International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Exploring Hybrid and Ensemble Models for Customer Churn Prediction in Telecom Sector

J. Pamina, T. Dhiliphan Rajkumar, S. Kiruthika, T. Suganya, Femila.F

Abstract: ***Most prominent challenges in all business is to retain and satisfy their valuable customers for sustain successfully in the market. Numerous Machine learning approaches are emerging to develop various customer retention models to solve this issue in many applications. This swing is more realized in telecom industry due its enormous significance. This article presents an elaborated survey on machine learning based churn prediction in telecom sector from the year 2000 to 2018. We also extracted the problems and challenges in Telecom Churn Prediction and reported suggestion and solutions. We believe this article helps the researches or data analysts in the telecom field to select optimal and appropriate methods and for designing improved novel model for churn prediction in future.***

Index Terms: Churn Prediction, Machine learning, Survey, Telecom.

I. INTRODUCTION

Literature Survey aims to produce the current idea about the topic, deliver the foundation and motivation for researchers to do new work in future. This paper presents a literature survey of various machine learning techniques in Telecom Industry. Due to the Worldwide development, Information Technology has shown great increase in various Service Providers which leads to high competition among them. The most common challenge for them to tackle customer churn, retain and satisfy their customers to sustain successfully in the market [9][35]. Churn is when a customer stops the relationships from current service provider and switches to another. This unceasing activity of churning affects the total business profit and image. So, it is always better to forecast and prevent customers from churning. The recent development in analyzing of customer records information are trending currently due to its huge significance, predominantly in telecom Sector. Churn Prediction is an important element as a cost for acquiring new customers is expensive than retaining the existing ones [18][20]. Thus, a minute upgrade and development in churn prediction model prevails good economic growth in organizations. This paper presents a detailed survey of Telecom churn Prediction works from the year 2000 to 2018. It is also noticed that, there has been continuous interest in this research area for creating and designing a churn prediction model for telecom [43]. Data analyzing for telecom churn prediction involves clustering, Pattern recognition, extraction, pre-processing and classification abiding the traditional classifiers, ensemble classifiers and other hybrid methods. This article mainly takes an elaborated survey of different churn prediction Machine Learning algorithm models that have been engaged in the sphere of telecom filed. The articles are analyzed and organized methodically by considering features, methods and machine learning techniques used. It has been observed that improvement in predicting accuracy in models are increased after the debut of ensemble and hybrid techniques. The structure of this paper is as follows: In section 2, We discussed about the selection of articles by Systemic Analysis Procedure for Electing Articles. Section 3, describes the taxonomy of articles. Section 4, presents a various data sources that have been employed for Telecom Churn Prediction. Section 5, reveals limitations, challenges and feature Selection Methods used in Telecom Churn Prediction. Lastly, Section 5, concludes this article.

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

J.Pamina*, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India.

T. Dhiliphan Rajkumar, Department of Computer Science and Engineering, , Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Srivilliputur Post-626126, Virudhunagar District, Tamilnadu, India

S.Kiruthika, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India.

T.Suganya, Department of Computer Science and Engineering, Sri krishna college of Technology, Coimbatore, Tamilnadu, India.

Femila F, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India.

II. SYSTEMATIC ANALYSIS PROCEDURE FOR ELECTING ARTICLES

The research articles in this paper are collected and elected according to Systemic Analysis Procedure (SAP). This Strategy helps to pick the standard articles to answer the research queries in effective and appropriate manner. Initially, we gathered 951 articles related to research queries. Next, we removed 476 papers due to irrelevant abstract and content outside the scope. The duplication phase 205 removed papers. Further, 217 papers have been eliminated by reviewer phase due to poor works.

A. *Research Questions:*

Research Queries carries three sets of questions:(a) Queries related to Machine Learning methods used in Telecom churn Prediction;(b) Questions related to Telecom churn datasets; and (C) Queries related to future trend and opportunities. Table 1 depicts the research questions for Telecom churn prediction.

Table 1. Research Queries

S. No	Questions
RQ1.	Which type of Machine Learning algorithm is employed for classification, clustering and optimization in churn Prediction?
RQ2.	What are the major kinds of ML methods used in churn Prediction?
RQ3.	What are the types of public and private datasets used in churn prediction? How many occurrences these datasets have been used?
RQ4.	How to integrate single classifiers to design hybrid classifier?
RQ5.	What is meant by hybrid ensembles? Why it is popular in recent days?
RQ6.	What are the major frequent challenges to perform customer churn prediction in Telecom? What are possibilities are developed to overcome the challenges?

B. *Articles Source:*

The papers are collected in the time duration between 2000 to 2018 from Standard sources mentioned below.

- IEEE Explorer
- Elsevier
- Springer
- Google Scholar
- ACM Digital Library.

C. *Search phrase:*

- Telecom churn Prediction
- Customer Churn in Telecom
- field
- Customer retention
- Churn prediction

D. *Inclusion and Exclusion Aspects:*

- Articles must from standard high-quality publishers and downloadable.
- Articles that report for application in Telecom industry only.
- Articles must possess quality work relevant to binary classification, clustering, prediction and identification of churners.
- It must propose idea or solutions to Telecom customer churn problems issues.
- Articles must relevant on Machine learning and its optimization algorithms.
- Papers should not be a review or scrutiny paper.
- Articles are other than English.
- Papers which has duplicates works, lack of effectiveness and not peer reviewed.

III. TAXONOMY OF ELECTED ARTICLES

The articles are investigated and sorted systematically based on their features, methods and machine learning techniques employed. In consideration of these criteria, the articles are divided into four main categories as traditional single methods, hybrid classifier methods, ensemble classifiers methods and hybrid ensemble classifiers. Fig 1. shows taxonomy of Various kinds of Churn Prediction Techniques from the year 2000 to 2018.

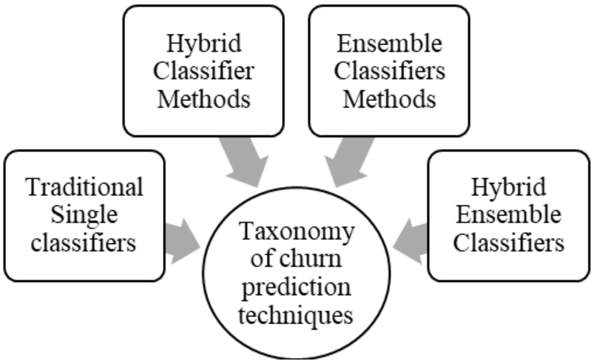
Hybrid Classifier Methods
Ensemble Classifiers Methods
Traditional Single classifiers
Taxonomy of churn prediction techniques
Hybrid Ensemble Classifiers

FIG 1. TAXONOMY OF CHURN PREDICTION TECHNIQUES

Traditional single classifier methods are common and standard bygone techniques such as regression, SVM, Decision Trees, etc. Hybrid classifiers are designed by integrating of two or more single classifiers. Ensemble classifiers are techniques such as boosting, stacking and bagging which is used for improving accuracy. It has been realized that increase in efficiency of models are after the introduction of ensemble and hybrid methods. Hybrid ensembles aggregate hybrid of multi classifiers with ensemble methods. Now a days, hybrid ensembles shines in telecom predictive data analytics and becomes very popular due its higher predicting ability.

A. *Traditional Single Classifier Methods*

Traditional Single Classifier Methods are most popular baseline classifiers such as Decision Trees, Support Vector Machines, Bayesian Network, Regression and Neural Networks. Churn Prediction Models have been designed using single classification algorithms and used for prediction of churners in datasets. Fig 2 represents various algorithms used in Traditional single classifier techniques from the year 2000 to 2018. In 2000, Michael et al. [47] introduced a churn prediction model by using Neural Networks and Linear Regression on a private wireless telecom dataset. Nath et al. [1] used Bayesian classifier in the year 2003, they applied the model on Teradata from Duke university. Their model acquired 68% of accuracy. In 2006, Shin-Yuan Hung et al. [2] selected K-means, Artificial Neural Networks (Back Propagation) and Decision Tree (C5.0) algorithms for research. These three algorithms are used in predictive modelling and customer segmentation. The data source they used was from Taiwan telecom company of one-year data. For performance evaluation they used hit ratio and Lift. Yu Zhao et al. (2005) proposed one class Support Vector Model which detects anomalies and predicted the accuracy of 87.1% on Teradata from Duke university [4].In 2008, Xia and Jin et al. [5] used Support Vector Machine on UCI churn Dataset. The conclusion was Radial Basis Function yields better results (90.9% of accuracy) than SVM with Radial Basis Function result (59% of accuracy). Pınar Kisioglu, et al. [6] used Bayesian Belief Network for identifying the effective churn management from customer’s behaviours. The model was applied on Turkish telecom dataset.

They used CHAID method for converting continues variables to discretize variables. In 2010, Marcin Owczarczuk et al. [16] used Logistic Regression in a Private dataset and selected lift curve for evaluation measure. He suggested future work shall be churn model for both prepaid and post-paid customers. In 2011, [12] Abbas Keramati et al. used Binomial Logistic Regression algorithm on Iranian mobile operator data. They calculated coefficients and hypothesis for variables present in the dataset. Wouter et al. (2012) [50] introduced a profit measure and conducted experiments with various classification algorithms such as LR, DT, NB etc applied on 11 telecom datasets. Decision trees performs well among others. Bingquan Huang et al. (2012) used six algorithms such as ANN, LR, DT, NB, SVM etc on a real-life Ireland telecom dataset [21]. They performed new feature selection approach in all above algorithms with the evaluation measure of true and false churn rate.

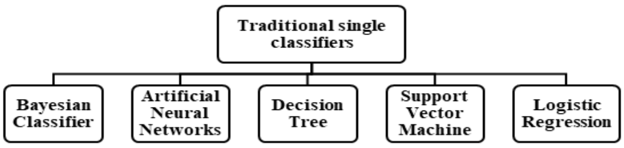

Fig 2. Traditional Single Classifiers

B. *Hybrid classifiers*

Hybrid classifier methods are developed by integration of two or more machine learning classifier algorithms. Since single predictor methods cannot perform well, hybrid classifiers are emerged to improve the prediction accuracy of the model in telecom field. Fig 3. represents various hybrid approaches of machine learning used in Telecom churn prediction from the year 2008 to 2018. In 2007, Bong-Horng Chu et al. [7] constructed a hybrid architecture of learning mode and usage mode. They used C5.0 for classification and GHSOM for clustering on Taiwan telecom dataset and they realized 85% of accuracy. Chih-FongTsai et al. (2009) [8] proposed a model with hybrid algorithms in combination ANN with ANN and ANN with SOM. They realized the accuracies of 94.32% and 93.06 %. They dint apply any feature selection methods and they entire model was tested by fuzzy testing data. In 2009, Pendharkar et al. proposed a model based on Neural Network and Genetic algorithm [17]. They used Tera duke datasets and used False positive rate for evaluation measure. Jiayin Qi et al. (2010) [9] integrated the advantages of ADTrees and Logistic Regression and applied on a private telecom dataset. They used ROC as evaluation measure and reported that variables selection and model selection are two main features for prediction churn. In 2010, Bingquan Huang et al. [15] use modified NASA II method for optimization for selecting sub features on real life Ireland Telecom data set. They used Decision Tree for fitness function and got 96% improved accuracy. WouterVerbeke et al (2010) [10] combined AntMiner+ with ALBA and realized the specificity of 99.71% and the best results are seen in ALBA combined with RIPPER or C4.5. In 2011, Adem Karahoca et al. [11] introduced a clustering algorithm called X-Means and Fuzzy C Means integrated with ANFIS for sensitive churn prediction. A comparison of many hybrid algorithms was executed and they reported 0.91 Sensitivity on GSM, Turkey dataset. In 2011, Hyeseon Lee et al. [13] built a model based on PLS techniques on highly correlated Tera Duke dataset. They reported PLS has performs well when compared to all other single classification models. In 2012, Zhen-YuChen et al. [22] proposed a novel approach called HMK-SVM to integrate static and longitudinal trends in customer data and reported 0.98 AUC value on Duke dataset. In 2013, Ying Huang et al. [25] introduced a hybrid approach of combining K-Means for grouping customers and FOIL algorithm for predicting churn. 5-fold cross validation is used as evaluating the model and it yields 89.70 as AUC value. Keramati et al. (2014) implemented a churn predictions models using four algorithms namely DT, ANN, KNN, SVM and reported ANN performs well among them [27]. Then constructed a hybrid of all above algorithms and reported a 95% of accuracy. Ammar A.Q et al.(2016) [31] proposed a hybrid firefly technique and reported 86.3% with 2.5 min. Hybrid firefly algorithms overcomes accuracy and run time of normal firefly algorithm. In 2016, [32] Wenjie et al. proposed a hybrid algorithm called SDSCM which is the combination of SCM and AFS and reported a clustering accuracy of 96% on Iris and wine dataset. Parallel SDSCM was developed and implemented in Hadoop tool on china telecom dataset. They fragmented the customers into 8 clusters and given priority based on churn rate of each clusters In 2017, M Azeem et al. [34] used fuzzy classifiers and stressed the significance of TP rate. They applied the fuzzy model in south Asian data set and reported AUC value of 0.68 by using OWANN classifier. In 2017, Long Zha et al. proposed a new KLMM algorithm for feature selection for high dimensional issue and used leave one out method as cross validation to evaluate the hyper parameter .In 2017, E. Sivasankar et al. used many clustering algorithms like K-Means, FCM, PFCM and reported that decision tree combined with K-Means gives higher accuracy when compared to all the combination [37]. In 2018, Adnan Amin et al. [40] developed a method based on the distance factor of classifiers. They applied this method on four different datasets and Naive Bayes was used as a baseline classifier. Bayesian Binomial method test was used to evaluate the entire system. J. Vijaya et al. [41] proposed a hybrid method of multi class clustering called PPFCM with ANN and reported an accuracy of 94%. They applied this novel hybrid method on tera duke dataset in 2017. Arno De Caigny et al. [42] proposed a hybrid method called LLM for classification of data. Decision Tree is used for segmentation of data and LLM is used in every leaf. The proposed model reported 0.62 AUC value.J. Vijaya et al. (2018) built a hybrid model using fuzzy clustering such as FCM, PCM, PFCM with DT, KNN SVM, NB & LDA. They made an ensemble combination of algorithms with bagging, boosting, and Random Subspace and reported the best ensemble hybrid as FPCM+ boosting with yields 98.40 % of accuracy. S Hoppner et al. [44] proposed a new classifier namely Proftree which is derived from Decision tree. They introduced this classifier for profitability and interpretability in churn prediction model. They used 9 different dataset and reported ProfTree algorithm yields good EMPC value when compared to other tree-based classifiers in the year 2018.In 2018, S. Babu et al. [52] proposed algorithms for class imbalance issue by enhanced

SMOTE and DT. They achieved higher accuracy on UCI churn dataset using those algorithms.

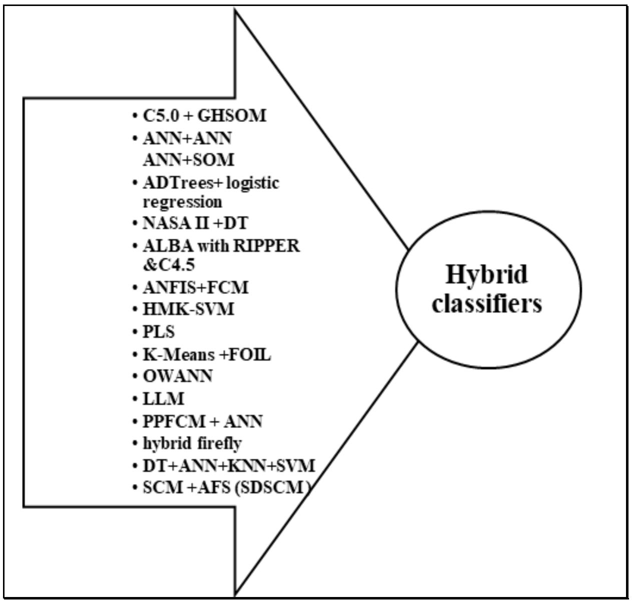

Fig 3. Hybrid Classifiers

C. *Ensemble classifiers*

Ensemble Methods are group of combined weak classifiers which yields better results on basis of voting Process [61]. Recently, ensemble classifiers such as boosting, bagging etc are used in Telecom field which are becoming popular for producing desired accurate results [46]. Fig 4. depicts various Ensemble Classifiers used in Telecom Churn Prediction from 2000 to 2018. Yong Seog Kim (2006) proposed an ensemble of ANN and Logit algorithms for better feature selection prediction. The dataset used was provided by Teradata Center for CRM at Duke University [3]. In 2006, Aurelie et al. [49] presented a comparison evaluation of three concepts namely Bagging, Boosting and Binary Logit model. They reported Bagging and Boosting yields good predictive power and its suitable for large datasets. In 2011, [14] Koen W.De Bock proposed two ensemble models namely Rot boost and Rotation Forest. The feature extraction methods like PCA, ICA and SPR are used with proposed techniques. They applied on real time European Telecom dataset and reported AUC value of 0.63 for combination of rotation forest with PCA.Adnan Idris et al. [18] integrated Genetic algorithm with Adaboost with two standard data of cell2cell and Tera dataset from Duke university. They reported AUC value of 0.89 evaluated by 10 f old cross validation. and in same year, they proposed an approach using random forest, mRMR &RF and reported a AUC value of 0.75. RF and KNN was used to evaluate the performance of reduced attributes. In 2012, [20] Koen W.De Bock et al. proposed an algorithm called GAMensplus and reported 63% of accuracy on European dataset. they compared with other techniques such as Bagging, RSM and Logistic Regression. Adnan Idris et al. [23] (2012) analysed a comparative study of tree-based ensemble algorithms with many feature selections techniques and reported Rotboost combines with mRMR gives higher AUC value of 0.86 oncell2cell dataset. In 2013, Adnan Idris et al. [24] combines RotBoost + mRMR and reported AUC value of 0.816 and 0.761 on Cell2Cell and orange dataset respectively. They used 10-fold validation for validating the performance of various feature extraction algorithms. In 2014, Ning Lu et al. [26] proposed a model to predict churn based on weights assigned by gentle Adaboost algorithm and Logistic Regression is used as a baseline algorithm. Gradient Descent technique is used for optimization and reported AUC value of 64.08. In 2015, T Vafeiadis et al. [28] used all baseline algorithms and evaluated the suitability using cross validation. In next phase, the performance is increased by boosting algorithm. Monte carlo simulation was applied to all baseline machine learning algorithms. The best algorithms were SVM_POLY with Adaboost which yields 84% of F-measure and 97% of accuracy. Jin Xiao et al. (2015) presented a feature selection technique based on GMDH Neural Network and classification is implemented for developing patterns from the data. Type 1 and type 2 accuracy are examined [29]. In 2015, Adnan Idris et al. [30] compared techniques in many phases, PSO, GA and mRMR was used for class imbalance, feature reduction process. SVM, Rotboost, Rotation forest and Random forest are used bring out feature space. Finally, ensemble methods are used based on voting. They reported AUC value of 0.85 and 0.82 for Orange and Cell2Cell datasets respectively. In 2017, [36] Bing Zhu et al. compared many techniques for feature selection, cost effective and ensemble techniques using many algorithms. They used eleven telecom public and private data from various sources. Adnan Idris et al. (2017) [33] proposed a combined technique of GP with Adaboost for higher level of classification and PSO was used to imbalance class issue. They reported AUC value of 0.63 and 0.91 for orange and Cell2cell dataset respectively. In 2018, J. Vijaya et al. [38] implemented a churn prediction model for feature selection using rough set, wrapper and filter techniques combined with ensemble techniques like bagging, boosting and random subspace for optimization.

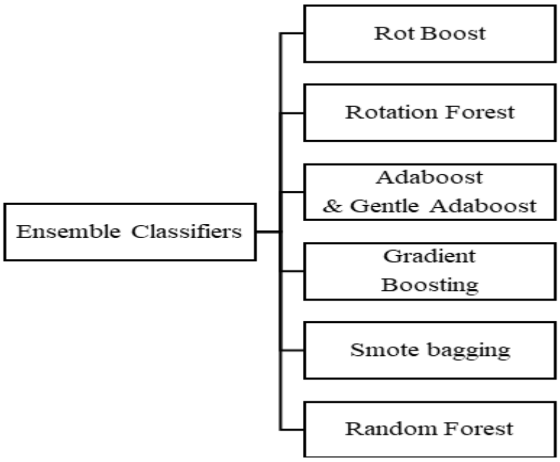

Fig 4. Ensemble Classifiers

D. *Hybrid ensemble classifiers*

Hybrid ensemble classifiers are made by new way integrating multiple classifiers. These classifiers yield optimal accuracy compared to bygone traditional methods. It is designed and developed by combination of two or more ensemble approaches like boost-stacked, bagged-stacked etc.

Many single classifiers are combined with various ensemble methods to form a hybrid of ensembles. Fig 5
In 2017, E. Sivasankar et al. [43] made hybrid of algorithms with PSO and simulated annealing in pre-processing stage and combined with hybrid of classifiers. They applied various models on small orange and large orange dataset and reported PSO with FSSA yields more accuracy than other hybrid models. In 2017, Adnan et al. [48] created hybrid of ensemble by heterogeneous and homogenous classification algorithms. They reported heterogeneous ensemble algorithms yields higher accuracy than individual and homogeneous ensemble methods. In 2018, Mahreen Ahmed et al. [45] used hybrid of ensembles of boost stacked and bagged stacked techniques with baseline algorithms. They reported the bagged stacked performs well in both datasets with 98.4% and 97.2% of accuracies. In 2018, [52] Ammar et al. created ensemble stacking with bench mark algorithms and integrated cost-effective mechanism. They applied on UCI churn dataset.

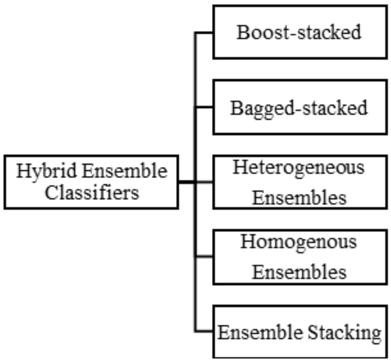

Fig. 5. Hybrid Ensemble Classifiers.

IV. DATASETS FOR CHURN PREDICTION:

Churn prediction in Telecom has been employed in both public and private datasets. The private churn datasets employed by researchers are gathered from various telecom operators. Most of the private datasets are unattainable due to proprietary issues.The summary of publicly available dataset used for telecom churn prediction are shown in table 2 and fig. 6 depicts the number of articles used for reseach using various Telecom datasets.

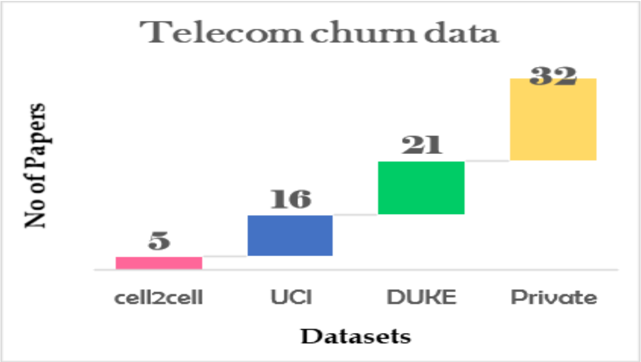

Fig 6. Telecom datasets Vs No of articles

Table 2. Publicly available Telecom datasets

No	DATASET	INSTANCES	FEATURES
1.	UCI/Big ML – University of California [60]	3333	21
2.	IBM Watson [53]	7043	21
3.	Sigtel Telecom (UK) [55]	5000	21
4.	Kaggle- private dataset [56]	100,000	100
5.	Orange dataset French Telecom company [54]	50,000	260
6.	SATO (2015) South Asian telecom company [57]	2000	13
7.	Cell2cell, Duke university Research Centre (CRM) [58]	71,047	58
8.	Telecom Churn Data for SE Asia Region (Kaggle) [59]	100,000	226

V. CHALLENGES

The prominent research challenge in Telecom churn prediction is data imbalance issue in Telecom dataset. Publicly available dataset for telecom are highly imbalance in nature. The algorithms proposed for this issue shows an effective act in churn prediction. Adnan et al. used PSO combined classifiers for class imbalance issue [19]. Bing Zhu et al. [36] used RUS method for class imbalance issue in 11 different datasets. Adnan et al. [33] applied PSO under sampling for imbalanced class distribution in two publicly available datasets. Another important challenge is integrating of multiple classifiers to form a hybrid one. Since single predictors doesn't perform well, there was shift from single predictors to hybrid classifiers. Many approaches [37] [39] [52] are introduced to solve this issue. Third challenge is about combination of multiple classifiers and ensemble methods to form hybrid ensemble. Recently introduced novel way method [45] performs well compared to bygone hybrid classifier methods. Selecting the correct feature for churn prediction also comes a challenging issue in telecom churn prediction analytics. The below Fig 7. summarize the various methods used for feature selection used in past studies.

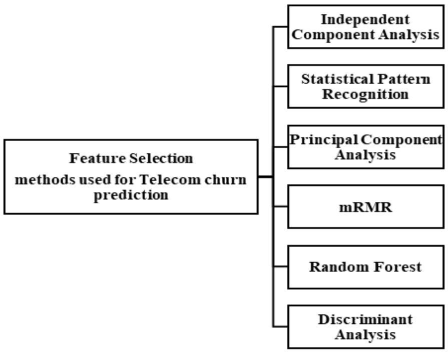
Feature Selection methods used for Telecom churn prediction
Independent Component Analysis
Statistical Pattern Recognition
Principal Component Analysis
mRMR
Random Forest
Discriminant Analysis

Fig 7. Feature Selection Methods

VI. CONCLUSION:

Telecom churn prediction is a trending area that is frequently employed in research to satisfy the valuable customers. Recently past, many Machine Learning models has been employed on different public and private telecom dataset. This article contributes an elaborated survey on various machine learning techniques employed between 2000 to 2018. Fig 8. shows a number of published standard articles between year 2000 to 2018. It has been observed that there is a continuous evolution of creating churn prediction models by researches especially in telecom field. This paper also reveals about public and private telecom churn datasets and major challenges in telecom sector. It is also perceived that more standard papers in the year 2017 and 2018. Currently, hybrid ensembles are becoming so popular due its higher prediction ability and huge significance. Table 3. Depicts the entire summary of various churn prediction carried out between the year 2000 to 2018.

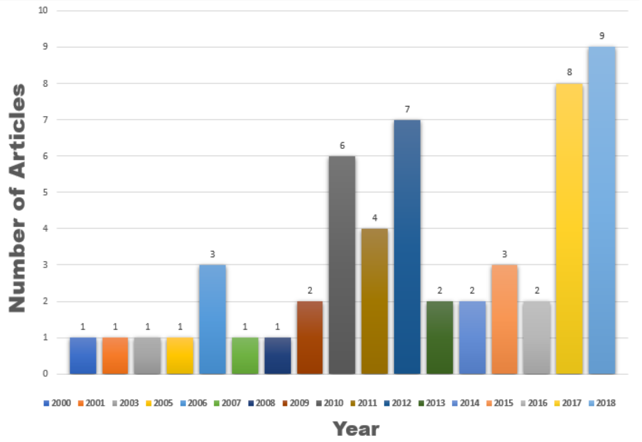

Fig 8. No Of Published Articles Per Year In Telecom Churn Prediction (2000 - 2018)

REFERENCES

1. Nath, Shyam V., and Ravi S. Behara. "Customer churn analysis in the wireless industry: A data mining approach." Proceedings-annual meeting of the decision sciences institute. Vol. 561. 2003.
2. Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." Expert Systems with Applications 31.3 (2006): 515-524.
3. Kim, YongSeog. "Toward a successful CRM: variable selection, sampling, and ensemble." Decision Support Systems 41.2 (2006): 542-553.
4. Zhao, Y., Li, B., Li, X., Liu, W. and Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. Int. Conf. Advanced Data Mining and Applications, Springer, pp. 300–306.
5. Xia, G.-E. and Jin, W.-D. (2008). Model of customer churn prediction on support vector machine. Syst. Eng. Theory Pract., **28**: 71–77.
6. Kisioglu, Pınar, and Y. Ilker Topcu. "Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey." Expert Systems with Applications 38.6 (2011): 7151-7157.
7. Chu, Bong-Horng, Ming-Shian Tsai, and Cheng-Seen Ho. "Toward a hybrid data mining model for customer retention." Knowledge-Based Systems 20.8 (2007): 703-718.
8. [8] Tsai, Chih-Fong, and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." Expert Systems with Applications 36.10 (2009): 12547-12553.
9. [9] Verbeke, Wouter, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert systems with applications 38.3 (2011): 2354-2364
10. [10] Qi, Jiayin, et al. "ADTreesLogit model for customer churn prediction." Annals of Operations Research 168.1 (2009): 247.
11. [11] Karahoca, Adem, and Dilek Karahoca. "GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system." Expert Systems with Applications 38.3 (2011): 1814-1822
12. [12] Keramati, Abbas, and Seyed MS Ardabili. "Churn analysis for an Iranian mobile operator." Telecommunications Policy 35.4 (2011): 344-356.
13. [13] Lee, Hyeseon, et al. "Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model." Decision Support Systems 52.1 (2011): 207-216.
14. [14] De Bock, Koen W., and Dirk Van den Poel. "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction." Expert Systems with Applications 38.10 (2011): 12293-12301
15. [15] Huang, Bingquan, Brian Buckley, and T-M. Kechadi. "Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications." Expert Systems with Applications 37.5 (2010): 3638-3646.
16. [16] Owczarczuk, Marcin. "Churn models for prepaid customers in the cellular telecommunication industry using large data marts." Expert Systems with Applications 37.6 (2010): 4710-4712.
17. Pendharkar, Parag C. "Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services." Expert Systems with Applications 36.3 (2009): 6714-6720.
18. Idris, Adnan, Asifullah Khan, and Yeon Soo Lee. "Genetic programming and adaboosting based churn prediction for telecom." 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2012.
19. Idris, Adnan, Muhammad Rizwan, and Asifullah Khan. "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies." Computers & Electrical Engineering 38.6 (2012): 1808-1819.
20. De Bock, Koen W., and Dirk Van den Poel. "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models." Expert Systems With Applications 39.8 (2012): 6816-6826.
21. Huang, Bingquan, Mohand Tahar Kechadi, and Brian Buckley. "Customer churn prediction in telecommunications." Expert Systems with Applications 39.1 (2012): 1414-1425.
22. Chen, Zhen-Yu, Zhi-Ping Fan, and Minghe Sun. "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data." European Journal of operational research 223.2 (2012): 461-472.
23. Idris, Adnan, and Asifullah Khan. "Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers." 2012 15th International Multitopic Conference (INMIC). IEEE, 2012.
24. Idris, Adnan, Asifullah Khan, and Yeon Soo Lee. "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification." Applied intelligence 39.3 (2013): 659-672.
25. Huang, Ying, and Tahar Kechadi. "An effective hybrid learning system for telecommunication churn prediction." Expert Systems with Applications 40.14 (2013): 5635-5647.
26. Lu, Ning, et al. "A customer churn prediction model in telecom industry using boosting." IEEE Transactions on Industrial Informatics 10.2 (2014): 1659-1665.
27. Keramati, Abbas, et al. "Improved churn prediction in telecommunication industry using data mining techniques." Applied Soft Computing 24 (2014): 994-1012.
28. Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.
29. Xiao, Jin, et al. "Feature-selection-based dynamic transfer ensemble model for customer churn prediction." Knowledge and information systems 43.1 (2015): 29-51.
30. Idris, Adnan, and Asifullah Khan. "Churn prediction system for telecom using filter–wrapper and ensemble classification." The Computer Journal 60.3 (2016): 410-430.
31. Ahmed, Ammar AQ, and D. Maheswari. "Churn prediction on huge telecom data using hybrid firefly based classification." Egyptian Informatics Journal 18.3 (2017): 215-220.

32. Bi, Wenjie, et al. "A big data clustering algorithm for mitigating the risk of customer churn." IEEE Transactions on Industrial Informatics 12.3 (2016): 1270-1281.
33. Adris, Adnan, Aksam Iftikhar, and Zia ur Rehman. "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling." Cluster Computing (2017): 1-15.
34. Azeem, Muhammad, Muhammad Usman, and Alvis Cheuk M. Fong. "A churn prediction model for prepaid customers in telecom using fuzzy classifiers." Telecommunication Systems 66.4 (2017): 603-614.
35. Zhao, Long, et al. "K-local mamum margin feature extraction algorithm for churn prediction in telecom." Cluster Computing 20.2 (2017): 1401-1409.
36. Zhu, Bing, Bart Baesens, and Seppe KLM vanden Broucke. "An empirical comparison of techniques for the class imbalance problem in churn prediction." Information sciences 408 (2017): 84-99.
37. Sivasankar, E., and J. Vijaya. "Customer Segmentation by Various Clustering Approaches and Building an Effective Hybrid Learning System on Churn Prediction Dataset." Computational Intelligence in Data Mining. Springer, Singapore, 2017. 181-191.
38. Vijaya, J., and E. Sivasankar. "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing." Cluster Computing (2017): 1-12.
39. Vijaya, J., and E. Sivasankar. "Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector." Computing 100.8 (2018): 839-860.
40. Amin, Adnan, et al. "Customer churn prediction in telecommunication industry using data certainty." Journal of Business Research 94 (2019): 290-301.
41. Sivasankar, E., and J. Vijaya. "Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network." Neural Computing and Applications: 1-20.
42. De Caigny, Arno, Kristof Coussement, and Koen W. De Bock. "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees." European Journal of Operational Research 269.2 (2018): 760-772.
43. Vijaya, J., E. Sivasankar, and S. Gayathri. "Fuzzy Clustering with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector." Recent Developments in Machine Learning and Data Analytics. Springer, Singapore, 2019. 261-274.
44. Höppner, Sebastiaan, et al. "Profit driven decision trees for churn prediction." European Journal of Operational Research (2018).
45. [45] Ahmed, Mahreen, et al. "Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry." Neural Computing and Applications (2018): 1-15.
46. Dietterich, T. G. (2000). Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, Springer, pp. 1–15.
47. Mozer, Michael C., et al. "Churn reduction in the wireless industry." *Advances in Neural Information Processing Systems*. 2000.
48. Amin, Adnan, et al. "Just-in-time Customer Churn Prediction: With and Without Data Transformation." *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2018.
49. Lemmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." *Journal of Marketing Research* 43.2 (2006): 276-286.
50. Verbeke, Wouter, et al. "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach." *European Journal of Operational Research* 218.1 (2012): 211-229.
51. Ahmed, A.A.Q. & Maheswari, D. Int. j. inf. tecnol. (2019) 11381.https://doi.org/10.1007/s41870-018-02483
52. Babu S., Ananthanarayanan N.R. (2018) Enhanced Prediction Model for Customer Churn in Telecommunication Using EMOTE. In: Dash S., Das S., Panigrahi B. (eds) International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 632.
53. https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/
54. http://kdd.org/kdd-cup/view/kdd-cup-2009/Data.
55. https://www.kaggle.com/akhilsaichinthala/telecom-churn-data-singtel
56. https://www.kaggle.com/abhinav89/telecom-customer/data
57. https://www.kaggle.com/mahreen/sato2015
58. [58https://www.kaggle.com/datasets?sortBy=updated&group=my&page=1&pageSize=20&size=all&filetype=all&license=all
59. https://www.kaggle.com/priyankanavgire/telecom-churn
60. https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383

AUTHORS PROFILE

J.Pamina, working as an assistant professor at Sri Krishna College of Technology. She received the B. Tech in computer Science and Engineering from kalasalingam university, Krishnankoil, Srivilliputur, Tamil Nadu, India in the year 2013 and received the M.E in K.L.N college, Madurai, Tamil Nadu, India in the year 2015. Her area of interest is Data analytics.

Dhiliphan Rajkumar Thambidurai working as an assistant professor in Kalasalingam Academy of Research and Education. He received this PhD degree in Manonmaniam Sundaranar University Tirunelveli, TamilNadu, India from 2012. He received his B.E (2009) in Computer Science and Engineering from Arulmigu Kalasalingam College of Engineering, KrishnanKovil, Srivilliputhur, TamilNadu, India. He received M.E(2011) in Computer Science and Engineering from Muthayammal Engineering College Rasipuram. He has a strong passion in Web Mining, Pattern recognition and Social networking.

S.Kiruthika, working as an assistant professor at Sri krishna college of Technology, coimbatore. She received her B.E in Computer Science and Engineering from Sri Eshwar college of Engineering,Coimbatore,Tamil Nadu,India in the year 2014 and received the M.E in Sri Krishna College of Engineering and Technology ,Coimbatore,Tamil Nadu,India in the year 2016.Her area of interest is Data Science.

.

T. Suganya, working as an assistant professor at Sri krishna college of Technology, coimbatore. She received her B.E in Computer Science and Engineering from Nandha Engineering college, Erode,Tamil Nadu,India in the year 2005 and received the M.E in SNS College of engineering, Coimbatore, Tamil Nadu,India in the year 2013.Her area of interest is Data Science.

Femila. F, Working as an assistant professor at Sri Krishna College of technology, Coimbatore. She received her B.E in Computer Science and Engineering from Jeppiaar Maamallan Institute of Technology, Chennai, Tamil Nadu, India in the year 2010 and ME in Computer Science and Engineering from SKR Engineering College, Chennai, Tamil Nadu, India in the year 2012. Her area of interest Data Science.

Table 3. Summary of Churn Prediction Models from 2000 to 2018

S.No	Author(s)	Year	Algorithms Used	Dataset	Measures
1.	Michael et al.	2000	Logit Regression Neural Network	Private dataset 47,000 observations	ROC
2.	Chih ping et al.	2002	Decision tree	Taiwan dataset (114,000 records)	Miss and false rate
3.	Shyam V. Nath	2003	Bayesian classifier	Teradata Center for CRM at Duke University (100,000 customers)	Accuracy
4.	Yu Zhao Bing Li	2005	Support Vector Machine	Teradata Center for CRM at Duke University	Accuracy
5.	Yong Seog Kim	2006	Ensemble of ANN and logit	Teradata Center for CRM at Duke University (100,000 examples)	Hypotheses and Coefficients
6.	Shin-Yuan Hung	2006	K-Means, artificial neural networks (back propagation) and decision tree (C5.0)	Private: Taiwan telecom company (160,000 subscribers)	Hit ratio Lift (%)
7.	Aurelie et al.	2006	Bagging, stochastic gradient & binary logit	Teradata Center for CRM at Duke University	Top decile & Gini coefficient
8.	Bong-HorngChu	2007	C5.0 with GHSOM	Taiwan telecom dataset (65516 business subscribers)	Accuracy
9.	XIA Guo-en, JINWei-dong	2008	support vector machine	UCI churn Data UCI (3333 customers)	Accuracy
10.	Parag C. Pendharkar	2009	Genetic algorithm with NN	Teradata Center for CRM at Duke University and Real life data of 195,956 customers	False Positive Rate
11.	Chih-FongTsai	2009	ANN and SOM	American telecom company dataset (51,306 Subscribers)	Accuracy
12.	Jiayin Qi	2010	ADTrees and logistic regression	Private dataset	ROC
13.	PınarKisioglu	2010	Bayesian Belief Network	Turkish telecom dataset (2000 instances)	Churn percentage
14.	Marcin Owczarczuk	2010	logistic regression	Private dataset (85,274 observations)	Lift curves
15.	Bingquan Huang	2010	Modified NSGA-II and C4.5	Ireland Telecom data (18,600 customers)	Overall Accuracy
16.	Wouter Verbeke, David Martens	2010	AntMiner+ and ALBA	Public dataset (5000 observations)	Specificity
17.	Adem Karahoca	2011	X-means, fuzzy C means and integrated with ANFIS	Turkey GSM operator (24,900 GSM subscribers)	Sensitivity Specificity
18.	Abbas Keramati	2011	Binomial Logistic Regression	Iranian mobile operator (3150 customers)	Coefficients
19.	HyeseonLee	2011	partial least squares	Teradata Centre for CRM at Duke University (100,000 observations)	Hit rate and Lift trend curve
20.	Koen W.De Bock	2011	Rotation Forest and Rot boost	European Telecom dataset (35,550 instances)	Accuracy, AUC, Top decile life.

21.	Adnan Idris	2012	Genetic algorithm with Adaboost	orange dataset (50,000 observations) and cell2cell dataset (40,000 samples)	AUC
22.	Adnan Idris et al.	2012	PSO+mRMR+RF	French telecom orange dataset	Accuracy AUC
23.	Koen W.De Bock et al.	2012	GAMensplus	European dataset (35,550 observations)	Accuracy AUC
24.	Bingquan Huang et al.	2012	ANN, LR, DT, NB, SVM etc	life Ireland telecom dataset (827,124 customers)	True & False churn rate
25.	Wouter et al.	2012	21 classification techniques	11 telecom datasets (both private & public)	AUC, Top decile lift
26.	ZY Chen et al.	2012	HMK-SVM	Tera Duke dataset (3399 instances)	AUC Lift criteria
27.	Adnan Idris	2012	Rotboost	Cell2cell (40000 instances)	AUC
28.	Adnan Idris et al.	2013	Rotboost+ + mRMR	Cell2cell (40000 instances) Tera Duke data(50,000)	AUC
29.	YingHuang et al.	2013	k-means + FOIL	Private dataset (104,199 customer records)	AUC
30.	Ning Lu et al.	2014	Adaboost + logistic regression	(Private dataset)7190 customers	AUC
31.	Keramati et al.	2014	DT, ANN, KNN, SVM	Iranian mobile company. (3150 customer data)	Accuracy F-Score
32.	T Vafeiadis	2015	SVM-PLOY with Adaboost	UCI ML Repository 5000 samples	Accuracy F-measure
33.	Jin Xiao et al.	2015	GMDH- NN	Churn (3333 observations)	Accuracy
34.	Adnan Idris et al.	2015	PSO, mRMR, Genetic Algorithm, Random Forest, Rotation Forest, RotBoost and SVM.	Orange datasets (50,000 observations) Cell2Cell (40,000 observations)	AUC
35.	Ammar A.Q et al.	2016	Hybrid firefly	Orange dataset50,000 observations)	Accuracy
36.	Wenjie Bi et al.	2016	SDSCM, AFS, K-Means	China Telecom	Accuracy
37.	Adnan Idris et al.	2017	PSO, GP, Adaboost,	Orange datasets (50,000 observations) Cell2Cell (40,000 observations)	AUC
38.	Adnan et al.	2017	SVM,bagging, KNN, NB, NN	UCI dataset	Accuracy, Kappa
39.	M Azeem et al.	2017	Fuzzy classifiers	south Asian Telecom (600000 Instances)	AUC
40.	Long Zha et al.	2017	KLMM	Orange datasets (50,000 observations)	Kappa, accuracy
41.	Adnan et al.	2017	homo and heterogenous ensembles	UCI , KDD cup2009	AUC
42.	Bing Zhu et al.	2017	RUS, SMOTE, Bagging,	11 data sets (4 public & 9 private)	EMP, AUC
43.	E. Sivasankar et al.	2017	FCM, PFCM & K-Means, DT	Churn dataset (50,000 observations)	Accuracy

44.	E. Sivasankar et al.	2017	PSO, NB, SVM, Random Forest and other hybrid models	Orange Small and Orange Large	Accuracy
45.	J. Vijaya et al.	2018	Baseline classifiers, Bagging, Boosting, RS, rough set, filter and wrapper	Teradata Centre for CRM at Duke University	Accuracy
46.	Adnan Amin et al.	2018	CCP method with distance factor	UCI Churn (3333 Observations), IBM Watson (7043 observations), Abinav Kaggle (100,000 records) and Pakdd2006(18,000 records)	Accuracy, and F-Measure
47.	J. Vijaya et al.	2018	PPFCM-ANN	Duke Tera Data	Accuracy
48.	ArnoDe Caigny et al.	2018	Logit leaf model, DT	European telecom (47,761 instances &50,000 instances)	AUC
49.	J Vijaya et al.	2018	Fuzzy clustering algorithms with baseline classifiers	Private dataset	Accuracy
50.	S Hoppner et al.	2018	ProfTree	9 Telecom datasets	EMPC
51.	S. Babu et al.	2018	EMOTE, DT	UCI Churn dataset	ACCURACY
52.	Ammar et al.	2018	Ensemble stacking	UCI Churn dataset	Accuracy
53.	Mahreen Ahmed et al.	2018	Boosted-Stacked Bagged-Stacked	UCI dataset (5000 samples) SATO dataset (2000 observations)	Accuracy