International Journal of Advanced Engineering and Technology
ISSN: 2456-7655; Impact Factor: RJIF 5.54
Received: 14-04-2019; Accepted: 17-05-2019
www.newengineeringjournal.com
Volume 3; Issue 3; July 2019; Page No. 31-35

Machine learning application to identify good credit customers

Carol Anne Hargreaves
Department of Statistics and Applied Probability National University of Singapore, Singapore

Abstract
In this paper, we focus on the application of machine learning algorithms to identify credit risk customers. The main objective is to firstly, identify the most important factors that may be associated with "good credit" customers, and to compare good credit customers with bad credit customers. The logistic regression model was used to score customers on their likelihood of being a "good credit" customer. The logistic regression model was highly accurate in predicting credit risk with a sensitivity and specificity of 74% and 75% respectively. Therefore, by following the classifications of the logistics regression model results, the bank was able to minimize losses of -$88 000 to a profit of $26 000.

Keywords: machine learning algorithms, logistic regression, credit risk, factors, accuracy, losses, profit

1. Introduction

Customers who default on loans can be a huge cost to banks, leading to a huge loss and drop in profit. Thus, we would like to use supervised machine learning methods to help banks minimize their loss and maximize their profit. Using a supervised machine learning technique, we would like to use a prospective applicants demographic and socio-economic details to predict whether the applicant is likely to default on the loan. Other than using a machine learning model to predict the prospective actions of the applicant, we would also like to gain insights on applicants who default and those who do not. These insights will help loan managers and banks to make more informed decisions when processing customer loans.

From the perspective of the bank, there are two possible types of losses that can be incurred should they make a wrong decision in approving a loan application. Firstly, the loss of profit and potential interest from rejected applicants with a good credit rating. Secondly, the financial loss born by the bank from approving loan applicants with bad credit ratings who are likely to default on their loan. Thus, it is crucial that the bank correctly classifies an applicant's profile so as to minimize the risk of approving a loan to a bad credit applicant and thereby maximize the profits earned.

The key objective of this study is to develop a good decision rule to classify the good creditors from the bad creditors. There are three performance measures looked at; the accuracy, the sensitivity and the specificity of the model. While a model with a high accuracy is desired, having sensitivity value of at least seventy percent is also important. We want the bank to minimize the risk they are approving applicants with actual bad credit ratings. In this case, the monetary loss and effort spent in following up with a defaulter is more than the opportunity cost of missing out on a potential credit-worthy applicant. In addition, we wish to identify the significant attributes that are indicative of a good credit customer.

This paper is structured into 5 sections. While Section 1 is the introduction, Section 2 gives a brief description of the data exploration, Section 3, a brief overview of the methods used, Section 4 the analysis results, Section 5, gives the impact of the logistic regression model results on the bank, after which Section 6 presents the conclusion.

Data Exploration

We used the real-world German Credit Risk data set for our analysis [1]. In total, this dataset contains 1000 observations with 31 variables. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases). New applicants for credit can also be evaluated on these 30 "predictor" variables, 6 of which are categorical, 6 of which are numerical and the rest being binary. The output RESPONSE is a binary variable, indicating whether the applicant has a good credit rating (1 for good, 0 for bad).

Categorical variables:
- Account Status
- Credit History
- Average balance in savings account
- Present Employment since
- Present resident since
- Nature of job

Numerical variables:
- Duration of credit
- Credit Amount
- Instalment rate as of percentage of disposable income
- Age
- Number of existing credits at this bank
- Number of people being liable to provide maintenance for

Binary variables:
- Purpose of credit - New Car?
- Purpose of credit - Used Car?
- Purpose of credit - Furniture?
- Purpose of credit - Radio/TV?
- Purpose of credit - Education?
- Purpose of credit - Retraining?
- Male and divorce?
- Male and single?
- Male and married/widowed

- Co-applicant present?
- Guarantor present?
- Owns a Real estate?
- Owns no property/unknown?
- Other instalment plan credit?
- Applicant rents?
- Owns a residence?
- Owns a telephone?
- Foreign worker?

We want to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. We first split our data into 70% Training and 30% Test using stratified sampling. Next, we used bar plots to find out which are the main contributors to credit rating.

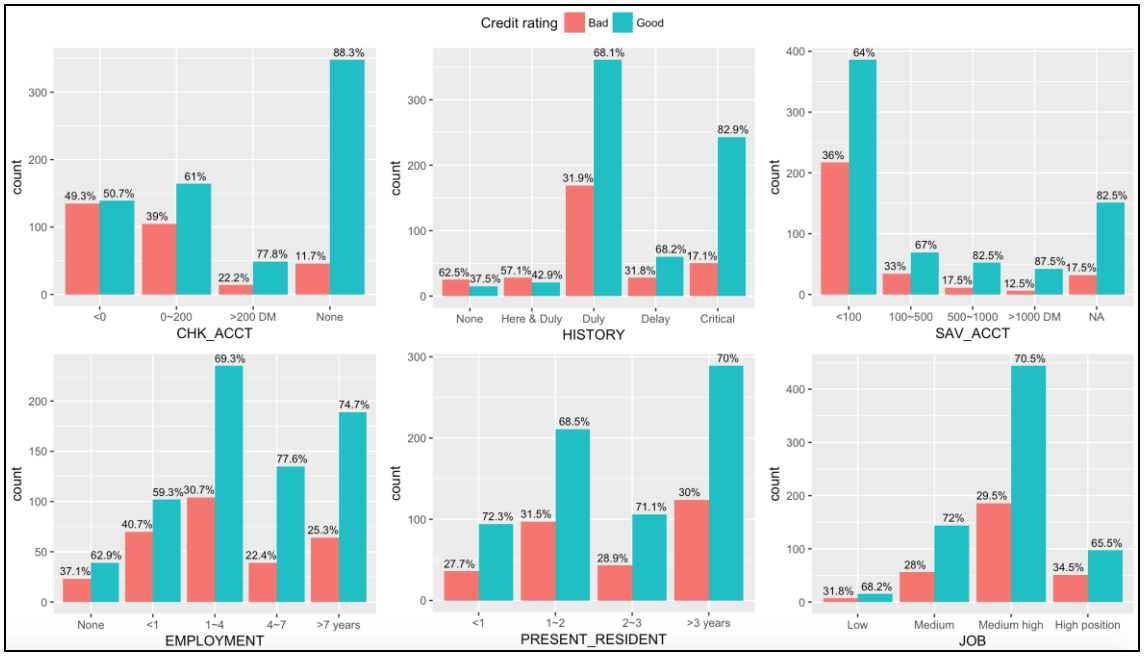


Fig 1: Data Exploration for Categorical Variables

Across these categorical variables, we can see that the proportion of good and bad credit record varies. From figure 1 above, suggest that checking account status, credit history and saving account balance are possible factors that influence credit rating, because there is a significant difference in the percentages of good and bad credit ratings across all levels of a categorical variable.

More concretely, we speculate that a higher balance in checking account favours a good credit rating. Chance of an applicant having a good credit rating is even higher if he or she does not have a checking account. However, for predictors like job, the percentage of a good rating is approximately 70% across all 4 categories. Therefore, it is likely that job is an insignificant variable, which means that it does not really matter whether an applicant is doing an unskilled job, skilled job, or being in a company's management position when assessing his or her credit.

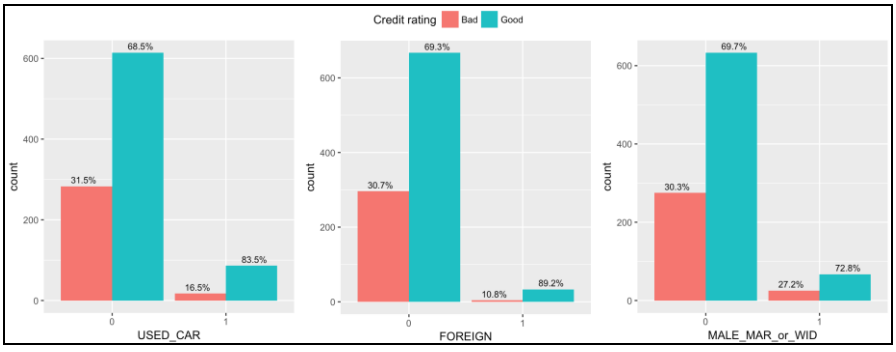


Fig 2: Data Exploration for Binary Variables

From figure 2 above, shows that people buying used car are a more credible than people with other purposes for loan. Similarly, although foreigners constitute only a little of the applicants, 89.2% of them have a good credit, while only 69.3% of local residents do. As a comparison, the variable of applicant being male and married or a widower does not seem to affect his rating because the percentages of good credits over level 0 and 1, 69.7% and 72.8%, are very similar.

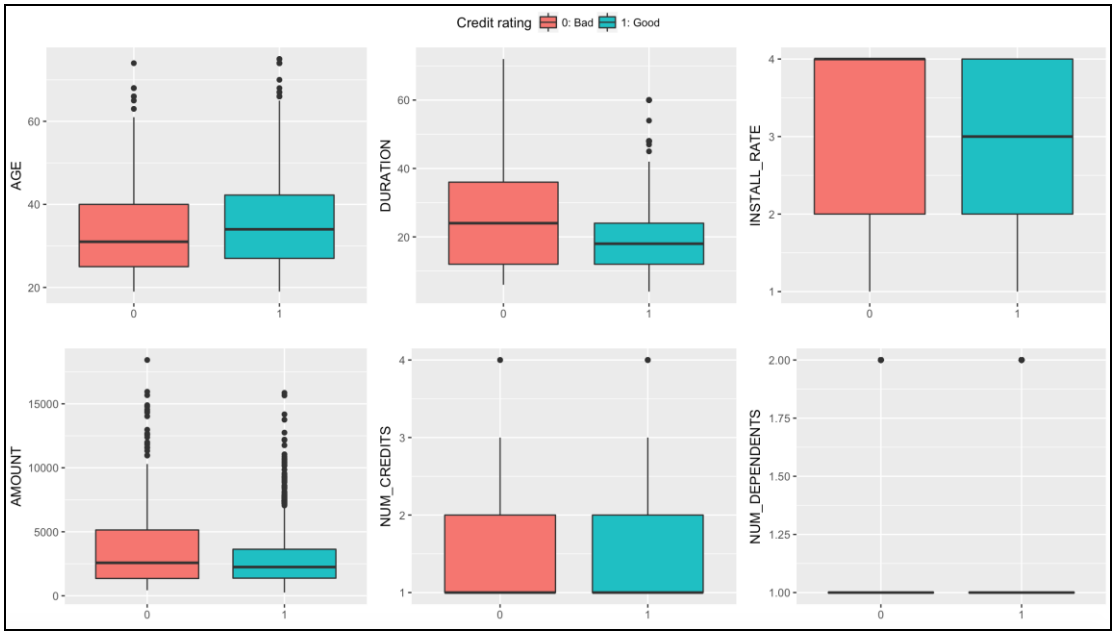


Fig 3: Data Exploration for Numerical Variables

From figure 3 above, people with good credit rating have a lower average instalment rate and shorter duration than people with bad credit rating. This could suggest that a high instalment rate as percentage of disposable income is likely to be a huge burden on the borrowers that drives them to Delay or not pay back the money. Duration is also likely to be a factor that influences credit rating, as longer duration of credits will be a higher risk for the bank.

Finally, we would like to check for collinearity and multi-collinearity before building model. According to the correlation plot below (figure 4), only OWN_RES and RENT (-0.74) have an absolute correlation greater than 0.7. We remove the input variable 'Rent' for our logistic regression modelling.

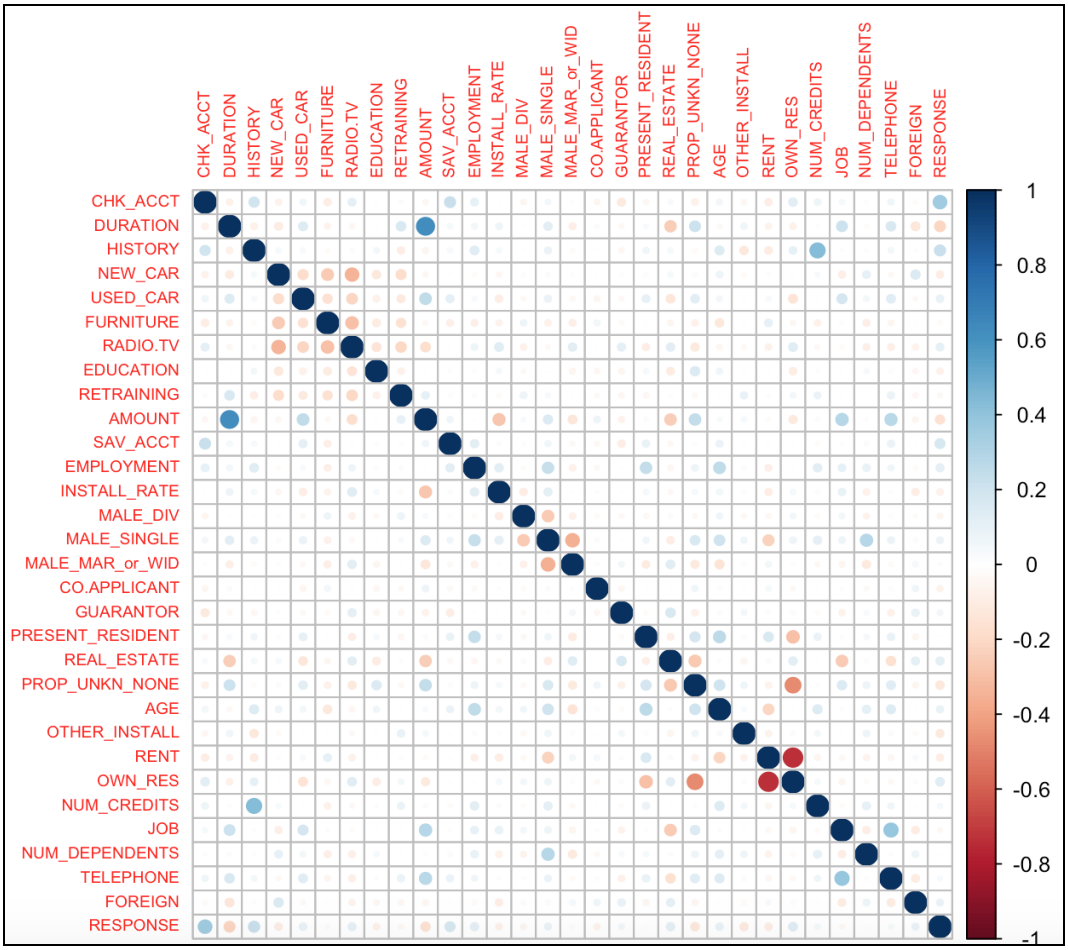


Fig 4: Data Exploration Correlation Plot

Method Used

The machine learning technique

Logistic regression model

The logistic regression model was selected for use in this study because it is the most basic and robust classification algorithm. The logistic regression is a special type of regression where the binary response variable is related to a set of explanatory predictor variables which can be continuous or discrete. The logistic regression model is perfect for situations where the aim is to predict r presence or absence of a characteristic or outcome based on the values of a set of predictor variables. The relationship between the target and input variables is not always a straight line, and so a non-linear or logistic regression model is used.

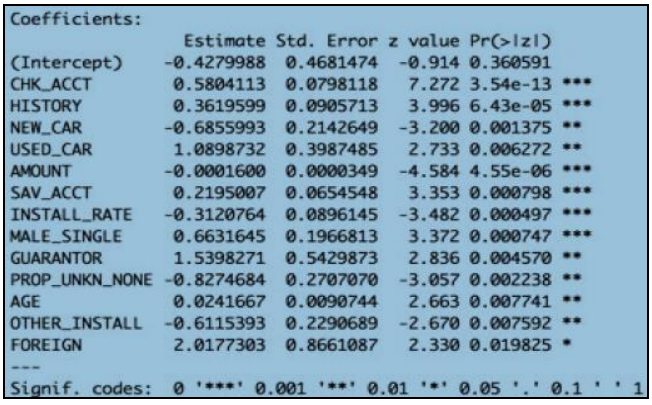
Further, the Logistic Regression model was chosen as it required little running time compared to other complicated machine learning algorithms and its output is also easy to interpret.

The main assumption for building a logistic regression model is that the independent variables do not have significant multi-collinearity [2]. Initially, there were thirty input variables. After multi-collinearity was handled, we were left with twenty-nine input variables for our Logistic Regression Model and was carried out using the backward stepwise regression method. Variables with a p-value greater than 0.05 was deemed as insignificant to "Good Credit" and was removed from the model. The process stopped when the model was left with variables with p-values less than 0.05.

Results and Findings

Logistic Regression Analysis Results

Insights about the input variables and their effect on 'Good Credit' was gained by understanding the signs and size of the coefficients of the Logistic Regression model and the resulting odds ratio. The size of the coefficient indicates the main drivers of "Good Credit". The larger the coefficient, the higher the significance of the variable to "Good Credit". A positive coefficient indicated that the predictor variable had a positive relationship with "Good Credit".



```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.4279988  0.4681474  -0.914 0.360591
CHK_ACCT        0.5804113  0.0798118   7.272 3.54e-13 ***
HISTORY         0.3619599  0.0905713   3.996 6.43e-05 ***
NEW_CAR        -0.6855993  0.2142649  -3.200 0.001375 **
USED_CAR        1.0898732  0.3987485   2.733 0.006272 **
AMOUNT         -0.0001600  0.0000349  -4.584 4.55e-06 ***
SAV_ACCT        0.2195007  0.0654548   3.353 0.000798 ***
INSTALL_RATE   -0.3120764  0.0896145  -3.482 0.000497 ***
MALE_SINGLE     0.6631645  0.1966813   3.372 0.000747 ***
GUARANTOR       1.5398271  0.5429873   2.836 0.004570 **
PROP_UNKN_NONE -0.8274684  0.2707070  -3.057 0.002238 **
AGE             0.0241667  0.0090744   2.663 0.007741 **
OTHER_INSTALL  -0.6115393  0.2290689  -2.670 0.007592 **
FOREIGN         2.0177303  0.8661087   2.330 0.019825 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig 5: Logistic Regression Coefficient Summary

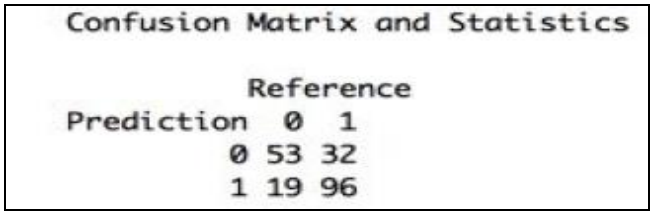
"Foreign" (2.0), "Guarantor" (1.5) and "Used Car" (1.1) had the highest positive relationship with "Good Credit".

An added advantage of using the Logistic Regression is that it calculates the logarithmic odds of "Good Credit". The odds ratio is calculated to study the effect of each variable in affecting the odds of "Good Credit" [4]. The odds ratio for "Foreign", "Guarantor" and "Used Car" was 7.4, 4.5 and 3.0 respectively. This means that holding all other variables constant, customers who are "Foreign" are 7.4 times more a "Good Creditor" than customers who are not "Foreign". Similarly, customers have a "Guarantor" are 4.5 times a "Good Creditor" than customers who do not have a "Guarantor". Similarly, customers who have a "Used Car" are 3 times a "Good Creditor" than customers who do not have a "Used Car".

Other than understanding how different variables would determine a "Good Creditor", there is a need to evaluate the training model to find out its predictive power. This is done by measuring the accuracy of the model. Accuracy measures such as the confusion matrix, recall, precision, specificity, overall accuracy, were calculated and are shown below.

The confusion matrix measured the ability of the logistic regression model to classify customers correctly as "Good Creditors" and "Bad Creditors". From the confusion matrix, 53 and 96 customers are classified correctly as "Bad Creditors" and "Good Creditors" respectively. However, 32 "Good Creditors" was classified as "Bad Creditors" and 19 "Bad Creditors" was classified as "Good Creditors". TABLE 1 below shows the "Sensitivity" and "Overall" Accuracy for classifying customers as "Good Creditors" and "Bad Creditors".

Table 1: Confusion Matrix Table



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53 32
         1 19 96
```

Our logistic regression model performance had an overall accuracy of 74.5%, sensitivity of 73.6%, and specificity of 75%. This is pretty good as they are all above 70%. We know that accuracy of 74.5% means the rate the training model

Correctly classifies the customers is 74.5%. And sensitivity of 73.6% or true positive rate of 73.6% means that 73.6% of the positive instances are correctly identified by the model. The specificity of 75% means that 75% of the negative instances are correctly identified by the model. The area under the receiver operating curve (AUC) was 0.7833116. So, we can conclude that the logistic regression model has good predictive ability.

Foreign, Guarantor and Used Car had the highest contribution in identifying customers as "Good Credit" or "Bad Credit". Perhaps, the banks usually require a higher income range for foreigners for a credit card to be issued

To them. Most of these foreigners with higher income may be highly-educated, so there is a high probability for them to pay back the credit on time.

Also, for "GUARANTOR", if the applicant has a guarantor, if the customer does not pay back the credit, the guarantor may need to pay back for him, so it makes sense why the variable "Guarantor" a high contribution to identifying "Good Creditors". Having a "Guarantor" puts more pressure on the customer to pay back his credit on time.

For USED CAR, we know that buying a used car rather than a new car can save money. To save the most money, you usually would buy a good used car, make sure its condition is still good, check the engine, maybe do a test drive, and

need to buy the car at the right time, as the car prices vary a lot. Customers who buy a "Used Car" are usually well organized and do a lot of work before buying a used car. This means that they are the kind of people who will plan their credit repayments well. Thus, customers who have a "Used-Car" may be more reliable, and more likely to pay back the credit.

As we can see that the variables "Foreign", "Guarantor" and "Used Car" do influence the identification of a "Good Creditor". Therefore, we are happy to conclude that the logistic regression model is highly acceptable for identifying whether new loan customers will be good creditors.

Impact of Logistic Regression Model

Results on the Bank

How do the results of the logistic regression model impact the bank? What does the bank have to gain from the model results?

If we calculate the business value, the value for the bank, based on the 200 unseen customer transaction data, we may assume that the average bank earnings per credible customer is $1000 and the average bank losses per non-credible customer is $2000 and that the bank loans are given to all predicted good customers and not the predicted bad customers, then looking at the confusion matrix in table 1, from our final logistic regression model, we have the following 4 classes of applicants:

C1: Actual credible applicants predicted as credible = 96

Earnings through C1: 96 x $1000 = $96 000

C2: Actual credible applicants predicted as non-credible = 32

Earnings through C2: 32 x -$1 000 = - $32 000

C3: Actual non-credible applicants predicted as no credible=53

Earnings through C3: $0

C4: Actual non-credible applicants predicted as credible = 19

Earnings through C4: 19 x -$2000 = -$38 000

Total Earnings = $26 000

Therefore, by following the classifications of the logistics regression model results, the bank was able to minimize losses of -$88 000 to a profit of $26 000.

Conclusion

The outcome from this study, confirms firstly that machine learning techniques can be used to predict which customers are likely to be "Good Credit" and which customers are likely to be "Bad Credit" as the logistic regression model performance of identifying "Good Creditors" on the unseen data was very good, 73.6%. In a nutshell, machine learning applications are valuable for investors as the analytical results can accurately identify "Good Creditors" and the bank was able to reduce losses of -$88 000 to produce a profit of $26 000.

There are other benefits for the bank besides monetary benefits as with the machine learning application for identifying "Good Creditors", the bank was able to approve loan customers much quicker, thereby saving time for other important business developments and decision making.

In addition, with the new machine learning application for approving customers for credit loans, the customer experience is enhanced as loan are approved much quicker, making the customer happy with the banks service. With this improved bank service, the bank is more likely to gain more new customers as word gets out that loan approvals are timely and efficient.

Lastly, the bank itself, will have more profit and therefore, will have more money to invest in innovation and be able to beat their competitors in customer service and new bank technology products. These new bank technology products will enhance the bank's productivity and sustainability in an environment that is strongly disruptive where survival of the fittest is dependent of data driven technology-based products, like the machine learning credit scoring application presented in this paper.

References

1. The German Credit data set (available at ftp. ics. uci. edu/pub/machine-learning-databases/stat log/)
2. Menard SW, NetLibrary I. Applied logistic regression analysis. Thousand Oaks, Calif: Sage Publications.
3. Tsai CF, Wang SP. Stock Price Forecasting by Hybrid Machine Learning Techniques. Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009; 978(988):17012-20
4. Hosmer DW, Lemeshow S. Sturdivant RX. Applied Logistic Regression (third ed.) Hoboken, N.J: Wiley.