International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST218498

Predictive Model on Churn Customers using SMOTE and XG-Boost Additive Model and Machine Learning Techniques in Telecommunication Industries

Bechoo Lal[1], **Suraj Kumar**[2]

[1]Department of Information Technology, Western College, University of Mumbai, Maharashtra, India

[2]College of Business, Westcliff University, United States of America (USA)

Article Info

Volume 8, Issue 4

Page Number : 646-661

Publication Issue

July-August-2021

Article History

Accepted : 07 Aug 2021

Published : 14 Aug 2021

ABSTRACT

In this research paper the researcher builds a predictive model on churn customers using SMOTE and XG-Boost additive model and machine learning techniques in Telecommunication Industries. Customer's churning is one of the global research issues in telecommunication industries. In somehow customers are not satisfying from telecommunication customer services, call rate, international plan, data pack, and others which are having a significant impact on customer's services. The researcher used the SMOTE and XGboost technique to handle the imbalanced dataset and gives the higher-level accuracy for predictive model to identify the category of customer whether they are in churn or not churn. The researcher used the comparative study between logistics regression and random forest algorithms to classify the category of churn customers and non-churn customers in Telecommunication Industries. The predictive model is verifying at 96% accuracy level and can be capable to handle imbalance dataset. As per the data analysis the score of the confusion matrix is such as accuracy 94%, Precision for " did not leave " is 0.97 whereas recall is 0.96, and F1score is 0.97 with the support features of 903. For the churn customers precision is 0.80, recall is 0.81, F1-score is 0.80 and support features is 160, the data analysis report shows that the predictive model is having 94% accuracy whereas at 6% does not predict accurately about the customers status. Finally, the researcher concluded that the predictive model is more accurate and can be capable to handle imbalance dataset. The researchers assure that the predictive model would be benefited for the telecommunication industries to categories the churn/ non-churn customers and accordingly the organization can make changes their business plan and policies which would be benefited for the customers.

Keywords : SMOTE, XG-Boost, Predictive, Machine Learning

I. INTRODUCTION

Customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition, customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers – earning business from new customer's means working leads all the way through the sales funnel, utilizing your marketing and sales resources throughout the process. Customer retention, on the other hand, is generally more cost-effective as you've already earned the trust and loyalty of existing customers [1]. Just like most things that can be calculated and measured, it is also possible to do this with your customer churn in several ways, so that you can find out:

The answer is simple: go the extra mile and do it right away by making a good first impression. If a customer is overwhelmed from the first moment, they meet your business they are less likely to continue to look for something better. Continue to meet your customers' expectations and beyond if you can. The quick way to lose a customer fails to deliver, but as customer knowledge begins with your business improves, their trust and commitment will be strengthened, which means that their risk of commotion after a while will be much greater [1]. However, sometimes you must let other customers go; you can't keep everyone happy. In deciding which ones to release, you should keep the benefits in mind. Some customers spend more money, flag your products on social media, and continue to visit your business longer. These customers are worth keeping [7].

The researcher is focused on new technologies, and new competitors are opening the telecommunications industry, portable speculation and management has become a major concern for mobile service providers. The mobile service provider who wishes to retain its subscribers needs to be able to predict which of them may be at risk of switching services and will turn those subscribers into customer retention efforts [5] [6]. In response to the limitations of existing churn forecasting systems and the unavailability of customer numbers at the investigated telecommunications provider, we propose, construct, and scrutinize the churn speculation process that predicts subscription contract information and telephone pattern changes from telephone details. This proposed approach can identify potential participants at the contract level over a period of prediction. In addition, the proposed process includes a multidisciplinary approach to address the challenge of highly distorted section distribution between churner and non-churner [3].

Advances in technology have made relationship marketing a reality in recent years. Technologies such as data storage, data mining, and campaign management software have made customer relationship management a new place where firms can gain competitive advantage. Automatic, future-oriented analysis of data mining operations in addition to the analysis of past events that are often provided with historical focus tools such as decision-making systems (Chris Rygielskia et. Al., 2002). Data mining tools answer business questions that have been time consuming to follow [8]. However, the answers to these questions make customer relationship management possible. Different strategies exist within the data mining software, each with its own advantages and challenges for different types of applications. A specific dichotomy exists between neural networks and the acquisition of automatic chi-square (CHAID) communication [7].

II. BACKGROUND OF THE RESEARCH STUDY

Su-YeonKima et.al (2006) the researcher has proposed a framework for analysing customer numbers and classifying clients according to their value. After the classification of customers according to their value, customer-building strategies will be demonstrated through case studies in wireless telecommunications company Esser Kandogan (2001) who specializes in multi-dimensional data is indicated by a point, where each attribute data contributes to its location by entering the same code. Star Coordinates interactive features give users the ability to apply a variety of dynamic conversions, aggregate and split sizes, multi-dimensional integration, view collections, styles, and locations for sale in data distribution, and query points based on data ranges.

Jae-Hyeon et. Al (2006) stated that quality-related factors contribute to customer frustration, however, customers participating in membership card programs are also more likely to arise, raising questions about the effectiveness of the system. Pınar Kisioglu and Y. Ilker Topcu (2011) emphasized that management, companies try to keep their existing customers, instead of getting new ones. A researcher mapped as a basis for the Bayesian Belief Network is presented with the results of integration analysis, multicollinearity testing and expert opinions. According to the results of the Bayesian Belief Network, intermediate call minutes, standard billing rate, frequency of calls to individuals from various providers and the type of tax are the most important variables that define customer fraud.

Chen-FuChiena and Li-FeiChenab (2008) developed a data mining framework based on decision-making rules and organizational rules for making useful workers' selection rules. The results can provide decision-making rules relating to employee knowledge of work and performance. Chih-Fong and TsaiaYu-HsinLub (2009) focused on hybrid models by combining two different neural churn predictive network techniques, namely artificial distribution networks (ANN) and self-organizing maps (SOM). Results as representative data are used to create a predictive model based on the second process. To test the effectiveness of these types, three different types of test sets are considered.

Yaya Xiea et. Al (2008) emphasized that improving predictive accuracy is highly comparable to other algorithms, such as artificial neural networks, decision trees, and vector basic weighted class (CWC-SVM) equipment. In addition, IBRF also produces better predictive results than other random algorithms such as moderate forests and random forests. Shin-Yuan et., Al (2006) stated that this study shows that both tree and network decisions can produce accurate predictor models using customer statistics, payment details, contract / service status, call details, and service log changes.

Hyunseok Hwang, and Taesoo Jung EuihoSuh (2004), a study conducted to calculate the number of customers based on the value of the Customer's life (LTV). However, there are limitations. It’s hard to imagine a customer revolt. The types of forecasts focus on the expected future cash flows based on past customer benefits. Miguel A.P.M. Lejeune, (2001) emphasized that data mining development is intended to play a major role in churn management. Relying on sensitivity analysis, we propose an analysis framework that can determine the potential impact of continuous data development on churn management and decision-making processes.

John Haddenaet et. al (2007) focuses on a variety of advanced churn management strategies in response to the above requirements. The focus of this paper is to review some of the most popular technologies found in the literature for the construction of customer management platform. Scott A. Neslin et.al (the accuracy of the 2006 speculation on all submissions could alter the profitability of a fraudulent management campaign by hundreds of thousands of dollars. Second, beauticians can stay. They experience a very small decrease in performance when predicting a data explosion compiled three months after power data.

Dudyala Anil Kumar. Ravi (2002) emphasized that Multilayer Perceptron (MLP), Logistic Regression (LR), Decision Trees (J48), Random Forest (RF), Radial Basis Function (RBF) and Russel Vector Machine (SVM) as the most measurable data elements with trusted clients 93% and 7%, Shu-Hsien et., al (2012) prescribes DMT, in terms of the following three areas: types of information, types of analysis, and types of structures, and their types of applications in various studies and functional domains.Bong - Horng et. al (2007) identified in the churn model to divide all 'churners' into distinct groups. Policy model builder is also responsible for developing a policy model for each churner group. In usage mode, the churn forecast uses the churn model to predict the churn potential of a given registrar.

Dirk Van den and Poel Bart Larivière (2004) focused on predictors becoming one complete model for retention including several 'new' types of various covariates related to real customer behaviours; (2) by analysing churn performance based on a random sample of the number of people using longitudinal data from the data warehouse. Kristof Coussementab et. al (2010) used Generalized Additive Models (GAM) such as Logistic Regression, GAM lowers the linearity limit that allows for non-linear equilibrium of data. Contributions that better identify risky customers; (ii) shows that GAM enhances churn-type interpretation by visualizing non-linear relationships with customer identification indicators of fundamental prominence, U, U modified or complex trends and (iii) market managers can significantly increase the value of the business by using GAM in this context. Amal M. Almana et al (2014) emphasized that customer fraud has emerged as one of the major problems in the Telecom Industry. Telecom research shows that it is more expensive to find a new customer than to keep an existing one. To retain existing customers, Telecom providers need to know the reasons for the fraud, which can be traced to the information extracted from Telecom data.

Rahul J. et.al., (2011) argued that data mining could be incorporated into the communications sector to identify Churn's predictions of customers leaving the company's platform called Churn Prediction. The company must take the necessary steps to maintain it. Gary M. Weiss (2005) who researches arrogant clients and finds a solution to research problems using data mining, is the first step in understanding data. Li-Shang Yang, Chaochang Chiu (2006) focused on preliminary research on the use of data mining techniques to solve the real-world customer problem in the communications sector. Ionut Brandusoiu, and Gavril Toderean (2013) emphasized that a high-performance, four-kernel Support Vector Machines algorithm has been used by churn and non-churn buyers. Amjad Hudaib et.al., (2015) was tested to develop existing models, three Hybrid models designed for active churn prediction in the telecommunications market.

III. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

Customers churn is one of the significant research issues in telecommunication industries. The customers affected by telecommunication services such call rate, voice message, voice mail, international plan, and customers services in case of any issues. Churning customers are big loss for the telecommunication industries and it is one of the global research issues how to predict churning customers before from the telecommunication industries. In this research study the research developing predictive model to predictive churn customer and non-churn. The predictive model is also capable to handle biased and imbalance dataset. The researcher formulated some of the research objectives which are stated as:

1. To study the churn/non-churn customers in telecommunication industries and its significant research issues
2. To develop a predictive model with high accuracy to classify churn/ non-churn customers.
3. To evaluate the accuracy of predictive model and optimized it.

IV. NATURE OF VARIABLES AND DATASET

The researcher used the data set with 4250 entries where 3400 for training dataset and 850 entries for testing dataset for classifier to classify the data in terms of 80:20 ratio. The entire datasets are represented as follows:

1. Let the complete datasets be represented D= {D1 ,D2, D3, D4………..D4250},
2. Let the training datasets be presented as Train={D1 ,D2, D3, D4…….D3400},
3. Let the test data be represented as Test={D326 ,D327, D328, D328…….D850},

The splitting of dataset is based on the random manner, system automatically divided the two different datasets in terms of ration 80:20 manner which is one of the standard mapping parameters to train and test the dataset in machine learning model.

DATA SOURCE: KAGGLE.COM, INDUSTRY: TELECOMMUNICATION

Table1.1: Nature of Variables and Dataset

S.NO	Variables/Features	Data Status	Category
0	state	4250 non-null	object
1	account_length	4250 non-null	int64
2	area_code	4250 non-null	object
3	international_plan	4250 non-null	object
4	voice_mail_plan	4250 non-null	object
5	number_vmail_messages	4250 non-null	int64
6	total_day_minutes	4250 non-null	float64
7	total_day_calls	4250 non-null	int64
8	total_day_charge	4250 non-null	float64
9	total_eve_minutes	4250 non-null	float64
10	total_eve_calls	4250 non-null	int64
11	total_eve_charge	4250 non-null	float64
12	total_night_minutes	4250 non-null	float64
13	total_night_calls	4250 non-null	int64
14	total_night_charge	4250 non-null	float64
15	total_intl_minutes	4250 non-null	float64
16	total_intl_calls	4250 non-null	int64
17	total_intl_charge	4250 non-null	float64
18	number_customer_service_calls	4250 non-null	int64
19	churn	4250 non-null	object

V. RESEARCH DESIGN AND METHODLOGY

In this research article the researcher used the SMOTE and XGboost technique to handle the imbalanced dataset and gives the higher-level accuracy for predictive model. With respect to the predictive model of machine learning the researcher used the comparative study between logistics regression and random forest algorithms to classify the category of churn customers and non-churn customers in Telecommunication Industries. There are several factors which are directly affected to the telecommunication services such as proper connectivity, international calls rate, day and night call rate, customer care/support services etc.

5.1 SMOTE: Synthetic Minority Oversampling Technique

SMOTE is a tool to control the imbalanced classification for a model to effectively learn the boundary decisions. To solve the data imbalanced problem, this can be achieved by simply duplicating examples from the minority class in the training datasets prior to fitting a model. This can balance the class distribution but does not any additional information to the predictive model. An improvement on duplicating examples from the minority class is to synthesize new examples from minority class. This is a type of data augmentation for tabular data and can be more effective to enhance the predictive model.

ALGORITHMS:

Step-1: Setting the minority class set A, for each x € A, the K-nearest neighbours of x are obtained by calculating Euclidean distance between x and every other sample in set A.

Step-2: The sampling rate N is according to the imbalance proportion. For each x € A, N examples (i.e x1,x2, x3……………….,xn) are randomly selected from its K-nearest neighbours, and construct the set A1.

Step-3: For each example xk € A1(k1,2,3,4,………..N), the following formula is used to generate a new example :

$X' = X + rand(0,1) * | X - Xk |$

in which rand(0,1) represents the number between 0 and 1.

SMOTE first select a minority class instance a at random and finds its k nearest minority class neighbours. The synthetic instance is then created by choosing one of the k nearest neighbours b at random and connecting a and b to form a line segment in the feature space. The synthetic instance is generated as a convex combination of the two chosen instances a and b.

5.2 XG-Boost:

It is an ensemble additive model that is composed of several base learners. XG-Boost uses the Taylor series to approximate the value of the loss function for a base learner ft(xi), thus , reducing the load on Emily to calculate the exact loss for different possible base learners.

Input training set $\{ (xi,yi)\}^{n}_{i=1,}$ a different loss function L(y,F(x)), number of iterations M.

Algorithms:

Step-1: Initialize model with a constant value

$F_0(x)=\arg\min \sum_{k=1}^{n} L(yi, y)$……………………(1)

Step-2: For m=1 to M

1. Compute so called pseudo residuals

$$rim=\left|\frac{[\text{Ə}L(yi, F(xi))]}{\text{Ə}F(xi)}\right|$$

$F(x)=F_{m\ 1}(x)$……………(2)

for i=1,2,3,4,…………..n.

2. Fit a base learners (e.g tree) hm(x) to pseudo residual i.e train it using the training set

$\{(xi, rim)\}_{i=1}^{n}$

3.γ compute multiplier γ_m by solving the one-dimension problem

$\gamma_m = \arg\min \sum_{k=1}^{n} L(yi, Fm - 1(x) + \gamma hm(x))$…..(3)

4. Update the model

$F_m(x)=Fm - 1(x) + \gamma mhm(x)$……..(4)

Step-3: Output $F_M(x)$

XG-Boost starts with an initial prediction and use the loss function to evaluate if the prediction works well or not. In this equation the first part represents the loss function which calculates the pseudo residuals of predicted value y_i with hat and true of y_i in each leaf whereas x_i represents the number of features which works like as independent variable.

$X_i = x1, x2, x3, x4 \ldots\ldots x_n$

$Y_i = y1, y2, y3, y4, \ldots\ldots y_n$

Where Xi represent independents variables datasets and $Y_{i\ \text{represents}}$ dependents datasets.

5.3 LOGISTIC REGRESSION

Logistic function is also called sigmoid function was developed by the statisticians to describe the properties of populations growth. It is S shaped curve that can take any real value and map it into a value between 0 and 1, but never exactly at those limits.

$$\text{Sigmoid Function} = \frac{1}{(1+e^{-\text{Value}})}$$

Where e is the base of the natural algorithms (Euler's number or EXP) and the value is the actual numeric value that it is going to transform between into the range 0 and 1 using logistic function.

Logistic regression uses an equation as the representation which is very much like as linear

$$y = \frac{1}{e^{(bo+b1^*x)}}$$

Input value (x) are combined linearly using weights or coefficients value to predict an output value (y) where y is the predicted output , b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in input datasets has an associated b coefficient that must be learn from training dataset.

5.4 RANDON FOREST CLASSIFIERS

Random forest is a supervised learning algorithms which is used for both classification as well as regression. It is mainly used for classification problems. As the researcher emphasized that forest is made up of trees and more trees means more robust forest. Random forest algorithms creates decision tree on the data samples and then get the prediction on each of them and finally select the best solutions by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over fitting by averaging the result.

Algorithm:

Step-1: First start the selection of random samples from a given dataset.

Step-2: Next, this algorithm will construct a decision tree for every sample then it will get the prediction result from every decision tree.

Stept-3: In this step, voting will be performed for every predicted results.

Step-4: At last, select the most voted prediction result as the final prediction result.

VI. RESULTS AND DISCUSSION

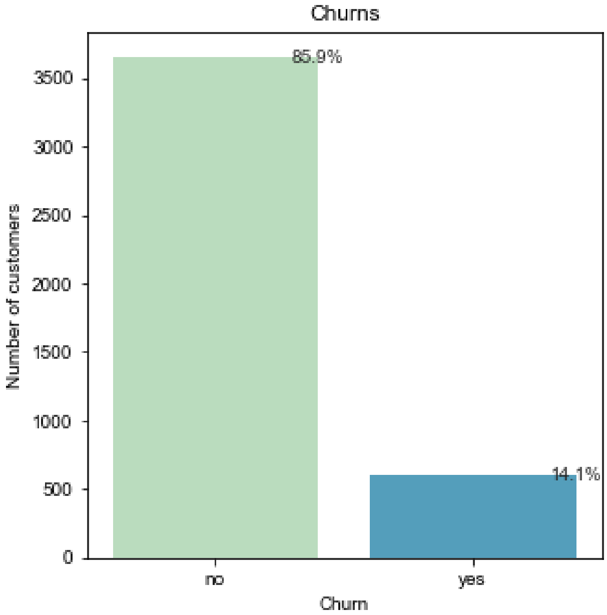

Fig.1.1: Statistics of Churn and Non-churn customers

The above data statistics shows that 86% customers are in non-churn category whereas 14% belong to churn customers. This is the current statistics of dataset in telecommunication industry. There are numbers of features who are responsible for this churn category. In this research article the research will develop and predictive model with high accuracy to predict whether customer would be churn or not in future (Fig.1.1).

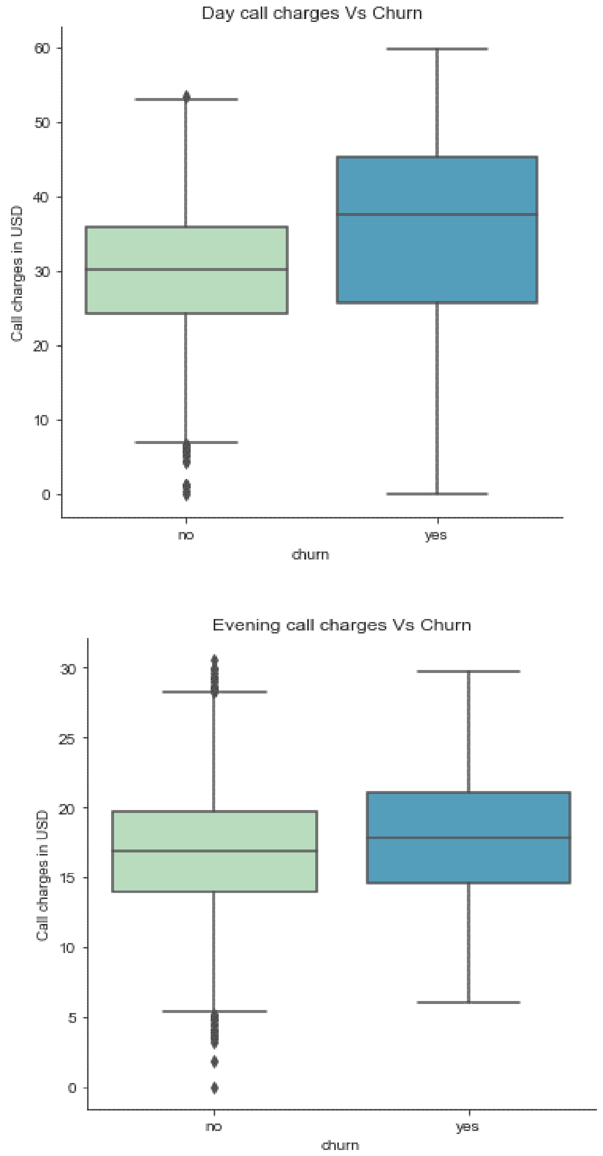

Fig.1.2: Statistics of Daily and Evening Calls- Churn Customers

The above data statistics shows that daily and evening calls where customer are moving in case of the churn category. The researcher used the boxplot to show that minimum, maximum, mean call charges which are significantly affected to customer which are in churn category. In daytime the statistics shows that call charges are high rather than non-churn customers whereas at night both are very similar category but still class charges are higher than non-churn customers. The researcher identify that calls charges is one of the significant factors which is responsible for churning a customer in telecommunication industries (fig.1.2).

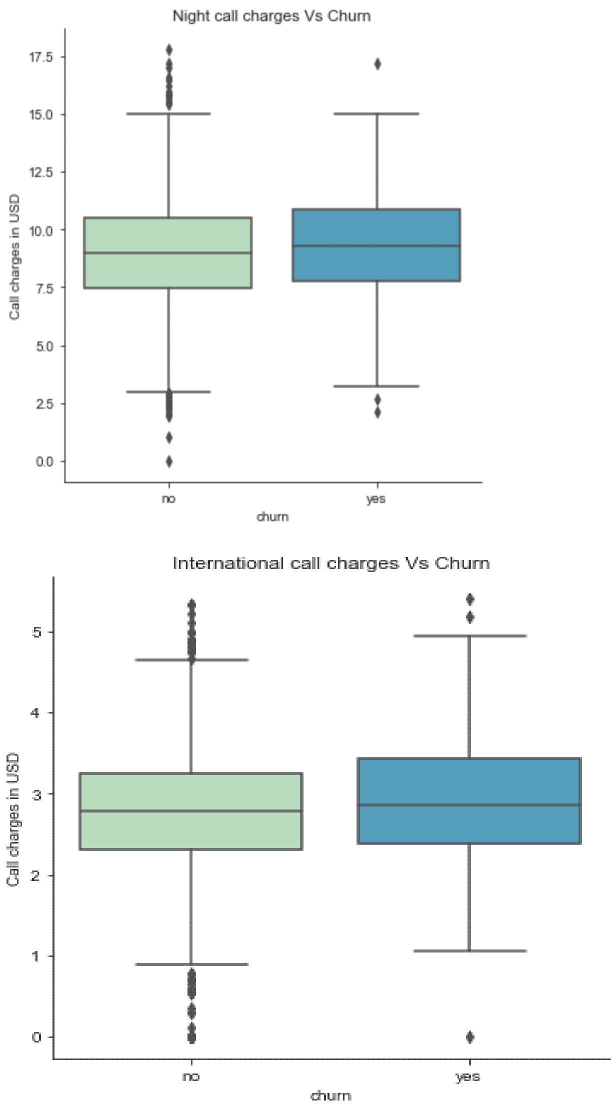

Fig.1.3: Statistics of Night and international Calls Charges- Churn Customers

The above data statistics shows that night and international calls where customer are moving in case of the churn category. The researcher used the boxplot to show that minimum, maximum, mean call charges which are significantly affected to customer which are in churn category. In night-time the statistics shows that call charges are high rather than non-churn customers whereas at international calls charges both are very similar category but still calls charges are higher than non-churn customers. The researcher identify that calls charges is one of the significant factors which is responsible for churning a customer in telecommunication industries (fig.1.3).

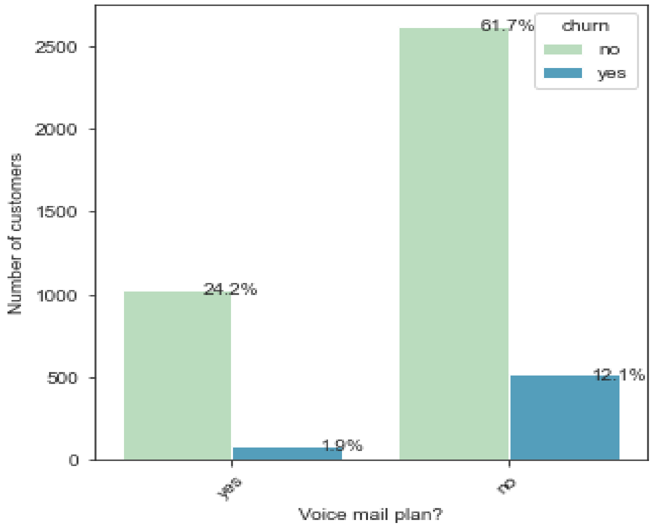

Fig.1.4: Statistics of Voice Mail Plan - Churn Customers

The above data statistics shows that customers are having the voice mail plan or not, as per the statistics shows that 25.1% customers are having voice mail plan whereas 73.9% customers are not having voice mail accessibility. In case of the churn statistics the data analysis report shows that the customer who are having voice mail plan 1.9% are in churn category whereas 24.2% are in non-churn category. The customer who are not having the voice mail plan the data statistics shows that 91.7% are not in churn category whereas 12.1% are in churn customer category (fig.1.4).

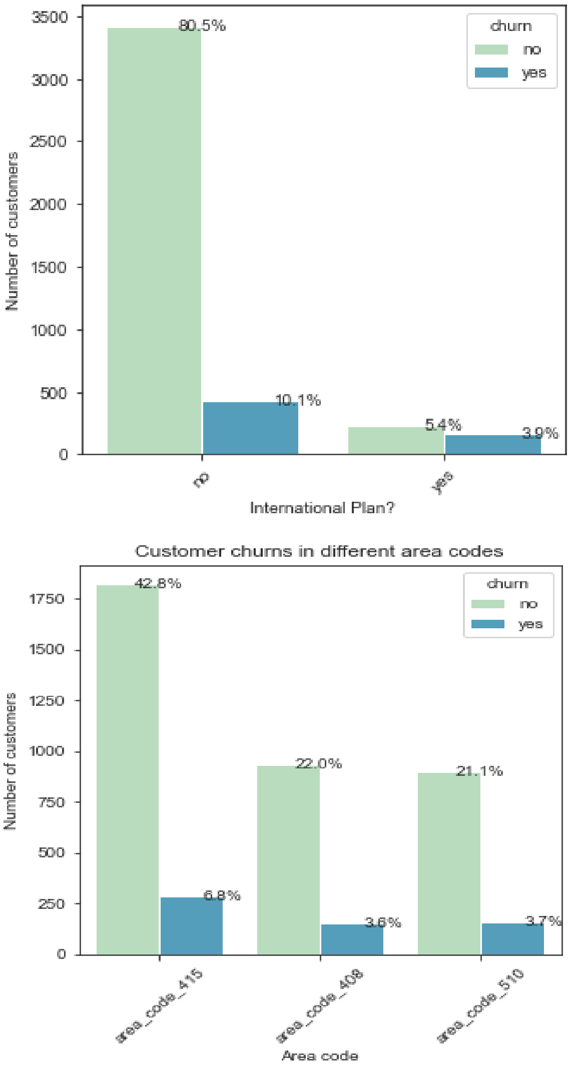

Fig.1.5: Statistics of International Plan and Different Areas - Churn Customers

The above data statistics shows that customers are having the international plan or not, as per the statistics shows that 90.6% customers are having international plan whereas 9.4% customers are not having international accessibility. In case of the churn statistics the data analysis report shows that the customer who are having international plan 10.1% are in churn category whereas 80.5% are in non-churn category. The customer who are not having the voice mail plan the data statistics shows that 5.4% are not in churn category whereas 3.5% are in churn customer category. The researcher concluded that the customers who are having international call plan are having higher churn rate rather than who are not having international call accessibility (fig.1.5).

The above statistics also shows the location wise churn prediction such as the data statistics identify that New Jersey has the maximum percentage of customer churns (27%) followed by California (25%) and Washington(22%), it is concluded that Area code 415 has the maximum customer churn of 6.8%.

As per the data statistics shows that the dataset is imbalanced with 86% data where customer has not churned the model prediction might be get biased. If the model only predict that customers have not churned, then the accuracy of the model will be 86%. To get the dataset balanced the researcher have implemented the SMOTE. It is first oversampled the minority class and then under sample the majority class (Fig.1.1).

6.1 LOGISTIC REGRESSION: ACCURACY STATISTICS

Table 1.1: Model Used : Logistics Regression

Best Score	0.8012218118347437
Best Hyper-parameters	{'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}
Score of the Logistic Regression	0.8184383819379115

The above data analysis results shows that how logistic regression analysis playing a significant role to category the churn and non-churn customers from the given dataset. The researcher used the logistic regression to predict the customers categories are in churn or not in churned. The score of the logistic regression shows that the predictive model is having 81% accuracy whereas at 19% does not predict accurately about the customers status.

RANDOM FOREST CLASSIFIERS: ACCURACY STATISTICS

Table 1.2: Model Used: Logistics Regression

Best Score	0.8538127074664548
Best Hyper-parameters	{'criterion': 'entropy', 'max_features': 'auto', 'min_samples_leaf': 4, 'n_estimators': 7}
Score of the Random forest	0.9040451552210724

The above data analysis results shows that how random forest classifiers playing a significant role to category the churn and non-churn customers from the given dataset. The researcher used the random forest classifiers to predict the customers categories are in churn or not in churned. The score of the random forest classifiers shows that the predictive model is having 90% accuracy whereas at 10% does not predict accurately about the customers status. At this level the researcher concluded that the random forest classifiers produced the more accurate result than logistics regression.

6.2 CONFUSION MATRIX-1

Table 1.3: Confusion Matrix-1

	Did not leave	Left
Did not leave	840	63
Left	39	121

$$\text{Accuracy} = \frac{(840+121)}{(840+121+39+63)}$$

$$= 0.90404515522$$

The researcher used the confusion matrix - random forest classifiers to predict the customers categories are in churn or not in churned. The score of the random forest classifiers shows that the predictive model is hav ing 90% accuracy whereas at 10% does not predict accurately about the customers status (Table 1.3).

Table 1.4: Confusion Matrix-1: Description

	Precision	Recall	F1-Score	Support
Did not leave	0.96	0.93	0.94	903
Left	0.66	0.76	0.70	160
Accuracy			0.90	1063
Macro age	0.81	0.84	0.82	1063
Weighted avg	0.91	0.90	0.91	1063

The above confusion matrix results shows that how it is playing a significant role to measuring the accuracy level of predictive model in case of churn and non-churn customers from the given dataset. The researcher used the confusion matrix to find the accuracy, precision, recall and F1-score to predict the customer's categories are in churn or not in churned and its accuracy level. As per the statistics given the score of the confusion matrix are such as accuracy 90%, Precision for " did not leave " is 0.96 where as recall is 0.93, and F1score is 0.94 with the support features of 903.

For the left customers precision is 0.66, recall is 0.76, F1score is 0.70 and support features is 160, the data analysis report shows that the predictive model is having 90% accuracy whereas at 10% does not predict accurately about the customers status. At this level the researcher concluded that the random forest classifiers produced the more accurate result than logistics regression (Table1.4.)

6.3 XG-Boost: RESULT ANALYSIS

Table 1.5: XG-Boost: Result Analysis

Iteration	Specification	Values
0	validation_0-auc	0.86366
50	validation_0-auc	0.99801
100	validation_0-auc	0.99945
150	validation_0-auc	0.99990
200	validation_0-auc	0.99999
250	validation_0-auc	1.00000
300	validation_0-auc	1.00000
Stopping. Best iteration:		
218	Validation_0-auc	1.00000

For more accuracy the researcher used the XG-Boost algorithms which started at iteration number 0 with v alue 0.6366 and passes through every 50 interval such as 0, 50, 100, 150,200,250, and 300 and until the predi ctive model get the stable result. As we can see that af ter 200 the predictive model produced the unique res ult and nothing changing in their status value. Let's n ow check the confusion matrix and the classification r eport. Certainly, there is a great improvement. The ac curacy is now 99% and the Precision and Recall has i mproved drastically (Table 1.5).

6.4 CONFUSION MATRIX-2:

Table 1.6: Confusion Matrix-2

	Did not leave	Left
Did not leave	870	33
Left	30	130

$$\text{Accuracy} = \frac{(870+130)}{(870+130+30+33)}$$

$$= 0.94073377234$$

The researcher used the confusion matrix-2 to predict the customers categories are in churn or not in churn ed. The results statistics shows that 870 out of 1063 w hich are truly predicted did not leave the customers w hereas 130 truly predicted the customers which are le ft the telecommunication services, the data value 33 a nd 30 which are wrongly predicted in case of positive false and true negative. The score of the confusion ma trix -2shows that the predictive model is having 94% accuracy whereas at 6% does not predict accurately a bout the customers status (Table 1.6).

Table 1.7: Confusion Matrix-2: Description

	Precisio n	Recall	F1-Score	Support
Did not l eave	0.97	0.96	0.97	903
Left	0.80	0.81	0.80	160
Accurac y			0.94	1063
Macro av g	0.88	0.89	0.89	1063
Weighte d avg	0.94	0.94	0.94	1063

The above confusion matrix-2 results shows that how it is playing a significant role to measuring the accuracy level of predictive model in case of churn and non-churn customers from the given dataset. The researcher used the confusion matrix-2 to find the accuracy, precision, recall and F1-score to predict the customer's categories are in churn or not in churned and its accuracy level. As per the statistics given the score of the confusion matrix are such as accuracy 94%, Precision for " did not leave " is 0.97 whereas recall is 0.96, and F1score is 0.97 with the support features of 903.

For the left customers precision is 0.80, recall is 0.81, F1score is 0.80 and support features is 160, the data analysis report shows that the predictive model is having 94% accuracy whereas at 6% does not predict accurately about the customers status. At this level the researcher concluded that the predictive model produced the more accurate result than previous (Table1.7).

6.5 DISCUSSION

Shin-Yuan et., Al (2006) emphasized that churn management focuses on cell phone owners to retain subscribers by satisfying their needs under resource constraints. Hyunseok Hwang, and Taesoo Jung EuihoSuh (2004), since the early 1980s, the concept of corporate relations management in the marketplace has gained its importance. Finding and retaining highly profitable customers is a major concern for the company in conducting highly targeted advertising campaigns. Miguel A.P.M. Lejeune, (2001), Churn's management is critical to business and the emergence of a digital economy has made the problem even more difficult. John Haddenaet et. al (2007), An entity has a very high cost of trying to win new customers rather than keeping existing ones. YayaXiea et. al (2008) Churn's speculation has become more focused on retaining customers by satisfying their needs under resource constraints. In churn forecasting, an important but challenging problem is the inequality in data distribution. Scott A. Neslin et.al (2006), provides a descriptive analysis of how trends contribute to the accuracy of customer speculation models.

Finally, in this research study the researcher handled a data equivalent to 86% of data where the customer did not consider the prediction of the model to be biased. If the model only predicts that customers have not been excluded, the accuracy of the model will be 86%. To create a database equate a researcher using SMOTE. According to statistical analysis matrix-2 confusing points such as 94% accuracy, Precision "did not leave" is 0.97 while memory is 0.96, while F1score is 0.97 with 903 support features. For left-hand customers the accuracy is 0.80, the memory is 0.81, the F1-score is 0.80 and the support features are 160, the data analysis report shows that the prediction model has 94% accuracy while 6% does not accurately predict customer status. The researcher confirmed that the prediction model provides 94% accuracy in predicting the client's status whether it is in churn or non-churn categories.

VII. LIMITATIONS

In this research study the researcher tries to cover all possible dimension of identification of churn customers in telecommunication industries using machine learning algorithms.

1. The predictive model is based on biased data.
2. Difficult to handle manually.

VIII. FUTURE RESEARCH

This research study is based on classification of churn customers in telecommunication industries using machine learning algorithms. The researcher developed a predictive model which is verified at 96% accuracy level to predict the status of customer whether they would belong to churn or non-churn categories. The researcher intended for further research study on churn customers in telecommunication industries.

1. Work on biased data and accuracy level.
2. To explore this research study at global level
3. Explore this research study to handle more than two categorical data

IX. CONCLUSION

Finally, the researchers concluded that the predictive model is more accurate to predict the category of customers whether they belong to churn or non-churn. The researcher used machine learning algorithms to build a predictive model on churn / non –churn customers in telecommunication industries. The predictive model which is based on logistic regression having 81% accuracy whereas at 19% does not predict accurately about the customers status. The accuracy score of the random forest classifiers shows that the predictive model is having 90% accuracy whereas at 10% does not predict accurately about the customers status. At this level the researcher concluded that the random forest classifiers produced the more accurate result than logistics regression. As per the statistics given the score of the confusion matrix are such as accuracy 90%, Precision for " did not leave " is 0.96 whereas recall is 0.93, and F1score is 0.94 with the support features of 903. For the left customers precision is 0.66, recall is 0.76, F1score is 0.70 and support features is 160, the data analysis report shows that the predictive model is having 90% accuracy whereas at 10% does not predict accurately about the customers status. For more accuracy and to handle the imbalance dataset the researcher used the SMOTE and XG-Boost algorithms which started at iteration number 0 with value 0.6366 and passes through every 50 interval such as 0, 50, 100, 150,200,250, and 300 and until the predictive model get the stable result. Finally, the confusion matrix and the classification report got a great improvement. The accuracy of predictive model is now 94% and the Precision and Recall has improved drastically.

X. REFERENCES

[1]. YayaXieaXiuLiaE.WTNgaibWeiyunYingc (2008). Customer Predicting using Advanced Informal Forests, Expert Systems with Applications, Volume 36, Issue 3, Part 1, April 2009, Pages 5445-5449, Copyright © 2008 Elsevier Ltd. https: //doi.org/10.1016/j.eswa.2008.06.121.

[2]. Shin-Yuan, HungaDavid, noC. YenbHsiu-YuWangc (2006). Applying Data Mining To Telecom Churn Management, Expert Systems with Applications, Volume 31, Issue 3, October 2006, Pages 515-524, Copyright © 2008 Elsevier Ltd. https://doi.org/10.1016/j.eswa.2005.09.080.

[3]. Hyunseok Hangman Taesoo Jung EuihoSuh (2004). An LTV model and customer segmentation based on value: a case study on the wireless telecommunication industry, Expert Systems with Applications, Volume 26, Issue 2, February 2004, Pages 181- 188, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0957-4174(03)00133-7

[4]. Miguel A.P.M. Lejeune, (2001). Measuring the impact of data mining on churn management, Internet Research, Vol. 11 Issue: 5, p. 375-387, https://doi.org/10.1108/10662240110410183.

[5]. Chih-Ping, Weia-Tang Chiub (2002). Transforming telecommunications data to predict prediction: data mining method, Application Systems System, Volume 23, Issue 2, August 2002, Pages 103-112, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0957-4174 (02) 00030-1.

[6]. John Haddena, Ashutosh Tiwari Raj Kumar, and Roya Dymitr Rutab (2007). Computer assisted customer churn management: State-of-the-art and future trends, Computers & Operations Research, Volume 34, Issue 10, October 2007, Pages 2902-2917, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.cor.2005.11.007.

[7]. Scott A. Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason (2006). Error Detection: Measuring and Understanding the Predictable Accuracy of Churn Customer Models. Marketing Research Journal: May 2006, Vol. 43, no. 2, pages 204-211, https: //doi.org/10.1509/jmkr.43.2.204.

[8]. Chris Rygielskia, Jyun-Cheng, and WangbDavid C. He (2002). Methods of data mining customer management data, Technology in Society, Volume 24, Issue 4, November 2002, Pages 483-502, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0160-791X (02) 00038-6.

[9]. Dudyala Anil Kumar, V. Ravi (2002). Predicting credit card customers in data mining banks, International Journal of Data Analysis Techniques and Strategies, Volume 1, Issue 1, 1 Institute for Development and Research in Banking Technology, Castle Hills Road # 1, Masab Tank, Hyderabad 500 057 (AP), India. https://doi.org/10.1504/IJDATS.2008.02002.

[10]. Shu-Hsien, LiaoPei-Hui, and ChuPei-YuanHsiao (2012). Data mining techniques and their use - Tenth Review from 2000 to 2011, Expert Systems with Applications, Volume 39, Issue 12, 15 September 2012, Pages 11303-11311, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2012.02.063.

[11]. Bong-Horng, ChuacMing-ShianTsaia, and Cheng-SeenHob (2007). Toward a hybrid data mining model for retention customer', Knowledge Based Knowledge, Volume 20, Issue 8, December 2007, Pages 703 -718, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.knosys.2006.10.003.

[12]. Su-YeonKima, Tae-Soo JungbEui-HoSuhc, and Hyun-SeokHwangd (2006). Customer segregation and strategic development

according to the value of customer life: Case studies, Application System, Volume 31, Issue 1, July 2006, Pages 101- 107, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2005.09.004.

[13]. Eser Kandogan (2001). Visualizing multi-dimensional cluster, trends, and outliers using star coordinates, KDD '01 Proceedings of ACM SIGKDD's seventh conference on data acquisition and data mining, Pages 107-116, San Francisco , California - August 26 - 29, 2001, ACM New York, NY, USA © 2001, ISBN: 1-58113-391-X, doi. 10.1145 / 502512.502530.

[14]. Jae-Hyeon, AhnaSang-PilHana, and Yung-SeopLeeb (2006). Customer Churn Analysis: Churn Symptoms and Consequences of Few Discrimination in the Korean Telecommunications Industry, Communication Policy, Volume 30, Issues 10 -11, November - December 2006, Pages 552-568, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.telpol.2006.09.006.

[15]. Pınar Kisioglu and Y. Ilker Topcu (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey, Expert Systems with Applications Volume 38, Issue 6, June 2011, Pages 7151 -7157, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2010.12.045.

[16]. Chen-FuChiena and Li-FeiChenab (2008). Data mining to improve staff selection and improve human performance: Study studies in the high technology industry, Systems System with Applications, Volume 34, Issue Pages 280-290, 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2006.09.003.

[17]. Chih-Fong and TsaiaYu-HsinLub (2009). Customer churn prediction by hybrid neural network, Expert Systems with Applications, Volume 36, Issue 10, December 2009, Pages 12547-12553, Copyright © 2008 Elsevier Ltd, https: //doi.org/10.1016/j.eswa.2009.05.032.

[18]. Dirk Van den and Poel Bart Larivière (2004). Customer analysis of financial services using risky forms, European Journal of Operational Research, Volume 157, Issue 1, 16 August 2004, Pages 196-217, Copyright © 2008 I -Elsevier Ltd, https://doi.org/10.1016/S0377-2217 (03) 00069-9.

[19]. Kristof Coussementab, Dries F. BenoitbDirk, and Van den Poelb (2010). Improving marketing decisions in the context of customer speculation using standard add-on models, Expert Systems with Applications, Volume 37, Issue 3, 15 March 2010, Pages 2132 -2143, Copyright © 2008 Elsevier Ltd., https://doi.org/10.1016/j.eswa.2009.07.029.

[20]. Amal M. Almana et al (2014). Research on Data Mining Methods in Churn For Customer Analysis For Industry Telecom, Int. Engineering Research and Applications Journal www.ijera.com, ISSN : 2248-9622, Vol. 4, Issue 5(Version 6), May 2014, pp.165-171.

[21]. Dr. Mamta Madan Dr. Meenu Dave Vani Kapoor Nijhawan (2015). A Review on: Data Mining for Telecom Customer Churn Management, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 9, September 2015.

[22]. Nabgha Hashmi, Naveed Anwer Butt and Dr. Muddesar Iqbal (2013). Customer Churn Prediction in Telecommunication A Decade Review and Classification, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013.

[23]. Rahul J. Jadhav, Usharani T. Pawar (2011). Churn Prediction in Telecommunication Using Data Mining Technology, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 2, February 2011.

[24]. Li-Shang Yang, Chaochang Chiu (2006). Subscriber Churn Prediction in Telecommunications, 2006.

[25]. Ionut Brandusoiu, Gavril Toderean (2013). Churn Prediction In The Telecommunications Sector Using Support Vector Machines Issue # 1, May 2013.

[26]. Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, Hossam Faris (2015). Hybrid Data Mining Models for Predicting Customer Churn, J. Communications, Network and System Sciences, May 2015, 8, 91-96.

Cite this article as :

Bechoo Lal, Suraj Kumar , "Predictive Model on Churn Customers using SMOTE and XG-Boost Additive Model and Machine Learning Techniques in Telecommunication Industries ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 4, pp. 646-661, July-August 2021. Available at doi : https://doi.org/10.32628/IJSRST218498
Journal URL : https://ijsrst.com/IJSRST218498