NEURAL NETWORK APPROACH TO LOAN DEFAULT PREDICTION

Manjeet Kumar¹, Vishesh Goel², Tarun Jain³, Sahil Singhal⁴, Dr. Lalit Mohan Goel⁵

¹,²,³,⁴ Students, CSE, Bharati Vidyapeeth’s College of Engineering, New Delhi
⁵Prof, CSE Department, Bharati Vidyapeeth’s College of Engineering, New Delhi

Abstract - Loan default prediction is one of the most critical and crucial problem faced by financial institutions and organizations as it has a noteworthy effect on the profitability of these institutions. In recent years, there is a tremendous increase in the volume of non – performing loans which results in a jeopardizing effect on the growth of these institutions. In this paper, we introduce an approach to implement a neural network model which will be used for loan default prediction. The prediction is performed by taking into consideration the financial and personal details provided by the potential debtor. The proposed neural network model is trained and tested using the dataset provided by the Lending Club bank. Before training the neural network, Principal Component Analysis is performed on the dataset in order to reduce the dimensionality of the dataset. Based on the results obtained, it is inferred that our proposed neural network model outperforms other classifiers that are traditionally used by financial institutions for loan default prediction.

Index Terms— Loan Default Prediction, Neural Network, Principal Component Analysis, Classifiers

I. INTRODUCTION

Along with the financial market, the credit business of banking industry is also rapidly expanding with extreme competition. At the same time, consumption method by the consumer have turned the consumer loan into major competitive market.

According to banking laws, “Consumer loan” is defined as personal credits provided by financial institution which is payed back by the method of instalment payment and is generally provided for personal or family consumption, to pay certain expenses, such as medical cost, education cost, travelling cost or to pay some other accumulated debt taken with some other purpose of consumption.

With increase in “Consumer loan” provided by financial institutions, the scenario of loan default is also increasing. With excessive competition in between different financial institutions to attract more consumers, it became important to not to deny the loan or simply provide the loan to consumer. To provide loan, it is important to know whether the consumer will be able to pay back the provided loan or not. Although, financial institutes provide “Consumer Loan” with extreme caution and after various affirmations, but still there are some cases in which consumer is not able to pay the loan back. These type of cases is what we call loan default cases. These cases can vary from extreme minor cases to major cases resulting in scams. Some of these cases may even lead the financial institution on the verge of bankruptcy. Thus, it became extreme important to predict whether loan default scenario can occur with some particular case or not on the basis of previous similar kind of consumers history whether they were defaulters or not.

To predict loan default, various prediction systems emerged. Every system created can be broadly divided into two parts: Predictor Variables and Predicting Models.

A. PREDICTOR VARIABLES

These are defined as those variables which may have influenced Consumer credit loan default. There are seven types of such variables which are the: Relationship between consumer and creditors, consumer’s annual income, debt-income ratio, occupation, house ownership, work duration or whether consumer possess the saving account or checking account. There are many more key factors or variables which play an important role in predicting the loan default scenario. According to Updegrave (1987) [1] there were eight key factors of the credit risk that influenced the credit or short-term loan. These factors were: Number of creditors, declared bankruptcy or not, previous payment records, consumer's income, work/resident duration, occupation, age, consumer possess the saving account or checking account. Similarly, according to Steenackers and Goovaerts (1989) [2] the key factors which may have influenced the credit loan are borrower's age, district, resident/work duration, owner of phone, whether working in public sector, house ownership, monthly income and it has also been pointed that loan duration and its numbers also have a significant relationship with paying back the loan to financial institution. Thus, it can be said that there are different variables which influence the loan default scenario which also includes demographic variables. In addition, as rightly pointed out by Chiang, Chow, and Liu (2002) [3] there are various individual characteristics such as attitude of borrower which also influences the default risk behavior of the borrower.

B. PREDICTING MODELS

It is important to use correct model from various different models present because the model chosen plays a crucial role in determining efficiency, accuracy and precision of prediction system. Predictor variables provide data which influences the credit loan risk but predicting
models use this data to predict whether the particular case may be loan default case or not. However, from various models, there is no specific model which can be said as the most optimal model.

Currently, the various models which are frequently used for prediction purposes are the statistic-oriented models such as Discriminant Analysis (DA) and Logistic Regression (LR). Also, various different models which are used are Neural Networks (NN) or gene algorithms (GA). There are also various kind of non-parametric models which are based on artificial technology.

As earlier said, various loan default prediction models have already been implemented in order to curb the increasing loan default cases.

Sueyoshi (1999) [4] used mathematic programming viewpoint to create a predictor system. This system integrated the programming concept of the integer with data envelopment analysis in order to create new and innovative kind of Data Envelopment Analysis–Discriminant Analysis (DEA–DA), which not only provided model for institutional bankruptcy, but also provided the good predicting capability with high amount of precision.

Chen and Huang (2003) [5] analysed credit risk by using actual borrower’s data of UCI Database as the sample and analysed it on DA, Classification and Regression Tree (CART) and BPN to discover that each and every one of them has their own speciality.

Noh et al. (2005) [6] used data of credit card center of South Korea to create a new predicting model of credit risk while considering time-dependent characteristics and considered this data as predicting variables, and then adapted survival analysis (SA) to compare capability with LR and NN. While, LR and NN have better precision rate for good borrower, but SA got better sensitivity of predicting default borrower.

Lee et al. (2006) [7] conducted a research by using data of credit customers in Taiwan, compared different models created applied using CART, DA, NN, MARS and LR models and the research concluded that MARS and CART have better average accuracy for classification than the other 3 compared models.

Thus, DA and LR can be used to analyse the predicting variables that may influence loan default significantly. NN has better adaptability than other predicting models and this model is able to construct non-linear model and can better adapt predicting variables than other models.

In this research paper, we use data provided by Lending Club bank and to identify most significant attributes or features which may influence in credit risk default or loan default, PCA is used. By using PCA, most influencing variables or attributes are highlighted which can further be used as input variables in predicting model while the other less significant influencing characteristics are excluded. This leads to significant decrease in processing time and also improves efficiency and provides efficient system.

For predicting models, Deep Neural Networks is used due to its better adaptability and its ability to create non-linear model based on predicting variables provided to this model as input(s) which in turn are already sorted using PCA.

II. IMPLEMENTATION

We move on to the crux of the paper. The following research on loan-default prediction system which has been implemented in a slightly different way than most other prediction systems. While there is no dataset available of Indian banking system but at time of writing this paper, Lending Club Bank has provided the dataset. However, the dataset provided by Lending Club Bank is in raw and unstructured form with lots of dimensions, some of which are correlated to each other and hence, what we have done here is that we created our own dummy dataset with reduced dimensions and implemented the loan prediction system based on that.

Although quite the large amount of work is done in the field of Loan Default Prediction System, it is not efficient and accurate. Here, we are trying to implement the prediction system using Multilayer Perceptron with multiple hidden layers and multiple perceptron within a single layer. The reason for using Multilayer Perceptron was its better adaptability and its ability to create non-linear model based on predicting variables provided to the model as inputs which are already filtered and reduced using PCA.

We used PCA to reduce the number of dimensions to remove the problem of underfitting as well as reduce the amount of time and complexity required to train as well as generating output.

Thus, the implementation first reduces the number of dimensions using PCA and then uses Multilayer Perceptron model to make a loan default prediction system.

Following are the steps in implementation:

1. Preparing Dataset: From the provided dataset by Lending Club, we have modified it by decreasing the number of dimensions in the dataset for the implementation ourselves. It contains a total of 9578 rows with 14 original columns converted to 19 with dummy columns for the “loan_purpose” attribute.
This 18 perceptron are connected to 20 perceptron in next immediate layer where they provide their own calculated output along with connection weight which was randomly selected for each and every connection initially. There are 2 such hidden layers which contain 20 perceptron each which are also connected to their immediate next layers. This perceptron in hidden layers are also activated using same sigmoid function. The last hidden layer is connected to output layer and output layer provides the final output.

During training of model this output is used to calculate the connection weight which was randomly assigned initially.

Let there be an error in output node j in n^{th} data point.

\[ e_j(n) = d_j(n) - y_j(n) \]

Where d is the expected output
\[ y \] is the produced output

The node weights are adjusted in such a way that the entire output error is minimized given by

\[ E(n) = \frac{1}{2} \sum_j e_j^2(n) \]

The cost is calculated using the sigmoid cross entropy function of tensorflow and optimizer function is the Adam Optimizer.

5. Metrics: Finally, we studied the model and its capability with the help of performance metrics.

III. RESULTS

We were able to achieve an accuracy of around 93% which is fairly accurate for such a large dataset.

The model is implemented using a multilayer perceptron with 2 hidden layers of 20 nodes each and it has been trained for 1000 epochs.

---

**Fig-3:** Result obtained after training the Neural Network for 1000 epochs
IV. FUTURE SCOPE

The whole purpose of this research is to automate the banking process of selecting the loan applicants which are not risky for their bank or financial institution. In the future, we can develop full-fledged early warning systems which will help a bank or any other financial institutions to reduce their losses and increase their profits. We can increase the accuracy of our model by training it with datasets of banks of different countries so that our model would be able to incorporate region or community-specific parameters, that sometimes play a huge role in the case of loan default prediction. We will also try to reduce the time required to do the predictions so that users of the early warning system can get results in real time, which will increase the productivity of the users as they will be able to concentrate on other useful things as well.

V. CONCLUSION

Loan default prediction is done using the Multilayer Perceptron Model with Adam Optimizer as the optimization function. The proposed neural network model is tested for effectiveness using dataset provided by the Lending club bank. Principal component analysis was done on this dataset to reduce the number of dimensions present in the dataset. For the neural network model, its accuracy rate within the regular cut off has already reached the mark of 93%. The accuracy and precision of the model can further increase if we increase the number of hidden layers as well as the number of perceptron in each layer within a certain threshold value.

REFERENCES


