Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction

Lakshmi Devasena C
Dept. of Operations and Systems,
ISB Hyderabad, IFHE University

ABSTRACT
Envisaging the Credit nonpayer is a risky task of Financial Industries like Banks. find out the defaulter before giving loan is a noteworthy and conflict-ridden task of the Bankers. Classification techniques are the superior choice for predictive analysis like finding the claimant, whether he/she is a modest customer or a cheat. Defining the excellent classifier is a tough assignment for any industrialist like a banker. This gives consent to computer science researchers to drill down efficient research works through evaluating different classifiers and finding out the best classifier for such predictive problems. This research work scrutinizes the efficiency of different Tree Based Classifiers (Random Forest, REP Tree and J48 Classifiers) for the credit risk prediction and compares their robustness through various measures. German credit dataset has been taken and used to envisage the credit risk with the help of open source machine learning tool.

Keywords
Credit Risk Forecast, J48 Classifier, Proficiency Comparison, Random Forest Classifier, REP Tree Classifier.

1. INTRODUCTION
The massive volume of business transactions made information processing automation an energizing factor for high quality standards, cost reduction, with high speed results. Data analysis automation and result of the relevant successes produced by state-of-the art computer algorithms have changed the opinions of many misanthropists. In the past, people thought that financial market analysis necessitates intuition, knowledge and experience and speculated how this job could be automated. Conversely, growth of scientific and technological advances, achieved the automation of financial market analysis. In recent days, credit defaulter prediction and credit risk evaluation have fascinated great deal of interests from regulators, practitioners, and theorists, in the financial industry. Since, the credit risk of an applicant could be predicted from the past giant database and the demographic data, it needs automation. Automation of credit risk forecast can be achieved using classification techniques. Selecting the classifier, which envisages credit risk in an efficient manner, is an imperative and critical task. This research work appraises the credit risk performance of three tree based classifiers, namely, Random Forest, REP Tree Classifier and J48 Classifier and compares their accuracy of credit risk prediction.

2. LITERATURE REVIEW
There are many research works made to predict credit risk using wide-ranging computing techniques. In [1], a neural network based algorithm for automatic provisioning to credit risk scrutiny in a real world problem is presented. An assimilated back propagation neural network (BPNN) with the customary discriminant analysis approach used to discover the performance of credit scoring is given in [2]. A comparative study of corporate credit rating analysis using back propagation neural network (BPNN) and support vector machines (SVM) is described in [3]. An uncorrelated maximization algorithm within a triple-phase neural network ensemble technique for credit risk evaluation to differentiate good creditors from bad ones are elucidated in [4]. An application of artificial neural network to credit risk assessment using two altered architectures are deliberated in [5]. Credit risk investigation using diverse Data Mining models like C4.5, NN, BP, RIPPER, LR and SMO is likened in [6]. The credit risk of a Tunisian bank through modeling the non-payment risk of its commercial loans is analyzed in [7]. Credit risk valuation using six stage neural network ensemble learning approach is argued in [8]. A modeling framework for credit calculation models is erected using different modeling procedures is explained and its performance is analyzed in [9]. Hybrid method for assessing credit risk using Kolmogorove-Smirnov test, Fuzzy Expert system and DEMATEL method is enlightened in [10]. An Artificial Neural Network centered methodology for Credit Risk supervision is proposed in [11]. Artificial neural networks using Feed-forward back propagation neural network and business rules to correctly determine credit defaulter is proposed in [12]. The performance comparison of Memory based classifiers for credit risk investigation is experimented and précised in [13]. The performance comparison between Instance Based and K Star Classifiers for Credit Risk Inspection is accomplished and pronounced in [14]. The performance comparison among Sequential Minimal Optimization and Logistic Classifiers for Credit Risk Calculation is specified in [15]. The performance comparison between Multilayer Perceptron and SMO Classifier for Credit Risk appraisal is described in [16]. The performance comparison between JRip and PART Classifier for Credit Risk Estimation is explored in [17]. This research work scrutinizes the efficiency of different Tree Based Classifiers (Random Forest, REP Tree and J48 Classifiers) for the credit risk prediction.

3. DATASET USED
The German credit data [18] is used to evaluate the performance of Random Forest, REP Tree and J48 Classifiers for credit risk prediction. This data set contains 20 attributes, namely, Duration, Credit History, Checking Status, Purpose, Credit Amount, Employment, Installment Commitment, Saving Status, Personal Status, Other parties, Property magnitude, Age, resident since, Other payment plans, existing credits, job, Housing, No. of dependents, Foreign worker and Own Phone. The data set comprises 1000 instances of client credit data with class detail. It discriminates the records into two classes, namely, good and bad.

4. METHODOLOGY USED

In this research work, different Tree Based Classifiers (Random Forest, REP Tree and J48 Classifiers) are compared for proficiency assessment of credit risk estimation.

4.1 REP Tree Classifier

Reduces Error Pruning (REP) Tree Classifier is a fast decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance [20]. This algorithm is first recommended in [21]. REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterwards it picks best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. Also, this algorithm prunes the tree using reduced-error pruning with back fitting method. At the beginning of the model preparation, it sorts the values of numeric attributes once. As in C4.5 Algorithm, this algorithm also deals the missing values by splitting the corresponding instances into pieces. [22].

4.2 Random Forest Classifier

Random Forests [23] are broadly believed to be the finest "off-the-shelf" classifiers for high-dimensional data. Random forests are a mixture of tree predictors such that each tree depends on the values of a random vector sampled autonomously and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the association between them. A different subset of the training data are selected, with replacement, to train each tree. Remaining training data are used to estimate error and variable importance. Class assignment is made by the number of votes from all of the trees and for regression the average of the results is used. It is similar to bagged decision trees with hardly some key differences as given below:

1.For each split point, the search is not over all p variables but just over mtry variables (where e.g. mtry = [p/3])

2. No pruning necessary. Trees can be grown until each node contains just very few observations (1 or 5).

Advantages of Random Forest over bagged decision trees are listed below:

1. better prediction.

2. almost no parameter tuning necessary with Random Forest.

4.3 J48 Classifier

J48 classifier is a straightforward C4.5 decision tree for classification, which creates a binary tree. It is most useful decision tree approach for classification problems. This technique constructs a tree to model the classification process. After the tree is built, the algorithm is applied to each tuple in the database and results in classification for that tuple [19].

Algorithm J48 [24]:

INPUT:

P//Training data

OUTPUT

DT //Decision tree

DTBUILD (*P)

{

DT=φ;

DT= Create root node and label with splitting attribute;

DT= Add arc to root node for each split predicate and label;

For each arc do

P= Database created by applying splitting

predicate to P;

If stopping point reached for this path, then

DT'= create leaf node and label with

appropriate class;

Else

DT'= DTBUILD(P);

DT= add DT' to arc;

}

While building a decision tree, J48 omits the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The key idea is to split the data into range based on the attribute values for that item that are identified in the training sample [19].

5. PERFORMANCE MEASURES USED

Various scales are used to gauge the performance of the classifiers.

5.1 Classification Accuracy

Any classifier could have an error rate and it may fail to categorize correctly. Classification accuracy is calculated as Correctly classified instances divided by Total number of instances multiplied by 100.

5.2 Mean Absolute Error

Mean absolute error is the average of the variance between predicted and actual value in all test cases. It is a good measure to gauge the performance.

5.3 Root Mean Square Error

Root mean squared error is used to scale dissimilarities between values actually perceived and the values predicted by the model. It is determined by taking the square root of the mean square error.

5.4 Confusion Matrix

A confusion matrix encompasses information about actual and predicted groupings done by a classification system.

6. RESULTS AND DISCUSSION

Open source machine learning tool is used to experiment the performance of different Tree based Classifiers (Random Forest, REP Tree and J48). The performance is tested out using the Training set as well as using different Cross Validation methods. The class is arrived by considering all 20 attributes of the dataset.

6.1 Performance of REP Tree Classifier

The overall assessment summary of REP Tree Classifier using training set and different cross validation methods is given in Table 1. The performance of REP Tree Classifier in terms of

Correctly Classified Instances and Classification Accuracy is shown in Fig. 1 and Fig. 2. The confusion matrix for different test mode is given in Table 2 to Table 6. REP Tree Classifier gives 80% accuracy for the training data set. Various cross validation methods are used to check its actual performance. On an average, it gives around 72% of accuracy for credit risk estimation.

Table 1. REP Tree Classifier Complete Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	800	200	80%	0.2905	0.3811	0.32
5 Fold CV	717	283	71.7%	0.3458	0.4437	0.78
10 Fold CV	718	282	71.8%	0.3417	0.4424	1.33
15 Fold CV	726	274	72.6%	0.3422	0.4382	0.16
20 Fold CV	719	281	71.9%	0.3368	0.4364	0.11

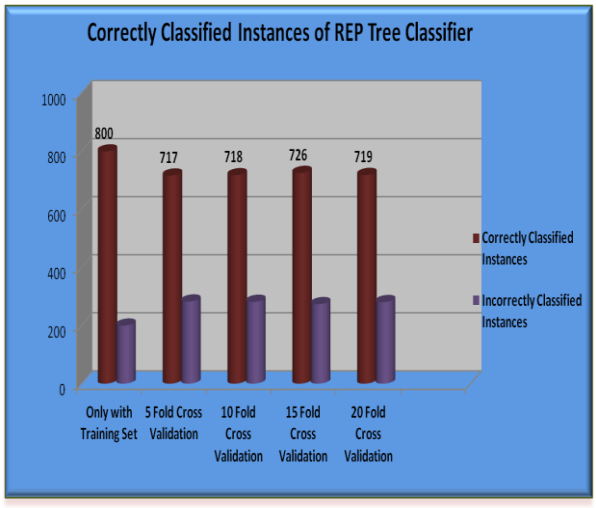

Fig 1: Correctly Classified instances of REP Tree Classifier

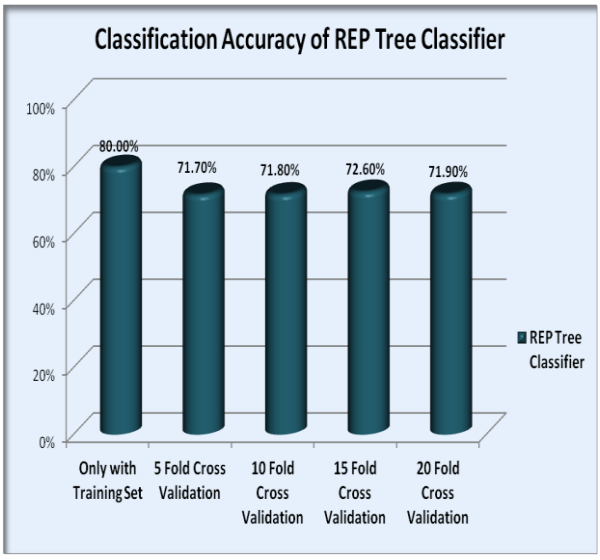

Fig 2: Classification Accuracy of REP Tree Classifier

Table 2. Confusion Matrix – REP Tree Classifier (On Training Dataset)

	Good	Bad	Actual (Total)
Good	649	51	700
Bad	149	151	300
Predicted (Total)	798	202	1000

Table 3. Confusion Matrix – REP Tree Classifier (5 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	616	84	700
Bad	199	101	300
Predicted (Total)	815	185	1000

Table 4. Confusion Matrix – REP Tree Classifier (10 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	601	99	700
Bad	183	117	300
Predicted (Total)	784	216	1000

Table 5. Confusion Matrix – REP Tree Classifier (15 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	612	88	700
Bad	186	114	300

Predicted (Total)	798	202	1000

Table 6. Confusion Matrix – REP Tree Classifier (20 Fold Cross Validation)

	Good	**Bad**	**Actual (Total)**
Good	605	95	700
Bad	151	149	300
Predicted (Total)	756	244	1000

6.2 Performance of Random Forest Classifier

The overall assessment summary of Random Forest Classifier using training set and different cross validation methods is given in Table 7. The performance of Random Forest Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 3 and Fig. 4. The confusion matrix for different test mode is given in Table 8 to Table 12. Random Forest Classifier gives 99% accuracy for the training data set. Various cross validation methods are used to check its actual performance. On an average, it gives around 73.4% of accuracy for credit risk estimation.

Table 7. Random Forest Classifier Overall Evaluation Summary

Test Mode	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Accuracy**	**Mean Absolute Error**	**Root Mean Squared Error**	**Time Taken to Build Model (Sec)**
Training Set	990	10	99%	0.1244	0.1761	0.13
5 Fold CV	741	259	74.1%	0.3403	0.4222	0.06
10 Fold CV	736	264	73.6%	0.3406	0.4232	0.03
15 Fold CV	730	270	73%	0.3427	0.4273	0.03
20 Fold CV	730	270	73%	0.3406	0.4273	0.03

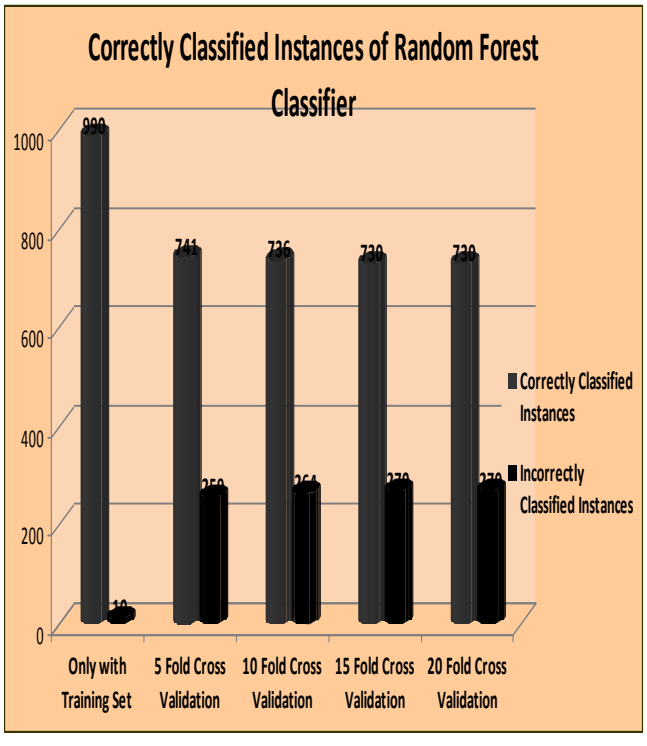

Fig 3: Correctly Classified instances of Random Forest Classifier

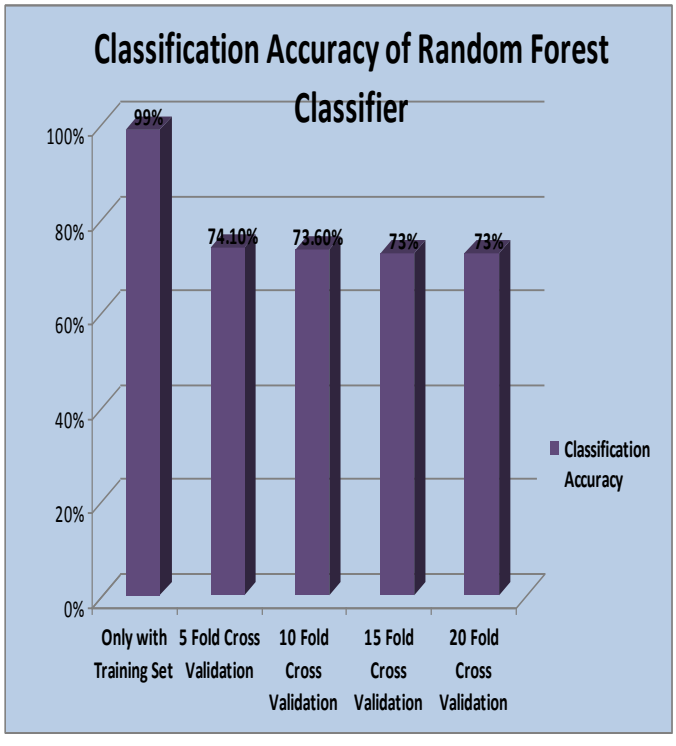

Fig 4: Classification Accuracy of Random Forest Classifier

Table 8. Confusion Matrix – Random Forest Classifier (On Training Dataset)

	Good	Bad	Actual (Total)
Good	699	1	700
Bad	9	291	300
Predicted (Total)	708	292	1000

Table 9. Confusion Matrix – Random Forest Classifier (5 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	**615**	**85**	**700**
Bad	**174**	**126**	**300**
Predicted (Total)	**789**	**211**	**1000**

Table 10. Confusion Matrix – Random Forest Classifier (10 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	**615**	**85**	**700**
Bad	**179**	**121**	**300**
Predicted (Total)	**794**	**206**	**1000**

Table 11. Confusion Matrix – Random Forest Classifier (15 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	**623**	**77**	**700**
Bad	**193**	**107**	**300**
Predicted (Total)	**816**	**184**	**1000**

Table 12. Confusion Matrix – Random Forest Classifier (20 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	**613**	**87**	**700**
Bad	**183**	**117**	**300**
Predicted (Total)	**796**	**204**	**1000**

6.3 Performance of J48 Classifier

The overall assessment summary of J48 Classifier using training set and different cross validation methods is given in Table 13. The performance of J48 Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 5 and Fig. 6. The confusion matrix for different test mode is given in Table 14 to Table 18. J48 Classifier gives 85.5% accuracy for the training data set. Various cross validation methods are used to check its actual performance. On an average, it gives around 71.4% of accuracy for credit risk estimation.

Table 13. J48 Classifier Overall Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	855	145	85.5%	0.2312	0.34	0.19
5 Fold CV	733	267	73.3%	0.3293	0.4579	0.06
10 Fold CV	705	295	70.5	0.3467	0.4796	0.02
15 Fold CV	719	281	71.9%	0.3348	0.4689	0.02
20 Fold CV	698	302	69.8%	0.3571	0.4883	0.02

Table 14. Confusion Matrix – J48 Classifier (On Training Dataset)

	Good	Bad	Actual (Total)
Good	669	31	700
Bad	114	186	300
Predicted (Total)	783	217	1000

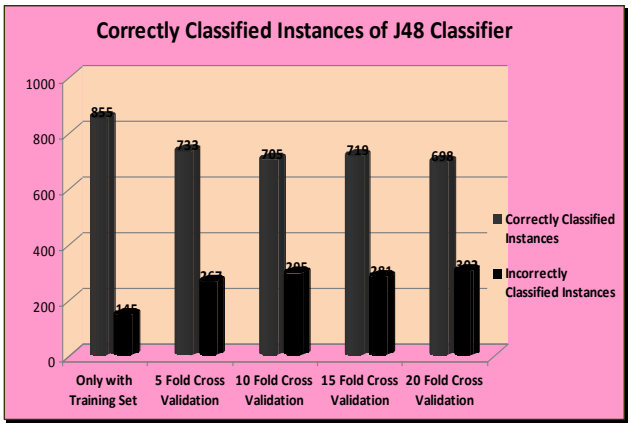

Fig 5: Correctly Classified instances of J48 Classifier

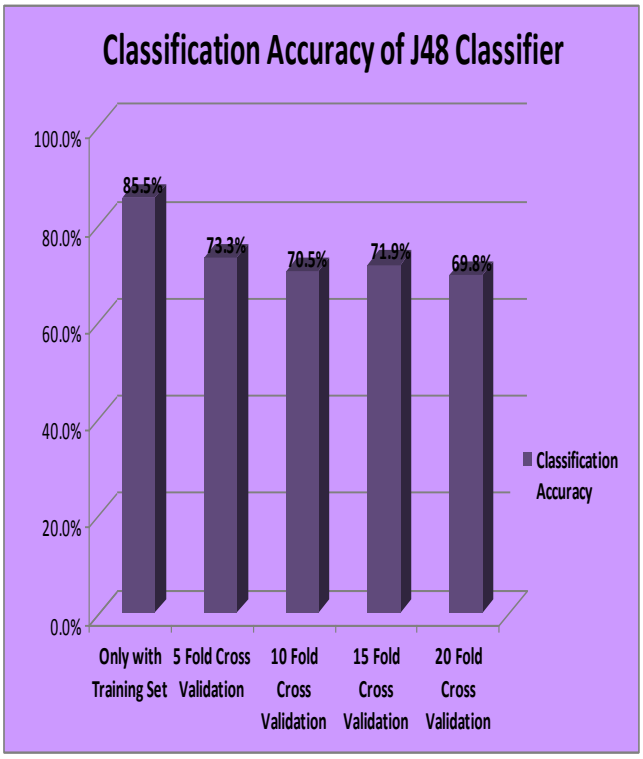

Fig 6: Classification Accuracy of J48 Classifier

Table 15. Confusion Matrix – J48 Classifier (5 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	596	104	700
Bad	163	137	300
Predicted (Total)	759	241	1000

Table 16. Confusion Matrix – J48 Classifier (10 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	588	112	700
Bad	183	117	300
Predicted (Total)	771	229	1000

Table 17. Confusion Matrix – J48 Classifier (15 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	606	94	700
Bad	187	113	300
Predicted (Total)	793	207	1000

Table 18. Confusion Matrix – J48 Classifier (20 Fold Cross Validation)

	Good	Bad	Actual (Total)
Good	586	114	700
Bad	188	112	300
Predicted (Total)	774	226	1000

6.4 Comparison of Random Forest, REP Tree and J48 Classifiers

The comparison of performance between Random Forest, REP Tree and J48 Classifiers is depicted in Fig 7, and Fig. 8 in terms of Correctly Classified Instances and Classification Accuracy. The complete ranking is prepared based on correctly classified instances, classification accuracy, MAE and RMSE values and other statistics found using Training Set result and Cross Validation Techniques. Consequently, it is perceived that Random Forest classifier outperforms the other two Classifiers.

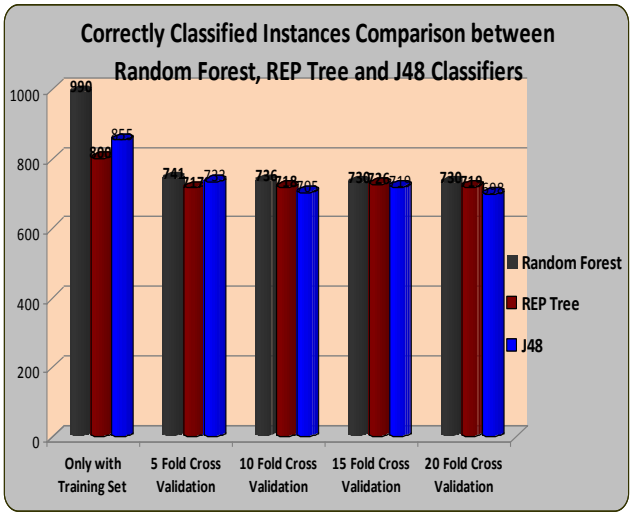

Fig 7: Correctly Classified Instances Comparison between Random Forest, REP Tree and J48 Classifiers

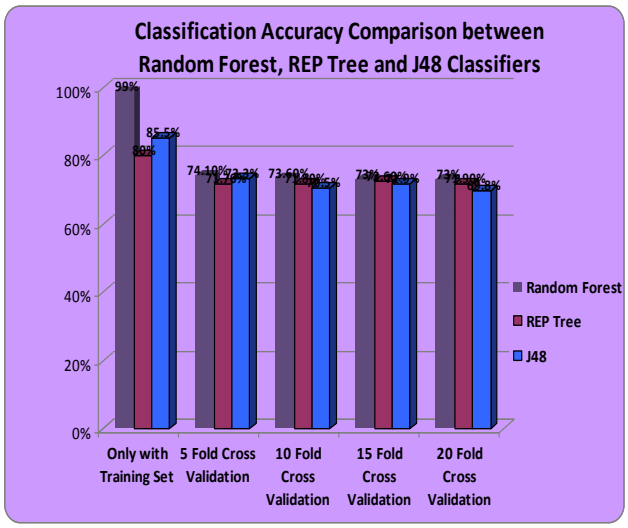

Fig 8: Classification Accuracy Comparison between Random Forest, REP Tree and J48 Classifiers

7. CONCLUSION

This work investigated the efficiency of three different classifiers namely, Random Forest, REP Tree and J48 Classifiers for credit risk prediction. Testing is accomplished using the open source machine learning tool. Also, effectiveness comparison of both the classifiers has been done in view of different scales of performance evaluation. At last, it is observed that Random Forest Classifier performs best, followed by REP Tree Classifier and then by J48 Classifier for credit risk prediction by taking various measures including Classification accuracy, Mean Absolute Error and Time taken to build the model.

8. ACKNOWLEDGMENTS

The author expresses her deep gratitude to the Management of IBS Hyderabad, IFHE University and Operations & IT Department of IBS Hyderabad for constant support and motivation.

9. REFERENCES

[1] Germano C. Vasconcelos, Paulo J. L. Adeodato and Domingos S. M. P. Monteiro. 1999. A Neural Network Based Solution for the Credit Risk Assessment Problem. Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais, (July 1999), 269-274.

[2] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen. 2002. Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications (Elsevier) 23, 245–254.

[3] Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu. 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study," Decision Support Systems (Elsevier) 37, 543– 558.

[4] Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. 2006. Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model. S. Kollias et al. (Eds.): ICANN 2006, Part II, Springer LNCS 4132, 682 – 690.

[5] Eliana Angelini, Giacomo di Tollo, and Andrea Roli. 2006. A Neural Network Approach for Credit Risk Evaluation," Kluwer Academic Publishers, 1 – 22.

[6] S. Kotsiantis. 2007. Credit risk analysis using a hybrid data mining model. Int. J. Intelligent Systems Technologies and Applications, Vol. 2, No. 4, 345 – 356.

[7] Hamadi Matoussi and Aida Krichene. 2007. Credit risk assessment using Multilayer Neural Network Models - Case of a Tunisian bank.

[8] Lean Yu, Shouyang Wang, and Kin Keung Lai. 2008. Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications (Elsevier) 34, pp.1434–1444.

[9] Arnar Ingi Einarsson. 2008. Credit Risk Modeling. Ph.D Thesis, Technical University of Denmark.

[10] Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi. 2011. A Hybrid Method for Credit Risk Assessment of Bank Customers. International Journal of Trade, Economics and Finance, Vol. 2, No. 2, (April 2011).

[11] Vincenzo Pacelli and Michele Azzollini. 2011. An Artificial Neural Network Approach for Credit Risk Management. Journal of Intelligent Learning Systems and Applications, 3, 103-112.

[12] A.R.Ghatge and P.P.Halkarnikar. Ensemble Neural Network Strategy for Predicting Credit Default Evaluation. International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 7, (January 2013), 223 – 225.

[13] Lakshmi Devasena, C. 2014. Adeptness Evaluation of Memory Based Classifiers for Credit Risk Analysis. Proc. of International Conference on Intelligent Computing Applications - ICICA 2014, 978-1-4799-3966-4/14 (IEEE Explore), 6-7 March 2014, 143-147.

[14] Lakshmi Devasena, C. 2014. Adeptness Comparison between Instance Based and K Star Classifiers for Credit Risk Scrutiny. International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, (March 2014).

[15] Lakshmi Devasena, C. 2014. Effectiveness Assessment between Sequential Minimal Optimization and Logistic Classifiers for Credit Risk Prediction. International Journal of Application or Innovation in Engineering & Management, Volume3, Issue 4, (April 2014), 55 - 63.

[16] Lakshmi Devasena, C. 2014. Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, 6156 - 6162.

[17] Lakshmi Devasena, C. 2014. Competency Assessment between JRip and Partial Decision Tree Classifiers for Credit Risk Estimation. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 (5), (May – 2014), 164-173.

[18] UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.

[19] Tina R. Patil, and S. S. Sherekar. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications Vol. 6, No.2, (Apr 2013), 256 - 261.

[20] Witten IH, and Frank E. 2005. Data mining: practical machine learning tools and techniques – 2nd ed. the United States of America, Morgan Kaufmann series in data management systems.

[21] Quinlan J (1987) Simplifying decision trees, International Journal of Man Machine Studies, 27(3), 221–234.

[22] S.K. Jayanthi and S.Sasikala. 2013. REPTree Classifier for indentifying Link Spam in Web Search Engines. IJSC, Volume 3, Issue 2, (Jan 2013), 498 – 505.

[23] Leo Breiman. 2001. Random Forests. Machine Learning. 45(1): 5-32.

[24] Margaret H. Danham, and S. Sridhar. 2006. Data mining, Introductory and Advanced Topics. Person education, 1st Edition