Analysis of Individual Loan Defaults Using Logit Under Supervised Machine Learning Approach

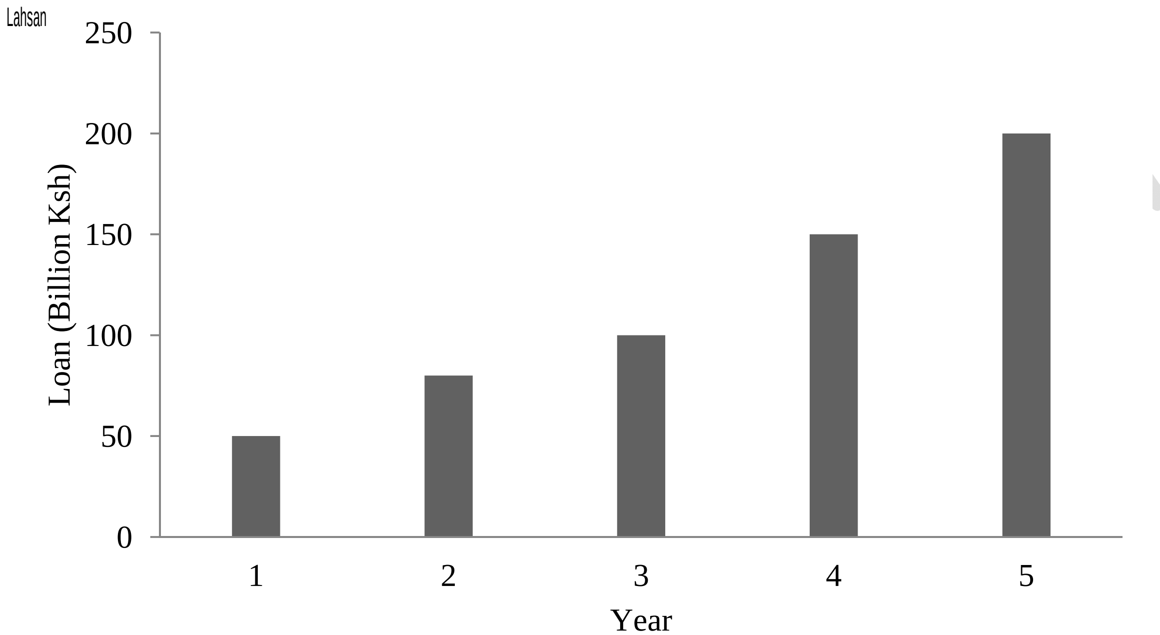
ABSTRACT

Financial institutions have a large amount of data on their borrowers, which can be used to predict the probability of borrowers defaulting their loan or not. Some of the models that have been used to predict individual loan defaults include linear discriminant analysis models and extreme value theory models. These models are parametric in nature since they assume that the response being investigated takes a particular functional form. However, there is a possibility that the functional form used to estimate the response is very different from the actual functional form of the response. The purpose of this research was to analyze individual loan defaults in Kenya using the logistic regression model. The data used in this study was obtained from equity bank of Kenya for the period between 2006 to 2016. A random sample of 1000 loan applicants whose loans had been approved by equity bank of Kenya during this period was obtained. Data obtained was on the credit history, purpose of the loan, loan amount, nature of the saving account, employment status, sex of the applicant, age of the applicant, security used when acquiring the loan and the area of residence of the applicant (rural or urban). This study employed a quantitative research design, it deals with individual loans defaults as group characteristics of a borrower. The data was pre-processed by seeding using R- Software and then split into training dataset and test data set. The train data was used to train the logistic regression model by employing Supervised machine learning approach. The R-statistical software was used for the analysis of the data. The test data set was used to do cross-validation of the developed logistic model which later was used for analysis prediction of individual loan defaults. This study focused on the analysis of individual loan defaults in Kenya using the logistic regression model in Machine learning. The logistic regression model predicted 303 defaults from train data set, 122 non-defaults and misclassified loans were 56 and 69. The model had an accuracy of 0.7727 with the train data and 0.7333 with the test data. The logistic regression model showed a precision of 0.8440 and 0.8244 with the train and test data respectively. The performance of the model with both the train and test data was illustrated using a plot of train errors and test errors against sample size on the same axes. The plot showed that the performance of the model increases with an increase in sample size. The study recommended the use of logistic regression in conjunction with supervised machine learning approach in loan default prediction in financial institutions and also more research should be carried out on ensemble methods of loan defaults prediction in order to increase the prediction accuracy.

Keywords: *loan defaults, loan default prediction, logistic regression model, Kenya*

1. Introduction

Loan defaults in Kenya are on the rise and this is a critical source of economic strain. For this reason, these defaults must be controlled and monitored (Divino *et al.,* 2013). The main importance of the financial institutions, particularly banks are to safeguard the money kept by their clients and make it accessible when need arises. They also advance loans to their customers (Kugiel & Jakobsen, 2009). There has been a growing concern about the relative regression on loans performance in commercial banks in Kenya (Evusa *et al.,* 2015).



Source: CBK, 2015

Figure 1: Total Non-performing loans in the bank industry (Billion Shilling).

In Kenya, several predictive models have been used to predict loan defaults (Ojala *et al.,* 2015). These models include; linear discriminant analysis, logistic regression models and generalized extreme value regression models. All these models are parametric since they assume the response being investigated takes a particular functional form. Logistic regression model has been used to analyze default risk. Martin *et al*., (2010) applied logit model as the basis for developing financial ratios and probabilistic prediction of bankruptcy. The results showed that coefficient estimates for this model were efficient in the use of relatively small samples because it overcomes problems arising from linear regression (Agbemava, 2016). Lahsana & Wah (2010) emphasized that credit risk decisions are key determinants for the success of financial institutions because of huge losses that result from wrong decisions. Hence, credit risk evaluation is essential before making any lending decision (Bekhet & Elletter, 2014). Due to the significance of credit risk, a number of studies have proposed embracing statistical modelling in banks to improve their risk assessment models and hence increase the prediction accuracy of existing models (Akkoc, 2012; Al-Kassar & Soileau, 2014; Jones & Hensher, 2004; Premachandra, Bhabra & Sueyoshi, 2009; Vuran, 2009; Mckee & Lensberg, 2002). Artificial Neural Networks, genetic algorithms, genetic programming, and some hybrid models have been used to evaluate credit risk with promising results in terms of performance accuracy. These models have several drawbacks: (1) lack of explanatory power; (2) reliance on the restrictive assumptions of statistical techniques; and (3) numerous variables, which result in multiple dimensions and complex data (Chen & Cheng, 2013). The hybrid models were found to perform better in terms of prediction accuracy and precision.

Survival analysis models have also been proposed to monitor credit risk modelling, such as Banasik *et al*. (1999), followed by Glennon & Nigro (2005), Bellotti and Crook (2009), Cao *et al.* (2009*).* Dirick *et al*. (2015) and concluded by Dirick *et al* (2017). These studies compared the methods on the development sample and on random cross-validation samples. From this point of view, it has been shown by Stepanova and Thomas (2002) and Tong *et al.* (2012), that the survival analysis models have a similar performance to the logistic regression in terms of precision. Classical linear technique models have also been employed to predict loan defaults (Zhou & Hastie, 2005). They fitted a decision rule based on the area under the curve, as well as root-mean-square error criteria with other non-parametric models classified as machine learning and deep learning, this includes, a random forest model, a gradient boosting machine and four deep learning models. The Ordinary Least Square (OLS) regression and calibrated Beta distributions for statistical inference have also been used to monitor the credit worthiness of a client (Zhang, 2014, 2016). The OLS regression model is simple with the normality assumption, which would not capture the typical features of loan defaults. Beta distributions offer a simple, parsimonious way of capturing a very broad range of distributional shapes over the unit interval.

Artificial neural network (ANN) has also been applied on credit prediction (Arisawa & Watanda, 1994). It is a stylish credit prediction model that draws attention from numerous modelers with its high forecast accuracy, from the past years. Although ANN has several flaws, for instance, a propensity to become trapped in a local optimum, short of descriptive power, expensive training time, overfitting, and requiring a huge amount of instances learning. These has been concurred by the introduction of Support Vector Machine (Vapnik, 1995). It is a comparatively new machine learning method and gained more popularity due to many gorgeous features and outstanding generalization performance on extensive applications. Support vector machine is designed to reduce structural risk by reducing the upper bound of the generalization error rather than the training error, hence solving the problem of overfitting. Support vector machine also solves linearly constrained quadratic programming problems by training it so that the solution is always distinctive and globally optimal, unlike neural networks' training which requires nonlinear optimization (Vapnik 1998). Zhou *et al.* (2010) used least square Support Vector Machine (SVM) with several parametric models for credit scoring and drew the conclusion that K-nearest neighbour outperformed on traditional measures of correctly classified samples, diagonal quadratic discriminant analysis on specificity and SVM on sensitivity for UK database. Voting ensemble outperformed on accuracy, diagonal linear discriminant analysis on specificity, neural network on sensitivity for German credit database. Hu & Ansell (2007) focused on US retail market credit prediction; using four methodologies with SVM, they concluded that different models had different classification abilities on the area under the receiver operating characteristics curve. Chen (2011) compared SVM with some traditional statistical methods and he found out that the rankings of the models differ on overall accuracy, precision, true positive rate and true negative rate. The analogous study did by Tinoco & Wilson (2013) on several logit models with different categories of explanatory variables using Gini index and Kolmogorov-Smirnov statistic as a measure of discriminatory power and concurred with the findings. Van Gestel *et al*, (2006) used least squares SVM with a Bayesian kernel to derive classifier for corporate bankruptcy and found out that there was no significant difference among Support vector machine, logistic regression and discriminant analysis. Zhong *et al*, (2014) used SVM with other two algorithms for credit rating analysis, and the results showed that SVM performs well on rating distributions and neural network approaches outperform SVM on reliability.

Logistic regression model had not been used to analyze individual loan defaults in Kenya. Logistic regression model is simple and flexible in terms of analysis and classification of loan defaults. This statistical analysis of individual loan defaults in Kenya was done by employing statistical learning in R under supervised machine learning algorithm.

This study used R-Statistical software (R-Core team 2017) to analyze secondary data obtained from Equity bank for a period between 2006-2016. Probabilities of loan defaults were determined by using logistic regression model in machine learning.

2. Methodology

This study was carried out at Equity bank headquarters. This study employed a mixed method research design. This design adopts both quantitative and qualitative approaches or methods in a single study (Tashakkori & Creswell, 2003). The study is not restricted by the use of traditional approaches to collect data but guided by foundation of enquiry that underlies the research activity. The data that was used for this study was obtained from the Equity Bank of Kenya headquarters from 2006-2016. This enabled the monitoring of long-term loans. The data were obtained for all applicants whose loans were approved at the Equity bank during this period. The sample size for this research represented 30 percent (30%) of the data collected from equity bank of Kenya. A stratified random sampling was used. According to Mugenda and Mugenda (1999), stratified random sampling achieves desired representation from various subgroups in the population. Data analysis was done using logistic regression model in R statistical software (R-core team, 2017) under a supervised machine learning approach. The first step was to filter the data by cleaning it through seeding in R-statistical software. The data was then coded for easy analysis using the R- software. The coding involved identification of a non-performing loan or a loan default with a value 1 and a performing loan with a value 0. Equivalent number of dummy variables were created for the purposes of coding independent variables. The clean data was then used for analysis and generation of descriptive statistics and also fit the models.

This study fitted the logistic regression model. This model was implemented by machine learning technique using the R software (R-Core team, 2017). In fitting the model by machine learning, the data set was divided into a training set and a testing set. The training set had a sample of 700 applicants. The machine was trained to divide the sample into seven sub samples. That is, a sample of 100, 200, 300, 400, 500, 600 and 700. The logistic regression model was fitted using each subsample and tests the behaviour of the model obtained against the test data in each case. The reason for this was to help in observing whether increasing the sample size increased the performance of the model. The behaviour of the model with both the test data was shown using a train error and test error curves against the sample size. This postulates the effect of the size of the sample on the effectiveness and performance of the models generated.

3. Results and Discussion

The data used to generate results in this study had a sample of 1000 applicants obtained from a data of 10,000 applicants whose loans were approved for equity bank of Kenya for the years 2006-2016. Eleven (11) variables were considered for the analysis of the data. Data visualization was done using R-statistical software. Logistic regression model was fitted using the data under supervised machine learning approach.

Analysis of loan performance showed that 70% of the individual loans approved by equity bank were performing and 30% were non-performing (Figure 2). The loans performing meant that the loans have been repaid in full or the repayment schedule was being adhered to by the borrowers. The non-performing loans were those that had not been serviced in 90 days. The percentages of non-performing loans across the world between the years 2000-2016 were consistently lower than the performing loans (IMF, 2017). The general over-view is that most individuals that apply for loans do repay. This agrees with the World Bank on Kenyan Report for the years 2006-2014 which states that the percentage of non-performing loans is generally lower compared to the percentage of performing loans. Credit Bank of Kenya credit survey report (October-December, 2017) also concurs with this study, it reported that the percentage of non-performing loans is lower than that of performing loans. Financial institutions have been employing the credit information sharing system in order to determine the creditworthiness of the borrowers before approving the credit, this system has drastically reduced the number of non-performing loans (Credit Reference Bureau, 2013). Some of the reasons that may make individuals default their loans are such as divorce, sickness which may lead to an inability to work, loss of a job, failed business among others (Signoriello, 2010).

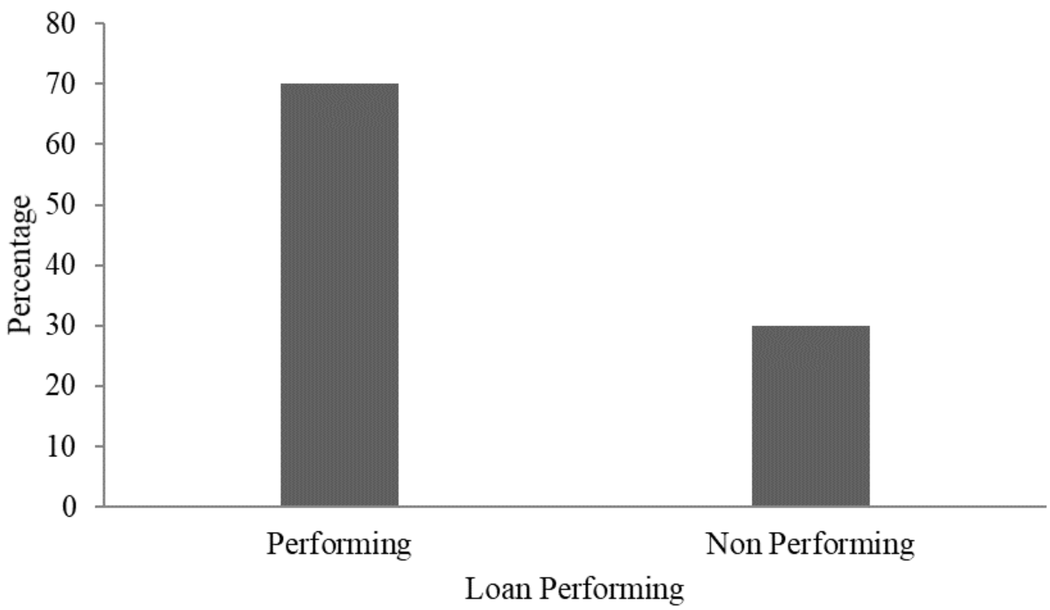


Figure 2: Summary of loan performance at Equity bank between 2006-2016.

This study found out that the average amount of money that the individuals applied for in terms of loans was ksh 327124.80 with a standard deviation of 282275.2 while the average duration was 20.9 months with a standard deviation of 1.06 (Table 1). The average age of the applicants was 28.9 years with a standard deviation of 20.08 while the average number of credits that the individuals had were 1.41 with a standard error of 0.58 (Table 1). The median amount of money borrowed by the applicants was ksh 231,950 while the median duration was 18 months. The median age of the applicants was 25 years while the median number of credits that the individuals had was 1 (Table 1). The skewness of the amount of loan was 1.94 while the kurtosis for duration was 1.09. The kurtosis for amount of loan applied was 4.25 while the one for duration was 0.9. The maximum amount of money applied for by an individual during the study period was ksh 184,400 while the minimum amount applied was ksh 25,000. The maximum duration used to repay a loan was 72 months while the minimum duration was 4 months (Table 1). The study shows that most financial institutions worldwide offer individual loans of up to ksh 5,000,000 though the amount can be higher depending on the value of security that an individual has placed (Arthur & Sheffrin, 2003). Duration of loan repayments are scheduled in months. The longer a person takes to pay a personal loan, the less the monthly payment but that means that the interest paid on the loan will be higher compared to if the loan was paid in a shorter period (Arthur & Sheffrin, 2003).

Table 1: Summary statistics of loans borrowed from equity between 2006-2016.

	Amount	Duration	Age	Number of credits
Mean	327124.8	20.9	28.9	1.41
Standard Deviation	282275.2	12.06	20.08	0.58
Median	231950	18	25	1
Skewness	1.94	1.09	1.09	1.27
Kurtosis	4.25	0.9	0.9	1.58
Maximum	1842400	72	72	4
Minimum	25000	4	4	1

The distribution of the amount borrowed showed that most individuals borrowed loan of up to ksh 500,000 (Figure 3). For an individual to acquire a loan, banks and other financial institutions require collateral. This could be the reason of having fewer persons borrowing loans of beyond ksh 1 million. As also observed on the purposes of the loans, individual's purposes may also not be very demanding to require huge amount of borrowing.

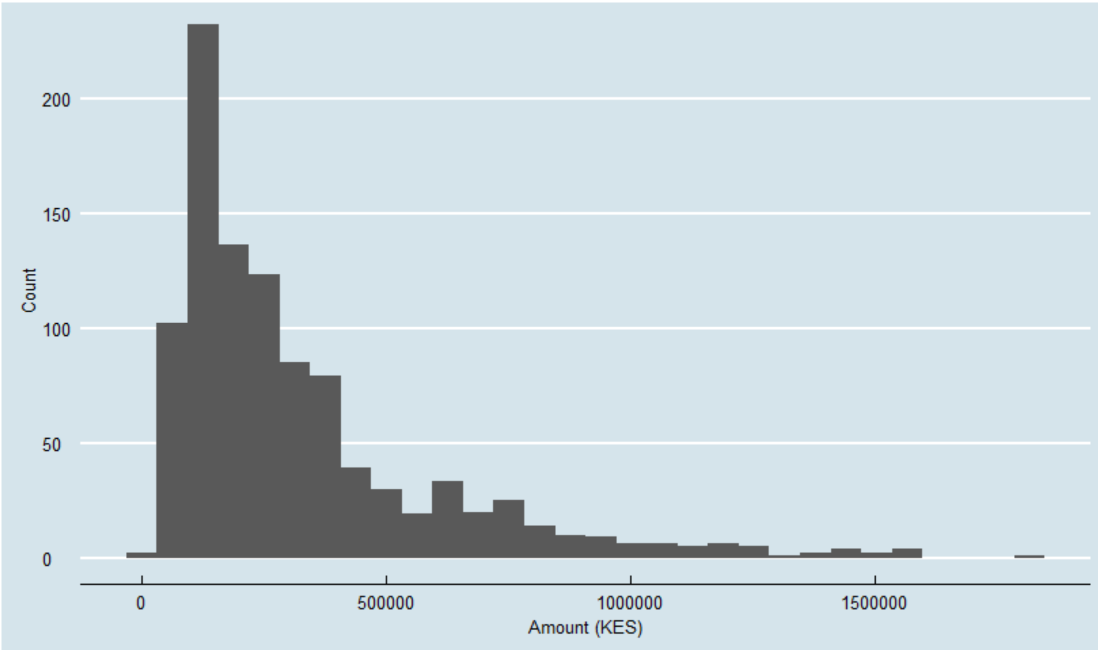


Figure 3: Distribution of amount of loan borrowed from Equity bank between 2006-2016.

Individuals who had acquired a loan for the purpose of financing education showed the highest percentage (44%) of loan defaults (Figure 4). This was followed by those whose purpose of the loan was not classified with a percentage of 42%. The individual loans that showed the best performance were loans acquired by people whose purpose was retraining. Retraining meant to acquire an extra skill such as in-service training. This can be associated to probably that people who go for retraining have already acquired jobs. Thus they do not struggle to repay their loans. Across all the purposes of borrowing the loans the percentages of those who honoured the repayment was always more than those who defaulted.

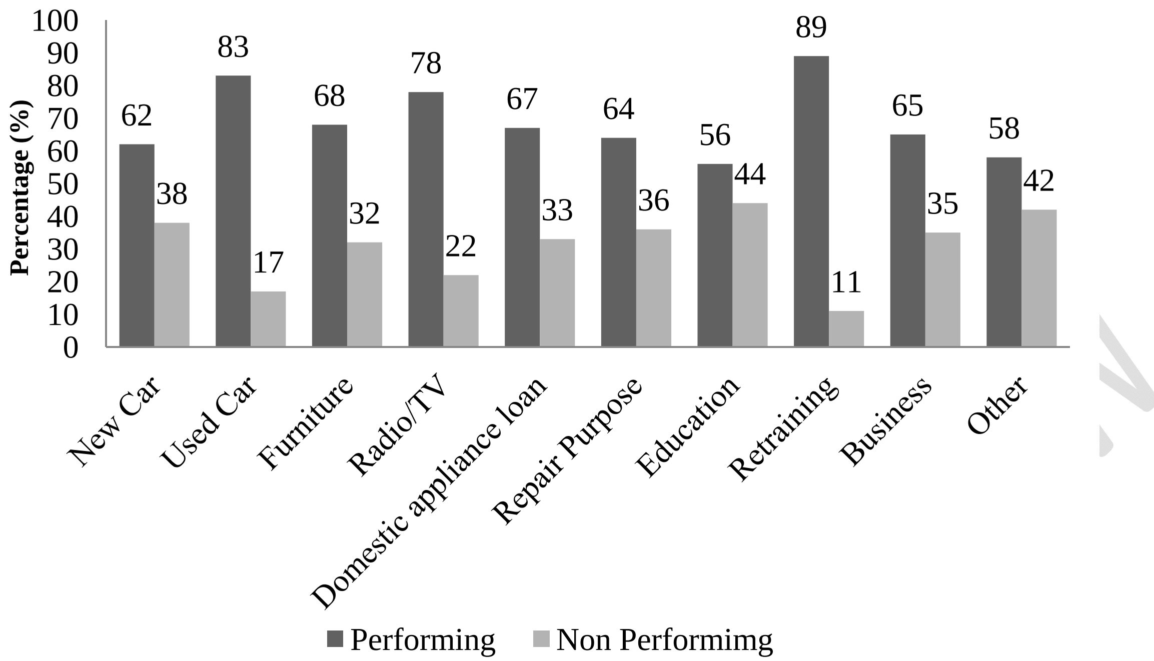


Figure 4: Performance of the loan by purpose of the loan of Equity bank between 2006-2016

Analysis of loans performance by the credit history showed that individuals who acquire loan for the first time showed defaults of 62.5% (Table 2). Those who had a loan before with the bank and had repaid fully showed 57.14 % of loan defaults. Those who had other loans and were still servicing them promptly showed a loan default of 31.89%. The individuals who had defaulted their loans in the past showed the best performance in repaying their loans. The general observation is that it was riskier to give a loan to a new borrower as compared to a borrower whose borrowing history was known. This agrees with Central Bank of Kenya's annual report (2016) that the credit history of a borrower is a key determinant in creditworthiness. The history of how an individual has been servicing the existing or previous loans will determine if he is likely to default. This finding also concurs with a study of Troy Segal (2017) (which showed that the credit history of an individual is vital to guarantee creditworthiness.

Table 2: Loan performance by credit history

Credit History	Preforming Loans		Non-Performing Loans	
	Frequency	Percentage	Frequency	Percentage
No credits are taken	15	37.5	25	62.5
All credits at this bank paid duly	21	42.86	28	57.14
Existing credits paid duly until now	361	68.11	169	31.89
Delay in paying in the past	60	68.18	28	31.82
Credits existing elsewhere	243	82.94	50	17.06

Analysis of duration of employment showed that clients who have been employed for only one year showed the highest percentage of loan default (Figure 5). This is because those who acquired a loan during early years of employment were unable to service the loan leading to defaults. Those who were unemployed have low financial power and this could lead to defaults. Those who have been employed for more than four years showed relatively less percentages in terms of defaults. This shows that employment increases the financial stamina of a person and thus increasing his ability to repay a loan.

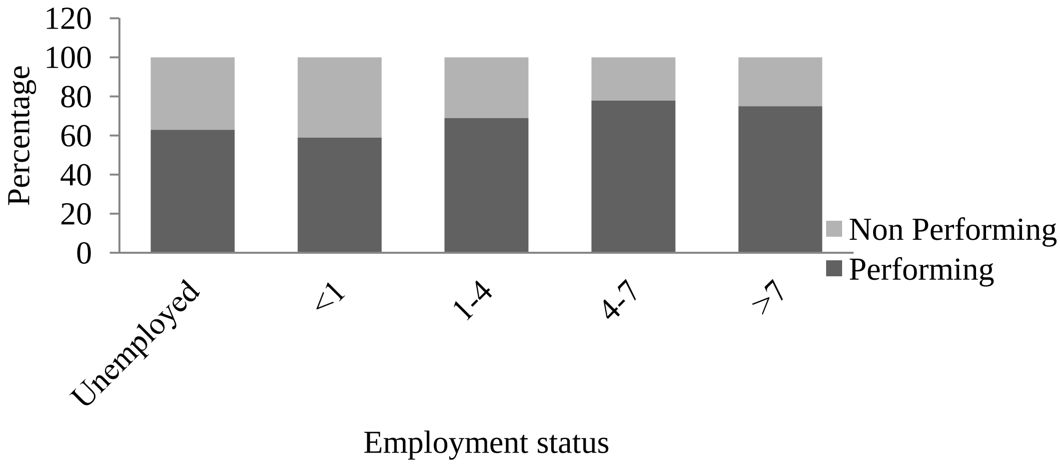


Figure 5: Loan defaults per duration of employment for Equity Bank between 2006-2016

In this study, females showed a poor performance in servicing their loans as compared to males (Figure 6). One of the reasons why individuals default loans are divorce (Appiah, 2011). When a divorce occurs, females are the most affected as compared to men. This could be one of the reasons why the percentages of women who defaulted their loans were more than males. This is in line with studies (Kono, 2006, Morduch, 2000) which found out that female borrowers tend to default more than male. This could be attributed to the way society depicts women in terms of property ownership and acquisition of wealth. Most financial decisions involving women are made by their husbands or their parents and this poses a risk to any amount of credit acquired.

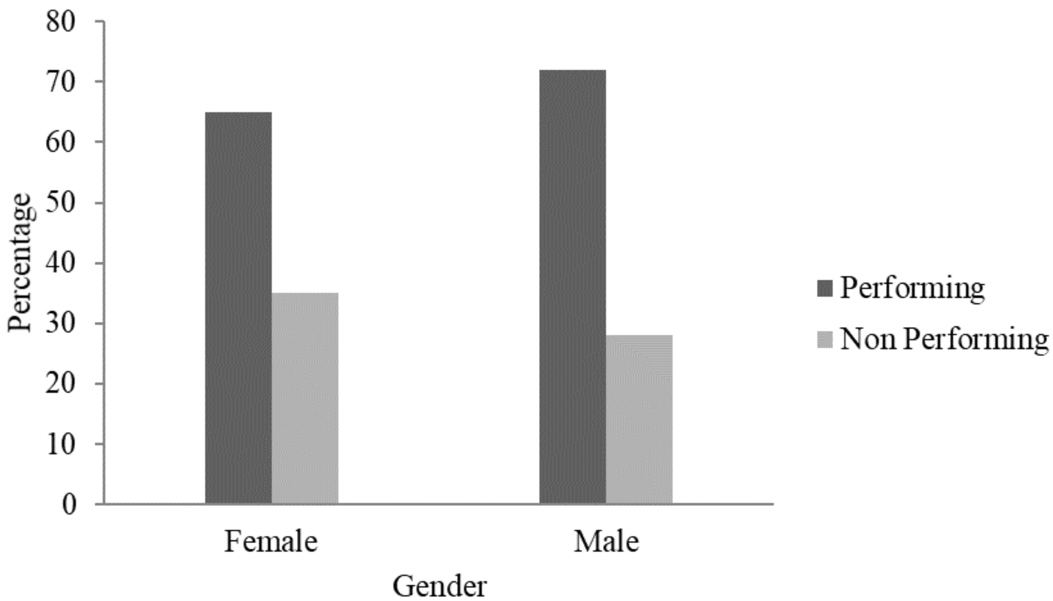


Figure 6: Loan performance by gender for Equity Bank between 2006-2016

Analysis of loans performance using operation status of the current account showed that individuals whose current accounts mostly operated with no money showed higher percentages of loan defaults when compared to individuals whose current accounts operated with some money (Table 3). Individuals who did not have current accounts at the bank showed better performance in repaying their loans but still loan defaults existed. The general observation is that the ability to repay a loan is determined by the financial power of an individual.

Table 3: Performance of loans by a current account at Equity Bank between 2006-2016

Current account	Preforming Loans		Non-Performing Loans	
	Frequency	Percentage	Frequency	Percentage
< 0	139	50.73	135	49.27
0 – 50000	164	60.97	105	39.03
> 50000	49	77.78	14	22.22
No current account	348	88.32	46	11.68

Individuals living in urban areas are more likely to default a loan as compared to individuals living in rural areas (Figure 7). Some of the reasons for loan defaults is the loss of employment or failed business (Trautmann, 2013). Most people in urban areas earn their livelihood through formal employment or business and do not have an immediate back up in case they lose a job or a business fails (Woolridge, 2003). This is unlike in rural areas where even the cost of living is low. This could be one of the reasons why people in urban areas are more likely to be defaulters as compared to people in rural areas.

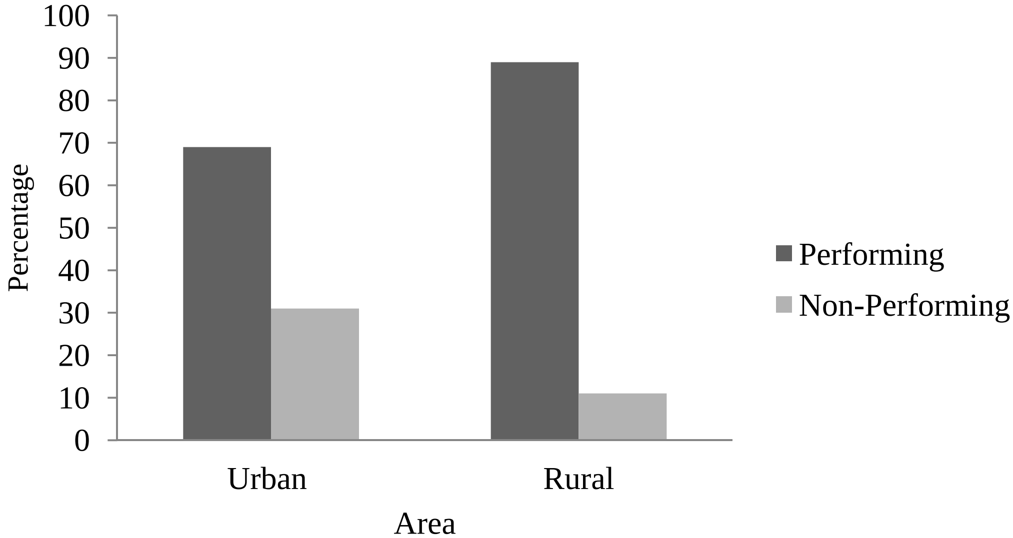


Figure 7: Loan performance by area of residence of loans from Equity Bank between 2006-2016.

Individuals who acquired loans using an item of a high value such as real estate or a farm showed better performance in repaying their loans as compared to individuals who acquired loans without the security (Table 4). It can be argued that when a person borrows a loan from an institution and he/she does not repay, the item placed as security can be confiscated by the institution. For this reason, if a person has used an item of high value as security he is likely to try by all means to repay the loan as compared to a person who has nothing to lose after defaulting apart from being listed with credit research bureau.

Table 4: Loan performance summary by security item at Equity Bank between 2006-2016.

Property	Preforming Loans		Non-Performing Loans	
	Frequency	Percentage	Frequency	Percentage
Real estate/farm	222	78.72	60	21.28
Savings/Insurance	161	69.4	71	30.6
Car	230	69.28	102	30.72
No Property	87	56.49	67	43.51

4. Validation of Logistic Regression Model.

The performance of the model with both the train and the test data was shown using a learning curve (Figure 8). This was a plot of the train errors and the test errors against the sample size on the same axes. This plot showed that the quality of the model increased as the sample size increased. This is in agreement with a study carried out by Dobson (2002). He carried out research on the effects of size of the sample on the performance of generalised linear models. The research revealed that the increase in size of the sample improves the performance of the models. The best model was produced with a sample size of 700. This was the entire data set. It can also be seen that probably by increasing the sample size a better logistic regression model could have been produced.

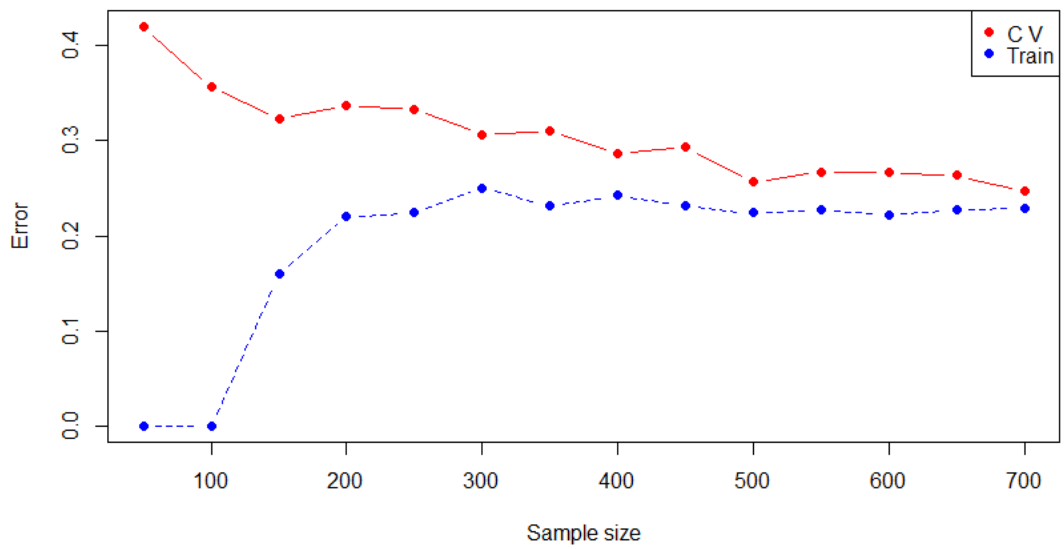


Figure 8: Train errors Vs Test errors plot for the logistic model

5. Conclusion

In this study, logistic model was used for the analysis of individual loan defaults. This study was motivated by the increasing need to explain how individual loan defaults relates to different variables of interest in the Kenyan financial institutions as well as determine how to mitigate the menace of loan defaults.

In order to achieve the objective of the study, the knowledge of machine learning was utilized and implemented for analysis of the data. The data was obtained from the equity bank of Kenya between 2006 - 2016. The data was cleaned and missing values removed through seeding in R., then coded according to the variables for easy analysis. The logistic regression model was fitted using R-statistical software. During the analysis, the data was split into two, train data set and test data set then the probabilities of loan defaults from the train data were developed which enabled data visualization. This helped to tell if an individual is likely to default an individual loan when compared to the Z-score in relation to the variables. A plot for train errors and test errors was developed (Figure 8). This was done in order to determine the effect of increasing the sample size of the study in relation to test errors.

Competing Interests

The authors declare that they have no competing interests.

References

Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic Regression Analysis of Predictors of Loan Defaults by Customers of Non- Traditional Banks in Ghana*. African Journal of Business Management* 10(2), 33-43.

Agbemava, Edinam, (2016). "Logistic Regression Analysis of Predictors of Loan Defaults by Customers of Non-Traditional Banks in Ghana. *"European Scientific Journal*, Esj 12.1 doi: 10.19044/esj. 2016.v12n1p175.

Akkoc, S. (2012). An Empirical Comparison of Conventional Techniques, Neural Networks and Three Stage Hybrid Adaptive Neuro Fuzzy Inference Systems (ANFIS) Model for Credit Scoring Analysis: The Case of Turkish Credit Card Data*. European Journal of Operational Research 222: 168-178.*

Al- Kassar, T.; Soileau, J. (2014). "Financial Performance Evaluation and Bankruptcy prediction (failure)". ARAB ECONOMICS AND BUSINESS JOURNAL, 9, 147-155.

Appiah,K, 2011, Corporate Failure Prediction: Some Empirical Evidence From Listed Firms in Ghana, China-USA Business Review, ISSN 1403-851X, Printed by Elanders Novum.

Arisawa, M., & Watada. J. (1994). Enhanced Learning in Neural Networks and it Application to Financial Statement Analysis. *Paper presented at IEEE International Conference on Neutral Networks.*

Arthur O., & Sheffrin, M. (2003). Economics: Principles in Action. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall. p. 512. *ISBN 0-13-063085-3*.

Banasik, J., Crook, J. N. and Thomas, L.C. (1999). Not If but When Will Borrowers Default. *Journal of the Operational Research Society, 50(12) pp. 1185-1190.*

Bekhet, H., Eletter, S. (2014). "Credit Risk Management for the Jordanian Commercial Banks: Neural Scoring Approach". *Review of Development Finance*, 4, 20-28.

Bellotti, T. & Crook, J. (2009). Credit Scoring with Macro- Economic Variables Using Survival Analysis. *Journal of the Operational Research Society*, 60(12), pp. 89-99.

Cao, R., Vilar, J.M & Devia, A. (2009). Modelling Consumer Credit Risk via Survival Analysis. SORT, 33(1), pp. 187-220.

Central Bank of Kenya, (2016). *Bank Supervision Annual Report*. Nairobi Kenya. Act Press.

Chen, Y.; Cheng, C. (2013). "*Hybrid Models based on Rough Set Classifiers for Setting Credit Rating Decision Rules in the Global Banking Industry*". Knowledge- Based Systems, 39(1), 224-239.

Dirick, L., Claeskens, G. & Baesens, B. (2017). Time to Default in Credit Scoring Using Survival Analysis: A Benchmark Study. *Journal of the Operational Research Society, 68(6)*, pp. 652-665.

Divino, J. A., Lima, E. S., & Orrillo, J. (2013). Interest Rates and Default in Unsecured Loan Markets. *Quantitative Finance*, *13*(12), 1925-1934.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models.* 2nd ed. Boca Rayon: Chapman & Hall/CRC.

Evusa, Z., Mudaki, J. S., & Ojala, D. O. (2015). Evaluation of the Factors Leading to Loan Default at Equity Bank, Kenya. *Journal of Economics and Sustainability.*

Glenon, D. C., & Nigro, P. (2005). Measuring the Default Risk of Small Business Loans: A Survival Analysis Approach*. Journal of Money, Credit, and Banking, 37(5) pp. 923-947.*

Hu, Y. C., & J. Ansell. (2007). Measuring Retail Company Performance Using Credit Scoring Techniques. *European Journal of Operational Research* 183: 1595-1606. doi: 10.1016/j.ejor.2006.09.101.

International Monetary Fund (2017). *Global Stability Report.* www.imf.org>2017> Documents>text*.*

J. Pinheiro, D., Bates, S DebRoy, D., Sarkar, R (2017) C Team R Package Version 3 (57), 1-89.

Jones, S.; Hensher, D. (2004). "Predicting Firm Financial Distress: A mixed Logit Model". The Accounting Review, 79(4), 1011-1038.

Kono, H. (2006). *Is Group Lending a Good Enforcement Scheme for Achieving High Repayment Rates?* Evidence from Field Experiments in Vietnam. Mimeo, Institute of Developing Economies, Chiba, Japan.

Kugiel, L., & Jakobsen, M. (2009). Fund Transfer Pricing in a Commercial Bank. *Master's Thesis, MSC in Finance and International Business*.

Lahsana, A.; Anion, R. & Wah, T. (2010). "Credit Scoring Models using soft Computing Methods: a survey". *International Arab Journal of Information Technology,* 7(2), 115-123.

Martin, Aruldos, Travis Miranda Lakshmi, & Venkatasamy Prasanna Venkatesan. (2010). A Framework to Develop Qualitative Bankruptcy Prediction Rules. *St. Joseph's Journal of Humanities and Science* 1:73-83

Mckee, T.E. & Lensberg, T. (2002). "Genetic Programming and Rough sets: A Hybrid Approach to Bankruptcy Classification". *European Journal of Operational Research*, 138, 436-51.

Morduch, J. (2000). The microfinance schism. World Development, 14(2) 273694.

Mugenda, A., & Mugenda, O. (1999*). Research Methods-Quantitative and Qualitative Approaches*, Nairobi. Act Press.

Premachandra, I.M.; Bhabra, G.S., & Sueyoshi, T. (2009). "DEA as a tool for bankruptcy assessment: A Comparative Study with Logistic Regression technique" *European Journal of Operational Research,* 193(2), 412-424.

Signoriello, J. (2010). *Commercial Loan Practices and Operations,* ISBN 978-1-55520-134-0

Stepanova, M., & Thomas, L. (2002). Survival Analysis Methods for Personal Loan Data. *Operations Research,* 50(2), pp.277-289.

Tashakkori, A., & Teddie, C. (Eds) (2003), *The Handbook of Mixed Methods in Social and Behavioural Research,* Sage, Thousand Oaks, CA.

Tinoco, M.H., & N. Wilson (2013). Financial Distress and Bankruptcy Prediction among Listed Companies Using Accounting, Market and Macroeconomic Variables. *International Review of Financial Analysis* 30: 394-419.

Tong, E. N. C, Mues, C. & Thomas, L. (2012). Mixture Cure Models in Credit Scoring: If and When Borrowers Default. *European Journal of Operational Research*, *218(1) pp. 132-139.*

Trautmann, T., & Vlahu, R. (2013). Strategic Loan Defaults and Coordination: An Experimental Analysis. *Journal of Banking & Finance*, *37*(3), 747-760.

Troy Segal (2017). The Perceived Relevance of Tax Risk-Management in a South African Context. *Meditari Accountancy Research, 25(1), 82-94.*doi:10.1108/medar-01-2016-0008.

Van Gestel, T., B. Baesens, J.A.K., Suykens, D., Van del Poel, D. Baestaens, & M. Willekens. (2006). Bayesian Kernel Based Classification for Financial Distress Detection. *European Journal of Operational Research* 172: 979-1003.

Van Gestel, T., Baesens, B., Suykens, J., Espinoza, M., Baestaens; D.E., Vanthienen, J. & De Moor, B. (2006). “Bankruptcy Predictor with least Squares Support Vector Machine Classifiers”, *Proceedings of the IEEE international Conference on Computational Intelligence for Financial Engineering*, Hong Kong, pp,1-8.

Vapnik, N. (1995). "Support-Vector Networks". *Machine Learning*. 20 (3): 273–297. doi*:*10.1007/BF0099401.

Vapnik, V. N. (1998). *Statistical Learning Theory*, New York; Wiley West, D. (2000) “Neural Network Credit Scoring”, Computer & Operations Research, 27(11): 1131-1152.

Woolrdge, J. (2003). Regression Analysis with Cross Sectional Data*. Introductory Econometrics*: *A Modern Approach (4th ed.).* Cengage Learning.

Woolridge, J. (2003). *Introductory Econometrics:* A Modern Approach. South-Western, Thomson Learning.

Zhang, X., Houzelot, V., Bani, A., Morel, J.L., Echevarria, G., Simonnot, M.O., (2014). *Selection and Combustion of Ni-hyperaccumulators for the Phytomining Process.* Int. J. Phytoremediat. 16, 1058-1072.

Zhong, H., C., Miao, Z., Shen, & Y. Feng. (2014). Comparing the Learning Effectiveness of BP, ELM, I-ELM, and SVM for Corporate Credit Ratings. *Neurocomputing* 128: 285-295.doi:10.1016/j. neucom.2013.02.054.

Zhou, Hui., & Trevor Hastie. (2005). Regulation and Variable Selection. *Via the Elastic Statistical net. Journal of the Royal Society*. 67(2):301-320,2005.

Zhou, L., K.K. Lai, & Yu. Lean. (2010). *Least Squares Support Vector Machines Ensemble Models for Credit Scoring*. Expert Systems with Applications 37: 127-133.doi: 10.1016/j.eswa.2009.05.024.