International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2020) pp. 265-275

Received: 18 Feb, 2020; Accepted: 15 July, 2020; Published: 1 August, 2020

Semi-Supervised Learning Approach to Improve Machine Learning Algorithms for Churn Analysis in Telecommunication

Bindu Rani[1], Shri Kant[2]

[1]Department of Computer Science and Engineering,Sharda University,
Knowledge Park-3, Greater Noida, India
bindu.var17@gmail.com

[2]Research and Technology Development Centre, Sharda University,
Knowledge Park-3, Greater Noida, India
Shrikant.ojha@gmail.com

***Abstract*: In semi supervised learning, knowledge is acquired with the help of unlabeled and labeled data both. Supervised classification predicts the labels of unknown data with the guidance of labeled data. To obtain the labeled data in sufficient amount and at low cost is challenging task. This paper presents comparative research on six most widely used machine learning classifiers for churn prediction in telecommunication. We also propose a pseudo label semi supervised learning model that could validate the improvement in the classifiers performance by exploiting large volume of unlabeled data with partnership of small-labeled data. In the first stage, six supervised algorithms like SVM(Support Vector Machine), Random Forest, Logistic Regression, AdaBoosting, Gradient Boosting and eXtreme Gradient Boosting are applied and assessed using cross validation technique along with external measures on telecom dataset for churn prediction. In second stage, the improvement in performance of all classifiers is evaluated using semi supervised learning. Empirical results demonstrate the competency of proposed model to these six baseline classifiers. The overall best classifiers are Gradient Boosting and eXtreme Gradient Bossting classifiers with semi supervised learning having 99.24% and 99.62% approximate accuracy respectively.**

***Keywords*:** Customer Churn, Data Mining, Machine Learning, Semi-supervised learning, Supervised Classification**.**

I. Motivation and Implications

With the excessive use of mobile phones and emerging technologies as Internet of Things (IoT) and big data in every dimension of life, telecom operators are experiencing a tremendous growth in mobile data. Telecom industry are holding remarkable place according to stored data level almost 715 PB (Petabyte) along with mass media field. Every day operators are gathering subscriber's data, which cater useful information about subscriber's background, participation, activities etc. in the form of Call Records, bill records, phone usage, server logs, network equipment detail, and social network data. Although CRM(Customer Relationship Management) strategies have essence from many decades, business world is going to shift from product centric strategies towards customer centric strategies. Customer segmentation and retention problems (predicting customer churn) are major targets for any organization.

In telecom market, competition also plays important role. As more number of competitors exists in business, mass marketing cannot play role significantly and segmentation becomes vital for target customers [1].Segmentation can be market segmentation and customer segmentation. Market segmentation is related to the division in homogeneous market containing homogeneous groups of customers having different behaviors. Customer segmentation form groups of customers having common characteristics so that organizations can approach each group effectively and appropriately. It falls into four categories namely segmentation according to customer value, customer behavior, customer life cycle and customer migration [2].

Although data mining and machine learning technologies have some level of capabilities to build effective predictive systems, advancement in functionalities of machine learning algorithms would facilitate the organizations for improving their infrastructure, customer services, customer retentions and revenues. Scope and possibilities can be increased with the application of big data analytics as expansion and heterogeneity in data sources are experienced day-to-day.

II. Introduction

Churn prediction is the indication for the customers who are about to cancel the subscription in future. It has been great indicator for improving customer service satisfaction and most CRM strategies are focusing on reducing rate of churn. According to past researches, retaining existing customer is

more beneficial to have a new customer and is an efficient marketing strategy to maximize the shareholder's value.

Generally, different classification algorithms as Random Forest Classifier, Support Vector Machine, AdaBoost Classifier, GradientBoosting Classifier, Logistic Regression etc.[3] are producing better results in telecommunication customer churn prediction field. Saran Kumar and Chandrakala(2016) produced a detailed study on the methods used for the process of customer churn prediction and reviewed the most popular machine learning algorithms used by researchers [4].

It has also observed that most of the previous work focused on finding best machine learning algorithm in terms of various parameters (Accuracy, Precision, F-Measure, and Recall) by building appropriate models for churn prediction. Researchers experimented churn prediction using either classification algorithm or clustering algorithms. Classification algorithms work on training the data by providing labels and then test the unlabelled data.

In today's scenario, it is not easy to find labeled data in sufficient amount and to train the model at low cost. In contrast, unlabelled data can be easily available and also have some important features to train the model.

Taking into consideration these concepts, we directed our work towards semi supervised machine-learning approach for telecom churn prediction. The key of proposed churn prediction model is Pseudo-Label semi supervised learning [3]. Semi supervised learning incorporates the features of supervised as well as unsupervised learning. The model is trained with a combination of small label data and large unlabelled data. The impact of proposed model is assessed in terms of external measures as precision,recall, f-score and accuracy. We also analyzed the model with cross validation accuracy measure. The analysis shows that our model is not predicting over fitted results. It has also observed that the model performed better than the previously used supervised algorithms.

The remaining paper is categorized as: Section III analyze the related work of machine learning techniques in churn prediction. Section IV explains traditional classification algorithms. Semi supervised pseudo label technique is introduced in Section V. Section VI depicts the flow of proposed model for predicting churn. Section VII shows experimental evaluation and results. Section VIII discusses conclusion and planning for future.

III. Related Work

This section contributes critical analysis of the literature with main point of concerns -customer segmentation and churn prediction. Over the last decades, for highly profitable customer segmentation and envision of churn behavior, research communities have proposed several classification algorithms as well as clustering algorithms.These techniques support industries for identifying customers who are supposed to churn in future and improve decision-making process in development of appropriate retention strategies.

A new framework has been proposed and implemented by Dahiya and Bhatia(2015) for churn prediction model. They compared Decision Tree and Logistic Regression algorithms using WEKA software and the performance is evaluated by calculating two terms as the accuracy and error rate[5].

Although several algorithms have been recommended for churn prediction, still there is room for improving performance of model. Saghir, Bibi, Bashir, Khan and Saghir (2019) evaluated existing individual and ensemble neural network based classifier in addition with utilizing bagging to improve the performance[6].

To focus on performance measures, Clemente, Giner-Bosch, Mat ás(2012) proposed the use of composite indicator for evalauting five classification models : Logistic Regression, Decision tree, Neural network, Adaboost and Random forest[7]. They considered four alternatives for input variable selection: original variables, aggregate variables , Prinicple component analysis and Stacking method. For performance measure, they considered other imporatnt parameters also with accuracy such as robustness, speed, interpretability and easy of use.

To cluster the mobile customers, Aheleroff and Gholamian(2011) focused on call period and segmented the customers into four loyal groups: plain loyal, not dependable, fence seated and loyal under incentive. A customer life cycle model was suggested by taking into consideration the past contributions, profitable values and churn prediction concurrently [8].

To deal the issue related to selection of appropriate cluster centers, Cheng,Cheng,Yuan, song, Xu, Ye and Zhang(2016) proposed a novel Multivariable Quantum Shuffled Frog Leaping Algorithm (MQSFLA)-k clustering for segmentation of telecom customers [9]. Their results showed more accurate value for convergence rate in comparison with other intelligent algorithms.

To improve the search capability and classification capability Idris, Iftikhar and Rehman (2017) combined genetic programming and Adaboost learning for generating high performance churn prediction model[10].

Ascarza,Neslin, Netzer, Anderson, Fader and Gupta(2017) identified not only churners but also drew attention for gaining observations to manage retention in conjunction with identifying areas for future research such as variety of metrics to measure and monitor retention with identification of agreements between different retention programs like reactive or proactive, remedies for retention programs and discrete campaigns and continuous process for managing retention[11].

Azeem and Usman(2018) classified churners using fuzzy classification and identified accurately according to different severity levels[12]. This model also analyzed customer usage data and complaint data to procreate intelligent retention campaigns in contrast to earlier efforts that were made on decreasing the rate of churn.

The existing prediction models in literature for churner identification cannot be applied directly on large size datasets due to their high volume and ignorance of important variables present in the data by feature selection methods. To solve the problem of large scale data, data warehouse system with data mining techniques was used but this method was not able to generate very convincing results. In addition, all the data sources producing large data were not considered due to intricacy in data.

Ahmad, Jafar, Aljoumaa(2019) handled these problems using Hadoop distributed file system[13]. Machine learning algorithms were used to build churn prediction model for telecom big data and performed feature engineering and selection with new aspects as customer social network analysis. Addition of SNA features with statistical features to classification algorithms, significant improvement can be

achieved in results. By including analysis attributes like degree of centrality measures, similarities values and network connectivity among customer's values, good improvement in AUC (Area under curve) results were achieved. The dataset used was about 70 terabytes with different file formats.

The reduction in data dimensionality also plays important role in accuracy improvement and processing time reduction of machine learning algorithms. High dimensional data has more noise and require more computation time. Though the reduction in dimensionality of data is a big assignment in the field of data analysis as reduction in number of attributes may cost in terms of meaningful information loss. One approach suggested assigning weights to attributes. Amin, Shah, Abbas, Anwar, Alfandi and Moreira (2019) assigned weights automatically by genetic algorithm to the attributes and classifications of churners are performed by Naive Bayes(NB) algorithm without involving domain experts[14].

Another approach proposed by Lin, Zhang, Wang, Xue and Liu (2019) is Spark based parallel large sum sub matrix biclustering algorithm (SP-PLSS) in which clustering is performed in two directions on customer samples as well as on attributes simultaneously and it identifies and segments eminent and beneficial telecom customers[15]. By implementing Map Reduce framework along with biclustering algorithm on Spark platform, performance of large datasets has shown improvement.

Ullah Raza, Malik, Imran, Islam and Kim(2019) proposed a prediction model based on machine learning algorithms and attribute selected classifier to find the root cause visualization for churn factors for telecom churners[16]. After classification, customer segmentation was performed using Cosine similarity coefficient to provide attractive offers for retention according to cluster formed.

Exploratory data analysis followed by feature engineering is performed by Halibas, Mathew, Pillai, Reazol, Delvo and Reazol (2019) and most appropriate classification algorithms as Na ïve Bayes, Logistic Regression algorithm, Decision Tree, random Forest classifier and Gradient boosted trees are applied for telecom churn prediction. Gradient Boosted is considered as best classifier by analyzing the classification performance measures[17].

In continuation with the study of classifiers, Vafeiadis, Diamantaras, Sarigiannidis, Chatzisavvas and Vafeiadis(2015) also compared five most widely used classifiers as Na ïve Bayes, Multilayer Artificial Neural Network, Decision Trees, Support vector Machines and Logistic Regression with their boosting version. They evaluated the suitability of classification algorithms on the problem of churning and experimented that boosting improves the classifier's performance[18].

Keramati, Jafari-Marandi, Aliannejadi, Ahmadian, Mozaffari, Abbasi(2014) compared the performance of 4 different soft computing techniques as Decision Tree, ANN(Artificial Neural Network),KNN(K-Nearest Neighbor) and SVM and found that ANN outperformed the other three. They proposed a new classifier able to decide about the final decision from four classifiers[19].

Customers personal connections and network properties among people can affect the churn prediction, Kim, Jun, Lee (2014) proposed new churn prediction procedure considering communication pattern and propagation of messages about churning. Hence, they evaluated the model based on social network and personal features both[20].

Earlier, Customer retention prediction problem was solved by mostly classification techniques only, a hybrid model of clustering and classification is suggested by Vijaya, Sivasankar and Gayathri (2019). According to authors, in first phase, clusters of churners are formed using fuzzy based clustering. In next phase, these clustered groups are splitted into training and testing data. In last phase, ensemble models are implemented for building the model[21].

Another hybrid model is suggested by Huang and Kechadi which is based on the combination of k means clustering algorithm and rule based First Order inductive learning technique(FOIL) . The advantage of this is results could be easily understood and interpreted[22].

Kisioglu and Topcu suggested that mathematical models are more appropriate than data mining technologies for simulation in case of incomplete and correlated data[23]. They used Bayesian Belief Network (BBN) to find important variables with CHAID (Chi-squared Automatic Interaction Detector) algorithm to discretised continuous variables.

In sharp, due to existence of high dimensional and variety of data, data mining techniques need improvements to develop efficient classification model on the problem of churning.

IV. Classification Algorithms

We selected six popular machine-learning classifiers to train the model:

A. Support Vector Machine

Support Vector Machine (SVM) is primarily a discriminate classifier that is based on the approach of decision planes (hyper plane) also known as support vector that can define decision boundaries in multi dimensional space.The algori thm separates labeled data sets having different class membership with an optimal hyper plane. It can perform linear and non linear classification efficiently but non linear classification with the help of kernel trick. If data is linear, it classifies the data into non-overlapping classes with the generation of linear hyper plane. In case of overlapping classes, it considers the hyper plane with maximum margin and minimum misclassification. There can be many possible hyper plane that can be chosen. The number of features affects the dimensions of hyper plane. If numbers of input features are two, hyper plane is a single straight line and for number of input feature three, hyper plane is two dimensional planes. Having large number of input feature, it becomes difficult to imagine.

To construct an optimum hyper plane, SVM exploits an iterative training algorithm by which is used to minimize the error function. In our experiment, we used linear SVC (Support Vector Classification) with parameter kernel 'linear. The optimization function is shown in (1)

$$\arg\min_{w} R(w) + c\sum_{i=1}^{N} L\left(y_i, w^T, x_i\right) \qquad (1)$$

where

$L(y_i, w^T, x_i)$ = loss function that measures the discrepancy between classifier's prediction and true output y_i for i^{th} training example

c = capacity constant

w = coefficient vectors

R(w) = regularization function that prevents parameters causing overfitting

x_i = independent variables.

B. Random Forest Classifier

Random forest is a sort of bagging ensemble classifier that comprises of various decision trees made at the training time and output class is resulted by individual tree with voting for predicted variable. Random forest selects the prediction that is having highest vote. Initially, the algorithm was based on stochastic discrimination approach that was further extended with the concept of bagging and feature selection randomly. Decision tree suffers from two problems: high variance or high bias. Random forest mitigate these problem as it is assortment of decision trees whose results are aggregated into final result to remove over fitting without increasing bias. Variance can be controlled by training on different samples of data or by selecting subset of features randomly.

Parameter Gini Importance is used to measure the importance of nodes for each decision tree by assuming decision tree as binary tree shown in (2)

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \qquad (2)$$

where
ni_j =the importance of node j
wj = weighted number of samples joining node j
Cj= the impurity value of node j
left (j) = child node from left split on node j
right (j) = child node from right split on node j

Each feature importance is calculated using (3)

$$fi_i = \frac{\sum_{j\text{node j splits on features } i} ni_j}{\sum_{k\text{òall nodes}} ni_k} \qquad (3)$$

Then the value is normalized [0, 1] using (4)

$$\text{norm } f_{i_i} = \frac{f_{i_i}}{\sum_{k\text{òall nodes}} n_{i_k}} \qquad (4)$$

Finally, the importance of feature i is calculated using (5)

$$RFFi = \frac{\sum_{j\text{òall trees}} norm\ fi_{ij}}{T\left(Total\ number\ of\ trees\right)} \qquad (5)$$

C. Logistic Regression

It is the technique acquired from statistics based on the function: logistic method that is sigmoid function. This function transforms any real value into a numeric value between 0 to1.Logistic regression equation is similar to linear regression equation. Weighted input values or input values using coefficient are combined to anticipate the output using linear transformation by logistic function. Output value produced is a numeric value instead of binary value (0/1) produced in linear regression.Logistic regression equation is as (6)

$$Y = \frac{e^{(b_0 + b_1 * X)}}{(1 + e^{(b_0 + b_1 * X)})} \qquad (6)$$

Where

Y = predicted output,

b_0 =bias or intercept term

b_1 = the coefficient for input value

D. AdaBoosting

Adaboost, the first successful boosting algorithm is invented by Freund et al., 1996, Freund and Schapire,1997. AdaBoost, ensemble classifier is also known as adaptive boosting which converts weak predictors into strong predictor sequentially but it learns from the mistakes by increasing the weight of misclassified patterns. Weight and prediction of each tree are multiplied and then values are added to build new prediction. The tree that has higher weight will have more effect in making final decision. Noisy data and outliers easily affect Adaboost. Let Ot(x) = output of weak classifier t for input x and Wt=weight assigned to classifier. Weight of weak classifier is calculated based on error rate E.

$$W_t = 0.5 * \ln\left(\frac{(1-E)}{E}\right) \qquad (7)$$

Weight of each data point is updated as (8)

$$w_{+1}(x_i, y_i) = \frac{w(x_i, y_i) \exp[-W_t\ y_i f(x_i)]}{Z} \qquad (8)$$

Where f(xi) is the weak classifier and Z is normalization factor to ensure the sum of all instance weights equal to 1.

E. Gradient Boosting

Adaboost technique has been generalized into Gradient Boosting in order to cope with a variety of loss functions by Friedman et al., 2000 and Friedman, 2001.Gradient boosting is ensemble type classification having boosting technique with sequential building of predictors, not independently as in Random forest classifier. In boosting technique, mistakes of previous predictors are considered as learning for succeeding predictors and these mistakes are identified by gradients. The concept behind gradient boosting technique is to develop a strong model by leveraging pattern from residuals and to make prediction better by using weak predictions. When no pattern in residual is left to make better prediction, process is stopped otherwise it may cause over fitting. The loss function to be minimized for Gradient Boosting algorithm is mean squared error defined as (9)

$$\boldsymbol{Loss = MSE = \sum (y_i - y_i^p)^2} \qquad (9)$$

Where y_i = i^{th} target value, y_i^p =ith prediction

By using gradient descent and updating our predictions based on learning rate, minimum value of MSE is calculated as (10).

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p \qquad (10)$$

This becomes (11)

$$y_i^P = y_i^P + \alpha * 2\sum\left(y_i - y_i^P\right)^2 \qquad (11)$$

Where α = learning rate

$$\sum\left(y_i - y_i^P\right) = sum \text{ of residuals}$$

So the predictions are updated such that the sum of residuals becomes minimum or nearly 0 and predicted values become close to actual values.

F. eXtreme Gradient Boosting(XGB)

XGB, developed by Tianqi Chen[24], is extension of gradient boosting, designed for improvement in speed and performance. Tianqi Chen developed scalable learning system for tree boosting. Implementation of several important sytems and by optimizing algorithms, scalibility has improved.

The features of XGB are:

(i) Sparse data handling using a new tree learning algorithm
(ii) Instance weights management by a weighted quantile sketch method
(iii)Performance improvement by faster learning using Parallel and distributed computing.
(iv) Out-of-core tree learning having adequate cache-aware block structure .

It gives faster performance than gradient boosting due to parallel processing process. To avoid the risk of overfitting, XGB includes modification in regularized learning objective function.
The following regularized objective is minimized.

$$\mathrm{L}\left(\phi\right) = \sum_i l(y_i^p, y_i) + \sum_k \Omega(f_k) \qquad (12)$$

$$\text{Where } \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \qquad (13)$$

Here l is a differentiable convex loss function to calculate the difference between prediction y_i^p and target y_i . This supplementry term Ω increase the model complexity but helps in smoothing the final weights learning procedure to reduce over-fitting.

V. Semi supervised Pseudo-Label technique

Classification technique is supervised learning which depends heavily on labeled data to achieve efficient classifications. But in this rapid changing world of big data, it is very difficult to find sufficient labeled data at low cost; in contrary of this, unlabeled data are in abundant. Semi supervised learning not only take advantage of labeled data, but also uses unlabeled data in large amount to create better learners and improve classification performance.

In this paper, the effectiveness of pseudo-label semi supervised technique will be measured to improve the accuracy of supervised classifiers. Lee [3] has proposed pseudo-label technique for neural network in 2013.According to Lee, Pseudo-labels are the predicted class of unlabeled data that has high probability and they are considered as true labels. It is similar to entropy regularization that means to take assistance from unlabeled data in the framework a posterior estimation. This scheme considers low-density separation between classes without any modeling of the density and by minimizing the conditional entropy of class probabilities for unlabeled data. The general assumptions behind Pseudo-label technique is that data points in high-density region belong to same classes and decision boundary lies in low-density region.

In our approach, the training procedure has access to a set of n labeled data(X_i, Y_i) (1⩽i⩽n) as well as to large number of unlimited data where we consider all predicted classes as pseudo-label classes. Figure 1 shows the concept of pseudo label semi supervised learning. Model is trained with labeled data to generate initial training model. This model generates pseudo-labels for test data (unlabeled data). Labeled data and pseudo-labeled data are combined and trained again on combined dataset and tested on validation dataset. By this technique, there may be change in position of initial decision boundary which will improve the performance of classification.

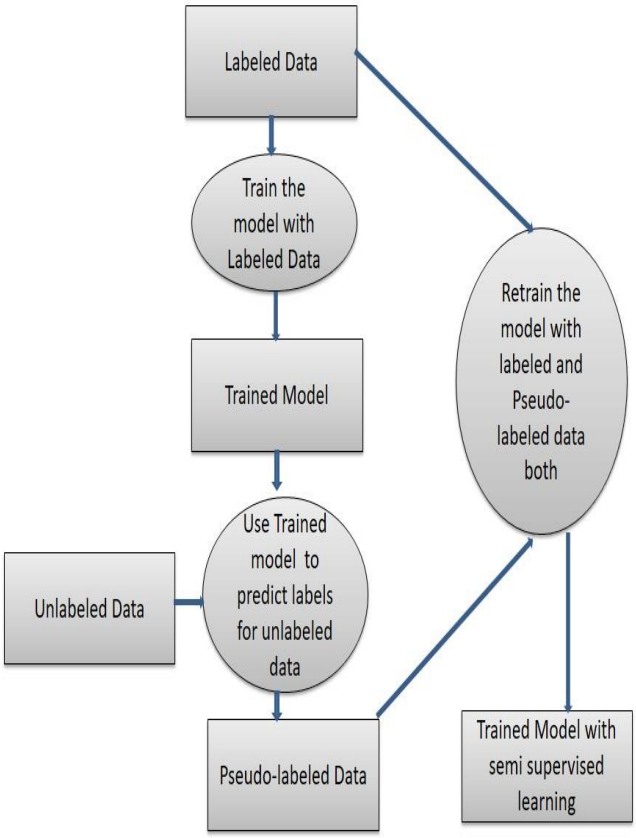

Figure 1. Classification using Psuedo-Label Technique

VI. Proposed model: Semi Supervised Learning Churn Prediction Model (SSL-CPM)

This section implies a churn prediction model using semi supervised learning in telecommunication data. The proposed model is trained using semi-supervised learning using labeled and unlabeled data both. Figure. 2 shows the representation of proposed model.

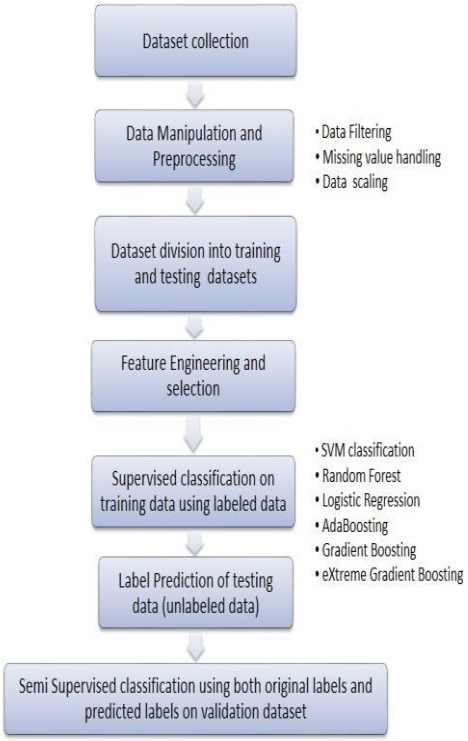

Figure 2. Flow Diagram for Proposed Churn prediction Model

In this study, we considered a combined dataset from two publically available datasets by considering the most valuable information from both datasets for churn prediction. Data set is divided into two parts with small amount of training data and large amount of testing data as unlabeled data. Classification based on training data utilizes known class labels. Model is then applied on testing phase to predict labels known as Pseudo-label data. Both labeled, Pseudo-label data are combined, and model is retrained again. Finally, we get a model with Pseudo-labeled semi supervised classification. Classification performance has improved with this technique.

VII. Experiment

A. Dataset

In telecom industries, various types of data available are Customer data as customer's service and contract information data, towers and complaints database coverage inquiries, complaint received etc, network log data as internal sessions for internet, calls and SMS, call detail records as all charging information about calls, SMSs, MMSs etc., Mobile IMEI information (brand, model, type of mobile phone). Since telecom data is very confidential, many researchers have suggested to use online available data or to use synthetic data. For churn prediction model, we found five common types of customer data available:

1) *Demographic data:* The demographic data contains gender, age, range, partner, dependent status of the customer.

2) *Customer Account Information:* Customer account information contains Tenure and contract of the customer.

3) *Services used by customer:* It contains following fields as Phone Service, Multiple Lines, Online Security service etc.

4) *Billing Information:* Monthly and total charges will come under this category.

5) *Customer who left means Predictor data:* Churn is the variable to be predicted so there is need to understand its interaction with other important variables.

However, online available single dataset does not have all the fields that can have important information; hence, in order to fulfill our requirement, we joined important features from two most popular telecom churn predictor dataset and used it for our model. Joined dataset contains 2666 customer data with 40 features. Validation process has performed on another validation dataset with 667 customer data with 40 features.

B. Data Manipulation and Preprocessing

Datasets can have number of missing values which require effective manipulations to get better result. Similarly data set is analyzed for preprocessing. Machine learning algorithms can be applied only on numerical features but our dataset has numerical and categorical features both, hence some standard techniques are used to convert all categorical features into numerical representations for further processing. Two common techniques are Label encoding and One-Hot encoding. In our data set, we used label encoding scheme for this purpose because One-Hot encoding technique increase the dimensionality of data.

C. Feature Selections

Data exploration and feature selections are needed to find the interesting trends in data. It visualizes correlations of variables to test hypothesis and to check assumptions with the help of statistics and graphical representation. Feature selection is also used for dimensionality reduction. Random Forests Regressor is best known for feature selection in data science. It works better because of tree based strategies. It improves the purity of the node and decrease impurity over the tree. Thus nodes with least decrease in impurity occur at the end of tree and a subset of most important features is created by pruning trees below a particular node. The selection of significant features can improve performance of proposed algorithm. We implemented Random Forest Regression method [25] to find the important features in dataset as shown in Figure 3.

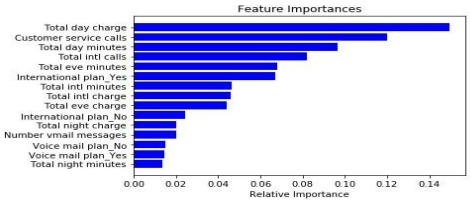

Figure 3. Feature Selection using Random Forest Regressor

D. Performance Evaluation Matrix

Validation of classification algorithms is as important as applying the best classifier. The proposed model is assessed using some external indices as accuracy, recall, precision, F1-score and cross validation score.

Notations used:

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

1) *Confusion Matrix for classification*

		Predicted Class	
		Churners	Non- Churners
Actual Class	Churners	TP	FN
	Non Churners	FP	TN

2) *Accuracy*

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (14)$$

3) *Precision*

$$\Pr ecision = \frac{(TP)}{(TP+FP)} \qquad (15)$$

4) *Recall*

$$\mathrm{Re}call = \frac{(TP)}{(TP+FN)} \qquad (16)$$

5) *F1-Score*

$$F1-score = 2*\frac{(\Pr ecision*\mathrm{Re}call)}{(\Pr ecision+\mathrm{Re}call)} \qquad (17)$$

6) *Cross validation Score*

Cross validation is resampling method used to evaluate the machine learning model on a limited data sample. We used k-fold cross validation technique to check the condition of over fitting. Over fitting is the condition when machine learning algorithms capture the noise of data and results in good accuracy for training data but poor results on new data set. In cross validation technique, data is partitioned into number of subsets, one set is hold out as test data and model is trained on remaining data sets; after training, model is tested on test partition. This process is repeated for each subsets of data set. K parameter refers to a group that a given data sample is to be split into. To validate the results, cross validation score function has been used and accuracy with precision is analyzed for proposed model by choosing the value of k from 5 to 10. The accuracy of the model is the average of accuracy for each fold.

E. Result and Analysis

In previous section, we introduced 6 different classification algorithms. In order to evaluate and compare these, the whole data has been separated into train and test sets. Train set and test set comprises 30% and 70% of data respectively. Experiment has conducted into three steps. In first step, classification algorithms are applied to train the model and evaluated by cross validation function for the value of k from 5 to 10. Results are measured and analyzed in terms of cross validation scores, Precision, Recall, F-measures and accuracy. In second step, trained model predicts the output of test set. The predicted output and original labels are combined with train and test set features. In third step, model is retrained again using pseudo label learning and evaluated on validation data set with cross validation function for the value of k from 5 to 10. Results are measured and analyzed again on same parameters as above.

1) Classifiers performance without pseudo-label learning

Cross validation scores of all six classifiers are depicted in table 1 to 6. Our simulation results showed that SVM classifier scores lowest values. AdaBoosting, Random Forest and Logistic Regression classifiers are showing similar scores and Gradient Boosting and eXtreme Gradient Boosting classifiers score highest values. SVM scores highest value at K=7, Random Forest at K=9, Logistic Regression at K=8, AdaBoosting at K=7, Gradient Boosting at K=10 and eXtreme Gradient Boosting at K=6. Performance of these classifiers are also measured on the basis of Precision, Recall, F1-score and Accuracy as depicted in table 7 .

K	Average	Variance
5	0.76926	0.35028
6	0.71077	0.37019
7	**0.72717**	**0.40605**
8	0.78766	0.18877
9	0.73606	0.39650
10	0.76477	0.21811

Table 1.Cross Validation scores of SVM

K	**Average**	**Variance**
5	0.87764	0.03493
6	0.88700	0.00197
7	0.87005	0.06204
8	0.87764	0.05642
9	**0.87791**	**0.06877**
10	0.88342	0.03856

Table 2. Cross Validation scores of Random Forest

K	Average	Variance
5	0.85309	0.01010
6	0.84184	0.04631
7	0.84744	0.01361
8	**0.85882**	**0.04126**
9	0.84062	0.03106
10	0.85200	0.02104

Table 3. Cross Validation scores of Logistic Regression

K	Average	Variance
5	0.86448	0.04770
6	0.84195	0.06501
7	**0.85125**	**0.08734**
8	0.85130	0.04966
9	0.85339	0.04979
10	0.84970	0.07698

Table 4. Cross Validation scores of AdaBoosting

K	Average	Variance
5	0.90404	0.03132
6	0.90587	0.04769
7	0.90772	0.03447
8	0.90600	0.05484
9	0.90240	0.04438
10	**0.90428**	**0.05976**

Table 5. Cross Validation scores of Gradient Boosting

K	Average	Variance
5	0.89834	0.02428
6	**0.91345**	**0.04584**
7	0.89266	0.06037
8	0.90403	0.05371
9	0.89077	0.06003
10	0.89462	0.04730

Table 6. Cross Validation scores of eXtreme Gradient Boosting(XGB)

Classifiers	Precision	Recall	F1-score	Accuracy
SVM	0.73	0.85	0.79	85.50
Random Forest	0.97	0.97	0.97	97.09
Logistic Regression	0.83	0.86	0.83	85.88
AdaBoosting	0.91	0.90	0.91	91.49
Gradient Boosting	0.99	0.99	0.99	99.06
XGB	0.99	0.99	0.99	99.10

Table 7. Precision, Recall, F1-score and Accuracy of SVM, Random Forest, Logistic Regression, AdaBoosting Gradient Boosting and XGB Classifiers without semi supervised learning

2) Classifiers performance with pseudo-label Semi Supervised learning

Next, we apply semi supervised learning technique to explore the impact of pseudo labels on the performance of the classifiers. Cross validation scores of all six classifiers with semi supervised learning are shown in table 8-13. Table 14 shows the external performance measures of six classifiers with semi supervised learning. Simulation results show significant improvements in cross validation scores and external performance meaures. Gradient Boosting and eXtreme Gradient Boosting performance is again highest at K=10 and K=6 respectively.

For SVM, cross validation score has high improvement at K=8 and accuracy has increased from 85.50% to 96.94 % .In Random Forest case, there is again high improvement at K= 8 and accuracy raised from 97.09% to 97.44% after applying semi supervised learning technique. Logistic Regression achieves cross validation score improvement at K= 8 and accuracy improvement is from 85.88% to 96.38%. AdaBoosting and Gradient Boosting cross validation scores with semi supervised learning is high at K=10. Accuracy improvement in case of AdaBoosting is from 91.49% to 96.87% and for Gradient Boosting is 99.06% to 99.24%. Using eXtreme Gradient Boosting with semi supervised learning, accuracy improvement is achieved from 99.10% to 99.62% at K=6.

K	Average	Variance
5	0.74878	0.47657
6	0.75385	0.32875
7	0.85683	0.24099
8	**0.83145**	**0.42854**
9	0.81426	0.41540
10	0.87388	0.24603

Table 8. Cross Validation scores of SVM with Semi Supervised Learning

K	Average	Variance
5	0.95933	0.01217
6	0.94878	0.01838
7	0.95292	0.01673
8	**0.95105**	**0.02522**
9	0.94803	0.00935
10	0.95179	0.01148

Table 9. Cross Validation scores of Random Forest with Semi Supervised Learning

K	Average	Variance
5	0.95933	0.01217
6	0.94878	0.01838
7	0.95292	0.01673
8	**0.95105**	**0.02522**
9	0.94100	0.01520
10	0.94605	0.02052

Table 10. Cross Validation scores of Logistic Regression with Semi Supervised Learning

K	Average	Variance
5	0.95220	0.05324
6	0.95369	0.04620
7	0.95559	0.05080
8	0.95900	0.04464
9	0.95561	0.04939
10	**0.95746**	**0.05563**

Table 11. Cross Validation scores of AdaBoosting with Semi Supervised Learning

K	Average	Variance
5	0.95971	0.01684
6	0.96008	0.01995
7	0.95971	0.02041
8	0.96271	0.02635
9	0.96196	0.02650
10	**0.96275**	**0.02922**

Table 12. Cross Validation scores of Gradient Boosting with Semi Supervised Learning

K	Average	Variance
5	0.92047	0.04048
6	0.91592	0.06081
7	0.90545	0.06329
8	0.91453	0.05853
9	0.91742	0.03666
10	0.91893	0.05231

Table 13. Cross Validation scores of XGB with Semi Supervised Learning

Classifiers	Precision	Recall	F1-score	Accuracy
SVM	0.94	0.97	0.96	96.94
Random Forest	0.97	0.97	0.97	97.44
Logistic Regression	0.96	0.96	0.96	96.38
AdaBoosting	0.97	0.97	0.97	96.87
Gradient Boosting	0.99	0.99	0.99	99.24
XGB	1.00	0.99	0.99	99.62

Table 14. Precision, Recall, F-score, Accuracy of SVM, Random Forest, Logistic Regression, AdaBoosting, Gradient Boosting and eXtreme Gradient Boosting Classifiers with semi supervised learning

Results are analyzed to compare the performance of various classification algorithms in the telecommunication sector . We got considerable improvements in performance measures and accuracy for all the classifiers. Experiments suggest the fact that the implication of pseudo-label semi supervised learning can significantly improve the classification performance. As result, Gradient Boosting and eXtreme Gradient Boosting can be best algorithms as compared to other algorithms for solving the problem of churn prediction in the field of telecommunication.

VIII. Conclusion and Future Scope

This work has placed emphasis on the performance of widely used machine learning classification techniques for churn prediction problem and promoted the advantage of pseudo-label semi supervised learning technique. Six most popular and traditional classification methods were experimented for predicting churn in telecommunication field on publically accessible datasets.

Initially all the methods were experimented without semi supervised learning under distinct parameter settings. Results

revealed that supervised Gradient Boosting technique shows highest accuracy 99.06% and precision 99%, recall 99% and F1-score 99% approximately with k fold cross validation at k=10 and supervised eXtreme gradient Boosting outperforms with k fold cross validation at k=6 achieving highest accuracy 99.10% and precision 99%, recall 99% and F1-score 99% approximately. The Support Vector Machine and Logistic Regression classifiers find lacking with approximate accuracy 85.50% and 85.88 % respectively.

Afterwards, we measured the impact of pseudo-label semi supervised learning to all the corresponding classifiers. Using semi supervised learning approach, classifiers achieved remarkable improvements in performance measures.

Comprehensively, it can be concluded that classification algorithms with semi supervised learning performed better than supervised classification algorithms. Among all the classifications, the best classifiers are Gradient Boosting and eXtreme Gradient Boosting classifires with semi supervised learning having 99.24% and 99.62% accuracy respectively.

In future aspects, the proposed model can be integrated with big data techniques to find more insights from high volume varied dataset of telecom industry. Clustering techniques can also be applied to maximize the significance of results.

`References

[1] A. Namvar, M. Ghazanfari and M. Naderpour, "A customer segmentation framework for targeted marketing in telecommunication", *Proc. IEEE International Conference on Intelligent system and Knowledge Engineering*, pp.1-6, 2017 10.1109/ISKE.2017.8258803.

[2] J.Bayer,"Customer Segmentation in the Telecommunications Industry", *Journal of Database Marketing & Customer Strategy Management*, Vol.17 (3/4) pp. 247–56, 2010.

[3] D.-H. Lee, "Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks", *ICML Workshop*, 2013.

[4] Saran. Kumar. A. and Chandrakala. D., "A Survey on Customer Churn Prediction using Machine Learning Techniques", *International Journal of Computer Applications*, pp.13-16, 2016 DOI: 10.5120/ijca2016912237.

[5] K. Dahiya and S. Bhatia, "Customer Churn Analysis in Telecom Industry", *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO*), pp. 1-6, 2015.

[6] M.Saghir, Z.Bibi, S.Bashir and F.H.Khan, "Churn Prediction using Neural Network based Individual and Ensemble Models" , *Proc. IEEE International Bhurban Conference on Applied Science and technology*, pp. 634-639, 2019.

[7] M. Clemente, V. Giner-Bosch and S. San Mat ás, "Assessing classification methods for churn prediction by composite indicators",2012.

[8] S. Aheleroff and M. R. Gholamian,"Customer Segmentation for a Mobile Telecommunications Company Based on Service Usage Behavior", *Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, pp. 308-313,2011.

[9] C.Cheng, X.Cheng, M.Yuan, C.song, L.Xu, H.Ye and T.Zhang, "A novel cluster algorithm for telecom customer segmentation", *Proc.IEEE International Symposium on Communications & Information Technologies*, pp. 231–237, 2016.

[10] A. Idris, A. Iftikhar and Z.u. Rehman, "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO under sampling", *Springer Science +Business media*, LLC,2017.

[11] E.Ascarza, S.A.Neslin, O.Netzer, Z. Anderson, P.S.Fader and S. Gupta, "In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions", *Springer Science Business media*LLC,2017, https://doi.org/10.1007/s40547-017-0080-0.

[12] M. Azeem and M. Usman, "A fuzzy based churn prediction and retention model for prepaid customers in telecom industry",*International Journal of Computational Intelligence Systems*, Vol.11, pp. 66-78, 2018.

[13] A.K. Ahmad, A. Jafar and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform", *Journal of Big Data*, https://doi.org/10.1186/s40537-019-0191-6, 2019.

[14] A Amin, B.Shah, A. Abbas, S. Anwar, O. Alfandi and F. Moreira, "Features Weight Estimation Using a Genetic Algorithm for Customer Churn Prediction in the Telecom Sector", *Springer Nature*, Switzerland,pp. 483-491, 2019.

[15] Q.Lin, H.Zhang, X. Wang,Y. Xue and H. Liu, "A novel Parallel Biclutsreing Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers, Special section on trends", *perspective and prospects of machine learning*, IEEE access,2019.

[16] I.Ullah, B. Raza, A.K. Malik, M.Imran, S.U. Islam and S.W.Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", *IEEE Access* 2019.

[17] A.S.Halibas,A.C.Mathew,I.G.Pillai,J.H.Reazol,E.G.Del voa, and L.B.Reazol,"Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling", IEEE Explore, *International Conference on Big data and smart city*,2019.

[18] T.Vafeiadis, K.I.Diamantaras, G.Sarigiannidis and K.Ch.Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", *Journal of Simulation Modelling Practice and Theory*, 2015 .

[19] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari and U. Abbasi, "Improved churn prediction in telecommunication industry using datamining techniques", *Journal of Applied Soft Computing*, 2014.

[20] K. Kim, C.Juna and J.Lee,"Improved churn Prediction in telecommunication industry by analyzing a large network", *Journal of Expert Systems with Applications*, 2014.

[21] J.Vijaya,E.Sivasankar and S.Gayathri, Fuzzy Clustering with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector", *Springer Nature Singapore* , 2019.

[22] Y. Huang and T. Kechadi,"An effiective hybrid learning system for telecommunication churn prediction", *Journal of Expert Systems with Applications*, 2013.
[23] P.Kisioglu and Y.L.Topcu, "Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey", *Journal of Expert Systems with Applications*, 2011.
[24] J. H. Friedman, "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, pp. 1189 – 1232, 2001.
[25] H. Kawakubo and H. Yoshida, "Rapid Feature Selection Based on Random Forests for High-Dimensional Data", *Information processing society of Japan*, SIG Technical Report, 2012.

Author Biographies

Bindu Rani is a Ph.D. scholar from Department of Computer Science and Engineering , Sharda University, Greater Noida, India and works as assistant professor in Information Technology Department in Inderprastha Engineering College, Ghaziabad, Dr. A.P.J Abdul Kalam Technical University, India. She received Master in Computer Science and Application degree from Aligarh Muslim Univerity(AMU), India. Her research interests are Data Mining, Big Data and Machine learning techniques.

Dr. Shri Kant has received his Ph. D. in applied mathematics from applied mathematics departments of institute of technology, Banaras Hindu University (BHU), Varanasi in 1981 He is working as a Professor at Research and Technology Development Centre (RTDC), Deptt. of Computer Science and Engineering of Sharda University, India and involved actively in teaching and research mainly in the area of cyber security and Machine learning.His areas of interest are Special Functions, Cryptology, Pattern Recognition, Cluster Analysis, Soft Computing Model, Machine Learning and Data Mining.