

## Bayesian credit ratings: A random forest alternative approach

Imad Bou-Hamad

To cite this article: Imad Bou-Hamad (2017) Bayesian credit ratings: A random forest alternative approach, Communications in Statistics - Theory and Methods, 46:15, 7289-7300, DOI: [10.1080/03610926.2016.1148730](https://doi.org/10.1080/03610926.2016.1148730)

To link to this article: <https://doi.org/10.1080/03610926.2016.1148730>



Published online: 13 Apr 2017.



Submit your article to this journal [↗](#)



Article views: 353



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)

# Bayesian credit ratings: A random forest alternative approach

Imad Bou-Hamad

Olayan School of Business, American University of Beirut, Beirut, Lebanon

## ABSTRACT

Cerciello and Giudici (2014) proposed a Bayesian approach to improve the ordinal variable selection in credit rating assessment. However, no comparison has been made with other methods and the predictive power was not tested. This study proposes an integrated framework of random forest (RF)-based methods and Bayesian model averaging (BMA) to validate and investigate the ordinal variable importance in evaluating credit risk and predicting default in greater depth. The proposed approach was superior to the Cerciello and Giudici method in terms of predictive accuracy and interpretability when applied to a European credit risk database.

## ARTICLE HISTORY

Received 20 April 2015  
Accepted 25 January 2016

## KEYWORDS



Bayesian model averaging;  
credit risk; default; random  
forests; variable selection.

## MATHEMATICS SUBJECT CLASSIFICATION

62H30

## 1. Introduction

Traditionally, credit risk has been assessed by experts according to the 5 Cs criteria: Consumer Character, Capital, Collateral, Capacity and the Condition of the economy (Wang et al., 2012). As markets expand, subjective assessment of credit risks can no longer be consistent, which has created a need for more scientific methods. In the wake of the most recent financial crisis, financial firms sought to gain a firmer grasp of risk evaluation. This boosted the demand for conventional statistical methods, and created opportunities for more innovative ones, such as artificial intelligence (AI). Along with technological advancement, different techniques of classification and prediction have been employed. The renewed interest in statistics was driven mostly by the desire to improve accuracy, as a slight improvement in predictions and classification drastically affected revenues. A fundamental task in credit scoring is also the selection of appropriate predictors that relate to default. Different methods like Stepwise regression, factor analysis, clustering and partial least square have been explored for variable selection in credit scoring (Leung, 2008), and considerable research has shown the relevance of financial factors to predict default (see Altman, 1968 for example). However, the role of non financial factors, in particular soft information variables such as management quality and industry, in predicting default remains vague, and there is a lack of quantitative research on this matter (Grunert et al., 2005). Recently, Cerciello and Giudici (2014) proposed a Bayesian non parametric approach to investigate the importance of soft information variables. They compared parametric and non parametric Bayesian approaches, where in both cases a priori distribution for the unobserved risk level of default “ $\theta$ ” is assumed. The parametric approach introduces another assumption about the parametric distribution of the covariates, given the value of  $\theta$ ,

**CONTACT** Imad Bou-Hamad  [ib12@aub.edu.lb](mailto:ib12@aub.edu.lb)  Olayan School of Business, American University of Beirut, P.O. Box: 11-0263, Riad El-Solh 1107-2020, Beirut, Lebanon.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/lsta](http://www.tandfonline.com/lsta).

whereas the non parametric approach uses an empirical distribution, conditional on  $\theta$ . Moreover, the predictive performance of the Bayesian model proposed by Cerciello and Giudici (2014) (hereinafter referred to as C&G) has not been explored. Additionally, no comparison has been made with other advanced techniques such as random forest (RF)-based models. We believe that, to the best of our knowledge, RF-based models have not been deeply explored in credit scoring, especially conditional inference random forest (CI-RF). These ensemble methods are popular predictive tools that handle complex data structure with few or no statistical assumptions and delivering high predictive accuracy (Strobl et al., 2008). Despite their high predictive power, RFs lack interpretability of how the predictor variables affect the dependent variable or the outcome. However, the variable importance measures yielded by RFs that are later developed for identifying relevant predictor variables could partially handle the problem of interpretability. Yet for some credit risk practitioners and analysts, interpretability is often as important as prediction accuracy for gaining a better understanding of the relationship between the predictor variables and the default (outcome). In this paper, I propose an integrated model of RFs and Bayesian model averaging (BMA) for logistic regression. In addition to the important predictor variables that the proposed framework suggests, its BMA component serves as a transparent credit scoring model that will be compared to the one proposed by C&G (2014). Thus, the motivation to conduct this research study is to (1) explore in greater depth the importance of soft information variables by using methods with maximal accuracy and minimal number of assumptions such as RFs; (2) validate their role in predicting default; and (3) compare the predictive performance of my final model to the Bayesian one proposed by C&G (2014). The paper is structured as follows. In Sec. 2, an overview of credit scoring-related literature is provided. In Sec. 3, I present the proposed framework and describe its components that include the data mining and statistical methods used in this study. The experimental design and data are presented in Sec. 4. In Sec. 5, I report and discuss the results. The last section draws conclusions and suggests future research directions.

## 2. Overview of credit scoring and related models

Credit scoring is the process of risk assessment associated with loaning to a client, whether individuals or enterprises. The individual credit score is calculated using variables like applicant age, income, and credit bureau variables. On the other hand, the enterprise credit score relies on audited financial variables and some internal or external variables, including credit bureau ones. The purpose of credit scoring is to classify clients into two types: those with good credit rating and those with bad credit rating. Clients with a good credit rating have a higher likelihood of repaying their financial obligations. However, poorly rated credit clients carry a major risk of defaulting (Wang et al., 2012). Many successful applications of credit scoring models have been reported in previous studies (Lewis, 1992; Bailey, 2001; Mays, 2001; Malhotra and Malhotra, 2003; Siddiqi, 2006; Chuang and Lin, 2009; Sustersic et al., 2009) and were regarded as an assessment tool for numerous institutions in different regions over the past few decades. Since the estimated credit scoring models only consider variables that statistically show significant correlation with repayment performance, the credit scoring decision-making process is relatively straightforward. However, judgment and decisions making, at first sight, have no statistical meaning since they are subjective and depend on the experience of the credit analyst; thus, variable reduction methods are not applicable in this case (Crook, 1996; Abdou and Pointon, 2011).

The main disadvantages of credit scoring models are indirect discrimination and misclassification, where in some cases both good and bad clients can have very similar characteristics

within the considered variables. Also, sometimes economic variables are not included in such models since individual variables are mostly used (Crook, 1996). Other criticisms include subjective choice of the cutoff point in order to classify a customer default or non default, as well as the limitation of the model to the historical data (Capon, 1982; Heffernan, 2005).

Conventional scoring models are built by employing a wide range of statistical techniques, such as linear discriminant analysis (Reichert et al., 1983; Karels and Prakash, 1987), regression analysis, probit analysis, and logistic regression. They are then evaluated (Orgler, 1971; Boyes et al., 1989; Steenackers and Goovaerts, 1989; Leonard, 1992; Greene, 1998; Banasik et al., 2001; Sarlija et al., 2004) and compared to other methods (Guillen and Artis, 1992; Greene, 1998; Banasik et al., 2003; Abdou, 2009). Other techniques used in credit scoring are Logistic Regression Analysis (LRA) (Thomas, 2000; West, 2000), mathematical programming, non parametric smoothing methods, Markov chain models, expert systems (Hand and Henley, 1997), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991), and others. The drawbacks of applying conventional statistical techniques quickly emerge; some assumptions, namely the normality of independent variables, are violated in the context of finite samples (Huang et al., 2004).

More recently, researchers have leaned toward the AI and machine learning (ML) techniques such as Artificial Neural Networks (ANN) (Trippi and Turban, 1993; Bishop, 1995; Masters, 1995; Desai et al., 1996; Gately, 1996; Reed and Marks, 1999; West, 2000; Dimla and Lister, 2000; Stefanowski and Wilk, 2001; Lee et al., 2002; Kim and Sohn, 2004; Zekic-Susac et al., 2004; Lee and Chen, 2005; Yim and Mitchell, 2005; Blöchlinger and Leippold, 2006; Seow and Thomas, 2006; Trinkle and Baldwin, 2007), Decision Trees (DT) (Makowski, 1985; Hung and Chen, 2009) a.k.a. recursive partitioning (Hand and Henley, 1997) or classification and regression trees (CART) (Breiman et al., 1984; Thomas, 2000; Fritz and Hosemann, 2000; Stefanowski and Wilk, 2001; Baesens et al., 2003), and Support Vector Machine (SVM) (Baesens et al., 2003; Huang et al., 2007; Schebesch and Stecking, 2005) and their combinations as innovative methods for credit scoring. The advantage of AI and ML methods is that they allow the data to talk without any prior assumptions for its distribution. According to Huang et al. (2004), these methods provide better-quality models for credit scoring, particularly when dealing with a non linear classification.

Although more accurate than conventional methods, several meta-analysis and comparative papers have shown that a clear map pinpointing the “best” AI or ML technique to build a credit scoring model has yet to be reached (Fritz and Hosemann, 2000; Lee et al., 2002; Malhotra and Malhotra, 2003; Zekic-Susac et al., 2004; Lee and Chen, 2005; Ong et al., 2005; Abdou et al., 2008; Abdou and Pointon, 2009; Abdou and Pointon, 2011; Wang et al., 2012). However, at least it is widely known that the best model would be bound by particular factors, such as the data structure, variable selection, and choice of cutoff point, problem details, and classification objective (Hand and Henley, 1997; Yu et al., 2008). In addition, it has been shown that for credit scoring models, AI ensemble methods achieve better accuracy than a single classifier (Yu et al., 2008; Hung and Chen, 2009).

### 3. Integrated framework

The proposed framework is a two-stage procedure consisting of a hybrid model of RFs and BMA for logistic regression. In the first stage, the framework identifies the important predictor variables by means of two types of RFs, namely, RF built using classification and regression trees (RF-CART) and the other type using conditional inference trees (RF-CI). The second stage consists of estimating a transparent and an explanatory model for credit scoring based on the selected predictor variables by means of BMA logistic regression. I describe below the

DTs and RFs employed in this study as well as the BMA for logistic regression. In addition, the Bayesian-based approach for credit rating proposed by C&G (2014) is briefly presented.

### **3.1. DTs: CART and CIT**

DTs are one of the most important data mining tools used to classify and predict a response or an outcome of interest. Their structure and ease of interpretability make them very useful in many business fields. DTs first appeared with Morgan and Sonquist (1963), but witnessed an increase in their popularity with Breiman et al. (1984). They are labeled classification trees when the response is categorical (here default and non default) and regression trees when the response is numerical. A classification tree is built by recursively partitioning the initial data set (root node) into more homogenous groups (or nodes) with respect to the response. Each split is based on logical if-then conditions on the predictors and selected according to a splitting criterion. Once the final tree is selected, the response for an observation can be classified or predicted by following the path from the root to the terminal nodes. The predicted value will be the class with the highest proportion in that terminal node. The first type of DT used in this study is the CART proposed by Breiman et al. (1984). The splitting criterion is based on maximum reduction in overall node impurity. The impurity of a group of observations or instances is designed to capture how different the instances are from each other. In the case of classification trees, CART uses the Gini index to measure impurity; however, growing a very large tree tends to overfit the data. To avoid this problem, Breiman et al. (1984) proposed a pruning algorithm based on cost complexity. The pruning algorithm generates a sequence of sub-trees from the full tree by eliminating branches that provide little information to classify instances. From this sequence of nested trees, an optimal tree will be selected as the final model. However, RF based on CART trees is grown without pruning.

The second type of tree employed here is the conditional inference trees (CIT) proposed by Hothorn et al. (2006b). These trees use a splitting criterion based on multiplicity-adjusted conditional tests (Hothorn et al., 2006a) rather than impurity reduction. For any node, the splitting procedure consists of conducting a global permutation test of no link between any predictor and the response within the node. If this global hypothesis of no link is not rejected, the node is not split further and remains as a terminal node. Otherwise, for each predictor, a null hypothesis of no association with the response will be conducted. The predictor with the lowest  $p$ -value is selected for splitting. In CIT, pruning is not required since the trees stop growing when the split is not statistically significant.

### **3.2. Random forests**

An RF is a non parametric data mining technique developed by Breiman (2001) and derived from CARTs. RF uses a combination of bootstrap aggregation (Breiman, 1996) and the random subspace method (Ho, 1998) at each split to generate multiple classes or trees. Each split of the tree is determined using a randomized subset of the predictors at each node and the final outcome is the average of the results of all the trees (Breiman, 2001). When the outcome is a class (here default or non default), the final prediction is the class with maximum votes. Hence, RF involves a combination of many trees, all built independently and generated by bootstrap samples, leaving about a third of the overall sample for validation (the out-of-bag predictions, OOB).

Fundamentally, RF selects a random subset of variables at each node of each tree, and only those variables are used to find the best split for the node. It uses two-step randomization to de-correlate the generated trees so that the forest ensemble will have low variance. The

first randomization is introduced to grow the tree using a bootstrap sample of the original data, whereas the second randomization is introduced at the node level when growing the tree. Typically, introducing the two-step randomization to RF trees yields lower bias because it grows deep trees without pruning. Recently, Strobl et al. (2007) proposed an RF based on conditional trees. Thus I refer by RF-CART and RF-CI to the two types of RFs based on CART and CITs, respectively.

The construction of RF is described in the following steps (Breiman, 2001):

- Draw  $ntree$  bootstrap samples from the original data;
- Grow a tree for each bootstrap dataset. At each node of the tree, randomly select a number of predictor variables for splitting. Grow the tree so that each terminal node has no fewer than node size cases;
- Aggregate information from the  $ntree$  trees for new data prediction such as majority voting for classification; and
- Compute an OOB error rate by using the data not in the bootstrap sample.

Both RF-CART and RF-CI provide variable importance measures that could be used to rank the predictors according to their association with the response. The most popular and reliable criterion is the decrease of classification accuracy when values in a node of tree are randomly permuted in the OOB sample for that tree (Breiman, 2001).

The variable importance measure for a predictor  $X_j$  in tree is defined as follows:

$$VI^{(t)}(x_j) = \frac{1}{ntree} \sum_{t=1}^{ntree} (\tilde{e}_{tj} - e_{tj}), \quad (1)$$

where  $ntree$  is the number of trees in the forest,  $\tilde{e}_{tj}$  and  $e_{tj}$  denote the means error rates over all OOB observations in tree  $t$  after and before permuting predictor  $X_j$ , respectively.

### 3.3. Bayesian model averaging for logistic regression

Logistic regression is widely used to predict or explain categorical outcomes, in particular a binary outcome whose probability is related to a set of predictor variables  $X_1, \dots, X_q$ . The standard logistic model is an equation of the form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q,$$

where  $p$  is the probability of the outcome of interest (in our case it is the default) and the  $\beta$  coefficients are the model parameters to be estimated (Hosmer and Lemeshow, 1989). However, logistic regression is a standard statistical approach that ignores the model uncertainty arising from the assumption that the outcome is related to the predictor variables by a single and prespecified structure model ( $M$ ). Moreover, it doesn't take into account the uncertainty in the parameter estimates. Ignoring uncertainty can impair the predictive power and lead to overconfident inferences and risky decisions (Hoeting et al., 1999; Hoeting, 2002). This criticism can be overcome by adopting the BMA approach (Draper, 1995; Chatfield, 1995).

The BMA estimate of a parameter  $\beta$  is

$$\hat{\beta}_{BMA} = \sum_{K=1}^K \hat{\beta}_k p(M_k|D)$$

where  $\hat{\beta}_k$  is a parameter estimate of model  $M_k$ ,  $p(M_k|D)$  is the posterior probability for model  $M_k$  given data  $D$ , and  $K$  is the number of models considered. For full details about BMA, I refer the reader to Hoeting et al. (1999). The implementation of BMA for generalized linear models became available recently in R language within the Package “BMA.”

### 3.4. C&G Bayesian approach

Recently, Cerciello and Giudici (2014) used a Bayesian approach to predict a default probability of firms, based on a collection of “K” ordinal covariates that are expected to explain the default event. In that paper, each covariate was considered separately, and similar firms were grouped to form “J” clusters according to the ordinal levels of the covariate in consideration (homogeneity assumption). Different groups should have significantly different probabilities of default (heterogeneity assumption). The predicted probability of default of the  $i$ th firm will be the posterior expected value of the risk level  $\theta$  and is obtained from the following equation:

$$E(\theta_i | X, Y) = \sum_{K=1}^K E(\theta_j | g_k, Y) \cdot p(g_k | Y) \quad (2)$$

where  $g_k$  is a partition induced by the covariate  $X_k$  that classifies each firm  $i$  into one level  $j$  and  $Y$  is the observed target variable with two levels: default or non default. Further mathematical details regarding the derivation of the final estimated probability of default and the variable selection methodology are omitted; therefore, I refer the reader to the full version of C&G’s (2014) paper.

## 4. Data

The dataset in this research study is the one used by C&G (2014). Provided by an Italian bank, it contains a set of ordinal and soft information variables measured on companies from diverse industries. The dataset contains 1000 European companies with 13 variables, including the response (default or non default event), and was collected at the end of 2010. Three of the 12 other explanatory variables are external ratings, measured on an ordinal scale of nine levels. The remaining nine predictors are internal and soft information variables (Q1–Q9) that shed light on the management and structure of the company itself as well as on the historical relationship between the firm and the bank (if any). The description for the variables in question is below as reported in C&G (2014).

	Variable	Description
External	Ai	Describes the banking transactions of the involved companies
	Dir	Contains macroeconomic scenarios
	Cebi	Includes balance sheet information
Internal	Q1	Represents the competitive position of the company
	Q2	Ability to change management board without financial consequences
	Q3	Payment times to suppliers
	Q4	Number of clients that the 50% of sales refer to
	Q5	Payment times to clients
	Q6	Trend of the demand for the goods produced by the company
	Q7	Historical relationship with the bank
	Q8	Financial ability of the board to face adverse economic conditions
	Q9	Professional experience of the management

**Table 1.** RF-CART variable importance.

Variable	Variable importance (VI)
AI	0.066
Q7	0.016
CEBI	0.008
Q1	0.008
Q8	0.006
Q5	0.005
DIR	0.004
Q3	0.003
Q6	0.003
Q2	0.001
Q4	0
Q9	0

The nine internal covariates have been scaled in decreasing order, meaning that higher levels refer to worse ratings.

## 5. Results and discussion

Since the results of RFs are subject to random variation, I built the two types of RFs (RF-CART and RF-CI) with a very large number of trees to avoid this problem. Thus, each RF is built with 5000 trees using the statistical programming language R.

The variables are listed in [Tables 1](#) and [2](#) in decreasing order according to their importance. Both RFs voted for AI and CEBI as the most important external variables. This finding is in line with the Bayesian results. Unlike the Bayesian method that identified Q3 as the most important among the internal variables, both RFs identified Q1, Q7, and Q8 as the most significant ones. In either case the selection of internal variables transmits an important signal to financial institutions to pay more attention to the information coming from their customers. Since both RFs found AI, CEBI, Q1, Q7, and Q8 to be important, it is highly recommended that these variables be considered in any model related to credit risk.

After the RF variable selection stage, three predictive models of default are built with the selected variables, namely RF-CART, RF-CI, and Bayesian model averaging for logistic regression (BMALR). The predictive performance of these models is compared to the C&G model. The full dataset is portioned into training and validation datasets. The training dataset contains around 70% of the observations. [Table 3](#) reports the accuracy measures of the predictive performance for each model, namely the total accuracy and the area under the ROC curve (AUC) that are calculated on the validation dataset. AUC is used as an overall measure of

**Table 2.** RF-CI variable importance.

Variable	Variable importance (VI)
AI	0.068
Q7	0.011
Q8	0.007
Q1	0.006
CEBI	0.001
Q3	0.001
Q5	0
Q2	0
Q6	0
Q9	0
Q4	0
DIR	0

**Table 3.** Predictive performance of the five models.

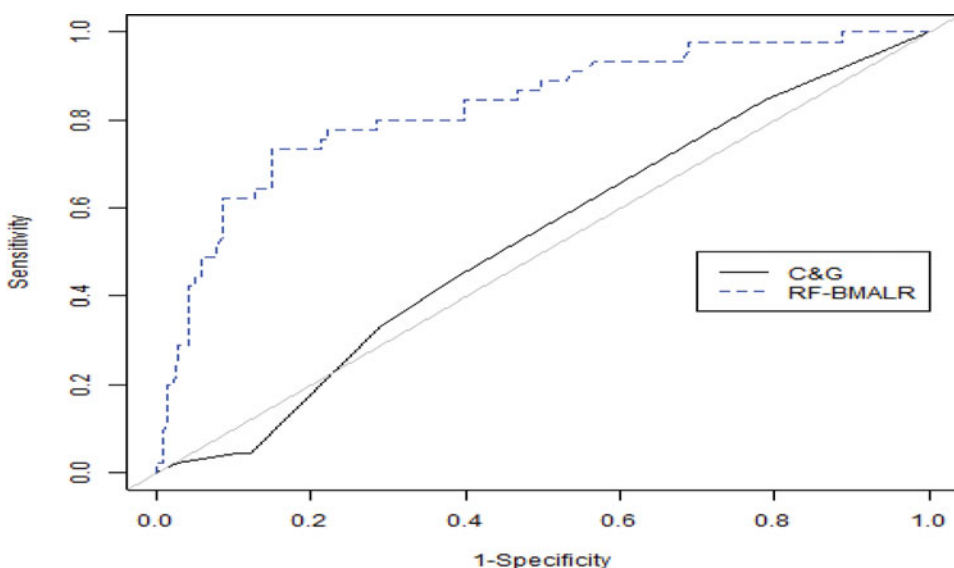
Model	Accuracy	AUC	AUC standard error
RF-CART	0.8333	0.8626	0.0332
RF-CI	0.8495	0.8548	0.0332
BMALR	0.8226	0.8297	0.0366
C&G	0.7581	0.5497	0.0457

**Table 4.** Coefficients and posterior probabilities.

Coefficients						
Intercept	CEBI	AI	Q1	Q7	Q8	
-14.08	0.16	0.56	0.05	0.30	0.84	
Posterior probabilities						
0.47	0.18	0.14	0.06	0.06	0.04	0.02

predictive performance. Moreover, AUC (Ling et al., 2003), which ranges from 0 to 1, is a better measure of overall performance and does not depend on any specific classification cutoff. Thus, the higher the AUC, the better a classifier performs.

As Table 3 shows, the RFs and BMALR noticeably outperform the C&G model with the highest accuracies and AUCs. RFs and BMALR show similar performance measures, which are much higher than those of the C&G model. More importantly, RFs do not require any variable transformation, and deal with all variable types, unlike the C&G model, whose partitioning of the ordinal variables into classes led to shrinkage of data and less accurate results. Although RFs outperform the predictive performance of the C&G model, they are treated as “black box” methods and fail to provide a clear interpretation of the relationship between the outcome and the predictor variables. However, this problem is solved here by estimating a BMALR as a second stage using the important predictor variables selected by the RFs in the first stage. This integrated framework outperformed the C&G in terms of straightforward

**Figure 1.** ROC curves.

interpretability and predictive power. The coefficients of BMALR were calculated by averaging over eight best models, and they are reported in Table 4 along with the individual posterior probabilities of the averaged models.

Aside from the numerical measures, the graphs in Figure 1 visually highlight the predictive comparison between the C&G model and BMALR based on ROC curves, which is helpful in visualizing the overall performance of a classifier. It maps the sensitivity against 1- specificity. The closer the curve is to the upper left corner, the higher the performance of the classifier.

According to Figure 1, BMALR produces much higher predictive performance than the C&G model. Once again, this proves that the C&G model lags behind the integrated framework proposed in this paper.

## 6. Conclusion

Credit rating is an essential measure of a company's repayment ability. In the literature, more attention is paid to financial variables rather than soft information variables as inputs for internal ratings. In this research study, an integrated framework of RF-based methods and Bayesian model averaging (RF-BMA) was proposed to investigate in greater depth the role of soft information variables in evaluating credit assessment. The proposed framework was compared to the Bayesian ratings model proposed by Cerciello and Giudici (2014) (C&G) on a European credit risk database. Although RFs showed high predictive performance, they remain “black-box” predictive models that do not adequately explain the relationship between the dependent variable and the predictor variables. However, their robust importance variable feature partially compensates for this interpretability problem. Therefore, RFs constituted the first stage of the proposed framework to select the important soft information variables. As a second stage, a BMA for logistic regression was estimated using RF selected variables. This model has advantages over the classical regression model since it accounts for the inherent uncertainty in the regression coefficients by averaging over the best models according to approximate posterior model probability.

Both models (the C&G and the proposed framework) proved that non financial variables are important to consider in credit scoring modeling. However, results showed that my proposed approach outperformed the C&G model in terms of the easiness of interpretability and predictive accuracy. Thus the integrated framework proposed in this study is a superior alternative to the C&G proposed model. However, the available dataset used in this work was limited to a one-year observation time and hence does not incorporate the credit rating change over time. Therefore, future research avenues could address the time-varying effects in credit rating by applying an adapted version of RFs or other appropriate data mining techniques to longitudinal data with additional soft information variables.

## References

- Abdou, H. (2009). Genetic programming for credit scoring: the case of Egyptian public sector banks. *Expert Syst. Appl.* 36(9):11402–11417.
- Abdou, H., Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intell. Syst. Accounting Finance Manage.* 18(2–3):59–88.
- Abdou, H., Pointon, J. (2009). Credit scoring and decision-making in Egyptian public sector banks. *Int. J. Managerial Finance.* 5(4):391–406.
- Abdou, H., Pointon, J., El Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Syst. Appl.* 35(3):1275–1292.
- Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance.* 23(4):589–609.

- Baesens, B., Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc.* 54(6):627–635.
- Bailey, M. (2001). *Credit Scoring: The Principles and Practicalities*. Bristol: White Box Publishing.
- Banasik, J., Crook, J., Thomas, L. (2003). Sample selection bias in credit scoring models. *J. Oper. Res. Soc.* 54(8):822–832.
- Banasik, J., Crook, J., Thomas, L. (2001). Scoring by usage. *J. Oper. Res. Soc.* 52(9):997–1006.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press.
- Blöchlinger, A., Leippold, M. (2006). Economic benefit of powerful credit scoring. *J. Bank. Finance.* 30(3):851–873.
- Boyes, W.D., Hoffman, D.S., Low, S. (1989). An econometric analysis of the bank credit scoring problem. *J. Econom.* 40(1):3–14.
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45:5–32.
- Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24:123–140.
- Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Belmont: Wadsworth.
- Capon, N. (1982). Credit scoring systems: a critical analysis. *J. Marketing.* 46(2):82–91.
- Cerciello, P., Giudici, P. (2014). Bayesian credit ratings. *Commun. Stat. Theory Methods.* 43: 867–878.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. Series A.* 158(3):419–466.
- Chuang, C., Lin, R. (2009). Constructing a reassigning credit scoring model. *Expert Syst. Appl.* 36(2/1):1685–1694.
- Crook, J. (1996). Credit scoring: an overview. *British Association, Festival of Science. University of Birmingham, The University of Edinburgh, Working Paper Series 96/13.*
- Desai, V., Crook, J., Overstreet, G. (1996). A Comparison of neural networks and linear scoring models in the credit union environment. *Eur. J. Oper. Res.* 95(1):24–37.
- Dimla, D., Lister, P. (2000). On-line metal cutting tool condition monitoring II: tool-state classification using multi-layer perceptron neural networks. *Int. J. Mach. Tools Manuf.* 40(5):769–781.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. R. Stat. Soc. Series B.* 57(1):45–97.
- Friedman, J. (1991). Multivariate adaptive regression splines. *Ann. Stat.* 19(1):1–141.
- Fritz, S., Hosemann, D. (2000). Restructuring the credit process: behavior scoring for German corporates. *Intell. Syst. Accounting Finance Manage.* 9(1):9–21.
- Gately, E. (1996). *Neural Networks for Financial Forecasting: Top Techniques for Designing and Applying the Latest Trading Systems*. New York: John Wiley and Sons, Inc.
- Greene, W. (1998). Sample selection in credit scoring models. *Japan World Econ.* 10:299–316.
- Grunert, J., Norden, L., Weber, M. (2005). The role of non-financial factors in internal credit ratings. *J. Bank. Finance.* 29(2):509–531.
- Guillen, M., Artis, M. (1992). Count data models for a credit scoring system. *The European Conference Series in Quantitative Economics and Econometrics on Econometrics of Duration: Count and Transition Models*, Paris.
- Hand, D., Henley, W. (1997). Statistical classification methods in consumer credit scoring: a review. *J. R. Stat. Soc. Series A (Statistics in Society)*. 160(3):523–541.
- Heffernan, S. (2005). *Modern Banking*. Chichester: John Wiley and Sons, Ltd.
- Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20:832–844.
- Hoeting, J.A. (2002). Methodology for Bayesian model averaging: an update. *Proceedings of the International Biometric Conference*, Freiburg, Germany, 231–240.
- Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian model averaging: a tutorial with discussion. *Stat. Sci.* 14:382–417.
- Hosmer, D.W., Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Hothorn, T., Hornik, K., Van de Wiel, A., Zeileis, A. (2006a). A Lego system for conditional inference. *Am. Stat.* 60:257–263.
- Hothorn, T., Hornik, K., Zeileis, A. (2006b). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* 15:651–674.

- Huang, Z., Chen, H., Chen, W., Hsu, C., Chen, C., Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decis. Support Syst.* 37(4):543–558.
- Huang, C., Chen, M., Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* 33(4):847–856.
- Hung, C., Chen, J. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert Syst. Appl.* 36(3):5297–5303.
- Karels, G., Prakash, A. (1987). Multivariate normality and forecasting of business bankruptcy. *J. Bus. Finance Accounting.* 14(4):573–593.
- Kim, Y.S., Sohn, S.Y. (2004). Managing loan customers using misclassification patterns of credit scoring model. *Expert Syst. Appl.* 26(4):567–573.
- Lee, T., Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.* 28(4):743–752.
- Lee, T., Chiu, C., Lu, C.I., Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Syst. Appl.* 23(3):245–254.
- Leonard, K.J. (1992). Credit scoring models for the evaluation of small-business loan applications. *IMA J. Math. Appl. Bus. Ind.* 4(1):89–95.
- Leung, K. (2008). A comparison of variable selection techniques for credit scoring. *Proceedings 7th International Conference on Computational Intelligence in Economics and Finance*, Taiwan.
- Lewis, E.M. (1992). *An Introduction to Credit Scoring*. San Rafael, CA: Fair, Isaac and Co., Inc.
- Ling, C., Huang, J., Zhang, H. (2003). AUC: A statistically consistent and more discriminating measure than accuracy. *Proc. Int. Joint Conf. Artificial Intell.* 519–526.
- Makowski, P. (1985). Credit scoring branches out. *Credit World.* 74(2):30–37.
- Malhotra, R., Malhotra, D.K. (2003). Evaluating consumer loans using neural networks. *Omega. Int. J. Manage. Sci.* 31(2):83–96.
- Masters, T. (1995). *Advanced Algorithms for Neural Networks: A C++ Sourcebook*. New York: John Wiley and Sons, Inc.
- Mays, E. (2001). *Handbook of Credit Scoring*. Chicago, IL: Glenlake Publishing Company, Ltd.
- Morgan, J., Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *J. Am. Stat. Assoc.* 58(302):415–434.
- Ong, C., Huang, J., Tzeng, G. (2005). Building credit scoring models using genetic programming. *Expert Syst. Appl.* 29(1):41–47.
- Orgler, Y.E. (1971). Evaluation of bank consumer loans with credit scoring models. *J. Bank Res.* 2(1):31–37.
- Reed, R.D., Marks, R.J. (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. London: The MIT Press.
- Reichert, A.K., Cho, C.C., Wagner, G.M. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *J. Bus. Econ. Stat.* 1(2):101–114.
- Sarlij, N., Bencic, M., Bohacek, Z. (2004). Multinomial model in consumer credit scoring. In *10th International Conference on Operational Research, Trogir, Croatia*.
- Schebesch, K.B., Stecking, R. (2005). Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *J. Oper. Res. Soc.* 56(9):1082–1088.
- Seow, H., Thomas, L.C. (2006). Using adaptive learning in credit scoring to estimate take-up probability distribution. *Eur. J. Oper. Res.* 173(3):880–892.
- Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken, NJ: John Wiley and Sons, Inc.
- Steenackers, A., Goovaerts, M.J. (1989). A credit scoring model for personal loans. *Insurance: Math. Econ.* 8(8):31–34.
- Stefanowski, J., Wilk, S. (2001). Evaluating business credit risk by means of approach-integrating decision rules and case-based learning. *Intell. Syst. Accounting Finance Manage.* 10(2):97–114.
- Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinf.* 9:307.
- Strobl, C., Boulesteix, A.L., Zeileis, A.T., Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf.* 8:25.

- Sustersic, M., Mramor, D., Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Syst. Appl.* 36(3):4736–4744.
- Thomas, L.C. (2000). A Survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *Int. J. Forecasting.* 16(2):149–172.
- Trinkle, B.S., Baldwin, A.A. (2007). Interpretable credit model development via artificial neural networks. *Intell. Syst. Accounting Finance Manage.* 15(3–4):123–147.
- Trippi, R.R., Turban, E. (1993). *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real-World Performance.* Chicago, IL: Irwin.
- Wang, G., Hao, J., Ma, J., Huang, L. (2012). Empirical evaluation of ensemble learning for credit scoring. *Management Associatio: Machine Learning: Concepts, Methodologies, Tools and Applications.* 1108–1127.
- West, D. (2000). Neural network credit scoring models. *Comput. Oper. Res.* 27(11–12):1131–1152.
- Yim, J., Mitchell, H. (2005). Comparison of country risk models: hybrid neural networks, logit models, discriminant analysis and cluster techniques. *Expert Syst. Appl.* 28(1):137–148.
- Yu, L.A., Wang, S.Y., Lai, K.K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Syst. Appl.* 34(2):1434–1444.
- Zekic-Susac, M., Sarlija, N., Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural networks and decision tree models. *In 26th International Conference on Information Technology Interfaces, Croatia.*