Performance Evaluation Criteria of Credit Scoring Models for Commercial Lenders

Bakker Daniel K[#1], Odundo F[*2], Nyakinda J[*3], Paul Samuel F[#4]

# Department of Mathematics, Chemistry and Physics, University of Eastern Africa, Baraton, Kenya
**Department of Applied Statistics, Financial Mathematics and Actuarial Science, Jaramogi Oginga Odinga University of Science and Technology, Kenya*

Abstract:
*Credit scoring has been regarded as a main tool of different companies or banks during the last few decades and has been widely investigated in different areas, such as finance and accounting. Different scoring techniques are being used in areas of classification and prediction, where statistical techniques have conventionally been used. We used ACC rate, which we believed was an important criterion, especially for new applications of credit scoring, because it highlighted the accuracy of the predictions. Confirmation of these values was done by the AUROC. Different Models were examined and the result showed that using Logistic Regression approach, 19.4% of the applicants were predicted false and 80.6% of them correct, which is relatively high compared to the other models, with the highest sensitivity and the lowest Type II error. That is to say, if we were credit officers, we would conclude that the model at hand, predicts 8 out of 10, the true status of each loan candidate***.**

Keywords: *Credit Scoring, True Positive Rate, Misclassification Errors.*

I. INTRODUCTION

Prediction of loan default has an obvious practical utility. The identification of default risk appears to be of paramount interest to banks. A lending major of a bank must evaluate tens or even hundreds of thousands of loan applications each year. These obviously cannot all be subjected to the scrutiny of a loan committee in the way that, say, a real estate loan might. Thus, statistical methods and automated procedures are essential, therefore banks typically should use “credit scoring models"[18]. In principle, the credit score could incorporate any amount of relevant business information. In practice, credit scoring for loan applications appears to be focused narrowly on default risk. Basically, through credit scoring, lenders use scores to determine who qualifies for a loan, at what interest rate, and what credit limits.

Lenders also use credit scores to determine which customers are likely to bring in the most revenue. The use of credit or identity scoring prior to authorizing access or granting credit is an implementation of a trusted system. Particularly, classification methods will provide results used in prediction or estimation [1]. The approaches for predicting qualitative responses is a process that is known as classification. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods, [17].

A. Credit Scoring

Credit scoring is a statistical method used to predict the probability that a loan applicant or existing borrower will default or become delinquent. In other words, credit scoring is a method of evaluating the credit risk of loan applications [10]. Decisions can be made faster and cheaper and more consumers can be approved. It helps spread risk more so vital resources, such as insurance and mortgages, are priced more fairly, [15]. For businesses, especially small and medium-sized enterprises, credit scoring increases access to financial resources, reduces costs and helps manage risk. For the national economy, credit scoring helps smooth consumption during cyclical periods of unemployment and reduces the swings of the business cycle. By enabling loans and credit products to be bundled according to risk and sold as securitized derivatives, credit scoring connects consumers to secondary capital markets and increases the amount of capital that is available to be extended or invested in economic growth [1], [14].

B. Objective

The objective of this study was to evaluate the performance criteria of the statistical credit scoring models.

II. PERFORMANCE EVALUATION CRITERIA

Performance evaluation criteria, such as the Confusion Matrix (CM) or the Average Correct Classification (ACC) rate, the Estimated Misclassification Cost, Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Receiver Operating Characteristics (ROC) curve, GINI coefficient, and other criteria are all used in credit scoring applications under different fields. We discuss some of them as follows:

A. Confusion Matrix (Average Correct Classification Rate Criterion)

This is one of the most widely used criteria in the area of accounting and finance for credit scoring applications in particular, and other fields, such as marketing and health in general. The average correct classification rate measures the proportion of the correctly classified cases as good credit and as bad credit in a particular data-set. The average correct classification rate is a significant criterion in evaluating the classification capability of the proposed scoring models. The idea of correct classification rates comes from a matrix, which is occasionally called “a confusion matrix", otherwise called a classification matrix, [1]. A classification matrix presents the combinations of the number of actual and predicted observations in a data-set. In Yu [20] study, the confusion matrix was compared with another two criteria: Mahalanobis Distance and Kolmogorov-Smirnov Statistics with reference to ROC curve. In other studies this matrix has been compared with MSE and RMSE.

Commonly the mainstream of credit scoring applications either in accounting and finance or other fields have used the average correct classification rate as a performance evaluation measure, [19]. It is believed that the average correct classification rate is an important criterion to be used, especially for new applications of credit scoring, because it highlights the accuracy of the predictions. Yet, the ACC rate criterion does not accommodate differential costs to a bank, arising from different types of error. Specifically, it ignores different misclassification costs for the actual good predicted bad and the actual bad predicted good observations. In the real field it is believed that the cost associated with Type II errors is normally much higher than that associated with Type I errors, [4]. The model performs better if it has a high percentage correctly classified.

Table 1: Classification of two groups

		Predicted Group	
Group	**Observations**	1	2
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

Considering table 1 above, among the n_1 observations in G_1, n_{11} are correctly classified into G_1 and n_{12} are misclassified into G_2 , where $n_1 = n_{11} + n_{12}$. Similarly, of the n_2 observations in G_1 , n_{21} are misclassified into G_1 , and n_{22} are correctly classified into G_2 , where $n_2 = n_{21} + n_{22}$ therefore, the Apparent Error Rate is given as:

$$AER = \frac{n_{12} + n_{21}}{n_1 + n_2} \tag{2.1}$$

B. Estimated Misclassification Cost Criterion

Simply measures the relative costs of accepting customer applications for loans that become bad versus rejecting loan applications that would be good. It is based on the confusion matrix; this criterion gives an evaluation of the effectiveness of the scoring models’ performance, which can cause a serious problem to the banks in the case of the absence of these estimations, especially with the actual bad predicted good observations. The estimated misclassification cost criterion, is a crucial criterion to evaluate the overall credit scoring effectiveness, and to find the minimum expected misclassification cost for the suggested scoring models, [13]. A few credit scoring applications have used the estimated misclassification cost criterion in the field of finance, [1]; [19] and in other fields. The reason, as noted by [16], is that the trustworthy or consistent estimates of the misclassification costs are a complicated and real challenging job to be provided, and, therefore, valid prediction might not be available.

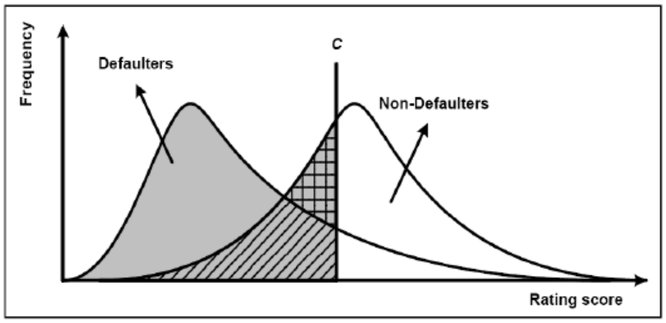
[20] stated that "it is generally believed that the costs associated with (both) Type I error (good credit misclassified as bad credit) and Type II error (bad credit misclassified as good credit) are significantly different" and "the misclassification costs associated with Type II errors are much higher than (the misclassification cost) associated with Type I errors". [19] noted that Dr Hofmann, who compiled his German credit data, reported that the ratio of misclassification costs, associated with Type II and Type I, is 5:1, which has been used by [7] as well. The use of this relative cost ratio has been extended, in terms of sensitivity analysis, to higher cost ratios (i.e. 7:1, 10:1 etc) as noted by [1].

C. The Receiver Operating Characteristics (ROC) curve

Sometimes called "Lorentz diagram", is a two-dimensional graph, which represents the proportion of bad cases classified as bad (called 'sensitivity' which is plotted on the vertical axis) versus the proportion of good cases classified as bad (called '1 minus specificity' which is plotted on the horizontal axis) at all cut-off score values. In fact, sensitivity is equal to 1 minus the Type II error rate, and specificity is equal to 1 minus the Type I error rate, as shown in Figure 1 [4,8,20]. The ROC curve illustrates the achieved overall performance with reference to all cut-off score points.

The construction of a ROC curve is illustrated in the figure 1 below which shows the possible distribution of the rating scores for default and non-default counter-parties. For a perfect rating model the distributions of defaulters and non-defaulters should be distinguished, but in the real world, perfect discrimination in general is not possible, then both distributions will overlap as shown in Figure 1 below. *C* is a cut-off value which provides a simple decision rule to divide counter-parties into potential defaulters and non-defaulters [12].

Figure 1: Rating Score distribution for defaulters & non-defaulters



The ROC curve illustrates the behaviour of classifiers with no regard to misclassification costs or different class distributions; therefore, it effectively separates classification performance from these features. The ROC curve identifies appropriate cut-off score points, whose scores can maximize the Kolmogorov- Smirnov statistic, but it visualizes the details from the Kolmogorov-Smirnov statistic if the ROC is illustrated, Chang [6]. Then four scenarios can occur which are summarized also in the Table 2 below.

Figure 2: Decision Results

		Observed:	
		Default	Non-default
	Above C	TPP	FPP
Scores	Below C	FNP	TNP

It should be emphasised that there are other performance evaluation criteria, such as the GINI coefficient, which "gives one number that summarizes the performance of the scorecard over all cut-off scores" Thomas [17], MSE, RMSE, MAE, and Goodness of Fit test (calibration). Table 3 summarizes some of the performance evaluation criteria investigated by [16].

Table 2: Frequently used performance evaluation criterion

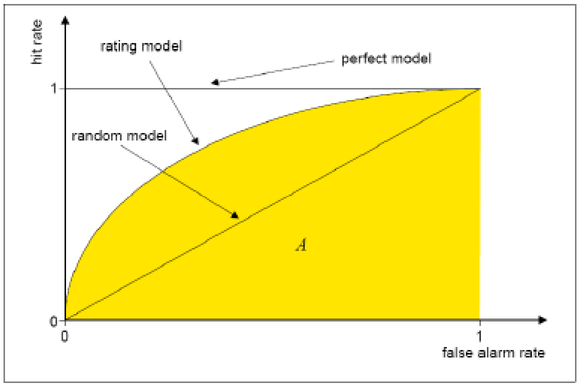
Error Measure	**No. of papers**
Confusion Matrix	36
MSE/RMSE	16
MAE	7
Mean Error	2
R/Adj R-Square	2
Sensitivity/Specificity/ROC	7
Goodness of fit test	3
Discrimination (C-Statistic/AUC)	5

The larger the area of the curve, the better the model. This area is called $AUROC$ denoted by A and calculated as:

$$A = \int_0^1 HR(FAR)d(FAR) \tag{2.2}$$

This area is *0.5* for a random model without discriminative power and *1* for a perfect model.

Figure 3: ROC Curve



III. THE MODELS

A wide range of statistical techniques are used in building the scoring models. Most of these statistical, and some of these non-linear, models are applicable to build an efficient and effective credit scoring system that can be effectively used for predictive purposes. Techniques, such as weight of evidence measure, regression analysis, discriminant analysis, probit analysis, logistic regression, linear programming, Cox's proportional hazard model, support vector machines, decision trees, neural networks, k-nearest-neighbour, genetic algorithms and genetic programming, are all widely used techniques in building credit scoring models by credit analysts, researchers, lenders and computer software developers and providers, [11]. This paper concentrated on Random Forest, Logistic regression, Linear discriminan Analysis, Quadratic discriminant analysis and k-NN models.

A. Validation criteria

Performance evaluation criteria, such as the Confusion Matrix (CM) or the Average Correct Classification (ACC) rate, the Estimated Misclassification Cost, Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Receiver Operating Characteristics (ROC) curve, GINI coefficient, and other criteria are all used in credit scoring applications under different fields, [2,5]. We intend to evaluate our data-set using ACC & AUC. Table 4 below reveals classification results of different scoring models investigated in literature. It can be observed from the table that the probit model has the highest correct total classification rate of 71.9%. Yet, it has the worst rate for classifying bad cases accepted in a good group (i.e. type II error), which are serious misclassifications in practice because of the default implications. By contrast the linear regression model has the lowest bad cases accepted in a good group even though its total correct classification rate is the worst amongst all models. It would be more meaningful to calculate both the type I and type II errors, applying a cost function to each on account of the different associated opportunity costs and produce an overall misclassification score, choosing the optimal model as the one with the lowest misclassification cost, [1,19].

Table 3: Classification results for different scoring models (%)

Model	**Total correct Classification**	**Correct classification of good**	**Correct classification of bad**	**The % of bad accepted in good group**
Discriminant Analysis	65.4	62.2	78	8.1
Linear regression Model	55.1	47	87.5	6.2
Probit Model	71.9	76.4	54.1	13.1
Poisson Model	62.4	57.7	81.8	7.3
Negative binomial II Model	63.3	58.9	80.6	7.6
Two step procedure	64.9	61.1	79.8	7.6

IV. RESULTS

A. Logistic Regression

With the **test** data, we got:

Table 4: Test Accuracy: LR

	Reference		
LogPred	**Not-worthy**	**Worthy**	**Total**
Not-worthy	73	21	**94**
Worthy	58	255	**313**
Total	**131**	**276**	**407**
	Accuracy = (73+255)/407 = 80.59%		

Hence the *LR* Model was *80.59%* accurate.

B. QDA Approach

Quadratic discriminant analysis (QDA) is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements from each class are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. When the normality assumption is true, the best possible test for the hypothesis that a given measurement is from a given class is the likelihood ratio test. The confusion matrix for the **test** data shows the lowest error rate in comparison with the *LDA* as below:

Table 5: Test Accuracy: QDA

	Reference		
Predicted	**Not-worthy**	**Worthy**	**Total**
Worthy	45	263	**308**
Not-Worthy	62	41	**103**
Total	**107**	**304**	**411**
	Accuracy = (263+107)/411 = 79.08%		

with an error rate of *20.92%*, meaning that *QDA* accurate prediction was *79.08%.*

C. LDA Approach

Linear discriminant analysis (LDA), is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. LDA works when the measurements made on independent variables for each observation are continuous quantities. Our **test** set against this model determined its accuracy as follows:

Table 6: Test Accuracy: LDA

	Reference		
Prediction	**Not-worthy**	**Worthy**	**Total**
Not-worthy	63	28	**91**
Worthy	62	245	**307**
Total	**125**	**273**	**398**
	Accuracy = (63+245)/398 = 77.39%		

D. k-NN Approach

K-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. This algorithm segregates unlabeled data points into well-defined groups. Choosing the number of nearest neighbors i.e. determining the value of k plays a significant role in determining the efficacy of the model. Thus, selection of *k* will determine how well the data can be utilized to generalize the results of the k-NN algorithm. A large k value has benefits which include reducing the variance due to the noisy data; the side effect being developing a bias due to which the learner tends to ignore the smaller patterns which may have useful insights. We constructed the **train** and **test** data sets and then predicted on a **test** set *400* observations and *600* we used as **train** set.

Table 7: Test Accuracy: k-NN

	Classified		
	Not-worthy	**Worthy**	**Total**
Not-worthy	2	86	**88**
Worthy	10	302	**322**
Total	**12**	**388**	**400**
	Accuracy = (2+302)/400 = 76.00%		

E. Random Forest Approach

A decision tree classifier uses a structure of branching decisions, which channel examples into a final predicted class value. This machine-learning approach is used to classify data into classes and to represent the results in a flowchart, such as a tree structure. This model classifies data in a data-set by flowing through a query structure from the root until it reaches the leaf, which represents one class. The root represents the attribute that plays a main role in classification, and the leaf represents the class. Splitting the data-set to **train** set and **test** sets, we found the less error rate because our intention was to find the less *Out of Bag* of error rate. For *Random Forest* we calculated the *OOB* error, such that for each bootstrap iteration and related tree, we get the prediction error using data not in bootstrap sample.

Table 8: Test Accuracy: RF

	Reference		
Prediction	**Not-worthy**	**Worthy**	**Total**
Not-worthy	47	25	**72**
Worthy	67	254	**321**
Total	**114**	**279**	**393**
	Accuracy = (47+254)/393 = 76.59%		

V. MODELS COMPARISON

The different models were examined in order to evaluate their performance. Correct classification method and Area Under curve methods were applied.

A. Correctly Classified Method

The table 10 below, reflected the summary of our models comparisons in terms of their correct classification of the applicants. The sensitivity rate is the true positive rate, that is, the percentage of defaulters predicted correctly as defaulters, and specificity is the true negative rate, that is, the percentage of non-defaulters being predicted correctly as non-defaulters. These values of indicators are based on the **training** and the **test** subsets respectively. In order to also compare their predictive models, we examined the **Type I error**, that is, a good credit customer being misclassified as bad credit customer and **Type II error**, that is, a bad credit customer being misclassified as a good credit customer of the models.

Table 9: Comparison of Credit Scoring Models

Model	**Sample**	**Sensitivity**	**Specificity**	**Type I Error**	**Type II Error**	**AUROC**
	Training	0.5691	0.5607	0.4393	0.4309	
k-NN	Validation	0.02	0.98	0.03	0.98	0.5475
	Training	1.000	0.9976	0.0024	0.0000	
RF	Validation	0.6528	0.7913	0.2087	0.3472	0.7652
	Training	0.6328	0.8017	0.1983	0.3672	
LDA	Validation	0.6796	0.8136	0.1864	0.3204	0.7319
	Training	0.6273	0.8030	0.1970	0.3727	
QDA	Validation	0.6019	0.7662	0.1461	0.3981	0.6996
	Training	0.6750	0.8086	0.1914	0.3250	
LR	Validation	0.7766	0.8147	0.1853	0.2234	0.7687

Generally, **Type II errors** are higher than **Type I errors** for the **training** data-sets, except from the **RF**. It was also found that**LR** has the highest *sensitivity* and the lowest *Type II error*. Looking at the **testing subset**, the **k-NN** had the highest *Type II error* and the lowest *sensitivity*. In as much as the **RF** has low *Type II error*, and *sensitivity* of *0.6528*, there was a great difference between the **training** and **testing** data sets. Therefore, **LR**, **LDA** and **QDA** performs best respectively. Testing the discriminatory power is one of the main tasks of the validation of a credit scoring model. It aims to assess the model's ability of separating good customers from bad customers. We could see that of the 5 different models applied, the value of the area under the ROC curves (AUROC) are all pretty high. It means that the model performed well from Logistic regression, Random Forest, Linear discriminant Analysis, Quadratic Discriminant Analysis and lastly k-NN which was almost without discriminatory power.

VI. CONCLUDING REMARKS

The main purpose of this paper was to analyze the data collected and discuss the results obtained by applying the credit scoring models and how they correctly classify cusomers. The data included *1000* personal loans collected from a Kenyan bank and *20* independent variables extracted from the application forms of *2011*. The results indicate that *15* variables were selected to be the best discriminative power between good and bad credits. The percentages correctly classified appeared to be high for both the training and validation data sets. The evaluation of these percentages resulted in the acceptable levels of classification accuracy. Correctly classified results indicated that Logistic Regression is superior to the other models. Other methods like MSE & RMSE can also be applied to such kind of data to check on the predictive power.

ACKNOWLEDGMENT

The authors would like to acknowledge all those who went on relentlessly through this document for proof reading. We thank Dr. Aminer and Lisso Tumaini for their guidance and corrections they have rendered on this document.

REFERENCES

[1] **Abdou, H., Pointon, J**. Credit scoring and decision-making in Egyptian public sector banks. International Journal of Managerial Finance 5 (4): 391-406., 2009

[2] **Al Amari, A.** The credit evaluation process and the role of credit scoring: A case study of Qatar. Ph.D. Thesis, University College Dublin., 2002

[3] **Arnur, I.E.,** Credit Risk Modeling: Journal of Banking and Finance, 77(4), 122-126,, 2013.

[4] **Bierens, H.,** The Logit Model: Estimation, Testing and Interpretation: The Pennsylvania State University, retrieved 18/8/2007,, http:// econ.la.psu.edu/, 2004

[5] **Boyle, M., Crook, J., Hamilton, R., and Thomas, L.,** Methods for credit scoring applied to slow payers. In: Thomas, L., Crook, J., and Edelman, D. (Eds),r, Credit scoring and credit control, Oxford University Press, Oxford, 75-90., 1992.

[6] **Chang, E.,** Developing and validating credit scoring model for Taiwan's enterprises., Paper presented to the 13 th Annual Conference on Pacific Basin Finance, Economics, and Accounting, Livingston Student Centre and Janice H. Levin Building Livingston Campus Rutgers University at New Brunswick, New Jersey, 2005

[7] **Cramer, J. S.,** Scoring bank loans that may go wrong:A case study. Statistica Neerlandica, 2004, 58(3), 365 -380.

[8] **Crook, J**. Credit constraints and US households. Applied Financial Economics, 6(6), 477-485, 1996.

[9] **David Bigman**., Globalization and the Developing Countries: Emerging Strategies for Rural Development and Poverty Alleviation. CABI. p. 136. ISBN 978-0-85199-575-5, 2002.

[10] **Fernandes., F. L.**. Classification methods applied to credit scoring: A systematic review and overall comparison., 2016

[11] **Gietzen, T.** Credit Scoring vs. Expert Judgment, A Randomized Controlled Trial. St. Gallen., 2017

[12] **Greene, W.,** Sample selection in credit-scoring models., Japan and the World Economy, 10(3), 299-316, 1998.

[13] **Hand, D., and Henley, W.,** Statistical classification methods in consumer credit coring: A review. Journal of the Royal Statistical Society, 160(3), 523, 541, 1997

[14] **Kern, A. M.** Credit Scoring Analysis, Springer, 2017.

[15] **Koehler, G., and Erenguc, S.,** Minimizing misclassifications in linear discriminant analysis: Decision Sciences, 21(1), 63-85, 2017.

[16] **Srinivasan, V., and Kim, Y.,** Credit granting: A comparative analysis of classification procedures: The Journal of Finance, 42(3), 665-681, 1987.

[17] **Thomas, L., Thomas, S., Tang, L., and Gwilym, O.,** The impact of demographic and economic variables on financial policy purchase timing decisions: Journal of the Operational Research Society, 56(9), 1051-1062, 2005.

[18] **Thompson, P.** Bank Lending and the Environment: Policies and Opportunities. International Journal of Bank Marketing 16 (6): 243-252., 1998

[19] **West, D., Dellana, S., Qian, J.** Neural network ensemble strategies for financial decision applications. Computers & Operations Research 32 (10): 2543-2559., 2010

[20] **Yu, L., Wang S., Lai, K.** An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring. European Journal of Operational Research 195 (3): 942-959., 2009