# Accuracies of Model Risks in Finance using Machine Learning

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

# Accuracies of Model Risks in Finance using Machine Learning

**Berthine Nyunga Mpinda**

AIMS - AMMI (Ghana)

GK- 0647-1372 Accra, Ghana

[*]bmpinda@aimsammi.org

**Jules Sadefo Kamdem**

MRE EA 7491 (Université de Montpellier),

Avenue Raymond Dugrand, 34970, Montpellier, France

Corresponding Author: jules.sadefo-kamdem@umontpellier.fr

**Salomey Osei**

AIMS-AMMI (Ghana)

GK- 0647-1372, Accra, Ghana

sosei@aimsammi.org

**Jeremiah Fadugba**

AIMS-AMMI (Ghana)

GK- 0647-1372 Accra, Ghana

jfadugba@aimsammi.org

## Abstract

There is increasing interest in using Artificial Intelligence (AI) and machine learning techniques to enhance risk management from credit risk to operational risk. Moreover, recent applications of machine learning models in risk management have proved efficient. That notwithstanding, while using machine learning techniques can have considerable benefits, they also can introduce risk of their own, when the models are wrong. Therefore, machine learning models must be tested and validated before they can be used. The aim of this work is to explore some

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

existing machine learning models for operational risk, by comparing their accuracies. Because a model should add value and reduce risk, particular attention is paid on how to evaluate it's performance, robustness and limitations. After using the existing machine learning and deep learning methods for operational risk, particularly on risk of fraud, we compared accuracies of these models based on the following metrics: accuracy, F1-Score, AUROC curve and precision. We equally used quantitative validation such as Back-testing and Stress-testing for performance analysis of the model on historical data, and the sensibility of the model for extreme but plausible scenarios like the Covid-19 period. Our results show that, Logistic regression out performs all deep learning models considered for fraud detection.

Keywords: Machine Learning, Model Risk, Credit Card Fraud, Decisions Support, Stress-Testing.

## 1 Introduction

The financial system plays a key role in the economy by stimulating economic growth. It includes three main components such as financial markets, financial institutions, and financial regulators, where each plays a particular role in the economy. Financial markets give the possibility for the flow of funds in order to invest in companies, governments as well as individuals. In the financial market place, financial institutions plays a crucial role, which is to serve as intermediates and determiners of the flow of funds, while financial regulators are responsibly monitoring and regulating the participants in the financial system [13].

Furthermore, financial instruments or securities (stocks, bonds, etc) are traded everyday, and it falls to the regulators to oversee these transactions between buyers and sellers, as well as determine the prices of the assets traded. This provides liquidity and opportunity for investors to sell a financial instruments at a fair value at any time in the market [13]. These transactions in the marketplace expose financial institutions to different types of risk and consequently, in need of various techniques for risk management and hedging.

Over the last 30 years financial risks has increased deeply and has led to the exponential growth of instruments designed to manage them [33]. Defined as a volatility of unexpected outcomes, risk can represent the value of assets, equity or earning [33]. Sometimes, risk may be seen as an uncertainty that can affect institutions. Financial institutions or firms are exposed to diverse types of risks which can be business or non-business risk. However, as part of non-business risk, financial risks are related to losses due to financial market activities such as interest rate, foreign exchange rate, insolvency and/or credit default. Knowing the impact of financial risks on an institution, the primary function of the institution becomes the careful management of such risks. This means they must measure financial risks [33] in order to monitor and manage them. In addition, financial institutions must plan for the consequences of unfavorable outcomes and be prepared for unavoidable uncertainty.

Following the global financial crisis between 2007 and early 2009, financial institutions have been facing different risks everyday [27]. These risks come from many sources such as financial uncertainty, legal liabilities, management errors, accidents and natural disasters, IT security, data quality or models misused. Nowadays, risk management is a high-profile and growing disciple for banks. The biggest concern of a risk manager is the capacity to figure out, quantify, report and reduce risks in order to limit possible losses. Efficient risk management is the key to a bank's performance. Given the role played by banks in the financial system as well, risks are subjects to regulatory attention [20]. There are four major types of financial risk faced by banks. These are; Credit risks, Market risks, Liquidity risks and Operational risks. According to the Basel Committee on Banking Supervision (BCBS), Operational risk is defined as a loss caused by an inappropriate or failed internal process. It can equally be caused by people, systems or from external events. Usually, operational risks are depending on human factors such as mistakes due to some actions. Operational risk can lead to credit risk or market risk [33] and they include fraud risk, cybersecurity risk, legal risk, technology risk, model risk, information and resilience risk, etc. Banks are dynamically engaged in risk management for monitoring, managing, and measuring them. Also, they are required by regulators to hold appropriate capital against these risks and that they are suitably indemnified for risks incurred. Among all the risks, credit risk requires the most capital [20].

From a quantitative point of view, model risk is an important notion that is universal in any quantification of risk and is a serious problem for financial institutions. In finance, models are widely used, any deficiencies in such models lead firms or financial institutions to price incorrectly assets, mishedge risks or enter into an disadvantage trades. The utilization of models invariably presents model risk, which is a possible loss an institution may undertake, as a consequence of decisions that could be based on the output of a model, due to errors in the development, implementation or misuse of models[2]. Because models are simplifications of reality, and real-world relationships among events, model risk can lead to serious financial losses and poor business decisions. For example, in November 2004, China Aviation Oil lost USD 550 million because of using an insufficient modeling

[2] www.deloitte.com

methods for derivatives [18]. In 2012, J.P. Morgan–The London Whale lost about USD 6 billion and was penalized with USD 1 billion due to the error of change of the Value at Risk (VaR) metric [3].

Traditional statistical models have been used for risk models such as linear regression techniques and Elastic Net for credit risk. Moving average model, Autoregressive Integrated Moving Average (ARIMA) models, Exponent Smoothing (ES) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models have been used for stock market predictions [3]. However, owing to incompleteness in traditional statistics methods, there is now an increased interest of using Artificial Intelligence (AI) and machine learning techniques to enhance risk management [24]. For example, there are some difficulties encountered when computing VaR on financial portfolios as explained by Acerbi [3]. Additionally, VaR is not able to estimate correctly the probabilities of future market events [21]. Without forgetting the short-term prediction of the ARIMA model in the stock market compared to the Artificial Neural Networks (ANNs) models [4]. Because of the advanced technology associated with big data, the availability of data and also the computing power, banks are innovating their business models with emerging new technologies [28]. The advance of data and machine learning has open on to a Fintech industry, wrapping digital innovation and technology-based business model innovations in the financial research development.

Recent applications of machine learning techniques in various fields show its efficiency. Also, from Lynn et al. [24] researchers have shown that machine learning can be applied in financial particularly in risk management. Classical statistics models explicitly trace uncertainty by considering and specifying the probabilitic model for the data, while machine learning models refrain from modelling the data generation process and rather learn from the data. Statistical models begin by assuming an additive assumption, which is a dominant way that predictors affect the outcome, and require time-consuming, data reduction steps when the number of candidate predictors is very large. Additionally, they require interactions of terms to be pre-specified, a lower sample size and they depend on analyst selection of input features contrary to machine learning models.

Drawing the line between machine learning models and classical statistics models, machine learning include models that highlight non-linearity, interactions and data-driven structure [12]. Also, machine learning techniques emphasize feature selection algorithms, discovery of terms interactions and creation of features from raw data. In December 2018, the Global Association of Risk Professionals (GARP) and analytic leader SAS conducted a survey (GARP/SAS suvey) about how AI technologies are used by financial institutions. Four out of five responded positively by saying that AI technology has been used in their institutions for forecasting (54%), optimization (51%) followed by machine learning (34%), robotic process automation (29%), Natural Language Processing (NLP) and Computer

[3] www.deloitte.com

Vision (23 % each), and virtual agents (22%). Over the next three years, they are planing to use machine learning and NLP up to 80% and 66% respectively [1].

The use of machine learning techniques to model risk by financial institutions can lead to both financial and non-financial risks. They are heavily regulated through the lifecycle model by regulatory guidance. Financial risk leads companies to loose money and it can be for credit, market, liquidity, capital management, investment, etc. And the non-financial risks can lead to compliance, legal and reputation. Machine learning techniques have considerable benefits and also can lead to many risks when the models used are wrong. According to George Box, all the models are wrong, and when they are wrong, they create harm and damages to the institution or to the customer. For example, when a hedging model is mis-hedging, it can create market risk. Also, for financial crimes or fraud detection, when the models are wrong, they can create compliance risk. A model should not only add values to the company, but it must also reduce risk even in long term period. That is why model risk management focuses on model validation, performance and improvement. Particular attention in this work is given to the notions around how to evaluate the performance, robustness and limitations of machine learning models to control risks, on how the model will fail and when it fails, what will be the damage generated. To do so, we used the quantitative validation test such as Back-testing for performance analysis of the model on historical data and Stress-testing for the sensibility of the model under a variety of conditions such as a Covid-19 crisis period.

In this work, after using the existing machine learning and deep learning methods for Credit Card Fraud detection, we tested the performance of models using historical data and the model failure using simulated data. Since it is important to have acceptance metric to compare the accuracies of the models, the following metrics have been used accuracy, F1-Score, precision and AUROC curve. Accuracies of the decision tree model, Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN) models for credit card fraud detection have been compared to the logistic regression accuracy, considered as the benchmark model. As we know that the Covid-19 crisis has left a deep mark on stock markets, the model robustness, adverse test to identify model risk have been done using simulated data.

## 2 Literature Review

The universal definition provided by regulators states that, a model is a quantitative method or approach that applies techniques and assumptions to process input data into quantitative estimates for decision making [4]. The techniques and assumptions can be statistical, financial, mathematical, economic or machine learning algorithm. The main model components are the inputs, which are

[4] www.deloitte.com

data, assumptions and scenarios, and the outputs, which can be forecasts, estimates or management decision support. Model risk essentially occurs for two main reasons: the first being that a model may have been built as it was intended, but could have fundamental errors and produce inaccurate outputs when compared to its design objective and intended use. Also, it may be used incorrectly or inappropriately, its limitations or assumptions may not be fully understood. Banks are heavily dependent on models to help them to make the best decision when it is about to quantify financial risk, to price financial instruments or derivatives. That is why the quality of the model matters. Model risk management should add value and reduce risk as well to drive better decision-making and business results.

Traditionally, classical statistics models such as risk rating models, Credit Valuation Adjustment (CVA) have been used for credit risk. But their limitations is that they should be used when they are accurate, not when they underestimate or overestimate risks. Most used methods by financial institutions for credit risk are the probability of default (PD), loss given default (LGD) and exposure at default (EAD) to compute the expected loss exposed to when a borrower backs out on a loan [28]. The Expected Loss (EL) is given by $\text{EL} = \text{P D} \times \text{LGD} \times \text{EAD}$. Also, classical statistics models such as linear regression and its extensions such as Lasso representation and Elastic Net penalty, logistic regression and multi-modal regression models [28] [24] have been used for credit risk. Value at Risk (VaR), Expected Shortfall (ES), which compute the maximum loss over a target horizon and the expected loss respectively, and Asset-Liability risk measures have been used to evaluate market or liquidity risk [3]. According to the Basel Committee on banking supervision, two others traditional methods have been introduced as principles and guideline for measuring, and managing bank liquidity risk: Liquidity Coverage Ratio (LCR) and Net Stable Funding Ratio (NSFR) [32] [36]. The problem of these methods is that they could not do the correct estimation of keys parameters involved in the calculation of these ratios [32], it is not implemented in practice in a banking system. Also, they are not able to identify and analyze the incidence of stress scenarios.

Another approach based on the inflow-outflow concept and an option-pricing approach have been introduced by [36] to evaluating banks' liquidity. Multiple regression models, probabilistic models, balance sheet ratio and calculation of the funding gap are also other measurements for Liquidity risk [15] [32][36]. Thus, the models such as Loss Distribution Approach (LDA) and integration models have been used to measure operational risk.

Machine learning models have attracted an important interest for different applications to financial institutions and its applications have received much attention from investors and remove to researchers. The main motivation of using machine learning models in financial sector is their ability to learn and improve productivity. Also, they are able to use high-dimensional data to make results more accurate and help to reduce risk and cost [20]. Machine learning is largely seen in financial markets as having the potential to implement the analytical capability that financial institutions wish. Machine learning techniques, in particular deep learning models have the ability to capture non-linear relationships in the data and also the ability to learn and adapt [20]. They have proven exceptionally successful results in different applications. There is a strong believe that machine learning will be a vital element of future strategies for operational and financial risk management [24]. Additionally, Lynn et al[24] affirms that machine learning is able to impact every aspect of financial institutions, their business model such us risk management, hedging, fraud detection and cybersecurity risk. According to the GARP/SAS survey, many institutions are planning to use machine learning up to 80% over the next three years [1].

Machine learning models are more accurate than traditional models and have a remarkable role in the credit risk modeling [28]. The most commonly used machine learning models in credit risk include Logistic Regression, k-Nearest Neighbor (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, Linear Discriminant Analysis (LDA), Boosting, Extreme Gradient Boost (XGBoost), Deep Multi-Layer Perceptron (DMLP), Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) [28] [20]. Since credit risk is about building binary classification models which means predict loan default probability, Martey et al. [28] focused on credit risk scoring where they examined the impact of Random Forest, Gradient Boosting and deep learning models for an enterprise to default. By using different approaches such as random forest model, gradient boosting and four deep learning models they mentioned that the gradient boosting model was able to provide stable results in deciding whether to provide loan or not.

Some research has been done using machine learning techniques for liquidity risk and other types of risk. From the literature review by Leo [20] for machine learning algorithms in banking risk management, a number of liquidity risk problems can be solved by using machine learning. Tavana [32] used Artificial Neural Networks (ANNs) and the Bayesian Network (BN) models for liquidity risk. The goal was to measure and predict liquidity risk, and also analyze the importance and occurrence of the risk indicators. Hu et al. [16] provided an overview of supervised machine learning techniques such as Random forest, Boosting and Neural Networks in credit risk.

From the literature review by Leo [20], models such as Gaussian Mixture Model, cluster methods and ANNs for volatility estimation have been used with the ability to extract the essential features from the past data in order to be used for predicting future values. Alkhatib [6] used the k-Nearest Neighbors (KNN) algorithm to predict stock price from the Jordanian stock exchange. Adebiyi et al. [8] developed the ARIMA model for stock price prediction using New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE) data. They proved the short-term prediction of ARIMA model using stock market. A systematic review has been done by Sezer et al.[30] showing the different deep learning models used for financial time series prediction. From this review, they proved the most used methods include: Deep Multi-layer Perceptron (DMLP), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), Auto-encoders (AEs), Deep Reinforcement Learning (DRL) for stock price, index, commodity price, volatility, forex exchange and trend. According to Sezer analysis, RNN with its variants LSTM and GRU rest the most preferred by researchers [30]. Shashi et al. [31] compared the performance of LSTM to GRU for market forecasting using Nepal stock. They proved that the improvement can be done by incorporating the financial news sentiments with the stock features as the input.

As part of operational risk management, fraud detection has been a challenge and a big concern for many financial institutions; billions of dollars are lost yearly because of this undetected fraud. According to Dixon [14] fraud detection is one of the earliest successful application of machine learning. It is important for financial institutions to be able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. The credit card fraud detection can be considered as pivotal asset for guaranteeing customer trust and saving money by fending off fraudulent loss [12]. In 2012 Zareapoor et al. [35] presented a survey of nine machine learning techniques used in credit card fraud detection such as Artificial Neural Networks (ANNs), fuzzy system, decision trees, Support Vector Machines (SVM), Artificial Immune System (AIS), genetic algorithms, K-Nearest Neighbor (KNN) algorithms, Hidden Markov Model (HMM) and Genetic Algorithm (GA). From the comparison analysis based on some criteria such as accuracy, speed and cost, they concluded that all the models were performing well. The review of Albashrawi [5] on financial fraud detection from 2004 to 2015 shown the logistic regression model as the most useful in detecting financial fraud with 13% followed by neural network and decision tree. Lucas [22] proposed a multi-perspective Hidden Markov models (HMM) based on automated features engineer-all using the transactions issued by Belgian credit cards between March and May 2015 in order to incorporate a broad spectrum of sequential information in the transactions feature sets. Breeden review [12], highlighted different approaches that have been utilized in order to apprehend the sequential properties of credit card transactions, and to use sequential modeling methods such as recurrent neural networks (LSTM) and Hidden Markov Models (HMM).

## 3 Machine Learning techniques

Machine learning is a core technique of Artificial Intelligence (AI) with the ability to learn without being explicitly programmed. As predictive models, they can learn from training datasets to make predictions. These predictive models can be built for regression or classification problems. Depending on the type of data used to build a model, several classes of algorithms exist: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Table 1: Summary of existing and present studies about different types of risk using machine learning techniques

| Authors | Year | Type of Risk | Methods used | Datasets | Outcome |
|---|---|---|---|---|---|
| Tavana et al [32] | 2018 | Liquidity Risk | ANNs and BN | Standard bank balance sheet | Distinguish most critical risk factors. |
| Wei XIONG [34] | 2018 | Market Risk | SVM | Simulated dataset | Develops a Machine Learning model to classify market risk VaR exceptions into "market move" and "VaR model issue". |
| Yan Liu [21] | 2005 | Market Risk | ANNs combination model | S&P 500 and Ford | VaR forecast combinations using Artificial Neural Networks (ANNs). |
| Joseph L. and Breeden [12] | 2020 | Credit Risk | Statistical and Machine Learning models | – | Highlighted the crucial ML methods used and developed in credit risk. |
| Leo Martin et al. [20] | 2019 | Credit risk, Market Risk, Liquidity and Risk Operational | Statistical and Machine Learning models | – | Shown the applications of ML in the management of banking risks. |
| Yves Lucas [23] | 2020 | Fraud Detection | Hidden Markov models with integration of contextual knowledge | Belgian Credit Cards | Proposed how to generate history-based features using hidden Markov models. |
| Mousa Albashrawi [5] | 2016 | Fraud Detection | Logistic Regression (LR), Decision Tree (DT), SVM, ANN and BN | – | Review of different data mining methods to detect financial fraud. |
| Jacobs and Michael [17] | 2018 | Credit risk | Stress-testing other Multivariate Adaptive Regression Splines model | | Examined a critical input into stress-testing process. |
| – | Present | Fraud detection | LR, DT, MLP, CNNS, Backtesting and Stress-testing | Credit Card Fraud detection and simulated data | Compare accuracies and performance. |

Given labeled data $(x_i, y_i)$ such that $x_i \in X$ and $y_i \in Y$ for $i = 1, \dots, n, n \in \mathbb{N}$, the objective of supervised learning model is to learn the correlation between the features $X$ and the target variable $Y$. The models in unsupervised learning are built with data points which are unlabeled, i.e $(x_i)$. Semi-supervised learning lies between supervised and unsupervised learning. Reinforcement learning consists in taking action based on each data point and then computing how good the decision is. In this work, we use supervised machine learning techniques such as logistic regression, decision trees and artificial neural networks to build more accurate and robust risk models by picking out complex, nonlinear patterns into large datasets for decision making.

### 3.1 Logistic Regression

The logistic model also called the logit model is used for modeling a binary dependent variable $y \in \{0, 1\}$ or $\{-1, 1\}$. Given the training data points $D = \{(x_i, y_i)\}_{i=1}^{n}$, a binary classification models use an S-shaped curve called sigmoid function [5] represented as: $\frac{1}{1+e^{-x}}$ for predicting binomial outcomes of the dependent variable. The hypothesis for the logistic regression is given by

$$h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}}$$

where $\theta^T$ is the transpose's vector of $\theta$ and $X$ the input matrix. The results of this model fall between the range of $[0, 1]$.

### 3.2 Decision Trees

As part of the most popular approaches for classifiers [29] and a specific type of probability tree, it represents a recursive partition of instance space. It is a directed tree with a root that has no incoming edges while the rest of the roots have a single incoming edge. A root which has outgoing edge is a test node while the remaining are called decision nodes. The internal nodes splits the instance space into sub-spaces using a discrete function of the input attribute values. Nodes are grown by adding incrementally based on the choices made and according to the problem. Decision trees classify data by posing series of questions about the associated features and an excellent way to deal with complex decisions. The questions form a hierarchy since each question is contained in a node and an internal node points to a single child node for a possible answer. Common measures for evaluation of decision trees are the entropy and the Gini index [19]. A decision tree is easy to interpret and understand but when used for categorical variables that have multiple levels, they accumulate more information gain and calculations become complex [6].

[5] https://www.kdnuggets.com/2018/02/logistic-regression-concise-technical-overview.html

[6] https://www.statisticshowto.com/decision-tree-definition-and-examples

# 4 Artificial Neural Networks

Artificial Neural Networks (ANN) are stimulated by biological neural in the sense that the artificial neuron is similar to the biological neuron both in structure and function [16] [37]. The artificial neuron obtains several inputs, $x_i$ from many other neurons and process them to get an output $y_k$ generally a 'real value'. But more specifically, in classification problems, it sends a binary output, 0 or 1. The function $f(z_k)$ that computes the outcome of the input vector (n-dimensional) $x_j, j = 1, \dots\ n$ is a linear combination of the product of inputs $x_j$ and the weights $(w_{jk})$ plus the bias term $b_k$. The index $k$ refers to the particular neuron that we are dealing with, while the $j$ refers to the input that the weight refers to. Mathematically, $z_k$ is given by

$$z_k = \sum_{j=1}^{n} w_{jk} x_j + b_k$$

Then, the output $y_k$ is obtained by

$$y_k = f(z_k)$$

where $f(.)$ is a non-linear function called activation function.

## 4.1 Multi-layer Perceptron (MLP)

A multi-layer perceptron (MLP) is a class of feed-forward Artificial Neural Networks (ANNs) built of layers that are organized together in such a way that every layer is composed of nodes that are connected to the nodes of the following layer. If the MLP is fully connected, each node in the previous layer is connected to each node in the next layer. Constituted of at least 3 different layers: input, hidden and output layers, its inputs are weighted and summed before to be distributed through the network and produce an output. MLP optimize its weights through back-propagation process and learning technique such as gradient descent in order to choose the weights that minimize the the cost function. However, non-linear activation function are used between hidden layer and output to scale the sum to a particular acceptable interval of real values (e.g [0, 1]) that can be an input for the next layer or the output at the last layer. In a binary classification problem, Sigmoid activation function is used at the last layer.

## 4.2 Convolutional Neural Networks (CNN)

A convolutional Neural Networks (CNN) are out of the successful artificial neural networks used for images particularly for image classification [16]. What makes CNN useful for image analysis is the pattern detection. CNN organizes neurons in a 3-dimensional manner, width, height, and depth.

CNN consists of an input layer, an output layer, and multiple hidden layers. The hidden layers usually contain one or more convolutional layers, pooling layers, and fully connected layers. An activation function Rectified Linear Unit (ReLU) is used between layer in order to increase non-linearity in the network. The convolutional layer performs dot products between filters and a receptive field of the neuron while the pooling layer consist to reduce the number of parameters and makes the features robust against noise and distortion [25]. Max-Pooling and Average-Pooling are the most common pooling used. The output layer acts as a classifier and computes the class scores using a Sigmoid or a Softmax function, depending on the number of classes.

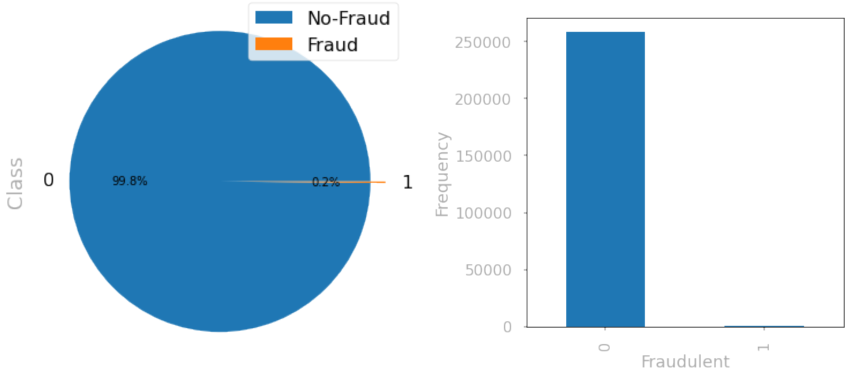
## 5 Data

### 5.1 Credit Card Fraud Dataset

There is no publicly available dataset that would sufficiently reflect the magnitude and variety of real-world card payments and that could therefore be considered as a basis for studying the many interesting challenges present in this domain. Thus, the payment data contains very sensitive exclusive information about individuals and institutions that make the access to such data highly limited to only the data holders and the companies manager.

Throughout this work, we explore the transactions dataset provided by Kaggle[2] containing transactions made by credit cards in September 2013 by European cardholders. It has been collected and analysed during a research collaboration of Worldline and the Université Libre de Bruxelles (ULB) Machine Learning Group on big data mining and fraud detection. The dataset presents transactions that occurred in two days made up of 284807 rows with 31 columns where there are 492 frauds out of 284,807 transactions as shown in Figure 1. The data is highly imbalanced and it contains only numerical input variables which are the result of a Principal Component Analysis (PCA) transformation. Due to confidentiality issues, the original features and more details about the data are not provided. Features V1, V2, . . . ,V28 are the PCA-transformed features, only two features 'Time' and 'Amount' which express the seconds passed between each transaction and the 'Amount' which represents the transaction amount respectively have not been transformed with PCA. The fraud label is binary and it indicates the authenticity of a transaction. The label was assigned by human investigators after consultation with the card holder. The label takes the value 0 when the fraud is not identified and 1 otherwise. Throughout this work, we will refer to the label as the class of the transaction. From the number of fraud transactions, 113 were for $1, 27 were for $99.99, 17 frauds for $0.76, 10 for $0.77 and 27 for $0.0. The highest fraud transaction amount was $2125.87 and lowest was $0.0.

Figure 1: Fraudulent and non-fraudulent transactions



### 5.2 Synthetic Data

Synthetic data plays an important role in deep learning and works well in different tasks and domains from computer vision with the generation of image data to finance. For more details see [26]. It is used for different purposes, such as data augmentation, model validation and model testing when there is a lack of real dataset. Assefa [10] highlighted the motivations beyond synthetic data generation in finance among others lack of historical data, data sharing restrictions and training machine learning models. Synthetic data has been also used for stress-testing and scenarios analysis. In machine learning, it plays an important role in preventing over-fitting, handle imbalanced data and to accommodate plausible scenarios. Therefore, to evaluate the performance of the built models in this work, we used the synthetic data generated following the Gaussian distribution, with mean zero and standard deviation one. Each of the 29 features is a sample of a Gaussian distribution and 20 of them contribute to the target value. We generated a sample dataset of 29 features, where 20 of them are a linear combination of the the target. Since, we want to test the performance during crisis, we assumed that during crisis 20% of transactions are fraudulent and 80% non-fraudulent.

## 6 Performance Measures

### 6.1 Confusion Matrix

Considered as an evaluation metric, the confusion matrix is an $n \times n$ matrix employed for evaluating the performance of classification algorithms in machine learning, where $n$ represent the total number of target classes. For example for a binary classification problem, the confusion matrix is a $2 \times 2$ matrix with four values such as the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in Figure 2. The false positive or type I error is the number of negative values predicted as positive by the model while false negative or type II is the number of positive values predicted as negatives by the model. It expresses how good the model is performing and also the kinds of errors the model is making. Thus, a perfect model would be the one that has all the positive examples predicted positive and all the negative example predicted negative.

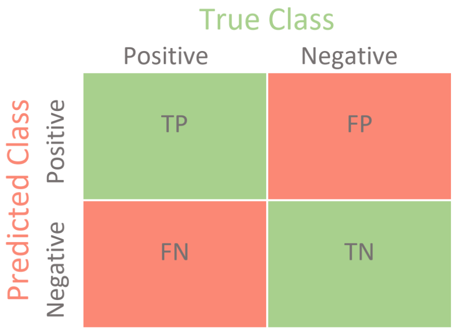


Figure 2: Confusion Matrices Illustration

### 6.2 Accuracy

The accuracy expresses the model's performance over all classes. In case of imbalanced classes, accuracy is not a good choice as metric. It is calculated by

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

### 6.3 Precision (Specificity)

The precision is defined as the proportion between true positives to the sum of true and false positives. It reflects the ability of the model to classify positive samples as positive and not misclassify negative samples. It determines how precise is the prediction on the positives. The higher the value is, the more accurate the model. It is calculated by

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

### 6.4 Recall (Sensitivity)

The recall metric is the capacity for a classifier to find all positive instances. For each class, it is described as the proportion of true positives to the sum of true positives and false negatives. A high value of the recall indicates that the model can categorize correctly all the positive samples.

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

### 6.5 F1-Score

The F1-score is the average of precision and recall. It is calculated by

$$\frac{2 \times Precision \times Recall}{Precision + Recall}$$

### 6.6 Area Under the Receiver Operating Characteristic (AUROC) curve

The AUROC curve is a metrics used in order to quantify the performance of a binary classifier. It tells how likely the model is capable of distinguishing between two classes. It interprets how good a model can differentiate positives and negatives. The larger the area, the good a model is at forecasting fraudulent and non-fraudulent transactions. The AUROC curve plots True Positive Rate (TPR) against the False Positive Rate (FPR) as shown below for every threshold $\alpha$.

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

## 7 Model Validation

Model validation is defined as kind of procedures to verify the performance of a model in line with institution objectives. All models that involve risks for a financial institution should be validated before to be used in decision making. Because of risks faced by banks for example, it is required to validate their models before to be used in practice. For machine learning models, investment managers should know what to expect with any machine learning techniques [9]. The validation includes benchmark test, stability test, convergence tests, sensitivity test , Back-Testing and Stress-Testing. The purpose of this work is about the two last tests.

### 7.1 Back-Testing

Back-testing is a very important model monitoring activity. It is a way to evaluate models on historical data and it is the only way to have confidence in a new model before to be deployed to real customers. Considered as a way to test the improvement of a new model on past experiences, Back-testing is important for the life-cycle of a model. It is expecting from a machine learning model to produce good results during back-testing. Arnott [9] developed a research protocol for investment strategy Back-testing using machine learning that include the respect of the temporal order of observations when splitting the data and the integrity of data. The methods have been used to back-test a machine learning models such as Train-Test split, Multiple Train-Test splits and Walk-Forward Validation, where a model can be updated every time when there is new data received. The Back-testing strategy for machine learning models consists on how to split the test data in a way that it adapts to the performance.

### 7.2 Stress-Testing

Stress-testing have become a central tool used by global regulators to manage financial stability of models. Since the financial crisis [11], regulators have been using Stress-testing as a means to evaluate the soundness of financial institutions'risk management procedures [17]. Stress-testing is defined as a type of performance test that checks the upper limits, by testing it under utmost loads. It is used for all types of risk and it helps to verify the stability and reliability of a model. It determines the model's robustness and error handling under extreme load conditions. Then, by stress-testing machine learning models before being used help to know the robustness of the model, to fix the bottlenecks and be prepared for any unexpected issues. One way to do the stress-testing is to proceed by splitting a given data into crisis and non-crisis period. And then train to analyze if the model risk changes in periods with greater variability in the return series. The second way is by testing a machine learning model on different data such synthetic or simulated data based on some assumptions. As a non-statistical risk measure [7], stress-test helps to determine the impact of extreme scenarios of a model. Also, the model's behavior in stressed extreme conditions [11]. In this work, to do the stress-testing of our models, using simulated data where we made some assumptions related to the crisis period such as Covid-19 period.

## 8 Experiments

### 8.1 Preprocessing

The data considered in this work is a class imbalanced problem with unequal number of data points across each class. Therefore, learning from an imbalanced dataset has been a challenge question for researcher that needs to be considered. Often in practice it is not common to have balanced data, the samples are often skewed to one class. Most of the existing classification machine learning algorithms assume balanced classed. Knowing the consequence of training on imbalanced data which is the probability to predict everything as the majority class, the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm has been used. The Up-sampling approach consists to duplicate the minority class with synthetic data.

Data preprocessing is a necessary step before building a model and it should be transformed in a way that it can be feed to the machine learning model. From the description of our data, all the features went through a PCA transformation except two features, "Time" and "Amount". To set up a good machine learning algorithms and because of the number of outliers detected in Amount column, we scale these features by standardization, which is a technique where the values are centered around the mean zero and with a unit standard deviation. That means, the new columns have been obtained by using the formula below:

$$z = \frac{x_i - \mu}{\sigma}$$

where $x_i, i = 1, \ldots, n$ is a data point, $\mu$ represents the mean, $\sigma$ the standard deviation of the data. We split the dataset in the training set and the validation set which is considered as the historical data. We split the data based on the time period (given in second) corresponding to the training and to the testing set. Since the data is a collection of transactions during two days, we decided to train our fraud detection model on the last 80% of the time period and test on the first 20% of the time period. We applied the up-sampling techniques on the training set only prevent data leakage and prevent over-fitting. After the up-sampling the ratio of fraudulent transactions changed for 0.14% to 50.0%. We scaled using the standardisation technique, also the feature time has been dropped since there was no correlation with the target and we ended up with a data set of 30 features including the target.

### 8.2 Models Training

We split our dataset into 80% for training and 20% for testing. We split the data before applying the up-sampling technique in order to avoid data leakage. During the training of Logistic regression and Decision trees, we used the search grid technique from Sci-kit learn in order to find the optimal parameters.

During the training process, the models learn some features of the data that are used for the predictions. The MLP architecture is made up with 3 hidden layers with 60 units each and an output layer. The activation function used between layers is the ReLU function, except the output layer where we used the Sigmoid function, as we are predicting two output values (fraud and non-fraud transactions). To avoid over-fitting, we used the dropout of 0.1 as regularizer and batch normalization. In addition, we used Stochastic Gradient Descent (SGD) as learning algorithm to optimize parameters of our model and binary cross-entropy as loss function. With an input size of 29, the model trained for 50 epochs with a batch size of 128 using a random seed of 10. The CNN architecture is made up of three convolutional layers each having a $2 \times 2$ kernel size, one fully connected layer with 512 neurons and two neurons in the output layer corresponding to the two classes (fraudulent and non-fraudulent transaction). In addition, batch normalization and max-pooling of $2 \times 2$ has been used at each layer. The ReLU function has been used as activation function in convolutional and hidden layers in order to introduce non-linearity in the network, and a Sigmoid function in the output layer. To avoid over-fitting, a dropout of 0.1, 0.2, 0.3, 0.4 and 0.5 has been applied as regularizer at each layer. The output of last convolution has been flatten and feed into a fully connected layer followed by a Sigmoid function to get the output class. The model trained for 50 epochs with a batch size of 64 and random seed of 10. The optimizer and the loss function are the same used for MLP model.

# 9 Results

## 9.1 Back-Testing

The data was trained using the Logistic Regression as benchmark model, Decision Trees, Multi-layer Perceptron (MLP) and Convolutional Neural Networks (CNN). For all the models, 80% of the data was used for training and the remaining 20% used validation since we wanted to see the performance on historical unseen data. As our dataset was highly imbalanced, we compute different metrics but we instead focused our comparison on the accuracy, F1-Score, precision score and AUROC curve. The performance of each model on validation data can be found in Table 2 and in Figure 3 of confusion matrices for each model.

Table 2: Performance metrics of the different models on Back-Testing

| **Metrics** | **Models** | | | |
|---|---|---|---|---|
| | **Logistic Regression** | **Decision Trees** | **Multi-Layer Perceptron** | **Convolutional Neural Networks** |
| Accuracy | 0.948719 | 0.98736 | 0.999508 | 0.996261 |
| F1-Score | 0.947296 | 0.987482 | 0.999509 | 0.99628 |
| Recall | 0.920474 | 0.995792 | 1.0 | 1.0 |
| Precision | 0.975729 | 0.97931 | 0.999019 | 0.992587 |
| AUROC | 0.948758 | 0.987348 | 0.999508 | 0.996256 |

Table 3: Accuracies and Losses of MLP and CNN models on Back-Testing

| **Models** | **Training** | | **Validation** | |
|---|---|---|---|---|
| | **Loss** | **Accuracy** | **Loss** | **Accuracy** |
| MLP | 0.0014 | 0.9997 | 0.0030 | 0.9995 |
| CNN | 0.0184 | 0.9941 | 0.0102 | 0.9963 |

Figure 3: Confusion Matrices results of the four models on Back-Testing

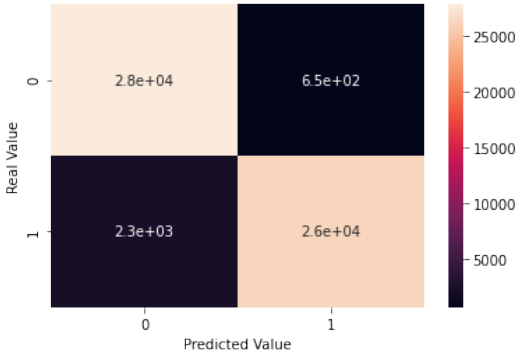


Figure 4: Logistic regression Confusion Matrix

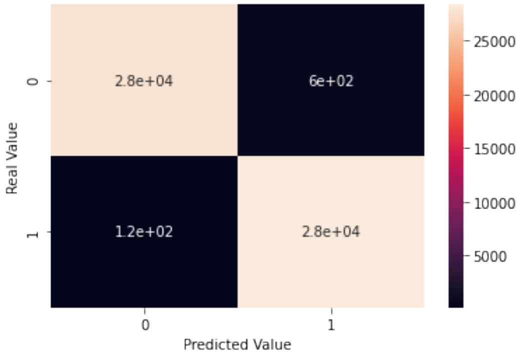


Figure 5: Decision Trees Confusion Matrix

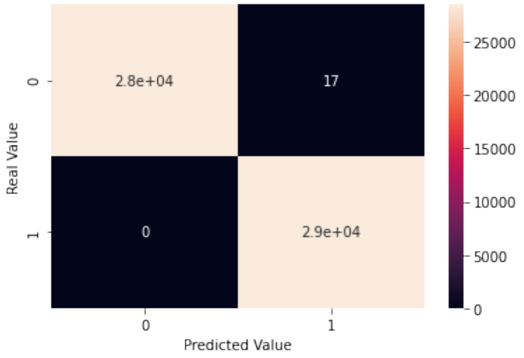


Figure 6: MLP Confusion Matrix

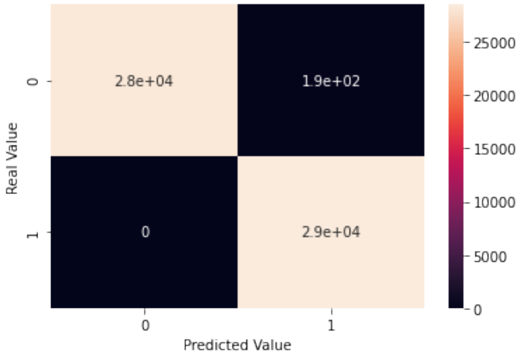


Figure 7: CNN Confusion Matrix

Figure 8: AUROC Curve results

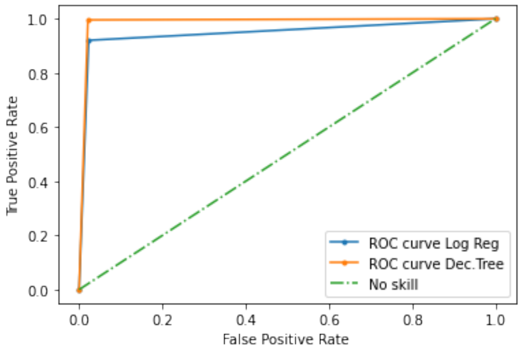


Figure 9: AUROC Curve of Logistic Regression and Decision trees

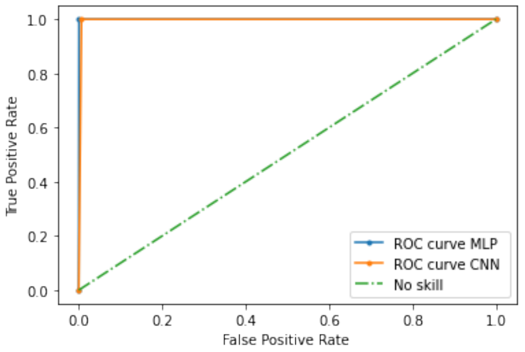


Figure 10: AUROC Curve of MLP and CNN.

### 9.2 Stress-Testing

We stressed our models on the generated data of 30 features including the target variable. For the generated data, we assumed the following scenario: 20% of transactions were fraudulent and 80% non-fraudulent. We obtained the results shown in Table 4.

Figure 11: Accuracy and loss of MLP during Back-Testing.

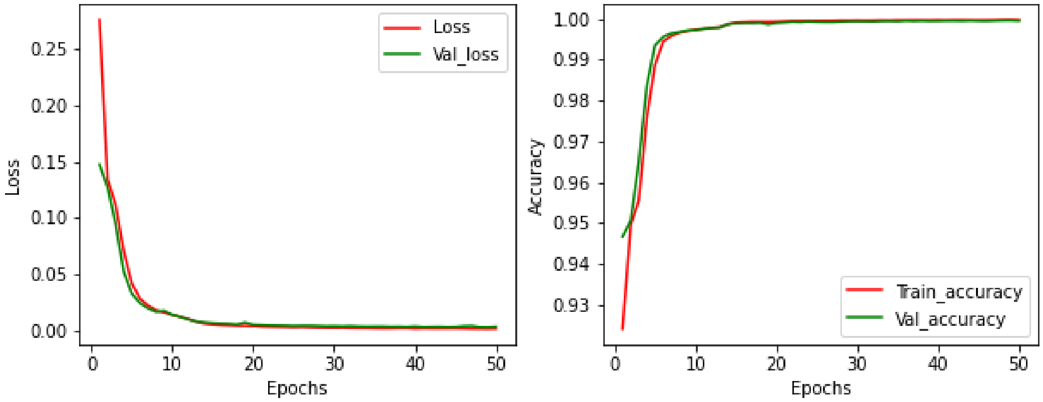


Figure 12: Accuracy and loss of CNN during training and Back-Testing

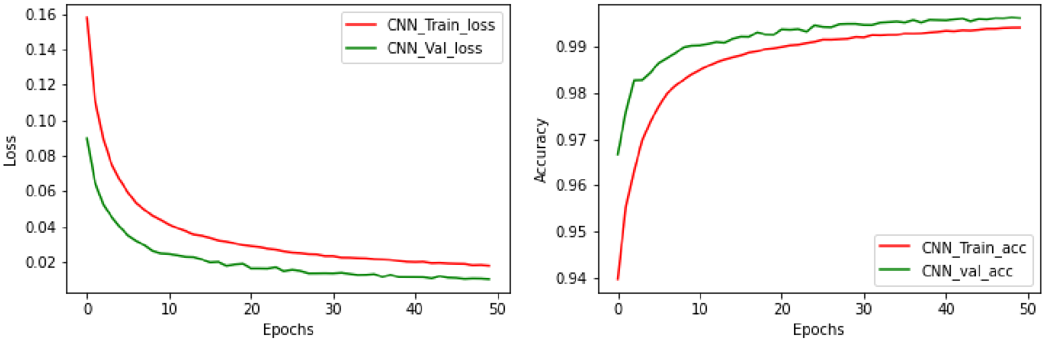


Table 4: Performance metrics of the different models on Stress-Testing

| **Metrics** | **Models** | | | |
|---|---|---|---|---|
| | **Logistic Regression** | **Decision Trees** | **Multi-Layer Perceptron** | **Convolutional Neural Networks** |
| Accuracy | 0.82324 | 0.20512 | 0.4108 | 0.4518 |
| F1-Score | 0.890843 | 0.012719 | 0.507226 | 0.478798 |
| Recall | 0.9016 | 0.0064 | 0.37905 | 0.31475 |
| Precision | 0.88034 | 1.0 | 0.766377 | 1.0 |
| AUROC | 0.70708 | 0.5032 | 0.458425 | 0.657375 |

From the Stress-Testing results obtained in the Table 4, we realised that all the models did not perform well as expected under the assumption of 20% of fraudulent transactions and 20 features where the linear combination of the target variable in the validation. This can be explained by the fact that the models have been trained on a transformed dataset from PCA and validated on simulated following the Gaussian distribution. According to Arnott [9], if a model can not perform as expected in out-of-sample testing, it can be improved by adding higher of powers of the inputs to the model.

Thus, we simulated another dataset of 25000 samples where we increased the proportion of number of fraudulent transactions to 30% against 70% of non-fraudulent transactions. We assumed that 20 features were a linear combinations of the target variable and were following the normal distribution. We obtained the results shown in Table 5.

Table 5: Performance metrics of the different models on Stress-Testing

| **Metrics** | **Models** | | | |
|---|---|---|---|---|
| | **Logistic Regression** | **Decision Trees** | **Multi-Layer Perceptron** | **Convolutional Neural Networks** |
| Accuracy | 0.70708 | 0.50332 | 0.50152 | 0.54236 |
| F1-Score | 0.755272 | 0.013192 | 0.00304 | 0.156828 |
| Recall | 0.904 | 0.00664 | 0.029511 | 0.08512 |
| Precision | 0.648568 | 1.0 | 1.0 | 0.995323 |
| AUROC | 0.70708 | 0.50332 | 0.50152 | 0.54236 |

## 10 Discussions

We considered a synthetic data to simulate future scenarios for stress-test of fraud detection risk. From the back-testing that was done, we found that MLP model performed better with an accuracy of 99%, F1-Score of 99%, AUROC of 99% and a precision of 99%. CNN also performed well giving an accuracy of 99%, F1-Score of 99%, ROC of 99% and a precision of 99% representing the second scoring model after the MLP. The logistic regression and decision tree gave a good accuracy and F1-score of 94% and 98% respectively. The both models performed well on the dataset, while giving a precision of 97 % each, also giving a good AUROC of 94% for logistic regression and 98% for decision tree. The confusion matrices results from Figure 3 show good predictions for the two classes, and the number of false negatives and false positives is not high. In general all the models performed well on the training and the historical data. From Table 3, the loss on the training and historical data is no different as they both show similar results of 0.0014 for MLP and 0.0184 for CNN.

Under stress scenarios, the logistic regression model was the one that performed well with an accuracy of 82%, F1-Score of 89%, precision of 88% and AUROC curve of 70%. The other models, decision tree, MLP and CNN gave a good precision of 100%, 76% and 100% respectively but not a better accuracy, F1-Score and AUROC curve. Also, by changing the assumptions on the stress-testing, logistic regression performed better by giving an accuracy and AUROC of 70% respectivly, F1-Score of 75%, precision of 64% compared to decision tree, MLP and CNN as shown in Table 5.

## 11 Conclusion

In this work, we have provided an overview of the applications of machine learning techniques used for financial risk models (Table 1) mainly operational risk, credit risk, market risk and liquidity risk. The main focus was on a subset of operational risk, risk of fraud for classification of a fraudulent transactions. We compared different machine learning such as logistic regression, decision tree, MLP and CNN models to detect financial fraud transactions and also to see which technique can yield higher accuracy and make good prediction. Besides, to test the performance of the models, Back-testing and Stress-testing have been done using historical data and simulated data respectively. From the Back-test comparison, we found that all the models performed well on historical data for all metrics but the MLP model consistently performed better than the other models followed by CNN model. Under stress scenarios, the logistic regression model was the one that performed well. The other models, decision tree, MLP and CNN gave a good precision but not a better accuracy, F1-Score and AUROC curve. Also, by changing the assumptions on the stress-testing, logistic regression performed better compared to decision tree, MLP and CNN. Logistic regression remains a robust model in different scenarios comparing to other models. This can be explained by the fact that logistic regression is very efficient to train and it makes no assumptions about distributions of classes in feature space. The logistic regression is fast at classifying unknown records and less inclined to over-fitting [7].

## References

[1] artificial intelligence in banking and risk management. `https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/third-party-whitepapers/en/artificial-intelligence-banking-risk-management-110277.pdf`. Accessed: 2020-11-20.

[2] kaggle competition. `https://www.kaggle.com/mlg-ulb/creditcardfraud`. Accessed: 2020-10-02.

[3] C. Acerbi, C. Nordio, and C. Sirtori. Expected shortfall as a tool for financial risk management. *arXiv preprint cond-mat/0102304*, 2001.

[4] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo. Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014, 2014.

[5] M. Albashrawi. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, 14(3):553–569, 2016.

[7] https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression

[6] K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi. Stock price prediction using k-nearest neighbor (knn) algorithm. *International Journal of Business, Humanities and Technology*, 3(3):32–44, 2013.

[7] S. L. Allen. *Financial risk management: A practitioner's guide to managing market and credit risk*, volume 721. John Wiley & Sons, 2012.

[8] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112. IEEE, 2014.

[9] R. Arnott, C. R. Harvey, and H. Markowitz. A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science*, 1(1):64–74, 2019.

[10] S. Assefa. Generating synthetic data in finance: opportunities, challenges and pitfalls. *Challenges and Pitfalls (June 23, 2020)*, 2020.

[11] C. Aymanns, J. D. Farmer, A. M. Kleinnijenhuis, and T. Wetzer. Models of financial stability and their application in stress tests. In *Handbook of Computational Economics*, volume 4, pages 329–391. Elsevier, 2018.

[12] J. L. Breeden. A survey of machine learning in credit risk. 2020.

[13] V. Darškuvienė. Financial markets. *Vytautas Magnus University. Kaunas*, 2010.

[14] M. F. Dixon, I. Halperin, and P. Bilokon. *Machine Learning in Finance*. Springer, 2020.

[15] S. Galletta and S. Mazzù. Liquidity risk drivers and bank business models. *Risks*, 7(3):89, 2019.

[16] L. Hu, J. Chen, J. Vaughan, H. Yang, K. Wang, A. Sudjianto, and V. N. Nair. Supervised machine learning techniques: An overview with applications to banking. *arXiv preprint arXiv:2008.04059*, 2020.

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

[18] V. Jokhadze and W. M. Schmidt. Measuring model risk in financial risk management and pricing. *International Journal of Theoretical and Applied Finance*, 23(02):2050012, 2020.

[19] C. Kingsford and S. L. Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[20] M. Leo, S. Sharma, and K. Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

[21] Y. Liu. Value-at-risk model combination using artificial neural networks. *Emory University Working Paper Series*, 2005.

[22] Y. Lucas. *Credit card fraud detection using machine learning with integration of contextual knowledge*. PhD thesis, Université de Lyon; Universität Passau (Deutscheland), 2019.

[23] Y. Lucas and J. Jurgovsky. Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479*, 2020.

[24] T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins. *Disrupting finance: fintech and strategy in the 21st century*. Springer Nature, 2019.

[25] S. Mahdavifar and A. A. Ghorbani. Application of deep learning to cybersecurity: A survey. *Neurocomputing*, 347:149–176, 2019.

[26] S. I. Nikolenko. Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*, 2019.

[27] B. C. on Banking Supervision and B. for International Settlements. *Principles for the management of credit risk*. Bank for International Settlements, 2000.

[28] A. Peter Martey, D. GUEGAN, and B. HASSANI. Credit risk analysis using machine and deep learning models.

[29] L. Rokach and O. Maimon. Decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005.

[30] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 2020.

[31] T. B. Shahi, A. Shrestha, A. Neupane, and W. Guo. Stock price forecasting with deep learning: A comparative study. *Mathematics*, 8(9):1441, 2020.

[32] M. Tavana, A.-R. Abtahi, D. Di Caprio, and M. Poortarigh. An artificial neural network and bayesian network model for liquidity risk assessment in banking. *Neurocomputing*, 275:2525–2554, 2018.

[33] E. Wipplinger. Philippe jorion: value at risk-the new benchmark for managing financial risk. *Financial Markets and Portfolio Management*, 21(3):397, 2007.

[34] W. XIONG. Machine learning in financial market risk: Var exception classification model. 2018.

[35] M. Zareapoor, K. Seeja, and M. A. Alam. Analysis on credit card fraud detection techniques: based on certain design criteria. *International journal of computer applications*, 52(3), 2012.

[36] J. Zhang, L. He, and Y. An. Measuring banks' liquidity risk: An option-pricing approach. *Journal of Banking & Finance*, 111:105703, 2020.

[37] J. Zupan. Introduction to artificial neural network (ann) methods: what they are and how to use them. *Acta Chimica Slovenica*, 41:327–327, 1994.