Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

Effective ML Techniques to Predict Customer Churn

Soumi De
Department of Computer Science
CHRIST (Deemed to be University)
Bangalore, India
soumi.de@res.christuniversity.in

Dr. Prabu P
Department of Computer Science
CHRIST (Deemed to be University)
Bangalore, India
prabu.p@christuniversity.in

Dr. Joy Paulose
Department of Computer Science
CHRIST (Deemed to be University)
Bangalore, India
joy.paulose@christuniversity.in

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544785

***Abstract*—Customer churn is one of the most challenging problems that affects revenue and growth strategy of a company. According to a recent Gartner Tech Marketing survey, 91% of C-level respondents rate customer churn as one of their top concerns. However, only 43% have invested in additional resources to support customer expansion. Hence, retaining existing customers is of paramount importance to a company's growth. Many authors in the past have presented different versions of models to predict customer churn using machine learning techniques. The aim of this paper is to study some of the most important machine learning techniques used by researchers in the recent years. The paper also summarizes the prediction techniques, datasets used and performance achieved in these studies for a deeper understanding of the domain. The analysis shows that although hybrid and ensemble methods have been widely successful in improving model performance, there is a need for well-defined guidelines on appropriate model evaluation measures. While most approaches used are quantitative in nature, there is lack of research that focuses on information-rich content in customer company interaction instances, like emails, phone calls or customer support chat records. The information presented in the paper will not only help to increase awareness in industry about emerging trends in machine learning algorithms used in churn prediction, but also help new or existing researchers position their research activity appropriately.**

Keywords—Customer churn, Machine learning, Class imbalance, Hybrid model, Ensemble model

I. Introduction

Churn is the event when a customer consciously decides to terminate its relationship with a service provider or vice versa. Customer churn prediction (CCP) is the study of identifying customers who are at risk of churn. There are three main types of churn – voluntary, involuntary and unavoidable [1]. Voluntary churn is when a customer decides to terminate the service relationship with provider. This may be due to switching to a competitor of the service provider. Involuntary churn is the case when the service provider terminates the relationship with customer. An example of this type of churn may be poor payment history of the customer. Unavoidable churn is when the relationship is terminated due to unavoidable reasons like when a customer is relocating to a new place or when a region faces natural disaster. It is possible to highlight "at risk" customers well in advance through in-depth analysis of differential behavioral patterns in the past and mapping these differential characteristics back to new customers. Clearly, churn prediction problem is a supervised classification problem. Machine learning algorithms that are trained to address customer churn, do so by means of a propensity score. The propensity score is important as it highlights how likely the customer is to churn in the near future. This forms the basis of prediction performance of CCP models. In the following section, we will describe various steps in the churn prediction workflow and corresponding machine learning techniques used in each step. A visual representation of the workflow is presented in Fig. 1.

II. Customer Churn Prediction Workflow

A. Feature Selection

Feature selection is one of the most important steps in CCP workflow. It is the foundation that one creates before working or experimenting with algorithms. After all, algorithms are as good as the input features with strong prediction capability. In the context of churn prediction, datasets often contain 100+ features. Most common features present in churn datasets are related to customer tenure and recency-frequency-monetary (RFM) characteristics [2–4]. Few researchers have experimented with novel features like customer's social network graph to predict churn in telecommunication industry [5–7]. While features are useful for model training, having too many input features may have a negative impact on model performance. Hence, feature selection techniques are employed to shortlist important independent variables that optimize model performance. Most often, the optimization is performed to improve accuracy, sensitivity, specificity or top-decile lift. However, in a few studies, feature selection is performed to optimize profit from customers [8–10]. The rationale behind the choice is that with limited company resources, such approach promotes targeted retention campaigns that are directed towards profit generating customers alone. The feature selection techniques are broadly divided into three main categories – filter, wrapper and embedded.

- Filter method considers each feature in the input space and scores these features based on a statistical metric.

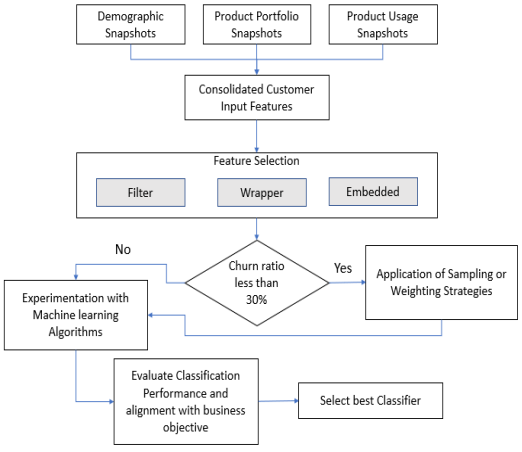

Fig. 1. Churn Prediction Machine Learning Workflow

This metric can be chi-square statistic, ANOVA coefficient, Kendall's rank coefficient or mutual information coefficients [11,12]. Finally, feature set with strong correlation with the categorical classification output is selected. Although filter methods are computationally efficient, they fail to consider inter-variable interaction to identify redundant features.

- Wrapper method on the other hand takes inter-feature interaction into consideration to select the best set of features to optimize model performance. Recursive feature elimination (RFE) and metaheuristic search algorithms, like Genetic Algorithm (GA), and bio inspired optimization algorithms, like Ant Colony Optimization (ACO), are part of this group of methods, and have been used in studies [13,14]. However, wrapper methods are prone to overfitting and are often computationally expensive.
- In Embedded method, feature selection takes place as part of model training itself. Random Forests (RF) and Lasso regularization are part of this group of methods [8,15]. Like wrapper methods, embedded methods are also computationally expensive.

B. Dimensionality Reduction

Dimensionality reduction is another option that researchers have applied to churn datasets with large number of attributes. It is the process of projecting input features to a lower dimensional space. Principal Component Analysis (PCA) and Self Organizing Maps (SOM) are used extensively and show promising results in churn prediction case studies [16,17].

C. Handling Class Imbalance

In many cases, churn is associated with rarity where there is significant gap between number of churners and non-churners. Although this is natural in many industries, it often leads to over-optimistic prediction results due to overfitting and lack of generalization capability in ML models. Hence, many researchers have tried to overcome this problem by means of sampling strategies or through cost sensitive learning.

There are three main sampling strategies that are adopted in the context of churn prediction – under-sampling, over-sampling and hybrid sampling. During training phase of churn prediction, under-sampling strategy, under samples the majority class of non-churners in order to have a balanced training data [18]. The disadvantage of this approach is that it often leads to information loss when valuable sample data is discarded. On the other hand, over-sampling strategy oversamples the minority class of churners for the same purpose. Synthetic Minority Oversampling Technique (SMOTE), is a popular over-sampling approach. Oversampling causes duplicate instances being reproduced and is therefore associated with the problem of overfitting. There is a third, and recent class of sampling strategy, which is a hybrid approach. This strategy attempts to combine the strengths of previously discussed sampling strategies, while trying to overcome the respective disadvantages. Synthetic Minority Oversampling Technique with Edited Nearest Neighbor (SMOTEENN) is a combination of over-sampling and under-sampling approach and has been used in a few papers [19,20].

Cost sensitive learning, on the other hand, addresses class imbalance without sampling the training data. In the context of churn prediction, cost sensitive learning is all the more relevant, because cost of misclassifying a churner as non-churner, and vice versa, can be significant. It may unnecessarily stretch limited headcount and monetary resources of a company. Experiments conducted by Glady et al. [21] adopted a cost-centric approach to train churn data of a financial service provider. The approach provides considerable boost to the model performance. Few studies have re-defined cost sensitive learning problem in churn as a profit centric problem. In this case, the target is to penalize the ML model for misclassification of customers that are more profitable for a company [8–10]. Profit centric ML models are found to be more aligned to business goals for revenue generation than other regular models.

D. Experimentation with Model and Framework

Few widely used ML models in CCP include Naïve Bayes (NB), k nearest neighbor (KNN), decision tree (DT), random forest (RF), logistic regression (LR) and artificial neural networks (ANN) [22,23].

1) Naïve Bayes (NB) Algorithm: Naïve Bayes algorithm calculates the probability of a customer to belong to a particular class (either churner class or non-churner class). Usually, churner class is considered as the positive class. Given a customer X with n features represented as a vector $\{x_1, x_2,\ldots,x_n\}$, this probability is given by (1):

$$p(y_j|X) = p(X|y_j)p(y_j) = p(x_1, x_2,\ldots,x_n|y_j) \quad (1)$$

where $p(y_j)$ is the prior probability of y_j and $y_j = \{1,0\}$. As NB assumes independence between conditional probabilities of independent variables $x_1, x_2,\ldots,x_n$, the posterior probability $p(y_j|X)$ can also be written as (2):

$$p(y_j|X) = p(y_j)\prod p(x_i|y_j) \quad (2)$$

where $i = \{1,2,\ldots,n\}$. Hence NB classifier classifies a customer instance by using (3). A probability score of greater than 0.5 suggests the instance belonging to the churner class.

$$\text{output} = \text{argmax}p(y_j|X) \quad (3)$$

2) K Nearest Neighbors (KNN): KNN classifiers predict an instance as churn by finding distance d between the concerned instance and another training instance X_i using the vector dot product of feature vectors and their respective moduli as shown in (4):

$$d(X,X_i) = 1 - (X \cdot X_i \div \|X\| \|X_i\|) \quad (4)$$

where $i = \{1,2,\ldots,k\}$. The training instance with minimum distance is identified as best match and churn class is subsequently assigned.

3) Random Forest (RF): Random forest is an ensemble classifier. The steps followed by RF classifier are given below:

Step 1: k training subsets are sampled from training data using bootstrap sampling.

$$S_{train} = \{S_1, S_2,\ldots,S_k\}$$

Step 2: In parallel, an out-of-bag (OOB) dataset is generated that is not part of S_{train}

$$S_{OOB} = \{OOB_1, OOB_2,\ldots,OOB_k\}$$

where $OOB_i \cap S_i = \{\}$ and $OOB_i \cup S_i = S_{train}$. The OOB sets will serve the purpose of a test set to check tree performance.

Step 3: Decision trees are developed from each S_i and while growing the tree, each split is carried out based on a smaller random subset of predictors m (usually, $m = \sqrt{n}$ where n is the total number of predictors) and best split is selected based on the gain ratio calculated for each predictor. The splitting process is repeated until a leaf node is reached. Finally, k such trees are generated using the same process and are included as part of the random forest.

Step 4: For a new instance to be classified, each tree casts its vote on an instance by predicting its class. Finally, a churn class is assigned to the instance based on majority votes.

4) Logistic Regression (LR): Logistic regression when applied to the case of churn, is a binary classifier. For an instance X_i with n independent features $x_1, x_2,\ldots,x_n$, LR calculates the probability of churn using (5):

$$P_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (5)$$

where $\beta_0, \beta_1, \beta_2,\ldots, \beta_n$ are the coefficients that will be learnt by the model via stochastic gradient descent process. These coefficients are iteratively updated post each error calculation.

5) Artificial Neural Network (ANN): Artificial Neural Network was initially developed for the purpose of modeling biological neural system, with an architecture that is similar to that of a human brain with various layers of interconnected nodes or neurons. The most commonly used architecture has an input layer, an output layer and a hidden layer, where all the processing takes place (Fig. 2). Each neuron in a layer is connected to another neuron in a subsequent layer through weighted connections. The neurons in the hidden layer perform two operations on the input $x_1, x_2,\ldots,x_n$. Firstly, it performs the summation operation as shown in (6):

$$S = \sum w_i x_i \quad (6)$$

Secondly, it applies an activation function to this net weighted input in order to make it tangible. Commonly used activation functions are sigmoid, tanh and RELU. An error criterion that is to be minimized is selected and weights w_i are constantly updated through iterations of error minimization through gradient descent.

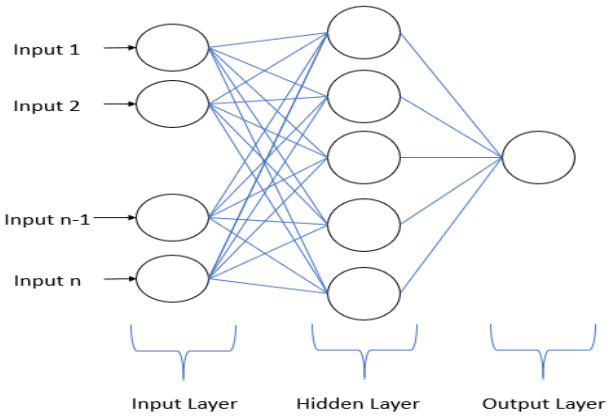

Fig. 2. Artificial Neural Network Architecture

While standalone models have achieved promising results through judicious feature selection, use of balanced training data, and appropriate cost or profit based optimization functions, recent work has focused more on experimenting with hybrid and ensemble framework of ML models to achieve better performance. The rationale behind choosing such strategy is mainly utilization of strengths of each type of model and integrating the output for a relatively unbiased and accurate result.

Hybrid framework combines clustering and classification in a sequential order as shown in Fig. 3. While clustering phase clusters the initial training set, this information is further used as input to the classification phase [24]. Caigny et al. [25] combine decision trees and logistic regression in a hybrid framework with decision trees performing the function of clustering followed by logistic regression for classification. Other studies using similar hybrid approach are [26,27].

In Ensemble framework, a parallel approach is adopted. In this framework, there are several weak base learners or classifiers [28,29] As shown in Fig. 4, each base learner is trained on the churn dataset to predict churn. Once trained, the output of the base learners is either combined as a weighted average, or majority voting is considered to predict the final churn decision. Deng et al. [30] combined ML models namely, Catboost, Lightgbm, and Random Forest in an ensemble framework to predict churn using bank dataset. They achieved an accuracy of 90% with an Area Under Curve (AUC) of over 80%. Few other studies using ensemble framework for churn prediction are [28,31] Ensemble solutions often employ bagging, boosting or stacking strategies. Bagging ensembles consider homogeneous base learners trained independently and simultaneously. While boosting ensembles contain homogeneous base learners combined sequentially and embed optimization function for cost within the boosting procedure itself. Stacking ensembles use a meta classifier to train on the output of bagged or boosted classifiers and finally predict churn.

III. Evaluation of Performance

A few widely used model evaluation measures in the context of churn prediction are described in this section. It is always important to choose an evaluation metric that is closely related to the business goal of an enterprise. Note that widely used evaluation metric, called accuracy, is not recommended in the case of churn datasets where class

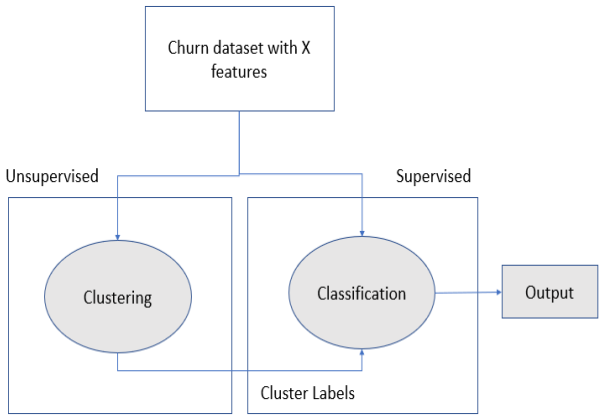

Fig. 3. Hybrid Framework for Churn Prediction

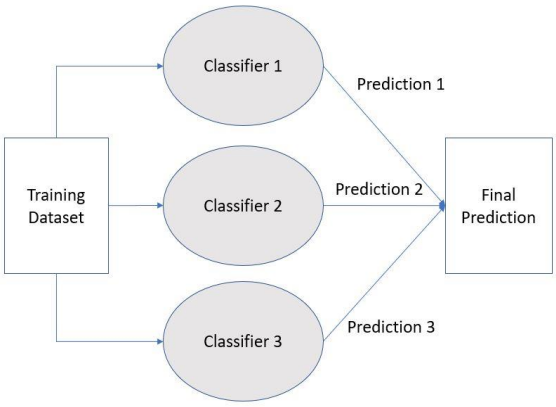

Fig. 4. Ensemble Framework for Churn Classification

imbalance is common. Using accuracy for model evaluation in such datasets can produce misleading and over optimistic results. Few evaluation parameters used in churn classification are derived from a table called confusion matrix shown in Fig. 5. The table is derived post model application on test dataset for which true values are known. The terms TP, TN, FP, FN stand for True Positive, True Negative, False Positive and False Negative, respectively. If churn is denoted as a positive class in churn dataset, TP denotes the number of actual churners who were correctly classified and TN represents the number of non-churners that were correctly classified. On the other hand, FP denotes the number of actual non-churners that were mis-classified by the model and FN represents the number of actual churners that were mis-classified by the model. In Fig. 5 as an example, we see that out of 50 customers, 20 are correctly classified churners or true positives, 10 are correctly classified non-churners or true negatives, 5 are incorrectly classified non-churners or false positives, and lastly, 15 are incorrectly classified churners or false negatives.

A. Sensitivity (Sn)

Sensitivity is defined as the proportion of actual churners that are correctly classified as churners. This metric should be chosen when business goal is solely to identify as many actual churners as possible. In a few studies, sensitivity is also known as recall or true positive rate.

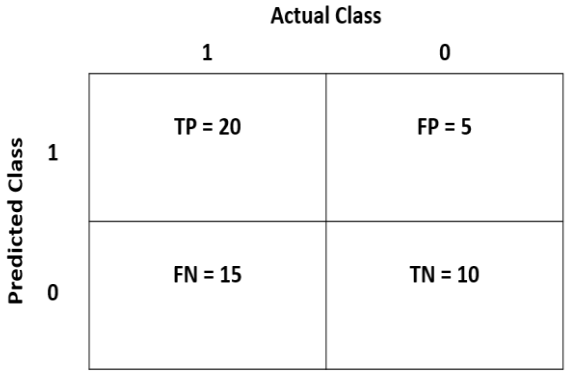

Fig. 5. Confusion Matrix for Churn Classifier

$$Sn = TP/(TP+FN)$$

In the example shown in Fig. 5, we see a sensitivity of 57%

B. Specificity (Sp)

Specificity is defined as the proportion of actual non-churners that are correctly classified. This metric is useful when identification of non-churners is more important to a business. This scenario may arise when an enterprise wants to avoid incurring unnecessary cost by including those customers in promotional retention campaigns who have little risk of churn.

$$Sp = TN/(TN+FP)$$

In the example shown in Fig. 5, specificity stands at 67%.

C. Precision (Pr)

Precision is defined as the proportion of predicted churners that are correctly classified. This metric should be chosen when a company has limited resources to target churners for retention. Precision is also known as positive predictive value.

$$Pr = TP/(TP+FP)$$

As per Fig. 5, precision is estimated to be 80%.

D. F1 Score

F-Measure is the harmonic mean of sensitivity and specificity. This measure is a good indicator of performance when both sensitivity and specificity, are equally important to the business objective.

$$F1\ Score = 2\times(Sn\times Sp)/(Sn+Sp)$$

In example shown in Fig. 5, F1 score is 62%

E. Area Under Receiver Operating Characteristic

Area Under Receiver Operating Characteristic (AUROC) curve or AUC, depicts the probability of a randomly chosen customer who is about to churn, to be ranked higher than a randomly chosen customer who is not going to churn. This is independent of any cut-off that demarcates churners from non-churners for churn classification. The curve is plotted with sensitivity in y-axis, and (1-specificity) in x-axis, for a range of decision threshold values of churn scores, that demarcate a churn customer from a non-churn customer. As shown in Fig. 6, AUC is depicted by the shaded area under the AUC. An ideal AUC extends towards the top left corner of the plot,

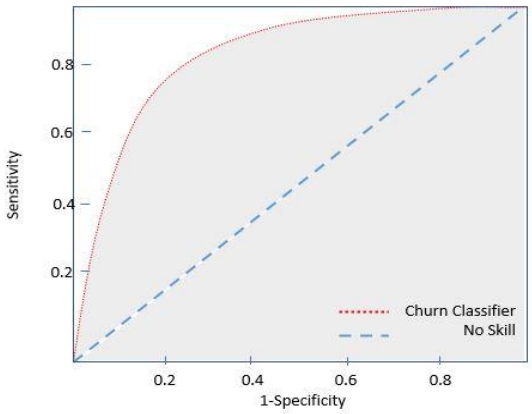

Fig. 6. Area Under Receiver Operating Curve

thereby, increasing the shaded area. This is often the metric used in most studies as it is independent of any decision threshold used for classification. The area under this curve is obtained by applying the trapezoidal rule of integration and is depicted as

$$AUROC = \int Sn\, d(1\text{-}Sp)$$

F. Top Decile Lift (L)

Top decile lift is another evaluation measure that is used in churn prediction. It is derived based on the process described below. A churn classifier assigns a probability score to each customer as an output. This score primarily depicts the probability of churn for that customer. Once this score is available, the customers are sorted in descending order of churn score. Top 10% of this sorted list of customers is selected and percentage of correct churn classifications is calculated. Let this percentage be represented as p. If percentage of churners in our entire customer dataset is c, then top decile lift is given by

$$L = p/c$$

It is understandable that top decile lift is primarily used when company has limited resources and the business goal is to identify as many true churners as possible in an optimized manner.

IV. Result and Discussion

Review of articles related to churn prediction, published in the past ten years, highlights a few important trends. The Appendix section presents a detailed summary of recent studies outlining the datasets used, algorithms applied and best performance achieved within the experimental setting. It should be noted that a fair comparison of "model performance" is challenging since each study is performed under varied experimental conditions. This is due to different dataset size, characteristics, granularity and complexity. It was found that while there are publicly available datasets for churn, most of them are dated and may not be relevant in current time and real-life scenario. Hence, many researchers have experimented with private datasets, mostly in the domain of telecom. Class imbalance was prevalent in almost all datasets. A churn percentage of 35% or less in datasets needed class re-balancing through sampling and weighting strategies during training phase to build a robust classifier. Features used in CCP are mainly demographic snapshots, product portfolio snapshots and product usage snapshots of a customer that is mostly combined with time series analysis techniques like window method to derive discriminatory patterns of a churner over a period.

It is evident that while models like NB, DT, LR and KNN have been very popular in the past, the current emphasis has been on experimentation with hybrid and ensemble techniques. These techniques combine deep learning models together with boosted/bagged versions of algorithms, yielding high performance in terms of AUC, sensitivity, precision and top decile lift. Using nested ensemble framework, which is a special case of ensemble model framework, researchers reported sensitivity as high as 93% along with precision of 96% [28]. This is particularly noteworthy because often high sensitivity comes at the cost of low specificity, and vice versa, in algorithmic iterations. Social network analysis has also shown promising results to predict churn in telecom. A study using this technique combined with time series through multivariate similarity forests reported decile lift of 7.0 [32]. This primarily implies that for a dataset with 11% churn, if top 10% of sorted output of classifier is analyzed, we are able to correctly identify 77% of actual churners. However, the time complexity of this implementation could be high for large datasets.

V. Conclusion and Future Research

The paper discusses popular ML techniques used to address the problem of customer churn classification and also describes the main areas where these techniques have been applied in churn prediction workflow. Problem of class imbalance and dimensionality is prevalent in CCP datasets and researchers have successfully overcome this problem using sampling techniques, cost-sensitive learning and ensemble solutions. The paper also presents few popular evaluation measures used widely in churn prediction. However, it is clear that although there are several evaluation measures for model selection, there is no single metric that universally satisfies all business goals. Hence, the final decision on model selection should be based on comparative study of a combination of performance evaluation measures, that are closely aligned to a business objective. This strategy, together with a good understanding of the financial situation of a company, can yield a favorable business outcome.

Despite many advancements in churn prediction using ML techniques, there are areas for improvement. There is lack of well-defined guidelines on appropriate model evaluation measures to be used in various scenarios of churn. Many articles use accuracy to evaluate model performance in datasets that clearly have class imbalance. In such cases, alternative evaluation parameters like AUC and F1 score are recommended. Above evaluation measures, combined with accuracy, provide a better indication on how robust the classifier is in detecting the rare class. It is evident that most articles that reported high AUC for churn classifiers, either are computationally expensive, or lack interpretability. Interpretability is an aspect that is highly relevant for business adaptation and trust. Hence, developing a high-performing churn classifier that is interpretable and computationally efficient, could be an important, yet challenging direction for future research. Also, the current classifiers in churn prediction are domain-specific. Studies that make an attempt towards developing "domain independent" and generic churn prediction models, witness inconsistent performance across domains. With transfer learning gaining popularity due to its power of reusability and saving training time, we see it as a

popular domain of study in future for churn classification problems.

While most approaches used in CCP are quantitative in nature, where features, related to customer behaviour and product usage are used on a daily or monthly aggregated level, there is lack of research that focuses on information-rich content in customer company interaction instances, like emails, phone calls or customer support chat records. With text analytics gaining immense popularity for its power to extract relevant action-driven qualitative information from unstructured data, its usage will become more and more popular to predict churn in future. This will be all the more relevant in scenarios where popularly used quantitative and demographic features fail to yield satisfactory results.

ACKNOWLEDGMENT

The authors wish to thank anonymous reviewers for their constructive feedback and comments that has helped to improve the paper.

REFERENCES

[1] J. Hadden, A. Tiwari, R. Roy, D. Ruta, Churn-Prediction-Does-Technology-Matter, Int. J. Ind. Manuf. Eng. (2008).

[2] K. Chen, Y.H. Hu, Y.C. Hsieh, Predicting customer churn from valuable B2B customers in the logistics industry: a case study, Inf. Syst. E-Bus. Manag. 13 (2015) 475–494. https://doi.org/10.1007/s10257-014-0264-1.

[3] S. Mitrović, B. Baesens, W. Lemahieu, J. De Weerdt, Churn prediction using dynamic RFM-augmented Node2vec, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer Verlag, 2017: pp. 122–138. https://doi.org/10.1007/978-3-319-71970-2_11.

[4] M. Dadfarnia, A.A. Matinpour, M. Abdoos, Churn Prediction in Payment Terminals Using RFM model and Deep Neural Network, in: 2020 11th Int. Conf. Inf. Knowl. Technol., IEEE, 2020. https://doi.org/10.1109/IKT51791.2020.9345626.

[5] A. Backiel, B. Baesens, G. Claeskens, Predicting time-to-churn of prepaid mobile telephone customers using social network analysis, J. Oper. Res. Soc. 67 (2016) 1135–1145. https://doi.org/10.1057/jors.2016.8.

[6] R. Pagare, A. Khare, Churn prediction by finding most influential nodes in social network, in: 2016 Int. Conf. Comput. Anal. Secur. Trends, IEEE, 2016. https://doi.org/10.1109/CAST.2016.7914942.

[7] L. Calzada-Infante, M. Óskarsdóttir, B. Baesens, Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts, Expert Syst. Appl. 160 (2020). https://doi.org/10.1016/j.eswa.2020.113553.

[8] E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, M. Snoeck, Profit maximizing logistic model for customer churn prediction using genetic algorithms, Swarm Evol. Comput. 40 (2018) 116–130. https://doi.org/10.1016/j.swevo.2017.10.010.

[9] S. Maldonado, J. López, C. Vairetti, Profit-based churn prediction based on Minimax Probability Machines, Eur. J. Oper. Res. 284 (2020) 273–284. https://doi.org/10.1016/j.ejor.2019.12.007.

[10] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, T. Verdonck, Profit driven decision trees for churn prediction, Eur. J. Oper. Res. 284 (2020) 920–933. https://doi.org/10.1016/j.ejor.2018.11.072.

[11] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Syst. Appl. 38 (2011) 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023.

[12] L. Zhao, Q. Gao, X.J. Dong, A. Dong, X. Dong, K- local maximum margin feature extraction algorithm for churn prediction in telecom, Cluster Comput. 20 (2017) 1401–1409. https://doi.org/10.1007/s10586-017-0843-2.

[13] P. Datta, B. Masand, D.R. Mani, Automated Cellular Modeling and Prediction on a Large Scale, Artif. Intell. Rev. 14 (2000) 485–502.

[14] A.Q. Ammar, D. Maheshwari, Churn prediction on huge telecom data using hybrid firefly based classification _ Elsevier Enhanced Reader, Egypt. Informatics J. (2017).

[15] K. Ng, H. Liu, Customer Retention via Data Mining, Artif. Intell. Rev. 14 (2000) 569–590.

[16] C.F. Tsai, Y.H. Lu, Customer churn prediction by hybrid neural networks, Expert Syst. Appl. 36 (2009) 12547–12553. https://doi.org/10.1016/j.eswa.2009.05.032.

[17] B.Q. Huang, T.M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, T. Rashid, A new feature set with new window techniques for customer churn prediction in land-line telecommunications, Expert Syst. Appl. 37 (2010) 3657–3665. https://doi.org/10.1016/j.eswa.2009.10.025.

[18] A. Idris, A. Iftikhar, Z. ur Rehman, Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling, Cluster Comput. 22 (2019) 7241–7255. https://doi.org/10.1007/s10586-017-1154-3.

[19] D.A. Kumar, V. Ravi, Predicting credit card customer churn in banks using data mining, 2008.

[20] B. Zhu, B. Baesens, S.K.L.M. vanden Broucke, An empirical comparison of techniques for the class imbalance problem in churn prediction, Inf. Sci. (Ny). 408 (2017) 84–99. https://doi.org/10.1016/j.ins.2017.04.015.

[21] N. Glady, B. Baesens, C. Croux, Modeling churn using customer lifetime value, Eur. J. Oper. Res. 197 (2009) 402–411. https://doi.org/10.1016/j.ejor.2008.06.027.

[22] D.S. S., D.A. Basar, D.H. Wang, Artificial Neural Network Based Power Management for Smart Street Lighting Systems, J. Artif. Intell. Capsul. Networks. 2 (2020). https://doi.org/10.36548/jaicn.2020.1.005.

[23] E. Sivasankar, J. Vijaya, Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network, Neural Comput. Appl. 31 (2019) 7181–7200. https://doi.org/10.1007/s00521-018-3548-4.

[24] [24] D.J.I.Z. Chen, D.S. S., Social Multimedia Security and Suspicious Activity Detection in SDN using Hybrid Deep Learning Technique, J. Inf. Technol. Digit. World. 2 (2020). https://doi.org/10.36548/jitdw.2020.2.004.

[25] A. De Caigny, K. Coussement, K.W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, Eur. J. Oper. Res. 269 (2018) 760–772. https://doi.org/10.1016/j.ejor.2018.02.009.

[26] R. Manivannan, R. Saminathan, S. Saravanan, An improved analytical approach for customer churn prediction using Grey Wolf Optimization approach based on stochastic customer profiling over a retail shopping analysis: CUPGO, Evol. Intell. (2019). https://doi.org/10.1007/s12065-019-00282-x.

[27] X. Hu, Y. Yang, L. Chen, S. Zhu, Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network, in: 2020 IEEE 5th Int. Conf. Cloud Comput. Big Data Anal., IEEE, 2020. https://doi.org/10.1109/ICCCBDA49378.2020.9095611.

[28] M. Ahmed, H. Afzal, I. Siddiqi, M.F. Amjad, K. Khurshid, Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry, Neural Comput. Appl. 32 (2020) 3237–3251. https://doi.org/10.1007/s00521-018-3678-8.

[29] Q.F. Wang, M. Xu, A. Hussain, Large-scale Ensemble Model for Customer Churn Prediction in Search Ads, Cognit. Comput. 11 (2019) 262–270. https://doi.org/10.1007/s12559-018-9608-3.

[30] Y. Deng, D. Li, L. Yang, J. Tang, J. Zhao, Analysis and prediction of bank user churn based on ensemble learning algorithm, in: 2021 IEEE Int. Conf. Power Electron. Comput. Appl., IEEE, 2021. https://doi.org/10.1109/ICPECA51329.2021.9362520.

[31] M. Malyar, M.V. Mykola Robotyshyn, M. Sharkadi, Churn Prediction Estimation Based on Machine Learning Methods, in: 2020 IEEE 2nd Int. Conf. Syst. Anal. Intell. Comput., IEEE, 2020. https://doi.org/10.1109/SAIC51296.2020.9239230.

[32] M. Óskarsdóttir, T. Van Calster, B. Baesens, W. Lemahieu, J. Vanthienen, Time series for early churn detection: Using similarity based classification for dynamic networks, Expert Syst. Appl. 106 (2018) 55–65. https://doi.org/10.1016/j.eswa.2018.04.003.

[33] N. Alboukaey, A. Joukhadar, N. Ghneim, Dynamic behavior based churn prediction in mobile telecom, Expert Syst. Appl. 162 (2020). https://doi.org/10.1016/j.eswa.2020.113779.

[34] A. Amin, B. Shah, A.M. Khattak, F. Joaquim, L. Moreira, G. Ali, A. Rocha, S. Anwar, Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods, Int. J. Inf. Manage. (2019). https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.08.015.

[35] S. Mitrović, B. Baesens, W. Lemahieu, J. De Weerdt, On the operational efficiency of different feature types for telco Churn prediction, Eur. J. Oper. Res. 267 (2018) 1141–1155. https://doi.org/10.1016/j.ejor.2017.12.015.

[36] R. Yu, X. An, B. Jin, J. Shi, O.A. Move, Y. Liu, Particle classification optimization-based BP network for telecommunication customer churn prediction, Neural Comput. Appl. 29 (2018) 707–720. https://doi.org/10.1007/s00521-016-2477-3.

APPENDIX

Year	Study	Domain	Dataset Specs	Learning Technique	Best Performance
2020	[28]	Telecom	UCI: 5,000 customers (14% churners) SATO: 1,000 customers (50% churners)	Boosted Stacked Learner and Bagged Stacked Learner of DT, LR, NB, ANN and KNN	Accuracy: 98.4% Precision: 96.1% Sensitivity: 92.5%
2020	[7]	Telecom	13,454 customers (10.2% churners)	Social network analysis and Similarity Forests	AUC: 84.3% Lift: 6.2 AUPR: 48.2%
2020	[33]	Telecom	15,00,000 customers (12% churn)	Convolutional Neural Network (CNN), Long Short Term Memory (LSTM) and Recurrent Neural Network (RNN)	AUC: 91.4% F1 score: 54.2% Lift: 5.1
2020	[9]	Nine real world churn datasets from Telecom including Duke	Duke1: 93,893 customers (50% churn) Duke2: 20,406 customers (2% churn)	ProfMEMPM, ProfLogit, ProfTree	AUC: 94.0%
2020	[10]	Telecom	889 customers (31% churn)	ProfTree	AUC: 82.4%
2019	[34]	Telecom	Dataset1: 3,333 customers (14% churn) Dataset 2: 18,000 customers (12% churn)	NB, KNN, Gradient Boosted Decision Tree (GBDT), Single Rule Induction (SRI) and ANN	Accuracy: 54.1% Sensitivity: 16.7% Specificity: 91.3%
2019	[18]	Telecom	Duke Cell2Cell: 40,000 customers (50% churn) Orange Telecom: 50,000 customers (7% churn)	PSO, AdaBoost	AUC: 91.0% Sensitivity: 87% Specificity: 89%
2019	[23]	Telecom	Duke: 1,00,000 customers (49% churn)	DT, KNN, SVM, NB, ANN, PPFCM-ANN	Accuracy: 97.9% Precision: 100%
2019	[29]	Search Ads	66,059 customers (37% churn)	GBDT	AUC: 84.1%
2019	[26]	E-commerce	12,730 customers (25% churn)	CUPGO, ACO, PSO	Accuracy: 89.3%
2018	[25]	Bank, Newspaper, Telecom, Energy	Sample size range: 3,827 - 6,31,627 customers (2% - 29% churn)	Logit Leaf Model (LLM), LR , RF and DT	AUC: 88.7% Top decile lift: 5.4
2018	[32]	Telecom	Dataset1: 86,000 customers (11% churn) Dataset2: 15,00,000 customers (10% churn) Dataset3: 1,70,000 customers (4% churn)	Multivariate similarity forests	AUC: 90% Top decile lift: 7.0
2018	[35]	Telecom	Dataset1: 7% churn Dataset2: 2% churn	RF, LR	AUC: 73.6%
2018	[36]	Telecom	CMCC Dataset: 1,00,000 customers	PBCCP	Sensitivity: 60%