



Credit Risk Evaluation in Banking and Lending Sectors Using Neural Network Model

Jane B. Gelindon^{1*}, Rose Mary A. Velasco², Dionicio D. Gante³

^{1*,2,3}College of Computing Studies, Information and Communication Technology, Isabela State University –Cauayan Campus, Cauayan City, Isabela Philippines

Email: ²rosemary.a.velasco@isu.edu.ph , ³dioniciogante@gmail.com

Corresponding Email: ^{1*}gelindon.jane@isu.edu.ph

Received: 05 February 2022

Accepted: 20 April 2022

Published: 23 May 2022

Abstract: *This study presents the results of an experiment made with google colaboratory. The Google Colaboratory is also known as google colab it is a free online cloud-based Jupyter notebook environment that allows us to train our machine learning and deep learning models on CPUs, GPUs, and TPUs. We can determine the accuracy of the credit risk evaluation based on the dataset that was run in google colab using a neural network. The dataset has been thoroughly evaluated in creating a test harness for evaluating candidate models by calculating accuracy using k-fold cross-validation. When compared to a single train-test split, the k-fold cross-validation technique provides a reasonable general approximation of model performance. Based on the result, the correct number of rows was loaded, and through the one-hot encoding of the categorical input variables, it increased the number of input variables from 20 to 61. That suggests that the 13 categorical variables were encoded into a total of 54 columns.*

The result of the evaluation can help the banking and other financial sectors to assess a person before they are given a loan and if they can pay on time. This is a big help to reduce the losses of a company. Not only in the banking and other financial sectors as well as in lenders of goods.

Keywords: *Neural Network, Simulation and Modelling, Credit Risk Evaluation, German Credit Dataset.*

1. INTRODUCTION

Background of the Study

Nowadays, people have become accustomed to borrowing money or things that can be borrowed for many reasons. Especially with lending companies, banks, and even in



cooperatives and other financial sectors. What exactly is borrowing and why is it so prevalent. It is often the cause of quarrels, especially among relatives. How do we trust our creditors or how do banks and other lending and financial sectors trust us to lend? Evaluating people in debt is very important. Especially with banks and other financial organizations. Often there is also a bias due to the emotions that people display. And another reason is that they do not provide the correct data in their loan application even though the financial organizations and banks are strict, there are still those who can get away with it. So often the creditors are not well assessed or they just based on someone's work and assets to get into debt. As a result, banks and financial organizations incur losses due to non-payment on time. These loans contain significant amounts of money, and their non-recovery can result in a significant loss for the financial institution. As a result, reliable risk assessment is critical for banks and other financial institutions. It is not always most effective critical to reducing the risks of extending credit, but also to reduce errors in denying legitimate clients. This accurate assessment will protect the banks from legal action.

In this study, the researcher wants to determine if a person can get a loan through Credit Risk Evaluation Using Neural Network in Google colab to be able to train and validate the model. Credit risk evaluation can be viewed as a continuation of the credit allocation process. When an individual or business asks for a loan from a bank or financial institution, the lending institution evaluates the prospective advantages and costs of the loan. Credit risk evaluation using neural network is used to calculate the costs of a loan. According to Jepkemei (2019), there are several model in existence that are used in credit risk prediction. Wang (2022) said that Commercial banks are of great value to social and economic development. Therefore, the accurate evaluation of credit risk and the establishment of a credit risk prevention system has its important theoretical and practical significance. However, Laura Maria Badea Stroe (n.d.) said that compared with trees and logistic regression, traditional credit risk estimation techniques such as decision neural networks provide the best results in terms of correctly classified firms, proving also flexibility when operating with large number of variables. The german credit risk data set will also be used in this experiment for credit risk evaluation. Through this, we can see the accuracy of the credit risk evaluation based on the dataset that was run in google colab using a neural network. It will also help the banking and other financial sectors to assess a person before they are given a loan and if they can pay on time. This is a big help to reduce the losses of a company. Not only in the banking and other financial sectors as well as in lenders of goods.

Research Elaborations

The goal of this project is to test and validate the model in order to determine the accuracy of Credit Risk Evaluation Using Neural Network in google colab. The german credit risk data set will also be used in this experiment for credit risk evaluation. Through this, we can see the accuracy of the credit risk evaluation based on the dataset that was run in google colab using a neural network. The result of the evaluation can help the banking and other financial sectors to assess a person before they are given a loan and if they can pay on time



Specific Objectives

1. To simulate a model of an experiment by creating a test harness for evaluating candidate models and calculate the accuracy using k-fold cross-validation conducted in a neural network using google collaborator.
2. To determine the specific applicability and accuracy of the result using the different identifiers true positive, false-negative and precision, recall, and f-measure.
3. To compare the difference in the accuracy obtained using the two types of algorithms.

Literature Review

Neural Network

Neural Network as defined by IBM Cloud (2020), also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning that serve as the foundation for deep learning techniques. Their name and structure are derived from the human brain, and they are designed to mirror the way organic neurons communicate with one another. Artificial neural networks (ANNs) are made up of node layers, each of which has an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, is linked to another and has its own weight and threshold. If the output of any particular node exceeds the given threshold value, that node is activated and begins transferring data to the network's next tier. Otherwise, no data is sent to the next network layer. Warren McCulloch, a neurophysiologist, and Walter Pitts, a young mathematician, took the first step toward artificial neural networks in 1943 when they published a paper on how neurons may work. They created an electrical circuit model of a basic neural network. Despite this, proponents of "thinking machines" continued to make their points. The Dartmouth Summer Research Project on Artificial Intelligence, which began in 1956, gave artificial intelligence and neural networks a boost. One of the effects of this process was the stimulation of research in both the intelligent side of the brain, or AI as it is known in the industry, and the much lower level neural processing aspect of the brain. Because of their human-like characteristics and ability to execute tasks in endless permutations and combinations, neural networks are especially suited to today's big data-based applications. Because neural networks have the unique capability (known as fuzzy logic) of making sense of ambiguous, contradictory, or incomplete input, they can use controlled processes when accurate models are not available. According to Statista, global data volumes hit close to 100,000 petabytes (one million gigabytes) each month in 2017 and are expected to reach 232,655 petabytes by 2021. With organizations, individuals, and devices creating massive amounts of data, big data is important, and neural networks can make sense of it.

Eliana et.al (2016), described the case of a successful application of neural networks to credit risk assessment. They developed two neural network systems, one with a standard feedforward network (Gupta, 2019), while the other with a special purpose architecture (Jouppi et. al. (2018). The application is tested on real-world data, related to Italian small businesses. It shows that neural networks can be very successful in learning and estimating the bonis/default tendency of a borrower, provided that careful data analysis, data pre-processing and training are performed.



Mahdi, M.K. and Fahimeh, K.(2017) stated that the demand of banks for predicting their customer's credit risk has significantly increased and has become more critical, still challenging than ever. They addresses the problem of credit risk evaluation of bank's customers utilizing data mining tools. Three classification techniques include: neural network, C5.0, and classification and regression trees (CART) algorithms. In order to evaluate the performance of the classification techniques, an innovative two-stage evaluation process. A standard German credit dataset are used to validate the performance of the algorithms. It is illustrated that the neural network algorithm is the superior algorithm to evaluate bank customers' credit risk according to the authors.

Asogbon, M.G., and Samuel, O.W. (2016) , according to the researchers, a neural network based on backpropagation learning algorithm and a fuzzy inference system based on Mamdani model were developed to evaluate the risk level of credit applicants. A comparative analysis of the performances of both systems was carried out and experimental results show that neural network with an overall prediction accuracy of 96.89% performed better than the fuzzy logic method with 94.44%. Finding from this study could provide useful information on how to improve the performance of existing credit risk evaluation systems.

Germannano et.al (2020), Bayesian networks are compared to artificial neural networks (ANNs) for forecasting recovered value in a credit operation. In machine learning, the credit score problem is often handled as a supervised classification problem. The current study investigates this issue and concludes that ANNs are a more effective tool for predicting credit risk than the naive bayesian (NB) technique. The most important aspect is that loan decisions are associated with a set of elements to the extent that probabilities are used to classify new applicants based on their features.

Khemakhem, S., Said, F.B., Boujelbene, Y., (2018) investigates the efficacy of ROS and SMOTE in conjunction with logistic regression, artificial neural networks, and support vector machines. The authors discuss the role of sampling tactics in the Tunisian credit market and how they affect credit risk. These sample procedures may assist financial institutions in lowering the costs of erroneous classification when compared to unbalanced original data, as well as boosting the bank's performance and competitiveness.

The fact that neural network layers can extract information from any type of data is fascinating. This means that it makes no difference whether you're using picture data or text data. For both instances, the procedure of extracting useful information and training the deep learning model is the same.

Simulation and Modelling

Simulation modeling is the process of creating and analyzing a digital prototype of a physical model to predict its performance in the real world. Simulation modeling is used to help designers and engineers understand whether, under what conditions, and in which ways a part could fail and what loads it can withstand. Simulation modeling can also help to predict



fluid flow and heat transfer patterns. It analyses the approximate working conditions by applying the simulation software.

Simulation modeling allows designers and engineers to avoid the repeated building of multiple physical prototypes to analyze designs for new or existing parts. Before creating the physical prototype, users can investigate many digital prototypes. Using the technique, they can:

- Optimize geometry for weight and strength
- Select materials that meet weight, strength, and budget requirements
- Simulate part failure and identify the loading conditions that cause them
- Assess extreme environmental conditions or loads not easily tested on physical prototypes, such as earthquake shock load
- Verify hand calculations
- Validate the likely safety and survival of a physical prototype before

According to Kim (2020), It is critical to understand how to model a complicated system to make correct predictions. A widely used classical simulation modeling method that isolates causation between inputs and outputs by employing knowledge such as physical principles or operational rules. However, if data collecting of the actual system is difficult, it may produce a problem with the model's validity. Machine learning, on the other hand, is a way for representing a relationship between two sets of data. The model can be developed utilizing the target system's big data. It has a restriction in that it cannot predict accurately using the learnt model if the parameters or operating rules are modified after the model is learned.

Credit Risk Evaluation

Credit risk is defined by an Article of CPA Tools. It is the risk of loss due to a borrower not repaying a loan. More specifically, it refers to a lender's risk of having its cash flows interrupted when a borrower does not pay principal or interest to it. Credit risk is considered to be higher when the borrower does not have sufficient cash flows to pay the creditor, or it does not have sufficient assets to liquidate make a payment. If the risk of nonpayment is higher, the lender is more likely to demand compensation in the form of a higher interest rate. The credit being extended is usually in the form of either a loan or an account receivable. In the case of an unpaid loan, credit risk can result in the loss of both interests' debt and unpaid principal, whereas in the case of an unpaid account receivable, there is no loss of interest. In both cases, the party granting credit may also incur incremental collection costs. Further, the party to whom cash is owed may suffer some degree of disruption in its cash flows, which may require expensive debt or equity to cover.

Credit risk is a lesser issue where the selling party's gross profit on a sale is quite high, since it is only running the risk of loss on the relatively small proportion of an account receivable that is comprised of its own cost. Conversely, if gross margins are



small, credit risk becomes a substantial issue, forcing sellers to engage in detailed credit analyses before allowing sales on credit.

According Li, Lin and Chen (2017) shows that lending approaches performs a key quarter in banking business. Due to the wealthy improvement of net economic service. The excessive income has pushed banks to endure the so-referred to as credit score chance that the borrower might also additionally default or postpone the reimbursement of the mortgage. To save you from or to expect the credit score chance, there are numerous methodologies extensively applied, along with the credit score scoring gadget with credit score score-card, the Logistic Regression (LR) for assessing reimbursement capacity or linear scoring feature like Bayesian Decision Rules for credit score chance of lending. These conventional strategies want vast facts to do the evaluation of client lending. Proposed a brand new technique for credit score-chance dimension of micro-lending. To study the sample of awful credit score, the technique of information cleansing and Back-Propagation Neural Network (BPN) are adopted. The dataset of mortgage facts from a Brazilian business financial institution is used to run the method of credit score default awareness. Based on the test outcomes, we proved that the Artificial Neural Network can drastically enhance the accuracy of the notion whilst lowering the threat of misjudgment of the lending.

According to Wu (2017), Conventional credit score processes in particular targeting binary classification, which lacks ok precision to carry out credit score threat critiques in practice, however, uncommon research mentioned the effect of information preprocessing and variable choice. The lower back propagation community, fuzzy neural community, and genetic set of rules neural community. It emphasizes the significance of adopting big samples and using extra than labeled companies the usage of variable choice to growth predictive accuracy. The experimental effects indicated that, in low-threat and medium-threat credit score prediction, the accuracy fee of the hybrid neural community exhibited advanced overall performance to that of the conventional neural community.

A version of augmenting credit score danger control within side the banking industry was developed by Kogeda and Vumane (2017) where a loss of dependable credit score danger measurements and terrible manipulation of credit score dangers has brought on large economic losses throughout a huge spectrum of business. Financial establishments like banks are not capable of manipulation and comprise the speedy that will increase the credit score defaulting. The researchers addressed the credit score lending demanding situations via way of means of casting off credit score defaulting confronted via way of means of the banking industry. Data from a financial institution of formerly commonplace and rejected mortgage candidates turned into used to assemble a credit score danger assessment community. The synthetic neural community approach with back-propagation set of rules turned into carried out to increase a version that helps the banks within side the credit score granting decision-making.

A test concerning organization credit score danger assessment (Xiaobing H., et.al, 2018), carried out primarily on neural community set of rules to discover the organization credit score danger assessment. The software impact of numerous not unusual place where the



type of accuracy and the applicability of the version had been in comparison. In the end, the not unusual place hassle of optimization neural community set of rules primarily based totally on populace turned into solved: want to decide the size in advance. The experimental outcomes confirmed that the probabilistic neural community (PNN) had the minimal mistakes charge and 2d varieties of errors, whilst the PNN version had the best AUC fee and turned into robust. To sum up, the set of rules makes a few contributions to remedy the financing hassle of small and medium-sized businesses in China. In 2018, Zamore et al. established an artificial BP neural network evaluation model, which realized the intelligentization of bank credit risk assessment and improved the scientific evaluation of bank credit evaluation and management.

According to Sohony, Pratap, and Nambiar (n.d.) proposed an ensemble learning approach for credit card fraud detection as the ratio of fraud a normal transaction is bit appropriate. They observed that Random forest is best suited to provide a higher accuracy and neural networks for detecting the fraud instances. They also experimented with the large real-world credit card transactions. Ensemble learning is combination of Random forest and neural networks.

Shruti Goyal (2018), proposed a credit risk prediction using artificial neural algorithm to predict the credit default, several methods have been created and proposed. The use of method depends on the complexity of banks and financial institutions, size and type of the loan. The commonly used method has been discrimination analysis. This method uses a score function that helps in decision making whereas some researchers have stated doubts about validity of discriminates analysis because of its restrictive assumptions; normality and independence among variables. Artificial neural network models have created to overcome the shortcomings of other inefficient credit default models.

According to Jepkemei (2019), there are several model in existence that are used in credit risk prediction. In this process, it is important to use correct model from various different models present because the model chosen plays a crucial role in determining efficiency, accuracy and precision of the system. Predictor variables provide data which influences the credit loan risk but predicting models uses this data to predict whether the particular instance may be loan default instance or not. However, from various models, there is no specific model which can be said as the best model. Currently, the various models which are frequently used for prediction purposes include statistic-oriented models such as Discriminant Analysis (DA) and Logistic Regression (LR). Neural Networks (NN), genetic algorithms (GA) are also used for this purpose.

Related to the machine learning approach to the modeling and understanding of consumer credit risk according to Di, Qi, and Zhang (2019), academic studies concerning retail credit are fewer comparing to the vast majority of the credit risk literature that is corporate, sovereign or mortgage oriented. One reason, is that there is little outright trading of individual personal loans, hence no public assessments of retail credit risk. Unlike corporate bonds, secondary trading of securities related to consumer credit are only in secularized form². Another reason is the lack of account-level data unless one has access to proprietary data owned



by commercial banks and credit card companies. In terms of risk metrics and the models used, the historical focuses are credit scoring and linear regression when it comes to consumer credit risk. However, as e-commerce plays an ever-larger role in retail credit insurance and much richer data becomes available, sophisticated credit models are needed for the management of retail credit risk.

A study from Balakrishnan and Thiagarajan (2020), wrote that a new credit risk model for Indian debt securities rated by major credit rating agencies in India using the ordinal logistic regression (OLR). The robustness of the model is tested by comparing it with classical models available for ratings prediction. They improved the model's accuracy by using machine learning techniques, such as the artificial neural networks (ANN), support vector machines (SVM) and random forest (RF). They found out that the accuracy of our model has improved from 68% using OLR to 82% when using ANN and above 90% when using SVM and RF.

Fonseca , et. al. (2020) explores and evaluates the use of soft computing systems for clients' credit risk assessment in a Brazilian private credit card provider through the development of an innovative two-stage process, both involving soft computing techniques (fuzzy and neural networks). They use commercially available credit score ratings both in the development of the method and for benchmarking. After describing the development of the method, they presented a discussion about the comparison of performances of their method and a number of other credit scoring methods described in literature (for e.g. statistical and soft computing-based). One of the analyzed existing methods for instance involves the use of a soft computing algorithm only – Artificial Neural Networks (ANN) – for client classification into solvent or non-solvent, having a market available credit score rating as input. One of the most relevant contributions of this study however is the development of what they consider an innovative approach for credit scoring. This is a two-stage process that involves the use of a fuzzy inference model as input for an ANN model (what we call a fuzzy-neural approach), using commercially available credit score ratings as response in order to conduct the fuzzy reasoning step of the analysis.

Analysis of the first basic development of China's P2P online lending and the credit risks of borrowers in the industry according to Hou, and Zhang, (2021) characterized P2P network lending a credit risk assessment indicators system for borrowers in P2P lending is formulated with 29 indicators. Finally, on the basis of the credit risk assessment indicators system constructed in this paper, BP neural network is built based on the BP algorithm, which is trained by the LM algorithm (Levenberg-Marquardt), Scaled Conjugate Gradient, and Bayesian Regularization respectively, to complete the credit risk assessment model. By comparing the results of three mentioned training methodologies, the BP neural network trained by the LM algorithm is finally adopted to construct the credit risk assessment model of borrowers in P2P lending, in which the input layer node is the hidden layer node is 11 and output layer node is 1. The model can provide practical guidance for China and other countries' P2P lending platforms, and therefore to establish and improve an accurate and effective borrower credit risk management system.



Zhao, Jing (2021), conducted a research based on the efficiency of corporate debt financing based on machine learning and convolutional neural network for the digital age today, the attack of financial data has a great risk, so it is necessary to establish several specific procedures to ensure the security and privacy of our financial data. To consider the security and reliability of financial data, we should consider the security of financial data transaction. Using the form of nodes to retain the copy data of financial accounting, in this way to prevent the failure of the network. At present, a new type of network data backup method is proposed, which can be applied to financial transactions at the present stage. Using machines to classify and distribute financial accounts and backup data on each node, assuming that one node's backup fails, it will not affect the failure of adjacent nodes. No longer afraid of losing transaction data and other problems. The system is backed up by convolution neural network (CNN) and ledger. It also includes backup of credit and debit transactions and backup of transaction ID mode in timestamp and simulation. It is used to ensure the classification of accounts, including block information of chain area, hash value of previous block and latter block, etc.

According to Badea Stroe (n.d.), comparing a traditional credit risk estimation techniques such as decision trees and logistic regression, a neural networks provide the best results in terms of correctly classified firms, proving also flexibility when operating with large number of variables. Moreover, in this study, neural networks generated significantly higher detection rates for “default” cases, which represent in fact the main focus when developing credit risk models. Even if it takes more time building and configuring them, with the proper architecture and an optimal stopping rule, neural networks become highly performant techniques with a good generalization power.

Jiboning Zhang(n.d). Proposed a literature research on government financing platforms, analyzes its risk characteristics, and tries to establish an appropriate early warning evaluation index system of credit risk. In the selection of risk assessment methods, it departs from the traditional logistic evaluation model of regression analysis based on historical data starting from the fuzzy neural network (FNN) model of artificial intelligence method. It aims to quantitatively estimates the early warning indicators of credit risk of sample enterprises from the financial perspective, draws conclusions through empirical comparative analysis, and puts forward corresponding policy recommendations.

Jiang, et. al. (n.d.) proposed a novel method with a couple of stages. First they accumulate the transactions made through card holder, then primarily based totally at the behavioral styles transactions are aggregated, subsequent the dataset is classified, similarly the version is trained and sooner or later the version is tested. If any odd behavior arises then a comments is supplied to machine approximately the odd behavior through comments mechanism.

According to Zao (2021), He said that for the virtual age today, the assault of monetary statistics has a tremendous risk, so it's far vital to set up numerous specific processes to make



sure the safety and privations of our monetary statistics. The safety and reliability of monetary statistics, have to keep in mind the safety of monetary statistics transaction. Using the shape of nodes to hold the replica statistics of monetary accounting, on this manner to save you the failure of the community. At present, a brand new kind of community statistics backup technique is proposed, which may be implemented to monetary transactions at the prevailing stage. Zao further emphasized that using machines to categorize and distribute monetary debts and backup statistics on every node have an effect on the failure of adjoining nodes. The device is subsidized up through convolution neural community (CNN) and ledger. It additionally includes backup of credit score and debit transactions and backup of transaction ID mode in timestamp and simulation. It is used to make sure the type of debts, together with block records of chain area, hash price of preceding block and latter block, etc.

Duan (2019) said that since the chance of mortgage defaulting in peer-to-peer (P2P) lending is notoriously tough to evaluate, a deep neural community-primarily based totally decision-making technique is proposed on this paintings for extra powerful evaluation of P2P lending risks. Although typically a dozen functions had been used for neural community modeling in preceding research done through different researchers on comparable topics, extra complete functions along with each numeric and specific ones (e.g. domestic possession and cause of mortgage), are taken into consideration on this paintings for progressed modeling. Since specific statistics can't be used without delay because the input of neural networks, they're transformed to numerical statistics the usage of one-warm encoding function. The deep neural community (DNN) used on this paintings is a multilayer perceptron (MLP) with 3 hidden layers skilled through the back-propagation algorithm. In empirical analysis, the mortgage statistics issued through the Lending Club via 2007–2015 are categorized into 3 classes, i.e. secure mortgage, unstable mortgage and horrific mortgage the usage of Tensor Flow. The schooling and check statistics units include 221,712 and 55,428 statistics observations, respectively. Since maximum of the statistics belong to the magnificence of secure mortgage, Synthetic Minority Over-Sampling Technique (SMOTE) is used to enhance the DNN prediction accuracy. It is proven that with the proposed technique the check statistics are categorized at accuracy of 93%, that's an awful lot better than the predication accuracy of 75% received the usage of MLP with most effective one hidden layer.

According to Zi-sheng et. al (2021), recommended an Attention-LSTM neural community version to observe the systemic hazard early caution of China. Based on textual content mining, the community public opinion index is built and used as a education set to be integrated into the early caution version to check the early caution effect. The effects display that the community public opinion is the non-linear Granger causality of systemic hazard. The Attention-LSTM neural community has robust generalization ability. Early caution outcomes were extensively advanced. Compared with the BP neural community version, the SVR version and the ARIMA version, the LSTM neural community early caution version has a better accuracy rate, and its common prediction accuracy for systemic hazard signs has been advanced over short, medium and lengthy terms. When the eye mechanism is included within side the LSTM, the Attention-LSTM neural community version is even extra correct in all of the cases.



According to Cai, et. al (2020) to observe the hazard control of deliver chain, sell collaboration among node firms, and make sure the ordinary and properly improvement of deliver chain, a hazard assessment version of deliver chain primarily based totally on BP (Back Propagation) neural network (BPNN) became built to discover the present elements affecting the hazard assessment of the deliver chain and to construct hazard assessment. Before defining the formal method for back propagation, I'd like to provide a visualization of the process. First, to compute the output of a neural network via forward propagation.

German Credit Dataset

Sivakumar (n.d.) defined the German Credit data set is a publically available data set downloaded from the UCI Machine Learning Repository. All the details about the data is available in the above link. So it won't be describing the variables here. The data contains data on 20 variables and the classification whether an applicant is considered a Good or Bad credit risk for 1000 loan applicants.

The dataset was used as part of the Statlog project, a European-based initiative in the 1990s to evaluate and compare a large number (at the time) of machine learning algorithms on a range of different classification tasks. The dataset is credited to Hans Hofmann. The German credit dataset describes financial and banking details for customers and the task is to determine whether the customer is good or bad. The assumption is that the task involves predicting whether a customer will pay back a loan or credit.

According to Brownlee, J. (2020), The German credit dataset is a conventional unbalanced classification dataset with varying penalty for misclassification errors. Models tested on this dataset can be evaluated using the Fbeta-Measure, which allows for both broad quantification of model performance and incorporates the condition that one sort of misclassification error is more costly than another. Ajjheh, A, (2020) said that based on the collection of criteria, use current loan application data to forecast whether or not an applicant will be able to repay a loan.

Hayashi, Y. (2020) claims that the German credit scoring dataset, which includes numerical, ordinal, and nominal variables. The varied structure of this dataset makes achieving very high accuracy difficult. For the Wisconsin Breast Cancer Dataset, DL-based approaches showed excellent accuracy (99.68 percent), but DL-inspired methods achieved high accuracy (97.39 percent) for the Australian credit dataset. It seeks to provide new insights into why DL-based and DL-inspired classifiers fail to perform effectively on categorical datasets with nominal features. We also go over the issues that come with employing nominal attributes to create high-performance classifiers.

2. RESULTS OR FINDINGS

This study was to determine the accuracy of credit risk evaluation of a customer who wants to get a loan from banks or other financial sectors using a neural network that runs in a Google colabory-jupyter-notebook.

The German credit risk dataset was used in this experiment for credit risk evaluation. Through this, we can see the accuracy of the credit risk evaluation based on the dataset that was run in google colab using a neural network. Based on the result of the accuracy, it will limit only on the simulation and modelling to decide whether to good (approve) or bad (disapproved) credit application, using ten input datasets from the German Credit dataset.

Conceptual Framework

The researcher uses a common unbalanced machine learning dataset known as the "German Credit" dataset or simply "German" in this project.

The German credit dataset describes financial and banking details for customers and the task is to determine whether the customer is good or bad. The assumption is that the task involves predicting whether a customer will pay back a loan or credit.

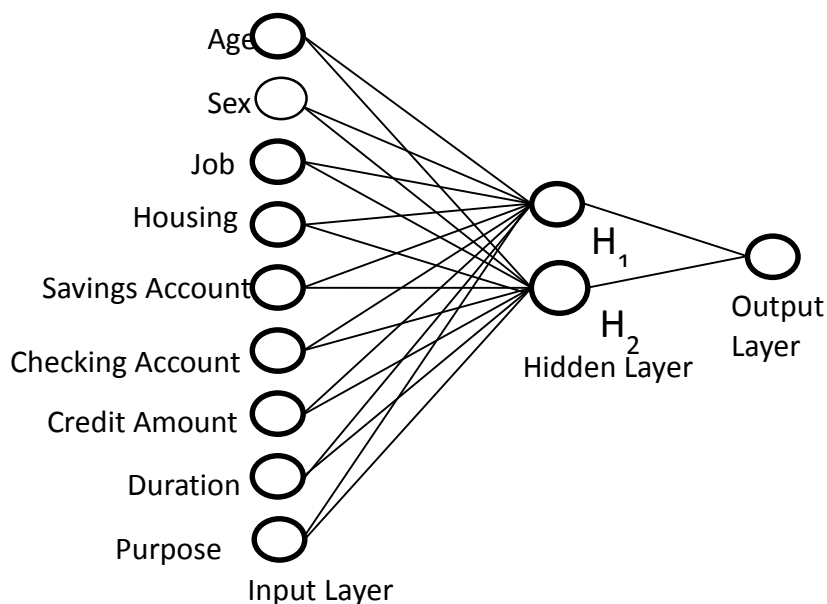


Figure 1. Conceptual Framework

Figure 1 shows the dataset includes 1,000 examples and 10 input variables, 4 of which are numerical (integer) and 3 are categorical. The hidden layers will test the accuracy of the dataset based on the risk whether a client is good or bad and to evaluate the dataset models to identify



a good or bad credit. In the output layer will identify the customers if it is good or bad customer. The different phases in exploring the data set was explained below:

Phase 1: Data Preparation

A standard unbalanced machine learning dataset known as the "German Credit" dataset or simply "German" will be utilized in this research. The dataset was used in the Statlog project, a European-based initiative in the 1990s that evaluated and compared a large number (at the time) of machine learning algorithms on a variety of classification tasks where Hans Hofmann is credited with creating the dataset. The German credit dataset contains financial and banking information about clients, and the aim is to evaluate whether the customer is good or bad. The work is assumed to involve estimating whether a consumer will repay a loan or credit. The German credit dataset is divided into two subsets, the training set, and the validation or testing set, which will be used in the development of neural network model for credit risk evaluation system. The training set is used to train the neural network model while the testing set is used for validating the neural network model that was trained. To avoid over-fitting in the training dataset, the cross validation method was used. The source dataset will be taken into 10 parts in each training process, wherein one part is used for the testing dataset and the remaining 9 parts are for the training dataset. Therefore, 9:1 is the training to validation ratio.

Phase 2: Explore the Dataset

Some of the categorical variables, such as "Savings account," have an ordinal relationship, but the majority do not. There are two classes: "1" for good customers and "2" for bad customers. Good customers are the default or negative class, whereas poor customers are the exception or positive class. Seventy percent of the cases are good customer, whereas the remaining thirty percent are bad customers.

Good Customers: Negative or majority class (70%).

Bad Customers: Positive or minority class (30%).

The dataset includes a cost matrix that assigns a different penalty to each misclassification error for the positive class. A cost of five is assigned to a false negative (classifying a bad client as a good customer), and a cost of one is allocated to a false positive (marking a good customer as bad).

Cost for False Negative: 5

Cost for False Positive: 1

This implies that the positive class is the emphasis of the prediction assignment, and that it is more expensive for the bank or financial organization to give money to a bad customer than it is to not give money to a good customer. This must be considered while choosing a performance metric. Prediction accuracy is the most commonly used metric for classification tasks in order to evaluate the model, although it is incorrect and sometimes dangerously deceptive when used on imbalanced classification tasks.



This is because if 98 percent of the data belongs to the negative class, you can achieve 98 percent accuracy on average by simply predicting the negative class all the time, resulting in a score that naively looks good but in practice has no skill. Instead, alternate performance metrics must be adopted.

Popular alternatives include precision and recall scores, which allow the model's performance to be evaluated by focusing on the minority class, known as the positive class. Precision is calculated by dividing the number of accurately predicted positive examples by the total number of positive examples anticipated. Maximizing precision will reduce false positives.

- **Precision** = TruePositives / (TruePositives + FalsePositives)

Recall predicts the ratio of the total number of correctly predicted positive examples divided by the total number of positive examples that could have been predicted. Maximizing recall will minimize false negatives.

- **Recall** = TruePositives / (TruePositives + FalseNegatives)

The performance of a model can be summarized by a single score that averages both the precision and the recall, called the F-Measure. Maximizing the F-Measure will maximize both the precision and recall at the same time. Using the F-Measure, you can compute the harmonic mean of precision and recall. This is a decent single number for comparing and selecting a model for this problem. The problem is that false negatives cause more harm than false positives.

- **F-Measure** = (2 * Precision * Recall) / (Precision + Recall)

This chapter presents the results based on the continuous testing, evaluation, and interpretation of the outcome. It also shows the results of the research conducted, and the tools being used by the researcher to determine and provide an answer to the objectives of the study.

Summary

Credit risk evaluation can be viewed as a continuation of the credit allocation process. When an individual or business asks for a loan from a bank or financial institution, the lending institution evaluates the prospective advantages and costs of the loan. Credit risk evaluation using neural network is used to calculate the costs of a loan.

3. RESULTS AND DISCUSSION

The German Credit Dataset was saved in the current working directory with the name "german_credit_data.csv" in order to view the contents of the file.

Running the dataset and confirming the number of rows and columns, that is 1,000 rows and 20 input variables and 1 target variable.

The class distribution is then summarized in table 1, confirming the number of good and bad customers and the percentage of cases in the minority and majority classes.

Table 1. Percentage Distribution



CLASS	COUNT	PERCENTAGE
1	700	70.000%
2	300	30.000%

In the Figure 3 shows the checking the NaN values of the dataset and check for an instance of missing value, whether it is numerical/categorical and in Figure 4.1 shows the numerical value of the data loaded in dataset pertaining to sex, housing, savings account, checking account, purpose and risk. You can see the total result in the said figure.

```
[3]:
Age          0
Sex          0
Job          0
Housing      0
Saving accounts 183
Checking account 394
Credit amount 0
Duration     0
Purpose      0
Risk         0
dtype: int64

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 0 to 999
Data columns (total 10 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Age             1000 non-null   int64
1   Sex             1000 non-null   object
2   Job             1000 non-null   int64
3   Housing         1000 non-null   object
4   Saving accounts  817 non-null    object
5   Checking account 606 non-null    object
6   Credit amount   1000 non-null   int64
7   Duration        1000 non-null   int64
8   Purpose         1000 non-null   object
9   Risk            1000 non-null   object
dtypes: int64(4), object(6)
memory usage: 85.9+ KB
```

Figure 3. Checking NaN Values

```
male      690
female    310
Name: Sex, dtype: int64

little    274
moderate  269
rich      63
Name: Checking account, dtype: int64

own       713
rent      179
free      108
Name: Housing, dtype: int64

car        337
radio/TV   280
furniture/equipment 181
business   97
education  59
repairs    22
domestic appliances 12
vacation/others 12
Name: Purpose, dtype: int64

little    603
moderate  103
quite rich  63
rich       48
Name: Saving accounts, dtype: int64

good      700
bad       300
Name: Risk, dtype: int64
```

Figure 4. Checking NaN Values by categorical



Taking a look at the distribution of the four numerical input variables by creating a histogram for each by showing the figure below that has been plotted.

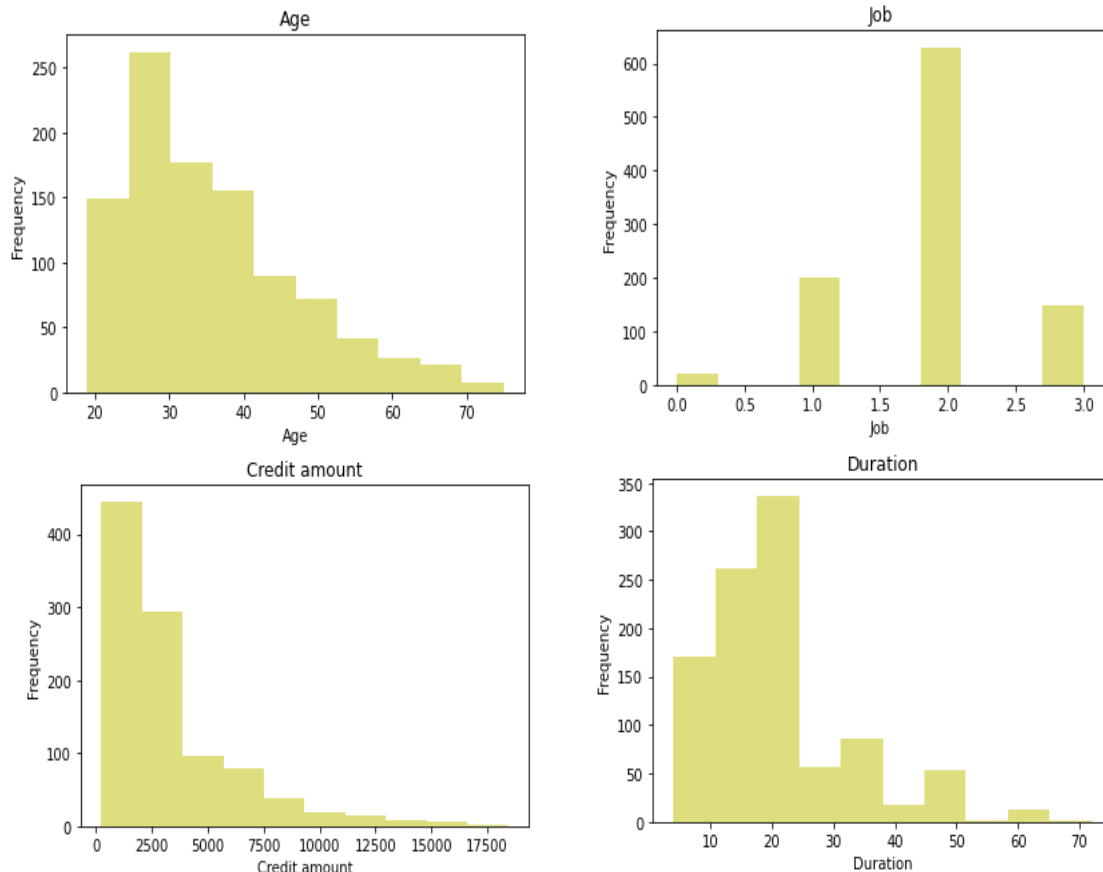


Figure 5. Histogram

Table 2. Distribution of Savings Account by Risk

Risk	Good	Bad
Little	498	105
Moderate	83	20
Quite Rich	58	5
Rich	43	5

Table 2 shows the description of distribution of savings account by risk values of the dataset.

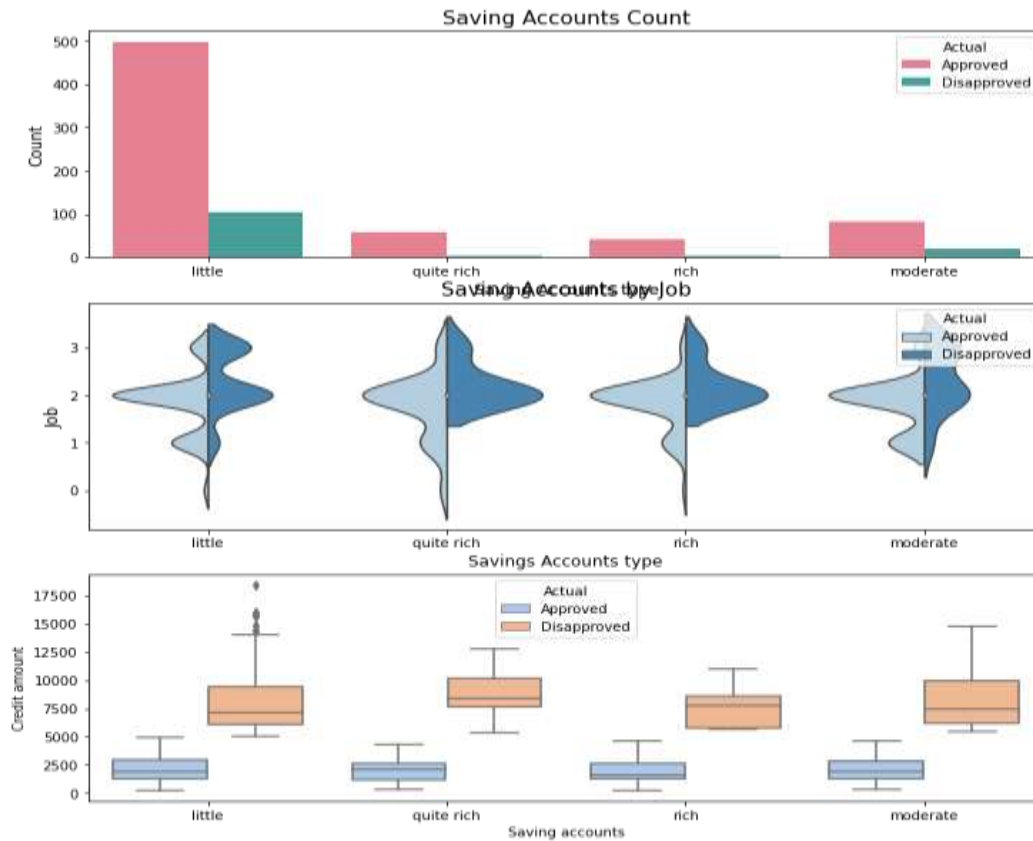


Figure 6. Savings Account Counts

Once the dataset has been thoroughly evaluated, next move is creating a test harness for evaluating candidate models by calculating accuracy using k-fold cross-validation. When compared to a single train-test split, the k-fold cross-validation technique provides a reasonable general approximation of model performance. We'll choose k=10, which means that each fold will include around 1000/10 or 100 epochs.

Next step is to determine whether a customer is excellent or bad and assign a class label to them. As a result, a suitable metric for assessing the anticipated class labels is required.

The task's emphasis is on the positive class (bad customers). Precision and recall are excellent places to begin. Maximizing precision will reduce false positives and maximizing recall will reduce false negatives in a model's predictions.

$$\text{Precision} = \frac{\text{TruePositives}}{(\text{TruePositives} + \text{FalsePositives})}$$

$$\text{Recall} = \frac{\text{TruePositives}}{(\text{TruePositives} + \text{FalseNegatives})}$$



Using the F-Measure, you can compute the harmonic mean of precision and recall. This is a decent single number for comparing and selecting a model for this problem. The problem is that false negatives cause more harm than false positives.

$$\mathbf{F\text{-Measure} = (2 * Precision * Recall) / (Precision + Recall)}$$

Remember that false negatives in this dataset are instances of a poor client being mislabeled as a good customer and receiving a loan. False positives occur when a good consumer is mistakenly labeled as a bad customer and denied a loan.

False Negative: Bad Customer (class 1) predicted as a Good Customer (class 0).

- **False Positive:** Good Customer (class 0) predicted as a Bad Customer (class 1).

False negatives are more costly to the bank than false positives.

- **Cost(False Negatives) > Cost(False Positives)**

In other words, we are interested in the F-measure, which summarizes a model's capacity to minimize misclassification errors for the positive class, but we prefer models that minimize false negatives over false positives.

This can be accomplished by employing a variant of the F-measure that computes a weighted harmonic mean of precision and recall while favoring greater recall scores over precision scores. This is known as the Fbeta-measure, which is a generalization of the F-measure, with "beta" defining the weighting of the two scores.

$$\mathbf{F\text{beta-Measure} = ((1 + \text{beta}^2) * Precision * Recall) / (\text{beta}^2 * Precision + Recall)}$$

A beta value of 2 will weight more attention on recall than precision and is referred to as the F2-measure.

- **F2-Measure = ((1 + 2^2) * Precision * Recall) / (2^2 * Precision + Recall)**

This metric will be used to evaluate models using the German credit dataset. We can write a function that will import the dataset and divide the columns into input and output variables. The categorical variables will be one-hot encoded, while the target variable will be label encoded. A one-hot encoding, as you may recall, substitutes the category variable with one new column for each variable value and marks values with a 1 in the column for that value.

As a result, divide the DataFrame into input and output variables first. Next, we need to select all input variables that are categorical, then apply a one-hot encoding and leave the numerical variables untouched. We can then label encode the target variable. The load_dataset() function below ties all of this together and loads and prepares the dataset for modeling. Next, we need a function that will evaluate a set of predictions using the fbeta_score() function with beta set to 2. We can then define a function that will evaluate a



given model on the dataset and return a list of F2-Measure scores for each fold and repeat. The `evaluate_model()` function below implements this, taking the dataset and model as arguments and returning the list of scores. Initially, we can evaluate a baseline model on the dataset using this test harness. A model that predicts the minority class for examples will achieve a maximum recall score and a baseline precision score. This provides a baseline in model performance on this problem by which all other models can be compared. Once the model is evaluated, we can report the mean and standard deviation of the F2-Measure scores directly.

Tying this together, the complete example of loading the German Credit dataset, evaluating a baseline model, and reporting the performance is listed in the figure below.

```
15 from pandas import read_csv
16 from sklearn.preprocessing import LabelEncoder
17 from sklearn.preprocessing import OneHotEncoder
18 from sklearn.compose import ColumnTransformer
19 from sklearn.model_selection import cross_val_score
20 from sklearn.model_selection import RepeatedStratifiedKFold
21 from sklearn.metrics import fbeta_score
22 from sklearn.metrics import mean_squared_error
23 from sklearn.dummy import DummyClassifier
24
25 # load the dataset
26 def load_dataset(full_path):
27     # load the dataset as a numpy array
28     dataframe = read_csv(full_path, header=None)
29     # split into inputs and outputs
30     last_ix = len(dataframe.columns) - 1
31     X, y = dataframe.drop(last_ix, axis=1), dataframe[last_ix]
32     # address categorical features
33     cat_ix = X.select_dtypes(include=['object', 'bool']).columns
34     # one-hot encode cat features only
35     ct = ColumnTransformer([('c', OneHotEncoder(), cat_ix)], remainder='passthrough')
36     X = ct.fit_transform(X)
37     # label encode the target variable to have the classes 0 and 1
38     y = LabelEncoder().fit_transform(y)
39     return X, y
40
41 # calculate F2 scores
42 def f2(y_true, y_pred):
43     return fbeta_score(y_true, y_pred, beta=2)
44
45 # evaluate a model
46 def evaluate_model(X, y, model):
47     # define evaluation procedure
48     cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
49     # define the model evaluation metric
50     metric = mean_squared_error
51     # evaluate model
52     scores = cross_val_score(model, X, y, scoring=metric, cv=cv, n_jobs=-1)
53     return scores
54
55 # define the location of the dataset
56 full_path = 'german.csv'
57 # load the dataset
58 X, y = load_dataset(full_path)
59 # summarize the loaded dataset
60 print(X.shape, y.shape, Counter(y))
61 # define the reference model
62 model = DummyClassifier(strategy='constant', constant=1)
63 # evaluate the model
64 scores = evaluate_model(X, y, model)
65 # summarize performance
66 print("Mean F2: %.3f (%.3f) % (mean(scores), std(scores))")
```

Figure 7. Evaluate



Figure 9 presents the measures of classification efficiency. The total number of instances is 1000, and the provided model accurately classifies 85.3% of instances is correctly classified.

```

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      853      85.3 %
Incorrectly Classified Instances    147      14.7 %
Kappa statistic                    0.5858
Mean absolute error                 0.2229
Root mean squared error             0.3221
Relative absolute error             58.7654 %
Root relative squared error         73.9903 %
Total Number of Instances          1000
    
```

Figure 9. Classification Efficiency Measures

In this type of confusion matrix, each cell in the table has a specific and well-understood name, summarized as follows:

Table 3. Confusion Matrix

Prediction	Actual	
	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Negative (FN)	True Positive (TP)

The precision and recall metrics are defined in terms of the cells in the confusion matrix, specifically terms like true positives and false negatives.

Table 3.1. Confusion Matrix Result

Prediction	Actual	
	Positive Prediction	Negative Prediction
Good Customer	697	49
Bad Customer	98	156

The Confusion Matrix results showed that 746 cases (697+49) were correctly identified and 254 (the remainder, out of 1000) were wrongly categorized. True positive has 697 clients who are really good debtors. True negative has 49 clients that are considered "bad" debtors (254 clients are considered "bad" clients). False negative and false positive classifications are incorrect. False negative shows 49 clients who were incorrectly predicted to be "bad" debtors, while False positive shows 98 clients who were incorrectly predicted to be "good" debtors.



Figure 10 shows the detailed accuracy result based on the weighted general average. True Positive Rate (0.853), False Positive Rate (0.305), Precision (0.847), Recall(0.853), F-Measure(0.847), MCC(0.591), ROC(0.912), PRC Area(0.926). As you can see on the result the number of significant figures is accurate.

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.934	0.386	0.877	0.934	0.905	0.591	0.912	0.965	c0
	0.614	0.066	0.761	0.614	0.680	0.591	0.912	0.812	c1
Weighted Avg.	0.853	0.305	0.847	0.853	0.847	0.591	0.912	0.926	

Figure 10. Detailed Accuracy Result

4. CONCLUSION

Financial firms employed credit risk evaluation models to estimate a potential borrower's probability of failure. The models provide information about a borrower's credit risk at any given time. If the lender does not notice the credit risk in advance, they are vulnerable to default and loss of funds. Lenders rely on the validation offered by credit risk evaluation models to make critical lending choices such as whether to grant credit to the borrower and the amount of credit to be charged. The dataset has been thoroughly evaluated in creating a test harness for evaluating candidate models by calculating accuracy using k-fold cross-validation. When compared to a single train-test split, the k-fold cross-validation technique provides a reasonable general approximation of model performance. Based on the result, the correct number of rows was loaded, and through the one-hot encoding of the categorical input variables, it increased the number of input variables from 20 to 61. That suggests that the 13 categorical variables were encoded into a total of 54 columns.

The class labels had the correct mapping to integers with 0 for the majority class and 1 for the minority class, customary for German credit dataset. The average of the F2-Measure scores is reported. In this case, the baseline algorithm was achieved with an F2-Measure of about 0.682. This score provides a lower limit on model skill.

- Based on the objectives, the dataset has been thoroughly evaluated, by calculating accuracy using k-fold cross-validation. When compared to a single train-test split, the k-fold cross-validation technique provides a reasonable general approximation of model performance. Choosing the k=10, which means that each fold will include around 1000/10 or 100 epochs.
- Once the model is evaluated, the mean and standard deviation of the F2-Measure scores directly reported. Using the F-Measure, you can compute the harmonic mean of



precision and recall. This is a decent single number for comparing and selecting a model for this problem.

- In this case, the baseline algorithm was achieved with an F2-Measure of 0.682. This score provides a lower limit on model skills whereas models that achieve a score below this value do not have skill on this dataset.

Acknowledgements

The researcher would like to show her heartfelt appreciation to everyone who contributed to the successful completion of her Thesis Project.

To **Mr. Dionicio D. Gante**, her thesis adviser, for his dedication to mentoring the researcher throughout the completion of her study.

To **Dr. Arnel C. Fajardo**, the Program Chair of the Graduate School, for his constructive comments that helped improve her work and for the never-ending support to finish this project. Lastly, the researcher also expresses her sincerest gratitude and appreciation to **Dr. Betchie A. Aguinardo**, Dean of CCSICT, for her unwavering guidance and support.

She could not be more appreciative of her efforts in making this academic work pleasurable.

5. REFERENCES

1. IBM Cloud Education (August 17, 2020) <https://www.ibm.com/cloud/learn/neural-networks#toc-what-are-n-2oQ5Vepe>
<http://www2.psych.utoronto.ca/users/reingold/courses/ai/cache/neural4.html#:~:text=The%20first%20step%20toward%20artificial,neural%20network%20with%20electrical%20circuits.>
2. Eliana, A., Giacomo, T., Andrea, R.(2016) A Neural Network Approach for Credit Risk Evaluation
3. Mahdi Massahi Khoraskani* and Fahimeh Kheradmand (2017) Application and comparison of neural network, C5.0, and classification and regression trees algorithms in the credit risk evaluation problem (case study: a standard German credit dataset) Download from: file:///C:/Users/JSONJHAINE-PC/Downloads/IJKEDM.2017.091013.pdf
4. Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabê, Sergei A. Kozlov (2020) Artificial neural network and Bayesian network models for credit risk prediction
5. Sihem Khemakem, Fatma Ben Said, Younes Boujelbene (2018) Credit Risk Assessment For Unbalanced Datasets Based On Data Mining, Artificial Neural Network And Support Vector Machines
6. Gupta, T. (2017). Deep Learning: Feedforward Neural Network. Published in Towards Data Science
7. Downloaded from <https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7>.



8. Jouppi, N.p. Young, C., Patil, N., Patterson, D. (2018). A Domain-Specific Architecture for Deep Neural Networks. Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59 10.1145/3154484.
9. Beong Soo Kim(2020) “Modeling and Simulation Using Artificial Neural Network-Embedded Cellular Automata”
10. CPA Tools(2022): <https://www.accountingtools.com/articles/what-is-credit-risk.html>
11. Li L. H.,Lin C. T.,Chen S. F.(2017), “Micro-lending Default Awareness Using ANN(Final)”
12. Wu H. C. (October, 2017),”Evaluating feature selection and neural network models in credit ratings”
13. Kogeda O. P.,Vumane N. N.(2017) “A Model Augmenting Credit Risk Management in the Banking Industry”
14. Xiaobing Huang †, Xiaolian Liu, Yuanqian Ren School (May ,2018),” Enterprise Credit Risk Evaluation Based On Neural Network Algorithm.”
15. 2018, Zamore et al “Credit Risk Research: Review and Agenda”
16. Ishan Sohony, Rameshwar Pratap, And Ullas Nambiar (N.D) “ENSEMBLE LEARNING FOR CREDIT CARD FRAUD DETECTION”
17. Shruti Goyal (2018), “Credit Risk Prediction Using Artificial Neural Algorithm”
18. Betty Jepkemei (July, 2019) “Effectiveness of Artificial Neural Network in Credit Risk Analysis.
19. Di Wang, Qi Wu, Wen Zhang (June,2019) “Neural Learning of Online Consumer Credit Risk”
20. Charumathi Balakrishnan* and Mangaiyarkarasi Thiagarajan(2020) “Credit Risk Model For Indian Debt Securities Rated By Major Credit Rating Agencies In India Using The Ordinal Logistic Regression (OLR).”
21. Fonseca, Diego PaganotiWanke, Peter Fernandes,Correa, Henrique Luiz(July,2020), “A Two-Stage Fuzzy Neural Approach For Credit Risk Assessment In A Brazilian Credit Card Company”
22. Zhengwei Ma ,Wenjia Hou,Dan Zhang (August ,2021) “Credit Risk Assessment Model Of Borrowers In P2P Lending Based On BP Neural Network”
23. Zhao, Jing (2021), “Efficiency Of Corporate Debt Financing Based On Machine Learning And Convolutional Neural Network”
24. Laura Maria Badea Stroie(n.d.) “Predicting Consumer Behavior with Artificial Neural Networks”
25. Jiboning Zhang(N.D) “Investment Risk Model Based On Intelligent Fuzzy Neural Network And Var” 1-s2.0-S2666285X21000066-main.pdf
26. Jing Duan(2019), “Financial System Modeling Using Deep Neural Networks (Dnns) For Effective Risk Assessment And Prediction”
27. Zi-Sheng Ouyang A , Xi-Te Yang B , Yongzeng Lai C, (2021) “Systemic Financial Risk Early Warning Of Financial Market In China Using Attention-Lstm Model”
28. Lulu Liu(2022) “A Self-Learning Bp Neural Network Assessment Algorithm For Credit Risk Of Commercial Bank”
29. SRISAI SIVAKUMAR; “German Credit Analysis”
30. Adam Hajjej (2020), German Credit Risk Classification : modeling and metrics



31. Yoichi Hayashi (2020), Does Deep Learning Work Well For Categorical Datasets With Mainly Nominal Attributes?
32. Dattachaudhuri, Abhinaba, Biswas, Saroj Kr., Sarkar, Sunita, Nath Boruah, Arpita, Chakraborty, Manomita, Purkayastha, Biswajit (2020) Transparent Neural based Expert System for Credit Risk (TNESCR): An Automated Credit Risk Evaluation System