A Non-Parametric-Based Computationally Efficient Approach for Credit Scoring

Afshin Ashofteh and Jorge Bravo

September 29, 2019

A Non-Parametric-Based Computationally Efficient Approach for Credit Scoring

Abstract

This research aimed at the case of credit scoring in risk management and presented the novel method for credit scoring to be used for default prediction. This study uses Kruskal-Wallis non-parametric statistic to form a computationally efficient credit-scoring model based on artificial neural network to study the impact on modelling performance. The findings show that new credit scoring methodology represents reasonable coefficient of determination and low false negative rate. It is computationally less expensive with high accuracy. Because of the recent respective of continues credit/behavior scoring, our study suggests to use this credit score for non-traditional data sources such as mobile phone data to study and reveal changes of client's behavior during the time. This is the first study that develops a non-parametric credit scoring, which is able to reselect effective features for continues credit evaluation and weighted out by their level of contribution with a good diagnostic ability.

Keywords: *Credit scoring; Neural network; Loan default; Kruskal_Wallis statistic.*

1. INTRODUCTION

Credit scoring involves the use of statistical methods to transform relevant data into numerical measures that inform and determine credit decisions. In recent years the use of credit scoring tools has expanded beyond their original purpose of assessing credit risk, being currently used for establishing the initial and ongoing credit limits available to borrowers, for assessing the risk-adjusted profitability of account relationships, and for assisting in a range of activities in loan servicing, including fraud detection, delinquency intervention, and loss mitigation (Tomas, 2000; Andersson, 2007). The Z-score model of Edward Altman for bankruptcy prediction and the FICO score for retail credit scoring are some of the oldest industry standards, which loan providers still use because of their high interpretability (Baesens et. al., 2016).

The regulatory changes brought by the revised Basel Accords (subsequently adopted by national legislation in many countries and regions) introduced stronger risk management requirements for banks. The main instruments of these regulations are the minimum capital requirements, the supervisory control mechanisms and the market discipline. Under this new regulation, the capital requirements are tightly coupled to estimated credit portfolio losses. According to the Basel II/III "internal ratings-based" (IRB) approach, financial institutions can use their own internal risk measures for key drivers of credit risk as key inputs in providing loss estimates for the mortgage book and in computing capital requirements (Basel 2006). The expected credit loss or impairment calculation rules imposed by the recently adopted IFRS9 and FASB's Current Expected Credit Loss (CECL) standards require financial institutions to calculate expected loss for the banking book over the entire life of the exposures, conditional on macroeconomic factors, on a point-in-time basis, that is, recalibrating PDs where necessary to reflect the effects of the current economic conditions.

To assess credit risk, in developed markets lenders typically consider historical loan application and loan performance data collected regularly from a small number of sources on the basis of long-standing banking and credit relationships to develop credit-scoring models to evaluate the ability to repay, the willingness to repay and identify fraud. These methods are less effective in emerging economies and among low-income unbanked segments of the population which often do not have access to formal financing and/or do not earn regular labour income. To cope with this constraints and to better assess credit risk, banks and loan providers are increasingly using non-traditional data sets (e.g., mobile operators, utilities, retailers, and direct-sales companies data) to sophisticate their credit bureaus and their credit rating services. This poses new challenges to credit scoring modellers since non-traditional data must typically be collected from different sources and its volume is several times that of traditional sources. By pursuing this approach, lenders seek to have more accurate information and incentive to grow the credit market under a robust credit control framework. By increasing their use of these new data sources, they try to provide more lending to the public customers and get to better analyze of loan requests, ultimately increasing the loan ratio and decreasing the decision time. People will then have more monthly disposable money for spending which will contribute to the economy, but it can also make risks for financial institutions. Therefore, non-traditional data sets provide credit market a chance to manage different data sources to boost credit analysis outcomes and follow the stipulated recommendations of standards in an appropriate way.

Credit scoring by using non-traditional data sets is a cost effective method of surveying personalities for risk management purposes of monetary institutions. It shifts credit scoring to high-tech to avoid the personal subjectivity of analysts or an underwriter (Fensterstock, 2005). It also helps in increasing the speed and consistency of the application processes and allows financial firms to automate their processes in this one-click new world (Rimmer, 2005). Credit scoring and new technologies helps loan providers to shorten the process time of loan applications and improve the allocation of resources (Jacobson & Roszback, 2003). Additionally, it can aid insurance firms in making better predictions on claims and determining the interest rate which the firms should charge their consumers as well as the pricing of portfolios and products (Avery et. al., 2000) (Kellison & Brockett, 2003).

Recently, mobile phones provide non-traditional data sources in the form of call-detail records (CDR) and many other log files. These new sources of Big Data attain much more importance to provide smarter credit-scoring models not only for customers of banks but also for largely unbanked population who has no regular credit history.

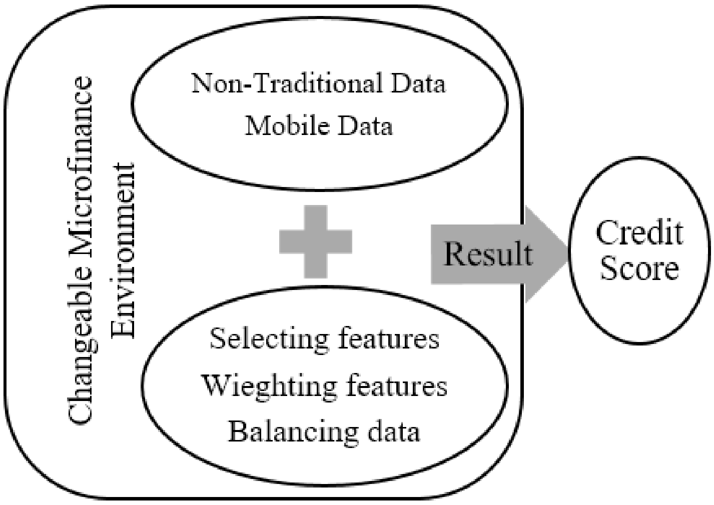

Fig. 1. Challenges in credit scoring.

While this approach presents some opportunities, it also carries some challenges. First, most of the mobile phone features are redundant and do not contribute to representing credit risk. Second, Banks and loan providers should follow the regulators and the important change factor in banking is regulation. It means that fluctuations in economy and regulations could change the behaviour of both banks and customers. Non-traditional data sets can reveal these fluctuations, but current methods are computationally expensive with high false negative rate. The third risk factor is the smart behaviour of fraudsters and it means conscious changes may exist in non-traditional data, which has influence on contribution level of features in credit scoring. However, the current methods are not wise enough to renew the credit scores during the time. Traditional credit-scoring models applying single-period classification techniques (e.g., logit, probit) to classify credit customers into different risk groups and to estimate the probability of default are still the most popular data mining techniques used in the industry (Chamboko & Bravo, 2019a,b). Despite their popularity, scoring models can only provide an estimate of the lifetime probability of default for a loan but cannot identify the existence of cures and or other competing transitions and their relationship to loan-level and macro covariates, and do not provide insight on the timing of default, the cure from default, the time since default and time to collateral repossession (Gaffney et al., 2014; Lessmann et al., 2015). Discriminant analysis, survival models, decision trees, support vector machines, artificial neural networks, genetic programming and standard models using external ratings provided by external credit assessment institutions have also been successfully applied (Arminger et al., 1997; Hand & Henley, 1997; Baesens et al., 2003; Kruppa et al., 2013; Lessmann et al., 2015; Butaru et al., 2016; Abellán & Castellano, 2017; Chamboko & Bravo, 2016, 2019a,b).

As the result, we need highly effective and computationally less expensive solutions to calculate an informative credit score for satisfying the accuracy expectation of financial institutions. Although there are large number of techniques employed in development of credit-scoring such as discriminant analysis, Probit/Logistic regression, kernel support vector machine (Bellotti & Crook, 2009) and different hybrid pre-processing approaches, empirical studies show that the false negative rate obtained is still not good enough for non-traditional data sets and they are time independent.

In this paper, we introduce a novel time dependent credit scoring method to determine good loans with low false negative rate. It reselects the most prominent features for credit evaluation over time and, as the result, it is computationally more efficient to implement. It also offers an internal mechanism to reselect prominent features for credit scoring to deal with changeable microfinance environment effectively.

The rest of the paper is structured as follows. In Section 2 we review credit-scoring models and new non-traditional data sources. Section 3, introduces a novel credit scoring method and discusses the calculation of the credit score. Then, using the credit risk data set, we compare the classification accuracy of credit score with available features in section 4. Additionally, an artificial neural network model is dedicated to the new method to show accuracy of predicting the probability of default. In section 5 we discuss the main managerial and theoretical implications of this research. Finally, Section 6 contains some concluding remarks.

2. LITERATURE REVIEW

Recent research in credit scoring mostly focuses on three dimensions: novel classification algorithms, novel performance measures, and statistical hypothesis tests. The first dimension concerns the development of scorecards, for example, an extreme learning machine, the second dimension assesses scorecards like H-measure, and the third dimension compares scorecard performance (García et al., 2010). However, mobile phone data is a new Big Data source for smarter credit-scoring models, independent of the usual databases of financial institutes. It can improve retail risk models and be used as a good solution not only for current customers of banks but also for the unbanked. Figure (2) represents how non-traditional data is emerging in credit scoring by using parallel computing, distributed computing and Big Data solutions. They represent how digitization in banking has gradually allowed financial institutions to use both Big Data and traditional credit records for managing risks. It shows how technology development is helping loan providers to create value from several huge volumes of non-traditional data (e.g., mobile phone data) with increasing computation efficiency.

Providing faster and consistent decisions for sub-prime customers with poor credit records, credit impairment, missing data in their credit histories, or difficulty in validating their income is another advantage of using non-traditional data in credit scoring modelling (Quittner, 2003). Some researchers use behavioural signatures in mobile phone data to predict default with an accuracy almost similar to that of credit-scoring methods that use financial history by Random Forest and Logistic regression (Björkegren & Grissen, 2017). Pedro et al. (2015) developed MobiScore, a methodology that models the user's financial risk using data collected from mobile usage using gradient boosting, support vector machine and linear regression models. Other studies have used boosted decision trees and logistic regression to create a credit score for under-banked populations considering information about people's usage of various mobile apps to make conclusions about their mood and personality traits (Skyler et. al., 2017; Chittaranjan et. al., 2012; Do & Gatica, 2010; Verkasalo, 2010). Recently, mobile phone data and social network analytics were used in credit scoring applications showing that incorporating telco data has the potential to increase the value of credit scoring (Oskarsfotti et. al., 2019).

Current credit scoring methods are computationally expensive. Additionally, non-traditional data for credit scoring is suffering the class imbalance problem. Because of minority of bad customers, class distributions are highly imbalanced and represented skewed distributions. Although the topic of imbalanced classification has gathered a wide attention of researchers during the last several years such as cost-sensitive learning technique by Douzas and Bacao (2018) (Douzas & Bacao, 2018), however, the emergence of mobile phone data brings new problems and challenges for the class imbalance problem in credit scoring, especially for unbanked individuals. Any rate of false negative in the models means big loss for loan providers. Recently, sound statistical and machine learning procedures that are scalable computationally to massive non-traditional datasets have been proposed (Jordan, 2013). Examples are subsampling-based approaches (Kleiner et. al., 2014, Liang et. al., 2013, Ma et. al., 2013 and Maclaurin & Adams, 2014), divide and conquer approaches (Song & Liang, 2014), and online updating approaches (Schifano et. al., 2015). Even though these methods are suitable for balanced data sets but in the case of using non-traditional data for credit-scoring systems, inner problems of imbalanced data, namely lack of data and small disjuncts are still accentuated especially when the amount of data is huge and during the data partitioning to fit the Map-Reduce programming model for data processing. Oskarsfotti et al. used undersampling method to reduce class imbalance for training set of data, which shows the intention to reduce the size of majority class when applying these analytics techniques (Oskarsfotti et. al., 2019).

3. DYNAMIC NONPARAMETRIC CREDIT SCORE

3.1 Kruskal-Wallis statistic for reduction and empowering features

The performance of classification system can be improved by picking up the optimized features of mobile phones and decreasing the complexity of model in preprocessing stage. There are many methods that have been developed for choosing significant features with high information such as Kruskal-Wallis method (Saeys et. al. 2007). The Kruskal-Wallis test as a nonparametric approach is useful to select informative features for loan default in credit risk management. Because it is sensitive to events which are far from the credit scores of good clients, we use Kruskal-Wallis non-parametric statistic in our proposed method, which is computationally less expensive and very simple to implement. We not only use these advantages but also introduce a credit scoring which is able to purify features and decrease the dimensions in real time. It also use Kruskal-Wallis statistic as a weight of feature to decrease the false negative rate.

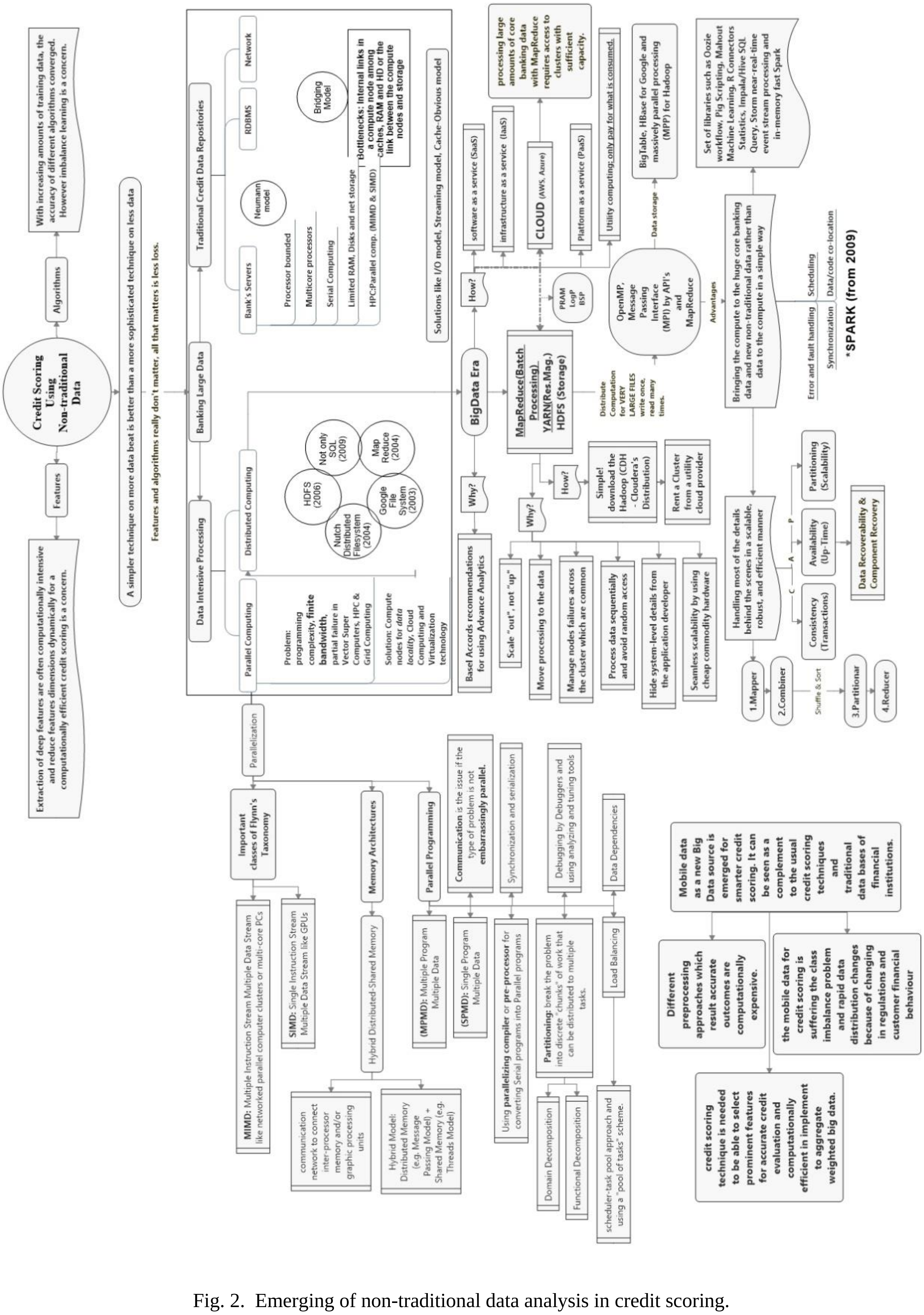

Fig. 2. Emerging of non-traditional data analysis in credit scoring.

Let us define the null hypotheses as mobile phone feature does not contain discriminative information to detect default possibility in loan request; otherwise it will be selected as informative feature. An assumption for this test is that the samples from the credit scores of good clients and credit scores of new clients are independent random samples from continuous distributions. In addition, we consider the time as an index of weights in our credit score method to make sure that distribution for training dataset of customers and new customers have the same shape at the time of analysis.

The computational procedure of the test can be considered as following. Let X_{ijt} denote an observation of feature j from mobile phone of client i at time t. If we let N be the total number of credit scores or total number of customers at time t by aggregating credit risk indices of $X_{ij}'s$, then we have a matrix $X^t{}_{N\times K}$ with N rows as number of clients and K as number of features at time t. Loans in banking credit risk literature usually divide into three categories of either being "Good" for good loans, "Medium" for substandard-loans/doubtful loans, and finally "Bad" for loss loans and Kruskal-Wallis test is appropriate for these kind of categorical variables with three or more than three groups. However, most of available datasets for credit purposes merge the first and the second groups.

We consider the number of categories as S representing the number of samples in Kruskal Wallis statistic. Therefore, consider vector $Y_{N\times 1}$ as label vector included three possible labels for each customer as "Good", "Substandard or Doubtful" and "Loss" or two groups as discussed above.

Computationally, Kruskal-Wallis statistic for j^{th} feature at time t for N customers is

$$H_j^t = \frac{12}{N(N+1)}\left(\sum_{i=1}^{S}\frac{R_{ij}^2}{n_{ij}}\right) - 3(N+1). \quad j = 1,2,\dots,K.$$

Then, as $R_{ij}$is the rank assigned to the j^{th} feature of i^{th} client

$$H_j^t \xrightarrow{d} \chi_1^2 \qquad \textit{in distribution}.$$

Where χ_1^2 is the χ^2 distribution with one degree of freedom. Null hypothesis will be rejected if the computed value of H_j^t for each j from 1 to k exceeds the value of chi-square for reselected confidence level and 1 degree of freedom. Simply, in credit scoring scenario, we define γ_j^t to find a balanced measure with less complexity to be followed with zero as base line and negative-positive values for making decisions about features based of following τ_j^t.

Let

$$\tau_j^t = 1 - \frac{H_j^t}{H_j^t + \chi_1^2} = \frac{\chi_1^2}{H_j^t + \chi_1^2} \ , \qquad (1)$$

$$0 \le \tau_j^t \le 1 \text{ ; for all } j, j = 1,2,\dots,k.$$

Hence, we obtain that

$$-0.5 \le \tau_j^t - 0.5 \le 0.5.$$

We define a nonparametric measure γ for feature j at time t as $\gamma_j^t = \tau_j^t - 0.5$ therefore

$$\gamma_j^t = \frac{\chi_1^2 - H_j^t}{2 \times (\chi_1^2 + H_j^t)} \ , \qquad (2)$$

and

$$-0.5 \le \gamma_j^t \le 0.5. \qquad (3)$$

If γ_j^t is positive then the feature is not able to differentiate the classes and if it is negative then the feature could be used for modeling phase to determine the creditworthiness of an applicant based on a set of selected features. It is clear that this change is only superficial to make it easier to understand and represent it as a control chart in business intelligence dashboards. Now, we study the behavior of γ_j^t by looking at its distribution function.

Let γ_j^t be as is introduced in Formula (2) and F be the corresponding cumulative distribution function (cdf).

$$F_{\gamma_j^t}(y) = P(\gamma_j^t \le y) = 1 - P\left(\frac{\chi_1^2 - H_j^t}{2\times\left(\chi_1^2 + H_j^t\right)} > y\right) = 1 - P\left(H_j^t < \frac{(1-2y)\chi_1^2}{(1+2y)}\right),$$

$$F_{\gamma_j^t}(y) = 1 - F_H\left[\left(\frac{0.5-y}{0.5+y}\right) \times \chi_1^2\right].$$

where F_H is the cumulative distribution function of Kruskal-Wallis statistic and $\left(\frac{0.5-y}{0.5+y}\right) \times \chi_1^2 > 0$ for any values of $y \in (-0.5, 0.5)$.

Furthermore, because $H \to x_1^2$ in distribution, the density function of γ_j^t will be

$$f_{\gamma_j^t}(y) = \frac{dF_{\gamma_j^t}(y)}{dy} = -\frac{d\left(\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right)}{dy} \times f_H\left(\left(\frac{0.5-y}{0.5+y}\right) \times \chi_1^2\right);$$

$$f_{\gamma_j^t}(y) = \sqrt{\frac{\chi_1^2}{2\pi(0.25 - y^2)(0.5 + y)^2}} \times Exp\left\{-\left[\left(\frac{0.5 - y}{1 + 2y}\right) \times \chi_1^2\right]\right\}, \quad -0.5 < y < 0.5$$

In 95% confidence level the $\chi^2_{1,0.05} = 3.841$ and the γ_j^t density function is

$$f_{\gamma_j^t}(y) = 0.7818 \times \sqrt{\frac{1}{(0.25 - y^2)(0.5 + y)^2}} \times Exp\left[\frac{3.841 \times (y - 0.5)}{1 + 2y}\right], \quad -0.5 < y < 0.5$$

The density function of γ_j^t is shown below (Fig. 3.).

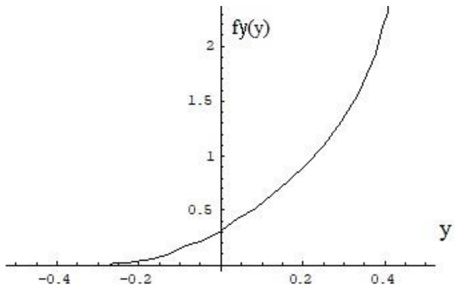

Fig. 3. The density plot of γ_j^t.

From Figure (3), it is clear that the majority of probability density based on area under the density function curve for statistically significant features is between -0.3 and 0. We can expect that most of magnificent fluctuations in effectiveness of features on our credit score model will happen in the small probability area under the density function between -0.3 and 0. Therefore, if we select the informative features based on γ_j^t, we can manage features and eliminate unnecessary ones in our computer program only by considering the positive sign of γ_j^t instead of its value. For instance, a char type of data needs only one byte of memory but decimal needs 12 bytes. It means less memory and data transfer in network will be occupied for huge non-traditional datasets with numerous features.

3.2 Credit score formulation

In this step, we use γ_j^t in credit score model to boost the effective features. It was shown that γ_j^t can be a part of credit scoring as an indicator for feature selection and transformed γ_j^t can be used as a transformation weight for k^{th} feature to improve the performance of classification.

Let us define w_k^t as the transformation weight of feature k at time t:

$$w_j^t = \begin{cases} 2 \times \left|\gamma_j^t\right| & \text{for} \quad \gamma_j^t < 0 \\ 0 & \text{for} \quad \gamma_j^t > 0 \end{cases}, \quad \text{j=1,2,…,N} \qquad (4)$$

then φ as impact factor of feature k at time t is

$$\varphi_j^t = \begin{cases} \frac{w_j^t}{\sum_{j=1}^{k} w_j^t} & \text{for} \quad w_j^t > 0 \\ 0 & \text{for} \quad w_j^t = 0 \end{cases}, \qquad \sum_{j=1}^{N} \varphi_j^t = 1. \qquad (5)$$

For 95% of confidence level, if $\left\{H_j^t \middle| H_j^t > \chi^2_{1,0.05} = 3.84\right\}$ then we reject the null hypotheses. It means that mobile phone feature contains discriminative information to detect default possibility in loan request. The equivalent values of γ_j^t and w_j^t to values of H_j^t is shown below in Table (1).

H_j^t	0	1	2	3	3.5	4	5	10	20	40	100	400	700
γ_j^t	0.5	0.29	0.15	0.06	0.02	-0.01	-0.06	-0.22	-0.33	-0.41	-0.46	-0.49	-0.495
w_j^t	0	0	0	0	0	0.020	0.131	0.445	0.678	0.825	0.926	0.981	0.989
φ_j^t	0	0	0	0	0	0.004	0.026	0.089	0.136	0.165	0.185	0.196	0.198

Table 1. $\boldsymbol{H_j^t}$ values and its equivalent in $\boldsymbol{\gamma_j^t}$, $\boldsymbol{w_j^t}$ and $\boldsymbol{\varphi_j^t}$

As it was shown in Figure (3), important fluctuations in effectiveness of features in credit scoring happens when γ_j^t is between -0.3 and 0. In Table (1) it is shown that w_j^t and φ_j^t react magnificently to the changes of γ_j^t in this interval. Therefore, w_j^t close to one shows the high capability of the feature to recognize the credit level of the customer and trend to zero means the probability of feature's influence is decreasing. Monitoring the trend of w_j^t for different t could show the level of model's reliability during the time. If it decreases to zero randomly and sequentially then it can be concluded that the reliability of trained model based on those features is decreasing and it is a sign of warning to renew the model by using new set of features. It helps to identify for how long our scoring model keep up with the initial performance and to discover the right time of redoing training step by new set of features. It would be also useful to monitor w_j^t by considering zero for excluding the feature from the credit score model. Additional to all mentioned advantages, we use w_j^t at time t as the power of attribute j to improve the accuracy of model in false negative rate.

For this purpose, we consider all attributes for defaulted customers to the power of their w_j^t in the training stage. The attributes with the higher H_j^t will have w_j^t closer to one. It means they experience less change than one with lower ability to determine the loan default. As the result, in training stage, we can make a difference between attributes of good and bad loans based on their power of contribution in the model. In fact, we are saving the high performance attributes and we use the less important attributes as labels for bad loans. It can help the model to use the attributes with high w_j^t for making model and the attributes with low w_j^t to differentiate between good and bad loans.

Now, we make a credit risk index (CRI) for attributes with interval or ratio scale of ith client at time t by using geometric mean as following

$$CRI_{it} = \prod_{j=1}^{k} \left[\frac{(x_{ij}^t)^{w_j^t}}{(\bar{x}_j^t)} \times 100 \right]^{\varphi_j^t}, \ i = 1,2,\dots,N.$$

Where $\bar{x}_j^t$ is the normal profile of attribute j extracted from data of good clients. If we investigate this credit score characteristics by using the most desirable axioms of axiomatic approach to index number theory, it satisfies most of them[1]. Thus, this credit index formula as a homogeneous symmetric average can be calculated as an accurate aggregate measure and it is able to renew features dynamically and weighted out by φ_j^t as their impact factors. The pseudo code of the proposed methodology is listed in Table 2.

```
INPUT attributes and Status variable;
OUTPUT weighted attributes, CRI;
STATEXPLORE attributes;
CHANGE[outliers] = FALSE;  {outliers are important in credit scoring.}
SET χ²_{1,0.05} = 3.84;
FOR each attribute DO
    Kruskal-Wallis.test;
    IF (Scale[attribute] is not Nominal) THEN
        IF (KW.statistic > χ²_{1,0.05})  THEN
            IF (Default=True) THEN
                SET Gama = (χ²_{1,0.05} – KW.statistic) / (2 * (χ²_{1,0.05} + KW.statistic ));
                SET W = 2 * ABS(Gama);
            ELSEIF (data=test.set) THEN
                SET Gama = (χ²_{1,0.05} – KW.statistic) / (2 * (χ²_{1,0.05} + KW.statistic ));
                SET W = 2 * ABS(Gama);
            ELSE
            SET attribute = unchanged; {equivalent to Set W = 1;  Set Phi = 1}
        ELSE
         SET attribute = excluded;  {equivalent to Set W = 0;  Set Phi = 0;}
    Else
    SET attribute = unchanged; {equivalent to Set W = 1;  Set Phi = 1}
END DO
FOR each attribute DO
    IF (Default=False) THEN  Mean_default_NO= AVERAGE(attribute);
    SET attribute.value.W = POWER [attribute.value , W]
    SET Phi = W / SUM(W's);
    SET attribute_J_CRI = POWER [(attribute.value.W / Mean_default_NO× 100) , Phi]
END DO
FOR each client DO
    CRI = Multiply(attribute_J_CRI)
END DO
```

Table 2. Pseudo Code of Proposed Methodology

In the test stage, obviously we will have a very large area under the AUC curve because the values of bad loans in the test set have experienced the same changes as the values of training set. However, in the validation stage, we are using an independent out-of-sample dataset and we have actually no idea which loan is default. As it is shown in table (2), we apply the transformation for attributes of all clients included good and bad loans to be used in the validation stage. It will shift the attributes of good loans to the bad loans and decrease the true

[1] . Axiomatic approach to index number theory such as Positivity test, Continuity Test, Identity Test, Homogeneity Test for Period t, Homogeneity Test for Period zero, Commodity Reversal Test, Invariance to Changes in the Units of Measurement or the Commensurability Test, Time Reversal Test, Circularity or Transitivity Test, Mean Value Test, Monotonicity Test with Respect to Period t and Monotonicity Test with Respect to Period zero.

positive rate, but it will also decrease the false negative rate dramatically. It could be interesting for loan providers especially when they want to offer a loan to clients without any credit history and only based on non-traditional data analysis. In this case, it is beneficial if we can detect the separated sections of good and the bad customers and struggle to detect the good customers from the muddy intersection of good and bad loans in dataset, where there are high similarity in attributes of different categories. Additionally, using CRI as an aggregate of features with interval/ratio scale will decrease the required computation for modeling significantly.

4. EXPERIMENTAL DESIGN

4.1 Data description

This research used a credit score data of a loan provider and the targets were loan request of its clients. The dataset is anonymized, and do not contain personal information. Among the total 6377 observations, 2217 observations (34.8%) are loan requests with default payment according to the Basel definition, which having three or more late payments implies default. The target vector is donated by a binary variable named "Status" which represent loan default (No=1, Yes=2). This study used the following 13 variables as explanatory variables, "V1: Seniority" for Job seniority (year), "V2: Home" for type of home ownership, "V3: Time" for time of requested loan, "V4: Age" for client age, "V5: Marital" for marital status, "V6: Records" for existence of records, "V7: Job" for type of job, "V8: Expenses" for amount of expenses, "V9: Income" for amount of income, "V10: Assets" for amount of assets, "V11: Debt" for amount of debt, "V12: Amount" for amount requested of loan and "V13: Price" for price of goods.

Descriptive statistics of categorical variable V2 is shown in Table (3).

LOAN DEFAULT	V2: TYPE OF HOME OWNERSHIP						
	RENT	OWNER	PRIVATE	IGNORE	PARENTS	OTHER	MEDIAN
NO	580	1697	159	11	544	1169	Owner
YES	1383	368	82	9	230	145	Rent

Table 3. Descriptive Statistics of "Type of home ownership"

and descriptive statistics of categorical variables V6 and V7 is shown in Table (4).

LOAN DEFAULT	V6: EXISTENCE OF RECORDS			V7: TYPE OF JOB				
	NO-REC	YES-REC	MEDIAN	FIXED	PART TIME	FREE LANCER	OTHERS	MEDIAN
NO	2820	1340	No-rec	3206	180	673	101	Fixed
YES	1802	415	No-rec	577	269	307	1064	Freelance

Table 4. Descriptive Statistics of "Existence of records" & "Type of job"

and categorical variables V5 is described in Table (5).

LOAN DEFAULT	V5: MARITAL STATUS					
	SINGLE	MARRIED	WIDOW	SEPARATE	DIVORCED	MEDIAN
NO	639	3383	47	67	24	MARRIED
YES	1318	806	19	60	14	SINGLE

Table 5. Descriptive Statistics of "Type of home ownership"

In addition, descriptive statistics of other variables is shown in following table (Table 6)

	LOAN DEFAULT	N	MEAN	MIN.	MAX.	STD. EV.	KURTOSIS	SKEWNESS
JOB SENIORITY (YEAR)	**NO**	4160	16.06	0	48	14.26	-1.01	0.643
	YES	2217	3.82	0	43	4.62	13.57	3.15
	TOTAL	6377	11.81	0	48	13.20	0.05	1.18
TIME OF REQUESTED LOAN	**NO**	4160	36.05	6	60	21.47	-1.48	-0.28
	YES	2217	59.39	6	72	14.81	0.29	-1.06
	TOTAL	6377	44.16	6	72	22.37	-1.01	-0.53
CLIENT AGE	**NO**	4160	43.13	18	70	14.00	-1.16	0.179
	YES	2217	29.06	18	65	10.37	0.59	1.18
	TOTAL	6377	38.23	18	70	14.50	-0.96	0.47

AMOUNT OF EXPENSES	**NO**	4160	80.71	35	173	48.22	-0.82	0.89
	YES	2217	50.16	35	165	17.66	3.70	1.80
	TOTAL	6377	70.09	35	173	42.86	0.49	1.38
AMOUN T OF INCOME	**NO**	4160	325.26	0	959	334.9	-0.66	1.07
	YES	2217	99.38	0	959	71.22	11.49	1.69
	TOTAL	6377	246.73	0	959	294.11	1.09	1.66
AMOUNT OF ASSETS	**NO**	4160	62212.56	0	250000	100457.06	-0.55	1.19
	YES	2217	6474.55	0	100000	7860.78	28.93	3.58
	TOTAL	6377	42834.27	0	250000	85491.53	1.49	1.85
AMOUN T OF DEBT	**NO**	4160	269.05	0	23500	1008.43	143.02	9.07
	YES	2217	9164.08	0	30000	10492.02	-1.3	0.555
	TOTAL	6377	33.61.45	0	30000	7541.20	3.37	2.18
AMOUNT REQUESTED OF LOAN	**NO**	4160	1776.21	100	4500	1446.13	-0.71	1.00
	YES	2217	951.72	105	4500	519.94	5.44	1.63
	TOTAL	6377	1489.57	100	4500	1261.64	0.81	1.52
PRICE OF FINANC	**NO**	4160	3632.35	125	11140	3903.34	-0.56	1.17
	YES	2217	1163.38	105	6802	608.47	11.86	2.36
	TOTAL	6377	2774.00	105	11140	3383.74	1.37	1.79

Table 6. Descriptive Statistics of scale variables

For the purpose of simplicity, we consider t equal to one and we calculated γ_j^1 and w_j^1. In this paper, we use receiver operating curves (ROC) to show statistical performance of the models. In the ROC chart, the horizontal axis represents the specificity and the vertical axis shows the sensitivity. The greater the area between the curve and the baseline, the better the feature performance in default prediction. After investigating the characteristics of the new credit score model, we employed area ratio of ROC curves to compare the classification accuracy and evaluate how well this credit scoring perform. Finally, the data set was randomly divided into two groups, 65% for model training and the other 35% to apply multilayer perceptron of artificial neural network to the novel credit scoring model.

4.2 Results

The results are organized in two parts starting with Kruskal-Wallis statistics, w_k^1 and φ_i^1 calculation to establish the artificial neural network model based on propose methodology. Subsequently, the results are detailed, first in the terms of models comparison and then in the terms of computation performance.

The φ gives positive values to statistically significant features based on their detection power of default and gives others zero value, which means excluding those features from model. The φ_i^1 based on w_k^1 is shown below (Table 7). In this dataset, all features are statistically significant.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
K-W	1361	1191	1795	1539	1103	131	1923	556	983	136	669	419	1175
$\mathbf{w_k^1}$	0.994	1	0.995	0.995	1	1	1	0.986	0.992	0.945	0.988	0.981	0.993
$\boldsymbol{\varphi_i^1}$	0.0772	0.0776	0.0773	0.0772	0.0776	0.0776	0.0776	0.0766	0.0770	0.0734	0.0767	0.0762	0.0771

Table 7. K-W, w_k^1 and φ_i^1

From Table (8), A Kruskal-Wallis test for credit risk index (CRI) resulted in a p-value of less than 0.0001, which means that CRI is able to differentiate the categories of good and default loans.

CREDIT RISK INDEX	N	MEAN	MEDIAN	MEAN RANK	K-W
GOOD LOANS	4160	1.64	1.28	3301.09	44.35
DEFAULT LOANS	2217	1.27	0.29	2978.67	P-VALUE=.0

Table 8. Credit score validation

Furthermore, CRI in artificial neural network can also be a candidate for representing variables with ratio scale, that is, the reduction in the features dimension and efficient computation resource management. Therefore, two credit-scoring models are built with the 13 explanatory variables (see Section 4.1) in each feature groups normal and weighted separately, as well as a model with a combination of nominal variables group and CRI. These three models study the main effects of each transformation on variables in accuracy and computation efficiency of credit scoring. Model A combines unchanged variables to categorize loans into groups of default and non-default, model B includes all variables belong to bad loans to the power of w_k^1 except the nominal features and in model C we consider all nominal variables and CRI as features of the model but it did not provide more significant results than model B except of less computation.

As is common practice in credit scoring, statistical model performance is measured by the area under the receiver operating curve (AUC) and it is represented in table (9).

MODEL		PREDICTED NEGATIVE	PREDICTED POSITIVE	PERCENT CORRECT	AUC	TIME (MILLISECOND)
A	TRAINING NEGATIVE	2540	174	93.6%	0.926	840
	POSITIVE	401	938	70.1%		
	OVERALL PERCENT	72.6%	27.4%	85.8%		
	HOLDOUT NEGATIVE	1370	76	94.7%		
	POSITIVE	197	681	77.6%		
	OVERALL PERCENT	67.4%	32.6%	88.3%		
B	TRAINING NEGATIVE	2715	0	100%	0.999	320
	POSITIVE	0	1339	100%		
	OVERALL PERCENT	67.0%	33.0%	100%		
	HOLDOUT NEGATIVE	514	932	35.5%		
	POSITIVE	2	876	99.8%		
	OVERALL PERCENT	22.2%	77.8%	59.8%		
C	TRAINING NEGATIVE	2714	0	100%	0.999	80
	POSITIVE	0	1339	100%		
	OVERALL PERCENT	67.0%	33.0%	100%		
	HOLDOUT NEGATIVE	504	942	34.9%		
	POSITIVE	1	877	99.9%		
	OVERALL PERCENT	21.7%	78.3%	59.4%		

Table 9. Classification table and statistical model performance (AUC).

In this credit scoring case, error rates are not the appropriate criteria to evaluate the performance of credit score model because most clients are classified into creditable customers (93.6%). From table (5), it is clear that there is not a significant difference in the performance of the three models, but by considering the area under the ROC curve as an important factor of credit risk cost, the model A perform the worst, of which models including proposed methodology (models B, C) perform best. To offer a loan based on non-traditional data analysis, the benefit of correctly identifying a defaulter plays a prominent role and it is interesting to see that having only CRI and pre-calculation of weighted features on bad loans section of dataset allows discriminating potentially better clients.

Additionally, by using the area ratio in the validation data, the classification result shows almost the same performance of the models B and C; however model C yield a better performance in computation time that is an important factor in performance of parallel and distributed computing for non-traditional datasets.

5. IMPACT OF RESEARCH

This section identifies various levels of impact based on the research findings.

5.1 Confidentiality and privacy

Transferring the sensitive data from data warehouses of financial institutions to different machines and nodes for parallel or distributed computation is always affected by privacy concerns. Financial service providers try to enhance trust in their systems as a fundamental policy of client right and confidentiality of personally identifiable information is crucial. Additionally, there are some standards and regulations such as General Data Protection Regulation (GDPR) in European Union. The result of this study shows that an index of features can be calculated as an aggregation before data distributing for mapping stage of MapReduce algorithm. There is as well an ethical concern in data anonymization because of outliers, which mostly belong to special well-known customers. This indexing can guarantee the confidentiality of sensitive data to provide easier access to parallel computing tasks.

5.2 Financial inclusion

Models based on non-traditional data sources such as mobile phone data in the form of call-detail records or log files of mobile phones, which are examples of Big Data sources typically suffer from complexity and time-consuming sophisticated algorithms. Despite of facilitating credit access to people without historical financial data, the models should be highly accurate to fulfill the expectations of loan providers. Using conservative models for these new sources of data can help loan providers to offer even small credits to under-banked populations, young people and immigrants, enhancing them to assess whether the new clients are creditworthy.

5.3 Compliance risk impact

Committee on payment and settlement systems in key consideration 3-4-7 of principles for financial market infrastructures2 and its explanatory notes declares that the financial systems should have clearly defined procedures for the management of credit and liquidity risks. It should specify the respective responsibilities of the system operator and the participants and which provide appropriate incentives to manage and contain those risks. Credit-scoring models also should be considered in provisions and capital buffers calculation by financial institutions according to standards such as Basel Accords and IFRS9. In this research, we illustrated how credit scoring have to be conservatively formulated to propagate in a non-traditional datasets with potential high risk of false negative rate to detect default. This insight pave the way for loan providers to be able to use new sources of data in a more sound and solid way and try to adapt with new emerged technologies without risk management concerns.

6. CONCLUSION

This study introduces a novel credit scoring methodology that reselects significant highly informative features and weighted out by their level of contribution in predicting credit categories of loans to be used in modelling phase. This new credit scoring uses Kruskal-Wallis non-parametric test, which can be used for two or more

[2] An FMI should establish explicit rules and procedures that address fully any credit losses it may face as a result of any individual or combined default among its participants with respect to any of their obligations to the FMI.(2012)

categories. Therefore, categories could be “good loans” and “default loans” or even more than two categories such as “good”, “doubtful” and “bad” loans as is recommended by Basel Accords. The proposed credit risk index is computationally less expensive with reasonable accuracy in compare with current computationally expensive hybrid algorithms in credit scoring or fix weights models in score cards. In the classification accuracy, the results show that this credit scoring method is more informative and conservative. It is able to predict default probability by ANN with showing good performance. Therefore, it is suitable for non-traditional data sets such as mobile phone data which selecting and extracting the information of features in one aggregated measure is needed. ■

REFERENCES

Abellán J. & Castellano J. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 73, 1–10.

Arminger G., Enache D. & Bonne T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. Computational Statistics, 12(2), 293-310.

Avery R.B., Rostic R. W., Calem P.S. & Canner G. B. (2000). Credit scoring: Statistical issues and evidence from credit-bureau files. Real Estate Economics, 28(3), pp. 524-526, 529-539, 544.

Baesens B., Gestel T., Viaene S., Stepanova M., Suykens J. & Vanthienen J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit scoring. The Journal of the Operational Research Society, 54(6), 627–635.

Baesens B., Scheule H., Rosch D. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS, John Wiley & Sons.

Bellotti T. & Crook J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36, pp. 3302–3308.

Björkegren D. & Grissen D. (2017). Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. Netmob.

Butaru F., Chen Q., Clark B., Das S., Lo A. & Siddique A. (2016). Risk and risk management in the credit card industry. Journal of Banking and Finance, 72, 218-239.

Chamboko R. & Bravo J.M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. Risk Management, 18, 264–287.

Chamboko R. & Bravo J.M., forthcoming. Frailty correlated default on retail consumer loans in Zimbabwe. International Journal of Applied Decision Sciences.

Chamboko R. & Bravo J.M., forthcoming. Modelling and forecasting recurrent recovery events on consumer loans. International Journal of Applied Decision Sciences.

Chittaranjan G., Blom J. & Gatica-Perez (2012). D. Mining large-scale smartphone data for personality studies. In: Personal and Ubiquitous Computing.

Committee on payment and settlement system (2012). Principles for financial market infrastructures. Page 37.

Do T.M.T. & Gatica-Perez D. (2010). By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia.

Douzas & Bacao (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks”, Expert Systems with Applications.

Fensterstock A. (2005). Credit scoring and the next step. Business Credit 107(3), pp. 46-49.

Gaffney E., Kelly R. & McCann F. (2014). A transitions-based framework for estimating expected credit losses. Central Bank of Ireland - Financial Stability Division Research Technical Paper 16RT14.

García S., Fernández A., Luengo J., & Herrera F. (2010). Advanced nonparametric for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information sciences, pp. 180, 2044–2064.

Hand D. J. & Henley W. E. (1997). Statistical Classification Methods in Consumer Credit scoring: a Review. Journal of the Royal Statistical Society Series A (Statistics in Society), 160 (3), 523-541.

Jacobson T. & Roszback K. (2003). Bank lending policy, credit scoring and value-at-risk. Journal of Banking and Finance 27(4), pp. 615-633.

Jordan M. I. (2013). On Statistics, Computation and Scalability. Bernoulli, 19, 1378–1390.

Kellison B. & Brockett P. (2003). Check the score: credit scoring and insurance losses: Is there a connection", Texas Business Review Special Issue, pp. 1-6.
Kleiner A., Talwalkar A., Sarkar P. & Jordan M. I. (2014). A Scalable Bootstrap for Massive Data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 795–816.
Kruppa J., Schwarz A., Arminger G. & Ziegler A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125–5131.
Lessmann S., Baesens B., Seow H. & Thomas L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247 (1), 124-136.
Liang F., Cheng Y., Song Q., Park J., & Yang P. (2013). A Resampling-based Stochastic Approximation Method for Analysis of Large Geostatistical Data. Journal of the American Statistical Association, 108, 325–339.
Ma P., Mahoney M. W., & Yu B. (2013). A Statistical Perspective on Algorithmic Leveraging. arXiv preprint arXiv:1306.5362.
Maclaurin D. & Adams R. P. (2014). Firefly Monte Carlo: Exact MCMC with Subsets of Data. arXiv preprint arXiv:1403.5693.
Oskarsdottir M., Bravo C., Sarraute C., Vanthienen J., Baesens B. (2019). The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion Using Mobile Phone Data and Social Network Analytics. Applied Soft Computing Journal: 74, pp. 26-39.
Pedro J. S. (2015). MobiScore: Towards Universal Credit scoring from Mobile Phone Data. pp. 195-207.
Quittner J. (2003). Credit cards: sub-prime's tech dilemma: with delinquencies and charge-offs on the rise, the industry examines the role of automated decisioning. Bank Technology 16(1), pp. 19-23.
Rimmer J. (2005). Contemporary changes in credit scoring. Credit Control 26(4), pp. 56-60.
Schifano E. D., Wu J., Wang C., Yan J., & Chen M. H. (2015). Online Updating of Statistical Inference in the Big Data Setting. Technometrics, forthcoming.
Saeys Y, Inza I, Larrañaga P. (2007). WLD: review of feature selection techniques in bioinformatics. Bioinformatics. 23(19):2507–2517.
Skyler S., Eric M., Isaac M. & Felix K. (2017). Mobile–phone based Credit scoring. NetMob2017.
Song Q. & Liang F. (2014). A Split-and-merge Bayesian Variable Selection Approach for Ultrahigh Dimensional Regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
Thomas L.C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, International Journal Forecast. 16(2), pp. 149-172.
Verkasalo H. (2010). Analysis of users and non-users of smartphone application. Telematics and Informatics 27(3), p. 242–255.