Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis

Miraz Al Mamun
Department of Economics & Decision Sciences
The University of South Dakota
Vermillion, South Dakota, USA
miraz.almamun@coyotes.usd.edu

Afia Farjana and Muntasir Mamun
Department of Computer Science
The University of South Dakota
Vermillion, South Dakota, USA
afia.farjana@coyotes.usd.edu, muntasir.mamun@coyotes.usd.edu

Abstract

As people's demands grow, so does the need for bank loans. Every day, banks get many loan applications from customers and other individuals but not every applicant is accepted. Typically, banks execute a loan application after verifying and evaluating the applicant's eligibility, which is a time-consuming and challenging process. When examining loan applications and making credit approval decisions, most banks use their credit score and risk assessment systems. Despite this, some applicants fail to pay their bills each year, causing financial institutions to lose a substantial amount of money. In this study, Machine Learning (ML) algorithms are employed to extract patterns from a common loan-approved dataset and predict deserving loan applicants. Customers' previous data will be used to undertake the study, including their age, income type, loan annuity, last credit bureau report, Type of organization they work for, and length of employment. ML methods such as Random Forest, XGBoost, Adaboost, Lightgbm, Decision tree, and K-Nearest Neighbor were used to discover the maximum relevant features, i.e., the elements that have the most impact on the prediction output. These mentioned algorithms are compared and assessed against one another using standard metrics. Among these, Logistic Regression achieved the highest accuracy of 92%. It was also determined as the best model and performed significantly well better than other machine learning methods in terms of F1-Score, which is 96%.

Keywords
Loan Sanction, Machine Learning, XGBoost, Adaboost, Lightgbm.

1. Introduction

People prefer to apply for loans on the internet because data is growing daily due to digitization in the financial sector. Artificial intelligence (AI) is gaining popularity as a common tool for data analysis. Individuals from diverse businesses are using AI calculations to solve problems based on their sector knowledge. Banks are having a difficult time getting loans approved. Every day, bank staff are faced with a large number of applications to manage, and the odds of making a mistake are significant. Almost every bank's fundamental operation is the distribution of loans. The profit earned from the loans distributed by the bank's accounts. So, one mistake can make a massive loss to a bank (Gupta et al 2020).

The primary goal in the banking sector is to place their funds in safe hands. Many banks and financial institutions now grant loans after a lengthy process of verification and validation, but there is no guarantee that the chosen applicant is the most deserving of all applicants. We can forecast whether a given applicant is safe or not using our method, and the entire feature validation process is automated using machine learning techniques. Loan Prediction is extremely beneficial to both bank employees and applicants (Kumar et al. 2016).

© IEOM Society International

The purpose of this paper is to provide a quick, straightforward, and efficient method of selecting qualified applicants. It may provide the bank with unique benefits. The Loan Prediction System can calculate the weight of each characteristic involved in loan processing automatically, and the same features are processed according to their associated weight on new test data. The applicant can be given a deadline to determine whether or not his or her loan will be approved. The Loan Prediction System allows you to jump to a specific application and review it on a priority basis [2]. This approach allows you to jump on specific applications that deserve to be accepted first. Gender.

Married, Dependents, Education, Self-Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan Amount Term, Credit History, Property Area, and Loan Status are some of the features used in the forecast.

There are six subsections in this research report. We've looked at the literature survey in the next part. A brief overview of our dataset follows that. A machine learning approach is suggested in the next section. The algorithms that were employed to create the model were then presented. After that, there will be a quick discussion of the findings and analysis, followed by the conclusion.

2. Literature Survey

A prediction is an assertion about what one believes will occur in the future. Predictions are made all the time. Some are highly serious and based on scientific calculations, while others are simply guessing. Prediction aids us in a variety of ways, such as predicting what will happen after a period of time, a year, or 10 years. Predictive analytics is a branch of advanced analytics that analyzes current data and makes forecasts using a variety of approaches from data mining, statistics, modeling, machine learning, and artificial intelligence. Kumar Arun et al. (2016) studied how to forecast how a bank will approve a loan. They presented a model using machine learning technologies such as SVM and neural networks. This assessment of the literature aided us in carrying out our research and developing a reliable bank loan prediction model.

Mohammad et al. (2010) proposed a study to predict whether or not a bank would give a loan to a customer. The goal of the model was to achieve classification; hence using Logistic Regression with sigmoid function was used for developing the model. The dataset for studying and prediction was obtained from Kaggle and consisted of two data sets, one for training and the other for testing. To avoid missing values in the data set, the data has to be cleansed first. After that, performance measures including sensitivity and specificity were used to compare the models. The model produced an accuracy of 81%, according to the final results. The model was marginally better because it included variables (such as a customer's age, purpose, credit history, credit amount, credit duration, and so on) other than checking account information (which indicates a customer's wealth) that should be considered when calculating the probability of loan default correctly. As a result, by calculating the chance of default on a loan, the suitable customers to target for loan giving might be simply identified using a logistic regression approach.

Pidikiti et al.(2019) designed an effective model, the major goal of this paper was to lower the risk element associated with picking a safe individual to assign the loan in order to save time and money for the bank. There were four sections to this paper. (i) Data collection (ii) Machine learning model comparison using the data acquired (iii) System training using the most promising model (iv) Testing. They forecasted loan data using machine learning algorithms such as classification, logistic regression, Decision Tree, and gradient boosting in this paper. When compared to other algorithms, the decision tree method was found to be the most accurate, with an accuracy of 82 percent. It was successful because it produced improved results in classification problem. It was incredibly user friendly, simple to install, and provided interpretable results.

According to Pandey et al. (2010) predicting loan defaulters is one of the most challenging challenges for any bank. However, by predicting loan defaulters, banks can significantly reduce their losses by lowering non-profit assets. As a result, the research of loan approval prediction became crucial. In the prediction of this type of data, machine learning techniques are extremely important and useful. Four classification-based machine learning algorithms, namely Logistic Regression, Decision tree, Support vector Machine, and Random Forest, were used in this study, with the Support Vector Machine approach being the most accurate in predicting loan acceptance with a high accuracy of 79.67%. They gathered a list (dataset) of past client’s information from numerous banks who had backed a series of boundary advances.

Ndayisenga et al. (2021) contributed to work with commercial banks to predict the behaviors of borrowers by developing and testing the accuracy of different models using data from Bank of Kigali. The data was divided intotwo categories: training and test, with the training dataset accounting for 70% of the total and the test dataset accounting for 30%. Ensembles were utilized to discover the best machine learning strategies to apply for predicting bank loan default. Gradient Boosting (Accuracy 80.40 %) was shown to be the best model for predicting bank loan default, followed by XGBoosting, with decision trees, random forest, and logistic regression performing badly.

In Tejaswini et al. (2020) a robust predictive modeling method was presented to approve or reject loan applications based on the customers' historical financial and credit scores. The purpose of this paper was to create a quick, straightforward, and efficient method of selecting qualified applicants. The data was gathered from a variety of financial institutions. The training data set was provided to the machine learning model, and the model was trained using that data set. Every new applicant's information entered on the application form serves as a test data set. In this paper, they used three machine learning methods to predict client loan approval: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The testing results show that the Decision Tree machine learning algorithm has a higher accuracy of 82.00 % when compared to Logistic Regression and Random Forest machine learning techniques.

KUMAR (2016) developed a model for predicting whether or not a person will be approved for a loan. The main goal of this work was to see if a person could acquire a loan or not by analyzing the data with the help of decision tree classifiers, which provided 76.40% accuracy to forecast. Datasets were acquired from Kaggle and separated into two categories: existing customers and new customers. Every new applicant's information serves as a set of fact tests.

MADANE et al. (2016) constructed a model using the decision tree induction technique and attempted to analyze credit score of mortgage loans and applicant requirements. The credit score plays a role in loan approval. They built a model to predict if loan sanctioning is safe or not, and it was discovered that most low-income applicants are approved for loans because they are more likely to repay them. The dataset was gathered from online. The model they developed for bankers in this research would assist them anticipate the trustworthy individuals who have sought for a loan, boosting the likelihood of maintaining their loans on time.

The authors of Shrishti et al.(2018) proposed a robust machine learning model to predict loan approval. This model's major goal was to approve loans to applicants in a short amount of time. They used three types of machine algorithms: Logistic Regression, Decision Tree, and Random Forest. After reviewing the data sets for various models, it was discovered that the Random Forest algorithm had the highest accuracy of all the models.

A review on machine learning classification strategy for bank loan clearance was proposed by Karthiban (2019).. Almost all applications in today's world are influenced and controlled by machine learning algorithms. Despite the fact that a number of researchers are working on various machine learning algorithms, the algorithms' performance and precision remain a difficulty. They obtained data from a bank. This research looked at the performance of various classification algorithms in terms of precision, recall, and f-measure in order to predict whether or not a bank loan will be approved. Gradient Boosting outperformed all other classifiers in terms of classification matrices (accuracy, precision, recall and F-1 score) which showed 98.06% accuracy and F1 score was 99.20% in table 1.

Table 1. Bank Loan Approval prediction model performance analysis

Aurthors (year)	Dataset Collection (samples)	Applied Models	Measures (Proposed model)
Mohammad et al. (2020)	Kaggle (1500 cases)	Logistic regression [Proposed]	Accuracy: 81.00%

Pidikiti Supriya et al. (2019)	From previous customers of Bank (1000 cases and 7 numerical and 6 categorical attributes.)	Logic regression, Decision Tree [proposed model]and Gradient Boosting	Accuracy: 82.00%
Nitesh Pandey et al. (2021)	From past clients of different banks	Logistic Regression, Decision tree, Support	Accuracy: 79.67%
		Vector Machine (SVM) [proposed model] and Random forest	Precision: 46.00% Recall: 95.00% F1-Score: 61.00%
Ndayisenga et al. (2021)	Bank of Kigali	Gradient Boosting[Proposed model] XGBoosting Decision trees Random forest, Logistic Regression	Accuracy: 80.40% Precision: 82.59% Recall: 80.25% F1-Score: 81.00%
TejaswiniIn et al. (2020)	Financial Institution	Logistic Regression (LR), Decision Tree (DT) [Proposed model] and Random Forest (RF)	Accuracy: 82.00% Precision: 83.00% Recall: 82.00% F1-Score: 75.00%
KUMAR, SOURAV et al.(2021)	Kaggle data source	Decision Tree (DT) [Proposed model]	Accuracy: 76.40% Precision: 59.00% Recall: 79.83%
NIKHIL MADANE et al. (2019)	Online	Decision Tree (DT) [Proposed model]	Accuracy: 85%
Shrishti et al. (2018)	Kaggle	Logistic Regression, Decision tree and Random Forest algorithm [proposed model]	Accuracy: 89.22%

Karthiban, R.el at (2019)	Bank	Logistic Regression, Decision tree, Naive Bayes, Random Forest Deep Learning, Gradient Boosting [Proposed model], Generated linear model	Accuracy: 98.06% Precision: 99.10% Recall: 99.30% F1-Score: 99.20%

3. Proposed Methodology

Data collection is the first step in the suggested methodology and then we moved to the data pre-processing. Using the standard hold-out approach, the selected classifiers such as XGBoost, AdaBoost, LighGBM, Random Forest, Decision Tree, and K-Nearest Neighbor are then trained and tested on the provided dataset. To establish the best effective Bank Loan eligibility prediction method, the findings are computed and analyzed. Figure 1 depicts the overview of the proposed strategy.

A. Dataset Collection

In this paper, the provided dataset has been collected from the Kaggle online website. This dataset has 10,128 instances, and 23 attributes, whereas 1 class attribute and 23 attributes are predictive. Proper Bank Loan eligibility prediction is conducted appropriately using attributes, where the attributes describe the eligibility. The predictive 23 attributes are associated mainly with the information of a person's age, gender, educational background, ownership, properties, financial status, types of income source, credit card information, etc. and the class attribute is bank loan eligibility prediction.

B. Dataset pre-processing

Dataset pre-processing has been done by using feature extraction, data cleaning, missing values handling, and categorical variables transformation.

C. Validation process:

Selecting the appropriate validation process for a particular dataset is crucial. The hold-out validation process is one of the effective methods for getting the appropriate results [12]. We applied the hold-out validation process by holding 70% data on training and 30% data on testing. Using this validation process, we figured out the performance by confusion matrix and found the results of accuracy, precision, recall, area under curve (AUC) and F1-Score for every machine learning technique.

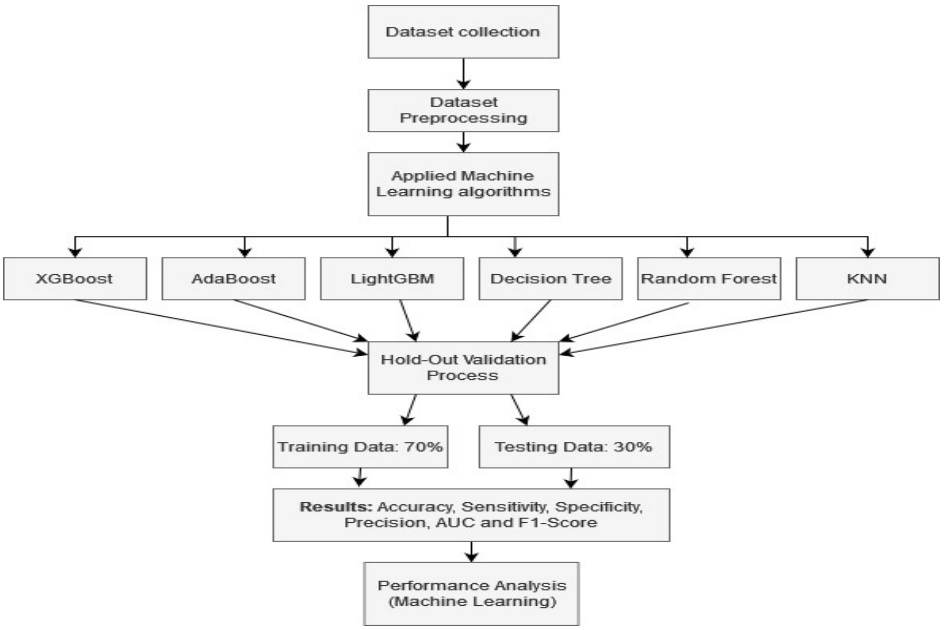

Figure 1. An overview of the study (Bank loan eligibility prediction)

4. Dataset Descriptions and Pre-processing

The bank loan prediction system dataset comes from the Kaggle competition and includes applicants of various ages and genders. The data set contains twenty-three attributes, such as education, marital status, income, assets, and so on, as shown in Table 2. There are total of 10,128 applicant records with the values of their relevant attributes in categorical and numerical data. We handle the missing value and normalize the data through pre-processing and feature engineering so that we may use it in an ML algorithm. The dataset is separated into two sections: training and testing. The model is trained using machine learning methods and forecasts the system using test data, as detailed in the following section.

Table 2. Some of dataset attribute names and information

Variable Name	Description of Variable	Data Type
Loan ID CLIENTNUM	Unique Loan ID	Integer
Customer_Age	Age of Customer	Integer
Gender	Male/ Female	Character
Dependents	Number of dependents	Integer
Married	Applicant married (Y/N)	Character
Education	Graduate/ Under Graduate	String
Income_Category	Income type	String
Card_Category	Card type	String
Self_Employed	Self Imployed (Y/N)	Character
ApplicantIncome	Applicant income	Integer
CoapplicantIncome	Coapplicant income	Integer

Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	credit history meets guidelines	Integer
Property_Area	Urban/ Semi Urban/ Rural	String
Loan_Status	Loan Approved(Y/N)	String

5. Result Analysis

Table 3 shows the average performance of selective machine learning classifiers such as XGBoost, LightGBM, Adaboost, Decision Tree, Random Forest, and KNN. After that, we looked at the results of the models in Figures 2 and 3. For observing the model's performances, we gave the results of Accuracy, Precision, Recall, F1Score, and AUC in table 3.

Table 3. Values of different measures for different machine learning classifiers for predicting the Bank loan eligibility

Model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-score**	**AUC**
XgBoost	0.9180	0.9223	0.4456	0.9969	0.9582	0.74
AdaBoost	0.9187	0.9217	0.4111	0.9976	0.9581	0.74
LightGBM	0.9189	0.9214	0.5316	0.9990	0.9586	0.75
Random forest	0.9188	0.9205	0.75	1.0	0.9586	0.70
Decision tree	0.8497	0.9252	0.1244	0.9088	0.9169	0.53
KNN	0.9167	0.9206	0.1400	0.9975	0.9575	0.54

Figure 2 shows that LightGBM has the highest accuracy score of 91.89 %, while Decision Tree has the lowest accuracy score of 84.97%. Furthermore, Random forest fared well with a score of 91.88 %. The results for XgBoost, AdaBoost, and KNN are 91%, 91.87%, and 91.67%, respectively.

© IEOM Society International

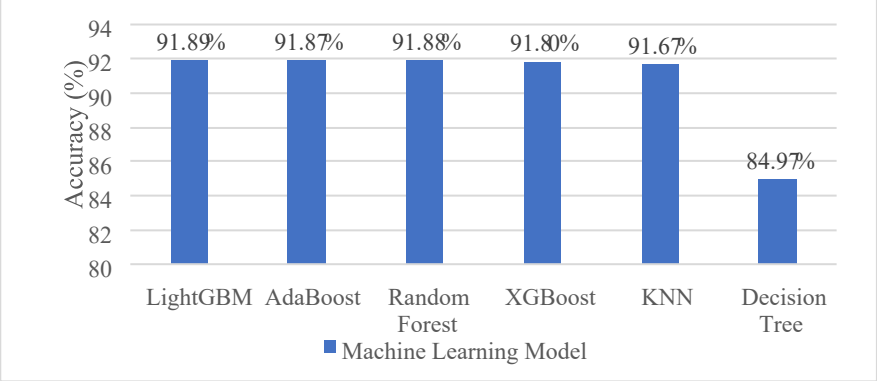

Figure 2. Accuracy analysis for predicting the Bank Loan eligibility using machine learning models

Accuracy, on the other hand, cannot be the only parameter used to assess model performance. As a result, the AUC value, which analyzes a model's ability to distinguish between classes, becomes an important metric for evaluating the model's performance. It's a probability curve that demonstrates how the True Positive Rate and False Positive Rate compare at different thresholds. The AUC assesses a model's ability to discriminate between positive and negative classifications. The AUC number should be as high as possible. The values vary from 0 to 1, with 0 being a fully inaccurate test and 1 representing a completely accurate test. AUC of 0.5 indicates no discrimination (i.e., the ability to distinguish a customer's eligibility probability or condition to get loan based on the test), 0.7 to 0.8 indicates acceptable performance, 0.8 to 0.9 indicates excellent results, and more than 0.9 indicates outstanding achievement for predicting the test. For the above machine learning models, we produced AUC graphs and mean results using holdout-validation in Figure 3.

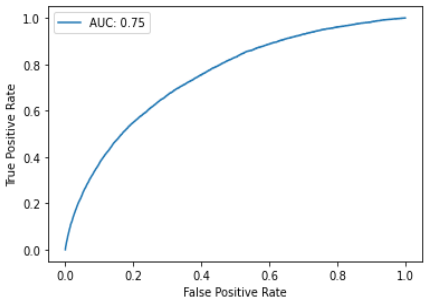

Figure 3a. LightGBM

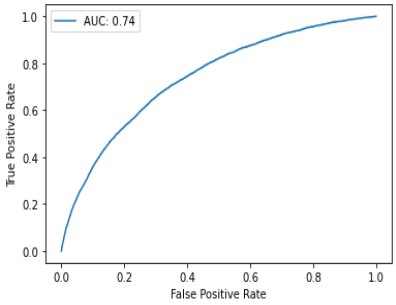

Figure 3b.Adaboost

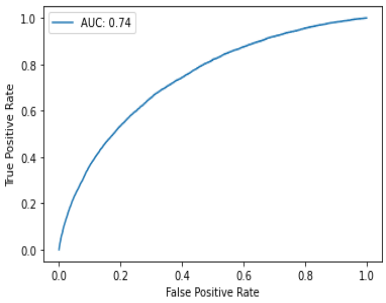

Figure 3c. **XGboost**

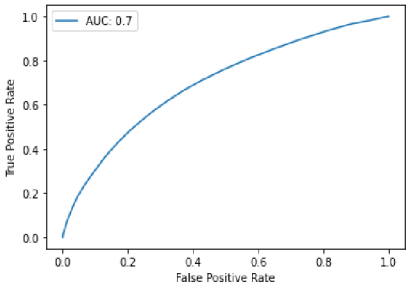

Figure 4. Random Forest

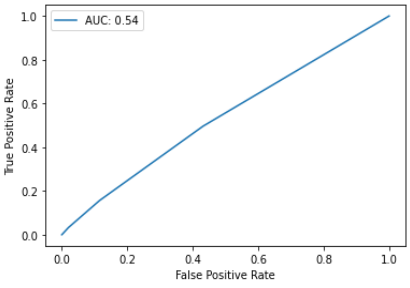

Figure 5. KNN

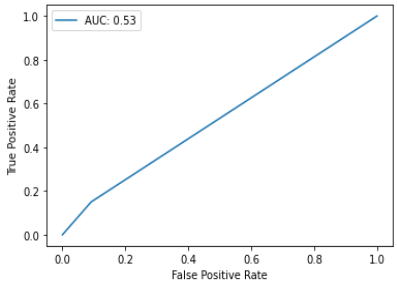

Figure 6. Decision Tree

Figure 3. Figure 4 and figure 5 , figure 6 Area under curve (AUC) output graph of LightGBM, XGBoost, Adaboost, Random forest, KNN and Decision Tree

As can be seen in Figure 4, LightGBM outperformed other machine learning classifiers in terms of AUC, which was 75 percent. XGboost and Adaboost both did well, scoring 74 percent, which is quite close to LightGBM's. (figure 7). Furthermore, AUCs of 70%, 54%, and 53% were reached using Random forest, KNN, and Decision tree, respectively. LightGBM outperformed other machine learning classifiers in terms of overall performance in terms of Accuracy and AUC.

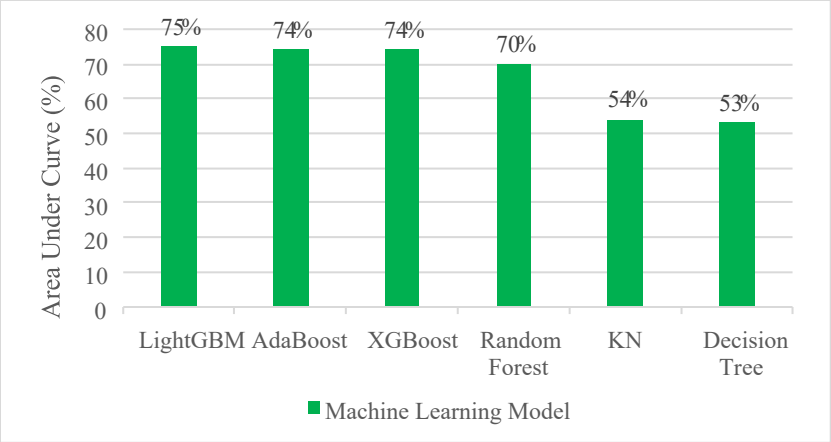

Figure 7. Area under curve (AUC) analysis for predicting the Bank Loan eligibility using machine lea rning models

6.Conclusion and Future Scope

Today's fast-growing IT sector requires the development of new technology and the updating of existing technology that allows us to eliminate human interference and boost job productivity. This model is used for the banking system or anyone who wants to apply for a loan. Based on the examination of the data, it is apparent that it reduces all frauds committed during the loan approval process. Time is valuable to everyone, and by doing so, not only the bank, but also the applicant's waiting time will be reduced. Cleaning and processing of data, imputation of missing values, experimental analysis of data set, model construction, and testing on test data are all steps in the prediction process. The best-case accuracy attained on the original data set is 0.9189 on Data set. After analyzing the data, the following conclusions were drawn: those applicants with the lowest credit scores will be denied a loan since they have a higher risk of defaulting on the loan. Most of the time, applicants with a high income and requests for a smaller loan are more likely to be approved, which makes sense because they are more likely to repay their debts. Other factors, such as gender and marital status, do not appear to be considered by the corporation. This prediction module can be enhanced and integrated in the future. The system is prepared on the previous training data but in the future, it is possible to make changes to software, which can accept new testing data and should also take part in training data and predict accordingly.

7. Reference

Gupta, Anshika, et al. "Bank Loan Prediction System using Machine Learning." 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, 2020.

Kumar, Arun, Garg Ishan, and Kaur Sanmeet. "Loan approval prediction based on machine learning approach." IOSR J. Comput. Eng 18.3, 18-21, 2016.

M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learni ng Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490-494, 2020.

Supriya, P. Usha et al. “Loan Prediction by using Machine Learning Models.” ,2019.

Loan Approval Prediction using Machine Learning Algorithms Approach. 2021 [Ebook]. Retrieved from https://ijirt.org/master/publishedpaper/IJIRT151769_PAPER.pdf

Ndayisenga, Theoneste. Bank Loan Approval Prediction Using Machine Learning Techniques. Diss. 2021.

Tejaswini, J., et al. "Accurate loan approval prediction based on machine learning approach." *Journal of Engineering Science* vol. 11, no.4, pp. 523-532. 2020.

KUMAR, SOURAV. "LOAN PREDICTION SYSTEM." 2021.

Nikhil Madane, Siddharth Nanda-Loan Prediction using Decision tree,Journal of the Gujrat Research History, Volume 21 Issue 14s, December , 2019

Shrishti Srivastava, Ayush Garg, Arpit Sehgal, Ashok kumar – Analysis and comparison of Loan Sanction Prediction Model using Python, International journal of computer science engineering and information technology research(IJCSEITR), Vol and issue 2, 2018

Karthiban, R. M. Ambika and K. E. Kannammal, "A Review on Machine Learning Classification Technique for Bank Loan Approval," *2019 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-6, 2019, doi: 10.1109/ICCCI.2019.8822014.

Yadav S. and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 78 -83, 2016,

Mandrekar, J. , Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal Of Thoracic Oncology,vol. 5, no.9, pp. 1315- 1316. 2010.

Biographies

Miraz Al Mamun is a recent graduate of the University of South Dakota (USD) and currently serves as Applications Support Analyst at Sanford Health under practical training. Sanford Health is a non-profit research affiliated organization with the University of South Dakota. At USD he studied Master of Science in Business Analytics. He earned bachelor's degree in business administration from North South University, Bangladesh. He also worked as a Graduate Research Assistant at the Beacom School of Business from 2019 to 2021. Under Beacom School of Business he had experience of working on different research projects on data analytics using machine learning methods. Currently, he is involved in several research utilizing machine learning models for business process automation and his research interests include financial fraud detection, customer privacy, customer churn modeling, customer sentiment analysis and business process automation.

Afia Farjana is a current graduate student of Computer Science at the University of South Dakota. She is working as a Research Assistant at the department of computer science and involved with different research project utilizing Machine Learning Algorithms. She completed Bachelor of science in Computer Science from American International University of Bangladesh (AIUB), Bangladesh. She has worked for sentimental analysis which is a survey on Machine learning for emotion and mental health detection, analysis, visualization using COVID-19 Social Media data. Apart from that recently she is doing her thesis related to federated learning on lung sound analysis. Her research interest includes data privacy, analysis of customer sentiment in business sector, image processing, pattern recognitions.

Muntasir Mamun is a Graduate student of University of South Dakota in Computer Science Department. Currently, he is working as Research and Teaching Assistant in University of South Dakota. He completed his bachelor's in electrical and Electronic Engineering at American International University of Bangladesh. However, he completed his research thesis and work on Covid-19 screening using Machine learning and Deep learning methods by cough sounds. This research work is already accepted in Springer Nature conference and another review work is accepted in peer J journal (impact factor:2.98). Currently, he is doing some research work on lung cancer and heart diseases predicting model using ensemble learning techniques. Apart from that, he has some other research publications in IEEE Xplore about Nanotechnology.