Consensus Based Bank Loan Prediction Model Using Aggregated Decision Making and Cross Fold Validation Techniques

Ibrahim Hadiza Ndanusa
Department of Computer Science
Federal University of Technology
Minna Niger State, Nigeria
deezah_dans@yahoo.com

Solomon Adelowo Adepoju
Department of Computer Science
Federal University of Technology
Minna Niger State, Nigeria
solo.adepoju@futminna.edu.ng

Adeniyi Oluwaseun Ojerinde
Department of Computer Science
Federal University of Technology
Minna Niger State, Nigeria
o.ojerinde@futminna.edu.ng

2022 5th Information Technology for Education and Development (ITED) | 978-1-6654-9370-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITED56637.2022.10051450

Abstract-- Considering the growth of the credit businesses, machine learning models for granting loan permissions with the minimum amount of risk are becoming increasingly popular among banking sectors. Machine Learning based models has proven to be useful in resolving a variety of banking risk prediction issues. ML Predictions are sometimes unfair and biased because they are heavily dependent on randomly selected training data sample for every prediction made. However, this problem can be address by utilizing a cross-validation strategy. Prediction can be improved by combining decisions from different machine learning algorithms (ensemble decision making). The proposed consensus-based prediction model is evaluated using standard performance metrics, and the proposed model achieved an accuracy of 83 percent.

Index Terms: - ML, Machine Learning, **K-NN**, K- Nearest Neighbor **DT**, Decision Tree, **SVM**, Support Vector Machine, **LR**, Logistic Regression.

I. INTRODUCTION

In the event of a global recession, the banking sectors is prepared to respond. Financial institutions are more reliant on loan interest rates acquired from businesses in distress. Such institutions are having difficulty authorizing loans and dealing with current loan defaulters [1]. Banks must take appropriate steps to reduce credit risks in order to reduce costs as much as possible. Customer's assets in banks are likely to weather the current crisis without too much difficulty. It is possible to determine the default status of payment or credit score based on a customer's portfolio [2]. Access to centralized information about bank customers is beneficial to any financial company that obtains and manages consumer data in order to analyze borrowing, purchasing, and repayment patterns [3]

However, predictive machine learning models are thought to be effective at classifying data that has never been seen before into various classes. The predictive models work by learning from labelled observations to estimate the most appropriate category to which a data sample belongs [4]. As a result, such models are widely employed in a variety of industries, including the financial sector [5].

Machine learning (ML) systems are capable of detecting patterns in data and predicting complicated outputs in the face of high uncertainty [6]. In most cases machine learning algorithms parameters has to be fine tune to attain good prediction result (thus, hyperparameter optimization) [7]. Human rely on algorithm precision and accuracy to handle complex tasks in varieties of field, including medicine, finance, and law. In many circumstances, machine learning algorithm can outperform human, especially when working with huge datasets or a large number of input characteristics [8]. Predicting criminal recidivism based on previous conviction of a previous crime, it's an area where machine learning algorithm and expert systems might help human make better judgments [4].

This study primarily tends to address the issue related to biased decision making and improve on existing prediction model using aggregated decision-making approach. However, Tolan and other researchers [4] identifies that precision and prediction tend to be unfair and biased based on a randomly selected training data set for each prediction. Hence, this motivate the adoption of cross fold validation approach to address the identified problem.

Furthermore, this paper aim to develop an improved consensus-based bank loan predication model and approval system using cross-fold validation techniques and aggregated decision-making approach.

978-1-6654-9370-3/22/$31.00 ©2022 IEEE

II. LITERATURE REVIEW

Khandani [9] built a nonlinear consumer credit risk prediction model using machine-learning algorithm. The researchers generate augmented data samples to improve the classification rates of credit card holder, delinquencies and credit historical information's. However, an accuracy of 85% is achieved using linear regression model. The time series patterns of predicted delinquency rates from our model during the recent financial crisis show that collective consumer- credit risk analytics might be useful for projecting systematic risk.

Furthermore, Bank databases contain a lot of client information on a regular basis. Credit risk can be assessed using these datasets. The notion of feature selection is popularly adopted for reducing the size of such datasets. A multi-stage feature selection strategy is suggested by Abdi [10] to lower the dimension of an Iranian bank's database, which includes 50 characteristics. The elimination of interrelated characteristics is carried out in the first stage. The second step adopt genetic algorithm to choose the most relevant feature. The third step proposes that the variables are weighted using various filtering approaches. The fourth stage uses a clustering method to filter features. Finally, the K-Nearest Neighbor (KNN) and Decision Tree (DT) classification methods are used to classify chosen characteristics. The researcher goal is to forecast the risk of individual client based on the most effective and optimal subset of characteristics obtainable.

Addo [11] identifies that most financial or lending institutions are updating their financial models based on the high technology requirement such as data availability, big data, and computing power. Credit risk prediction, credit risk tracking, model integrity, and efficient loan preparation are essential in making successful loan approval decisions. Hence, the researcher utilize data collected from real world to develop a binary classification model using both machine and deep learning algorithm to forecasting loan default risk. The researcher filter 10 most relevant attribute for the modeling process to authenticate the binary classification model stability using the performance difference of the dataset. Hence, the researcher identifies that the tree-based machine model is more consistent than the deep neural network models. As a result, this present numerous questions about the widespread usage of deep learning systems in businesses.

Default rates on airtime loans are often minimal in comparison to those faced by the micro finance and banking institutions, but they are expected to improve as the service becomes generally available. Loan threshold methodologies are explored in this research, and the information's are used to approve lending of loan using machine learning model. More than 3 million credit information of 41 thousand consumers was examined with a three-month payback period. Dushimimana [12] uses multiple cross-validation methodologies, Logistic Regression, Decision Trees, and Random Forest to categorize defaulters. All the model was assessed and evaluated, and the latter model performed best. When the default rate is less than 2%, it is preferable to provide everyone a loan. The model improves profitability significantly when default rates are greater. The allowable amount of default rate for breaking even is quadrupled in the model, from 8% to 32%. Nonlinear classification models have a lot of potential for credit scoring since they can handle higher degrees of default and hence handle more consumers.

Ensemble method always performs better than model trained with single machine learning or deep learning algorithm.

Feng [13] proposed a weighted assemble based approach for scoring credit approval. The method proof that dynamic weighted ensemble approach can efficiently outperform benchmark models via experiment.

The development of credit rating system is essential for most financial bases institution.

Wigfield [14] proposed four varieties of hybrid models, which include classification and clustering prediction techniques. The study reveals that hybridizing logistic regression and neural network produce highest level of prediction accuracy.

III. AGGREGATED DECISION-MAKING AND CROSS FOLD VALIDATION METHOD

A. Aggregated Decision-Making System

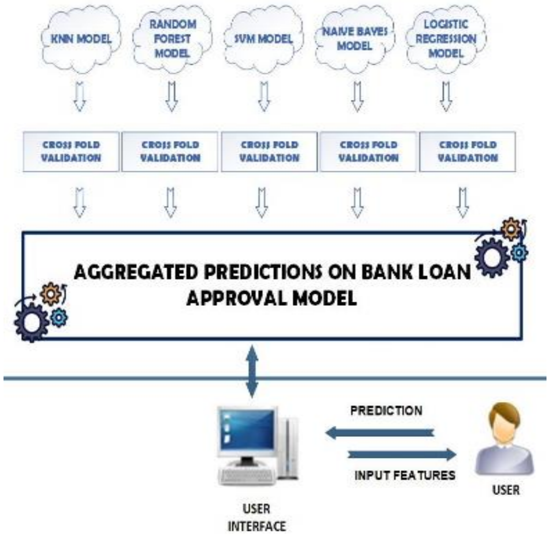


Fig 1. Conceptual Illustration of the System

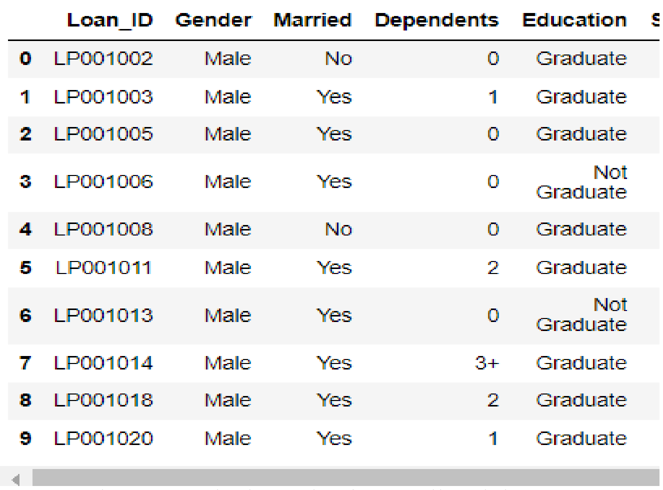
The Fig 1 illustrate the conceptual architecture of the proposed system using cross fold validation and aggregated decision-making model. As illustrated, user can interact (submit feature and receive prediction result) with the proposed system via a user interface. However, the model uses aggregation of cross fold validation result of each machine learning algorithm (KNN, Random Forest, SVM, Naïve Bayes, and Logistic Regression). Finally, each result is collected, and a consensus prediction is made on the adopted machine learning model using the aggregated decision-making approach.

B. Tools used

The model was developed using the Jupiter notebook IDE. Python programming language as the language of choice. Modules such as NumPy (array manipulations), Pandas (dataset exploration, preprocessing data, and data manipulation in a tabular form), and Sci-kit module for implementing various machine learning algorithm [15]. However, Microsoft Vision is used for designing various conceptual diagram and describing various aspect of the proposed system. The consensus base bank loan prediction system is implementation and design using python and KIVY programming language.

C. Data Collection

The loan approval or credit worthiness dataset was downloaded from Kaggle repository, an open data science resource. Kaggle is an online subsidiary of LLC Google that enables researchers to explore, develop, and publish datasets [16]. It also makes it much easier for data scientists and data engineers to construct models in a web platform environment [17] [18]. The fig 2 shows 10 data sample for exploration purpose.



	Loan_ID	Gender	Married	Dependents	Education
0	LP001002	Male	No	0	Graduate
1	LP001003	Male	Yes	1	Graduate
2	LP001005	Male	Yes	0	Graduate
3	LP001006	Male	Yes	0	Not Graduate
4	LP001008	Male	No	0	Graduate
5	LP001011	Male	Yes	2	Graduate
6	LP001013	Male	Yes	0	Not Graduate
7	LP001014	Male	Yes	3+	Graduate
8	LP001018	Male	Yes	2	Graduate
9	LP001020	Male	Yes	1	Graduate

Fig 2. Sample datapoint from collated dataset.

D. Methodology

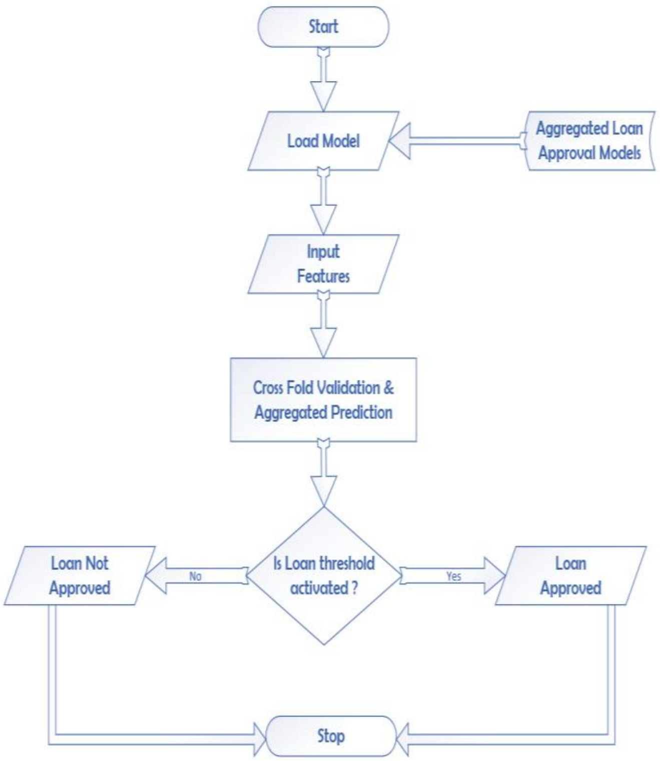
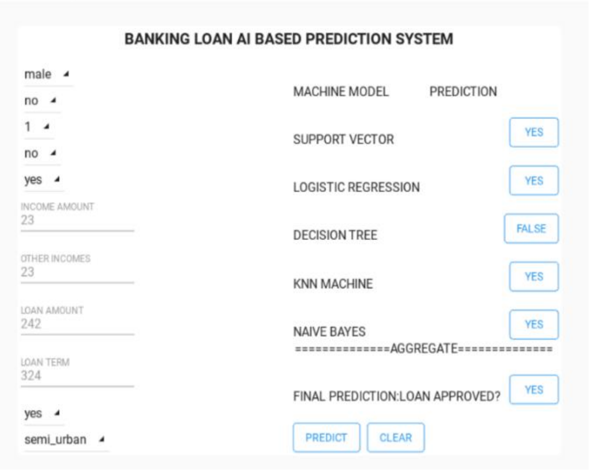

Start
Load Model
Aggregated Loan Approval Models
Input Features
Cross Fold Validation & Aggregated Prediction
Is Loan threshold activated ?
No
Yes
Loan Not Approved
Loan Approved
Stop

Fig 3 flow chart of the aggregated decision bank approval system

Fig 3 describe the data flow process of the consensus-based bank loan prediction model using cross fold validation and aggregated decision-making techniques. The first step includes data loading from an external data repository into system, then data feature can be inputted for predictions. However, the system can make collective decision based on cross-fold validation prediction from each machine learning algorithm. Thus, loan is issued only to customers that meet the loan approval threshold.

IV. EXPERIMENTAL RESULT AND ANALYSIS

This section presents the implemented system, model training, evaluation, and result from performance evaluation with the benchmark models. However, a comparative result from each machine learning model is also evaluated individually. Each model cross-fold validation scores and average mean score are compared. In addition, the accuracy of the final aggregated result is presented.



BANKING LOAN AI BASED PREDICTION SYSTEM

male
no
1
no
yes
INCOME AMOUNT 23
OTHER INCOMES 23
LOAN AMOUNT 242
LOAN TERM 324
yes
semi_urban

MACHINE MODEL	PREDICTION
SUPPORT VECTOR	YES
LOGISTIC REGRESSION	YES
DECISION TREE	FALSE
KNN MACHINE	YES
NAIVE BAYES	YES

=============AGGREGATE=============

FINAL PREDICTION:LOAN APPROVED? YES

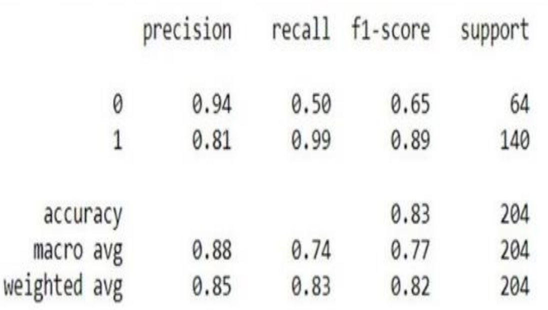
Fig 4 Consensus Based Bank Loan Prediction Model

The system was implemented using KIVY programming language to create the graphical user interface, this contains buttons, labels, menu buttons, and a textbox for the client to select and define data for new loan request. The system takes in the inputted data and perform classification using the pre-trained models (SVM, LR, DT, and Naive Bayes). Finally, the aggregated prediction model seeks consensus among the pretrained models before making final decision based on majority vote.

TABLE 1. K-Fold Score and Final accuracy Score (Mean)

S/N	Machine Learning Model	K-Fold Score Accuracy Scores	Final Accuracy Score (Mean)
1	Support Vector Machine	80%, 79%, 83%	80%
2	Logistic Regression	83%, 79%, 80%	80%
3	Random Forest	77%, 75%, 78%	77%
4	Naïve Bayes	78%, 79%, 82%	79%
5	KNN	68%, 68%, 68%	68%

Each model's K-fold validation score and final Accuracy score are compared in Table 1. Support Vector Machine and Logistic Regression perform best, according to the tabular result, the dataset is randomly sampled 3 times (shuffling the loan approval dataset into training and testing set in three instances). Consequently, the unbiased final accuracy (the mean of each random samples) of each machine model is safely utilized for the aggregated decision making.



```
              precision    recall  f1-score   support

           0       0.94      0.50      0.65        64
           1       0.81      0.99      0.89       140

    accuracy                           0.83       204
   macro avg       0.88      0.74      0.77       204
weighted avg       0.85      0.83      0.82       204
```

Fig 5. Propose aggregated prediction Classification report

In comparison to the various machine learning models examined in this work, the aggregated prediction model shows a better performance, with an overall accuracy of 83 percent.

TABLE 2. Comparative Analysis

S/N	Author/Year	Machine Model	Accuracy
1	(Motwani et al., 2018)	Linear Regression (LR)	79%
2	(Arun et al., 2021)	Decision Tree (DT)	82%
3	Proposed	Aggregated ML model	83%

Table 2 indicate the classification performance comparison between the proposed model and the benchmark model. Motwani [2] adopted Linear Regression model with 79% accuracy, Arun [20] adopt Decision Tree model for their classification with an accuracy of 82%. While the proposed model achieves an accuracy of 83% using an aggregated decision-making approach. Fig 6 visually shows the existing classification model approach and the proposed model with their corresponding accuracy score using the bar char diagram.

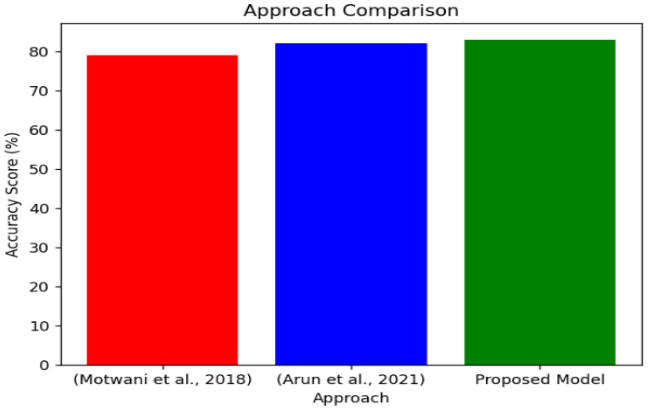


Fig 6. Pictorial Comparative analysis of models

V. CONCLUSION AND FUTURE WORK

The performance of any Machine Learning model highly depends on the training data samples been supply. Trusting the model performance will be difficult when the machine learning model is trained using all the available dataset for every data splitting and training, prediction accuracy tends to vary due to the randomly selected sample data for every data splitting (training and testing). Thus, this research introduces a cross-fold validation techniques to generate a random training and testing set from the main dataset. This result in different accuracy for every fold validation, and the mean accuracy is selected. However, the resulted accuracy can confidently depend on decision making. The aggregated decision-making uses the cross-fold validation result by multiple ML consensus prediction. Hence, this proposed approach performed better than the existing models.

Additionally, interested researchers can improve on this work by generating more data sample using sophisticated augmented techniques.

REFERENCES

[1] Moula, Fahmida E., Chi Guotai, and Mohammad Zoynul Abedin. "Credit default prediction modeling: an application of support vector machine." *Risk Management* 19.2 (2017): 158-187. https://doi.org/10.1057/s41283-017-0016-x

[2] Motwani, Anand, Goldi Bajaj, and Sushila Mohane (2018). Predictive Modelling for Credit Risk Detection using Ensemble Method. *International Journal of Computer Sciences and Engineering*, *6*(6), 863–867. https://doi.org/10.26438/ijcse/v6i6.863867

[3] Singh, Mukesh, and Gireesh Kumar Dixit. *Modelin g Customer ' s Credit Worthiness using Machine Learning Models : A Review*. International Journal of Scientific Engineering and Science, http://ijses.com/wp-content/uploads/2018/06/140-IJSES-V2N5.pdf

[4] Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019*, *June*, 83–92, https://doi.org/10.1145/

[5] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*(351), ALGORITMI Research Centre, Univ. of Minho, 4800-058 Guimar˜aes, Portugal22–31. https://doi.org/10.1016/j.dss.2014.03.001

[6] Al-Shiakhli, S. (2019). Big Data Analytics: A Literature Review Perspective. *Luleå University of Technology*, *1*(1), 1–57. https://www.diva-portal.org/smash/get/diva2:1320182/fulltext01.pdf

[7] Priyadarshini, I., & Cotton, C. (2021). A novel LSTM – CNN – grid search - based deep neural network for sentiment analysis. *The Journal of Supercomputing*, *0123456789*. https://doi.org/10.1007/s11227-021-03838-w

[8] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. tom.mitchell@cs.cmu.edu (T.M.M.) https://doi.org/10.1126/science.aaa8415

[9] Khandani, A. E., Kim, A. J., & Lo, A. W. (2012). Consumer Credit Risk Models Via Machine-Learning Algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1568864

[10] Abdi, F. (2021). A new multi-stage feature selection and classification approach: Bank customer credit risk scoring. *Journal of Industrial Engineering International*, *17*(1), 78–87. https://jiei.stb.iau.ir/article_684240_f3c1f63e431239 0c2e26c515189315f6.pdf

[11] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 1–20. www.mdpi.com/journal/risks https://doi.org/10.3390/risks6020038

[12] Dushimimana, B., Wambui, Y., Lubega, T., & McSharry, P. E. (2020). Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans. *Journal of Risk and Financial Management*, *13*(8), 180. https://doi.org/10.3390/jrfm13080180

[13] Feng, X., Xiao, Z., Zhong, B., Dong, Y., & Qiu, J. (2019). Dynamic weighted ensemble classification for credit scoring using Markov Chain. *Applied Intelligence*, *49*(2), 555–568. https://doi.org/10.1007/s10489-018-1253-8

[14] Wigfield, A., Eccles, J., & Rodriguez, D. (2013). Credit rating by hybrid machine learning techniques. 1568-4946/$ – see front matter ß 2009 Elsevier B.V. C.-F. Tsai, M.-L. Chen / Applied Soft Computing 10 (2010) 374–380. doi:10.1016/j.asoc.2009.08.003

[15] Jakhar, K., & Hooda, N. (2018). *Big Data Deep*

Learning Framework using Keras : A Case Study of Pneumonia Prediction. *December*. 4th International Conference on Computing Communication and Automation (ICCCA)Big https://doi.org/10.1109/CCAA.2018.8777571

[16] Salman, F. M., Abu-naser, S. S., Alajrami, E., Abu-nasser, B. S., & Ashqar, B. A. M. (2020). *COVID-19 Detection using Artificial Intelligence*. *4*(3), 18–25. International Journal of Academic Engineering Research (IJAER). http://dstore.alazhar.edu.ps/xmlui/bitstream/handle/123456789/587/IJAER200304.pdf?sequence=2&isAllowed=y

[17] Bhattacharya, Sweta, et al. "Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset." *Multimedia Tools and Applications* (2020): 1-25. DOI: https://doi.org/10.1007/s11042-020-09988-y

[18] Mangal, Ankita, and Nishant Kumar. "Using big data to enhance the bosch production line performance: A kaggle challenge." *2016 IEEE international conference on big data (big data)*. IEEE, 2016. https://arxiv.org/pdf/1701.00705.pdf

[19] Sulistiana, & Muslim, M. A. (2020). Support Vector Machine (SVM) Optimization Using Grid Search and Unigram to Improve E-Commerce Review Accuracy. *Journal of Soft Computing Exploration*, *1*(1), 8–15.https://shmpublisher.com/index.php/joscex/article/download/3/2

[20] Arun, K., Garg, I., & Kaur, S. (2021). Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering*, *18*(3), 79–81. https://d1wqtxts1xzle7.cloudfront.net/47121285/O1803017981