Selective Ensemble Model for Telecom Churn Prediction

Ahmad Hammoudeh, Malak Fraihat, Mahmoud Almomani
Princess Sumaya University for Technology
Amman, Jordan
ahmad.hammoudeh@ieee.org

Abstract— **Customer based companies are concerned about costumers who decide to stop using their services (churn) because the cost of acquiring new customer is much higher than satisfying an existing customer. With the rapid development of the telecom industry, churn prediction emerges as one of the fundamental tasks for gaining the competitive advantage in the market. This paper introduces Selective Ensemble Model (SEM) as a powerful technique for churn prediction. Among a set of machine learning models, SEM dynamically selects a combination of models to participate in forming the final outcome. Experimental results show that SEM outperforms its constituent models and the averaging ensemble model.**

Keywords—machine learning, ensemble learning, predictive modeling, churn, telecom

I. Introduction

Churn is defined as the turnover of customers of a business with a company or service. Churn rate is the proportion of customers who cut ties with a company or service in a given period of time. On the other hand, gross adds is the number of new subscribers. Companies' growth occurs when gross adds rate exceeds churn rate, which in turn yields a positive net adds. Continuous positive net adds ensures that a company is expanding its clientele.

While churn is a helpful metric for detecting weaknesses of a company, it can lead to financial burden in the long run [1] as the cost of acquiring new customer is much higher than pleasing an unsatisfied customer. Churn management process is a branch of customer relationship management. One way to ensure effective customer relationship management is through carefully monitoring churn.

Customers are one of the most important assets in telecom industry. With the rapid grow of the telecom market, customers can switch easily to another operator. Predicting customers who are most likely to churn is fundamental for telecommunication companies. As a result, churn prediction is a vital business metric as well as one of the vastly important machine learning applications in telecommunication industry.

This paper introduces Selective Ensemble Modeling (SEM) to predict customer churn precisely. SEM achieved high performance as shown in the experimental results.

The rest of this paper is structured as follows: Section 2 provides an overview of ensemble modeling. Section 3 explains the concept of SEM. Section 4 shows the experimental setup which includes data preparation, models implementation and results evaluation. Finally, conclusions are drawn in Section 5.

II. Ensemble model

The idea of ensemble learning is to utilize a combination of trained models to achieve better performance than any constituent model alone. An ensemble method that takes the average of N models can reduce the error ideally by a factor of N if the errors are uncorrelated. In practice the errors between the models are highly correlated, and therefore the expected enhancement is relatively smaller. However, the performance of the averaging ensemble model remains better than any of the constituent models as can be shown using Cauchy–Schwarz's inequality [2]. Ensemble modeling reduces single model limitations such as bias or high variability.

For example, an averaging ensemble model of 3 learners, 2 learners predict that a costumer will churn and 1 predicts that the costumer will not churn. The averaging ensemble model takes the vote of the majority. In this case, the final prediction will be yes this is a potential churn. Ensemble models have been used to win many machine learning competitions such as Netflix [3], KDD [4], and Kaggle [5]. For more about ensemble modeling, the reader is referred to [6].

III. Selective Ensemble Model (SEM)

Among a set of trained machine learning models, SEM only involves the models predicted to be correct and filters out the models predicted to provide false outcome. The element that dynamically selects which models to involve and which to eliminate is called the selector. The selector is a machine learning model that learns the correctness behavior of each model. Only the models nominated by the selector participate in predicting the final outcome.

Fig. 1 illustrates an SEM that selects among four trained models (A, B, C, D). The models are trained to predict a target value y. The symbols y1, y2, y3, y4 represent the predicted target values of models A, B, C, D respectively.

SEM comes in two phases: the selection phase and the ensembling phase.

The selection mechanism is demonstrated through dynamic switches controlled by the selector. s1, s2, s3, s4 are the switches associated with models A, B, C, D, respectively. When a switch is open, its associated model is filtered out and it will not be taken into account when voting takes place. On the other hand, if a switch is closed, then its model will pass to the next phase where an ensemble method is applied to the selected models.

978-1-5386-7942-5/19/$31.00 ©2019 IEEE

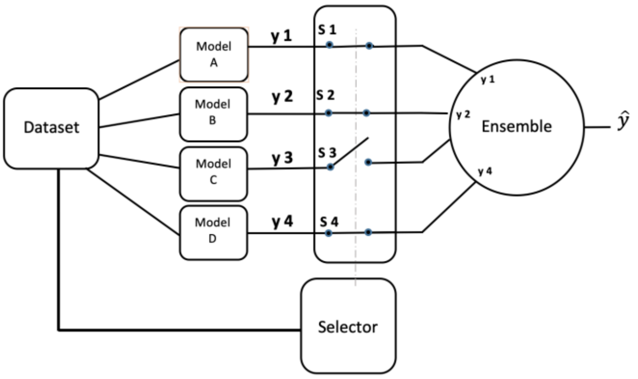

Fig. 1. SEM archetiucture.

As shown in Fig. 1, for a certain input s1, s2, and s4 are closed and s3 is open. Which indicates that the selector got rid of model C, and nominated models A, B, and D to the next phase. Hence, only y1, y2 and y4 contribute to predicting the final outcome. As shown in Table I, y1, y2, y3, y4 are assumed to be: [yes, no, no, yes], respectively. Eliminating y3 results in [yes, no, _, yes]. An averaging ensemble model finds yes 2 times and no once. therefore, the final prediction will be the vote of the majority which is 'yes'.

TABLE I. SEM EXAMPLE

Model	Model Prediction	Selector decision	Votes for 'yes'	Votes for 'no'
A	Yes	OK	1	0
B	No	OK	0	1
C	No	×	×	×
D	Yes	OK	1	0
Total votes for each class			**2**	**1**

The procedure to train the selector is by learning the state of the switch from comparing the prediction of each model against the correct outcome. For the cases where a model matches the correct answer, the selector learns to keep the associated switch closed and when a model's prediction is false, the selector learns to open the associated switch. After training, the selector will be able to predict dynamically which models are more likely to be correct and which will be false. For example, in Table I, if the correct answer is 'yes'. The selector should learn to open both s2 and s3 in this case.

IV. Experimental Setup and Results

A. Dataset

Churn dataset [7] consists of 21 attributes for 3,333 customers with a label indicating churn status of each customer. The data is divided into two parts: 30% for testing and 70% for training and development. The test set is held out to evaluate the performance of the model. Hence, machine learning model do not see the test set before the final evaluation.

B. Data preparation

Data preparation steps include:

1) Feature Selection

Irrelevant attributes were completely dropped, namely: State, Phone number, and Area code.

2) Feature transorfmation

Categorical data of 2 distinct categories were transformed into binary encoding; where one of the two categories is replaced by 1 and the other by 0.

Numerical data were standardized (Z-score normalization) by subtracting the arithmetic mean of that attribute and dividing over the standard deviation.

3) Imbalance data

The dataset is imbalanced; less than 17% of the records are labeled as churn. The problem of imbalanced data is one of the obstacles for many machine learning algorithms, it arises when the data are dominated by a majority class and a minority class is rarely detected. As a result, the classifier performance on the minority may be insufficient when compared to the majority. For example, a dumb classifier that always predicts the majority class can achieve high accuracy. In this work, synthetic minority over-sampling technique (SMOTE) [8] is implemented to overcome the problem of imbalanced classes. For the impact of over-sampling in enhancing the performance of NNs, the reader is referred to [9]. Implementing SMOTE has resulted in increasing the records of the minority class from 333 to 2000 as shown in Table 2.

TABLE II. NUMBER OF RECORDS BEFORE AND AFTER SMOTE

Churn Status	Before Balancing	After Balancing
Yes	333	2000
No	2000	2000

C. Evaluation metrics

The performance of machine learning models is evaluated based on data that never seen before (test data). a model predicts among two classes one of them is considered "positive" or "churn/yes" and the other is "negative" or "not churn/no". Some of the widely used class output evaluation measures [10] for binary predictive models are:

1) Aaccuracy

Accuracy is the percentage of correctly predicted instances to the total instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where, True Positive (TP): total instances that are actually positive and predicted correctly as positive. True Negative (TN): total instances that are actually negative and predicted correctly as negative. False Positive (FP): total instances that are actually negative and predicted mistakenly as Positive. False Negative (FN): total instances that are actually positive and predicted mistakenly as negative.

2) Precision

Precision measures how many of the record classified as positive are actually positive.

$$P.recision = \frac{TP}{TP+FP} \quad (2)$$

3) Recall

Recall measures how many of the total positive records were classified correctly as positive.

$$R.ecall = \frac{TP}{TP+FN} \quad (3)$$

D. Machine learning models

Four machine learning models were trained to predict the churn: Support Vector Machine (SVM), traditional Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Random Forests (RF). The four models were used in both the averaging ensemble model (AEM) and the SEM. The selector is an ANN. Fig. 2. illustrate the SEM developed for churn prediction.

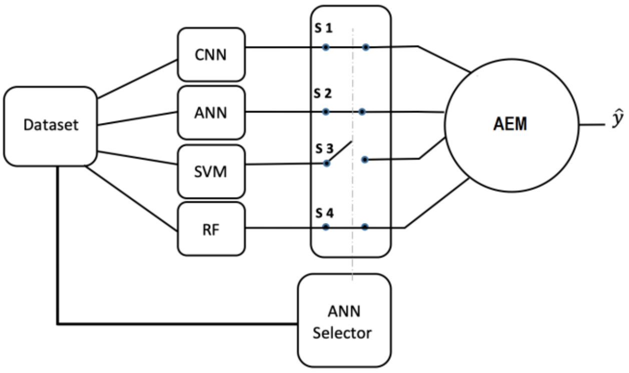

Fig. 2. SEM developed for churn prediction

Table 3 summarizes the results of all the models. SEM outperforms all models as well as the averaging ensemble model in terms of different evaluation metrics: accuracy, precision and recall.

TABLE III. EVALUATIO OF MACHINE LEARNING MODELS DEVELOPED FOR CHURN PREDICTION

Method	Accuracy	Precision	Recall
CNN	89.9	77.6	76.7
ANN	85.1	70.1	78.7
SVM	84.9	83.6	73.6
RF	87.4	73.2	80.6
AEM	90.9	79.5	81.3
SEM	**93.9**	**83.8**	**84.6**

V. CONCLUSION

This paper proposed a selective ensemble model for predicting the customers who are more likely to churn in the mobile telecommunication industry. Compared against the averaging ensemble model, SEM achieved higher accuracy, precision and recall. The SEM achieves better performance than its constituent models and allows taking marketing decisions based on accurate predictions.

REFERENCES

[1] R. Jadhav and T. Usharani, “Churn Prediction in Telecommunication Using Data Mining Technology,” *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 2, 2011.

[2] M. Sewell, "Ensemble learning," RN, vol. 11, no. 02, 2008.

[3] E. Chen, “Winning the Netflix Prize: A Summary,” Edwin Chens Blog Atom. [Online]. Available: http://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/. [Accessed: 17-Dec-2018].

[4] A. Niculescu-Mizil, C. Perlich, G. Swirszcz, V. Sindhwani, Y. Liu, P. Melville, W. Shang, “Winning the KDD cup orange challenge with ensemble selection,”. In Proc. of the 2009 Int. Conf. on KDD-Cup 2009, vol. 7, pp. 23-34, 2009.

[5] M. Asseck, “Bird song classification in field recordings: winning solution for NIPS4B 2013 competition,” In Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada , pp. 176-181, 2013.

[6] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. Boca Raton, FL: Taylor & Francis, 2012.

[7] "Churn in Telecom's dataset", Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/becksddf/churn-in-telecoms-dataset. [Accessed: 02- Aug- 2018].

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jan. 2002.

[9] P. Hensman, and D. Masko, “The Impact of Imbalanced Training Data for Convolutional Neural Networks.” KTH Royal Institute of Technology, 2015.

[10] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.