

Customer Churn Prediction using Machine Learning

Kalyani Nivrutti Mahajan¹, Dr. Ram Joshi² and Prof. K. V. Deshpande³

¹⁻³JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, Maharashtra, India- 411033

Email: kalyanimahajan.1398@gmail.com, rbjoshi_it@jspmrscoe.edu.in, kvdeshpande_comp@jspmrscoe.edu.in

Abstract—The basic random forests are combined with a sampling method, cost sensitive learning, and improved balanced random forests (IBRF), which we utilize to predict client churn. We may repeatedly learn the best features of IBRF by changing the distribution of the classes and toughening the punishment for misclassifying the minority class. They are demonstrated to considerably increase prediction accuracy when compared to other methods such as artificial neural networks, decision trees, and class-weighted core support vector machines when applied to a database of credit debt clients from an anonymous commercial bank in China (CWC-SVM). To evaluate these algorithms' qualities, they are compared and graded. The sampling plan and data processing are explained in great detail. This article demonstrates how machine learning handles such Random Forest can effectively forecast which customers are mostly will leave the company or organization in the future by using data such as personal information, bank information, etc. The end goal of churn prediction is to understand when exactly a customer is expected to churn so from there, organization can prioritize at-risk customers which are valuable for growth and execute personalized strategies that to reengage those customers again.

Index Terms— Random forest classifier, Support vector machine, Prediction, Churn, Machine learning.

I. INTRODUCTION

This essay will examine customer churn, one of the most significant problems now affecting businesses and a popular topic in CRM. Customer churn, or the likelihood that customers will stop journey of doing work with a company within a predetermined period of time, has become a significant issue for many businesses, including those in the publishing, investment, insurance, electric utility, health care, credit card, banking, Internet, telephone, and cable services sectors. Customer turnover, of course, has a direct impact on customer retention rates and, thus, a customer's lifetime worth to a business.

It is easy to discover that devoted consumers contribute the majority of a company's earnings and that acquiring new customers is more costly than keeping old ones by looking at the present of a customer's lifetime profit to the business. So, it's essential for businesses to keep up their relationships with current clientele.

II. MOTIVATION

We may choose which consumers to target and the maximum benefit price at which the investment will still be worthwhile using the Churn Prediction Model.

The SaaS and IT companies with the great performance and revenue growth also had the highest retention rates and the lowest net revenue churn, according to a McKinsey analysis.

III. LITERATURE SURVEY

[1]Paper : Customer Churn Prediction by Classification Models in Machine Learning

Author : Heng Zhao, Xumin Zuo , Yuanyuan Xie

Publication Year : 2022

Description:

The classification problem in machine learning is used to solve and mark various problems. Machine Learning Classification is useful method to realize customer churn prediction. Through classification methods based on machine learning , customer churn can be predicted accurately so to provide solution to these companies. In this paper random forest and decision tree models are used to predict customer churn using data provided by chemical company from 2012 to 2020.

[2]Paper : Machine Learning Based Customer Churn Prediction In Banking

Author : Manas Rahman, V Kumar

Publication Year : 2020

Description:

The number of service providers are increasing in each piece of business. Currently , there is no shortage of ways for customers in banking origin when choosing where to put their money. As a result , customer churn and engagement has become trending issues in bank. In this paper , through machine learning method prediction of the customer churn is done. The KNN, SVM, Decision Tree, and Random Forest classifiers are used in this paper. As a result, the use of the Random Forest model after oversampling is better compared to other models in terms of prediction of correctness.

[3]Paper : Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier

Author : V.Geetha A. Punitha A.Nandhini

Publication Year : 2020

Description:

In today's world, data is become a crucial component in every industry. In this, data on the telecom sector is gathered,

and the unclassified data is divided into churn and non-churn clients. Consumers who regularly use the same resource

signals are said to churn, whereas non-churn customers use resources in accordance with the services offered by the

specific company. The LDT and UDT algorithms used in the current system teach the system blindly with too many

attributes that are not required for computation . As a result, the system requires a lot of time to train and only achieve

a performance of 84 percent with respect to accuracy. . However, this level of performance is insufficient for a company

to deliver reliable services. Therefore, in order to fix this issue in the current system, we propose the system be equipped

with effective method known as Random Forest Classifier and Support Vector Machine. These methods choose the key

attributes that improve system performance, and by putting these two algorithms into practice, we can achieve an

efficiency of about 95%. Because of the company's increased performance efficiency, the telecommunications business

may be sustained by providing the right services and keeping customers who do not churn.

[4] Paper : Analysis of Machine Learning Methods for Churn Prediction and Factor Identification in the Telecom Sector: A Churn Prediction Model Using Random Forest

Author : Irfan Ullah , Basit Raza

Publication Year : 2019

Description:

Due to a sizable customer base, the telecom industry generates enormous amounts of data every day. Business analysts and decision-makers stressed that acquiring new clients is more costly than keeping the ones you already have. Business analysts and CRM analysts need to understand the causes of customer churn as well as behaviour patterns from the data on existing turnover clients. In order to identify churning consumers and explain the causes of customer churn in the telecom industry, this research develops a churn prediction model

that makes use of classification and clustering algorithms. The suggested model first uses classification methods to categorize customer turnover data, and the Random Forest (RF) algorithm did well with 88.63% of correctly categorized occurrences.

[5]Paper : The gamma CUSUM chart method for online customer churn prediction

Author : Ssu-Han Chen

Publication Year : 2016

Description:

In this paper individual customer churn prediction mechanism is invented. This term is based on gamma CUSUM chart and hierarchical Bayesian model. Integration of Inter-arrival time and recency are used to improve TPR and reduce ATS. Customer churn in online organization is difficult to manage because customers are so biased. The data in this analysis having source from an online dating website in Taiwan. The results show that the accuracy rate for a gamma CUSUM chart is 5.2% higher where as the average time to signal is about two days longer than required for the best CQC-v.

[6]Paper : A Customer Churn Prediction Model in Telecom Industry Using Boosting

Author : Ning Lu, Hua Lin, Jie Lu and Guangquan Zhang

Publication Year : 2012

Description:

This research reveals a real-world study on customer churn prediction and proposes the use of boosting to enhance a customer churn prediction model. This paper helps to separate customers into two different clusters depending on the weight assigned by the boosting method. As a solution, a higher risk customer cluster has been identified. In this research, Logistic regression is used as a basis learner, and a churn prediction model is built on each cluster. Comparison of the result is done with a single logistic regression model. Practical analysis tells that boosting provides a good differentiation of churn data so, boosting is suggested for churn prediction analysis.

[7] Paper : Preventing Customer Churn by Using Random Forests Modeling

Author : Weiyun Ying , Xiu Li , Yaya Xie

Publication Year : 2008

Description :

In this paper , improved balance random forests(IBRF) is used to predict the customer churn , where as integrating a sampling technique and cost-sensitive learning into standard random forest to achieve a better performance than most existing algorithms. algorithms like CWC , SVM, artificial neural networks, decision trees evaluated and compared in order to examine their characteristics. The sampling plan and data processing are also explained in detail.

IV. EXISTING SYSTEM

In order to keep customers on board with businesses and boost their profitability, churn prediction from customers became a crucial component in the telecommunications sector. Due of this, it is more important than ever to have an accurate, clear, and reasonable churn forecast system. The domain experts make recommendations based on this computation, taking into account both projected churn and non-churn clients. To increase profits, a shareholder or businessperson must predict which customers will go. The dataset is iterated for each of these procedures in the current system, and the best iterated samples are used. There are now four different dataset sets used in the system. Their view is that the effectiveness of the best employees is what matters most. Churn prediction is one of the best methods for gathering customer input and also generates offers to keep customers happy. Based on their purchases, consumers are divided into churn and non-churn customers according to the churn projection. The vast majority of datasets undergo pre-processing after being gathered from churning customers. Prior to sorting, data items is prepared by being processed and normalized. In order for each sample to be independently estimated using the LDT and UDT procedures, the samples are then divided into smaller groups (Lower Distance Zone and Upper Distance Zone).

The data is also divided and used for training. The classifier receives the LDT, UDT, and training process results. On the basis of LDT and UDT techniques, the classifier provides us with various outcomes. This should be done with smaller datasets, which have an efficiency of 60% to 70%. The suggested approach uses both of the Random Forest classifier and support vector machine to increase efficiency over the current system.

V. METHODOLOGY

The Breiman invented random forests approach outperforms many other classifiers by adding an extra layer of randomization to bootstrap aggregating (known as “Bagging”). It is extremely user-friendly and strong against overfitting. With random forests, subsets of m predictors are chosen at random to construct trees, with each tree being formed on a bootstrap sample of the training set. The number m , which is substantially less than entire number of variables available for analysis is used to divide the nodes. Since every tree in the forest has the same distribution and is dependent on values from a randomly sampled vector. The dataset for predicting customer turnover is an example of a dataset. Two solutions are there for random forest’s imbalanced data classification problem that is balanced random forest which is based on sampling technique while weighted random forest is based on cost sensitive learning. Both techniques outperform the current algorithms and increase the minority class’s forecast accuracy. They do, however, have some drawbacks. By merging balanced random forest and weighted forest we can obtain the enhanced balanced random forest.

VI. FLOW DIAGRAM

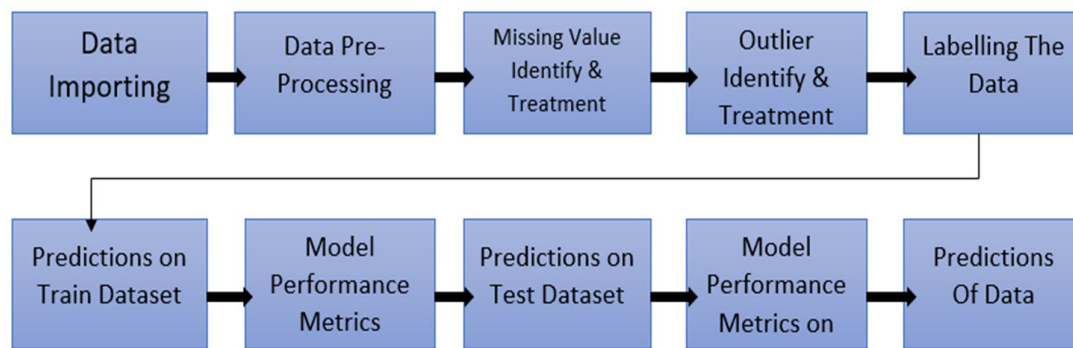


Fig 1. System flow diagram

VII. CONCLUSION

In this article, we provide an original approach called IBRF for foretelling churn in the banking sector. IBRF has the advantage of changing the class distribution and penalizing minority class misclassification more severely by combining a sampling technique with cost-sensitive learning. By artificially equating class priors, which the best features are then iteratively trained from, the best weak classifiers are produced. Our approach outperforms prior random forest algorithms like balanced random forests and weighted random forests, according on experimental results on bank databases. In terms of top-decile lift, IBRF exceeds ANN, DT, and CWC-SVM. Compared to conventional methodologies, IBRF offers a lot of promise because of its scalability and faster training and running speeds. Current research should improve efficacy and generalizability. IBRF

uses internal variables to analyze sample distribution. Although it has been demonstrated that the findings are not sensitive to the values of these variables, reducing their values in future testing may enhance the method's ability to more accurately predict results. Here, key objective of the paper is to develop a unique Customer churn prediction model which help to predict customers who are most likely to churn and therefore warnings can help to take measurable actions to retain them. Here, we evaluated and examined the performance of various tree-based machine learning algorithms and methods and identified the Random Forest Classifier as the most optimal solution to Customer churn prediction problem.

REFERENCES

- [1] S Neslin, S Gupta, W Kamakura, J Lu, C Mason, “Defection Detection: Improving Predictive Accuracy of Customer Churn Models”, Working Paper, Teradata Center at Duke University, 2004.3
- [2] Wells, Melanie, “Brand ads should target existing customers” Advertising Age, pp26-47, 1993.
- [3] CP Wei, IT Chiu, “Turning telecommunications call details to churn prediction: a data mining approach”, Expert Systems With Applications, Vol. 23, Issue 2, pp103-112, 2002.
- [4] Mozer, M. C., Wolniewicz, R., Grimes, D. B., etc, “Churn reduction in the wireless industry”, Advances in Neural Information Processing Systems, pp935- 941, 2000(12).

- [5] Eiben, A.E.; Koudijs, A.E.; Slisser, F. Source, "Genetic modeling of customer retention", *Lecture Notes in Computer Science*, Vol. 1391, pp178, 1998
- [6] J. R. Quinlan, "Decision trees as probabilistic classifiers", *Proc. 4th Int. Workshop Machine Learning*, pp31–37, Irvine, CA, 1987.
- [7] Wai-Ho Au, Keith C. C. Chan, and Xin Yao, "A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction", *IEEE Transactions on Evolutionary Computation*, Vol. 7, Issue 6, pp532-545, 2003.
- [8] M.A.H. Farquad, Vadlamani Ravi and S. Bapi Raju, "Churn prediction using comprehensible support vector machine: An analytical CRM application", *Applied Soft Computing*, vol. 19, pp. 31–40, 2014.
- [9] Chih-Fong Tsai and Yu-Hsin Lu, "Customer churn prediction by hybrid neural networks", *Expert Systems with Applications*, vol. 36, pp. 12547-12553, 2009.
- [10] Ning Lu, Hua Lin, Jie Lu and Guangquan Zhang, "A Customer Churn Prediction Model in Telecom Industry Using Boosting", *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, may 2014.
- [11] Ssu-Han Chen, "The gamma CUSUM chart method for online customer churn prediction", *Electronic Commerce Research and Applications*, vol. 17, pp. 99-111, 2016.
- [12] Benlan He, Yong Shi, Qian Wan and Xi Zhao, "Prediction of customer attrition of commercial banks based on SVM model", *Proceedings of 2nd International Conference on Information Technology and Quantitative Management (ITQM) Procedia Computer Science*, vol. 31, pp. 423-430, 2014.