2017 International Conference on Frontiers of Information Technology

Resolving class imbalance and feature selection in customer churn dataset

Aamer Hanif
Department of Computer Science
Air University
Islamabad, Pakistan
ahanif@mail.au.edu.pk

Noor Azhar
Department of Computer Science
Air University
Islamabad, Pakistan
noorieazhar@gmail.com

***Abstract*— Churn prediction datasets pertaining to telecom sector often have the class imbalance problem. Due to large number of features, dimensionality reduction (or feature selection) and dataset balancing become important data preprocessing steps. This research utilizes a real dataset to classify defecting customers in the telecom sector. Three different feature selection and dataset balancing techniques are applied for data preprocessing before classification model building. The results show that random oversampling performed better to balance the dataset and the three feature selection techniques used performed equally well. Customer call related features are extracted as features that are more important. The classification model is built using random forest technique and model evaluation measures are computed and reported. Conduct of experiments on a real dataset that does not have any customer demographic variables is a significant contribution of this paper.**

Keywords — customer churn, dataset balancing, dimensionality reduction, feature selection

I. INTRODUCTION

Retaining existing customers in the face of extreme market competition is a big challenge for mobile telecom companies in the world and in Pakistan as well. Various mobile telecom advertisements running to attract customers also highlight this fact since these companies offer attractive packages to customers and try to win back defecting customers by offering them lucrative plans at low costs. In Pakistan, quality of service has been of major concern and customers readily exercise the option to quit a company's service for similar reasons and to subscribe to another service provider. This generates a challenge for each company to be able to predict customer behavior, to understand causes of customer attrition and keep it in control. Another reason companies would want to predict churn for customer retention is because loyal customers are of substantial value to the business as they bring in more revenue by disseminating positive word of mouth to others as compared to the dissatisfied customers. Besides, it costs less to retain existing customers than to acquire new customers [1][2]. Therefore, it is imperative to identify those customers in advance who are likely to leave the company in future to avoid significant loss of revenues for the business and this process is called churn prediction. Since the telecom companies have their customer base in millions, even a small percentage of churners will cause financial losses to the company who will find it even more costly to attract new customers. Therefore, customer retention is crucial for business to remain profitable [3].

To solve this problem of classifying customers, data mining techniques and statistical tools are employed on customer data that is generated and gathered by the telecom companies in large volumes. Usually, this data includes customer billing information and usage in terms of calls, texts and data plans. To extract valuable knowledge from this customer data, past researches have made extensive use of data mining processes employing machine learning algorithms, statistics, pattern recognition, and visualization techniques [4][5][6]. By using these techniques, classification and prediction models are implemented to discover patterns in customer data that allow telecom companies to make informed decisions based upon knowledge extracted from the data. The classification model performs a binary classification task differentiating between churners and non-churners and can then classify future unlabeled records of customers.

In this paper, we present a classification model which helps in identifying customers that are at the risk of churning and must be retained, while dealing with the problem of class imbalance through three re-sampling methods and by employing feature selection techniques as well.

II. PREVIOUS WORK

Although customer churn prediction is not a new research area, there is still a continued focus of researchers in this area to employ and evaluate data mining techniques [7]. A Google scholar search with the keywords "customer churn prediction" yielded over four dozen papers published since 2015 to date in this area displaying this as an active research track. Comparative analysis of various techniques for customer churn prediction specifically in the telecommunications sector happens to be frequently studied [8, 9, 10, 11].

Many techniques have been proposed and used by the research community to solve the customer churn prediction problem. The most popular techniques used include Artificial Neural Networks [12, 13, 14, 15, 16], Decision Trees learning [17, 18, 18, 20], Regression Analysis [21, 22, 23] and Support Vector Machines [24, 25, 26, 27] besides other techniques. Moreover, feature section and handling the class imbalance problem have also been studied extensively [28, 29, 30].

0-7695-6347-3/17/$31.00 ©2017 IEEE
DOI 10.1109/FIT.2017.00022

The motivation behind this paper is to study dataset balancing techniques along with feature selection before building a classification model using the random forest technique on actual customer churn data from the telecom sector.

III. CLASS IMBALANCE PROBLEM

Class imbalance problem has been comprehensively researched in literature [11, 29]. This problem occurs when the occurrences of one class outnumbers the occurrences of other classes. The majority class may be extensively larger in number than the minority class in a dataset and the minority class is relatively rare because it does not occur as widely as the majority class. The problem arises when the minority class in more interesting to the researcher and of more value to the organization. Some examples where imbalanced datasets exist are customer churn cases, fraud detection cases, network attacks detection and medical diagnosis cases involving relatively low percentage of sick people as compared to healthy individuals.

Sampling is the most common technique to deal with imbalanced classes and it works by altering the distribution of training examples [31]. This research uses three techniques to deal with class imbalance problem. These techniques are random oversampling, random under-sampling and synthetic minority oversampling.

A. Random oversampling

This technique works with the minority class as it increases the number of instances in the minority class by randomly replicating them in order to provide larger representation of the minority class in the sample.

B. Random under-sampling

This method works with majority class by reducing the number of observations from majority class to make the data set balanced. The strength of this method is improving run time and reducing storage requirements of large datasets.

Both of these techniques reduce the class imbalance by decreasing the rarity of the minority class. These techniques have drawbacks as under-sampling discards potentially useful majority class samples thereby degrading classifier performance while oversampling may increase time to build the classifier. Moreover, oversampling introduces no new data thereby making under-sampling a better choice [32]. Due to these issues, the third advanced sampling technique was also studied.

C. Synthetic minority oversampling technique (SMOTE)

The over-sampling approach called SMOTE works in a way in which the minority class is over-sampled by creating “synthetic” examples rather than by randomly replicating the minority class real data examples [33]. Depending upon the amount of over-sampling required, the minority class examples are generated by adding examples from line segments that join the k minority-class nearest neighbors (SMOTE uses k=5). This method generalizes the examples as compared to the specialization that occurs due to replicating examples.

D. Model Evaluation Measures

Evaluation measures are used to evaluate model performance in machine learning. Accuracy is the most widely used measure but it is inappropriate to handle imbalanced classes as the overall accuracy tends to be biased towards the majority class. Consider a dataset with 90 percent examples from majority class and 10 percent from minority class. A model classifying everything from the majority class will have an accuracy of 90% which is misleading as the model does not classify anything from the minority class. Hence other measures like sensitivity, specificity, precision and F measure are used.

TABLE I. CONFUSION MATRIX

		Predicted Class	
Actual Class		+ve	-ve
	+ve	True Positive (TP)	False Negative (FN)
	-ve	False Positive (FP)	True Negative (TN)

From the confusion matrix provided in Table I, the following model evaluation measures can be derived:

Sensitivity (recall) = (TP)/(TP+FN)

Specificity = (TN) /(TN+FP)

Precision = (TP)/(TP+FP)

F measure = (2 x precision x recall) / (precision + recall)

IV. FEATURE SELECTION

Feature selection is another critical issue in machine learning and data mining. The idea is to choose important features aiming to improve performance of the classification model. Presence of irrelevant features in high dimensional data may reduce the performance of the classifier and increase the misclassification rate especially in imbalanced data sets [34, 35]. Two reasons for using feature selection are (a) reducing the number of features, to reduce overfitting and improve the generalization of models, and (b) understanding the data features and their relationship to the class labels. Three techniques described below have been used in this research for feature selection.

A. Random forest

Being an ensemble method for classification, regression and other tasks, random forests construct a mass of decision trees during training phase and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. At each node in the tree, a random selection of subset of features is made and the best split available with those features is selected for that node. Thus, by pruning trees below a particular node, we can create a subset of the most important features [31].

B. *Gradient Boosting*

Gradient boosting involves a loss function to be optimized, a weak learner to make predictions and an additive model to add weak learners to minimize the loss function. Decision trees constructed in greedy manner are used as the weak learner in gradient boosting. A gradient descent procedure is used to minimize the loss when adding trees. The output for the new tree is then added to the output of the existing sequence of trees to correct or improve the final output of the model [36].

C. *MRMR Feature Selection*

In mRMR (minimum redundancy maximum relevance) technique, mutual information defining the correlation among features is taken as the basic criterion to find the feature relevance and redundancy. The mutual information between a feature and class labels defines the relevance of that feature. The goal of this method is to find the most relevant and least redundant features set. The features contained by the final feature set have maximum relevance and least correlation [37].

V. DATA AND METHODOLOGY

A. *Data*

The dataset used for this research is a real dataset provided by a local telecom company. The original dataset contained customer data for three months. It had 800,000 instances and 1947 attributes. The data was primarily based upon counts and revenues for call, SMS and GPRS services differentiated over weekdays, weekends, day, night and aggregated over weeks and months grouped by various plans that the company offers for off net, on net, friends and family etc. The dataset was imbalanced as the proportion of churners was only 8% as compared to the majority class. The data was subjected to preprocessing in which columns that had over 80% null values were removed and columns that were used in building aggregates were also dropped to avoid duplication of information. Thus a new and modified dataset was formed that had 248 attributes and randomly sampled down to 75,000 rows for ease of conducting the experiments and ensuring that this sample size was still larger than that used in many published papers. The original proportion of churners was preserved in this modified smaller dataset. It is pertinent to mention here that no demographic attributes were available in the provided real dataset making this research different from most previous ones which have some demographic variables in the datasets.

B. *Experiments*

The dataset was subjected to three techniques for handling the class imbalance problem. Random oversampling, under-sampling and SMOTE techniques were used to balance the dataset. After balancing the dataset, results of a classification model using random forest technique were obtained and model evaluation measures were calculated. After this step, feature selection was done using three techniques (random forest, gradient boosting and mRMR) and a classification model was built using random forest technique again. All experiments were conducted using the R language with help of ROSE, DMwR, caret, randomForest, mRMRe and ROCR packages [38].

VI. RESULTS

In the first experiment, sampling techniques were performed to balance the dataset. Sampling is done with three different proportions i.e. {80%, 20%}, {70%, 30%}, {75%, 25%} and with three different sampling methods as discussed above. The evaluation measures (F measure, sensitivity, specificity and precision) are compared by building a random forest model on datasets sampled with the three proportions mentioned.

Random oversampling method has given the best results with highest sensitivity as well as highest specificity among all. SMOTE, also a type of oversampling method has given the best specificity along with the highest precision. On the other hand, random under-sampling performed well detecting true non churners but didn't perform well detecting true churners which in our case is our main concern.

Seeing the results above, we can say that random oversampling has performed the best. The results of these experiments are given in Table II.

TABLE II. COMPARISON OF BALANCING TECHNIQUES

Technique	**Sampling ratio**	**Evaluation Measures**			
	Non-churn, churn	*F meas*	*Sens*	*Spec*	*Prec*
Random oversampling	[70%,30%]	0.96	0.95	0.98	0.98
	[75%,25%]	0.93	0.91	0.98	0.96
	[80%,20%]	0.90	0.84	0.99	0.96
Random undersampling	[70%,30%]	0.45	0.32	0.94	0.77
	[75%,25%]	0.60	0.25	0.95	0.65
	[80%,20%]	0.30	0.19	0.97	0.63
SMOTE	[70%,30%]	0.87	0.83	0.99	0.93
	[75%,25%]	0.90	0.84	0.99	0.98
	[80%,20%]	0.91	0.85	0.99	0.99

The results of low sensitivity in the case of under-sampling need to be explained more. One reason could be training the classifier using few samples because more samples are discarded when under-sampling when the dataset is highly imbalanced. That way, potentially useful information is lost. Therefore, this lack of data results in development of a poor classifier as evident here. Moreover, these results do not incorporate feature selection yet. That means relying on potentially bad features which do not have sufficient discriminative power, therefore, having more data of the same type is not helping.

The second experiment involved feature selection. Table III shows the results of using three methods as discussed above. Three different feature selection techniques were applied on the sampled data to get the top 10 features out of 284 features that were most contributing towards predicting true churners. It can be seen that all the techniques have given the highest variable importance to the features consisting of calling data. i.e. inbound calling, outbound calling, on network calling, total calling minutes etc. Hence, call related features are more important as compared to SMS or other features when it comes to predicting the churners.

TABLE III. FEATURE SELECTION RESULTS

Feature Sel Technique	Top ten variables selected
mRMR	W4_N_V_IB_min, WD_N_V_IB_min, WE_N_V_IB_min, W3_N_V_OnNet_IB_min, W2_N_V_OnNet_IB_min, W2_N_V_IB_min, W3_N_V_IB_min, W1_N_V_IB_min, WE_OnNet_V_IB_min, W2_N_V_OB_revenue
Gradient boosting	W1_D_V_IB_cnt, W2_ON_V_IB_cnt, W2_D_V_IB_cnt, W1_ON_V_IB_cnt, W3_D_V_IB_cnt, W4_D_V_IB_cnt, W1_V_OB_CD, W3_ON_V_IB_cnt, W4_V_OB_CD, M3_D_CD
Random Forest	W1_D_V_IB_cnt, W2_D_V_IB_cnt, W1_ON_V_IB_cnt, W1_D_V_IB_min, W2_ON_V_IB_cnt, W4_D_V_IB_cnt, W3_D_V_IB_cnt, Total_Revenue, W1_ON_V_IB_min, W2_D_V_IB_min
Codes: W week, N night, V voice, min minutes, IB/OB in/outbound, D day, M month, cnt count, CD call duration	

Data was divided randomly into proportions of 70-30, with 70% for training and 30% for testing by createDataPartition() method of caret library of R software. The dependent variable was converted to factor because then the random sampling occurs within each class but does not disturb the overall class distribution of the dataset.

TABLE IV. CLASSIFICATION MODEL OUTCOME

Evaluation Measures	Feature Selection Technique		
	MRMR	Gradient Boosting	Random Forest
F-measure	0.95	0.97	0.96
Sensitivity	0.99	0.99	0.99
Specificity	0.95	0.97	0.97
Precision	0.93	0.96	0.95

The results of three different feature selection techniques were applied on the sampled dataset. mRMR being a filter technique while Random forest and gradient boosting being wrapper techniques.

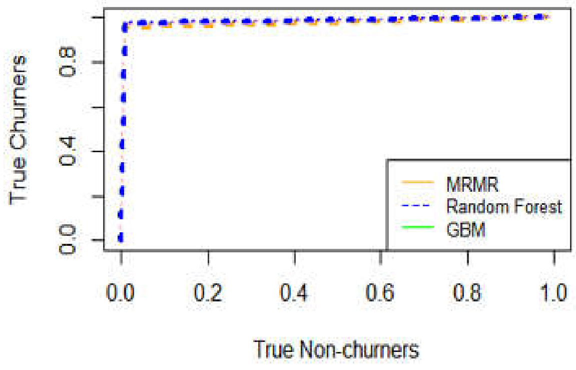

Fig. 1. ROC curves for classification model

All the method have performed very well in detecting the churners giving high sensitivity values. Table IV provides the model evaluation measures for this experiment. Gradient boosting and Random forest performed better as compared to mRMR when classifying non churners giving high specificity as well as precision.

In order to assess the accuracy of a classifier independent of any threshold, ROC (Receiver operating characteristics) analysis is used. The ROC curve results of three different feature selection techniques are given in Fig. 1. The measure of quality of a probabilistic classifier can be provided by the area (AUC) under ROC curve. Any random classifier for example using a coin to decide the classes has an area under curve 0.5, while a perfect classifier has 1. Hence, that AUC value in practice should be close to 1 for a good classifier. AUC values for the three models are given below. All values are close to 1 showing good results.

- mRMR: Area under the curve (AUC): 0.974
- Gradient boosting: Area under the curve (AUC): 0.985
- Random Forest: Area under the curve (AUC): 0.981

VII. Conclusions

Classifying churners is still an area of active research specially in the telecom business. Presence of class imbalance problem in these datasets and solution to these problems has received great interest of the research community. In this research, three methods were utilized to handle the class imbalance problem. Moreover, three methods for feature selection have also been used. To conclude the research, a random forest classification model was built to evaluate and compare the studied techniques. To study the impact of these methods, a real life dataset was used making it a positive aspect of this research since it does not use artificial data like many other published papers. Lack of any demographic variables was another challenge in the dataset as it made the classification task more challenging. Random oversampling gave the best results in this research to balance the dataset, however, all three feature selection techniques highlighted only the call related features as most important and performed equally well.

In future, the authors plan to use deep learning neural networks to predict customer churn using the same dataset.

References

[1] Yang, Zhilin, and Robin T. Peterson. "Customer perceived value, satisfaction, and loyalty: The role of switching costs." Psychology & Marketing 21, no. 10 (2004): 799-822.

[2] Kim, Moon-Koo, Myeong-Cheol Park, and Dong-Heon Jeong. "The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services." Telecommunications policy 28, no. 2 (2004): 145-159.

[3] Dalvi, Preeti K., Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, and V. A. Kanade. "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression." In Colossal Data Analysis and Networking (CDAN), Symposium on, pp. 1-4. IEEE, (2016).

[4] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications23, no. 2 (2002): 103-112.

[5] Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications 38, no. 3 (2011): 2354-2364.

[6] Huang, Bingquan, Mohand Tahar Kechadi, and Brian Buckley. "Customer churn prediction in telecommunications." Expert Systems with Applications39, no. 1 (2012): 1414-1425.

[7] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications23, no. 2 (2002): 103-112.

[8] Vafeiadis, Thanasis, Konstantinos I. Diamantaras, George Sarigiannidis, and K. Ch Chatzisavvas. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[9] Hassouna, Mohammed, Ali Tarhini, Tariq Elyas, and Mohammad Saeed AbouTrab. "Customer Churn in Mobile Markets A Comparison of Techniques." arXiv preprint arXiv:1607.07792 (2016).

[10] Coussement, Kristof, Stefan Lessmann, and Geert Verstraeten. "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry." Decision Support Systems 95 (2017): 27-36.

[11] Amin, Adnan, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, and Amir Hussain. "Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study." IEEE Access 4 (2016): 7940-7957.

[12] Tsai, Chih-Fong, and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." Expert Systems with Applications 36, no. 10 (2009): 12547-12553.

[13] Sharma, Anuj, Dr Panigrahi, and Prabin Kumar. "A neural network based approach for predicting customer churn in cellular network services." arXiv preprint arXiv:1309.3945 (2013).

[14] Pendharkar, Parag C. "Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services." Expert Systems with Applications 36, no. 3 (2009): 6714-6720.

[15] Mozer, Michael C., Richard Wolniewicz, David B. Grimes, Eric Johnson, and Howard Kaushansky. "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry." IEEE Transactions on neural networks 11, no. 3 (2000): 690-696.

[16] Song, Guojie, Dongqing Yang, Ling Wu, Tengjiao Wang, and Shiwei Tang. "A mixed process neural network and its application to churn prediction in mobile communications." In Data Mining Workshops, (2006). ICDM Workshops 2006. Sixth IEEE International Conference on, pp. 798-802. IEEE, 2006.

[17] Bin, Luo, Shao Peiji, and Liu Juan. "Customer churn prediction based on the decision tree in personal handyphone system service." In Service Systems and Service Management, International Conference on, pp. 1-5. IEEE, (2007).

[18] Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying. "Customer churn prediction using improved balanced random forests." Expert Systems with Applications36, no. 3 (2009): 5445-5449.

[19] Lemmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research 43, no. 2 (2006): 276-286.

[20] Glady, Nicolas, Bart Baesens, and Christophe Croux. "Modeling churn using customer lifetime value." European Journal of Operational Research 197, no. 1 (2009): 402-411.

[21] Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason. "Defection detection: Measuring and understanding the predictive accuracy of customer churn models." Journal of marketing research 43, no. 2 (2006): 204-211.

[22] Günther, Clara-Cecilie, Ingunn Fride Tvete, Kjersti Aas, Geir Inge Sandnes, and Ørnulf Borgan. "Modelling and predicting customer churn from an insurance company." Scandinavian Actuarial Journal 2014, no. 1 (2014): 58-71.

[23] Backiel, Aimée, Bart Baesens, and Gerda Claeskens. "Mining telecommunication networks to enhance customer lifetime predictions." In International Conference on Artificial Intelligence and Soft Computing, pp. 15-26. Springer, Cham, (2014).

[24] Farquad, Mohammed Abdul Haque, Vadlamani Ravi, and S. Bapi Raju. "Churn prediction using comprehensible support vector machine: An analytical CRM application." Applied Soft Computing 19 (2014): 31-40.

[25] Rodan, Ali, Hossam Faris, Jamal Alsakran, and Omar Al-Kadi. "A support vector machine approach for churn prediction in telecom industry." International Information Institute (Tokyo). Information 17, no. 8 (2014): 3961.

[26] Chen, Zhen-Yu, Zhi-Ping Fan, and Minghe Sun. "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data." European Journal of operational research 223, no. 2 (2012): 461-472.

[27] He, Benlan, Yong Shi, Qian Wan, and Xi Zhao. "Prediction of customer attrition of commercial banks based on SVM model." Procedia Computer Science 31 (2014): 423-430.

[28] Xiao, Jin, Ling Xie, Changzheng He, and Xiaoyi Jiang. "Dynamic classifier ensemble model for customer classification with imbalanced class distribution." Expert Systems with Applications 39, no. 3 (2012): 3668-3675.

[29] Maldonado, Sebastián, Richard Weber, and Fazel Famili. "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines." Information Sciences 286 (2014): 228-246.

[30] Xiao, Jin, Yi Xiao, Anqiang Huang, Dunhu Liu, and Shouyang Wang. "Feature-selection-based dynamic transfer ensemble model for customer churn prediction." Knowledge and information systems 43, no. 1 (2015): 29-51.

[31] Burez, Jonathan, and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." Expert Systems with Applications 36, no. 3 (2009): 4626-4636.

[32] Weiss, Gary M. "Mining with rarity: a unifying framework." ACM Sigkdd Explorations Newsletter 6, no. 1 (2004): 7-19.

[33] N. V. Chawla et al., “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, (2002).

[34] Chomboon , K. Kerdprasop and N. Kerdprasop, Rare Class Discovery Techniques for Highly Imbalance Data. Proc. International multi conference of engineers and computer scientists, vol. 1, (2013).

[35] L. Lusa and R. Blagues, “ The Class-imbalance for highdimensional class prediction,” in 11th International Conference on Machine Learning and Application, IEEE, (2012).

[36] Xu, Zhixiang, Gao Huang, Kilian Q. Weinberger, and Alice X. Zheng. "Gradient boosted feature selection." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 522-531. ACM, (2014).

[37] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence27, no. 8 (2005): 1226-1238.

[38] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2003) URL http://www.R-project.org/.