International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 25th Jun 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 05

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 05/22

Title **CUSTOMER CHURN PREDICTION USING MACHINE LEARNING**

Volume 11, SPL ISSUE 05, Pages: 144-150

Paper Authors

Dr. Satyabrata Dash , Moka Neeraja, Sepeni Rahul, Putti Santhi Priyanka, Sugasani Trinath

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

Dr. Satyabrata Dash[1] , Moka Neeraja[2], Sepeni Rahul[3], Putti Santhi Priyanka[1], Sugasani Trinath[5]
[1]AssociateProfessor, Dept. of CSE ,[2]18ME1A0563,[3]18ME1A05A1, [4]19ME5A0503, [5]18ME1A05A6
Ramachandra College of Engineering, A.P.India
satyabrata.cse@rcee.ac.in,mokaneeraja@gmail.com,sepenirahul@gmail.com,santhipriyankaputti@gmail.com,trinathsugasani0811999@gmail.com

Abstract
All over the world, in different sectors churn prediction plays a very important role in the growth of the organization. For the company's revenue and profit, customer churn is very harmful. The most important step to avoid churn is to detect churn and its reason, accordingly initiate the prevention measures. Nowadays machine learning plays a vital role to get rid of this problem. The objective of this project is to predict the churn in banking sectors, by using well known machine learning techniques like Logistic Regression. The classification model is built by analyzing historical data and then applying the prediction model based on the analysis. Withholding the customer in an organization is one of the primary growth in today's world. Predicting the customer churn rate will help the bank in knowing which category of customers generally tend to leave the bank. Churn is based on various factors, including changing the position to a competitor, canceling their subscription because of poor customer service, or discontinuing all contact with a brand due to insufficient interaction between customers. Being connected for a long period of time with customers is more effective than trying to attract new customers. Dealing to figure out the amiss issues can make the customers happy. In this project finding the major factors that affect the customers to churn and analyze them by using Machine learning algorithms. Then churn gives the information of how many existing customers tend to leave the business, so lowering churn has an immense positive impact on the revenue streams. On the basis of this the Churn rates track lost customers, and growth rates track new customers comparing and analyzing both of these metrics tells exactly how much the business is growing over time. In this predictive process popular models have been used to achieve a decent level of accuracy.

Keywords—Machine Learning(ML), Logistic Regression.

Introduction

To get and keep loyal customers for every business organization is a big challenge. Correct prediction about a customer is going to churn or not and then successfully convincing him to stay with that company can increase the revenue of that company. Therefore, predicting customer churn, i.e. if a customer is about to leave for a better service, that is an important part for analyzing the customer behavior. The churn model is a representation

of various calculations that are built on existing historical data. The customer churn can be defined in other ways also, like low switching cost, deregulation motivates a customer to replace the sector. The churn is also classified into two: voluntary and involuntary churn Voluntary churn is defined as the termination of services by the customer itself, whereas involuntary churn is defined as the termination of services by the bank for fraud, non-payment services.

The customer churn is very risky if it is not managed carefully, it may bring a company to its knees for poor maintenance services. Cost of customer churn also includes loss of revenue, profit. Previous case study has shown that the cost of maintaining a new customer is higher than the cost of maintaining the old one. There are various banks who are suffering with this customer churn problem. So the most defined way to deal with this problem is developing a predictive model that can reliably and easily identify the possible churner. In the recent past the most frequently used technique is data mining to develop such models with satisfactory results Day by day when this churn prediction problem gets importance, much more research efforts are generated towards improving churn prediction rates.

Most frequently used features that have been used previously include credit score, geography, having a credit card or not, active member, estimated salary Due to business privacy and policy it is difficult to find and use public dataset for churn prediction In this project, a new subset of features has been presented in order to improve the prediction accuracy. Logistic Regression has been used to predict the results. Customer churn prediction is the technique of allocating a likelihood of churn to each customer in the corporate database based on an anticipated correlation between that customer's historical data and its anticipated future churning behaviour.
In practise, the likelihood that a customer will leave the organisation is used to rank customers from most to least likely to churn, and those with the highest propensity to do so are subjected to marketing initiatives aimed at keeping them there. It reveals the behavior of the accuracy of the models' predictions in comparison to the actual outcomes of those consumers over a period of time.

The interests of the client have been prioritised by all reputable organisations.
Due to strong competition among service providers, customers have a variety of options, and the best services are available without end. The biggest obstacles to attracting new clients are a lack of data, targeted marketing, and company modernization. It has been discovered that customer value and rising revenue are what influence current customers' retention rather than gaining new ones.

The key to boosting profits and customer value is for businesses to get to know their current customers well, build close relationships with them, and collect a tonne of data about them. It is very important to find whether the customer will churn in the near future or stay within the bank, Which affects the revenue streams of the bank.

Related Work

The phrase has a broad use in the financial sector and is today used in a number of various business domains.

When a consumer stops using their credit card within a predetermined duration, it is known as credit card churn. A consumer who ceases using a network bank's online (home banking) service is referred to as a churning customer. Chiang, Wang, Lee, and Lin (2003) studied this subject by tracking the users' transaction times on a regular basis.

Additionally, Glady et al. (2008) established two distinct definitions of churn that exist in the organisation that will be the subject of this paper: the idea of voluntary churn and involuntary churn. Glady et al. (2008) defined a churner as a customer with less than 2.500 Euros of assets at the bank (savings, securities, or other types of products).
In order to identify certain trends that might suggest a customer is at risk of churning, the current study will be especially focused on tracking the behavioural history of previous churners within a specific time frame.

Churning is a significant issue that has been researched in a number of fields, including mobile and phone, insurance, and healthcare, as noted by Oyeniyi and Adeyemo (2015). Online social network churn study and retail banking are two such industries where the customer churn issue has been studied.The concept of churn must be examined in light of the context in which it is used, even though the broadest or most often accepted definition refers to a company losing a customer.

Customer attrition was defined by Eichinger, Nauck, and Klawonn (2006) as when a customer switches to a competitor. Qiasi, Roozbehani, and Minaei-bidgoli (2002) support this idea by defining churn as the process by which a client stops using an organization's goods and services in favour of those offered by a rival.

On the other side, customer churn was defined as the tendency by Neslin et al. (2006).

To address the shortcomings of general SVM models that produce black box models, M.A.H. Farquad [4] presented a hybrid technique (i.e., it does not reveal the knowledge gained during training in human understandable form). Three phases make up the hybrid approach: SVM-RFE (SVM-recursive feature elimination) is used in the first step to condense the feature set. In the second stage, a dataset with less characteristics is utilised to create an SVM model and extract support vectors. Naive Bayes Tree is then used to generate rules in the final phase (NBTree which is combination of Decision tree with naive Bayesian Classifier).

The dataset employed in this study is the bank credit card customer dataset (Business Intelligence Cup 2004), which has a highly imbalanced client retention rate of 93.24 percent and a customer turnover rate of 6.76 percent. The experimental results demonstrated that the model is not scalable to big data sets.

Wouter Verbeke [6] proposed the deployment of Ant-Miner+ and ALBA algorithms on a publically available churn prediction dataset in order to develop accurate as well as clear classification rule-sets churn prediction models. Ant-Miner+ is a high performing data mining method based on the principles of Ant Colony Optimization which allows to add domain knowledge by placing monotonicity restrictions on the final rule-set. High accuracy, understandability of the created models, and the ability to request intuitive prediction models are all benefits of Ant-Miner. A rule extraction approach called Active Learning Based Approach (ALBA) for SVM rule

extraction combines the ruleset format's readability with a non-linear support vector machine model's high prediction accuracy.

Results that are benchmarked against C4.5, RIPPER, SVM, and logistic regression reveal that ALBA paired with RIPPER yields the highest accuracy, while C4.5 and RIPPER used on an oversampled dataset produce the maximum sensitivity. Ant-Miner produces understandable rule sets that are far smaller than the rule sets generated by C4.5, but it also produces less sensitive rule sets that allow for the inclusion of domain knowledge. The rule sets produced by RIPPER are also concise and understandable, but they produce models that defy domain knowledge and are not intuitive.

Customers are divided into two clusters based on the weights supplied by the boosting algorithm in Ning Lu's proposal to employ boosting algorithms to improve a customer churn prediction model. A high-risk consumer cluster has been identified as a result. A churn prediction model is created for each cluster individually using logistic regression as a basic learner. When compared to a single logistic regression model, the testing findings shown that boosting algorithm effectively separates churn data.

By taking into account the imbalance characteristics of customer data sets, Benlan proposed a methodology for predicting customer attrition based on the SVM model. In a high- or infinite-dimensional space, a support vector machine creates a hyper-plane that can be utilised for classification.

To alter the data's distribution and lessen the dataset's imbalance, random sampling can be performed. Due to the low percentage of churners, the dataset is unbalanced.

By introducing a finite mixture model to design the reference value and decision interval of the chart and by using a hierarchical Bayesian model to capture the heterogeneity of customers, Ssu-Han Chen developed a novel mechanism based on the gamma Cumulative SUM (CUSUM) chart that monitors each individual customer's Inter Arrival Time (IAT). The model incorporates Recency, a second time interval variable that complements IAT and tracks the most recent state of the login behavior. The graphical interface for every customer is another benefit of the suggested method, in addition to advantages from the fundamental nature of control charts. The results showed that the accuracy rate (ACC) for gamma CUSUM chart is 5.2% higher than exponential CUSUM and the Average Time to Signal (ATS) is about two days longer than required for exponential CUSUM.

There are various banks who are suffering with this customer churn problem. So the most defined way to deal with this problem is developing a predictive model that can reliably and easily identify the possible churner. In the recent past the most frequently used technique is data mining to develop such models with satisfactory results Day by day when this churn prediction problem gets importance, much more research efforts are generated towards improving churn prediction rates. The interests of the client have been prioritised by all reputable organisations. Due to strong competition among service providers, customers have a variety of options, and the best services are available without end. The biggest obstacles to attracting new clients are a lack of data, targeted marketing, and company modernization. It has been discovered that

customer value and rising revenue are what influence current customers' retention rather than gaining new ones.

Rotation Forest and Rotboost are two rotation-based ensemble classifiers that Koen W. De Bock proposed as modelling approaches for customer churn prediction. An ensemble classifier aggregates many member classifier models into a single model, using the fusion rule to integrate the outputs of the member classifiers.
RotBoost combines Rotation Forest with AdaBoost, whereas in Rotation Forests feature extraction is done to feature subsets in order to transform the input data for training base classifiers.
We use four data sets from actual customer churn prediction initiatives.
The findings revealed that RotBoost outperforms Rotation Forests in terms of top-decile lift and area under the curve (AUC), but RotBoost performs better in terms of accuracy. On the classification performance of both RotBoost and Rotation Forest, they also contrasted the three alternative feature extraction algorithms Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Sparse Random Projections (SRP).
In general, the applied feature extraction technique and the performance criteria used to quantify classification performance determine how well a rotation-based ensemble classifier performs. By utilising a Partial Least Squares (PLS) based method on highly correlated data sets among variables, Lee et al. concentrated on developing an accurate and concise predictive model with the objective of churn prediction. They deploy a simple but effective churn marketing programme in addition to presenting a prediction model to precisely forecast customers' churning behaviour. The suggested methodology enables the marketing managers to successfully and efficiently maintain an optimal (or at least a near optimal) level of churners through the marketing campaigns. PLS is used in this instance as the predictive modelling technique.

Methods

Logistic Regression

The generalised linear regression model includes the binary classification procedure known as logistic regression. Logistic regression is a typical predictive model used for binary categorization and outcome prediction. Additionally, it can be applied to problems involving more than two classes. In order to predict whether a certain customer or a group of customers will stop using the service, a model can be developed using logistic regression and the customer churn data.

Attribute Name	Attribute description
CreditScore	reliability of the customer
Geography	where is the customer from
Gender	Male or Female
Age	Age of the client
Tenure	number of years of customer history in the company
Balance	the money in the bank account
NumOfProducts	Number of products of the customer in the bank

Fig. Attributes

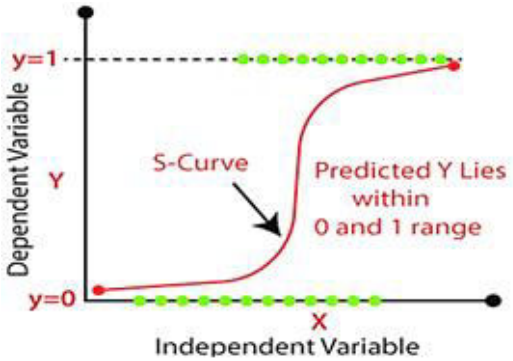

Fig. Logistic Regression

A sigmoid curve can be used to map the relationship between the dependent variable and independent variables.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Implementation

In this Study, a bank data is considered where a huge number of customers are leaving the bank. Almost 10000 records of the bank, collected from Kaggle repository, are going to help the model to investigate and predict which of the customers are about to leave the bank soon. To test and evaluate the features the total dataset sliced into two subsets, training and testing dataset. Training dataset can be used to define the statistical model and the testing dataset can be used to predict the result and calculation of accuracy metrics for determining the model accuracy.

In Fig 1, box plots of some important attributes are given. When it comes to the distribution of all data points over mean, box plots are used to identify the median and also respective qualities in a well-structured manner.

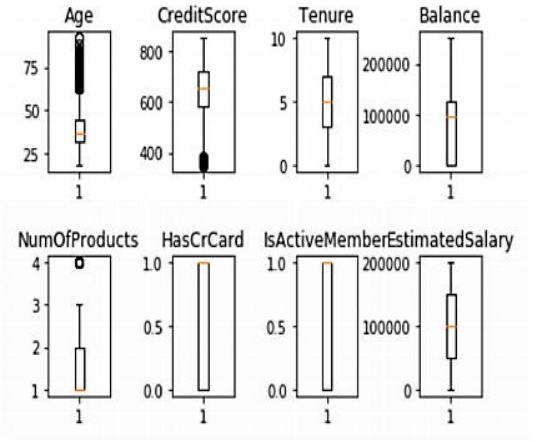

Fig.1. Box-Plots of Important Attributes

Results

The customers of a bank are described in the data set, and the target variable is a binary variable that indicates whether the customer has left the bank (closed his account) or not. It consists of 10,000 records with demographic and Bank history information from customers from three countries, France, Germany and Spain. Continuing with splitting the data into separate training and test sets. 30% of observations will be set aside for the test set the rest, 70%, will be used as the training set. In the output, it displays an application containing the customer details to be filled.The first output is the application given input by the user with the appropriate details to predict the result.

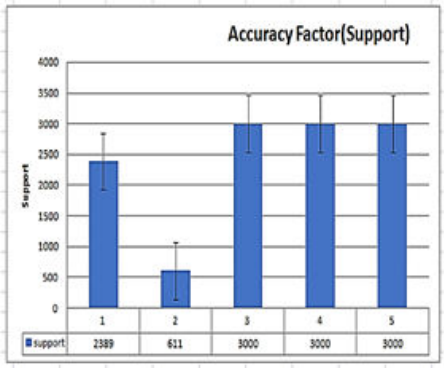

Fig. Accuracy Factor(support)

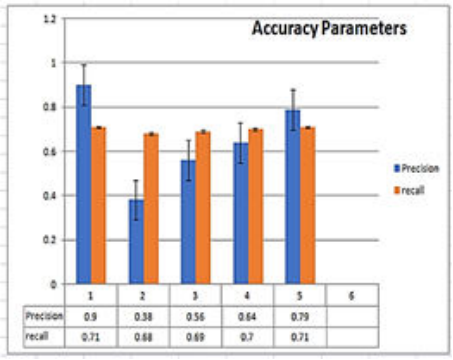

Fig Accuracy Parameter

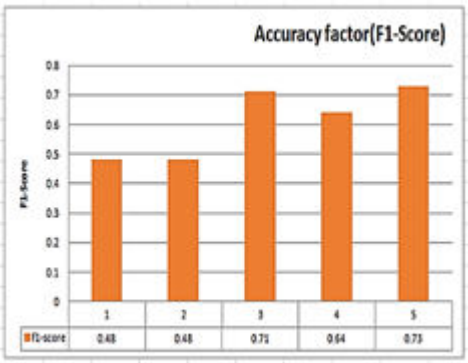

Fig Accuracy Factor(F1-Score)

Conclusion

In this project, we proposed an algorithm that can predict Customer Churn and gives the outcomes, contrasting other used techniques. This algorithm which deals with the large amount of datasets. Based on the given customer details it predicts whether the customer will stays in the bank or left the bank.. It gives an accuracy of 71.00 %. The proposed system would be helpful in predicting if a customer will stays or left the bank. And also helps in the growth of a company.

References

1. Seo D, Ranganathan C, Babad Y. Two-level model of customer retention in the US mobile telecommunications service market. Telecommun Policy 2008; 32 (3):182–96.
2. G. M. Weiss, “Mining with rarity: a unifying framework,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7–19, 2004.
3. J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, 2009.
4. M. T. K. a. B. B. B. Huang, "Customer churn prediction in telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414-1425, 2012.
5. W.-H. a. C. K. C. a. Y. X. Au, "A novel evolutionary data mining algorithm with applications to churn prediction," IEEE transactions on evolutionary computation, vol. 7, no. 6, pp. 532-545, 2003.
6. M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju “Churn prediction using comprehensible support vector machine: An analytical CRM application”, Applied Soft Computing 19 (2014) 31–40.
7. Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens “Building comprehensible customer churn prediction models with advanced rule induction techniques”, Expert Systems with Applications 38 (2011) 2354–2364.
8. Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang “A Customer Churn Prediction Model in Telecom Industry Using Boosting”, IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2014.
9. Benlan He, Yong Shi, Qian Wan, 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 31 (2014) 423 – 430.
10. https://www.kaggle.com/filippoo/deep-learning-az-annn
11. Chu, B. H., Tsai, M. S., and Ho, C. S., "Towards a hybrid data mining model for customer retention", Knowledge-Based Systems, 20, 2007, pp. 703–718.
12. Shui Hua Han a, ShuiXiuLu a, Stephen C.H. Leung., "Segmentation of telecom customers based on customer value by decision tree model", Expert Systems with Applications, 39, 2012, 3964–3973
13. https://learnpython.com/blog/python-customer-churn-prediction/
14. https://neptune.ai/blog/how-to-implement-customer-churn-prediction.
15. Koen W. De Bock, Dirk Van den Poel, “An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction”, Expert Systems with Applications 38 (2011) 12293–12301.