Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

An XGBoost Based Ensemble Model for Customer Churn Prediction in Telecommunications Industry

Sagar Maan Shrestha [a], Aman Shakya [b]

[a, b] *Department of Electronics & Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
✉ [a] 076msdsa014.sagar@pcampus.edu.np, [b] aman.shakya@ioe.edu.np

Abstract

Telecommunications Industry are one of the fastest evolving business sectors. They demand huge financial investment even at the onset of business, unlike others. And the repay of their investment is driven by the number of customers they have garnished over time. With the increasing global and national competition this has become a major challenge for all the Telecommunication companies to retain their existing customers. Therefore, the telecommunication operators have major concern over identifying customers who are at risk of churning all the time. An analysis of call detail records inside the telecom will provide an insight on how customer behaviors are affected by the available services and provide an analysis part on whether they will be potential churners. In this research paper, we propose an ensemble-based Machine Learning model with the help of Stacking Technique on the Logistic Regression and Random Forest algorithms as base classifiers and XGBoost as final classifier for churn prediction using call detail records of a fictional telecommunication company that contains 7043 records of customers. The performance metrics like accuracy and F1-Score thus, obtained by this proposed model on this publicly available dataset are 80.88% and 62.69% respectively.

Keywords

Machine Learning, XGBoost, Ensemble, Customer Churn, Stacking Classifier

1. Introduction

Customer churn problem has attracted lots of researchers from different backgrounds. But their target always remains same i.e., to predict the customers who are going to churn and prevent them from doing so. An effective churn prediction model not only strengthens the financial position of company but also facilitates the good relationship between customers and company by helping the decision makers to address the issues related to customers' demands thus, increasing loyalty of customers toward the company. There are various researches done using both qualitative analysis like the research done by V. Umayaparvathi et al.[1] using survey of different literature on customer churn prediction and quantitative methods like the researches conducted by J. Pamina et al.[2] and Shuli Wu et al.[3] using machine learning algorithms or research done by Piotr Sulikowski et al. [4] using collinearity testing but due to the dependencies of algorithms on the datasets, there is no best algorithm that works well on all datasets. Churn is a binary classification problem and the datasets will be mostly unbalanced type.

In this research, the XGBoost classifier has been chosen for final classifier in the stacking model as it can address the class imbalance problem inherently by using the function scale_pos_weight where the user can define appropriate weight to the positive or negative levels and can make use of parallel processing for relatively faster computation.

2. Literature Review

There has been done a significant amount of research for churn prediction in telecommunications industry. Most of the researches are focused on use of standalone machine learning algorithms while some researches have presented comparative analysis of both standalone and hybrid models.

A comparative study of ten different machine learning algorithms was done in the research work of Sahar F. Sabbeh (2018)[5] on the dataset of 3333 records and obtained an accuracy of 96.39% by using Random Forest. In the research of Fahd Idrissi Khamlichi et al.

(2019)[6], the use of different machine learning algorithms on the dataset consisting 5000 number of customers was done and concluded XGBoost performed well with an accuracy of 95% among the suggested standalone models. They have also proposed hybrid model by splitting dataset based on similar attributes into different clusters on span basis and the obtained clusters were trained using Support Vector Machine, Logistic Regression and Decision Tree. Likewise,a comparative study of different standalone models with a hybrid model for customer churn prediction was conducted by Qi Tang et al. (2020)[7]. This research was done on two datasets; one from music industry and another from telecom industry. The telecom dataset used has 51047 samples and they have obtained an maximum accuracy of 72.07% using hybrid model i.e combination of XGBoost and MLP.

Hemlata Jain et al. (2020)[8] has proposed two standalone models using Logistic Regression and Logit Boost on the dataset containing 3333 samples and obtained an accuracy of 85.2385% using LR. In the research of Shuli Wu et al. (2021)[3], three different datasets of telecom companies were taken and they performed a comparative study of different machine learning algorithms on each datasets. They have obtained an accuracy of 80.19% on dataset containing 7032 samples using LR, an accuracy of 95.34% on dataset containing 4031 samples using Random Forest and accuracy of 71.05% on dataset containing 51047 samples using LR. In the research done by J.Pamina et al. (2019)[2] on the publicly available dataset containing 7043 instances, they have obtained an accuracy of 79.8% using XGBoost algorithm.

3. Methodology

The methodological framework for this research is shown in Figure 1. The detail explanation of this research process is as follows.

3.1 Dataset Acquisition and Description

The dataset for this research has been obtained publicly from data repository website Kaggle [9]. This dataset contains 7043 customers' call detail records of a fictional telecommunication company in California in Quarter three. This dataset has 21 attributes as shown in table 1.

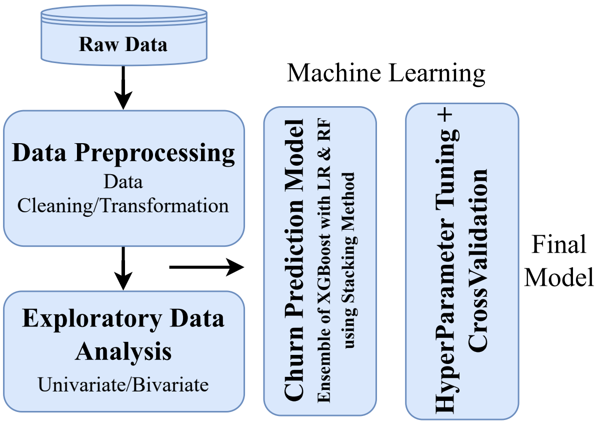

Raw Data
Machine Learning
Data Preprocessing
Data Cleaning/Transformation
Exploratory Data Analysis
Univariate/Bivariate
Churn Prediction Model
Ensemble of XGBoost with LR & RF using Stacking Method
HyperParameter Tuning + CrossValidation
Final Model

Figure 1: Proposed Model

Table 1: Dataset Features Description

S.N	Featrures	Data Format	Description
0	Customer ID	string	ID of the Customer
1	gender	Male/Female	Male/Female
2	Senior Citizen	(0 / 1)	Senior Citizen or not
3	Partner	(Yes / No)	Has partner or not
4	Dependents	(Yes / No)	Has dependents or not
5	tenure	numerical	Stayed number of months with the company
6	Phone Service	(Yes / No)	Subscription of phone service
7	Multiple Lines	(Yes / No / No phone service)	Subscription of multiple lines
8	Internet Service	(DSL / Fiber optic / No)	Internet Service Technology
9	Online Security	(Yes / No / No internet service)	Subscription of Online security
10	Online Backup	(Yes / No / No internet service)	Subscription of Online backup
11	Device Protection	(Yes / No / No internet service)	Subscription of Device protection
12	Tech Support	(Yes / No / No internet service)	Subscription of Tech Support
13	Streaming TV	(Yes / No / No internet service)	Subscription of Streaming TV
14	Streaming Movies	(Yes / No / No internet service)	Subscription of Streaming movies
15	Contract	(Month-to-Month / One Year / Two Year)	Conract term
16	Paperless Billing	(Yes/No)	Billing method
17	Payment Method	(Bank Transfer / Credit Card / Electronic Check / Mailed Check)	Payment Method
18	Monthly Charges	numerical	Monthly charge to customer
19	Total Charges	numerical	Total amount charged to customer
20	Churn	(Yes/No)	if customer is churned or not

3.2 Data Preprocessing and EDA

After the dataset is acquired, the raw data are cleaned and transformed to eliminate the various inconsistencies and noise. In this research we have made use of Jupyter notebook, Pandas, Numpy, Matplotlib, Seaborn and Sklearn libraries. The steps are as follows:

a) Load the csv/excel data into dataframe.

b) Check for shape and general info.

c) Check for inconsistency of the data types and feature names.
It is found that the raw data has inconsistency in TotalCharges columns, the entries are in number but obtained as object type. This has been changed to numeric datatype. In SeniorCitizen column, number has been entered which is reverted back to categorical type.

d) Check for null and duplicated values.
Null and duplicated values can cause over fitting to our model. It is found 11 rows have missing values out of 7043, which is quite insignificant. The rows containing missing values have been dropped. The final record obtained is of 7032 entries.

e) Removing unnecessary columns
CustomerID does not contribute in our ML model performance. Hence dropped.

f) Converting numerical variables to a common data type.
All the numerical parameters are changed to float type.

g) Check if some numerical columns can be assigned to group.
Tenure Column is of Ordinal nature, so it is binned with width of 12 months.

h) Encoding and Mapping of Categorical attributes.
Most of the categorical features are binary labelled which can be simply mapped into binary label 0 and 1. Categorical features more than two labels are OneHotEncoded using OneHotEncoder from Sklearn. Some Features like Contract has been Ordinally encoded.

i) Finding the outliers and distribution of data and choosing normalization technique.
Histogram plot as shown in Figure 2 and Box plot as shown in Figure 3 of the numerical features show that most of the numerical variables are not normally distributed and have no outliers. The MonthlyCharges Column is negatively skewed and Total Charges is positively skewed. Thus, MinMaxScaler from Sklearn is chosen for scaling/normalization purpose. Minmaxscaler scales the data in the range between [-1,1]. The equation is given as:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

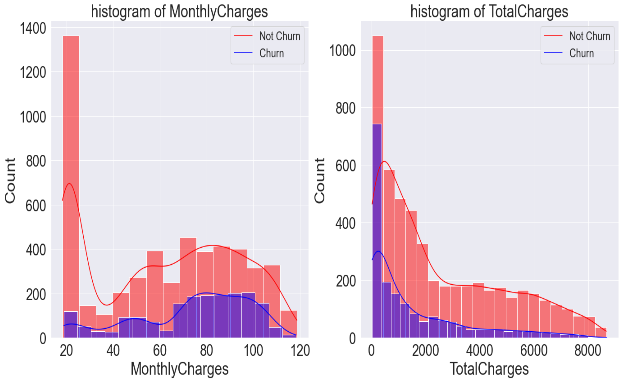


Figure 2: Bivariate Analysis of Numerical and Dependent Variables using histplot

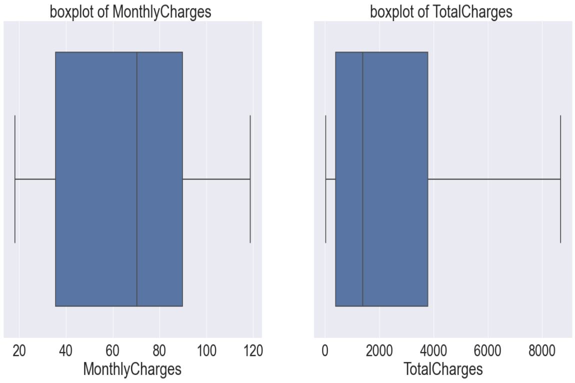


Figure 3: Univariate Analysis of Numerical Variables using boxplot

3.3 Evaluation Metrics

Following Performance metrics have been considered for the evaluation of our proposed model in this research. Accuracy is the closeness of measurement to a specific value, given as,

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

Precision is the probability that the model correctly identifies the true positive case out of total number of predicted positives., given as,

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall, also known as true positive rate or sensitivity, is the probability that the model correctly identifies the true positive case out of total number of actual positives. It is a useful metric for cases including False Negative.

$$\text{Recall} \;=\frac{TP}{TP+FN} \tag{4}$$

F1-Score is the harmonic mean of precision and recall. It gives the combined idea about precision and recall metrics.

$$F1-Score=\frac{2*\text{ Precision }*\text{ Recall}}{\text{Precision }+\text{ Recall}} \tag{5}$$

Where: TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

3.4 Model Building

After all the Preprocessing and Exploratory Data Analysis steps have been carried out, the dataset is splitted into train and test sets into 80/20 ratio. The train set is used for model training and test set is used for evaluation of the performance metrics. At first the Logistic Regression and Random Forest Classifier are individually trained and the results are validated using 10 fold StratifiedKFold Cross Validation. The hyperparameters of each classifiers are tuned using GridSearchCV. Then these two classifiers are used as base learners for our ensemble model using Stacking method. The predicted output probabilities of these two classifier are then provided to the XGBoost classifier for final prediction. The hyperparameters of the XGBoost are tuned and results are cross validated using similar approach.

4. Results and Discussion

The Binary Classifiers were trained using python3 and scikit-learn library using Intel® core™ i7- 10th Gen CPU @2.9 GHz, 16 CPUs, and 16 GB RAM.

4.1 Experimental Setup

The optimal parameters of each of the classifier are obtained using GridSearchCV from Sklearn library with 10-Folds Stratified Cross-Validation. Then, these parameters are finally used during the model evaluation as shown in Tables 2,3 and 4.

Table 2: Hyperparameters of Logistic Regression

Hyperparameters	Base Estimator 1(Logistic Regression)
C	2
penalty	l2
solver	lbfgs
Class_Weight	1: 1.2

Table 3: Hyperparameters of Random Forest

Hyperparameters	Base Estimator 2(Random Forest)
n_estimators	200
max_depth	6
max_leaf_nodes	50
Class_Weight	1: 1.2

Table 4: Hyperparameters of XGBoost

Hyperparameters	Final Estimator(XGBoost)
colsample_bytree	0.6
learning_rate	0.1
max_depth	6
min_child_weight	2
n_estimators	250
scale_pos_weight	1.2

4.2 Results

As shown in Table 5, all the performance metrics has been increased when XGBoost is used as final estimator in stacking classifier as compared to the individually trained standalone model using Logistic Regression and Random Forest. The accuracy obtained using ensemble method is 80.88% and f1 is 62.69% on the test dataset.

Table 5: Performance Metrics of Classifiers

Classifier	Accuracy	Recall	Precision	f1	Support
Logistic Regression	0.8053	0.5989	0.6437	0.6205	1407
Random Forest	0.8109	0.5909	0.6617	0.6243	
Stacking Classifier (XGBoost as Final Estimator)	0.8088	0.6043	0.6513	0.6269	

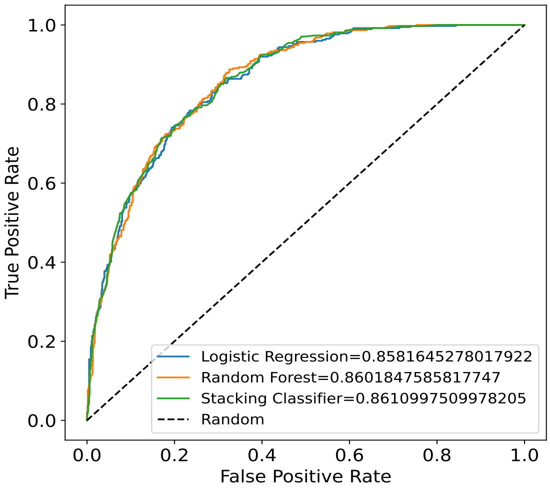


Figure 4: Roc_Auc curve of Classifiers

The Roc_Auc Curves of base classifiers and Stacking Classifier are shown in Figure 4. The Area Under Curve obtained by the final model is 86.109% which is higher than the AUC (84.51%)as obtained by Shuli Wu et al. [3] on the same dataset using AdaBoost algorithm without use of SMOTE.

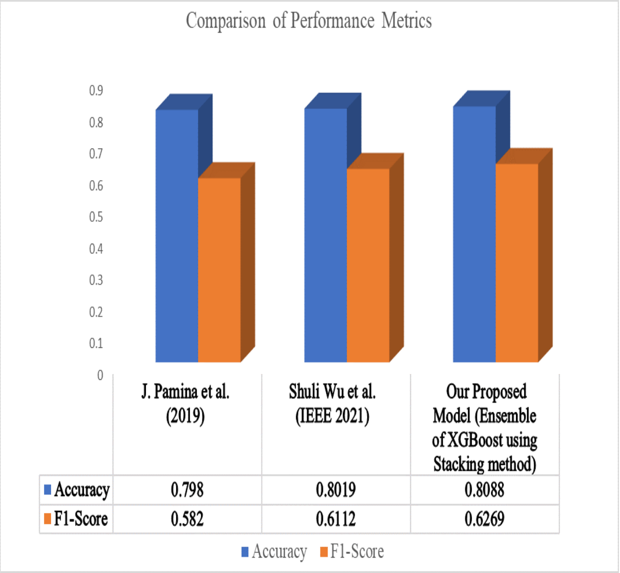
Table 6: Results of 10-fold StratifiedKFold CV

Fold No	fit time	score time	test accu	test rec	test prec	test f1
0	5.972	0.100	0.794	0.572	0.622	0.596
1	5.941	0.084	0.813	0.572	0.673	0.618
2	6.003	0.085	0.799	0.559	0.638	0.596
3	5.676	0.070	0.814	0.610	0.663	0.635
4	5.704	0.090	0.784	0.508	0.613	0.556
5	5.582	0.092	0.787	0.572	0.605	0.588
6	5.666	0.072	0.809	0.583	0.661	0.619
7	5.725	0.102	0.809	0.567	0.667	0.613
8	5.604	0.092	0.805	0.567	0.654	0.607
9	5.468	0.078	0.804	0.561	0.652	0.603

The results of 10-Fold StratifiedKFold Cross-Validation for the Final Model is as shown in Table 6.

4.3 Performance Comparision

The accuracy and f1-score obtained by the final model is 80.88% and 62.69% respectively which is better than the accuracy and f1 (79.8%, 58.2%) as obtained by J. Pamina et al. [2] and the accuracy of 80.19% as obtained by Shuli Wu et al. [3] using Logistic Regression and f1 of 61.12% using Naive Bayes individually as shown in Figure 5.



	J. Pamina et al. (2019)	Shuli Wu et al. (IEEE 2021)	Our Proposed Model (Ensemble of XGBoost using Stacking method)
■ Accuracy	0.798	0.8019	0.8088
■ F1-Score	0.582	0.6112	0.6269

Figure 5: Performance Comparision

5. Conclusion and Future Work

This paper proposes an ensemble based method of churn prediction in telecommunications industry using stacking technique. The final estimator in stacking classifier is XGBoost and base estimators are Logistic Regression and Random Forest. Compared with the performance metrics obtained by using single classifiers, the performance metrics obtained by using ensemble method is better in all aspects as shown in Table 5. Also, the proposed model in this research paper has obtained accuracy and f1-score of 80.88% and 62.69% respectively outperforming the researches done by J. Pamina et al. [2] and Shuli Wu et al. [3] as shown in Figure 5.

The features importance has not been considered in this research paper. A further study of churn prediction model can be studied on this basis. Moreover, as the machine learning algorithms depends on the nature of dataset, the performance of this proposed model may not guarantee the same results on every dataset but by implementing the stacking technique, the overall performance of standalone model can be improved.

References

[1] V Umayaparvathi and K Iyakutti. A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. *International Research Journal of Engineering and Technology (IRJET)*, 3(04), 2016.

[2] Jeyakumar Pamina, Beschi Raja, S SathyaBama, MS Sruthi, Aiswaryadevi VJ, et al. An effective

classifier for predicting churn in telecommunication. *Jour of Adv Research in Dynamical & Control Systems*, 11, 2019.

[3] Shuli Wu, Wei-Chuen Yau, Thian-Song Ong, and Siew-Chin Chong. Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, 9:62118–62136, 2021.

[4] Piotr Sulikowski and Tomasz Zdziebko. Churn factors identification from real-world data in the telecommunications industry: case study. *Procedia Computer Science*, 192:4800–4809, 2021.

[5] Sahar F Sabbeh. Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, 9(2), 2018.

[6] Fahd Idrissi Khamlichi, Doha Zaim, and Khalid Khalifa. A new model based on global hybridization of machine learning techniques for "customer churn prediction". In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–4. IEEE, 2019.

[7] Qi Tang, Guoen Xia, Xianquan Zhang, and Feng Long. A customer churn prediction model based on xgboost and mlp. In *2020 International Conference on Computer Engineering and Application (ICCEA)*, pages 608–612. IEEE, 2020.

[8] Hemlata Jain, Ajay Khunteta, and Sumit Srivastava. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 167:101–112, 2020.

[9] Telco customer churn. `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`, 2017. [Online; accessed 2022].