Comparative Analysis of Linear, Non Linear and Ensemble Machine Learning Algorithms for Credit Worthiness of Consumers

David .O. Oyewola[1], Emmanuel Gbenga Dada[2], Oluwatosin Temidayo Omotehinwa[3],Isa.A.Ibrahim[1]
[1]Department of Mathematics & Computer Science, Federal University Kashere, Gombe, Nigeria.
[2]Department of Computer Engineering Faculty of Engineering, University of Maiduguri, Maiduguri, Nigeria.
[3]Department of Mathematical Sciences, Achievers University, Owo, Nigeria.
Corresponding author: David.O.Oyewola (davidakaprof01@yahoo.com)

ABSTRACT - We apply machine-learning techniques to construct linear, non-linear and ensemble machine learning algorithms. The linear comprises of logistic regression(LR), linear discriminant analysis(LDA) and least absolute shrinkage and selection operator (LASSO) while non-linear comprises of support vector machine(SVM), neural network(NN) and decision tree while ensemble learning are boosting, bagging and weighted average. We compared linear, non-linear and ensemble learning using the entire feature in consumer credit data sets. In the first experiment, the study revealed that ensemble learning performs significantly well follow by non-linear and linear machine learning. In the second experiment, we use the selected features from genetic algorithms. It was observed that there was a slight improvement in the result obtained from some of the linear and non-linear machine learning but slight decrease in the result obtained from ensemble learning. The grading scale was used on the overall accuracy of area under the curve(AUROC) in which bagging and weighted average score between 90-100 grade and performs excellently well, boosting obtained 80-90 grade, support vector machine obtained 70-80 grade, logistic regression, linear discriminant analysis, neural network and decision tree obtained 60-70 grade and lasso performs worse with 50-60 grade.

Keywords: Credit Worthiness, Consumers, Machine Learning, Ensemble Learning, Genetic Algorithms.

I. INTRODUCTION

The critical factor in macroeconomic and fundamental risk is delinquency and defaults of buyer credit [1]. Credit agencies gather consumer credit information in order to compute a score to assess the credit value of the client; In any case, these measurements do not adjust rapidly to changes in customer behaviours and economic situations after some time [25][26]. The requirement for credit analysis was conceived in the beginnings of trade in collaboration with the borrowing and lending of cash, and the buying authorization to pay any obligation in future. However, the advanced ideas and thoughts of credit scoring analysis rose around 70 years back with [8]. From that point forward, merchants have begun to accumulate data on the candidates for credit and list them to decide between loan or not certain measure of cash [4][22][20]. As indicated by Thomas [31] credit scoring is "a set of choice models and their basic methods that guide credit banks in the conceding of credit". A more extensive definition is considered in the present work: credit scoring is a numerical articulation in light of a level investigation of client credit value, an accommodating tool for appraisal and aversion of default risk, a vital technique in credit risk assessment, what's more a functional exploration region in financial risk management [6]. In the meantime, the modern statistical and data mining techniques have given a huge commitment to the field of data science and are fit for building models to quantify the risk level of a single client conditioned to his attributes, and after that characterize him as a good or a bad payer as per his risk level. In this way, the primary thought of credit scoring models is to recognize the highlights that impact the installment or the non-installment conduct of the costumer as well as his default risk, occurring the classification into two groups portrayed by the decision on the acknowledgement or dismissal of the credit application [14]. Ref. [17] employed support vector machines. They utilized backpropagation neural network (BNN) as a benchmark and obtained prediction accuracy around 80% for both BNN and SVM strategies for the United States and Taiwan markets. Be that as it may, only slight change of SVM was observed. Rule extraction methods was proposed for SVM and present two others taken from the artificial neural networks area, being Trepan and G-REX. The depicted techniques are compared utilizing publicly accessible datasets, such as, Ripley's synthetic Dataset, multi-class iris dataset, medical diagnosis and credit scoring where intelligibility is a key prerequisite and even an administrative proposal. Their outcomes demonstrate that the SVM rule extraction techniques lose just a little percentage in performance compared to SVMs and in this manner rank at the highest point of comprehensible classification strategies by [7].

Ref. [15] and [24] noticed that in building practical scoring models, an extensive variety of statistical and more recently non-linear methods have been utilized. Here, the utilization of more complex non-linear techniques, such as neural networks, and support vector machines, to build credit scoring applications has seen critical increment in the detailed accuracy and performance on benchmark datasets [3]. Ref. [18] and [28] concurred that such progressed statistical techniques give a better option to traditional statistical methods, for example, discriminant analysis, probit analysis and logistic regression, when building practical models. This perspective was additionally embraced by [23] who believed that the utilization of advanced techniques, such as neural networks, was fundamental on account of the capacity to model credit scoring data that exhibit cooperations and curvature. This can be contrasted with traditional linear techniques, such as, linear/logistic regression and linear discriminant analysis. Ref. [2] apply machine learning methods to develop nonlinear non parametric forecasting models of buyer credit risk. By collaborating client transactions and credit agencies information for a sample of a noteworthy commercial bank's clients.

They are capable to build out-of-sample forecasts that essentially improve the classification rates of credit card holder delinquencies and defaults, with linear regression of delinquencies of 85%. Ref. [32] use least squares SVMs with a Bayesian kernel to infer classifiers for corporate insolvency. They find no significant contrast between SVM, LR and LDA in terms of proportion of test cases correctly classified and no distinction between LR and SVM in terms of area under the ROC curve. Ref. [19] find SVMs outperform multi-layer perception for consumer credit information, but their outcomes are additionally based on a small sample size. Ref. [16] compare SVMs and a back-propagation neural network to predict corporate credit ratings however find a significant differences in performance. Ref. [17] find SVMs classify credit applications no more accurately than artificial neural nets (ANN) , decision trees or genetic algorithms (GA), and compared the relative significance of utilizing features chose by GA and SVM alongside with ANN and genetic programming. However, they utilize data sets far smaller and with fewer features than would be used by a financial institution and do not compare the features selected by SVM alone, nor do they compare with methods used in practice such as LR.

II. METHODOLOGY

The real world data sets use in this study is presented in Table 1, the Australian credit data sets, are accessible from the UCI Repository of Machine Learning Databases [27] and are received thus to evaluate the predictive accuracy. The Australian credit data comprises of 307 instances of creditworthy applicants and 383 instances where credit is not creditworthy. Each occurrence contains 6 nominal, 8 numeric attributes, and 1 class attribute (accepted or rejected). This dataset is fascinating in light of the fact that there is a descent blend of attributes: continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a couple of missing values.

To secure the secrecy of information, the attributes names and values have been changed to meaningless representative data. In this study, linear (logistic regression, linear discriminant analysis and least absolute shrinkage and selection operator) and nonlinear (support vector machine, artificial neural network and decision tree) machine learning algorithms, along with three ensemble techniques (bagging, boosting and weighted average) are built and compared to each other using their predictive accuracy. To evaluate the performance of the predicted models a 10 fold cross validation approach was utilized. In 10 fold cross validation the whole dataset is partitioned into 10 mutually exclusive folds. Each fold is utilized once to test the performance of the predicted model that is generated from the consolidated information of the remaining nine folds, leading to 10 independent performance estimates. Classification accuracy highly dependents on the nature of the features in a dataset which may contain unessential or redundant data.

Genetic Algorithms was utilized to upgrade the accuracy of the listed predictive model and to determine the best features among the dataset and all the predictive models was subjected to the selected feature and the best accuracy was also acquired. The fundamental point of utilizing Genetic Algorithm for feature selection is to works on the feature set to

reduce the number of features and improve the classification accuracy simultaneously.

A pictorial depiction of this evaluation process is shown in Fig. 1.

TABLE 1: Australian Credit data on UCI Repository of Machine Leaning Databases

Number Attributes	Number of Instances	Nominal	Numeric
14	690	6	8

A. Logistic Regression (LR)

Binary Logistic regression is a traditional statistical technique that is appropriate for examining the connection between a binary categorical response variable and at least one categorical or continuous independent variables. The model is generally presented in the following format:

$$\ln\left(\frac{p}{1-p}\right) = b_o + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (1)$$

Where $\ln(\frac{p}{1-p})$ defines the natural logarithm of the odds ratio, b signifies the coefficients of parameters and x represents the independent variables.

B. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a strategy created by [10]. It can likewise be called Fisher Discriminant Analysis. The primary goal of LDA is to isolate samples of distinct groups. Basically, it transforms data to an alternate space which optimally recognizes classes which can be allured to as the "between class (S_b)" and "within class (S_w)" are characterized as:

$$S_b = \sum_k(\mu_k - \mu)(\mu_k - \mu)^T \quad (2)$$

$$S_w = \sum_k \sum_{i\in k}(x_i - \mu_k)(x_i - \mu_k)^T \quad (3)$$

Where μ_k is the mean of class k and μ is the global total mean. The aggregate covariance matrix is $S_t = S_b + S_w$. The central idea is to separate distinctive classes as much as could be expected (maximize the between – class scatter S_b) while gather each class as much as possible (minimize within-class scatter S_w)

C Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO is a shrinkage technique like ridge, with subtle but vital differences. Assume we have a set of training data set $y = (y_1, \ldots, y_n)^T$ is the response vector and $x_j = \left(x_{1j}, \ldots, x_{nj}\right)^T$, $j = 1, \ldots, p$ are linearly independent predictors. The LASSO estimate is defined by:

$$\beta^{lasso} = argmin \sum_{\iota=1}^{N} \acute{\left(y_\iota - \beta_0 - \sum_{J=1}^{p} x_{\iota J\beta_J}\right)}^2 \quad (4)$$

$$subject\ to\ \sum_{j=1}^{p}|\beta_j| \le t \quad (5)$$

where t is the upper bound for the whole of the coefficients. This LASSO problem can likewise be written in Lagrangian form equation as follows:

$$\beta^{lasso} = argmin\{\sum_{\iota=1}^{N} \acute{\left(y_\iota - \beta_0 - \sum_{J=1}^{p} x_{\iota J\beta_J}\right)}^2 + \lambda \sum_{J=1}^{p}|\beta_J|\} \quad (6)$$

Where λ is the non negative regularization parameter.

D. Support Vector Machine (SVM)

This procedure is a statistical classification technique and presented by [34]. Given a training set $\{x_i, y_i\}$, with $i = \{1, \ldots, n\}$ where x_i is the explanatory variable vector, and y_i represents the binary category of interest, and n signifies the number of dimensions of input vectors. SVM attempts to locate an optimal hyper-plane, making it a non-probabilistic binary linear classier. The optimal hyper-plane could be written as follows:

$$\sum_{i=1}^{n} w_i x_i + b = 0 \quad (7)$$

where $w = w_1, w_2, \ldots, w_n$ is the normal of the hyper-plane, and b is a scalar threshold.

E. Artificial Neural Network (ANN)

A neural network was presented by [29] is a framework based on input variables, moreover known as explanatory variables, joined by linear and non-linear interactions through one or more hidden layers, resulting in the output variables, also called response variables. Neural networks were made trying to reproduce the human brain, since it is depends on

sending electronic signals between an immense number of neurons. The NN structure have components which get a measure of stimuli (the input variables), create synapses in a few neurons (activation of neurons in hidden layers), and results in reactions (output variables).

F. Decision Trees (DT)

Decision trees are a straight forward model that makes prediction by splitting the training data into pieces and essentially retained the outcome for each piece [11]. Likewise called classification and regression trees or CART, it is a natural non-parametric supervised learning model that produces exact predictions by effortlessly interpretable rules. The guidelines can be composed in plain English and can be effectively translated by human beings. The transparency of the models makes them extremely material to economic and financial applications. Besides, it can deal with both continuous and discrete data.

G. Ensembles

An ensemble comprises of a set of independently trained classifiers whose predictions are joined while characterizing novel instances. Past research has demonstrated that an ensemble
is often more accurate than any of the single classifiers in the ensemble. Bagging [5], Boosting [11]; [30] and Stacking are moderately yet prevalent strategies for delivering ensembles.

G.1 Bagging Bagging (Treebag) [5] is a "bootstrap" [9] Bagging, or bootstrap aggregation, utilizes bootstrap sampling to make numerous data sets on Collections on which method is then performed. The final prediction depends on an average of all the predictions made for each observation. By and large, bagging decreases the variance while leaving bias unaffected. The bagging algorithms conceptual procedure is outline:

Algorithm 1

For *t* number of iterations:

1. Test *n* observations with replacement *t* times to make *t* data sets of size *n*.
2. Apply the regression tree to every one of *t* data sets to make *m* models.
3. Store the *m* results.
    1. Get the estimate from the predicted the class of *n* observations of the original data set.
    2. Average the class predicted.

In bagging, the resampling of the training set is not subject to the performance of the prior classifiers. In this work we inspect powerful algorithms Bagging Classification and Regression Tree (CART) (treebag).

G.2 Boosting

Boosting (Stochastic Gradient Boosting) [11]; [30] envelops a group of strategies. The focal point of these methods is to create a series of classifiers. The training set utilized for every individual of the series is chosen based on the performance of the earlier classifier(s) in the series. In Boosting, examples that are erroneously predicted by past classifiers in the series are chosen more regularly than examples that were correctly predicted. Along these lines, Boosting attempts to produce new classifiers that are better to predict cases for which the present ensemble's performance is poor. In this research, we analyse powerful types of Boosting named Stochastic Gradient boosting which depends on a gradient descent search for optimizing the underlying loss function to determine both the weights and the learner at each iteration [12]. The Stochastic Gradient Boosting algorithm with the gradient boosting modification depends on the regularization parameter $\nu \in [0, 1]$, the alleged learning rate. A reasonable model of the algorithm is given:

Algorithm 2

1: Initialize $F(x) := 0$
2: for $m = 1$ to M do
3: Set $w_i = -\frac{\delta L(y,g)}{\delta g}|_{g=F(x)}$
4: Fit $y = \eta(h_m(x))$ as the base weighted classifier utilizing $|w_i|$, with training sample π_m
5: Compute line search step $\alpha_m = argmin_\alpha \sum_{i \epsilon \pi_m} L(y_i, F(x) + \alpha\eta(h_m(x_i)))$
6: Update $F(x) = F(x) + \nu\alpha\eta(h_m(x_i))$
7: end for

On account of exponential loss, the line search step solution Step 3 can be written as:

$$\alpha_k = \alpha_{k-1} - (\eta^T P(\alpha_{k-1})\eta)^{-1} (y^T P(\alpha_{k-1})\eta) \quad (8)$$

where $P(\alpha_{k-1}) = diag(P_i(\alpha_{k-1}))$, $P_i(\alpha_{k-1}) = \omega_i e^{-\alpha y_i \eta_i}$ and $w_i = e^{-y_i F(x_i)}$

G.3 Weighted Average(WA)

The weighted average rule joins the mean and the weighted majority voting rules, where the weights are connected not to class labels, but rather to the actual continuous outputs. The weights can be acquired during the ensemble generation as part of the regular training, or a different training can be utilized to acquire the weights, for example, in Mean error. Normally, every classifier gets a weight, despite the fact that it is possible to allot a weight to every class output apply distinctive weights to every one of the algorithms [14]. The algorithms models for the weighted average is provided:

Algorithm 3

1: K closest neighbourhood, Logistic regression and Random forest algorithms are trained on training data. These are the three classifiers utilized in this research.

2: Apply cross validation (k-fold) on all the three models utilizing training data.

3: Calculate and save predicted probabilities from every one of the 10 folds.

4: Combine the predicted probabilities.

5: Determine least error on data set prepared in step 4.

6: Assign weights on every one of the coefficients based on the least error.

7: Predict on test data using trained models.

8: Calculate Ensemble Learning Prediction Probability Score by multiplying weights with predicted scores.

Ensembling Learning = $w_1 p_{knn} + w_2 p_{lr} + w_3 p_{rf}$

Where w_1 is the weight of k closest neighbourhood, w_2 weight of logistic regression, p_{knn} predicted probability of k closest neighbourhood algorithm, p_{lr} predicted probability of logistic regression and p_{rf} predicted probability of random forest.

H. Genetic Algorithms (GA)

Performing feature selection with GAs requires conceptualizing the procedure of feature selection as an optimization problem and after that mapping it to the genetic structure of random variation and natural selection. Individuals from a given generation of a population mate to create offspring who acquires genes (chromosomes) from both parents. Random mutation modifies a small part of child's genetic material. The offspring of this new generation who are genetically most fit produce the next generation to come. In the feature selection setting, individuals progress towards becoming a prediction problem. Chromosomes (sequences of genes) are modelled as vectors of 1's and 0's with a 1 showing the presence of a feature and a 0 its absence. The reproduced genetic algorithm at that point does the following: it chooses two individuals, randomly chooses a split point for their chromosomes, maps the front of one chromosome to the back of the other (and vice versa) and after that randomly transforms the subsequent chromosomes according to some predetermined probability. In this study, table 2 give point of interest parameters utilized as a part of choosing the best features of the Australian credit Data sets:

TABLE 2: Parameters Set for GA

Parameter	Value
Population of Chromosomes	[0,1]
Total number of Chromosomes	200
Maximum Generation	50
Mutation Probability	0.9
Crossover	Single point crossover
Mutation	Single point mutation
K fold cross validation	10

III. RESULTS

In the first set of experiments, we utilized every features in the Australian credit data set and the 10 fold cross validation on all the machine learning utilized in this paper. Based on the10-fold cross-validation for Linear machine learning, the Least Absolute Shrinkage and Selection Operator (LASSO) produced the best results with an overall prediction rate of 67.49%, and the Linear Discriminant Analysis (LDA) came out as the runner up with an overall prediction rate of 67.29%, followed by Logistic Regression(LR) with overall prediction rates of 67.08% (see Table 3).

A careful examination of these results uncovers that the prediction accuracy for the "Good" class is significantly higher than the prediction accuracy of the "Bad" class. In fact, all the three linear model types predicted the customers who are likely to receive good loan with better than 66% accuracy while they did poorly on predicting the bad loan. In non-linear machine learning, Decision Tree (DT) produced the best results with an overall prediction rate of 71.64%, and Neural Network (NN) came out second with an overall prediction of 67.91%, followed by Support Vector Machine (SVM) with an overall prediction rate of 67.49% (see Table 4).

Ensemble model was also considered in this paper, Bagging (treebag) produced the best result with overall prediction rate of 99.59%, while Weighted Average (WA) came out second with an overall prediction of 94.41% and boosting (sgb) fell on third position with an overall prediction of 76.19%. The results revealed that bagging and Weighted average perform excellently well than boosting, results is shown in Table 5. By comparing Ensemble, Nonlinear and linear models of machine learning. The results showed a better performance in bagging and weighted average of ensemble learning to a total above 94%. This proved that combining different machine learning can definitely improve the accuracy of the model.

TABLE 3: Linear Prediction results for 10 fold cross validation

	LR		LDA		LASSO	
	Bad	Good	Bad	Good	Bad	Good
Bad	1.04%	1.45%	1.03%	1.24%	0%	0%
Good	31.47%	66.05%	31.47%	66.25%	32.56%	67.49%
Balanced Accuracy	50.52%		50.67%		50%	
Overall Accuracy	67.08%		67.29%		67.49%	

TABLE 4: Non Linear Prediction results for 10 fold cross validation

	SVM		NN		DT	
	Bad	Good	Bad	Good	Bad	Good
Bad	0%	0%	8.90%	8.49%	12.42%	8.28%
Good	32.56%	67.49%	23.60%	59.01%	20.08%	59.21%
Balanced Accuracy	50%		57.41%		62.97%	
Overall Accuracy	67.49%		67.91%		71.64%	

TABLE 5: Ensemble Models Prediction results for 10 fold cross validation

	Boosting(sgb)		Bagging(treebag)		WA	
	Bad	Good	Bad	Good	Bad	Good
Bad	10.55%	1.86%	32.09%	0%	26.92%	0%
Good	21.95%	65.63%	0.41%	67.49%	0.41%	67.49%
Balanced Accuracy	64.86%		99.36%		91.4%	
Overall Accuracy	76.19%		99.59%		94.41%	

In the next round of experiments, we utilized Genetic Algorithms for feature selection of the dataset in which nine out of fifteen was selected from Australian credit data set. In realizing this, we utilized all the selected features for all the algorithms which include linear, non linear and ensemble models. Using a 10-fold cross-validation methodology, we developed and tested prediction models for all the model types. The results of these experiments are shown in Table 6, 7 and 8. Based on linear prediction results, logistic regression (LR) and linear discriminant analysis (LDA) shows a slight improvement rate of 67.49% (see Table 6) when compared with results obtained in Table 3 but there is no improvement in least absolute shrinkage and selection operator (LASSO). In nonlinear prediction results, there is an improvement in decision trees (DT) and Neural network (NN) as shown in Table 7 but no improvement in the accuracy results of support vector machine (SVM). Also, in ensemble models it turns out that there was a slight improvement in overall accuracy of boosting methods (see Table 8) but there was a slight decrease of overall accuracy of 99.38% in bagging and wider decrease overall accuracy of 88.41% in weighted average. By comparing the results of Table 3 – 8, the results obtained shows that feature selection can improve the overall accuracy of linear (logistic regression (LR) and linear discriminant analysis (LDA), non-linear (decision tree (DT) and neural network (NN)) and ensemble models (boosting(sgb)) but no improvement in the other two ensemble model which includes bagging and weighted average.

TABLE 6: Linear Prediction results for 10 fold cross validation for GA Selection

	LR		LDA		LASSO	
	Bad	Good	Bad	Good	Bad	Good
Bad	1.04%	1.04%	1.04%	1.04%	0%	0%
Good	31.47%	66.46%	31.47%	66.46%	32.56%	67.49%
Balanced	50.83%		50.83%		50%	

Accuracy Overall Accuracy	67.49%	67.49%	67.49%

TABLE 7: Non Linear Prediction results for 10 fold cross validation for GA Selection

	SVM		NN		DT	
	Bad	Good	Bad	Good	Bad	Good
Bad	0%	0%	3.52%	2.69%	10.97%	5.8%
Good	32.51%	67.49%	28.99%	64.80%	21.53%	61.7%
Balanced Accuracy	50%		53.42%		62.58%	
Overall Accuracy	67.49%		68.32%		72.67%	

TABLE 8: Ensemble Models Prediction results for 10 fold cross validation for GA Selection

	Boosting(sgb)		Bagging(treebag)		WA	
	Bad	Good	Bad	Good	Bad	Good
Bad	11.59%	2.07%	29.41%	0.41%	31.64%	0%
Good	20.91%	65.42%	3.11%	67.08%	0.83%	67.49%
Balanced Accuracy	66.3%		99.04%		82.17%	
Overall Accuracy	77.02%		99.38%		88.41%	

AUROC is a shortened form for area under receiver operation characteristic curve. Accuracy of the models is measured by AUROC. Fig. 1 is the graphical representation of values acquired from AUROC, the red shading bar showed all the feature used in the data sets while the blue shading bar displayed the selected nine features from the credit data sets. The blue and red shading is grouped in pairs based on the results obtained from AUROC. Table 9 comprises of percentage grade obtained from Fig.1 and models is categorized in the grades they belong. Bagging and WA is graded in A category due to its excellent performance; Boosting is graded in B category; SVM is graded in C category; LR, LDA, NN, DT is in D category while LASSO perform woefully categorized in E.

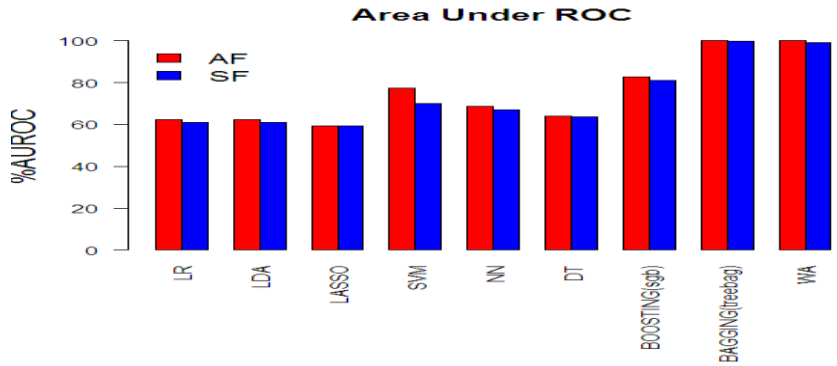

FIGURE 1. Graphical representation of the area under receiver operating characteristics curve

TABLE 9: Grading Scale for AUROC

Grade	Scale	Grade Description	Prediction Models
A	90-100	Excellent	Bagging, WA
B	80-90	Good	Boosting
C	70-80	Fair	SVM
D	60-70	Pass	LR, LDA, NN, DT
E	50-60	Fail	LASSO

IV. DISCUSSION AND CONCLUSION

In this comparative study we have experiments on various machine learning models, and studied their performance in real-life Australian credit data sets. Linear, Non linear and Ensemble method that have yet to be fully researched in credit risk worthiness were also chosen to give a broader review of the techniques available. The results of these experiments show that the bagging and weighted average classifiers performed well with overall performance of between 94-99%. It does appear that in extreme cases the ability of bagging and weighted average can be used to modeled all the features in the data set.

The linear and nonlinear prediction model used in this paper gave results that were reasonably competitive between the two techniques and this competitive performance continued even after feature selection of the data sets. This would suggest that the currently most popular approaches are fairly robust. On the other hand, techniques such as LASSO and SVM did

not showed any better improvement after feature selection. It can also be concluded that the use of Bagging and weighted average would be beneficial in the scoring of data sets where a very large feature was used.

Further work that could be conducted, as a result of these findings, would be to firstly consider other ensembles approach which are average, majority voting and stacking approach. Such an approach would allow in picking a base-learner to select the best model to classify an observation. However, another interesting extension to the research would be to apply these techniques on much larger data sets which display a wider variety of class distributions. It would also be of interest to look into the effect of not only the percentage class distribution but also the effect of the actual number of observations in a data set.

REFERENCES

[1] A. Husejinovic, D. Keco and Z. Masetic, "Application of Machine Learning Algorithms in Credit Card Default Payment Prediction", *International Journal Of Scientific Research*, Vol. 7, Issue 10, pp.1-2, 2018.

[2] A. E. Khandani, A. J. Kim and A. W. Lo, "Consumer credit risk models via machine learning Algorithms", *journal of banking & finance*, vol. 34, pp. 2767-2787, 2010.

[3] B. Baesens, T. V. Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring", *Journal of the Operational Research Society*, vol. 54, pp. 1082-1088, 2003.

[4] J. Banasik, J. Crook, L.Thomas, "Not if but when will borrowers default," *Journal of the Operational Research Society*, 50 (12): 1185-1190, 1999.

[5] L. Breiman, "Random forests", *Machine Learning*, vol. 45, issue 1, pp. 5–32, 2001.

[6] B. Florentin, Q. Chen, B. Clark, S. Das, A. W. Lo and A. Siddique. "Risk and risk management in the credit card industry", *Journal of Banking and Finance*, vol. 72, pp. 218–39, 2016

[7] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines", *European Journal of Operational Research*, pp. 1-11, 2007.

[8] D. Durand, "Risk elements in consumer instalment financing." *In: National Bureau of Economics. New York*, 1941.

[9] B. Efron, and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

[10]R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, vol. 7, pp. 179-188, 1986.

[11]Y. Freund, and R. Schapire, "Experiments with a new boosting algorithm", *In Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, pp. 148–156, 1996.

[12]J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, vol. 29, issue 5, pp. 1189–1232, 2001.

[13]M. Graczyk, T. Lasota, B. Trawiński, K Trawiński, "Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal." *In: ACIIDS'10 Proceedings of the Second international conference on Intelligent information and database systems: Part II Proceeding.* Springer-Verlag Berlin, Heidelberg, 2010.

[14] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Morgan kaufmann, 2006.

[15] W. E. Henley, *Statistical aspects of credit scoring*, Open University, 1994.

[16] C. Huang, M. Chen, and C. Wang, "Credit Scoring With A Data Mining Approach Based On Support Vector Machines", *Expert Systems with Applications*, vol. 33, pp. 847–856, 2006.

[17] Z. Huang, H. Chen, C.J. Hsu, W.H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study", *Decision Support Systems*, vol. 37, issue 4, pp. 543-558, 2004.

[18] W. E. Irwin, K. Warwick, and K. J. Hunt, *Neural network applications in control*, 309 pages, 1995.

[19] S. Li, W. Shiue, and M. Huang, "The Evaluation Of Consumer Loans Using Support Vector Machines", *Expert Systems with Applications*, vol. 30, pp. 772–782, 2006.

[20] F. Louzada, P. Ferreira-Silva, and C. Diniz, "On the impact of disproportional samples in credit scoring models: An application to a brazilian bank data", *Expert Systems with Applications*, vol. 39 , issue 9, pp. 8071-8078, 2012.

[21]M. D. Muhlbaier, A. Topalis, and R. Polikar, "LearnCC.NC: combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes", *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 152–168, 2009.

[22] D. Marron, "Lending by numbers': Credit scoring and the constitution of risk within american consumer credit", *Economy and Society*, vol. 36, no. 1, pp. 103-133, 2007.

[23]T. Masters, *Advanced algorithms for neural networks: A C++ sourcebook*: John Wiley & Sons, Inc, 1995.

[24]E. Mays, *Handbook of credit scoring*, Global Professional Publishing, 1995.

[25]M. H. Flood, *Early identification of high-risk credit card customers based on behavioral data*, Master Thesis, Department of Computer Science, Norwegian University of Science and Technology, 2017.

[26]A. Motwani, P. Chaurasiya, and G. Bajajm, "Predicting Credit Worthiness of Bank Customer with Machine Learning over Cloud", *International Journal of Computer Sciences and Engineering*, vol. 6, issue 7, pp. 1471-1477, 2018.

[27]P.M. Murphy, and D. W. Aha, *UCI repository of machine learning databases*, Department of Information and Computer Science, University of California Irvine, CA., 2001.

[28]M. Paliwal, and U.A. Kumar, "Neural networks and statistical techniques: A review of applications", *Expert Systems with Applications*, vol. 36, issue 1, pp. 2-17, 2009.

[29] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[30]R. Schapire, "The strength of weak learnability", *Machine Learning,* vol. 5, issue 2, pp. 197–227, 1990.

[31]L. C. Thomas, D. Edelman, J. Crook, "Credit Scoring and its applications", *Philadelphia, Pa. : Society for Industrial and Applied Mathematics,* 2002.

[32]T. V. Gestel , B. Baesens , J. A.K. Suykens , D. V. D. Poel, D. E. Baestaens, and M. Willekens. "Bayesian kernel based classification for financial distress detection," *European Journal of Operational Research*, vol. 172, pp. 979–1003, 2006.

[33]U. Bhuvaneswari, P. J. D. Paul, and S. Sahu, "Financial Risk Modelling in Vehicle Credit Portfolio", 978-1-4799-4674-7/14/$31.00, *IEEE* 2014.

[34]V. Vapnik, *Statistical learning theory*, Volume 2 of Adaptive and learning systems for signal processing, communications, and control, 1998.