IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Logistic Regression for Detection of Bankruptcy

Sagar Kumar[1*], Shubhajit Mukherjee[2], Shubham Agrawal[3], Ila Chandrakar[4]

[1,2,3]School of Computing and Information Technology, REVA University, Bangalore, India
[4] School of Computing and Information, REVA University, Bangalore, India

DOI: https://doi.org/10.26438/ijcse/v7si14.131133 | Available online at: www.ijcseonline.org

Abstract— Bankruptcy is a legal procedure that claims a person or organization as a debtor. It is essential to ascertain the risk of bankruptcy at initial stages to prevent financial losses. In this perspective, different soft computing techniques can be employed to ascertain bankruptcy. This study proposes a bankruptcy prediction system to categorize the companies based on extent of risk. The prediction system acts as a decision support tool for detection of bankruptcy.

Keywords— Bankruptcy, soft computing, decision support tool

I. INTRODUCTION

Machine learning is evolved from artificial intelligence which is a very recent and important field of computing technology. It permits our computer system to make analytical models of knowledge and see unknown insights automatically, whereas not being unambiguously coded, it has been applied to various aspects such as in stylish society, ranging from DNA sequences classification, master card fraud detection, mechanism locomotion, to communication method. It will be used to solve many problems/tasks like classification. Bankruptcy detection is a common example of all classification problems.

II. RELATED WORK

Pattern recognition has given birth to machine learning. Machine learning existing work which is used for bankruptcy detection is done using different techniques like logistical regression, genetic rule, and inductive learning algorithms. Logistic regression may be a way permitting researchers to create prophetic operate supported a sample. This model give idea to deduce single outcome variable from several independent variables [1]. Logistic regression algorithm is very useful in some applications, but it also has many limitations. In Genetic rule, natural selection is done and then evolution happens. Extraction of first order rules are generated by Genetic algorithm and then these rules are taken for complex classification problems [2]. Inductive learning is done using decision tree algorithm. It takes some information for training to gain knowledge about existing patterns and then test the other set of data to generate standard rules used for problem solving [2]. To determine if the accuracy of bankruptcy detection is often more improved, we tend to suggest 3 current models—support vector machine, neural networks, and auto encoder. SVM may be a supervised learning technique which is particularly efficient in problems of high dimensions and is memory economical as a result of which it uses a set of coaching points in the decision function. It also specifies kernel operates consistent with the choice function [3]. Its mathematical property guarantees a straightforward convexo-convex improvement downside to converge to one world downside. Neural network generates models that improves learning by existing problems. They work on various hidden layers at the same time within the existing data, therefore they are very efficient in learning even very complex connection between inputs and outputs. And that they operate considerably quicker than typical methods**.** However, the size of given data is small and limited, accuracy will be less due to problem arise of overfitting. To overcome this, a way referred to as dropout—temporarily and every which way clears units (hidden and visible)—to the neural network [4]. Auto encoder, conjointly called Diabolo network, is associate degree unattended learning rule that orders the target values in order to be capable of the inputs.

Since this is done, it overcomes the calculation of rendering several functions that help improving correctness. Also, the quantity of coaching knowledge needed to find out these functions are reduced [5]. This paper uses the concept of logistic regression for bankruptcy prediction.

III. METHODOLOGY

The various recently developed models which can be used for machine learning are SVM, neural network, logistic regression etc and it is used in many applications. Logistic regression is very useful technique for applications like text grouping, classification of cancer disease, object tracking etc. Text grouping is particularly useful in our everyday life—web looking out and email filtering offer vast convenience and work potency.

Neural networks gain from the examples rather than algorithms; therefore, they are used to solve issues wherever

it is laborious or not possible to use algorithmic ways [6]. As an example, finger print recognition is one of the useful application of all time People nowadays uses distinct fingerprints as keys to unlock their phones and payment accounts, so it solves the problem of remembering long passwords for protection.

Basically logistic regression is the suitable statistical procedure to regulate once the variable is split (binary). Logistical regression is used to clarify information and to elucidate the link between one dependent binary variable and one or plenty of nominal, ordinal, interval or ratio-level freelance variables. Occasionally logistical regressions area unit troublesome to translate; the Intellects Statistics tool simply permits you to manage the analysis, then in plain English interprets the output.
Below are the assumptions for Binary logistic regression are:

1. The dependent variable should have either of two values: (e.g., presence vs. absent).
2. The outliers present in the data should be deleted from the data set. This can be done by the conversion of continuing predictors to standardized scores and by deleting values above 3.29 and below -3.29.
3. Correlation must be low between the predictors.

This can be evaluated by using a matrix among the predictors. Tabachnick and Fidell (2013) counsel that as long correlation coefficients among freelance variables are but zero.90 the assumption is met. At the middle of the provision multivariate analysis is that the task estimating the log odds of a happening. Mathematically, provision regression estimates a multiple regression toward the mean performance outlined as:
Logit p = ln (p/1-p)

Overfitting. When choosing the model for the supplying multivariate analysis, one more necessary thought is that the model work. Attaching freelance variables to a logistical regression model can perpetually expand the quantity of variance described within the log odds (commonly expressed as R^2). However, adding a lot of and a lot of variables to the model may end up in overfitting that decreases the generalizability of the model on the far side the information on that the model is work.

Reporting the R2: Many number and varieties of pseudo-R2 values have been developed for binary logistic regression.
This ought to be understood with extreme caution as they need several machine problems that cause them to be by artificial means high or low. A better approach is to gift any of the goodness of work checks available; Hosmer-Lemeshow may be a normally used live of goodness of work supported the Chi-square test. Figure 1 represents the linear and logistic regression.

Logistic regression: Logistic regression is basically a supervised classification algorithm .It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables

Logistic regression is a technique borrowed by machine learning from the field of statistics.It is the go-to method for binary classification problems.

Logistic regression can be classifies as:-

1.Binomial: Target variable can have only two possible types:"0" or "1" which may represent "win" vs "loss" , "pass" vs "fail", "dead" vs "alive" etc.

2. Multinomial: Target variable can have 3 or more possible types which are not ordered. For example-"disease a" vs "disease b" vs "disease c"

3.Ordinal: It deals with target variables with ordered categories. For example-A test score can be categorized as "very poor" ,"poor" ,"good" ,"very good".

Logistic Function: Logistic regression is named for the function used at core of the method, the logistic function

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S shaped curve that can take any real valued number and map it into a value between 0 and 1, but never exactly at those limit.

1/ (1+e^-value)

Where e is the base of the natural logarithms and is the actual numerical value that you want to transform.

Linear regression: Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task.Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting

Linear regression performs the task to predict a dependent variable value(y) based on a given independent variable(x).So, this regression technique finds out a linear relationship between x(input) and y(output). Hence the name is Linear regression.

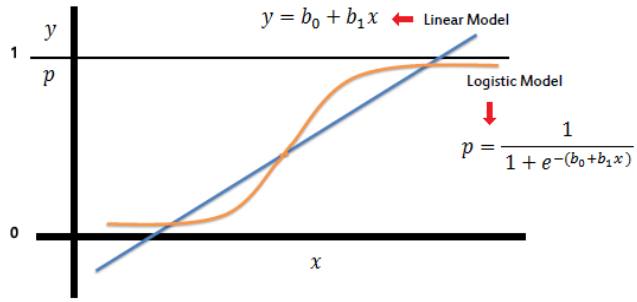

Figure 1: Logistic regression and Linear regression

IV. RESULTS AND DISCUSSION

The average AUC score of the best model, isolation forest, is about 0.95. The higher AUC shows that both precision and recall of the algorithm are also high at the same time. Besides, the results illustrate that random forest gained the best results for the reference supervised approaches. The result of logistic regression is as shown below in the form of confusion matrix.

Table 1: Result

		Target			
		Positive	Negative		
Model	Positive	727	2110	Positive Predictive Value	0.256257
	Negative	384	4772	Negative Predictive Value	0.925524
		Sensitivity	Specificity		
		0.654365437	0.69340308		Accuracy
					0.687977

V. CONCLUSION AND FUTURE SCOPE

Support vector machine and neural network with dropout, and autoencoder unit of measurement has three relatively latest models used in bankruptcy detection problems. Their accuracies overcome the accuracies of the three older models like robust offer regression, inductive learning algorithms and genetic algorithms. The better aspects embrace the management for overfitting, the better chance of searching the globe maxima, and the capacity to hold big feature areas. This paper used and finished the progress of machine leaning models related to bankruptcy detection and checked the performance of comparatively latest models within the conditions of bankruptcy detection that have not been used in that field. Somehow, the 3 models even have disadvantages. SVM does not directly offer the chance evaluate but rather uses a modern five-fold cross-validation.

REFERENCES

[1] Hauser, R.P. and Booth, D., "Predicting Bankruptcy with Robust Logistic Regression", Journal of Data Science , Vol 9, pp. 565-584, 2011.
[2] Kim, M.-J. and Han, I. , "The Discovery of Experts' Decision Ruels from Qualitative Bankruptcy Data Using Genetic Algorithms", Expert Systems with Application , Vol 25, pp. 637-646, 2003.
[3] Pedregosa, et al., "Scikit-Learn: Machine Learning in Python", Journal of M achine Learning Research , Vol 12, pp. 2825-2830, 2011.
[4] Sirvastava, N., et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting, "Journal of Machine Learning Research, Vol 15, 1929-1958, 2011.
[5] Dev, D., "Deep Learning with Hadoop", Packet Publishing, Birmingham, 52, 2017.
[6] Nielsen, F., "Neural Networks—Algorithms and Applications", https://www.mendeley.com/research-papers/neural-networks-algorithms-applicatio ns-5/
[7] Robinson, N., "The Disadvantages of Logistic Regression. http://classroom.synonym.com/disadvantages-logistic-regression-8574447.html "
[8] Sima, J. (1998) Introduction to Neural Networks. Technical Report No. 755.
[9] Baldi, P. (2012) Autoencoders, Unsupervised Learning, and Deep Architectures. Journal of Machine Learning Research , 27, 37-50.
[10] Martin, A., Uthayakumar, J. and Nadarajan, M. (2014) Qualitative Bankruptcy Data Set, UCI.
[11] Ramosacaj, Miftar & Hasani, Vjollca & Dumi, Alba. (2015). Application of Logistic Regression in the Study of Students' Performance Level (Case Study of Vlora University). Journal of Educational and Social Research.
[12] Abedin, Tasnima & Chowdhury, Mohammad & Afzal, Arfan & F, Yeasmin & Turin, Tanvir. (2016). Application of Binary Logistic Regression in Clinical Research. Journal of National Heart Foundation of Bangladesh.

