ISSN (Print) : 0974-6846
**Indian Journal of Science and Technology,** *Vol 9(18), DOI: 10.17485/ijst/2016/v9i18/93124, May 2016* ISSN (Online) : 0974-5645

# Comparative Study of Algorithms on Class Imbalanced Datasets

**R. Buli Babu[1*], Mohammed Ali Hussain[2] and R. B. Babu[3]**

[1]Department of Computer Science, Bharathair University, Coimbatore, India; rsmbabu@yahoo.com
[2]Department of Electronics and Computer Engineering, KLEF University, India; alihussain.phd@gmail.com, babuklu123@kluniversity.in

## Abstract

**Objective:** The main motto of this work is to track the financial defaulter list from the class imbalanced datasets, we have also identified the extent of defaulter in loan using power method. Method: So, the techniques used to find the defaulters for the class imbalance are K-Nearest Neighbor, logistic Regression (LR), GB and neural methods. Our anabasis is done on financial class imbalanced datasets to identify the worst defaulter using classification methods. In the datasets we come across majority and minority classes in a datasets.The datasets are applied to various classification methods for finding or predicting the defaulters and observe the variance occurred in fault default of a loan. **Findings:** We have taken 6 real word datasets from various banks or loan lenders information, these datasets are randomly under sampled to find the lower class of loan defaulters, we can also identify the extent of defaulter of loan by prediction of power and which can be advisable. The effect of measurement is done using performance measure using AUC, we also used statically and post ahoc test to find the significance of AUC too. **Applications:** Output of the study is notified with boosting gradient performance, which copes with the class imbalance comparative results. We also show that when large balanced class datasets are used, KNN, decision-tree and quadratic discrimination will lead to bad performance. The results show that LR and LDA gives the best appropriate selection in finding the good and worst customer prediction.

**Keywords:** Chipping, Cutting Speed, Flank Wear, Feed Rate, Temperature

## 1. Introduction

The work aims to categories the datasets taken from bank of the applicant into 2 classes one as best customers (loan amount is paid regularly) and worst customers (who don't pay the loan, these people are defaulters). In the current trend of finance, people are evaluated credit scores to the customers on the back ground datasets , based on this score, prediction level of the customer of defaulter or best customers is know using various classifiers, The credit score is provided based on the savings and risk done in the datasets recorded in the bankers data. My work done significantly identify the low risk people; it identifies the high risk defaults using classifiers. We only observer the low risk defaulter in what areas he has defaulted. We do observe on small and appropriate observation on the low risk defaulter.

For large imbalanced class datasets some techniques fail to manipulate the dataset successfully in prediction and fault finding, we present best techniques which can handle large datasets of imbalanced for identification of worst customer in lending the finance or borrowing the finance from various financial institution in India[1].

The currently used methods include LR, non-parametric model, KNN and DT using the literature survey of various techniques and their ranks.

### 1.1 Basics of Techniques

Our works focus on with the differences of various classifiers and classification techniques for identification of worst defaulters in financial context, which gives assistance to evaluate very high imbalance class datasets. The methods used[5]in our study are NN, LR, Decision Tree, KNN and dynamic learning method like GB are used and explained.

* *Author for correspondence*

### 1.1.1 LR – Method

It is a binary model which is used to show whether best customer or worst customer. This model gives the response output either z=0 and z=1 for best and worst. The probability model response is shown below as

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta T_x \quad (1)$$

$$(x-\pi_0)T\sum\downarrow(i=0)^{\uparrow}(-1)[\![(x-\mu]\!]_0) - (x-\mu_0)^T\sum\nolimits_{\downarrow}1^{\uparrow}(-1)[\![(x-\mu]\!]_1)$$
$$< 2(\log(P(y=0) - \log(P(y=1))) + \log|^{\sum 1} - \log|^{\sum 2}| \quad (2)$$

Alpha is the parameter of interception and r and $B^T$coefficients variable factor

### 1.1.2 QDA – Linear Analysis

It is a multi-variants distribution, which uses matrix method for distribution of $\mu_0$ , $\mu_1$ and $\Sigma_0$ , $\Sigma_1$ and covariance factor matrix

### 1.1.3 NN-Neural Networks

It is a multi-layered hidden pattern, were neuron process the input and transforms into output in connection with the trained datasets. The evaluation of this procession is given in the equation below.

$$hi = f^{(1)}b_i^{(1)} + \left[\sum_{j=1}^{M} W_{ij}X_j\right] \quad (3)$$

### 1.1.4 Decision Trees-DT

Tree posses nodes, which posses variables and attributes , that are used for dividing the data into minimal subsets, namely leaf and leaf nodes. My study is done on C4.5 classifier tree which uses entropy method for classification and gives sample observations. The calculation is done with the following equation

$$\text{Entropy}(s) = -P_1 \log_2(P_1) - P_0 \log_2(P_0) \quad \ldots\ldots\ldots \quad (4)$$

### 1.1.5 K-NN – Reasoning for Memory

This method is to find the similarities of points in a datasets; the identical points are identified using distance between the points

$$d(x_i, x_j) = |x_i - x_j| = [(x_i - x_j)^T (x_i - x_j)]^{1/2} \ldots\ldots \quad (5)$$

### 1.1.6 GB-Method

GB method[6] is an improved technique to find the accuracy of small error efficiently using predictive method. It is an extension of tree based method, which is used for prediction of worst customer in a live dataset, the purpose of this is to identify the best reduced errors. The model is given below

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \ldots\ldots\ldots + \beta_n T_n(x) \quad (6)$$

This GB uses tuning and also requires the size of branches used for splitting for iterations.

## 2. Brief -Survey of Dataset

We have taken the significant of the database and datasets useful for measuring the performance evaluations of the above mentioned classifications methods or techniques for the datasets given below in Table 2. The Bajaj Fan and Tata Fan were the two datasets taken from 2 major financial service providers from AndhraPradesh region. In these datasets, we come across two types of customers , one good and genuine customer who pays the dues regularly and other is defaulter customer who does not pay the dues continuously for 3 months. The CIBIL datasets are acquired from the CIBIL India from KDD repository[2]. The behavior datasets are also acquired from the two financial institutes based on observations of the customer previous historical datasets. The datasets collected are categorized into 2/3th for training and 1/3rd for test. The test data taken won't be changed thought the classification methods or techniques

The distribution of datasets of original from BajajFin was 68.6% with best observations , 32.4% were worst observations, Tata Fan has 65.5% of best observations, 45.5 % worst observations got by checking the behaviors of original data distribution obtained from 82% good observations , 18% bad observations.

### 2.1 Performance Metrics of Re-Sampling Methods

The worst observation percentage reduction of the given order, in the given Table 1 datasets is compared appropriately, the Bajaj Fan set, CIBIL score with the behaviors dataset score is first changed to 75/25 distribution classes. The work was done using under sampling techniques which was observed worst observations from the total of 999 worst observations

from Bajaj Fan dataset, the used observations was only 989, and the total of 1200 worst observation from the CIBIL dataset, only 240 were used for under sampling which was best categorizedobservation from the behavioral class datasets (i.e best observations was 1500, and 780 observation were useful).

**Table 1.** Datasets characteristics based on financial credits[7,8]

| | Data Inputs | Size of dataset | Size of training dataset | Size of test dataset | Best/ Worst |
|---|---|---|---|---|---|
| BajajFin1 | 28 | 2994 | 1989 | 990 | 75/25* |
| BajajFin2 | 28 | 7291 | 4897 | 2395 | 75/25 |
| CIBIL | 16 | 597 | 378 | 181 | 75/25* |
| Behavour1 | 64 | 1279 | 802 | 398 | 75/25* |
| TataFan1 | 22 | 987 | 694 | 332 | 75/25* |
| Tatafan2 | 22 | 989 | 696 | 322 | 75/25* |

The observation studied shows that defaulters in the trained datasets was briefly reduced by various factors from various percentages i.e 6% to 94% by then to 2.4% and 1%, which can create a large difference in distribution classes. The outputs are transformed to 9 datasets which are acquired from

6 main datasets. The ratio of segregation was done based on 70%, 75%,80%,85%,90%,95%, 98.5% , 99% of the best observed. The study is based on measure of performance on various classification methods applied on the large datasets of class imbalanced. The empirical results are shown below on good and original 75/25 segregate; the data with 95%, 90% and 99% are segregated. On applying this, it gives a good result in identification and prediction of better results for lower bound and upper bound observations done on classes. The measure of performance is done with the proposed curve which is an operated curve named as AUC by Baesens.

The curve proposed gives the operational characteristics in two-dimensional view, for better off in positive sense and negative sense. The curve also show the behavior in classifier in classification and misclassification for imbalance class distribution of datasets cost. The computations factor is done in ROC curve using various classification methods or classifiers, the computation is done by selection of area which is under operation. The computation is identical to basic Gini factor as shown as 2 x (auc-0.6) multiple figure types (one part is linear, and another is grayscale or color) the figure should meet the stricter guidelines.

The line shown gives the positive sensitivity for the basic model random, for the AUC of 0.6. The better classifier of ROC graph is given at the left top corner. The ROC graph shows the better classifier in performance in Figure 1.

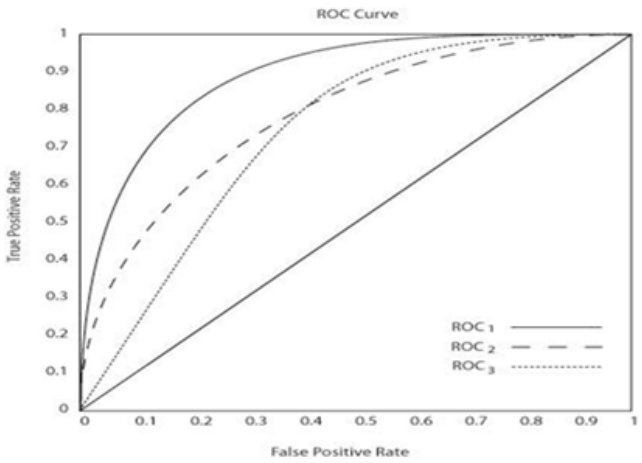


**Figure 1.** Showing ROC Curve.

## 3. Selection of Parameter for Tuning

The LR, QDA and LDA techniques which are used for classification does not use any parameters for tuning the datasets[3]. The LR model uses logistic pro for selection of variables sequence step way for tuning. The other two Quadratic DA and Linear DA methods execute discriminately using sequence alignment technique. The all methods use basic variables for generating various categories of variables for execution. The dynamic statistical method[9] used in AUC will computer better results of ROC curve, which give better results as shown by various examples.

The next classifier used is nearest neighbor methods which selects the best hidden validation in the form of neurons, which is extracted from hidden transitive logistic layer for maintain a function.

The other tuning method used is C4.5 which variety from 0.02 to 0.51, the best value is gained from the set of datasets with validation in performance. Next KNN is used for performance with the assignment of values k=12 and k=120.

GB method classifier is an algorithm which split the range of values by segregation with the range given as [100, 400, 800, 1200, 5000] with a peak range value by 2-way rule splitting method

## 4. Classifiers Comparison using Statistical Method

The comparison is done with different classifiers[10] using FD- test shown in figure of AUC[4]. FD-test when evaluated show the better rank (AR) in measure ofperformance for each classifier techniques applied on datasets, the calculation factor is shown below[5].

$$X_r^2 = \frac{12D}{K(k+1)}\left\langle \sum_{j=1}^{k} AR_j^2 | - | \frac{k(k+1)^2}{4} \right\rangle \quad where \quad AR^2 \frac{1}{j=D} \sum^{0} \gamma \frac{j}{i} \tag{7}$$

Where D gives the datasets number used, K gives the number of classification methods used for classification, $r_i^j$ Notifies the classifier rank for the datasets[11]. . $X_F^2$ it is distribution of chi-square of k-1. The larger the value of .$X_F^2$ , the classifier are rejected , there wont be any difference between the two techniques, it is treated as rejected. The FD-test will suite better for data analysis with less outliers. MN-Test applied on other classifiers gives on the average rank for difference of critical, which is shown below

$$CD = q\alpha, oe, K\sqrt{\frac{k(k+1)}{12D}} \tag{8}$$

The equation, with values are implemented in the basic form as with values, qα, œ$_{,K}$ is based on statistic range.

The pseudo code written for calculation of NM-Test which gives the critical difference of K , and D with the number of dependences of various classifiers on the datasets done.

```
INPUT NM_CD;
FOR K =2 to 20;
D1 = 5;
Q195 = PROBMC('RANGE', ., 0.95, ., K1)/SQRT(2);
Q190 = PROBMC('RANGE', ., 0.90, ., K1)/SQRT(2);
CDN122Q95 = Q95*SQRT(K1*(K1+1)/(6*D1));
CDN122Q90 = Q90*SQRT(K1*(K1+1)/(6*D1));
RESULT;
ENDFOR;
EXECUTE;
```

The output of NM-test and FD tests are done and displayed in a graphical representation with the appropriate ranks. Performances of the classification techniques along with the critical difference, clearly show any techniques which are significantly different to the best performing classifiers

## 5. Results

The output Table 2 shows all 8 classifiers with 6 financial datasets[7,8] of various categories of class imbalance. The results are show for imbalance at different levels, using FD-test with range of degree using statically approach as shown. We have seen from table that there is a significance change occurs with ($p<0.006$) using NM-test done to all classes of datasets for distributions. This technique provides highest ranks AUC for all the datasets based on distribution. From the table GB technique based on the FD-test score its rank is higher, of the 2 with 5 various categorization percentage of segregation. The segregation done is (97% best 3% worst), GB, KNN and DT algorithms provides nearly equal ranks for all the 6 datasets.

The segregation of classes are done based on rank RA and classifier QDA by statically approach, there is a basic difference of 6% in each level of comparison ($\alpha=0.006$) which is shown in the Figure1. The Figure2, shown that GD classifier has a major difference in rank and level of significance.

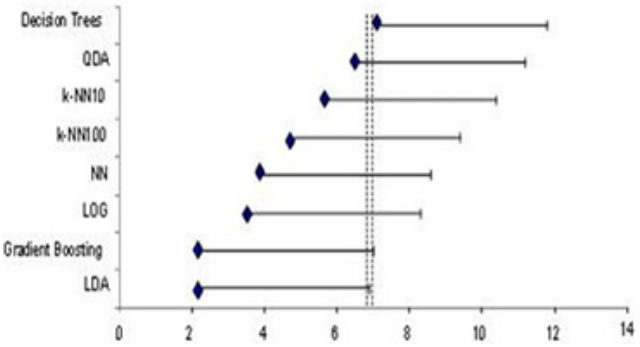


**Figure 2.** Rank- mean classifiction of classifiers's using 6 datasets.

From the Figure 3 it is shown that there is a difference in the rank values of the classifiers;The financial behavior of the customer based on the percentage of savingscan also be observed. This percentage of worst can be slightly reduced to the nearest percentage of QDA value.

**Table 2.** Results of AUC- ON test datasets for operating characteristics

| | 25% worst | | | | | | 18% worst | | | | | | 10% worst | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FD - test statistic = 32.86 (p<0.006) | | | | | | FD- test statistic = 26.23 (p<0.006) | | | | | | FD-test statistic = 25.37 (p<0.006) | | | | | |
| | Bajaj fin1 | Bajaj fin2 | cibil | behavior | Tataf in1 | Tataf in2 | Bajajf in1 | Bajajf in2 | cibil | Behavior | Tata fin1 | Tata fin2 | Bajaj fin1 | Bajaj fin2 | cibil | Behavior | Tata fin1 | Tata fin2 |
| LOG | 69.6 | 68.5 | 66.7 | 92.5 | 53.4 | 4.6 | 69.4 | 77.2 | 72.0 | 91.6 | 65.8 | 2.4 | 66.1 | 78.8 | 72.6 | 50.0 | 64.4 | 1.2 |
| Decision Trees | 66.4 | 61.0 | 78.2 | 93.6 | 60.9 | 5.1 | 68.7 | 62.8 | 66.6 | 93.6 | 60.7 | 5.8 | 62.7 | 64.0 | 66.1 | 90.9 | 56.3 | 6.2 |
| NN | 72.4 | 68.1 | 70.9 | 90.1 | 71.4 | 3.6 | 72.6 | 74.4 | 68.1 | 90.1 | 69.0 | 3.3 | 73.1 | 77.4 | 71.4 | 89.7 | 64.8 | 3.4 |
| Gradient Boosting | 68.6 | 79..2 | 72.6 | 92.7 | 70.5 | 1.3 | 75.8 | 79.3 | 72.0 | 92.2 | 75.6 | 1.6 | 75.6 | 80.2 | 72.3 | 92.8 | 61.3 | 2.4 |
| LDA | 69.2 | 74.4 | 74.5 | 91.4 | 72.4 | 2.4 | 77.4 | 76.6 | 74.2 | 91.8 | 66.6 | 1.4 | 76.3 | 75.1 | 64.2 | 94.5 | 72.1 | 2.4 |
| QDA | 75.6 | 72.7 | 70.8 | 75.5 | 61.6 | 6.3 | 66.4 | 71.4 | 57.7 | 67.7 | 54.4 | 6.6 | 65.1 | 68.8 | 57.8 | 83.9 | 54.7 | 7 |
| k-NN10 | 71.2 | 77.0 | 74.3 | 72.8 | 60.5 | 4.7 | 72.4 | 65.4 | 70.7 | 92.3 | 56.7 | 5.3 | 66.4 | 63.4 | 65.8 | 90.5 | 54.3 | 4.6 |
| k-NN100 | 73.4 | 70.9 | 74.5 | 83.0 | 54.2 | 2.7 | 73.7 | 71.6 | 76.1 | 91.6 | 60.7 | 2.6 | 72.3 | 71.9 | 79.5 | 90.3 | 59.7 | 2.8 |
| | 6% worst | | | | | | 1.5% worst | | | | | | 0.9% worst | | | | | |
| | FD - test = 23.29 (p<0.006) | | | | | | FD-test statistic = 26.43 (p<0.006) | | | | | | FD-test = 29.86 (p<0.006) | | | | | |
| | Bajaj fin1 | Bajaj fin2 | cibil | behavior | Tataf in1 | Tataf in2 | Bajajf in1 | Bajaj fin2 | cibil | Behaiour | Tata fin1 | Tata fin2 | Bajajfin1 | Bajaj fin2 | cibil | Behaviour | Tata fin1 | Tata fin2 |
| LOG | 76.2 | 71.4 | 72.7 | 56.0 | 52.0 | 4.5 | 70.7 | 72.9 | 55.1 | 50.6 | 51.0 | 5.6 | 51.0 | 63.7 | 53.0 | 53.0 | 52.0 | 6.3 |
| Decision Trees | 52.6 | 63.9 | 55.5 | 76.4 | 53.0 | 6.7 | 64.8 | 66.9 | 61.4 | 58.7 | 52.9 | 5.7 | 50.0 | 54.5 | 65.2 | 55.0 | 53.0 | 5.6 |
| NN | 66.4 | 69.7 | 64.3 | 85.4 | 65.4 | 4.1 | 70.2 | 71.2 | 54.2 | 70.5 | 65.3 | 3.9 | 52.0 | 64.5 | 53.2 | 87.7 | 55.0 | 5.5 |
| Gradient Boosting | 68.8 | 78.0 | 77.6 | 90.1 | 51.7 | 2.9 | 66.1 | 75.7 | 76.4 | 88.2 | 55.6 | 2.6 | 54.1 | 69.4 | 56.4 | 76.5 | 54.0 | 1.7 |
| LDA | 72.1 | 75.1 | 72.8 | 92.5 | 62.5 | 2.5 | 77.7 | 71.2 | 60.6 | 81.5 | 62.6 | 2.7 | 52.2 | 67.0 | 57.3 | 85.8 | 56.6 | 1.6 |
| QDA | 65.8 | 78.0 | 49.0 | 55.7 | 54.5 | 6.7 | 62.5 | 66.3 | 51.0 | 51.4 | 52.5 | 7.0 | 51.0 | 55.0 | 54.0 | 57.0 | 50.7 | 5.6 |
| k-NN10 | 68.2 | 66.0 | 66.1 | 86.6 | 52.5 | 6 | 58.0 | 54.3 | 55.3 | 72.7 | 52.7 | 6.5 | 51.5 | 55.3 | 52.8 | 68.2 | 52.0 | 5.4 |
| k-NN100 | 72.7 | 70.3 | 72.8 | 90.3 | 57.8 | 3 | 72.6 | 66.8 | 66.3 | 87.6 | 57.3 | 3.1 | 65.2 | 62.2 | 62.6 | 92.0 | 55.0 | 2.3 |

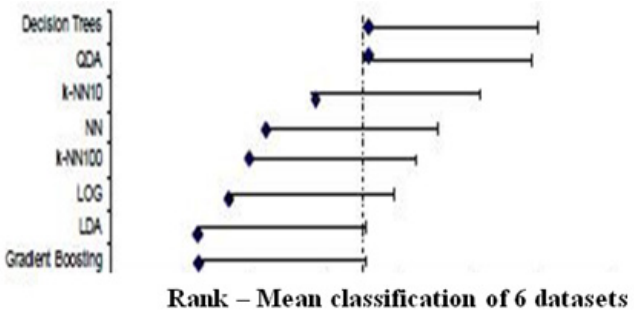


**Figure 3.** Rank average comparison of 85/15 segregation of best/worst.

Figure 4shows the performance mean ranks of the all the classifiers which is displayed in AUC, and also the difference of critical values for all the classifiers which is nearly equal to 4.60. Figure 1 shows classification techniques displayed in order of performance given by the rank in y-axis, as well as the rank mean of all 6 classified datasets displayed in x-axis. The parallel lines show the start and end point of the classifier which give appreciable performance.

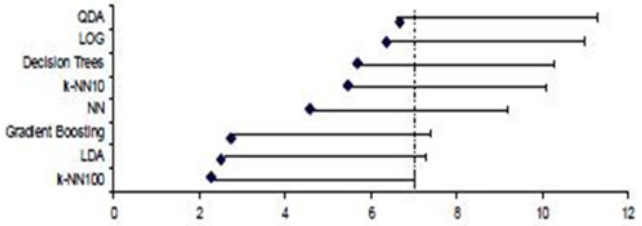


**Figure 4.** Rank average comparison of best/worst splitting.

Figure 2 shows the ranks values of average for all the classifiers. Based on the imbalance class distribution of 75% best and 25% worst are segregated.

In Figure 2 the difference of percentage 75/25 segregation of best/worst datasets lookup. FromFigure 3 we can observe that compare of all other LDA reaches the 75/25 with the average Rank value of 2.3 which is the best classification in performance measure. We can also observe that the good classifier than the worst case which is significantly gives the value of 7.2 shown in the diagram LDA , C4.5 and GD are better than the QDA algorithms which is worse shown in Figure 3 at most 15% are bad splits and 85 are good splits.

The graphics without converting to a PS, EPS, TIFF, PDF, or PNG file: Microsoft Word, Microsoft PowerPoint, or Microsoft Excel. Though it is not required, it is recommended that these files be saved in PDF format rather than DOC, XLS, or PPT. Doing so will protect your figures from common font and arrow stroke issues that occur when working on the files across multiple platforms[11]. When submitting your final paper, your graphics should all be submitted individually in one of these formats along with the manuscript.

Last and final split gives avery good results i.e 98% good and 2% bad split. This seems that incase of best performing technique any other classification is significantly worse in the classification of analysis in distribution

Gradient boosting classifiers yields, the measures AUC performance a very good class imbalance to reach the extreme levels.LDA and LOG gives a very good measure in linear techniques of classification, which is not relative to the gradient classifier. The evaluations made are used to confirm the non linear measure weakly. The performing techniques of QDA and decision trees made worst performance identifying the reduction of the percentage. The most of the classification techniques gives the performance measures which are competitive to each other.

## 6. Conclusions

The basis study is done to check the performance of imbalance dataset distributions of 6 real datasets. The LR and LDA gives the best appropriate selection in finding the good and worst customer prediction, when datasets are taken small of imbalanced type. Where are QDA and C4.5 give the better performance compared to the previous, when datasets are large and complex enough.

Future, we can identify the class range distribution of customer behavior in acting as defaulter using class range distribution method.

## 7. References


1. Friedman J. Stochastic gradient boosting. Comput Statist Data Analysis. 2002 Feb; 38(4):367–78.
2. Van Der BM. Calibrating Low-Default Portfolios, using the Cumulative Accuracy Profile. ABN AMRO.2007 Mar; 1(4):1–17.
3. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach LearnRes. 2006 May; 7:1–30.
4. Wiginton JC. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. J Finan Quant Anal. 1980Sep; 15(3):757–70.
5. Hastie T, Tibshirani R,Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. New York: Springer: 2009 Feb.
6. Friedman J .Greedy function approximation: A gradient boosting machine. Ann Statist. 2001 Oct; 29(5):1189–232.
7. Defaultors profile and data from Bajaj Financial Services, 2015.
8. The datasets from TATA Financial Services, 2015 Mar.
9. Chawla N, Bowyer K, Hall L,Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. J Artificial Intelligence Res. 2002; 16:321–57.
10. SasirekhaD,PunithaA.A Comprehensive Analysis on Associative Classification in Medical Datasets. Indian Journal of Science and Technology. 2015 Dec;8(33):1–9.
11. Agrawal R. Design and Development of Data Classification Methodology for Uncertain Data. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–12.