Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression

Article · November 2015
DOI: 10.19026/rjms.7.2206

3 authors, including:

Charles Kwofie
University of Energy and Natural Resources
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Caleb Boadi
University of Ghana
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Research Journal of Mathematics and Statistics 7(4): 46-52, 2015
ISSN: 2042-2024, e-ISSN: 2040-7505

Submitted: January 14, 2015 Accepted: February 14, 2015 Published: November 25, 2015

Research Article

Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression

[1]Charles Kwofie, [1]Caleb Owusu-Ansah and [2]Caleb Boadi
[1]Department of Statistics, University of Ghana, P.O. Box LG 25,
[2]School of Business, University of Ghana, P.O. Box LG 25, Legon-Accra, Ghana

Abstract: This study examines the performance of logistic regression in predicting probability of default using data from a microfinance company. A logistic regression analysis was conducted to predict default status of loan beneficiaries using 90 sampled beneficiaries for model building and 30 out of sample beneficiaries for prediction. Age, marital status, gender number of years of education, number of years in business and base capital were used as predictors. The predictors that were significant in the model were marital status, number of years in business and base capital. The explained variability in the response variable in the logistic regression was very weak.

Keywords: Default, loan, logistic regression, response variable

INTRODUCTION

The issue of granting loans to loan applicants has been a problem that microfinance companies in developing countries most especially Ghana are trying to overcome over the years. This is due to the fact that losses that companies incur are largely due to default of these loans and some due to the fact that they are unable to pay the loans on time. In order to grant a loan to a person or a group of individuals it is necessary on the part of the institution to assess the risk of default of payment by the borrower (s). This assessment of the customer is based on the banks' ability to collect data on the customer's credit history such as income level, age, number of years of education and so on as well as the expected profitability of the particular project. The company can then assess these data before loan is granted to a borrower.

Amelie and Allen (2011) argued that financial risk can be divided into credit, market and operational risk but the largest component is credit risk [9]. By developing an accurate credit risk rating system, banks will be able to identify loans that have lower probability of default versus loans that have a higher probability of default. Andrea (2010) stated that poorest people are often considered "unbankable," because they do not have characteristics of traditional borrowers, such as reliable credit histories or high levels of collateral. However, over the past three decays, many microfinance institutions have emerged across the globe and compared to traditional banks, many Microfinance institutions boast high repayment rates from borrowers without formal credit histories (Morduch, 1999). Some of these rates, however, are deceiving (Andrea, 2010). Banks often need to charge large interest rates because small loans can be expensive to service and do not return large profits per loan (Andrea, 2010). Gary and Tang (2001) study the microcredit challenge in California. The authors revealed that most of the Microfinance institutions are not close to reaching any measure of financial sustainability. They attribute part of this problem to excessive operating costs some of which can be three times the size of the loan amounts. These operating costs can include the time a loan officer spends investigating the borrower's background, any paperwork and other administrative tasks.

All lenders do some sort of risk analysis before underwriting a loan. Artur (2008) stated that the two types of risk analysis are quantitative and qualitative. Loan officers perform a qualitative risk analysis when they interview the potential borrower, look over the business plan (if available) and review past financial history. Quantitative risk analyses are more expensive and time consuming, because they require keeping track of loan data both during loan origination and monitoring (Andrea, 2010). Many works have been done on predicting the default rate in both developed and developing countries like Ghana.

Amelie and Allen (2011) assessed the Probability of Default in Agricultural Loans using logistic regression and came out with the probability of default as $p = \frac{\exp(b_0+\sum x_{ij})}{1+\exp(b_0+\sum x_{ij})}$. However, their results showed that leverage, profitability and liquidity at loan origination are statistically significant indicators of the probability of default. According to Dadson (2012) andrea (2010), Amiram (2011), Allen *et al*. (2006) and

Corresponding Author: Charles Kwofie, Department of Statistics, University of Ghana, P.O. Box LG 25, Legon-Accra, Ghana, Tel.: +233 240643442/574432331

Amelie and Allen (2011) the probability of default increases as the length of the loan increases.

Dmytro and Venzhyk (2013) studied the micro level causes of retail loan defaults in Ukraine. They discussed the reasons for loan defaults within car loans and mortgages with some specific variables, such as foreign currency usage and the housing bubble. The authors used logit models and neural networks; however their results failed to confirm directly whether the variables studied affect credit risk. As it turns out, neural networks outperform logistic regression, but not significantly. The authors found factors that affect the probability of a retail loan default and recommended, however, using neural networks because of their better ability to identify potential defaulters.

Oluwarotimi *et al*. (2006) emphasized that credit institutions have made several attempts at modeling and forecasting default rate using approaches like credit scoring, discriminant analysis, conventional econometric models and neural networks, but with great challenge in incorporating human knowledge into these technologies. Existing literature shows that results on the performance of these models remain diverse (Turetken, 2004) leading many to suggest that the superiority of a specific method may be case-specific. This research will make use of two methods in predicting the probability of default, namely; logistic regression and discriminant analysis in order to compare their predictive abilities. A comparative examination of these two estimation methods will be conducted in terms of their predictive accuracies of credit default incidences for loan applicants.

Most of these works made use of the borrower's profitability of business, liquidity, leverage, borrower's equity, borrower's working capital and so on to build models. However, the data to be used in this study will be data collected on the individual before the loan is granted by the microfinance company that is data such as age, average salary and income etc.

Jaime (2008) built a quantitative model that sort to estimate the probability that a US issuer will default on public debt within a year using logistic regression. The author found that all market variables considered were significant in the model. He concluded that logistic model can be used to predict default but, however warned that high correlation exist between the models.

MATERIAL AND METHODS

A secondary data from a sample of 120 clients was obtained using systematic sampling from a micro finance institution in Greater Accra Region of Ghana. Out of those 90 clients were used in building the models and 30 clients to validate the models.

The statistical tool used in the study was binary logistic regression. The choice of the models was a result of the fact that the response variable is a dichotomous variable.

Binary logistic regression: Many social phenomena are qualitative rather than quantitative in nature thus an event occurs or it does not occur, a person makes one choice but not the other, an individual or group passes from one state to another (Pampel, 2007). Binary discrete phenomena usually take the form of a dichotomous variable. The mean of the dichotomous variable equals the proportion of cases with a value of 1 and can be interpreted as a probability.

Logistic regression analyses the relationship between multiple independent variables and a single dichotomous dependent variable. The choice of this model was based on the fact that the desired result "Default Status" has two possible outcomes coded as 0 and 1.The response variable Y is a dichotomous variable with possible values of 0 and 1 thus:

$$Y = \begin{cases} 0 \ Default \\ \\ 1 \ Non\ Default \end{cases}$$

We consider k independent variables. Where (k = 1, 2, 3, 4, 5, 6) and then our prediction equation have the form:

$$E(Y/x_1, x_2, ..., x_6) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_6 x_6)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_6 x_6)}$$

where, they β_0, β_1, β_2, …, β_6 are the estimated logistic regression coefficients. The logistic regression slope will have the usual interpretation, except that it will be in probability terms: for every 1 unit change in a given independent variable there will be a change in probability of being in a category. In this research the categories are default and repay categories. The predicted probability for each case can be derived from the log odds and consequently the residual (the difference between the response for that case and their actual 1 or 0 statuses) can be calculated. Unlike multiple linear regression models, logistic regression does not assume linearity of relationship between dependent and independent variables. Also the error term (ε) is not normally distributed since Y takes only values 0 and 1. In addition, the probability of occurrence of the event Y lies between 0 and 1; that is $0 \leq P(Y) \leq 1$. The logistic regression was used to calculate the probability of success over the probability of failure; the results of the analysis were in the form of an odds ratio and will help in the prediction of group. Moreover, the logistic regression also provided knowledge of the relationships and strengths among the variables.

The goal of a statistical model is to select the most parsimonious variable that still explains the data very well. A univariate logistic regression model was used to

obtain the estimated coefficient, Wald statistics and p-value. Any variable whose univariate test has p-value <0.25 should be considered as likely candidate for the multivariate model. The use of 0.25 as a screening criterion is based on works by Bendel and Afifi (1997), Mickey and Greenland (1989). This is because the use of large p-value has the disadvantage of including variables that are of questionable importance.

By fitting the model, we will be able to estimate the logistic regression coefficients β_0, β_1, β_2, …, β_6 of the variables selected. The coefficient of the logistic regression model is estimated using the maximum likelihood estimation.

After estimating the coefficients, there is the need to assess the significance of the variables in the model. This usually used formulation and testing statistical hypothesis to determine whether the independent variable in the model is significantly related to the response variable. The various tests and measures below help to test for the significance of the models and its parameters.

R^2 is the proportion of the variation in the dependent variables accounted for by the logistic regression model:

$$R^2_{logistic} = \frac{-2LL_N - 2LL_K}{-2LL_N}$$

where, $-2LL$ is the -2log likelihood value? Therefore $-2LL_N$is -2log likelihood value of the logistic regression model with just the constant and $2LL_K$ is the value for the model with all the predictors.

The Hosmer Lemeshow goodness of fit was used in assessing the fit of logistic regression model. The Hosmer Lemeshow statistic was obtained together with expected frequencies. The hypothesis for the Hosmer Lemeshow statistics is:

H_0: There is no difference between the observed and model predicted values

H_1: There is a difference between the observed and model predicted values

Wald statistics which is the ratio of the estimated coefficient to its standard error was used to test the significance of individual logistic regression coefficients for each independent variable. This assisted us in determining the quality of the overall model. The hypothesis of interest was:

H_0: $\beta_i = 0$ i.e., X_i has no significant effect on the log-odds ratio

H_1: $\beta_i \neq 0$ i.e., X_i has a significant effect on the log-odds ratio

The Wald (W) statistics of the β_i coefficient is used as the test statistics where:

$w = \left(\frac{\hat{\beta}_i}{SE\,\hat{\beta}_i}\right) \sim x_1^2$, SE is the standard error of $\hat{\beta}_i$

Since the main objective is to obtain the best model, which minimizes the number of parameters, a test is carried out to find out if the variables that have been eliminated are truly insignificant in the model. To do this we fit a model containing all the variables. The likelihood ratio estimate has a chi-square distribution on j degrees of freedom being the number of variables eliminated from the model. If for example the p-value exceeds $\alpha = 0.05$, then we can conclude that the reduced model contains fewer variable, but yet explains the data very well. Hence we choose the reduced model.

ANALYSIS AND DISCUSSION

A sample of size 120 was obtained of which 90 was used in building the models and 30 to validate the model. Figure 1 shows the distribution of age group of beneficiaries with 41.1% of those who benefitted from the loan are between the ages of 30-39 years, 37.8% are between the ages of 40-49, 11.9% are between the ages of 50-59 while 1.1% are also between the ages of 60-69.This means majority of the people who received the loan are between the ages of 30-39.

Figure 2 shows that out of the 90 beneficiaries of the loan, 34.4% are single while 65.6% are married. This indicates that majority of those who were given the loans were married. From Fig. 3 it can be seen that out of the sample of 90 who received the loans 70% were females while 30% were males. This means that females received more loans than males. Figure 4 gives the percentages of different ranges of base capitals that the loan beneficiaries use in conducting their various businesses. It shows that 75.6% of the beneficiaries operate with a capital base of between GH¢ 0-1999, 18.9% operate with base capital of between GH¢2000-3999, 3.3% operate with base capital of between GH¢ 10000-30000 while 1.1% also operate with base capital of between GH¢4000-5999 and GH¢8000-9999. This shows that majority of the loan beneficiaries operate with base capital of between GH¢ 0-1999. Figure 5 gives an indication that 42.2% of the loan beneficiaries had 10-13 years of education, 26.7% had 7-9 years of education, 25.6% had 0-6years of education while 5.6% had 14-17 years of education. This means that those who had 10-13 years of education received most loans than the rest of the years. Figure 6 shows that those who have been in business for few numbers of years (0-9) were given most of loans than those who had been in business for long (i.e. 10-19 and 20-29).

Table 1 presents the results with only the constant included, before any coefficients are entered into the equation.

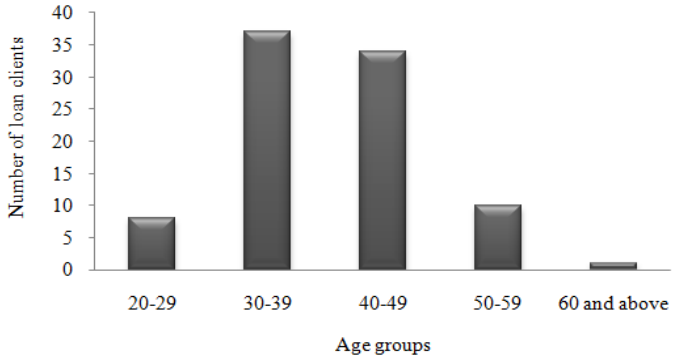

Fig. 1: Bar graph showing the distribution of age of loan clients

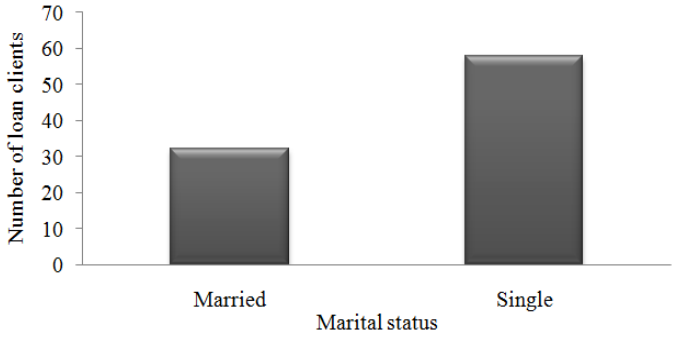

Fig. 2: Bar graph showing the Marital Status of loan clients

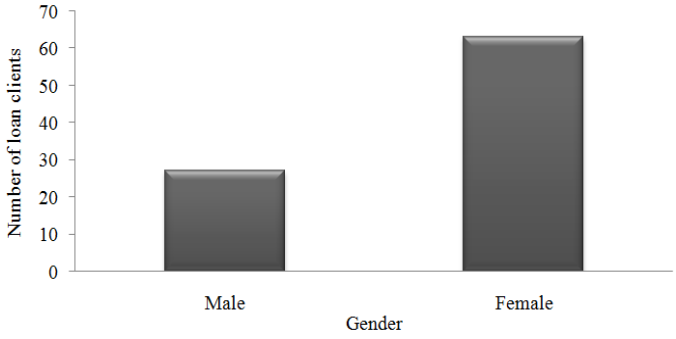

Fig. 3: Bar graph showing the distribution of gender

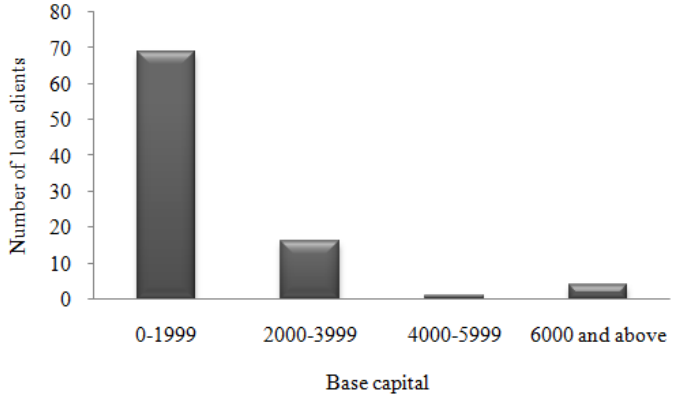

Fig. 4: Bar graph showing the distribution of base capital

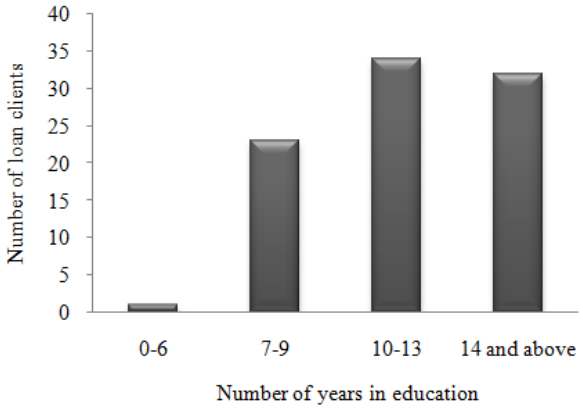

Fig. 5: Number of years in education

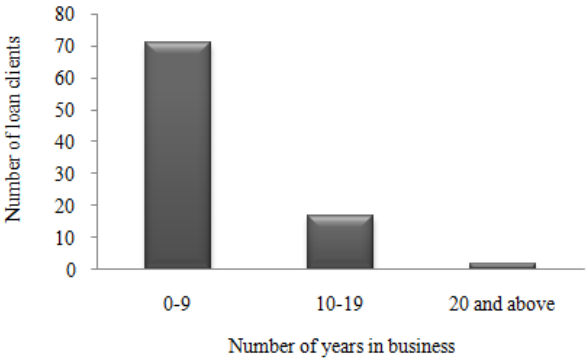

Fig. 6: Bar graph showing the distribution of number of years in business

Table 1: Constant fit in the equation

	B	S.E.	Wald	Degree of freedom	Significant	Exp (B)
Constant	-0.314	0.213	2.16	1	0.142	0.731

Table 2: Model summary

-2 Log likelihood	Cox and Snell R^2	Nagelkerke R^2
116.234[a]	0.068	0.092

Table 3: Omnibus test of model coefficients

	Chi-square	Degree of freedom	Significant
Step	6.346	1	0.012
Block	6.346	1	0.012
Model	6.346	1	0.012

Table 4: Hosmer and lemeshow test

Chi-square	Degree of freedom	Significant
9.247	8	0.322

The overall significance of the model is tested using what SPSS calls the *Model Chi square*, which is derived from the likelihood of observing the actual data under the assumption that the model that has been fitted is accurate. There are two hypotheses to test in relation to the overall fit of the model:

$\boldsymbol{H_0}$**:** The model is a good model. (i.e., model without any predictor is appropriate)

$\boldsymbol{H_1}$**:** The model is not a good model (i.e., the predictors have a significant effect)

The difference between -2 Log Likelihood (–2*LL*) values for models with successive terms added also has a chi squared distribution, so when we use a stepwise procedure, we can use chi-squared tests to find out if adding one or more extra predictors significantly improves the fit of our model. The –2LL value from Table 2 is 116.324 which is distributed as x_1^2 with a significant probability of p = 0.012<0.05 (i.e. we reject the null hypothesis). Thus, the indication is that the model has a poor fit. The model containing only the constant also implies that the predictors do have a significant effect on the response variable and create essentially a different model. So we need to look closely at the predictors and determine which of them will be significant in the model.

From Table 2 although there is no close analogous statistic in logistic regression to the coefficient of determination R^2, Table 2 provides some approximations. *Cox and Snell's R-Square* attempts to imitate multiple R-Square based on 'likelihood', but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. The Cox and Snell's R-square obtained is 0.068 indicating that 6.68% of the variation

Table 5: Variables in the equation

	B	S.E.	Wald	Degree of freedom	Significant	Exp(B)
Marital status	0.375	0.593	2.757	1	0.042	1.455
Base capital	0	0	1.459	1	0.027	1
Number of years in business	-0.023	0.017	2.778	1	0.036	0.977
Constant	0.043	1.231	0.082	1	0.045	1.044

Table 6: Number of correctly predicted default status out of 10 out-samples

	Number of years in business (NYE)			Marital status	
	High NYE	Moderate NYE	Low NYE	Married	Single
Logistic regression	4	3	5	2	6

in the dependent variables is explained by the logistic model. The Nagelkerke modification that does range from 0 to 1 is a more reliable measure of the relationship. Nagelkerke's R^2 will normally be higher than the Cox and Snell measure. Nagelkerke's R^2 is part of SPSS output in the 'Model Summary' table and is the most-reported of the R-squared estimates. In our case it is 0.092, indicating a very weak relationship of 9.2% between the predictors and the response variable.

From Table 3 an alternative to model Chi-square is the Hosmer and Lemeshow test which divides subjects into 10 ordered groups of subjects and then compares the number actually in each group (observed) to the number predicted by the logistic regression model. The 10 ordered groups are created based on their estimated probability; those with estimated probability below 0.1 form one group and so on, up to those with probability 0.9 to 1.0. Each of these categories is further divided into two groups based on the actual observed outcome variable (success, failure). The expected frequencies for each of the cells are obtained from the model. A probability (p) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model.

$\boldsymbol{H_0}$**:** There is no difference between the observed and predicted values

$\boldsymbol{H_1}$**:** There is a difference between the observed and predicted values

Since the p-value (0.05) for the Hosmer-Lemeshow goodness-of-fit test statistic is less than the significance value of 0.322, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level.

Table 4 shows the variables that are significant in the final model. The variables left after the forward likelihood selection are base capital, marital status and number of years in business. The table has several important elements. The Wald statistic and associated probabilities provide an index of the significance of the predictor in the equation. The Wald statistics has a chi-square distribution. The simplest way to assess Wald is to take the significance values and if it is less than 0.05 reject the null hypothesis and accept the alternative hypothesis. Where the null and the alternate hypothesis are:

$\boldsymbol{H_0}$**:** $\beta_i = 0$ i.e., X_i makes no significant contribution in the model

$\boldsymbol{H_0}$**:** $\beta_i \neq 0$ i.e., X_i makes a significant contribution in the model

The EXP(B) is the exponential of the logistic coefficients. The EXP(B) column in Table 5 presents the extent to which raising the corresponding measure by one unit influences the odds ratio. If the value of EXP(B) exceeds 1 then the odds of an outcome occurring increases while if the Figure is less than 1 then any increase in the predictor leads to a drop in the odds of the outcome occurring. For example, the EXP(B) value associated with base capital is 1.000. Hence when base capital is raised by one unit (GH¢1) the odds ratio is 1 time as large and therefore loan applicants are 1 time likely to belong to the *repay* group. The 'B' values are the logistic coefficients that can be used to create a predictive equation (similar to the *beta* values in linear regression). Hence the predictive equation is:

$$Probability\ of\ a\ case = \frac{\exp\{(0.375*MS)+(-0.023*NYB)+0.0430\}}{1+\exp\{(0.375*MS)+(-0.023*NYB)+0.0430\}}$$

where,

MS = Marital status

NYB = Number of years in business

CONCLUSION

A logistic regression analysis was conducted to predict default status of loan beneficiaries for 90 sampled beneficiaries using, age, marital status, gender, and number of years of education, number of years in business and base capital as predictors. The predictors that were significant in the model were marital status, number of years in business and base capital. The explained variability in the response variable in both the logistic regression was very weak.

The Cox and Snell's R-Square and Nagelkerke's R^2 of the logistic regression were 6.68% and 9.92% respectively, both indicating a very weak relationship between prediction and predictors.

From Table 6 the logistic regression model predicted well among beneficiaries with low number of years in business and predicted poorly among the beneficiaries with high and moderate number of years. Logistic regression also predicted well among single beneficiaries while predicting poorly for married beneficiaries. Generally, the logistic regression predicted 40.0% default status correctly.

REFERENCES

Allen, M.F., M.R.Laura and J.B.Peter, 2006. Determining the probability of default and risk-rating class for loans in the seventh farm credit district portfolio. Rev. Agr. Econ., 28(1): 4-23.

Amelie, J.and M.F.Allen, 2011. Determining the Probability of Default of Agricultural Loans in a French Bank. Proceeding of the American AgriculturalEconomics Association Annual Meeting.International Scientific Press, Long Beach, California,pp: 1799-6599.

Amiram, D., 2011. Debt contracts and loss given default. Job Market Paper,Working Paper, Columbia University, pp: 2-54.

Andrea, R.C., 2010. Measuring the likelihood of small business loan default: Community Development Financial Institutions (CDFIs) and the use of credit-scoring to minimize default risk. Honours Thesis, Department of Distinction in Economics, Duke University, Durham, North Carolina, pp: 1-74.

Artur, R., 2008. IT risk assessment: Quantitative and qualitative approach. Proceedings of the World Congress on Engineering and Computer Science (WCECS'08), San Francisco, USA.

Dadson, A.V., 2012. Determinants of loan repayment default among farmers in Ghana. J. Dev. Agric. Econ., 4(13): 339-345.

Dmytro, G.and K.Venzhyk, 2013. Loan default prediction in Ukrainian retail banking. Procceding of the Economic Education and Research Consortium,pp: 1-39, ISSN 1561-2422.

Gary, P.andS.Y.Tang, 2001. The microcredit challenge: A survey of programs in California. J. Dev. Entrep., 6(1): 1-16.

Jaime, F., 2008. Credit Risk Modeling: Default Probabilities. pp: 1-34.http://stat.fsu.edu/~jfrade/HOMEWORKS/STA5168/FRADE_STA5168_Project.pdf.

Morduch, J., 1999. The microfinance promise. J. Econ. Lit., 37(4): 1569-1614.

Oluwarotimi, O.O., M.F.Allen and S.Das, 2006. Predicting credit default in an agricultural bank: Methods and issues. Proceeding of the Southern Agricultural Economics Association Annual Meetings.Southern Agricultural Economics Association,Orlando, Florida, pp: 1-21.

Pampel, F.C., 2007. Quantitative Applications in the Social Sciences. SAGE,Boulder.

Turetken, O.,2004. Predicting Financial Performance of Publicly Traded Turkish Firms: A Comparative Study. Retrieved from:mis.temple.edu: http://mis.temple.edu/research/Documents/TuretkenOct2004-NNPrediction.pdf.(Accessed on: December 4, 2014.