Performance Analysis of Under-sampling Approaches for Solving Customer Churn Prediction *

Geeta Mahadeo Ambildhuke[a]**, G.Rekha [a], Amit Kumar Tagi [b]

[a] Koneru Lakshmaiah Education Foundation, Hyderabad, India. [b] Vellore Institute of Technology, Chennai, India

Abstract. With the increasing growth of technologies and digitization in telecom industries, retention of customers has been a serious concern. The prediction of customer retention/customer churn is a challenging task for telecom industry. This problem greatly affects the revenue of telecommunication industry. A good customer churn prediction model is needed to reduce the revenue loss and to rapidly increase the business. However, the development of good model is a challenging task due to imbalanced nature of data, large size of data, high dimensional features and many more. In this paper, we focus on imbalanced data distribution by presenting a solution in balancing it in a very effective way. We employed undersampling techniques using decision tree and with boosting. Thus, this paper outlines the various undersampling techniques using single and ensemble classifiers in solving the credit card churn prediction problem.

Keywords: Churn prediction· Machine Learning· Class imbalance · Sampling · Under-sampling techniques · Ensemble techniques.

1 INTRODUCTION

The term 'Customer churn' is defined as a customer terminates his/her usage of service from the service provider. It is been frequently used in telecom industries. Over the last decade, telecommunication industry witnessed rapid growth in service subscription. By the beginning of year 2015, the number of cell phone customers has arrived at about 8 billion around the globe roughly same as total population [2]. Therefore, the major problem faced by the telecom industry is retention of customers in order to maintain good revenue. Also, it is broadly acknowledged that holding of existing customer is more valuable than finding new customer. As, acquiring a new customer is costly affair pertaining to certain factors like satisfaction of demands, loyalty with the service providers and many more. So, retention of existing customers is very important for service providers and does not involve any expenses but only concern is to address the customer's

* Supported by KL University.

** Geeta Mahadeo Ambildhuke. Email: geeta@klh.edu.in

concerns in time. Further, by holding long term customer the industry benefits not only in terms of profit but only the existing customers can refer new customers. Therefore, the telecom industry needs an effective customer churn prediction model to detect the churners and it became a important research topic.In the literature, different approaches have been proposed by the research community using data mining and machine learning techniques. In general, the customers data like personal demographics, call details, account details, billing details and many more features were consider by the classification algorithm for predicting churn prediction model. So, customer churn prediction is a complex and challenging as most of data used for prediction is messy, noisy and imbalance in nature. The customer churn prediction model is a binary classification problem where in minority/ positive class consist of churners and majority/ negative class consist of non-churners.

Voluntary and involuntary churns are two main categories of churners [12]. Voluntary churners are those customers who make a decision to quit their services from the service providers. It is very difficult to decide/ determine these type of customers. The second type, involuntary churns are those customers whom the organizations decide to discard from the service because of many reasons such as non-payment, fraud or non usage of phone. The former is more tough to identify in general. It occurs when a decision was made by the service user to terminate his/her service with the provider. Several machine learning algorithms focus on the classification accuracy improvement to predict minority class. A need for automated churn prediction model to predict customer churn is much needed by telecommunication industry.

In this paper, we employed various undersampling techniques using decision tree and with ensemble technique (Boosting) to investigate the credit card churn prediction problem in telecom. Furthermore, we trained the resultant data using single and ensemble technique.

The paper is organized as follows. In Section 2, gives an overview of the customer churn prediction literature. In Section 3, we describe the methodology used in the study. We present numerical experiments and results in Section 4, and Section 5 concludes the paper.

2 Literature Survey

Many techniques in the literature were proposed to predict customer churn prediction in telecom industries. Most of these approaches/techniques applied machine learning algorithms to predict churn. As data is imbalance in nature, many techniques were proposed in the literature [7] [17] [6]. Gavril al. [3] presented machine learning technique to predict the prepaid customers churn using 3333 customers data. The data consists of more than 20 features including the number of messages (both incoming and outgoing) and voicemail for each service user and a dependent variable called churn with binary outcomes (Yes/No). They applied Principal Component Analysis (PCA) for dimension reduction. To predict the prepaid churn three machine learning algorithms were used: Neural Networks

(NN), Support Vector Machine (SVM), and Bayes Networks. The AUC metric were used to measure the algorithm performance. The author stated 99.10%, 99.55% and 99.70% respectively for the above algorithms.

A Neural network model was proposed by He et al [9] to solve customer churn problem for large Chinese telecom industry. The dataset was huge with 5.25 million customer's records. The achieved an overall accuracy nearer to 91.1

Idris [11]proposed genetic programming using AdaBoost for churn prediction. The model was tested on Orange and cell2cell datasets. The accuracy reported by the model was 89% for cell2cell dataset and 63% for Orange dataset. Huang et al [10] stated that applications of big data techniques improves the performance of the churn prediction model. They test the model on China's largest telecommunication company using Random Forest Algorithm. They proved better Area Under Curve (AUC) results. Author [14] proposed rough set theory for customer churn prediction in mobile industry. As revealed in this article rough set classification algorithm outperformed the other algorithms like LR (Linear Regression), DT (Decision Tree), and NN (Neural Network). Various researchers studied the imbalance nature of the customer churn data as a major concern. Amin et al [1] proposed rules-based genetic algorithms and compared six different OS (Over Sampling) techniques to balance the telecom data. The results outperformed with that of other oversampling algorithms. Burez et al [4] present comparison study using different sampling techniques such as Random Sampling, Gradient Boosting Model, and Weighted Random Forests using AUC and Lift metrics and the results showed better performance for undersampling techniques than other techniques.

The various technqiues studied in this work are as follows: Tomek-link undersampling technique is used to eliminate boundary instances considering them to be getting misclassified most often [19]. By definition, two instances y_i and y_j where class of y_i is equal to class of y_j, are said to form Tomek-link pair, if there is no instance k such that $d(y_i, y_k) < d(y_i,y_j)$. Basically, instances creating Tomek-link pair, promote noise along the data distribution. Cluster-based undersampling approaches selects important samples from training data to improve the minority class accuracy and also reports the effect of under-sampling methods on skewed class distribution. C-Nearest Neighbor Hart [8] proposed CNN undersampling technique using nearest neighbour method. Wilson's edited nearest neighbor rule (ENN) [15] identify noisy data and removes instances whose class differs from the majority class of the three nearest neighbors. The author stated that ENN retains most of the data, while maintaining a good classification accuracy. The author proposed Near Miss wherein the data samples from C ($C \geq 1$) clusters initially, and determine the number of selected majority class samples for each cluster. The sampling based on clustering with NearMisss-1 selects the majority class samples whose average distances to M nearest minority class samples ($M \geq 1$) in the ith cluster ($1 \leq i \leq C$) are the smallest Random under-sampling [13] is a popular method that aims at balance skewed distribution through the eliminating majority class samples randomly. The ma-

jor drawback of random undersampling is it will discard potential data samples that may be the important samples in learning process.

Hence in this section, we reviewed the article of various techniques to handle customer churn prediction.

3 Commonly Used Machine Learning Algorithms used for Customer Churn Prediction

In this section, we briefly present the popular machine learning techniques used for customer churn prediction. The most popular algorithms used by the research community in the past decade are Decision Tree, Support Vector Machine, Artificial Neural Network, Naïve Bayes and Regression analysis. These algorithms are considered due to their efficiency, reliability and popularity. [18], [16]

- Decision Tree: Decision Trees (DT) is a classification model. The tree-liked structure of DT with set of decisions for generating classification rules for a specific data. There are different variations such as C4.5, Classification and Regression Trees (CART). In these tree structures, the decision variable are presented as leaf nodes and branches of the tree presents the outcomes of attributes/features. DT presents good performance accuracy when applied to customer churn problem.
- Support Vector Machines: Support Vector Machines (SVM), also known as Support Vectors used for supervised learning. It was proposed by Boser, Guyon, and Vapnik [5]. SVM uses support vectors and analyze the given data to derive useful information. It is for both classification and regression analysis. SVM employ kernel functions for improving the performance. Selecting the proper hyperplane or combination of different hyperplane is still an open research. For customer churn prediction problem, mostly SVM outperform DT.
- Artificial Neural Network: Artificial Neural Networks (ANN) is a well known approach to deal with typical classification problems, such as the customer churn prediction. Neural networks works on neurons associated with weights for each neuron. Different topology have been defined to make the learning system work for varied problems.
- Naive Bayes: A Bayes classifier is a probabilistic algorithm. It depends on Bayes' hypothesis. The features are independent of model, with priori and posteriori likelihood estimations. A Naive Bayes (NB) classifier expect that the frequency of a specific element of a class. The NB classifier accomplished great outcomes on the customer churn prediction for the telecom industry.
- Regression Analysis: Regression analysis depends on statistical model. It is utilized for assessing the connections among attributes/ features. It incorporates numerous systems for developing model up a few features.

Hence, this section discusses several exiting algorithms which are useful in prediction churn prediction. Now, next section will discuss several practices related to imbalance data-sets in many applications (in this smart era).

4 Numerical Experiments and Results

The data set is openly available on internet and widely used in the literature for customer churn prediction in telecom. Table 1 show the information about the characteristics of the data used in the work. Orange dataset without preprocessing are provided online and has an Imbalance Ratio (IR) of 7.3% .

Table 1. Telecom datasets characteristics

	Orange
Total Instances	50,000
Total Features	260
IR Rate	7.3%

4.1 Performance measures

In this section, we present the important evaluation metrics to assess the performance of the classifiers when trained on telecom datasets. F-measure, Area Under Curve (AUC) and accuracy based methods are used to evaluate the prediction performance. The formula for calculating F-Measure is shown in equation 1 whereas AUC is computed as shown in equation 2.

$$F - Measure = 2 * (Precision * Recall) \,/\, (Precision + Recall) \tag{1}$$

$$AUC = \frac{1}{2} \times \left(\frac{TruePositive}{TruePositive + FalseNegative} + \frac{TrueNegative}{TrueNegative + FalsePositive} \right) \tag{2}$$

4.2 Results

In this section, we compared the results using different state-of-the-art techniques namely, Tomek Link, Cluster based Undersampling, C-Nearest Neighbors, Edit Nearest Neighbors, Near Miss, RUS on single (J48) and ensemble (Adaboost) classification algorithms. The metric used are AUC, F-Score and G-Mean. Table 2-3 shows the results of different state-of-the-art techniques compared on two different classification algorithms.

From the experimental results we observe that the performance of the Adaboost algorithm using different under sampling techniques is consistent. Figure 1-2 shows the comparison of various state-of-the-art techniques using two different classifiers. From the results, we observe that ensemble (Adaboost) algorithm outperform when compared with single (J48) classifier. It is also observed that the classifier show bias towards majority class and showed high performance results.

Table 2. Performance Results of J48 classification algorithm

Classification Algorithm (J48)	TP	Precision	Recall	F-Measure	ROC
Complete Data	0.923	0.922	0.923	0.922	0.896
Tomek Link	0.923	0.923	0.923	0.923	0.888
Cluster based Undersampling	0.881	0.882	0.881	0.881	0.885
C-Nearest Neighbors	0.877	0.877	0.877	0.877	0.89
Edit Nearest Neighbors	0.877	0.877	0.877	0.877	0.89
Near Miss	0.877	0.877	0.877	0.877	0.89
RUS	0.875	0.877	0.875	0.875	0.872

Table 3. Performance Results of AdaBoost classification algorithm

Classification Algorithm (AdaBoost)	TP	Precision	Recall	F-Measure	ROC
Complete Data	0.93	0.929	0.93	0.929	0.945
Tomek Link	0.936	0.935	0.936	0.935	0.948
Cluster based Undersampling	0.895	0.895	0.895	0.895	0.945
C-Nearest Neighbors	0.896	0.897	0.896	0.896	0.95
Edit Nearest Neighbors	0.896	0.897	0.896	0.896	0.95
Near Miss	0.896	0.897	0.896	0.896	0.95
RUS	0.904	0.905	0.904	0.904	0.94

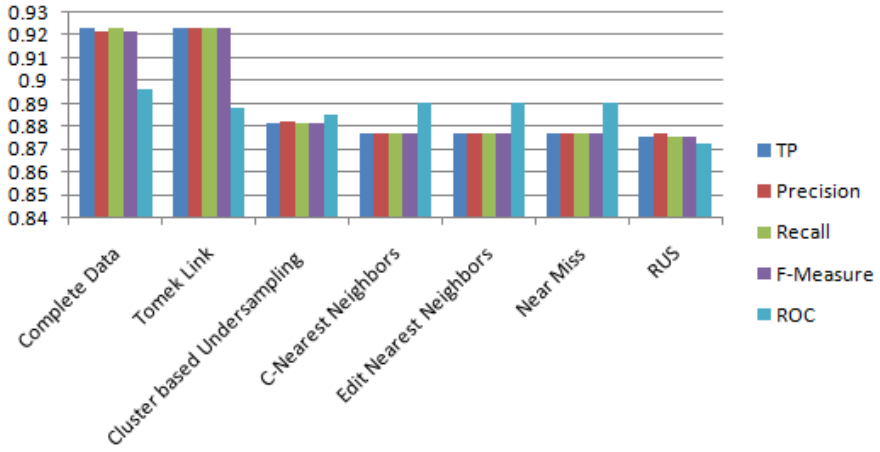

Fig. 1. Performance Results of J48 classification algorithm

5 Conclusion

In this paper, we conducted a study using various under sampling techniques on customer churn prediction problem. We investigated undersampling techniques like using decision tree and with boosting. Thus, this paper outlines the

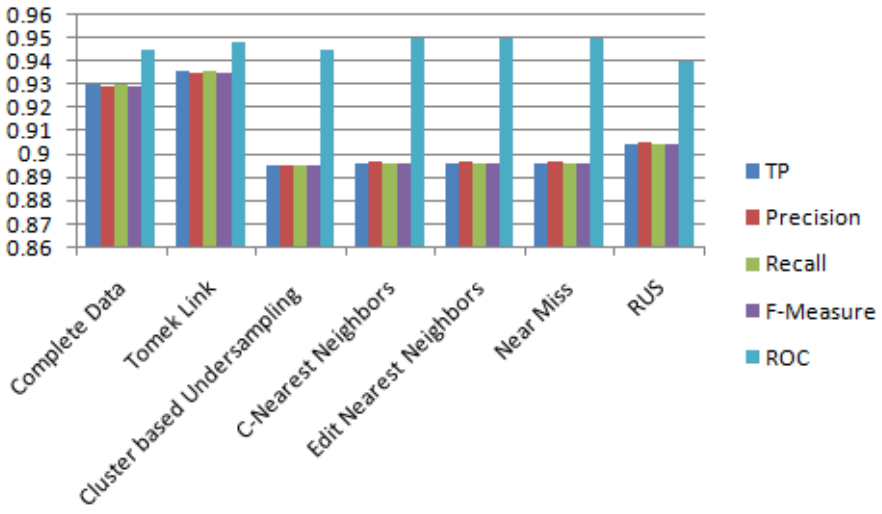

Fig. 2. Performance Results of AdaBoost classification algorithm

various undersampling techniques Tomek-link, Cluster-based under-sampling, C-Nearest Neighbor, edited nearest neighbor, Near Miss, Random Under-sampling using single and ensemble classifiers in solving the credit card churn prediction problem. From the results, we observe that ensemble (Adaboost) algorithm outperform when compared with single (J48) classifier.

References

1. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., Hussain, A.: Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access **4**, 7940–7957 (2016)
2. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ data science **4**(1), 10 (2015)
3. Brânduşoiu, I., Toderean, G., Beleiu, H.: Methods for churn prediction in the pre-paid mobile telecommunications industry. In: 2016 International conference on communications (COMM). pp. 97–100. IEEE (2016)
4. Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Systems with Applications **36**(3), 4626–4636 (2009)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
6. Gillala Rekha, Tyagi, A.K.: Cluster-based under-sampling using farthest neighbour technique for imbalanced datasets. In: 10th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2019). Springer (2019)

7. Gillala Rekha, V Krishna Reddy, A.K.T.: A novel approach for solving skewed classification problem using cluster based ensemble method. Mathematical Foundations of Computing (2020)
8. Hart, P.: The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory **14**(3), 515–516 (1968)
9. He, Y., He, Z., Zhang, D.: A study on prediction of customer churn in fixed communication network based on data mining. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. vol. 1, pp. 92–94. IEEE (2009)
10. Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J.: Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data. pp. 607–618 (2015)
11. Idris, A., Khan, A., Lee, Y.S.: Genetic programming and adaboosting based churn prediction for telecom. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 1328–1332. IEEE (2012)
12. Kraljević, G., Gotovac, S.: Modeling data mining applications for prediction of prepaid churn in telecommunication services. Automatika **51**(3), 275–283 (2010)
13. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2008)
14. Makhtar, M., Nafis, S., Mohamed, M., Awang, M., Rahman, M., Deris, M.: Churn classification model for local telecommunication company based on rough set theory. Journal of Fundamental and Applied Sciences **9**(6S), 854–868 (2017)
15. Penrod, C., Wagner, T.: Another look at the edited nearest neighbor rule. Tech. rep., TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING (1976)
16. Rekha, G., Tyagi, A.K.: Necessary information to know to solve class imbalance problem: From a user's perspective. In: Proceedings of ICRIC 2019, pp. 645–658. Springer (2020)
17. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. International Journal of Hybrid Intelligent Systems (Preprint), 1–10 (2019)
18. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: A wide scale classification of class imbalance problem and its solutions: A systematic literature review. Journal of Computer Science **15**, 886–929 (2019)
19. Tomek, I., et al.: Two modifications of cnn. (1976)