“AN EXPLORATORY STUDY BASED ANALYSIS ON LOAN PREDICTION”

A DISSERTATION

Submitted in partial fulfilment of the requirements for the award of the degree
of
Master of Technology
in
Computer Science and Engineering

Submitted by
ANKIT SHARMA (Roll No: **2K20/CSE/04**)

under the guidance of
Dr. Vinod Kumar
Professor
Dept. of Computer Science and Engineering

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY, DELHI,
MAY 2022

DECLARATION

"I, Ankit Sharma, 2K20/CSE/04 student of M.Tech (Computer Science and Engineering), hereby declare that the project Dissertation titled "**An Exploratory Study Based Analysis on Loan Prediction**" which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition".

New Delhi
23-05-2022

Ankit Sharma

CERTIFICATE

I hereby certify that the Project Dissertation titled "**An Exploratory Study Based Analysis on Loan Prediction**" which is submitted by Ankit Sharma, 2K20/CSE/04, Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place : Delhi
Date : 23-05-2022

Dr. VINOD KUMAR
SUPERVISOR
Professor
Department of Computer Science and Engineering
Delhi Technological University

ABSTRACT

With the advancement in the financial sector, more people are seeking for bank loans. However, banks have limited resources that they must allocate to certain persons, analyzing and evaluating as to who would be a potential risk to the bank and who would not be, determining the credit to be given to the risk-free consumer. So, this research is an attempting to reduce the risk element associated with selecting the protected individual in order to save a large number of bank asset, endeavors and resources. It's done by trawling through previous records of those who have received advances previously granted, and the machine was constructed based on these records using the AI model that provides the most exact result. This research deals with the motive whether or not it is safe to lend a loan to a consumer keeping in mind the amount should be returned on time and the credit is safeguarded or not. It is one of the most pressing and significant factors related to the financial institutions and equivalent businesses, since it has a considerable influence on their net revenue and profitability. The presence of multi non-performing mortgages has risen considerably in recent years, jeopardizing the growth of these Banking Institutions. We present a method for implementing a neural nets model that will be used to forecast and predict loan mortgage default. The projection is made by considering the personal and monetary information given by the probable loan taker. The data-set which is used to train and evaluate the suggested neural network model. Our suggested neural network model surpasses alternative classifiers that are usually employed by monetary firms for loan default forecasting, based on the findings obtained. The accuracy obtained by our model on data set is close to 97.8%.

ACKNOWLEDGEMENT

"I might want to show my appreciation and thankfulness to my mentor, **Dr. Vinod Kumar**, Professor, Delhi Technological University for giving us the chance and the necessary rules to take a shot at this task alongside various counsels. I value my seniors for their thoughtful collaboration and significant consolation that assisted us with finishing this mission. I additionally offer our thanks to all other employees of our specialty for their steady consolation, and genuine help for this task work. Numerous individuals have offered significant remark recommendations on this proposition which gave us the motivation to improve our task. I thank all the individuals for their assistance, straightforwardly and in a roundabout way, in finishing our significant venture.

Ankit Sharma
(2K20/CSE/04)

CONTENTS

List of Tables

List of Figures

List of Abbreviations and Symbols

1. DL - Deep Learning
2. ANN - Artificial Neural Network
3. VM - Virtual Machine
4. NN - Neural Networks
5. BIns - Banking Institutions

CHAPTER 1: INTRODUCTION

1.1 A BRIEF OVERVIEW

Considering the financial institutions situation whether or not they are willing to lend a loan to a client is a risky factor for them. The following are the two most important banking issues: 1) What is the borrower's risk level? 2) Given the danger, should we credit to the borrower? The lender's interest rate is determined by the answer to the first question. The interest rate, together with other factors (such as the payback period), assesses the borrower's riskiness; the higher the rate of interest, the riskier the consumer. depending upon the interest rate, we will determine if the applicant is eligible for the loan. Lenders (investors) provide loans to creditors in exchange for interest-bearing repayment guarantees. The lender only gets paid (interest) if indeed the borrower pays back the loan. Whether he or she repays the loan or not, the lender loses money. Customers are given loans by banks in return for a guarantee of payback. Some people would fail on their loans because they failed to properly repay them for various reasons. In the event of a default, the bank keeps insurance to reduce the risk of collapse. The insured amount might be used to reimburse the whole loan or only a part of it. Banking operations rely on manual methods to evaluate whether or not a client is qualified for a loan. When there were a significant number of loan applications, manual techniques were usually successful, but they were inadequate. Making a choice at the moment would take a long time. As a consequence, the machine learning model for loan prediction may be used to analyze a customer's loan condition and develop plans. This system and approach extract and presents the key characteristics of a borrower that determine the loan status of the consumer. Finally, it provides the desired result (loan status). These reports make the work of a bank management easier and faster.

1.2 MOTIVATION

The Financial Banking Institutions have a huge overhead of managing their credit allotment to so and so clients, they require a mechanism which would automate their process of managing this problem statement of granting the loan to clients. This research is based on making the job of financial banking institutions easier for them to assess the nature of clients and their previous credit history.

Asset prices are steadily rising, and the money necessary to acquire an entire asset is quite expensive. So, we won't be able to buy it with your funds. Applying for a loan is the simplest approach to get the finances needed. However, obtaining a mortgage is a lengthy procedure. The application must go through many steps before being granted, and approval is not guaranteed. Many loan prediction algorithms have been devised to reduce the clearance time and risk connected with the loan. The goal of this research was to collect and analyze multiple Loan Prediction Models and determine which one had the least margin of uncertainty and can be utilized by financial institutions in the real world to forecast whether a loan should be accepted or not while considering risk. After analyzing and comparing the systems and approaches, it was discovered that the feed forward neural network-based prediction model was the most precise and fitting of all. This may save time and money by lowering the time and people necessary to secure loans and screening out the best applicants for lending.

There has already been a tremendous job done in this field and excellent algorithmic research have been devised to solve this problem statement, most of them delve upon regression algorithms and related to it, this research is based off a neural network approach to solve the problem statement, which would perform better in accuracy than traditional algorithms.

1.3 PROBLEM STATEMENT

All types of house loans are handled by some or the other financial organizations. They are present in all urban, semi-urban, and rural settings. The consumer first qualifies for a house loan, and the firm then verifies the customer's loan eligibility. The firm seeks to automate (in real-time) the loan qualifying procedure based on information supplied by customers while completing out online application forms. Sex, relationship status, schooling, family size, income, loan balance, good credit past, and other characteristics are included. They have offered a dataset to determine the client categories that are qualified for loan amounts, allowing them to target these customers directly. You can get more information on the issue statement and obtain the testing and the training data.

It could be clearly seen that; this is a Binary Classification issue in which we must forecast the label "Loan Status" as our Target label.

There are two possible loan statuses: Yes or NO.

If the loan is authorized, yes.

NO: if the loan is rejected.

So, we'll use the training dataset to train our model and attempt to predict our target column on the test dataset, which is "Loan Status.

CHAPTER 2: TECHNOLOGY STACK

2.1 TEXT EDITOR

The Visual Studio Code content management was used as an IDE. This Integrated Development Environment (IDE) supports a variety of programming languages. It includes an attached code editorial management, compiler, CLI which is the command line, and breakpoint debugger that can debug both machine and source code. It has been flattened out for the purpose of developing and researching current development online and cloud Apps.

2.2 PROGRAMMING LANGUAGE

Python 2.7 and Python 3.6 are the programming languages that we employ. Python offers for rapid system integration and a simple working environment which even a novice can grasp and operate quickly and efficiently. It may be used to create networks, servers, and a variety of other applications since it is an interpretive, broad sense general-purpose, and high-level programming language. It is built on object-oriented computing. It also offers a simple syntax that helps developers to produce precise, logical code quickly for both big and small projects.

2.3 GOOGLE COLAB

Google Colab is a free cloud administration for Artificial learning and research, analogous to Jupyter Notebooks. It offers an expanded platform for the execution of fully configured deep learning applications, as well as complementary access to a powerful GPU. We may use Google Collaboratory to construct a broad variety of data - intensive applications using the inexpensive Tesla K80 GPU, leveraging Keras, Machine learning, and TensorFlow. The project's benefit from Google Colab.

Free GPU support.

1. It allows distant users and developers to share Jupyter Notebooks as well as other files, as well as Google Docs.
2. The most important Python libraries are already installed.
3. It is built and developed on the Jupyter Notebook platform.
4. It enables for the free training of DL models.

2.4 OS VERSION

Windows 11 and Ubuntu 18.04 LTS were the operating systems utilized in our tests.

CHAPTER 3: LITERATURE REVIEW

3.1 Machine and Deep Learning

3.1.1 INTRODUCTION

Banks are vitally crucial in a market economy. Whether or whether a company succeeds is determined by the industry's ability to analyze credit risk. The bank assesses if a borrower is excellent or bad before issuing credit (defaulter)/(non defaulter). Predicting whether a debtor will fail or not default in the future is a challenging task for any firm or bank. Loan default prediction is fundamentally a binary classification problem. The credit history of the consumer determines the loan amount. The problem is assessing whether or not a borrower is in default. However, building such a system is challenging owing to the increased demand for loans. Companies may use model prototypes to make the right or accurate decision on whether to accept or refuse consumer loan requests. This project includes the building of an ensemble classifier by combining three distinct machine learning models. Banks compete fiercely for a competitive edge in order to grow their entire company. Client retention and fraudulent prevention are critical strategic instruments for banks to maintain healthy competition. Banks use a variety of risk assessment processes to determine whether or not to offer a loan to a person, if it would be profitable to them, and what dangers are associated if the consumer somehow doesn't pay back the loan on schedule, such as high interest rates. With the use of these types of mechanisms beneficial for BIns, the chance of defaulters might be assessed and investigated. As a result, they will be able to enhance their income creation while also being able to lend money to their consumers. As a result, while analyzing loan forecast, take all risk elements in mind. It becomes critical to develop a full-fledged model that can accurately foresee this, easing the burden on numerous financial BIns. As a consequence, financial institutions and Bins have undergone transformations and have grown exponentially. Customers' spending patterns have also radically transformed and affected the lending business in general.

When it comes to credit/money granted to a person, what precisely is a loan or a consumer loan? The loan may be defined as credit or money that a financial institution gives to a client in one go so that the customer can pay his or her obligation on time, providing that the customer is aware of all the risks associated if he or she is unable to pay the debt on moment to the Finance Ins. Education loans, different sorts of healthcare loans, travel fees to distant regions, or any staggered debt committed for various other causes are some of the loan reasons. It is clear that as the number of financial institutions offering loans to clients increases, so do the hazards.

As financial institutions develop exponentially, it becomes more important for BIns to differentiate themselves from the competition by offering consumers the finest loan schemes possible, which is a challenging undertaking in and of itself and would need extensive investigation. Furthermore, it

is crucial from the customer's perspective to comprehend whether or not he or she will be able to pay back the debt. These are usually made with extreme care and subjected to extensive research before being given to someone. Despite this research, it is clear that the problem has not yet been addressed, and that an automated model would be required, which would need to be trained and refined on a regular basis.

There are a large number of bad loan instances, which may be classified into a broad range of minor to major frauds. Some resulted in massive fraud, resulting in BIns becoming bankrupt. With these factors in mind, it is critical to predict loan default situations, which may be determined by looking at the consumer's previous payment history, if they failed to properly pay it on time or not.

To forecast loan default, many prediction approaches have emerged. The strategy used in this study is centered on using a neural network to improve the prediction model's accuracy.

Machine Learning

It is a subfield of artificial intelligence. It's a strong method of teaching computers to develop and learn from their experiences without having to be explicitly programmed. The machine may learn from the information available in its surroundings (or experience). This is also developed to increase overall performance.

Supervised Learning

Supervised learning techniques learn from previous data and generalize it to new data. In supervised methods, the data is 'labelled.' Let's imagine a training dataset contains X-Y pairs (where X is the source and Y is the result), and the system trains an algorithm that assesses a suitable output for a given input. A sample set that reflects the device in observation is required for such a paradigm, and it may also be used to test the approach's correctness.

Unsupervised Learning

Unsupervised machine learning is a kind of machine learning in which the model discovers patterns and information on its own. The data used to train the system is not categorized or labelled in any way. Without any previous training, the computer is responsible for organizing the data as per patterns, similarities, and differences.

The purpose of this type of learning would be to model the organization in order to get a better understanding of the data. Unsupervised learning is divided into two categories: Association and Clustering.

Deep Learning

Deep Learning is an ANN-based subset of machine learning. The human brain is represented as a NN. An Artificial Neural Network is inspired by the structure and function of the brain. Even though the information is unlabeled or disorganized, Deep Learning AI can learn. Deep Learning is often used to object recognition, understand language, discern dialects, and make decisions.

Supervised Deep Learning

Based only on raw-data representation learning, a computer may find out the parameters necessary to identify or categorize activities. Other traditional machine learning algorithms, on the other hand, are incapable of dealing with raw sample data. The transition from a low-level to a higher-level abstract model is performed thanks to various layers of representation. Machines have been able to learn increasingly complicated functions thanks to large numbers of these modifications. Deep learning approaches have also outperformed traditional algorithms for a variety of machine learning applications, including voice recognition, object identification, intrusion prevention, and NLP comprehension, to name a few.

There are three types of DL models:

- Both discriminative and generative models are used in hybrid models.
- Unsupervised learning algorithms are used in-generative models.
- Discriminative systems use supervised learning approaches.

3.1.2 Neural Network

Neural Networks are artificial intelligence computations that mimic the human sensory perceptions. A neuronal network (NN) is a collection of neurons arranged into a circuit. On the basis of false neurons, a neural network (NN) is termed a counterfeit (Artificial) neural Network (ANN). Neural structures are non-direct quantifiable information exhibiting or adaptive apparatus in everyday words. They may be used to show intricate linkages between sources of data and outputs or to find patterns in data.

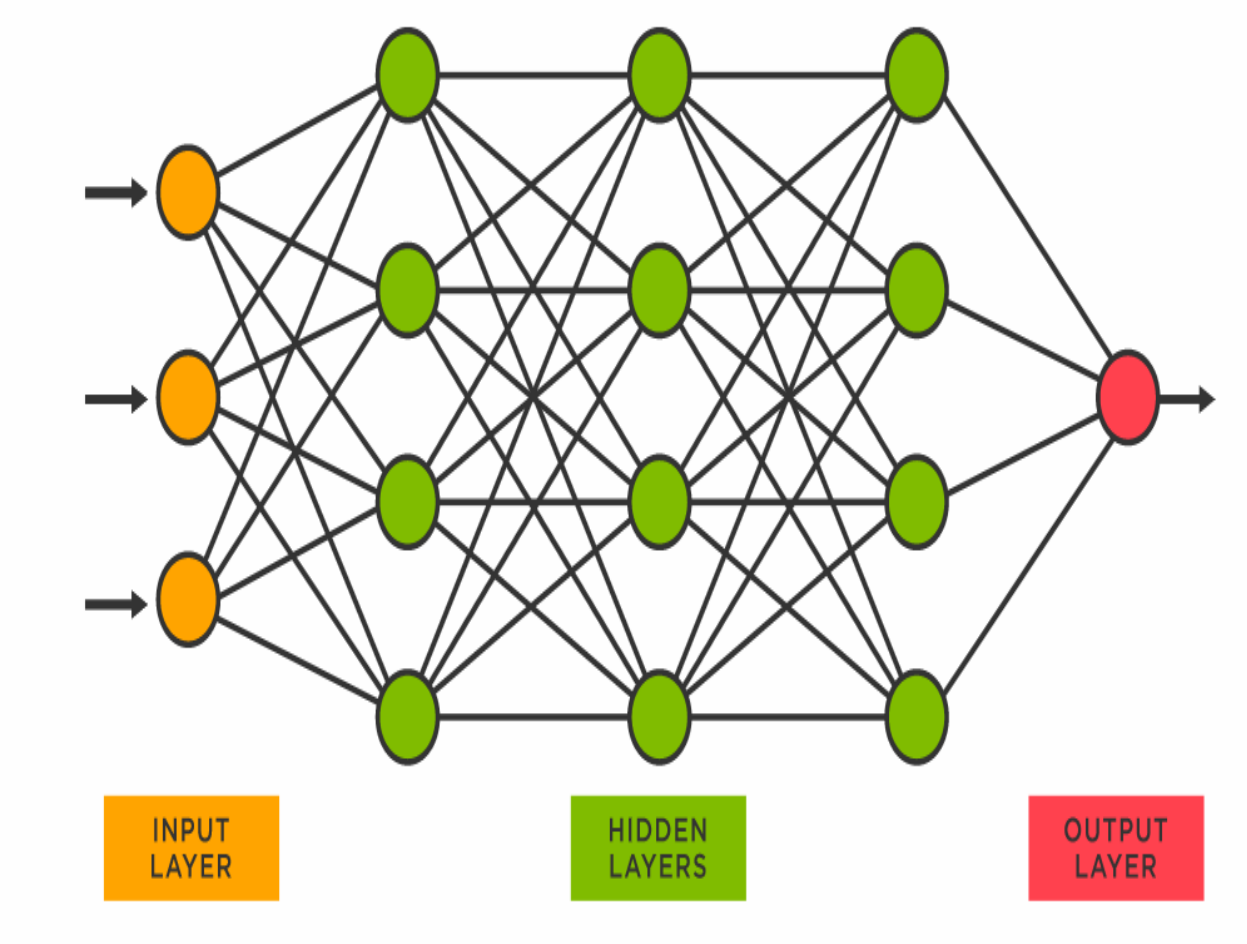

Figure 3.1 – A simple NN

In its most basic form, an artificial neural network includes three layers of neurons. Like in the human brain, information travels from one to another:
The input layer is where data enters the system.
The information is handled in the hidden layer.
The output layer is where the system makes decisions based on data.
Artificial neural networks with more complexity will have numerous layers, some of which will be hidden.
The neural network, like artificial neurons, is made up of a group of nodes or linked units. The neuron network in the animalistic instincts is roughly modelled by these nodes. An artificial neuron, like its organic counterpart, accepts a signal in terms of a stimulus, analyses it, and sends messages to other neurons linked to it. The similarities, however, stop there.

3.2 RELATED WORK

In this segment, Different researchers have worked on techniques to anticipate reliable outcomes utilizing many computational intelligence learning models in similar research with loan prediction. [1] It is clear from this research that a prediction system was implemented from the mathematical programming point of view and direction, which used coding with integers and data envelopment research findings into an innovative, inventive, and novel breed of Dimension Reduction, which assisted in loan forecasting. It properly predicted the model with excellent accuracy. [2] Examined loan default utilizing data from the UCI database and exploratorily examined it using discrimination analysis, classification, and regression tree, Back propagation to show that they either have their own expertise. [3] took Taiwan consumer loan data and displayed it, as well as compared it to several models such as LR models, DA, Neural Networks, and CART models, and came to the conclusion that the CART model attained and outperformed some other three different models in terms of average classification accuracy. [4] employed SK or South Korean data, which was acquired from the cc center and used to construct a prediction model; it was a unique strategy that used a scheduling dependent feature and was also used to forecast non-constants. After that, they used SA, or survival analysis, to examine the data using linear regression & neural nets. Their study revealed that regression analysis and neural networks were more accurate, although the SA was more tactful. [5] examined and collected data on certain former banks customers, calculating how loans were provided based on a set of exact criteria. They utilized certain data to train the models (ML) and a specific set of data to evaluate them, resulting in reliable results. The main goal of this research was to predict the loan's/outcome. safety's mortgage's They cleaned the data they used by modifying it to eliminate specific missing variables. They have 1500 samples with ten numerical and eight categorical characteristics in their data set. These qualities were developed even more. The examination included several characteristics such as BV, Customer Assets, and CIBIL score. The precision reached was 81.1 percent. [6] investigated logistic regression and how to mathematically describe it. Vaidya used a machine learning strategy called logistic regression to combine probabilistic and predictive approaches to tackle a problem with loan approval prediction. His study used logistic regression to determine whether or not a loan for data collection from an application would be accepted. It covers logistic regression, Random Forest, and other real-world applications of machine learning models. He specialized in logistic regression and developed statistical models for it. He also determined that logistic regression was limited to small sample estimates and needed specific independent factors to avoid bias toward the dependent variables. He also observed that using the ANN model will help since it is more visual and has more layers of nodes, resulting in a more sophisticated structure and a better prediction model. [7] uses Python to do a logistic regression study on skewed datasets, creating a variety of categorization criteria based on the fraction of skewed datasets. This research concentrated on a handful of the categories data sets. The purpose of this research was to focus on the data set rather than improving accuracy. The criteria were largely centered on the imbalanced data set since it was used. [8] concentrated on the logistical math equation, error probability formulation, gradient descent technique for calculating the regression coefficient, and Nonlinear activation function augmentation. According to this paper,

which focused on the logistic model and utilized the gradient descent technique, the sigmoid function might be improved by exceeding the value of n.

[9] evaluated the data gathered using a data mining approach. Because it quickly detects customers who are capable of taking the loan and repaying it within a certain time limit, the data mining technique delivers improved vision in loan prediction systems. The "J48," "NB," and "Bayes net" algorithms were employed. Their datasets were trained using these approaches. They determined that the "J48" algorithm had an accuracy of 78.3784 percent throughout their investigation, enabling the lender to decide whether or not to supply the customer with a loan.

CHAPTER 4: PROPOSED WORK

4.1 Introduction

This is where the research implementation section begins. The following study examines a loan default predictive model that differs from most others in its implementation. We utilized a dataset, which is unstructured and in raw format with a huge number of features, some of which are related to one another; thus, we built our substitute dataset with lower dimensionality and based the loan prediction engine on it.

Despite the fact that much work has been put into bad loan prediction systems, they are ineffective and erroneous. We're working on a prediction system that uses a neural network technique with several hidden layers and convolutions inside a single layer. The neural network technique was chosen because of its greater flexibility and ability to create non-linear predictions based on predictive factors.

4.2 Dataset and its Structure

In our feed forward Neural Network model, the dataset that we used consist of 5000 clients or the consumers. The description about the nature of clients that are there in the dataset are collected from different resources.

Variable	Description	Type
ID	Customer ID	INT64
Age	Customer Age	INT64
Experience	Amount of work experience in years	INT64
Income	Amount of annual income (in thousands)	INT64
Zipcode	Zipcode of where customer lives	INT64
Family	Number of family members	INT64
CCAvg	Average monthly credit card spendings	FLOAT64
Education	Education level (1-Bachelor, 2-Master, 3-Advanced Degree)	INT64
Mortgage	Mortgage of house (in thousands)	INT64
Securities Account	Boolean of whether customer has a securities account	INT64
CD Account	Boolean of whether customer has Certificate of Deposit account	INT64
Online	Boolean of whether customer uses online banking	INT64
CreditCard	Does the customer use credit card issued by the bank?	INT64
Personal Loan	This is the target variable (Binary Classification Problem)	INT64

Table 4.1: - Dataset Structure

4.3 Exploratory Analysis

The data was displayed, and the trend could be seen; the correlation chart, which is used to discover the correlation between the DF or data frames, is shown below. We may see how our various qualities relate to each other using the data provided. The significant correlation between experience and age may be seen. As you become older, you get more experience.

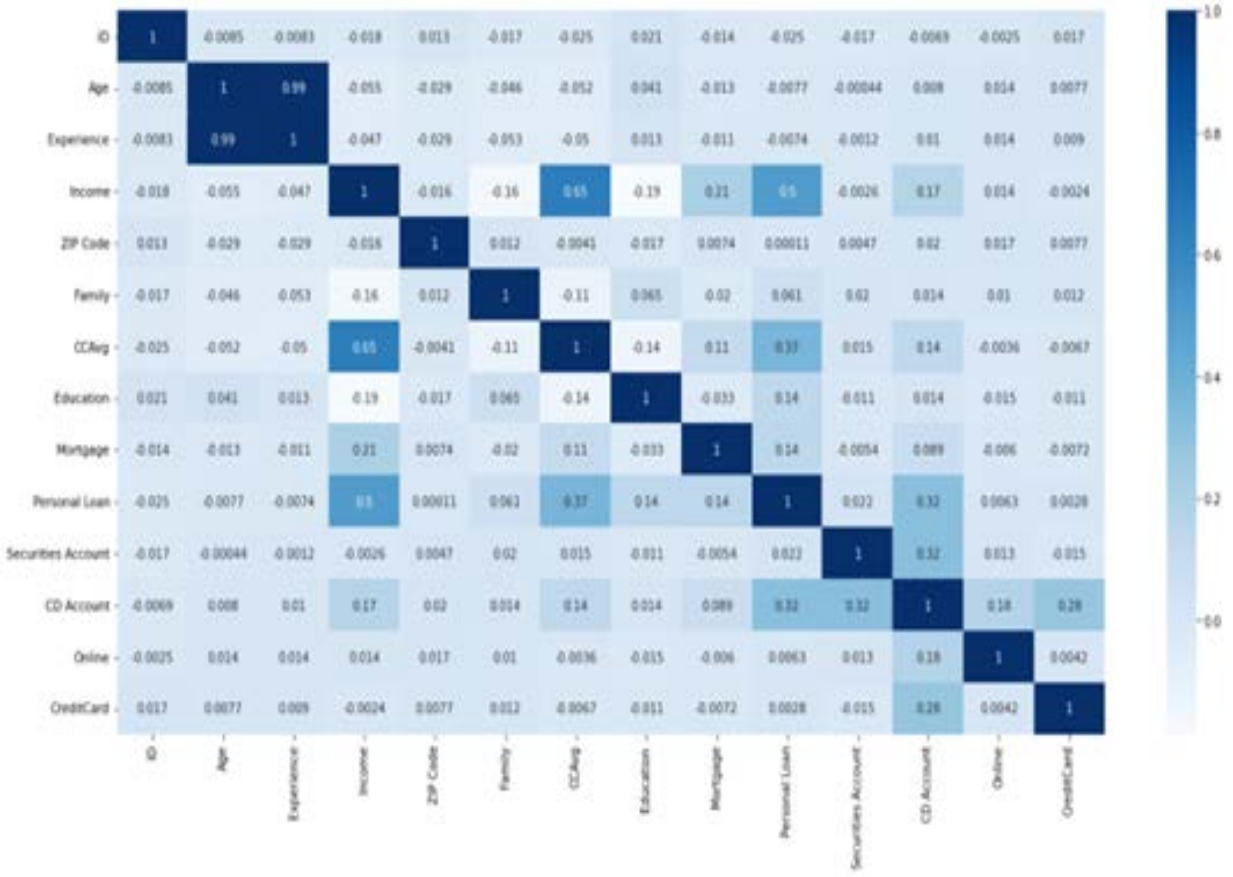

Figure 4.1 Correlation Matrix

Another visualization might be observed in the income and number of people who were authorized for loans and those who were not. The tendency can also be seen here: as a person's or individual's income rises, the likelihood of the loan being accepted rises as well. It can also be seen that more and more individuals sought to acquire their loans even though their revenue was less than the desired amount. Individuals with higher incomes did not apply for a loan in larger numbers, and they are also quite few in number.

4.4 Data Preparation

As part of the approach, a dataset is separated into two subsets. The training dataset is the starting point for fitting the model. Instead of using the next subset to build the system, the dataset's inputs element is supplied to the model, which produces predictions and compared them to the estimated parameters.
We will first transform some of the known factors for evaluating the behavior back to data attributes like the Personal Loan, which were updated throughout the Data exploration to produce the output into binary form. Following that, the information must be separated into two categories: training and testing. The data is divided into 85 percent testing and 15 percent training for ease of use.

4.5 Building the Model

A Neural Network has three layers, but there might be more depending on the number of hidden layers, which is variable. The input layer, hidden layers, and the output layer, which is the ultimate one, are the layers. The equation below depicts a generic neural network model made up of several neurons.

$$h_i = \sigma(\sum_{j=1}^{N} V_{i_j} x_j + T_i^{hid}) \qquad \text{....(1)}$$

The result which we get from the above equation is the hidden layer (h_i), σ is the activation function, N are the count of neurons in the input, V_{ij} are the weights, x_j is what is fed to the neuron which is the input, T_i^{hid} are the concealed neurons' threshold terms. [12]

```
Epoch 20/25
120/120 [==============================] - 2s 17ms/step - loss: 0.0867 - f1: 0.9677 - val_loss: 0.0860 - val_f1: 0.9787
Epoch 21/25
120/120 [==============================] - 2s 17ms/step - loss: 0.0784 - f1: 0.9719 - val_loss: 0.0885 - val_f1: 0.9759
Epoch 22/25
120/120 [==============================] - 2s 17ms/step - loss: 0.0868 - f1: 0.9688 - val_loss: 0.0881 - val_f1: 0.9773
Epoch 23/25
120/120 [==============================] - 2s 17ms/step - loss: 0.0753 - f1: 0.9719 - val_loss: 0.0871 - val_f1: 0.9744
Epoch 24/25
120/120 [==============================] - 2s 16ms/step - loss: 0.0814 - f1: 0.9698 - val_loss: 0.0871 - val_f1: 0.9773
Epoch 25/25
120/120 [==============================] - 2s 17ms/step - loss: 0.0791 - f1: 0.9719 - val_loss: 0.0870 - val_f1: 0.9787
```

Figure 4.2 25 Epochs of Training yielded Results.

In general, the ANN employs one of two models: sequential or functional. For a basic stack of layers with only one input and one output node, a sequential method is perfect. The functional method can handle models with non-linear design, overlapping layers, and even multiple inputs and outputs. We employed a sequential Neural Network technique in our implementation. We supplied the model the modified features, which made it much easier for the model to analyze the data. To boost the depth and accuracy of our output, we called the add many times to the dense and dropout layers. We discovered that there were 1,188,952 total trainable parameters after getting the many layers to work.

Layer (type)	Output Shape	Param #
dense_21 (Dense)	(None, 250)	3500
dropout_18 (Dropout)	(None, 250)	0
dense_22 (Dense)	(None, 550)	138050
dropout_19 (Dropout)	(None, 550)	0
dense_23 (Dense)	(None, 550)	303050
dropout_20 (Dropout)	(None, 550)	0
dense_24 (Dense)	(None, 550)	303050
dropout_21 (Dropout)	(None, 550)	0
dense_25 (Dense)	(None, 550)	303050
dropout_22 (Dropout)	(None, 550)	0
dense_26 (Dense)	(None, 250)	137750
dropout_23 (Dropout)	(None, 250)	0
dense_27 (Dense)	(None, 2)	502

Total params: 1,188,952
Trainable params: 1,188,952
Non-trainable params: 0

Figure 4.3 Trainable Parameters Obtained.

The model's training data was divided into two portions for cross validation, with 15% of it utilized in cross validation. In addition, the accuracy, recall, and f1 were determined. After then, two optimizers, Adam's optimizer and the Stochastic Gradient Descent Optimizer, were used to visualize the model during compilation (SGD). The next section delves further into these optimizers. The SGD optimizer was utilized in this implementation. With loss as categorical cross entropy, we trained the model for around 25 epochs and 15% cross validation split. After training the model, the total epochs were used to assess its performance.

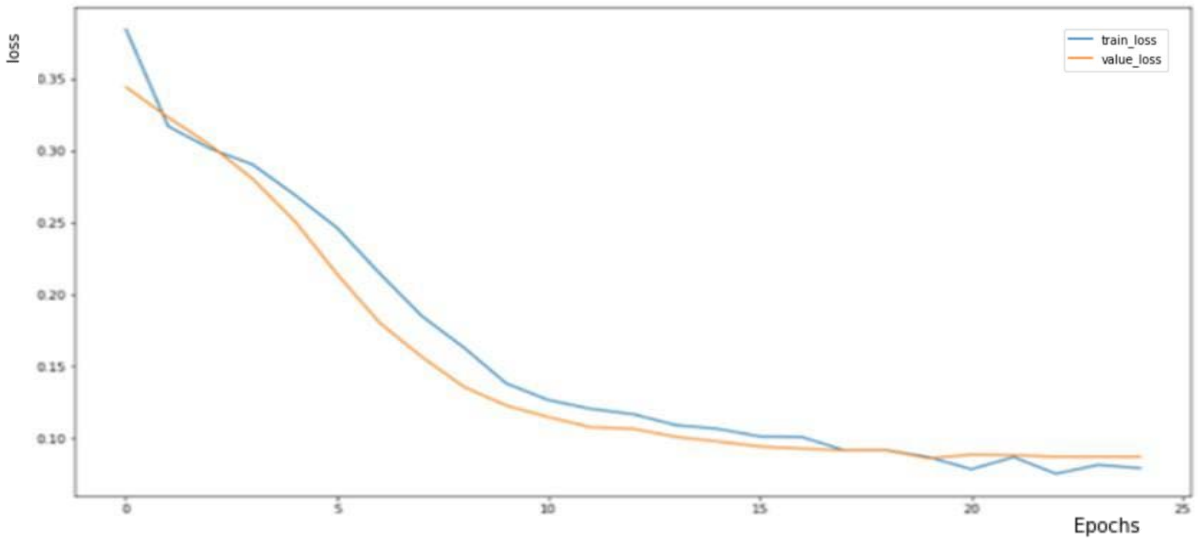

Figure 4.4 Model Loss vs Epochs.

As may be seen in the diagram, The value loss is shown by the yellow line, while the training loss is represented by the blue line. The experimental outcomes section discusses the final assessment results.

4.6 Optimizers

Deep convolutional neural networks have been built to perform complex tasks in vision, speech, natural language processing, clustering, classification, and other areas. In order to create predictions, we may expand the power of neural networks to tackle challenges that are less observable and more connected to mathematics and statistics. Optimizers are methods or strategies for adjusting the characteristics of a neural network, such as parameters and learning rate, to reduce losses. We employed two optimizers in this experiment, which are detailed below:

4.6.1 Stohastic Gradient Descent Optimizer

A version is Gradient Descent. It tries to keep the model's characteristics updated more often. After each training example's loss computation, the model's parameters are modified. As a consequence, if the dataset has 1000 rows, SGD will modify the system parameters 1000 times in a single dataset loop, rather than once like Gradient Descent. By increasing error gradient by a proportionate gain coefficient and then adding that value to previous rates, SGD increases network weights.[10]

$$\theta = \theta - \eta.\nabla_\theta J(\theta)$$

… (2)

In this equation, θ is parameters vector, η is the rate of learning coefficient estimating the number of steps required to attain the local minima, where J(θ) is the objective function and ∇ θ J(θ)is the slope of the objective function defined by θ. The SGD could be further defined as

$$\theta = \theta - \eta.\nabla_\theta J(\theta, x^{(i)}, y^{(i)})$$

… (3)

(x(i), y(i)) might be gleaned from the practice data Because the SGD approach converges faster than batch-based gradient descent, each update is made with relation to a large number of training samples n or a minibatch instead of one training example. The length of the mini - batch varies per application, although it is usually between 50 and 256 bytes.[11]

4.6.2 Adam's Optimizer

Adam works with the first and 2nd order momentums (Adaptive Moment Estimation). Adam's intuition appears to be that we shouldn't move so quickly only to get past the bare minimum; instead, we should slow down a bit and look more carefully. Adam, like AdaDelta, keeps an exponential decay average of prior squared gradients while also keeping an exponential function average.

Adam maintains track of both the declining average of prior gradients and squared gradients, as specified in the equations.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t^1$$

… (4)

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

… (5)

where mt denotes for both the mean vector estimate, v_t for the gradients' uncentered variance, and β1 and β2 for the decay rates.[11]

CHAPTER 5: EXPERIMENTAL RESULTS

5.1 Results

To anticipate loan defaults, the Neural Network Model with SGD Optimizer as the optimization function is utilized. The data is used to assess the efficacy of the proposed neural network model. The neural network model's accuracy rate within the normal cutoff hit 98 percent, as seen below. It is also possible to understand the confusion matrix.

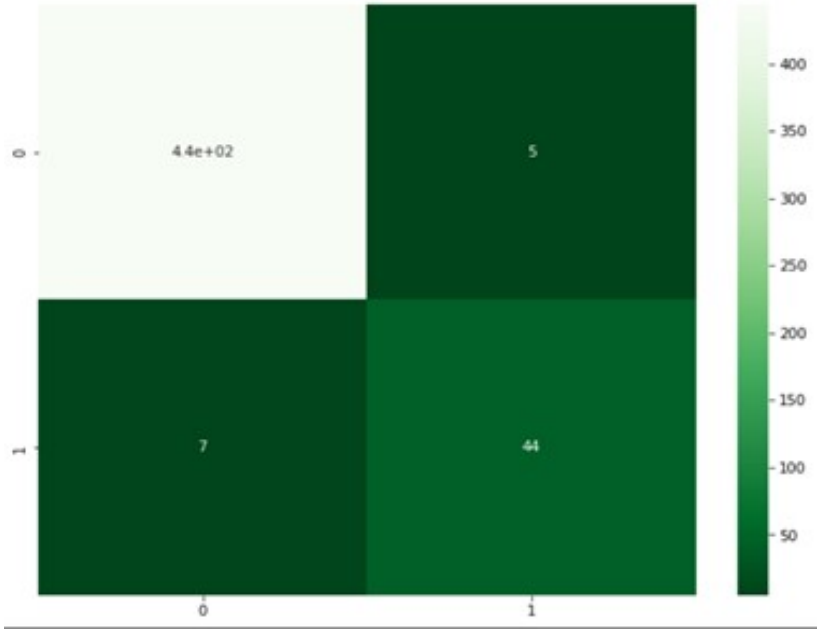

Figure 5.1 Confusion Matrix for the given model

The accuracy of recall, precision, f1 score could be interpreted as below:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	449
1	0.90	0.86	0.88	51
accuracy			0.98	500
macro avg	0.94	0.93	0.93	500
weighted avg	0.98	0.98	0.98	500

Fig 5.2: Accuracy for the Model

5.2 Summary of Results Obtained

Model	Precision	Recall	F1-Score	Accuracy
NNwSGD	**94**	**93**	**93**	**98**

Table 5.1: Result Obtained

Graph Plot for the Model

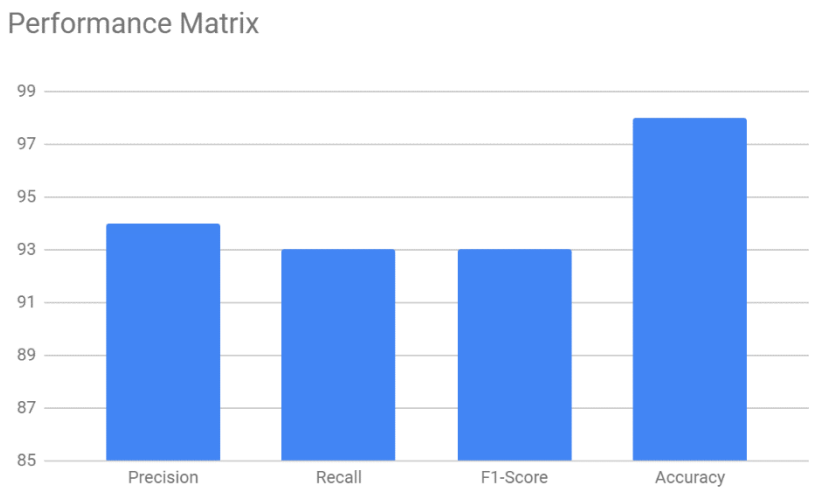

Fig 5.3 Graph Plot of Performance Matrix

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

The study analysis and information cleansing and processing, attribution of missing features, exploratory investigation of informational collecting, and finally model structure to model assessment and testing on test data are all steps in the process. It was also discovered that applicants with very high wages and asks for a smaller advance are much more likely to be granted and keep their credit. We looked into loan prediction using a neural network, which is one of the many ML and DL methods available. Increase the number of hidden layers and even the number of convolutional within a certain threshold value to increase the accuracy and precision of this developed model. The model's accuracy was estimated to be approximately 98 percent.

REFERENCES

[1] Sueyoshi, T. (1999). DEA–discriminant analysis in the view of goal programming. European Journal of Operational Research, 115, 564–582.

[2] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, 24, 433–441.

[3] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics and Data Analysis, 50, 111.

[4] Noh, P. J., Rohb, T. H., \& Hana, I. (2005). Prognostic personal credit risk model considering censored information. Expert Systems with Applications, 28, 753–762.

[5] Sheikh MA, Goel AK, Kumar T., "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020.

[6] Vaidya A., "Predictive and probabilistic approach using logistic regression: application to prediction of loan approval", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2017

[7] TejaswiniJ, Kavya TM, Ramya RD, Triveni PS, Maddumala VR, "Accurate Loan Approval Prediction Based On Machine Learning Approach", Journal of Engineering Science. 2020.

[8] Zhang H, Li Z, Shahriar H, Tao L, Bhattacharya P, Qian Y, "Improving prediction accuracy for logistic regression on imbalanced datasets", IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) 2019

[9] A. Goyal and R. Kaur, “A survey on Ensemble Model for Loan Prediction”, International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[10] Sebastian Ruder, An overview of gradient descent optimization algorithms, 2017, pp. 1-14; https://arxiv.org/pdf/1609.04747.pdf.

[11] B. Keegan, "Using First-Order Stochastic Based Optimizers in Solving Regression Models," 2018 IEEE MIT Undergraduate Research Technology Conference (URTC), 2018, pp. 1-4, doi: 10.1109/URTC45901.2018.9244791.

[12] Wang, SC. (2003). Artificial Neural Network. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science, vol 743. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0377-4_5.

[13] M. Bayraktar, M. S. Aktaş, O. Kalıpsız, O. Susuz and S. Bayracı, "Credit risk analysis with classification Restricted Boltzmann Machine," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.doi: 10.1109/SIU.2018.8404397.

[14] Y Y. Shi and P. Song, "Improvement Research on the Project Loan Evaluation of Commercial Bank Based on the Risk Analysis," 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2017, pp. 3-6.doi: 10.1109/ISCID.2017.60.

[15] Raj, J. S., Ananthi, J. V., “Recurrent neural networks and nonlinear prediction in support vector machine” Journal of Soft Computing Paradigm (JSCP),1(01), 33-40, 2019.

[16] Z. Huang, H. Chen, C. Hsu, W. Chen and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," Decis. Support Syst., vol. 37, pp. 543-558, 2004

Short Description of Work done in Paper 1 (Survey):-

Abstract— The number of persons seeking loans in India has increased significantly, and the reasons for this might be several. Employees in the banking industry lack the information to assess or predict whether a client (good or bad) would be able to repay the loan obligation at the agreed-upon interest rate. Banking institutions provide a number of services across the financial system, but lines of credit remain their primary and largest source of revenue. As a result, the money generated by the mortgages they provide will benefit banking enterprises. The financial statement of a banking organization is affected by lending, or whether customers repay money or fail on their loans. Estimating mortgages will reduce the non-performing investments of financial organizations. As a result, greater investigation of the current incident is required. Because precise estimates are necessary for adequate service, numerous approaches must be evaluated and studied. Our research and study aims to present a complete overview of loan estimating systems and structures that use prediction methods and procedures that have grown and evolved in recent years. Researchers looked at learning approaches as well as the raw datasets used for training and testing in this study and article.

The accuracy of the system model is also examined. Our research also gives a concise rundown of a few datasets that may be utilized to predict loan/mortgage analyses. Trends from the past and future are also highlighted.

Results and Discussion

Taking into account the previous years' research that was examined, as well as all of the observations that were made. We may particularly state that we have a good understanding of the studies conducted in recent years. The authors initially attempted an exploratory analysis of the data in order to change the prescribed data and arrive at a relevant conclusion. After acquiring the data, we attempted to use algorithms such as Logistic Regression, SVM, J48, KNN, and Tree Model, all of which offered high accuracy when compared to the training data. Furthermore, logistic regression may only be used to estimate a small number of variables and needs some of the variables to be independent; otherwise, it will favour the dependent variables. The prediction error and accuracy might be increased by utilising an ANN model since it is more graphical and has numerous layers of nodes, resulting in a more complex structure and a better prediction model. SVM and LR have limitations, and alternative models such as decision trees and Bayesian classifiers might be used to improve prediction.

Ensemble techniques might be used to a wide range of data sets. The accuracy and error prediction of the ANN model might be enhanced by doubling the number of hidden layers and perceptron to a threshold value.

Publication Details:

- Ankit Sharma, Vinod Kumar, "An exploratory study-based analysis on loan prediction" . Accepted at the **International Conference on Inventive Communication and Computational Technologies (ICICCT 2022)**

  Accepted in Springer Lecture Notes in Networks and Systems. ISSN : 2367-3370

  INDEXED BY – **Scopus**, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, WTI Frankfurt eG, zbMATH

Short Description of Work done in Paper 2 (Implementation): -

Abstract— With the advancement in the financial sector, more people are seeking for bank loans. However, banks have limited resources that they must allocate to certain persons, analyzing and evaluating as to who would be a potential risk to the bank and who would not be, determining the credit to be given to the risk-free consumer. So, this research is an attempting to reduce the risk element associated with selecting the protected individual in order to save a large number of bank asset, endeavors and resources. It's done by trawling through previous records of those who have received advances previously granted, and the machine was constructed based on these records using the AI model that provides the most exact result. This research deals with the motive whether or not it is safe to lend a loan to a consumer keeping in mind the amount should be returned on time and the credit is safeguarded or not. It is one of the most pressing and significant factors related to the financial institutions and equivalent businesses, since it has a considerable influence on their net revenue and profitability. The presence of multi non-performing mortgages has risen considerably in recent years, jeopardizing the growth of these Banking Institutions. We present a method for implementing a neural nets model that will be used to forecast and predict loan mortgage default. The projection is made by considering the personal and monetary information given by the probable loan taker. The data-set which is used to train and evaluate the suggested neural network model. Our suggested neural network model surpasses alternative classifiers that are usually employed by monetary firms for loan default forecasting, based on the findings obtained.

Results and Discussion

The Neural Network Model with SGD Optimizer as the optimization function is used to forecast loan defaults. The efficacy of the proposed neural network model is evaluated using the data. The accuracy rate of the neural network model inside the usual cut off reached 98 percent as shown below. The confusion matrix can also be interpreted. The accuracy of recall, precision, f1 score could be interpreted as below:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	449
1	0.90	0.86	0.88	51
accuracy			0.98	500
macro avg	0.94	0.93	0.93	500
weighted avg	0.98	0.98	0.98	500

Publication Details:

- Ankit Sharma, Vinod Kumar, "Predicting Loan Grant using Feed Forward Neural Nets and SGD optimizer". Accepted at the **International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–22)**

  Accepted in IEEE Explore.

  INDEXED BY – **Scopus**, **IEEE and Google Scholar.**

PAPER NAME

Modified-Thesis-Final.pdf

WORD COUNT

6996 Words

CHARACTER COUNT

36811 Characters

PAGE COUNT

38 Pages

FILE SIZE

575.1KB

SUBMISSION DATE

May 23, 2022 4:10 PM GMT+5:30

REPORT DATE

May 23, 2022 4:11 PM GMT+5:30

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 6% Internet database
- 0% Publications database
- Crossref Posted Content database
- 15% Submitted Works database

Excluded from Similarity Report

- Crossref database
- Bibliographic material

Acceptance Of Paper 1

6th International Conference on
Inventive Communication and Computational Technologies
(ICICCT 2022)

12-13 May, 2022 | icicct.org/2022 | icicct.conf@gmail.com

Acceptance Letter

To: **Ankit Sharma and Vinod Kumar**

Paper id: **ICICCT049**

Title: **An Exploratory Study Based Analysis on Loan Prediction**

Dear Author,

With the heartiest congratulations, we are happy to inform you that based on double blind review process and the recommendations of the conference review committee, your paper mentioned above has been accepted for publication and oral presentation at the 6th International Conference on Intelligent Computing and Communication Technologies [ICICCT 2022].

ICICCT 2022 is a Springer approved conference and all the registered papers will be recommended for publication in springer "Lecture Notes in Networks and Systems". ICICCT 2022 gives due recognition to great achievements of students, researchers and industrialists in the promotion and effective utilization of their research works. Herewith, the conference committee sincerely invites you for oral presentation at ICICCT 2022 to be held in Namakkal, India, 12-13 May, 2022. For more information on the conference kindly visit ICICCT 2022.

Yours' Sincerely

Dr.G.Ranganathan
Conference Chair ICICCT 2022

Proceedings by

Source details

Lecture Notes in Networks and Systems

Scopus coverage years: from 2016 to Present

Publisher: Springer Nature

ISSN: 2367-3370 E-ISSN: 2367-3389

Subject area: Engineering: Control and Systems Engineering | Computer Science: Signal Processing | Computer Science: Computer Networks and Communications

Source type: Book Series

View all documents > | Set document alert | Save to source list | Source Homepage

Acceptance Of Paper 2

Microsoft CMT <email@msr-cmt.org> Mon, May 16, 9:43 AM (8 days ago)
to me

Dear Ankit Sharma,
Delhi Technological University

Greetings from ICAC3N-22 ...!!!

Congratulations....!!!!!

On behalf of the 4th ICAC3N-22 Program Committee, we are delighted to inform you that the submission of "Paper ID- 718 " titled " Predicting Loan Grant using Feed Forward Neural Nets and SGD optimizer " has been accepted for presentation at the ICAC3N- 22 and will be sent for the submission in the conference proceedings to be published by the IEEE.

Please complete your registration by clicking on the following Link: https://forms.gle/8acy23i3UbtwLkFXA on or before 20 May 2022.

Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Minimum paper length should be 5 pages.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences.
5. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
6. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp/ UPI-Paytm/PhonePay) or write us at icac3n.22@gmail.com.

Regards:
Organizing committee
ICAC3N – 22