Ann. For. Res. 65(1): 7521-7528, 2022
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

A REVIEW ON COMMERCIAL BANK LOAN STATUS PREDICTION USING MACHINE LEARNING CLASSIFICATION

Dr. Ashish Mishra
Professor, Dept. of CSE Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.)
ashishmishra@ggits.org

Yashi Patel
Research Scholar, Dept of CSE Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.), yashipatel020@gmail.com

Abstract— A credit institution needs an accurate and consistent credit risk rating system in order to operate effectively and regularly. If accurate projections are not maintained, they will not be able to conduct their organization with any acceptable levels of inaccuracy or continue in a financially viable and transparent manner. Financial organizations are finding it harder and harder to distinguish between different loan requests as default rates rise. The topic of credit risk is now a source of widespread concern on a global scale. Alternatives have been tried, but no clear solution has emerged. This study suggests a machine learning strategy that can both detect and control borrower risk. It provides a method to precisely anticipate whether or not a loan request would be approved taking into account the borrowers' financial and social history, which reduces future losses. Our supervised learning approach was employed in this experiment to examine the distinctions between former clients and clients who had defaulted on loan payments. This model will help a lender decide whether to approve or deny a loan application. Several algorithms and classifiers have been developed, and a preference has been given to one based on several metrics such AUC, F1, etc., in order to determine the combination of functions that results in the best overall measure of precision and class discrimination. The best number of components for prediction by cross-validation was determined using PCA, and the final features were found using cross-validation cross-recursive feature selection (CVEC) and subspace eversion (SV). This, by definition, enables us to use our scarce resources more effectively. Extreme Gradient Boosting Regression, Random Forest, and Support Vector Machines (SVM) (and general support vectors). A k-fold cross validation has been used to ensure that all possible combinations have been taken into account. To produce the best model output possible, Grid Search CV[was also employed. The outcomes were then shown in a table to show the most effective and trustworthy techniques for evaluating loan requests.

Keywords—Cloud Computing, Data Security, Data Classification, File Splitting Security.

I. INTRODUCTION

Through data, input in the form of interactions and observations from the actual world, machine learning enables computers to behave and learn like humans do [14]. The area of artificial intelligence research known as machine learning teaches computers through real-world interactions, which ultimately enables the computer to adapt to new environments [13]. In their

study, R.H. Davis et al. examined the use of machine learning algorithms in determining the risk associated with credit cards. Both neural networks and other classifiers were used to compare the accuracy results. The research was concluded with the idea that while the accuracy of all the algorithms tested was comparable, the neural networks' time complexity was higher [11].

However, the evaluation of loan borrowers' credit risk is the focus of this essay. There are more loan defaulters and charged-off loans than ever before. Transactions are being blocked, assets are being frozen, and lending institutions like banks and finance corporations are suffering significant losses. Around 9 million loan defaulters were said to be present in China alone in 2018 [28-29]. Since 2011, the number of defaulted loans in Bangladesh has nearly tripled [26]. In the United States, more than a million student loans default each year, and in the last ten years, the debt owed for education has increased to three times the original amount [27]. In the case of India, from the year of 2013 to 2017 the amount of money owed to banks by loan defaulters had quadrupled. The same study also informs us that default loans increased alarmingly at a rate of 27% in 2017. According to experts, Bangladesh's current situation will hinder business expansion and halt the implementation of numerous plans to create jobs for the general populace [19]. As a result, it is abundantly clear that loan defaults not only negatively affect financial institutions but also a nation's overall economy.

Carefully selecting the persons who merit a loan is a workable answer to this issue. Only those loan applicants with the lowest likelihood of defaulting should be chosen by banks or other loan providers. And this is where data science and machine learning really shine. In this situation, machine learning can be used to create a model that can recognize and learn from the behavioral patterns of both successful clients and loan defaulters. Based on the patterns it learnt in advance, the model can accurately estimate a new applicant's likelihood of defaulting on the loan. Credit institutions like banks or other lending firms can use this probability to determine whether or not to approve the applicant's loan request. The effectiveness of using machine learning algorithms for credit risk assessment has been shown in numerous studies. The authors of [21] also discussed the application of neural networks to the prediction of loan defaults. In this case, they examined three distinct artificial neural network models in nine different ways. The number of nodes in the input layer [3] and the output layer (1) of each neural network will be the same, however the number of nodes in the hidden layer (39, 4) will vary. For nine distinct assessment methods, they used nine different learning ratios. The ratio of train to test split is referred to as the learning ratio. In order to determine the ideal train test split ratio, the author divided their data set into nine different groups. The three neural networks listed above were compared in nine distinct scenarios, and the final analysis revealed that the neural network with 23 nodes in its hidden model was the most effective option. They further stated that a learning ratio of 4:6 produced the best results.

Numerous different supervised classifiers, including neural networks, have been applied in this field. The authors of the paper [17] have chosen to employ the J48 algorithm, Bayes net, and Naive

Bayes for this. The authors have made predictions on whether or not a new applicant will default on a loan based on factors like gender, history of prior credit, occupation, loan purpose, age, type of home, and credit amount. Their research indicated that the j48 algorithm, which had an accuracy of 78.3784 percent, was their top pick. SVMs, or support vector machines, are another well-liked solution for classification issues. In light of this, this paper will discuss the use of various supervised algorithms and feature selection techniques to determine the best investment for a lending institution by correctly predicting whether or not a new loan applicant will default on their loan or not. Both the applicant's personal history and credit history are included in our data set. The complex patterns that exist in the prior borrowers, whether successful or defaulters, will be identified using classifiers like extreme gradient boosting, support vector machines, random forests, and logistic regression. With this information, we can classify a new applicant into a defaulter or a non-defaulter category. To reduce computational expense and time, PCA and RFECV will be used to extract the ideal number of features and the precise features to use for the classification process. Later, 5 folds are crossed.

The ideal answer for each model will be chosen using grid search and validation. Finally, a comparison of each model will be done in order to choose the one that is best for assessing credit risk. The remainder of the essay briefly discusses some pertinent prior research in this area. Then, we'll go into great detail about our proposed model and data set. The steps of data pre-processing, the outcomes, and experimental analysis will all be covered in later sections. Finally, the paper will be concluded with a section on future work and closing remarks.

II. LITERATURE REVIEW

The topic of credit risk assessment is one that receives a lot of attention and discussion in the world of banking and financing. This subject has become even more important in light of the recent boom in data science and several significant developments in the discipline of machine learning. Numerous significant research findings have been made in this area, serving as a springboard for ongoing and future investigations. In this discipline, artificial neural networks are frequently employed.

The information processing paradigm known as an Artificial Neural Network (ANN) is modelled after how biological nervous systems, such as the brain, process information [1]. In [8], the authors employed a logistic regression model and an RBF multilayer feed forward neural network to compare their results. 492 test instances were included in their data collection and were gathered from Jordanian Commercial Banks. The regression model was shown to be more reliable at correctly identifying accepted applications, whereas the neural network has the upper hand when it comes to rejecting situations. Similar comparisons of these two models were also conducted in [6], where the authors evaluated the unsatisfactory consumers using the chi square test. The data set included 1,000 instances when the logistic regression performed better than the neural network.

In the realm of machine learning, the support vector machine is also a very common technique for categorizing.

Based on the data set and the desired computational time, research has demonstrated that a variety of additional strategies can be applied to obtain superior results. On a Brazilian bank data collection, the Gradient Boosting technique (GBM) is applied in [24]. Additionally, In this work, a generalized linear model and distributed random forest were both used. In this study, more than 20,000 examples were used, and 70% of the data set was used to train the model. It was shown that the GBM significantly outperformed the aforementioned approaches with an AUC value of nearly 98%. Another well-known classification algorithm that relies on Bayes' theorem is naive Bayes. In [5], the authors compared a Nave Bayes model against an LDA model, a k-nearest neighbour model (knn), a logistic regression model, a model based on classification trees, and a model based on neural networks. It was found that the Nave Bayes model performed poorly while the KNN model did the best. Despite the small and very inconsequential difference in scores, the authors contend that the Naive Bayes algorithm's low performance is largely due to the size of the data set. The dependencies of the categories also had a role in the Nave Bayes model's poor performance.

Classification and regression trees (CART), commonly referred to as decision trees in popular culture, are another common type of supervised learning method that is based on a tree structure. Internal nodes in the tree structure stand in for the data set's features, while related leaf nodes stand in for the target label. Such a model was employed by the authors in [20] to evaluate the credit risk of loan borrowers. The authors employed adaptive boosting to more heavily weight the less popular labels in order to address the issue of an unbalanced data set. The data set comprises account-balance-related information, transactional data, and credit bureau data. For a better understanding of the trends and underlying developments, the data set has been separated into various train and test sets. Results were assessed using instances from the following 90 days (3 months) for each training set, which included both successful and unsuccessful cases (3 months). Excellent results were obtained in this model using the performance metrics of precision, recall, accuracy, and AUC score. To obtain a more dependable and uniform score, ten folds cross validation was also carried out.

The data set was reduced from a starting count of 4520 instances after data pre-processing to 3271 instances. Information gain technique served as the attribute evaluator, and the default search algorithm for feature selection in Weka was ranked. Based on the size test set, accuracy values ranging from 85 to 90 percent were attained. The publication [4] also illustrates a situation in which tree-based models outperformed neural network-based models. In this proposal, the models of logistic regression, gradient boosting, random forest, and neural networks have all been compared. The algorithms were changed to improve accuracy and reduce computational expense. The elastic net approach was used to fine-tune the alpha and lambda hyperparameters for the logistic

regression model. This aids in improving accuracy and preventing "excessive regularization" of the LogR model [3]. For both the random forest and gradient boosting methods, 120 trees were used. Finally, four distinct neural network models were developed, each with a different number of hidden layers and regularization function values. The best settings for the drop out ratio, activation functions, hidden layers, and regularization functions were determined using a grid search and applied to one of the neural networks. AUC score and root mean square error (RMSE) were computed to evaluate the outcomes. The results revealed that the random forest and gradient boosting models, which are tree-based models, performed significantly better than the neural networks and LogR model.

III. GENERALIZED METHODOLOGY

Using supervised learning algorithms, the dataset from Lending Club will be used in our proposed model to assess credit risk. From this point on, the data set pre-processing phase will start. In order to handle categorical values, dummy variables afre first created, and the output label has been binarized. Currently, the desired output is either a "1" for "Fully Paid" or a "0" for "Charged Off." RFECV, a potential candidate for feature selection methods, and PCA, a member of the feature extraction family, will be used to perform dimensionality reduction. Prior to PCA and following RFECV, feature scaling will be carried out. To assess the model's performance, cross validation and a train test split will be used. Prior to that, however, SMOTE will only be used on the training set to address class imbalance. The hyperparameters will then be tuned using grid search with cross validation and fed into the classifiers for prediction. The predicted results will then be assessed using various evaluation metrics, and the findings will be presented in tabular and graphical form. The workflow that was used to create the suggested model is depicted in the flow chart below.

FIG:1 FLOWCHART OF THE MODEL

IV. CONCLUSION & FUTURE WORK

Support vector machines (SVM) or extreme gradient boosting models can outperform other tree-based or linear models in the discussion over which supervised learning model to adopt, if the experiment setup is comparable to ours. Furthermore, our model has shown that in the discussion about the optimal dimensionality reduction technique to employ, models based on recursive feature removal with a fivefold cross validation can outperform models based on PCA.

As we have previously indicated, computational cost is a concern in this situation. It should be noticed that the SVM model had required the longest training time. We would like to use Apache Hadoop in the future to build additional supervised models in order to shorten computation times. Additionally, we used data spanning the years 2007 through 2011. We want to use all the data to illustrate a better understanding of the trends in this field in order to make future improvements.

We have already mentioned that the idea of using neural networks has been dropped in order to lower computational costs. However, if we have the opportunity to work with more data in this area, we'd like to conduct a comparative analysis using neural networks as well. Since it is well-known that neural networks typically perform better when exposed to large amounts of data, we plan to use this hypothesis in our upcoming research. In order to assess their effectiveness, we would also like to integrate alternative dimensionality reduction strategies, such as evolutionary algorithms, univariate feature selection techniques, tree-based feature selections, etc. This is due to the fact that we are also talking about how feature selection and extraction strategies contribute.

V. REFERENCES

[1] network of neurons. Access the report at https://www.doc.ic.ac.uk/nd/surprise 96/journal/vol4/cs11.html.
[2] Alternative investments and peer-to-peer financing. https://www.lendingclub.com/ .
[3] (2017). (2017). elastic net with variable selection. Variable selection using an elastic net can be found at https://www.r-bloggers.com/ .
[4] P. M. Addo, D. Guegan, and B. Hassani (2018). employing deep learning and machine learning models for credit risk assessments. Risks, 6(2):38.
[5] M. Sfakianakis and A. Antonakis (2009). evaluating naive bayes as a tool for credit applicant screening. Applied Statistics Journal, 36(5), 537-545.
[6] G. V. Attigeri, M. Pai, and R. M. Pai (2017). assessment of credit risk via machine learning methods. 23(4):3649–3653 in Advanced Science Letters.
[7] G. E. Batista and M. C. Monard (2003). an examination of four supervised learning techniques for handling missing data. 17(5-6):519-533 in Applied Artificial Intelligence.
[8] H. A. Bekhet and S. F. K. Eletter (2014). Neural scoring approach for credit risk assessment model for Jordanian commercial banks. Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia, Review of Development Finance, 4(1):20-28.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote stands for synthetic minority over-sampling. Research in artificial intelligence journal, 16(3), 321-357.
[10] Benesty, M., Chen, T., He, T., et al (2015). Extreme gradient boosting, or Xgboost. Pages 1-4 of the R package version 0.4-2.
[11] Gammerman, A., Edelman, and R. H. Davis (1992). Algorithms for machine learning in credit card applications. 4(1):43–51 in the IMA Journal of Management Mathematics.
[12] Keerthi, S. S. and K.-B. Duan (2005). Which multiclass svm technique is the best? a statistical analysis. Nanyang Technological University, Nanyang Avenue, Singapore, pages 278–285. International Workshop on Multiple Classifier Systems Springer.
Faggella, David [13]. (2017). A discussion with Yoshua Bengio about how neural networks and deep learning are becoming more prevalent in our daily lives.
[14] D. Faggella (2018). An educated definition of what machine learning is. https://www.techemergence.com/what-is-machine-learning/.
[15] J. H. Friedman (2001). A gradient boosting machine for avaricious function approximation. Stanford University, USA, Annals of Statistics, pages 1189–1232.
[16] J. H. Friedman (2002). boosting using a stochastic gradient. Stanford University, Stanford, California 94305, USA, Computational Statistics & Data Analysis, 38(4):367-378.
[17] A. J. Hamid and T. M. Ahmed (2016). Bank loan risk prediction models are being developed utilising data mining. University of Khartoum, Sudan, Machine Learning and Applications: An International Journal, 3(1):1–9.
[18] Chang, C.-C., Lin, C.-J., et al., Hsu, C.-W. (2003). A helpful manual for support vector classification University of National Taiwan, Taipei 106, Taiwan
[19] S. Islam (2017). Banks are hampered by bad loans.
[20] A. E. Khandani, A. J. Kim, and A. W. Lo (2010). models for consumer credit risk using machine learning techniques. pp. 2767–2787 in Journal of Banking & Finance, 34(11).
Khashman, A.
[21] (2010). Neural networks for credit risk assessment: A comparison of several neural models and training approaches. Lefkosa, Mersin 10, Turkey: Expert Systems with Applications, 37(9):6233-6239.
[22] R. Kohavi and others (1995). a study of bootstrap and cross-validation for model selection and accuracy estimation. Volume 14, pages 1137–1145 of Ijcai, Stanford University, Stanford, California Canada's Montreal.
[23] Liaw, A., M. Wiener, et al (2002). Randomforest classification and regression. R news, 2(3):18-22.
[24] R. G. Lopes, R. N. Carvalho, M. Ladeira, and R. S. Carvalho (2016). estimating the return of credit activities at a Brazilian bank. Pages 780–784 in 15th IEEE International Conference on Machine Learning and Applications. IEEE.
[25] A. Mackiewicz and W. Ratajczak (1993). Analysis of the primary components (pca). Department of Mathematics, Technical University of Poznan, Piotrowo 3a, Poznan, Poland, 19:303-342, Computers and Geosciences.

[26] G. Mowla (2018). Banks are plagued by defaulted loans.
[27] Nova, A. (2018). Every year, over 1 million people miss payments on their student loans.
P. T.
[28] Deepak Singh Rajput, Saurabh Sharma, Shiv Kumar Tiwari, AK Upadhyay, Ashish Mishra (2020),"Medical data security using blockchain and machine learning in cloud computing".
[29] P.T.of India (2018). China has blacklisted 9 million loan defaulters and frozen 27 billion dollars.

.