**Ensemble - Based Credit Risk Assessment Using Machine Learning**

**Srinivas B.V1, Nikita Kapil2, Prerna Kumari3, S Vivek4**

*Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India 1*

*Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India2, 3, 4*

# -------------------------------------------------------------\*\*\*--------------------------------------------------------------

**Abstract:** Credit safety plays an important part while lending money, banks and other financial institutions make sure that they assess their customers carefully before lending money. To ensure unbiased nature and reduce human error in the process of money lending, many tools and techniques are used. The most commonly used approach is to apply machine learning algorithms. Most of the tools do not provide applicably accurate and precise results since they only use supervised learning algorithms, which label data based on the behavior of instances of each class in the training data. This project implements supervised, semi-supervised and unsupervised machine learning algorithms to increase the accuracy. The model that is proposed in this endeavor is an ensemble of one unsupervised clustering stage, one supervised classification stage and one semi-supervised consensus stage. Model performance is be compared with different types of models such as individual model, individual model with consensus model, clustering and individual model, clustering with individual model and consensus model. The end result is expected to be a Credit Risk Assessment machine with a higher accuracy than currently used models*.*

***Keywords:*** Kohonen’s self-organizing maps (SOM), Machine Learning, Supervised Learning, and Unsupervised Learning.

1. **INTRODUCTION**

A credit risk is the risk of default on a debt that arises from a borrower failing to make required payments. Therefore, credit risk is considered a crucial factor by commercial banks and financial institutions to grant loans, enabling reliable evaluation models for credit risk playing an important role in loss control and revenue maximization. A tangible parameter is required to evaluate the user as a defaulter, being the bad applicant, or a non-defaulter, being the good applicant. A defaulter is one who has missed or has been missing their regular payments for a span of more than ninety days. The parameter being used for labeling an applicant as good or bad is their credit score.

Credit scoring is a statistical analysis performed by banks and financial institutions to access an applicant’s credit worthiness, i.e., whether applicant is genuine or not. Credit scoring is used by the banks and financial institutions to help decide on whether to approve or deny credit, and differs between regions depending on countries, governments etc. The credit score by FICO (Fair, Isaac and Company), for example, is a number that ranges between 300 and 850, where 850 is the highest credit and 300 lowest. Used in India, CIBIL (Credit Information Bureau (India) Limited) is a three-digit number ranging between 300 and 900 where 300 is the lowest and the 900 is the highest. A CIBIL score above 750 is considered as a good score. The CIBIL score is basically a reflection of the transactions made by the person by their credit cards based on their activities, that is, whether a person has completed payments for loans in the past on time or not. This way, it assists commercial banks and financial institutions in making the decision of giving a person a loan. A person with a good CIBIL score that is above 750 is given loan without any inconvenience whereas the person with bad CIBIL score that is below 500 may have to face inconvenience for the loan approval. The company collects and maintains credit records of individuals and commercial entities as well. This contains borrowing and payments related to loans and credit cards.

Traditionally, statistical algorithms, such as Linear Discriminate Analysis (LDA) and Logistic Regression (LR) were used to tackle this problem. These statistical methods are widely used due to their simple interpretability and easy implementation; however, their relatively poor predictive performance limit their usefulness, especially on large datasets with a vast number of feature dimensions, leading to the need of a model with good classification performance and interpretability. Classification performance has become crucial for credit scoring, because even a small fraction of a percentage of improvement in performance could lead to a considerable amount of profits for financial institutions.

With the development of machine learning (ML) algorithms and accumulation of a great amount of multi- dimensional customer data, developing credit scoring models with ML methods has become easier and more reliable. The ML methods are divided into two categories, namely supervised ML and unsupervised ML.

Based on the example input-output pairs, supervised learning algorithms map an input to an output. Supervised machine learning algorithms are designed to learn by using examples. The name “supervised” learning comes from the idea that training such an algorithm is like having a teacher who supervises the whole process. When training a supervised learning algorithm, training data will consist of inputs paired with correct outputs. During training, the algorithm will search for patterns in the data that correlates with the desired outputs. After training the model, a supervised learning algorithm will take in inputs and it will determine which label the inputs will be classified based on the training data. The main objective of a supervised learning model is to predict the correct label for newly presented input data.

Unsupervised learning is employed to draw inferences from datasets which consists of data with none labeled responses. Sometimes the training data consists of a group of input vectors with none any target values. Main goal of unsupervised learning problems is to make groups of comparable or similar examples within the given data, which is understood as clustering, or to see how the data is being distributed in space, which is understood as density estimation*.* To put it in simpler terms, true class labels are not provided for any sample, hence known as learning without teacher*.*

The fundamental difference between the two learning methods is whether the examples given to the learning algorithm are labeled or not. Supervised ML, applied to labeled examples, has a wide range of available algorithms, such as support vector machines (SVM), decision trees (DT), random forest (RF), artificial neural network (ANN), each having its strengths and weaknesses.

Unsupervised ML, applied to unlabeled examples, includes algorithms such as k-means clustering, hierarchical clustering; Kohonen’s self-organizing maps (SOM). Both supervised machine learning and unsupervised machine learning have been extensively applied in credit risk assessment.

Supervised machine learning algorithms are used in credit scoring models to find relationship between the customer features and credit default risk and then to predict the default classification usually in a binary format. In a large body of literature, the implementation of supervised ML algorithms in credit scoring models has shown good predictive accuracy. Unsupervised ML algorithms, particularly clustering algorithms, are used as an important data mining technique to cluster examples into groups of similar objects instead of giving predictions directly. Therefore, they are often used as complimentary tools to supervised ones.

There have been different models proposed by people before, all of them either using supervised learning or unsupervised learning but not both. Since both methods have their own drawbacks, the models weren’t reliable as even small inaccuracies can lead to big implications when applied on large datasets. The proposed model uses both supervised as well as unsupervised learning algorithms to give highly accurate results and also help improve the base performance of future credit risk assessment systems.

The dataset being used for this endeavor comes in a masked form from a private bank, unnamed for confidentiality. The dataset has missing values in data, as many people don’t fill optional fields. These missing values can lead to some pattern which may help in identifying the defaulter. The intention of the proposed strategy for handling data missing is to make use of the missing state instead of filling or handling the missing data with data mining techniques, since this pattern of missing data can be a good indicator for us to segregate and find the defaulters at the earliest stages possible.

## RESEARCH MOTIVATION

This work integrates unsupervised learning algorithms with supervised learning algorithms to improve performance. The performance of this model’s flow has been compared with different, previously used flows such as using individual classification models, using a combination of different classification models followed by a consensus model, and clustering before applying the individual classification models. Individual classification models are based on supervised learning and may include logistic regression, support vector machines, decision trees, and artificial neural networks, among others. The reason individual models underperform is that no single ML algorithm can perform best on all practical learning problems, as No Free Lunch theorem indicates. When classification models are combined with a consensus model, the accuracy is considerably better, but still not good enough, since the data itself is not being understood before classification, which could lead to misclassification. Clustering before applying individual classification models does cover this front, but still doesn’t provide enough accuracy to surpass the No Free Lunch theorem. Hence, the endeavor in the discussed model is to apply clustering on the data to make data more meaningful to the system, then apply an array of classification models to figure out the base probability of the models before finally taking a consensus of these class predictions, which is done using a semi-supervised consensus model that weighs the relevance of the output of each classifier and takes a scaled average to determine the output.

## METHODOLOGY

In this section, the methods related to the work are presented. The diagram below shows the workflow of the model considering the following aspects: credit datasets, data cleaning and feature selection, machine learning algorithms and evaluation measures of model performance.



*Figure 1: Overview of the proposed model*



*Figure 2: Flow diagram of the proposed system*

## Credit datasets

The detailed information about the dataset considering the number of samples, defaults/ non-defaults and feature dimension is presented in the table below.

Private bank credit dataset

The private bank credit dataset contains the information on the default payments by customers, credit data that tells the nature of credit habits, history of payment that tells the payments made and timings, and the bill statements of credit card of the clients, while excluding any and all personal information that could pertain to the individual’s identity beyond the system. The dataset has 11 columns. Some of the features include marital status, number of dependents, whether the person has delayed payments beyond certain time frames etc. The dataset, post- cleaning, contains 40000 samples out of which ~4000 samples are defaulters.

## Data cleaning

Incorrect or inconsistent data leads to false conclusions. Clean and understandable data has a high impact on the quality of the results. For cleaning process, two issues raise much concern – the empty points and the outliers. For handling the empty points, four kinds of methods can be used: case deletion, missing data imputation, model-based procedures and machine learning methods. In this work the features that are not filled by most of the applicants that is 95% or above are removed and the new feature was added to describe the feature status, whether it is empty or not , by assigning status 1 for empty fields and status 0 otherwise. Empty fields are filled with the mean values of these features. The outliers, it has been demonstrated that the use of filters for outliers can help improve the model performance. In this work, the rationality of abnormal points was checked manually, and then the outliers, if rational, are kept, and otherwise replaced by the upper or lower values of box plots. Additionally, normalization was performed to scale the feature values so that they can fall into specified range, typically from 0 to 1.

## Supervised Machine Learning Methods

 Logistic Regression (LR)

Logistic regression uses a logistic function to model a binary dependent variable, although there are many complex extensions that exist. In regression analysis, LR is estimating the parameters of a logistic model (a form of binary regression). Logistic regression has been considered as the industry standard in the field of credit scoring [1]. It is used to solve binary classification problems (default and non-default in this work) and regression problems. The target of LR model is to get the logarithm of the ratio of two probability outcomes of interest.

Random Forest (RF)

A random forest is an estimator that is used to fit a number of decision trees that classifies on various sub- samples of the datasets and uses averaging to improve the predictive accuracy and control over-fitting problems. RF is considered an advanced technique of DTs [2], with the idea behind RF being combining feature selection to merge individual DTs. Randomness is used in two stages in RF – first, while selecting the subsets from the datasets, and then to randomly select subsets of features. This reduces the correlation between the DTs and the forest is reduced.

Gradient Boosting Decision Trees (GBDT)

Gradient boosting algorithm [3] is used to solve classification and regression problems. Boosting is the method of converting weak learners into strong learners. In boosting, each new tree is a fit on the modified version of original dataset. The modeling process is to add decision trees at a time, and then the next tree is added and trained to reduce the loss by moving in the right direction. The model keeps adding trees until the number of trees reaches a fixed number or the loss reaches an acceptable level or no longer improves.

Support Vector Machines (SVM)

SVM [4] has been extensively applied in the field of credit scoring owing to its powerful predictive capabilities. Support vector machines attempt to pass a linearly separable hyper plane through a dataset in order to classify the data into the two groups. This hyper plane is a linear separator of any dimension; it could be linear, polynomial, radial basis function (RBF) and sigmoid [5]. Support Vector Machines have been looked up to in various fields, due to its capability to solve a huge variety of problems.

k-Nearest Neighbors (kNN)

The k-nearest neighbors (KNN) algorithm is a simple and easy to implement supervised machine learning algorithm that can be used to solve both classification and regression problems. kNN is one of the most venerable algorithms in statistical pattern recognition [6]. It has been extensively used in constructing credit scoring models [7], [8]. The main idea behind the kNN model is that it predicts the labels of the new input samples according to the nearest set (or k-nearest neighbors) of previously labeled samples. Euclidean distance is commonly used in kNN models to measure the distance between the new sample and the previous training samples. The model requires only one parameter, being the size of neighborhood - k in kNN algorithm. [9]

## Unsupervised machine learning methods

K-means

K-means is a simple and efficient clustering (or unsupervised learning) method [10]. The basic steps for k- means clustering are as follows: a) randomly select k cluster centers; b) assign each data points to its nearest cluster center;

c) replace the original center with the position center in each cluster; d) relocate each data points to a new cluster which it is nearest to; e) repeat the previous steps c and d until no data point changes position or some convergence criterion is met. There is only one main parameter in k-means model, that is, the number of clusters k. In this work, k-means was used to clustering samples according to their presence conditions (missing or not) to split the dataset into several subsets based on which supervised machine learning models were built.

Kohonen’s self-organizing maps (SOM)

SOM is an unsupervised neural network [11]. The idea of SOM is to project a nonlinear data vector from high- dimensional space into a two-dimensional space, which makes the patterns graphically visualized and easily recognizable. In an SOM map, the neurons are arranged in a two-dimensional array where a winner neuron will be found for each data vector. In the process of training maps, the winner neuron and its neighbors are adjusted according to topology distance, consequently the topological similarity is preserved between the neighboring neurons.

That is to say, if two samples are projected to adjacent neurons, these two samples are similar in terms of input feature descriptions. In this work, we labeled each neuron in an SOM map based on the training set according to the majority voting principle and then predicted the samples in the test set with the pre-defined labels. In cases where the samples from the test set were projected to the neurons without any sample from the training set being projected to, we then turned to look at their surrounding neurons and labeled these undefined neurons according to the majority voting principle based on their surrounding neurons.

## Semi-Supervised Learning

Semi-supervised learning is an approach to machine learning that, as the name suggests, pertains to both the supervised and unsupervised learning methodologies. It combines a small amount of labeled data with a large amount of unlabeled data during the training process, and performs both data clustering and classification, although it can be used to operate on a different form of the data, which, in this case, pertains to the class probabilities or labels predicted by each of the classification models in the previous stage. Voting Classification has been used in this ensemble for consensus, which is a semi-supervised learning algorithm that takes votes of each of the classification models fit into it, and depending on whether the method is hard or soft, takes a scaled consensus of the class labels or class probabilities, and produces a clear and distinct class label per data point as output.

## EXPERIMENTAL RESULTS AND DISCUSSIONS

The algorithms have been tested on the main dataset which is the private bank dataset, along with which Australian and German credit datasets have been used to make comparisons. The performance metrics used to appraise this model are Accuracy, Recall, Precision, F1 Score and Confusion Matrix.

The results of implementation of this model have been considerable, but hindered by the skewed nature of the private bank dataset, which has been derived from practical and real- time records, and while the maximum number of positive records was provided, they do not provide good enough results. Hence, on using a more balanced dataset with higher number of records that represent defaulting behavior, the accuracy can be improved further. The accuracy of the model is 93%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 Score** |
| **Defaulter** | 0.93 | 1.00 | 0.96 |
| **Non-****defaulter** | 0.68 | 0.03 | 0.05 |
| **Average** | 0.80 | 0.51 | 0.51 |
| **Weighted****Average** | 0.91 | 0.93 | 0.90 |

## Table 1. Performance evaluation of the experimental results

The confusion matrix obtained for the same is as follows:

|  |  |  |
| --- | --- | --- |
|  | **Predicted Non-****defaulter** | **Predicted****Defaulter** |
| **Actual Non- defaulter** | 8954 | 11 |
| **Actual Defaulter** | 1017 | 18 |

## Table 2. Confusion Matrix of experimental results

As can be inferred from both the tables below, the prediction of defaulters is affected mostly by the fact that the number of samples that exist for them are low. However, the overall performance of the model is significantly better than many currently used models, and with a few more improvements can be useful for progressing research in credit risk assessment.

## CONCLUSIONS AND FUTURE WORK

Credit risk assessment has always been an important part in banks and financial institutions. And predictive models for credit scoring play an important role in making unbiased decisions. To improve the accuracy and

performance this model, in which unsupervised learning algorithms have been integrated with supervised learning algorithms, has been implemented. The model is tested on three datasets using four routes. It can be observed that the model performs better when using consensus model and dataset clustering method. This means that integration of unsupervised learning algorithms with supervised learning algorithms at different stages leads to increase in model performance.

# This model performance can be improved more by removing all the features that will not contribute in predicting accurate results, can also use clustering algorithms that does not require pre-set number of clusters at all and can also identifies noisy data and does not throw them into a cluster even if the data point is very different.

## REFERENCES

1. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithm for credit scoring.
2. Breiman, L. (1999). Random forest.
3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.
4. Cortes, C., & Vapnik, V. (1995). Support vector machine.
5. Zhou, L., Lai, K. K., & Yu, L. (2010). Least square support vector machines ensemble models for credit scoring.
6. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification.
7. Henley, W. E., & Hand, D. J. (1996). A k-nearest- neighbor classifier for assessing consumer credit risk.
8. Islam, M. J., Wu, Q. M. J., Ahmadi, M., & Sid-Ahmed, M.

A. (2007). Investigating the performance of Naive- Bayes classifier and K- nearest neighbor classifiers.

1. Holmes, C. C., & Adams, N. M. (2002). A probabilistic nearest neighbour method for

statistical pattern recognition.

1. Asgharbeygi, N., & Maleki, A. (2008). Geodesic K- means clustering.
2. Kohonen, T. (1998). The self-organizing map