Ala'raj *et al. J Big Data* (2021) 8:69
https://doi.org/10.1186/s40537-021-00461-7

Journal of Big Data

RESEARCH Open Access

Modelling customers credit card behaviour using bidirectional LSTM neural networks

Maher Ala'raj[1*], Maysam F. Abbod[2] and Munir Majdalawieh[1]

*Correspondence:
maher.alaraj@zu.ac.ae
[1] Department of Information Systems, College of Technological Innovation, Zayed University, 19282 Dubai, United Arab Emirates
Full list of author information is available at the end of the article

Abstract

With the rapid growth of consumer credit and the huge amount of financial data developing effective credit scoring models is very crucial. Researchers have developed complex credit scoring models using statistical and artificial intelligence (AI) techniques to help banks and financial institutions to support their financial decisions. Neural networks are considered as a mostly wide used technique in finance and business applications. Thus, the main aim of this paper is to help bank management in scoring credit card clients using machine learning by modelling and predicting the consumer behaviour with respect to two aspects: the probability of single and consecutive missed payments for credit card customers. The proposed model is based on the bidirectional Long-Short Term Memory (LSTM) model to give the probability of a missed payment during the next month for each customer. The model was trained on a real credit card dataset and the customer behavioural scores are analysed using classical measures such as accuracy, Area Under the Curve, Brier score, Kolmogorov–Smirnov test, and H-measure. Calibration analysis of the LSTM model scores showed that they can be considered as probabilities of missed payments. The LSTM model was compared to four traditional machine learning algorithms: support vector machine, random forest, multi-layer perceptron neural network, and logistic regression. Experimental results show that, compared with traditional methods, the consumer credit scoring method based on the LSTM neural network has significantly improved consumer credit scoring.

Keywords: Behavioural scoring, Neural networks, Bidirectional LSTM, Classification

Introduction

The case with many financial institutions such as banks is that credit lending products such as credit cards, personal loans and mortgages are the center of their dealings, and proper lending will yield huge gains. As a result, it is important for financial institutions and banks to get new customers and ensure to keep profitable ones. Banks have created a wide customer database over the years, which can be used to analyze a bank's performance and make progressive business decisions. It is not possible that all customers will act the same way when it comes to financial performance, therefore, there should be distinguishable treatment between customers who qualify for certain profitable requirements, based on their repayment and purchasing behaviour customers exhibiting such behaviour can be offered greater incentives and rewards [1]. Banks need to know their

good or bad customers, and they will need credit scoring and behavioural scoring to do so. Article [2] defined credit scoring as the means of analyzing the likelihood of applicant to falter in their repayments, or not. In Anderson [3] authors defined it by dividing the term into two parts: the first is 'credit', which means to buy an item and pay afterwards, and the second is 'scoring', which is alike with the method used for credit cards.

There are two major kinds of credit scoring, and they are application credit scoring, where a score is applied to provide a decision on a new credit application; and behavioural scoring, is where the score is used to address existing customers after they have been given a loan. Liu [4] banks use behavioural scoring to guide their decisions about lending in credit limit management strategies; managing debt collection and recovery; retaining future profitable customers; predicting accounts likely to close or settle early; offering new financial products and interest rates; managing dormant accounts; optimizing telemarketing operations; and predicting fraudulent activity [3, 5–8], the number of risk payment and the future risk of payment [9].

Furthermore, Lim and Sohn [10] have emphasized the benefits of having multifaceted models that predict when customers will fail to pay or repay debts, as follows: (1) Calculating the profitability over a customer's lifetime and doing profit scoring; (2) Making available to the bank an average of default levels over time, which is beneficial for debt provisioning; (3) Assisting in arriving at the terms of the loan; and (4) Adapting more to changing economic conditions. Banks usually try to estimate a borrower's credibility and give a safe probability when a customer may miss a payment generally, and subsequent payments particularly [11]. These models help the bank to take actions quickly against any risk that ends up in unfavorable behaviour by borrowers [12].

This paper focuses on behavioural scoring. According to Hsieh [13], behavioural scoring is utilized to examine the behaviour of existing customers, considering their attitudinal variables and estimate their payment behaviour or credit status. Behavioural scoring lets lenders to consistently monitor the changing behaviour or features of customers and help to direct customer level decision making.

Motivations

The primary origin of a credit card related risk for banks is client default, which is the inability to reimburse a debt on a loan or security. A default can happen when a borrower cannot make convenient payments, misses payments, or dodges or quits making payments. In the case of credit cards, no assets are securing the debt, but the lender still has legal recourse in the event of default. Credit card corporations regularly give few months before an account goes into default. However, if after 6 months or more there have been no instalments, the account will get feed off, meaning the lender takes a loss on the account [14]. Consecutive missed payments for credit card debt are an early sign of customer bankruptcy. Following the Basel II convention, consumer credit default is commonly defined as delinquency beyond a period of 90 days [15]. Therefore, the research of this paper is motivated by the necessity of automatically scoring the customer's behaviour on repayments to make risk decisions, and the use of credit card scores to make necessary financial security decisions. Using such scores, banks can classify customers into "risk groups", which could help to detect potential bankruptcy early and block the customer's card in time to limit losses.

Hence, the task of estimating the missed payment probability for clients who already have one or more missed payments turns out to be important for bank management.

The main drawback of existing automatic scoring solutions lies in the necessity for bank management to manually extract features from raw transactional data. This process is subjective and can lead to the loss of information in the data. On the other hand, LSTM extracts features internally and in the way which is hidden from outside observers.

The main aim of this paper is to help bank management in scoring credit card clients using machine learning techniques. The main contributions and objectives of this paper, based on the above motivations, are:

(1) Introduce a deep learning neural network architecture based on Long-Short Term Memory (LSTM) bidirectional neural networks as a method of customer behaviour score estimation.
(2) Prove the feasibility of LSTM model and test it on the real credit cards dataset by comparison with other classifiers.

The developed LSTM model is compared to four classical machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), Bagged Neural Network (NN), and Logistic Regression (LOGR). The paper discusses the importance of performing a detailed comparison procedure while proving high accuracy using LSTM model that best fulfils the users' interest.

The remainder of the paper is organized as follows: Section "Machine learning approaches in behavioural scoring" gives a preview of the relevant literature on machine learning models in credit and behavioural scoring. Section "Methodology" describes the proposed methodology that is used in this paper. Section "Experimental design" explains the experimental setup, whereas Section "Results and discussion" presents the experimental results and analysis. Finally, in Section "Conclusion", conclusions are drawn, and future work prospects are discussed.

Machine learning approaches in behavioural scoring

The field of credit scoring has become a broadly investigated subject by researchers and the financial industry [16], with numerous models having been proposed and created utilizing measurable methodologies, for example, LOGR [6] and Linear Discriminant Analysis (LDA) [17, 18]. Because of the financial crisis, the Basel Committee on Banking Supervision demanded all banks to apply thorough credit assessment models in their frameworks while conceding a loan to an individual customer or a company. Appropriately, research have shown that Artificial Intelligence (AI) procedures (e.g., neural networks, SVM, and RF) can be a decent exchange for measurable methodologies in building credit scoring models [19–21].

Behavioural scoring applies characteristics of customers' ongoing behaviour to predict whether they are prone to default during a specific outcome period. Often the outcome period and fixed performance period are subjectively selected, which causes instability in the prediction-making process.

Most papers in the literature were centred on behavioural scoring with respect to customer loans [22–24]. However, behavioural scoring of client's credit card payments has not been appropriately investigated. Behavioural scoring models support to analyse purchasing behaviour of existing customers [25]. Only a few works have studied the mining of bank databases from the viewpoint of customer behavioural scoring [26]. To alleviate this, Hsieh et al. [27] have used a Taiwanese bank credit card dataset to demonstrate the effectiveness of behavioural scoring. The authors use three commonly discussed data mining techniques: LDA, SVM, and Back Propagation Neural Networks (BPNN).

In recent years, loan and credit card transactions information has become significantly larger. Therefore, it is often difficult to use traditional mathematical and statistical models for such types of problems. To construct behavioural scoring models, professionals must think about a few significant issues, such as the extensiveness of the dataset to model, the planning horizon, and drivers of unwanted behaviour [6]. The literature does not contain solid suggestions on the most proficient method to respond to these questions.

One of the approaches is to use feature selection on features, generated from raw transactional data. Feature selection was used in credit scoring problems [28]. In general, feature selection is very important to use such as for knowledge discovery in databases (KDDs). Some of the applications are for Colorectal Cancer Cases Phenotype [29], breast cancer identification [30], household poverty [31], air pollution [32]. Meanwhile, the extended version of SVM-DHGLM increased the accuracy, precision, recall, for feature selection and classification [33].

Hence, this paper explores a portion of the issues influencing the structure of a behavioural scoring model using machine learning by investigating the performance of a large pool of credit card transactions dataset.

Pereira [34] examined the conduct of a credit card purchaser relying upon whether they do payments involving a tremendous measure of cash. In Alborzi and Khanbabaei [35], a new hybrid model of behavioural scoring and credit scoring based on data mining and neural network techniques is introduced for both banking and marketing purposes. A two-stage scoring approach with wide and deep learning usage suggested in Bastani et al. [36] is an integration of credit scoring and profit scoring. Stage 1 was designed to identify non-default loans, which were then moved to stage 2 for probability prediction, wide and deep learning were used to build the predictive models in both stages to achieve both memorization and generalization. In the study by Akkoç [37], the author has proposed a three-stage hybrid Adaptive Neuro Fuzzy Inference System credit scoring model, which is based on statistical and neuro-fuzzy techniques. Addo et al. [38] have built binary classifiers based on machine and deep learning models were built on real data to predict loan-default probability. In Gui [39], the author intends to apply multiple machine learning algorithms to analyse the default payment of credit cards. Based on the user operation behaviour data of the P2P lending industry, a consumer credit scoring method based on the attention mechanism LSTM was offered by Wang [21].

Considering the relevant literature and to the best of our knowledge, there are no studies which apply LSTM neural networks to the task of predicting consecutive missed payments and defaults for customers' credit cards. For example, in [21] an LSTM neural network was used, but the application differs from the field of this

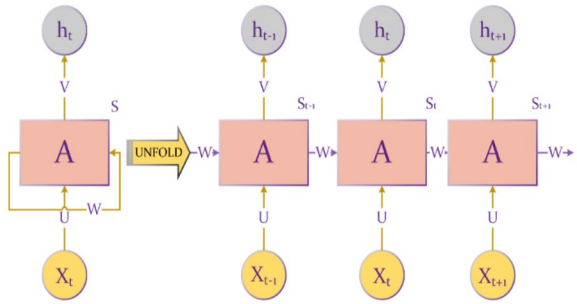

Fig. 1 RNN model structure [21]

research; in Heryadi and Warnars [40] and Graves et al. [41] various architectures of neural networks was used, but research topic was credit card fraud detection, which is different from ours. Also, we show that scores of the model can be treated as probabilities, which is significant fact.

This paper discovery is contributing to the literature of credit and behavioural scoring since as the application of LSTM neural networks to missed payment analysis with concurrent use of customer information has not be studied previously.

Methodology

Recureent and LSTM nueral networks

Recurrent neural networks (RNNs) are a special class of supervised machine learning models. They are made of a sequence of cells with hidden states which have non-linear dynamics. RNNs are used mostly with time series data, for example, speech recognition [42], unsupervised anomaly detection [43], and automated translation [44]. LSTM is also used in economics to forecast time series data as an alternative to the ARIMA model [45]. As transactional data in credit cards has a temporal nature, it is advisable to use RNNs instead of other types such as fully connected or convolutional neural networks.

In a recurrent neural network, connections between cells form directed cycles. Each cell contains a hidden state, which is updated on each iteration using its previous values. Such a structure creates an internal network state and works as a memory. The RNN equations are:

$$\begin{cases} s_t = f(U \cdot x_t + W \cdot s_{t-1}), \\ h_t = g(V \cdot s_t), \end{cases} \tag{1}$$

where x is an input vector, s is a hidden vector of RNN layer values, h is an output vector of RNN layer values, U is a weight matrix of the input layer to the hidden layer, V is a weight matrix of the hidden layer to the output layer, W is a weight matrix for the previous time point to the current time point of the hidden layer, and g and f are activation functions for output and hidden layers respectively. The structure of a standard RNN model is shown in Fig. 1.

In Fig. 1, the work of one RNN cell is illustrated. We feed time series signal X to the cell element by element. The vector X can be an input vector or output from other

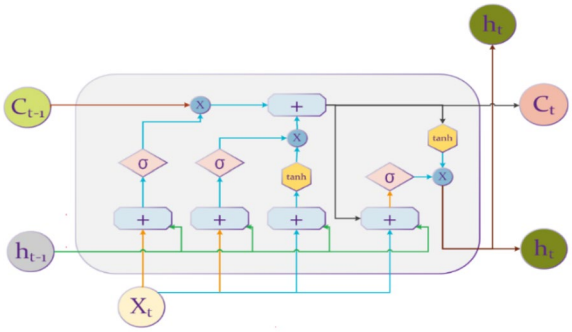

Fig. 2 LSTM model structure [21]

RNN cell from the previous layer. The RNN cell holds its state s. At each iteration t, the state s_t and output h_t are calculated by Eq. (1). Because of their architecture, RNNs can [21]:

(1) Recognize patterns, characteristics, and dependencies in sequential and time series data;
(2) Store, remember, and process past complex signals for long time periods;
(3) Map an input sequence to the output sequence at the current timestep and predict the sequence in the next timestep; and
(4) Replicate any target dynamics after the training process, even with adjusted accuracy.

However, there are issues with learning long-term dependencies. Because RNN is prone to vanishing gradients during training, it is difficult to learn long-term dependencies [46, 47]. To solve this problem, Hochreiter and Schmidhuber [48] have proposed an LSTM based on RNN. As with RNNs, LSTM predictions are always conditioned by the experience of the network's inputs. Its distinguishing feature is the existence of special units called memory blocks in the recurrent hidden layer, which perform like accumulators of the state information. Every memory block has memory cells with self-connections, which store the temporal network state, and special multiplicative units called gates, which can control the stream of information. These cells and gates allow the LSTM to trap the gradient in the cell (also known as constant error carousels) and prevent it from vanishing. The gate activation functions are sigmoid, thus output value ranges from 0 to 1, and denotes how much information can be allowed to pass outside. The structure of a single LSTM cell is shown in Fig. 2.

As seen in Fig. 2, an LSTM cell consists of three gates, namely an input gate, that controls how many cell states need to be stored an output gate that controls how many cell states are sent to the next cell have to, and a forget gate, that controls how much information needs to be removed [49, 50]. Two of these gates contain internal states. It can be seen that on each iteration t, the LSTM cell is using the previous values of the candidate vector C_{t-1} and output vector h_{t-1} to calculate their next values. The output of each gate is post-processed using activation functions. The shape of the activation function is important and can significantly affect the efficiency of the neural network [43].

By default, the activation function of the recurrent gates is a sigmoid function [48], which is a non-linear activation function that is used mostly in feedforward neural networks. It is a bounded monotonically increasing differentiable real function, defined for all real input values, as given by the following sigmoid function equation:

$$\sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{2}$$

The sigmoid function is applied to the output layers of the deep learning architectures in binary classification problems, modelling logistic regression tasks as well as other neural network domains. However, the sigmoid activation function suffers major drawbacks which include sharp damp gradients during back propagation from deeper hidden layers to the input layers, gradient saturation, slow convergence, and non-zero-centred output, thereby causing the gradient updates to propagate in different directions [28].

The hyperbolic tangent function is the default activation function for an LSTM cell's output gate [48]. The hyperbolic tangent function, tanh, is a smooth antisymmetric function with the range of values [− 1,1]. The output of the tanh function is given by:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3}$$

The main advantage provided by tanh is that it produces zero-centred output, thereby aiding the back-propagation process. The detailed procedure of an LSTM cell is explained as follows:

On the first step, LSTM should decide which information to forget. For this purpose, the information of the previous memory state is processed through the forget gate f_t:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{4}$$

On the second step, input gates i_t decide which information should be updated, and the tanh layer updates the candidate vector $\tilde{C}_t$:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{6}$$

On the next step, memory states C_t are updated as a combination of the two parts above:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{7}$$

Finally, output gates o_t are used for controlling the output h_t:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \times \tanh\left(C_t\right) \tag{9}$$

Therefore, each LSTM layer is characterized by [48]:

(1) Matrix W_f andb_f, vector, which are parameters of the forget gate;

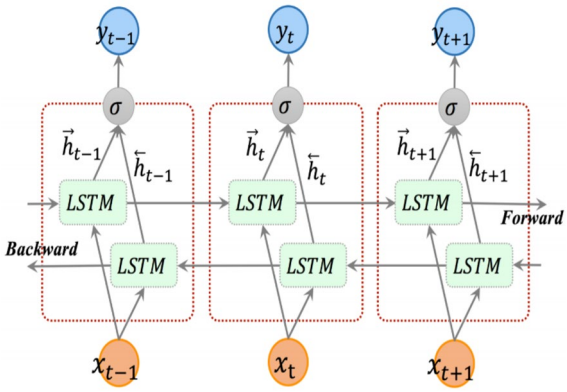

Fig. 3 Bidirectional LSTM architecture [53]

(2) matrix W_C and vector b_C, which are parameters of the input gate; and
(3) matrix W_o and b_o, vector, which are parameters of the output gate.

To increase the performance and learning speed of LSTM neural networks, in the research [51] bidirectional LSTM neural networks were proposed. According to Schuster and Paliwal [51], bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. In problems where all time steps of the input sequence are available, bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

According to Fig. 3, the forward layer output sequence, h, is iteratively calculated using inputs in a positive sequence from time $t=0$ to time $t=T$, while the backward layer output sequence, $\overleftarrow{h}$, is calculated using the reversed inputs from time $t=T$ to $t=0$. Both the forward and backward layer outputs are calculated by using the standard LSTM updating equations, Eqs. (2–7). The Bidirectional LSTM layer generates an output vector, Y_t, in which each element is calculated by using the following equation:

$$y_t = \omega\left(\overrightarrow{h}_t, \overleftarrow{h}_t\right), \tag{10}$$

One more extension of stacked LSTM neural networks is the "Attention" mechanism. The Attention Mechanism in the deep learning model is a model that simulates the attention of the human brain. When people observe images, they do not carefully look at every pixel of the image. Instead, they focus their attention selectively on some important parts of the image, ignoring other unimportant parts. Initially, attention mechanism was developed for automatic translation challenges [52], but then its usage was enhanced to image recognition and classification problems.

Proposed model

Even though the LSTM neural network principles are already well studied, choosing the architecture is often up to the researcher [21, 45, 53]. This includes choosing the number and type of layers, number of cells in each layer, activation functions, etc. In order to use the LSTM architecture in the behavioural scoring task, it must be modified to make it possible to use not only transactional data but also other customer data (age, salary, country of origin, etc.).

Usually neural network architecture is chosen with respect to data used for training. That's why it is important to use spatial structure and order of input data to make it possible to build efficient model with low number of parameters (weights). For temporal input usually RNN's and LSTM's neural networks are used. However, for mixed temporal and non-temporal data LSTM network is not applicable. One solution is to feed non-temporal data into dense layers at the top of LSTM, but in this case non-temporal features are used only in final stage of model.

Attention layer require optional query input which is used as a context of temporal input. We use non-temporal data as a query input to this layer to add a context of customer good or bad payment behaviour. Hence, such layer is able to distinguish financial behaviour of customers with taking into account their educational and marital status, as well as gender and age.

As it is seen from Fig. 4, the first two layers are bidirectional LSTM, next layer is Attention. The two last layers are the concatenation of output of Attention layer and the non-temporal client data. The last layer consists of only one neuron.

Table 1 shows the hyperparameters for the developed models. As it can be seen, the model for monthly purchase estimation is more complex than the one for missed payment prediction. This can be explained by the fact that, in general, regression problems are more complex than the classification ones. Number of neurons in each layer was selected using grid search, activation functions were selected by adopting the most used from similar research [21, 40, 53].

Time window parameter is important, but it belongs to the input data rather than model, so it will be defined in Section "Data description".

Experimental design

The aim of the LSTM model is to automate credit card behaviour scoring for customers as well as to trigger an early alert for credit card default. The framework of the proposed model is presented in Fig. 5. The workflow presented will let us fully investigate the model performance to make reliable conclusions.

The proposed framework consists of several steps. Firstly, the dataset is pre-processed and formatted to be used by Bidirectional LSTM classifier. As a next step, fivefold validation technique is used to get prediction for all customers in dataset. Then the performance measures are calculated for different groups of customers which is of financial interest to the bank institutions (banks are especially interested in customers with unsatisfactory history of payments). To outline performance of the model it is compared to benchmark models using various performance measures. Results are discussed in the final section.

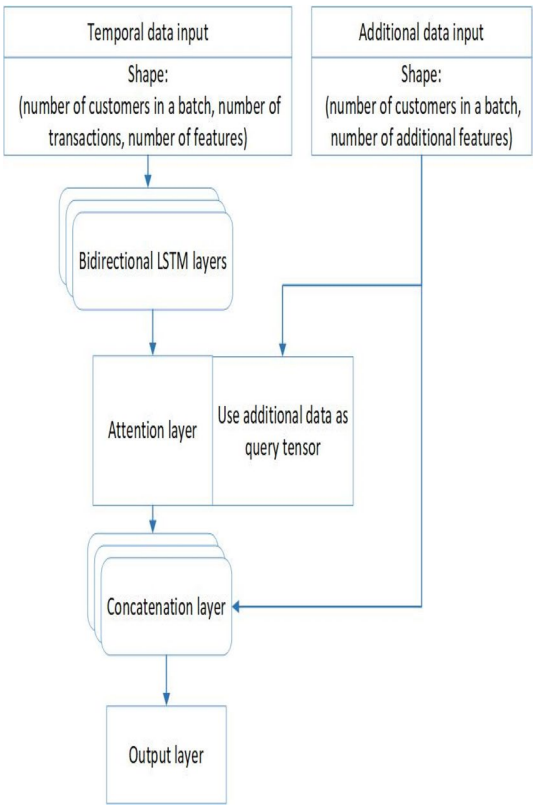

Fig. 4 The data processing flow in LSTM Neural Network model

Table 1 Hyperparameters for the developed models

Model task	Parameter value
Number of transactional features	3
Number of additional features	8
Number of bidirectional LSTM layers	2
Number of cells in each LSTM layer	4
Number of attention layers	1
Activation function of hidden LSTM layers	Hyperbolic tangent
Number of fully connected layers (dense)	1
Activation function for all dense layers except the output one	Sigmoid
Activation function of the output layer	Sigmoid
Loss function	Binary cross-entropy
Optimizer	Adam

Dataset

There are only few open source transactional datasets that can be used to test efficiency of proposed model. Majority of datasets are either non-temporal or they are from different field of research. To verify the practicality and effectiveness of the proposed LSTM

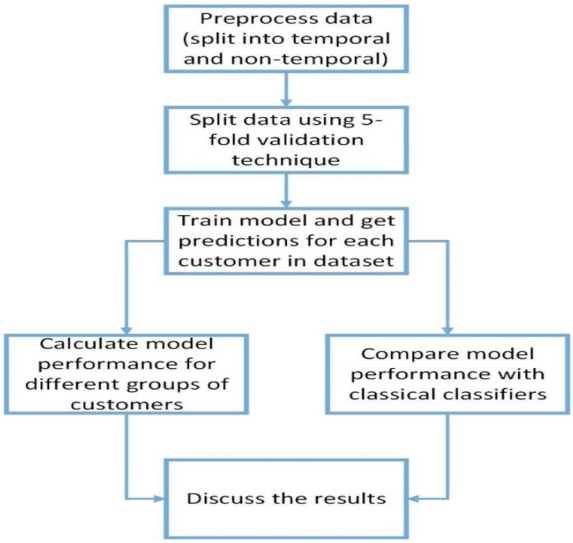

Fig. 5 Model framework

model we use a public[1] real credit cards dataset used in Bahdanau et al. [52] and can be easily converted to temporal form.

Dataset description

The dataset used in this paper is a public non-transactional credit cards dataset that reflects customer's default payments in Taiwan [54]. It has been widely used in validating credit and behavioural scoring models [55–57], also in deep learning models [58, 59]. Usually, banks do not disclose transactional databases in raw form, and thus majority of datasets in the open access are in processed form. Hence, we used this dataset because this is the only publicly available dataset which can be converted into temporal form (customer payment statistics for each month rather than aggregated values).

The size of the data set is 30,000 records, which is large enough to test the efficiency of the proposed model. The number of non-default payments is 23,364, while the number of default payments is 6636 (proportion of default payments in dataset is 22%). There were no missing values in dataset.

In the dataset the following 23 variables are used as explanatory:

(1) X1: Amount of the given credit, which includes both the individual consumer credit and his/her family (supplementary) credit.
(2) X2: Gender (1 = male; 2 = female).
(3) X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
(4) X4: Marital status (1 = married; 2 = single; 3 = others).
(5) X5: Age (year).
(6) X6–X11: History of past payment. Tracked payment records are denoted from September to April 2005 by X6–X11, respectively. The measurement scale for

[1] The dataset is available at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

the repayment status is: − 1 = pay duly; 1 = payment delay for 1 month; 2 = payment delay for 2 months; ...; 8 = payment delay for 8 months; 9 = payment delay for 9 months and above.

(7) X12–X17: Amount of bill statement. The amount of bill statement is denoted from September to April 2005 by X12–X17, respectively.
(8) Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005; ...; X23 = amount paid in April 2005.

The variables can be divided into two groups: numerical and categorical. The examples of the first are: X1 (amount of given credits), X5 (age), X6–X11 (history of past payment), etc. The second group contains such variables: X2 (gender), X3 (education), X4 (marital status).

Dataset pre-processing and partitioning

Before feeding into a neural network, it was split into two parts: temporal data and non-temporal data. Columns X6–X23 as temporal data that reflect customer behaviour in time were reshaped into a three-dimensional array of shape (number of customers, number of months, number of features). According to the data set description, for each customer we have information about his payment behaviour during 6 previous months. Therefore, the second dimension of the array is equal to six. The number of temporal features available for each customer is equal to three, namely:

(1) Payment delay by the end of each past month;
(2) Amount of bill statement by the end of each past month;
(3) Amount of the payment in each month.

Non-temporal categorical data was split into binary, thus for each customer there are eight non-temporal features:

(1) Amount of given credit.
(2) Gender.
(3) Education—graduate school.
(4) Education—university.
(5) Education—high school.
(6) Education—others.
(7) Marital status.
(8) Age.

To properly test the performance of the model we use fivefold cross validation as partitioning technique. All customers were randomly split into five groups, and during each fold each group become testing once.

As it can be seen, most of the information in the dataset is stored in temporal features of past credit card activity and payments. On the other hand, non-temporal features are too general and are, in fact, categorical features. That is why without using temporal features it is impossible to predict future missed payment probability.

Attention mechanism is used to provide a context. Hence, age and gender provide such context for temporal financial information. It means that similar payment behaviour for young and old customers can lead to different payment outcomes (e.g., young customers can forget or skip to pay in some month and have bad payment history, but they would pay eventually).

Benchmark models development

To measure how well the proposed approach has performed, the results of the proposed model are compared to five benchmark models, namely, GB, BNN, RF, SVM and LOGR. The latter model is the industry standard for developing credit scoring models [60, 61]. However, [61] has stated that it is beneficial to compare a new method with the standard one as well as other established techniques. MLP, RF, and SVM have been used in several studies as a benchmark model [62]. The theoretical backgrounds of the models are described in the following sections.

Gradient boosting

Gradient Boosting (GB) machines are a group of powerful machine learning techniques that have demonstrated impressive accomplishment in a wide scope of practical applications. They are highly customizable to the particular needs of the application, like being learned with respect to different loss functions. The fundamental thought of boosting is to add new models to the ensemble consecutively. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the full ensemble learnt up to the last iteration [63].

Bagging neural network

Neural Networks (NN) are machine learning frameworks motivated by the scheme of the biological neuron [64]. These are shown so as to have the option to copy the human brain capacities regarding discovering complex connections between the inputs and outputs [65]. One of the most well-known designs for NNs is the multi-layer perceptron, which comprises of one input layer, at least one hidden layer, and one output layer. As per [66], central points of contention waiting be tended to in building NNs are their topology, structure, and learning algorithm. The most used MLP topology for credit scoring is three-layer feedforward back propagation network. Consider the input of a credit scoring training set $x = \{x_1, x_2, \ldots, x_n\}$; the MLP model works in one direction, starting from feeding the data x to the input layer (x includes the customer's attributes or characteristics). These inputs are then sent to a hidden layer through links, or synapses, associated with the random initial weight for every input. The hidden layer will process what it has received from the input layer and, accordingly, will apply an activation function to it. The result is worked as a weighted input to the output layer, which will further process weighted inputs and apply the activation function, take the lead to a final decision [67]. In recent years ensemble models became more popular, so instead of a single NN, Bagging NN is used with 10 neural networks.

Support vector machines

A SVM is another ground-breaking machine learning method utilized in order and credit scoring issues. SVMs are used for binary classification to make the best separation that splits the input data into two classes (good and bad credit). SVMs were first proposed by Cortes and Vapnik [68], adapting the form of a linear classifier. The primary distinction of the SVM model from the linear one is the occurrence of a function that is used to map the data into a higher dimensional space. To achieve this, linear, polynomial, radial basis, and sigmoid kernel functions were suggested. An SVM maps non-linear data of two classes to a high-dimensional feature space, with a linear model then being used to implement the non-linear classes. The linear model in the new feature space will denote the non-linear decision margin in the original space. Consequently, the SVM will build an optimal line or hyperplane that can perfectly separate the two classes in the space. SVMs are being widely used in credit scoring and other fields owing to the method's exceptional results [69, 70].

Random forests

A random forest (RF), as proposed by Breiman [71], is considered an innovative decision tree (DT) technique which consists of a large number of trees that are created by generating n subsets from the core dataset, with each subset being a tree created based on randomly selected variables, therefore the name "random forest". After all the DTs are generated and trained, the final decision class is based on a voting method, where the most popular class decided by the trees is selected as the final output class by the RF.

Logistic regression

Logistic Regression (LOGR) has been considered until now to be the industry standard for credit scoring model development [68]. It is a broadly used statistical technique that is popular for solving classification and regression problems. LOGR is used to model a binary outcome variable, usually characterized by 0 or 1 (good and bad loans). The LOGR formula is expressed in Atiya and Parlos [19].

Performance measure metrics

To validate the proposed model and in order to reach a reliable and strong conclusion on the predictive accuracy of the proposed method, five performance indicator measures are implemented, specifically: (1) accuracy, (2) Area Under the Curve (AUC), (3) H-measure, (4) Kolmogorov–Smirnov (KS) chart, and (5) Brier's score. These are chosen because they are popular in credit scoring and they give a comprehensive view on all facets of model performance. The accuracy stands for the proportion of correctly classified good and bad loans, which measures the predictive power of the model. As such, this is a standard that measures the discriminating ability of the model [68]. The accuracy can be defined as the percentage of correctly classified instances

$$\frac{TP + TN}{TP + TN + FP + FN},$$

where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives, respectively.

AUC is a tool used in binary classification analysis to determine which of the models used predicts the classes the best. According to Hand [72], the AUC can be used to estimate the model's performance without any preceding evidence about the error costs. However, it assumes different cost distributions among classifiers depending on their actual score distribution, which prevents them from being compared effectively. As a result, Hand [72] proposed the H-measure as an alternative to AUC for measuring classification performance, which assumes different cost distributions between classifiers without depending on their scores. In other words, this measure finds a single threshold distribution for all classifiers. AUC is evaluated as area under the ROC-curve for measured classifier.

The KS distribution was originally formulated as an observance hypothesis test for distribution-fitting to data. In binary classification problems, it has been used as a divergence metric for assessing the classifier's discriminant power by measuring the distance that its score produces between the cumulative distribution functions of the two data classes [73].

Lastly, the Brier score, which is also known as the mean squared error [74], measures the accuracy of the probability predictions of the classifier by taking the mean squared error of the probability. In other words, it shows the average quadratic possibility of a mistake. The main difference between the Brier score and accuracy is that it directly takes the probabilities into the account, while accuracy transforms these probabilities into zero or one based on a predetermined threshold or cut-off score. The lower the Brier score, the better the classifier performance. The most common formulation of the Brier score is:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left(f_t - \sigma_t\right)^2$$

in which f_t is the probability that was forecast, σ_t the actual outcome of the event at instance t (zero if it does not happen and one if it does happen) and N is the number of forecasting instances.

To check whether a model's behavioural score can be considered as the likelihood of missed payment, calibration curves are used. Well-known as reliability diagrams, they can be applied to classifiers which predictand obtain a probability of the respective class. Reliability diagrams offer a diagnostic to check whether the scores are trustworthy. Thus, a prediction is considered as trustworthy if the event happens with an observed relative frequency consistent with the forecast value [75]. A calibration curve works by sorting the output scores of the classifier. In Particular, the forecasts are apportioned into a fixed number of buckets along the x-axis. The number of classes or labels are then counted for each bin (e.g., the relative observed frequency). After All, the counts are normalized. The results are then plotted as a line plot. If the classifier is forecasting accurately, then it is expected that the percentage of dominant class classifications and the mean probabilities assigned to the dominant classes in each bin to be close to one another. If it is not doing so accurately, these two values

diverge. The point positions on the curve relative to the diagonal help to interpret the forecasts, for example:

(1) Below the diagonal: the model has over-forecast; the probabilities are too large.
(2) Above the diagonal: the model has under-forecast; the probabilities are too small.

Statistical significance tests

As indicated by Witten et al. [76], it is not adequate to demonstrate that one model accomplishes results in a way that is better than another, because of the different performance measures or splitting techniques used. For complete performance evaluation, it would appear to be proper to actualize some some hypothesis testing to stress that the experimental differences in performance are statistically significant and not just due to random splitting influences. Selecting the right test for detailed experiments depends on factors such as the number of datasets and the number of classifiers to be contrasted.

According to Demšar [77], statistical tests can be parametric (e.g., paired *t*-test) and non-parametric (e.g., Wilcoxon, McNemar). However, the author recommended that non-parametric tests are desirable to parametric tests as the last can be conceptually unsuitable and statistically unsafe. Non-parametric tests may be more applicable and safer than parametric tests since they do not presume the normality of data or homogeneity of variance [77]. Accordingly, in this study, the McNemar test to compare the ranking performance of all the models measured across a unique dataset is adopted [78]. According to Kavzoglu [79], the McNemar test investigates the statistical significance of the differences in classifiers' performances. The test is a Chi-square (χ^2) test for goodness of fit, comparing the distribution of counts expected under the null hypothesis to the observed counts. It is applied to a 2×2 contingency table, the cells of which include the number of cases correctly and incorrectly classified by both models and the number of samples classified correctly by only one model.

The aim of the McNemar test is to check the null hypothesis, which says that neither of the two models performs better than the other. The alternative hypothesis asserts that the performance of the two models are not equal. The McNemar statistic is as illustrated in Eq. (11):

$$\chi^2 = \frac{\left(\left|n_{ij} - n_{ji}\right| - 1\right)^2}{n_{ij} - n_{ji}} \tag{11}$$

where n_{ij} indicates the number of cases misclassified by model i but classified correctly by model j, and n_{ji} indicates the number of cases misclassified by model j but not by model i.

The computed statistic is thought as a value from the χ^2 distribution with 1 degree of freedom. Based on this assumption, the p-value is calculated. If this p-value is smaller than predefined significance level α, then we fail to reject the null hypothesis. Otherwise, we reject the null hypothesis, and accept the alternative hypothesis. For example, if the value of test statistic is greater than 3.84, then (according to the χ^2 table at 95% confidence interval) it can be stated that the two methods differ in their performances.

In other words, the difference in performance between the methods i and j is said to be statistically significant [78, 79].

Results and discussion

In this section, the results of the proposed LSTM model are presented along with comparisons to the benchmark classifiers. The model is validated over the above-described dataset across five performance measure metrics. In addition, several tables and figures regarding the proposed model results and comparison to traditional models are provided and discussed. All the experiments for this study were performed using Python 3.8×64 on a PC with an AMD 8-core Ryzen™ 7 3700X 3.6–4.4 GHz processor and 32 GB RAM, running Microsoft Windows 10 operating system.

To outline the discrimination power of the Bidirectional LSTM model performance measures are calculated not only for all active customers, but for different subsets of them:

(1) Customers with one missed payment during the last 2 months are the group that generally have a low risk of default, but the recent missed payment is a reason to look at those in this group more closely.
(2) Customers with a missed payment during the last month is a subset of the first group. Whilst one missed payment can be made by chance, here there is a need to look at this group to distinguish riskier customers from other ones.
(3) Customers with two consecutive missed payments form a group in which most customers might have financial problems because it is unlikely to forget to pay during more than 1 month.
(4) Customers with three consecutive missed payments are those on a verge of default. For this group, a fourth missed payment is equivalent to default, so a Bidirectional LSTM prediction of the fourth missed payment is a prediction of default.

As a next step, the model was compared with five classical classifiers: Gradient Boosting, Bidirectional Neural Network, Logistic Regression, SVM, and Random Forest. Comparisons were made not only using the performance measures but also using the statistical McNemar test.

The LSTM model provides probability of missed payment for next month for each customer based on previous 6 months, and it does not use future data to predict past.

Bidirectional LSTM model results

To prove that the results obtained on the testing set are sound and to make the results of Bidirectional LSTM significant, different measures need to be evaluated, each of which reflect different aspects of the model performance:

(1) Accuracy is the simplest method of evaluating the model preciseness. It does not consider any misclassification loss and simply displays the proportion of correctly classified missed payments for the default score threshold, which is equal to 0.5.
(2) Specificity measures the proportion of missed payments that are correctly identified.

(3) Specificity measures the proportion of payments made on time that are correctly identified.
(4) The balanced accuracy in binary and multiclass classification problems to deal with imbalanced datasets. It is defined as the average of recall obtained on each class. For binary classification problems, balanced accuracy is evaluated using Eq. (12).

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{12}$$

(5) AUC tells us how the model will perform for different selected thresholds.
(6) Brier score reflects the discriminatory power of the model (i.e., how certain the model is about the customer's predicted missed payment).
(7) KS reflects the maximum difference between the fraction of correctly classified customers, those who missed a payment, and incorrectly classified customers, those who did not miss a payment. The value tells us that model correctly classifies not only the presence of a missed payment, but also absence of it.
(8) H-measure is an integral measure over all misclassification costs. A high H-measure value tells us that, regardless of actual cost of misclassification, the total loss cost of model is low.

As shown in Table 2, the correctness of the LSTM model prediction ability is shown in "Accuracy" column. Performance measures for the customers with three or more consecutive missed payments is much lower than for other groups. It could be explained by the fact, that some proportion of customers drastically change its behaviour in the risk of bankruptcy and trial. So, based on its past behaviour they should have fourth missed payment, but pressure from the bank forces them to pay. The table shows that that the model considers consumers with payment problems as those who are more prone to them in the future. The classifier accuracy is lower than for the transactional dataset, which can be explained by initial data pre-processing which might lead to information loss.

Sensitivity (ability of model of identifying missed payments) is around 40% for first two groups of customers rise to 90% for the last one. On other hand, specificity for all groups except the last one is more than 90%. It tells us that if model identifies customer as "low risk", bank management should not worry about future payments from him. As mentioned earlier, the higher the AUC value, the better the classifier is capable of distinguishing between classes. The proposed model shows similar prediction ability on all subsets of active customers except the last one. For those except the last it is higher than 77%, which proves good classifier separability. The lower the Brier score is, the better classifier performs. an increase can be seen in the Brier score for the customers with three missed payments. The higher the Kolmogorov–Smirnov chart statistics, the better the discriminative power of the model. As was mentioned before, for all subsets except the last one, this value is sufficiently high to prove good discriminative model ability.

As was mentioned before, the H-measure is a measure of the misclassification loss, and this depends on the relative proportion of objects belonging to each class. The influence of the different number of customers with missed payment fee can be seen from the table. But generally, the higher H-measure, the better the classifier is in terms of

Table 2 Performance measures for LSTM classifier for the non-transactional dataset

Description	Total	Missed Payments	Proportion	Accuracy (%)	Sensit-ivity (%)	Specif-icity (%)	AUC (%)	Brier Score (%)	KS	H-Measure
All customers	30,000	6636	22.12	82.4	37.51	95.15	78.47	13.28	0.43	0.3
Customers with at least one missed payment during the last 2 months	15,265	3997	26.18	79.85	39.58	94.13	78.37	14.81	0.43	0.29
Customers with missed payment during last month	13,714	3567	26.01	79.95	39.22	94.27	78.23	14.78	0.43	0.29
Customers with two consecutive missed payments	7974	2592	32.51	77.33	51.04	89.99	79.81	16.15	0.46	0.32
Customers with three or more consecutive missed payments	313	210	67.09	73.16	90	38.83	76.86	17.62	0.46	0.27

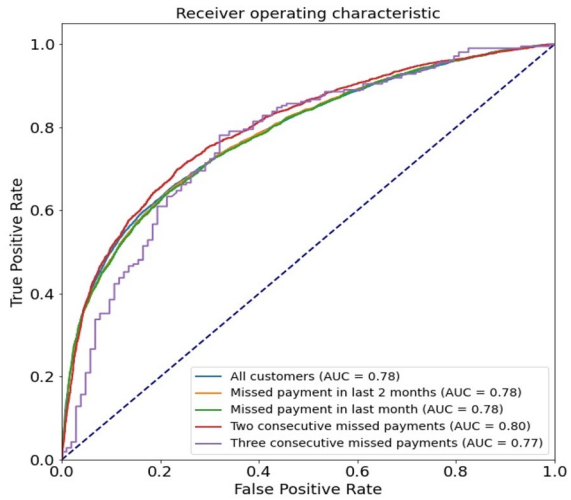

Fig. 6 ROC and AUC values for different customer groups

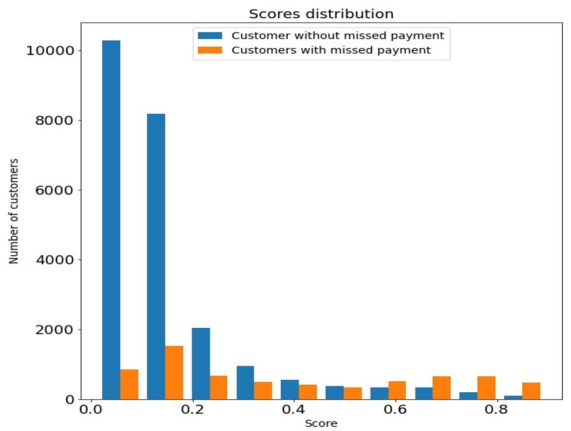

Fig. 7 Distribution of scores for different customer groups

performance over different misclassification costs. For all subsets of customers that are investigated, this value is good enough.

As it can be seen from Fig. 6, the AUC value for the Bidirectional LSTM model is high for all customers as well as for specific risk groups except the last one (with three consecutive missed payments), despite the fact that proportion of missed payments for all customers and for customers with missed payment differs greatly (see Table 2). The shape of the ROC curve is round for all customer groups. The highest AUC is for the customers with two consecutive missed payments. Bank can use this group to early put pressure on such customers and prevent third missed payment.

Figure 7 represents the behaviour score distribution for different customer groups along with the observed behaviour. The splitting process into ten buckets along the x-axis was based on the customer missed payment prediction. Thus, it is expected that the number of customers without a missed payment will decrease along the axis, while

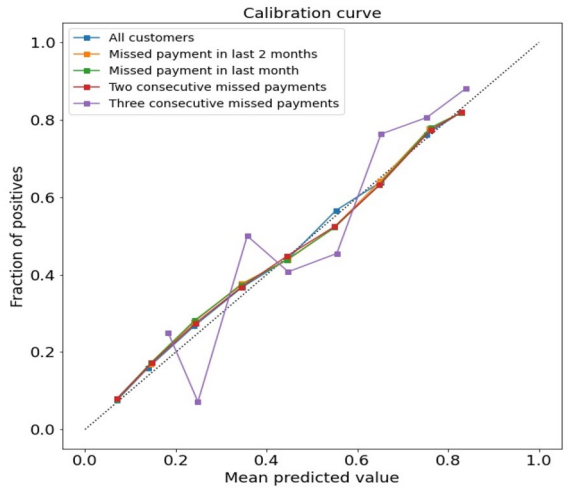

Fig. 8 Calibration curves for different customer groups

Table 3 Comparison of performance measures for all classifiers for the non-transactional dataset

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	AUC (%)	Brier score (%)	KS	H-measure
GB	82.07	36.47	95.02	65.75	78.03	13.43	0.43	0.29
BNN	81.78	37.07	94.48	65.77	77.53	13.58	0.42	0.28
SVM	81.46	28.56	96.49	62.52	69.57	14.33	0.37	0.23
RF	80.18	17.01	98.12	57.57	76.54	14.38	0.41	0.27
LOGR	80.88	22.56	97.44	60	71.82	14.55	0.37	0.24
Bidirectional LSTM	82.4	37.51	95.15	66.33	78.47	13.28	0.43	0.3

the number of customers with missed payment will increase, and the histogram reflects this tendency. So, whilst there are, of course, misclassified clients in every group, their percentage is significantly less than correctly classified. Thus, the proposed model can be considered as reliable.

Figure 8 compares how well the probabilistic predictions of Bidirectional LSTM for the different client groups are calibrated using 10 bins. The calibration curve for all clients shows that it is the best calibrated among the others. It fits the line almost perfectly, which means that missed payment scores can be considered as probabilities. The only group with the curve far from the central line is customers with three or more consecutive missed payments. This curve has small over-forecast for low scores, but in general it also lies close enough to other curves.

Benchmark model results and comparison

To verify the strength and discriminative power of proposed model, its performance was compared to five benchmark models, namely GB, BNN, RF, SVM and LOGR. As all benchmark classifiers do not accept temporal input, all temporal data was flattened

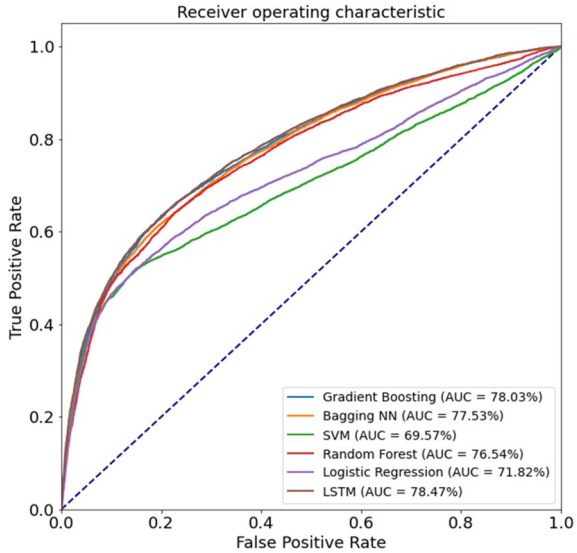

Fig. 9 ROC and AUC values for all classifiers

before feeding into each classifier. The comparison results are show in Table 3 which represents the performance indicator measures for the different classifiers on the same input data.

On the first sight, the correctness of the predictions is similar and high enough for all the models. The closest by performance model to proposed bidirectional LSTM model is the GB model. The reason why performance of all classifiers is so close to each other lies in lower dimensionality of the input data. For each customer there is only 23 features instead of hundreds of transactions in previous data set. So, simple classifiers have less problems in extraction useful information from feature space. Threshold changes can improve classifier accuracy; maximum accuracy can be achieved by applying the optimal threshold. So, as it can be seen, there is a slight increase for all of them when applying the optimal threshold, but the highest value still belongs to proposed model. It is obvious that bidirectional LSTM and GB have similar KS value, which is slightly higher than corresponding value for other classifiers. Similar pattern can be observed with H-measure. Brier score for bidirectional LSTM model is the lowest, which proves the quality of this model.

Proposed model has the highest sensitivity among all other models. Its specificity is equal to 95%. Despite some other classifiers like Random Forest, SVM and LOGR have higher specificity, their sensitivity is much lower. That is why balanced accuracy for Bidirectional LSTM classifier is the highest among other classifiers. Therefore, from the Fig. 9 it can be conducted that performance of Bidirectional LSTM is the best among all considered classifiers. The worst AUC value is from the SVM classifier (especially in the second part of the plot), which means that it is acceptable to use it to increase the True Positive Rate value.

For such complex problem even half of the percent of increase in accuracy or AUC of classifier leads to significant loss decrease for bank due to missed payments and bankruptcies of customers. That's why we think that results are significant.

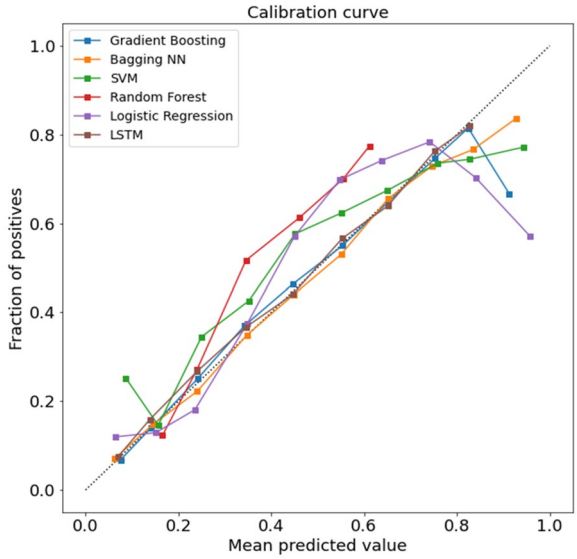

Fig. 10 Calibration curves for all classifiers

Table 4 McNemar test for Bidirectional LSTM pairwise comparison with other classifiers

	p-value	Statistic
GB	3.62×10^{-5}	17
BNN	8.01×10^{-11}	42
SVM	2.77×10^{-18}	76
RF	1.02×10^{-47}	211
LOGR	1.15×10^{-29}	128

Figure 10 compares how well the probabilistic predictions of the different classifiers are calibrated, using a calibration curve with ten bins. The plot shows that there are two perfectly calibrated classifiers: Bidirectional LSTM and Bagging NN. That is why scores of this classifier can be used as probabilities. The worst curves have RF and LOGR classifiers. To make sure that the difference in performance measures are statistically significant and are not caused by chance, the McNemar test is used.

Table 4 represents the results of applying the McNemar test for pairwise comparison of the LSTM model and the other classifiers. During the application of McNemar test the same value of significance threshold is used for the transactional data set $\alpha = 0.05$. According to the results above, every classifier pair shows a statistically significant performance difference. The closest to Bidirectional LSTM classifier is GB, which has p-value equal to 3.62×10^{-5}. Current results of McNemar test combined with the previously mentioned performance indicator measures prove that all the traditional classifiers show worse prediction ability in contrast with the Bidirectional LSTM model.

As a last step, we provide auxiliary table with standard deviations of the most important measures across all folds was measured. In the last column we provide training time of each model.

Table 5 Comparison of performance measures standard deviation

Classifier	Accuracy STD (%)	AUC STD (%)	Brier score STD	Time of training in s
GB	0.41	0.52	0.25	7.4
BNN	0.47	0.42	0.22	306
SVM	0.46	1.03	0.29	84
RF	0.47	0.47	0.18	8.5
LOGR	0.41	0.71	0.19	0.2
Bidirectional LSTM	0.37	0.41	0.2	127

As we can see from Table 5, LSTM neural network has the lowest values for accuracy and AUC, and third lowest for Brier Score. Despite LSTM is more complex model than others, it takes comparable time for training and because it utilizes GPU capabilities of testing PC, which makes training process much faster comparing to CPU-driven models.

The LSTM model can deal with real time data in addition, evaluation of LSTM model is very fast (seconds of computational time).

Conclusion

The paper emphasizes the importance of credit card scoring for assessing and decreasing bank losses. By conducting a detailed comparison procedure it was proven that the LSTM model is the one that gives the highest accuracy in predicting late fees and mispayments, and that is why it is the best for banks' interests. In this paper, Bidirectional LSTM model was presented and validated on non-transactional open dataset.

To prove the effectiveness of the proposed model, it was compared to five other traditional classification models. The following performance measures were used for the comparison, specifically: accuracy, AUC, H-measure, Kolmogorov–Smirnov test, Brier score, calibration curves, and the McNemar test. On Taiwanese bank credit card dataset, it has 82.4% accuracy, whilst the best of other models has 81.8%. It seems not so much, however in banking business even 1% of difference in bad credit card behaviour prediction makes huge difference in terms of bank losses.

All measures prove outperformance by the Bidirectional LSTM model. Therefore, it can be concluded that Bidirectional LSTM performs statistically better than other classifiers. Its calibration curve shows that the output of the model can be considered as the probability of default without any additional improvements.

Banks can use outcome of the model not only as a binary output (whether customer will have missed payment in each next month), but also can make use of scores of each client.

LSTM gives the probability of user to be insolvent in next month. It is up to management to set up thresholds above which bank moves this user into group of high or medium risk with corresponding consequences to the user (decreasing credit card limit, blocking card etc.).

In other words, the scores provided by LSTM model can be used to group customers into different risk groups. Thus, bank can use different security and service level for each of these risk groups. Moreover, such scores can be used as missing payment

probabilities, so bank management can calculate potential losses of each customer and even credit portfolios. This will allow management to efficiently assess financial risks and make bold financial decisions.

In future work, the model will be tested on other datasets that are transactional and non-transactional in nature to prove its efficiency. Moreover, the proposed model will be extended to customer credit scoring for consumer loans.

Abbreviations
LSTM: Long short term memory; GB: Gradient boosting; NN: Neural networks; BNN: Bagged neural network; SVM: Support vector machines; RF: Random forests; LOGR: Logistic regression; RNN: Recurrent neural networks; ROC: Receiver operating curve; AUC: Area under the curve; KS: Kolmogorov Smirnov.

Acknowledgements
Not applicable.

Authors' contributions
MA designed and carried out experiments and data analysis and drafted the manuscript. MM and MA participated in research coordination and checked, read and approved the final manuscript. All authors contributed in revising the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Office of Research, Zayed University under Grant Number R20053.

Availability of data and materials
The dataset used for supporting the conclusions of this paper is available from the public data repository at http://archive.ics.uci.edu/ml/index.php.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Information Systems, College of Technological Innovation, Zayed University, 19282 Dubai, United Arab Emirates. [2]Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, Kingston Lane, Uxbridge UB8 3PH, UK.

Received: 26 February 2021 Accepted: 3 May 2021
Published online: 19 May 2021

References

1. Dyché J. The CRM handbook: a business guide to customer relationship management. Boston: Addison-Wesley Longman Publishing; 2001.
2. Hand DJ, Henley WE. Statistical classification methods in consumer credit scoring: a review. J R Stat Soc Ser A. 1997;160(3):523–41. https://doi.org/10.1111/j.1467-985x.1997.00078x.
3. Anderson R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford: Oxford University Press; 2007.
4. Liu Y. New issues in credit scoring application. Institut für Wirtschaftsinformatik, Abteilung Wirtschaftsinformatik II, Georg-August-Universität, Göttingen. 2001.
5. Bensic M, Sarlija N, Zekic-Susac M. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intell Syst Acc Fin Manag. 2005;13(3):133–50. https://doi.org/10.1002/isaf.261.
6. Kennedy K, Mac Namee B, Delany SJ, O'Sullivan M, Watson N. A window of opportunity: Assessing behavioural scoring. Expert Syst Appl. 2013;40(4):1372–80. https://doi.org/10.1016/j.eswa.2012.08.052.
7. Malik M, Thomas LC. Modelling credit risk of portfolio of consumer loans. J Oper Res Soc. 2010;61(3):411–20. https://doi.org/10.1057/jors.2009.123.
8. McNab H, Wynn A. Principles and practice of consumer credit risk management. Ottawa: CIB Publishing; 2000.
9. So MMC, Thomas LC. Modelling the profitability of credit cards by Markov decision processes. Eur J Oper Res. 2011;212(1):123–30. https://doi.org/10.1016/j.ejor.2011.01.023.

10. Baesens B, Van Gestel T, Stepanova M, Van den Poel D, Vanthienen J. Neural network survival analysis for personal loan data. J Oper Res Soc. 2005;56(9):1089–98. https://doi.org/10.1057/palgrave.jors.2601990.
11. Lim MK, Sohn SY. Cluster-based dynamic scoring model. Expert Syst Appl. 2007;32(2):427–31. https://doi.org/10.1016/j.eswa.2005.12.006.
12. Sarlija N, Bensic M, Zekic-Susac M. Comparison procedure of predicting the time to default in behavioural scoring. Expert Syst Appl. 2009;36(5):8778–8. https://doi.org/10.1016/j.eswa.2008.11.042.
13. Hsieh N-C. An integrated data mining and behavioural scoring model for analyzing bank customers. Expert Syst Appl. 2004;27(4):623–33. https://doi.org/10.1016/j.eswa.2004.06.007.
14. Bertola G, Disney R, Grant C. The economics of consumer credit. Cambridge: MIT Press; 2008.
15 Kim H, Cho H, Ryu D. An empirical study on credit card loan delinquency. Econ Syst. 2018. https://doi.org/10.1016/j.ecosys.2017.11.003.
16. Kumar PR, Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques: a review. Eur J Oper Res. 2007;180(1):1–28. https://doi.org/10.1016/j.ejor.2006.08.043.
17. Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J. Benchmarking state-of-the-art classification algorithms for credit scoring. J Oper Res Soc. 2003;54(6):627–35. https://doi.org/10.1057/palgrave.jors.2601545.
18. Mylonakis J, Diacogiannis G. Evaluating the likelihood of using linear discriminant analysis as a commercial bank card owners credit scoring model. Int Bus Res. 2010. https://doi.org/10.5539/ibr.v3n2p9.
19. Atiya AF, Parlos AG. New results on recurrent network training: unifying the algorithms and accelerating convergence. IEEE Trans Neural Netw. 2000;11(3):697–709. https://doi.org/10.1109/72.846741.
20. Bellotti T, Crook J. Credit scoring with macroeconomic variables using survival analysis. J Oper Res Soc. 2009;60(12):1699–707. https://doi.org/10.1057/jors.2008.130.
21. Wang C, Han D, Liu Q, Luo S. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. IEEE Access. 2019;7:2161–8. https://doi.org/10.1109/access.2018.2887138.
22. Thomas LC. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. Int J Forecast. 2000;16(2):149–72. https://doi.org/10.1016/s0169-2070(00)00034-0.
23. Thomas LC, Ho J, Scherer W. Time will tell: behavioural scoring and the dynamics of consumer credit assessment. IMA J Manag Math. 2001;12(1):89–103. https://doi.org/10.1093/imaman/12.1.89.
24. Louzada F, Ara A, Fernandes GB. Classification methods applied to credit scoring: systematic review and overall comparison. Surv Oper Res Manag Sci. 2016;21(2):117–34. https://doi.org/10.1016/j.sorms.2016.10.001.
25. Setiono R, Thong JYL, Yap C-S. Symbolic rule extraction from neural networks. Inf Manag. 1998;34(2):91–101. https://doi.org/10.1016/s0378-7206(98)00048-2.
26. Sharda R, Wilson RL. Neural network experiments in business failures predication: a review of predictive performance issues. Int J Comput Intell Organ. 1996. https://doi.org/10.1109/hicss.1993.284245.
27. Hsieh H-I, Lee T-P, Lee T-S. Data mining in building behavioural scoring models. 2010. https://doi.org/10.1109/cise.2010.5677005.
28. Ha S, Nguyen H-N. Credit scoring with a feature selection approach based deep learning. MATEC Web Conf. 2016;54:5004. https://doi.org/10.1051/matecconf/20165405004.
29. Cenggoro TW, Mahesworo B, Budiarto A, Baurley J, Suparyanto T, Pardamean B. Features importance in classification models for colorectal cancer cases phenotype in Indonesia. Procedia Comput Sci. 2019;157:313–20. https://doi.org/10.1016/j.procs.2019.08.172.
30. Hassan MR, Hossain MM, Begg RK, Ramamohanarao K, Morsi Y. Breast-cancer identification using HMM-fuzzy approach. Comput Biol Med. 2010;40(3):240–51. https://doi.org/10.1016/j.compbiomed.2009.11.003.
31. Sani NS, Abdul Rahman M, Bakar A, Sahran S, Sarim H. Machine learning approach for bottom 40 percent households (B40) poverty classification. Int J Adv Sci Eng Inf Technol. 2018;8:1698. https://doi.org/10.18517/ijaseit.8.4-2.6829.
32. De Vito S, Piga M, Martinotto L, Di Francia G. CO, NO_2 and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. Sensors Actuators B Chem. 2009;143(1):182–91. https://doi.org/10.1016/j.snb.2009.08.041.
33. Caraka R, Lee Y, Chen R, Toharudin T. Using hierarchical likelihood towards support vector machine: theory and its application. IEEE Access. 2020. https://doi.org/10.1109/ACCESS.2020.3033796.
34. Pereira S. Modelling credit card customer behaviour. Work project presented as a partial requirement for Degree of Master of Statistics and Information Management, with a specialization in Information Analysis and Management. 2019.
35. Alborzi M, Khanbabaei M. Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. Int J Bus Inf Syst. 2016;23(1):1. https://doi.org/10.1504/ijbis.2016.078020.
36. Bastani K, Asgari E, Namavari H. Wide and deep learning for peer-to-peer lending. Expert Syst Appl. 2019;134:209–24. https://doi.org/10.1016/j.eswa.2019.05.042.
37. Akkoç S. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. Eur J Oper Res. 2012;222(1):168–78. https://doi.org/10.1016/j.ejor.2012.04.009.
38. Addo P, Guegan D, Hassani B. Credit risk analysis using machine and deep learning models. Risks. 2018;6(2):38. https://doi.org/10.3390/risks6020038.
39. Gui L. Application of machine learning algorithms in predicting credit card default payment, University of California. 2019.
40. Heryadi Y, Warnars HL. Spits Warnars, Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, stacked LSTM, and CNN-LSTM. 2017.
41 Jurgovsky J, et al. Sequence classification for credit-card fraud detection. Expert Syst Appl. 2018. https://doi.org/10.1016/j.eswa.2018.01.037.
42. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE International conference on acoustics, speech and signal processing. 2013; p. 6645–9. https://doi.org/10.1109/ICASSP.2013.6638947.
43. Malhotra P, Vig L, Shroff G, Agarwal P. Long short-term memory networks for anomaly detection in time series. 2015.
44. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst. 2014;4.

45. Siami-Namini S, Namin AS. Forecasting economics and financial time series: ARIMA vs. LSTM. 2018.
46. Bengio Y, Frasconi P, Simard P. The problem of learning long-term dependencies in recurrent networks. In: IEEE international conference on neural networks. 1993. p. 1183–8.
47. Srinivasan K, Cherukuri AK, Vincent DR, Garg A, Chen BY. Chen, an efficient implementation of artificial neural networks with K-fold cross-validation for process optimization. J Internet Technol. 2019;20:1213–25. https://doi.org/10.3966/160792642019072004020.
48. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
49. Lai CY, Chen RC, Caraka RE. Prediction average stock price market using LSTM. 2019.
50. Toharudin T, Pontoh R, Caraka R, Zahroh S, Lee Y, Chen R. Employing long short-term memory and facebook prophet model in air temperature forecasting. Commun Stat Simul Comput. 2021. https://doi.org/10.1080/03610918.2020.1854302.
51. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Proces. 1997;45:2673–81. https://doi.org/10.1109/78.650093.
52. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.3215. 2014.
53. Cui Z, Ke R, Pu Z, Wang Y. Deep Bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. 2018.
54. Yeh I, Lien C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl. 2009;36:2473–80.
55. Zhang W, Yang D, Zhang S, Ablanedo-Rosas JH, Wu X, Lou Y. A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. Expert Syst Appl. 2021;165:113872.
56. Tripathi D, Edla DR, Bablani A, Shukla AK, Reddy BR. Experimental analysis of machine learning methods for credit score classification. Prog Artif Intell. 2021;15:1–27.
57. Jadhav S, He H, Jenkins K. Information gain directed genetic algorithm wrapper feature selection for credit rating. Appl Soft Comput. 2018;1(69):541–53.
58. Hamori S, Kawai M, Kume T, Murakami Y, Watanabe C. Ensemble learning or deep learning? Application to default risk analysis. J Risk Financial Manag. 2018;11(1):12. https://doi.org/10.3390/jrfm11010012.
59. Shen F, Zhao X, Kou G, Alsaadi FE. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. Appl Soft Comput. 2021;98:106852.
60. Bellotti T, Crook J. Modelling and predicting loss given default for credit cards. Work Pap Quant Financ Risk Manag Cent. 2007.
61. Lessmann S, Baesens B, Seow H-V, Thomas L. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur J Oper Res. 2015. https://doi.org/10.1016/j.ejor.2015.05.030.
62. Bhatia S, Sharma P, Burman R, Hazari S, Hande R. Credit scoring using machine learning techniques. Int J Comput Appl. 2017;161:1–4.
63. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21. https://doi.org/10.3389/fnbot.2013.00021.
64. Haykin SS. Neural networks and learning machines. 3rd ed. Pearson Education: Upper Saddle River; 2009.
65 Bhattacharyya S, Maulik U. Soft computing for image and multimedia data processing. Berlin: Springer; 2013.
66. Angelini E, Tollo G, Roli A. A neural network approach for credit risk evaluation. Q Rev Econ Financ. 2008;48:733–55. https://doi.org/10.1016/j.qref.2007.04.001.
67. Malhotra R, Malhotra DK. Evaluating consumer loans using neural networks. Omega. 2003;31:83–96. https://doi.org/10.2139/ssrn.314396.
68. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.
69. Lahsasna A, Ainon R, Wah T. Credit scoring models using soft computing methods: a survey. Int Arab J Inf Technol. 2010;7:115–23.
70. Huang C-L, Chen M-C, Wang C-J. Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl. 2007;33(4):847–56. https://doi.org/10.1016/j.eswa.2006.07.007.
71. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
72. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77(1):103–23. https://doi.org/10.1007/s10994-009-5119-5.
73. Adeodato PJ, Melo SB. On the equivalence between Kolmogorov–Smirnov and ROC curve metrics for binary classification. 2016.
74. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.
75 Bröcker J, Smith L. Increasing the reliability of reliability diagrams. Weather Forecast. 2007. https://doi.org/10.1175/WAF993.1.
76. Witten IH, Frank EF, Hall MA. Credibility: evaluating what's been learned. In: Witten IH, Frank E, Hall MA, editors. Data mining: practical machine learning tools and techniques. 3rd ed. Boston: Morgan Kaufmann; 2011. p. 147–87.
77. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
78. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895–923. https://doi.org/10.1162/089976698300017197.
79. Kavzoglu T. Object-oriented random forest for high resolution land cover mapping using quickbird-2 imagery. In: Samui P, Sekhar S, Balas VE, editors. Handbook of neural computation. Cambridge: Academic Press; 2017. p. 607–19.

Publisher's Note