International Journal of Advanced Science and Technology
Vol.72 (2014), pp.19-30
http://dx.doi.org/10.14257/ijast.2014.72.02

# Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study

Nazeeh Ghatasheh

*Department of Business Information Technology, The University of Jordan*
*Aqaba, 77110, Jordan*
*n.ghatasheh@ju.edu.jo*

***Abstract***

*In the era of stringent and dynamic business environment, it is crucial for organizations to foresee their clients' delinquency behavior. Such environment and behavior create unreliable base for strategic planning and risk management. Business Analytics combines the business expertise and computer intelligence to assist the decision makers by predicting an individual's credit status. This empirical research aims to evaluate the performance of different Machine Learning algorithms for credit risk prediction with more focus on Random Forest Trees. Several experiments inspired by observation and literature illustrate the potentials of computer-based model in classifying a number of bank history records. However, enhanced classification outcomes require tuning the randomness and tree growing parameters of the Random Forests algorithm. The model based on Random Forest Trees overperformed most of the other models. Moreover, such a model has various advantages to business experts as the ability to help in understanding the relations between the analyzed attributes.*

***Keywords:*** *Business Analytics; Decision Trees; Machine Learning; Random Decision Forest; Risk Prediction; Strategic Planning*

## 1. Introduction

In the period of unstable business environment, it is critical for organizations to predict the behaviors of their clients. One of the main concerns of funding organizations or banks is their clients' adherence to payback the debts as expected. For that it is important to assess the clients' credit suitability before authorizing a loan for example. According to the study in [1] it is apparent how United States and Europe loans and mortgages tend to raise over years, though a credible risk assessment of credit needs to take place. The variables affecting the risk factor vary and the effects associated with overdue or unpaid debts by consumers may have unwanted consequences that may exceed the organizational level [1, 2].

Business Analytics (BA) [3] is a convenient approach that utilizes Business Intelligence (BI) techniques to fulfill business needs, as predicting behaviors and outcomes. Therefore BA benefits business needs using the capabilities of Information Technology (IT) including Data Mining. The overall idea behind BA is to integrate the potentials of IT domain expertise with business domain expertise to reach an effective collaboration.

Machine learning (ML) has been one of the IT domains contributing effectively to business prediction problems. An interesting proposed machine learning approach in [4] seeks predicting customers churn in telecommunication industry using Genetic Programming (GP). Churn management gains high importance in business domain as it is related directly to customer retention strategies. Though analyzing the behaviors of customers is critical and acts

ISSN: 2005-4238 IJAST

as an early alarming mechanism. Afterwards it triggers the business activities related to risk prevention and thus business continuity.

In literature, many researches proposed the application of Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Fuzzy Learning for predicting credit risk. Table 1 lists a number of attempts to predict credit risk using different approaches. However, some of these approaches do not give much insight on the dynamics and interaction between the underlying factors, and they are commonly considered as black box modeling techniques. For these reasons other category of machine learning approaches is investigated, this category contains approaches that produce more interpretable models that are easy to evaluate and explain more the effect of each factor in the prediction. For instance, some researchers [5] suggest using Random Forest for being more descriptive.

**Table 1. ML Approaches for Credit Risk Prediction**

| |
|---|
| Support Vector Machines [6] |
| Artificial Neural Networks [7, 8] |
| Feature Selection Based using ANN [9] |
| Ensemble ANN [10] |
| ANN and Decision Tables [11] |
| Evolutionary Product-ANN [12] |
| Fuzzy Immune Learning [13] |
| Genetic Programming [14] |
| Genetic Programming and SVM [15] |
| Wavelet Networks and Particle Swarm Optimization [16] |
| Various AI Techniques [17, 18] |

ML homogenous and hybrid approaches show promising results; nevertheless they do not overperform simpler approaches significantly. Per contra, simpler approaches ask for attention with an opportunity for efficient prediction performance. This research attempts to investigate the opportunities of Decision Trees with a focus on Random Forest Trees. The main objectives of this work are tailored towards studying the performance of Forest Trees, comparing different approaches, finding the best configurations for higher prediction accuracy, and spotting the drawbacks of the selected prediction approach.

This work is organized in five main sections in which the first part tries to show the importance and the motivations, the second part explores the literature to summarize the existing related work and prediction attempts, the third part explains in details the Random Forest Trees algorithm, the used dataset and research methodology are presented in the fourth part along with evaluation measures, followed by the discussion of the empirical findings.

## 2. Related Work

A number of Data Mining algorithms were investigated in [18] for credit risk evaluation. DT, ANN, SVM, and Logistic Regression (LR) outcomes for risk prediction were compared over Australian Credit, German Credit and refined German Credit datasets. The authors used 10-fold cross validation strategy with different dataset splitting configurations. The overall findings based on prediction Accuracy, Precision, and error rate of Type I/II, show that SVM and LR models dominate in terms of performance. More precisely SVM based models are robust and have higher ability for better generalization with small training sets. Nonetheless, DT models as the case of C4.5 trees are easier to interpret due to their higher level of clarity and showed comparable results. The authors, referring to the empirical evidence, stress the

 

idea that DT demonstrate better aid to the decision making of business domain experts. Clear and easy understanding of a prediction model reveals greater details for business domain experts who would be glad to comprehend. Moreover, they proposed possible future work to include a hybrid approach incorporating DT and SVM for credit risk assessment.

A model based on Random Survival Forests was compared to Logit model for predicting credit risk for a number of German Small and Medium Enterprises (SMEs) in [5]. The dataset contains a set of variables that show whether the financial situation is good or not, relying on financial calculation called "Financial Ratio". An error measurement encompassing the number of trees and variables importance was used to evaluate the performance of the used models. The empirical results show that Logit models gave better predictions, while a Random Survival Forests were more explanatory.

The authors in [17] investigated the suitability of several AI techniques in predicting credit risk. Ten classification algorithms were used to predict the credit risk in a German dataset. The set of measures used to compare the performance of the different algorithms was "Type I, Type II, and total accuracy [19]", while ranking the algorithms relied on the area under Receiver Operating Characteristic (ROC) curve. The overall results nominate SVM based models to be more accurate in prediction. However, a hybrid approach was used that includes clustering before the classification step. Such approach would be relatively complex and applying it in business world could not be understood easily.

Even though there are many complex approaches to predict credit risk, none of their complexities is justified by a significant accuracy level. For that the literature lacks a significant improvement in terms of performance to accuracy ratio. That would raise the question of whether simple approaches would suit more the credit risk prediction issue, especially of being relatively more descriptive.

## 3. Random Forest Trees

Random Forest Trees are based on a number of prediction trees that are less tolerant to noise compared to "Adaboost" and utilize random selection of features in splitting the trees. "Random Forests" is a voting procedure for the most popular class among a large number of trees. Thus, a random forest classifier is composed of a set of tree-structured classifiers [20-22]. Equation [1] represents the classifiers where $\Theta_i$ represents a number of independent random vectors distributed identically, such that every tree has a vote for most popular class of input *X*.

$$Space = \{ h( X, \Theta_i ); \ i= 1,2,3,.... nT \} \quad (1)$$

Using Random Forests for prediction has many advantages such as their immunity to overfitting, an appropriate selection of randomness type leads to accurate classification or regression, the correlation and strength of predictors makes a good estimate of the ability for prediction, faster than boosting and bagging, better estimation of internal errors, not complicated, and can perform well in parallel processing. Based on the empirical results in [20, 21], Random Forests could compete with similar approaches in terms of accuracy. Moreover, the author proved that Random Forest could give better results with "Boosting" and "Bagging" [20]. Accordingly, opening the door for an investigation area of the injected randomness effects on prediction accuracy. The pseudo code of a modified version of Random Decision Forests (RDF) used in Heursticlab[1] environment is shown next[2].

[1] HeuristicLab is a framework for heuristic and evolutionary algorithms that is developed by members of the Heuristic and Evolutionary Algorithms Laboratory (HEAL)
[2] http://www.alglib.net/dataanalysis/decisionforest.php

**Algorithm 1:** Modified RDF Algorithm

Input: training set of size *N*, having *M* independent variables
Input Parameters: *r*, *m*, *nT*
T1 = generate *n* by *nT* random and unique samples of training set
G1 = rest of training set
for each node
randomly choose *m* variables
calculate best split in *T1* according to *m*
repeat *nT* times

## 4. Data and Methodology

The German Credit dataset used in this study is publicly available at the University of California, Irvine (UCI) Machine Learning Repository [23]. In which there are 1000 instances divided into two classes; 700 "good credit" and 300 "bad/refused credit request". The original credit dataset contains 20 variables that fall into 13 categorical and 7 numerical ones listed in Table 2. However, this research work is conducted entirely on a processed copy (by the Strathclyde University) that is also available at UCI repository. The processed dataset is a conversion of the originals into 25 numerical variables, in which number 25 is an output variable (Good/Bad). Figure 1 illustrates the distribution of the numerical variables.

**Table 2. Original Attributes [23]**

| No. | Attribute | Type |
|---|---|---|
| 1 | Status of existing checking account | Qualitative |
| 2 | Duration in month | Numerical |
| 3 | Credit history | Qualitative |
| 4 | Purpose | Qualitative |
| 5 | Credit amount | Numerical |
| 6 | Savings account/bonds | Qualitative |
| 7 | Present employment since | Qualitative |
| 8 | Installment rate in percentage of disposable income | Numerical |
| 9 | Personal status and sex | Qualitative |
| 10 | Other debtors / guarantors | Qualitative |
| 11 | Present residence since | Numerical |
| 12 | Property | Qualitative |
| 13 | Age in years | Numerical |
| 14 | Other installment plans | Qualitative |
| 15 | Housing | Qualitative |
| 16 | Number of existing credits at this bank | Numerical |
| 17 | Job | Qualitative |
| 18 | Number of people being liable to provide maintenance for | Numerical |
| 19 | Telephone | Qualitative |
| 20 | Foreign worker | Qualitative |

 

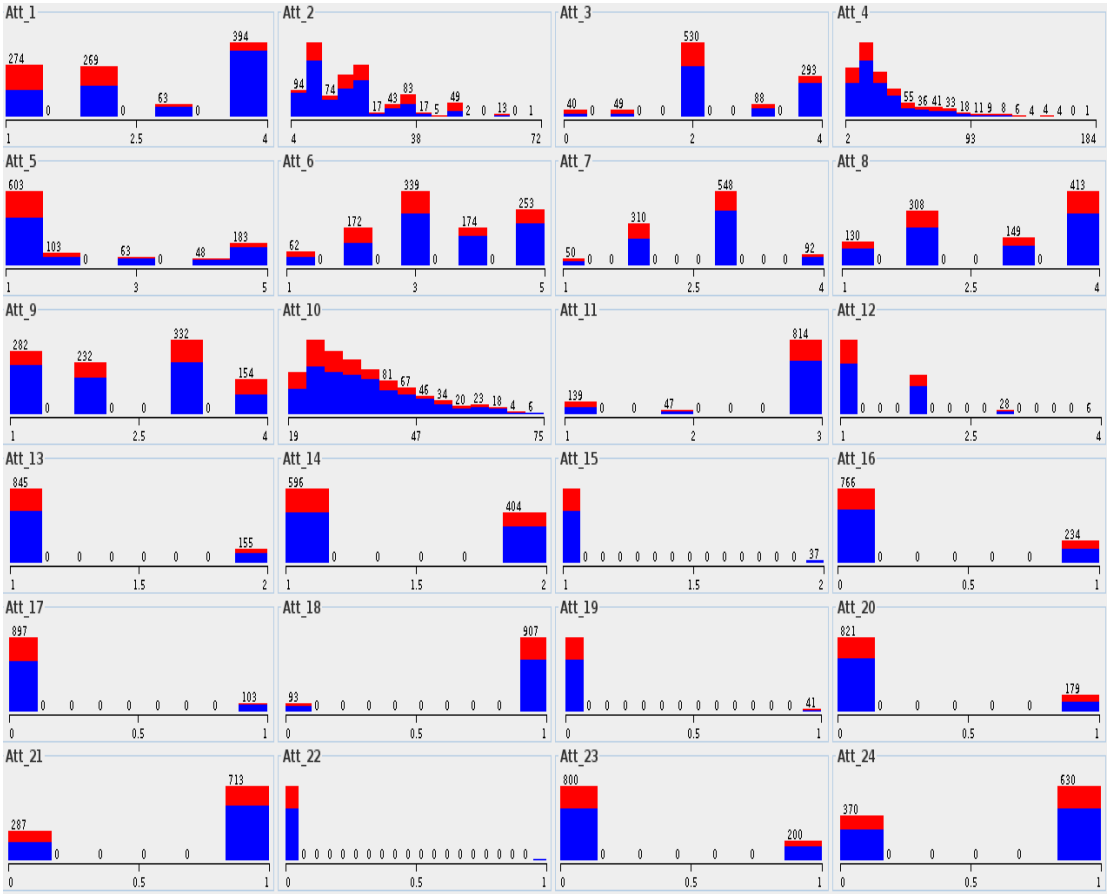


**Figure 1. Distribution of the Numerical Variables in the Processed Dataset**

All classification runs use an approach of 10-Fold cross validation using different tools that are HeuristicLab, Weka, and Keel. Default configurations are used in terms of model creation and testing to compare with the experimental results of the Random Forest Trees. The benchmark algorithms in Weka are C4.5 Decision Trees, Adaboost with C4.5, RealAdaboost with C4.5, Bagging with C4.5, Dagging with C4.5, Decorate with C4.5, and SVM using Linear Kernel. Keel algorithm is Evolutionary Product Neural Network for Classification (NNEP-C). In Heuristiclab the algorithms are Neural Network Ensemble Classification (NNEC), Multinomial Logit Classification (MN Logit), Genetic Programming.

The main experiments are conducted using Heuristiclab using a modified Random Forest Trees algorithm for classification. Relying on the author's propositions in [20], the modified algorithm utilizes RDF. There are three main parameters to configure: *r* is a ratio between 0 and 1, *m* number of attributes, and *nT* trees. *r* represents the ratio of the individual tree samples to be constructed, *m* represents the number of variables used in growing the trees, and *nT* is the number of the model's trees. Proper selection of *r* and *m* affects the noise tolerance issue in the training set, for that these variables need to be adjusted carefully. In this research empirical tests find the most suitable parameters *r*, *m*, and *nT*.

Several measures are used to evaluate the outcomes of the classification algorithms relying on the classified instances [19]. After processing the input instances each one is classified into one of four possibilities that are presented visually in a matrix, by confronting the actual and predicted instances. Figure 2 represents the matrix that is called the "Confusion or Contingency Matrix".

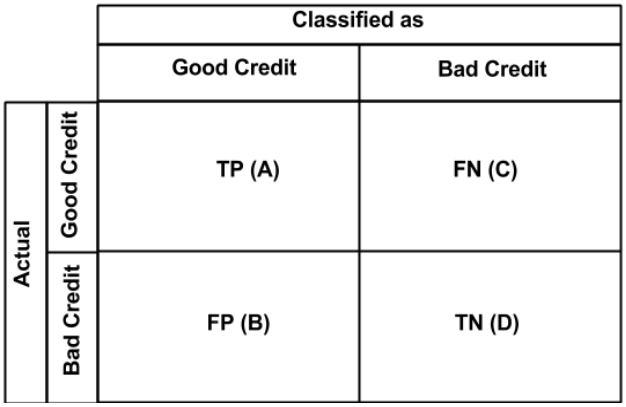


**Figure 2. Confusion Matrix**

This confusion matrix illustrates the classification's "confusion" or what is called classification error, in which the rows represent the actual classes and the columns represent the predicted classes. Total number of correctly classified instances will be represented diagonally in True Positive (TP) and True Negative (TN) cells. TP represents "Actual Good" classified as "Good" while TN represents "Actual Bad" classified as "Bad". The higher TP and TN the better is the performance of the classification algorithm. Incorrectly classified instances go to False Positive (FP) and False Negative (FN) that are "Actual Good" classified as "Bad" and "Actual Bad" classified as "Good" respectively. Total sum of the correctly and incorrectly classified classes should match the total number of the input instances.

Several mathematical measures based on the confusion matrix make it easier to assess deeply the performance of the classification algorithm, also make it easier to compare the performance of different algorithms. Due to the variety of measures most of the researchers use and for better comparison, this study will report a number of measures that are:

- Total Accuracy (Correctly Classified Instances) = TP + TN / ( TP + TN + FP + FN )
- Sensitivity (Recall, Hit Rate, TP Rate, or Type II Error) = TP / ( TP + FN )
- Precision (Confidence or Type I Error) = TP / ( TP + FP )
- F-Measure = (2 * Precision * Sensitivity) / (Precision * Sensitivity)
- Area Under Receiver Operating Characteristics Curve (AUC) [19, 24], for some of the algorithms. It is calculated automatically in Weka environment.

This study tries empirically to find out the potentials of Random Forest Trees in classification by tuning the used models. That includes relying on the literature and arbitrary selection of model parameters.

 

## 5. Results and Discussion

### 5.1. Tuning of the Parameters

The first set of experiments reported the outcomes of the classification under a systematic tuning of the Random Forests (modified RDF in Heuristiclab) algorithm. Several runs were analyzed with arbitrary selection of the parameters *R* and *M* for different number of trees. Starting from 50 to 500 trees with a 50 increment each time. Changing the parameters does not affect significantly the test performance, thus two tests were selected for deeper analysis. Figures 3 and 4 show the results of the classification model on the test dataset over the 10-fold cross validation with *R* = 0.3 and *M*= 0.5 for the first, *R*= 0.66 and *M* = 0.3 for the second.

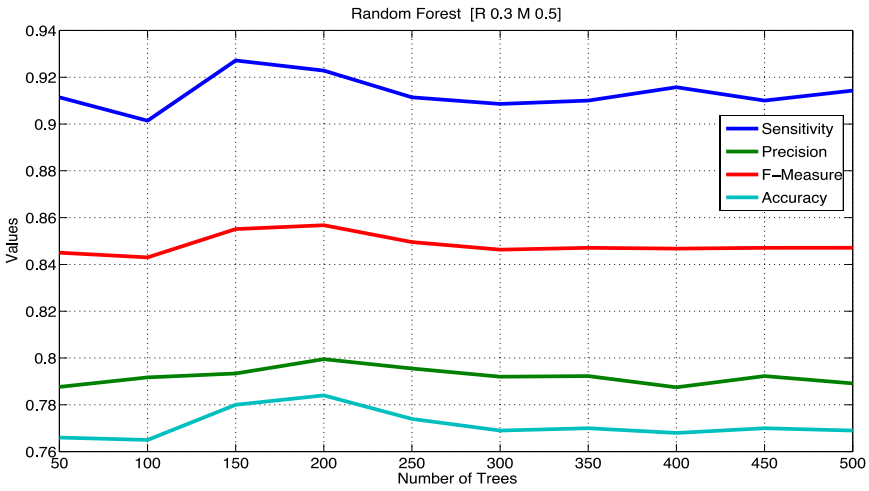


**Figure 3. Test Dataset Results using 10-fold Cross Validation for 10 runs (R=0.3, M=0.5)**

The modified RDF algorithm makes it easier to study the effects of constructed trees and variables used. There is no clear relation between the number of the trees and the performance of the model, it is clear that the performance is relatively better using 200 trees. Table 3 shows the mean, standard deviation (S.D.), median, maximum and minimum of the reported measures for the test set. That shows no high extremes among the total results. Moreover, in terms of training set accuracy all reported measures exceeded 90% and at 200 trees precisely the total accuracy was 93.4, sensitivity 0.917, precision 0.996, and F-Measure reached 0.955.

**Table 3. Test Dataset Statistics using 10-fold Cross Validation for 10 Runs (R=0.3, M=0.5)**

| | *Mean* | *S.D.* | *Median* | *Maximum* | *Minimum* |
|---|---|---|---|---|---|
| Sensitivity | 0.913 | 0.007 | 0.911 | 0.927 | 0.901 |
| Precision | 0.792 | 0.004 | 0.792 | 0.8 | 0.787 |
| F-Measure | 0.848 | 0.004 | 0.847 | 0.857 | 0.843 |
| Accuracy | 0.772 | 0.006 | 0.77 | 0.784 | 0.765 |

Also for the second attempt, there is no clear relation between the number of the trees and the performance of the model, it is clear that the performance is relatively better using 300 trees. Table 4 shows the mean, standard deviation (S.D.), median, maximum and minimum of the reported measures for the test set. That shows no high extremes among the total results. Moreover, the training set performance reached 100% classification accuracy.

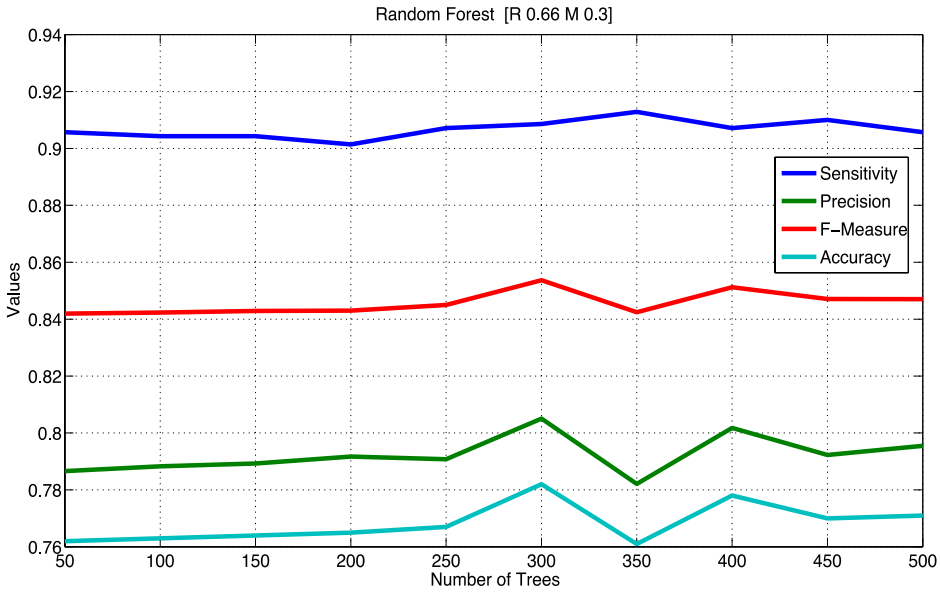


**Figure 4. Test Dataset Results using 10-fold Cross Validation for 10 Runs (R=0.66, M=0.3)**

**Table 4. Test Dataset Statistics using 10-fold Cross Validation /10 runs (R=0.66, M=0.3)**

| | *Mean* | *S.D.* | *Median* | *Maximum* | *Minimum* |
|---|---|---|---|---|---|
| Sensitivity | 0.907 | 0.003 | 0.906 | 0.913 | 0.901 |
| Precision | 0.792 | 0.007 | 0.791 | 0.805 | 0.782 |
| F-Measure | 0.846 | 0.004 | 0.844 | 0.854 | 0.842 |
| Accuracy | 0.768 | 0.007 | 0.766 | 0.782 | 0.761 |

The empirical results of changing the parameters are summarized in Table 5; it shows the relatively best results of four tests using the same methodology of analyzing the RDF algorithm. It is clear that using half of the variables in building the model's trees led to a competitive performance. While changing the number of individual tree samples during model creation has no clear or significant effect on the classification performance.

 

**Table 5. Test Dataset Results using 10-fold Cross Validation for Four Tests - Ten Runs Each**

| *Parameters* | *200, 0.3, 0.5* * | *200, 0.5, 0.5* * | *200, 0.66, 0.5* * | *300, 0.66, 0.3* * |
|---|---|---|---|---|
| Sensitivity | 0.923 | 0.904 | 0.899 | 0.909 |
| Precision | 0.8 | 0.799 | 0.796 | 0.805 |
| F-Measure | 0.857 | 0.849 | 0.844 | 0.854 |
| Accuracy | 0.784 | 0.774 | 0.768 | 0.782 |

* *Tuning parameters: nT, r, m*

Relying on the outcomes of the first set of experiments, a second set was conducted to empirically find an adequate number of variables to grow the trees using Random Forest Classification. The objective was to use about 50% of the variables for a number of trees between 100 and 150. The tests were conducted in Weka environment by changing the number of trees *T* and selected variables for growing trees *V*. Table 6 shows three relatively well performing approaches. Bagging of Random Forest Trees using 100 Trees and 16 variables outperformed in most of the measures, and has 80% area under the ROC curve.

**Table 6. Test Dataset Results of 10-fold Cross Validation on Random Forests**

| *Tuning* | *Algorithm* | *Sensitivity* | *Precision* | *F-Measure* | *Accuracy* | *AUC* |
|---|---|---|---|---|---|---|
| 120T 15V | RF | 0.874 | 0.806 | 0.839 | 0.765 | 0.773 |
| 150T 15V | RF Adaboost | 0.897 | 0.807 | 0.850 | 0.778 | 0.800 |
| 100T 16V | Bagging RF | 0.910 | 0.806 | 0.855 | 0.784 | 0.800 |

### 5.2. Comparison

The final set of experiments was conducted to benchmark the performance of the Random Forest Trees algorithm with related ones. For unbiased benchmark all other tests were conducted over the same data set using default configurations; default parameters were used. The performance of the classification algorithms is summarized in Table 7, including also the best results of the first two sets of experiments.

**Table 7. Test Dataset Results using 10-fold Cross Validation for Different Algorithms**

| *Tuning* | *Algorithm* | *Sensitivity* | *Precision* | *F-Measure* | *Accuracy* | *AUC* |
|---|---|---|---|---|---|---|
| 200T R0.3 M0.5 | RF (RDF) | **0.923** | 0.800 | **0.857** | **0.784** | - |
| 120T 15V | RF | 0.874 | 0.806 | 0.839 | 0.765 | 0.773 |
| 150T 15V | RF Adaboost | 0.897 | 0.807 | 0.850 | 0.778 | **0.800** |
| 100T 16V | Bagging RF | 0.91 | 0.806 | 0.855 | **0.784** | **0.800** |
| Default | C4.5 | 0.847 | 0.794 | 0.82 | 0.739 | 0.688 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Default | Adaboost C4.5 | 0.810 | 0.795 | 0.803 | 0.721 | 0.748 |
| Default | RealAdaboost C4.5 | 0.856 | 0.789 | 0.821 | 0.739 | 0.739 |
| Default | Bagging C4.5 | 0.861 | 0.803 | 0.831 | 0.755 | 0.781 |
| Default | Dagging C4.5 | 0.891 | 0.766 | 0.824 | 0.733 | 0.750 |
| Default | Decorate C4.5 | 0.881 | 0.787 | 0.832 | 0.75 | 0.758 |
| Default | NNEC | 0.895 | 0.806 | 0.848 | 0.772 | - |
| Default | SVM NU Linear | 0.854 | 0.801 | 0.827 | 0.749 | - |
| Default | NNEP | 0.877 | 0.794 | 0.834 | 0.755 | - |
| Default | MN Logit | 0.795 | **0.883** | 0.837 | 0.759 | - |
| Default | GP | 0.731 | 0.676 | 0.702 | 0.599 | - |

** In bold, the best result of each measure.*

Many researchers reported the performance of different classification approaches over the same dataset [6-9, 11-15, 17, 18]. However, the outcomes of this research exceeded many of the previously obtained results and relatively competitive to the rest. In summary, classification using Random Forest Trees has been the focus of many researchers and its empirical results are competitive, without omitting the advantages and disadvantages of any related work.

## 6. Conclusions and Future Work

Due to the importance of understanding and managing the risks in volatile business domains, it is required to find an effective aid in making decisions. The results of this research show that Random Forest Trees algorithm is a promising opportunity for Business Analytics in predicting credit risk. The main advantages of using Random Forest Trees in prediction are the competitive classification accuracy and simplicity. Such simplicity makes it easier for decision makers to understand more the underlying relations, especially for the fact that none of the classification approaches achieved significant accuracy. The pluses make the results of decision trees more useful and appealing for business domain experts than other approaches. A noteworthy finding is the effect of injected randomness and how to grow the individual trees on producing better classification results.

The empirical findings of this research and others open the door for deeper future work to improve the performance of decision trees. Firstly, to improve the classification models by enhancing the way to grow the decision trees, and better variable selection. Secondly, hybrid approaches incorporating Random Forest Trees need thorough investigation and testing. The possibilities vary starting from the results of this research as the use of different datasets, the redesign of the datasets to include or exclude affecting variables, study the impact of each variable on the overall performance, and model different problems as the bankruptcy prediction.

## Acknowledgements

The members of the Heuristic and Evolutionary Algorithms Laboratory (HEAL) for HeuristicLab framework for heuristic and evolutionary algorithms, the developers of WEKA

 

Data Mining software, and the developers of Knowledge Extraction based on Evolutionary Learning (KEEL).

## References

[1] J. Crook and J. Banasik, "Forecasting and explaining aggregate consumer credit delinquency behaviour", Int. J. of Forecasting, vol. 28, no. 1, **(2012)**, pp. 145-160.
[2] S. Jeanneau, "Financial stability objectives and arrangements - what's new?," in The Role of Central Banks in Macroeconomic and Financial Stability, Edited B. for Int. Settlements, B. for Int. Settlements, vol. 76, **(2014)**, pp. 47-58.
[3] R. Bartlett, "A Practitioner's Guide To Business Analytics: Using Data Analysis Tools to Improve Your Organizations Decision Making and Strategy", McGraw Hill Professional, **(2013)**.
[4] H. Faris, B. Al-Shboul and N. Ghatasheh, "A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry", LNCS, vol. 8733, **(2014)**, pp. 253-362.
[5] D. Fantazzini and S. Figini, "Random Survival Forests Models For SME Credit Risk Measurement," Methodology and Computing in Applied Probability, vol. 11, no. 1, **(2009)**, pp. 29-45.
[6] J. Shi, S. Zhang and L. Qiu, "Credit Scoring by Feature-Weighted Support Vector Machines", Zhejiang University Science C, vol. 14, no. 3, **(2013)**, pp. 197-204.
[7] A. Khashman, "Neural Networks for Credit Risk Evaluation: Investigation of Different Neural Models and Learning Schemes", Expert Systems with Applications, vol. 37, no. 9, **(2010)**, pp. 6233-6239.
[8] M. Nazari and M. Alidadi, "Measuring Credit Risk of Bank Customers using Artificial Neural Network", Management Research, vol. 5, no. 2, **(2013)**, pp. 17-27.
[9] P. O'Dea, J. Griffith and C. O'Riordan, "Combining Feature Selection and Neural Networks for Solving Classification Problems" Proc. 12th Irish Conf. Artificial Intell. Cognitive Sci., **(2001)**, pp. 157-166.
[10] A. Ghatge and P. Halkarnikar, "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation", Int. J. of Eng. and Innovative Tech. (IJEIT), vol. 2, **(2013)**, pp. 223-225.
[11] B. Baesens, R. Setiono, C. Mues and J. Vanthienen, "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation", Management Sci., vol. 49, no. 3, **(2003)**, pp. 312-329.
[12] F. Martnez-Estudillo, C. Hervs-Martnez, P. Gutirrez and A. Martnez-Estudillo, "Evolutionary Product-Unit Neural Networks Classifiers", Neurocomputing, vol. 72, no. 13, **(2008)**, pp. 548-561.
[13] E. Kamalloo and M. S. Abadeh, "Credit Risk Prediction using Fuzzy Immune Learning," Adv. Fuzzy Sys., vol. 2014, **(2014)**.
[14] J. J. Huang, G. H. Tzeng and C. S. Ong, "Two-Stage Genetic Programming (2SGP) for the Credit Scoring Model", Applied Mathematics and Computation, vol. 174, **(2006)** March 15, pp. 1039-1053.
[15] A. Satsiou, M. Doumpos and C. Zopounidis, "Genetic Algorithms for the Optimization of Support Vector Machines in Credit Risk Rating", Proceedings of the 2nd Int. Conf. on Enterprise Systems and Accounting, **(2005)** July.
[16] Y. Zhaoji, M. Qiang and W. Wenjuan, "The Application of WN Based on PSO in Bank Credit Risk Assessment", Proceedings of the Int. Conf. on Artificial Intelligence and Computational Intelligence (AICI), vol. 3, **(2010)** October, pp. 444-448.
[17] A. Ghodselahi and A. Amirmadhi, "Application of Artificial Intelligence Techniques for Credit Risk Evaluation", Int. J. of Modeling and Optimization, vol. 1, no. 3, **(2011)**, pp. 243-249.
[18] H. Yu, X. Huang, X. Hu and H. Cai, "A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation", Proceedings of the 4th Int. Conf. on Management of e-Commerce and e-Government (ICMeCG), **(2010)** October, pp. 35-38.
[19] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", Tech. Rep. SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia, **(2007)**.
[20] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, **(2001)**, pp. 5-32.
[21] M. R. Segal, Machine Learning Benchmarks and Random Forest Regression, **(2004)**.
[22] T. K. Ho, "Random Decision Forests", Proceedings of the 3rd Int. Conf. on Document Analysis and Recognition, vol. 1, **(1995)** August, pp. 278-282.
[23] K. Bache and M. Lichman, UCI machine learning repository, **(2013)**.
[24] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers. Tech. Rep., HP Laboratories, **(2004)**.

 

## Author

**Nazeeh Ghatasheh** received the B.Sc. degree in Computer Information Systems from The University of Jordan, Amman, Jordan, in 2004, then he was awarded merit-based full scholarships to continue his M.Sc. Degree in Electronic Business Management and the Ph.D. Degree in Electronic Business at the University of Salento, Lecce, Italy, in 2008 and 2011 respectively. He conducted research activities in the aerospace field related applications, information and communication technologies in the telecommunication industry, and corporate knowledge management. His research interests include image processing and its applications, knowledge representation and management, corporate learning, Machine Learning, and e-Business. Dr. Ghatasheh, at present, is the Head of Business Information Technology and Computer Information Systems Departments at The University of Jordan, Aqaba, Jordan.

 