Accuracies of some Learning or Scoring Models for Credit Risk Measurement

Salomey Osei, Berthine Nyunga Mpinda, Jules Sadefo-Kamdem, Jeremiah Fadugba

Accuracies of some Learning or Scoring Models for Credit Risk Measurement

Preprint · March 2021

DOI: 10.13140/RG.2.2.22472.44803

4 authors, including:

Salomey Osei
Kwame Nkrumah University Of Science and Technology
12 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Jules SADEFO KAMDEM
Université de Montpellier
90 PUBLICATIONS **268** CITATIONS
SEE PROFILE

Berthine Nyunga
African Masters in Machine Intelligence (AMMI)
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Accuracies of some Learning or Scoring Models for Credit Risk Measurement

Salomey Osei
African Masters of Machine Intelligence
Accra, Ghana
sosei@aimsammi.org

Prof. Jules Sadefo Kamdem
University of Montpelier
Montpelier, France
jules.sadefo-kamdem@umontpellier.fr

Berthine Nyunga Mpinda
African Masters of Machine Intelligence
Accra, Ghana

Jeremiah Fadugba
African Masters of Machine Intelligence
Accra, Ghana

Abstract

Given the role played by banks in the financial system as well, risks are subject to regulatory attention, and Credit risk is one of the major financial risks faced by banks. According to Basel I to III, banks have the responsibility to implement the credit risk strategy. Nowadays, machine learning techniques have attracted an important interest for different applications to financial institutions and its applications have received much attention from investors and researchers. Hence in this paper, we discuss existing literature by shedding more light on a number of techniques and examine machine learning models for Credit risk by focusing on Multi-Layer Perceptron (MLP) and Convolutional Neural Networks (CNN) for credit risk. Different test performances of these models such as back-testing and stress-testing have been done using Home Credit historical data and simulated data respectively. We realized that the MLP and CNN models were able to predict well with an accuracy of 91% and 67% respectively for back-testing. To test our models in stress scenarios and extreme scenarios, we consider a generated imbalanced data with 80% of defaults and 20% of non-default. Using the same model trained on Home Credit data, we perform a stress-test on the simulated data and we realized that the MLP model did not perform well compared to the CNN model, with an accuracy of 43% as against 89% obtained during the training. Thus, the CNN model was able to perform better during stressed situations for accuracy and for other metrics such as ROC AUC curve, recall, and precision.

Keywords: Model Accuracy; Machine Learning; Credit Risk; Basel III; Risk Management

1 Introduction

Despite the raging dangers and insecurities associated with giving out loans, financial institutions such as banks, insurance companies, micro-finance, etc. still consider giving out loans to individuals and corporate industries since it is a major source of income to several banks. Arguably, most financial institutions would choose not to give credit to customers who are not able to pay back the credit given (13). Several statistical methods have been used in the past for credit risk assessments even though these methods have shown some levels of difficulty in the modeling of complex financial systems due to the fact that some functions need to be fixed and other statistical assumptions have to be made (26). It has been shown that the history of credit scoring dates back to the late 1950s where credit decisions were taken based on certain judgmental methods which were commonly called the 5C's techniques

i.e. Character, Capital, Collateral, Capacity, and Condition. The Character is related to how much knowledge the lender has on the borrower such as knowing the person's family or the borrower. The Capital explains the amount that the borrower is asking for from the lender with the Collateral being the amount the borrower is prepared to give based on their resources. Meanwhile, the Capacity goes further to tell the ability of the lender to repay what has been taken. Lastly, the Condition outlines the situation or circumstances in the market. With the 5C's techniques, it was identified that the number of applications processed daily was small resulting in the discovery of scorecards to handle what will represent fair judgments (9). The most commonly used statistical and machine learning techniques in credit scoring include Linear Discriminant Analysis (LDA), Logistic Regression, Naive Bayes, Artificial Neural Networks (ANN), Random Forests, Bagging, Boosting, etc. (9). As soon as credit cards were introduced, the importance of credit scoring models was activated.

A credit scoring model should be able to accurately classify customers into default or non-default groups to save costs incurred by financial institutions. Generally, a credit scoring model is used for modeling and assessing each example in a dataset. This has reduced the use of purely judgmental credit-granting decisions with old methods such as statistical models due to the continuous increase of customer data (15). The elasticity of deep learning models can be able to produce strong support in credit scoring (26) due to the existence of data and the reason for the frequent use of data mining techniques which has received widespread attention in the world (23). For the past years, there has not been a clear best technique for the credit scoring domain. Several research has been performed on feature selection using genetic algorithm as a wrapper to improve the performance of credit scoring models. However, the challenge lies in finding an overall best method in credit scoring problems and improving the time-consuming process of feature selection(13). According to the World Bank Group (WGB) and the International Committee on Credit Reporting, credit scoring has been understood broadly to have enormous potential in helping with the economic growth of the world's economy which is also useful for enhancing financial inclusion, model accuracy improvements and access to credit for individuals. It is also used for discovering the minimum levels of regulatory and economic capital and to support the management of customer relationships to seek potential customers and business opportunities. The technologies that support innovative credit scoring are continuously evolving with time and most financial institutions are still operating based on the judgments of credit officers, judgmental scorecards, or the use of traditional regression models. Debtor's details are applied to the credit scoring model which helps in predicting the chances of repayment of loans on time. Therefore, it is important to develop a model for credit scoring which have the right variables. The main indicators to consider in credit scoring are the financial indicators, demographic, employment, and behavioral indicators. The financial indicators deal with the capacity of a borrower to repay a loan whiles the employment indicators show whether the borrower is employed or not. On the other hand, the demographic indicators give an idea of measurable characteristics such as race, sex, age, and others, and the behavioral indicators give a clear credit history of the borrower. Aside from these factors, there are certain factors that are mostly considered to also give an in-depth knowledge about the borrower to the lender. These factors include life events such as loss of jobs, unforeseen problems, sickness and disability, and credit lending practices where lenders charge excessive interest and bank charges, huge up-front to start repayment plans which can lead to late loans (13). A credit score may be determined by five categories namely; the types of credit, the payment history, the length of credit, new credit, and the current debt (11).

Machine learning as has been evident gives us powerful tools for quantitative finance but the dangers in misapplying the techniques lead to disappointments (3). In as much as we need to be careful of the use of these methods, the tools offered by machine learning give many potential applications to finance. To address these challenges and to have highly reliable models, it is important to do testing of the models. This is extremely important because according to (27), in the wake of the 2008/2009 financial crises in which stress testing exercise was made on large US banks mainly to know the ability of these banks to withstand crises, several criticisms were made about the existing methods that were used. In this case, it has now become difficult to tell how a model is performing without testing it against shocks and impacts for the future. The main and important question that is being asked is whether banks can survive the next financial crisis (27). Before these crises, most of the financial institutions were considered to be well-capitalized according to the standards that had been set by regulators and believed that the internal risk models of the financial institutions were not excessively out of line with what the regulators had (21). Hence in this paper, we will use stress testing and back-testing to examine and evaluate the performance of some of the deep learning techniques for credit risk knowing that banks generate most of their profits through lending. Our

aim will be to compare the accuracies of these methods.The form of model evaluation in machine learning is mainly splitting the data into train and test set, while the training is used in preparing the model and the test set for evaluating the model. In doing this, the idea is that the model will use the test set for making predictions for that period to provide insight on how the model will perform operationally. The reason why these evaluations are not considered to be the best is that any estimate of the performance on the train data is optimistic, meaning decisions made based on this performance will be biased. In credit risk, we are interested in the performance of the model on data that the model has not seen or out of sample data. In back-testing, we use historical data by splitting up the data. We use part of the data to prepare the model and the other for making predictions. This is done so that the model will remember the timestamps and the value for each observation to enhance performance. Also, other methods for testing historical data assumes that there is no relationship between the observations and that each of them is independent. The data used for credit risk is seen as a time series data where the dimension of observations means that we will not be able to split them randomly into groups but rather split the data while respecting the temporal order in which the values are observed. This research enables us to further strengthen the risk measurement and management of deep learning models for credit risk [1].

2 Literature Review

We have witnessed the significant and rapid growth of machine learning to credit scoring and the variety of scoring applications mainly because of better access to a wide variety of large datasets, the advancement of programming architectures and the demand for improvements in efficiencies. Before giving a loan financial institutions performs what is called credit worthiness, an evaluation which is described as credit scoring. This evaluation is a procedure based on numeric expressions which allows banks to represent the likelihood of default for a customer (20). Before the advancement of technology or the consideration of machine learning techniques for credit scoring, there existed methods that were used to appropriately score a customer or individual's potential of passing for a loan. These terms were known to be the statistical methods that have been used for decades and now needs improvement. For instance, according to the Xolani Dastile and Turgay Celik's paper which discusses several commonly used statistical and machine learning techniques (9), logistic regression has been mostly used for evaluating credit worthiness because it is simple and transparent during predictions. They pointed out that machine learning models if discovered are a great replacement for the logistic regression model which has been considered by financial institutions for credit scoring. In as much as the applications of machine learning models in credit scoring are a great replacement, it is pointed out that these models are unable to explain predictions and the issue of imbalanced datasets. Their conclusion and final results shows an ensemble of classifiers performs better than single classifiers. Besides, it was shown that deep learning models such as Convolutional Neural Networks performed better compared to statistical and classical machine learning models.

Although both statistical and machine learning models have a common objective to learn from the data given, both methods investigates the underlying relationships with the help of training data. Statistical methods assume formal relationships between variables in the form of mathematical equations while machine learning methods can learn from data without requiring any rules based programming. Due to its flexibility, machine learning methods can better fit the patterns in the data (4). Addo et al (1) contributed by building a binary classifier for the prediction of loan default probability with six approaches using elastic net approach as a benchmark. It was observed after choosing the ten most important features that, tree-based models are stable than the multi-layer artificial neural networks. Because existing statistical methods such as logistic regression and different architectures (support vector machine, multi-layer perceptrons) focuses mainly on the output of the classifiers at the abstract stage. This enables the neglecting of several rich information unseen in the confidence degree and causes difficulties especially with real world applications. To address these setbacks, Hinton (18) introduces a deep belief network as an ensemble technique which captures relevant information in the confidence degree.

In relation to Deep Beliefs, Luo (26) also performed a comprehensive study using credit default swaps with the main goal being to produce an illustrative outcome and test that provides a foundation for later theoretical research on Deep Belief Networks in corporate credit scoring related to credit default swaps. The outcome of this paper according to classification accuracy and the area under

[1] https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting

the receiver operating characteristic curve shows deep belief network with restricted Boltzmann Machines performs better than the baselines such as logistic regression, multi-layer perceptron, and support vector machines and consistent on various rating categories. R. Y. Goh (15) presented credit scoring models and credit-granting decisions and suggested that the approach of artificial intelligence is a successful technique in credit scoring and concluded that the SVM is an active domain where works are still ongoing in credit scoring, and Hybrid MA-DM has been the leading model type throughout the years, and it is deemed that this approach will somehow persist shortly since hybrid models formulation is considered as a direct way to propose new models. Also, they extended their findings and proposed that the use of Support Vector Machines is widely seeking attention based on different finders and researchers for helping institutions in identifying defaulters and non-defaulters. They also mentioned that SVM is the most data mining technique used for building credit scoring and has four main classes namely, standard SVM and its variants, modified SVM, hybrid SVM, and ensemble models.

In (34), the authors uses the fact that there exist relatively large amounts of irrelevant and redundant features in credit data which leads to inaccurate predictions to come up with a proposed model called NCSM (Novel Credit Scoring Model), which regards the information entropy as heuristic for selecting the optimal features. This model is based on feature selection and grid search and it is intended to optimize the random forest algorithm. The result using the UCI datasets shows that the proposed method has high performance in improving the prediction accuracy.In Yosi Lizar (13), they proposed Logistic Regression as one of the most used methods in predicting non-performing loans. In cases where there is no trend of data available, it was suggested that Decision-Making Trial and Evaluation Laboratory (DEMATEL) technique and the Analytic Hierarchy Process (AHP) technique is preferable for credit scoring. Finally, they were able to show that the two mechanisms which are AI-based and statistical-based techniques are the high-ranking credit scoring models and the choice for bankers. (19) introduces a preprocessing step, a method for designing the architecture of an efficient ensemble classifier called class-wise classification by incorporating different data mining methods such as Bayesian Network for augmenting the ensemble classifier. The Markov blanket concept enables a natural form for selecting features thereby providing a basis for mining association rules. Gabriella et al (30) used a discriminant analysis approach which positively indicates the kind of randomly selected data to be used for credit scoring, and the presume data is a specimen from a multivariate normal distribution. It also reduces the dimensions without losing relevant information. In their conclusion, multivariate data analysis was proposed to be a productive modus operandi when pulling out information while using a discriminant analysis approach.

In credit scoring analysis, the main goal is to get a good accuracy for the model which is being used. From these highly driven and illustrated techniques used for credit risk, it is well established that machine earning techniques however have been used for analyzing and solving problems related to credit risk. As such, there is the need to evaluate these models and to find out whether these models are efficient for the tasks. What we fail to understand is the fact that these models are not perfect and hence we need a criterion that can tell us when a model is performing well or not. In this case, it is necessary to have a look at the models' accuracy by performing several tests. In relation to this, (27) proposed a generative deep learning method for bank stress-testing with adversarial modules which allows the objective function of the neural network to rely on banks' performance prediction independently and also makes it possible to assess the risks that are seen in smaller banks. Since financial crises have risen in the past years, (21) discusses "The validation of machine learning models for stress testing of credit risk" and proposes a method, Multivariate Adaptive Regression Splines(MARS), and validate against Vector Autoregression (VAR). It is discovered that MARS exhibited higher accuracies in terms of the out-of-sample data using various metrics and produces good forecasts. Unfortunately, not much work has been done for credit risk using stress-testing and back-testing, a method that has been shown to be effective.

Table 1: Summary of existing review articles and the present study

Authors	Year	Journal	Model Used	Datasets	Review Type	Outcome
R.Goh and Lee (15)	2019	Advances in operations research	SVM and Meta-heuristic approaches (MA)	UCI credit datasets	Comprehensive literature review	Review of AI techniques with detailed discussions
Yosi and Engku(13)	2019	Journal of Advanced Research in Business and Management Studies	Statistical and ML based models	—	Brief literature review	Ai-based methods gives higher-ranking scoring models but not user friendly
Frederico(20)	2020	ResearchGate	—	—	Comprehensive literature review	Discusses most common ML methods in previous years
Siddharth et al(6)	2020	Journal of Banking and Financial Technology			Systematic literature review	
Usha and Neera (11)	2020	Journal of Critical Review	SVM, MDA(Multiple discriminant analysis),RS(Rough sets), LR(Logistic regression), ANN, CBR(case based reasoning), DT(Decision tree), GA(Genetic algorithm), KNN, XGBoost and DGHNL(Deep Genetic Hierarchical Network of Learners)	—	Systematic literature review	Framework for selecting the best fit tool for CRPM developers to make informed decisions.
—	Present	—	—	Existing and new datasets	Comprehensive	Using methods such as stress testing, back testing, etc., to examine and evaluate the performance on the existing models for credit risk

3 Background

3.1 Credit Risk

The risk that the value of a portfolio will decrease due to the fact that a borrower defaulted or because they are not meeting obligations is termed as credit risk (29). One of the common methods used for credit analysis is Value at Risk (VaR). In order to cover unexpected losses, financial institutions especially the banks are restricted to have some form of capital cover, this is to improve the credit risk of financial institutions under the regulations set by the Basel II and III. Credit risk models can be used for underwriting, loan acceptance, pricing, provisioning, and capital calculation.

Credit Risk = Default Probability $\times$ Exposure $\times$ Loss Rate

The terms in the credit risk formulae are defined below:

1. Default Probability which is the probability of a debtor going back on debt payment
2. Exposure which is the total amount including interest payments that is borrowed by the debtor
3. Loss Rate which is calculated as 1 - Recovery Rate. To further explain, recovery rate is the total proportion that the lender receives from the borrower in case of default.

Credit scoring shows the conventional mathematical and statistical methods that are used for credit granting (17). Credit risk has been considered by most researchers in the field of AI and machine learning due to its implication on the economy of the world if not managed well. Some of the implications have been tied to climate as evident by (7) paper which investigates climate change and credit risk and points out that climate risk is a factor that influences the creditworthiness of loans by cooperates. Several criteria may be required for credit risk prediction models to gain proper efficiency. This criterion is mostly dependent on the model developer whose intention is to achieve better accuracy. In other words, there is the need to have models that are transparent and high accuracy is sort after. According to (11), thirteen criteria has been identified by researchers to be the most common and important namely; accuracy, transparency, and interpretability of the results, sample size, model failure, selection of good variables, dispersed data handling, sensitivity to collinear variables, updates, integration, avoiding over-fitting, assumptions imposed by the model and types of variables.

3.2 Deep Learning

Deep learning, first introduced as a term by (10) and then called neural networks by Igor Aizenberg and others (2) has been used in several areas including credit risk. Deep learning is known to be an artificial intelligence function that mimics the human brain in its processing of data and pattern creations inspired by the information processing and distributed communication between nodes in biological systems. Under machine learning, it is capable of learning from unsupervised data which is not structured or labeled. It is often able to learn without the supervision of humans and has been used in several areas including fraud detection, credit risk, and many more. Deep learning methods are made up of adding several layers to a neural network with the word "deep" coming from the use of several layers in the network. A standard neural network is made up of several simple connected processors known as neurons which produce a sequence of real-valued activation (1) or neurons put together in a form of a web. The basic unit for computation in a neural network is the neuron that receives input from other nodes to compute an output and is also associated with weights that are assigned to each layer. Deep learning is useful for huge amounts of unstructured data which in the normal sense can take humans decades to unravel by extracting high-level features from the raw data and also enables the computer to learn from experience in terms of hierarchical concepts (16).

Each level of the neural network leans to transform the input data into a theoretical and complex representation. Deep learning methods depend typically on four types of structures, namely; Convolutional Neural Networks (CNN's), Recurrent Neural Networks (RNN's), Recursive Neural Networks (RNN's) and the normal Standard Deep Neural Networks (1). Deep learning uses a strategy called backpropagation which was first introduced by Yann Lecun and others (25) in a paper that utilized backpropagation to handwritten zip code recognition. According to (32), deep learning has to do with credit assignment with long chains of potentially causal links between actions and consequences. The neurons in a neural network are represented by circles that are interconnected. The neurons are

grouped into three types of different layers namely; the input layer (receives input data), hidden layer (mostly responsible for performing several mathematical computations on the input data), and output layer (returns the output) in that order. The connections between the neurons are associated with an initial weight indicating the importance of an input value. The neurons also have "Activation functions" which help in standardizing the output from the neuron. After each iteration, the weights between the neurons are adjusted using a technique called Gradient Descent whose task is to reduce the cost function.

3.3 Back-testing

Back-testing is known to give traders and investors an option to evaluate trading models before implementation. Moreover, there is a risky assumption that is a key part of back-testing which indicates that the performance of the past determines the performance in the future. The fear is the notion that the method that has worked poorly in the past will work poorly in the future. Back-testing is generally a process in which a trading strategy or method is applied to historical data to determine how the method or strategy would have predicted the real results. In essence, back-testing is done to determine whether a model can give a predictive value. A back-test on historical data allows researchers to evaluate the risk profile of an investment strategy before funds are committed (5). Back-testing is for monitoring the performance of the default client and implies that the model is trained on data from a certain time period and then tests its performance on older data.

3.4 Stress-testing

Stress-testing is defined as a type of performance test that checks the upper limits of a model, by testing it under utmost loads. Research on stress-testing of financial risk models has increased since the financial crises in 2008 and it has become a central tool used by global regulators to manage the financial stability of models. In the light of model risk management in the 1990s, stress-testing as a tool has been used to tackle the question of how exposures or models behave under unfavorable conditions. Stress-testing was first introduced within the Basel 1 (14) through the 1995 Market Risk Amendment (21) and hence stress-testing with respect to credit risk has evolved as a separate discipline. In the framework of both Basel II and III, credit risk remains the greatest form of risk for financial institutions (27), therefore, stress-testing of PD, EAD, and LGD is mandatory for larger financial institutions. It helps to determine the model's robustness and error handling under extreme load conditions. Then, stress-testing machine learning models before being used helps to know the model's robustness, to fix the bottlenecks, and also to make sure if the model can be still used to solve another problem. Given our model, find another data where we can test the performance and analyze if the model risk changes in a period with greater variability.

4 Some advanced Learning techniques for Credit Risks and Scoring

4.1 Multi-Layer Perceptron (MLP)

A multi-layer perceptron, consisting of one or more hidden layers is a feed-forward neural network that mainly consists of interconnected neurons that transfer information to each other. As a deep learning technique, it consists of three main layers (input, hidden, and output) with each neuron assigned a value and the connections between each layer assigned weights. The values of the neurons are gotten using mathematical functions that take into account the weights and values of the previous layer. The weights propagate values through the network to produce an output. MLP learns non-linear functions through a process called back-propagation to update the weights accordingly. In a classification task, activation functions such as soft-max are used at the last layer to squash the values into a vector of values that sums up to one (probabilities). Since it is a deep learning method, it is characterized by several layers in a directed graph where each node has a nonlinear activation function with the exception of the input nodes.

4.2 Convolutional Neural Network (CNN)

CNN is one of the most popular types of deep neural networks which convolves learned features with the input data. CNN gets rid of the need for manual feature extraction by extracting the features

directly. In literature, CNN's have been used for image classifications, recommender systems, Natural Language Processing, and Finance. It is also a regularized version of multi-layer perceptrons that uses the hierarchical pattern in the data to handle over-fitting. The multiple plains in each layer for detecting multiple features are the convolutional layers (24). One dimensional convolution is between a vector of weights $a \in \mathcal{R}^a$ and a sequential input $s \in \mathcal{R}^s$ (22) while the vector a is the filter for the convolution. In mathematical operations, convolution is used in place of the normal matrix multiplication. Common activation function i.e. Rectified Linear Unit (ReLU) is used for CNN because it does not saturate. Activation functions are followed by pooling layers, fully connected and normalized hidden layers masked by the activation function. With pooling (33), there is some form of down-sampling by giving maximum outputs in case of max pooling. There is also a fully connected layer that is connected to all the previous activations next to the last layer (loss function layer) where predictions are made. The input of the CNN is a tensor with shape (N, H, W, D), N; number of images, which becomes the feature map after convolution, H; image height. W; image width and D; image depth. The process of CNN is similar to the response of the visual cortex to a stimulus where the convolutional neuron processes data for its receptive field (12). The neurons receive inputs from each element in the previous layer. The architectures of CNN's are formed by stacking several layers which later transform the input into output through differentiable functions.

5 Methodology

5.1 Cases Studies with Data for Back-Testing and Stress-Testing

5.1.1 DATA for Home Credit Risk

Home credit is an international consumer financial service provider founded in 1997 in the Czech Republic with its headquarters in the Netherlands and operates in countries such as Russia, Kazakhstan, Ukraine, Belarus, China, India, Indonesia, Philippines, Vietnam, and the Czech Republic [2]. The data is a historical dataset on actual clients and loans provided for Home Credit Default Risk for Kaggle competition, collected over different time frames and sources. The dataset contains 7 different details of sources of data in csv files; namely, application_{train/test}.csv, bureau.csv, bureau_balance.csv, previous_application.csv, POS_CASH_balance.csv, installments_payments.csv and credit_card_balance.csv but for this project, we will use 3 datasets due to computational limitations i.e. application_{train/test}.csv, bureau.csv and bureau_balance.csv.

1. **application_train/application_test**: This contains useful information on each of the loan application with each row representing one loan and each loan uniquely identified as "SK_ID_CURR". The training data contains a target with indications: 0 as non default and 1 as default. It contains 122 fields, including number of children, contract type, total income, occupation type, etc. The test dataset contains the same fields as the application dataset with the exception that the target column is missing.
2. **bureau.csv**: This contains information about the clients previous credits and this is mainly provided by other financial institutions that has been reported to a credit bureau. Each previous credit is identified by "SK_ID_BUREAU".
3. **bureau_balance**: This file contains a row for each month where the clients is having data. Each row is linked to the bureau dataset by "SK_ID_BUREAU" and contains 3 fields. It also contains information about how a client repaid monthly.

5.1.2 Exploratory Analysis on Data

We go through each column in the data frame and replace all missing values (in all there are 67 columns with missing values) with a random value chosen among available values in the columns, and also, label-encode non-numerical values. The target distribution was imbalanced hence we up-sampled the minority class. From the correlation that was done, we discovered that the older the clients, the less likely they are to default and vice-versa [3].

[2] https://www.homecredit.net

[3] https://github.com/edpolanco/home_credit_competition

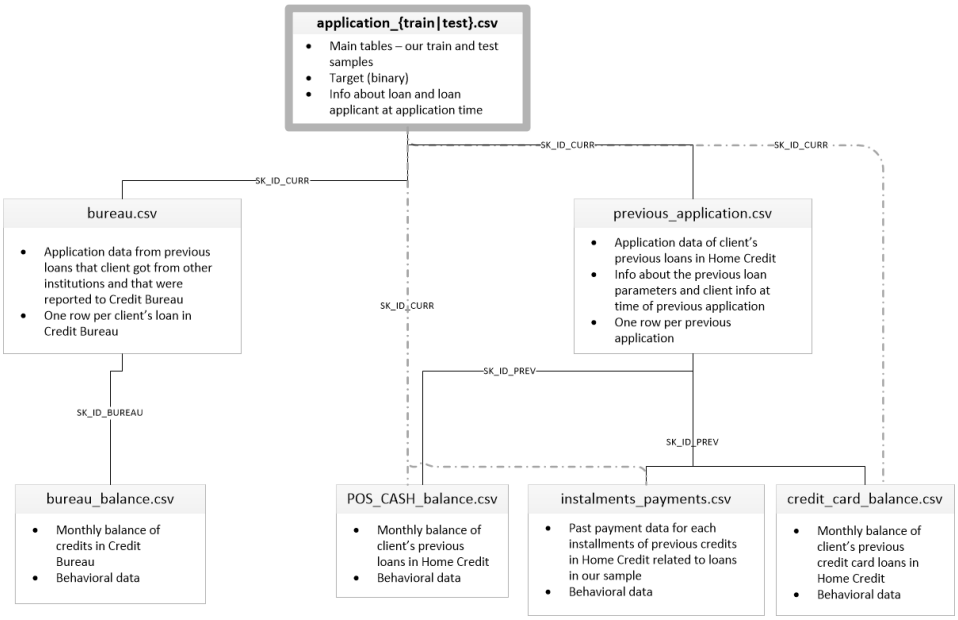

Figure 1: Home Credit Data.

5.1.3 Synthetic Dataset

A synthetically generated dataset has been used for stress-testing and scenario analysis. The data was generated using scikit-learn where each feature is a sample of a Gaussian distribution, in such a way that the 120 features obtained are correlated to the target. A sample size of 1000 was generated with 199 features. The data contained an imbalanced target (20% of non-default 80% of default) to replicated the case where individuals default as a result of crises, e.g., a scenario where there is a pandemic.

6 Metrics

6.1 Confusion Matrix

The confusion matrix describes the difference between the ground truth of the dataset and the model prediction. It remains the traditional way to evaluate classification machine learning problems. From the confusion matrix, a variety of metrics can be computed in order to compare classifier performances such as Precision, recall (sensitivity), specificity, accuracy, and the F1 Score. The precision is the fraction of the true positive examples that the model classified as positive while the recall is the fraction of the examples among the total number of positive examples that the model classified as positive. The true positive and false negatives are the number of positives and negatives classified by the model respectively. This is given by:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Sensitivity = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative}$$

And the specificity is defined as the fraction of actual negatives that are classified as Negatives,

$$Specificity = \frac{True\ \ Negative}{True\ \ Negatives + False\ \ Positive}.$$

6.2 F1 Score

The F1 score, also known as F score is the harmonic mean of the precision and recall of a model. A perfect model has an F1 score of 1. A way to combine the precision and recall of a model is given as:

$$F1\ \ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6.3 Accuracy

The accuracy of a model is the ratio between the number of samples that are correctly classified against the total number of samples (8).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Examples}$$

It is often not considered because it does not perform very well on an imbalanced dataset but it gives a good generalization of how a model will perform.

6.4 Receiver Operating Characteristic Area Under the Curve (ROC AUC/AUROC)

The Receiver Operating Characteristic (ROC) curve, a common classification metric is used to evaluate the loan default prediction of the models used. ROC curve is useful for the prediction of the probability of binary outcomes typical in our case for default and non-default. What the ROC curve does is plot the false positive rate against the true positive rate. The benefit of the ROC curve is that different models can be compared or different thresholds can be compared using the Area Under the Curve (AUC) measurements with a higher AUC implying a better model. Each line on the plot shows the curve for a single model while the movement along a line shows the change of threshold used for the classification of a positive instance. The threshold begins at 0 (upper right) and ends at 1 (lower left). The AUC is the area under the ROC curve which is between 0 and 1 with a good model scoring higher. If a model guesses at random, the outcome is a ROC AUC of 0.5. Thus, the ROC Curve is defined as a plot between Sensitivity and (1 - Specificity), which intuitively is a plot between True Positive Rate and False Positive Rate.

7 Model Evaluation and Validation

7.1 Training of MLP and CNN

The model is trained using the first three datasets–application_train, bureau, and bureau_balance, due to computational limitations. The deep multi-layer perceptron model is made of 8 layers with two activation functions; the rectified linear unit (ReLU) applied to the intermediate layers, and the Sigmoid function applied to the output layer. Also, half of the layers include a dropout layer each to prevent over-fitting. The number of epochs chosen for this model is 200 and an input dimension of 199 with a learning rate of 0.001. The training was done with the last portion (80%) of the dataset, and the rest of the data was used for testing. Validation is done to verify if the model performs well on previous data from different years. Since the dataset does not specify up to which year it was collected, it is assumed that the latter is the beginning from which the dataset was collected. This is similar to that of CNN except for the number of layers which is 2 (CNN trains longer).

Table 2: MLP model during training

MLP			
Training		Validation	
Training Loss	0.25	Val Loss	0.37
Training F1 Score	0.89	Val F1 Score	0.84
Training Accuracy	0.89	Val Accuracy	0.84

Table 3: CNN model during training

CNN			
Training		Validation	
Training Loss	0.60	Validation Loss	0.60
Training F1 Score	0.67	Validation F1 Score	0.67
Training Accuracy	0.67	Val Accuracy	0.67

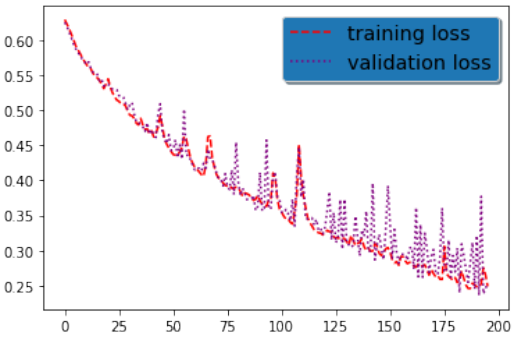

Figure 2: MLP loss

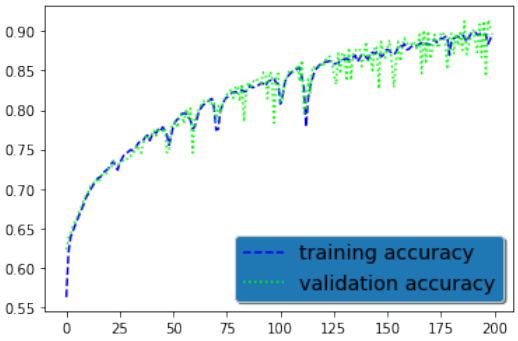

Figure 3: MLP accuracy

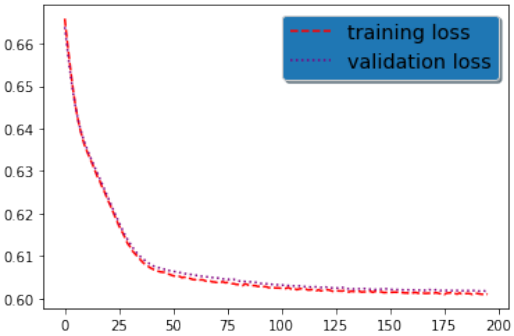

Figure 4: CNN loss

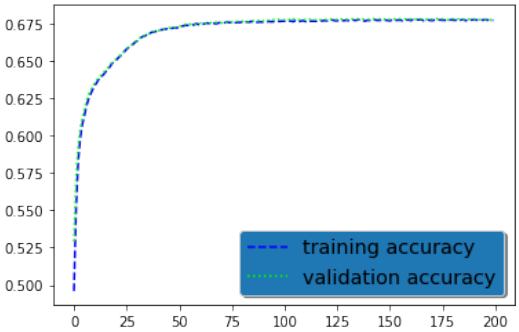

Figure 5: CNN accuracy

8 Stress-Testing of MLP and CNN

The stress-testing of the two models has been done using the synthetic dataset. We use ROC AUC, accuracy, recall, and precision as the metrics in the order to check the performance. We use the weights saved from MLP and CNN model trained on the Home Credit dataset. After testing we obtained the following results shown in Table 4 and Table 5.

Table 4: Results for MLP

MLP			
Back-testing		Stress-testing	
ROC AUC	0.963	ROC AUC	0.605
Accuracy	0.91	Accuracy	0.43
Recall	[0.8481 0.9811]	Recall	[0.7411 0.3450]
Precision	[0.9783 0.8655]	Precision	[0.2512 0.8180]

Table 5: Results for CNN

CNN			
Back-testing		Stress-testing	
ROC AUC	0.662	ROC AUC	0.664
Accuracy	0.67	Accuracy	0.69
Recall	[0.6843 0.6726]	Recall	[0.5483 0.7394]
Precision	[0.6771 0.6798]	Precision	[0.3841 0.8466]

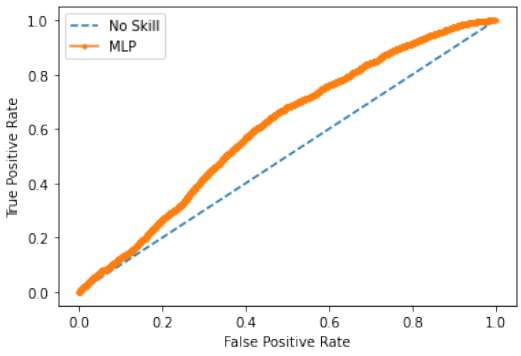

Figure 6: Stress-testing MLP vs No skill (AUCROC Curve)

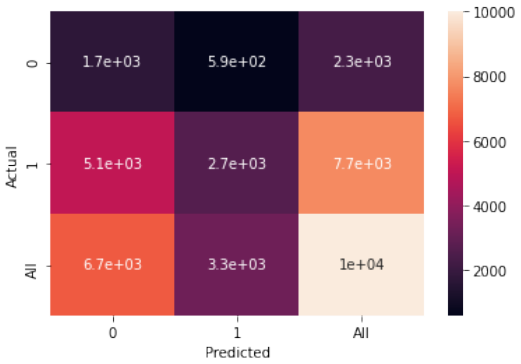

Figure 7: Confusion matrix for MLP

9 Back-testing of MLP and CNN

For the back-testing, we use the 20% portion of the Home credit dataset as the historical data. Since the dataset does not specify up to which year it was collected, it is assumed that the latter is the beginning from which the dataset was collected. It is observed that the MLP model performed well on the historical data with an accuracy of 91% against 89% on the training. This is the same for the CNN model with an accuracy of 66% on the historical data against 60% on the training. The results obtained can be seen in Tables 4 and 5.

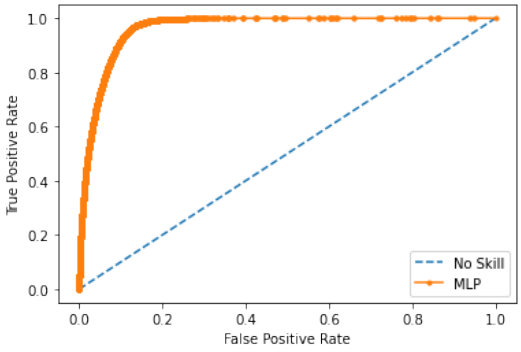

Figure 8: MLP vs No skill (AUCROC Curve)

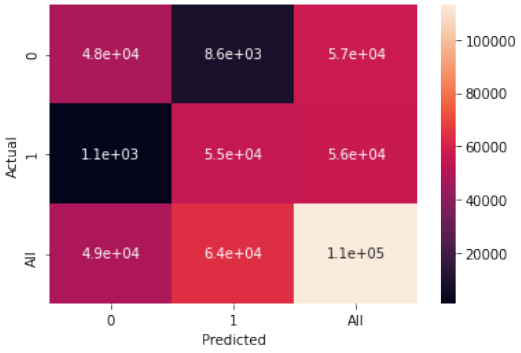

Figure 9: Confusion matrix for MLP

10 Discussion

During the back-testing of the MLP, it was realized that the model trained fast following an increase in the validation accuracy, hence training was successfully done. It was also realized that MLP is a good method for credit risk and does better with parameter tuning. As the plots show, the loss

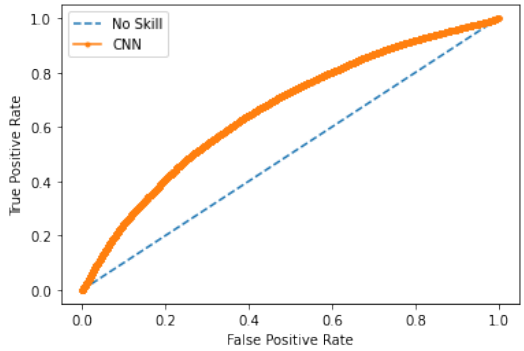

Figure 10: CNN vs No skill (AUCROC Curve)

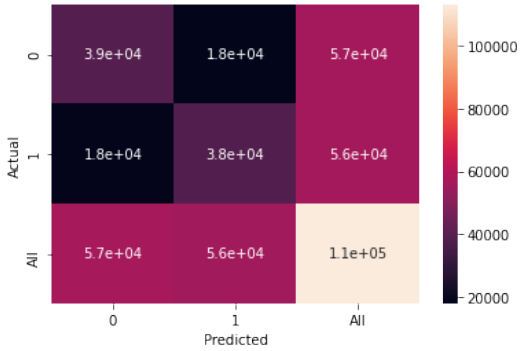

Figure 11: Confusion matrix for CNN

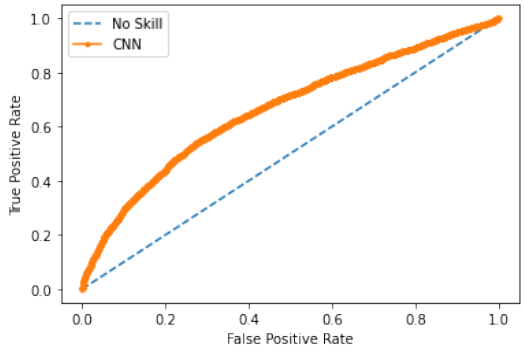

Figure 12: Stress-testing CNN vs No skill (AUCROC Curve)

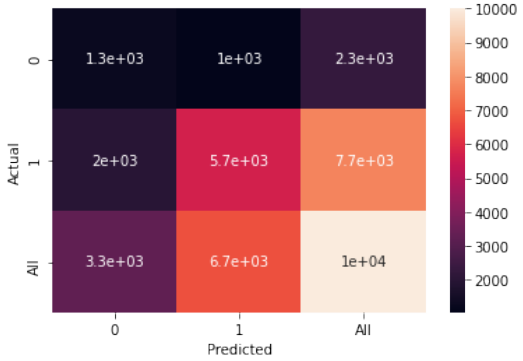

Figure 13: Stress-testing Confusion matrix for CNN

decreased at the set epoch, meaning with an increase in the number of epochs, the model will perform slightly well. The decreasing loss compared to the no-skill point is significant and shows a good performance for the MLP. For CNN, the model trains at a slower speed, and the plots also show improvements for the set number of epochs. At an accuracy of 67%, it is obvious that the method does well on credit data. Yet, there is still a percentage gap between CNN and the MLP. This can be associated with the fact that CNN performs better on images (31) whiles MLP also performs better on tabular datasets (28) as often represented in csv files. A typical MLP is more appropriate for classification problems where the inputs are assigned a label, and for regression with a set of inputs for the prediction of a real-valued quantity. Since datasets related to credit risk, in general, fall under the category of the tabular datasets, we can conclude that this has effects on the results that have been seen for both methods. From the stress-testing experiments conducted for MLP and CNN, we are convinced that the MLP model did not perform well on the synthetic data compared to the CNN model. The architecture choices we have made are all fairly reasonable, although they can still be improved.

11 Recommendations

- The lack of availability of credit risk datasets has been a big challenge in the domain of finance, more resources should be available to help researchers with their studies.
- Make available existing works for Back-testing and Stress -testing for credit risk. To the best of our knowledge, there is little scientific literature on back-testing or stress testing for credit risk. In particular, there is no back-testing and stress testing for credit risk based on learning models such as MLP and CNN.

- Existing codes used for different techniques of deep learning models should be available for improved scientific researches.
- Making clear what state-of-the-art techniques available for credit risks are to be able to keep track, for reproducibility of results and its evaluation or the choice of baselines.

12 Conclusion

Credit risk is the most common risk faced by financial institutions. Using machine learning techniques can lead to some financial and non-financial risks and the reason why model validation or model performance is indispensable for them before making decisions. In this paper, we have presented some learning techniques for credit risk; namely, MLP and CNN. We have also discussed works of literature available for deep learning methods in general for credit risks. We have also compared metrics such as F1 Score, accuracy score, and ROC AUC that have been used to test the performances of the models on the historical data. We performed back-testing and stress-testing for both methods (MLP and CNN) and realized that the MLP model performed well than the CNN for back-testing according to the Home credit risk dataset. By using simulated data and extreme scenarios for stress-testing, we found that the CNN model performs well for accuracy and the other performance metrics used. This can be explained by the fact that Convolutional Neural Network models learn the pattern on the data even in stressed situations.

References

[1] ADDO, P. M., GUEGAN, D., AND HASSANI, B. Credit risk analysis using machine and deep learning models. *Risks 6*, 2 (2018), 38.

[2] AIZENBERG, I., AIZENBERG, N. N., AND VANDEWALLE, J. Multi-valued and universal binary neurons: Theory, learning, and applications. *IEEE Transactions on Neural Networks 12*, 3 (2001), 647.

[3] ARNOTT, R., HARVEY, C. R., AND MARKOWITZ, H. A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science 1*, 1 (2019), 64–74.

[4] BACHAM, D., AND ZHAO, J. Machine learning: Challenges, lessons, and opportunities in credit risk modeling. *Moody's Analytics Risk Perspectives 9* (2017), 30–35.

[5] BAILEY, D. H., BORWEIN, J., LOPEZ DE PRADO, M., AND ZHU, Q. J. The probability of backtest overfitting. *Journal of Computational Finance, forthcoming* (2016).

[6] BHATORE, S., MOHAN, L., AND REDDY, Y. R. Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology* (2020), 1–28.

[7] CAPASSO, G., GIANFRATE, G., AND SPINELLI, M. Climate change and credit risk. *Journal of Cleaner Production* (2020), 121634.

[8] CHICCO, D., AND JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics 21*, 1 (2020), 6.

[9] DASTILE, X., CELIK, T., AND POTSANE, M. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* (2020), 106263.

[10] DECHTER, R. Learning while searching in constraint-satisfaction problems.

[11] DEVI, M. U., AND BATRA, N. Exploration of credit risk based on machine learning tools. *Journal of Critical Reviews 7*, 19 (2020), 4698–4718.

[12] DHEIR, I. M., METTLEQ, A. S. A., ELSHARIF, A. A., AND ABU-NASER, S. S. Classifying nuts types using convolutional neural network.

[13] EDDY, Y. L., AND BAKAR, E. M. N. E. A. Credit scoring models: Techniques and issues. *Journal of Advanced Research in Business and Management Studies 7*, 2 (2017), 29–41.

[14] FOR INTERNATIONAL SETTLEMENTS, B. Basel committee on banking supervision (bcbs). *International convergence of capital measurement and capital standards:: a revisedframework* (2006).

[15] GOH, R., AND LEE, L. S. Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research 2019* (2019).

[16] GOODFELLOW, I., BENGIO, Y., COURVILLE, A., AND BENGIO, Y. *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[17] HAND, D. J. Credit scoring. *Wiley StatsRef: Statistics Reference Online* (2014).

[18] HINTON, G. E., AND SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *science 313*, 5786 (2006), 504–507.

[19] HSIEH, N.-C., AND HUNG, L.-P. A data driven ensemble classifier for credit scoring analysis. *Expert systems with Applications 37*, 1 (2010), 534–545.

[20] INNOCENTI, F. Machine learning in credit scoring.

[21] JACOBS JR, M. The validation of machine-learning models for the stress testing of credit risk. *Journal of Risk Management in Financial Institutions 11*, 3 (2018), 218–243.

[22] KALCHBRENNER, N., GREFENSTETTE, E., AND BLUNSOM, P. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).

[23] KOH, H. C., TAN, W. C., AND GOH, C. P. A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information 1*, 1 (2006).

[24] LAWRENCE, S., GILES, C. L., TSOI, A. C., AND BACK, A. D. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks 8*, 1 (1997), 98–113.

[25] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation 1*, 4 (1989), 541–551.

[26] LUO, C., WU, D., AND WU, D. A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence 65* (2017), 465–470.

[27] MALIK, N., SINGH, P. V., AND KHAN, U. Can banks survive the next financial crisis? an adversarial deep learning model for bank stress testing. *An Adversarial Deep Learning Model for Bank Stress Testing (June 30, 2018)* (2018).

[28] MARAIS, J. A. *Deep learning for tabular data: an exploratory study*. PhD thesis, Stellenbosch: Stellenbosch University, 2019.

[29] MERCKEL, W. Developing a stress testing model for the credit risk exposure of a leasing company. Master's thesis, University of Twente, 2017.

[30] MIRCEA, G., PIRTEA, M., NEAMTU, M., AND BAZAVAN, S. Discriminant analysis in a credit scoring model. *Paper of Faculty of Economics and Business Administration West University of Timisoara, Romania* (2011).

[31] O'SHEA, K., AND NASH, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[32] SCHMIDHUBER, J. Deep learning. *Scholarpedia 10*, 11 (2015), 32832.

[33] YAMAGUCHI, K., SAKAMOTO, K., AKABANE, T., AND FUJIMOTO, Y. A neural network for speaker-independent isolated word recognition. In *First International Conference on Spoken Language Processing* (1990).

[34] ZHANG, X., YANG, Y., AND ZHOU, Z. A novel credit scoring model based on optimized random forest. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (2018), IEEE, pp. 60–65.