Predictive Risk Analysis For Loan Repayment of Credit Card Clients

Anirudh Bindal
Department of Computer Science and Engineering
Manipal University Jaipur, Rajasthan
anirudh312bindal@gmail.com

Sandeep Chaurasia
Department of Computer Science and Engineering
Manipal University Jaipur, Rajasthan
sandeep.chaurasia@jaipur.manipal.edu

Abstract— **Defaults on credit card loans lead to a lot of consumer debt and other major problems for banks and financial institutions. This paper compares five different data mining techniques to differentiate between those credit card customers who default on their credit card payments and those who don't. Some techniques use classifiers to identify whether a customer will default on his or her payment such as multi-layer perceptron classifier, while other techniques calculate the probability of defaulting such as logistic regression. Among the classifier techniques used, MLP classifier gives us the best accuracy of identifying the defaulters and Logistic Regression helps us identify some significant factors which have a direct impact on predicting whether a customer will be able to make his or her payment.**

Keywords—Multilayer Perceptron, Logistic Regression, Classifier, Data Mining.

I. INTRODUCTION

After the profits of the Taiwanese banks [1] stagnated due to the saturation of the real estate industry, who were their major customers, the banks thought of expanding their customer base to rake in more profits and in lieu of that they started issuing credit cards and encouraging more and more people to apply for them. Eventually, the youth of Taiwan became their target customers. Because of the low income of youth, customers started defaulting on their payments and by February,2006, debt due to credit cards and other cash cards was around 260 billion USD. This gave rise to many problems in Taiwan such increased suicide and rated and other illegal activities to repay the loans.

In retrospect, it seems that predictive or risk analysis of the customers before issuing them credit cards would have been much more productive and efficient and would have gone a long way to avoid such debt problems [2]. Banks now use many classification techniques such as naïve bayes and KNN to analyse risk prediction [3]. Whether or not a credit card should be issued to a client, they can calculate the probability of risk a client has of defaulting on his payments. Other way of tacking this problem can be identifying the significant factors which affect the risk factor of payment and representing them visually so as to get a picture of the situation as we have further done in this paper. Next, we will be analysing the five different data mining techniques, K Nearest Neighbours, Logistic Regression, Naïve Bayes Classifier, Multi-layer Perceptron Classifier and Decision Tree(CART).

II. LITERATURE

Supervised Learning

According to [4] supervised learning is application of those algorithms that form a hypothesis based on externally supplied data i.e. training data and then make future predictions on test data. Basically, the algorithms build a model on the basis of the initial information and then classify the unknown data according to the predictor attributes of the training data.

A. *K Nearest Neighbours Classfier*

K Nearest Neighbour is a non-parametric classifier and is one of the simplest classification techniques. Ref. [3] states that Non-Parametric means that it does not make any assumptions about the data and classifies it into groups based on some attribute. This is done on the training dataset. When we have want to classify the new data (test dataset), the algorithm compares the results of the training data set with the new data and makes predictions based on the K nearest neighbours of the data [5]. According to [6] the new observation is classified based on its K nearest neighbours.

The popular distance measure used for KNN classifier is Euclidian distance

$$\text{Euclidean Distance } (\chi, \chi\imath) = \chi, \chi\imath = \sqrt{\Sigma(\chi\imath - \chi\imath j)^2} \cdots (1)$$

Here x is the new point, x_i is the exiting point and j refers to all the input attributes. The K nearest neighbours are identified based on the distance calculated according to (1).

B. *Logistic Regression*

According to [7] Logistic Regression is a popularly used credit scoring and risk analysis method). It is a specialized case of generalized linear model [8],[9] and the output variable or the dependent variable is categorical/binary and the input variables can be continuous [10]. The classifier predicts the output by fitting the new data into the logit function. The generalized equation of the glm function is

$$\ln \frac{p(x)}{1 - p(x)} = \beta o + \beta\imath\chi \cdots (2)$$

Here p is the probability of the event being a success and B_0 and B_1 are parameters. Logistic regression gives us good

978-1-5386-2440-1/18/$31.00 ©2018 IEEE

results for this paper as it most accurately calculates the percentage of people who didn't default on their loans, and it also helps us in identifying significant factors which have an impact on the credit risk to be visually represented from a business perspective.

C. Naïve Bayes Classifier

According to [5] this classification technique is based on Bayes Theorem in probability and works to predict the class of the unknown dataset. It is relatively very fast as compared to other classification techniques and hence is useful when a large amount of data has to be classified in a time crunch. This technique also assumes independence between the predictors i.e. it considers all the input attributes in the dataset to be independent of each other. The generalized formulation for Bayes classifier according to [11] is:

$$\hat{y} = argmax\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \cdots (3)$$

Here y is the classifier, $x(x_{1\ldots}x_n)$ is the input vector which represents n features and C_k is the given classes for each of k possible outcomes. According to [11] the Bayes classifier model combines with a decision rule and selects the hypothesis with the most probability. The limitations of this techniques are that it assumes all the attributes to be independent which is almost impossible in real life [5]. Also, if a categorical variable has a category which was unobserved in the training dataset, it will be unable to make a prediction and assign the value 0 to it.

D. *Multi-layer Perceptron Classifier*

MLP classifier is a feed forward artificial neural network consisting of at least three layers of nodes, with one input layer, one or more intermediate layer and one output layer [12]. According to [13] the MLP function is as shown in the figure:

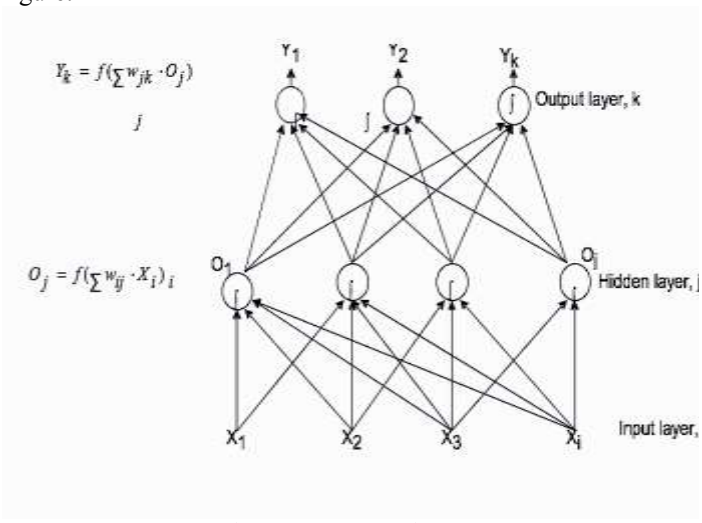

Figure 2.1 MLP Function.

This function is learned by training on the training dataset. According to [14], the sigmoid activation function is

$$f(t) = \frac{1}{1 + e^{-t}} \cdots (4)$$

This classifier gives us the best results among all the classification techniques with an accuracy of 88%. The best part about this classifier is its ability to learn or fit non-linear model and its capability to work in real-time situations as proved by [15]. MLP uses back propagation algorithm for instance classification [16]. Also, MLP requires fine tuning of large number of parameters. MLP produces best results when the scaling of data is done.

E. *Classification And Regression Trees(CART)*

This Decision tree algorithm is recursive partitioning algorithm which splits each input node into child nodes [17]. At each level of the decision tree, the algorithm identifies which variable and level to split in to further child nodes. Here the splitting is done using a greedy approach. This means that all the splits are tried and tested using the cost function and the split with the best cost (min cost) is selected. The splitting procedure needs to know when to stop splitting through the dataset. Usually, a minimum count is set to stop the procedure. Here we use the depth as minimum count. CART algorithm also provides good results on this dataset.

III. Experimental Setup

A. Database

This dataset set is from the Taiwanese bank crisis of the year 2006. It has been used previously by [18] for calculating credit risk for clients. *The dataset has 30000 thousand entries out of which 23364 have successfully paid their loans and the remaining 6636 are defaulters.*

Attribute Information:

Database is available from link [19]: This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

-X1: Amount of the given credit (NT$): it includes both the individual consumer credit and his/her family (supplementary) credit.

-X2: Gender (1 = male; 2 = female).

-X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

-X4: Marital status (1 = married; 2 = single; 3 = others).

-X5: Age (year).

-X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:

-X6 = the repayment status in September, 2005;

-X7 = the repayment status in August, 2005;

. .;

-X11 = the repayment status in April, 2005.

The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for

two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

-X12-X17: Amount of bill statement (NT$).

-X12 = amount of bill statement in September, 2005;

-X13 = amount of bill statement in August, 2005; . . .;

-X17 = amount of bill statement in April, 2005.

-X18-X23: Amount of previous payment (NT$).

-X18 = amount paid in September, 2005;

-X19 = amount paid in August, 2005; . . .;

-X23 = amount paid in April, 2005.

B. Dataset Features

The significant factors identified via logistic regression in the data were age, repayment status in September 2005, repayment status in august 2005 and amount of bill statement in September. The size of the red circles in the graph represent the size of people in that particular region corresponding to the x and y axis. The bigger the circle, the more number of people it represents.

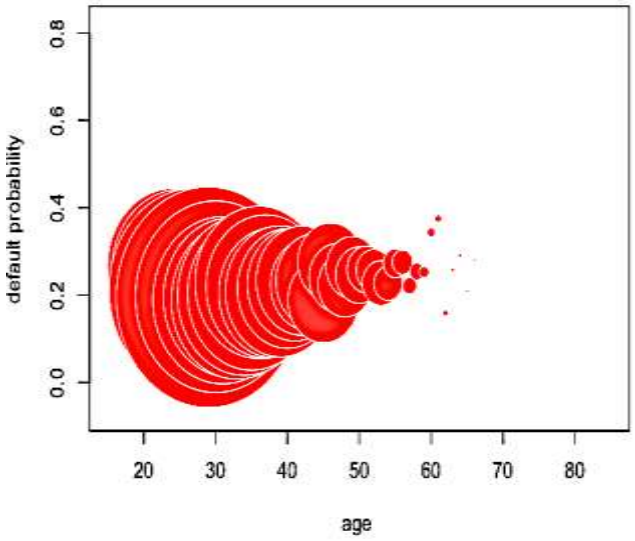

Figure 2.2 This graph depicts the relationship of age with defaulting probability. The lower the age, lesser income, hence more chances of defaulting

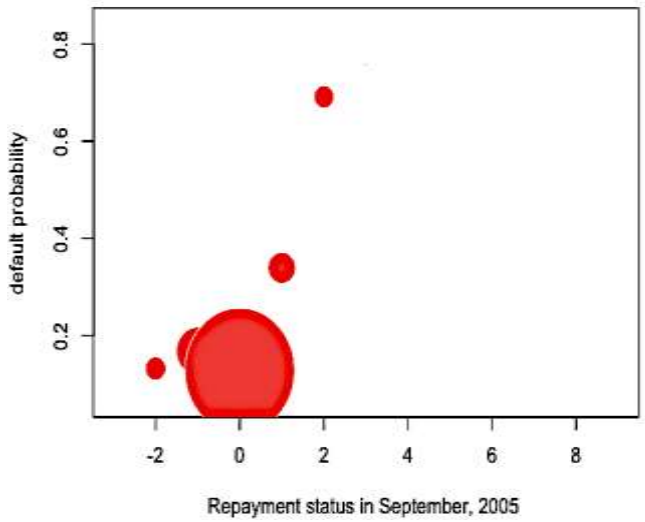

Figure 2.3 This is a representation of the repayment status in September to the probability of defaulting. As it can be clearly seen, more the number of months payment is delayed, greater is the default probability

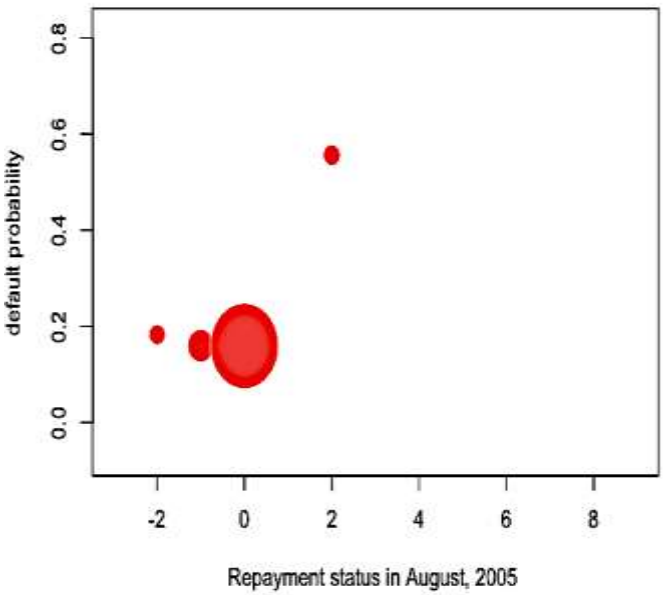

Figure 2.4 This is a represents the repayment status in August to the probability of defaulting. As it can be clearly seen, more the number of months payment is delayed, here is a steep increase in the default probability

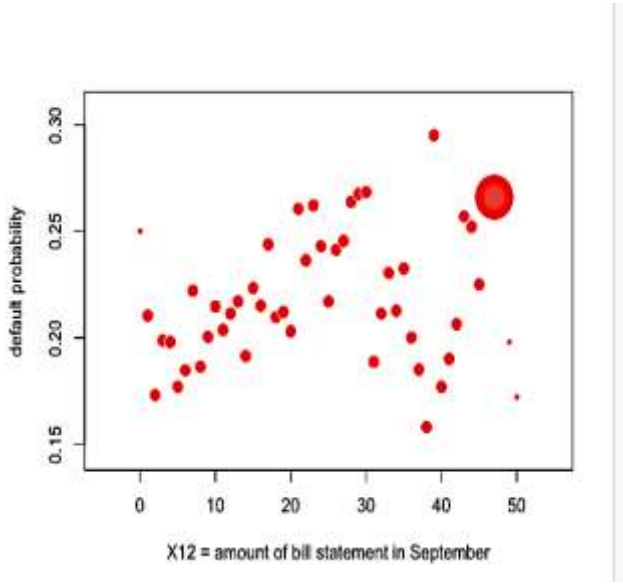

Figure 2.5 This depicts the amount of bill statement in September. As it can be clearly seen, more the amount of bill statement, greater is the default probability

Clearly it can be seen from the age graph that lesser the age of the client, the more chances he or she has on defaulting on his or her payment. This is quite a good analysis in fact because the majority of the youth in Taiwan had very low income back in 2005 and hence were unable to pay off their debts. The other thing that can be seen from the repayment status graphs visually is that those who had made their payments on time for the respective months of august and September had significantly low probability of defaulting than those people who had a delay in their repayment status. Finally, the last graph shows us that the more amount of bill statement people had in September higher was their probability of defaulting on their loans. Visually this type of representation of data is of significant helps to banks when assessing the credit risk of a client From a business perspective, bank officials have absolutely no idea about the working of an algorithm, but visual representation such as this helps them understand the intricacies of the problem.

C. Algorithms' Setup

Table 2.1 This table shows all the model parameters used for different classifiers.

CLASSIFIER	Model Parameters
Logistic Regression	*C=1.0* *max_iter=100* *solver='liblinear'*
Naïve Bayes	*alpha=1.0*
MLP	*activation='tanh'* *alpha=1* *beta_1=0.9* *beta_2=0.999*
	hidden_layer_sizes=(10,10) *max_iter=1000* *solver='lbfgs'*
Decision Tree(CART)	*min_samples_split=1000* *min_samples_leaf=1000*
KNN	*leaf_size=30* *n_neighbors=5*

D. Results

As done in the paper by [18] we use the area ratio terminology to calculate the classification accuracy because 77% of the clients in the data are non-risky i.e. they have paid their debts, and therefore this leads to error insensitivity in the classification accuracy. The accuracy of each model is estimated by calculating the area between the base curve (ROC curve of the category) and the best curve (micro-average ROC curve).

KNN classifier gives us an area ratio of 0.75 while the naïve Bayes classifier has an accuracy of 0.77. MLP classifier gives us the best performance with the area ratio of 0.88 with CART algorithm close behind with 0.85. Logistic regression has an area ratio of 0.77.

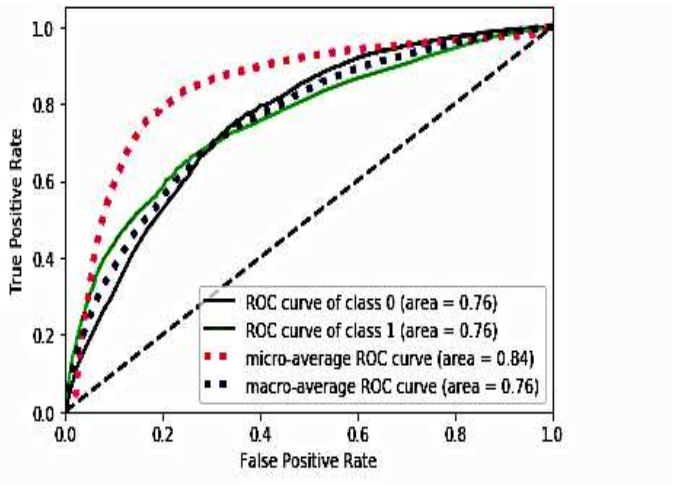

Figure 2.6 ROC curve for Naïve Bayes classifier with an area ratio of .85

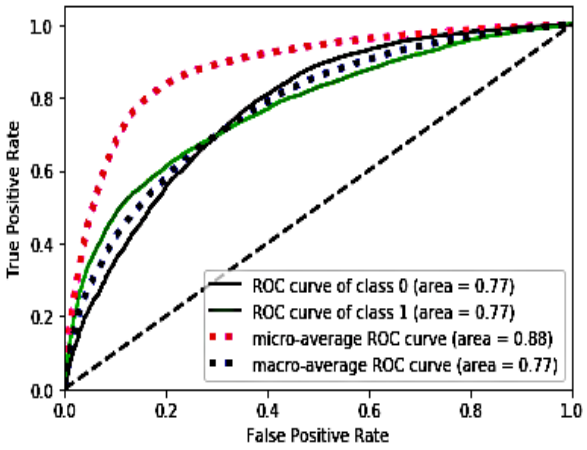

Figure 2.7 ROC curve for MLP classifier with an area ratio of .88

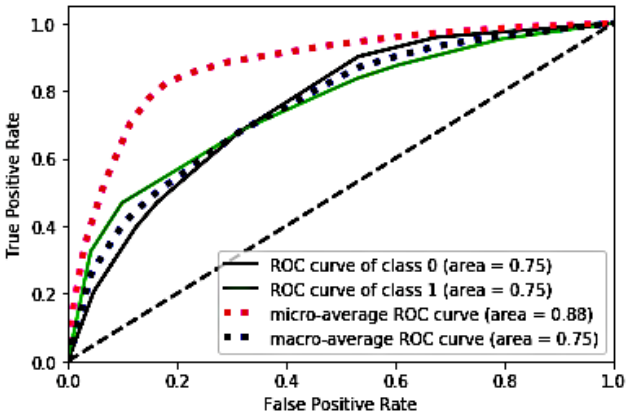

Figure 2.8 *ROC curve for CART algorithm with an area ratio of .85*

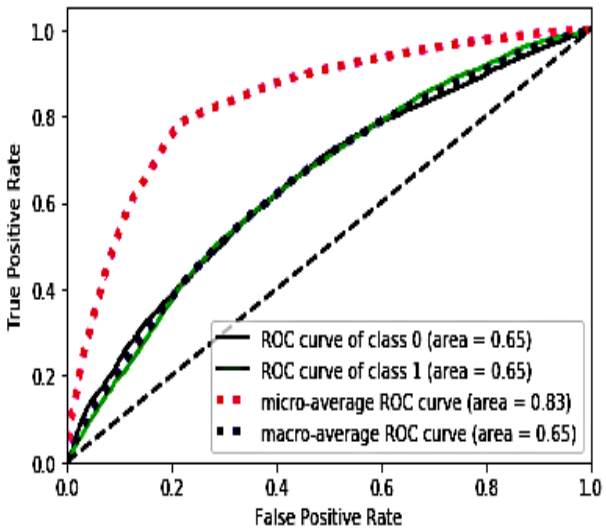

Figure 2.9 ROC curve for Logistic Regression classifier with an area ratio of .77

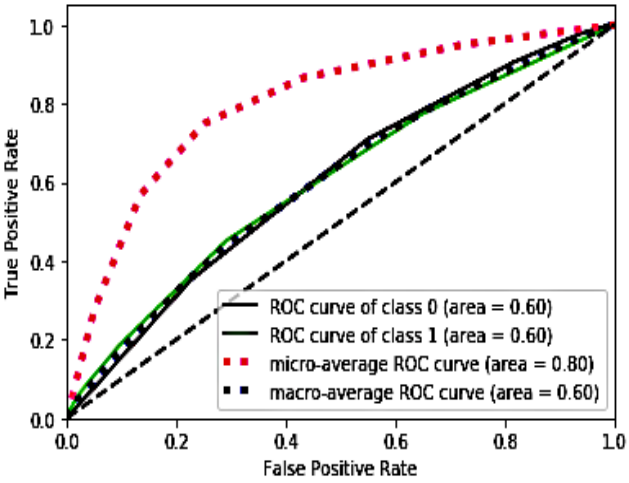

Figure 2.10 ROC curve for KNN classifier with an area ratio of .75

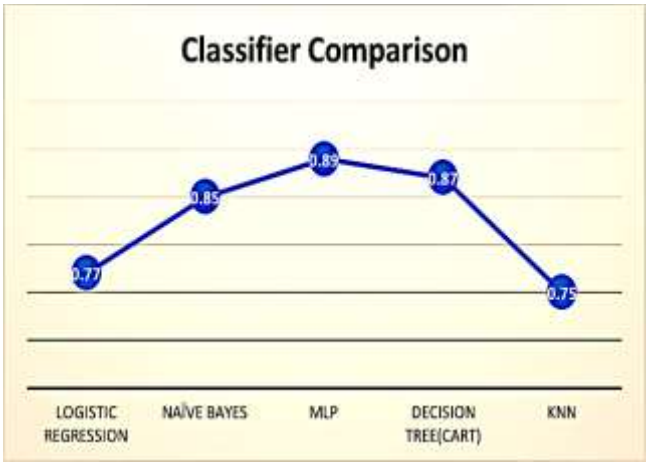

Figure 2.11 This is a graphical comparison among the classifier techniques used.

Table 2.2 This table compares the area ratio of the classifiers used in this paper.

CLASSIFIER	AREA RATIO
Logistic Regression	*0.77*
Naïve Bayes	*0.85*
MLP	*0.89*
Decision Tree(CART)	*0.87*
KNN	*0.75*

IV. Conclusion

Clearly among the 5 data mining techniques that were compared in the paper, MLP classifier gave us the best performance with an area ratio of .88. Apart from that logistic Regression also helped us in identifying the significant factors which are important for analysing the credit risk of a client and thus representing it visually with the help of graphs. The novelty of this paper is that apart from using the scientific classification algorithms to calculate the credit score or credit risk accuracy, we have also used these methods to identify the

significant factors which impact the credibility of loan repayment of a client and represented them visually with the help of graphs which is extremely crucial for a business audience as they don't understand the scientific methods or techniques used, but can easily relate to a graphical representation.

References

[1] https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/.
[2] Koh, H. C., & Chan, K. L. G. (2002). Data mining and customer relationship marketing in the banking industry. Singapore Management Review, 24(2), 1–27.
[3] Berry, M., & Linoff, G. (2000). Mastering data mining: The art and science of customer relationship management. New York: John Wiley & Sons, Inc.
[4] S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007) 249-268
[5] Pranab Kumar D. G., Radha Krishna, P., (2013) «Database management system oracle SQL AND PL/SQL" PHI Learning Pvt. Ltd., 576 pages.
[6] Keramati, A., N. Yousefi, 2011. A Proposed Classification of Data Mining Techniques in Credit Scoring.International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.
[7] Rosa, P. T. M. (2000) Modelos de Credit Scoring: Regressão Logística, CHAID e REAL. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
[8] Dobson, A. (1990) An Introduction to Generalized Linear Models. London: Chapman & Hall.
[9] Paula, G. A. (2002) Modelos de Regressão com Apoio Computacional. Material disponível em http://www.ime.usp.br/~giapaula/livro.pdf acesso em 05/12/2004.
[10] Gouvêa, M. A., & Gonçalves, E. B. (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. Paper presented at the Production and Operations Management Society (POMS), Dallas, Texas, U.S.A.
[11 https://en.wikipedia.org/wiki/Naive_Bayes_classifier.
[12] Pacelli, V., & Azzollini, M. (2011). An Artificial Neural Network Approach for Credit Risk Management. Journal of Intelligent Learning Systems and Applications, 3(2), 103-112.
[13] http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/kms/4-MLP.pdf
[14] Subrata Saha,& Sajjad Waheed. Credit Risk of Bank Customers can be Predicted from Customer's Attribute using Neural Network. International Journal of Computer Applications (0975 – 8887) Volume 161 – No 3, March 2017.
[15] Mehdi Khashei * and Akram Mirahmadi. A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification. Int. J. Financial Stud. 2015, 3, 411-422; doi:10.3390/ijfs3030411
[16] Lakshmi Devasena C. Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 4, April 2014.
[17] Bowen Baker. Consumer Credit Risk Modeling. MIT Departments of Physics and EECS, 70 Amherst Street, Cambridge, MA 02142 (Dated: December 17, 2015).
[18] I-Cheng Yeh a,*, Che-hui Lien b. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36 (2009) 2473–2480
[19] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients