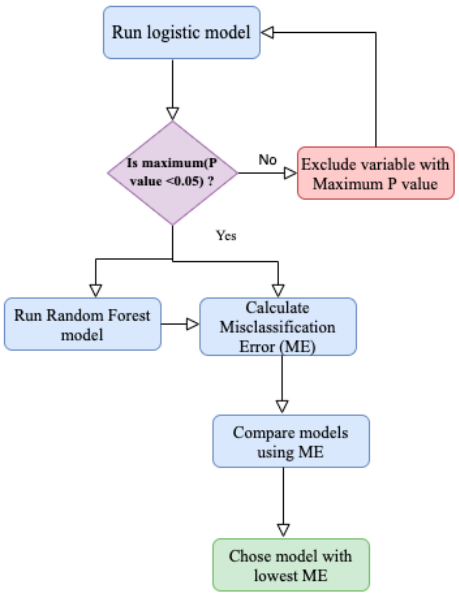


Figure 2: The figure illustrates the process of variable selection and model comparison

4.4 Fitting Random Forest Model

After the preliminary comparison between the two models in which random forest performs better than logistic regression based on the misclassification errors, we fit a random forest model for all variables available for each country. To do this, we divide the data set into two mutually exclusive training and testing sets allocating 75% to the training set and 25% to the test set. Splitting the data is aimed at facilitating cross validation while minimising the risk of over-fitting which is associated to in-sample validation

To optimise the performance of the algorithm, we set the number of parameters as discussed earlier in section 4.2.2.

4.5 Boosting Random Forest using SMOTE

A common challenge from the Early warning literature is the imbalanced nature of the data on which models are built.The data used in this study is no exception, the crisis periods account for only approximately 5% of the total years available.

Applying machine learning algorithms to highly unbalanced data poses the challenge of biasing the algorithm towards the majority class. We seek to minimise this problem by increasing the share of crisis ($C_{t,n} = 1$) in the data using the Synthetic Minority Oversampling Technique (SMOTE).

This technique which was pioneered by Chawla et al (2002) proposes creating additional examples of the minority class using the bootstrapping and K-nearest neighbours through the process of under sampling the majority class while oversampling the minority class.

We implement the SMOTE algorithm in R software using the "SMOTE" function from "DMwR" library which takes two key parameters; "perc.over" and "perc.under" which control oversampling and under sampling of the the minority and majority category respectively. We set these two parameters differently for each country depending on existing imbalance in the country data. In most

cases, We increase the share of crisis to 15% by over-sample the minority class while under-sampling the majority class.

Because the SMOTE algorithm depends on the K-nearest neighbour, we normalize the data using

$$\bar{x} = \frac{x - x_{min}}{x_{max} - X_{min}} \tag{3}$$

Normalizing data improves the performance of algorithms that depend on distance between the data points.

We fit a new model as describe in section 4.4 using the data set transformed using SMOTE and obtain the misclassification error for the new model. We then compare the ME of the new model to the initial one.

4.6 Variable Importance

To identify variable importance, we assess the variable's impact on out-of-bag (OOB) accuracy each time the variable is permuted. Changes in OBB rate[3] when a variable is randomly permuted indicates high importance of the variable.

[3]subtracting the OBB rate with variable j permuted minus OBB rate without the permutation of variable j

5 Discussion of the results

5.1 Comparing logistic regression and Random forest

Table 4 in the appendix shows the misclassification errors for both models for all countries. Overall, Random forest performs better than logistic regression for all countries except Denmark were the two models have the same error rate.Moreover the choice of variables is limited to only variables that are significant using logistic regression.This finding is inline with previous studies such as Alessi and Detken (2018),Holopainen and Sarlin (2017),anaka et al., 2016 but differs from Beutel, List and von Schweinitz, 2019 who concluded that logistic regression outperformed random forest. The difference could be attributed to the difference in sample size employed. The data set used in this study covers a span of 146 years and includes more crisis episodes while in their study, the sample size covers 45 years.

5.2 Boasting prediction using SMOTE

Table 5 shows the misclassification errors for random forest before and after implementing the SMOTE algorithm. The results show that the random forest model built on data with reduced imbalance using the SMOTE performs slightly better than the model built on the original highly imbalanced data set.This finding is consistent with previous related studies such as (Shrivastava, Jeyanthi and Singh, 2020). Reducing the decision space of the majority class while increasing that of minority class improves prediction.

5.3 Variable Importance

Figures 7 to 22 show the variable importance of random forest models for different countries. The results show that the importance of variables varies from country to country. Credit variables such as total loans to the non-financial private sector, mortgage loans to the non-financial private sector, total loans to households and total loans to business emerge as very important in detecting a financial crisis in Australia, Belgium, Denmark, France,Italy, Norway, Switzerland and Portugal. This is inline with findings by previous studies such as Schularick and Taylor,2012; Fricke, 2017 who concluded that credit growth is key in predicting financial crisis.

Rates of return on assets is important in detecting financial crisis in Netherlands, Norway and Portugal. Housing prices are very important in detecting crisis in Norway, Australia,Sweden and USA. This is inline with the findings of Beutel et al., 2019; Kindleberger et al., 2011; Jord‘a et al., 2015 who concluded that real estate prices as well as asset prices drive crisises especially if they are debt-financed.

Money prices and interest rates are important in detecting financial crisis in Portugal,Spain, USA and UK. Similar findings have been made by Sevim et al., 2014.Real economy variables are generally important but appear specifically important in Australia, Belgium, Finland,France,Germany and Switzerland. Public debt to GDP ratio, govern-

ment revenue and expenditure are important in Belgium,Italy,Japan,Netherlands, Sweden,USA and UK.

The difference in variable importance across countries points to the heterogeneity in crisis causing factors across countries. Some caution should however be taken when interpreting this results since the variables included in the model differ from country to country depending on availability. Thus some variables that appear very important for some country may not have been available for another country. Table 6 in the appendix shows the variables included in each country model. Generally, in additional to the general real economy variables, credit and monetary variables emerge as very important variables for detecting a financial crisis 1 to 3 years from it's onset.

6 Conclusion

In this study, we have identified variables that are important for detecting that a financial crisis may occur 1 to 3 years from it is onset. To do this, first we show that random forest performs better than our benchmark model, logistic regression on long historical macroeconomic data.

We have minimised class imbalance in the data which is a major problem in modeling crisis due to the irregular nature of their occurrence. We have shown that the SMOTE technique improves the performance of random forest. Future studies may focus on adopting methods that optimize machine learning techniques by complimenting them with better methods that minimize the data imbalance which is still a problem.

The key finding of the study is that whereas variables that are important in detecting that a financial crisis may occur in a country 1 to 3 years from it is onset vary from country to country, some similarities are observed. Credit and monetary variables for instance emerge as very important in detecting financial crisis across a number of countries. Asset and housing prices in addition to the traditional real economy variables were also found to be specifically important among

countries.

7 References

Alessi, L. and Detken, C. (2018). Identifying excessive credit growth and leverage. Journal of Financial Stability, 35, pp.215-225.

Asanović, Ž. (2017). Predicting Systemic Banking Crises Using Early Warning Models: The Case of Montenegro. Journal of Central Banking Theory and Practice, 6(3), pp.157-182.

Aydin, Alev çalışkan çavdar, şeyma. (2015). Prediction of Financial Crisis with Artificial Neural Network: An Empirical Analysis on Turkey. International Journal of Financial Research. 6. 10.5430/ijfr.v6n4p36.

Beutel, Johannes List, Sophia Von Schweinitz, Gregor. (2018). An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions?.

Beutel, J., List, S. and von Schweinitz, G., 2019. Does machine learning help us predict banking crises?. Journal of Financial Stability, 45, p.100693.

Bordo, M., Eichengreen, B., Klingebiel, D. and Martinez-Peria, M., 2001. Is the crisis problem growing more severe?. Economic Policy, 16(32), pp.52-82.

Bussiere, M. and Fratzscher, M. (2006). Towards a new early warning system of financial crises. Journal of International Money and Finance, 25(6), pp.953-973.

Breiman, L. (1996). Bagging Predictors. Machine Learning 24 (2), 123–140.

Breiman, L. (2001). Random Forests. Machine Learning 45 (1), 5–32.

Candelon, B., Dumitrescu, E. and Hurlin, C. (2014). Currency crisis early warning systems: Why they should be dynamic. International Journal of Forecasting, 30(4), pp.1016-1029.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, pp.321-357.

Coke, Rebecca Berg, Andrew. (2004). Autocorrelation-Corrected Standard Errors in Panel Probits: An Application to Currency Crisis Prediction. IMF Working Papers. 04. 10.5089/9781451845860.001.

Demirguc-Kunt, A. and Detragiache, E. (2000). Monitoring Banking Sector Fragility: A Multivariate Logit Approach. The World Bank Economic Review, 14(2), pp.287-307.

Demirgüç-Kunt, A. and Detragiache, E. (2005). Cross-Country Empirical Studies of Systemic Bank Distress: A Survey. National Institute Economic Review, 192(1), pp.68-83.

Duca, M. and Peltonen, T. (2013). Assessing systemic risks and predicting systemic events. Journal of Banking Finance, 37(7), pp.2183-2195.

Fricke, D. (2017). Financial Crisis Prediction: A Model Comparison. Deutsche Bundesbank; University College London; London School of Economics Political Science (LSE) - Systemic Risk Centre.

Holopainen, M. and Sarlin, P. (2017). Toward robust early-warning models: a horse race, ensembles and model uncertainty. Quantitative Finance, 17(12), pp.1933-1963.

Jordà, Ò., Schularick, M. and Taylor, A. (2011). Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons. IMF Economic Review, 59(2), pp.340-378.

Jordà, Ò., Schularick, M. and Taylor, A., 2015. Leveraged bubbles. Journal of Monetary Economics, 76, pp.S1-S20.

Kaminsky, Graciela Lizondo, Saul Reinhart, Carmen. (1998). Leading Indicators of Currency Crises. International Monetary Fund. 45. 10.1596/1813-9450-1852.

Kumar, M., Moorthy, U. and Perraudin, W. (2003). Predicting emerging market currency crashes. Journal of Empirical Finance, 10(4), pp.427-454.

Michie, R., 2012. Charles P. Kindleberger and Robert Z. Aliber, Manias, panics and crashes: a history of financial crises (New York: Palgrave Macmillan, 6th edn., 2011. Pp. viii + 356. 3 tabs. ISBN 9780230365353 Pbk. . The Economic History Review, 65(4), pp.1609-1611. Neunhoeffer, M. and Sternberg, S. (2018). How Cross-Validation Can Go Wrong and What to Do About It. Political Analysis, 27(1), pp.101-106.

Nicole, M. (2016). Predicting Financial Crises. Wharton Research Scholars. 136.

Olivier, B., Angela, D. (2010). Euro area GDP forecasting using large survey datasets. A random forest approach. Euroindicators working papers

Òscar Jordà, Moritz Schularick, and Alan M. Taylor. 2017. "Macrofinancial History and the New Business Cycle Facts." in NBER Macroeconomics Annual 2016, volume 31, edited by Martin Eichenbaum and Jonathan A. Parker. Chicago: University of Chicago Press.

Pattillo, C. and Berg, A. (1998). Are Currency Crises Predictable? a Test. IMF Working Papers, 98(154), p.1.

Rose, A. and Spiegel, M. (2012). Cross-country causes and consequences of the 2008 crisis: Early warning. Japan and the World Economy, 24(1), pp.1-16.

Sevim, C., Oztekin, A., Bali, O., Gumus, S. and Guresen, E. (2014). Developing an early warning system to predict currency crises. European Journal of Operational Research, 237(3), pp.1095-1104.

Schularick, M. and Taylor, A., 2012. Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870–2008. American Economic Review, 102(2), pp.1029-1061.

Tanaka, K., Kinkyo, T. and Hamori, S. (2016). Random forests-based early warning system for bank failures. Economics Letters, 148, pp.118-121.

Tudela, Merxe Falcetti, Elisabetta. (2006). Modelling Currency Crises in Emerging Markets: A Dynamic Probit Model with Unobserved Heterogeneity and Autocorrelated Errors. Oxford Bulletin of Economics and Statistics. 68. 445-471. 10.1111/j.1468-0084.2006.00172.x.

van den Berg, J., Candelon, B. and Urbain, J. (2008). A cautious note on the use of panel models to predict financial crises. Economics Letters, 101(1), pp.80-83.

Shrivastava, S., Jeyanthi, P. and Singh, S., 2020. Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. Cogent Economics Finance, 8(1).

8 Appendix

8.1 Table 1: Table showing Summary literature review

Author and paper	Dependent	Sample and variables	Methods	Results
Alessi and Detken (2018)	Banking Crisis	EU countries 1970-2010	Random Forest	Random forest Early warning model outperforms the logit model under the K-cross validation setup.
Jordà et al., (2010)	Financial Crisis	14 developed countries 1870-2008	Logit, ROC	An increase in credit growth and a decrease in the natural interest rate tends to precede gloabal financial crisis Including External imbalances improve financial crisis prediction
Candelon et al (2014)	Financial Crisis	16 emerging countries, 1985-2011	Logit	Taking into accoun the endogenus persisitence of the crisis in predicting it improves prediction compared to static models that consider only exogenous macroeconomic
Aydin et al, (2015)	Financial Crisis	Turkey, 1990-2004	Artificial Neural Network	Contagion effects and foreign excahnage rates play a role in the occurrence of crisis in emerging economies
Beutel et al, (2018)	Systematic Banking Crisis	15 advanced countries, 1970-2016	Logit, Random Forest, Knn	machine learning Early warning stems need improvement to out perform traditional econometric methods notably the logit model.Credit Expansions, asset price booms and external imbalances are key indicators of a
Jeffrey and George, (2011)	Financial crisis	All countries, 2007-2011	Regression	Bank reserves and exchange rate movements are the leading indicators of a crisis
Sevim et al.,(2014)	Currency crisis	Turkey, 1992-2011	Artificial neural networks (ANN), decision trees, and	artificial neural networks, decision trees, and logistic regression out perform other econometric methods in terms of classification
Daniel, 2017	Financial Crisis	14 developed countries 1870-2008	Logit , Classification trees and forests,KNN, Neural Networks (NN), , Support Vector Machines (SVM),	Forests reduce the prediction volatilities associated with trees.They avergae predictions over alarge number of trees which reduces the variance of individual trees. Credit growth is the most important indicator of finanial crisis
van den Berg et al., (2008)	Financial Crisis	13 countries from South America and South East Asia,	Logit	Pooling country data reduces the quality of predictions. Preliminary country clustering is necessary to obtain better predictions.
Bussiere and Fratzscher, (2006)	Financial Crisis	32 emerging countries, 1993-2006	Multi-nominal logit	Probit/Logit based Early warning system models are prone to post crisis bias. The inclusion of contagion ariables improvves prediction of financial crisis
Duca and Peltonen, (2013)	Financial Crisis	28 Emerging and developed countries 1990-2009	Logit	Inclussion of both domestic and global macro-financial indicators improves the prediction of financnial crisis substancially

8.2 Table 2: Table showing Crisis years per country 1870-2008

Country	Crisis years								
Italy	1873	1887	1893	1907	1921	1930	1935	1990	2008
Spain	1883	1890	1913	1920	1924	1931	1977	2008	
Belgium	1870	1885	1925	1931	1934	1939	2008		
Denmark	1870	1885	1925	1931	1934	1939	2008		
USA	1873	1893	1907	1929	1984	2007			
Sweden	1878	1907	1922	1931	1991	2008			
Germany	1873	1891	1901	1907	1931	2008			
Japan	1871	1890	1907	1920	1927	1997			
Switzerland	1870	1910	1931	1991	2008				
Netherlands	1893	1907	1921	1939	2008				
Portugal	1890	1920	1923	1931	2008				
Finland	1877	1900	1921	1931	1991				
France	1882	1889	1930	2008					
UK	1890	1974	1991	2007					
Norway	1899	1922	1931	1988					
Australia	1893	1989							

8.3 Table 3: Variable names and description

Category	Variable label	Variable description
Credit Data	tloans	Total loans to non-financial private sector (nominal, local currency)
	tmort	Mortgage loans to non-financial private sector (nominal, local currency)
	thh	Total loans to households (nominal, local currency)
	tbus	Total loans to business (nominal, local currency)
Government	debtgdp	Public debt-to-GDP ratio
	revenue	Government revenues (nominal, local currency)
	expenditure	Government expenditure (nominal, local currency)
House Prices	hpnom	House prices (nominal index, 1990=100)
International	ca	Current account (nominal, local currency)
	imports	Imports (nominal, local currency)
	exports	Exports (nominal, local currency)
	xrusd	USD exchange rate (local currency/USD)
Money, Prices & Interest Rates	cpi	Consumer prices (index, 1990=100)
	narrowm	Narrow money (nominal, local currency)
	money	Broad money (nominal, local currency)
	stir	Short-term interest rate (nominal, percent per year)
	ltrate	Long-term interest rate (nominal, percent per year)
Rates of Return	eq_tr	Equity total return, nominal. r[t] = [[p[t] + d[t]] / p[t-1]] - 1
	housing_tr	Housing total return, nominal. r[t] = [[p[t] + d[t]] / p[t-1]] - 1
	bond_tr	Government bond total return, nominal. r[t] = [[p[t] + coupon[t]] / p[t-1]] - 1
	bill_rate	Bill rate, nominal. r[t] = coupon[t] / p[t-1]
	housing_capgain	Housing capital gain, nominal. cg[t] = [p[t] / p[t-1]] - 1
	housing_rent_rtn	Housing rental return. dp_rtn[t] = rent[t]/p[t-1]
	housing_rent_yd	Housing rental yield. dp[t] = rent[t]/p[t]
	eq_capgain	Equity capital gain, nominal. cg[t] = [p[t] / p[t-1]] - 1
	eq_dp	Equity dividend yield. dp[t] = dividend[t]/p[t]
	bond_rate	Gov. bond rate, rate[t] = coupon[t] / p[t-1], or yield to maturity at t
	eq_div_rtn	Equity dividend return. dp_rtn[t] = dividend[t]/p[t-1]
	capital_tr	Tot. rtn. on wealth, nominal. Wtd. avg. of housing, equity, bonds and bills
	risky_tr	Tot. rtn. on risky assets, nominal. Wtd. avg. of housing and equity
	safe_tr	Tot. rtn. on safe assets, nominal. Equally wtd. avg. of bonds and bills
Real Economy	pop	Population
	rgdpmad	Real GDP per capita (PPP)
	rgdppc	Real GDP per capita (index, 2005=100)
	rconpc	Real consumption per capita (index, 2006=100)
	gdp	GDP (nominal, local currency)
	iy	Investment-to-GDP ratio

8.4 Inspecting stationarity using Auto correlation Function (Before detrending)

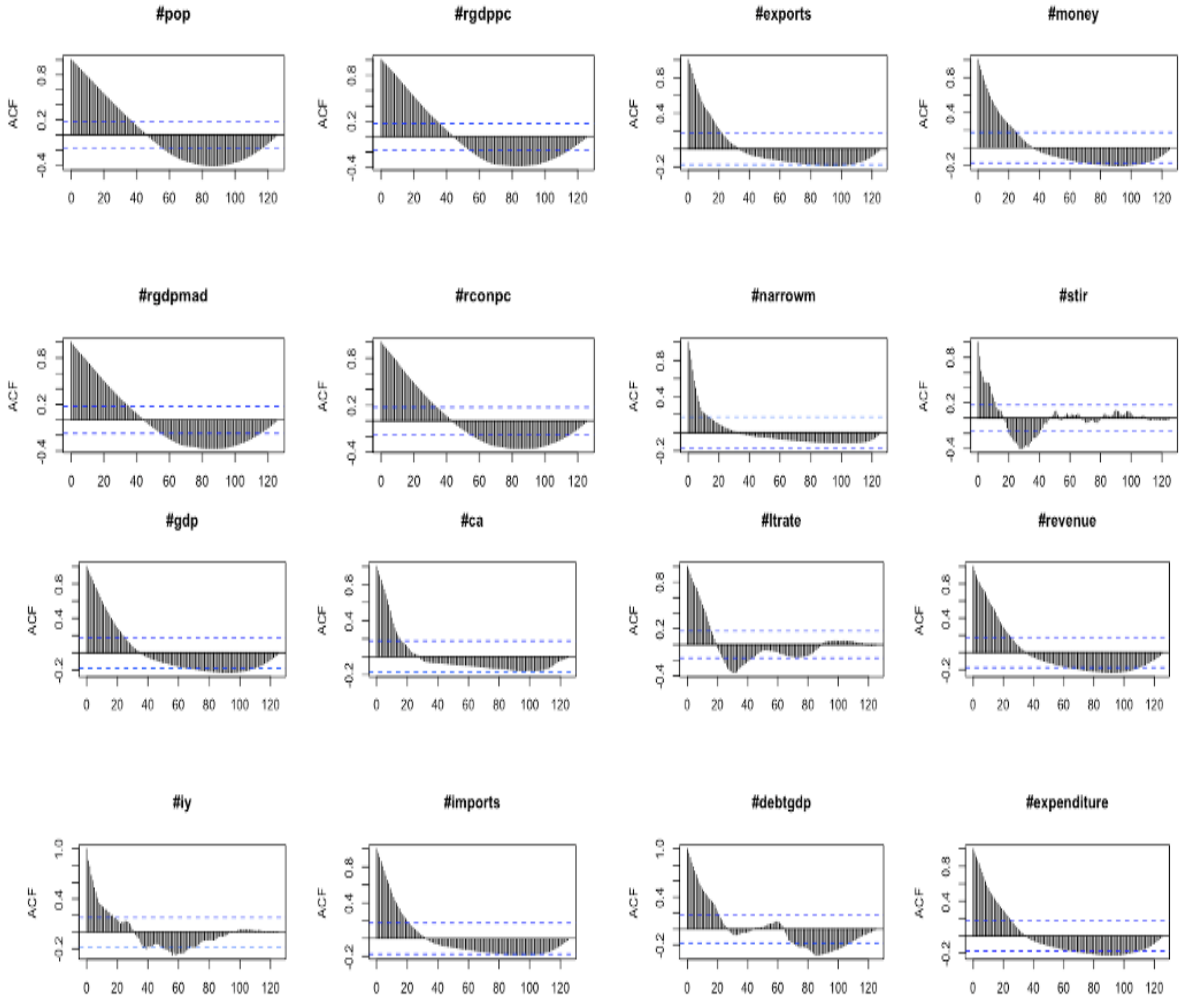


Figure 3: The figure shows the ACF plots for the different series. For stationary series, a decay in lags overtime is expected

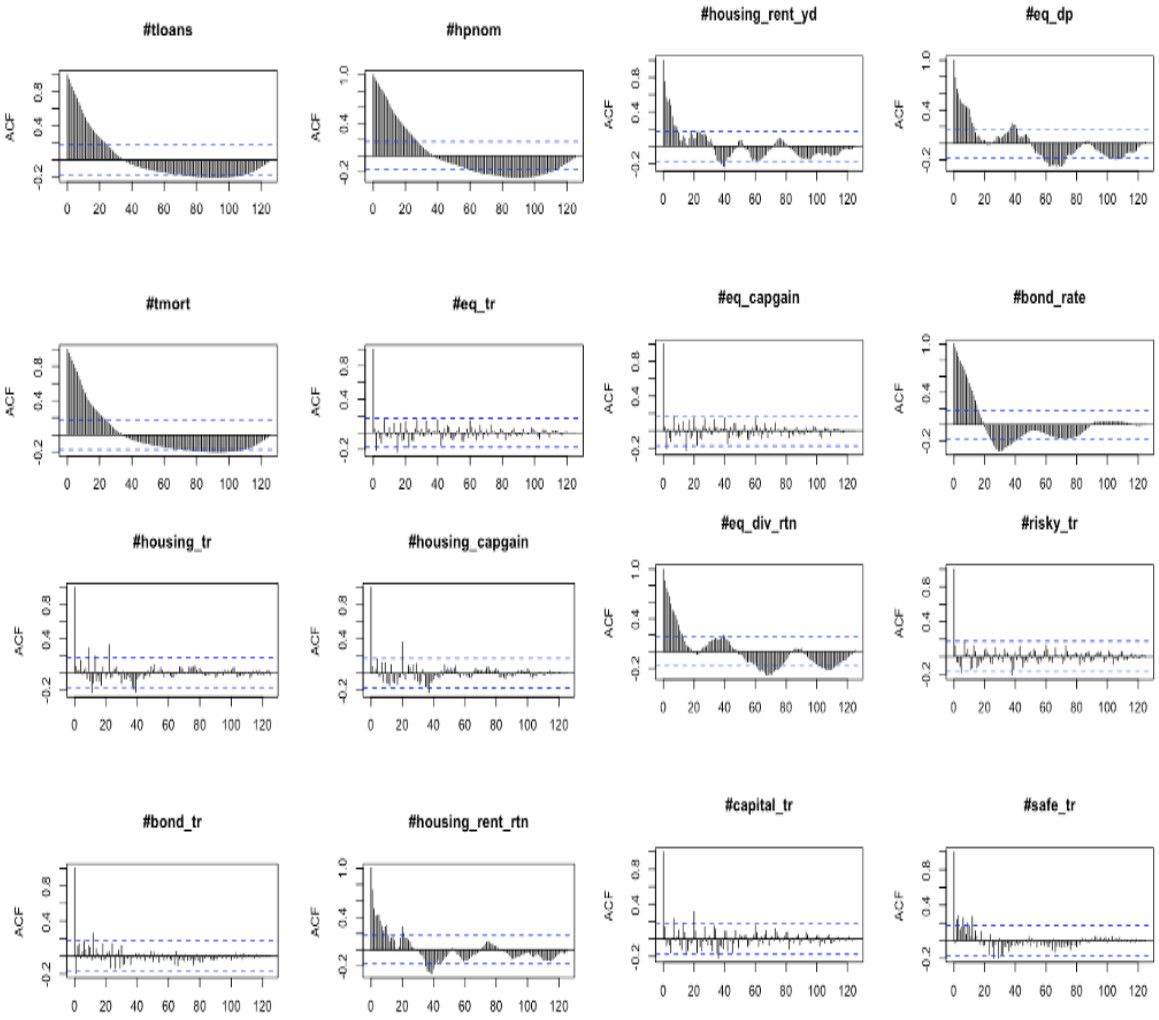


Figure 4: 7.4 continued

8.5 Inspecting stationarity using Auto correlation Function (After detrending)

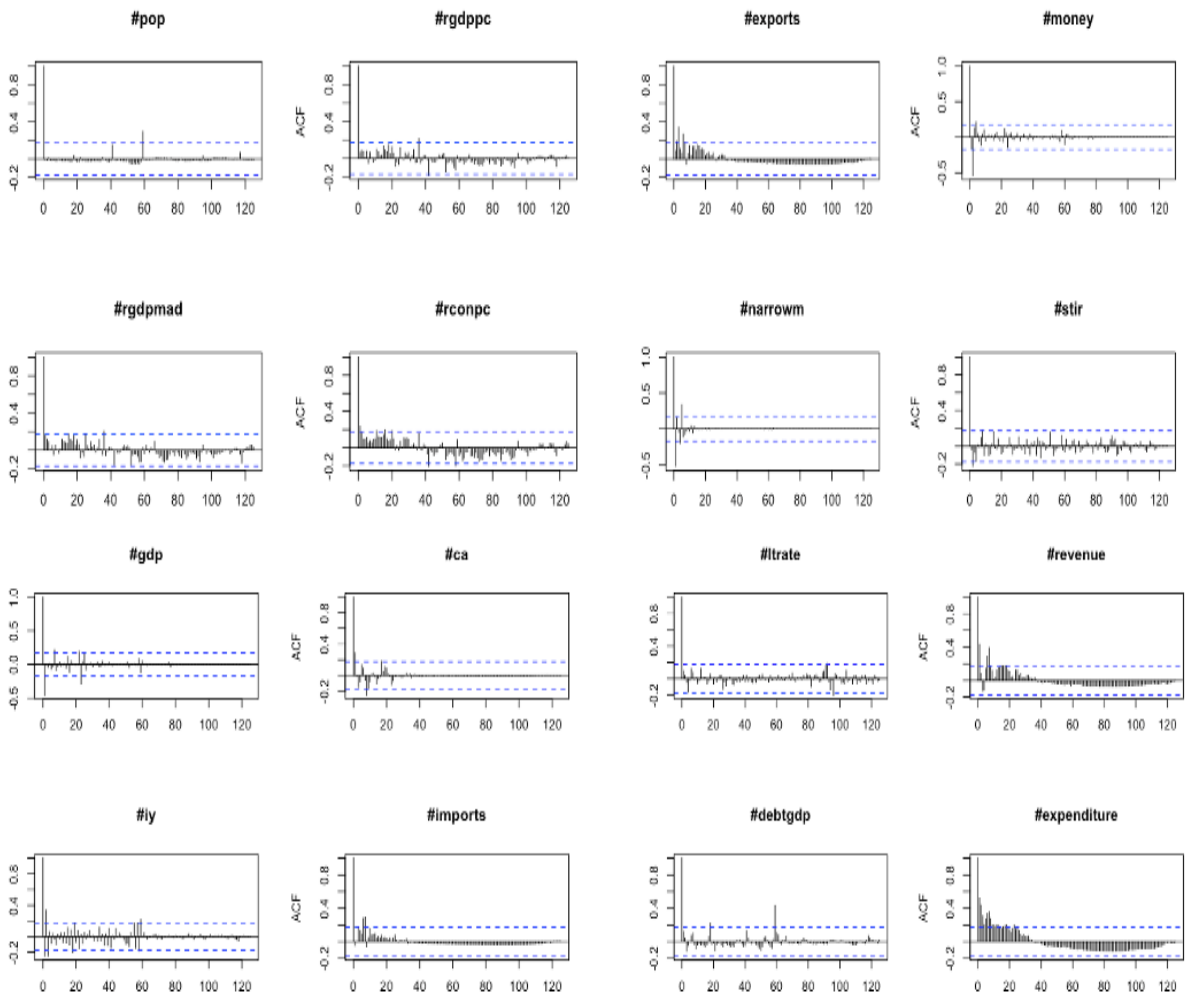


Figure 5: The figure shows the ACF plots for the different series.The lags are observed to decay to zero pointing to stationarity

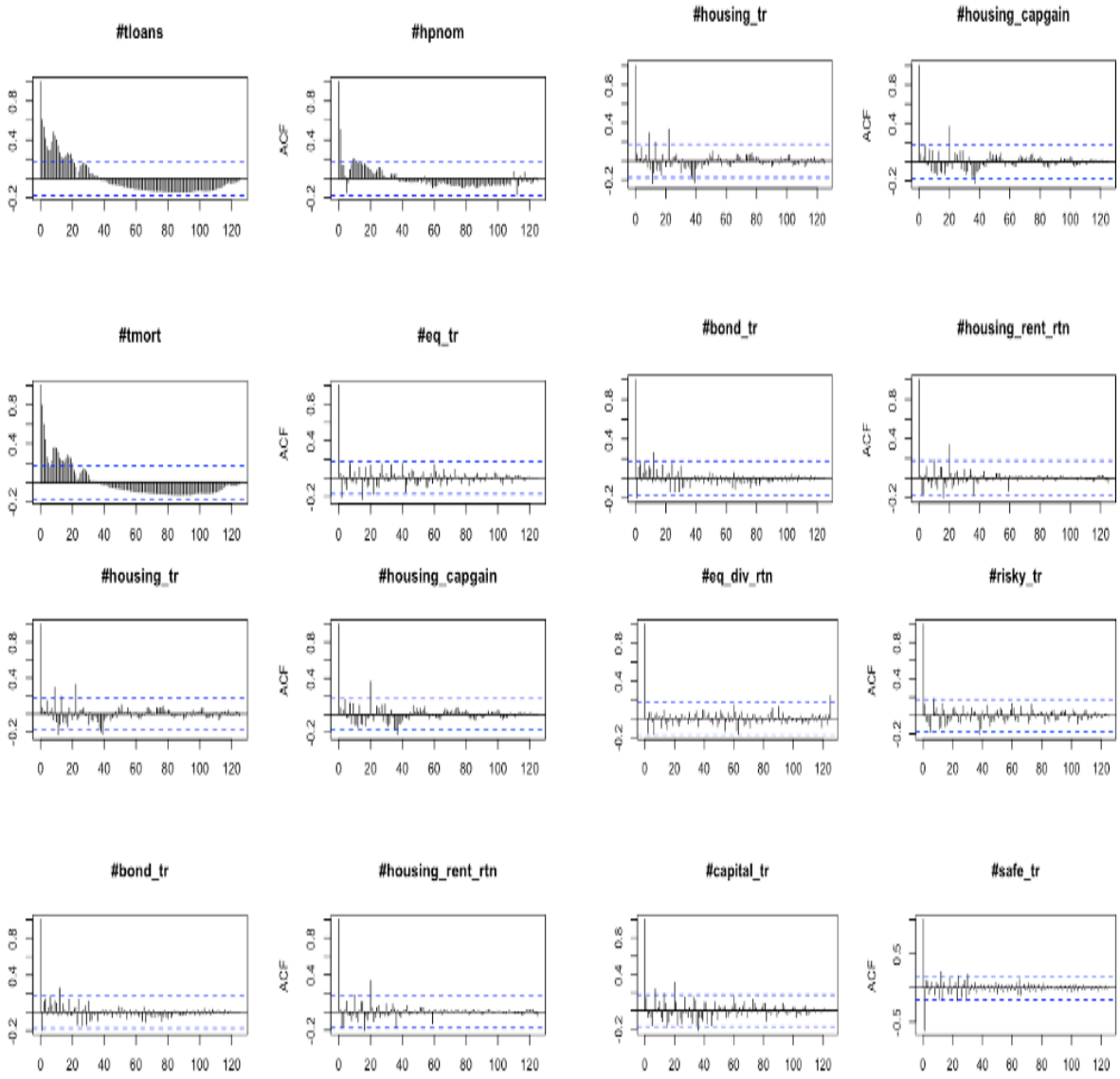


Figure 6: 7.5 continued

8.6 Table 4: Misclassification error for logistic regression and random forest on significant variables from imbalanced data

Model	Australia	Belgium	Denmark	Finland	France	Germany	Italy	Japan	Netherlands	Norway	Portugal	Spain	Sweden	Switzerland	UK	USA
Logistic Regression	0,1	0,13	0,13	0,2	0,22	0,2	0,23	0,21	0,11	0,13	0,12	0,13	0,18	0,18	0,3	0,16
RandomForest	0,03	0,09	0,13	0,11	0,08	0,09	0,2	0,16	0,08	0,11	0,07	0,11	0,09	0,07	0,07	0,11

8.7 Table 5: Misclassification error for random forest before and after SMOTE

Model	Australia	Belgium	Denmark	Finland	France	Germany	Italy	Japan	Netherlands	Norway	Portugal	Spain	Sweden	Switzerland	UK	USA
RandomForest	0,03	0,1	0,13	0,11	0,08	0,1	0,18	0,12	0,09	0,11	0,06	0,11	0,1	0,06	0,06	0,1
RandomForest SMOTE	0,02	0,08	0,12	0,1	0,04	0,07	0,12	0,04	0,07	0,05	0,05	0,05	0,06	0,04	0,03	0,07

8.8 Variable Importance

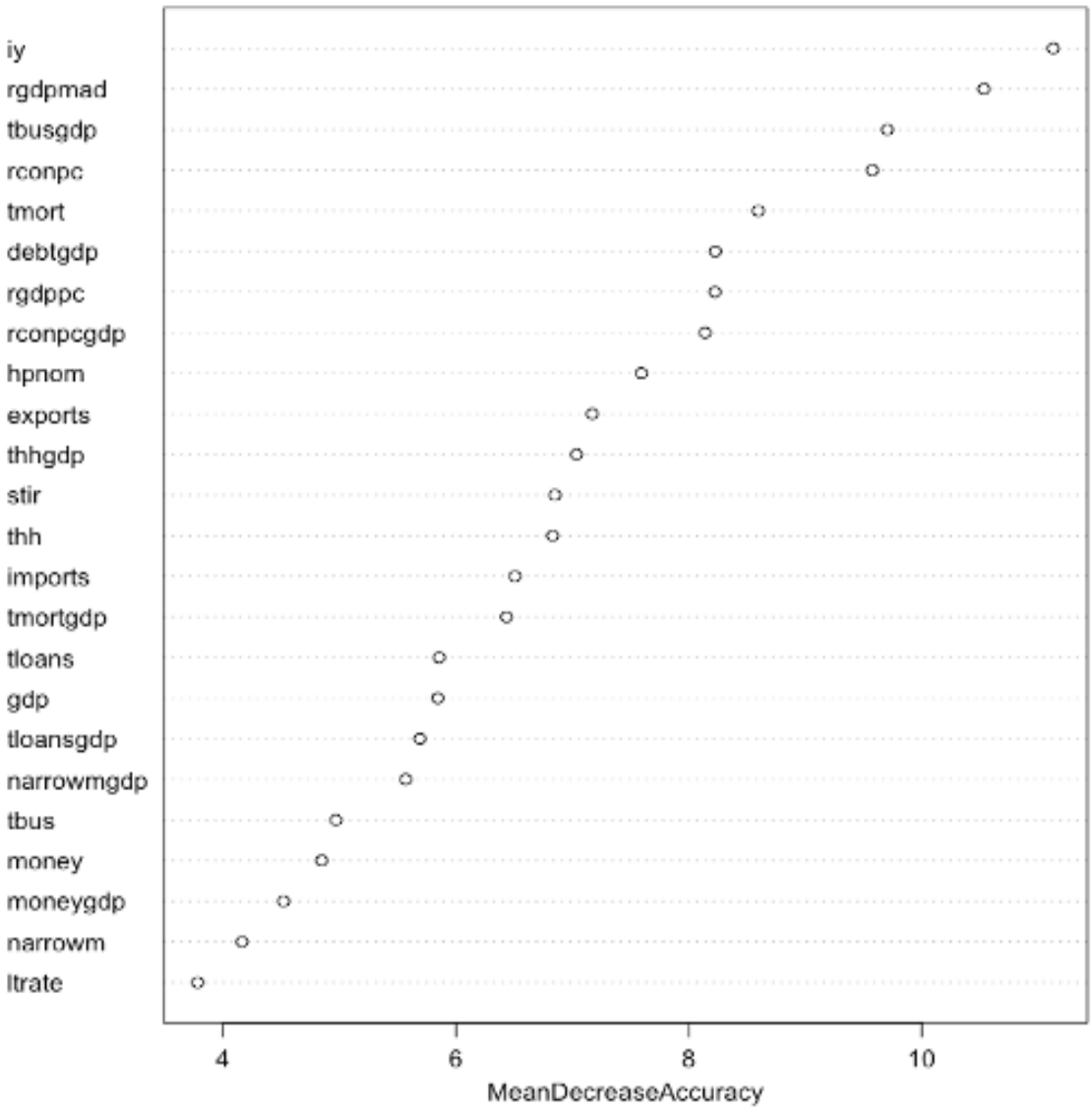


Figure 7: Variable importance - Australia

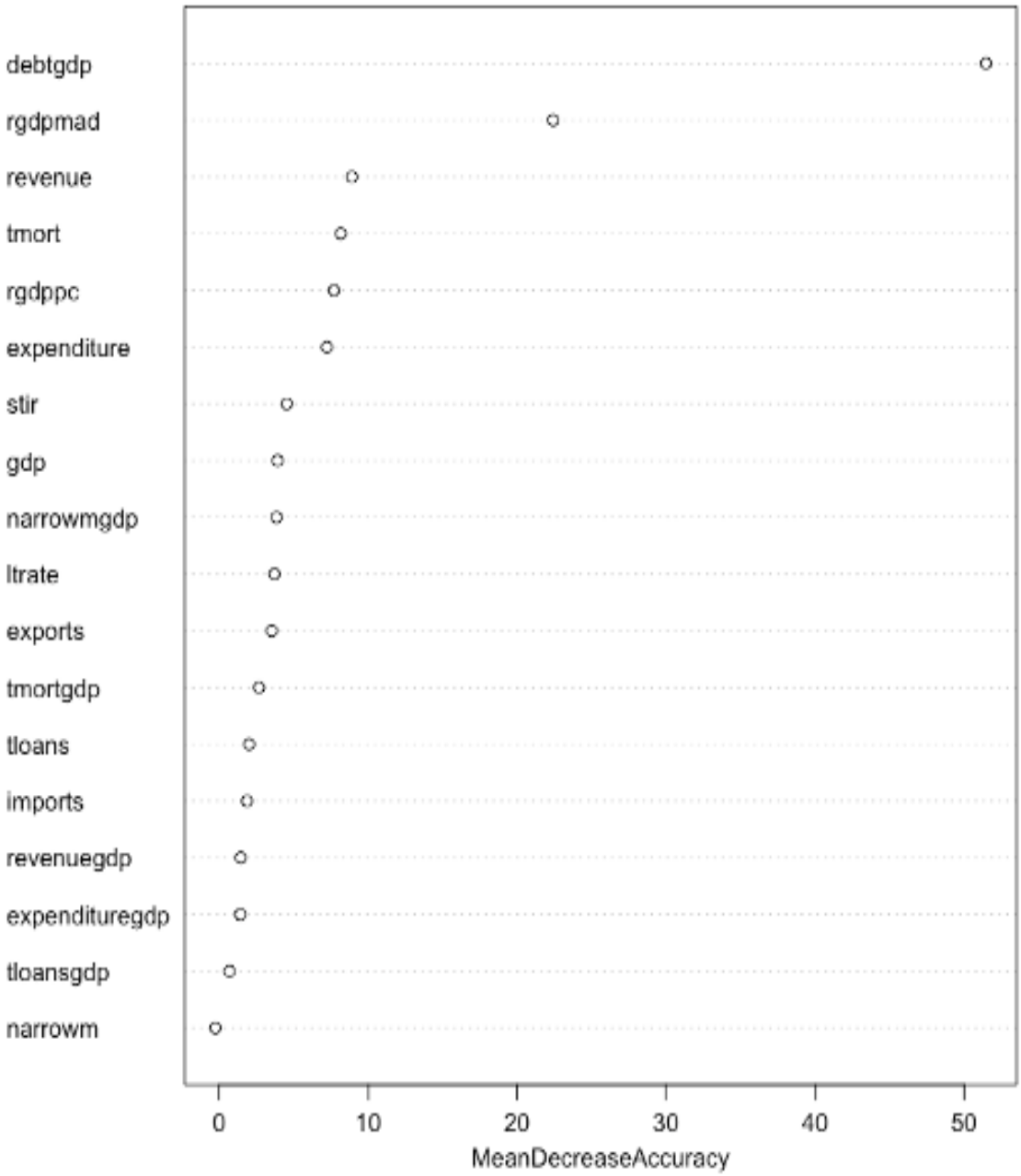


Figure 8: Variable importance - Belgium

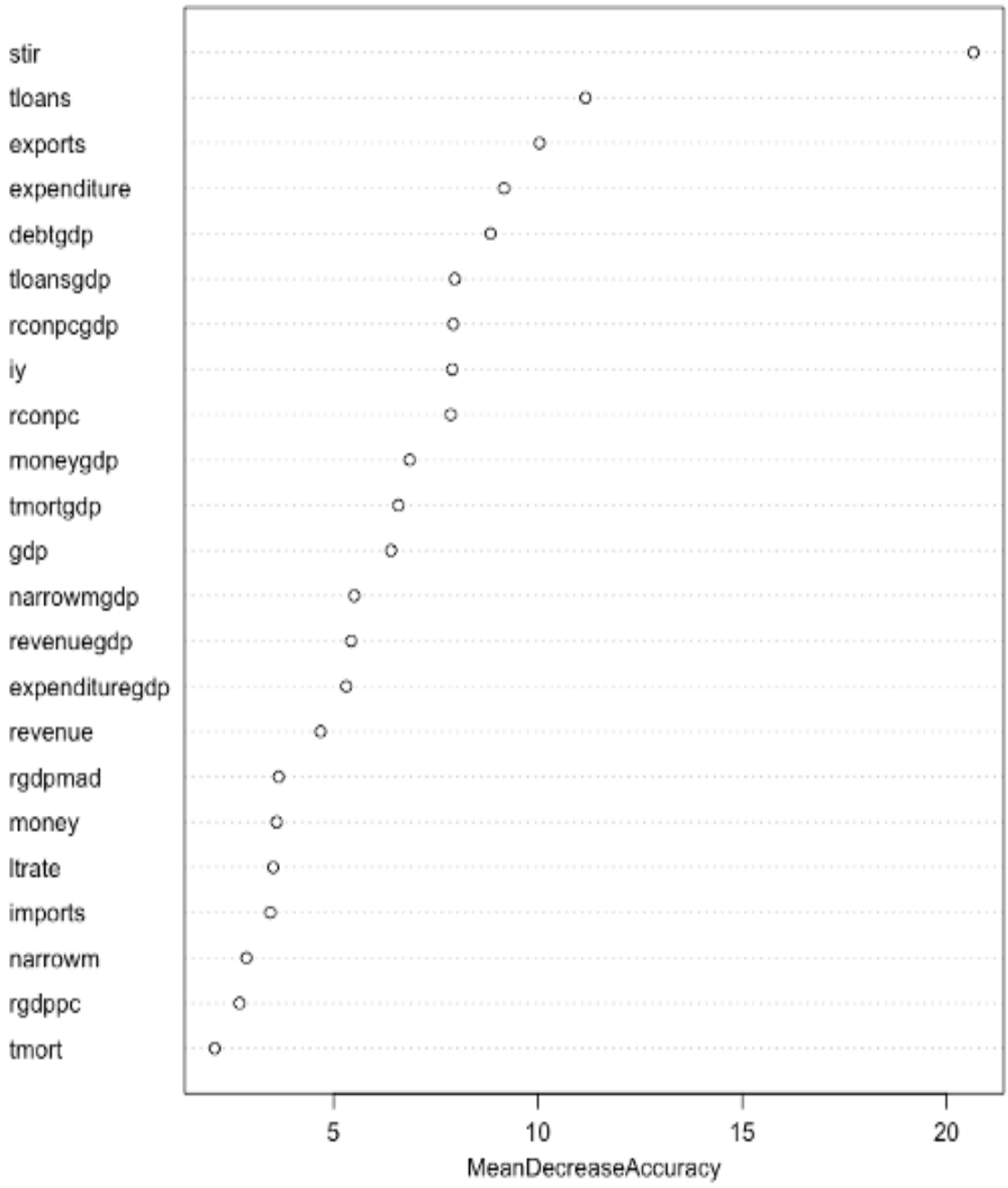


Figure 9: Variable importance - Denmark

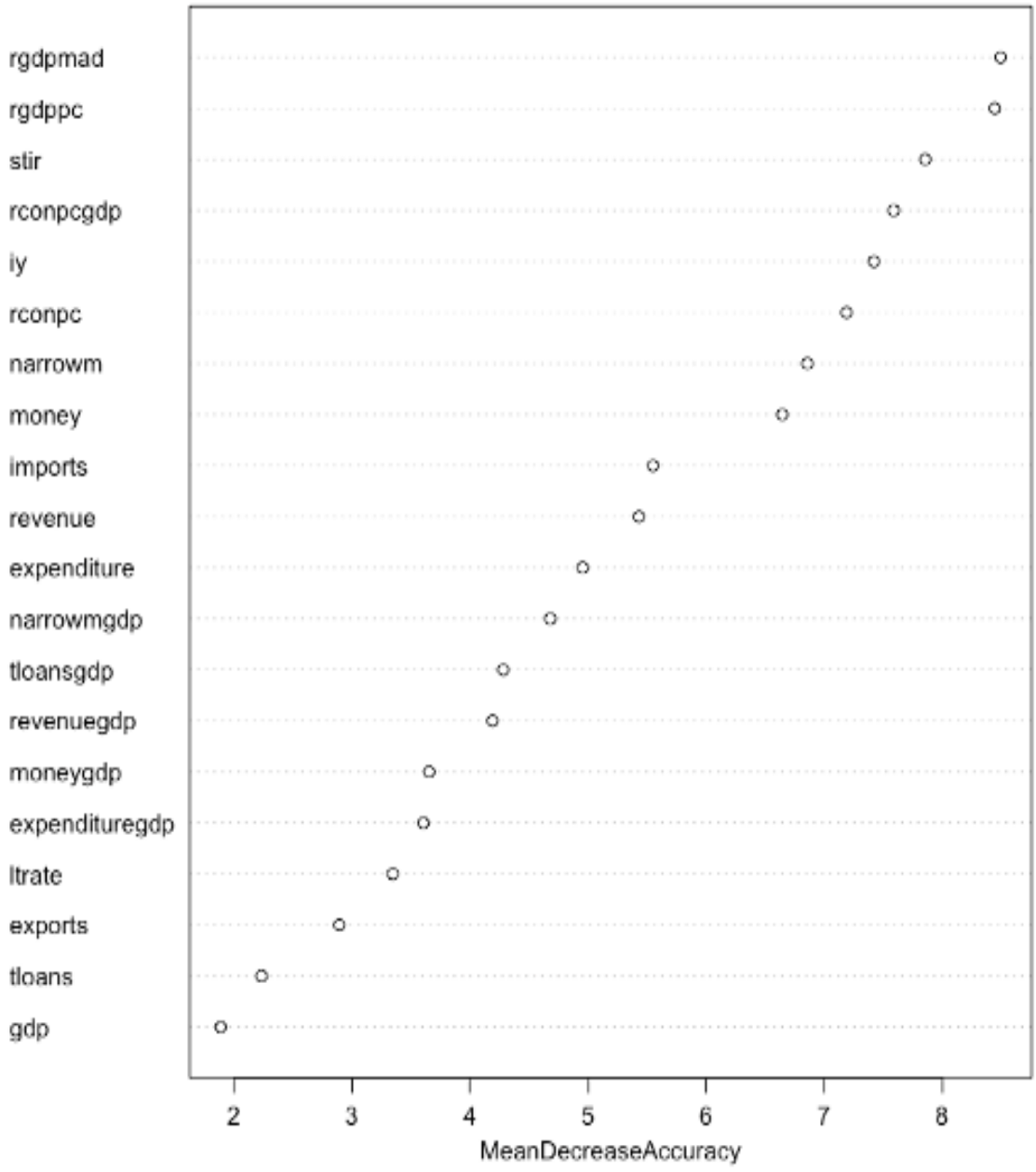


Figure 10: Variable importance - Finland

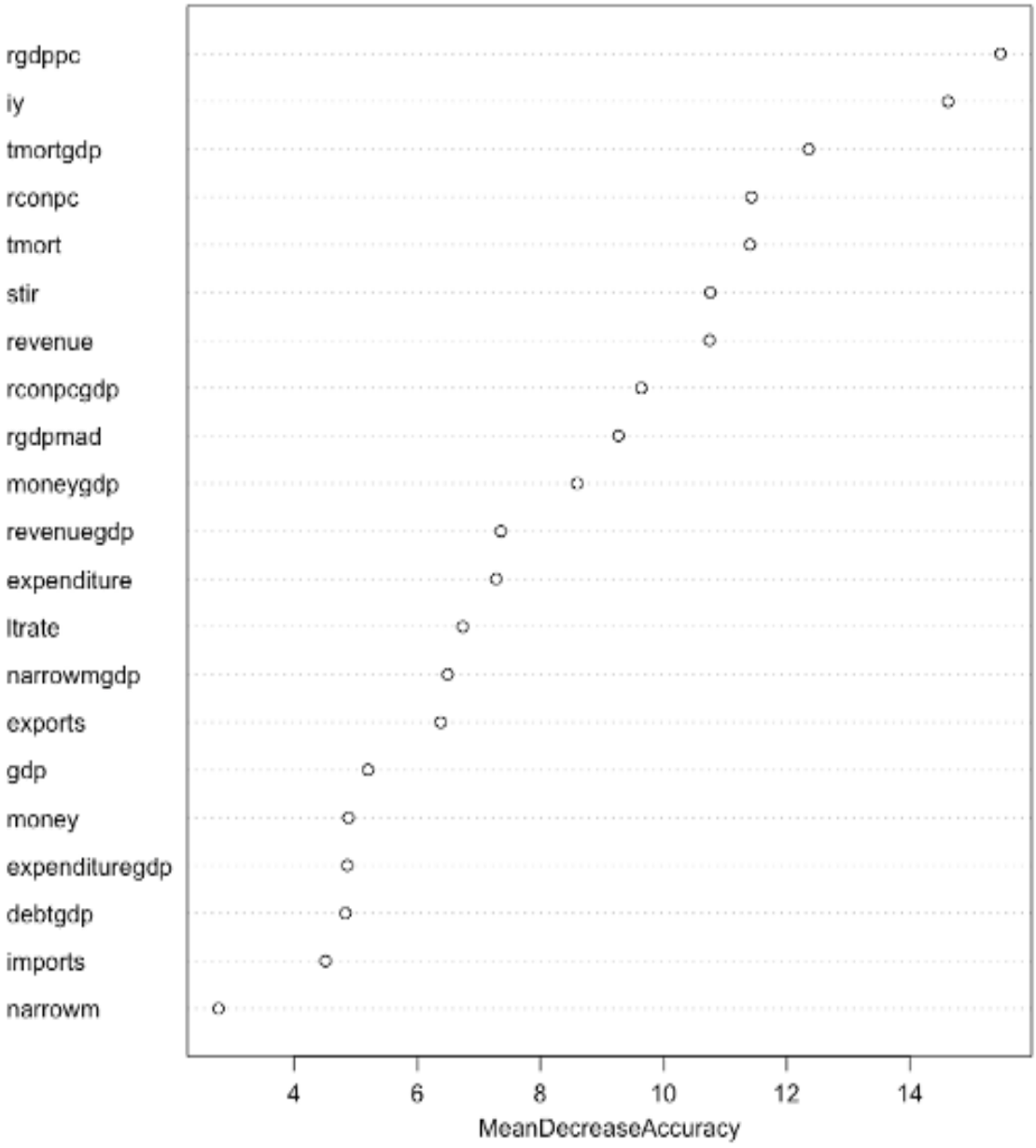


Figure 11: Variable importance - France

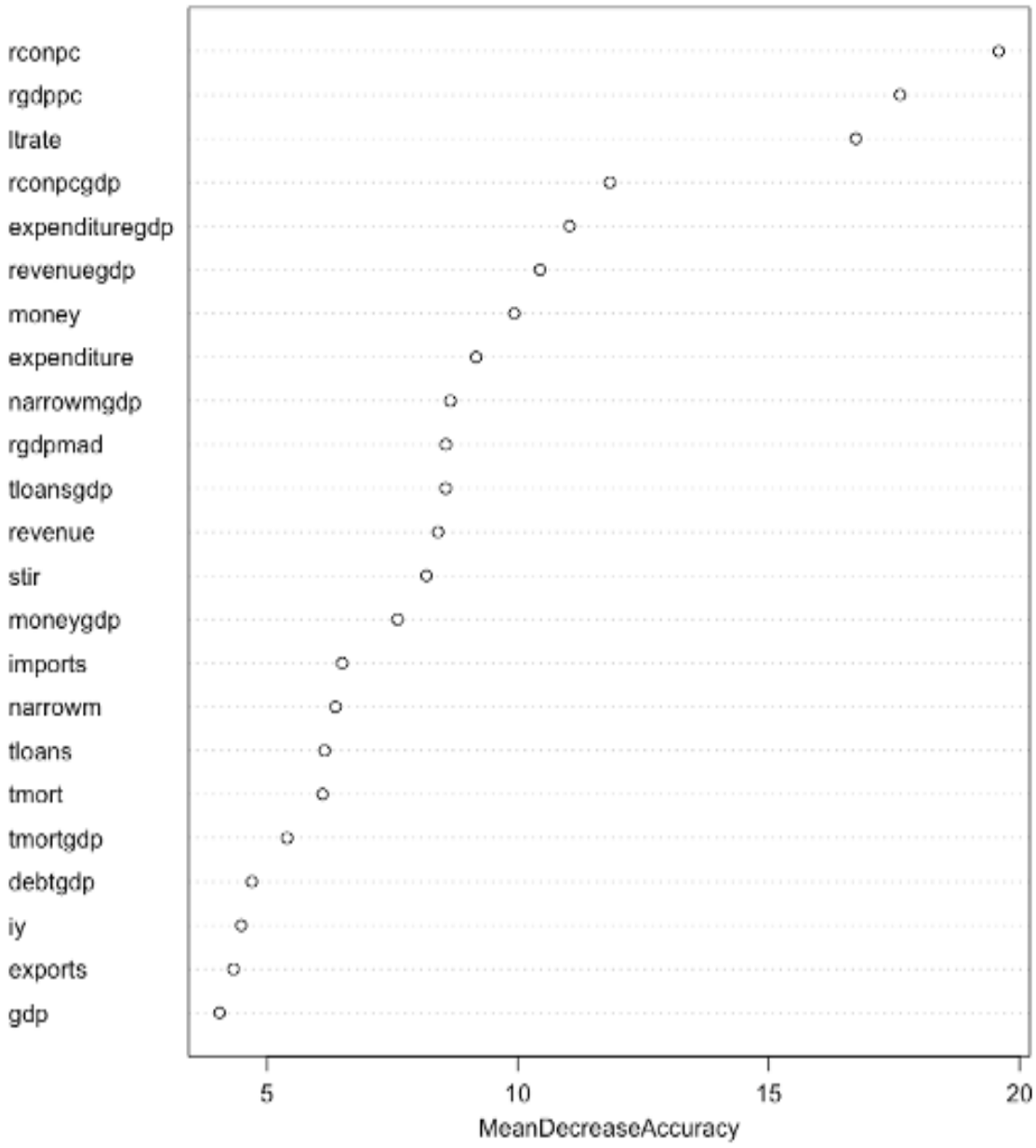


Figure 12: Variable importance - Germany

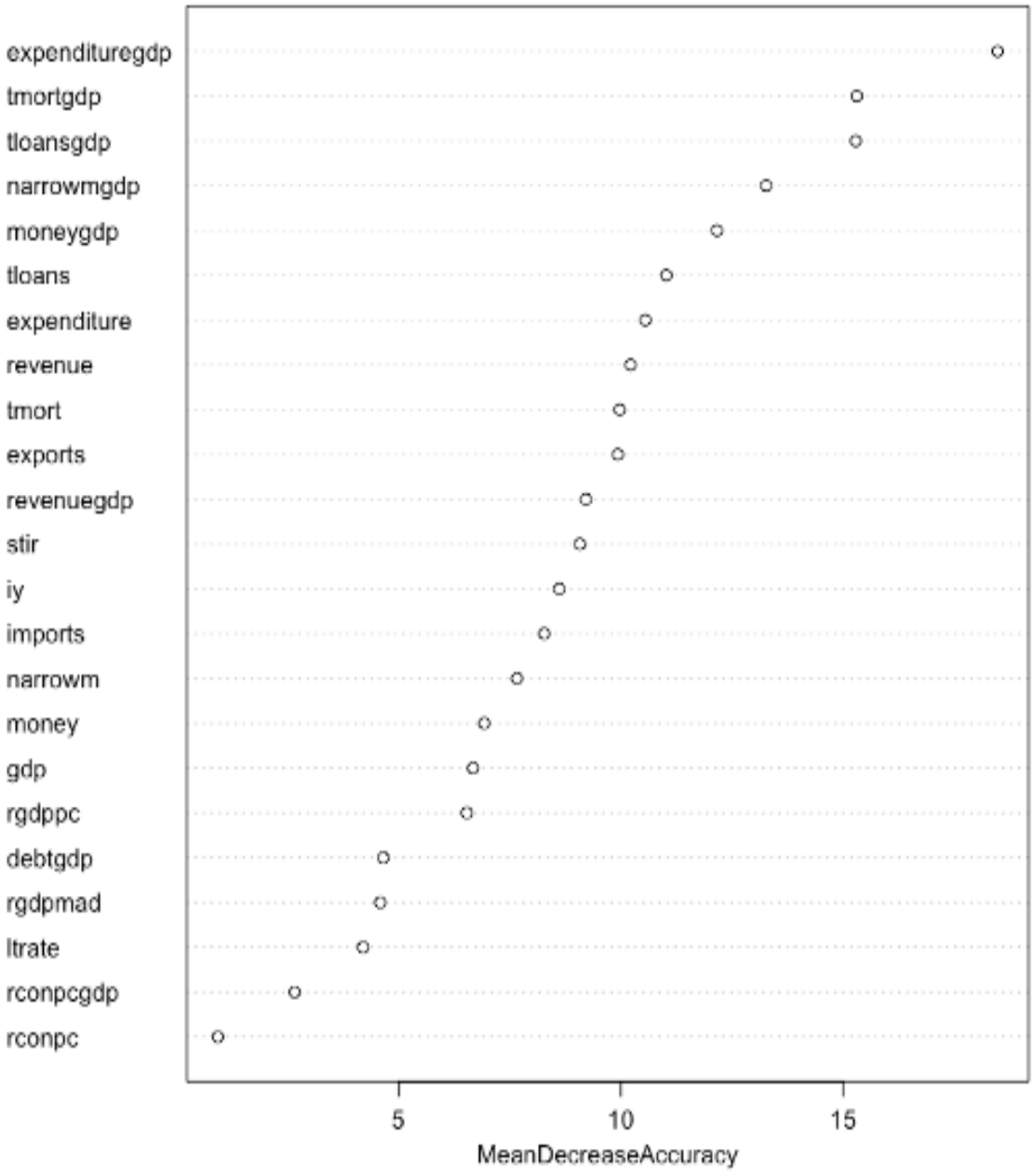


Figure 13: Variable importance - Italy

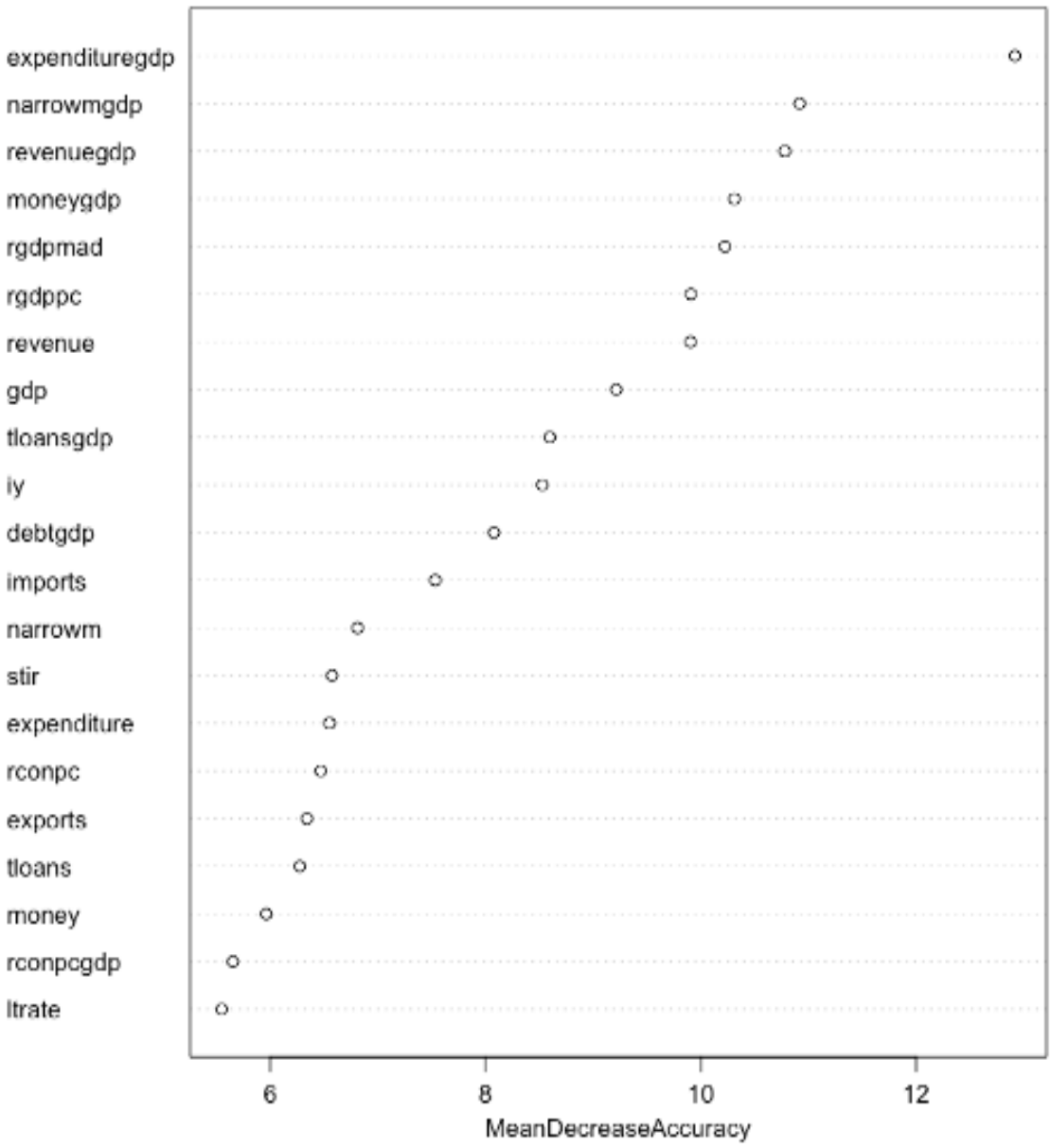


Figure 14: Variable importance - Japan

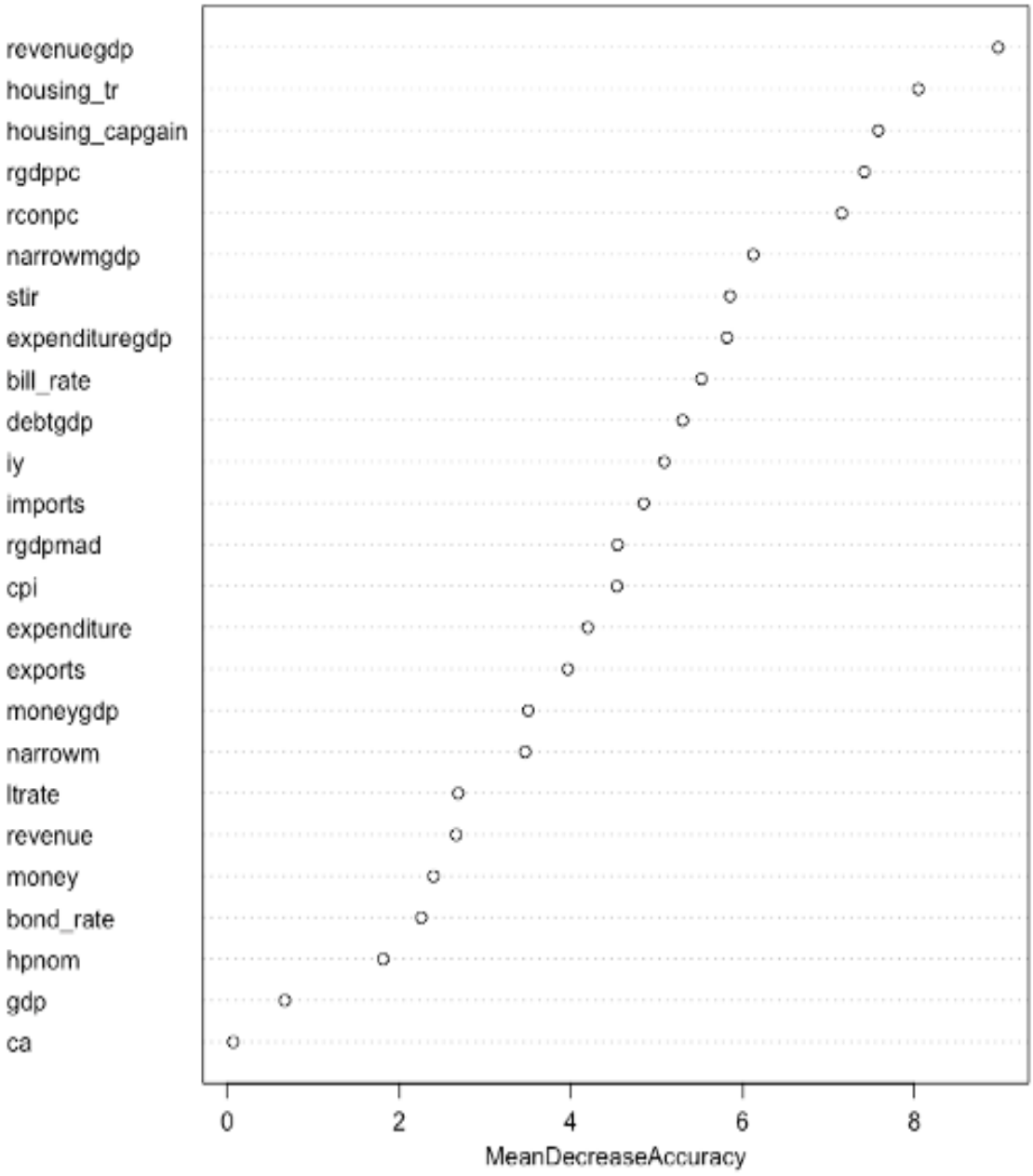


Figure 15: Variable importance - Netherlands

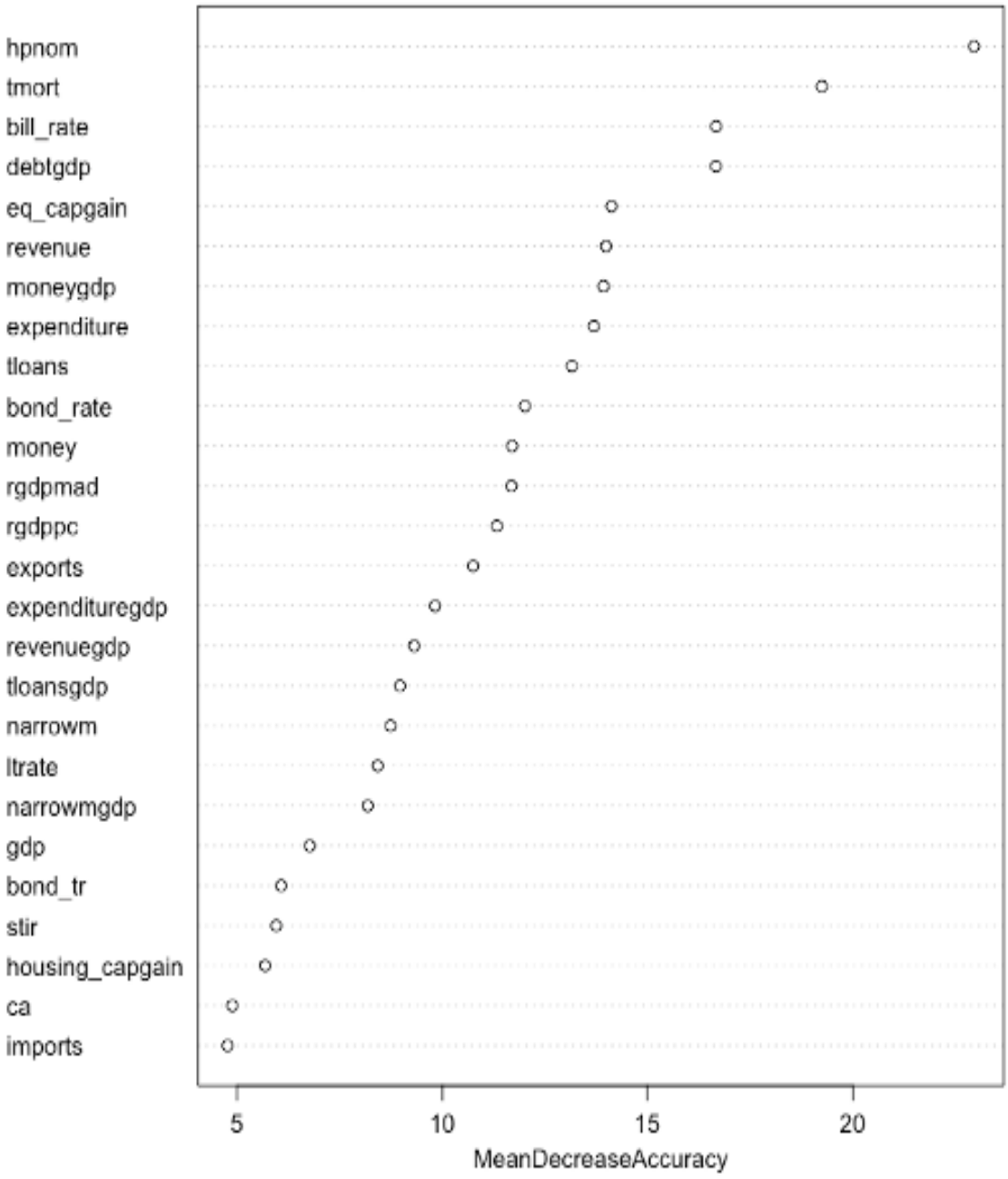


Figure 16: Variable importance - Norway

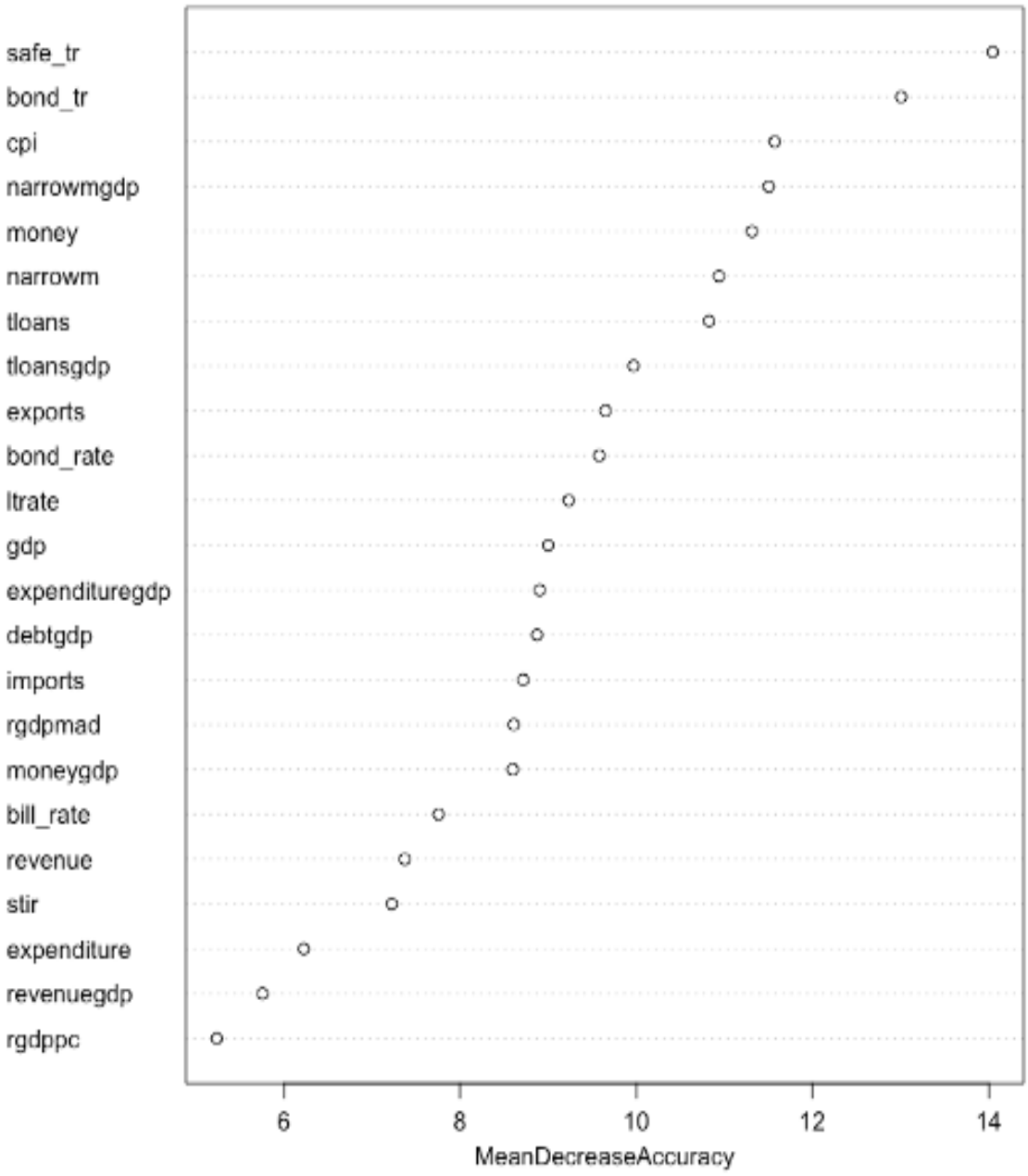


Figure 17: Variable importance - Portugal

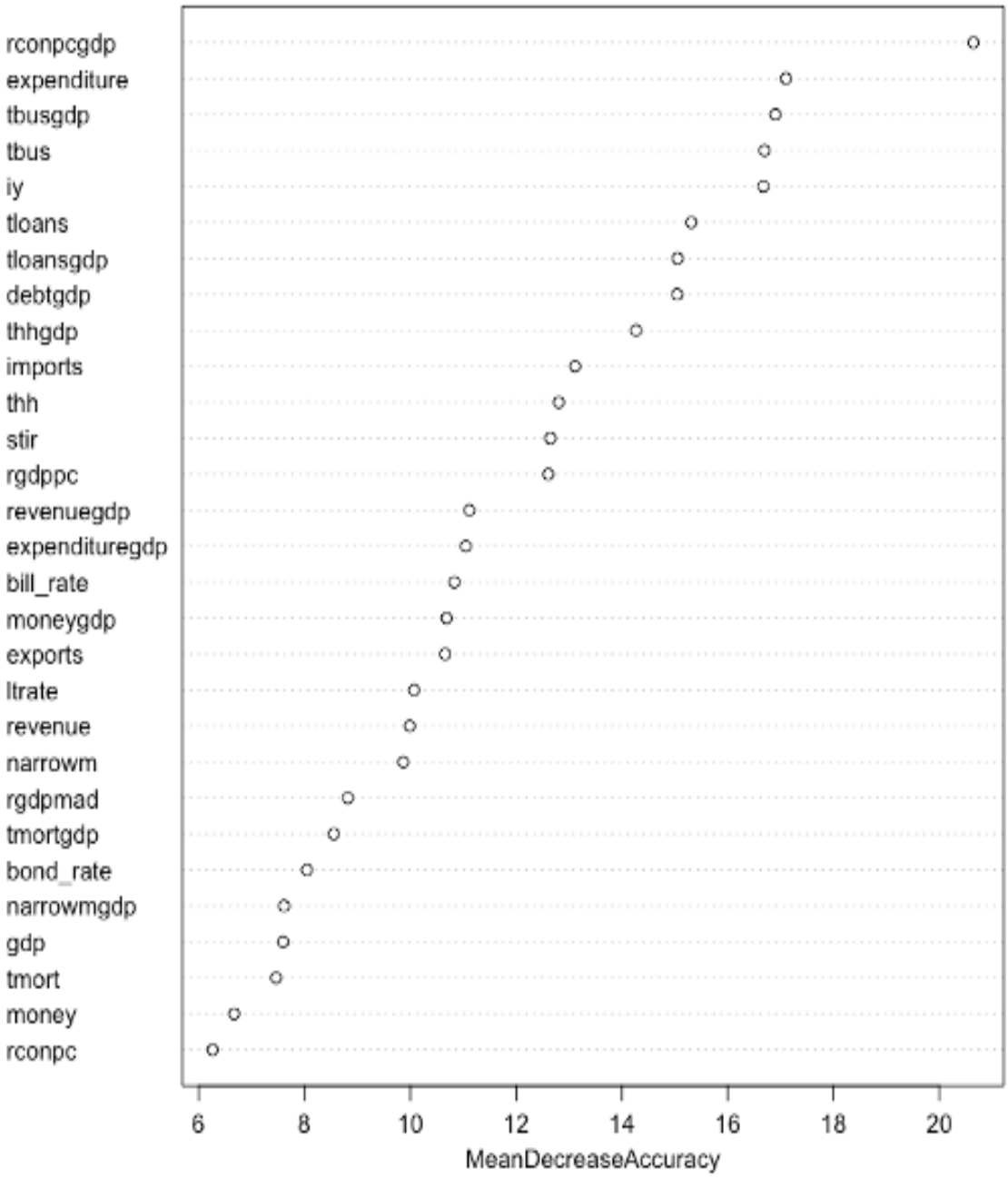


Figure 18: Variable importance - Switzerland

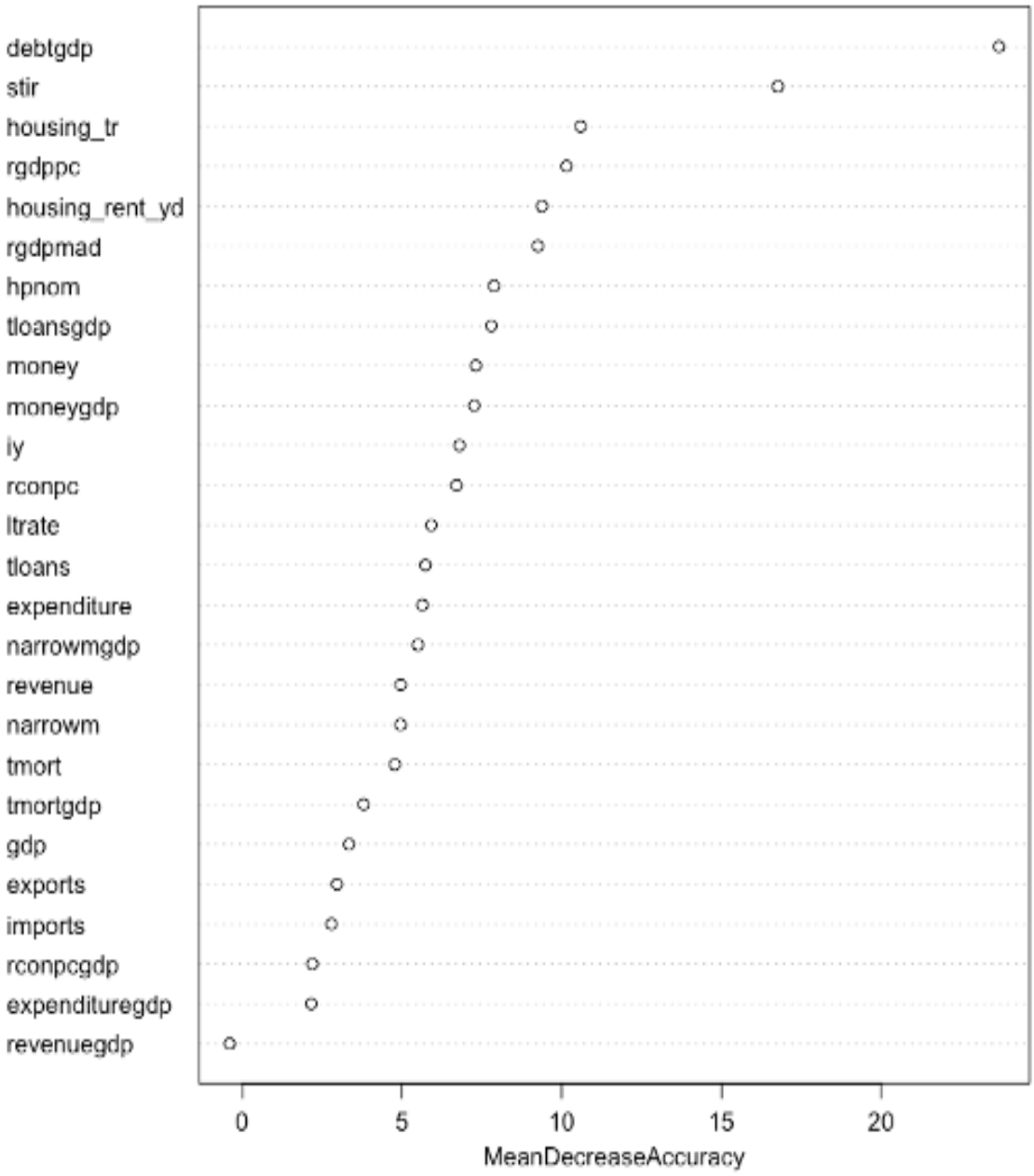


Figure 19: Variable importance - Sweden

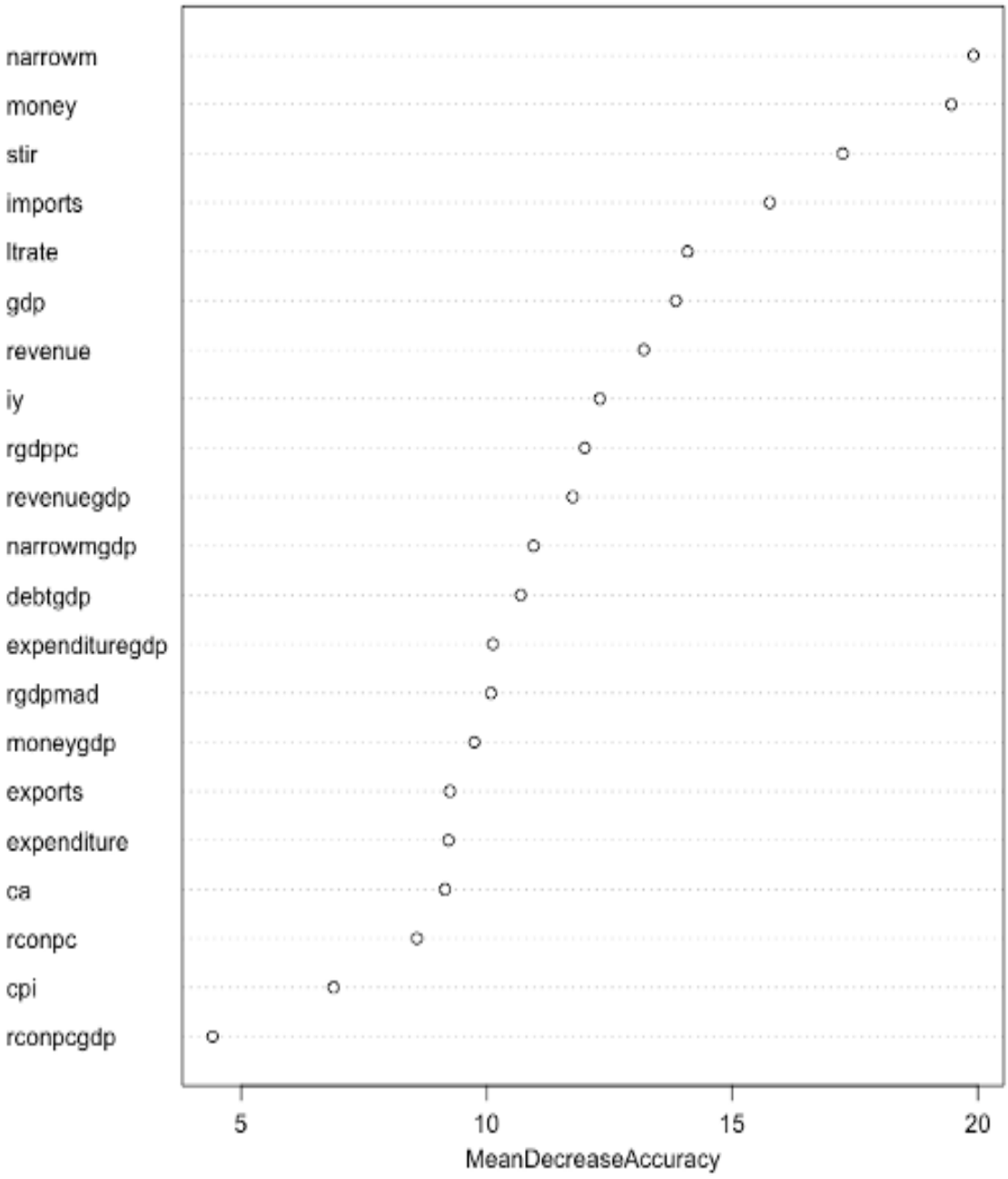


Figure 20: Variable importance - Spain

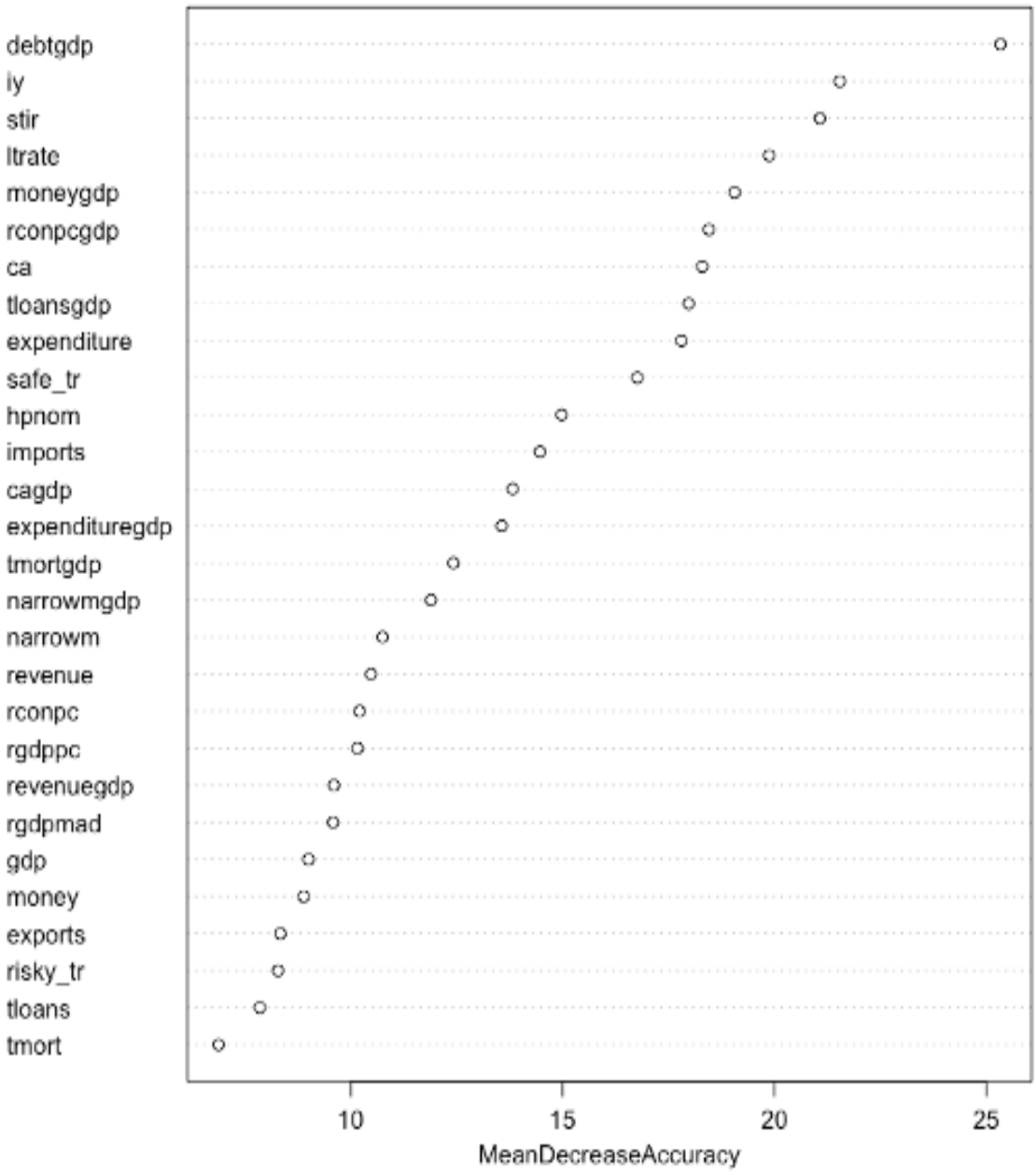


Figure 21: Variable importance - USA

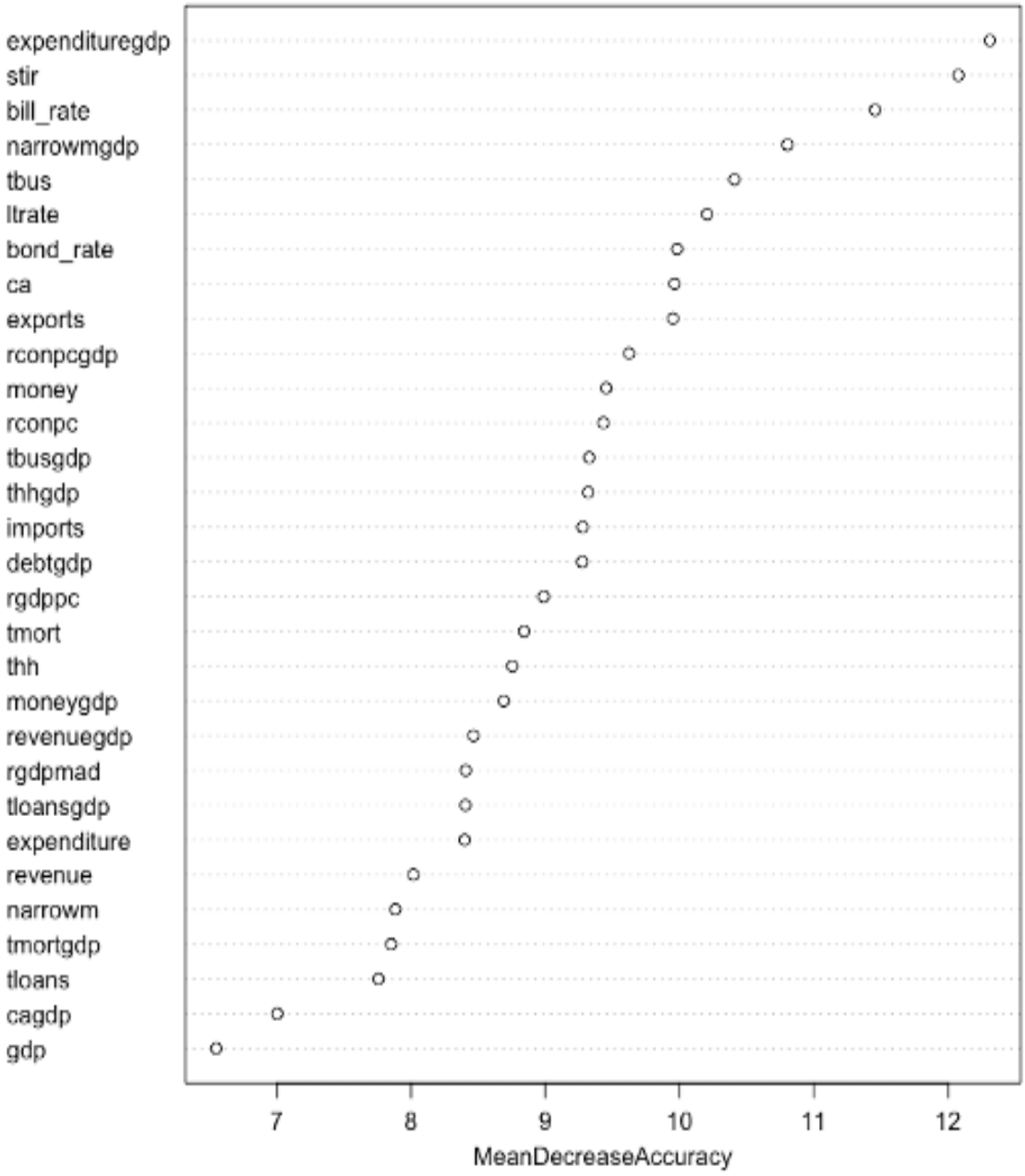
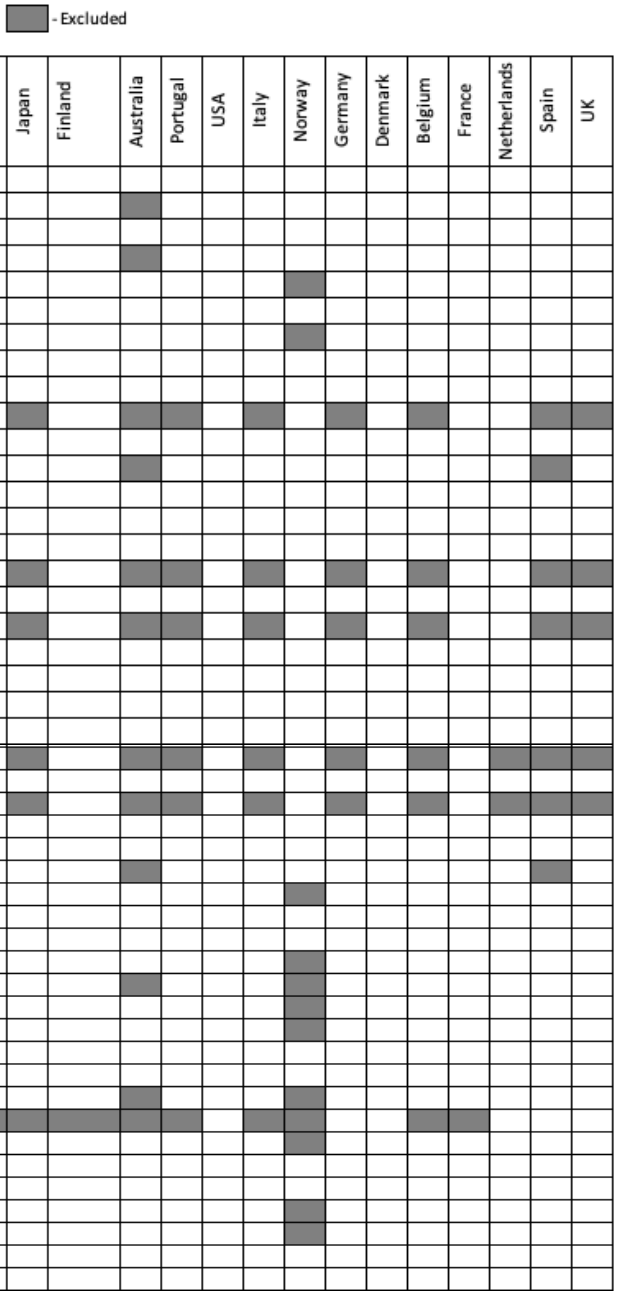


Figure 22: Variable importance - UK

8.9 Table 6: Variables included in each country model



Non-exclusive licence to reproduce thesis and make thesis public

I, Geofrey Wanyama

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Early warning system for financial crisis: application of random forest

supervised by Mustafa Hakan Eratalay and Luca Alfieri

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Geofrey Wanyama
25/05/2020