Zimbabwe Journal of Science & Technology pp 26 – 34 Vol.13 [2018]
e-ISSN 2409-0360
Zimbabwej.sci.technol

Credit scoring using machine learning algorithims

Evander E.T. Nyoni[1], Ntandoyenkosi Matshisela[2]

[1]Department of Applied Mathematics
National University of Science and Technology, Box AC 939 Ascot
Bulawayo, Zimbabwe
[2]Department of Operations Research and Statistics
National University of Science and Technology, Box AC 939 Ascot
Bulawayo, Zimbabwe

Email: evandernyoni@gmail.com

ABSTRACT

Credit risk mitigation is an area of renewed interest due to the 2007-2008 financial crises and thus masses of data are collected by the financial institutions. This has left the risk analysts with a daunting task of adequately determining the credit worthiness of an individual. In the search for highly efficient credit scoring models, financial institutions can adopt sophisticated machine learning techniques. We employ the AUROC approach to make a comparative analysis of machine learning methods of classification by performing 10-fold cross validation for model selection on the German Credit data set from the UCI database. The results show that Lasso regression provides the best estimation for default with an AUROC of 0.8048 followed by the Random Forest model with 0.7869 AUROC. The widely used logit model performed better than the Support Vector Machine (Linear) with 0.7678 and 0.7581 AUROC respectively. Moreover, by the Kolmogorov-Smirnov test, we proved that the other machine learning techniques outperform the widely used logit model in how well the model is able to classify "good" class from "bad" class.

Keywords: Machine Learning; Credit Risk; Random Forests; Lasso regression; Support Vector Machine; Logit regression.

Received: 07.08.17 Accepted: 30.11.18

1.0 INTRODUCTION

There is a global competition of banks on market share and an antagonism to gain competitive advantage. Banks and financial bodies have been seen to suffer from high levels of non-performing loans (Moti, Masinde, Mugenda *et al.*, 2013). Such customer behaviour has affected the viability and sustainability of these financial bodies and weakened their growth. A number of risks can cause a bank to fail (Moti, Masinde, Mugenda *et al.*, 2013), but failure to manage credit risk can cause banks to close and lead to its failure to compensate its clients.

The financial markets are dynamic and spontaneous this then calls for constant monitoring and perpetual change of the firms' credit policy. Loaning money to a bad client is costly not only to the banks but a loss of equities by the stakeholders (Hooman, Marthandan and Karamizadeh, 2013). The loss has always been the failure to predict payment defaulting prior the event. (Wehinger, 2012) assumed that the financial crisis has brought other financial woes such as fraud and banking scandals. Such maladies have brought low confidence in the financial industry and raised anxiety in the structural flaws in the methods used by banks to function and the way they are run.

One of the critical arms of banking is the credit function. The interest percentage is the main source of revenue for each and every bank. It is also a reality that risk is intrinsic in each and every loan transaction. Credit risk comes as a result of a debt not being cleared by the borrower. This causes cash flow jams, principal and interest leakages and a lot of collection costs. Credit risk assessment aids

in objective decision making, to decide whether to lend or not and how much to charge for the loan. The construction and implementation of predictive models have shown to be a powerful strategy tool (Moin and Ahmed, 2012). At the heart of modern predictive analytics are various machine learning algorithms that extract hidden insights from masses of data. The data may be multimedia data, text data, web data, time series data and spatial data (Moin and Ahmed, 2012).

Harnessing this huge data helps the bank management to make profitable decisions daily (Sudhakar, Reddy and Pradesh, 2016). The masses of structured and unstructured data leave the human analyst with a daunting of transforming this data into information. As a result, data mining techniques continue to gain popularity (Sudhakar, Reddy and Pradesh, 2016). The growth of mobile and internet banking in Africa has left the door ajar to financial institution leveraging big-data in credit risk mitigation.

In this paper we seek to explore the customer characteristics that signal the capability of a customer to default, and also discover the best credit modelling algorithm that effectively models loan delinquency and credit worthiness using German Credit Data from the UCI Machine Learning Repository.

2.0 MATERIALS AND METHODS

2.1 Logit Regression

The logit model is one of the most widely used algorithms in credit scoring. It is a unique case of the general linear model and in certain respects comparable to linear regression (Bhatia *et al.*, 2017). However, unlike general linear regression, logit models are primarily for predicting dichotomous dependent outcome rather than a continuous outcome. This is achieved by restricting the output from $[-\infty, +\infty]$ to a probability between 0 and 1, owing to a logistic transform. The logistic function is given by the inverse-logit:

$$logit^{-1}(\alpha) = \frac{1}{1 + e^{(-\alpha)}} = \frac{e^{\alpha}}{e^{\alpha} + 1} \qquad [1]$$

For a training set of N data points $H = \{(x_i, y_i)\}_{i=1}^{N}$ and $x_i \in \mathbb{R}^n$ as the input variables. The aim of the logistic regression technique corresponding to a binary outcome $y_i \in \{0, 1\}$ is to estimate $P(y = 1 \mid \boldsymbol{x})$ as follows:

$$P(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{\left(-(w_0 + \boldsymbol{w}^T \boldsymbol{x})\right)}}, \qquad [2]$$

and,

$$P(y = 0 \mid \boldsymbol{x}) = \frac{e^{-(w_0 + \boldsymbol{w}^T \boldsymbol{x})}}{1 + e^{-(w_0 + \boldsymbol{w}^T \boldsymbol{x})}}, \quad [3]$$

where w_0 is the intercept, $\boldsymbol{w}$ is the parameter vector and $\boldsymbol{x} \in \mathbb{R}^n$ is a n-dimensional vector.

To estimate the parameters w_0 and $\boldsymbol{w}$, we employ the maximum likelihood technique. With the probability of observing either outcome given as

$$P(y \mid \boldsymbol{x}) = P(y = 1 \mid \boldsymbol{x})^{y} (1 - P(y = 1 \mid \boldsymbol{x})^{1-y}, \qquad [4]$$

The rationale of this procedure describes the maximization of the likelihood of observing the data set H, given the observations are drawn independently, which yields

$$\prod_{i=1}^{N} P(y_i = 1 \mid \boldsymbol{x}_i)^{y_i} (1 - P(y_i = 1 \mid \boldsymbol{x}_i)^{1-y_i}, \qquad [5]$$

The log-likelihood is then given as:

$$LL = \sum_{i=1}^{N} y_i \log(P(y_i = 1 \mid \boldsymbol{x}_i)) + (1 - y_i) \log(1 - P(y_i = 1 \mid \boldsymbol{x}_i)) \quad [6]$$

The log-likelihood statistic is a performance measure of unexplained information after model fitting, making it comparable to the residual sum of squares. The criterion of a performance measure is given by the magnitude of log-likelihood, where the larger

the statistic, the more unexplained information there is.(Field, Miles and Field, 2013).

2.2 Lasso-Regularized Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is a regression technique that couples regularization with variable selection in order to improve the prediction power of the resulting model. For an input space $X \in \mathbb{R}^N$ and a measurable $Y \in \mathbb{R}$ we consider a family of linear hypotheses $G = \{x \mapsto \boldsymbol{w} \cdot \boldsymbol{x} + b: \boldsymbol{w} \in \mathbb{R}^{\boldsymbol{N}}, b \in \mathbb{R}\}$. Let training sample $S = \{(x_1, y_1), \ldots\ldots, (x_m, y_m)\} \in (X \times Y)^m$. The objective of Lasso is to minimise the empirical squared error on S with the regularization term regulated by the L_1-norm of the weight vector.

$$\min_{(\boldsymbol{w},b)} F(\boldsymbol{w}, b) = \lambda \|\boldsymbol{w}\|_1 + \sum_{i=1}^{m} (\boldsymbol{w} \cdot x_i + b - y_i)^2 \quad [7]$$

where λ is a positive parameter. Equation [7] is an optimization problem since both the $\|\cdot\|_1$ and the empirical error are convex. (The subject of convex norms is beyond the scope of this paper). Therefore, the optimization for equation [7] can be written as

$$\min_{(\boldsymbol{w},b)} \sum_{i=1}^{m} (\boldsymbol{w} \cdot x_i + b - y_i)^2 \quad subject\ to: \|\boldsymbol{w}\|_1 < \psi \quad [8]$$

where ψ is a positive parameter.

2.3 Support Vector Machines (SVMs)

SVM is one of the most effective Artificial Intelligence (AI) algorithms for typical two-class classification problems. SVMs algorithm comes with a unique set of key concepts:

- Maximum margin hyperplane to find a linear classifier through optimization
- Kernel trick to expand up from linear classifier to a nonlinear one
- Soft-margin to cope with the noise in the data.

(Haltuf, 2011). However, we shall delve into the basics for typical binary classification tasks. For a training sample $S = \{(x_1, y_1), \ldots\ldots, (x_m, y_m)\}$ where $x_i \in \mathbb{R}^{\boldsymbol{n}}$ and $y_i \in \{+1, -1\}$. For a linearly separable scenario the data points are properly separated by,

$$\langle w \cdot x_i \rangle + b \geq +1 \; for \; y_i = +1 \quad [9]$$

$$\langle w \cdot x_i \rangle + b \leq +1 \; for \; y_i = -1 \quad [10]$$

Combining [9] and [10] yields the following inequality:

$$y_i(\langle w \cdot x_i \rangle + b) - 1 \geq 0 \; for \; i = 1, \ldots\ldots, m \quad [11]$$

The objective of SVM is to find a hyperplane that that maximizes its distance from the nearest point x_i while separating the data points for which $y_i = +1$ and $y_i = -1$. This is achieved by finding the optimal solution to:

$$\min_{\boldsymbol{w},b} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} \quad subject\ to: \; y_i(\langle \boldsymbol{w} \cdot x_i \rangle + b) - 1 \geq 0 \quad [12]$$

To solve equation [12] one has to find the saddle point of the Lagrange function:

$$L_p(w, b, \alpha) = \frac{1}{2} w^T \cdot w - \sum_{i=1}^{m} (\alpha_i y_i(\langle w \cdot x_i \rangle + b) - 1) \quad [13]$$

where Lagrange multipliers $\alpha_i \geq 0$. In finding the optimal saddle point the L_p should be maximised with respect to the dual feature α_i and minimized with respect to the optimal features w and . The L_p is then transformed into a dual Lagrangian $L_D(\alpha)$:

$$\min_{\alpha} L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \quad s.t : \alpha_i \geq 0, i = 1, \dots, m \; and \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad [14]$$

Introducing the generalized the Karush-Kuhn-Tucker (KKT) constraints and taking the derivative with respect to w and b. The resulting α_i for the optimization problem ascertains the optimal margin classifier and its parameters w^* and b^*. Hence the optimal decision hyper-plane $f(x, \alpha^*, b^*)$:

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{m} y_i \, {\alpha_i}^* \langle x_i \cdot x \rangle + b^* - \sum_{i=sv}^{m} y_i \, {\alpha_i}^* \langle x_i \cdot x \rangle + \; b^* \qquad [15]$$

and indicator decision function sign $[f(x, \alpha^*, b^*)]$. The optimal margin classifier depends exclusively upon the set of vectors for which $\alpha_i > 0$, hence the name support vectors.

Now we extend the above concepts to a case of non-separable data i.e. the case for which equation [15] has no solution. Such a case is solved by introducing slack variables $\vartheta_i \geq 0, i = 1, \dots, m$. The resulting optimization problem

$$\min_{w,b,\vartheta} \frac{1}{2} w^T w + \; C \sum_{i=1}^{m} \vartheta_i \;, \quad st: y_i(\langle w \cdot x_i \rangle + b) + \vartheta_i - 1 \geq \; 0, \qquad \vartheta_i \geq 0 \qquad [16]$$

where C is the penalty feature on the training error and ϑ_i the slack variable, is solved by the Lagrangian approach as in the separable scenario. To find the optimal classifier, the dual Lagrangian $L_D(\alpha)$,

$$\max_{\alpha} L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \qquad [17]$$

is maximised with respect to $\alpha_i$under the constraints:

$$\sum_{i=1}^{m} \alpha_i \, y_i = 0 \; and \; 0 \leq \alpha_i \leq C, i = 1, \dots, m. \qquad [18]$$

where upper bound C (the penalty parameter) is decided by the user. The optimal margin classifier is the same as in [15].

In the case of a nonlinearly separable data SVM employs a kernel (or mapping) function ϕ to map the input space of the training sample into a feature space of higher-dimension. To solve the dual Lagrangian $L_D(\alpha)$, the kernel function ϕ :

$$\left(\phi(x_i) \cdot \phi(x_j)\right) := k(x_i \cdot x_j) \qquad [19]$$

replaces the inner products in [17] giving birth to a nonlinear SVM dual Lagrangian:

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k\langle x_i \cdot x_j \rangle, \quad st: \sum_{i=1}^{m} \alpha_i \, y_i = 0 \; and \; 0 \leq \alpha_i \leq C, i = 1, \dots, m. \qquad [20]$$

Employing the same approach of solving the optimization model in the separable scenario, the resulting optimal decision function is:

$$f(x) = \; sgn\left(\sum_{i=1}^{m} y_i {\alpha_i}^* k\langle \phi(x_i) \cdot \phi(x_j) \rangle + b^* \right) = \; sgn\left(\sum_{i=1}^{m} y_i {\alpha_i}^* k(x_i . x_j) + b^* \right) \qquad [21]$$

Classification accuracy depends on the kernel parameters which should be properly set by the user.

2.4 Random Forest

A random forest is an ensemble learning technique of classification where the classifier takes the form of a collection of tree-structures

$$\{h(\mathbb{x}, \Theta_k), k = 1, ...\} \quad [22]$$

Where the random vectors Θ_k are independent and identically distributed (iid), with each tree registering a decision for the most popular category at input $\mathbb{X}$ and output Θ.

An input random vector $\mathbb{X} \in \mathcal{X} \subset \mathbb{R}^p$ is observed and with this input the aim is to predict the square integral random response $Y = \mathbb{R}$ by estimating the regression function:

$$m(\mathbb{x}) = \mathbb{E}[Y|\mathbb{X} = \mathbb{x}] \quad [23]$$

The assumption is that there is a training data set, $\mathrm{D}_n = ((\mathbb{X}_1, Y_1), \ldots, (\mathbb{X}_n, Y_n))$ which is independent and identically distributed. Using the data set D_n, the following function is created: $m_n = \mathcal{X} \to \mathbb{R}$. In such a case, consistency of the regression function m_n holds if $\mathbb{E}[m_n(\mathbb{X}) - m(\mathbb{X})]^2 \to 0$ as n $\to \infty$. The mean is calculated over $\mathbb{X}$ and data set D_n.

The random decision forests can be thought of as a collection of randomized base regression trees

$$\{r_n(\mathbb{x}, \Theta_m, \mathrm{D}_n), m \geq 1\} \quad [24]$$

Where Θ_k are independent and identically distributed random vectors.

$$r_n(\mathbb{x}, \Theta_j, \mathrm{D}_n) = \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \frac{\mathbb{I}_{\mathbb{X}_i \in A_n(\mathbb{x}, \Theta_j, \mathrm{D}_n)} Y_i}{N_n(\mathbb{x}, \Theta_j, \mathrm{D}_n)} \quad [25]$$

Where $\mathcal{D}_n^*(\Theta_j)$ is the training data set. $A_n(\mathbb{x}; \Theta_j, \mathcal{D}_n)$ is the cell containing input random vector $\mathbb{x}$ and $N_n(\mathbb{x}; \Theta_j, \mathcal{D}_n)$ being the number of points selected in $A_n(\mathbb{x}; \Theta_j, \mathcal{D}_n)$. The trees are then combined to form the finite forest estimate

$$m_n(\mathbb{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{m} m_n(\mathbb{x}; \Theta_j, \mathcal{D}_n) \quad [26]$$

A large number of trees in the forests (M) makes sense in an experimental point of view, the infinite estimate then becomes

$$m_{\infty,n}(\mathbb{x};\ \mathcal{D}_n). = \mathbb{E}_\Theta[m_n(\mathbb{x}; \Theta_j, \mathcal{D}_n)] \quad [27]$$

Where $\mathbb{E}_\Theta$ is the expectation with respect to random feature Θ, conditional on $\mathcal{D}_n$.

2.5 Performance measure

There are many ways of assessing the predictive power of a model i.e. its ability to generalise the rules it has learned from the training data-set to the validation set. To evaluate model performance we compare the Area Under Receiver Operating Characteristic (AUROC) curve, Gini, and Anderson Darling statistic. ROC curve is the commonly used tool in evaluating binary classification problems. It is a graphical representation of the sensitivity against the specificity at various threshold settings. The larger the AUROC curve the better the model. Generally, a very good model would have an area of 0.80-0.89. The Gini coefficient above 0.6 indicates a good model. The Kolmogorov-Smirnov test (KS-test) evaluates how well the algorithm model is able to classify “good” class from “bad

2.6 Data description.

The German credit data set classifies loan applicants as good or bad risks. The data set consists of 1000 instances with 70% accepted and 30% rejected. Each instance (or applicant) is described by 20 attributes, thirteen categorical features and seven numerical. The complete list of attributes is summarized in the table below.

chk_acc_status	c	4	Status of an existing checking account.

duration_month	n		Duration of the loan in a month.
credit_history	c	5	Client's credit history.
purpose	c	10	The purpose of the loan.
credit_amount	n		Credit amount.
savings_acc_bond	c	5	Savings account/bonds.
p_employment_since	c	5	Client's present employment since.
instalment_pct	n		Installment rate in percentage of disposable income.
personal_status	c	5	Personal status and gender.
other_debtors_or_grantors	c	5	Other debtors/guarantors.
residence_since	n		Present residence since.
property_type	c	4	Property.
age_in_yrs	n		Age in years.
other_instalment_type	c	3	Other installment plans.
housing_type	c	3	Housing.
number_cards_this_bank	n		Count of existing credits at this bank.
job	c	4	Job type.
no_people_liable_for_mntnance	n		Count of people liable for maintenance.
telephone	c	2	If the client owns a telephone.
foreign_worker	c	2	Foreign worker.
good_bad	c	2	Risk classification.

Table 1: List of data set features. Second column: c = categorical, n = numerical variable. The third column represents a number of categories for respective variable.

3.0 Results and Discussion.

To avoid over fitting (failing to generalize a pattern) we perform 10-fold cross validation. Only the highly significant variables (predictors) were considered in the models.

In evaluating variable importance we employ the random forests based plot as shown in figure 1 below.

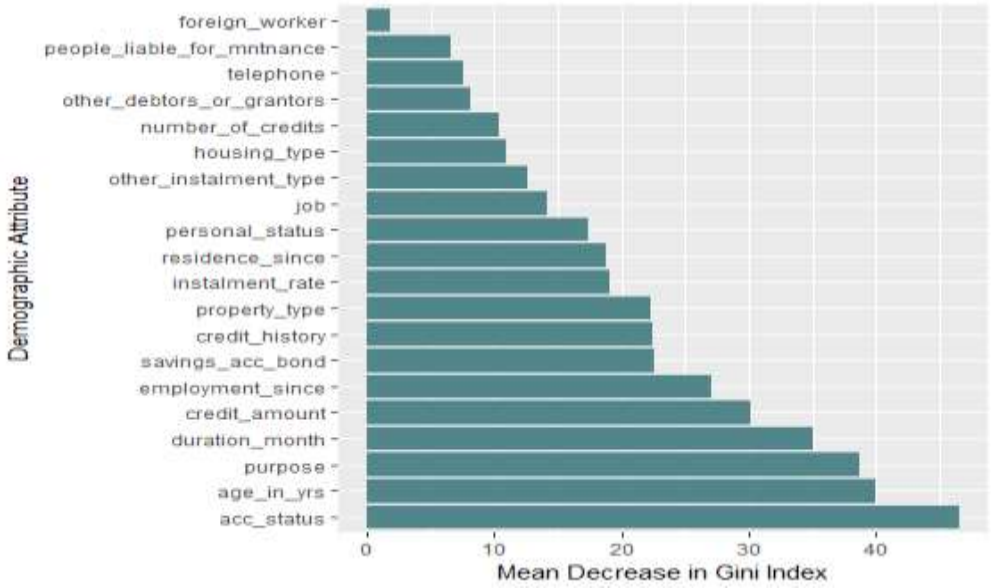

Figure 1: Graph of variable importance.

The higher decrease in Gini (i.e. low Gini) means the variable plays a greater role in partitioning the data into the probable classes. We select the top 10 most significant variables.

Table 2: Table of model performance.

Model	AUC	KS	Gini
Logistic Regression	76.78	42.06	53.56
Random Forest	78.69	43.81	57.38
Support Vector Machine (linear)	75.81	42.22	51.62
Lasso Regression	80.48	48.90	60.96

Using the Area Under the Curve (AUC) it was evident that the Lasso regression model correctly classified 80% of the instances. This high percentage means that the regression model is a very good model in classifying the credit loans. The Lasso was followed by the Random Forest model which accounted for 78.69% and tracked by the traditional classification model, the logistic regression model. It can be clearly seen that the machine learning algorithms perform better than the traditional model accounting up to 4% greater than the traditional model.

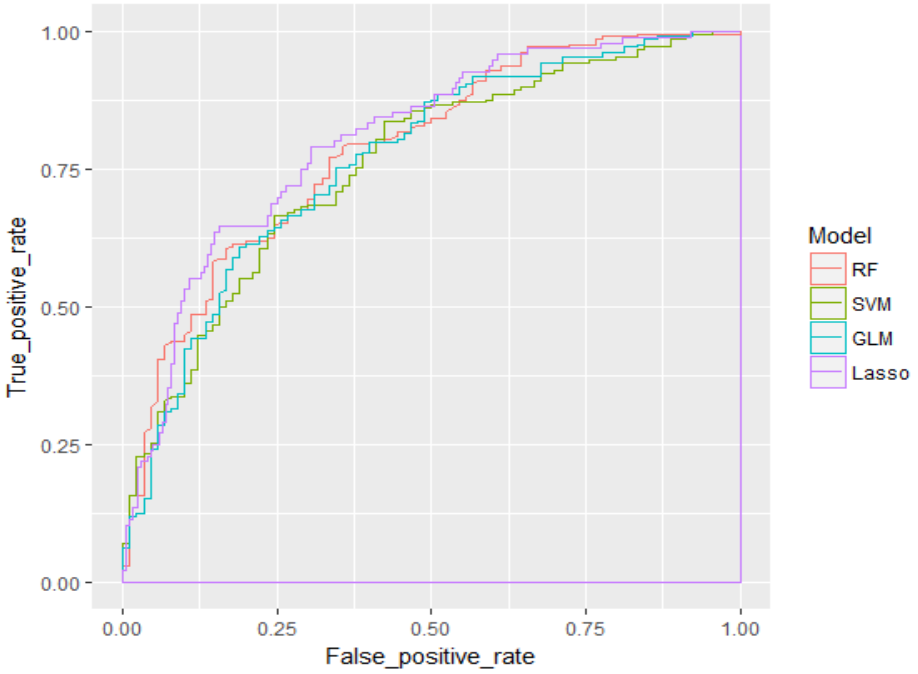

Figure 2: Model performance by AUROC

The ROC curve represents sensitivity (true positive rate, TPR) against specificity (false positive rate, FPR) corresponding to a precise decision threshold. The AUROC curve evaluates how well an algorithm can distinguish between probable groups (good and bad).

To test if the two classes, the bad loans, and good loans, are significantly differentiated by the models, a Kolmogorov Smirnov test is performed. If they are completely dispersed a value of 100 is returned. This then means

that the higher the score the better the model in differentiating the two. It can be seen in Table 2 that the Lasso regression accounted for the highest Kolmogorov-Smirnov score of 48.90. The Support Vector Machine, kernel being Linear, had a higher Kolmogorov-Smirnov score than Logistic regression with a score of 42.22 versus 42.06 despite turning up weaker in terms of the AUC.

Gini is a measure that evaluates the goodness of fit for a binary classification model. Like the other tests, the higher the value means the better the model. Looking at all the evaluation methods it is evident that the lasso is superior followed by the random forest, then the logistic regression and lastly the Support Vector Machine (linear) model.

4.0 Conclusion

The goal of this study was to develop and evaluate the classification data mining techniques. By analysis, it is concluded that the best classifying credit model is the lasso regression algorithm. Our results indicate that the machine learning techniques outperform the traditional method that is the logistic regression. The 4% increase may yield millions of savings for a financial company. We recommend the use of these techniques and also the hybrid models may be investigated in order to increase the performance of the algorithms.

5.0 Acknowledgements

We acknowledge the encouragement and support from friends given during the course of the project. We wish to acknowledge UC Irvine Machine Learning Repository for contributing the dataset used in the study.

References

Haron O. Moti, Justo Simiyu Masinde, Nebat Galo Mugenda, M. N. S., Ismail, S., Thesis, A., Philosophy, D. O. F., Vaish, A. K., Kiplimo, K. S., Kalio, A. M. and Liv, D. (2013) „Effectiveness of credit Management System on loan performance: Empirical Evidence from micro Fianance Sector i Kenya", International Journal of Business, Humanities and Technology, 2(6), pp. 99–108. Available at: http://www.blueorchard.com/wp-content/uploads/2013/05/OID-Final-Report.pdf.

Hooman, A., Marthandan, G. and Karamizadeh, S. (2013) „Statistical and Data Mining Methods

Moin, K. I. and Ahmed, Q. B. (2012) „Use of Data Mining in Banking", International Journal of Engineering Research and Applications (IJERA), 2(2), pp. 738–742. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.7821&rep=rep1&type=pdf.

Sousa, M. de M. and Figueiredo, R. S. (2014) „Credit Analysis Using Data Mining: Application in the Case of a Credit Union", Journal of Information Systems and Technology Management, 11(2), pp. 379–396. doi: 10.4301/S1807-17752014000200009.

Sudhakar, M., Reddy, C. V. K. and Pradesh, A. (2016) „TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE Research Scholar , D epartment of Computer Science and Technology Professor , D epartment of Physics , Rayalaseema University Kurnool , Andhra", 5(3).

Wehinger, G. (2012) „Banking in a challenging environment: Business models, ethics and approaches towards risks", OECD Journal: Financial Market Trends, 2012(2), pp. 79–88. doi: 10.1787/fmt-v2012-2-en.

Bhatia, S. et al. (2017) „Credit Scoring using Machine Learning Techniques", International Journal of Computer Applications, 161(11), pp. 975–8887. Available at: http://www.ijcaonline.org/archives/volume161/number11/bhatia-2017-ijca-912893.pdf.

Field, A., Miles, J. and Field, Z. (2013) Discovering Statistics Using SPSS, Sage. doi: 10.1111/insr.12011_21.

Haltuf, M. (2011) „UNIVERSITY OF ECONOMICS IN PRAGUE Faculty of Business Administration".