American Journal of Data Mining and Knowledge Discovery
2022; 7(2): 5-12
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20220702.11
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

Deep Learning for Sentiment Analysis to Predict the Probability of Bank Loan Default

Katleho Makatjane

Department of Statistics, University of Botswana, Gaborone, Botswana

Email address:
makatjanek@ub.ac.bw

To cite this article:
Katleho Makatjane. Deep Learning for Sentiment Analysis to Predict the Probability of Bank Loan Default. *American Journal of Data Mining and Knowledge Discovery*. Vol. 7, No. 2, 2022, pp. 5-12. doi: 10.11648/j.ajdmkd.20220702.11

Received: July 20, 2022; **Accepted**: August 10, 2022; **Published**: January 9, 2023

Abstract: Social networks have taken the world by storm with their fast and commendable speed. It could be social, political, or present with all sorts of situations that arise. People's opinions around the globe are articulated through social media, making it apposite for drawing out opinions. Organizations that aim at refining their products and services use sentimental analysis methods to increase their resources. In the banking and financial industry, it is much easier to get feedback from customers through Twitter and or Facebook sentimental analysis. The elements associated with Twitter or consumers and services providers who want to know who they are, and what they are in their daily life towards their bank and financial portfolios cannot suppress Facebook sentimental analysis. Hence, this study aims to predict the probability of bank loan default and classify the Twitter messages by exhibiting the results of deep learning algorithms. High-performance computing with hyper-parameter space for grid-search (HPSGS) and hyper-parameter optimization (HPO) are developed and compared with the effectiveness of three gradient boosting decision trees. The results reveal that the XGboot algorithm has a better prediction or features a score that is better as compared to other algorithms at 91 percent in the test data and 93 percent performance in the validation data. It is also seen that women are more likely to default than men as across all the algorithms, their likelihood of risk or default is higher than that of men. These results are useful for decision-makers and the financial sector for future use and planning in credit risk and bank loan default-prone areas.

Keywords: Bank Loan, Credit Risk, Data Mining, Deep Learning, Machine Learning, Sentiment Analysis

1. Introduction

In their daily lives, both organizations and individuals had found lending a loan a significant part of life. With ever-increasing financial competition and a substantial number of financial constraints, the activity of taking out a loan has become more or less unavoidable [1]. Around the globe, most people rely on loan lending for a quiet number of reasons. This incorporates the incapacitating of financial constraints and achieving personal goals. Similarly, small to large businesses count on borrowing money for the basic purpose of managing their affairs and operating smoothly during times of financial constriction [2]. Some significant risks in financial institutions are said to be carried by loan lending which for both lenders and receivers is beneficial [3]. This type of risk, Perera, H. and Premaratne S [1] refer to it as credit risk, which means that the borrower is unable to repay the loan within a specified period agreed by the lender and the borrower. Credit risk is known to be a major concern of financial institutions because it can lead to a serious credit default situation and can be disastrous for lenders [4]. Even after identifying the risks, financial institutions around the world consider lending to be an important activity.

Although profit and risk are directly proportional, credit risk factors increase [5-6]. Throughout the financial world, lenders try to reduce credit risk by carefully evaluating and verifying the borrower's ability to repay their loans. Thus, an effective and comprehensive credit risk assessment leads to lower default scenarios for financial institutions [6].

Historically, financial institutions have focused on recruiting highly skilled talent for the sole purpose of assessing a candidate's eligibility for a loan based on two factors. (i) A risk assessment to help verify applicants' eligibility for loan approval. (ii) Or denial based on numerical scores. Over the years, various measures of default risk have been developed to assess the likelihood of default for individual and corporate borrowers and to help banks better manage credit risk and allocate economic capital. This means that various quantitative models have been developed [7].

According to the underlying technology, hypothetical risk models and algorithms can be broadly classified into traditional statistical categories and categories of intelligence or machine learning [8]. Examples of traditional statistical methods include simple multivariate analysis (MDA), multiple regression analysis, linear discriminant analysis (LDA), and logit models. On the other hand, the other category includes machine learning (ML) techniques such as recursive division algorithms, neural network models (NN), decision trees, k-nearest neighbors (KNN), genetic programming, approximate clusters, fuzzy neural techniques, soft and multiprogramming criterion, and classification and regression trees (CART). An overview and comparative study of these credit risk models are provided in [9-12]. Although comparative analysis of default classification models has been intensively carried out, the results of these studies are not always consistent with each other. A default classification model that performs better than another using one dataset and its performance criteria may be undesirable in various situations. For example, Naidu, H. P. and K. Govinda [13] reported that the neural network model generally provides better results than the default prediction using CART-based decision tree algorithms, while Aziz, S. and M. Dowling [14] concluded that the CART decision tree model is better than neural networks for predicting mortgage defaults. However, the process of credit scoring in the past required specialists with statistical algorithms to accurately predict the qualification of a loan or deviation. Though, researchers and companies have recently chosen training classifiers based on machine learning algorithms and neural networks which automatically predict their credit rating that is based on historical data and authenticity; and, according to Perera, H. and Premaratne S [1], this is divided by credit disorders.

Machine learning algorithms are employed to analyze all individuals' historical credit scores or authorized passed bank loans and report if an individual can currently be granted a loan/credit. Professionals investigate these reports and provide feedback to an automated system, which is then used to train and update the algorithm, resulting in an improved loan default detection performance over time. Therefore, this study contributes to the existing methods by building a Chrome extension that flags "defaulter" that will help banks to know if the person who is asking for a loan using online systems has a good or bad credit score and gives a precise likelihood of defaulting. This is a new system that banks will be able to use for both online and in-bank loan approval systems. Even though detecting credit default is considered a high priority for many organizations, recent literature lacks updated and comprehensive in-depth reviews that can help organizations with their decisions in selecting appropriate data mining methods. In the past, some scholars such as Aziz, S. and M. Dowling [15] have indicated that neural networks were introduced to detect credit card frauds. Therefore, in this study, the focus is based on the subset of ML which is deep-learning to predict the probability of credit (loan) defaulter and significantly classify sentiments of a defaulter. The deep learning method used in this study is based on sentiment analysis or opinion mining and is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [24]. In this case, people's opinions are analyzed to get deeper information about who is likely to have a bad credit score or who is likely to default before authorizing a credit or bank loan. Deep learning algorithms are cast-off to extract knowledge about the processing of quantitative and qualitative data hence; the use of sentiment analysis and extreme gradient boosting decision tree (GBDT) to predict the probability of bank loan default. To accurately label people’s opinions and those that might be a defaulter or not, a crowd-sourcing rating application instead of team labels is used for this task. This, therefore, lead to supervised learning alone, not unsupervised learning. This paper will help banks to know who among women and men are likely to default with the given loan or resonates well in the banks and financial sector. Sentiment analysis may provide answers to many of the company's questions regarding its success and growth or decline of sales. It can help the banks and other financial institutions to know their service performance and handle customer complaints through which strategy analysis can be performed [25].

2. Methods and Procedures

The proposed system is used to predict the probability of credit or bank loan default on a real-time basis by analyzing incoming credit or loan transactions and this designed system consists of two components for default prediction. These are described or discussed in the next two subsections.

2.1. Designing a Framework for Data Pre-Processing

This constituent is legally responsible for the usage of big data effectively and bids it to the analytical server for predictive modeling. The configuration of the system mainly consists of Spark which introduces a data structure called resilient distributed dataset (RDD). Data is repeated and outcomes are stored in the cache of the machines cluster, and Altman, E. I [9] have declared that this machine's cluster is concerned with how to create a computer that enhances itself based on its experiences. This machine learning cluster has a lot of potential for data analysis and according to Kumar, P. R., Ravi, V. and Verikas, A., Kalsyte, Z., Bacauskiene, M., and Gelzinis [10, 11], is typically used to perform data pre-processing, learning, and evaluation. Data pre-processing aids in the transformation of atomic data into a usable format. The pre-processing phase cleans, extracts, transforms, and fuses these datasets into a shape that may be used as learning inputs. The learning step chooses learning algorithms and modifies parameters of the model to produce desirable outputs from pre-processed input data. Furthermore, data pre-processing can be accomplished by using a variety of learning approaches, including representational learning. The taught models will next be evaluated to determine their performance. A classifier's performance evaluation, for example, may include dataset selection, and performance measurement. The evaluation outputs will lead to changes in the factors of the chosen learning algorithms and/or the selection of different algorithms.

2.2. Gradient Boosting Decision Tree

Boosting algorithms were originally introduced by the machine learning community for classification problems. The main approach is to combine iteratively, several simple models; called 'weak learners', to obtain a 'strong learner' with improved prediction accuracy [16]. The Boosting algorithm combines simple classification rules with 'mediocre' performance in terms of the misclassification error rate to produce a highly accurate classification rule. Stochastic gradient boosting on the other hand provides an enhancement that incorporates a random mechanism at each boosting step, showing improvement in performance and speed while at the same time generating the ensemble. As declared by Son, J., Jung, I., Park, K., and Han, B [17], the gradient boosting algorithm utilizes decision stumps or regression trees as weak classifiers. These weak learners measure the observed error in each node, and split the node using the following test function $k: \mathbb{R}^n \to \mathbb{R}$ with a threshold τ, and return values η^{ι} and η^{r} respectively. To minimize the error after a split, the following triplet $(\eta^{\iota}, \eta^{r}, \tau)$ is identified and thereafter, an optimal split is obtained; and, according to Son, J., Jung, I., Park, K., and Han, B [17] an optimal split is given by

$$\varepsilon(\tau) = \sum_{i:k(x_i)<\tau} \omega_i^j \left(r_i^j - \eta^{\iota}\right)^2 + \sum_{i:k(x_i)\geq\tau} \omega_i^j \left(r_i^j - \eta^{r}\right)^2 \quad (1)$$

where ω_i^j and r_i^j are the weights and responses of x_i for the j[th] iteration. By minimizing the error in equation (1), an optimal triplet is attained as (η^*, η^*, τ^*) over all possible $\tau's$ at each node. Note that $\tau(\eta^{\iota}, \eta^{r})$ is found simply by working out the weighted average of r_i^j's over the training data that tumble on the resultant lateral of the fragmented. The training process of gradient boosting decision tree is presented in Algorithm 1[1]

Algorithm 1. Gradient Boosting Decision Tree

Initialize: $f_0(X) = 0, \eta, d_0 = 0$

for $j = 1; \cdots; M$ do

$\omega_i = exp\left(-y_i f_{j-1}(x_i)\right), \mathrm{i} = 1, \cdots, \mathrm{N}$

$r_i = -y, \mathrm{i} = 1, \cdots, \mathrm{N}$

$S = \{(x_i, \omega_i, r_i)\}_{1,\cdots,\mathrm{N}}$ and $v = \{x \mid x \in \mathbb{R}\}$

$\mathrm{R} = \mathrm{GROWTREE}(S, v, \eta_0, d_0)$

$$f_j(x) = f_{j-1}(x) - v\blacksquare \sum_{k=1}^{|R|} \eta_k \delta(x \in \Re_k)$$

where $\Re_k, \eta_k \in \mathrm{R}$

end

Procedure SPLITLEARNING (S)

$(\eta^{\iota*}, \eta^{r*}, \tau^*) = argmin_{\eta^{\iota},\eta^{r},\tau}\ \varepsilon(\tau)$ in equation 3

Return $(\eta^{\iota*}, \eta^{r*}, \tau^*)$

where,

$$\varepsilon_{t+1}\left(\tau_{(t+1,\rho)}\right) = \sum_{i=1}^{N_\rho+1} \omega_{t+1,i}\left(r^i - \eta_{t+1,\rho}\right)^2. \quad (2)$$

Primarily, $\eta_{t+1,\rho}$ is augmented in equation (2) for the reason that it is assumed that $\tau_{(t+1,\rho)}$ is presently acknowledged in online learning, and $\left\{\left(\omega_{t+1,i}, r_i\right)\right\}_{i=1,\cdots,N_\rho}$ is unattainable. For this reason, Touzani, S., Granderson, J., and Fernandes, S [16] blatant that it is therefore unbearable to minimize the error directly by adjusting $\tau_{t+1,\rho}$ and also by figuring the slanted average of r_i^j as in the off-line learning. Subsequently, Liu, B [18] updated a weak classifier based on the new classifier as well as the limited information of the present classifier. Note that equation (3) is derived by signifying the right-hand side of equation (2) by $\Delta\eta$

$$\varepsilon_{t+1}\left(\tau_{(t+1,\rho)}\right) = \sum_{i=1}^{N_\rho+1}\left\{\omega_{t+1,i}\left(r_i - \eta_{t,\rho}\right)^2 + \omega_{t+1,i}\left(-2\left(r_i - \eta_{t,\rho}\right)\Delta\eta + (\Delta\eta)^2\right)\right\} \quad (3)$$

which helps to minimize the above quadratic function to Δη. Employing a recursive procedure to find η, Liu, B [18] obtained these two models

$$\Delta\eta^{\iota}\left(\tau_{t+1,\rho}\right) = \alpha\left(\omega_{t+1,i}N_\rho + 1rN_{\rho+1} - \eta_{t,\rho}^{\iota}\right) \quad (4)$$

and

$$\Delta\eta^{r}\left(\tau_{t+1,\rho}\right) = \alpha\left(\omega_{t+1,i}N_\rho + 1rN_{\rho+1} - \eta_{t,\rho}^{r}\right), \quad (5)$$

where α denotes a learning rate. The difference between the weighted response and the previous return value is the amount used to bring up-to-date the return value η. For readings on gradient boosting algorithm see for instance [17].

The algorithm starts with a single leaf, and then the learning rate is optimized for each node and each record [20]. The eXtreme Gradient Boosting (XGBoost) which is proposed in this study, is an extremely scalable, flexible, and versatile tool. It is designed to make better use of resources and to overcome the limitations of previous gradient boosting. The main difference between XGBoost and other gradient boosting trees is that it uses a new adjustment method to control the over-fitting of parameters. So, it is faster and stronger when one adjusts the model. The regularization technique is achieved by adding a new term to the loss function as

$$L(f) = \sum_{i=1}^{n} L(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \quad (6)$$

1 v is a shrinkage factor to training data and the classifier.

$$\Omega(\delta) = \alpha|\delta| + 0.5\beta\|w\|^2 \tag{7}$$

where $|\delta|$ is the number of branches, w is the value of each leaf and Ω is the regularization function. XGBoost uses a new gain function, as

$$Gain = 0.5\left[\frac{G_L^2}{H_L+\beta} + \frac{G_R^2}{H_R+\beta} - \frac{(G_R+G_L)^2}{H_R+H_L+\beta}\right] - \alpha \tag{8}$$

2.2.1. Sentimental Analysis

Deep learning to collected sentiments is now applied. Consequently, this section delivers an ephemeral introduction to two core tasks of sentiment analysis. For additional details, please refer to Liu, B [18] on sentiment analysis. Researchers mainly study sentiment analysis at three levels of granularity: document level, sentence level, and aspect level. At the document level, opinionated documents are categorized by venting an inclusive positive or negative opinion. In the case of the current paper, these are defaulter and non-defaulter. The sentiment analysis tool considers the document as a whole and assigns a sentiment level to each sentence. However, it cannot be assumed that every sentence is self-righteous. Opinion statements obtained are classified as positive or negative opinions. Sentiment rating can also be formulated at the sentence level as a three-category classification problem that ranks sentences as neutral, positive, or negative. Sentiment analysis at the aspect level is more detailed than sentiment analysis at the document or sentence level. It focuses on extracting and synthesizing people's views on specific entities and aspects or attributes of those entities. The general activity of aspect-based sentiment analysis consists of several secondary activities such as aspect extraction, entity extraction, and classification of lateral feeling [19]. Therefore, this study uses document-level sentiment assessments to extract positive and negative options and predict the probability of loan or credit default among men and women in Botswana.

Instead of using the Bag of Words (BoW) model to classify the sentiments, boosting decision tree algorithms are used to represent input words of a document as a dense vector (or called a dense document vector). The proposed procedures also can encode approximately semantic and syntactic word chattels [21, 22]. Despite its popularity, BoW has its drawbacks. First, word order is ignored, meaning that two documents can have the same representation if they contain the same words. Bag-of-n-grams, an extension of BoW, can consider sequences of words in a narrow context (n-gram) but suffers from data scarcity and high dimensionality. Second, BoW can rarely encode verbal meanings [20]. Methods that are grounded on neural networks, were prosed to overcome the inadequacies of BoW and these engender impenetrable vectors for presenting some words.

2.2.2. Cross-Validation on Classification Problems

To estimate the test error rate, various ways use math concepts to adjust the training error rate. A Bayesian Leave-One-Out Cross-Validation (LOOCV) method, similar to that used by Magnusson M, Vehtari A, Jonasson J., and Andersen M [26] in their empirical analysis is utilized in this study. This is a cross-validation procedure where each observation of a dataset acts as a validation set and the remaining $n-1$ observations serve as a training set. The LOOCV uses a single set of observational validations to fit models and make predictions. As an affirmation set, the process is repeated N times per observation. The model is then evaluated against the missing data point, and the prediction's test error is recorded. The total prediction error is computed by averaging the test error estimates for all data points [27]. The number of folds in this form of K-fold cross-validation is directly proportional to the number of observations ($K = N$). Bürkner P. C, Gabry J, and Vehtari A [28] added that the set is verified by a single observation (x_1, y_1), whereas the training set is composed of the remaining observations $\{(x_2, y_2), \cdots, (x_n, y_n)\}$. A deep learning process fits $n-1$ training observations and omitted observations x_1 are used to produce predictions $\hat{y}_1$. Because (x_1, y_1) is not included in the fitting process, then;

$$MSER_1 = (y_1 - \hat{y}_1)^2 \tag{9}$$

provides a somewhat accurate measure of the test error. Notwithstanding being neutral for the test error, MSER_1 is an insufficient estimate because it is dependent on a single observation (x_1, y_1) and has great diversity. This is remedied by repeating the method n times, yielding n squared errors like $\mathrm{MSER}_1, \cdots, \mathrm{MSER}_\mathrm{n}$.

The aggregate of these n test error estimates, as per James, G., Witten, D., Hastie, T., and Tibshirani, R [29], is the LOOCV estimate for the test MSER; given by

$$\mathrm{CV_n} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{MSER_i} \tag{10}$$

Bürkner P. C, Gabry J, and Vehtari A [28] obtained the MSE (Mean squared error) by fitting the entire dataset;

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2, \tag{11}$$

where $\hat{y}_i$ is defined as the $\mathrm{i^{th}}$ fitted value from the original least-squares fit, and h_i represents the amount of influence an observation has on its fit, ranging from 0 to 1 punishing the residual because it divides by a tiny integer and h_i raises the residual value [28].

3. Empirical Analysis

The data is obtained from tweets based on the bank loan and credit risk controversy and then it is classified in Rapid miner software and RStudio for stemming and cleaning. The extracted data consists of 16 attributes where 6 are categorical and 10 are numeric with a total of 100000 transactions. The distribution of credit amount by gender as depicted in Figure 1, clearly indicates a skewed distribution for both genders, and this skewness is also confirmed by Figure 2 where the highest credit is being taken by both males and females aged between 20 to 40 years.

Figure 1. *Distribution of credit amount by Gender.*

Moreover, looking again at Figure 1, there is a good association between male and female loan applications, and female applicants are getting slightly higher credit than males. This is because male applicants almost double the number of female applicants and it can be seen that female applicants are more likely to be risky than male applicants.

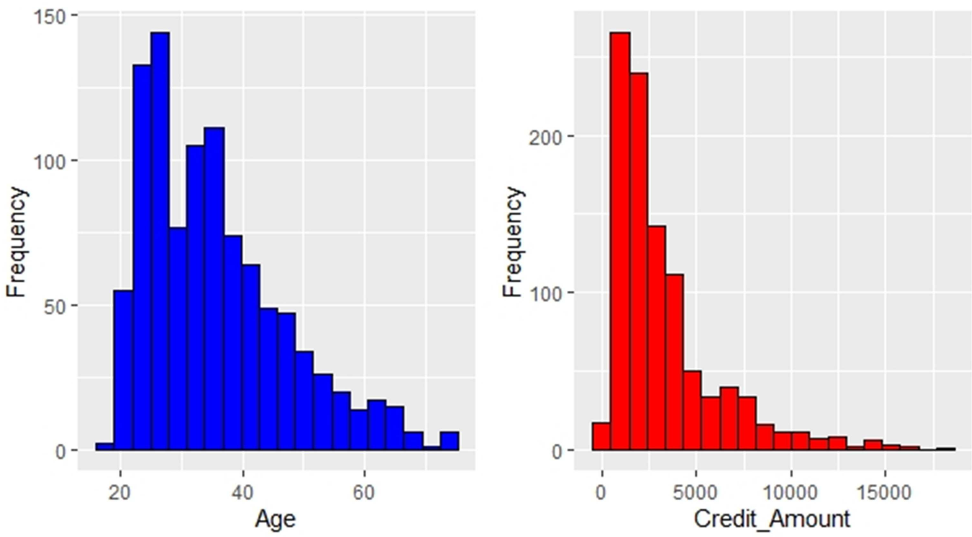

Figure 2. *Age and Credit Amount Distribution.*

Using the Shapiro-Wilk test, the null hypothesis of normality is rejected at a 1% level of significance; hence the conclusion that the data is not normally distributed and it has to be transformed before further analysis. Using box transformation, the observed skewness is transformed into a normal distribution using the log1p function. The log transformation of age helps in improving the distribution of the data which now gives an age variable to be slightly skewed but it has a much better distribution as compared to the earlier distribution. Moreover, Figure 3 displays the proportion of good and bad loans by sex. It is seen in this figure that males have the highest proportion of having good loans as compared to the females and also about 20% of them are having a bad loan. While the proportion of women with a bad loan is at about 10%. This indicates that men are likely to be defaulters compared to women.

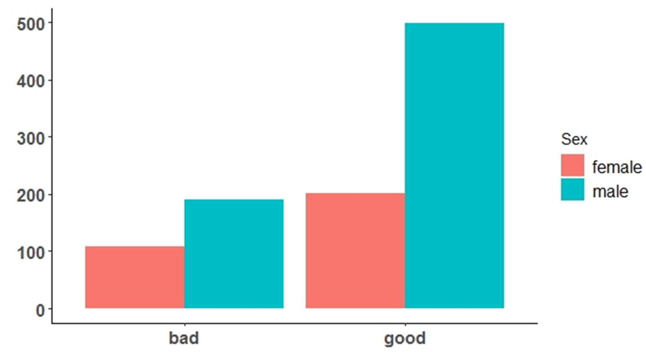

Figure 3. *The proportion of Good and Bad loans by Sex.*

Three deep learning algorithms namely, LightGBM, XGBoost, and Catboost are used to determine customers' likelihood of evasion; and three samplings' scenarios are also established. Firstly, a database is modeled and it has a distribution of good and bad credit risk variable values with a proportion of 70 percent to 30 percent. This data is highly imbalanced because there is a class that has scarcer annotations than the other group; and, in the modeling phase, problems may arise due to this imbalance incurring the biased default prediction event of the model. As a result, the model may not capture enough information from the available data, favoring the prediction of the non-default class. Modeling with this imbalanced data can bring about high specificity for the mainstream class but, for a smaller group, brings about poor results with the same trials. In addition, there is a problem of efficiency on accuracy and error rate while assessing the performance of classifiers because of imbalance in a data set. For instance, if there is a 99 percent that accounts for a group of majority predictions on the data, then the most naïve classifier that predicts only this class would still have the same accuracy of 99 percent whereas the minority will only have 1 percent classification accuracy. This type of classifier becomes inoperable when predicting the event of the minority group. Furthermore, when applied to imbalanced datasets, traditional classifiers' results are influenced for the reason that they are inclined to heighten accuracy although, developing a model that is analogous to the naïve classifier described earlier. To reduce this awkwardness, the second scenario aims to reduce the disparity between the proportion of non-defaulters and defaulters' cases by under-sampling the majority class to achieve an 80 percent to 20 percent ratio between non-defaulter and defaulter values. In conclusion, the imbalanced data set from under-sampling the majority class and over-sampling the minority class with 50 percent replicates of the defaulter's cases, leads to balancing the data set. The performance of the model is now assessed by the percentage of 80 training set and 20 validations set. The models are only trained on the training sets, so the other observations can be used to assess the likelihood of entering default, just as if the models are put into production (unseen data).

A distributed grid search and Hyper-parameter optimization framework are employed just as in the empirical analysis of Anghel, A., Papandreou, N., Parnell, T., De Palma, A., and Pozidis, H [23], to assess how well the algorithms take a broad view to unobserved data and to fine-tune the model parameters through Bayesian optimization. Bayesian optimization consists of iteratively evaluating new parameters according to an acquisition function and updating the surrogate model with their results until a certain evaluation budget is exhausted. To avoid the assumption of prior and maximize the probability of including the finest formation, iterations in the HOP framework are set wider as evidenced in Table 1. Identical parameters are shared by the GBDT. Catboost does not have a feature fraction parameter, and LightGBM has two boosting types' gbdt and goss. Finally, the number of leaves in LightGBM is set to 2^{depth}.

Table 1. *Hyper-Parameter Space for Grid-Search and HPO.*

	Iterations	Depth	Regularizer	Learn Rate	Feature Fraction	Boosting
Grid Search						
Catboost	40,80,160,320,480	4,8,10,12	0,1,100	0.1,0.3		
XGBoost	40,80,160,320,480	4,8,10,12	0,1,100	0.1,0.3	0.8,1.0	
LightGBM	40,80,160,320,480	4,8,10,12	0,1,100	0.1,0.3	0.8,1.0	gbdt, goss
Hyper-Parameter Optimization (HPO)						
Catboost	[20,1500]	[4,18]	[0.001, 2]	$[10^{-4},10^{10}]$		
XGBoost	[20,1500]*	[4,18]*	[0.001, 2]	$[10^{-4},10^{10}]$	[0.001, 2]	
LightGBM	[20,1500]	[4,18]	[0.001, 2]	$[10^{-4},10^{10}]$	[0.001, 2]	gbdt, goss

In Table 2 the best validation score while using HPO is reported; as well as the score on the unseen test data for the resulting hyper-parameter configuration. As a reference, the score obtained by the logistic model tree (LMT) similarly includes only the labels of the classes based on their characteristics and frequencies. From Table 2, it can be seen that the XGboot algorithm has better predictions or scores than the other algorithms with 91% performance on test data and 93% performance on validation data. It can also be seen that men are more likely to default than women as across all the algorithms, their likelihood of risk or default is higher than that of men. This confirms the results reported in Figure 3 that, there are more men with bad credit scores than women. The cause of men to be likely to default is that experimental asset market design by Reynal-Querol, M., and Montalvo, J. G [31] has shown that all-male markets generate significant price bubbles, while all-female markets generate smaller bubbles or none at all. Women's price expectations can explain this behavior, as they are significantly lower than those of men. In contrast, when the experiment is repeated without revealing the single-sex composition of the groups, gender differences disappear, suggesting that common expectations and stereotyping can lead to bubble formations [32]. This gives an overall rate average of 50 percent for both genders with men being 1 time more likely to default than women. The odds of men to women are given as 0.152.

Table 2. *Best test scores across algorithms and Risk detection rate.*

	Baseline	XGBoost	LightGBM	Catboost
Test	0.7802	0.9076	0.7983	0.7965
Validation	0.8106	0.9254	0.8416	0.8142
	Risk			
Women	0.432	0.534	0.492	0.381
Men	0.492	0.555	0.521	0.421

4. Conclusion

The analysis of this study confirmed predictions about how effective Twitter sentiment analysis is. The GBDT classifier used in the algorithm accurately reflects public sentiment, allowing banks and other credit issuers to easily interpret and use the data in an attempt to improve defaults or dislikes. A detailed empirical analysis of the three latest GBDT packages (XGBoost, LightGBM, and Catboost) is in this study. The level of HPC acceleration is determined and evaluated by the ability of each package to give predictions faster in the Bayesian HPO context. The observation made is that for a fixed set of hyper-parameters, XGBoost provides the largest reduction in training time giving high prediction accuracy of defaults. It is further shown that XGBoost has accurately detected the default or bank loan risk at 53 percent for women and 55 percent for men. This shows that men are more likely to default than women. This analysis is very new and has not been tested with many other classification models, so there is still room for enhancement. The main drawback is that the maximum number of tweets analyzed by Rapid Miner is 100000 tweets per day for free users. Otherwise, Hadoop and Spark for big data are needed to broaden the scope of research and analysis in the future by collecting a large amount of data and expanding the data mining related to the analysis method.

References

[1] Perera, H. and Premaratne S (2016), An artificial neural network approach for the predictive accuracy of payments of leasing customers in Sri Lanka. H. A. P. L. Perera/BESSH-2016/Full Paper Proceeding. 285 (2): 1-11.

[2] Abellán, J. and Castellano J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert systems with applications 73: 1-10.

[3] Adewusi, A. O., Oyedokun, T. B., and Bello, M. O. (2016). Application of artificial neural network to loan recovery prediction. International Journal of Housing Markets and Analysis. 9 (2): 222-238.

[4] Netzer, O., A. Lemaire, and Herzenstein M. J. J (2019). When words sweat: Identifying signals for loan default in the text of loan applications. Journal of Marketing Research, 56 (6): 960-980.

[5] Aslam, U., Tariq Aziz, H. I., Sohail, A., and Batcha, N. K. (2019). An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16 (8): 3483-3488.

[6] Jiang, C., Wang, Z., Wang, R., and Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research, 266 (1): 511-529.

[7] Tanoue, Y., Kawada, A., and Yamashita, S. (2017). Forecasting loss given default of bank loans with multi-stage model. International Journal of Forecasting, 33 (2): 513-522.

[8] Song, Y., and Peng, Y. (2019). A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. IEEE Access, 7: 84897-84906.

[9] Altman, E. I. (1984). The success of business failure prediction models: An international survey. Journal of Banking and Finance, 8 (2), 171-198.

[10] Kumar, P. R., and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. European journal of operational research, 180 (1): 1-28.

[11] Verikas, A., Kalsyte, Z., Bacauskiene, M., and Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Computing, 14 (9): 995-1010.

[12] Makatjane, K., N. Moroke, and B. Ncube (2020) Detecting Financial Fraud in South Africa: A Comparison of Logistic Model Tree and Gradient Boosting Decision Tree. In Proceedings of 36th International business information management association, 1331-1340.

[13] Naidu, G. P. and K. Govinda. (2018) Bankruptcy prediction using neural networks. in 2018 2nd International Conference on Inventive Systems and Control (ICISC. 248-251 IEEE.

[14] Aziz, S. and M. Dowling (2019). Machine learning and AI for risk management. In Disrupting finance, 33-50. Palgrave Pivot, Cham.

[15] Pumsirirat, A., and Liu, Y. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine. International Journal of advanced computer science and applications, 9 (1): 18-25.

[16] Touzani, S., Granderson, J., and Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158: 1533-1543.

[17] Son, J., Jung, I., Park, K., and Han, B. (2015). Tracking-by-segmentation with online gradient boosting decision tree. In Proceedings of the IEEE international conference on computer vision, 3056-3064.

[18] Liu, B. (2020). Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge university press.

[19] Al-Shabi, M. A. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. IJCSNS, 20 (1): 1.

[20] Kumar, R. S., Saviour Devaraj, A. F., Rajeswari, M., Julie, E. G., Robinson, Y. H., and Shanmuganathan, V. (2022). Exploration of sentiment analysis and legitimate artistry for opinion mining. Multimedia Tools and Applications, 81 (9): 11989-12004.

[21] Aquino, P. A., López, V. F., Moreno, M. N., Muñoz, M. D., and Rodríguez, S. (2020). Opinion mining system for Twitter sentiment analysis. in International Conference on Hybrid Artificial Intelligence Systems. 465-476 Springer Cham.

[22] Van Looy, A. (2022). Sentiment analysis and opinion mining (business intelligence 1). In Social Media Management. 147-163. Springer Cham.

[23] Anghel, A., Papandreou, N., Parnell, T., De Palma, A., and Pozidis, H. (2018). Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint arXiv: 1809.04559.

[24] Hancock, J. T., and Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of big data, 7 (1): 1-45.

[25] Dixit, P., and Silakari, S. (2021). Deep learning algorithms for cybersecurity applications: A technological and status review. Computer Science Review, 39: 100317.

[26] Magnusson M, Vehtari A, Jonasson J. and Andersen M, (2020). Leave-one-out cross-validation for Bayesian model comparison in large data. In International Conference on Artificial Intelligence and Statistics. 341-351. PMLR.

[27] Kaplan, D. (2021). On the Quantification of Model Uncertainty: A Bayesian Perspective. Psychometrika, 86, 215–238.

[28] Bürkner P. C, Gabry J, and Vehtari A, (2020). Approximate leave-future-out cross-validation for Bayesian time series models. Journal of Statistical Computation and Simulation, 90 (14), 2499-2523.

[29] James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). An Introduction to Statistical Learning 112, 18. New York Springer.

[30] Ma, Xiaomeng, and Shuliang Lv. (2019): Financial credit risk prediction in internet finance driven by machine learning." Neural Computing and Applications 31 (12): 8359-8367.

[31] Reynal-Querol, M., and Montalvo, J. G. (2020). Gender and Credit Risk: A View From The Loan Officer's Desk. Barcelona GSE Working Paper Series Working Paper no 1076.

[32] Eckel, C. C., and F¨ullbrunn, S. C. (2017). Hidden vs. known gender effects in experimental asset markets. Economics Letters, 156: 7-9.