International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Misclassified Reduced Instance and Stochastic Gradient Descent with Logistic Regression Model for Customer Churn Prediction

Isabella Amali, R. Arunkumar, R. Madhan Mohan

Abstract*: Customer Churn Prediction (CCP) is a difficult problem found to be helpful to make decisions due to the rapid growth in the number of telecom providers. At present, deep learning models are familiar because of the significant improvement in different areas. In this paper, a deep learning based CCP is introduced by the use of Stochastic Gradient Boosting (SGD) with Logistic regression (LR) classifier model. By the integration of SGD and LR, effective classification can be accomplished. To further improve the classifier efficiency, misclassified instances are removed from the dataset. Then, the processed data is again provided as input to the classification model. The presented SGD-LR model is validated on a benchmark dataset and the results are examiner with respect to different measures. The experimental outcome pointed out the projected model is superior to available CCP models on the identical dataset.

***Keywords* : CCP; Classifier; Machine learning; Deep learning .**

I. INTRODUCTION

Nowadays, the telecommunication providers have been increased which leads to heavy competition as well as churn users. Because of maximum churn users, many firms have been focusing to the clients individually. Churn is stated that [1] it is the ability of a customer to terminate the business over alternate company. The prior need is to analyze the customers who have high possibility to transfer to other products. A firm should be capable of tracking the user to avoid the cause for churn. Churn occurs mainly due to the dissatisfaction of the customer. In order to detect this complaint different types of parameters has been utilized. In general, a customer would not become as a churn for single aspect of any complaint regarding the product [2].

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Isabella Amali, *, Research Scholar, Department of Computer and Information Science, Annamalai University. Email: isabella.amali@gmail.com
Dr. R. Arunkumar, Assistant Professor, Department of Computer Science and Engineering, Annamalai University. Email: arunkumar_an@yahoo.com
Dr. R. Madhan Mohan, Assistant Professor, Department of Computer Science and Engineering, Annamalai University. Email: madhanmohan_mithu@yahoo.com

Generally, before ceasing the entire transaction, several factors of dissatisfaction associated with the company has been existing.

Different operation models with the firm as well as a feature which is connected along with the customer is stores by an organization, this process illustrates the nature of the information. An efficient concept is achieved by examining the data and present condition of a user might be gained [3]. Churn prediction could use these data as basic knowledge. The dense nature of information, depicts the whole product by considering the firm. Therefore, the necessity of information in a structural form, all instances are included with whole property corresponding to common user in the industry [4].

Based on the sparse data, customers are connected only with few properties and not with all properties. While predicting churn, sparsity and massiveness of a data is promising issue. To provide various facilities, several firms interact with their clients [5]. The ability of forecasting while a customer moves to other product and this operation could increase the customer service. But this problem is very complex whenever the nature of information is consecutive and different. In any type of organization, churn could not be removed. Therefore, there are some possibilities to analyze the cause of churn. In recent times, machine learning (ML) techniques have been applied in different domains to achieve best results. The ML model could be used to resolve CPP.

By analyzing the probable customers of churn with the threat detection model is proposed by [6]. Generalized Additive Models (GAM) is employed in this technique. This model stabilizes linear limitations by allowing the non-linear fits to information. Recognizing the dangerous users and suggest the non-linear relationships, it helps to improve the marketing solutions. In [7] deployed a profiling technique based on Neural Network (NN) that could be applied to detect churn. Hence this models is different from alternate methods, because several models have been proposed to analyze the clients could rapidly move as churn. To compute the anticipation activities, it detects the user's advanced churn nature, providing enough data which is essential buffer to the company. [8] incorporates the similar NN based method. The scheme in [9] is based on the 80-20 pattern for recognizing the main attributes influencing churn as well as analyzing the key features of the data to decide churn. In [9] introduced a regression depended method of predicting churn. The above models deploy the user characteristics data to analyze and present extended performance.

Class imbalance is treated a major role in impact across the trustworthiness of the classification model. Based on minority class, the major issue is due to the class imbalance as well as it does not have sufficient training. In [10] presented a model by applying transfer learning schemes. From the related parts, the method proposed in [10] works with the application of classification process by providing the training for the user behavior.

It includes the aim over the banking field and results have been developed to enhance the operation. In [11], other model has been suggested that considers the imbalance behavior to work with churn detection. A sample technique that works effectively across the churn data proposed in [12]. In Game theory based CCP models are established. By employing several approaches, multiple publications enable the problematic nature of churn behavior. In [13], a method for predicting churn over decision tree and cluster analysis is proposed. It is applied in China's Telecom data.

In [14], alternate scheme applying several prediction models were presented. An integration of frequent alignment and k-Nearest Neighbor technique is employed in this method. In order to forecast churn, this models includes temporal categorical features. In a challenging nature of information, there is a heavy demand for utilizing heuristic for prediction process. Through In [15], a rule generation model is established that applies heuristics to predict user churn in telecom sector. For predicting and analyzing churn, [16] employed the integration of Genetic Programming (GP) and Self Organizing Maps (SOM). SOM is helpful in collecting the user as clusters and then outliers are eliminated to obtain clusters representing user behaviors. By applying GP, an enhanced classification tree is introduced.

[17] proposes a boosting model which has the objective of improving classification model accuracy from classifier model. This technique inspires the training patterns by employing a concatenation of LR and clustering. The similar prediction boosting method employing Genetic Algorithm is proposed. This is an ensemble technique applying different methods for predicting churn.

In this paper, a deep learning based CCP is introduced by the use of Stochastic Gradient Boosting (SGD) with Logistic regression (LR) classifier model. By the integration of SGD and LR, effective classification can be accomplished. To further improve the classifier efficiency, misclassified instances are removed from the dataset. Then, the processed data is again provided as input to the classification model. The presented SGD-LR model is validated on a benchmark dataset and the results are examiner with respect to different measures. The experimental outcome pointed out the projected model is superior to available CCP models on the identical dataset.

II. PROPOSED ALGORITHM

A. Misclassified reduced instance (MRI) models

1. AdaBoost

AdaBoost is an optimal classifier which is used to enhance the performance of decision tree on binary classifier issues. Recently, it is known as discrete AdaBoost as it is applied for classifying process and not for regression. It is helpful in improving the performance of different machine learning (ML) technique, which helps for weak learners. These techniques obtain accuracy in a arbitrary manner in classification issue. Most suitable and common technique employed with AdaBoost classifier is decision tree algorithm. Due to short trees and it consist of single decision tree it is referred as decision stumps. Every instance present in the training dataset is checked for its weight. The primary weight is set to:

$$weight(xi) = \frac{1}{n} \quad (1)$$

where xi is i'th training instance and n is the number of training instances.

2. J48 classifier

Decision Tree Algorithm is used to identify the behavior of attribute-vector in any number of instances. The basic training instance have been identified for newly originated instances. This technique provides few strategy for predicting the target parameter. By the usage of classification technique a serious distribution of information could be known in a simple manner. J48 is the expansion of ID3. The extra characteristics of J48 is considered for decision trees pruning, missing values, derivation of patterns, frequent attribute value ranges, and so on. WEKA is a data mining tool, where J48 is open source Java implementation of C4.5 technique. This tool offers a count of chances linked with pruning tree. While the process of potential over fitting takes place, then pruning could be applied as précising tool. In alternate models the classifying process is performed recursively unless all leaf becomes pure, where data classification must be efficient. This algorithm generates the rules from specific identity of data. The key objective is to generalize a decision tree till it achieves equal accuracy and flexibility.

3. LR

LR is a supervised classification technique. In a classification based issue, the target parameter, y, could consider the discrete values from provided set of features, X. This method creates a regression model for predicting the possibility where the given data entry belongs to the class which is numbered as "1". Similar to Linear regression the data follows linear function. LR develops the information with the help of sigmoid function.LR is designed as classification model if threshold decision is comprised within the image. Fixing threshold value is an important aspect of LR which is based on classification problem. The result obtained from threshold value affects the values of precision and recall. But, the recall and precision should be 1 which is rare in this case.

4. Olex-GA classifier

Generally, GA consists of 3 components like population which is the set of candidate solutions referred as individual or chromosome that is increased in iteration count. Fitness function utilized to allocate a value for all individuals of population. Evolution strategy is based on the operators such as crossover, mutation and selection.

Population Encoding

Based on GA technique, each individual in population could be shown as individual rule or set of rules. The initial method organizes the individual in a simple manner, but the semantic indicator of rule quality.

Simultaneously, different rules for single model was an individual that denotes the entire classification, and it requires efficient individual coding, thus the fitness does not provide any reliable symptom. Therefore, a balance is managed between the simple encoding and effective fitness function.

Fitness Function

The fitness of a chromosome K, denotes $H_c(Pos,Neg)$, which is the value of F-score derived by applying $H_c(Pos,Neg)$ to the training set TS. This result is obtained from formulating the statement of $MAX-F$. Recently, while $D(K) \subseteq TS$indicates the collection of all document including the positive term in Pos and there is no negative term in Neg, i.e.,

$$D(K) = \cup_{t_{\in Pos}} \Delta(t) \setminus \cup_{t_{\in Neg}} \Delta(t) \quad (2)$$

starting from the definition of $F_{c,\alpha}$, following few algebras, the representation of $F_{c,\alpha}$ is obtained:

$$F_{c,\alpha}(K) = \frac{|D(K) \cap TS_c|}{(1-\alpha)|TS_c| + \alpha|D(K)|} \quad (3)$$

Evolutionary Operators

The selection process is taken place with the help of roulette-wheel model and crossover is carried out by the even crossover approach. Mutation captures the process to flip all individual bit provided using the given probability. In order to leave effective chromosome, elitism is employed to verify that best individuals originate from current generation is provided to consecutive one without modifying the genetic operator.

5. RBF

A Radial Basis Function Network (RBFN) is a specific model of Neural Network (NN). RBFN model is intuitive than MLP. RBFN performs classification process by calculating the input's identity from the training dataset. Every RBFN neuron records a "prototype", that is a single instance form training set. Whenever a novel input to be classified, all neurons to be processed with the Euclidean distance between input and its prototype. Therefore, when the input resembles as class A prototype than class B prototype , then is known as class A.

B. SGD-LR classification model

SGD is a popular ML and DL depend optimization methods. SGD is effective technique which requires unique monitoring to save the memory. For example, assume a dataset that comprises 1 million inputs. Generally, GD contains around 1 million observations in all rounds. Hence, in SGD, the parameter evaluation gets maximized with single observation at same time and the application to calculate the online learning, where new observations have been involved every time. LR is a ML technique that is motivated from statistics concept. It is a mainly employed to solve binary classification, i.e. a problem that comprises a set of two classes. In LR model, the response parameter c_i is categorical. The intention of work depends upon binary actions as well as it could be described in 2 classes only churn and non-churn. Hence, $c_i \in \{0,1\}$, where 0 denotes the churn and 1 represents non-churn. The logistic function $h(z) = 1/(1+e^{-Z})$, when $z = \theta^T A$ is expressed as follows [8]

$$h(\theta^T A) = \frac{1}{1+e^{-\theta TA}} \quad (4)$$

where θ is the parameter vector and $h(z)$ is limited in [0, 1]

In this work, the SGD with LR is applied to CCP and the pattern incorporated in SGD-LR is shown as follows.

Step 1: Start with Rocchio-like linear classifier:
$\hat{c} = \text{sign}(\text{a} \cdot \text{b})$

Step 2: Replace $sign(\dots)$ with other differentiable:
$\hat{c} = \sigma(\text{a} \cdot \text{b}) = \text{p}$

Step 3: Then, scaling takes place from 0 to 1 not 1- to +1, i.e.

$$\sigma(s) = \frac{1}{1+e^{-s}} \quad (5)$$

Step 4: Optimization is carried out

$$LCL(c|b,a) = \begin{cases} log\sigma(\text{b,a}) & c=1 \\ \log(1-\sigma(\text{b,a}) & c=0 \end{cases} \quad (6)$$

Step 5: Differentiate….$= \log\left(\sigma(\text{b,a})^c\left(1-\sigma(\text{b,a})\right)^{1-c}\right)$

$$logP(C=C|A=a,b) = \begin{cases} \log p & if(c=1) \\ \log(1-p) & if(c=0) \end{cases} \quad (7)$$

$P = \sigma(A.B)$ (4)

Step 6: If it is differentiated, the result would be

$$\frac{\partial}{\partial B} L(B|c,A) = (c-p)A \quad (8)$$

$$\frac{\partial}{\partial b^j} L(B|c,A) = (c-p)x^i \quad (9)$$

Step 7: The update of gradient descent with rate λ:

$$b^{(t+1)} = b^{(t)} + \lambda (c-p)A \quad (10)$$

$$\frac{\partial}{\partial a^j} \log P(C=c|A=A,B) = (c-p)a^i \quad (11)$$

A group of 2 key computational points requires to be managed:

- if $a^j = 0$, the gradient of b^j becomes zero
- Weights should be updated to non-zero features of an example.

III. PERFORMANCE VALIDATION

A. Dataset

To assess the effectiveness of the introduced MIR-SGD-LR approach, a benchmark dataset is applied. The dataset includes a collection of 7043 instances under the existence of 21 features [18]. Besides, a pair of class labels namely positive and negative instances is present in the dataset.

A collection of 26.54% of samples comes under positive label and the remaining 73.46% of samples comes under negative samples. Table 2 shows the parameters involved under the applied classifier models.

Table 1 Dataset Description

Description	Dataset
Instance count	7043
Feature count	21
Class count	2
% of Positive Samples	26.54%
% of Negative Samples	73.46%

Table 2 Parameter Settings

Methods	Class Index	Maximum Iterations	Number of Folds	Thres hold
AdaBoost	-1	>=1	<2	>=0
J48	-1	>=1	<2	>=0
LR	-1	>=1	<2	>=0
OlexGA	-1	>=1	<2	>=0
RBF	-1	>=1	<2	>=0

B. Results analysis of MRI

Fig. 1 shows the distribution of the instances on the applied dataset under different attributes. Figs. 2-6 shows the distribution of the MRI dataset of various techniques. Table 3 shows the results attained by different classifier model. The table indicated that the collection of 7043 instances. It is noted that the J48 classifier model exhibits ineffective reduction with the reduced number of 5850 instances. The LR classification model showed that unproductive diminution with the lessen number of 5674 instances. Then, the AdaBoost classier model offered moderate MRI with the reduced number of 5570 instances. Simultaneously, the Olex-GA technique offered effective classification with the MRI with the number of 5351 instances. Moreover, the RBF model offered superior classification with the minimal number of 5265 instances.

Table 3 MIR results of diverse models

Methods	Original Instances	Reduced Instances
AdaBoost		5570
J48		5850
LR	7043	5674
OlexGA		**5351**
RBF		**5265**

C. Classifier results analysis under diverse measures

Table 4 offered the results attained by diverse models interms of various measures on the tested dataset. Fig. 7 investigates the results attained by diverse models based on sensitivity. At this point, voting classification technique shows incompetent for classifying the data and is evident by noticing the least sensitivity value of 73.46. Afterwards, NBTree technique started to exhibit somewhat enhanced classification model over the vote classifier by achieving a sensitivity value of 83.99.

At the same time, the SGD+LR model offered almost identical results with the sensitivity value of 83.73. But, the sensitivity value of the SGD+LR model does not outperform all the other compared methods. At the same time, it is not that much better than the other comparison methods. On continuing with, the NB model tried to showcase better results over the vote, NBTree and SGD+LR models with the sensitivity value of 90.93. This value seems to be higher than many of the compared methods except presented model. Subsequently, the presented MIR-SGD-LR model works well and attains closer classifier results to SGD+LR model by attaining the highest sensitivity value of 99.95. On all of the above, the MIR-SGD-LR model exhibits excellent classification by offering the maximum sensitivity value.

Fig. 7 investigates the results attained by diverse models based on specificity. At this point, NB technique shows incompetent for classifying the data and is evident by noticing the least specificity value of 48.28. Afterwards, NBTree started to exhibit somewhat enhanced classification model over the NB classifier by achieving a specificity value of 61.88.

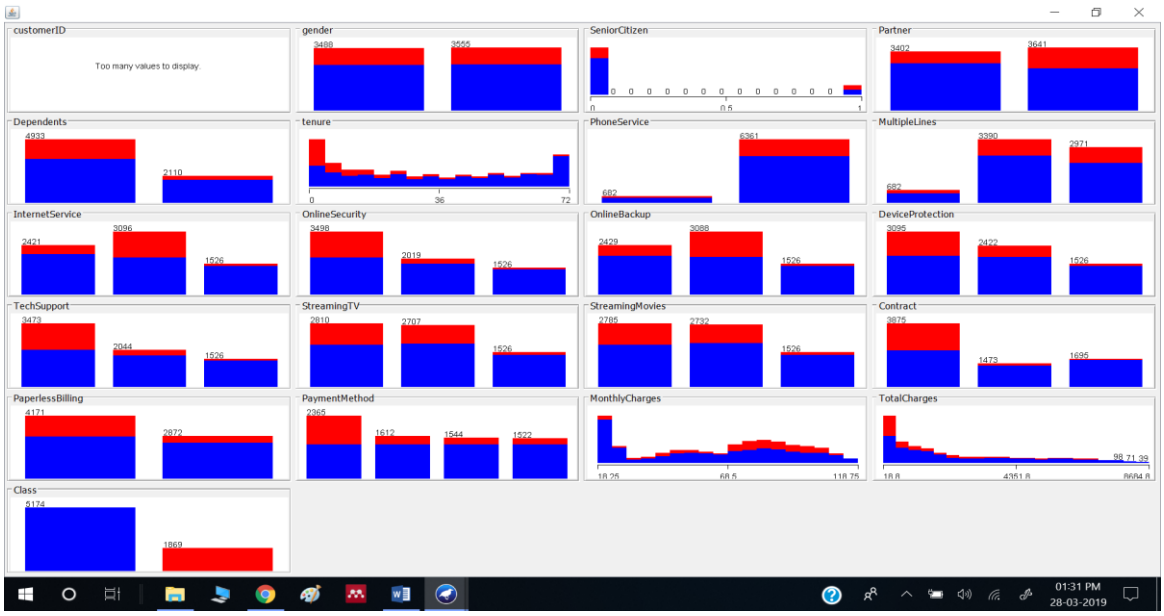

Fig. 1. Frequency Distribution of Original Dataset for all Attributes

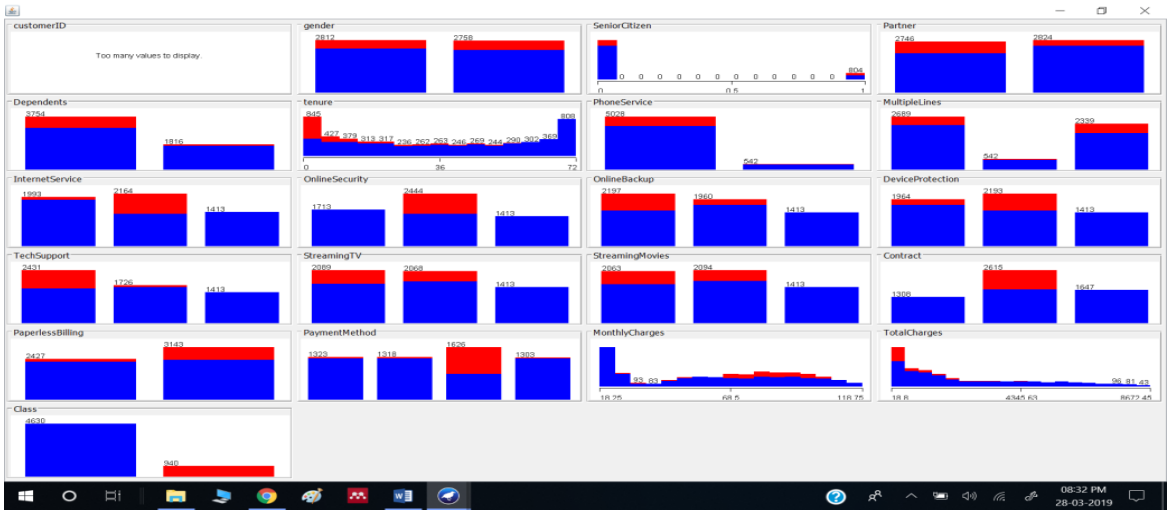

Fig. 2. Frequency Distribution of Reduced Dataset using AdaBoost

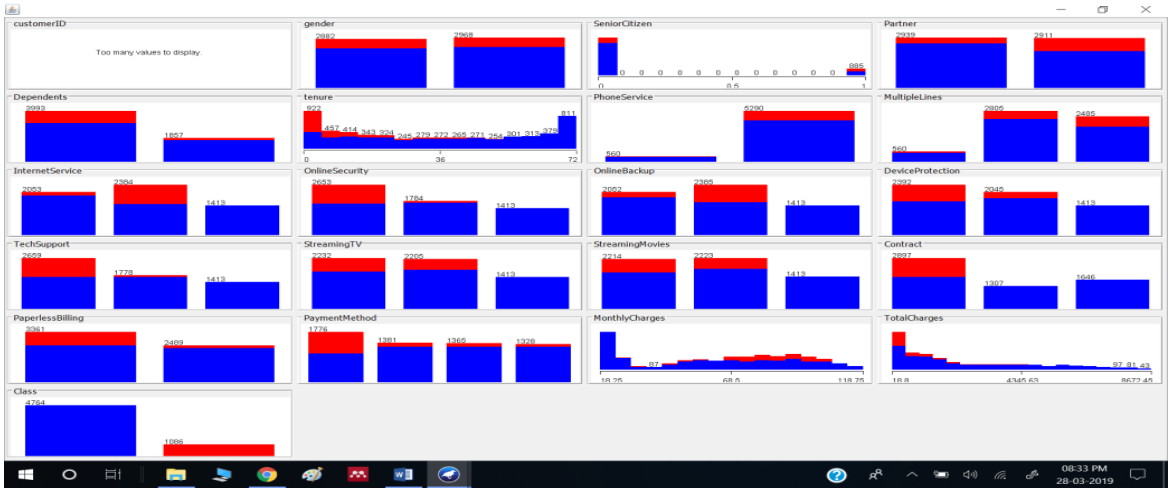

Fig. 3. Frequency Distribution of Reduced Dataset using J48

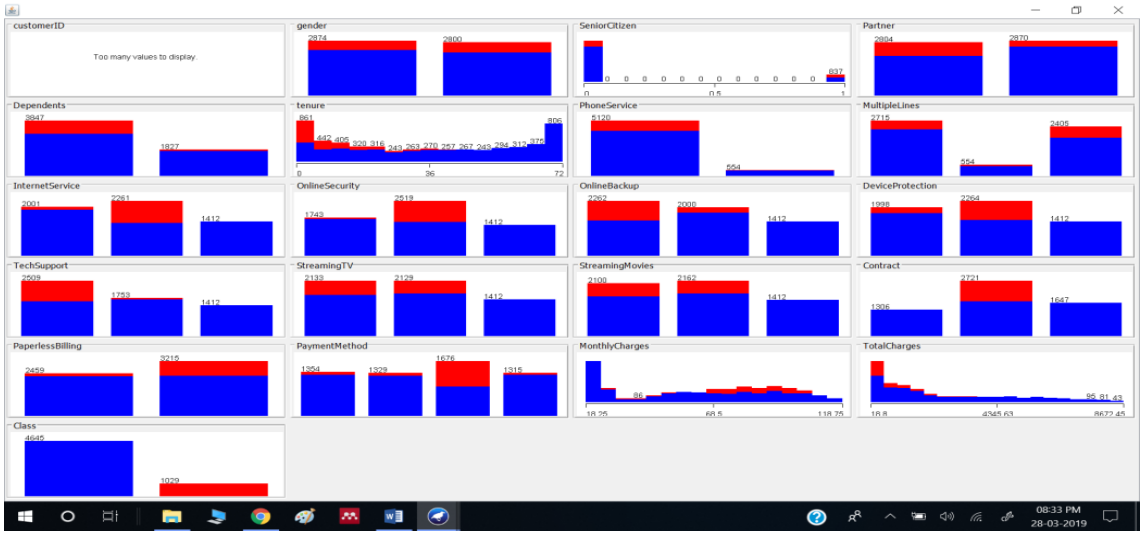

Fig. 4. Frequency Distribution of Reduced Dataset using LR

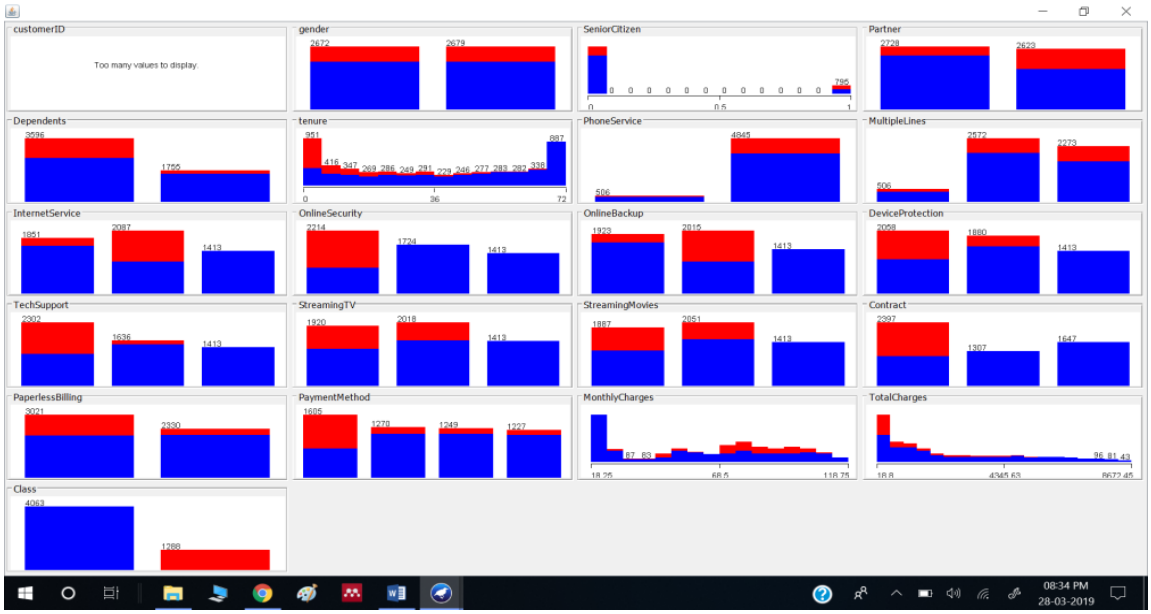

Fig. 5. Frequency Distribution of Reduced Dataset using OlexGA

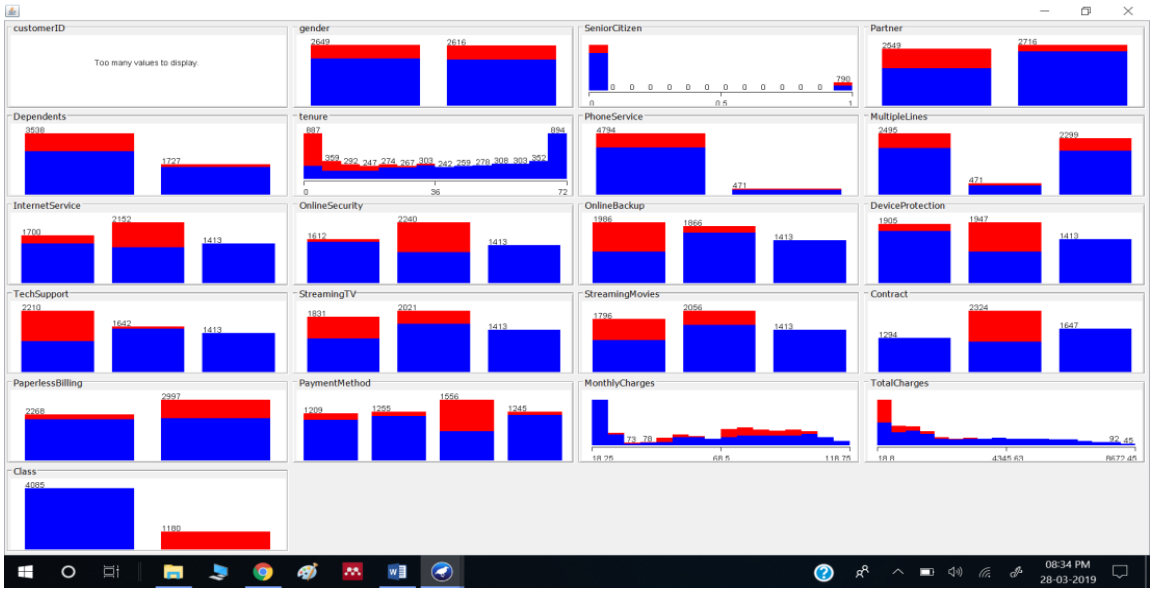

Fig. 6. Frequency Distribution of Reduced Dataset using RBF

Table 4 Performance Evaluation of Different Classifier Algorithms with Proposed on Applied Dataset

Methods	Sensitivity	Specificity	Accuracy	F-Score	Kappa
Proposed+IR	99.95	99.45	99.83	**99.88**	99.54
SGD+LR	83.78	67.56	80.52	**87.29**	45.84
NB	90.93	48.28	71.93	**78.22**	40.84
NBTree	83.99	61.88	78.91	**85.99**	43.47
Vote	73.46	-	73.46	**84.70**	0

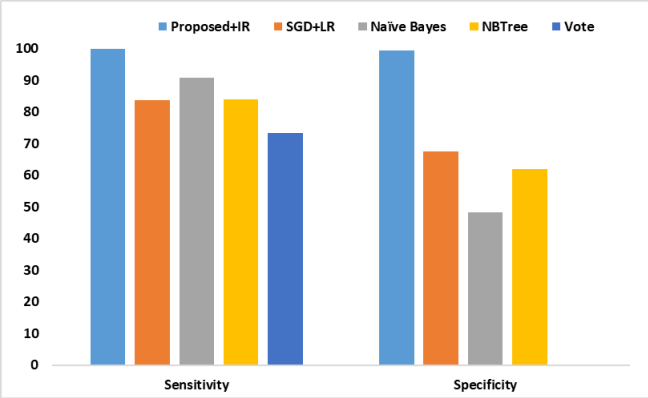

Fig. 7. Sensitivity and Specificity analysis

At the same time, the SGD+LR model offered almost somewhat identical results with the specificity value of 67.56. But, the specificity value of the SGD+LR model does not outperform all the other compared methods. At the same time, it is not that much better than the other comparison methods. On continuing with, the presented MIR-SGD-LR model works well and attains closer classifier results to SGD+LR model by attaining the highest specificity value of 99.45. On all of the above, the MIR-SGD-LR model exhibits excellent classification by offering the maximum specificity value.

Fig. 8 investigates the results attained by diverse models based on accuracy. At this point, NB technique shows incompetent for classifying the data and is evident by noticing the least accuracy value of 71.93. Afterwards, vote technique started to exhibit somewhat enhanced classification model over the NB classifier by achieving a accuracy value of 73.46. At the same time, the NBTree model offered almost somewhat identical results with the accuracy value of 78.91. But, the accuracy value of 80.52 by the SGD+LR model does not outperform all the other compared methods. On continuing with, the presented MIR-SGD-LR model works well and attains closer classifier results to SGD+LR model by attaining the highest accuracy value of 99.83. On all of the above, the MIR-SGD-LR model exhibits excellent classification by offering the maximum accuracy value.

Fig. 8 examines the results attained by diverse models based on F-score. At this point, NB technique shows incompetent for classifying the data and is evident by noticing the least F-score value of 78.22. Afterwards, vote technique started to exhibit somewhat enhanced classification model over the NB classifier by achieving a F-score value of 84.70.

At the same time, the NBTree model offered almost somewhat identical results with the F-score value of 85.99. But, the F-score value of 87.29 by the SGD+LR model does not outperform all the other compared methods. On continuing with, the presented MIR-SGD-LR model works well and attains closer classifier results to SGD+LR model by attaining the highest F-score value of 99.88. On all of the above, the MIR-SGD-LR model exhibits excellent classification by offering the maximum F-score value.

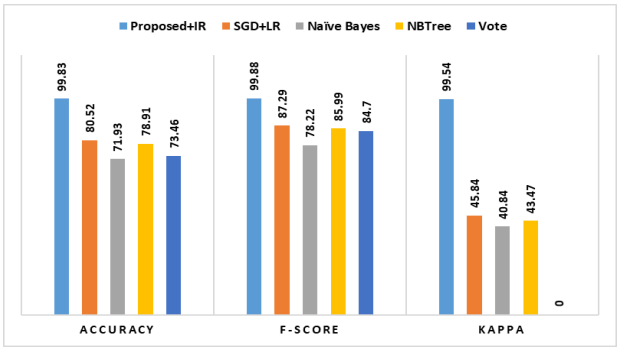

Fig. 8. Classifier results analysis

Fig. 8 displays the results attained by diverse models based on Kappa. At this point, voting technique shows incompetent for classifying the data and is evident by noticing the least Kappa value of 0. Afterwards, NB technique started to exhibit somewhat enhanced classification model over the vote classifier by achieving a Kappa value of 40.84. At the same time, the NBTree model offered almost somewhat identical results with the Kappa value of 43.47. But, the Kappa value of 45.84 by the SGD+LR model does not outperform all the other compared methods. On continuing with, the presented MIR-SGD-LR model works well and attains closer classifier results to SGD+LR model by attaining the highest Kappa value of 99.54. On all of the above, the MIR-SGD-LR model exhibits excellent classification by offering the maximum Kappa value.

IV. CONCLUSION

In this paper, a deep learning based CCP has been presented by the use of SGD-LR classifier model.

By the integration of SGD and LR, effective classification can be accomplished. To further improve the classifier efficiency, misclassified instances are removed from the dataset. Then, the processed data is again provided as input to the classification model. The presented SGD-LR model is validated on a benchmark dataset and the results are examiner with respect to different measures. The experimental outcome pointed out the projected model is superior to available CCP models on the identical dataset.

REFERENCES

1. Effendy V, Baizal ZA. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In: 2014 2nd International Conference on Information and Communication Technology (ICoICT), IEEE; 2014. p. 325–30.
2. Seo D, Ranganathan C, Babad Y. Two-level model of customer retention in the US mobile telecommunications service market. Telecommun Policy 2008;32 (3):182–96.
3. Hung SY, Yen DC, Wang HY. Applying data mining to telecom churn management. Expert SystAppl 2006;31(3):515–24.
4. Canning G. Do a value analysis of your customer base. Ind Mark Manage 1982;11(2):89–93.
5. Bhattacharya CB. When customers are members: customer retention in paid membership contexts. J Acad Mark Sci 1998;26(1):31–44.
6. Coussement K, Benoit DF, Van den Poel D. Preventing customers from running away! Exploring generalized additive models for customer churn prediction. In: The Sustainable Global Marketplace. Springer International Publishing; 2015. p. 238-238.
7. Tiwari A, Hadden J, Turner C. A new neural network based customer profiling methodology for churn prediction. In: Computational Science and Its Applications–ICCSA 2010. Berlin Heidelberg: Springer; 2010. p. 358–69.
8. Shen Q, Li H, Liao Q, Zhang W, Kalilou K. Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction. In: The 26th Chinese Control and DecisionConference (2014 CCDC), IEEE; 2014. p. 2250–55
9. Awang MK, Rahman MNA, Ismail MR. Data mining for churn prediction: multiple regressions approach. In: Computer applications for database, education, and ubiquitous computing. Berlin Heidelberg: Springer; 2012. p.318–24.
10. Zhu B, Xiao J, He C. A balanced transfer learning model for customer churn prediction. In: Proceedings of the eighth international conference on management science and engineering management. Berlin Heidelberg: Springer; 2014. p. 97–104.
11. Xiao J, Teng G, He C, Zhu B. One-step classifier ensemble model for customer churn prediction with imbalanced class. In: Proceedings of the eighth international conference on management science and engineering management. Berlin Heidelberg: Springer; 2014. p. 843–54.
12. Amin A, Rahim F, Ali I, Khan C, Anwar S. A comparison of two oversampling techniques (SMOTE vs MTDF) for handling class imbalance problem: a case study of customer churn prediction. In: New contributions in information systems and technologies. Springer International Publishing; 2015. p. 215–25.
13. Li G, Deng X. Customer churn prediction of China telecom based on cluster analysis and decision tree algorithm. In: Emerging Research in Artificial Intelligence and Computational Intelligence. Berlin Heidelberg: Springer; 2012. p. 319–27.
14. Le M, Nauck D, Gabrys B, Martin T. KNNs and sequence alignment for churn prediction. In: Research and development in intelligent systems XXX. Springer International Publishing; 2013. p. 279–85.
15. Huang Y, Huang B, Kechadi MT. A rule-based method for customer churn prediction in telecommunication services. In: Advances in knowledge discovery and data mining. Berlin Heidelberg: Springer; 2011. p. 411–22.
16. Faris H, Al-Shboul B, Ghatasheh N. A genetic programming based framework for churn prediction in telecommunication industry. In: Computational collective intelligence, technologies and applications. Springer International Publishing; 2014. p. 353–62.
17. Lu N, Lin H, Lu J, Zhang G. A customer churn prediction model in telecom industry using boosting. IEEE Trans Indust Inform 2014;10(2):1659–65.
18. https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_FnUseC-Telco-Customer-Churn.xlsx