Predicting Churn Customer in Telecom using Peergrading Regression Learning Technique

M.HemaLatha, S.Mahalakshmi

***Abstract:** Customer churn is an important issue which is faced by telecom industries daily. It is an essential concern for enterprises. Most of the telecom companies suffer a lot for voluntary churn. Here the churn rate is a significant impact for the industries on the lifetime value, and it affects the length of the service and also the revenue of the company. Because of the direct effect of the company revenue, especially in the telecom field, the companies are requesting to implement the development of predicting the potential customer to churn. Here the churn is an essential factor. Hence, analyzing the churn factor of increasing customer churn is as necessary as it is needed to reduce the churn. Recently telecom service facing the churn problem, for this analysis this research article focused on prediction of customer churn. In this research work, the main contribution is to develop a prediction model of churn, that supports the telecom operators to predict the customer, who is going to be churn. Here the model used is the Machine learning technique with the big data analysis. Our methodology used here is Peer Grading Regression (PGR). To provide a prediction of churn customer in telecom industries. To validate the performance of this proposed model, the Area under Curve (AUC) is used, and it is one of the standardized approaches. This model will be tested under some benchmark dataset; working on these vast datasets creates a transformation of raw data by telecom companies. The dataset values are provided for testing and training in the ensemble classifier. For experimentation, four algorithms are considered here: Peer Grading Regression (PGR), Random Forest, Decision Tree, and k-NN classifier. Moreover, the best outcomes are attained with the use of the boosting algorithm. Here this algorithm is applied for the classification of the churn prediction model.*

***Keywords:** Machine Learning, Telecommunication, Customer Churn, Area Under Curve and Prediction*

I. INTRODUCTION

Today numerous telecom companies are prompt all over the world. The telecommunications sector has become one of the main industries in developed countries. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers [1]. Companies are working hard to survive in this competitive market depending on multiple strategies.

In this telecommunication, churn is a major issue where Churn is the activity of the telecommunication industry is the customers leaving the current company and moving to another telecom company. The technical progress and the increasing number of operators raised the level of competition. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up with the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning [2]. In the telecmmunication industry each company provides the customers with huge incentives to attract them to switch to their services, it is one of the reasons that customer churn is a big problem in the industry nowadays. Churn management is very important for reducing churns as acquiring a new customer is more expensive than retaining the existing ones [3]. To prevent this, the company should know the reasons for which the customer decides to move on to another telecom company. Prediction is the only process to analyze the problem of churn activity. Customers churn prediction in telecom is generally a challenging task for the large feature space and imbalanced nature of the dataset. Churn prediction is used to recognize customers who are most probable to churn. Churn prediction and analysis can help a company to develop a sustainable strategy for customer retention programs [4].

Churn prediction and management have become of great concern to the mobile operators. Mobile operators wish to retain their subscribers and satisfy their needs. Hence, they need to predict the possible churners and then utilize the limited resources to retain those customers. The basic layer for predicting future customer churn is data from the past.

We look at data from customers that already have churned (response) and their characteristics / behavior (predictors) before the churn happened. By fitting a statistical model that relates the predictors to the response, we will try to predict the response for existing customers. This method belongs to the supervised learning category, just in case you needed one more buzzing expression [5]. The minority class comprises of fewer instances in the whole dataset that leads to biased training of classifiers due to dominating presence of majority class instances. The telecom companies also acquire lots of information about customers including billing, payments, call records, demographics, etc. which turns large number of features to account for in the dataset. Consequently, the predictors suffer from the curse of dimensionality and imbalanced distribution of the telecom dataset for predicting churners. Thus, a churn prediction method is highly desirable that can affectively mitigate the imbalanced nature and high dimensionality issues present in the telecom dataset [6].

Churn occurs mainly due to customer dissatisfaction. Identifying customer dissatisfaction requires several parameters.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author

Dr. M. Hema Latha*, Professor, Ramakrishna College of Arts & Science Coimbatore
S.Mahalakshmi, Research Scholar, Bharathiar University,Coimbatore.

A customer usually does not churn due to a single dissatisfaction scenario [7].

There usually exist several dissatisfaction cases before a customer completely ceases to do transactions with an organization. Several properties associated with the customer and their mode of operations with the organization are recorded by the organizations. This represents the customer's behavior data. Analyzing this data would present a clear view of the customer's current status [8]. Hence this can be used as the base data for churn prediction. The major difficulty arising from this mode of operation is that the data under discussion tends to be very huge. The hugeness can be attributed to the behavioral nature of the data, depicting all the product lines dealt with by the organization. Further, due to the requirement of structural representation of the data, all the instances are bound to contain all the properties corresponding to a generic customer in the organization [9]. This leads to data sparseness, since customers will be associated with only a few properties and not all the properties pertaining to the organization. The hugeness of data and sparsity acts as the major difficulties in the process of churn prediction.

Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data. Machine learning is subset of artificial intelligence which enables computers to learn (i.e. improvise) from data without any intervention. Machine learning is often coined with the terms pattern recognition and computational learning theory. Machine learning involves constructing algorithms that can learn from data available and can be used to make predictions on data. Here a model is built from the given input data which is then used to make predictions on new data. As the data being collected is drastically increasing each day, this calls for the need of machine learning. In machine learning, the learning process is extracting knowledge from the given data thus there is no need for a human to specify any kind of knowledge [10]. Because of the hierarchy of concepts, complex concepts can be learned by building them from simpler ones thus reducing the complexity. Prediction of churn customer is analyzed and implemented using machine learning approach. Large companies interact with their customers to provide a variety of services to them [11]. Customer service is one of the key differentiators for companies. The ability to predict if a customer will leave in order to intervene at the right time can be essential for pre-empting problems and providing high level
of customer service. The problem becomes more complex as customer behavior data is sequential and can be very diverse. Churn is an unavoidable process in any industry. However, though difficult, it is possible to identify the causes of churn using Machine learning approaches.

Having a good churn prediction model becomes extremely useful in order to minimize the churn rate because tailored promotions can be offered to specific customers that are not satisfied. Thus, there is no formal notification from the customer of ending a contract term. Our goal is to predict the customer churn. The sooner these changing patterns are detected the more opportunities and time the company will have to retain the customer. For this accuracy Area Under Curve (AUC) is used in classification analysis in order to determine which of the used models predicts the classes best. Although machine learning algorithms are usually designed to improve accuracy by reducing error, not all of them take into account the class balance, and that may give bad results. In general, classes are considered to be balanced in order to be given the same importance in training.

II. RELATED WORK

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning that were discussed in this section.

Adnan Amin et al 2015 [12] study the problem of Customer churn in telecom industry and understand that predicting the customer behavior and to preserve those customers that will churn or conceivable may churn. Author's another attempt is to make use of rough set a rule based decision making technique for mining the rules for predicting the customer. For this approach author used four different algorithms (Exhaustive, Genetic, Covering, and LEM2). From this 4 algorithm genetic algorithm produces most suitable performance. Dataset was used here is publicly available dataset. This approach fully predicts the customers who will churn or possibly may churn.

Hanif, E., 2019 [13] presents a predictive analytics approach to improve customer churn in the telecom industry for "cross-selling" or "market basket analysis". Here 4 algorithms were used K-Nearest Neighbor, Decision Tree, Naïve Bayes and Random Forest and predict customer churn in Rapid Miner. Datasets was used here is publically available datasets. From their implementation Decision Tree and Random Forest are the two algorithm which predict most accurate. Overall, the key drivers of churn are identified in this study and useful associations between products are established. This information can be used by companies to create personalized offers and campaigns for customers who are at risk of churning.

Abdelrahim Kasem Ahmad 2019 [14] is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. The model developed in this work uses machine learning techniques on big data platform. For experimenting their performance Area Under Curve (AUC) standard measure is adopted and obtained 93.3%. then another contribution is extracting Social Networ
Analysis. This model used four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST". However, the best results were obtained by applying XGBOOST algorithm. This algorithm was used for classification in this churn predictive model.

Vafeiadis, T., et al 2015 [15], present a comparative study on the most popular machine learning methods for the problem of customer churning prediction in telecom. They performed a series of Monte Carlo simulations to determine the most efficient parameter combinations. They analyze the classifier algorithm by using SVM-POLY in AdaBoost with accuracy of almost 97% and F-measure over 84%.

Amin, A., Khan, C., Ali, I. and Anwar, S., 2014 [16], to formalize customer churn prediction where rough set theory is used as one-class classifier and multi-class classifier to investigate the trade-off in the selection of an effective classification model for customer churn prediction.

Experiments were performed to explore the performance of four different rule generation algorithms (i.e. exhaustive, genetic, covering and LEM2).

It is observed that rough set as one-class classifier and multi-class classifier based on genetic algorithm yields more suitable performance as compared to the other three rule generation algorithms.

Bi, W., Cai, M., Liu, M. and Li, G., 2016 [17], find he problem of dealing with big data in the industry, existing churn prediction models cannot work very well and decision makers are always faced with imprecise operations management. So proposed clustering algorithm called semantic-driven subtractive clustering method (SDSCM). Author results indicate that SDSCM has stronger clustering semantic strength than subtractive clustering method (SCM) and fuzzy c-means (FCM). Then, a parallel SDSCM algorithm is implemented through a Hadoop MapReduce framework. From their approach, the proposed algorithm enjoys a fast running speed when compared with the other methods.

Lu, N., Lin, H., Lu, J. and Zhang, G., 2012 [18] in their research work conducts a real-world study on customer churn prediction and proposes the use of boosting to enhance a customer churn prediction model. Most of the research work that uses boosting as a method to boost the accuracy of a given basis learner. Lu, N., Lin, H., Lu, J. and Zhang, G., tries to separate customers into two clusters based on the weight assigned by the boosting algorithm. From this a higher risk customer cluster has been identified. Logistic regression is used as a basis learner in this research, and a churn prediction model is built on each cluster, respectively. The result is compared with a single logistic regression model. Experimental evaluation reveals that boosting also provides a good separation of churn data; thus, boosting is suggested for churn prediction analysis.

Prashanth, R., Deepak, K. and Meher, A.K., 2017 [19], find the Churn prediction is an important factor to consider for Customer Relationship Management. This paper statistical and data mining technique were used for churn prediction.

Prashanth, R.,et al use linear (logistic regression) and non-linear techniques of Random Forest and Deep Learning architectures including Deep Neural Network, Deep Belief Networks and Recurrent Neural Networks for prediction. This is the first time that a comparative study of conventional machine learning methods with deep learning techniques have been carried out for churn prediction. It is observed that non-linear models performed the best. Such predictive models have the potential to be used in the telecom industry for making better decisions and customer management.

Huang, B., Kechadi, M.T. and Buckley, B., 2012 [20] presents a new set of features for land-line customer churn prediction, including 2 six-month Henley segmentation, precise 4-month call details, line information, bill and payment information, account information, demographic profiles, service orders, complain information, etc. Then the seven prediction techniques (Logistic Regressions, Linear Classifications, Naive Bayes, Decision Trees, Multilayer Perceptron Neural Networks, Support Vector Machines and the Evolutionary Data Mining Algorithm) are applied in customer churn as predictors, based on the new features. Finally, the comparative experiments were carried out to evaluate the new feature set and the seven modelling techniques for customer churn prediction.

Saran Kumar, A. and Chandrakala, D., 2016 [21] reviews the most popular machine learning algorithms used by researchers for churn predicting, not only in banking sector but also other sectors which highly depends on customer participation. Author analyze suggested two Genetic Algorithm(GA) based neural network (NN) models to predict the customer churn, improved balance random forest (IBFR) model which is a combination of balanced random forests and weighted random forests in order to overcome the data distribution problems, generalized additive models and many other researchers work, hence many approaches of churn prediction models has low accuracy and prediction. Later a good prediction model is required in order to avoid the customer churn problem. This can be achieved by combining SVM with boosting algorithms for higher accuracy and performance which can be considered as a future work for Churn prediction.

Gavril et al. 2016 [22] presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. The author applied principal component analysis algorithm “PCA” to reduce data dimensions. Three machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset used in this study is small and no missing values existed.

Huang F, Zhu M, Yuan K, Deng EO. 2015 [23] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China’s largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

Burez, J. and Van den Poel, D., 2009 [24] studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. the result showed that under sampling technique outperformed the other tested techniques.

Awang, M.K., Rahman, M.N.A. and Ismail, M.R., 2012 [25] A regression based churn prediction model was presented. This method identifies churn by using multiple regressions analysis. This technique utilizes the customer’s feature data for analysis and proposes to provide good performance. This technique differs from the other proposed techniques by the fact that most of the techniques are only able to identify the customers who will instantaneously churn.

Zhu, B., Xiao, J. and He, C., 2014 [26]. Class imbalance plays a major role in affecting the reliability of a classifier. The major issue existing due to class imbalance is that the minority class is not well represented and hence the classifier is undertrained on the minority classes. The technique proposed by Zhu et al. proposes to eliminate this issue by using transfer learning techniques.

The approach presented operates by training the classifier using customer related behavioral data obtained from related domains. This approach has its major focus on the banking industry and the results are proposed to exhibit enhanced performance.

III. PROPOSED METHODOLOGY

Customer churn is the action of the customer who is like to leave the company and it is one of the mounting issues of today's rapidly growing and competitive telecommunication industry. To minimize the customer churn, prediction activity to be an important part of the telecommunication industry's vital decision making and strategic planning process. Customers' churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase. The main objective of this research is to produce a predictive model with better results that assess customer churn rate of telecommunication companies. The model developed in this work uses machine learning techniques on customer churn data. Proposed approach of peer grading regression method where used to predict the customer churn. Peer grading regression is the process of customer churn reviewing work which is scientifically and empirically investigation and testing of prediction model which encompassing relationships between different constructs such as important churn related variables, switching reasons, service usage, costs and behavior. This is based on the method of X ⊆C and B is an equivalence relation in information system IS= (C, B). X belongs to customer; C indicates the customer of the specific telecom industry. Here B specifies the balancing customer, which shows the will be churn or not churn.

A. Dataset

The data was processed to convert it from its raw status into features to be used in machine learning algorithms. This process took the longest time due to the huge numbers of columns. Firms in telecommunication sector have detailed call records. These firms can segment their customers by using call records for developing price and promotion strategies from that complaint call also registered, by analyzing the customer complaints and their customer care executive's answers rating we can easily identify the customer who will be going to churn. Table 1 and Table 2 shows some of the data from 2018- 2019, which is used for processing this implementation from cell 2 cell telecom industries the link where referred in [27].

B.Data Filter and Noise Removal

In the next step, we remove the noisy data from the sample taken in our last step. Noise hinders in most of the data analysis, so removing noise is most important goal of data cleaning. Noisy data can lead to inappropriate and erroneous results, so removing such data before training a classifier is very important step [28]. Duplicate or missing data may also cause incorrect or even misleading results. In telecom dataset, generally, there exist a lot of garbage, missing, inappropriate and incomplete values like spaces, 'Null' or any other special characters that need to be removed or replaced with appropriate values. For instance, most of the CDR data columns have "Null" values and those could be replacing with "0" for efficient processing. In order to remove inconsistencies, outliers are identified and removed from sample at this stage.

C.Feature Selection

Once noisy data is removed, the next step is to select the important features from the sample taken for processing. Here, we suggest that firstly the features should be selected on the basis of domain knowledge [29, 30]. For instance, there are some common churn indicators such as spend rate decline, outbound call decrease, increase on days with zero balance, decrease in recharge frequency, balance burn out rate, increase in calls to one competitor, decrease in voice call frequency and increase in inactive days. Such indicators are very helpful in filtering out the variables which are not relevant to the prediction task then to correlate the data. In general, these are the main indicators which help analysts and domain experts in identifying the potential churners before they actually decide to churn. With the help of these indicators, analysts can easily remove the non-informative features. Then select the information rich variables to process further. Then feature correlation will be done, this can be extract by using 2 variables A and B to the correlation parameter ρ .
period of time a specific customer is likely to churn or receive some probability estimate of churn per customer [34].

$$Correlation(X) = \frac{\rho\ (A, B)}{\sqrt{Attributes(A)Attributes(B)}}$$

D.Imbalance in output class labels

An important observation from our data-sets is the imbalance in data. On an average about 5-10% customers churn year-on-year basis depending on the segment we are looking at [31, 32]. This imbalance in distribution consisting returning/non-returning customers is a good recipe for learning algorithms to classify a large number of customers under returning and still attain high overall accuracy. We employ cross-validation technique based on confusion matrix AUC curve to provide a best predictor model. Main advantage of this model is proportion responders. It gives the rate of successful classification. It is the metric that is used to measure how well the model can distinguish two classes.

$$\text{Count class C} \propto \frac{W\square ole\_Samples}{(W\square ole\_Classes)(Number\ of\ occurance\ C)}$$

Number of occurrence C represents the total number of occurrences of samples belonging to a given class

E.Classification using Peer Grading Regression

In this step, identified churners are categorized based on complaints data provided by the operator. Complaints dataset contain the information like complaint type, number of complaints and resolution time that is whether complain resolved within agreed time or not. Complaint type is marked by operator either critical or normal, based on nature of complaint registered by customer [33].

In general, identified churners with critical complaints have been categorized into 'High' whereas the customer with 'Normal' complaint types are categorized as 'Medium' and rest are marked as 'Low'. There are three factors involved in deciding the categorization of a churner. First is complaint type, followed by the number of complaints registered by the same customer and finally the time taken to resolve the problem.

In case complaint is not resolved within agreed time and customer has registered the same complaint more than once with type ‘Critical’, then churner severity is marked as ‘High’. Similarly, churn severity is marked as ‘Medium’ against which there is only a single complaint is registered under complaint type ‘Normal’ and resolved within agreed time.

Churn customer classification is applied by a peer grading regression task, Regression analysis is a statistical technique to estimate the relationship between a target variable and other data values that influence the target variable, expressed in continuous values. If that’s too hard – the result of regression is always some number, while classification always suggests a category. In addition, regression analysis allows for estimating how many different variables in data influence a target variable. With regression, businesses can forecast in what

Peer grading is the process of reviewing each variable of data. Reviewing of data will helps to analyze the task fully. Here customer care call is to be analyzed then if the customer provides bad review for that call then we apply the peer grading algorithm to get the updates after that customer call, here we collect all the records like their call activity, data activity and message activity for verifying that the customer will be churn or not to churn.

We assume that the true scores of all submissions S. Each grader g has an inherent bias(g) and a certain reliability customer(C).

$$N=\frac{\text{Submission of churn customer}(S,true)+\text{bias}(g).}{\text{stable customer } C}$$

True difference in means is not equal to 0

Fix analysis of customer after their bad review calls and provide the submission s1,s2,... finding their activity for minimum 1 month. To this exercise have been handed in by peer grading mechanism. Consider a set of graders g1,g2,.... By score(s,g) we denote the score given to submission s by grader g. here the grades tells that the activity to our predictor churn. Here the average prediction is calculated by following equation.

Average Prediction = Average of Labeled data - Average predicted data

Following algorithm shows the peer grading algorithm

```
INPUT:
Average Non-Churn Voice Data
Average Non-Churn Message Data
Average Non-Churn Data Data
Complaint Cruelty (low, High, Medium)
Possible_Churners PC
VoiceCus, DataCus, and SmsCus of all PCs
OUTPUT:
Churner Data_Classification (‘Voice’, ‘Sms’, and ‘Data’)

FOR each NC
// Finding the priority of customer complaints
Check Complaint_Seriousness (High, Medium, Low)
// Analyzing the complaints and if the complaints is in high then further process is continuing
        IF Complaint_Seriousness is High then
Check their activity(peer)
// Check condition of Voice, SMS and Data
IF (Voice_call < AvgVoice_Call add PC to VoiceList AND
Data < AvgData add PC to DataList AND
Msg < AvgMsg add PC to MsgList Add Submission s1 check till s30) // process continues daily up to 30 days
IF PC drop in more than 8 List then
    and add to separate List Grade G which belongs to Churns
ELSE IF PC Complaint Severity is Medium
getMaxUsage(Voice_Call, Data, Sms)
            and add to separate List
ELSE
No Campaign
  END_IF
  END_IF
  END_IF
  END_IF
END FOR
Return List[Voice], List[Data] and List[Msg]
```

In order to validate and demonstrate the accuracy of our proposed model, implemented and that were shown in next section with the telecom dataset. Proposed work flow diagram is shown in Fig 1.

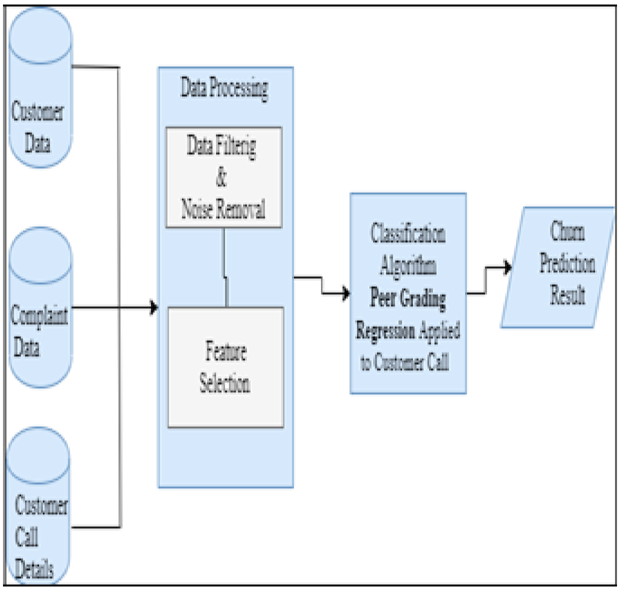


Fig 1: Flow diagram of PGR approach

IV. EXPERIMENTAL RESULT

We focused on evaluating and analyzing the performance of the machine learning methods and algorithms for predicting churn in telecommunications companies. We have experimented a number of algorithms such as Peer Grading Regression (PGR), Random Forest, Decision Tree and k-NN classifier to build the predictive model of customer Churn after developing Noise Removal, Feature Selection, Imbalance in output class labels and Classification methods. Here, simulation is carried out in MATLAB environment as it is a user friendly tool. This tool establishes proper validation of classification in terms of accuracy, sensitivity, specificity, F-measure and so on. This software runs on Intel core i3 processor with certain configuration. To evaluate the performance of tested classifiers, we use the churn dataset from the UCI Machine Learning Repository. These techniques analyze the data and identify reasons behind customer churning.

CRM can employ these techniques to maximize their profit. Furthermore, it may be used to design retention strategies to reduce the ratio of customers that are going to churn.

A.Dataset

Here, data were collected from online available Telecom industries of cell 2 cell dataset for validation. Customer data contains all data related to customer's services and convention information like Customer ID, Monthly Minutes, Roaming Calls, Blocked Calls, Customer Care Calls, Received Calls, Outbound Calls, Inbound Calls, Call Forwarding, Months in Service, Service Area. Some of the data were attached here which is shown in table 1.

Table I: Dataset parameters of telecom customer

Customer ID	Monthly Minutes	Roaming Calls	Blocked Calls	Customer Care Calls	Received Calls	Outbound Calls	Inbound Calls	Call Forwarding Calls	Months In Service	Service Area
3000006	483	0	1	1.7	55.3	46.3	6.3	0	56	MILMIL414
3000018	570	0	0.7	8.7	106.3	14.7	0.7	0	57	NNYSYR315
3000034	1039	0	3	11.3	213.9	103.3	0.3	0	55	MILMIL414
3000070	153	0	0.7	0	33.1	8	4.7	0	59	SLCSLC801
3000074	1213	1.3	6	4.3	490.1	50.7	8.3	0	52	OKCTUL918
3000086	1424	0	0	0.3	939.3	7.7	39	0	54	SEAPOR503
3000098	300	0	0.7	1	77.5	6.3	5	0	52	SLCSLC801
3000110	972	0	0	0	244.6	20	8	0	58	PITHOM412
3000246	406	0	1.7	0	47.7	6.3	4.3	0	50	SANMCA210
3000254	2961	62.3	9.3	0	871.5	210.7	96.3	2.3	52	NSHNSH615
3000258	244	0	2.7	0	48.8	13	1	0	54	KCYKCK913
3000274	380	0	0.7	0	76.2	21.7	4.7	0	56	DENDEN303
3000334	1088	0	1	3.7	335.1	11	6	0	54	OKCLRK501
3000338	1348	1.9	7.3	0	436.5	59.3	23	0	56	SANAUS512
3000342	1804	0	11	0	742.4	133	31.7	0	50	NSHNSH615
3000366	365	0	7	3.3	99.9	9.3	2	0	50	KCYWIC316
3000370	1306	0	1	1.7	809.1	25.3	0	0	52	KCYKCM816
3000438	1656	0	9.7	9.3	197.9	42.7	0.3	0	52	DALSHR903
3000450	25	0	0	0	1.6	0	0	0	50	NSHNSH615

There are many things brands may do wrong, from complicated onboarding when customers aren't given easy-to-understand information about product usage and its capabilities to poor communication, e.g. the lack of feedback or delayed answers to queries. Another situation: Longtime clients may feel unappreciated because they don't get as many bonuses as the new ones. And also longtime customer may feel for dissatisfied customer care support call. For this Churn Analysis is applied to this research why customersswitch service provider based on some data like customer call data. In a Churn analysis application, the first thing is to access to the customer data. Table 1 shows the customer data to process or identify the churn. Then, factors are classified to decide which factor or factors affect customer churn decision. In general, it's the overall customer experience that defines brand perception and influences how customers recognize value for money of products or services they use.

TableII: Churn identification dataset through customer care calls

Customer ID	Total Recurring Charge	Customer Care Calls	Director Assisted Calls	Un answered Calls	Not respond calls	Un satisfied customer	Unique Subs	Active Subs	Service Area	Credit Rating
3000002	22	6.3	0.25	0.7	0.7	0	2	1	SEAPOR503	5-satisfied
3000010	17	2.7	0	0.3	0	0	1	1	PITHOM412	5-satisfied
3000014	38	0	0	0	0	0	1	1	MILMIL414	3-Good
3000022	75	76	1.24	52	7.7	4.3	2	2	PITHOM412	4-Medium
3000026	17	0	0	0	0	0	2	2	OKCTUL918	1-Highest

3000030	52	13	0.25	9	1.7	0.7	1	1	OKCTUL918	3-Good
3000038	30	2.3	0.25	0	1	0	2	2	OKCTUL918	5-satisfied
3000042	66	4	2.48	0	0.3	4	2	2	OKCOKC405	0- Dissatisfied
3000046	35	1	0	0	0	0	3	3	SANMCA210	5-satisfied
3000050	75	0	0	0	0	0	1	1	PITHOM412	3-Good
3000054	25	0.3	0	0	0	0	2	2	SANMCA210	5-satisfied
3000058	85	43.7	2.23	9	0	0.3	5	1	SLCSLC801	5-satisfied
3000062	37	7.7	0.25	3.3	1.7	1	2	2	OKCOKC405	3-Good
3000066	60	17.3	0	5	0	3.7	1	1	SLCSLC801	3-Good
3000078	70	9	0	1.7	0.3	0.3	1	1	MILMIL414	3-Good
3000082	100	114.3	0	7.3	18	4	2	1	LOULOU502	0- Dissatisfied
3000102	30	2.3	0	0.3	0	4	3	3	SLCSLC801	0- Dissatisfied
3000118	30	0.7	0	0	0	0	1	1	SANMCA210	0- Dissatisfied
3000122	17	0	0	0	0	0	2	1	KCYKCK913	1-Highest
3000126	30	5.3	0	2	1.7	0	1	1	SANMCA210	4-Medium
3000130	35	3.3	0.25	0	0.7	0	1	1	OKCOKC405	1-Highest
3000134	17	7.3	0	1.3	3	0	1	1	KCYNEW316	1-Highest
3000138	30	1.3	0	0.3	0	0	2	1	SLCSLC801	1-Highest
3000142	75	53.3	0.5	11	5	2.7	1	1	KCYKCM816	1-Highest
3000146	30	3	0	2	2	0	2	2	KCYKCM816	1-Highest
3000158	30	6.3	0	6.7	0.3	0	2	1	DENDEN303	1-Highest

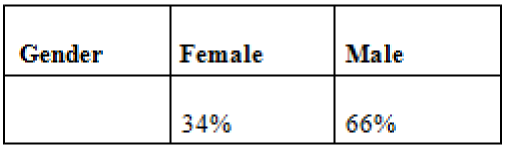
Table 2 shows the dataset of customer care call data like their satisfactory level, dissatisfactory level and the level of their opinion from customer care support, from this it can analyze that what problem is made and that solution is satisfied or not, from their opinion if the person provides dissatisfactory level then we have to watch their future response about the calls, data. If the response gets low means it is possible for churn and in that we can predict the customer churn easily.

B.Discussion

This section is going to discuss the comparison of the most prominent classification methods on churn prediction. There are a lot of forms of the boosting algorithms, but the most popular is AdaBoost, where the weak classifiers are decision trees. Linear Regression (LR), classification based on Support Vector Machine (SVM), Association Rules, advanced rule induction, Decision Tree (DT), ensemble of hybrid methods, churn prediction by feature selection techniques, Bayesian network classifiers and improved balanced random forests. The primary objective of these previously reported prediction models is to utilize large amount of telecom data to identify potential churners. However, this suffer from a number of limitations which put strong barriers towards the direct applicability of these in a real world environment where large amount of data is present. which don't provide a true representation of real world telecom data consisting of noise and large number of missing values. The presence of noise and missing values in the variables degrades the performance of the prediction models. Here 50% of the 'control list' customers have churned, retention would be 50% effective. From C5 Model incorrectly predicted 19% of churners, who were not actually churners, but it marked them as churners. For False cases approach, the model was able to correctly predict 71% of the non-churners correctly. As one of the important steps to ensure the model generalizes well, the performance of the predictive churn model has to be evaluated. That is, the prediction rates of a predictor are needed to be considered. In this work, the prediction rates refer to true churn rate (TP) and false churn rate (FP). The objective of the application is to get high TP with low FP.

To assess the classification results we count the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The FN value actually belongs to Positive P (e.g. TP + FN = P) but wrongly classified as Negative N (e.g. TN + FP = N).

TableIII: Gender Computation



Gender	Female	Male
	34%	66%

Table IV: Churn Age Factor Computation

Age	No Churn presented	Churn presented
15- 25	45	5
26- 35	35	4
36-45	18	3
46-55	22	1
Above 55	13	2

Table V: Prediction based on Customer complaint

Month	Repeated Customer Complaint call	Rare Customer Complaint call
Jan- Mar	2	.2
Apr- Jun	1.1	.5
Jul- Sep	2.3	.4
Oct- Dec	2.4	.5

Table VI: Churner categorization based on usage pattern

Category	Data	SMS	Voice
High	12	83	74
Medium	566	712	1,349
Low	812	834	2,209

Table VII: Prediction comp

Prediction Calculus	Peer Grading Regression	Existing work
	89	72

Churn prediction = Number of cases satisfied – whole
Customer call count

Accuracy: Overall accuracy of the classifier can calculate by the given formula

Accuracy = TP + TN / P + N

TableVIII: Recall computation

Method	Recall
Peer Grading Regression(PGR)	88.642
Random Forest	74.444
Decision Tree	80.678
k-NN classifier	82.909

Sensitivity(Recall): It measures the fraction of churn customers who are correctly identified as true churn

Recall = TP / P

Table IX: Precision computation

Method	Precision
Peer Grading Regression(PGR)	89.022
Random Forest	74.444
Decision Tree	68.678
k-NN classifier	79.909

Precision: It is characterized the number of correctly predicted churns over the total number of churns predicted by proposed approach. It can formally express as;

Precision = TP / TP + FP

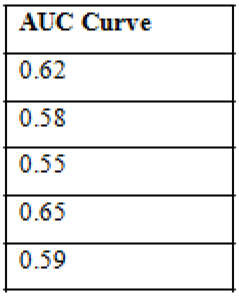
Precision or recall alone cannot describe the efficiency of a classifier since good performance in one of those indices does not necessarily imply good performance on the other. For this reason, F-measure, a popular combination.

Table X: F-measure computation

Method	F-measure
Peer Grading Regression (PGR)	86.022
Random Forest	74.444
Decision Tree	78.678
k-NN classifier	69.909

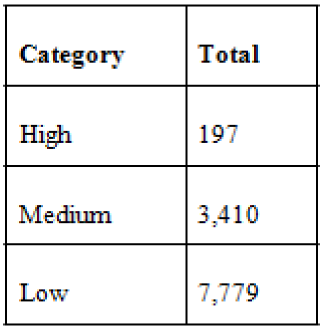
F- Measure = precision * recall / precision + recall

Table XI: Computation of AUC Curve



AUC Curve
0.62
0.58
0.55
0.65
0.59

TableXII: Churner categorization based on complaints



Category	Total
High	197
Medium	3,410
Low	7,779

Table XIII: AUC performance measure for churn predictive model

Months	Jan, Feb	Mar, Apr	May, Jun	Jul, Aug	Sep, Oct	Nov, Dec

PGR	5	5.5	4.2	3.6	6	8
Random Forest	2	2.3	3.4	3.9	1.9	4
Decision Tree	1	4	3	.9	2	2
k-NN classifier	4	3	5	4	2	2

A classified data sample is taken from original CDR dataset. Proposed approach exactly find the consistent ratio of churner and non-churners from the dataset. There are total 20000 datasets were selected in which 9.1% are churners. Below the chart shows the graphical representations of our work.

F-Measure: A composite measure of precision and recall to compute the test's accuracy. It can be interpreted as a weighted average of precision and recall.

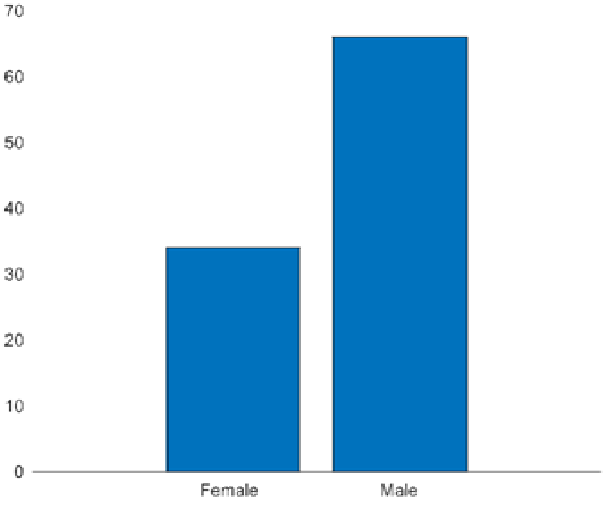


Fig 1: Gender Computation

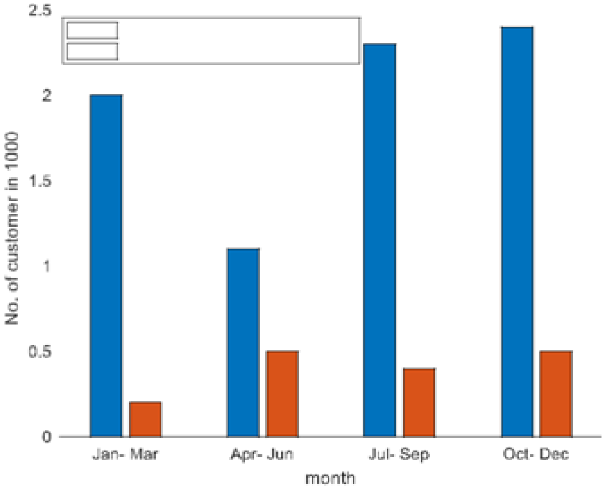


Fig 3: Prediction based on Customer complaint

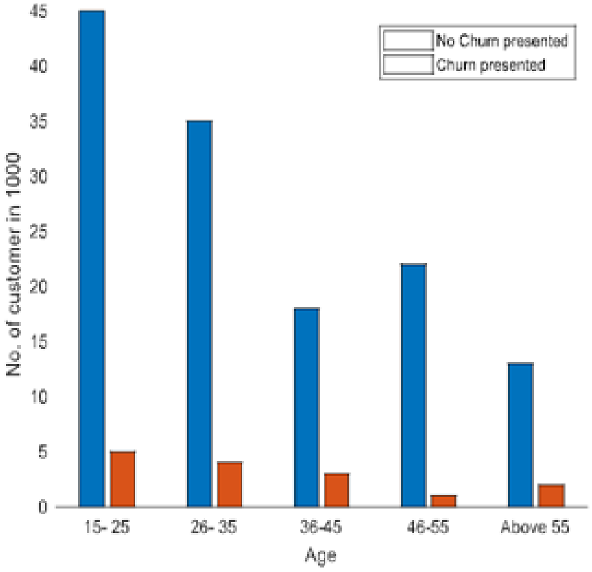


Fig 2: Churn Age Factor Computation

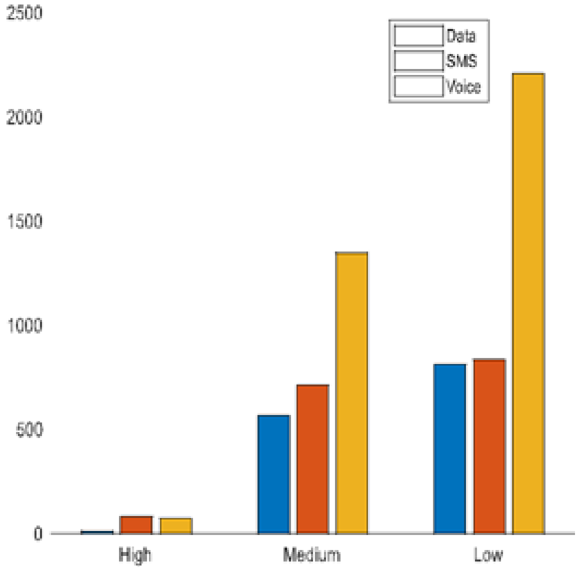


Fig 4: Churner categorization based on usage pattern

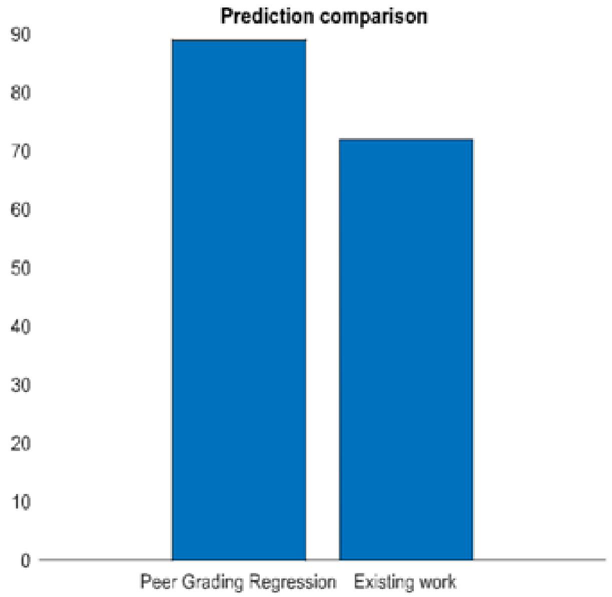


Fig 7: Precision computation

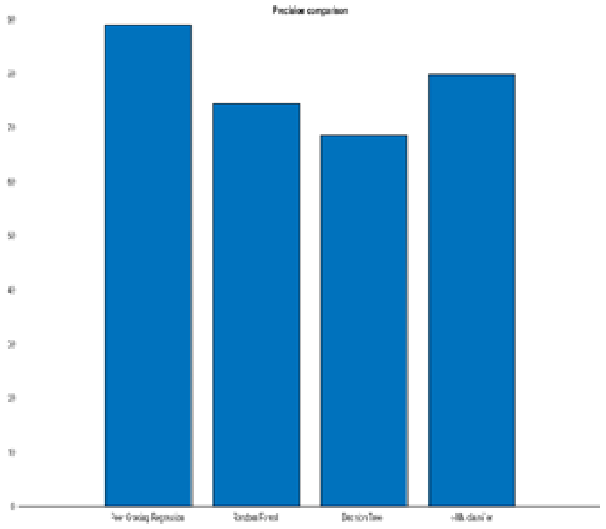


Fig 5: Prediction comparison

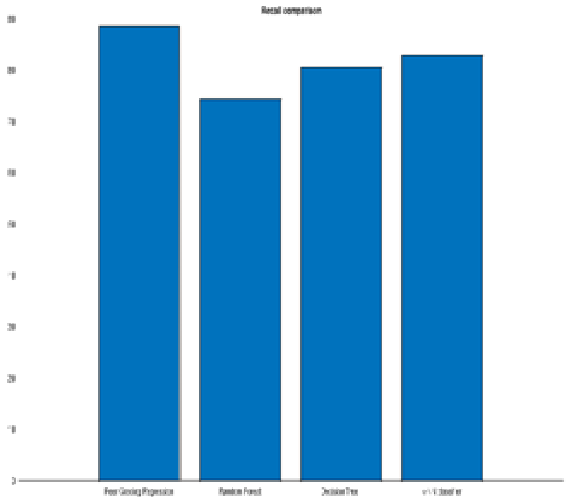


Fig 6: Recall computation

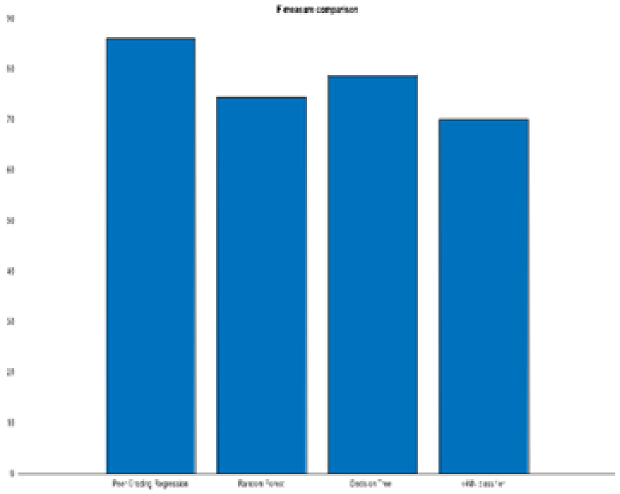


Fig 8: F-measure computation

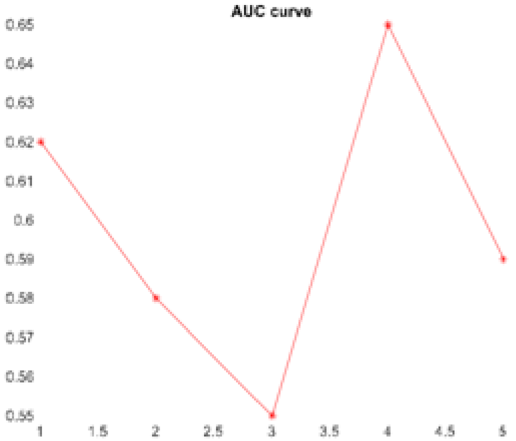


Fig 9: AUC Computation

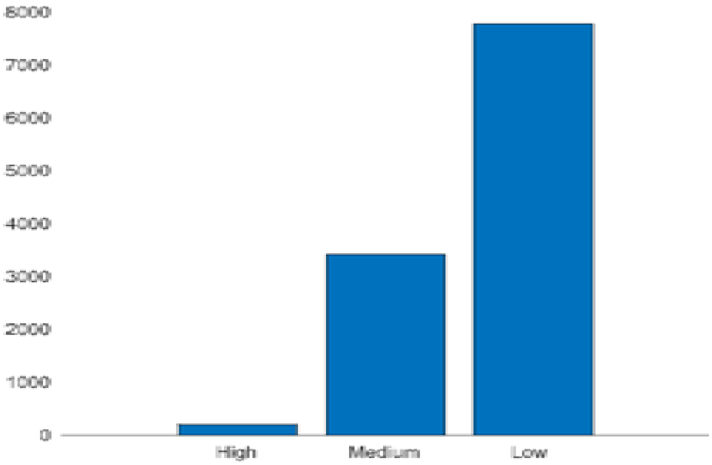


Fig 10: Churner categorization based on complaints

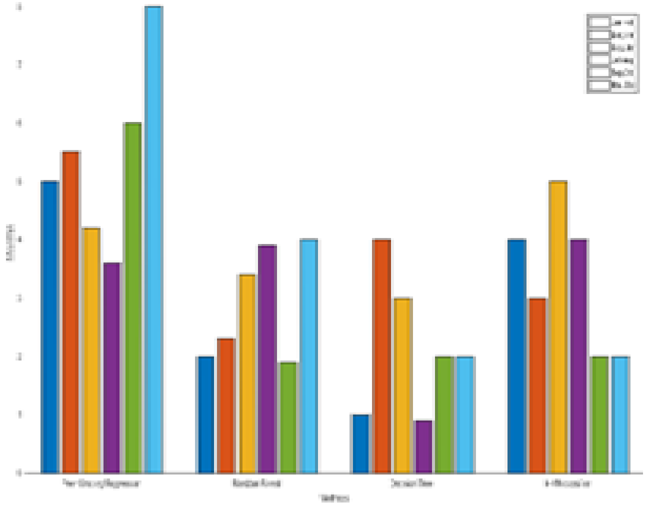


Fig 11: AUC performance measure for churn predictive model

V. CONCLUSION

Churn is an unavoidable procedure in the telecom industry. In any case, however, troublesome, it is conceivable to recognize the reasons for churn usages by a few methodologies. Prediction plays a vital role in identifying the churn customer. In this research work, the prediction model has been proposed for the telecommunication sector for the customer churn using peer grading regression method. This model provides that reviewing the customer process in daily activity and analyze the maximum possible churn activity and predict the customer churn process accurately.

By applying our proposed method, it predicts the churn customer exactly by analyzing the activity of the customer. As many approach said that holding the existing customer is more precious than getting new customer. So to provide a enhancing approach our proposed model exploits the classifiers to accurately predict the churners from an enormous arrangement of customer records dependent on the machine learning approach. Our result of experimental performed on the CDRs dataset of a cell two cell telecom organization proved that the proposed model achieves best result by applying AUC score of true positive range of 89% accuracy.

REFERENCES

1. Idris, A. and Khan, A., 2017. “Churn prediction system for telecom using filter–wrapper and ensemble classification”. IEEE, 60(3), pp.410-430.
2. Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W., 2019. “A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom secto”r. IEEE Access, 7, pp.60134-60149.
3. Azeem, M. and Usman, M., 2018. A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. International Journal of Computational Intelligence Systems, 11(1), pp.66-78.
4. Anjum, A., Zeb, A., Afridi, I.U., Shah, P.M., Anjum,A., Raza, B. and Anwar, Z., 2017. “Optimizing Coverage of Churn Prediction in Telecommunication Industry”. International Journal of Advanced Computer Science and Applications, 8(5), pp.179-188.
5. https://towardsdatascience.com/hands-on-predict-customer-churn-5c2a42806266
6. Adwan, O., Faris, H., Jaradat, K., Harfoushi, O. and Ghatasheh, N., 2014. Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. Life Science Journal, 11(3), pp.75-81.
7. Seo, D., Ranganathan, C. and Babad, Y., 2008. Two-level model of customer retention in the US mobile telecommunications servicemarket. Telecommunications policy, 32(3-4), pp.182-196.
8. Hung, S.Y., Yen, D.C. and Wang, H.Y., 2006. Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), pp.515-524.
9. Coussement, K., Benoit, D.F. and Van den Poel, D., 2015. Preventing customers from running away! Exploring generalized additive models for customer churn prediction. In The Sustainable Global Marketplace (pp. 238-238). Springer, Cham.

10. Qureshii SA, Rehman AS, Qamar AM, Kamal A,Rehman A. 2013 “Telecommunication subscribers’ churn prediction model using machine learning”. In:Eighth international conference on digital information management.. p. 131–6.
11. Tiwari, A., Hadden, J. and Turner, C., 2010, March. A new neural network based customer profiling methodology for churn prediction. In International Conference on Computational Science and ItsApplications (pp. 358-369). Springer, Berlin,
12. Heidelberg.
13. Adnan Amin, Shehzad, S., Khan, C., Ali, I. and Anwar,S., 2015. “Churn prediction in telecommunication industry using rough set approach”. In New trends in computational collective intelligence (pp. 83-95). Springer, Cham.
14. Hanif, E., 2019. “Applications of data mining techniques for churn prediction and cross-selling in the telecommunications industry” (Doctoral dissertation, Dublin Business School).
15. Ahmad, A.K., Jafar, A. and Aljoumaa, K., 2019.“Customer churn prediction in telecom using machine learning in big data platform”. Journal of Big Data, 6(1), p.28. Springer.
16. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. andChatzisavvas, K.C., 2015. “A comparison of machinelearningtechniquesforcustomerchurnprediction”. Simulation Modelling Practice and Theory, 55, pp.1-9.
17. Amin, A., Khan, C., Ali, I. and Anwar, S., 2014,November. “Customer churn prediction in telecommunication industry: With and withoutcounter-example”. In Mexican international conference on artificial intelligence (pp. 206-218). Springer, Cham.
18. Bi, W., Cai, M., Liu, M. and Li, G., 2016. “A big data clustering algorithm for mitigating the risk of customerchurn”. IEEE Transactions on IndustrialInformatics, 12(3), pp.1270-1281.
19. Lu, N., Lin, H., Lu, J. and Zhang, G., 2012. “A customer churn prediction model in telecom industry using boosting”. IEEE Transactions on Industrial Informatics, 1(2), pp.1659-1665.
20. Prashanth, R., Deepak, K. and Meher, A.K., 2017, July.“High accuracy predictive modelling for customer churn prediction in telecom industry”. In International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 391-402). Springer, Cham.
21. Huang, B., Kechadi, M.T. and Buckley, B., 2012.“Customerchurnpredictionin telecommunications”. Expert Systems with Applications, 39(1), pp.1414-1425.
22. Saran Kumar, A. and Chandrakala, D., 2016. “A Survey on Customer Churn Prediction using Machine Learning Techniques”. International Journal of Computer Applications, 975, p.8887.
23. Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.
24. Huang F, Zhu M, Yuan K, Deng EO. 2015 “Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data”. p.607–18.
25. Burez, J. and Van den Poel, D., 2009. Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), pp.4626-4636.
26. Awang, M.K., Rahman, M.N.A. and Ismail, M.R., 2012. Data mining for churn prediction: multiple regressions approach. In Computer Applications for Database, Education, and Ubiquitous Computing (pp. 318-324). Springer, Berlin, Heidelberg.
27. Zhu, B., Xiao, J. and He, C., 2014. A balanced transferlearning model for customer churn prediction. In Proceedings of the Eighth International Conference on Management Science and Engineering
28. Management (pp. 97-104). Springer, Berlin,
29. Heidelberg.https://www.kaggle.com/vpfahad/telecom-churn-data-sets
30. Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A.,Rehman, A.: Telecommunication subscribers’ churnprediction model using machine learning. In: Eighth International Conference on Digital Information Management (ICDIM 2013), pp. 131–136. IEEE(2013)
31. Kandel, Ibrahem Hamdy Abdelhamid. "A comparative study of tree-based models for churn prediction: a case study in the telecommunication sector." PhD diss., 2019.
32. Le, M., Nauck, D., Gabrys, B. and Martin, T., 2013, December. KNNs and sequence alignment for churn prediction. In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 279-285). Springer, Cham.
33. Zhu, B., Xiao, J. and He, C., 2014. A balanced transfer learning model for customer churn prediction.In Proceedings of the Eighth International Conference on Management Science and EngineeringManagement (pp. 97-104). Springer, Berlin, Heidelberg.
34. Jayaswal, P., Prasad, B.R., Tomar, D. and Agarwal, S., 2016. An Ensemble Approach for Efficient Churn Prediction in Telecom Industry. International Journal of Database Theory and Application, 9(8), pp.211-232.
35. T. Verbraken, W. Verbeke, B. Baesens, “A novel profit maximizing metric for measuring classification performance of customer churn prediction models”,IEEE Transaction on Knowlee and Data Engineering 25 (2013) 961–973
36. Koen W. De Bock, Dirk Van den Poel, “An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction”, Expert Systems withApplications 38 (2011) 12293–12301

AUTHORS PROFILE

Dr. M. Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph. D, Mother Terasa women's University, Kodaikanal). She is currently working as a professor in Ramakrishna College of Arts & Science,Coimbatore .Having fifteen years of experience in teaching and published more than two hundred papers in International Journals and also presented more than hundred papers in various national and international conferences. She received best researcher award in the year 2012 from Karpagam University. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering, Bioinformatics and Neural Network. She is a reviewer in several National and InternationalJournals.She received Women Excellance Award for the year 2020.

S.Mahalakshmi Completed MCA.,Mphil.,SET and currently pursuing PhD in Bharathiar University,Coimbatore.Published and presented several papers in various national and International conferences.Her field of interest include Data Mining ,Machine Learning,Software Engineering etc.Having seven years of experience in the field of teaching as a Asst.Professor.