Credit Risk Assessment Using AI and Deep Learning

Priya Mehrotra
The Hong Kong University of Science and Technology
pmehrotra@connect.ust.hk

Haojun Chen
The Hong Kong University of Science and Technology
hchenbw@connect.ust.hk

Abstract

Machine Learning and AI are getting increasingly popular in the recent years in the financial services community. They are being used in a range of applications in financial industry. Uses of these technologies are in their evolving phase. These technologies have dramatically improved analytical capabilities and have automated all kinds of business lines including credit underwriting, compliance, interaction with clients, and risk management. AI and machine learning [13] techniques are nowadays used to manage credit risk, market risk, operational risk and other risks involved in the FinTech market. Credit Risk assessment is an important topic in financial risk management. Reliable models are needed by the financial institutions such as the banks to grant consumer credits accurately. Deep learning is by far the most accurate state-of-the-arts approach in the field of artificial intelligence. Many supervised and unsupervised problems have been well resolved by deep learning. Credit scoring is one of the most important challenges to the decision-making process for the lending institutions. A reliable model can help the financial institution to increase its profits and reduce losses. In our research report we have carefully investigated the ability of the artificial neural networks involving deep learning as a decision support system to automatically detect and predict credit risks based on the customer characteristics. In this report we aim to provide an overview of credit scoring models, artificial neural networks as means to detect credit risk.

1 Introduction

Innovation in the field of finance is not a new concept and there are major advancements focussing on technological innovations. Solutions making use of artificial intelligence[12], big data analytics, neural networks and blockchain technologies are introduced at an unprecedented rate in the finance markets. Introduction to such solutions leave the door open to many risks hampering consumer protection and financial stability such as underestimation of creditworthiness, fraud detection, market risk noncompliance and cyber-attacks.

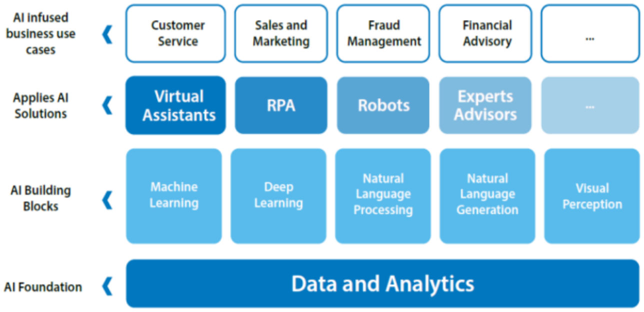


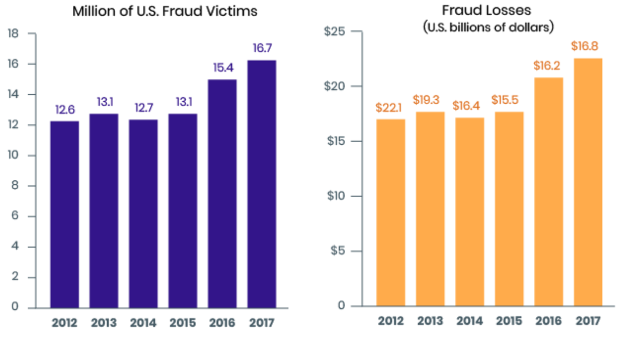
Figure 1: Subfields of AI used in various FinTech Segments [1]

Credit risk is one of the major challenges being faced in the banking systems. Traditional models focus mainly on the borrower's financial status, categorizing the customers on the same. However, spending, saving and borrowing money methods and ways in customers are changing and so is the technology.

1.1 Objectives of the Research

According to a research by PwC, the industry that is capable to take the maximum advantage of machine learning algorithms in finance. One of the basic tasks which the financial institutions such as the banks must deal with is to minimize the credit risk which is involved while granting it to consumers. Predicting the probability that an applicant, consumer[14] or the existing borrower will default or become delinquent is important. In credit risk analysis, credit scoring is a technique that helps the organizations to decide whether to grant the credit to the consumers or not. One approach can be to apply classification by training the model with data of both good and delinquent customers.

There are a number of classification techniques used to evaluate the risks concerned with default applicants such as logistic regression, support vector machine, artificial neural networks and so on. Also, the credit data contain a lot of irrelevant and redundant information as features which generally reduces the classification accuracy and also can lead to incorrect results. Therefore, it is important that we filter out the redundant information and extract the required features before applying models. More and more technological advancements lead to more risks of fraud cases in financial organizations. The industry which is most suffering from fraud-related losses is the financial services industry. As an example, almost 17 million customers and organizations have experienced fraud in the United States according to Javelin's 2018 Identity Fraud Report.



Source: 2018 Identity Fraud Study, Javelin Strategy & Research

Figure 2: U.S. Fraud Victims and Losses

Credit scoring techniques deal with utilizing the past information of consumers credit history and its current economic conditions to classify which customer is liable to pay back and which customer is not. ANNs can be efficiently used as such classifiers to solve complex problems in credit risk analysis. Main objective and significance of the research is to improve the decision making in financial institutions using ANNs as Decision Support Systems (DSS) and determine if ANNs can improve the accuracy and efficiency of the existing credit rating models. Through our research we aim to find reliable models that can improve the accuracy as well as the efficiency of predicting the consumer credit default cases and improving the decision making by using ANNs as Decision Support Systems (DSS).

1.2 The Research Problem

AI technology learns the data obtained from the past, hence the AI solutions in the financial institution are destined to succeed. A company which sells credit cards uses a credit score to decide who is eligible and who is not. The company even provides a customized interest rate on the credit card for which it makes use of the individual's data such as

- number of active loans he has
- number of active credit cards he has
- his loan repayment habit etc.

AI with deep learning has found its place of being capable to go through thousands of personal financial records and provide the figures and results very quickly and accurately as well.
Earlier Back-Propagation Neural Network (BNN), Multi-Layer Perceptron (MLP) and Radial Basis Neural Networks were found to be used for credit scoring. Here we aim to evaluate the performance of deep neural networks in accurately analyzing consumer credit risk.
AI is functioning as a data analyst by saving millions and at the same time increasing the accuracy. AI technology makes use of machine learning algorithms which can learn over time and can analyze large amount of data accurately.

Credit scoring is defined as a method to evaluate the credit risks in case of loan applications. This method outputs a score which is used by a bank to rank the applicants of loan or borrowers aiming on the risk factor.

1.4 Need for Credit Risk Assessment

Bank Systems Technology contributor Chuck Nwokocha demanded a dynamic system for lenders to assess the borrowers. Following considerations need to be made:

- Evaluating credit risk at loan origination point also over time.
- Considering the character of the borrower with appropriate credentials and references.
- Management ability
- Environment conditions and factors

Rules-based approach	**AI-based approach**
Long-term processing	Real-time processing
Requires much manual work	Automatic detection of all possible anomalies
Multiple verification steps that harm the user experience	Reduced number of verification measures
Catches obvious fraudulent activities	Finds hidden fraudulent activities

Table 1: A comparison between Rule-based approaches and AI-based approaches in Fintech Industry

Clients can be both individuals or organizations, a comprehensive loan management software can help the lenders to determine if borrowers are creditworthy of the loan or they pose excessive risk. Credit risk assessment model should be able to capture the "Five Cs of Credit" which are required by lenders to evaluate the borrowers. The five C's of credit include:

- Borrower's reputation and character

- Their capacity to repay the loan along with their borrowing history
- Their total capital
- The conditions of the loan
- Other collateral that could serve as another repayment option

1.5 Challenges to Successful Credit Risk Assessment

There are number of challenges that are faced in order to achieve successful credit risk assessment:

- Inefficient data management: There is an ability to gain knowledge of the right data at the right time.
- Insufficient risk tools: Tools available to manage risk are not very efficient.
- No groupwide risk modeling framework: Big picture of groupwide risk is not possible without proper risk assessment measures.
- Cumbersome reporting: Manual reporting processes as on a spreadsheet are tedious for analysts.
- Constant rework: Changing model parameters is impossible every time as it wastes the efforts and also it has a negative effect on the efficiency of the bank.

A comparison between Rules-based and AI-based approaches used in financial institutions is given in table 1.

2 Credit Scoring Models

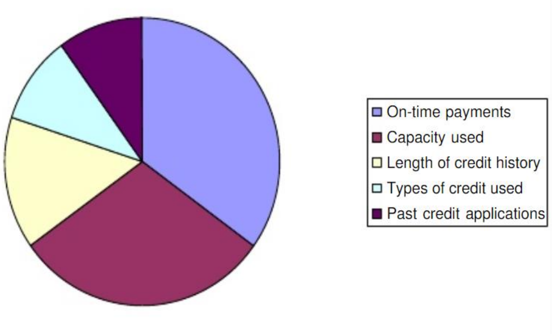


Figure 3: Information used to calculate credit scores

A numerical expression that represents the creditworthiness of an individual is called the credit score of that person. It is the expression that denotes how likely will that individual repay his or her debts. Banks and other lenders use this credit score to analyse the risks that are involved in lending the money to the customer. Credit Scores determines who qualifies for a loan and who does not along with the credit limit.

Following components are used for the calculation of an individual's credit score:

- Length of credit history
- Type of credit used
- Past credit applications
- Punctuality of payments in the past
- Amount of debt (ratio of outstanding balances to total available credit)

The exact formulas for calculating the credit score using above information is guarded as secret by Fair Isaac Corporation (FICO). These components stated above are fed as input into the model which outputs the credit score.

2.1 Problems with credit scores

Several problems are associated with credit scoring:

- If an individual has a higher credit score, he may be offered some different services than the one with less credit score.
- Inaccuracies in terms of credit score can be disadvantageous to the customer as well as the corporation.
- Maintaining past data of individuals if difficult as well as costly.

3 Artificial Neural Networks

3.1 Introduction to ANNs

Artificial Neural Networks are networks which can acquire, store and utilize experimental knowledge. ANNs are systems that are inspired by the biological neural network and replicate the way the human brain works. It consists of an input layer and an output layer as well as a hidden layer consisting of many simple elements called neurons that operate in parallel.

3.2 Capabilities of ANNs

ANNs often employ supervised learning. Training data containing both the input and the target value is fed into the system and the machine learns patterns on its own. For this huge amount is data is needed to be fed. Also, data should be numeric. ANNs usually operate with numeric data only.

3.3 ANN Architectures

In common practice, the neural networks are trained with data so that it produces a specific target output. The ANN is adjusted until its output matches the target. This is done by adjusting its weights as shown in figure 4. This is called supervised network learning.

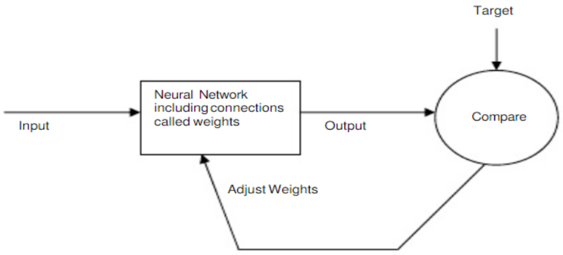


Figure 4: Neural network training

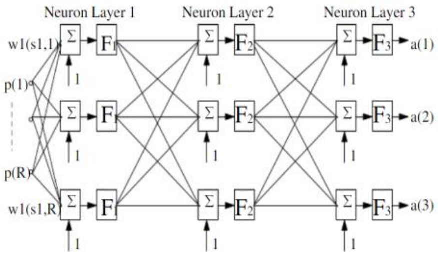


Figure 5: Multiple layers of neurons (Single hidden layer)

Figure 5 shows a network with R inputs, S1 neurons in the first layer. Each layer of ANN has a transfer function and can have different number of neurons. Next layer's input is the output of the previous layer. The transfer function to be used and the architecture to be adopted depends on the nature of the problem to be solved.

3.4 ANN Problems

- **Hidden layer neurons:** If large errors are produced even if we train the network for a long time, the problem is most likely due to the lack of neurons in the hidden layer.
- **Overfitting:** A complex network is more likely to fit the noise in the training data leading to overfitting. This usually results due to too many neurons in the hidden layer. Therefore, the ANNs do not generate reasonable results.
- To represent more complex features and to learn increasingly complex models, a greater number of hidden layers are needed. Such neural networks which consist of more than three layers of neurons (including the input and output layer) are called **Deep Neural Networks [2]**.

4 Need for Deep Neural Networks

Deep Learning has emerged as a novel and crucial model in recent researches which is able to transform raw data into high-level representation of the data using simple non-linear functions. It automatically extracts features from each layer of the network. In real life, data corresponding to bad customers correspond to a very small part of the entire data. An important goal of credit risk assessment models is to build an efficient classification model for a specific dataset. Deep neural networks have been successfully applied to evaluate credit risks. DNNs include multiple hidden layers to automatically extract the features from the data of the customer and determine whether to grant loan to that customer or not.

Deep learning approaches include adding multiple layers to a neural network, though these layers can be repeated. In this regard, most deep learning strategies are based on the following four types of architectures [3]:

- convolutional neural networks, which, essentially, are standard neural networks that have been extended across space using shared weights;
- recurrent neural networks, which are standard neural networks that have been extended over time by having edges that feed into the next time step instead of into the next layer in the same time step;
- recursive neural networks, which are apparently hierarchical networks where there is no time aspect to the input sequence, but inputs have to be processed hierarchically as in a tree;
- standard deep neural networks, which are a combination of layers of different types without any repetition or particular order. This is the type we research on the most

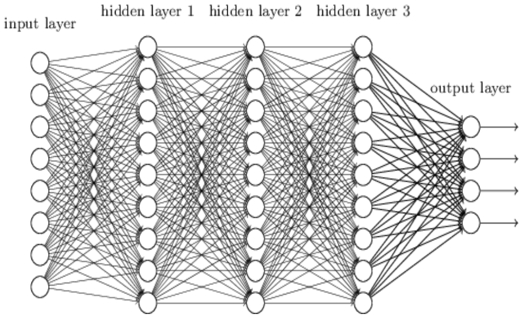


Figure 6: Structure of Deep Neural Network
(Multiple hidden layers)

There are number of advantages of using AI and deep learning techniques for credit risk analysis:

- The machine can learn without being explicitly programmed.
- Unexpected forms of dependencies between the variable are unveiled.
- We can tackle with non-linear relationships.
- A deep network is difficult to train with current algorithms as compared to shallow networks (networks with single hidden layer).

5 The Deep Neural Network Models for Credit Risk Management

Usually the credit data is highly imbalanced. Deep Neural Networks can be used for credit risk assessment to achieve a balanced dataset and also to judge whether the customer can be granted loans or not. Based on the experiments, several model parameters and data sampling methods are analysed, and the classification ability of different models are compared. Specifically, commercial banks faced many risk management processes, among which credit risk assessment is a significant part of commercial banks and application for loans.

A good credit risk assessment approach can aid the financial institutions distinguish “good” customers from “bad” customers to reduce losses. Among many classification methods and techniques that can be used for default risk evaluation, deep learning [1,2] is a new machine learning model that receive scholars’ attention lately. In view of the credit risk assessment issue, there are several credit risk assessment models used currently. The experiment results show that the proposal algorithm achieves a higher prediction accuracy in credit risk assessment.

Shallow neural networks that contain single hidden layers are limited in representative capacity and so is more difficult to train with the algorithms (more local minima, slower convergence rate) [28]. With multiple hidden layers, DNNs improve the feature representative capacity. After intensive training, DNNs usually show better feature representations. Feedforward DNN model contains an input layer, multiple hidden layers and an output layer, where the input layer is used for data input, the hidden layers transforming raw data to high-dimensional non-linear features, and the output layers for classification and prediction of the
output results. The model structures illustrated are as follows:

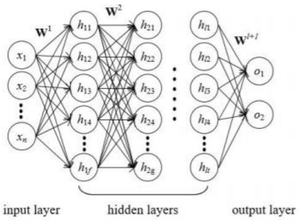


Figure 7: Structure of Deep Feedforward Neural Network

6 Deep Learning Models Review

Deep neural networks have a depth factor which distinguishes them from the single hidden layer networks. Depth means the number of layers of nodes through we the data is passed in order to find the patterns.

Traditionally, credit risk measurement methods focus on estimating the probability of default (PD), rather than on the magnitude of potential losses in the event of default, loss given default (LGD), which is also called LIED, loss in the event of default [5]. Also, the traditional models show failure as bankrupt filing [25], or liquidation and hence they ignore considering credit quality. Traditional models are divided into three broad categories to estimate the PD [9]: (1) expert systems (including artificial neural networks); (2) rating systems and (3) credit scoring model[24].

Altman in 1968 devised a multiple discriminant credit scoring analysis method which is the most widely used traditional credit risk assessment model. In 1997, Mester[26] found that 97 percent banks improved their credit card applications by using credit scoring and 70 percent of the banks used credit scoring for small business lendings.

Four methodological forms of multivariate credit scoring models are as follows: (1) the linear probability model; (2) the logit model; (3) the probit model and (4) the multiple discriminant analysis model

These above models assigned a Z-score to the applicants classifying them as ‘good’ or ‘bad’. This Z-score can then be converted into a PD.

6.1 Clustering and Merging Sampling Algorithm

In credit risk management, most of applicants are non-defaulting. However, there is still a small number of applicants that are defaulting users who are needed to be identified and predicted whether or not they will default in the future. Hence, k-means, as one of the most popular and simple clustering algorithms, can be used in clustering subgroups.

Generally, standard learning algorithms performed better on majority (negative) samples while poorly on minority (positive) samples. In this research report we have proposed a data sampling method based on clustering and merging. This alleviates performance degradation of the model due to class imbalance and increases number of sets to make the classifiers more diverse. The algorithm structure is shown below in figure 8.

And the process consists of the steps [4]:

- The raw data is divided into training set and test set, in which the learning model is trained by the training set and the performance of the model is tested and evaluated by the test set.

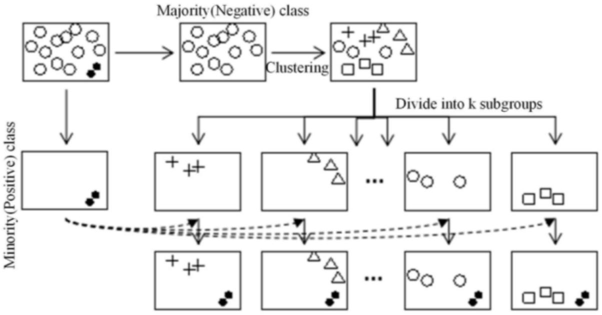


Figure 8: Process of Clustering and Merging Sampling Algorithm

- The majority class samples are clustered into k subgroups so the samples of each subgroup must come from completely different data if k-means algorithm is used in the training set.

- The k subgroups of majority class data all other minority class data are then merged orderly into k balanced subgroups to create diverse sets.

Considering the issue of imbalanced data class and redundant data in credit risk assessment and based on the findings above, a credit risk assessment method using DNNs with clustering and merging sampling algorithm is proposed as shown as figure 9. This method consists of two parts: Data equalization and DNNs classification.

In the first part, the data is clustered into different subgroups to ensure the difference between input data by clustering and merging sampling algorithm. In the second part, multiple DNNs are trained to achieve the classifier [21] diversity by using different training sets that are clustered into multiple subgroups. This method is based on the majority voting that takes majority prediction class as prediction results. The algorithm of the model is shown as below in table 2:

6.2 Classification Restricted Boltzmann Machines (ClassRBM)

The credit scoring problem is explained as a discrimination between those applicants whom the lender is confident will repay credit and those applicants who are considered by the lender as insufficiently reliable. There is a novel method called Classification Restricted Boltzmann Machines (ClassRBM)[15][16]. It is purposed for constructing comprehensible scoring model.

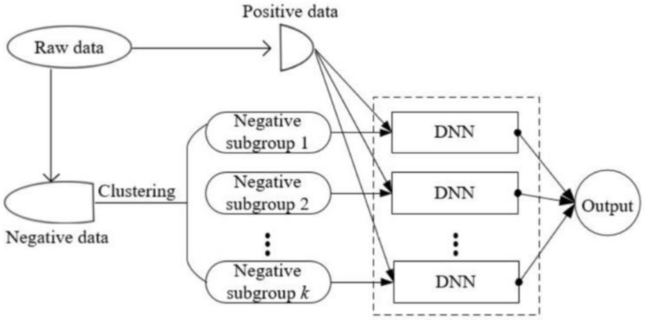
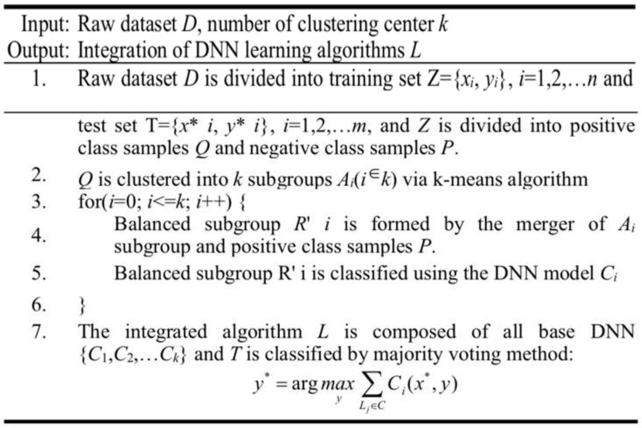


Figure 9: Structure of Credit Risk Assessment Model



```
Input:  Raw dataset $D$, number of clustering center $k$
Output: Integration of DNN learning algorithms $L$
1.  Raw dataset $D$ is divided into training set Z={$x_i$, $y_i$}, $i$=1,2,…$n$ and
    test set T={$x^*$ $i$, $y^*$ $i$}, $i$=1,2,…$m$, and $Z$ is divided into positive
    class samples $Q$ and negative class samples $P$.
2.  $Q$ is clustered into $k$ subgroups $A_i(i \in k)$ via k-means algorithm
3.  for($i$=0; $i$<=$k$; $i$++) {
4.      Balanced subgroup $R'$ $i$ is formed by the merger of $A_i$
        subgroup and positive class samples $P$.
5.      Balanced subgroup R' i is classified using the DNN model $C_i$
6.  }
7.  The integrated algorithm $L$ is composed of all base DNN
    {$C_1$,$C_2$,…$C_k$} and $T$ is classified by majority voting method:
        $y^* = \arg \underset{y}{max} \sum_{L_j \in C} C_i(x^*, y)$
```

Table 2: Credit Risk Assessment Model Algorithm

A novel machine learning technique takes advantage of Classification Restricted Boltzmann Machine (ClassRBM) to construct the credit scoring table. ClassRBM can be used as a universal approximator over the binary random variables. It can be applied to determine weights (scoring points) in the scoring table. Moreover, the scoring tables are the simplest models to interpret and can be easily implemented in any bank system. Additionally, purposed ClassRBM method deals with the imbalanced data by selecting the cutting point with the highest geometric mean of specificity and sensitivity. This is different from the standard methods, purposed approach combines separate issues[17][18]:

- Makes use of the strong classifier (ClassRBM)

- Deals with the uneven class distribution problem, and

- Constructs highly comprehensible and easy-to-implement scoring model.

ClassRBM (Larochelle & Bengio, 2008; Larochelle, Mandel, Pascanu, & Bengio, 2012) is a three-layer undirected graphical model. The first layer consists of visible input variables, the second layer consists of hidden variables (units), and the third layer represents observable output variable. A 1-to-K coding scheme is conducted and the results show that the representing output as a binary vector of length K denoted by y, such that if the output (or class) is k, then all elements are zero except element y_k that takes the value 1. We allow only the inter-layer connections, i.e., there are no connections within a layer.

There are several advantages for ClassRBM. One is that crucial conditional probabilities which are further used in the inference that can be calculated analytically (Larochelle & Bengio, 2008; Larochelle et al., 2012). One important advantage of the ClassRBM is that for enough number of hidden units this model can represent any distribution over binary vectors and its likelihood can be improved by adding new hidden units, unless the generated distribution already equals the training distribution (Le Roux & Bengio, 2008; Martens, Chattopadhya, Pitassi, & Zemel, 2013).

ClassRBM can be used as a standalone classifier for credit repayment status but since it is a hard-to-interpret black box model, it is not commonly used in credit scoring. It can be also used to determine relevancy of inputs by calculating conditional probabilities.

In the credit scoring context this conditional probability expresses the probability of occurring i-th applicant's characteristic for given other inputs (characteristics) and the class label, e.g., when y=1 represents the unpaid credit, we can assume the important (relevant) inputs among the unreliable credit applicants

The volume of credit consumers have been drastically increased by several magnitudes (Crook et al., 2007) since a long time ago. These large increases in lending and credit applications impose development of automatic tools for consumer credit risk assessment[7] including statistical and machine learning methods that constitute a general class of methods known as credit scoring.

In general, credit scoring is the assessment of the risk associated with a consumer (an organization or an individual) that apply for the credit (Crook et al., 2007; Hand & Henley, 1997). Therefore, the problem of credit scoring can be stated as a discrimination between those applicants whom the lender is confident will repay credit and those applicants who are considered by the lender as insufficiently reliable.

One advantage of the scoring table is that it is a comprehensible because of its simple interpretation. Each i^{it} input can get a fixed amount of points equal w_i and the cutting point corresponds to a minimal sum of points at each the positive decision about the credit consumer is made.

Additionally, in the case of refusing the credit to an applicant the reason of the decision can be easily explained by the weights and the cutting point. Another advantage of the scoring table is that it can be almost effortlessly modified by a human expert. Both the weights w and the cutting point s can be tuned by hand, i.e., if the domain expert decides that some value is inappropriate, it can freely increase or decrease the values of the parameters.

6.3 Multilayer Neural Network

The first deep learning that has been utilized is multilayer neural network (MLP) [20]. Selection is done here because MLPs pass information forth to output by using feedforward links. The loss function correlates to the number of links (weight parameters)[8]. Therefore, the performance of deep learning model depends on the loss function and optimization method. When the number of hidden layers and nodes are large, the model will overfit.It is more not suitable to learn from given small dataset [10]. Hence, we implement a grid search with cross-validation so that we can define the optimal number of nodes in three hidden layers neural network for each dataset.

Figure 10 specifies the framework of our proposed algorithm to build the model. Furthermore, the details are shown in algorithm I. which consists of two stages. First stage, according to mentioned above we define the optimal number of nodes in each hidden layer. Second stage, in order to improve the performance of our model we search effective batch size and epoch number using grid search based on the selected networks [11]. In the end, we evaluate our constructed model using accuracy and area under the Receiver operating characteristic curve (AUC).

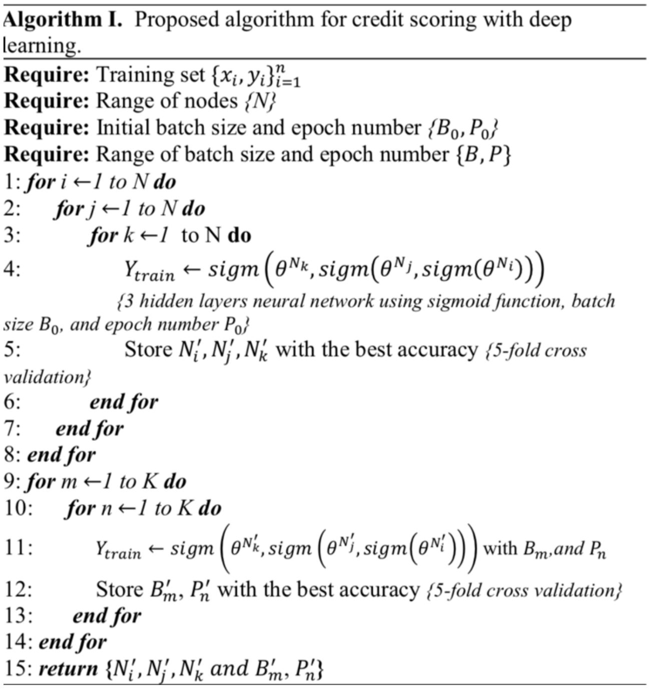
6.4 Classification Algorithms

Financial institutes need formal tools for lending decisions. A credit score is a model-based estimate of the probability that how the borrowers will behave in the future, whether or not he/she will default. In application scoring, for example, lenders employ predictive models, called scorecards, to evaluate how likely an applicant is to default. This PD (probability of default) scorecards are developed using classification algorithms (e.g., Hand & Henley, 1997).

In 2003, alternative classifier Baesens was found promising. Multiple classifier systems[27] were more promising theoretically yet no attempt has been made to analyse the potential of such approaches for credit scoring.

Nonetheless, PD modelling literature indicates that these developments have received little attention in credit scoring, and shows further limitations of previous studies as follows:

- Lack of using large data sets/using small data sets,
- Not comparing different state-of-the-art classifiers to each other,
- Using only a small set of conceptually similar accuracy indicators.



Algorithm I. Proposed algorithm for credit scoring with deep learning.

```
Require: Training set $\{x_i, y_i\}_{i=1}^{n}$
Require: Range of nodes {N}
Require: Initial batch size and epoch number $\{B_0, P_0\}$
Require: Range of batch size and epoch number $\{B, P\}$
for i ←1 to N do
    for j ←1 to N do
        for k ←1 to N do
            $Y_{train} \leftarrow sigm\left(\theta^{N_k}, sigm\left(\theta^{N_j}, sigm\left(\theta^{N_i}\right)\right)\right)$
            {3 hidden layers neural network using sigmoid function, batch size $B_0$, and epoch number $P_0$}
            Store $N_i', N_j', N_k'$ with the best accuracy {5-fold cross validation}
        end for
    end for
end for
for m ←1 to K do
    for n ←1 to K do
        $Y_{train} \leftarrow sigm\left(\theta^{N_k'}, sigm\left(\theta^{N_j'}, sigm\left(\theta^{N_i'}\right)\right)\right)$ with $B_m$,and $P_n$
        Store $B_m', P_n'$ with the best accuracy {5-fold cross validation}
    end for
end for
return $\{N_i', N_j', N_k'$ and $B_m', P_n'\}$
```

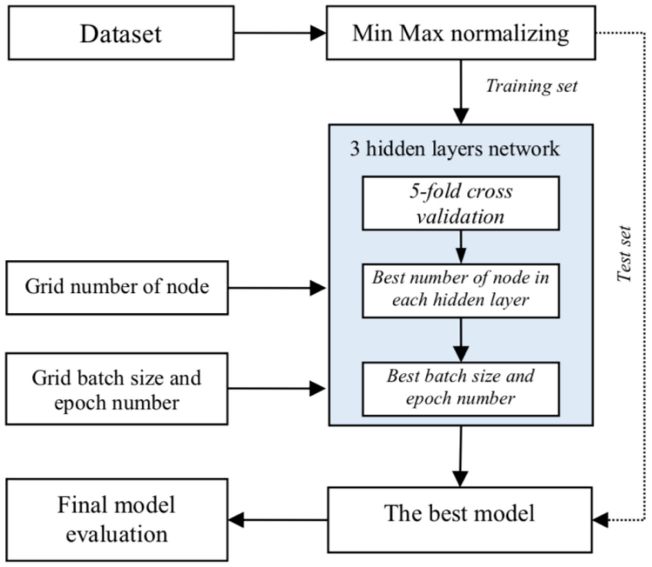
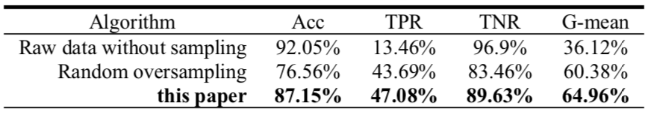


Figure 10: Proposed Algorithm Structure

7 Evaluation

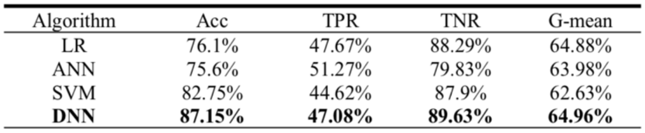
In order to evaluate the performance of the clustering and merging sampling, we conducted the analysis of DNNs model parameter, with different data sampling algorithms and evaluation algorithms while doing the experiments. The training set and test set are divided into a 70% to 30% ratio randomly and DNN-classifiers adopts the 5-fold cross validation, with a learning rate of 0.001, number of neurons in each layer 200, ReLu activation function, iteration of 50 and mini-batch size of 100. With all of these parameters, we compare the clustering and merging sampling algorithm against other sampling algorithms and find out that the total accuracy, TPR (Total Positive Rate), TNR(Total Negative Rate) and G-mean classification for the test set of clustering and merging sampling algorithm results are higher than that of raw data without sampling and random sampling approach. Finally, the results are shown below in table 3.



Algorithm	Acc	TPR	TNR	G-mean
Raw data without sampling	92.05%	13.46%	96.9%	36.12%
Random oversampling	76.56%	43.69%	83.46%	60.38%
this paper	**87.15%**	**47.08%**	**89.63%**	**64.96%**

Table 3: Model Classification Results for Different Data Sampling Methods

For different evaluation algorithms, we compare the classification accuracy of DNNs model against other models such as LR, ANN, SVM, accuracy, TPR, TNR and G-mean are also compared, and it shows that DNN model achieves the highest accuracy of 87.15% among all other algorithms as shown in table 4:



Algorithm	Acc	TPR	TNR	G-mean
LR	76.1%	47.67%	88.29%	64.88%
ANN	75.6%	51.27%	79.83%	63.98%
SVM	82.75%	44.62%	87.9%	62.63%
DNN	**87.15%**	**47.08%**	**89.63%**	**64.96%**

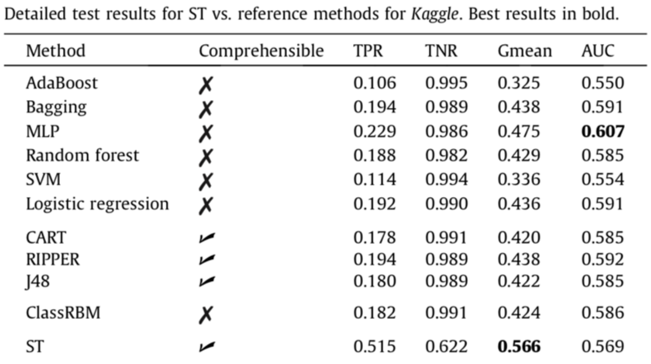
Table 4: Credit Evaluation Resultsof different models

Therefore, DNNs can be leveraged to build multiple hidden layer network models in the application of credit risk assessment. From customer credit data, we can then obtain future information to improve the classification accuracy. Clustering and merging sampling algorithm is not limited to class imbalance problem for classifiers, thus is proved to generate balanced data and provide data diversity. The experiments show that the model purposed has relatively higher accuracy and better performance on imbalanced credit data.

Currently, in supervised classification, the classification model evaluation is an integral part of the process of developing a classification model, and there are well-accepted evaluation measures and procedures. However, cluster evaluation is not a still under development and is not commonly used in part of cluster analysis because of its very nature. Nonetheless, cluster evaluation, or cluster validation is important because it is a necessary part of an exploratory data analysis process.

For ClassRBM, the method for constructing the scoring table is comprehensible and achieves high predictive performance. In the proposed approach the ClassRBM is trained directly on the data which is further used to compute the relevancy of each attribute that is essential component of the scoring table. We compared the performance of the credit scoring table (ST) trained using the approach proposed in this paper with the results from the following non-interpretable reference methods: AdaBoost, Bagging, Multilayer Perceptron (MLP), Support Vector Machines (SVM)[6], and Logistic Regression. We also took under consideration typical comprehensible models like classification and regression trees (CART), RIPPER and J48. In order to determine model parameters we performed the model selection using the validation set and collection of parameters' values, e.g., number of base learners of an ensemble. Moreover, we evaluated the performance of ClassRBM, as a standalone classifier, that can be also used for credit repayment prediction.

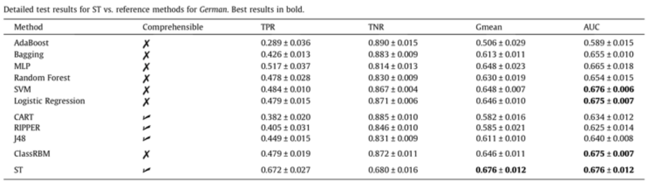
The experiment results are shown by different regions in tables 6-8 below. Clearly, ST approach outperformed the reference methods on German, Kaggle and Short-Term Loans datasets, which are characterized by high imbalance ratio and low general predictive accuracy.



Detailed test results for ST vs. reference methods for *Kaggle*. Best results in bold.

Method	Comprehensible	TPR	TNR	Gmean	AUC
AdaBoost	✗	0.106	0.995	0.325	0.550
Bagging	✗	0.194	0.989	0.438	0.591
MLP	✗	0.229	0.986	0.475	**0.607**
Random forest	✗	0.188	0.982	0.429	0.585
SVM	✗	0.114	0.994	0.336	0.554
Logistic regression	✗	0.192	0.990	0.436	0.591
CART	✓	0.178	0.991	0.420	0.585
RIPPER	✓	0.194	0.989	0.438	0.592
J48	✓	0.180	0.989	0.422	0.585
ClassRBM	✗	0.182	0.991	0.424	0.586
ST	✓	0.515	0.622	**0.566**	0.569

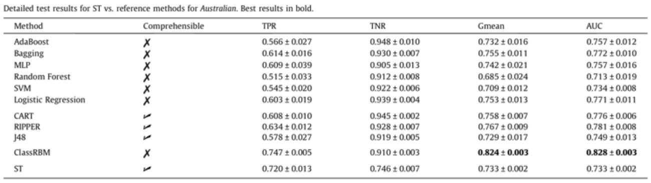
Table 5: ST vs. reference methods for Kaggle



Detailed test results for ST vs. reference methods for *German*. Best results in bold.

Method	Comprehensible	TPR	TNR	Gmean	AUC
AdaBoost	✗	0.289 ± 0.036	0.890 ± 0.015	0.506 ± 0.029	0.589 ± 0.015
Bagging	✗	0.426 ± 0.013	0.883 ± 0.009	0.613 ± 0.011	0.655 ± 0.010
MLP	✗	0.517 ± 0.037	0.814 ± 0.013	0.648 ± 0.023	0.665 ± 0.018
Random Forest	✗	0.478 ± 0.028	0.830 ± 0.009	0.630 ± 0.019	0.654 ± 0.015
SVM	✗	0.484 ± 0.010	0.867 ± 0.004	0.648 ± 0.007	**0.676 ± 0.006**
Logistic Regression	✗	0.479 ± 0.015	0.871 ± 0.006	0.646 ± 0.010	**0.675 ± 0.007**
CART	✓	0.382 ± 0.020	0.885 ± 0.010	0.582 ± 0.016	0.634 ± 0.012
RIPPER	✓	0.405 ± 0.031	0.846 ± 0.010	0.585 ± 0.021	0.625 ± 0.014
J48	✓	0.449 ± 0.015	0.831 ± 0.009	0.611 ± 0.010	0.640 ± 0.008
ClassRBM	✗	0.479 ± 0.019	0.872 ± 0.011	0.646 ± 0.011	**0.675 ± 0.007**
ST	✓	0.672 ± 0.027	0.680 ± 0.016	**0.676 ± 0.012**	**0.676 ± 0.012**

Table 6: ST vs. reference methods for German

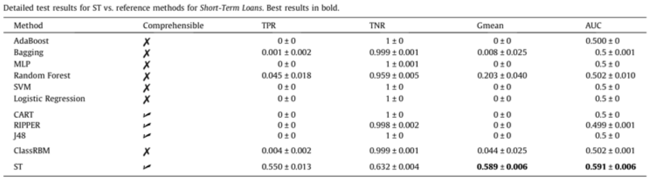


Detailed test results for ST vs. reference methods for *Australian*. Best results in bold.

Method	Comprehensible	TPR	TNR	Gmean	AUC
AdaBoost	✗	0.566 ± 0.027	0.948 ± 0.010	0.732 ± 0.016	0.757 ± 0.012
Bagging	✗	0.614 ± 0.016	0.930 ± 0.007	0.755 ± 0.011	0.772 ± 0.010
MLP	✗	0.609 ± 0.039	0.905 ± 0.013	0.742 ± 0.021	0.757 ± 0.016
Random Forest	✗	0.515 ± 0.033	0.912 ± 0.008	0.685 ± 0.024	0.713 ± 0.019
SVM	✗	0.545 ± 0.020	0.922 ± 0.006	0.709 ± 0.012	0.734 ± 0.008
Logistic Regression	✗	0.603 ± 0.019	0.939 ± 0.004	0.753 ± 0.013	0.771 ± 0.011
CART	✓	0.608 ± 0.010	0.945 ± 0.002	0.758 ± 0.007	0.776 ± 0.006
RIPPER	✓	0.634 ± 0.012	0.928 ± 0.007	0.767 ± 0.009	0.781 ± 0.008
J48	✓	0.578 ± 0.027	0.919 ± 0.005	0.729 ± 0.017	0.749 ± 0.013
ClassRBM	✗	0.747 ± 0.005	0.910 ± 0.003	**0.824 ± 0.003**	**0.828 ± 0.003**
ST	✓	0.720 ± 0.013	0.746 ± 0.007	0.733 ± 0.002	0.733 ± 0.002

Table 7: ST vs. reference methods for Australians

For Multilayer Perceptron Network (MLP), in the experiment, the optimal number of nodes, batch size and epoch are listed in table 10.



Detailed test results for ST vs. reference methods for *Short-Term Loans*. Best results in bold.

Method	Comprehensible	TPR	TNR	Gmean	AUC
AdaBoost	✗	0 ± 0	1 ± 0	0 ± 0	0.500 ± 0
Bagging	✗	0.001 ± 0.002	0.999 ± 0.001	0.008 ± 0.025	0.5 ± 0.001
MLP	✗	0 ± 0	1 ± 0.001	0 ± 0	0.5 ± 0
Random Forest	✗	0.045 ± 0.018	0.959 ± 0.005	0.203 ± 0.040	0.502 ± 0.010
SVM	✗	0 ± 0	1 ± 0	0 ± 0	0.5 ± 0
Logistic Regression	✗	0 ± 0	1 ± 0	0 ± 0	0.5 ± 0
CART	✓	0 ± 0	1 ± 0	0 ± 0	0.5 ± 0
RIPPER	✓	0 ± 0	0.998 ± 0.002	0 ± 0	0.499 ± 0.001
J48	✓	0 ± 0	1 ± 0	0 ± 0	0.5 ± 0
ClassRBM	✗	0.004 ± 0.002	0.999 ± 0.001	0.044 ± 0.025	0.502 ± 0.001
ST	✓	0.550 ± 0.013	0.632 ± 0.004	**0.589 ± 0.006**	**0.591 ± 0.006**

Table 8: ST vs. reference methods for Short-Term Loans

Dataset	**Nodes 1**	**Nodes 2**	**Nodes 3**	**Batch size**	**Epoch**
German	16	64	8	64	250
Australian	32	64	16	32	150
Japanese	16	32	4	16	150
Taiwan	32	4	4	8	200

Table 9: The optimal number of nodes, batch size and epoch.

By using the grid search algorithm of 125 results with a different number of nodes for each dataset as shown in figure 11. We chose the number of nodes which has the highest accuracy and then taken into next step. The effective batch size and epoch number were selected in this step based on the optimal number of nodes.

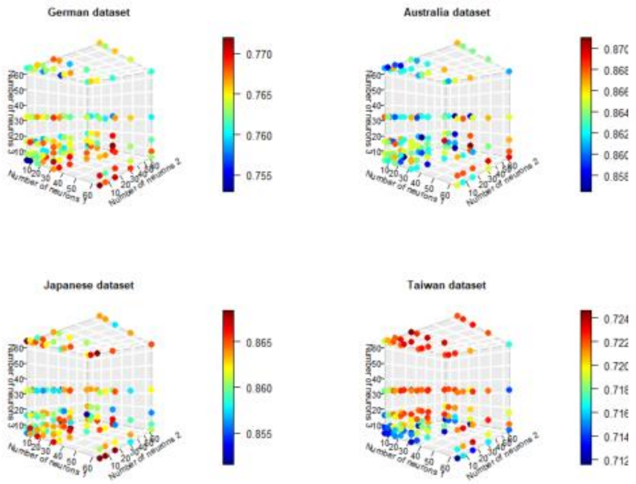
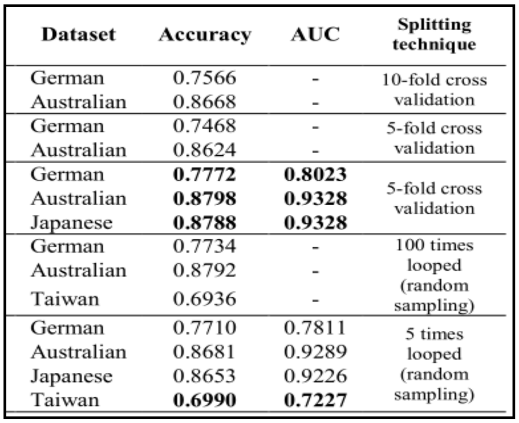


Figure 11: The result of grid search algorithm. The color dynamics represent the accuracy

Finally, we assessed the model using the optimal number of nodes, batch size, and epoch number based on the test dataset. The experiments are looped for five times based on the evaluation criteria for these experiments are averaged to compare with other studies. Table 10 shows the model accuracy and performance comparison.



Dataset	Accuracy	AUC	Splitting technique
German	0.7566	-	10-fold cross validation
Australian	0.8668	-	
German	0.7468	-	5-fold cross validation
Australian	0.8624	-	
German	**0.7772**	**0.8023**	5-fold cross validation
Australian	**0.8798**	**0.9328**	
Japanese	**0.8788**	**0.9328**	
German	0.7734	-	100 times looped (random sampling)
Australian	0.8792	-	
Taiwan	0.6936	-	
German	0.7710	0.7811	5 times looped (random sampling)
Australian	0.8681	0.9289	
Japanese	0.8653	0.9226	
Taiwan	**0.6990**	**0.7227**	

Table 10: The result and performance comparison

For German, Australia and Japanese datasets, the deep learning model achieved the accuracy 0.771, 0.8681, and 0.8653, respectively. The proposed algorithm showed better performance for certain datasets but performed poorly on the other. The reason is that the sample size is too small. Noticeably, for Taiwan dataset, the deep learning model achieved the best performance. Taiwan dataset contains 600 instances and the sample size is much bigger than any other datasets. Consequently, the model outperformed the best previous model by 0.22% [30]. For Classification algorithms, the author concludes that the paper of Baesens et al. (2003) is in need of some updates. Hence, in the paper, he resolves several of the abovementioned issues during recent developments through the following means:

- Conducting a large-scale comparison of many established and novel classifiers including selective ensembles,
- Using multiple data sets of considerable size
- Considering several conceptually different performance criteria
- Using suitable statistical testing techniques.

The paper compares 41 different classification algorithms, in which the inspiration of them comes from several previous studies (e.g., Baesens et al., 2003; Finlay, 2011; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012) and it covers several distinct approaches (linear/nonlinear, parametric/non-parametric, etc.). The algorithms can be split into individual and ensemble classifiers[22] and the final scorecard consists of a single classification model in the first group while the ensemble classifiers integrate the prediction of multiple models, called base models. Then we discern and separate homogeneous ensembles, which create the base models using the same algorithm, and heterogeneous ensembles, which employ different algorithms. Fig.7 demonstrates the whole modelling process using different classifiers.

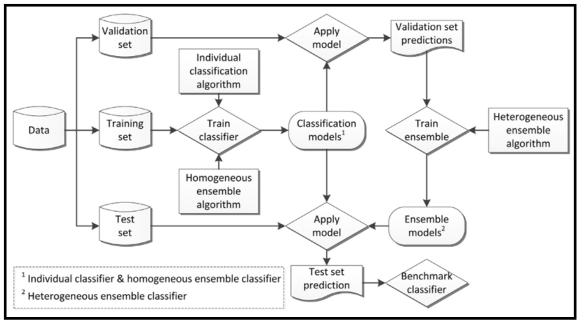

Validation set
Apply model
Validation set predictions
Individual classification algorithm
Data
Training set
Train classifier
Classification models¹
Train ensemble
Heterogeneous ensemble algorithm
Homogeneous ensemble algorithm
Test set
Apply model
Ensemble models²
¹ Individual classifier & homogeneous ensemble classifier
² Heterogeneous ensemble classifier
Test set prediction
Benchmark classifier

Figure 12: Classifier development and evaluation process

Since it is not possible to list and describe all algorithms in detail, due to the large number of classifiers, only the main classification algorithm approaches underneath different classifier families are listed and compared. First and foremost we consider individual classifiers. They pursue different objectives to develop a classification model. The second one is Homogeneous ensemble classifiers. They pool the predictions of multiple base models. The third one is Heterogeneous ensemble classifiers, which combine multiple classification models but create these models using different classification algorithms. The goal of an ensemble is to predict and achieve a high accuracy. Many selective ensembles chose base models to maximize predictive accuracy (e.g., Caruana et al., 2006).

8 Conclusion

It is worth mentioning that the proposed approach has a few practical advantages. First of all, once the parameters of the scoring table are determined, the model can be easily implemented in typical bank system. What's more, the prior knowledge of a human expert can be easily incorporated into the model by modifying scoring points. Additionally, some comprehensible models, such as decision trees, may have complex structure and thus it may become difficult to interpret. In our approach the structure is fixed and depends on the number of attributes. The proposed scoring table has some limitations. The need of binarizing the input variables may lead to losing information in data. However, the input binarization is necessary in learning the ClassRBM and the scoring table. Furthermore, the ClassRBM, which is essential in constructing the scoring table, consists of relatively large number parameters which may result in overfitting the model during learning. Therefore, it is important to apply regularization technique.

Several improvements are needed to be addressed in the outlined approach. First, the quality of the proposed model using large number of datasets from other domains should be carefully addressed. Second, criterions other than G-mean in the training procedure of the scoring table will be much more interesting. Moreover, the training procedure could be modified to increase other evaluation metrics, for example., AUC. Last but not least, it is worth considering incorporating cost-matrix in the process of constructing the scoring table.For Multilayer perceptron network (MLP), the author Munkhdalai [19] applied deep learning approach with grid search method for credit scoring. They trained 4 different datasets on 3-hidden layers neural network. The proposed algorithm demonstrated promising results particularly for bigger sample dataset. It achieves the best performance by deep learning. This paper presented the case where deep learning with bigger data sample can have better performance than other. We can expect more potential future work in this area that includes developing deep learning model for credit scoring based on the bigger dataset.

For classification algorithms, the results indicate that several classifiers predict credit risk significantly more accurately than the industry standard LR particularly heterogeneous ensembles classifiers perform well. It is also proved that more accurate scorecards will facilitate sizeable financial returns. Finally, it can be concluded that several common performance measures have similar signals as to which scorecard is the most effective and the use of two rarely employed measures are recommended since they contribute additional information.

From a managerial perspective, it is important to note whether the excellent performance that we observe for some classifiers generalizes to real-world applications, and to what extent they can be adopted so as to increase returns. From this study, we learn some insightful ideas; First, we show that advancements in computer power, classifier learning, and statistical testing facilitate rigorous classifier comparisons. This does not guarantee external validity. Though there are several concerns on why laboratory experiments may overestimate the advantage of advanced classifiers remain valid and might not be able to be resolved (e.g., Hand, 2006). Secondly, the results make use of some remarks related to the organizational acceptance of advanced classifiers. Specifically, a lack of acceptance can be caused by concerns that much expertise is needed to handle such classifiers.

Results in work [23] show that this is not the case. Lastly, the business value of more accurate scorecard predictions is a crucial issue. The preliminary simulation provides some foundation that the “higher (statistical) accuracy equals more profit equation” might hold. In addition, retail scorecards support a vast number of business decisions.

Regulatory frameworks and organizational acceptance constrain occasionally prohibit the use of advanced scoring techniques today; at least for classic credit products. Nevertheless, given that the current interest in data-centric decision grows and the richness of online-mediated forms of credit granting boost, we foresee a bright future for advanced scoring methods in credit scoring. We have covered several deep learning models in our research report, but the use of deep learning for credit scoring is at its initial phase[29]. It is in the recent times that deep learning has been used for credit risk modelling.

Major financial institutions use AI to detect fraud cases for several years now. For example, JPMorgan Chase has introduced Coin, a platform which uses machine learning. It could review 12,000 credit agreements within few seconds. The staff on the other hand took 360,000 hours per year to analyze the same number of agreements. Morgan Stanley in their article in 2019[30] have claimed that “machine learning is changing the game for the next generation of consumer borrowers”. They have used machine learning to find the curve to learn the income levels of their customers. FICO [31] is the best practitioner in our opinion since they use scorecards, which are a powerful tool that consists of non-linearity in the input layer. The scorecards are transparent, explainable and accurate. The FICO scorecard model is very reliable as it has the best track record. There have been numerous revisions since 1989 to determine an accurate credit score. All companies use different deep learning models that distinguish them from each other. We do not have access to the exact models that different financial companies are using, however, we have mentioned a few models in our research report which are generally adopted by big institutions.

The deep learning methods that we have talked about has its both advantages as well as disadvantages. But in recent researches, work has been done in improving the prediction accuracy. Focus is given on obtaining diverse data to improve our models. There are endless number of possibilities in this field and we are just getting started with deep learning and AI in credit risk assessment.

References

[1] Puneet Chhahira. *AI Adoption*(2018) Access at: (https://www.edgeverve.com/finacle/finacleconnect/trulydigital-2018/ai-adoption/)

[2] Schmidhuber J. Deep learning in neural networks: An Overview[J]. Neural Networks, 2015, 61:85-117

[3] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553):436-444

[4] Y. Li, X. Lin, X. Wang, F. Shen and Z. Gong, "Credit Risk Assessment Algorithm Using Deep Neural Networks with Clustering and Merging," *2017 13th International Conference on Computational Intelligence and Security (CIS)*, Hong Kong, 2017, pp. 173-176.

[5] Martey Addo, Peter & Guegan, Dominique & Hassani, Bertrand. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks. 6. 38. 10.3390/risks6020038.

[6] Harris T. Credit scoring using the clustered support vector machine[J]. Expert Systems with Applications,2015,42(2): 741-750

[7] Crook J N, Edelman D B, Thomas L C. Recent developments in the consumer credit risk assessment[J]. Proceedings of European Journal of Operational Research,2007,183(3): 1447-1465

[8] Angelini, Eliana, Giacomo di Tollo, and Andrea Roli. 2008. A neural network approach for credit risk evaluation. The Quarterly Review of Economics and Finance 48: 733–55.

[9] Bahrammirzaee, Arash 2010. A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications 19: 1165–95

[10] Dominique Guegan, Peter Addo, Bertrand Hassani. Credit Risk Analysis Using Machine and Deep Learning Models. Risks, MDPI, 2018, Computational Methods for Risk Management in Economics and Finance, 6 (2), pp.38.

[11] FRB, SR 11-7: Guidance on Model Risk Management (Apr. 4, 2011); OCC, Bulletin 2011-12, Supervisory Guidance on Model Risk Management (Apr. 4, 2011).

[12] Brainard, L. *What Are We Learning about Artificial Intelligence in Financial Services?* (Nov. 13, 2018).

[13] FSB, Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications (Nov. 1, 2017).

[14] Press, G, *Equifax and SAS Leverage AI and Deep Learning to Improve Consumer Access to Credit*, Forbes (Feb. 20, 2017).

[15] Jakub M. Tomczak⇑, Maciej Zie ˛ba Classification Restricted Boltzmann Machine for comprehensible credit scoring model

[16] Le Roux, N., & Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. Neural Computation, 20, 1631–1649.

[17] Larochelle, H., & Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In Proceedings of the 25th international conference on machine learning (pp. 536–543).

[18] Larochelle, H., Mandel, M., Pascanu, R., & Bengio, Y. (2012). Learning algorithms for the classification restricted Boltzmann machine. Journal of Machine Learning Research, 13, 643–669.

[19] Munkhdalai, Lkhagvadorj & Namsrai, Oyun-Erdene & Ryu, Keun. (2018): Credit Scoring with Deep Learning.

[20] Haykin, Simon S., et al.: Neural networks and learning machines. Vol. 3. Upper Saddle River, NJ, USA: Pearson, 2009.

[21] Hand, David J. “Rejoinder: Classifier Technology and the Illusion of Progress.” *Statistical Science*, vol. 21, no. 1, 2006, pp. 30–34.

[22] Rich Caruana; Art Munson; Alexandru Niculescu-Mizil: Getting the Most Out of Ensemble Selection, 2006

[23] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow , Lyn C. Thomas :Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research

[24] Linda Allen, Ph.D: Credit Risk Modeling of Middle Markets

[25] Altman, E.I., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." Journal of Finance, September 1968, 589-609.

[26] Mester, L., “What’s the Point of Credit Scoring?” Federal Reserve Bank of Philadelphia Business Review, September/October 1997, pp. 3-16.

[27] Finlay, Steven. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research. 210. 368-378. 10.1016/j.ejor.2010.09.029.

[28] Anastasios Petropoulos, Vasilis Siakoulis, Evaggelos Stavroulakis and Aristotelis Klamargias: A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. Ninth IFC Conference on “Are post-crisis statistical initiatives completed?” Basel, 30-31 August 2018

[29] Zhu, Bing & Yang, Wenchuan & Wang, Huaxuan & Yuan, Yuan. (2018). A hybrid deep learning model for consumer credit scoring. 205-208. 10.1109/ICAIBD.2018.8396195.

[30] Machine Learning Is Changing the Game for the Next Generation of Consumer Borrowers. 2019 Access at: https://www.morganstanley.com/im/en-us/registered-investment-advisor/insights/investment-insights/the-impact-of-machine-learning-on-consumer-lending.html

[31] How to Build Credit Risk Models Using AI and Machine Learning. Access at: https://www.fico.com/blogs/analytics-optimization/how-to-build-credit-risk-models-using-ai-and-machine-learning/