Customer Churn Prediction using Weighted Soft-voting Ensemble Learning

V. Umayaparvathi
Assistant Professor
Department of Computer Science
Kamarajar Government Arts and Science College,
Surandai, Tamilnadu, India

Prof. K. Iyakutti
Professor-Emeritus,
Department of Physics and Nanotechnology,
SRM University, Chennai, Tamilnadu, India

Abstract:

Customer churn prediction is the process of identifying the possible churners in advance before they leave the network. It helps the Customer Relationship Management (CRM) department prevent the subscribers who are likely to churn in the future by taking the required retention policies to attract the likely churners and retain them. Machine learning techniques are widely applied to the churn prediction problem, and recent studies compare their performances. In recent years, ensemble learning techniques gained much attention due to their superior performance than traditional algorithms. A voting ensemble involves combining multiple models' predictions to output the final class, e.g., churner or not. This work presents an improved weighted soft-voting ensemble classifier. The weights are assigned to the heterogeneous ensemble members based on their performance during the training phase. The proposed ensemble classifier achieved better accuracy when compared with ten standard machine learning and ensemble models.

Keywords: Customer relationship management, Churn prediction, Machine learning, and Ensemble learning,

1. INTRODUCTION

Customer churn prediction is an increasingly important problem in the highly competitive telecommunication sector [1]. According to the telecom market, the process of subscribers (either prepaid or post-paid) leaving or switching from one service provider to another is called customer churn. If churning continues for any telecom industry, it will lead to significant revenue loss. Moreover, acquiring a new subscriber is more expensive than retaining the existing ones. As numerous factors affect a subscriber to become churn, telecom operators seek alternative ways of using advanced analytical tools to identify the cause in advance.

Customer churn prediction is the process of identifying the possible churners before they leave the service. In general, the churn prediction problem is formulated as a supervised binary classification problem. Customer's historical data is used to train a model about customer behavior (churners and non-churners). Then the learned model is used to identify chances of a customer becoming a churner. The efficiency of the churn prediction system depends highly on the availability of detailed customer attributes representing the customer behavior and the modeling accuracy [2].

Numerous machine learning algorithms have been used to model the customers' past behavior, such as Decision Trees, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Artificial Neural Networks [2]. Though these techniques are widely adopted, they are often inadequate in capturing the complex relationships between the outpour and input variables. These traditional machine learning techniques also performed poorly on large and diverse datasets and lacked robustness in their prediction.

Recently, ensemble learning gained much attention due to its superior performance in various domains. Unlike the traditional machine learning algorithms, ensemble learning algorithms combine the predictions from two or more models, resulting in improved performance and robustness. There are different ensembles, including bagging, boosting, and stacking. In order to combine the predictions from two or more models, a voting mechanism is used. There are two categories of voting-based ensemble models: soft-voting and hard-voting. One of the issues with voting-based ensemble learning is finding the optimal weights for each base model to achieve superior performance and robustness (high variance and low bias) in the final model.

This study presents an improved weighted voting ensemble classifier to improve the performance of the churn prediction system. The proposed model is thoroughly evaluated on two public datasets. The performance of the

churn prediction model is compared with ten recent algorithms using four standard evaluation metrics. Through experimental validation, it is found that the proposed model achieved better accuracy than the other models.

This paper is organized as follows: Literature review is presented in Section 2. The proposed churn prediction methodology is presented in Section 3. In Section 4, dataset details are described. Experimental setup and results are presented in Section 5. Finally, conclusions are made in Section 6.

2. LITERATURE REVIEW

Many machine learning techniques are used to predict customer churn. Decision Trees (DT) are commonly used techniques for churn prediction [4]. Probabilistic algorithms such as Naïve Bayes (NB) classifiers are also used for churn prediction in the telecom industry [5]. In addition, Artificial Neural Network (ANN) models are used for predicting churn in cellular networks [6]. In [7], researchers proposed a Multilayer Perceptron (MLP) neural network and compared its performance with logistic regression classifiers. The results showed that neural networks achieved better accuracy than the statistical models. In [8], researchers have proposed a hybrid approach for churn prediction using KNN and logistic regression.

In a recent study [9], the performance of multi-layer perceptron, Decision Tree, Support Vector Machine (SVM), Naïve Bayes, and Logistic regression were compared. All the models were evaluated using cross-validation and Monte Carlo simulations, and SVM has outperformed other models with an accuracy of 97% and F-measure of 84%. In a similar comparative study, researchers have compared the performance of seven contemporary machine learning techniques on two public datasets [10]. It was concluded that the gradient boosting classifier outperformed the other models.

Recently, ensemble learning techniques have been leveraged to improve the accuracy of the churn prediction models [10]. Although they gained much attention in many fields, such as computer vision, due to their superior performance, the efficiency of ensemble models in churn prediction is not yet been investigated thoroughly. Unlike the traditional machine learning algorithms, ensemble learning algorithms combine the predictions from two or more models. There are different ensembles, including bagging, boosting, and stacking. In order to combine the predictions from two or more models, a voting mechanism is used. One of the challenges with using voting-based ensemble learning is finding the optimal weights for each base model. Therefore, a voting-based ensemble classifier using adaptive weights is presented in this work to improve churn prediction accuracy.

3. METHODOLOGY

Customer churn prediction is modeled as a supervised binary classification problem. The objective is to construct an efficient churn prediction model using various customer attributes available with the telecom service providers. In addition, the churn prediction system should also identify the factors that influence customer churn.

3.1 Ensemble learning

Ensemble learning is a meta-approach that combines the predictions from two or more models. The main idea behind ensemble learning is that combining the predictions from two or more models, each trained on different samples or whole datasets, will improve prediction accuracy than using a single model. Ensemble learning techniques have been applied in various machine learning problems and achieved superior performance than the traditional algorithms. Although there are many ways to combine the predictions, there are three main types available: 1) *bagging, 2) boosting*, and 3) *stacking.*

3.2 Voting ensembles

Voting ensembles are similar to stacking, except a voting mechanism is used to make the final prediction instead of a meta learner. As voting ensembles often involve heterogeneous base models, it is often used to improve the model performance than any single model used in the ensemble.

Hard and Soft voting: A voting mechanism is commonly used to combine the predictions from two or more base models. Two types of voting are broadly used: a) hard voting and b) soft voting.

Hard voting is the simplest case of the ensemble in which the final class label $\hat{y}$ is predicted based on majority voting of each classifier C_i.

$$\hat{y} = mode\{C_1(\mathrm{x}), C_2(\mathrm{x}), \dots, C_n(\mathrm{x})\}$$

Where x is the training dataset and $C_i(\mathrm{x})$ denotes the output from the classifier C_i for $i = 1,2, \dots, n$. As an example, let us consider there are three binary classifiers with the following output:

$C_1(\mathrm{x}) = 1\ (churn),\ C_2(\mathrm{x}) = 0\ (not\ churn),\ and\ C_3(\mathrm{x}) = 1\ (churn)$

With majority voting, the final prediction $\hat{y} = 1$ (churn).

$$\hat{y} = \{1, 0, 1\} = 1(churn)$$

In addition to the simple majority voting, wherein there are no weights assigned to the classifiers, each classifier C_i can be assigned with weights w_i, before calculating the majority voting. This variant is called weighted majority voting.

$$\hat{y} = \arg\max_j \sum_{i=1}^{n} w_i \chi A\ (C_i(\mathrm{x}) = j)$$

Where w_i is the weight assigned to ith classifier and χA denotes the function $[C_i(\mathrm{x}) = j \in A]$ and A = {"churn", "not churn"}.

As an example, if we assign the weights $w = \{0.2, 0.6, 0.2\}$ to classifiers C_1, C_2, and C_3 in the previous example, then the final prediction would be:

$$\hat{y} = \arg\max_j [0.2 \times j_1 + 0.6 \times j_0 + 0.2 \times j_1] = 0\ (not\ churn)$$

3.3 Weighted Soft-voting

In soft-voting, the final output is based on the prediction probability p of the classifiers C_i. The prediction probabilities are summed for each class label, and the final class label $\hat{y}$ is the one with the highest average probability.

$$\hat{y} = \arg\max_j \sum_{i=1}^{n} w_i p_{ij}$$

Where, w_i is the weight assigned to ith classifier and p_{ij} is the prediction probability from ith classifier and $j \in \{0,1\}$ are the class labels. Here, $j = 1$, denotes the churn class.

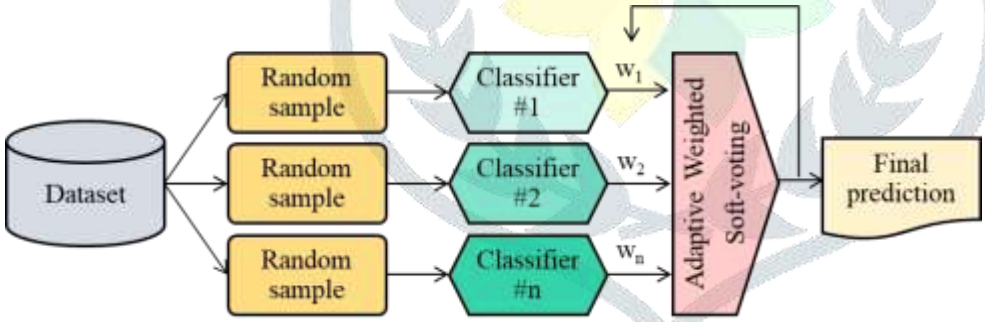

Figure 1: The Weighted Soft-Voting Ensemble learning method.

It is crucial to find optimal weights for the base classifiers. Assigning uniform weight is simple and easy, but this may not improve accuracy, mainly when the underlying classifiers are heterogeneous. In this work, the weights for members of the ensemble classifiers are assigned based on the training error of the individual classifier. Each classifier C_i is assigned with the weight $w_i = \log\left(\frac{1}{\beta_i}\right)$ where $\beta_i = \frac{\varepsilon_i}{1-\varepsilon_i}$ and $\varepsilon_i$is the adaptive weighted based on the training error of the classifier C_i.

4. DATASET

The telecommunication datasets usually contain confidential customer information; therefore, it is challenging to access them outside the company due to the data protection policies. Therefore, this study used Orange Telecom's churn dataset available in the public domain [11]. This dataset has already been used in many studies to evaluate the performance of the various churn prediction models. The Orange dataset contains a total of 21 attributes from 7043 customers. The list of customer attributes includes demographic information such as gender number of dependents, account information such as tenure, type of contract, total charges, and service information, including call data records. Table 1 shows the summary of this dataset.

Dataset/properties	Orange Telecom's churn dataset
Total number of consumers	7,043
Total number of variables	21
Total number of non-churners	5174
Total number of churners	1869
Percentage of non-churners	73.5
Percentage of churners	26.5
Total number of usable features	19

The Orange dataset was already cleaned, and there were no missing values. Hence, no manual data cleaning was done. All customer attributes that could predict churn are selected as features except the customer identifier and churn label. After preprocessing, there were 19 usable features. This dataset contains both numeric and categorical variables. All categorical variables were transformed into numeric using one-hot encoding. After that, all the numerical variables were standardized by applying a min-max scaler.

5. RESULTS

5.1 Performance metrics

Three standard metrics were used to evaluate and compare the performance of the churn prediction models. They are defined as below:

1. ***Prediction accuracy***: Accuracy of the given prediction model is defined as below

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

Here, TP is true positives, TN is true negatives, TP is false positives, and TN is false negatives. All these values are calculated based on the actual and predicted class labels. Accuracy provides the overall performance of the model.

2. ***Precision and Recall***: Precision measures, "What proportion of positive identifications were actually correct (churners)?", whereas the recall "What proportion of actual positives (churners) was identified correctly?".

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

3. ***F1-score***: It is defined as the harmonic mean between precision and recall. It is calculated as below:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

5.2 Experimental setup

In general, input dataset is split into training and test sets using random sampling. However, model validation using a single split may introduce bias because not all combinations of samples were included into training and test datasets. Therefore, a five-fold cross-validation approach is used in this study, which is more robust than using a single split. The given dataset is shuffled randomly and split into five groups. A model is trained on the samples from the remaining four groups and tested on the current group (hold-out set). This process is repeated for all five groups, and the average of their predictions is the final score.

Ensemble members: In this study, seven heterogeneous models are used as ensemble members. They are: (1) Decision Trees, (2) Extra Trees, (3) k-Nearest Neighbors, (4) Naive Bayes, (5) SVM, (6) AdaBoost, and (7) Gradient Boosting. Final models were selected after turning the hyper-parameters using 5-fold cross-validation. The weights for each classifier were calculated based on the procedure described in Section 3.4. After calculating the weights, the prediction probabilities from these classifiers and the weights were given as input to the voting ensemble to get the final prediction.

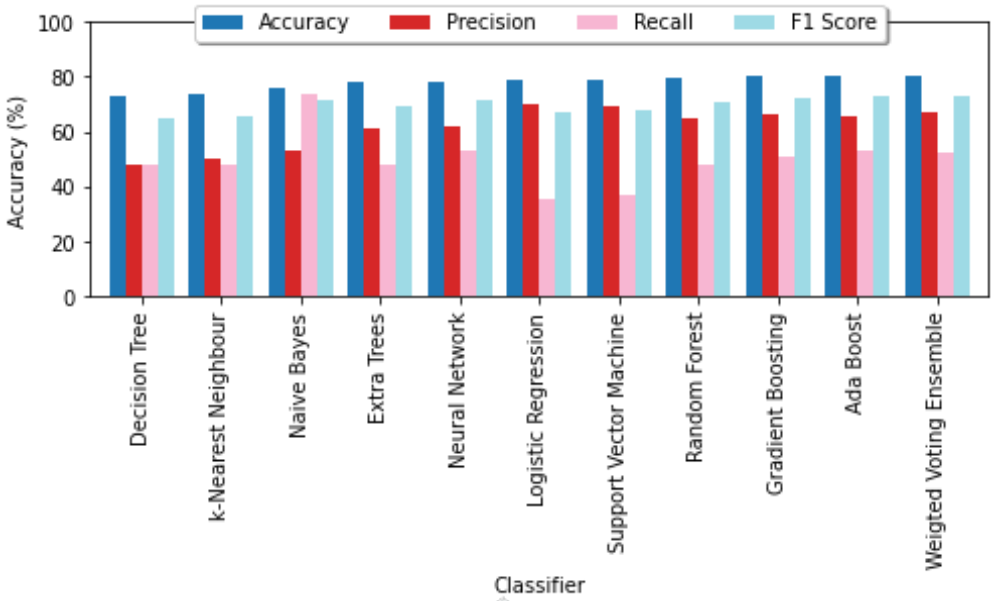

Figure 2: Comparison of accuracy, precision, recall, and F1-score between the voting classifier and ten standard classification models in ascending order (left to right). The voting classifier achieved the highest accuracy of 80.38%.

Classifier	Accuracy	Precision	Recall	F1 Score
Decision Tree	72.68	48.15	47.58	64.97
k-Nearest Neighbour	73.68	50.43	47.62	65.62
Naive Bayes	75.55	52.83	73.36	71.75
Extra Trees	77.81	61.19	47.52	69.16
Neural Network	78.36	61.90	53.29	71.36
Logistic Regression	78.84	69.90	35.64	66.98
Support Vector Machine	78.93	69.21	37.19	67.56
Random Forest	79.67	64.77	48.16	70.98
Gradient Boosting	80.09	66.25	50.90	72.23
Ada Boost	80.31	65.87	53.38	73.00
Weigted Voting Ensemble	80.38	66.75	52.32	72.91

Table 1: Comparison of accuracy, precision, recall, and F1-score between the voting classifier and ten standard classification models in ascending order (top to bottom). The voting classifier achieved the highest accuracy of 80.38%.

The performance of the weighted soft-voting ensemble classifier and ten standard classification models are compared in Figure 2 and Table 1. One can observe that the SVM achieved the highest accuracy of 78.93% among the standard machine learning models. Whereas AdaBoost achieved the highest accuracy of 80.31% among the ensemble learning group, including Random Forest and Gradient Boosting. Overall, the weighted voting ensemble classifier achieved the highest accuracy of 80.38%, better than both standard machine learning and ensemble models.

6. CONCLUSION

Customer churn prediction is an essential yet challenging task in many telecom companies. Predictive models are widely sought to model customer behavior using heterogeneous attributes. These churn prediction models are expected to be highly accurate and robust to identify the possible churners in advance. In this paper, a more accurate churn prediction model is proposed using weighted soft-voting ensemble learning. The model is tested on a large public dataset, and its performance was compared with ten standard models. The proposed ensemble classifier achieved better accuracy and showed promising to be incorporated into the company's Customer Relationship Management (CRM) department to prevent the subscribers from becoming churn.

REFERENCES

[1] Gary Cokins, Ken King, "Managing Customer Profitability and Economic Value in the Telecommunication Indutry", SAS Institute White paper.

[2] Sumathi, Sai, and S. N. Sivanandam. Introduction to data mining and its applications. Vol. 29. Springer, 2006.

[3] Yu-Teng Chang, "Applying Data Mining To Telecom Churn Management", IJRIC , 2009 67 – 77.

[4] Umayaparvathi, V., and K. Iyakutti. "Applications of data mining techniques in telecom churn prediction." International Journal of Computer Applications 42.20 (2012): 5-9.

[5] Kirui, Clement, Li Hong, Wilson Cheruiyot, and Hillary Kirui. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining."IJCSI International Journal of Computer Science Issues 10, no. 2 (2013): 1694-0784

[6] Sharma, A.; Panigrahi, D.P.K. A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. Int. J.Comput. Appl. 2013, 27, 26–31.

[7] Ismail, Mohammad Ridwan, Mohd Khalid Awang, M.Nordin A. Rahman, and Mokhairi Makhtar. "A MultiLayer Perceptron Approach for Customer Churn Prediction." International Journal of Multimedia and Ubiquitous Engineering 10, no. 7 (2015): 213-222.

[8] Zhang, Y.; Qi, J.; Shu, H.; Cao, J. A hybrid KNN-LR classifier and its application in customer churn prediction. In Proceedings of the 2007 IEEE International Conference on Systems, Man and Cybernetics, Montréal, QC, Canada, 7–10 October 2007; pp.3265–3269. [CrossRef]

[9] Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[10] Umayaparvathi, V., and K. Iyakutti. "Attribute selection and Customer Churn Prediction in telecom industry." 2016 international conference on data mining and advanced computing (sapience). IEEE, 2016.

[11] Telecom Churn Dataset - Cleaned Orange Telecom Customer Churn Dataset, url= https://www.kaggle.com/mnassrib/telecom-churn-datasets (last accessed on 20th February 2019).