INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 0975-4520

DEVELOPING A PREDICTION MODEL FOR CUSTOMER CHURN FROM ELECTRONIC BANKING SERVICES USING DATA MINING

PADMASRI ARDHALA[1], AKHILA MAGULURI[2], DEEPTHI ATMAKURI[3], GAYATHRI KURAPATI[4], P.VENU BABU[5]

[1, 2, 3]UG Scholars, Department of CSE, ***MALINENI LAKSHMAIAH WOMEN'S ENGINEERING COLLEGE, GUNTUR***, India.

[4] Assistant Professor, Department of CSE, ***MALINENI LAKSHMAIAH WOMEN'S ENGINEERING COLLEGE, GUNTUR***, India.

ABSTRACT:

A new method for customer churn analysis and prediction has been proposed. The method uses data mining model in banking industries. This has been inspired by the fact that there are around 1,5 million churn customers in a year which is increasing every year. Churn customer prediction is an activity carried out to predict whether the customer will leave the company or not. One way to predict this customer churn is to use a classification technique from data mining that produces a machine learning model. This study tested 5 different classification methods with a dataset consisting of 57 attributes. Experiments were carried out several times using comparisons between different classes. Support Vector Machine (SVM) with a comparison of 50:50 Class sampling data is the best method for predicting churn customers at a private bank in Indonesia. The results of this modeling can be utilized by company who will apply strategic action to prevent customer churn.

Keywords: SVM, churn, Banking.

1. INTRODUCTION:

ISSN : 0975-4520

Our case study (XYZ Bank) is one of the largest banks in Indonesia with dozens of millions of customers who must be considered well so that they want to continue to use the facilities provided by the company. Companies have realized that they must strive not only to get new customers but also to retain existing customers because if existing customers become churn customers, the number of customers will decrease if there are no more new customers. At our case study (XYZ Bank), there are around 1,5 million churn customer in a year and increasing every year Although it can have an impact on the decline of new customers, to get new customers costs five to six times greater than retaining existing customers [1],[2]. Some techniques can be done to defend old customers, which is to predict customers who will churn. Predicting churn customers aims to identify prospective churn customers based on past information and previous behaviour so that incentives can be offered to survive. Data analysis can be described as an in-depth examination of the meaning and important values available in the data to identify important information using specific methods and techniques [3]. One technique that can be used is data mining techniques. Some previous research[1],[4],[5],[6] many have shown that data mining techniques can be used to predict churn customers. The purpose of this study is to obtain the best data mining learning model that can be implemented by XYZ Bank to prevent customers from leaving them.

2 RELATED STUDY

Churn prediction is the process of using transaction data to identify customers who are likely to cancel their subscriptions. Popular among many service industries like telecommunication, finance and e-commerce, churn prediction can help organizations to prevent loss of future revenue and to increase customer

ISSN : 0975-4520

loyalty (NGData 2013, Fiworks 2019). To date, most work in churn prediction focuses on sampling strategies, feature engineering and supervised modeling over a fixed period of time. Predictors used in existing work, such as user demographics and credit status, are static in nature and are not adequate to provide accurate prediction (Ali and Ariturk 2014). Few studies have explored the area of using dynamic predictors to mine customer behavior over longitudinal data. This research aims to develop and validate a dynamic classification approach to accurately capture customer behavior for churn prediction. The approach considers dynamic time-series predictors, multiple time periods, and rare event detection and modeling to enable accurate and long-term dynamic prediction. The approach was applied to churn prediction on a unique three-year, fine-grained dataset consisting of 32,000 transaction records provided by a retail bank located in Florida, USA. The banking industry needs churn prediction because each lost customer costs on average $750 while adding new customers is expensive (acquiring a new customer costs $200 on average) in today's saturated market (Fiworks 2019). Using the approach, the study finds that its trend modeling helped to capture the change of customer behavior over multiple periods of time. The approach's rare event detection and its use of data from dynamic time periods to aggregate the number of churn cases helped to improve model precision and recall. The empirical results demonstrate a strong potential to improve profitability of businesses that need accurate and scalable churn prediction. The research provides a useful approach to optimizing churn prediction modeling and a unique case study and empirical findings for the banking industry.

3. PROPOSED SYSTEM:

ISSN : 0975-4520

Regarding modeling approaches, the performance of seven statistical modeling approaches were compared in churn prediction (Umayaparvathi and Iyakutti 2016): logistic regression, k-nearest neighbor, random forest, support vector machine, ridge classifier, decision tree and gradient boosting. Gradient boosting and random forest outperformed other approaches in terms of accuracy, precision and recall. Similar results were shown in analyzing participants in a churn prediction analytics tournament, logistic regression and tree methods performed better than other methods like discriminant analysis and clustering (Neslin et al. 2006). Modified models like using lasso regression to extract features that were highly correlated with churn, then fed those features into radial basis function neural network (RBF). This method resulted in better performance than using RBF, Log-R or boosting alone (Xiong et al. 2019). Another modified model was improved balanced random forest (IBRF) which combined cost-sensitive learning with random forest to alter class distribution and penalized more heavily on misclassification of the minority class. This technique showed improvement over other methods like SVM, random forest and decision tree (Ying et al. 2008). However, these two models are computationally intensive, were applied to small datasets without considering the temporal dimension, and may not be suitable for large, dynamic transaction data used by banks. Furthermore, these processes are complicated with low interpretability. The banking industry is highly regulated and transparency is essential. There is a need to provide a clear and efficient way to process large volume of longitudinal data.

The six phases can be explained as follows:

ISSN : 0975-4520

1) Business understanding focuses on understanding the objectives of the research and the views of the business.

2) Data understanding to understand the data to be used.

3) Data preparation includes all activities to build the final dataset of raw data.

4) Modeling for processing data using selected data mining methods and conducting experiments with parameters calibrated to optimal values.

5) Evaluation is evaluating all the steps implemented to build the model reviewed and ensuring its business objectives are achieved.

6) Deployment uses the model obtained for existing business processes and also the development of the model so that it remains valid to be used onwards.

RESULTS:

Some data must be cleaned so as not to cause performance values from modeling to decrease. The data that is cleaned up is VINTAGE data where there is data that has a value below zero. This can be called dirty data because basically the range of a person being a customer must start from zero years whereas in this data it starts from minus 1. This study tries to use three types of class comparison the data is by Stratified Sampling of 1%, 50% versus 50% and 30% versus 70% to see which comparison produces the highest value of accuracy and sensitivity. This is made a comparison because not always the ratio of 50% to 50% is the best choice.

METODE	SAMPLING	ACCURACY	RECALL	PRECISION	AUC	VALIDATION TIME (dalam detik)
DECISION TREE	Stratified	91,58%	13,73%	61,05%	0,735	60
	50:50	69,04%	70,58%	16,39%	0,745	59
	30:70	88,47%	39,09%	31,94%	0,759	60
NEURAL NETWORK	Stratified	89,47%	34,61%	34,02%	0,780	122
	50:50	70,79%	69,35%	17,05%	0,769	166
	30:70	81,64%	50,60%	21,79%	0,767	226
SVM	Stratified	**92,65%**	13,92%	**68,41%**	0,750	494
	50:50	73,68%	**73,24%**	19,38%	0,811	580
	30:70	89,17%	45,01%	35,63%	0,804	595
NAÏVE BAYES	Stratified	79,65%	60,27%	21,79%	0,793	63
	50:50	76,33%	65,72%	19,96%	0,795	**55**
	30:70	78,58%	62,43%	21,20%	0,795	**55**
LOGISTIC REGRESSION	Stratified	92,18%	21,03%	52,05%	0,804	64
	50:50	74,57%	**72,68%**	19,89%	0,815	58
	30:70	89,65%	40,52%	36,41%	**0,816**	61

In this study there are several attributes that are not used due to the high value of stability. This stability

ISSN : 0975-4520

value is obtained by looking at the distribution of data whether it is evenly distributed in each class used. If this attribute is considered to have a high stability value or in each class the value of this attribute is the same then these attributes will be excluded from the modeling because it is considered not to affect because all data has the same value.

4. CONCLUSION:

The use of data mining is proven to be used in predicting customer churn in the banking business. This research produces several conclusions such as:

1) The number of samples of data used for learning greatly influences the results of modeling. The number of inter-class comparisons greatly influences the recall results where the comparison of the 50:50 data will result in a greater recall value (average 70%) compared to the other two settings. In this study using around 15.949 samples of data so for each class around 7.975 samples of data. Accuracy values cannot be fully used as a reference for comparison if the distribution of data is very unbalanced.

2) The best model is the model with the highest profit value, namely the 50:50 SVM sampling model with a profit value of 456 billion with loss and benefit calculations, with the five most significant attributes is vintage, volume of EDC (Electronic Data Capture) transaction, amount of EDC (Electronic Data Capture) transaction, average balance in one month and age. This is in line with the research of Dolatabadi et al. which obtained SVM as modeling with the best accuracy in its research, but Logistic Regression is also worth considering because it results in smaller losses.

REFERENCES:

[1] Oyeniyi, A., & Adeyemo, A. (2015). Customer churn analysis in banking sector using data mining techniques. African Journal of Computing & ICT Vol 8, 165-174.

ISSN : 0975-4520

[2] Peng, S., Xin, G., Yunpeng, Z., & Ziyan, W. (2013). Analytical model of customer churn based on bayesian network. Ninth International Conference on Computational Intelligence and Security (pp. 269-271). IEEE.

[3] Banarescu, A. (2015). Detecting and preventing fraud with data nalytics. ScienceDirect, 2.

[4] Dolatabadi, S. H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. The 2nd International Conference on Computer and Communication Systems (pp. 74-77). IEEE.

[5] Keramati, A., Ghaneei, H., & Mohammad Mirmohammadi, S. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation, 2-10.

[6] Zoric, A. B. (2016). Predicting customer churn in banking industry using neural networks. Interdisciplinary Description of Complex Systems, 116-124.

[7] Chitra, K., & Subashini, B. (2011). Customer retention in banking sector using predictive data mining technique. International Conference on Information Technology. ICIT.

[8] De Caigny, A., Coussement, K., & W. De Bock, K. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269, 760-772.

[9] Larose, D. T. (2006). Data mining methods and models. New Jersey: John Wiley & Sons, Inc.

[10] Turban, E., Aronson, J. E., & Liang, T.-P. (2005). Decision support systems and intelligent systems. New Jersey: Pearson Education, Inc.

[11] Han, j., Kamber, M., & Jian, P. (2012). Data mining concepts and techniques third edition. Morgan Kaufmann Publishers.