|| Volume 6 || Issue 2 || February 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

CUSTOMER CHURN PREDICTION USING NLP AND MACHINE LEARNING: AN OVERVIEW

Farhad Shaikh[1], Brinda Pardeshi[2], Ajay Jachak[3], Akash Bendale[4], Nandakishor Sonune[5], Prof Mangal Katkar[6]

Department of Information Technology, Dhole Patil College of Engineering, Wagholi, Pune, India[1,2,3,4,5,6]

Abstract:- Churn prediction system using classification as well as clustering techniques to classify churn customers and the reasons behind the churning of telecom customers. In telecom industry should we generate large amount of data on daily basis, it is very tedious task to mine such a kind of last data using specific data mining techniques, while hard to interpret the prediction on classical techniques. Various researchers already described search a work to eliminate churn from large data sets fusion static as well as dynamic approaches, but still such systems are facing many problems actual identification of churn. Sometime such telecommunication data may be containing some churn and, it is much necessary to identify search problems. To successful identification of churn from large data is providing effectiveness to customer relationship management (CRM). In this paper we proposed churn identification as well as prediction from large scale telecommunication data set using Natural Language Processing (NLP) and machine learning techniques. First system deals with strategic NLP process which contains data preprocessing, data normalization, feature extraction and feature selection respectively. Feature extraction techniques have been proposed like TF-IDF, Stanford NLP and occurrence correlation techniques. Where machine learning classification algorithms are has used to train and test the entire module. Finally experiment analysis shows performance evaluation of proposed system and evaluate with some existing systems.

Keywords: *Natural language processing, churn prediction, machine learning, telecom industry, customer relationship management*

I INTRODUCTION

In today's computer environment writing comments to churn more frequently while voice mail plan customers can dispose to churn less frequently. Customers with four or more customer service calls churn as often as other customers churn more than four times. We calculate the average churn rate during model training using different machine learning approaches and evaluate the for testing. To maximize the organization's sales, as we suggested in our study, predicting accuracy churn is very critical. Rest of the paper is organized as follows. Section 2 gives brief overview of latest research, section 3 explains proposed work, system overview; datasets description section 4 observations Section 5 research contribution Section 6 application of churn prediction systems 7 concludes the paper section 8 future works.

1.1 Background

According to [1] Clustering algorithms are clustered input functions with k-means and fuzzy c-means to position subscribers in independent, distinct classes. Using these groups the Adaptive Neuro Fuzzy Inference Framework (ANFIS) is implemented to construct a predictive model for successful churn management. The first step towards prediction starts with the parallel classification of Neuro soft. FIS then uses the outputs of Neuro fuzzy classifiers as feedback to settle on the behaviors of the churners. Progress metrics can be used to identify issues of inefficiency. Churn reduction indicators are concerned with the facilities, processes and performance of customer support network. Versatility of GSM numbers is a critical criterion for churner's determination.

In System [2] a current collection of software to increase the standard of detecting possible churners. The roles are extracted from request information and client accounts and are classified as deal, request pattern and call pattern adjustments overview functionality. The characteristics are evaluated using two probabilistic data mining algorithms from Naïve Bayes and Bayesian Network, and their findings compared to those obtained by the use of C4.5

decision tree, an algorithm widely used in many classification and prediction tasks. Among other reasons these have led to the possibility that consumers will quickly turn to competitors. One of the techniques that can be used to do this is to improve churn prediction from large amount of data with extraction in the near future.

According to [3] formalization of time-window of the collection process, coupled with literature review. Second, by expanding the duration of consumer events from one to seventeen years using logistic regression, classification trees and bagging together with classification trees, this analysis analyzes the rise in churn model accuracy. The practical result is that researchers may substantially reduce the data-related pressures, such as data collection, preparation, and analysis. The price customers are expected to pay depends on the length and the pro-motional nature of the subscription. The newspaper business is sending a letter telling them that the service is ending. Then ask them if they want to renew their subscription, along with guidance on how to do that. Customers are unable to cancel their subscription and have a grace period of four weeks once they have subscribed lapsed.

According to [4] the most efficient consumer engagement strategies can be used to high the client satisfaction level efficiently. The study indicates a Multilayer Perceptron (MLP) neural network method to estimate client turnover in one of Malaysia's leading telecommunications firms. The results were contrasted with the most traditional churn prediction strategies such as Multiple Regression Analysis and Analyzing Logistic Regression. The maximal neural network architecture includes 14 input nodes, 1 concealed node and 1 output node with the learning algorithm Levenberg Marquardt (LM). Multilayer Perceptron (MLP) neural network approach to predict client churn in one of the leading telecommunications companies in Malaysia compared to the most common churn prediction techniques, such as Multiple Regression Analysis and Logistic Regression Analysis.

In system [5] on creating an efficient and descriptive statistical churn model utilizing a Partial Least Square (PLS) approach focused on strongly associated intervals in data sets. A preliminary analysis reveals that the proposed model provides more reliable results than conventional forecast models and recognizes core variables in order to better explain churning behaviors. Additionally, network administration, overage administration and issue handling approaches are introduced in certain simple marketing campaigns and discussed.

Burez and Van den Poel [6] Unbalance data sets studies in churn prediction models, and contrasts random sampling performance, Advanced Under-Sampling, Gradient Boosting Method, and Weighted Random Forest. The concept was evaluated using Metrics (AUC, Lift). The study shows that the methodology under sampling is preferable to the other techniques evaluated.

Gavril et al. [7] Describes an innovative data mining method to explain the broad dataset type of consumer churn detection. About 3500 consumer details is analyzed based on incoming number as well as outgoing input call and texts. Specific machine learning algorithms were used for training classification and research, respectively. The system's estimated average accuracy is about 90 percent for the entire dataset.

He et al. [8] in a with approximately 5.23 million subscribers, a major Chinese telecommunications corporation developed a predictive model focused on the Neural Network method to address the issue of consumer churn. The average degree of precision was the extent of predictability of 91.1%.

Idris [9] Suggested a genetic engineering solution to modeling AdaBoost-churning telecommunications problems. Two Standard Data Sets verified the series. With a precision of 89%, one from Orange Telecom and the other from cell2cell and 63% for the other one.

Huang et al. [10] the customer churn studied on the big data platform. The researchers ' aim was to show that big data significantly improves the cycle of churn prediction, based on the quantity, variety and pace of the data. A broad data repository for fracture engineering was expected to accommodate data from the Project Support and Business Support Department at China's biggest telecommunications firm. AUC used the forest algorithm at random and assessed.

Makhtar et al. [11] proposed the usage of rough set theory in telecom as a statistical concept of churn. As stated in this post, the Rough Set classification algorithm has outperformed the other algorithms.

Different work has Tackled the problem of unbalanced data sets where the churned customer groups are below the active customer levels, rendering churn estimation a big

concern. Amin et al.[12] compared the issue of the telecom churn forecast with six separate oversampling techniques. The findings revealed that the other algorithms (MTDF and rules creation dependent on genetic algorithms) exceeded the others. fantastic screening algorithms.

II LITERATURE SURVEY

We have surveyed several recent trends in this field and tabulated the techniques, datasets used and research gap in Table 1

Table1.Brief overview of survey

No	Technique	Dataset	Extracted Features	Research Gap
1	x-Means clustering algorithms and Neuro Fuzzy algorithm [1]	GSM operation data, 24,900 customers 22 attributes Turkey dataset	Some value-added services and some values added services	System reflects good accuracy on structured dataset only.
2	Naïve Bayes, Decision Tree[2]	European operator 106,405 customers 112 attributes	Contract, usage pattern patterns, and calls pattern	High error rate to detect actual churn due to redundant features.
3	Neural network, Regression [3]	Unknown 129,892 customers 113 attributes	Demographic, Value added, usage pattern	Heterogeneous dataset tedious to handle in similar patterns environments.
4	Neural network, Regression [4]	Unknown, 169 customers 10 attributes	Demographic, Billing data, usage pattern, customer relationship	High space complexity generate in each layer
5	Stepwise variable selection partial least squares [5]	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral information, Customer care and demographics	Redundant features should be generating high error rate.
6	Artificial Neural Network [6]	ML Dataset of UCI 2,427 user's information with 20 attributes	Demographics, Usage pattern, Value added services	It works only define statically parameters.
7	Binomial logistic regression model [7]	Iranian telco operator 3150 customers 15 attributes	Demographic, call usage pattern, customer care service	Language influence should be generate irrelevant features vector.
8	Generalized additive models (GAM) [8]	Belgian 134, 120 customers 27 attributes	Demographic Usage patter, bill and payment	High error rate during unknown text prediction.
9	Logistic regression Decision tree [9]	Polish mobile operator 122098 customers 1381 attributes	Demographic, call data records, customer care services	Its works only synthetic data only and high data reduction rate.
10	Decision tree as well as machine learning [10] algorithms has used.	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral data, of customer care and feature information	Behaviors information generate the churn possibility sometime it generate false ratio.

III PROPOSED WORK

In this research we proposed churn prediction from large scale data, system initially deals with telecommunication synthetic data set which contains some imbalance meta data. To apply data preprocessing, data normalization, feature extraction as well as feature selection respectively. During this execution some Optimization strategies have been used to eliminate redundant features which sometimes generate high error rate during the execution. The below figure 1 shows propose system execution for training and testing. After completion both phases system describe classification accuracy for entire data set.

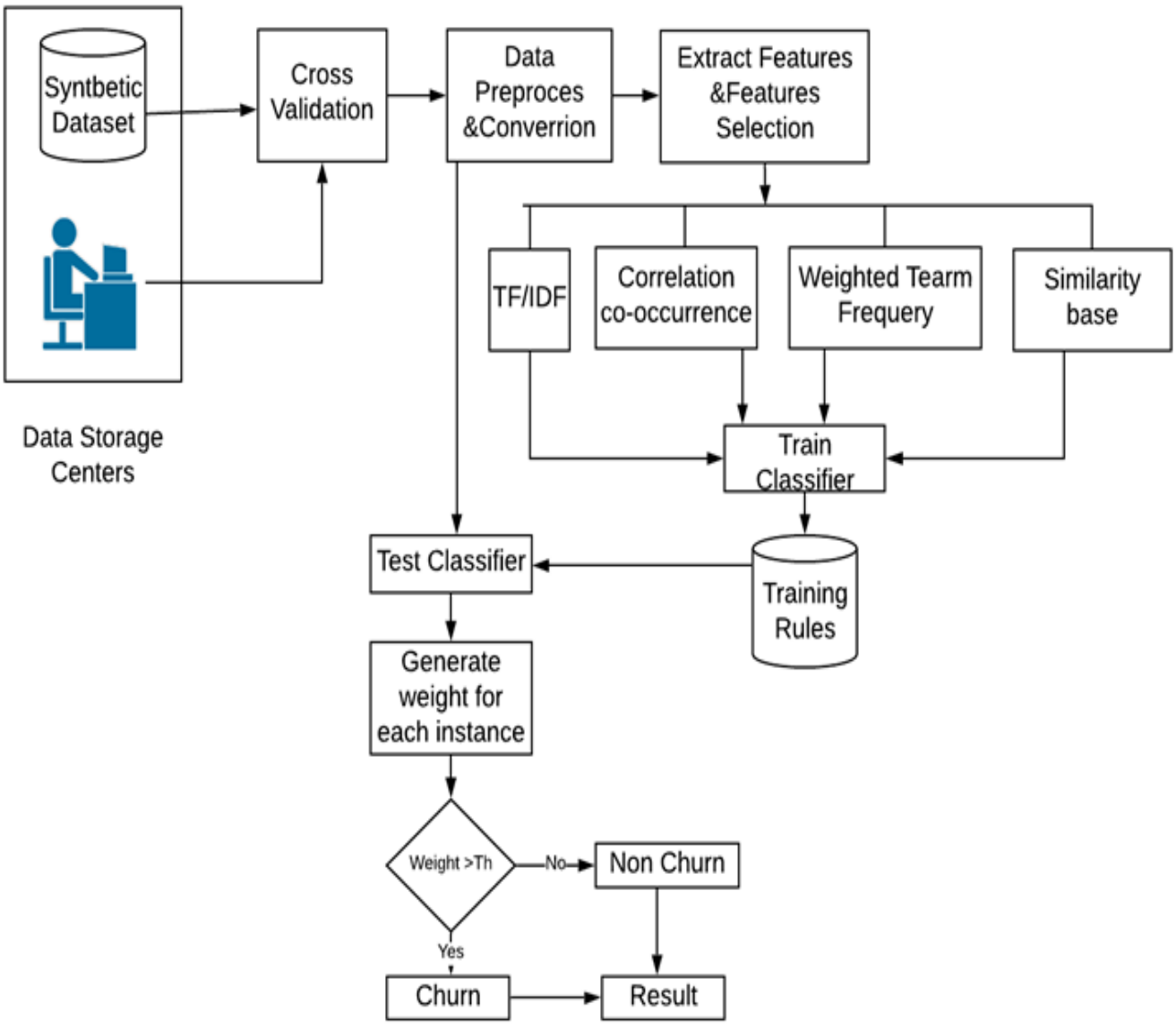

Fig.2 Proposed system overview

3.1 System overview

The aim of this such kind of research in the telecommunications industry is to help businesses make more profit. Telecom companies have become known to forecast turnover as one of the most important sources of income. Therefore, this research was aimed at building a system in the Telecom Company that predicts customer churn. Such prediction models will achieve high AUC values. The sample data was divided into 70% for training and 30% for testing to evaluate and develop the model. We chose 10-fold cross-validation for evaluating and optimizing hyper parameters. We used engineering tools, effective function transformation and selection approach. Making the interface fit for machine learning algorithms. Another concern was also found: the data was not balanced. Only about 5% of the entries are customers ' churn. A problem has been solved by under-sampling or using trees algorithms that are not affected by this issue. In detecting the churn in large data and providing accurate prediction, our different classifiers can be more accurate. This work contributes to suggesting a supervised approach to the extraction of dimensional categories, selecting suitable characteristics and avoiding duplication by measuring correlation between characteristics. The results obtained show that there is a comparatively higher f-score in the weighted frequency of the term with the correlation process. In this regard, selecting features using weighted word frequency is more important. The overlap between features in a category of aspect is avoided by measuring the association.

Datasets used

We used a telecom sector dataset available on Kaggle.com for prediction of churn customers as it contains data of both the customers i.e. churn as well as no churn. It contains around 21 attributes and 7043 rows with class label as churn as yes or no. The class label is the last attribute defined in numeric value like 1 and 2.

IV OBSERVATIONS

- The rule generation provides better classification accuracy than other classification techniques which is define in [12].
- System [7] provide accuracy for churners as well as non-churners model around 99.10% and for BN algorithm it should be around, 99.55% as well as MLP, and 99.0 0% for SVM respectively.
- Hybrid method has used for churn prediction which generates around 90% classification accuracy in [1]
- Neural Network has used for classification as well as accuracy prediction in [4] which provides around 91.28% accuracy on large dataset

V RESEARCH CONTRIBUTION

- Evaluate the system with various kind of heterogeneous dataset from client reviews and predict the accuracy.
- To implement the proposed system with various feature extraction as well as feature selection techniques which will reduce the redundant feature set.
- To develop the system with various machine learning algorithms for increased the churn prediction accuracy.

VI APPLICATIONS

- BPO centers churn prediction systems.
- Service application churns prediction systems.
- Customer behaviors mapping system using churn prediction.

VII CONCLUSION

This research basically focuses on identification and detection of churn customers from large telecommunication data set, state of art describes churn prediction systems which is developed by various researches. Many systems still facing a linguistic data conversion issues, which may occur high error rate during the execution. Many researchers have been proposed Natural Language Processing (NLP) techniques as well as different machine learning algorithms such a combination probably generate high accuracy when data is structured. Whenever any machine learning algorithm deals with such a kind of system it is mandatory to evaluate or validate entire data set with even sampling technique which eliminate data imbalance problem and provide consistent data flow for prediction.

VIII FUTURE WORK

To implement a proposed system with large heterogeneous dataset in Hadoop distribution File System (HDFS) will be future interesting task of this system,

REFERENCES

1 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.

2 Kirui, Clement, et al. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining." International Journal of Computer Science Issues (IJCSI) 10.2 Part 1 (2013): 165.

3 Ballings, Michel, and Dirk Van den Poel. "Customer event history for churn prediction: How long is long enough?." Expert Systems with Applications 39.18 (2012): 13517-13522.

4 Ismail, Mohammad Ridwan, et al. "A multi-layer perceptron approach for customer churn prediction." International Journal of Multimedia and Ubiquitous Engineering 10.7 (2015): 213-222.

5 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.

6 Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.

7 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.

8 He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.

9 Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.

10 Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.

11 Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

12 Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57