Customer Churn Prediction in Telecommunication Industry Using Machine Learning Approach

Sunjare Zulfiker
Department of ECE
North South University
Dhaka, Bangladesh
sunjare.zulfiker@northsouth.edu

S.M Ashraful Hasan
Department of ECE
North South University
Dhaka, Bangladesh
sm.ashraful@northsouth.edu

Dip Roy
Department of ECE
North South University
Dhaka, Bangladesh
dip.roy@northsouth.edu

***Abstract*—Nowadays, there is high competition in the telecommunication industry. So, this research is for predicting customer churn by analyzing some causes and factors using machine learning. In this study, we have used six different classifiers for predicting the churn. Moreover, we have used Synthetic Minority Oversampling Technique (SMOTE) to reduce the imbalance and we have also standardized our data to get more accuracy in predicting customer churn.**

***Index Terms*—churn, classifiers, machine learning, SMOTE, imbalance, standardized**

I. Introduction

In this study, we explored telecommunication industry's one of the fastest growing service the mobile phone service industry. The telecom operators keep retaining their existing customers as a prime concern to make a good profit. The competition between the mobile operators now a days, is more than ever. So, in order to predict customers subscription, the company applies data analysis for the churn prediction. They analyze frequency of use, seconds of use, tariff plan, charge amount etc. These data analysis help them to understand which customer is more important. The companies often come up with new ideas to the telecom market to attract customers and win them from intense competition. This type of analysis is called churn prediction and customer retention strategies. Companies creates multiple churn models from an economical perspective as profit is the main concern in business environments.

Churn prediction requires different statical and data analytic methods. Churning helps not only focusing on bringing new customers but also reducing churn on the areas where company losing customers. Companies who offers poor pricing range definitely need a different developed strategy than one that is churning customer due to broken support system [1]. We will work on this paper's most important contribution which is the churn prediction by implementing well known algorithms. We will be able to predict which algorithm has the most accuracy rate after the implementation of those five algorithms.

II. Literature Review

Customers are the main asset of any industry or business. Currently, the demand for telecommunication is increasing very fast. As a result, there is tight competition among the telecommunication industries. Therefore, it is very important for the industry or business to predict the customer churn and keep their heads high in this competitive market.

Keramati et al. [2] have used four well-known classifiers for predicting customer churn for the telecommunication industry. They have collected this data from an Iranian call center. In this data mining process first, they have collected the raw data and prepared the data for machine learning by doing data cleaning and data preprocessing. After that, they used Decision tree, Artificial Neural Network, K-Nearest Neighbors, and Support Vector Machine classifier for predicting churn. They have also proposed a hybrid method using all the classifiers that have been used in this research. This hybrid method is very much accurate and it can return the required result with a precision and recall of 95% and this method outperforms all the four base classifiers.

Huang et al. [3] used some new feature set and used seven classifiers to predict customer churn and shows that this is a quite efficient technique for predicting the churn in this field. This dataset is collected from the telecommunication company of Ireland. They have prepared their raw data by doing data cleaning and data normalization. Their data set was highly imbalanced. So, they overcome this problem by applying the sampling method. Then they applied Logistic Regression, Decision Trees, Naive Bayes, Linear Classifiers, Artificial Neural Networks, Support Vector Machines, and the last one is the Evolutionary data mining algorithm. In this research, they did cross validation of 10 folds. This modelling technique and use of new feature set are perfect for customer churn. Therefore, they got a decent performance from this technique. Although, they have stated that there are some limitations to this project. They should use some new features with their dataset. Moreover, they just used the sampling method for removing the class imbalance. They should use some more methods for solving this imbalance problem of their dataset.

Zhang et al. [4] predicts the customer churn for a telecommunication company and used the effects of interpersonal influence in it. They identified the customer churn in two sections first one is network attributes. Where they applied the interpersonal influence and the connection of a customer with his

neighbors and others. The second one is traditional attributes that are often used in predicting customer churn. They have collected this dataset from a prominent mobile service provider company. Here, they used Logistic Regression, Decision Tree, and Neural Network and combined those traditional attributes with the network attributes and significantly improves their accuracy in predicting customer churn. After combining the interpersonal influence with traditional attributes, the hit rate increases from 19.57% - 24.58%. They confirmed that this combined method is very much effective than the traditional classification method and this method outperforms all other models.

Samira et al. [5] predict the customer churn for grocery stores using machine learning techniques. Here, they used some effective attributes for predicting churn and got a decent accuracy in this field. They collected this data from an Iranian grocery shop. First of all, they prepared their dataset for machine learning using some necessary methods and they are data cleaning, integration, and data transformation. Then they selected important attributes from the dataset and starts training and testing using some well-known models. They used Support Vector Machine, Artificial Neural Network, Decision Tree, and finally used Ensemble methods for higher accuracy in predicting the churn. In this study, they have got an accuracy of 97.92% from their proposed model where Artificial Neural Network gave the highest accuracy. However, I think this study is not complete. Because this study and their method are only suitable for grocery shops also, they can not use this method for other industries. Therefore, they should extract some new attributes or variables so that they could apply their study or method to other different sectors.

Caigny et al. [6] have used a novel hybrid algorithm and predicted the customer churn in the telecommunication industry. Here, he used the logit leaf model (LLM) for predicting churn and it performed well. LLM model is the combination of logistic regression and decision tree. It contains two steps in the first step it creates decision tree and classifies homogenous customer segmentation. In the next step, they apply logistic regression in each segment. But first, they prepare their data by doing preprocessing, and then they did the variable selection and parameter optimization for better accuracy. They confirmed that LLM outperforms all the traditional models and it is more accurate. Hit rate and lift also increased after applying the LLM method. In this research, they stated that the dataset they have used in this study is quite small. They should have used large data for better validation.

Kirui et al. [7] used some distinctive unified modeling language to predict customer churn in Mobile Telephony Industry using Probabilistic Classifiers in Data Mining. They have obtained the data from a European telecommunications company and collected three months of data from August to October 1997. In this process, they used data mining as a collectively efficient way. Their approach includes data sampling, data preprocessing, model construction, and model evaluation phases. They used two probabilistic data mining algorithms Naïve Bayes and Bayesian Network to construct their model for this research and for their result they used a famous algorithm call Decision Tree for classification and prediction. This algorithm is widely used and almost accurate. Decision Tree classifier gave an accuracy of 91% and this method is better than the other two probabilistic data mining algorithms.

Jahromi et al. [8] applied some unique classifiers to predict B2B customer churn, retention, and profitability their four distinguish classifiers are quite decent to predict customer churn, retention, and profitability. The data used for this study was adapted from the transactional records of 11,021 business customers of an Australian online retailer. In this study, they prioritized predicting inactivity instead of predict churn as a principal phenomenon, they generated the idea of churn as being inactive in the second half of the year and being active in the first half of the year. They have used four types of classifiers in this study and they are Simple decision tree, Decision tree with cost-sensitive learning, Boosting as an ensemble learner method, and Logistic regression. To ensure better predictive results they used two efficient criteria like characteristic curve and cumulative lift curve. They also used random classifier to compare with those four classifiers to see churn prediction without any model. So, if the company aims to target 40% of its customer base, then the boosting model will outperform the other three models to capture real churner while using random classifier (no model) will capture only 40% of the real churner.

Zhao et al. [9] used a popular SVM (Support vector machine) algorithm with different types of Kernel function in SVM to predict Churn Prediction Using Support Vector Machine. They have collected their data from the carrier is stored in an Oracle database. They have used some effective strategies to find the churn analyze subscriber dissatisfaction. Their study includes Demographics, Usage level. Call detail records, Quality of Service (QOS), and Features-Marketing analysis method. In this study, they have used different types of Kernel functions like Linear, Polynomial, Gaussian Kernel function to improve their prediction in their SVM model. They used well-known classifiers like ANN, Decision Tree, Naïve Bays and compared them with Gaussian Kernel function. According to their study, the Gaussian Kernel function can detect more churners than the other three algorithms and it has the highest accuracy rate 87.15%.

Dahiya et al. [10] used some useful ideas in their data mining model and they generated their ideas with the two most popular classifiers for their study to construct the churn prediction model. They acquired this data from KDD Cup 2009 and it was used to analyze large databases from French telecom company Orange. They have used some simple but fruitful data mining processes like Data Preparation, Data Preprocessing, and Data Extraction to arrange their study in a more precise way. They have used Decision Tree and Logistic Regression classifier to predict their churn analysis. Both of these techniques are quite useful and popular in terms of predicting accurate results but in this study, decision tree outperforms Logistic Regression. For a large dataset, decision

tree is showing more accurate prediction churn analysis than Logistic Regression. Apart from this, they should include some other classifiers and data mining model methods to compare the prediction result and enrich their churn prediction analysis in a more appropriate way.

Xie et al. [11] used lift curve and top-decile methods to measure out the accuracy of the prediction model and they have used four types of classifiers to build their churn prediction model. They collected this data set from a major Chinese bank who provided them records of more than 20,000 customers. They have used three popular classifiers for this study and they are: Artificial Neural Network (ANN), Decision Tree, and CWC-SVM, apart from that they have used a special technique called Improved Balanced Random Forests (IBRF). They applied it to the real-world database for this study. They used two measurements method called lift curve and top-decile method for evaluating and comparing model performance and correct accuracy. Their experimental result for this is showing that their novel method called IBRF is showing more accuracy than ANN, Decision tree, and CWC-SVM. It is showing an accuracy rate of about 93.2% with a top-decile lift of 7.1 which is the highest among all of them. According to their research, IBRF has great potential compared to other conventional approaches due to its manageability and quick training and handling speeds.

III. Data Set

This data is collected from an Iranian telecom company's database [2]. In this database, there are a total of 14 columns. 13 of them are features and the last one is the target or class label. The dataset contains the information of 3150 subscribers. The attributes of this dataset are

- Call Failure: This represents the number of Call Failures in one year.
- Complains: It is a binary attribute where 0 means no complaints and 1 means there is a complaint last year.
- Subscription Length: Which is the total number of subscriptions in months.
- Charge Amount: It is the amount of charge during the subscription. It is ranked from zero to nine. Where zero is the lowest amount and 9 is the highest.
- Seconds of Use: It represents the total call in seconds.
- Frequency of Use: It is the total number of calls in last year.
- Frequency of SMS: Number of SMS (short message service).
- Distinct Called Numbers: It represents the total distinct number of call in a year.
- Age Group: All the subscribers are grouped by their age.

Group	Age
1	less than 15
2	15-30
3	30-45
4	45-60
5	60-75

- Tariff Plan: It is a binary attribute. Here we have two types of tariff plan. 1 - Pay as you go 2 - Contractual
- Status: It is a Binary attribute. 1 – active, 2 – non active
- Customer Value: It is the calculated value of each customer. Basically, it is the ratio between perceived benefits and cost.

IV. Methodology

In order to predict the customer churn here, we have split our data into train and test. Then we have used feature scaling to standardize the features. The dataset that we have used is highly imbalanced. Here, we can see that in our training dataset there are 1879 data which are of class 0 (no churn) and 326 data which are of class 1 (churn). We have handled this class imbalance problem using Synthetic Minority Oversampling Technique(SMOTE). We have used six different classifiers. They are Logistic Regression, Random Forest Classifier, Decision Tree, K-nearest neighbor (KNN), AdaBoost Classifier and Voting Classifier. At last, we have used Weighted Voting Classifier by combining the outputs from different base classifiers.

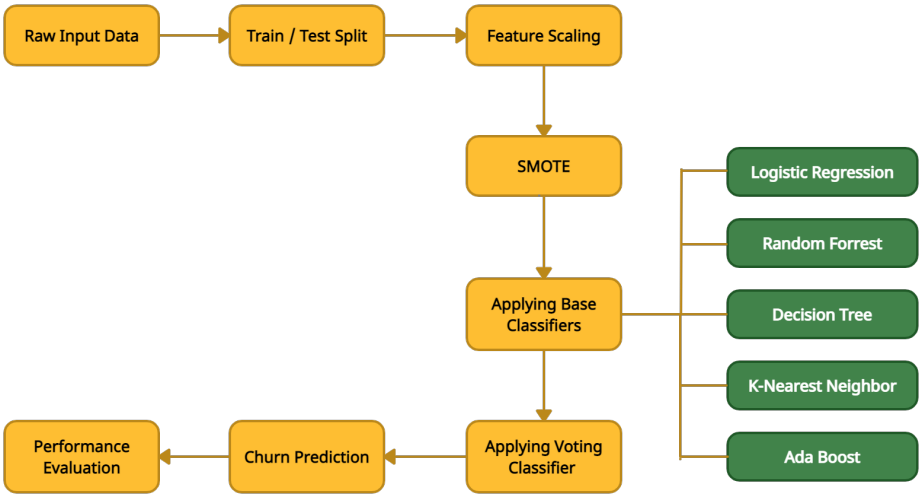

Fig. 1. Step by step method for churn prediction

V. RESULTS AND ANALYSIS

Here in this test data, we have the records of 945 customers. Between them 776 customers are non-churners and 169 customers are churners. Here, table 1 shows the confusion matrix of the different classifiers we have used in this study.

Table 1: Confusion Matrix of the classifiers

Classifiers	*TN*	*FP*	*FN*	*TP*
Logistic Regression	659	117	25	144
Random Forest	753	23	23	146
Decision Tree	744	32	38	131
KNN	726	50	21	148
AdaBoost	709	67	21	148
Voting Classifier	754	22	23	146

For analyzing and evaluating the performance of the classifiers, we have calculated the accuracy, precision and the F1-score of the models.
Here,

- True Negative(TN) = When a customer is predicted as non-churner by the models and the known outcome is also non-churner.
- False Negative(FP) = When a customer is predicted as churners but actually the known outcome is non-churner
- False Negative(FN) = When a customer is a churner but our models predicts as non-churner
- True Positive(TP) = When a customer is churner and our models also predicts as churner

Here, In this table 2 we have shown the accuracy, precision, F1-score and AUC for all the classifiers

Table 2: Performance of the classifiers

Classifiers	**Accuracy (%)**	**Precision (%)**	**F1-score (%)**	**AUC**
Logistic Regression	84.9%	55.1%	66.0%	0.851
Random Forest	95.1%	86.3%	86.3%	0.917
Decision Tree	92.5%	80.3%	78.9%	0.867
KNN	92.4%	74.7%	80.6%	0.906
AdaBoost	90.6%	68.8%	77.0%	0.895
Voting Classifier	95.2%	86.9%	86.6%	0.918

From this table 2, we can see the performance of models. First, of all, we have applied Logistic regression. It did not perform well. Though it has an accuracy of 84.9% but its precision is 55.1%, F1-score is 66.0%, and AUC 0.851. The reason behind this low precision and F1-score is the number of False Positives (FP) is pretty high which is 117. The next classifier is Random Forest. We got 95.1% of accuracy in this classifier. Moreover, its precision and F1-score are 86.3% with an AUC of 0.917. Random forest outperforms all the base classifiers. Then we applied the Decision Tree and got an accuracy of 92.5%. Decision Tree also performed

well. Its precision and accuracy are respectively 80.3% and 78.9%. Its AUC score is 0.917. We got an accuracy of 92.4% using K-Nearest Neighbor (KNN) classifier. Its precision and F1-score are 74.7% and 80.6% with an AUC score of 0.906. The last base classifier is AdaBoost. It has an accuracy of 90.6% but its precision is low which is 68.8%, F1-score 77.0%, and AUC 0.895. Finally, we used a weighted voting classifier to accelerate the overall performance of this study. Voting classifiers outperform all the base classifiers and its accuracy is 95.2%. Moreover, its precision and F1-score are also high which are 86.9% and 86.6%. Voting classifiers also give us the lowest number of False Positive (FP) and False Negative (FN). Which are only 22 and 23 where the number of True Positives (TP) is 146 and True Negative (TN) is 754.

Here, we can see the ROC curve for different classifiers

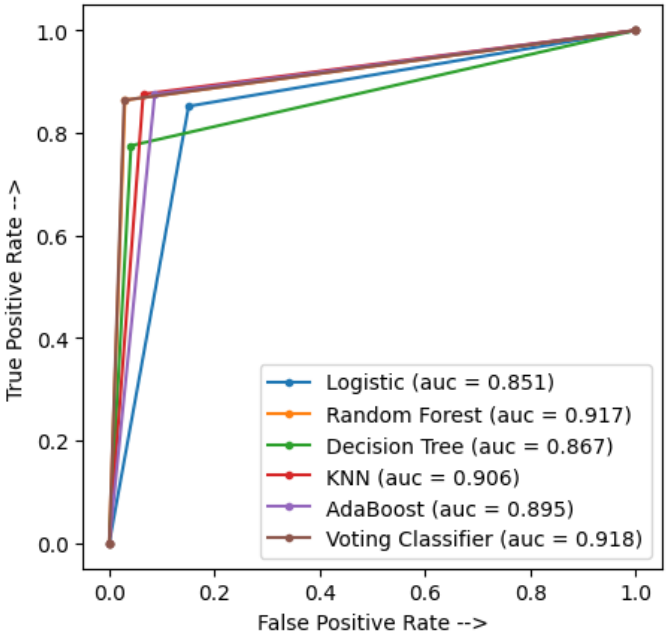

Fig. 2. ROC Curve of the classifiers

In this ROC curve the x-direction represents the False Positive Rate and y-direction represents the True Positive Rate. We know that when a ROC curve is closer to the top-left corner of the graph then that specific classifier is giving better performance. Therefore, in this study Voting Classifier is in the most top-left corner of this graph. So, it is giving the best performance among all the classifiers.

VI. CONCLUSION

For phase 2/3.

REFERENCES

[1] PATRICK CAMPBELL, "Customer Churn Analysis: One of SaaS's Most Important Processes," DEC 8 2020. Available: https://www.profitwell.com/customer-churn/analysis

[2] A. Keramati, R. JafariMarandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," Applied Soft Computing, vol. 24, pp. 994–1012, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494614004062

[3] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications, "Expert Systems with Applications ,vol. 39, no. 1, pp. 1414–1425,2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411011353

[4] X. Zhang, J. Zhu, S. Xu, and Y. Wan, "Predicting customer churn through interpersonal influence," Knowledge-Based Systems, vol. 28, pp. 97–104, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705111002693

[5] S. Khodabandehlou and M. Z. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, 2017.

[6] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," European Journal of Operational Research, vol. 269, no. 2, pp. 760–772,2018.

[7] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 2 Part 1, p. 165,2013

[8] A. T. Jahromi, S. Stakhovych, and M. Ewing, "Managing b2b customer churn, retention and profitability," Industrial Marketing Management, vol. 43, no. 7, pp.1258–1268, 2014.

[9] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Customer churn prediction using improved one-class support vector machine," in International Conference on Advanced Data Mining and Applications. Springer, 2005, pp. 300–306.

[10] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," in 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions). IEEE, 2015, pp. 1–6.

[11] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests" Expert Systems with Applications, vol. 36, no. 3, pp.5445–5449, 2009.