UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SCHOOL OF INFORMATION TECHNOLOGY

STUDENT NAME: Phathutshedzo Tiny Mudau
STUDENT NO: 22783352
EMAIL: u22783352@tuks.co.za

MODULE CODE: COS801
MODULE NAME: Big Data
ASSIGNMENT NO: 01
DATE: 29 October 2022

Contents

Introduction

Credit scoring is a method for determining how likely it is that a counterparty, current borrower, or loan applicant would default or go past due. For consumer loans, credit card use, and mortgage lending, it provides a forecast of the chance of default or delinquent. Credit scoring is a tool that lenders use to decide whether or not to grant borrowers' requests for credit. Condensing all of the information that is known about the borrower into a score is the goal of any credit scoring procedure. Credit is only granted if it is determined that the borrower's credit score is higher than a predetermined cutoff. Application scoring and behavioral scoring are both used to assess new applicants and monitor current borrowers to determine whether their creditworthiness has changed. In order to assess both applications and behavioral data efficiently, lending organizations should have access to a significant sample of past customers as well as information on their applications and subsequent credit histories.

The most extensively studied risk management techniques in credit scoring literature over the years have been statistical and classical machine learning models; more recently, deep learning models have come into prominence. This change is the result of deep learning models' improved results across a variety of domains. Despite improved performances, it is still necessary to describe how deep learning models generate their predictions. An explanation, as described by Liu et al. [3], is the capacity to present relationships between input information and output predictions visually or textually.

Deep learning algorithms' central idea is automating the extraction of representations from the data. Massive amounts of supervised data are used by deep learning algorithms to automatically extract complicated representation. Statistical and conventional machine learning models, as well as deep learning models, have been the most thoroughly investigated risk management strategies in the literature on credit scoring over time. For high stakes judgments like those involving the criminal justice system, healthcare, and credit scoring, Rudin (C. Rudin, 2019) offers an opposing perspective on XAI and emphasizes the benefits of using interpretable models over explainable black box models. (Chari et al., 2022) created a taxonomy of explanations from the literature and compiled it in their review study. Nine different explanation kinds are included in this taxonomy, including contextual, case-based, contrastive, statistical, scientific, trace-based, practical, and simulation-based explanations. It was intended to aid in the development of explanations that fit the requirements of the circumstantial evidence. They used both categorical and continuous features to systematically discretize the data into

the best possible categories based on the weights of the evidence (X. Dastile and T. Celik, 2019). In this project we are going to replicate thesis done by (Le Quy Tai and Giang Thi Thu Huyen, 2019) who did models of deep neural network (DNN) and Convolutional neural network (CNN) in 1D. We are going to extend the paper by applying other deep neural networks such as Long Short-Term Memory neural network.

Problem statement

Fintech's advancements have made it possible for the financial services sector to conduct business in new ways. Financial institutions that lend money are currently under pressure to make their products more accessible to the market, to conduct credit underwriting that is more open and non-discriminatory, and to process loan applications more quickly than before. As a result, the processes for determining whether to provide credit or not need to be improved. The future of machine learning models is becoming more widely discussed due to ongoing debates by the European Banking Authority (EBA). The question we still have to answer is that, of the models that have been used and deep neural networks, which of them give the best results between good and bad borrowers? The aim of this current study is to explain predictions made by some of the commonly used methods and deep learning model in a credit scoring setting.

Approaches/Model/Methods/Algorithms description

For categorization, we'll utilize ANN. These independent variables are used by the business to assess client risk. We're going to prepare the dataset and then use Keras to construct an artificial neural network. ANN will then be trained via compilation and fitting to the training set. The accuracy of the test set prediction is 70%. We moreover forecast the outcomes of the test set; if y pred is greater than or equal to 0.5, then the new y pred will turn into 0. (True). Additionally, if y pred is greater than 0.5, new y pred will turn into 1. (False). Last but not least, we created a confusion matrix and obtained 70%. We use the sequential neural network model of deep neural networks (DNN) in this study. We employ a unique deep learning algorithm application called the attention mechanism LSTM to present a consumer credit scoring system. Bidirectional LSTM are an extension of conventional LSTM that can enhance model performance on sequence classification issues, claim Schuster and Paliwal. The LSTM architecture must be altered to allow for the usage of other customer data in addition to transactional data in order to be used in the behavioural scoring task. The architecture of a neural network

is typically chosen in relation to the training data. In order to create an effective model with a small number of parameters, it is crucial to leverage the spatial organization and order of the input data. RNN and LSTM neural networks are typically employed for temporal input. The LSTM network, however, is not appropriate for combining temporal and non-temporal data. One alternative is to inject non-temporal data into the thick layers on top of the LSTM, however in this case, non-temporal features are only employed in the final stages of the model. With the exception of the Support Vector Machine on the German credit dataset, we will also employ standard techniques like Decision Trees, k-Nearest Neighbor, Naive Bayes, Multi-Layer Perceptrons, and Random Forest. As a result of the library's ability to address the imbalance in the label distribution, we chose to train ANN in Keras. In our trained models, we took into account the ensuing factors. The training set refers to the portion of the data that was divided into two and used for training. The test set refers to the additional portion that was set aside for model evaluation. The amount of data we had was taken into account when deciding how to split training and testing. A significant balance between parameter estimate variance and performance statistic variance was the goal of the split. We utilized a split module for train and test. We also going to implement CNN. We construct a model. The model has two hidden layers, two MaxPooling layers, and two Convolution layers. The flatten function is mostly used to combine the previous layer's 3-D array into a single layer (Because the Ann model only take 1-D array as input). Therefore, flattening produces the input layer. We have now fitted the model to the data. By increasing the number of epochs, the Conv layer, and the Maxpool layers, you can improve accuracy. (In this instance. However, using an image data generator is the ideal option if you're working with images. This generates a large variety of images for the same label using real-time data augmentation. The principal benefit is a decrease in data volume.

Experimental description of the dataset and results

We use four real world datasets namely German Credit data, Australian Credit Approval, and HMEQ in credit scoring, to take the experiments with our deep neural networks. German Credit dataset was obtained from UCI Machine Learning repository, and Australian Credit Approval, and HMEQ datasets was obtained from Kaggle. This has been widely used in validating credit and behavioural scoring models, also in deep learning models. The German dataset has 7 features that are numerical and 13 that are categorical. Dataset have 700 cases of good credits and 300 cases of bad credits. The HMEQ dataset has 11 features that are numeric and two that are categorical. The target vari-

able for the credit scoring datasets mentioned above is binary, the applicants classified either bad or good which are denoted by 0 and 1 respectively. Accuracy rate: For each sample, the ratio of correct samples to total samples is anticipated; the mathematical expression is as follows. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Precision: The percentage of samples that are genuinely in the positive class and the prediction result that is likewise in the positive class number is true for all samples with positive class prediction results; the mathematical formula is as follows:

$Precision = \frac{TP}{TP+FN}$

Recall rate: In all samples that are in fact positive, the fraction of samples that are expected to be positive is expressed mathematically as follows:

$Recall = \frac{TP}{TP+FN}$

Table 1: German credit scoring performance

	kNN	CART	NB	SVM	ANN	LSTM	CNN
Accuracy	66%	71%	72.5%	71%	73.6%	72.5%	87.5%
Precision	72.6%	50%	71.4%	71.4	79.9%	77.7%	75%
Recall	83.8 %	71 %	66.8%	98.6%	87.2%	85.9%	25%
F1	59 %	77.8 %	80.4 %	82.8%	77.8%	80.7%	75%

Table 2: HMEQ credit scoring performance

	kNN	CART	NB	SVM	ANN	LSTM	CNN
Accuracy	82.13%	85.5%	88.25%	80.1%	75.5%	72.5%	93.18%
Precision	64%	77%	77.5%	100%	79.9%	77.7%	87.5%
Recall	23.9%	78%	58%	0.4%	87.2%	85.9%	37.2%
F1	34.9%	77%	66.3%	0.8%	83.4%	81.6%	32.5%

Table 3: Australia credit scoring performance

	kNN	CART	NB	SVM	ANN	LSTM	CNN
Accuracy	71.1%	52.8%	80.4%	71.1%	52.9%	72.5%	87.5%
Precision	68.1%	28%	95.2 %	75	82 %	77.7%	25%
Recall	53	75.4%	61.5%	60%	76.9%	85.9%	75%
F1	37 %	71.5%	74.8%	68.1%	79.4%	81.6%	37%

On the German credit data set, Table I shows how classifiers performed. With an accuracy measurement of almost 93%, ANN and CNN display their impressive capabilities. ANN has a classification accuracy of 87.2% when measured by F1 score. Contrary to popular belief, ANN performance outperforms SVM in terms of F1 score, despite SVM having the highest score on the Recall measurement. The performance of the ANN and CNN algorithms on an experiment with Australian credit data is excellent; Table II provides more specific results. The functionality of our approaches on the HMEQ data set is described in Table III. Every classifier performs well, scoring around 93.18% on the accuracy measurement. The greatest F1 score of our approaches, however, is 81%, indicating that they are sensitive to imbalance data.

Conclusion/Discussion/Future work

In this research, three Deep Neural Network designs have been modified to address the credit rating issue. Although they exhibit a number of poor outcomes in data sets with unbalanced classes, deep neural networks exhibit high performance in comparison to other classifiers in the majority of studies. We think deep neural networks have a lot of potential for credit rating. The performance of deep learning methods for credit scoring can be improved in a number of ways, including by applying 2D CNN with max-pooling stack to a sequence model and turning tabular datasets into pictures.

Reference

C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

S. Chari, D. M. Gruen, O. Seneviratne, and D. L.Mcguinness, Directions for Explainable Knowledge-Enabled Systems, 2020.

X. Dastile and T. Celik, Model-Agnotic counterfactua, explanations in credit score, open access journal, 2022.

L.Q. Tai and G.T Huyen, Deep learning techniques for credit scoring, *Journal of Economics, Business and Management*, 7(3), 2019.

DECLARATION OF ORIGINALITY / DECLARATION ON PLAGIARISM

The Department of Computer science (University of Pretoria) places great emphasis upon integrity and ethical conduct in the preparation of all written work submitted for academic evaluation. While academic staff teaches you about referencing techniques and how to avoid plagiarism, you too have a responsibility in this regard. If you are at any stage uncertain as to what is required, you should speak to your lecturer before any written work is submitted. You are guilty of plagiarism if you copy something from another authorâs work (e.g. a book, an article or a website) without acknowledging the source and pass it off as your own. In effect you are stealing something that belongs to someone else. This is not only the case when you copy work word-for-word (verbatim), but also when you submit someone elseâs work in a slightly altered form (paraphrase) or use a line of argument without acknowledging it. You are not allowed to use work previously produced by another student. You are also not allowed to let anybody copy your work with the intention of passing if off as his/her work. Students who commit plagiarism will not be given any credit for plagiarised work. The matter may also be referred to the Disciplinary Committee (Students) for a ruling. Plagiarism is regarded as a serious contravention of the Universityâs rules and can lead to expulsion from the University. The declaration which follows must accompany all written work submitted while you are a student of the Department of Geology (University of Pretoria). No written work will be accepted unless the declaration has been completed and attached. I, the undersigned, declare that: 1. I understand what plagiarism is and am aware of the Universityâs policy in this regard. 2. I declare that this assignment (e.g. essay, report, project, assignment, dissertation, thesis, etc) is my own original work. Where other peopleâs work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with Departmental requirements. 3. I have not used work previously produced by another student or any other person to hand in as my own. 4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work. Full names of student: Phathutshedzo Tiny Mudau Student number: 22783352 Date submitted: 2022/10/29 Topic of work: Project Signature: mudau pt Supervisor: Abi