PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING ALGORITHMS

Chaitanya Kamlekar School of Graduate Studies

University of Central Missouri

Lee’s Summit, United States of America

[cxk12480@ucmo.edu](mailto:cxk12480@ucmo.edu)

Murali Yaswanth Motamarri

School of Graduate Studies

University of Central Missouri

Lee’s Summit, United States of America

mxm16470@ucmo.edu

Hirthik Rahul Siripurapu

School of Graduate Studies

University of Central Missouri

Lee’s Summit, United States of America

[hxs38450@ucmo.edu](mailto:hxs38450@ucmo.edu)

Uttam Kumar Gowda Matta

School of Graduate Studies

University of Central Missouri

Lee’s Summit, United States of America

uxm05810@ucmo.edu

***Abstract:* With the upgrade in the financial area, numerous people in their day-to-day life are applying for bank loans. Finding out who can be approved with a loan and who will be the most secure option for the bank is a frequent cycle, but the bank has limited resources that it must admit to specific people only depending upon their given information. Therefore, in this project, we experiment to reduce the risk of borrowers who would be unable to repay the loan and choose the ideal/valid person to preserve the assets of the banks. This is accomplished by obtaining information from the past records of those to whom credit has already been awarded, and based on these records and information, a machine was constructed using the AI model that produces the most trustworthy results.**

**The principal objective of this undertaking is to foresee regardless of whether appointing the advance to a specific individual will be protected. This venture is chiefly divided into four areas (i)Data Assortment (ii) Connecting the AI models on accumulated data (iii) Preparing of framework on generally encouraging model (iv) Testing. In this undertaking we are foreseeing the credit information by utilizing AI calculation. Even while the process of approving a loan is crucial, the actual decision-making process is complex and fraught with uncertainty. By using various statistical learning techniques, statisticians and data scientists have recently attempted to automate this process in an effort to reduce risk and boost profitability. The project experimented with various tree methods ranging from simplified decision trees to complex Random Forest. In terms of performance, relevance and interpretation, boosting performed better. The loan applications were approved depending upon the boosting-based decision tree models. The model's findings suggest that a bank should not just give loans to wealthy customers but also consider other customer characteristics that are crucial for determining credit decisions and identifying loan defaulters. With the help of decision trees, this work intended to create a high-performance predictive model for predicting loan acceptance.**

**Keywords: Machine learning, Decision Tree, prediction, Python, KNN Classifier, SVM Classifier.**

1. INTRODUCTION

Loans play an important role in the banking sector as they are the main source of revenue for the industry. The loans would be granted after a lot of verification and investigation. But they would not be able to predict if the loan granted is for a genuine person or if the loan would be repaid by the person. The approval process for the loans is done based on factors such as income and type of application. Even though the approval is done based on rigorous verification, there is still no guarantee that the chosen applicant is the most deserving candidate among all applicants**.** In the banking industry, a candidate is required to provide proof or backup before the credit amount is approved. Whether or not the application is approved relies on the system's analysis of the candidate's historical data. Multiple applications for loans are received every day, but banks can only provide limited money. In this

scenario, class-function algorithms would be very helpful in making the correct forecast. For the banking industry, debt recovery is of utmost importance and the improvement procedure is crucial here. This is a categorization issue where we must categorize whether the loan will be accepted. A classification challenge is one that involves predicting a class label for a specific example of input data. The primary goal of this study is to forecast the loan's safety. [1][2]. A machine learning model was developed using various classification methods using the historical data of the candidates. For example, KNN Classifier, Logistic Regression, Decision tree, Support Vector Machine Classifier, etc., Using machine learning models trained on historical data collection, the primary goal of this project is to predict whether a new applicant will be approved for a loan or not.

The goal of the current project is to automate the loan approval process using statistical learning techniques based on data about applicants from a bank. The main reason behind this is to minimize risk and allow banks to take proper decisions regarding loan approval. This effort sought to create a high-performance decision tree-based predictive model for loan approval prediction. In terms of simple decision trees, the results yielded were impotent due to correlated and complex feature space. However, using the testing dataset's importance chart scoring accuracy, boosting outperformed the control group in terms of performance, relevance, and interpretation. Due to this, boosting based decision tree models are the most recommended models in decision making of the applications based on their characteristics.

For classification issues, the popular and effective machine learning approach known as logistic regression is used. Logistic regression has the benefit of being a predictive analysis. It is employed to describe data and to clarify the connections between a single binary variable and one or more nominal, ordinal, and ration level variables. In this model, firstly data is cleaned so that the missing values are not included in the dataset. The dataset is divided into training and testing data and the engineering techniques are used to gain data analysis, where statistic concepts are applied. The understanding of the internal dependent and independent variables will come from studying the univariate, bivariate, and multivariate analyses [9][10]. A statistical machine learning technique known as logistic regression classifies data by considering the extremes of the outcome variables and attempting to draw a logarithmic line that distinguishes between them. Logistic regression can be used to predict in this fashion.

The KNN classifier algorithm is used to predict the class of the unlabeled data, a labeled training dataset with data points divided into different classes is provided. Different criteria are used in classification to identify the class to which the unlabeled data belongs. Most often, KNN is employed as a classifier. It is used to categorize data based on nearby or close-by training examples in a certain area. This approach is employed due to its speedy computation and ease of operation. KNN uses Euclidean distance to calculate nearest neighbors. The classifier can be run multiple times to validate the value of the K. It is known as lazy learning algorithm as the data set has to be stored and memorized when the data is being trained. It can also be used in the cases where there is no prior information about the dataset. In this method, we prepare a classifier using training data. When a new data is added into the set, it considers the K point near to the new data point, then using the neighbors classes, KNN decides which category the new data should be placed in. In this way, this algorithm increased clarity in more accurately classifying the data inputs into different categories. Because all calculations are done when classifying the training data rather than when it appears in the dataset, the computation cost is a little higher. Because it only stores the training data and memorizes the dataset when the dataset is being trained, it is a lazy learning method. On the training dataset, generalization is not done. Therefore, when in the testing stage, the whole training fundamental dataset is needed. KNN forecasts continuous values in regression. This number is the average of its K-nearest neighbors' values. K-nearest-neighbor classification was created to carry out characteristic analysis when it was uncertain or challenging to obtain clear parametric approximations of probability densities. KNN is more important for a study when the data being used has never been known before. The training and validation set must be separated from the initial dataset in order to obtain a favorable value of K. The result is uncertain if the two Nearest Neighbors belong to two separate classes. Therefore, the number of nearest neighbors is increased to the largest variable, which provides clarity. There are many variants such as Locally Adaptive KNN [4], Weight Adjusted KNN [5], KNN with K-Means [6], KNN with Shared Nearest Neighbours [10] which can make this algorithm more efficient. KNN and other models differ primarily in that KNN requires more computing time than other techniques.

The concepts of classification and regression are the core topics covered by support vector machines, which are an application of the supervised learning paradigm. One of the most popular algorithms for supervised learning, which uses datasets with features and class labels, is SVM. SVM chiefly has a place with the managed learning procedure which is planned as given in [11]. Finding the function f that minimizes the error as specified in [8], or a function that minimizes the expectation of error on the new data, SVM creates hyper-planes in a multidimensional space to demarcate various class boundaries; the feature vector of the dataset is the number of dimensions. It can handle the continuous and categorical variables [3]. There are two types of SVM, linear and nonlinear. In linear SVM, there are n training datasets (x1, y1 to xn, yn). Between the two classes of data, a classifier with a wide margin will be found. SVM plays an important role in Non-linear type, later on with the help of kernels, it was used for nonlinear classification. It uses kernels such as Polynomial, Sigmoid and Radial basis function kernels. When compared to other algorithms, SVM provides the results that are the most accurate and optimized in its domain. Thus, the approval of loan prediction can be done easily by using the above-mentioned algorithms as it is the key factor in the banking sector.

1. Motivation

The loan's acceptance is crucial in the banking industry. Banks receive multiple applications every day. Accepting or rejecting the loans is an important activity for banking organizations, and loan recovery is an essential factor in a bank's financial performance. Even though the approval is done based on rigorous verification, there is still no guarantee that the chosen applicant is the most deserving candidate among all applicants**.** A machine learning model will be helpful in categorizing the right candidate to receive a loan. Predicting whether a consumer will pay back a loan is exceedingly tough. Also at the same time, a genuine person who is eligible for the loan has to be granted the loan within the given amount of time, as it leads to the faith of the person in the bank, which directly reflects the business of the bank. Thus, loan approval plays a crucial role in every bank application. Evaluating the loan applications manually is very difficult as it requires a large manpower and a lot of time, using this process, the loan applications can be verified correctly based on certain factors and loans can be approved in the stipulated time. It also helps in increased customer satisfaction, which plays a major role in running any business successfully in the present day. However, to reduce the risk of loan default, the benefits can only be realized if the bank has a reliable model to anticipate which loans from customers it should approve and which to reject. To make this happen, precise calculation and approval is required, which can be achieved using machine learning.

1. Main Contribution and Objectives

The following are the contributions and objectives achieved in this project.

This project aims to provide a quick, easy, and effective method of choosing eligible applicants. This expedites the loan approval process. The primary goal of this project is to approve loans for the right candidates so that the genuine person will not be affected by the approval procedure and do not get declined for the loan. The dataset [12] will be divided into two sets, training and testing respectively, Models are developed using the training data, and their performance is evaluated using the testing data. By using this project, we can predict if an application can be approved or not for the loan.

*i.) Analysis*: Analyzed several research papers and gathered relevant information from the papers. Got to know the advantages and disadvantages from the previous research works and proposed methods to overcome the drawbacks.

*ii.) Collection of Datasets*: This involves gathering and preparing an appropriate dataset for developing and testing machine learning models. The dataset must contain various attributes essential for loan approval like gender, marital status, education, employment, property area etc. Clean and prepare the data and develop the proper feature sets.

*iii.) Selection of Model:* This includes choosing appropriate models and implementing them on the training and testing data. In this project we choose “Logistic Regression”, “Decision Tree”, “Random Forest”, “SVC”, “K Nearest Neighbor”.

*iv.) Implementation*: During implementation the dataset is divided into testing and training datasets. Calculating metrics like “Accuracy”, “Precision”, “Recall”, “F1 score”, evaluating the models performance using these metrics and improving the models' hyperparameters for performance.

*V.) Analysis and Comparing various Models*: Using the metrics like “Accuracy”, “Precision”, “Recall”, “F1 score” we can get the advantages and disadvantages of various models we use and evaluating the best model for predicting the loan approval.

*vi.) Presentation and Visualization of results*: visualization includes presenting the output of the various algorithms in the form of charts, graphs, tables etc. The presentation and visualization of results are crucial components of any research or analytical work. The objective of results visualization is to concisely and clearly express the conclusions and insights drawn from the data analysis.

1. Related *work*

To determine whether to authorize a loan for a certain borrower, certain standards need to be considered by all money lending organizations. The decision-making criteria need not be limited to a single attribute; they might include any number of attributes that must be taken into consideration.

The features contained in this dataset will be used to build a model that will determine whether a loan should be accepted for a particular borrower. There are two possible outcomes: approval and rejection.

The built model must draw conclusions faster than would be possible if the decision were made manually. The model should be precise enough to meet the requirements of the loan providers [14].

Numerous institutions have been processing loans up to this point using paper and pen. When many clients seek bank loans, these banks take a long time to approve their loans. There is no guarantee that the selected applicant will be able to repay the loan once it has been approved by the banks. For the loan approval process, many banks employ their own software. The current system relies on data mining techniques, which is an outdated method for loan acceptance. Different machine learning methods are used to get results after combining multiple data sets to create a generalized dataset. These methods, however, fall short of expectations. As a result, large banks are experiencing financial crises. We present a novel method for loan approval to solve this problem.

The banking industry uses a software program called a loan approval system to approve loans. We have a machine learning algorithm that we have employed in this proposed system. In the machine learning process, an identical model is created from the previous dataset and used to evaluate the new dataset. The system comprises a trained dataset and a test dataset. The model is built using the trained dataset. This model is used on a testing dataset to produce the desired outcome. For the model's construction, we employed the ensemble approach.

Here in this section, we discuss the related work where machine learning techniques are used in financial applications [13].

*a.) Risk of Corporate Credit*: Majority of a bank's lending business comes from corporations. Banks will be seriously impacted once default happens because of the large average loan made to corporations as clients. As a result, a lot of study on the risk of corporate credit was done in the past, among which some of used machine learning techniques.

Support vector machines (SVM) are used to the corporate credit rating problem in [16], and the author uses a grid search strategy with 5-fold cross validation to determine the ideal parameter value for the RBF kernel function of the SVM. To determine the prediction accuracy of SVM, the author also contrasts its performance with that of “multiple discriminant analysis” (MDA), case-based reasoning (CBR), and three-layer fully connected “back-propagation neural networks” (BPNs). The outcomes demonstrate that SVM is superior to other current approaches.

In [17], the authors carried out analytical study on the performance of four learning algorithms – “single hidden layer feedforward networks” (SLFN), “Backpropagation” (BP), “Extreme learning machine” (ELM), “Incremental extreme learning machine” (I-ELM). The outcomes proved that SVM performs better on output distributions, whereas SLFN approach is highly reliable when compared with SVM.

In [18], the authors put out a fresh approach to multi-classification that relies on the fuzzy clustering algorithm and support vector domain to complete corporate credit rating.

*b.) Risk of Consumer credit:* To anticipate and manage risk for various loans, including consumer loans and mortgage loans, numerous models are needed: The rise of the big data era has made it possible to conduct research on consumer credit risks.

Convolutional neural networks are used by the authors of [19] to forecast mortgage defaults using consumer transaction data. Convolutional neural network and random forest classifier models can both achieve ROC AUCs of 0.98 and 0.926, respectively, in the absence of any further data about each customer.

In paper [20] the author proposed “Gradient Boosting Survival Tree (GBST), to process heterogeneous data gathered in Chinese consumer market.

The authors of [21] offer a standard methodology for evaluating a consumer's credit risk using machine learning techniques. The dataset of complete payment histories for short-term instalment credits is analyzed using machine learning techniques as well as optimized logic regression.

*c.) Risk of Person-to-Person Lending:* Peer-to-peer (P2P) lending has long been a prominent application of innovation in the financial sector. However, P2P lending is one of the industries where defaults happen most frequently due to incomplete personal information, a lack of risk control, and a rapidly expanding business. There have been numerous attempts to use machine learning models to carry out studies related to the credit risk of P2P lending to promote the healthy development of P2P and lower default rates.

In [22], the authors compare the conventional machine learning technique with the CatBoost algorithm using loan data from Australia's P2P network. The examination of the AUC value and accuracy of these algorithms as evaluation indicators shows that CatBoost performs better and is more accurate at credit scoring than typical machine learning methods.

In [24], the authors use real P2P transaction data from Lending Club and the cutting-edge machine learning algorithms LightGBM and XGBoost to assess the outcomes of various approaches to creatively anticipate the chance of loan default. LightGBM is the best, according to the results of the classification prediction using multiple observational data sets.

To categorize P2P loan applications into groups of default and non-default, the authors suggest a credit scoring model using artificial neural networks in [23]. The results show that the neural networks-based credit scoring models are very effective.

1. Proposed Framework

For determining whether a person is eligible for acquiring a loan or not, we have used the machine learning algorithms such as logistic regression, KNN classifier, Radom Forest classifier, support vector machine, decision tree classifiers determine the eligibility of an individual. The implementation includes gathering the data, cleaning the data, training the model with the training data, and predicting the testing data against the models, and we have evaluated the quality of the prediction using the quality metrics such as F1 score, accuracy, precision, recall. The flow of the project is illustrated in the figure below fig.1. It involves collection of data, feature extraction, implementation and then evaluating.

Diagram

Description automatically generated

Fig.1 framework.

*a.) Feature Extraction*: a legitimate bank customer who can repay his loan we have features such as good credit history, annual income, marital status etc., The other features in our dataset are as below:

1. *Loan ID*: This is the id generated by the Bank for a loan request.

2. *Gender*: This attribute gives information about the gender of the loan applicant.

3. *Married*: It indicates the marital status of the applicant.

4. *Dependents*: it shows the number of dependents of the primary applicant.

5. *Education*: It is a binary variable which indicates whether the primary applicant has completed his high school.

*6. Self-employed*: It is a Binary variable indicating whether the primary applicant is self-employed.

*7. Applicant Income:* It shows the income of the primary applicant.

*8. Co-Applicant Income:* It shows the income of the co-applicant.

*9. Loan Amount*: It is the loan amount that the applicant wants to borrow from the Bank.

*10. Loan Amount Term*: The period during which the borrower would repay the loan.

*11. Credit History:* It is a Binary variable which tells whether the applicant had a good history or a bad history.

*12. Property Area*: It is a Categorical variable which shows whether the applicant is from an urban, semi urban, or a rural area.

*13. Loan Status:* This is the final output attribute which indicates whether the loan is approved.

*b.) Feature Analysis:* while analyzing the data we have seen that some of features are categorical which are later converted to which we have later converted to Boolean data so that it can be easier for computation. Using correlation, we have determined the features which have high impact or effect on the target variable we have choose hyperparameters accordingly and train our models accordingly to get higher accuracy, precision values.

*c.) Random Forest Classifier*: Random Forest Classifier uses several decision trees to make a reliable prediction. A group of decision trees are trained using subset of input features and data samples and the prediction made by each tree is considered as a vote. The class which gets more votes will be assigned to the data point. The hyperparameters such as depth of each decision tree in a random forest classifier or the number of decision trees in a random forest classifier play a major role while determining computational time and the precision of the model. The prediction using Random Forest classifier is increased because of the combination of various decision trees.

*d.) decision tree:* The decision tree algorithm is based on divide and conquer methodology, the features which have high information gain made as root node and a subtree is drawn from it. The node in the subtree the which high information game selected as a root and the subtree will be drawn as long as subtree completely containing pure class The feature which allows them division of data to subclass which contains homogeneous labels is the main objective off this algorithm.

e) KNN algorithm: KNN algorithm is a productive machine learning algorithm which is majorly used for predictive classification. kNN algorithm classifies the training data into groups or classes, the newly inputted data or the testing data, based on its similarity with previously trained data, the k nearest neighbors around the new data will be considered and if the newly inputted data share mores number of neighbors with a particular class, then that class label will be assigned to the test subject. The hyper-parameters in KNN algorithm is k-value and the distance metric. The value of k will determine the model overfits or underfits the given data.

1. Data Description

The dataset consists of testing and training datasets. The training dataset has 614 data points with 11 features for each observation. Each data point will be labelled as positive or negative classes, where positive class means that the candidate is eligible for acquiring the loan and negative class means that the individual is not eligible for the loan. Positive class means that the credit company can trust the individual that he can repay his loan. We have obtained the data from KAGGLE website using the following link [12]. The data set contains training and testing data separately. We have two different datasets for the training and the testing data.

1. Results & Analysis

The dataset [12] is taken from Kaggle. The dataset consists of two csv files, train.csv and test.csv. The train data consists of 615 rows and 13 columns. The test data consists of 368 rows and 12 columns. The train.csv has the final attribute ‘Loan\_Status’ which is not available in test.csv.

We start with the test-train split when training the model. The training and test sets are randomly selected from the dataset using cross validation by specifying the number of folds. Following the division of the dataset, the model is trained on the training set and its performance is evaluated on the test set. Some of the metrics used to judge the model's performance include accuracy, precision, recall, and F1 score. The following points have been raised: One Vs One SVC Classifier is used for Support Vector Machine to overcome the multi class functionality because the target variable belongs to four different classes, and Multinomial is specified as the multi class argument in Logistic Regression.

A confusion matrix provides a highly detailed breakdown of the number of true positives, true negatives, false positives, and false negatives for a certain classification task. The true positives (TP) measure the proportion of instances that are correctly identified as positive, whereas the true negatives (TN) measure the proportion of instances that are correctly classified as negative. False positives (FP) are circumstances that are categorized wrongly as positive, whereas false negatives (FN) are situations that are classed incorrectly as negative. The sum of the counts in each row and column yields the total number of instances for each real class, while the sum of the counts in each row and column yields the total number of instances for each predicted class. A confusion matrix, which aids in evaluating the efficacy of a classification approach, offers information on the model's accuracy, precision, recall, and F1 score.

1. *Accuracy:* The percentage of cases that are correctly classified out of all instances is called accuracy. In other words, it assesses the frequency with which the model predicts correctly given all predictions.

Accuracy = *TP* + *TN*

*P* + *N*

where P represents the total number of instances in Positive class and N represents the total number of instances in Negative Class, which together give the total instances in the data set.

1. *Precision:* Precision is the percentage of cases correctly categorized as positive (true positives) among all instances the model classified as positive. In other words, it assesses the frequency with which the model predicts positively given all positively predicted cases.

Precision = TP

TP+FP

1. *Recall:* The fraction of actual positive incidents that are true positives is known as recall. It assesses the model’s ability to recognize positive cases given all actual positive instances.

Recall = TP

TP+FN

1. *F1 score:* The harmonic mean of recall and precision is the F1 score. It offers a single statistic and strikes a com- promise between the trade-off between recall and precision.

*F* 1*score* = 2 ∗ *Recall* ∗ *Precision*

*Recall* + *Precision*

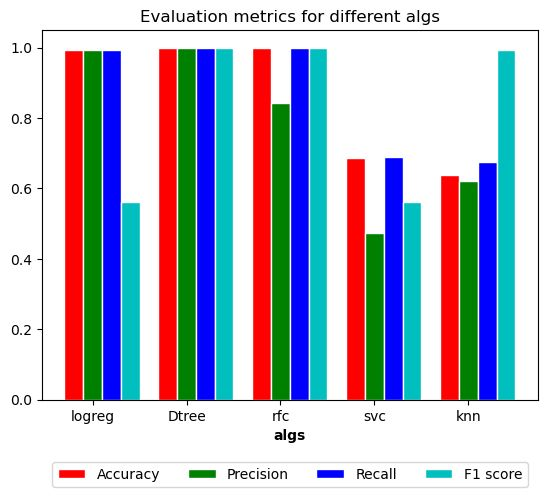
Table 1 tabulates the results obtained for the algorithms implemented containing the accuracy, precision, recall and f1 score values calculated.

TABLE I

Evaluation Metrics of The Classifiers Implemented

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning**  **Classifiers** | **Evaluation Metrics** | | | |
| ***Accuracy (%)*** | ***Precision*** | ***Recall*** | ***F1 score*** |
| Logistic Regression | 99.35 | 0.9935 | 0.9935 | 0.9934 |
| Decision Tree | 100 | 1.0 | 1.0 | 1.0 |
| Random Forest Classifier | 100 | 1.0 | 1.0 | 1.0 |
| Support Vector Machine | 68.83 | 0.4737 | 0.6883 | 0.5612 |
| KNeighbors Classifier | 67.53 | 0.6201 | 0.6753 | 0.9934 |

The visualization for the comparison of evaluation metrics of the classifiers implemented has been illustrated in Fig 2.

Fig. 2. Comparison of Evaluation Metrics

According to our testing and evaluation of several models, as shown in Table 1 and Figure 4, the Random Forest Classifier outperforms the other models in terms of accuracy, precision, recall, and F1 score. We also observed that the classifier's accuracy rose as the dataset grew. We took an easy technique to extract the information from the URLs using simple regular expressions. The system's accuracy might be enhanced still further by testing other components. Additionally, we found that the process of feature selection plays a significant role in determining how well machine learning models perform. By selecting the most crucial elements and eliminating the unnecessary ones, we were able to significantly improve the performance of the models. The overall results highlight the relevance of using the best machine learning algorithm and optimizing the feature selection process to achieve the best outcomes in phishing website detection. In the future, more accurate and effective anti-phishing solutions that protect consumers from online fraud and frauds will be made possible thanks to the knowledge gathered via this endeavor.

References:

[1] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications.” O’Reilly Media.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,”The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” Springer ,Kindle

[3] C.J.C Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[4] D. Wettschereck and D. Thomas G., “Locally adaptive nearest neighbour algorithms,” Adv. Neural Inf. Process. Syst., pg. 184–186, 1994.

[5] Han EH., Karypis G., Kumar V. (2001) “Text Categorization Using Weight Adjusted k-Nearest Neighbour Classification”. In: Cheung D., Williams G.J., Li Q. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2001. Lecture Notes in Computer Science, vol 2035. Springer, Berlin, Heidelberg.

[6] Shengyi Jiang,Guansong Pang,Meiling Wu,Limin Kuang, “An improved K-nearest-neighbour algorithm for text categorization”,Expert Systems with Applications,Elsevier(2012)

[7] Wei Zheng,HaiDong Wang,Lin Ma,RuoYi Wang,”An Improved k-Nearest Neighbour Classification Algorithm Using Shared Nearest Neighbour Similarity” ,Metallurgical & Mining Industry . (2015), Issue 10, pg. 133-137. 5p.

[8] V. Vapnik, “Statistical Learning Theory”, Wiley-Interscience, Publication, New York, 1998.

[9] X.Frencis Jensy, V.P.Sumathi,Janani Shiva Shri, “An exploratory Data Analysis for Loan Prediction based on nature of clients”, International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[10] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,“Loan Prediction by using Machine Learning Models”, International Jo

[11] J. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods”, 2004.

[12]<https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset?select=train_u6lujuX_CVtuZ9i.csv>

[13] Lili Lai, Loan Default Prediction with Machine Learning Techniques. 2020 International Conference on Computer Communication and Network Security (CCS)

[14] Ashlesha Vaidya. Predictive And Probabilistic Approach Using Logistic Regression: Application to Prediction of Loan Approval. 8th ICCCNT 2017 July 3-5, 2017, IIT Delhi, Delhi, India.

[15] P. L. Srinivasa Murthy1, G. Soma Shekar2, P. Rohith3, G. Vishnu Vardhan Reddy4. Loan Approval Prediction System Using Machine Learning. Journal of Innovation in Information Technology. Vol. 4(1), Jan–Jun 2020@ ISSN: 2581-723X.

[16] Lee Y C. Application of support vector machines to corporate credit rating prediction [J]. Expert Systems with Applications, 2007, 33(1): 67-74.

[17] Zhong H, Miao C, Shen Z, et al. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings [J]. Neurocomputing, 2014, 128: 285-295.

[18] Guo X, Zhu Z, Shi J. A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm[J]. Mathematical Problems in Engineering, 2012, 2012.

[19] Kvamme H, Sellereite N, Aas K, et al. Predicting mortgage default using convolutional neural networks[J]. Expert Systems with Applications, 2018, 102: 207-217.

[20] Bai M, Zheng Y, Shen Y. Gradient Boosting Survival Tree with Applications in Credit Scoring (J. arXiv preprint arXiv: 1908.03385, 2019.

[21] Kruppa J, Schwarz A, Arminger G, et al. Consumer credit risk: Individual probability estimates using machine learning (J. Expert Systems with Applications, 2013, 40(13): 5125-5131.

[22] Li H, Huang H, Zheng Z. Research on Credit Risk of P2P Lending Based on CatBoost Algorithm[J]. 2019.

[23] Byanjankar A, Heikkila M, Mezei J. Predicting credit risk in peer-to-peer lending: A neural network approach[C]//2015 IEEE Symposium Series on Computational Intelligence. IEEE, 2015: 719-725.

[24] Ma X, Sha J, Wang D, et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J].

[25] Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems, ISBN: 978-1-7281-4108-4.

[26] Kashvi Taunk, Sanjukta De, Srishti Verma, Aleena Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification”, 2019 International Conference on Intelligent Computing and Control Systems, ISBN: 978-1-5386-8113-8.

[27] Mohamed Alaradi, Sawsan Hilal,” Tree-Based Methods for Loan Approval”, 2020, International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy

[28] Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma, “A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification”, 2019, International Conference on Intelligent Sustainable Systems, ISBN: 978-1-5386-7799-5

[29] S. Zaman, S. M. Uddin Deep, Z. Kawsar, M. Ashaduzzaman and A.I. Pritom,” Phishing Website Detection Using Effective Classifiers and Feature Selection Techniques,” 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2019, pp. 1-6