2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS) | 979-8-3503-0345-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICERA/ICIS59538.2023.10420221

Loan Default Prediction Using Machine Learning Techniques and Deep Learning ANN model

A.Lakshmanarao
Department of IT
Aditya Engineering College
Surampalem,India
a.lakshmanarao@aec.edu.in

Chabi Gupta
School of Commerce
Finance and Accountancy
Christ University,India
chabi.gupta@christuniversity.in

Chandra Sekhar Koppireddy
Assistant professor,CSE department,
Pragati Engineering College
Surampalem, A.P,India
chandrasekhar.koppireddy@gmail.com

U.V.Ramesh
Assistant Professor,Department of CSE
Aditya Engineering College
Surampalem, India
veerarameshu@aec.edu.in

D.Rajendra Dev
Assistant professor
Vignan's Institute of engineering for women
rajendra0511@gmail.com

Abstract— **Loan default prediction is a critical task in the financial sector, aimed at assessing the creditworthiness of borrowers and minimizing potential losses for lending institutions. Online loans continue to reach the public spotlight as Internet technology develops, and this trend is expected to continue in the foreseeable future. In this paper, the authors proposed loan default loan prediction system based on ML and DL models. This work makes use of the information on loan defaults provided by Lending Club. The dataset is preprocessed by applying various data preprocessing techniques and preprocessed dataset is generated. Later, we proposed four ML algorithms decision tree, random forest, logistic regression, K-NN and Feed forward neural network. The experimental results shown that proposed feed forward neural network achieved good accuracy for loan default prediction with an accuracy of 99%.**

Keywords— Loan prediction, Kaggle, Lending Club, Artificial Neural Network.

I. INTRODUCTION

Individuals and small companies need access to supplementary loan options, and it is crucial for online loan podiums to successfully decrease the credit emergency that is associated with customer loan defaults. People in every region of the globe are dependent, at least partially, on the ability of banks to offer them finance for a diversity of purposes, with assisting them in overwhelming their economic bounds and accomplishing their individual ambitions. The practise of obtaining a loan is now inevitable because of the dynamic nature of the economy and the ever-increasing level of competition that exists in the world of finance. In addition, financial institutions of all sizes, from the most modest to the most expansive, are dependent on the business of making loans to customers in order to generate profits, oversee their operations, and continue operating normally in the face of fluctuating financial conditions. In a society where people of all income levels need loans to reach their personal and professional objectives, it's not surprising that borrowing has become commonplace [2]. Lending money usually works out well for both parties involved. There is still the possibility of debt default, which poses a serious threat and might lead to a financial disaster. Therefore, it is crucial to determine whether a borrower meets the necessary criteria to be approved for a loan. The need to borrow money has become inevitable in today's financially challenging environment. Loans are used by both individuals and businesses for a variety of reasons, including the ability to attain objectives that would otherwise be out of reach financially and the ability to manage day-to-day operations when resources are limited. Although loan lending is advantageous for both the lender and the borrower, and is thus an integral aspect of financial transactions, it nonetheless involves substantial dangers. Credit risk, often known as default risk, refers to this risk. Traditional methods of assessment relied heavily on time-consuming and labor-intensive manual review. For this reason, financial institutions have increasingly turned to ML techniques for automated loan default prediction. In this paper, we applied four ML classifiers and deep learning model for loan prediction. The algorithms used in this work are described below:

DTR is a powerful ML model used for both classification and regression. In order to function, it divides the input data repeatedly into subgroups depending on the feature values provided, ultimately leading to a prediction or decision. The tree's interior nodes, each of which reflects a separate feature-based decision, and each leaf indicates a category label in the case of classification or a continuous value in the case of regression.The process of building a decision tree involves selecting the finest feature to fragmented the data at individual node, based on criteria such as Gini impurity or entropy for classification, mse for regression tasks. By partitioning the data into subsets, the algorithm aims to maximize the purity of each subset, ensuring that the resulting tree effectively separates different classes or predicts target values.

LGR is a fundamental statistical method in the field of ML for binary classification tasks. Logistic regression is mainly used for classification and not for regression. It is especially suited for scenarios where the dependent variable is binary, meaning it has two possible outcomes, typically denoted as 0 and 1.

In logistic regression, The method simulates the connection

979-8-3503-0345-2/23/$31.00 ©2023 IEEE

between the input characteristics and the likelihood of a binary result. It routines the logistic function to convert a linear combination of the input features into a probability score between 0 and 1. This probability score represents the likelihood that a given data point belongs to the positive class (1) in binary classification.

Random Forest is a popular and robust ensemble learning technique used in machine learning for both classification and regression. It is combination of several DTs, each trained on a separate sets of the data and making individual estimations. The final prediction from the Random Forest is determined by aggregating the predictions of its constituent trees, resulting in a more accurate and stable model. The Random Forest algorithm addresses some of the limitations of a single decision tree, such as overfitting and instability. By combining the predictions of multiple trees, it reduces the risk of overfitting and increases the overall predictive power of the model. Every tree in the forest is trained on a arbitrary subsection of the data (called as bootstrap aggregating or "bagging"), and a arbitrary subset of structures is considered for each split in the tree. This introduces randomness and diversity into the forest, leading to improved generalization on unseen data.

The k-Nearest Neighbors (k-NN) classifier is a simple yet effective ML model used for classification. It belongs to the category of instance-based learning, where predictions are made based on the similarity of input data points to their k-nn in the training dataset. The k-NN algorithm operates by comparing the input data point to all other points in the training dataset to identify its k-nearest neighbours. The "k" represents the number of neighbours to consider, which is a user-defined hyperparameter. Once the nearest neighbours are identified, the algorithm assigns a class label to the input data point based on the majority class among its neighbours. In other words, the predicted class is determined by the most common class among the k-nearest neighbours.

ANN, also called as a MLP, is a foundational and widely used architecture in deep learning. It is designed to model complex relationships in data and is particularly effective for tasks such as classification, regression, and pattern recognition. The network consists of numerous levels of interrelated nodes, or neurons, each contributing to the transformation and processing of input data. The structure of a feedforward NN was characterized by its layers: an input layer, one or more hidden layers, and an output layer. Data flows strictly in one direction from the input level, through the secret layers, and out to the output level. Neurons in each layer are connected to neurons in subsequent layers through weighted connections, and each connection has an associated weight that determines its influence on the network's output. During the forward pass, input data is fed into the network, and computations occur layer by layer. Neurons in a layer receive inputs, perform a weighted sum of those inputs, apply an activation function, and then pass the output to the next layer. Activation functions introduce non-linearity into the network, allowing it to capture complex patterns that linear models may struggle to represent. Forward and reverse propagation are the two fundamental phases of training a ANN. Data is sent into the network and predicted outputs are produced during forward propagation. Then, the model's predictions are compared to the actual target values using a loss function. Backpropagation calculates the gradients of the loss with respect to the model's weights, allowing the network to update its weights using optimization algorithms like gradient descent. This iterative process helps the network learn to make better predictions over time.

II. LITERATURE SURVEY

The authors in [1] present a comprehensive analysis of credit risk modelling for small and medium-sized enterprises (SMEs) in the U.S. market. They discuss the challenges and nuances associated with SME credit risk assessment and propose effective modelling techniques using financial ratios and machine learning algorithms. In [2], authors introduced a novel approach to credit-risk evaluation using neural network model. They demonstrate how these techniques enhance interpretability while maintaining predictive accuracy, a critical aspect in the financial sector. In [3], the authors explore the applicability of bankruptcy prediction models to microfinance institutions. They discuss the challenges unique to microfinance and evaluate the effectiveness of traditional credit scoring models in this context. In [4], the authors aim to provide an in-depth analysis of the research landscape surrounding loan default prediction using machine learning methods. They reviewed and discussed a selection of notable journal papers that contribute to the understanding, methodologies, and advancements in this domain. Through this survey, we aim to offer insights into the key approaches, challenges, and future directions in predicting loan defaults. In [5], the authors introduced a novel approach to credit risk prediction by combining the strengths of DNNs and SVMs in a cascaded architecture. The proposed cascade model leverages the feature learning capabilities of DNNs to extract high-level representations from credit data, which are subsequently used as input to an SVM for final risk assessment. Through extensive experiments and comparisons with standalone DNN and SVM models, we demonstrate the effectiveness and advantages of the cascade model in achieving enhanced credit risk prediction accuracy [6]. In [7], various ML algorithms applied for loan prediction and achieved good results. P.M. Addo [8] et al. investigated credit risk assessment using a deep learning approach. They explore the use of a ML and DL to extract features from credit-related data and showcase the potential of deep learning in enhancing accuracy of loan prediction.

III. PROPOSED METHOD

The proposed method is shown in figure 1. Initially, lending club dataset [9] collected. It consists of 55 features with 10,000 samples of data. The data contains some missing values also. First several data preprocessing techniques applied and a clean dataset was generated. It consists of 45 features in which one variable is dependent and all others are independent. Later, we applied two techniques for handling data imbalance. The techniques applied are smote and adasyn. In order to compensate for class imbalance in machine learning datasets, data augmentation techniques like SMOTE (Synthetic Minority Over-sampling Technique) are often used. In order to achieve a more equitable class composition, it creates fictitious samples for underrepresented groups. SMOTE helps improve the performance of machine learning model, especially in situations where one class is significantly underrepresented. ADASYN is a data sampling model designed to address class imbalance in machine learning datasets, similar to SMOTE.

ADASYN goes a step further by adaptively generating synthetic samples for the minority class, focusing on regions of the feature space that are more challenging to classify correctly. This makes ADASYN particularly effective in scenarios where the class imbalance is accompanied by overlapping regions between classes. After creation of preprocessed dataset, several ML classification methods are applied on dataset. Later deep learning ANN model applied for loan prediction.

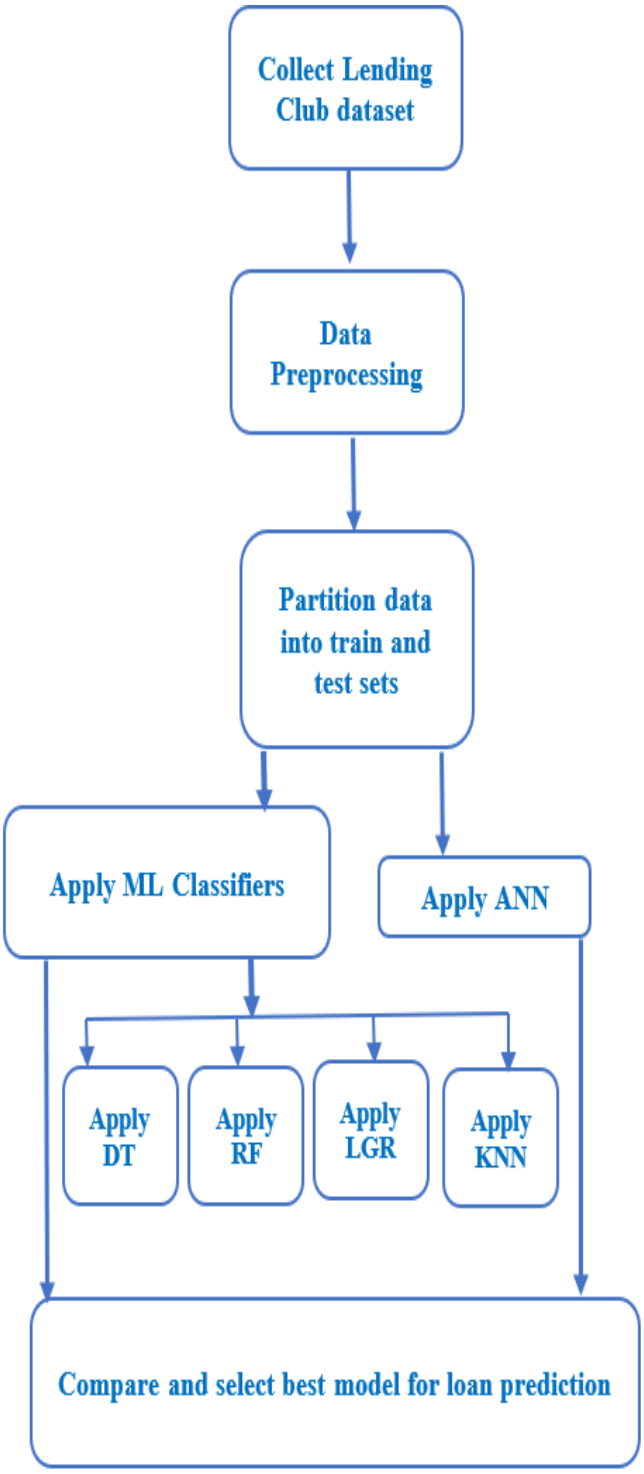

Fig. 1. Proposed Method

In the proposed system for medical cost prediction, the initial step involves gathering comprehensive medical data encompassing patient demographics, medical history, diagnostic codes, treatment specifics, and associated cost information. Later, data cleaning and preprocessing techniques are applied to check missing values, outliers, and inconsistencies. Numeric features are normalized or standardized, while categorical variables are encoded to prepare the data for further analysis. Later, careful consideration is given to the selection of ML and DL algorithms, based on the inherent characteristics of the problem. We applied four ML regression algorithms namely MLR, Gradient Boosting Regression algorithm, DTR and Random Forest Regressor. After applying ML algorithms, we also performed regression analysis with backpropagation for identifying the best features. Later, DNN applied for medical cost prediction and achieved good results. This approach proves valuable when attribute values exhibit discrepancies in their magnitudes.

IV. Experiments and Results

A. Data collection & cleaning

A loan dataset from Lending Club [9] is collected. It consists 10,000 samples. The dataset contains missing values. It is not possible apply ML algorithm with missing values. So, all the features with missing values are removed. The collected dataset also contains categorical features. ML algorithm needs input to be in numeric. So, all the categorical features are encoded using one hot encoding technique. After that the dataset ended with 133 features as input features and one feature as target variable.

B. Apply ML Techniques

After getting cleaned dataset, various ML classification methods applied. In this paper, four classifiers namely KNN, DTC, RF and logistic regression applied. But before applying these algorithms, we also applied smote and adasyn techniques for solving data imbalance issue. So, with smote and adasyn, we separately applied all the four classifiers.

Results with SMOTE:

Results of outputs with smote is shown in Table-I and Fig.2.

TABLE I. Results with SMOTE,ML algs

Alg	Acc(%)	Prec(%)	Recal(%)
KNN	89.6	91	90
DT	97.9	98	98
RFC	98.4	98.3	98.5
LGR	75.7	76	76

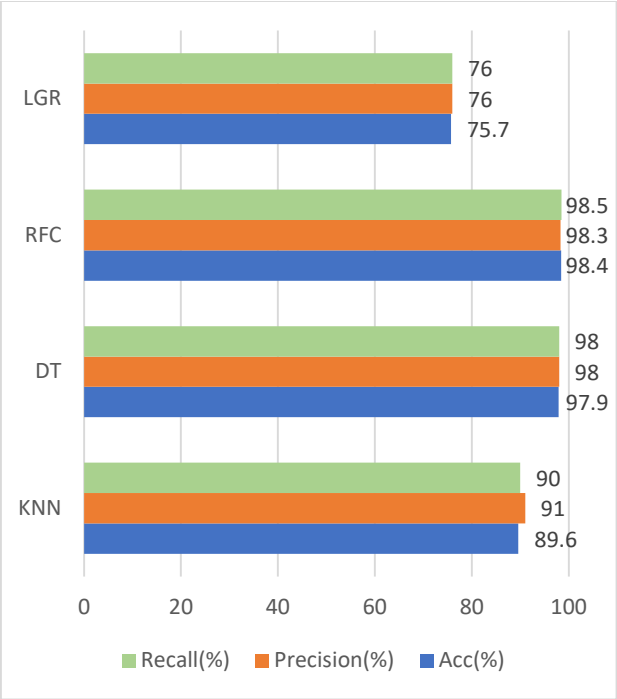

Fig. 2. Results with smote and ML Classifiers

After applying all the ML models, it is observed that Random Forest well among ML models. It had given accuracy of 98.4% for loan prediction. The precision and recall values are also good with RF. Next, Decision tree given an accuracy of 97.9. Later, KNN given an accuracy of 89.6%. Logistic regression was not able to perform well for loan prediction. It had given accuracy of 75% only. The precision and recall values also less for LGR.

Results with ADASYN:

Results of outputs with ADASYN is shown in Table-II and Fig.3.

TABLE II. RESULTS WITH ADASYN,ML ALGS

Alg	Acc(%)	Prec(%)	Recal(%)
KNN	89.4	91	89
DT	98.2	98	98.4
RFC	99	99	99
LGR	76.5	77	77

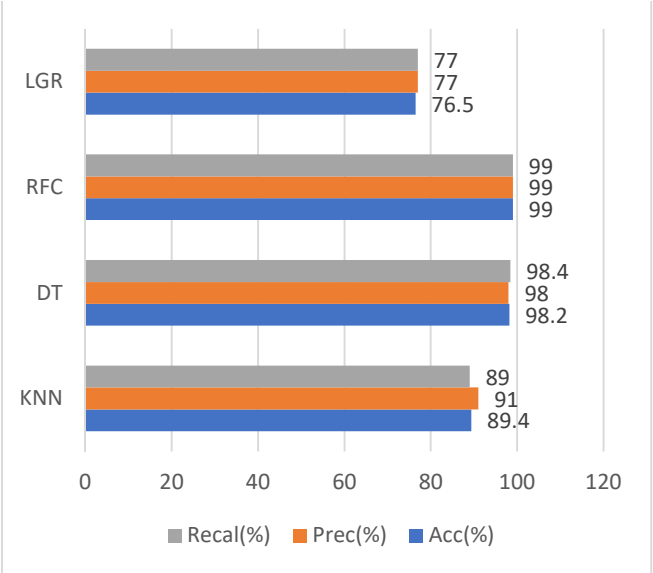

Fig. 3. Results with adasyn and ML Classifiers

After applying all the ML models with adasyn, it is observed that Random Forest well among ML models. It had given accuracy of 99% for loan prediction. The precision and recall values are also 99%, 99%. Next, Decision tree given an accuracy of 98.2%. Later, KNN given an accuracy of 89.4%. Logistic regression was not able to perform well for loan prediction. It had given accuracy of 76.5% only.

C. Apply ANN

The best accuracy achieved with ML classifiers is 99% with ADASYN and RF. To enhance, a deep learning ANN. The ANN model is shown in Fig.4. It contains input layer followed by four hidden layers followed by output layer. The no.of neurons in hidden levels are 260, 130, 60, 30 respectively. The output function is sigmoid as it is a binary classification problem. The activation function used in the hidden layers is relu function.

The accuracy achieved with ANN is 99.4%. It is better than accuracy of RF. The comparison of best ML classifier (RF) and ANN is shown in Fig.5.

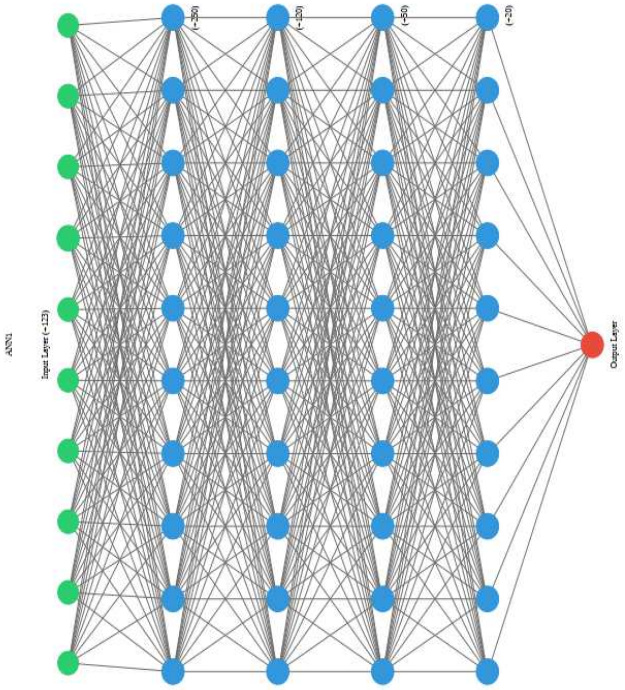

Fig. 4. Proposed ANN model

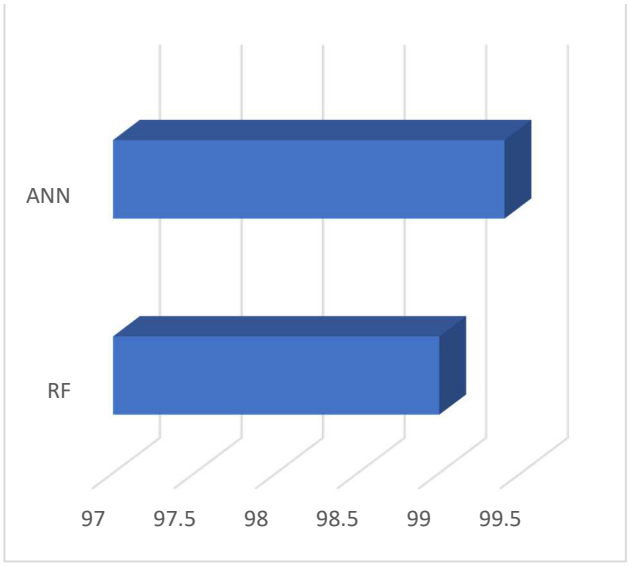

Fig. 5. Accuracy Comparison

V. CONCLUSION

In this paper, Loan prediction system is built with ML and ANN models. First a dataset from lending club was collected and several data preprocessing techniques applied. After that, two data imbalance techniques namely smote and adasyn applied along with four ML classifiers namely RF, DT, LGR and KNN. Among ML classifiers, RF with ADaSYN given best accuracy. To increase accuracy, a deep learning ANN with four hidden layers applied and achieved an accuracy of 99.4%. The results shown that proposed ANN performed well for load default prediction.

REFERENCES

[1] K Gomathyet al., “The Loan Prediction Using Machine Learning,” International Research Journal of Engineering and Technology, Volume: 08 Issue: 10 ,Oct 2021.

[2] Bart Baesens et al., “Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation.” Management Science 49, no. 3,2003.

[3] M. A. Sheikh et al.,“An Approach for Prediction of Loan Approval using Machine Learning Algorithm,” ICESC, Coimbatore, India, 2020.

[4] Kim, Hyeongjun & Cho, Hoon & Ryu, Doojin, “Corporate Default Predictions Using Machine Learning: Literature Review,” Sustainability. 12. 6325. 10.3390/su12166325.

[5] O. Awodele et al., "Cascade of Deep Neural Network And Support Vector Machine for Credit Risk Prediction," 2022 5th Information Technology for Education and Development (ITED), Abuja, Nigeria, 2022, pp. 1-8, doi: 10.1109/ITED56637.2022.10051312.

[6] Feng Shen et al.,“A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique, ” Applied Soft Computing,volume 98,2021,106852,ISSN 1568-4946..

[7] M. V. Rajesh et al., “An Efficient Machine Learning Classification model for Credit Approval,”International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 499-503.

[8] P. Addo et al.,“Credit Risk Analysis Using Machine and Deep Learning Models,” Risks, vol. 6, no. 2, p. 38, Apr. 2018, doi: 10.3390/risks6020038.

[9] https://www.openintro.org/data/index.php?data=loans_full_schema