Customer Churn Prediction System using Machine Learning

Article *in* International Journal of Advanced Science and Technology · January 2020

CITATION
1

READS
482

2 authors, including:

Amit Savyanavar
MIT-WPU
31 PUBLICATIONS **49** CITATIONS

SEE PROFILE

International Journal of Advanced Science and Technology
Vol. 29, No. 5, (2020), pp. 7957-7964

Customer Churn Prediction System using Machine Learning

Vrushabh Jinde and Prof. Amit Savyanavar
School of Computer Engineering and Technology
MIT World Peace University,Kothrud, Pune
School of Computer Engineering and Technology
MIT World Peace University, Kothrud, Pune
jindevrushabh@gmail.com, amit.savyanavar@mitwpu.edu.in

Abstract

As the technology is evolving day by day, businesses are also evolving and changing their strategies to gain more profit. They are using these developed technologies along with their business experience to get succeed. Especially, in the field of telecommunication where the service providers have to face a large number of customer's data and this gets more difficult as they cannot focus on each individual customer's need with respect to services. And if customers are not getting their requirements fulfilled they switch their service provider. So to overcome such problem the idea of churn prediction system comes in picture. Service provider companies are using customer's data in order to understand and improve Customer Relationship Management (CRM). This paper proposes such various churn prediction systems developed by researchers which uses machine learning approaches that will help telecommunication industries to understand their customer's need in order to fulfil their requirements. 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: Customer relationship management, Churn prediction system, Machine learning, Telecommunication industry, Retention, Natural language processing

1. Introduction

Customers are considered as the most important assets in any industry since in most of the scenarios the profit in business is directly proportional to the number of customers. The telecommunication sector is made up of companies that make communication possible on a global scale, whether it is through the phone or the Internet, through cables or wirelessly. These companies created the infrastructure that allows data in words, voice, audio or video to be sent anywhere in the world. The largest companies in the sector are telephone (both wired and wireless) operators, satellite companies, cable companies, and internet service providers. But as per the modern technologies invented, most of the wired communication is converted into wireless communication. So the number of customers using wireless communication media is increasing day by day. Even though majority customers are using wireless way of communication they still can be classified into two types, based on their way of subscription as post-paid and pre-paid customers. Post-paid customers first uses different set of services and then service provider will charge them for the services at end of the month or a year. And they are bound to service provider on monthly or yearly basis. Whereas Pre-paid customers have to first purchase the required

ISSN: 2005-4238 IJAST

services from service provider in order use them. And they are not bounded with service providers like post-paid users.

Customer churn is defined as the customer who terminates the services of their service provider or switches to another service provider. It really decreases the profit margin for such service provider if it loses its customers. To avoid such problem of losing the customers churn prediction system is used. Telecom industries should know the reasons behind the churning of customers so that they can avoid it in the future. For this they use the customer's behavior data from the day he/she subscribed to the services and using them till present. The main reasons for churn are that customers not satisfied with the services are poor network facilities, expensive services, plans are suitable, bad customer help support. [1]

In today's computer environment every business has their own website which is one of the way to advertise, expand their business and also keep them connected with their customers. And websites can be helpful in both the ways since customer can express their experience while using services as a form of feedback or review. But till now the existing churn prediction systems were using historical data of the customers and identified churn customers using machine learning. Feedbacks of the customers could play crucial role in understanding their needs from a service provider. And later, the analyst can use this feedback data and processes it with Natural Language Processing (NLP), it can give valuable insights through their sentiments.

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 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 Overview of Machine Leaning

Machine learning is a method of data analysis which creates analytical model building. Using algorithms that iteratively learn from data, machine learning allows systems to explore hidden patterns without being explicitly programmed where to look. There are three types of machine learning approaches: unsupervised, semi-supervised, and supervised. Supervised learning is the machine learning deals with labelled data and the hidden patterns can be found from it. Unsupervised learning deals with unlabeled data and the hidden patterns from can be found from it. Semi-supervised learning is a combination of supervised approach and unsupervised approach. Basically, it deals with both the type of data labelled as well as unlabeled. But in this case the labelled data is very small in size as compared to unlabeled data.

1.2 Natural Language Processing

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding. By using NLP, one can train a machine learning model which can process the text data and interpret the sentiment behind it. In general sentiments can be divided into three types as positive sentiment, neutral sentiment and negative sentiment. By analyzing the sentiments one can understand the persons emotion whether he or she is happy or sad. [2]

ISSN: 2005-4238 IJAST

Rest of the paper is organized as follows. Section 2 gives the background of work done in this field by various researchers, Section 3 explains proposed work, and Section 4 gives system overview, section 6gives observations, Section 7 research contribution and Section9 concludes the paper.

2. Background

In this paper they proposed a customer churn prediction system using Random Forest algorithm. And this system not only predicts churn customer but also identifies the reason behind it. They have used South Asian telecom company's data in for prediction and it show great accuracy. But since they used Random Forest algorithm it is very time consuming. Also number features that has been used are very large. [1]

Karahoca Adem, and Dilek Karahoca [3] proposed a clustering method in which 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.

Kirui, Clement, et al. [4] created a system in which data mining techniques has been applied on statistical data and probabilistic model is developed. 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 Ballings, Michel, and Dirk Van den Poel [5] 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 analyses 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 [6] 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

ISSN: 2005-4238 IJAST

most traditional churn prediction strategies such as Multiple Regression Analysis and Analysing 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 [7] 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 behaviours. Additionally, network administration, overage administration and issue handling approaches are introduced in certain simple marketing campaigns and discussed.

Burez and Van den Poel [8] 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. [9] Describes an innovative data mining method to explain the broad dataset type of consumer churn detection. About 3500 consumer details is analysed 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. [10] 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 [11] 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. [12] 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. [13] 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. [14] 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.

ISSN: 2005-4238 IJAST

2.1 Literature Survey

Table 1.Brief overview ofsurvey

Reference No.	Year of Publication	Technique Used	Dataset used	Extracted Features	Research Gap
[1]	2019	Random Forest Algorithm	South Asian GSM's data	Calling services, Value added services	Time consuming, High error rate
[15]	2016	Gradient Boosting, Decision Tree, Support Vector Machine, Random Forest, K-NN, Ridge Regression and Logistic Regression	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral information, Customer care and demographics	Behaviors information generate the churnpossibility sometime it generate false ratio.
[6]	2015	Neural network, Regression	Unknown, 169 customers 10 attributes	Demographic, Billing data, usage pattern, customer relationship	High space complexity generate in each layer
[4]	2013	Naïve Bayes, Decision Tree	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.
[5]	2012	Neural network, Regression	Unknown 129,892 customers 113 attributes	Demographic, Value added, usage pattern	Heterogeneous dataset tedious to handle in similar patterns environments.
[3]	2011	x-Means clustering algorithms and Neuro Fuzzy algorithm	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.

3. Proposed work

In this research we proposed churn prediction system from large scale data, system initially deals with telecommunication synthetic data set which contains some imbalance Meta data. To apply data pre-processing, 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.

ISSN: 2005-4238 IJAST

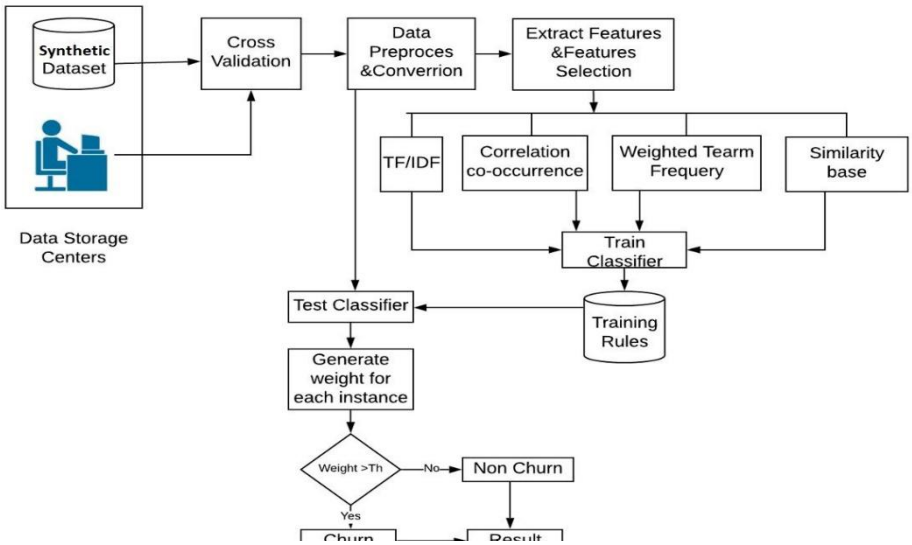

Figure 1. Proposed system overview

4. 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.

5. 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.

6. 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

ISSN: 2005-4238 IJAST

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.

7. 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.

8. Applications

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

9. 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.

Acknowledgments

Sincere thanks to my guide Prof. Amit Savyanavar Sir for providing immense support in this research. Also, I thank Dr. Siddhivinayak Kulkarni Sir for helping me the best way possible.

References

[1] Irfan Ullah, Basit Raza , Ahmad Kamran Malik , Muhammad Imran, Saif Ul Islam And Sung Won Kim"A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector" IEE ACCESS.2019.2914999, VOLUME 7, 2019
[2] Xing Fang, Xing Fang Xing Fang "Sentiment analysis using product review data."
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.
[9] 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.

ISSN: 2005-4238 IJAST

[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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
[15] V. Umayaparvathi, K. Iyakutti,, "Attribute Selection and Customer Churn Prediction in Telecom Industry", Proceedings of the IEEE International Conference On Data Mining and Advanced Computing, 2016 (to be appeared).
[16] Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying, "Customer churn prediction using improved balanced random forests", Expert Systems with Applications 36 (2009) 5445–5449.

Authors Profile

Vrushabh Jinde, completed **B.E.** from Pune University and currently pursuing**M.Tech** from MIT World Peace University, Pune.
Research Interests: Machine Learning, Data Analysis, Deep Learning, Cloud Computing.

Prof. Amit Savyanavar, completed **M.E** from Pune University Campus and pursuing **Ph.D**in Computer Science & Engineering. Currently Faculty at MIT World Peace University Pune.
Research Interests: Mobile Computing, Distributed Systems.

ISSN: 2005-4238 IJAST