Modelling Sovereign Credit Ratings: Evaluating the Accuracy and Driving Factors using Machine Learning Techniques*

Bart H.L. Overes†
Erasmus University Rotterdam

Michel van der Wel
Erasmus University Rotterdam, ERIM, Tinbergen Institute, and CREATES

February 1, 2021

Abstract

Sovereign credit ratings summarize the creditworthiness of countries. These ratings have a large influence on the economy and the yields at which governments can issue new debt. This paper investigates the use of a Multilayer Perceptron (MLP), Classification and Regression Trees (CART), and an Ordered Logit (OL) model for the prediction of sovereign credit ratings. We show that MLP is best suited for predicting sovereign credit ratings, with an accuracy of 68%, followed by CART (59%) and OL (33%). Investigation of the determining factors shows that roughly the same explanatory variables are important in all models, with regulatory quality, GDP per capita and unemployment rate as common important variables. Consistent with economic theory, a higher regulatory quality and/or GDP per capita are associated with a higher credit rating, while a higher unemployment rate is associated with a lower credit rating.

Keywords: Sovereign Credit Ratings; Machine Learning; Determining factors; Ordered Logit

JEL classification: G12, C32.

*The authors would like to thank Dick van Dijk, Hüseyin Özturk and Corne Vriends for their comments and suggestions.

†Corresponding author. Mail: `Overes.Bart@gmail.com`.

1. Introduction

A sovereign credit rating is an evaluation of the credit risk of a country and gives an indication of the likelihood that the country will be able to make promised payments. These ratings have a large influence on the interest rate at which governments are able to issue new debt and thereby a big effect on government spending and the government deficit. Sovereign credit ratings are usually given by one of three Credit Rating Agencies (CRAs): Moody's, S&P, and Fitch. These agencies use a combination of objective and subjective factors to determine the rating, however, unfortunately, the exact rating methodology and the determining factors remain unknown. This lack of transparency has resulted in widespread criticism of the CRAs. They have, among other things, been accused of giving biased ratings (Luitel et al., 2016), reacting slowly to changing circumstances (Elkhoury, 2009), and behaving procyclically (Ferri et al., 1999).

Getting an understanding of the rating methodology and the determining factors would be very helpful for governments, investors, and financial institutions. Governments would be able to anticipate possible rating changes, while investors and financial institutions could check if ratings deviate from what the fundamentals of a country imply. In order to get an understanding of the credit rating process, a model is needed that can predict the ratings, ideally with high accuracy. Research has, up until now, mostly focussed on modelling sovereign credit ratings using various forms of the Ordered Probit/Logit (OP/OL) model, which assumes a particular functional form for the relation between a linear combination of the input variables and the continuous output variable, or other related models, see, for example, Cantor and Packer (1996); Dimitrakopoulos and Kolossiatis (2016); Reusens and Croux (2017).[1] These models allow for easy interpretation of the determining factors and prove to be fairly accurate, but come at the cost that the linear relation they assume might not always hold. A recent branch of research has therefore focussed on using Machine Learning (ML) techniques to model sovereign credit ratings (Bennell et al., 2006; Ozturk et al., 2015, 2016). Ozturk et al. (2015, 2016) show that ML models outperform linear models on predictive accuracy, sometimes by a large margin. Especially the Multilayer Perceptron (MLP) and Classification and Regression Trees (CART) prove to be well suited for modelling sovereign credit ratings. However, getting an insight into the inner workings of the models and their determining factors is difficult.

This paper focusses on obtaining the determining factors of two ML models used for sovereign credit ratings; MLP and CART, which has, up until now, not been done for ML

[1] For ease of reference, we will refer to the probit/logit variants simply as linear forms because of the linear relation among variables.

models in the sovereign credit rating setting. This will give insight into the way in which the ML models give the ratings and what variables are important in the process, lack of these insights has been the main weakness of the ML models to date. In order to obtain the determining factors, we use so called Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). SHAP allow for the isolation of each variable's effect and can pick up on non-linear relations, making them well suited for the interpretation of ML models. Getting an understanding of these more accurate models will help figure out the driving factors and methodologies for sovereign credit ratings, as interpreting a model is only useful when that model accurately represents reality. We contrast these approaches to an OL model, this allows for examining how different the insights are for Machine Learning methods compared to a more econometric approach.

This study uses Moody's credit ratings for a set of 62 developed and developing countries, such as Argentina, China, Germany and New Zealand, for the period 2001-2019, to train and evaluate the models. The explanatory variables are similar to those of Dimitrakopoulos and Kolossiatis (2016): GDP growth, inflation, unemployment, current account balance, government balance, government debt, political stability, regulatory quality and GDP per capita.

We document that MLP is the most accurate model for sovereign credit ratings with an accurate rating prediction of 68%, and 86% of ratings correct within 1 notch. Where the percentage within 1 notch indicates what fraction of the ratings given by the model does not deviate more than 1 class from the actual rating. CART follows relatively closely with an accuracy of 59%, and 76% correct within 1 notch. OL, however, significantly underperforms with correct predictions for only 33% of the observations, and 57% within 1 notch. Analysis of the determining factors shows that regulatory quality and GDP per capita are very important explanatory variables in the credit rating process, especially for the MLP and CART. The unemployment rate also proves to be influential in all the models. The relation between these explanatory variables and the credit rating is as expected in all models, with regulatory quality and GDP per capita having a positive influence and unemployment rate having a negative influence.

The structure of our paper is as follows. We begin by discussing the methodology used in this study in Section 2, directly followed by a discussion of the data in Section 3. Section 4 gives an overview of the results obtained in this study. Section 5 concludes.

2. Methodology

In this section, we discuss the methods used in this study, starting off with the modelling techniques. Thereafter, we discuss the so called SHAP values, which allow us to isolate the influence of individual variables in complex models. Lastly, the methods used to evaluate and compare the accuracy of the different models are discussed.

2.1. Modelling techniques

This section gives an overview of the different models. These models are used to predict the sovereign credit rating denoted by y_i for observation i, which represents a rating class, with m being the total number of classes. In this research we use Moody's credit ratings, Moody's gives categorical credit ratings ranging from Aaa (highest) to C (lowest), with 19 categories in between. As algorithms in general cannot handle categorical ratings, they are transformed to numeric ratings from 17 (Aaa) to 1 (Caa1 and lower), where all the ratings of Caa1 and lower have been grouped in $C_{combined}$ because of their infrequent occurrence. The structure of y_i therefore is as follows:

$$y_i = \begin{cases} Aaa\ (17) \\ Aa1\ (16) \\ \quad \vdots \\ C_{combined}\ (1) \end{cases} \tag{1}$$

with the numerical value corresponding to a rating given in brackets. Thus, a high numerical value corresponds to a high credit rating. The explanatory variables are contained in X_i, and n is the total number of observations. Consistent with the main modelling approaches in the literature that we follow (see, e.g, Dimitrakopoulos and Kolossiatis (2016); Ozturk et al. (2015, 2016)), the panel structure of the credit rating data is not taken into account. The main reason for this is that the ML models used in this study do not support panel data, although there are developments in this area.

Multilayer Perceptron

The Multilayer Perceptron (MLP) is a form of an Artificial Neural Network (ANN) which mimics the way that the human brain processes information. MLPs, or similar Neural Network type of algorithms, are often found to perform very well in classification problems involving corporate and sovereign credit rating, see, for example, Baesens et al. (2003); Lessmann et al. (2015) and Ozturk et al. (2015, 2016). A MLP is able to model non-linearities in the data, and can therefore handle very complex classification problems with heterogeneous groups. However, interpretation of the MLP is extremely difficult and, up to a certain degree, it will always remain a "black box".

The MLP consists of an input layer, an output layer and a certain number of hidden layers in between. The input layer contains a number of neurons equal to the number of explanatory variables, here the set of explanatory variables (X_i) are fed into the model. This layer is followed by a certain number of hidden layers, which contain neurons that get an input signal from all the neurons in the previous layer and process that information in order to generate an output that is passed on to every neuron in the next layer. The output layer is the final layer in the MLP structure and has a number of neurons equal to the number of desired outputs, in this case the probability of belonging to each of the credit rating categories in y_i.

The output for each neuron j in hidden layer i is given by

$$h_j^{(i)} = \sigma^{(i)}\left(z_j^{(i)}\right) = \sigma^{(i)}\left(b_j^{(i)} + \sum_{k=1}^{n^{(i-1)}} W_{jk}^{(i-1)} \cdot h_k^{(i-1)}\right), \tag{2}$$

where $b_j^{(i)}$ presents the bias term, $W_{jk}^{(i-1)}$ gives the weight connecting neuron k from layer $i-1$ to neuron j in layer i and $n^{(i-1)}$ is the total number of neurons in layer $i-1$. The activation function $\sigma^{(i)}(z)$ enables the algorithm to model non-linearities that might be present in the data, and can be varied for each layer (Baesens et al., 2003). We use a Rectified Linear Unit (ReLU) function for the hidden layers, given by

$$\sigma(z) = max(0, z), \tag{3}$$

since it is often found to perform best (Ramachandran et al., 2017). For the output layer, we use the Softmax function, given by

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{m} e^{z_k}} \quad \text{for } j = 1, ..., m \text{ and } z = (z_1, ..., z_m), \tag{4}$$

where m is the number of desired output categories. This activation function gives us, for every country, the probability of belonging to each rating class and is therefore well suited for multiclass classification problems. In the end, the estimate for every country $\hat{y}_i$ is set to the numerical class for which it has the highest probability.

The MLP is optimized by minimizing the categorical cross-entropy function, given by

$$C(y_{i,j}, \hat{y}_{i,j}) = -\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{i,j} \cdot ln(\hat{p}_{i,j})\right) \tag{5}$$

where the number of categories in y_i is given by m and the total amount of observations by n. The true value of observation i for class j is given by $y_{i,j}$, which is 1 if observation i belongs to class j and 0 otherwise. The predicted probability that observation i belongs to class j is given by $\hat{p}_{i,j}$. The algorithm is trained by backward propagation of error information through the network. That is, the partial derivative of the cost function with respect to all weights and biases is determined. Thereafter, the weights connecting all the nodes, and the biases are adjusted in such a way that the cost is minimized.

The MLP architecture, that is, the number of hidden layers and neurons per hidden layer is optimized through a grid search. In this grid search, we also determine the optimal dropout rate, which is the fraction of neurons dropped at random to prevent overfitting to the training data. The optimal performance-complexity trade-off for this data set is given by the MLP with 1 hidden layer, 256 neurons and a dropout rate of 0.1. Estimation of this MLP is done using a batch size of 8 and 400 epochs. Details of the grid searches can be found in Appendix A.3. The MLP is implemented in Python's Keras package (Chollet et al., 2015).

Classification and Regression Trees

The idea behind the Classification and Regression Tree (CART) is quite simple, the algorithm finds the optimal splits based on the values of the explanatory variables in order to classify the observations. CARTs have shown to be well suited for credit rating, see, for example Moor et al. (2018) and Ozturk et al. (2015, 2016). A few of the advantages of CARTs are that they can handle outliers, do automatic feature selection and allow for easy interpretation of the model. However, CARTs can be very prone to overfitting.

A CART consists of a root, one or more nodes and several leaves. The first split of the data, based on one of the explanatory variables in X_i, is made at the root, that split leads either to a node, where the remaining data is split further, again based on one of the explanatory variables in X_i, or a leaf, meaning a decisions is made for these observations. Every observation moves through the tree until it ends up at a leaf, which in our case represents one of the different rating categories in y_i.

In this research, we use an algorithm that splits the data in two at every node. The sequential data splits are determined using the Gini method. That is, for each variable the algorithm calculates the weighted average Gini impurity, e.g. how effective the different categories can be separated based on that variable, using the following formula

$$Gini = \sum_{j=1}^{2} \left(\frac{n_j}{n_n} \sum_{i=1}^{m} p(i) * (1 - p(i)) \right), \tag{6}$$

where m is the number of different categories in y_i and $p(i)$ is the probability of picking a data point of class i within that branch of the split. Furthermore, n_j is the number of data points assigned to branch j and n_n gives the total number of data points entering that node. The split that leads to the largest decrease in Gini Impurity is used at that node. This means that the CART is greedy, i.e. it does not care about future splits and does not take them into account.

CARTs are notorious for overfitting, and therefore sometimes need to be restricted. There are two ways of doing this: restricted growth and pruning. In the case of restricted growth, constraints that limit the growth of the tree in certain ways are implemented, which prevent it from overfitting. Whereas with pruning, the tree is left to grow unrestricted and is decreased in size afterwards. Both methods show no improvement on the cross-validated out-of-sample accuracy, and thus an unrestricted CART is used in this study. The details of the CART optimization can be found in Appendix A.4. The CART is implemented in Python's scikit-learn package (Pedregosa et al., 2011).

Ordered Logit

The Ordered Logit (OL) model is, together with the Ordered Probit model, the most frequently used model in literature (Dimitrakopoulos and Kolossiatis, 2016; Afonso et al., 2011; Reusens and Croux, 2017). It therefore provides a good benchmark for the Machine Learning models, because these more complex models are only useful when they are able to outperform the OL model. As opposed to OLS, the OL model can deal with unequal distances between rating classes and the presence of a top and bottom category. Furthermore, the OL model allows for interpretation and significance testing of the explanatory variables' coefficients, which makes it easy to obtain the determining factors.

A pooled OL model is implemented. Here, the latent continuous variable y_i^* has the following specification

$$y_i^* = \alpha + X_i'\beta + \epsilon_i, \tag{7}$$

where the intercept is given by α, X_i contains the explanatory variables for data point i, β is a vector containing the coefficients and the idiosyncratic errors are given by ϵ_i, which has a standard logistic distribution.

However, rating categories are not continuous and our continuous variable therefore needs to be transformed into a categorical rating using

$$y_i = \begin{cases} Aaa\ (17) & \text{if } y_i^* \geq \tau_{16} \\ Aa1\ (16) & \text{if } \tau_{16} > y_i^* \geq \tau_{15} \\ \vdots & \\ C_{combined}\ (1) & \text{if } \tau_1 > y_i^* \end{cases} \tag{8}$$

where the boundaries between the different classes are given by τ_j. The OL model is implemented in Python's Mord package (Pedregosa-Izquierdo, 2015).

2.2. SHAP values

Getting insight into the inner workings of complex models is difficult. Therefore, Lundberg and Lee (2017) came up with a method to approximate the effects that the individual explanatory variables have on the model outcome, called SHAP. This method, based on Shapley values (Shapley, 1953), evaluates how model outcomes differ from the baseline by tuning all the explanatory variables individually, or in combination with a selection of other explanatory variables, while keeping the others constant.

The basic framework for explaining a model $f(x)$ using SHAP values is the explanation model

$$g(x) = \phi_0 + \sum_{i=1}^{n_{input}} \phi_i x_i, \tag{9}$$

where x is a vector containing all the explanatory variables, ϕ_0 is the baseline prediction, ϕ_i is the weight of the i^{th} explanatory variable in the final prediction, and n_{input} is the total number of explanatory variables. The explanation model $g(x)$ gives an approximation of the output of the real model $f(x)$ by using a linear combination of the input variables and a baseline prediction. Calculating the contribution of each variable x_i to the explanation model $g(x)$ is done using

$$\phi_i(f(x), x) = \sum_{v \subseteq x} \underbrace{\frac{|v|!(n_{input} - |v| - 1)!}{n_{input}!}}_{no.\ of\ permutations} \underbrace{(f_x(v) - f_x(v \backslash i))}_{contribution\ of\ i}, \tag{10}$$

where, $v \subseteq x$ represents all the possible v vectors where the non-zero elements are a combination of the non-zero elements in x, $f_x(v \backslash i)$ is the model output of the original model with the i^{th} element of v set to zero, and $|v|$ gives the total number of non-zero elements in v (Lundberg and Lee, 2017). The SHAP values are now given by the solution to Equation 10 that satisfies

$$f_x(v) = E[f(v)|v_S], \tag{11}$$

where S represent the set of non-zero indices in v. This constraint ensures that the SHAP values do not violate the consistency and/or the local accuracy properties, for more information see Lundberg and Lee (2017). Thus, in the end, we get a specific contribution of each explanatory variable to the prediction for every individual observation considered. As the OL model allows for easy interpretation through its coefficients, this method is only used for the MLP and CART. For the SHAP values Python's SHAP package is used (Lundberg and Lee, 2017).

2.3. Model evaluation

Following common practice in literature (Ozturk et al., 2015; Reusens and Croux, 2017), for each model, we determine what percentage of the predictions was exactly right, 1 or 2 notch(es) too high and 1 or 2 notch(es) too low. Where a credit rating prediction is said to be u notch(es) too low (high) if the predicted class is the u class(es) below (above) the actual rating class.

Predictions are made using random split 10-fold cross-validation. That is, the data is split into 10 approximately equal subsets of which 9 are used to train the model, and the subset that was left out is used for evaluation of the out-of-sample predictive accuracy. By rotating the 10-folds, we obtain the out-of-sample accuracy of the model on the entire data set. We use the averages of 100 replications of this procedure, each time using different 10-fold data splits, thus making sure that results are not dependent on one specific random split.

3. Data

We use Moody's' sovereign credit ratings for a variety of 62 developed and developing countries, among which Brazil, Canada, Morocco and Thailand, from 2001 to 2019.[2] A histogram of the ratings with their alphabetical and numerical rating is shown in Figure 1. In this figure, we see that the data set contains a good mixture of the different categories, however, class 17 (Aaa) is, with 286 observations, significantly overrepresented. This is due to the fact that most of the Aaa countries stayed in this category throughout the entire period. There is therefore a trade-off between having enough different countries with an Aaa rating to train on and making sure the share of Aaa ratings does no become too large. The share

[2]Obtained from countryeconomy.com.

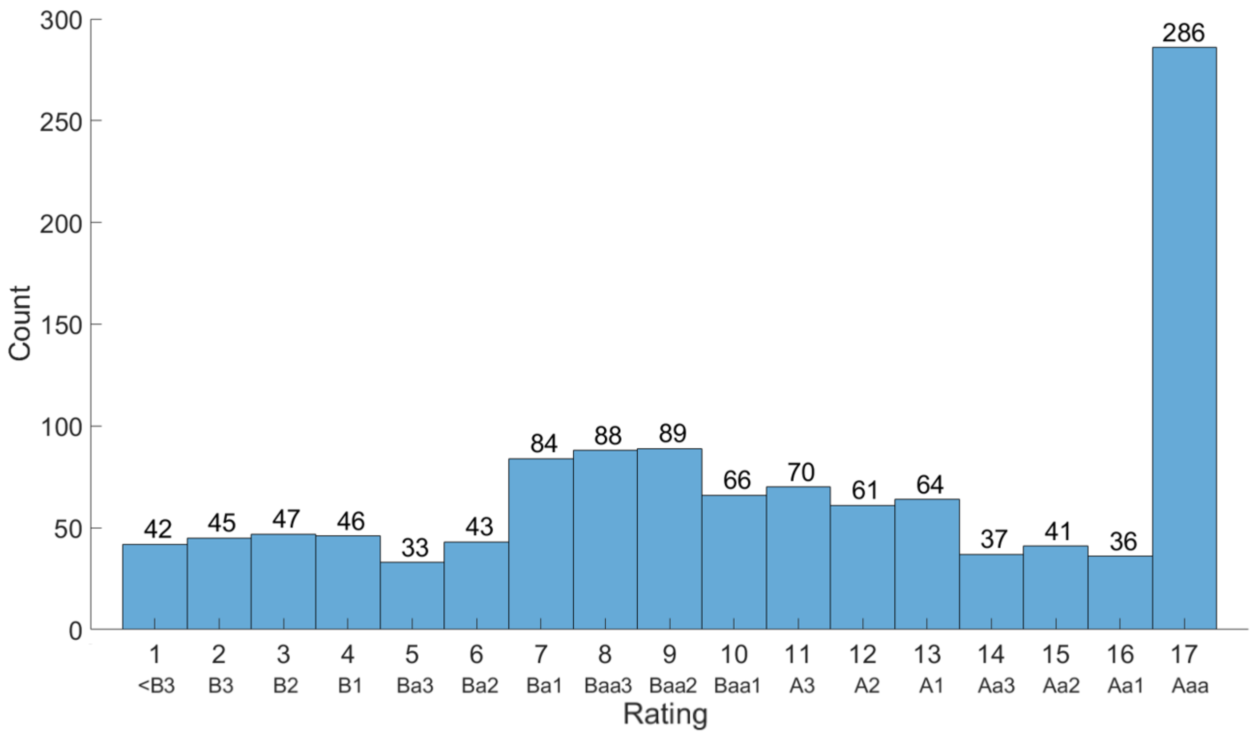

Fig. 1. Histogram of ratings given by Moody's with numerical conversion for the period 2001-2019, ratings as of January 1 of each year are used. Full list of countries included can be found in Appendix A.1. 1178 observations in total.

of other ratings seems to be well balanced, with numeric ratings 7, 8 and 9 appearing a little more often (approximately 85 times). The full list of countries can be found in Appendix A.1, and the transformation from Moody's to numerical ratings in Table 5.

Previous research investigating the determining factors of sovereign credit ratings, mostly using linear regressions or an Ordered Logit/Probit model, shows that there are factors that frequently prove to be important. Cantor and Packer (1996) found that GDP per capita, GDP growth, inflation, external debt, level of economic development, and default history are determining factors in the credit rating process. The importance of these factors is also found by other researchers (Afonso et al., 2011; Ferri et al., 1999; Gaillard, 2009; Dimitrakopoulos and Kolossiatis, 2016), although there are some contrasting outcomes, for example, Ferri et al. (1999) finds GDP per capita to be unimportant. Apart from economic and fiscal indicators, governance indicators prove to be important in the credit rating process, as was shown by Ozturk (2014). Other factors that also commonly proved to be influential in these studies are: government balance, current account balance, and government effectiveness. We have included all these factors in our study, except for default history, because very few sovereign defaults have occurred during the period and estimating its effect would thus be very difficult.

In accordance with the above mentioned literature, we use a combination of economic figures, fiscal indicators, and governance indices as explanatory variables. These variables are: unemployment rate[3](measured in %, with expected sign –), government balance[3] (% of GDP, +), current account balance[3] (% of GDP, +), inflation as measured by CPI[3] (%, –), GDP per capita[3] ($, +), government debt[3] (% of GDP, –), GDP growth[4](annual %, +), regulatory quality index[4] (+), and political stability and absence of violence/terrorism index[4] (+). Where the variables up to and including GDP growth are economic & fiscal indicators, while the latter two are measures of governance. The regulatory quality index measures perceptions of the government's ability to formulate and implement policies and regulations that permit and promote private sector development. While the political stability and absence of violence/terrorism index captures the perceptions of likelihood of political instability and/or politically-motivated violence. Both these variables have values ranging from approximately -2.5 to 2.5, where a higher score indicates better regulatory quality or higher political stability. The values of all the explanatory variables for year t are used to model Moody's sovereign credit ratings as of January 1 of year $t + 1$.

This set of explanatory variables represents three main factors in the credit rating process: the strength of the economy, the level of debt, and the willingness to repay. A strong

[3]Obtained from the International Monetary Fund.

[4]Obtained from the World Bank.

Table 1: Descriptive statistics of the macroeconomic and socioeconomic variables for the period 2000-2018, 1178 observations. Data obtained from the IMF and the World Bank. Full list of countries included can be found in Appendix A.1.

Variable	Med.	Mean	Std	1%	99%
GDP growth (%)	3.2	3.2	3.4	-7.4	10.9
Inflation (%)	2.7	60.1	1904.7	-1.1	31.1
Unemployment rate (%)	6.9	7.9	4.6	1.5	25.2
Current acc. (% of GDP)	-0.4	9.6	50.2	-91.0	236.5
Gov. balance (% of GDP)	-2.2	-1.9	4.2	-11.6	12.0
Gov. debt (% of GDP)	46.9	55.0	33.8	7.4	182.6
Political stability	0.4	0.3	0.9	-2.3	1.6
Regulatory quality	0.8	0.7	0.8	-1.3	2.0
GDP per capita (1000$)	13.5	22.1	22.0	0.9	101.8

economy is expected to be better capable of repaying its debt and preventing the debt burden to get out of control. Typical for strong economies are: a low unemployment rate, low (though not negative) and stable inflation, a high GDP per capita, and a high GDP growth. The debt position of a country is given by the government debt, and the government balance shows if the total debt sum (in $) is increasing or decreasing. Finally, regulatory quality and political stability can give an indication of the willingness of a country to repay their debt, but also of the economic climate for the private sector in a country.

The descriptive statistics for the explanatory variables are shown in Table 1. This table reports the median, mean, standard deviation, and 1% & 99% percentiles of all variables. We immediately observe that the average country experienced an economic expansion during this period, as can be seen from the positive average GDP growth of 3.2%. Furthermore, we observe that the average unemployment rate for the period is 7.9%, but very low unemployment rates of below 2.0% are also observed, for example for Thailand and Singapore. The average government debt is 55.0% for the period, with a small number of countries, such as Japan and Venezuela, having a debt of over 180%, which can be considered extremely large. Lastly, the gap between the very rich and very poor countries is large, with some of the rich countries (Luxembourg, Norway) having a GDP per capita that is up to 80 times higher than some of the poor countries (Honduras, Pakistan).

4. Results

In this section, we present the results obtained in this study. First, we discuss the accuracies of the different models when evaluated using cross-validation. Second, the determining factors for each model are analysed individually, and compared to those of the other models.

4.1. Cross-validated accuracy

The accuracies of the MLP, CART, and OL, determined using 100 replications of 10-fold random cross-validation, are shown in Table 2. In this table, for every model, we present the percentage of predictions exactly right, 1 or 2 notch(es) too high or too low, the number of predictions correct within 1 and 2 notch(es), and the Mean Absolute Error (MAE).

The MLP performs best with an accuracy of 68.3%, and 85.7% of predictions correct within 1 notch. MLP outperforms CART and OL, with respective accuracies of 58.6% and 33.1%, significantly on a 99% significance level. CART outperforms the OL model significantly and is, based on performance, much closer to the MLP than to the OL model. These results confirm earlier findings that MLP and CART outperform linear models based on accuracy, see, for example, Bennell et al. (2006); Moor et al. (2018); Ozturk et al. (2015, 2016).

A nice symmetry in over- and underrating is observed for all models. This shows that none of the models has a tendency to consistently rate higher or lower than Moody's. Additional related results, available upon request, show that no country is persistently under- or overrated by MLP and CART. OL, on the other hand, has that tendency and, for example, continuously underrates France and Belgium, and overrates Bulgaria and Cyprus compared to Moody's.

There are multiple possible causes for the relatively large difference in accuracy between the ML techniques and the OL model. First, ML techniques are able to pick up on non-linear relations, where the OL model with its assumption of linear relations cannot. Research has shown that there are non-linear effects in the sovereign credit rating process, so assuming linear relations is likely to harm performance (Reusens and Croux, 2016). Second, the ML techniques have more modelling freedom to pick up on subjective factors of the CRAs, which Moor et al. (2018) show to be especially large for low-rated countries.

4.2. Determining factors

In order to get an insight into the sovereign credit ratings, we analyse the determining factors for every model. We obtain the determining factors of the MLP and CART by using SHAP values, as discussed in Section 2.2, and those of the OL model by looking at each variables' coefficients and their significance.

Table 2: Averages of 100 replications of 10-fold cross-validated predictions for MLP, CART, and OL. All numbers, except for MAE, given in %.

	Correct prediction percentage							
	2 below	1 below	Exact	1 above	2 above	Within 1	Within 2	MAE
MLP	3.9	8.4	68.3	9.0	3.6	85.7	93.2	0.64
CART	5.6	8.7	58.6	9.1	5.2	76.4	87.2	1.00
OL	9.8	10.3	33.1	13.2	10.3	56.6	76.7	1.60

Multilayer Perceptron

SHAP values are calculated for every variable used in the MLP to isolate their effects, and are shown in Figure 2. We immediately observe clear patterns for the regulatory quality and GDP per capita, the most important and second most important variable respectively. A higher value for either variable is associated with an increase in the credit rating, which is in line with economic theory. The importance of regulatory quality is perhaps surprising, since one would expect financial indicators to be most important in an assessment of credit risk. However, regulatory quality might be the best indicator of the economic climate for the private sector in a country, which in turn might be the most relevant factor in separating creditworthy from non-creditworthy countries. Furthermore, regulatory quality also gives an indication of the willingness to pay. That GDP per capita turns out to be an important factor in the credit rating process is not unexpected, since it is a good measure of the relative size of the economy and wealth of a country, and has proven to be important in previous studies (Bissoondoyal-Bheenick, 2005; Gaillard, 2009; Afonso et al., 2011).

The next variable, current account balance, shows a positive influence on the credit rating when the value is either relatively low or relatively high, and a negative influence on the credit rating for an average value. This non-linear relation is also visible in the data, as stronger economies are more towards the extremes. The Netherlands and Germany for example have a very high current account balance, while that of the United Kingdom and Australia is very low. Current account balance is directly followed by government debt, where a higher debt is associated with a lower rating, which is in line with economic intuition. Political stability and unemployment rate, ranking 5$^{\text{th}}$ and 6$^{\text{th}}$, also show a pattern although less pronounced than the previously discussed variables. Here, a higher political stability and/or a lower unemployment rate are associated with an increase in the credit rating, and vice versa.

The three least important variables, being government balance, GDP growth, and inflation, show no clear effect. Inflation even seems to have no influence at all. The relative unimportance of these three factors is quite sensible. A negative government balance is

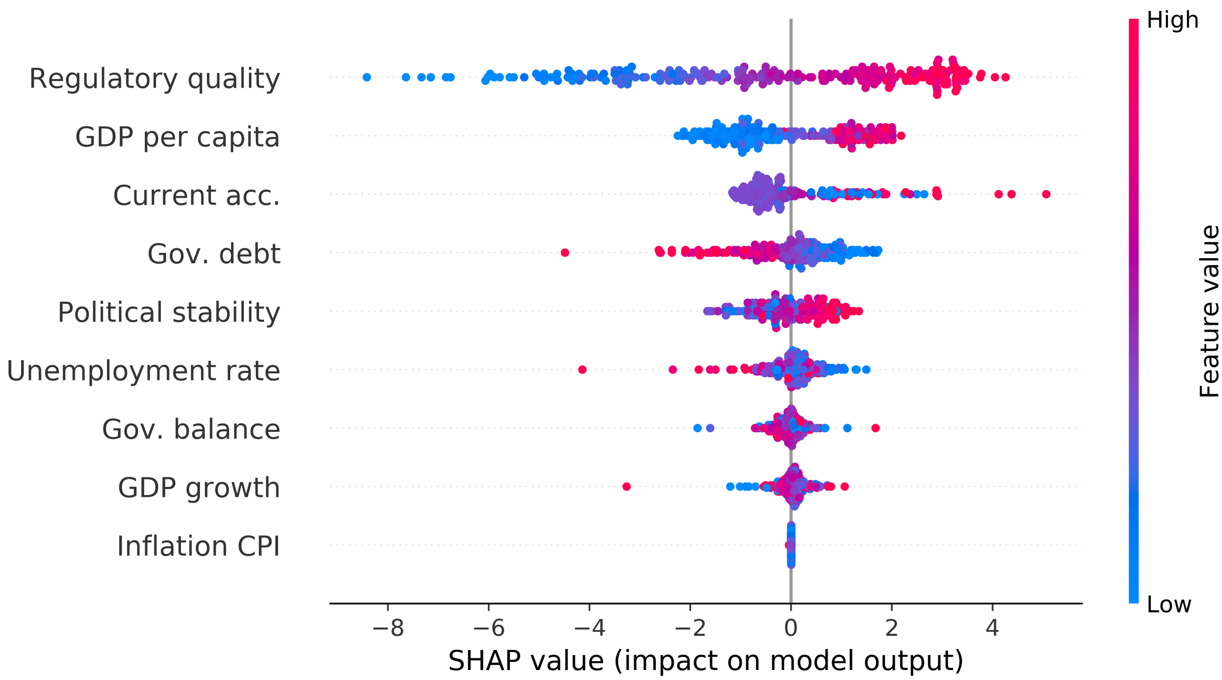

Fig. 2. SHAP values plot for the MLP, explanatory variables are ranked from highest mean absolute SHAP value (regulatory quality) to lowest (inflation). Individual dots represent data points that have been evaluated, where the color indicates whether the value is relatively high (red) or low (blue) for that explanatory variable. The x-axis shows the impact of the particular feature on the prediction, i.e. the number of notches the prediction deviates from the baseline prediction when that feature is included.

generally a bad sign, because it increases the government debt. However, as previously discussed, a higher rating leads to a lower interest rate and therefore less inclination to keep the debt low. Government balance is thus not such a helpful factor in the credit rating process. GDP growth and inflation do not lend themselves very well for distinction between creditworthy and non-creditworthy countries. In the case of GDP growth, we observe that lower rated countries have on average a higher GDP growth, but a lower cumulative GDP growth over long periods, which in the end determines the long-term growth of the economy. While hyperinflation is obviously a bad sign, and should lead to a low rating, inflation offers no clear guidance for the other values.

Classification and Regression Trees

The same procedure as for the MLP is repeated to isolate the influence of variables in the CART, the plot containing SHAP values for the CART is shown in Figure 3. Furthermore,

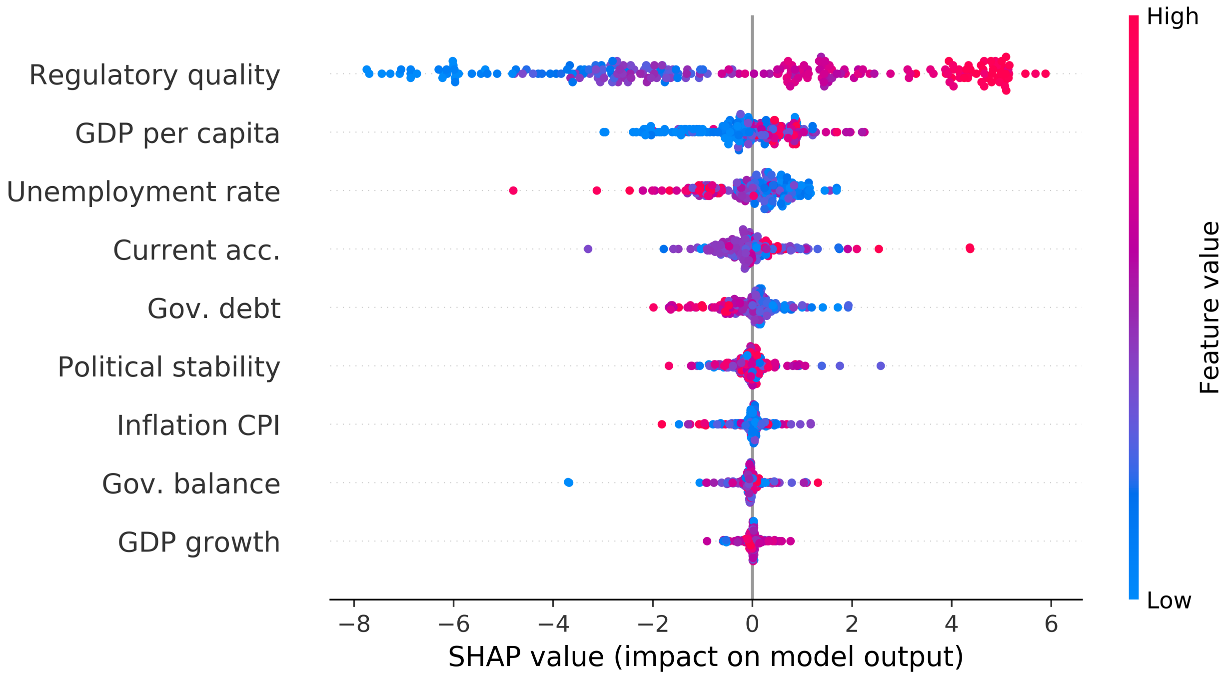

Fig. 3. SHAP values plot for the CART, explanatory variables are ranked from highest mean absolute SHAP value (regulatory quality) to lowest (GDP growth). Individual dots represent data points that have been evaluated, where the color indicates whether the value is relatively high (red) or low (blue) for that explanatory variable. The x-axis shows the impact of the particular feature on the prediction, i.e. the number of notches the prediction deviates from the baseline prediction when that feature is included.

Table 3: Ranking of the variables based on influence in the predictions of MLP, CART, and OL. The higher the rank, the more important the variable, with 1 being the most influential variable. As gov. debt is excluded from the OL model, no rank can be assigned to the variable, however, this could be interpreted as ranking last.

	MLP	CART	OL
Regulatory quality	1	1	1
GDP per capita	2	2	5
Current acc.	3	4	7
Gov. debt	4	5	-
Political stability	5	6	6
Unemployment rate	6	3	2
Gov. balance	7	8	8
GDP growth	8	9	3
Inflation	9	7	4

to facilitate comparison with the MLP, Table 3 shows the explanatory variables ranked on importance for every model. In the CART, similar to the MLP, regulatory quality and GDP per capita are the most important and second most important variables respectively. As expected, a higher regulatory quality and a higher GDP per capita are both associated with a higher credit rating.

In the CART, just as in the MLP, unemployment rate, current account balance, government debt, and political stability rank 3rd to 6th, although the exact order differs. The unemployment rate shows a clear relation to the credit rating, where a higher unemployment rate is associated with a lower credit rating. The 4th most important variable, current account balance, shows the same non-linear behaviour as the MLP, with average values resulting in a lower credit rating, and the extremes in a higher one. However, in this case, the effect is much less pronounced than in the MLP. Government debt proves to have a negative effect on the credit rating, which is in line with economic theory. In contrast to the MLP, political stability shows no clear relation to the credit rating in the CART.

The three least important variables in the CART match those of the MLP. Inflation, government balance, and GDP growth show no distinct relation to the credit rating.

Ordered Logit

Extracting the determining factor of the OL model is relatively simple, as the model only uses linear relations. The significance of the coefficients of different variables, combined with the sign, tells us how important a variable is and if the relation to the credit rating is positive or negative. The estimated coefficients, together with their standard error and p-value, are shown in Table 4. No coefficient for government debt is estimated because excluding government debt from the model results in a higher cross-validated accuracy. The ranking of importance for all the variables in the OL model is also shown in Table 3, together with those of the other models.

Again, regulatory quality proves to be the most important variable, where the positive sign of the coefficient shows that the relation is positive, as was the case for the MLP and CART. That regulatory quality is most important in all models is a strong indication that it is a very important factor in the credit rating process. The second most important variable in the OL model is unemployment rate, where a higher unemployment rate is associated with a lower rating. The unemployment rate is followed by GDP growth and inflation, which ranked very low in the other two models, and in the case of GDP growth, the sign is counter to expectation. However, as previously discussed, lower rated countries have a higher GDP growth on average, just not a higher cumulative growth. The OL model, with its linear relations, therefore finds a negative relation between GDP growth and the credit rating. GDP per capita ranks 5^{th} in the OL model, where it ranked 2^{nd} in the other two models. The sign does match expectations with a higher GDP per capita associated with a higher credit rating. The next variable is political stability, here the three models more or less agree on the importance of the factor. However, the sign of the coefficient in the OL model is counter to economic theory, which could be an effect of the inclusion of a lot of variables in a linear setting. The last two variables, current account balance and government balance have a positive influence on the credit rating, which is in line with economic theory. That current account balance is relatively unimportant in the OL models compared to the other two makes sense, as the OL model cannot pick up on the non-linear relation that the other models find.

The rank of government balance in the OL model is similar to that of the other two models. While the coefficients of current account balance and government balance are insignificant, they do contribute to the cross-validated accuracy and are therefore included in the model. The only factor that does not contribute to a higher cross-validated accuracy is government debt. This is strange, since government debt is commonly assumed to be a very important factor in the creditworthiness of a country. Countries that already have a lot of debt might be less able to repay new debt. It is nonetheless not a clear distinguishing factor on its own. There are also Aaa rated countries that have a lot of debt, since they have less inclination to keep the debt low due to the low interest rates that they pay. Government debt therefore seems to only be influential when taking into account other factors at the same time, and is thus not useful for the OL model. These results confirm earlier findings that government debt is a useful variable to split data on, but not necessarily useful in a regression model, see, for example, Bozic and Magazzino (2013); Reusens and Croux (2016).

Table 4: Coefficients, standard errors and p-values for the Ordered Logit model.

	Coefficients	S.E.	p-value
GDP growth (%)	-0.0012	0.0000	0.0000
Inflation (%)	-0.0467	0.0118	0.0001
Unemployment rate (%)	-0.1082	0.0014	0.0000
Current acc. (% of GDP)	0.0124	0.0182	0.4951
Gov. balance (% of GDP)	0.0409	0.1200	0.7331
Political stability	-0.2623	0.1567	0.0941
Regulatory quality	3.6001	0.0045	0.0000
GDP per capita (1000$)	0.0337	0.0182	0.0642

The large similarities in determining factors, especially between CART and MLP, are surprising, since the modelling techniques are quite different. This makes it more likely that some of variables found to be important in this study, such as regulatory quality, have a large influence on the credit rating.

5. Conclusion

This paper investigates the use of two Machine Learning techniques, Multilayer Perceptron (MLP) and Classification and Regression Trees (CART), and an Ordered Logit (OL) model, for prediction of sovereign credit ratings. MLP proves to be most suited for predicting Moody's ratings based on macroeconomic variables. Using random 10-fold cross-validation it reaches an accuracy of 68%, and predicts 86% of ratings correct within 1 notch. Thereby, it significantly outperforms CART and OL with their respective accuracies of 59% and 33%.

Investigation of the determining factors, which has so far not been done for Machine Learning models in the sovereign credit rating setting, shows that there are common influential factors across the models. Regulatory quality and GDP per capita are respectively the most important and second most important factor in the MLP and CART, with, as expected, a positive relation between both variables and the predicted credit rating. This behaviour is also reflected by the signs of the respective coefficient in the OL model. Other, slightly less, influential variables are: current account balance, government debt, political stability and unemployment rate. The behaviour of MLP and CART with respect to most of these variables is similar. A higher government debt and unemployment rate are associated with a lower credit rating, and for both models an average current account balance value leads to a lower rating while a relatively low or high value leads to a higher credit rating. The models differ on the interpretation of political stability. In the MLP, a higher value for political stability leads to a higher credit rating, but there is no clear relation in the CART. Most of the previously mentioned effects are also observed in the signs of OL coefficients. However, the signs of GDP growth and political stability are in contrast to economic theory, where a higher GDP growth and/or political stability are associated with a lower credit rating, possibly due to inclusion of all variables jointly in the restrictive linear setting.

In short, we advice governments wanting to check their rating or investors deliberating an investment to use a MLP model, as this model proves to be most accurate. Sovereign credit ratings are heavily influenced by the regulatory quality and GDP per capita of a country. Expected changes in either of these factors could thus result in a credit rating change. Anticipating this possible change can be very valuable, as the credit rating has a major influence on the interest rate at which governments can issue new debt, and thus on the government budgets.

We end this paper with a few recommendations for future research. First, the determining factors of other Machine Learning techniques, such as Support Vector Machines (SVM), Naive Bayes (NB), and Bayes Net (BN), in the sovereign credit rating setting could be investigated. Second, the inclusion of more explanatory variables might increase accuracy of some methods (most likely CART) and might lead to more insights into the relevant variables.

References

Afonso, A., Gomes, P., and Rother, P. (2011). Short- and long-run determinants of sovereign debt credit ratings. *International Journal of Finance & Economics*, 16(1):1–15.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Bennell, J. A., Crabbe, D., Thomas, S., and ap Gwilym, O. (2006). Modelling sovereign credit ratings: Neural networks versus ordered probit. *Expert Systems with Applications*, 30(3):415 – 425. Intelligent Information Systems for Financial Engineering.

Bissoondoyal-Bheenick, E. (2005). An analysis of the determinants of sovereign ratings. *Global Finance Journal*, 15(3):251 – 280. Special Issue.

Bozic, V. and Magazzino, C. (2013). Credit rating agencies: The importance of fundamentals in the assessment of sovereign ratings. *Economic Analysis and Policy*, 43(2):157 – 176.

Cantor, R. and Packer, F. (1996). Determinants and impact of sovereign credit ratings. *Economic Policy Review*, 2(2).

Chollet, F. et al. (2015). Keras. https://keras.io.

Dimitrakopoulos, S. and Kolossiatis, M. (2016). State dependence and stickiness of sovereign credit ratings: Evidence from a panel of countries. *Journal of Applied Econometrics*, 31(6):1065–1082.

Elkhoury, M. (2009). Credit rating agencies and their potential impact on developing countries. *UNCTD Compendium on Debt Sustainability*, pages 165–180.

Ferri, G., Liu, L.-G., and Stiglitz, J. E. (1999). The procyclical role of rating agencies: Evidence from the east asian crisis. *Economic Notes*, 28(3):335–355.

Gaillard, N. (2009). The determinants of moody's sub-sovereign ratings. *International Research Journal of Finance and Economics*, 31(1):194–209.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124 – 136.

Luitel, P., Vanpée, R., and Moor, L. D. (2016). Pernicious effects: How the credit rating agencies disadvantage emerging markets. *Research in International Business and Finance*, 38:286 – 298.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Moor, L. D., Luitel, P., Sercu, P., and Vanpée, R. (2018). Subjectivity in sovereign credit ratings. *Journal of Banking & Finance*, 88:366 – 392.

Ozturk, H. (2014). The origin of bias in sovereign credit ratings: Reconciling agency views with institutional quality. *The Journal of Developing Areas*, 48(4):161–188.

Ozturk, H., Namli, E., and Erdal, H. I. (2015). Reducing overreliance on sovereign credit ratings: Which model serves better? *Computational Economics*, 48:59 – 81.

Ozturk, H., Namli, E., and Erdal, H. I. (2016). Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample. *Economic Modelling*, 54:469 – 478.

Panchal, G., Ganatra, A., Kosta, Y., and Panchal, D. (2011). Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3(2):332–337.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pedregosa-Izquierdo, F. (2015). *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *CoRR*, abs/1710.05941.

Reusens, P. and Croux, C. (2016). Sovereign credit rating determinants: the impact of the european debt crisis. *Available at SSRN 2777491.*

Reusens, P. and Croux, C. (2017). Sovereign credit rating determinants: A comparison before and after the european debt crisis. *Journal of Banking & Finance*, 77:108 – 121.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Appendix A. Appendix

A.1. List of countries

Countries included in the data set: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, China, Colombia, Costa Rica, Cyprus, Czech Republic, Denmark, Dominican Republic, El Salvador, Fiji Islands, Finland, France, Germany, Greece, Honduras, Hungary, Iceland, Indonesia, Ireland, Israel, Italy, Japan, Jordan, Korea, Latvia, Lithuania, Luxembourg, Malaysia, Malta, Mauritius, Mexico, Moldova, Morocco, Netherlands, New Zealand, Norway, Pakistan, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovenia, South Africa, Spain, Sweden, Switzerland, Thailand, Tunisia, United Kingdom, Venezuela.

Table 5: Conversion of Moody's ratings into numeric ratings.

Moody's rating	Numeric rating
Aaa	17
Aa1	16
Aa2	15
Aa3	14
A1	13
A2	12
A3	11
Baa1	10
Baa2	9
Baa3	8
Ba1	7
Ba2	6
Ba3	5
B1	4
B2	3
B3	2
Caa1	1
Caa2	1
Caa3	1
Ca	1
C	1

A.2. Rating transformations

Table 5 shows the transformation of all Moody's ratings into numerical ratings.

A.3. MLP optimization

There are basically two ways of optimizing model hyperparameters: grid search and Bayesian model-based optimization. While the Bayesian methods are more likely to give you the optimal setting, they give no insight into the different performance-complexity trade-offs. As that trade-off is important in this study, since interpretation suffers for more complex models, a grid search is used to optimize the MLP.

In this grid search, five hyperparameters are optimized: number of hidden layers, number of neurons per hidden layer, dropout rate, number of epochs and batch size. The number of hidden layers and number of neurons per hidden layer, as previously explained, determine the structure of the MLP. The dropout rate gives the fraction of neurons that is dropped from the model at random. Randomly dropping neurons from the model prevents overfitting, as an overfitted model would perform very poorly when neurons are left out. The number of epochs and batch size determine how the internal model parameters are estimated. When a MLP is trained, it updates the parameters after working through a number of data points. That is, the internal parameters are not updated after evaluating every individual data point, but after evaluating a certain number of data points, a batch. A larger batch size thus means that the algorithm evaluates more data points before updating the parameters and vice versa for a small batch size. The number of epochs determines how many times the algorithm goes through the entire data set, especially for smaller data sets this number can be very large, often hundreds or thousands.

Setting up a full grid, where all the different combinations are tested, is computationally extremely expensive, since the number of possible combinations becomes very large. We have therefore opted for two separate grid searches. First, one where the optimal structure is investigated: hidden layers, neurons and dropout rate. Thereafter, a second grid search in which the estimation of the optimal structure found in the first grid search is analysed: epochs and batch size.

In the first grid search the following parameters are considered: hidden layers [1, 2, 3], neurons [8, 16, 32, 64, 128, 256, 512] and dropout rate [0, 0.1, 0.2] using a batch size of 8 and 400 epochs. In general, one hidden layer suffices, only in cases where there are discontinuities in the data is more than one hidden layer required (Panchal et al., 2011). Therefore, the grid search is limited to three hidden layers, to make sure that additional hidden layers do not improve performance. There are rules of thumb for selecting the number of neurons, such as that it should be between the size of the input and the size of the output layer. However, deviating from these rules often results in drastically improved performance. Even though 512 neurons seems excessively large, and is unlikely to improve performance compared to lower numbers, it is still evaluated to make sure no performance increase is obtained. The dropout parameters are set in such a way that we can see if dropout is needed, or if dropping a significant, but not too large, fraction of the neurons improves performance.

Thereafter, for the optimal structure, we investigate the estimation parameters using: batch size [4, 8, 16, 32] and epochs [100, 200, 400, 800]. Keskar et al. (2016) show that the batch size should be much smaller than the total number of data points in the set, and that using a large batch size decreases the ability of the model to generalize. For these reasons, we have decided to set an upper bound of 32 on the batch size. There are no clear guidelines for the optimal number of epochs, as this is highly dependent on the data set. The number of epochs is thus increased until performance of the MLP stops improving. If optimal performance in this grid is found at 800 epochs, the use of an even higher number of epochs is investigated.

The results of the two grid searches are shown in Tables 6 and 7. The optimal performance-complexity trade-off is in our view given by the MLP with 1 hidden layer, 256 neurons, and a dropout rate of 0.1. Even though the MLP with 2 hidden layers, 256 neurons, and a dropout rate of 0.2 has a slightly higher accuracy, we deem the increase in accuracy too small to justify addition of a hidden layer. The results of the estimation grid search show that performance increases with more epochs, but levels off at about 200 epochs. Since we rather be on the safe side, we opted for 400 epochs. There is very little variation in performance between the different batch sizes, although combinations of a low number of epochs with large batches perform poorly. We have therefore decided to use a batch size of 8, for which the MLP was structure was optimized.

A.4. CART Optimization

There are two ways in which a CART can be restricted: restricted growth and pruning. When restricting the growth of the CART, we limit the growth of the tree a priori, while with pruning we allow the tree to grow unobstructed but cut off branches afterwards.

Limiting the growth of the CART can be done in multiple ways. In this study, we optimize the following settings: maximum depth, minimum samples for a split and minimum impurity decrease. Maximum depth limits the number of splits the tree is allowed to make by stopping after a certain depth is reached. That is, it limits the number of sequential splits the algorithm is allowed to make, counting from the root node. The minimum samples for a split restrict splitting, if the minimum number for a split is not reached, the algorithm is forced to make a leaf there. Lastly, a restriction can be set on the minimum impurity decrease, which means that the algorithm is only allowed to make a further split if that leads to a certain decrease in the Gini impurity (Equation 6).

For the CART, just as for the MLP, we use a grid search instead of Bayesian hyperparameter optimization techniques to get insight into the performance of the CART. Selecting the hyperparameter values to be included in the grid search requires some preliminary investigation, since the restrictions have to be adjusted to the size the tree would grow to when left unrestricted. Being too restrictive compared to unrestricted growth will significantly harm performance, whereas restrictions that do not restrict maximum growth have no effect at all. Initial investigation shows that a tree grown unrestrictedly on the full data set ends up with 320 leaves and a maximum depth of 20. Therefore, we have decided to use the following grid parameters: maximum depth [10-20], minimum samples for a split [2, 3, 4, 5] and minimum impurity decrease [0-0.0002] in steps of 0.00001.

Instead of limiting the growth of the tree, we can also prune it after letting it grow unre-

Table 6: MLP model structure optimization with hidden layers [1, 2, 3], neurons [8, 16, 32, 64, 128, 256, 512] and dropout [0, 0.1]. All models are estimated using batch size 8 and 400 epochs. Optimal structure, in terms of accuracy, consists of 2 hidden layer with 256 neurons with a dropout rate of 0.2. The best performance-complexity trade-off, underlined in the table, is obtained by the MLP with 1 hidden layer, 256 neurons and a dropout rate of 0.1. All numbers are given in %.

Batch size Epochs	8 400	No dropout		Dropout 0.1		Dropout 0.2	
Number of hidden layer(s)	Neurons per hidden layer	Accuracy Mean	Std	Accuracy Mean	Std	Accuracy Mean	Std
1	8	43.9	5.8	41.6	5.1	38.5	5.0
	16	50.9	4.6	49.4	3.5	46.1	4.9
	32	59.2	4.5	56.0	4.2	53.7	6.4
	64	63.7	3.7	64.1	3.3	63.3	6.9
	128	66.2	3.4	68.5	3.7	68.2	5.2
	256	67.1	4.2	<u>69.7</u>	3.7	69.4	4.5
	512	67.1	3.8	68.3	2.2	68.2	4.5
2	8	40.8	5.3	38.8	4.7	37.0	5.0
	16	48.7	4.1	43.7	4.3	41.3	4.7
	32	55.6	4.6	56.6	4.4	51.9	5.3
	64	62.1	4.0	64.3	4.2	65.7	6.0
	128	65.5	4.5	68.9	4.4	68.3	4.3
	256	67.7	2.8	69.5	3.1	70.0	4.2
	512	68.8	2.2	68.1	1.9	69.6	2.9
3	8	39.5	8.3	38.5	5.9	34.0	4.8
	16	45.6	5.4	45.0	5.8	38.9	4.8
	32	52.8	3.0	54.7	5.2	46.3	5.8
	64	62.8	5.1	64.8	3.8	63.7	4.6
	128	65.5	4.5	67.9	4.0	69.6	4.8
	256	66.4	2.5	68.1	3.9	68.4	4.9
	512	68.1	1.8	68.3	4.1	66.9	4.9

Table 7: MLP model estimation optimization with epochs [20, 50, 100, 200, 400, 800] and batch size [4, 8, 16, 32] on the MLP with 1 hidden layer, 256 neurons and dropout rate 0.1. Optimal estimation is achieved using 200 epochs and a batch size of 8. All numbers are given in %.

Epochs	Batch size	Mean acc.	Std
20	4	56.9	4.8
	8	54.3	6.2
	16	52.5	4.8
	32	50.3	4.3
50	4	65.4	3.7
	8	65.0	3.5
	16	60.9	4.4
	32	56.4	4.4
100	4	67.7	4.1
	8	67.5	5.9
	16	65.4	4.1
	32	63.2	4.7
200	4	67.5	3.8
	8	<u>68.9</u>	5.2
	16	67.1	3.7
	32	67.6	3.3
400	4	68.0	4.8
	8	68.7	5.1
	16	68.0	4.1
	32	67.2	5.6
800	4	68.3	4.8
	8	68.2	5.1
	16	68.2	5.4
	32	68.2	4.1

stricted. In this study, we make use of minimal cost-complexity pruning, that is, minimizing the cost-complexity criterion

$$C_\alpha(L) = \sum_{l=1}^{|L|} \left(\sum_{u_i \subset R_l} (y_i - \hat{y}_l)^2 \right) + \alpha |L|, \tag{12}$$

where each l represents a leaf and $|L|$ is the total number of leaves. The set R_l contains all the data points in leaf l, $\hat{y}_l$ is the prediction for leaf l and α is the factor punishing for complexity (Hastie et al., 2009). In words, the cost-complexity criterion is the sum of squared errors with an additional factor that punishes for tree complexity in the form of the number of leaves. Thus, optimizing tree complexity is done by optimizing the factor α.

The results of the CART optimisation are very straight forward, any a priori limitation tree growth or pruning results in a decreases out-of-sample accuracy. An unrestricted CART is therefore most suited for sovereign credit rating predictions on this data set and is thus used in this research.