Predicting Default Probabilities for Stress Tests: A Comparison of Models

Martin Guth*
February 8, 2022

Abstract
Since the Great Financial Crisis (GFC), the use of stress tests as a tool for assessing the resilience of financial institutions to adverse financial and economic developments has increased significantly. One key part in such exercises is the translation of macroeconomic variables into default probabilities for credit risk by using macrofinancial linkage models. A key requirement for such models is that they should be able to properly detect signals from a wide array of macroeconomic variables in combination with a mostly short data sample. The aim of this paper is to compare a great number of different regression models to find the best performing credit risk model. We set up an estimation framework that allows us to systematically estimate and evaluate a large set of models within the same environment. Our results indicate that there are indeed better performing models than the current state-of-the-art model. Moreover, our comparison sheds light on other potential credit risk models, specifically highlighting the advantages of machine learning models and forecast combinations.

Keywords: stress test, credit risk, default probability, model comparison

JEL codes: C53, E58, G32

*Oesterreichische Nationalbank (OeNB), Otto-Wagner-Platz 3, A-1090 Vienna, Austria. Email: martin.guth@oenb.at. The views expressed in this paper are those of the author and do not necessarily reflect those of the Eurosystem or the OeNB. The author would like to thank Florian Huber and Robert Ferstl for helpful comments that significantly improved the article.
1 Introduction

Stress tests have a twenty-year history as tools for micro- and macroprudential supervision and are now used regularly by financial institutions and those who supervise them. The aim of such tests is to assess institutions’ resilience to adverse financial and economic developments, as well as to contribute to the overall assessment of systemic risk in the financial system. In order to assess the behavior of financial institutions to stress, a multi-year macrofinancial scenario needs to be designed. It captures relevant systemic risks and the materialization thereof to generate stress in the system.

The present study focuses on the use of macrofinancial linkage models to translate the country-level scenario into bank-level risk parameters, which are an essential input to every stress test. These so-called satellite models thus provide scenario-conditional forecasts for the probabilities of default (PD). What is key in these models is the proper detection of signals from an array of global variables which are not directly linked to the financial health of companies.

Outside the context of stress testing, such credit risk models have a long-standing history (Keeton & Morris, 1987; Wilson, 1998) and still represent a very active research area in which various models in different setups have been proposed. Specifically, the literature covers linear models (see, e.g., Aver, 2008; Bofondi & Ropele, 2011), vector autoregressive (VAR) models (see, e.g., Gambera, 2000; Pesaran, Schuermann, Treutler, & Weiner, 2006; Castren, Dees, & Zaher, 2010), panel models (see, e.g., Pesola, 2001; Castro, 2013), latent factor models (see, e.g., Koopman & Lucas, 2005; Kerbl & Sigmund, 2011), quantile regressions (Schechtman & Gaglianone, 2012) and machine learning models (Jacobs, 2018). Interestingly, besides Jacobs (2018), in the machine learning literature, there seems to be little to no coverage of credit risk models that estimate or predict PDs. Even in very recent literature surveys that specifically cover specifically machine learning in banking risk management (Leo, Sharma, & Maddulety, 2019) and machine learning in credit risk (Breeden, 2020), there seems to be just a few papers in the approximate vicinity of this topic.

The current state-of-the-art satellite model for PD translation is Bayesian model averaging (BMA) (Raftery, 1995). It has a long track record as being a reliable tool for generating scenario-conditional projections for credit risk and is being adopted by more and more central banks and institutions. The inherent advantage of BMA is the explicit tackling of model uncertainty by operating on a large pool of competing models which are weighted by their predictive performance and combined to one final model. However, with easier access to sophisticated regression approaches provided by open source programming languages such as R (R Core Team, 2020), Julia (Bezanson, Edelman, Karpinski, & Shah, 2017) or Python (Van Rossum & Drake, 2009) and the advent of new predictive models in the field of machine learning, the question arises if there are other models that could deliver better results.

The aim of this paper is to conduct a systematic forecast comparison with a large number of different regression models to find the best performing credit risk satellite model. The winning model is evaluated for the ability to precisely forecast default
probabilities conditional on a standardized set of macroeconomic variables as provided to financial institutions by the ESRB (2020) for the EU-wide banking sector stress test. We implemented a total of 43 models, which can be assigned to 9 categories, ranging from conventional statistical models to more recent machine learning methods. We tried to encompass as many models as possible with a proven track record in forecasting linear and non-linear relationships and which were readily available within an R library (R Core Team, 2020). Additionally, we also combine the models with different forecast combination approaches to further gauge their potential accuracy. For the purpose of this paper, we implement a framework that allows us to conduct this comparison with a standardized data set for all models, to tune the respective hyperparameters for each model and to cross-validate the results based on recursive pseudo out-of-sample forecasts.

There are only two papers in the literature that are related to our work. Papadopoulos, Papadopoulos, and Sager (2016) created a composite satellite model for stress testing by weighting candidate models from the full space of all possible variable combinations. Grundke, Pliszka, and Tuchscherer (2019) used a combination of BMA to select relevant variables and OLS to regress the selected variables onto a credit default index. They further analyzed different modifications to various steps in their estimation framework and assessed the different outcomes in terms of out-of-sample default rate forecasts. Both papers focus on the proper identification of the model given one estimation technique but lack comparisons across different procedures.

Our paper contributes to the literature in the following ways: To the best of our knowledge, it represents the first systematic model comparison in the field of credit risk and stress testing. First and foremost, we deliver insights on a large number of potential credit risk satellite models across a wide range of modelling techniques. Since it is not evident which modelling technique will achieve the best results, we take a naive approach by testing many different models without prejudice on how well they will fare. Our results indicate that there are better performing models than the current state-of-the art model (i.e. BMA), which have not been mentioned in the literature until now. Second, due to the large variation in models, we are able to shed light on the potential positive effects of machine learning models and thus extend the scarce literature in this area. Moreover, the use of different forecast combination techniques across different sets of models shows that not only can these techniques help to hedge against model uncertainty, but they can also enhance the predictive accuracy.

The remainder of the paper is structured as follows. In Section 2 our estimation and evaluation framework are explained in detail, including the used data, models, hyperparameter tuning and performance measures. Thereafter, the results are shown in Section 3, which also presents a deep dive into the winning model. Section 4 concludes the paper.
2 Design of the Forecasting Exercise

This section outlines the setup of our forecasting exercise. First, we present the underlying data set and the applied transformations. Second, we give a short summary of the 43 models within the 9 overarching categories. Third, as many models need a prior setting of parameters, we discuss the process of tuning the respective hyperparameters. Fourth, we briefly discuss the measure used to evaluate the forecasting performance.

2.1 Data

The data involved in the modelling exercise refer to Austria and include as dependent variable a measure for the probability of default for the non-financial corporate sector and macroeconomic and financial data as independent variables. The deployed default probabilities are based on Moody’s KMV Expected Default Frequency (EDF) measure, available at quarterly frequency from 2002Q2 to 2019Q2 ($T = 1, \ldots, 69$). The EDFs have been successfully deployed in multiple credit risk models (see, e.g., Alves, 2005; Castren et al., 2010; Gross & Población, 2019). In adherence to the IMF’s approach of credit risk modelling in their Financial Sector Assessment Programs (FSAP), we also take the provided mean PD measures and not the median (see IMF, 2020). Hence, the measure relates to the average default probability across non-financial corporations.

The independent covariates are based on the variables within the macroeconomic scenario as designed for the EU-wide stress tests by the European Systemic Risk Board (ESRB, 2020). These variables are real GDP growth ($GDP$), the unemployment rate ($UNE$), the inflation rate ($INF$), real estate price growth ($RRE$), stock price growth ($EQP$), exchange rates ($EXR$) and short-term ($STR$) and long-term interest rates ($LTR$). However, the actual scenario figures are not needed for the estimation as we conduct pseudo out-of-sample forecasts for which the actual default probability time series is needed. Hence, the scenario serves as guidance for the choice of covariates, but the models are evaluated on the basis of historical figures.

Although the range of the time series is limited by the availability of the default probability, it still includes non-linear events such as the financial crisis of 2008 and the European sovereign debt crisis of 2011, which is an important feature for credit risk satellite models, as they must be able to estimate and predict such structural breaks. We get data on real GDP, Harmonized Index of Consumer Prices (HICP) and the unemployment rate from Eurostat. The GDP figure is seasonally and calendar adjusted and transformed to year-on-year (YoY) growth rates to fit the figures used by the ESRB. Similarly, the HICP is also transformed to YoY growth rates to match the ECB definition of inflation (ECB, 2020). The real estate prices and EUR/USD exchange rate are sourced from the ECB’s Statistical Data Warehouse (SDW). The house prices cover all new and existing residential properties across all dwelling types and are also transformed into YoY growth rates. Finally, we take 10-year government bond yields as long-term interest rates, 3-month Euribor as short-term interest rates and the equity price index for Austria, the ATX, from the OECD database. The equity price index is also transformed into YoY growth rates.
A first descriptive analysis of the variables and potential correlations among them can be drawn from Figure 1. The chart depicts the variables without further transformations and the grey shaded areas mark the two previous crisis periods – the GFC (2008 Q2 - 2009 Q2) and the European sovereign debt crisis (2011 Q1 - 2013 Q1). The sample starts exactly at the peak of the crisis that unfolded in the aftermath of the burst of the dotcom bubble in 2000, the uncertainty triggered by the 9/11 attacks and the very volatile years after the introduction of the euro as a new currency from 2000 - 2001. The economic shockwave led to a significant increase in expected and actual corporate defaults around the globe (Altman & Bana, 2003).

In the case of Austria, the default probability starts with a value of 5.3% and decimates until the start of the financial crisis, when it reaches the sample maximum of 6%. The pattern during the European sovereign debt crisis is not as clear, even though there is a small increase in the EDFs towards the end of 2012. The movements of the macro and financial variables behave as expected during the downturns. In both periods, we see significant drops in GDP growth and inflation, an uptick in the unemployment rate, large negative distortions on the stock market, devaluations of the euro vis-a-vis the US dollar and an increase in real estate prices reflecting a flight to safe investments, and we also see the ECB reacting to these events by pushing the short-term interest rate and indirectly the long-term interest rates towards zero and beyond. Even though the co-movements of the variables within the structural breaks seem to be going in sensible directions (e.g. GDP down, PD up), the credit risk satellite models in our setup will need to be able to pick up correct signals in calmer periods. In the majority of European countries, the same co-movement patterns have been observed in the last two decades. Therefore, our conclusions based on Austria can be generalized to other regions.

As a last transformative step, we need to make sure that the variables have optimal properties and behave well in predictions. First, to ensure that the point forecasts of the default probability is bounded between a 0% and 100% interval, we apply the following logit transformation for the regression and calculation of the performance measures,

\[
y = \logit(PD) = \log \left( \frac{PD}{100 - PD} \right)
\]  

(1)

Second, we analyze the trend and cyclical components of the variables and perform a series of unit root tests. The seasonal decomposition by loess (Robert, William, & Irma, 1990) shows that all variables, except GDP (since it has already been adjusted), exhibit a form of cyclicality, which we remove in due course. In order to get a clear picture of the stationarity of the variables we deploy the augmented Dickey-Fuller (ADF) test (Dickey & Fuller, 1979), the Elliott, Rothenberg & Stock (ERS) test (Elliott, Rothenberg, & Stock, 1996), the Phillips-Perron (PP) test (Phillips & Perron, 1988) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Without further transformations, all variables would suffer from unit roots. Hence, by taking the first difference for each variable, the test statistics in each
Figure 1: Overview variables and recession indicators

Note: the grey shaded areas mark Austria’s economic crisis periods based on the recession indicator of the Federal Reserve Bank of St. Louis. These two areas are the Great Financial Crisis (2008 Q2 - 2009 Q2) and the European sovereign debt crisis (2011 Q1 - 2013 Q1).

Source: Moody’s KMV EDF, Eurostat, ECB SDW, OECD, Federal Reserve Economic Data (FRED).

unit root test indicate robust level and trend stationarity.

2.2 Econometric Models

In this section we give a short overview of, and introduction to, the adopted credit risk satellite models. In order not to artificially expand the paper, the formal definitions
of the models are not mentioned below. The interested reader is referred to the publications cited in each paragraph. The selection of the models was based, on the one hand, on the desire to cover as many models as possible with a wide array of different features. In doing so, we want to extend the existing model space in the literature from mainly linear models to non-linear, data-driven models with a special focus on regularization. Especially the latter point will turn out to be very important when analyzing the results. On the other hand, as the implementation of so many models is time-intensive, we focus our attention on proven models that are readily available, using open source computing environments, such as R (R Core Team, 2020).

In total, we have implemented 43 models which are placed in 9 overarching groups to give the reader a better overview of the models at hand. All satellites use the same basic model structure in which the dependent variable \( y = (y_1, \ldots, y_T)' \) is described as a function of contemporaneous and lagged predictors \( X = (x_1, \ldots, x_T)' \), such that

\[
y = f(X_t, X_{t-1}, \ldots, X_{t-p}) + \varepsilon
\]  

(2)

The actual approximation of \( f() \) is dependent on the respective model in our comparison. Notice that the equation is ordered by the lags, i.e. \( p = 0, 1, \ldots P \). This fact is important for certain stepwise regressions and it achieves overall better prediction results than with the equation being sorted by the variables. We set \( P = 4 \) as a standard choice for quarterly data. We follow Gross and Población (2019) by forcing a "closed" lag structure without gaps between the first and fourth lag for the initial equation. Gross and Población (2019) found in their analysis that this type of structure appeared to be more meaningful and robust. However, models with regularization or variable selection are not constrained in their choice of the proper equation.

We do deviate from some papers in the literature on credit risk satellite models in that we do not include an autoregressive lag of the dependent variable. When we tested both options, nearly all models showed a worse performance with the autoregressive lag than without, while the overall ranking across the models remained stable. Moreover, we did not want to increase the already large set of contemporaneous and lagged regressors \( (n = 40) \) in combination with the limited sample length \( (T = 64) \).

(Generalized) Linear Models

The first class of competitors are (generalized) linear models with a proven track record in a number of disciplines. We start with naive benchmarks based on a standard ordinary least squares (OLS) regression (short name: \texttt{lm}) and its robust alternative based on the M-estimator (\texttt{rlm}) as defined in P. Huber (1992).

Nevertheless, due to the high number of variables relative to the sample, overfitting in the context of predictions is problematic. Thus, the next group of models uses regularization techniques to reduce the number of covariates. More precisely, we implement a forward selection algorithm (\texttt{lmfs}) (Hocking, 1976) and a least-angle regression

\footnote{For a compact overview of the models and R libraries see Table 2 in the annex.}
Shrinkage estimators deviate from the stepwise approach as they penalize the coefficients to reduce multicollinearity in the equation. Here we chose the ridge estimator (ridge) (Tikhonov, 1943; D. Phillips, 1962) and, as the ridge cannot yield sparse equations because no covariates are dropped due to the regularization procedure, the least absolute shrinkage and selection operator (lasso) by Tibshirani (1996).

We further implement three refinements to the lasso estimator that try to deal with certain shortcomings of the original estimator. First, if one takes the temporal structure of the data into account, we get the fused lasso (flasso) by Tibshirani, Saunders, Rosset, Zhu, and Knight (2005), which adds a second penalty to the differences of the coefficients. Second, Meinshausen and Bühlmann (2006) showed that using only one penalty factor implies an inherent conflict between model selection and shrinkage estimation, leading to many noise variables in the final variable set. Thus, Meinshausen (2007) introduced the relaxed lasso (relaxo) with a new parameter controlling the applied shrinkage. Third, lasso introduces a potential bias for large coefficients (Fan & Li, 2001), which can be tackled by adding weights to the regularization term – i.e. the adaptive lasso (adalasso) by Zou (2006).

Building on the strengths and weaknesses of the ridge and lasso estimator and combining both penalties, Zou and Hastie (2005) introduced the elastic net (glmnet). This algorithm includes a complexity parameter controlling the strength of the regularization and a mixing parameter between ridge and lasso regression.

The next three models also belong in the class of shrinkage estimators, yet they do not alter the coefficients but the covariates themselves. Specifically, they assume that the variables can be described as a linear combination of a reduced set of factors and loadings. These are principal component regressions (pcr), independent component regression (icr) by Comon (1994) and partial least squares (pls) by Wold, Sjöström, and Eriksson (2001).

Until now the models assumed that the estimated parameters are fixed and driven by an unknown underlying data-generating process. With the upcoming models we want to venture into Bayesian territory and thereby assume that the parameters are random variables following certain distributions for which we can apply prior knowledge. As before, starting with a simple alteration of the linear model in which the coefficients follow a Student-t distribution (bayesglm) (Gelman, Jakulin, Pittau, & Su, 2008). Additionally, we introduce shrinkage via the Bayesian ridge regression (bridge) and the spike-and-slab (spikeslab) prior (Ishwaran & Rao, 2005).

Unsurprisingly, there is also a Bayesian alternative for the lasso estimator (blasso) introduced by Park and Casella (2008). The results are very similar to the original lasso algorithm, but the Bayesian treatment has the advantage that the penalty factor has not to be determined via cross-validation but can be implicitly derived in a fully Bayesian fashion. To further reduce the time-intensive and computationally demanding

---

2The corresponding backward selection procedure yields exactly the same best model subset. Using the Akaike Information Criterion (AIC) instead of the BIC, diminishes the out-sample performance as too many predictors are chosen as being relevant.
calculations, Cai, Huang, and Xu (2011) combined an empirical Bayes (EB) method with lasso to create the empirical Bayesian lasso (eblasso).

As a last subgroup of the Bayesian models, we introduce "global-local" shrinkage estimators that, as the name suggests, introduce a global shrinkage parameter pushing the coefficients uniformly towards the origin, while the local parameter allows for coefficient-specific deviations. Although there are many different priors to choose from, we settled for the Dirichlet–Laplace prior (dbayes) by Bhattacharya, Pati, Pillai, and Dunson (2015), the horseshoe prior (horseshoe) by Carvalho, Polson, and Scott (2010) and the extended horseshoe prior by Bhadra, Datta, Polson, and Willard (2017) named horseshoe+ (horseshoePlus), which exhibits significant improvements in the case of "ultra-sparse" signals (i.e. nearly all coefficients are zero).

Now coming to the last subgroup of the (generalized) linear models: ensembles of linear models. We deploy a straightforward gradient boosting algorithm (bstlm), which is based on linear models as weak learners.

Model Averaging

Variable selection, either in the way of regularization or shrinkage, can lead to an over-correction due to too many variables being penalized and thus biased estimates of the remaining covariates and too narrow confidence intervals as the inherent model uncertainty is not taken into account by focusing on only one equation (Lukacs, Burnham, & Anderson, 2010). Therefore, we introduce three model averaging models that can overcome these issues by combining multiple equations of the same base model. First, Mallow’s model averaging (mma) by Hansen (2007), which forms the final model by weighting multiple nested models based on minimized mean squared forecasting errors. Second, jackknife model averaging (jma), introduced by Hansen and Racine (2012), calculates the weights by minimizing the leave-one-out (LOO) cross-validated residuals. Third, we introduce the current state-of-the-art credit risk satellite model: Bayesian model averaging (bma) (Raftery, 1995). With this technique, the weights are based on Bayes’ theorem and the posterior model probabilities thereof.

Exponential Smoothing

Forecasts based on exponential smoothing have a long and successful track record going back to the late 1950s (Brown, 1957). We implement a simple exponential smoothing (es) algorithm with additive errors and no trend or seasonal component based on Hyndman, Koehler, Ord, and Snyder (2008). As the choice of model types is crucial for the performance of the model, Svetunkov (2016) introduced complex exponential smoothing (ces), which avoids the artificial distinction of a time series in level, trend, and seasonality.

---

3The model type for error, trend and seasonality – in our case ANN – is implicitly chosen by the sample-size-corrected AIC. This is inline with our data transformations, which include de-seasonalizing and differencing. The exogenous variables are also chosen based on this criterion.
The class of generalized additive models (GAMs) marks the point from which the presented models become more non-parametric and data-driven in their estimation routines. A feature which is often attributed to machine learning models. GAMs allow more flexibility compared to a standard linear model due to the built-in smoothing function. Additive models were originally proposed by J. Friedman and Stuetzle (1981), and the first model in this class will be the one which they proposed in their seminal paper, namely Projection Pursuit Regression (\texttt{ppr}). Furthermore, we implement boosted smoothing spline (\texttt{bstpline}), which utilizes the same gradient boosting algorithm as the linear version (i.e. \texttt{bstlm}). Lastly, a boosted generalized additive model (\texttt{boostgam}) as outlined by Schmid and Hothorn (2008) is implemented, combining the features of the first and second models.

Multivariate Adaptive Regression Splines

Another regression technique, which was proposed by Jerome Friedman, is the multivariate adaptive regression spline (\texttt{mars}) (J. Friedman, 1991). Somewhat similarly to (generalized) additive models, the setup is now fully non-parametric, consisting of linear combinations of hinge functions. This extension of linear models allows the implicit modelling of non-linearities and interactions between variables.

Support Vector Machines

Support vector machines (SVM) have a long-standing history in providing robust classifications for linear and non-linear problems. The work of Drucker, Burges, Kaufman, Smola, and Vapnik (1997) introduced the concept of support vectors to regression problems. For our model comparison, we utilize a specific SVM model called a L2-regularized L1-loss support vector regression (\texttt{svr}), which uses a linear kernel and, as the name states, performs ridge regularization on the covariates (Hsieh, Chang, Lin, Keerthi, & Sundararajan, 2008).

The second model in this class is a relevance vector machine (RVM) which has the same functional form as SVMs but uses Bayesian inference to estimate the equation (Tipping, 2001). The advantages are that, compared to SVMs, RVMs are highly sparse and no prior cross-validation is needed to tune the cost function. We deploy the model with the standard Gaussian radial basis kernel (\texttt{rvm}), which delivered the best predictive performance.

Gaussian Process

A Gaussian process (GP) regression (Williams & Rasmussen, 1996) is a Bayesian kernel machine which bridges the gap between Bayesian linear models or spline models and SVMs. A GP builds on a covariance matrix which is used to assess how much informa-
tion contiguous observations convey about each other. By applying a kernel matrix, in our case a polynomial kernel (\texttt{gp}), the estimation is extended into a non-linear realm (Karatzoglou, 2006).

**Tree-Based Models**

Similar to the class of (generalized) linear models, tree-based models also contain a multitude of different approaches, ranging from simple to more complex. The first subclass are rule-based tree structures such as classification and regression trees (\texttt{cart}) (Breiman, Friedman, Olshen, & Stone, 1984) that recursively split the data set to make predictions about the outcome variable. As a further advancement in this field, Hothorn, Hornik, and Zeileis (2006) introduced conditional inference trees (\texttt{ctree}) which tackle the issues of overfitting and selection bias prevalent in CART by introducing permutation tests during the partitioning.

In the realm of tree-based models, ensemble learning methods are more common than in the case of linear models. One of the most well-known ensemble models is random forests (\texttt{rf}) by Breiman (2001). The method combines multiple decision trees which are trained on different subsampled parts of the data and with random subsets of variables. The results of all trees are averaged across to form the final prediction (i.e. bootstrap aggregation or bagging).\footnote{We want to note that the standard implementation of bagging in the used R library \texttt{ranger} is not sensitive to time series data as it assumes iid data. Currently, there exists no available library that includes bootstrap methods for dependent data.} The low bias in the results comes at the price of large variance. In order to tackle these issues, Geurts, Ernst, and Wehenkel (2006) suggested extremely randomized trees (\texttt{ert}) which use the whole data set for each tree instead of subsampling and randomize the splitting rule for each node.

Another popular modelling technique is gradient boosting (J. Friedman, 2001), which is again an ensemble method, but in contrast to random forests, gradient boosted trees (or gradient boosting machines, GBM) are built sequentially. Each new tree improves the shortcomings of former trees, combining the results along the way. We opt for a standard Gaussian loss function (\texttt{bsttree}). It may come as no surprise that there also exists a Bayesian version of GBM called Bayesian additive regression trees (\texttt{bart}) established by Chipman, George, and McCulloch (2010). Instead of combing the trees via a learning rate, an iterative backfitting Markov chain Monte Carlo (MCMC) algorithm is used. As we will see in Section 3, this flexible setup turns out to be the winning model.

The last competitor in this class of models is called node harvest (Meinshausen, 2010) and settles itself between easy-to-understand regression trees and more accurate ensembles like random forests (\texttt{enstree}). The model delivers sparse results by initially creating random nodes and then finding suitable weights for each node based on an empirical loss function.
The last class of competitor models is one of the earliest and most commonly used techniques in machine learning: neural networks (McCulloch & Pitts, 1943). As a first model, we deploy a deep neural network (nn) with three hidden layers using resilient backpropagation with weight backtracking (Riedmiller, 1994). The last model will be again Bayesian, namely a Bayesian regularized neural network (brnn). This model fits a two-layer neural network as described in Foresee and Hagan (1997). In contrast to classical neural networks, Bayesian networks are graphical models in which each node represents a variable with probabilistic relationships among them.

2.3 Hyperparameter Tuning

The increasing complexity of the models outlined above is also reflected in the number of parameters that need to be set before a model can be estimated. These hyperparameters control complexity and are thus crucial ingredients to the overall outcome of the model. Although many machine learning libraries provide default values for most parameters, Olson, Cava, Mustahsan, Varik, and Moore (2018) showed that tuned hyperparameters can significantly reduce the variance compared to the out-of-the-box values.

There are different methods to find the most suitable set of hyperparameters (see Feurer & Hutter, 2019 for an overview), again with different layers of complexity. We choose a model-free, non-black-box method for this task: grid search. After specifying a set of values for each parameter, grid search evaluates each set combination. The drawback of grid search is the possible large number of combinations that must be evaluated. However, to tackle this problem, we combine grid search with expert judgment. Specifically, on the one hand, we conduct research on the proper parameter space for each model and, on the other hand, we manually fine tune the grid to reduce the computational burden. In combination with a highly efficient implementation by the caret library (Kuhn, 2020), the whole process stays feasible and very transparent.\(^5\)

For the performance measure we follow Hyndman and Koehler (2006) and stay within the realm of the well-known scale-dependent measures. The used data set stays the same across the models and we thus do not need to take the possible different data characteristics in consideration. Therefore, we choose the mean absolute error (MAE) as our indicator on how well a model predicts the default probability. Since some models exhibit an unstable forecasting behavior, we disregarded the more widely used mean square error (MSE) or root mean square error (RMSE) as they are more sensitive to outliers (Armstrong, 2001).\(^6\)

Finally, in order to generalize the parameters for different sample lengths and time periods, a cross-validation strategy is introduced. More precisely, we apply a rolling-origin evaluation, starting with an initial training set of \(T_{t,1} = 1, \ldots, 41\) and a fixed

\(^5\)We implement our own grid search for models which are not implemented in caret.

\(^6\)For completeness’ sake, we also implemented the RMSE, mean absolute percentage error (MAPE) and mean absolute scaled error (MASE). However, the tuning parameters and model rankings are nearly identical.
test (or holdout set) of $T_h = 12$, representing the 12 quarters we want to forecast and base our model comparison upon. In each iteration of the cross-validation, the training set is extended by one observation, leading to a total of 12 estimation rounds and a final training set of $T_{t,12} = 1, \ldots, 52$. The best performing sets of hyperparameters are saved and used in the following estimations. The fairly large initial training set is justified by our sizeable set of $n = 40$ predictors, which allow us to estimate each model without a constraint on their ability to restrict the set of predictors with regularization. As a result, to obtain the hyperparameters for 37 models, around 15,000 estimations have to be carried out.\footnote{Due to the limited data availability and the large set of predictors, we do not deploy a fixed, moving window.}

3 A Comparison of Forecasts

For the same reasons as outlined above, the estimation of the 43 models is carried out by deploying the same cross-validation strategy. The only difference is that the initial training set is $T_{t,1} = 1, \ldots, 4$. Hence, we provide for the initial estimation only one full year of data to the models. In combination with the large set of variables, this enables us to gauge the performance of the models under such extreme overfitting conditions. Using such as small data set also reflects reality, in which data availability is often limited to incomplete data or only a few recent observations. Given these circumstances, we would assume that especially regularized models and machine learning models, which are often advertised for their ability to handle such cases well (Bzdok, Altman, & Krzywinski, 2018), will fare well in the comparison.

As with the hyperparameter tuning, the MAE is again used as our main evaluation factor. After each cross-validation window, the pseudo out-of-sample forecast is computed and compared to the actual PD time series. We thereby follow the idea outlined by Hastie, Tibshirani, and Tibshirani (2017) that for the purpose of model comparisons based on conditional forecasts, we do not need to focus on the causal relationship between the variables but identify statistical dependencies that are stable over time. Thus, this section will not present any results on the estimated coefficients, but only focus on the forecasting performance.

Before we come to the evaluation of the models in terms of their predictive accuracy, we need to make sure the comparison is as fair as possible. Specifically, if models exhibit certain instabilities during the estimation or deliver no proper output at all, combining these predictions in one final score can lead to distorting effects on the overall ranking of the models.\footnote{An overview of the tuned hyperparameters per model can be found in Table 2.} Thus, if 25% of the predictions of a model fall in such a category, the whole model is dropped for further processing. As a consequence, 21 models have been dropped due process – 16 of which due to estimation instabilities and 5 could not estimate Eq. (2) in the case of $T \ll n$. Due to this process, the multivariate adaptive

\footnote{In the majority of cases the instabilities occurred due to over-regularization effects, which led to flat forecasts or algorithms not being estimable in the case of $T \ll n$.}
regression splines model has been dropped and thus the whole category, leaving 22 models in eight categories.

Figure 2 plots the evolution of the out-of-sample performance criterion across all cross-validation windows. The dashed vertical lines indicate the overfitting threshold after which sample $T$ is larger than the number of covariates $n$. At first glance, the figure indicates a significant variation of prediction accuracy across model categories, but also within the groups. Another key takeaway, without diving into the details, is that nearly all models are able to improve and stabilize forecast accuracy once the overfitting threshold is overcome. This is especially true for models which have no built-in feature to treat the overfitting problem. However, even models with such features, like BMA, struggle at first until a certain length of the training set is reached. This confirms our expectation that there is indeed a large variation in results across models and that thus a proper model comparison is needed to gain more insight into the driving factors.

Indeed, there is more to be learned from Figure 2. The group of (generalized) linear models still contains most competitors (a total of eleven). For the first 35 iterations, the models depict a somewhat similar, erratic forecasting behavior. However, from then onwards, there is a clear separation of the majority of the models towards a MAE region of 0.2, whereas the Bayesian generalized linear model shows an increasingly worse performance. In contrast, (generalized) additive models attain a more stable forecasting pattern earlier in the process. The worst predictions are obtained from exponential smoothing and the Gaussian process, in both we see a large increase in forecasting error at around the 20th iteration. Interestingly, the overall best performance stems from the tree-based models. Specifically, classical random forests and Bayesian additive regression trees are able to accurately pick up the correct signals even in a remarkable overfitting setting, thereby keeping a very stable profile across the iterations. Lastly, the categories of model averaging, support vector machines and neural networks show a somewhat similar pattern of irregularities for the first 30 to 35 iterations and stable performance afterwards. At first glance, it seems that the state-of-the-art credit risk satellite model BMA (red line) already has strong competition.

After descriptively reviewing the evolution of the out-of-sample MAE it would be difficult to state which model performed best and what the actual difference between the models would be. In order to gain more insight into the ranking of the results, we could resort to parametric statistical tests (F-test or t-test) or to their non-parametric alternatives (Friedman test (M. Friedman, 1937) or Wilcoxon signed-ranks test (Wilcoxon, 1945)). However, all of these allow only a pairwise comparison, which would mean 462 comparisons per cross-validation iteration and hence 22,638 in total. Moreover, the parametric tests rely on strong assumptions about the distribution of prediction errors, which seem to be at least partly violated for most models, as can be seen in Figure 2. Another popular method to compare the accuracy of forecast methods is the pairwise Diebold-Mariano test (Diebold & Mariano, 1995) and its multivariate alternative (Mariano & Preve, 2012). However, even with the multivariate test, we would still be left with one test per cross-validation iteration, i.e. 49 tests.
if we were to combine all models at once. Even without conducting the test, we can already assume that the equal predictive accuracy (EPA) hypothesis would not hold on such a diverse set of prediction models and thus refrain from implementing the tests.

However, the question of how to compare multiple forecasting models across multiple cross-validation runs has been tackled before. Most notably, Koning, Franses, Hibon, and Stekler (2005) ranked the results of the M3 competition using two non-parametric tests: multiple comparisons with the mean (ANOM) and multiple comparisons with the best (MCB). On the one hand, the new testing strategy allowed them to provide new insights on the statistical significance of the comparative model performance. On the other hand, both tests have received criticism due to the binary nature of the conclusion why certain models performed better or worse. Thus, Demšar (2006) introduced a third test as an alternative based on the Nemenyi test (Nemenyi, 1963). The test ranks the performance of each model across the various data sets or cross-validation iterations, averages over the ranks and produces confidence bounds. From the (non-)overlapping confidence bounds one can deduce if the models are statistically

---

Figure 2: Evolution of out-of-sample MAE

Note: The chart depicts the Mean Absolute Error (MAE) for the remaining 28 models across the 9 categories. The performance criterion is calculated for each cross-validation window. The dashed vertical lines indicate the overfitting threshold after which the sample $T$ is larger than the number of covariates $n$. The red line depicts the current state-of-the-art model Bayesian Model Averaging (bma).
different from each other.\footnote{We are aware of the possible drawbacks of using null hypothesis significance testing (see, e.g., Benavoli, Corani, Demšar, & Zaffalon, 2017 for an overview) and the Bayesian alternatives that could help with such issues (see, e.g. Calvo, Ceberio, & Lozano, 2018). However, to the best of our knowledge, there is currently no (publicly available) Bayesian hypothesis testing strategy that can handle multiple models and multiple cross-validation results at once.}

Figure 3 depicts the results of the Regression for Multiple Comparison with the Best (RMCB), which is the regression-based version of the Nemenyi test introduced by Svetunkov (2020). The test constructs a simple linear regression with the model ranks as dummy variables and uses the estimated coefficients and confidence intervals to determine the performance differences. The main difference between the Nemenyi test and RMCB is the underlying critical distance. For the former it is a studentized range distribution, while the latter uses a Student’s $t$-distribution. This leads to narrower confidence bounds for the RMCB, which can be helpful with smaller sample sizes.

The models on the x-axis in Figure 3 are sorted by the mean rank which they achieved across the cross-validation, depicted by the points in the figure, while the vertical lines reflect the confidence bounds. Thus, this test allows us to reveal the winning model: Bayesian additive regression trees (bart). However, as indicated by the dashed line, there are seven more models that are not statistically different from the winning model on a 5% significance level. Particularly, BART is closely followed by the spike-and-slab prior (spikeslab) and Random Forests (rf). The following section will provide a deep dive into BART and give more details on the estimation and results.

Within these eight best performing models, two belong to the group of tree-based models (bart, rf), four are (generalized) linear models (spikeslab, icr, lasso, pcr), one is a neural network (nn) and one is a support/relevance vector machine (rvm). Given the fairly long cross-validation period in which overfitting prevailed (36 out of 49 iterations), it is remarkable that the tree-based models, neural network and relevance vector machine are able to provide such accurate forecasts without the need for regularization. In contrast, it comes as no surprise that all of the linear models use some form of regularization to tackle the overfitting issue.

Lastly, the red indicator depicts again the current state-of-the-art satellite model among central bankers. Within our framework and given these results, we can deduce that Bayesian model averaging (bma) is significantly worse than the first eight models, i.e. the winning group. Hence, for the use case of Austrian corporate default probabilities, these eight models would, on average, deliver more precise forecasts compared to BMA.

### 3.1 Deep dive into the winning model

From the 43 implemented models we started out with and the 22 models that remained, Bayesian additive regression trees (BART) by Chipman et al. (2010) turned out victorious. The following paragraphs will introduce the model in more detail, give insights into why the model worked that well and provide detailed results.
Figure 3: Regression for Multiple Comparison with the Best

Note: Regression for Multiple Comparison with the Best (RMCB) is the regression-based version of the Nemenyi test (Demšar, 2006) introduced by Svetunkov (2020). The models on the x-axis are sorted by the mean rank which they achieved across the cross-validation results, represented by the solid dot. The vertical lines indicate the confidence bounds per model. Bayesian Additive Regression Trees (bart), marked in blue, is the best model. The red indicator depicts the current state-of-the-art model Bayesian Model Averaging (bma). All models that have intersecting confidence bounds with BART or BMA, as shown by the matching colors, are not statistically different from each other. The results are evaluated on a 5% significance level.

The framework consists of two parts, a sum-of-trees model and regularization priors on the parameters that constraint each tree. BART approximates the function $f$ from Eq. (2) by

$$f(X) \approx \sum_{j=1}^{N} g(X|\mathcal{T}_j, \mathbf{m}_j) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{1}),$$

whereas $N$ binary trees $\mathcal{T}_j$ are used, each $j^{th}$ tree with a vector of $\mathbf{m}_j = (\mu_{j1}, \ldots, \mu_{jb_j})'$ terminal nodes and $b_j$ leaves.

The second part imposes a set of regularization priors over the grown trees $p(\mathcal{T}_j)$, the model parameters $p(\mu_{ij}|\mathcal{T}_j)$ and error variance $p(\sigma)$. These priors ensure that no individual tree is too influential in the sum of trees. Specifically, we want to highlight the prior on the terminal node parameter $\mu_{ij}$, representing the effect of a tree,

$$\mu_{ij} = \mathcal{N}(0, \sigma^2_{\mu}), \quad \sigma_{\mu} = 0.5/k\sqrt{N}.$$
The prior ensures that the tree parameters are shrunken towards zero, thereby constraining each weak learner. The variable \( k \) is the prior probability that \( E(y|X) \) is within the range of \( y \). The prior standard deviation \( \sigma_{\mu} \) is related to the gradient boosting shrinkage parameter of J. Friedman (2001), which also balances the effect of each tree. For more details on the model and inference, we refer the interested reader to the original paper.

The first step in our forecasting framework is the tuning of the hyperparameters. BART, as many other machine learning models, offers the possibility to tune all prior hyperparameters. However, Chipman et al. (2010) also specify out-of-the-box parameters that work well on a range of different data sets. More precisely, one can tune the number of trees \( N \), the base \( (\alpha) \) and power \( (\beta) \) parameter for the prior \( p(T_j) \), the above-mentioned variable \( k \) for \( p(\mu_{ij}|T_j) \) and the parameters for the inverse chi-square distribution on \( p(\sigma), \nu \) and \( q \).

As outlined in Section 2.3, we use a grid of potential hyperparameters and the combinations thereof, centered around the default values by Chipman et al. (2010). Nevertheless, the default values returned the best prediction accuracy, and we will thus not go into the details of the tuning process. There is one exception for which we diverted from the default value, namely the number of trees \( N \). Chipman et al. (2010) state a default value of \( N = 200 \) as a larger number of trees increases BART’s representation flexibility and thus predictive capabilities. However, they also state that BART can be used for variable selection when the number of trees is reduced. The more trees are grown, the more irrelevant covariates are mixed with relevant ones, diminishing its selection effectiveness. When the number of trees is reduced, BART endogenously picks the more relevant variables. Given our large overfitting period in the training sample, a lower number of trees \( (N = 50) \) achieved the best results.

Until now, for the purpose of a unified model comparison, we concentrated on point forecasts and the deviation from the true default probability. This was a deliberate choice in order to focus on the prediction accuracy of each model and to keep the set of results manageable. Nevertheless, the uncertainty surrounding the point forecasts is at least as important as the forecast itself. Thus, Figure 4 shows the 12-step ahead forecast from the BART model including prediction intervals. The forecast covers the whole length of the training sample, starting after the initial four quarters as the first cross-validation iteration. The solid blue line is the estimated posterior mean, while the dark blue shaded area represents the 80\% prediction interval and the light blue area the 95\% interval. The solid black line is the actual PD time series. Plotting all cross-validation results (49 in total) would lead to many overlapping points forecasts and indistinguishable prediction intervals. We thus decided to only show five non-overlapping predictions, which nonetheless span all forecasted quarters. The vertical dotted lines indicate the predicted region. Especially in this aspect we can gauge the Bayesian inference as the predictive distribution can simply be calculated from the posterior draws, thereby incorporating the inherent parameter uncertainty.

Given our long forecasting horizon, there are some deviations between the solid blue and solid black line. Especially in the first two segments – from around 2004
Figure 4: 12-step ahead forecasts with prediction intervals

Note: the figure shows the evolution of the 12-step ahead point prediction and prediction intervals of the BART model. The PD on the y-axis is kept in the same transformation as during the estimation. The solid blue line is the estimated posterior mean, the dark blue shaded area represents the 80% prediction interval and the light blue area the 95% interval. The solid black line is the actual PD time series. In order to avoid overlapping lines and ribbons, only five out of 49 cross-validation runs are depicted. The vertical dotted lines indicate the predicted region.

to 2007 and 2007 to 2010 – BART was not able to pick up proper signals from the data to detect the surge in default probabilities in the first quarter of 2006 and during the financial crisis of 2007 and 2008. In the subsequent segments, the accuracy of the model increases as the blue line starts to trace the peaks and troughs of the actual PD. The prediction interval keeps a reasonable width over the whole time span. However, the uncertainty around the estimates is undeniable and would need special attention when being used as a credit risk satellite model.

Overall, the non-parametric Bayesian approach coupled with the adaptive weak learners seems to be a very sensitive, capable model that worked well in our setting. Given that many machine learning models depend critically on the chosen set of hyperparameters, the performance of BART with default setting is quite remarkable. The success of BART can be seen in the growing literature in various fields and the technical extensions that have been proposed. Particularly, BART with heteroscedastic errors (Pratola, Chipman, George, & McCulloch, 2020), BART in (non-linear) VAR settings (F. Huber, Koop, Onorante, Pfarrhofer, & Schreiner, 2020), multinominal logistic regression via BART (Murray, 2017) and survival models (Sparapani, Logan, McCulloch, & Laud, 2016). A more detailed overview of the technical extensions can be found in Hill, Linero, and Murray (2020).
3.2 Forecast Combinations

As we have seen in last section, even the best models are fraught with uncertainty regarding the point prediction. This inherent uncertainty regarding econometric models is undisputed in the literature and relates to unknown data generating processes, misspecified models and a generally complex reality the models try to replicate. In order to hedge against such uncertainties, Bates and Granger (1969) introduced in their seminal paper the concept of forecast combinations. The idea is straight-forward: there is no one true model, there are only different approximations of the data generating process. The underlying models have their own strength and weaknesses, which, when combined, should yield an overall better forecast. Even though such forecast combinations can be hard to interpret in terms of marginal effectiveness of the coefficients, we can again take up the point made by Hastie et al. (2017) that especially with comparisons among prediction models, a stable statistical dependency outweighs the underlying causal relationships.

In the last five decades since Bates and Granger (1969), a wide range of combination methods have been suggested. For the purpose of this paper, we focus on three groups: simple combination methods, regression-based combination methods and eigenvector-based combination methods. For the simple methods, we choose the naive average, which weights all models equally (AVG) and the Newbold/Granger method (Newbold & Granger, 1974), which calculates the weights from the estimated mean squared prediction error (MSPE) matrix (NG). The second group is still based on linear functions of the individual forecasts, but the weights are determined using a constrained least squares (CLS) regression. Finally, the standard eigenvector-based approach by Hsiao and Wan (2014) uses, unlike Newbold and Granger (1974), a normalization condition that leads to an unconstrained minimum of the MSPE (SEA). These methods (and more) have been implemented by Weiss, Raviv, and Roetzer (2018).

Moreover, we conduct the analysis with three different scenarios: first, in a naive approach we combine all 22 models; second, we only use the winning eight models which have been determined by the Nemenyi test in Figure 3; third, we combine the best model of each category, as outlined in section 2.2.

Figure 5 shows, similar to Figure 2, the evolution of the mean absolute errors (MAE) across the cross-validation iterations. The four panels reflect the forecast combinations group outlined above with the three different model scenarios – all models (green line), top eight winning models (purple line) and the best models of each category (orange line). In order to properly frame the results of the forecast combinations, we additionally plot the performance of BART (blue line) and BMA (red line). First of all, the sole comparison between BMA and BART again emphasizes the significant increase in terms of predictive accuracy that BART delivers. However, while the simple average

---

11 Besides a wide array of combination method, the implementation by Weiss et al. (2018) also allows a dynamic version of combinations, which is related to the idea of time series cross-validation. However, in our setting the normal static version achieved better results than the dynamic version.

12 Using all 22 models with the Newbold/Granger and the eigenvector-based method leads to mathematical problems and is thus excluded.
Figure 5: Evolution of out-of-sample MAE for forecast combinations

Note: The chart depicts the Mean Absolute Error (MAE) for BART (blue line), BMA (red line) and the forecast combination methods: the naive average (AVG), the Newbold/Granger method (NG), the constrained least squares approach (CLS) and the standard eigenvector-based approach (SEA). The forecast combination have been calculated for the scenario using all 22 models (green line), utilizing only the top 8 winning models (purple line) and for the best model per group (orange line).

Method (AVG) is not able to fully outperform BART, the constrained least squares version (CLS) is able to combine the underlying models in a way that beats BART across nearly all cross-validation iterations. The difference becomes even clearer in the case of the Newbold/Granger method (NG) and the standard eigenvector-based approach (SEA). While the eigenvector approach depicts a somewhat erratic behavior, driven by the combined models, the Newbold/Granger approach delivers a significant improvement across all cross-validation iterations.

In order to get a clearer picture of the results, Table 1 shows the average out-of-sample MAE and ranks for BART, BMA and the forecast combinations, calculated over the cross-validation iterations. The numbers in parentheses indicate the average forecasting accuracy and rank relative to BART. On average, BART is able to beat two of the three simple average scenarios and is nearly on par with the top eight scenarios. However, all other combination methods are able to improve the average prediction accuracy significantly. The best accuracy is achieved by the Newbold/Granger method using the group-wise best performing models. Although this setup includes rather badly performing models, such as exponential smoothing and the Gaussian process,
|                  | bart | bma | AVG all models | AVG top 8 | AVG top group | NG top 8 |
|------------------|------|-----|----------------|----------|---------------|----------|
| Avg. MAE         | 0.20 | 0.29| 0.22           | 0.20     | 0.22          | 0.12     |
|                  | (1.00)| (0.70)| (0.90)       | (1.01)   | (0.91)        | (1.62)   |
| Avg. Rank        | 8.57 | 10.94| 10.02          | 8.10     | 8.78          | 2.55     |
|                  | (1.00)| (0.78)| (0.86)       | (1.06)   | (0.98)        | (3.36)   |
| NG top group     | 0.11 | 0.17| 0.19           | 0.18     | 0.17          | 0.17     |
|                  | (1.83)| (1.18)| (1.07)       | (1.12)   | (1.20)        | (1.20)   |
| Avg. Rank        | 1.49 | 4.14| 6.20           | 5.02     | 6.35          | 5.84     |
|                  | (5.75)| (2.07)| (1.38)       | (1.71)   | (1.35)        | (1.47)   |

Note: The table shows the average Mean Absolute Error (MAE) and average rank across the cross-validation results for BART, BMA and the forecast combination methods. The values in the parenthesis are the errors and ranks relative to BART – values above 1 indicate a better performance.

the Newbold/Granger approach is able to calculate the weights in a way that extract only the positive features from the underlying models.

### 4 Conclusion

Since the Great Financial Crisis, the use of stress tests as a tool for assessing the resilience of financial institutions to adverse financial and economic developments has increased significantly. One key part in such exercises is the translation of macroeconomic variables into default probabilities for credit risk by using macrofinancial linkage models. A key requirement for such models is that they should be able to properly detect signals from a wide array of macroeconomic variables in combination with a mostly short data sample.

The current state-of-the-art satellite model for PD translation is Bayesian model averaging (BMA) (Raftery, 1995). It has a long track record as being a reliable tool for generating scenario-conditional projections for credit risk and is being adopted by more and more central banks and institutions. However, with the easier access to regression models and the advent of new predictive models in the field of machine learning, the question arises if there are other models that could deliver better results.

The aim of this paper is to conduct a systematic forecast comparison with a large number of different regression models to find the best performing credit risk satellite model. The best model is evaluated for the ability to precisely forecast default probabilities conditional on a standardized set of macroeconomic variables as provided to financial institutions by the ESRB (2020) for the EU-wide banking sector stress test. We implement a total of 43 models which we assigned to 9 categories, ranging from
conventional statistical models to more recent machine learning methods. Additionally, we combine the models with different forecast combination approaches to further gauge their potential accuracy. For the purpose of this paper, we implement a framework that allows us to conduct this comparison with a standardized data set for all models, to tune the respective hyperparameters for each model and to cross-validate the results based on recursive pseudo out-of-sample forecasts. The data used in the modelling exercise refer to Austria and include as dependent variable a measure for the probability of default for the non-financial corporate sector and macroeconomic and financial data as independent variables.

Our results indicate that there are indeed better performing models than the current state-of-the-art model. Specifically, a group of eight models significantly outperforms BMA in terms of their capability to forecast default probabilities. Five of these eight models belong to the category of machine learning models. Among these models, there is also the overall winner of the model comparison: Bayesian additive regression trees (BART) by Chipman et al. (2010). The combination of a flexible sum-of-trees model and well calibrated regularization priors gives BART the advantage needed to outperform all other models, especially in situations where overfitting is prevalent. Additionally, given that the majority of winning models has not been explicitly covered in the literature yet, our comparison sheds light on potential other credit risk models to be further investigated. We specifically highlight the advantages of machine learning models in the context of default probability prediction and more generally their applicability in high dimensions where overfitting prevails. Lastly, as most forecast combinations even outperform BART, we show that simple combination techniques can help to further hedge against model uncertainty and boost predictive accuracy.
# Appendix

## Table 2: Overview of implemented models and used R libraries

| Base Model       | Extension                     | Model Name            | Short Name | Libraries       | Parameter          |
|------------------|-------------------------------|-----------------------|------------|-----------------|--------------------|
| Linear           |                               | Linear Regression     | lm         | stats           | intercept          |
| Robust           |                               | Robust Linear Model   | rlm        | MASS            | intercept, psi     |
| Regularized      |                               |                       |            |                 |                    |
| (Generalized)    |                               | Linear Regression with Forward Selection | lmfs       | leaps           | nvmax              |
| Linear Models    | Least Angle Regression        | lars                  | lars       | lars            | fraction           |
| Regularized      | Ridge Regression              | ridge                 | glmnet     |                 |                    |
|                  | Least Angle Regression        | lars                  | lars       | lars            | fraction           |
|                  | Ridge Regression              | ridge                 | glmnet     |                 |                    |
|                  | Least Angle Regression        | lars                  | lars       | lars            | fraction           |
|                  | Adaptive Lasso                | adalasso              | glmnet, HDecomometrics, glmnet | crit             |
| Bayesian         | Relaxed Lasso                 | relaxo                | relaxo     | lambda, phi     |                    |
|                  | Penalized GLM                 | glmnet                | glmnet     | alpha, lambda   |                    |
| Feature Extraction| Independent Component Regression | icr                  | fastICA    | n.comp          |                    |
|                  | Principal Component Analysis  | pca                   | pls        | ncomp           |                    |
|                  | Partial Least Squares         | pls                   | pls        | ncomp           |                    |
| Bayesian         | Bayesian GLM                  | bglm                  | arm        |                 |                    |
|                  | Bayesian Ridge Regression     | bridge                | monomvn    |                 |                    |
|                  | Spike and Slab Regression     | spikeslab             | spikeslab  | vars            |                    |
|                  | The Bayesian lasso            | blasso                | monomvn    |                 |                    |
| Base Model | Extension | Model Name | Short Name | Libraries | Parameter |
|------------|-----------|------------|------------|-----------|-----------|
| Empirical Bayesian Lasso | eblasso | eblasso | a, b |
| Dirichlet Laplace shrinkage prior | dlbayes | dlbayes | - |
| Horseshoe Prior | horseshoe | horseshoe | method.tau |
| Horseshoe+ Prior | horseshoe+ | bayesreg | - |
| Ensembles Boosted Linear Model | bstlm | bst | mstop, nu |
| Model Averaging | Linear Mallow’s Model Averaging | mma | mami | - |
| Jackknife Model Averaging | jma | mami | - |
| Bayesian Bayesian Model Averaging | bma | BMS | - |
| Exponential Smoothing | Linear Exponential Smoothing | es | smooth | ic, xregDo |
| Complex Exponential Smoothing | ces | smooth | initial, ic, xregDo |
| (Generalized) Additive Models | Feature Extraction Projection Pursuit Regression | ppr | stats | nterms |
| Boosted Smoothing Spline | bstpline | bst | mstop, nu |
| Boosted GAM | boostgam | mboost | mstop, prune |
| Multivariate Adaptive Regression Splines | Non-parametric MARS | mars | earth | nprune, degree |
| Support Vector Machines | Regularized SVM with Linear Kernel | svm | LiblineaR | cost, Loss |
| Bayesian RVM with Gaussian Kernel | rvm | kernlab | - |
| Gaussian Process | Bayesian GP with Polynomial Kernel | gp | kernlab | degree, scale |
| Base Model          | Extension | Model Name                      | Short Name | Libraries     | Parameter               |
|--------------------|-----------|---------------------------------|------------|---------------|-------------------------|
| Tree-Based Model   | Trees     | Classification and Regression Trees | cart       | rpart         | cp                      |
|                    |           | Conditional Inference Tree      | ctree      | party         | maxdepth, mincriterion  |
|                    | Bayesian  | Bayesian Additive Regression Trees | bart       | bartMachine   | num_trees, kvar, alpha, beta, nu |
|                    | Ensembles | Tree-Based Ensembles            | enstree    | nodeHarvest   | maxinter, mode          |
|                    |           | Gradient Boosted Tree           | bsttree    | bst           | mstop, maxdepth, nu     |
|                    | Random Forest | Random Forest                  | rf         | ranger, e1071 | mtry, min.node.size     |
|                    |           | Extremely randomized trees      | ert        | ranger        | mtry, splitrule, min. node size |
| Neural Network     | Linear    | Neural Network                  | nn         | neuralnet     | layer1, layer2, layer3  |
|                    | Bayesian  | Bayesian NN                     | brnn       | brnn          | neurons                 |
References

Altman, E. I., & Bana, G. (2003). Defaults and returns on high yield bonds: The year 2002 in review and the market outlook.

Alves, I. (2005). Sectoral fragility: factors and dynamics. *Investigating the relationship between the financial and real economy*, 22, 450–80.

Armstrong, J. S. (2001). Evaluating forecasting methods. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 443–472). Springer.

Aver, B. (2008). An empirical analysis of credit risk factors of the Slovenian banking system. *Managing Global Transitions*, 6(3), 317–334.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4), 451–468.

Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research*, 18(1), 2653–2688.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1), 65–98.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4), 1105–1131.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512), 1479–1490.

Bofendi, M., & Ropele, T. (2011). Macroeconomic determinants of bad loans: evidence from Italian banks. *Bank of Italy Occasional Paper* (89).

Breeden, J. L. (2020). A survey of machine learning in credit risk.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth.

Brown, R. (1957). Exponential smoothing for predicting demand. In *Operations research* (Vol. 5, pp. 145–145).

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature methods*, 15(4), 233–234.

Cai, X., Huang, A., & Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC bioinformatics*, 12(1), 1–13.

Calvo, B., Ceberio, J., & Lozano, J. A. (2018). Bayesian inference for algorithm ranking analysis. In *Proceedings of the genetic and evolutionary computation conference companion* (pp. 324–325).

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465–480.

Castren, O., Dees, S., & Zaher, F. (2010). Stress-testing euro area corporate default probabilities using a global macroeconomic model. *Journal of Financial Stability*, 6(2), 64–78.

Castro, V. (2013). Macroeconomic determinants of the credit risk in the banking system: The case of the GIPS. *Economic Modelling*, 31, 672–683.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3), 287–314.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, 1–30.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a), 427–431.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3), 253–263.

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression
machines. *Advances in neural information processing systems*, 9, 155–161.

ECB. (2020). *The definition of price stability*. Retrieved 2020-11-06, from https://www.ecb.europa.eu/mopo/strategy/pricestab/html/index.en.html

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of statistics*, 32(2), 407–499.

Elliott, G., Rothenberg, T. J., & Stock, J. H. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, 64(4), 813–836.

ESRB. (2020). *Macro-financial scenario for the 2020 EU-wide banking sector stress test*. Retrieved 2020-11-06, from https://www.esrb.europa.eu/mppa/stress/shared/pdf/esrb.stress.test200131-09dbe748d4.en.pdf

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348–1360.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In *Automated machine learning: methods, systems, challenges* (pp. 3–33).

Foresee, D., & Hagan, M. (1997). Gauss-newton approximation to bayesian learning. In *Proceedings of international conference on neural networks (icnn’97)* (Vol. 3, pp. 1930–1935).

Friedman, J. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376), 817–823.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American statistical association*, 32(200), 675–701.

Gambera, M. (2000). *Simple forecasts of bank loan quality in the business cycle* (Vol. 230). Federal Reserve Bank of Chicago Chicago, IL.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of applied Statistics*, 2(4), 1360–1383.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3–42.

Gross, M., & Población, J. (2019). Implications of model uncertainty for bank stress testing. *Journal of Financial Services Research*, 55(1), 31–58.

Grundke, P., Pliszka, K., & Tuchscherer, M. (2019). Model and estimation risk in credit risk stress tests. *Review of Quantitative Finance and Accounting*, 1–37.

Hansen, B. (2007). Least squares model averaging. *Econometrica*, 75(4), 1175–1189.

Hansen, B., & Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1), 38–46.

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

Hill, J., Linero, A., & Murray, J. (2020). Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and Its Application*, 7, 251–278.

Hocking, R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 1–49.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), 651–674.

Hsiao, C., & Wan, S. K. (2014). Is there an optimal forecast combination? *Journal of Econometrics*, 178, 294–309.

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S., & Sundararajan, S. (2008). A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on machine learning* (pp. 408–415).

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., & Schreiner, J. (2020). Nowcasting in a pandemic using non-parametric mixed frequency vars. *Journal of Econometrics*.

Huber, P. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics* (pp. 492–518). Springer.
Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International journal of forecasting, 22*(4), 679–688.

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

IMF. (2020). *Publication of Financial Sector Assessment Program Documentation - Technical Note on Financial Stability Analysis, Stress Testing, and Interconnectedness*. Retrieved 2020-11-06, from https://www.imf.org/~/media/Files/Publications/CR/2020/English/1AUTEA2020006.ashx

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of statistics, 33*(2), 730–773.

Jacobs, M. (2018). The validation of machine-learning models for the stress testing of credit risk. *Journal of Risk Management in Financial Institutions, 11*(3), 218–243.

Karatzoglou, A. (2006). *Kernel methods software, algorithms and applications* (Unpublished doctoral dissertation).

Keeton, W. R., & Morris, C. S. (1987). Why do banks’ loan losses differ. *Economic review, 72*(5), 3–21.

Kerbl, S., & Sigmund, M. (2011). What Drives Aggregate Credit Risk? *Oesterreichische Nationalbank Financial Stability Report, 22.*

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting, 21*(3), 397–409.

Koopman, S. J., & Lucas, A. (2005). Business and default cycles for credit risk. *Journal of Applied Econometrics, 20*(2), 311–323.

Kuhn, M. (2020). *caret: Classification and Regression Training* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-86)

Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics, 54*(1-3), 159–178.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks, 7*(1), 29.

Lukacs, P., Burnham, K., & Anderson, D. (2010). Model selection bias and Freedman’s paradox. *Annals of the Institute of Statistical Mathematics, 62*(1), 117.

Mariano, R. S., & Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of econometrics, 169*(1), 123–130.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115–133.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis, 52*(1), 374–393.

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics, 2049–2072.

Meinshausen, N., & Bühlmann, P. (2006). Variable selection and high-dimensional graphs with the lasso. *Ann Stat, 34*, 1436–1462.

Murray, J. S. (2017). Log-linear bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503, 3.*

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.

Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General), 137*(2), 131–146.

Olson, R. S., Cava, W. L., Mustahsan, Z., Varik, A., & Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. In *Pacific symposium on biocomputing 2018: Proceedings of the pacific symposium* (pp. 192–203).

Papadopoulos, G., Papadopoulos, S., & Sager, T. (2016). *Credit risk stress testing for eu15 banks: a model combination approach* (Tech. Rep.).

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association, 103*(482), 681–686.

28
Pesaran, M. H., Schuermann, T., Treutler, B.-J., & Weiner, S. M. (2006). Macroeconomic dynamics and credit risk: a global perspective. *Journal of Money, Credit and Banking*, 1211–1261.

Pesola, J. (2001). The role of macroeconomic shocks in banking crises. *Bank of Finland discussion paper* (6).

Phillips, D. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1), 84–97.

Phillips, P., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2), 335–346.

Pratola, M. T., Chipman, H. A., George, E., & McCulloch, R. (2020). Heteroscedastic bart via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2), 405–417.

R Core Team. (2020). R: A Language and Environment for Statistical Computing [Computer software manual].

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Riedmiller, M. (1994). Rprop-description and implementation details.

Robert, C., William, C., & Irma, T. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of official statistics*, 6(1), 3–73.

Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*, 53(2), 298–311.

Sparapani, R. A., Logan, B. R., McCulloch, R., & Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*, 35(16), 2741–2753.

Svetunkov, I. (2016). *Complex exponential smoothing*. Lancaster University (United Kingdom).

Svetunkov, I. (2020). greybox: Toolbox for Model Building and Forecasting [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=greybox (R package version 0.6.0)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. akad. nauk sssr* (Vol. 39, pp. 195–198).

Tipping, M. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun), 211–244.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Weiss, C. E., Raviv, E., & Roetzer, G. (2018). Forecast Combinations in R using the ForecastComb Package. *R Journal*, 10(2).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

Williams, C., & Rasmussen, C. E. (1996). Gaussian processes for regression.

Wilson, T. C. (1998). Portfolio credit risk. *Economic Policy Review*, 4(3).

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109–130.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301–320.