Credit scoring in the age of Big Data -A State-of-the-Art

Article *in* International Journal of Computer Science and Information Security, · July 2017

CITATIONS
8

READS
1,676

3 authors:

Youssef Tounsi
University of Hassan II Casablanca
5 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Larbi Hassouni
University of Hassan II Casablanca
64 PUBLICATIONS **653** CITATIONS
SEE PROFILE

Houda Anoun
Polytechnical Institute of Lisbon
43 PUBLICATIONS **202** CITATIONS
SEE PROFILE

Credit scoring in the age of Big Data – A State-of-the-Art

Youssef TOUNSI, Larbi HASSOUNI, Houda ANOUN
RITM Laboratory, CED Engineering Sciences
Ecole Superieure de Technologie
Hassan II University of Casablanca, Morocco
tounsi@gmail.com, lhassouni@hotmail.com, houda.anoun@gmail.com

Abstract **- The banking sector has become very competitive, and increasingly sensitive to political and economic circumstances in each country and around the world. In addition to the traditional strategy which aims to reduce expenses and increase profits, many banks are looking for new methods to reduce credit risks, in order to improve their performance. In this sense, to resolve the most common problem, which is the lack of data, several researches are interested in using new data sources such as social networks which are used by all kind of users and more particularly by the young population. These new sources store data in large quantities and in a non-traditional format, hence the need to look for new methods of processing. Techniques of Big Data allow to store any voluminous amount of structured, semi-structured and unstructured data also providing many solutions to mine these data in order to extract relevant information.**
This paper has multiple goals. The first, which is the main, is to examine the role of big data in predicting the creditworthiness of consumers. The second is to explore the main machine learning methods used in credit scoring. The third is to investigate what type of data is relevant in determining consumer creditworthiness. And the last is to determine how various inherent characteristics of big data – volume, velocity, variety, variability and complexity – are related to the assessment of credit risk.
To address these topics, we have conducted a detailed and careful study of several researches works in the field. We envisage that our work can be of great usefulness to academics and professionals in the field of finance and especially to microfinance organizations whose main activity is to grant microcredits to people with limited incomes. These people are numerous in Morocco and often do not even have bank accounts.

Index Terms - Credit Scoring, Big Data, Machine Learning Techniques, Social data.

I. INTRODUCTION

It was in the banking sector that the credit risk assessment was first established in the mid-twentieth century, now some banks use more than one type of score [1] [2]: application (Credit) scoring, propensity score, behavioral scoring, Collection scoring, recovery score, attrition scores, fraud detection, etc.
Usual credit scoring are typically constructed using data extracted from traditional transactional systems such as OLTP (online transaction processing), Core Banking, ERP (enterprise resource planning), CRM (customer relationship management) applications and credit bureaus. The basic data, about consumers on which these systems operate, is very conventional, such as birthdate, gender, income, employment status, etc. All this data is precisely stored in relational databases or data warehouses. Nevertheless, this history is not rich enough in most countries even in developed regions, and especially for young customers, given this category tend to be riskier and they do not have sufficient credit history [3]. Therefore, other sources of data as "Big Data" is significantly necessary to bring value to the consumers' scoring systems performance.

One of the most important lessons learned from the recent global financial crises is that information technology and data architectures of financial institutions are inadequate to support the overall management of financial risks. In this context, the Basel Committee on Banking Supervision published in 2013 a set of principles under the name BCBS 239, the objective is to enable banks to improve their production capacities and improve the reliability of regulatory reporting. BCBS 239 mandates that banks adhere to a set of core principles for effective Risk Data Aggregation and Risk Reporting (RDARR) practices. As a result, systems integrators, big data firms and business consultants actively help banks prepare their processes and IT infrastructures for compliance. Improving risk assessment involves using more information to construct a complete and relevant client profile. Furthermore, the possibilities are explored in overcoming the hurdles encountered by credit scoring system through the use of Big Data.

The remainder of this article is organized as follows. We present some related work in Section II. In section III, we present the main classification methods in classical credit scoring. In section IV, we summarize principles and insights of big data for credit scoring. Then we present credit scoring using non-traditional data. Finally, in section V, we end with a conclusion and discuss possible future working directions in Section 5.

II. RELATED WORK

There are many researches works made to predict credit risk using wide-ranging computing [4], but we have to mention that there is not much academic literature covering big data for Credit Scoring [5] [6]. However, there is a number of papers analyzing big data analytics in banking sector in general terms [7] [8], and other works studying the social networking data from the sociological point of view. In fact, a few of available studies have mentioned the impact of offline and online customer's information on the credit scoring. In addition, no paper presents the state of the art of the credit scoring based on massive volume data.

A. Masyutin [6] clarifies that credit history is not rich enough for young clients but the social data using big data can bring value to the scoring systems performance. More, Y. Wei, P. Yildirim, C. Van and C. Dellarocas [10], highlight that credit

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

scoring using social network data can reduce lenders' misgivings about engaging applicants with limited personal financial history. Further, G. Guo, F. Zhu, E. Chen, Q. Liu, L. Wu and C. Guan [3], illuminate that quantitative investigation of the extracted features shows that online social media data does have good potential in discriminating good credit users from bad.

Likewise, Y. Yang, J. Gu and Z. Zhou [11], analyze the opinions about some enterprises transmitted through social media to predict their future credit risk. Additionally, D. Ntwiga and P. Weke [12], present the limitations of the traditional consumer lending models due to the use of historical data, and checking the benefits that could arise by incorporating social media data in credit scoring process for consumer lending. In another empirical analysis, Y. Zhanga, H. Jiaa, Y. Diaoa, M. Haia and H. Lia [13], construct a credit scoring model in the case of online Peer-to-Peer (P2P) lending, by fusing social media information based on decision tree, their result shows that their model has good classification accuracy.

Moreover, D. Björkegrena and D. Grissenb [14], establish a method to predict default among borrowers without formal financial histories, using behavioral patterns revealed by mobile phone usage. Additionally, N. Kshetri [4], indicates that, the main reason why low-income families and micro-enterprises in emerging economies such as China, lack access to financial services is not because creditworthiness does not exist but merely because banks and financial institutions lack data. Reading-through another case, M. Hurley and J. Adebayo [9] explore the problems posed by big data credit scoring tools and analyze the gaps in existing laws of the United States of America.

The amount of data is exploding at a remarkable rate because of developments in web technologies, social media and other activities generated data (mobile device, log file, IoT, etc.). The conclusion is that credit risk assessment can greatly benefit from using non-traditional information with big data. Nevertheless, traditional approaches are struggling when faced with these massive data [15].

Many studies have tried to address the challenges and opportunities of big data in general terms. E.Fortuny, D.Martens and F. Provost [16], provide a clear illustration that larger data indeed can be more valuable assets for predictive analytics. This implies that institutions with larger data assets—plus the skill to take advantage of them—potentially can obtain substantial competitive advantage over institutions without such access or skill. Furthermore, L. Wang, C. Alexander [17], establish a comparison of several machine learning algorithms and an evaluation of big data technologies. Additionally, A. L'Heureux, K. Grolinger, H. ElYamany and M. Capretz [18], address machine learning challenges in the era of big data with the ultimate objective of helping practitioners select appropriate solutions for their use cases. Limited existing studies have mentioned the effect of using big data for credit scoring. Our contribution consists in reviewing these hurdles and their solutions in the case of credit scoring (Section 4).

In view of these elements, which credit scoring models will benefit most from the advantages of the dig data in terms of performance? In addition, what are the non-traditional data that are most relevant for improving credit rating? These are the questions we attempt to answer in this paper.

III. The main Machine Learning methods used in Credit Scoring

Machine learning (ML) is continuously unleashing its power in a wide range of applications. Credit scoring using predictive analytics techniques, is one of these solicitations. ML has been pushed to the front line in late years partly due to the rise of big data. It can be broadly categorized based on two factors (Figure 1):

- Learning types: This is to do with what type of response variable (supervised, unsupervised, semi-supervised, and reinforcement learning).
- Subjective grouping: This grouping is driven by "what" the model is trying to achieve.

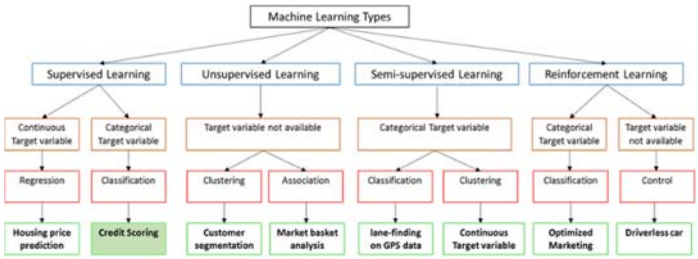

Fig. 1 Machine Learning Categories.

The supervised learning algorithms are a subcategory of the family of machine learning algorithms, which are mainly used in predictive modeling. These algorithms try to model relationships between the target prediction output and the input features based on those independencies that it learned from the historical or previous data sets. The main types of supervised learning algorithms are:

- Regression: the output variable takes continuous values.
- Classification: the output variable takes class labels.

The aim of the credit scoring model is to perform a classification: To distinguish the "good" applicants from the "bad" ones. In practice this means the statistical models is required to find the separating line distinguishing the two categories, in the space of the explanatory variables (age, salary, education, etc.). For the banking industry, the scoring methodology has played an important role in developing internal rating systems. There are several reasons that can explain the widespread use of scoring models: First, since credit scoring model are established upon statistical models and not on opinions, it offers an objective way to measure and manage risk. Second, the statistical model used to produce credit scores can be validated. Data have to be carefully used to check the accuracy and performance of the predictions. Third, the statistical models used by credit scores can be improved over time as additional data are collected.

The scoring system are made up of three major parts:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Problem Definition: This initial phase project focuses on understanding the project objectives and requirements.
- Data Gathering and Preparation: The data understanding involves internal and external data collection and storage in traditional systems (Relational databases, data warehouses, etc) or big data platform, selection variables, data cleansing and data transformation if required, also data exploration.
- Model Building and Evaluation: Splitting data into training and test sets, selecting algorithm, tuning algorithm, building Model using training data and Evaluating the model using test data (Figure 2)

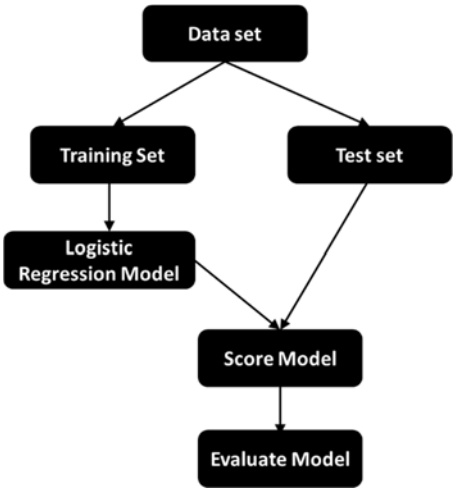

Fig. 2 Model building Steps (Logistic Regression e.g.)

There are many evaluation measurements of predictive performance of models in the area of credit scoring. These measurements are: ROC Curves, average accuracy, Type I and Type II errors.

The ROC (Receiver Operating Characteristics) curve was first applied to assess how well radar equipment in WWII distinguished random interference or "noise" from the signals that were truly indicative of enemy planes (Swets, et al., 2000). The ROC curve plots the sensitivity or "hits" (e.g., true positives) of a model on the vertical axis against 1-specificity or "false alarms" (e.g., false positives) on the horizontal axis. The area under the ROC curve is a convenient way to compare different predictive binary.

TABLE I
CONFUSION MATRIX FOR CREDIT SCORING

	Predicted class (%)	
Actual class (%)	Good loans	Bad loans
Good loans	TP	FN (Type II error)
Bad loans	FP (Type I error)	TN

From the confusion matrix table the following calculations are defined:

Average Accuracy (ACC) = TP+TN/TP+FN+TN+FP (1)

Type I error = FP/ TN+FP (2)

Type II error = FN/ TP+FN (3)

A TP stand for good applicant correctly classified as good, TN stands for bad applicant correctly classified as bad, FN (Type II) stands for good applicant incorrectly classified as bad customer and FP (Type I) stands for Bad customer incorrectly classified as Good customer (high risk).

There are several statistical methods for building and estimating scoring models, including linear regression models, logit models, probit models, and neural networks. We introduce them in the remainder of this section.

Linear regression (LR)

Linear regression is a supervised machine learning technique to identify the linear relationship between target variables and explanatory variables. A general linear regression problem can be explained by assuming some dependent or response variable yi which is influenced by inputs or independent variables xi1, xi2,....,xiq. A regression model can express this relation:

$$yi= \beta 1\ xi1 + \beta 2\ xi2 + ...+ \beta q\ xiq + \varepsilon \quad (4)$$

Where $\beta 1, \beta 2,..., \beta q$ are fixed regression parameters and ε is a random error or noise parameter. To get a more accurate prediction, we need to reduce this error term as soon as possible. Perhaps the most primitive classification method is formally introduced by Ronald Fisher in 1936 [19]. Orgler (1970) used regression analysis in credit scoring for commercial loans. The use of regression analysis extended such applications to include further aspects (Lucas, 1992; Henley, 1995; Hand & Henley, 1997; Hand & Jacka, 1998). Furthermore, other authors have been studying linear regression models or its generalizations in credit scoring (Hand and Kelly, 2002; Banasik et al., 2003; Karlis and Rahmouni, 2007; Efromovich, 2010) [20]. Besides and in the age of big data, S. Jun, S. Lee and J. Ryu (2015), propose a new analytical methodology for big data analysis in linear regression problem for reducing the computing burden [21].

Discriminant analysis (DA)

Discriminant analysis is a credit scoring technique established to discriminate between two groups. It is commonly agreed that the discriminant method is still one of the most generally industrialized techniques to classify customers as good credit or bad credit.

Since the crisis of 1929-1933, more and more academics and practitioners have studied this phenomenon of bankruptcy prediction, developing or using DA model (Fisher, 1936; 1986; Altman, 1968; 1977; 2000; Edmister, 1972; Deakin, 1977; Conan and Holder, 1979; Beaver, 1996; West, 2000; Anghel, 2002; Gestel et al., 2006; Yang, 2007; Falangis and Glen, 2010; Dinca and Gidinceanu, 2011; Armeanu et al., 2012; Akkoc, 2012) [19] [22].

Logistic regression (LG)

Logistic regression proposed by Berkson (1944), LG is a classification model that, in the specific context of binary

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

classification, estimates the posterior probability of the positive class, as the logistic sigmoid of a linear function of the feature vector [22]. It can be implemented using logistic functions:
To predict the log odds ratios, use the following formula:

$$\text{logit}(p) = \beta 0 + \beta 1 \times x1 + \beta 2 \times x2 + \ldots + \beta n \times xn \quad (5)$$

The probability formula is as follows:

$$p = e^{\text{logit}(p)} / 1 + e^{\text{logit}(p)} \quad (6)$$

logit(p) is a linear function of the explanatory variable, X (x1,x2,x3..xn), which is similar to linear regression. So, the output of this function will be in the range 0 to 1. Based on the probability score, we can set its probability range from 0 to 1. In a majority of the cases, if the score is greater than a threshold (e.g. 0.5), it will be considered as 1, otherwise 0. Also, we can say it provides a classification boundary to classify the outcome variable. Logistic regression has been largely used in credit scoring applications [23]. The major benefit of this method is that it can generate a simple probabilistic formula for classification. LG is compared with other credit scoring techniques (Li and Hand, 2002; Hand, 2005; Lee and Chen, 2005; Abdou et al., 2008; Yap et al., 2011; Pavlidis et al., 2012) [19].

Decision trees (DT)

Decision tree is one of the most broadly used classification and prediction method of machine learning. DT can deal with both numerical data and categorical data (such as gender), so it is very applicable for personal credit rating [13]. A decision tree can be defined as a tree in which each branch node specifies a choice between a number of alternatives, and each leaf node distinguishes a classification or decision. Most algorithms such as ID3, C4.5, and CART for decision tree induction follow a greedy, top-down recursive divide-and-conquer approach, which starts with a training set of tuples and their associated class labels [24]. Kao et al. (2012) proposes a combination of a Bayesian behavior scoring model and a CART-based credit scoring model. Other possible and particular methods of decision trees are C4.5 decision trees algorithm and J4.8 decision trees algorithm.

Artificial Neural networks (ANNs)

ANNs are inspired by the functionality of the nerve cells in the brain. Just like humans, ANNs can learn to recognize patterns by repeated exposure to many different examples. They are non-linear models that can classify based on pattern recognition capabilities [25]. A neural network is a highly interconnected network with abundant neurons and mutual links between them. NN has several layers consisted of neurons having similar characteristics. The neurons in one layer are connected with those in contiguous layers. The value of the connection between two neurons in different layers is called 'weight' (Figure 3).

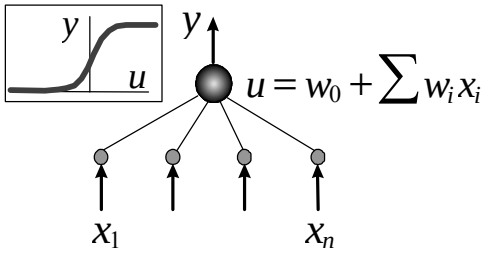

Fig. 3 Activation functions in use with neural networks.

$$y = f\left(\sum_{j=1}^{d} w_j x_j + w_0\right) \equiv f\left(\sum_{j=0}^{d} w_j x_j\right) \quad (7)$$

More recently, different artificial neural networks have been suggested to tackle the credit scoring problem.
In some datasets, the neural networks have the highest average correct classification rate when compared with other traditional techniques, such as discriminant analysis and logistic regression, taking into account the fact that results were very close (Abdou et al., 2008).

Deep Learning (DL)

Since 2006, deep structured learning, or more commonly called deep learning or hierarchical learning, has emerged as a new area of machine learning research [26]. Deep learning is a generic term for multilayer neural networks. Multilayer neural networks decrease the overall calculation time by performing calculation on hidden layers. Thus, they were prone to excessive over training, as an intermediate layer was often used for approximately every single layer.
Deep learning provides training stability, generalization, and scalability with big data. Deep Learning is quickly becoming the algorithm of choice for the highest predictive accuracy [27]. Niimi (2015) conduct Credit Card Data Analysis using deep Learning and confirm that deep learning has the same accuracy as the Gaussian kernel SVM using German Data set (the most used public data set in credit scoring) [28].

Support vector machine (SVM)

Support vector machine is described as a technique of classification that was first introduced by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. The key benefits of this model are based on the nonparametric case. The main idea of the Support Vector Machines [29] algorithm is that given a set of points which belong to one of the two classes, it is needed an optimal way to separate the two classes by a hyperplane as seen in the below figure. This is done by:

- maximizing the distance (from closest points) of either class to the separating hyperplane (W=Gap)
- minimizing the risk of misclassifying the training samples and the unseen test samples.

$$\min_{\vec{w},b} \frac{1}{2}\|w\|, \quad (8)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$s.t \quad y_i = +1 \Rightarrow \vec{w} \cdot \vec{x}_i + b \geq +1$$
$$y_i = -1 \Rightarrow \vec{w} \cdot \vec{x}_i - b \leq -1 \qquad (9)$$
$$s.t\ y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \quad \forall i$$

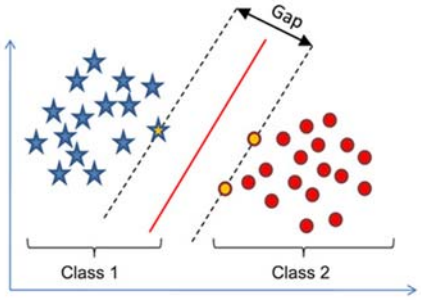

Fig. 4 Support Vector Machines.

The identification of the each data point xi is yi, which can take a value of +1 or -1 (representing positive or negative respectively).

Depending on the way the given points are separated into the two available classes, the SVMs can be linear SVM or Non-Linear SVM. Lately, support vector machine was used in credit scoring (Chen et al., 2009; Li et al., 2006; Gestel et al., 2006; Xiao and Fei, 2006; Yang, 2007; Chuang and Lin, 2009; Zhou et al., 2009; 2010; Feng et al., 2010; Hens and Tiwari, 2012; Ling et al., 2012).

Fuzzy logic (FUZZY)

Zadeh (1965) introduced the Fuzzy Logic as a mathematical system, which there are not just two alternatives but a whole continuum of truth values for logical propositions. Unlike the binary logic, fuzzy logic uses the notion of membership to handle the imprecise information. Several fuzzy-based approaches have been suggested to evaluate credit worthiness. Zimmermann and Zysno (1980) used fuzzy operators to aggregate evaluation results from a four-level hierarchy of criteria. Using a fuzzy connectionist model, Romaniuk and Hall established an expert system, FUZZNET system, to acquire a knowledge base for classifying the credit worthiness of loan applicants [30]. Furthermore, Hoffman et al. (2002) proposed a genetic fuzzy for credit scoring and compared it with neuro-fuzzy algorithm NefClass [31]. Other authors have been studying Fuzzy logic models or its generalizations in credit scoring (Lahsasna et al., 2010; Nosratabadi, Nadali and Pourdarab, 2012; Bazmara and Donighi, 2013; Duc and Thien, 2013) [32] [33]. Possible methods in fuzzy logic are regularized adaptive network based Fuzzy inference systems and fuzzy Adaptive Resonance [19].

Genetic Algorithms (GA)

GAs try and replicate the natural selection process where genes are passed from one generation to the next generation. The theory of GAs was developed in the late 1960s and early 1970s by John Holland and his associates as a means to study evolutionary processes in nature (Holland, 1975) [34]. The use of GAs is now growing rapidly with successful applications in finance trading, fraud detection and other areas of credit risk. Yobas et al. (2000) compared the predictive performances of four techniques, one of which is GAs, GA faired quite well coming in second place [19]. Genetic programming (GP) is one kind of evolutionary algorithms and can be considered as an extension of genetic algorithms, which employ a complication representation formal to code individuals. In 2002, Koza proposed that GP can be employed in the credit scoring problems. GA is usually used to attribute reduction in the credit scoring problem, GA is frequently used to attribute reduction in the credit scoring problem. Zhang et al. (2008) used a GA to reduce the attributes before the classifier [35]. In the age of big data, we can conclude that GA is very suitable for solving problems with higher dimensions.

Bayesian networks (BN)

Bayesian networks (BN). A Bayesian classifier (Friedman et al., 1997) is based on calculating a posterior probability of each observation belongs to a specific class. The Bayesian methodology is built upon the well known Bayes' Rule, which is itself derived from the fundamental rule for probability calculus.

$$P(a,b) = P(a \mid b) * P(b) \qquad (10)$$

P(a,b) is the joint probability of both events a and b occurring, P(a|b) is the conditional probability of event a occurring given that event b occurred, and P(b) is the probability of event b occurring.

$$P(b \mid a) = \frac{P(a \mid b) * P(b)}{P(a)} \qquad (11)$$

The Bayesian network is a probabilistic model that represents a set of random variables and their conditional dependencies via a direct acyclic graph. The most remarkable characteristic of BN is the capability to encode both quantitative and qualitative knowledge. This means, we can rely in this model on statistical data analysis and experience of domain experts as well. Sometimes referred to as causal networks, Bayesian models are constructed as graph models, where nodes are used to encode parameter characteristics (descriptive variables), and directional links between them encode often complex correlations, usually of causal nature. Several academics have studied Bayesian Credit Scoring Models (Whittaker, 1990; Sewart and Whittaker, 1998; Hand et al., 1997; Chang et al., 2000; Baesens et al., 2001; Gemela, 2001; Zhu et al., 2002; Abramowicz et al., 2003; Thomas et al., 2005; Antonakis and Sfakianakis, 2009; Bier et al., 2010; Wu, 2011; Hand and Adams; 2014). Possible methods in Bayesian networks are naive Bayes, tree augmented naive Bayes and Gaussian naive Bayes [19] [36].

Hybrid methods

Hybrid methods combine different techniques to improve the performance capability. There are four different hybrid methods: Classification+Clustering, Clustering+Classification, Classification+ Classification and Clustering + Clustering (Figure 5):

Fig. 5 Hybrid methods.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The first classifier/Cluster can be trained in order to classify good and bad clients, detect and filter outlier or data reduction. The second uses the output of the first classifier conducive to perform and provide better results [37]. Several authors have been studying Hybrid approaches (Lee et al., 2002; Huang et al., 2007; Lee and Chen, 2005; Hsieh, 2005; Huysmans et al., 2006; Shi , 2009; Chen et al., 2009; Liu et al., 2010; Ping and Yongheng, 2011; Capotorti and Barbanera, 2012; Vukovic et al., 2012; Akkoc, 2012; Pavlidis et al., 2012) [1] [19] [36].

Ensemble methods

An ensemble method is a set of classifiers that learn a target function, and their individual predictions are combined to classify new examples. Ensembles generally improve the generalization performance of a set of classifiers on a domain [17]. Some of the most frequently used methods are: bagging (Breiman, 1996), boosting (Schapire, 1990), stacking (Wolpert, 1992) and random forest (Breiman and Cutler, 2001). Many authors have used these combined approaches in credit scoring problems (Homann et al., 2007; Hsieh and Hung, 2010; Paleologo et al., 2010; Zhang 10 et al., 2010; Finlay, 2011; Louzada et al., 2011; Xiao et al., 2012; Marques et al., 2012; Wang, Ma, Huang and Xu, 2012; Malekipirbazari and Aksakalli, 2015) [17] [19] [38] [39] [40] [41].

IV. UNDERSTANDING BIG DATA AND THEIR CHALLENGES FOR CREDIT SCORING

A. What Is Big Data?

The volume of data which needs storage every day is increasing exponentially. It is now possible to acquire these vast amounts of information on low cost platforms such as Hadoop. The "big data" phenomenon brings challenge to empower predictive methods for credit scoring. Indeed, Big Data has to deal with large and complex datasets that can be structured, semi-structured, or unstructured and will typically not typically fit into memory to be processed.

Doug Laney[41] was the first one in talking about 3 V's in Big Data management (Figure 6):

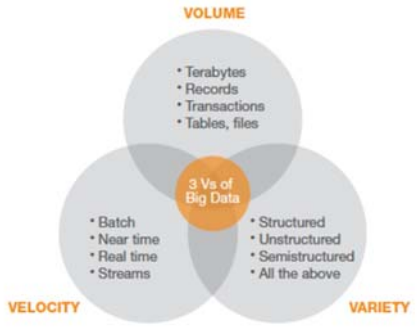

Fig. 6 Three V's Big Data.

First, the volume at which new data is being generated is surprising, for this reason, this "V" is the most associated with Big Data. We live in an age when the amount of data we expect to be generated in the world is measured in exabytes and zettabytes. As the database grows, applications and architecture designed to support data need to be re-evaluated quite often. Furthermore, the amount of data stored by the financial institutions is rapidly increasing and provides the opportunity for them to conduct predictive analytics and enhance its businesses. However, data scientists are facing large challenges, handling the massive amount of data efficiently and generating insights with real business value.

Secondly, Data velocity measures the speed of data creation, streaming, and aggregation. eCommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue, it is also an ingest issue (extract-transform-load)[43].

Then, data variety is the challenge of having disparate data sets, from different sources in different formats, all in silos – text, images video, audio, etc. Thereby neglecting the benefits a unified view of data brings, from an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl [44].

Nowadays, 15 characteristics are defined by some professionals and academics (Doug Laney, SAS, Oracle, Oguntimilehin A and Data Science Central) [42]: Volume, Velocity, Variety, Value, Veracity, Validity, Volatility, Visualization, Virality, Viscosity, Variability, Venue, Vocabulary, Vagueness and Complexity.

The methodologies and frameworks behind the Big Data concept have become very common in wide number of research and industrial areas. This subsection introduces Hadoop and Spark.

B. Apache Hadoop

A Hadoop distribution is made of number of separate frameworks that are designed to work together. The frameworks are extensible as well as the Hadoop framework platform. Hadoop has evolved to support fast data as well as big data. Big data was initially about large batch processing of data. Now banks also need to make lending decisions in real time or near real time as the data arrives. Fast data involves the capability to act on the data as it arrives. Hadoop's flexible framework architecture supports the processing of data with different run-time characteristics. The Hadoop framework platform includes these principals' modules (Figure 7):

- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.
- Hadoop Common: The common utilities that support the other Hadoop modules.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

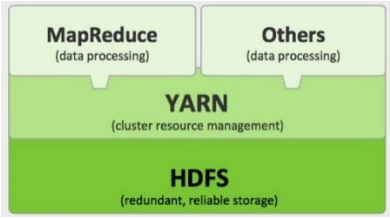

Fig. 7 Hadoop framework.

Hadoop Distributed File System (HDFS) is a file system that provides reliable data storage and access across all the nodes in a Hadoop cluster (Figure 8). It links together the file systems on many local nodes to create a single file system. Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout various nodes in the cluster. This way, the map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability that is needed for big data processing. This powerful feature is made possible through the HDFS of Hadoop.

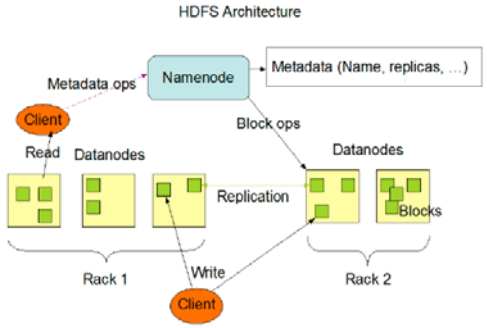

Fig. 8 Hadoop Distributed File System.

MapReduce is a programming framework of Hadoop suitable for writing applications that process large amounts of structured and unstructured data in parallel across a cluster of thousands of machines, in a reliable, fault-tolerant manner. MapReduce is the heart of Hadoop. This programming paradigm allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The idea behind this programming model is to design map functions (or mappers) that are used to generate a set of intermediate key/value pairs, after which the reduce functions will merge(reduce can be used as a shuffling or combining function) all of the intermediate values that are associated with the same intermediate key. The key aspect of the MapReduce algorithm is that if every map and reduce is independent of all other ongoing maps and reduces, and then the operations can be run in parallel on different keys and lists of data. Although three functions, Map(), Shuffling(), and Re-duce(), are the basic processes in any MapReduce approach (Figure 9).

- Map Step: Each worker node applies the Map() function to the local data and writes the output to a temporary storage space. The Map() code is run exactly once for each key value Mapping, generating output that is organized by key values Shuffling. A master node arranges it so that for redundant copies of input data only one is processed.
- Shuffle Step: Data belonging to one key is redistributed to one worker, such that each worker node contains data related to one key only.
- Reduce Step: Each worker node now processes data belonging to same key in parallel.

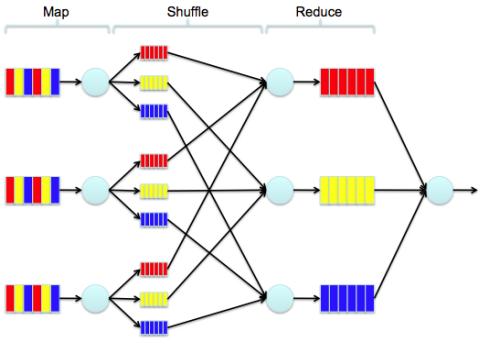

Fig. 9 MapReduce framework.

YARN is a cluster management technology. It is one of the key features in second-generation Hadoop. It is the next-generation MapReduce, which assigns CPU, memory and storage to applications running on a Hadoop cluster. It enables application frameworks other than MapReduce to run on Hadoop, opening up a wealth of possibilities. Part of the core Hadoop project, YARN is the architectural center of Hadoop that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform.

C. Apache Spark

Spark is designed to run on top of Hadoop and it is an alternative to the traditional batch map and reduce model that can be used for real time stream data processing and fast queries that finish work within a seconds. In addition to Map and Reduce functions spark take supports of SQL queries, streaming data, machine learning and graph data processing for big data analysis. The spark platform is implemented in Scala and is hence run within the Java Virtual Machine (JVM). In addition to a Scala API, interfaces in Java and Python are available. Spark provides two options for running applications. Firstly, interpreter in the Scala language distribution allows users to run queries on large data sets through Spark engine. Secondly applications can be written as Scala programs called driver programs and can be submitted to the cluster's master node after compilation [45].

Apache spark consists of a driver program (SparkContext), workers also called executors, cluster manager, and the HDFS. Driver program is the main program of Spark. Spark applications run as independent sets of processes on a cluster, coordinated by the Spark Context object called the driver program. Whereas each application gets its own processes and run tasks in multiple threads and must be network addressable from worker nodes. Once connected, Spark acquires executors

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

on nodes in the cluster, which are worker processes that run computations and store data for your application. Next, it sends your application code (defined by JAR or Python files passed to SparkContext) to the executors. Finally, SparkContext sends tasks for the executors to run (Figure 10).

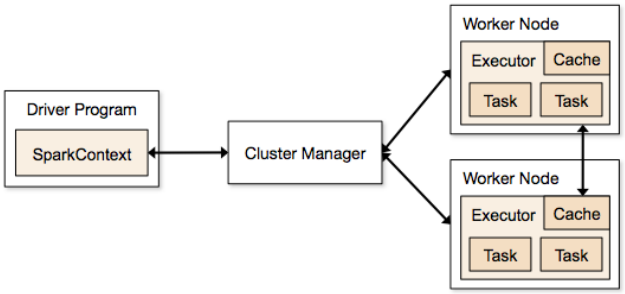

Fig. 10 Driver program (SparkContext).

Apache Spark is based on two key concepts: Resilient Distributed Datasets (RDD) and directed acyclic graph (DAG) execution engine. With regard to datasets, Spark supports two types of RDDs: parallelized collections are based on Scala collections and Hadoop datasets that are created from the files which are stored on HDFS.

Spark jobs perform work on Resilient Distributed Datasets (RDD), an abstraction for a collection of elements that can be operated on in parallel (Figure 11). When Spark is running on a Hadoop cluster, RDDs are created from files in the distributed file system in any format such as text and sequence files or anything supported by a Hadoop forma.

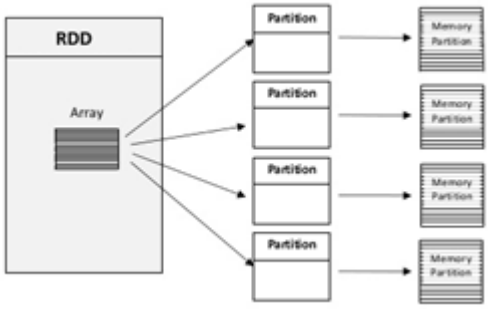

Fig. 11 Resilient Distributed Datasets.

A resilient distributed dataset (RDD) is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. The elements of an RDD need not exist in physical storage. RDDs support two kinds of operations: transformations and actions [45].

TABLE II
HADOOP VS SPARK

Aspects	Hadoop	Spark
Difficulty	MapReduce is difficult to program and needs abstractions.	Spark is easy to program and does not require any abstractions
Interactive Mode	There is no in-built interactive mode.	It has interactive mode.
Streaming	Hadoop MapReduce just get to process a batch of large stored data.	Spark can be used to modify in real time through Spark Streaming.
Performance	MapReduce does not leverage the memory of the Hadoop cluster to the maximum.	Spark has been said to execute batch processing jobs about 10 to 100 times faster than Hadoop MapReduce.
Latency	MapReduce is disk oriented completely.	Spark ensures lower latency computations by caching the partial results across its memory of distributed workers.
Ease of coding	Writing Hadoop MapReduce pipelines is complex and lengthy process.	Writing Spark code is always more compact.
Supported languages	Java	Java, Python, R, Scala

D. Big data challenges in credit scoring

Supported by the increase of "Big Data" platforms such as Apache Hadoop and Spark, banks are collecting and analyzing ever larger datasets. Thus, the Big Data presents opportunities and challenges for credit scoring as application of the predictive modeling [44]. First the topic with volume, many studies indeed continue to see increasing predictive performance from more data as the datasets become massive [44] [45] etc. Another benefit of big data to machine learning lies in the fact with more samples available for learning, the risk of overfitting becomes smaller. The subject with variety, absolutely, interesting. A fast flood of unstructured data, such as social media, image, audio, video, in addition to the structured data, is providing novel overtures for credit scoring especially for a population whose data are missing. Velocity implies data are streaming at rates faster than that can be handled by traditional systems. Veracity suggests that despite the data being available, the quality of data is still a major concern. On the other hand, there are several bottlenecks in designing credit scoring system based on the big data such as [46]:

- Time complexity: Finishing the computation within acceptable time on a single computer is very hard.
- Memory restrictions: It is difficult to keep the completely training data set or most of it in memory on one computer.

Indeed, SVM for example is extremely powerful and widely accepted classifier in the field risk assessment due to its better generalization capability. However, SVM is not suitable for large scale dataset due to its high computational complexity [47].

The following table presents a summary of the challenges and some proposed solutions for credit scoring with big data [18]:

TABLE III
BIG DATA CHALLENGE AND SOME PROPOSED SOLUTIONS

Challenges	Examples / explanation	Some proposed solutions
Volume		
Processing Performance	• SVM algorithm has a training time complexity of $O(m^3)$ and a space complexity of $O(m^2)$	Resilient distributed datasets (RDDs)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Challenges	Examples / explanation	Some proposed solutions
	• Logistic regression $O(mn^2+n^3)$ • Linear regression $O(mn^2+n^3)$ • Gaussian discriminative analysis $O(mn^2+n^3)$ *m* is the number of training samples n the number of features	
Curse of Modularity	Many learning algorithms rely on the assumption that the data being processed can be held entirely in memory or in a single file on a disk	
Class Imbalance	• This problem is especially prominent when some classes are represented by a large number of samples and some by very few. • The performance of a machine learning algorithm can be negatively affected when large datasets contain data from classes with various probabilities of occurrence.	Japkowicz and Stephen showed that decision trees, neural networks, and support vector machine algorithms are all very sensitive to class imbalance.
Curse of Dimensionality	• Difficulties encountered when working in high dimensional space. Specifically, the dimensionality describes the number of features or attributes present in the dataset. • the time complexity of the principal component analysis is $O(mn^2+n^3)$ • logistic regression $O(mn^2+n^3)$,	Feature selection and feature Engineering
Feature selection and feature Engineering	• feature selection (dimensionality reduction) aims to select the most relevant features • the selection of the most appropriate features is one of the most time consuming pre-processing tasks in machine learning • As the dataset grows, both vertically and horizontally, it becomes more difficult to create new, highly relevant features.	
Non-Linearity	• the correlation coefficient is often cited as a good indicator of the strength of the relationship between two or more variables • This problem is not exclusive to Big Data, non-linearity can be expected to be more prominent in large datasets. • In the case of Big Data, the large number of points often creates a large cloud, making it difficult to observe relationships and assess linearity.	
Generalization error	• Generalization error can be broken down into two components: variance and bias • Variance describes the consistency of a learner's ability to predict random things • Bias describes the ability of a learner to learn the wrong thing • As the volume of data increases, the learner may become too closely biased to the training set and may be unable to generalize adequately for new data	
Variety		
Data locality	• It is difficult to keep the whole training data set or most of it in memory on one computer.	MapReduce-based approaches encounter difficulties when working with highly iterative algorithms.
Data Heterogeneity	• Machine learning approaches were not developed to handle semantically diverse data types, file formats, data encoding, data model, and similar. • The business value of data analytics typically involves correlating diverse datasets, and integration is crucial for carrying out machine learning over such datasets.	
Velocity		
Data Availability	• Incremental learning is a relatively old concept, it is still an active research area due to the difficulty of adapting some algorithms to continuously arriving data	
Real-Time Processing/Streaming	• Traditional machine learning approaches are not designed to handle constant streams of data • The business value of real-time processing systems lies in their ability to provide instantaneous reaction	
Concept Drift	• Big Data are non-stationary; new data are arriving continuously • Machine learning models are built using older data that no longer reflect the distribution of new data accurately. • The challenges typically lie in quickly detecting when	

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Challenges	Examples / explanation	Some proposed solutions
	concept drift is occurring and effectively handling the model transition during these changes.	
Veracity		
Data Provenance	• Data provenance is the process of tracing and recording the origin of data and their movements between locations • The provenance dataset itself becomes too large, therefore, while these data provide excellent context to machine learning, the volume of these metadata creates its own set of challenges. • Not only is this dataset too large, but the computational cost of carrying this overhead becomes overwhelming.	Reduce and Map Provenance (RAMP) developed for MapReduce
Data Uncertainty	• For example, sentiment data are being collected through social media, but although these data are highly important because they contain precious insights into subjective information, the data themselves are imprecise. • Machine learning algorithms are not designed to handle this kind of imprecise data, thus resulting in another set of unique challenges for machine learning with Big Data.	
Dirty and Noisy Data	• Noisy data contain various types of measurement errors, outliers, and missing values. • From the machine learning perspective this is different from imprecise data; having an unclear picture is different from having the wrong picture • Noisy data one of the three main challenges of Big Data analysis in addition to multiple sources and Dependent data challenge (the samples are dependent with relatively weak signals).	

In conclusion, Banks with bigger data assets may have an important strategic advantage over their smaller competitors if they can exceed the above bottlenecks.

V. CREDIT SCORING USING NON-TRADITIONAL DATA

The Exponential growth of using the social networks like Facebook, YouTube, Instagram, Twitter and Weibo, as we have witnessed during the past few years, generally, the total number of people using Social Media continues to rise and therefore online user footprints are accumulating rapidly on the social web. However, compared with traditional financial data, diverse social data presents both opportunities and challenges for Credit scoring [3].

Facebook was the first social network to surpass 1 billion registered accounts and currently sits at 1.97 billion monthly active users (See Figure 12 according to www.statista.com).

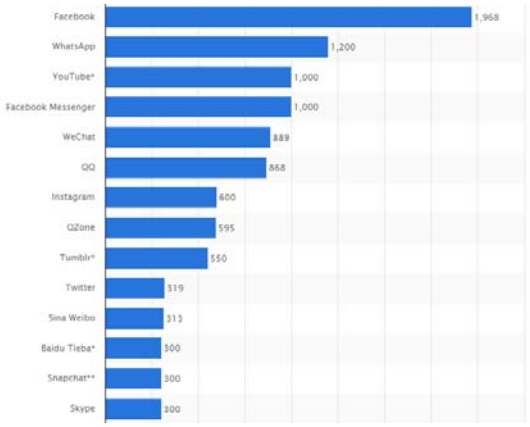

Fig. 12 Social network sites ranked by number of active users (in millions/April 2017).

With developments in internet connectivity alongside rises in smartphone adoption, social media platforms have been able to enrich the data they collect. This has particular relevance to the variety and frequency aspects of data value. Geographic data logging across 4G, for example, supports platforms to identify a user's routine movements in real time. Social data can be retrieved from any point where a user has a traceable interaction with an accessible technology. Therefore, social data analysis could enable the banks to establish a credit scoring for individuals with no credit history and to identify relevant products for specific individuals.

Indeed, we are addressing various non-traditional data for credit scoring: technical information, public databases, web searches, location, session tracking, social networks, mobile data, browser data, telecom data, e-commerce, financials transactions, etc. The table below presents the comparison of the analyzed papers, which addresses the credit scoring using non-traditional data.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE IV
NON-TRADITIONAL DATA

Autors	Type loan / ML method	Social network / Region	Data
A. Masyutin [6]	Personal Credit / Logistic Regression	Vkontakte / Russia	Age, Gender, Marital status, Number of days since last visit, Number of subscriptions, Number of days since the first post, Number of user's posts with photos, Number of user's posts with video, Number of children, Major things in life and Major qualities in people
Y. Wei, P. Yildirim, C. Van and C. Dellarocas [10]	Personal Credit		Tie among customers
G. Guo, F. Zhu, E. Chen, Q. Liu, L. Wu and C. Guan [3]	Personal Credit / Decision Tree, Naive Bayes, Logistic Regression and SVM	Weibo / China	Demographic features Tweet features Network features High-Level features
Y. Yang, J. Gu and Z. Zhou [11]	Enterprise Credit / Logistic and Probit Regression	Hexun.com and Finance.sina.com.cn / China	Financial status: Worries, predictions, explanations, etc. Operation: News, technologies, strategies changes, etc. Executive: Risk attitude, experience, etc. Marketing: Prospect, competitors, upstream and downstream, etc. Major Issues: Mergers, restructuring, major investment, etc.
D. Ntwiga and P. Weke [12]	Personal Credit / Linear reg, Logistic reg, Neural nets and Genetic Alg	Kenya	Ntwiga (2016) : Age, Gender, Trust, Interactions, Risk factor, Sociability, Relationship strength, Private data and Return on private data
Y. Zhanga, H. Jiaa, Y. Diaoa, M. Haia and H. Lia [13]	Online Peer-to-Peer (P2P) lending / Decision Tree	PPDai / China	membership score, prestige, forum currency, contribution and group
D. Björkegrena and D. Grissenb [14]	Personal Credit	Caribbean country	Mobile phone use preceding loan Weekly : Calls out, number, Calls out, minutes, SMS sent, Data use, Top Ups, Spend, Balance and Days of mobile phone data preceding loan Features derived from phone usage : Variation in usage, Periodicity of usage and Mobility
M. Hurley and J. Adebayo [9]	Personal Credit	USA	Media sites, browser activity, blogs, retail information, Internet Service Protocol (ISP) address and other data points obtained from public platforms

Information posted or broadcast on social networks may be inaccurate or misleading. For example, a position of an unhappy customer may be inaccurate and may not be indicative of a company's success or its solvency. As a result, not only social media data must be predictive of the solvency of the applicant, but the reliability of social media data needs to be confirmed. Whether the use of such information in a credit decision is appropriate must be established by the bank. Failure to use the correct information to make a credit decision can lead to problems of security and reliability which the bank needs to take into consideration [3] [8] [9].

VI. CONCLUSIONS AND OPEN PROBLEMS

We have presented a survey of credit scoring using non-traditional data in area of big data. As soon as credit history (which usually serves as input data in the classical credit scoring) is not exist or not rich enough for young clients, many academics and professional find that the social data can improve the scoring systems performance. Nowadays, Credit scoring using big data emerge as a way to ensure greater efficiency in underwriting while expanding access to the underbanked and to historically neglected groups. We have to mention that there is not much academic work covering the use of non-traditional data in credit scoring with the big data. However, There are some firms and start-ups whose domain is online/offline data retrieval, aggregation and customer analytics for credit organizations: Wonga, Kreditech, Big Data Scoring, Lenddo, SOCSCOR, Crediograph, etc[6].

Thereafter we presented the summarized principles and insights of big data for credit scoring. Machine Learning is at the core of data analysis in credit scoring. The accuracy of these algorithms depends on the size and the importance of the data. One of the challenges of "big data credit scoring" is the need for scalable Machine Learning algorithms implementation on very large data sets. Executing MapReduce jobs using Hadoop or spark and Machine Learning give best results for optimal time efficiency. Finally, the open discussion at the end can help researchers to have more general understanding the use of big data in credit scoring and motivate them to get involved and contribute. We hope that this survey will pave the way for subsequent research in big data for credit scoring. There is no reason to believe that adapting traditional credit scoring to Big Data could benefit banks substantially.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] S. Sadatrasoul1, M. Gholamian1, M. Siami1, Z. Hajimohammadi. Credit scoring in banks and financial institutions via data mining techniques, Journal of AI and Data Mining, Article 12, Volume 1, Issue 2, Summer and Autumn 2013, Page 119-129 (2013).

[2] S. Tuffery. Data mining and statistics for decision making, John Wiley & Sons, Ltd (2011).

[3] G. Guo, F. Zhu, E. Chen, Q. Liu, L. Wu, C. Guan. From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, ACM Transactions on the Web, Vol. 10, No. 4, Article 22 (2016).

[4] N. Kshetri. Big data's role in expanding access to financial services in China. International Journal of Information Management, 297-308 (2016).

[5] L. Devasena. Proficiency comparison ofladtree and reptree classifiers for credit risk forecast, International Journal on Computational Sciences & Applications (IJCSA) Vol.5, No.1 (2015).

[6] A. Masyutin. (2015) Credit scoring based on social network data. Business Informatics, No. 3 (33), pp. 15-23

[7] D. Andrea, A. Alessia. An overview of methods for virtual social network analysis Computational Social Network Analysis: Trends, Tools and Research Advances. Springer. P. 3–25 (2009).

[8] N. Sun, J. G. Morris, J. Xu, X. Zhu, M. Xie. iCARE: A framework for big data-based banking customer analytics, IBM Journal of Research and Development (Volume: 58, Issue: 5/6, Sept.-Nov) (2014).

[9] M. Hurley and J. Adebayo. Credit scoring in the era of big data, Yale Journal of Law and Technology: Vol. 18 : Iss. 1 , Article 5 (2016).

[10] Y. Wei, P. Yildirim, C. Van and C. Dellarocas. Credit Scoring with Social Network Data, Marketing Science Vol. 35, No. 2, pp. 234–258 ISSN 0732-2399 ISSN 1526-548X (2016).

[11] Y. Yang, J. Gu and Z. Zhou. Credit risk evaluation based on social media,Procedia Computer Science, Elsevier Volume 55, 2015, Pages 725-731 (2015).

[12] D. Ntwiga and P. Weke. Consumer lending using social media data, International Journal of Scientific Research and Innovative Technology ISSN: 2313-3759 Vol. 3 No.2 (2016).

[13] Y. Zhanga, H. Jiaa, Y. Diaoa, M. Haia and H. Lia. Research on Credit Scoring by fusing social media information in Online Peer-to-Peer Lending, Procedia Computer Science 91/168 – 174 (2016).

[14] D. Björkegrena, D. Grissenb. Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment, Entrepreneurial Finance Lab, Brown University (2015).

[15] T. Asha, U. Shravanthi, N. Nagashree, M. Monika. Building Machine Learning Algorithms on Hadoop for Bigdata, International Journal of Engineering and Technology Volume 3 No. 2 (2013).

[16] E. Fortuny, D. Martens, F. Provost. Predictive Modeling With Big Data: Is Bigger Really Better?, Big Data Mary Ann Liebert, Inc doi:10.1089/big.2013.0037, VOL. 1 NO. 4 (2013).

[17] L. Wang, C. Alexander. Machine Learning in Big Data, International Journal of Mathematical, Engineering and Management Sciences, Vol. 1,No.2,52-61 (2016).

[18] A. L'Heureux, K. Grolinger, H. ElYamany, M. Capretz. Machine Learning with Big Data: Challenges and Approaches. NSERC CRD at Western University (CRD 477530-14), DOI 10.1109/ACCESS.2017.2696365, IEEE (2017).

[19] F. Louzadaa, A. Araa, G. Fernandes. Classification methods applied to credit scoring: A systematic review and overall comparison, Surveys in Operations Research and Management Science, Elsevier (2016).

[20] H. Abdou, H, J. Pointon. Credit scoring, statistical techniques and evaluation criteria: A review of the literature, Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59-88 (2011).

[21] S. Jun, S. Lee and J. Ryu. A Divided Regression Analysis for Big Data, International Journal of Software Engineering and Its Applications Vol. 9, No. 5 (2015), pp. 21-32 (2015).

[22] A. Bahnsen, D. Aouada and B. Ottersten. Example Dependent Cost-Sensitive Logistic Regression for Credit Scoring, 13th International Conference on Machine Learning and Applications, 978-1-4799-7415-3/14 IEEE (2014).

[23] H. Nguyen. Default Predictors in Credit Scoring: Evidence from France's Retail Banking Institution, Journal of Credit Risk, Vol. 11, No. 2, Pages 41–66 (2015).

[24] Y. Jiang. Credit Scoring Model Based on the Decision Tree and the Simulated, Second International Conference on Computer Modeling and Simulation (2009).

[25] K. Leung, F. Cheong, C. Cheong. Consumer Credit Scoring using an Artificial Immune System Algorithm, 1-4244-1340-0/07 IEEE.Annealing Algorithm, World Congress on Computer Science and Information Engineering (2008).

[26] Y. Bengio. Learning deep architectures for AI, in Foundations and Trends in Machine Learning, 2(1):1–127 (2009).

[27] V. Ha, H. Nguyen. Credit scoring with a feature selection approach based deep learning. MATEC Web of Conferences 54, 05004, MIMT (2016).

[28] A. Niimi. Deep Learning for Credit Card Data Analysis, World Congress on Internet Security (WorldCIS-2015).

[29] W. Chen, C. Ma, L. Ma. Mining the customer credit using hybrid support vector machine technique. Expert Systems with Applications 36 (4), 7611–7616 (2009).

[30] L. Chen, T. Chiou. A fuzzy credit-rating approach for commercial loans: a Taiwan case, OMEGA - International Journal of Management Science, 27,407-419 (1999).

[31] U. Farouk, J. Panford, J. Hayfron-Acquah. Fuzzy Logic Approach to Credit Scoring for MicroFinances in Ghana, International Journal of Computer Applications (0975 – 8887) Volume 94 – No.8 (2014).

[32] A. Lahsasna, R. Ainon, T. Wah. Credit Scoring Models Using Soft Computing Methods: A Survey, The International Arab Journal of Information Technology, Vol. 7, No. 2 (2010).

[33] S. Mammadlia. Fuzzy logic based loan evaluation system, 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS (2016).

[34] S. Finlay. Are we modelling the right thing? The impact of incorrect problem specification in credit scoring, Expert Systems with Applications Volume 36, Issue 5, Pages 9065–9071(2008).

[35] B. Chen, W. Zeng, Y. Lin. Applications of Artificial Intelligence Technologies in Credit Scoring: a Survey of Literature, 2014 10th International Conference on Natural Computation. IEEE 978-1-4799-5151-2/14 (2014).

[36] C. Leong. Credit Risk Scoring with Bayesian Network Models, Springer Science Business Media New York, Comput Econ, Volume 47, Issue 3, pp 423–446 (2015).

[37] C. Tsai a, M. Chen. Credit rating by hybrid machine learning techniques, Applied Soft Computing 10 (2010) 374–380(2010).

[38] A. Fensterstock, J. Salters, R. Willging. On the Use of Ensemble Models for Credit Evaluation, The Credit and Financial Management Review (2013).

[39] L. Breiman. Random forests. Machine Learning, Statistics Department University of California Berkeley, CA 94720 45, 5–32 (2001).

[40] G. Wang, J. Ma, L. Huang, K. Xu. Two credit scoring models based on dual strategy ensemble trees, Knowledge-Based Systems 26 (2012) 61–68 (2012).

[41] M. Malekipirbazari, V. Aksakalli. Risk assessment in social lending via random forests, Expert Systems with Applications 42 4621–4631 (2015).

[42] G. Kapil, A. Agrawal and R. Khan, A Study of Big Data Characteristics, Communication and Electronics Systems (ICCES) International Conference (2016).

[43] G. Bello-Orgaza, J. Jungb, D. Camachoa. Social big data: Recent achievements and new challenges, Information Fusion 000(2015)1–15 (2015).

[44] Z. Zhou, N. Chawla, Y. Jin, G. Williams. Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives, IEEE Computational intelligence magazine (2013).

[45] A. Verma, A. Mansuri, N. Jain. Big Data Management Processing with Hadoop MapReduce and Spark Technology: A Comparison, Symposium on Colossal Data Analysis and Networking (CDAN) (2016).

[46] J. Wang, D. Crawl, S. Purawat, M. guyen, I. Altintas. Big Data Provenance: Challenges, State of the Art and Opportunities, IEEE International Conference on Big Data (2015).

[47] A. Priyadarshini, S. Agarwa. A Map Reduce based Support Vector Machine for Big Data Classification (2015).