International Journal of Economics & Strategic Management of Business Process (ESMB)
Vol. 6, 2016 - pp. 1-6

SUPPORT VECTOR MACHINE ESTIMATION OF DROPOUT RISK

Fathi Abid[#1], Moncefhabibi[*2]
[#]*Department of finance*
University of Sfax, Tunisia
[1]fathi.abid@fsegs.rnu.tn
[*]*Department of business administration*
Bizerte High Institute of Technological Studies, Tunisia
[2]moncef.habibi@fsegs.rnu.tn

***Abstract*—Theaim of this paper is to compare non parametric and parametric methods to estimate the dropout risk in a microfinance institution. We tune the kernel function of a support vector machine classifier (SVM) and compare its performance to the standard logistic regression using the confusion matrix and the non-parametric McNemar test. Extensive quantitative analysis is applied to a randomly stratified sample of 100 customers drawn from an NGO microlender in Tunisiashow that the support vector machine classifier with a radial basis Kernel outperform linear kernel SVM and logistic regression in forecasting dropout risk.**

***Keywords*—Dropout, Support vector machine, logistic regression, Microfinance**

I. INTRODUCTION

By the end of 2011 the interim government in Tunisia has set up a new regulatory framework[1] that aims to foster microfinance investment for both domestic and foreign operators. Previously, the market for microfinance is solely monopolized by two operators with the tightly scope of providing microloans for microentrepreneurs and low-income households: a governmental bank, the BTS, and a well-established NGO, ENDA Inter Arab. In March 2013, more than twenty agreement applications for approval have been submitted within the ministry of finance for the exercise of a broader microfinance activity with different legal forms. By the virtue of this new juridical microfinance frame, existing microfinance institutions (MFI), will bear a new major risk on how to deal with client dropout.

Dropout (attrition, desertion or defection) occurs whenever a customer does not renew a loan after repaying an old one. Dropout implies a cut in market share and an increase in operating costs. Its drain on profitability and long term viability has been well documented in previous research [2], [10], [23], [21], [1]. Desertion is also more pronounced in competitive markets and for maturate MFIs where new client entrance might be offset by large client exit. Nonetheless, many MFIs still address the issue of customer desertion on a reactive basis and only few of them have designed business strategies to build customer loyalty [2], [3]. Early discovery of customers who are at risk of attrition is the basis of the start point of any customer retention strategy.

In this paper we propose a linear and a non parametric dropout risk estimation methods. Thelinear model is the standard logistic regression and thenon parametric method is the support vector machine (SVM) algorithm. The different classifiers allowthe forecastingof prior dropout probabilities. A fundamental contribution of the proposed approaches is that they proposes a quantitative assessment of dropout risk in terms of probabilities that can be used to monitor client attrition. Moreover, they can serve as a scoring device to rank clients on the basis of their risk of leaving.

The layout of the paper is as follows: section two reviews the potential variables that can influence customer's decision to stay or retrieve from a microfinance program. Section three exposesthe methodology with an emphasis on the statistical foundations of the different classification schemas. Section four describes the data and setting. Section five exhibits the main findings. Section six concludes.

II. FACTORS INFLUENCING CUSTOMER DROPOUT

Previous research on client desertion worldwide has reported that dropout varies with the socio-economic context. Pagura[15] provides a dropout rate in the range on 10-20% in BRAC, ASA, Grameen, and WWB affiliate in Bangladesh. Tedeschi and Karlan[19] reported a dropout rate among clients of a microcredit organization in Peru,

[1]Ministry of economic and finances: Decree number 2011-117.

Mibanco, of 56% over a two-year period and they claim that this dropout rate is quite normal. Wright [23] finds that the rate of dropout ranges from 25% to 60% in East Africa. Hulme[9] analyzes annual dropout rate from thirteen MFIs in Kenya, Tanzania, and Uganda. He finds an average annual dropout rate of 36.5%. Falco and Leher [5] find a dropout of 46% in 2 years (between 2008 and 2009) from the ADRA microfinance program in Ghana. To conclude, higher dropout is observed for African MFIs [15].

In what follows we draw upon the previous researches to give theoretical background on the factors leading client to exit a given microfinance program.Basically, seven factors seem to influence client decision to exit a microfinance program: Gender, Poverty level, age, loan term and amount, interest rate and satisfaction with respect to staff attitude.

A. Gender

Shreiner[17] examined the case of a microlender in Bolivia and showed that women are more likely to exit than men. A 2001 survey conducted by the United Nations' Special Unit for Microfinance (SUM) for a sample of 29worldwide MFIs, showed that women are more loyal to their micro finance program than men [4]. However, the evidence is not clear for Hulme et al. [10] and Hulme[9] who find no correlation between gender and dropout in Kenya and east Africa, respectively.

B. Poverty level

Hulme[9] analyzes dropouts from thirteen MFIs in Kenya, Tanzania, and Uganda, and the reasons of dropout. He finds that socio-economic status is a significant factor that explain client dropout. Furthermore, he concludes that poorer clients dropout if the average loan size within their group rises, requiring them to guarantee larger loans than they can take themselves. By contrast, wealthier dropouts complain that the available loan is too small given the system of weekly meetings which demand a significant amount of their time.

Typically, healthier clients require microfinance loans only for temporary financial shock such as consumption, education and housing. As long as they cope with their financial needs they stop doing business with the MFI.

In a study of dropout in Ughanda, Wright et al. [22] noticed that poorer leave or are pushed out from MFIs primarily because they find problems in repaying their loans.

C. Age

Client age can exert an effect on dropout. Based on a client survey applied to financial service industry, PriceMetrix[16] reported a negative relationship between client age and exit probability. However, the study does not report whether this causality is significant or not. The same conclusions have been derived by Hulme [9] in analyzing attrition from a sample of 30 MFIs in Kenya, Tanzania and Ughanda.

D. Loan term

Credit policy, through loan term, loan size and interest rate can impact to a large extend customer retention rate. Musona and Coetzee [14] used focus group methodology to depict the reasons behind observed high dropout in Zambian MFIs. They concluded that the repayment schedule was perceived as too rigid and, therefore, not adequately taking into account the realities of micro businesses. In a comparative study between two South African MFIs, Stark and Nyirumuringa[18] assessed that rigidity in loan term is a main reason of customers' dropout. Maximabali et al. [12] reported identical results for Tanzanian MFIs. Pagura[15] examined the case of Piyeli, a Malian microfinance NGO, and asked clients to rank 13 exit reasons according to their importance. He found that credit term is secondly ranked.

E. Loan size

Borrower who does not find the exact loan size he requires might be more likely to exit. The evidence is clear in a competitive market structure. These statements have been empirically validates in [17] for a Bolivian microlender and in [18] for two South African MFIs. Hulme et al. [10] pointed out that a small loan size leads wealthier clients to dropout. The opposite holds insofar that when the loan size is increased, poorer clients voluntarily dropout. Musona and Coetzee [14] highlighted the importance of increasing the loan size to reduce client exit.

As a consistent measure of the effect of loan size on the borrower decision to exit we propose the difference between the amounts of loan disbursed and required.

F. *Loan fees*

Many clients voluntarily withdrew from MFIs due to the loan fees. Maximabali et al. [12] pointed out that clients complained that the interest being charged is too high, particularly when taken together with other costs e.g. application fees, disbursement fees, etc…Pagura[15] reported similar results for a Malian microfinance NGO, however, interest fees seems to be not perceived by clients. In financial theory, interest rate fees should rise as MFI's experience high credit risk and operation

expenditures. A major challenge facing is how to reduce these fees given the risk level to which they are exposed.

As a direct measure of the loan fees we consider interest rate charged to client.

G. *Satisfaction with respect to staff attitude*

The relationship between client satisfaction and dropout is straightforward. High dropout rates are the result of client dissatisfaction with respect to loan officer attitude and with the services and products being offered [18], [17], [10], [23]. Today, the microfinance industry is becoming more client or market driven. Customer relationship is the result of a sustainable improved quality of service and a clear client driven management strategy. A satisfied customer creates a strong business relationship with the firm which translates into loyalty and invariably retention.

Given the huge number of financial products offered by the microlender case study, we examine client satisfaction only with respect to loan officer attitude. Loan officer can substantially reduce dropout likelihood by sustaining good relationship with customers.

III. METHODOLOGY

Let X a set of *n* possible examples (covariates, attributes or features): $X = \{x_1, \dots, x_i, \dots, x_n\}$.

Each example x_i is a $(d \times 1)$vector of the customer's attributes (gender, age, educational level…). Let Y the set of all possible output values. Based on several measurements of attributes or features, we want to classify customers into one of the two categories: to exit (dropout) or renew a microfinance loan. Then Y is a dichotomous variable and takes a value of 1 in the first case and -1 otherwise. This refers to a problem of two class pattern classification or a binary classification problem.

A. *Logistic regression classification*

The first suitable model for binary classification is the standard logistic regression. In this model the dropout conditional probability $\pi(X) = Pr(\mathrm{Y} = 1|\mathrm{X})$ is given by

$$\pi(X) = \frac{e^{\mathrm{x}^{\mathrm{T}}\mathrm{w}}}{1+e^{\mathrm{x}^{\mathrm{T}}\mathrm{w}}} \qquad (1)$$

A logit transformation of $\pi(X)$is :

$$Z = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) \quad (2)$$

The functional form relating the logit Z with the set of covariates X is therefore:

$$Z = \mathrm{x}^{\mathrm{T}}\mathrm{w} + \varepsilon \qquad (3)$$

withw stands for the slope coefficient and measure the change in the logit per unit change of the covariate.

$\varepsilon$is the error term which is standard Gaussian.

Maximum likelihood method is used to estimate equation (3) [7].

B. *Support vector machine classification*

The SVM model is a new statistical technique for binary classification [20]. It is based on linear classifier that simultaneously maximizes the margin or the distance between the classes and minimizes empirical risk related to misclassification. Recent applications of the SVM algorithm in financial time series forecasting [6], credit scoring [8] and bankruptcy prediction [11] have reported high prediction and classification accuracy compared to linear models and other machine learning algorithms.

Support vector machine have two main advantages over simple logistic regression: (1) It is robust to very large number of variables and small samples, and (2) It can learn both simple and highly complex classification models. The basic idea is to find a linear decision surface called hyperplane in between data sets to indicate which class it belongs to. This is achieved by training the machine to understand structure from data and mapping with the right class label, for the best result. The best hyperplane for an SVM means the one with the largest margin between the two classes. Fig. 1 shows such a hyperplanethat separate two classes to the boundary.

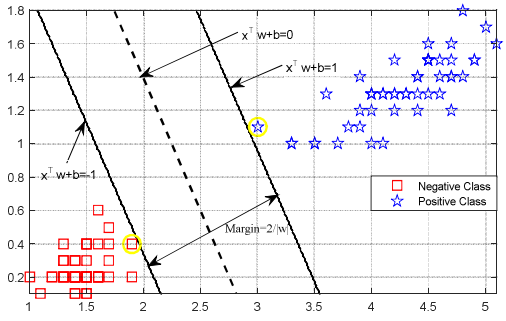

Fig. 1. Separating hyperplanefortwoseparableclasses.

If such linear decision surface does not exist, the data is mapped into a much higher dimensional space (feature space) where the separating decision surface is found. The feature space is constructed

via simple mathematical projection using different kernels.

The separating function generated by a linear SVM is given by:

$$x^{\intercal}w + b = 0 \quad (4)$$

The scalar b is a location parameter,w is a $(n \times 1)$ vector of weights that determines the slope of the separating function.

For the linear separable case, the following constraint must hold:

$$\begin{cases} x^{\intercal}w + b \geq 1 \text{ for } y_i = 1 \\ x^{\intercal}w + b \leq -1 \text{ for } y_i = -1 \end{cases} \quad (5)$$

for $i = \{1, 2, \dots, N\}$.

In the case where the data is not linearly separable, *x* should be replaced by$\varphi(x)$, with$\varphi(.)$ is the mapping from the input space $\mathbb{R}^n$ to a higher dimensional feature space $\mathcal{H}$:

$$\varphi(.): \mathbb{R}^n \longrightarrow \mathcal{H}$$

The SVM optimization problem isformulated as follows:

$$\underset{\alpha_i}{\text{Min}}\, Z = -\sum_{i,j=1}^{N} y_i y_j < \varphi(x_i)\varphi(x_j) > \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i$$

subject to (6)

$$\begin{cases} \sum_{i=1}^{N} \alpha_i \alpha_j = 0 \\ 0 \leq \alpha_i \leq C \end{cases}$$

$\alpha_i$are Lagrange multipliers and are referred to the support values and give the relative weight of their corresponding support vector x_i.C is a penalty factor that rules the tradeoff between high complexity of decision rule (high C) and low frequency of error (low C).The kernel function is defined as:

$$K(x_i, x_j) = \varphi(x_i)\varphi(x_j) \quad (7)$$

In order to capture the implicit patterns hidden in the data set we use two different kernels: The linear kernel (LSVM) and the Gaussian radial basis kernel (GSVM)[2]. The diverse kernels will produce different errors:

(a) Linear kernel function: $K(x_i, x_j) = x^{\intercal}x$

(b) Gaussian RBF kernel function: $K(x_i, x_j) = exp\left(\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$ with$\sigma \in \mathbb{R}^+$

For practical considerations we choose, the confusion matrix as the performance criteria. Table 1 illustrates this matrix:

TABLE 1. CONFUSION MATRIX

		Predicted Class	
		1	-1
True Class	1	True positive (TP)	False positive (FP)
	-1	False negative (FN)	True negative (TN)

The average accuracy ratio of a given classifier is defines as:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

IV. SETTING AND DATA

A. *Microfinance in Tunisia*

The market for microfinance in Tunisia goes back for more than three decades. The first microlending experience has been carried out by the NGO Enda Inter Arabe. Enda is an international NGO and a member of the ENDA third world family based in Dakar. Enda is created in 1990 with the primary mission of poverty alleviation through credit and micro-enterprise support. Enda is, exclusively, the largest MFI in Tunisia with a network of more than 70 branches that serves nearly 230,000 micro-entrepreneurs and an estimated portfolio value of about 90 million USD.

In 1999, policymakers recognized micro-lending activity by establishing the so called Tunisian Bank of Solidarity (BTS). The BTS provides highly subsidized credit either directly or indirectly through more than 280 local associations. By facilitating access to credits, especially for microentrepreneurs, the BTS contributes to the whole national objective of supporting income generation and unemployment reduction.

Due to restrictive regulations, the microfinance industry structure has remained unchanged with the limited scope to microlending activity. It is until 2011 when the interim government has established a road map to promote microfinance operations through a plethora of juridical measures. The cornerstone of these regulations is to create MFIs with a broad range of microfinance services such as microinsurance and microsaving. These institutions can take different legal forms and are supervised by the ministry of finance under different governance and control standards.

The new regulation has attracted new domestic and foreign operators in search for financial profit in an underexplored market. In March 2013, the Tunisian finance minister reported that more than 20 agreement applications for approval are submitted for the exercise of the microfinance activity. Many operators have already started operations such as Taysir microfinance, AdvansTunisie and Microcred.

For existing MFIs, the new juridical order gives the opportunity to extend operations to new financial products such as microinsurance and to improve the financial viability by giving access to

[2] Other kernels are tested such as the polynomial and the sigmoid kernels andseems to produce low accuracies.

new funding sources such as microsaving. However, they should undertake a full revision of the risk management process, by firstly reclassifying the risks they bear. For instance, desertion risk which has been neglected for long time before should move to the high concern of risk management as like as any other viability linked risk metrics.

B. Data

Data are drawn from Enda Inter Arabe. Enda is interested on dropout to design proactive measures on how to deal with desertion risk. The data sampling method consist of two steps. First, we construct a random sample of 100 clients and stratify them into two equal groups based on the decision to exit or renew a loan. Dataare collected coherently by loan officers from Enda branches and from Enda risk department. Second, to get more in-depth insights about client's attitudes towards loan officers, we rely on direct interview with clients. Accordingly, an independent research team is sent to each applicant to insure the credibility of the latter's responses. A questionnaire is then administrated to the 100 respondents. The questionnaire is developed using insights from the literature.

In order to avoid possible distortions in the results, we choose only clients that have definitely and voluntarily leaved a given microcredit program. To reach the target sample, we first screen customers based on their spells of arrears. This allows us to distinguish between good clients and bad clients. Good clients are those that had proven a zero spell arrears whereas bad clients are those for them spell arrears are at least thirty days. Respondents are then asked explicitly if they had left the lender, either temporarily or permanently.

The set of covariatesand their corresponding measures are given in table 2:

TABLE 2. COVARIATE DEFINITION AND MEASUREMENT

Covariates	Definition and measurement
Gender (GDR)	Dummy, takes 1 if the customer is male and 0 otherwise
Poverty level (INC)	Or income group. Ordinal and ranges from 1 (less poor) to 5 (more poor)
Age (AG)	Age of the costumer. Measured in years
Loan term (LT)	Length of the loan. Measured in months
Loan size (LS)	Measured by the difference between the disbursed loan amount minus the required loan amount
Interest rate (INT)	A proxy for the loan cost
Satisfaction (SAT)	Dummy, takes one if costumer is satisfied and 0 otherwise

V. ESTIMATION RESULTS

Figure 2 compares the dropout probability estimated from the different classifiers with the true exit probability. A first striking result is that the GSVM probability estimate seems to superpose on the true dropout probability. The complex structure in the data seems to be captured by the Gaussian transformation of the initial input data.

The LSVM and logistic classification models produce high estimation errors compared to the GSVM. Furthermore they display similar pattern.

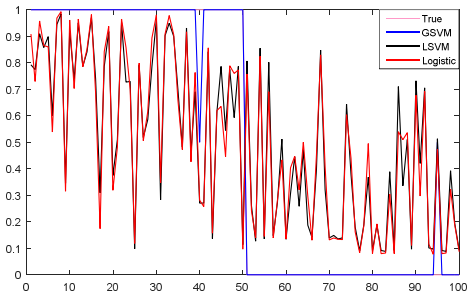

Fig. 2. Dropout probability estimation.

Tables 3 to 5 report the confusion matrix of the different classifiers.

TABLE 3. CONFUSION MATRIX FOR LOGISTIC REGRESSION

		Predicted Class		Accuracy
		1	-1	
True Class	1	**40**	**10**	80%
	-1	**13**	**37**	74%

Note on table: C(i,j) is a count of observations known to be in group i but predicted to be in group j

TABLE 4. CONFUSION MATRIX FOR LINEAR SVM

		Predicted Class		Accuracy
		1	-1	
True Class	1	**39**	**11**	78%
	-1	**11**	**39**	78%

Note on table: C(i,j) is a count of observations known to be in group i but predicted to be in group j.

TABLE 5. CONFUSION MATRIX FOR GAUSSIAN SVM

		Predicted Class		Accuracy
		1	-1	
True Class	1	**50**	**0**	100%
	-1	**1**	**49**	98%

Note on table: C(i,j) is a count of observations known to be in group i but predicted to be in group j.

The average accuracy for the logistic, LSVM and GSVM are, respectively, 77%, 78% and 99%.As expected, the Gaussian SVM produces100% accuracy in forecasting exit customers.Only one example from the non-dropout group is misclassified. This estimation error can be neglected if the manager's focus is on the exit risk.

In what follows we assess the significance between the different classifiers using the non-parametricMcNemar test. The null hypothesis is that the predicted class labels resulting from two separate classifiers have equal average accuracy for

predicting the true class labels Y. The alternative hypothesis is that the labels have unequal accuracy.

TABLE 6. P-VALUE FOR THE MCNEMAR TEST

	Logistic/LSVM	GSVM/Logistic
P-value	0.6875	0

The p-value for the difference in accuracy between the linear SVM and logistic regression classifiers is 0.6875 which is highly superior to the 0.05 cutoff. However the Gaussian SVM seems to outperform the logistic regression classifier. The difference in performance is significant with p-value equal to 1.

VI. CONCLUSION

In this paper we compared parametric and non-parametric methods to estimate the dropout risk for a microlender. The non-parametric estimation uses the support vector machine method with linear and Gaussian radial basis kernel. Empirical results show that the Gaussian kernel produce a significant high accuracy compared to standard logistic regression. As a policy rule, the model can be used as a scoring device to classifycostumers based on their likelihood of exit.

REFERENCES

[1] M.Brand, and J. Gerschick, "Maximizing Efficiency: The Path to Enhanced Outreach and Sustainability,"Accion International, Washington, D.C., 2000.

[2] C.Churchill, "Banking on Customer Loyalty,"*the journal of microfinance*, 2, pp. 1-21, 2000

[3] F. E. Credle, "Want to increase Profits? Stop customer defections,"*Quality Progress*, 1995.

[4] R.Deshpanda, "Increasing Accessand Benefits for Women: Practices and Innovations among Microfinance Institutions: Servey Results", New York, UNCDF, 2001.

[5] P.Falco, and K.Leher,"Understanding Microcredit Membership and Retention: Evidence from Ghana," Working Paper, 2011.

[6] E. H. T. Francis and C. Lijuan, " Application of support vector machines in financial time series forecasting," *Omega: The International Journal of Management Science*, Vol 29, Issue 4, pp. 309-317, 2001.

[7]D. Hosmer, S.Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression*, 3rd ed. 2013, John Wiley & Sons.

[8] C. L. Huang, M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications*, Vol 33, Issue 4, pp. 847–856, 2007.

[9] D.Hulme,"Client Drop-outs from East African Microfinance Institutions,"MicroSave Research Paper, 1999.

[10] D.Hulme, J. Kashangaki, and H. Mugwanga, "Dropouts amongst Kenyan Microfinance Institutions,"MicroSave Research Paper, 1999.

[11] H.K. Kim and S. Y. Sohn, "Support Vector Machines for Default Prediction of SMEs based on technology credit," *European Journal of Operational Research, 2010*, Issue 201, pp. 838-846.

[12] F.Maximambali, C. Lwoga, and S. Rutherford, "Client Exits (Drop-outs) Amongst Tanzanian Microfinance Institutions,"MicroSave Research Paper, 1999.

[13] Ministère de l'économie et des finances,"Décret-loi n° 2011-117 du 5 novembre 2011 Portant Organisation de l'Activité des Institutions de Microfinance", 2011.

[14] D.Musona, and G. Coetzee, "Drop-outs Among Selected Zambian Microfinance Institutions: Causes And Potential Impact On Product Design,"MicroSave Research Paper, 2001.

[15] M.Pagura, "Examining Client Exit in Microfinance : Theoretical and Empirical perspectives," PhD Dissertation, The Ohio State University, 2003.

[16] Pricemetrix, "Putting Some Numbers Behind Client Retention", 2013, available at:http://www.sifma.org.

[17] M.Schreiner,"Scoring Drop-out at a Microlender in Bolivia,"*Saving and* Development, Vol 27, Issue 2, pp. 101-118, 2003.

[18] E.Stark, and P.Nyirumuringa,"Dropouts in Northern Province, South Africa,"Microsave working paper, 2002.

[19] G.A., Tedeschi, and D., Karlan,"Cross Sectional Impact Analysis: Bias from Dropouts," *Perspectives on Global Development and Technology*, Volume 9, Issue 3, pp. 270-291, 2010.

[20] V.N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.

[21] K. E.Wilson, "Exodos:Why Customers Leave,"*The Microbanking Bulletin*, Issue 6, pp. 17-19, 2001.

[22] D.Wright, Mutesasira, L.Sempangi, H.Hulme, D.andS. Rutherford, , "Drop-outs Amongst Ugandan Microfinance Institutions,"Microsave working paper, 1998.

[23] G. Wright, "Drop-outs and Graduates – Lessons from Bangladesh,"*The Microbanking Bulletin*: Focus on productivity, 2001, Issue 6, pp.14-16.