Covenant Journal of Informatics & Communication Technology. Vol. 10 No. 2, June, 2022
ISSN: *print* 2354 – 3566 *electronics* 2354 – 3507 DOI:

An Open Access Journal Available Online

A Prediction Model for Bank Loans Using Agglomerative Hierarchical Clustering with Classification Approach

Micheal Olaolu Arowolo[1], Oluwatosin Faith Adeniyi[2], Marion Olubunmi Adebiyi[3], Roseline Oluwaseun Ogundokun[4, 5]

[1]Lab 110, Bondlife Sciences Center, Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri
[2, 3, 4]Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria.
[5]Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania

Received: 06.01.2022 Accepted: 14.11.2022
Date of Publication: December, 2022

Abstract—Businesses depend on banks for financing and other services. The success or failure of a company depends in large part on the ability of the industry to identify credit risk. As a result, banks must analyze whether or not a loan application will default in the future. To evaluate if a loan application was eligible for one, financial firms used highly competent personnel in the past. Machine learning algorithms and neural networks have been used to train class-sifters to forecast an individual's credit score based on their prior credit history, preventing loans from being provided to individuals who have failed on their obligations but these machine learning approaches require modification to solve difficulties such as class imbalance, noise, time complexity. Customers leaving a bank to go to a competitor is known as churn. Customers who can be predicted in advance to leave provide a firm an edge in client retention and growth. Banks may use machine learning to predict the behavior of trusted customers by assessing past data. To retain the trust of those clients, they may also introduce several unique deals. This study employed agglomerative hierarchical clustering, Decision Trees, and Random Forest Classification techniques. The data with decision tree obtained an accuracy of 84%, the data with the Random Forest obtained an accuracy of 85% and the clustered data passed through the agglomerative hierarchical clustering obtained an accuracy of 98.3% using random forest classifier and an accuracy of 98.1 % using decision tree classifier.

Keywords/Index Terms— Churn prediction, Decision Tree, Random Forest, bank, agglomerative hierarchical clustering

1. Introduction

Throughout the business economy, the bank plays a critical role. The organizational success or breakdown rests to a considerable extent on the capacity of the industry to assess credit risk (Alabi et al., 2021). The bank evaluates if the borrower is bad or excellent before issuing the credit loan to creditors (non-defaulter). The borrower status forecast, which means that the borrower in the future will be default or non-defeating in the case of any company or bank, is a challenge. (Kalyani and Tijare, 2017). To determine if a loan application is likely to default in the future, banks use credit evaluation to determine whether or not to proceed with the loan. The difficulty is that a loan must be classified as either default or non-defaulting. This allows the banks to reduce probable loss and can boost the amount of credit, as well as the application that enables banks to anticipate the future of the loan and its condition. (Kumar et al. 2019). Machine learning techniques like classification and prediction have been adopted to solve bank loan problems. They are widely employed in the banking industry to enable them to compete in the market and to reduce the risk of the appropriate product to the appropriate consumer (Adebiyi et al., 2022). The main source of risk to the banking business is credit risks that account for the potential of losses or loan defaults. Machine learning can allow banks to forecast trusted consumers by evaluating prior data (Adebiyi et al., 2022). They might also arrange for the launch of several unique offers to maintain the credibility of those clients (Kalyani and Tijare, 2017).

Computer-aided prediction systems benefit greatly from machine learning, which is an artificial intelligence approach (Behera et al., 2020). It develops a model based on data collected during training. To make a forecast, a model created by a machine-learning algorithm is employed. After training the system with a small percentage of the available data, the algorithm uses the remaining data to test it. To make predictions, machine learning techniques can be used to a sample set of test data. (Kumar et al.2019). Several machine learning methods were used for loan prediction systems, e. g. logistic regression, linear model, decision tree (DT), neural network (NN), random forests (RF), vector support machines, model tree, multivariable adaptive regression splines, bagged cart model. Problems with the use of lean models include several computer faults, content problems, and weight fixing in the computerized prediction system. (Blessie and Rekha, 2019).

This study suggests a machine learning approach using an agglomerative hierarchical clustering algorithm with decision trees and random forest classification algorithm. The clustering algorithm is used for segmenting or clustering different groups of the data and the classification algorithm is used to remove those who are most likely to default.

2. Related Works

Sudhamathy (2016) The method of risk analysis to sanction a customer loan using the R package has been proposed. Data selection, pre-processing, extraction and selection of functions, model construction, prediction, followed by assessment included. The data set utilized in this procedure was used in the UCI repository for review. The pre-processing activities have included: identification, rating, and elimination of outliers, elimination of imputations, and equilibrium of the data set in proportionate bifurcation for the testing and training process to improve its accuracy. In addition, the selection procedure enhances the accuracy of the forecast. The decision tree model was assessed and the

prediction accuracy of 94.3 percent was achieved.

A technique for identifying loan risk using data mining was proposed by Aboobyda and Tarig (2016). The predictive model for predicting and classifying loan applications that have been brought by the customers to a good or poor loan using customer behavior and past credit history was designed by three algorithms - j48, Net of the Bayes Net, and naive algorithms. Three algorithms. Weka has designed the model. After using the algorithms j48, Bayes Net, and naïve Bayes data mining approaches of classification, the best loan method for classification was discovered to be j48. This is optimal because of the high precision and low mean absolute error. J48 algorithm.

Goyal and Kaur, (2016) Proposed a consumer loan prediction process based on ensemble technical techniques. In the current technique, the following sub-processes included: data collecting, data filtering, feature extraction, model application, and outcomes analysis.

The random forest, vector support, and tree models with a genetic algorithm were the different prediction processes utilized in the current technique. To evaluate the models, the parameters studied were precision, GINI coefficient, an area under curve, receiver curve, CRT, Kolmogorov - Simonov chart, minimal cost - weighted error rate, minimal error rate, and CCK parameters. The experimental findings revealed that three approaches integrated into the random forest, the vector support machine, and the genetic algorithm tree model were enhanced by loans - predictive outcomes rather than the prediction by a single approach.

Goyal and Kaur, (2016) submitted a model loan prediction employing many machine learning algorithms (ML). To determine loan eligibility to the loan sanction, the data package was used with features such as gender, relationship status, schooling, family size, work status, earnings, combined income, loan balance, the duration for loans, loan record, current loan status and area of the property. Various models used in this technique are Linear model, Decision Tree (DT), Neural Network (NN), Supermarket Forest (RF), SVM, Model Tree, Multivariate Adaptive Record Splines, Bagged Cart Model, NB, and STR. TGA resulted in superior loan predicting performance than the other techniques when these models were examined in five runs using R Environment.

Tejaswini et al. (2016) applied Three techniques for machine learning, Logistic Regression (LR), Random Forest (RF), and Decisions Tree (DT) are intended to forecast consumer loan acceptance. The outcomes of the experiment revealed that in comparison with logistic regression and Random Forest Machine learning techniques the accuracy of the algorithm for the verdict is improving. Many incidents of computer crashes have occurred, content mistakes have occurred, and the most essential weight of characteristics is remedied in automated prediction systems so that so-called software adjustment might be made more safe, dependable, and dynamic. This module can soon be added to the Module for automated system processing. In future software, the system can be trained using old training data so that fresh tests After a while, dates should also be included in training data.

Kalyani and Tijare (2017) The suggested model was designed to forecast the trustworthiness and behavior of clients in terms of loan payback. For data preparation and classification model building, several R functions and packaging were utilized. The work has shown that the R package is an excellent tool to visualize data extraction technology. R package libraries help in successful data analysis and feature selection. Using this approach bank, the necessary information can be simply identified from large numbers of data sets and helps to successfully loan average the number of problematic loan issues. The technology of

database mining for banking services is very important to improve the targeting and acquisition of new clients, to ensure that customer retention is of paramount value, to provide automatic credit approval for fraud prevention purposes, to identify fraud in real-time, to provide market-based products, to analyze customers, to maintain and to market transaction patterns over time.

Vimala and Sharmili (2018) Proposed a model of loan prediction utilizing Naïve Bayes and Support Vector Machine methodologies. An independent speculation methodology, Naïve Bayes includes the notion of probability concerning data categorization. On the other hand, Support Vector Machine utilizes a prediction classification statistic learning model. To assess the suggested strategy, a data set was adopted from the UCI repository with 21 characteristics. The combination has been identified via experiments of Naïve Bayes and Support Vector Machine leads to efficient classification of loan forecasts rather than independent classifying performances (NB and SVM).

Jency et al. (2018) proposed an EDA analysis of loan predictions based on the character and demand of clients. Exploratory data analyzing During the data analysis the main factors concentrated on: yearly revenue vs. credit, the confidence of the customer duration of the loan compared to the criminal months, loan tenure vs. type of credit, and loan tenure versus current employment number of years.

Kumar et al. (2019) applied machine learning in Loan approval prediction. The client loan acceptance status for banking credits was predicted with 3 machine learning algorithms. The results indicate that for logistic regression, decision books, and random forest algorithms the prediction accuracy is 93.04%, 95% respectively 92.53%. Amongst three, it is better to forecast lending to the accuracy of the decision tree algorithm. The Decision Tree Algorithm Could be utilized in the future to further assess its correctness in different data sets available for loan approvals. In addition, a rigorous investigation of the power of machine learner algorithms for loan approval prediction may be performed further than these three.

Blessie and Rekha, (2019) implemented four models for loan predictions which are Logistic Regression, Decision Tree, and Support Vector Machine, and Naïve Bayes method. By studying positively positive characteristics and limitations, the model Naïve Bayes has been concluded with confidence that it is very efficient and results better than other model models. It operates properly and satisfies all bankers' standards and can be linked to many other systems. In automated forecast systems, there were several faults, contain inaccuracies and weight fixation banking software may become more reliable, accurate, and dynamic shortly. Initially, old data sets are provided to the system, and then new data sets are added afterward. The learning of machines helps to identify the aspects that most affect the individual results. Other models such as neural networks and discriminatory analyses may be used on their own or coupled to increase reliability and predictive accuracy.

Vangaveeti et al. (2020) proposed a model using the logistic regression model of the machine learning techniques which falls under supervised learning. Using the logistic regression model, they were able to predict whether the loan is approved or not. They were able to predict whether the loan is accepted or not using the logistic regression model. To apply these different input variables, the output was obtained. If the software receives the input data, the result is given as binary, i.e., as both 0 and 1. If the output is 1, '1' is shown and the loan is accepted. It is shown. If the output is 0, '0' will be shown and the loan is disallowed. The loan prediction system has been created to assist firms to choose the appropriate choice to approve or reject clients' loan requests that

will undoubtedly assist the bank sector build efficient supply channels. In this model, the procedure for logistic regression is applied. Implementation and testing of the domain using different approaches that outperform common data mining methods.

Amruta S. Aphale, (2020), Used a machine learning method to analyze credit data in the bank, to forecast the value of consumers' credit. They used several machine algorithms to investigate the bank credit dataset, to find what methods were most suited. Aside from the closest Centroid and Gaussian Naive Bayes algorithms, all of the other algorithms performed well in terms of accuracy and other performance measurement techniques. Each of these algorithms obtained a precision rate between 76% and more than 80%.

The most crucial characteristics that impact consumers' credit value have also been identified. Some of the Their performance accuracy compared to the specified algorithms to the case where All characteristics are employed then employed these most significant characteristics. No difference in their prediction accuracy and other measurement was found in the experimental findings. They constructed a predictive model for the prediction of credit value using linear regression, which consisted of the main characteristics. Predict bank loan approval to include the most significant functions to estimate consumer's credit worth to develop an automated method of bank risk.

For Gautam et al. (2020) to handle or reject the loan request or loan forecast, they have applied exploratory data analyzes techniques. To tackle the bank lending problem, two machine learning models (decision tree and random forest) were deployed. In the future, the document can be extended to a higher level, so that the program may be somewhat safer and more accurate. Finally, in the automated prediction system, there have been several computer failures, content faults, and the main weight of the characteristics. The program may be adjusted to make weight adjustments safer, more dependable, and more dynamic. In the future the with the automated processing system module, the prediction module may be incorporated.

Patel et al. (2020) employed data mining methods to forecast potential defaulters in a home loan application dataset. Different methods to forecast loan defaults have been implemented. Using logistic regression, the random forest, gradient boosting and cat boost classification, optimum results were obtained. In contrast to logistic regression, gradient boosting provides better or equal outcomes.

Sheikh et al. (2020) employed logistic regression to explore the problem of loan default forecasts as an extremely essential method in predictive analytics. The prediction procedure started with data clearing and handling, Missing values imputation, data set, and model experimental analysis construction for model assessment and testing of test data. Data analysis was performed. The best accuracy in the data set was 0.811. On the original data set. After analyzing the following findings, those candidates with the poorest loan score will fail to receive loan approvals since the loan amount would be more likely to not be reimbursed. As a general rule, applicants with higher incomes and smaller loan requests are more likely to be approved, which makes sense, and are more likely to repay their obligations.

3. Methodology

The model suggested in this research focuses on the model implementation utilizing techniques of clustering and classification. The dataset is a dataset from the UCI learning machine in Taiwan. The dataset is pre-processed utilizing the agglomerative hierarchical clustering that filters the data. The characteristics of the data set can be categorized using decision tree and random

forest classification algorithms, and the results are compared with precision, accuracy, specificity, f1 score, and computational time.

3.1 Dataset

A dataset from the UCI (https://archive.ics.uci/ml/datasets/defaults+of+credit+cart+clients) is proposed in this project for Taiwan. This dataset includes default payments by customers in Taiwan and examines the predicted accuracy of default likelihood across the six methods of data mining. The dataset includes multivariate features, 3000 occurrences, 24 attributes, and no values that are missing.

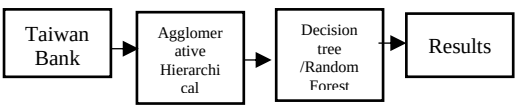

Figure 1. Proposed System Workflow

3.2 Cluster Technique

Data tuples are considered objects for clustering methods. They divide the items into groups and clusters, making them "similar" to each other within one cluster and "different" from the items of other clusters. The similarity is often characterized by how the objects "near" on a distance-based basis are within space. The quality which is the greatest distance between any two items in the cluster can be represented by its diameter. Centroid distance is an alternate cluster quality metric and is defined as the mean distance between each cluster item and the centroid cluster. The cluster representation of data is used to substitute the real data in the reduction of the data. The efficacy of this technology depends on the type of data. For data that can be sorted into several clusters rather than for smeared data, it is significantly more effective. Agglomerative hierarchical is used for the clustering technique.

3.2.1 Agglomerative Hierarchical Clustering

Most often, items are clustered based on their similarity using agglomerative clustering, a hierarchical clustering technique. Alternatively, it is known as AGNES (Agglomerative Nesting). Each item is first treated as a singleton cluster. All clusters are then combined into one large cluster, which contains every item. Items are represented as trees in a dendrogram, which is a tree-based representation of objects. A "bottom-up" approach is used in agglomerative clustering. So, each item starts as its cluster (leaf). The two most comparable clusters are merged into a new larger cluster at each stage of the algorithm (nodes). One large cluster is created until all points are members (root). (Chung. et.al, 2015)

A measure of dissimilarity between sets of data is necessary to identify whether clusters should be merged (for agglomerative) or where a cluster should be divided (for divisive). To do this, most techniques of hierarchical clustering employ a distance measure between pairs of observations and a dissimilarity criterion that specifies the dissimilarity of sets as a function of the pairwise distances between observations in those sets, respectively. Because certain components may be closer together under one metric than another, the form of the clusters will depend on the metric used. (Zhang and Guo,2007)

Algorithm 3.1: Agglomerative Hierarchical Clustering (Chung. et.al, 2015)

```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3     do C[n][i] ← SIM(d_n, d_i)
 4     I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do ⟨i, m⟩ ← arg max_{⟨i,m⟩: i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8     A.APPEND(⟨i, m⟩)  (store merge)
 9     for j ← 1 to N
10     do C[i][j] ← SIM(i, m, j)
11        C[j][i] ← SIM(i, m, j)
12     I[m] ← 0  (deactivate cluster)
13  return A
```

3.3 Classification Algorithm

The categorization approach is one of the main elements for the evaluation of food quality using computer vision, as the objective of computer vision is to substitute automatic methods for the visual decision-making process. Supported by potent categorization systems, computer vision offers a framework for artificial stimulation of the human thought process and may assist people to make difficult opinions precisely, fast, and extremely consistently over a long period. (Abdullah et al., 2004). With the use of sample data, a classification system can produce an updated foundation for better categorization of later data from the same source. (Michie, 1991). Moreover, with a set of training data, it may automatically acquire significant or non-trivial associations and generalize these connections to comprehend fresh, unsightly test data (Mitchell et al., 1996).

Classification usually identifies items by classification into one of the finite sets of classes, including comparing the measurable characteristics of a new thing to those of a known object or other known criteria, and if that new entity is in a certain category of items. The classification algorithm obtains the essential knowledge for making choices in unknown situations once the training set has been received. Intelligent judgments are formed, based on knowledge, as results and at the same time as the knowledge base, which generalizes how inspectors do their functions (Abayomi-Alli, Misra, & Abayomi-Alli, 2021). A classifier is inducing the computationally difficult portion of the classification - i.e., the ideal values of the parameters that the classificatory will employ. Classifiers can offer simple answers to yes or no and can also estimate the likelihood that an object is in each class (Odusami et al., 2021). Decision trees and random forest are used for classification techniques.

3.3.1 Decision Tree

The decision tree is a non-parametric supervised machine learning approach. The target variable has been pre-defined and is usually utilized in issue categorization. It is useful to both classify and regress. It works for both input and output variables categorically and continuously (Kumar et al. 2019, Ogundokun et al., 2021). Decision Trees employ tree representation for problem prediction, the external node of the tree and the node of the leaf represent attribute and class label, respectively. This section describes the pseudo-code for the Decision Tree model:

Algorithm 3.2: Decision Tree (Chandra Blessie, 2019)

Step 1: The best tree root characteristic is selected.

Step 2: Training is separated into sub-sets such each sub-set has an attribute with the same value.

Step 3: Step 1 or Step 2 are repeated for all sub-sets until every single node in a tree passes

3.3.2 Random Forest

Random Forests are a group way to learning for classification, regression, and other activities, which function through the construction of a multitude of Decision-making trees and class (classification) or average class mode (regression)of the various trees. Random forests are the (Kumar et al. 2019). Random forest decisions are right to overfit their training habits by decision-makers. In general, random forests outweigh choice trees, though their precision is less than gradient enhanced trees. However, its performance might have an impact on data properties. Naturally, the random forest forecasters lead to a difference between observations.

Unstructured data can be measured using a

random forest differential measure. The data presented are the original, unmarked data that were obtained from a distribution of references. Dissimilarity calculated using a random forest is advantageous since it can handle mixed variable types effectively and is invariant to repeated changes of the input variables. For many semi-continuous variables, the random forest difference is a simple treatment because of its inherent variable selection properties. For example, "Addcl 1" random forest disseminations weigh the contribution to each variable by its dependency on other factors. Random forest discrepancies were used in several applications, for instance in the identification of tissue marker information for patient groups.

Algorithm 3.3 Random Forest

Step 1: From a total of "m" features, choose "k" at random. In such a scenario, km would be appropriate.
Step 2: Find the optimal division point between the "k" features to determine the "d" node.
Step 3: Split the node into the best-performing daughter nodes that result from the split.
Step 4: "L" node number is found by repeating steps 1 through 3.
Step 5: To create a forest of Trees with "n" digits, repeat steps 1 through 4 n times.

3.4 Performance Evaluation

Accuracy is the number of right forecasts provided by the model over predictions of all kinds in the categorization tasks.

Accuracy is a good metric in the almost equilibrated target variable classes of the data. Accuracy = (TP+TN)/(TP+FP+FN+TN)

Precision is a metric that shows us how much the forecasts are right.
Precision= TP/(TP+FP)

The fraction of true positives that are accurately identified as positives is measured by sensitivity.
4 Sensitivity=TP/(TP+FN)

Specificity is defined as the percentage of genuine negatives that are accurately detected as opposed to positive, known as selectivity or real negative rate (TNR).
Specificity=TN/(FP+TN)

The F1 Score measures the accuracy of a test, defined as the harmonic mean of precision and recall. F1 score=2TP/((2TP+FP+FN)).

3.5 Research Tools

This project proposes to develop the implementation using python on an icore7 processor, with 1.1 GHz speed, 4 GB RAM, 20 GB Hard disk, and Windows 7 OS. Machine learning algorithms were used in the implementation of this system.

4. Results and Discussion

The Taiwan bank dataset was obtained from the UCI repository (https://archive.ics.uci/ml/datasets/defaults+of+credit+cart+clients) and includes default payments by customers in Taiwan and examines the predicted accuracy of default likelihood across the six methods of data mining. Agglomerative hierarchical clustering technique was implemented on Google Collab platform, thereafter, random forest and decision tree classification techniques were performed. Results of the research for the suggested model are presented in this chapter. The Taiwan bank dataset was utilized and was found to consist of 25 attributes and 30,000 instances.

The dataset is cleaned by applying data pre-processing techniques and the data is transformed to be used in the models. The “column ID” was observed to have no significance to our model so it is dropped from the dataset. The pre-processed

dataset is split into testing and training set by each algorithm and 70 percent of the dataset is used for training, while 30 percent is used for testing. An ideal mix of variables for a successful prediction model is determined or learned using a training dataset. On the training dataset, the final model fit is evaluated impartial using the testing data.

The preprocessed dataset is classified using a decision tree classifier. the decision tree classifier is utilized to discover the most efficient way to ask if/then questions to arrive at the correct answer. Finding the most useful test regarding the target variable is the result of a thorough search of all tests available. The decision classifier obtained 84% of classification accuracy when passed into the dataset. Figure 4.5 shows the result of the decision tree.

Figure 2 shows the confusion matrix plot for the classification using decision tree classifier. The false-positive rate yields = 0.1808 and the false-negative rate yields = 0.1322.

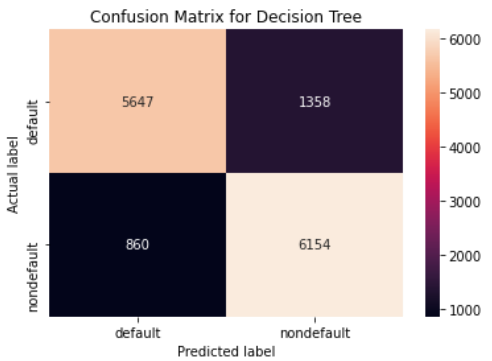

Figure 2: Taiwan Bank Data Confusion Matrix Classification Using Decision Tree (TP= 5647 FP= 1358 FN= 860 TN=6154;)

The pre-processed dataset is classified using the Random Forest classifier. The n estimators' parameter of Random Forest Regressor or Random Forest Classifier determines the number of trees be constructed. The random forest classifier generally performs well without a lot of parameter adjustment, and it doesn't require scaling of the data to work effectively. When applied to the dataset, the random forest classifier had a classification accuracy of 85%.

Figure 3 shows the confusion matrix plot using the Random Forest classifier. The false-positive rate yields =0.1684 and the false-negative rate yields =0.1195.

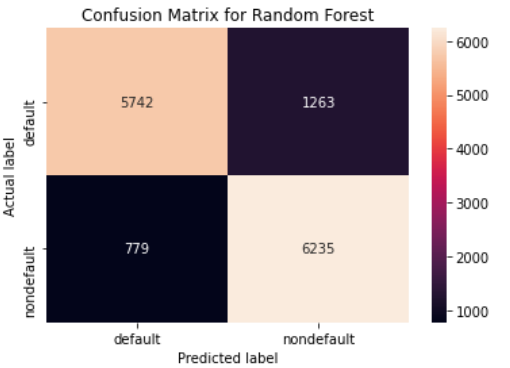

Figure 3: Taiwan Bank Data Confusion Matrix Classification Using Random Forest (TP=5742, FP= 1263, FN= 779, TN=6235)

The agglomerative hierarchical clustering is used for segmenting or clustering different groups of the data. Items are clustered based on their similarity using agglomerative clustering and all clusters are then combined into one large cluster, which contains every item. The clustered data is passed through the decision tree classifier and obtains an accuracy of 98.1%.

Figure shows the confusion matrix plot using the decision tree classifier. The false-positive rate yields = 0.0240 and the false-negative rate yields = 0.0145.

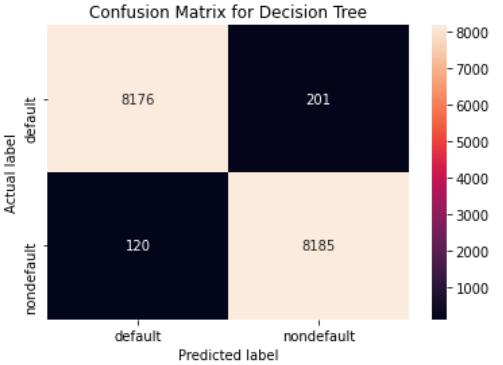

Figure 4. Taiwan Bank Data Confusion Matrix Using Decision Tree with AHC (TP= 8176 FP= 201 FN= 120 TN=8185).

The clustered data is passed through the random forest classifier and obtains an accuracy of 98.3%. Figure 5 shows the confusion matrix plot using the random forest classifier. The false-positive rate yields = 0.0216 and the false-negative rate yields = 0.0122

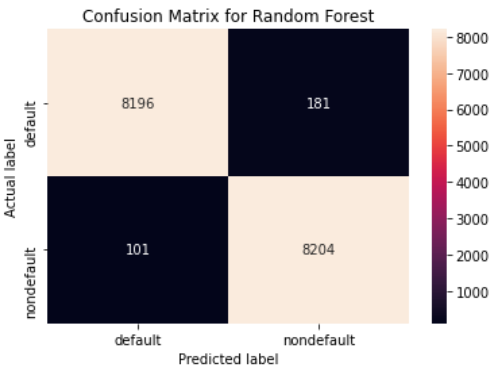

Figure 5. Taiwan Bank Data Confusion Matrix Using Random Forest with AHC (TP= 8196 FP= 181 FN= 101 TN=8204).

The confusion metrics obtained are evaluated using agglomerative hierarchical clustering, Decision Tree, and Random Forest with evaluation such as Accuracy, Precision, Specificity, Sensitivity, F1 score, and Matthew Correlation Coefficient. Table 1 shows the Result evaluations of the experiments.

Table 1. Performance Evaluation

Performance Measures (%)	**AHC + Random Forest**	**AHC + Decision Tree**	**Data + Random Forest**	**Data+ Decision Tree**
Accuracy	98.3	98.1	85.4	84.2
Specificity	97.8	97.6	83.2	81.9
Sensitivity	98.8	98.6	88.1	86.8
Precision	97.8	97.6	81.9	80.6
F1 Score	98.3	98.1	84.9	83.5
Matthews Correlation Coefficient	96.6	96.2	71.0	68.5

In this study, several experiments were carried out and table 4.1 shows the evaluation, however, the AHC + random forest outperformed others with an accuracy of 98.3%. Table 2 shows the accuracy comparison of the results obtained with the state-of-the-art.

Table 2. Comparative Algorithm Used

Authors	**Algorithms/ Methods used**	**Result (Accuracy)**
Aboobyda and Tarig (2016)	J48+Bayes Net+ Naïve Bayes	78.3%,77.7%,73.8%

Arutjothi and Sentham arai (2017)	K-Nearest Neighbour	75.08%
Sheikh et al. (2018)	Logistic Regressio n	81.1%
Blessie and Rekha, (2019)	Logistic Regressio n+ Decision Tree +Naïve Bayes+ Support Vector Machine	78.91%, 71.92%, 65.27%, 80.42%
Kumar and Goel (2020)	Decision Tree	76.4%

5. Conclusion

In the banking industry, loan prediction and evaluating a customer's eligibility for a loan are of paramount importance. As a result, banks must analyse whether or not a loan application will default in the future. The difficulty is that a loan must be classified as either default or non-defaulting. The goal of this project is to develop a Machine learning approach for predicting worthy/ not worthy applicants to be issued a loan in the banking sector. In this study, Machine learning models were adopted such as agglomerative hierarchical clustering, Decision Tree, and Random Forest was used in the development of a Taiwan bank dataset for bank loan prediction system to help the banking sector in determining the eligibility of customers for a loan. An ideal combination of variables for a decent prediction model was determined on the training dataset (70%) and evaluated on the testing dataset (30%). The data was passed into the Decision Tree and Random Forest classification algorithms and results were obtained with an accuracy of 84% and 85% respectively. This work is enormous and of advantage to the banking sector in loan prediction. There was difficulty in obtaining a dataset within a small reach but a good dataset was obtained. This work can be used by the banking sector to improve the prediction of loans and assist them in selecting worthy applicants with a fast, immediate and easy approach. It also allows future researchers to build, enhance approaches to solving the bank loan system problem and also save any financial institution from undergoing huge losses. It is crucial to note that the loan prediction system is being developed to improve it. This study would also recommend that other algorithms can be introduced such as Naïve Bayes, X-means, Logistic Regression in other to improve the robustness of the system.

References

Abayomi-Alli, O., Misra, S., & Abayomi-Alli, A. (2022). A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset. Concurrency and Computation: Practice and Experience, e6989.

Adebiyi, M. O., Adeoye, O. O., Ogundokun, R. O., Okesola, J. O., & Adebiyi, a. A. (2022). Secured loan prediction system using artificial neural network. Journal of Engineering Science and Technology, 17(2), 0854-0873.

Alabi, K. O., Abdulsalam, S. O., Ogundokun, R. O., & Arowolo, M. O. (2021). Credit risk prediction in commercial bank using chi-square with SVM-RBF. In International Conference on Information and Communication Technology and

Applications (pp. 158-169). Springer, Cham.

Amuda, K. A., & Adeyemo, A. B. (2008). Customers Churn Prediction in Financial Institution Using Artificial Neural Network.

Aphale, A. S. (2020). Predict Loan Approval In Banking System Machine Learning Approach for Cooperative Banks Loan Approval. 9(8), 991–995.

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan Approval Prediction based on Machine Learning Approach. IOSR Journal of Computer Engineering (IOSR-JCE), 18(3), 79–81. https://doi.org/10.9790/0661-1803017981

Arutjothi, G., & Senthamarai, C. (2018). Prediction of loan status in the commercial bank using machine learning classifier. Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017, Iciss, 416–419. https://doi.org/10.1109/ISS1.2017.8389442

Behera, R. K., Das, S., Rath, S. K., Misra, S., & Damasevicius, R. (2020). Comparative Study of Real Time Machine Learning Models for Stock Prediction through Streaming Data. J. Univers. Comput. Sci., 26(9), 1128-1147.

Biswas, G. (2014). Clustering Sequence Data using Hidden Markov Model Representation. December. https://doi.org/10.1117/12.339979

Blessie, E. C., & Rekha, R. (2019). Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process. 1, 2714–2719. https://doi.org/10.35940/ijitee.A4881.119119

C. Wang and Y. Tzeng, "Prediction Model for Policy Loans of Insurance Company," in 2007 9th IEEE International Conference on e-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, e-Commerce, and e-Services, Tokyo, 2007 pp. 653-658. DOI: 10.1109/CEC-EEE.2007.81

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. Machine Learning, 32(1), 41–62. https://doi.org/10.1023/A:1007469218079

Gautam, K., Singh, A. P., Tyagi, K., & Kumar, S. (2020). Loan Prediction using Decision Tree and Random Forest. 853–856.

G. Shingi, “A federated learning-based approach for loan defaults prediction," in 2020 International Conference on Data Mining Workshops (ICDMW), Sorrento, Italy, 2020 pp. 362-368.DOI: 10.1109/ICDMW51313.2020.00057

H. Park, K. Kwon, A. Khiati, J. Lee and I. Chung, "Agglomerative Hierarchical Clustering for Information Retrieval Using Latent Semantic Index," in 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 2015 pp. 426-431. DOI: 10.1109/SmartCity.2015.108

Jafar Hamid, A., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. Machine Learning and

Applications: An International Journal, 3(1), 1–9. https://doi.org/10.5121/mlaij.2016.3101

Khan, A. A. (2010). Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. 9(7), 8–14.

Kumar, R., Jain, V., Sharma, P., Awasthi, S., & Jha, G. (2019). Prediction of Loan Approval using Machine Learning. 28(7), 455–460.

L. Boiko Ferreira, J. Barddal, H. Gomes and F. Enembreck, "Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Imbalanced Learning Techniques," in 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, USA, 2017 pp. 175-181. DOI: 10.1109/ICTAI.2017.00037

Liu, B. (2010). Software Design Document, Testing, Deployment And Configuration Management, And User Manual of the UUIS -- A-Team 4 COMP5541-W10 Project Approach.

L. Lai, "Loan Default Prediction with Machine Learning Techniques," in 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020 pp. 5-9.DOI: 10.1109/CCNS50731.2020.00009

Madane, N., & Siddharth, N. (2019). Loan Prediction Analysis Using Decision Tree. Journal of the Gujarat Research Society, 21(14), 214–221.

Martín-Oliver, A., Ruano, S., & Salas-Fumás, V. (2020). How does bank competition affect credit risk? Evidence from loan-level data. Economics Letters, 196, 109524. https://doi.org/10.1016/j.econlet.2020.109524.

Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Sharma, M. M. (2021). A hybrid machine learning model for predicting customer churn in the telecommunication industry. In International conference on innovations in bio-inspired computing and applications (pp. 458-468). Springer, Cham.

Ogundokun, R. O., Awotunde, J. B., Sadiku, P., Adeniyi, E. A., Abiodun, M., & Dauda, O. I. (2021). An Enhanced Intrusion Detection System using Particle Swarm Optimization Feature Extraction Technique. Procedia Computer Science, 193, 504-512.

Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020). Loan Default Forecasting using Dat2a Mining. 7–10.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. 77(2).

Rawate, K. R., & Tijare, P. P. A. (2017). International Journal of Advanced Engineering and Research. 860–867.

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. Icesc, 490–494.

Supriya, P., Pavani, M., & Saisushma, N. (2019). Loan Prediction by using Machine Learning Models. 5(2),

144–148.
Vangaveeti, S. A., Venna, N. L., Naga, P., Ramyayajamanam, S., Marni, H., Satish, N., & Maganti, K. (n.d.). Logistic regression-based loan approval prediction. XIII(V), 319–325.
X. Sun, "Prediction of the Borrowers' Payback to the Loan with Lending Club Data," in 2020 International Conference on Modern Education and Information Management (ICMEIM), Dalian, China, 2020 pp. 375-379. DOI: 10.1109/ICMEIM51375.2020.00092
Y. Zhang and Y. Guo, "Notice of Retraction: Agglomerative Mechanism and Spatial Evolution of FDI," in 2007 International Conference on Fuzzy Systems and Knowledge Discovery, Haikou, 2007 pp. 573-577. DOI: 10.1109/FSKD.2007.136.