International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Identification of Default Payments of Credit Card Clients using Boosting Techniques

S. Sathya Bama, A. Maheshwaran, S. KishoreKumar, K. RaghulKumar, M. Yogeshwaran

Abstract: ***Understanding the history of clients will act as a valuable screening method for banks by providing information that can categorize clients as defaulters on a loan. Customer credit rating is a grade process where the consumer is categorized by the grade. Credit scoring model used to ascertain credit risk from new and existing customer. Credit rating is an assessment used to measure the creditworthiness of the customer. For the huge customers related dataset we can use various classification techniques used in the field of data mining. The main idea is by analyzing the customer data and by combining machine-learning algorithm to identify the default credit card user. Default is a keyword, used for predicting the customer who cant repay the amount on time. Predicting future credit default accounts in advance is highly tedious task. Modern statistical techniques are usually unable to manage huge data. The proposed work focus mainly on ensemble learning and other artificial intelligence technique.***

Index Terms: Customers, Classification Techniques, Credit Card, Ensemble methods

I. INTRODUCTION

Customer record provides many valuable information about used to identify the customer who will fail to pay the loan. The data related to customer are very huge so various data classification techniques are to be included. The Machine learning domain plays a major role in various application like medical, theft identification, forest fire, human detection, networking and so on[4][12][16][19]. Ensemble approaches are meta-algorithms incorporating many techniques of machine learning into one predictive model to reduce unreliability (bagging), bias (boosting), or improve predictions (stacking). Ensemble methods can be classified into two groups namely sequential set methods and parallel set method.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
S. SathyaBama*, Assistant professor at Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: s.sathyabama@skct.edu.in
A. Maheshwaran, Department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs110@skct.edu.in
S. KishoreKumar, Department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs105@skct.edu.in
K. RaghulKumar, Department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs145@skct.edu.in
M. Yogeshwaran, Department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs258@skct.edu.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

The fundamental motive of sequential methods is to manipulate the dependency among the base learners. Weighing previously mislabeled examples of higher weight will improve overall performance. Parallel ensemble methods where parallel baseline learners are created (e.g. Random Forest). The basic motive of parallel methods is to maximize independence among the base learners, as the error can be dramatically reduced by averaging. Bagging stands for grouping of bootstraps. One way to lower an estimate's variance is by combining several estimates together. The main objective of boosting is to fit into weighted versions of the data a series of poor learners – models which are only slightly better better than random guessing, such as small decision trees. Examples which were wrongly predicted in previous rounds are given more weight. The predictions are then combined to produce the final prediction by means of a weighted majority vote (classification) or a weighted sum (regression). The main difference between boosting methods and committee methods, like bagging, is that base learners are trained sequentially on a weighted version of the data. Stacking is the machine learning technique that combines multiple classification or regression model through meta classifier. Soft computing is used in various fields like medical, online and for human detection[4]. The base level techniques are trained based on full trained set, then the surrogate-model is trained as features on the expected outputs from the base level model.

II. RELATED WORKS

Credit score is a statistical and numerical representation based on the analysis of customer data which introduced in early 1980's[17]. Classification is a technique or process of grouping set of data into group and then perform analysis[18]. The static and dynamic model statistical models are under in decision making method and data analysis to test the credit score. For identification of credit card default client it can be analyzed by other method too like techniques like Neural Network, Support Vector Machine and so on[7][11]. The ensemble learning was proposed for further better classification performance and for imbalanced credit scoring dataset and to provide empirical evaluation using multiple data mining approaches[10]. These researches are done under the batch training and few researches done based on streaming dataset. Online machine learning is a machine learning technique used to restore the best analysis for future data. One characteristics of online learning is no need of availability of sufficient dataset before training[1].

Online learning represents a adequate and devastating algorithms compared to other algorithms and tackles the problem with memory consumption and retain cost with incoming data[9].

III. DATASET DESCRIPTION

The suggested system uses the original UCI repository report. There are 25 factors and 30,000 documents for customers. This dataset contains information on the credit card clients, regular charges, demographic factors, credit records, payment.

IV. MACHINE LEARNING CLASSIFERS

3.1 ADA BOOST CLASSIFIER

Ada-boost is otherwise known as Adaptive Boosting, is a machine learning algorithm to increase or improve performance. It is used to boost performance of decision tree for binary classification problem. Ada-boost is a classifier used for converting weaker dataset into stronger dataset. The base idea of ada-boost classifier to iterate by adding weights to misclassified observation.

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)$$

3.2 CAT BOOST CLASSIFIER

Cat Boost is abbreviated as “Category” and “Boosting”. Cat Boost is a recently open-sourced machine learning algorithm and it can be easily integrated with other framework. Cat boost has various features like converting categorical values into numbers using various statistics. It produces efficient result in accuracy result. It provides one more major functionality which is, it reduces the need for maximum hyper-parameter tuning and lower the chances of over fitting models. It is mainly used within organization for ranking and has wide application in various domain.

$$Q(X, j, t) = F(X) - \frac{|X_l|}{|X|}F(X_l) - \frac{|X_r|}{|X|}F(X_r)$$

3.3 XG BOOST CLASSIFIER

XG Boost is otherwise as eXtreme Gradient Boosting which is one of the machine learning boosting classifier models. The XG boost use **plot_importance()** function which is a build in function to generate feature importance, which improves the performance and efficiency by algorithmic optimization and system optimization.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(\bar{y}_i, \hat{y}_i{}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

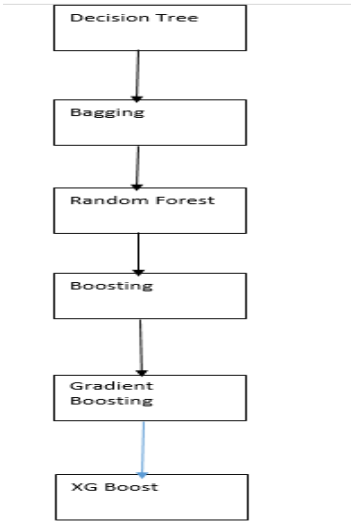

Fig.1.Evolution Of Xg Boost

3.4 LIGHT GRADIENT BOOST MACHINE CLASSIFIER

Light GBM is otherwise as Light Gradient Boosting Method is also known as fast,highly efficient gradient boosting methodology ,which is a tree based algorithm. The word light is derived since it is can process faster compare to other classifier. which uses tree based algorithm and its trees grows vertically whereas the other algorithms grows the opposite. The Light gradient boosting algorithm can be very efficient on large data sets which uses very low memory and it is very fast to compare other algorithms.

V. PROPOSED FRAMEWORKS:

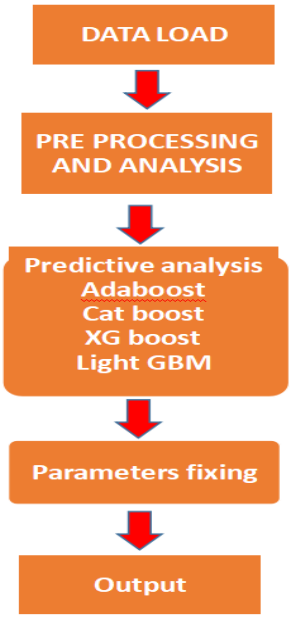

Fig.2. Proposed Flow Of The Work

The above diagram represents the proposed flow of the work. The data load is the first phase of the project, which is loading the data set into the algorithm, The next stage is nothing but pre processing. It performs the operation like data cleaning, finding missing values and to ensure that there is no redundant data.

Next comes the analysis method. It involves the algorithm like adaboost, Cat boost, XG boost and light GBM. It is used for performing or analyzing the huge dataset or imbalanced dataset and to provide the accurate result. The parameter fixing is nothing but removing the another fields. Then comes the final result phase. That is displaying the result.

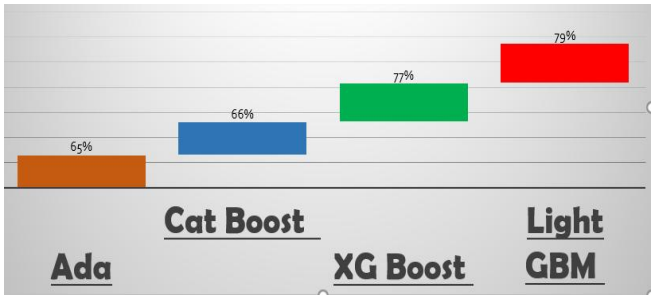

Fig.3. Auc Value Comparision

The above figure describes the comparison of four boosting techniques like adaboost, Cat boost, XG boost and light GBM. When compared to other techniques light GBM performance is good and provide accuracy result.

VI. RESULTS

5.1 CURRENT FINANCIAL STATUS OF THE BANK

Generally as per reports the bank is generating profit of nearly 22 million dollars using credit card business. The profit of the people in the data set approximately -600000 dollars. By this we can analyze each customer in data set produce loss of 20 dollars for the bank. The in-active customers brings profit to the bank steadily by paying their yearly fee. This following image shows profit of each customer in the data set.

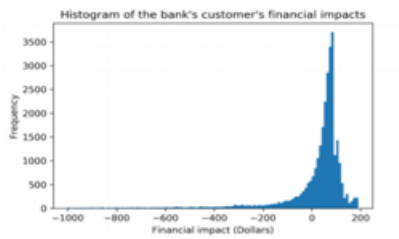

Fig.4.Profit Per Customer

The above figure implies that the profitable customer brings below 200 dollars to the bank while the individual loss per customer is over 5000 dollars. By analyzing the data set, we can know that the bank has about 23000 profit bringing customers and about 7000 non-profit bringing customers. It also further implies that the average profit is only 72 dollars however the loss is 330 dollars. By this we can imply that although there are more profit bringing customers the loss is very high.

5.2 FILTERING CUSTOMER BASE USING BOOSTING TECHNIQUES

5.2.1 ADA BOOST

Ada Boost is one of the boosting techniques which is used to convert weekly performing classifier to strong classifier. We run the data set using Ada boost algorithm. The confusion matrix of training the data set using the Ada Boost classifier is given below:

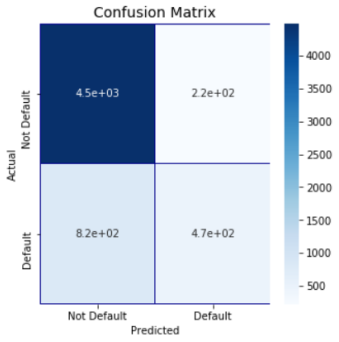

Fig.5.Confusion Matrix Of Ada Boost

From Ada Boost confusion matrix the ROC(Receiver Operating Characteristics Curve) score is 0.6588048536053512.

5.2.2 CAT BOOST

Cat Boost has various features like converting categorical values into numbers using various statistics. It produces efficient result in accuracy result.We run the data set using Cat boost algorithm. The confusion matrix of training the data set using the cat Boost classifier is given below:

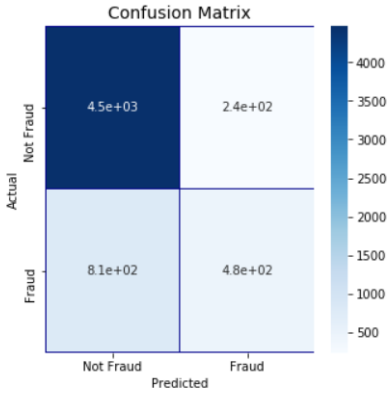

Fig.6.Confusion Matrix Of Cat Boost

From Cat Boost confusion matrix the ROC(Receiver Operating Characteristics Curve) score is 0.6619619855275654.

5.3 FPREDICTED DEFAULT RISK IN BOOSTING TECHNIQUES

The boosting algorithms are used to predict the credit default payments in the data set. The feature importance result is generated in each boosting algorithms and accuracy of each boosting techniques is also calculated.

5.3.1 ADA BOOST

The Ada boost uses parameters that are **base_estimator, n_estimator, learning_rate** which are very important in predict the feature importance.

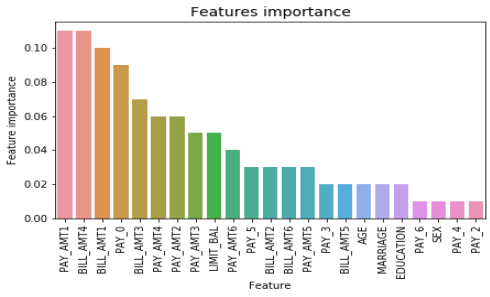

Fig.7.Feature Importance Of Ada Boost

The F1 Score of each attribute is calculated and it is further split based on those values. The accuracy obtained running the data set using Ada boost algorithm is 65%.

5.3.2 CAT BOOST

The Cat boost uses parameters that are **prettified,thread_count,verbose** which are very important in predict the feature importance.

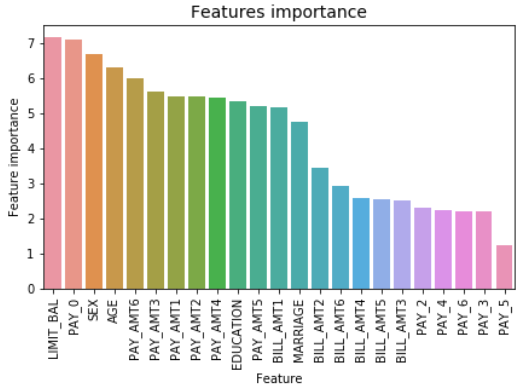

Fig.8.Feature Importance Of Cat Boost

The individual feature importance value of each input attributes is calculated and it is used in specifying in loss function which will produce accurate results. The accuracy obtained running the data set using Cat boost algorithm is 66%.

5.3.3 XG BOOST

The XG boost use **plot_importance()** function which is a build in function to generate feature importance of the input attributes.

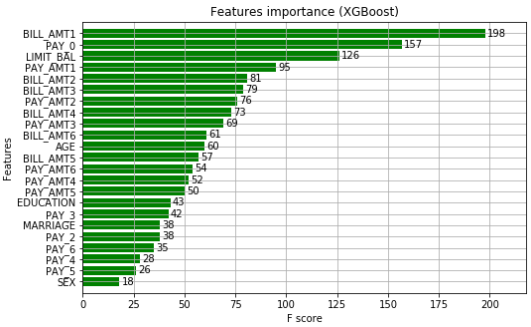

Fig.9.Feature Importance Of Xg Boost

XG Boost is a extreme gradient boosting method which improves the performance and efficiency by algorithmic optimization and system optimization. The accuracy obtained running the data set using XG boost algorithm is 77%.

5.2.4 LIGHT GBM BOOST

The Light gradient boost use top level of **eli5.explain_weights()** calls are dispatched to **eli5.lightgbm.explain_weights_lightgbm()** function to generate feature importance of attributes.

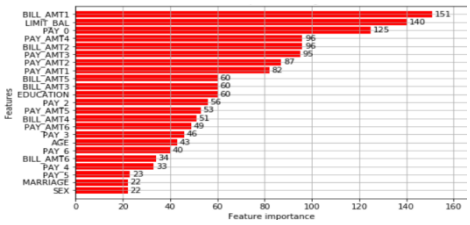

Fig.10.Feature Importance Of Light Gbm Boost

The Light gradient boosting is a gradient boosting framework which uses tree based algorithm and its trees grows vertically whereas the other algorithms grows the opposite. The Light gradient boosting algorithm can be very efficient on large data sets which uses very low memory and it is very fast to compare other algorithms. Its accuracy also higher than the other algorithms, since it has the highest accuracy of 79% which is higher than the above used algorithms.

VII. CONCLUSION

To identify the default payment of credit card clients of huge data set data analysis should be involved. Data analysis allows cultivation and learning based on model build, feature extraction, and various conditions that can improve the trait of customer acquirement. Here we use boosting technique which provides better performance and accuracy. The four boosting techniques mentioned can analysis the huge data set and to provide the accurate result. The boosting techniques which are included here can perform analysis for imbalanced dataset. Analysis of data set that allow 80% data set for learning and 20% of data set for training to improve the ability of customer. By using Predictive analysis model for estimating the default payment and loss of extend and for predicting losses.

REFERENCES

1. Ade, R.R., and Deshmukh, P.R., "Methods for incremental learning: a survey," International Journal of Data Mining and Knowledge Management Process, 3(4), 119.
2. Baesens, B., Setiono, R., Mues, C., and Vanthienen, J., 2003, "Using neural network rule extraction and decision tables for credit-risk evaluation," Management science, 49(3), 312-329.
3. Bellotti, T., and Jonathan, C., 2009, "Support vector machines for credit scoring and discovery of significant features," Expert Systems with Applications, 36(2), 3302-3308.
4. Bhuvaneswari, K. and Rauf, H.A., 2009, June. Edgelet based human detection and tracking by combined segmentation and soft decision. In 2009 International Conference on Control, Automation, Communication and Energy Conservation (pp. 1-6). IEEE.
5. Brown, L., and Christophe M., 2012, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," Expert Systems with Applications, 39(3), 3446-3453.
6. Freund, Y., and Llew, M., 1999, "The alternating decision tree learning algorithm," International Conference of Machine Learning, 99-109.
7. Giudici, P., 2001, "Bayesian data mining, with application to benchmarking and credit scoring," Applied Stochastic Models in Business and Industry, 17(1), 69-81.

8. Hand, D., and William, H., 1997, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.
9. Hoi, S.C., Wang, J., and Zhao, P., 2014, "Libol: A library for online learning algorithms," The Journal of Machine Learning Research, 15(1), 495-499. Huang, G., Huang, G.B., Song, S., and You, K., 2005, "Trends in extreme learning machines: a review," Neural Networks, 61, 32-48.
10. Huang, G., Qin, Z., and Siew, C., 2006, "Extreme learning machine: theory and applications," Neurocomputing, 70(1), 489-501.
11. Krishnan, M.S., Ragavi, S., RamKumar, M.S. and Kavitha, D., 2019. Smart Asthma Prediction System using Internet of Things. Indian Journal of Public Health Research & Development, 10(2), pp.1103-1107.
12. Lee, T.S., Chiu, C.C., Lu, C.J., and Chen, I.F., 2002, "Credit scoring using the hybrid neural discriminant technique," Expert Systems with applications, 23(3), 245-254.
13. Liang, N., Huang, G.B., and Saratchandran, P., 2006, "A fast and accurate online sequential learning algorithm for feedforward networks," IEEE Transactions on Neural networks, 17(6), 1411-1423.
14. Oza, N., 2005, "Online bagging and boosting," IEEE international conference on systems, man and cybernetics, May 25, 2340-2345.
15. Punithavathani, D.S. and Sankaranarayanan, K., 2009. IPv4/IPv6 transition mechanisms. European Journal of Scientific Research, 34(1), pp.110-124.
16. Rosenberg, E., and Alan, G., 1994, "Quantitative methods in credit management: a survey," Operations research, 42(4), 589-613.
17. Sreeja, N.K. and Sankar, A., 2015. Pattern matching based classification using ant colony optimization based feature selection. Applied Soft Computing, 31, pp.91-102.
18. Tamije, P., Palanisamy, V. and Purusothaman, T., Performance Analysis of Clustering Algorithms in Brain Tumor Detection of MR Images. European journal of scientific research, ISSN, pp.321-330.
19. Wang, G., Hao, J., Ma, J., and Jiang, H., 2011, "A comparative assessment of ensemble learning for credit scoring," Expert systems with applications, 38(1), 223-230.
20. Yeh, C., and Lien C., 2009, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications 36(2), 2473-2480.
21. Yuan, L., Soh, Y., and Huang, G.B., 2009, "Ensemble of online sequential extreme learning machine," Neurocomputing, 72(13), 3391-3395.
22. Ahmed, K., Ahmed, F., Roy, S., Paul, B.K., Aktar, M.N., Vigneswaran, D. and Islam, M.S., 2019. Refractive index-based blood components sensing in terahertz spectrum. IEEE Sensors Journal, 19(9), pp.3368-3375.

AUTHORS PROFILE

Ms. S.Sathya Bama is currently working as Assistant Professor in the Department of Computer at Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. She received his Master's degree in the year 2018 from Anna University, Coimbatore, Tamil Nadu, India. Her research interests include Data Analytics. She has published 2 papers in Journals and 2 Conferences.

S.KishoreKumar is pursuing her Bachelor's Degree in Engineering in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. His research interests include Machine Learning.

A.Maheshwaran is pursuing her Bachelor's Degree in Engineering in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. His research interests include Machine Learning.

K.Raghul Kumar is pursuing her Bachelor's Degree in Engineering in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. His research interests include Machine Learning.

M. Yogeshwaran is pursuing her Bachelor's Degree in Engineering in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. His research interests include Machine Learning