Proceedings of IOE Graduate Conference, 2019-Summer
Peer Reviewed
Year: 2019 Month: May Volume: 6
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Customer Churn Prediction Using ADASYN Sampling Technique and Ensemble Model

Sabina Pun [a], Sharan Thapa [b], Suresh Timilsina [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pashchimanchal Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] sabina22pun@gmail.com, [b] sharant@ioepas.edu.np [c] timilsinasurace@gmail.com

Abstract

With the growing number of companies in service-based industries, such as telecommunication industry, banking industry, and insurance industry, there is high competition in the market and such companies are prone to customer churn, which is a serious problem. Customer churn reduces the company's revenue and slows down the company's growth. Early prediction of the potential churning customers can be helpful in timely retaining them. Churn datasets often suffer from class imbalance problem, which means the number of non–churn data samples are much more than that of churn data samples. This introduces learning bias towards the non–churn class. The customer churn prediction model presented in this paper implements ADASYN (Adaptive Synthetic Sampling Technique) sampling technique to address the class imbalance problem, and uses the ensemble of three popular classifiers Support Vector Machine, Logistic Regression, and Backpropagation algorithm, to improve the prediction results. The performance of the model is evaluated using different evaluation metrics such as F1-score, and AUC. As per the experimental results, the model has F1-score 0.68, and AUC 0.90 respectively.

Keywords

ADASYN, backpropagation, customer churn, ensemble, class imbalance, logistic regression, support vector machine

1. Introduction

Customers are the crucial parts of any company and for the companies in service-based industries, such as telecommunication industry, banking industry, and insurance industry, it is important to have strong customer base. Customers have multiple options for service they want to use since there are many companies providing the same service. Customers can switch between the companies anytime they wish to. So, customer churn is a serious concern. Customer churn is when the customer ends his or her relationship with the company and/ or switches to the competitor's company. Due to the exponential growth of the use of such services and high competition in the market, these companies are very much prone to customer churn. Several reasons for the customer churn are dissatisfaction with service, high service cost, lack of quality in service, lack of customer value, interesting offers from competitor's companies, and so on. Customer churn has negative impact on the revenue of the company, and slows down the company's growth. Also, the cost of acquiring the new customer is five to six times more than retaining the existing one [1]. Customer retention is useful to increase the company's revenue and for the company's growth .Timely identifying the churning customers helps company save its resources and time, and also helps company take appropriate steps for retaining those customers. Therefore, customer churn prediction can be helpful in timely retaining the churning customers. Customer churn prediction is the two-class classification problem that classifies the customer as either churn (the minority class) or non-churn (the majority class). Building the churn prediction model with highly accurate prediction is very crucial. Real life datasets are often imbalanced, i.e. there is vast different between the number of majority class samples and the number of minority class samples. So, some sort of techniques needs to be used in order to deal with such imbalanced datasets. Also, there are numbers of classification methods in machine learning that can be used to build the prediction model. Integrating some of the popular classifiers and combining the decision of each classifier helps in providing more accurate prediction.

2. Literature Review

SMOTE (Synthetic Minority Oversampling Technique) is the over-sampling approach for focused learning that introduces a bias towards the minority class, thus improving the accuracy of classifiers for a minority class in the imbalanced dataset [2]. In SMOTE approach, the minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. He *et al.* introduced ADASYN sampling technique, one of the modified version of SMOTE, is adaptive learning approach for dealing with imbalanced datasets by adaptively generating the synthetic minority class samples according to their level of difficulty in learning [3].

Aditsania *et al.* have used adaptive synthetic sampling technique to deal with class imbalance problem and backpropagation algorithm for classification model [4]. Do *et al.* have implemented SMOTE for handling imbalanced dataset, and several classification models such as AdaBoost, Extra Trees, KNN, Neural Networks and XGBoost are separately used for building predictive models [5]. Mitkees *et al.* have implemented different data mining techniques such as clustering, classification and association rule to build customer churn prediction model [6]. Wu and Meng have presented customer churn prediction model based on improved SMOTE and AdaBoost [7]. Xia *et al.* have proposed weighted selective ensemble model [8]. In this model, the base classifiers used were Bayes, C4.5, ANN and SVM, and the final classification decision was the weighted combination of the individual classifiers within the ensemble.

3. Methodology

3.1 Proposed Framework

Figure 1 presents the proposed customer churn prediction framework. The goal of this framework is to predict whether the data samples fall into either churn class (minority class) or non-churn class (majority class). The process starts with preparing the dataset. Often, the real life datasets are in raw form. So, it is important to clean and inspect the dataset in order to make the dataset appropriate for further processing. The tasks such as checking for missing values, converting categorical values into numeric values, etc., are performed in this stage. After the dataset is ready, informative features are selected from the feature space in feature selection stage. Thereafter, the dataset is split into the training set and test set. Training set is used to train classifiers and test set is used for making prediction. Now, only the training set is resampled using ADASYN sampling technique to balance the number of minority class samples and majority class samples in the training set. Once the training set is resampled, the resampled training set and the test set are fed to the base classifiers. The base classifiers used in this ensemble model are Support Vector Machine (SVM), Logistic Regression, and Backpropagation algorithm. The prediction output of each base classifier is then passed to the voting classifier. Here, the soft voting method is used, in which the average predicted probabilities of the base classifiers are used for predictions. The output of the voting classifier is the final prediction result. Finally, the performance of the proposed framework is evaluated using different evaluation metrics such as F1-score, and AUC.

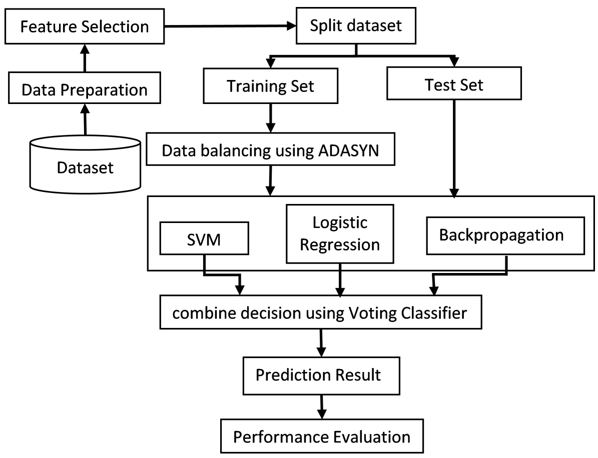

Figure 1: The proposed customer churn prediction framework

3.2 ADASYN Sampling Technique

ADASYN sampling technique is one of the sampling techniques for handling imbalanced class problem. In this approach, synthetic minority data samples are adaptively generated according to their level of difficulty in learning, where more synthetic data is generated for minority class samples that are harder to learn compared to those minority class samples that are easier to learn. It not only reduces the bias

introduced by the class imbalance, but also adaptively shifts the classification decision boundary toward those difficult to learn samples.
The algorithm for ADASYN sampling technique is executed as follows:

1. Calculate the number of synthetic samples that needs to be generated as:

$$G = (|s_{\text{maj}}| - |s_{\text{min}}|) \times \beta, \beta \in [0,1] \quad (1)$$

   where s_{maj} and s_{min} are the number of majority class samples and the number of minority class sample, and β is a parameter used to specify the desired balanced level after generation of the synthetic data.

2. For each sample $x_i \in s_{\text{min}}$, find K–nearest neighbors and calculate ratio

$$r_i = \Delta_i \div K, i = 1, ..., s_{\text{min}} \quad (2)$$

   where Δ_i is the number of samples in the K-nearest neighbors of x_i that belong to the majority class.

3. Calculate number of synthetic samples to be generated for each x_i using

$$g_i = r_i \times G \quad (3)$$

4. For each minority class sample x_i, generate g_i synthetic data samples according to the following steps:
   Do the loop from 1 to g_i

   (a) Randomly choose one minority data sample x_{zi} from K nearest neighbors for x_i.
   (b) Generate the synthetic data sample using

$$s_i = x_i = (x_{zi} + x_i) \times \lambda, \lambda \in [0,1] \quad (4)$$

   λ is a random number.

   End loop

3.3 Base Classifiers

Support Vector Machine is a supervised machine learning algorithm often used for classification tasks. In this algorithm, each data sample is plotted as a point in n–dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the optimal hyperplane that differentiate two classes very well. Hyperplane can be a point in case of 1-dimensional data, line in case of 2-dimensional data, plane in case of 3-dimensional data, and so on. Selecting the optimal hyperplane maximizes the margin of the training data. Such optimal hyperplane is likely to generalize better and correctly classifies new unseen data samples.
Logistic Regression is another popular technique in machine learning for binary classification problems. It uses the logistic function (also called the sigmoid function), which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
Backpropagation Algorithm is a supervised learning algorithm for training Multi-layer Perceptrons. It is a network of two or more layers of nodes where one layer is fully connected to the next layer. The different layers are input layer, hidden layers, and output layer. The principle of the backpropagation approach is to model a given function by modifying internal weights of inputs to produce an expected output. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state.

3.4 Voting Classifier

Voting classifier combines the predictions of base classifiers by averaging those predictions. The soft voting method or the hard voting method can be used for voting classifier. Soft voting averages the base classifiers' predictions based on the predicted probabilities and the class label with the highest average probability is the final prediction result, whereas hard voting averages the predicted class labels and the class label with majority vote is the prediction made by the ensemble.

3.5 Performance Measure

For the imbalanced dataset, accuracy is can be misleading to evaluate the performance of the classifier because in the imbalanced dataset, the accuracy of the learner is high even when the classifier classifies all the majority samples correctly and misclassifies all the minority samples since the number of majority samples are much more than the number of minority samples, and under such circumstance, accuracy cannot reflect reliable prediction for the minority class. Therefore, metrics beyond accuracy such as F1-score, and AUC are used to evaluate the performance of the proposed framework.

Table 1: Confusion Matrix

		Predicted churn No	Predicted churn Yes
Actual churn	No	TN	FP
	Yes	FN	TP

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1-score = \frac{2*Precision*Recall}{Precision+Recall}$$

AUC (Area under ROC curve) represents a model's ability to distinguish between positive (churn) class and negative (non-churn) class. It measures the entire two-dimensional area underneath the entire ROC curve from (0, 0) to (1, 1). It can be interpreted as the probability that the model ranks a random positive sample more highly than a random negative sample. AUC has a range of [0, 1]. An area of 1.0 represents a model that made all predictions perfectly and area of 0.0 represents that all predictions are wrong. The higher the value of AUC, the better is the classification performance of the model.

4. Results and Analysis

4.1 Dataset

Telecom churn dataset is used which is taken from Kaggle website. The dataset can be downloaded via the link[1]. Kaggle is an online platform for obtaining dataset. The dataset contains 20 telecom features and 3333 data samples, out of which 483 samples belongs to churn class and 2850 samples belongs to non-churn class. Among 20 features, 18 relevant features were used. The whole dataset is divided into 70% training set and 30%test set .

[1] https://www.kaggle.com/becksddf/churn-in-telecoms-dataset

4.2 Experimental Results

The proposed framework was implemented in Python 3. Different python packages such as imbalanced-learn [9], and scikit-learn [10] were used. Since confusion matrix gives a better idea of what the classification model is getting right and what types of errors it is making, here it is used to show the effectiveness of the resampling technique on the dataset.

Table 2: Confusion matrix for Logistic regression (without resampling)

		Predicted churn No	Predicted churn Yes
Actual churn	No	839	27
	Yes	104	30

Table 3: Confusion matrix for Logistic regression (with resampling)

		Predicted churn No	Predicted churn Yes
Actual churn	No	677	189
	Yes	33	101

Table 4: Confusion matrix for SVM (without resampling)

		Predicted churn No	Predicted churn Yes
Actual churn	No	852	14
	Yes	65	69

Table 5: Confusion matrix for SVM (with resampling)

		Predicted churn No	Predicted churn Yes
Actual churn	No	804	62
	Yes	46	88

Table 6: Confusion matrix for Backpropagation algorithm (without resampling)

		Predicted churn	
		No	Yes
Actual churn	No	850	16
	Yes	47	87

Table 7: Confusion matrix for Backpropagation algorithm (with resampling)

		Predicted churn	
		No	Yes
Actual churn	No	773	93
	Yes	40	94

Table 8: Confusion matrix for Ensemble model (with resampling)

		Predicted churn	
		No	Yes
Actual churn	No	793	73
	Yes	27	107

As it is more costly to miss a churn label than to falsely label a non-churn sample, our interest is in predicting the churn class correctly. With the use of resampling technique, we want to increase the number of true positive and decrease the number of false negative i.e. we want increase recall . From the Figure 2, it is observed that after using the resampling technique, recall is increased.

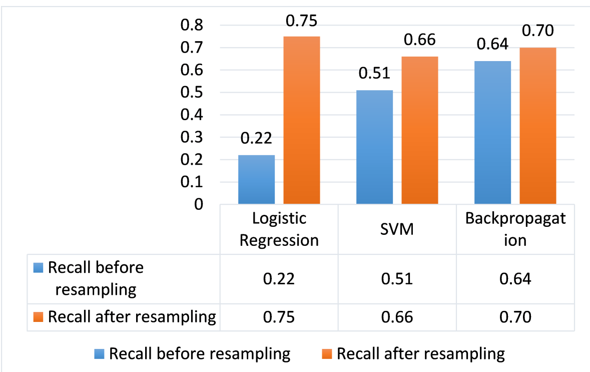

Figure 2: Recall before and after sampling

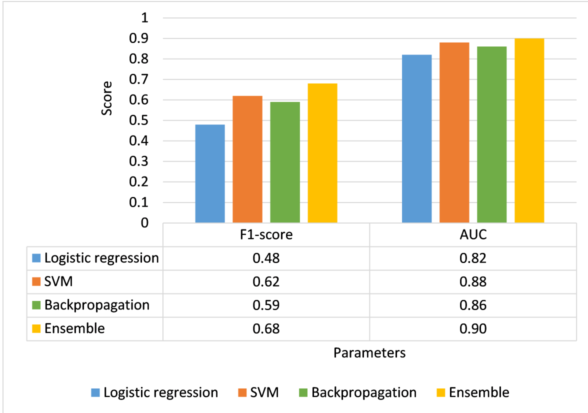

Figure 3: Performance metrics of the base classifiers and the ensemble model

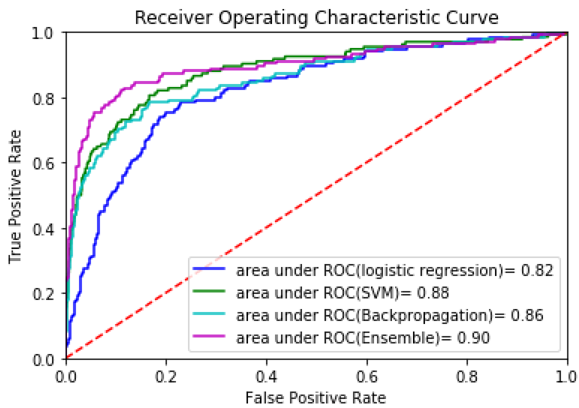

Figure 4: ROC curve for base classifiers and the ensemble model

The performance of the proposed framework is evaluated using F1-score and AUC. In the Figure 3, the different performance metrics of the base classifiers and the ensemble model are compared with each other, and Figure 4 shows the ROC curve of the base classifiers and the ensemble model . The Figure 3 shows that among the base classifiers, SVM has the highest F1-score of 0.62 and the highest AUC of 0.88. But, when all the classifiers are compared with each other, the ensemble model has the highest F1-score and AUC which is 0.68 and 0.90. The highest F1-score and AUC of the ensemble model indicates that the ensemble model performs better in comparison with the base classifiers.

5. Conclusion

In this paper, ADASYN sampling technique is used for resampling the dataset in order to handle class imbalance problem in the dataset. As per the experimental results, using resampling technique helps increase the number of true positive (i.e. increase the number of correctly predicted churn labels)and decrease the number of false negative. Also, the results show that the ensemble model improves the classification performance of the customer churn prediction model.

Acknowledgments

This work was supported by the University Grants Commission [UGC Masters Research Support award no.: MRS-75/76-Engg-3]. The authors are thankful to University Grants Commission for the research support fund.

References

[1] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, 2012.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

[4] Annisa Aditsania, Aldo Lionel Saonard, et al. Handling imbalanced data in churn prediction using adasyn and backpropagation algorithm. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pages 533–536. IEEE, 2017.

[5] Duyen Do, Phuc Huynh, Phuong Vo, and Tu Vu. Customer churn prediction in an internet service provider. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3928–3933. IEEE, 2017.

[6] Ibrahim MM Mitkees, Sherif M Badr, and Ahmed Ibrahim Bahgat ElSeddawy. Customer churn prediction model using data mining techniques. In *2017 13th International Computer Engineering Conference (ICENCO)*, pages 262–268. IEEE, 2017.

[7] Xiaojun Wu and Sufang Meng. E-commerce customer churn prediction based on improved smote and adaboost. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–5. IEEE, 2016.

[8] Guo-en Xia, Hui Wang, and Yilin Jiang. Application of customer churn prediction based on weighted selective ensembles. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*, pages 513–519. IEEE, 2016.

[9] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.