International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

# A Method for System Runtime Improvement of Credit Classification for E-Banking Customers

Seyed Shahin Ghaffari,
Supervisor: Ramin Nasiri
Advisor: Reza Ravanmehr

*Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran Iran*

Date of Submission: 15-11-2023 Date of Acceptance: 25-11-2023

**ABSTRACT:** The reasonable speed of runtime performance of the credit classification systems is one of the goals of any credit institutions and banks**.** Based on this, this issue has been addressed in the present research. So far, many investigations have been done in this field, but no one has been done on the execution time of these models. In the current research, credit classification model (Spark Credit Classification) is proposed including the ability to increase the speed of credit classification while maintaining the accuracy. In this model, a spark cluster has been developed with the ability of two ensemble models. One is a random forest with a random selection of variables, and a penalized logistic regression to increase the speed, and accuracy of the ensemble model, and the other is a decision tree with a random selection of variables, and a logistic regression (SLF) to increase the accuracy of the ensemble model. Based on the evaluations, execution time of this model is 0.085s, and according to the confusion matrix the amount of true positive 269, false negative 38, true negative 333, and false positive 50 are all promising, where on the other hand the accuracy 87.24%, area under the ROC curve (AUC) 93%, true positive rate 87.62%, false positive rate 13.05%, and according to the precision-recall, precision 84.32%, recall 87.62% are acceptable too. The results show that the proposed model is superior to the PLTR in terms of speed and accuracy.

**Keywords**: Runtime- Spark cluster- Random forest- Bagging trees - Logistic regression (SLF)

## I. INTRODUCTION

The credit classification of the loan applicant customers by the model PLTR is considered as one of the most accurate, and interpretable models. This model consists of a decision tree, and a penalized logistic regression, so after the initial classification of the loan applicant customers with the decision tree model, the concluded output of the decision tree is to be entered into the penalized logistic regression model to make a more accurate classification [1]. In this article, first we review the spark cluster model [2-6], so by this model we can increase the speed of the ensembled model, and then keep on the journey by reviewing the random forest model, so by this model we may be able to increase the speed of the ensembled model of the spark cluster ensemble model, then the k-fold cross-validation method [7, 8], so by this method we can do random selection of variables that later on we can increase the accuracy of the random forest model [8, 9], the penalized logistic regression model [1], and the decision tree model [1], later on a bagging trees model that with this model we can implement the decision tree model for random selection of variables to increase the accuracy of the decision tree model, and finally we end up by the review the logistic regression (SLF) model [10]. By considering and scrutinizing the aforementioned models, we proposed the ensemble model of spark cluster with two ensemble models embedded in its branches, one of them is the random forest with a random selection of variables, and the penalized logistic regression [1], and the other one is the decision tree [1] with a random selection of variables, and the logistic regression (SLF) [10].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

## II. SPARK CREDIT CLASSIFICATION MODEL

### 2.1. Description of the methodology

The (Spark Credit Classification) model can speed up the PLTR [1] while preserving the accuracy of the model. This model is formed from one spark cluster that has two ensemble models embedded in its branches. One of them is a random forest with the ability to a random selection of variables, and the penalized logistic regression [1], and the other one is the decision tree [1] with the ability to a random selection of variables, and the logistic regression saturated load forecasting [10].

#### 2.1.1. Spark Cluster

**Our proposed model is shown in Fig.1**

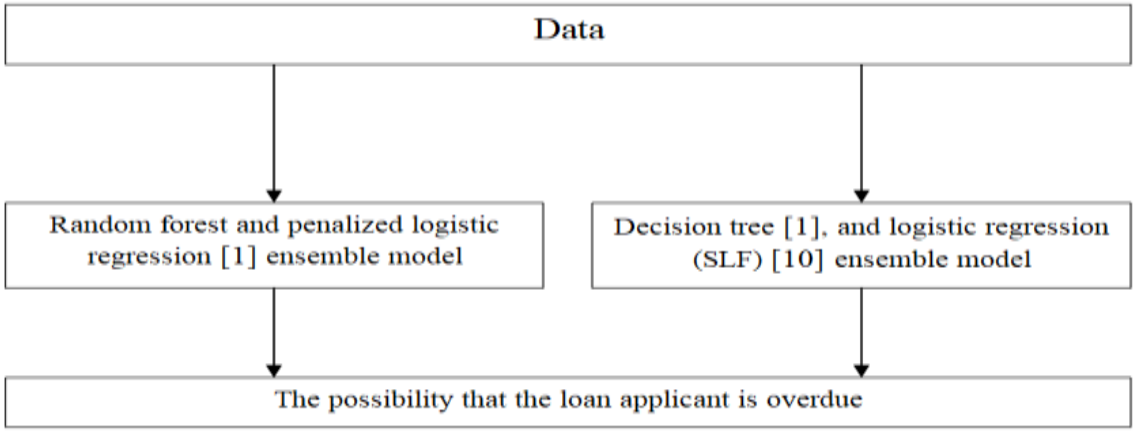


**Fig.1. Proposed spark cluster ensemble model**

**According to the spark cluster model, data are entered in both ensemble models contemporaneously.**

The random forest with a random selection of variables, and the penalized logistic regression [1] that is the ensemble model

The algorithm is processed in seven steps. In the first step of the modeling the random forest model, three components should be considered. One of them is the node size that we considered 6, next is the number of trees (ntree) that we considered 2, and finally is the mtry [8] that we considered 4. In the second step, we select variables according to the amount of mtry and put this value in the k-fold cross-validation method, so that each time the variables are selected only the selected numbers are selected completely randomly. With this method we will be able to increase the random forest model accuracy. Fig.2 shows an example of the proposed random forest model. In the third step, threshold effects are identified from the random forest model. The proposed random forest model creates 6 binary prediction variables for each loan applicant customer i. In Fig.2 job status variable is the mth prediction variable, income is the lth, housing status is the nth, and the savings account status fth, so that each of the prediction variables indicates a threshold effect.

First prediction variable in the random forest model

1. We create two leaf nodes for the first variable.
2. In the upper nodes, we put variables with the maximum ability to divide the loan applicant customers into overdue and non-overdue.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

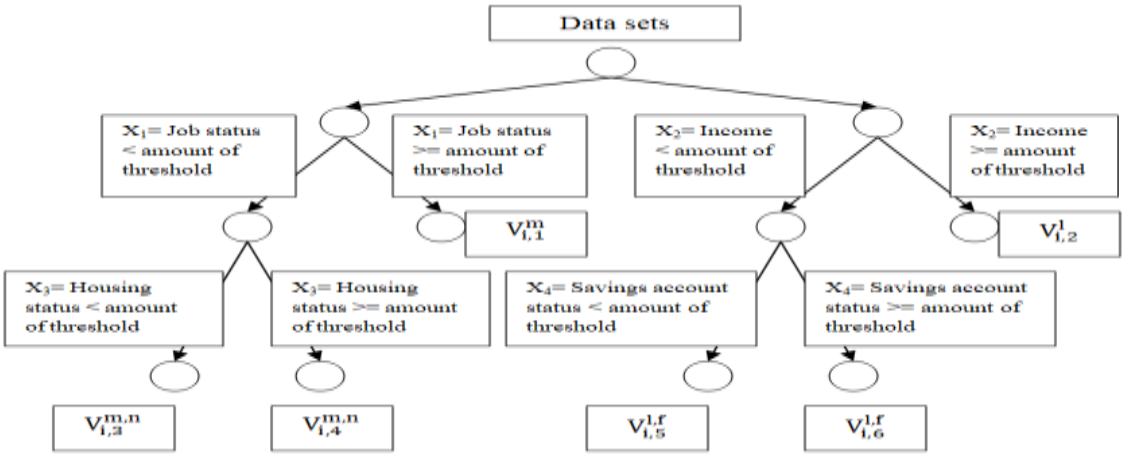


**Fig.2. Example of the proposed random forest model**

Variable $V_{i,1}^{m}$ is the first one-variable output of the random forest model, and if the loan applicant customer's job status is greater than or equal to the amount of threshold, so it takes the value one, and otherwise, it takes zero.

Second prediction variable in the random forest model

We do the same process as above for the second tree of the random forest model, and we achieve the second output of the random forest model $V_{i,2}^{l}$. The value of this variable is equal to one when the income of the loan applicant customer is greater than or equal to the amount of threshold, and otherwise, it takes zero.

Third prediction variable in the random forest model

We keep the leaf of the longest branch of the tree which is a sign that the loan applicant customer is overdue as a two-variable output. Variable $V_{i,3}^{m,n}$ is the first two-variable output of the random forest model. In the example presented in Fig.2 , this variable takes the value one if the loan applicant customer's job status is smaller than the amount of threshold and at the same time his housing status is smaller than the amount of threshold, and otherwise, it takes zero. Variable $V_{i,4}^{m,n}$ is the second two-variable output of the random forest model that according to the example presented in Fig.2 when the job status of the loan applicant customer is smaller than the amount of threshold, and at the same time his housing status is greater than or equal to the amount of threshold, so the value of this variable is equal to one, and otherwise, it takes zero.

Fourth prediction variable in the random forest model

Similar to the process above, we get $V_{i,5}^{l,f}$ , and $V_{i,6}^{l,f}$ .

Note that in the first tree of the random forest model for both outputs $V_{i,1}^{m}$ , $V_{i,4}^{m,n}$ the loan will be granted to the loan applicant customer i, and $V_{i,3}^{m,n}$ is definition as overdue, and so in the second tree for $V_{i,2}^{l}$ , $V_{i,6}^{l,f}$ outputs loan will be granted to the loan applicant customer i, and $V_{i,5}^{l,f}$ is definition as overdue.

Note that if one of the random forest model trees doesn't have enough information to divide the loan applicant customers into overdue and non-overdue the random forest model will rely on another tree that has enough information to divide the loan applicant customers (That is if the information of that random forest model tree is not available for any reason).

In the fourth step, we count one-variable threshold effects, and name them p, and then we count two-variable threshold effects, and name them q [1].

In the fifth step, we put all of the threshold effects that have been obtained from the random forest model into the penalized logistic regression model [1] to classify the remaining loan applicant customers more accurately [1].

$$\Pr\left(y_i = 2 \mid V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{l,f}; \theta\right) = \frac{1}{1+\exp\left[-\eta\left(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{l,f}; \theta\right)\right]}$$

**(1)**

1- In the formula above $(x_i,y_i)$, i=1,2,3,…,n are data set observations that in this observations $x_i$ is independent variable and $y_i$ is dependent variable that $x_i \in R^p$ and p is a p dimensional of predictors

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

and on the other hand $y_i \in \{1,2\}$ are binary predictor variables.

2- The goal of logistic regression model is obtaining an estimate of the probability of the loan applicant customer is overdue according to his related variables in the data set. That's mean $Pr(y_i=2|\ x_i)$.

3- $\eta\left(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{l,f}; \Theta\right)$ is an index function.

4- $\boldsymbol{\Theta}$ is a vector of parameters with a random value of [-1,1].

$$\eta(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{lf}; \theta) = \rho_0 + \sum_{m=1}^{p} \tau_m x_i + \sum_{m=1}^{p} \rho_m V_{i1}^{m} + \sum_{m=1}^{p-1} \sum_{n=m+1}^{p} \zeta_{m,n} V_{i,4}^{m,n} + \lambda_0 + \sum_{l=1}^{p} \psi_l x_i + \sum_{l=1}^{p} \lambda_l V_{i,2}^{l} + \sum_{l=1}^{p-1} \sum_{f=l+1}^{p} \varphi_{lf} V_{i6}^{lf}$$

(2)

1- In the formula above $\rho_0$, $\tau_m$, $\rho_m$, $\zeta_{m,n}$ are the parameters of $\boldsymbol{\Theta}$ for the first tree of the random forest model and on the other hand $\lambda_0$, $\psi_l$, $\lambda_l$, $\varphi_{l,f}$ are parameters of $\boldsymbol{\Theta}$ for the second tree of the random forest model.

2- m, n, l, f are binary prediction variables that select by random forest model.

3- p is the number of one-variable output and since the number of selected variables for each random forest model is 4, we consider this value equal to 4.

4- $V_{i,1}^{m}$ , $V_{i,4}^{m,n}$ are threshold effects outputs from the first tree of the random forest model and $V_{i,2}^{l}$, $V_{i,6}^{l,f}$ are threshold effects outputs from the second tree of the random forest model.

$\boldsymbol{\Theta}$ is equal to below [1]

$$\Theta = \left(\rho_0, \tau_1, \dots, \tau_p, \rho_1, \dots, \rho_p, \zeta_{1,2}, \dots, \zeta_{p-1,p}, \lambda_0, \psi_1, \dots, \psi_p, \lambda_1, \dots, \lambda_p, \varphi_{1,2}, \dots, \varphi_{p-1,p}\right)'$$

(3)

In the sixth step, we get the logarithm of the probability of the logistic regression model according to [1]

$$\mathcal{L}(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{lf}; \theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i \log \left[ F\left(\eta(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{lf}; \theta)\right)\right] + (1 - y_i) \log \left[1 - F(\eta(V_{i,1}^{m}, V_{i,4}^{m,n}, V_{i,2}^{l}, V_{i,6}^{lf}; \theta))\right]\right]$$

(4)

1- n is the number of data set observations.

In the seventh step, we add a penalty term to the negative value of the logarithm of probability [1].

$$\mathcal{L}_p\left(v_{i,1}^{m}, v_{i,4}^{m,n}, v_{i,2}^{l}, v_{i,6}^{lf}; \theta\right) = -\mathcal{L}\left(v_{i,1}^{m}, v_{i,4}^{m,n}, v_{i,2}^{l}, v_{i,6}^{lf}; \theta\right) + \lambda p(\theta)$$

(5)

1- $p(\boldsymbol{\Theta}) = \sum_{v=1}^{V} w_v |\theta_v|$ is the penalized phrase has been added.

2- $\boldsymbol{\lambda}$ It is an adjustment parameter that controls the intensity of regularization to reduce the out-of-sample error.

$$\hat{\theta}_{adaptive-lasso}(\lambda) = \arg_{\theta}\min - \mathcal{L}\left(v_{i,1}^{m}, v_{i,4}^{m,n}, v_{i,2}^{l}, v_{i,6}^{lf}; \theta\right) + \lambda \sum_{v=1}^{V} w_v |\theta_v|$$

(6)

1- $\arg_{\boldsymbol{\Theta}}\min$ is equal to the Quantities of $\boldsymbol{\Theta}$ that the function is minimized.

2- $w_v = |\hat{\theta}_v^{(0)}|^{-v}$, $\hat{\theta}_v^{(0)}$ is a consistent initial estimator of the parameters $\boldsymbol{\Theta}$, v=1,2,3,…,V is a positive constant value.

The decision tree [1] with a random selection of variables, and logistic regression (SLF) [10] that is the ensemble model

The algorithm is processed in four steps.

In the first step, we create the decision tree model [1] with the weighted outputs.

In the second step, we use the bagging trees model in order to implement the decision tree model, so that the random selection of variables occurs in order to increase the accuracy of the model. These variables have the most relationship with each other in term of weight.

In the third step, we give to the non-overdue outputs weight of one, and to the others zero.

An example of the decision tree produced by this model is shown in (Fig.3).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

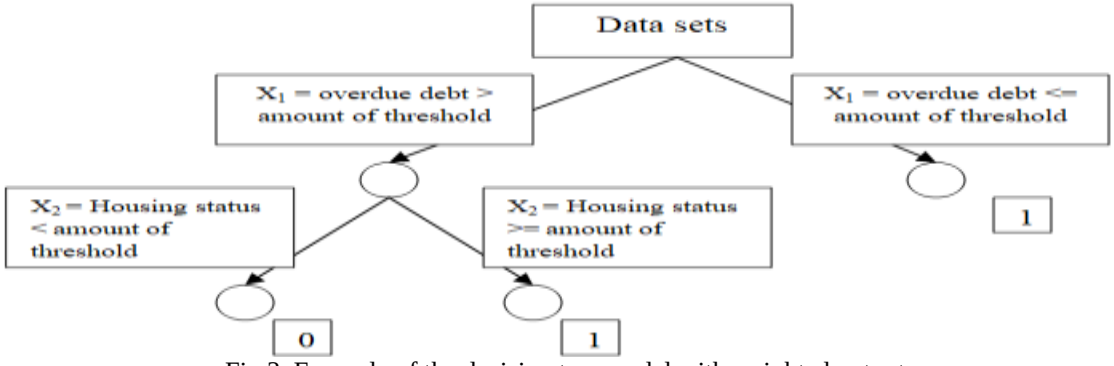

Fig.3. Example of the decision tree model with weighted outputs

Suppose that the loan applicant customer has overdue debt smaller than or equal to the amount of threshold, so in that case one-variable output takes the value one, and otherwise it takes zero, and if his overdue debt is greater than the amount of threshold, and at the same time his housing status is greater than or equal to the amount of threshold then two-variable output takes the value one, and otherwise, it takes zero, and it will be classified in the overdue category.

In the fourth step, we put the information obtained from the decision tree model of the previous step inside the logistic regression (SLF) model [10] for more detailed analysis.

The logistic regression (SLF) model [10] is expressed as follows (With this model, we will be able to simplify the logistic regression model [1]).

$$\frac{1}{y_t} = m + ae^{bt} \quad (7)$$

In the formula above $y_t$ represents entity weight (Customer) and r entity progress rate r>0, k maximum entity weight, $Y_0$ weight in ideal conditions, e Euler number (2.71828).

Suppose that on the above formula $m = \frac{1}{k}$, $a = (k - y_0) / (ky_0)\, e^{rt0}$ and $b = -r$, $m > 0$, $a > 0$, $b < 0$.

Suppose that:

$$x_t = m + aT_t^b \quad (8)$$

Suppose that on the above formula $x_t = \frac{1}{y_t}$ and $T_t = e^t$.

Now we add the Gaussian white noise $\eta \sim N(0,\delta^2)$ to the (8) formula:

$$x_t = m + a\,T_t^b + \eta \quad (9)$$

1- δ is equal to 1.

Suppose that $X = [X_1, X_2, X_3, \dots, X_N]^T$, its alternative probability model can be expressed as follows (X's are the weights of the loan applicant customer's variables)

$$P(x, a, m) = \frac{1}{2\pi\delta^2} * e^{\frac{\sum_{t=1}^{N}\left(x_t - m - a\,T_t^b\right)^2}{2\delta^2}} \quad (10)$$

1- π is equal to 3.141592.

2- N is the current year in which the loan applicant wants a loan.

3. Evaluation of the model

We evaluate the model according to [4] that the results show the superiority of the proposed model.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 202-207 www.ijaem.net ISSN: 2395-5252

**Table.1.The evaluation of the proposed model with the PLTR model [1]**

| PLTR model [1] | Proposed model | Evaluations |
|---|---|---|
| 0.335 | 0.085 | Execution time |
| 268 | 269 | True positive |
| 39 | 38 | False negative |
| 328 | 333 | True negative |
| 55 | 50 | False positive |
| 86.37% | 87.24% | Accuracy |
| 93% | 93% | Area under the ROC curve (AUC) |
| 87.29% | 87.62% | True positive rate |
| 14.36% | 13.05% | False positive rate |
| 82.97% | 84.32% | Precision |
| 87.29% | 87.62% | Recall |

## III. CONCLUSION

According to the current research, the spark cluster model, and the random forest model increase the speed of the ensemble model, and on the other hand the penalized logistic regression model [1], and (SLF) [10], and random selection of variables by the K-fold cross-validation method in the random forest model [8], and the use of the bagging trees in the decision tree model [1] increase the accuracy of the spark cluster ensemble model that the results show the superiority of spark cluster model. According to this article the proposed model is 0.25s faster and 0.87% more accurate to the PLTR model therefore this model can be used of credit institutions and banks.

## REFERENCES

[1] Dumitrescu, E., et al., Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 2022. **297**(3): p. 1178-1192.

[2] Liu, W., H. Fan, and M. Xia, Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. Engineering Applications of Artificial Intelligence, 2021. **97**: p. 104036.

[3] Pławiak, P., M. Abdar, and U.R. Acharya, Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. Applied Soft Computing, 2019. **84**: p. 105740.

[4] Ruyu, B., H. Mo, and L. Haifeng, A Comparison of Credit Rating Classification Models Based on Spark-Evidence from Lending-club. Procedia Computer Science, 2019. **162**: p.811-818.

[5] Yu, H., et al., A three-way cluster ensemble approach for large-scale data. International Journal of Approximate Reasoning, 2019. **115**: p. 32-49.

[6] Expósito, R.R., J. Gonzalez-Dominguez, and J. Tourino, SMusket: Spark-based DNA error correction on distributed-memory systems. Future Generation Computer Systems, 2020. **111**: p. 698-713.

[7] Saud, S., et al., Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach. Sustainable Energy Technologies and Assessments, 2020. **40**: p. 100768.

[8] Wang, Z., et al., Random Forest based hourly building energy prediction. Energy and Buildings, 2018. **171**: p. 11-25.

[9] Speiser, J.L., et al., A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 2019. **134**: p. 93-101.

[10] Feng, R., et al., Saturated load forecasting based on clustering and logistic iterative regression. Electric Power Systems Research, 2022. **202**: p .107604.