NEURAL NETWORK APPROACH TO LOAN DEFAULT PREDICTION

Manjeet Kumar[1], Vishesh Goel[2], Tarun Jain[3], Sahil Singhal[4], Dr. Lalit Mohan Goel[5]

[1,2,3,4] Students, CSE, Bharati Vidyapeeth's College of Engineering, New Delhi
[5]Prof, 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 uses 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 corelated 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.

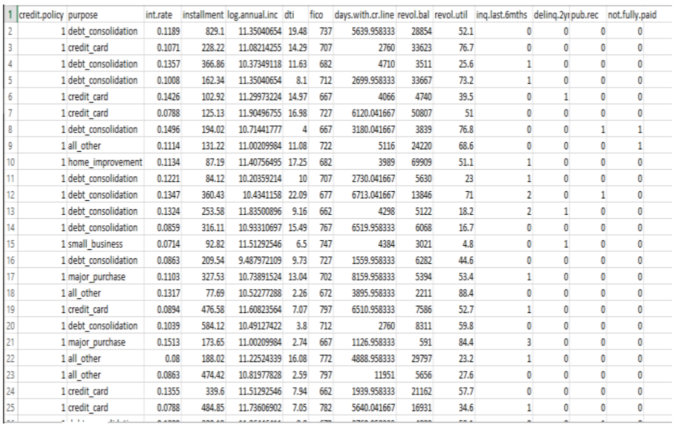
Although quite the large amount of work is done in the field of Loan Default Prediction System, it is not efficient and accurate. Here, we are trying to implement the prediction system using Multilayer Perceptron with multiple hidden layers and multiple perceptron within a single layer. The reason for 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:

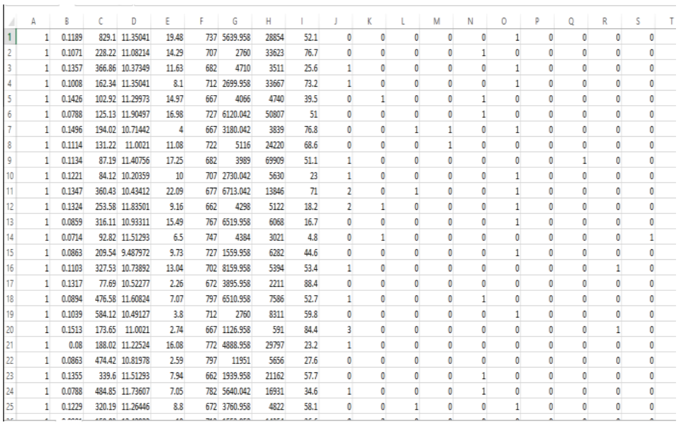
1. Preparing Dataset: From the provided dataset by Lending Club, we have modified it by decreasing the number of dimensions in the dataset for the implementation ourselves. It contains a total of 9578 rows with 14 original columns converted to 19 with dummy columns for the "loan_purpose" attribute.



1	credit.policy	purpose	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yr	pub.rec	not.fully.paid
2	1	debt_consolidation	0.1189	829.1	11.35040654	19.48	737	5639.958333	28854	52.1	0	0	0	0
3	1	credit_card	0.1071	228.22	11.08214255	14.29	707	2760	33623	76.7	0	0	0	0
4	1	debt_consolidation	0.1357	366.86	10.37349118	11.63	682	4710	3511	25.6	1	0	0	0
5	1	debt_consolidation	0.1008	162.34	11.35040654	8.1	712	2699.958333	33667	73.2	1	0	0	0
6	1	credit_card	0.1426	102.92	11.29973224	14.97	667	4066	4740	39.5	0	1	0	0
7	1	credit_card	0.0788	125.13	11.90496755	16.98	727	6120.041667	50807	51	0	0	0	0
8	1	debt_consolidation	0.1496	194.02	10.71441777	4	667	3180.041667	3839	76.8	0	0	1	1
9	1	all_other	0.1114	131.22	11.00209984	11.08	722	5116	24220	68.6	0	0	0	1
10	1	home_improvement	0.1134	87.19	11.40756495	17.25	682	3989	69909	51.1	1	0	0	0
11	1	debt_consolidation	0.1221	84.12	10.20359214	10	707	2730.041667	5630	23	1	0	0	0
12	1	debt_consolidation	0.1347	360.43	10.4341158	22.09	677	6713.041667	13846	71	2	0	1	0
13	1	debt_consolidation	0.1324	253.58	11.83500896	9.16	662	4298	5122	18.2	2	1	0	0
14	1	debt_consolidation	0.0859	316.11	10.93310697	15.49	767	6519.958333	6068	16.7	0	0	0	0
15	1	small_business	0.0714	92.82	11.51292546	6.5	747	4384	3021	4.8	0	1	0	0
16	1	debt_consolidation	0.0863	209.54	9.487972109	9.73	727	1559.958333	6282	44.6	0	0	0	0
17	1	major_purchase	0.1103	327.53	10.73891524	13.04	702	8159.958333	5394	53.4	1	0	0	0
18	1	all_other	0.1317	77.69	10.52277288	2.26	672	3895.958333	2211	88.4	0	0	0	0
19	1	credit_card	0.0894	476.58	11.60823564	7.07	797	6510.958333	7586	52.7	1	0	0	0
20	1	debt_consolidation	0.1039	584.12	10.49127422	3.8	712	2760	8311	59.8	0	0	0	0
21	1	major_purchase	0.1513	173.65	11.00209984	2.74	667	1126.958333	591	84.4	3	0	0	0
22	1	all_other	0.08	188.02	11.22524339	16.08	772	4888.958333	29797	23.2	1	0	0	0
23	1	all_other	0.0863	474.42	10.81977828	2.59	797	11951	5656	27.6	0	0	0	0
24	1	credit_card	0.1355	339.6	11.51292546	7.94	662	1939.958333	21162	57.7	0	0	0	0
25	1	credit_card	0.0788	484.85	11.73606902	7.05	782	5640.041667	16931	34.6	1	0	0	0

Fig-1: Sample Dataset

Modified dataset with dummy columns for with descriptive columns converted into categorical.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	1	0.1189	829.1	11.35041	19.48	737	5639.958	28854	52.1	0	0	0	0	0	1	0	0	0	0	
2	1	0.1071	228.22	11.08214	14.29	707	2760	33623	76.7	0	0	0	0	1	0	0	0	0	0	
3	1	0.1357	366.86	10.37349	11.63	682	4710	3511	25.6	1	0	0	0	0	1	0	0	0	0	
4	1	0.1008	162.34	11.35041	8.1	712	2699.958	33667	73.2	1	0	0	0	0	1	0	0	0	0	
5	1	0.1426	102.92	11.29973	14.97	667	4066	4740	39.5	0	1	0	0	1	0	0	0	0	0	
6	1	0.0788	125.13	11.90497	16.98	727	6120.042	50807	51	0	0	0	0	1	0	0	0	0	0	
7	1	0.1496	194.02	10.71442	4	667	3180.042	3839	76.8	0	0	1	1	0	1	0	0	0	0	
8	1	0.1114	131.22	11.0021	11.08	722	5116	24220	68.6	0	0	0	1	0	0	0	0	0	0	
9	1	0.1134	87.19	11.40756	17.25	682	3989	69909	51.1	1	0	0	0	0	0	0	1	0	0	
10	1	0.1221	84.12	10.20359	10	707	2730.042	5630	23	1	0	0	0	0	1	0	0	0	0	
11	1	0.1347	360.43	10.43412	22.09	677	6713.042	13846	71	2	0	1	0	0	1	0	0	0	0	
12	1	0.1324	253.58	11.83501	9.16	662	4298	5122	18.2	2	1	0	0	0	1	0	0	0	0	
13	1	0.0859	316.11	10.93311	15.49	767	6519.958	6068	16.7	0	0	0	0	0	1	0	0	0	0	
14	1	0.0714	92.82	11.51293	6.5	747	4384	3021	4.8	0	1	0	0	0	0	0	0	0	1	
15	1	0.0863	209.54	9.487972	9.73	727	1559.958	6282	44.6	0	0	0	0	0	1	0	0	0	0	
16	1	0.1103	327.53	10.73892	13.04	702	8159.958	5394	53.4	1	0	0	0	0	0	0	0	1	0	
17	1	0.1317	77.69	10.52277	2.26	672	3895.958	2211	88.4	0	0	0	0	0	0	0	0	0	0	
18	1	0.0894	476.58	11.60824	7.07	797	6510.958	7586	52.7	1	0	0	0	1	0	0	0	0	0	
19	1	0.1039	584.12	10.49127	3.8	712	2760	8311	59.8	0	0	0	0	0	1	0	0	0	0	
20	1	0.1513	173.65	11.0021	2.74	667	1126.958	591	84.4	3	0	0	0	0	0	0	0	1	0	
21	1	0.08	188.02	11.22524	16.08	772	4888.958	29797	23.2	1	0	0	0	0	0	0	0	0	0	
22	1	0.0863	474.42	10.81978	2.59	797	11951	5656	27.6	0	0	0	0	0	0	0	0	0	0	
23	1	0.1355	339.6	11.51293	7.94	662	1939.958	21162	57.7	0	0	0	0	1	0	0	0	0	0	
24	1	0.0788	484.85	11.73607	7.05	782	5640.042	16931	34.6	1	0	0	0	1	0	0	0	0	0	
25	1	0.1229	320.19	11.26446	8.8	672	3760.958	4822	58.1	0	0	1	0	0	1	0	0	0	0	

Fig-2: Processed Dataset

2. Pre-processing: After preparation of dataset, the data was divided into testing and training set. 90% data was used to train the model and the rest 10% to test it.

3. Selecting the algorithm: As explained at the earlier Multilayer Perceptron model of Deep Neural Network is used because of its ability to create non-linear model based on predicting variables provided to the model as inputs

4. Implementation of Algorithm on prepared data:

   Multilayer Perceptron is implemented.

   With reduced dimensions using PCA, the number of inputs provided to Deep Neural Network are reduced and 18 inputs have been provided to the model. These 18 inputs are provided to 18 perceptron of the input layer or first layer of the model which are non-linearly activated using sigmoid function.

$$y(v_i) = (1 + e^{-v_i})^{-1}$$

   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

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

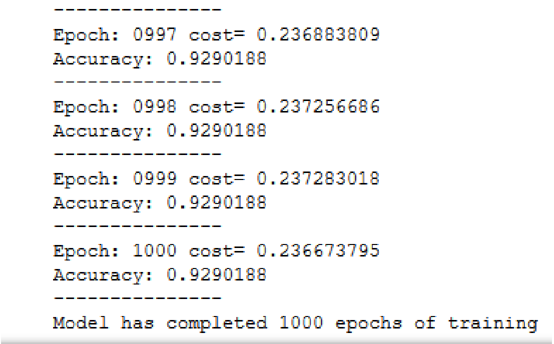
   The cost is calculated using the sigmoid cross entropy function of tensorflow and optimizer function is the Adam Optimizer.

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.



```
---------------
Epoch: 0997 cost= 0.236883809
Accuracy: 0.9290188
---------------
Epoch: 0998 cost= 0.237256686
Accuracy: 0.9290188
---------------
Epoch: 0999 cost= 0.237283018
Accuracy: 0.9290188
---------------
Epoch: 1000 cost= 0.236673795
Accuracy: 0.9290188
---------------
Model has completed 1000 epochs of training
```

Fig-3: Result obtained after training the Neural Network for 1000 epochs

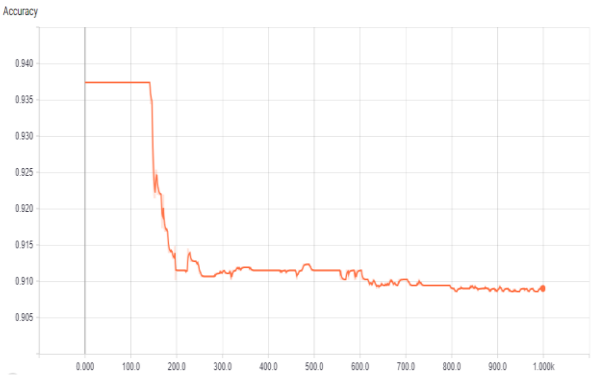


Fig-4: Variation of accuracy with training epochs

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

[1] Updegrave, W. L. (1987). How lender size you up. Money.

[2] Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. Insurance: Mathematics and Economics, 8(1), 31–34.

[3] Chiang, R. C., Chow, Y. F., & Liu, M. (2002). Residential mortgage lending and borrower risk: The relationship between mortgage spreads and individual characteristics. Journal of Real Estate Finance and Economics, 25(1), 5–32.

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

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

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

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